GWAS identifies candidate regulators of *in planta* regeneration in *Populus trichocarpa*

Michael F. Nagle[1#], Jialin Yuan[2], Damanpreet Kaur[2], Cathleen Ma[1], Ekaterina Peremyslova[1], Yuan Jiang[3], Christopher J. Willig, Greg S. Goralogia, Alexa Niño de Rivera[1], Megan McEldowney[1], Amanda Goddard[1], Anna Magnuson[1], Wellington Muchero[4], Li Fuxin[2], Steven H. Strauss[1@]

[1] Department of Forest Ecosystems and Society, Oregon State University, Corvallis, OR, USA

[2] Department of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA

[3] Statistics Department, Oregon State University, Corvallis, OR, USA

[4] Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

@ Correspondence to steve.strauss@oregonstate.edu

Co-correspondence to michael.nagle@oregonstate.edu

Keywords: GWAS, poplar, *Populus trichocarpa*, regeneration, transformation, SNP, callus, shoot, transformation

Summary

Plant regeneration is an element of natural and horticultural plant propagation, and a key step in the production of transgenic plants. However, regeneration capacity varies widely among genotypes and species, the molecular basis of which is largely unknown. To shed light on the causes of variation in natural regeneration capacity, we undertook a GWAS of shoot regeneration from dormant cut stems in *Populus trichocarpa*. Using estimates of callus and shoot regeneration provided by a novel computer vision system, and using a variety of GWAS pipelines and statistical approaches, our analyses revealed over 200 candidate genes. The candidates each explained small fractions of the total genetic variance and many appeared to be members of genetic regulatory networks, showing regeneration to be a highly polygenic trait. The top candidates included regulators of cell adhesion, stress signaling, and hormone signaling pathways, as well as other diverse functions**.** These candidates provide new insights into the biological complexity of plant regeneration, and may serve as new reagents for improving regeneration and transformation of recalcitrant genotypes and species.

Introduction

Plant genetic engineering and gene editing have produced new varieties of crops with a variety of valuable traits (NAS, 2016; Jaganathan *et al.*, 2018). However, the ability to impart new traits by these methods is limited to crop species with genotypes that can reliably undergo regeneration and transformation (RT). RT requires developmental responses to a series of hormone treatments and amenability to gene insertion, and the capacity for both varies greatly between and within species (Altpeter *et al.*, 2016). The causes of this great variation in recalcitrance are poorly known; however, GWAS – with its potential to identify genes whose variation plays a key role in capacity for RT – should greatly enhance understanding of the RT process. In addition, the identified genes may serve as "reagents" for overcoming recalcitrance, similar to how overexpression of morphogenic regulator (MR) genes can enhance *in vitro* regeneration of transgenic shoot in a variety of species (Gordon-Kamm *et al.*, 2019). *In planta* transformation methods can also be enhanced by MR genes, including in *Populus tomentosa* (Deng *et al.*, 2009), *Nicotiana benthamiana*, tomato, potato and grape (Maher *et al.*, 2020). However, given the complexity and genotypic variation in RT capacity, it is likely that only a fraction of the potentially useful MR genes have been identified.

To help identify the genes responsible for variation in RT, we conducted GWAS in a population of 1,219 wild cottonwoods that had been resequenced by the US Department of Energy, up to 917 of which were previously studied for a variety of traits (Zhang *et al.*, 2018a; Tuskan *et al.*, 2018; Muchero *et al.*, 2018; Bdeir *et al.*, 2019; Chhetri *et al.*, 2020). We focused on regeneration from cut stems, while considering it may be a direct substrate for accelerated *in planta* transformation systems, and because of the expectation that regeneration processes are likely to share many elements whether induced *in vivo* or *in vitro*. GWAS has previously been applied to study variation in the rate of *in vitro* regeneration in Arabidopsis, cotton, wheat, sorghum and poplar (reviewed by Lardon et al. 2020).

Regeneration phenotypes are notoriously difficult to quantify, whether *in vivo* or *in vitro*. Calli and emerging shoots are often highly variable and complex in shape, color, and size, and sequential measurements are hard to take without damaging or contaminating regenerating tissues. This appears to have limited sample sizes in prior GWAS studies of regeneration. For example, Tuskan et al. (2018) selected only 280 genotypes to phenotype callus growth from a resequenced GWAS population of 1,084 *P. trichocarpa* genotypes. A similar GWAS of callus dedifferentiation into shoots in *P. euphratica* was limited to 297 genotypes (Zhang *et al.*, 2020). Nguyen et al. (2020) noted the "extremely laborious" nature of phenotyping *in vitro* traits as a constraint in their GWAS of callus formation across 96 rose genotypes (Nguyen *et al.*, 2020).

Because of the importance of a large and precise sample for statistical power in GWAS (López-Cortegano & Caballero, 2019), we developed a computer vision (CV) method to measure regeneration from sequential images of cut, regenerating stems. Over 40 published studies have made use of diverse CV methods in GWAS of plants, including Arabidopsis, maize, wheat, rice, sorghum, soybean and barley (reviewed by Xiao *et al.*, 2021). They have used high-throughput scanners and thresholding to phenotype leaf traits such as size, shape, and color (Yang *et al.*, 2015), and employed a wide range of sensors (e.g., RGB, hyperspectral, CT, infrared) and algorithms (e.g., thresholding-based methods, support vector machines, and neural networks). In recent years neural networks similar to those employed in the present study have outperformed earlier methods and become the dominant approach used for diverse CV tasks. In the context of plant phenotyping, this was demonstrated by the unparalleled performance of neural networks for the Leaf Segmentation Challenge benchmark dataset (Aich & Stavness, 2017; Dobrescu *et al.*, 2017).

Here, we report identities of numerous potential regulators of *in planta* regeneration in *Populus trichocarpa* through application of several GWAS pipelines. We employed a population with over 1,200 wild genotypes whose SNPs display extremely low linkage disequilibrium (LD), used a very high number and density of SNP markers (up 34 million depending on GWAS method), and phenotyped regeneration precisely using a high-throughput CV pipeline. We report a large number of statistically-supported gene candidates with diverse physiological roles that include hormone signaling, plant stress response, control of cell division, and cell wall structure – as well as many genes whose function is yet to be determined.

Materials and Methods

An overview of the experimental population and analysis pipeline is shown in Fig. 1.

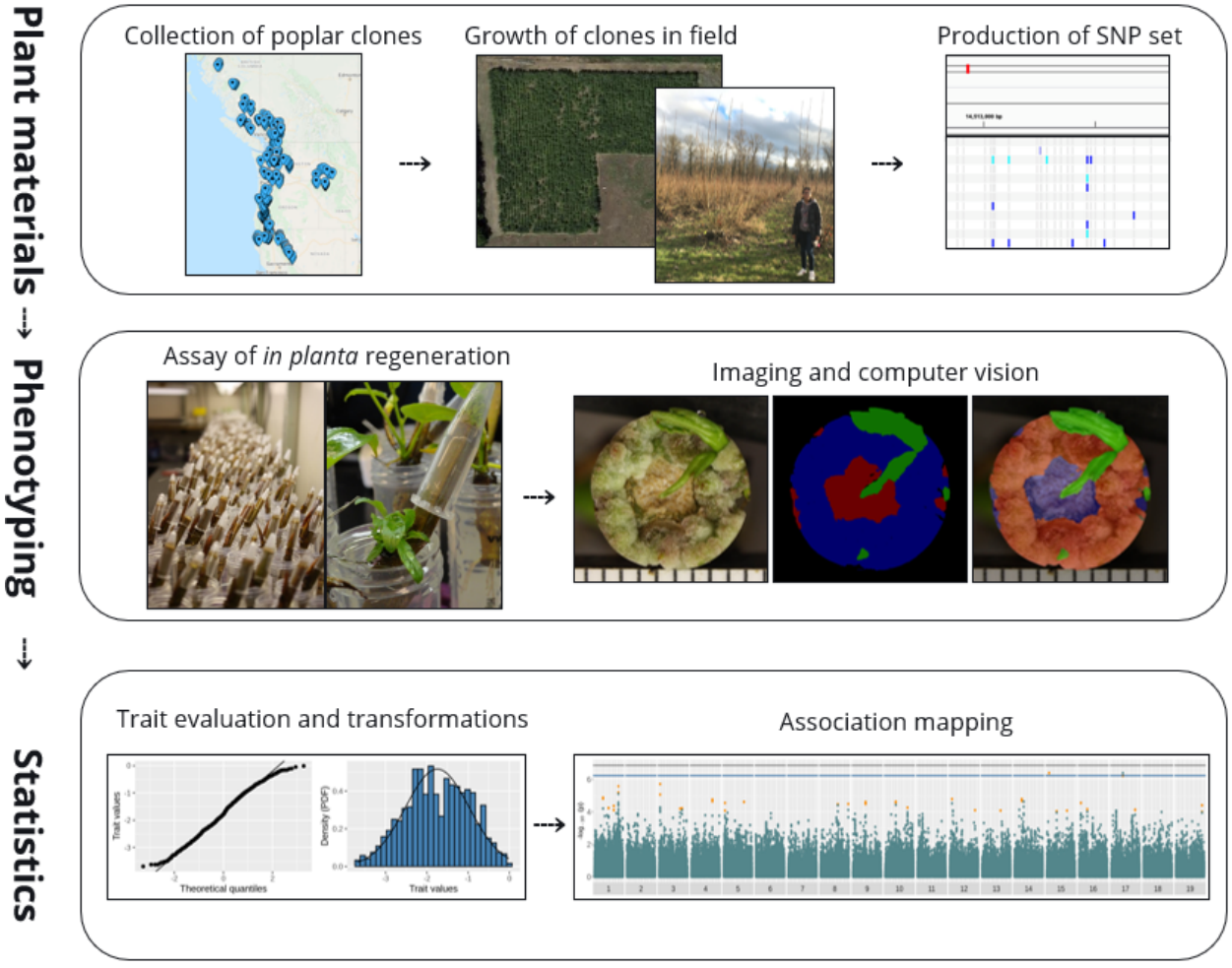

Fig. 1. Overview of experimental workflow. Plots in the "statistics" panel are shown for

transformation of the trait of Callus Area at week two and analysis of this trait via MTMC-SKAT.

Plant materials

We utilized an expanded version of the previously reported re-sequenced *P. trichocarpa* GWAS population (Tuskan *et al.*, 2018; Bdeir *et al.*, 2019; Weighill *et al.*, 2019; Chhetri *et al.*, 2020; Chen *et al.*, 2021). The population was expanded to include an additional 441 genotypes, particularly from Northern California, Oregon and Idaho, filling a geographical gap that existed in the previous population (Fig. 2). While this clone bank is kept at multiple locations, phenotyping in this study only made use of the replicate in Corvallis, OR, featuring a total of 1,307 clones in the population (out of 1,323) and for 1,219 of which regeneration phenotyping was performed. Clones were grown at two field locations in Corvallis, OR: one location planted in 2009 featuring the original GWAS population, and another planted in 2015 featuring the newly added clones. Dormant cuttings were taken in the winter of 2018, 2019, and 2020, frozen, and then rooted up to one year later. Plants were then regularly pruned and fertilized to ensure there were healthy green leaves suitable for sterilization and introduction into tissue culture. A second greenhouse population was established and allowed to go dormant in winter; plants from this source were occasionally used to replace plants in the main greenhouse population that provided plant materials throughout the year.

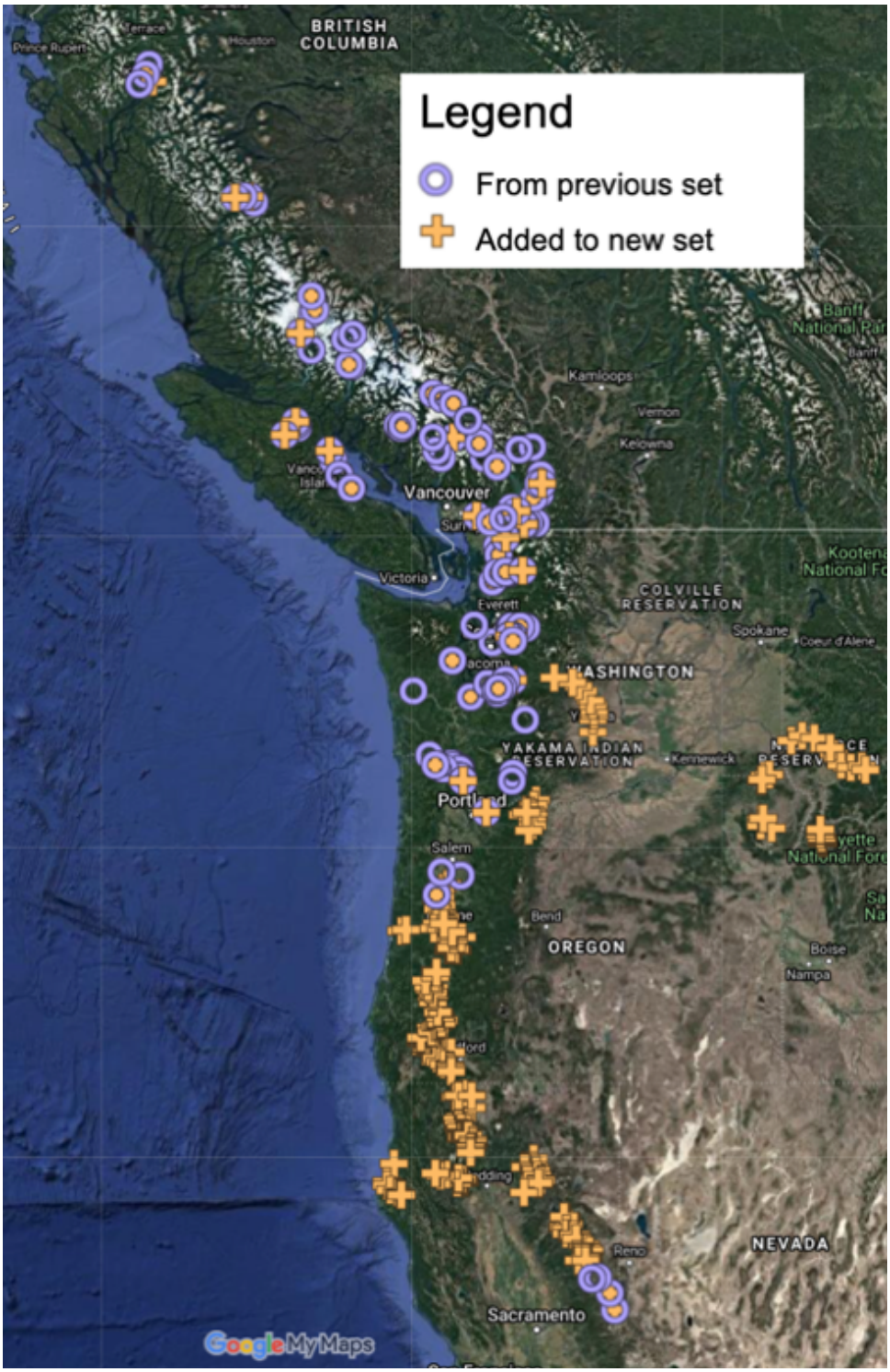

Figure 2. Origins of *P. trichocarpa* clones used to generate SNP set. A total of 1,323 wild clones were collected over a geographical range across the pacific northwest region of the USA and the southwest of Canada. Tree location is shown for 1,301 genotypes for which precise location data is available.

Sequencing and SNP set preparation

We analyzed the distribution of SNPs after resequencing of 406 additional genotypes by the DOE Joint Genome Institute. SNP calling was done at Oak Ridge National Laboratory (Yates *et al.*, 2021). There was a total of 40.4M SNPs prior to filtering for minor allele frequency (MAF) and additional quality criteria. The density and consistency of SNP data on each chromosome were assessed using the R package CMplot (Fig. S1) and by producing histograms of gap sizes for each chromosome.

Assay of regeneration

Frozen stem cuttings were incubated at 4° C for 2-4 weeks, then placed in 50mL falcon tubes with water for five weeks. Based on preliminary experiments (data not shown), we found that treatment of the cut top of each stem with 10µL of 0.5mg/mL thidiazuron (TDZ) in water improved callus regeneration considerably (37% of genotypes produced shoots, compared to 24% without TDZ). After application of TDZ to a given stem tip, a 1.5mL microcentrifuge tube was inverted over the stem tip to prevent desiccation during regeneration (as shown in Fig. 1). On a weekly basis beginning the second week, stem tips were imaged from overhead using a Canon Rebel XSi DSLR camera attached to a rack mount.

Due to practical limitations on the numbers of clones that could be assayed for regeneration simultaneously, subsets of the study genotypes (termed “phases”) were assayed at one time, with no more than 400 cuttings per phase. Images were taken on a weekly basis from the second week through the fifth week, with the exceptions of weeks four and five in the first phase and week four in the third phase. There were two replicate plants measured for all but the first three phases, where only a single replicate was used.

CV pipeline

To perform annotation of images for CV, 249 images were randomly sampled from the first seven phases and manually annotated using the Intelligent Deep Annotator for Segmentation (IDEAS) graphical user interface (Yuan *et al.*, 2022). As described in our prior work, these samples were used to train a convolutional neural network (PSPNet) to segment images of regenerating stem tips with each pixel labeled as one of four classes: callus, shoot, unregenerated

stem and background. At each timepoint, two traits were computed: the proportion of total plant area which consists of callus (henceforth, "callus area"), and of shoot ("shoot area").

Data preparation

For replicated samples, the mean value of each trait across the two replicates was computed and used in downstream analysis. For genotypes lacking replication, the single unreplicated trait value was used.

Additional traits were computed by performing principal component analysis (PCA) using `stats::princomp` in R over three groups of traits: 1) callus area traits at all timepoints; 2) shoot area traits at all timepoints; and 3) both callus area and shoot area at all timepoints. Genotypes missing data for a given trait at any timepoint were excluded from a given PCA. Scree plots were evaluated to estimate the number of PCs representing significant proportions of trait variation.

The normality of traits was assessed using Q-Q plots, histograms, Shapiro-Wilks tests and Pearson correlation coefficients computed against theoretical normal distributions with the same mean and standard deviation as the given trait. To avoid severe violations of normality that may lead to inflated error rates, all traits were transformed prior to statistical analysis. The most basic transformation applied was a removal of zero values followed by Box-Cox transformation. For certain PC traits, a spike was observed at particular values, which corresponded to genotypes with zero values for all traits used in the given PCA; these genotypes were consequently removed. In cases where we determined that thresholding or extreme outlier removal was necessary, these treatments were performed prior to Box-Cox. In addition, as an alternative to Box-Cox transformations, rank-based inverse normal (RB-INV) transformations were performed for difficult distributions (Fig. S2, Table S1-2).

Association mapping

Because of the distinct assumptions and data types for which various GWAS methods are suited, we employed an analysis pipeline that made use of four GWAS methods. First, Genome-wide Efficient Mixed Model Association (GEMMA) (Zhou & Stephens, 2012) was used to perform single-marker tests with continuous traits (following transformations toward normality) using a kinship matrix generated from genome-wide SNPs as a covariate to adjust for population

structure. Prior to GEMMA, SNPs were filtered based on minor allele frequency (MAF) > 0.05 and a missing rate of given SNPs across genotypes > 0.10 using PLINK, resulting in ~13.2 million SNPs. GEMMA was used to compute Wald *p*-values for SNP effects, using the `-lmm 1` option. To speed computation, GEMMA was parallelized using the GNU Parallel (Tange, 2020) framework to simultaneously run each given trait on a CPU core. In addition to performing association mapping, GEMMA was used to provide an estimate of narrow-sense SNP heritability (h^2_{SNP}) for each trait. Downstream GWAS and gene candidate evaluation was performed for traits with estimated h^2_{SNP} above 0.10.

Second, the Generalized Mixed Model Association Test (GMMAT) (Chen *et al.*, 2016) was used for single-marker tests with the same kinship covariate; however, rather than using continuous trait variables, GMMAT applies logistic regression and works with binarized traits. Due to the computational expense of computing Wald *p*-values via logistic regression, we first performed the GMMAT variance component score test (`glmm.score`) for a genome-wide screen and then extracted a subset of 100 or 1,000 SNPs with the lowest score test *p*-values from each run and computed Wald *p*-values for these (using `glmm.wald`). This GMMAT workflow was performed with two SNP subsets prepared by PLINK: one had a missing rate threshold of 0.10 and an MAF threshold of 0.05 (7.7 million SNPs), and the second had the same missing rate threshold but an MAF threshold of 0.01 (13.2 million SNPs).

Third, we applied Fixed and Random Model Circulating Probability Unification (FarmCPU) (Liu *et al.*, 2016), which provided single-marker tests for continuous, transformed traits similarly to GEMMA, but with an adjusted kinship covariate for improved statistical power. The package FarmCPUpp (Kusmec & Schnable, 2018) was used, together with an R function to apply resampling for optimization of significance threshold (`p.threshold`) for inclusion of SNPs in the kinship matrix calculation. To avoid singular or near-singular matrix errors that can result when multiple SNPs passing this threshold are in strong LD, we performed this workflow using a SNP set that was filtered by PLINK on the basis of LD (using parameters `--indep-pairwise 100kb 10 0.7`) after filtering by MAF (0.05) and missing rate (0.10), resulting in ~2.3M SNPs.

Finally, for multiple-marker tests we applied the SNP-set (sequence) Kernel Association Test (SKAT) (Ionita-Laza *et al.*, 2013) with untransformed traits. SKAT was performed on overlapping 3kb windows staggered by 1kb, using a set of 34.0 M SNPs filtered for a missing

rate of 15%. The R extension Multi-Threaded Monte Carlo SKAT (MTMCSKAT) was used to run SKAT on a high-performance cluster, COMET (made available through NSF XSEDE (Towns *et al.*, 2014). We calculated empirical *p*-values for top associations to avoid Type I and Type II error resulting from the non-normal distributions of untransformed traits. Two means of controlling for population structure were tested and compared with this workflow. We compared a "P" model in which structure is represented by principal components derived from SNPs (computed with PLINK) to a "Q" model in which structure is alternatively represented by subpopulation estimates produced by fastSTRUCTURE (Raj *et al.*, 2014).

To produce PCs for the P model, we employed a filtered set of ~10.3M SNPs with MAF > 0.05 and consulted scree plots and used K-means clustering to inform about the number of PCs appropriate for representing population structure; as a result we used 6 PCs for the P model.

To produce a Q matrix for use with SKAT Q models, we used fastSTRUCTURE using a subset of ~72k SNPs filtered based on LD, MAF, and missing rate using PLINK with parameters ` --indep-pairwise 100kb 10 0.05 --maf 0.05 --geno 0.1`. Ten replicates were performed with fastSTRUCTURE for each possible number of subpopulations (K) ranging from 3 to 12. To understand subpopulations in an evolutionary context, we used SNPhylo (Lee *et al.*, 2014) to produce a dendrogram from our SNP data. SNPhylo was run with a subset of ~129k SNPs prepared by PLINK with parameters ` --indep-pairwise 10kb 10 0.05 --maf 0.05 --geno 0.1`.

Geographical locations (longitude and latitude) were recorded for 1,301 of 1,323 genotypes in the SNP set and plotted against traits, SNP-derived PCs (for SKAT "P" model), primary subpopulation information (for SKAT "Q" model), and dendrogram information (from SNPhylo) using the `phylo.to.map` function in Phytools (R) and Google Maps "My Maps". Phytools was also used to cross-reference dendrograms with traits, SNP-derived PCs and primary subpopulation information (using function `phylo.heatmap`) (Revell, 2012).

To inform about the appropriate window size for SKAT, as well as to inform about the likelihood of genes proximal to associated SNPs or SNP windows being directly involved in affecting traits (vs. being associated as a result of genetic linkage), we evaluated LD decay. To facilitate efficient computation of LD decay, a reduced SNP set (~78k SNPs) was prepared by PLINK with parameters `--maf 0.05 –geno 0.1 –thin 0.01`. Further reduced SNP files were prepared with PLINK to only include genotypes in the "Oregon" and "California" subpopulations (named based on general location of most genotypes in each). PLINK was further

used to compute pairwise LD between all SNPs on each given chromosome with each SNP set. Using R, the average LD for each possible distance (e.g. 1bp, 2bp, 3bp… up to 50kb) was computed and plotted for the whole population as well as each of the two selected subpopulations.

Statistically significant associations from the various pipelines were first determined by computing FDR ($\alpha = 0.10$) and Bonferroni thresholds ($\alpha = 0.05$). The Bonferroni thresholds were computed given the number of tests equal to the number of SNPs (for single-marker tests GEMMA, GMMAT and FarmCPU) and the number of SNP windows (in the case of SKAT). We then extracted lists of SNPs with *p*-values below these thresholds for interrogation.

We then evaluated the extent to which multiple SNPs supported the association of a nearby gene, whether individual SNPs met the FDR or Bonferroni statistical thresholds or not. We implemented the augmented rank truncation (ART) method (Vsevolozhskaya *et al.* 2019) to scan Wald *p*-values from GEMMA and GMMAT and identify cases where a SNP produces a *p*-value below $1*10^{-5}$ and is within 500bp of at least 5 additional SNPs with *p*-values below $1*10^{-4}$ when considering the upper half of top-ranking SNPs. For each of these windows, a combined *p*-value was computed for the extracted SNPs. A Bonferroni threshold for ART *p*-values was computed ($\alpha = 0.05$) from the approximate number of independent tests (contiguous assembled genome size / ART window size). The Bonferroni threshold of $\sim 1.27*10^{-7}$ was computed using the number of independent tests of ($\sim 3.94*10^{5}$ 1kb windows spanning the ~394 Mb of contiguous assembled chromosomes) and is notably less conservative than the Bonferroni threshold used for raw *p*-values from GEMMA/GMMAT (henceforth, "conservative Bonferroni"), as computed from the total number of tests (as low as $\sim 3.79*10^{-9}$, given up to 13.2 million SNPs. However, it is well known that Bonferroni thresholds for individual SNPs erroneously consider each SNP as an independent test, though very large numbers of SNPs are of course in LD.

To determine on a high-throughput scale which genes are likely to be responsible for statistically significant quantitative trait loci (QTLs; either SNPs or SNP windows), we used R scripts to reference genome and genome annotation data available through Phytozome (phytozome.doe.gov) (Tuskan *et al.*, 2006). In this workflow, the position of loci were evaluated for candidate genes only when these loci represent the "peak" of a signal, determined by checking for any other loci within 3kb with a more significant *p*-value. The candidate gene

responsible for the significance of a given locus was assumed by the workflow to be the gene that encompasses or is closest to the locus, where one exists within 5kb. The R package InterMineR (Kyritsis *et al.*, 2019) was used to collect Phytozome data on gene function, Arabidopsis homologs, and gene ontology terms and organized these by locus. The GreeNC database was used to identify possible noncoding regulatory RNAs among gene candidates (Di Marsico *et al.*, 2022). For the top gene candidates, particularly those passing the conservative Bonferroni or FDR ($\alpha = 0.10$) thresholds, or those passing the less conservative Bonferroni thresholds used for ART and among the five most-significant GEMMA-ART or GMMAT-ART associations for a given trait, Integrative Genomics Viewer (IGV) (Robinson *et al.*, 2011; Thorvaldsdóttir *et al.*, 2013) was used to manually investigate gene position relative to significant SNPs, including consideration of other nearby genes, distance to the putative transcription start site, and direction of transcription.

Evaluation of possible adaptive role of regeneration traits

Following the identification of subpopulation structure when fastSTRUCTURE was used to produce covariates for the SKAT "Q" model, we aimed to further investigate the relationships between traits, geography and theoretical ancestral subpopulations to gain insights into the possible adaptive evolution of these regeneration traits. To this end, we used `lm` (R) to construct linear models regressing each trait over latitude and the Q matrix featuring estimates of each theoretical ancestral subpopulation's contribution to each individual's genome (from fastSTRUCTURE). We then visualized relationships, latitude and subpopulation using `ggplot2` (R).

Results

SNP set for *P. trichocarpa* provides comprehensive view of natural variation

The SNP set produced for this population displays polymorphism across all regions of contiguous chromosomes represented in the reference genome (Fig. S1). Poplar clones collected for the GWAS clone bank represent a wide range of geographic diversity, nearly spanning the natural range of *P. trichocarpa* across British Columbia and the Pacific Northwest of the United States, including Idaho and northern California (Fig. 2). There is clearly very strong natural

intrachromosomal recombination, as LD decay occurs rapidly reaching $R^2 = 0.2$ within 2kb whether computed for the whole population or either of two prominent subpopulations (Fig. 3).

We attempted to gain insights into the possible role of adaptive evolution in regeneration traits via relationships between the traits, latitude and theoretical ancestral subpopulation (Methods). At $\alpha = 0.005$, there appears to be a significant effect of latitude of clone origin on the trait of callus area at week four, while controlling for subpopulation. Several other relationships are significant at 0.05, between various callus traits and latitude and/or subpopulation (Table S3). Visualization of the relationships between traits and latitude along with regression trendlines showed a positive relationship between many regeneration traits and increasing latitude, but the significance of these trends was lost when considering clones of each given primary subpopulation independently (Fig. S3). Considering the lack of independence between variables of theoretical ancestral subpopulation and latitude, we advise caution in overinterpreting these results as evidence of an adaptive role of regeneration, but also note several significant or borderline-significant trends indicating such a role may exist.

Relationships between evolutionary clades, geography, and population structure suggest that that *P. trichocarpa*, despite its dioecy and long-distance gene flow, exists with a number of subpopulations that are statistically distinct albeit highly admixed. A total of 110 fastSTRUCTURE runs were performed, including 10 replicates for each value of K (subpopulation number) ranging from 2-13. The log marginal likelihood appears to be maximized with K equal to 6 or 7 (Fig. S4). For each individual in the population, the most closely related subpopulation was extracted and considered the primary subpopulation. Geographic and evolutionary patterns were revealed by cross-referencing of a dendrogram (SNPhylo) with primary subpopulation and geographic location. These plots were evaluated with primary subpopulations from fastSTRUCTURE models both with K=6 and K=7 (Fig. S5); the K=7 model showed the strongest alignment between phylogeny and geography. Approximately from Seattle northward, individuals display a heavy degree of admixture and fail to cluster into clear subpopulations. Otherwise, the existence of several subpopulations is supported by agreement between phylogenetic clades, geographic location, and primary subpopulation label from fastSTRUCTURE. These include distinct subpopulations in the western region of Idaho and nearby eastern Oregon and Washington (and extending all the way to the eastern Washington Cascades near Yakima), the Willamette Valley of central western Oregon and

nearby Western Washington, southwest Oregon and nearby northern California, northwestern Washington extending into southwestern Canada, and central western to northwestern Canada (Fig. 2).

We further attempted to summarize population structure by performing PCA over SNP data using PLINK. Similar to fastSTRUCTURE subpopulation estimates, PCs explaining a substantial portion of variance show clear relationships with geography and most of the same phylogenetic clades (Fig. S6-7). The use of 6 PCs to represent population structure in SKAT models, as discussed below, was supported by the scree plot (Fig. S8) and the relatively minor contributions of subsequent PCs to k-means clusters computed from PCs (Fig. S9).

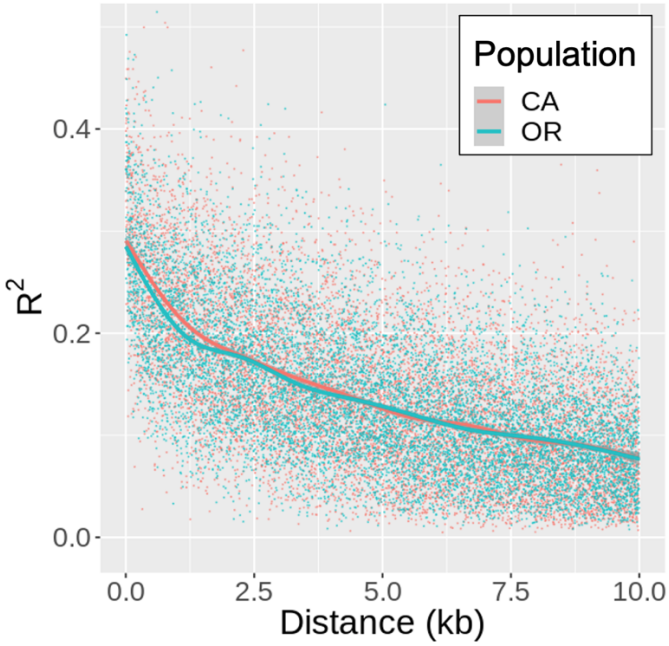

Figure 3. Linkage disequilibrium decay curves for Oregon ("OR") subpopulation and California ("CA") subpopulation. Primary subpopulations were determined using fastSTRUCTURE, with a K = 7 model (Methods).

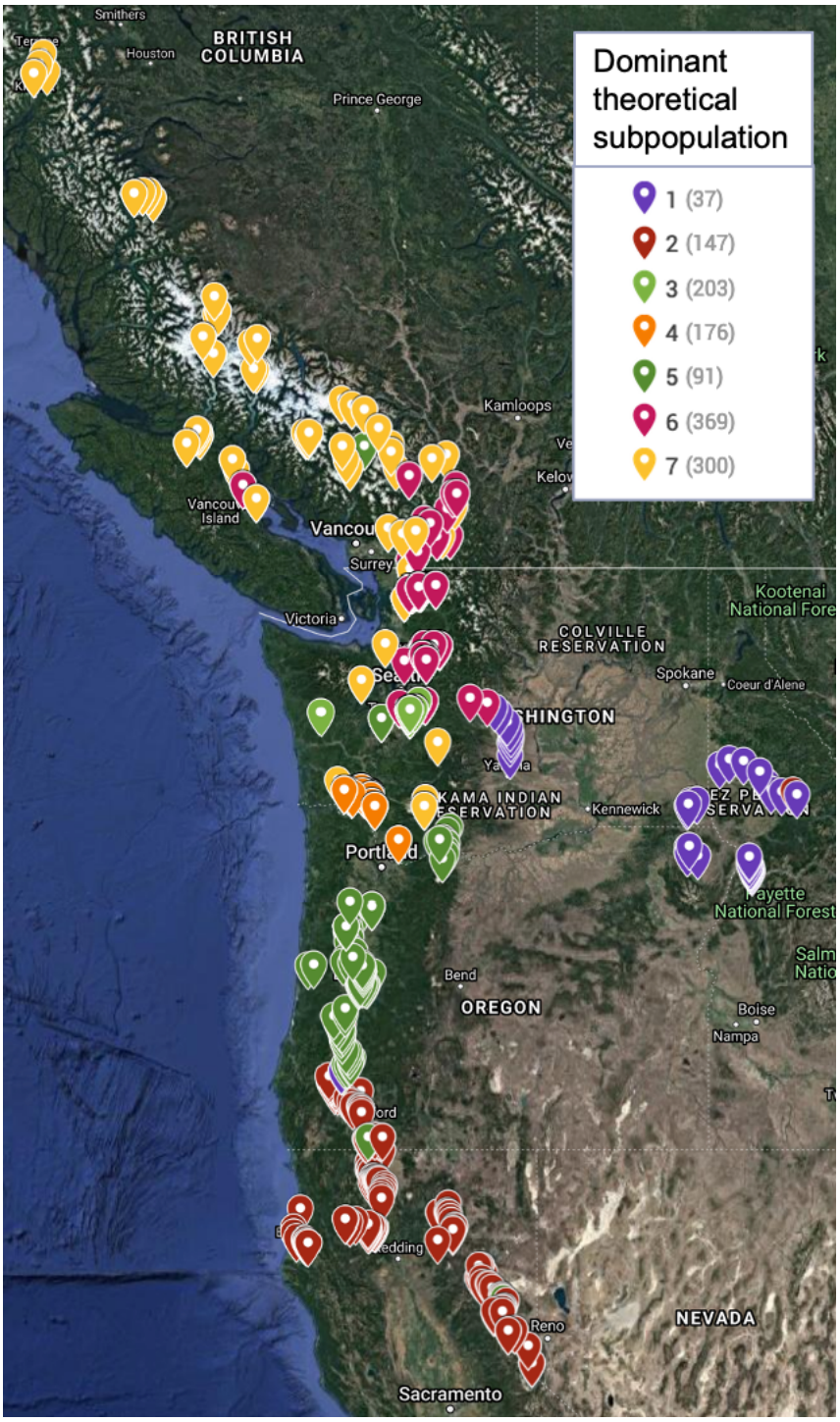

Figure 4. Information on theoretical ancestral subpopulations (fastSTRUCTURE, with K=7 model), cross-referenced with geographical locations of clones. Data is shown for the 1,301 clones for which location data is available (out of 1,323). Points are labeled by the theoretical subpopulation accounting for the largest portion of ancestry for each clone. This plot was produced with Google Maps MyMaps.

314 Trait transformations

315 Prior to transformations, most traits displayed marked non-normal characteristics as
316 indicated by Q-Q plots, histograms, Shapiro-wilk tests, and Pearson correlation coefficients of
317 each distribution with a normal distribution featuring the same mean and standard deviation. In
318 most cases the improvement in normality after transformation was marked (Table S1; Fig. S2;
319 data not shown). Non-normal characteristics were reduced substantially in most cases by
320 excluding genotypes with zero values and applying a Box-Cox transformation (e.g., Fig. S2). For
321 the traits of callus or shoot area at each timepoint, based on visual inspection and consult with a
322 statistical consultant, the improvement in metrics of normality was deemed adequate for linear
323 models. All PCA-derived traits necessitated additional treatments to avoid severe violations of
324 the normality assumption of linear models, including removal of outliers and in some cases
325 removal of values below an elbow in the frequency distribution (estimated as the position where
326 the second derivative of the probability frequency distribution is maximum) (Table S2).
327
328 Principal components as proxies for complex patterns of regeneration
329 Scree plots and heat maps of loadings revealed common trends in regeneration across
330 timepoints and regenerating tissue types (callus and shoot). These results were obtained for three
331 different PCA analyses: first, for both callus and shoot area at all timepoints (Fig. 5), and then
332 with callus and shoot data analyzed independently over all timepoints (Fig. S10). In all three
333 cases, the PC explaining the most variation (PC1) represented a tendency of the tissue(s)
334 included in PCA to regenerate well across all timepoints. Latter PCs provided proxies for more
335 complex patterns of regeneration. PC2 from the PCA over callus traits appears to represent high
336 levels of callus regeneration at early, but not later timepoints. PC2 from the PCA over all callus
337 and shoot traits appears to represent a tendency for callus to regenerate robustly, but to fail to
338 develop into shoots. Subsequent PCs, for each batch of traits, represented a relatively small
339 proportion of variance explained and were thus not analyzed for gene candidates.
340

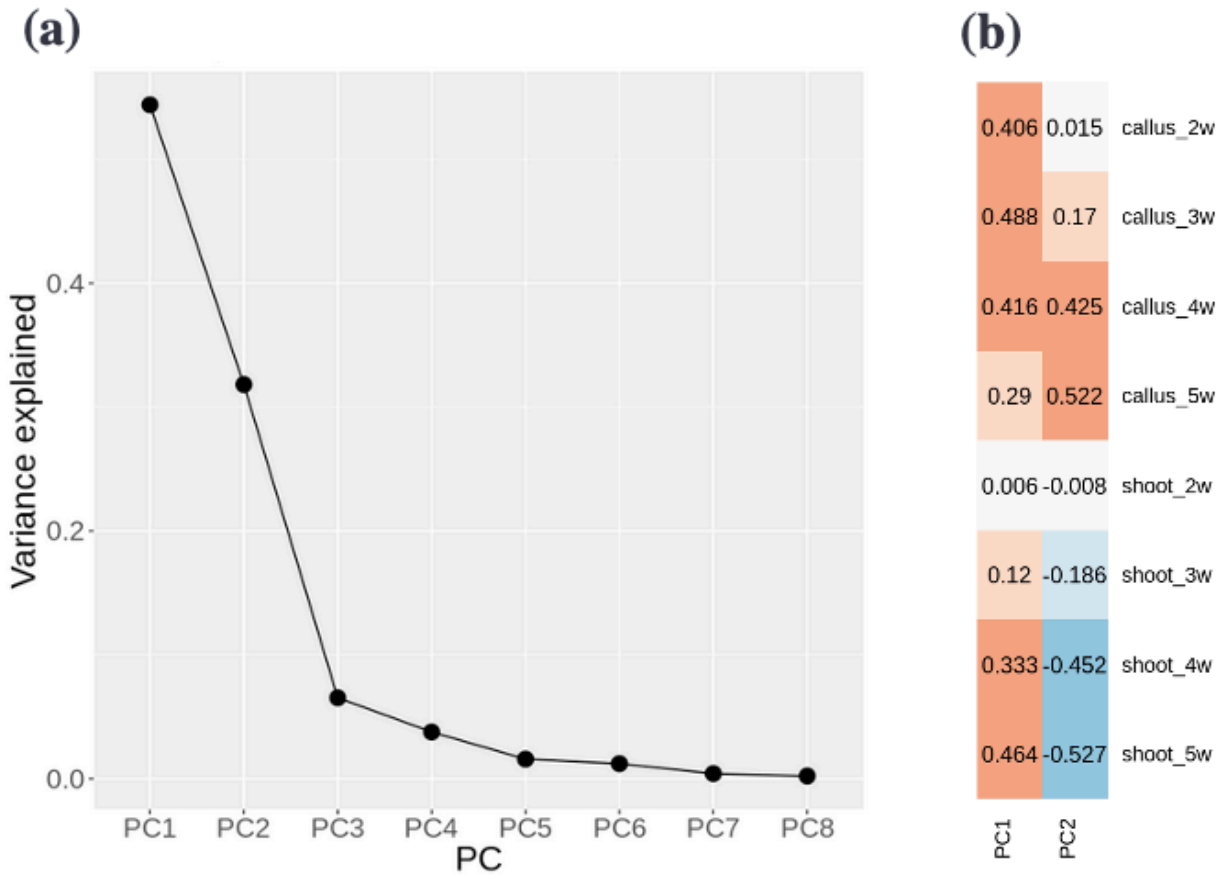

Figure 5. Results from PCA over all callus and shoot traits. A) Scree plot; B) Heat map of loadings from PCA.

341 Genes implicated by significant quantitative trait loci (QTLs)

342 We interrogated traits with h^2_{SNP} above 0.10 for candidate genes (Table S4). Across all
343 four GWAS models applied (GEMMA, GMMAT, FarmCPU and SKAT), we report a total of 8
344 unique QTL peaks with *p*-values passing the Bonferroni significance threshold, as well as 46
345 passing the FDR (alpha = 0.10) threshold. All Bonferroni-significant associations are inside or
346 within 5kb of a gene found in the genome annotation, as well as 34 associations (73.91%)
347 meeting the latter threshold (Fig. 6-8, Table S5-6). We found 139 unique QTL peaks from
348 applying our implementation of ART to GEMMA results (Table S7), as well as 48 from applying
349 ART to GMMAT results (Table S8).

350 We compared results from complementary SKAT models with population structure
351 represented either by the fastSTRUCTURE Q matrix with 7 subpopulations ("Q model") or by
352 the first 6 PCs ("P model") for a subset of four traits (callus area at wk. 4 and wk. 5; shoot area at

353 wk. 4 and wk. 5). These models displayed a remarkable level of agreement, especially for *p*-
354 values that met thresholds of significance and were thus selected for validation by computing
355 empirical *p*-values with MTMCSKAT (Fig. S11). Several of the most promising candidates,
356 based on the biology of their homologs in Arabidopsis, are shown in Table 1.
357

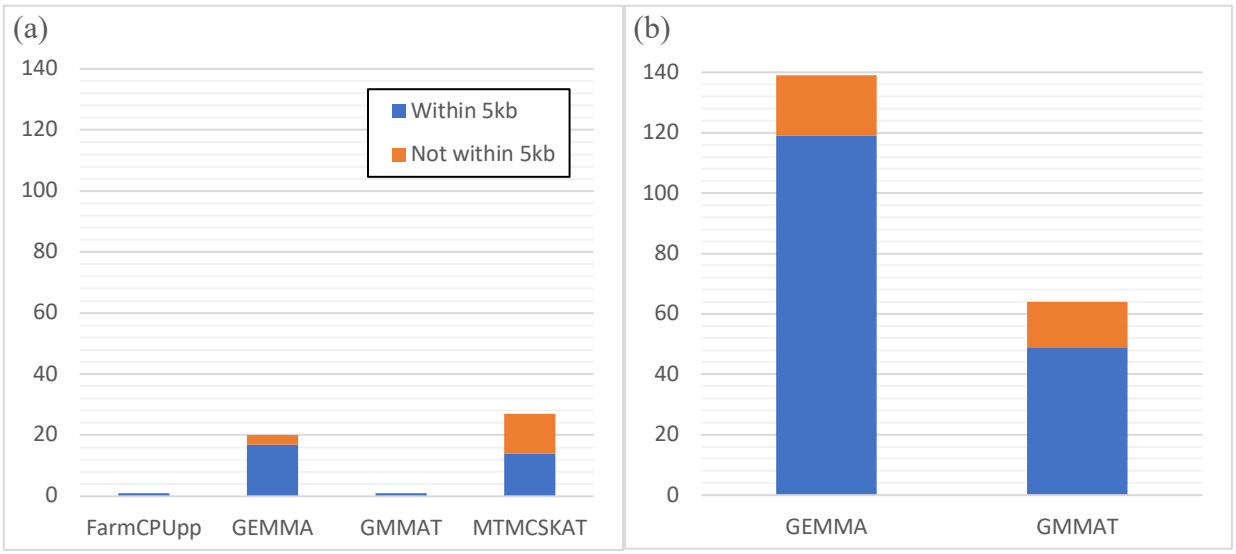

Figure 6. Barplots summarizing the numbers of associations from each GWAS method, with two types of significance thresholds, as well as within a 5kb distance threshold of the nearest gene. QTL peaks were taken as the point with the lowest *p*-value at any given peak, where multiple points within the same peak may otherwise pass a given significance threshold. A) QTL peaks passing the Benjamini-Hochberg threshold (FDR; $\alpha = 0.10$); B) QTL peaks passing ART-Bonferroni threshold ($\alpha = 0.05$, N of # 1kb windows in genome).

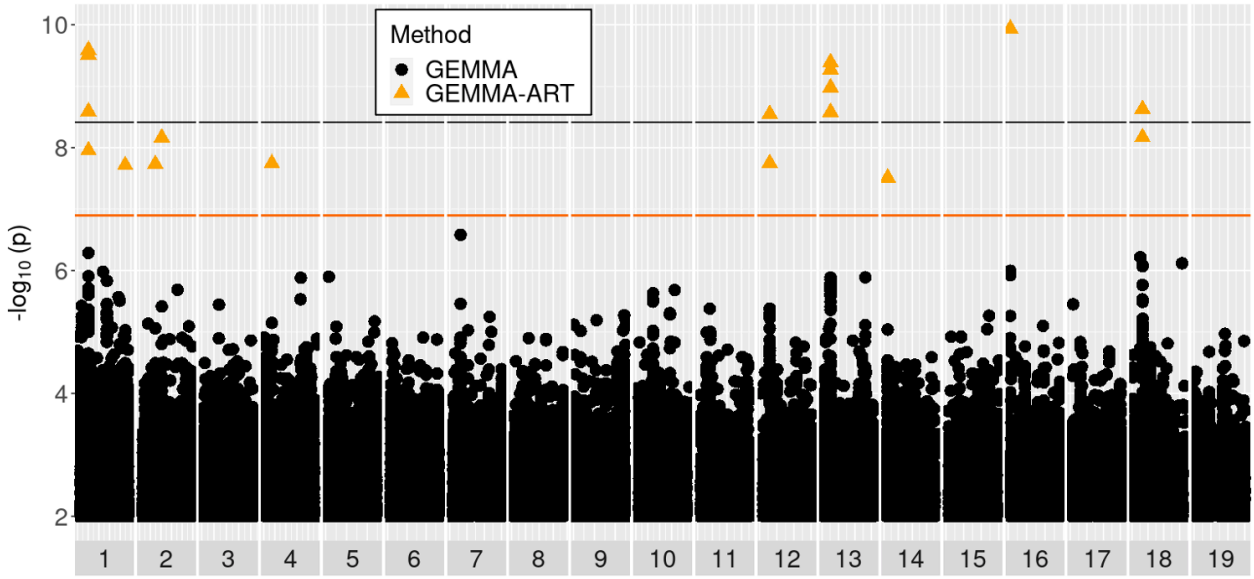

Figure 7. Manhattan plot for GEMMA results for the trait of callus area at week four: Black and orange lines show Bonferroni significance thresholds for GEMMA results with independent SNPs, and for ART applied to GEMMA over 1kb windows of SNPs, respectively. Black circles represent tests of individual SNPs by GEMMA, while orange triangles represent 1kb windows tested by ART applied to GEMMA results.

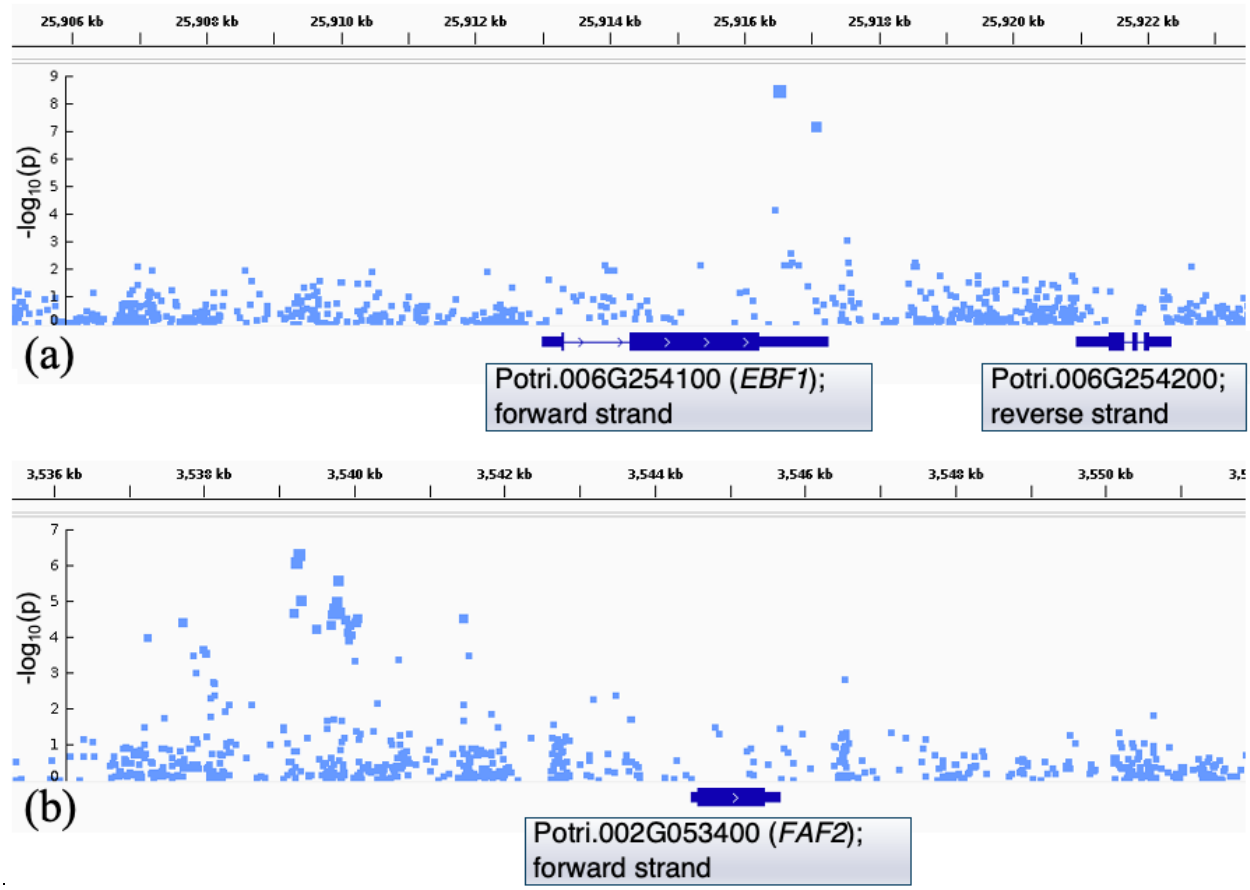

Figure 8. Plots produced by Integrated Genome Viewer (IGV) show zoomed-in portions of Manhattan plots aligned to the genome annotation for P. trichocarpa (v3.1). Introns, untranslated regions and exons are respectively visualized with increasing thickness of bars. Labels in gray boxes were manually added to show gene IDs and the strand on which genes are found. A) Results on chromosome 6 for GEMMA of Box-Cox transformed trait Shoot PC2; B) Results on chromosome 2 for GEMMA of Box-Cox transformed trait Shoot PC1, showing an association found significant via ART. Examples of plots for additional loci can be found in Supp. Materials 1.

Gene candidates							*Arabidopsis homologs*			
Threshold	**Trait**	**Method**	**Transf.**	**Dist. (bp)**	**QTL Pos.**	**Accession ID**	**Accession ID**	**Description**	**Score**	**Similarity**
Bonf.	Callus 2w	GMMAT	Binarized trait	3	5’	Potri. 006G276200	AT 3G12660	FASCICLIN-like arabinogalactan protein 14 precursor (FLA14)	156	60.40%
FDR (α=0.1)	Shoot PC1	MTMC-SKAT	Untransformed	0	Exonic	Potri. 019G101900	AT 4G28250	expansin B3	409	86.90%
Bonf.	Shoot PC2	GEMMA	Outliers removed, Box-Cox	0	Exonic	Potri. 006G254100	AT 2G25490	EIN3-BINDING F BOX PROTEIN 1 (EBF1)	812	80.70%
FDR (α=0.1)	Shoot PC1	MTMC-SKAT	Untransformed	0	Intragenic, non-exonic	Potri. 003G194600	AT 3G12250	TGACG MOTIF-BINDING FACTOR 6 (TGA6)	520	91%
FDR (α=0.1)	Shoot PC1	MTMC-SKAT	Untransformed	4877	3’	Potri. 015G041800	AT 3G18165	modifier of snc1,4 (MOS4)	363	84.70%
FDR (α=0.1)	Callus 3w	FarmCPU	Box-Cox	2839	5’	Potri. 001G177801	AT 1G80410	EMBRYO DEFECTIVE 2753 (EMB2753)	246	94.40%
FDR (α=0.1)	Shoot PC1	MTMC-SKAT	Untransformed	0	Exonic	Potri. 002G070600	AT 1G21326	VQ motif-containing protein 3 (VQ3)	105	65%
FDR (α=0.1)	Shoot PC2	GEMMA	Outliers removed, thresholding, Box-Cox	0	Exonic	Potri. 002G173300	AT 2G46560	transducin family protein / WD-40 repeat family protein	2357	66.80%
Bonf.	Callus 5w	GEMMA	Box-Cox	3947	3’	Potri. 018G049600	AT 5G35550	TRANSPARENT TESTA 2 (TT2)	196	69.90%
FDR (α=0.1)	Shoot PC1	MTMC-SKAT	Untransformed	0	Exonic	Potri. 004G155400	AT 1G75250	RADIALIS-LIKE	129	82.40%

								SANT/MYB 3 (RSM3)		
FDR (α=0.1)	Callus 5w	GEMMA	Box-Cox	1192	5'	Potri. 010G105600	AT 4G16110	ARABIDOPSIS RESPONSE REGULATOR 2 (ARR2)	382	75.70%
FDR (α=0.1)	Shoot PC2	GEMMA	Outliers removed, thresholding, Box-Cox	137	5'	Potri. 011G031100	AT 1G11530	C-TERMINAL CYSTEINE RESIDUE IS CHANGED TO A SERINE 1; thioredoxin	96	69.30%
ART-Bonf.	Callus, Shoot PC1	GEMMA	Box-Cox	5222	5'	Potri. 002G053400	AT 1G03170	FANTASTIC FOUR 2 (FAF2)	103	54.60%
ART-Bonf.	Callus, Shoot PC1	GEMMA	RB-INV	5222	5'	Potri. 002G053400	AT 1G03170	FANTASTIC FOUR 2 (FAF2)	103	54.60%
ART-Bonf.	Callus 4w	GEMMA	Box-Cox	5749	5'	Potri. 012G032900	AT 4G27950	CYTOKININ RESPONSE FACTOR 4 (CRF4)	227	62.60%
ART-Bonf.	Shoot PC1	GEMMA	RB-INV	2346	5'	Potri. 012G070400	AT 3G52960	PEROXIREDOXIN-II-E (PRXIIE)	105	78.90%

Table 1. Fifteen gene candidates with Arabidopsis homologs that have putative roles in biological processes related to *in vitro* regeneration. Relevant literature is discussed for each of these candidates (Discussion). Distance (Dist.) of QTLs from the transcription start site is shown for intergenic associations. Score and similarity percentage is shown for Smith-Waterman alignment of poplar gene candidates with Arabidopsis homologs. Remaining gene candidates are summarized in Table S6-8.

Discussion

Distinct subpopulations correlate with phylogeography

The existence of distinct geographical subpopulations of *P. trichocarpa* is supported by cross-referencing of results from population structure analysis (fastSTRUCTURE), phylogenetics (SNPhylo), and geographical information for genotypes. These distinct subpopulations appear clearly in the southern portion of the population, whereas the northern portion displays a remarkable degree of admixture with mixed origins across the southern subpopulations. We speculate that, following the establishment of distinct southern subpopulations during the Last Glacial Period (Armstrong *et al.*, 1965), the recession of glaciers allowed for these subpopulations to spread to the northern region—where there has not yet been sufficient time or subdivision for distinctive populations to form. In contrast, the disjunct nature of many of the southern population groups is likely to have provided historical opportunities for differentiation. While previous work using approximately 12 isozyme loci did not reveal distinct subpopulations of *P. trichocarpa* over a more narrow, but similar geographical range (Weber & Stettler, 1981), our work demonstrates the much-increased power of genome-scale SNP data—where millions of loci are surveyed—to detect subpopulations.

<u>High-throughput phenomics support scale and precision of GWAS</u>

The high-throughput phenomics workflow used for this work was described, in part, by Yuan et al. (2022). The IDEAS graphical interface for image annotation enabled the production of a large set of training examples (249 images in total) with pixelwise labels for callus, shoot and unregenerated tissues. This training set enabled a deep segmentation model that was used to automatically segment the 4,647 remaining images. Although generation of the training samples was time-consuming, performing manual segmentation for all images would have been time-prohibitive, and summarizing traits with an ordinal scale instead of pixelwise statistics would have risked the introduction of subjective biases and violation of linear model assumptions—while sacrificing much precision and detail. This system or others that are functionally comparable (Russell *et al.*, 2008; Dutta & Zisserman, 2019) can be made more accessible and practical with innovations to reduce the number of clicks needed for image annotation by further semi-automation of annotation. Overall accuracy in segmentation of the "validation" set of images was 79.21% as measured by Intersection over Union (IoU), while relatively homogenous stem tissues had IoU of 88.14%, and highly heterogenous callus tissues had 67.40%. Advances

in the architectures of deep segmentation neural networks can contribute to improved accuracy in segmenting complex and heterogenous tissues of interest to biologists.

Complementary GWAS approaches provide variety of insights

Transformation of traits to approximate normality is commonly employed for biological data during GWAS to avoid linear models' assumption of normality of residuals. In our study, because traits were computed as the proportion of plant tissue labeled by CV as callus or shoot, and many genotypes failed to develop either tissue, the resulting distributions feature a mix of a zero and nonzero values. Among traits in our study, the proportion of genotypes with zero values ranged from 89 (for callus area at week five) to 1,106 (for shoot area at week two). To help avoid violations of the normality assumption, genotypes featuring zero values were excluded from GEMMA and FarmCPU tests for each trait, but presence/absence tests were performed using GMMAT that employed the observations of a complete absence of callus and/or shoot. GMMAT and SKAT offer two complementary approaches to avoid this assumption altogether, thus obviating the need to exclude totally recalcitrant genotypes and thus suffer reduced statistical power.

Single-SNP methods including GEMMA and GMMAT share the advantage of providing insights into the specific SNPs most likely to be causative with respect to the effect of a gene on a trait. In most cases in our results, these appear to be regulatory SNPs in promoters, suggesting that variation in gene expression, rather than sequence, is the primary cause of trait variation. However, single-SNP methods suffer from relatively low statistical power since by their nature they treat each SNP-trait relationship as an independent test and do not consider combined effects of nearby SNPs. In contrast, SKAT provides improved statistical power by allowing tests for the combined effects of adjacent SNPs grouped into SNP windows, but only provides a single *p*-value for a whole SNP window. Thus, our SKAT results do not make clear which SNPs in a given window are responsible for trait variation, and as windows often overlap coding and regulatory regions, we lack insight into whether SKAT-implicated candidates are responsible for trait differences due to variation in their regulation or protein structure. Moreover, even when a given window is entirely intergenic, we lack an ability for straightforward investigation of specific promoter motifs that may be implicated by SKAT due to the lack of single-SNP

resolution. Finally, SKAT involves the upweighting of rare SNPs and results are therefore less likely to feature top gene candidates regulated by common variation (Wu *et al.*, 2011).

We therefore sought to employ a "best of both worlds" approach to improve the statistical power of GEMMA and GMMAT by considering combined effects of adjacent common SNPs without losing clarity into the specific SNPs most likely to be causative. To this end, we employed ART as a post-hoc analysis of GEMMA and GMMAT results. As ART involves the computation of combined *p*-values over SNP windows and does not assume independence of SNPs, we obtained an increase in statistical power both via both reduced *p*-values for SNP windows compared to individual SNPs (Vsevolozhskaya *et al.*, 2019), and by the ability to use a less-stringent Bonferroni threshold due to the number of tests being equal to the number of 1kb SNP windows rather than the number of individual SNPs. Our usage of ART enabled the detection of candidate genes including FAF2, CRF4 and PRXIIE (Table 1) that otherwise would have been missed in our study. Although we are unaware of applied GWAS studies utilizing ART, our results demonstrate the potential for this method to increase effective statistical power in GWAS.

Whereas prior work describes improved statistical power of FarmCPU relative to less complex Mixed Linear Models (MLM) methods such as GEMMA (Liu *et al.*, 2016; Kaler *et al.*, 2020), we report only a single significant association from our FarmCPU tests. This is likely due to loss in statistical power resulting from LD-based pruning to avoid singular matrix errors, which can affect highly structured populations such as ours in which multiple pseudo-QTNs added to FarmCPU models match between genotypes. Nonetheless, the single gene candidate revealed by FarmCPU, *RADIALIS-LIKE SANT/MYB 3* (*RSM3*), may be among the most promising for use as a biotechnological tool to enhance regeneration (discussed below).

Candidate genes have diverse roles in signaling and development

Our results indicate that natural variation in capabilities for *in planta* regeneration in poplar is controlled by numerous genes with functionally diverse roles, including in cell wall and membrane structure, hormone signaling, anthocyanin production and reactive oxygen species (ROS) regulation. Several of the most promising gene candidates, organized by biological function of orthologs in Arabidopsis, are briefly discussed below.

453 *Regulation of cell wall adhesion*

454 Potri.006G276200 encodes a member of the FASICLIN-LIKE ARABINOGALACTAN
455 (FLA) PROTEIN family and is implicated by a QTL three bases upstream of the transcription
456 start site. We report this association from GMMAT of callus area at week two, the trait with
457 greatest trait with greatest h^2_{SNP} as estimated by GEMMA. The significance of this QTL passes
458 the most stringent multiple testing correction method applied – the Bonferroni threshold (α =
459 0.05) with each individual SNP considered an independent test. No other QTLs associated with
460 this trait meet the same threshold, nor do any other QTLs from GMMAT with any trait in our
461 study.

462 The FLA gene family (~18 genes in Arabidopsis) is differentially expressed during *in*
463 *planta* embryogenesis (Costa *et al.*, 2019), but regulation in the context of *in vitro* regeneration
464 has received little study. AtFLA1 was found to be upregulated during CIM incubation media,
465 while AtFLA2 upregulation occurred upon transfer of explants to SIM. Knockout of AtFLA1
466 was reported to confer an ability for efficient *in vitro* shoot regeneration to the otherwise
467 recalcitrant Col-0 ecotype, while contrarily leading to loss of efficient regeneration in the
468 regenerable ecotype W52. Thus, effects of differential expression, as is likely to be a
469 consequence of the polymorphism from the SNP location, may be genotype-dependent in poplar
470 as well.

471 We found an association of shoot development (week four area and PC1) with a window
472 of SNPs including a portion of the promoter and first exon of Potri.019G101900 that is related to
473 Arabidopsis EXPANSIN B3 (86.9% similarity by Smith-Waterman alignment). Expansins
474 facilitate the process of cell wall loosening by regulating pH in cell walls, with various expansins
475 expressed during different stages of development. Mutations of this gene superfamily have been
476 studied in several plant species, including Arabidopsis, tomatoes, rice, soybean, and tobacco.
477 Overexpression typically produces phenotypes of enhanced growth, such as increased size of
478 plant cells and tissues, as well as reduced fruit firmness. Knockouts, in contrast, lead to reduced
479 growth and increased firmness (Marowa *et al.*, 2016). Expansins are believed to be key
480 regulators of cell wall expansion downstream of auxin, a key hormone for control of
481 regeneration (Majda & Robert, 2018).
482

Regulators of wound-responsive hormone signaling

Potri.006G254100 encodes a putative homolog of EIN3-binding F box protein 1 (EBF1). Molecular evidence from Arabidopsis suggests that EBF1 facilitates ubiquitin-mediated degradation of ETHYLENE-INSENSITIVE 3 (EIN3) and EIN3-LIKE 1 (EIL1) and that this degradation is prevented when EIN3 and EIL1 are stabilized by ETHYLENE-INSENSITIVE 2 (EIN2) (An *et al.*, 2010). Arabidopsis knockouts of EIN2 (*ein2)* were used to supply cotyledon explant material for an *in vitro* regeneration assay, which revealed an approximate fourfold reduction in shoot regeneration in the mutants. The same assay revealed a roughly threefold increase in shoot regeneration with knockout of HOOKLESS1 (HLS1; Chatfield & Raizada, 2008), a gene encoding a putative n-acetyltransferase with a mechanistically uncharacterized role downstream of EIN3 in regulating a range of ethylene-regulated traits including apical hook development and *in vitro* regeneration. Also downstream of EIN3 is positive and negative regulation of numerous genes across nine hormone pathways, suggesting that EIN3 represents a key modulator of hormone crosstalk (Chang *et al.*, 2013). In support of this, we present at least eight gene candidates implicated as interacting directly or indirectly with EIN3 and upstream regulators of EIN3 (Fig. 9). Our results, considered together with mutant studies in Arabidopsis, suggest that these candidates mediate crosstalk between ethylene, jasmonic acid (JA), and salicylic acid (SA) signaling pathways.

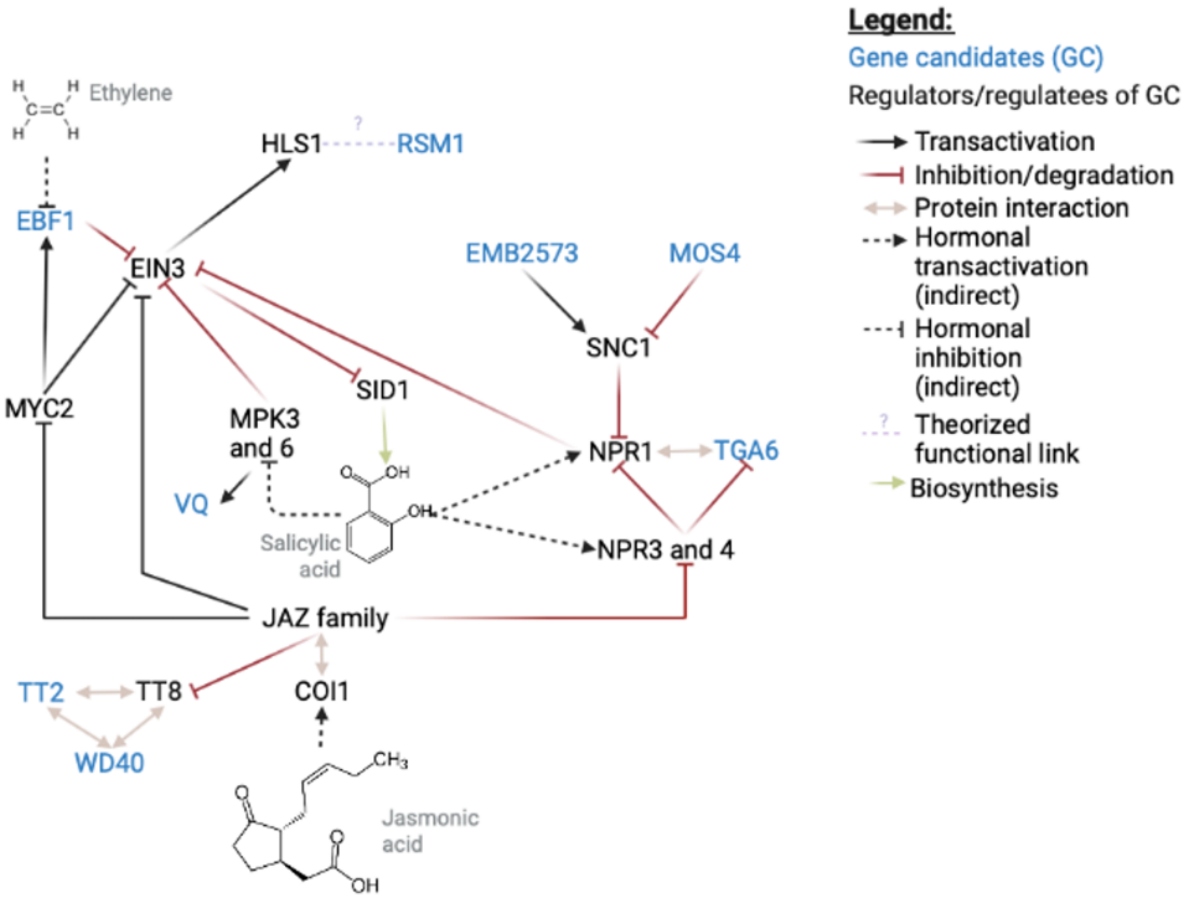

Figure 9. Interactions involving Arabidopsis homologs of eight gene candidates and associated regulators were identified by literature review, providing an understanding of the broader context of hormone crosstalk between ethylene, JA, and SA pathways as they relate to regeneration. Node placement was assisted by the Force Atlas 2 algorithm as implemented in Gephi. This Fig. was produced using BioRender (biorender.com). Standard acronyms and abbreviations can be found on The Arabidopsis Information Resource (TAIR; 2022) and are listed in Table S9. Evidence for interactions is summarized in Table S10.

Our GWAS results suggest a central role for salicylic acid (SA) and related genes. NPR1 is a regulator of salicylic acid signaling via a mechanism that depends on at least three genes with homologs implicated by QTLs in our GWAS (Fig. 9). Gene candidate Potri.003G194600 encodes a homolog of TGACGT motif transcription factor TGA6. TGA6 and other redundant members of the TGA family have been reported to regulate transcription of *NPR1* (Hussain *et al.*, 2018), in addition to interacting with NPR1 (Boyle *et al.*, 2009) to form a histone

acetyltransferase complex responsible for SA-associated epigenetic reprogramming (Jin *et al.*, 2018). Simultaneous knockout of functionally redundant TGAs (Zhang *et al.*, 2003b) or of NPR1 (Cao *et al.*, 1997) confers a loss of SA signaling, including SA-mediated pathogen resistance. Contrarily, constitutive SA signaling, dwarf morphology, and enhanced pathogen resistance results from knockout of the upstream regulator SUPPRESSOR OF NPR1, *CONSTITUTIVE 1 (SNC1*) (Zhang *et al.*, 2003a; Yang & Hua, 2004). This phenotype is reversed by concurrent knockout of MODIFIER OF SNC1,4 (MOS4), a homolog of our gene candidate Potri.015G041800 (Palma *et al.*, 2007). Whereas *mos4* reverses the dwarf phenotype of *snc1*, this double-mutant phenotype is itself reversed with concurrent partial loss-of-function of the n-acetyltransferase EMBRYO DEFECTIVE 2573 (EMB2573; a homolog of gene candidate Potri.001G177801), restoring the dwarf morphology. Knockout of EMB2573 also confers a wide range of defects including in embryo differentiation, notably in the shoot apical meristem (SAM), as indicated by abolished expression of the SAM marker SHOOT MERISTEMLESS (Chen *et al.*, 2018). MOS4 and EMB2573 are believed to regulate degradation of SNC1 in addition to other genes involved in related SA-signaling roles (Xu *et al.*, 2015).

Additional regulation of EIN3 is believed to exist via phosphorylation of EIN3 protein, which is mediated by two known mechanisms, one of which is via the SA-regulated MAP KINASE 3 (MPK3) and MAP KINASE 6 (MPK6). MPK3 and MPK6 are also responsible for phosphorylation of VQ MOTIF PROTEIN 3 and 4 (VQ3 and VQ4), which are the two most similar homologs of our gene candidate Potri.002G070600. Although VQ3 and VQ4 have not been studied in the context of *in vitro* regeneration, they are believed to function downstream of pathogen-associated molecular patterns (PAMPs) and upstream of pathogen defense genes (Yoo *et al.*, 2008; Pecher *et al.*, 2014). Finally, we note one additional gene among our candidates with a likely role in SA signaling. Potri.004G047700 is a homolog of NECROTIC SPOTTEN LESIONS 1 (NLS1), knockouts of which display a phenotype of increased SA accumulation and necrosis of leaves, particularly upon infection (Noutoshi *et al.*, 2006; Fukunaga *et al.*, 2017).

Our GWAS results also suggest a central role for anthocyanin and related genes. The salicylic acid and jasmonic acid pathways are linked with anthocyanin signaling by the activity of JAZ proteins in negatively regulating MYB/bHLH/WD40 (MBW) protein complexes responsible for transcriptional regulation of anthocyanin biosynthesis genes (Qi *et al.*, 2011). We report two gene candidates homologous to MBW components, Potri.002G173300 (encoding a

WD-40 repeat family protein) and Potri.018G049600 (encoding a homolog of TRANSPARENT TESTA 2). Although these genes have not been studied in the context of *in vitro* regeneration, MBWs regulate steps of anthocyanin biosynthesis immediately downstream of naringenin chalcone, which is produced by CHALCONE SYNTHASE (CHS); CHS knockout in Arabidopsis confers deficient *in vitro* shoot regeneration, with a light-dependent effect. The effects of anthocyanins on shoot regeneration may be mediated by their effects of ROS scavenging (Nameth *et al.*, 2013) and/or auxin accumulation (Brown *et al.*, 2001).

A functional relationship between HLS1 (previously described; downstream of EIN3) and RSM1 (homolog of gene candidate Potri.004G155400) has been proposed due to phenotypic similarities between *hls1* and RSM1-overexpressing Arabidopsis. Etiolated seedlings of both mutant lines presented various degrees of reduced hypocotyl length, reduced IAA content, defective hook formation and defective gravitropism (Hamaguchi *et al.*, 2008). However, whereas HLS1 knockout is known to confer enhanced shoot regeneration in Arabidopsis(Chatfield & Raizada, 2008), the effects of RSM1 or RSM family overexpression or knockout on shoot regeneration have not yet been reported.

Several gene candidates from GWAS appear to affect cytokinin signaling. Potri.010G105600 is a homolog of ARABIDOPSIS RESPONSE REGULATOR 2 (ARR2) that functions shortly downstream of cytokinin signaling. B-type ARRs such as ARR2 share some degree of functional redundancy and may each positively regulate *in vitro* regeneration via transcriptional upregulation of key developmental genes such as WUSCHEL (WUS) (Xie *et al.*, 2018; reviewed by Nagle *et al.*, 2018). An additional level of regulation over WUS expression exists via the FANTASTIC FOUR (FAF) gene family. Overexpression of any of the four FAF genes (including FAF1, homolog of gene candidate Potri.002G053400) leads to arrest of vegetative shoot meristem development, possibly by inhibiting WUS expression via an interaction with the feedback loop of regulation between WUS and the WUS inhibitor CLUVATA3 (Wahl *et al.*, 2010). Shoot meristem development is also regulated by the CYTOKININ RESPONSE FACTOR (CRF) gene family (featuring CRF4, a homolog of candidate Potri.012G032900), as shown by increased or reduced rosette growth when other members of the CRF family are knocked out or overexpressed, respectively. However, these experiments did not feature mutant analysis of the closely related CRF4 (Raines *et al.*, 2016).

Reactive oxygen species (ROS) signaling

At least two genes among our candidates appear to have roles in ROS regulation, which may affect regeneration and other developmental processes by mediating post-translational modifications of proteins involved in hormone signaling and/or by affecting levels of oxidative damage to developing tissues. Potri.011G031100 and Potr.012G070400 encode a putative thioredoxin-like protein and a peroxiredoxin, respectively. Although we did not find reports of mutant phenotypes for closely related genes in Arabidopsis in the context of regeneration or related processes, the thioredoxin DCC1 has been reported to affect *in vitro* shoot regeneration capacity in mutant lines as well as across natural ecotypes of Arabidopsis (Zhang *et al.*, 2018b).

Overlap with genes implicated from published GWAS analyses of regeneration

The candidates we identified showed very little similarity to results from related work. In prior work, GWAS was performed in 280 genotypes of *P. trichocarpa* to study traits related to *in vitro* callus regeneration. This study yielded eight candidate genes, none of which appear among our results (Tuskan *et al.*, 2018). A GWAS of traits related to roots and vegetative shoots in *Populus deltoides x simonii* with 434 genotypes produced 224 QTLs and multiple gene candidates were considered within proximity of each QTL, yielding a total of 595 unique gene candidates, only three of which were also found among traits analyzed in our study. Potri.015G018200, encoding a putative protein kinase, is a gene candidate from our analysis of callus area at week two as well as a prior analysis of a measurement of the number of leaves per vegetative shoot in *P. euphratica*. This leaf number trait also yields an association for Potri.004G156900, a putative RETICULATA-related protein also appearing as a candidate in our analysis of shoot area at week four. Another association is with Potri.019G035200, which encodes an oxygenase involved in heme degradation within chloroplasts; it was found among our gene candidates for callus at week two as well as in the same work for average stem diameter (Sun *et al.*, 2019).

In a review of GWAS of regeneration in diverse species, Lardon and Geelen (2020) noted that gene candidates identified across studies are non-overlapping to a great extent. Some of the potential causes for the low degree of overlap include genetic differences between study populations, variation in tissue or explant physiology, variation in the treatments used to promote regeneration, random variation in detection given underpowered statistics and numerous genes

602 under polygenic traits control, and differing statistical approaches (Lardon & Geelen, 2020). All
603 of these factors would apply to our study vs. the other published work in Populus. Another likely
604 contributor to lack of overlap is that our GWAS is the only one studying *in planta* regeneration,
605 as opposed to *in vitro* regeneration or vegetative shoot development, and the genetic control of
606 these developmental processes is likely to vary significantly.

607 **Conclusions**

608 We report a GWAS of *in planta* regeneration in *P. trichocarpa* using a novel system for
609 phenotyping regeneration with computer vision, along with four complementary statistical
610 methods for association mapping. These analyses revealed over 200 candidate genes, strongly
611 implicating regulators of cell adhesion and stress signaling. While canonical regulators of *in vitro*
612 regeneration tend to be involved in auxin and cytokinin signaling pathways, our results suggest
613 that stress pathways downstream of ethylene, salicylic acid, and jasmonic acid are of greatest
614 importance to the mode of *in planta* regeneration that we studied in *P. trichocarpa*. These
615 pathways have received little attention in studies where developmental regulator genes are used
616 to promote regeneration, and would appear to be promising avenues to pursue, at least in woody
617 species. Furthermore, at least eight top candidates are members of a genetic regulatory network,
618 separated from one another by no more than four degrees of direct interactions. This, considered
619 along with the complex nature of *in vitro* regeneration traits, suggests that emerging multi-locus
620 methods and epistasis tests may provide significantly greater insights into the polygenic control
621 of these traits.
622

623 **Acknowledgements**

624 We thank the National Science Foundation Plant Genome Research Program for support
625 (IOS #1546900, Analysis of genes affecting plant regeneration and transformation in poplar),
626 and members of GREAT TREES Research Cooperative at OSU for its support of the Strauss
627 laboratory.
628 Support for the Poplar GWAS dataset is provided by the U.S. Department of Energy,
629 Office of Science Biological and Environmental Research (BER) via the Center for Bioenergy
630 Innovation (CBI) under Contract No. DE-PS02-06ER64304. The Poplar GWAS Project used
631 resources of the Oak Ridge Leadership Computing Facility and the Compute and Data

Environment for Science at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

This work used the COMET high-performance cluster at the San Diego Supercomputing Center (University of California, San Diego) made available through the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562.

We thank biostars.org contributor "rmf" for providing a tutorial for generating LD decay curves using PLINK and R (https://www.biostars.org/p/300381/). This method was used to compute LD data for the three LD decay curves presented in this work.

Author Contributions

Strauss, Li, Jiang, and Muchero designed and directed the overall study, and obtained funding for its execution; Ma, Peremyslova, Magnuson, and Goddard designed and/or executed the phenotypic analyses; Nagle, Yuan, and Damanpreet created, adapted, and executed the machine vision, computation, and data analysis pipelines; Niño de Rivera assisted with inspecting results in IGV. Nagle wrote the manuscript with editing from Strauss, and all others contributed further edits and revisions.

Data Availability

Raw data and code used for this project is available upon request to the authors. MTMC-SKAT is available on GitHub (https://github.com/naglemi/mtmcskat).

References

Aich S, Stavness I. 2017. Leaf Counting With Deep Convolutional and Deconvolutional Networks. In: 2080–2089.

Altpeter F, Springer NM, Bartley LE, Blechl AE, Brutnell TP, Citovsky V, Conrad LJ, Gelvin SB, Jackson DP, Kausch AP, *et al.* 2016. Advancing Crop Transformation in the Era of Genome Editing. *The Plant Cell* **28**: 1510–1520.

An F, Zhao Q, Ji Y, Li W, Jiang Z, Yu X, Zhang C, Han Y, He W, Liu Y, *et al.* 2010. Ethylene-induced stabilization of ETHYLENE INSENSITIVE3 and EIN3-LIKE1 is mediated by proteasomal

663 degradation of EIN3 binding F-box 1 and 2 that requires EIN2 in Arabidopsis. *The Plant Cell* **22**:
664 2384–2401.

665 **Armstrong JE, Crandell DR, Easterbrook DJ, Noble JB**. **1965**. Late Pleistocene Stratigraphy and
666 Chronology in Southwestern British Columbia and Northwestern Washington. *GSA Bulletin* **76**:
667 321–330.

668 **Bdeir R, Muchero W, Yordanov Y, Tuskan GA, Busov V, Gailing O**. **2019**. Genome-wide
669 association studies of bark texture in Populus trichocarpa. *Tree Genetics & Genomes* **15**: 14.

670 **Boyle P, Le Su E, Rochon A, Shearer HL, Murmu J, Chu JY, Fobert PR, Després C**. **2009**. The
671 BTB/POZ Domain of the Arabidopsis Disease Resistance Protein NPR1 Interacts with the
672 Repression Domain of TGA2 to Negate Its Function. *The Plant Cell* **21**: 3700–3713.

673 **Brown DE, Rashotte AM, Murphy AS, Normanly J, Tague BW, Peer WA, Taiz L, Muday GK**.
674 **2001**. Flavonoids Act as Negative Regulators of Auxin Transport in Vivo in Arabidopsis. *Plant*
675 *Physiology* **126**: 524–535.

676 **Cao H, Glazebrook J, Clarke JD, Volko S, Dong X**. **1997**. The Arabidopsis NPR1 gene that
677 controls systemic acquired resistance encodes a novel protein containing ankyrin repeats. *Cell*
678 **88**: 57–63.

679 **Chang KN, Zhong S, Weirauch MT, Hon G, Pelizzola M, Li H, Huang SC, Schmitz RJ, Urich MA,**
680 **Kuo D, *et al.* 2013**. Temporal transcriptional response to ethylene gas drives growth hormone
681 cross-regulation in Arabidopsis (D Weigel, Ed.). *eLife* **2**: e00675.

682 **Chatfield SP, Raizada MN**. **2008**. Ethylene and shoot regeneration: hookless1 modulates de
683 novo shoot organogenesis in Arabidopsis thaliana. *Plant Cell Reports* **27**: 655–666.

684 **Chen H, Li S, Li L, Wu W, Ke X, Zou W, Zhao J**. **2018**. Nα-Acetyltransferases 10 and 15 are
685 Required for the Correct Initiation of Endosperm Cellularization in Arabidopsis. *Plant and Cell*
686 *Physiology* **59**: 2113–2128.

687 **Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, Szpiro AA, Chen W, Brehm JM, Celedón**
688 **JC, *et al.* 2016**. Control for Population Structure and Relatedness for Binary Traits in Genetic
689 Association Studies via Logistic Mixed Models. *The American Journal of Human Genetics* **98**:
690 653–666.

691 **Chen Y, Wu H, Yang W, Zhao W, Tong C**. **2021**. Multivariate linear mixed model enhanced the
692 power of identifying genome-wide association to poplar tree heights in a randomized complete
693 block design. *G3 Genes|Genomes|Genetics* **11**.

694 **Chhetri HB, Furches A, Macaya-Sanz D, Walker AR, Kainer D, Jones P, Harman-Ware AE,**
695 **Tschaplinski TJ, Jacobson D, Tuskan GA, *et al.* 2020**. Genome-Wide Association Study of Wood
696 Anatomical and Morphological Traits in Populus trichocarpa. *Frontiers in Plant Science* **11**: 1391.

Costa M, Pereira AM, Pinto SC, Silva J, Pereira LG, Coimbra S. **2019**. In silico and expression analyses of fasciclin-like arabinogalactan proteins reveal functional conservation during embryo and seed development. *Plant Reproduction* **32**: 353–370.

Deng W, Luo K, Li Z, Yang Y. **2009**. A novel method for induction of plant regeneration via somatic embryogenesis. *Plant Science* **177**: 43–48.

Di Marsico M, Paytuvi Gallart A, Sanseverino W, Aiese Cigliano R. **2022**. GreeNC 2.0: a comprehensive database of plant long non-coding RNAs. *Nucleic Acids Research* **50**: D1442–D1447.

Dobrescu A, Valerio Giuffrida M, Tsaftaris SA. **2017**. Leveraging Multiple Datasets for Deep Leaf Counting. In: 2072–2079.

Dutta A, Zisserman A. **2019**. The VIA Annotation Software for Images, Audio and Video. In: MM '19. Proceedings of the 27th ACM International Conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2276–2279.

Fukunaga S, Sogame M, Hata M, Singkaravanit-Ogawa S, Piślewska-Bednarek M, Onozawa-Komori M, Nishiuchi T, Hiruma K, Saitoh H, Terauchi R, *et al.* 2017. Dysfunction of Arabidopsis MACPF domain protein activates programmed cell death via tryptophan metabolism in MAMP-triggered immunity. *The Plant Journal: For Cell and Molecular Biology* **89**: 381–393.

Gordon-Kamm B, Sardesai N, Arling M, Lowe K, Hoerster G, Betts S, Jones T. **2019**. Using Morphogenic Genes to Improve Recovery and Regeneration of Transgenic Plants. *Plants* **8**: 38.

Hamaguchi A, Yamashino T, Koizumi N, Kiba T, Kojima M, Sakakibara H, Mizuno T. **2008**. A small subfamily of Arabidopsis RADIALIS-LIKE SANT/MYB genes: a link to HOOKLESS1-mediated signal transduction during early morphogenesis. *Bioscience, Biotechnology, and Biochemistry* **72**: 2687–2696.

Hussain RMF, Sheikh AH, Haider I, Quareshy M, Linthorst HJM. **2018**. Arabidopsis WRKY50 and TGA Transcription Factors Synergistically Activate Expression of PR1. *Frontiers in Plant Science* **9**: 930.

Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. **2013**. Sequence Kernel Association Tests for the Combined Effect of Rare and Common Variants. *American Journal of Human Genetics* **92**: 841–853.

Jaganathan D, Ramasamy K, Sellamuthu G, Jayabalan S, Venkataraman G. **2018**. CRISPR for Crop Improvement: An Update Review. *Frontiers in Plant Science* **9**.

Jin H, Choi S-M, Kang M-J, Yun S-H, Kwon D-J, Noh Y-S, Noh B. **2018**. Salicylic acid-induced transcriptional reprogramming by the HAC–NPR1–TGA histone acetyltransferase complex in Arabidopsis. *Nucleic Acids Research* **46**: 11712–11725.

Kaler AS, Gillman JD, Beissinger T, Purcell LC. **2020**. Comparing Different Statistical Models and Multiple Testing Corrections for Association Mapping in Soybean and Maize. *Frontiers in Plant Science* **10**: 1794.

Kusmec A, Schnable PS. **2018**. FarmCPUpp: Efficient large-scale genomewide association studies. *Plant Direct* **2**: e00053.

Kyritsis KA, Wang B, Sullivan J, Lyne R, Micklem G. **2019**. InterMineR: an R package for InterMine databases. *Bioinformatics (Oxford, England)* **35**: 3206–3207.

Lardon R, Geelen D. **2020**. Natural Variation in Plant Pluripotency and Regeneration. *Plants* **9**: 1261.

Lee T-H, Guo H, Wang X, Kim C, Paterson AH. **2014**. SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics* **15**: 162.

Liu X, Huang M, Fan B, Buckler ES, Zhang Z. **2016**. Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. *PLOS Genetics* **12**: e1005767.

López-Cortegano E, Caballero A. **2019**. Inferring the Nature of Missing Heritability in Human Traits Using Data from the GWAS Catalog. *Genetics* **212**: 891–904.

Maher MF, Nasti RA, Vollbrecht M, Starker CG, Clark MD, Voytas DF. **2020**. Plant gene editing through de novo induction of meristems. *Nature Biotechnology* **38**: 84–89.

Majda M, Robert S. **2018**. The Role of Auxin in Cell Wall Expansion. *International Journal of Molecular Sciences* **19**: 951.

Marowa P, Ding A, Kong Y. **2016**. Expansins: roles in plant growth and potential applications in crop improvement. *Plant Cell Reports* **35**: 949–965.

Muchero W, Sondreli KL, Chen J-G, Urbanowicz BR, Zhang J, Singan V, Yang Y, Brueggeman RS, Franco-Coronado J, Abraham N, *et al.* 2018. Association mapping, transcriptomics, and transient expression identify candidate genes mediating plant–pathogen interactions in a tree. *Proceedings of the National Academy of Sciences* **115**: 11573–11578.

Nagle M, Déjardin A, Pilate G, Strauss SH. **2018**. Opportunities for Innovation in Genetic Transformation of Forest Trees. *Frontiers in Plant Science* **9**.

Nameth B, Dinka SJ, Chatfield SP, Morris A, English J, Lewis D, Oro R, Raizada MN. **2013**. The shoot regeneration capacity of excised Arabidopsis cotyledons is established during the initial hours after injury and is modulated by a complex genetic network of light signalling. *Plant, Cell & Environment* **36**: 68–86.

National Academies of Sciences, Engineering, and Medicine, Division on Earth and Life Studies, Board on Agriculture and Natural Resources, Committee on Genetically Engineered Crops: Past Experience and Future Prospects. **2016**. *Genetically Engineered Crops: Experiences and Prospects*. p.406-443: National Academies Press.

Nguyen THN, Winkelmann T, Debener T. **2020**. Genetic analysis of callus formation in a diversity panel of 96 rose genotypes. *Plant Cell, Tissue and Organ Culture (PCTOC)* **142**: 505–517.

Noutoshi Y, Kuromori T, Wada T, Hirayama T, Kamiya A, Imura Y, Yasuda M, Nakashita H, Shirasu K, Shinozaki K. **2006**. Loss of Necrotic Spotted Lesions 1 associates with cell death and defense responses in Arabidopsis thaliana. *Plant Molecular Biology* **62**: 29–42.

Palma K, Zhao Q, Cheng YT, Bi D, Monaghan J, Cheng W, Zhang Y, Li X. **2007**. Regulation of plant innate immunity by three proteins in a complex conserved across the plant and animal kingdoms. *Genes & Development* **21**: 1484–1493.

Pecher P, Eschen-Lippold L, Herklotz S, Kuhle K, Naumann K, Bethke G, Uhrig J, Weyhe M, Scheel D, Lee J. **2014**. The Arabidopsis thaliana mitogen-activated protein kinases MPK3 and MPK6 target a subclass of 'VQ-motif'-containing proteins to regulate immune responses. *The New Phytologist* **203**: 592–606.

Qi T, Song S, Ren Q, Wu D, Huang H, Chen Y, Fan M, Peng W, Ren C, Xie D. **2011**. The Jasmonate-ZIM-domain proteins interact with the WD-Repeat/bHLH/MYB complexes to regulate Jasmonate-mediated anthocyanin accumulation and trichome initiation in Arabidopsis thaliana. *The Plant Cell* **23**: 1795–1814.

Raines T, Shanks C, Cheng C-Y, McPherson D, Argueso CT, Kim HJ, Franco-Zorrilla JM, López-Vidriero I, Solano R, Vaňková R, *et al.* 2016. The cytokinin response factors modulate root and shoot growth and promote leaf senescence in Arabidopsis. *The Plant Journal* **85**: 134–147.

Raj A, Stephens M, Pritchard JK. **2014**. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**: 573–589.

Revell LJ. **2012**. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**: 217–223.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. **2011**. Integrative genomics viewer. *Nature Biotechnology* **29**: 24–26.

Russell BC, Torralba A, Murphy KP, Freeman WT. **2008**. LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision* **77**: 157–173.

Sun P, Jia H, Zhang Y, Li J, Lu M, Hu J. **2019**. Deciphering Genetic Architecture of Adventitious Root and Related Shoot Traits in Populus Using QTL Mapping and RNA-Seq Data. *International Journal of Molecular Sciences* **20**: 6114.

Tange O. **2020**. *GNU Parallel 20201122 ('Biden'); GNU Parallel is a general parallelizer to run multiple serial command line programs in paralle lwithout changing them.* Zenodo.

The Arabidopsis Information Resource (TAIR). **2022**.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. **2013**. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**: 178–192.

Towns J, Cockerill T, Dahan M, Foster I, Gaither K, Grimshaw A, Hazlewood V, Lathrop S, Lifka D, Peterson GD, *et al.* 2014. XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering* **16**: 62–74.

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, *et al.* 2006. The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **313**: 1596–1604.

Tuskan GA, Mewalal R, Gunter LE, Palla KJ, Carter K, Jacobson DA, Jones PC, Garcia BJ, Weighill DA, Hyatt PD, *et al.* 2018. Defining the genetic components of callus formation: A GWAS approach. *PLoS ONE* **13**: e0202519.

Vsevolozhskaya OA, Hu F, Zaykin DV. **2019**. Detecting Weak Signals by Combining Small P-Values in Genetic Association Studies. *Frontiers in Genetics* **0**.

Wahl V, Brand LH, Guo Y-L, Schmid M. **2010**. The FANTASTIC FOUR proteins influence shoot meristem size in Arabidopsis thaliana. *BMC Plant Biology* **10**: 285.

Weber JC, Stettler RF. **1981**. Isoenzyme variation among ten populations of Populus trichocarpa Torr. et Gray in the Pacific Northwest. *Silvae Genetica* **30**: 82–87.

Weighill D, Tschaplinski TJ, Tuskan GA, Jacobson D. **2019**. Data Integration in Poplar: 'Omics Layers and Integration Strategies. *Frontiers in Genetics* **10**: 874.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. **2011**. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *American Journal of Human Genetics* **89**: 82–93.

Xiao Q, Bai X, Zhang C, He Y. **2021**. Advanced high-throughput plant phenotyping techniques for genome-wide association studies: A review. *Journal of Advanced Research*.

Xie M, Chen H, Huang L, O'Neil RC, Shokhirev MN, Ecker JR. **2018**. A B-ARR-mediated cytokinin transcriptional network directs hormone cross-regulation and shoot development. *Nature Communications* **9**: 1604.

Xu F, Huang Y, Li L, Gannon P, Linster E, Huber M, Kapos P, Bienvenut W, Polevoda B, Meinnel T, *et al.* 2015. Two N-terminal acetyltransferases antagonistically regulate the stability of a nod-like receptor in Arabidopsis. *The Plant Cell* **27**: 1547–1562.

Yang W, Guo Z, Huang C, Wang K, Jiang N, Feng H, Chen G, Liu Q, Xiong L. 2015. Genome-wide association study of rice (Oryza sativa L.) leaf traits with a high-throughput leaf scorer. *Journal of Experimental Botany* **66**: 5605–5615.

Yang S, Hua J. 2004. A Haplotype-Specific Resistance Gene Regulated by BONZAI1 Mediates Temperature-Dependent Growth Control in Arabidopsis. *The Plant Cell* **16**: 1060–1071.

Yates TB, Feng K, Zhang J, Singan V, Jawdy SS, Ranjan P, Abraham PE, Barry K, Lipzen A, Pan C, *et al.* 2021. The Ancient Salicoid Genome Duplication Event: A Platform for Reconstruction of De Novo Gene Evolution in Populus trichocarpa. *Genome Biology and Evolution* **13**: evab198.

Yoo S-D, Cho Y-H, Tena G, Xiong Y, Sheen J. 2008. Dual control of nuclear EIN3 by bifurcate MAPK cascades in C2H4 signalling. *Nature* **451**: 789–795.

Yuan J, Kaur D, Zhou Z, Nagle M, Kiddle NG, Doshi NA, Behnoudfar A, Peremyslova E, Ma C, Strauss SH, *et al.* 2022. Robust High-Throughput Phenotyping with Deep Segmentation Enabled by a Web-Based Annotator. *Plant Phenomics* **2022**.

Zhang Y, Goritschnig S, Dong X, Li X. 2003a. A gain-of-function mutation in a plant disease resistance gene leads to constitutive activation of downstream signal transduction pathways in suppressor of npr1-1, constitutive 1. *The Plant Cell* **15**: 2636–2646.

Zhang Q, Su Z, Guo Y, Zhang S, Jiang L, Wu R. 2020. Genome-wide association studies of callus differentiation for the desert tree, Populus euphratica. *Tree Physiology* **40**: 1762–1777.

Zhang Y, Tessaro MJ, Lassner M, Li X. 2003b. Knockout Analysis of Arabidopsis Transcription Factors TGA2, TGA5, and TGA6 Reveals Their Redundant and Essential Roles in Systemic Acquired Resistance. *The Plant Cell* **15**: 2647–2653.

Zhang J, Yang Y, Zheng K, Xie M, Feng K, Jawdy SS, Gunter LE, Ranjan P, Singan VR, Engle N, *et al.* 2018a. 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* **220**: 502–516.

Zhang H, Zhang TT, Liu H, Shi DY, Wang M, Bie XM, Li XG, Zhang XS. 2018b. Thioredoxin-Mediated ROS Homeostasis Explains Natural Variation in Plant Regeneration1[OPEN]. *Plant Physiology* **176**: 2231–2250.

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**: 821–824.