Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in *Arabidopsis*

Jose M. Alonso and Joseph R. Ecker‡*

Abstract | Genome sequencing, in combination with various computational and empirical approaches to sequence annotation, has made possible the identification of more than 30,000 genes in *Arabidopsis thaliana*. Increasingly sophisticated genetic tools are being developed with the long-term goal of understanding how the coordinated activity of these genes gives rise to a complex organism. The combination of classical forward genetics with recently developed genome-wide, gene-indexed mutant collections is beginning to revolutionize the way in which gene functions are studied in plants. High-throughput screens using these mutant populations should provide a means to analyse plant gene functions — the phenome — on a genomic scale.

Whole-genome tiling microarray
A high-density oligonucleotide array that represents the majority of DNA sequences of an organism's genome.

Massively parallel signature sequencing
A sequencing procedure that allows the reading, in parallel, of short sequence segments, about 17 or 12 nt long, from hundreds of thousands of microbead-attached cDNAs.

**North Carolina State University, Department of Genetics, Raleigh, North Carolina 27695-7614, USA.*
‡The Salk Institute for Biological Studies, Plant Biology and Genomic Analysis Laboratories, La Jolla, California 92037, USA.
Correspondence to J.M.A. and J.R.E.
e-mails: jmalonso@unity.ncsu.edu; ecker@salk.edu
doi:10.1038/nrg1893
Published online 6 June 2006

After completion of the sequencing of the *Arabidopsis thaliana* genome in 2000 (REF. 1), the plant biology community faced the new challenge of assigning biological functions to all of the genes in this 120 Mb genome. Computational annotation of the genome was initiated to predict the locations of the genes and their basic structural elements (introns, exons and putative regulatory sequences) and led to rapid annotation of more than 25,000 *Arabidopsis thaliana* genes[1]. Although extremely useful, such *ab initio* annotation generates numerous inaccuracies — at least 40% of gene predictions were subsequently found to be erroneous[2,3]. Further refinement and validation of the computational gene models and identification of additional genes not predicted by gene-finding algorithms has been achieved using various experimental approaches[4,5].

The main source of empirical information about gene structure has been the capture and characterization of RNA transcripts (FIG. 1). A variety of high-throughput methodologies have been successfully used in *A. thaliana*, including ESTs and full-length cDNA sequences[2,6,7], whole-genome tiling microarrays[2] and gene-expression arrays[2,8], a massively parallel signature sequence technique[9], and serial analysis of gene expression[10,11]. Combined, they have provided conclusive evidence for the existence of thousands of new genes, as well as complete gene-structure information for ~18,500 genes (see The Arabidopsis Information Resource homepage). At the time of completion of the genome sequence, only ~10% of the 25,500 genes that were initially predicted had an experimentally assigned function[1]. Determination of the functions of the remaining 90% of genes presents a tremendous challenge, not only because of the large number of genes to be examined, but also because defining what constitutes a 'gene' is itself a complex problem[12].

Recent improvements in genomic technology allow genome-wide capture of some basic information, such as the determination of gene or protein expression levels using microarrays or shotgun mass spectrometry[13–15]. Although informative, these types of data alone are typically not sufficient to define the function of a gene, as by its very nature this information is largely correlative. In fact, studies in yeast suggest that functional inferences that are based on gene-expression data can be misleading in many cases; for example, Giaever *et al.* found that most mutants that showed altered fitness in particular stress conditions corresponded to genes that were not transcriptionally regulated by that type of stress[16].

Mutant analysis provides an alternative and typically more reliable means to assign gene function[17]. However, this 'phenotype-centric' process, classically known as forward genetics, typically is not suitable for systematic genome-wide gene analysis, primarily

owing to the enormous effort required to identify each gene responsible for a particular phenotype[18]. In spite of improvements in the cloning of genes on the basis of phenotype (such as the availability of whole-genome sequences, large numbers of mapped polymorphisms, and faster and cheaper genotyping technologies), it can often take over a year for a skilled scientist to move from a mutant to the affected gene[19] (BOX 1).

Faster and easier ways to test hypotheses about specific gene functions are expected in the post-genome era. Various reverse genetic approaches have been developed that leverage the flood of available information from whole-genome sequences in order to understand gene function. To identify a mutant of interest, these technologies primarily rely on predictions of function that are based on DNA sequence homology[20–24]. In the process of generating tools for reverse genetics, large collections of gene-indexed mutations have been created. The expected improvement and expansion of these mutant collections in the near future promises to aid biologists in meeting the main challenge of the post-genome era: the systematic analysis of the functions of all genes in a genome.

In this review, we discuss the challenges associated with characterizing gene function in the post-genome era in the reference plant *A. thaliana*. Classical and more recent genome-wide experimental approaches are described and compared in the context of the complexities of systematic studies of gene function in this model organism. We conclude with a discussion of how several new and up-and-coming technologies could profoundly affect this field in the near future.

Perturbing gene activity

Random mutagenesis. In both forward and reverse genetic approaches, the function of a gene is typically assigned to a specific biological process by analysing the phenotypic consequences of altering that gene's activity. A wide variety of physical, chemical and biological agents can be used to create genetic alterations in *A. thaliana*[17], each of which has a different efficiency (the number of events per genome that are inherited in a population that has been mutagenized under standard conditions) and molecular outcome (ranging from a single nucleotide substitution to an insertion or a rearrangement of large chromosomal regions[25]).

Fast neutrons, γ-radiation and carbon ions are most commonly used for mutagenesis in *A. thaliana*[25,26]. In general, the efficiency of high-energy ionizing radiation is intermediate, and the induced alterations typically include chromosomal deletions and rearrangements, although point mutations are not uncommon[25,26]. Although high mutagenicity means that fewer plants need to be screened to identify mutants with the desired phenotype, it also means that numerous unrelated mutations must be removed before the mutant phenotypes can be analysed (typically, by several rounds of backcrossing to the parental non-mutagenized line).

The nature of the damage that is caused by mutagenesis determines the functional class of genetic alterations that are produced. Deletions, insertions and rearrangements are more likely to result in loss-of-function alleles, whereas point mutations can lead to a broader range of effects, including hypomorphic, hypermorphic and neomorphic effects (that is, alleles of reduced, enhanced or novel gene function, in corresponding order). The nature of the damage also has implications for the experimental

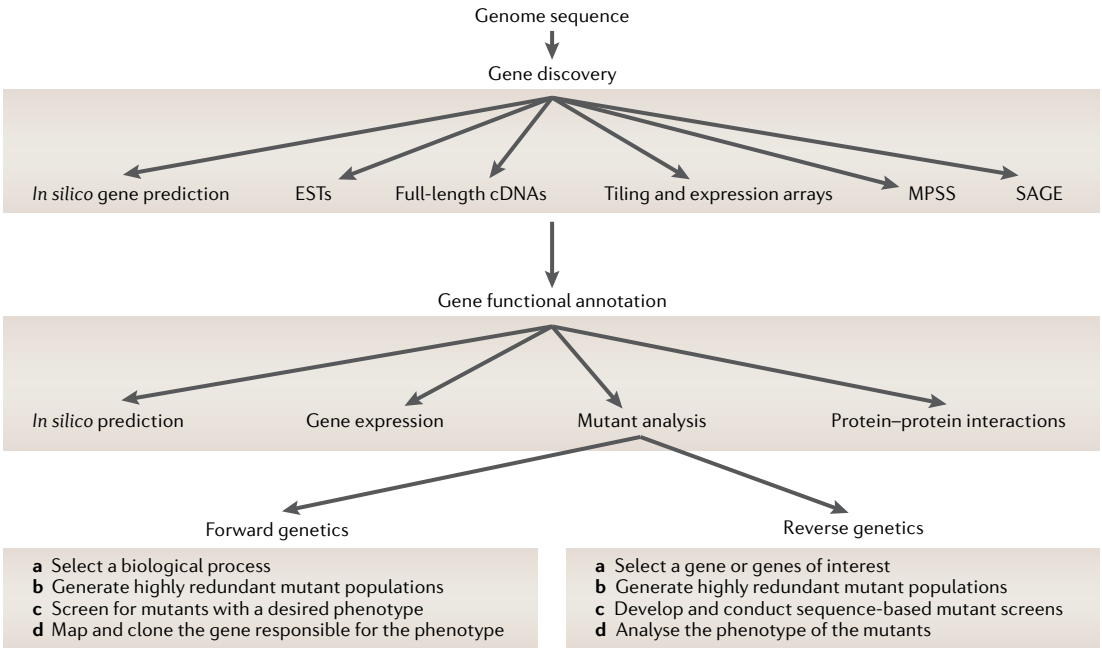


Figure 1 | **From genome sequence to gene function.** Steps and experimental approaches that are used in the functional annotation of the genome. MPSS, massively parallel signature sequencing; SAGE, serial analysis of gene expression.

Serial analysis of gene expression
A technique that is used to obtain short sequence tags (typically 16 nt long) from large numbers of cDNA clones by cutting, concatemerizing and finally sequencing cDNA fragments.

Shotgun mass spectrometry
A bottom-up proteomics approach that is used in the identification of individual components of a complex protein mixture by the identification of peptide fragments on the basis of their mass.

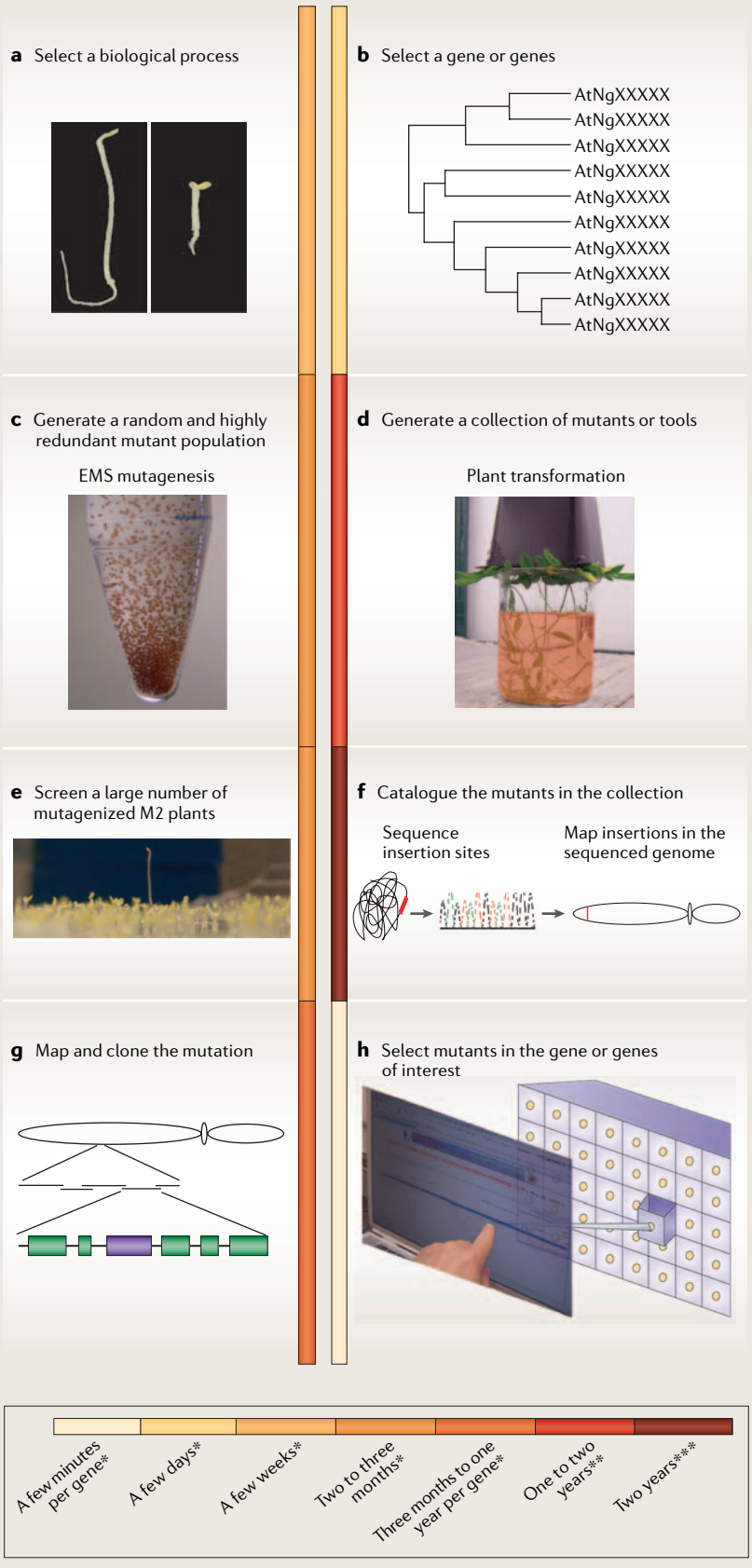
Box 1 | **Forward versus reverse genetic approaches**

The selection of a biological process is the first step in forward genetics, illustrated in panel **a** by the response to ethylene. Three-day-old *Arabidopsis thaliana* seedlings that were grown in the dark in the absence (left panel) or presence (right panel) of ethylene are shown. The clear difference between treated and untreated plants is an excellent morphological trait for use in a screen. By contrast, a typical reverse genetic screen starts with a set of genes of known sequence that are of particular interest (in panel **b** a hypothetical family of genes (generic *Arabidopsis* nomenclature is used) is selected on the basis of their sequence similarity).

The next step in both forward and reverse genetics is to generate a population of mutants. After mutagen treatment (panel **c**), seeds are germinated in soil and grown in families (with ~1,000 individual mutagenized plants per family) to maturity to produce the M1 generation. Their progeny (M2) are typically used for phenotypic screening. For transferred DNA (T-DNA) mutagenesis, adult plants are transformed by submerging the floral buds in a suspension of *Agrobacterium tumefaciens* cells that harbour an appropriate T-DNA-containing vector (panel **d**). The seeds of these plants are grown in the presence of a selective medium, so that only the plants expressing the resistance marker gene that is supplied by the T-DNA survive. Each transformant (T1) results from an independent transformation event and is hemizygous for one or more insertion. The progeny of a selfed T1 plant (T2 family) is composed of a segregating population of wild-type plants, and hemizygous and homozygous mutants. The T2 generation is usually used in screens. Panel **e** shows a typical phenotypic forward genetic screen of M2 seeds. A mutant resistant to the hormone can be easily identified among thousands of normally responding plants.

In a reverse genetic screen, once the T-DNA mutants are obtained, two different approaches can be used: T-DNA pools are screened by PCR (FIG. 2) or the insertion sites are mapped in individual T2 families (shown in panel **f**). To map the insertion sites genomic DNAs are obtained from individual T2 families, and DNA fragments flanking the insertion sites are amplified by PCR and sequenced. Sequencing of hundreds of thousands of insertion sites results in a gene-indexed catalogue of mutants.

The last step in a forward genetic approach is the identification of the gene that is responsible for the mutant phenotype. Panel **g** illustrates the series of mapping steps required to determine the position of the mutation in the genome, starting from a large chromosomal region and zooming in to the gene level. Sequencing and complementation studies are required for conclusive mutant identification. The reverse genetics counterpart is extremely fast and easy. Once a catalogue of gene-indexed mutants has been obtained, it can be searched for the presence of mutants in any gene of interest. Once found, seeds of the corresponding mutant(s) can be ordered from one of the public stock centres (panel **h**). It is important to point out, however, that in the reverse screening approach, this step is the starting point of a sometimes tedious search for the mutant phenotype, whereas in the forward genetic approach, this step is already accomplished (panel **a**).
*Requires one skilled person. **Requires two to five skilled people. ***Requires five to ten skilled people. EMS, ethyl methanesulphonate.



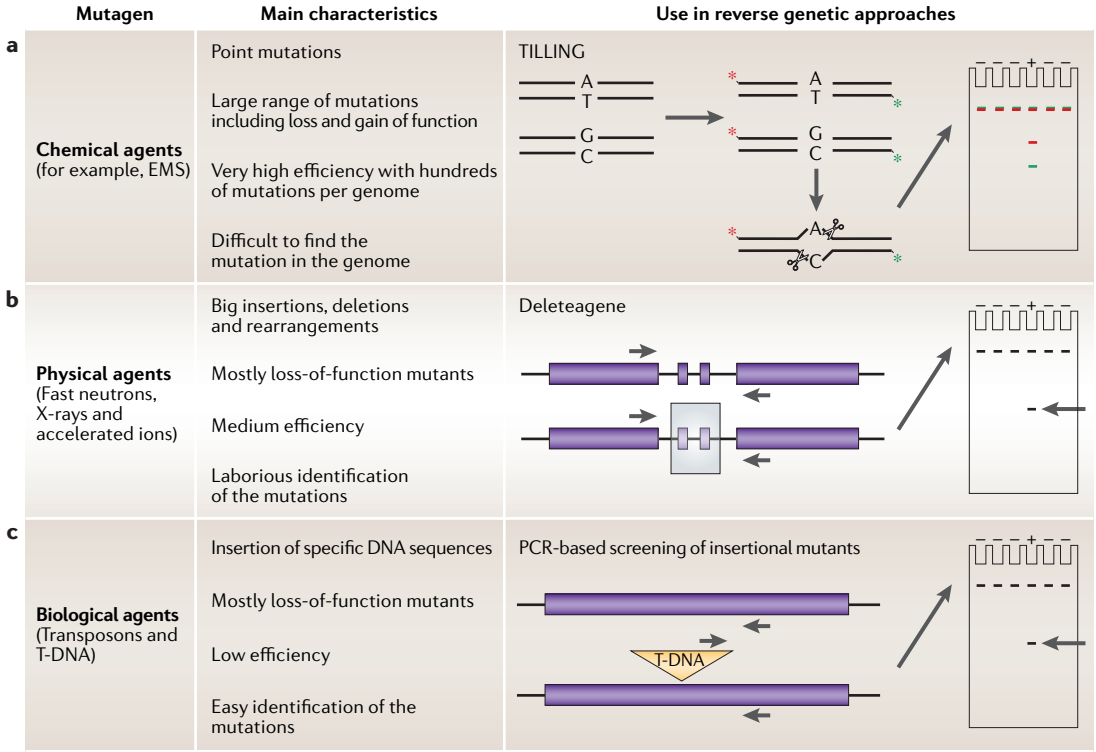


Figure 2 | **Generation of genetic diversity. a** | Point mutations can be identified using TILLING (targeting induced local lesions in genomes). DNAs from pools of EMS (ethyl methanesulphonate)-mutagenized plants are PCR-amplified using primers that are specific for the gene of interest. Because each primer is labelled with a different fluorescent dye, both DNA strands can be visualized simultaneously. Amplified DNA products are denatured and then allowed to renature, leading to the formation of homodimers and heterodimers between the wild-type and mutant DNA (only heterodimers are shown for simplicity). Mismatches are recognized and cleaved by the mismatch-specific celery nuclease CELI. The presence of a mutation in a pool is then determined by separating the products of the CELI reaction in a denaturing gel. The presence of two different size bands that are labelled with different fluorescent dyes in a given pool indicates the presence of a mutation in the gene of interest. The specific sizes of the bands provide information about the position of the mutation in the gene. Mutations are confirmed by sequencing. **b** | Identification of deletions from pools of plants mutagenized with fast neutrons. DNA from the pools of mutants is PCR-amplified using primers that flank the gene of interest. The amplification products are then analysed by gel electrophoresis. In some pools, a band that is smaller than that predicted for the wild type (marked with an arrow) can be detected, which indicates the presence of a deletion in one of the plants in that pool. Secondary screening of individual plants in the pool confirms the presence of the mutation and identifies individual mutants. **c** | The presence of a transferred DNA (T-DNA) insertion in a gene can be detected using primers that are specific for the gene of interest and for the T-DNA. DNA from mutant plants is used as a template for PCR amplification. The presence of an amplification product (marked with an arrow) indicates the possible existence of an insertion in or near the gene of interest. Individual plants from the pool are tested to identify the mutant line. Alternatively, the insertion sites can be systematically identified, as indicated in BOX 1.

strategy that is best suited to identifying the genomic location of the mutation. Methods such as Deleteagene[23] (see below for details), subtractive hybridization (illustrated by the cloning of a gibberellin biosynthetic mutant *ga1* using a fast neutron allele *ga1-3* (REF. 27)), or high-density microarray-based mapping[28,29] take advantage of chromosomal deletions for rapid gene identification. Other types of DNA alteration, such as point mutations, require different and, traditionally, more tedious experimental approaches (FIG. 2).

Chemical agents, such as ethyl methanesulphonate (EMS) and nitrosomethylurea (NMU), are extremely efficient mutagens in *A. thaliana*. Under optimal conditions, EMS treatment of seeds can generate about 400 mutations per genome, compared with an average of 1.5 insertions per transferred DNA (T-DNA) mutant (see below for details)[30,31]. Owing to their high efficiency, chemical agents are the first choice in exploratory mutagenesis, where the frequency of specific mutant phenotypes needs to be evaluated[32]. Chemical agents generate a broad range of DNA alterations; these are predominantly single base-pair substitutions, but also include small insertions and deletions, as demonstrated in a recent large-scale study of more than 1,900 EMS-induced mutations in *A. thaliana*[31,33]. Importantly, the distribution of EMS-induced mutations is unbiased. But chemical mutagenesis has one main drawback — finding the DNA alteration that is responsible for a mutant phenotype, which is typically a single base change, is difficult in a large genome. This handicap of chemical

Table 1 | **Methods used for reverse genetics in plants**

Method	Advantages	Disadvantages
Homologous recombination*	• Allows for exact replacement or modification of the targeted gene • Highly specific to the target gene (no off-target effects) • Results in stable mutations	• Very low efficiency • Low throughput
Gene silencing‡	• Possibility of generating allelic series, allowing the study of essential genes • Possibility of restricting the alterations to specific tissues or developmental stages • Study of gene families with high degree of functional redundancy • Can be adapted for high-throughput screens	• The degree of gene silencing is unpredictable • Risk of off-target effects • Instability of phenotypes
Ectopic expression‡	• Similar to gene silencing • Allows for the analysis of gain-of-function alleles • Can be adapted for high-throughput screens	• Similar to gene silencing • Possibility of generating misleading neomorphs • Its use is limited to transformable species
Zinc-finger nucleases*	• Highly specific • Results in stable mutations	• Low throughput • Its use is limited to transformable species
TILLING‡	• Allows the identification of loss-of-function alleles, hypomorphs and gain-of-function alleles • Can be used in non-transformable species • Results in stable mutations	• Based on random mutagenesis, so the desired mutation might never be found • Low to medium throughput
Deleteagene*	• Allows the identification of two or more genes in close proximity • Can be used in non-transformable species • Results in stable mutations	• Based on random mutagenesis, so the desired mutation might never be found • Limited to loss-of-function mutations • Low to medium throughput
Insertional mutagenesis§	• High throughput • Can be adapted for both loss-of-function and gain-of-function studies • Results in stable mutations • Few unwanted mutations	• Based on random (T-DNA) or non-targeted (transposon) mutagenesis, so the desired mutation might never be found • Cannot be used to study tandemly repeated genes (T-DNA mutagenesis) • Only limited information can be obtained for essential genes

*Each of these approaches has so far been of limited practical use, with fewer than 10 successful examples reported. ‡Each of these approaches was used as a tool to study gene function in more than 50 publications. §There are more than 500 publications that used this technique. T-DNA, transferred DNA.

mutagenesis, which is the method of choice in forward genetic screens, is even more serious when applied in the reverse genetic mode, where the detection of the sequence alteration is the only way of finding the mutant. However, the development of highly sensitive methodologies for point-mutation detection — such as targeting induced local lesions in genomes (TILLING) (see below for details)[31,34,35] — is now allowing the incorporation of the power of chemical mutagenesis into the reverse genetic toolbox. In addition, new methods for efficient genome-wide detection of point mutations, such as mismatch-repair detection on tag arrays, are appearing on the horizon[36].

Biological agents that mediate the transfer of DNA or RNA molecules into plant cells can also be used as mutagens. Depending on whether the DNA or RNA is used to disrupt the sequence of a gene or to alter its activity indirectly, one can differentiate between insertional and gene-replacement mutagenesis[37] versus overexpression and gene-silencing methods[38,39]. In either case, the first step is the introduction of an engineered DNA, or sometimes RNA, molecule into living cells. The most common method of introducing a DNA molecule into plant cells relies on *Agrobacterium tumefaciens* as a shuttle[40]. These soil-born bacteria can infect plant cells and transfer specific sequences that are contained in their Ti plasmid into the plant genome[40,41]. Most of this plasmid's sequence can be replaced without affecting transfer efficiency[42]. Several other types of plant-associated bacteria, such as *Rhizobium* spp., *Sinorhizobium meliloti* and *Mesorhizobium loti*, are also capable of horizontal transmission and integration of plasmid DNA into plant chromosomes, opening up the possibility of gene transfer to plant species that are resistant to *Agrobacterium*-mediated transformation[43]. It is also possible to use DNA or RNA viruses to deliver genetic information into plant cells[44]. In this case, the new genetic material remains associated with the viral genome, which limits its use in the transient alteration of gene function (TABLE 1; see below for further discussion). Alternatively, cells can be bombarded with DNA-coated microparticles[45].

Arabidopsis thaliana can be stably transformed by submerging floral buds into a suspension of the appropriate *Agrobacterium* strain. Owing to the simplicity and ease of this 'floral dip' method[46], *Agrobacterium*-mediated transformation has been widely used in gene-function studies. In the simplest case, a selectable marker, such as a herbicide resistance gene or an antibiotic resistance gene, is inserted into the T-DNA, in addition to the T-DNA left and right border sequences that mediate the T-strand movement from the bacterial plasmid to the plant cell nucleus[41]. If the T-DNA inserts within the boundaries of a gene, it can drastically alter, and in many cases completely abolish, gene function[47]. Such simple Ti plasmids that carry selectable markers have often been the preferred type of vector in the generation of large T-DNA mutant collections that are designed exclusively to yield 'knockouts'[30,48].

Mismatch-repair detection on tag arrays
A hybridization-based technique that combines a bacterial selection assay and microarray analysis to rapidly scan large genomic regions for the presence of DNA polymorphisms.

Random insertional mutagenesis can be used in many other ways. T-DNA or transposons can carry promoterless reporter genes (such as β-glucuronidase, *GFP* or luciferase), which can 'trap' the regulatory sequences of the 'tagged' gene[49,50] — as a result, the reporter gene will display the original expression patterns of the tagged gene[49]. Promoter trap lines provide an excellent tool for studying gene-expression patterns[49]. Systematic analysis of trap lines provides molecular markers that can be used to monitor *in vivo* developmental processes, such as flower[51] or root[52] morphogenesis, responses to stress[53] and circadian rhythms[54].

Owing to its disruptive nature, T-DNA mutagenesis is commonly associated with loss-of-function or hypomorphic mutations. However, it can be adapted to generate gain-of-function alleles by activation tagging[38]. To achieve this, several copies of a strong transcriptional enhancer are introduced into the T-DNA. On integration, the enhancers stimulate the transcription of a nearby gene and cause its ectopic expression. Assigning biological function to a gene on the basis of the phenotypic effects of its overexpression should be done with caution, as illustrated by the initial assignment of the histidine kinase CKI1 (CYTOKININ-INDEPENDENT 1) as a putative receptor for the hormone cytokinin[55]. Subsequent studies have not only failed to confirm the role of CKI1 in cytokinin reception, but instead showed that three histidine kinases that are related to CKI1 (CRE1 (CYTOKININ RESPONSE 1; also known as WOL), AHK2 and AHK3 (ARABIDOPSIS HISTIDINE KINASE 2 and 3)) are the *bona fide* cytokinin receptors[56]. Nevertheless, there are many examples that highlight the utility of activation tagging in gene-function studies[57–60], especially in plant genomes in which high levels of gene duplication and compensatory mechanisms make the analysis of loss-of-function mutations challenging[61–63].

In even more sophisticated approaches, transposons can be introduced into the plant genome using T-DNA-mediated transformation. Once inserted, the transposon can 'hop' from one chromosomal location to another, as long as an active transposase is present, with the potential of creating mutations at both the landing and excision sites[64]. Although most transposons tend to hop to linked sites, a strategy has been devised to select for transpositions that land at unlinked loci[65]. The mobile element and the T-DNA integration site launch pad have been engineered to harbour positive and negative selectable markers, respectively. Plants that are able to grow in the presence of both the positive and negative selection should be greatly enriched for unlinked transposition events.

In most cases, it is desirable to fix the insertions after a single hop, so that the phenotype of a stable mutant can be studied. The most common way to regulate transposon activity is to introduce both transposon and transposase into the same plant (typically by co-transforming plants with T-DNA constructs bearing both[66] or by crossing a transposon-harbouring plant with a plant harbouring an active transposase[67]) and then to separate them (usually by crossing) to stabilize the new mutations. Typically, once the transposon has been allowed to jump, plants with new insertions that do not contain an active transposase are selected from the progeny of the cross to create stable mutant (transposant) lines. Therefore, although the generation of large numbers of transposon insertions is straightforward, considerable effort is required to select plants that have stable insertions[65].

The mobile properties of transposons have been exploited in several other applications. For example, reversion of the mutant phenotype after remobilization of a transposon has been used to prove causal relationships between a gene and a mutant phenotype[49,68,69]. Transposon remobilization has also been used to generate sectors of heterozygous cells in a homozygous mutant plant to determine whether or not a phenotype is cell autonomous[49,70,71]. Transposons can also be engineered to contain simple selectable markers, more elaborate gene traps or transcriptional enhancer systems[49,72].

A recently developed application takes advantage of the predominance of local transposition to study loss of function within gene families[73]. Approximately 17% of the predicted genes in *A. thaliana* are tandemly repeated[1]. Functional analysis of tandemly repeated genes is problematic because generating double mutants of tightly linked loci by crossing is difficult. Local transposition, together with the high frequency of DNA lesions at the site of transposon excision, allow for the simultaneous disruption of several genetically adjacent loci[64,74]. Furthermore, a specialized T-DNA system composed of three elements — a transposon, two *loxP* recombination sites (one inside the transposon and one in the immobile part of the T-DNA), and an inducible *loxP* recombinase — has been developed to generate deletions of selected genomic regions[73]. Using a similar approach, Krysan and colleagues have generated a population of ~10,000 randomly inserted *loxP* T-DNAs for which the insertions sites have been determined (see the New Wisconsin T-DNA Lines — pDs–Lox Vector web site). Once transposition is activated in a selected line (by crossing in an active transposase) the recombinase is activated, which leads to permanent removal of the genomic region that is contained between the *loxP* elements in the transposon and the immobile part of the T-DNA. This strategy could be useful for studying functionally redundant genes that are located in adjacent chromosomal locations, which is a common phenomenon in *A. thaliana*.

The aforementioned approaches, all of which produce lesions in unpredictable chromosomal locations, can be used in forward and reverse genetic approaches. There are, however, three main requirements that must be met for a collection of random mutations to be efficiently used in a reverse genetic screen. First, the number of mutations in the collection should exceed (by five to tenfold) the number of genes in the genome. This redundancy is necessary to ensure that mutations in a particular gene will be found with a sufficiently high probability. Second, each individual plant in the mutagenized population should be catalogued, propagated and pooled so that it can then be effectively screened. Third, and perhaps most important, a DNA

loxP
A sequence that is specifically recognized by the Cre recombinase from the bacteriophage P1. Cre catalyses the recombination between two *loxP* sequences.

sequence-based screening approach needs to be developed that is sensitive enough for a single plant with a specific sequence alteration to be detected within a pool of wild-type individuals. Different types of DNA lesion (deletions, insertions and point mutations) require different screening methodologies (see below).

Targeted mutagenesis. True directed mutagenesis approaches, in which the researcher chooses the gene to be perturbed, are widely used in *A. thaliana*. The most elegant and precise targeted mutagenesis approach relies on homologous recombination to target foreign DNA to homologous sequences in the host genome. Although a few successful attempts have been reported in plants[75,76], the extremely low frequencies of recombination events make this approach ineffective for gene-function studies in *A. thaliana* (and in most other eukaryotes)[77]. Improved selection strategies and a better understanding of the basic molecular mechanisms involved in homologous recombination indicate that targeted gene replacement could become a reality in the near future[78,79]. For example, Shaked and co-workers[80] recently found that overexpression of a yeast SWI/SNF chromatin remodelling enzyme, RAD54, in *A. thaliana* resulted in a 27-fold increase in gene targeting. If these results can be reproduced for other genes, this improvement in efficiency, in combination with stringent selection procedures, will make gene targeting a routine approach in *Arabidopsis* research. Meanwhile, alternative approaches have been developed to alter the expression of selected genes. There are two main variants of directed mutagenesis (or methods of alteration of gene expression): gene-silencing approaches and zinc-finger nucleases, which are still in an experimental phase[81]. In these strategies, specific sequences that are unique for each gene to be disrupted must be engineered *in vitro* and then introduced into the plant. *Agrobacterium tumefaciens*[40], viruses[44] or particle bombardment[45] have all been used to mediate the transfer of such sequences.

First described in plants[82,83], gene silencing seems to have evolved as a defence mechanism in most eukaryotes against viruses and active endogenous transposons[44]. This mechanism, also called RNA interference (RNAi), co-suppression or post-transcriptional gene silencing, can be exploited for attenuating gene expression by introducing specific sequences into the plant cell. Once a dsRNA of choice is expressed, a silencing signal is generated that can spread throughout the plant, causing a reduction in the transcript levels of the endogenous gene or genes in a sequence-specific manner[39]. In contrast with insertional mutagenesis, RNAi lines typically show a wide range of effects on gene expression, from complete inhibition to no reduction[39]. Although this approach allows the analysis of allelic series that could be important in the study of essential genes, the variability of the silencing effects in different plants and tissues, and even between genes or gene constructs, complicates the interpretation of results and necessitates the analysis of several independent RNAi lines. Another drawback stems from the fact that during the processing of the dsRNA large numbers of different small interfering RNAs (siRNAs) can be generated, making it difficult to predict and therefore avoid off-target effects[84]. Therefore, the sequences that are used to generate the dsRNA must be carefully selected, because even stretches as short as 10–11 nucleotides that are homologous between the RNA and an endogenous sequence can result in silencing[84].

A possible solution to some of the problems associated with using siRNAs might come from the use of microRNAs (miRNAs)[85]. miRNAs can regulate the expression of genes to which they are only partially complementary, making the identification of all potential miRNA targets challenging. However, although such sequence promiscuity has been found for both plant and animal miRNA targets, genome-wide studies in *A. thaliana*[8] indicate that plant miRNAs have fewer targets than animals, perhaps as a result of tighter mechanisms governing target recognition. Recent genome-wide studies of the effects of several artificial miRNAs (amiRNAs) on gene expression support this hypothesis and identify key sequence determinants that underlie miRNA specificity in plants[86]. It should now be possible to design amiRNAs that selectively target genes of interest, making this technology an excellent tool for dissecting the function of essential genes. Most importantly, these types of technology (amiRNA and RNAi) will facilitate the functional characterization of a potentially large number of *Arabidopsis* genes that have overlapping or partially redundant functions by making RNA molecules that are specific to a family of related genes rather than to a single gene[87–89].

One important obstacle to a more widespread use of the sequence-directed mutagenesis approaches described above is that altering the function of the targeted gene relies on the activity of a transgene. The interpretation of these studies is also complicated by the fact that transgene expression can vary markedly from one plant to another and from one generation to the next[90]. Recently, a novel sequence-directed mutagenesis methodology has been described that uses zinc-finger proteins to alter the sequence of a targeted gene and, therefore, results in stable mutations[81]. Zinc-finger proteins that recognize specific DNA sequences can be engineered. Specific sequences can be cleaved by combining these zinc-finger domains with a dsDNA endonuclease in a single protein[91]. It has recently been shown that the expression of a sequence-specific zinc-finger nuclease in *A. thaliana* generates mutations (deletions and insertions) in the target gene *in planta*[81]. Stable mutations in the gene of choice can be isolated because the permanent DNA damage that is inflicted by the nuclease is heritable[81]. The large battery of well-characterized zinc fingers, each with different and specific DNA-recognition sequences, should allow the use of this methodology to target almost any gene in the *A. thaliana* genome. Chromosome breaks that are induced in this way can enhance the rate of homologous recombination in specified target locations[92], opening another door to the future use of homologous recombination in plants.

RNA interference
A form of gene silencing in which dsRNA induces the degradation of homologous endogenous mRNA transcripts, thereby mimicking the effect of reduction, or loss, of gene activity.

Co-suppression
Silencing the expression of an endogenous gene, which is caused by the expression of a transgene bearing high levels of sequence identity with the endogenous gene.

Post-transcriptional gene silencing
Refers to the general mechanisms that are involved in gene silencing at the post-transcriptional level independently of the initial silencing agent, that is, double-stranded, antisense or sense RNA.

Small interfering RNA
20–25 nucleotide long RNA molecules that are generated during the post-transcriptional gene-silencing process and act as key determinants of the sequence specificity of the silencing mechanism.

MicroRNA genes
Genome-encoded or artificial genes; the RNAs that are transcribed from them are processed to generate small RNA fragments (20–25 nt long) that target specific mRNAs for degradation or inhibit their translation into proteins.

Ectopic expression. In addition to triggering silencing, gene-specific sequences can be introduced into the plant genome to enhance gene expression. In the simplest case, a complete gene, including its regulatory and coding regions, is introduced by *Agrobacterium*-mediated transformation into the plant genome. This approach is typically used in complementation assays, in which a loss-of-function mutant is transformed with the wild-type gene to confirm the phenotype–genotype relationship or to study the effect of gene dosage.

In another application, the regulatory region of one gene is fused with the coding region of another. Strong constitutive promoters, such as cauliflower mosaic virus 35S (*CaMV35S*), are usually preferred in initial studies in which the objective is to obtain general functional information[93,94]. In other cases, it is often desirable to analyse the effects of ectopic activation of a gene of interest in a specific cell type or developmental stage. This laborious process is facilitated by use of the *GAL4–GFP* enhancer trap system, which was initially developed in *Drosophila melanogaster*[87] and later adapted for *A. thaliana*[88,89]. In this approach, wild-type plants are transformed with a trap construct that harbours a selectable marker and two other genes: a promoterless gene that encodes the yeast transcription factor *GAL4*, and a reporter gene (typically *GFP*), which is under the control of a minimal promoter and a UAS (upstream activating sequence), which is recognized by the GAL4 transcription factor. When the T-DNA that harbours the reporter construct lands in a gene, the regulatory sequences of that native gene are 'hijacked' by the T-DNA and used to drive the expression of *GAL4* in a manner characteristic of the particular 'trapped' gene promoter. The GAL4 transcription factor will bind to the UAS and activate the expression of *GFP* or any other gene that is placed under the control of the UAS. By monitoring GFP fluorescence, one can easily determine where and when *GAL4* is expressed in a particular trap line. A catalogue of lines that display specific *GAL4–GFP* expression patterns (*GAL4–GFP* trap lines) can be generated by crossing a characterized *GAL4–GFP* line to a plant in which the transgene of interest is under the control of the UAS promoter. Although such collections have been generated, the frequent loss of GFP expression in the trap lines, which is due to silencing in later generations, limits their utility. It might be possible to alleviate such problems by using promoters other than *CaMV35S* to drive *GAL4* expression, which could result in less frequent activation of the cellular gene-silencing machinery.

With this tremendous battery of tools at hand, three main strategies can be used to study gene function: classical forward genetics, classical reverse genetics and systematic reverse genetics.

Classical forward genetics

Forward genetics has been a powerful approach to studying the genetic components of almost every signalling or biochemical process in *A. thaliana*[95,96]. One classical example is the identification of mutants that express the nuclear chlorophyll a/b binding protein (*CAB*) gene in the absence of functional chloroplasts to study nucleus–chloroplast communication[97]. Another involves forward screens for mutations that affect specific and marked morphological changes induced in dark-grown seedlings by ethylene to uncover a large number of ethylene-related loci that are involved in either biosynthesis, perception, signalling or response to this gaseous plant hormone[98]. In another example, the circadian cycling of *CAB* gene expression has been exploited to develop sensitive molecular markers to study biological rhythms in plants[99]. By fusing a *CAB* promoter with a *luciferase* reporter gene, Kay *et al.* were able to closely monitor the circadian oscillations *in planta*, allowing them to isolate mutants with altered circadian rhythmicity[99]. High reproducibility, sensitivity and specificity, together with the potential for high throughput, are the main desirable characteristics of a phenotype in forward genetic screens. Commercial and community resources such as Lehle Seeds, Arabidopsis Biological Resource Center Stocks, and the Nottingham Arabidopsis Stock Center now provide different types of 'ready-to-screen' mutagenized seed, including lines that have been mutagenized using EMS, fast neutrons, T-DNA and transposons. Most have been generated in a limited number of widely used genetic backgrounds (such as Columbia, Wassilewskija and Landsberg *erecta*), rendering these established collections unsuitable for highly specialized screens in which the use of a particular molecular reporter or mutant background is desired. Therefore, individual research groups often have to generate their own mutagenized plant populations.

One of the most attractive features of forward genetic approaches is that they represent truly unbiased gene-discovery processes — no preconceived idea about the nature of the gene involved in the process is required.

Reverse genetics

The large amount of sequence information that has been generated by genome projects for *A. thaliana*[1] and other plant species[62,63], together with the implementation of genome-wide approaches to study gene expression, protein–protein interactions and other aspects of plant biology, have resulted in an increased interest in reverse genetic methodologies. Numerous examples highlight the recent success of these reverse genetic approaches. Analysis of multigenic families, in which functional redundancy has made forward genetic approaches difficult, has particularly benefited from reverse genetic approaches. Entire gene families have been knocked out, uncovering overlapping and specific functions among their members[100–102]. If the size of a gene family is too large to disable all of its members, criteria other than sequence similarity are often applied to investigate their functions. For example, a combination of sequence comparison and expression profiling was used to identify, out of over a hundred related genes, a small subfamily of four transcription factors that are involved in the early response to ethylene[30]. Alternatively, approaches based on gene silencing can, in principle, be used to alter the expression of several members of a gene family. However, the utility of this approach is yet to be demonstrated.

There are two main approaches to finding mutations within a gene on the basis of its DNA sequence: using one of the targeted techniques such as RNAi or ectopic expression, or screening a collection of randomly generated mutants for a knockout. The ease of producing large collections of random mutants, the genetic stability of the resulting mutations, and the recent development of high-throughput sequence-based screening strategies such as TILLING and Deletagene (see below) have made screening for knockouts an essential tool in reverse genetics[103]. Perhaps, the most important single prerequisite for choosing a collection of randomly generated mutations for a reverse genetic screen is the availability of highly efficient screening methods to identify sequence alterations in genomic DNA. The number of individual mutants that are screened tends to be very high (in some cases, in the order of tens of thousands), making it necessary to pool individual mutants before testing for the presence of the mutation of interest. The optimal strategy for screening these DNA pools depends on the type of mutagen used or, more specifically, the nature of the DNA lesion.

TILLING. TILLING has been developed to identify mutations in EMS-mutagenized populations[34,35]. Detection relies on the amplification of the gene of interest from pooled DNA. After PCR amplification, the DNA is denatured, renatured and then digested with the mismatch-specific celery nuclease CELI, which recognizes base-pair mismatches[104]. The existence of a mutation in a particular DNA pool can be identified by the presence of two DNA fragments on a denaturing acrylamide gel — which is a result of nuclease activity at the site of the mismatch between the wild-type and mutant heteroduplexed DNA fragments (FIG. 2a). DNA from the individual mutants that comprise the positive DNA pool is screened to identify the individual plant that carries the mutation. The main advantage of this strategy is that it integrates a reverse genetic approach with the benefits of chemical mutagenesis, which enables a wide range of genetic alterations to be produced: not only loss-of-function alleles, but also hypomorphic, hypermorphic and neomorphic mutations[105]. Nonetheless, as this technology remains relatively labour intensive, it is typically relegated to second choice and is usually relied on when sequence-indexed mutants are not available or are not useful for specific studies (for example, in the case of lethal mutations).

Deleteagene. Experimental approaches for identifying DNA deletions in pools of mutants that are generated by high-energy ionizing radiation have also been developed[23]. Although not yet widely popular in *A. thaliana*, Deleteagene has several appealing advantages. For example, it can be applied to plants in which transformation is inefficient; it might also provide a means to simultaneously mutate (delete) tandem duplicated genes[23,73]. Deleteagene uses PCR primers that flank the gene(s) of interest and DNA from pools of mutagenized plants as a template. The presence of an amplification product that is smaller than that expected for the wild type identifies a deletion mutation (FIG. 2b). In spite of the potential advantages of this technology, its use in *A. thaliana* has been limited[106,107], possibly in light of competition with other successful reverse genetic approaches. In non-model species, however, the wide use of Deleteagene is hindered by the prerequisite of knowing the sequence of the genes of interest and the need for a large infrastructure to generate big mutant populations.

Insertional mutagenesis. Identifying a T-DNA or transposon insertion in a gene of interest is even more straightforward. Gene-specific primers and T-DNA-specific or transposon-specific primers are used in PCR amplification from a DNA template that is derived from pools of mutant plants. The presence of specific PCR products in a particular pool indicates that a T-DNA or transposon has inserted in or near the gene of interest in one of the plants (FIG. 2c). One drawback of this approach is that PCR using such complex pooled samples often yields a high background of nonspecific amplification products. This problem can be overcome by introducing a hybridization step (Southern blotting with a gene-specific label) to identify the true-positive pools. This is by far the most widely used reverse genetic approach in *A. thaliana*. It has enabled the identification of hundreds of specific mutants since the first successful screens were carried out more than 10 years ago[108].

Insertional mutagenesis also has its limitations, such as the predominance of loss-of-function alleles, the biased distribution of insertions in the genome, the inability to characterize lethal mutations, and the difficulty of generating populations that are large enough to reach complete saturation of the genome (saturation mutagenesis)[109]. The identification of mutations in a specific gene requires the screening of large numbers of mutagenized lines, necessitating the construction and assaying of large numbers of plant pools. This requirement has prompted the development of more sophisticated pooling strategies that minimize the number of assays required, but still allow the identification of an individual mutant line in one-step or two-step screens[30,110].

Systematic reverse genetics

A major drawback of reverse genetic methods is that the screen has to be repeated for every gene. To overcome this problem, alternative and more effective mutation detection approaches have been developed, and in many cases they have supplanted the individual gene-by-gene methods. High-throughput identification of genome insertion sites in T-DNA-mutagenized and transposon-mutagenized plant populations can be achieved by taking advantage of the known sequence of the inserted DNA and the availability of a complete genome sequence. Various PCR-based strategies have been adapted for this purpose (including thermal asymmetrical interlaced PCR (TAIL PCR)[48] and the adaptor ligation approach[30]). So far, more than 360,000 insertion sites have been mapped in the *A. thaliana* genome, covering ~90% of the genes (see the SIGnAL — Salk Institute Genomic

Saturation mutagenesis
Mutagenesis as a result of which there should be a 100% probability of identifying at least one mutation in any given gene.

Thermal asymmetrical interlaced PCR
A PCR-based strategy that utilizes nested primers, which are complementary to a known DNA sequence, and degenerate primers to amplify the unknown flanking DNA regions.

Adaptor ligation PCR
A PCR-based approach that is used to determine the sequence flanking a DNA region of known sequence. It utilizes two sets of nested primers and a DNA adaptor that, after its ligation to fragmented DNA, facilitates the amplification of the desired DNA region.

Analysis Laboratory web site). However, owing to the near-random nature of insertional mutagenesis and the selection of the lines to be analysed, these approaches are no longer cost effective and more gene-directed strategies are now needed to identify mutations in the remaining 10% of 'untagged' genes.

One solution is to make further use of the existing T-DNA-tagged and transposon-tagged collections of mutants that contain hundreds of thousands of uncharacterized (unsequenced) insertion sites (a typical T-DNA line contains on average 1.5 'expressed' insertions with nearly 50% of the lines containing 2 or more T-DNAs). By using gene-specific and T-DNA/transposon-specific primers, a directed gene-by-gene approach (as described above) could identify mutations in the missing genes. However, the most useful catalogue of mutations in all genes in a plant genome will probably depend on use of a range of mutagenesis approaches. Zinc-finger nucleases, TILLING or Deleteagene strategies can be used to obtain permanent, heritable mutations, and gene-silencing approaches can be used to knock down the activity levels of selected genes. RNAi constructs that target every *Arabidopsis* gene are being developed[111] and could be used to disrupt gene expression in any genetic background, including the ever-growing collection of natural accessions[112]. This tool will not only facilitate the analysis of genes that have not been disrupted by insertional mutagenesis methods (including essential genes that can not be properly studied in the homozygous knockout T-DNA/transposon lines owing to lethality) but will also generate allelic series that represent different levels of silencing.

Forward genetics with reverse genetic tools

One of the most exciting uses of the near complete collection of gene-indexed *A. thaliana* mutations is the ability to carry out whole-genome forward genetic screens[113]. Unlike classical forward genetic approaches, phenotypic screening of gene-indexed collections can directly identify all the genes with knockout phenotypes that are related to the process under investigation. Therefore, dozens of genes that are involved in the process of interest can be rapidly identified[16,106–109].

Two technical limitations currently stand in the way of implementing genome-wide systematic screening in *A. thaliana*. First, one (or ideally two) homozygous mutations per gene need to be selected from the more than 370,000 hemizygous sequenced mutants available. Two are preferable because of the possibility that there might be multiple mutations per plant, owing not only to additional insertions but also to frequent deletions and chromosomal rearrangements[114], which prevent direct gene–phenotype association without obtaining further genetic confirmation. A set of homozygous mutants that cover 25,000 *Arabidopsis* genes is currently being developed (see the Arabidopsis 2010 web page). Second, systematic screening of tens of thousands of individual mutant seed lines under many experimental conditions will require the development of seed-handling instrumentation that is not currently available (FIG. 3).

Equally important to overcoming these two limitations is the development of alternatives to stable insertion and collections of point or deletion mutants in fixed genetic backgrounds. Large collections of constructs for use in RNAi and VIGS (virus-induced gene silencing)-based techniques are also being developed[111]. When applied in reverse genetic screens, as has been done in *D. melanogaster* and *Caenorhabditis elegans*, and more recently in human cells, these approaches have proved effective in global gene-function studies[113]. For implementation in plants, several independent RNAi lines per gene need to be obtained and characterized before systematic gene analysis is carried out. For these types of global study, VIGS (a transient assay) is an attractive and rapid alternative, because the generation of stable transformants can be avoided[115]; in principle, it can therefore be applied to non-model organisms in which transformation is difficult or time-consuming[116]. However, current VIGS systems can produce undesirable phenotypic effects that are an unavoidable consequence of the viral infection itself. Nonetheless, continued improvement of this technology, along with the development of high-throughput infection procedures, should allow this approach to be adapted to genome-wide phenotypic screens in various plant species.

The functional characterization of all genes in a genome represents a formidable task that will require the use of multiple methodologies in parallel. Typical functional genomic approaches such as the use of microarrays provide the desired high throughput, but only limited functional information, whereas classical genetic approaches are not suitable for genome-wide studies. Recent advances in sequence-based mutant detection have allowed for the generation of large gene-indexed mutant collections and the concomitant popularization of reverse genetic approaches. Recombination-based cloning is facilitating the creation of large collections of vectors that are suitable for RNAi, VIGS and the use of amiRNAs. In combination, these tools are allowing the functional characterization of individual genes of interest in many laboratories.

Several important advances towards gene-function analysis in *A. thaliana* are on the horizon: the ability to do systematic forward genetics using reverse genetic tools (simultaneous phenotypic analysis of all gene-indexed mutants), the development of new phenomic platforms, improvements in targeted mutagenesis (specifically, homologous recombination), and the utilization of natural variation in gene-function studies. Mutants or vectors that cover all of the genes in the genome might become available in 2 or 3 years. It is more difficult to predict whether the advances required for the systematic screening of these mutants will be generated and/or how widely accessible they will become. New platforms to examine not only morphological but also molecular phenotypes are being developed[117,118]. The goal of these new phenomic approaches is to enable systematic characterization of gene-indexed mutants, yielding detailed information about gene function. On the other hand, although

Accession
A sample of a plant variety that is collected at a specific location and time. The terms ecotype, wild type and accession are not uniformly used in the *Arabidopsis* field and often cause confusion. The term accession is probably the most appropriate way to describe the *Arabidopsis* laboratory lines that are collected initially from the wild.

Virus-induced gene silencing
Gene silencing that is triggered by a viral vector that encodes a dsRNA with high sequence identity to an endogenous gene.