A comparative approach for selecting orthologous candidate genes 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 causative genes is still challenging. In genome-wide association studies (GWAS) measuring elemental accumulation (ionomic) traits, only 5% of loci overlap genes known to control the ionome - indicating that many causal genes are still unknown. To

identify candidates for the remaining 95% loci, we developed a method that uses GWAS studies across multiple species to identify conserved causative genes. By Filtering the Results of Multi-species, Analogous, GWAS Experiments (FiREMAGE) we were able to take the GWAS of 19 ionomic traits in Arabidopsis, soybean, rice, maize, and sorghum, and identify alleles affecting trait variation at conserved genes. Permutation testing demonstrated that GWAS loci affecting the same trait contained homologs more often than expected by chance. Most of the top 10% most significant conserved candidate sets encoded alleles in all five species, highlighting the conservation of ionomic genetic regulators in flowering plants. The candidates include proteins with known biochemical functions that regulate the ionome, validating the approach. In addition to genes with known functions, this approach also identified many conserved genes underlying GWAS loci affecting the same trait in multiple species that have no previously identified function in regulating the ionome, providing a path to discover the as yet unknown mechanisms of element accumulation in plants. This method enables the identification of conserved genes of previously unknown function via GWAS.

Author summary

Quantitative genetics identifies a genomic region of interest but not the causal gene. We developed an approach to narrow these gene lists using genetic loci affecting elemental (i.e., calcium, iron, zinc) accumulation. Comparative genomics and GWAS demonstrates that alleles at evolutionarily conserved genes alter the same phenotype in multiple species. This produced a list of conserved candidate genes including previously known elemental regulators and genes whose elemental accumulation mechanism has yet to be determined. Combining datasets across species boosted the signal at these loci. This approach accelerates the discovery of new functional roles of genes.

Introduction

High-throughput phenotyping of elemental accumulation (ionomics) paired with genome-wide association studies (GWAS) has identified many genetic loci associated with variation in the accumulation of elements in plants [1–5]. However, identifying the causal gene among the group of genes linked to the phenotypically-relevant variants, often called a GWAS locus, is challenging for any trait or organism. Typically, many genes are linked to a GWAS locus. Thus each GWAS results in long lists of genes that are equally-weighted candidates for affecting variation in the phenotype. This is especially true for species with large linkage disequilibrium ranges. This has limited the molecular insights provided by studies of natural variation. Complex loci, multiple alleles, rare alleles, incomplete variant information due to a diversity of coding capacity across the species (aka pangenome variation), and sparse marker data further complicate the assignment of causality and can result in the most statistically significant variant being encoded at a position displaced from the causal gene [6]. Additional, and complementary, information about allele consequences and gene function are necessary to identify the molecular cause of detected quantitative genetic variation. Often this relies on labor- and resource-intensive experiments using induced alleles and single-gene testing to provide information of gene functions at candidate loci. Bridging this trait-to-gene gap is necessary to deliver molecular mechanistic insights from studies of natural variation and understand the molecular basis of diversity and adaptation.

Current gene selection methods include examining the functional annotations of genes in a region. Gene ontologies (GO) and annotations derive from our current understanding of known (or predicted) gene functions but are prone to human biases [7,8] and oversights from automatic computational algorithms [9]. GO terms are largely uninformative for novel or under-described traits, including some elemental accumulation phenotypes. GO terms require ongoing curation and submission of newly discovered gene functions from researchers to maintain relevance and utility.

80 Known genes linked to an association are more likely to be selected for further studies [7,10] which
81 produces a form of positive publication bias that reduces the discovery of new gene functions.
82 Additionally, these methods are most effective in species with the most annotations, like *Arabidopsis*
83 *thaliana* (Arabidopsis), with well-annotated phenotypes. Exploration of candidate genes in a region
84 is less common, and typically results when GWAS identifies a locus with large trait effects.
85 Several approaches have been used to identify causal genes from genes of unknown function.
86 Guided methods that can work without gene annotations include machine learning techniques like
87 PILGRM [15], GenAMap [16], network analysis, gene expression experiments, or a combination of
88 techniques [17]. These methods rely on extensive computation and resources like gene expression
89 data. As a result, they are most valuable in model systems with large collections of preexisting data.
90 Model systems like Arabidopsis also benefit from short generation times and small growth
91 requirements enabling complex and multigenerational designs [18].
92 Distinguishing likely causative genes among many linked candidates without relying on existing
93 annotations, ontologies, or preexisting data is significantly more complex and time consuming. For
94 example, linkage mapping in experimental line cross populations has been used to differentiate
95 causal candidates following trait-associated GWAS [11,12]. Fine mapping to find a causal gene is
96 expensive and laborious and is not feasible for alleles with moderate or small effect sizes [13,14].
97 Candidate gene testing by phenotyping knockout and overexpression mutants is an alternative
98 approach and the identification of additional alleles with impacts on the same phenotype is seen as
99 evidence of the correct specification of causal loci. This is facilitated by large collections of mutants,
100 only available for the best resourced model systems such as Arabidopsis and *Zea mays* (maize) with
101 substantial mutant resources. Gene editing has made these methods less dependent on pre-existing
102 populations and expanded the possibilities of candidate gene testing outside of traditional model
103 systems. The facility of these experiments is still dependent on generation time and tissue culture,
104 transformation, and plant regeneration. This makes the identification and prioritization of short lists
105 of candidate genes essential for feasibly testing candidates.

106

107 Developments in whole genome and next-generation sequencing have improved *de novo* genome
108 assembly of complex genomes and the identification of genetic markers (i.e., SNPs, indels) within
109 populations [19]. This makes it conceivable that comparison of alleles discovered in multiple species
110 might form the basis of validation experiments. When independent alleles in homologs have similar
111 phenotypic impacts, this is functionally equivalent to annotation of function by inference from
112 sequence similarity. Many agriculturally important crop species have excellent genome assemblies
113 and annotations. Arabidopsis still has the highest quality plant genome, annotations, and resources
114 that inform annotations in other organisms. For some agricultural species (e.g., soybean), gene
115 annotations are inferred directly by homology searches against Arabidopsis. Seeking gene functions
116 using resources beyond arabidopsis can dramatically enhance our understanding of plant biology,
117 including the biology of crops.

118

119 The ionome describes the composition of elements in plant tissue, all of which is acquired from its
120 environment [20]. Obtaining, excluding, and homeostatically maintaining the concentration of
121 elements in cells are functions shared by organisms across all domains of life. Elements obtained
122 from the environment are critical cofactors and substrates for multiple biochemical reactions, and
123 solving the challenges of balancing the ionome in cells was a necessary step in the evolution of
124 cellular life. As a result, many of the mechanisms that achieve elemental homeostasis are ancient,
125 and experimental approaches for mechanistic validation, such as cross-kingdom and cross-domain
126 molecular complementation (e.g., complementation of bacterial or yeast mutants with plant genes),
127 are highly effective [21,22]. Moreover, as elements are the same regardless of which species they
128 are measured in, we can directly compare elemental accumulation traits across species. As a result,
129 the field has repeatedly demonstrated that natural alleles at homologous genes affect the levels of
130 the same elements in multiple plants [22–31]. This independent evolution of alleles with similar
131 phenotypic consequences in multiple species is the hallmark of conserved critical regulators. The

direct comparability of ionomics traits, repeated observation of natural variation in homologs affecting phenotypic variation, and the deep conservation of mechanisms affecting its variation suggests that we could use homology to identify the causative genes for ionomic trait variation within GWAS loci across multiple species.

We created a computational method to take the genes linked to GWAS data sets from multiple species and detect orthologous genes present among the linked candidates. To assess the statistical significance of the homology overlaps, and determine if orthology occurred more often than expected by chance, we built a permutation-based assessment of the effectiveness of this tool that calculates the expected background distribution of orthologs contained in a set of genomic regions drawn from multiple species. Comparison of permuted sets to observed orthology overlap demonstrated that GWAS detects alleles at orthologous genes more often than expected by chance. This is the expectation if variation in elemental accumulation is accomplished by conserved genes across flowering plants. We validated these results using lists of genes of known ionomic function [26] and demonstrated that our combination of GWAS and orthology detected alleles at genes known to affect the ionome at a higher rate than expected by chance. However, genes of known function were overwhelmingly in the minority and it identified as likely candidates many genes with no previously known functions. This approach, which we call Filtering Results of Multi-species, Analogous, GWAS Experiments (FiREMAGE), identifies candidate genes without regard for prior evidence of function in element accumulation. This frees the process of candidate selection from the constraint of prior demonstration of function and permits gene function discovery through the study of natural variation directly. This approach can be applied at any dataset and encourages exploration of novel gene functions affecting trait natural variation.

Methods

Summary of Approach

We created a pipeline to compare GWAS results across species and determine if orthologous candidate genes underlie loci of phenotypic impact. The first step is processing of GWAS results to generate genomic intervals affecting the same trait in each member of a set of species. The genes within these intervals are tabulated and the orthologs of those genes in each member of the species set are determined. Permuted genomic regions, matched to each species' input, are generated and processed in the same manner as the experimental data. Comparison of the background distribution of ortholog overlap, from the permutations, and the experimental observations are made for each set of species and for every trait.

Briefly:

1. Following a GWAS in species 1, regions containing linked SNPs affecting trait variation are identified and retained as windows. The trait's top associated SNPs in each retained genomic window are identified and used to uniquely name the GWAS locus and summarize the significance and effect direction of the quantitative trait locus.

2. This is repeated for every species. At this point a set of genomic regions, linked to GWAS hits is established for the trait across all species.

3. Candidate genes are identified as encoded within these experimentally determined windows. At this point, overlaps between the candidates and known ionomic genes can take place (this is

separate from FIREMAGE and is a standard approach with GWAS made highly effective by the existence of a curated list of known ionomic genes).

4. Orthology relationships for all candidates from step 3 are determined across the five species using Orthofinder. At this point, the number of GWAS loci containing orthologous candidate genes were determined. These orthologous genes are returned as FIREMAGE candidates. This is the first step that is unique to our FiREMAGE approach.

5. FiREMAGE assesses the statistical significance of the orthology overlaps using permutation testing. 1000 sets of genomic regions with size-matched windows, from these five genomes, are selected at random. The rate at which these randomly permuted regions return overlapping orthologs is then used as the expected distribution of orthology overlaps for each trait given the observed number and size of genomic regions across all five species. This is a permutation test of the expected number of overlaps for windowed regions of the sizes returned from the GWAS for each species.

6. The False Discovery Rate of GWAS candidate gene ortholog overlaps is determined by comparing the true number of orthologous genes among the genomic regions affecting a trait across these species (Step 4) to the number determined for a permuted set of genomic regions across these species (Step 5).

7. This pipeline is repeated for every trait.

Data availability

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). R scripts created for this newest version of the Known Ionome Gene (KIG) List used to validate FiReMAGE results can be found at https://github.com/danforthcenter/KIG_v2.

GWAS datasets

Single nucleotide polymorphism (SNP) filtering

We used published elemental profiling GWAS datasets from *Arabidopsis thaliana* (Arabidopsis) [32], *Oryza sativa* (rice) [33–36], *Sorghum bicolor* (sorghum) [1], and *Zea mays* (maize) [3]. The *Glycine max* (soybean) data is described below. The p-value thresholds used are intentionally permissive to reduce false negatives in our GWAS/ortholog overlaps. False positive control is provided by requiring orthologous genes to exceed multiple thresholds and estimated by permutation, see Random Permutation, below. 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; $p < 1e\text{-}5$), *Sorghum bicolor* (sorghum; $p < 1e\text{-}4$), and *Zea mays* (maize; $p < 0.01$).

The fundamental units that FIREMAGE uses are genomic ranges from each species corresponding to each locus identified by GWAS. To provide SNP-to-locus identification, and delimit the genomic range corresponding to each locus, SNPs from each dataset were collapsed into loci using linkage ranges from each species (Table 1). The SNP with the lowest p-value (e.g. the most significant or “top SNP”) for each range was used to name and annotate each locus. The p-value of the top SNP was

224 retained as the collapsed loci's p-value. This SNP-to-locus mapping procedure results in no
225 overlapping ranges and avoids any problems associated with genes occupying more than one
226 overlapping locus. All references to "loci" provided to FIREMAGE refer to these genomic intervals
227 and the use of the word "locus" and not "gene" is deliberate as these locations can comprise
228 multiple genes. In addition to GWAS SNP-to-locus mapping, the FIREMAGE pipeline can accept any
229 file that describes loci as genomic intervals. As a result, it is already constructed to accept the
230 outputs of procedures such as joint linkage analyses in nested-association mapping populations such
231 as the maize NAM dataset.
232
233

Table 1. Overview of GWAS datasets.

Species Annotation	general linkage range	input as loci	SNPs/loci	collapsed 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

234
235 **Soybean GWAS and 2015-2017 field experiments**
236
237 *Glycine max* genotypes were obtained from the USDA-ARS GRIN collection (https://npgsweb.ars-
238 grin.gov/gringlobal/). Field experiments were performed over three years (2015, 2016, and 2017)
239 with three field replicates at each field experiment in a random complete block design. The 2015 and
240 2016 field experiments were performed at the University of Missouri Bradford Research farm
241 located near Columbia, Missouri, (38°53'38.4"N 92°11'47.1" W) which has a Mexico silt loam (fine,
242 smectitic, and mesic Vertic Epiaqualf) and the 2017 field experiment was performed at the Rollins
243 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 [37]. 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 [38] 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 [39] and the multiple-Bonferroni criterion [40]. A third set of SNPs 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) [41]. A resulting SNP was said to overlap with a neighboring SNP if the r^{-2} linkage disequilibrium was > 0.2 as calculated using the LD function in the genetics R package [42] 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 [43].

OrthoFinder orthologs

OrthoFinder v2.0 [44] was downloaded from https://github.com/davidemms/OrthoFinder and used with default parameters and best practices outlined here: https://davidemms.github.io/orthofinder_tutorials/orthofinder-best-practices.html. Protein reads of

294 primary transcripts were obtained from Phytozome (v13) [45]. To make the Known Ionome Gene
295 (KIG) list, we used files for all species in the list except for wheat due to its low-quality genome
296 annotation. *Liriodendron tulipifera* was included as the most closely related outgroup to the eudicots
297 and monocots in Phytozome v13. The resulting ortholog table included one-to-one, one-to-many,
298 and many-to-many ortholog relationships.

299 Loci/gene overlaps

300 We obtained all the Phytozome (v13) [45] gene coordinates for each species via the BioMart tool
301 (v0.7) located on the Phytozome homepage (https://phytozome-next.jgi.doe.gov). Using the
302 coordinates, we search for genes overlapping trait loci that are also in our 5-species ortholog groups.
303 For each trait, FiReMAGE retains genes from ortholog groups with overlaps in at least 3/5 species
304 and produces a list of prioritized candidate genes. The number of overlapping orthologs is recorded
305 for each locus to calculate false discovery rates.

306 Random permutations

307 GWAS loci permutations were 1000 selections of random locations equal to the actual datasets'
308 species and trait composition. Each species' chromosome start and end coordinates were
309 downloaded from the European Nucleotide Archive (ENA) database (GCA_000001735,
310 GCA_000004515, GCA_001433935, GCA_000003195, GCA_000005005). Because random base pair
311 locations are unlikely to be in range for collapsing into "loci", we assigned range lengths randomly
312 chosen from the actual GWAS loci. If random locations are close together, their randomly assigned
313 ranges may overlap. FiReMAGE runs the random permutations through the same loci/gene overlap
314 pipeline as the actual data, calculates the permutation's distributions to evaluate the likelihood of a
315 locus overlapping any nearby ortholog, and produces a false discovery rate (FDR) to prioritize

316 candidate genes. FDR was calculated as the mean number of orthologs overlapping random
317 locations divided by the number of actual orthologs overlapping loci.

318 Permutations were also carried out to assess the significance of the overlap with the KIG. Primary
319 KIG list permutations (S3 and S4 Figs, S6 Table) were formed by selecting random genes to replicate
320 the same species and element compositions of the actual primary KIG list and repeating 1000 times.
321 The extended KIG list permutations were formed by putting the primary genes permutations
322 through the same ortholog-finding process as the actual KIG list, with one deviation. In the actual
323 extended lists, a primary KIG's paralogs (if any) and orthologs are sourced from its ortholog group. In
324 the permutations, we only take the random gene's orthologs from its ortholog group. After
325 tabulating all orthologs of each set of randomly selected genes in each permutation iteration, we
326 add a random selection of genes from Arabidopsis, rice, and maize that equal the number of
327 paralogs in their actual extended KIG lists. The number of paralogs added to the permutation sets
328 equalled the number of paralogs that exist in the extended KIG list, and does not expand groups
329 where multiple primary genes are also paralogs to each other. This choice was made to avoid
330 random gene selections in which large, single-species, expansions of a gene family filled up the
331 permutation set as this would not correctly replicate the structure of the KIG list and would not do a
332 good job of estimating the background rate of overlap. Known ionomic genes are largely conserved
333 across species, only about 9% of the primary KIG list in the first version of the KIG list were single-
334 species genes [26].

335 Results

336 Known Ionomic Genes as GWAS candidates

337 Elemental accumulation, or ionomic, datasets from association mapping populations of five species
338 (Table 2): Arabidopsis [32], *Glycine max* (soybean; see methods), *Oryza sativa* (rice) [33–36],

339 *Sorghum bicolor* (sorghum) [1], and maize [3] were reanalyzed to identify alleles at conserved
340 orthologous genes. These GWAS data sets were selected because they are publicly available and
341 contain the same 19 elemental accumulation traits, measured with the same type of instrument
342 (ICP-MS), and all measurements were obtained from seed samples. In all cases except Arabidopsis,
343 data sets were generated from plants grown in multiple environments. SNPs were filtered by p-value
344 thresholds specific to each dataset (Arabidopsis $p < 0.001$, soybean $p < 0.05$, rice $p < 1e-5$, sorghum p
345 $< 1e-4$, and maize $p < 0.01$) and, when possible, across multiple environments (Table 2). The p-value
346 thresholds varied for each species to reflect differences in statistical power due to different
347 experimental designs. Linkage ranges estimated from genome wide assessments for each population
348 were used to determine locus sizes and tabulate linked genes (Table 1).
349

Table 2. 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

350

351 To help identify candidate genes linked to elemental uptake loci in plants, we had previously
352 compiled a list of known ionomic genes (KIG) experimentally shown to alter elemental uptake in
353 plants [26]. As these genes were sourced from the primary literature, we called this our "primary"
354 KIG list. Since the original publication, we have added 19 new primary genes to the list (S1 Table;
355 https://github.com/danforthcenter/KIG_v2), for a total of 201 primary genes. Most primary genes
356 are in Arabidopsis (59%) andrice (33%) with a small number contributed from work in maize,
357 *Medicago truncatula*, and *Triticum aestivum* (wheat). Loci and gene overlaps were considered on a
358 trait-by-trait basis, i.e., Fe uptake GWAS loci were only compared to KIG genes listed for Fe. The
359 overlap between the identified GWAS loci and the KIG identified alleles at genes already known to
360 affect changes in the ionome (Fig 1). Permutations were used to determine the statistical
361 significance of these overlaps by selecting an equal number of random base pair locations across the
362 genome and assigning a "locus range" randomly selected from the actual loci. The primary KIG list
363 covered a low percentage of GWAS loci (<2%) but overlapped with ionomic loci more often than the
364 random locations (80th percentile in Arabidopsis, 95th in rice, >99th in maize).

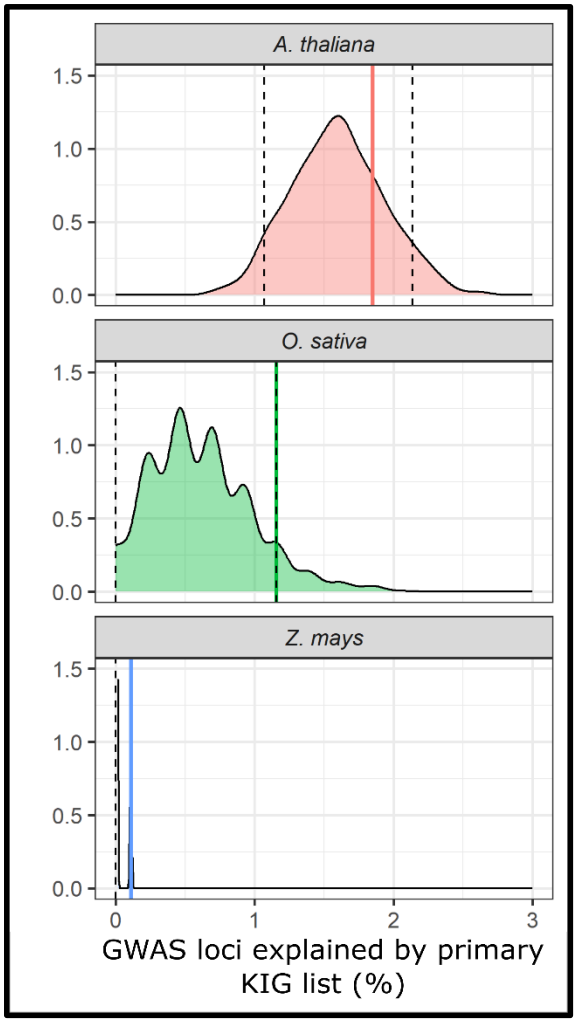


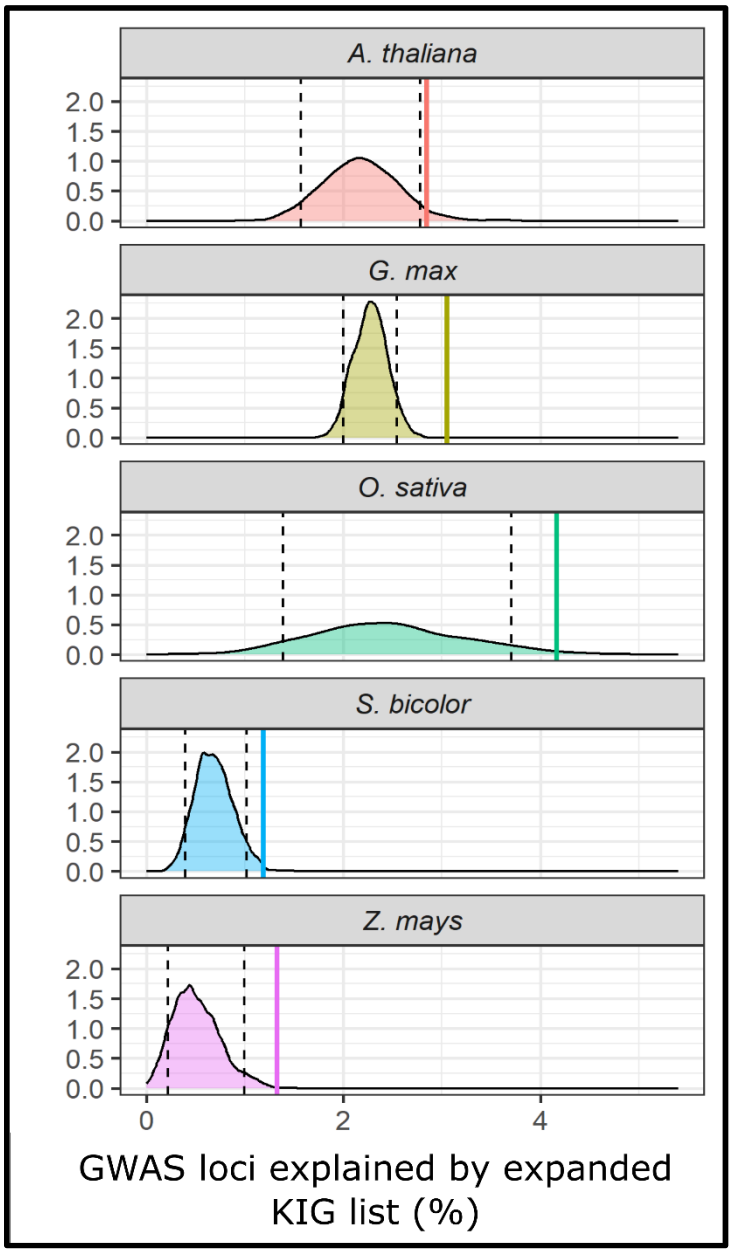
Fig 1. Primary KIG list coverage of ionomic GWAS loci. Solid, vertical lines represent the percentage of GWAS loci explained by a gene from the primary KIG list. Kernel Density Estimate (KDE) plots display the distribution of 1000 random permutations with the 5th/95th percentiles shown as dashed lines. 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?

KIGs overlapped the ionomic GWAS loci more often than random genome locations but provided no insight regarding likely candidate genes at 98% of GWAS loci. We hypothesized that inferring orthologs across species could 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

376 in separate species, orthologs tend to retain similar functionality [46], making them good candidate
377 genes when their function is known in one species. Indeed, for many KIGs, orthologs have been
378 tested in multiple species and found to encode conserved functions affecting element accumulation
379 for S, Fe, Cd, Na, K, Rb, Mo, Ni, Cu, P, Zn, As, B, Mn, and aluminum (Al) [26]. For example, alleles at
380 orthologous genes encoding the *MOT1* transporter affect variation in the level of Mo in multiple
381 species [3,28,34,47] and the same has been observed at *HKT1* orthologs for Na [29,48,49]. Inferring
382 the orthologs of KIGs across all of our study species should allow us to discover alleles in genes
383 affecting the ionome, including genes that have not yet been tested in every species in our list.
384
385 Ortholog inference algorithms vary by their use of pairwise similarity scores, phylogenetic analysis,
386 and multiple sequence alignment - all of which produce variations in the ortholog groups. We
387 previously published an expanded KIG list [26] made using InParanoid [50,] which uses species-pair
388 comparisons, because the ortholog tables were readily available on Phytozome [45]. For this
389 analysis, we switched to OrthoFinder v2.0 [44] because it performed better than other ortholog
390 group inference methods in benchmarking tests as determined by Emms et al. (2019) and Emms et
391 al. (2020), and forms ortholog groups between 3 or more species. With the ortholog tables from
392 OrthoFinder, we produced “extended” KIG lists that identify orthologs of the primary KIGs in the
393 original species and any species with a sequenced genome, including the five we are studying here.
394 Expanding the list of taxa in the extended KIG list created soybean and sorghum gene lists so we
395 could repeat the KIG/GWAS overlap comparison in all five of our ionomic GWAS species.
396
397 The extended KIG lists improved loci coverage across all five species and resulted in significant
398 overlap between the observed GWAS loci and the KIG ortholog lists (>95th percentile) for all species
399 (Fig 2). As expected, this is true whether we randomized the loci (Figs 1 and 2) or randomized the
400 genes in the list (Supplemental Fig 5). While the likelihood that we observed an enrichment of true
401 positives was significant (Fig 2), per-gene-set assessment of false discovery rates (FDR) was relatively

402 high (Table 3), partly due to the large number of genes linked to GWAS loci. Our FDR calculation is a
403 ratio of expected/observed, where the expected average number of extended KIGs linked to random
404 locations from permuted datasets is divided by the observed number of KIGs linked to GWAS loci.
405 Thus, the overlaps between KIG and GWAS loci represent high-value candidates for alleles affecting
406 ionomic variation at conserved genes (Table 3; S6 Table).



407

408 **Fig 2. Extended KIG list coverage of ionomic GWAS loci.** Solid, vertical lines represent the
409 percentage of GWAS loci explained by a gene from OrthoFinder extended KIG lists. Kernel Density
410 Estimate (KDE) plots display the distribution of 1000 random permutations with the 5th/95th

411 percentiles shown as dashed lines. Permutations were made with random genome selections of the
412 same size and species composition as the actual GWAS datasets (Table 2).
413

Table 3. Summary of extended KIG list overlaps with GWAS loci.

				KIG/GWAS overlaps		KIG/random overlaps		False discovery rate (FDR)	
Species	GWAS loci	random locations	Extended 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

414
415 Among the extended KIGs that overlap GWAS loci are conserved genes known to regulate elemental
416 accumulation across these plant species (Table 4; S4 Table). Identification of alleles linked to
417 orthologous candidate genes via GWAS for the same element provides evidence that these genes
418 not only encode conserved determinants of elemental homeostasis but also indicate that
419 phenotypically relevant, standing variation exists in these genes that could be used to enhance plant
420 mineral nutrition (e.g., P, Fe, Zn, Mn, Mo), improved mineral nutrition for humans (e.g., Fe and Zn),
421 or increased food safety (e.g., Cd).
422

Table 4. KIG ortholog groups overlapping GWAS loci. KIGs were only compared to GWAS loci for traits they were cited for in the primary literature. KIG citations are listed in S1 Table.

Known ionomic gene family	Cited traits	GWAS trait	Orthologs
AtPEN3	Cd	Cd	*AT3G16340* *Glyma.07G015800* *LOC_Os01g52560, LOC_Os06g36090*
AtHMA2, AtHMA3, AtHMA4 *OsHMA3*	Zn Cd	Cd	*AtHMA4, AtHMA2, AtHMA3* *Glyma.09G055600* *OsHMA3* *Sobic.002G083000, Sobic.002G083100* *Zm00001d005189, Zm00001d005190*

AtZIF1, AtZIFL2 *OsZIFL12* *ZmYS3*	Cs Zn Fe	Fe	*AtZIFL2* *Glyma.06G313600, Glyma.06G313700* *Sobic.008G015700*
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*
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*
AtPHT1;1, AtPHT1;4, AtPHT1;5 *OsPT2*	P As Se	P	*AtPHT1;4* *Glyma.10G186400, Glyma.10G186500* *LOC_Os08g45000*

The KIG orthologs linked to GWAS loci include several transporters with known roles in element uptake, sequestration, and remobilization. Notably, alleles were discovered in orthologs of transporters in the zinc-induced major facilitator superfamily (*ZIFL1* and *ZIFL2*) genes that affect the accumulation of Fe in seeds (Table 4). Fe is a human nutrient that is limited in many diets [51–53]—making the discovery of these alleles high-priority targets for breeding efforts to improve nutrient accumulation in soybean and sorghum (Table 4). KIGs *MOT1* and *MOT2* encode molybdate transport proteins [28,54–57] 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

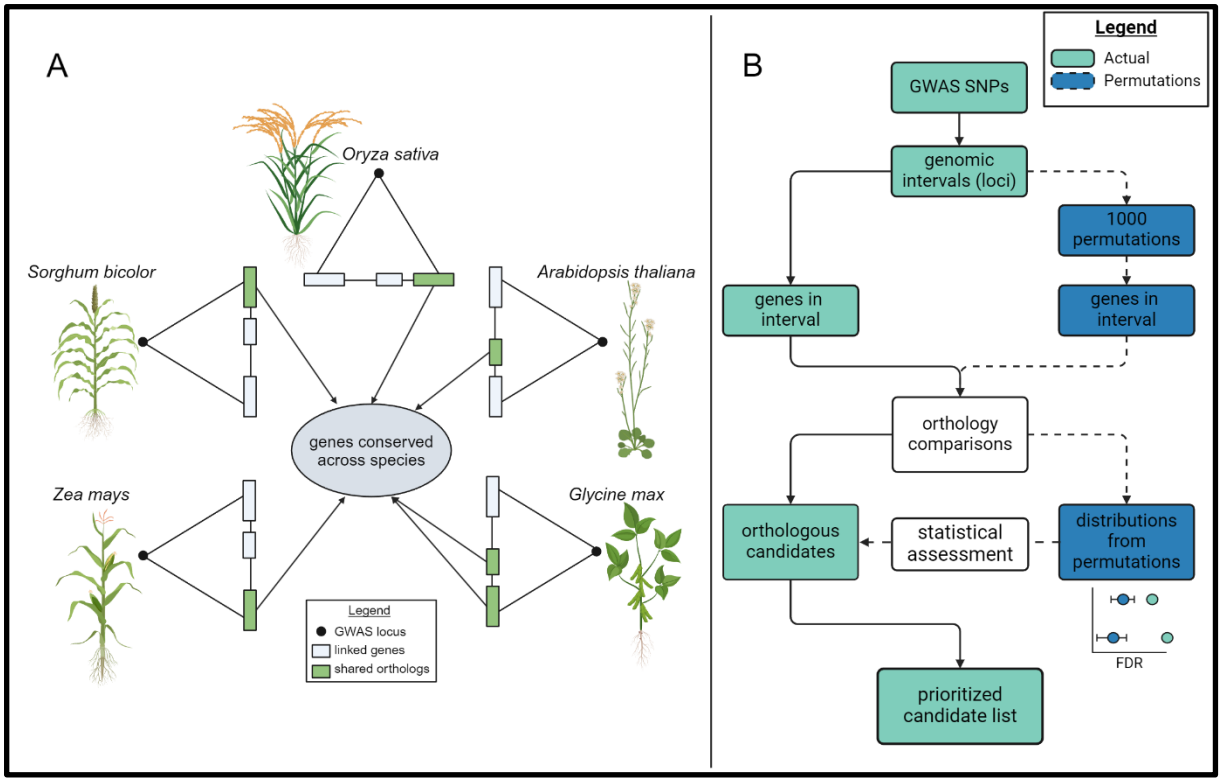
433 [11,58–60]. In addition to direct Cd transport, alleles at *PEN3/PDR8*, implicated in Cd transport via
434 chelation with the glutathione-derived phytochelatin [61], affected Cd in three species. The alleles in
435 these transporters and chelator biosynthetic enzymes represent immediate breeding targets for
436 improved human food safety. The phosphate transporter (*PHT1*) family has been shown to affect
437 phosphate and As accumulation in Arabidopsis [62–64] and Se uptake in rice [65]. Orthologs of this
438 transporter were linked to alleles that affect P accumulation in Arabidopsis, soybean, and rice. Thus,
439 these alleles represent possible targets to enhance nutrient uptake for this yield-limiting nutrient.
440 *MTP11* encodes a Mn transporter involved in Mn tolerance and detoxification with the Golgi
441 apparatus [66], and orthologs to *AtMTP11* are linked to Mn loci in four species. Alleles in the
442 orthologs of yet another transporter family, including *AtNRAMP3*, which is involved in Mn
443 homeostasis via the mesophyll vacuoles [67], were also associated with Mn accumulation in
444 Arabidopsis, soybean, sorghum, and maize.
445
446 Alleles at regulatory genes from conserved ortholog groups were also identified among multiple
447 species. These included alleles at orthologs of the transcription factor BRUTUS (*BTS*), which affects
448 the expression of genes affecting Fe accumulation in Arabidopsis, including the transporter *IRT1* and
449 the regulatory transcription factor *FIT1* [22,68,69]. Alleles linked to the *BTS* orthologs affected Fe
450 accumulation in soybean, sorghum, and maize (Table 4), demonstrating conservation of function for
451 this transcription factor family as regulators of iron accumulation response across species with
452 substantial variation in Fe transport mechanisms and biochemically distinct Fe chelation systems. As
453 these alleles affect the accumulation of Fe, a nutrient that limits human health, the alleles described
454 here are alleles of immediate value for breeding food staples with improved seed Fe and benefiting
455 global human nutrition. Our approach also detected alleles at a regulator of the ionome, the
456 *Enhanced Suberin 1* (*ESB1*) gene, a dirigent-like protein previously only known via a knockout mutant
457 in Arabidopsis which exhibits altered Casparian strip biosynthesis [70], Ca, Mn, Zn, Na, S, K, As, Se,
458 and Mo levels [71]. GWAS detected alleles at this gene affecting Mn in Arabidopsis, soybean, and

459 sorghum (Table 4), demonstrating that this gene performs conserved functions in regulating ion
460 levels across species and that disruption of the Casparian strip can have similar impacts on the
461 ionome across species with dramatically different root systems and root vascular anatomies.
462
463 This analysis demonstrated the utility of combining GWAS across species by using orthology and
464 conservation for genes of known functions. The number of GWAS loci overlapping a KIG ortholog (Fig
465 2, Table 3) was greater than the overlap with a primary KIG gene (Fig 1, S3 Table). However, this
466 overlap only captured less than 5% of the ionomic GWAS loci. To identify high-quality candidates for
467 the remainder of the GWAS results, we sought a formal procedure for assessing the likelihood of
468 GWAS results at orthologous loci across these five species regardless of previously demonstrated
469 functions.

470 Orthology as a key factor for candidate selection

471 We hypothesized that conservation of orthologs within GWAS loci across multiple species could be a
472 criterion to identify genes of previously unknown function. In the expanded KIG ortholog/GWAS
473 overlap comparison (Fig 2, Tables 3 and 4, S4 Table), we frequently detected phenotypically-
474 impactful alleles in multiple species' GWAS. GWAS is "forward" genetics and will also detect alleles
475 at genes of previously unknown functions. Therefore, ionomic GWAS datasets will detect alleles at
476 unknown regulators of elemental accumulation. Alleles at conserved but previously unknown
477 regulators of the ionome should also appear more than expected by chance in comparisons between
478 GWAS loci from multiple species.
479
480 To test this hypothesis, and identify conserved ionomic genes, we developed a method we call
481 FiReMAGE: Filtering Results of Multi-species, Analogous, GWAS Experiments. This approach
482 identifies the orthologous genes linked to GWAS associations in multiple species and compares their
483 occurrence to a background distribution to assess the false discovery rate. This increases confidence

484 in specific candidate genes linked to GWAS loci without relying on previously known functions.
485 FiReMAGE works by tabulating the genes within trait loci in each species, determining their
486 orthologous relationships across species via OrthoFinder, and retaining ortholog groups with
487 candidate hits from at least three species (Fig 3). A threshold of three or more species was used to
488 reduce the false positives due to the high synteny between some species pairs (e.g.. Sorghum and
489 maize). Like the permutations used with the KIG/GWAS comparisons, FiReMAGE creates 1000
490 permutations equivalent to the GWAS data by randomly selecting an equal number of genomic
491 locations and ranges from each species' genome. FiReMAGE runs the permutated sets through the
492 same orthology comparison as the actual data. By comparing the expected (permuted) and observed
493 (GWAS data) orthologs returned from these genomes, this background distribution evaluates the
494 significance of the observed overlaps and provides a false discovery rate. This is a more accurate null
495 distribution for comparison than determining the number orthologous genes in a set of randomly
496 selected genes. Perfectly matched permutation sets were calculated for each set of species and
497 traits allowing FiReMAGE to output FDR for every comparison set (e.g. error control for the 3/5 set
498 of genes associated with Fe in O. sativa from the rice, maize, Arabidopsis comparison). These can be
499 seen as likelihood weights for the candidates in every comparison and are arrived at by performing
500 the same FDR calculation: average number of GWAS-linked orthologs expected by permutations
501 divided by the observed GWAS-linked orthologs in the actual data.
502



503

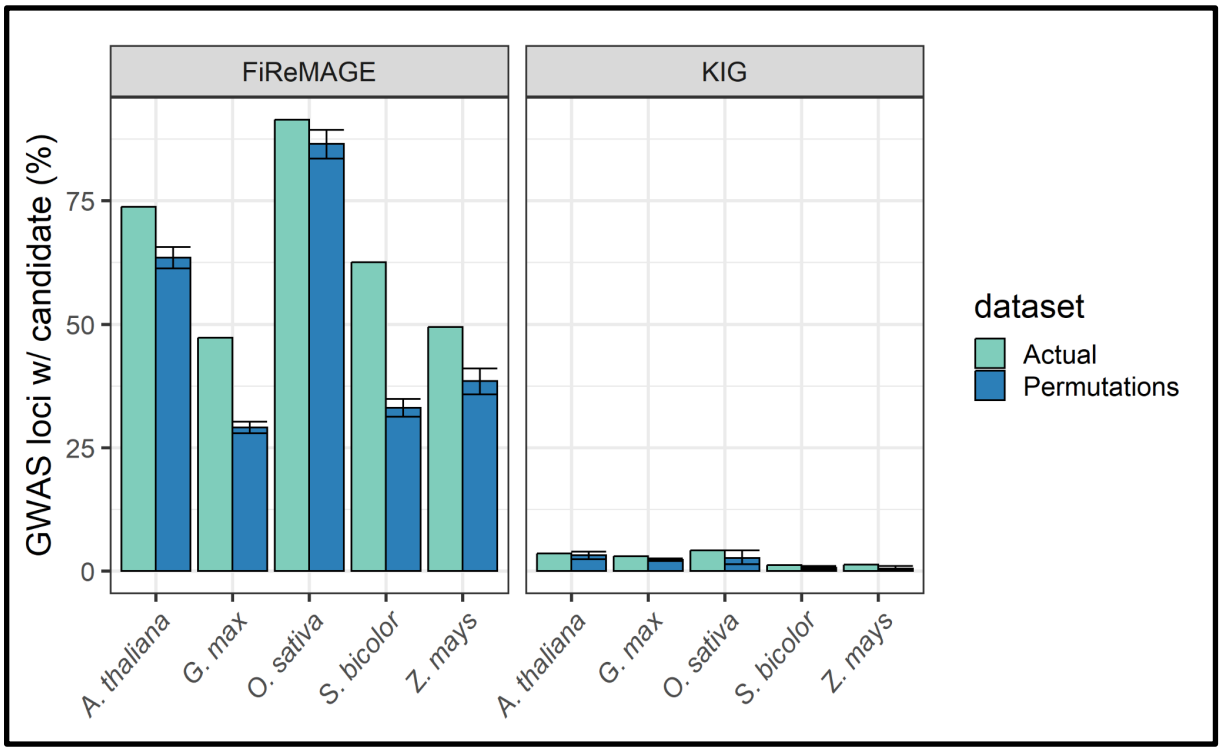
504 **Fig 3. Overview of the FiReMAGE pipeline.** (A) FiReMAGE orthology comparison identifies genes
505 linked to SNPs in each GWAS that are also orthologous to each other. (B) FiReMAGE starts by
506 identifying genes within range of GWAS loci intervals. These genes are recorded and filtered through
507 the orthology comparison to identify orthologous candidates. The original GWAS loci then are used
508 to generate 1000 random permutations, which go through the same process as the actual data.
509 Distributions from the permutations are used to assess genes in the real dataset and return a
510 prioritized candidate list. Created in BioRender (https://BioRender.com/c20l148).

511

512 With FiReMAGE, we identified candidates in ortholog groups with hits in at least three of the five
513 species for over 30% of soybean and maize loci and over 50% in Arabidopsis, rice, and sorghum (Fig
514 4). The orthologous loci identified from our GWAS datasets was greater than the 95th percentile of
515 the permutations in each species demonstrating enrichment of conserved genes underlying GWAS
516 hits. However, the individual candidate gene false discovery rates (the ratio of genes expected from
517 permutations/genes observed in the data) was still relatively high (Arabidopsis = 0.58, soybean =
518 0.43, rice = 0.62, sorghum = 0.39, maize 0.67). Across the 19 elemental traits for the 3 out of 5

519 species comparison, 3866 unique orthologs were linked to loci in Arabidopsis, 7852 in soybean, 2630

520 in rice, 2982 in sorghum, and 872 in maize (S5 Fig; S7 Table).



521

522 **Fig 4. FiReMAGE identifies candidates for a larger portion of GWAS loci than the KIG list.**

523 Percentage of ionomic GWAS intervals with a candidate from FiReMAGE's comparative approach

524 (left) and the OrthoFinder KIG list (right). The actual GWAS dataset is in green; 1000 random

525 permutations of the actual dataset are in blue, with error bars representing 5th/95th percentiles.

526

527 The size of candidate lists and false discovery rates vary between species and traits (Figs 5 and 6, S1

528 and S2 Figs). Requiring more species to be present in an ortholog group returned shorter candidate

529 lists (Fig 5) with lower false discovery rates (Fig 6). While still encompassing a large number of genes,

530 the 3/5 list is a substantial reduction in the number of potential candidates.

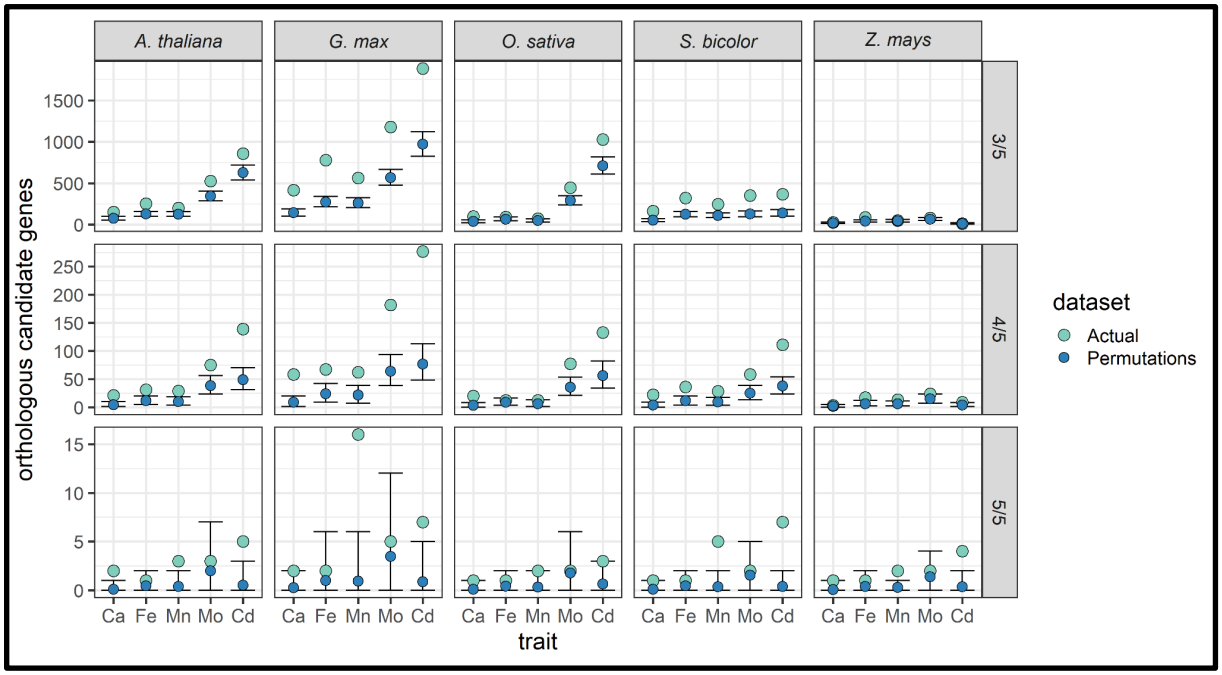


Fig 5. Orthologs from ortholog groups linked to Ca, Cd, Fe, Mn, or Mo loci in 3/5, 4/5, and 5/5 species. Actual dataset counts are in green, and 1000 permutations of the actual data are in blue, with the point representing the mean and error bars representing the 5th/95th percentiles. Columns are by 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.

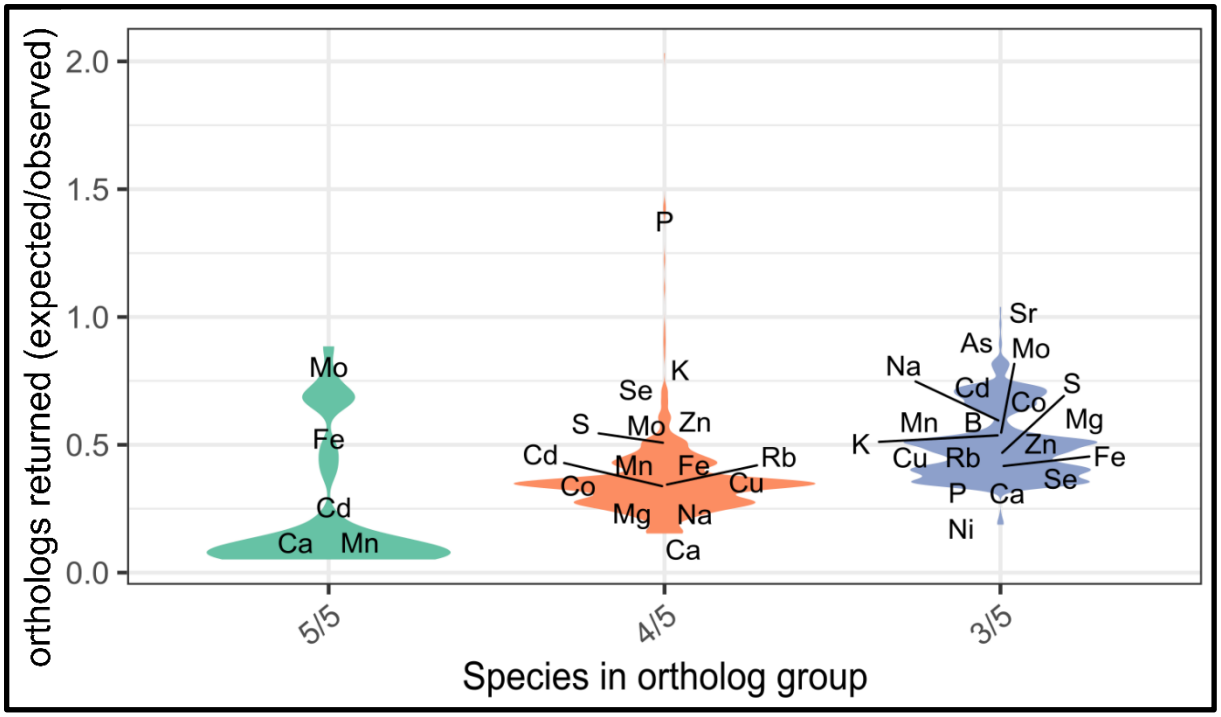


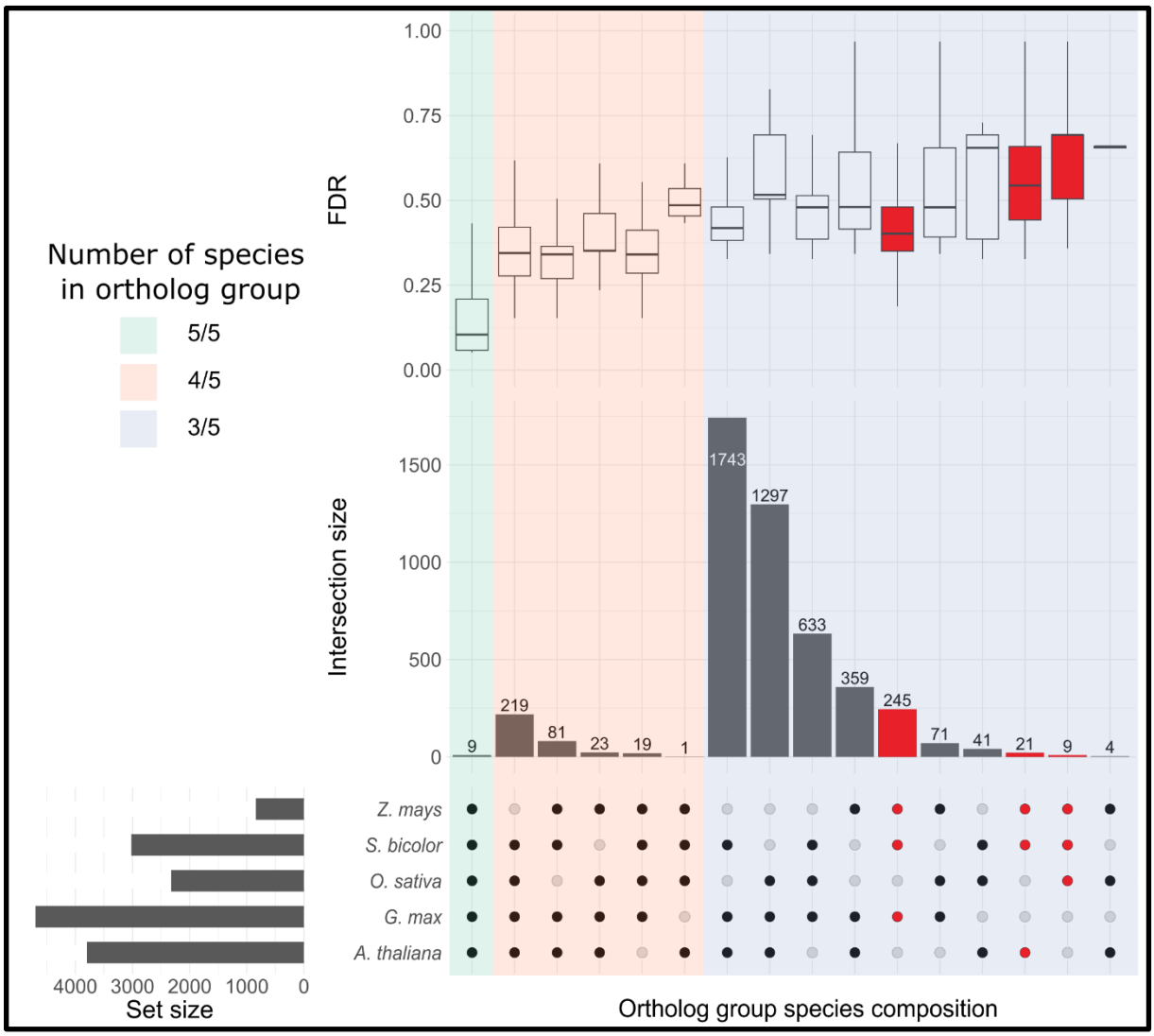
Fig 6. False discovery rate (FDR) of FiReMAGE's selected candidate genes from ortholog groups with representation from 3/5, 4/5, and all five species for each ionomic trait. FDR was calculated by taking the average number of orthologous candidates returned in the 1000 permutation datasets and dividing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

High synteny between two species results in many ortholog pairs (syntelogs) at each locus. We carefully examined the influence that the maize-sorghum comparison had on our FiReMAGE results as this was the most closely related species-pair in the data set. We expected the increased synteny between maize and sorghum to increase the number of orthologous genes found (highlighted red in Fig 7). This might increase the number of false positive (FDR) ortholog/GWAS overlaps and obscure the signal at true causative genes. This high synteny did not, however, dominate the results. GWAS overlap at ortholog groups from species with increased synteny do not make up the majority of the loci detected. The largest intersection with maize, sorghum, and third species (maize-sorghum-soybean) detected 245 ortholog groups, and was only the fifth largest intersection out of all ten ortholog group species combinations. Additionally, the FDR of these genes is comparable to 3/5

556 ortholog groups derived from species with less synteny, suggesting that this is not a major concern

557 for this experiment.

558



559

560 **Fig 7. UpSet plot of FiReMAGE's selected candidates' ortholog group composition.** The intersection

561 matrix shows the unique compositions of all ortholog groups determined by OrthoFinder. Ortholog

562 groups are ordered by the degree (number of species) of the intersection and then by group size.

563 Intersection size represents the number of ortholog groups matching the species composition below.

564 Set size shows the total number of instances in which a species appears in all ortholog groups. The

565 top panel displays the individual genes' false discovery rate (FDR) distribution in their respective

566 ortholog groups for all ionomic traits. Note that some genes may be represented twice in the boxplot

567 distribution if they appear for multiple traits. Thus, they may have multiple FDRs due to the different
568 SNP distributions and FDR of the ionomic traits. Because of the theory that high levels of synteny
569 between sorghum and maize may increase the number of false positives returned, ortholog groups
570 with two of the three species being sorghum and maize are highlighted red.

571 FiReMAGE identifies highly likely candidates and novel hypotheses
572 for the regulation of element accumulation

573 For each ionomic trait and species combination, FiReMAGE produced a list of orthologous
574 candidates and an FDR estimated by permutation (S7 Table). If elemental accumulation is controlled
575 by conserved genes, then the FiReMAGE procedure should identify more orthologous gene/loci
576 overlaps than expected by chance via the permutations. FDRs differed across the element
577 accumulation traits and species combinations. Generally, comparisons with low FDRs were observed
578 for elements with a previously demonstrated high heritability [72–76] and with GWAS hits in all five
579 species: Cd, Ca, Fe, Mn, and Mo (Fig 5). Of 23,117 FiReMAGE candidates, only 81 belong to these 5/5
580 ortholog groups. If we rank all candidate genes using the FDR for each gene set from the perfectly
581 matched permutation scores, candidates 1-56 and 115-116 are linked to GWAS loci at orthologous
582 candidates in all five species. The top 10% of candidate genes ranked in this way correspond to 233
583 genes from 18 element/species/comparison combinations (S8 Table). That a lower FDR was
584 observed when genes were present in more species demonstrates the value of using orthology, the
585 power of replication across the species, and the utility of our permutation-calculated FDR to rank
586 candidates.
587
588 Among the candidates from orthologous groups linked to loci in all five species are some of the most
589 significant GWAS SNPs in the input set. The GWAS loci linked to these 5/5 candidate genes (n = 64)
590 were enriched for the top 10% most significant SNP-trait associations from their respective GWAS

591 (Chi-square p-value > 1.25 e-10; obs: 21 True, 43 False; exp: 6 True, 58 False). Low p value loci within
592 an individual species' GWAS (n=21), might have been considered targets for candidate verification.
593 FiReMAGE improves upon this by identifying likely causative conserved genes at each of these
594 locations, accelerating the process of gene function discovery. In addition, FiReMAGE also outputs
595 the likely alleles in the species where the effects at these conserved genes were less strong. Thus, by
596 using orthology, FiReMAGE identified genes linked to GWAS SNPs with "middling" signal to identify
597 the remaining conserved candidates.
598
599 The identities of the genes encoded by the FIReMAGE candidates in our 5/5 species lists implicate a
600 wide variety of biological processes in elemental homeostasis. The 81 candidates with alleles in all
601 five species belong to nine orthologous groups. For Fe, among the 5/5 candidates with no previously
602 known role in Fe homeostasis, is an ortholog group that encode Mevalonate kinases across these
603 five species, suggesting a currently unknown role for the mevalonic acid pathway in regulating Fe
604 homeostasis in flowering plants. Among the top Cd candidates, we find the *HMA* ortholog group
605 mentioned above in the KIG list (Table 4), RAB GTPase homologs of *RABA1i* that are involved in
606 regulating cell wall composition [77], and an ortholog group of Domains of Unknown Function 567
607 (DUF567) protein family. In Arabidopsis, DUF567 is expressed in the roots and identified among a list
608 of 129 genes associated with metal hyperaccumulation in *Arabidopsis halleri* [78]. Thus, FireMAGE
609 provides additional experimental evidence, leveraged across five species, that this DUF567 family
610 affects ion homeostasis and may underlie ecological adaptation via hyperaccumulation in
611 Arabidopsis relatives. For Mo, the 5/5 candidates are the molybdate transporters, *MOT2*, listed in
612 the KIG list for Mo accumulation [54], and adenine nucleotide transporters (*NTT1* and *NTT2*).
613 Nucleotide metabolism is tied to Mo via molybdopterin, which is synthesized from nucleotide
614 precursors [79] to form reaction cofactors for four plant enzymes [80] suggesting a mechanism for
615 these associations. For Ca, the orthologous candidates are annotated as encoding a Ring/Ubox
616 superfamily with no known role in Ca accumulation. Lastly, for Mn, the candidate ortholog groups

617 are annotated as histone proteins and cellulose synthase A9 (*CESA9*). *CESA9* has been shown to
618 influence the cellulose content and cell wall biosynthesis in the epidermal seed coat in Arabidopsis
619 seeds [81]. In cellulose synthesis, Mn ions are used to catalyze glycosidic bond formation and CESA
620 protein coordinates Mn ions with the glycosyltransferase substrate uridine diphosphate glucose [82].
621 Given the large number of Mn-dependent processes in the cell, the association is unlikely the result
622 of alleles affecting Mn levels via the size of the cellulose synthase sink. It may be that Mn
623 homeostasis in the golgi and trans-golgi networks [83], are sensitive to the binding state for
624 coordinated Mn in CESA. Thus, highly likely candidates identified as conserved genes underlying
625 GWAS peaks in this species set have reasonable proposed mechanisms of action, are worthy of
626 further scientific investigation and, again, represent potential immediate-use targets for breeding
627 and crop improvement.
628
629 We also compared the FiReMAGE candidates to another candidate selection method performed in
630 maize: Camoco [17]. From the 47,000 candidate observations derived from multiple expression
631 network and elemental trait combinations, we selected the top 200 observations as ranked by the
632 FDR calculations in Camoco's multiple expression-network-based prioritizations and ionomics trait
633 analysis [17]. This selection contained 122 unique candidate genes to compare against FiReMAGE's
634 evolutionary-based prioritizations. Only two of these112 candidates were detected as conserved
635 regulators by FiReMAGE. One gene, *Zm00001d04461*, identified through both methods as affecting
636 Mo, encodes an F-box domain containing protein; the other, *Zm00001d012274*, affecting Fe
637 accumulation, is an enzyme encoding an Omega-hydroxypalmitate O-feruloyl transferase involved in
638 the production of suberinized hydrophobic barriers [84] that are known to influence ionomic traits.
639 Despite the clear mechanism implicated, the finding that only two genes affected variation detected
640 by the two methods indicates that single-species expression networks and orthology selection
641 methods boost different genetic signals to tie GWAS results to candidate genes.

642 **Discussion**

643 We have demonstrated that combining orthology and GWAS from multiple species detects
644 conserved regulators of phenotypic variation. Using the Known Ionomic Genes (KIG) as a validation
645 set of curated annotations confirmed our hypothesis that combining GWAS and orthology would
646 correctly identify variation affecting elemental accumulation. This resulted in the discovery of
647 phenotypically-impactful alleles affecting ionomic variation in genes of previously known functions
648 encoded in these five species. Additionally, the FiReMAGE approach went beyond the KIG list by
649 identifying numerous candidates in genes without previously determined roles in element
650 accumulation (S7 Table). This ability to identify and prioritize candidate genes for conserved roles in
651 trait variation using the natural variation of multiple species transcends one of the key limitations of
652 the post-genomic era: ascertainment bias due to prior study. Prior research bias has resulted in gene
653 discovery focused on a handful of elements due to researcher interests, leaving a majority of
654 elements under-studied [26]. Similar biases for prior demonstration of gene function, result in the
655 preferential publication of additional alleles at homologs of known function rather than the
656 exploration of natural alleles at locations with no genes of prior molecularly demonstrated effects on
657 the studied trait. The overlap of less than 5% between GWAS outputs and the KIG indicate that the
658 KIG list is vastly incomplete [8], and many more plant genes that regulate the ionome are waiting to
659 be discovered. The overlap between the FiReMAGE output and genes of previously known function
660 as well as the identification of orthologous high likelihood produced a comprehensive public list of
661 alleles at candidate conserved regulators of 19 elemental accumulation for future molecular
662 mechanistic studies or marker-assisted selection (S2 and S7 Tables).
663
664 Instead of selecting candidates with prior annotations, FiReMAGE annotates orthologous genes as
665 high-value candidate genes for further molecular function characterization. Experimental error, high
666 genome-wide statistical thresholds, and low statistical power in a GWAS results in high false
667 negative rates (e.g. insignificant p-values at a true causal gene). Meanwhile, every gene linked to
668 SNPs with a significant p-value will not be a true causal gene and assuming so would increase

669 false positives and result in spurious conclusions. The permutation results in Figs 1 and 2
670 demonstrate the power of collectively assessing GWAS datasets across species. When the same
671 gene is linked to GWAS SNPs in multiple species, these independent experiments provide
672 multiple lines of evidence for that gene's function and increase the likelihood that it is causal. T

673 Orthology-based FiReMAGE is highly effective for identifying gene 674 candidates

675 The integration of orthogroups into gene annotation provided improved estimates of gene function,
676 as demonstrated by the improved GWAS/KIG overlap when orthologs were included (Figs 1 and 2,
677 Table 3, S3 Table). As a result of this success, we hypothesized that orthology, even in the absence of
678 gene function annotations, would also identify conserved genes as candidates without any prior
679 demonstration of ionomic function. FiReMAGE identified more orthologous genes linked to GWAS
680 loci for the same trait than expected by chance (Figs 4 and 5), demonstrating that this approach
681 works. In addition to KIGs (Table 4), FiReMAGE also returned non-KIG ortholog groups that encoded
682 natural variation in elemental accumulation (S7 Table). These FiReMAGE candidates are likely to
683 encode conserved regulators of elemental accumulation, for which our analysis is the first genetic
684 demonstration of this role. Finding conserved critical regulators of ion homeostasis by GWAS
685 validates the expectation that conserved regulators (e.g., hub genes) will repeatedly accumulate
686 phenotypically effective variation. It also dramatically broadens the capacity of GWAS to work as a
687 gene function discovery tool in the forward-genetics mode. Thus, comparing element accumulation
688 association studies across organisms can help us discover additional element homeostasis
689 mechanisms. Our approach relies on both annotated genomes and GWAS data, but it should be
690 generalizable to any homologous traits across taxa. FiReMAGE does not require any information
691 other than a list of GWAS SNPs and gene positions in genomes, thus avoiding prior knowledge bias
692 and enabling accelerated gene function discovery for traits and species with little to no functional
693 annotations or prior work.

694

695 One of the non-KIG ortholog groups with alleles in all five species detected by FiReMAGE for Cd
696 encodes a domain of unknown function 567 protein (DUF567) family. This DUF567 is just one of the
697 198 elemental accumulation candidate genes identified by FiReMAGE in Arabidopsis that has no
698 previously known function in any biological processes (S9 Table). No other genes whose best
699 annotation is as domain of unknown function (DUF) containing gene are in 5/5 ortholog groups, but
700 19 are in the top 10% of Arabidopsis candidates for all traits, and 46 unknown genes are in the top
701 25% when sorted by their permutation estimated FDRs. None of these unknown genes would be
702 considered candidates if relying only on gene annotations. That these DUFs are within the gene sets
703 with the lowest permutation-estimated FDRs which also include known conserved regulators from
704 multiple species (Table 4) suggest these DUFs are conserved, previously unknown, regulators of the
705 ionome. Future single-gene mechanistic experiments may result in the annotation of these genes as
706 elemental regulators. Problems with annotating genes and Gene Ontology (GO) terms in general,
707 and plants in particular, have been extensively discussed [7–9]. FiReMAGE side-steps this problem,
708 as is illustrated by the observation that it finds 602 orthologous gene candidates for Mo and 116 for
709 Rb (S7 Table), yet there are no GO terms for the transport, accumulation, or metabolism of either
710 Mo or Rb.

711

712 Using data from multiple experiments, FiReMAGE emphasizes loci that would be passed over in a
713 single-species experiment. In many cases, FiReMAGE outputs strong signals when orthologs are
714 considered across taxa for loci with moderate effects in each species. Five traits, Ca, Fe, Mn, Mo, and
715 Cd (Fig 5, S7 Table), have ortholog groups with candidate gene hits in all five species. The federation
716 of data across multiple species implemented by FiReMAGE can boost the signal at genes encoding
717 alleles that were below the p-value thresholds typically employed for single-species GWAS
718 assessment. This is illustrated by the results located within GWAS loci at the ortholog groups
719 containing *MOT2* and the Cd/Zn transporting *HMAs* (Table 4). Some of the p-values in single species

720 comparisons are very significant, but not all. The worst performing taxa for these ortholog groups
721 were soybean for Mo accumulation at *MOT2* (p-value = 0.0104) and Arabidopsis for Cd accumulation
722 at the tandem gene duplication comprising AtHMA2, AtHMA3, and AtHMA4 (p-value = 5.6e-05).
723 Neither of these associations would meet a multiple testing corrected p-value cutoff for significance
724 if considered alone in their respective single species GWAS (S7 Table). Finding associations at these
725 loci in all five of the species demonstrates that variation at these loci is present. Indeed, AtHMA3 is
726 at a major effect QTL affecting the level of Cd in leaves at a p-value < 10e-20 [11]. The cause of this
727 association, and our detection of a smaller effect on seed Cd accumulation, is a complicated
728 collection of loss-of-function alleles of various frequencies encoded at this locus [11]. One null
729 mutant allele in HMA3 is present in the reference Col-0 accession [58]. This loss of function allele
730 induces sensitivity to metals, including Cd which it directs to the vacuole [85]. The dramatic effect of
731 these alleles on Cd accumulation in leaves and weaker effects in seeds likely results from a strong
732 effect of HMA3 on vacuolar sequestration, and a weaker contribution to Cd remobilization during
733 seed filling in Arabidopsis.
734
735 For those genes that are in the KIG orthologs set, these have been highlighted as genes affecting the
736 ionone independently via FiReMAGE, demonstrating phenotypically affective variation co-
737 segregating with identified SNPs, and via prior work on molecular genetic demonstration. These
738 alleles all represent utility-ready breeding targets that are the highest confidence hits in our study.
739 Of particular note are the alleles identified by FiReMAGE that overlap the KIG and alter the levels of
740 minerals known to limit human health. Because the GWAS data sets employed here are all studies of
741 seeds, all of the SNPs affecting variation in elements important for human nutrition can be
742 incorporated into breeding programs. For example, all Fe and Zn alleles at KIG can be put to use
743 immediately in public breeding seeking to increase the concentration of these critical human
744 nutrients to overcome health-limiting deficiencies. Similarly, the Cd and As associations identified in

745 these materials at KIG orthologs (Table S7) can be put to use immediately to reduce the levels of
746 these toxic elements in the food supply.

747 FiReMAGE approach and utility for gene discovery

748 Our FiReMAGE approach combines a wide array of information. It has many steps (functions) and
749 dozens to hundreds of user-defined parameters. Many of these lack obvious correct or best choices
750 for parameter values. These include the p-value thresholds for the GWAS input data sets, the linkage
751 range for SNP-to-gene relationships, and the outgroups for constructing ortholog relationships. In
752 addition, several variables are inherent in the data and experimentation that also affect FiReMAGE,
753 including differences in measurement error in each experiment, variation in the statistical power of
754 each GWAS panel, the total genetic variation in a species, and the association panel selection. By
755 using permutations instead of an idealized null distribution, FiReMAGE estimated how likely the
756 results were to happen by random chance, given this set of input parameters, and provided
757 estimates of the likelihood that our predictions are true. The FiReMAGE approach calculates these
758 permutations for every trait, species, and dataset combination. As researchers get better estimates
759 of parameters or add additional data, the permutations must also be updated. Preliminary testing
760 determined that random selection parameters were replicating real data structure and composition,
761 and our permutations serve as consistent internal controls for benchmarking results.
762
763 One of the reasons to employ permutations of regions instead of random sets of genes, was the
764 collinearity of genes in genomes. We expected that the high synteny between the three Poaceae
765 species, rice, sorghum, and maize, would inflate false positives. To the extent that multiple linked
766 genes are shared across taxa due to conservation of gene order, we expected the ortholog groups
767 detected as 3/5 hits with FiReMAGE in only the Poaceae to have high FDR. However, these ortholog
768 groups did not have dramatically increased FDR compared to other 3/5 ortholog groups (Fig 7). The
769 combination of the recent genome duplications and uneven loss of duplicate genes well

770 demonstrated in maize, Arabidopsis, and others [86–91] and the prevalence of chromosomal
771 rearrangements and transposition of genes in plant genomes may have shuffled these genomes
772 enough to permit the detection of single conserved genes within our selected window sizes.
773 Although we did not see a huge influence of synteny in our comparison, other species combinations
774 may find controlling for synteny critical [92]. Any undue influence of synteny will be reflected in the
775 permutation FDR calculations, highlighting the value of empirically estimated expectations over
776 naive null hypotheses.
777
778 The success of FiReMAGE with these particular data sets and user-set parameters demonstrates that
779 this approach is viable. One feature of this analysis was the reliance on user-reported p-values for a
780 number of the data sets. This resulted in uneven p-value thresholds and presumably influenced false
781 positive and negative rates due to differences in statistical power and error across the input GWAS
782 experiments. Because FiReMAGE federates data together, a search for the optimal p-value
783 thresholds for each experiment would require determining how changing the input data affected the
784 performance of the comparison by FiReMAGE. One interesting feature of FiReMAGE is that it is
785 highly sensitive to false-negative rates. Using too stringent of a p-value cutoff for each species GWAS
786 results will fail to input into FiReMAGE the true associations of modest effect near conserved
787 regulators. FiReMAGE cannot then include this gene in an overlap and loses one true overlap as a
788 result. This problem increases as the number of species compared goes up, when stringent pvalues
789 are used. Thus, we chose p-value thresholds for the single-species inputs that are considered
790 permissive to reasonably limit false negative rates and still retain signal in this FiReMAGE
791 comparison. Again, the use of permutations to benchmark the comparison and determine
792 significance of FiReMAGE validated this choice. As candidate genes suggested by FiReMAGE are
793 tested, the genes that are confirmed and genes that were false positives will provide an additional
794 mechanism to compare and refine parameters.
795

796 The high FDRs from the KIG comparison demonstrate why common practices filter SNPs with
797 stringent thresholds (Bonferroni, FDR, etc.) before referencing functional annotations of linked
798 genes. However, these recommendations do not address that even with generous genome coverage,
799 KIG candidates cannot explain a large portion of ionomic variation (Figs 1 and 2). While some loci
800 highlighted by GWAS at these thresholds are due to spurious associations and some KIGs do not
801 drive elemental phenotypes in all populations in all species, we propose that the majority of loci lack
802 KIG candidates because there are a vast number of gene functions and allele consequences yet to be
803 discovered [8]. Candidate selection methods reliant on prior information have inherently narrow
804 scopes that increase false negatives and, by focusing efforts on highly likely confirmation
805 experiments at genes of known function, inhibit gene discovery. The KIG list is slightly better at
806 discovery, as it is valuable in transferring information between species with high quality annotations
807 to species with fewer genomic/genetic resources. However, identifying novel ionomic candidates or
808 functional annotations requires novel approaches, such as FiReMAGE.

809 Limitations of FiReMAGE

810 One effect of the FiReMAGE approach on candidate gene assessment is that it ignores alleles and
811 false positives that occur in GWAS results from a single species or in a limited number of species.
812 While of great potential utility, this approach can potentially contribute to false negative
813 assessments in at least two ways. Firstly, the FiReMAGE approach will necessarily miss
814 taxonomically-restricted solutions and limitations to element homeostasis. For example,
815 glucosinolates in Arabidopsis, benzoxazinoids in maize, and nodulation in soybeans, all have impacts
816 on the ionome [93,94] and are not evolutionarily conserved processes across these taxa. As a result,
817 alleles affecting variation in these processes, even if they have substantial impacts in one taxon,
818 would not be prioritized by FiReMAGE. The KIG list or more involved experiments like gene transcript
819 expression would be better selection methods in these cases [17]. Secondly, the requirement to
820 appear in multiple lists and the stringency of the criteria for inclusion in the GWAS outputs

821 contribute to false negatives intrinsic in FiReMAGE. Researchers particularly worried about false
822 positive assessments in their GWAS outputs may not have considered the effect that the high rate of
823 false negatives will have on downstream meta-analysis and data federation approaches, like
824 FiReMAGE. Because detecting conserved candidates by FiReMAGE relies on true positives in multiple
825 species, providing too-stringent lists of loci will result in poor performance as more species are
826 added to the comparison. Communities that make raw data readily available for re-analysis are likely
827 to see approaches like FiReMAGE return better results.
828
829 Another limitation is that FiReMAGE requires strict trait matches across taxa and we only compared
830 KIGs to loci of traits for which they had been cited. Many elemental uptake traits are related and
831 often elements are co-transported by the same transporters. The same gene could have different
832 elemental phenotypes in taxa due to many factors, including the interaction with species-specific
833 mechanisms (see above). The rules governing elemental interactions are poorly understood [95,96].
834 Rather than attempting to relate ortholog groups to similar elements they have yet to be tested for
835 with our (likely) biased understanding of the ionome, we chose a strict comparison between traits
836 across species. We have provided a list of KIGs overlapping any trait loci (S4 Table) to guide future
837 researchers interested in exploring their potential functions in related traits.
838
839 A functional requirement for FiReMAGE is the collection of the same trait across multiple diverse
840 taxa. Many traits, such as flower morphology and resistance to specific pathogens, are not
841 comparable across species. Additionally, collecting trait data from well-powered (i.e. large) GWAS
842 populations across multiple species can be time and resource-intensive. However, advances in
843 phenotyping technologies are enabling the creation of multi-species datasets, and more datasets are
844 publicly available in GWAS databases like AraGWAS [97] for Arabidopsis, RiceVarMap2 [98] for rice,
845 and GWAS Atlas [99] for studies in cotton, maize, soybean, rice, and sorghum. FiReMAGE is able to
846 identify candidates for any trait where the data is available in three or more species. We note that

847 the threshold of three or more species excludes eudicot-specific candidates in this five-species
848 comparison. As more eudicot species with ionomic GWAS experiments become available it may be
849 possible to carry out FiReMAGE searches for candidate genes using phylogenetically targeted sub-
850 sets of taxa.
851
852 Lastly, because FiReMAGE relies on orthology, the limitations surrounding current ortholog group
853 inference methods apply to the candidate lists produced by FiReMAGE. While OrthoFinder v2.0 is
854 the best algorithm for whole genome comparisons [44], there are likely some gene trees with
855 evolutionary histories that make accurate ortholog inference challenging [100]. We recommend
856 following best practices for operating OrthoFinder (see methods), including proper species sampling
857 and validating gene tree phylogenies for FiReMAGE candidates of interest when appropriate.
858

859 Future mutant screens

860 Unbiased candidate selection methods are necessary to transition into research using non-model
861 species and with genes of unknown function. Our approach has identified the conserved genes
862 associated with genetic variation via analogous GWAS in multiple species and their likelihood as
863 causal genes in ionomic phenotypic variation. In future studies, testing mutant knock-out libraries in
864 Arabidopsis [101], sorghum [102], and maize [103] will allow us to validate FiReMAGE predictions
865 and improve candidate selection parameters.
866
867 Our research demonstrates that the disparate approaches of the KIG list and FiReMAGE both
868 present valuable ionomic candidates, each with distinct advantages and weaknesses, and highlight
869 the importance of strategic decisions surrounding candidate gene identification for downstream
870 quantitative genetic analysis. Future experiments will further explore best practices and outcomes of
871 comparative GWAS methods, including parameters to balance false negatives and false positives.
872 We also can imagine implementing a ranking equation that includes variables other than

permutation calculated false discovery rates. We hope this tool facilitates gene discovery in genetics experiments across all organisms.

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

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

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

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

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

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

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

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

914 14. Spain SL, Barrett JC. Strategies for fine-mapping complex traits. Hum Mol Genet. 2015;24: R111-
915 9.

916 15. Greene CS, Troyanskaya OG. PILGRM: an interactive data-driven discovery platform for expert
917 biologists. Nucleic Acids Res. 2011;39: W368-74.

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

920 17. Schaefer RJ, Michno J-M, Jeffers J, Hoekenga O, Dilkes B, Baxter I, et al. Integrating
921 Coexpression Networks with GWAS to Prioritize Causal Genes in Maize. Plant Cell. 2018;30:
922 2922–2942.

923 18. Koornneef M, Meinke D. The development of Arabidopsis as a model plant. Plant J. 2010;61:
924 909–921.

925 19. Park ST, Kim J. Trends in Next-Generation Sequencing and a New Era for Whole Genome
926 Sequencing. Int Neurourol J. 2016;20: S76-83.

927 20. Salt DE, Baxter I, Lahner B. Ionomics and the study of the plant ionome. Annu Rev Plant Biol.
928 2008;59: 709–733.

929 21. Dhankher OP, Rosen BP, McKinney EC, Meagher RB. Hyperaccumulation of arsenic in the shoots
930 of Arabidopsis silenced for arsenate reductase (ACR2). Proc Natl Acad Sci U S A. 2006;103:
931 5413–5418.

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

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

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

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

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

944 27. Véry A-A, Nieves-Cordones M, Daly M, Khan I, Fizames C, Sentenac H. Molecular biology of K+
945 transport across the plant cell membrane: what do we learn from comparison between plant
946 species? J Plant Physiol. 2014;171: 748–769.

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

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

953 30. Rubio F, Nieves-Cordones M, Alemán F, Martínez V. Relative contribution of AtHAK5 and
954 AtAKT1 to K+ uptake in the high-affinity range of concentrations. Physiol Plant. 2008;134: 598–
955 608.

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

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

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

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

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

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

973 37. Ziegler G, Terauchi A, Becker A, Armstrong P, Hudson K, Baxter I. Ionomic screening of field-
974 grown soybean identifies mutants with altered seed elemental composition. Plant Genome.
975 2013;6: lantgenome2012.07.0012.

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

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

980 40. Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, et al. An efficient multi-locus
981 mixed-model approach for genome-wide association studies in structured populations. Nat

982 Genet. 2012;44: 825–830.

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

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

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

990 44. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics.
991 Genome Biol. 2019;20: 238.

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

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

997 47. Asaro A, Dilkes B, Baxter I. Multivariate analysis reveals environmental and genetic
998 determinants of element covariation in the maize grain ionome. bioRxiv. 2017. p. 241380.
999 doi:10.1101/241380

1000 48. Campbell MT, Bandillo N, Al Shiblawi FRA, Sharma S, Liu K, Du Q, et al. Allelic variants of
1001 OsHKT1;1 underlie the divergence between indica and japonica subspecies of rice (Oryza sativa)
1002 for root sodium content. PLoS Genet. 2017;13: e1006823.

1003 49. Rus A, Baxter I, Muthukumar B, Gustin J, Lahner B, Yakubova E, et al. Natural variants of AtHKT1
1004 enhance Na+ accumulation in two wild populations of Arabidopsis. PLoS Genet. 2006;2: e210.

1005 50. Remm M, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from
1006 pairwise species comparisons. J Mol Biol. 2001;314: 1041–1052.

1007 51. Haydon MJ, Kawachi M, Wirtz M, Hillmer S, Hell R, Krämer U. Vacuolar nicotianamine has
1008 critical and distinct roles under iron deficiency and for zinc sequestration in Arabidopsis. Plant
1009 Cell. 2012;24: 724–737.

1010 52. Remy E, Cabrito TR, Batista RA, Teixeira MC, Sá-Correia I, Duque P. The major facilitator
1011 superfamily transporter ZIFL2 modulates cesium and potassium homeostasis in Arabidopsis.
1012 Plant Cell Physiol. 2015;56: 148–162.

1013 53. Che J, Yokosho K, Yamaji N, Ma JF. A Vacuolar Phytosiderophore Transporter Alters Iron and
1014 Zinc Accumulation in Polished Rice Grains. Plant Physiol. 2019;181: 276–288.

1015 54. Gasber A, Klaumann S, Trentmann O, Trampczynska A, Clemens S, Schneider S, et al.
1016 Identification of an Arabidopsis solute carrier critical for intracellular transport and inter-organ
1017 allocation of molybdate. Plant Biol . 2011;13: 710–718.

1018 55. Gil-Díez P, Tejada-Jiménez M, León-Mediavilla J, Wen J, Mysore KS, Imperial J, et al. MtMOT1.2
1019 is responsible for molybdate supply to Medicago truncatula nodules. Plant Cell Environ. 2018.
1020 doi:10.1111/pce.13388

1021 56. Tejada-Jiménez M, Gil-Díez P, León-Mediavilla J, Wen J, Mysore KS, Imperial J, et al. Medicago
1022 truncatula Molybdate Transporter type 1 (MtMOT1.3) is a plasma membrane molybdenum
1023 transporter required for nitrogenase activity in root nodules under molybdenum deficiency.
1024 New Phytol. 2017;216: 1223–1235.

1025 57. Huang X-Y, Liu H, Zhu Y-F, Pinson SRM, Lin H-X, Guerinot ML, et al. Natural variation in a
1026 molybdate transporter controls grain molybdenum concentration in rice. New Phytol.
1027 2019;221: 1983–1997.

1028 58. Hussain D, Haydon MJ, Wang Y, Wong E, Sherson SM, Young J, et al. P-type ATPase heavy metal
1029 transporters with roles in essential zinc homeostasis in Arabidopsis. Plant Cell. 2004;16: 1327–
1030 1339.

1031 59. Olsen LI, Hansen TH, Larue C, Østerberg JT, Hoffmann RD, Liesche J, et al. Mother-plant-
1032 mediated pumping of zinc into the developing seed. Nat Plants. 2016;2: 16036.

1033 60. Tanaka N, Nishida S, Kamiya T, Fujiwara T. Large-scale profiling of brown rice ionome in an ethyl
1034 methanesulphonate-mutagenized hitomebore population and identification of high- and low-
1035 cadmium lines. Plant Soil. 2016;407: 109–117.

1036 61. Kim D-Y, Bovet L, Maeshima M, Martinoia E, Lee Y. The ABC transporter AtPDR8 is a cadmium
1037 extrusion pump conferring heavy metal resistance. Plant J. 2007;50: 207–218.

1038 62. Shin H, Shin H-S, Dewbre GR, Harrison MJ. Phosphate transport in Arabidopsis: Pht1;1 and
1039 Pht1;4 play a major role in phosphate acquisition from both low- and high-phosphate
1040 environments. Plant J. 2004;39: 629–642.

1041 63. Nagarajan VK, Jain A, Poling MD, Lewis AJ, Raghothama KG, Smith AP. Arabidopsis Pht1;5
1042 mobilizes phosphate between source and sink organs and influences the interaction between
1043 phosphate homeostasis and ethylene signaling. Plant Physiol. 2011;156: 1149–1163.

1044 64. Catarecha P, Segura MD, Franco-Zorrilla JM, García-Ponce B, Lanza M, Solano R, et al. A mutant
1045 of the Arabidopsis phosphate transporter PHT1;1 displays enhanced arsenic accumulation.
1046 Plant Cell. 2007;19: 1123–1133.

1047 65. Zhang L, Hu B, Li W, Che R, Deng K, Li H, et al. OsPT2, a phosphate transporter, is involved in the
1048 active uptake of selenite in rice. New Phytol. 2014;201: 1183–1191.

1049 66. Peiter E, Montanini B, Gobert A, Pedas P, Husted S, Maathuis FJM, et al. A secretory pathway-
1050 localized cation diffusion facilitator confers plant manganese tolerance. Proc Natl Acad Sci U S
1051 A. 2007;104: 8532–8537.

1052 67. Lanquar V, Ramos MS, Lelièvre F, Barbier-Brygoo H, Krieger-Liszkay A, Krämer U, et al. Export of
1053 vacuolar manganese by AtNRAMP3 and AtNRAMP4 is required for optimal photosynthesis and
1054 growth under manganese deficiency. Plant Physiol. 2010;152: 1986–1999.

1055 68. Hindt MN, Akmakjian GZ, Pivarski KL, Punshon T, Baxter I, Salt DE, et al. BRUTUS and its
1056 paralogs, BTS LIKE1 and BTS LIKE2, encode important negative regulators of the iron deficiency
1057 response in Arabidopsis thaliana. Metallomics. 2017;9: 876–890.

1058 69. Yuan Y, Wu H, Wang N, Li J, Zhao W, Du J, et al. FIT interacts with AtbHLH38 and AtbHLH39 in
1059 regulating iron uptake gene expression for iron homeostasis in Arabidopsis. Cell Res. 2008;18:
1060 385–397.

1061 70. Hosmani PS, Kamiya T, Danku J, Naseer S, Geldner N, Guerinot ML, et al. Dirigent domain-
1062 containing protein is part of the machinery required for formation of the lignin-based Casparian
1063 strip in the root. Proc Natl Acad Sci U S A. 2013;110: 14498–14503.

1064 71. Baxter I, Hosmani PS, Rus A, Lahner B, Borevitz JO, Muthukumar B, et al. Root suberin forms an
1065 extracellular barrier that affects water relations and mineral nutrition in Arabidopsis. PLoS
1066 Genet. 2009;5: e1000492.

1067 72. Asaro A, Ziegler G, Ziyomo C, Hoekenga OA, Dilkes BP, Baxter I. The Interaction of Genotype and
1068 Environment Determines Variation in the Maize Kernel Ionome. G3 . 2016;6: 4175–4183.

1069 73. Baxter IR, Ziegler G, Lahner B, Mickelbart MV, Foley R, Danku J, et al. Single-kernel ionomic
1070 profiles are highly heritable indicators of genetic and environmental influences on elemental
1071 accumulation in maize grain (Zea mays). PLoS One. 2014;9: e87628.

1072 74. Campos ACAL, van Dijk WFA, Ramakrishna P, Giles T, Korte P, Douglas A, et al. 1,135 ionomes
1073 reveal the global pattern of leaf and seed mineral nutrient and trace element diversity in
1074 Arabidopsis thaliana. Plant J. 2021;106: 536–554.

1075 75. Taliercio E, Scaboo A, Baxter I, Locke AM. The Ionome of a Genetically Diverse Set of Wild
1076 Soybean Accessions. Crop Sci. 2019. doi:10.2135/cropsci2019.02.0079

1077 76. Norton GJ, Deacon CM, Xiong L, Huang S, Meharg AA, Price AH. Genetic mapping of the rice
1078 ionome in leaves and grain: identification of QTLs for 17 elements including arsenic, cadmium,
1079 iron and selenium. Plant Soil. 2010;329: 139–153.

1080 77. Lunn D, Gaddipati SR, Tucker GA, Lycett GW. Null mutants of individual RABA genes impact the
1081 proportion of different cell wall components in stem tissue of Arabidopsis thaliana. PLoS One.
1082 2013;8: e75724.

1083 78. Filatov V, Dowdle J, Smirnoff N, Ford-Lloyd B, Newbury HJ, Macnair MR. Comparison of gene
1084 expression in segregating families identifies genes and genomic regions involved in a novel
1085 adaptation, zinc hyperaccumulation. Mol Ecol. 2006;15: 3045–3059.

1086 79. Hover BM, Loksztejn A, Ribeiro AA, Yokoyama K. Identification of a cyclic nucleotide as a cryptic
1087 intermediate in molybdenum cofactor biosynthesis. J Am Chem Soc. 2013;135: 7019–7032.

1088 80. Schwarz G, Mendel RR. Molybdenum cofactor biosynthesis and molybdenum enzymes. Annu
1089 Rev Plant Biol. 2006;57: 623–647.

1090 81. Stork J, Harris D, Griffiths J, Williams B, Beisson F, Li-Beisson Y, et al. CELLULOSE SYNTHASE9
1091 serves a nonredundant role in secondary cell wall synthesis in Arabidopsis epidermal testa cells.
1092 Plant Physiol. 2010;153: 580–589.

1093 82. Qiao Z, Lampugnani ER, Yan X-F, Khan GA, Saw WG, Hannah P, et al. Structure of *Arabidopsis*
1094 CESA3 catalytic domain with its substrate UDP-glucose provides insight into the mechanism of
1095 cellulose synthesis. Proc Natl Acad Sci U S A. 2021;118. doi:10.1073/pnas.2024015118

1096 83. Kosuth T, Leskova A, Castaings L, Curie C. Golgi in and out: multifaceted role and journey of
1097 manganese. New Phytol. 2023;238: 1795–1800.

1098 84. Mertz RA, Ellsworth P, Ellsworth P, Lori Tausta S, von Caemmerer S, Howard Berg R, et al. The
1099 maize Aliphatic Suberin Feruloyl Transferase genes affect leaf water movement but are
1100 dispensable for bundle sheath CO2 concentration. bioRxiv. 2020. p. 2020.09.30.321265.

1101 doi:10.1101/2020.09.30.321265

1102 85. Morel M, Crouzet J, Gravot A, Auroy P, Leonhardt N, Vavasseur A, et al. AtHMA3, a P1B-ATPase
1103 allowing Cd/Zn/Co/Pb vacuolar storage in Arabidopsis. Plant Physiol. 2009;149: 894–904.

1104 86. Birchler JA, Yang H. The multiple fates of gene duplications: Deletion, hypofunctionalization,
1105 subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation.
1106 Plant Cell. 2022;34: 2466–2474.

1107 87. Gout J-F, Lynch M. Maintenance and Loss of Duplicated Genes by Dosage Subfunctionalization.
1108 Mol Biol Evol. 2015;32: 2141–2148.

1109 88. Schnable JC, Springer NM, Freeling M. Differentiation of the maize subgenomes by genome
1110 dominance and both ancient and ongoing gene loss. Proc Natl Acad Sci U S A. 2011;108: 4069–
1111 4074.

1112 89. Edger PP, Pires JC. Gene and genome duplications: the impact of dosage-sensitivity on the fate
1113 of nuclear genes. Chromosome Res. 2009;17: 699–717.

1114 90. Panchy N, Lehti-Shiu M, Shiu S-H. Evolution of Gene Duplication in Plants. Plant Physiol.
1115 2016;171: 2294–2316.

1116 91. Thomas BC, Pedersen B, Freeling M. Following tetraploidy in an Arabidopsis ancestor, genes
1117 were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive
1118 genes. Genome Res. 2006;16: 934–946.

1119 92. Murat F, Van de Peer Y, Salse J. Decoding plant and animal genome plasticity from differential
1120 paleo-evolutionary patterns and processes. Genome Biol Evol. 2012;4: 917–928.

1121 93. Harun S, Abdullah-Zawawi M-R, Goh H-H, Mohamed-Hussein Z-A. A comprehensive gene
1122 inventory for glucosinolate biosynthetic pathway in Arabidopsis thaliana. J Agric Food Chem.
1123 2020;68: 7281–7297.

1124 94. Chu Q, Watanabe T, Shinano T, Nakamura T, Oka N, Osaki M, et al. The dynamic state of the
1125 ionome in roots, nodules, and shoots of soybean under different nitrogen status and at
1126 different growth stages. J Plant Nutr Soil Sci. 2016;179: 488–498.

1127 95. Baxter I. Should we treat the ionome as a combination of individual elements, or should we be
1128 deriving novel combined traits? J Exp Bot. 2015;66: 2127–2131.

1129 96. Baxter I. Ionomics: studying the social network of mineral nutrients. Curr Opin Plant Biol.
1130 2009;12: 381–386.

1131 97. Togninalli M, Seren Ü, Meng D, Fitz J, Nordborg M, Weigel D, et al. The AraGWAS Catalog: a
1132 curated and standardized Arabidopsis thaliana GWAS catalog. Nucleic Acids Res. 2018;46:
1133 D1150–D1156.

1134 98. Zhao H, Yao W, Ouyang Y, Yang W, Wang G, Lian X, et al. RiceVarMap: a comprehensive
1135 database of rice genomic variations. Nucleic Acids Res. 2015;43: D1018-22.

1136 99. Tian D, Wang P, Tang B, Teng X, Li C, Liu X, et al. GWAS Atlas: a curated resource of genome-
1137 wide variant-trait associations in plants and animals. Nucleic Acids Res. 2020;48: D927–D932.

1138 100. Emms DM, Kelly S. Benchmarking Orthogroup Inference Accuracy: Revisiting Orthobench.

1139 Genome Biol Evol. 2020;12: 2258–2266.

1140 101. Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, et al. Genome-wide insertional
1141 mutagenesis of Arabidopsis thaliana. Science. 2003;301: 653–657.

1142 102. Addo-Quaye C, Tuinstra M, Carraro N, Weil C, Dilkes BP. Whole-Genome Sequence Accuracy Is
1143 Improved by Replication in a Population of Mutagenized Sorghum. G3 . 2018;8: 1079–1094.

1144 103. McCarty DR, Settles AM, Suzuki M, Tan BC, Latshaw S, Porch T, et al. Steady-state transposon
1145 mutagenesis in inbred maize. Plant J. 2005;44: 52–61.

1146 104. Krishnakumar V, Contrino S, Cheng C-Y, Belyaeva I, Ferlanti ES, Miller JR, et al. ThaleMine: A
1147 Warehouse for Arabidopsis Data Integration and Discovery. Plant Cell Physiol. 2017;58: e4.

1148

1149 Supplemental Files

1150 **S1 Fig. Orthologs in groups with 3/5 species representation.** Error bars represent the 5th/95th

1151 percentiles of the permutations.

1152

1153 **S2 Fig. Orthologs in groups with 4/5 species representation.** Error bars represent the 5th/95th

1154 percentiles of the permutations.

1155

1156 **S3 Fig. Distribution of list sizes in the inferred KIG permutations depending on ortholog tables.** The

1157 old distribution of permutations' list sizes was from using OrthoFinder to call ortholog groups only

1158 between the species in the inferred KIG. The new distribution of permutations' list sizes shows the

1159 list size when *Liriodendron tulipifera* is added to our OrthoFinder run as an outgroup to our target

1160 species. The green dot represents the number of inferred KIGs in the actual list.

1161

1162 **S4 Fig. Inferred KIG and random permutations of the KIG list compared to GWAS loci.** Solid, vertical

1163 lines represent the percentage of GWAS loci explained by a gene from OrthoFinder inferred KIG lists.

1164 Kernel Density Estimate (KDE) plots display the distribution of 1000 random permutations with the

1165 5th/95th percentiles shown as dashed lines. Inferred KIG list permutations were made from the

1166 orthologs of a list of random genes equal to the species and elemental composition of the primary
1167 KIG list (see Random permutations methods section).
1168
1169 **S5 Fig. Distribution of the number of orthologous candidates returned per loci.**
1170
1171 **S1 Table. The current primary KIG list.** Genes with functional characterizations from primary
1172 literature sources that are linked to changes in the ionome.
1173
1174 **S2 Table. Updated KIG list with OrthoFinder inferred orthologs.** *T. aestivum* (wheat) is not included
1175 in this version of the inferred KIG list.
1176
1177 **S3 Table. False discovery rates of the genes and loci overlaps in Fig 1.** False discovery rates are
1178 calculated by taking the average number of genes/loci in random permutation overlaps and dividing
1179 by the number of genes/loci in the actual GWAS datasets.
1180
1181 **S4 Table. KIGs overlapping GWAS loci.** KIGs compared to all elemental uptake trait loci regardless of
1182 KIG list citations.
1183
1184 **S5 Table. KIGs in FiReMAGE's candidate lists.**
1185
1186 **S6 Table. False discovery rate table for gene and loci overlaps in S4 Fig.** False discovery rates are
1187 calculated by taking the average number of genes/loci in random permutation overlaps and dividing
1188 by the number of genes/loci in the actual GWAS datasets.
1189
1190 **S7 Table. The complete FiReMAGE priority lists for each trait and species.**
1191

1192 **S8 Table. Top 10% of all FiReMAGE candidates.** FiReMAGE candidates were sorted by their
1193 permutation calculated false discovery rates, and taking the top 10% with the lowest FDR.
1194
1195 **S9 Table. FiReMAGE candidates in Arabidopsis with no previously known biological processes**
1196 **function.** We used ThaleMine [104] to analyze Arabidopsis candidates for each trait to identify genes
1197 with unknown function.
1198
1199 **S10 Table. Soybean SNPs associated with ionomic traits.** SNP associations generated with methods
1200 outlined in "Soybean GWAS and 2015-2017 field experiments".
1201
1202