Spectrum of Chemically Induced Mutations From a Large-Scale Reverse-Genetic Screen in Arabidopsis

Elizabeth A. Greene,* Christine A. Codomo,* Nicholas E. Taylor,* Jorja G. Henikoff,* Bradley J. Till,* Steven H. Reynolds,† Linda C. Enns,† Chris Burtner,† Jessica E. Johnson,† Anthony R. Odden,* Luca Comai† and Steven Henikoff*,‡,[1]

*Fred Hutchinson Cancer Research Center and ‡Howard Hughes Medical Institute, Seattle, Washington 98109 and †Department of Biology, University of Washington, Seattle, Washington 98195

Manuscript received January 18, 2003
Accepted for publication March 26, 2003

ABSTRACT

Chemical mutagenesis has been the workhorse of traditional genetics, but it has not been possible to determine underlying rates or distributions of mutations from phenotypic screens. However, reverse-genetic screens can be used to provide an unbiased ascertainment of mutation statistics. Here we report a comprehensive analysis of ~1900 ethyl methanesulfonate (EMS)-induced mutations in 192 *Arabidopsis thaliana* target genes from a large-scale TILLING reverse-genetic project, about two orders of magnitude larger than previous such efforts. From this large data set, we are able to draw strong inferences about the occurrence and randomness of chemically induced mutations. We provide evidence that we have detected the large majority of mutations in the regions screened and confirm the robustness of the high-throughput TILLING method; therefore, any deviations from randomness can be attributed to selectional or mutational biases. Overall, we detect twice as many heterozygotes as homozygotes, as expected; however, for mutations that are predicted to truncate an encoded protein, we detect a ratio of 3.6:1, indicating selection against homozygous deleterious mutations. As expected for alkylation of guanine by EMS, >99% of mutations are G/C-to-A/T transitions. A nearest-neighbor bias around the mutated base pair suggests that mismatch repair counteracts alkylation damage.

THE ability to induce mutations has been a major driving force in genetics for the past 75 years (Muller 1930). Among the mutagens that have been used to induce mutations, chemical mutagens administered in various ways have become especially popular. Alkylating agents, such as ethyl methanesulfonate (EMS), are particularly effective, because they form adducts with nucleotides, causing them to mispair with their complementary bases, thus introducing base changes after replication (Haughn and Somerville 1987; Ashburner 1990). EMS mutagenesis results in high point mutational densities with only low levels of chromosome breaks that would cause aneuploidy, reduced fertility, and dominant lethality. Therefore, chemical mutagenesis has become the method of choice for genetic studies, remaining popular even with the advent of sophisticated transgenic technologies that allow for tagging or precise targeting of mutational lesions.

Despite geneticists' heavy reliance on chemical mutagenesis, traditional genetic screens do not readily reveal the underlying mutational process. This is because geneticists select for phenotypes, and as a result, only a small minority of mutations within a target gene are examined. This situation is changing. With the availability of large amounts of DNA sequences from model organisms and the incentives to determine the functions of genes discovered from DNA sequence, reverse-genetic approaches are becoming increasingly important. Among these are genome-wide mutagenesis methods followed by screening within individual gene segments, which is made possible by using PCR (Henikoff and Comai 2003). Although PCR-based detection of insertions and deletions is straightforward, detection of point mutations, such as those introduced by chemicals, is challenging, because the amplified fragment does not change in size. Nevertheless, detection of single-base changes has improved rapidly with advances in single-nucleotide polymorphism (SNP) detection technologies (Kwok 2001), and this has fueled the application of new technologies to reverse-genetic mutational screening.

One example of SNP detection technology being applied to reverse genetics is TILLING (*t*argeting *i*nduced *l*ocal *l*esions *in g*enomes), in which chemical mutagenesis is followed by screening for point mutations (McCallum *et al.* 2000). TILLING has been streamlined for high throughput with the use of the CEL 1 endonuclease (Colbert *et al.* 2001), which cleaves at mismatches within heteroduplexes formed between mutant and wild-type strands (Oleykowski *et al.* 1998). This allows for cleaved fragments to be detected on electrophoretic gels, revealing the mutation and its approximate posi-

[1] *Corresponding author:* Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N., Seattle, WA 98109. E-mail: steveh@fhcrc.org

tion in the fragment. Using this technology, we have established a public TILLING facility for the general Arabidopsis community, the Arabidopsis TILLING Project (ATP; Till *et al.* 2003b). ATP uses the CEL 1 mismatch cleavage method to screen pooled DNAs from EMS-mutagenized plants and subsequently from the individual DNAs that constitute each pool in which a mutation was detected. ATP then determines the sequence of the mutation using its approximate location determined from the sizes of CEL 1-cleaved fragments as a guide. The ~1900 mutations that ATP has delivered to users are an exceptionally rich resource for ascertaining the spectrum of EMS-induced mutations in a way that is not compromised by phenotypic selection.

Here we analyze the spectrum of mutations reported by ATP. We test the assumption that these mutations are generated at random and that detection is robust, and we discover a local compositional bias. Our findings have practical implications for the application of EMS mutagenesis to reverse genetics and also provide insights into chemical mutagen damage and repair in the germplasm.

MATERIALS AND METHODS

High-throughput TILLING detection and sequencing of mutations: Mutagenesis, growth, screening, and sequencing procedures are described elsewhere (Till *et al.* 2003a,b). Briefly, seeds were mutagenized by soaking 10–20 hr in 20–40 mm EMS and sown. The resulting M_1 plants were self-fertilized, and M_2 individuals were used to prepare DNA samples for mutational screening and seeds for distribution. DNA samples were pooled and arrayed in microtiter plates, and the pools were amplified using gene-specific primers. Amplification products were heated and cooled to form heteroduplexes and incubated with the CEL 1 endonuclease, which cleaves precisely 3′ to mismatches. Cleavage products were electrophoresed using LI-COR IR^2 gel analyzers, and gel images were analyzed using Photoshop (Adobe Systems). The use of two different dye labels on primers allowed mutations to be detected on complementary strands, facilitating confirmation. For each mutation detected in a pool, the eight individual DNA samples were similarly screened to identify the plant carrying the mutation. For DNA sequencing, individual genomic DNA samples were amplified with the gene-specific primers in 96-well plates and subjected to terminator sequencing using the ABI (Applied Biosystems, Foster City, CA) Big-Dye system by capillary sequencing at the Fred Hutchinson genomics facility.

Sequence data processing: Sequencing traces were processed using Sequencher (Gene Codes, Ann Arbor, MI). Traces were aligned and compared with the reference sequence. Sequencher was first asked to report anomalous peak heights, which typically identified all homozygotes and many heterozygotes. We then confirmed mutations by comparing the Sequencher anomalies to the table of CEL 1 mobilities, which indicates the approximate positions of the mutations. Finally, we examined the remaining traces using the table of CEL 1 mobilities as a guide. Homozygous changes were assigned if replacements of single chromatogram peaks relative to the reference were observed. Heterozygous changes were assigned if a mixed peak in which one was the reference was observed and the height of the reference peak was reduced relative to neighboring peaks.

Data analysis and interpretation: Users gain access to ATP via a "welcome" page that explains TILLING (http://tilling.fhcrc.org:9366/Welcome_to_ATP.html) and describes situations in which an allelic series is useful for determining gene function. Users then proceed to the interactive CODDLE (*co*dons *o*ptimized to *d*etect *d*eleterious *le*sions) analysis system (http://www.proweb.org/input), which facilitates the acquisition of genomic sequence, gene model, and protein conservation information, identifies regions most likely to yield deleterious lesions, and runs Primer3 (Rozen and Skaletsky 2000) to design primers that are optimal for TILLING. A form is filled out and primers that amplify the chosen ~1-kb gene segment are ordered, while storing the sequence of the segment and the gene model. This information is used to calculate GC content and codon and splicing statistics. Following identification of sequence changes, which can be scored on either strand, a table corresponding the TILLed fragment and its gene model is generated and PARSESNP (project aligned related sequences and evaluate single-nucleotide polymorphisms; http://www.proweb.org/parsesnp) parses the mutation data. Regardless of the strand on which the mutation was scored, mutations are reported on the coding strand only. All data are accessible from The Arabidopsis Information Resource (http://arabidopsis.org) and the ATP website (http://tilling.fhcrc.org:9366).

RESULTS

EMS-induced mutations are randomly distributed: Our data set derives from accumulation of mutation data generated by the Arabidopsis TILLING Project over its first 18 months of operation (Till *et al.* 2003b). Operations began with an announcement to the general Arabidopsis community that TILLING would be available as a public service, and potential users were encouraged to request a region of their single favorite gene for reverse-genetic analysis. Interactive web-based tools were made available for users to choose ~1-kb segments within their genes, favoring selection of regions where mutations were predicted to damage the protein, such as conserved missense and protein truncation mutations. Primers were chosen and orders were placed. Using these primers, ATP typically screened 3072 EMS-mutagenized plants pooled eightfold in a 96-well format, using the CEL 1 mismatch-cleavage method. Whenever a positive pool was discovered, the eight individuals were similarly screened using the CEL 1 method to find the mutated individual in the pool. The size of the CEL 1-cleaved fragment approximated the location of the mutation, which was then determined by a single-pass sequence from either end. This led to an allelic series of 1890 mutations from 192 fragments distributed on all euchromatic chromosome arms, which indicates that mutations can be found throughout the genome (Figure 1).

On average, we identified 10 mutations per gene fragment, and for two-thirds of the genes, 8–12 mutations were reported to users (Figure 2). The distribution of genes with truncation mutations, consisting of nonsense

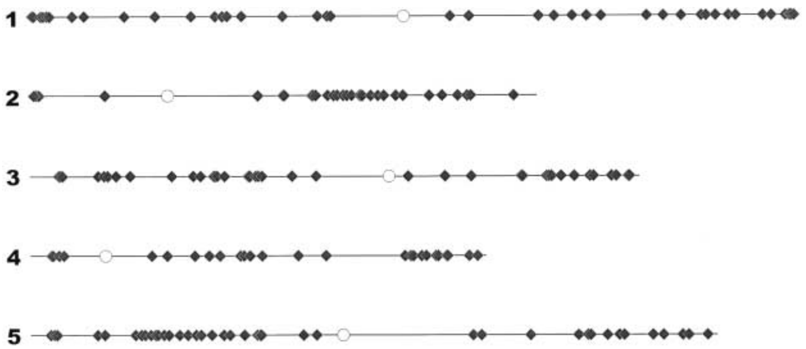


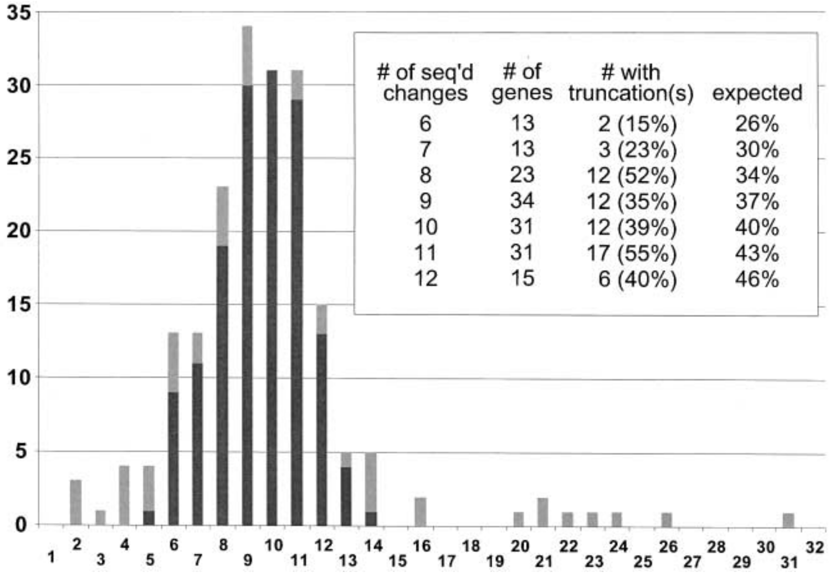
Figure 1.—Distribution of TILLed fragments. The five *Arabidopsis thaliana* chromosomes are shown with centromeres as open circles and 192 TILLed fragments as diamonds.

and splice junction changes, was as expected by chance (Figure 2, inset). For example, in genes for which 10 mutations were reported, 39% of the time at least 1 truncation mutation was discovered, which is expected on the basis of an overall predicted truncation frequency of 5% $[1 - (1 - 0.05)^{10} = 0.40]$. The fact that the expected number of the most severely deleterious mutations was found suggests that the large majority of gene segments chosen can tolerate the full spectrum of EMS-induced mutations.

In total, we reported 1890 mutations in the 192 fragments screened. Taking the average number of individual plant DNAs screened (~3000), we can calculate the overall mutation density as $1890/(192 \times 3000) = 1$ mutation/300 kb screened. There are caveats to this rough estimate, including the possibilities that not all DNA fragments were effectively screened and that mutations were missed. However, by carefully examining the data set, as described below, we can deduce mutation rates in ways that are not subject to these caveats.

EMS mutagenesis delivers >99% G/C-to-A/T transition mutations: EMS alkylates guanine residues, producing O^6-ethylguanine, which pairs with T but not with C (Ashburner 1990). As a result, replication of unrepaired alkylation damage will effectively replace the G/C base pair with an A/T. This mechanism predicts a strong G/C-to-A/T bias in EMS-induced mutations, as observed in numerous mutagenesis studies (Vidal *et al.* 1995). However, the degree to which G/C-to-A/T transitions are favored cannot be accurately determined from forward genetic screens because of selection for phenotypes of interest. In a reverse genetic screen, where selection was probably nonexistent, Bentley *et al.* (2000) found that all 16 mutations in the single Drosophila gene that was targeted were G/C-to-A/T changes. Our much larger data set on 192 genes extends and generalizes this bias. Indeed, 1890 of 1906 (99%) of the changes are G/C-to-A/T, with 53% of changes at G and 47% at C on the coding strand, at the frequencies expected for these TILLed fragments.

We can ask whether or not the 16 non-G/C-to-A/T exceptions are likely to have been induced by EMS. We first note that the spontaneous mutation rate in Arabidopsis has been reported to be between 10^{-7} and 10^{-8} bp/generation (Kovalchuk *et al.* 2000), which is high enough to account for all of our exceptions. However, this estimate is very uncertain, and so we need to consider the possibility that some exceptions are



# of seq'd changes	# of genes	# with truncation(s)	expected
6	13	2 (15%)	26%
7	13	3 (23%)	30%
8	23	12 (52%)	34%
9	34	12 (35%)	37%
10	31	12 (39%)	40%
11	31	17 (55%)	43%
12	15	6 (40%)	46%

Figure 2.—Frequency histogram of mutations reported per TILLed fragment. Results of typical TILLING screens on 3072 plants are shown as solid bars, and results of exceptional screens are shown as shaded bars. Low exceptions include cases in which mutations were discovered but only a subset sequenced. High exceptions include cases in which a typical TILLING screen was performed, but further mutations were obtained by screening additional plants. Inset shows the percentages of fragments with truncation mutations.

TABLE 1

Distribution of missense and truncation mutations in heterozygotes and homozygotes

	All	Silent	Missense	Truncation
n	1890	851	946	93
Distribution				
% expected	100	44.4	48.3	5.3
% observed	100	45	50.1	4.9
Heterozygous	1276	566	637	73
Homozygous	614	285	309	20
Ratio	2.08	1.99	2.06	3.6 ($P < 0.05$)

caused by seed contamination of the TILLed population. During our study, we identified 8 plants that were obviously homozygous contaminants in our Columbia *er105* mutagenized population, on the basis of the presence of multiple non-G/C-to-A/T changes, including small insertions and deletions. Therefore, we also expect a class of homozygous contaminating plants, each with a single detected polymorphic difference from Columbia. Of the 16 non-G/C-to-A/T plants, 11 are homozygotes, whereas we expected only 5–6 homozygotes by chance, and so some of these are likely to be non-Columbia contaminants. Two of the changes are found on the same fragment as G/C-to-A/T mutations and so may have resulted from error-prone repair. It is possible that some of the other exceptional changes lie near mutations on adjacent fragments and so also resulted from error-prone repair. The remaining 5 changes are heterozygotes, and these may have been spontaneous mutations that occurred in the progeny of the single plant that was used to collect seed for EMS mutagenesis or in their M_1 or M_2 descendants. We conclude that most, if not all, of the exceptions are likely to be spontaneous mutations or contaminants and that EMS is a nearly perfect mutagen for inducing G/C-to-A/T mutations in Arabidopsis.

Negative selection is inferred from a deficiency of homozygous protein truncation mutations: Mutations can be categorized as missense, truncation, or silent depending on how they affect the encoded protein. From the segments ordered for TILLING, we expected to find 48.3% missense mutations and found 50.1% (Table 1). Truncation mutations are of two types: mutations to nonsense codons and mutations to splice junction losses, either of which will lead to truncation or loss of protein and/or mRNA. We expected 4.3% nonsense mutations and found 3.4%, and we expected 1% splice junction losses and found 1.5%. Therefore, we observed 55% nonsilent mutations, which closely matches our expectation of 53.6%. This correspondence supports our assertion, based on the number of TILLed fragments recovered with truncation mutations, that all classes of EMS-induced mutations can be recovered at the expected frequencies.

We also expected to find twice as many heterozygotes as homozygotes owing to the selfing of M_1 plants to yield a 1:2:1 ratio of wild type:heterozygote:homozygote in the screened M_2 individuals. This expected ratio is unlikely to be biased by chimeric M_1 flowers, because relatively few cells make up the apical meristem in Arabidopsis (Koornneef 1994), and so almost every M_2 zygote should be from a single lineage. Indeed, ATP detected 2.08 times as many heterozygotes as homozygotes (Table 1), potentially fulfilling this expectation and suggesting that detection of heterozygotes relative to homozygotes is not noticeably compromised by pooling. In other words, detecting a heterozygote that is one-sixteenth of an eightfold pool appears to be as reliable as detecting a homozygote that is one-eighth of the pool.

Despite this close correspondence to expectation for the heterozygous:homozygous ratio, there were categorical differences. Both silent and missense mutations showed 2:1 ratios, but truncation mutations were significantly skewed in favor of heterozygotes (Table 1). Nonsense changes were discovered 3.6 times as often in heterozygotes ($n = 51$) as in homozygotes ($n = 14$) and splice junction losses were discovered 3.7 times as often (22 heterozygotes and 6 homozygotes). This skew is especially notable given that detection of heterozygotes in pools could be more difficult than detection of homozygotes. We attribute this relative deficiency in homozygous truncation mutations of both types to their severely deleterious effects on plants that inherit them in most cases. The strong skew found in these fragments most likely reflects the intention on the part of ATP users to TILL genes for which knockout changes are known or suspected to be lethal, where less severe hypomorphic mutations are most needed for functional studies.

EMS mutagenesis shows a local compositional bias: The large size of the TILLING data set and the singularity of the lesion caused by EMS allows us to sensitively detect local compositional biases. When we examine nucleotide positions flanking the mutated G, we detect deviations from random expectation on both sides (Figure 3). In both the -1 and the $+1$ positions from the mutated G, purines are more frequent and pyrimidines are less frequent than expected ($P \ll 10^{-12}$). The purine bias is slightly stronger for A (1.4) than for G (1.25) at both -1 and $+1$, and the pyrimidine bias is stronger for T at -1 (0.40) and for C at $+1$ (0.60; Table 2). Somewhat weaker, but still highly significant biases are seen at both -2 and $+2$ ($P < 10^{-8}$), but they are quite different from those at -1 and $+1$. Especially striking is the deficiency of G at -2 (0.75) and the excess of G at $+2$ (1.36). Weak biases ($P > 0.01$) are detected at -3 and $+3$ and at positions farther out to -10 and $+10$ (data not shown).

Finding biases on both sides of the mutated base raises the possibility that the biases are correlated, such as would be the case if particular motifs spanning the

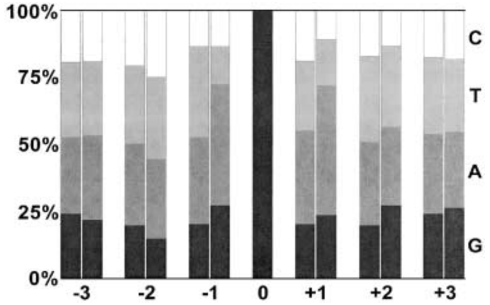


FIGURE 3.—Local compositional biases around the mutated G for all 1890 mutations. The expected frequencies (left bar in each pair) determined from the neighborhood of all Gs in the 192 fragments are compared to the observed frequencies (right bar in each pair) for all mutated Gs. Note that in 10 random samplings of 1890 Gs from the reference fragments the frequencies of the bases in each position differed by no more than 1%.

mutation were preferentially mutated. However, the most frequent motif, AGA, is seen no more frequently than would be expected from the product of the ratio of frequencies of A at both −1 and +1 (1.35 × 1.47 = 2.0 expected *vs.* 2.0 observed; Table 3). Indeed, the biases seen for the most overrepresented triplet (TGC) and the most underrepresented triplet (AGC) were not statistically significant. Because we are not able to discern a pattern to these biases, we tentatively conclude that compositional biases arise primarily from independent influences on the target G residue from its neighbors in the −2, −1, +1, and +2 positions.

We noted that the same mutation sometimes appeared in different plants at almost precisely the same frequency that we had initially expected by chance (55 observed *vs.* ~56 expected occurrences). The discovery of local biases raises the possibility that repeated mutations are similarly biased. For example, we might expect to find that AGA, which is overrepresented among the mutations discovered, is likewise overrepresented among the repeated mutations. Consistent with this hypothesis, we find that mutations within AGA account for 15 of the 55 repeated mutations (27%), which is even higher than its representation in the entire data set (24%) (Table 4). Adjusting for compositional biases at −1 and +1, we find that, overall, the 55 repeated mutations are slightly fewer than expected by chance. Therefore, no individual G residues in our screened target fragments appear to be hotspots for EMS-induced mutations.

Estimation of missing mutations: Although we detected no hotspots beyond the compositional bias, it is possible that individual plants differ in their susceptibility to mutation genome-wide. To test for this, we examined the distribution of mutations among the plants in the screening population. For any gene, only a fraction of available pools were screened, because suitable allelic series were usually obtained by screening fewer than the 6912 M_2 DNA samples that were prepared and arrayed for the project. For simplicity, we first consider only the five 96-well eightfold pool plates (representing 5 × 96 × 8 = 3840 plants) that were used for the bulk of the screening. These plates yielded 1564 mutations for 183 gene fragments. If all plants were equally likely to have yielded mutations, then we would have expected the 1564 mutations to be distributed among 1285 different plants $\{3840 \times [1 - (1 - 1/3840)^{1564}]\}$, whereas we observed mutations in 1184 plants, which is 92% of the expectation. We infer that mutations were missed in 8% of the plants.

One possible cause of the missing mutations is that not all pools were homogeneously screened. Examining the distribution of positives on pool plates, we see inhomogeneities on the 12 × 8 array for the five plates analyzed. In particular, well H12 showed only three mutations (Table 5). We can rationalize the deficiency: because the 96 samples were loaded on each 100-lane electrophoretic gel into lanes 4–99, with H12 transferred to lane 99, CEL 1-generated bands may have been overlooked occasionally because of edge effects. By the same token, we suspect that the remainder of the deficit can be explained by inhomogeneities of one sort or another, for example, by the addition of lane markers exclusively to samples arrayed in rows D and H, where the total number of mutations is lowest.

TABLE 2

Ratios of observed/expected frequencies on either side of the mutated G

	−3	−2	−1	0	+1	+2	+3
A	1.10	0.98	1.35		1.47	0.93	0.95
C	0.99	1.21	0.96		0.60	0.77	1.03
G	0.91	0.75	1.28		1.23	1.36	1.09
T	0.99	1.06	0.40		0.71	0.94	0.93
$\chi^2_{(3)}$	9.00	40.1	348		225	69.1	7.46
P	0.029	9.9×10^{-9}	$\ll 10^{-12}$		$\ll 10^{-12}$	2×10^{-12}	0.06

These values are derived by taking the ratio of the frequency of a particular nucleotide at the indicated distance from the mutated G to the frequency of that nucleotide at that distance from any available G found in the 192 TILLed fragments.

TABLE 3

Frequency of NGN triplet motifs

		GA				GC				GG				GT			
		n	obs	exp		*n*	obs	exp		*n*	obs	exp		*n*	obs	exp	Total
AG	AGA	454	2.0	2.0	AGC	71	0.55	0.81	AGG	202	1.7	1.7	AGT	123	0.91	1.0	850
CG	CGA	102	1.2	1.4	CGC	27	0.61	0.58	CGG	73	1.2	1.2	CGT	52	0.79	0.68	254
GG	GGA	251	1.8	1.9	GGC	50	0.74	0.77	GGG	122	2.0	1.6	GGT	93	0.78	0.91	516
TG	TGA	100	0.49	0.59	TGC	54	0.46	0.24	TGG	54	0.37	0.50	TGT	62	0.36	0.29	270
Total		907				202				451				330			1890

The observed values are derived by taking the ratio of the frequency of a particular triplet centered on the mutated G to the frequency of that triplet centered on any available G found in the 192 TILLed fragments. The expected values are the product of the appropriate NG and GN ratios calculated in Table 2.

Another way that mutations may have been missed arises from inhomogeneities in screening along the length of each fragment. Such missing mutations would not contribute to the 8% estimate because they would not be expected to cause variability among pools. To arrive at an estimate of losses from this type of inhomogeneity, we asked whether locations along each fragment are noticeably low in the number of mutations reported. Indeed, striking inhomogeneities are seen from a frequency histogram of mutations per fragment-length interval (Figure 4). It is clear that detection falls off toward both ends of fragments. In large part, falling off is expected because of weaker fluorescence toward the top of each lane and increasing fluorescence "noise" toward the bottom. This falling off appears to be independent of pooling ratios, because we see essentially the same histogram shape when homozygotes and heterozygotes are plotted independently (data not shown). To estimate the magnitude of underreporting caused by such inhomogeneities, we assume that optimal detection occurred within the fourth sextile of the histogram, where the greatest density of mutations was reported overall (Figure 4). If the density of mutations reported had been uniformly as high as we saw in the fourth sextile peak, then we would have reported a total of 2532 mutations, rather than the total of 1890 in our data set. This suggests that we have reported 25% fewer mutations than expected as a result of inhomogeneities along the lengths of fragments.

Coincident mutations provide estimates of mutation rates: We next asked whether there is an excess of eightfold pools with multiple mutations. Of 1847 eightfold pools with at least one positive, we found 43 pools with two mutant individuals, and no pools with three or more mutant individuals. We can use this number of coincident mutations in pools to estimate a mutation rate. Importantly, by basing our estimation only on positive pool samples, we avoid uncertainties caused by missed mutations in each pool screen. We estimate the average target to be 840 bp, which excludes the 80 bp from each end in which few mutations are discovered because of priming and systematic gel artifacts. This means that the 43 second mutations were found by screening 1847 × 860 × 8 = 1.27 × 10^4 kb for an overall density of 1 mutation/295 kb. The correspondence of this estimate to our initial rough estimate of 1/300 kb suggests that few mutations were missed in the pool screen. When corrected for (1) the estimated 8% of plants that did not contribute, (2) the estimated 25% of mutations not reported because of inhomogeneities along the length of fragments, (3) the dilution by one-fourth wild-type individuals because of the 1:2:1 Mendelian segregation in the M_2 generation, and (4) the higher G + C content of TILLed fragments (41%) relative to the genome as a whole (36%), we arrive at a mutation density of 1/170 kb (295 kb × 0.92 × 0.75 × 3/4 × 41/36).

We can also estimate the mutation density from coincidence within a fragment representing a single individual. Five such coincidences were discovered from among the 1890 positive fragments that were sequenced. This corresponds to a density of 1 mutation/300 kb [(1890 × 840)/5], which again is the same as our rough estimate.

TABLE 4

Percentages of NGN triplet motifs found as repeated mutations

	n	% obs	% exp		*n*	% obs	% exp		*n*	% obs	% exp		*n*	% obs	% exp
AGA	15	27	24.0	AGC	0	0	3.8	AGG	9	16	10.7	AGT	3	5	6.5
CGA	4	7	5.4	CGC	0	0	1.4	CGG	1	2	3.9	CGT	2	4	2.8
GGA	9	16	13.3	GGC	3	5	2.6	GGG	4	7	6.5	GGT	1	2	4.9
TGA	0	0	5.3	TGC	2	4	2.9	TGG	2	4	2.9	TGT	0	0	3.3

TABLE 5

Distribution of mutations projected on the 96-well plate map

	1	2	3	4	5	6	7	8	9	10	11	12	Total
A	7	13	15	16	19	20	24	14	23	13	19	11	194
B	13	14	17	15	22	26	16	17	12	23	11	13	199
C	9	15	16	21	26	30	22	8	10	23	20	24	223
D	9	10	8	16	23	14	14	7	14	12	17	15	159
E	13	7	11	25	18	22	22	18	20	13	19	14	202
F	11	17	18	25	21	27	18	20	18	12	14	22	223
G	14	15	12	12	13	25	22	13	19	25	14	8	192
H	12	9	8	18	18	17	18	16	18	23	12	3	172
Total	88	100	105	148	160	180	156	113	134	144	126	110	1564

This close agreement of mutation rates calculated from different parameters of the data set provides powerful confirmation of both the mutation density estimated and the high quality of TILLING data.

DISCUSSION

We have explored a data set of EMS-induced mutations that is nearly two orders of magnitude larger than that of previous reverse-genetic analyses. Bentley *et al.* (2000) reported on a series of 16 EMS-induced mutations in the *awd* gene in Drosophila, and Wienholds *et al.* (2002) reported on a series of 23 *N*-ethyl-*N*-nitrosourea (ENU)-induced mutations in the *Rag-1* gene in zebrafish. By contrast, we report on 192 genes with an average of 10 mutations per gene, and this much more broadly based and larger data set allows for a thorough analysis of a chemical mutagenesis spectrum *in vivo.* From this analysis, we can draw conclusions about the efficiency with which our method detects mutations and about their frequency in Arabidopsis, and we can speculate about the processes that take place in germplasm subjected to mutagen treatment.

High-throughput TILLING with CEL 1 is a robust detection method: Detection of heterozygotes is notoriously difficult using sequence traces (Nickerson *et al.* 2001), and so methods that are customized for mutation discovery have become popular (Kwok 2001). We use the CEL 1 mismatch-cleavage endonuclease for fragment detection on electrophoretic gels, because we judge that it would allow for robust detection of both heterozygous and homozygous point mutations (Colbert *et al.* 2001). Indeed, we find that screening 1-kb fragments in eightfold pools succeeded in detecting heterozygotes at the expected 2:1 ratio relative to homozygotes.

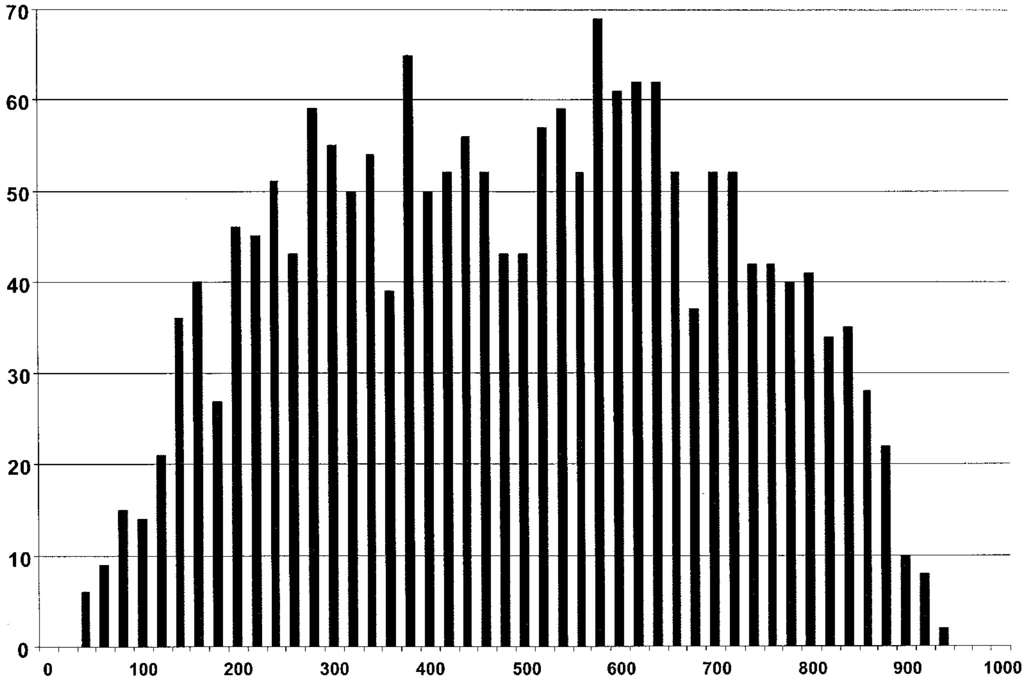


Figure 4.—Distribution of mutations as a function of scaled fragment coordinates. The length of each TILLed fragment was scaled to 1000 bp, and scaled coordinates of sequenced mutations were divided into 50 bins of equal length.

Features of our method make it inherently robust. First, double-end labeling allows for independent detection of the two cleavage products in different fluorescence channels, which guards against false positives. Second, we reamplify and retest each individual from a pool in the same way, which further eliminates false positives. Third, from fragment mobilities we estimate the position of the mutation within the fragment; we find that this estimation is accurate to ±10 bp (data not shown). We use both the near certainty of a mutation being in the fragment and the mobility information to help in identifying the heterozygous changes. Other technologies have been applied to single test samples and have demonstrated high accuracy in detection of heterozygotes (Spiegelman *et al.* 2000; Kwok 2001; Li *et al.* 2002). However, methods that simply report the presence of a migration or retention anomaly, such as single-strand conformational polymorphism, temperature-gradient capillary electrophoresis, and denaturing high performance liquid chromatography, do not provide positional information that is key to robustness of the CEL 1 method. Furthermore, we are unaware of any method that has been tested on pools of individuals in a production setting, where gene-specific customizations are impractical. It is notable that ATP has delivered allelic series for 93% of primer pairs submitted by users, where all failures were caused by insufficient PCR product, and not by inability to detect mutations in the target. Therefore, ours may be the only currently available technology that is suitable for discovery of rare heterozygous changes in large populations.

Our analysis also reveals that production-scale operation did not substantially compromise the discovery of rare changes. We could deduce that, at most, 8% of the plant DNAs screened may have been overlooked in a screen. Some of these missed samples came from inhomogeneities at one edge of the electrophoretic gel, and we think that other aspects of the screening procedure, such as adding lane markers to every fourth sample, can account for the other missed samples. Sample-to-sample variations in the amount of DNA in each pool would have also compromised detection. However, samples had been carefully normalized to prevent such inhomogeneities in the pools, a procedure that very likely led to detection of nearly all mutations. We also found that we had underreported mutations toward both ends of fragments. We attribute these losses primarily to poor fluorescence signal; however, it is likely that most of these mutations were initially detected, but not followed up because of ATP's practice of pursuing only the 12 best positive pools. Because of this policy, we cannot distinguish mutations that were not detected from those that were detected in pools but not reported.

Although we detected essentially all heterozygous mutations, on the basis of finding the expected ratio of heterozygotes to homozygotes overall, it is striking that truncation mutations were severely depleted in homozygotes, but not heterozygotes. Specifically, the ratio of heterozygotes to homozygotes for truncation mutations averaged 3.6:1, as opposed to 2:1 overall. This depletion indicates that for many of the TILLed genes, severely deleterious EMS-induced lesions were present in the M_1 generation, and yet we still obtained heterozygous mutations in the M_2 generation at the expected level. This statistical evidence that a large fraction of TILLed genes are viable and fertile as heterozygotes, but are severely deficient as homozygotes, supports the notion that TILLING mutations will be generally useful for determining gene function.

EMS-induced mutations are randomly distributed G/C-to-A/T changes: Our analysis provides compelling evidence that the Arabidopsis TILLING data set is of high quality and that there were relatively few false negatives, and probably no false positives. Therefore, we can use this data set to draw firm conclusions about the EMS-induced mutational spectrum. At least 99.5% of mutations are G/C-to-A/T changes and the rest might have been secondary mutations caused by error-prone repair. In previous work, G/C-to-A/T transitions have been found to dominate, but exceptions have been reported. For example, of the nine EMS-induced changes at the Arabidopsis *sos1* locus, only seven were G/C-to-A/T mutations and the other two were deletions of 1 and 16 bp (Shi *et al.* 2000). We suggest that, when compared to all mutations that occurred in *sos1*, the two deletions were actually much rarer events, but by causing a knockout phenotype, these mutants were readily selected for. Our data reveal just how rare these exceptional but selectable mutations are. Forward genetic screens have also revealed examples of repeated mutations, such as the several independently EMS-mutagenized plants that have been found to carry the same LEAFY mutation (Weigel *et al.* 1992). Our data do not confirm these observations, as we failed to detect even a single example in which more than two independently mutagenized plants have the same mutation. These examples illustrate the advantages of using comprehensive reverse-genetic data for drawing inferences about the distribution of mutations.

While this article was being submitted, a single-gene study reporting 16 EMS-induced gain-of-function mutations for Arabidopsis estimated a mutation density that is not significantly different from our measurement of 1/170 kb (Jander *et al.* 2003). Furthermore, a single-gene reverse-genetic screen in Drosophila reporting 16 mutations estimated an unbiased mutation density for EMS mutagenesis of 1/210 kb (Bentley *et al.* 2000), which also is not significantly different from our measurement. We do not consider this close correspondence to be coincidental, because for both Arabidopsis and Drosophila, mutagenesis protocols attempt to maximize the mutational density up to the point that M_1 viability and fertility become severely compromised. The close correspondence in EMS-induced mutation densities between a plant and an animal might reflect a balance between the costs and benefits of mutation and repair. If so, then similar mutational densities will be found for other alkylating agents and other organisms, regardless of genome size. Accordingly, we and others are applying TILLING to organisms with larger genomes, such as EMS-mutagenized maize and ENU-mutagenized zebrafish (our unpublished results).

Local compositional biases suggest that DNA repair is rate limiting for EMS mutagenesis: Although we detected no hotspots for mutations and all parts of the genome appeared to be equally susceptible, we did detect local compositional biases. The nearest two neighbors on either side of the mutated G showed strong skews, decreasing in degree from −1 and +1 to −2 and +2 and continuing weakly beyond. At −1 and +1, purines were in excess, with adenines slightly favored over guanines; at −2, guanines were deficient; and at +2, guanines were in excess, with other weaker biases detected at these positions. For Drosophila, Bentley *et al.* (2000) described a stronger purine nearest-neighbor bias, but theirs was essentially all guanine. In addition, they report weaker biases, including hotspots. However, on the basis of a data set that is 100-fold larger, we can rule out hotspots and other local preferences of the type reported by Bentley and co-workers. It is possible that the spectrum of EMS-induced mutations differs between Arabidopsis and Drosophila. Alternatively, we may be detecting the same −1 and +1 purine bias as Bentley and co-workers reported, but the weaker biases that they pointed out were based on too few data, or the single gene that they examined lies in an atypical region. Therefore, our results might yet generalize to animals.

What caused the bias? Three general possibilities come to mind. One is that the bias reflects a preference for CEL 1 cleavages. For example, CEL 1 might not cleave well at the G of CGC, which is the least common motif found. If so, then it would be more difficult to detect this mutation in heterozygotes, which are one-sixteenth of each pool, than in homozygotes, which are one-eighth. However, we detect identical local biases for both heterozygotes and homozygotes (data not shown). Another possibility is that EMS adducts are sensitive to the identities of the nearest two neighboring base pairs on either side. This possibility would require a level of complex chemical specificity over a 5-bp duplex region that seems implausible for a small miscible organic molecule.

A third possibility is that adducts occur with similar probabilities at all G residues, but that the adducts are removed (Vidal *et al.* 1995) or the resulting mutations are repaired (Marsischky and Kolodner 1999) at different rates depending on their local environment. We believe that this is a very likely explanation for the bias. Mismatch repair of alkylation damage is a well-studied process that displays local biases in model systems. Sites of damage are recognized by MutS in bacteria and by MutS homologs (MSH proteins) in eukaryotes, and the damaged region is excised and repaired by a special DNA polymerase. In the case of very short patch repair, a yeast MSH2-MSH6 dimer has been shown to recognize a G/T mismatch with different affinities depending on the identities of several base pairs on either side (Marsischky and Kolodner 1999). We suggest that recognition by the counterparts of yeast MSH2-MSH6 is responsible for the compositional biases seen in EMS mutagenesis. Following EMS treatment, the number of adducts will be larger than the number of available MSH2-MSH6 dimers in the germplasm, and so some damage will go unrepaired. In this way, the unrepaired damage, which is the adducts in regions that dimers disfavor, would show the bias that we found.

Given that DNA repair pathways are very similar among diverse organisms (Tuteja *et al.* 2001), we expect that compositional biases of the type that we found will generalize. In Arabidopsis, MSH2-MSH6 and MSH2-MSH7 dimers are thought to recognize G:T mismatch damage (Culligan and Hays 2000), and a homolog of the OGG1 DNA glycosylase excises alkylated nucleotides (Dany and Tissier 2001; Garcia-Ortiz *et al.* 2001). It will be interesting to determine the effects of mutations in these proteins. Because it is thought that these heterodimers counteract alkylation damage, mutations in them should allow for higher mutational densities to be obtained using the same dosages of EMS. As a result, mismatch repair lesions might make TILLING more effective. T-DNA insertional mutations have been recovered in *msh2*, *msh6*, and *msh7* (http://signal.salk.edu), and we are currently TILLING for hypomorphic mutations, which should provide us with a series of mutants that display a range of repair phenotypes. Among these, we hope to identify healthy lines that will accommodate greater densities of point mutations. In this way, TILLING might be used to improve the TILLING process itself.

We thank past members of the high-throughput TILLING team, including Terri Bryson, Nina Accornero, Trent Colbert, Rebecca Lechalk, and Mike Steine. We also thank Elisabeth Spitzer, Emily Kerr, Tracy Cunningham, Bill Orr, Elisabeth Schnackenberg, Cresanna Puffer, Angela Jacobs, Laura Vasquez, Steve Hentel, Ernest Cho, Sabrina Anderson, Paul Beeman, Michelle Acupanda, Brianna Borders, Amy Holmes, and Amber Kost for planting, harvesting, and preparing the DNA samples. This work was funded by the National Science Foundation Plant Genome Research Program.

LITERATURE CITED

Ashburner, M., 1990 *Drosophila: A Laboratory Handbook.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Bentley, A., B. MacLennan, J. Calvo and C. R. Dearolf, 2000 Targeted recovery of mutations in Drosophila. Genetics **156:** 1169–1173.

Colbert, T., B. J. Till, R. Tompa, S. Reynolds, M. N. Steine *et al.*, 2001 High-throughput screening for induced point mutations. Plant Physiol. **126:** 480–484.

Culligan, K. M., and J. B. Hays, 2000 Arabidopsis MutS homologs—AtMSH2, AtMSH3, AtMSH6, and a novel AtMSH7—form three distinct protein heterodimers with different specificities for mismatched DNA. Plant Cell **12:** 991–1002.

Dany, A. L., and A. Tissier, 2001 A functional OGG1 homologue from *Arabiodopsis thaliana.* Mol. Gen. Genet. **265:** 293–301.

Garcia-Ortiz, M. V., R. R. Ariza and T. Roldan-Arjona, 2001 An OGG1 orthologue encoding a functional 8-oxoguanine DNA glycosylase/lyase in *Arabidopsis thaliana.* Plant Mol. Biol. **47:** 795–804.

Haughn, G., and C. R. Somerville, 1987 Selection for herbicide resistance at the whole-plant level, pp. 98–107 in *Applications of Biotechnology to Agricultural Chemistry*, edited by H. M. LeBaron, R. O. Mumma, R. C. Honeycutt and J. H. Duesing. American Chemical Society, Easton, PA.

Henikoff, S., and L. Comai, 2003 Single-nucleotide mutations for plant functional genomics. Annu. Rev. Plant Biol. **54:** 375–401.

Jander, G., S. R. Baerson, J. A. Hudak, K. A. Gonzalez, K. J. Gruys *et al.*, 2003 Ethylmethanesulfonate saturation mutagenesis in Arabidopsis to determine frequency of herbicide resistance. Plant Physiol. **131:** 139–146.

Koornneef, M., 1994 *Arabidopsis* genetics, pp. 89–120 in *Arabidopsis*, edited by E. M. Meyerowitz and C. R. Somerville. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Kovalchuk, I., O. Kovalchuk and B. Hohn, 2000 Genome-wide variation of the somatic mutation frequency in transgenic plants. EMBO J. **19:** 4431–4438.

Kwok, P. Y., 2001 Methods for genotyping single nucleotide polymorphisms. Annu. Rev. Genomics Hum. Genet. **2:** 235–258.

Li, Q., Z. Liu, H. Monroe and C. T. Culiat, 2002 Integrated platform for detection of DNA sequence variants using capillary array electrophoresis. Electrophoresis **23:** 1499–1511.

Marsischky, G. T., and R. D. Kolodner, 1999 Biochemical characterization of the interaction between the *Saccharomyces cerevisiae* MSH2-MSH6 complex and mispaired bases in DNA. J. Biol. Chem. **274:** 26668–26682.

McCallum, C. M., L. Comai, E. A. Greene and S. Henikoff, 2000 Targeted screening for induced mutations. Nat. Biotechnol. **18:** 455–457.

Muller, H. J., 1930 Types of visible variations induced by X-rays in *Drosophila.* J. Genet. **22:** 299–334.

Nickerson, D. A., N. Kolker, S. L. Taylor and M. J. Rieder, 2001 Sequence-based detection of single nucleotide polymorphism. Methods Mol. Biol. **175:** 29–35.

Oleykowski, C. A., C. R. Bronson Mullins, A. K. Godwin and A. T. Yeung, 1998 Mutation detection using a novel plant endonuclease. Nucleic Acids Res. **26:** 4597–4602.

Rozen, S., and H. Skaletsky, 2000 Primer3 on the WWW for general users and for biologist programmers. Methods Mol. Biol. **132:** 365–386.

Shi, H., M. Ishitani, C. Kim and J. K. Zhu, 2000 The *Arabidopsis thaliana* salt tolerance gene SOS1 encodes a putative Na+/H+ antiporter. Proc. Natl. Acad. Sci. USA **97:** 6896–6901.

Spiegelman, J. I., M. N. Mindrinos and P. J. Oefner, 2000 High-accuracy DNA sequence variation screening by DHPLC. Biotechniques **29:** 1084–1090.

Till, B. J., T. Colbert, R. Tompa, L. Enns, C. Codomo *et al.*, 2003a High-throughput TILLING for functional genomics, pp. 205–220 in *Plant Functional Genomics: Methods and Protocols*, edited by E. Grotewald. Humana Press, Clifton, NJ.

Till, B. J., S. H. Reynolds, E. A. Greene, C. A. Codomo, L. C. Enns *et al.*, 2003b Large-scale discovery of induced point mutations with high throughput TILLING. Genome Res. **13:** 524–530.

Tuteja, N., M. B. Singh, M. K. Misra, P. L. Bhalla and R. Tuteja, 2001 Molecular mechanisms of DNA damage and repair: progress in plants. Crit. Rev. Biochem. Mol. Biol. **36:** 337–397.

Vidal, A., N. Abril and C. Pueyo, 1995 DNA repair by Ogt alkyltransferase influences EMS mutational specificity. Carcinogenesis **16:** 817–821.

Weigel, C., J. Alvarez, D. R. Smyth, M. F. Yanofsky and E. M. Meyerowitz, 1992 LEAFY controls floral meristem identity in Arabidopsis. Cell **69:** 843–859.

Wienholds, E., S. Schulte-Merker, B. Walderich and R. H. Plasterk, 2002 Target-selected inactivation of the zebrafish rag1 gene. Science **297:** 99–102.

Communicating editor: V. Sundaresan