Arabidopsis Map-Based Cloning in the Post-Genome Era

Georg Jander*, Susan R. Norris, Steven D. Rounsley, David F. Bush, Irena M. Levin[1], and Robert L. Last

Cereon Genomics LLC, 45 Sidney Street, Cambridge, Massachusetts 02139

Map-based cloning is an iterative approach that identifies the underlying genetic cause of a mutant phenotype. The major strength of this approach is the ability to tap into a nearly unlimited resource of natural and induced genetic variation without prior assumptions or knowledge of specific genes. One begins with an interesting mutant and allows plant biology to reveal what gene or genes are involved. Three major advances in the past 2 years have made map-based cloning in Arabidopsis fairly routine: sequencing of the Arabidopsis genome, the availability of more than 50,000 markers in the Cereon Arabidopsis Polymorphism Collection, and improvements in the methods used for detecting DNA polymorphisms. Here, we describe the Cereon Collection and show how it can be used in a generic approach to mutation mapping in Arabidopsis. We present the map-based cloning of the *VTC2* gene as a specific example of this approach.

Map-based cloning, also called positional cloning, is the process of identifying the genetic basis of a mutant phenotype by looking for linkage to markers whose physical location in the genome is known. The amount of effort required for map-based cloning of genes in Arabidopsis has dropped dramatically in recent years (Fig. 1). Only a few years ago, it was necessary to build a physical map, develop markers, and iteratively zero-in on the gene by “chromosome walking.” This was followed by cloning, complementation by transformation, and de novo determination of the sequence of the entire region of interest to high quality without a previously determined wild-type DNA sequence as a guide (Arondel et al., 1992; Giraudat et al., 1992; Leung et al., 1994; Meyer et al., 1994; Mindrinos et al., 1994).

Many of the steps of chromosome walking have been eliminated or have been made much easier by three nearly simultaneous breakthroughs during the past 2 years: sequencing of the entire Columbia (Col-0) Arabidopsis genome (The Arabidopsis Genome Initiative, 2000), the availability of tens of thousands of randomly distributed genetic markers to registered users of the Cereon Arabidopsis Polymorphism Collection (http://www.arabidopsis.org/cereon/), and advances in the methods used to detect DNA polymorphisms. One can now proceed from a mutant with a desirable phenotype to an identified mutation in a gene with less than one person-year of effort (Fig. 1). The minimal start-to-finish time of a mapping project has also been shortened significantly, making it possible to find a gene using an iterative approach taking approximately 1 year (Fig. 2).

In the process of map-based cloning, one starts with a mutant and eventually identifies the gene responsible for the altered phenotype, allowing the plant to tell you what genes are important in the physiological process of interest. This is in contrast to reverse genetic approaches, which tend to rely on some sort of prior knowledge that the gene that is being mutated will be interesting. When using reverse genetic approaches, such as tilling for point mutations (McCallum et al., 2000) or searching for T-DNA insertion mutations (Sussman et al., 2000), one starts with a gene of interest, finds a mutation in that gene, and then looks for a phenotype.

The big advantage to map-based cloning is that it is a process without prior assumptions. Essentially, one is looking at all of the genes in the genome at the same time to find the ones that affect the phenotype of interest. It is a process of discovery that makes it possible to find mutations anywhere in the genome, including intergenic regions and the 40% of Arabidopsis genes that do not resemble any gene with known or inferred function (The Arabidopsis Genome Initiative, 2000).

Insertional mutagenesis using T-DNA or transposons has become increasingly popular as a tool for gene discovery. Pools of lines representing more than 200,000 insertional mutations are available from Arabidopsis stock centers (http://www.Arabidopsis.org/abrc; http://nasc.nott.ac.uk). Large-scale projects are under way for disrupting most genes in Arabidopsis by insertional mutagenesis (Sussman et al., 2000). Mutant screens performed using these populations are undoubtedly worthwhile and can lead to rapid identification of the gene of interest if it is actually has a T-DNA or transposon insertion. However, there are also several good reasons to screen for mutants in chemically mutagenized populations and to isolate the affected genes by map-based cloning.

Insertional mutations tend to result in complete knockouts of the gene, making it difficult to associate a phenotype other than death with essential genes. In contrast, chemical mutagenesis, e.g. with ethyl methane sulfonate, can produce promoter mutations or

[1] Present address: Syrrx Inc., 10410 Science Center Drive, San Diego, CA 92121.

* Corresponding author; e-mail georg.jander@cereon.com; fax 617–551–1950.

www.plantphysiol.org/cgi/doi/10.1104/pp.003533.

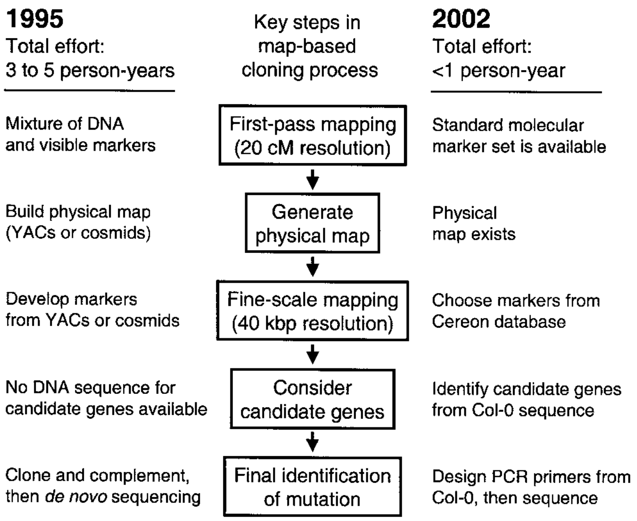

Figure 1. Comparison of effort involved in map-based cloning. Key steps that have become easier between 1995 and 2002 are presented.

mis-sense mutations in the coding region, resulting in a hypomorphic knock-down rather than an amorphic knockout of a protein function. Many interesting but essential genes have been found through such hypomorphic mutations. For instance, "leaky" mutations in *VTC1* (*CYT1*) can result in ozone sensitivity and reduced vitamin C levels in Arabidopsis (Conklin et al., 1999), but knockout mutations cause embryo lethality (Lukowitz et al., 2001). Key regulatory steps in biochemical pathways are often found through dominant point mutations that prevent feedback inhibition of an enzyme, e.g. anthranilate synthase (Kreps et al., 1996; Li and Last, 1996) or Asp kinase (Heremans and Jacobs, 1997). Such dominant mutations would not be found by insertional mutagenesis.

Chemical mutagenesis, in addition to generating a greater diversity of mutations than insertional mutagenesis, also results in many more mutations in each individual plant. Plants mutagenized with T-DNA typically have only one to three insertions per line. Even in a best-case scenario (insertion of three T-DNAs per line in a completely random manner, which is not likely), more than 100,000 plants are needed for a 95% likelihood of having a mutation in a given gene of average size. Screening this many plants can be prohibitive if the mutant screen being performed is laborious or slow. In contrast, ethyl methane sulfonate mutagenesis typically introduces dozens of mutations in each plant line, and it is generally possible to find a mutation in any given gene by screening fewer than 5,000 plants (Feldman et al., 1994).

The techniques of map-based gene identification are also essential for the identification of the genetic basis of phenotypic variation among Arabidopsis ecotypes (natural isolates). The genomes of Arabidopsis ecotypes differ from one another at many thousands of locations and represent a level of genetic variation that is not achievable in the laboratory (Alonso-Blanco and Koornneef, 2000). Hundreds of ecotypes collected from around the world are available to researchers through Arabidopsis stock centers (http://www.Arabidopsis.org/abrc; http://nasc.nott.ac.uk). Phenotypic variation for almost any trait of interest can be found in progeny of crosses made between these ecotypes. In many cases this variation is due to the effects of several genes and is quantitative in nature. Statistical methods developed in the 1990s (Haley and Knott, 1992; Jansen, 1993; Zeng, 1994) and the availability of an almost unlimited set of genetic markers (see below) make it feasible to map and clone such quantitative trait loci (QTL). We will not describe QTL mapping here, but other recent reviews have covered this subject (Kearsey and Farquhar, 1998; Alonso-Blanco and Koornneef, 2000; Yano, 2001).

In this paper, we present a large set of DNA markers identified at Cereon Genomics, we describe how these markers can be applied to a generic map-based cloning project, and we introduce the *VTC2* gene as an example of a specific mapping project.

THE CEREON ARABIDOPSIS POLYMORPHISM COLLECTION

Positional cloning of genes in Arabidopsis is greatly facilitated by the recent sequencing of Col-0 and Landsberg *erecta* (L*er*). These two ecotypes were sequenced because they are among the most commonly used ecotypes in Arabidopsis research. George Redei, one of the founders of modern Arabidopsis genetics, began working with Col and L*er* in the 1950s (Redei, 1992). Since then, they have been the subjects of literally thousands of papers that have been published on the genetics, molecular biology, and biochemistry of Arabidopsis. Col-0 and L*er* are also the parents of a widely used collection of recombinant inbred lines (Lister and Dean, 1993). Hundreds of markers have been analyzed in these lines, and the genetic map produced from this work has become the standard against which other Arabidopsis genetic maps are aligned.

The Col-0 ecotype was the subject of a large international sequencing project, which has produced a nearly complete sequence using a clone by clone approach (The Arabidopsis Genome Initiative, 2000). This high-quality sequence (less than one error in 10,000 bp) is a permanent resource for all future Arabidopsis sequencing efforts. Partial genomic sequence data generated from other ecotypes can be positioned on the framework of Col-0 genome sequence. Sequencing of individual genes from mutants or from other ecotypes has become routine; it is simply a matter of designing PCR primers based on the Col-0 sequence, amplifying the desired gene, and sequencing the product.

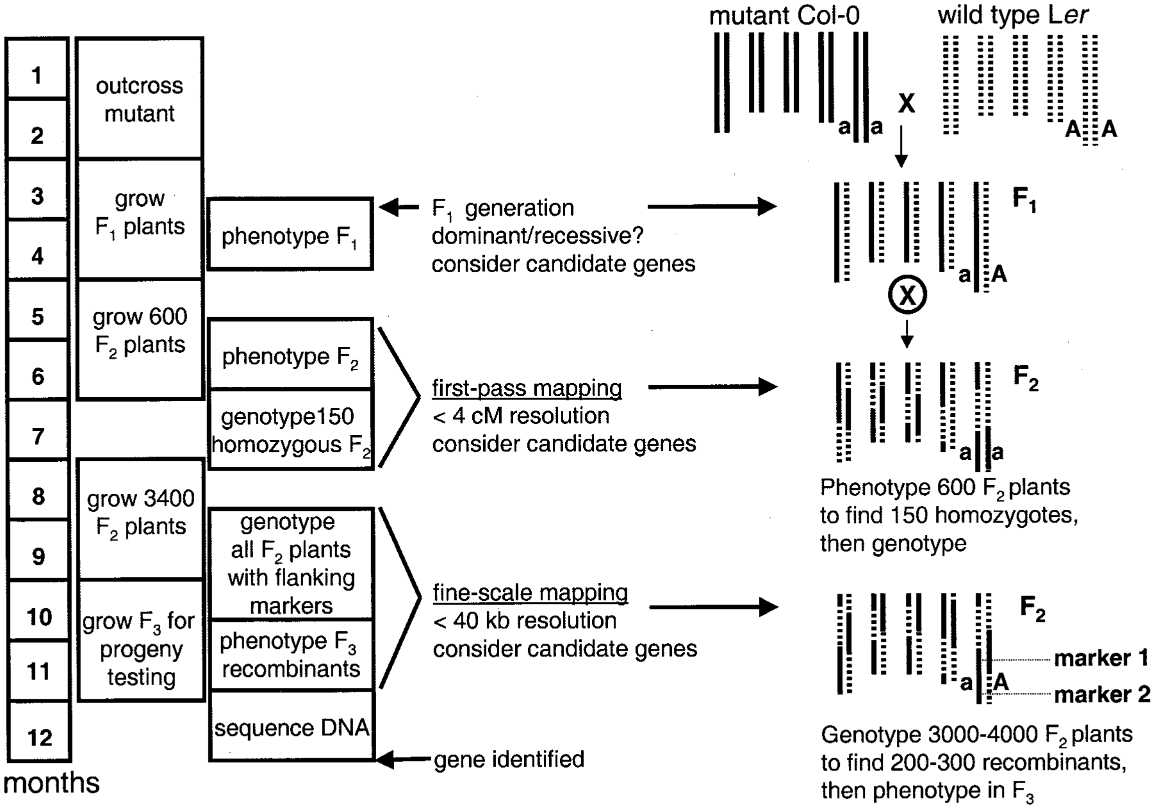
months
1
2
3
4
5
6
7
8
9
10
11
12
outcross mutant
grow F1 plants
grow 600 F2 plants
grow 3400 F2 plants
grow F3 for progeny testing
phenotype F1
phenotype F2
genotype150 homozygous F2
genotype all F2 plants with flanking markers
phenotype F3 recombinants
sequence DNA
F1 generation dominant/recessive? consider candidate genes
first-pass mapping < 4 cM resolution consider candidate genes
fine-scale mapping < 40 kb resolution consider candidate genes
gene identified
mutant Col-0
wild type Ler
a
A
X
F1
F2
Phenotype 600 F2 plants to find 150 homozygotes, then genotype
marker 1
marker 2
Genotype 3000-4000 F2 plants to find 200-300 recombinants, then phenotype in F3

Figure 2. Schematic of the map-based cloning process. Left, Typical 1-year mapping timeline for a mutation whose phenotype can be measured as the plants are growing. Right, Schematic of the five pairs of Arabidopsis chromosomes during critical stages of a sample mapping of a recessive mutation on chromosome 5 in the Col-0 background.

The L*er* ecotype was the subject of a very different genome sequencing effort, low coverage shotgun sequencing at Cereon Genomics. This project generated approximately 700,000 500-bp sequence traces. Of these, more than 200,000 were chloroplast, mitochondrial, or ribosomal DNA and were not used for the assembly. This left 498,037 traces totaling 263 Mbp of good quality raw sequence, representing approximately 2-fold coverage of the Arabidopsis genome. Assembly of the sequences produced 50,262 contigs (average size, 1.5 kb) and 31,044 single-read sequences. The size of the assembled dataset totaled 92.1 Mbp, suggesting that approximately 70% of the genome is covered at the nucleotide level. To assess the coverage at the gene level, more than 2,000 cDNA sequences from GenBank were extracted and searched against the L*er* shotgun dataset using the BLASTn algorithm (Altschul et al., 1990). A total of 96.5% of the cDNAs were at least partially detected using a 95% identity cutoff, indicating that at least some sequence from over 95% of all genes is present in the data assembled from the low coverage shotgun approach.

For Arabidopsis researchers who are interested in map-based cloning, the value of two genome sequences greatly exceeds that of only one such sequence. Whereas the availability of the genome sequence of a single ecotype mainly facilitates DNA sequencing in the final stages of a mapping project (Fig. 1), data from two genomes make it possible to develop a database of DNA polymorphisms that can be used as genetic markers. A high-density map of DNA markers greatly facilitates fine-scale genetic mapping. To generate such a map, we compared stretches of L*er* shotgun sequence with Col-0 genomic sequence determined from cloned bacterial artificial chromosomes (BACs; we will refer to all large DNA clones sequenced by the Col-0 genome project collectively as BACs). Differences between the ecotypes were classified into two types: single nucleotide polymorphism (SNP) changes, which alter a single nucleotide present at specific location in the genome (Fig. 3), and insertion-deletion (InDel) differences, where one ecotype has an insertion of a number of nucleotides relative to the other (Fig. 3).

To detect SNPs and InDels, one must be able to accurately predict true polymorphisms against a background of sequencing errors. This is of particular concern for the L*er* data, which are unedited shotgun sequence, in contrast to the high quality "finished" Col-0 sequence. To increase the likelihood of detecting real ecotypic differences, fairly stringent criteria were applied to a single base difference before calling it a bioinformatically predicted SNP. The aligned

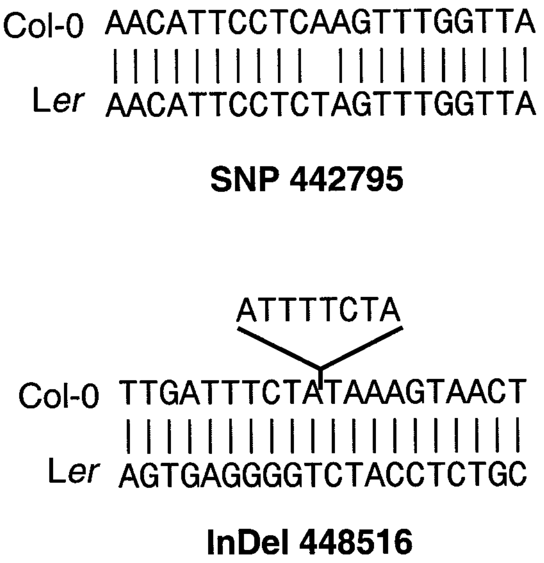
Col-0 AACATTCCTCAAGTTTGGTTA
Ler AACATTCCTCTAGTTTGGTTA
SNP 442795
ATTTTCTA
Col-0 TTGATTTCTATAAAGTAACT
Ler AGTGAGGGGTCTACCTCTGC
InDel 448516

Figure 3. Examples of SNP and InDel polymorphisms. Two markers from the Cereon Arabidopsis Polymorphism Collection are shown. Marker 442795 has a single-nucleotide change from A to T, whereas marker 448516 has an eight-nucleotide insertion in Col-0 versus L*er*.

region of L*er* and Col-0 sequence had to be longer than 200 bp and to include more than 75% of the length of the L*er* sequence. In addition, the polymorphic base must be unambiguous in L*er*, covered by at least two reads, and be greater than 25 bp from any single coverage region. The quality of the local sequence must be high: The SNP-containing base must have a phrap consensus quality score (Green, 1996, Version 0.980812, downloaded 1999) of at least 40, and the surrounding 25 nucleotides must have consensus scores of at least 30. Re-sequencing of the L*er* allele of a representative sample of SNPs predicted in this way showed that the success rate was close to 100%. Single-basepair InDels were found using the same methods as those used for SNP prediction. Less stringent criteria were applied for the detection of larger InDels. A gapped alignment between L*er* and Col-0 was required to be greater than 90% identical over the matched region, with an insertion of at least 2 bp in either Col-0 or L*er*. Unlike with SNP polymorphisms, we did not confirm a representative sample of predicted InDels by resequencing the L*er* allele. Given the less stringent selection criteria, the error rate for predicted InDel polymorphisms is likely to be higher than the error rate for predicted SNP polymorphisms.

At the time of writing, sequence for 1,501 Col-0 BACs representing 123 Mbp of Col-0 genome sequence had been compared against the assembled L*er* shotgun sequence. This resulted in the identification of 37,344 SNPs, 18,579 small InDels (less than or equal to 100 bp), 747 large InDels (larger than 100 bp), or a total of 56,670 polymorphisms. On average, there is one bioinformatically predicted SNP every 3.3 kb and one predicted InDel every 6.6 kb. The SNPs and InDels are distributed throughout the genome, with most BACs having several polymorphisms that could be used for genetic mapping (Fig. 4). Because of the stringent selection criteria and the partial L*er* sequence, these numbers represent an underestimate of the true frequency of SNP and InDel differences that exist. For instance, a screen of 500 kb of Arabidopsis sequence by denaturing HPLC (DHPLC) found polymorphisms at a frequency of close to one per kilobasepair (Cho et al., 1999). The Cereon Arabidopsis Polymorphism Collection is made available to registered users at non-profit and educational institutions for non-commercial research. Access is obtained by one-time registration through The Arabidopsis Information Resource Web site (http://www.arabidopsis.org/cereon/). At the time

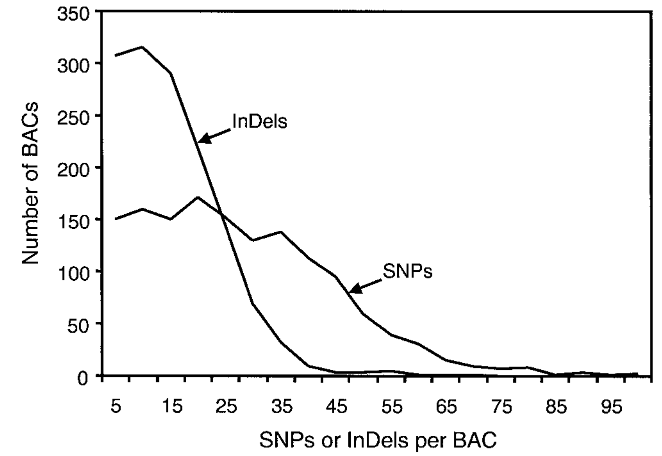

Figure 4. Frequency of SNPs and InDels by BAC. A total of 56,668 SNP and InDel polymorphisms between Col-0 and L*er* were identified. These polymorphisms were assigned to 1,501 sequenced Col-0 BAC clones (The Arabidopsis Genome Initiative, 2000). Data are presented as bins of 5, i.e. 1 to 5 polymorphisms/BAC, 6 to 10 polymorphisms BAC, etc. Nineteen BACs have no predicted InDel or SNP polymorphisms.

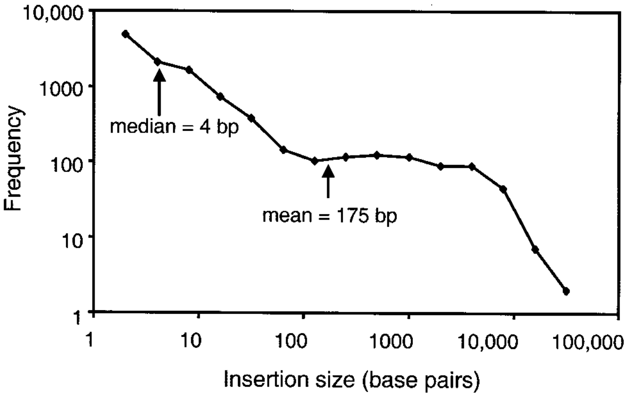

Figure 5. Insertions in Col-0. A total of 10,578 insertions in Col-0 relative to L*er* were identified. Insertion size data are presented as bins of 0.3 $\log_{10}$(no. of basepairs), i.e. $\log_{10}$(no. of basepairs) < 0.3, 0.3 < $\log_{10}$(no. of basepairs) < 0.6, etc. The median (4 bp) and mean (175 bp) insertion sizes are indicated.

of writing, 890 researchers from 40 countries had registered to use this database.

The five chromosomes of Arabidopsis have approximately equal densities of SNP polymorphisms. Not surprisingly, SNP frequency varies between exons and introns, with one SNP every 3.1 and 2.2 kb, respectively. Transitions (A/T to G/C) account for 52.8% of the SNPs, and transversions occur with frequencies of 17.4% (A/T to T/A), 23.0% (T/A to G/C), and 7.9% (C/G to G/C). There is no Col-0 or L*er* bias in the directionality of the transitions or transversions.

InDel polymorphisms between Col-0 and L*er* range from 1 bp to greater than 38 kb. Due to the average 1.5-kb contig size of the L*er* shotgun sequence, large insertions can only be detected in the Col-0 background and not in the L*er* background. Insertions in Col-0 relative to L*er* have an average size of 175 bp and a median size of 4 bp (Fig. 5). Approximately 10% of the InDels were associated with polymorphisms in the length of simple sequence repeats that were identified with the Sputnik program (Abajian, 1994, downloaded 1999), but most were found in non-repetitive sequences. Most InDels (93%) are smaller than 100 bp, making them suitable for PCR-based marker detection methods (see below).

The Cereon Col-0/L*er* SNPs and InDels sequences should be very informative for discovering polymorphisms between other ecotype pairs. If one assumes a random genetic reassortment of polymorphisms among Arabidopsis ecotypes, then 50% of the Col-0/L*er* polymorphisms should be useful for genetic mapping in any other pair of ecotypes. Work done with amplified fragment length polymorphism (AFLP) markers, which generally are due to underlying SNPs, indicates that there is such a random assortment of polymorphisms. Approximately 50% of Col-0/L*er* AFLP polymorphisms can also be used for segregation analysis in Col-0/C24, Col-0/Wassilewskij, or Col-0/Cape Verde Islands crosses (Peters et al., 2001). Analysis of 79 AFLP markers in 142 ecotypes shows a high degree of recombination in the evolution of these ecotypes, such that it is not possible to draw an “ecotype phylogeny” (Sharbel et al., 2000). Thus, the Cereon Arabidopsis Polymorphism Collection will be useful for mapping QTLs or mutations in most and perhaps all other pairs of Arabidopsis ecotypes. It is a relatively minor disadvantage that one-half of all attempted markers will fail and the average marker density is reduced by 50%, i.e. one SNP every 6.6 kb instead of one SNP every 3.3 kb.

Overall, the density of both SNP and InDel markers is high enough that it is theoretically possible to map most mutations within a few thousand basepairs using either type of marker in any combination of ecotypes. The availability of genetic markers is no longer the limiting factor for the fine-scale genetic mapping needed for map-based cloning in Arabidopsis. Instead, this process is limited by our ability to generate recombination events at a high enough density and to rapidly and inexpensively genotype plants using these markers.

METHODS FOR DETECTION OF DNA POLYMORPHISMS

A critical aspect of map-based cloning is the ability to accurately detect DNA markers at an appropriate cost and throughput. In the past few years, a number of new technologies for high-throughput detection of DNA polymorphisms have been developed. Most of these advances were driven by the field of human genetics, but all of the methods can be applied equally well to plant systems. Because they tend to require a relatively large initial investment, these fast and highly automatable methods are best suited to research settings where large numbers of genotypes need to be determined in a short period of time and with minimal human intervention.

Because SNPs are more common than InDels in biology and are more amenable to automation strategies, most high-throughput genotyping approaches are designed for SNP rather than InDel detection. Oligonucleotide arrays (Gene Chips) contain thousands of oligonucleotides annealed to a glass slide. Such arrays allow the detection of SNP polymorphisms by differential hybridization in a highly parallel and automated manner (Lipshutz et al., 1999). The Taq-Man PCR assay is designed to detect SNPs in a high-throughput manner through the release of fluorescent reporter dye from a quencher on the same oligonucleotide by 5′ nuclease activity (Livak, 1999). By using more than one reporter dye, it is possible to detect different alleles of a SNP in a single reaction. The relatively high price of oligonucleotides tagged with reporter and quencher dyes makes this method cost-effective only if a large number of reactions need to be run with each SNP marker. In pyrosequencing, an enzymatic cascade and luminometric detection system is used to measure the pyrophosphate that is released as a result of nucleotide incorporation (Ahmadian et al., 2000; Alderborn et al., 2000). Because 20 or more nucleotides are determined by this method, it is possible to detect several closely linked SNPs at once. The pyrosequencing method can be automated but has the disadvantage that it does not work well on stretches of repeated nucleotides. DHPLC allows the detection of SNPs through different retention time of heteroduplex and homoduplex DNA in reversed-phase HPLC under partially denaturing conditions (Spiegelman et al., 2000). DHPLC allows detection of SNP polymorphisms in PCR-amplified DNA up to about 1,000 bp in size. Although not inherently high-throughput, DHPLC lends itself nicely to bulked segregant analysis. The method of fluorescence resonance energy transfer combines PCR and oligonucleotide ligation to detect SNPs (Chen et al., 1998). Dye-labeled oligonucleotide probes are used in this assay, and allele-specific ligation is detected by fluorescence resonance energy transfer, which only occurs when two dye-labeled oligos are joined by ligation. Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry can be used to rapidly detect SNPs in short DNA pieces by differences in molecular mass (Wada and Yamamoto, 1997).

A disadvantage of most high-throughput methods for detecting DNA polymorphisms is the high initial equipment cost, which results in a high per-assay cost for a lab that does not need to perform large numbers of genotyping reactions on a routine basis. In contrast, both InDels and SNPs can be detected using gel-based methods, which have a relatively low start-up cost and moderate throughput. InDels of small to moderate size can be detected by PCR amplification and gel electrophoretic separation (Bell and Ecker, 1994). Pairs of PCR primers are designed to amplify a segment of DNA spanning the InDel, and size differences in the amplified products are detected using either agarose or acrylamide gels. Agarose gels are easier and less expensive to use, but size differences of less than 5 bp are difficult to detect reliably. Acrylamide gels, on the other hand, give single-basepair resolution and allow the detection of even very small InDels. In either case, InDels are scored as codominant markers with one band seen on the gel for either homozygous class and two bands seen for heterozygous individuals. To reduce the number of PCR reactions needed for a mapping project, it is possible to pool DNA samples for bulked segregant analysis (Michelmore et al., 1991; Lukowitz et al., 2000), or multiple primer pairs can be added to one reaction tube to amplify several markers at once (Ponce et al., 1999).

Several gel electrophoresis-based strategies for detecting SNP markers have been devised. Many SNPs alter sites cleaved by restriction enzymes and can be used as cleaved-amplified polymorphic sequence (CAPS; Konieczny and Ausubel, 1993) markers. CAPS markers are amplified by PCR, the amplified DNA is cleaved with the appropriate restriction enzyme, and the cleavage products are examined on agarose gels. Just as with InDels, such markers are codominant, allowing the differentiation of heterozygotes and either homozygote class. If there is no suitable restriction site at a SNP, it is possible create a site during PCR amplification with suitably designed primers (dCAPS [Michaels and Amasino, 1998; Neff et al., 1998]). Disadvantages of using CAPS and dCAPS for genotyping include the extra time and cost involved in the restriction enzyme digestion and the possibility of a false result attributable to incomplete digestion by the restriction enzyme.

It is also possible to detect SNPs using allele-specific PCR primers, where the 3′ end of a primer has a perfect match with one allele and a mismatch with the other allele (Ugozzoli and Wallace, 1991). In theory, such primers can be used to preferentially amplify one allele of a SNP, but in practice a single-basepair change is often not enough to allow reliable differentiation between the two alleles of an SNP (Kwok et al., 1990; Cha et al., 1992). A modification of the allele-specific amplification procedure (single nucleotide amplified polymorphism [SNAP]) has recently been described (Drenkard et al., 2000). In this method, additional mismatches are introduced in the amplifying primers to maximize the difference in the amplification efficiencies of the two alleles of the SNP. Primer basepair changes that allow differential amplification of SNP sites can be predicted using the SNAPER program. Both the SNAPER program and a collection of primers that have been used successfully to amplify Arabidopsis SNAP markers can be found at http://patho.mgh.harvard.edu/ausubel-web. As with CAPS, SNAP markers are codominant and can be detected on agarose gels. However, it is necessary to run two PCR reactions—one for each allele of the SNP—to get complete SNAP genotyping data.

The detection of SNPs and InDels is an essential part of the map-based cloning process. Because marker discovery is no longer a problem in Arabidopsis, the selection of an efficient genotyping platform plays a critical role in the mapping timeline that we describe in the next section. We have mentioned several commonly used genotyping methods, and the choice of which method to apply will depend on the resources of an individual laboratory and the number of genotyping reactions that will need to be performed.

MAP-BASED CLONING PROCESS

Given a sequenced genome and a dense collection of genetic markers, map-based cloning becomes a relatively straightforward process. Figure 2 illustrates a time-efficient approach to map-based cloning in Arabidopsis, a variant of the "chromosome landing" method proposed by Tanksley et al. (1995). Starting with a mutation in the Col-0 or L*er* background, it is possible to proceed from having a mutant plant to identifying the affected gene in approximately 1 year. The overall length of this cloning process is dictated largely by the fact that it incorporates five cycles of plant growth (we assume 2 months/cycle).

As a first step in the mapping process, the mutant is out-crossed to the opposite ecotype (Col-0 or L*er*). In most cases, it is not necessary to "clean up" the genetic background of the mutant by back-crossing and it does not matter whether the mutant is used as the male or the female parent in the out-cross. F_1 seeds are planted and, as the plants are growing, it is possible to perform phenotype and genotype analysis. Presence or absence of the phenotype in the F_1 generation will suggest whether the mutation of interest is likely to be dominant or recessive. We recommend genotyping the F_1 plants with a few markers to make sure that they are truly heterozygous and that there was no mistake made during the cross. Similarly, it is worthwhile to genotype the original mutant to make sure that it is in the presumed ecotype background. Contamination with other ecotypes is a surprisingly frequent cause of "mutants" that arise in screens.

F_2 seeds are collected from self-pollination of the F_1 plants, and a population of approximately 600 individuals is planted for first-pass mapping (Fig. 2). As they are growing, the phenotype of the F_2 plants is determined, unless the trait can only be scored in the progeny (F_3) seed. It should be possible to identify approximately 150 plants in this population as homozygous: homozygous mutant in the case of a recessive mutation or homozygous wild type in the case of a dominant mutation. DNA for genotype analysis is prepared from the leaves or other tissue of these 150 plants. Initially, the 150 plants are genotyped with 25 markers, spaced roughly every 20 centiMorgan (cM) apart on the five chromosomes. Genetic linkage to one or more of the 25 markers is determined and a three-point cross is used to define a 20-cM interval that contains the gene of interest. Once a 20-cM interval has been found, additional markers are used to narrow down the region of interest to approximately 4 cM. Given a population of 150 plants, it should be possible to determine this 4-cM interval with a high degree of certainty. The two markers closest to the mutation on either side will be used as flanking markers in further work.

Next, it is necessary to plant a larger F_2 population for fine-resolution mapping (Fig. 2). The ultimate goal of fine mapping is to narrow down the region containing the gene of interest to 40 kb or less (approximately 0.16 cM genetic distance in Arabidopsis). There would ideally be several recombination events in this interval to define the position of the mutation that is being mapped. Unfortunately, the number of F_2 plants needed to have a 95% chance of recombination events in a given genetic interval increases rapidly as the size of the interval decreases (Fig. 6). We recommend having a fine mapping population of 3,000 to 4,000 plants (including the original 600 lines grown for first-pass mapping) to give a high probability of mapping the gene of interest to less than 40 kb. In areas of the genome with reduced meiotic recombination, e.g. near the centromeres, larger F_2 populations will be necessary to map a mutation to an equivalent physical interval on the chromosome. Many Arabidopsis mapping projects have been successful with fewer than 3,000 to 4,000 F_2 plants (Lukowitz et al., 2000), but when planting fewer plants one runs the risk of extending the mapping timeline by having to plant an additional F_2 population later on.

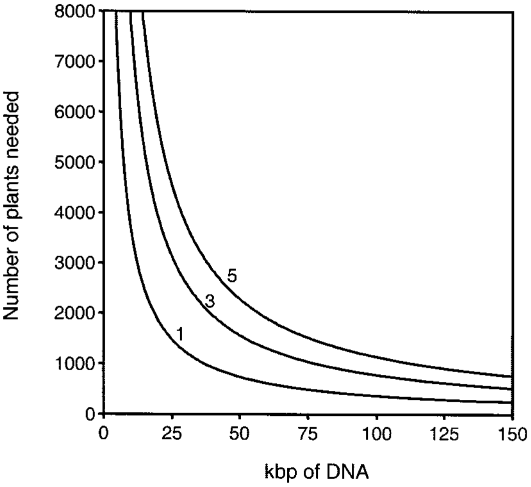

Figure 6. Number of plants needed to find recombinants. The curves show the number of F_2 plants needed to have a 95% chance of finding at least one plant (1), at least three plants (3), or at least five plants (5) with recombination events in a given physical interval of DNA. The calculations assume an average 250 kb/cM for Arabidopsis (Lukowitz et al., 2000). The possibility of multiple recombination events in one individual plant has a negligible effect and is not included in the calculations.

At this point, plants that are recombinant in the 4-cM interval determined by first-pass mapping are sought for use in fine mapping. DNA is isolated from the mapping population of 3,000 to 4,000 plants and the genotype of the two flanking markers is determined. This should identify 200 to 300 plants that have genetic recombination events in the region of interest (Fig. 2). The allelic state of the mutation being mapped (homozygous mutant, homozygous wild type, or heterozygous) in these recombinant plants is determined by looking at the phenotype in a representative sample of progeny in the F_3 generation. Additional markers in the 4-cM interval are used to look for increasingly tight linkage to the mutation. In most cases, it should be possible to define a pair of markers flanking the mutation that are less than 40 kb apart.

Once an interval of less than 40 kb containing the mutation of interest has been determined, this entire region is sequenced to find the mutation. In theory, it is possible to map a mutation to the single-gene level using the Cereon Arabidopsis Polymorphism Collection, but the number of F_2 plants needed to find recombinants in such a small interval would be very large (Fig. 6). It is faster and less expensive to sequence a larger interval. Because the sequence of the Col-0 genome is known, one efficient way to sequence the mutant region is to design PCR primers to amplify overlapping segments of about 500 bp spanning the entire 40 kb. These segments are then sequenced and assembled, the sequence is compared with that of a wild-type plant (Col-0 or L*er*), and the mutation is identified. In the case of a mutation in the L*er* background, it is necessary to also sequence the L*er* wild type for comparison at every location where a difference to the wild-type Col-0 is found. In the case of a mutation in Col-0, a published sequence is available. However, it is necessary to confirm that any nucleotide that diverges from the published Col-0 sequence was induced by the mutagenesis treatment and is not present in the wild-type progenitor strain. This is because strain to strain differences exist in "Col-0 wild type," and even at the high quality standard of the Col-0 sequence, sequencing errors are expected and found.

APPLICATION OF CEREON MARKERS TO CLONING *VTC2*

The identification of the *VTC2* gene is a specific example of a map-based cloning project using the Cereon Arabidopsis Polymorphism Collection. The *vtc2-1* mutation was isolated in a screen for ozone-sensitive mutants of Arabidopsis (Conklin et al., 1996). Further work showed that this mutant was deficient in ascorbic acid (vitamin C), and an additional three alleles (*vtc2-2*, *vtc2-3*, and *vtc2-4*) were isolated based on this phenotype. A first-pass map position for the *vtc2-1* mutation between CAPS markers WU95 (74 cM) and PRHA (78 cM) on chromosome 4 was reported (Conklin et al., 2000).

The CAPS markers WU95 and PRHA are relatively difficult to score. Instead, we used the nearby InDel markers (449235 and 450577 from the Cereon Arabidopsis Polymorphism Collection) as flanking markers for fine mapping (Fig. 7A). These markers are approximately 980 kb apart on chromosome 4. DNA segments spanning these markers were amplified by PCR, and the amplified products were detected by PAGE. A population of 3,700 Col-0 *vtc2-2* × L*er* F_2 plants was analyzed with markers 449235 and 450577. A total of 52 recombinants were identified and confirmed by repeating the genotyping with the same markers in the F_3 generation. The number of recombinants is considerably less than one would expect given the genetic separation previously reported for the CAPS markers WU95 and PRHA (4 cM apart, expected approximately 280 recombinants). We do not have a good explanation for this observation, but it does illustrate the utility of generating a mapping population that is larger than the theoretical minimum needed.

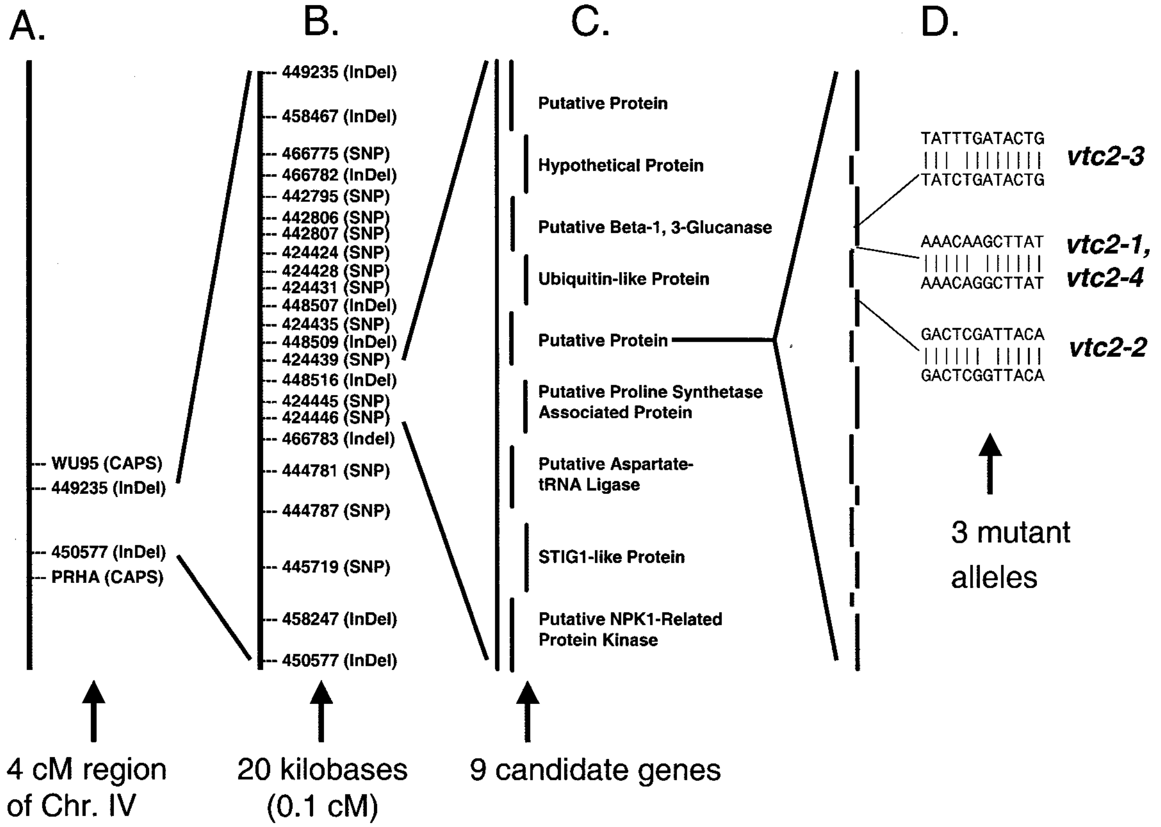

Figure 7. Map-based cloning of the *VTC2* gene. A, First-pass mapping of the *VTC2* identified flanking CAPS markers WU95 and PRHA (4 cM apart). B, Fine mapping of *VTC2* using SNP and InDel markers identified markers 424439 and 424446 in the Cereon Arabidopsis Polymorphism Collection (20 kb apart) as the closest flanking markers based on the available recombinants. C, Nine candidate genes between the SNP markers 424439 and 424446 were identified from the Col-0 sequence in GenBank. D, Mutations in *vtc2-1*, *vtc2-2*, *vtc2-3*, and *vtc2-4* were identified by sequencing. Staggered lines represent the predicted exons and introns of the *VTC2* gene. The 5′ end of the gene is at the bottom.

Additional markers between 449235 and 450577 were chosen from the Cereon Arabidopsis Polymorphism Collection (Fig. 7B) for fine mapping. All 52 recombinants were genotyped with these 21 markers to narrow down the positions of the recombination events. Pieces of DNA containing the marker of interest were amplified by PCR, and the polymorphisms were detected by PAGE (for InDels) or DNA sequencing (for SNPs). Vitamin C levels of individual F_3 progeny (at least 20 per line) were measured to determine whether the 52 F_2 recombinants were homozygous mutant, homozygous wild type, or heterozygous at the *VTC2* locus. This information was combined with the marker data to identify markers 424439 and 424446, which are contained in BAC F10 M23 (GI:4756963), as the closest markers flanking the mutation.

Markers 424439 and 424446 are approximately 20 kb apart. In the Col-0 genomic sequence, there are nine predicted genes in this region (Fig. 7C), but none are annotated as enzymes of the proposed Wheeler-Smirnoff Pathway for vitamin C biosynthesis in plants (Wheeler et al., 1998). We designed primer pairs to amplify overlapping segments of DNA spanning the 20-kb region from the *vtc2-2* mutant. Sequencing of these fragments and comparison with the wild-type Col-0 sequence identified a mis-sense change in the putative gene F10M23.190 (GI:7452423; Fig. 7D), resulting in a Gly to Asp change in the predicted exon 5 (new GenBank ID AF508793). This gene was also sequenced from the three other *vtc2* mutants. A mis-sense mutation was identified in *vtc2-3* (Fig. 7D), resulting in a Ser to Phe change in the predicted exon 6. Both *vtc2-1* and *vtc2-4* had the same mutation, which changed the 3′ splice site of the predicted intron 5 from AG to AA (Fig. 7D). These last two mutations are almost certainly independently generated, because one was isolated in wild-type Col-0 and a the other was from a strain of Col-0 carrying a *PAT1-GUS* transgene (Rose and Last, 1997). Together, these four mutations show that putative gene F10 M23.190 is *VTC2.* As additional confirmation, all four mutant alleles of *VTC2* were complemented using genomic clones of F10M23.190 isolated from Col-0 by PCR (I. Levin and S. Norris, unpublished data).

The F10M23.190 gene (*VTC2*) was previously annotated as an undefined protein (GI:7452423; Mayer, 1999). The most similar proteins in the GenBank database are as follows: Arabidopsis protein MCO15.7, *Caenorhabitis elegans* protein C10F3.4, and fruitfly (*Drosophila melanogaster*) protein CG3552, none of which have a demonstrated function. Thus, although we have a phenotype associated with mutations in *VTC2*, the regulatory or biosynthetic pathways leading to the reduced vitamin C levels in these mutants remain to be discovered.

DISCUSSION

We have outlined a map-based cloning strategy, which leads to the identification of an Arabidopsis mutation in a straightforward manner in approximately 1 year. Our timeline assumes that it is possible to determine the phenotype of F_2 plants as they are growing. If the phenotype of interest is measured on seeds (i.e. F_3 seeds from F_2 plants), then the mapping time will be increased by 3 months. The strategy that we propose is designed to minimize the number of plants that have to be subjected to phenotypic analysis. In most cases, DNA based markers can be determined faster and more accurately than individual plant phenotypes. Obviously, if phenotyping is easier than genotyping, this procedure can be changed by identifying a large number of homozygous mutant, or wild type in the case of dominant mutations, plants and genotyping these alone.

Modifications of the process that we have outlined can speed up the mapping timeline. In many cases, as the mapping region is narrowed down, candidate genes become obvious, and it is possible to shift to sequencing at any stage during the process (Fig. 2). For rare examples of very reliable phenotypes, it may not be necessary to grow an F_3 generation for progeny testing, thus, shortening the timeline by 2 months. It is also possible to grow a single large F_2 population, rather than two sequentially grown populations (first-pass mapping and fine-scale mapping). However, this may result in wasted effort because some mutations are recalcitrant to genetic mapping. Situations that can make a given mutation difficult or impossible to map include: QTL variation for the trait of interest in the Col-0/L*er* F_2 population, phenotypes caused by multiple mutations, sensitivity of the phenotype to environmental variation in the greenhouse or growth chamber, and non-nuclear mutations.

The mapping timeline that we have outlined depends on the ability to rapidly genotype large numbers of plants. It may be difficult to maintain this timeline by using gel-based methods for SNP and InDel detection. High-throughput SNP detection methods are available, but they involve a high initial equipment cost that could make them prohibitive to set up and use in an individual laboratory. One solution to this problem may be for universities or academic departments to set up genotyping centers, similar to those that currently exist for DNA sequencing. Similar to a DNA sequencing center, a genotyping center could serve a large number of researchers working in all areas of molecular genetics.

The current rate-limiting step for map-based cloning in Arabidopsis is the number of F_2 plants that must be analyzed to find recombinants in a sufficiently small interval of DNA. There are no known methods for increasing meiotic recombination frequency in Arabidopsis (or any other plant). However, both ecotype-specific variation (Barth et al.,

2001) and mutations that decrease meiotic recombination frequency (Masson and Paszkowski, 1997; Grelon et al., 2001) have been reported. It is plausible that it will be possible to selectively alter meiotic recombination frequency at some point in the not too distant future by crossing QTLs from other ecotypes into standard laboratory strains, by overexpressing proteins necessary for elevated meiotic recombination, or perhaps by physical or chemical treatments that increase the recombination rate.

Sequencing of the Arabidopsis genome, the availability of the Cereon Arabidopsis Polymorphism Collection, and advances in the methods used for DNA marker detection have made map-based cloning of mutations in Arabidopsis a routine process. Mutation mapping will play a central role in the process of assigning a function to the thousands of plant genes that currently are known only as predicted open reading frames. Given the advantages of map-based cloning that we have outlined in the introduction, this is a viable approach for gene discovery that can be used in any laboratory.

Received January 31, 2002; accepted February 5, 2002.

LITERATURE CITED

Abajian C (1994) Sputnik program, http://rast.abajian.com/sputnik/

Ahmadian A, Gharizadeh B, Gustafsson AC, Sterky F, Nyren P, Uhlen M, Lundeberg J (2000) Single-nucleotide polymorphism analysis by pyrosequencing. Anal Biochem **280:** 103–110

Alderborn A, Kristofferson A, Hammerling U (2000) Determination of single-nucleotide polymorphisms by real-time pyrophosphate DNA sequencing. Genome Res **10:** 1249–1258

Alonso-Blanco C, Koornneef M (2000) Naturally occurring variation in Arabidopsis: an underexploited resource for plant genetics. Trends Plant Sci **5:** 22–29

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol **215:** 403–410

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796–815

Arondel V, Lemieux B, Hwang I, Gibson S, Goodman HM, Somerville CR (1992) Map-based cloning of a gene controlling omega-3 fatty acid desaturation in Arabidopsis. Science **258:** 1353–1355

Barth S, Melchinger AE, Devezi-Savula B, Lubberstedt T (2001) Influence of genetic background and heterozygosity on meiotic recombination in *Arabidopsis thaliana*. Genome **44:** 971–978

Bell CJ, Ecker JR (1994) Assignment of 30 microsatellite loci to the linkage map of Arabidopsis. Genomics **19:** 137–144

Cha RS, Zarbl H, Keohavong P, Thilly WG (1992) Mismatch amplification mutation assay (MAMA): application to the c-H-ras gene. PCR Methods Appl **2:** 14–20

Chen X, Livak KJ, Kwok PY (1998) A homogeneous, ligase-mediated DNA diagnostic test. Genome Res **8:** 549–556

Cho RJ, Mindrinos M, Richards DR, Sapolsky RJ, Anderson M, Drenkard E, Dewdney J, Reuber TL, Stammers M, Federspiel N et al. (1999) Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. Nat Genet **23:** 203–207

Conklin PL, Norris SR, Wheeler GL, Williams EH, Smirnoff N, Last RL (1999) Genetic evidence for the role of GDP-mannose in plant ascorbic acid (vitamin C) biosynthesis. Proc Natl Acad Sci USA **96:** 4198–4203

Conklin PL, Saracco SA, Norris SR, Last RL (2000) Identification of ascorbic acid-deficient *Arabidopsis thaliana* mutants. Genetics **154:** 847–856

Conklin PL, Williams E, Last RL (1996) Environmental stress sensitivity of an ascorbic acid-deficient Arabidopsis mutant. Proc Natl Acad Sci USA **93:** 9970–9974

Drenkard E, Richter BG, Rozen S, Stutius LM, Angell NA, Mindrinos M, Cho RJ, Oefner PJ, Davis RW, Ausubel FM (2000) A simple procedure for the analysis of single nucleotide polymorphisms facilitates map-based cloning in Arabidopsis. Plant Physiol **124:** 1483–1492

Feldman KA, Malmberg RL, Dean C (1994) Mutagenesis in Arabidopsis. *In* E Meyerowitz, C Somerville, eds, Arabidopsis. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp 137–182

Giraudat J, Hauge BM, Valon C, Smalle J, Parcy F, Goodman HM (1992) Isolation of the Arabidopsis *ABI3* gene by positional cloning. Plant Cell **4:** 1251–1261

Green P (1996) Phrap program, http://bozeman.mbt.washington.edu/phrap.docs/phrap.html

Grelon M, Vezon D, Gendrot G, Pelletier G (2001) *AtSPO11-1* is necessary for efficient meiotic recombination in plants. EMBO J **20:** 589–600

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315–324

Heremans B, Jacobs M (1997) A mutant of *Arabidopsis thaliana* (L.) Heynh. with modified control of aspartate kinase by threonine. Biochem Genet **35:** 139–153

Jansen RC (1993) Interval mapping of multiple quantitative trait loci. Genetics **135:** 205–211

Kearsey MJ, Farquhar AG (1998) QTL analysis in plants; where are we now? Heredity **80:** 137–142

Konieczny A, Ausubel F (1993) A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers. Plant J **4:** 403–410

Kreps JA, Ponappa T, Dong W, Town CD (1996) Molecular basis of alpha-methyltryptophan resistance in *amt-1*, a mutant of *Arabidopsis thaliana* with altered tryptophan metabolism. Plant Physiol **110:** 1159–1165

Kwok S, Kellogg DE, McKinney N, Spasic D, Goda L, Levenson C, Sninsky JJ (1990) Effects of primer-template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies. Nucleic Acids Res **18:** 999–1005

Leung J, Bouvier-Durand M, Morris PC, Guerrier D, Chefdor F, Giraudat J (1994) Arabidopsis ABA response gene *ABI1*: features of a calcium-modulated protein phosphatase. Science **264:** 1448–1452

Li J, Last RL (1996) The *Arabidopsis thaliana trp5* mutant has a feedback-resistant anthranilate synthase and elevated soluble tryptophan. Plant Physiol **110:** 51–59

Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ (1999) High density synthetic oligonucleotide arrays. Nat Genet Suppl **1:** 20–24

Lister C, Dean C (1993) Recombinant inbred line for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. Plant J **4:** 745–750

Livak KJ (1999) Allelic discrimination using fluorogenic probes and the 5′ nuclease assay. Genet Anal **14:** 143–149

Lukowitz W, Gillmor CS, Scheible WR (2000) Positional cloning in Arabidopsis: why it feels good to have a genome initiative working for you. Plant Physiol **123:** 795–805

Lukowitz W, Nickle TC, Meinke DW, Last RL, Conklin PL, Somerville CR (2001) Arabidopsis *cyt1* mutants aredeficient in a mannose-1-phosphate guanylyltransferase and point to a requirement of N-linked glycosylation for cellulose biosynthesis. Proc Natl Acad Sci USA **98:** 2262–2267

Masson JE, Paszkowski J (1997) *Arabidopsis thaliana* mutants altered in homologous recombination. Proc Natl Acad Sci USA **94:** 11731–11735

Mayer K, Schuller C, Wambutt R, Murphy G, Volckaert G, Pohl T, Dusterhoft A, Stiekema W, Entian KD, Terryn N et al. (1999) Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. Nature **402:** 769–777

McCallum CM, Comai L, Greene EA, Henikoff S (2000) Targeted screening for induced mutations. Nat Biotechnol **18:** 455–457

Meyer K, Leube MP, Grill E (1994) A protein phosphatase 2C involved in ABA signal transduction in *Arabidopsis thaliana*. Science **264:** 1452–1455

Michaels SD, Amasino RM (1998) A robust method for detecting single-nucleotide changes as polymorphic markers by PCR. Plant J **14:** 381–385

Michelmore RW, Paran I, Kesseli RV (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci USA **88:** 9828–9832

Mindrinos M, Katagiri F, Yu GL, Ausubel FM (1994) The *A. thaliana* disease resistance gene *RPS2* encodes a protein containing a nucleotide-binding site and leucine-rich repeats. Cell **78:** 1089–1099

Neff MM, Neff JD, Chory J, Pepper AE (1998) dCAPS, a simple technique for the genetic analysis of single nucleotide polymorphisms: experimental applications in *Arabidopsis thaliana* genetics. Plant J **14:** 387–392

Peters JL, Constandt H, Neyt P, Cnops G, Zethof J, Zabeau M, Gerats T (2001) A physical amplified fragment-length polymorphism map of Arabidopsis. Plant Physiol **127:** 1579–1589

Ponce MR, Robles P, Micol JL (1999) High-throughput genetic mapping in *Arabidopsis thaliana*. Mol Gen Genet **261:** 408–415

Redei GP (1992) A heuristic glance at the past of Arabidopsis genetics. *In* C Koncz, N Chua, J Schell, eds, Methods in Arabidopsis Research. World Scientific, Singapore, pp 1–15

Rose AB, Last RL (1997) Introns act post-transcriptionally to increase expression of the *Arabidopsis thaliana* tryptophan pathway gene *PAT1*. Plant J **11:** 455–464

Sharbel TF, Haubold B, Mitchell-Olds T (2000) Genetic isolation by distance in *Arabidopsis thaliana*: biogeography and postglacial colonization of Europe. Mol Ecol **19:** 2109–2118

Spiegelman JI, Mindrinos MN, Oefner PJ (2000) High-accuracy DNA sequence variation screening by DHPLC. Biotechniques **29:** 1084–1092

Sussman MR, Amasino RM, Young JC, Krysan P, Austin-Phillips S (2000) The Arabidopsis knockout facility at the University of Wisconsin-Madison. Plant Physiol **124:** 1465–1467

Tanksley SD, Ganal MW, Martin GB (1995) Chromosome landing: a paradigm for map-based cloning in plants with large genomes. Trends Genet **11:** 63–68

Ugozzoli L, Wallace RB (1991) Allele-specific polymerase chain reaction. Methods Enzymol **2:** 42–48

Wada Y, Yamamoto M (1997) Detection of single-nucleotide mutations including substitutions and deletions by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Rapid Commun Mass Spectrom **11:** 1657–1660

Wheeler GL, Jones MA, Smirnoff N (1998) The biosynthetic pathway of vitamin C in higher plants. Nature **393:** 365–369

Yano M (2001) Genetic and molecular dissection of naturally occurring variation. Curr Opin Plant Biol **4:** 130–135

Zeng Z-B (1994) Precision mapping of quantitative trait loci. Genetics **136:** 1457–1468