

Plant Adaptation to the Environment: Required Cyber Infrastructure

Group Leader: John McKay (John.McKay@ColoState.edu) & John Willis (jwillis@duke.edu)

Community group members: Sally Aitken (Sally.Aitken@ubc.ca), Amy Angert (Amy.Angert@colostate.edu), Monica Geber (mag9@cornell.edu), Tom Juenger (tjuenger@mail.utexas.edu), Leonie Moyle (lmoyle@indiana.edu), Johanna Schmitt (Johanna_Schmitt@brown.edu), Sharon Strauss (systrauss@ucdavis.edu), Outi Savolainen (outi.savolainen@oulu.fi)

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

There is clear and broad agreement within the scientific community that one of the most pressing grand challenges in biology **is to understand how plants adapt to their complex and often unpredictable biotic and abiotic environments**. The proposed CI involves researchers involved in investigating the mechanistic basis of plant adaptation to address this challenge. The group members include individuals with expertise in climatic and soil databases, spatial statistics and plant physiology, ecology, and evolution.

The proposed CI will allow this diverse group of participants to coalesce around the concept that an integrated cyberinfrastructure would improve the group's ability to more efficiently determine the fundamental genetic, biochemical and physiological mechanisms underlying plant adaptation, by improving cooperation and integration across laboratories, field sites and species. The cyberinfrastructure would also allow for better training in the integrative concepts underlying evolutionary genetics and plant adaptation, and also improve the education of the public on the significance of this work for agriculture and future sustainable utilization of natural resources.

The significant questions the proposed CI will help address are;

- What loci and traits underlie adaptation?
- What is the Evolutionary history of these alleles?
- What climatic and edaphic factors are strongly correlated with functional genetic variation?
- What environmental and genetic factors limit species ranges and adaptation?

2. The societal significance of the challenge.

Establishing what natural genetic variants underlie plant adaptation at a molecular mechanistic level will provide breakthroughs in our understanding of plant physiology, ecology and evolution that will allow us to explore community and ecosystem responses to both local and global environmental changes. Importantly, such fundamental discoveries will **greatly advance our ability to conserve and subsequently to exploit genetic diversity to produce new crops and ecological services with greater resilience to emerging changes in pathogen, water, temperature, salinity and mineral nutrient stress, and higher level environmental impacts**. Such information will also help us better predict the extent and limits of possible future adaptations of plants to changes in the surface chemistry of the Earth driven by a changing global climate.

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

We propose to create a CI for relating the natural variation in a plant species' genotype and phenotype to an environmental variation across the landscape. The goal is to build an intuitive graphical user interface that will allow both novice and expert users to explore the interconnections between the landscape of environmental variables and a geo-referenced catalog of natural genetic variation within a plant species.

The CI will enable the user both to query the data using combinations of criteria, and to browse the data sets using a geographic map browser, much like a Google map, that can display environmental, genetic and phenotypic data. Under both query modes the user can simply explore the data qualitatively, or analyze patterns using statistical methodology. The CI will provide layers of climatic, edaphic, and biological data for the specific locations identified. This will include some information on data quality for each location to avoid spurious effects of interpolation etc. Layers of interest include climate data (using a model that includes topology), species, genotype, as well as soil type, solar radiation, slope, aspect, hydrology and other relevant data that are available. For climate and other data types that are sampled temporally, the user will be able to define the dates and timeline of the data.

The proposed CI will also include sophisticated tools for visualization, with the major goal having the ability to map and visualize any environmental or plant genetic variable (including summary statistics) onto a geographic map of a defined location. For example, many users would want to truncate the maps based on the range area of a given species. The graphics created by the CI will be of publication quality. The visualization will also include the ability to create dynamic movies.

Finally, the CI will include tools for analysis, ideally by calling well defined statistical procedures in R, and other programming languages. This is necessary so that users do not have to switch programs in order to analyze the data. Analyses to be included are multivariate and spatial stats, PCA, Structure. This CI will include the ability to save an analysis session and will automatically create a log of the workflow that the user followed.

As a starting point, we would like to build on an existing tool like ClimateWNA <http://www.genetics.forestry.ubc.ca/cfcg/ClimateWNA/ClimateWNA.html>, a program developed by Sally Aitken's group that can generate scale-free climate data for ecological genomic and climate change studies in western North America. ClimateWNA extracts and downscales historical and future monthly climate data and then calculates climate variables for specific latitude, longitude, and elevation locations, with the output easily viewed as high-resolution maps of particular geographic regions of western North America. Currently ClimateWNA is limited to climate data in western North America, but our goal would be to extend it to other geographic regions (esp. in North America and Europe) and to other classes of environmental data, especially edaphic factors such as soil chemistry. And we would like to incorporate plant species' genetic and phenotypic data for a comprehensive analysis of plant adaptation. Along these lines, David E Salt's group at Purdue has 1-year of funding to support a computer science graduate student to develop the Ionomic Atlas, which will be a graphical interface, similar to a Google map, that can analyze natural genetic variation in *Arabidopsis thaliana* populations in terms of the mineral nutrient and trace element content of plant tissues obtained through experiment (ionomic data), and the soil composition and climate data of the native habitat of each population. Where possible the adaptation group will leverage this work by generalizing it to a broader set of genotypes, phenotypes and environments, and incorporating data analysis capacities.

4. Design, development and implementation time line.

The timeline to design, develop and implement the Adaptation CI prototype is expected to be one year.

5. Management plan.

This prototype project will be co-managed by Dr's Willis and McKay and the iPlant development team. Willis and McKay in collaboration with the IT staff person supported by iPlant and representatives of the broader adaptation community, including Dr Aitken's, will develop the

initial schema and identify all necessary data resources. The iPlant development team will do the programming and interface development to develop the CI.

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.

Understanding how genome-wide variation relates to plant adaptation will require major advances in bioinformatics, computational biology, statistical analysis and modeling. These tools will need to be able to handle the deluge of genomic sequence data from hundreds, thousands, or even millions of plants that is about to result from advances in next generation genome sequences, and to relate it to massively detailed phenomic (diverse morphological, developmental, cellular, physiological, metabolomic, transcriptomic, and proteomic phenotypes scored in native and experimental settings), geographical, climatological, and geological databases constructed for those individuals and populations. Such integration can be visualized in Figure 1, in which data is grouped logically into sets that can be anchored to **Genomic space** (e.g. genetic polymorphism, gene, transcript, and protein), **Landscape space** (e.g. soil type, rainfall, temperature and plant genotype) and **Phenotypic space** (e.g. metabolic pathways, water use efficiency, drought tolerance).

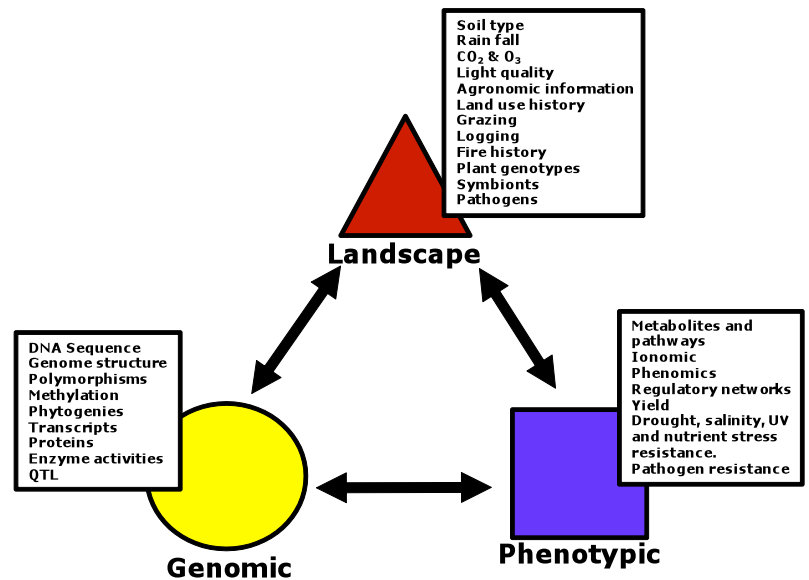


Figure 1. Future more comprehensive adaptation CI

Furthermore, the proposed adaptation CI provides a key bridge point connecting the biome and community levels of organization through to the genome. The adaptation CI fits into the workflow of the genotype to phenotype, integrating the connections between genotype and phenotype upwards to the landscape scale. The adaptation CI also integrates downwards in scale, connecting the community and biome information in the proposed biodiversity CI to the level of individual.