bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Ofaim Shany[1], Menichetti Giulia[1,2], Sebek Michael[1] and Barabási Albert László[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 wellbeing, 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 genomics to predict the chemical composition of natural ingredients. We therefore developed tools to unveil the chemical composition of 75 edible plants' genomes, finding that genome-based annotations increase the number of compounds linked to specific plants by 42 to 100%. We rely on Gibbs free energy to identify compounds that accumulate in plants, i.e., those that are more likely to be detected experimentally. 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 find 59 novel compounds in corn, predicted by genomics annotations and supported by our experiments,

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

but previously not assigned to the plant. Our study shows that genome-based 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.

Background

"Make every bite count", recommends the U.S Departments of Agriculture (USDA) Dietary guidelines for Americans (2020-2025)[1], reminding us of the multiple roles food, and specifically fruits and vegetables, play in our wellbeing, serving as a source of energy and nutrients, modulating our health, and affecting disease[2–4]. Plants are complex organisms characterized by large genomes. For example, the corn genome (2,280 Mb) encodes 57,181 proteins, annotated to Gene Ontology (GO) categories, helping unveil the biological processes these proteins participate in and their potential molecular functions[5]. The genome can also be a strong predictor of phenotypic traits[6], like color, taste and aroma, known indicators of the nutritional value of a plant[7]. Indeed, color and pigments including carotenoids, betalains and anthocyanins are well known for their bioactive properties[8]. Furthermore, the genetically predetermined polyphenols, alkaloids, carotenoids and phytosterols have well-documented antioxidant and anti-inflammatory activities effecting multiple diseases, from Cancer to diabetes or hypertension[9].

Our current knowledge on food composition is limited to 150 nutrients catalogued by USDA, despite the fact that the true number of chemical compounds in food ranges from tens[10] to hundreds of thousands across all known plant species[11]. The bulk of our knowledge on the chemical composition of food comes from mass spectrometry and other low-throughput

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

analytical methods and is compiled in repositories such as FooDB[12] and The Dictionary of Food Compounds[13] (DFC), cataloguing comprehensive information on the detected compounds, including both evidence-based and predicted annotations.

Our work is driven by the hypothesis that the full list of known and yet unknown chemicals present in plants are encoded in the genome of the respective organism, encapsulating its metabolic capacities. Recent advancements in genomics have resulted in the emergence of extensive annotation efforts to decipher the genetic potential and the metabolic capacities of edible plants. For example, KEGG[14] links genes to their functional annotations such as enzymes, reactions, and chemical compounds and catalogues them in metabolic pathway maps, offering functional annotations for 7,254 organisms across the tree of life, out of which 56 are edible plants. Another contributor to plant genome metabolic annotations is PlantCyc[15], a BioCyc[16] based platform adapted to annotate the functional diversity of plant genomes.

Here we explore to what degree genome-based annotations can offer a valuable resource to expand the knowledge of the compound composition of foods. To do so we rely on metabologenomics[17–19], to integrate genomics and metabolomics, used in the past to discover novel natural products[17,20]. To be specific, we develop a systematic metabologenomics pipeline, coupled with thermodynamic feasibility analysis aiming to predict the composition of edible plants. We validate our predictions by comparing them to the chemical knowledge curated by food composition databases like FooDB, DFC, and USDA[4]. We also collect new experimental data to explore the chemical composition of 13 plants. Our findings indicate that genomics-based annotations offer a predictive platform capable of systematically capturing the chemical

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

composition of plant-based food, from fruits to vegetables and allow us to predict and experimentally test the presence of novel compounds in plants.

Results

The existing knowledge on food composition

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

To estimate the currently available knowledge on the chemical composition of edible plants, we collected compound annotations from FooDB, DFC, and USDA, cataloguing 5,834, 5,151, and 140 compounds respectively, across all plants (Figure S1A). FooDB carries 723±452 compounds on average per plant (median: 850), a number that can be as low as three compounds for clementine (*citrus clementina*) and as high as 2,181 compounds for tea (*Camellia sinensis*). DFC carries 83±121 compounds on average per plant (median:33) with as low as one compound for vegetable marrow (*Cucurbita pepo subsp. pepo*) and red rice (*Oryza punctata*) and as high as 697 compounds for tea. Finally, USDA carries 88±46 compounds on average per plant (median:110), with a single compound for false flax (*Camelina sativa*) and Chinese white pear (*Pyrus × bretschneideri*) and 128 compounds for apple (*Malus domestica*).

Genomics contribution to food composition

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To estimate the contribution of genome-based annotations to the existing knowledge on food composition, we collected plant-related compounds from KEGG and PlantCyc. KEGG stores information about 7,245 organisms, out of which 546 are eukaryotes and 92 are plants, including 56 edible plants. PlantCyc is a plant-oriented database storing 126 genomes, out of which 58 are edible. These databases overlap and complement each other (39 plants overlap), together covering 75 metabolically annotated plant genomes. Overall, KEGG and PlantCyc contributed 1,201 and 3,737 new unique compounds respectively, adding a total of 5,224 new compounds (unique and common) to the composition of plants in our catalogue (Figure S1B).

To illustrate the contribution of genomics-based annotations to edible plants (Figure 1b), we focused on corn (*Zea Mays*), a highly consumed staple crop[21] worldwide and in the US. Existing knowledge for corn includes 1,221 compounds from FooDB, 311 compounds from DFC and 127 compounds from USDA. Considering overlaps between all sources, this compiled to a total of 1,038 unique compounds. Next, we set out to explore the value of adding genome-based annotations to corn's existing knowledge.

One contribution of genomics-based annotations is the metabolic context of these compounds, carried by a network of pathways. Some known pathways are only partially annotated even after considering multiple databases. For example, in Monoterpenoid biosynthesis in corn (Fig 2B) six out of nine compounds are annotated in both databases. Of the three remaining compounds, (R)-Ipsdienol is known to be present in food but was not annotated to any plant in our collection. The two remaining compounds, Ipsdienon, a product of the reaction catalyzed by EC 1.1.1.386 directly from (R)-Ipsdienol, and (6E)-8-Oxolinalool, a product of the reaction catalyzed by EC 1.14.14.84, are currently documented in food but not in corn

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(white circles, Figure 2b). To strengthen the stringency of our work, these compounds, whose presence is documented in food, but not known to be associated with corn, are not included in the plant's catalogue.

Consider another example, the tocopherol biosynthesis pathway. Tocopherols are an important class of compounds for health and nutrition[22] (α-tocopherol better known as vitamin E). Figure 2c shows the tocopherol biosynthesis pathway in corn and the delineated contribution of each database to its compounds. We find that while databases such as FooDB and DFC annotate the lower half of the pathway, capturing products such as vitamin E and its derivatives. In contrast, genomics-based annotations offers a full pathway annotation, adding 4 new intermediates: 3-(4-hydroxyphenyl)pyruvate, homogenistate, phytyl diphosphate and 2-methyl-6-phytyl-1,4-benzoquinol and 2 new cofactors: S-adenosyl-L-homocysteine and S-adenosyl-L-methionine, shedding light on the metabolic processes leading to the production of vitamin E in corn.

Taken together, genomics-based annotations added 3,021 compounds to the list of chemical compounds potentially present in corn, increasing its compound library by 64% (Figure 2a). Across our collection, genomics-based annotations increased the number of compounds by 2,363±728 chemicals on average per plant, an increase of 75%±15%. After this increase, our database documents 3,239±1,054 compounds per plant. We find that some plants are well annotated in FooDB, DFC, and USDA, while others are poorly annotated (Figure 3). Tea (*Camellia sinensis*) contains the highest number of FooDB, DFC, and USDA annotations (2,547 compounds out of a total of 4,379) showing an increase of 42% in new compounds. Some varieties of rice (*Oryza glaberrima*, *Oryza longistaminata*, *Oryza barthii*), white yam (*Dioscorea rotundata*), wild

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tomato (*Solanum pennellii*) and woodland strawberry (*Fragaria vesca*) are not catalogued by FooDB, DFC or USDA, hence for these plants genomics-based annotations contributed 100% of the compounds. Other plants like clementine (*citrus clementina*), lotus (nelumbo nucifera) and african oil palm (*elaeis guineensis*) are poorly annotated (<250 compounds) in FooDB, DFC, and USDA, hence the addition of genomics-based annotations increased our knowledge about their chemical composition by more than 85%.

Pathway enrichment analysis

Genomics-based annotations not only increase our knowledge about the chemical composition of plants but also help us unveil the network of pathways responsible for the production of the newly predicted chemical compounds. Indeed, metabolic pathway mapping allows for a better understanding of the metabolic mechanisms responsible for the synthesis and modulation of natural products and offer a knowledge base towards the prediction of currently undetected compounds.

Metabolic pathways are divided into two main classes: primary and secondary metabolism. Primary metabolism is the collection of pathways involved in growth, energy, and reproduction of a plant, while secondary metabolism captures all other functions[23], like flavonoid biosynthesis, xenobiotics metabolism or plant-hormone metabolism to name only a few. For corn we collected 789 pathways out of which 35% belong to primary and 65% belong to secondary metabolism. We asked if certain pathways are represented more than others in our catalogue. To test for pathway bias we performed a hypergeometric enrichment test capturing the chance for a set of compounds mapped to a pathway in a certain plant to exceed the expected overlap

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

with the general reference pathway (p-value <0.05) (Figure S2). In corn, we found 479 enriched pathways spanning both primary (45%) and secondary metabolism (55%), indicating that our corn dataset is metabolically diverse.

We next asked about specialized metabolism occurring only in corn, scanning our plant collection for enriched pathways specific to it (Figure 4a). Corn-specific pathways include specialized metabolism like kauralexin and zealexin biosynthesis[24,25], maysin biosynthesis[26], bergamotene biosynthesis[27] and all-trans-farnesol biosynthesis[28]. Each of these natural products of the diterpenoid, volatile sesquiterpenes and flavone families are produced by the plant to acquire resistance against biotic and abiotic stress, such as herbivore attack.

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

We found 762 enriched pathways across all plants (p-value<0.05), out of which 34% are related to primary and 66% are related to secondary metabolism. The distribution of the number of enriched pathways per plant shows two peaks (Figure 4c), corresponding to the two databases, KEGG and PlantCyc. The peaks capture the fact that KEGG has fewer, more complex and generic pathway map representations while PlantCyc has a larger number of smaller pathways. For

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

example, KEGG represents the metabolism of alanine, aspartate and glutamate in one map while PlantCyc breaks down this process into 10 smaller pathways.

Different plants are known for their production of specialized compounds and natural products. Examples include the production of curcumin by turmeric, thymol by thyme and vanillin by vanilla plants[29], prompting us to identify pathways enriched in single plants, pointing towards unique functionalities. Indeed, we observed a bimodal distribution of the number of plants per enriched pathway, one of the peaks being closer to 60 plants and another around 5 plants, indicating the presence of specialized plant metabolism in our catalogue (Figure 4c). Overall, we find that 30 out of 75 plants have at least one plant-specific enriched pathway (Figure S3) out of which we explored seven (Figure 4a). An example of such unique pathway is anthocyanin biosynthesis (delphinidin 3-O-glucoside) in grapes (*Vitis vinifera*), present in seeds and grape skins and used to differentiate between types of wine[30,31]. Other examples include: (1) hordatine biosynthesis in Barley (*Hordeum vulgare subsp. vulgare*). Hordatine, an antifungal compound, highly abundant in young barley shoots, reported to be a potential inhibitor of two main COVID-19 proteins, a protease (PDB ID: 7BQY) and a RNA polymerase (PDB ID: 7bV2)[32] and found in measurable quantities in different types of beer[33]. (2) Ricinoleate biosynthesis, found in castor bean (*Ricinus communis*) and the main constituent of castor oil, was shown to have antibacterial activities, and its elastic properties make it a candidate for packaging polymers with potential applications in biomedical and food technology[34]. (3) Capsaicin, a specialized component in pepper (*Capsicum annuum*)[35] was recently shown to have positive dietetic effects and beneficial antioxidant activity through association with the gut microbiome[36]. (4) Sorgoleone, an allelochemical exuded from sorghum (*Sorghum bicolor*) roots is known to effect both microbial

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

communities and neighboring plant growth[37]. Finally, we find a variety of acyl-sugar biosynthesis pathways in tomato (*Solanum lycopersicum*). Acyl-sugars are created in tomato trichomes and have commercial and medicinal uses[38]. Another unique pathway in tomato is phenylpropanoid volatiles glycoconjugation, a pathway describing specialized volatiles found in tomato fruits contributing to its signature smell[39].

In summary, genomics-based annotations bring diverse metabolic information from both primary and secondary metabolism, offering evidence of new specialized compounds.

Experimental confirmation of genomics contribution to food knowledge

To experimentally test the capacity of plant genomics to predict the presence of compounds in plants, we performed untargeted metabolomics experiments on 13 out of the 75 plants in our collection, resulting in a catalogue of 939 detected compounds (see Methods). On average, 371±130 compounds were detected per plant, ranging from 264 in pear to 652 in apple, and 370 compounds for corn (39.4% of our experimental catalogue). These experimental results allow us to evaluate the accuracy of genomics-based predictions. For this, we measured the overlap of the chemical structures between experimentally detected compounds and our known genomics-based compound collection, finding that genomics-based annotations show a significant overlap with experiments (p-value= 0.018, SI section 1).

To be specific, we identify 59 compounds that are found only in genomics-based annotations and were also detected in our experiments results (Figure 5a). We clustered these new compounds based on their chemical structure and classified them into primary and secondary metabolism (Figure 5b, Figure S5), finding that the majority of compounds were

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

attributed to primary metabolism (30) and might represent pooled intermediates and their possible fragments. Metabolite detection may also vary depending on the compound and spectra libraries used in identification. As primary compounds are better studied and annotated, secondary metabolites and their fragmentation products are less abundant in spectra libraries[40]. Overall, we observed well characterized fractions of lipids, cofactors, sugar and amino acid derivatives.

We identified seven compounds, present in corn that might potentially affect human well-being (Figure 5c). For example, Citrulline, a non-essential amino acid known to effect cardiovascular health and dilate blood vessels, and primarily found in watermelon and in smaller amounts in other fruits including a variety of corn species[41], is marketed as a dietary supplement for bodybuilders and athletes to improve exercise endurance. N-Acetyl-D-Glucosamine, previously detected in the shell that protects the first leaf of a corn shoot[42], is known to help support the joints and may help promote healthy skin[43]. Another compound detected is nicotinamide ribonucleotide (NMN), reported to be detected together with NAD+, a well-known cofactor, always present in the cell[44]. Well-being benefits related to this compounds stem from its effects on NAD+ content. This compound has been well studied and as a member of the vitamin B3 family[45] and is being used for the treatment of a number of cardiovascular, neurodegenerative and metabolic disorders[46]. Other compounds detected in corn and unveiled by genomics-based annotations have potential cancer related effects. For example, 5-methylthioadenosine (MTA), a sulfur-containing nucleoside has recently reported as tumor suppresser[47]. Another cancer associated compound, hypotaurine, a sulfinic acid with antioxidant properties, derived from cysteine and an intermediate in taurine production, is one of the top

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ranked metabolites for differentiating low and high grade tumors[48] but was also shown to evoke a malignant phenotype in glioma, the most common primary brain malignancies in adults[49]. It is also marketed as a dietary supplement, together with taurine and L-carnitine and associated with semen quality improvement[50]. Some other compounds observed in this subset were suggested to have both a well-being and economic/industrial importance. For example, Pipecolic acid, a product of lysine metabolism, is an important regulator of immunity in plants and humans. In plants, it accumulates upon pathogen infection and is associated with systemic acquired resistance (SAR)[51]. Pipecolic acid is also an important intermediate of pharmaceutically and biologically derived compounds such as immunosuppressive agents and antibiotics[52,53].

Our experimental investigation also unveiled 22 compounds found only in FooDB, DFC and USDA. These include: (1) 6-Methoxy-2(3H)-benzoxazolone (MBOA), a degradation product of the known bioactive compound 2,4-dihydroxy-7-methoxy-2H-1,4-benzoxazin-3(4H)-one (DIMBOA). DIMBOA is the main benzoxazinone synthesized in young corn tissues and accumulates in the cells. It is exuded by the roots and acts as a biocide against pests and as an attractant for soil bacteria. MBOA, the more stable form, is often detected in corn soils[54]. Interestingly, while DIMBOA is annotated to the corn genome, MBOA, its derivative is not. (2) Vanillic acid, a phenolic detected in corn grits[55], (3) Trigonelline, an active alkaloid known to be found in corn and associated with antioxidant, anti-carcinogenic, anti-diabetic and anti-hypercholesterolemia properties[56], (4) Syringic acid, a phenolic compound found in fruits and vegetables including corn and reported to have anti-oxidant, antimicrobial, anti-inflammatory and antiendotoxic properties[57], and (5) Feruloylputrescine, a polyamine monoconjugate previously detected in corn kernels[58].

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, we considered all the plants in our catalogue for which we have experimentally detected compounds. We find that the number of compounds added by genomics-based annotations is typically higher than the number already catalogued by FooDB, DFC, and USDA (49±15 and 26±14 compounds respectively, Figure S6). In other words, our analysis shows that genomics-based annotations significantly enhance existing knowledge of edible plants' chemical composition, helping us uncover potentially novel bioactive compounds.

Genomics annotations contribute to the feasibility of compound accumulation

The number of currently identified compounds detected by metabolomics is limited by instrumentation, standard libraries, and analysis pipelines. As a final step, we set out to explore genomics-based annotations' contribution to our ability to predict compounds that accumulate in a plant. Indeed, chemicals that accumulate are more likely to be present in sufficient quantity to be experimentally detected or to enter the bloodstream, potentially modulating health. Indeed, transient compounds may be harder to detect. To study the likelihood of a compound to accumulate we used Gibbs free energy (ΔG) reaction values combined with a genome-scale metabolic network topology. To establish thermodynamic feasibility, we determine the probability that a compound accumulates given all the reactions it takes part in the network context (Figure 6a). In other words, thermodynamic feasibility offers a method for compound ranking based on cumulative ΔG values. We collected ΔG values from modelSEED[59] and PlantCyc and calculated the cumulative ΔG value (score) for each compound. If the compound acted as a reactant in a reaction, we assume that it is consumed, hence it is a transient compound, assigned a negative ΔG value. In contrast, if the compound is a product of the reaction, it is assigned a

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

positive value. Compounds with positive scores are produced more than consumed, thus they are likely to accumulate. The reaction representing the largest ΔG value is called a sink reaction, as it shifts the balance largely towards either the consumption or production of a compound. In corn, we scored 2,985 compounds (63% of total compounds), out of which 1,460 have positive ΔG values, i.e., are expected to accumulate.

To illustrate our findings, we explored the compounds on the vitamin E biosynthesis pathway in corn (Figure 6b). We observed likely accumulation of pathway products α and β tocopherol and of the intermediates γ and δ tocopherol. Intermediates shown to accumulate are likely involved in more than one reaction outside the pathway. For example, phytyl diphosphate is involved in 7 reactions and 5 pathways. The largest ΔG value measured for this set of reactions representing the sink was annotated to the phytyl salvage pathway, describing the conversion of degraded chlorophyl to phytyl phosphate, contributing to its high likelyhood to accumulate. A compound could likely accumulate if it is an intermediate appearing in two reactions with a large ΔG value difference. For example, 2,3-dimethyl-6-phytyl-1,4-benzoquinol is involved in two reactions, both annotated to the vitamin E biosynthesis pathway, where the ΔG value for the reaction producing it (20.64 kcal/mol) is larger than the ΔG value for the reaction consuming it (-12.05 kcal/mol). Overall, 1,754±588 compounds were scored per plant, covering an average of 79±17% of its compounds. The fraction of scored compounds was as high as 96.5% of the total compounds in *Oryza longstaminata*, a species of rice, and as low as 50% in adzuki bean (*Vigna angularis*).

To estimate the predictive power of our approach, we first asked if kinetics-based annotations have increased the predictive power of our platform compared to total genomics-

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

based annotations. We find that Kinetics-based annotations show a more significant overlap with the experiments compared to genomics (p-value=0.0018, SI section 1, Figure S4), and they are characterized by a high degree of structural similarity, significantly different from a random sample of the same size from genomics annotations (p-value<0.001, SI section 1).

Next, we used the experimentally detected compounds in the 13 plants for which we performed untargeted metabolomics, to estimate the performance of our approach. Similar to known machine learning methods, we use ΔG scores as a 'classifier' predicting the likelihood of a compound to accumulate. We then compare it to our experimentally detected compound catalogue as ground truth values (a binary classification denoting presence or absence). We calculated standard performance metrics, such as the true positive and false positive rates and the area under the receiver-operator curve (ROC), AUC_{ROC}. In addition, we calculated the precision, recall and F-1 scores. Since our data may be imbalanced, we initially set the threshold of prediction to be larger than zero (positive values). We then performed a moving threshold analysis (see Methods) to determine the optimal threshold for best performance in each plant found in both our annotation and experimental catalogue (Table S1). Most AUC_{ROC} values were above the discrimination line (0.5), several representing acceptable discrimination (AUC_{ROC} values between 0.69 to 0.76) (Figure 6c). Overall, AUC_{ROC} values were better for the compounds that are expected to accumulate than the compounds predicted based on the whole genome, confirming the predictive power of this thermodynamics-based analysis.

Finally, we selected the top ranked 110 corn compound families with the best performance of our thermodynamics-based approach (AUC_{ROC}=0.74). This list contained 15 experimentally detected compounds including AMP, S-adenosylhomocysteine (SAH) and 5-

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

methylthioadenosine (MTA), which are cofactors maintained as constant pools in the cell. Other compounds such as succinate, Glycerol 3-phosphate, and 3-phospho-D-glycerate are key products of major energy producing pathways, such as carbon fixation by photosynthesis, glycolysis and the TCA cycle. Interestingly, L-glutamate and D-Galacturonic acid were also detected. L-glutamate was previously reported as a key metabolite in corn, measured in the large amounts in the endosperm[60,61]. D-Galacturonic acid is the main component of pectin, a polysaccharide naturally found in plant cell walls. Thus, both compounds are likely to accumulate and be detected in metabolomics measurements, supporting the predictive power of our approach.

Discussion

Here we developed a systematic methodology to extract the contribution of genomics-based annotations to the molecular composition of foods. We found that genomics-based annotations not only boosted, in some cases by more than 85%, the number of compounds known to be present in a plant (in comparison to FooDB, DFC, and USDA databases) but also offered valuable mechanistic knowledge in the form of chemical structures and the metabolic pathways responsible for their production. Using multiple types of annotations (compound, reaction and pathway) we surveyed the contribution of genomics in depth. These annotations, combined with experimentally detected compounds, were used to gain new insights into the chemical composition of edible plants, specifically corn.

The molecular composition of plants changes in time and in response to biotic and abiotic stresses such as environmental conditions, herbivore and pathogen attacks and hormonal

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

signals, all governed by complex genetic and metabolic regulation. We therefore set out to explore the feasibility of a compound to accumulate, affecting the likelihood of being experimentally detectable. We find good discriminative power, supporting the large-scale use of thermodynamic annotations. As experimental data and thermodynamics annotations expand, they may lead to significant enhancement in the predictive power of metabologenomics.

While the advent of genomics offers new insights into the composition of edible plants, it is not without limitations. Various biases that might arise from such data were explored throughout this work, representing only a few of the multiple factors that might affect our knowledge of food composition. One major limitation is the availability of annotated plant genomes. While the cost of sequencing dropped significantly, we find that the number of non-model plant genomes annotated remains limited and grows slowly. Other factors such as annotation quality and database standardization can also limit omics-based analysis. As shown here, two major databases, KEGG and PlantCyc, has introduced some redundancy in pathway mappings, partially rooted in the different pathway definitions used. As data continues to accumulate, standardization and mappings between the different data sources is key to deriving new insights. Finally, the use of metabolomics to detect compounds in edible plants highly depends on the instrumentation used, standard libraries and identification methods. For example, it is known that mass-spectrometers have poor detection of stereoisomers[62–64]. Even though considerable efforts have been made to establish spectral analysis pipelines, high-throughput metabolomics compound identification remains limited. Future improvements in metabolomics analysis and annotation could significantly enhance our knowledge of specialized metabolites and natural products in edible plants.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, the genomics-based annotation analyzed here greatly contribute to our existing knowledge of the composition of edible plants. The integration of genomics and metabolomics has been suggested as a promising combination towards the and identification of promising compounds[65]. Thus, this work offers a steppingstone towards a better understanding of food composition, offering insights based on fast-growing computational and experimental datasets.

Methods

Data collection

Compound annotations were collected from open-source (FooDB and USDA) and proprietary (DFC) databases. Genomics based compound, reaction and pathway annotations were collected from KEGG and PlantCyc. ΔG values were gathered from PlantCyc and modelSEED. All annotations were completed with SMILES, InChIKey and mass. Plant diversity was analyzed and visualized using the ETE3 toolkit[66] tree functions (ete3, version 3.1.2). All taxonomic data originated from NCBI. Overall, our catalogue includes 15,296, out of which 12,662 first block InChi keys representing chemical families. The database files are available on https://github.com/Barabasi-Lab/Plant-genomics

Mass spectrometry experiments

A selection of 13 produce items were purchased from two local grocery stores (Whole Foods Market and Stop & Shop): apple, banana, basil, black bean, carrot, chickpea, corn, garlic, lettuce, olive, onion, peach, pear, pepper, potato, spinach, soybean, strawberry, sugar beet, and

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tomato. Each produce item sample contained the combined material of six units (for example 6 apples) and was prepared in a humidity-controlled room with minimum light exposure. Sample preparation included peeling, chopping, freeze drying (-80°C for 24 hrs, Catalog No. 10-269-56B from LabConco/Fisher) and pulverizing into a fine homogenized powder (Kitchen Aid, 170W, Model No. BCG111OBO). All samples were prepared by the Giese lab (Northeastern University, Boston MA USA). The final samples were stored at -80 °C in vials containing 200 mg of powder and argon gas. The samples were shipped to two metabolomics centers (West Coast Metabolomics Center, UC Davis, CA USA and Metabolon, Morrisville, NC USA) for analysis on multiple platforms including UHPLC-CSH C18-HRMS-Orbitrap, UHPLC-BEH Amide-HRMS-Orbitrap, UHPLC-PFP-HRMS-Orbitrap, and Shimadzu LC and SelexION QTRAP MS and annotation (see SI section 2).

Experimental results were annotated by spectral matching or identified by a reference standard library. Results from both metabolomics centers were merged and standardized to InChIKeys (PubChem). Since metabolomics methodologies does not account for stereochemistry, only the first block of the InChIKey is used to compare two entries. Therefore, the resulting list is one of unique compound structures found in each food item.

Statistics and enrichment tests

Enrichment analysis was performed with the hypergeometric distribution test (python 3.8, scipy[67] 1.6.3). Pathway enrichment test was performed per organism. To that end, we first collected all the compound annotations for the reference pathways from KEGG and PlantCyc. including all known annotations attributed to that pathway. We calculated the number of

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

compounds found in each pathway in the organism and tested it for significance against the reference. To account for multiple testing on the same pathway, Bonferroni correction was applied to pathways' p-values.

Compound similarity across genomics, kinetics, and experiments

To investigate the degree of structural similarity and overlap between molecules retrieved by different techniques, we performed similarity search and clustering on a variety of molecular fingerprints (FP). We leveraged the python package RDKit[68] to standardize SMILES and InChIKeys associated with each chemical annotation, using Morgan fingerprints (FP). To 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[69].

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* chemical compounds, by calculating the *intrinsic dimension* of their Jaccard similarity matrix $\{S_{ij}\}$, a function of the spectrum $\{\lambda_i\}$ of $\{S_{ij}\}$,

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

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. The higher is n the more chemically diverse is the 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[59] and PlantCyc[15].

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

Performance evaluation

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

This publication has received funding from the Rockefeller Foundation's 2109 FOD 026 grant. 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Author contributions

S.O performed the analysis and wrote the manuscript. G.M performed structural analysis, structural enrichment, developed the intrinsic dimensionality approach to characterize the structural redundancy of the Jaccard similarity matrix, and contributed to writing the manuscript. M.S collected and analyzed the metabolomics experimental results. A.L.B conceived the project, supervised, and contributed to the final version of this manuscript. All authors discussed the results and contributed to the final manuscript.

Competing interests

A.L.B. is a scientific founder of Scipher Medicine, Inc., which applies network medicine strategies to personalized drug selection, Foodome, Inc., which applies data science to food and health, and Datapolis, Inc., which focuses on human mobility.

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

References

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. Dietary Guidelines for Americans | USDA-FNS.

2. Luo, Y., Shang, P. & Li, D. Luteolin: A Flavonoid that has multiple cardio-protective effects and its molecular mechanisms. *Front. Pharmacol.* **8**, 1–10 (2017).

3. Barabási, A.-L., Menichetti, G. & Loscalzo, J. The unmapped chemical complexity of our diet. *Nat. Food* **1**, 33–37 (2020).

4. Hooton, F., Menichetti, G. & Barabási, A.-L. Exploring food contents in scientific literature with FoodMine. *Sci. Rep.* **10**, 16191 (2020).

5. Wimalanathan, K. & Lawrence-Dill, C. J. Gene Ontology Meta Annotator for Plants (GOMAP). *Plant Methods 2021 171* **17**, 1–14 (2021).

6. Choi, I. Y., Kwon, E. C. & Kim, N. S. The C- and G-value paradox with polyploidy, repeatomes, introns, phenomes and cell economy. *Genes and Genomics* **42**, 699–714 (2020).

7. Nyonje, W. A. *et al.* Precision phenotyping and association between morphological traits and nutritional content in Vegetable Amaranth (Amaranthus spp.). *J. Agric. Food Res.* **5**, 100165 (2021).

8. Sharma, S., Katoch, V., Kumar, S. & Chatterjee, S. Functional relationship of vegetable colors and bioactive compounds: Implications in human health. *J. Nutr. Biochem.* **92**, (2021).

9. Otunola, G. A. & Martiryosan, D. Choosing suitable food vehicles for functional food products. *Funct. Foods Heal. Dis.* **11**, 44–55 (2021).

10. Moghe, G. D., Leong, B. J., Hurney, S. M., Jones, A. D. & Last, R. L. Evolutionary routes to biochemical innovation revealed by integrative analysis of a plant-defense related

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

specialized metabolic pathway. *Elife* **6**, (2017).

11. Vermeulen, R., Schymanski, E. L., Barabási, A.-L. & Miller, G. W. The exposome and health: Where chemistry meets biology. *Science (80-.).* **367**, 392–396 (2020).

12. FooDB. Available at: https://foodb.ca/. (Accessed: 25th June 2019)

13. Yannai, S. *Dictionary of food compounds with CD-ROM*. (Crc Press, 2012).

14. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).

15. Schläpfer, P. *et al.* Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiol.* **173**, 2041–2059 (2017).

16. Karp, P. D. *et al.* The BioCyc collection of microbial genomes and metabolic pathways. *Brief. Bioinform.* **20**, 1085–1093 (2019).

17. Soldatou, S. *et al.* Comparative Metabologenomics Analysis of Polar Actinomycetes. *Mar. Drugs 2021, Vol. 19, Page 103* **19**, 103 (2021).

18. Parkinson, E. I. *et al.* Discovery of the Tyrobetaine Natural Products and Their Biosynthetic Gene Cluster via Metabologenomics. *ACS Chem. Biol.* **13**, 1029–1037 (2018).

19. Saad, H. *et al.* Nocathioamides, Uncovered by a Tunable Metabologenomic Approach, Define a Novel Class of Chimeric Lanthipeptides. *Angew. Chemie Int. Ed.* **60**, 16472–16479 (2021).

20. Fox Ramos, A. E., Evanno, L., Poupon, E., Champy, P. & Beniddir, M. A. Natural products targeting strategies involving molecular networking: Different manners, one goal. *Nat. Prod. Rep.* **36**, 960–980 (2019).

21. Ranum, P., Peña-Rosas, J. P. & Garcia-Casal, M. N. Global maize production, utilization,

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and consumption. *Ann. N. Y. Acad. Sci.* **1312**, 105–112 (2014).

22. Milanlouei, S. *et al.* A systematic comprehensive longitudinal evaluation of dietary factors associated with acute myocardial infarction and fatal coronary heart disease. *Nat. Commun.* **11**, 1–14 (2020).

23. Aharoni, A., Galili, G., Avni, A. & Blazquez, M. Metabolic engineering of the plant primary-secondary metabolism interface This review comes from a themed issue on Plant biotechnology Edited. *Curr. Opin. Biotechnol.* **22**, 239–244 (2011).

24. Murphy, K. M., Ma, L. T., Ding, Y., Schmelz, E. A. & Zerbe, P. Functional characterization of two class ii diterpene synthases indicates additional specialized diterpenoid pathways in maize (zea mays). *Front. Plant Sci.* **871**, 1–12 (2018).

25. Schmelz, E. A. *et al.* Biosynthesis, elicitation and roles of monocot terpenoid phytoalexins. *Plant J.* **79**, 659–678 (2014).

26. Moore, V. M. & Tracy, W. F. Combining ability of husk extension, maysin content, and corn earworm resistance. *J. Am. Soc. Hortic. Sci.* **146**, 14–23 (2021).

27. Köllner, T. G., Degenhardt, J. & Gershenzon, J. The product specificities of maize terpene synthases tps4 and tps10 are determined both by active site amino acids and residues adjacent to the active site. *Plants* **9**, (2020).

28. Schnee, C., Köllner, T. G., Gershenzon, J. & Degenhardt, J. The maize gene terpene synthase 1 encodes a sesquiterpene synthase catalyzing the formation of (E)-β-farnesene, (E)-nerolidol, and (E,E)-farnesol after herbivore damage. *Plant Physiol.* **130**, 2049–2060 (2002).

29. Chiorcea-Paquim, A. M., Enache, T. A., De Souza Gil, E. & Oliveira-Brett, A. M. Natural

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

phenolic antioxidants electrochemistry: Towards a new food science methodology. *Compr. Rev. Food Sci. Food Saf.* **19**, 1680–1726 (2020).

30. Kyraleou, M. *et al.* Discrimination of five Greek red grape varieties according to the anthocyanin and proanthocyanidin profiles of their skins and seeds. *J. Food Compos. Anal.* **92**, 103547 (2020).

31. Tang, K., Liu, T., Han, Y., Xu, Y. & Li, J. M. The Importance of Monomeric Anthocyanins in the Definition of Wine Colour Properties. *South African J. Enol. Vitic.* **38**, 1–10 (2017).

32. Dahab, M. A., Hegazy, M. M. & Abbass, H. S. Hordatines as a Potential Inhibitor of COVID-19 Main Protease and RNA Polymerase: An In-Silico Approach. *Nat. Products Bioprospect.* **10**, 453–462 (2020).

33. Pihlava, J. M., Kurtelius, T. & Hurme, T. Total hordatine content in different types of beers. *J. Inst. Brew.* **122**, 212–217 (2016).

34. Totaro, G. *et al.* Elastomeric/antibacterial properties in novel random Ricinus communis based-copolyesters. *Polym. Test.* **90**, 106719 (2020).

35. Scossa, F., Roda, F., Tohge, T., Georgiev, M. I. & Fernie, A. R. The Hot and the Colorful: Understanding the Metabolism, Genetics and Evolution of Consumer Preferred Metabolic Traits in Pepper and Related Species. *CRC. Crit. Rev. Plant Sci.* **0**, 1–43 (2019).

36. Sinisgalli, C. *et al.* The Beneficial Effects of Red Sun-Dried Capsicum annuum L. Cv Senise Extract with Antioxidant Properties in Experimental Obesity are Associated with Modulation of the Intestinal Microbiota. *Mol. Nutr. Food Res.* **65**, 1–13 (2021).

37. de Oliveira, I. F. *et al.* Sorgoleone concentration influences mycorrhizal colonization in sorghum. *Mycorrhiza* 259–264 (2020). doi:10.1007/s00572-020-01006-1

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

38. Schilmiller, A. L., Gilgallon, K., Ghosh, B., Jones, A. D. & Last, R. L. Acylsugar acylhydrolases: Carboxylesterase-catalyzed hydrolysis of acylsugars in tomato trichomes. *Plant Physiol.* **170**, 1331–1344 (2016).

39. Tikunov, Y. M., de Vos, R. C. H., Paramás, A. M. G., Hall, R. D. & Bovy, A. G. A role for differential glycoconjugation in the emission of phenylpropanoid volatiles from tomato fruit discovered using a metabolic data fusion approach. *Plant Physiol.* **152**, 55–70 (2010).

40. Covington, B. C., McLean, J. A. & Bachmann, B. O. Comparative mass spectrometry-based metabolomics strategies for the investigation of microbial secondary metabolites. *Natural Product Reports* **34**, 6–24 (2017).

41. Trehan, S., Singh, N. & Kaur, A. Diversity and relationship among grain, flour and starch characteristics of Indian Himalayan colored corn accessions. *J. Food Sci. Technol.* **57**, 3801–3813 (2020).

42. Martínez-Cruz, M., Zenteno, E. & Córdoba, F. Purification and characterization of a galactose-specific lectin from corn (Zea mays) coleoptyle. *Biochim. Biophys. Acta - Gen. Subj.* **1568**, 37–44 (2001).

43. Belcaro, G. *et al.* Management of symptoms, pain and mobility with supplementary managements (including Movardol Forte) in osteoarthrosis: A 6-month, morphology supplement study. *Minerva Ortop. e Traumatol.* **71**, 160–167 (2021).

44. Trammell, S. A. J. & Brenner, C. Targeted, LCMS-based metabolomics for quantitative measurement of NAD+ metabolites. *Computational and Structural Biotechnology Journal* **4**, e201301012 (2013).

45. Conze, D. B., Crespo-Barreto, J. & Kruger, C. L. Safety assessment of nicotinamide

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

riboside, a form of vitamin B 3. doi:10.1177/0960327115626254

46. Mehmel, M., Jovanovi´c, N. J. & Spitz, U. Nicotinamide Riboside-The Current State of Research and Therapeutic Uses. doi:10.3390/nu12061616

47. Li, Y., Wang, Y. & Wu, P. 5'-Methylthioadenosine and Cancer: old molecules, new understanding. *J. Cancer* **10**, 927–936 (2019).

48. Shen, D. *et al.* Cell Death Discovery ADO/hypotaurine: a novel metabolic pathway contributing to glioblastoma development. *Cell Death Discov.* **7**, 21 (2021).

49. Gao, P. *et al.* Hypotaurine evokes a malignant phenotype in glioma through aberrant hypoxic signaling. *Oncotarget* **7**, 15200–15214 (2016).

50. Partyka, A., Rodak, O., Bajzert, J., Kochan, J. & Nihanski, W. The Effect of L-Carnitine, Hypotaurine, and Taurine Supplementation on the Quality of Cryopreserved Chicken Semen. (2017). doi:10.1155/2017/7279341

51. Wang, C. *et al.* Pipecolic acid confers systemic immunity by regulating free radicals. *Sci. Adv.* **4**, (2018).

52. Cheng, J., Chen, · Peng, Song, A., Wang, D. & Wang, Q. METABOLIC ENGINEERING AND SYNTHETIC BIOLOGY-REVIEW Expanding lysine industry: industrial biomanufacturing of lysine and its derivatives. *J. Ind. Microbiol. Biotechnol.* **45**, 719–734 (2030).

53. Cheng, J. *et al.* An economically and environmentally acceptable synthesis of chiral drug intermediate l-pipecolic acid from biomass-derived lysine via artificially engineered microbes. *J. Ind. Microbiol. Biotechnol.* **45**, 405–415 (2018).

54. Schulz, M. *et al.* Pantoea ananatis converts MBOA to 6-methoxy-4-nitrobenzoxazolin-2(3H)-one (NMBOA) for cooperative degradation with its native root colonizing microbial

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

consortium. *Nat. Prod. Commun.* **13**, 1275–1278 (2018).

55. Fenz, R., Galensa, R. & Ernst, L. Phenolcarbonsäuren und ihre Glycerinester in Maisgrits. *Zeitschrift für Leb. und Forsch. 1992 1943* **194**, 252–258 (1992).

56. Mahajan, N. *et al.* High fructose induced adipogenesis and inhibitory potential of trigonelline on murine mesenchymal stem cells: A morphological study. *Int. J. Pharm. Sci. Res.* **10**, 528–536 (2019).

57. Srinivasulu, C., Ramgopal, M., Ramanjaneyulu, G., Anuradha, C. M. & Suresh Kumar, C. Syringic acid (SA) – A Review of Its Occurrence, Biosynthesis, Pharmacological and Industrial Importance. *Biomed. Pharmacother.* **108**, 547–557 (2018).

58. Moreau, R. A., Nuñez, A. & Singh, V. Diferuloylputrescine and p-coumaroyl-feruloylputrescine, abundant polyamine conjugates in lipid extracts of maize kernels. *Lipids* **36**, 839–844 (2001).

59. Seaver, S. M. D. *et al.* The ModelSEED Biochemistry Database for the integration of metabolic annotations and the reconstruction, comparison and analysis of metabolic models for plants, fungi and microbes. *Nucleic Acids Res.* **49**, D575–D588 (2021).

60. Wang, L., Xu, C., Qu, M. & Zhang, J. Kernel amino acid composition and protein content of introgression lines from Zea mays ssp. mexicana into cultivated maize. doi:10.1016/j.jcs.2007.09.014

61. Li, K. *et al.* Large-scale metabolite quantitative trait locus analysis provides new insights for high-quality maize improvement. *Plant J.* **99**, 216–230 (2019).

62. Opialla, T., Kempa, S. & Pietzke, M. Towards a More Reliable Identification of Isomeric Metabolites Using Pattern Guided Retention Validation. *Metabolites* **10**, 1–16 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

63. Claes, B. S. R., Takeo, E., Fukusaki, E., Shimma, S. & Heeren, R. M. A. Imaging Isomers on a Biological Surface: A Review. *Mass Spectrom.* **8**, (2019).

64. Kranenburg, R. F. *et al.* Mass-Spectrometry-Based Identification of Synthetic Drug Isomers Using Infrared Ion Spectroscopy. *Anal. Chem.* **92**, 7282–7288 (2020).

65. Van Der Hooft, J. J. J. *et al.* Linking genomics and metabolomics to chart specialized metabolic diversity. *Chem. Soc. Rev.* **49**, 3297–3314 (2020).

66. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).

67. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

68. Landrum, G. Rdkit: Open-source cheminformatics software. *URL http://www. rdkit. org/, https://github. com/rdkit/rdkit* (2016). Available at: https://www.rdkit.org/. (Accessed: 30th December 2020)

69. Patten, J. J. *et al.* Multidose evaluation of 6,710 drug repurposing library identifies potent SARS-CoV-2 infection inhibitors In Vitro and In Vivo. *bioRxiv Prepr. Serv. Biol.* 2021.04.20.440626 (2021). doi:10.1101/2021.04.20.440626

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

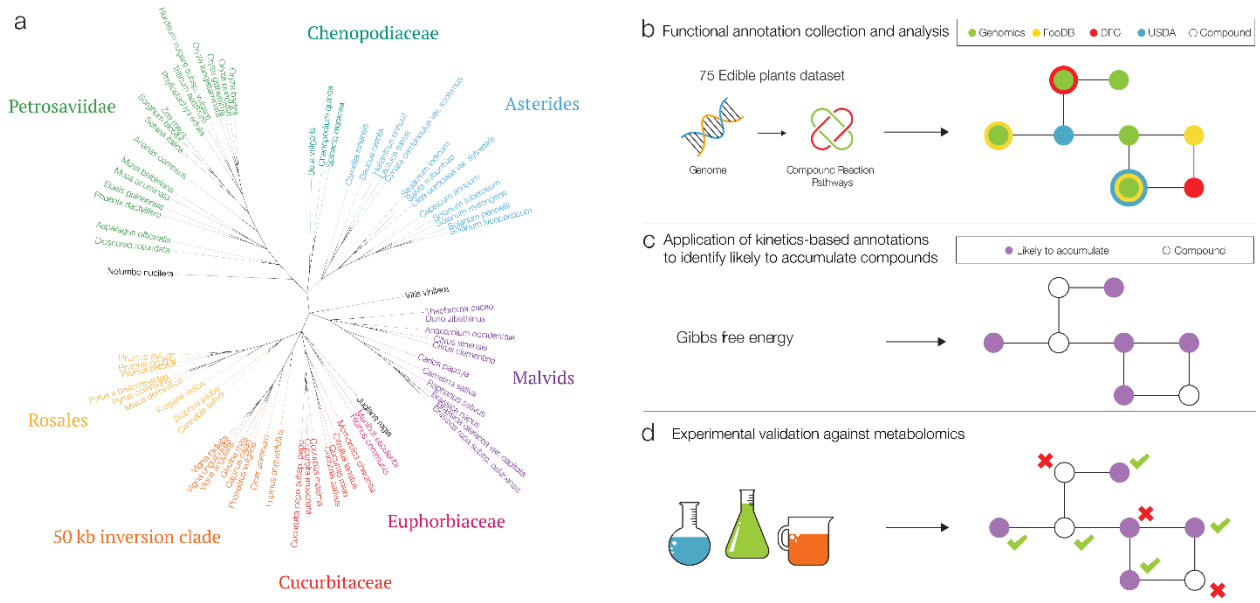

Figure 1 – Plant phylogenetic diversity and schematic overview of genomics contribution to food composition knowledge. (a) Phylogenetic tree representing the 75 plants in our collection colored by plant order. **(b)** Collection and evaluation of genomics-based annotations to food composition knowledge.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

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

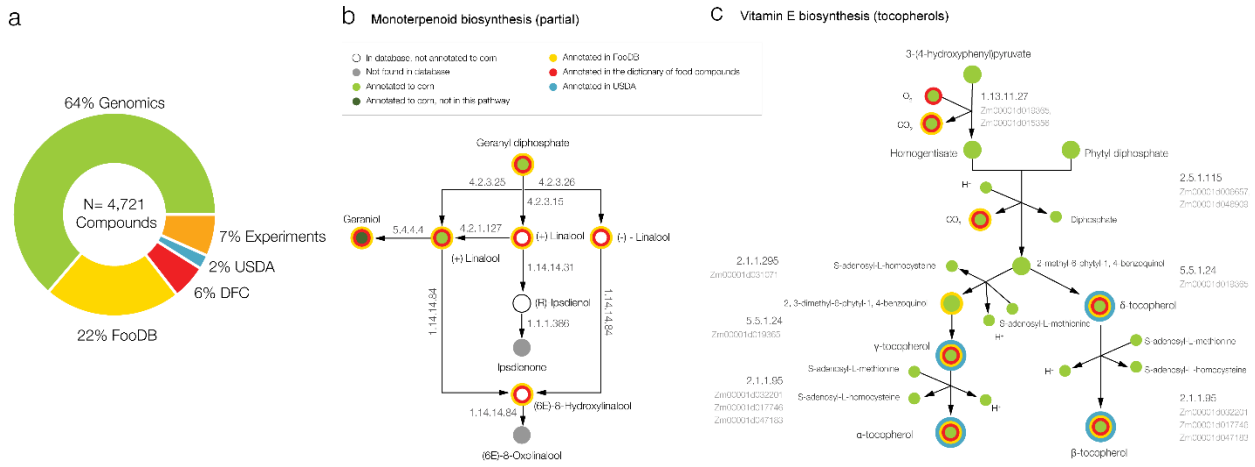

Figure 2 – Genomics-based annotations boost corn composition knowledge. **(a)** Database annotations for corn, indicating that some compounds are annotated in multiple databases. The total number (N=4,721) represents the number of unique compounds for corn after the addition of genomics-based annotations.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

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

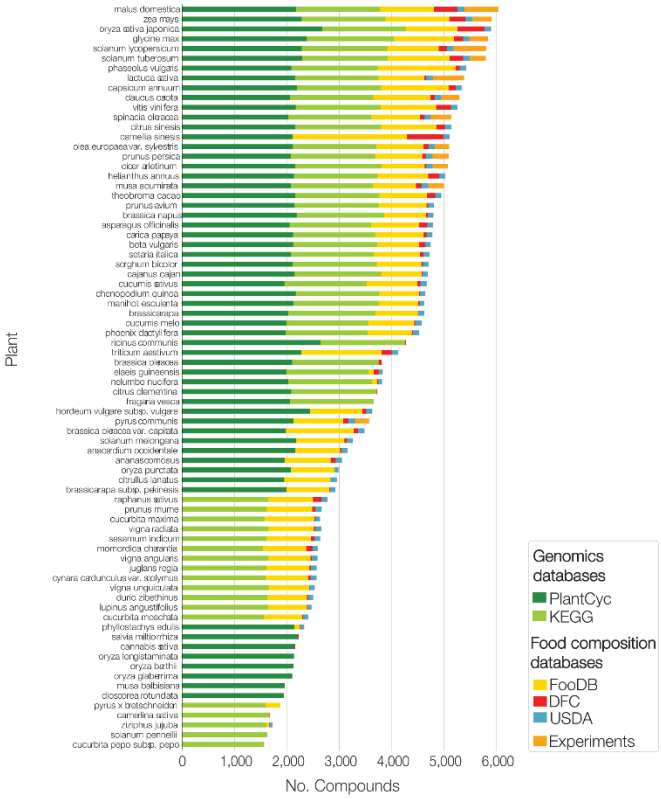

Figure 3 – Contribution of the different sources of food composition across the entire edible plant catalogue. Genomics-based annotation are presented in two shades of green and include the KEGG (light green) and PlantCyc (dark green). Food composition databases include FooDB (yellow), DFC (red), USDA (blue). Compounds detected in metabolomics experiments are shown in orange.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

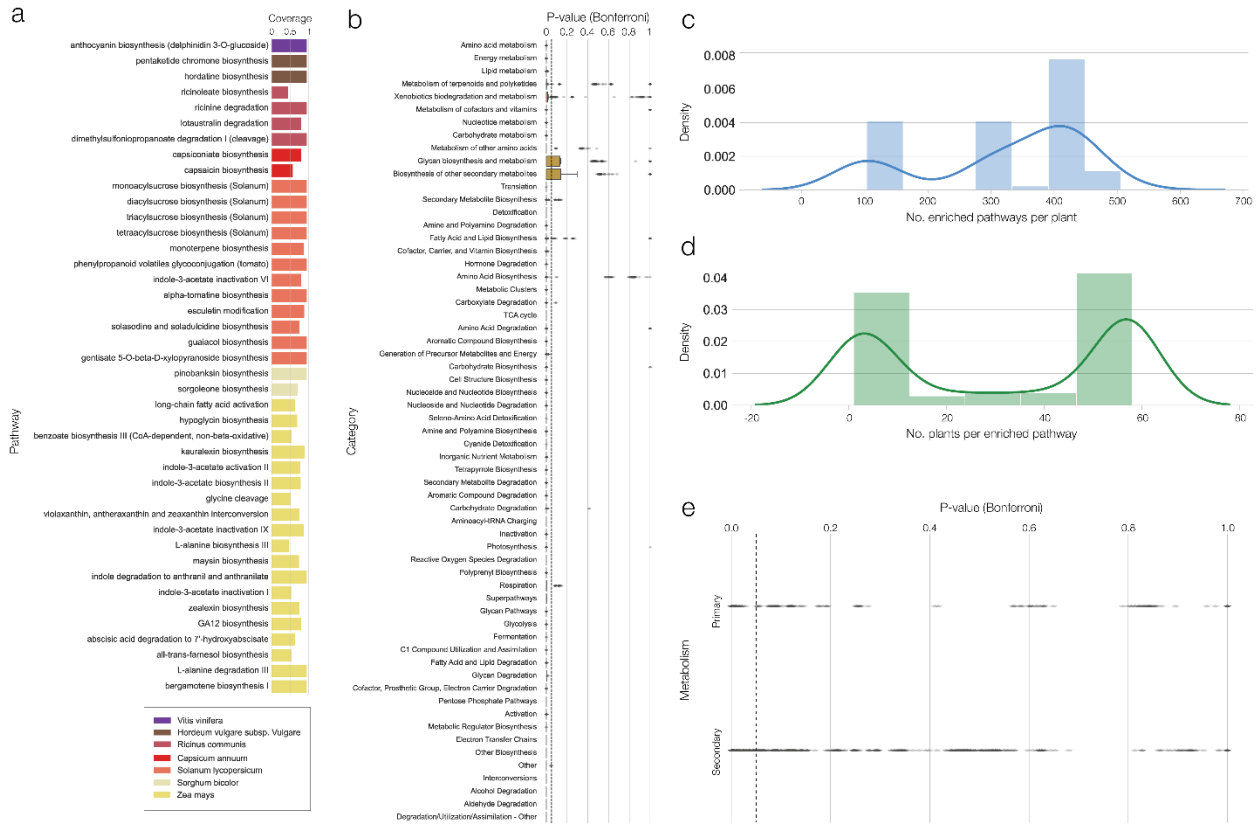

Figure 4 – Pathway enrichment analysis. (a) Plant specific significantly enriched pathways (hypergeometric test, p<0.05, Bonferroni multiple testing correction). The coverage of each pathway is defined by the ratio of present compounds to the total number of compounds in the reference pathway. These 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.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

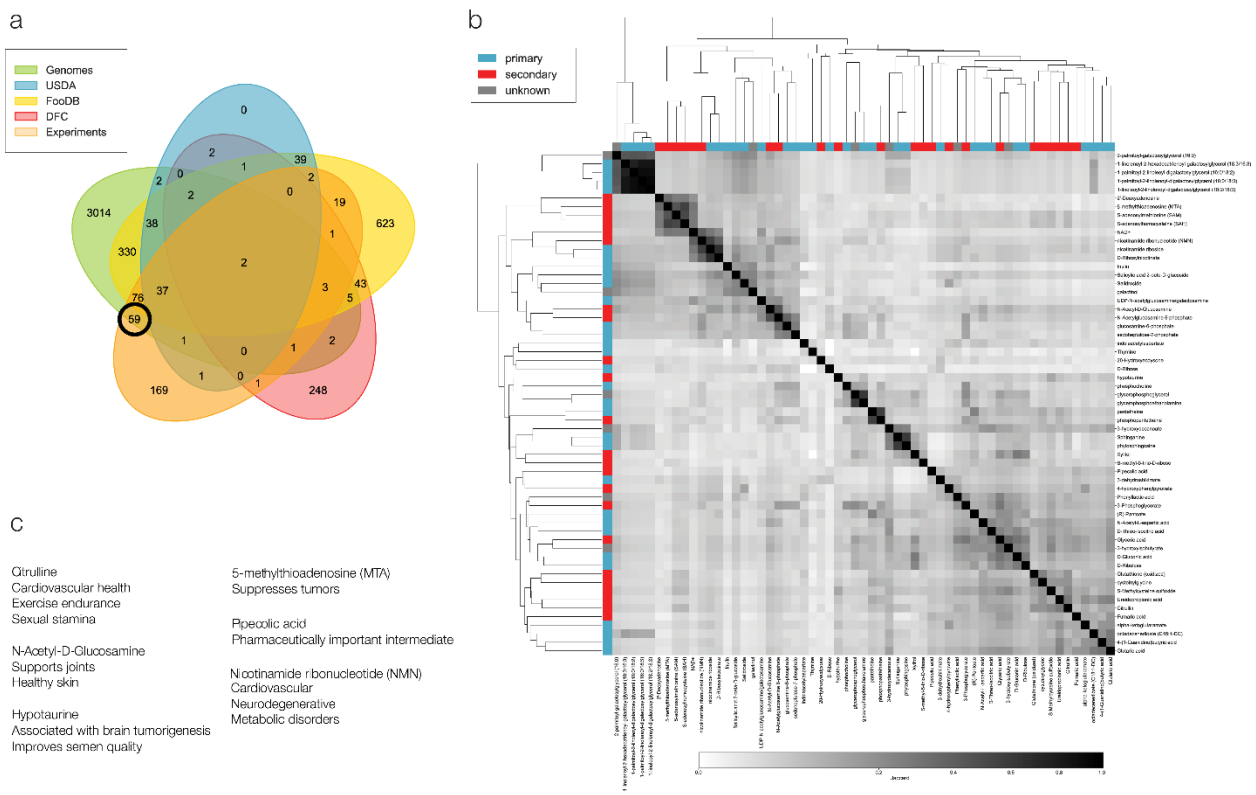

Figure 5 –Bioactive compounds in corn unveiled by genomics. (a) Venn diagram comparing all sources of data contributing to the corn compound collection. The black circle marks the fraction of compounds unique to genomics-based annotations and metabolomics experiments reported here. **(b)** A structural 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 wellbeing as learned with the addition of genome based-annotations, associated with antioxidant and anti-inflammatory activity promoting, heart, skin and metabolic health.

bioRxiv preprint doi: https://doi.org/10.1101/2022.01.24.477528; this version posted January 25, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

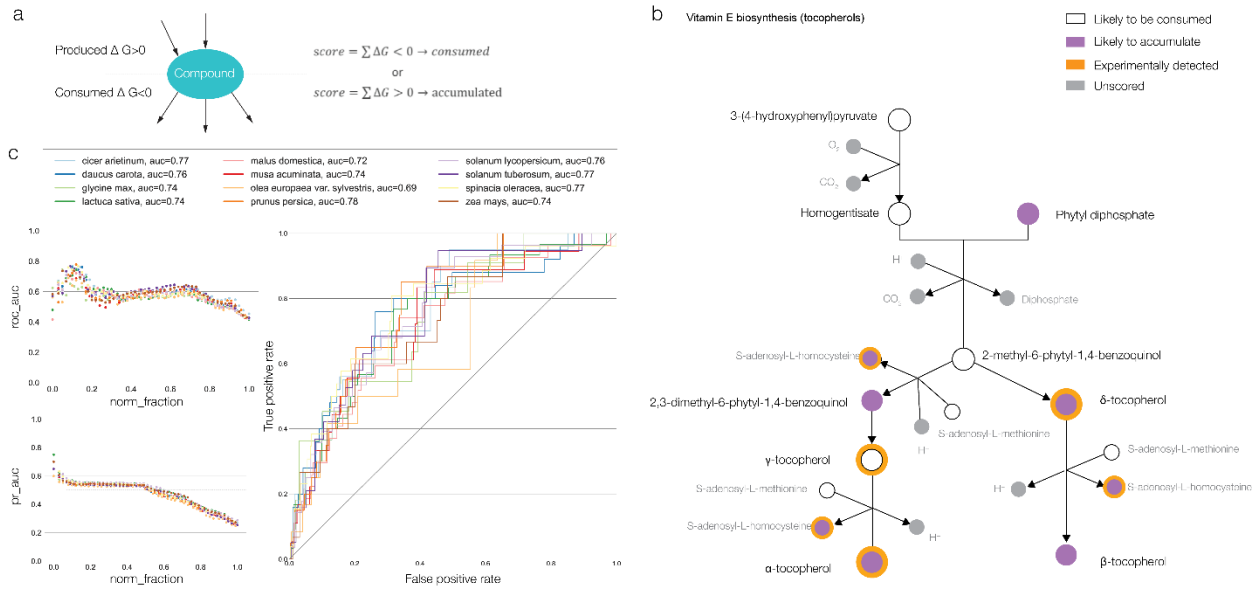

Figure 6 – Application of kinetics-based annotations to predict likely to accumulate compounds. **(a)** Schematic description of the kinetics-based annotations approach. Gibbs free energy values, ΔG, were collected for each reaction in each plant and used to calculate the cumulative score of a compound. 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.