A comparative approach for selecting orthologous candidate genes underlying signal in genome-wide association studies across multiple species

Lauren Whitt[1,#a], Elizabeth H. Mahood[1], Greg Ziegler[1], Collin Luebbert[1], Jason D. Gillman[2], Gareth J. Norton[3], Adam H. Price[3], David E. Salt[4], Brian P. Dilkes[5], Ivan Baxter[1*]

[1]Donald Danforth Plant Science Center, St. Louis, Missouri, United States of America

[2]USDA-ARS Plant Genetics Research Unit, Columbia, Missouri, United State of America

[3] School of Biological Sciences, University of Aberdeen, Aberdeen, United Kingdom

[4]Future Food Beacon of Excellence and School of Biosciences, University of Nottingham, Nottingham, United Kingdom

[5]Department of Biochemistry, Purdue University, West Lafayette, Indiana, United States of America

[#a]Plant Science and Technology, University of Missouri, Columbia, Missouri, United States of America

* Corresponding author

Email: ibaxter@danforthcenter.org (IB)

Abstract

Advances in quantitative genetics have enabled researchers to identify genomic regions associated with changes in phenotype. However, genomic regions can contain hundreds to thousands of genes, and progressing from genomic regions to candidate genes is still challenging. In genome-wide association studies (GWAS) measuring elemental accumulation (ionomic) traits, a mere 5% of loci are associated with a known ionomic gene - indicating that many causal genes are still unknown. To select candidates for the remaining 95% of loci, we developed a method to identify conserved genes underlying GWAS loci in multiple species. For 19 ionomic traits, we identified 14,336 candidates across Arabidopsis, soybean, rice, maize, and sorghum. We calculated the likelihood of candidates with random permutations of the data and determined that most of the top 10% of candidates were orthologous genes linked to GWAS loci across all five species. The candidate list also includes orthologous genes with previously established ionomic functions in Arabidopsis and rice. Our methods highlight the conserved nature of ionomic genetic regulators and enable the identification of previously unknown ionomic genes.

Author summary

Identifying the genes contributing to changes in a given trait is challenging. Many genes can be near the region of interest, but proximity does not always translate to causality. We use other methods to narrow our focus to the genes most likely involved in our trait of interest before confirming their involvement through confirmation experiments. However, these other methods are often time, labor, and resource-consuming. We developed an approach to narrow these gene lists before these laborious methods are required. Through testing with genetic markers for elemental (i.e., calcium, iron, zinc) uptake, we found that comparing markers across multiple species for nearby evolutionarily conserved genes is a successful approach. We've produced a list of candidate genes likely to be involved in elemental uptake traits, including previously known elemental uptake genes

and genes whose potential elemental uptake function has yet to be observed. Some of these genes would not have been considered according to the significance threshold within a single GWAS, but combining comparable datasets across species has collectively boosted their signal. Methods like our approach are useful for reducing candidate lists to conserve resources spent in functional characterization experiments and encouraging the discovery of new functional roles of these genes.

Introduction

High-throughput phenotyping (HTP) paired with genome-wide association studies (GWAS) has identified many genetic loci associated with variation in the accumulation of elements in many plant species [1–5]. However, identifying the causal gene associated with a group of linked variants, often called a GWAS peak, is challenging for any trait or organism. Typically, many genes are linked to a GWAS peak, especially in species with large linkage disequilibrium ranges, and each is a candidate for the allele affecting variation in the phenotype. Complex loci, multiple alleles, rare alleles, and incomplete variant information (due to technology or sparse marker data) can also result in the most statistically significant variant being encoded at a position displaced from the causal gene [6]. As a result, additional and complementary information about allele consequences and gene function is necessary to identify the molecular cause of detected quantitative genetic variation. Bridging these trait-to-gene gaps is necessary to deliver molecular mechanistic insights into natural variation.

Current gene selection methods range from quick functional annotation checks to more extensive experimental confirmations. Gene ontologies (GO) and annotations provide knowledge of our current understanding of known (or predicted) genes but are prone to human biases [7,8] and oversights from automatic computational algorithms [9]. Excluding a supremely significant GWAS peak where the single nucleotide polymorphism (SNP) resides within a coding region, known genes are more likely to be selected for further studies than unknown genes [7,10]. Additionally, these methods are most effective in species with the most annotations, like *Arabidopsis thaliana*

(Arabidopsis), with well-annotated phenotypes. GO terms are largely uninformative for novel or under-described traits, like some elemental accumulation phenotypes. GO terms require ongoing curation and submission from researchers to maintain relevance and utility.

More in-depth approaches look for candidates with additional signals that indicate their candidacy, regardless of a gene's existing annotations or ontologies. For example, linkage mapping in experimental populations has been used to differentiate causal candidates in trait-associated GWAS peaks [11,12]. However, fine mapping to find a causal gene requires expensive and laborious methods (for example, high-resolution sequencing and cloning) [13,14] and will not work for alleles with moderate effect sizes. Candidate gene testing by working with knockout and overexpression mutants is an alternative approach to differentiate candidates from causal loci. Despite advances in gene editing, these methods are much easier in species like Arabidopsis and *Zea mays* (maize) with substantial mutant resources. Model systems like Arabidopsis also benefit from short generation times and small growth requirements [15]. Developments in whole genome and next-generation sequencing have improved *de novo* genome assembly of organisms with complex genomes and identification of genetic markers (i.e., SNPs, indels) within populations [16]. Now, most agriculturally important crop species have genome assembly and annotations. Still, Arabidopsis has the highest quality plant genome, annotations, and resources that inform annotations in other organisms. Indeed, for several agricultural species (e.g., soybean), gene annotations are inferred from homology searches against Arabidopsis. Several approaches have been used to identify causal genes from genes of unknown function. Guided methods that can work without gene annotations include machine learning techniques like PILGRM [17], GenAMap [18], network analysis, gene expression experiments, or a combination of techniques [19]. These methods rely on extensive computation and resources like gene expression data. As a result, they are most valuable in model systems with large collections of preexisting data.

The ionome describes the composition of elements in plant tissue acquired from its environment [20]. The challenge of obtaining, excluding, and homeostatically maintaining the concentration of elements in cells is a key function shared by organisms across all domains of life. Elements obtained from the environment are critical cofactors and substrates for multiple biochemical reactions, and solving these challenges was a necessary step in the evolution of cellular life. As a result, many of the mechanisms that achieve elemental homeostasis are ancient, and experimental approaches for mechanistic validation, such as cross-kingdom molecular complementation (e.g., yeast complementation, bacterial complementation), are highly effective. Moreover, as elements are the same regardless of which species they are measured in, we can directly compare elemental accumulation traits across species. As a result, the field has repeatedly demonstrated that homologous genes affect the levels of the same elements in multiple plants [21–30]. The independent evolution of alleles with similar phenotypic consequences in multiple species is the hallmark of conserved critical regulators. The direct comparability of ionomics traits and the deep conservation of mechanisms affecting its variation raises the possibility that we could identify homologous genes affecting variation in ionomic traits linked to GWAS peaks across multiple species.

We created a computational method to take the genes linked to GWAS data sets from multiple species and detect orthologous genes when present among the linked candidates. With this, we built a permutation-based assessment of the effectiveness of this tool that calculates the expected background distribution of orthologs for each collection of GWAS results. Together, this demonstrated that GWAS detects alleles at orthologous genes more often than expected by chance, including genes of previously unknown function in regulating the ionome. We further validated these results using lists of genes of known ionomic function [24] and demonstrated that these are also detected by our combination of GWAS and orthology at a higher rate than expected by chance. One key advantage of our approach is that it is powered by phylogenetic analysis and, therefore,

132 detects high-quality candidate genes for traits without requiring prior demonstration of function.
133 This approach utilizes existing datasets and encourages exploration into novel functional
134 annotations.

135 Results

136 Known Ionomic Genes as GWAS candidates

137 Elemental accumulation, or ionomic, datasets from association mapping populations of five species
138 (Table 1): Arabidopsis [31], *Glycine max* (soybean; see methods), *Oryza sativa* (rice) *[32–35]*,
139 *Sorghum bicolor* (sorghum) [1], and maize [3] were reanalyzed to identify alleles at conserved
140 orthologous genes. These GWAS were selected because they are publicly available and contain 19
141 independent traits measured with the same type of instrument (ICP-MS). In all cases, element
142 accumulation was measured in the seed, and in all cases except Arabidopsis, data sets were
143 generated from plants grown in multiple environments. SNPs were filtered by p-values specific to
144 each dataset (see methods) and, when applicable, the SNP's significance across multiple
145 environments (Table 1). Loci from the maize NAM dataset from Ziegler et al. 2017 were filtered
146 more stringently than the other GWAS datasets due to the large number of SNPs. General linkage
147 ranges for each dataset were used to determine loci and linked genes.
148

Table 1. Ionomic loci from GWAS across five species.

	Arabidopsis	soybean	rice	sorghum	maize
Environments	single	multiple	multiple	multiple	multiple
boron (B)	0	401	0	63	10
sodium (Na)	118	331	40	53	0
magnesium (Mg)	65	355	10	163	54
phosphorus (P)	58	372	14	87	39
sulfur (S)	57	362	13	78	54

potassium (K)	49	380	12	93	53
calcium (Ca)	54	360	16	102	43
iron (Fe)	55	354	20	210	67
manganese (Mn)	59	380	15	175	72
cobalt (Co)	149	373	0	106	105
nickel (Ni)	0	374	0	73	56
copper (Cu)	53	353	26	82	72
zinc (Zn)	51	384	24	103	58
arsenic (As)	68	360	20	0	28
selenium (Se)	126	359	0	114	7
rubidium (Rb)	72	352	0	53	59
strontium (Sr)	60	387	0	0	37
molybdenum (Mo)	109	362	70	115	73
cadmium (Cd)	114	381	152	98	11
Total loci	1404	7308	432	1768	905

To help identify candidate genes underlying elemental uptake loci in plants, we had previously compiled a list of known ionomic genes (KIG) experimentally shown to alter elemental uptake in plants [24]. As these genes were sourced from the primary literature, we called this our “primary” KIG list. Since the original publication, we have added 19 new primary genes to the list (S1 Table), for a total of 201 primary genes. Most primary genes are in Arabidopsis (59%), with smaller contributions from rice (33%) and a small number in maize, *Medicago truncatula*, and *Triticum aestivum* (wheat). The overlap between the identified GWAS loci and the KIG list was used to determine if these GWAS datasets identified alleles at genes already known to affect changes in the ionome (Fig 1). Loci and gene overlaps were considered on a trait-by-trait basis, i.e., Fe uptake GWAS loci were only compared to KIG genes listed for Fe. Permutations of our GWAS loci were created by selecting an equal number of random base pair locations across the genome and assigning a “locus range” randomly selected from the actual loci. The primary KIG list covered a low percentage of

GWAS loci (<2%), but overlapped with ionomic loci more often than the random locations (80th percentile in Arabidopsis, 95th in rice, >99th in maize).

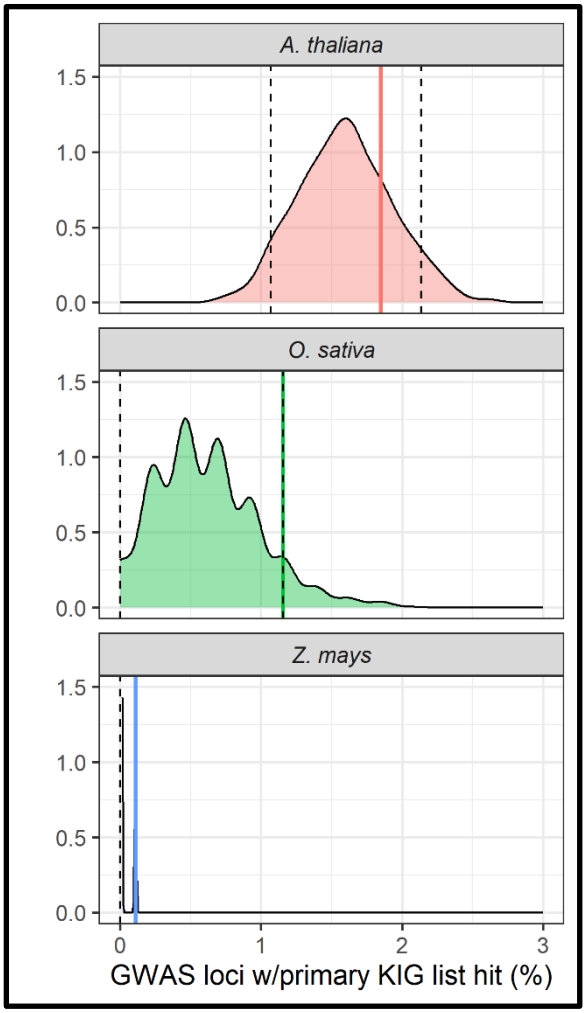


Fig 1. Primary KIG list coverage of ionomic GWAS loci. Solid, vertical lines represent the percentage of GWAS loci explained by a gene from the primary KIG list. Kernel Density Estimate (KDE) plots display the distribution of 1000 random permutations with the 5th/95th percentiles shown as dashed lines. Permutations were made with random genome selections of the same size and species composition as the actual GWAS datasets (S3 Table).

Can orthology improve the KIG list?

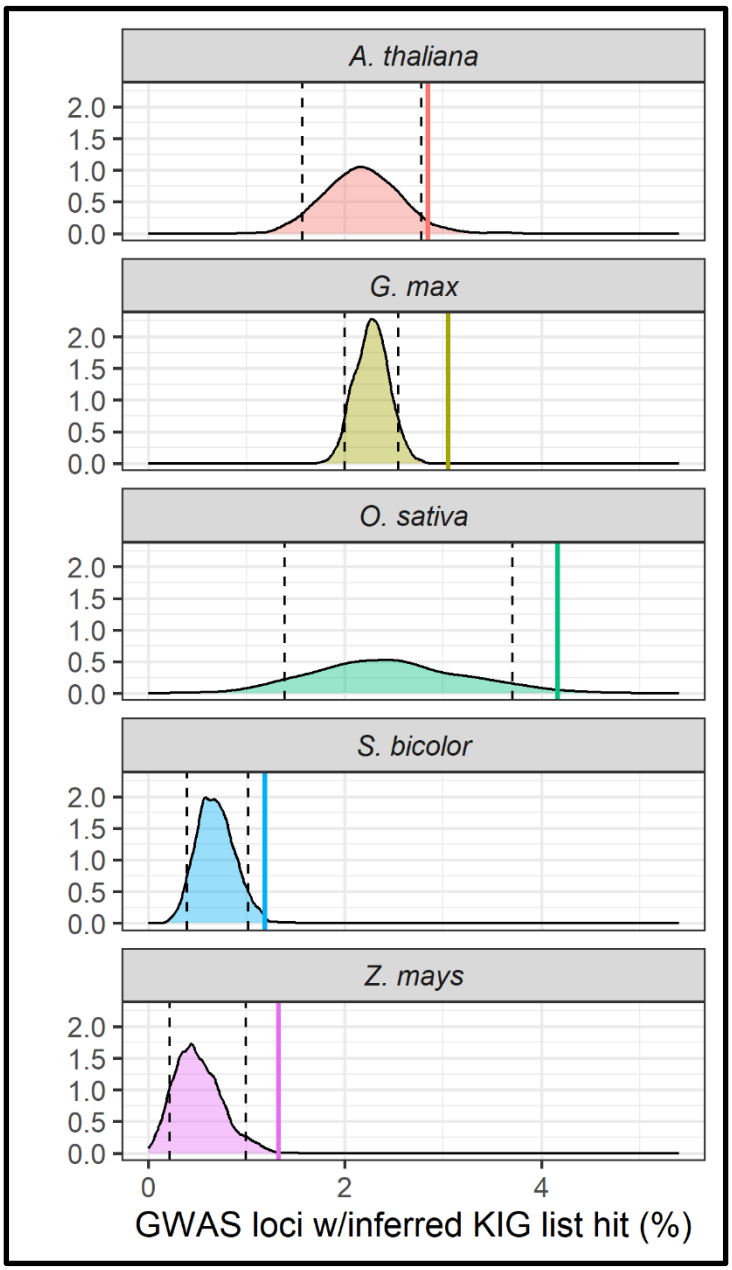
While KIGs overlapped with ionomic loci more often than random genome locations, a large portion of GWAS loci are left unexplained by the KIG list. We hypothesized that inferring orthologs across

species would increase the overlap between GWAS loci and expand the list of KIGs and orthologs. Orthologs are genes across species inherited from a common ancestor. Although now in separate species, orthologs tend to retain similar functionality [36], making them good candidate genes when their function is known in one species. Indeed, for many KIGs, orthologs have been tested in multiple species and found to encode conserved functions affecting element accumulation for S, Fe, Cd, Na, K, Rb, Mo, Ni, Cu, P, Zn, As, B, Mn, and aluminum (Al) (Whitt et al., 2020). For example, alleles at orthologous genes encoding the *MOT1* transporter affect variation in the level of Mo in multiple species [26,33,37,38] and the same has been observed at *HKT1* orthologs for Na [27,39,40]. Inferring the orthologs of KIGs across all of our study species should allow us to discover alleles in genes affecting the ionome, including genes that have not yet been tested in every species in our list.

There are various ortholog inference algorithms that vary by their use of pairwise similarity scores, phylogenetic analysis, and multiple sequence alignment - all of which produce variations in the ortholog groups. The original KIG list [24] was made with InParanoid [41] because the ortholog tables were readily available on Phytozome [42]. For this analysis, we switched to OrthoFinder v2.0 [43] because it performed better than other ortholog group inference methods in benchmarking tests [43,44], and it can form ortholog groups between 3 or more species in one comparison. With the ortholog tables from OrthoFinder, we produce “inferred” KIG lists by identifying orthologs of the primary KIGs in the original species and any species with a sequenced genome, including the five we are studying here. Expanding the list of taxa in the inferred KIG list created soybean and sorghum gene lists so we could repeat the KIG/GWAS overlap comparison in all five of our ionomic GWAS species.

The inferred KIG lists improved loci coverage across all five species and resulted in significant overlap between the observed GWAS loci and the KIG ortholog lists (>95th percentile) for all species (Fig 2). As expected, this is true whether we randomized the loci (Figs 1 and 2) or randomized the genes in

199 the list (Supplemental Fig 5). While the likelihood that we observed an enrichment of true positives
200 was significant (Fig 2), per-gene assessment of false discovery rates (FDR) was relatively high (Table
201 2), partly due to the large number of genes linked to GWAS peaks. Thus, the overlaps between KIG
202 and GWAS loci represent high-value candidates for alleles affecting ionomic variation at conserved
203 genes (Table 2; S6 Table).



204

205 **Fig 2. Inferred KIG list coverage of ionomic GWAS loci.** Solid, vertical lines represent the percentage
206 of GWAS loci explained by a gene from OrthoFinder inferred KIG lists. Kernel Density Estimate (KDE)
207 plots display the distribution of 1000 random permutations with the 5th/95th percentiles shown as

208 dashed lines. Permutations were made with random genome selections of the same size and species
209 composition as the actual GWAS datasets (Table 2).
210

Table 2. Summary of inferred KIG list overlaps with GWAS loci.

				KIG/GWAS overlaps		KIG/random overlaps		False discovery rate (FDR)	
Species	GWAS loci	random locations	Inferred genes	genes	loci	genes	loci	genes	loci
Arabidopsis	1404	1404	163	45	40	32.86	30.55	0.73	0.76
soybean	7308	7308	378	264	223	170.03	166.23	0.64	0.75
rice	432	432	181	22	18	11.57	10.63	0.53	0.59
sorghum	1768	1768	201	23	21	13.36	12.13	0.58	0.58
maize	905	905	222	13	12	4.89	4.59	0.38	0.38

211
212 Among the genes that overlap between the inferred KIG orthologs and GWAS loci are conserved
213 genes known to regulate elemental accumulation across these plant species (Table 3; S4 Table).
214 Identification of alleles linked to orthologous candidate genes via GWAS for the same element
215 provides evidence that these genes not only encode conserved determinants of elemental
216 homeostasis but also indicate that phenotypically relevant, standing variation exists in these genes
217 that could be used to enhance plant mineral nutrition (e.g., P, Fe, Zn, Mn, Mo), improved mineral
218 nutrition for humans (e.g., Fe and Zn), or increased food safety (e.g., Cd).
219

Table 3. KIG ortholog groups overlapping GWAS loci for their cited traits. KIG citations are listed in S1 Table.

KIG family	Cited traits	GWAS trait	Orthologs
AtZIF1, AtZIFL2 *OsZIFL12* *ZmYS3*	Cs Zn Fe	Fe	*AtZIFL2* *Glyma.06G313600, Glyma.06G313700* *Sobic.008G015700*
AtPHT1;1, AtPHT1;4, AtPHT1;5 *OsPT2*	P As Se	P	*AtPHT1;4* *Glyma.10G186400, Glyma.10G186500* *LOC_Os08g45000*
AtPEN3	Cd	Cd	*AT3G16340* *Glyma.07G015800*

			LOC_Os01g52560, LOC_Os06g36090
AtNRAMP3	Fe Mn Zn	Mn	*AtNRAMP3* *Glyma.04G044000* *Sobic.001G170000* *Zm00001d030846*
AtMTP11	Mn	Mn	*AtMTP11* *Glyma.18G286100* *Sobic.003G349200* *Zm00001d042939*
AtMOT2	Mo	Mo	*AtMOT2* *Glyma.08G208700* *LOC_Os01g45830* *Sobic.003G237000* *Zm00001d044068*
AtMOT1 *MtMOT1.2, MtMOT1.3* *OsMOT1;1*	Mo	Mo	*AtMOT1* *Glyma.06G074100, Glyma.17G203000, Glyma.17G203200, Glyma.17G203500* *Zm00001d033053*
AtBTS	Fe Mn Zn	Fe	*Glyma.05G237500, Glyma.08G044700* *Sobic.009G222200* *Zm00001d038916*
AtESB1	Ca, Mn, Zn, Na, S, K, As, Se, Mo	Mn	*AT1G07730, AT2G39430* *Glyma.08G258300, Glyma.12G067400, Glyma.18G282600* *Sobic.003G064800*
AtHMA2, AtHMA3, AtHMA4 *OsHMA3*	Zn Cd	Cd	*AtHMA4, AtHMA2, AtHMA3* *Glyma.09G055600* *OsHMA3* *Sobic.002G083000, Sobic.002G083100* *Zm00001d005189, Zm00001d005190*

Among the orthologs of KIGs linked to GWAS peaks for the same element are several transporters with known roles in element uptake, sequestration, and remobilization. Notably, alleles were discovered in orthologs of the *ZIFL1* and *ZIFL2* genes that affect the accumulation of Fe in seeds (Table 3), a human nutrient that is limited in many diets [45–47]—making these alleles high-priority targets in breeding materials for improved nutrient accumulation in food. KIGs *MOT1* and *MOT2* encode molybdate transport proteins [26,48–51] and overlap Mo loci in multiple species. Alleles affecting the accumulation of the toxic heavy metal Cd were discovered in all five species at the heavy metal ATPases *AtHMA2*, *AtHMA3*, *AtHMA4*, and *OsHMA3*, which belong to a subset of *HMAs* specific to Zn/Cd transport [11,52–54]. In addition to direct Cd transport, alleles at *PEN3/PDR8*,

implicated in Cd transport via chelation with the glutathione-derived phytochelatin [55], affected Cd in three species. The alleles in these transporters represent additional breeding targets for improved human food safety. The phosphate transporter (*PHT1*) family has been shown to affect phosphate and As accumulation in Arabidopsis [56–58], Se uptake in rice [59] and was linked to P loci in Arabidopsis, soybean, and rice. Thus, these alleles represent possible targets to enhance nutrient uptake for the crop-limiting nutrient P. *MTP11* encodes a Mn transporter involved in Mn tolerance and detoxification with the Golgi apparatus [60], and orthologs to *AtMTP11* are linked to Mn loci in four species. Alleles in the orthologs of yet another transporter family, including *AtNRAMP3*, which is involved in Mn homeostasis via the mesophyll vacuoles [61], were associated with Mn accumulation in Arabidopsis, soybean, sorghum, and maize.

Alleles at regulatory genes from conserved ortholog groups were also identified among multiple species. These included alleles at orthologs of the transcription factor BRUTUS (*BTS*), which affects the expression of genes affecting Fe accumulation in Arabidopsis, including the transporter *IRT1* and the regulatory transcription factor *FIT1* [29,62,63]. Alleles linked to the *BTS* orthologs affected Fe accumulation in soybean, sorghum, and maize, demonstrating that this approach can identify regulators of transporters and the proteins themselves. Thus, these alleles and genes represent high-value breeding targets for improving seed Fe content in these three crops. The enhanced suberin 1 (*ESB1*) gene involved in Casparian strip biosynthesis [64] was previously known to affect Ca, Mn, Zn, Na, S, K, As, Se, and Mo levels in Arabidopsis [65], but GWAS detected alleles at this gene affecting Mn in Arabidopsis, soybean, and sorghum, suggesting that disruption of the Casparian strip has similar impacts on the ionome across these species with dramatically different root systems.

The utility of combining GWAS across species using orthology and conservation for the known genes was clear. The number of GWAS loci overlapping a KIG ortholog (Fig 2, Table 2) was greater than the overlap with a primary KIG gene (Fig 1, S3 Table). However, this overlap only captured less than 5%

of the ionomic GWAS loci. Prior research bias has resulted in gene discovery focused on a handful of elements due to researcher interests, leaving a majority of elements under-studied [24]. The overlap of less than 5% may indicate that the KIG list is vastly incomplete [8], and more plant ionome genes are waiting to be discovered. The permutation results in Figs 1 and 2 also demonstrate the naive assumption of using GWAS results as independent experiments. Jointly considering the p-values returned for orthologous genes across all five GWAS would increase false positives and result in spurious conclusions. To overcome this and identify high-quality candidates for the remainder of the GWAS results, we sought a formal procedure for assessing the likelihood of GWAS results at orthologous loci across these five species regardless of previously demonstrated functions.

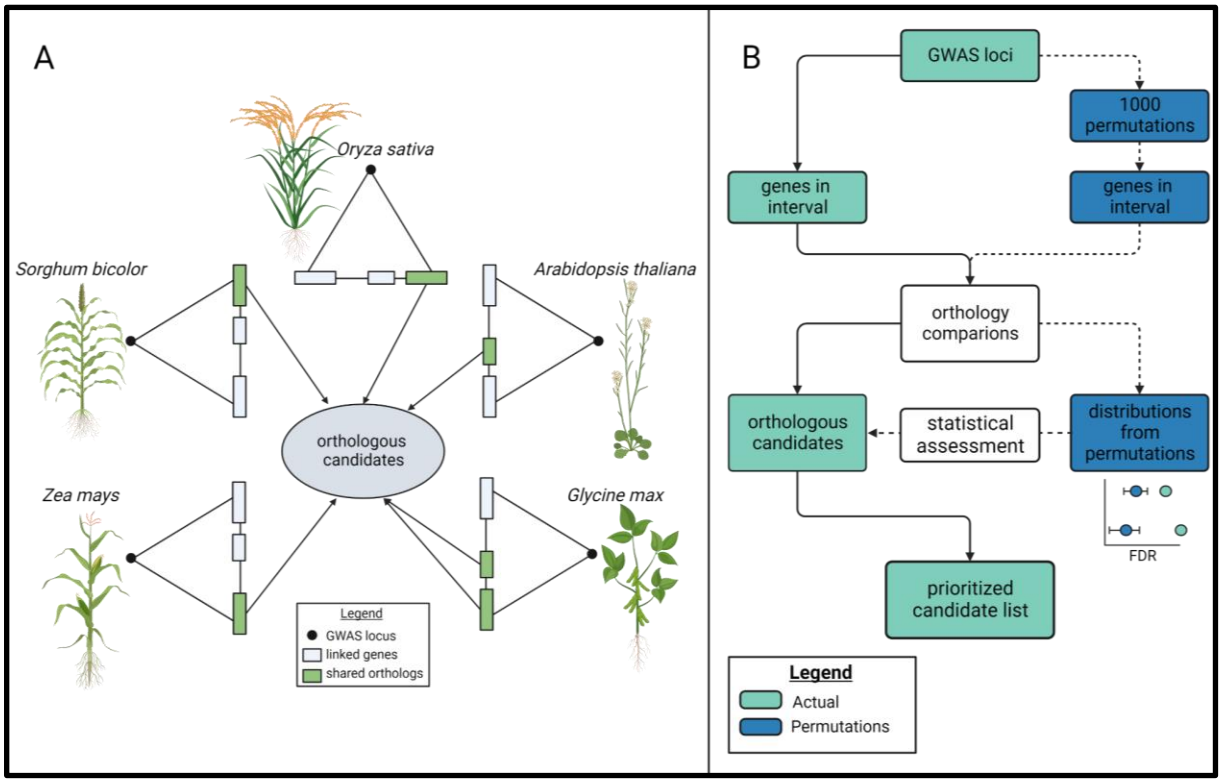
Orthology as a key factor for candidate selection

We hypothesized that conservation of orthologs within GWAS peaks across multiple species could be a criterion to identify genes of previously unknown function. In the expanded KIG ortholog/GWAS overlap comparison (Fig 2, Tables 2 and 3, S4 Table), we frequently detected phenotypically-impactful alleles in multiple species GWAS. GWAS is “forward” genetics and will also detect alleles at genes of previously unknown functions. Therefore, ionomic GWAS datasets will detect alleles at unknown, conserved regulators of elemental accumulation. If alleles at previously unknown conserved regulators of the ionome are present, comparisons between GWAS loci from multiple species will occur at orthologous genes more than expected by chance.

To determine if this was possible and identify conserved ionomic genes, we developed a method we call FiReMAGE: Filtering Results of Multi-species, Analogous, GWAS Experiments. This approach identifies the orthologous genes linked to GWAS associations in multiple species and compares their occurrence to a background distribution to assess the false discovery rate. Additionally, it increases confidence in a candidate gene linked to a GWAS locus without relying on previously known functions. FiReMAGE works by tabulating the genes underlying trait loci in each species, determining

281 their orthologous relationships via OrthoFinder, and retaining ortholog groups with candidate hits
282 from at least three species (Fig 3). The threshold of three or more species is to reduce the false
283 positives due to the likelihood of increased synteny between two closely related species. Like the
284 permutations used with the KIG/GWAS comparisons, FiReMAGE creates 1000 permutations of our
285 actual GWAS data by randomly selecting an equal number of genomic locations and ranges from
286 each species' genome. FiReMAGE runs the random permutations through the same orthology
287 comparison as the actual data and uses the permutation distributions to evaluate the likelihood that
288 loci overlap orthologs. This background distribution is used to evaluate the significance of the
289 overlap and produces a false discovery rate for candidate genes identified this way.

290



291 **Fig 3. Overview of the FiReMAGE pipeline.** (A) FiReMAGE orthology comparison identifies genes
292 linked to SNPs in each GWAS that are also orthologous to each other. (B) FiReMAGE starts by
293 identifying genes within range of GWAS loci intervals. These genes are recorded and filtered through
294 the orthology comparison to identify orthologous candidates. The original GWAS loci are used to
295 generate 1000 random permutations, which go through the same process as the actual data.

Distributions from the permutations are used to assess genes in the real dataset and return a prioritized candidate list. Created with BioRender.com.

With FiReMAGE, we identified candidates in ortholog groups with hits in at least three of the five species for over 30% of soybean and maize loci and over 50% in Arabidopsis, rice, and sorghum (Fig 4). The signal from the actual datasets was above the 95th percentile of the permutations in each species, although the false discovery rate was still relatively high. Across the 19 elemental traits for the 3 out of 5 species comparison, 3866 unique orthologs were linked to loci in Arabidopsis, 7852 in soybean, 2630 in rice, 2982 in sorghum, and 872 in maize (S5 Fig; S7 Table).

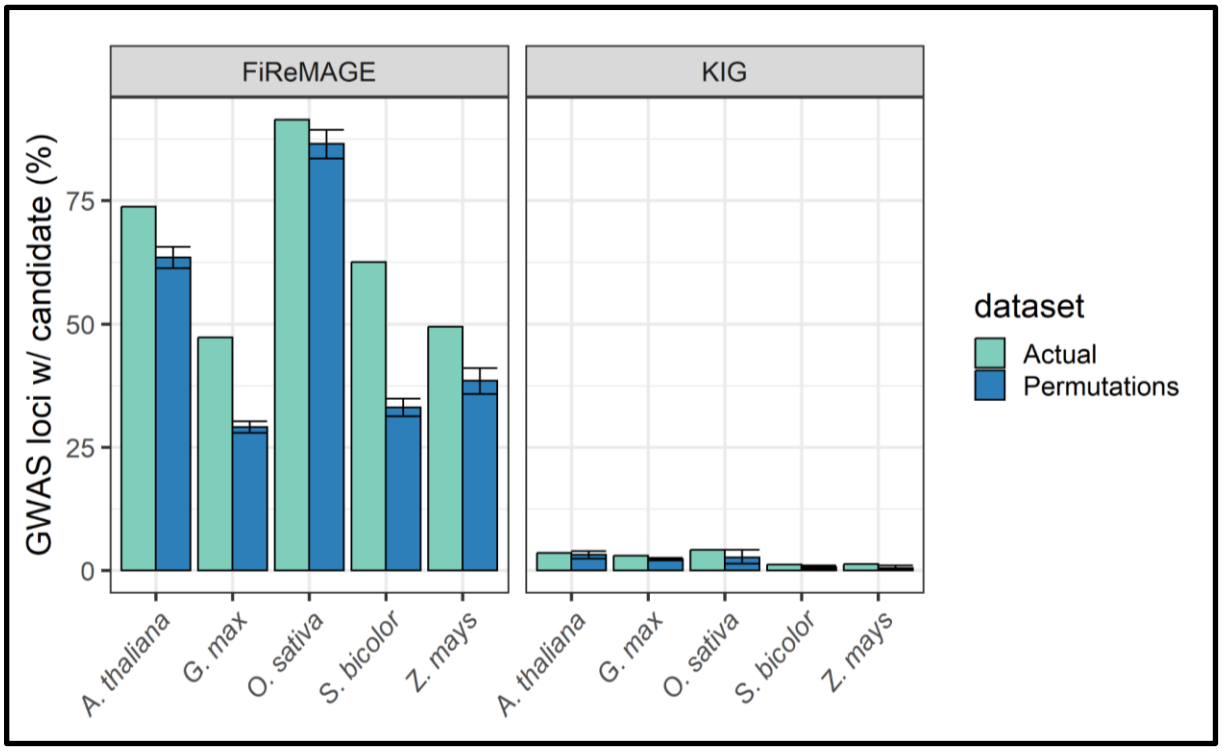


Fig 4. FiReMAGE identifies candidates for a larger poriton of GWAS loci than the KIG list. Percentage of ionomic GWAS intervals with a candidate from FiReMAGE's comparative approach (left) and the OrthoFinder KIG list (right). The actual GWAS dataset is in green; 1000 random permutations of the actual dataset are in blue, with error bars representing 5th/95th percentiles.

The size of candidate lists and false discovery rates vary between species and traits (Figs 5 and 6, S1 and S2 Figs), but overall, requiring more species to be present in an ortholog group returns a shorter candidate list (Fig 5) with lower false discovery rates (Fig 6). While still encompassing a large number of genes, the 3/5 list is a significant reduction in the number of potential candidates.

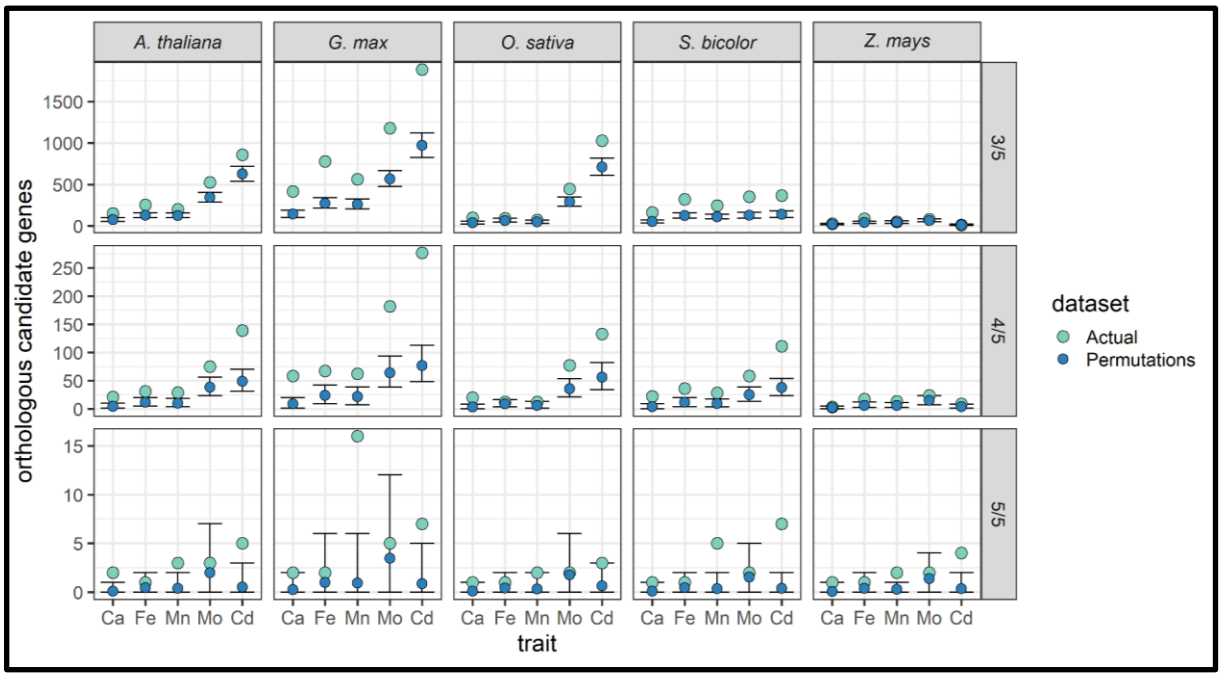


Fig 5. Orthologs from ortholog groups linked to Ca, Cd, Fe, Mn, or Mo loci in 3/5, 4/5, and 5/5 species. Actual dataset counts are in green, and 1000 permutations of the actual data are in blue, with the point representing the mean and error bars representing the 5th/95th percentiles. Columns are by species; rows are the number of species represented in the returned genes' ortholog groups. Distribution plots for ionomic traits found only in 3/5 or 4/5 species ortholog groups are in S1 and S2 Figs.

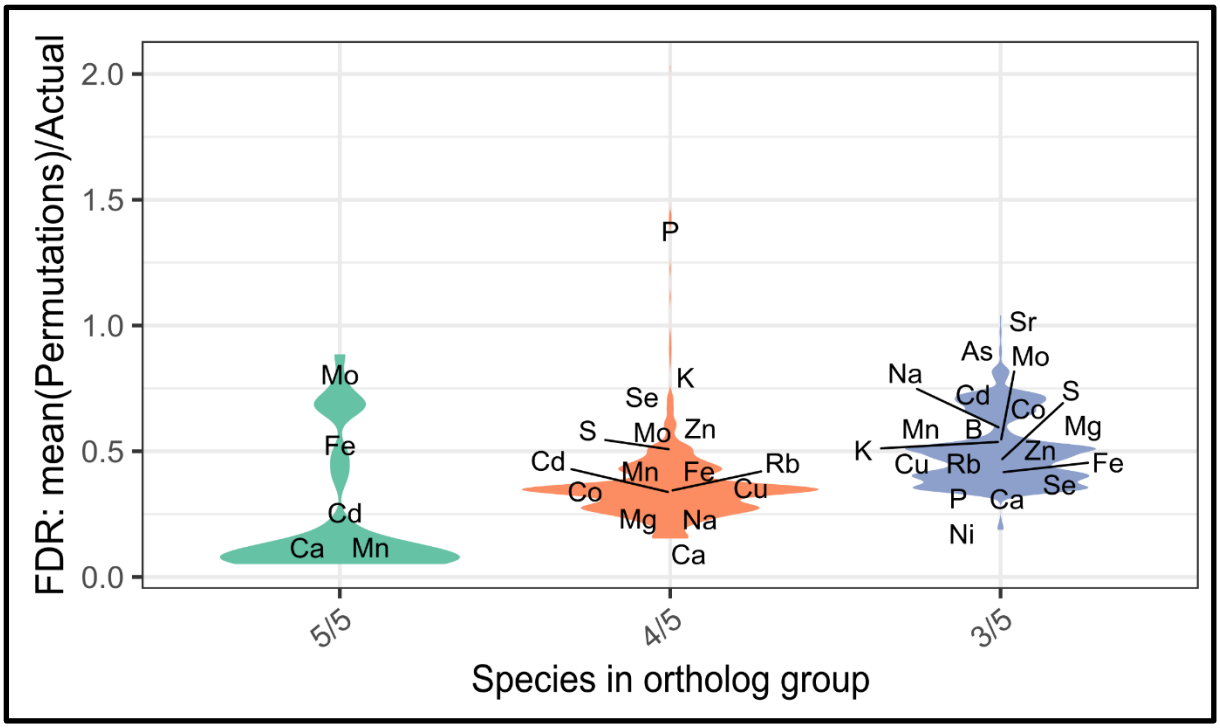


Fig 6. False discovery rate (FDR) of FiReMAGE's selected candidate genes from ortholog groups with representation from 3/5, 4/5, and all five species for each ionomic trait. FDR was calculated by taking the average number of orthologous candidates returned in the 1000 permutation datasets and dividing it by the number of orthologous candidates returned in the actual dataset for each combination of species, ionomic trait, and species representation in the ortholog group.

Including maize and sorghum in our FiReMAGE comparison enabled the observation of how two closely related species impacted the results. We expected the increased synteny between maize and sorghum to increase the number of orthologous genes found, especially in the 3/5 ortholog groups where both maize and sorghum are used (highlighted red in Fig 7). However, these ortholog groups from species with increased synteny do not make up the majority of the results. The largest highlighted intersection was between maize, sorghum, and soybean, with 245 ortholog groups. However, the FDR of these genes is comparable to 3/5 ortholog groups with less syntenic species compositions, suggesting that this is not a major concern. Including a perfectly matched permutation set enabled the weighting of candidates (through calculation of FDR) across all comparisons.

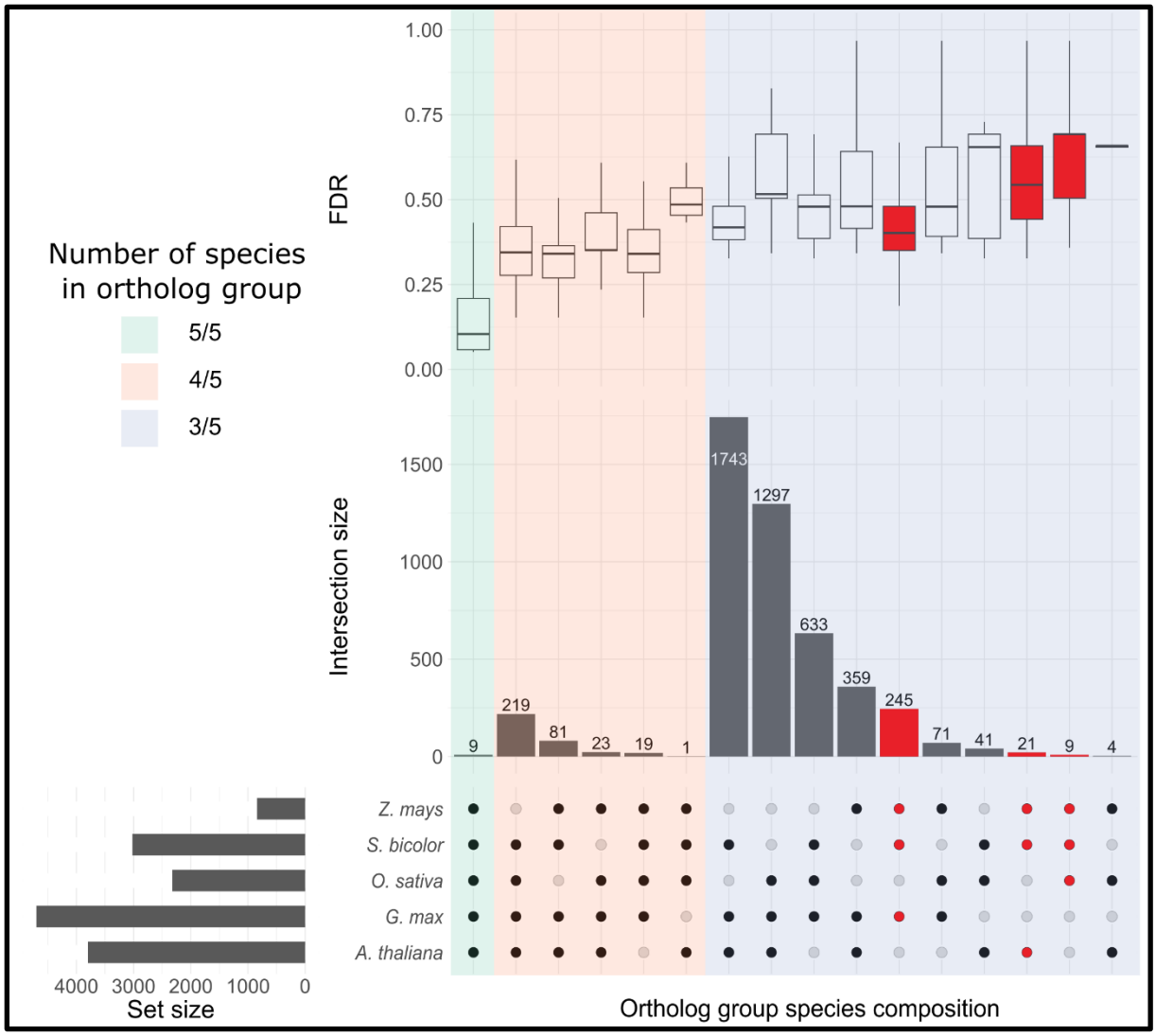


Fig 7. UpSet plot of FiReMAGE's selected candidates' ortholog group composition. The intersection matrix shows the unique compositions of all ortholog groups determined by OrthoFinder. Ortholog groups are ordered by the degree (number of species) of the intersection and then by group size. Intersection size represents the number of ortholog groups matching the species composition below. Set size shows the total number of instances in which a species appears in all ortholog groups. The top panel displays the individual genes' false discovery rate (FDR) distribution in their respective ortholog groups for all ionomic traits. Note that some genes may be represented twice in the boxplot distribution if they appear for multiple traits. Thus, they may have multiple FDRs due to the different SNP distributions and FDR of the ionomic traits. Because of the theory that high levels of synteny

between sorghum and maize may increase the number of false positives returned, ortholog groups with two of the three species being sorghum and maize are highlighted red.

Notable candidates

For each ionomic trait, FiReMAGE produced a list of orthologous candidates prioritized by likelihood: the FDRs as estimated by permutation (S7 Table). If elemental accumulation is controlled by conserved genes, then the FiReMAGE procedure should identify more orthologous gene/loci overlaps than expected by chance via the permutations. FDRs differed by element accumulation trait and species combinations. Generally, high-likelihood candidates were observed for elements with a previously demonstrated high heritability [66–70]. These elements also had ortholog groups with GWAS hits in all five species: Cd, Ca, Fe, Mn, and Mo (Fig 5). Of 23,117 FiReMAGE candidates, only 81 belong to these 5/5 ortholog groups. The value of orthology is highlighted by looking at these genes' positions in ranked candidate likelihoods. When ranked by likelihood, the top 56 FiReMAGE GWAS candidates are from the 81 candidates with GWAS hits in all five study species (S7 Table). When the top 10% of candidate genes were considered (232 genes; S8 Table), candidates 1-56 and 115-116 are linked to GWAS hits at orthologous candidates in all five species. The GWAS hits linked to these 5/5 candidate genes (n = 64 SNPs) were also enriched for the top 10% of the most significant SNPs from their respective GWAS at a Chi-squared estimated p-value of 1.25 e-10 (obs: 21 True, 43 False; exp: 6 True, 58 False). By design, the signal in GWAS is determining portions of our top 10% of candidates. Considered individually, candidates linked to the 21 GWAS hits that are the most significant SNPs would be targets for candidate verification. Using orthology, we can filter through the GWAS SNPs with "middling" signal to identify the remaining conserved candidates linked to the other 43 GWAS hits.

The identities of the genes encoded by the FIReMAGE candidates in our 5/5 species lists implicate a wide variety of biological processes in elemental homeostasis. The 81 candidates belong to nine

374 orthologous groups. For Fe, among the 5/5 candidates with no previously known role in Fe
375 homeostasis, is an ortholog group that encodes the Mevalonate kinases across these five species,
376 suggesting a currently unknown role for the mevalonic acid pathway in regulating Fe homeostasis in
377 flowering plants. Among the top Cd candidates, we find the *HMA* ortholog group mentioned above
378 in the KIG list (Table 3), RAB GTPase homologs of *RABA1i* that are involved in regulating cell wall
379 composition [71], and an ortholog group of Domains of Unknown Function 567 (DUF567) protein
380 family. In Arabidopsis, DUF567 is expressed in the roots and identified among a list of 129 genes
381 associated with metal hyperaccumulation in *Arabidopsis halleri* [72]. Thus, FireMAGE provides
382 additional experimental evidence, leveraged across five species, that this DUF567 family affects ion
383 homeostasis and may underlie ecological adaptation via hyperaccumulation in Arabidopsis relatives.
384 For Mo, the 5/5 candidates are the molybdate transporters, *MOT2*, listed in the KIG list for Mo
385 accumulation [48], and adenine nucleotide transporters (*NTT1* and *NTT2*). Nucleotide metabolism is
386 tied to Mo via molybdopterin, which is synthesized from nucleotide precursors [73] to form reaction
387 cofactors for four plant enzymes [74]. For Ca, the orthologous candidates are annotated as encoding
388 a Ring/Ubox superfamily with no known role in Ca accumulation. Lastly, for Mn, the candidate
389 ortholog groups are annotated as histone proteins and cellulose synthase A9 (*CESA9*). *CESA9* has
390 been shown to influence the cellulose content and cell wall biosynthesis in the epidermal seed coat
391 in Arabidopsis seeds [75]. In cellulose synthesis, Mn ions are used to catalyze glycosidic bond
392 formation, and in Arabidopsis, *AtCESA3* coordinates Mn ions with the glycosyltransferase substrate
393 uridine diphosphate glucose [76].
394
395 We also compared the FiReMAGE candidates to another candidate selection method performed in
396 maize: Camoco [19]. Out of over 47,000 candidate observations over multiple expression networks
397 and elemental trait combinations, we selected the top 200 observations determined by FDR rankings
398 from Camoco's multiple expression-network-based prioritizations and ionomics trait analysis. This
399 selection contained 122 unique candidate genes to compare against FiReMAGE's evolutionary-based

prioritizations. Of these, only two affected variations for the same trait in both methods, indicating that single-species expression networks and orthology selection methods boost different genetic signals to tie GWAS results to candidate genes. One gene, *Zm00001d04461*, identified through both methods as affecting Mo, encodes an F-box domain containing protein; the other, *Zm00001d012274*, affecting Fe accumulation, is an enzyme encoding an Omega-hydroxypalmitate O-feruloyl transferase involved in the production of suberinized hydrophobic barriers [77].

Discussion

We have demonstrated that combining orthology and GWAS from multiple species detects conserved regulators of phenotypic variation. Using the Known Ionomic Genes (KIG) as a set of curated annotations validates our hypothesis that combining GWAS and orthology will correctly identify variation affecting elemental accumulation. This resulted in the discovery of phenotypically-impactful alleles affecting ionomic variation in genes of previously known functions encoded in these five species. Additionally, the FiReMAGE approach goes beyond the KIG list by identifying numerous candidates in genes without previously determined roles in element accumulation (S7 Table). This ability to identify and prioritize candidate genes for conserved roles in trait variation using the natural variation of multiple species transcends one of the key limitations of the post-genomic era: ascertainment bias due to prior study. Instead of selecting candidates with prior annotations, these orthologs are annotated as high-value candidate genes for further molecular function characterization through the demonstrated orthology of genes linked to natural variation in multiple species. The overlap of GWAS results with genes of previously known function and the identification of orthologous variants using FiReMAGE produced a comprehensive public list of alleles at candidate conserved regulators of 19 elemental accumulation for future molecular mechanistic studies or marker-assisted selection (S2 and S7 Tables). Together, these two approaches can help identify candidate genes for traits that are common between species.

Orthology-based FiReMAGE is highly effective for identifying gene candidates

To detect unknown ionomic candidates that do not meet the criteria of the KIG list [24], we implemented the multi-species comparison we called FiReMAGE. Orthology clearly provides a good estimate of function, as demonstrated by the improved GWAS/KIG overlap (Figs 1 and 2, Table 2, S3 Table). As a result of this success, we hypothesized that orthology, even in the absence of gene function annotations, would also identify conserved genes as candidates without any prior demonstration of ionomic function. FiReMAGE identified more orthologous genes underlying GWAS for the same trait than expected by chance (Figs 4 and 5), demonstrating that this approach does work. In addition to KIGs (Table 3), FiReMAGE also returned non-KIG ortholog groups that encoded natural variation in elemental accumulation (S7 Table). These FiReMAGE candidates are likely to encode conserved regulators of elemental accumulation, for which our analysis is the first genetic demonstration of this role. Finding conserved critical regulators of ion homeostasis by GWAS validates the expectation that conserved regulators (e.g., hub genes) will repeatedly accumulate phenotypically effective variation. It also dramatically broadens the capacity of GWAS to work as a gene function discovery tool in the forward-genetics mode. Thus, comparing element accumulation association studies across organisms can help us discover additional element homeostasis mechanisms. Our approach relies on both annotated genomes and GWAS data, but it should be generalizable to any homologous traits across taxa. FiReMAGE does not require any information other than a list of GWAS SNPs and gene positions in genomes, thus avoiding prior knowledge bias and enabling accelerated gene function discovery for traits and species with little to no functional annotations or prior work.

One of the non-KIG ortholog groups with alleles in all five species detected by FiReMAGE for Cd encodes a domain of unknown function 567 protein (DUF567) family. This DUF567 is just one of 198

elemental accumulation candidate genes identified by FiReMAGE in Arabidopsis that has no previously known function in any biological processes (S9 Table). The remaining unknown genes are not in 5/5 ortholog groups, but when ranked by likelihood (permutation estimated FDRs), 19 are in the top 10% of Arabidopsis candidates for all traits, and 46 unknown genes are in the top 25%. None of these unknown genes would be considered candidates if relying only on gene annotations. The likelihood rankings returned by FiReMAGE and the fact that known conserved regulators also underlie significant GWAS peaks in multiple species (Table 3) suggest these are conserved, previously unknown regulators of the ionome. Further single-gene mechanistic experiments may result in the annotation of these genes as elemental regulators. Problems with annotating plant genes and Gene Ontology (GO) terms in general, and for plants in particular, have been extensively discussed [7–9]. FiReMAGE side-steps this problem, as is illustrated by the observation that it finds 602 orthologous gene candidates for Mo and 116 for Rb (S7 Table), yet there is no GO term explicitly for either Mo or Rb.

Using the data from multiple experiments, FiReMAGE emphasizes loci that would be passed over in a single experiment but have a strong signal when considered across taxa. Five traits, Ca, Fe, Mn, Mo, and Cd (Fig 5, S7 Table), have ortholog groups with candidate gene hits in all five species. Interestingly, these ortholog groups are substantially enriched for hits within the top 10% of most significant associations (p = 1.25e-10). FiReMAGE uses the federation of data across multiple species to boost signal below the p-value thresholds typically employed for single-species GWAS assessment. This is illustrated by the results underlying GWAS hits at the ortholog groups containing *MOT2* and the Cd/Zn transporting *HMAs* (Table 3). The worst performing taxa for these ortholog groups were soybean for Mo accumulation at *MOT2* (p-value = 0.0104) and Arabidopsis for Cd accumulation at the tandem gene duplication comprising AtHMA2, AtHMA3, and AtHMA4 (p-value = 5.6e-05) - neither meeting strict stringency cutoffs in their respective GWAS assessments (S7 Table).

FiReMAGE approach and utility for gene discovery

Our FiReMAGE approach combines a wide array of information. It has many steps (functions) and dozens to hundreds of user-defined parameters with no obvious correct choice. These include the p-value thresholds for each GWAS, the linkage range for SNP-to-gene relationships, and the outgroups for constructing ortholog relationships. In addition, several variables are inherent in the data and experimentation that also affect FiReMAGE, including differences in measurement error in each experiment, variation in the statistical power of each GWAS panel, the total genetic variation in a species, and the association panel selection. By using permutations instead of an idealized null distribution, we can estimate how likely the results are to happen by random chance and provide estimates of the likelihood that our predictions are true. The FiReMAGE approach calculates these permutations for every trait, species, and dataset combination. As researchers get better estimates of parameters or add additional data, the permutations must also be updated. Preliminary testing determined that random selection parameters were replicating real data structure and composition, and our permutations serve as consistent internal controls for benchmarking our real results.

We expected that the high synteny between the three Poaceae species, rice, sorghum, and maize, would inflate false positives. To the extent that multiple linked genes are shared across taxa due to conservation of gene order, we expected the ortholog groups detected as 3/5 hits in FiReMAGE in only the Poaceae to have high FDR. However, these ortholog groups did not have increased FDR compared to other 3/5 ortholog groups (Fig 7). The combination of the recent genome duplications and uneven loss of duplicate genes well demonstrated in maize, Arabidopsis, and others [78–83] and the prevalence of chromosomal rearrangements and transposition of genes in plant genomes may have shuffled these genomes enough to permit the detection of single conserved genes within our selected window sizes. Although we did not see a huge influence of synteny in our comparison, other species combinations may find controlling for synteny critical [84]. In all cases, the undue

499 influence of synteny will be reflected in the permutation FDR calculations, highlighting the value of
500 empirically estimated expectations over naive null hypotheses.
501
502 The success of FiReMAGE with the particular data sets and user-set parameters we chose here
503 demonstrates that the approach is viable. One feature of this analysis was the reliance on user-
504 reported p-values for a number of the data sets. This resulted in uneven p-value thresholds and
505 presumably false positive and negative rates across the input experiments. Because FiReMAGE
506 federates all the data together, the optimal p-value thresholds for each experiment are dependent
507 on the performance of the comparison. One interesting feature of FiReMAGE is that it is highly
508 sensitive to false-negative rates, a problem made larger as more species are brought together. Using
509 too stringent of a p-value cutoff for the GWAS results input into FiReMAGE can fail to include a true
510 association of modest effect near a conserved regulator. FiReMAGE cannot then include this gene in
511 an overlap or loses one true overlap as a result. We chose p-value thresholds considered permissive
512 by community standards (see methods) to reasonably limit false negative rates. As candidate genes
513 suggested by FiReMAGE are tested, the genes that are confirmed and genes that were false positives
514 will provide an additional mechanism to compare and refine parameters.
515
516 The high FDRs from the KIG comparison demonstrate why common practices filter SNPs with
517 stringent thresholds (Bonferroni, FDR, etc.) before referencing functional annotations of linked
518 genes. However, these recommendations do not address that even with generous genome coverage,
519 KIG candidates cannot explain a large portion of ionomic variation (Figs 1 and 2). While some GWAS
520 loci are due to spurious associations and some KIGs do not drive elemental phenotypes in all
521 populations in all species, we believe the majority of loci lack KIG candidates because the KIG list is
522 incomplete [8]. Candidate selection methods reliant on prior information have inherently narrow
523 scopes that increase false negatives and inhibit gene discovery. The KIG list is slightly better at
524 discovery, as it is valuable in transferring information between species with few and abundant

525 annotations. However, identifying novel ionomic candidates or functional annotations requires novel
526 approaches, such as FiReMAGE.

527 Limitations of FiReMAGE

528 One effect of the FiReMAGE approach on candidate gene assessment is that it rejects alleles and
529 false positives that occur in GWAS results from a single species or in a limited number of species.
530 While of great potential utility, this approach can potentially contribute to false negative
531 assessments in at least two ways. 1) The FiReMAGE approach will necessarily miss taxonomically-
532 restricted solutions and limitations to element homeostasis. For example, glucosinolates in
533 Arabidopsis, benzoxazinoids in maize, and nodulation in soybeans, all have impacts on the ionome
534 and are not evolutionarily conserved processes across these taxa. As a result, alleles affecting
535 variation in these processes, even if they have substantial impacts in one taxon, would not be
536 prioritized by FiReMAGE. The KIG list or more involved experiments like gene transcript expression
537 would be better selection methods in these cases. 2) The requirement to appear in multiple lists and
538 the stringency of the criteria for inclusion in the GWAS outputs contribute to false negatives intrinsic
539 in FiReMAGE. Researchers particularly worried about false positive assessments in their GWAS
540 outputs may not have considered the effect that the high rate of false negatives will have on
541 downstream meta-analysis and data federation approaches, like FiReMAGE. Because detecting
542 conserved candidates by FiReMAGE relies on true positives in multiple species, providing too-
543 stringent lists of loci will result in poor performance as more species are added to the comparison.
544
545 Another possible limitation is that FiReMAGE requires strict trait matches across taxa and we only
546 compared KIGs to loci of traits for which they had been cited. However, many elemental uptake
547 traits are related and often co-transported by the same transporters. The same gene could have
548 different elemental phenotypes in taxa due to many factors, including the interaction with species-
549 specific mechanisms (see above). The rules governing elemental interactions are poorly understood

[85,86]. Rather than attempting to relate ortholog groups to similar elements they have yet to be tested for with our (likely) biased understanding of the ionome, we chose a strict comparison between traits across species. We have provided a list of KIGs overlapping any trait loci (S4 Table) for exploring their potential functions in related traits.

A functional requirement for FiReMAGE is the collection of the same trait across multiple diverse taxa. Many traits, such as flower morphology and resistance to specific pathogens, are not comparable across species. Additionally, collecting trait data from GWAS size populations across multiple species can be time and resource-intensive. However, advances in phenotyping technologies are enabling the creation of multi-species datasets, and more datasets are publicly available in GWAS databases like AraGWAS [87] for Arabidopsis, RiceVarMap2 [88] for rice, and GWAS Atlas [89] for studies in cotton, maize, soybean, rice, and sorghum. FiReMAGE is able to identify candidates for any trait where the data is available in three or more species. We note that the threshold of three or more species excludes eudicot-specific candidates in this five-species comparison, and we may include more eudicot species as their ionomic GWAS experiments become available.

Lastly, because FiReMAGE relies on orthology, the limitations surrounding current ortholog group inference methods apply to the candidate lists produced by FiReMAGE. While OrthoFinder v2.0 is the best algorithm for whole genome comparisons [43], there are likely some gene trees with evolutionary histories that make accurate ortholog inference challenging [44]. We recommend following best practices for operating OrthoFinder (see methods), including proper species sampling and validating gene tree phylogenies for FiReMAGE candidates of interest when appropriate.

Future mutant screens

Unbiased candidate selection methods are necessary to transition into research using non-model species and with genes of unknown function. Our approach has identified the conserved genes underlying genetic variation via analogous GWAS in multiple species and their likelihood as causal genes in ionomic phenotypic variation. In future studies, testing mutant knock-out libraries in Arabidopsis [90], sorghum [91], and maize [92] will allow us to validate FiReMAGE predictions and improve candidate selection parameters.

Our research demonstrates that the disparate approaches of the KIG list and FiReMAGE both present valuable ionomic candidates, each with distinct advantages and weaknesses, and highlight the importance of strategic decisions surrounding candidate gene identification for downstream quantitative genetic analysis. Future experiments will further explore best practices and outcomes of comparative GWAS methods, including parameters to balance false negatives and false positives. We will also explore implementing a ranking equation that considers other variables besides the number of species in ortholog groups. We hope this tool facilitates gene discovery in plant genetics.

Methods

Data availability

R scripts created for this version of the KIG list can be found at https://github.com/danforthcenter/KIG_v2. FiReMAGE scripts, GWAS SNPs used in this paper, ortholog tables, and gene coordinates can be found at https://github.com/danforthcenter/FiReMAGE. Gene coordinates can alternatively be downloaded from Phytozome (https://phytozome-next.jgi.doe.gov/phytomine/begin.do).

GWAS datasets

Single nucleotide polymorphism (SNP) filtering

The SNPs from each GWAS dataset were selected by p-value thresholds that were separately determined for each species: *Arabidopsis thaliana* (Arabidopsis; $p < 0.001$), *Glycine max* (soybean; $p < 0.05$), *Oryza sativa* (rice; $p < 1e-5$), *Sorghum bicolor* (sorghum; $p < 1e-4$), and *Zea mays* (maize; $p < 0.01$). The p-value thresholds used are intentionally permissive, especially in soybean, to limit spurious SNPs but mostly to reduce false negatives in our GWAS/ortholog overlaps.

SNPs from each dataset are collapsed into loci using general linkage ranges for each species (Table 4). The p-value from the most significant SNP is retained as the collapsed loci's p-value. Collapsing SNPs reduces computational load and avoids genes overlapping neighboring/redundant SNPs. The pipeline also accepts files already in loci intervals, like the maize NAM dataset.

Table 4. Overview of GWAS datasets.

Species Annotation	general linkage range	input as loci	SNPs	collapsed SNPs/loci
A. thaliana_TAIR10	75000	FALSE	7167	1404
G. max_Wm82.a2.v1	200000	FALSE	37854	7308
O. sativa_v7.0	243000	FALSE	17109	432
S. bicolor_v3.1.1	100000	FALSE	2139	1768
Z. mays_RefGen_V4	100000	TRUE	905	905

Soybean GWAS and 2015-2017 field experiments

Glycine max genotypes were obtained from the USDA-ARS GRIN collection (https://npgsweb.ars-grin.gov/gringlobal/). Field experiments were performed over three years (2015, 2016, and 2017) with three field replicates at each field experiment in a random complete block design. The 2015 and 2016 field experiments were performed at the University of Missouri Bradford Research farm

located near Columbia, Missouri, (38°53'38.4"N 92°11'47.1" W) which has a Mexico silt loam (fine, smectitic, and mesic Vertic Epiaqualf) and the 2017 field experiment was performed at the Rollins Bottom research farm located near Columbia, MO, ((38°55’37.5” N, 92°20’44.6” W) on a Haymond silt loam soil (course-silty, mixed, superactive, mesic Dystric Fluventic Eutrudepts). Seeds were planted at a rate of 8 seed/ft, over 6 ft with a 2 ft gap between plots in 2015 and 2016, and 3 seed/ft over 8 ft with a 4 ft gap between plots in 2017. Row spacing was 30” between rows in all experiments. Field conditions were typical of soybean production in the Midwest USA, with NPK fertilizer applied at rates appropriate per soil analyses, and pre-emergent and post-emergent herbicides were applied as recommended to control weeds for conventional soybeans. Field plots were planted on 06/30/2015, 05/06/2016, and 06/06/2017. Plots were hand-harvested and threshed using a small plot thresher by 10/29/2015, 09/23/2016, and 11/17/2017. Flower color for each field plot was recorded and compared to known colors for each genotype. Similarly, seeds were inspected after threshing and compared to known hila and seedcoat coloration, with non-matching samples discarded. Approximately 25 seeds were lyophilized for each field plot and ground to a fine powder using a Perten 3310 lab mill using a “type 0” grinding disc (Perten Instruments, Sweden).

Powdered samples from each grow-out were analyzed concurrently for concentrations of 20 elements using an inductively coupled plasma mass spectrometry (ICP-MS) pipeline previously described [93]. Batch effects inherent in the ICP-MS process were confounded with grow-out in the initial analysis. Thus, a subset of samples from each batch of the initial ICP-MS analysis were re-analyzed for elemental concentrations, and a batch-specific correction factor was generated and subtracted from the first run’s values. Analytical outliers identified as > 10 median absolute deviations away from the median were removed within grow out.

For analysis within each grow-out, phenotypic values for each genotype were taken as the median value within that grow-out. To better meet normality assumptions both for the GWAS and for aggregation of the grow-outs using a mixed model method, phenotypes were Box-Cox transformed

within the out and across grow-out and then standardized. Lambda values for the Box-Cox were chosen using a 95% confidence interval. To aggregate data from different grow-outs, best linear unbiased predictors (BLUPs) for each genotype were calculated from a random intercept mixed model: $Phenotype \sim (1|Genotype) + (1|Grow\ out) + (1|Genotype{:}\ Grow\ out)$.

Association mapping was performed for each phenotype in the three grow-outs and the combined grow-out BLUPs. Genotype sample sizes for each grow-out are 355, 363, and 364 for 2015, 2016, and 2017, respectively. To account for population structure, kinship matrix and principal components were included in the mixed model as fixed effects. The kinship matrix was calculated using all 365 genotypes present in at least one grow-out following the realizedAB method in the synbreed R package [94] as implemented in the R package ionomicsUtils (v1.0 https://github.com/gziegler/ionomicsUtils). Principal components included in the model were selected for each phenotype if they were correlated significantly with that phenotype at the alpha = .001 level.

The multi-locus mixed model (MLMM) used performs forward inclusion and backwards elimination until all of the heritable variation is explained or a maximum number of cofactors has been reached, which in this study was 40. Model selection was performed using the extended Bayesian information criterion [95] and the multiple-Bonferroni criterion [96]. A third set of SNP's was also returned based on inclusion in the final forward inclusion step (max cofactor). Effect sizes for SNPs returned were calculated using the mlmm.gwas package (S10 Table) [97]. A resulting SNP was said to overlap with a neighboring SNP if the r^{-2} linkage disequilibrium was > .2 as calculated using the LD function in the genetics R package [98] and implemented in the R package ionomicsUtils (v1.0 https://github.com/gziegler/ionomicsUtils). Average linkage disequilibrium values for all pairs of SNPs were calculated across the genome by first separating SNP pairs into 100 kB bins based on distance and then calculating LD within these bins [99].

667 ## OrthoFinder orthologs

668 OrthoFinder v2.0 [43] was downloaded from https://github.com/davidemms/OrthoFinder and used
669 with default parameters and best practices outlined here:
670 https://davidemms.github.io/orthofinder_tutorials/orthofinder-best-practices.html. Protein reads of
671 primary transcripts were obtained from Phytozome (v13) [42]. To make the KIG list, we used files for
672 all species in the list except for wheat due to its low-quality genome annotation. *Liriodendron*
673 *tulipifera* was included as the most closely related outgroup to the eudicots and monocots in
674 Phytozome v13. The resulting ortholog table included one-to-one, one-to-many, and many-to-many
675 ortholog relationships.

676 ## Loci/gene overlaps

677 We obtained all the Phytozome (v13) [42] gene coordinates for each species via the BioMart tool
678 (v0.7) located on the Phytozome homepage (https://phytozome-next.jgi.doe.gov). Using the
679 coordinates, we search for genes overlapping trait loci that are also in our 5-species ortholog groups.
680 For each trait, FiReMAGE retains genes from ortholog groups with overlaps in at least 3/5 species
681 and produces a list of prioritized candidate genes. The number of overlapping orthologs is recorded
682 for each locus to calculate false discovery rates.

683 ## Random permutations

684 GWAS loci permutations were 1000 selections of random locations equal to the actual datasets'
685 species and trait composition. Each species' chromosome start and end coordinates were
686 downloaded from the European Nucleotide Archive (ENA) database (GCA_000001735,
687 GCA_000004515, GCA_001433935, GCA_000003195, GCA_000005005). Because random base pair
688 locations are unlikely to be in range for collapsing into "loci", we assigned range lengths randomly

chosen from the actual GWAS loci. If random locations are close together, their randomly assigned ranges may overlap. FiReMAGE runs the random permutations through the same loci/gene overlap pipeline as the actual data, calculates the permutation's distributions to evaluate the likelihood of a locus overlapping any nearby ortholog, and produces a false discovery rate (FDR) to prioritize candidate genes. FDR was calculated as the mean number of orthologs overlapping random locations divided by the number of actual orthologs overlapping loci.

The primary KIG list permutations (S3 and S4 Figs, S6 Table) were formed by selecting random genes to replicate the same species and element compositions of the actual primary KIG list and repeating 1000 times. The inferred KIG list permutations were formed by putting the primary genes permutations through the same ortholog-finding process as the actual KIG list, with one deviation. In the actual inferred lists, a primary KIG's paralogs (if any) and orthologs are sourced from its ortholog group. In the permutations, we only take the random gene's orthologs from its ortholog group. After finding all the orthologs for the random genes, we add a random selection of genes from Arabidopsis, rice, and maize that equal the number of paralogs in their actual inferred KIG lists. Specifically, it is the number of paralogs only in the list as inferred genes - not primary genes that are also paralogs to each other. This deviation was made to avoid random gene selections in large, single-species gene families, as this would no longer randomly replicate the structure of the KIG list.

Funding

This study was supported by the National Science Foundation (www.nsf.gov) grant IOS-2309932 (IB, BPD), the United States Department of Energy Office of Science (BER; https://science.osti.gov/ber) grant DE-SC0023305 (BPD), the Biotechnology and the Biological Sciences Research Council (https://www.ukri.org/councils/bbsrc/) grants BB/J003336/1 (AHP, GJN, DES) and BB/L000113/1 (DES), and Donald Danforth Plant Science Center (https://www.danforthcenter.org) funds (LW). The funders had no role in

study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We thank Ryan Hartsock for contributions to the early development of our orthology comparison methods.

References

1. Shakoor N, Ziegler G, Dilkes BP, Brenton Z, Boyles R, Connolly EL, et al. Integration of Experiments across Diverse Environments Identifies the Genetic Determinants of Variation in Sorghum bicolor Seed Element Composition. Plant Physiol. 2016;170: 1989–1998.

2. Ziegler G, Nelson R, Granada S, Krishnan HB, Gillman JD, Baxter I. Genomewide association study of ionomic traits on diverse soybean populations from germplasm collections. Plant Direct. 2018;2: e00033.

3. Ziegler G, Kear PJ, Wu D, Ziyomo C, Lipka AE, Gore M, et al. Elemental Accumulation in Kernels of the Maize Nested Association Mapping Panel Reveals Signals of Gene by Environment Interactions. bioRxiv. 2017. p. 164962. doi:10.1101/164962

4. Zhang M, Pinson SRM, Tarpley L, Huang X-Y, Lahner B, Yakubova E, et al. Mapping and validation of quantitative trait loci associated with concentrations of 16 elements in unmilled rice grain. Theor Appl Genet. 2014;127: 137–165.

5. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature. 2010;465: 627–631.

6. Cubillos FA, Coustham V, Loudet O. Lessons from eQTL mapping studies: non-coding regions

735 and their role behind natural phenotypic variation in plants. Curr Opin Plant Biol. 2012;15: 192–
736 198.

737 7. Stoeger T, Gerlach M, Morimoto RI, Nunes Amaral LA. Large-scale investigation of the reasons
738 why potentially important genes are ignored. PLoS Biol. 2018;16: e2006643.

739 8. Baxter I. We aren't good at picking candidate genes, and it's slowing us down. Curr Opin Plant
740 Biol. 2020;54: 57–60.

741 9. Wimalanathan K, Friedberg I, Andorf CM, Lawrence-Dill CJ. Maize GO Annotation—Methods,
742 Evaluation, and Review (maize-GAMER). Plant Direct. 2018;2: e00052.

743 10. Stoeger T, Nunes Amaral LA. COVID-19 research risks ignoring important host genes due to pre-
744 established research patterns. Elife. 2020;9. doi:10.7554/eLife.61981

745 11. Chao D-Y, Silva A, Baxter I, Huang YS, Nordborg M, Danku J, et al. Genome-wide association
746 studies identify heavy metal ATPase3 as the primary determinant of natural variation in leaf
747 cadmium in Arabidopsis thaliana. PLoS Genet. 2012;8: e1002923.

748 12. Chao D-Y, Chen Y, Chen J, Shi S, Chen Z, Wang C, et al. Genome-wide association mapping
749 identifies a new arsenate reductase enzyme critical for limiting arsenic accumulation in plants.
750 PLoS Biol. 2014;12: e1002009.

751 13. Gupta PK, Kulwal PL, Jaiswal V. Association mapping in plants in the post-GWAS genomics era.
752 Adv Genet. 2019;104: 75–154.

753 14. Spain SL, Barrett JC. Strategies for fine-mapping complex traits. Hum Mol Genet. 2015;24:
754 R111–9.

755 15. Koornneef M, Meinke D. The development of Arabidopsis as a model plant. Plant J. 2010;61:
756 909–921.

757 16. Park ST, Kim J. Trends in Next-Generation Sequencing and a New Era for Whole Genome
758 Sequencing. Int Neurourol J. 2016;20: S76–83.

759 17. Greene CS, Troyanskaya OG. PILGRM: an interactive data-driven discovery platform for expert
760 biologists. Nucleic Acids Res. 2011;39: W368–74.

761 18. Xing EP, Curtis RE, Schoenherr G, Lee S, Yin J, Puniyani K, et al. GWAS in a box: statistical and
762 visual analytics of structured associations via GenAMap. PLoS One. 2014;9: e97524.

763 19. Schaefer RJ, Michno J-M, Jeffers J, Hoekenga O, Dilkes B, Baxter I, et al. Integrating
764 Coexpression Networks with GWAS to Prioritize Causal Genes in Maize. Plant Cell. 2018;30:
765 2922–2942.

766 20. Salt DE, Baxter I, Lahner B. Ionomics and the study of the plant ionome. Annu Rev Plant Biol.
767 2008;59: 709–733.

768 21. Chen A, Hu J, Sun S, Xu G. Conservation and divergence of both phosphate- and mycorrhiza-
769 regulated physiological responses and expression patterns of phosphate transporters in
770 solanaceous species. New Phytol. 2007;173: 817–831.

771 22. Chen A, Gu M, Sun S, Zhu L, Hong S, Xu G. Identification of two conserved cis-acting elements,
772 MYCS and P1BS, involved in the regulation of mycorrhiza-activated phosphate transporters in
773 eudicot species. New Phytol. 2011;189: 1157–1169.

774 23. Ram H, Sardar S, Gandass N. Vacuolar Iron Transporter (Like) proteins: Regulators of cellular
775 iron accumulation in plants. Physiol Plant. 2021;171: 823–832.

776 24. Whitt L, Ricachenevsky FK, Ziegler GZ, Clemens S, Walker E, Maathuis FJM, et al. A curated list
777 of genes that affect the plant ionome. Plant Direct. 2020;4: e00272.

778 25. Véry A-A, Nieves-Cordones M, Daly M, Khan I, Fizames C, Sentenac H. Molecular biology of K+

779 transport across the plant cell membrane: what do we learn from comparison between plant
780 species? J Plant Physiol. 2014;171: 748–769.

781 26. Baxter I, Muthukumar B, Park HC, Buchner P, Lahner B, Danku J, et al. Variation in molybdenum
782 content across broadly distributed populations of Arabidopsis thaliana is controlled by a
783 mitochondrial molybdenum transporter (MOT1). PLoS Genet. 2008;4: e1000004.

784 27. Baxter I, Brazelton JN, Yu D, Huang YS, Lahner B, Yakubova E, et al. A coastal cline in sodium
785 accumulation in Arabidopsis thaliana is driven by natural variation of the sodium transporter
786 AtHKT1;1. PLoS Genet. 2010;6: e1001193.

787 28. Rubio F, Nieves-Cordones M, Alemán F, Martínez V. Relative contribution of AtHAK5 and
788 AtAKT1 to K+ uptake in the high-affinity range of concentrations. Physiol Plant. 2008;134: 598–
789 608.

790 29. Eide D, Broderius M, Fett J, Guerinot ML. A novel iron-regulated metal transporter from plants
791 identified by functional expression in yeast. Proc Natl Acad Sci U S A. 1996;93: 5624–5628.

792 30. Huang X-Y, Deng F, Yamaji N, Pinson SRM, Fujii-Kashino M, Danku J, et al. A heavy metal P-type
793 ATPase OsHMA4 prevents copper accumulation in rice grain. Nat Commun. 2016;7: 12138.

794 31. Campos ACA, van Dijk WFA, Ramakrishna P, Giles T, Korte P, Douglas A, et al. Full species-wide
795 leaf and seed ionomic diversity of Arabidopsis thaliana. bioRxiv. 2020. p. 2020.11.09.373282.
796 doi:10.1101/2020.11.09.373282

797 32. Norton GJ, Travis AJ, Douglas A, Fairley S, Alves EDP, Ruang-Areerate P, et al. Genome Wide
798 Association Mapping of Grain and Straw Biomass Traits in the Rice Bengal and Assam Aus Panel
799 (BAAP) Grown Under Alternate Wetting and Drying and Permanently Flooded Irrigation. Front
800 Plant Sci. 2018;9: 1223.

801 33. Norton GJ, Douglas A, Lahner B, Yakubova E, Guerinot ML, Pinson SR, et al. Genome wide
802 association mapping of grain arsenic, copper, molybdenum and zinc in rice (Oryza sativa L.)
803 grown at four international field sites. PLoS One. 2014;9: e89685.

804 34. Norton GJ, Travis A, Ruang-Areerate P, Nicol GW, Adeosun AA, Hossain M, et al. Genetic loci
805 regulating cadmium content in rice grains. Euphytica. 2021;217: 35.

806 35. Talukdar P, Travis AJ, Hossain M, Islam MR, Norton GJ, Price AH. Identification of genomic loci
807 regulating grain iron content in aus rice under two irrigation management systems. Food
808 Energy Secur. 2022;11: e329.

809 36. Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. Resolving the ortholog conjecture:
810 orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS
811 Comput Biol. 2012;8: e1002514.

812 37. Asaro A, Dilkes B, Baxter I. Multivariate analysis reveals environmental and genetic
813 determinants of element covariation in the maize grain ionome. bioRxiv. 2017. p. 241380.
814 doi:10.1101/241380

815 38. Ziegler G, Kear P, Wu D, Ziyomo C, Lipka A, Gore M. Elemental Accumulation in Kernels of the
816 Maize Nested Association Mapping Panel Reveals Signals of Gene by Environment Interactions.
817 BioRxiv. 2017. Available: https://www.biorxiv.org/content/10.1101/164962v1.abstract

818 39. Campbell MT, Bandillo N, Al Shiblawi FRA, Sharma S, Liu K, Du Q, et al. Allelic variants of
819 OsHKT1;1 underlie the divergence between indica and japonica subspecies of rice (Oryza
820 sativa) for root sodium content. PLoS Genet. 2017;13: e1006823.

821 40. Rus A, Baxter I, Muthukumar B, Gustin J, Lahner B, Yakubova E, et al. Natural variants of AtHKT1
822 enhance Na+ accumulation in two wild populations of Arabidopsis. PLoS Genet. 2006;2: e210.

823 41. Remm M, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from
824 pairwise species comparisons. J Mol Biol. 2001;314: 1041–1052.

825 42. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative
826 platform for green plant genomics. Nucleic Acids Res. 2012;40: D1178–86.

827 43. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics.
828 Genome Biol. 2019;20: 238.

829 44. Emms DM, Kelly S. Benchmarking Orthogroup Inference Accuracy: Revisiting Orthobench.
830 Genome Biol Evol. 2020;12: 2258–2266.

831 45. Haydon MJ, Kawachi M, Wirtz M, Hillmer S, Hell R, Krämer U. Vacuolar nicotianamine has
832 critical and distinct roles under iron deficiency and for zinc sequestration in Arabidopsis. Plant
833 Cell. 2012;24: 724–737.

834 46. Remy E, Cabrito TR, Batista RA, Teixeira MC, Sá-Correia I, Duque P. The major facilitator
835 superfamily transporter ZIFL2 modulates cesium and potassium homeostasis in Arabidopsis.
836 Plant Cell Physiol. 2015;56: 148–162.

837 47. Che J, Yokosho K, Yamaji N, Ma JF. A Vacuolar Phytosiderophore Transporter Alters Iron and
838 Zinc Accumulation in Polished Rice Grains. Plant Physiol. 2019;181: 276–288.

839 48. Gasber A, Klaumann S, Trentmann O, Trampczynska A, Clemens S, Schneider S, et al.
840 Identification of an Arabidopsis solute carrier critical for intracellular transport and inter-organ
841 allocation of molybdate. Plant Biol . 2011;13: 710–718.

842 49. Gil-Díez P, Tejada-Jiménez M, León-Mediavilla J, Wen J, Mysore KS, Imperial J, et al. MtMOT1.2
843 is responsible for molybdate supply to Medicago truncatula nodules. Plant Cell Environ. 2018.
844 doi:10.1111/pce.13388

845 50. Tejada-Jiménez M, Gil-Díez P, León-Mediavilla J, Wen J, Mysore KS, Imperial J, et al. Medicago
846 truncatula Molybdate Transporter type 1 (MtMOT1.3) is a plasma membrane molybdenum
847 transporter required for nitrogenase activity in root nodules under molybdenum deficiency.
848 New Phytol. 2017;216: 1223–1235.

849 51. Huang X-Y, Liu H, Zhu Y-F, Pinson SRM, Lin H-X, Guerinot ML, et al. Natural variation in a
850 molybdate transporter controls grain molybdenum concentration in rice. New Phytol.
851 2019;221: 1983–1997.

852 52. Hussain D, Haydon MJ, Wang Y, Wong E, Sherson SM, Young J, et al. P-type ATPase heavy metal
853 transporters with roles in essential zinc homeostasis in Arabidopsis. Plant Cell. 2004;16: 1327–
854 1339.

855 53. Olsen LI, Hansen TH, Larue C, Østerberg JT, Hoffmann RD, Liesche J, et al. Mother-plant-
856 mediated pumping of zinc into the developing seed. Nat Plants. 2016;2: 16036.

857 54. Tanaka N, Nishida S, Kamiya T, Fujiwara T. Large-scale profiling of brown rice ionome in an
858 ethyl methanesulphonate-mutagenized hitomebore population and identification of high- and
859 low-cadmium lines. Plant Soil. 2016;407: 109–117.

860 55. Kim D-Y, Bovet L, Maeshima M, Martinoia E, Lee Y. The ABC transporter AtPDR8 is a cadmium
861 extrusion pump conferring heavy metal resistance. Plant J. 2007;50: 207–218.

862 56. Shin H, Shin H-S, Dewbre GR, Harrison MJ. Phosphate transport in Arabidopsis: Pht1;1 and
863 Pht1;4 play a major role in phosphate acquisition from both low- and high-phosphate
864 environments. Plant J. 2004;39: 629–642.

865 57. Nagarajan VK, Jain A, Poling MD, Lewis AJ, Raghothama KG, Smith AP. Arabidopsis Pht1;5
866 mobilizes phosphate between source and sink organs and influences the interaction between
867 phosphate homeostasis and ethylene signaling. Plant Physiol. 2011;156: 1149–1163.

868 58. Catarecha P, Segura MD, Franco-Zorrilla JM, García-Ponce B, Lanza M, Solano R, et al. A mutant
869 of the Arabidopsis phosphate transporter PHT1;1 displays enhanced arsenic accumulation.
870 Plant Cell. 2007;19: 1123–1133.

871 59. Zhang L, Hu B, Li W, Che R, Deng K, Li H, et al. OsPT2, a phosphate transporter, is involved in
872 the active uptake of selenite in rice. New Phytol. 2014;201: 1183–1191.

873 60. Peiter E, Montanini B, Gobert A, Pedas P, Husted S, Maathuis FJM, et al. A secretory pathway-
874 localized cation diffusion facilitator confers plant manganese tolerance. Proc Natl Acad Sci U S
875 A. 2007;104: 8532–8537.

876 61. Lanquar V, Ramos MS, Lelièvre F, Barbier-Brygoo H, Krieger-Liszkay A, Krämer U, et al. Export of
877 vacuolar manganese by AtNRAMP3 and AtNRAMP4 is required for optimal photosynthesis and
878 growth under manganese deficiency. Plant Physiol. 2010;152: 1986–1999.

879 62. Hindt MN, Akmakjian GZ, Pivarski KL, Punshon T, Baxter I, Salt DE, et al. BRUTUS and its
880 paralogs, BTS LIKE1 and BTS LIKE2, encode important negative regulators of the iron deficiency
881 response in Arabidopsis thaliana. Metallomics. 2017;9: 876–890.

882 63. Yuan Y, Wu H, Wang N, Li J, Zhao W, Du J, et al. FIT interacts with AtbHLH38 and AtbHLH39 in
883 regulating iron uptake gene expression for iron homeostasis in Arabidopsis. Cell Res. 2008;18:
884 385–397.

885 64. Hosmani PS, Kamiya T, Danku J, Naseer S, Geldner N, Guerinot ML, et al. Dirigent domain-
886 containing protein is part of the machinery required for formation of the lignin-based Casparian
887 strip in the root. Proc Natl Acad Sci U S A. 2013;110: 14498–14503.

888 65. Baxter I, Hosmani PS, Rus A, Lahner B, Borevitz JO, Muthukumar B, et al. Root suberin forms an
889 extracellular barrier that affects water relations and mineral nutrition in Arabidopsis. PLoS
890 Genet. 2009;5: e1000492.

891 66. Asaro A, Ziegler G, Ziyomo C, Hoekenga OA, Dilkes BP, Baxter I. The Interaction of Genotype
892 and Environment Determines Variation in the Maize Kernel Ionome. G3 . 2016;6: 4175–4183.

893 67. Baxter IR, Ziegler G, Lahner B, Mickelbart MV, Foley R, Danku J, et al. Single-kernel ionomic
894 profiles are highly heritable indicators of genetic and environmental influences on elemental
895 accumulation in maize grain (Zea mays). PLoS One. 2014;9: e87628.

896 68. Campos ACAL, van Dijk WFA, Ramakrishna P, Giles T, Korte P, Douglas A, et al. 1,135 ionomes
897 reveal the global pattern of leaf and seed mineral nutrient and trace element diversity in
898 Arabidopsis thaliana. Plant J. 2021;106: 536–554.

899 69. Taliercio E, Scaboo A, Baxter I, Locke AM. The Ionome of a Genetically Diverse Set of Wild
900 Soybean Accessions. Crop Sci. 2019. doi:10.2135/cropsci2019.02.0079

901 70. Norton GJ, Deacon CM, Xiong L, Huang S, Meharg AA, Price AH. Genetic mapping of the rice
902 ionome in leaves and grain: identification of QTLs for 17 elements including arsenic, cadmium,
903 iron and selenium. Plant Soil. 2010;329: 139–153.

904 71. Lunn D, Gaddipati SR, Tucker GA, Lycett GW. Null mutants of individual RABA genes impact the
905 proportion of different cell wall components in stem tissue of Arabidopsis thaliana. PLoS One.
906 2013;8: e75724.

907 72. Filatov V, Dowdle J, Smirnoff N, Ford-Lloyd B, Newbury HJ, Macnair MR. Comparison of gene
908 expression in segregating families identifies genes and genomic regions involved in a novel
909 adaptation, zinc hyperaccumulation. Mol Ecol. 2006;15: 3045–3059.

910 73. Hover BM, Loksztejn A, Ribeiro AA, Yokoyama K. Identification of a cyclic nucleotide as a cryptic
911 intermediate in molybdenum cofactor biosynthesis. J Am Chem Soc. 2013;135: 7019–7032.

912 74. Schwarz G, Mendel RR. Molybdenum cofactor biosynthesis and molybdenum enzymes. Annu

913 Rev Plant Biol. 2006;57: 623–647.

914 75. Stork J, Harris D, Griffiths J, Williams B, Beisson F, Li-Beisson Y, et al. CELLULOSE SYNTHASE9
915 serves a nonredundant role in secondary cell wall synthesis in Arabidopsis epidermal testa cells.
916 Plant Physiol. 2010;153: 580–589.

917 76. Qiao Z, Lampugnani ER, Yan X-F, Khan GA, Saw WG, Hannah P, et al. Structure of *Arabidopsis*
918 CESA3 catalytic domain with its substrate UDP-glucose provides insight into the mechanism of
919 cellulose synthesis. Proc Natl Acad Sci U S A. 2021;118. doi:10.1073/pnas.2024015118

920 77. Mertz RA, Ellsworth P, Ellsworth P, Lori Tausta S, von Caemmerer S, Howard Berg R, et al. The
921 maize Aliphatic Suberin Feruloyl Transferase genes affect leaf water movement but are
922 dispensable for bundle sheath CO2 concentration. bioRxiv. 2020. p. 2020.09.30.321265.
923 doi:10.1101/2020.09.30.321265

924 78. Birchler JA, Yang H. The multiple fates of gene duplications: Deletion, hypofunctionalization,
925 subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation.
926 Plant Cell. 2022;34: 2466–2474.

927 79. Gout J-F, Lynch M. Maintenance and Loss of Duplicated Genes by Dosage Subfunctionalization.
928 Mol Biol Evol. 2015;32: 2141–2148.

929 80. Schnable JC, Springer NM, Freeling M. Differentiation of the maize subgenomes by genome
930 dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci U S A. 2011;108: 4069–
931 4074.

932 81. Edger PP, Pires JC. Gene and genome duplications: the impact of dosage-sensitivity on the fate
933 of nuclear genes. Chromosome Res. 2009;17: 699–717.

934 82. Panchy N, Lehti-Shiu M, Shiu S-H. Evolution of Gene Duplication in Plants. Plant Physiol.

935 2016;171: 2294–2316.

936 83. Thomas BC, Pedersen B, Freeling M. Following tetraploidy in an Arabidopsis ancestor, genes
937 were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive
938 genes. Genome Res. 2006;16: 934–946.

939 84. Murat F, Van de Peer Y, Salse J. Decoding plant and animal genome plasticity from differential
940 paleo-evolutionary patterns and processes. Genome Biol Evol. 2012;4: 917–928.

941 85. Baxter I. Should we treat the ionome as a combination of individual elements, or should we be
942 deriving novel combined traits? J Exp Bot. 2015;66: 2127–2131.

943 86. Baxter I. Ionomics: studying the social network of mineral nutrients. Curr Opin Plant Biol.
944 2009;12: 381–386.

945 87. Togninalli M, Seren Ü, Meng D, Fitz J, Nordborg M, Weigel D, et al. The AraGWAS Catalog: a
946 curated and standardized Arabidopsis thaliana GWAS catalog. Nucleic Acids Res. 2018;46:
947 D1150–D1156.

948 88. Zhao H, Yao W, Ouyang Y, Yang W, Wang G, Lian X, et al. RiceVarMap: a comprehensive
949 database of rice genomic variations. Nucleic Acids Res. 2015;43: D1018–22.

950 89. Tian D, Wang P, Tang B, Teng X, Li C, Liu X, et al. GWAS Atlas: a curated resource of genome-
951 wide variant-trait associations in plants and animals. Nucleic Acids Res. 2020;48: D927–D932.

952 90. Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, et al. Genome-wide insertional
953 mutagenesis of Arabidopsis thaliana. Science. 2003;301: 653–657.

954 91. Addo-Quaye C, Tuinstra M, Carraro N, Weil C, Dilkes BP. Whole-Genome Sequence Accuracy Is
955 Improved by Replication in a Population of Mutagenized Sorghum. G3 . 2018;8: 1079–1094.

956 92. McCarty DR, Settles AM, Suzuki M, Tan BC, Latshaw S, Porch T, et al. Steady-state transposon

mutagenesis in inbred maize. Plant J. 2005;44: 52–61.

93. Ziegler G, Terauchi A, Becker A, Armstrong P, Hudson K, Baxter I. Ionomic screening of field-grown soybean identifies mutants with altered seed elemental composition. Plant Genome. 2013;6: lantgenome2012.07.0012.

94. Wimmer V, Albrecht T, Auinger H-J, Schön C-C. synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics. 2012;28: 2086–2087.

95. Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. Biometrika. 2008;95: 759–771.

96. Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet. 2012;44: 825–830.

97. mlmm.gwas: Pipeline for GWAS Using MLMM. In: Comprehensive R Archive Network (CRAN) [Internet]. [cited 23 Mar 2023]. Available: https://cran.r-project.org/web/packages/mlmm.gwas/index.html

98. Warnes G, Warnes MG. The genetics package. Locus. 2007;26: 1.

99. Barria A, López ME, Yoshida G, Carvalheiro R, Lhorente JP, Yáñez JM. Population Genomic Structure and Genome-Wide Linkage Disequilibrium in Farmed Atlantic Salmon (Salmo salar L.) Using Dense SNP Genotypes. Front Genet. 2018;9: 649.

100. Krishnakumar V, Hanlon MR, Contrino S, Ferlanti ES, Karamycheva S, Kim M, et al. Araport: the Arabidopsis information portal. Nucleic Acids Res. 2015;43: D1003–9.

Supplemental Files

S1 Fig. Orthologs in groups with 3/5 species representation. Error bars represent the 5th/95th percentiles of the permutations.

S2 Fig. Orthologs in groups with 4/5 species representation. Error bars represent the 5th/95th percentiles of the permutations.

S3 Fig. Distribution of list sizes in the inferred KIG permutations depending on ortholog tables. The old distribution of permutations' list sizes was from using OrthoFinder to call ortholog groups only between the species in the inferred KIG. The new distribution of permutations' list sizes shows the list size when *Liriodendron tulipifera* is added to our OrthoFinder run as an outgroup to our target species. The green dot represents the number of inferred KIGs in the actual list.

S4 Fig. Inferred KIG and its permutations compared to GWAS loci. Solid, vertical lines represent the percentage of GWAS loci explained by a gene from OrthoFinder inferred KIG lists. Kernel Density Estimate (KDE) plots display the distribution of 1000 random permutations with the 5th/95th percentiles shown as dashed lines. Inferred KIG list permutations were made from the orthologs of a list of random genes equal to the species and elemental composition of the primary KIG list (see Random permutations methods section).

S5 Fig. Distribution of the number of orthologous candidates returned per loci.

S1 Table. The current primary KIG list. Genes with functional characterizations from primary literature sources that are linked to changes in the ionome.

1002 **S2 Table. Updated KIG list with OrthoFinder inferred orthologs.** *T. aestivum* (wheat) is not included
1003 in this version of the inferred KIG list.
1004
1005 **S3 Table. False discovery rates of the genes and loci overlaps in Fig 1.** False discovery rates are
1006 calculated by taking the average number of genes/loci in random permutation overlaps and dividing
1007 by the number of genes/loci in the actual GWAS datasets.
1008
1009 **S4 Table. KIGs overlapping GWAS loci.** KIGs compared to all elemental uptake trait loci regardless of
1010 KIG list citations.
1011
1012 **S5 Table. KIGs in FiReMAGE's candidate lists.**
1013
1014 **S6 Table. False discovery rate table for gene and loci overlaps in S4 Fig.** False discovery rates are
1015 calculated by taking the average number of genes/loci in random permutation overlaps and dividing
1016 by the number of genes/loci in the actual GWAS datasets.
1017
1018 **S7 Table. The complete FiReMAGE priority lists for each trait and species.**
1019
1020 **S8 Table. Top 10% of all FiReMAGE candidates.** FiReMAGE candidates were sorted by their
1021 permutation calculated false discovery rates, and taking the top 10% with the lowest FDR.
1022
1023 **S9 Table. FiReMAGE candidates in Arabidopsis with no previously known biological processes**
1024 **function.** We used ThaleMine [100] to analyze Arabidopsis candidates for each trait to identify genes
1025 with unknown function.
1026
1027 **S10 Table. Soybean SNPs associated with ionomic traits.** SNP associations generated with methods

1028 outlined in “Soybean GWAS and 2015-2017 field experiments”.