Natural Variation in Arabidopsis. How Do We Find the Causal Genes?

Detlef Weigel* and Magnus Nordborg

Department of Molecular Biology, Max Planck Institute for Developmental Biology, D–72076 Tuebingen, Germany (D.W.); Plant Biology Laboratory, Salk Institute for Biological Studies, La Jolla, California 92037 (D.W.); and Molecular and Computational Biology Program, University of Southern California, Los Angeles, California 90089 (M.N.)

According to PubMed, more than one-third of all papers on natural genetic variation in Arabidopsis (*Arabidopsis thaliana*) have been published since the beginning of 2004, underscoring the rapid growth of this area of Arabidopsis research. Motivations for studying variation found in wild strains range from simply exploiting it in order to find new genes involved in specific aspects of plant physiology or development, to trying to understand the molecular basis of adaptations to the local environment. However, irrespective of our ultimate goal, we first have to solve two more immediate problems. Which are the genes that affect variation in a specific trait? And, what is the nature of the allelic differences?

These are unfortunately difficult questions to answer. The main reason is that although major genetic differences that segregate as Mendelian factors are sometimes found (e.g. Bowman et al., 1993; Grant et al., 1995), alleles with major effects are an exception. Even drastic phenotypic differences are often due to allelic variation at several loci, and the contribution of each locus to the phenotype can be quite small. Thus, while the most direct approach toward identifying the causal genes is genetic mapping, one needs to use statistical methods to find regions of the genome associated with the trait of interest, an approach that is known as quantitative trait locus (QTL) mapping. Initial QTL mapping in an F_2, backcross, or recombinant inbred line (RIL) population is usually followed by the creation of near-isogenic lines, in which only one QTL region segregates in an otherwise identical genetic background. One can then apply conventional fine-mapping methods to identify the gene(s) responsible for the QTL and ultimately the causal changes at the nucleotide level (hence the term quantitative trait nucleotide). It goes without saying that this is a slow process.

Because phenotypic effects of individual genes are often small and because of complex interactions between the genes, it is generally not easy to obtain definitive proof for the identification of a QTL or quantitative trait nucleotide. For this reason, the mammalian quantitative genetics community (Members of the Complex Trait Consortium, 2003) has proposed several lines of experimentation that can be combined to provide evidence for identification of a QTL, once genetic linkage of a particular genomic region to a trait of interest has been established.

1. DNA polymorphisms that distinguish alleles with different phenotypic effects.
2. A mechanistic link between function of the gene and the trait of interest.
3. Functional studies showing that one allele has, for example, different biochemical properties.
4. Transgenic complementation.
5. Allele replacement through homologous recombination.
6. Deficiency complementation test (showing that one allele has a different phenotypic effect when in trans to a knockout of the QTL candidate).
7. Mutational analysis (demonstrating that a knockout of the gene affects the trait of interest).
8. Natural genetic variation at a homologous locus affecting the same trait in another species.

Obviously, it will be rare (and rarely necessary) that all of these criteria are fulfilled. In addition, some of the criteria will provide stronger evidence than others. Interestingly, unambiguous fine mapping, which has been achieved in a particularly heroic effort with tomato (*Lycopersicon esculentum*; Fridman et al., 2000), is not among this set of criteria, apparently because this is unrealistic in mammals. Short of mapping a QTL to a single gene, however, we agree that several of the criteria listed above need to be met in order to claim a causal link between allelic variation and variation at a trait.

As difficult as it is to demonstrate that a particular allelic variant is indeed causally related to the phenotype, the rate-limiting step typically remains the initial mapping and identification of the QTL. Several types of whole-genome analyses across a large panel of wild strains have been proposed as shortcuts. One is the direct identification of genes whose expression is correlated with a trait. Another is the identification of sequence variants that correlate with a particular phenotype, so-called linkage disequilibrium (LD) or

* Corresponding author; e-mail weigel@weigelworld.org; fax 49–7071–601–1410.

www.plantphysiol.org/cgi/doi/10.1104/pp.104.900157.

association mapping. Because LD typically decays rapidly in Arabidopsis (over 25–50 kb; Nordborg et al., 2005), sequence variants identified through this approach are expected to be very closely linked to the QTL. In the canonical example from human genetics, Hästbacka and colleagues (Hästbacka et al., 1992) used LD mapping to refine the location of the gene responsible for diastrophic dysplasia (a Mendelian disease in humans) from over 1 Mb to about 60 kb. We have demonstrated that the *FRIGIDA* gene, which had previously been shown to underlie natural genetic variation in flowering response of Arabidopsis (Johanson et al., 2000), could have been mapped to within about 30 kb by marker association (Hagenblad et al., 2004).

Panels for genome-wide association scans are becoming available through sequencing efforts targeting thousands of regions across the entire genome (Nordborg et al., 2005) and through DNA array hybridization studies (Borevitz et al., 2003). In addition, we are currently undertaking an effort to discover at least half of all single nucleotide polymorphisms in 20 different Arabidopsis strains, exploiting the same DNA array hybridization technology that has been used to assay human genetic variation (Hinds et al., 2005). With this information in hand, one can then genotype hundreds of Arabidopsis strains for many (i.e. thousands to hundreds of thousands) of markers, which will greatly increase the power of association studies. Because strains are naturally inbred, Arabidopsis is uniquely suited for these kinds of studies: once genotyped, a collection of strains can be repeatedly assayed for many different phenotypes.

However, it is important to remember that LD mapping does not provide direct evidence. A confounding factor in these analyses is that marker-trait associations may not be due to causal relationships (or even linkage), but rather due to an unexpected statistical association between a (unknown) causal gene and other genes. This would be the case, for example, if strains that show a certain phenotype are more likely to be related to each other (genome wide, as opposed to with respect to the causative loci) than expected by chance. This is known as population structure and may occur because strains from the same geographic area, where they may be exposed to similar environments, are also often more related than those that come from geographically distant regions. Genome-wide marker information allows us to gain insight into the severity of this problem. When many more markers than expected are identified as significantly associated with a trait, one can safely assume that many of these associations are not due to causal marker-trait relationships. Statistical methods that help to correct for this problem have been proposed (e.g. Devlin and Roeder, 1999; Pritchard et al., 2000), but they will not bypass the need for experimental studies.

The true power of association studies lies in the combination with other approaches, especially the analysis of experimental populations. If association studies point to alleles with opposite effects on a trait of interest, one can quickly generate multiple F_2 populations from parents that harbor contrasting alleles and determine whether differences in phenotype cosegregate with the locus in question. Even more powerful will be the use of RILs, which are immortalized F_2 populations that need to be genotyped only once but can be repeatedly phenotyped (just like natural strains). Some 50 RIL sets have been produced or are currently in production in different laboratories (www.inra.fr/qtlat/NaturalVar/RILSummary.htm), and these will be an invaluable resource not only for conventional QTL mapping but also in combination with association studies. Expression studies can also provide valuable information by further narrowing the list of candidate genes likely to be causally linked to the trait.

In summary, we feel that the plant genetic community would do well to adopt a set of standards for what does and what does not constitute acceptable proof that naturally occurring genetic variants cause differences in phenotype. Once we have solved this problem, we can turn to the difficult questions, such as the following. Is an unusual allele simply a deleterious variant that has become fortuitously fixed in the population? Or, does this allele indeed provide a fitness advantage in the local environment?

LITERATURE CITED

Borevitz JO, Liang D, Plouffe D, Chang H-S, Zhu T, Weigel D, Berry CC, Winzeler E, Chory J (2003) Large-scale identification of single-feature polymorphisms in complex genomes. Genome Res **13:** 513–523

Bowman JL, Alvarez J, Weigel D, Meyerowitz EM, Smyth DR (1993) Control of flower development in *Arabidopsis thaliana* by *APETALA1* and interacting genes. Development **119:** 721–743

Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics **55:** 997–1004

Fridman E, Pleban T, Zamir D (2000) A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. Proc Natl Acad Sci USA **97:** 4718–4723

Grant MR, Godiard L, Straube E, Ashfield T, Lewald J, Sattler A, Innes RW, Dangl JL (1995) Structure of the *Arabidopsis RPM1* gene enabling dual specificity disease resistance. Science **269:** 843–846

Hagenblad J, Tang C, Molitor J, Werner J, Zhao K, Zheng H, Marjoram P, Weigel D, Nordborg M (2004) Haplotype structure and phenotypic associations in the chromosomal regions surrounding two *Arabidopsis thaliana* flowering time loci. Genetics **168:** 1627–1638

Hästbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E (1992) Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nat Genet **2:** 204–211

Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, Frazer KA, Cox DR (2005) Whole-genome patterns of common DNA variation in three human populations. Science **307:** 1072–1079

Johanson U, West J, Lister C, Michaels S, Amasino R, Dean C (2000) Molecular analysis of *FRIGIDA*, a major determinant of natural variation in *Arabidopsis* flowering time. Science **290:** 344–347

Members of the Complex Trait Consortium (2003) The nature and identification of quantitative trait loci: a community's view. Nat Rev Genet **4:** 911–916

Nordborg M, Hu T, Ishino Y, Toomajian C, Jhaveri J, Bakker E, Calabrese GJ, Goyal RP, Jakobsson M, Jhaveri J, et al (2005) The genomic pattern of polymorphism in *Arabidopsis thaliana.* PLoS Biol **3:** e196

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet **67:** 170–181