Metabolomic selection–based machine learning improves fruit taste prediction

Alisdair R. Fernie[a,b,1] and Saleh Alseekh[a,b]

At least with regard to horticultural crops, we are currently experiencing a step change in crop breeding targets. Historically, breeders focused on high-yielding, resilient varieties; however, this has led to considerable dissatisfaction with modern varieties of fruits and vegetables (1). The recent increasing willingness of consumers to pay a premium for quality is, however, driving a renaissance of breeding for quality traits. That said, flavor is a highly complex composite trait made up of the interactions between the chemical composition of the crop as well as the taste, olfaction, and psychology of the consumer (2, 3). In recent years, flavor has been assessed by costly consumer sensory panels or by breeders themselves in the field. Both approaches have disadvantages. Field evaluation, while allowing the evaluation of many varieties in a day, is highly subjective and error prone. Although population-based sensory panels are well established and accurate, they are difficult to scale to large breeding programs. These limitations in flavor phenotyping are elegantly addressed via the employment of metabolomic selection–based machine learning in the report by Colantonio et al. (4).

Machine learning has been gaining increasing traction as a means to analyze various high-throughput phenotyping applications, which enable researchers to identify meaningful patterns in relevant plant data. For example, it has proven utility in two-dimensional light imaging as a proxy for plant biomass, reflectance ratios as proxies for yield, hyperspectral reflectance as proxies for leaf chlorophyll and nitrogen content, and canopy temperatures as proxies for the drought response (reviewed in ref. 5). It is additionally already being used in breeding, particularly in the form of genomic selection, in which genome-wide marker data are used to predict the genetic value of an unobserved candidate in a breeding population via estimating the effects of all markers (6). A recent refinement of this approach—genome optimization via virtual simulation—simulates a virtual genome encompassing the most abundant advantageous alleles in a genetic pool, thereby helping plot the optimal route for breeding (7).

In their study, Colantonio et al. (4) used a combination of metabolomic profiles and consumer taste panel information to train machine learning models such that they can predict how flavorsome a fruit will be from its chemical composition. To do so, they took target metabolomic profiling and consumer panel ratings from previous studies in tomato and blueberry (1, 3, 8) and used a suite of 18 different statistical and machine learning models to predict various taste sensations, including liking, sweet, sour, and taste intensity. The data used encompassed sugars, acids, volatiles, and the taste sensations mentioned above (as well as umami in the case of tomato). As a first approach, the metabolites were partitioned according to compound class. Interestingly, when the results of the consumer tests were assessed following this partitioning, it was apparent that the proportion of variance of each trait that was explained by the sugars and acids varied across the flavor attributes as well as between species. For instance, while sugars and acids predominantly explain blueberry sweetness, the volatiles were the main contributors to this trait in tomato. Colantonio et al. (4) next applied 18 statistical and machine learning methods to predict sensory traits from the metabolite levels; the highest prediction accuracies were observed for the XGBoost library, gradient boosting machines, and random forest models, with the XGBoost model recording accuracies of 0.62 to 0.87 across all traits and in both species. Using these approaches, sweetness, flavor intensity, and sourness were the most predictable traits in tomato, and sourness and sweetness were the most predictable traits in blueberry.

While the above-described studies demonstrated the utility of metabolite data for the prediction of taste, metabolomics data are far less easily obtained (both experimentally and with regard to access from

[a]Department of Molecular Physiology, Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam-Golm, Germany; and [b]Department of Metabolomics, Center for Plant Systems Biology and Biotechnology, 4000 Plovdiv, Bulgaria

Author contributions: A.R.F. and S.A. wrote the paper.

The authors declare no competing interest.

See companion article, “Metabolomic selection for enhanced fruit flavor,” 10.1073/pnas.2115865119.

[1]To whom correspondence may be addressed. Email: fernie@mpimp-golm.mpg.de.

Published February 25, 2022.

databases) than genomics data. That said, using available data encompassing whole-genome sequencing data, chemical profiles, and sensory panel data for 70 varieties of tomato (1) allowed Colantonio et al. (4) to evaluate the prediction potential of metabolomic and genomic selections. In order to do so, they used the genomic best linear unbiased prediction method (9) to predict the consumer sensory ratings from a subset of almost 80,000 single-nucleotide polymorphisms as well as metabolomic information from the same 70 varieties to predict the panel ratings. Metabolomic selection was found to greatly outperform the genomic selection in the prediction of all traits, thereby highlighting the potential of this approach to support breeding (Fig. 1). Indeed, relatively high accuracies for certain traits could be obtained when using the metabolome data from as few as 50 individuals. Finally, it was demonstrated that BayesA and gradient booster machines were able to identify which sugars, acids, or volatiles enhanced or suppressed consumer sensory perceptions of flavor. These analyses revealed that, for example in tomato, glucose and fructose are the most important sensory perception enhancers, while the volatiles 1-penten-3-one and 2-phenylethanol as well as E-2-pentenal and 4-carene were also important for sweetness with a different (although sometimes overlapping) set of metabolites influencing these sensory perceptions in blueberry.

As described by the authors, the use of metabolomic selection is an excellent complement to a molecular breeding program since this enables quantitative trait loci (QTL) mapping or genome-wide association studies. As such, the flavor-related metabolites identified by metabolomic selection could be used to identify the causal genes (or at least the genetic locus) that influence their abundance level and create markers for molecular breeding. Indeed, a legion of studies has been published across a wide range of crop species in which metabolic QTL or gene associations have been assessed (10, 11). That said, while many of these have additionally assessed yield-associated traits, the linkage of this information to the perception of taste or alternatively, to health benefits following consumption remains relatively scarce.

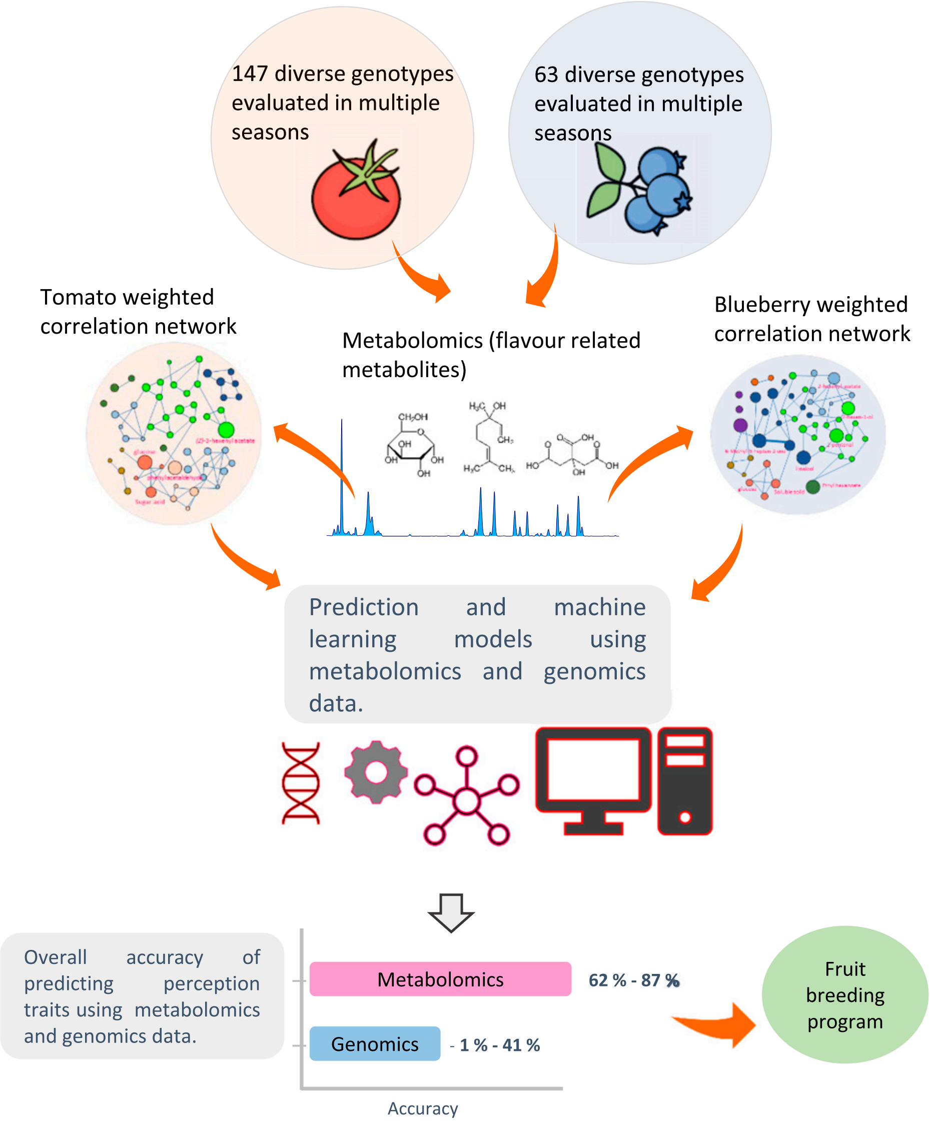
147 diverse genotypes evaluated in multiple seasons
63 diverse genotypes evaluated in multiple seasons
Tomato weighted correlation network
Metabolomics (flavour related metabolites)
Blueberry weighted correlation network
Prediction and machine learning models using metabolomics and genomics data.
Overall accuracy of predicting perception traits using metabolomics and genomics data.
Metabolomics
62 % - 87 %
Genomics
- 1 % - 41 %
Accuracy
Fruit breeding program

Fig. 1. Metabolic-based prediction model in tomato and blueberry. The weighted correlation network analysis of tomato and blueberry metabolites and their clusters is based on biochemical classification. Eighteen statistical and machine learning methods were used to predict fruit flavor based on its chemistry.

Given that metabolite traits often display large variability and low heritability and are subject to complex interaction effects (12, 13), gaining success in modifying compositional traits involved in either taste or nutrition relies deeply on identifying the correct metabolite targets. The integration of metabolomic selection with marker-assisted selection provides a powerful route to ensure that this identification is correct. While linear regression, random forest models, and partial least square regression methods have been recorded to have variable success in predicting flavor (12, 14, 15), the machine learning models employed by Colantonio et al. (4) displayed superior performance over these methods, irrespective of the fruit or trait to which they were applied.

The idea of metabolomics-assisted breeding has long been postulated (16); however, for this purpose, its cost was often deemed an insurmountable barrier. Taken alongside the rapid development of metabolomics technologies and in particularly, their increased coverage (17), the results of Colantonio et al. (4) suggest that when coupled with machine learning, this may no longer be the case. Computation will be a major driver in this process as well as in the perspective use of this approach to improve dietary-based aspects of human health. It can be anticipated that these advances will pave the way to knowledge-informed breeding of both tastier and more nutritious food.

1 D. Tieman *et al.*, A chemical genetic roadmap to improved tomato flavor. *Science* **355**, 391–394 (2017).

2 G. M. Shepherd, Smell images and the flavour system in the human brain. *Nature* **444**, 316–321 (2006).

3 D. Tieman *et al.*, The chemical interactions underlying tomato flavor preferences. *Curr. Biol.* **22**, 1035–1039 (2012).

4 V. Colantonio *et al.*, Metabolomic selection for enhanced fruit flavor. *Proc. Natl. Acad. Sci. U.S.A.*, 10.1073/pnas.2115865119 (2022).

5 A. D. J. van Dijk, G. Kootstra, W. Kruijer, D. de Ridder, Machine learning in plant science and plant breeding. *iScience* **24**, 101890 (2020).

6 R. K. Varshney *et al.*, Fast-forward breeding for a food-secure world. *Trends Genet.* **37**, 1124–1136 (2021).

7 Q. Cheng *et al.*, Genome optimization via virtual simulation to accelerate maize hybrid breeding. *Brief. Bioinform.* **23**, bbab447 (2022).

8 L. F. V. Ferrão *et al.*, Genome-wide association of volatiles reveals candidate loci for blueberry flavor. *New Phytol.* **226**, 1725–1737 (2020).

9 P. M. VanRaden, Efficient methods to compute genomic predictions. *J. Dairy Sci.* **91**, 4414–4423 (2008).

10 C. Fang, J. Luo, Metabolic GWAS-based dissection of genetic bases underlying the diversity of plant metabolism. *Plant J.* **97**, 91–100 (2019).

11 A. R. Fernie, T. Tohge, The genetics of plant metabolism. *Annu. Rev. Genet.* **51**, 287–310 (2017).

12 C. Eggink *et al.*, Predictio of sweet pepper (*Capsicum annuum*) flavor over different harvests. *Euphytica* **187**, 117–131 (2012).

13 A. B. Kouassi *et al.*, Estimation of genetic parameters and prediction of breeding values for apple fruit-quality traits using pedigreed plant material in Europe. *Tree Genet. Genomes* **5**, 659–672 (2009).

14 E. G. Abegaz, K. S. Tandon, J. W. Scott, E. A. Baldwin, R. L. Shewfelt, Partitioning taste from aromatic flavor notes of frech tomato (*Lycopersicum esculentum*, Mill) to develop predictive models as a function of volatile and nonvolatile compoents. *Postharvest Biol. Technol.* **34**, 227–235 (2004).

15 J. L. Gilbert *et al.*, Identifying breeding priorities for blueberry flavor using biochemical, sensory, and genotype by environment analyses. *PLoS One* **10**, e0138494 (2015).

16 A. R. Fernie, N. Schauer, Metabolomics-assisted breeding: A viable option for crop improvement? *Trends Genet.* **25**, 39–48 (2009).

17 S. Alseekh *et al.*, Mass spectrometry-based metabolomics: A guide for annotation, quantification and best reporting practices. *Nat. Methods* **18**, 747–756 (2021).