Genomics-based annotations help unveil the molecular composition of edible plants

Shany Ofaim[1], Giulia Menichetti[1,2], Michael Sebek[1] and Albert László Barabási[1,2,3*]

[1]Network Science Institute and Department of Physics, Northeastern University, Boston, USA; [2]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, USA; [3]Department of Network and Data Science, Central European University, Budapest, Hungary.

Corresponding author

Correspondence to: a.barabasi@northeastern.edu

Abstract

Given the important role food plays in health and well-being, the past decades have seen considerable experimental efforts dedicated to mapping the chemical composition of food ingredients. As the composition of raw food is genetically predetermined, here we ask, to what degree can we rely on genome-associated metabolic annotations to predict the chemical composition of natural ingredients? To that end, we developed an approach to unveil the chemical composition of 75 edible plants' genomes, finding that genome-associated functional annotations substantially increase the number of compounds linked to specific plants. To account for the gap between the metabolic potential represented by functional annotations and the compounds likely to be accumulated and detected experimentally, we employed a predictive thermodynamic feasibility approach to systematically identify compounds that are likely to accumulate in plants. To quantify the accuracy of our predictions, we performed untargeted metabolomics on 13 plants, allowing us to experimentally confirm the detectability of the

predicted compounds. For example, we observed 59 potentially novel compounds predicted and experimentally detected in corn. Our study shows that the application of genome-associated functional annotations can lead to an integrated metabologenomics platform capable of unveiling the chemical composition of edible plants and the biochemical pathways responsible for the observed compounds.

Author Summary

Nutrition and well-being take a central role in today's high pace life, but how much do we really know about the food we eat? Here, we harness existing metabolic knowledge encrypted in staple food ingredients' genome to help us explore the composition of raw edible plants. We first show the benefit and value of looking into genome-associated functional annotations on a wide scale. Next, we rely on new experimental data to develop a framework that helps us reveal new, potentially bioactive compounds in staple food ingredients. This has significance in, first, extending current food composition knowledge and second in discovering newly detected bioactive compounds, shedding light on the potential impacts of common food ingredients beyond their nutritional value as described in food labels. Finally, we show that staple foods that are already included in our daily diets might have the potential to contribute to our well-being.

Background

"Make every bite count" is one of the main guidelines the U.S Departments of Agriculture (USDA) Dietary recommends for Americans (2020-2025)[1], reminding us of the multiple roles of the food we consume, and how it contributes to our well-being. Fruit and

47 vegetables, serve as a source of energy and nutrients, modulating our health, and affecting
48 disease [2–4]. Plants are complex organisms characterized by large genomes. For example, the
49 corn genome (2,280 Mb) encodes 57,181 proteins, annotated to Gene Ontology (GO) categories,
50 helping unveil the biological processes these proteins participate in and their potential molecular
51 functions [5]. The genome can also be a strong predictor of phenotypic traits [6], like color, taste
52 and aroma, known indicators of the nutritional value of a plant [7]. Indeed, color and pigments
53 including carotenoids, betalains, and anthocyanins are well known for their bioactive properties
54 [8]. Furthermore, the genetically predetermined polyphenols, alkaloids, carotenoids, and
55 phytosterols have well-documented antioxidant and anti-inflammatory activities effecting
56 multiple diseases, from cancer to diabetes or hypertension [9].

57 The composition of edible plants has been extensively mined in the context of bioactivity,
58 focusing on antioxidants [10], anti-microbial potential as an alternative food preservatives [11],
59 and most recently as holistic sources of bioactive compounds in work describing the use of the
60 edible and non-edible parts of plant foods. For example, the non-edible parts of Bell pepper, like
61 the seeds and leaves, were shown to be a valuable source of bioactive phytochemicals [12]. All
62 these reports either rely on a collection of pre-existing work done on single plants or focus on a
63 specific plant. To the best of our knowledge, the work presented here is the first systematic,
64 large-scale approach developed for the exploration of edible plant composition.

65 Our current knowledge on food composition is limited to 150 nutrients catalogued by
66 USDA, despite the fact that the true number of chemical compounds in food ranges from tens
67 [13] to hundreds of thousands across all known plant species [14]. The bulk of our knowledge
68 on the chemical composition of food comes from mass spectrometry and other low-throughput
69 analytical methods and is compiled in repositories such as FooDB [15] and The Dictionary of

70 Food Compounds [16] (DFC), cataloguing comprehensive information on the detected
71 compounds, including both evidence-based and predicted annotations.

72 Our work is driven by the hypothesis that the full list of known, and yet unknown
73 chemicals present in plants are encoded in the genome of the respective organism, encapsulating
74 its metabolic capacities. Recent advancements in genomics have resulted in the emergence of
75 extensive annotation efforts to decipher the genetic potential and the metabolic capacities of
76 plants. For example, KEGG [17] links genes to their functional annotations such as enzymes,
77 reactions, and chemical compounds and catalogues them in metabolic pathway maps, offering
78 functional annotations for 7,254 organisms across the tree of life, out of which 56 are edible
79 plants. Another contributor to plant genome-associated metabolic annotations is PlantCyc [18], a
80 BioCyc [19] based platform adapted to annotate the functional diversity of plant genomes.
81 Although the number of plant genomes sequenced, assembled and annotated has dramatically
82 increased in recent years, the number of available genomes is relatively small (approx. 1,045
83 genomes in NCBI) and represents less than 0.16% of the known plant population [20]. As recent
84 estimates indicate, approximately thirty domesticated plant species represent a significant portion
85 of dietary diversity and three major grains (rice, wheat, and corn) contribute to more than half of
86 the world's caloric intake [21]. Thus, the exploration of existing genome-associated metabolic
87 annotations of staple plant ingredients like corn, rice, soybean, potatoes, and other highly
88 consumed edible plants offer a potentially good resource for describing their composition and
89 hidden benefits.

90 Here, we explore to what degree existing genome-associated metabolic annotations can
91 offer a valuable resource to deepen the knowledge of food composition. To do so, and focus on
92 the edible parts of plants, we rely on metabologenomics [22–24], integrating genomics and

93 metabolomics, used in the past to discover novel natural products [22,25]. To be specific, we
94 develop a systematic metabologenomics approach, coupled with thermodynamic feasibility
95 analysis, aiming to predict the composition of edible plants. We validate our predictions by
96 comparing them to the existing chemical knowledge curated by food composition databases like
97 FooDB, DFC, and USDA. We also collect new experimental data to explore the chemical
98 composition of 13 plants. Our findings indicate that existing genome-associated metabolic
99 annotations offer a predictive platform capable of systematically capturing the chemical
100 composition of plant-based food, from fruits to vegetables, and allow us to predict and
101 experimentally test the presence of novel compounds in plants.
102

103 **Results**

104 **The existing knowledge on food composition**

105 We collected data for 75 edible plants with published and metabolically annotated
106 genomes from two well established databases: KEGG and PlantCyc. Our collection represents
107 plants from 28 families including monocots and dicots (Figure 1a), covering major plant food
108 groups: fruits (apple, banana and orange), grains (rice, corn and quinoa), vegetables (tomato,
109 potato and spinach) and proteins (soy, chickpeas and pigeon pea).

110 To estimate the currently available knowledge on the chemical composition of edible
111 plants, we collected compound annotations from FooDB, DFC, and USDA, cataloguing 5,834,
112 5,151, and 140 compounds respectively, across all plants (Figure S1A). FooDB carries 723±452
113 compounds on average per plant (median: 850), a number that can be as low as three compounds
114 for clementine (*Citrus clementina*) and as high as 2,181 compounds for tea (*Camellia sinensis*).
115 DFC carries 83±121 compounds on average per plant (median: 33) with as low as one compound

116 for vegetable marrow (*Cucurbita pepo subsp. pepo*) and red rice (*Oryza punctata*) and as high as
117 697 compounds for tea. Finally, USDA carries 88±46 compounds on average per plant (median:
118 110), with a single compound for false flax (*Camelina sativa*) and Chinese white pear (*Pyrus* ×
119 *bretschneideri*) and 128 compounds for apple (*Malus domestica*).
120

121 **Genome-associated metabolic annotations' contribution to food composition**

122 To estimate the contribution of available genome-associated metabolic annotations to the
123 existing knowledge on food composition, we collected plant-related compounds from KEGG and
124 PlantCyc, FooDB, USDA, and DFC. KEGG stores information about 7,245 organisms, out of
125 which 546 are eukaryotes and 92 are plants, including 56 edible plants. PlantCyc is a plant-
126 oriented database storing 126 genomes, out of which 58 represent edible plants. As all of
127 databases detailed above are well-established and are of the highest quality of data, they provide
128 annotations with sequence and literature evidence. Thus we consider the collection of plant-
129 associated annotations as a comprehensive representation of the existing metabolic knowledge in
130 the plant genome. These databases overlap and complement each other (39 plants overlap),
131 covering 75 metabolically annotated plant genomes when combined. Overall, KEGG and
132 PlantCyc contributed 1,201 and 3,737 new unique compounds respectively, adding a total of
133 5,224 new compounds (unique and common) to the composition of plants in our catalogue
134 (Figure S1B).

135 To illustrate the contribution of genome-associated metabolic annotations to edible plants
136 (Figure 1b), we focused on corn (*Zea mays*), a highly consumed staple crop [26] worldwide and
137 in the US. Additionally, corn is known to have nutritional and medicinal qualities used in natural
138 medicine to alleviate numerous symptoms, including pain and rheumatism [27]. Existing

139 knowledge for corn includes 1,221 compounds from FooDB, 311 compounds from DFC, and
140 127 compounds from USDA. Considering overlaps between all sources, this compiled to a total
141 of 1,038 unique compounds. Next, we set out to explore the value of adding genome-based
142 annotations to corn's existing knowledge.

143 One contribution of genome-associated metabolic annotations is the metabolic context of
144 these compounds, carried by a network of pathways. Some known pathways are only partially
145 annotated even after considering multiple databases. For example, in Monoterpenoid
146 biosynthesis in corn (Fig 2B) six out of nine compounds are annotated in both databases. Of the
147 three remaining compounds, (R)-Ipsdienol is known to be present in food but was not annotated
148 to any plant in our collection. The two remaining compounds, Ipsdienon, a product of the
149 reaction catalyzed by EC 1.1.1.386 directly from (R)-Ipsdienol, and (6E)-8-Oxolinalool, a
150 product of the reaction catalyzed by EC 1.14.14.84, are currently documented in food but not in
151 corn (white circles, Figure 2b). To strengthen the stringency of our work, the compounds, whose
152 presence is documented in food but not known to be associated with corn, are not included in the
153 plant's catalogue.

154 Consider another example, the tocopherol biosynthesis pathway. Tocopherols are an
155 important class of compounds for health and nutrition [28] (α-tocopherol better known as
156 vitamin E). Figure 2c shows the tocopherol biosynthesis pathway in corn and the delineated
157 contribution of each database to its compounds. We find that databases such as FooDB and DFC
158 annotate the lower half of the pathway, capturing products such as vitamin E and its derivatives.
159 In contrast, genomics-based annotations offer a full pathway annotation, adding 4 new
160 intermediates: 3-(4-hydroxyphenyl)pyruvate, homogenistate, phytyl diphosphate, and 2-methyl-
161 6-phytyl-1,4-benzoquinol and 2 new cofactors: S-adenosyl-L-homocysteine and S-adenosyl-L-

162 methionine, shedding light on the metabolic processes leading to the production of vitamin E in
163 corn.

164 Taken together, existing genome-associated metabolic annotations added 3,021
165 compounds to the list of chemical compounds potentially present in corn, increasing its
166 compound library by 64% (Figure 2a). Across our collection, genome-associated metabolic
167 annotations increased the number of compounds by 2,363±728 chemicals on average per plant,
168 an increase of 75%±15%. After this increase, our database documents 3,239±1,054 compounds
169 per plant. We find that some plants are well annotated in FooDB, DFC, and USDA, while others
170 are poorly annotated (Figure 3). Tea (*Camellia sinensis*) contains the highest number of FooDB,
171 DFC, and USDA annotations (2,547 compounds out of a total of 4,379) showing an increase of
172 42% in new compounds. Some varieties of rice (*Oryza glaberrima*, *Oryza longistaminata*, *Oryza*
173 *barthii*), white yam (*Dioscorea rotundata*), wild tomato (*Solanum pennellii*) and woodland
174 strawberry (*Fragaria vesca*) are not catalogued by FooDB, DFC, or USDA, hence for these
175 plants genome-associated metabolic annotations contributed 100% of the compounds. Other
176 plants like clementine (*Citrus clementina*), lotus (*Nelumbo nucifera*) and African oil palm
177 (*Elaeis guineensis*) are poorly annotated (<250 compounds) in FooDB, DFC, and USDA, hence
178 the addition of genome-associated metabolic annotations increased our knowledge about their
179 chemical composition by more than 85%.

180

181 **Pathway enrichment analysis**

182 Genome-associated metabolic annotations do not only increase our knowledge about the
183 chemical composition of plants but also help us unveil the network of pathways responsible for
184 the production of the newly predicted chemical compounds. Indeed, metabolic pathway mapping

185 allows for a better understanding of the metabolic mechanisms responsible for the synthesis and
186 modulation of natural products and offer a knowledge base towards the prediction of currently
187 undetected compounds.

188 Metabolic pathways are divided into two main classes: primary and secondary
189 metabolism. Primary metabolism is the collection of pathways involved in growth, energy, and
190 reproduction of a plant, while secondary metabolism captures all other functions [29], like
191 flavonoid biosynthesis, xenobiotics metabolism or plant-hormone metabolism, to name only a
192 few. For corn, we collected 789 pathways, of which 35% belong to primary, and 65% belong to
193 secondary metabolism. We asked if certain pathways are represented more than others in our
194 catalogue. To test for pathway bias, we performed a hypergeometric enrichment test capturing
195 the chance for a set of compounds mapped to a pathway in a certain plant to exceed the expected
196 overlap with the general reference pathway (p-value <0.05) (Figure S2). Although the use of the
197 hypergeometric test is a common and well-established tool, it is not without limitations,
198 including recurring tests on the same pathway and the use of pathways as singular units with no
199 inter-pathway biological dependence. To account for recurring tests on the same pathway, we
200 applied the Bonferroni correction to the p-values in this analysis. The choice of database was
201 shown to strongly affect pathway enrichment tests [30]. To maximize the biological relevance of
202 our analysis and compensate for the effects that might rise from database selection, including
203 biological dependence between pathways, we integrated pathway information from two
204 inherently different pathway schemes; KEGG, with larger, more generalized pathway maps, and
205 PlantCyc, a BioCyc-based mapping scheme with smaller, more specific pathway maps. In corn,
206 we found 479 enriched pathways spanning both primary (45%) and secondary metabolism
207 (55%), indicating that our corn dataset is metabolically diverse.

208 We next asked about specialized metabolism occurring only in corn, scanning our plant
209 collection for enriched pathways specific to it (Figure 4a). Corn-specific pathways include
210 specialized metabolism like kauralexin and zealexin biosynthesis [31,32], maysin biosynthesis
211 [33], bergamotene biosynthesis [34] and all-trans-farnesol biosynthesis [35]. Each of these
212 natural products of the diterpenoid, volatile sesquiterpenes, and flavone families are produced by
213 the plant to acquire resistance against biotic and abiotic stress, such as herbivore attack.

214 Generalizing, the analysis performed for corn, we next classified pathways to
215 primary/secondary in our entire collection. We identified 789 pathways across the full plant
216 catalogue, 35% of which are related to primary and 65% to secondary metabolism, a fraction
217 similar to the one observed for corn, offering evidence of diversity in our catalogue (Figure 4b,
218 e). We find most pathways to be enriched (p-value <0.05), with the exception of pathways
219 belonging to glycan biosynthesis and secondary biosynthesis metabolism. Other pathways
220 belonging to secondary metabolism showed a large variation in p-values as multiple outliers
221 were observed above the enrichment line (Figure 4b).

222 We found 762 enriched pathways across all plants (p-value<0.05), out of which 34% are
223 related to primary and 66% are related to secondary metabolism. The distribution of the number
224 of enriched pathways per plant shows two peaks (Figure 4c), corresponding to the two databases,
225 KEGG and PlantCyc. The peaks capture the fact that KEGG has fewer, more complex, and
226 generic pathway map representations, while PlantCyc has a larger number of smaller pathways.
227 For example, KEGG represents the metabolism of alanine, aspartate, and glutamate in one map,
228 while PlantCyc breaks down this process into 10 smaller pathways.

229 Different plants are known for their production of specialized compounds and natural
230 products. Examples include the production of curcumin by turmeric, thymol by thyme, and

231 vanillin by vanilla plants [36], prompting us to identify pathways enriched in single plants,
232 pointing towards potential unique functionalities. Indeed, we observed a bimodal distribution of
233 the number of plants per enriched pathway, one of the peaks being closer to 60 plants and
234 another around 5 plants, indicating the presence of specialized plant metabolism in our catalogue
235 (Figure 4c). Overall, we find that 30 out of 75 plants have at least one plant-specific enriched
236 pathway (Figure S3), out of which we explored seven (Figure 4a). An example of a unique
237 pathway is anthocyanin biosynthesis (delphinidin 3-O-glucoside) in grapes (*Vitis vinifera*),
238 present in seeds and grape skins and used to differentiate between types of wine [37,38]. Other
239 examples include: (1) hordatine biosynthesis in Barley (*Hordeum vulgare subsp. vulgare*).
240 Hordatine, an antifungal compound, highly abundant in young barley shoots, reported to be a
241 potential inhibitor of two main COVID-19 proteins, a protease (PDB ID: 7BQY) and a RNA
242 polymerase (PDB ID: 7bV2) [39] and found in measurable quantities in different types of beer
243 [40]. (2) Ricinoleate biosynthesis, found in castor bean (*Ricinus communis*) and the main
244 constituent of castor oil, was shown to have antibacterial activities, and its elastic properties
245 make it a candidate for packaging polymers with potential applications in biomedical and food
246 technology [41]. (3) Capsaicin, a specialized component in pepper (*Capsicum annuum*) [42],
247 was recently shown to have positive dietetic effects and beneficial antioxidant activity through
248 association with the gut microbiome [43]. (4) Sorgoleone, an allelochemical exuded from
249 sorghum (*Sorghum bicolor*) roots, is known to affect both microbial communities and
250 neighboring plant growth [44]. Finally, we find a variety of acyl-sugar biosynthesis pathways in
251 tomato (*Solanum lycopersicum*). Acyl-sugars are created in tomato trichomes and have
252 commercial and medicinal uses [45]. Another unique pathway in tomato is phenylpropanoid

253 volatiles glycoconjugation, a pathway describing specialized volatiles found in tomato fruits
254 contributing to its signature smell [46].

255 In summary, publicly-available genome-associated metabolic annotations bring diverse
256 metabolic information from both primary and secondary metabolism, offering evidence of new
257 specialized compounds with bioactivity potential.

258

259 **Experimental confirmation of genome-associated metabolic annotations' contribution to**
260 **food knowledge**

261 To experimentally test the capacity of existing plant genomics metabolic annotations to
262 predict the presence of compounds in plants, we performed untargeted metabolomics
263 experiments on 13 out of the 75 plants in our collection, resulting in a catalogue of 939 detected
264 compounds (see Methods). All experiments were done on the edible parts of these plants. On
265 average, 371±130 compounds were detected per plant, ranging from 264 in pear to 652 in apple,
266 and 370 compounds for corn (39.4% of our experimental catalogue). These experimental results
267 allow us to evaluate the accuracy of genome-associated metabolic annotations in the context of
268 the edible parts of the plants analyzed. For this, we measured the overlap of the chemical
269 structures between experimentally detected compounds and our known genomics-based
270 compound collection, finding that genome-associated metabolic annotations significantly overlap
271 with experiments (p-value= 0.018, SI section 1).

272 To be specific, we identify 59 compounds that are found only in corn genome-associated
273 functional annotations and experimentally detected (Figure 5a). We clustered these new
274 compounds based on their chemical structure and classified them into primary and secondary
275 metabolism (Figure 5b, Figure S6), finding that the majority of compounds were attributed to

276 primary metabolism (30) and might represent pooled intermediates and their possible fragments.
277 Metabolite detection may also vary depending on the compound and spectra libraries used in
278 identification. As primary compounds are better studied and annotated, secondary metabolites
279 and their fragmentation products are less abundant in spectra libraries [47]. Overall, we observed
280 well-characterized fractions of lipids, cofactors, sugar, and amino acid derivatives.

281 We identified seven compounds present in corn that might potentially affect human well-
282 being (Figure 5c). For example, Citrulline, a non-essential amino acid known to affect
283 cardiovascular health and dilate blood vessels, and primarily found in watermelon and in smaller
284 amounts in other fruits including a variety of corn species [48], is marketed as a dietary
285 supplement for bodybuilders and athletes to improve exercise endurance. N-Acetyl-D-
286 Glucosamine, previously detected in the shell that protects the first leaf of a corn shoot [49], is
287 known to help support the joints and may help promote healthy skin [50]. Another compound
288 detected is nicotinamide ribonucleotide (NMN), reported to be detected together with NAD+, a
289 well-known cofactor, always present in the cell [51]. Well-being benefits related to this
290 compounds stem from its effects on NAD+ content. This compound has been well studied and as
291 a member of the vitamin B3 family [52] and is being used for the treatment of several
292 cardiovascular, neurodegenerative, and metabolic disorders [53]. Other compounds detected in
293 corn and unveiled by genomics-based annotations have potential cancer-related effects. For
294 example, 5-methylthioadenosine (MTA), a sulfur-containing nucleoside, was recently reported as
295 a tumor suppresser [54]. Another cancer-associated compound, hypotaurine, a sulfinic acid with
296 antioxidant properties, derived from cysteine and an intermediate in taurine production, is one of
297 the top-ranked metabolites for differentiating low and high-grade tumors [55] but was also
298 shown to evoke a malignant phenotype in glioma, the most common primary brain malignancies

299 in adults [56]. It is also marketed as a dietary supplement, together with taurine and L-carnitine,
300 and associated with semen quality improvement [57]. Some other compounds observed in this
301 subset were suggested to have both well-being and economic/industrial importance. For
302 example, Pipecolic acid, a product of lysine metabolism, is an important regulator of immunity
303 in plants and humans. In plants, it accumulates upon pathogen infection and is associated with
304 systemic acquired resistance (SAR) [58]. Pipecolic acid is also an important intermediate of
305 pharmaceutically and biologically derived compounds such as immunosuppressive agents and
306 antibiotics [59,60]. The compound discussed above are known to have potential bioactivity for
307 several applications. The potential effects related to medicinal, industrial and other applications,
308 require further validation. Furthermore, rigorous quantitative analysis is needed to determine
309 their contribution in the scope of quantitative consumption of raw or cooked corn or the
310 concentrated form of a compound as a nutritional supplements.

311 Our experimental investigation also unveiled 22 compounds found only in FooDB, DFC,
312 and USDA. These include: (1) 6-Methoxy-2(3H)-benzoxazolone (MBOA), a degradation
313 product of the known bioactive compound 2,4-dihydroxy-7-methoxy-2H-1,4-benzoxazin-3(4H)-
314 one (DIMBOA). DIMBOA is the main benzoxazinone synthesized in young corn tissues and
315 accumulates in the cells. It is exuded by the roots and acts as a biocide against pests and as an
316 attractant for soil bacteria. MBOA, the more stable form, is often detected in corn soils [61].
317 Interestingly, while DIMBOA is annotated to the corn genome, MBOA, its derivative is not. (2)
318 Vanillic acid, a phenolic detected in corn grits [62]. (3) Trigonelline, an active alkaloid known to
319 be found in corn and associated with antioxidant, anti-carcinogenic, anti-diabetic, and anti-
320 hypercholesterolemia properties [63]. (4) Syringic acid, a phenolic compound found in fruits and
321 vegetables including corn, is reported to have anti-oxidant, antimicrobial, anti-inflammatory, and

322 antiendotoxic properties [64]. (5) Feruloylputrescine, a polyamine monoconjugate, previously
323 detected in corn kernels [65].

324 Finally, to complete our survey of experimentally-detected compounds also found in our
325 genomics-based annotations, we considered the remaining 12 out of 75 plants for which we have
326 both experimental data and plant-genomics metabolic annotations. We find that the number of
327 compounds added by genome-associated metabolic annotations is typically higher than those
328 already catalogued by FooDB and DFC (49±15 and 26±14 compounds respectively, Figure S7).
329 In other words, our analysis shows that genome-associated metabolic annotations significantly
330 enhance existing knowledge of edible plants' chemical composition, helping us uncover
331 potentially novel bioactive compounds.

332

333 **Available genome-associated metabolic annotations annotations contribute to the feasibility**
334 **of compound accumulation**

335 The number of currently identified compounds detected by metabolomics is limited by
336 instrumentation, standard libraries, and analysis pipelines. Recent work presented predictive
337 approaches, applying existing knowledge as a basis for compiling lists of potential compounds
338 that might be detectable by metabolomics experiments. One notable example of such an
339 approach is MINEs [66], a tool that employs pathway and enzyme-based candidate rules to
340 increase the amount and accuracy of compound annotations from raw mass-spectromtery data.
341 Other approaches employ the existing knowledge together with thermodynamic principles to test
342 the feasibility of metabolic pathways like glycolysis, explaining its activity under different
343 conditions and constrains [67]. As thermodynamics rules and constrains are often used to provide
344 insights on the pathway level, here, we hypothesize that the application of standard

345 thermodynamic annotations, and specifically reaction Gibbs free energy values on the compound
346 level can shed light on the likelihood of compounds to accumulate in different plants. Thus, we
347 would be able to identify not only active metabolic routs but the compounds that might play a
348 metabolic role or serve as important metabolic junctions in the plant cell. To that end, we set out
349 to explore genome-associated metabolic annotations' contribution to our ability to predict
350 compounds that accumulate in a plant's edible-parts. Indeed, chemicals that accumulate are more
351 likely to be present in sufficient quantity to be experimentally detected or to enter the
352 bloodstream, potentially modulating health. Indeed, transient compounds may be harder to
353 detect.

354 To study the likelihood of a compound to accumulate, we used Gibbs free energy (ΔG)
355 reaction values combined with genome-scale metabolic network topology. To establish
356 thermodynamic feasibility, we determine the probability that a compound accumulates given all
357 the reactions it takes part in a plant's network context (Figure 6a). In other words,
358 thermodynamic feasibility offers a method for discriminating compounds as likely to accumulate
359 based on cumulative ΔG values of the reactions it is either a reactant or a product in. We
360 collected ΔG values from modelSEED [68] and PlantCyc and calculated the cumulative ΔG
361 value (score) for each compound. ModelSEED values are calculated as described in eQuilibrator
362 [69], a well-established source for Gibbs free energy values. PlantCyc values are calculated
363 according to the BioCyc schema of computing Gibbs free energy as described in [18]. If the
364 compound acted as a reactant in a reaction, we assume that it is consumed, hence it is a transient
365 compound, assigned a negative ΔG value. In contrast, if the compound is a product of the
366 reaction, it is assigned a positive value. Compounds with positive scores are produced more than
367 consumed, thus they are likely to accumulate. The set of available reactions was considered for

368 each compound, out of this set, the reaction representing the largest absolute ΔG value may shift
369 the balance towards either the consumption or production of that compound. This, in turn may
370 determine the metabolic pathway activity and the creation of intermediates or byproducts that
371 may or may not accumulate if they are not consumed by other thermodynamically feasible
372 reactions. In corn, we scored 2,985 compounds (63% of total compounds), out of which 1,460
373 have positive ΔG values, i.e., are potentially expected to accumulate given the context of
374 reactions they are involved in.

375 To illustrate our findings, we explored the compounds on the vitamin E biosynthesis
376 pathway in corn (Figure 6b). We observed a potentially likely accumulation of pathway products
377 α and β tocopherol and of the intermediates γ and δ tocopherol. Intermediates with the potential
378 to accumulate are likely involved in more than one reaction outside the pathway. These reactions
379 may have larger absolute ΔG values that can either lead to the rerouting of the compound to
380 another metabolic pathway, creating a temporary hold on the production of certain compounds.
381 For example, phytyl diphosphate is involved in 7 reactions and 5 pathways. The largest ΔG value
382 measured for this set of reactions representing the largest ΔG value was annotated to the phytyl
383 salvage pathway, describing the conversion of degraded chlorophyll to phytyl phosphate,
384 contributing to its potentially high likelihood to accumulate. A compound could potentially
385 accumulate if it is an intermediate appearing in two reactions with a large ΔG value difference in
386 which the reaction with the larger ΔG value favors the production of the compound in question.
387 For example, 2,3-dimethyl-6-phytyl-1,4-benzoquinol is involved in two reactions, both annotated
388 to the vitamin E biosynthesis pathway, where the ΔG value for the reaction producing it (20.64
389 kcal/mol) is larger than the ΔG value for the reaction consuming it (-12.05 kcal/mol). The use of
390 thermodynamic feasibility to predict potentially accumulating compounds that might be

391 experimentally detected is not without its limitations. One example is pathway shifts due to ΔG
392 value differentiation. These shifts are governed not only by thermodynamic feasibility but by
393 various other cellular considerations such as genetic and metabolic regulation, and environmental
394 conditions including nutrient availability, temperature, osmolarity, and pressure. To further
395 estimate the power of our approach we condensed the compounds to compound families
396 represented by the first block of their InchIKey. This allows for better comparison to
397 metabolomics experimental results (see methods). Overall, 1,754±588 compound families were
398 scored per plant, covering an average of 79±17% of its compounds. The fraction of scored
399 compounds was as high as 96.5% of the total compounds in *Oryza longstaminata*, a species of
400 rice, and as low as 50% in adzuki bean (*Vigna angularis*).

401 To estimate the predictive power of our approach, we first asked if kinetics-based
402 annotations have increased the predictive power of our approach compared to total genome-
403 associated metabolic annotations. We find that thermodynamics-based annotations show a more
404 significant overlap with the experiments compared to genome-associated metabolic annotations
405 (p-value=0.0018, SI section 1, Figure S4), and they are characterized by a high degree of
406 structural similarity, significantly different from a random sample of the same size taken from
407 genome-associated metabolic annotations (p-value<0.001, SI section 1).

408 Next, we used the experimentally detected compounds in the 13 plants for which we
409 performed untargeted metabolomics, to estimate the performance of our approach. Similar to
410 known machine learning methods, we look for compounds that are likely to accumulate, since
411 our experimental data used as validation does not contain concentration information, we use
412 compound score as means to establish the best performing threshold for solving for a binary
413 classification (a compound might likely accumulate or not). After establishing the score

414 threshold, all compounds with scores above it are considered as likely to accumulate. We then
415 compare it to our experimentally detected compound catalogue as ground truth values (a binary
416 classification denoting presence or absence). We calculated standard performance metrics, such
417 as the true positive and false positive rates and the area under the receiver-operator curve (ROC),
418 AUC_{ROC}. In addition, we calculated the precision, recall, and F1 scores. Since our data may be
419 imbalanced, we initially set the threshold of prediction to be larger than zero (positive values).
420 We then performed a moving threshold analysis (see Methods) to determine the optimal
421 threshold for best performance in each plant found in both our annotation and experimental
422 catalogue (Table S1). Most AUC_{ROC} values were above the discrimination line (0.5), several
423 representing acceptable discrimination (AUC_{ROC} values between 0.69 to 0.76) (Figure 6c).

424 Finally, we selected the top ranked 110 corn compound families with the best
425 performance of our thermodynamics-based approach (AUC_{ROC}=0.74). This list contained 15
426 experimentally detected compounds including AMP, S-adenosylhomocysteine (SAH), and 5-
427 methylthioadenosine (MTA), which are cofactors maintained as constant pools in the cell. Other
428 compounds such as succinate, Glycerol 3-phosphate, and 3-phospho-D-glycerate are key
429 products of major energy producing pathways, such as carbon fixation by photosynthesis,
430 glycolysis, and the TCA cycle. Interestingly, L-glutamate and D-Galacturonic acid were also
431 detected. L-glutamate was previously reported as a key metabolite in corn, measured in large
432 amounts in the endosperm [70,71]. D-Galacturonic acid is the main component of pectin, a
433 polysaccharide naturally found in plant cell walls. Thus, both compounds are likely to
434 accumulate and be detected in metabolomics measurements, supporting the predictive power of
435 our approach. To further examine the possibility of detecting natural products in both our
436 experiments and predictions, we compared our results to CoconutDB [72], an open collection of

437 natural products. Indeed, we found that most of our predicted and experimentally detected
438 compounds overlap with this dataset (Figure S5, Table S2). Surprisingly, CoconutDB contains
439 many annotations for primary metabolites, including TCA cycle intermediates and simple sugars.
440 Interestingly, the fraction of experimentally detected compounds that does not overlap with
441 CoconutDB is comprised mostly of fatty acid derivatives. Indeed, the detection and annotation of
442 lipids by mass spectrometry is limited. One of the limitations is the existence of thousands of
443 structurally diverse lipids in large gradient of concentrations, some being below the experimental
444 detection level. Another limitation in the detection and annotation of these compounds is
445 ingrained in the differences between mammalian and plant lipids, as most of the existing lipid
446 measurement technologies were developed for mammalian lipid analysis. Finally, lipid
447 annotations are limited to well-studied and well-captured lipids. For example, the KEGG
448 database contains 4,132 lipid annotations, representing only a small fraction of the expected
449 diversity of hundreds of thousands of plant lipid species [73]. In many cases lipids are
450 represented as a general backbone that can carry different residues (R groups). One example is
451 the KEGG annotation of Glycerophosphoglycoglycerolipid (KEGG compound C06041), a
452 generic compound with an unknown R group that still takes part in two pathways; (1) Teichoic
453 acid biosynthesis and (2) Glycerolipid metabolism. Considering these and other limitations, the
454 presented approach still captures a significant signal describing the compounds that can
455 potentially accumulate in edible plants.
456

457 **Discussion**

458 Here, we developed a systematic methodology to extract the contribution of publicly
459 available genome-associated metabolic annotations to the molecular composition of foods. We

460 found that genome-associated metabolic annotations not only boosted, in some cases by more
461 than 85%, the number of compounds known to be present in a plant (in comparison to FooDB,
462 DFC, and USDA databases) but also offered valuable mechanistic knowledge in the form of
463 chemical structures and the metabolic pathways responsible for their production. Using multiple
464 types of annotations (compounds, reactions, and pathways), we surveyed the contribution of
465 genome-associated metabolic annotations in depth. These annotations, combined with
466 experimentally detected compounds, were used to gain new insights into the chemical
467 composition of edible plants, specifically corn.

468 The molecular composition of plants changes in time and in response to biotic and abiotic
469 stresses such as environmental conditions, herbivore and pathogen attacks and hormonal signals,
470 all governed by complex genetic and metabolic regulation. We therefore set out to explore the
471 feasibility of a compound to accumulate, affecting the likelihood of being experimentally
472 detectable. We find good discriminative power, supporting the large-scale use of thermodynamic
473 annotations. As experimental data and thermodynamics annotations' availability increase, they
474 may lead to significant enhancement in the predictive power of metabologenomics.

475 While the advent of genome-associated metabolic annotations offers new insights into the
476 composition of edible plants, it is not without limitations. Various biases that might arise from
477 such data were explored throughout this work, representing only a few of the multiple factors
478 that might affect our knowledge of food composition. One major limitation of a whole-genome
479 holistic 'metabolic-soup' approach, specifically in complex organisms such as plants, is the
480 spatial allocation of compounds. It is known that in plants, a compound can be formed in one
481 tissue and accumulate in another. For example, the spatiotemporal distribution of amino acids,
482 flavonoids, and more was shown to be differential across different tissues of citrus fruits with a

483 larger fraction of flavonoids in the peel and an increased content of amino acids in the juice sacs
484 [74] . Another challenge is whole-genome metabolic approaches is the integration of genetic and
485 metabolic regulation. The use of transcriptomics in plants has been recently rising. As
486 transcriptomics analysis advances, technologies such as single-cell transcriptomics and
487 transcriptomics integrated with metabolomics are more widely used in plant biology [75,76].
488 One notable example is the integration of transcriptomics and metabolomics to study
489 anthocyanin biosynthesis in cabernet sauvignon grape berries in response to growth condition
490 changes [76]. A natural extension of transcriptomics is plant proteomics, describing the proteins
491 that are encoded by activated or transcribed genes. Plant proteomics has also been shown to be
492 valuable in agricultural and food applications [77,78]. The integration of transcriptomics and
493 proteomics data into the approach developed here, may improve and augment the prediction of
494 compounds that are likely to accumulate and be detected by mass spectrometry analysis in the
495 future. Annotations of regulation of gene expression and metabolic feed-forward or back loops
496 are limited. With the advancement of the collection, analysis, and standardized curation, these
497 data will improve our predictions. Another major limitation is the availability of annotated plant
498 genomes. One limitation is the ability of existing annotation tools to comprehensively annotate
499 compounds that might be metabolized by enzymes encoded by orphan genes or genes with
500 unknown functions. Recent work suggested such genes to contribute to a wide variety of
501 biological functions and playing essential roles in plant phenotype [79]. As many annotation
502 tools rely on homolog information to infer metabolic annotations (notable examples are
503 plantiSMASH [80] and PlantCyc), these functions might be overlooked. Other metabolic
504 annotation repositories like KEGG curate their annotations based on experimental and published
505 work and therefore provide a more accurate collection. Still these data repositories rely heavily

506 on published work that in turn might overlook orphan gene contributions. As plant orphan gene
507 identification and study of function is rapidly advancing [81], improved annotations may soon be
508 available for a more accurate investigation into the metabolic annotation of edible plants.

509 While the cost of sequencing dropped significantly, we find that the number of non-
510 model plant genomes annotated remains limited and grows slowly. Other factors such as
511 annotation quality and database standardization can also limit omics-based analysis. As shown
512 here, two major databases, KEGG and PlantCyc, has introduced some redundancy in pathway
513 mappings, partially rooted in the different pathway definitions used. As data continues to
514 accumulate, standardization and mappings between the different data sources is key to deriving
515 new insights.

516 The use of metabolomics to detect compounds in edible plants highly depends on the
517 instrumentation used, standard libraries, and identification methods. Sample variation in the form
518 of produce origin, growth conditions, and post-harvest treatments can also increase the
519 variability and introduce environmental compounds that might be experimentally detected. For
520 example, it is known that mass-spectrometers have poor detection of stereoisomers [82–84]. In
521 our results, we apply the most stringent gold standard metabolomics annotation method,
522 comparing the resulting spectra to known standards. To determine this approach's comparability
523 to the benchmark of massSpec annotation methodologies used for food metabolomics, we
524 assessed the methodologies on the three most common experimental platforms (HILIC, RP, GC)
525 and five food items (seaweed, Sichuan peppercorn, Miso, Pig's Ear mushroom, Macaroni and
526 cheese) and three annotation approaches: gold standard, spectra matching, and computational
527 annotation. We find that the gold standard method tends to annotates up to 597 compounds,
528 spectra library matching annotates up to 2,855 compounds, and the computational method

529 annotates up to 5,225 compounds. We find that spectra library matching offers a 5-fold increase
530 to gold standard and computation annotation provides an 8-fold increase. In our pilot the existing
531 single-peak based tools annotated 2,426 peaks of the total 28,000 peaks combined over three
532 experimental methods. Studies have shown that about 50% of the peaks in any MassSpec
533 experiment are adducts [85–87], molecules formed by the addition of two or more molecules
534 during sample preparation. This means that about 14,000 peaks correspond to individual
535 compounds of which currently only 17.3% are annotated. Plants are known to produce around
536 200,000 compounds, including primary and secondary metabolites [88]. As new metabolomics
537 annotation tools are being developed, with a focus on computational analysis, we expect more
538 and more compounds to be annotated, increasing the discovery of a greater fraction of the vast
539 diversity produced by plants. Even though considerable efforts have been made to establish
540 spectral analysis pipelines, high-throughput metabolomics compound identification remains
541 limited. Future improvements in metabolomics analysis and annotation could significantly
542 enhance our knowledge of specialized metabolites and natural products in edible plants.
543 Additionally, wider scale surveys including samples of produce from different points in the
544 supply chain, starting from the field and ending in the grocery store will be useful to further
545 study the accumulation patterns of potential bioactive compounds and their availability to the
546 consumer.

547 Finally, the genome-associated metabolic annotations analyzed here greatly contribute to
548 our existing knowledge of the composition of edible plants. Furthermore, this approach can be
549 extended to be used for predicting microbial natural products and compound with bioactivity or
550 industrial potential. Microbial genomes are smaller and easier to sequence and assemble.
551 Moreover microbial species are better represented in the metabolic annotation databases and

552 potentially may lead to more accurate predictions. The integration of genomics and
553 metabolomics has been suggested as a promising combination towards the identification of
554 promising compounds [89]. Here, we employ metabolomics of the edible parts of staple plant
555 food ingredients and specifically corn to show the value this approach in the discovery of new
556 compounds in specific plant tissues. Thus, this work offers a stepping stone towards a better
557 understanding of food composition, offering insights based on fast-growing computational and
558 experimental datasets.

559

560

561 **Methods**

562 **Data collection**

563 Compound annotations were collected from open-source (FooDB and USDA) and
564 proprietary (DFC) databases. Genomics based compound, reaction and pathway annotations
565 were collected from KEGG and PlantCyc. ΔG values were gathered from PlantCyc and
566 modelSEED. All annotations were completed with SMILES, InChIKey, and molecular weight.
567 Plant diversity was analyzed and visualized using the ETE3 toolkit [90] tree functions (ete3,
568 version 3.1.2). All taxonomic data originated from NCBI. Overall, our catalogue includes
569 15,296, out of which 12,662 first block InChIKeys representing chemical families. To compare
570 our catalogue and experimentally detected compounds (detailed in the next section), we collected
571 first block InChikey compound annotations from CoconutDB (January 2022 release).
572 CoconutDB identifiers were converted to BioCyc and KEGG identifiers for further functional
573 analysis using CTS – The chemical translation service [91]. The database files are available on
574 https://github.com/Barabasi-Lab/Plant-genomics.

575

576 **Mass spectrometry experiments**

577 A selection of 13 produce items was purchased from two local grocery stores (Whole
578 Foods Market and Stop & Shop): apple, banana, carrot, chickpea, corn, lettuce, olive, peach,
579 pear, potato, spinach, soybean, and tomato. Produce available in grocery stores represent the
580 product consumed by the populations. Indeed, this produce may be affected by post-harvest
581 treatments and preservation methods that might introduce foreign non-plant produced
582 compounds. To standardize and minimize these potentially foreign compounds, we treated the
583 samples as they are eaten by the customer. For example, a common post-harvest treatment is
584 applying a wax coating to fruit like apples, this protects the fruit from both mechanical and
585 microbial damage. Removing the peel, removes these waxy compounds that otherwise would
586 potentially be detected by mass spectrometry. Each produce item sample contained the combined
587 material of six units (for example 6 apples) and was prepared in a humidity-controlled room with
588 minimum light exposure. Sample preparation included peeling, chopping, freeze drying (-80ºC
589 for 24 hrs, Catalog No. 10-269-56B from LabConco/Fisher) and pulverizing into a fine
590 homogenized powder (Kitchen Aid, 170W, Model No. BCG111OBO). All samples were
591 prepared by the Giese lab (Northeastern University, Boston MA USA). The final samples were
592 stored at -80 °C in vials containing 200 mg of powder and argon gas. The samples were shipped
593 to two metabolomics centers (West Coast Metabolomics Center, UC Davis, CA USA and
594 Metabolon, Morrisville, NC USA) for analysis on multiple platforms including UHPLC-CSH
595 C18-HRMS-Orbitrap, UHPLC-BEH Amide-HRMS-Orbitrap, UHPLC-PFP-HRMS-Orbitrap,
596 and Shimadzu LC and SelexION QTRAP MS and annotation (see SI section 2).

597 Experimental results were annotated by spectral matching or identified by a reference
598 standard library. Results from both metabolomics centers were merged and standardized to
599 InChIKeys (PubChem). Since metabolomics methodologies does not account for
600 stereochemistry, only the first block of the InChIKey is used to compare two entries. Therefore,
601 the resulting list is one of unique compound structures found in each food item. All mass
602 spectromtery data is available in MetabolomicsWorkbench as ST002493 (for data acquired from
603 UC Davis) and ST002492 (for data acquired from Metabolon).
604
605 **Statistics and enrichment tests**
606 Enrichment analysis was performed with the hypergeometric distribution test (python 3.8,
607 scipy [92] 1.6.3). Pathway enrichment test was performed per organism. To that end, we first
608 collected all the compound annotations for the reference pathways from KEGG and PlantCyc,
609 including all known annotations attributed to that pathway. We calculated the number of
610 compounds found in each pathway in the organism and tested it for significance against the
611 reference. To account for multiple testing on the same pathway, Bonferroni correction was
612 applied to pathways' p-values.
613
614 **Compound similarity across genomics, kinetics, and experiments**
615 To investigate the degree of structural similarity and overlap between molecules retrieved by
616 different techniques, we performed similarity search and clustering on a variety of molecular
617 fingerprints (FP). We leveraged the python package rdkit [93] to standardize SMILES and
618 InChIKeys associated with each compound annotation, using Morgan fingerprints (FP). To
619 generate a Morgan FP, all substructures around the heavy atoms of a molecule within a defined

radius are generated and assigned to unique identifiers. These identifiers are then hashed to a vector with a fixed length. The chemoinformatic community employs 1024- or 2048-bit vectors, populating them with fragments up to radius 1 or 2. For our analysis we increased the resolution up to 8192 bits and radius 3, to capture fragments of bigger size and reduce the potential bit collision.

Since the first block of the InChIKey represents several stereoisomers, we assign a bit vector to each first block as the union of all bit vectors representing the related isomers. We then compare the degree of structural similarity between any pair of compounds by computing the Jaccard similarity between binary vectors.

We assess structural similarity for a given set of N compounds, by calculating the *intrinsic dimension* of their Jaccard similarity matrix $\{S_{ij}\}$, a function of the spectrum $\{\lambda_i\}$ of $\{S_{ij}\}$,

$$n = e^{-\sum_{i=1}^{N} \frac{\lambda_i}{\sum_{k=1}^{N} \lambda_k} \log\left(\frac{\lambda_i}{\sum_{k=1}^{N} \lambda_k}\right)} = e^{H}, \quad (1)$$

where the Shannon entropy H of the normalized spectrum is used to estimate the number of independent components synthetizing the same amount of structural similarity observed in the sample. A higher n indicates a more chemically diverse sample. We used n to assess how different the kinetics annotations are compared to a random sample from the genomics set (comparison with 1000 subsamples) (SI Section 1).

Thermodynamic feasibility analysis

We collected available thermodynamic annotations for all the plants in our collection. Thermodynamic annotations in the form of ΔG values (Kcal/mol) were collected from ModelSEED[68] and PlantCyc [18].

642 As our validation set is based on metabolomics identified compounds, we used the first
643 block of the InChIKey string as a compound identifier. First block identifiers collapse
644 stereoisomer information, a known limitation of mass spectrometry. To calculate the cumulative
645 score of ΔGs for a compound we collected all the reaction annotations in which it is involved. If
646 the compound was acting as reactant in a reaction its ΔG value would be assigned a negative
647 value and if it would be a product, a positive value. All the values were then summed to
648 represent a likelihood to accumulate score where a positive value indicated a likely to
649 accumulate compound and a negative value a likely to be consumed value. This approach was
650 applied for each plant in our plant collection with corresponding experimental results. A
651 schematic overview of the developed pipeline is described in Figure S7.
652

653 **Performance evaluation**

654 To evaluate the performance of the thermodynamics feasibility approach, we used the
655 first block of the InChIKey of our experimentally detected compound catalogue as a validation
656 set and ground truth. For each plant we created a binary matrix representing the predicted score
657 (calculated likely to accumulate value) and the experimental outcome (detected/not detected)
658 using a cutoff threshold of a positive score (>0). We then created an ROC and precision-recall
659 curve and calculated the area under the curves. To test for optimal performance, we first ranked
660 our data according to the highest scores, implying better likelihood to accumulate and/or be
661 detected experimentally. Next, we applied a moving threshold analysis to establish the cutoff
662 threshold leading to optimal performance. Briefly, we systematically increased the portion of
663 ranked data, calculating performance metrics for each step and identified the threshold leading to
664 optimal performance by finding the maximum geometric mean of sensitivity and specificity, and

updated performance metrics accordingly. All calculations were performed and plotted using the sklearn (version 0.24.1) and seaborn (version 0.11.1) python packages and Python 3.8).

Acknowledgements

Experiments performed at the West Coast Metabolomics Center at University of California, Davis were led by Dr. Oliver Fiehn and Dr. Arpana Vaniya. Experiments at Metabolon were led by Nino Esile, Dr. Brian Ingram, and Nathan Testa. We thank Dr. John de la Parra, and Rebekah Carlson for preparing the food samples analyzed by UC Davis. We thank Dr. Roger Giese and Dr. Pushkar M. Kulkarni at the Department of Pharmaceutical Sciences, Northeastern University for preparing the food samples analyzed by Metabolon, Inc.

Author contributions

Shany Ofaim: Data Curation, Formal Analysis, Software, Methodology, Validation, Visualization, Writing – Original Draft Preparation. **Giulia Menichetti:** Conceptualization, Formal analysis, Methodology, Software, Writing – Review & Editing. **Michael Sebek**: Data Curation, Investigation, Resources, Writing – Review & Editing. **Albert László Barabási:** Conceptualization, Funding Acquisition, Supervision, Writing – Review & Editing.

Competing interests

A.L.B. is a scientific founder of Scipher Medicine, Inc., which applies network medicine strategies to personalized drug selection and Naring, Inc., which applies data science to food and health. All other authors declare no competing interests.

Data availability

The datasets generated during and/or analyzed during the current study are available on https://github.com/Barabasi-Lab/Plant-genomics.

Code availability

Code and scripts are available on https://github.com/Barabasi-Lab/Plant-genomics.

696 **References**

697

698 1. Dietary Guidelines for Americans | USDA-FNS. [cited 29 Jun 2021]. Available:
699 https://www.fns.usda.gov/cnpp/dietary-guidelines-americans

700 2. Luo Y, Shang P, Li D. Luteolin: A Flavonoid that has multiple cardio-protective effects and its
701 molecular mechanisms. Frontiers in Pharmacology. 2017. pp. 1–10. doi:10.3389/fphar.2017.00692

702 3. Barabási AL, Menichetti G, Loscalzo J. The unmapped chemical complexity of our diet. Nature
703 Food. 2020. pp. 33–37. doi:10.1038/s43016-019-0005-1

704 4. Hooton F, Menichetti G, Barabási A-L. Exploring food contents in scientific literature with
705 FoodMine. Sci Rep. 2020;10: 16191. doi:10.1038/s41598-020-73105-0

706 5. Wimalanathan K, Lawrence-Dill CJ. Gene Ontology Meta Annotator for Plants (GOMAP). Plant
707 Methods 2021 171. 2021;17: 1–14. doi:10.1186/S13007-021-00754-1

708 6. Choi IY, Kwon EC, Kim NS. The C- and G-value paradox with polyploidy, repeatomes, introns,
709 phenomes and cell economy. Genes and Genomics. 2020. pp. 699–714. doi:10.1007/s13258-020-
710 00941-9

711 7. Nyonje WA, Schafleitner R, Abukutsa-Onyango M, Yang RY, Makokha A, Owino W. Precision
712 phenotyping and association between morphological traits and nutritional content in Vegetable
713 Amaranth (Amaranthus spp.). J Agric Food Res. 2021;5: 100165.
714 doi:10.1016/J.JAFR.2021.100165

715 8. Sharma S, Katoch V, Kumar S, Chatterjee S. Functional relationship of vegetable colors and
716 bioactive compounds: Implications in human health. J Nutr Biochem. 2021;92.
717 doi:10.1016/J.JNUTBIO.2021.108615

718 9. Otunola GA, Martiryosan D. Choosing suitable food vehicles for functional food products. Funct
719 Foods Heal Dis. 2021;11: 44–55. doi:10.31989/FFHD.V11I2.764

720 10. Arias A, Feijoo G, Moreira MT. Exploring the potential of antioxidants from fruits and vegetables
721 and strategies for their recovery. Innov Food Sci Emerg Technol. 2022;77: 102974.
722 doi:10.1016/J.IFSET.2022.102974

723 11. Oulahal N, Degraeve P. Phenolic-Rich Plant Extracts With Antimicrobial Activity: An Alternative
724 to Food Preservatives and Biocides? Front Microbiol. 2022;12: 3906.
725 doi:10.3389/FMICB.2021.753518/XML/NLM

726 12. Anaya-Esparza LM, de la Mora ZV, Vázquez-Paulino O, Ascencio F, Villarruel-López A. Bell
727 peppers (Capsicum annum l.) losses and wastes: Source for food and pharmaceutical applications.
728 Molecules. MDPI; 2021. doi:10.3390/molecules26175341

729 13. Moghe GD, Leong BJ, Hurney SM, Jones AD, Last RL. Evolutionary routes to biochemical
730 innovation revealed by integrative analysis of a plant-defense related specialized metabolic
731 pathway. Elife. 2017;6: 1–33. doi:10.7554/eLife.28468

732 14. Vermeulen R, Schymanski EL, Barabási A-L, Miller GW. The exposome and health: Where
733 chemistry meets biology. Science (80-). 2020;367: 392–396. doi:10.1126/science.aay3164

734 15. FooDB, www.foodb.ca. [cited 11 Feb 2023]. Available: https://foodb.ca/

735 16. Yannai S. Dictionary of food compounds with CD-ROM. 2012.

736 17. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res.
737 2000;28: 27–30.

738 18. Schläpfer P, Zhang P, Wang C, Kim T, Banf M, Chae L, et al. Genome-wide prediction of
739 metabolic enzymes, pathways, and gene clusters in plants. Plant Physiol. 2017;173: 2041–2059.
740 doi:10.1104/pp.16.01942

741 19. Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, et al. The BioCyc
742 collection of microbial genomes and metabolic pathways. Brief Bioinform. 2019;20: 1085–1093.
743 doi:10.1093/BIB/BBX085

744 20. Marks RA, Hotaling S, Frandsen PB, VanBuren R. Representation and participation across 20
745 years of plant genome sequencing. Nat Plants. 2021;7: 1571–1578. doi:10.1038/s41477-021-
746 01031-8

747 21. Ray A, Ray R, Sreevidya EA. How Many Wild Edible Plants Do We Eat—Their Diversity, Use,
748 and Implications for Sustainable Food System: An Exploratory Analysis in India. Front Sustain
749 Food Syst. 2020;4. doi:10.3389/fsufs.2020.00056

750 22. Soldatou S, Eldjárn GH, Ramsay A, van der Hooft JJJ, Hughes AH, Rogers S, et al. Comparative
751 Metabologenomics Analysis of Polar Actinomycetes. Mar Drugs. 2021;19: 1–21.
752 doi:10.3390/md19020103

753 23. Parkinson EI, Tryon JH, Goering AW, Ju K-S, McClure RA, Kemball JD, et al. Discovery of the
754 Tyrobetaine Natural Products and Their Biosynthetic Gene Cluster via Metabologenomics. ACS
755 Chem Biol. 2018;13: 1029–1037. doi:10.1021/ACSCHEMBIO.7B01089

756 24. Saad H, Aziz S, Gehringer M, Kramer M, Straetener J, Berscheid A, et al. Nocathioamides,
757 Uncovered by a Tunable Metabologenomic Approach, Define a Novel Class of Chimeric
758 Lanthipeptides. Angew Chemie Int Ed. 2021;60: 16472–16479. doi:10.1002/ANIE.202102571

759 25. Fox Ramos AE, Evanno L, Poupon E, Champy P, Beniddir MA. Natural products targeting
760 strategies involving molecular networking: Different manners, one goal. Nat Prod Rep. 2019;36:
761 960–980. doi:10.1039/c9np00006b

762 26. Ranum P, Peña-Rosas JP, Garcia-Casal MN. Global maize production, utilization, and
763 consumption. Ann N Y Acad Sci. 2014;1312: 105–112. doi:10.1111/nyas.12396

764 27. Narayan A, Kumar D, Narayan Jhariya A. Mini Review Paper Nutritional, Medicinal and
765 Economical importance of Corn: A Mini Review. Res J Pharm Sci. 2013;2: 1–6.
766 doi:10.13140/RG.2.2.30058.62408

767 28. Shahidi F, Pinaffi-Langley ACC, Fuentes J, Speisky H, de Camargo AC. Vitamin E as an essential
768 micronutrient for human health: Common, novel, and unexplored dietary sources. Free Radic Biol
769 Med. 2021;176: 312–321. doi:10.1016/J.FREERADBIOMED.2021.09.025

770 29. Aharoni A, Galili G, Avni A, Blazquez M. Metabolic engineering of the plant primary-secondary
771 metabolism interface. Curr Opin Biotechnol. 2011;22: 239–244. doi:10.1016/j.copbio.2010.11.004

772 30. Karp PD, Midford PE, Caspi R, Khodursky A. Pathway size matters: the influence of pathway
773 granularity on over-representation (enrichment analysis) statistics. BMC Genomics. 2021;22.
774 doi:10.1186/s12864-021-07502-8

775 31. Schmelz EA, Huffaker A, Sims JW, Christensen SA, Lu X, Okada K, et al. Biosynthesis,
776 elicitation and roles of monocot terpenoid phytoalexins. Plant J. 2014;79: 659–678.
777 doi:10.1111/tpj.12436

778 32. Murphy KM, Ma LT, Ding Y, Schmelz EA, Zerbe P. Functional characterization of two class ii
779 diterpene synthases indicates additional specialized diterpenoid pathways in maize (zea mays).
780 Front Plant Sci. 2018;871: 1–12. doi:10.3389/fpls.2018.01542

781 33. Moore VM, Tracy WF. Combining ability of husk extension, maysin content, and corn earworm
782 resistance. J Am Soc Hortic Sci. 2021;146: 14–23. doi:10.21273/JASHS04974-20

783 34. Köllner TG, Degenhardt J, Gershenzon J. The product specificities of maize terpene synthases
784 tps4 and tps10 are determined both by active site amino acids and residues adjacent to the active
785 site. Plants. 2020;9. doi:10.3390/plants9050552

786 35. Schnee C, Köllner TG, Gershenzon J, Degenhardt J. The maize gene terpene synthase 1 encodes a
787 sesquiterpene synthase catalyzing the formation of (E)-β-farnesene, (E)-nerolidol, and (E,E)-
788 farnesol after herbivore damage. Plant Physiol. 2002;130: 2049–2060. doi:10.1104/pp.008326

789 36. Chiorcea-Paquim AM, Enache TA, De Souza Gil E, Oliveira-Brett AM. Natural phenolic
790 antioxidants electrochemistry: Towards a new food science methodology. Compr Rev Food Sci
791 Food Saf. 2020;19: 1680–1726. doi:10.1111/1541-4337.12566

792 37. Kyraleou M, Kallithraka S, Gkanidi E, Koundouras S, Mannion DT, Kilcawley KN.
793 Discrimination of five Greek red grape varieties according to the anthocyanin and
794 proanthocyanidin profiles of their skins and seeds. J Food Compos Anal. 2020;92: 103547.
795 doi:10.1016/j.jfca.2020.103547

796 38. Tang K, Liu T, Han Y, Xu Y, Li JM. The Importance of Monomeric Anthocyanins in the
797 Definition of Wine Colour Properties. South African J Enol Vitic. 2017;38: 1–10.
798 doi:10.21548/38-1-661

799 39. Dahab MA, Hegazy MM, Abbass HS. Hordatines as a Potential Inhibitor of COVID-19 Main
800 Protease and RNA Polymerase: An In-Silico Approach. Nat Products Bioprospect. 2020;10: 453–
801 462. doi:10.1007/s13659-020-00275-9

802 40. Pihlava JM, Kurtelius T, Hurme T. Total hordatine content in different types of beers. J Inst Brew.
803 2016;122: 212–217. doi:10.1002/jib.311

804 41. Totaro G, Sisti L, Bozzi Cionci N, Martinez GA, Di Gioia D, Celli A. Elastomeric/antibacterial
805 properties in novel random Ricinus communis based-copolyesters. Polym Test. 2020;90: 106719.
806 doi:10.1016/j.polymertesting.2020.106719

807 42. Scossa F, Roda F, Tohge T, Georgiev MI, Fernie AR. The Hot and the Colorful: Understanding
808 the Metabolism, Genetics and Evolution of Consumer Preferred Metabolic Traits in Pepper and
809 Related Species. CRC Crit Rev Plant Sci. 2019;0: 1–43. doi:10.1080/07352689.2019.1682791

810 43. Sinisgalli C, Vezza T, Diez-Echave P, Ostuni A, Faraone I, Hidalgo-Garcia L, et al. The
811 Beneficial Effects of Red Sun-Dried Capsicum annuum L. Cv Senise Extract with Antioxidant
812 Properties in Experimental Obesity are Associated with Modulation of the Intestinal Microbiota.
813 Mol Nutr Food Res. 2021;65: 1–13. doi:10.1002/mnfr.202000812

814 44. de Oliveira IF, Simeone MLF, de Guimarães CC, Garcia NS, Schaffert RE, de Sousa SM.
815 Sorgoleone concentration influences mycorrhizal colonization in sorghum. Mycorrhiza. 2020;
816 259–264. doi:10.1007/s00572-020-01006-1

817 45. Schilmiller AL, Gilgallon K, Ghosh B, Jones AD, Last RL. Acylsugar acylhydrolases:
818 Carboxylesterase-catalyzed hydrolysis of acylsugars in tomato trichomes. Plant Physiol. 2016;170:
819 1331–1344. doi:10.1104/pp.15.01348

820 46. Tikunov YM, de Vos RCH, Paramás AMG, Hall RD, Bovy AG. A role for differential
821 glycoconjugation in the emission of phenylpropanoid volatiles from tomato fruit discovered using
822 a metabolic data fusion approach. Plant Physiol. 2010;152: 55–70. doi:10.1104/pp.109.146670

823 47. Covington BC, McLean JA, Bachmann BO. Comparative mass spectrometry-based metabolomics
824 strategies for the investigation of microbial secondary metabolites. Natural Product Reports. 2017.
825 pp. 6–24. doi:10.1039/c6np00048g

826 48. Corleto KA, Singh J, Jayaprakasha GK, Patil BS. A sensitive HPLC-FLD method combined with
827 multivariate analysis for the determination of amino acids in L-citrulline rich vegetables. J Food
828 Drug Anal. 2019;27: 717–728. doi:10.1016/j.jfda.2019.04.001

829 49. Martínez-Cruz M, Zenteno E, Córdoba F. Purification and characterization of a galactose-specific
830 lectin from corn (Zea mays) coleoptyle. Biochim Biophys Acta - Gen Subj. 2001;1568: 37–44.
831 doi:10.1016/S0304-4165(01)00196-9

832 50. Belcaro G, Cesarone MR, Feragalli B, Cotellese R, Hu S, Dugall M, et al. Management of
833 symptoms, pain and mobility with supplementary managements (including Movardol Forte) in
834 osteoarthrosis: A 6-month, morphology supplement study. Minerva Ortop e Traumatol. 2021;71:
835 160–167. doi:10.23736/S0394-3410.20.04008-4

836 51. Trammell SAJ, Brenner C. Targeted, LCMS-based metabolomics for quantitative measurement of
837 NAD+ metabolites. Computational and Structural Biotechnology Journal. Research Network of
838 Computational and Structural Biotechnology; 2013. p. e201301012. doi:10.5936/csbj.201301012

839 52. Conze DB, Crespo-Barreto J, Kruger CL. Safety assessment of nicotinamide riboside, a form of
840 vitamin B 3. doi:10.1177/0960327115626254

841 53. Mehmel M, Jovanovi´c NJ, Spitz U. Nicotinamide Riboside-The Current State of Research and
842 Therapeutic Uses. [cited 8 May 2021]. doi:10.3390/nu12061616

843 54. Li Y, Wang Y, Wu P. 5'-Methylthioadenosine and Cancer: old molecules, new understanding. J
844 Cancer. 2019;10: 927–936. doi:10.7150/jca.27160

845 55. Shen D, Tian L, Yang F, Li J, Li X, Yao Y, et al. Cell Death Discovery ADO/hypotaurine: a novel
846 metabolic pathway contributing to glioblastoma development. Cell Death Discov. 2021;7: 21.
847 doi:10.1038/s41420-020-00398-5

848 56. Gao P, Yang C, Nesvick CL, Feldman MJ, Sizdahkhani S, Liu H, et al. Hypotaurine evokes a
849 malignant phenotype in glioma through aberrant hypoxic signaling. Oncotarget. 2016;7: 15200–
850 15214. doi:10.18632/oncotarget.7710

851 57. Partyka A, Rodak O, Bajzert J, Kochan J, Nihanski W. The Effect of L-Carnitine, Hypotaurine,
852 and Taurine Supplementation on the Quality of Cryopreserved Chicken Semen. 2017 [cited 8 May
853 2021]. doi:10.1155/2017/7279341

854 58. Wang C, Liu R, Lim GH, De Lorenzo L, Yu K, Zhang K, et al. Pipecolic acid confers systemic
855 immunity by regulating free radicals. Sci Adv. 2018;4. doi:10.1126/sciadv.aar4509

856 59. Cheng J, Chen · Peng, Song A, Wang D, Wang Q. Expanding lysine industry: industrial
857 biomanufacturing of lysine and its derivatives. J Ind Microbiol Biotechnol. 2030;45: 719–734.
858 doi:10.1007/s10295-018-2030-8

859 60. Cheng J, Huang Y, Mi L, Chen W, Wang D, Wang Q. An economically and environmentally
860 acceptable synthesis of chiral drug intermediate l-pipecolic acid from biomass-derived lysine via
861 artificially engineered microbes. J Ind Microbiol Biotechnol. 2018;45: 405–415.
862 doi:10.1007/s10295-018-2044-2

863 61. Schulz M, Hofmann D, Sicker D, Hennig L, Schütz V, Preusche M, et al. Pantoea ananatis
864 converts MBOA to 6-methoxy-4-nitrobenzoxazolin-2(3H)-one (NMBOA) for cooperative
865 degradation with its native root colonizing microbial consortium. Nat Prod Commun. 2018;13:
866 1275–1278. doi:10.1177/1934578x1801301010

867 62. Fenz R, Galensa R, Ernst L. Phenolcarbonsäuren und ihre Glycerinester in Maisgrits. Z Lebensm
868 Unters Forsch. 1992;194: 252–258. doi:10.1007/BF01198417

869 63. Mahajan N, Raina K, Gupta A, N. M, K. R, A. G, et al. High fructose induced adipogenesis and
870 inhibitory potential of trigonelline on murine mesenchymal stem cells: A morphological study. Int
871 J Pharm Sci Res. 2019;10: 528–536. Available: http://ijpsr.com/bft-article/high-fructose-induced-
872 adipogenesis-and-inhibitory-potential-of-trigonelline-on-murine-mesenchymal-stem-cells-a-
873 morphological-study/?view=fulltext

874 64. Srinivasulu C, Ramgopal M, Ramanjaneyulu G, Anuradha CM, Suresh Kumar C. Syringic acid
875 (SA) – A Review of Its Occurrence, Biosynthesis, Pharmacological and Industrial Importance.
876 Biomed Pharmacother. 2018;108: 547–557. doi:10.1016/J.BIOPHA.2018.09.069

877 65. Moreau RA, Nuñez A, Singh V. Diferuloylputrescine and p-coumaroyl-feruloylputrescine,
878 abundant polyamine conjugates in lipid extracts of maize kernels. Lipids. 2001;36: 839–844.
879 doi:10.1007/S11745-001-0793-6

880 66. Jeffryes JG, Colastani RL, Elbadawi-Sidhu M, Kind T, Niehaus TD, Broadbelt LJ, et al. MINEs:
881 Open access databases of computationally predicted enzyme promiscuity products for untargeted
882 metabolomics. J Cheminform. 2015;7: 1–8. doi:10.1186/s13321-015-0087-1

883 67. Greinert T, Vogel K, Maskow T, Held C. New thermodynamic activity-based approach allows
884 predicting the feasibility of glycolysis. Sci Reports |. 123AD;11: 6125. doi:10.1038/s41598-021-
885 85594-8

886 68. Seaver SMD, Liu F, Zhang Q, Jeffryes J, Faria JP, Edirisinghe JN, et al. The ModelSEED
887 Biochemistry Database for the integration of metabolic annotations and the reconstruction,
888 comparison and analysis of metabolic models for plants, fungi and microbes. Nucleic Acids Res.
889 2021;49: D575–D588. doi:10.1093/nar/gkaa746

890 69. Beber ME, Gollub MG, Mozaffari D, Shebek KM, Flamholz AI, Milo R, et al. EQuilibrator 3.0: A
891 database solution for thermodynamic constant estimation. Nucleic Acids Res. 2022;50: D603–
892 D609. doi:10.1093/nar/gkab1106

893 70. Wang L, Xu C, Qu M, Zhang J. Kernel amino acid composition and protein content of
894 introgression lines from Zea mays ssp. mexicana into cultivated maize. [cited 22 May 2021].
895 doi:10.1016/j.jcs.2007.09.014

896 71. Li K, Wen W, Alseekh S, Yang X, Guo H, Li W, et al. Large-scale metabolite quantitative trait
897 locus analysis provides new insights for high-quality maize improvement. Plant J. 2019;99: 216–
898 230. doi:10.1111/tpj.14317

899 72. Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. COCONUT online: Collection of
900 Open Natural Products database. J Cheminform. 2021;13: 1–13. doi:10.1186/S13321-020-00478-
901 9/FIGURES/4

902 73. Kehelpannala C, Rupasinghe T, Hennessy T, Bradley D, Ebert B, Roessner U. The state of the art
903 in plant lipidomics. Molecular Omics. Royal Society of Chemistry; 2021. pp. 894–910.
904 doi:10.1039/d1mo00196e

905 74. Wang S, Tu H, Wan J, Chen W, Liu X, Luo J, et al. Spatio-temporal distribution and natural
906 variation of metabolites in citrus fruits. Food Chem. 2016;199: 8–17.
907 doi:10.1016/J.FOODCHEM.2015.11.113

908 75. Seyfferth C, Renema J, Wendrich JR, Eekhout T, Seurinck R, Vandamme N, et al. Advances and
909 Opportunities in Single-Cell Transcriptomics for Plant Research. https://doi.org/101146/annurev-
910 arplant-081720-010120. 2021;72: 847–866. doi:10.1146/ANNUREV-ARPLANT-081720-010120

911 76. Yang B, He S, Liu Y, Liu B, Ju Y, Kang D, et al. Transcriptomics integrated with metabolomics
912 reveals the effect of regulated deficit irrigation on anthocyanin biosynthesis in Cabernet
913 Sauvignon grape berries. Food Chem. 2020;314. doi:10.1016/j.foodchem.2020.126170

914 77. Agregán R, Echegaray N, López-Pedrouso M, Aadil RM, Hano C, Franco D, et al. Proteomic
915 advances in cereal and vegetable crops. Molecules. MDPI AG; 2021.
916 doi:10.3390/molecules26164924

917 78. Carrera M. Proteomics and food analysis: Principles, techniques, and applications. Foods. MDPI;
918 2021. doi:10.3390/foods10112538

919 79. Ma D, Lai Z, Ding Q, Zhang K, Chang K, Li S, et al. Identification, Characterization and Function
920 of Orphan Genes Among the Current Cucurbitaceae Genomes. Front Plant Sci. 2022;13.
921 doi:10.3389/fpls.2022.872137

922 80. Kautsar SA, Suarez Duran HG, Blin K, Osbourn A, Medema MH. PlantiSMASH: Automated
923 identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic
924 Acids Res. 2017;45: W55–W63. doi:10.1093/nar/gkx305

925 81. Jiang M, Li X, Dong X, Zu Y, Zhan Z, Piao Z, et al. Research Advances and Prospects of Orphan
926 Genes in Plants. Frontiers in Plant Science. Frontiers Media S.A.; 2022.
927 doi:10.3389/fpls.2022.947129

928 82. Opialla T, Kempa S, Pietzke M. Towards a More Reliable Identification of Isomeric Metabolites
929 Using Pattern Guided Retention Validation. Metabolites. 2020;10: 1–16.
930 doi:10.3390/METABO10110457

931 83. Claes BSR, Takeo E, Fukusaki E, Shimma S, Heeren RMA. Imaging Isomers on a Biological
932 Surface: A Review. Mass Spectrom. 2019;8. doi:10.5702/MASSSPECTROMETRY.A0078

933 84. Kranenburg RF, Van Geenen FAMG, Berden G, Oomens J, Martens J, Van Asten AC. Mass-
934 Spectrometry-Based Identification of Synthetic Drug Isomers Using Infrared Ion Spectroscopy.
935 Anal Chem. 2020;92: 7282–7288. doi:10.1021/ACS.ANALCHEM.0C00915

936 85. Baker ES, Patti GJ. Perspectives on Data Analysis in Metabolomics: Points of Agreement and
937 Disagreement from the 2018 ASMS Fall Workshop. J Am Soc Mass Spectrom. 2019;30: 2031–
938 2036. doi:10.1007/S13361-019-02295-3/ASSET/IMAGES/MEDIUM/JS8B06238_0004.JPG

939 86. Determination of Abundant Metabolite Matrix Adducts Illuminates the Dark Metabolome of
940 MALDI-Mass Spectrometry Imaging Datasets. Anal Chem. 2021;93: 8399–8407.
941 doi:10.1021/acs.analchem.0c04720

942 87. Stricker T, Bonner R, Lisaek F, Hopfgartner G. Adduct annotation in liquid chromatography/high-
943 resolution mass spectrometry to enhance compound identification. doi:10.1007/s00216-020

944 88. Vinay CM, Udayamanoharan SK, Prabhu Basrur N, Paul B, Rai PS. Current analytical
945 technologies and bioinformatic resources for plant metabolomics data. Plant Biotechnology
946 Reports. Springer; 2021. pp. 561–572. doi:10.1007/s11816-021-00703-3

947 89. Van Der Hooft JJJ, Mohimani H, Bauermeister A, Dorrestein PC, Duncan KR, Medema MH.
948 Linking genomics and metabolomics to chart specialized metabolic diversity. Chem Soc Rev.
949 2020;49: 3297–3314. doi:10.1039/d0cs00162g

90. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol. 2016;33: 1635–1638. doi:10.1093/molbev/msw046

91. Wohlgemuth G, Haldiya PK, Willighagen E, Kind T, Fiehn O. The Chemical Translation Service —a web-based tool to improve standardization of metabolomic reports. Bioinformatics. 2010;26: 2647. doi:10.1093/BIOINFORMATICS/BTQ476

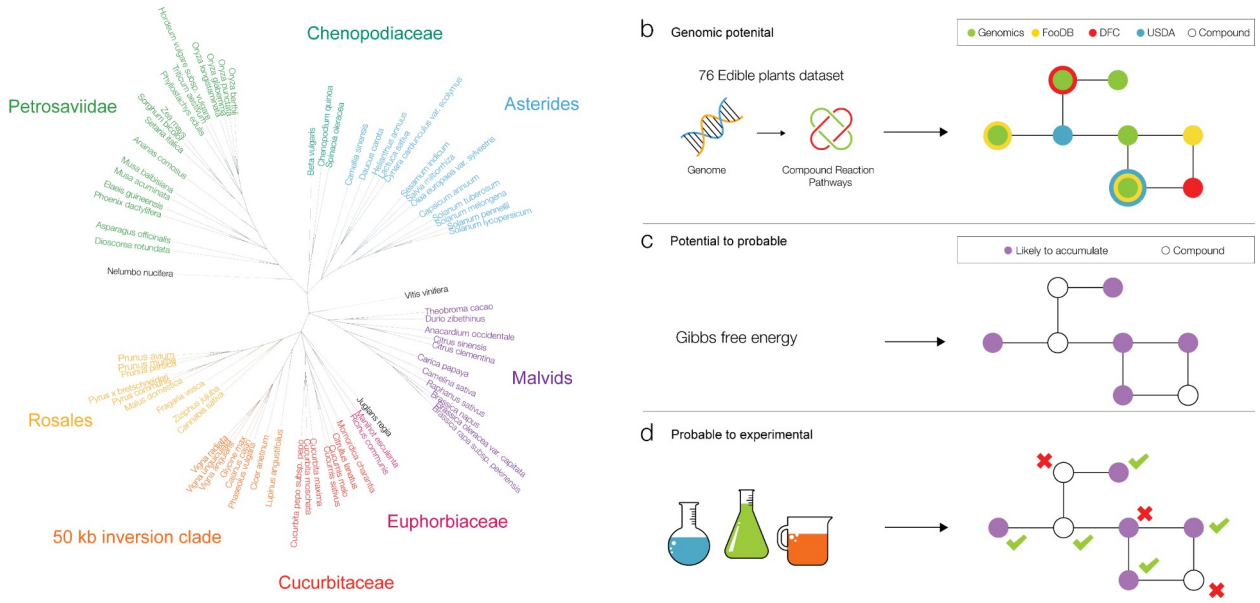
92. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17: 261–272. doi:10.1038/s41592-019-0686-2

93. Landrum G. Rdkit: Open-source cheminformatics software. In: URL http://www. rdkit. org/, https://github. com/rdkit/rdkit. 2016.

List of Legends for Supporting Information Files

1. Supporting Information (PDF) – Contains all supporting text and supporting figures S1-S8.

2. Table S2 (CSV) – Contains supporting metabolomics data used in this work.

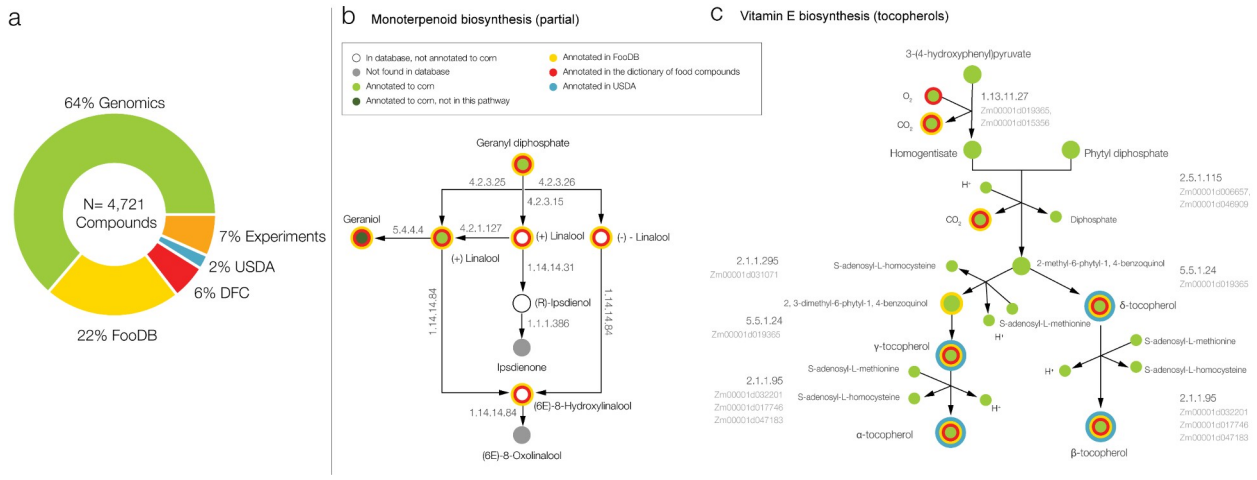


974

975 **Figure 1 – Plant phylogenetic diversity and schematic overview of genomics contribution to**
976 **food composition knowledge. (a)** Phylogenetic tree representing the 75 plants in our collection
977 colored by plant order. **(b)** Collection and evaluation of genomics-based annotations to food
978 composition knowledge. Functional annotations include compounds, reactions, and pathways. **(c)**
979 Kinetics-based annotations help us to infer compounds likely to accumulate and hence
980 experimentally detectable. **(d)** Validation of our kinetics-based approach against new
981 metabolomics experiments, which detected compounds for 13 plants in our collection.

982

983



984 **Figure 2 – Genomics-based annotations boost corn composition knowledge**. **(a)** Database
985 annotations for corn, indicating that some compounds are annotated in multiple databases. The
986 total number (N=4,721) represents the number of unique compounds for corn after the addition
987 of genomics-based annotations. Genomics databases are represented by KEGG and PlantCyc.
988 Other food related databases used in this work are DFC (the Dictionary of Food Compounds),
989 FooDB, and USDA. Finally, experiments denote the set of compounds collected by
990 metabolomics experiments reported here for corn. **(b)** A partial adaptation of the Monoterpenoid
991 biosynthesis pathway in corn (KEGG) showing annotation availability, overlap and gaps in the
992 coverage of different databases and genomics-based annotations. The different colors denote
993 annotation sources as single or multiple concentric circles. **(c)** Vitamin E biosynthesis
994 (tocopherols) pathway in corn (PlantCyc). Circles denote compounds and edges denote reactions.
995 The different colors denote annotation sources as single or multiple concentric circles.

996

997

998

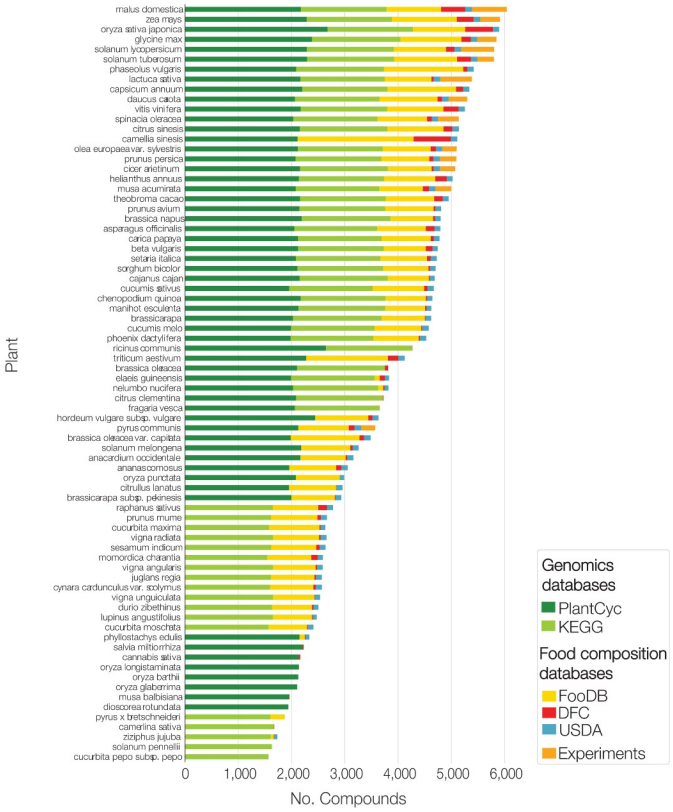


Figure 3 – Contribution of the different sources of food composition across the entire edible plant catalogue. Genomics-based annotation are presented in two shades of green and include the KEGG (light green) and PlantCyc (dark green). Food composition databases include FooDB (yellow), DFC (red), and USDA (blue). Compounds detected in metabolomics experiments are shown in orange. The stacked bars represent here convey the proportions of annotations available from each of the databases/experiments. Annotations from each data source might overlap.

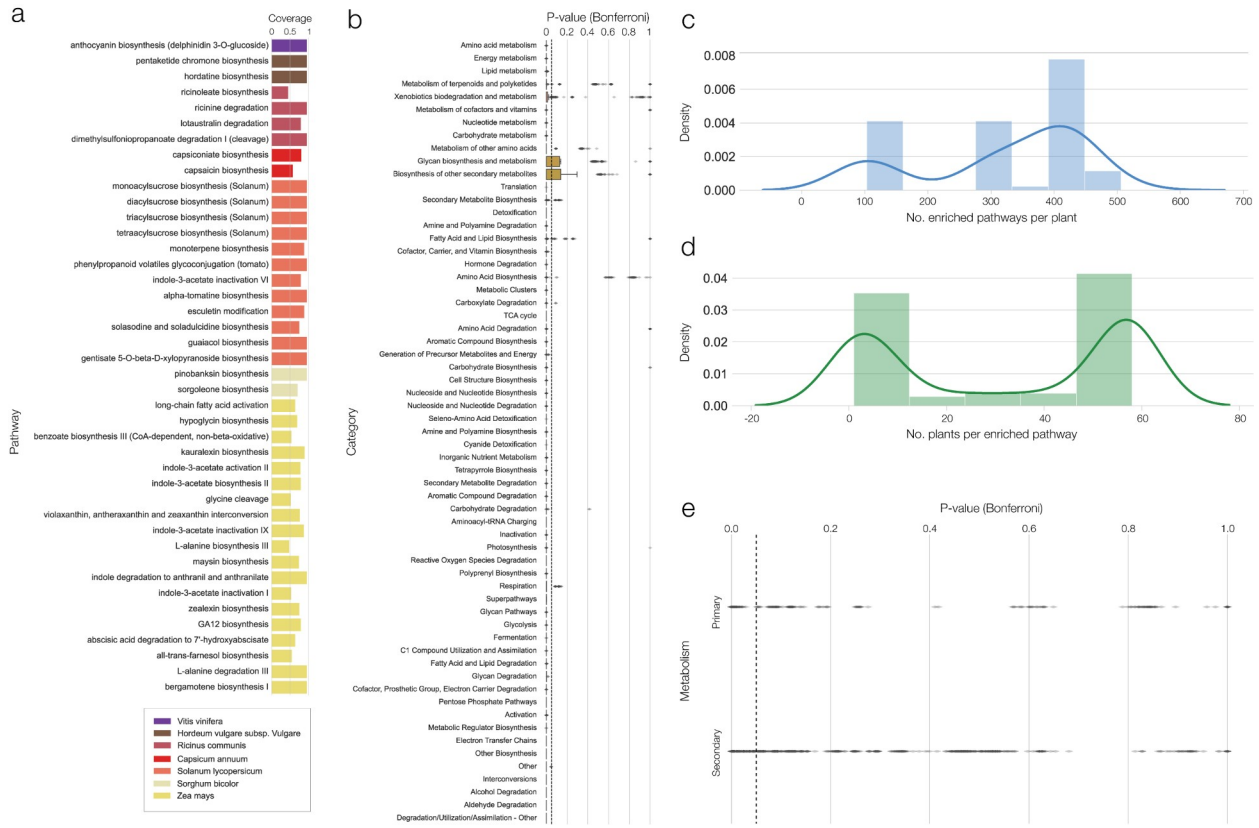


Figure 4 – Pathway enrichment analysis. (a) Plant specific significantly enriched pathways (hypergeometric test, p<0.05, Bonferroni multiple testing correction). The coverage of each pathway is defined by the ratio of present compounds to the total number of compounds in the reference pathway. These include signature known pathways like Capsaicin biosynthesis in pepper (red bars), acyl sugars and alpha tomatine pathways in tomato (light red bars) and maysin and zealexin biosynthesis (yellow bars). **(b)** A boxplot of the p-values (hypergeometric test, p<0.05, Bonferroni multiple testing correction) of pathway categories in corn. Medians found below the dashed line (0.05 enrichment line) represent the enriched pathways. We find that most pathways are enriched, indicating the diversity in our pathway coverage. **(c)** A density plot describing the distribution of enriched pathways per plant. We observed two peaks compatible with the two genomics databases included, KEGG (lower number of maps, each containing a larger number of reactions) and PlantCyc (larger number of maps, each containing a smaller number of reactions) **(d)** A density plot describing the distribution of the number of plants per enriched pathway. While most pathways are found in many plants, we observe several plant specific pathways described in detail in (a) and Figure S3. **(e)** A boxplot of enriched pathways (hypergeometric test, p<0.05, Bonferroni multiple testing correction) classified as primary and secondary metabolism. Medians found below the dashed line (0.05 enrichment line) represent an enriched class.

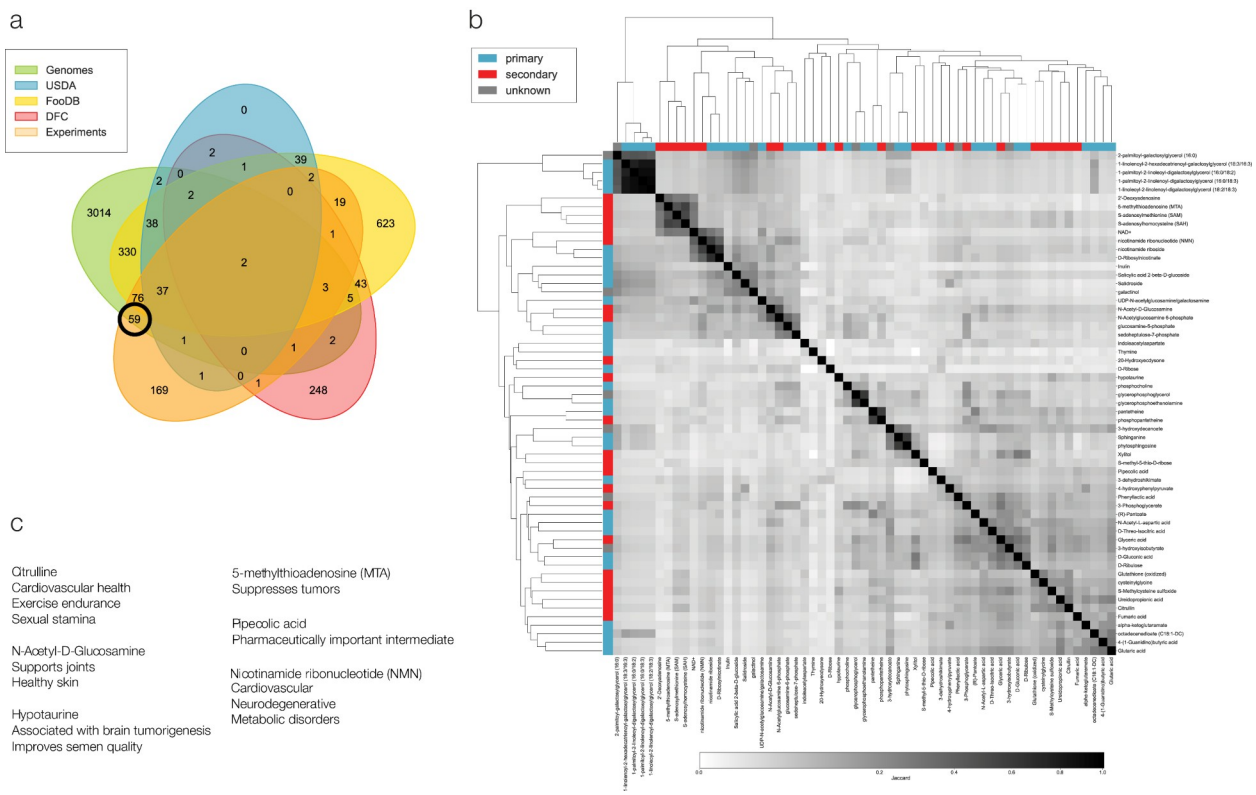


Figure 5 –Bioactive compounds in corn unveiled by genomics. (a) Venn diagram comparing all sources of data contributing to the corn compound collection. The black circle marks the fraction of compounds unique to genomics-based annotations and metabolomics experiments reported here. **(b)** A structural similarity based clustermap of the 59 compounds highlighted by the black circle in panel (a). Compounds are classified to primary (blue) and secondary (red) metabolism according to the pathways they are part of. Greyscale denotes structural similarity as Jaccard distance, 0 being not similar and 1 being composed of the same bits in their vector representation. **(c)** The potential of corn as a nutritional influencer on well-being as learned with the addition of genome-based annotations, associated with antioxidant and anti-inflammatory activity promoting, heart, skin, and metabolic health.

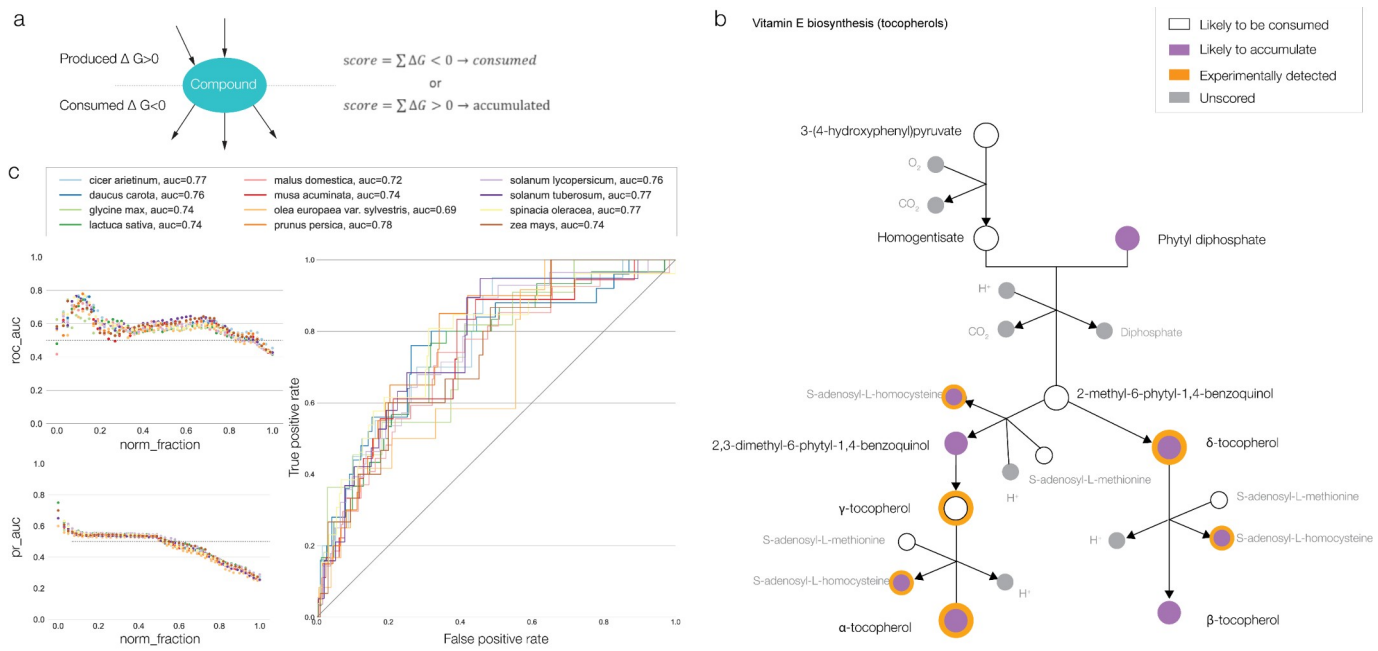


Figure 6 – Application of kinetics-based annotations to predict likely to accumulate compounds. **(a)** Schematic description of the kinetics-based annotations approach. Gibbs free energy values, ΔG, were collected for each reaction in each plant and used to calculate the cumulative score of a compound. If a compound is a reactant in a reaction, it gets a negative ΔG value and if it is a product, it gets a positive ΔG value. **(b)** Vitamin E biosynthesis (tocopherols) pathway of corn (PlantCyc), describing the possible outcomes of our approach. Circles are compounds and edges are reactions. **(c)** Performance of our approach including (i) optimal threshold analysis on the top-ranking compounds predicted to accumulate, establishing the window of best performance denoted by the area under the receiver operator curve, AUC_{ROC} and the area under the precision recall curve, AUC_{PR} for each plant, and (ii) the optimal receiver operating curves for plants included in this analysis.