Cellular and Molecular Life Sciences
https://doi.org/10.1007/s00018-021-03844-4

REVIEW

Characterization of effects of genetic variants via genome-scale metabolic modelling

Hao Tong[1,2,3] · Anika Küken[1,3] · Zahra Razaghi-Moghadam[1,3] · Zoran Nikoloski[1,2,3]

Received: 30 January 2021 / Revised: 15 April 2021 / Accepted: 23 April 2021

Abstract

Genome-scale metabolic networks for model plants and crops in combination with approaches from the constraint-based modelling framework have been used to predict metabolic traits and design metabolic engineering strategies for their manipulation. With the advances in technologies to generate large-scale genotyping data from natural diversity panels and other populations, genome-wide association and genomic selection have emerged as statistical approaches to determine genetic variants associated with and predictive of traits. Here, we review recent advances in constraint-based approaches that integrate genetic variants in genome-scale metabolic models to characterize their effects on reaction fluxes. Since some of these approaches have been applied in organisms other than plants, we provide a critical assessment of their applicability particularly in crops. In addition, we further dissect the inferred effects of genetic variants with respect to reaction rate constants, abundances of enzymes, and concentrations of metabolites, as main determinants of reaction fluxes and relate them with their combined effects on complex traits, like growth. Through this systematic review, we also provide a roadmap for future research to increase the predictive power of statistical approaches by coupling them with mechanistic models of metabolism.

Keywords Single-nucleotide polymorphisms · Metabolic models · Genome-wide association studies · Genomic selection

Introduction

Advances in genotyping have provided unprecedented insights in the genetic variations among individuals of the same species. Allelic variations within a genome of the same species include the differences in the number of tandem repeats at a particular locus, segmental insertions/deletions (indels), and single-nucleotide polymorphisms (SNPs) [1]. Since SNPs represent the most abundant form of allelic variations [2], they represent the predominant factor that induces phenotypic differences among individuals. Usage of SNPs with modern machine-learning approaches have revolutionized molecular plant breeding, both with respect to applied research in prediction of traits and basic research in the mechanisms governing a trait [3–5]. Hence, characterising the effects of SNPs on agronomically relevant traits is a key problem in the interlinked fields of plant systems biology, quantitative genetics, and plant breeding.

Depending on their genomic location, SNPs have the potential to alter all steps of transcription and translation. For instance, if a SNP lies in a transcriptional regulatory element, it can alter mRNA expression; in addition, SNPs that do not lie in protein-coding regions can affect splicing, mRNA degradation, and the sequence of non-coding RNA. If a SNP that lies in a protein-coding region is synonymous, i.e. does not cause amino acid change, it can affect the translation rate and turnover of mRNA, ultimately reflected in changes of the protein abundance; finally, if the SNP is nonsynonymous (missense or nonsense), i.e. leads to amino acid change, it can modify the protein activity. Through their effects on mRNA, enzyme abundance and stability as well as enzyme activity, SNPs have direct effect on metabolic reactions catalysed by the respective enzymes.

Metabolism represents the entirety of biochemical reactions through which nutrients are imported from

✉ Zoran Nikoloski
zniko@uni-potsdam.de

1 Bioinformatics Group, Institute of Biochemistry and Biology, University of Potsdam, Potsdam, Germany

2 Bioinformatics and Mathematical Modeling Department, Centre for Plant Systems Biology and Biotechnology, Plovdiv, Bulgaria

3 Systems Biology and Mathematical Modeling Group, Max Planck Institute of Molecular Plant Physiology, Potsdam, Germany

Published online: 05 May 2021

the environment and are transformed into the building blocks of biomass, ensuring growth, as well as other cellular components that support defence, reproduction, and, ultimately, survival [6]. A quantitative characteristic of a metabolic reaction is its rate. The rate or flux of a reaction denotes the speed at which it transforms the substrates into products [7]. The flux of a reaction depends on the abundance, E, of the enzyme that catalyses the reaction, its turnover number, k_{cat}, representing the number of substrate molecules that each active site of the enzyme converts to product molecules per unit time, and the concentration of metabolites, x, acting as substrates and/or effectors (e.g. allosteric regulators, inhibitors). In the most general form, the flux of a metabolic reaction r can be mathematically written as:

$$v\left(k_{cat}, \boldsymbol{K}, E, x\right) = k_{cat} E \eta(\boldsymbol{K}, x) = V_{max} \eta(\boldsymbol{K}, x),$$

where $\boldsymbol{K}$ denotes a set of parameters (e.g. Michaelis–Menten constants, K_m, equilibrium constants, K_{eq}), V_{max} is the maximal enzyme activity, and $\eta(\boldsymbol{K}, x)$ is a function that models the effect of metabolite concentration on the flux.

Metabolic reactions do not operate in isolation and jointly affect the temporal change of metabolite concentrations (Fig. 1a). A metabolic reaction can be described by the stoichiometry of its substrates and products, yielding the stoichiometric matrix, N, over all reactions (Fig. 1b). The change of metabolite concentrations over time can then be modelled as $\frac{dx}{dt} = Nv$, where v gathers the fluxes of all reactions in the modelled metabolic network. Correspondingly, we can categorise the effect of SNPs on reaction fluxes into local, affecting k_{cat}, and global, via transient effects of SNPs on enzyme abundance, E, metabolite concentrations, x. Given the role of reaction fluxes in shaping the main components of growth and other cellular tasks important for survival, it is paramount to determine the effects of SNPs on reaction fluxes and to further dissect them into local and global.

Reactions can be divided into extra- and intracellular based on whether or not they facilitate the exchange (i.e. import or export) of metabolites with the environment. Monitoring the change of extracellular metabolite concentrations over time can be readily used to estimate extracellular reaction fluxes [8]. However, intracellular fluxes are more difficult to quantify, and require setting up isotope labelling experiments and measurement of metabolite labelling patterns which are then fitted to a metabolic model [9, 10]. In plants, the problem is further complicated by the fact that time-resolved metabolite labelling patterns from feeding $^{13}CO_2$ are required to infer intracellular reaction fluxes in photoautotrophic growth [11, 12]. Therefore, isotope labelling experiments are currently too laborious to allow estimation of fluxes in a population of individuals from a given species, rendering it infeasible to dissect the genetic architecture of fluxes in different model plants and crops following this approach.

As a result, other computational approaches have been developed to predict/estimate fluxes in the constraint-based modelling framework based on the assumption that an organism optimises a cellular task (e.g. growth) under a set of physicochemical constraints [13] (Fig. 1c). This is the essence of flux balance analysis (FBA) which provides efficient means to estimate fluxes based on constraints from measurement of extracellular fluxes and growth of microorganisms [14]. Extension of FBA has led to variants that include additional assumptions capturing efficient usage of cellular resources [15]. Interestingly, this parsimonious strategy is also often followed in application of constraint-based modelling approaches with plant metabolic networks [16, 17]. In contrast to the isotope labelling experiments above, constraint-based approaches provide a feasible means to begin to unravel the genetic determinants of reaction fluxes in plants and to use them in plant breeding.

Here, we review a collection of recent modelling approaches, which allow the dissection of the genetic basis of reaction fluxes by identifying their association with SNPs that are integrated into metabolic networks. Since these approaches can be grouped based on whether or not they rely on the principles underlying genome-wide association and genomic selection, we also describe the basic methodology underlying these machine-learning and statistical approaches. Focusing on the global effects of SNPs, we also provide a succinct review of studies that examine SNP effects on maximal enzyme activity in model plants and crops. We then offer a perspective for determining local effects of SNPs on k_{cat}'s by coupling of proteomics technologies and modelling approaches in diversity panels. Finally, we point out how these modelling approaches can help address the transferability of statistical models to make predictions of traits in unseen environments by their integration into mechanistic models of metabolism.

Constraint-based metabolic models of model plants and crops

Access to a high-quality metabolic model of an organism is key to accurate estimation of fluxes. Genome-scale metabolic models (GEMs) gather the entirety of documented metabolic reactions assembled based on annotation of the genome of an organism [18]. GEMs are further refined to include cellular compartments by considering information of protein localization and intracellular transporters. GEMs usually include a synthetic reaction, called biomass reaction, that expresses biomass as a defined ratio of macromolecules synthesised from metabolites, assembled from genome

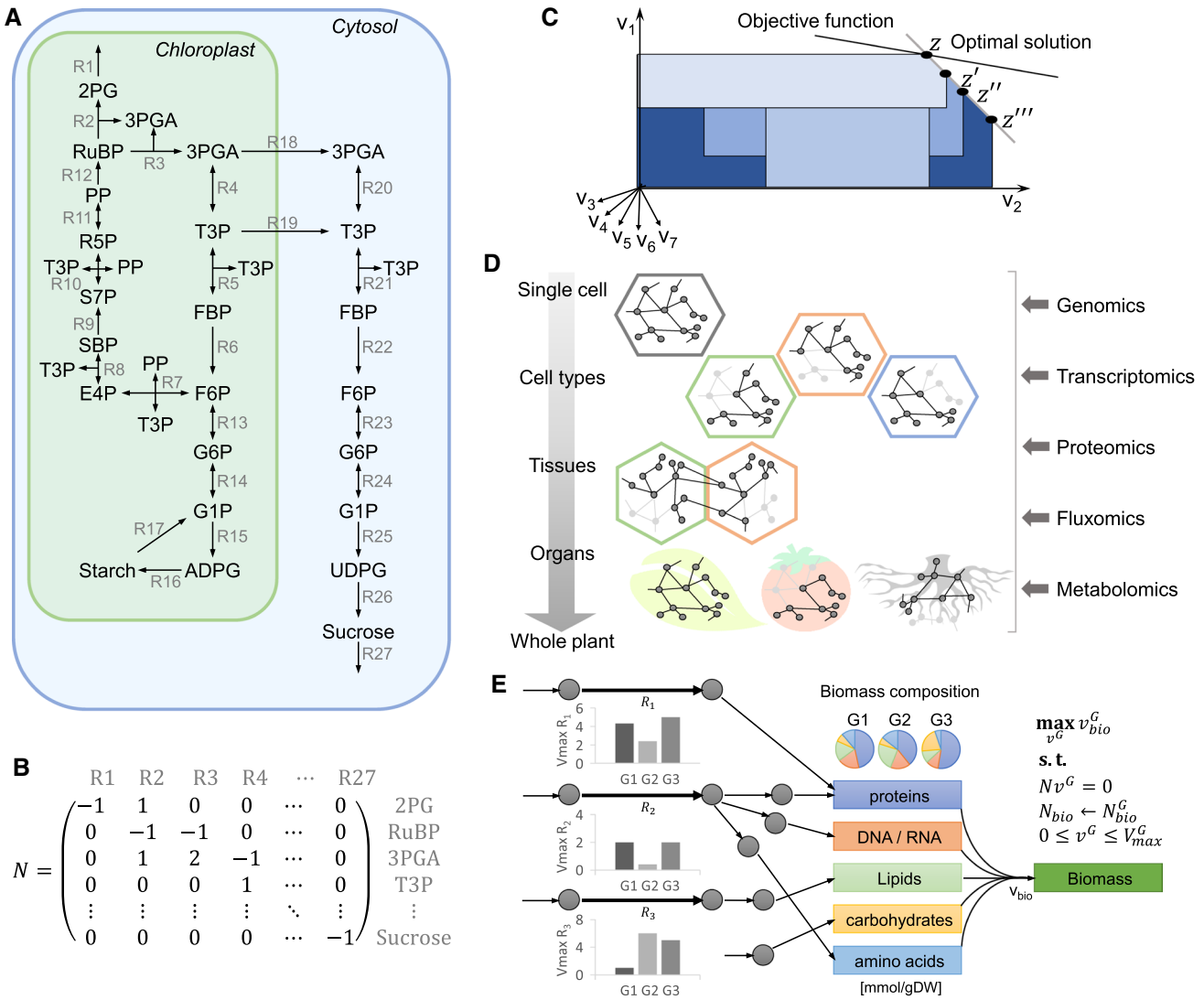

Fig. 1 Concepts from constraint-based modelling of metabolic networks. **a** Simplified metabolic network of the Calvin–Benson cycle, starch and sucrose synthesis including two compartments (chloroplast and cytosol), 27 reactions and 24 compartment-specific metabolites. All triose-3-phosphates are lumped in a common pool denoted by T3P. **b** The concept of the stoichiometric matrix N on reactions R_1 to R_4 and R_{27} from (**a**). **c** The system of linear equations representing the metabolic model has multiple solutions, forming the solution space. Data-driven constraints can be included to reduce the solution space, each resulting in a smaller subspace. **d** Integration of data from various technologies/approaches (genomics, transcriptomics, proteomics, fluxomics, and metabolomics) allow the reconstruction of cell type-, tissue- or organ-specific metabolic networks. **e** Data on maximal reaction rates (V_{max}) and biomass composition for different genotypes (here G_1, G_2, and G_3) can be used to further refine the predictions from metabolic networks to obtain genotype-specific flux estimates. Metabolite abbreviations: 2PG—2-phosphoglycerate, RuBP—ribulose-1,5-bisphosphate, 3PGA—3-phosphoglycerate, T3P—triose-3-phosphates, FBP—fructose-1,6-bisphosphate, F6P—fructose 6-phosphate, G6P—glucose 6-phosphate, G1P—glucose 1-phosphate, ADPG—ADP-glucose, UDPG—UDP-glucose, PP—pentose-5-phosphates, E4P—erythrose-4-phosphate, SBP—sedoheptulose-1,7-bisphosphate, S7P—sedoheptulose-7-phosphate, R5P—ribulose-5-phosphate

annotation and metabolomics measurements [19]. Since metabolism differs between cell types, tissues, and organs, omics data (e.g. transcriptomics, proteomics, and metabolomics) from these cellular context have been used in combination with constraint-based approaches to extract respective context-specific metabolic networks [20] (Fig. 1d).

Efforts in the last decade have resulted in the assembly of high-quality GEMs and metabolic models of central carbon metabolism for key model plants and crops, including: *Arabidopsis thaliana* (Arabidopsis) [21–31], *Oryza sativa* (rice) [32–37], *Zea mays* (maize) [23, 30, 38–40], *Solanum lycopersicum* (tomato) [41], *Solanum tuberosum* (potato) [42], *Hordeum vulgare* (barley) [43], *Brassica napus* (oilseed rape) [44, 45], *Medicago truncatula* [46], *Glycine max* (soybean) [47], *Setaria viridis* [48] and *Populus trichocarpa* [49], as well as generic models for CAM,

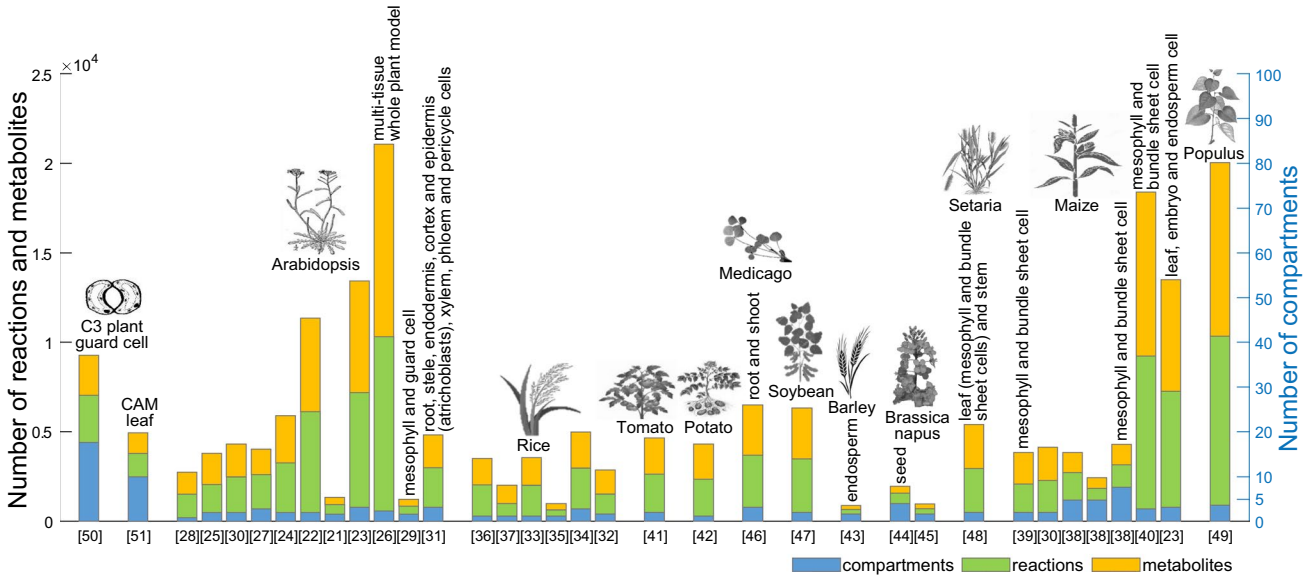

Fig. 2 Overview of available plant metabolic network reconstructions. The existing stoichiometric models of model plants and crops are compared based on the number of compartments, metabolites and reactions included

C_3, and C_4 plant species [50–53] (Fig. 2). The models differ with respect to whether they only include pathways from primary metabolism (e.g. AraCore in Arabidopsis [21]) or they also consider pathways of secondary metabolism (e.g. the Arabidopsis model of Mintz-Oron et al. [27]). Further, the models include different details of representation of the underling biochemical reactions, which particularly holds for lipid metabolism [54]. They also differ with respect to the number of cellular compartments modelled and genes included (Fig. 2). The latter is particularly important if missense SNPs are to be integrated in these models following the gene–protein–reaction (GPR) rules, modelling the relation between genes, their products, and the reactions they catalyse [18]. Further, based on these GEMs, context-specific metabolic networks have already been extracted for Arabidopsis cotyledon, flower bud, open flower, root, juvenile leaf and silique [27], mesophyll and guard cells [29, 50], root cell types [31], as well as mesophyll and bundle sheath cells in maize, along with models of maize leaf, embryo, and endosperm [23, 40]. These models have been used to make predictions and further analyse genome-scale flux distributions under different growth scenarios [16].

Determining the genetic basis of the flux for a given reaction requires that it is quantified in multiple individuals (i.e. genotypes) whose metabolic networks may differ (Fig. 1e). To this end, availability of quantitative metabolomics data from different individuals allow the possibility to set up genotype-specific biomass reactions [55, 56]. In addition, measurements of extracellular fluxes across individuals as well as maximal enzyme activities, V_{max}, can be used to establish genotype-specific constraints on the model's input and output and intracellular fluxes [55]. Further, omics data from different genotypes in conjunction with context-specific model extraction approaches, mentioned above, can be used to extract more refined genotype-specific models. The resulting genotype-specific metabolic models together with constraint-based modelling approaches [14] can be used to obtain genotype-specific flux distributions, as a first step in dissecting the genetic basis of individual fluxes.

Statistical approaches for linking SNPs with metabolic traits

Establishing a link between genetic markers (e.g. SNPs) and a trait of interest is carried out by application of machine-learning and statistical approaches. Two principal questions can be posed: (1) is the trait statistically associated with a genomic region or position? (2) Are the genetic markers predictive of the trait? (in the sense of predicting a major proportion of the variance). These questions can be used to group the statistical approaches to link genetic markers with (metabolic) traits into those that aim to conduct genetic mapping and those that devise models for genomic selection, respectively.

Genetic mapping approaches

Genetic mapping of a given trait can be used to determine and dissect the genetic architecture of the trait. Therefore, it provides a useful approach to improve crop breeding towards generation of better performing genotypes [57]. An

essential requirement for genetic mapping is having access to a population with available genotypic data, describing the genetic variation, and phenotypic data for studied traits. Genetic mapping consists of five steps: (1) design or select a population, (2) collect the genotypic and phenotypic data, (3) conduct a screen based on statistical genetic models, (4) prioritise the significant signal for candidate genes, and (5) validate candidate genes [58]. Based on the population employed, the statistical models used to link the genotypic with phenotypic variation can drastically differ: the approach using biparental populations (e.g. F_2 populations, backcross, and recombinant inbred lines (RILs)) is termed as quantitative trait loci (QTL) mapping (Fig. 3a), while that using natural populations (i.e. diversity panels) is at the core of genome-wide association studies (GWAS) [59] (Fig. 3b). Preprocessing of data on multiple traits based on principal component analysis can be also used to derive linear combinations of traits as latent variables, which can also be used in mapping [60].

QTL mapping

QTL for a studied trait denotes genomic regions that control the trait. QTL mapping relies on using low-density genetic markers, e.g. amplified fragment length polymorphism (AFLP), restriction fragment length polymorphism (RFLP), and simple sequence repeat (SSR), because the recombination blocks in biparental populations are relatively big. This approach has provided powerful means to identify loci that co-segregate with the studied trait in the employed biparental population, due to the smaller number of false positive candidates [59]. However, the resolution of QTL mapping is relatively low as it depends solely on the recombination events that occur during the process of generating the population [61]. Multi-parent populations can increase the mapping resolution [62–64], but require high-density genetic markers which can readily be obtained with modern cost-effective genotyping technologies. The statistical approaches for QTL mapping are based on the linkage map, which is the order of markers on chromosome and genetic distance between marker pairs. The most widely used QTL mapping model is composite interval mapping (CIM) model that considers the covariates to eliminate the effect of markers outside the tested interval [65] (Fig. 3a).

GWAS

In contrast to QTL mapping, GWAS has a relatively higher resolution, often down to a single gene level, since it relies on high-density SNPs covering the entire genome [66]. Therefore, GWAS has been the method of choice to dissect the genetic architecture of quantitative traits in animals and plants in the last decade [67–69]. The evolutionary history of diversity panels leads to accumulation of alleles that are in linkage disequilibrium, allowing to detect association between genotype and phenotype. However, the power of GWAS to detect true associations can be affected by at least five factors: (i) the mapped trait should exhibit (approximately) normal distribution, (ii) size of the population, which is related to the proportion of associations

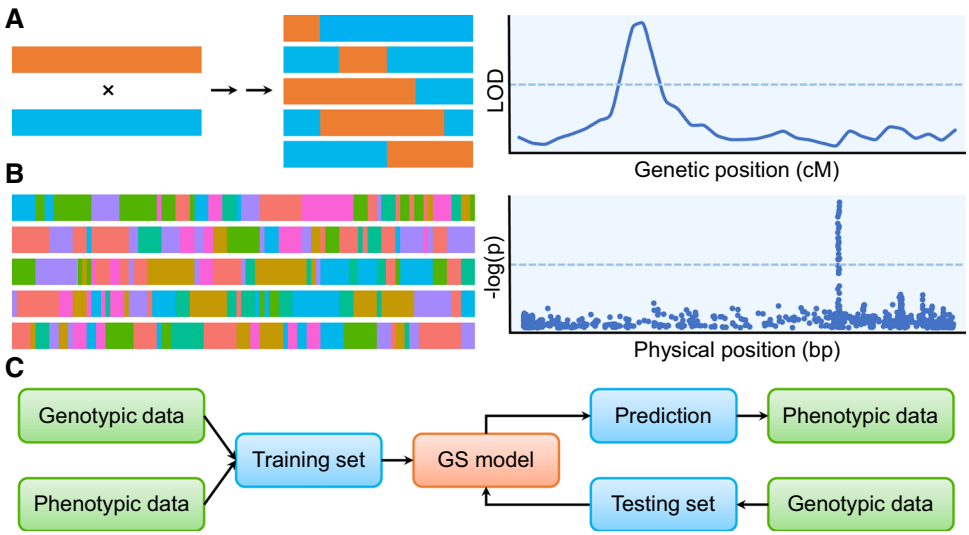

Fig. 3 Statistical approaches for linking SNPs to (metabolic) traits. **a** Biparental mapping population based on crossing of parents that show differing values for a trait of interest together with a LOD scores for regions associated with the trait. **b** GWAS population composed of genetically diverse genotypes along with a Manhattan plot showing the *p* value of the SNPs used in mapping. **c** The process underlying genomic selection, in which genotypic and phenotypic data in a training set is used to train a statistical model for a studied trait, followed by application of the model to a testing population that is only genotype to predict respective phenotypes

of higher effects, (iii) population structure, which leads to identification of spurious associations, (iv) allele frequency, that affects the resolution power, and (v) linkage disequilibrium, that assists in defining the significantly affected loci [59].

To address the issues of population structure and relatedness present in natural population, a mixed linear model (MLM) including kinship matrix and population structure was proposed [70], which is the most common used GWAS model in plants. The model is given by

$$y = X\beta + S\alpha + Qv + \xi + e,$$

where y is a vector of phenotypic data, $X\beta$ is the intercept other than SNP effect and population structure, S is a design vector for each SNP, α is the SNP effect, Q is the population structure matrix, v is the population structure effect, $\xi \sim MVN(0, K\sigma_u^2)$ is the polygenic effect, and e is the residual error. The population structure can be revealed by approaches based on principal component analysis [71]. The polygenic effect and residual error are treated as random effects, while the others are fixed effects. Therefore, the variance of y is

$$V = K\sigma_u^2 + I\sigma_e^2,$$

where K is the kinship matrix, I is the identity matrix, σ_u^2 and σ_e^2 are the variance component of polygenic effect and residual errors, respectively. These variance components are estimated by restricted maximum likelihood (REML) approach. The best linear unbiased estimation (BLUE) of fixed effects and best linear unbiased prediction (BLUP) of random effects are then calculated. The test of significance is performed by the F test or likelihood ratio test between the model without consideration of the SNP effect and the model that includes the tested SNP (Fig. 3b). The test of significance is carried out in a single locus analysis, so a multiple test correction must also be performed.

However, application of the MLM approach is computationally challenging with the increase in the number of samples and SNPs that are required to improve the resolution and power of the genetic mapping. Several efficient GWAS algorithms have been devised to handle the population structure and kinship by employing elegant matrix transformations (e.g. the efficient mixed model association (EMMA) [72], genome-wide EMMA (GEMMA) [73], and factored spectrally transformed linear mixed model (FaST-LMM) [74]). In contrast to the above methods, other algorithms estimate the polygenic effect only once, and keep it constant for every tested SNP (e.g. population parameters previously determined (P3D) [75]). In addition, to avoid control the population stratification via kinship and population structure matrix, the multi-locus mixed model (MLMM) has also been used in GWAS [76].

From this brief review of computational approaches for genetic mapping based on GWAS, it is evident that they are all based on statistical approaches of association and do not consider mechanistic insights and constraints. Several pressing questions arise: can the coupling of the basic principles of GWAS with mechanistic models of metabolism help in detecting causal SNPs with local effects on reaction fluxes? If so, could this be done with smaller population sizes, without reducing the power of the detected associations? These questions will be addressed in "Approaches based on GWAS".

Genomic selection

Genomic selection (GS) is considered the most promising breeding method to speed up the development and release of improved genotypes [77]. It is based on a model to arrive at genomic estimated breeding value (GEBV) based on usage of genome-wide markers with various machine learning [78]. More specifically, GS uses machine learning to integrate phenotypic data of a given trait with molecular markers in a statistical model for a training population. The model is then employed to predict traits values of genotypes in a testing population, which have been genotyped but not phenotyped [79] (Fig. 3c). The predictions for unseen genotypes can be used for selection without any further phenotyping. Therefore, an increase in GS accuracy for agronomically important traits can accelerate genetic gain by shortening the breeding cycles [77].

In contrast to GWAS, GS forgoes statistical testing for the effect of SNPs, as the goal is to devise a model of high predictive power. Nevertheless, like GWAS, the accuracy of GS is affected by several factors, including: (i) the sample size, (ii) genetic relationship within and between the training and testing population, (iii) marker density, (iv) heritability of the trait, (v) linkage disequilibrium between markers and quantitative trait loci controlling the trait of interest, and (vi) non-additive genetic effects (e.g. epistasis) [80, 81]. It has been observed that increases in the sample size, but also changes in the structure of the training set have a strong effect on the prediction accuracy of GS [82]. Further, increases in accuracy of GS have been found to plateau after certain level of marker density [83]. GS models that take into consideration multi-environment data allow for sharing information across environments and usually lead to increase in accuracy in comparison to models derived from single-environment data [84]. However, the generation of such data takes considerable resources, so the question remains if the performance of single-environment models can be improved by modifying the modelling strategy. Finally, several studies have pointed out that epistasis is an important contributor to the long-term response to selection [85, 86]. However, while consideration of two-locus epistatic effects has led to

improvements in GS accuracy [87], general consideration of epistasis in GS models remains computationally challenging and deserves further method development.

Based on the machine-learning/statistical techniques employed, GS approaches can be roughly divided into those relying on regression, classification, and deep learning techniques [5]. Ridge regression best linear unbiased prediction (rrBLUP) is one of the most common used GS models in plants [78]; it is a mixed-effect linear model, given by

$$y = Xb + Zu + e,$$

where y is a vector of phenotype, X is the fixed effect design matrix, b is the fixed effect, Z is a matrix of genetic markers, u is the marker effect as random effect and e is the residual error. The variance of y is

$$V = ZZ^T\sigma_u^2 + I\sigma_e^2,$$

where σ_u^2 is the marker effect variance and σ_e^2 is the residual error variance. Since the number of markers is considerably larger than the number of observations (i.e. genotypes), regularisation techniques are usually used to estimate the model parameters. In comparison to ridge regression, the parameter λ of the l_2 norm is equivalent to $\lambda = \sigma_e^2/\sigma_u^2$ and penalises the ratio between the two random effect variance components. According to the mixed model theory, the value of GEBV can be solved and used to predict the phenotypic value in testing population.

This approach can shrink all effects toward zero equally across markers, under the assumption that all markers have a common variance. Other approaches, like genomic best linear unbiased prediction (GBLUP), estimate the kinship matrix from genomic markers to represent the pedigree information, then estimate GEBV in a mixed linear model that is equivalent to the rrBLUP model [88]. The GEBV can also be obtained from Bayesian statistics [78]. To compare the model performance in GS, one usually uses the prediction accuracy. It is determined by k-fold cross-validations, whereby one fold is treated as testing population and other folds as training population. The prediction accuracy is the correlation coefficient between the predicted phenotypic value and measured phenotypic value in the testing population.

Like the GWAS approaches outlined above, GS is based on machine-learning algorithms and the transferability of the resulting models to unseen scenarios, i.e. different population, different environments, and the combination of the two, remains one of the biggest challenges in the application of GS. Therefore, it is of interest to investigate if the prediction accuracy of GS for metabolic traits can be improved if the mentioned approaches are coupled with mechanistic models of metabolism, discussed in "Approach based on genomic selection".

Application of genetic mapping approaches for maximal enzyme activity in plants

Determining the genetic architecture of metabolism entails genetic mapping of metabolic traits, including: metabolite levels (relative and absolute content, concentrations), protein abundances and activities, and reaction fluxes. There are plethora of studies that use GWAS and QTL mapping approaches in diverse plants and crops based on measurement of metabolite levels and protein abundances [89, 90]. However, these studies rely on relative quantification of these traits, rendering it difficult to interpret the findings in terms of effects on reaction fluxes. A reaction flux depends linearly on the maximal activity, V_{max}, of the respective enzyme, and is fully determined by it when the enzyme is substrate-saturated [91]. Thus, it may be expected that the results of genetic mapping of V_{max} would coincide with those of the corresponding reaction fluxes. However, due to the interconnectedness of gene regulatory and protein–protein interaction networks that affect metabolism, QTL or associated SNPs can be found not only in *cis* position (i.e. on the same chromosome and proximal) to the location of the corresponding structural genes (coding of structural proteins, rather than regulatory proteins), but also in *trans* position (i.e. on a different chromosome), denoting regulatory QTL.

To this end, all the statistical approaches for genetic mapping mentioned above can be readily used to determine the genetic architecture of V_{max} of different enzymes as well as reaction fluxes, if these are measured in an investigated population. For instance, the only study to date that has performed QTL mapping of reaction fluxes uses flux estimations from a small model of *Saccharomyces cerevisiae* (yeast) central carbon metabolism [92] based on bounds of measured extracellular fluxes and profiling of dry weight in 125 F_2-segregants (genotyped by 3727 SNPs) from a cross of two yeast strains [93]. These approaches identified four flux QTL and two gene variants that contribute to the explanation of the variations in the flux distributions in the population.

Since intracellular fluxes are more challenging to quantify (see "Introduction"), majority of QTL mapping studies in plants have focused on dissecting the genetic basis of maximal enzyme activities. However, genome-wide profiling of maximal enzyme activities is currently not feasible, due to the limitations of the assays used [94]. As a result, these studies usually involve a handful to two dozens of enzymes, mostly covering key pathways in primary metabolism in maize, Arabidopsis, and tomato. For instance, Causse et al., Prioul et al., Thevenot et al., and Pelleschi et al. [95–98] measured the maximal enzyme activities of four enzymes, sucrose-phosphate-synthase,

sucrose-synthase, sucrose-invertase, and ADP-glucose pyrophosphorylase, covering key steps in carbohydrate metabolism in sources (i.e. leaves) and/or sinks (i.e. grains) in maize RIL populations. Colocation of QTL for maximal enzyme activity and structural gene were found for sucrose-phosphate-synthase and the invertase. Limami et al. [99] measured the activity of enzymes from nitrogen metabolism, including: glutamine synthase, NAD(H)-glutamate dehydrogenase, the ferredoxin-dependent as well as the NAD(H)-dependent glutamate synthase, and phosphoenolpyruvate carboxylase in a population of 140 maize RILs, and identified QTL for glutamine synthase in the early and late stages of germination. An intermated RIL maize population was used to map QTL for the activity of ten enzymes, six from carbon and four from nitrogen metabolism [100]. All identified QTL for enzyme activities in this study were in *trans* to the respective structural genes, except for single *cis*-QTL for nitrate reductase, glutamate dehydrogenase, and shikimate dehydrogenase.

In addition, Mitchell-Olds and Pedersen [101] performed QTL mapping of maximal activity for ten enzymes (i.e. six glycolytic enzymes, glucose-6-phosphate dehydrogenase, fructose bisphosphatase, phosphoglucose isomerase, phosphoglucomutase, glucose-6-phosphatase, and hexokinase, as well as four enzymes putatively involved in defence pathways, peroxidase, shikimic dehydrogenase, myrosinase, and chitinase) in an Arabidopsis RIL population. In another Arabidopsis RIL population, Sergeeva et al. [102, 103] mapped the activity of phosphoglucomutase and sucrose-invertase. The same population was later used to dissect the genetic architecture for the maximal activity of 15 enzymes [104]; QTL were detected for 10 of the 15 enzyme activities, which exhibited higher heritability, and involved respective structural genes as well as other genes with *cis*- and *trans*-acting control. A tomato introgression population, generated by introgressing segments of the genome of the wild relative *Solanum pennellii* into the modern tomato cultivar *Solanum lycopersicum*, was used to investigate QTL for the maximal enzyme activities of 28 enzymes from central carbon metabolism [105]. To this end, measurements were conducted in the pericarp tissue of ripe tomato fruits from two field trial experiments. The identified QTL support the observations from Arabidopsis that maximal enzyme activity is under the control of *trans*-acting genes.

The only GWAS with maximal enzyme activities as a trait was carried out in an Arabidopsis diversity panel composed of 349 accessions. To this end, associated SNPs for 24 maximal enzyme activities in central metabolism were detected [106]. The study identified *cis*-QTL of moderate effects for maximal enzyme activity of five enzymes, including UDP-glucose pyrophosphorylase, ADP-glucose pyrophosphorylase, fumarase, and phosphoglucose isomerase. The remaining QTL were *trans*-acting of smaller effects than the *cis*-acting, and were found in genomic regions that include components involved in transcriptional and post-translational modifications.

Genetic mapping of maximal enzyme activities in different plant species demonstrates that genetic variants in both regulatory and structural genes can affect this trait of different enzymes in central metabolism. Therefore, consideration of missense SNPs may only identify a small fraction of the phenotypic variance in this trait. The latter implies that the integration of SNPs into mechanistic models should consider the action of *trans*-acting genes for accurate predictions of their effects on metabolic traits.

Integration of SNPs in genome-scale metabolic models

The approaches that integrate SNPs into a metabolic network can be grouped based on two criteria: (i) if they investigate the positioning of SNPs in metabolic network, using the GPR rules and (ii) if they characterize the effect of a SNP on reaction fluxes. With respect to the second criterion, one can further subdivide these approaches based on whether they rely on principles of GWAS or genomic selection, as principal statistical approaches for linking SNPs with traits.

Approaches based on the metabolic network structure

A first approach to investigate the role of SNPs in metabolic networks is to characterize their position in the metabolic network. Due to the possibility that a metabolic reaction is catalysed by isoenzymes and protein complexes, as well as due to the promiscuity of some enzymes, whereby they can catalyse multiple reactions [107], the product of one gene can affect the flux through multiple reactions [108]. As a result, the effects of a nonsynonymous SNP on such a gene can be readily determined by investigating its position in the metabolic network. Jamshidi and Palsson indicated that the effect of SNPs that reside in genes whose products catalyse reactions that form co-sets can be readily obtained [109]. A co-set is a maximal set of reactions whose fluxes are perfectly correlated across any steady-state that the network can support [110], and coincide with fully coupled reactions from flux coupling analysis [111]. We note that a co-set can be composed of a single reaction if that reaction is not fully coupled to any other in the network. As a result, SNPs in the same co-set are expected to have similar effects. A co-set can consist of a single reaction, reactions on a linear chain, or subnetworks of more intricate structure which may also be disconnected, denoted as co-sets of types A, B, and C, respectively (Fig. 4a). While this approach is useful in providing a partitioning of SNPs based on their participation in

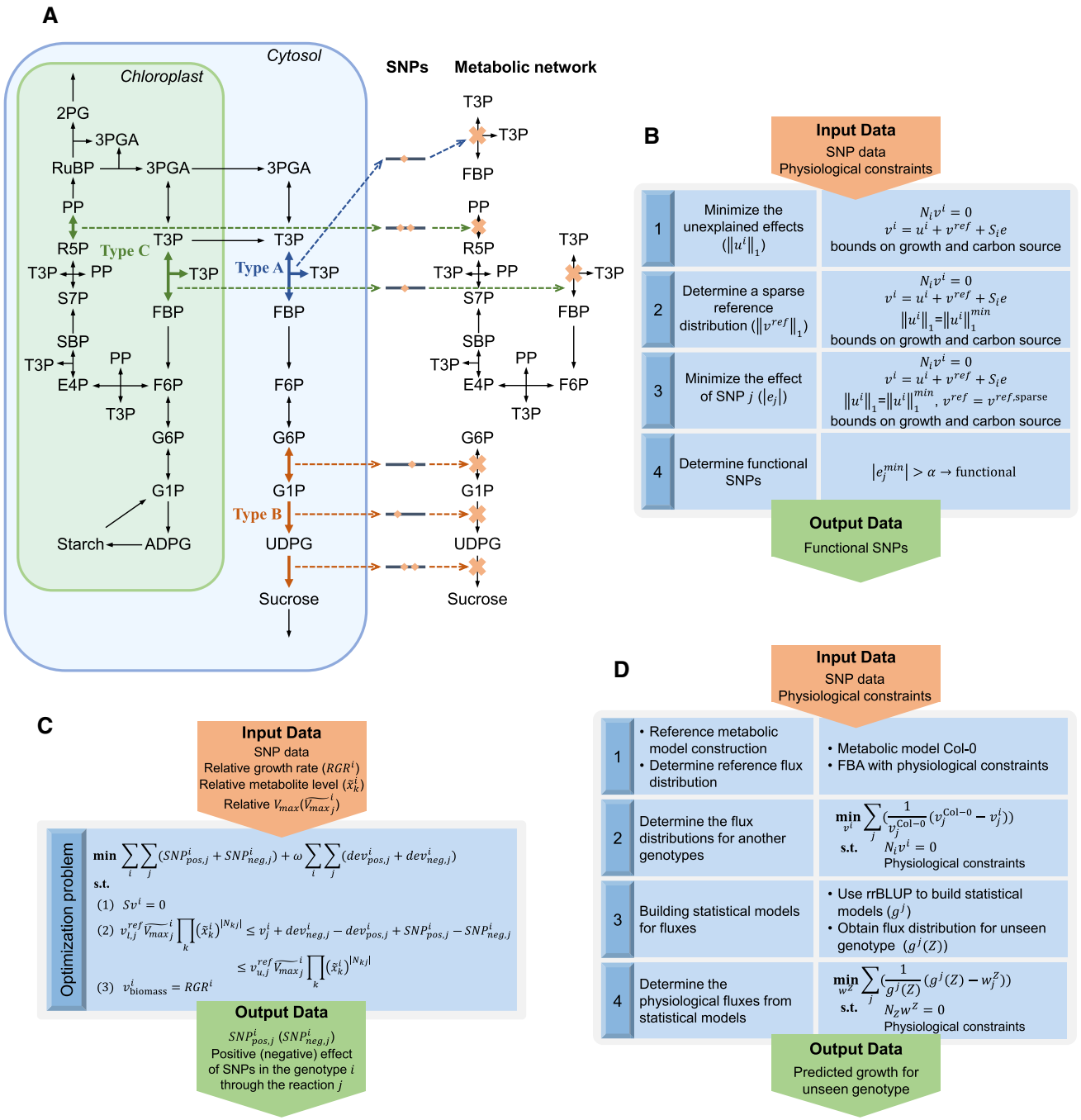

Fig. 4 Approaches that integrate SNPs into metabolic models. **a** Examples of different types of co-sets on the metabolic network of Fig. 1a are presented by different coloured arrows. Orange diamonds show the SNPs in the gene coding for the proteins that catalyse reactions in the co-sets. The causal SNPs affect the reactions, marked by an orange x symbol for knock-out, and result in the inability of the network to produce particular products. **b** The SNP effects in [113] are predicted through three optimization steps: (1) minimising the unexplained effects, (2) finding sparse reference flux distribution, and (3) minimising the flux effect of each SNP. In the fourth step, SNPs with the minimum effects larger than the threshold of α are considered as functional. **c** The positive or negative effect of SNPs are captured in SNPeffect [49] by an optimization problem, in which mass-action kinetics is assumed and relative growth rate, relative metabolite level and relative $V_{\max}$ are given. **d** Four steps presented in netGS [56] allows for prediction of growth in unseen genotypes: (1) the construction of reference metabolic model and the prediction of reference flux distribution, (2) prediction of flux distributions in other genotypes by finding the closest flux distributions to the reference one, which are compatible to physiological constraints, (3) building statistical models for fluxes based on SNPs, and (4) prediction of physiological flux distributions from statistical models by finding the closest steady-state flux distribution to that obtained from the statistical models

specific subsets of reactions, it does not provide a quantification of the effect of SNPs on reaction fluxes. Interestingly, to date, there has been no characterisation of the effect of SNPs with respect to other types of dependencies between steady-state reaction fluxes [112].

Approaches based on GWAS

One of the critical factors that determine the power of GWAS is the population size. Integration of SNPs in a metabolic network can facilitate characterisation of their effect on fluxes even with a very small population [113]. Here, we review two approaches, one based on structural sensitivity analysis and the other based on incorporation of metabolomics datasets under the assumptions about enzyme kinetics.

Structural sensitivity analysis and GWAS

This constraint-based approach is based on structural sensitivity analysis [114], whereby the effect of a SNP in a gene is the same for every reaction that the gene product catalyses. Based on structural sensitivity analysis, the propagation of the SNP effect to the rest of the fluxes in the network can be determined. To this end, the problem is cast as a least-square adjustment of steady-state fluxes whose solution results in a sensitivity matrix, S_i, for the genotype i. The following restrictions and assumptions apply: (i) only nonsynonymous SNPs in the genes included in the metabolic network are considered, (ii) the considered SNPs are allowed to only decrease reaction fluxes (as deleterious effects of mutations are more likely), and (iii) the effect of a SNP is the same in all analysed genotypes.

With these assumptions, the approach is based on representing the genotype-specific flux distribution, v^i, in terms of a reference distribution, v^{ref}, and deviations from it; the deviations can either be explained by the nonsynonymous SNPs, $S_i e$, or are unexplained by them, u^i, i.e.

$$v^i = u^i + v^{\text{ref}} + S_i e.$$

In this sense, u^i can be seen as a residual error that cannot be explained by the sensitivity matrix (via the effect of e) and the reference flux distribution. The flux distributions v^i are determined for all genotypes jointly by enforcing steady-state in each, i.e. $N_i v^i = 0$, using experimentally determined, genotype-specific exchange rates for a subset of metabolites. The simultaneous solving of the steady-state equations is needed due to the relation between flux distributions of different genotypes via the nonsynonymous SNP effects, given by e, $e \leq 0$. The four steps include: (1) finding a sparse solution for the unexplained effects (via minimization of the first norm); (2) determining a sparse reference distribution (that specifies reaction fluxes in absence of SNPs); (3) minimising the effect, e_j, of SNP j under the constraints of the spare solutions found in first two steps. This is needed due to the variability of e_j in the feasible space, and helps with the interpretation of the SNP effects; and (4) only SNPs whose minimum effects are larger than an arbitrarily selected threshold α are considered to have functional effects (Fig. 4b). This algorithm was tested with 18 strains of the *Mycobacterium tuberculosis* complex with 556 nonsynonymous SNPs, and 88 SNPs were classified as functional with the used threshold value.

The approach can be viewed as a multi-locus GWAS [76], but does not provide statistics for associations, as it relies on the predictions from the integration of SNPs in the metabolic network. The findings from this approach depend on: (i) the number of nonsynonymous SNPs in the genotyping data, which would lead to different sensitivity matrices if the set of SNPs is altered, (ii) the number of genotypes used, as the number of variables grows linearly, leading to numerical issues with models of larger sizes, (iii) the order in which the factors, u^i, v^{ref}, and e, of the genotype-specific flux distribution are estimated, (iv) the norm used to arrive at a sparse solution. Further, it is challenging to validate the predictions for the reference flux distribution, as it is a concept that is not tied with a particular genotype. Moreover, the SNPs are modelled as present/absent, and no distinction can be made between homozygous and heterozygote genotypes for a gene of interest. Therefore, refinements of this approach are needed to apply it in plant and crop breeding.

SNPeffect

Like the constraint-based approach above, SNPeffect aims to determine whether a SNP is functional or not by characterising its effect on reaction fluxes [49] (Fig. 4c). The flux of a reaction j in genotype i is assumed to follow mass-action-like kinetics while considering enzyme action [115]:

$$v_j^i\left(k_{\text{cat}}, \boldsymbol{K}, E, x\right) = {k_{\text{cat}}}_j^i E_j^i \eta(\boldsymbol{K}, x) = {V_{\max}}_j^i \prod_k \left(x_k^i\right)^{|N_{kj}|}.$$

As a result, the flux v_j^i can be expressed relative to a reference flux distribution as:

$$v_j^i = v_j^{\text{ref}} \frac{{V_{\max}}_j^i}{{V_{\max}}_j^{\text{ref}}} \prod_k \left(\frac{x_k^i}{x_k^{\text{ref}}}\right)^{|N_{kj}|}.$$

With measurements of available relative changes in maximal enzyme activities and metabolite levels with respect to a reference genotype, one can obtain lower and upper bounds. Deviation of the steady-state flux is then attributed to (positive/negative) additive effects of SNPs and saturation effects of the enzyme. There are three assumptions on which

SNPeffect is based: (i) a SNP is assumed to have consistent effect across all genotypes, i.e. it either increases or decreases reaction fluxes, (ii) the effect of a SNP is allowed to vary across genotypes, (iii) only nonsynonymous SNPs in genes included in the metabolic network are considered. Here, the effect of a SNP are simultaneously determined over all genotypes, by including constraints of steady-state and relative growth rate with respect to the reference genotype. Implementation of the approach clearly requires setting up a reference flux distribution or specifying lower and upper bounds, $v_{l,j}^{\mathrm{ref}}$ and $v_{u,j}^{\mathrm{ref}}$, for the fluxes in the reference genotype, resulting in the following constraint:

$$v_{l,j}^{\mathrm{ref}} \tilde{V}_{\max_j}^{\ i} \prod_k \left(\tilde{x}_k^i\right)^{|N_{kj}|} \leq v_j^i + \mathrm{dev}_{\mathrm{neg},j}^i - \mathrm{dev}_{\mathrm{pos},j}^i + \mathrm{SNP}_{\mathrm{pos},j}^i - \mathrm{SNP}_{\mathrm{neg},j}^i \leq v_{u,j}^{\mathrm{ref}} \tilde{V}_{\max_j}^{\ i} \prod_k \left(\tilde{x}_k^i\right)^{|N_{kj}|},$$

where $\tilde{V}_{\max_j}^{\ i}$ and $\tilde{x}_k^i$ are the relative maximal enzyme activity and relative metabolite content in genotype i with respect to the reference genotype, $\mathrm{dev}_{\mathrm{neg},j}^i$ and $\mathrm{dev}_{\mathrm{pos},j}^i$ denote deviations from the assumed enzyme kinetic and $\mathrm{SNP}_{\mathrm{pos},j}^i$ and $\mathrm{SNP}_{\mathrm{neg},j}^i$ are linear combinations of SNPs denoting their negative and positive, additive effects, respectively. In the actual implementation, these constraints are simplified by assuming that $\tilde{V}_{\max_j}^{\ i} = 1$.

Like in the structural sensitivity approach, above, SNPeffect can be regarded as a multi-locus GWAS in which the SNPs as present/absent, i.e. without making distinctions between different alleles. Its performance depends on: (i) the optimization function used, which in the existing implementation minimises the effects of the deviations from steady-state flux distribution that respect constraints from relative enzyme activities and metabolite levels, (ii) the reference flux distribution, determined by parsimonious FBA [15], and (iii) the number of metabolites and enzyme activities for which lower and upper bounds appearing in the expression above can be determined. In addition, SNPeffect inherits the factors that make its application challenging at a genome-scale level due to the sheer number of SNPs that can be considered. The approach was tested with models of Arabidopsis and *Populus trichocarpa* (poplar) [49], and identified functional SNPs in purine and amino acid biosynthesis pathways as well as lignin biosynthesis, respectively.

Approach based on genomic selection

Availability of flux distributions from a population of genotypes whose size is preclusive to conduct GWAS can still be used in GS for reaction fluxes. Tong et al. [56] developed an extension to GS, called netGS, based on integration of the machine-learning models of GS in a metabolic network. netGS relies on training a machine-learning model for steady-state fluxes obtained from genotype-specific metabolic models in particular conditions. The genotype-specific models are obtained by modifying the biomass function and applying constraints with respect to the growth relative to a reference phenotype. netGS is a four-step approach: (1) a model of a reference genotype is developed and is used to obtain a reference flux distribution following constraint-based approaches, like FBA [14]; (2) a flux distribution for another genotype is obtained by assuming that the difference to the reference is minimised, while ensuring that the ratio of predicted growth rates for the two accessions matches the ratio of measured fresh weights. This steps quantifies the flux of every reaction in the investigated genotypes; (3) each reaction flux is used as a trait for GS statistical modelling (implemented as rrBLUP), resulting a model with a specific predictability; (4) since the statistical models for each flux do not result in a steady-state flux distribution when applied to an unseen genotype, netGS next finds a flux distribution compatible with biochemical constraints given the flux predictions obtained from the statistical models based on the genomic data for the unseen genotype (Fig. 4d). In such a way, netGS allows prediction of growth, via the respective biomass reaction included in the model. This constraint-based approach has also been extended to consider predictions across environments. This extension is based on the assumption that the ratio between exchange fluxes for the reference genotype in two different environments is maintained across genotypes. With this additional constraint, the developed models in one environment can be used in another.

The statistical models that are devised in the third step of netGS inherits the shortcomings of GS models. However, through forcing these models to jointly respect physicochemical constraints, netGS aims to improve the model performance for unseen genotypes and in scenarios when there are large differences between training and testing populations. The imposing of these constraints can be regarded as adjusting for epistatic interactions between SNPs, which are otherwise difficult to integrate in a statistical framework due to the large number of SNPs considered. In contrast to the approaches above, netGS is not limited to investigating only nonsynonymous SNPs, but can also consider SNPs which lie in non-coding regions of the genome—which boosts the usage of genomic data. netGS was tested with 67 Arabidopsis accessions for which genotype- and condition-specific biomass reactions were developed based on measurements. The results showed that, in comparison to classical GS, it improves the prediction accuracy of growth within and

across nitrogen environments by 32.6% and 51.1%, respectively, as well as from optimal nitrogen to low carbon environment by 50.4%. The approach can readily be applied to any plant species for which metabolic models of high-quality exist and can be coupled with constraints from phenotypic data of specific genotypes.

Roadmap for future research

The brief review of the approaches for linking SNPs with metabolic and complex traits highlighted the division of two sets of approaches rooted in different methodologies. On one side, approaches for QTL mapping, GWAS, and genomic selection are solely based on statistics; moreover, genomic selection can be regarded as a black box, machine-learning approach that does not provide mechanistic insights or candidates for further testing. On the other hand, constraint-based approaches are applicable with large-scale models of metabolism and allow to establish a link between fine-grained metabolic processes and complex traits, such as biomass accumulation and growth.

Our systematic review indicated the possibility for merging the two complementary types of approaches to overcome their principle drawbacks, namely, the need for large populations, in the case of quantitative genetics approaches, and the need for depicting phenotypic diversity in a population of genotypes, in the case of the constraint-based modelling framework. While these approaches seem to have a great potential, demonstrating their added value necessitates addressing the following issues: first, studies should be planned to compare and contrast the findings between the purely statistical approaches and those based on consideration of SNPs in metabolic models. The existing studies have not performed this comparison due to the small sizes of the populations employed. Such comparative studies would require development of approaches for extraction of genotype-specific metabolic models for which no pipelines are yet freely available. Second, as shown on Fig. 2, there exist different metabolic models for the same plant species; these models different with respect to size, details, and modelled metabolic functionalities. Thus, it will also be important to investigate the effect of the model used for integration of genotypic data. Third, the consideration of SNP effect in constraint-based modelling can potentially introduce a lot of variables; thus, it is necessary to investigate how the preselection of SNPs may affect the findings from these approaches. In addition, since constraint-based approaches are marked with alternative solutions, one would have to design procedures to explore and/or further reduce the space of alternative solutions in a meaningful way.

The prospects for coupling mechanistic and statistical modelling approaches offer several new research avenues. First, one can aim to determine the statistical significance of a SNP effect obtained from constraint-based approaches. This can be accomplished by usage of permutation tests along with the aforementioned exploration of the space of alternative solutions. As a result, one would not need to rely on arbitrarily set threshold values to classify SNPs as functionally significant. Second, similar to netGS, one can use other types of machine-learning approaches for genomic selection to partition the reactions into active/inactive or into those carrying large or small fluxes, opening the possibility for other modelling directions. Third, with the availability of algorithmic procedures for estimation of turnover numbers of enzymes in a given genotype (e.g. *A. thaliana* Col-0 [116]), one can also aim to obtain such estimates in different genotypes, opening the possibility for using genetic mapping approaches and genomic selection. The resulting statistical models can, in turn, be employed to better constraint genotype-specific models using computational approaches, such as FBA with molecular crowding [117], MOMENT [118], or GECKO [119], or by incorporating macromolecular expression (so-called ME-models) [120] that are, however, still only applied to microbes.

We envision that these milestones can be achieved in the next 5–10 years of research in metabolic modelling of crops. Altogether, such prospects for a synergistic combination of machine-learning and metabolic models will pave the way for mechanistic modelling of complex traits in populations that involve both inbred and hybrid genotypes.

Acknowledgements ZN acknowledges the support from the Max Planck Society.

Author contributions ZN designed the systematic review. HT, AK and ZR contributed to writing the manuscript in figures creation. ZN wrote the manuscript. All the authors read, edited, and approved the final version of the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. ZN and HT acknowledge the funding from the European Union's Horizon 2020 research and innovation program, project PlantaSYST (SGA-CSA No. 739582 under FPA No. 664620) as well as the BG05M2OP001-1.003-001-C01 project, financed by the European Regional Development Fund through the Bulgarian 'Science and Education for Smart Growth' Operational Programme. ZN and AK acknowledge the funding from HFSP, project RGP0046.

Declarations

Consent for publication The authors give consent for publication.

Conflict of interest The authors declare no competing interests.

References

1. Mammadov J, Aggarwal R, Buyyarapu R, Kumpatla S (2012) SNP markers and their impact on plant breeding. Int J Plant Genom 2012:728398. https://doi.org/10.1155/2012/728398
2. Rafalski JA (2002) Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Sci 162:329–333. https://doi.org/10.1016/S0168-9452(01)00587-8
3. Rasheed A, Hao Y, Xia X et al (2017) Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. Mol Plant 10:1047–1064. https://doi.org/10.1016/j.molp.2017.06.008
4. Varshney RK, Graner A, Sorrells ME (2005) Genomics-assisted breeding for crop improvement. Trends Plant Sci 10:621–630. https://doi.org/10.1016/j.tplants.2005.10.004
5. Tong H, Nikoloski Z (2020) Machine learning approaches for crop improvement: leveraging phenotypic and genotypic big data. J Plant Physiol 257:153354. https://doi.org/10.1016/j.jplph.2020.153354
6. Stitt M, Sulpice R, Keurentjes J (2010) Metabolic networks: how to identify key components in the regulation of metabolism and growth. Plant Physiol 152:428–444. https://doi.org/10.1104/pp.109.150821
7. McMurry J, Fay RC, Robinson JK (2015) Chemistry. Pearson, Boston
8. Mahadevan R, Edwards JS, Doyle FJ (2002) Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. Biophys J 83:1331–1340. https://doi.org/10.1016/S0006-3495(02)73903-9
9. Basler G, Fernie AR, Nikoloski Z (2018) Advances in metabolic flux analysis toward genome-scale profiling of higher organisms. Biosci Rep 38:BSR20170224. https://doi.org/10.1042/BSR20170224
10. Antoniewicz MR (2015) Methods and advances in metabolic flux analysis: a mini-review. J Ind Microbiol Biotechnol 42:317–325. https://doi.org/10.1007/s10295-015-1585-x
11. Ma F, Jazmin LJ, Young JD, Allen DK (2014) Isotopically nonstationary 13C flux analysis of changes in *Arabidopsis thaliana* leaf metabolism due to high light acclimation. Proc Natl Acad Sci USA 111:16967–16972. https://doi.org/10.1073/pnas.1319485111
12. Szecowka M, Heise R, Tohge T et al (2013) Metabolic fluxes in an illuminated Arabidopsis rosette. Plant Cell 25:694–714. https://doi.org/10.1105/tpc.112.106989
13. O'Brien EJ, Monk JM, Palsson BO (2015) Using genome-scale models to predict biological capabilities. Cell 161:971–987. https://doi.org/10.1016/j.cell.2015.05.019
14. Orth JD, Thiele I, Palsson BØ (2010) What is flux balance analysis? Nat Biotechnol 28:245–248. https://doi.org/10.1038/nbt.1614
15. Lewis NE, Hixson KK, Conrad TM et al (2010) Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. Mol Syst Biol 6:390. https://doi.org/10.1038/msb.2010.47
16. Nikoloski Z, Perez-Storey R, Sweetlove LJ (2015) Inference and prediction of metabolic network fluxes. Plant Physiol 169:1443–1455. https://doi.org/10.1104/pp.15.01082
17. Küken A, Nikoloski Z (2019) Computational approaches to design and test plant synthetic metabolic pathways. Plant Physiol 179:894–906. https://doi.org/10.1104/pp.18.01273
18. Thiele I, Palsson BØ (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 5:93–121. https://doi.org/10.1038/nprot.2009.203
19. Feist AM, Palsson BO (2010) The biomass objective function. Curr Opin Microbiol 13:344–349. https://doi.org/10.1016/j.mib.2010.03.003
20. Estévez RS, Nikoloski Z (2014) Generalized framework for context-specific metabolic model extraction methods. Front Plant Sci 5:491. https://doi.org/10.3389/fpls.2014.00491
21. Arnold A, Nikoloski Z (2014) Bottom-up metabolic reconstruction of Arabidopsis and its application to determining the metabolic costs of enzyme production. Plant Physiol 165:1380–1391. https://doi.org/10.1104/pp.114.235358
22. Cheung CYM, Poolman MG, Fell DA et al (2014) A diel flux balance model captures interactions between light and dark metabolism during day-night cycles in C3 and crassulacean acid metabolism leaves. Plant Physiol 165:917–929. https://doi.org/10.1104/pp.113.234468
23. Seaver SMD, Bradbury LMT, Frelin O et al (2015) Improved evidence-based genome-scale metabolic models for maize leaf, embryo, and endosperm. Front Plant Sci 6:142. https://doi.org/10.3389/fpls.2015.00142
24. Cheung CYM, Williams TCR, Poolman MG et al (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 75:1050–1061. https://doi.org/10.1111/tpj.12252
25. de Oliveira Dal'Molin CG, Quek LE, Palfreyman RW et al (2010) AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. Plant Physiol 152:579–589. https://doi.org/10.1104/pp.109.148817
26. de Oliveira Dal'Molin CG, Quek LE, Saa PA, Nielsen LK (2015) A multi-tissue genome-scale metabolic modeling framework for the analysis of whole plant systems. Front Plant Sci 6:4. https://doi.org/10.3389/fpls.2015.00004
27. Mintz-Oron S, Meir S, Malitsky S et al (2012) Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity. Proc Natl Acad Sci USA 109:339–344. https://doi.org/10.1073/pnas.1100358109
28. Poolman MG, Miguet L, Sweetlove LJ, Fell DA (2009) A genome-scale metabolic model of Arabidopsis and some of its properties. Plant Physiol 151:1570–1581. https://doi.org/10.1104/pp.109.141267
29. Robaina-Estévez S, Daloso DM, Zhang Y et al (2017) Resolving the central metabolism of Arabidopsis guard cells. Sci Rep 7:8307. https://doi.org/10.1038/s41598-017-07132-9
30. Saha R, Suthers PF, Maranas CD (2011) Zea mays irs1563: a comprehensive genome-scale metabolic reconstruction of maize metabolism. PLoS ONE 6:e21784. https://doi.org/10.1371/journal.pone.0021784
31. Scheunemann M, Brady SM, Nikoloski Z (2018) Integration of large-scale data for extraction of integrated Arabidopsis root cell-type specific models. Sci Rep 8:7919. https://doi.org/10.1038/s41598-018-26232-8
32. Chatterjee A, Huma B, Shaw R, Kundu S (2017) Reconstruction of *Oryza sativa* indica genome scale metabolic model and its responses to varying RuBisCO activity, light intensity, and enzymatic cost conditions. Front Plant Sci 8:2060. https://doi.org/10.3389/fpls.2017.02060

33. Chatterjee A, Kundu S (2015) Revisiting the chlorophyll biosynthesis pathway using genome scale metabolic model of *Oryza sativa* japonica. Sci Rep 5:14975. https://doi.org/10.1038/srep14975
34. Lakshmanan M, Lim SH, Mohanty B et al (2015) Unraveling the light-specific metabolic and regulatory signatures of rice through combined in silico modeling and multiomics analysis. Plant Physiol 169:3002–3020. https://doi.org/10.1104/pp.15.01379
35. Lakshmanan M, Mohanty B, Lee DY (2013) Identifying essential genes/reactions of the rice photorespiration by in silico model-based analysis. Rice 6:20. https://doi.org/10.1186/1939-8433-6-20
36. Poolman MG, Kundu S, Shaw R, Fell DA (2013) Responses to light intensity in a genome-scale model of rice metabolism. Plant Physiol 162:1060–1072. https://doi.org/10.1104/pp.113.216762
37. Shaw R, Kundu S (2015) Flux balance analysis of genome-scale metabolic model of rice (*Oryza sativa*): aiming to increase biomass. J Biosci 40:819–828. https://doi.org/10.1007/s12038-015-9563-z
38. Bogart E, Myers CR (2016) Multiscale metabolic modeling of C4 plants: connecting nonlinear genome-scale models to leaf-scale metabolism in developing maize leaves. PLoS ONE 11:e0151722. https://doi.org/10.1371/journal.pone.0151722
39. de Oliveira Dal'Molin CG, Quek LE, Palfreyman RW et al (2010) C4GEM, a genome-scale metabolic model to study C4 plant metabolism. Plant Physiol 154:1871–1885. https://doi.org/10.1104/pp.110.166488
40. 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. https://doi.org/10.1104/pp.114.245787
41. Yuan H, Cheung CYM, Poolman MG et al (2016) A genome-scale metabolic network reconstruction of tomato (*Solanum lycopersicum* L.) and its application to photorespiratory metabolism. Plant J 85:289–304. https://doi.org/10.1111/tpj.13075
42. Botero K, Restrepo S, Pinzón A (2018) A genome-scale metabolic model of potato late blight suggests a photosynthesis suppression mechanism. BMC Genom 19:863. https://doi.org/10.1186/s12864-018-5192-x
43. Grafahrend-Belau E, Schreiber F, Koschützki D, Junker BH (2009) Flux balance analysis of barley seeds: a computational approach to study systemic properties of central metabolism. Plant Physiol 149:585–598. https://doi.org/10.1104/pp.108.129635
44. Hay J, Schwender J (2011) Metabolic network reconstruction and flux variability analysis of storage synthesis in developing oilseed rape (*Brassica napus* L.) embryos. Plant J 67:526–541. https://doi.org/10.1111/j.1365-313X.2011.04613.x
45. Pilalis E, Chatziioannou A, Thomasset B, Kolisis F (2011) An in silico compartmentalized metabolic model of *Brassica napus* enables the systemic study of regulatory aspects of plant central metabolism. Biotechnol Bioeng 108:1673–1682. https://doi.org/10.1002/bit.23107
46. Pfau T, Christian N, Masakapalli SK et al (2018) The intertwined metabolism during symbiotic nitrogen fixation elucidated by metabolic modelling. Sci Rep 8:12504. https://doi.org/10.1038/s41598-018-30884-x
47. Moreira TB, Shaw R, Luo X et al (2019) A genome-scale metabolic model of soybean (*Glycine max*) highlights metabolic fluxes in seedlings. Plant Physiol 180:1912–1929. https://doi.org/10.1104/pp.19.00122
48. Shaw R, Maurice Cheung CY (2019) A mass and charge balanced metabolic model of *Setaria viridis* revealed mechanisms of proton balancing in C4 plants. BMC Bioinform 20:357. https://doi.org/10.1186/s12859-019-2941-z
49. Sarkar D, Maranas CD (2020) SNPeffect: identifying functional roles of SNPs using metabolic networks. Plant J 103:512–531. https://doi.org/10.1111/tpj.14746
50. Tan XLJ, Cheung CYM (2020) A multiphase flux balance model reveals flexibility of central carbon metabolism in guard cells of C3 plants. Plant J 104:1648–1656. https://doi.org/10.1111/tpj.15027
51. Shameer S, Baghalian K, Cheung CYM et al (2018) Computational analysis of the productivity potential of CAM. Nat Plants 4:165–171. https://doi.org/10.1038/s41477-018-0112-2
52. Töpfer N, Braam T, Shameer S et al (2020) Alternative crassulacean acid metabolism modes provide environment-specific water-saving benefits in a leaf metabolic model. Plant Cell 32:3689–3705. https://doi.org/10.1105/tpc.20.00132
53. Blätke MA, Bräutigam A (2019) Evolution of C4 photosynthesis predicted by constraint-based modelling. Elife 8:e49305. https://doi.org/10.1101/670547
54. Correa SM, Fernie AR, Nikoloski Z, Brotman Y (2020) Towards model-driven characterization and manipulation of plant lipid metabolism. Prog Lipid Res 80:101051. https://doi.org/10.1016/j.plipres.2020.101051
55. Cañas RA, Yesbergenova-Cuny Z, Simons M et al (2017) Exploiting the genetic diversity of maize using a combined metabolomic, enzyme activity profiling, and metabolic modeling approach to link leaf physiology to kernel yield. Plant Cell 29:919–943. https://doi.org/10.1105/tpc.16.00613
56. Tong H, Küken A, Nikoloski Z (2020) Integrating molecular markers into metabolic models improves genomic selection for Arabidopsis growth. Nat Commun 11:2410. https://doi.org/10.1038/s41467-020-16279-5
57. Huang X (2016) From genetic mapping to molecular breeding: genomics have paved the highway. Mol Plant 9:959–960. https://doi.org/10.1016/j.molp.2016.06.001
58. Burghardt LT, Young ND, Tiffin P (2017) A guide to genome-wide association mapping in plants. Curr Protoc Plant Biol 2:22–38. https://doi.org/10.1002/cppb.20041
59. Korte A, Ashley F (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9:29. https://doi.org/10.1186/1746-4811-9-29
60. Worley B, Powers R (2013) Multivariate analysis in metabolomics. Curr Metabolomics 1:92–107. https://doi.org/10.2174/2213235x130108
61. Mitchell-Olds T (2010) Complex-trait analysis in plants. Genome Biol 11:113. https://doi.org/10.1186/gb-2010-11-4-113
62. Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539–551. https://doi.org/10.1534/genetics.107.074245
63. Huang BE, Verbyla KL, Verbyla AP et al (2015) MAGIC populations in crops: current status and future prospects. Theor Appl Genet 128:999–1017. https://doi.org/10.1007/s00122-015-2506-0
64. Xiao Y, Tong H, Yang X et al (2016) Genome-wide dissection of the maize ear genetic architecture using multiple populations. New Phytol 210:1095–1106. https://doi.org/10.1111/nph.13814
65. Zeng Z-B (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1468
66. Brachi B, Morris GP, Borevitz JO (2011) Genome-wide association studies in plants: the missing heritability is in the field. Genome Biol 12:232. https://doi.org/10.1186/gb-2011-12-10-232
67. Liu HJ, Yan J (2019) Crop genome-wide association study: a harvest of biological relevance. Plant J 97:8–18. https://doi.org/10.1111/tpj.14139

68. Sharmaa A, Lee JS, Dang CG et al (2015) Stories and challenges of genome wide association studies in livestock—a review. Asian-Australas J Anim Sci 28:1371–1379. https://doi.org/10.5713/ajas.14.0715
69. Huang X, Han B (2014) Natural variations and genome-wide association studies in crop plants. Annu Rev Plant Biol 65:531–551. https://doi.org/10.1146/annurev-arplant-050213-035715
70. Yu J, Pressoir G, Briggs WH et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208. https://doi.org/10.1038/ng1702
71. Price AL, Patterson NJ, Plenge RM et al (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909. https://doi.org/10.1038/ng1847
72. Kang HM, Zaitlen NA, Wade CM et al (2008) Efficient control of population structure in model organism association mapping. Genetics 178:1709–1723. https://doi.org/10.1534/genetics.107.080101
73. Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. Nat Genet 44:821–824. https://doi.org/10.1038/ng.2310
74. Lippert C, Listgarten J, Liu Y et al (2011) FaST linear mixed models for genome-wide association studies. Nat Methods 8:833–835. https://doi.org/10.1038/nmeth.1681
75. Zhang Z, Ersoz E, Lai CQ et al (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42:355–360. https://doi.org/10.1038/ng.546
76. Segura V, Vilhjálmsson BJ, Platt A et al (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44:825–830. https://doi.org/10.1038/ng.2314
77. Crossa J, Pérez-Rodríguez P, Cuevas J et al (2017) Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci 22:961–975. https://doi.org/10.1016/j.tplants.2017.08.011
78. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829
79. Desta ZA, Ortiz R (2014) Genomic selection: genome-wide prediction in plant improvement. Trends Plant Sci 19:592–601. https://doi.org/10.1016/j.tplants.2014.05.006
80. Nakaya A, Isobe SN (2012) Will genomic selection be a practical method for plant breeding? Ann Bot 110:1303–1316. https://doi.org/10.1093/aob/mcs109
81. Wang X, Xu Y, Hu Z, Xu C (2018) Genomic selection methods for crop improvement: current status and prospects. Crop J 6:330–340. https://doi.org/10.1016/j.cj.2018.03.001
82. Berro I, Lado B, Nalin RS et al (2019) Training population optimization for genomic selection. Plant Genome 12:190028. https://doi.org/10.3835/plantgenome2019.04.0028
83. Norman A, Taylor J, Edwards J, Kuchel H (2018) Optimising genomic selection in wheat: effect of marker density, population size and population structure on prediction accuracy. G3 Genes Genomes Genet 8:2889–2899. https://doi.org/10.1534/g3.118.200311
84. Burgueño J, Crossa J, Cotes JM et al (2011) Prediction assessment of linear mixed models for multienvironment trials. Crop Sci 51:944–954. https://doi.org/10.2135/cropsci2010.07.0403
85. Wang D, Salah El-Basyoni I, Stephen Baenziger P et al (2012) Prediction of genetic values of quantitative traits with epistatic effects in plant breeding populations. Heredity (Edinb) 109:313–319. https://doi.org/10.1038/hdy.2012.44
86. Jiang Y, Reif JC (2015) Modeling epistasis in genomic selection. Genetics 201:759–768. https://doi.org/10.1534/genetics.115.177907
87. Hu Z, Li Y, Song X et al (2011) Genomic value prediction for quantitative traits under the epistatic model. BMC Genet 12:15. https://doi.org/10.1186/1471-2156-12-15
88. Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics 194:597–607. https://doi.org/10.1534/genetics.113.152207
89. Le Signor C, Aimé D, Bordat A et al (2017) Genome-wide association studies with proteomics data reveal genes important for synthesis, transport and packaging of globulins in legume seeds. New Phytol 214:1597–1613. https://doi.org/10.1111/nph.14500
90. Luo J (2015) Metabolite-based genome-wide association studies in plants. Curr Opin Plant Biol 24:31–38. https://doi.org/10.1016/j.pbi.2015.01.006
91. Kacser H, Burns JA (1981) The molecular basis of dominance. Genetics 97:639–666
92. Celton M, Goelzer A, Camarasa C et al (2012) A constraint-based model analysis of the metabolic consequences of increased NADPH oxidation in Saccharomyces cerevisiae. Metab Eng 14:366–379. https://doi.org/10.1016/j.ymben.2012.03.008
93. Eder M, Nidelet T, Sanchez I et al (2020) QTL mapping of modelled metabolic fluxes reveals gene variants impacting yeast central carbon metabolism. Sci Rep 10:2162. https://doi.org/10.1038/s41598-020-57857-3
94. Stitt M, Gibon Y (2014) Why measure enzyme activities in the era of systems biology? Trends Plant Sci 19:256–265. https://doi.org/10.1016/j.tplants.2013.11.003
95. Causse M, Rocher JP, Henry AM et al (1995) Genetic dissection of the relationship between carbon metabolism and early growth in maize, with emphasis on key-enzyme loci. Mol Breed 1:259–272. https://doi.org/10.1007/BF02277426
96. Pelleschi S, Leonardi A, Rocher J-P et al (2006) Analysis of the relationships between growth, photosynthesis and carbohydrate metabolism using quantitative trait loci (QTLs) in young maize plants subjected to water deprivation. Mol Breed 17:21–39. https://doi.org/10.1007/s11032-005-1031-2
97. Prioul JL, Pelleschi S, Séne M et al (1999) From QTLs for enzyme activity to candidate genes in maize. J Exp Bot 50:1281–1288. https://doi.org/10.1093/jxb/50.337.1281
98. Thévenot C, Simond-Côte E, Reyss A et al (2005) QTLs for enzyme activities and soluble carbohydrates involved in starch accumulation during grain filling in maize. J Exp Bot 56:945–958. https://doi.org/10.1093/jxb/eri087
99. Limami AM, Rouillon C, Glevarec G et al (2002) Genetic and physiological analysis of germination efficiency in maize in relation to nitrogen metabolism reveals the importance of cytosolic glutamine synthetase. Plant Physiol 130:1860–1870. https://doi.org/10.1104/pp.009647
100. Zhang N, Gibon Y, Gur A et al (2010) Fine quantitative trait loci mapping of carbon and nitrogen metabolism enzyme activities and seedling biomass in the maize IBM mapping population. Plant Physiol 154:1753–1765. https://doi.org/10.1104/pp.110.165787
101. Mitchell-Olds T, Pedersen D (1998) The molecular basis of quantitative genetic variation in central and secondary metabolism in *Arabidopsis*. Genetics 149:739–747
102. Sergeeva LI, Vonk J, Keurentjes JJB et al (2004) Histochemical analysis reveals organ-specific quantitative trait loci for enzyme activities in *Arabidopsis*. Plant Physiol 134:237–245. https://doi.org/10.1104/pp.103.027615
103. Sergeeva LI, Keurentjes JJB, Bentsink L et al (2006) Vacuolar invertase regulates elongation of *Arabidopsis thaliana* roots as revealed by QTL and mutant analysis. Proc Natl Acad Sci USA 103:2994–2999. https://doi.org/10.1073/pnas.0511015103
104. Keurentjes JJ, Sulpice R, Gibon Y et al (2008) Integrative analyses of genetic variation in enzyme activities of primary carbohydrate metabolism reveal distinct modes of regulation in

Arabidopsis thaliana. Genome Biol 9:R129. https://doi.org/10.1186/gb-2008-9-8-r129
105. Steinhauser MC, Steinhauser D, Gibon Y et al (2011) Identification of enzyme activity quantitative trait loci in a *Solanum lycopersicum* x *Solanum pennellii* introgression line population. Plant Physiol 157:998–1014. https://doi.org/10.1104/pp.111.181594
106. Fusari CM, Kooke R, Lauxmann MA et al (2017) Genome-wide association mapping reveals that specific and pleiotropic regulatory mechanisms fine-tune central metabolism and growth in Arabidopsis. Plant Cell 29:2349–2373. https://doi.org/10.1105/tpc.17.00232
107. Leveson-Gower RB, Mayer C, Roelfes G (2019) The importance of catalytic promiscuity for enzyme design and evolution. Nat Rev Chem 3:687–705. https://doi.org/10.1038/s41570-019-0143-x
108. Razaghi-Moghadam Z, Nikoloski Z (2020) GeneReg: a constraint-based approach for design of feasible metabolic engineering strategies at the gene level. Bioinformatics. https://doi.org/10.1093/bioinformatics/btaa996
109. Jamshidi N, Palsson BØ (2006) Systems biology of SNPs. Mol Syst Biol 2:38. https://doi.org/10.1038/msb4100077
110. Thiele I, Price ND, Vo TD, Palsson BØ (2005) Candidate metabolic network states in human mitochondria. Impact of diabetes, ischemia, and diet. J Biol Chem 280:11683–11695. https://doi.org/10.1074/jbc.M409072200
111. Burgard AP, Nikolaev EV, Schilling CH, Maranas CD (2004) Flux coupling analysis of genome-scale metabolic network reconstructions. Genome Res 14:301–312. https://doi.org/10.1101/gr.1926504
112. Basler G, Nikoloski Z, Larhlimi A et al (2016) Control of fluxes in metabolic networks. Genome Res 26:956–968. https://doi.org/10.1101/gr.202648.115
113. Øyås O, Borrell S, Trauner A et al (2020) Model-based integration of genomics and metabolomics reveals SNP functionality in *Mycobacterium tuberculosis*. Proc Natl Acad Sci USA 117:8494–8502. https://doi.org/10.1073/pnas.1915551117
114. Uhr M, Stelling J (2008) Structural sensitivity analysis of metabolic networks. IFAC Proc 41:15879–15884. https://doi.org/10.3182/20080706-5-kr-1001.02684
115. Sajitz-Hermstein M, Töpfer N, Kleessen S et al (2016) iReMetflux: constraint-based approach for integrating relative metabolite levels into a stoichiometric metabolic models. Bioinformatics 32:i755–i762. https://doi.org/10.1093/bioinformatics/btw465
116. Küken A, Gennermann K, Nikoloski Z (2020) Characterization of maximal enzyme catalytic rates in central metabolism of *Arabidopsis thaliana*. Plant J 103:2168–2177. https://doi.org/10.1111/tpj.14890
117. Beg QK, Vazquez A, Ernst J et al (2007) Intracellular crowding defines the mode and sequence of substrate uptake by *Escherichia coli* and constrains its metabolic activity. Proc Natl Acad Sci USA 104:12663–12668. https://doi.org/10.1073/pnas.0609845104
118. Adadi R, Volkmer B, Milo R et al (2012) Prediction of microbial growth rate versus biomass yield by a metabolic network with kinetic parameters. PLoS Comput Biol 8:e1002575. https://doi.org/10.1371/journal.pcbi.1002575
119. Sánchez BJ, Zhang C, Nilsson A et al (2017) Improving the phenotype predictions of a yeast genome-scale metabolic model by incorporating enzymatic constraints. Mol Syst Biol 13:935. https://doi.org/10.15252/msb.20167411
120. Fang X, Lloyd CJ, Palsson BO (2020) Reconstructing organisms in silico: genome-scale models and their emerging applications. Nat Rev Microbiol 18:731–743. https://doi.org/10.1038/s41579-020-00440-4