Running head: High efficiency TILLING population

Luca Comai
Plant Biology and Genome Center
UC Davis
Davis, CA 95616
USA
Tel: 530-752-8485
Email: lcomai@ucdavis.edu

Research area: Breakthrough technology
Production of a high efficiency TILLING population through polyploidization

Helen Tsai\textsuperscript{1}, Victor Missirian\textsuperscript{1,2}, Kathie J. Ngo\textsuperscript{1}, Robert K. Tran\textsuperscript{1}, Simon R. Chan\textsuperscript{3}, Sundaresan Vankatesan\textsuperscript{3}, and Luca Comai\textsuperscript{1,*}

\textsuperscript{1}Dept. of Plant Biology and Genome Center, \\
\textsuperscript{2}Dept. of Computer Science \\
\textsuperscript{3}Dept. of Plant Biology \\
University of California, Davis, CA 95616, USA

One-sentence summary:
Conversion of diploid Arabidopsis to autotetraploidy enables denser mutagenesis resulting in a highly efficient population for reverse genetics
Footnotes

This work was supported by NSF Plant Genome award DBI-0822383, TRPGR: Efficient identification of induced mutations in crop species by ultra-high-throughput DNA sequencing.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Luca Comai (lcomai@ucdavis.edu).

Some figures in this article are displayed in color online but in black and white in the print edition.

Supplemental material is available on line at [TO BE PROVIDED]. The sequence reads from this study are deposited in the NCBI SRA as SRS387455, the CAMBA2 and other software is available at http://comailab.genomecenter.ucdavis.edu/index.php/Data_methods.

Current addresses:

Victor Missirian: Dept. of Plant Biology, UC Davis, Davis, CA 95616

Robert K. Tran: San Mateo County Public Health Laboratory, San Mateo, CA 94403

Simon R. Chan passed away August 23, 2012
Abstract

Targeting Induced Local Lesions in Genomes (TILLING) provides a non-transgenic method for reverse genetics that is widely applicable, even in species where other functional resources are missing or expensive to build. The efficiency of TILLING, however, is greatly facilitated by high mutation density. Species vary in the number of mutations induced by comparable mutagenic treatments, suggesting that genetic background may affect the response. Allopolyploid species have often yielded higher mutation density than diploids. To examine the effect of ploidy, we autotetraploidized the Arabidopsis thaliana Col-0, whose diploid has been TILLED extensively, and mutagenized it with 50 mM EMS. While the same treatment sterilized diploid Col-0, the tetraploid M1s produced good seed. To determine the mutation density, we searched 528 individuals for induced mutations in 15 genes for which few or no knock-out alleles were previously available. We constructed tridimensional pools from the genomic DNA of M2 plants, amplified target DNA and subjected them to Illumina sequencing. The results were analyzed with an improved version of the mutation detection software CAMBa that accepts any pooling scheme. This small population provided a rich resource with ~28 mutations per TILLED 1.5kb fragment, including in average 4 severe missense and 1.3 truncation. The overall mutation density of 19.4/Mb is four times that achieved in the corresponding diploid accession, indicating that genomic redundancy engenders tolerance to high mutation density. Polyploidization of diploids will allow production of small populations, such as less than 2000, that provide allelic series from knock-out to mild loss of function for virtually all genes.
Growing availability of whole genome sequence is spurring functional gene studies in species where specific tools for reverse genetic are not available. However, developing suitable functional genetic resources is often challenging and expensive. For example, targeted gene inactivation through T-DNA (Alonso et al., 2003) or transposable elements (Krishnan et al., 2009) typically requires 300,000 to 500,000 tagged individuals to approach saturation, even for a small genome (Arabidopsis), and both tools may be needed to avoid intrinsic insertional bias (Pan et al., 2005). Two other tools, RNA interference and targeted endonucleases, require transformation and cannot be scaled easily. Furthermore, transgenic plants require regulatory clearance hindering phenotypic characterization. Targeting Induced Local Lesions in Genomes (TILLING) is a functional genomics method that discovers chemically-induced mutations in populations and presents considerable advantages: it is applicable to many sexual species, it requires relatively small populations, it is not transgenic and can target potentially any gene (Comai and Henikoff, 2006; Wang et al., 2012). TILLING consists of mutagenesis, DNA isolation and pooling, and high-throughput mutation discovery in targeted genes. First described in Arabidopsis (McCallum et al., 2000) and Drosophila (Bentley et al., 2000), it has been successfully extended to multiple model and economic species, thus becoming an important tool for functional genomics. Originally, TILLING discovered mutations through the detection of mismatched sites in PCR products (Oleykowski et al., 1998; Till et al., 2004a; Dong et al., 2009). The advent of low-cost high-throughput sequencing has added another powerful method (Rigola et al., 2009; Tsai et al., 2011). Tsai et al. recently described the use of Illumina sequencing and SNP analysis in multidimensional pools as a method for efficient mutation discovery.

TILLING efficiency, i.e. the cost of obtaining informative mutations per gene, depends on the characteristics of the population used and particularly on the mutation density (Comai and Henikoff, 2006). This, in turn, depends at least in part on the intensity used to mutagenize the target species and on its response. Mutagenesis is usually applied by treatment of seed with a chemical mutagen in a manner that produces tolerable lethality and sterility of the treated individuals, the M1s, while allowing sufficient production of fertile M2. M2 DNA and M3 seed are typically inventoried to preserve the resource. Optimal conditions for mutagenesis vary from species to species and mutation densities determined by TILLING differ as much as 100-fold. Since similar treatments yield different mutation densities in different species, cellular or
developmental characters depending from the genetic background must play a role in the outcome. The highest mutation densities were obtained in polyploids (Fig. 1), hexaploid wheat (Slade et al., 2005; Uauy et al., 2009) and tetraploid Brassica napus (Wang et al., 2008; Harloff et al., 2012) displaying the highest number at ~40 mutations per Mb of diploid DNA (i.e. for each of the ancestral genomes of the polyploid). Allotetraploid peanut, on the other hand, is an exception with about 1 mutation/Mb (Knoll et al., 2011). Paleopolyploids have yielded diverse mutation rates: Brassica rapa yielded a near polyploid-like density at 16.6/MB (Stephenson et al., 2010), while soybean and maize displayed lower mutation rates (Till et al., 2004b; Cooper et al., 2008) and within the range of diploids responses. Arabidopsis thaliana acc. Ler yielded about 11 mutations/mb (Martin et al., 2009), one of the highest density described for a diploid. Comparable mutagenic treatments on A. thaliana Col, however, yielded about 4/Mb (Greene et al., 2003).

There is no information on molecular mechanisms underlying variable mutation yields in plants. Such knowledge should be useful for improving TILLING populations. Available data suggest a possible role for polyploidy, but not genome size, in conferring tolerance to high mutation density, a finding consistent with early analysis (Stadler, 1929). There are at least two possible explanations for such an effect. First, genetic redundancy may shield polyploids from the deleterious consequences of mutation making them more tolerant to mutagenic agents. Second, polyploids may be physiologically more tolerant of genotoxic treatments or more susceptible to mutagenesis. Such changes, for example, could result from adaptive changes affecting DNA repair and genome maintenance taking place after polyploidization.

Here we show that an autotetraploid derivative of A. thaliana acc. Col-0 resists an EMS concentration that sterilizes its diploid progenitor while at the same time accumulating mutations at a density four times that previously achieved with this ecotype. We discuss the implications of this finding in the context of whole genome duplication and functional genomics.

Results

Mutagenesis
To investigate the effect of ploidy on the response to EMS mutagenesis, we treated wild-type Col-0 (2X = 2N = 10) and its autotetraploid derivative Col-4X (4X = 2N = 20) with 30 and 50 mM EMS using our standard mutagenesis protocol (Fig. 2). The M1 seed (i.e. treated with EMS) was germinated and grown. During seed development in the M1 plants, immature siliques (seed pods) were dissected and the number of healthy vs. dead or dying seed (presumed to harbor embryo lethal mutations) was scored. Treatment of the diploid line with 30 mM EMS resulted in high frequency (75%) of individuals segregating embryo lethal mutations, i.e. displaying siliques with seed counts consistent with a 3:1 ratio of live to dead seed. By contrast, we observed no tetraploid siliques consistent with a 3:1 ratio. Instead, 1-4 dead embryos occurred in 35% of M1s (see Methods). At 50 mM most diploid siliques were empty or had few shrunken dead seed. At 50 mM, 1-4 dead seed could be observed in 40% of the tetraploid M1s. Control untreated plants of either ploidy displayed negligible counts of embryo lethals. M2 seed from the 50 mM treatment germinated with good efficiency and produced plants whose phenotype was moderately variable in size and growth habit. Most produced seed. In contrast, diploid M2s resulting from treatment with 30 mM EMS displayed 20-40% sterility. In conclusion, autotetraploid Col-0 demonstrated higher tolerance to EMS treatment than its isogenic diploid.

Tri-dimensional pooling, sequencing and identification of mutations

To identify the mutation rate resulting from 50 mM EMS treatment we grew a single M2 progeny from each of about 600 M1 families. Genomic DNA and seed was obtained from 528 individuals. Twenty-four overlapping DNA pools were constructed from 528 M2 individuals by mixing equal amounts of 33 individual genomic DNA (see Methods). Autotetraploidy requires decreasing the pooling ratio because each individual has four alleles. In a pool made from the DNA of 33 tetraploids, a mutant allele in an individual with a simplex genotype (AAAa) is diluted by a factor of 1/132. The same dilution factor is achieved in a pool of 66 diploids. To test the suitability of such a population for mutation discovery, we chose genes that for which knockouts were not available at the time when this study initiated (2007) in the Signal T-DNA insertion collection and searched mutations in an optimally positioned amplicon ~1.5 kb in size. From each pool, genomic DNA fragments were amplified for each of 15 target genes, processed for Illumina sequencing (see Methods), marked with 48 different barcoded adapters and sequenced in two Illumina GAIIx lanes using 100bp paired end reads. The sequence was processed to divide the reads according to barcodes and filter them for quality (see Methods).
The reads were then aligned to the reference, i.e. the expected sequence for the TILLED amplicons. The alignment was processed with CAMBa2, a program that compares the frequency of changes at each nucleotide position across the 24 libraries, identifying putative mutations. The analysis pipeline plots the frequency of changes versus their position on the amplicon (Fig. 3). In positions where no mutations are found all 24 pools display background noise. In positions where candidate mutations are present 3 pools are outliers. CAMBa2 compares these outliers pattern to call mutations.

The 3D pooling scheme employed here differed from that previously described. The original CAMBa was written for one bidimensional and one tridimensional pooling scheme (Tsai et al., 2011), both involving 768 individuals (Missirian et al., 2011). To fit the present study and to prepare CAMBa for any possible population size and dimensional pooling, parts of the program were extensively rewritten. Gridded and non-gridded pooling schemes are now supported to an arbitrary number of dimensions, multiple individuals can be placed in the same combination of libraries, and the number of individuals can vary between different library combinations. One important restriction on pooling scheme designs is that all individuals must be placed in the same number of sequencing libraries. Using this improved version of the program, we identified 413 candidate mutations.

**Statistical analysis**

For each candidate mutation CAMBa2 calculates the probability of the null hypothesis, i.e. that the mutation is false. This probability is expressed as a logarithmic conversion, F(t), for which higher values correspond to higher confidence in a specific mutation. The distribution of F(t) scores (Fig. 4-A) forms a prominent peak centered around F(t)=10 and displaying a long left tail. The F(t) score was essentially independent from the depth of sequence coverage (Fig. 4-B) indicating that CAMBa2 was not biased by sequence read depth. The False Discovery Rate can be derived from F(t), but, to strengthen our calls, we investigated what score distribution could be expected from noise. In a second analysis, we applied CAMBa2 both on non-overlapping pools (i.e. do not share any individuals) and on correct pools. Mutation calls derived from non-overlapping pool sets are false (implausible) unless the same mutation occurred independently in different individuals. We then compared the distribution of F(t) scores for plausible and implausible calls (Fig. 4-C). The two distributions are clearly distinct. Their overlap suggests the
optimal placement of a threshold separating high confidence calls from rejected ones. Of the 413 mutations, 384 had a score of 2 or higher, which is connected with an empirically determined false discovery rate (FDR) rate of 0.02 (using the number of implausible mutations with $F(t) \geq 2$, see Fig. 4-C) and a statistical FDR of 0.005 (Missirian et al., 2011; Tsai et al., 2011). A fraction of the 29 mutations below the threshold is also expected to be true, but are associated with higher FDR. Considering the 384 high-probability mutations (Table 1), our screen of 528 individuals in fifteen 1.5kb sized-amplicons (one per gene) discovered on average a total of 25 mutations/gene. All changes were GC to AT transitions and their effect ranged from predicted knock-outs (19 in 10 targets, 1.3 per target) to missense changes of predicted varying severity (Kumar et al., 2009) (four severe per target) to silent or intronic changes. Four of these mutations were tested independently by PCR amplification and Sanger sequencing, and confirmed (results not shown). In conclusion, TILLING of the 528 lines yielded extensive allelic series for each target gene. The 384 mutations in an adjusted query space of 9.9 Mb of tetraploid DNA (19 kb of TILLED DNA x 528 individuals), is equal to 1 mutation/25.8 kb or 38.8/Mb of the tested tetraploid DNA. This correspond to a mutation density of 19.4/Mb of diploid DNA.

**Discussion**

528 tetraploid individuals provide functional discovery for a majority of genes

*A. thaliana* accession Col-0 has been employed in its natural diploid state for the development of an extensive and well characterized population from which 9,600 mutations have been TILLed. The mutation rate in the diploid Col-0 population was calculated at 6/Mb (Greene et al., 2003), but the estimate of 4/Mb was also obtained using a slightly different calculation approach used commonly in most TILLING publications (see above). A study in accession Ler used M2 mutagenized at different concentrations of EMS (20-40 mM) and yielded an average mutation rate of about 10/Mb, one of the highest rate reported for a diploid. In comparison, the high mutation rate of 19.4/Mb achieved here approaches that found in tetraploid wheat and *Brassica napus*. Given the low toxicity and sterility displayed by autotetraploid Col-0, higher EMS concentrations could be used and one would expect that an even higher mutation rate could be achieved. Truncations, which are most likely to result in knock outs, were found for 2/3 of the tested genes and predicted deleterious mutations for all. One knock out, for example, affected centromeric histone 3 (CENH3) gene, for which KO alleles were not present in other functional
resources at the time of this study (Table 2). This *cenh3-1* allele was used to investigate the function of the encoded histone variant demonstrating its role in chromosomal inheritance, genome maintenance as well as utility for haploid induction, artificial apomixis and genetic analysis (Ravi and Chan, 2010; Ravi et al., 2010; Marimuthu et al., 2011; Marimuthu et al., 2011; Ravi et al., 2011; Seymour et al., 2012; Wijnker et al., 2012). The potential yield of null- and hypomorphic mutations from this easily assembled population compares favorably to the one provided by all the tagging populations available worldwide for Arabidopsis. A moderate expansion to 1,500 M2 individuals would ensure virtual saturation of the *A. thaliana* genes (Fig. 1). The TILLING efficiency displayed when using polyploids makes polyploidization an attractive strategy for development of populations in new species.

**Effect of polyploidy**

Our study differs from previously published work on polyploids because we used a newly-made (synthetic) autopolyploid instead of a natural allopolyploid species. We can thus rule out evolved mechanisms that may stabilize polyploid genomes and affect responses to mutagenic treatments. Because the response reported in this study was measured in the immediate generations after polyploidization, it is a direct effect of ploidy. We believe that the best explanation involves the masking of recessive deleterious mutations by redundant copies. Masking occurs in sporophytes because individuals homozygous for a given mutation are rare (see below). Masking also occurs in the diploid gametophytes, buffering selection against gametophytic lethal mutations and boosting fertility. This is potentially an important feature of the autotetraploid approach, as such mutations may not be transmitted through the germline of diploids, and therefore never recovered. Additionally, DNA damage from mutagenesis may cause chromosomal breakage and segmental aneuploidy compromising health of the M1 plant. The chimeric nature of the M1 meristems after seed mutagenesis (Comai and Henikoff, 2006) could result in clonal selection (Fagerström, 1992), but no investigation addressing this possibility is available in the literature. Aneuploidy and chromosomal deletions would complicate transmission of the mutagenized genome because unbalanced parental gametes can result in failure of the endosperm and seed death. Polyploidy would lessen the impact of this hypothetical occurrence by reducing imbalance and providing intact copies of each chromosome (Henry et al., 2010).
Analysis of mutations discovered in autotetraploids

Autotetraploid inheritance has important consequences on the analysis of mutant alleles. In the M1 generation of an autotetraploid any induced mutation (represented as A -> a) will affect one of four alleles resulting in the simplex genotype AAAa. Inheritance in an autopolyploid is affected by linkage to centromere and the number of homologous chromosomes in a synaptic group (disomic vs multisomic pairing). When the mutant locus is not involved in crossing over or pairing is disomic, a situation defined as chromosome segregation (Burnham, 1962), an AAAa plant can only form two types of gametes each with equal probability: AA and Aa. If meiotic pairing involves multivalents and crossing overs can occur between the A locus and the centromere, a scenario defined as maximum equational segregation can be applied (Burnham, 1962). Following recombination events in which the "a" allele is moved to a different chromosome, both mutant alleles can move to the same pole and at anaphase II can enter the same gamete resulting in an "aa" genotype, a process called double reduction (p = 1/24). Thus, individuals with AAAa genotype can produce "aaaa" recessive zygotes, albeit at a frequency (1/576) much lower than the 3:1 at which diploid "Aa" can form an "aa" zygote. Therefore, the "a" mutant allele will be present in ~70% of the M2s and will be arranged in simplex (AAAa), duplex (AAaa) and, sometime, triplex and quadruplex genotypes according to the ratios presented in Table 3. The two most abundant classes are simplex and duplex, present at approximately 2:1 ratio. TILLING populations are usually stored as M3 seed families derived by the selfing of a single M2. Since most M2 are simplex or duplex, the ratios expected can be derived from Table 3. In conclusion, a minimum of 70% of the M3 in a family should carry at least one allele and production of a homozygous recessive individual is laborious as it requires either double reduction associated with lack of centromeric linkage, or duplex parental genotype. For the reasons above, analysis of a mutation is best carried out through conversion to diploidy.

Conversion to diploidy can be achieved using the triploid bridge (two crosses, Fig. 5-A) or using a haploid inducer (one cross, Fig. 5-B). Using the triploid bridge, an AAAa tetraploid mutant is crossed to the diploid wild type producing a triploid progeny, ~50% of which carries the mutant allele. Some times, a strong interploidy mating barrier exists and may complicate the use of polyploids for TILLING. In certain cases, these barriers are accession specific. In the Col-0 accession, for example, a lethal male excess barrier (Dilkes et al., 2008) can be avoided by using
the tetraploid Col-0 as a female or by employing a tetraploid male from a different accession. Once a triploid carrying the mutation has been obtained, it is crossed to the diploid wild type. In *A. thaliana*, this type of cross produces about a third diploids. The frequency of Aa heterozygotes in the diploid progeny will be 2/3 for Aaa parent and 1/3 for AAa parent. The presence of a gametophytic lethal mutation can be tested at the same time by observing transmission through diploid (or disomic) gametes, but failure in haploid ones. This strategy was used to obtain diploid plants carrying a null mutation generated by autotetraploid TILLING for *CENH3*, a gene for which insertional alleles were unavailable (Ravi and Chan, 2010). Manipulation of this mutant enabled production of a haploid inducer. Its use to diploidize a tetraploid involves crossing the tetraploid mutant to a line expressing a modified centromeric histone 3. The modified *CENH3* causes genome elimination during hybridization to an individual with normal centromeric histone (Fig. 5-B). About 1/4 of the progeny will lose the genome marked by the modified histone resulting in diploids, half of which will be Aa. Selfing of the Aa heterozygote will produce a segregating progeny where the effect of the mutation can be determined by comparing homozygous mutants to individuals carrying wild-type alleles. Haploid induction through natural crosses is available in some species and may be extended soon to others via engineering of *CENH3*.

Given the high density of mutations, will linkage to deleterious mutations complicate the use of these resources? The experimental design described above uses two outcrosses to transfer a mutation into a diploid genome followed by selfing to generate a homozygote required to investigate the phenotype of a recessive mutation. We consider only mutations in the exome (including splice sites) as those in intergenic and intronic regions are very unlikely to be deleterious. With a density of 19.4 mutations/Mb calculated for a diploid genome equivalent (see above), and assuming that most of the mutation are in a heterozygous state, an outcross will transfer about half of the mutations, or 19.4/2 x 40 Mb exome in the *A. thaliana* genome yielding 388 mutations per gamete. Assuming that 15% of these are highly to moderately deleterious (Greene et al., 2003; Comai and Henikoff, 2006) results in 58 mutations or 0.11 mutations/cM in the total 520 cM of the genetic map (Hauge et al., 1993; 2000). Modeling the occurrence of mutations according to the Poisson distribution using p=0.11, the probability of one or more deleterious mutations occurring within 1, 2, 4, and 8 cM of a mutation of interest is, respectively, 0.11, 0.2, 0.36, and 0.59. Only 2400 Arabidopsis genes yield a detectable phenotype when mutated, which is a fraction of all that were tested (Lloyd and Meinke, 2012). Therefore, the
probability of a phenotype-perturbing, tightly linked mutation is manageably small, but still such to require an appropriate genetic design in the analysis. A conclusion on a given gene will be both facilitated and greatly strengthened by the availability of two independent alleles, a common sense requirement. These could be characterized independently, or could be combined in a trans-heterozygous arrangement to cancel the effect of most other genetic lesions.

**Conclusion**

The use of polyploid lines would be effective for TILLING in species recalcitrant to mutagenesis and to increase mutation density in any species. Populations with high density of mutations lower the cost of discovery and are ideally suited for genomic approaches that, instead of TILLING a gene at time, sequence the whole genome or exome to collect all significant mutations. In conclusion, this work indicates that the genomic redundancy of polyploids confers tolerance to the alkylating mutagen EMS, demonstrating the production and exploitation of populations with mutation rates considerably higher than those possible in the isogenic diploids.
Materials and methods

Mutagenesis and plant growth

The autotetraploid Col-0 (Col-0-4x) line was described previously (Dilkes et al., 2008). A seed batch produced from a single plant that was either 2 or 3 generations from the primary colchicine-induced tetraploid were washed in water for 1 hour, then transferred to a solution of EMS in distilled water, and shaken gently in glass drum vials (~3cm diameter) for 17 hours at about room temperature (~23°C). Seeds were rinsed 5 times in distilled water and sown on potting mix. They were grown to maturity in growth chamber at 21°C and 16 hours daylight. Seed was harvested separately from each M1 individual. A single M2 plant was grown for DNA isolation and M3 banking. Hypothetical colchicine-induced point mutations (AAAa in the original tetraploid) would most likely be shared by many of the pooled individuals and would have violated a requirement imposed by CAMBa2: for any mutation to be considered it should be carried by a single individual in the population or, exceptionally, by two. Therefore, these mutations would have been ignored in the analysis.

Phenotyping

Siliques of M1 plants were dissected about 2 weeks after pollination and the seed health was evaluated by color and shape. Dead seed were dark or transparent instead of displaying a well shaped green embryo at the expected stage of development. The occurrence of embryo lethal mutations in the diploid was estimated by counting 10 siliques in 10 individuals and scoring as positive any silique that displayed 6-12 dead seed, which in the average silique of 40 seeds approximates a 1:3 ratio of dead to live seed, thus fulfilling the expected behavior for a recessive mutation present in heterozygous state in the germ cells of the flower. The same expectation could not be applied to tetraploid plants because under the assumption of a simple dominant-recessive pair of alleles, a simplex AAAa flower would yield a maximum 1:575 dead:live seed ratio (Table 3). Therefore, in the tetraploid M1 siliques embryo lethal mutations could not be estimated. Other factors, such as haploinsufficiency and chromosomal aberrations, may contribute to death.

DNA Isolation

Genomic DNA was isolated as described (Tsai et al., 2011). DNA was quantified using SYBR
green1 dye fluorescence, and standardized.

**DNA Pooling Strategy**

Our method is based on overlapping genomic DNA pools (Tsai et al., 2011). We wanted the ability to sequence all pools in a minimal number of Illumina lanes (unit of sequencing). Each pool was prepared by combining equal amounts of genomic DNAs from 33 individuals, providing a mix of 132 alleles (= 33 tetraploid individuals x 4 alleles). Different pools shared a given set of individuals in such a way that each individual should be uniquely identified by three intersecting pools. To fit the 528 individuals of this library in 24 pools we chose a mixed pooling scheme in which 480 individuals were uniquely identified and 48 were not, but could be traced to a trio of individuals and subsequently deconvolved. This scheme was an acceptable compromise, but for future work we designed a custom python program (pooling_capacity.py), which takes a range of pool numbers (number of available barcodes), maximum pooling depth (individuals/pool), and the total number of individuals in order to elaborate possible pooling solutions. The program outputs a table of individuals and pools that can be used by robotic pipettors to prepare the pool from plates containing individual DNAs. A Biomek 2000 robotic pipettor was used to prepare the 24 template pools. TILLING targets amplified by PCR from the same pool were processed to make an Illumina library.

**PCR amplification of targets and library preparation**

We followed the methods described by Tsai et al. (Tsai et al., 2011). We used 5-base barcoded adapters so that 24 libraries with different barcodes could be pooled into one 100-base paired-read sequencing lane. The name and sequence of the adapters were adA2_nnnn: P-nnnnnAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAG and adB2_nnnnn: ACACTCTTTCCCTACACGACGCTCTTCCGATCTuuumuT, where nnnnn is the 5 nucleotide barcode, u is the complementary base, and P represents 5’ phosphorylation. The expected coverage per individual was calculated as follows: lane yield/(library ratio x input DNA x pooling ratio).

**Bioinformatics**

A Python program was used to convert Illumina sequence reads from Illumina quality to Sanger quality, group sequence reads by barcodes, remove adapter sequences from sequencing reads,
filter sequencing reads by a base quality cutoff (Phred 20), and filter sequencing reads by a minimum length (25bp). We used Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009) in conjunction with SAMtools (Li et al., 2009) to align the sequencing reads to the amplicon sequences. Another Python program was used to extract alignment information from the pileup table (output file format from SAMtools) and generate a human readable output, parsed pileup table.

We used an updated version of CAMBa for mutation detection, CAMBa2. The inputs to CAMBa2 are the parsed pileup table, a pooling scheme file, and the reference sequences (amplicon sequence, genomic sequence trimmed to begin at the start codon, and coding sequence) for each TILLed fragment. CAMBa2 processes each position along a TILLed amplicon fragment separately, considering each possible configuration (assignment of mutations to individuals) that satisfies the assumption of at most one mutation per individual. It then computes the probability of obtaining the observed base calls under each candidate configuration, assuming a binomial model of sequencing error. CAMBa2 incorporates the prior probability of each configuration in order to compute the posterior probabilities of those configurations, using Bayes' Theorem. We used CAMBa2 to assign a probability score and identify the mutant individual or mutation-containing subpool, and finally produced a list of candidates and predicted effects. Plausibility analysis (the search of mutations in pools that do not overlap) was run separately by rerunning the CAMBa2 program while allowing any library to be considered overlapping with any other one. The effect of the mutation was predicted by the CAMBa2 software. The severity of amino acid substitutions was predicted using SIFT (Kumar et al., 2009).

Mutation density was calculated by dividing the total number of high-confidence mutations by the total sampled DNA sequence, followed by further division by 2 to account for tetraploidy. Sampled DNA sequence was calculated by multiplying the number of sampled individuals by the sum of the well-covered sequence in each TILLed fragment. The two terminal 50 bases of each TILLed fragment were subtracted from the fragment size because the termini correspond to the PCR primers and the rest is underrepresented in the sequencing reads. Sequence reads have been deposited in the NCBI SRA with identification number SRS387455. All software used here is freely available on the Comai laboratory Methods web page (http://comailab.genomecenter.ucdavis.edu/index.php/Data_methods).
**Figure legends**

**Figure 1.** Effect of mutation rate and ploidy on functional discovery through TILLING. **A.** The probability of getting at least one severe missense mutation (assumed to be 15% of all mutations) and one knock-out (5% of all mutations) is plotted versus the total mutation number identified in the coding region of a gene (redrawn from Henikoff et al., (Henikoff et al., 2004)). **B.** The relationship between mutant yield and mutation rate is illustrated by the number of individuals in a population required to yield a given number of mutations (in A) in a 1kb fragment. For example, considering mutations in a 1kb coding region, the hatched blue stripe highlights how a population with a mutation rate of 2/Mb requires screening more than 15,000 individuals for a 0.8 confidence of obtaining at least a single knock out. The same result can be obtained with 768 individuals of a population that has a mutation rate of 40/Mb such as hexaploid wheat (Ta). The number of mutations expected for a given population and mutation density should be scaled according to the gene size. For example, for a 2kb coding region and a mutation density of 2/Mb 60 mutations are expected. A population of 15,000 individuals would yield ~95% chance of at least one KO. **C.** Published mutation rates in TILLING populations of different species organized according to ploidy. The vertical bar connects instances of the same species. AtC=Arabidopsis thaliana Col-0 (Greene et al., 2003), AtL=A. thaliana Ler (Martin et al., 2009), Hv=Hordeum vulgare (Talame et al., 2008), Os=Oryza sativa (Till et al., 2007), Mt=Medicago truncatula (Le Signor et al., 2009), Ps=Pisum sativum (Dalmais et al., 2008), Zm=Zea mays (Till et al., 2004b), Br=Brassica rapa (Stephenson et al., 2010), Gm=Glycine maxima (Cooper et al., 2008), Ai=Arachis hypogaea (Knoll et al., 2011) Bn=Brassica napus (Harloff et al., 2012), Td=Triticum durum (4x) (Slade et al., 2005; Uauy et al., 2009), Ta=T. aestivum (6x) (Slade et al., 2005; Uauy et al., 2009).
Figure 2. Embryo lethality in mutagenized diploid and autotetraploid Arabidopsis. Two genetically identical strains of Col-0, but differing in ploidy, were treated as seed with the chemical mutagen ethymethane sulfonate (EMS). The effect of the treatment was monitored by observing development of M2 seeds in the siliques of the M1 plants. Mutagenesis-induced changes affect one allele in any target gene: the corresponding genotype is Aa in the diploid and AAAa (simplex) in the tetraploid. M1 individuals are chimerae because the mature embryo shoot meristem is a multicellular structure and each cell is mutagenized independently. Genetic evidence indicates that the same cell lineage forms male and female gametes in a single flower. The occurrence of failed seed in 1:3 (failed:live) ratio in the 30mM EMS-treated diploid can be attributed to homozygous recessive mutants in genes required for embryo development and formed by selfed Aa lineages of the M1 plant. Diploid Arabidopsis was effectively sterilized by 50mM EMS, which only marginally affected seed set in the autotetraploid. In tetraploids, the probability of a homozygous recessive embryo (aaaa) produced from an AAAa flower lineage is minuscule (1:575, Table 3). Therefore, the dead seed progeny of the polyploid M2 is not readily explained by mutations in genes where the wild-type allele has a dominant action. More likely, the dead seed in polyploids are the result of mutations, or perhaps chromosomal aberrations, that have a lethal dosage effect.
Figure 3. Mutations detected in the Ala-tRNA Ligase gene (At1G50200). The change frequencies for the 24 sequenced pools are plotted versus the base position on the TILLed DNA fragment. The triplet pattern formed by three outliers, corresponding to a unique mutant individual shared by three pools, is evident in the G>C and C>T frequency tracks. No triplets, i.e. candidate mutations, are visible in the A>C and A>G tracks. This is consistent with the observation that EMS in Arabidopsis is specific for GC>AT base pair changes (Greene et al., 2003) and with the notion that the triplet pattern is not generated by random sequencing errors. Frequency plots for the remaining base changes were comparable to the latter two in pattern and in range (data not shown). Background noise consists of high quality sequence changes and is characteristic of each change type: for example, higher in A>G than in A>C.
Figure 4. Discovery of mutations through Illumina sequencing of overlapping pools. A. The probability score F(t) is the mean centered log of the posterior probability 't' of the null hypothesis. Higher F(t) scores are desirable, as they correspond to lower probability of false discovery. The figure illustrates the distribution of 413 mutation candidates. The mutations were detected by TILLING 15 amplicons corresponding to 15 A. thaliana genes in a population of 528 autotetraploid individuals. Changes detected by aligning Illumina reads to the TILLED sequence reference were subjected to CAMBa2 analysis to discover mutation candidates. The vertical grey line indicates the threshold of F(t)=2, corresponding to an empirically determined False Discovery Rate (FDR) = 0.02 (see C), and to a statistical FDR of 0.005. B. Probability is independent of sequencing coverage. Each point corresponds to a putative mutation and displays F(t) versus mean coverage. The lack of correlation between F(t) and coverage (RSquare = 0.02) indicates that once sequencing coverage is above a critical threshold discovery is reliable. C. Setting the threshold probability for mutation calling. To shorten the time required by the computationally complex step, half the dataset (264 individuals) used in A was subjected to mutation detection using CAMBa2 allowing all potential pool overlaps, both true and false. Mutations were then divided according to the overlap (i.e. the shared individuals from the population) type and are displayed both as frequency histogram and as dithered dots. Those detected in true overlaps (had common individuals) were defined plausible (i.e. true). Those detected in false overlaps (did not share individuals) were defined implausible (i.e. false). "Double mutations" are consistent with two identical changes occurring in independent individuals and thus found in two sets of three correctly overlapping pools. The threshold at F(t)=2 was arbitrarily chosen to provide a predicted FDR of 2% and corresponds to the F(t) value above most implausible mutations and below most plausible ones. Using that threshold, 384 mutations are called in the whole dataset.
Figure 5. Diploidization strategies for mutations discovered in autotetraploids. Two alternatives are illustrated for a simplex autotetraploid mutant with genotype $CCCc$. A, Diploidization via triploid bridge (Henry et al., 2009). All progeny of cross 1 are expected to be triploid. Triploids transmit either one or two copies of each chromosome type (Table 3). The progeny of cross 2 will fit in three classes: diploids (~30%, (Henry et al., 2009)), aneuploids and some triploids. B, Diploidization via genomic elimination achieved through the use of a *haploid inducer* variety, such as Col-0 *cenh3-1, GFP-tailswap* (Ravi and Chan, 2010). Cross 3 will produce tetraploids, diploids and aneuploids. The direction of the cross influences the percentage of diploids produced. Furthermore, some plant varieties, such as Col-0, carry male-specific interploidy lethality factors (Dilkes et al., 2008). Reciprocal crosses should be tested for optimal results. $c =$ mutant allele, $C =$ wild-type allele, $2X =$ diploid, $4X =$ autotetraploid, $[ ] =$ gametic genotypes, thick lettered genotypes in haploid inducers refer to alleles on genomes targeted for elimination.
### Tables

**Table 1. Summary of induced mutations discovered in a 528 tetraploid Arabidopsis population**

<table>
<thead>
<tr>
<th>Gene</th>
<th>Total mutations</th>
<th>Intronic</th>
<th>Synonymous</th>
<th>Missense</th>
<th>Predicted truncations</th>
<th>Median Coverage</th>
<th>Median F(t) score</th>
<th>Fragment Size (bp)</th>
<th>Queried Fragment Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G01370</td>
<td>22</td>
<td>12</td>
<td>3</td>
<td>6</td>
<td>4</td>
<td>9396</td>
<td>20.1</td>
<td>1507</td>
<td>1357</td>
</tr>
<tr>
<td>AT1G07350</td>
<td>21</td>
<td>13</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>14234</td>
<td>19.7</td>
<td>1136</td>
<td>986</td>
</tr>
<tr>
<td>AT1G09620</td>
<td>26</td>
<td>0</td>
<td>10</td>
<td>16</td>
<td>9</td>
<td>15422</td>
<td>14.7</td>
<td>1490</td>
<td>1340</td>
</tr>
<tr>
<td>AT1G15660</td>
<td>28</td>
<td>5</td>
<td>7</td>
<td>14</td>
<td>0</td>
<td>10989</td>
<td>15.8</td>
<td>1490</td>
<td>1340</td>
</tr>
<tr>
<td>AT1G25580</td>
<td>40</td>
<td>8</td>
<td>8</td>
<td>21</td>
<td>6</td>
<td>12859</td>
<td>15.3</td>
<td>1527</td>
<td>1377</td>
</tr>
<tr>
<td>AT1G50200</td>
<td>34</td>
<td>13</td>
<td>6</td>
<td>14</td>
<td>6</td>
<td>18056</td>
<td>25.6</td>
<td>1503</td>
<td>1353</td>
</tr>
<tr>
<td>AT1G58200</td>
<td>28</td>
<td>11</td>
<td>5</td>
<td>12</td>
<td>5</td>
<td>11326</td>
<td>25.1</td>
<td>1498</td>
<td>1348</td>
</tr>
<tr>
<td>AT1G69170</td>
<td>11</td>
<td>0</td>
<td>4</td>
<td>7</td>
<td>3</td>
<td>12442</td>
<td>15.8</td>
<td>1080</td>
<td>930</td>
</tr>
<tr>
<td>AT1G72550</td>
<td>24</td>
<td>5</td>
<td>10</td>
<td>8</td>
<td>2</td>
<td>16558</td>
<td>22.2</td>
<td>1471</td>
<td>1321</td>
</tr>
<tr>
<td>AT2G35670</td>
<td>17</td>
<td>4</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>11281</td>
<td>16.6</td>
<td>1499</td>
<td>1349</td>
</tr>
<tr>
<td>AT3G10440</td>
<td>15</td>
<td>5</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>12221</td>
<td>14.3</td>
<td>1511</td>
<td>1347</td>
</tr>
<tr>
<td>AT3G54350</td>
<td>37</td>
<td>0</td>
<td>14</td>
<td>23</td>
<td>3</td>
<td>14356</td>
<td>14.8</td>
<td>1497</td>
<td>1344</td>
</tr>
<tr>
<td>AT4G10320</td>
<td>30</td>
<td>5</td>
<td>1</td>
<td>23</td>
<td>7</td>
<td>12569</td>
<td>14.9</td>
<td>1494</td>
<td>1336</td>
</tr>
<tr>
<td>AT4G13780</td>
<td>23</td>
<td>9</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>11167</td>
<td>13.8</td>
<td>1486</td>
<td>1356</td>
</tr>
<tr>
<td>AT5G16630</td>
<td>28</td>
<td>7</td>
<td>4</td>
<td>16</td>
<td>5</td>
<td>14947</td>
<td>15.8</td>
<td>1506</td>
<td>1361</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>384</strong></td>
<td><strong>97</strong></td>
<td><strong>85</strong></td>
<td><strong>183</strong></td>
<td><strong>59</strong></td>
<td><strong>183</strong></td>
<td><strong>10</strong></td>
<td><strong>9</strong></td>
<td><strong>21695</strong></td>
</tr>
</tbody>
</table>

| **Mean**   | **25.1**        | **3.9**  | **1.3**    | **15603** | **22.7**              |
### TABLE 2. Comparison of tetraploid TILLING to available tag resources

<table>
<thead>
<tr>
<th>Locus</th>
<th>Function</th>
<th>Salk T-DNA tags(^1)</th>
<th>All tags(^2)</th>
<th>4X At TILLING</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Exon</td>
<td>Intron</td>
<td>Exon</td>
</tr>
<tr>
<td>At1g01370</td>
<td>CEN-H3 DNA binding</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>At1g07350</td>
<td>transformer</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>At1g09620</td>
<td>Leu-tRNA ligase</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>At1g15660</td>
<td>CENP-C protein</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>At1g25580</td>
<td>ANAC8, transcription factor</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>At1g50200</td>
<td>ALA-tRNA ligase</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>At1g58200</td>
<td>MSL3 ion channel</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>At1g69170</td>
<td>SPL6 transcription factor</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>At1g72550</td>
<td>tRNA synthetase</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>At2g35670</td>
<td>FIS2, Polycomb factor</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>At3g10440</td>
<td>Shugosin-like protein</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>At3g54350</td>
<td>emb1967</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>At4g10320</td>
<td>Ile-tRNA ligase, putative</td>
<td>1</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>At4g13780</td>
<td>Met-tRNA ligase</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>At5g16630</td>
<td>RAD4</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

\(^1\) Flanking sequence tag (FST) reported as of July 2012 in the Signal Insertional Database (signal.salk.edu) as SALK_000000. Tags in the 5' and 3' region of the target genes were not included because their effect is not easily predicted.

\(^2\) All FST reported as of July 2012 in the Signal Insertional Database (CSHL, FLAG, GABI-KAT, RIKEN, SALK, SK, Wis)

\(^3\) Amino acid substitutions with SIFT probability ≤ 0.05 (Kumar et al., 2009)
<table>
<thead>
<tr>
<th>Parental genotype</th>
<th>Stage</th>
<th>Filial genotype</th>
<th>Segregation type¹</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Chromosome²</td>
</tr>
<tr>
<td>AAAa</td>
<td>Gametophyte</td>
<td>AA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aa</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>aa</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Sporophyte</td>
<td>AAAA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>from selfcross</td>
<td>AAAa</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AAaa</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aaaa</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>aaaa</td>
<td>0</td>
</tr>
<tr>
<td>AAaa</td>
<td>Gametophyte</td>
<td>AA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aa</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>aa</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Sporophyte</td>
<td>AAAA</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>from selfcross</td>
<td>AAAa</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>AAaa</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Aaaa</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>aaaa</td>
<td>1</td>
</tr>
</tbody>
</table>

¹ According to Burnham (Burnham, 1962)
² Mutant locus is linked to centromere
³ Mutant locus is unlinked to centromere
For example, if the A locus is tightly linked to the centeromere, a AAAa individual produces two types of gametes: Aa and AA, in a 1:1 ratio. Self-fertilization will produce three genotypes: AAAA, AAAa, and AAaa in a 1:2:1 ratio.
Supplemental Data

The following materials are available in the online version of this article.
Supplemental Table S1. Mutations identified in the autotetraploid Arabidopsis population
Supplemental Table S2. Fasta sequences of TILLED amplicons

Acknowledgements

We thank Jong-A Park for assistance with the EMS mutagenesis and seed collection from the tetraploid plants. We thank Charlie Nicolet, Vanessa Rashbrooks and Heather Witt at the UCD Genome Center DNA Technology Core for their assistance with Illumina GA sequencing, Smit Shah and Wasinee Pongprayoon for technical help in DNA and Illumina library preparation.
Literature Cited


Burnham CR (1962) Discussions in cytogenetics. 375 p


Downloaded from on July 15, 2017 - Published by www.plantphysiol.org Copyright © 2013 American Society of Plant Biologists. All rights reserved.


Stadler IJ (1929) Chromosome number and the mutation rate in Avena and Triticum. Proc Natl Acad Sci USA 15: 876–881


Individuals necessary to find the given mutation num.
**diploid Col-0**

- **colchicine**
  - Leads to **diploid Col-0** and **autotetraploid Col-0**

---

**Seeds**

- EMS 30mM
- 50mM
- 30mM
- 50mM

---

**M1 plants**

- Viável sementes
- Mortas

---

**Seed set**

- 80%
- 0%
- 92%
- 90%

- Live:dead ratio
  - 3:1
  - n.a.
  - 15:1
  - 15:1

---

**M2 seeds**

- Sementes

---

**M2 plants**

- Plantas
The image contains four scatter plots labeled C>T, G>A, G>A, and A>C, each representing the frequency of different SNP transitions at various positions of a TILL Fragment. The x-axis of each plot corresponds to the position of the TILL Fragment, ranging from 0 to 1600 base pair positions, while the y-axis shows the frequency of SNP transitions from 0.00 to 0.08. The plots are color-coded, with data points in red, green, and blue, indicating different categories or data sets.