A collection of INDEL markers for map-based cloning in seven Arabidopsis accessions

Daniel Ioan Pacurar1,2,*, Monica Lacramioara Pacurar1,2, Nathaniel Street†, John Desmond Bussell1,4, Tiberia Ioana Pop†, Laurent Gutierrez1,‡ and Catherine Bellini1,3,†

1 Umeå Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, SE-90183 Umeå, Sweden
2 University of Agricultural Sciences and Veterinary Medicine, 400372 Cluj Napoca, Romania
3 Institut Jean-Pierre Bourgin, UMR1318 INRA-AgroParisTech, F-78026 Versailles Cedex, France
4 Australian Research Council Centre of Excellence in Plant Energy Biology, University of Western Australia, Crawley, WA 6009, Australia

† Present address: Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, SE-90187 Umeå, Sweden.
‡ Present address: Centre de Ressources Régionales en Biologie Moléculaire, Université de Picardie Jules Verne, F-80039 Amiens, France.
* To whom correspondence should be addressed. E-mail: Daniel.Pacurar@plantphys.umu.se

Received 19 October 2011; Accepted 24 November 2011

Abstract

The availability of a comprehensive set of resources including an entire annotated reference genome, sequenced alternative accessions, and a multitude of marker systems makes Arabidopsis thaliana an ideal platform for genetic mapping. PCR markers based on INsertions/DELetions (INDELs) are currently the most frequently used polymorphisms. For the most commonly used mapping combination, Columbia × Landsberg erecta (Col-0 × Ler-0), the Cereon polymorphism database is a valuable resource for the generation of polymorphic markers. However, because the number of markers available in public databases for accessions other than Col-0 and Ler-0 is extremely low, mapping using other accessions is far from straightforward. This issue arose while cloning mutations in the Wassilewskija (Ws-4) background. In this work, approaches are described for marker generation in Ws-4 x Col-0. Complementary strategies were employed to generate 229 INDEL markers. Firstly, existing Col-0/Ler-0 Cereon predicted polymorphisms were mined for transferability to Ws-4. Secondly, Ws-0 ecotype Illumina sequence data were analyzed to identify INDELs that could be used for the development of PCR-based markers for Col-0 and Ws-4. Finally, shotgun sequencing allowed the identification of INDELs directly between Col-0 and Ws-4. The polymorphism of the 229 markers was assessed in seven widely used Arabidopsis accessions, and PCR markers that allow a clear distinction between the diverged Ws-0 and Ws-4 accessions are detailed. The utility of the markers was demonstrated by mapping more than 35 mutations in a Col-0 × Ws-4 combination, an example of which is presented here. The potential contribution of next generation sequencing technologies to more traditional map-based cloning is discussed.

Introduction

The function of a gene can be addressed via two strategies, forward and reverse genetics (Alonso and Ecker, 2006; Alonso-Blanco et al., 2009). Although positional cloning is a widely used forward genetics approach to isolate genes in different organisms (Choi et al., 2008), its utility can only be fully exploited in model systems, such as Arabidopsis thaliana. The principle behind positional cloning is to systematically narrow down the genetic interval containing a causal mutation by sequentially excluding all the other regions in the genome (Lukowitz et al., 2000). This can be achieved by the use of available and/or newly generated genetic markers that are polymorphic between the
accessions used for generating the mapping population(s). Different map-based cloning strategies have been described (reviewed in Lukowitz et al., 2000; Jander et al., 2002; Peters et al., 2003), and all rely on the availability of a highly dense genetic marker collection to provide adequate mapping resolution. This is a major limiting factor to the rate of mapping progress. Balancing the available marker systems can compensate for the lack of the preferred marker type (reviewed in Peters et al., 2003). In the last decade, DNA-based marker systems such as restriction fragment length polymorphism (RFLP) have progressively been replaced by PCR-based markers such as random amplified polymorphic DNA (RAPD), simple sequence repeat (SSR), and amplified fragment length polymorphisms (AFLP) (reviewed in Peters et al., 2003) and recently there have been several proposals for the use of next-generation sequencing (NGS) to exploit SNPs for mapping (Lister et al., 2009; Schneeberger and Weigel, 2011). Indeed, mass sequencing of new Arabidopsis accessions by the 1001 Genomes Project (http://www.1001genomes.org/accessions.html) has dramatically expanded the possibilities for sequence comparisons for mapping.

In Arabidopsis, Insertion/DELetions (INDELS) and single nucleotide polymorphisms (SNPs) have become the most commonly used markers because they are easy to use, PCR based, co-dominant (fully informative) and relatively abundant. Importantly, these markers are also readily accessible; either as designed and tested PCR markers deposited at The Arabidopsis Information Resource (TAIR; http://www.arabidopsis.org/) or as an indexed list of polymorphisms in direct sequence comparisons (Cereon collection, also available at TAIR). By systematically exploiting the available predicted polymorphic sequences in the Cereon collection, Hou et al. (2010) generated a maker database as an alternative to TAIR that can be used for mapping in a Col-0×Ler-0 combination. Although Columbia-0 and Landsberg erecta are the most commonly used accessions for genetic studies, there are often compelling reasons to isolate new mutants in other ecotypes. Firstly, screens for suppressor mutations rely on the re-mutagenesis of existing mutants that may be in backgrounds other than Col-0 or Ler-0. Secondly, diverse accessions are increasingly being used to unravel complex biological mechanisms by exploiting natural genetic variation (reviewed in Alonso-Blanco et al., 2009). Mapping traits in these accessions is clearly hampered by the fact that most of the documented polymorphism in the public databases is between Col-0 and Ler-0. Although these documented polymorphisms can serve as a starting point for mapping in other segregating combinations, only approximately 50% of the Col-0/Ler-0 polymorphisms can be used for other pair combinations (Peters et al., 2001). Thus, additional new markers need to be identified for each particular new combination.

In an attempt to map over 35 Arabidopsis mutants generated in the Ws-4 background, it was soon realized that publicly deposited markers polymorphic between Ws and Col-0 were far too few. This deficiency was addressed by identifying new polymorphisms as follows. Firstly, it was tested if the available Col-0/Ler-0 polymorphic INDELS from Cereon were conserved between Col-0 and Ws-4. Secondly, and in lieu of a Ws-4 sequence, advantage was taken of the available Ws-0 sequence (Gan et al., 2011; http://www.1001genomes.org/), and three different computational methods were used to identify nearly 13 500 INDELS between Col-0 and Ws-0. A selection of these was tested by PCR to generate new markers for Col-0/Ws-0 and their transferability to Ws-4 was assessed. Finally, shotgun sequencing was used for direct comparison of Ws-4 and Col-0 in selected regions. In addition, all 229 markers were tested for polymorphism amongst seven Arabidopsis accessions including the classical Col-0/Ler-0 combination. Thus, polymorphisms have been verified among seven widely used Arabidopsis accessions, increasing the number of markers available in TAIR for any given pair of accession by a minimum of 60% (Col-0/Ler-0) to more that 630% (No-0/C24), and so providing an invaluable tool for mapping mutations amongst these accessions. Moreover, the existing database has been updated with new accessions, first by including Shahdara (Sha) in the list, and second by differentiating two of the Wassilewskija accessions (Ws-0 and Ws-4).

Materials and methods

Plant material

Seven commonly used Arabidopsis thaliana accessions: Columbia (Col-0, N1092); Landsberg erecta (Ler-0, NW20); Wassilewskija (Ws-0, N1602 and Ws-4, N5390); C24 (N906); Nossen (No-0, CS8521); and Shahdara (Sha, N929) were included in this study. The mutant designated as 420 was previously identified in a screen for suppressors of the sur2-1 mutation (DI Pácurar et al. unpublished data). To map the suppressor mutation (Ws-4 background), phenotyped mutant seedlings making fewer adventitious roots than sur2-1 were identified in a F2 population obtained by crossing the mutant with atr4-1, an allele of the sur2 mutant in a Col-0 background (Smolen and Bender, 2002). Using standard protocols, genomic DNA was extracted from entire mutant seedlings grown in vitro as previously described by Sorin et al. (2005), and from the different Arabidopsis accessions, and used as template for mapping and testing the newly developed markers, respectively.

Identification and validation of the polymorphic INDELS

The INDEL markers described in this study were identified/generated from three different sources. First, the Monsanto Arabidopsis Polymorphism and Ler Sequence Collections were used to identify polymorphisms. Using the described Col-0/Ler-0 polymorphism, INDELS of at least 5 bp in length in the regions of interest were identified and they were subsequently verified by amplifying the region spanning the INDEL in all the accessions included in the study. For visualization of polymorphic INDELS as short as 5 bp in length, the optimal size of the fragment spanning the INDEL was determined to be approximately 10 times the size of the respective INDEL. Using the Primer3 software (http://frodo.wi.mit.edu/primer3/), the primers were designed accordingly to match the characteristics of each INDEL.

Second, access to paired-end sequence data for the Ws-0 accession was kindly provided from the 1001 Genomes project (Gan et al., 2011; http://www.1001genomes.org/). The data consisted of 36 bp paired-end reads, with an insert size of 380 bp, generated using the Illumina GAII platform. There were a total of 121 million reads and
4.4 Gbp sequence, which is ~36-fold coverage. Insert size was estimated by mapping the reads to the reference *Arabidopsis thaliana* genome (TAIR9) using the Burrows–Wheeler Aligner (bwa; Li and Durbin, 2009) allowing two mismatches and two gaps. As some sequence files had quality values on different scales, all values were rescaled to Phred33 using BioPerl (Stajich et al., 2002). Three different approaches that all utilize paired-end mapping information to identify potential INDELS were then used. For the SHORE pipeline (Ossowski et al., 2008) analysis, the genome mapper read alignment software was used and two mismatches and two INDELS (gaps) were allowed within read alignments. Additional steps were performed as detailed in the SHORE documentation. The ‘shore structure’ function was used to identify large INDEL events. Read alignments and insert distribution estimates generated using bwa were used as input for BreakDancerMax (Chen et al., 2009) and Pindel (Ye et al., 2009) analysis. Output from both software tools was post-filtered to only consider insertions or deletions of >15 bp.

Finally, blasting sequenced fragments of the Ws-4 genome against the Col-0 reference sequence generated a small set of markers. If an INDEL of at least 5 bp in size was identified, the polymorphism was subsequently verified in all the accessions included in the study.

**Nomenclature**

In order to facilitate the association of a marker with its location on the reference *Arabidopsis* genome, the markers were named in the format UPSC-N-XXXXX, where UPSC stands for Umeå Plant Science Centre, N for the chromosome number, and the X represents each marker’s physical position on the reference genome, in kb.

**PCR amplification and gel electrophoresis**

Template DNA from the seven analysed accessions was amplified, on a BioRad S1000™ Thermal Cycler, with the primers designed for each INDEL, using standard PCR conditions: 5 min at 95 °C, followed by 40 cycles of 20 s at 95 °C, 20 s at 55–60 °C, and 20 s at 72 °C, with a final extension of 5 min at 72 °C. The PCR products were subsequently separated in 4% agarose gels, or 2% agarose gels for INDELS bigger than 100 bp.

**Results and discussion**

**Alternative resources and strategies to generate new INDEL markers**

Since the success of a map-based cloning project depends on the availability of a high marker density between the ecotypes used to generate the mapping population(s), the ability to detect new polymorphic markers in the region of interest is critical. Moreover, this detection should be accurate and at an appropriate cost and throughput (Jander et al., 2002). Several high-throughput strategies have been developed to detect polymorphism (reviewed in Jander et al., 2002). However, the majority of them detect only SNPs. Detecting INDELS is a more challenging task and requires substantial bioinformatics analysis. In this study, three approaches to this problem were taken that are described in detail below. Firstly, a selection of the predicted Cereon INDELs (http://www.arabidopsis.org/browse/Cereon/index.jsp; Jander et al., 2002) were tested for transferability to detect Col-0 versus Ws-4 markers. Secondly, deep sequencing of the Ws-0 accession was utilized for computational prediction of INDELS between Col-0 and Ws-0. Finally, shotgun sequencing was used in selected regions that required markers and for which none were readily found with the other two methods. In total, these methods have yielded 229 new confirmed markers that are variously polymorphic amongst seven commonly used *Arabidopsis* accessions (Col-0, Ler-0, Ws-0, Ws-4, C24, No-0, and Sha; see Tables 1, 2; Fig. 1).

The Cereon collection as a classical resource for identification of INDEL markers

In the first approach, which yielded about two-thirds of our markers (Table 1), predicted Col-0/Ler-0 polymorphisms were taken from the Monsanto Arabidopsis Polymorphism and Ler Sequence Collections (http://www.arabidopsis.org/browse/Cereon/index.jsp; Jander et al., 2002). INDELs that matched our selection criteria (see the Materials and methods) were amplified by flanking primers, and polymorphism was assessed in the extended accession set. All of the 163 predicted Cereon INDELs that were tested were confirmed to be polymorphic between Col-0 and Ler-0. By comparison, confirmation of a maximum 90% of the tested predicted single nucleotide polymorphisms (SNPs) has been reported (Rounsley, 2003). However, only 111 (68%) of the tested INDELS were polymorphic between Col-0 and Ws-4 (Table 1), making this approach somewhat inefficient for generating polymorphic INDEL markers between combinations other than Col-0/Ler-0 (Table 1).

**Table 1. Summary of UPSC marker sources**

| Source                  | Number of predicted markers from each source | The accuracy of prediction<sup>a</sup> | % of markers polymorphic between Col-0 and Ws-4<sup>b</sup> |
|-------------------------|---------------------------------------------|--------------------------------------|------------------------------------------------------------|
| 1                       | Cereon                                      | 163                                  | 68.1                                                       |
| 2                       | Method 1 (bwa+Pindel)                       | 27                                   | 66.7                                                       |
|                         | Method 2 (SHORE)                           | 11                                   | 45.4                                                       |
|                         | Method 3 (Breakdancer)                     | 8                                    | 37.5                                                       |
|                         | Method 2+3                                 | 2                                    | 50.0                                                       |
| 3                       | In-house sequencing                        | 18                                   | 94.4                                                       |
| Total                   |                                             | 229                                  | 67.7                                                       |

<sup>a</sup> Markers that were polymorphic between Col-0 and the accession used for prediction (Ler, for source 1; Ws-0, for source 2; Ws-4, for source 3).

<sup>b</sup> The % relates to the number of markers generated from respective source/ method.
Markers identified using next-generation sequencing data

In the second approach, pre-releaseWs-0, CS6891 accession, sequence from 1001 Genomes project (Gan et al., 2011; http://www.1001genomes.org/) was used to computationally predict INDELs between Col-0 and Ws-0. In order to maximize the number of predicted INDEL markers, three different Structural Variation (SV) software methods (Pindel, SHORE, and BreakDancer) were used. These methods respectively identified 932, 711, and 40 insertions and 188, 9488, and 2068 deletions between Col-0 and Ws-0. To assess the accuracy of these prediction methods, a set of 46 non-overlapping predicted INDELs and an additional two INDELs that were predicted independently by two methods (Table 1) were selected for confirmation. Primers were designed to flank each predicted INDEL and PCR products from the seven ecotypes were visualized on agarose gels. All but two of the 48 predicted INDELs were confirmed to be polymorphic between Col-0 and Ws-0. The two predictions, although monomorphic between Col-0 and Ws-0 (probably due to additional insertion/deletion that complemented the size of the targeted one), were polymorphic between other ecotype combinations. Deep-sequencing alignment yielded 48/229 (21%) of the markers in our set. Although, in the current work, only a small number of predicted INDELs have been tested, the fact that all of the tested INDEL events yielded viable mapping markers highlights the potential of using paired-end next generation sequence data to develop high-density maps of a desired marker type and accession.

Markers derived from direct sequencing of Ws-4 and Col-0

Finally, primers were designed based on Col-0 sequence in regions where insufficient Col-0/Ws-4 polymorphic markers had been identified using the other methods. These primers were designed to amplify approximately 1.6 kb of (usually) non-coding genomic DNA which was sequenced directly in both Col-0 and Ws-4. In addition, in the process of map-based-cloning of mutations in the Ws-4 background, a candidate gene approach was taken and Ws-4 sequence was obtained by sequencing the candidate genes in the corresponding suppressor mutants. Aligning shotgun or targeted sequenced fragments with reference to Col-0 sequence generated the remaining 18 (8%) of the markers. The identified INDELS were subsequently tested in all seven accessions included in the study.

Map position of UPSC markers

The relative chromosomal position of the 229 newly generated UPSC (UPSC stands for Umea Plant Science Centre) markers is shown in Fig. 2. The marker distribution over the five chromosomes shows some regions with high clustering and other regions with less coverage. This situation does not reflect relative degrees of polymorphism but rather that our mutations of interest were located in the densely covered regions (DI Păcurar et al., unpublished data). The number of markers generated in this study that were polymorphic in pairwise comparisons of the seven Arabidopsis accessions is shown in Table 2. Although the Col-0/Ler-0 combination is relatively well represented at TAIR, a very limited number of markers are available there for the additional accessions included in the current study (Table 3). An overview of the polymorphisms between the pairs of Arabidopsis accessions revealed by our marker collection is given in Fig. 1. Some loci were able to distinguish all or most of the seven accessions, but many of them (67%) yielded only two allele sizes distributed among the ecotypes. Despite this, a high degree of definition between the reference genome (Col-0) and the others was possible (Table 2). Some markers could not be amplified in an ecotype-specific manner, most likely due to polymorphisms in (or deletions of) primer binding sites compared with the reference sequence used for primer design. Alternatively, insertions may have been large enough to preclude amplification. The complete resource information, including primer sequences, polymorphism size, and PCR conditions is detailed in Supplementary Table 1, and has also been deposited at TAIR.

Polymorphism between Wassilewskija accessions Ws-0 and Ws-4

Arabidopsis lines originating from the same ecotype are often used and circulated between laboratories and research groups without a clear specification of their exact origin or accession number. In a recent extensive study, Anastasio et al. (2011) uncovered the existence of many misidentified Arabidopsis accessions in stock centres and recommended caution when using particular accessions. Of five Wassilewskija accessions available in stock centres, two (Ws-2 and Ws-4) have been used as parental lines in individual tagging projects, one (Ws-1) as background for recombinant inbred (RI) lines, and two (Ws-0 and Ws-3) are available as donations. A high degree of polymorphism is evident between Ws-0 and Ws-4 (Fig. 1). This finding, also reported by Aukerman et al. (1997) and recently by Anastasio et al. (2011), is significant for Arabidopsis geneticists because these two accessions have been used in major projects: Ws-4 was used as background for the FLAG lines generated at INRA Versailles; (Samson et al., 2002) and Ws-0 has been sequenced as part of the 1001

Table 2. Number of INDEL markers from a total of 229 generated in this study that were polymorphic in pairwise comparisons of seven Arabidopsis accessions

| Accession | Col-0 | Ler-0 | Ws-4 | Ws-0 | C24 | No-0 | Sha |
|-----------|-------|-------|------|------|-----|------|-----|
| Col-0     | •     | 209   | 155  | 165  | 151 | 151  | 154 |
| Ler-0     | •     | 106   | 104  | 107  | 104 | 97   |     |
| Ws-4      | •     | 83    | 98   | 88   | 93  |      |     |
| Ws-0      | •     | 93    | 82   | 91   |     |      |     |
| C24       | •     | 88    | 96   |      |     |      |     |
| No-0      | •     | 79    |      |      |     |      |     |
| Sha       | •     |       |      |      |     |      |     |

| Aridopsis Accession | Number of Genomically Polymorphic Loci | Number of Genomically Polymorphic and Phylogenetically Distinct Loci |
|---------------------|---------------------------------------|---------------------------------------------------------------|
| Col-0               | 209                                   | 151                                                           |
| Ler-0               | 106                                   | 104                                                           |
| Ws-4                | 83                                    | 98                                                            |
| Ws-0                | 93                                    | 82                                                            |
| C24                 | 88                                    | 96                                                            |
| No-0                | 79                                    |                                                               |
| Sha                 |                                       |                                                               |
Genomes project; (Gan et al., 2011; http://1001genomes.org/accessions.html). Documented PCR-based markers are provided here that can be used to distinguish the two accessions. The percentage of Col-0/Ws-4 polymorphic markers generated by using the Col-0/Ws-0 predicted INDELs was lower than the percentage of Col-0/Ws-4 polymorphic markers generated by using the Cereon Col-0/Ler-0 predictions (Table 1), suggesting that the two Wassilewskija accessions are more similar than the Col-0/Ws-4 accessions. Fig. 1. Matrix representation of the polymorphism revealed by the UPSC markers amongst seven Arabidopsis accessions. Each UPSC marker’s position on the five Arabidopsis chromosomes is shown in kilobase pairs (kb). Each different allele size is represented by a different colour. Green, Col-0 allele; blue, Ler-0 allele; yellow, light and dark orange represent new alleles amplified with the UPSC markers. Markers that failed to amplify for particular accessions are represented in grey.
divergent than expected. As shown in Table 2, a high degree of polymorphism was observed, with 83 markers being polymorphic between the two lines. The question of Wassilewskija ecotype definition was explored further by testing a selection of classical SSLP markers indexed at TAIR. For these markers, the originating Wassilewskija accession is not specified (the ecotype is abbreviated on TAIR only as ‘Ws’) and it was possible to show, based on the size of the amplified fragments, that different Wassilewskija accessions were used in defining the marker sizes for ‘Ws’ (Table 4).

Together, our results, and those of others (Torjek et al., 2003; Anastasio et al., 2011) accentuate the need for a careful evaluation of the genetic background prior to assuming that a line is in fact of the implied origin. Such
genotyping can be readily achieved by using accession-diagnostic PCR markers such as the INDELs reported here.

**High-resolution mapping of superroot2 suppressor mutants using the UPSC marker set**

A screen for suppressor mutations of the Arabidopsis *superroot2-1* (*sur2-1*) mutant, previously identified by our group (Delarue et al., 1998), was performed to isolate new mutants affected in adventitious root formation. The mutants were characterized and subsequently mapped using the UPSC marker collection described here. For mapping, the *sur2-1* suppressor mutants (Ws-4 background) were crossed with *atr4-1*, an allele of *sur2-1* in the Col-0 background (Smolen and Bender, 2002). By application of the UPSC markers and following the strategy described in Fig. 3D, it was possible successfully to fine map in parallel 37 mutations (DI P. Cuar et al., unpublished data).

The phenotype of one of the *sur2-1* suppressors, designated 420, is shown in Fig. 3A–C. Suppressor mutant seedlings, germinated in vitro and etiolated for 72 h, showed shorter hypocotyls and roots than *sur2-1* (Fig. 3A). In addition, all suppressor seedlings displayed a triple-response phenotype, indicative of ethylene overproduction. Seven days after transfer to light, mutant seedlings showed a strong suppression of the *sur2-1* phenotype and significantly fewer adventitious roots developed on the hypocotyl compared with *sur2-1* (Fig. 3B). Grown in soil, in short day conditions (8/16 h light/darkness) suppressor plants developed a compact rosette with wrinkled leaf blades (Fig. 3C). Segregation analysis of F$_2$ progeny from a *sur2-1*×420 cross showed a 3:1 ratio of superroot:suppressor phenotype, consistent with a single recessive mutation (not shown).

The map-based cloning of the *superroot2-1* suppressor mutant 420 is described here as an example of the application of the UPSC markers. Initially, a mapping population of approximately 100 phenotyped mutant plants was collected. DNA was extracted from 24 individuals and used in first-pass mapping. For practical reasons, the DNA from the 24 seedlings was not pooled because it would have made it impossible to trace incorrectly phenotyped seedlings or contaminants that occasionally occurred due to incomplete penetration of the *sur2-1* phenotype or as a result of growth conditions that influence the *sur2* phenotype (Delarue et al., 1998). Marker usage and mapping progress was continually updated in a Microsoft Excel template, as shown in Fig. 3D. For first-pass mapping, classical Col-0/Ws polymorphic markers from TAIR were used, and the marker NGA1139 was shown to be linked to the mutation. Subsequently, a three-point cross analysis identified NGA1107 as a flanking marker. For comparison, segregation analysis of two unlinked markers (CIW12, on Ch1 and NGA151, on Ch5) is shown. As shown in Fig. 3D, eight new internal markers were subsequently used to map the mutation. Using the UPSC marker resource, the mutation was mapped to the bottom of chromosome 4, between the markers UPSC_4-17326 and UPSC_4-17432 (i.e. a region of 106 kb). For two nested

| Table 3. Number of SSLP markers available on TAIR prior to our study that were polymorphic between pairs of *Arabidopsis* accessions included in this study. The specific Ws accession used to define these markers is not given on TAIR, and there are no markers indexed on TAIR for Sha ecotype. |
|---|---|---|---|---|---|
| Col | Ler | Ws | C24 | No |
| Col | ● | 338 | 99 | 44 | 36 |
| Ler | ● | 83 | 45 | 30 |  |
| Ws | ● | 32 | 33 |  |  |
| C24 | ● | 14 |  |  |  |
| No | ● |  |  |  |  |

| Table 4. Allele sizes of PCR products amplified from Col-0, Ws-4, and Ws-0 for 15 selected SSLP markers from TAIR. The accession of origin of the Ws marker (Ws-0, Ws-4, other) detailed on TAIR is inferred based on the size of amplified product compared to the allele size registered on TAIR. |
|---|---|---|---|---|
| Marker | Chromosome | Col-0 | Ws-4 | Ws-0 | Origin of Ws TAIR marker |
| NGA59 | 1 | 111 | 141 | 111 | 141, 83 | Ws-4, other |
| CW12 | 1 | 128 | 120 | 115 | 120, 115 | Ws-4, Ws-0 |
| NGA111 | 1 | 128 | 146 | 180 | 146 | Ws-4 |
| NGA280 | 1 | 105 | 85 | 85 | 85 | Ws-4/Ws-0 |
| NGA168 | 2 | 151 | 135 | 135 | 135, 130 | Ws-4/Ws-0, other |
| NGA6 | 3 | 143 | 147 | 154 | 131 | Other |
| NGA162 | 3 | 107 | 85 | 97 | 85 | Ws-4 |
| NGA172 | 3 | 162 | 138 | 180 | 138 | Ws-4 |
| NGA8 | 4 | 154 | 166 | 188 | 166 | Ws-4 |
| NGA1107 | 4 | 150 | 140 | ~145 | 140 | Ws-4 |
| NGA1139 | 4 | 114 | ~145 | ~100 | 118 | Other |
| NGA151 | 5 | 150 | 102 | 110 | 102 | Ws-4 |
| CA72$^a$ | 5 | ~225 | ~210 | ~205 | 110 | Other |
| NGA249 | 5 | 125 | 115 | 115 | 115 | Ws-4/Ws-0 |
| CW9 | 5 | 165 | 145 | 140 | 140 | Ws-0 |

$^a$ In our hands the marker CA72 gives these allele sizes. By comparison the sizes registered on TAIR for Col and Ws are 124 bp and 110 bp, respectively.
Fig. 3. Phenotype of a superroot2 suppressor mutant and mapping using the UPSC marker set. (A,B,C) Phenotype of the suppressor 420, compared to the sur2-1 mutant: 3-d-old etiolated seedlings (A), adventitious roots on etiolated hypocotyls 8 d after transfer to the light (B), and 40-d-old sur2-1 and 420 suppressor seedlings grown in soil, in short day conditions (B/16 h light/dark) (C). Arrows indicate the hypocotyl-root junction. Bar, 1 cm. (D) Mapping strategy for the suppressor 420. An F$_2$ mapping population was generated by

![Fig. 3](image_url)

### Table 1: Phenotype of the suppressor 420

| Chr Marker | Position (bp) | Phenotyped (F$_2$) plants of the suppressor 420 | RF(%) |
|------------|---------------|-----------------------------------------------|-------|
| I          | CIW12         | 9621357                                       | 43.75 |
| V          | NGAl51        | 4689920                                       | 50    |

(D)
markers (UPSC_4-17345 and UPSC_4-17363) no additional recombinants were found after increasing the mapping population to 450 individuals (900 chromosomes). Evidently these two markers were closely linked to the mutation (Fig. 3D, E). As our suppressor showed a very similar phenotype to the previously characterized mutant rce1-1 (Dharmasiri et al., 2003), the locus At4g36800, encoding the RUBI-conjugating enzyme1 (RCE1), is proposed as a potential suppressor gene. Sequencing of the candidate gene revealed a C-to-T substitution in the mutant 420 but not in sur2-1. The mutation, localized in exon 4, modified the Trp 121 to a premature STOP codon (Fig. 3E), potentially generating a truncated protein. RCE1 was confirmed as the suppressor gene by identifying a new mutation in a second allele (1375) isolated in our screen (Fig. 3E). The example provided above, together with the successful mapping of 36 other suppressor mutants (DI Piacuru et al., unpublished data), shows the potential of the UPSC marker resource for mapping.

Future prospects for map-based cloning

Despite the recent advances made in developing tools to facilitate map-based cloning, fine mapping per se still remains a research step many would prefer to avoid because it can be tedious work beset by complications. Primarily, high-resolution mapping relies on the availability of a high density of genetic markers (Lukowitz et al., 2000). A number of recent papers have proposed pipelines for next generation sequencing-based approaches to mutation mapping as a remedy for this. These approaches highlight the virtues of virtually limitless detection of SNPs for cost-effective increased mapping throughput and, consequently, the possibility to use new or non-reference accessions to generate the F2 mapping populations (Lister et al., 2009; Schneeberger et al., 2009; Laitinen et al., 2010; Austin et al., 2011; Schneeberger and Weigel, 2011; Uchida et al., 2011). Such mapping relies on computationally intensive assignment across the parental genomes of high-density SNP data (Deschamps and Campbell, 2010) and association of the SNPs of each accession with the phenotype. Linkage is deduced by the finding of a region where SNPs of the mutant accession are enriched. However, in the particular case of mutants generated by ethyl methane-sulphonate (EMS), direct sequencing of the mutant genome will not be sufficient to detect the mutation, unless two or more alleles are isolated from the screen (Schneeberger and Weigel, 2011; Uchida et al., 2011). As the likelihood of detecting only single alleles is higher (Pollock and Larkin, 2004), direct sequencing of the mutant will have to be supported by mapping (Schneeberger and Weigel, 2011). Moreover, although mapping by next generation sequencing may prove reliable in compatible genetic backgrounds and with clearly identifiable phenotypes, it is potentially sensitive in cases where these conditions are not met (Schneeberger and Weigel, 2011).

Another approach for using NGS data in mapping, and one that we are advocating here utilizes deep sequenced genomes to rapidly facilitate marker design for application in more traditional mapping methodologies (Lukowitz et al., 2000; Jander et al., 2002; Jander, 2006). Coarse mapping provides an approximate chromosomal location for the mutation and markers can be rapidly generated for fine mapping without the requirement for sequencing or high investment, low return prospecting for markers traditionally associated with map-based cloning. During fine mapping, crossing 420 (Ws-4 background) into atr4-1, an allele of sur2 in the Col-0 background. Twenty-four phenotyped mutant plants were used for first-pass mapping. DNA was extracted and stored on a 96-well plate for easy tracking of each individual. Using polymorphic markers from TAIR, the recombination frequency for the 24 individuals was assessed with the markers CIW12 and NGA151 located on chromosomes 1 and 5, respectively. As the calculated recombination frequencies (RF) were close to 50%, it was assumed that there was no linkage in between the two markers and the mutation. A third marker, NGA1139 (chromosome 4), appeared to be linked to the mutation: the RF dropped to 12.5%. Subsequently another chromosome 4 marker, NGA1107, was localized to the south of the mutation by observation of different individuals carrying recombination events. Subsequently, new UPSC markers confirmed the initial flanking markers, as the number of recombination events increased for the UPSC_4-13880, UPSC_4-11238, and UPSC_4-16516 markers, respectively. Once the initial interval was delimited, new internal markers were used to narrow down the genomic region containing the mutation. Thus, four markers to the north (UPSC_4-16853, UPSC_4-17110, UPSC_4-17251, UPSC_4-17326) and two to the south (UPSC_4-17544, UPSC_4-17432) were used to localize the mutation to a 106 kb genomic region containing 26 annotated loci. Only individuals carrying one (1, light grey) or two (2, dark grey) recombination events were kept for fine mapping, those homozygous for Ws-4 (0, lighter grey) being discarded, or excluded from further mapping. Two individuals, 9 and 19, were heterozygous for the last identified flanking markers, UPSC_4-17326 and UPSC_4-17432, respectively, but homozygous for UPSC_4-17345 and UPSC_4-17363. The region delimited by the two flanking markers (dotted line) was eventually confirmed by analysing a bigger mapping population. Note that individual 23 was heterozygous for all the markers in the mapped region and consequently was concluded to be a phenotypically mis-scored plant. (E) Positional cloning of the suppressor mutant 420 using UPSC markers. Recombination mapping localized the 420-suppressor allele to the bottom of chromosome 4 in between the INDEL markers UPSC_4-16853 and UPSC_4-17444, for which, respectively, 18 and 12 recombination events were identified. Additional internal UPSC markers used to score the recombinants located the suppressor mutation in between the markers UPSC_4-17326 (2 recombination events/900 tested chromosomes) and UPSC_4-17432 (6 recombination events/900 tested chromosomes). A candidate gene approach subsequently identified the locus At4g36800 (RCE1) as the suppressor gene. Sequence analysis of the candidate gene (exons are represented as black boxes, and introns as lines) using 420 mutant DNA revealed a C-to-T substitution in the 4th exon, converting a Trp to STOP. A second allele identified in our screen (1375) was shown to carry a C-to-T mutation at the border between the 3rd intron and the 4th exon, causing a splicing defect.
a candidate gene approach can be adopted to speed the process further. Given that about 50% of Arabidopsis genes have a documented function (Iida et al., 2011), and that systems studied with genetic screens are often a priori very well characterized, mutant genes can often be identified from a limited set of candidates without the need for generation of large fine-mapped pools. By way of example, it was possible to isolate sur2-1 suppressor 420 from 24 phenotyped F$_{25}$ by (i) conventional coarse mapping, followed by (ii) intensive marker design using existing INDEL databases and new INDELs identified from assembled Illumina sequence reads, and (iii) intelligent candidate gene selection informed by knowledge of the study system. Such success can readily lead to a search for other alleles or to complementation for confirmation. There will still be mutants that are hard to map, for example, due to genetic background incompatibilities or regions of substantial genomic rearrangement (Jander, 2006). In these cases, the availability of NGS data facilitates the ready design of markers for coarse and fine mapping in crosses with alternative non-reference accessions.

Next generation sequencing technologies offer an unprecedented possibility to sequence numerous Arabidopsis accessions, thereby enabling different biological processes to be investigated by uncovering the molecular basis for natural variation. Mapping QTLs in these accessions requires a good coverage with polymorphic markers. However, although a significant drop in the cost of next generation sequencing technologies will allow rapid generation of sequence data, the subsequent bioinformatics analyses to pinpoint the mutated gene requires highly specialized expertise that may be of limited availability and, consequently, the cost and pipeline savings may not live up to the initial promise. The sort of next generation sequencing-assisted map-based cloning described here is likely to provide a useful marriage of the two approaches.

Supplementary data
Supplementary data can be found at JXB online.

Supplementary Table S1. The Arabidopsis UPSC marker collection.

Acknowledgements
The authors would like to thank Dr Jonathan Jones and Dr Eric Kemen from The Sainsbury Laboratory for facilitating our access to Ws-0 sequence. This work was supported by the Swedish Natural Sciences Research Council (VR), the Swedish Foundation for Strategic Research (SSF), the Swedish Research Council for Research and Innovation for Sustainable Growth (VINNOVA), the K&A Wallenberg foundation, and the Carl Trygger foundation (CTS08:298).

References
Alonso JM, Ecker JR. 2006. Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in Arabidopsis. Nature Reviews Genetics 7, 524–536.

Alonso-Blanco C, Aarts MG, Bentsink L, Keurentjes JJ, Reymond M, Vreugdenhil D, Koornneef M. 2009. What has natural variation taught us about plant development, physiology, and adaptation? The Plant Cell 21, 1877–1896.

Anastasio AE, Platt A, Horton M, Grotewold E, Scholl R, Borevitz JO, Nordborg M, Bergelson J. 2011. Source verification of mis-identified Arabidopsis thaliana accessions. The Plant Journal 67, 554–566.

Aukerman MJ, Hirschfeld M, Wester L, Weaver M, Clack T, Amasino RM, Sharrock RA. 1997. A deletion in the PHYD gene of the Arabidopsis Wassilewskija ecotype defines a role for phytochrome D in red/far-red light sensing. The Plant Cell 9, 1317–1326.

Austin RS, Vidaurre D, Stamatiou G, et al. 2011. Next-generation mapping of Arabidopsis genes. The Plant Journal 67, 715–725.

Chen K, Wallis JW, McEllan MD, et al. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature Methods 6, 677–681.

Chi XF, Lou XY, Shu QY. 2008. Progressive fine mapping in experimental populations: an improved strategy toward positional cloning. Journal of Theoretical Biology 253, 817–823.

Delarue M, Prinsen E, Onckelen HV, Caboche M, Bellini C. 1998. Sur2 mutations of Arabidopsis thaliana define a new locus involved in the control of auxin homeostasis. The Plant Journal 14, 603–611.

Deschamps S, Campbell MA. 2010. Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. Molecular Breeding 25, 553–570.

Dharmasiri S, Dharmasiri N, Hellmann H, Estelle M. 2002. Arabidopsis map-based cloning in the post-genome era. The Plant Cell 14, 253–264.

Jander G. 2006. Gene identification and cloning by molecular marker mapping. Methods in Molecular Biology 323, 115–126.

Jander G, Norris SR, Rounsley SD, Bush DF, Levin IM, Last RL. 2002. Arabidopsis map-based cloning in the post-genome era. Plant Physiology 129, 440–450.

Laitinen RA, Schneeberger K, Jelly NS, Ossowski S, Weigel D. 2010. Identification of a spontaneous frame shift mutation in a non-reference Arabidopsis accession using whole genome sequencing. Plant Physiology 153, 652–654.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754–1760.

Lister R, Gregory BD, Ecker JR. 2009. Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. Current Opinion in Plant Biology 12, 107–118.
Lukowitz W, Gillmor CS, Scheible WR. 2000. Positional cloning in Arabidopsis. Why it feels good to have a genome initiative working for you. Plant Physiology 123, 795–805.

Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. 2008. Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Research 18, 2024–2033.

Peters JL, Cnudde F, Gerats T. 2003. Forward genetics and map-based cloning approaches. Trends in Plant Science 8, 484–491.

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 Physiology 127, 1579–1589.

Pollock DD, Larkin JC. 2004. Estimating the degree of saturation in mutant screens. Genetics 168, 489–502.

Rounsley S. 2003. Sharing the wealth. The mechanics of a data release from industry. Plant Physiology 133, 438–440.

Samson F, Brunaud V, Balzergue S, Dubreucq B, Lepiniec L, Pelletier G, Caboche M, Lecharny A. 2002. FLAGdb/FST: a database of mapped flanking insertion sites (FSTs) of Arabidopsis thaliana T-DNA transformants. Nucleic Acids Research 30, 94–97.

Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, Jorgensen JE, Weigel D, Andersen SU. 2009. SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nature Methods 6, 550–551.

Schneeberger K, Weigel D. 2011. Fast-forward genetics enabled by new sequencing technologies. Trends in Plant Science 16, 282–288.

Smolen G, Bender J. 2002. Arabidopsis cytochrome P450 cyp83B1 mutations activate the tryptophan biosynthetic pathway. Genetics 160, 323–332.

Sorin C, Russel JD, Camus I, et al. 2005. Auxin and light control of adventitious rooting in Arabidopsis require ARGONAUTE1. The Plant Cell 17, 1343–1359.

Stajich JE, Block D, Boulez K, et al. 2002. The Bioperl toolkit: perl modules for the life sciences. Genome Research 12, 1611–1618.

Torjek O, Berger D, Meyer RC, Mussig C, Schmid KJ, Rosleff Sorensen T, Weisshaa B, Mitchell-Olits T, Altmann T. 2003. Establishment of a high-efficiency SNP-based framework marker set for Arabidopsis. The Plant Journal 36, 122–140.

Uchida N, Sakamoto T, Kurata T, Tasaka M. 2011. Identification of EMS-induced causal mutations in a non-reference Arabidopsis thaliana accession by whole genome sequencing. Plant and Cell Physiology 52, 716–722.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25, 2865–2871.