GWAS supported by computer vision identifies large numbers of candidate regulators of *in planta* regeneration in *Populus trichocarpa*

Michael F. Nagle [ID],[1] Jialin Yuan,[2] Damanpreet Kaur,[2] Cathleen Ma,[1] Ekaterina Peremyslova,[1] Yuan Jiang,[3] Alexa Niño de Rivera,[1] Sara Jawdy [ID],[4,5] Jin-Gui Chen [ID],[4,5,6] Kai Feng,[4,5] Timothy B. Yates,[4,5,6] Gerald A. Tuskan,[4,5] Wellington Muchero [ID],[4,5,6] Li Fuxin,[2] Steven H. Strauss[1,]*

[1]Department of Forest Ecosystems and Society, Oregon State University, 321 Richardson Hall, Corvallis, OR 97311, USA
[2]Department of Electrical Engineering and Computer Science, Oregon State University, 1148 Kelley Engineering Center, Corvallis, OR 97331, USA
[3]Statistics Department, Oregon State University, 239 Weniger Hall, Corvallis, OR 97331, USA
[4]Biosciences Division, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831, USA
[5]Center for Bioenergy Innovation, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831, USA
[6]Bredesen Center for Interdisciplinary Research, University of Tennessee-Knoxville, 310 Ferris Hall 1508 Middle Dr, Knoxville, TN 37996, USA

*Corresponding author: Email: Steve.Strauss@OregonState.Edu

Plant regeneration is an important dimension of 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. Association mapping methods such as genome-wide association studies (GWAS) have long demonstrated abilities to help uncover the genetic basis of trait variation in plants; however, the performance of these methods depends on the accuracy and scale of phenotyping. To enable a large-scale GWAS of *in planta* callus and shoot regeneration in the model tree *Populus*, we developed a phenomics workflow involving semantic segmentation to quantify regenerating plant tissues over time. We found that the resulting statistics were of highly non-normal distributions, and thus employed transformations or permutations to avoid violating assumptions of linear models used in GWAS. We report over 200 statistically supported quantitative trait loci (QTLs), with genes encompassing or near to top QTLs including regulators of cell adhesion, stress signaling, and hormone signaling pathways, as well as other diverse functions. Our results encourage models of hormonal signaling during plant regeneration to consider keystone roles of stress-related signaling (e.g. involving jasmonates and salicylic acid), in addition to the auxin and cytokinin pathways commonly considered. The putative regulatory genes and biological processes we identified 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.

Keywords: GWAS; poplar; regeneration; computer vision; machine vision; phenomics

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 transformation and regeneration (TR), typically via *in vitro* tissue culture methods. Various *in planta* approaches show promise to enable efficient TR without labor-intensive tissue culturing or the materials and sterile conditions it requires, but nonetheless require plant transformation and organ regeneration (Maher *et al.* 2020; Kesiraju *et al.* 2021; Bomzan *et al.* 2022; Lian *et al.* 2022). TR requires developmental responses to a series of hormone treatments and amenability to gene insertion, and the capacities for both vary greatly between and within species (Altpeter *et al.* 2016). The causes of this great variation in recalcitrance are poorly known. However, genome-wide association studies (GWAS)—with their potential to identify genetic

variants with key roles in capacity for TR—can greatly enhance understanding of the TR process. In addition, putatively causal genes represented by variants may serve as "reagents" for overcoming recalcitrance. Overexpression of morphogenic regulator (MR) genes has been found to improve transformation and/or regeneration in over 30 plant species (Dai *et al.* 2017; Lee and Wang 2023), including *Populus* spp. (Deng *et al.* 2009; Liu *et al.* 2018; Pan *et al.* 2022), in many cases with the same MRs or their homologs proving effective, and suggesting at least some level of mechanistic conservation across both *in vitro* and *in planta* pathways (Deng *et al.* 2009; Maher *et al.* 2020; Lian *et al.* 2022). However, given the complexity and genotypic variation in TR 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 TR, 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

(Muchero *et al.* 2018; Tuskan *et al.* 2018; Bdeir *et al.* 2019; J. Zhang *et al.* 2018; Weighill *et al.* 2019; Chhetri *et al.* 2020). We focused on regeneration from cut stems, while considering these 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 planta* or *in vitro*. GWAS and related association mapping methods have previously been applied to study variation in the rates of *in vitro* regeneration in Arabidopsis, rice, wheat, rose, sorghum, and poplar, among other plants (reviewed by Lardon and Geelen 2020).

Regeneration phenotypes are notoriously difficult to quantify, whether *in planta* 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 differentiation 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 consideration in their GWAS of callus formation across 96 rose genotypes.

Because of the importance of large and precise phenotype datasets for statistical power in GWAS (López-Cortegano and Caballero 2019), we developed a computer vision method to measure regeneration from sequential images of decapitated, regenerating stems. Over 40 published studies have made use of diverse computer vision methods in GWAS of plants, including Arabidopsis, maize, wheat, rice, sorghum, soybean, and barley, among others (Xiao *et al.* 2021; Affortit *et al.* 2022; Guo *et al.* 2022). Developmental traits related to biomass and growth of whole plants or specific tissues have commonly been studied using various methods for segmentation of RGB images, including those involving color thresholds (Das *et al.* 2015; Yang *et al.* 2015; Al-Tamimi *et al.* 2016; Gage *et al.* 2018; Guo *et al.* 2018; Pham *et al.* 2019; Campbell *et al.* 2020; Seethepalli *et al.* 2020; Ogawa *et al.* 2021; Affortit *et al.* 2022), machine learning approaches such as support vector machines and random forests (She *et al.* 2019; Zou *et al.* 2019; Carlier *et al.* 2022; Xu *et al.* 2022), and deep convolutional neural networks (DCNNs; Ma *et al.* 2019; Yasrab *et al.* 2019; Kamath *et al.* 2022; Xu *et al.* 2022). In recent years, DCNNs similar to that employed in the present study have demonstrated abilities to outperform earlier methods (Adams *et al.* 2020) and have become the dominant approach used for diverse computer vision 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 and Stavness 2017; Dobrescu *et al.* 2017).

Association mapping methods such as GWAS rely on several key assumptions and considerations related to statistical characteristics of data. These include population stratification, which is common in wild and experimental populations, and can lead to inflation of error rates if not adequately controlled by covariates such as principal components (PCs) or relatedness (kinship) matrices derived from genomic data (Price *et al.* 2006; Kang *et al.* 2008). Moreover, nonnormality of model residuals can compromise the accuracy of *P*-values computed from regression models. This concern can be addressed by methods such as transformation of trait data to normality (Beasley *et al.* 2009), or alternatively by permutation-based methods to compute empirical *P*-values (Good 2013).

We report a GWAS of *in planta* regeneration in poplar, using a computer vision workflow based on deep segmentation to quantify specific regenerating tissues, and four association mapping methods to identify genetic markers associated with regeneration traits (Fig. 1). The non-normal and skewed characteristics of segmentation-based traits presented a challenge for GWAS, which commonly relies on an assumption of normally distributed residuals. Due to this, we tested multiple GWAS methods with adaptations to avoid violating this assumption, while avoiding or mitigating overcorrections that may result from transformations. We found over 200 statistically supported quantitative trait loci (QTLs), with suspected causal genes representing diverse physiological roles that include hormone signaling, plant stress response, control of cell division, and cell wall structure—as well as many genes of unknown function.

Materials and methods
Experimental design
A main objective of our work here was to develop, test, and refine workflows using computer vision to measure developing tissues during regeneration. The system also needed to be sensitive enough to enable detailed genetic characterization of regeneration traits. Toward this end, we first adapted a DCNN of the PSPNet architecture (Zhao *et al.* 2017) to compute segmentation masks for regenerating plant tissues (callus, shoot, and unregenerated stem) on a high-throughput scale. Second, we evaluated the statistical considerations of using these traits in GWAS, whether by models often relying on trait transformation to avoid violating assumptions of normally distributed residuals, or by models that avoid this assumption altogether by dichotomization or permutation and computation of empirical *P*-values. Finally, we applied these methods in a population with over 1,200 wild poplar genotypes ideal for GWAS due to a remarkable number and density of single-nucleotide polymorphisms (SNPs; up 34 million depending on GWAS method) and extremely low linkage disequilibrium (LD) (Muchero *et al.* 2018; Tuskan *et al.* 2018; J. Zhang *et al.* 2018; Bdeir *et al.* 2019; Weighill *et al.* 2019; Chhetri *et al.* 2020). An overview of the experimental population and analysis pipeline is shown in Fig. 1.

Plant materials
We utilized an expanded version of the previously reported resequenced *P. trichocarpa* GWAS population (Muchero *et al.* 2018; Tuskan *et al.* 2018; J. Zhang *et al.* 2018; Bdeir *et al.* 2019; Weighill *et al.* 2019; Chhetri *et al.* 2020). The population was recently expanded to include an additional 441 genotypes, filling a geographical gap that existed in the first release with 882 genotypes (Fig. 2). While this clone bank is kept at multiple locations, phenotyping in this study only made use of the replicate in Corvallis, Oregon, 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, Oregon: 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 winters of 2018, 2019, and 2020, frozen, and then rooted up to 1 year later. A second greenhouse population was established and allowed to go dormant in the winter; plants from this source were occasionally used to replace plants in the main greenhouse population that provided plant materials throughout the year.

Sequencing and SNP dataset preparation
We performed SNP calling and analyzed the distribution of SNPs after resequencing was completed for the population, with a total of 1,323 genotypes in the complete SNP dataset. DNA short-read
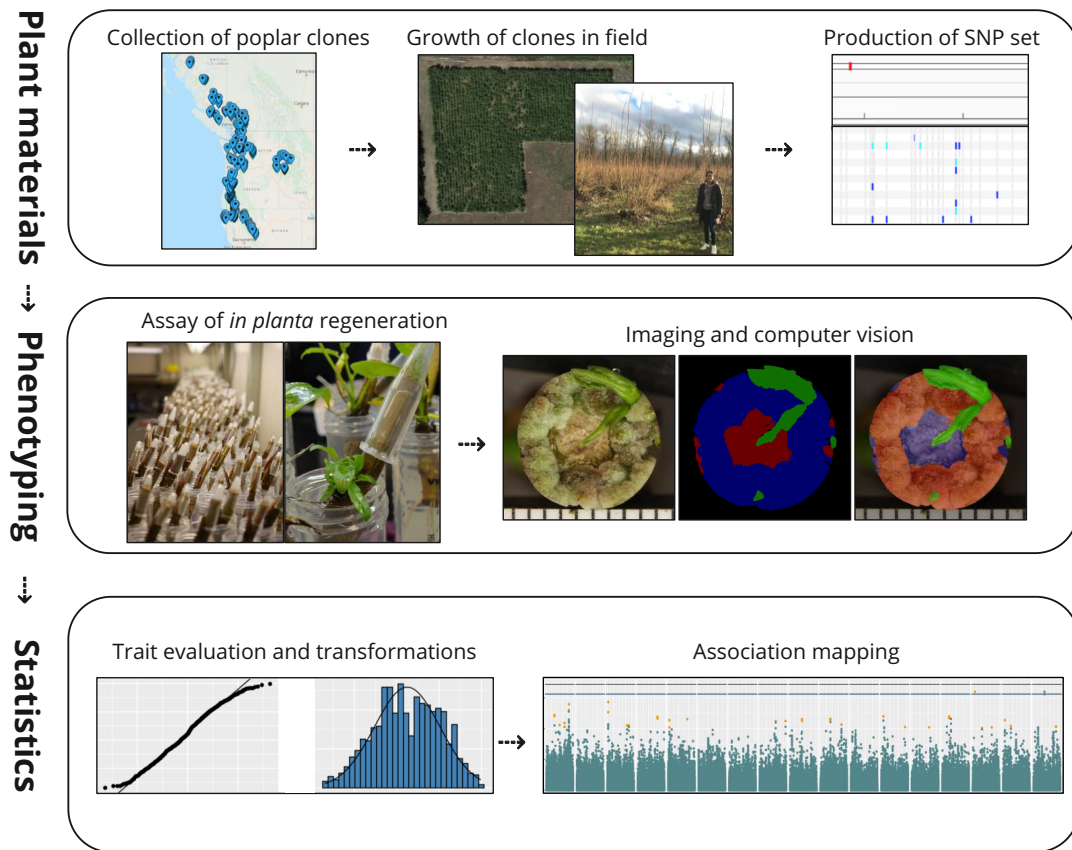
Fig. 1. Overview of *in planta* callus and shoot regeneration GWAS workflow. The first row describes resources that were provided through other work prior to undertaking the GWAS. Images on the right of the "Phenotyping" panel are the RGB image, ground truth semantic label and overlay for one sample.

sequencing data for all 1,323 genotypes was filtered and trimmed with BBDuk (https://sourceforge.net/projects/bbmap/). Parameters used for BBDuk included "kmer = 25", "ktrim = r", and "trimq = 6". Clean reads were then aligned to the *P. trichocarpa* v3.0 reference genome with bwa-mem v0.7.17 (Li and Durbin 2009). Duplicate reads were then marked with Sambamba markdup v0.7.0 (Tarasov *et al.* 2015). Next, base quality score recalibration was performed with GATK v4.1.3.0 BaseRecalibrator (Van der Auwera and O'Connor 2020). Variants were then called by sample with GATK v4.1.3.0 HaplotypeCaller and genomic variant call format files were generated (Van der Auwera and O'Connor 2020). The GVCFs were then consolidated with GATK v4.1.3.0 GenomicsDBImport and joint calling was performed with GATK v4.1.3.0 Genotype GVCFs. Next, SNPs and insertion–deletion variants (indels) were separated with GATK v4.1.3.0 SelectVariants. Indels were filtered with GATK v4.1.3.0 VariantFiltration with the following thresholds and flags: "-filter "QD < 2.0" –filter-name "QD2" -filter "FS > 200.0" –filter-name "FS200" -filter "ReadPosRankSum < −20.0" –filter-name "ReadPosRankSum-20". SNPs were filtered with GATK v4.1.3.0 VariantFiltration with the following thresholds and flags, -filter "QD < 2.0" –filter-name "QD2" -filter "SOR > 4.0" –filter-name "SOR4" -filter "FS > 60.0" –filter-name "FS60" -filter "MQ < 40.0" –filter-name "MQ40" -filter "MQRankSum < −12.5" –filter-name "MQRankSum-12.5" -filter "ReadPosRankSum < −8.0" –filter-name "ReadPosRankSum-8". There was a total of 40.4 M SNPs prior to filtering for minor allele frequency and additional quality criteria. The density and consistency of SNP data on each chromosome were assessed using the R package Cmplot (Yin *et al.* 2021; Supplementary Fig. 1) 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 50 mL Falcon tubes with water for 5 weeks. Based on preliminary experiments (data not shown), we found that treatment of the decapitated surface of each stem with 100 μL of 0.5 mg/mL thidiazuron (TDZ) in water improved regeneration considerably (37% of genotypes produced shoots, compared to 24% without TDZ). After application of TDZ to a given stem tip, a 1.5 mL 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, decapitated stem surfaces 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, randomly selected 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 4 and 5 in the first phase and week 4 in the third phase. There were two replicate plants measured for all but the first three phases, where only a single replicate was used.

Image segmentation

To perform annotation of images for computer vision, 249 images were randomly sampled from the first seven phases and manually annotated using the Intelligent Deep Annotator for Segmentation (IDEAS) graphical user interface that we previously developed (Yuan *et al.* 2022). As described in our prior work, these samples
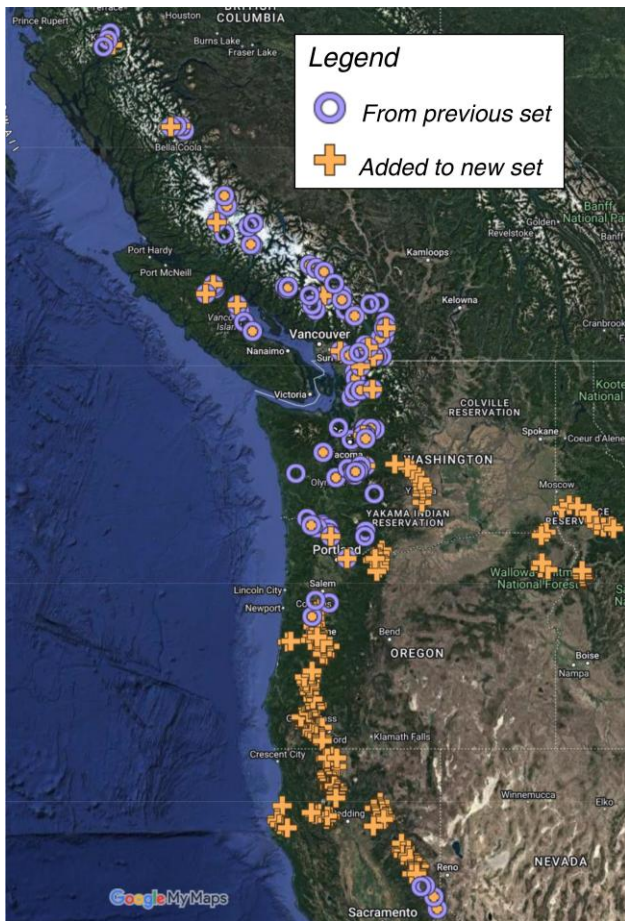
Fig. 2. Origins of *P. trichocarpa* clones used to generate SNP dataset. A total of 1,323 native genotypes were collected over a geographical range across the Pacific Northwest region of the USA and the southwest of Canada. Tree location is shown for 1,301 genotypes for which precise location data are available.

were used to train a convolutional neural network (PSPNet) via TensorFlow v1.14 to segment images of regenerating stem tips with each pixel labeled as one of four classes: callus, shoot, unregenerated stem and background. The trained model was deployed to produce inferred semantic masks for 11,791 images in the complete GWAS dataset, including 1,219 clones, most of which had two replicates, across four weekly timepoints (Fig. 3). 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").

Computer vision trait preparation for GWAS

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

Additional traits were computed by performing principal component analysis (PCA) using "stats::prcomp" in R (R Core Team 2022) 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. Since traits entered into PCA were computed as proportions of given tissues out of total plant area (lower limit of 0 and upper limit

of 1) and were thus on the same meaningful scale, rescaling was not performed before PCA, although traits were centered. Scree plots were evaluated to estimate the number of PCs representing significant proportions of trait variation, and loadings were inspected to interpret the trends across traits represented by PCs.

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, using base R and the "ggplot2" package (Wickham 2016; R Core Team 2022). To avoid severe violations of normality that may lead to inflated error rates, all traits were transformed prior to GWAS methods that depend on normality and linear model assumptions. The most basic transformation applied was a removal of zero values followed by Box–Cox transformation (Box and Cox 1964) using the "car" package in R (Fox and Weisberg 2018). 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 the Box–Cox transformation. In addition, as an alternative to Box–Cox transformations, rank-based inverse normal (RB-INV) transformations were performed for difficult distributions (Supplementary Fig. 2; Supplementary Tables 1 and 2).

Association mapping

Because of the distinct assumptions and data types for which various GWAS methods are suited, along with the non-normality of computer-vision-based traits and the risk of overcorrection with transformations, we employed an analysis pipeline that made use of four GWAS methods. These include Genome-wide Efficient Mixed Model Association (GEMMA) (Zhou and Stephens 2012), which was parallelized using the GNU Parallel (Tange 2022) framework to simultaneously run each given trait on a CPU core. Due to an assumption of normally distributed model residuals for GEMMA, this method was only used to study versions of traits that had been transformed toward normality, including the exclusion of all samples with zero values. The Generalized Mixed Model Association Test (GMMAT; Chen *et al.* 2016) was used for single-marker tests with the same kinship covariate as GEMMA; however, rather than using continuous trait variables, GMMAT applies logistic regression and works with dichotomized traits. Finally, as a means of performing tests that avoid any transformation or dichotomization of traits, we applied our Multi-Threaded Monte Carlo Sequence Kernel Association Test (MTMC-SKAT; github.com/naglemi/mtmcskat), an R extension of the established SKAT method (Wu *et al.* 2011; Ionita-Laza *et al.* 2013) extending support for efficient resampling on high-performance clusters. Using MTMC-SKAT, we tested for the combined effect of adjacent SNPs within 3 kb windows (staggered by 1 kb) and computed empirical P-values for SNPs indicated by an initial parametric test as being strongly associated with a trait. With MTMC-SKAT, we tested and compared the control for population structure via two types of covariates: a matrix of PCs from PCA of SNP data (P model; Supplementary Figs 3 and 4) and a matrix of theoretical subpopulation ancestry estimates from fastSTRUCTURE (Q model; Supplementary Fig. 5) (Raj *et al.* 2014). MTMC-SKAT was deployed on the COMET high-performance cluster made available through the NSF XSEDE program (Towns *et al.* 2014). For all GWAS methods, we included covariates to represent the diameters of stem cuttings and the "phase" of phenotyping (as samples were divided into different phases due to practical limitations in concurrent phenotyping,
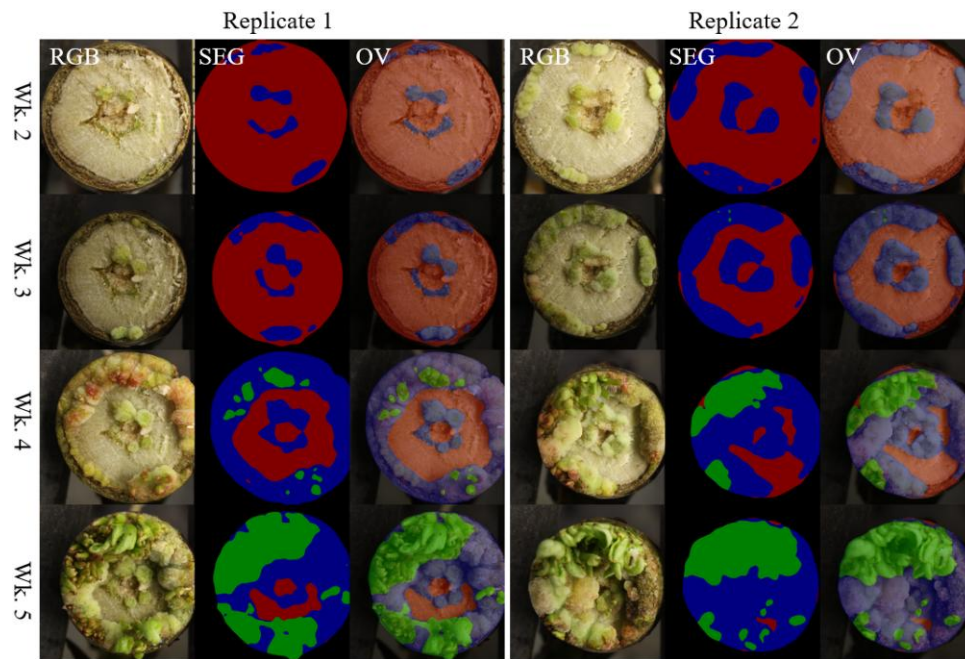
Fig. 3. False-color examples of segmentation from regenerating stems shown in cross-section. For each section, shown are RGB images (left panels), inferred semantic mask labels (SEG, middle panels), and overlays of images and masks (OV, right panels) for two replicates of genotype BESC-1013 across four timepoints of development, from week 2 (top) to week 5 (bottom). Semantic masks show unregenerated stem tissue (red), callus (blue), and shoot (green).

as previously discussed). Additional details on GWAS methods can be found in Supplementary Methods (File 1).

Statistically significant associations from the various pipelines were first determined by computing Benjamini–Hochberg false discovery rate (FDR; computed at 0.10 using "stats::p.adjust" from base R; Benjamini and Hochberg 1995; R Core Team 2022) and Bonferroni thresholds ($\alpha = 0.05$) (Dunn 1959). SNPs not found on contiguous assembled chromosomes were excluded from computation from this point forward (approximately 3.6% of total SNPs). The Bonferroni thresholds were computed given the number of tests equal to the number of filtered SNPs (for single-marker tests GEMMA and GMMAT) or the number of 3 kb SNP windows (staggered by 1 kb; in the case of SKAT). We then extracted lists of SNPs with P-values meeting these thresholds for interrogation.

We next evaluated the extent to which multiple nearby SNPs provided additive support for regional associations, 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 P-values for marginal effects of individual QTLs from GEMMA and GMMAT and identify cases where a SNP produces a P-value below $1*10^{-5}$ and is within 500 bp of at least 5 additional SNPs with P-values below $1*10^{-4}$ when considering the upper half of top-ranking SNPs (according to P-values). Due to the computational expense of computing relatively high-accuracy Wald P-values as opposed to score tests in GMMAT, we applied ART to GMMAT results over two rounds. First, ART was applied to GMMAT score test results, and then GMMAT Wald P-values were computed for SNPs composing significant ART windows, followed at last by recomputation of ART P-values. While we initially computed Wald P-values with GMMAT for only the top 100 or 1,000 SNPs according to a score test, additional Wald P-values were computed for use in ART. For each of these 1 kb 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 $\sim1.27*10^{-7}$ was computed using the number of independent tests ($\sim3.94*10^6$ 1 kb 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 $\sim4*10^{-9}$, given up to ~13 million SNPs on contiguous chromosomes remaining after filtering; Supplementary File 1, Supplementary Methods). To be emphasized, it is well-known that Bonferroni thresholds for individual SNPs consider each SNP as an independent test, leading to overcorrection since large numbers of SNPs are in LD.

To determine on a high-throughput scale which genes are possibly causal with respect to statistically significant QTLs, we used R scripts to reference genome and genome annotation data available through Phytozome (phytozome.jgi.doe.gov) (Tuskan et al. 2006). In this workflow, the positions 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 30 kb 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. We note that the confidence in this assumption is generally limited, but considering the extremely rapid LD decay in the poplar GWAS population (Fig. 5), the assumption can reasonably be expected to be safe for the majority of QTLs and candidate genes in this study. The R package InterMineR (Kyritsis et al. 2019) was used to collect Phytozome data on gene function, Arabidopsis homologs, and gene ontology terms. The GreeNC database was used to identify possible noncoding regulatory RNAs among candidate genes (Di Marsico et al. 2022). For the top candidate genes, particularly those passing the conservative Bonferroni or FDR (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
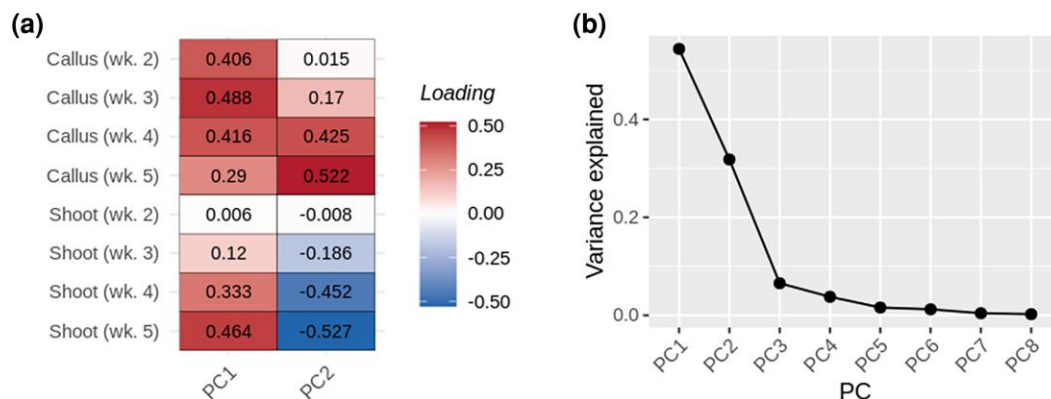
(a)

(b)

Fig. 4. PCA of *in planta* callus and shoot regeneration traits. Results are shown from PCA over all callus and shoot traits across four weekly timepoints. a) Heat map of loadings from PCA; b) Scree plot.
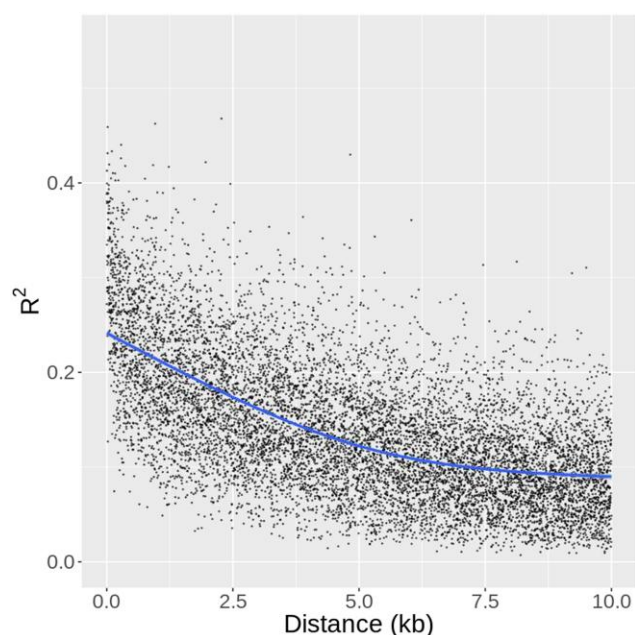

Fig. 5. Linkage disequilibrium decay. Each point represents the mean LD between SNPs of a given distance (x-axis). The blue line represents LD decay as computed with a spline function.

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.

Results
Trait transformations

Prior to transformations, most traits displayed marked non-normal characteristics as indicated by Q–Q plots, histograms, Shapiro–Wilk tests, and Pearson correlation coefficients of each distribution with a normal distribution with the same mean and standard deviation. The improvement in normality after transformation was marked in most cases (Supplementary Tables 1 and 2; Supplementary Fig. 2; data not shown). Non-normal characteristics of most traits were reduced substantially by excluding genotypes with zero values and applying a Box–Cox

transformation (e.g. Supplementary Fig. 2). For the traits of callus or shoot area at each timepoint, based on visual inspection of histograms and Q–Q plots, the improvement in metrics of normality was deemed adequate for linear models. All PCA-derived traits necessitated additional treatments to avoid severe deviations from normality, including removal of outliers and in some cases removal of values below an elbow in the frequency distribution (estimated as the position where the second derivative of the frequency distribution is maximum; Supplementary Table 2).

PCs as proxies for complex patterns of regeneration

Scree plots and heat maps of loadings revealed common trends in regeneration across timepoints and regenerating tissue types (callus and shoot). These results were obtained for three different sets of PCA with different groups of traits: (1) callus area across four timepoints; (2) shoot area across four timepoints (Supplementary Fig. 6); and (3) callus area and shoot area across four timepoints (Fig. 4). In all three cases, the PC explaining the most variation (PC1) represented a tendency of the tissue(s) included in PCA to regenerate across all timepoints. Latter PCs provided proxies for more complex patterns of regeneration. PC2 over callus traits appears to represent high levels of callus regeneration at early, but not later timepoints. PC2 over all callus and shoot traits appears to represent a tendency for callus to regenerate robustly, but fail to develop into shoots. Subsequent PCs, for each batch of traits, represented a relatively small proportion of variance explained and were thus not analyzed to find associations.

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

The SNP dataset produced for this population displays polymorphism across all regions of contiguous chromosomes represented in the reference genome (Supplementary Fig. 1). Poplar clones collected for the GWAS clone bank represent a wide range of geographic diversity, nearly spanning the natural range of *P. trichocarpa* from British Columbia to the Pacific Northwest of the United States (Fig. 2). There is clearly very strong natural intra-chromosomal recombination, as LD decay occurs rapidly and reaches $R^2 = 0.2$ within 2 kb whether computed for the whole population or either of the two most prominent subpopulations (Fig. 5; Supplementary Fig. 7). Population structure analysis with fastSTRUCTURE, cross-referenced with geography and phylogenetics, supported the existence of 6 or 7 ancestral subpopulations, which were most distinct in southern regions, while a greater
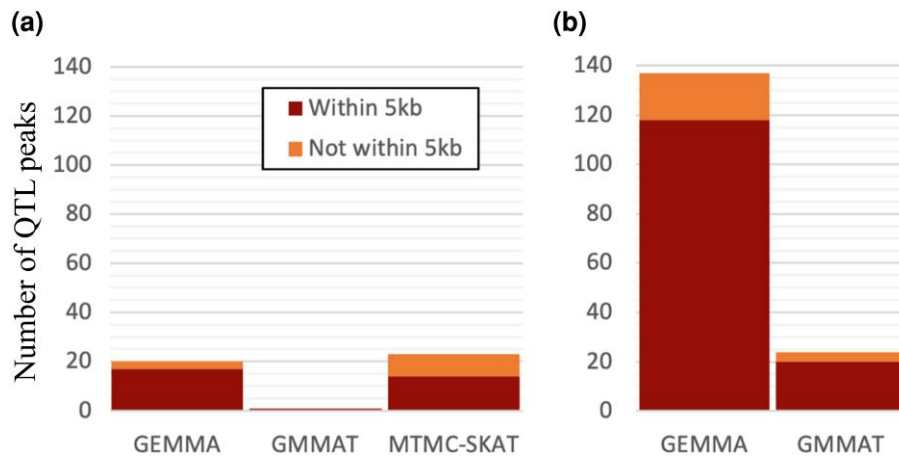
Fig. 6. Tallies of *in planta* callus and shoot associations across GWAS methods. Shown are barplots summarizing the numbers of associations from each GWAS method, with two types of significance thresholds, as well as within a 5 kb distance threshold of the nearest gene. QTL peaks were taken as the point 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 = 0.10) and/or conservative Bonferroni threshold; b) QTL peaks passing ART-Bonferroni threshold ($\alpha = 0.05$, N of 1 kb windows in genome).

degree of admixture was prevalent farther north and may be responsible for the significantly lesser LD in the "Oregon" subpopulation (*P* < 0.001; Supplementary Figs. 7–12). Our linear models supported a positive relationship between latitude and callus traits, but were unable to detect such a relationship for shoot traits (Supplementary Fig. 13; Supplementary Table 3; Supplementary File 1, Supplementary Results and Discussion).

Genes supported by significant QTLs

We performed association mapping with several distinct statistical approaches. Tests for associations of traits with individual SNPs were first undertaken with Genome-wide Efficient Mixed Model Association (GEMMA; Zhou and Stephens 2012) for continuous and transformed versions of traits, and with Generalized Mixed Model Association Test (GMMAT; Chen *et al.* 2016) for dichotomized traits. In addition, combined variant tests for SNP windows were performed using two methods. Our Multi-Threaded Monte Carlo Sequence Kernel Association Test (MTMC-SKAT; github.com/naglemi/mtmcskat), an R extension of the established SKAT method (Wu *et al.* 2011; Ionita-Laza *et al.* 2013), relies on computation of a combined test statistic for SNP windows integrated via kernelization, followed by an empirical *P*-value for each. In addition, we applied Augmented Rank Truncation (ART), which is a post hoc method applied after computation of initial *P*-values (in our case via GEMMA or GMMAT) for individual SNPs. It takes advantage of cases where the co-occurrence of significant SNPs within a window provides combined support for an association in the window (Vsevolozhskaya *et al.* 2019).

GEMMA also provided measures of the proportion of trait variance explained by additive genetic effects (modeled via a relatedness matrix under the null model), which we consulted as a form of narrow-sense SNP heritability (h^2_{SNP}). We interrogated traits with h^2_{SNP} above 0.10 for candidate genes (Supplementary Table 4), except for shoot area at week 2 (first timepoint) due to a sparse distribution and high level of noise for this trait resulting from very low levels of shoot regeneration at this early timepoint. We utilized three multiple-testing correction methods (from most to least conservative: conservative Bonferroni, Benjamini–Hochberg FDR, and ART-Bonferroni) to provide three levels of confidence in associations to support prioritization of candidate

genes. Across the three GWAS methods GEMMA, GMMAT, and (MTMC-)SKAT, we report a total of 13 unique QTL peaks with *P*-values passing the conservative Bonferroni significance threshold, as well as 44 passing the FDR (0.10) threshold. Among Bonferroni-significant associations, 6 are inside or within 5 kb of a gene found in the genome annotation, as well as 32 associations (~72.7%) meeting the FDR threshold (Figs. 6–8, Supplementary Tables 5 and 6). We found 137 unique QTL peaks from applying our implementation of ART to GEMMA results (Supplementary Tables 5 and 7), as well as 24 from applying ART to GMMAT results (Supplementary Tables 5 and 8). Several of the most promising candidates, based on the biology of suspected causal gene homologs in Arabidopsis, are shown in Table 1.

Discussion

High-throughput phenomics support scale and precision of GWAS

Our mapping of complex developmental traits relied on computer vision methods to precisely quantify traits that would be impractical to measure manually. Although the use of an ordinal scoring system instead of segmentation is an alternative approach, this would have risked the introduction of subjective biases and violation of linear model assumptions—while sacrificing much precision and detail. 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 and validation examples (249 images in total) with pixelwise labels for callus, shoot, and unregenerated tissues. Here, we applied this method on a high-throughput scale for GWAS and evaluated statistical consequences and necessary downstream adaptations, such as appropriate trait transformations and GWAS methods. Due to a dependence of linear models on an assumption of normally distributed residuals, which is prone to violation when dependent variables are themselves non-normal, transformation of dependent variables towards normality is a common strategy (McCaw et al. 2020; Schielzeth et al. 2020). We employed this approach for our highly skewed traits related to callus and shoot growth. We also avoided the assumption via use of MTMC-SKAT to compute
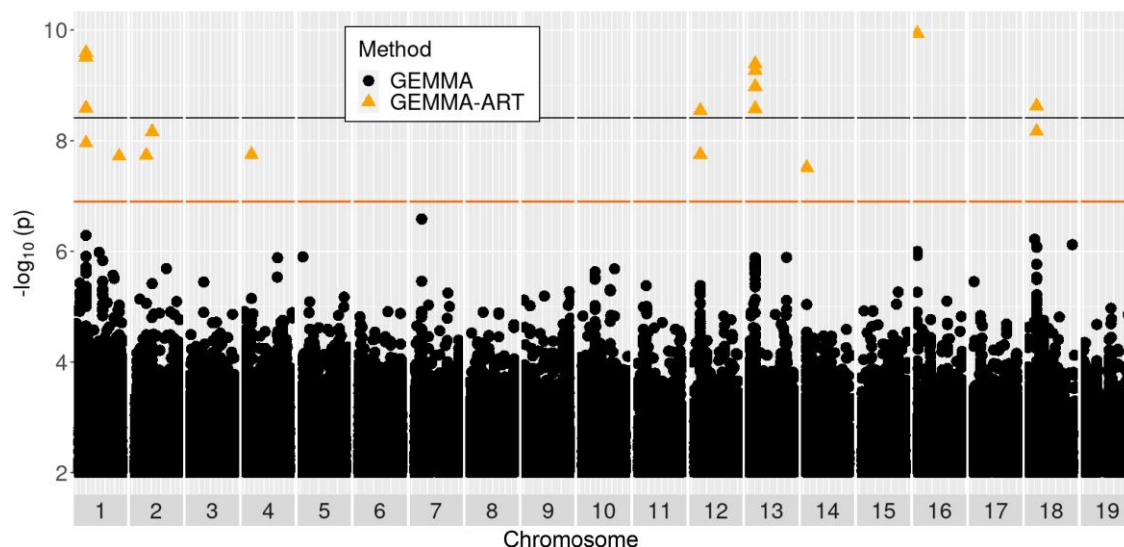
Fig. 7. Manhattan plot of GEMMA-ART results for week four callus area. Shown is a Manhattan plot for GEMMA results for the trait of callus area at week 4. Black and orange lines show Bonferroni significance thresholds for GEMMA results with independent SNPs, and for ART applied to GEMMA over 1 kb windows of SNPs, respectively. Black circles represent tests of individual SNPs by GEMMA, while orange triangles represent 1 kb windows tested by ART applied to GEMMA results.
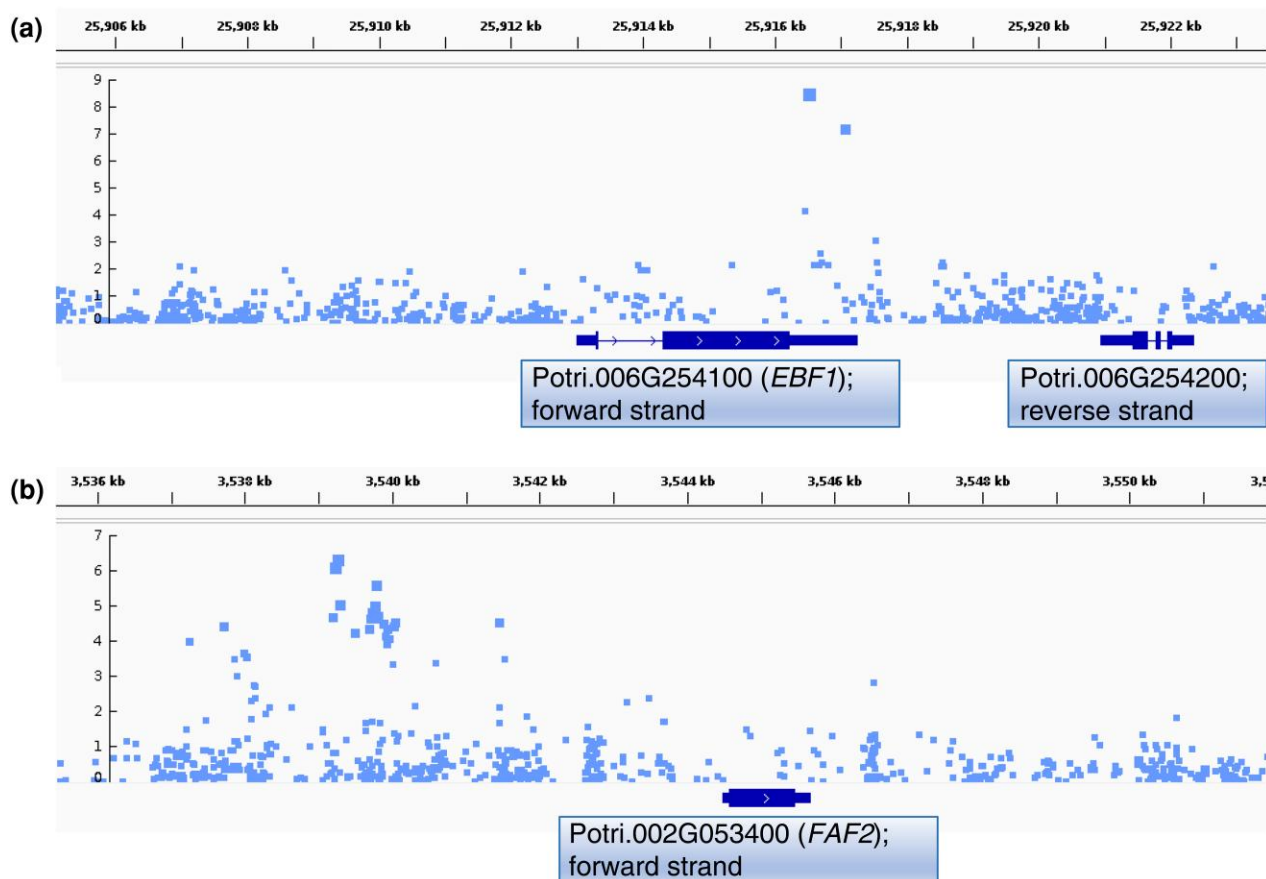

Fig. 8. Close-up view of Manhattan plots for example *in planta* regeneration traits aligned with genome annotation. Shown are zoomed-in portions of Manhattan plots aligned to the genome annotation for *P. trichocarpa* (v3.1). Introns, untranslated regions and exons are respectively visualized with increasing thickness of bars. Labels in rectangles 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, with a peak in the 3′ UTR of *EBF1*; b) Results on chromosome 2 for GEMMA of Box–Cox-transformed trait Callus/Shoot PC1, showing an association found significant via ART. These plots were made with Integrative Genomics Viewer (IGV) and boxes with Accession ID, gene name, and strand were manually added underneath. Examples of plots for additional loci can be found in Supplementary Fig. 14.

Table 1. Fourteen *in planta* regeneration GWAS candidates highlighted for discussion, and Arabidopsis homologs relevant to regeneration.

Candidate genes							Arabidopsis homologs			
Threshold	Trait	Method	Transf.	Dist. (bp)	QTL Pos.	Accession ID	Accession ID	Description	Score	Similarity
Bonf.	Callus Area wk. 2	GMMAT	Dichotomized trait	3	5'	Potri.006G276200	AT3G12660	FASCICLIN-like arabinogalactan protein 14 precursor (FLA14)	156	60.40%
FDR = 0.10	Shoot Area wk. 4	MTMC-SKAT	Untransformed	0	Exonic	Potri.019G101900	AT4G28250	Expansin B3	409	86.90%
Bonf.	Shoot PC2	GEMMA	Outliers removed, Box–Cox	0	Exonic	Potri.006G254100	AT2G25490	EIN3-BINDING F BOX PROTEIN 1 (EBF1)	812	80.70%
FDR = 0.10	Shoot PC1	MTMC-SKAT	Untransformed	0	Intragenic, nonexonic	Potri.003G194600	AT3G12250	TGACG MOTIF-BINDING FACTOR 6 (TGA6)	520	91.00%
FDR = 0.10	Shoot PC1	MTMC-SKAT	Untransformed	4,877	3'	Potri.015G041800	AT3G18165	Modifier of snc1,4 (MOS4)	363	84.70%
FDR = 0.10	Shoot PC1	MTMC-SKAT	Untransformed	0	Exonic	Potri.002G070600	AT1G21326	VQ motif-containing protein 3 (VQ3)	105	65.00%
FDR = 0.10	Shoot PC2	GEMMA	Outliers removed, thresholding, Box–Cox	0	Exonic	Potri.002G173300	AT2G46560	transducin family protein/WD-40 repeat family protein	2357	66.80%
Bonf.	Callus Area wk. 5	GEMMA	Box–Cox	3,947	3'	Potri.018G049600	AT5G35550	TRANSPARENT TESTA 2 (TT2)	196	69.90%
FDR = 0.10	Shoot PC1	MTMC-SKAT	Untransformed	0	Exonic	Potri.004G155400	AT1G75250	RADIALIS-LIKE SANT/MYB 3 (RSM3)	129	82.40%
FDR = 0.10	Callus 5w	GEMMA	Box–Cox	1,192	5'	Potri.010G105600	AT4G16110	ARABIDOPSIS RESPONSE REGULATOR 2 (ARR2)	382	75.70%
FDR = 0.10	Shoot PC2	GEMMA	Outliers removed, thresholding, Box–Cox	137	5'	Potri.011G031100	AT1G11530	C-TERMINAL CYSTEINE RESIDUE IS CHANGED TO A SERINE 1; thioredoxin	96	69.30%
ART-Bonf.	Callus, Shoot PC1	GEMMA	Box–Cox	5,222	5'	Potri.002G053400	AT1G03170	FANTASTIC FOUR 2 (FAF2)	103	54.60%
ART-Bonf.	Callus, Shoot PC1	GEMMA	RB-INV	5,222	5'	Potri.002G053400	AT1G03170	FANTASTIC FOUR 2 (FAF2)	103	54.60%
ART-Bonf.	Callus Area wk. 4	GEMMA	Box–Cox	5,749	5'	Potri.012G032900	AT4G27950	CYTOKININ RESPONSE FACTOR 4 (CRF4)	227	62.60%
ART-Bonf.	Shoot PC1	GEMMA	RB-INV	2,346	5'	Potri.012G070400	AT3G52960	PEROXIREDOXIN-II-E (PRXIIE)	105	78.90%

Relevant literature is discussed for each of these candidates (Discussion section). Distance (Dist.) of individual SNPs or SKAT/ART window centers from the transcription start site is shown for intergenic associations. Score and similarity percentage is shown for Smith–Waterman alignment of poplar candidate genes with Arabidopsis homologs, extracted from Phytozome (phytozome.jgi.doe.gov). Remaining candidate genes are summarized in Supplementary Tables 6–8.

empirical *P*-values. We speculate that the skewness of tissue growth traits studied here results from a combination of widespread recalcitrance to regeneration expressed as a threshold trait (often zero vs observable regeneration), and complex multiplicative and exponential patterns of growth.

Our image segmentation training set enabled a deep segmentation model that was used to automatically segment a total of 11,791 images. Although generation of the training samples was time-consuming, performing manual segmentation for all images would have been time-prohibitive. Image annotation using the GUI required approximately five minutes per image using IDEAS, for a total of approximately 21 hours with 249 images. Without a trained model for image segmentation, annotation of the whole dataset would have required nearly 1,000 h. This system or others that are functionally comparable (Russell *et al.* 2008; Dutta and Zisserman 2019) can be made more accessible and practical with innovations to reduce the number of clicks needed for image annotation by further semiautomation of annotation. While there have been many successes in the use of color thresholding methods that segment plants while requiring relatively little or no training data (Das *et al.* 2015; Yang *et al.* 2015; Al-Tamimi *et al.* 2016; Gage *et al.* 2018; Guo *et al.* 2018; Pham *et al.* 2019; Campbell *et al.* 2020; Seethepalli *et al.* 2020; Ogawa *et al.* 2021; Affortit *et al.* 2022), the highly variable colors observed in callus and shoot rendered these methods inappropriate for our study. Furthermore, in our previous work we reported a comparison of random forests (RF) to our trained PSPNet model and found RF to underperform in segmentation accuracy (Yuan *et al.* 2022).

As reported in our prior work (Yuan *et al.* 2022), 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 more heterogenous tissues had up to 67.40% IoU. Imperfections in defining segmentation boundaries, whether manually or using a trained model, leads to the introduction of statistical noise that will tend to reduce the computed heritability of traits. However, the use of PCA over batches of traits reduced the impact of this noise and yielded traits that tended to display greater heritability. Indeed, while we studied six PC traits and eight non-PC traits, four of the six traits with the greatest heritabilities were PCs (Supplementary Table 4).

Clearly, there is room for further advancement to make these phenomic workflows more efficient and user-friendly. Advances in the architectures of segmentation neural networks can contribute to improved accuracy in segmenting complex and heterogenous tissues of interest to biologists. Further innovation in the interfaces used to generate training sets can reduce the time cost needed for training these models. Once a training set is obtained, the training of neural networks can require a labor-intensive process of hyperparameter tuning. Recent developments such as Automated Machine Learning (AutoML) can be used to automate the hyperparameter tuning process and simplify training (Waring *et al.* 2020; Karmaker *et al.* 2022). Finally, the continued development of robust and efficient GWAS methods that avoid linear model assumptions can help to maximize the genetic insights gained from non-normal traits based on image segmentation.

Complementary GWAS approaches provide variety of insights

Transformation of traits to approximate normality is commonly employed for biological data during GWAS, as one approach to avoid violating linear model assumption of normality of residuals.

In our study, because traits were computed as the proportion of plant tissue labeled by segmentation masks as callus or shoot, and many genotypes did not produce either tissue at early timepoints or any timepoints, the resulting distributions featured 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 5) to 1,106 (for shoot area at week 2). Since spikes at discrete values such as zero can introduce discontinuity to distributions and thus lead to violations of parametric assumptions, genotypes with zero values were excluded from GEMMA tests for each trait. We used two complementary approaches to avoid continuity assumptions, provided by GMMAT (logistic regression) and MTMC-SKAT (efficient computation of empirical *P*-values), thus avoiding the exclusion of totally recalcitrant genotypes from analysis. Although we are unaware of a means to directly and quantitatively compare statistical power between set-based (SKAT, GEMMA-ART, GMMAT-ART) and single-variant (GEMMA, GMMAT) methods as they differ in their explanatory variables and the hypotheses being tested, we note that the majority of our associations were found with the set-based methods.

Single-SNP methods including GEMMA and GMMAT share the advantage of providing insights into the specific SNPs more likely to be causative with respect to the effect of a gene on a trait (Zhou and Stephens 2012; Chen *et al.* 2016). In most cases in our results, these appear to be regulatory SNPs in promoters, suggesting that variation in gene expression, rather than coding sequence, is the primary driver of trait variation. However, single-SNP methods may fail to detect significant associations, in part because they treat each SNP-trait relationship as an independent test and do not consider combined effects of nearby SNPs. In contrast, SKAT can identify additional associations by testing for combined effects of SNPs grouped into SNP sets, but only provides a single *P*-value for each SNP set (Wu *et al.* 2011; Ionita-Laza *et al.* 2013). Thus, our SKAT results do not make clear which SNPs in a given window are responsible for trait variation, as windows often span coding and regulatory regions. We therefore lack insight into whether SKAT-implicated candidates are responsible for trait differences due to variation in their expression or coding sequence. Moreover, even when a given window is entirely intergenic, we lack the ability for straightforward investigation of specific promoter motifs that may be implicated by SKAT due to the lack of single-SNP resolution.

We therefore employed a "best of both worlds" approach to find additional statistically significant associations from GEMMA and GMMAT results by considering combined effects of adjacent common SNPs, without losing the capacity to identify 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 associations both via 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 1 kb 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 the reach of GWAS, which may be of particular benefit to studies using segmentation-based traits that follow non-normal distributions and undergo transformations with a risk of overcorrection.

Our work relied on mapping to the Nisqually-1 reference genome. Thus, we are unsure of whether any given SNP is truly causal

or is only in LD with a causal SNP that was not observed, such as due to not being found in the Nisqually-1 reference genome. Greater certainty can be provided with the use of pan-genome SNP datasets, which are under rapid development for many plant species (Li *et al.* 2022).

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. Very few of the candidate genes we found were previously reported in two previous GWAS of related poplar traits, including several in a GWAS of adventitious shoot and root in hydroponic culture (Sun *et al.* 2019) and none in a GWAS of *in vitro* callus regeneration (Tuskan *et al.* 2018), likely due in part to the different phenotyping methods, traits studied, and dependence of GWAS associations on statistical approaches and experimental settings (Supplementary Tables 6–9; File 1, Supplementary Results and Discussion). Several of the most promising candidate genes we found, organized by biological function of orthologs in Arabidopsis, are briefly discussed below.

Regulation of cell wall adhesion

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

Many genes in the *FLA* family are differentially expressed during embryogenesis (Costa *et al.* 2019), but their regulation in the context of *in vitro* regeneration has received little study. In Arabidopsis, *FLA1* was found to be upregulated during incubation on callus induction media, while *FLA2* upregulation occurred upon transfer of explants to shoot induction media (Johnson *et al.* 2003). Loss-of-function of *FLA1* was reported to confer an ability for efficient *in vitro* shoot regeneration to the otherwise recalcitrant Col-0 ecotype, while contrarily leading to loss of efficient regeneration in the regenerable ecotype WS (Johnson *et al.* 2011).

We found an association of shoot development (week 4 shoot area and PC1) with a window of SNPs including a portion of the promoter and first exon of Potri.019G101900, related to Arabidopsis *EXPANSIN B3*. Expansins facilitate the process of pH-dependent cell wall loosening, with various expansins expressed during different stages of development. Perturbations of this gene superfamily have been studied in several plant species, including Arabidopsis, tomatoes, rice, soybean, and tobacco. Overexpression typically produces phenotypes of enhanced growth, such as increased size of plant cells and tissues, as well as reduced fruit firmness. Knockout or knockdown, in contrast, leads to reduced growth and increased firmness (Marowa *et al.* 2016). Expansins are believed to be key regulators of cell wall expansion downstream of auxin (Majda and Robert 2018).

Regulators of wound-responsive hormone signaling

Potri.006G254100 is 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) (Potuschak *et al.* 2003; Shen *et al.* 2016) and that this degradation is prevented when EIN3 and EIL1 are stabilized by ETHYLENE-INSENSITIVE 2 (EIN2) (An *et al.* 2010). Arabidopsis loss-of-function mutants of *EIN2* were used to supply cotyledon explant material for an *in vitro* regeneration assay, which revealed an approximate 4-fold reduction in shoot regeneration in the mutants. The same assay revealed a roughly threefold increase in shoot regeneration with loss-of-function of *HOOKLESS1* (*HLS1*), a gene encoding a putative *N*-acetyltransferase with a role downstream of *EIN3* in regulating a range of ethylene-regulated traits including apical hook development and *in vitro* regeneration (Chatfield and Raizada 2008). Also downstream of *EIN3* is positive and negative regulation of numerous genes across at least eight hormone pathways including ethylene signaling and the related jasmonate (JA) and salicylic acid (SA) signaling pathways among others, suggesting that *EIN3* represents a key modulator of hormone crosstalk (Chang *et al.* 2013). Although not found among our candidate genes, *WOUND-INDUCED DEDIFFERENTIATON* (*WIND*) transcription factors also regulate ethylene, JA and SA pathways (Iwase *et al.* 2021) and their perturbation is known to influence callus and shoot regeneration (Iwase *et al.* 2011, 2017). In support of a key role of *EIN3* in these pathways and in regeneration, we present at least seven candidate genes implicated as interacting directly or indirectly with *EIN3* or upstream regulators of EIN3 (Fig. 9). Our results, considered together with mutant studies in Arabidopsis, suggest that these candidates regulate regeneration by mediating crosstalk between ethylene, JA and SA signaling pathways.

Our GWAS results suggest a central role for SA and genes involved in SA signaling. *NPR1* is a regulator of SA signaling via a mechanism that depends on several genes with homologs implicated by QTLs in our GWAS (Fig. 9). Candidate gene Potri.003G194600 encodes a homolog of TGACGT motif transcription factor *TGA6*. *TGA6* and other members of the TGA family have been reported to interact with *NPR1* (Zhang *et al.* 1999; Subramaniam *et al.* 2001; Fan and Dong 2002; Rochon *et al.* 2006; Boyle *et al.* 2009; Hussain *et al.* 2018) to form a histone acetyltransferase complex responsible for SA-associated epigenetic reprogramming (Jin *et al.* 2018). Simultaneous knockout of functionally redundant *TGA* genes (Zhang, Tessaro, *et al.* 2003) or of *NPR1* (Cao *et al.* 1997) confers a loss of SA signaling, including SA-mediated pathogen resistance. Moreover, a dominant mutation of the putative upstream regulator *SUPPRESSOR OF NPR1, CONSTITUTIVE 1* (*SNC1*) confers constitutive SA signaling, dwarf morphology, and enhanced pathogen resistance (Zhang, Goritschnig, *et al.* 2003). This phenotype is reversed by loss-of-function of *MODIFIER OF SNC1,4* (*MOS4*), a homolog of our candidate gene Potri.015G041800 (Palma *et al.* 2007).

Additional regulation of *EIN3* is believed to exist via phosphorylation of EIN3 by the SA-regulated MAP KINASE 3 (MPK3) and MAP KINASE 6 (MPK6) (Yoo *et al.* 2008). MPK3 and MPK6 are also responsible for phosphorylation of many substrates in a VQ MOTIF PROTEIN family related to our candidate gene Potri.002G070600. Although the VQ proteins have not been studied in the context of *in vitro* regeneration to our knowledge, many members are believed to function downstream of pathogen-associated molecular patterns (PAMPs) and upstream of pathogen defense genes (Pecher *et al.* 2014).

Our GWAS results also suggest a central role for anthocyanin and related genes. The SA and JA 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 (Baudry *et al.* 2004; Qi *et al.* 2011). We report two candidate genes
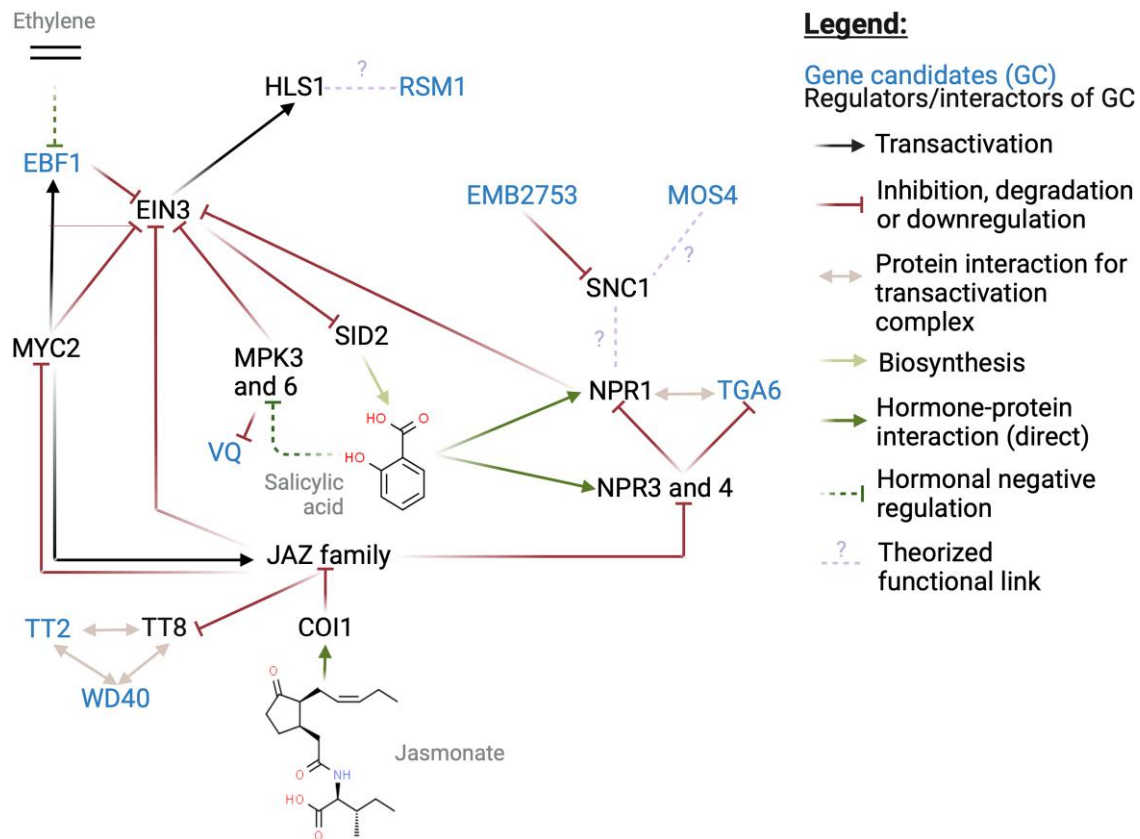
Fig. 9. Proposed model integrating roles of *in planta* regeneration candidate genes. Interactions involving Arabidopsis homologs of seven candidate genes (blue) and associated regulators were identified by literature review, providing an understanding of the broader context of hormone crosstalk between ethylene, JA, and SA pathways as they relate to regeneration. Node placement was assisted by the Force Atlas 2 algorithm as implemented in Gephi. Chemical structures for hormones were retrieved from ChemSpider (chemspider.com). Jasmonates are represented by (+)-7-Iso-Jasmonyl-L-isoleucine. This figure was produced using BioRender (biorender.com). Details for each of these candidate genes (blue) can be found in Table 1 and Discussion section. Standard acronyms and abbreviations can be found on The Arabidopsis Information Resource (Berardini *et al.* 2015) and are listed in Supplementary Table 9. Evidence for interactions is summarized in Supplementary Table 10 (Zhang *et al.* 1999; Subramaniam *et al.* 2001; Wildermuth *et al.* 2001; Fan and Dong 2002; Potuschak *et al.* 2003; Zhang, Goritschnig, *et al.* 2003; Baudry *et al.* 2004; Zimmermann *et al.* 2004; Rochon *et al.* 2006; Zhang *et al.* 2006; Chini *et al.* 2007; Palma *et al.* 2007; Thines *et al.* 2007; Hamaguchi *et al.* 2008; Yoo *et al.* 2008; Boyle *et al.* 2009; Chen *et al.* 2009; An *et al.* 2010; Qi *et al.* 2011; Zhu *et al.* 2011; Fu *et al.* 2012; Chang *et al.* 2013; Shi *et al.* 2013; Pecher *et al.* 2014; Song *et al.* 2014; Zhang *et al.* 2014; Kuai *et al.* 2015; Xu *et al.* 2015; Liu *et al.* 2016; Shen *et al.* 2016; An *et al.* 2017; Ding *et al.* 2018; Hussain *et al.* 2018; Jin *et al.* 2018; Huang *et al.* 2020).

homologous to MBW components, Potri.002G173300 (encoding a WD-40 repeat family protein) and Potri.018G049600 (homolog of *TRANSPARENT TESTA 2*). Although we are unaware of these genes being studied in the context of *in vitro* regeneration, MBWs regulate several steps of anthocyanin biosynthesis downstream of naringenin chalcone, which is produced by CHALCONE SYNTHASE (CHS) (Yan *et al.* 2021). 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 on 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 candidate gene Potri.004G155400) has been proposed due to phenotypic similarities between *hls1* and *RSM1*-overexpressing Arabidopsis. Light-deprived seedlings of both mutant lines presented various degrees of reduced hypocotyl length, defective hook formation, and defective gravitropism (Hamaguchi *et al.* 2008). However, whereas *HLS1* knockout is known to confer enhanced shoot regeneration in Arabidopsis (Chatfield and Raizada 2008), the effects of *RSM1* or *RSM* family overexpression or knockout on shoot regeneration have not yet been reported to our knowledge.

Regulators of cytokinin signaling

Several candidate genes from GWAS appear to affect cytokinin signaling. Potri.010G105600 is a homolog of *ARABIDOPSIS RESPONSE REGULATOR 2* (*ARR2*), which functions shortly downstream of cytokinin. B-type *ARR* genes 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*) (Nagle *et al.* 2018; Xie *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 *FAF2*, homolog of candidate gene Potri.002G053400) leads to arrest of shoot apical 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).

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 (H. Zhang *et al.* 2018).

Conclusions

We report a GWAS of *in planta* regeneration in *P. trichocarpa* using a computer vision system that we customized for determining regeneration phenotypes. We also utilized four complementary statistical methods for association mapping. These analyses revealed over 200 candidate genes, strongly implicating regulators of cell adhesion and stress signaling. However, each of these associations is based on statistical inference in an observational study, and we therefore cannot be certain of the causal genes without functional validation. Further research using gene-editing or other transgenic technologies, while using carefully selected hosts and candidate gene alleles expressed in an appropriate cell-specific manner, could be used to validate these associations. The candidate genes could also be used as reagents for biotechnology research aimed at enhancement of regeneration in poplar and other plants. While canonical regulators of *in vitro* regeneration tend to be involved in auxin and cytokinin signaling pathways, our results suggest that stress pathways downstream of ethylene, SA, and jasmonates may be of greatest importance to the mode of *in planta* regeneration that we studied in *P. trichocarpa*. These pathways have received relatively little attention in studies where developmental regulator genes are used to promote regeneration and would appear to be promising avenues to pursue. Furthermore, at least seven top candidates are likely closely linked members of a genetic regulatory network, separated from one another by no more than four degrees of direct interactions. This, considered along with the complex nature of *in vitro* regeneration traits, suggests that emerging multilocus methods and epistasis tests may provide significantly greater insights into the polygenic control of these traits. The scale and precision of this study would not have been possible without computer vision, and our success in identifying established regulators of regeneration demonstrates that semantic segmentation using deep convolutional neural networks shows promise in supporting characterization of regeneration and other complex developmental plant traits.

Data availability

The MTMC-SKAT R package is available on GitHub (https://github.com/naglemi/mtmcskat), as is other R code used for this study, including phenotype data parsing, association mapping, and downstream interrogation of results (https://github.com/naglemi/inplantaGWAS). Our SNP and image datasets are also available for public download (Nagle et al. 2023).

Supplemental material is available at G3 online.

Author contributions

SHS, LF, YJ, and WM designed and directed the overall study, and obtained funding for its execution; CM and EP designed and/or executed the phenotypic analyses; MFN, YJ, JY, and DK created, adapted, and executed the machine vision, computation, and data analysis pipelines; MFN investigated candidate genes. ANR assisted with inspecting results in IGV; SJ, J-GC, KF, TBY, GAT, and WM supported acquisition and preparation of the GWAS population and SNP dataset; MFN wrote the manuscript with editing from SHS, and all others contributed further edits and revisions.

Conflicts of interest

The author(s) declare no conflict of interest.

Literature cited

Adams J, Qiu Y, Xu Y, Schnable JC. 2020. Plant segmentation by supervised machine learning methods. Plant Phenome J. 3(1): e20001. doi:10.1002/ppj2.20001.

Affortit P, Effa-Effa B, Ndoye MS, Moukouanga D, Luchaire N, Cabrera-Bosquet L, Perálvarez M, Pilloni R, Welcker C, Champion A, *et al.* 2022. Physiological and genetic control of transpiration efficiency in African rice, *Oryza glaberrima* Steud. J Exp Bot. 73(15): 5279–5293. doi:10.1093/jxb/erac156.

Aich S, Stavness I. 2017. Leaf counting with deep convolutional and deconvolutional networks. In: IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy. p. 2080–2089. doi:10.1109/ICCVW.2017.244.

Al-Tamimi N, Brien C, Oakey H, Berger B, Saade S, Ho YS, Schmöckel SM, Tester M, Negrão S. 2016. Salinity tolerance loci revealed in rice using high-throughput non-invasive phenotyping. Nat Commun. 7(1):13342. doi:10.1038/ncomms13342.

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. Plant Cell. 28(7):1510–1520. doi:10.1105/tpc.16.00196.

An C, Li L, Zhai Q, You Y, Deng L, Wu F, Chen R, Jiang H, Wang H, Chen Q, *et al.* 2017. Mediator subunit MED25 links the jasmonate receptor to transcriptionally active chromatin. Proc Natl Acad Sci USA. 114(42):E8930–E8939. doi:10.1073/pnas.1710885114.

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 degradation of EIN3 binding F-box 1 and 2 that requires EIN2 in Arabidopsis. Plant Cell. 22(7):2384–2401. doi:10.1105/tpc.110.076588.

Baudry A, Heim MA, Dubreucq B, Caboche M, Weisshaar B, Lepiniec L. 2004. TT2, TT8, and TTG1 synergistically specify the expression of BANYULS and proanthocyanidin biosynthesis in *Arabidopsis thaliana*. Plant J. 39(3):366–380. 10.1111/j.1365-313X.2004.02138.x.

Bdeir R, Muchero W, Yordanov Y, Tuskan GA, Busov V, Gailing O. 2019. Genome-wide association studies of bark texture in *Populus trichocarpa*. Tree Genet Genomes. 15(1):14. doi:10.1007/s11295-019-1320-2.

Beasley TM, Erickson S, Allison DB. 2009. Rank-based inverse normal transformations are increasingly used, but are they merited? Behav Genet. 39(5):580–595. doi:10.1007/s10519-009-9281-0.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 57(1):289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, Huala E. 2015. The *Arabidopsis* information resource: making and mining the "gold standard" annotated reference plant genome. Genesis. 53(8):474–485. doi:10.1002/dvg.22877.

Bomzan DP, Shilpashree HB, Nagegowda DA. 2022. Agrobacterium-mediated in planta transformation in periwinkle. In: Courdavault V, Besseau S, editors. Catharanthus roseus: Methods and Protocols. New York (NY): Springer US. p. 301–315. doi:10.1007/978-1-0716-2349-7_22.

Box GEP, Cox DR. 1964. An analysis of transformations. J R Stat Soc Ser B Methodol. 26(2):211–243. doi:10.1111/j.2517-6161.1964.tb00553.x.

Boyle P, Le Su E, Rochon A, Shearer HL, Murmu J, Chu JY, Fobert PR, Després C. 2009. The BTB/POZ domain of the Arabidopsis disease resistance protein NPR1 interacts with the repression domain of TGA2 to negate its function. Plant Cell. 21(11):3700–3713. doi:10.1105/tpc.109.069971.

Brown DE, Rashotte AM, Murphy AS, Normanly J, Tague BW, Peer WA, Taiz L, Muday GK. 2001. Flavonoids act as negative regulators of auxin transport in vivo in Arabidopsis. Plant Physiol. 126(2):524–535. doi:10.1104/pp.126.2.524.

Campbell MT, Grondin A, Walia H, Morota G. 2020. Leveraging genome-enabled growth models to study shoot growth responses to water deficit in rice. J Exp Bot. 71(18):5669–5679. doi:10.1093/jxb/eraa280.

Cao H, Glazebrook J, Clarke JD, Volko S, Dong X. 1997. The Arabidopsis NPR1 gene that controls systemic acquired resistance encodes a novel protein containing ankyrin repeats. Cell. 88(1):57–63. doi:10.1016/s0092-8674(00)81858-9.

Carlier A, Dandrifosse S, Dumont B, Mercatoris B. 2022. Wheat ear segmentation based on a multisensor system and superpixel classification. Plant Phenomics. 2022. doi:10.34133/2022/9841985.

Chang KN, Zhong S, Weirauch MT, Hon G, Pelizzola M, Li H, Huang SC, Schmitz RJ, Urich MA, Kuo D, *et al.* 2013. Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in Arabidopsis. eLife. 2:e00675. doi:10.7554/eLife.00675.

Chatfield SP, Raizada MN. 2008. Ethylene and shoot regeneration: hookless1 modulates de novo shoot organogenesis in *Arabidopsis thaliana*. Plant Cell Rep. 27(4):655–666. doi:10.1007/s00299-007-0496-3.

Chen H, Wang C, Conomos Matthew P, Stilp AM, Li Z, Sofer T, Szpiro AA, Chen W, Brehm JM, Celedón JC, *et al.* 2016. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. Am J Hum Genet. 98(4):653–666. doi:10.1016/j.ajhg.2016.02.012.

Chen H, Xue L, Chintamanani S, Germain H, Lin H, Cui H, Cai R, Zuo J, Tang X, Li X, *et al.* 2009. ETHYLENE INSENSITIVE3 and ETHYLENE INSENSITIVE3-LIKE1 repress SALICYLIC ACID INDUCTION DEFICIENT2 expression to negatively regulate plant innate immunity in Arabidopsis. Plant Cell. 21(8):2527–2540. doi:10.1105/tpc.108.065193.

Chhetri HB, Furches A, Macaya-Sanz D, Walker AR, Kainer D, Jones P, Harman-Ware AE, Tschaplinski TJ, Jacobson D, Tuskan GA, *et al.* 2020. Genome-wide association study of wood anatomical and morphological traits in *Populus trichocarpa*. Front Plant Sci. 11. doi:10.3389/fpls.2020.545748.

Chini A, Fonseca S, Fernández G, Adie B, Chico JM, Lorenzo O, García-Casado G, López-Vidriero I, Lozano FM, Ponce MR, *et al.* 2007. The JAZ family of repressors is the missing link in jasmonate signalling. Nature. 448(7154):666–671. doi:10.1038/nature06006.

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 Reprod. 32(4):353–370. doi:10.1007/s00497-019-00376-7.

Dai X, Liu Z, Qiao M, Li J, Li S, Xiang F. 2017. ARR12 promotes de novo shoot regeneration in *Arabidopsis thaliana* via activation of WUSCHEL expression. J Integr Plant Biol. 59(10):747–758. doi:10.1111/jipb.12567.

Das A, Schneider H, Burridge J, Ascanio AKM, Wojciechowski T, Topp CN, Lynch JP, Weitz JS, Bucksch A. 2015. Digital imaging of root traits (DIRT): a high-throughput computing and collaboration platform for field-based root phenomics. Plant Methods. 11(1):51. doi:10.1186/s13007-015-0093-3.

Deng W, Luo K, Li Z, Yang Y. 2009. A novel method for induction of plant regeneration via somatic embryogenesis. Plant Sci. 177(1):43–48. doi:10.1016/j.plantsci.2009.03.009.

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 Res. 50(D1):D1442–D1447. doi:10.1093/nar/gkab1014.

Ding Y, Sun T, Ao K, Peng Y, Yaxi Z, Li X, Yuelin Z. 2018. Opposite roles of salicylic acid receptors NPR1 and NPR3/NPR4 in transcriptional regulation of plant immunity. Cell. 173(6):1454–1467.e15. doi:10.1016/j.cell.2018.03.044.

Dobrescu A, Valerio Giuffrida M, Tsaftaris SA. 2017. Leveraging multiple datasets for deep leaf counting. In: 2017 IEEE International Conference on Computer Vision Workshop (ICCVW), Venice, Italy. p. 2072–2079. doi:10.1109/ICCVW.2017.243.

Dunn OJ. 1959. Confidence intervals for the means of dependent, normally distributed variables. J Am Stat Assoc. 54(287):613–621. doi:10.2307/2282541.

Dutta A, Zisserman A. 2019. The VIA annotation software for images, audio, and video. In: Proceedings of the 27th ACM International Conference on Multimedia (MM '19). New York (NY): Association for Computing Machinery. p. 2276–2279. doi:10.1145/3343031.3350535.

Fan W, Dong X. 2002. In vivo interaction between NPR1 and transcription factor TGA2 leads to salicylic acid-mediated gene activation in Arabidopsis. Plant Cell. 14(6):1377–1389. doi:10.1105/tpc.001628.

Fox J, Weisberg S. 2018. An R Companion to Applied Regression. 3rd ed. Thousand Oaks (CA): SAGE Publications.

Fu ZQ, Yan S, Saleh A, Wang W, Ruble J, Oka N, Mohan R, Spoel SH, Tada Y, Zheng N, et al. 2012. NPR3 and NPR4 are receptors for the immune signal salicylic acid in plants. Nature. 486(7402): 228–232. doi:10.1038/nature11162.

Gage JL, White MR, Edwards JW, Kaeppler S, de Leon N. 2018. Selection signatures underlying dramatic male inflorescence transformation during modern hybrid maize breeding. Genetics. 210(3): 1125–1138. doi:10.1534/genetics.118.301487.

Good P. 2013. Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. New York (NY): Springer. doi:10.1007/978-1-4757-2346-5.

Guo Z, Yang W, Chang Y, Ma X, Tu H, Xiong F, Jiang N, Feng H, Huang C, Yang P, et al. 2018. Genome-wide association studies of image traits reveal genetic architecture of drought resistance in rice. Mol Plant. 11(6):789–805. doi:10.1016/j.molp.2018.03.018.

Guo S, Zhou G, Wang J, Lu X, Zhao H, Zhang M, Guo X, Zhang Y. 2022. High-throughput phenotyping accelerates the dissection of the phenotypic variation and genetic architecture of shank vascular bundles in maize (Zea mays L.). Plants. 11(10):1339. doi:10.3390/plants11101339.

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. Biosci Biotechnol Biochem. 72(10):2687–2696. doi:10.1271/bbb.80348.

Huang P, Dong Z, Guo P, Zhang X, Qiu Y, Li B, Wang Y, Guo H. 2020. Salicylic acid suppresses apical hook formation via NPR1-mediated repression of EIN3 and EIL1 in Arabidopsis. Plant Cell. 32(3):612–629. doi:10.1105/tpc.19.00658.

Hussain RMF, Sheikh AH, Haider I, Quareshy M, Linthorst HJM. 2018. Arabidopsis WRKY50 and TGA transcription factors synergistically activate expression of PR1. Front Plant Sci. 9:930. doi:10.3389/fpls.2018.00930.

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. Am J Hum Genet. 92(6):841–853. doi:10.1016/j.ajhg.2013.04.015.

Iwase A, Harashima H, Ikeuchi M, Rymen B, Ohnuma M, Komaki S, Morohashi K, Kurata T, Nakata M, Ohme-Takagi M, et al. 2017. WIND1 promotes shoot regeneration through transcriptional activation of ENHANCER OF SHOOT REGENERATION1 in Arabidopsis. Plant Cell. 29(1):54–69. doi:10.1105/tpc.16.00623.

Iwase A, Kondo Y, Laohavisit A, Takebayashi A, Ikeuchi M, Matsuoka K, Asahina M, Mitsuda N, Shirasu K, Fukuda H, et al. 2021. WIND transcription factors orchestrate wound-induced callus formation, vascular reconnection and defense response in Arabidopsis. New Phytol. 232(2):734–752. doi:10.1111/nph.17594.

Iwase A, Mitsuda N, Koyama T, Hiratsu K, Kojima M, Arai T, Inoue Y, Seki M, Sakakibara H, Sugimoto K, et al. 2011. The AP2/ERF transcription factor WIND1 controls cell dedifferentiation in Arabidopsis. Curr Biol. 21(6):508–514. doi:10.1016/j.cub.2011.02.020.

Jaganathan D, Ramasamy K, Sellamuthu G, Jayabalan S, Venkataraman G. 2018. CRISPR for crop improvement: an update review. Front Plant Sci. 9:985. doi:10.3389/fpls.2018.00985.

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 Res. 46(22):11712–11725. doi:10.1093/nar/gky847.

Johnson KL, Jones BJ, Bacic A, Schultz CJ. 2003. The fasciclin-like arabinogalactan proteins of Arabidopsis. A multigene family of putative cell adhesion molecules. Plant Physiol. 133(4):1911–1925. doi:10.1104/pp.103.031237.

Johnson KL, Kibble NAJ, Bacic A, Schultz CJ. 2011. A fasciclin-like arabinogalactan-protein (FLA) mutant of Arabidopsis thaliana, fla1, shows defects in shoot regeneration. PLoS One. 6(9): e25154. doi:10.1371/journal.pone.0025154.

Kamath R, Balachandra M, Vardhan A, Maheshwari U. 2022. Classification of paddy crop and weeds using semantic segmentation. Cogent Eng. 9(1):2018791. doi:10.1080/23311916.2021.2018791.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. 2008. Efficient control of population structure in model organism association mapping. Genetics. 178(3):1709–1723. doi:10.1534/genetics.107.080101.

Karmaker ("Santu") SK, Hassan M, Smith MJ, Xu L, Zhai C, Veeramachaneni K. 2022. AutoML to date and beyond: challenges and opportunities. ACM Comput Surv. 54(8):1–36. doi:10.1145/3470918.

Kesiraju K, Tyagi S, Mukherjee S, Rai R, Singh NK, Sreevathsa R, Dash PK. 2021. An apical meristem-targeted in planta transformation method for the development of transgenics in flax (Linum usitatissimum): optimization and validation. Front Plant Sci. 11:562056. doi:10.3389/fpls.2020.562056.

Kuai X, MacLeod BJ, Després C. 2015. Integrating data on the Arabidopsis NPR1/NPR3/NPR4 salicylic acid receptors; a differentiating argument. Front Plant Sci. 6:235. doi:10.3389/fpls.2015.00235.

Kyritsis KA, Wang B, Sullivan J, Lyne R, Micklem G. 2019. InterMineR: an R package for InterMine databases. Bioinformatics. 35(17): 3206–3207. doi:10.1093/bioinformatics/btz039.

Lardon R, Geelen D. 2020. Natural variation in plant pluripotency and regeneration. Plants. 9(10):1261. doi:10.3390/plants9101261.

Lee K, Wang K. 2023. Strategies for genotype-flexible plant transformation. Curr Opin Biotechnol. 79:102848. doi:10.1016/j.copbio.2022.102848.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 25(14):1754–1760. doi:10.1093/bioinformatics/btp324.

Li W, Liu J, Zhang H, Liu Z, Wang Y, Xing L, He Q, Du H. 2022. Plant pangenomics: recent advances, new challenges, and roads ahead. J Genet Genomics. 49(9):833–846. doi:10.1016/j.jgg.2022.06.004.

Lian Z, Nguyen CD, Liu L, Wang G, Chen J, Wang S, Yi G, Wilson S, Ozias-Akins P, Gong H, et al. 2022. Application of developmental regulators to improve in planta or in vitro transformation in plants. Plant Biotechnol J. 20(8):1622–1635. doi:10.1111/pbi.13837.

Liu B, Zhang J, Yang Z, Matsui A, Seki M, Li S, Yan X, Kohnen MV, Gu L, Prasad K, et al. 2018. PtWOX11 acts as master regulator conducting the expression of key transcription factors to induce de novo shoot organogenesis in poplar. Plant Mol Biol. 98(4–5):389–406. doi:10.1007/s11103-018-0786-x.

Liu L, Sonbol F-M, Huot B, Gu Y, Withers J, Mwimba M, Yao J, He SY, Dong X. 2016. Salicylic acid receptors activate jasmonic acid signalling through a non-canonical pathway to promote effector-triggered immunity. Nat Commun. 7(1):13099. doi:10.1038/ncomms13099.

López-Cortegano E, Caballero A. 2019. Inferring the nature of missing heritability in human traits using data from the GWAS catalog. Genetics. 212(3):891–904. doi:10.1534/genetics.119.302077.

Ma X, Deng X, Qi L, Jiang Y, Li H, Wang Y, Xing X. 2019. Fully convolutional network for rice seedling and weed image segmentation

at the seedling stage in paddy fields. PLoS One. 14(4):e0215676. doi:10.1371/journal.pone.0215676.

Maher MF, Nasti RA, Vollbrecht M, Starker CG, Clark MD, Voytas DF. 2020. Plant gene editing through de novo induction of meristems. Nat Biotechnol. 38(1):84–89. doi:10.1038/s41587-019-0337-2.

Majda M, Robert S. 2018. The role of auxin in cell wall expansion. Int J Mol Sci. 19(4):951. doi:10.3390/ijms19040951.

Marowa P, Ding A, Kong Y. 2016. Expansins: roles in plant growth and potential applications in crop improvement. Plant Cell Rep. 35(5):949–965. doi:10.1007/s00299-016-1948-4.

McCaw ZR, Lane JM, Saxena R, Redline S, Lin X. 2020. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. Biometrics. 76(4):1262–1272. doi:10.1111/biom.13214.

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. Proc Natl Acad Sci USA. 115(45):11573–11578. doi:10.1073/pnas.1804428115.

Nagle M, Déjardin A, Pilate G, Strauss SH. 2018. Opportunities for innovation in genetic transformation of forest trees. Front Plant Sci. 9:1443. doi:10.3389/fpls.2018.01443.

Nagle MF, Yuan J, Kaur D, Ma C, Peremyslova E, Jiang Y, Niño de Rivera A, Jawdy S, Chen J-G, Feng K, *et al.* 2023. Expanded genetic variation (SNP) and phenomics (image based) dataset for Populus trichocarpa. Center for Bioenergy Innovation, Oak Ridge National Laboratory. doi:10.25983/2263365.

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 Environ. 36(1):68–86. doi:10.1111/j.1365-3040.2012.02554.x.

National Academies of Sciences, Engineering, and Medicine. 2016. Genetically Engineered Crops: Experiences and Prospects. Washington (DC): The National Academies Press. doi:10.17226/23395.

Nguyen THN, Winkelmann T, Debener T. 2020. Genetic analysis of callus formation in a diversity panel of 96 rose genotypes. Plant Cell Tissue Organ Cult. 142(3):505–517. doi:10.1007/s11240-020-01875-6.

Ogawa D, Sakamoto T, Tsunematsu H, Kanno N, Nonoue Y, Yonemaru J. 2021. Haplotype analysis from unmanned aerial vehicle imagery of rice MAGIC population for the trait dissection of biomass and plant architecture. J Exp Bot. 72(7):2371–2382. doi:10.1093/jxb/eraa605.

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 Dev. 21(12):1484–1493. doi:10.1101/gad.1559607.

Pan C, Li G, Malzahn AA, Cheng Y, Leyson B, Sretenovic S, Gurel F, Coleman GD, Qi Y. 2022. Boosting plant genome editing with a versatile CRISPR-Combo system. Nat Plants. 8(5):513–525. doi:10.1038/s41477-022-01151-9.

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. New Phytol. 203(2):592–606. doi:10.1111/nph.12817.

Pham A-T, Maurer A, Pillen K, Brien C, Dowling K, Berger B, Eglinton JK, March TJ. 2019. Genome-wide association of barley plant growth under drought stress using a nested association mapping population. BMC Plant Biol. 19(1):134. doi:10.1186/s12870-019-1723-0.

Potuschak T, Lechner E, Parmentier Y, Yanagisawa S, Grava S, Koncz C, Genschik P. 2003. EIN3-dependent regulation of plant ethylene hormone signaling by two arabidopsis F box proteins: EBF1 and EBF2. Cell. 115(6):679–689. doi:10.1016/s0092-8674(03)00968-1.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 38(8):904–909. doi:10.1038/ng1847.

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*. Plant Cell. 23(5):1795–1814. doi:10.1105/tpc.111.083261.

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. Plant J. 85(1):134–147. doi:10.1111/tpj.13097.

Raj A, Stephens M, Pritchard JK. 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics. 197(2):573–589. doi:10.1534/genetics.114.164350.

R Core Team. 2022. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol. 29(1):24–26. doi:10.1038/nbt.1754.

Rochon A, Boyle P, Wignes T, Fobert PR, Després C. 2006. The coactivator function of Arabidopsis NPR1 requires the core of its BTB/POZ domain and the oxidation of C-terminal cysteines. Plant Cell. 18(12):3670–3685. doi:10.1105/tpc.106.046953.

Russell BC, Torralba A, Murphy KP, Freeman WT. 2008. Labelme: a database and web-based tool for image annotation. Int J Comput Vis. 77(1–3):157–173. doi:10.1007/s11263-007-0090-8.

Schielzeth H, Dingemanse NJ, Nakagawa S, Westneat DF, Allegue H, Teplitsky C, Réale D, Dochtermann NA, Garamszegi LZ, Araya-Ajoy YG. 2020. Robustness of linear mixed-effects models to violations of distributional assumptions. Methods Ecol Evol. 11(9):1141–1152. doi:10.1111/2041-210X.13434.

Seethepalli A, Guo H, Liu X, Griffiths M, Almtarfi H, Li Z, Liu S, Zare A, Fritschi FB, Blancaflor EB, *et al.* 2020. RhizoVision crown: an integrated hardware and software platform for root crown phenotyping. Plant Phenomics. 2020. doi:10.34133/2020/3074916.

She Y, Wang X, Gui T, Cai J. 2019. Lawn plant identification and segmentation based on least squares support vector machine and multifeature fusion. J Electron Imaging. 28((02|2)):023034. doi:10.1117/1.JEI.28.2.023034.

Shen X, Li Y, Pan Y, Zhong S. 2016. Activation of HLS1 by mechanical stress via ethylene-stabilized EIN3 is crucial for seedling soil emergence. Front Plant Sci. 7:1571. doi:10.3389/fpls.2016.01571.

Shi Z, Maximova S, Liu Y, Verica J, Guiltinan MJ. 2013. The salicylic acid receptor NPR3 is a negative regulator of the transcriptional defense response during early flower development in Arabidopsis. Mol Plant. 6(3):802–816. doi:10.1093/mp/sss091.

Song S, Huang H, Gao H, Wang J, Wu D, Liu X, Yang S, Zhai Q, Li C, Qi T, *et al.* 2014. Interaction between MYC2 and ETHYLENE INSENSITIVE3 modulates antagonism between jasmonate and ethylene signaling in Arabidopsis. Plant Cell. 26(1):263–279. doi:10.1105/tpc.113.120394.

Subramaniam R, Desveaux D, Spickler C, Michnick SW, Brisson N. 2001. Direct visualization of protein interactions in plant cells. Nat Biotechnol. 19(8):769–772. doi:10.1038/90831.

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. Int J Mol Sci. 20(24): 6114. doi:10.3390/ijms20246114.

Tange O. 2022. GNU Parallel 20220522 ('NATO'). https://doi.org/10.5281/zenodo.6570228.

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba: fast processing of NGS alignment formats. Bioinformatics. 31(12): 2032–2034. doi:10.1093/bioinformatics/btv098.

Thines B, Katsir L, Melotto M, Niu Y, Mandaokar A, Liu G, Nomura K, He SY, Howe GA, Browse J. 2007. JAZ repressor proteins are targets of the SCFCOI1 complex during jasmonate signalling. Nature. 448(7154):661–665. doi:10.1038/nature05960.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 14(2):178–192. doi:10.1093/bib/bbs017.

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. Comput Sci Eng. 16(5): 62–74. doi:10.1109/MCSE.2014.80.

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(5793):1596–1604. doi:10.1126/science.1128691.

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(8):8. doi:10.1371/journal.pone.0202519.

Van der Auwera GA, O'Connor BD. 2020. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. Sebastopol (CA): O'Reilly Media, Inc.

Vsevolozhskaya OA, Hu F, Zaykin DV. 2019. Detecting weak signals by combining small p-values in genetic association studies. Front Genet. 10. doi:10.3389/fgene.2019.01051.

Wahl V, Brand LH, Guo Y-L, Schmid M. 2010. The FANTASTIC FOUR proteins influence shoot meristem size in Arabidopsis thaliana. BMC Plant Biol. 10(1):285. doi:10.1186/1471-2229-10-285.

Waring J, Lindvall C, Umeton R. 2020. Automated machine learning: review of the state-of-the-art and opportunities for healthcare. Artif Intell Med. 104:101822. doi:10.1016/j.artmed.2020.101822.

Weighill D, Tschaplinski TJ, Tuskan GA, Jacobson D. 2019. Data integration in poplar: 'omics layers and integration strategies. Front Genet. 10:874. doi:10.3389/fgene.2019.00874.

Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. New York (NY): Springer-Verlag. https://ggplot2.tidyverse.org.

Wildermuth MC, Dewdney J, Wu G, Ausubel FM. 2001. Isochorismate synthase is required to synthesize salicylic acid for plant defence. Nature. 414(6863):562–565. doi:10.1038/35107108.

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. Am J Hum Genet. 89(1):82–93. doi:10.1016/j.ajhg.2011.05.029.

Xiao Q, Bai X, Zhang C, He Y. 2021. Advanced high-throughput plant phenotyping techniques for genome-wide association studies: a review. J Adv Res. 35:215–230. doi:10.1016/j.jare.2021.05.002.

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. Nat Commun. 9(1):1604. doi:10.1038/s41467-018-03921-6.

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. Plant Cell. 27(5):1547–1562. doi:10.1105/tpc.15.00173.

Xu Z, York LM, Seethepalli A, Bucciarelli B, Cheng H, Samac DA. 2022. Objective phenotyping of root system architecture using image augmentation and machine learning in alfalfa (Medicago sativa L.). Plant Phenomics. 2022. doi:10.34133/2022/9879610.

Yan H, Pei X, Zhang H, Li X, Zhang X, Zhao M, Chiang VL, Sederoff RR, Zhao X. 2021. MYB-mediated regulation of anthocyanin biosynthesis. Int J Mol Sci. 22(6):3103. doi:10.3390/ijms22063103.

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. J Exp Bot. 66(18):5605–5615. doi:10.1093/jxb/erv100.

Yasrab R, Atkinson JA, Wells DM, French AP, Pridmore TP, Pound MP. 2019. RootNav 2.0: deep learning for automatic navigation of complex plant root architectures. GigaScience. 8(11):giz123. doi:10.1093/gigascience/giz123.

Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, Yuan X, Zhu M, Zhao S, Liu X, et al. 2021. rMVP: a memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. Genomics Proteomics Bioinformatics. 19(4):619–628. doi:10.1016/j.gpb.2020.10.007.

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(7180):789–795. doi:10.1038/nature06543.

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. doi:10.34133/2022/9893639.

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 Physiol. 40(12):1762–1777. doi:10.1093/treephys/tpaa098.

Zhang H, Zhang TT, Liu H, Shi DY, Wang M, Bie XM, Li XG, Zhang XS. 2018. Thioredoxin-mediated ROS homeostasis explains natural variation in plant regeneration. Plant Physiol. 176(3):2231–2250. doi:10.1104/pp.17.00633.

Zhang J, Yang Y, Zheng K, Xie M, Feng K, Jawdy SS, Gunter LE, Ranjan P, Singan VR, Engle N, et al. 2018. 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 Phytol. 220(2):502–516. doi:10.1111/nph.15297.

Zhang X, Zhu Z, An F, Hao D, Li P, Song J, Yi C, Guo H. 2014. Jasmonate-activated MYC2 represses ETHYLENE INSENSITIVE3 activity to antagonize ethylene-promoted apical hook formation in Arabidopsis. Plant Cell. 26(3):1105–1117. doi:10.1105/tpc.113.122002.

Zhang Y, Cheng YT, Qu N, Zhao Q, Bi D, Li X. 2006. Negative regulation of defense responses in Arabidopsis by two NPR1 paralogs. Plant J Cell Mol Biol. 48(5):647–656. doi:10.1111/j.1365-313X.2006.02903.x.

Zhang Y, Fan W, Kinkema M, Li X, Dong X. 1999. Interaction of NPR1 with basic leucine zipper protein transcription factors that bind sequences required for salicylic acid induction of the PR-1 gene. Proc Natl Acad Sci USA. 96(11):6523–6528. doi:10.1073/pnas.96.11.6523.

Zhang Y, Goritschnig S, Dong X, Li X. 2003. 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. Plant Cell. 15(11):2636–2646. doi:10.1105/tpc.015842.

Zhang Y, Tessaro MJ, Lassner M, Li X. 2003. Knockout analysis of Arabidopsis transcription factors TGA2, TGA5, and TGA6 reveals their redundant and essential roles in systemic acquired resistance. Plant Cell. 15(11):2647–2653. doi:10.1105/tpc.014894.

Zhao H, Shi J, Qi X, Wang X, Jia J. 2017. Pyramid scene parsing network. arXiv. doi:10.48550/arXiv.1612.01105.

Zhou X, Stephens M. 2012. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 44(7):821–824. doi:10.1038/ng.2310.

Zhu Z, An F, Feng Y, Li P, Xue L AM, Jiang Z, Kim J-M, To TK, Li W, *et al.* 2011. Derepression of ethylene-stabilized transcription factors (EIN3/EIL1) mediates jasmonate and ethylene signaling synergy in Arabidopsis. Proc Natl Acad Sci USA. 108(30):12539–12544. doi:10.1073/pnas.1103959108.

Zimmermann IM, Heim MA, Weisshaar B, Uhrig JF. 2004. Comprehensive identification of *Arabidopsis thaliana* MYB transcription factors interacting with R/B-like BHLH proteins. Plant J Cell Mol Biol. 40(1):22–34. doi:10.1111/j.1365-313X.2004.02183.x.

Zou K, Ge L, Zhang C, Yuan T, Li W. 2019. Broccoli seedling segmentation based on support vector machine combined with color texture features. IEEE Access. 7:168565–168574. doi:10.1109/ACCESS.2019.2954587.

Editor: L. McIntyre