

IN PRESS - THE PLANT JOURNAL

Shotguns and SNPs: how fast and cheap sequencing is revolutionizing plant biology

Steven D. Rounsley^{a,b} and Robert L. Last^{c,d,*}

^aSchool of Plant Sciences, University of Arizona, Tucson AZ 85721

^bBIO5 Institute, University of Arizona, Tucson AZ 85721

^cDepartment of Biochemistry and Molecular Biology, Michigan State University, East Lansing MI 48824 USA

^dDepartment of Plant Biology, Michigan State University, East Lansing MI 48824 USA

Correspondence to *(fax 517-353-9334; email: lastr@msu.edu)

KEYWORDS

Whole genome shotgun sequencing, whole genome association, polymorphisms, Landsberg *erecta*, Columbia, WGA

5200 Words total, including references

Note: The authors were employees of Cereon Genomics LLC, a wholly owned subsidiary of Monsanto Co., when they participated in the generation and analysis of the Landsberg *erecta* shotgun sequence.

Summary

In 1998 Cereon Genomics LLC, a subsidiary of Monsanto Co., performed a shotgun sequencing of the *Arabidopsis thaliana* Landsberg *erecta* genome to a depth of two-fold coverage using 'classic' Sanger sequencing. This sequence was assembled and aligned to the Columbia ecotype sequence being produced by the Arabidopsis Genome Initiative. The analysis provided tens of thousands of high confidence predictions of polymorphisms between these two varieties of *A. thaliana*, and the predicted polymorphisms and Landsberg *erecta* sequence were subsequently made available to the not-for-profit research community by Monsanto. These data have been used for a wide variety of published studies including map-based gene identification from forward genetic screens, studies of recombination and organelle genetics and gene expression studies. The combination of resequencing approaches with next-generation sequencing technology has led to an increasing number of similar studies of genome-wide genetic diversity in *A. thaliana* including the 1,001 genomes project (<http://1001genomes.org/>). Similar approaches are becoming possible in any number of crop species as DNA sequencing costs plummet and throughput rapidly increases, promising to lay the groundwork for revolutionizing our understanding of the relationship between genotype and phenotype in plants.

Introduction

In 1998, the Arabidopsis Genome Initiative was well underway in its international, coordinated effort to generate the first plant genome sequence. The consortium focused on the Columbia (Col) ecotype of *Arabidopsis thaliana* using a BAC-by-BAC sequencing approach and their goal was to create a high quality reference sequence of the euchromatic regions of the genome (Arabidopsis Genome Initiative 2000). Also at that time, scientists (including these two authors) at Cereon Genomics LLC, a subsidiary of Monsanto Co., embarked on a project to sequence a second ecotype, Landsberg *erecta* (*Ler*), using a whole genome shotgun approach. While this effort was thought by many to be competing with the public project, the reality was that it was complementary and in addition to serving the company's internal needs, it led to the first large-scale, genome-wide polymorphism database for any plant species. Thus, it provided a first glimpse at the nature of large-scale genomic variation that exists within a plant species. Here, we review its immediate impact, and how similar approaches using today's technologies are advancing our understanding of plant biology and evolution.

Using the *Ler* sequence data at Cereon

The sequencing and analysis of the *Ler* genome was among the first projects at Cereon. While we appreciated that the sequence would be of long-term broad utility to the community, there were two short-term goals for the project: accessing the majority of the genes for a flowering plant and developing markers for map-based cloning. Both goals were part of a broad functional genomics strategy to use *Arabidopsis* to find genes for Monsanto's transgenic and molecular breeding programs. The strategy also included

forward genetics screening for a diverse set of mutants altered in phenotypes directly related to Monsanto's commercial targets. Thus, in addition to providing gene sequences directly, a primary goal of the *Ler* project was to provide tools to enable map-based cloning on mutants with a wide variety of phenotypes, including some that required analytical chemistry techniques and were therefore difficult to score (e.g. seed metabolite traits).

The sequencing phase of the project generated over 700,000 Sanger sequencing reads. While this seems modest by today's standards, it was very ambitious in 1998, requiring over 7,000 96-lane sequencing runs, which after filtering for mitochondrial and chloroplast contamination represented approximately 2x coverage of the *Ler* genome. Along with the large amount of sequence data came an assembly challenge. Not only was the coverage relatively low, but software tools for attempting whole genome assembly of large genomes were still in their infancy. Ultimately, with various pre- and post-processing strategies and the phrap assembler (Green, 1996), these data were assembled into a total of 92.1 Mbp that contained at least a portion of over 95% of genes in the genome (Jander *et al.* 2002). This collection of *Ler* sequences was regularly aligned against the Col sequences from the public sequencing project as they were produced to identify putative polymorphisms that could be used as markers (Fig. 1). Two distinct types of polymorphisms were predicted – single nucleotide polymorphisms (SNPs) and insertion-deletion polymorphisms (InDels). Due to the low coverage of the *Ler* sequence data, stringent thresholds were used to maximize the quality of the predicted polymorphisms, and thereby minimize the resources spent on markers that were

unlikely to be useful. Indeed, a random sampling of the SNPs showed a surprisingly high validation rate (Jander *et al.* 2002).

To adapt map-based cloning to an industrialized 'one size fits all' environment, a strategy was implemented that minimized the number of plants that were phenotyped. Instead, relatively large numbers of plants were genotyped and recombinant progeny tested for their phenotype (Jander *et al.* 2002). Genes affected in dozens of mutants were identified, including unannotated genes for enzymes involved in seed amino acid (Jander *et al.* 2004, Lee *et al.* 2008), glucosinolate (Kim *et al.* 2004, Kliebenstein *et al.* 2007) and tocopherol (Valentin *et al.* 2006, Van Eenennaam *et al.* 2003) metabolism, as well as a gene encoding a key enzyme of ascorbate biosynthesis (Jander *et al.* 2002) and *ESK* genes, whose loss of function mutants are constitutively freezing tolerant (Xin *et al.* 2007).

Making the data available to the public

As the internal successes grew, there was increasing realization that placing the polymorphism dataset in the hands of the academic community would have a mutually beneficial outcome. Individual scientists would benefit by being able to clone their genes of interest more rapidly, and Monsanto would benefit by access to that knowledge through the scientific literature. In total, the knowledge generated by the community was likely to greatly complement what Monsanto could generate internally, and at a much reduced cost. Finding the ideal mechanism and legal framework for providing access to the data was not trivial, but ultimately a partnership with The Arabidopsis Information

Resource (TAIR; www.arabidopsis.org) provided access with a "click agreement" to a license that protected only the dataset as a whole, and allowed polymorphisms to be used and published freely (Rounsley 2003). The final dataset contained 56,670 polymorphisms including 37,344 single nucleotide polymorphisms (SNPs) and 18,579 insertion-deletions (InDels), at an average density of 1 SNP per 3.3kb of genome, and 1 InDel per 6.6kb of genome. This was the largest set of genetic markers available for any plant species at the time, and remained so until similar resources for rice were made available in 2008 via an array-based resequencing platform (McNally *et al.*, 2009). Following the release of the polymorphism database, Monsanto also made the full set of *Ler* sequence contigs available – also through the TAIR website.

Use of the data by the broader Arabidopsis community - examples

The tens of thousands of predicted polymorphisms have been used in many published studies, ranging from mapping of interesting mutations to studying genome structure and function. Most notably, more than one hundred papers have appeared describing genetic mapping and map-based cloning approaches using the Cereon/Monsanto collection of markers. Not surprisingly, the fields impacted span a wide range of plant biology, with recent examples including mutants altered in cytokinesis (Thiele *et al.* 2009), root tropic responses (Miyazawa *et al.* 2009), female gametophyte development (Moll *et al.* 2008) and disease resistance (Wawrzynska *et al.* 2008). Identification of alleles (including quantitative trait loci) controlling phenotypes observed in crosses between Col and *Ler* has also benefited from the availability of dense marker maps (Edwards *et al.* 2005,

Hoekenga *et al.* 2006, Staal *et al.* 2008). The high-density marker map facilitates studies of genetic mechanisms such as genome wide patterns of recombination. For example, Drouaud and coworkers used these markers to study the dynamic pattern of meiotic recombination across chromosome 4 of Arabidopsis and characterized sequences associated with 'hot' and 'cold' spots in euchromatin (Drouaud *et al.* 2006). Sites that are polymorphic between Col and Ler are also a good starting place for efficiently finding genetic differences between Col/Ler and other ecotypes; for a recently published example, see the paper by Huang and coworkers (Huang *et al.* 2008).

Because the shotgun Ler sequence contained deep coverage of organelle DNA, it has been useful for studies of structure and expression of organelle genomes. In a recently published example, Forner and coworkers used the Ler-derived markers and reciprocal crosses to analyze the genetic basis for differences in mitochondrial mRNA terminus processing (Forner *et al.* 2008). Both maternal (likely *cis*-acting effects of mitochondrial DNA polymorphism) and *trans*-effects due to differences in nuclear genes were found. On a utilitarian but important note, polymorphisms that 'uniquely' identify an ecotype or mutant allele are very useful in confirming the identity of seed stocks, individual plants or cell cultures. This is especially useful for a plant like Arabidopsis where the tiny seeds and availability of tens of thousands of mutants and hundreds of wild accessions can lead to a nearly limitless amount of contamination and confusion of lab stocks.

The expanding universe of Arabidopsis genetic polymorphisms

While the *Ler*-Col comparison was the first published example of a genome-wide set of insertion-deletion and SNP markers, the resources for studying and using *Arabidopsis* sequence variation is expanding at a rapidly increasing rate. An early effort to mine expressed sequence tags (ESTs) and sequence tagged sites (STS) led to the identification of nearly 9000 polymorphisms across 12 *A. thaliana* accessions (Schmid *et al.* 2003). This general approach was expanded by Nordborg *et al.*, who sequenced >870 fragments in 96 different accessions of *A. thaliana*, providing an early view of the overall pattern of genetic variation across many loci and a substantial number of isolates of the species (Nordborg *et al.* 2005). These data provided insight into the overall population structure of *A. thaliana* from around the world. Their results also indicated that linkage disequilibrium decays over 25-50 kb, suggesting that genetic association mapping could be used in this or similar populations of *A. thaliana*.

The use of information about genome-wide and transcriptome-wide variation in a diverse set of individuals to discover genes of interest continues to gain popularity in plants (Nordborg and Weigel 2008), and is being applied in a wide range of studies in *A. thaliana*. For example, high resolution mapping of the Col X *Ler* recombinant inbred lines (RILs) by array hybridization provided detailed information about recombination behavior and created a durable tool for fast and high resolution mapping of QTLs in *Arabidopsis* (Singer *et al.* 2006). This idea was taken a step further by West and coworkers who combined genome-wide DNA polymorphisms and gene expression markers to comprehensively characterize RILs from a cross between the Bay-0 and Sha ecotypes (West *et al.* 2006).

Such tools that allow efficient pan-genome surveys are becoming increasingly important in harnessing 'natural variation' to associate genes with traits. Screening of diverse germplasms (linkage disequilibrium mapping) promises to become increasingly important for associating genes with phenotype. A recent report described a combination of ecotype screening, genetic segregation analysis and resequencing of a candidate gene in 92 ecotypes to demonstrate genetic association of the *MOT1* gene and shoot molybdenum content (Baxter *et al.* 2008). These types of results also facilitate studies in evolutionary and population genetics. For example, Toomajian and coworkers (Toomajian *et al.* 2006) were able to rigorously ask whether the *FRIGIDA* locus, controlling the requirement for vernalization for flowering, was under selection in *A. thaliana*. By comparing polymorphism patterns at *FRI* to more than a thousand other loci in 96 accessions, they concluded that this locus was under strong selection; such a conclusion is made compelling due to sampling of a very large number of other loci. The increasing availability of DNA sequence from related taxa allows comparisons of evolutionary and population biological processes. For instance, Foxe and coworkers examined rates of purifying (stabilizing) and positive selection in the outcrossing species *A. lyrata* with the self-pollinating *A. thaliana* (Foxe *et al.* 2008).

Recent breakthroughs in resequencing throughput and affordability have led to efforts to attempt to sequence 1,001 isolates of *A. thaliana* (<http://1001genomes.org/>). A step in this direction was reported using high density array resequencing of several dozen accessions (Borevitz *et al.* 2007, Clark *et al.* 2007). These studies provide an unprecedented view of genome evolution and relationships among individuals and populations; for example

huge variations in patterns of genetic polymorphism including large regions of low polymorphism consistent with recent 'selective sweeps' (Clark *et al.* 2007). Recently, deep coverage short read sequencing approaches have been used for several technical breakthroughs in Arabidopsis research: the resequencing of three ecotypes (Ossowski *et al.* 2008); the sequencing of the floral epigenome (Lister *et al.*, 2008); and the demonstration of its application to simultaneous mapping and mutation identification (Schneeberger *et al.*, 2009). Plummeting costs and longer read length next generation sequencing technologies should provide a staggering amount of sequence data for Arabidopsis over the next few years.

Applications to crops

Until recently, studies of genetic diversity in crops and other larger genome plant species focused on sequencing of cDNAs and analysis of specific genomic regions (Ganal *et al.* 2009). This was due to the large size and complexity of their genomes as well as lack of reference sequences for plants other than Arabidopsis and rice. A common approach was to develop primers for specific (typically evolutionarily conserved) genomic or cDNA sequences and perform PCR amplification and sequencing of these amplicons on diverse cultivars, natural isolates or related species of interest. Random sequencing of EST collections from diverse germplasms and bioinformatic detection of polymorphisms in orthologous genes is also commonly employed. High levels of redundancy in plant genomes, such as large gene families or polyploidization, present special challenges for distinguishing between polymorphic alleles and paralogous gene family members. Conversely, for diploid outcrossing species, analysis of a single individual can identify

genome-wide collections of polymorphisms due to the wide-spread heterozygosity in the genome.

As with *Arabidopsis*, second generation sequencing technologies are revolutionizing crop genomics, including our understanding of genome diversity, development of mapping resources, and studies of ecological and evolutionary biology. By lowering costs and increasing the rate of sequence acquisition, these technologies are causing researchers to rethink how crop genomes and transcriptomes are analyzed. Some of these approaches represent natural extensions of past approaches. For example, 454 EST sequencing in species as phylogenetically diverse as maize (Barbazuk *et al.* 2007) and eucalyptus (Novaes *et al.* 2008) led to the discovery of SNPs in thousands of genes at much lower cost than with dideoxy sequencing. These markers can then be deployed in molecular breeding for variety improvement or gene discovery by genetic mapping.

RILs and nearly isogenic lines (NILs) are widely used in genetic mapping of naturally occurring variation and in plant breeding. In the past, genetic analysis of RILs and NILs in crops and model plants required tedious and expensive methods for marker discovery and genotyping of the lines with these individual markers, and the resulting maps were often of relatively low resolution (Eshed and Zamir 1995, Loudet *et al.* 2002, Simon *et al.* 2008). The use of multiplexing strategies combined with fast and cheap DNA sequence analysis enables very high resolution genetics. In a recently published example (Huang *et al.* 2009), low pass Illumina sequencing was performed on 150 rice RILs derived from a cross between indica and japonica cultivars. This approach permitted construction of a

high-resolution map of the recombination events in these RILs and allowed efficient mapping of traits associated with individual RILs.

Whole genome scans of genetic polymorphism are increasingly being used to search for sequences under natural and artificial selection. While pioneering work is being done using maps with millions of polymorphisms to analyze genetic variation in human (Sabeti *et al.* 2007), this approach is also being successfully applied to studying complex traits in crop plants. Scanning populations derived from the Illinois high and low kernel oil lines with nearly 500 genetic makers revealed the influence of >50 QTL influencing the trait, with each locus responsible for small amounts of genetic variance (Laurie *et al.* 2004). This study indicates the value of genome-wide genetic analysis in revealing the genetic basis for complex traits in maize, though the existence of large numbers of small effect alleles precludes identification of the genes underlying the effects. Whole genome scan association mapping with ~8500 SNP markers was used to analyze the genetic basis of kernel oleic acid (18:1) content (Beló *et al.* 2008). In this case the fatty acid desaturase gene *fad2* was found very close to the SNP marker genetically associated with the trait, confirming the value of this approach in gene discovery in maize. Studies using genome scans of maize and its progenitor teosinte are revealing candidates for genes under artificial selection. In an early study, analysis of DNA sequences of 774 gene fragments in 14 maize inbreds and 16 teosinte inbreds led Wright and coworkers to estimate that more than 1,000 genes have been influenced by artificial selection in the evolution of teosinte to the varieties they analyzed (Wright *et al.* 2005). A similar study of genetic variation in cultivated and wild sunflower (*Helianthus annuus*) accessions revealed evidence for several dozen regions being under selection during the time since sunflower

domestication (Chapman *et al.* 2008). These studies suggest that genome-wide genetic scans will be a useful approach to identifying genes influencing important traits in a variety of crop plants.

Concluding comments

In the last ten years, we have seen the dramatic impact that an available reference genome sequence can have on an entire scientific community. In addition to all the intrinsic information that is present in that reference, it also provides a framework to which other data resources can add. In particular, the sequencing of additional related genomes can provide practical utility and immensely rich datasets for studying genetic variation. With the unprecedented changes in sequencing technologies over the last few years, production of the Cereon Col-*Ler* dataset now seems almost trivial. However, its value has been long lasting, and has seeded a burgeoning field whose current proposals seemed outrageous just a few years ago. It is staggering to consider where sequencing technologies may be in 5 years time, the potential volume of sequence data that will be collected from complex crop genomes and from the biota of complex ecosystems. With these new datasets will come tremendous challenges associated with their analysis and presentation.

Acknowledgements

We thank Ivan Baxter for helpful comments on the manuscript. Research in RLL's group is supported by NSF grants DBI-0604336 and MCB-0519740 and research in SDR's

group is supported by NSF grants DBI-0638541, DBI-0822284, and DEB-0918758.

REFERENCES

- Arabidopsis Genome Initiative** (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796-815.
- Barbazuk, W.B., Emrich, S.J., Chen, H.D., Li, L. and Schnable, P.S.** (2007) SNP discovery via 454 transcriptome sequencing. *Plant Journal*, **51**, 910-918.
- Baxter, I., Muthukumar, B., Park, H.C., Buchner, P., Lahner, B., Danku, J., Zhao, K., Lee, J., Hawkesford, M.J., Guerinot, M.L. and Salt, D.E.** (2008) Variation in molybdenum content across broadly distributed populations of *Arabidopsis thaliana* is controlled by a mitochondrial molybdenum transporter (MOT1). *Plos Genetics*, **4**.
- Beló, A., Zheng, P., Luck, S., Shen, B., Meyer, D., Li, B., Tingey, S. and Rafalski, A.** (2008) Whole genome scan detects an allelic variant of *fad2* associated with increased oleic acid levels in maize. *Molecular Genetics and Genomics*, **279**, 1-10.
- Borevitz, J.O., Hazen, S.P., Michael, T.P., Morris, G.P., Baxter, I.R., Hu, T.T., Chen, H., Werner, J.D., Nordborg, M., Salt, D.E., Kay, S.A., Chory, J., Weigel, D., Jones, J.D. and Ecker, J.R.** (2007) Genome-wide patterns of single-feature polymorphism in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A*, **104**, 12057-12062.
- Chapman, M.A., Pashley, C.H., Wenzler, J., Hvala, J., Tang, S., Knapp, S.J. and Burke, J.M.** (2008) A Genomic Scan for Selection Reveals Candidates for Genes Involved in the Evolution of Cultivated Sunflower (*Helianthus annuus*). *Plant Cell*, **20**, 2931-2945.
- Clark, R.M., Schweikert, G., Toomajian, C., Ossowski, S., Zeller, G., Shinn, P., Warthmann, N., Hu, T.T., Fu, G., Hinds, D.A., Chen, H., Frazer, K.A., Huson, D.H., Scholkopf, B., Nordborg, M., Ratsch, G., Ecker, J.R. and Weigel, D.** (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science*, **317**, 338-342.
- Drouaud, J., Camilleri, C., Bourguignon, P.Y., Canaguier, A., Berard, A., Vezon, D., Giancola, S., Brunel, D., Colot, V., Prum, B., Quesneville, H. and Mezard, C.** (2006) Variation in crossing-over rates across chromosome 4 of *Arabidopsis thaliana* reveals the presence of meiotic recombination "hot spots". *Genome Res.*, **16**, 106-114.
- Edwards, K.D., Lynn, J.R., Gyula, P., Nagy, F. and Millar, A.J.** (2005) Natural allelic variation in the temperature-compensation mechanisms of the *Arabidopsis thaliana* circadian clock. *Genetics*, **170**, 387-400.
- Eshed, Y. and Zamir, D.** (1995) An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. *Genetics*, **141**, 1147-1162.
- Forner, J., Holzle, A., Jonietz, C., Thuss, S., Schwarzlander, M., Weber, B., Meyer, R.C. and Binder, S.** (2008) Mitochondrial mRNA polymorphisms in different *Arabidopsis* accessions. *Plant Physiology*, **148**, 1106-1116.

- Foxe, J.P., Dar, V.U., Zheng, H., Nordborg, M., Gaut, B.S. and Wright, S.I.** (2008) Selection on amino acid substitutions in Arabidopsis. *Mol Biol Evol*, **25**, 1375-1383.
- Ganal, M.W., Altmann, T. and Röder, M.S.** (2009) SNP identification in crop plants. *Current Opinion in Plant Biology*, **12**, 211-217.
- Hoekenga, O.A., Maron, L.G., Piñeros, M.A., Canãşado, G.M.A., Shaff, J., Kobayashi, Y., Ryan, P.R., Dong, B., Delhaize, E., Sasaki, T., Matsumoto, H., Yamamoto, Y., Koyama, H. and Kochian, L.V.** (2006) AtALMT1, which encodes a malate transporter, is identified as one of several genes critical for aluminum tolerance in Arabidopsis. *Proceedings of the National Academy of Sciences*, **103**, 9738-9743.
- Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., Guan, J., Fan, D., Weng, Q., Huang, T., Dong, G., Sang, T. and Han, B.** (2009) High-throughput genotyping by whole-genome resequencing. *Genome Res*.
- Huang, Y.D., Li, C.Y., Biddle, K.D. and Gibson, S.I.** (2008) Identification, cloning and characterization of sis7 and sis10 sugar-insensitive mutants of Arabidopsis. *BMC Plant Biology*, **8**.
- Jander, G., Norris, S.R., Joshi, V., Fraga, M., Rugg, A., Yu, S., Li, L. and Last, R.L.** (2004) Application of a high-throughput HPLC-MS/MS assay to Arabidopsis mutant screening; evidence that threonine aldolase plays a role in seed nutritional quality. *Plant J*, **39**, 465-475.
- Jander, G., Norris, S.R., Rounsley, S.D., Bush, D.F., Levin, I.M. and Last, R.L.** (2002) Arabidopsis map-based cloning in the post-genome era. *Plant Physiol*, **129**, 440-450.
- Kim, J.H., Durrett, T.P., Last, R.L. and Jander, G.** (2004) Characterization of the Arabidopsis TU8 glucosinolate mutation, an allele of TERMINAL FLOWER2. *Plant Mol Biol*, **54**, 671-682.
- Kliebenstein, D.J., D'Auria, J.C., Behere, A.S., Kim, J.H., Gunderson, K.L., Breen, J.N., Lee, G., Gershenzon, J., Last, R.L. and Jander, G.** (2007) Characterization of seed-specific benzoyloxyglucosinolate mutations in Arabidopsis thaliana. *Plant J*, **51**, 1062-1076.
- Laurie, C.C., Chasalow, S.D., LeDeaux, J.R., McCarroll, R., Bush, D., Hauge, B., Lai, C.Q., Clark, D., Rocheford, T.R. and Dudley, J.W.** (2004) The genetic architecture of response to long-term artificial selection for oil concentration in the maize kernel. *Genetics*, **168**, 2141-2155.
- Lee, M., Huang, T., Toro-Ramos, T., Fraga, M., Last, R.L. and Jander, G.** (2008) Reduced activity of *Arabidopsis thaliana* HMT2, a methionine biosynthetic enzyme, increases seed methionine content. *Plant J*, **54**, 310-320.
- Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H. and Ecker, J.R.** (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, **133**(3), 523-536.
- Loudet, O., Chaillou, S., Camilleri, C., Bouchez, D. and Daniel-Vedele, F.** (2002) Bay-0 x Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in Arabidopsis. *Theor Appl Genet*, **104**, 1173-1184.

- [McNally, K.L.](#), [Childs, K.L.](#), [Bohnert, R.](#), [Davidson, R.M.](#), [Zhao, K.](#), [Ulat, V.J.](#), [Zeller, G.](#), [Clark, R.M.](#), [Hoen, D.R.](#), [Bureau, T.E.](#), [Stokowski, R.](#), [Ballinger, D.G.](#), [Frazer, K.A.](#), [Cox, D.R.](#), [Padhukasahasram, B.](#), [Bustamante, C.D.](#), [Weigel, D.](#), [Mackill, D.J.](#), [Bruskiewich, R.M.](#), [Rätsch, G.](#), [Buell, C.R.](#), [Leung, H.](#) and [Leach, J.E.](#) (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proceedings of the National Academy of Sciences* **106**, 12273-12278.
- [Miyazawa, Y.](#), [Takahashi, A.](#), [Kobayashi, A.](#), [Kaneyasu, T.](#), [Fujii, N.](#) and [Takahashi, H.](#) (2009) GNOM-Mediated Vesicular Trafficking Plays an Essential Role in Hydrotropism of Arabidopsis Roots. *Plant Physiol.*, **149**, 835-840.
- [Moll, C.](#), [von Lyncker, L.](#), [Zimmermann, S.](#), [Kagi, C.](#), [Baumann, N.](#), [Twell, D.](#), [Grossniklaus, U.](#) and [Gross-Hardt, R.](#) (2008) CLO/GFA1 and ATO are novel regulators of gametic cell fate in plants. *Plant Journal*, **56**, 913-921.
- [Nordborg, M.](#), [Hu, T.T.](#), [Ishino, Y.](#), [Jhaveri, J.](#), [Toomajian, C.](#), [Zheng, H.](#), [Bakker, E.](#), [Calabrese, P.](#), [Gladstone, J.](#), [Goyal, R.](#), [Jakobsson, M.](#), [Kim, S.](#), [Morozov, Y.](#), [Padhukasahasram, B.](#), [Plagnol, V.](#), [Rosenberg, N.A.](#), [Shah, C.](#), [Wall, J.D.](#), [Wang, J.](#), [Zhao, K.](#), [Kalbfleisch, T.](#), [Schulz, V.](#), [Kreitman, M.](#) and [Bergelson, J.](#) (2005) The Pattern of Polymorphism in *Arabidopsis thaliana*. *PLoS Biol*, **3**, e196.
- [Nordborg, M.](#) and [Weigel, D.](#) (2008) Next-generation genetics in plants. *Nature*, **456**, 720-723.
- [Novaes, E.](#), [Drost, D.R.](#), [Farmerie, W.G.](#), [Pappas, G.J., Jr.](#), [Grattapaglia, D.](#), [Sederoff, R.R.](#) and [Kirst, M.](#) (2008) High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *Bmc Genomics*, **9**, 312.
- [Ossowski, S.](#), [Schneeberger, K.](#), [Clark, R.M.](#), [Lanz, C.](#), [Warthmann, N.](#) and [Weigel, D.](#) (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res*, **18**, 2024-2033.
- [Rounsley, S.](#) (2003) Sharing the wealth. The mechanics of a data release from industry. *Plant Physiol*, **133**, 438-440.
- [Sabeti, P.C.](#), [Varilly, P.](#), [Fry, B.](#), [Lohmueller, J.](#), [Hostetter, E.](#), [Cotsapas, C.](#), [Xie, X.](#), [Byrne, E.H.](#), [McCarroll, S.A.](#), [Gaudet, R.](#), [Schaffner, S.F.](#), [Lander, E.S.](#), [Frazer, K.A.](#), [Ballinger, D.G.](#), [Cox, D.R.](#), [Hinds, D.A.](#), [Stuve, L.L.](#), [Gibbs, R.A.](#), [Belmont, J.W.](#), [Boudreau, A.](#), [Hardenbol, P.](#), [Leal, S.M.](#), [Pasternak, S.](#), [Wheeler, D.A.](#), [Willis, T.D.](#), [Yu, F.](#), [Yang, H.](#), [Zeng, C.](#), [Gao, Y.](#), [Hu, H.](#), [Hu, W.](#), [Li, C.](#), [Lin, W.](#), [Liu, S.](#), [Pan, H.](#), [Tang, X.](#), [Wang, J.](#), [Wang, W.](#), [Yu, J.](#), [Zhang, B.](#), [Zhang, Q.](#), [Zhao, H.](#), [Zhou, J.](#), [Gabriel, S.B.](#), [Barry, R.](#), [Blumenstiel, B.](#), [Camargo, A.](#), [Defelice, M.](#), [Faggart, M.](#), [Goyette, M.](#), [Gupta, S.](#), [Moore, J.](#), [Nguyen, H.](#), [Onofrio, R.C.](#), [Parkin, M.](#), [Roy, J.](#), [Stahl, E.](#), [Winchester, E.](#), [Ziaugra, L.](#), [Altshuler, D.](#), [Shen, Y.](#), [Yao, Z.](#), [Huang, W.](#), [Chu, X.](#), [He, Y.](#), [Jin, L.](#), [Liu, Y.](#), [Sun, W.](#), [Wang, H.](#), [Wang, Y.](#), [Xiong, X.](#), [Xu, L.](#), [Waye, M.M.](#), [Tsui, S.K.](#), [Xue, H.](#), [Wong, J.T.](#), [Galver, L.M.](#), [Fan, J.B.](#), [Gunderson, K.](#), [Murray, S.S.](#), [Oliphant, A.R.](#), [Chee, M.S.](#), [Montpetit, A.](#), [Chagnon, F.](#), [Ferretti, V.](#), [Leboeuf, M.](#), [Olivier, J.F.](#), [Phillips, M.S.](#), [Roumy, S.](#), [Sallee, C.](#), [Verner, A.](#), [Hudson, T.J.](#), [Kwok, P.Y.](#), [Cai, D.](#), [Koboldt, D.C.](#), [Miller, R.D.](#), [Pawlikowska, L.](#), [Taillon-Miller, P.](#), [Xiao, M.](#), [Tsui, L.C.](#), [Mak, W.](#), [Song, Y.Q.](#), [Tam, P.K.](#), [Nakamura, Y.](#), [Kawaguchi, T.](#), [Kitamoto, T.](#)

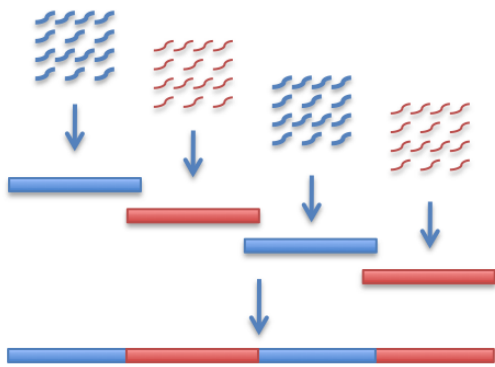
- Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C.P., Delgado, M., Dermitzakis, E.T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B.E., Whittaker, P., Bentley, D.R., Daly, M.J., de Bakker, P.I., Barrett, J., Chretien, Y.R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D.J., Sabeti, P., Saxena, R., Sham, P.C., Stein, L.D., Krishnan, L., Smith, A.V., Tello-Ruiz, M.K., Thorisson, G.A., Chakravarti, A., Chen, P.E., Cutler, D.J., Kashuk, C.S., Lin, S., Abecasis, G.R., Guan, W., Li, Y., Munro, H.M., Qin, Z.S., Thomas, D.J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L.R., Clarke, G., Evans, D.M., Morris, A.P., Weir, B.S., Johnson, T.A., Mullikin, J.C., Sherry, S.T., Feolo, M., Skol, A., Zhang, H., Matsuda, I., Fukushima, Y., Macer, D.R., Suda, E., Rotimi, C.N., Adebamowo, C.A., Ajayi, I., Aniagwu, T., Marshall, P.A., Nkwodimmah, C., Royal, C.D., Leppert, M.F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I.F., Knoppers, B.M., Foster, M.W., Clayton, E.W., Watkin, J., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G.M., Yakub, I., Birren, B.W., Wilson, R.K., Fulton, L.L., Rogers, J., Burton, J., Carter, N.P., Clee, C.M., Griffiths, M., Jones, M.C., McLay, K., Plumb, R.W., Ross, M.T., Sims, S.K., Willey, D.L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J.C., L'Archeveque, P., Bellemare, G., Saeki, K., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A.L., Brooks, L.D., McEwen, J.E., Guyer, M.S., Wang, V.O., Peterson, J.L., Shi, M., Spiegel, J., Sung, L.M., Zacharia, L.F., Collins, F.S., Kennedy, K., Jamieson, R. and Stewart, J. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913-918.
- Schmid, K.J., Sorensen, T.R., Stracke, R., Torjek, O., Altmann, T., Mitchell-Olds, T. and Weisshaar, B. (2003) Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Res*, **13**, 1250-1257.
- Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A.H., Nieldsen, K.L., Jorgenson, J., Weigel, D. and Uggerho, S. (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nature Methods*, **6**, 550-551.
- Simon, M., Loudet, O., Durand, S., Berard, A., Brunel, D., Sennesal, F.-X., Durand-Tardif, M., Pelletier, G. and Camilleri, C. (2008) Quantitative Trait Loci Mapping in Five New Large Recombinant Inbred Line Populations of *Arabidopsis thaliana* Genotyped With Consensus Single-Nucleotide Polymorphism Markers. *Genetics*, **178**, 2253-2264.
- Singer, T., Fan, Y., Chang, H.-S., Zhu, T., Hazen, S.P. and Briggs, S.P. (2006) A High-Resolution Map of *Arabidopsis* Recombinant Inbred Lines by Whole-Genome Exon Array Hybridization. *PLoS Genet*, **2**, e144.
- Staal, J., Kaliff, M., Dewaele, E., Persson, M. and Dixelius, C. (2008) RLM3, a TIR domain encoding gene involved in broad-range immunity of *Arabidopsis* to necrotrophic fungal pathogens. *Plant Journal*, **55**, 188-200.

- Thiele, K., Wanner, G., Kindzierski, V., Jurgens, G., Mayer, U., Pahl, F. and Assaad, F.F.** (2009) The timely deposition of callose is essential for cytokinesis in Arabidopsis. *Plant Journal*, **58**, 13-26.
- Toomajian, C., Hu, T.T., Aranzana, M.J., Lister, C., Tang, C., Zheng, H., Zhao, K., Calabrese, P., Dean, C. and Nordborg, M.** (2006) A nonparametric test reveals selection for rapid flowering in the Arabidopsis genome. *PLoS Biol*, **4**, e137.
- Valentin, H.E., Lincoln, K., Moshiri, F., Jensen, P.K., Qi, Q., Venkatesh, T.V., Karunanandaa, B., Baszis, S.R., Norris, S.R., Savidge, B., Gruys, K.J. and Last, R.L.** (2006) The Arabidopsis vitamin E pathway gene5-1 mutant reveals a critical role for phytol kinase in seed tocopherol biosynthesis. *Plant Cell*, **18**, 212-224.
- Van Eenennaam, A.L., Lincoln, K., Durrett, T.P., Valentin, H.E., Shewmaker, C.K., Thorne, G.M., Jiang, J., Baszis, S.R., Levering, C.K., Aasen, E.D., Hao, M., Stein, J.C., Norris, S.R. and Last, R.L.** (2003) Engineering vitamin E content: from Arabidopsis mutant to soy oil. *Plant Cell*, **15**, 3007-3019.
- Wawrzynska, A., Christiansen, K.M., Lan, Y., Rodibaugh, N.L. and Innes, R.W.** (2008) Powdery Mildew Resistance Conferred by Loss of the ENHANCED DISEASE RESISTANCE1 Protein Kinase Is Suppressed by a Missense Mutation in KEEP ON GOING, a Regulator of Abscisic Acid Signaling. *Plant Physiology*, **148**, 1510-1522.
- West, M.A.L., van Leeuwen, H., Kozik, A., Kliebenstein, D.J., Doerge, R.W., St. Clair, D.A. and Michelmore, R.W.** (2006) High-density haplotyping with microarray-based expression and single feature polymorphism markers in Arabidopsis. *Genome Res.*, **16**, 787-795.
- Wright, S.I., Bi, I.V., Schroeder, S.G., Yamasaki, M., Doebley, J.F., McMullen, M.D. and Gaut, B.S.** (2005) The effects of artificial selection on the maize genome. *Science*, **308**, 1310-1314.
- Xin, Z., Mandaokar, A., Chen, J., Last, R.L. and Browse, J.** (2007) Arabidopsis ESK1 encodes a novel regulator of freezing tolerance. *Plant J*, **49**, 786-799.

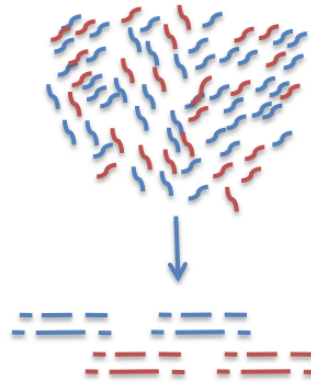
Figure Legends:

Figure 1: A schematic representation of the process by which putative Col-Ler polymorphisms were identified. (A): The publicly funded Arabidopsis Genome Initiative sequenced the Col ecotype in a stepwise BAC-by-BAC manner (AGI, 2000). Each clone was sequenced and assembled independently, and combined to form the high quality reference genome sequence. (B): Cereon Genomics used a whole genome shotgun approach to sequence the Ler ecotype. Shotgun sequence reads from the entire genome were assembled to form short contigs of overlapping sequence. The two collections of sequences were then aligned to each other at high stringency to identify either (C) putative single nucleotide polymorphisms or (D) putative insertion/deletions. The entire collection of predicted polymorphisms was provided to the not-for-profit research community through the TAIR database (Jander *et al.*, 2002; Rounsley, 2003).

(A) Columbia (Col)
BAC-by-BAC



(B) Landsberg *erecta* (Ler)
Whole genome shotgun



Col TTTTGAG CTCGCTAC AATGATTGATAAGTTAAGTCCTAT GAGAGAGA TCGACGCATCAGC
Ler TTTTGAG CTCGTTAC AATGAT AAGTCCTAT GAGA-GA TCGACGCATCAGC

(C)

(D)