the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2020) **103,** 512–531 doi: 10.1111/tpj.14746

SNPeffect: identifying functional roles of SNPs using metabolic networks

Debolina Sarkar and Costas D. Maranas*

Department of Chemical Engineering, Pennsylvania State University, University Park, PA, USA

Received 30 October 2019; accepted 20 February 2020; published online 13 March 2020.
*For correspondence (e-mail: costas@psu.edu).

SUMMARY

Genetic sources of phenotypic variation have been a focus of plant studies aimed at improving agricultural yield and understanding adaptive processes. Genome-wide association studies identify the genetic background behind a trait by examining associations between phenotypes and single-nucleotide polymorphisms (SNPs). Although such studies are common, biological interpretation of the results remains a challenge; especially due to the confounding nature of population structure and the systematic biases thus introduced. Here, we propose a complementary analysis (SNPeffect) that offers putative genotype-to-phenotype mechanistic interpretations by integrating biochemical knowledge encoded in metabolic models. SNPeffect is used to explain differential growth rate and metabolite accumulation in *A. thaliana* and *P. trichocarpa* accessions as the outcome of SNPs in enzyme-coding genes. To this end, we also constructed a genome-scale metabolic model for *Populus trichocarpa*, the first for a perennial woody tree. As expected, our results indicate that growth is a complex polygenic trait governed by carbon and energy partitioning. The predicted set of functional SNPs in both species are associated with experimentally characterized growth-determining genes and also suggest putative ones. Functional SNPs were found in pathways such as amino acid metabolism, nucleotide biosynthesis, and cellulose and lignin biosynthesis, in line with breeding strategies that target pathways governing carbon and energy partition.

Keywords: flux balance analysis, SNPs, metabolic networks, Arabidopsis, poplar, plant metabolic modeling, complementary GWAS.

INTRODUCTION

Tremendous natural variation in growth and development is seen in a variety of plant species (Maloof, 2003). Harnessing this diversity intelligently necessitates an understanding of the causal relationship between genetic polymorphisms and phenotypic differences observed within a species. In the past decade, driven by the enormous progress in genomic sequencing and data processing abilities, genetic approaches using natural variation to identify genes underpinning quantitative traits have gained traction. In particular, genome-wide association studies (GWAS) are used to identify causative/predictive factors for a given trait (Nielsen *et al.*, 2011; Prasad *et al.*, 2012; Soltis and Kliebenstein, 2015). This is commonly implemented by evaluating the statistical significance of the association between quantitative differences of a phenotype and the genetic polymorphisms seen in a set of genetically distinct individuals. However, the efficacy of a GWAS analysis in identifying truly causative SNPs is heavily dependent on the phenotypic variance observed in the population, with both rare and common loci presenting problems in results interpretation (Gibson, 2012).

Efforts to augment GWAS using prior biological knowledge began with the introduction of pathway-based approaches (Wang *et al.*, 2007; Wang *et al.*, 2010). These methods increased power by grouping genes by their co-occurrences in metabolic pathways or gene ontologies. As genes do not work in isolation but instead interact via complex molecular networks and cellular pathways, such an analysis examines whether test statistics for a group of related genes consistently deviates significantly from that obtained by chance. Wang *et al.* (2007) were the first to propose a pathway-based GWAS which was akin to Gene Set Enrichment Analysis. The authors re-examined data from published GWAS by pre-defining gene sets for pathways and then calculating the significance of each pathway based on the association of markers in or near genes. In doing so, they were able to identify biologically plausible signals on multiple datasets, such as the glutamate receptor in Parkinson's disease (Fung *et al.*, 2006) was identified

with statistical significance and is known to be a susceptible pathway (Marino *et al.*, 2003). Interestingly, the original GWAS by Fung *et al.* was highly discordant with another GWAS study of Parkinson's disease (Maraganore *et al.*, 2005), underscoring the importance of looking beyond the most-significant genes. Since then, set-based approaches have been widely used to reduce the GWAS multiple-testing burden while enriching the association signal in individual SNP-based tests (Allen *et al.*, 2010; Nurnberger *et al.*, 2014; Locke *et al.*, 2015). However, these methods assumed independence between pathways, whereas in reality genes from distal pathways may be functionally connected (Kelley and Ideker, 2005; Tong *et al.*, 2004; Xu *et al.* 1994; Zhou *et al.*, 1998). Recognition of this limitation led to the development of network-based analyses. A scaffold of protein–protein interaction networks (PPIs) was used to infer connections between gene products, ultimately honing-in on subnetworks enriched with SNP-associated genes (Leiserson *et al.*, 2013). Network 'guilt by association' (Lee *et al.*, 2011) exploits the fact that genes associated with a trait are more likely to organize into functional groups. Protein interaction network-based pathway analysis (PINBPA) was used to identify over-represented modules associated with multiple sclerosis using aggregated gene-wise statistics (Wang *et al.*, 2014b). Potential disease-associated mechanisms (such as DNA damage response and cell waste disposal) were identified in frontotemporal dementia using a weighted protein–protein interaction network analysis (WPPINA), which analyzes tissue-specific interactomes implicated by known disease-associated PPIs (Ferrari *et al.*, 2017). The prioritization of genes was further improved upon, such as in Sharma *et al.* in which a putative neighborhood was identified using a degree-adjusted random walk around known disease-causing genes (Sharma *et al.*, 2018). Network-based analysis has also been used to amplify marker signals from existing GWAS studies. Liu *et al.* (2017) identified genes that contributed significantly to childhood-onset asthma by searching for consistent gene modules between two large GWAS datasets.

In this paper, we take the next step by not simply *a priori* assigning genes to pathways and known phenotypic traits, but also explaining changes in growth rate, and differential metabolite and enzyme levels as the outcome of one or more SNPs in the coding regions of the genotype. We supplement the purely data-driven discovery in GWAS using mechanistic information that underpins known biochemistry and metabolic network structure. SNPeffect, in essence, focuses on explaining SNPs in only enzyme-coding (or regulatory) regions by constructing scenarios for their mechanistic role by integrating all available metabolomics, proteomics, and metabolic network information into a self-consistent narrative. SNPeffect can thus be used *a posteriori* to evaluate the functional explanation of GWAS-obtained hits or *a priori* to generate linkages between SNPs, genes, enzymes, metabolites, and various traits that can be used as priors in GWAS and other plant breeding techniques. In the present study, we used SNPeffect to investigate the basis behind growth in different *Arabidopsis thaliana* and *Populus trichocarpa* accessions by exploiting naturally occurring genotypic and phenotypic variations. A set of putative causal loci ranging from genes in protein synthesis to lignin metabolism was identified with distinct patterns of activating and deactivating SNPs (see Supporting Information Table S1 for *Arabidopsis* and Table S2 for poplar SNPeffect results). Faster growing *Arabidopsis* genotypes were predicted to have higher fluxes through the amino acid metabolism pathways and preferentially employed the energy-efficient purine salvage pathway as opposed to the *de novo* purine synthesis pathway for generating the energy metabolites AMP and GMP. Putative growth-related SNPs were found in multiple pathways such as folate biosynthesis, branched-chain amino acid biosynthesis, and shikimate metabolism. In *P. trichocarpa,* we observed that SNPs that were significantly associated with genotype growth were responsible for increasing flux through amino acid, pigment, cellulose, and nucleotide biosynthetic reactions. For example, the genes encoding for sucrose synthase were found to have more activating SNPs in faster growing genotypes, indicating that the degradation of sucrose to produce cellulose via UDP-glucose enhances growth. We also investigated epistasis controlling biomass production between enzymes in *Arabidopsis* and poplar, and captured known epistatic interactions as well as predicted putative ones. Interestingly, we identified more genetic interactions between pathways rather than between genes within them, contrary to studies in microbial systems such as yeast (Segrè *et al.*, 2005).

RESULTS

Integrating systems-level 'omics data to assign functional roles to SNPs

We developed a method called SNPeffect for identifying functional SNPs among multitude of polymorphisms that typically occur in natural populations. Genotypic and phenotypic data superimposed on a plant metabolic model serve as inputs. Then, the most parsimonious distribution of functional SNPs is identified that is required to explain variations in growth and metabolite levels. This modeling approach takes into account the direct link between genes with SNPs, encoded enzymes, network representation of metabolism, and ultimately biomass formation through the availability of precursors. The presence of a SNP in a gene can either over-regulate or under-regulate the V_{max} of the corresponding enzyme by either increasing or decreasing its activity and/or abundance. We require that the same

SNP in different genotypes causes the same direction of change (either enhancing or suppressing enzyme activity) but not at the same magnitude. This effect is propagated at the metabolic level by increasing or decreasing the corresponding reaction flux. This, in turn, may ultimately affect growth by replenishing or limiting the availability of a biomass precursor. For example, consider a case when the concentration of a metabolite in a genotype is reduced as compared with the reference. However, this genotype does not suffer a growth defect despite the metabolite producing a limiting biomass precursor. This is possibly because the enzymatic turnover (i.e. V_{max}) of the reaction using that metabolite as a substrate is increased by polymorphism(s) that either directly affect its kinetics (i.e. k_{cat}) or the total abundance [E]. Thus, SNPeffect will flag a SNP in the corresponding gene as a putative activator that increases enzyme V_{max}. Conversely, if this genotype exhibited a lower productivity despite having high levels of the limiting metabolite, then associated SNP(s) would be flagged as inhibiting. One could envision more complicated patterns of metabolite concentration changes superimposed on SNPs in enzyme-coding regions, causing either enhanced or suppressed growth across genotypes. Thus, making sense of these large heterogeneous sets of data requires taking an algorithmic approach. Obviously, not all changes in metabolite concentrations, growth rates, and location of SNPs would be perfectly consistent with one another. Therefore, SNPeffect finds causal outcomes for SNPs so as to explain the maximal amount of variance in the data while obtaining consistent flux distributions. The output of SNPeffect is the identification of a subset of SNPs that consistently have either an activating (or inactivating) effect in all genotypes. SNPeffect only makes use of SNPs in enzyme-coding regions and a 10 kbp flanking region (as gene activity can be affected by the presence of SNPs in enhancer/repressor regions; Biscarini *et al.*, 2016; Brodie *et al.*, 2016; Lee and Shatkay, 2008; Torkamani *et al.*, 2008). This is typically only about 0.3% of the SNPs in the population. In addition, SNPeffect parsimoniously assigns a functional role to only a fraction (from 1.24% to 4.1%) of the metabolic SNPs. This fraction depends on the number of genotypes for which sequence and 'omics data are available in the population.

Figure 1 illustrates how SNPeffect works for a three genotype (G_1, G_2,and the reference genotypeG_{ref}) toy example (Figure 1a). Note that the production of metabolite B from A is limiting for biomass synthesis. Both genotypes G_1 and G_2 have half of the concentration of metabolite A compared with the reference genotype G_{ref}. However, genotype G_2 grows twice as fast as G_{ref}, whereas G_1 half as fast as G_{ref} (Figure 1b). SNP data for the corresponding gene sequences are shown in a Boolean matrix, in which rows correspond to genotypes and columns refer to the genomic co-ordinate. The presence of a SNP is recorded with an entry of one in this matrix. Because genotype G_2 grows faster than the reference despite having the same levels of metabolite A, SNPeffect will assign one or more activating SNP(s) in G_2 responsible for increasing the V_{max} of the corresponding enzyme. The putatively activating SNP is selected from the entries in red in the last row of the matrix shown in Figure 1(b). Figure 1(c) shows the resultant flux distributions in the three genotypes. No functional SNPs are assigned by SNPeffect for genotype G_1 as its reduced relative growth rate can be explained solely by the reduced levels of metabolite A. However, in genotype G_2 flux through the biomass reaction is 20 mmol/gDW h^{-1} which is twice that in G_{ref}. Mass-action kinetics dictates that the flux through v_{G_2} can be at most 5 mmol/gDW h^{-1} as the concentration of A has not changed. Therefore, the required flux increase must be attributed to another factor such as an activating SNP in the sequence of the corresponding gene. This is encoded in SNPeffect by the departure variable $SNPdev^{pos}_{G_2}$ which assumes a value of 15 mmol/gDW h^{-1} to be consistent with the higher final relative growth rate (RGR) in genotype G_2. The magnitude of variable $SNPdev^{pos}_{G_2}$ is explained as the additive action of the two SNPs at genomic positions 23 and 26. The relative influences of the two SNPs are encoded by variables X^{pos}_{23,G_2} and X^{pos}_{26,G_2}, respectively. The SNP at position 26 is absent in genotype G_1 and only present in genotype G_2. It is assigned an activating role by SNPeffect as it provides the most parsimonious explanation for the increased relative growth of genotype G_2.

In SNPeffect, slack variables $SNPdev^{pos}_{jl}$ and $SNPdev^{neg}_{jl}$ are used to capture the effect of SNP(s) in a genotype l on flux through the enzymatic reaction j. A negative deviation from the reference flux distribution (i.e. a non-zero value of $SNPdev^{neg}$) indicates that the value of the reaction flux must decrease to satisfy the imposed metabolomics-obtained constraints. This can be putatively attributed to an inactivating SNP (if one is present in the corresponding gene). Otherwise, this discrepancy remains unexplained and is assigned to variable dev^{neg}. Similarly, a positive deviation from the reference flux distribution (i.e. a non-zero value of $SNPdev^{pos}$) indicates that the reaction flux must increase to satisfy the imposed metabolomics-obtained constraints. This can then be explained as the effect of an activating SNP (if one is present), and otherwise the flux discrepancy remains unexplained and is stored in dev^{pos}. Only SNP(s) in the gene(s) encoding for the catalyzing enzyme (for reaction j) are considered; this information is encoded in a three-dimensional sparse Boolean matrix A_{jkl} which includes an entry of one if reaction j is associated with SNP k in genotype l.

Capturing known and putative growth-affecting SNPs in Arabidopsis

Arabidopsis thaliana was selected to benchmark SNPeffect as it has one of the best collection of functional genomics resources in plants (Togninalli *et al.*, 2018). Across all

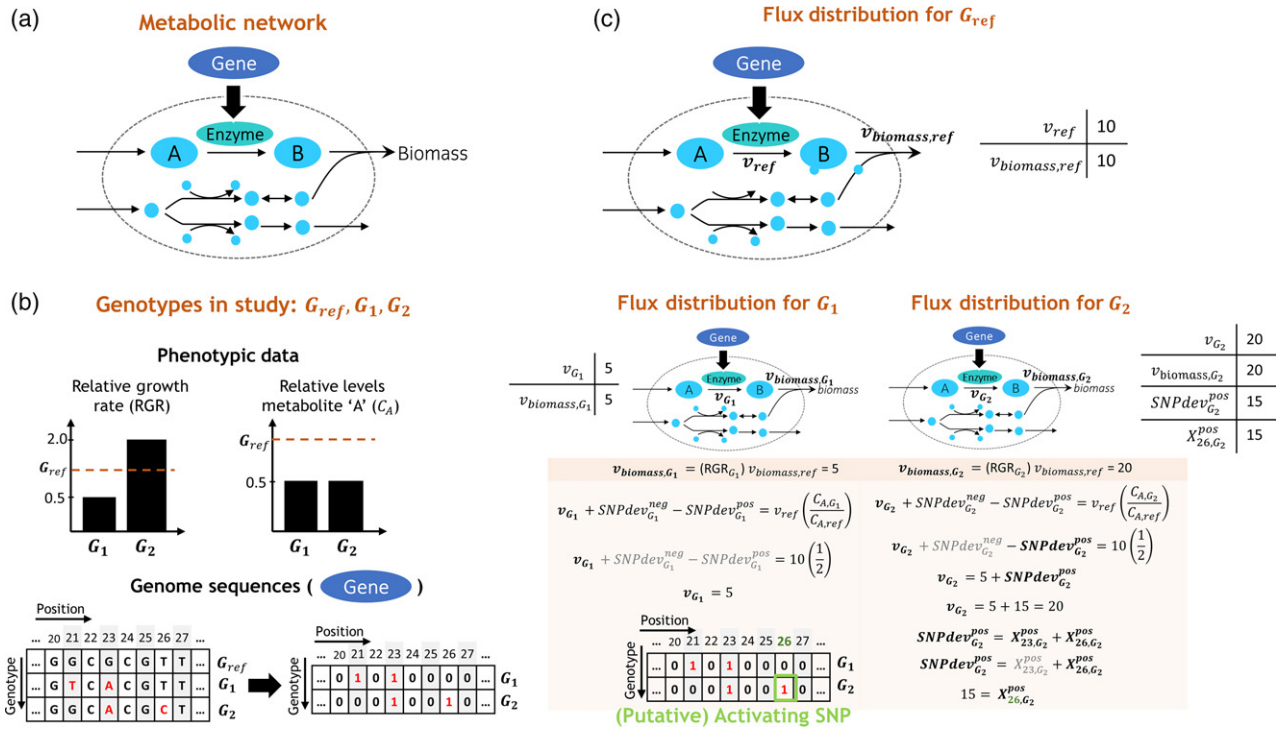

Figure 1. Illustrative SNPeffect example on a toy metabolic network and three genotypes.
(a) Toy metabolic network showing the reaction converting metabolite A to B which is limiting for biomass synthesis. (b) Sample SNPeffect input data for genotypes G_1, G_2, and the reference genotype G_{ref}. Note that the phenotypic data consist of relative growth rates (RGR) and metabolite levels with respect to the reference genotype. Genotypic data show SNP positions (an entry of one denotes the presence of a SNP, whereas zero for no difference from the reference). (c) SNPeffect-predicted flux distributions in the three genotypes. The relative concentration of metabolite A in genotype G_1 is consistent with the reduced growth rate, which is why no functional SNPs are identified. However, genotype G_2 exhibits a doubling in RGR, while retaining the same level for metabolite A. The increase in RGR is explained by SNPeffect by the effect of an activating SNP at position 26.

genotypes, 72 unique genes were identified as being differentially affected by the presence of 340 SNPs (Table S1). Corroborating literature evidence was found for 37 genes (*c.* 51%) for which perturbations in their expression (upregulation/downregulation or knockout) caused growth-related phenotypic changes in *Arabidopsis* (Table S1). We structurally characterized the identified SNP set (using the CDD database; Marchler-Bauer *et al.*, 2017) to determine if they featured in conserved protein residues such as those coding for active and catalytic sites, substrate and co-factor binding domains, and dimer interfaces as these are known to affect protein function and thus enzymatic activity (Bhatnager and Dang, 2018; Esaki *et al.*, 2012; Evnouchidou *et al.*, 2011; Liu *et al.*, 2011). SNPeffect identified 73 within gene and 267 SNPs in the 10 kb flanking region across 72 unique genes in *Arabidopsis*, out of which functional domains could be identified in 30 genes for 32 SNPs (*c.* 43% of within gene SNPs) (Table S1). To assess algorithm performance, we calculated the precision and recall rates of SNPeffect (Guinot *et al.*, 2018). Alternate optimal solutions were found iteratively using integer cuts. An integer cut (implemented as a model constraint) serves to render a previously identified integer solution (i.e. combination of y_k^{pos} and y_k^{neg} in this case) infeasible without excluding any other feasible solutions. After appending 100 integer cuts, 661 SNPs were identified across all solution sets (all of which had the same objective function value). Here, 105 SNPs (*c.* 16%) were common to all solution sets out of which 98 had the same role (i.e. strictly activating/inactivating); 169 SNPs (*c.* 26%) appeared in at least two solution sets and maintain consistent activity, indicating that although there exist multiple optima there is considerable overlap between the SNPs identified in each case. The recall rate and precision were then calculated as a function of the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). The precision (defined as TP/(FP + TP)) was on average 0.48 and the recall (defined as TP/(FN + TP)) was 0.47. The enrichment of literature evidence seen in the SNPeffect-predicted functional SNPs is significant with a *P*-value of 1.06e-63 (hypergeometric test). This demonstrates that SNPeffect captures information encoded within SNP sets, metabolomics, and RGR datasets to make prediction that rises well above any random occurrence threshold.

The SNPeffect-identified gene set (i.e. genes with functional SNPs) for every genotype was significantly enriched for essential genes (*P*-value < 0.05, hypergeometric test). *In silico* essential genes are associated with reactions

which, upon deletion, prevent biomass synthesis in the metabolic model. This indicates that SNPs in these genes cause (on average) a greater phenotypic change than ones in non-essential genes. This is not surprising as essential genes are required for growth, and thus any reduction in their activity can have a proportionally adverse effect on fitness. Essentiality investigations in plants have mainly focused on investigating the loss-of-function phenotypes associated with gene knockouts (Lloyd and Meinke, 2012; Savage *et al.*, 2013; Wang *et al.*, 2014a), but microbes such as *S. cerevisiae* have been used to compare the impact of perturbations in essential versus non-essential genes. Variations in gene expression for essential genes gave rise to greater phenotypic changes than non-essential gene deletions (Bauer *et al.*, 2015). These conclusions, of course, remain untested in plants.

Next, we compared the predicted functional SNP set to previous GWA studies and found a number of SNPs to be concordant between these two complementary methods (see Table S1 for a full list). Two functional SNPs identified in shikimate kinase (SK) were found in a previous *Arabidopsis* GWAS (Atwell *et al.*, 2010) to be significantly associated with serrated leaves and flowering time, which has been linked to leaf growth (Cookson *et al.*, 2007). The shikimate pathway is involved in the synthesis of L-tryptophan, L-phenylalanine, and L-tyrosine (Figure 2) alongside many aromatic secondary metabolites such as alkaloids, flavonoids, and lignin (Tohge *et al.*, 2013). It is estimated that 20–50% of the total fixed carbon in land plants passes through this pathway (Ni *et al.*, 1996; Corea *et al.*, 2012), indicating its central role in *Arabidopsis* growth. Functional SNPs were also found in genes coding for 3-dehydroquinate synthase (DHQS) from the same pathway. Although the connection between perturbations in this pathway and their impact on *Arabidopsis* growth has not yet been investigated, the predicted phenotype is consistent with a recent study (Guo *et al.*, 2018). Guo *et al.* found DHQS and SK2 to be upregulated in plants with chronic JAZ (JASMONATE ZIM DOMAIN) deficiency, resulting in reduced growth rate and root length.

Genes belonging to other pathways of amino acid metabolism, encoding DAHP synthase, chorismate synthase, anthranilate synthase, phosphoribosylanthranilate isomerase (PRAI), ketol-acid reductoisomerase (KARI), and 2-isopropylmalate synthase (IPMS) were predicted to have functional SNPs which have also been implicated in a previous GWA analysis to be associated with growth-related phenotypes (Atwell *et al.*, 2010). SNPs in the genes AT1G29410 and AT1G07780 encoding for PRAI were associated with days to flowering time and leaf number, respectively. The SNP at 21 680 457 in the gene AT3G58610 encoding for KARI was significantly associated (*P*-value = 2.67e-5) with vegetative growth rate and the inactivating SNP at 16 121 370 in AT4G33510 (encoding DAHP synthase) was previously found to be significantly associated (*P*-value = 9.71e-5) with flowering time (Atwell *et al.*, 2010). This is possibly because amino acid metabolism is known to be intrinsically linked to plant productivity and constitutes a primary growth requirement. Increasing amino acid availability by either increasing the expression of specific amino acid transporters or by supplementing them in the growth medium has been shown to increase *Arabidopsis* growth (Forsum *et al.*, 2008).

In addition to verifying SNPeffect predictions, comparing overlapping SNPs with GWAS can also help provide *a posteriori* narratives for GWAS associations by mechanistically explaining their impact on plant metabolism. For example, two functional SNPs flagged in IPMS have also been implicated in numerous GWA studies as being significantly associated with flowering time, aphid resistance, and leaf number (Atwell *et al.*, 2010; Li *et al.*, 2010; Alonso-Blanco *et al.*, 2016). IPMS catalyzes the first committed step of leucine biosynthesis, converting acetyl-CoA and 3-methyl-2-butanoate into 2-isopropylmalate. SNPeffect predicted slower-growing genotypes to have inactivating SNPs in genes encoding for IPMS, but faster growing genotypes to have compensatory activating SNPs, thereby restoring flux through the reaction (Figure 2). This indicates that IPMS is pivotal for controlling flux through the leucine biosynthetic pathway and contributes to biomass production. A similar phenotype has also been characterized experimentally, albeit in rice. Proteomic analysis showed that IPMS is involved in releasing seed dormancy (Xu *et al.*, 2016) and IPMS knockout mutants exhibit deficient amino acid synthesis, low glycolytic activity, and reduced ATP and AMP production (He *et al.*, 2019).

Lastly, a number of putative SNPs identified in aminoacyl-tRNA synthesis have also been found in previous GWA studies to be related to growth phenotypes such as leaf number (Atwell *et al.*, 2010), leaf chlorosis (Atwell *et al.*, 2010), and ion concentration in leaves (Atwell *et al.*, 2010; Chao *et al.*, 2014) (Table S1). This particular pathway is known to be associated with plant growth, as was observed in a screening of *Arabidopsis* accessions (Berg *et al.*, 2005). Loss-of-function mutations in aminoacyl-tRNA synthetase genes had adverse effects on growth, causing gametophytic lethality and interfering with protein synthesis causing ovule abortion.

Energy allocation is a primary growth determinant in *A. thaliana*

Next, we examined functional SNPs flagged in pathways of energy metabolism. Upon comparing SNPs to previous GWA studies, we found that an inactivating SNP in the gene encoding for phosphoribosylamine and glycine ligase in purine metabolism has been associated previously with seed dormancy using GWAS (Atwell *et al.*, 2010) (Table S1). The SNPs in genes coding for IMP

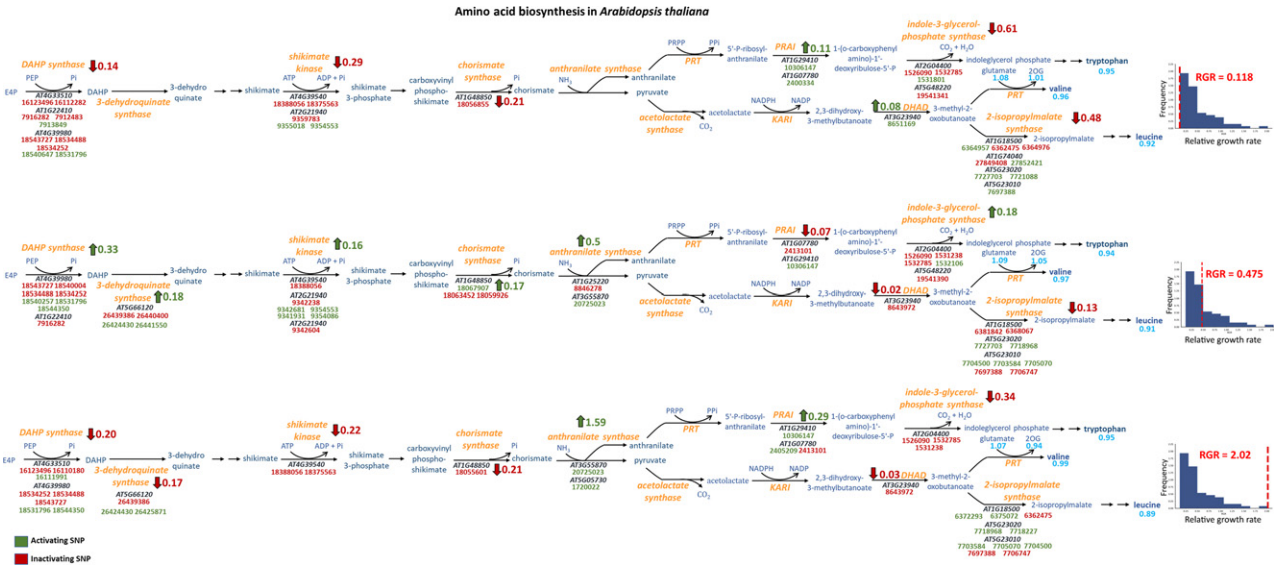

Figure 2. Functional SNPs in amino acid biosynthesis pathway of *A. thaliana* in three representative genotypes.
Distribution of SNPs in three representative genotypes which have a low (RGR = 0.118), average (RGR = 0.475), and high (RGR = 2.02) growth rate. Metabolite names are shown in blue (with relative metabolite levels shown in cyan below them, wherever available), gene IDs in black and corresponding enzymes in orange. Activating SNPs are marked in green and inactivating SNPs in red, below the gene that they feature in (marked in black). The net fold-change in flux (increase/decrease) through a reaction (with respect to the reference genotype) due to SNPs is given next to it, with a flux increase shown in green with an upward arrow and a flux decrease shown in red with a downward arrow. 2OG, 2-oxoglutarate; DAHP, 3-deoxy-D-arabino-hept-2-ulosonate 7-phosphate; DAHP synthase, 3-deoxy-7-phosphoheptulonate synthase; DHAD, dihydroxy-acid dehydratase; E4P, erythrose 4 phosphate; KARI, ketol-acid reductoisomerase; PEP, phosphoenolpyruvate; PRAI, phosphoribosylanthranilate isomerase; PRT, anthranilate phosphoribosyltransferase.

cyclohydrolase (the ultimate step in IMP biosynthesis) have been associated with flowering time in *Arabidopsis* (Atwell *et al.*, 2010; Li *et al.*, 2010), which has been linked to leaf growth (Cookson *et al.*, 2007). Interestingly, inactivating SNPs were found in genes AT4G38880 and AT2G16570 (Figure 3) that encode for amidophosphoribosyltransferase, indicating that plant growth is associated with decreased flux through purine metabolism. This enzyme catalyzes the first dedicated step of *de novo* purine biosynthesis by transferring an amido group from glutamine to 5-phosphoribosyl-1-pyrophosphate (PRPP). As purine nucleotides are major energy carriers of the cell and also serve as precursors for the synthesis of co-factors such as NADH and SAM, decreasing flux through this reaction would propagate downstream to reduce the levels of AMP and GMP. This would retard growth, in contrast with what is observed. This contradictory behavior may be explained by the fact that there are two routes for nucleotide synthesis in *Arabidopsis*, the *de novo* and salvage pathways. The *de novo* pathway converts PRPP and precursors such as CO_2 and tetrahydrofolate into purines while the salvage pathway interconverts purine bases, nucleosides, and nucleotides, which are released as catabolic by-products. The salvage pathway for purine synthesis is energetically favorable as only a single reaction requires ATP, whereas in the *de novo* pathway five of the 12 steps require ATP or GTP hydrolysis (Moffatt and Ashihara, 2002). SNPeffect predicts an energy-efficient mechanism at play wherein faster growing genotypes have downregulated *de novo* purine biosynthesis while the purine salvage pathway has no inactivating SNPs. Indeed, previous studies have established purine synthesis to be the growth-limiting step for both prokaryotic and eukaryotic cells (Liechti and Goldberg, 2012). Investigations of nucleotide metabolism in plants have focused on either the levels of nucleotides or the expression of genes involved in these pathways (Zrenner *et al.*, 2006), but for microbes the rate at which they generate GMP and AMP pools often directly correlates with growth (Hedstrom, 2009). It would be interesting to see if this observation holds for plants, but indirect evidence of the relative impact of these two pathways can be found in transgenic potato tubers. Decreasing UMP synthase expression, a key enzyme in the pyrimidine *de novo* biosynthetic pathway, stimulated the uridine salvage pathway, increased total carbon flux to starch and cell wall synthesis, and elevated tuber growth (Geigenberger *et al.*, 2005).

Overall, SNPeffect provided a narrative for the putative effect of 340 SNPs in 73 genes in the *Arabidopsis* dataset. Corroborating evidence was found for *c.* 51% of them after a cursory literature search. We anticipate that putative causal roles for many more SNPs can be deciphered if the SNPeffect analysis is repeated with additional genotypes, phenotypes (such as nutrient exchange fluxes or rates of photosynthetic oxygen evolution) and/or datasets (e.g. proteomics or transcriptomics).

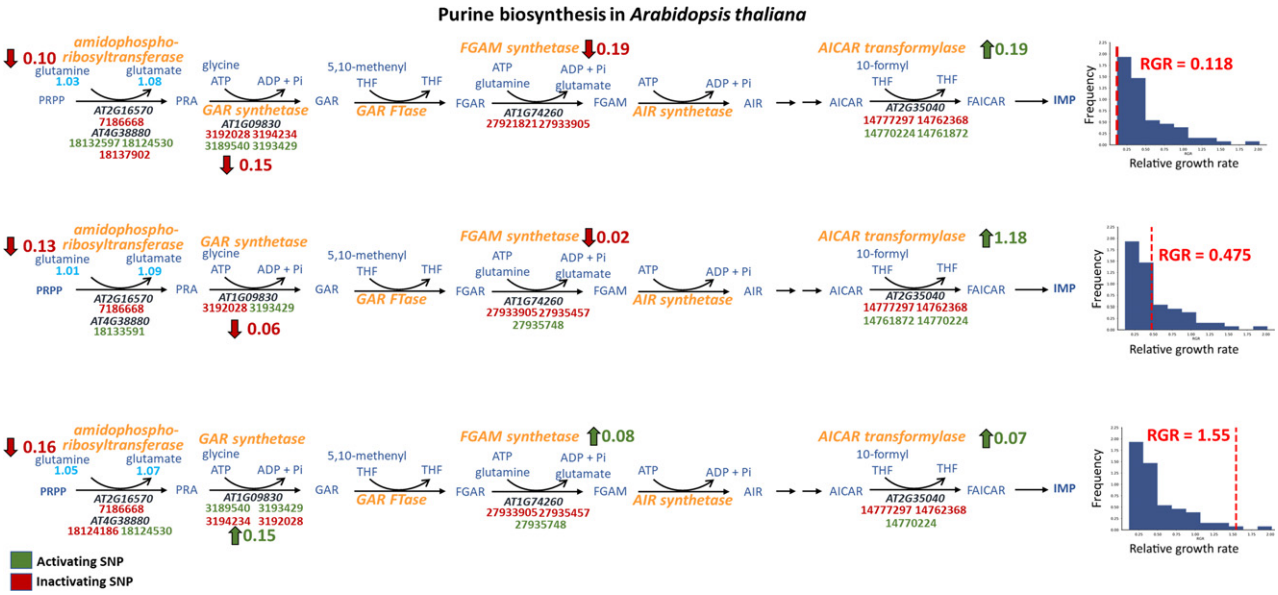

Figure 3. Functional SNPs in the purine biosynthesis pathway of *A. thaliana* in three representative genotypes.
Distribution of SNPs in three representative genotypes which have a low (RGR = 0.118), average (RGR = 0.475), and high (RGR = 2.02) growth rate. Metabolite names are shown in blue (with relative metabolite levels shown in cyan below them, wherever available), gene IDs in black and corresponding enzymes in orange. Activating SNPs are marked in green and inactivating SNPs in red, below the gene that they feature in. The net fold-change in flux (increase/decrease) through a reaction due to SNPs is given next to it, with a flux increase shown in green with an upward arrow and a flux decrease shown in red with a downward arrow. AIR, 5-aminoimidazole ribonucleotide; AICAR, aminoimidazole carboxamide ribonucleotide; AICAR transformylase, phosphoribosyl aminoimidazole carboxamide formyltransferase; FAICAR, 5-phosphoribosyl-formamido-carboxamide; FGAR, 5-phosphoribosyl-*N*-formylglycinamide; GAR, 5-phospho-β-D-ribosyl-glycinamide; GAR synthetase, phosphoribosylamine—glycine ligase; GAR TFase, phosphoribosylglycinamide formyltransferase; PRA, 5-P-β-D-ribosyl-amine; THF, tetrahydrofolate.

SNPs affecting cell wall biosynthesis in *P. trichocarpa*

Similar to *Arabidopsis*, we used SNPeffect to identify SNPs associated with metabolic flux changes in reactions that generate biomass precursors and thus affect plant growth in *P. trichocarpa.* For this, we employed the rich dataset of SNPs and growth phenotypes available in literature (Muchero *et al.*, 2015; BESC, 2017). SNPeffect identified 843 putatively functional SNPs in 365 genes (Table S2), out of which 583 SNPs feature within a gene and 260 in the 10 kb flanking region. The SNPeffect-identified gene set was significantly enriched for essential genes (P-value < 0.05, hypergeometric test). The predicted functional SNPs also showed a significant overlap (P-value = 0.0086, hypergeometric test) (Table S2) with a previous GWAS (Evans *et al.*, 2014). As Evans *et al.* used the same mapping population as SNPeffect and examined the same growth trait, it is expected that phenotypic traits should be controlled by the same set of loci. We structurally characterized the SNP set to identify those featuring in functional protein domains. At least one functional site could be identified in 191 genes using the CDD database (Marchler-Bauer *et al.*, 2017), and 149 SNPs (*c.* 25% of within gene SNPs) among 101 unique genes were predicted to be in functional residues (Table S2).

Pathways of cell wall metabolism were predicted to be enriched in functional SNPs, indicating that flux through them is potentially growth regulating. The total flux increase through sucrose synthase (SuSy; E.C. 2.4.1.13) due to activating SNPs was found to be directly related to the genotype's biomass yield (Pearson's correlation coefficient = 0.964, P-value = 0.008165). Cellulose is produced from UDP-glucose, which in turn is made from the cleavage of sucrose catalyzed by SuSy. SuSy plays a central role in plant metabolism by modulating the carbohydrate sink (Sun *et al.*, 1992; Zrenner *et al.*, 1995; Dejardin *et al.*, 2015) and its overexpression in tobacco increases plant height, cellulose content, fiber length, and soluble sugar accumulation (Wei *et al.*, 2015). Conversely, suppressing SuSy activity in cotton plants produces a fiberless cotton phenotype and shrunken seeds (Ruan, 2003). Thus, poplar seems to be similarly dependent on sucrose synthase for cellulose biosynthesis and our results indicate that upregulating this particular gene would increase growth.

In addition to SuSy, SNPeffect flagged SNPs in other enzymes involved in cell wall biosynthesis such as shikimate *O*-hydroxycinnamoyltransferase (HCT), phenylammonia-lyase (PAL), ferulate-5-hydroxylase (F5H), and trehalase. Trehalase hydrolyzes the sugar trehalose to produce two glucose molecules (Blázquez *et al.*, 1998; Vogel *et al.*, 1998). Blocking its activity resulted in stunted growth and lancet-shaped leaves in tobacco plants (Goddijn *et al.*, 2002). The other cell wall-producing enzymes belong to the

lignin biosynthetic pathway (Figure 4), and lignin is known to be a significant metabolic sink for carbon and energy. It is a major contributor to woody biomass recalcitrance as it prevents the access of extracellular enzymes to the degradable sugar moieties. Thus, there is considerable interest in developing plants with reduced lignin content that do not incur a growth penalty (Chanoca *et al.*, 2019) and SNPeffect-identified SNPs can serve as potential tunable levers for controlling flux through this pathway. Indeed, growth-related phenotypes have been reported for lignin biosynthetic genes flagged as having functional SNPs. PAL is responsible for converting phenylalanine to cinnamic acid, which constitutes the first dedicated step of lignin biosynthesis in poplar. PAL overexpression has been observed to negatively impact plant growth in transgenic poplar (Rueda-López *et al.*, 2017). HCT catalyzes the transfer of the caffeoyl moiety of 5-*O*-caffeoylquinate onto CoA (producing caffeoyl-CoA and quinate) and transgenic poplar with downregulated HCT genes exhibit slower growth and thinner stems than wild-type (Peng *et al.*, 2014; Zhou *et al.*, 2018). Two (putatively) activating SNPs were predicted in the HCT paralog PtHCT2 (Potri.018G105500), and both of these lie in the same interval on chromosome 18 (Chr18:13219799-13252693) that was found to be significantly associated with levels of *cis*- and *trans*-3-*O*-caffeoylquinic acid by GWAS and eQTL analysis (Zhang *et al.*, 2018a). The population used in SNPeffect is a subset of that used by Zhang *et al.* (2018a) due to which the same loci are expected to shape plant phenotypes. Ferulate 5-hydroxylase (F5H) was also predicted to have functional SNPs. This enzyme features further downstream in the lignin pathway, catalyzing the conversion of coniferaldehyde into 5-hydroxy-coniferaldehyde. Transgenic poplar overexpressing F5H show increased wood density (Koehler and Telewski, 2006), consistent with SNPeffect's prediction of F5H being a growth-related gene.

SNPs shaping secondary metabolism in poplar

Genes belonging to pathways of energy metabolism were found to have SNPs whose role was explainable by SNPeffect (Figure 5), such as dihydroorotase and dUTP diphosphatase. Activating SNPs were found in these genes, with the resultant flux increase being proportional to the genotype's biomass yield. Thus, SNPeffect results imply that upregulating the expression of these genes in poplar would increase downstream production of energy-generating precursors and thus enhance growth. Dihydroorotase catalyzes the reversible conversion of carbamoyl-L-aspartate into dihydroorotate and constitutes the third step of the pyrimidine *de novo* biosynthetic pathway. This pathway produces the nucleotide UMP from carbamoyl phosphate, aspartate, and PRPP. It has been reported that downregulation of this gene led to reduced height in potato plants (Schröder, 2005). dUTP diphosphatase features downstream of dihydro-orotase and was also predicted to have growth-affecting SNPs. This enzyme hydrolyzes dUTP into its corresponding monophosphate (dUMP) and its expression has been found to be directly proportional to growth rate in onion (*Allium cepa*) (Pardo and Gutiérrez, 1990).

SNPeffect predicted faster growing genotypes to have activating SNPs and thus a higher flux through pathways of amino acid metabolism. Serine hydroxymethyltransferase (SHMT) is a key enzyme associated with one-carbon metabolism in higher plants (Besson *et al.*, 1995) and catalyzes the reversible conversion of glycine and 5,10-methylenetetrahydrofolate into serine and tetrahydrofolate. Functional SNPs were identified in SHMT by SNPeffect and, interestingly, SHMT was recently explored as a potential genetic engineering target in *Populus* (Zhang *et al.*, 2019). Zhang and colleagues overexpressed a SHMT-encoding gene (*PtSHMT2*) in *P. deltoides* that increased biomass and sugar yields, while concomitantly decreasing the total lignin content. Figure 6 shows the functional SNPs identified in the lysine biosynthesis pathway, in genes encoding for aspartate kinase (AK), aspartate-semialdehyde dehydrogenase, 4-hydroxy-tetrahydrodipicolinate synthase (DHPS), LL-diaminopimelate aminotransferase, diaminopimelate aminotransferase, diaminopimelate decarboxylase, and homoserine dehydrogenase. These predictions are supported by lysine's role as an essential amino acid that serves as a precursor for proteins and for glutamate, which regulates plant growth and environmental responses (Galili, 2002). *Arabidopsis* mutants deficient in lysine biosynthesis exhibit retarded growth (Sarrobert *et al.*, 2000) and we predict a similar trend in poplar in which the net flux through DAP-decarboxylase increases with the genotype's growth rate (Figure 6). Phenotypes relating to perturbations in expression for other genes in this pathway have not been analyzed experimentally but AK and DHPS have been seen to be most abundant in growing tissues in Arabidopsis (Vauterin *et al.*, 1999; Zhu-Shimoni *et al.*, 1997). These can serve as candidate gene targets in future experimental investigations.

Epistatic interactions in *A. thaliana* and *P. trichocarpa* among SNPeffect-identified SNPs

SNPeffect putatively classifies a point mutation in a gene as being activating (i.e. increasing reaction flux), inactivating (i.e. decreasing reaction flux), or neutral. However, additional information on fitness can be gleaned by looking at pairs of SNPs in different genes. For example, consider the case where there are SNPs upstream and downstream in the synthesis pathway of a limiting biomass precursor. In such a case, the most positive influence on growth will be when the SNP upstream is activating and the SNP downstream is inactivating. Similarly, if both these SNPs were found in parallel pathways that produce

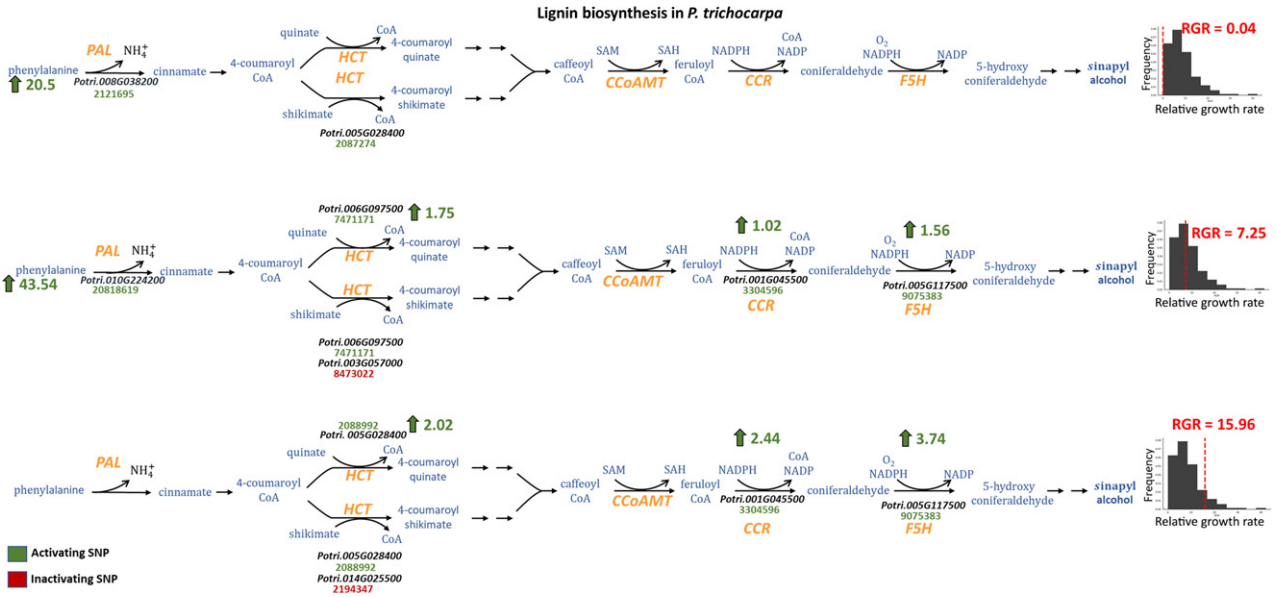

Figure 4. Functional SNPs in the lignin biosynthesis pathway of *P. trichocarpa* in three representative genotypes.
Distribution of SNPs in three representative genotypes which have a low (RGR = 0.04), average (RGR = 7.25), and high (RGR = 15.96) growth rate. Metabolite names are shown in blue, gene IDs in black and corresponding enzymes in orange. Activating SNPs are marked in green and inactivating SNPs in red below the gene that they feature in. The net fold-change in flux (increase/decrease) through a reaction due to SNPs is given next to it, with a flux increase shown in green with an upward arrow and a flux decrease shown in red with a downward arrow. CCoAMT, caffeoyl-CoA O-methyltransferase; CCR, cinnamoyl-CoA reductase; F5H, ferulate 5-hydroxylase; HCT, shikimate hydroxycinnamoyltransferase; PAL, phenylalanine ammonia-lyase.

the same limiting biomass precursor, maximum beneficial effect on growth will be when both are activating. Such interactions in which genetic mutations affect each other's phenotypic consequences are called epistatic. Epistasis between two mutations is said to be positive when a double mutant causes a weaker mutational defect than the individual mutations, and is negative when the double mutant causes a larger defect (Boone *et al.*, 2007; Phillips, 2008). If epistasis is prevalent, then a new mutant allele will interact with many other loci and alleles in the genetic background. The resultant genotype fitness might be dependent not only upon its direct effects on the phenotype, but also upon its effects through these interactions. With epistatic interactions, the effect of a SNP on a phenotype becomes the collective property of a network of SNPs.

Here we studied the spectrum of epistatic interactions between functional SNPs identified by SNPeffect in *A. thaliana* and *P. trichocarpa* by first calculating all possible epistatic interactions and then examining the distribution of SNPs among them. For modeling the inactivation of an enzyme (and hence its associated reaction(s)), the reaction flux(es) corresponding to that enzyme were constrained to be zero. For enzymes encoded by multiple genes, we included all that were associated with at least one gene with a causal SNP. We then calculated the maximal rate of biomass production with single and double deletions relative to the rate of the biomass production of the unperturbed wild-type network. Following previous flux balance analysis (FBA) studies examining epistasis (Xu *et al.*, 2012), the total flux through the network was also constrained to be less than or equal to the total flux through the unperturbed wild-type network (Schuetz *et al.*, 2012). When deleting an enzyme X, the fitness is defined as $W_X = v^{\Delta X}_{biomass} / v^{wild-type}_{biomass}$. For a pair of enzymes X and Y, we then evaluate the level of epistasis by comparing the fitness of the double mutant (W_{XY}) with the fitness of the single mutants W_X and W_Y. Similar to Segrè *et al.* (Segrè *et al.*, 2005), epistasis was calculated as shown in Table 1.

In *Arabidopsis*, 12 distinct epistatic interactions were identified across all genotypes in the study (Table S3) belonging to pathways of lignin biosynthesis, BCAA synthesis, pantothenate, and CoA biosynthesis, phenylalanine, tyrosine and tryptophan biosynthesis, and pyrimidine metabolism. Caffeoyl-CoA 3-*O*-methyltransferase and 4-coumarate-CoA ligase 1 (Figure 7b) were found to constitute a synthetic lethal pair. This is because simultaneous deletion of both genes would negate synthesis of feruloyl CoA and its downstream product coniferyl alcohol a precursor for the guaiacyl subunit of lignin. Such negative epistasis between the two genes has indeed been found previously in *P. tomentosa* using association mapping (Gong *et al.*, 2018). Four other synthetic lethal pairs were also identified between enzymes belonging to BCAA metabolism and pantothenate and CoA biosynthesis.

Using the constructed metabolic network for *P. trichocarpa*, we conducted a pairwise deletion scan to

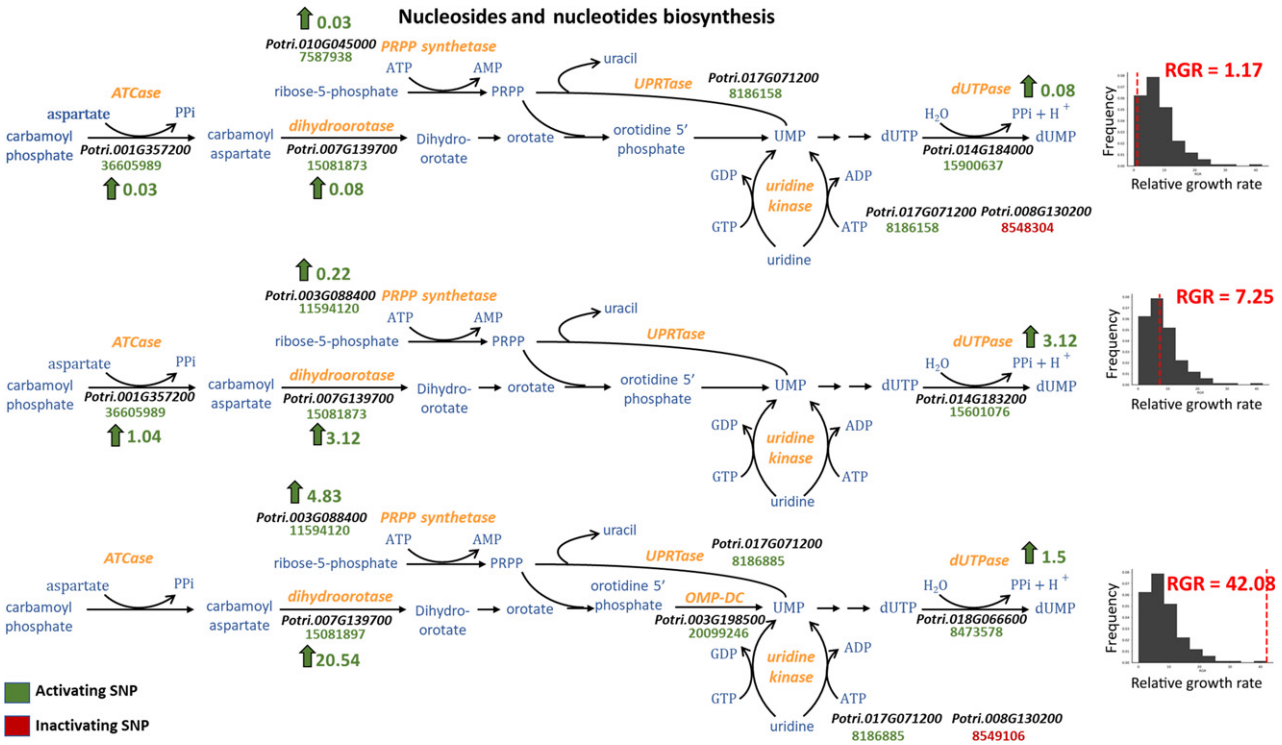

Figure 5. Functional SNPs in the purine and pyrimidine biosynthesis pathway of *P. trichocarpa* in three representative genotypes.
Distribution of SNPs in three representative genotypes which have low (RGR = 1.17), average (RGR = 7.25), and high (RGR = 42.08) growth rates. Metabolite names are shown in blue, gene IDs in black and corresponding enzymes in orange. Activating SNPs are marked in green and inactivating SNPs in red, below the gene that they feature in. The net fold-change in flux (increase/decrease) through a reaction due to SNPs is given next to it, with a flux increase shown in green with an upward arrow and a flux decrease shown in red with a downward arrow. ATCase, aspartate carbamoyltransferase; dUTPase, dUTP diphosphatase; OMP-DC, orotidine-5′-phosphate decarboxylase; PRPP synthetase, 5-phosphoribosyl-1-pyrophosphate synthetase; UPRTase, uracil phosphoribosyltransferase.

decipher the underlying network of epistatic interactions. Previously, studies investigating epistasis in poplar have focused mainly on cell wall-related pathways such as lignin and cellulose biosynthesis (Du *et al.*, 2015; Quan *et al.*, 2018) which also showed up in the present study (Table S6). Similar to *Arabidopsis*, both interpathway and intrapathway epistatic interactions among SNPs were found in poplar. Figure 7(a) summarizes predicted intrapathway epistasis in poplar terpenoid biosynthesis. Isoprenoid synthesis starts with the condensation of isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP). Both are produced by two distinct pathways in plants, the mevalonate (MVA) pathway and the methylerythritol 4-phosphate (MEP) pathway. As these two pathways represent complementary routes of producing the same essential precursors (IPP and DMAPP), a double mutation (one in each pathway) is likely to be more detrimental than a single one. As expected, our analysis also identified epistatic interactions between SNPs in genes belonging to these two pathways. 3-Hydroxy-3-methylglutaryl-coenzyme A reductase (HMGR) and mevalonate kinase, which catalyze the third and fourth steps of the MVA pathway, respectively, were found to interact with DXP synthase, MEP synthase, MECDP synthase, and 4-hydroxy-3-methylbut-2-enyl-diphosphate synthase from the MEP pathway. Evidence of metabolite exchanges have also been observed between them in spinach and tobacco, indicating that one pathway is capable of functionally compensating for the other (Hemmerlin and Bach, 1998; Bick and Lange, 2003; Hemmerlin *et al.*, 2003). Although the extent of this metabolic complementation is yet to be fully explored in plants, *Arabidopsis* genotypes carrying a mutation in hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate synthase (HDS) (the penultimate enzyme of the MEP pathway) exhibit an albino phenotype but are viable (Gutierrez-Nava, 2004).

Instances of epistasis between genes belonging to the same pathway were also identified, such as between SNPs in PAL and HCT, that were predicted to have a strong negative epistatic interaction. Previously, Quan *et al.* (2018) also predicted prevalent epistasis between genes in the lignin biosynthesis pathway, including PAL and HCT. We also predicted negative epistatic interactions between cellulose synthase (UDP-forming) and cellulose synthase (GDP-forming), whereby inactivating both of them had a much greater impact on poplar growth than a simple additive

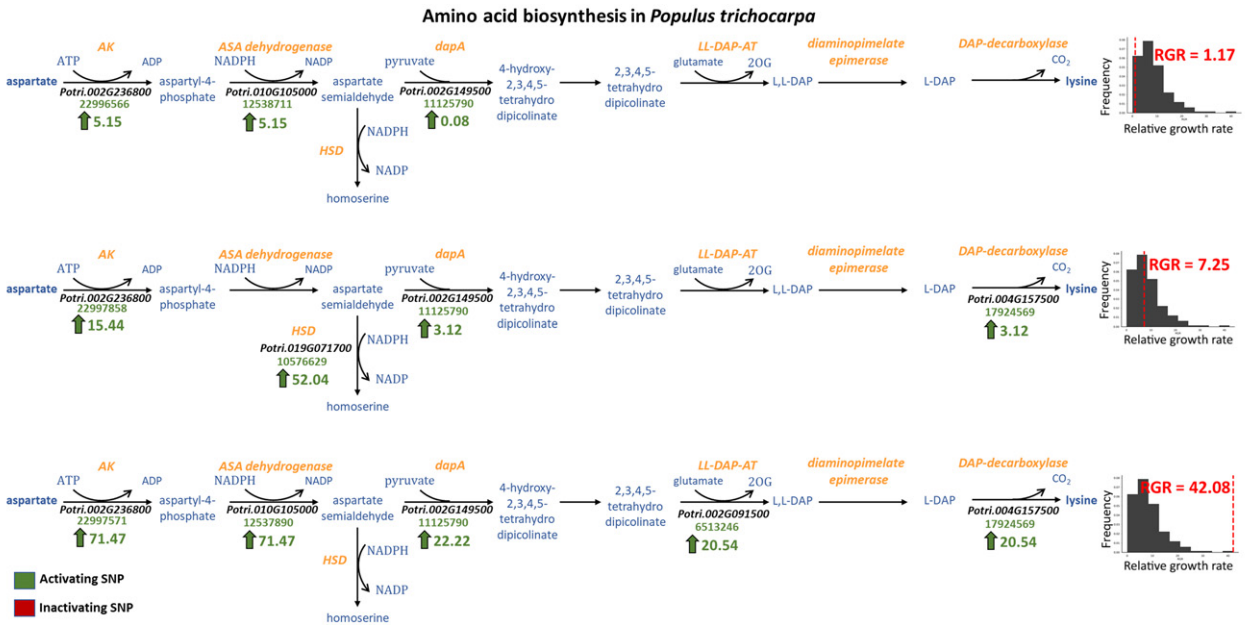

Figure 6. Functional SNPs in the amino acid biosynthesis pathway of *P. trichocarpa* in three representative genotypes.
Distribution of SNPs in three representative genotypes which have a low (RGR = 0.04), average (RGR = 7.25), and high (RGR = 42.08) growth rate. Metabolite names are shown in blue, gene IDs in black and corresponding enzymes in orange. Activating SNPs are marked in green and inactivating SNPs in red, below the gene that they feature in. The net fold-change in flux (increase/decrease) through a reaction due to SNPs is given next to it, with a flux increase shown in green with an upward arrow and a flux decrease shown in red with a downward arrow. AK, aspartate kinase; ASA dehydrogenase, aspartate-semialdehyde dehydrogenase; dapA, 4-hydroxy-tetrahydrodipicolinate synthase; DAP-decarboxylase, diaminopimelate decarboxylase; HSD, homoserine dehydrogenase; LL-DAP-AT, LL-diaminopimelate aminotransferase.

Table 1 Definitions of epistasis

	$\varepsilon = W_{XY} - W_X W_Y$
No epistasis	$\varepsilon = 0$
Negative epistasis (aggravating)	$\varepsilon < 0$
Positive epistasis (buffering)	$\varepsilon > 0$

When deleting an enzyme X, the fitness was defined as $W_X = v^{\Delta X}_{biomass} / v^{wild-type}_{biomass}$. For a pair of enzymes X and Y, epistasis is evaluated by comparing the fitness of the double mutant (W_{XY}) with the fitness of the single mutants W_X and W_Y.

contribution. This is because inactivation of both of these enzymes blocks cellulose biosynthesis altogether. Other studies have also captured this predicted interaction, but in *P. tomentosa* instead of *P. trichocarpa* (Du *et al.*, 2015; Tian *et al.*, 2016). Thus, FBA is a useful tool for predicting epistasis among metabolic genes and can be used to provide focused targets for experimental investigations.

DISCUSSION

The results presented here demonstrate the ability of SNPeffect to assign putative roles to SNPs in coding regions of the genome as activating or inhibiting. A follow-up epistasis analysis provided additional information by assessing how these SNPs interact when considered in pairs. We used *Arabidopsis* and poplar datasets as test cases. *Arabidopsis* has the best gene annotations, largest repository of genome sequences collected from different accessions, and numerous studies collecting large-scale phenotypic data. Similarly, *Populus trichocarpa* (*Arabidopsis* for forestry) is the first woody plant to have its genome sequenced (Tuskan *et al.*, 2006), can be propagated vegetatively (making it easy to study mapping populations), has one of the largest re-sequenced populations in plants (Chhetri *et al.*, 2019), and has functional CRISPR/Cas genome editing systems (Novaes *et al.*, 2010).

SNPeffect minimizes flux value deviations from mass-action predicted reaction kinetics to find the most parsimonious explanation of phenotypic variations as a function of SNPs. It integrates the biochemistry encoded in genome-scale metabolic models with genotypic and phenotypic data to identify SNPs that are putatively causal to a given phenotype, while also identifying the underlying mechanisms behind the perturbed metabolic processes. For example, we find that SNPs that increase flux through pathways of amino acid metabolism also contribute positively to growth in both *Arabidopsis* and *Populus*. Faster growing *Arabidopsis* genotypes were predicted to employ the more energy-efficient purine salvage pathway for AMP/GMP production over *de novo* synthesis. Indeed, a similar mechanism has been observed in potato tubers previously, when downregulating the UMP synthase enzyme (present in *de novo* pyrimidine synthesis), the uridine salvage pathway was activated and tuber growth was increased

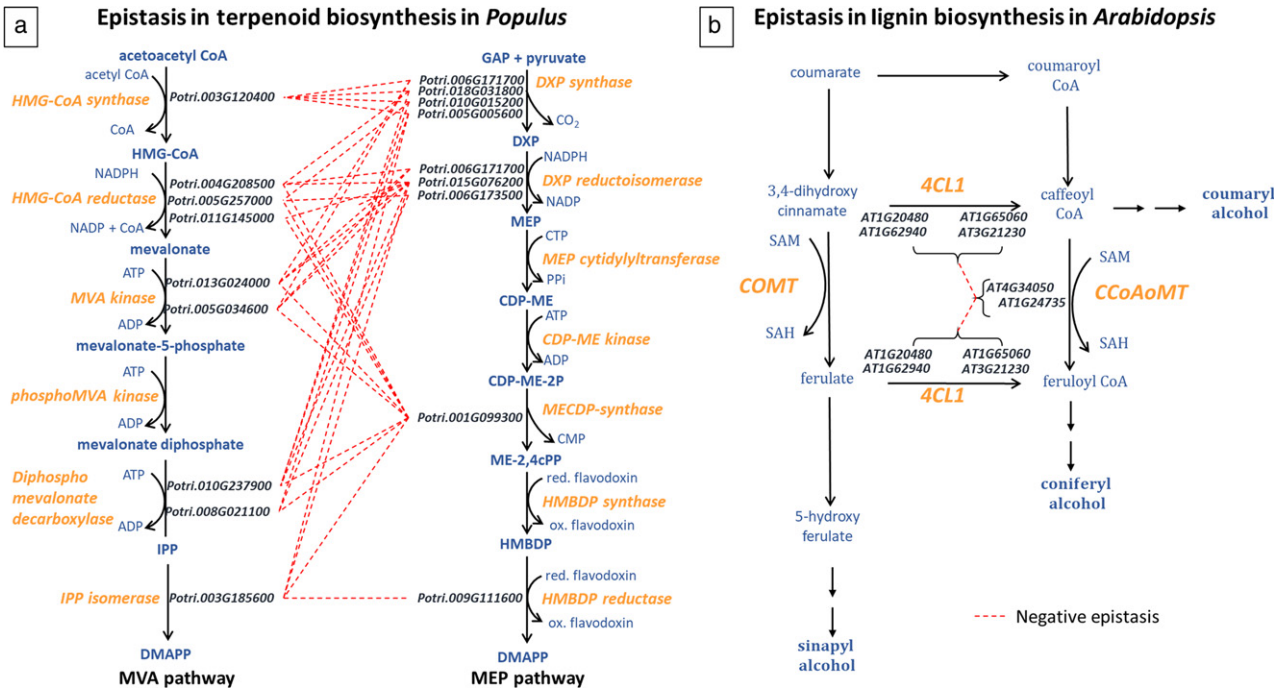

Figure 7. Interpathway and intrapathway epistatic interactions in *Populus* and Arabidopsis.
(a) Negative epistatic interactions between genes (red dotted lines) identified to have functional SNPs in the mevalonate (MVA) and methylerythritol 4-phosphate (MEP) pathways of terpenoid biosynthesis in *P. trichocarpa*. (b) Intrapathway epistatic interactions between genes identified to have functional SNPs in the lignin biosynthesis pathway of *A. thaliana*. 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase; CDP-ME, CDP-methyl-D-erythritol; CDP-ME-2P, CDP-methyl-D-erythritol 2-phosphate; DXP, deoxyxylulose-5-phosphate; DXP reductoisomerase, deoxyxylulose-5-phosphate reductoisomerase; HMG-CoA: 3-hydroxy-3-methylglutaryl-CoA; HMG-CoA synthase: hydroxymethylglutaryl-CoA synthase; IPP, isopentenyl diphosphate; ME-2, 4cP, 2-C-methyl-D-erythritol-2,4-cyclodiphosphate; MECDP synthase, HMBDP, (e)-4-hydroxy-3-methylbut-2-en-1-yl diphosphate; MVA kinase, mevalonate kinase; phosphoMVA kinase, phosphomevalonate kinase. Metabolite names are shown in blue, gene IDs in black and corresponding enzymes in orange.

(Geigenberger *et al.*, 2005). The UMP biosynthetic pathway also emerged as a potential target in *Populus*, in which SNPs in genes encoding for dihydroorotase and dUTPase were found to have functional SNPs. Some putative growth-determining SNPs were also predicted for both plant species in core metabolic processes such as folate, shikimate, and pyrimidine metabolism. These SNPs and combinations thereof form the basis of SNPeffect guided strategies for improved growth yield.

The combined effect of multiple genetic mutations on the overall phenotype was further assessed using epistasis. By examining the distribution of epistatic interactions between functional SNPs, negative epistatic pairs can be avoided when designing breeding experiments that combine genotypic traits from multiple accessions (Segrè *et al.*, 2005). There are several experimental approaches for measuring epistatic effects (Koornneef *et al.*, 1994; Michaels, 1999), the prevalent one in plant systems being quantitative trait locus (QTL) mapping (Paterson *et al.*, 1988; Jansen and Stam, 1994; Zeng, 1994). Although it has helped elucidate functional relationships between many genes, such studies have rarely considered epistasis at the resolution of individual genes on a genome-wide scale. Even though studies in model organisms have revealed significant epistasis (Clark and Wang, 1997; Elena and Lenski, 1997), this is because mapping epistatic interactions is time and resource intensive. Large population sizes are required to sample the landscape of possible genetic interactions and detect significant interactions. Even in genetically tractable systems such as *S. cerevisiae*, a genome-wide interaction screen would entail testing *c.* 18 million pairwise interactions. For *P. trichocarpa* which has *c.* 40 000 genes and a much longer lifespan, such a scan would be prohibitive. In the present work, we explored epistatic effects between SNPeffect-identified SNPs using FBA, which offers a complementary path for predicting a subset of such epistatic interactions occurring within metabolic genes. It has been used previously to investigate the fitness consequence of single-deletion mutations (Ibarra *et al.*, 2002; Papp *et al.*, 2004) and elucidate epistatic relationships between metabolic genes, reactions, and pathways (Chowdhury *et al.*, 2015; Deutscher *et al.*, 2006; Harrison *et al.*, 2007; He *et al.*, 2010; Segrè *et al.*, 2005; Xu *et al.*, 2012).

It is important to emphasize that SNPeffect accounts only for SNPs in the enzyme-coding (or promoter) regions. However, a majority (*c.* 88%; MacArthur *et al.*, 2016) of GWAS-identified SNPs lie in intergenic or intronic regions

and possibly influence gene regulation. Regulatory interactions such as enzyme activation and feedback inhibition were not included in the current version of SNPeffect. Extensions of FBA (O'Brien *et al.*, 2015) incorporating aspects of gene regulation offer a way forward for incorporating SNPs even in non-coding regions within SNPeffect in future. Ultimately, SNPeffect provides a complementary avenue of analysis to GWAS as it can: (i) provide a mechanistic interpretation of the deciphered genotype-to-phenotype relations; (ii) is immune to the confounding effects caused by population structure and multiple-testing, (iii) can handle both monogenic and polygenic traits; (iv) resultant SNP hits can be used as priors in genomic selection; and (v) SNPeffect meta-analysis can evaluate and explain GWAS hits.

EXPERIMENTAL PROCEDURES

GENOME-SCALE METABOLIC MODEL RECONSTRUCTION

In SNPeffect, we make use of a genome-scale metabolic (GSM) model to encode the connection between genes, enzymes, and metabolites participating as substrates or products. GSMs contain the complete list (known up to the present) of the chemical repertoire of an organism, the set of gene(s) needed to be expressed to generate a functional enzyme for a given reaction, and the complete stoichiometry of all such reactions (Varma and Palsson, 1994; Orth *et al.*, 2010). The process of genome-scale model reconstruction naturally lends itself to an iterative approach accompanied by several rounds of model refinement and curation. Multiple plant metabolic models already exist for species such as Arabidopsis (de Oliveira Dal'Molin *et al.*, 2010), maize (Saha *et al.*, 2011; Simons *et al.*, 2014), barley (Grafahrend-Belau *et al.*, 2009), and rice (Lakshmanan *et al.*, 2013).

Arabidopsis thaliana

For *Arabidopsis*, we adopted the previously published and extensively used leaf-specific metabolic model AraGEM (Gomes de Oliveira Dal'Molin *et al.*, 2015; de Oliveira Dal'Molin *et al.*, 2010) and altered 50 reaction directions so as to remove thermodynamically infeasible cycles (Table S4). AraGEM captured more metabolites present in the metabolomics dataset than the more recently published model (Beckers *et al.*, 2016) thereby enabling us to better constrain the feasible solution space using reaction rate expressions derived from mass-action kinetics. The *Arabidopsis* GSM constructed by (Mintz-Oron *et al.*, 2012) had a greater number of reactions participating in thermodynamically infeasible cycles.

Populus trichocarpa

Unlike *Arabidopsis* there is currently no metabolic model for poplar. We used a workflow extensively used by our group (Simons *et al.*, 2014) during the reconstruction of the metabolic model for *P. trichocarpa* *i*Pop7188. A draft model was first created using the PoplarCyc database consisting of 2502 metabolites participating in 3282 reactions catalyzed by enzymes coded by 7188 genes. This draft model was curated so as to remove metabolites and associated reactions with ambiguous atomic compositions (i.e. generic reactants such as 'carboxylates' and those having an R group in their chemical formula such as 'aldose'). Every reaction was subsequently elementally balanced (Chan *et al.*, 2017) and network gaps filled in to enable biomass production (Satish Kumar *et al.*, 2007) (see Table S5 for details on the biomass equation reconstruction).

The draft model directly constructed from PoplarCyc did not contain compartmental annotations for reactions and their participating metabolites. *P. trichocarpa*-specific enzyme localizations were downloaded from BRENDA and UniProt databases for approximately half (*c.* 48%) of the reactions. Compartmentalization information was also called from *Arabidopsis* (SUBA database), with gene homologs being established using a bidirectional protein–protein BLAST. The highest priority was given to *Populus trichocarpa*-specific annotations, followed by *Populus*-specific and then *Arabidopsis* annotations. Compartment-designation for the remaining reactions was carried out computationally using a metabolic network-based procedure (S. Mintz-Oron *et al.* 2009). The method predicts localizations while maximizing flux through known reactions and parsimoniously adding cross-membrane metabolite transporters. The procedure takes as input the *a priori* localization for a subset of reactions and the set of non-localized reactions is duplicated in every subcellular compartment. Next, cross-membrane transporter reactions are added for every metabolite to enable metabolite exchange. Binary variables are assigned to every localized reaction to count the number of reactions that carry a non-zero flux. Then, a mixed-integer linear programming (MILP)-based optimization procedure is implemented with a tilted objective that maximizes the number of flux-carrying localized reactions while minimizing the total flux through all added transporters. A flux-variability analysis is finally carried out to determine the flux ranges of every non-localized reaction in every compartment. A reaction is then allocated to the compartment where it is found to carry the maximum flux.

Gene–protein reaction (GPR) relationships were established using the latest annotated *P. trichocarpa* genome (version 3.1). For reactions that are found in multiple compartments, as attempts of predicting subcellular compartmentalization using computational tools such as WoLFPSORT have been previously unsuccessful due to ambiguous and inconsistent assignments (Dal'Molin *et al.*, 2010), we followed the conservative approach taken by previous reconstruction efforts (Dal'Molin *et al.*, 2010; Gomes de Oliveira Dal'Molin *et al.* 2015; Saha *et al.*, 2011). All genes coding for the catalyzing enzyme(s) were assigned to all subcellular organelles that the reaction features in.

Constraint-based modeling such as FBA poses restrictions on possible metabolic flux distributions by imposing mass balance imperatives for all metabolite species in the system. The flux for biomass precursors is largely determined by knowing the total biomass production flux and the ratio of precursors to form one unit of biomass. However, the synthesis of biomass precursors only consumes a fraction of the cell's total energy budget.

Masakapalli *et al.* (Masakapalli *et al.*, 2010) calculated that only 10–13% of the total ATP produced in heterotrophic *Arabidopsis* cells was being used for biomass production with the remaining used for processes such as transporting metabolites between compartments, substrate uptake, and maintaining transmembrane ion and electrical gradients. Hence, other energy costs such as metabolite transport and maintenance costs must be accounted for in order to accurately estimate subcellular fluxes (Cheung *et al.*, 2013). Maintenance cost for the current model was calculated by introducing an ATP drain in the form of an ATPase into the model (reaction ID 'ATPM') similar to Poolman *et al.* (2009). Flux through this ATP demand reaction was then varied to determine the level at which the glucose uptake rate matched the experimentally observed value recorded by Zhang *et al.* (2018b) using metabolic flux analysis in hybrid poplar. The ATP maintenance demand thus estimated was 11.99 mmol gDW^{-1} h^{-1}, which is within the range of values used before in plant metabolic models (Poolman *et al.*, 2009; Cheung *et al.*, 2013; Yuan *et al.*, 2016). The poplar metabolic model (iPop7188) is included in SBML format (Data S1).

Identifying causal SNPs using SNPeffect

SNPeffect uses as a scaffold the plant metabolic network on which heterogeneous proteomic, metabolomic, and phenotypic datasets are superimposed. Let $\boldsymbol{I}$ and $\boldsymbol{J}$ be the sets of metabolites and reactions present in the plant metabolic model, respectively. The network stoichiometry is captured using a stoichiometric matrix S_{ij}, each entry of which denotes the stoichiometric coefficient of metabolite i in reaction j. Thus, rows in the matrix correspond to metabolites and columns to reactions present in the model. Flux through a reaction j is denoted by v_j, which is constrained to be between an upper and lower bound (determined by thermodynamics) (Equation 1):

$$v_j^{LB} \leq v_j \leq v_j^{UB}, \quad \forall j \in \boldsymbol{J} \tag{1}$$

We implemented the pseudosteady state condition associated with FBA (Orth *et al.*, 2010), where the flux through all reactions producing a metabolite is equated to the flux through all reactions that catabolize it (Equation 2):

$$\sum_j S_{ij} v_j = 0, \quad \forall i \in \boldsymbol{I} \tag{2}$$

This can be viewed as averaging over the 24-h diurnal cycle. Although metabolite concentrations change during plant growth, a study comparing metabolic flux analysis (MFA)-obtained flux distributions across different conditions and tissue types found that the resultant fluxes varied more between different tissues than when the same tissue was subjected to different environments (Sweetlove *et al.*, 2013). This indicates that metabolic fluxes are driven primarily by the demands placed on the system (such as biomass composition and ATP yield), and since it is the fluxes that govern the biological activity of a cell, Equation (2) is a reasonable assumption to make in order to calculate aggregated fluxes.

Let $\boldsymbol{L}$ denote the set of genotypes in the study. Each genotype is represented in SNPeffect by its own metabolic model, bearing the same network stoichiometry S_{ij} but a distinct flux distribution ($v_{jl}, \forall j \in \boldsymbol{J}, \forall l \in \boldsymbol{L}$). A genotype is chosen as the fixed reference ('*ref*') for the population and every other genotype is evaluated with respect to it (such as while determining SNPs, relative metabolite levels, and RGRs). Columbia-0 was chosen as the reference genotype for *Arabidopsis* and Nisqually-0 for poplar in the present study.

Phenotypic constraints are incorporated in SNPeffect from metabolomics and growth data. The RGR of genotype l (RGR_l) (with respect to the reference genotype) is used to set (in relative proportion) the flux through the biomass reaction for each genotype (Equation 3), i.e.:

$$v_{biomass,l} = RGR_l v_{biomass,ref}, \forall l \in \boldsymbol{L} \tag{3}$$

where $v_{biomass,ref}$ is the maximum biomass production flux. Metabolite levels obtained from metabolomics data are incorporated using mass-action kinetics (Sajitz-Hermstein *et al.*, 2016), where the flux v_{jl} through an irreversible reaction j in a genotype l is expressed as (Equation 4):

$$v_{jl} = k_{jl} E_{jl} \prod_{i|S_{ij}<0} (C_{il})^{|S_{ij}|} \tag{4}$$

Here k_{jl} is the rate constant for the reaction j in genotype l, E_{jl} is the abundance of the catalyzing enzyme j in genotype l, and S_{ij} is the stoichiometric coefficient of metabolite i in reaction j. C_{il} is a parameter representing the amount of metabolite i in genotype l whose value is obtained from genotype-specific metabolomics data.

Equation (3) expressed with respect to the reference genotype is (Equation 5):

$$v_{jl} = v_{j,ref} \frac{k_{jl} E_{jl}}{k_{j,ref} E_{j,ref}} \prod_{i|S_{ij}<0} \left(C_{il}^{rel}\right)^{|S_{ij}|} = v_{j,ref} \frac{V_{jl}^{max}}{V_{j,ref}^{max}} \prod_{i|S_{ij}<0} \left(C_{il}^{rel}\right)^{|S_{ij}|} \tag{5}$$

where V_{jl}^{max} is the maximum rate of reaction j, and C_{il}^{rel} is a parameter representing the amount of metabolite i in genotype l (relative to the reference genotype) whose value is obtained from genotype-specific metabolomics data. Similar to (Sajitz-Hermstein *et al.*, 2016), the largest and smallest metabolite ratios (over all measured metabolites for a given genotype) were used for metabolites not present in the metabolomics dataset.

Implementing parsimonious flux balance followed by flux-variability analysis, we obtain reaction bounds for a reaction j in the reference genotype ($LB_{j,ref}, UB_{j,ref}$) (hereby referred to as the 'reference flux distribution'). We can then replace $v_{j,ref}$ in (4) by ($LB_{j,ref}, UB_{j,ref}$) and thus constrain the flux through a reaction j in a genotype l as (Equation 6):

$$LB_{j,ref} \left(\frac{V_{jl}^{max}}{V_{j,ref}^{max}}\right) \prod_{i|S_{ij}<0} \left(C_{il}^{rel}\right)^{|S_{ij}|} \leq v_{jl} \leq UB_{j,ref} \left(\frac{V_{jl}^{max}}{V_{j,ref}^{max}}\right) \prod_{i|S_{ij}<0} \left(C_{il}^{rel}\right)^{|S_{ij}|} \tag{6}$$

Several studies have demonstrated that enzymes in plant central metabolism are not substrate saturated (i.e. substrate concentrations are lower than the enzyme's K_m) (Harris and Königer, 1997; Mettler *et al.*, 2014), to which our approach is readily applicable. However, this may not be the case for all enzymes considered in a large-scale model (Sajitz-Hermstein *et al.*, 2016). In this case, the imposed metabolite ratio constraints are expected to lead to infeasibilities, especially for reactions catalyzed by enzymes without any SNPs in the corresponding genes. Thus, for a reaction with no SNPs in the corresponding genes, the reactions bounds are constrained as (Equation 7):

$$LB_{j,ref}\prod_{i|S_{ij}<0}\left(C_{il}^{rel}\right)^{|S_{ij}|}\le v_{jl}+\boldsymbol{dev}_{jl}^{\boldsymbol{neg}}-\boldsymbol{dev}_{jl}^{\boldsymbol{pos}}\le UB_{j,ref}\prod_{i|S_{ij}<0}\left(C_{il}^{rel}\right)^{|S_{ij}|},\quad \forall j\in\boldsymbol{J}\&\forall l\in\boldsymbol{L} \tag{7}$$

where $\boldsymbol{dev}_{jl}^{\boldsymbol{pos}}$, and $\boldsymbol{dev}_{jl}^{\boldsymbol{neg}}$ represent deviations from mass-action kinetics (Sajitz-Hermstein *et al.*, 2016).

SNPeffect casts the effect of a SNP as affecting reaction rate by changing the product of an enzyme's rate constant and its abundance (i.e. V_{max}) with respect to the reference genotype. The presence of a SNP can either over-regulate or under-regulate enzyme V_{max} by either increasing or decreasing its activity (k) and/or abundance (E), respectively. This effect is ultimately propagated at the metabolic level by increasing or decreasing the corresponding reaction flux. Thus, for a reaction which has SNPs in the corresponding genes, the reaction bounds are expressed as (Equation 8):

$$LB_{j,ref}\prod_{i|S_{ij}<0}\left(C_{il}^{rel}\right)^{|S_{ij}|}\le v_{jl}+\boldsymbol{SNPdev}_{jl}^{\boldsymbol{neg}}-\boldsymbol{SNPdev}_{jl}^{\boldsymbol{pos}}+\boldsymbol{dev}_{jl}^{\boldsymbol{neg}}-\boldsymbol{dev}_{jl}^{\boldsymbol{pos}}\le UB_{j,ref}\prod_{i|S_{ij}<0}\left(C_{il}^{rel}\right)^{|S_{ij}|},\quad \forall j\in\boldsymbol{J}\&\forall l\in\boldsymbol{L} \tag{8}$$

where $\boldsymbol{SNPdev}_{jl}^{\boldsymbol{pos}}$ and $\boldsymbol{SNPdev}_{jl}^{\boldsymbol{neg}}$ represent the positive and negative deviation from the reference flux distribution due to the presence of SNPs, respectively. These slacks allow the flux through a reaction to increase (via $\boldsymbol{SNPdev}_{jl}^{\boldsymbol{pos}}$ when the reaction upper bound $UB_{j,ref}\prod_{i|S_{ij}<0}\left(C_{il}^{rel}\right)^{|S_{ij}|}$ is limiting) or decrease (via $\boldsymbol{SNPdev}_{jl}^{\boldsymbol{neg}}$ when the reaction lower bound $LB_{j,ref}\prod_{i|S_{ij}<0}\left(C_{il}^{rel}\right)^{|S_{ij}|}$ is limiting) in order to satisfy all imposed phenotypic data. Thus, (7) is the limiting case of (8) when there are no functional SNPs in the genes for reaction j and the ratio $\left(\frac{V_{jl}^{max}}{V_{j,ref}^{max}}\right)$ from (5) equals one. However, if there is a functional SNP in the corresponding genes, its effect on $\left(\frac{V_{jl}^{max}}{V_{j,ref}^{max}}\right)$ can be captured by the slack variables $\boldsymbol{SNPdev}_{jl}^{\boldsymbol{pos}}$ and $\boldsymbol{SNPdev}_{jl}^{\boldsymbol{neg}}$. These updated reaction bounds translate downstream into increased (or reduced) biomass synthesis, thereby presenting a genotype as having a growth benefit (or penalty) with respect to the reference.

Next, we decompose reaction slacks $\boldsymbol{SNPdev}_{jl}^{\boldsymbol{pos}}$ and $\boldsymbol{SNPdev}_{jl}^{\boldsymbol{neg}}$ in terms of the contributions from the individual SNPs present in the gene(s) encoding for the enzyme catalyzing reaction j. This is done using the GPR relationships for reaction j by defining a matrix A_{jkl} with entries such that (Equation 9):

$$A_{jkl}=\begin{cases}1, & \textit{if reaction j is associated with SNP k in genotype l}\\ 0, & \textit{otherwise}\end{cases} \tag{9}$$

SNPeffect does not account for any structural protein information and can thus only comment on the final effect of a non-synonymous SNP on reaction flux. Thus, we assume that reaction flux is increased (or decreased) proportionally to the activity of SNP(s) found in the corresponding genes. For isozymes, this entails an increase (or decrease) in the net enzymatic activity (Wang *et al.*, 2006). For enzymes encoded by protein subunits, nsSNPs can both increase enzymatic activity (such as by stabilizing the catalytic conformation (Clifton *et al.*, 2018)) or decrease it (such as by destabilizing protein–protein interactions by disrupting existing salt bridges (Richard *et al.*, 2003) or introducing charged residues that disrupt hydrophobic interactions (Colombo *et al.*, 1994)). We model the final impact of SNPs on reaction flux by combining effects of multiple SNPs additively, and $\boldsymbol{SNPdev}_{jl}^{\boldsymbol{pos}}$ $\boldsymbol{SNPdev}_{jl}^{\boldsymbol{neg}}$ are thus written as (Equation 10):

$$\boldsymbol{SNPdev}_{jl}^{\boldsymbol{pos}}=\sum_{K}A_{jkl}\boldsymbol{X}_{kl}^{\boldsymbol{pos}},\quad \boldsymbol{SNPdev}_{jl}^{\boldsymbol{neg}}=\sum_{K}A_{jkl}\boldsymbol{X}_{kl}^{\boldsymbol{neg}} \tag{10}$$

where $\boldsymbol{K}$ is the set of all SNPs and $\boldsymbol{X}_{kl}^{\boldsymbol{pos}}$ (or $\boldsymbol{X}_{kl}^{\boldsymbol{neg}}$) is the contribution of SNP k in genotype l toward increasing (or decreasing) flux through reaction j.

We further require that the effect of a particular SNP across all genotypes (i.e. same amino acid change at the same genomic position) be consistent (i.e. either enhances or suppresses enzyme activity). For this, we employ binary variables y_k^{pos} and y_k^{neg} such that (Equations 11–13):

$$y_k^{pos}=\begin{cases}1, & \textit{if SNP k is activity}-\textit{enhancing}\\ 0, & \textit{otherwise}\end{cases} \tag{11}$$

and

$$y_k^{neg}=\begin{cases}1, & \textit{if SNP k is activity}-\textit{suppressing}\\ 0, & \textit{otherwise}\end{cases} \tag{12}$$

and a matrix B_{kl} to map SNPs to genotypes such that

$$B_{kl}=\begin{cases}1, & \textit{if SNP k is present in genotypel}\\ 0, & \textit{otherwise}\end{cases} \tag{13}$$

The following constraints are implemented to ensure consistent SNP consequence across genotypes (Equations 14–16):

$$0\le\boldsymbol{X}_{kl}^{\boldsymbol{pos}}\le My_k^{pos}B_{kl},\quad \forall k\in\boldsymbol{K}\&l\in\boldsymbol{L}, \tag{14}$$

$$0\le\boldsymbol{X}_{kl}^{\boldsymbol{neg}}\le My_k^{neg}B_{kl},\quad \forall k\in\boldsymbol{K}\&l\in\boldsymbol{L}, \tag{15}$$

$$y_k^{pos}+y_k^{neg}\le 1,\quad \forall k\in\boldsymbol{K}, \tag{16}$$

Here, M is a scalar large enough so as not to artificially constrain any possible value of $\boldsymbol{X}_{kl}^{\boldsymbol{pos}}$ or $\boldsymbol{X}_{kl}^{\boldsymbol{neg}}$ (taken to be 100 in the current study). The final constraint (16) ensures that a SNP k either increases or decreases catalytic activity.

In order to obtain a parsimonious description of the phenotypic differences seen between genotypes as a function of the SNPs present in the metabolic genes, we formulate and solve the following mixed-integer minimization problem (Equation 17):

$$\text{minimize} \sum_l \sum_j (SNPdev_{jl}^{pos} + SNPdev_{jl}^{neg}) + \omega \sum_l \sum_j \left(dev_{jl}^{pos} + dev_{jl}^{neg} \right) \quad (17)$$

such that (Equations 18–26):

$$\sum_j S_{ij} v_{jl} = 0, \quad \forall i \in \boldsymbol{I} \& \forall l \in \boldsymbol{L} \quad (18)$$

$$LB_{j,ref} \prod_{i|S_{ij}<0} \left(C_{il}^{rel}\right)^{|S_{ij}|} \leq v_{jl} + SNPdev_{jl}^{neg} - SNPdev_{jl}^{pos} + dev_{jl}^{neg} - dev_{jl}^{pos} \leq UB_{j,ref} \prod_{i|S_{ij}<0} \left(C_{il}^{rel}\right)^{|S_{ij}|}, \quad \forall j \in \boldsymbol{J} \& \forall l \in \boldsymbol{L} \quad (19)$$

$$v_{biomass,l} = (\mathrm{RGR}_l) v_{biomass,ref}, \quad \forall l \in \boldsymbol{L} \quad (20)$$

$$SNPdev_{jl}^{pos} = \sum_K A_{jkl} X_{kl}^{pos}, \quad \forall j \in \boldsymbol{J} \& \forall l \in \boldsymbol{L} \quad (21)$$

$$SNPdev_{jl}^{neg} = \sum_K A_{jkl} X_{kl}^{neg}, \quad \forall j \in \boldsymbol{J} \& \forall l \in \boldsymbol{L} \quad (22)$$

$$0 \leq X_{kl}^{pos} \leq MB_{kl} y_k^{pos}, \quad \forall k \in \boldsymbol{K} \& \forall l \in \boldsymbol{L} \quad (23)$$

$$0 \leq X_{kl}^{neg} \leq MB_{kl} y_k^{neg}, \quad \forall k \in \boldsymbol{K} \& \forall l \in \boldsymbol{L} \quad (24)$$

$$y_k^{pos} + y_k^{neg} \leq 1, \quad \forall k \in \boldsymbol{K} \& \forall l \in \boldsymbol{L} \quad (25)$$

$$SNPdev_{jl}^{pos}, SNPdev_{jl}^{neg}, dev_{jl}^{pos}, dev_{jl}^{neg} \geq 0, \quad \forall j \in \boldsymbol{J} \& \forall l \in \boldsymbol{L} \quad (26)$$

Parameter ω is a penalty factor that over-penalizes (100 fold in the present simulations) deviations ($dev_{jl}^{pos}, dev_{jl}^{neg}$) away from the assumed mass-action kinetics (Sajitz-Hermstein *et al.*, 2016) whenever there is no SNP(s) to explain discrepancies. Deviations that can be explained as the outcome of SNP(s) are quantified by variables $SNPdev_{jl}^{pos}$ and $SNPdev_{jl}^{neg}$. The same parsimony criterion is imposed to both $SNPdev_{jl}^{pos}, SNPdev_{jl}^{neg}$ and $dev_{jl}^{pos}, dev_{jl}^{neg}$; however, the penalty for discrepancies in $SNPdev_{jl}^{pos}, SNPdev_{jl}^{neg}$ is 100 times larger.

Solving the above optimization problem provides a parsimonious set of functional SNPs across all genotypes (i.e. non-zero values of X_{kl}^{pos} and X_{kl}^{neg}), flux departures from the reference genotype due to SNPs ($SNPdev_{jl}^{pos}, SNPdev_{jl}^{neg}$), flux departures from mass-action kinetics ($dev_{jl}^{pos}, dev_{jl}^{neg}$), and the flux distribution across all genotypes (v_{jl}). The parameters obtained from data are S_{ij} (stoichiometric coefficient of metabolite *i* in reaction *j*), $LB_{j,ref}$, $UB_{j,ref}$ (reaction bounds for reaction *j* in the reference genotype), RGR_l (Relative Growth Rate with respect to the reference genotype (Columbia-0 for *Arabidopsis* and Nisqually-0 for poplar)), A_{jkl} (sparse matrix used to map SNPs to reactions), B_{kl} (sparse matrix used to map SNPs to genotypes), and C_{il}^{rel} (level of metabolite *i* in genotype *l* with respect to the reference obtained from metabolomics data).

Implementing SNPeffect for *A. thaliana* and *P. trichocarpa*

Arabidopsis genome sequences and non-synonymous SNPs were downloaded from the 1001 Genomes project (Alonso-Blanco *et al.*, 2016) and leaf-specific metabolomics data were taken from Wu *et al.* (2018) for *c.* 54 unique metabolites in every genotype. Growth data (relative leaf area increase) (Atwell *et al.*, 2010) were found for 69 genotypes and used to constrain the flux through the biomass reaction, but additional phenotypic constraints such as photosynthetic CO_2 evolution or nutrient uptake rates can also be added if available. SNPeffect requires SNP and phenotypic data for every genotype, which is why, although SNP data are available for 1001 genotypes (Alonso-Blanco *et al.*, 2016) and metabolomics data for 309 genotypes (Wu *et al.*, 2018), the final number of *Arabidopsis* genotypes in the study is 69 (genotypes with RGR data). In accordance with previous GWAS studies, SNPs with a MAF < 1% were excluded from the study (Scuteri *et al.*, 2007; Hamblin and Jannink, 2011) and SNPs within 10 kbp of a gene were assigned to that gene (due to the possibility of gene activity being affected by the presence of SNPs in enhancer/repressor regions) (Biscarini *et al.*, 2016; Brodie *et al.*, 2016; Lee and Shatkay, 2008; Torkamani *et al.*, 2008). In total, 8269 non-synonymous SNPs were found across 69 genotypes, with *c.* 73% of *Arabidopsis* metabolic genes having at least one non-synonymous SNP. 16.6% of the SNPs are within coding regions and 83.4% within the 10 kbp stretch, with the average gene length being 3276 bp. A genotype-wise distribution of SNPs can be found in Figure S1, gene-wise distribution of unique SNPs (across all genotypes) in Figure S5, and metabolite abundances have been summarized in Figure S3. The leaf-specific GSM AraGEM served as the metabolic network for *Arabidopsis* and was used for all calculations.

Similarly, genome sequences for 882 poplar clones were obtained from the Bioenergy Science Center database (https://bioenergycenter.org/besc/gwas/) and phenotypic data (i.e. crown biomass and glucose and xylose sugar release) for 160 of those were available in Muchero *et al.* (Muchero *et al.*, 2015). The sequence data were processed and SNPs within 10 kbp of all genes in the metabolic model identified using pyVCF (https://github.com/jamescasbon/PyVCF). SNPs with a MAF <1% were excluded from the study (Scuteri *et al.*, 2007; Hamblin and Jannink, 2011). Then, non-synonymous SNPs were extracted using SnpEff (http://snpeff.sourceforge.net/). The poplar input dataset had 24 877 SNPs in 160 genotypes, with *c.* 68% of poplar metabolic genes having at least one non-synonymous SNP. 42.5% of these SNPs are within coding regions and 57.5% within the 10 kbp stretch, with the average gene length being 3107 bp. A genotype-wise distribution of SNPs can be found in Figure S2, gene-wise distribution of unique SNPs (across all genotypes) in Figure S6 and metabolite abundances have been summarized in Figure S4.

For both the organisms, the reference flux distribution was calculated using parsimonious FBA (pFBA) (Lewis *et al.*, 2010) as its predictions have been found to be as good or better than those obtained using methods integrating transcriptomics/proteomics data (Machado and Herrgård, 2014). Briefly, pFBA calculates a flux distribution that minimizes the total flux through all metabolic reactions in the model. The General Algebraic Modeling System (GAMS) (using the Cplex solver) was used to implement SNPeffect and Python 2.7 used to generate all input files. A GAMS

implementation alongside the necessary Python code can be downloaded from the research github repository (https://github.com/maranasgroup) and group webpage (http://www.maranasgroup.com/software.htm). All computations were carried out on dual 10-core and 12-core Intel Xeon E5-2680 and Intel Xeon E7-4830 quad 10-core processors that are part of the ACI cluster of High-Performance Computing Group of Pennsylvania State University. SNPeffect requires the solution of a sequence of mixed-integer linear MILP problems whose size is determined by the number of SNPs included in the study as two binary variables need to be defined for every SNP present in the study. For the *Arabidopsis* dataset (with *c.* 8500 SNPs) SNPeffect took *c.* 30 min to run, and for the poplar dataset (with *c.* 25 000 SNPs) the algorithm took *c.* 70 min.

DATA STATEMENT

All relevant data can be found within the manuscript and its supporting materials.

ACKNOWLEDGEMENTS

Funding provided by the BioEnergy Science (BESC) and the Center for Bioenergy Innovation (CBI). U.S. Department of Energy Bioenergy Research Centers supported by the Office of Biological and Environmental Research in the DOE Office of Science. Support for the Poplar GWAS dataset is provided by the U.S. Department of Energy, Office of Science Biological and Environmental Research (BER) via the BESC under Contract No. DE-PS02-06ER64304. The Poplar GWAS Project used resources of the Oak Ridge Leadership Computing Facility and the Compute and Data Environment for Science at Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. We thank Stephen DiFazio, Lin Wang, Ratul Chowdhury, Patrick Suthers, and Jonathan Cumming for immensely helpful discussions. The authors would like to thank all reviewers for their insightful and detailed comments that have helped greatly to improve the clarity of the manuscript.

AUTHOR CONTRIBUTIONS

CDM designed the research, DS and CDM performed the research, the analysis, and wrote the article. CDM agrees to serve as the author responsible for contact and ensures communication.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. (a) Genotype-wise distribution of non-synonymous SNPs (nsSNP) in the *Arabidopsis* input dataset. (b) Genotype-wise distribution of non-synonymous SNPs (nsSNP) in *Arabidopsis* in the genes identified by SNPeffect to have functional SNPs. SNPs present within a gene are shown in orange while SNPs present in the 10 kbp region are shown in blue.

Figure S2. (a) Genotype-wise distribution of non-synonymous SNPs (nsSNP) in the poplar input dataset. (b) Genotype-wise distribution of non-synonymous SNPs (nsSNP) in poplar in the genes identified by SNPeffect to have functional SNPs. SNPs present within a gene are shown in orange while SNPs present in the 10 kbp region are shown in blue.

Figure S3. Genotype-wise distribution of metabolites in the *Arabidopsis* input dataset (taken from Wu *et al.*, 2018). Metabolite levels are shown on the x-axis while genotype IDs are listed on the y-axis.

Figure S4. Genotype-wise distribution of glucose (shown in blue) and xylose (shown in orange) in the poplar input dataset (taken from Muchero *et al.*, 2015). Metabolite levels are shown on the x-axis while genotype IDs are listed on the y-axis.

Figure S5. Gene-wise distribution of SNPs (across all genotypes) in the *Arabidopsis* input dataset. SNPs present within a gene are shown in orange while SNPs present in the 10 kbp region are shown in blue.

Figure S6. Gene-wise distribution of SNPs (across all genotypes) in the poplar input dataset. SNPs present within a gene are shown in orange while SNPs present in the 10 kbp region are shown in blue.

Table S1. List of all functional SNPs identified in *Arabidopsis thaliana* alongside their activities, and the corresponding gene, the minor allele frequency (with respect to the reference genotype), experimental evidence for growth-related effects and functional characterization using the CDD database (wherever applicable), and the associated metabolic pathway(s).

Table S2. List of all functional SNPs identified in *Populus trichocarpa* alongside their activities, and the corresponding gene, the minor allele frequency (with respect to the reference genotype), experimental evidence for growth-related effects and functional characterization using the CDD database (wherever applicable), and the associated metabolic pathway(s).

Table S3. Summary of the unique epistatic interactions found among genes associated with functional SNPs in *Arabidopsis thaliana*.

Table S4. List of changes made to the AraGEM model so as to remove thermodynamically infeasible cycles.

Table S5. Details of the *i*Pop7188 biomass equation.

Table S6. Summary of the unique epistatic interactions found among genes associated with functional SNPs in *Populus trichocarpa*.

Data S1. Poplar metabolic model (iPop7188) in SBML format.

REFERENCES

Alonso-Blanco, C., Andrade, J., Becker, C. *et al.* (2016) 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*, **166**(2), 481–491.

Atwell, S., Huang, Y.U.S, Bjarni, J. *et al.* (2010) Genome-wide association study of 107 Phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, **465**, 627–631.

Bauer, C.R., Li, S. and Siegal, M.L. (2015) Essential gene disruptions reveal complex relationships between phenotypic robustness, pleiotropy, and fitness. *Mol. Syst. Biol.* **11**, 773.

Beckers, V., Dersch, L., Lotz, K., Melzer, G., Bläsing, O., Fuchs, R., Ehrhardt, T. and Wittmann, C. (2016) *In silico* metabolic network analysis of Arabidopsis leaves. *BMC Syst. Biol.* **10**, 102.

Berg, M., Rogers, R., Muralla, R. and Meinke, D. (2005) Requirement of Aminoacyl-TRNA Synthetases for Gametogenesis and Embryo Development in Arabidopsis. *Plant J.* **44**, 866–878.

BESC, BioEnergy Science Center. (2017) 'GWAS Dataset.' https://bioenergycenter.org/besc/gwas//

Besson, V., Neuuburger, M., Rebeille, F. and Douce, R. (1995) EVIDENCE for 3 serine hydroxymethyltransferases in green leaf-cells - Purification and characterization of the mitochondrial and chloroplastic isoforms. *Plant Physiol. Biochem.* **33**, 665–673.

Bhatnager, R., Dang, A.S. (2018) Comprehensive in-silico prediction of damage associated SNPs in Human Prolidase gene. *Sci. Rep.* **8**, 1–14.

Bick, J.A., Lange, B.M. (2003) Metabolic cross talk between cytosolic and plastidial pathways of isoprenoid biosynthesis: unidirectional transport of intermediates across the chloroplast envelope membrane. *Arch. Biochem. Biophys.* **415**, 146–154.

Biscarini, F., Cozzi, P., Casella, L. *et al.* (2016) Genome-wide association study for traits related to plant and grain morphology, and root architecture in temperate rice accessions. *PLoS ONE*, **11**, e0155425.

Blazquez, M.A., Santos, E., Flores, C.L., Martínez-Zapater, J.M., Salinas, J. and Gancedo, C. (1998) Isolation and molecular characterization of the Arabidopsis TPS1 gene, encoding trehalose-6-phosphate synthase. *Plant J.* **13**. 685–689.

Boone, C., Bussey, H. and Andrews, B.J. (2007) Exploring genetic interactions and networks with yeast. *Nat. Rev. Genet.* **8**, 437–449.

Brodie, A., Azaria, J.R. and Ofran, Y. (2016) How far from the SNP may the causative genes be? *Nucleic Acids Res.* **44**, 6046–6054.

Chan, S.H., Cai, J., Wang, L., Simons-Senftle, M.N. and Maranas, C.D. (2017) Standardizing biomass reactions and ensuring complete mass balance in genome-scale metabolic models. *Bioinformatics* **33**, 3603–3609.

Chanoca, A., de Vries, L. and Boerjan, W. (2019) Lignin engineering in forest trees. *Front. Plant Sci.* **10**, 912.

Chao, D.-Y., Chen, Y., Chen, J. *et al.* (2014) Genome-wide association mapping identifies a new arsenate reductase enzyme critical for limiting arsenic accumulation in plants. *PLoS Biol.* **12**, e1002009

Cheung, C.M., Williams, T.C., Poolman, M.G., Fell, D.A., Ratcliffe, R.G. and Sweetlove, L.J. (2013) A Method for Accounting for maintenance costs in flux balance analysis improves the prediction of plant cell metabolic phenotypes under stress conditions. *Plant J.* **75**, 1050–1051.

Chhetri, H.B., Macaya-Sanz, D., Kainer, D. *et al.* (2019) Multi-trait genome-wide association analysis of populus trichocarpa identifies key polymorphisms controlling morphological and physiological traits. *New Phytol.* **23**, 293–309.

Chowdhury, R, Chowdhury, A. and Maranas, C. (2015) Using gene essentiality and synthetic lethality information to correct yeast and CHO cell genome-scale models. *Metabolites* **5**, 536–570.

Clark, A.G. and Wang, L. (1997) Epistasis in measured genotypes: drosophila P-element insertions. *Genetics.* **47**, 157–163.

Clifton, B.E., Kaczmarski, J.A., Carr, P.D., Gerth, M.L., Tokuriki, N. and Jackson, C.J. (2018) Evolution of Cyclohexadienyl Dehydratase from an Ancestral Solute-Binding Protein Article. *Nat. Chem. Biol.* **14**, 542–547.

Colombo, I., Finocchlaro, G., Garavaglia, B., Garbugllo, N., Yamaguchl, S., Frerman, F., Berra, B. and DiDonato, S. (1994) Mutations and polymorphisms of the gene encoding the β-subunit of the electron transfer flavoprotein in three patients with glutaric acidemia type II. *Hum. Mol. Genet.*, **3**, 429–435.

Cookson, S.J., Chenu, K. and Granier, C. (2007) Day length affects the dynamics of leaf expansion and cellular development in *Arabidopsis thaliana* partially through floral transition timing. *Ann. Bot.* **99**, 703–711.

Corea, O., Ki, C., Cardenas, C., Kim, S., Brewer, S., Patten, A., Davin, L. and Lewis, N. (2012) Arogenate dehydratase isoenzymes profoundly and differentially modulate carbon flux into lignins. *J. Biol. Chem.* **287**, 11446–11459.

Dejardin, A., Sokolov, L.N. and Kleczkowski, L.A. (2015) Sugar/osmoticum levels modulate differential abscisic acid-independent expression of two stress-responsive sucrose synthase genes in Arabidopsis. *Biochem. J.* **344**, 503-509.

Deutscher, D., Meilijson, I., Kupiec, M. and Ruppin, E. (2006) Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat. Genet.* **38**, 993–998.

Du, Q., Tian, J., Yang, X., Pan, W., Xu, B., Li, B., Ingvarsson, P.k and Zhang, D. (2015) Identification of additive, dominant, and epistatic variation conferred by key genes in cellulose biosynthesis pathway in Populus tomentosa. *DNA Res.* **22**, 53–67.

Elena, S.F. and Lenski, R.E. (1997) Test of synergistic interactions among deleterious mutations in bacteria. *Nature*, **390**, 395–398.

Esaki, S., Malkaram, S.A. and Zempleni, J. (2012) Effects of single-nucleotide polymorphisms in the human holocarboxylase synthetase gene on enzyme catalysis. *Eur. J. Hum. Genet.* **20**, 428–433.

Evans, L.M., Slavov, G.T., Rodgers-Melnick, E. *et al.* (2014) Population genomics of Populus trichocarpa identifies signatures of selection and adaptive trait associations. *Nat. Genet.* **46**, 1089–1096.

Evnouchidou, I., Kamal, R.P., Seregin, S.S. *et al.* (2011) Cutting edge: coding single nucleotide polymorphisms of endoplasmic reticulum aminopeptidase 1 can affect antigenic peptide generation in vitro by influencing basic enzymatic properties of the enzyme. *J. Immunol.* **186**, 1909–1913.

Ferrari, R., Lovering, R.C., Hardy, J., Lewis, P.A. and Manzoni, C. (2017) Weighted protein interaction network analysis of frontotemporal dementia. *J. Proteome Res.* **16**, 999–1013.

Forsum, O., Svennerstam, H., Ganeteg, U. and Näsholm, T. (2008) Capacities and constraints of amino acid utilization in Arabidopsis. *New Phytol.* **179**, 1058–1069.

Fung, H.C., Scholz, S., Matarin, M. *et al.* (2006) genome-wide genotyping in Parkinson's disease and neurologically. Normal controls: first stage analysis and public release of data. *Lancet Neurol.* **5**, 911–916.

Galili, G. (2002) New insights into the regulation and functional significance of lysine metabolism in plants. *Annu. Rev. Plant Biol.* **53**, 27–43.

Geigenberger, P., Regierer, B., Nunes-Nesi, A., Leisse, A., Urbanczyk-Wochniak, E., Springer, F., van Dongen, J.T., Kossmann, J. and Fernie, A.R. (2005) Inhibition of de novo pyrimidine synthesis in growing potato tubers leads to a compensatory stimulation of the pyrimidine salvage pathway and a subsequent increase in biosynthetic performance. *Plant Cell*, **17**, 2077–2088.

Gibson, G. (2012) Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135–145.

Goddijn, O.J., Verwoerd, T.C. and Voogd, E. (2002) Inhibition of trehalase activity enhances trehalose accumulation in transgenic plants. *Plant Physiol.* **113**, 181–190.

Gong, C., Du, Q., Xie, J., Quan, M., Chen, B. and Zhang, D. (2018) Dissection of insertion–deletion variants within differentially expressed genes involved in wood formation in populus. *Front. Plant Sci.* **8**, 2199.

Grafahrend-Belau, E., Schreiber, F., Koschützki, D. and Junker, B.H. (2009) Flux balance analysis of barley seeds: a computational approach to study systemic properties of central metabolism. *Plant Physiol.* **149**: 585–598.

Guinot, F., Szafranski, M., Ambroise, C. and Samson, F. (2018) Learning the optimal scale for GWAS through hierarchical SNP aggregation. *BMC Bioinformatics*, **19**, 459.

Guo, Q., Yoshida, Y., Major, I.T. *et al.* (2018) JAZ Repressors of Metabolic Defense Promote Growth and Reproductive Fitness in Arabidopsis. *Proc. Nati. Acad. Sci.* **115**, E10768–E10777.

de la Luz Gutiérrez-Nava, M. (2004) CHLOROPLAST BIOGENESIS Genes Act Cell and Noncell Autonomously in Early Chloroplast Development. PLANT PHYSIOLOGY.

Hamblin, M.T. and Jannink, J.-L. (2011) Factors affecting the power of haplotype markers in association studies. *Plant Genome J.* **4**, 145–153.

Harris, G.C. and Köniğer, M. (1997) The 'high' concentrations of enzymes within the chloroplast. *Photosynth. Res.* **54**, 5–23.

Harrison, R., Papp, B., Pal, C., Oliver, S.G. and Delneri, D. (2007) Plasticity of genetic interactions in metabolic networks of yeast. *Proc. Natl Acad. Sci.* **104**, 2307–2312.

He, X., Qian, W., Wang, Z., Li, Y. and Zhang, J. (2010) Prevalent positive epistasis in Escherichia coli and Saccharomyces cerevisiae metabolic networks. *Nature Genet.* **42**, 272–276.

He, Y., Cheng, J., He, Y., Yang, B., Cheng, Y., Yang, C., Zhang, H. and Wang, Z. (2019) Influence of isopropylmalate synthase OsIPMS1 on seed vigour associated with amino acid and energy metabolism in rice. *Plant Biotechnol. J.* **17**, 322–337.

Hedstrom, L. (2009) IMP dehydrogenase: structure, mechanism, and inhibition. *Chem. Rev.* **109**, 2903–2928.

Hemmerlin, A. and Bach, T.J. (1998) Effects of mevinolin on cell cycle progression and viability of tobacco BY-2 Cells. *Plant J.* **14**, 65–74.

Hemmerlin, A., Hoeffler, J.-F., Meyer, O., Tritsch, D., Kagan, I.A., Grosdemange-Billiard, C., Rohmer, M. and Bach, T.J. (2003) Cross-talk between the cytosolic mevalonate and the plastidial methylerythritol phosphate pathways in tobacco bright yellow-2 cells. *J Biol Chemist*, **278**, 26666–26676.

Ibarra, R.U., Edwards, J.S. and Palsson, B.O. (2002) Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature*, **420**, 186–189.

Jansen, R.C. and Stam, P. (1994) High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, **136**, 1447–1455.

Kelley, R. and Ideker, T. (2005) Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.* **23**, 561–566.

Koehler, L and Telewski, F.W. (2006) Biomechanics and transgenic wood. *Am. J. Bot.* **93**, 1433–1438.

Koornneef, M., Blankestijn-de Vries, H., Hanhart, C., Soppe, W. and Peeters, T. (1994) The phenotype of some late-flowering mutants is enhanced by a locus on chromosome 5 that is not effective in the Landsberg erecta wild-type. *Plant J.* **6**, 911–919.

Lakshmanan, M., Zhang, Z., Mohanty, B. *et al.* (2013) Elucidating rice cell metabolism under flooding and drought stresses using flux-based modeling and analysis. *Plant Physiol.* **162**, 2140–2150.

. **Lango, H.A., Estrada, K., Lettre, G. *et al.*** (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, **467**, 832–838.

Lee, P.H. and Shatkay, H. (2007) F-SNP: computationally predicted functional SNPs for disease association studies. *Nucleic Acids Res.* **36**, D820–D824.

Lee, I., Blom, U.M., Wang, P.I., Shim, J.E. and Marcotte, E.M. (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.* **21**, 1109–1121.

Leiserson, M.D.M., Eldridge, J.V., Ramachandran, S. and Raphael, B.J. (2013) Network analysis of GWAS data. *Curr. Opin. Genet. Dev.* **23**, 602–610.

Lewis, N.E., Hixson, K.K., Conrad, T.M. *et al.* (2010) Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Mol. Systems Biol.* **6**, 390.

Li, Y., Huang, Y., Bergelson, J., Nordborg, M. and Borevitz, J.O. (2010) Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. *Proc Nati Acad Sci.* **107**, 21199–21204.

Liechti, G. and Goldberg, J.B. (2012) Helicobacter pylori relies primarily on the purine salvage pathway for purine nucleotide biosynthesis. *J. Bacteriol.* **194**, 839–854.

Liu, Y.-L., Chiang, Y.-H., Liu, G.-Y. and Hung, H.-C. (2011) Functional role of dimerization of human peptidylarginine deiminase 4 (PAD4). *PLoS ONE*, **6**, e21314.

Liu, Y., Brossard, M., Sarnowski, C. *et al.* (2017) Network-assisted analysis of GWAS data identifies a functionally-relevant gene module for childhood-onset asthma. *Sci Rep.* **7**, 1–10.

Lloyd, J. and Meinke, D. (2012) A comprehensive dataset of genes with a loss-of-function mutant phenotype in Arabidopsis. *Plant Physiol.* **158**, 1115–1129.

Locke, A.E., Kahali, B., Berndt, S.I. *et al.* (2015) Genetic studies of body mass index yield new insights for obesity biology. *Nature*. **18**, 197–206.

MacArthur, J., Bowler, E., Cerezo, M., *et al.* (2016) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901.

Machado, D. and Herrgård, M. (2014) Systematic Evaluation of Methods for Integration of Transcriptomic Data into Constraint-Based Models of Metabolism' ed. Costas D. Maranas. *PLoS Computat. Biol.* **10**, e1003580.

Maloof, J.N. (2003) QTL for plant growth and morphology. *Curr. Opin. Plant Biol.* **6**, 85–90.

Maraganore, D.M., de Andrade, M., Lesnick, T.G. *et al.* (2005) High-resolution whole-genome association study of Parkinson disease. *Am. J. Human Genet.* **77**, 685–693.

Marchler-Bauer, A., Bo, Y., Han, L. *et al.* (2017) CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45**, D200–D203.

Marino, M.J., Valenti, O. and Conn, P.J. (2003) Glutamate receptors and Parkinson's Disease. *Drugs Aging*, **20**, 377–397.

Masakapalli, S.K., Le Lay, P., Huddleston, J.E., Pollock, N.L., Kruger, N.J. and Ratcliffe, R.G. (2010) Subcellular flux analysis of central metabolism in a heterotrophic Arabidopsis cell suspension using steady-state stable isotope labeling. *Plant Physiol.* **152**, 602–619.

Mettler, T., Mühlhaus, T., Hemme, D. *et al.* (2014) Systems analysis of the response of photosynthesis, metabolism, and growth to an increase in irradiance in the photosynthetic model organism chlamydomonas reinhardtii. *Plant Cell*, **26**, 2310–2350.

Michaels, S.D. (1999) FLOWERING LOCUS C Encodes a Novel MADS Domain Protein That Acts as a Repressor of Flowering. *Plant Cell*, **11**, 949.

Mintz-Oron, S., Meir, S., Malitsky, S., Ruppin, E., Aharoni, A. and Shlomi, T. (2012) Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity. *Proc. Nati. Acad. Sci. USA*, **109**, 339–344.

Mintz-Oron, S., Aharoni, A., Ruppin, E. and Shlomi, T. (2009) Network-Based Prediction of Metabolic Enzymes' Subcellular Localization. *Bioinformatics* **25**, i247–1252.

Moffatt, B.A. and Ashihara, H. (2002) Purine and pyrimidine nucleotide synthesis and metabolism. *Arabidopsis Book*, **1**, e0018.

Muchero, W., Guo, J., DiFazio, S.P. *et al.* (2015) High-resolution genetic mapping of allelic variants associated with cell wall chemistry in Populus. *BMC Genom.* **16**, 24.

Ni, W., Fahrendorf, T., Ballance, G.M., Lamb, C.J. and Dixon, R.A. (1996) Stress Responses in Alfalfa (Medicago Sativa L.). XX. Transcriptional Activation of Phenlpropanoid Pathway Genes in Elicitor-Induced Cell Suspension Cultures. *Plant Mol. Biol.* **30**, 427–438.

Nielsen, R., Paul, J.S., Albrechtsen, A. and Song, Y.S. (2011) Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12**, 443–451.

Novaes, E., Kirst, M., Chiang, V., Winter-Sederoff, H. and Sederoff, R. (2010) Lignin and biomass: a negative correlation for wood formation and lignin content in trees. *Plant Physiol*, **154**, 555–561.

Nurnberger, J.I., Koller, D. L., Jung, J. *et al.* (2014) Identification of pathways for bipolar disorder: a meta-analysis. *JAMA Psychiatry*, **71**, 657.

O'Brien, E.J., Monk, J.M. and Palsson, B.O. (2015) Using genome-scale models to predict biological capabilities. *Cell*. **161**, 971–987.

Gomes de Oliveira Dal'Molin, C., Quek, L.E., Saa, P.A. and Nielsen, L.K. (2015) A multi-tissue genome-scale metabolic modeling framework for the analysis of whole plant systems. *Front. Plant Sci.* **6**, 4.

de Oliveira Dal'Molin, C.G., Quek, L.E., Palfreyman, R.W., Brumbley, S.M. and Nielsen, L.K. (2010) C4GEM, a genome-scale metabolic model to study C4 plant metabolism. *Plant Physiol.* **154**, 1871–85.

de Oliveira Dal'Molin, C.G., Quek, L.-E., Palfreyman, R. W., Brumbley, S. M. and Nielsen, L. K. (2010) AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. *Plant Physiol*, **152**, 579–589.

Orth, J.D., Thiele, I. and Palsson, B.Ø. (2010) What is flux balance analysis? *Nat. Biotechnol.* **28**, 245–248.

Papp, B., Pál, C. and Hurst, L.D. (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature*, **429**, 661–664.

Pardo, E.G. and Gutiérrez, C. (1990) Cell cycle- and differentiation stage-dependent variation of DUTPase activity in higher plant cells. *Exp. Cell Res.* **186**, 90–98.

Paterson, A.H., Lander, E.S., Hewitt, J.D., Peterson, S., Lincoln, S.E. and Tanksley, S.D. (1988) Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature*, **335**, 721–726.

Peng, X.-P., Sun, S.-L., Wen, J.-L., Yin, W.-L. and Sun, R.-C. (2014) Structural characterization of lignins from hydroxycinnamoyl transferase (HCT) down-regulated transgenic poplars. *Fuel*, **134**, 485–492.

Phillips, P.C. (2008) Epistasis - The essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **9**, 855–867.

Poolman, M.G., Miguet, L., Sweetlove, L.J. and Fell, D.A. (2009) A genome-scale metabolic model of arabidopsis and some of its properties. *Plant Physiol.* **151**, 1570–1581.

Prasad, K.V.S.K., Song, B.-H., Olson-Manning, C. *et al.* (2012) A gain-of-function polymorphism controlling complex traits and fitness in nature. *Science*, **337**, 1081–1084.

Quan, M., Du, Q., Xiao, L., Lu, W., Wang, L., Xie, J., Song, Y., Xu, B. and Zhang, D. (2018) Genetic architecture underlying the lignin biosynthesis pathway involves noncoding RNAs and transcription factors for growth and wood properties in Populus. *Plant Biotechnol. J.* **17**, 302–315.

Richard, P., Charron, P., Carrier, L. *et al.* (2003) Hypertrophic Cardiomyopathy: Distribution of Disease Genes, Spectrum of Mutations, and Implications for a Molecular Diagnosis Strategy. *Circulation*, **107**(17), 2227–2232.

Ruan, Y.-L. (2003) Suppression of sucrose synthase gene expression represses cotton fiber cell initiation, elongation, and seed development. *Plant Cell*, **15**, 952–964.

Rueda-López, M., Pascual, M.B., Pallero, M., Henao, L.M., Lasa, B., Jauregui, I., Aparicio-Tejo, P.M., Cánovas, F.M. and Ávila, C. (2017) Overexpression of a pine Dof transcription factor in hybrid poplars: a comparative study in trees growing under controlled and natural conditions. *PLoS ONE*, **12**, e0174748.

Saha, R., Suthers, P.F. and Maranas, C.D. (2011) Zea Mays IRS1563: A Comprehensive Genome-Scale Metabolic Reconstruction of Maize Metabolism' ed. Mikael Rørdam Andersen. *PLoS ONE*, **6**, e21784.

Sajitz-Hermstein, M., Töpfer, N., Kleessen, S., Fernie, A.R. and Nikoloski, Z. (2016) iReMet-flux: constraint-based approach for integrating relative metabolite levels into a stoichiometric metabolic models. *Bioinformatics*, **32**, i755–i762.

Sarrobert, C., Thibaud, M.-C., Contard-David, P., Gineste, S., Bechtold, N., Robaglia, C. and Nussaume, L. (2000) Identification of an *Arabidopsis thaliana* mutant accumulating threonine resulting from mutation in a new dihydrodipicolinate synthase gene. *Plant J.* **24**, 357–368.

Satish Kumar, V., Dasika, M.S. and Maranas, C.D. (2007) Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, **8**, 212.

Savage, L.J., Imre, K.M., Hall, D.A. and Last, R.L. (2013) Analysis of essential arabidopsis nuclear genes encoding plastid-targeted proteins. *PLoS ONE*, **8**, e7329.

Schröder, M. (2005) Functional analysis of the pyrimidine de novo synthesis pathway in solanaceous species. *Plant Physiol.* **138**, 1926–1938.

Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M. and Sauer, U. (2012) Multidimensional optimality of microbial metabolism. *Science*, **336**, 601–604.

Scuteri, A., Sanna, S., Chen, W.-M. *et al.* (2007) Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet.* **3**, e115.

Segrè, D., DeLuna, A., Church, G.M. and Kishony, R. (2005) Modular epistasis in yeast metabolism. *Nat. Genet.* **37**, 77–83.

Sharma, A., Kitsak, M., Cho, M.H. *et al.* (2018) Integration of molecular interactome and targeted interaction analysis to identify a COPD disease network module. *Sci. Rep.* **8**, 1–14.

Simons, M., Saha, R., Amiour, N. *et al.* (2014) Assessing the metabolic impact of nitrogen availability using a compartmentalized maize leaf genome-scale model. *Plant Physiol.* **166**, 1659–1674.

Soltis, N.E. and Kliebenstein, D.J. (2015) Natural variation of plant metabolism: genetic mechanisms, interpretive caveats, evolutionary and mechanistic insights. *Plant Physiol.* **169**(3), 1456–1468.

Sun, J., Loboda, T., Sung, S.J.S. and Black, C.C. (1992) Sucrose synthase in wild tomato, lycopersicon chmielewskii, and tomato fruit sink strength. *Plant Physiol.* **8**, 1163–1169.

Sweetlove, L.J., Williams, T.C.R., Maurice Cheung, C.Y. and George Ratcliffe, R. (2013) Modelling metabolic CO_2 evolution - a fresh perspective on respiration. *Plant cell Environ.* **36**, 1631–1640.

Tian, J., Song, Y., Du, Q. *et al.* (2016) Population genomic analysis of gibberellin-responsive long non-coding RNAs in populus. *J. Exp. Bot.* **67**, 2467–248

Togninalli, M., Seren, Ü., Meng, D. *et al.* (2018) The AraGWAS catalog: a curated and standardized *Arabidopsis thaliana* GWAS catalog. *Nucleic Acids Res.* **46**, D1150–D1156.

Tohge, T., Watanabe, M., Hoefgen, R. and Fernie, A. R. (2013) The evolution of phenylpropanoid metabolism in the green lineage. *Crit. Rev. Biochem. Mol. Biol.* **48**, 123–152.

Tong, A.H.Y. *et al.* (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808–813.

Torkamani, A., Topol, E.J. and Schork, N.J. (2008) Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*, **92**, 265–272.

Tuskan, G.A., Difazio, S., Jansson, S. *et al.* (2006) The Genome of Black Cottonwood, Populus Trichocarpa (Torr. & Gray). *Science.* **313**, 1596–1604.

Varma, A. and Palsson, B.O. (1994) Metabolic flux balancing: basic concepts, scientific and practical use. *Bio/Technol.* **12**(10): 994–998.

Vauterin, M., Frankard, V. and Jacobs, M. (1999) The *Arabidopsis thaliana* dhdps gene encoding dihydrodipicolinate synthase, key enzyme of lysine biosynthesis, is expressed in a cell-specific manner. *Plant Mol. Biol.* **39**, 695–708.

Vogel, G., Aeschbacher, R.A., Müller, J., Boller, T. and Wiemken, A. (1998) Trehalose-6-phosphate phosphatases from *Arabidopsis thaliana*: identification by functional complementation of the yeast tps2 mutant. *Plant J.* **13**, 673–683.

Wang, W.H., Takano, T., Shibata, D., Kitamura, K. and Takeda, G. (2006) Molecular basis of a null mutation in soybean lipoxygenase 2: substitution of glutamine for an iron-ligand histidine. *Proc. Nati. Acad. Sci.* **91**, 5828–5832.

Wang, Y., Bouwmeester, K., Beseh, P., Shan, W. and Govers, F. (2014a) Phenotypic analyses of *Arabidopsis* T-DNA insertion lines and expression profiling reveal that multiple L-type lectin receptor kinases are involved in plant immunity. *Molecular Plant–Microbe Interactions.* **27**, 1390–1402.

Wang, L., Matsushita, T., Madireddy, L., Mousavi, P. and Baranzini, S. (2014b) PINBPA: Cytoscape app for network analysis of GWAS data. *Bioinformatics*, **31**, 262–264.

Wang, K., Li, M. and Bucan, M. (2007) Pathway-based approaches for analysis of genome-wide association studies. *Am J Human Genet.* **81**: 1278–1283.

Wang, K., Li, M. and Hakonarson, H. (2010) Analysing biological pathways in genome-wide association studies. *Nat. Rev. Genet.* **11**, 843–854

Wei, Z., Qu, Z., Zhang, L., Zhao, S., Bi, Z., Ji, X., Wang, X. and Wei, H. (2015) Overexpression of poplar xylem sucrose synthase in tobacco leads to a thickened cell wall and increased height. *PLoS ONE*, **10**, e0120669.

Wu, S., Tohge, T., Cuadros-Inostroza, Á. *et al.* (2018) Mapping the Arabidopsis metabolic landscape by untargeted metabolomics at different environmental conditions. *Mol. Plant*, **11**, 118–134.

Xu, H.-H., Liu, S.-J., Song, S.-H., Wang, W.-Q., Møller, I. M. and Song, S.-Q. (2016) Proteome changes associated with dormancy release of dongxiang wild rice seeds. *J. Plant Physiol.* **206**, 68–86

Xu, L., Barker, B. and Gu, Z. (2012) Dynamic epistasis for different alleles of the same gene. *Proc. Nati. Acad. Sci.* **109**, 10420–10425.

Xu, Y., Chang, P.F.L., Liu, D., Narasimhan, M.L., Raghothama, K.G., Hasegawa, P.M. and Bressan, R.A. (1994) Plant defense genes are synergistically induced by ethylene and methyl jasmonate. *Plant Cell*, 1077–1085

Yuan, H., Cheung, C.M., Poolman, M.G., Hilbers, P.A. and van Riel, N.A. (2016) A Genome-Scale Metabolic Network Reconstruction of Tomato (Solanum Lycopersicum L.) and Its Application to Photorespiratory Metabolism. *Plant J.* **85**, 289–304.

Zeng, Z.B. (1994) Precision mapping of quantitative trait loci. *Genetics*, **36**, 1457–1468

Zhang, J., Yang, Y., Zheng, K. *et al.* (2018a) Genome-Wide Association Studies and Expression-Based Quantitative Trait Loci Analyses Reveal Roles of HCT2 in Caffeoylquinic Acid Biosynthesis and Its Regulation by Defense-Responsive Transcription Factors in Populus. *New Phytol.* **220**, 502–516.

Zhang, X., Misra, A., Nargund, S., Coleman, G.D. and Sriram, G. (2018b) Concurrent isotope-assisted metabolic flux analysis and transcriptome profiling reveal responses of poplar cells to altered nitrogen and carbon supply. *Plant J.* **93**:472–488.

Zhang, J., Li, M., Bryan, A.C. *et al.* (2019) Overexpression of a serine hydroxymethyltransferase increases biomass production and reduces recalcitrance in the bioenergy crop: populus. *Sustain Energ Fuels*, **3**, 195–207

Zhou, L., Jang, J.C., Jones, T.L. and Sheen, J. (1998) Glucose and ethylene signal transduction crosstalk revealed by an Arabidopsis glucose-insensitive mutant. *Proc. Nati. Acad. Sci.* **95**, 10294–10299

Zhou, X., Ren, S., Lu, M., Zhao, S., Chen, Z., Zhao, R. and Lv, J. (2018) Preliminary study of cell wall structure and its mechanical properties of C3H and HCT RNAi transgenic poplar sapling. *Sci. Rep.* **8**(1), 1–10.

Zhu-Shimoni, J.X., Lev-Yadun, S., Matthews, B. and Galili, G. (1997) Expression of an aspartate kinase homoserine dehydrogenase gene is subject to specific spatial and temporal regulation in vegetative tissues, flowers, and developing seeds. *Plant Physiol.* **113**, 695–706.

Zrenner, R., Salanoubat, M., Willmitzer, L. and Sonnewald, U. (1995) Evidence of the crucial role of sucrose synthase for sink strength using transgenic potato plants (Solanum Tuberosum L.). *Plant J.* **7**, 97–107

Zrenner, R., Stitt, M., Sonnewald, U. and Boldt, R. (2006) Pyrimidine and purine biosynthesis and degradation in plants. *Annu. Rev. Plant Biol.* **57**, 805–836.