2008 Natural Variation and Comparative Genomics Subcommittee Report

Prepared by Julin Maloof (Co-chair, jnmaloof@ucdavis.edu) and J. Chris Pires (Co-chair,piresjc@missouri.edu)

Arabidopsis thaliana serves not only as a model system for understanding the genetic, molecular and biochemical functions underlying plant life, but also for determining the mechanisms by which these functions (and variation in them) contribute to ecological and evolutionary success. The ease of genetic manipulation, abundant natural variation, and rich understanding of genetic and biochemical pathways all point to the suitability of Arabidopsis and its relatives for ecological, quantitative genetic, and evolutionary studies. Indeed, Arabidopsis and its relatives represent an ideal system for understanding environmental adaptation, quantitative genetic variation, and microevolution at the mechanistic level.

Natural variation and comparative genomics studies are required for true understanding of how genes function. For example, understanding how genes are used to build an *A. thaliana* plant requires knowledge not only about molecular functions in *A. thaliana*, but also an understanding of why *A. thaliana* genes don't make a plant that looks more like *Capsella*, or *Brassica*, or *Cleome*, or cotton. Thus, understanding the genetic basis of developmental, metabolic, or physiological differences between species is at the very crux of plant biology. Finally, diverse species with different structures, life histories, and environmental adaptations provide tools for exploring gene function (in the molecular sense), that complement those traditionally deployed in *A. thaliana*. More generally, *A. thaliana* is only second to humans when it comes to knowledge and ability to exploit sequence variation for understanding biological processes, and indeed might serve as a useful model for developing methods that will be applicable in medical genetics.

Notable Advances and Publications

A number of publications in the last year demonstrated significant advancement in the field. We will highlight a few that represent new directions or resource development. Clark et al. [1] published results from resequencing twenty Arabidopsis accessions. These data give a genome-wide view of nucleotide variation in *Arabidopsis thaliana*, highlight genomic regions with substitution patterns suggestive of non-background selection, allow evaluation of population genetic statistics against empirical data, and provide a rich database of polymorphisms for further genetic study. Further analysis of these data by Kim et al. [2] revealed that linkage disequilibrium (LD) decays within an average of 10kb, notably faster than previous estimates. Data from these studies has been used by the Borevitz, Nordborg, and Weigel groups to design an Affymetrix™ genotyping chip that assays 250,000 SNPs. Importantly, statistical methods for association mapping in Arabidopsis have been developed to help cope with the complex population structure often present in these types of mapping populations [3,4].

Microarrays have been used to assay gene expression across accessions and in recombinant inbred line (RIL) mapping populations [5-8]. This allowed mapping of loci (eQTL) controlling variation in gene expression. Combining eQTL and traditional QTL mapping provides testable hypotheses about the mechanisms of QTL action. Further studies have used high-throughput methods to study variation in metabolites [9,10], ion content [11], and genome methylation [12,13]. In regards to comparative genomics, recent publications of the grape genome (and forthcoming publication of the papaya genome) have revised our understanding of the timing of the three nested whole genome duplication events contained within *A. thaliana*. The gamma duplication event is now thought to occur not at the origin of the angiosperms but with the origin of the eudicots or rosids, and the beta duplication event is also now considered to be much more recent.

New Resources

- The above mentioned resequencing data are integrated into the TAIR sequence viewer; the SNP genotyping chip is available from Affymetrix™. *Arabidospis lyrata* genomic sequence is in assembly at JGI; *Capsella rubella* genomic sequence is in production.
- Densely genotyped mapping populations (RILs, near isogenic lines (NILs), heterogeneous inbred families (HIFs), and panels for association mapping) are critical to the field. New RIL populations continue to become available (www.inra.fr/vast/RILs.htm) [14,15], including an interesting 19-parent advanced intercross population [16]. Genome-wide NIL introgression sets are also available for two populations [17,18]. HIFs have been helpful in fine-mapping in the Bay x Shah RIL set and could be exploited elsewhere. An expanded association mapping panel genotyped at 250,000 SNPs using the SNP chip is being developed and will be available from the stock center.

Needs and recommendations

- New resources are needed to aid the process of demonstrating functional polymorphisms for QTLs.
- Longer funding cycles (4-5 years minimum) are needed to allow QTL mapping and identification.
- Better access to mapping populations and data is needed. We encourage the deposition of all mapping populations in the stock centers as institutional Material Transfer Agreements (MTAs) may impede resource sharing. An organized effort to collect and organize NILs and HIFs for each RIL population is needed. All genotyping data should be provided in a common and easily transformable format.
- New collections of wild genotypes with documented location information, particularly from native regions, are needed. This will enable understanding of selective influence of environment and is key for ecological evolutionary questions.
- Detailed phenotyping in lab and field environments, with multiple measurements over time and detailed environmental sensing, is needed.
- There is a need for an inexpensive "fingerprinting" method for identifying *A. thaliana* stocks. SNP chip genotyping will provide the reference data but we need a method for individual labs to fingerprint their own stocks.
- An integrated database for storing and retrieving QTL and eQTL data and results is needed; ideally data will be incorporated into TAIR. Ideally this would use a common mapping framework to facilitate comparison. This is best carried out at the community level.
- Tracing the exact origin of the ancient duplication events will require a multi-gene nuclear phylogeny of the major lineages of the *Brassicaceae* and closely related families. Characterizing these whole genome duplication events can be done in non-model organisms by molecular cytogenetics (Comparative Genome Hybridization, CGH) and increasingly by transcriptome and whole genome sequencing. A *Brassiaceae* "genome browser" needs to be developed that could serve as a plant cyberinfrastructure model that eventually extends to other rosid genomes and eventually to all eudicots and beyond. This is also critical for researchers leveraging Arabidopsis knowledge to study morphological and physiological evolution in *Brassicaceae*. The phylogenetic project will require a consortium to deploy new approaches to solve the problem.
- High-throughput sequencing enables sequencing of multiple accessions within and across species. Full genome sequence from additional *A. thaliana* accessions and *Brassicaceae* species will aid QTL mapping and cloning and facilitate understanding of diversity, innovation, and selective pressure. Creating fixed homozygous lines of an array of diverse species and varieties within species is the first step to leverage these new genomic technologies. It is worth considering a hierarchical approach to genome sequencing (multiple very high quality references, higher number of genomes surveyed at reduced quality, very high number of genomes of low quality [e.g., just SNPs]). It is important that (re)sequencing studies allow discovery of genes absent from reference genomes such as *A. thaliana* Columbia. This is particularly important because "Next Generation" methods are particularly poor in this regard. Cost-effective strategies for accessing this very interesting aspect of variation need to be discussed and implemented.

References

1 Clark et al (2007) Science 317: 338-342.
2 Kim et al (2007) Nat Genet 39: 1151-1155.
3 Kang et al (2008) Genetics 178: 1709-1723.
4 Zhao et al (2007) PLoS Genet 3: e4.
5 van Leeuwen et al (2007) Plant Cell 19: 2099-2110.
6 Kliebenstein et al (2006) BMC Bioinformatics 7: 308.
7 West et al (2006) Genetics 175: 1441-1450.
8 Keurentjes et al (2007) Proc Natl Acad Sci U S A 104: 1708-1713.
9 Meyer et al (2007) Proc Natl Acad Sci U S A 104: 4759-4764.
10 Wentzell et al (2007) PLoS Genet 3: e162.
11 Rus A et al (2006) PLoS Genet 2: e210.
12 Zhang et al (2008) PLoS Genet 4: e1000032.
13 Vaughn et al (2007) PLoS Biol 5: e174.
14 Pfalz et al (2007) PLoS ONE 2: e578.
15 O'Neill et al (2008 Theor Appl Genet .
16 Scarcelli et al (2007) Proc Natl Acad Sci U S A 104: 16986-16991.
17 Keurentjes et al (2006) Genetics 175: 891-905.
18 Torjek et al (2008) J Hered : esn014-esn014.