Discovery of Rare Mutations in Populations: TILLING by Sequencing


Department of Plant Biology and Genome Center (H.T., T.H., B.W., K.J.N., M.L., R.K.T., A.A.K., L.C.), Department of Plant Sciences (R.N., C.U., T.H.T., J.D.), Department of Computer Sciences (V.M., V.F.), Bioinformatics Core, Genome Center (J.F.), and United States Department of Agriculture Agricultural Research Service, Crops Pathology and Genetics Research Unit (T.H.T.), University of California, Davis, California 95616

Discovery of rare mutations in populations requires methods, such as TILLING (for Targeting Induced Local Lesions in Genomes), for processing and analyzing many individuals in parallel. Previous TILLING protocols employed enzymatic or physical discrimination of heteroduplexes from homoduplexed target DNA. Using mutant populations of rice (Oryza sativa) and wheat (Triticum durum), we developed a method based on Illumina sequencing of target genes amplified from multidimensionally pooled templates representing 768 individuals per experiment. Parallel processing of sequencing libraries was aided by unique tracer sequences and barcodes allowing flexibility in the number and pooling arrangement of targeted genes, species, and pooling scheme. Sequencing reads were processed and aligned to the reference to identify possible single-nucleotide changes, which were then evaluated for frequency, sequencing quality, intersection pattern in pools, and statistical relevance to produce a Bayesian score with an associated confidence threshold. Discovery was robust both in rice and wheat using either bidimensional or tridimensional pooling schemes. The method compared favorably with other molecular and computational approaches, providing high sensitivity and specificity.

The discovery of rare mutations is critical for basic science as well as for biomedical and biotechnological progress. TILLING (for Targeting Induced Local Lesions in Genomes) is a functional genomics method that discovers rare mutations in populations. It consists of mutagenesis, DNA isolation and pooling, and high-throughput mutation discovery in target genes. First described in Arabidopsis (Arabidopsis thaliana; McCallum et al., 2000) and Drosophila (Bentley et al., 2000), it has been successfully extended to multiple model and economic species (Comai and Henikoff, 2006; Gilchrist and Haughn, 2010), becoming a major tool for functional genomics. Traditionally, TILLING discovers mutations through the detection of PCR products digested at mismatched sites in heteroduplexes with CELI endonuclease (Oleykowski et al., 1998) or with enzymatic mixes from celery (Apium graveolens) extracts (celery juice extracts [CJE]; Till et al., 2004). Heteroduplexes can also be detected by high-resolution melting (Dong et al., 2009). These methods, while effective, are labor intensive and are challenged by pools deeper than eight individuals, and once mismatches have been detected, sequencing is required to characterize the mutations.

Alternatively, mutations and polymorphisms can be discovered through sequencing followed by comparison of a putative new allele with a reference sequence. This method requires the discrimination of sequencing errors from real changes, a task facilitated by sufficient sequencing coverage and bioinformatic analysis (Druley et al., 2009; Koboldt et al., 2009; Bansal, 2010; Bansal et al., 2010). High-throughput sequencing produces millions of DNA sequence reads in a single experiment and opens the possibility of developing mutation discovery methods based on massive parallel sequencing (Gilchrist and Haughn, 2010). Recently, the rate and spectrum of spontaneous mutations in Arabidopsis were analyzed by whole-genome resequencing (Ossowski et al., 2010). Detection of single-nucleotide polymorphisms (SNPs), however, requires substantial sequencing coverage (20×–30× for a diploid genome), and the cost of resequencing a large population is still substantial. To decrease resequencing costs, it is ad-
vantageous to focus on a subset of genes. This can be done by sequencing PCR products (Porreca et al., 2007; Rigola et al., 2009) or by using oligonucleotide-mediated sequence capture, followed by release and subsequent sequencing (Nijman et al., 2010). Furthermore, protocols that address the challenges in finding rare polymorphisms in populations have been developed. Rigola et al. (2009) described the use of 454 sequencing of overlapping pools coupled with a computational pipeline to identify mutations in tomato (Solanum lycopersicum). A number of computational tools exist for finding SNPs in nonoverlapping pools sequenced with next-generation technology, such as VarScan (Koboldt et al., 2009), CRISP (Bansal, 2010; Bansal et al., 2010), SNPSeeker (Druley et al., 2009), SAMtools (Li et al., 2009), and MAQ (Li et al., 2008). However, these methods and computational tools were not designed for an overlapping, multidimensionally pooled, Illumina GA-sequence based system, and they have limited performance in this context (V. Missirian, L. Comai, and V. Filkov, unpublished data).

We describe here the use of next-generation sequencing coupled with multidimensional pooling for the identification of rare alleles in populations using rice (Oryza sativa) and wheat (Triticum durum and Triticum aestivum) as examples. Rice is a diploid with a 389-Mb genome (Zhang et al., 2008). Durum wheat is an allotetraploid that combines two ancestral parental genomes (A and B) adding to 12,800 Mb (Rees and Walters, 1965). In wheat, PCR amplification can be targeted to one or the other homeologous gene by primer design (Uauy et al., 2009). To establish the feasibility of the approach, we targeted genes of interest in characterized populations that carry frequent mutations. By endonuclease digestion of mismatched heteroduplexes, we previously characterized mutations and attributed events to sequencing errors. These averaged change at position 808 [A808G] in RNA-Directed RNA polymerase2 (RDR2, Os02g50330), C1326T in Inositol 1,3,4-Triphosphate 5/6-Kinase (PITPK, Os10g01480), T469C in putative retrotransposon protein TGT2A (Os11g37750). Using a large pool of wild-type individuals, the three mutant individuals were diluted in ratios of 1:64, 1:96, 1:128, and 1:192 (or 1:128, 1:192, 1:256, and 1:384 mutant allele to wild-type allele). These numbers were chosen because they are based on convenient pooling combinations of wells on 96-well plates and cover a range compatible with high-throughput processing. The particular gene containing the mutation in each of the three individuals was then amplified from each of the four template pools. Each template pool was amplified in triplicate to allow additional assessment of technical variation and then processed for Illumina sequencing.

Sequence reads were aligned to the target sequence reference, and both coverage (the number of times a given base was represented in the sequencing results) and nucleotide variation were assessed per reference base. We excluded changes that were associated with Solexa sequencing quality scores lower than 20 (equivalent to $P = 0.01$). The observed base change frequencies were plotted against base position for visual identification of mutations (Fig. 1A). As seen in Figure 1A, coverage varied within genes with higher coverage at the gene termini corresponding to the primers, followed by a decline and then a gradual increase to average coverage level. Although gene inputs were standardized, the average coverage varied between genes as well. Background changes, referred to as noise, were evident at virtually all positions and could be attributed to sequencing errors. These averaged below one change in 500 bases (0.0020 frequency) and could not be eliminated even by stringent quality filtering, indicating that they are unlikely to be caused by errors in sequencing. Decreased coverage increased the noise level (Fig. 1, A and B). This noise hindered the discovery of mutations in pool sizes of 128 (0.0039 frequency, or one allele in 256) and 192 (0.0026 frequency) individuals, because real mutations and noise could display similar frequencies (Fig. 1C). A pool size of 96 has an expected heterozygous mutant frequency (0.0052) that is 2.6-fold higher than the average background noise and appeared acceptable for mutation discovery as long as coverage stayed above a critical level estimated to be at least 10× (per bp) per allele (or 96 × 2 × 10 = 1,920 for a pool of 96 individuals). The observed mutation frequency averaged above the expected frequency (Fig. 1C, left panel), probably because the noise signal still emitted by other individuals in the pool and the mutation signal emitted by a single mutant individual were additive.

The average noise levels appeared to be base-change dependent (Fig. 1C, right panel). Transitions were noisier than transversions. Changes from AT base pair to GC (AT > GC transitions) were more costly. Changes from GC base pair to AT (GC > AT transitions), characteristic of the G-alkylating action of ethyl

RESULTS

Pilot Experiments

Efficient identification of mutants in a multidimensionally pooled large population requires the resolution of complex samples. Large pools allow the screening of more individuals for a given labor input, but pool size is limited by the ability to detect a single mutation in a large background. To establish the resolution limit of this approach, we selected three rice individuals, each with a different heterozygous mutation (an A-to-G change at position 808 [A808G] in RNA-Directed RNA polymerase2 (RDR2, Os02g50330), C1326T in Inositol 1,3,4-Triphosphate 5/6-Kinase (PITPK, Os10g01480), T469C in putative retrotransposon protein TGT2A (Os11g37750). Using a large pool of wild-type individuals, the three mutant individuals were diluted in ratios of 1:64, 1:96, 1:128, and 1:192 (or 1:128, 1:192, 1:256, and 1:384 mutant allele to wild-type allele). These numbers were chosen because they are based on convenient pooling combinations of wells on 96-well plates and cover a range compatible with high-throughput processing. The particular gene containing the mutation in each of the three individuals was then amplified from each of the four template pools. Each template pool was amplified in triplicate to allow additional assessment of technical variation and then processed for Illumina sequencing.

Sequence reads were aligned to the target sequence reference, and both coverage (the number of times a given base was represented in the sequencing results) and nucleotide variation were assessed per reference base. We excluded changes that were associated with Solexa sequencing quality scores lower than 20 (equivalent to $P = 0.01$). The observed base change frequencies were plotted against base position for visual identification of mutations (Fig. 1A). As seen in Figure 1A, coverage varied within genes with higher coverage at the gene termini corresponding to the primers, followed by a decline and then a gradual increase to average coverage level. Although gene inputs were standardized, the average coverage varied between genes as well. Background changes, referred to as noise, were evident at virtually all positions and could be attributed to sequencing errors. These averaged below one change in 500 bases (0.0020 frequency) and could not be eliminated even by stringent quality filtering, indicating that they are unlikely to be caused by errors in sequencing. Decreased coverage increased the noise level (Fig. 1, A and B). This noise hindered the discovery of mutations in pool sizes of 128 (0.0039 frequency, or one allele in 256) and 192 (0.0026 frequency) individuals, because real mutations and noise could display similar frequencies (Fig. 1C). A pool size of 96 has an expected heterozygous mutant frequency (0.0052) that is 2.6-fold higher than the average background noise and appeared acceptable for mutation discovery as long as coverage stayed above a critical level estimated to be at least 10× (per bp) per allele (or 96 × 2 × 10 = 1,920 for a pool of 96 individuals). The observed mutation frequency averaged above the expected frequency (Fig. 1C, left panel), probably because the noise signal still emitted by other individuals in the pool and the mutation signal emitted by a single mutant individual were additive.

The average noise levels appeared to be base-change dependent (Fig. 1C, right panel). Transitions were noisier than transversions. Changes from AT base pair to GC (AT > GC transitions) were more costly. Changes from GC base pair to AT (GC > AT transitions), characteristic of the G-alkylating action of ethyl
methanesulfonate (EMS), were moderately noisy. Errors resulting in transversions were less frequent, with GC > CG being the least common.

In conclusion, noise increased with lowered coverage, and mutation detection was inefficient at dilutions of one heterozygous mutant in 192 wild-type
individuals (one mutant allele in 384 = 0.0026), which were only slightly above the average noise level (0.0020). We concluded that pools of 64 and 96 are acceptable for mutation detection using this methodology.

Bidimensional Pooling

We expanded the scope of our mutation discovery by searching for new mutations in characterized rice (Till et al., 2007) and wheat (Uauy et al., 2009) populations. The genomic DNA of 768 individuals was used to create two types of pools, eight “row pools” (96 individuals each) and 12 “column pools” (64 individuals each), for a total of 20 “pools” (Fig. 2). Each individual was represented twice, once in the row pool and once in the column pool. A mutation found in both the row and column pools identified a pool of eight individuals where one of the eight individuals carries the mutation. A total of 18 targets (13 rice and five wheat), averaging 1.5 kb each, were sequenced. The experiment was designed so that the average coverage per gene would be 30× per base for the row pools and 45× per base for the column pools, providing a substantial margin of error (Supplemental Fig. S1).

Data for target gene VRN3 are displayed in Figure 3. Mutations can be visually identified by observing two outlier points in a frequency plot (Fig. 3A) that correspond to the two pools sharing the mutation. Variance in sequence coverage along a gene produced noise, which diminished the ability to detect mutations (Fig. 3B). Similar to what was observed in the pilot experiment, the combination of primer location and low coverage resulted in a blind region for mutation detection of about 100 bp at each PCR fragment terminus.

Figure 2. Bidimensional pooling strategy for mutation discovery. Micropools of eight individuals arrayed on 96-well plates were super-pooled by combining all the wells in a row or those in a column and using the pooled template for PCR amplification of the targets. Super-pools consisted of 96 or 64 individuals for row and column, respectively. PCR products for multiple targets were amplified from each super-pool and combined to make 20 different Illumina libraries. Experiments 1 and 2 employed this strategy. [See online article for color version of this figure.]

Figure 3. Coverage, mutation signal, and noise. A, Identification of a hypothetical G > A mutation. The expected frequency of G > A changes at five base positions is illustrated. Residues that are wild type for all individuals of the population display background frequency of changes (i.e. noise), while a residue corresponding to a mutant individual in the population displays a pair of outliers corresponding to the libraries derived from the row and column pools, where such a mutant shifts the frequency in proportion to its contribution to the pool.

B, Wheat VRN3 analysis from 20 bidimensional pools encompassing 96× or 64× combinations of 768 individuals was sequenced (experiment 2), achieving the illustrated coverage (top track). Coverage per individual is 1/64th or 1/96th of that indicated, depending on the pooling ratio of the template DNA used. The percentage changes in the given base are plotted versus the base position (lower tracks of each panel). Pairs of outlier points corresponding to mutations identified in one row pool and one column pool out of the total of 20 pools are obvious in the G > A (marked by circles) and C > T tracks (not marked).

A > G changes, which are not expected from EMS, are plotted for comparison (bottom track), demonstrating the absence of outlier pairs. Two types of noisy positions are evident: those corresponding to lower coverage adjacent to the termini of the fragment and those corresponding to a microsatellite-rich region with decreased coverage (sequence: 748-GTGTGTGTGTGTGTGTGCGCGCGGTGTGCATCTACATGTGTGTGCGGCCCTTGAC-800). [See online article for color version of this figure.]
In the case of wheat genes, where EMS-induced changes were almost exclusively GC > AT, outlier pairs were frequent only in the G > A and C > T tracks (Fig. 3B, two middle panels) and not, for example, in the A > G change track (Fig. 3B, bottom panel). By the relatively simple criterion of having a base change frequency greater than 2.5× higher than the average background noise, many candidate mutations were identified in both rice and wheat.

**Tridimensional Pooling**

The bidimensional pooling scheme described above assigned mutations to pools of eight individuals, which then need to be sequenced and analyzed individually to assign the mutation to a specific individual and seed stock. This task complicates a high-throughput workflow based on next-generation sequencing. Therefore, for high-throughput workflows, we developed an approach that identifies single mutant individuals through the addition of a third pooling dimension. We chose a maximum pool of 64 (one in 128 alleles). A total of 768 rice individuals were arrayed in a tridimensional grid to form 16 row pools (48 individuals), 16 column pools (48 individuals), and 12 “dimensional” (or “D”) pools (64 individuals). The names are arbitrary and reflect the arraying strategy (Fig. 4). Thirty-one different target-gene segments were amplified from each of the 44 rice pools, processed for Illumina GA sequencing (see “Materials and Methods”), marked with four different barcoded adapters, and sequenced, four pools at a time, in 12 Illumina GA lanes using 40-base reads. To address the possibility of contamination during the construction of multiple libraries, we added to each library before DNA fragmentation a unique tracer sequence (the “spike”), whose presence was thus expected in only one of the 44. Since the spike was added in the initial step of library construction and before the addition of barcoded adapters, and because libraries processed in different lanes of Illumina GA flow cells may share the same barcodes, the spikes provided valuable verification that samples were properly processed and contamination free. As in the bidimensional analysis, candidate base changes could be identified using the criterion of frequency outliers present in three different pools. Figure 5 displays examples of these analyses. The average coverage for each rice target in each of 44 pools is presented in Figure 5A. The effect of high GC content on coverage along a gene is illustrated in Figure 5B. Triplets of outliers are demonstrated in Figure 5C.

**A Bioinformatic Pipeline for Mutation Identification**

Discovery of mutations presented two challenges. First, the volume of data was large and could not easily be processed manually. Second, separating mutation signal from noise was complicated because both the clarity of the signal and the level of the noise could vary considerably according to position and coverage. We established a bioinformatic pipeline that evolved over multiple experiments. Its general steps include quality filtering of the reads and trimming to remove adapter sequences, identification of multiplexed samples according to barcode, and the use of the software BWA (Li and Durbin, 2009; or MAQ [Li et al., 2008] in early experiments) to align the sequence to the reference, identify and parse changes from the BWA output, compare their frequency across the multiple libraries, and derive a probability for each change considering quality scores and other parameters. The pipeline incorporated CAMBa (for Coverage Aware Mutation Calling Using Bayesian Analysis [pronounced like the dance]; V. Missirian, L. Comai, and V. Filkov, unpublished data), a program that employs Bayesian interpretation to estimate the probability that a mutation is falsely called with an F1 score, the mean-centered log of

---

**Figure 4.** Tridimensional pooling and sampling strategies for mutation discovery. Individual genomic DNAs represented by the small cubes were arrayed on plates of 8 × 8 wells. Pools were formed by combining individual DNA along the three-dimensional axis, here called arbitrarily D-pools, column pools, and row pools. Each of these three pooling arrangements is illustrated by the color arrays (green, red, or blue). The pooling resulted in 44 mixed templates. Forty-four gene targets amplified from each template mix were pooled and processed in a single Illumina library. Four libraries indexed by unique barcodes were combined and sequenced in a single Illumina lane. Plausible and implausible changes for mutation identification were derived by comparing libraries as exemplified and used for setting the probability threshold, as described in “Results.”

---
Figure 5. Sequencing coverage and mutation discovery in tridimensionally pooled gene fragments. In this experiment (no. 3), target PCR products amplified from 44 template pools encompassing 64× or 48× combinations of 768 individuals were sequenced. A, The average sequencing depth in the 44 pools of each rice gene fragment is displayed. The boxed data sets correspond to the genes in B and C. B, Example of intragenic fluctuation in coverage associated with extreme GC contents. The top panel displays the percentage GC in a 10-bp window moved in 1-bp increments along the rice gene Os04g56580. The bottom panel reports the 0 to 5,000× coverage range. High GC content regions (arrows) correspond to low sequencing coverage. C, Inositol kinase-like (Os09g34300) gene of rice. The top panel displays sequencing coverage. The change frequencies are
the posterior probability of mutation. CAMBa uses pool overlap in the multidimensional pooling experiments to identify both the mutations and the individuals (or subpools) in which they occur. CAMBa takes into consideration the pooling depth, the expected frequency and type of mutations, the probability of heterozygosis, the sequencing quality of the call, and the frequency of forward and reverse reads. For example, to produce a list of candidate changes in bidimensional pooling, CAMBa compares the frequency of each change in all pool libraries, evaluating the probability that two outliers, one corresponding to a row pool (first dimension) and the other to a column pool (second dimension), identified a unique well containing the mutant individual. CAMBa works similarly for tridimensional and higher dimensional pooling experiments.

Comparison of Unidimensional with Bidimensional and Tridimensional Pooling

Our strategy for mutation discovery was enhanced by the double or triple validation of intersecting pools. In bidimensional and higher dimensional designs, the overlapping pools allow CAMBa to identify the subpool or individual carrying the mutation and also to increase its detection sensitivity by requiring that each mutation be present in multiple pools (thus, higher dimensional setups would likely yield a higher number of true positives). By changing the parameters of CAMBa’s search to unidimensional pooling, we looked for and found singletons (i.e. outlier changes that were not in pairs but occurred in a single library), which we termed orphans (Fig. 5C). In wheat, some of the orphans could be accounted for as instances where the cognate change was less easily detected. In rice, however, most of the orphans were inconsistent with mutations, because their frequency varied unexpectedly from gene to gene and was higher than the orphan frequency in wheat, despite having a lower expected mutation rate (3.3 per Mb in rice versus 19.7 per Mb in wheat; Supplemental Fig. S2). For these reasons, the occurrence of orphans in rice was best explained as some kind of artifact. As a consequence, in rice, unidimensional pooling would have resulted in a very high false-positive rate. We have since been able to decrease the incidence of orphans by increasing the input of template DNA in PCR (see “Discussion”).

Mutation Calls Are Associated with a Probability Value

CAMBa yields candidate mutations with an associated probability of the null hypothesis of no mutation (or a different candidate mutation at the same position). The measure $F_t$ is a logarithmic conversion of $P$. The output list starts with mutations that have a high $F_t$ (typically in the 30s, corresponding to very low probability [i.e. changes that are very unlikely to be false positive]) and ends with mutations that have a negative $F_t$ (higher probability that change may be a false positive). From a practical point of view, however, one would like to know where in the ranked list a line should be drawn separating high-confidence calls from less reliable ones. The impact of setting such a line, the probability threshold, at different places on a list is illustrated in Supplemental Figure S3. As the threshold increased, the mutant predictions were more reliable. At the same time, fewer mutations were discovered. Decreasing the probability threshold (allowing less probable changes) increased the yield of putative mutations, but it did so more pronouncedly in rice than in wheat, indicating that the two experimental systems differed in noise level.

For both rice and wheat, an estimate of the expected number of mutations could be derived from previous studies. By setting the threshold $F_t$ to a value that yielded the expected number of mutations, we identified 81 predicted rice mutations. Of these, 11 of 14 tested positive in an independent assay. The false positive had lower $F_t$ scores (Supplemental Tables S1 and S2). In wheat, with four genes (five amplicons), we identified 112 predicted mutations, from $F_t = 0.14$ to 8. Some of these wheat genes were independently TILLED by the CJE method (Uauy et al., 2009), so a direct comparison could be made for selected DNA regions. Of the mutations predicted by CAMBa that were directly comparable to the CJE set, 37 were tested and confirmed by other methods, including 20 that had not been identified by CJE. In addition, CJE had discovered two mutations that were not predicted by CAMBa. Therefore, sequencing compared favorably with the established method.

The difference in the wheat and rice data was also evident in the performance of previously published SNP discovery methods. By choosing settings that predict the same total number of mutations, alternative SNP discovery methods could be compared by counting how many known mutations were yielded (V. Missirian, L. Comai, and V. Filkov, unpublished data). In wheat, where CAMBa yielded 36, others yielded 32 (Rigola et al., 2009), 12 (Koboldt et al., 2009), or none (Bansal, 2010; a modified CRISP program yielded 32). In rice, where CAMBa yielded 10, others yielded two (Rigola et al., 2009) or none (Koboldt et al., 2009; Bansal, 2010). In conclusion, discovery could be made robust by optimizing detection algorithms and choosing the optimal probability threshold. The risk of

Figure 5. (Continued.)
plotted versus the base position on the TILLED DNA fragment. In rice, GC > AT are induced most commonly, but not exclusively. G288A and C1319T were tested and confirmed to be heterozygous and homozygous, respectively. Numbers next to mutations are CAMBa $F_t$ scores, where higher values correspond to increased probability of true positive. The arrow points to a singleton signal (orphan). [See online article for color version of this figure.]
investigating changes with lower scores than the optimal threshold could in some cases be counterbalanced by the potential reward of confirming a significant mutant, such as a predicted knockout.

Establishing a Probability Threshold for Mutation Calling by Other Methods

The expected mutation rate may not be known for a given population. An alternative approach is empirical testing. From the TILLING of 32 rice fragments sequenced in a tridimensional pooling scheme (Figs. 4 and 5), a mutant candidate table was obtained and featured 129 mutations ranked according to $F_t$ score. Displayed on a frequency plot (Fig. 5C), each candidate corresponded to a more or less evident set of three outlier frequency points. We tested 79 candidates by Sanger sequencing, establishing the zygosity of discovered changes by the presence of single or double peaks in the chromatograms. Figure 6 illustrates the distribution of the $F_t$ probability score of the candidates. False-positive candidates were uniformly located below a threshold $F_t$ of 5, indicating high reliability for probability scores greater than 5. The ratio of homozygotes to heterozygotes, 23:32, was not different from the 1:2 expectation ($P_{\text{ch}i\text{test}} = 0.18$).

Empirical testing is a laborious approach for threshold determination, and we evaluated other methods. The “plausibility” approach estimates the experimental noise by detecting mutations in pools that do not intersect, comparable to estimating noise in a control experiment. Since these mutations cannot exist unless they occur independently in unrelated individuals, the lowest $F_t$ score to which they are associated can be used as a threshold. Mutations identified by CAMBa in the intersecting tridimensional pools were defined as plausible, and those that were identified in nonintersecting pools were defined as implausible (Fig. 4). The threshold indicated by the implausible candidate with the lowest $F_t$ score was very close to that determined empirically, about 5.

In this case, the estimated mutation rate for rice was known to be about 3.3 per Mb of DNA (Till et al., 2007), and this information can be used to derive an expected mutation number. The 32 TILLED rice gene fragments add up to 42,034 bp. Subtracting the 100 bases at each terminus that escape query (Figs. 3 and 5) leaves 35,634 bp, which queried in 768 individuals correspond to a total of about 27 Mb of searched diploid DNA and yielded an expected number of 89 mutants (3.3 mutations $\times$ 27 Mb). The corresponding threshold probability determined by selecting the 89th candidate of the mutation table is 3, which would have yielded three false-positive and three implausible mutations (Supplemental Table S3). For example, in the inositol kinase-like gene Os09g34300 (Fig. 5C), four mutations were predicted by CAMBa. The least probable candidate had a score of 3.3 and was thus below the plausibility-determined threshold. It was nevertheless confirmed as a heterozygous mutation. The G408A and C1319T changes were supported by strong scores, 9.8 and 20.6, respectively. The C1319T change was also tested and confirmed to be a homozygous mutation. Notably, CAMBa identifies true mutations that were not readily identifiable by the visual outlier pattern (V. Missirian, L. Comai, and V. Filkov, unpublished data). In conclusion, the combination of molecular and bioinformatic procedures resulted in an efficient tool for the discovery of rare mutations in complex pools through amplicon resequencing.

Figure 6. Probability score distribution of mutation candidates in 32 rice TILLING targets of experiment 3, involving tridimensional design. Changes detected by aligning Illumina reads to the TILLED sequence reference were subjected to CAMBa analysis to discover mutation candidates. Higher scores correspond to lower probability of false discovery. Plausible mutations are those that could be true because they are identified by comparing pools that intersect and define unique individuals. Implausible mutations, on the other hand, are unlikely to be true because they were identified in pools that do not intersect. Their identification thus constitutes a measurement of the error noise. A set of plausible mutations were subjected to independent testing by Sanger sequencing, yielding confirmed positives (homozygous or heterozygous) or false positives. Implausible calls and false positives indicate a nearly identical threshold of $F_t = 5$, above which mutations can be detected with high reliability.

[See online article for color version of this figure.]
DISCUSSION

We have addressed the challenge of identifying rare mutations by sequencing pooled DNA amplicons, enabling the high-throughput screening of large populations for targeted lesions. For this, the method combines an efficient template and product pooling systems, multiplexing, sequencing on an Illumina platform, and processing of sequence reads through a bioinformatic pipeline.

The method starts with the pooling of genomic samples, each derived from a single mutagenized individual. Accounting for heterozygosity (Supplemental Fig. S4), the optimal pooling depth was determined to be between 64 and 96 individuals per pool. The amount of input DNA in the pooled templates used for PCR amplification of target genes must be such that each individual is represented a sufficient number of times (we use a minimum of 40 copies per allele). For example, in wheat, one heterozygous mutant allele copy is diluted in 26 pg of DNA per nucleus (for tetraploid wheat, \( n = 13 \) pg). Using 100 ng of initial template for 96 individuals in a DNA pool results in 1.04 ng per individual, which has only 40 copies of a mutant molecule. Reduction of the initial number of molecules increases the fluctuation of copy number across individuals and in extreme cases results in the absence of molecules of the desired allele (e.g. for five copies, there is a 5% probability of having zero copies of the mutant allele). The negative effect of reducing the number of copies below the 40 per allele threshold is supported by our anecdotal observation of orphans (singletons restricted to a single library) in cases where representation was accidentally reduced. Consequently, this reduced sensitivity and specificity (Supplemental Fig. S2; V. Missirian, K. Ngo, B. Watson, T.H. Tai, and L. Comai, unpublished data). We hypothesize that orphans are not real mutations but are DNA polymerase errors that occur very early during amplification and become more frequent when subrepresentative amounts of templates are used. Furthermore, low-concentration templates may further exacerbate the problem by reducing polymerase accuracy (Akbari et al., 2005).

The balancing of DNA input is critical in the creation of pools of template genomic DNA and the pools of amplified TILLING fragments. The balancing of the template genomic DNA is particularly important, as individuals with higher representation can dilute the signal of individuals with lower representation. Balancing steps involve quantification, standardization, and pooling. Experimental variation results from errors in balancing inputs for equal representation and is ameliorated by the establishment of reliable and strict protocols such as those involving automation. Nonetheless, fluctuations are expected from any enrichment method, whether PCR or sequence-capture based.

A general sequencing noise was observed even with stringent quality filtering (Fig. 1), and it increased with low coverage (Fig. 1). We plan our sequencing so that, on average, each allele will yield 30× sequencing coverage (three times the minimal requirement for SNP detection). Nonetheless, coverage varies within target a sequence and across targets despite standardization of input (Figs. 3 and 5). Within a target, some sequences are noisy. Some of these sites, such as position 750 of the VRN3 gene (Fig. 3B), correspond to simple sequences, which are known to be prone to replication-dependent errors (Pearson et al., 2005). Others are GC-rich regions (Fig. 5B), which tend to be underrepresented in Illumina sequences. As a consequence, some mutations are easier to detect than others. It is also possible that certain noisy sites may be explained by varietal contamination or residual heterozygosity in the TILLING line.

Reads produced by sequencing are processed in a bioinformatic pipeline that combines an efficient alignment tool for short reads, BWA, with custom scripts that extract the changes and subject them to the program CAMBa for identification and evaluation of mutation candidates. CAMBa takes our pooling scheme into account and addresses the problems connected with high-variance data sets. For comparison purposes, V. Missirian, L. Comai, and V. Filkov (unpublished data) designed an alternative program, Outlier, which mimics visual inspection of change frequency plots by detecting data points that visually stand out (i.e. so-based outliers) and is efficient with clean data sets. Outlier, however, is less effective with high-variance samples, such as our rice TILLING population. Other approaches to the discovery of mutations in pooled samples (Rigola et al., 2009; Bansal, 2010; Bansal et al., 2010) also proved less effective with our data set. Therefore, our method was best suited to our experimental design and TILLING data sets (V. Missirian, L. Comai, and V. Filkov, unpublished data).

After CAMBa analysis, we obtain a list of candidates with associated mutational effects on the encoded protein and F1 probability scores. Higher F1 values indicate that a call is more probable and, therefore, estimate reliability. To determine the F1 threshold, the value below which the number of false positives is unacceptable, we tried three methods. The first method involves empirically testing each candidate. However, this method is too laborious to be practical. The second method uses the expected number of mutations to set the threshold, but it requires previous knowledge of the mutation frequency in the studied population. The third method, the plausibility test, uses pools that do not overlap to call mutations that are not possible. The F1 score close or corresponding to the highest scoring “unplausible” mutations is then chosen as a threshold. The plausibility test estimates noise and is well suited for our high-throughput system.

Our computational methods can be adapted to any pooling dimension, including single. Unidimensional pooling, however, is inefficient for high-throughput screening. Our tridimensional pooling system yields the mutant individual, while our bidimensional pooling
system yields a pool of eight individuals that must be further deconvolved. For high-throughput screening, the tridimensional scheme may be preferable, as it avoids the cost of deconvolution and reinforces the identification of the mutant through the requirement for three calls. The multiplexing depth achieved through barcoding of the Illumina adapters can be increased with the throughput of the sequencing system to fit the optimal combination of gene numbers and pools.

The target species (we have applied this method with success to the discovery of mutations in tomato, Arabidopsis, and camelina [Camelina sativa; H. Tsai, K.J. Ngo, and L. Comai, unpublished data]) and experimental design for our method can vary considerably. We chose PCR amplification rather than potentially more parallel approaches (Krishnakumar et al., 2008; Choi et al., 2009) because PCR allowed higher accuracy in balancing the input of each target, thus affording a better evaluation of the method’s intrinsic potential. Additionally, PCR amplification and sequence capture differ in their scope: the first is best suitable for population-deep searches for mutations in a few genes, while the second is well suited for broad searches that are carried out in fewer individuals (Nijman et al., 2010).

In conclusion, we have described here a flexible and effective combination of molecular and bioinformatic methods that leverage next-generation sequencing to enable a deep search for mutations in targeted loci. The method is immediately applicable to TILLING populations available in multiple crops. It should also be applicable to any search for rare SNPs or mutations in a complex sample. The method features the following highlights: (1) it is robust and provides reliable mutant identification through the application of multidimensional pooling and a probability threshold; (2) it identifies a mutation with associated base change and effect; (3) it does not rely on labile fluorescent primers or potentially variable endonuclease digestion; (4) it allows flexible choice of pooling methods and species; and (5) it is scalable in scope and experimental combination.

MATERIALS AND METHODS

DNA Isolation

Rice genomic DNA was isolated using the MP Biomedicals FastDNA Kit from 768 M2 rice (Oryza sativa ‘Nipponbare’) plants from a population treated with sodium azide and methyl nitrosourea (Till et al., 2007). We obtained about 1 to 2 μg of DNA from 30 mg of lyophilized tissue. DNA was quantified using the Labworks program (UVP; http://uvp.com), an agarose gel-based image quantification software, and standardized. The DNA used for the pilot experiment was quantified using SYBR Green I dye fluorescence. DNA isolation from wheat (Triticum durum and Triticum aestivum) has been described (Uayu et al., 2009).

DNA Pooling

Pilot Experiment

The genomic DNA of three heterozygous individuals, each carrying a different mutation (A808G in RDR2, C1326T in P1TPK, and T469C in TGT2A), was mixed with the DNA of wild-type individuals to yield pools of 64, 96, 128, and 192 individuals. The TILLING targets for each gene were amplified from each pool in triplicate and processed to make 12 Illumina libraries, each containing the three gene targets for a given dilution point and replicate. The libraries were barcoded, mixed with other samples, and sequenced to yield an average coverage of 5,113 ± 3,672 for each of the 36 sampled sequences (3 TILLED genes × 4 dilutions × 3 replicates; all but two yielded more than 1,000× coverage).

Multiplexing Overview

The number of Illumina GAII lanes needed for a given experiment was calculated from the total number of input bases divided by the minimum expected base sequence yield for a flow cell lane. The input was derived from the product of desired coverage per base, number of targets, average target size, number of individuals, and ploidy number. In addition, an adjustment factor was employed to compensate for unexpected fluctuations in the mixing ratio of target genes and individuals. Our method is based on overlapping genomic DNA pools. Each pool was prepared by combining equal amounts of 48, 64, or 96 individual genomic DNAs. Different pools shared a given set of individuals in such a way that a group of eight was uniquely identified by two intersecting pools in our bidimensional scheme (Fig. 2) and a single individual by three intersecting pools in our tridimensional scheme (Fig. 4). TILLING targets amplified by PCR from the same pool were processed to make an Illumina library. Thus, each library combined a variable number of TILLING targets, depending on the experiment and on the yield of the sequencing platform, which changed over time. The targets combined in a single library could be derived from a single species or from multiple species. For example, in experiment 2 (wheat bidimensional pooling), the expected sequencing output allowed 40 targets of 1.5 kb each to be stacked in a single library. To take advantage of this capacity, targets from multiple species were combined, and these additional amplicons present in our experiments are referred to here for simplicity as originating from “other species” (tomato [Solanum lycopersicum], Arabidopsis [Arabidopsis thaliana], and camelina [Camelina sativa]; these results are not described here). Depending on the experiment, the library was sequenced in a single lane of Illumina GA, or multiple libraries that had been barcoded by adding unique nucleotides to their adapters were combined and sequenced in a single Illumina GA lane. The expanding sequencing capacity of the Illumina platform dictates how many libraries can be sequenced in a single lane. For example, assuming 40 million reads of 40 bases in length (= 1,600,000 kb of sequence) and a desired coverage of 2,500× per base, the number of 1.5-kb targets amplified from a pool of 96 individuals that could be sequenced in a single lane is 1,600,000 kb/(2,500 coverage × 1.5-kb length × 2 alleles × 96 individuals).

For bidimensional pooling (rice experiment 1, bidimensional, and wheat, all experiments), we combined equivalent amounts of DNA from eight individual plants to make one 96-well × 8× pool plate. Genomic DNAs were further pooled by collapsing rows (12 wells × 8 individuals = 96 individuals) and columns (8 wells × 8 individuals = 64 individuals) of this plate to generate 20 template DNA superpools. Thus, this produced a bidimensional array where a single pool of eight individuals was common to two superpools (Fig. 2). After sequencing indicated the presence of a mutation in an eight-individual subpool, further deconvolution (i.e. individual testing of single DNA samples) was carried out to identify the mutant individual.

A tridimensional pooling scheme was used in a second rice experiment to eliminate the deconvolution step, allowing the identification of a single mutant individual. Equivalent amounts of DNA from individual plants were pooled to give 44 libraries according to Figure 4. The DNAs were arrayed on 12 plates of 8× 8 wells. Mixing the 64 individuals on each plate produced the D-pools. The stack of 12 plates was divided by a horizontal plane in two halves of six plates each, then vertical arrays of 6 × 8 = 48 individuals were combined to make the column pools. The stack of 12 plates was figuratively bisected by a vertical plane in two halves, each 12 deep. Vertical arrays of 4 × 12 = 48 individuals, orthogonal to the column pools, were combined to provide the row pools.

PCR

Template and Genomic Units

The number of template DNA molecules representing an individual chromosome in a PCR can be derived by the following formula: template
cycles/allele (TCG) = pg input DNA × (pg/2C nucleus)^3 × pool dilution^2 × 2. For example, rice has a 1-pg 2C genome, and 1 pg of DNA contains two alleles of a single-copy target. The input and TCA for the different experiments were as follows. The pilot experiment used 80 pg of the target heterozygous individual plus wild-type DNA as needed for the desired ratio (1:63, 1:95, 1:127, 1:191). Rice experiment 1 (bidimensional pooling) used 1.5 ng TCA = 16•x for row and 23•x for column. Rice experiment 2 (bidimensional pooling) used 0.5 ng TCA = 5•x in a row and 8•x in a column. Rice experiment 3 (tridimensional pooling) used 5 ng TCA = 104•x for row and column pools and 64•x for the D-pool. Wheat has a 12-pg genome. All experiments used 50 ng of input, TCA = 43•x in the row pool and 65•x in the column pool.

Reaction Conditions
The PCR reagents used were from Takara ExTaq. Primers (Supplemental Table S4) were designed as described previously (Till et al., 2007; Uaay et al., 2009). In some cases, two or more TILLING fragments for the same gene were designed to cover a single gene. For a 30•µL reaction, we mixed 3 µL of 10× ExTaq buffer, 2.4 µL of 2.5 m each deoxyribonucleotide triphosphate mixture (0.2 m in reaction), 0.9 µL of 10 µM forward primer (0.3 µM in reaction), 0.9 µL of 10 µM reverse primer (0.3 µM in reaction), 0.15 µL of 5 units µL^−1 Takara ExTaq HS enzyme (0.75 unit), and template and water as needed. The 10× ExTaq buffer contains MgCl2 at a final concentration of 2.0 mM. The cycling program for rice targets entailed a 2-min 95°C preincubation step followed by eight touchdown cycles at 94°C for 20 s, 73°C for 30 s (increment at −1°C cycle^−1, ramp to 72°C at 0.5°C s^−1), and 72°C for 1 min. Twenty-five more cycles were done at 94°C for 20 s, 65°C for 30 s (ramp 72°C at 0.5°C s^−1), and 72°C for 1 min. Reactions were held at 8°C until retrieval. Programs for wheat targets varied by primer set as needed for genome specificity.

PCR Amplicon Stacking
PCR products were quantified using Labworks (UVP Inc.) and standardized to uniform concentration. Using equivalent amounts of PCR product, targets were pooled together according to their template pool source. We describe in detail three experiments: rice experiment 1, wheat T5, and rice T7. In each of these experiments, the Illumina GA libraries included additional TILLING targets derived from other species not covered here. The complete composition of each experiment was as follows. Experiment 1, 13 rice genes (described here) and five genes from other species (not described here), 20 libraries. It entailed a bidimensional amplification. Experiment 2, 33 rice genes (not described here, except for orphan frequency in Supplemental Fig. S2), five wheat genes (six TILLING targets; described here), two genes from other species (not described here), 20 libraries. Experiment 3, 32 rice TILLING targets from 17 genes (described here), six genes from other species × 768 individuals (not described here), and seven wheat TILLING targets from six genes × 1,536 individuals (data not shown). This experiment involved 44 libraries with identity spikes.

Spiking
To track samples and assess purity, we used marker sequences, called spikes, each consisting of a different PCR-amplified genomic fragment about 1 to 1.5 kb in length. The spike PCR products were quantified using SYBR Green1 and normalized to 1 ng µL^−1 DNA. Each individual spike was added as 2 ng of DNA to the set of target PCR products (1.5 µg total) before any additional step in library construction was performed. As a result, each library was uniquely labeled by a different sequence.

Illumina Library Preparation and Multiplexing
All libraries were fragmented using the Diagenode Bioruptor, which uses ultrasonic waves in a water bath to randomly fragment DNA. Fragmented DNA was processed through an alternative protocol developed by the University of California, Davis, DNA Technologies Core. Ends were repaired using the End-It DNA End Repair Kit from Epicentre (catalog no. ER0720) as specified by the manufacturer. Addition of A base to 3′ ends used DNA Polymerase Klenow Fragment 3′ → 5′ exonuclease (New England Biolabs) (catalog no. M0212a). Adapter oligonucleotides were as provided in the Illumina genomic library kit or were purchased from LifeTech (desalted purity). Ligation employed the LigaFast kit from Promega (catalog no. M8221). Excess adapters can interfere with sequencing, so the optimal adapter concentration was derived by titrating relative to the starting DNA. Size selection followed the Illumina protocol. Enrichment PCR was carried out for 12 to 18 cycles and used Phusion DNA polymerase (New England Biolabs catalog no. F531) and the primers specified by Illumina.

In experiment 3, instead of using the original Illumina adapters, we used four-base barcoded adapters so that four libraries with different barcodes could be pooled into one 80-base single-read sequencing lane. The names and sequences of the adapters were adT2_nnn (5′-PnnnGATCCGAAGACGGGTTTCACGACGCTCTTCCGATCTuuuuT-3′), where nnnn is the four-nucleotide barcode, u is the complementary base, and P represents 5′ phosphorylation. The barcodes used were AGCT, CTAG, GCTA, and TAGC.

The expected coverage per individual was calculated as follows: lane yield/library ratio × input DNA × pooling ratio. For example, for a total lane yield of 20 million reads × 80 bases, four barcodes per lane, 40 TILLING targets of 1.5 kb each, and a 64-individual pool: 1.2 × 10^6 kb/(4 × 60 kb × 64) = 78× per individual, or 37.5× per base.

Mutation Confirmation
To identify the mutant individual from the pool, or to confirm a putative mutant, individual genomic DNA was subjected to PCR amplification with primers specific for the target gene. The presence of the putative change was confirmed through differential restriction enzyme digestion predicted by the pipeline. Sanger sequencing, or CEJ analysis (Uaay et al., 2009). For restriction enzyme, PCR products were incubated with the restriction enzymes according to the manufacturer’s instructions and analyzed by gel electrophoresis on agarose gels. For sequencing, PCR products were purified using the QiAquick PCR Purification Kit (Qiagen) according to the manufacturer’s instructions and subsequently quantified by spectrophotometry (Nanodrop). Samples were sequenced using the BigDye Terminator sequencing kit and the ABI 3730 Capillary Electrophoresis Genetic Analyzer (Applied Biosystems).

Bioinformatics
For each TILLED fragment, its sequence, primer to primer, the genomic sequence of the target gene from start to stop codon, and the inferred or known cDNA sequence from start to stop codon were provided to the pipeline. Illumina generated-sequence was processed through MAQ (earlier) and BWA (later). Initial analyses used a Perl script that parsed the “pile-up table” output of MAQ. Ad hoc analysis employed the JMP software by SAS. Later analysis involved a pipeline in which a Python program processed the raw Illumina GA reads, converted them to FASTQ format, sorted them by barcode, trimmed them to remove adapter sequences, filtered them setting the base call quality cutoff to be 10 below the mean of the quality distribution of the reference base calls, ran BWA to align the reads to the reference, extracted base calls and qualities from the pile-up table, ran CAMBa 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 for mutations in pools that do not overlap) was run separately by returning the CAMBa program while allowing any library to be considered overlapping with any other one. The resulting mutations were then categorized as plausible or implausible depending on the existence of a real overlap between the pools that shared the putative mutation. GC content was measured and graphed through a custom Python program that used the Enthought package distribution of the Matplotlib and NumPy libraries. Google Sketchup was used to draw tridimensional pooling models. SAS-JMP and Adobe Illustrator were used for custom graphing.

The sequence reads from this study, processed input data, and related software are available at http://comailab.genomecenter.ucsd.edu/index.php/TILLING_by_Sequencing.

Supplemental Data
The following materials are available in the online version of this article.

Supplemental Figure S1. Sequencing coverage in experiments 1 and 2.
Supplemental Figure S2. Variable occurrence of orphans.
Supplemental Figure S3. Discovery of mutations versus probability cutoff.
Supplemental Figure S4. Relationship between zygosity and Illumina-measured frequency.

Supplemental Table S1. Mutations discovered in wheat using bidimensional pooling.

Supplemental Table S2. Mutations discovered in rice using bidimensional pooling.

Supplemental Table S2. Mutations discovered in rice using tridimensional pooling.

Supplemental Table S4. Gene and primer list.

ACKNOWLEDGMENTS

We thank Charlie Nicolet, Vanessa Rashbrooks, and Heather Witt at the University of California, Davis, Genome Center DNA Technology Core for their assistance with Illumina GA sequencing and Smit Shah and Wasinee Pongprayoon (University of California, Davis) for technical help.

Received November 20, 2010; accepted April 28, 2011; published April 29, 2011.

LITERATURE CITED


