

Plant Nutrition Cyber Infrastructure: A comparative genomic browser to enable rapid identification of candidate genes controlling mineral nutrition related traits.

Group Leader: Ivan Baxter (USDA, Danforth Center) ivan.baxter@ars.usda.gov
Martin Broadley (Nottingham University) Martin.Broadley@nottingham.ac.uk
Matt Hudson (University of Illinois) mhudson@illinois.edu

Community group members: Martin Broadley, Ivan Baxter, Keyan Zhao, Philip White, Malcolm Hawkesford, Chris Johnson, Gerrit Hoogenboom, Stuart Roy, Ramil Mauleon, Sylvie Brouder, Matt Hudson

1. The biological challenge the seed CI is aimed at addressing.

A major challenge for plant scientists is to understand how gene function and regulation integrates from subcellular to whole-plant (and community) scales. Multiscale data integration is essential for translating knowledge from model eukaryotes for the purposes of crop improvement and environmental stewardship. The plant nutrition community has generated vast legacy data, including detailed knowledge of gene function through to whole plant and community datasets. There is an additional wealth of legacy data within the agronomy and soil geochemistry communities. However, efforts to integrate these data have been minimal due to a lack of multiscale CI tools.

Genes affecting mineral element homeostasis in plants are now being identified at an accelerating rate. Large datasets of mineral element composition (*ionomes*) of plant and cell tissues are facilitating this process; millions of open-source data points are available for model organisms including *Arabidopsis*, *Saccharomyces cerevisiae*, and crops such as maize, rice and *Brassica*. To date, hundreds of loci affecting elemental uptake and tolerances to environmental mineral stresses have been identified from the ionomes of mutagenized and inbred mapping populations, and from natural variants and single nucleotide polymorphism (SNP) panels. Although the causal gene and polymorphisms underlying these loci have been identified in only a small number of cases we believe that the nature of these traits makes them highly suitable for CI-based molecular discovery, since the underlying genetics is often conserved and relatively simple across distantly related species.

Identifying causal polymorphisms on a locus by locus basis is intractable, especially in crop plants, due to labor costs and difficulties with quantitative phenotyping in outcrossed populations. However, the recent rapid acceleration in sequencing capacity has delivered a wealth of information on orthologs and syntenic relationships between genomes. If comparative genomics could be combined with outputs from genetic mapping studies for a wide range of phenotypes, candidate genes could be identified *in silico*, based on an assumption of conserved mechanisms. Such comparisons could accelerate the pace of gene identification and allow for discoveries in model organisms to flow more easily into species which provide the food, fuel and fiber that society requires. Discoveries resulting from these approaches will increase our understanding of plant nutrition and provide trait locus information for marker assisted breeding approaches.

2. The societal significance of the challenge.

The Grand Challenge being addressed by the proposed CI ***is to develop sustainable and nutritious crops from finite mineral resources for a growing human population in the context of a globally changing climate.***

3. A detailed description of the functionalities of the seed CI.

For this seed CI, we propose to develop a data integration platform that will allow users to identify orthologous candidate genes that lie within intervals identified using QTL or association mapping for any two phenotypes in any two sequenced organisms.

An immediate opportunity for multiscale data integration is afforded by existing public domain datasets. For example, candidate SNPs, genes and loci from one genome can be compared with orthologous loci in other, related genomes to identify likely conserved gene function by integrating genomic and phenotypic data. Mineral nutrient traits (and likely other molecular traits) appear to be relatively simple, with multiple examples of genetic variation in orthologs conferring the bulk of phenotypic variation in populations of widely divergent species.

An Example Case Study: identifying candidate genes for sodium tolerance. A researcher wants to integrate three datasets: 1) QTLs for Na accumulation identified in inbred populations of diploid *Brassica rapa* grown at Nottingham, UK, 2) SNPs associated with Na accumulation identified in natural variants of *Arabidopsis* at Purdue, US; 3) QTLs for Na tolerance and accumulation identified in rice insertion mutants at a high-throughput crop phenomics platforms in Adelaide, Australia. In the proposed CI, the researcher uploads QTL traces from (1) and (3) along with the genetic maps and tells the system the significance cutoff, and the amount of buffer to add to the confidence intervals to account for uncertainties in the physical/genetic map conversion. The system maps the markers onto the physical map, converts the genetic distances to inferred physical distances, and draws QTL maps of both genomes. Either the system or the user then define regions of the genome corresponding to significant QTL (e.g. those that meet a user-defined threshold based on LOD score) and the software assembles gene lists from those regions. The system then asks the user for a stringency level for the orthology comparison (an e value cutoff plus a qualitative metric such as 'reciprocal best hit' or 'partial BLAST hit') and produces a table of orthologous genes (with annotations) and a graphical display of the significant regions (in a second, scalable window). When candidate genes are selected, ortholog relationships and the location of candidate genes in the QTL map window are displayed. The user then saves the comparison and uploads dataset (2). The gene space from this dataset can then be compared to either dataset (1) or (3) or to the tables resulting from the original comparison. The user can adjust the stringency of these comparisons by altering the size of the genome windows under consideration or by altering the stringency of the ortholog comparison. Thus, the combination analysis of the quantitative genetics, comparative sequence and annotation information enabled by the CI would allow a much more powerful search for candidate loci for detailed molecular exploration.

As demonstrated by the above case study, the seed CI would have the following capabilities:

1. The ability to translate between physical and genetic maps using known markers.
2. The ability to define different levels of orthology/quasi orthology/gene families
3. An easy to use graphical user interface in addition to allowing for web-services based text queries
4. The ability to display other data about genes using web services.
5. The ability to handle both QTL traces and Association mapping data.

While this tool is urgently needed in the mineral nutrition field, it is inherently generalizable, as any trait can be used to generate the QTL or association mapping traces. As such, this tool will be a good complement to the GtoP efforts already existing in iPlant.

4. Design, development and implementation time line.

Design 1 month

Development: 4 months

Beta testing: 1 month

5. Management plan.

A full time programmer (Brandon Smith) in Matt Hudson's lab will be dedicated to the project and Dr. Hudson will be the primary supervisor of the development of the browser. Brandon Smith has a BS in Computer Science and MS in Bioinformatics, and has successfully developed both a graphical web-based tool for orthology assignment and reciprocal blast, and a rapid and user-friendly stand-alone graphical genome browser. Manuscripts describing these programs are in preparation. The open-source code of the genome browser (developed to view second-generation sequencing data with funding from the Energy Biosciences Institute, and presented at the 2010 iPlant conference) will form the basis of the proposed QTL comparative genomic browser. Given the base of source code currently in place, we estimate that 6 months of full time work should be sufficient to deliver a working beta version of the proposed browser. Drs Baxter and Broadley will participate in the design phase and will provide input on necessary features and functions throughout development. Beta testing will be carried out by members of the working group and their laboratories (Specifically, Baxter, Broadley, Tester, Hawkesford and White). Further programmer time will be necessary for debugging and building a public-release version.

6. Brief vision of the future more comprehensive CI needs of the discipline, and how the seed CI being developed will be integrated and help facilitate the larger CI

Aspects that would be implemented in later versions

1. Association mapping data. For association mapping data, the user would enter the significance cutoff for the chosen association score and the window around each significant search (which should be ~the rate of LD decay in the organism).
2. Web outlinks or web-services based queries. The user can get more information about the genes in the intervals through web- or web-services based queries to other online resources (for example: queries of T-DNA insertion line phenotypes in Arabidopsis and yeast mutants) which would be directly accessible via the proposed CI.

To truly integrate mineral nutrition from the DNA to the globe, we will require models that incorporate the current datasets of plant mineral profiles and genetic data and link them to: other phenotypic data, environmental data such as farm nutrient inputs, soil and water characteristics, and data on the nutrition of the people and animals consuming the products. CI challenges inherent in this process will include ways of obtaining and digitizing historical agronomic trials and finding ways to identify, link, and integrate the diverse types of data listed above. This CI is a first step towards the integration of species and trait specific datasets, using the genome as a bridge to identify common mechanisms of related traits.