Statistical Mutation Calling from Sequenced Overlapping DNA Pools in TILLING Experiments

Victor Missirian[1], Luca Comai[2], Vladimir Filkov*[1]

[1] Department of Computer Science, UC Davis, 1 Shields Ave., Davis, CA 95616, USA
[2] Department of Plant Biology and Genome Center, 1 Shields Ave., Davis, CA 95616, USA

Email: filkov@cs.ucdavis.edu;

*Corresponding author

Abstract

Background: TILLING (Targeting induced local lesions IN genomes) is an efficient reverse genetics approach for detecting induced mutations in pools of individuals. Combined with the high-throughput of next-generation sequencing technologies, and the resolving power of overlapping pool design, TILLING provides an efficient and economical platform for functional genomics across thousands of organisms.

Results: We propose a probabilistic method for calling TILLING-induced mutations, and their carriers, from high throughput sequencing data of overlapping population pools, where each individual occurs in two pools. We assign a probability score to each sequence position by applying Bayes' Theorem to a simplified binomial model of sequencing error and expected mutations, taking into account the coverage level. We test the performance of our method on variable quality, high-throughput sequences from wheat and rice mutagenized populations.

Conclusions: We show that our method effectively discovers mutations in large populations with sensitivity of 92.5% and specificity of 99.8%. It also outperforms existing SNP detection methods in detecting real mutations, especially at higher levels of coverage variability across sequenced pools, and in lower quality short reads sequence data. The implementation of our method is available from: http://www.cs.ucdavis.edu/~filkov/CAMBa/.

Background

TILLING (Targeting Induced Local Lesions IN Genomes) [1] is a reverse genetics approach to detect effects of globally induced mutations in a population and identify the individuals that have mutations in genes of interest. Mutations discovered through TILLING allow the functional characterization of genes known only by their sequence. Furthermore, because TILLING is applicable to any species that can be mutagenized, it can be used to knock out undesirable characters in crops [2]. A refinement of previous approaches, TILLING-by-Sequencing (Comai et al., unpublished) follows up the mutagenesis with deep sequencing of individuals or populations of interest. Because of the high throughput of current sequencing technologies, deep sequencing to hundred and thousand fold coverage is possible [3]. This allows unprecedented precision when identifying the induced mutations.

The TILLING-by-sequencing setup in one of our labs (Comai) uses the mutagen ethyl methanesulphonate (EMS) or the combination of sodium azide and methyl-nitrosourea (Az-MNU) to induce mutations in a population of 1500-6000 individuals. The mutations induced will be heterozygous in 2/3 of the cases and homozygous in the rest. Units of 768 individuals arrayed in a 96 well-plate, 8 individuals per well, are then screened. The row and column samples are pooled to yield 8 row- and 12 column-pools (for a total of 20 pools). Thus, the row and column pools overlap in their DNA content in such a way that each individuals DNA is present in exactly 2 pools. This arrangement allows for the identification of both the mutated positions and the individuals that carry them. We call this setup *bi-dimensional pooling*, and illustrate it in Fig. 1. DNA from each of the 20 pools is PCR-amplified with primers designed to amplify 1-1.5 DNA segments from up to 40 genes of interest, and subsequently sequenced using Illumina GA machines. The reads are then mapped onto reference genomes.

Given a stretch of DNA of interest from a reference genome and a complete set of deep sequenced, bi-dimensionally overlapping pools (20 pools in our case) we want to identify the positions with mutations along the DNA and their individual carriers. The computational problem, then, is to identify the position and the row- and column-pool for each mutation.

Any solution to this problem would of course focus on identifying significant differences between the reference genome and the sequenced DNA. The problem is complicated by the non-independent pools of the experimental setup, the infrequency of the mutation's occurrences with respect to the size of the population under study, and also by variability, or non-uniformity in the sequencing coverage, which is not uncommon for 2nd generation sequencing technologies [4]. For example, in our experimental setting, a mutation in a single individual is expected to cause a higher base change frequency in one row and one column library, and many mutations can be recognized in this way, by visual detection of outliers (we call this the *Outlier approach*). In Fig. 2, we show the base change frequency for each library at three positions with confirmed mutations from mutagenized wheat and rice. From left to right, there is apparent increased difficulty in identifying a mutation. The accuracy of calls made by the Outlier approach depends on the coverage, or number of nucleotide calls per position per library. Given a fixed probability of base change due to error, at high coverage levels, libraries with real mutations will usually stand out clearly from the noise. As coverage drops, a larger range of base change frequencies may reasonably occur due to chance in the absence of a real mutation, thus increasing the number of false positives. The Outlier approach cannot distinguish these cases because it does not take coverage levels into account, so a single gene that has low coverage on a few libraries can cause a high overall false positive rate.

Here we propose a new method, *Coverage Aware Mutation calling using Bayesian analysis*, *CAMBa*, (read like the dance) which directly considers the pooled setup and coverage levels when calculating mutation and noise probabilities. Using data from two TILLING experiments, one on rice and one on tetraploid wheat, we validate CAMBa's efficacy in identifying mutations, and demonstrate that it does at least as well as other mutation calling methods, and that it outperforms the other methods on the rice experiment which has a lower quality and higher variance in coverage across libraries. We also test and confirm the hypothesis that CAMBa is insensitive to lower data quality and variable sequencing coverage across libraries dur to the overlapping pools experimental design.

Related Work

Rigola et al. [5] use a Poisson distribution based approach to identify mutations and natural variations in individuals using bi- and three-dimensional pooling schemes coupled with high-throughput sequencing. We compare their method to CAMBa in the Results section.

A variety of approaches exist for calling SNPs from non-overlapping pooled samples, e.g. VarScan [6], CRISP [7], SNPseeker [8], the MAQ alignment tool [9], and others; and non-pooled samples, e.g. POLYBAYES [10], PolyScan [11], the method by Stephens et al. [12], and others. Our approach is specifically geared to working on pooled experiments with overlap between the DNA pools, i.e. DNA from the same individual is present in two pools. That is not the case for these other approaches, so we could not compare them directly to CAMBa. Moreover, these other approaches identify mutations but not the individuals in the populations that carry them. We modified some of them (VarScan and CRISP) in order to compare them to CAMBa and report those studies in the Results section.

Overlapping pool designs for high-throughput resequencing have been recently proposed by Prabhu and Pe'er [13], where they focus on optimizing overlaps to increase design efficiency (as compared to optimal), lower necessary sequencing coverage, decrease false positives and false negatives, and identify mutation carrier with lower ambiguity. They do not provide software for testing their designs and it is not immediately clear that their designs could easily fit into standard wet lab protocols (e.g. with respect to standard well plates, etc.) Our overlapping pooling scheme can be evaluated in their theoretical framework, and in terms of the "code efficiency" it is 50% worse than the theoretically optimal binary design (although it is not clear if that optimum is achievable in practice).

Methods

The experimental TILLING setup encompasses a 2D well-plate, with i_{well} individuals pooled per well and i_l individuals in each library l, the total number of wells n_w in the experiment, the reference base r at the current position, the probabilities p_c and p_{nc} with which the mutagen, EMS or MNU, will induce a specific canonical ($G \rightarrow A$ or $C \rightarrow T$) or non-canonical base change at a given position in a single individual, and the fraction of induced mutations t_z for each zygosity z.

The input data, D, is comprised of a set, L, of row and column libraries of short reads covering the sequences of interest. The reads for each library are aligned to their reference sequences using the MAQ (Mapping and Assembly with Quality) alignment tool [9], associating each position in each tilling sequence with a set of nucleotide calls: either the reference nucleotide base, r or a base change $r \rightarrow m$, $m \neq r$. In the alignment, or pileup, of reads, for each position we count the total number of reads, i.e. coverage, in a given library, denoted by n_l, and separately the number of reads that have base b at that position in that library, k_{lb} (so $n_l = k_{lA} + k_{lT} + k_{lC} + k_{lG}$).

To find the carriers and the mutations, given the data, D, and the well-plate experimental setup for each sequence position we model the posterior probabilities of each possible mutation in each well. We assume that at most one individual will have a mutation at any given sequence position. [1] Thus, at most one well can have a mutation with respect to one specific base change. We denote these possibilities, or (configurations), as $c_{w,m}$, where w is the well, and m is the base change from reference. (In our setup, we distinguish 288 mutation possibilities since we have 96 wells and 3 possible base changes different than the reference). The probabilities corresponding to the configurations are $p(c_{w,m}|D)$. We call a mutation at a given position if the probability of at least one mutant configurations $c_{w,m}$ exceeds a predefined threshold indicating that well w contains an individual with base change m at the current position. If more than one $c_{w,m}$ pass the threshold, then the one with highest probability is chosen. The threshold is determined based on the expected number of mutations in an experiment, as described in Sec. .

In the following we calculate the probabilities $p(c_{w,m}|D)$. Since the experimental procedure makes the expected number of heterozygous mutations equal to twice the number of homozygous mutations, we further distinguish configurations by zygosity, and use $c_{w,m,z}$ to model the probabilities of heterozygous, $z = het$, and homozygous, $z = hom$, mutations separately, and $p(c_{w,m}|D) = p(c_{w,m,het}|D) + p(c_{w,m,hom}|D)$. Then, from Bayes' Theorem we get:

$$p(c_{w,m,z} \mid D) = \frac{p(c_{w,m,z})p(D \mid c_{w,m,z})}{\sum_{c' \in C} p(c')p(D \mid c')},$$

where C is the set of all possible configurations $c_{w,m,z}$ at the given position ($288 * 2 = 576$ in our setup). Since we exclude all configurations with more than one mutant individual for the current position, the sum of the prior probabilities $p(c' \mid s)$ do not add up to 1, but normalizing does not affect the result.

Next, we calculate both the prior and conditional likelihood probabilities. We model the prior probability of a base change $r \rightarrow m$ at the current position in exactly i out of the i_{well} individuals in a given well as:

[1] This is supported by evidence from a previous TILLING experiment in tetraploid wheat using the mutagen EMS, where Slade et al. [2] identified 50 positions for which at least one of the 768 individuals contained a mutation but only 3 for which there was a mutation in two individuals. In rice, which has a significantly lower expected mutagenesis rate at each possible reference base [2, 14, 15], we expect an even smaller percentage of the positions for which there is a mutation in one individual to have a mutation in more than one individual.

$$p_{im} = \begin{cases} B(i \mid i_{well}, p_c) & \text{if } r \to m \text{ is canonical} \\ B(i \mid i_{well}, p_{nc}) & \text{if } r \to m \text{ is non-canonical} \end{cases}$$

where $B(i \mid i_{well}, p_c)$ is the Binomial distribution, i.e. the probability of having i successes out of i_{well} trials given an individual success probability of p_c.

A given well has prior probability $p_{tm} = p_{1m}$ of being a mutant well for base change m and $p_{fm} = p_{0m}$ of being a non-mutant well for base change m. The prior probability of a given configuration $c_{w,m,z}$ is the product of the prior probabilities of each well w on each base change m, multiplied by the fraction of induced mutations with the given zygosity, if applicable. Thus, since there are three possible base changes at the current position, m, m', and m'', $p(c_{w,m,z}) = t_z p_{tm} p_{fm}^{n_w-1} p_{fm'}^{n_w} p_{fm''}^{n_w}$. Note that these probabilities are the same for all wells (i.e. they don't depend on w).

To calculate the likelihood $p(D \mid c_{w,m,z})$, we use the fact that each well w is an intersection of a row and a column library. First, we assume that all base change counts are conditionally independent given n_l and $c_{w,m,z}$, and so we simplify the likelihood to the individual libraries:

$$p(D \mid c_{w,m,z}) = \prod_{\substack{m \in \{A,T,C,G\} \\ m \neq r \\ l \in L}} p(k_{lm} \mid n_l, c_{w,m,z}).$$

Each term in the product on the right is the conditional probability of observing k_{lm} reads having base m at that position, given n_l coverage for that library and configuration $c_{w,m,z}$. We model this with the Binomial distribution, as a function of probability r_{lmc}, the expected rate of reading base change m at the current position in library l, given configuration $c_{w,m,z}$:

$$p(k_{lm} \mid n_l, c_{w,m,z}) = B(k_{lm} \mid n_l, r_{lmc}).$$

We model r_{lmc} in terms of m_{lc}, the fraction of mutant alleles in library l under configuration c, and $r_{r \to m}$, the rate at which a base r is read as m in well w, as: $r_{lmc} = (1 - m_{lc}) r_{r \to m,c} + m_{lc}$. We compute m_{lc} from the number of individuals i_l in library l and the zygosity z of the candidate mutation. Finally, we estimate $r_{r \to m}$:

$$r_{r \to m,c} = \sum_{l \in L_c} k_{lm} / \sum_{l \in L_c} n_l.$$

where L_c is the set of all libraries excluding the two (one row and one column library) which intersect at well w, and have mutation m at the given position.

Pre-processing

For each library, we compute a low quality cutoff for base calls to be one standard deviation below the mean quality of the reference base calls; we throw out all base calls below this quality threshold. After filtering out low quality base calls, we do not search for candidate mutations at the current position if the mean coverage over all but two libraries is less than 10,000, to avoid inaccurate estimates of $r_{r \to m}$.

The orientation bias of a specific base is the fraction of base reads for that base coming from reads that map to the forward rather than the reverse strand of the TILLING sequence. If the reference base orientation bias for a given library at the current position is different from the orientation bias of base change m with pvalue < 0.01, then we set $p(D \mid c_{w,m,z}) = 0$ to exclude each configuration c for which a well represented in that library is a mutant well for base change m. We also set $p(D \mid c_{w,m,z}) = 0$ for these configurations if the reference base orientation bias for the given library is greater than 10 or less than 0.1, since a strong reference base orientation bias can make it difficult to detect a significant difference between the orientation biases of the reference base and a given base change. In addition, if a given library has more base reads for the candidate base change than for the reference base at the current position, then we set $p(D \mid c_{w,m,z}) = 0$ for each configuration where a well in that library is a mutant well for any base change.

Number of Predictions

We construct our initial estimate of the number of real mutations in a given experiment by adding up the probabilities of each possible induced base change at each position across all TILLING sequences in all individuals, where the probability of a given canonical or non-canonical base change is determined from CEL-1 screening of an experiment on the same organism using the same mutagen [14, 15], as described in the data section below. By this method, we estimate 47 real mutations in Rice and 69 real mutations in wheat. Since CEL-1 has a significant false negative rate, we correct our initial estimate using additional validation information from the wheat experiment. When an older version of our approach was run on the wheat experiment, 8 of the 10 predictions ranked 86 to 95 were tested and all 8 were confirmed. We drop below this ranking to give the semi-conservative estimate of 107 real mutations. We divide 107 by 69 to get a candidate scaling factor of 1.55. We predict the number of real mutations for a given experiment to be 1.55 times our initial estimate of the number of real mutations from CEL-1 screening. The predicted number of mutations is 107, by definition, for wheat, and it is 75 for rice. This determines our threshold.

The above approach for determining the appropriate threshold yields very good bounds for our data and can be applied whenever previous CEL-1 screening experiments have been done. In the absence of such prior experiments, one can apply the following method, although the results may include higher false positive rates. The false positive rate at a given number of predictions can be estimated by running CAMBa with input a scaled down bi-dimensional arrangement using only the row pools. E.g., the row pools in the new scheme could be half of the actual row pools, and the new column pools could be the other half of the original row pools. Since we expect few or no instances where the same mutation occurs in two independent row libraries, the number of row/row calls serves as an upper bound on the number of false positives among the row/column candidates. Similarly, we could scale down the original arrangement using the original column pools instead of the row pools. We scale up by the ratio of the number of row/column pools versus the number of row/row pools, and choose the largest number of candidates for which the estimated false positive rate is nearest to our goal threshold. As an illustration of this method, we split the wheat data set 12 column pools into two groups of 6 pools each, and ran CAMBa on this new bi-dimensional pooled data of 6 rows and 6 columns. At a false positive rate of 0.05, this method yields a threshold for CAMBa of 105 mutations.

We note that although CAMBa yields posterior probabilities for each of thousands or tens of thousands of positions, we never use hypothesis testing to determine the threshold in either of the two approaches above, and thus we need not correct for multiple hypothesis testing.

Methods for Comparison

We compare the performance of CAMBa to those of a number of methods below. Only the Outlier and the Poisson outlier methods predict both the mutated positions and the individual carriers, whereas the others only predict the mutation positions. Hence, either they or CAMBa have been modified to allow for the comparison. In each case we specify the modification undertaken.

The *Outlier method* uses the same preprocessing techniques as CAMBa, and is inspired by simple visual identification of a row and column library pair that stand out from the rest in terms of the frequency of a given base change. When considering a given position in a TILLING sequence, if at least one well has a score greater than a fixed threshold t on some base change, then we predict a mutation for the base change and well combination with the highest score. For a given well w and base change m, we find the z-score of the $r \to m$ base change frequency for both the associated row and column library with respect to the distribution of the $r \to m$ base change frequencies for the remaining libraries, and we set the score of well w on base change m to be the lower of these two z-scores. We add 0.0001 to the sample standard deviation to avoid division by zero.

The *Poisson outlier method* is described in a TILLING-by-sequencing pipeline by Rigola et al. [5]. This method consider only G to A and C to T base changes for the purposes of mutation detection. Since MNU can induce any type of base change in rice, we modified the Poisson outlier method to search for all possible base changes when detecting mutations in rice. Rather than following the procedure for detecting natural variation, which considers all base changes as a whole, we tested for each base change individually, to reflect the assumption that an individual will have at most one mutation at a given position.

VarScan is a SNP identification method in individual or pools of massively parallel sequence data by Koboldt et al [6]. It identifies variants based on read counts, base quality, and allele frequency. We used VarScan with varying p-values (its only parameter setting). VarScan does not take into account overlap in pools and it does not identify the individuals carrying the mutations. To compare it to CAMBa in a bi-dimensional setting, we ran VarScan separately on the row pools and again on the column pools. We took the maximum of the row and column p-values as the combined p-value for a mutation present in both the rows and the columns. The results were mostly unchanged when we ran VarScan on the 20 pools together.

CRISP [7] is a statistical method for variant detection in pooled DNA samples, shown to dominate a number of other methods in direct comparison of SNP detection [7]. Like VarScan, CRISP does not identify the mutation carriers, so when comparing CAMBa and CRISP by the number of candidates that overlap with the set of confirmed mutations, we considered only the position and base change of each candidate. Since it does not account for pool overlap, to compare it to CAMBa we ran CRISP under three different scenarios: on all 20 libraries, separately on the 8 row-pools, and separately on the 12 column-pools. We only include results for CRISP on its default parameters, due to the non-intuitive performance of CRISP when running it with relaxed parameters. (As CRISP has two user-modifiable parameters, we performed two-dimensional search to determine the threshold combination that gives us the set of top candidates that has the highest overlap with the set of confirmed mutations. Unfortunately, even after this optimization step, the performance was weaker than when CRISP was run with the default parameters. Modifying CRISP more substantially was beyond the scope of this project and certainly beyond expectations of any practicing biologist.)

We also attempted comparisons with MAQ [9] but we could not get any mutation predictions on our data sets with their default settings.

Results and Discussion

Using data from two TILLING-by-sequencing experiments we analyze the performance of *CAMBa* and compare it to those of other approaches. We investigate the effect of sequencing quality, sequencing coverage variability, and the overlapping pool design on the fidelity of our and the other methods in resolving mutations from the data.

Two TILLING Experiments

Using the setup described earlier (768 individuals arrayed evenly into a 96-well plate, 8 row and 12 column pools sequenced) a total of 13 rice genes (avg. TILLING seq. length = 1393 bp) and 5 wheat genes (avg. TILLING seq. length = 934 bp) of interest were sequenced using Illumina GA machines to look for mutations in a population of 768 individuals. The reads on the average were of length 35 bp for the rice and 40 bp for the wheat data. There was a significant difference in the read quality between the two: the rice had an average Phred quality score of 13 and the wheat of 31. There was also a larger variance in coverage between individual libraries in the rice data set than in the wheat data set. Also, on average, the coverage was 140× per individual in rice and 270× in wheat. The full data and experiments will be detailed elsewhere (Comai et al., unpublished). The differences in quality, coverage, etc. between these two data sets make them very good case studies for our method.

To evaluate the performance of the mutation calling algorithms, we used two sets of mutations, one set for wheat and one for rice, which have been previously confirmed using an independent method (PCR amplification followed by sequencing) [16]. In total we had 39 confirmed mutations from the wheat experiment and 11 from the rice experiment. [2] We note that the confirmed mutations are a fraction of the total expected mutations in these data sets. Ideally, all predicted mutations should be tested, but practical resource constrains dictate limits on the validation.

Using prior TILLING experiments we determine the probabilities of a canonical mutation, p_c, non-canonical mutation, p_{nc}. We assume position independent values for p_c and p_{nc}, as indicated by prior experiments [2,14,15]. We estimate p_c and p_{nc} from i_c and i_{nc}, the induced mutation rates for canonical and non-canonical mutations, and $p_{til}(b)$, the fraction of TILLING reference sequences with base b, for each b (in prior corresponding experiments): $p_c = i_c/p_{til\,G,C}$ and $p_{nc} = i_{nc}/p_{nc}\big(2p_{til\,G,C} + 3p_{til\,A,T}\big)$. We compute i_c and i_{nc} using previously described methods [14, 15]. Thus, we get for rice, $p_c = 5.6\times10^{-6}$ and $p_{nc} = 4.93\times10^{-7}$, and for wheat, $p_c = 3.88\times10^{-6}$ and $p_{nc} = 0$. The fraction of heterozygous mutations is $t_{het} = 2/3$, and homozygous mutations is $t_{hom} = 1/3$.

[2] These sets of confirmed mutations come from predictions using prior iterations of our approach, and prior experimental approaches, using CEL-1, on these data sets and all consequently confirmed with PCR amplification.

Validation

Due to the apparent bimodality of the calculated posterior probabilities, and their clustering around the values of 0 and 1, we apply the following function to transform t, the posterior probabilities returned by CAMBa:

$$F(t) = \begin{cases} -\big(log_{10}(1-t) - log_{10}(0.5)\big) & \text{if } t \geq 0.5 \\ log_{10}(t) - log_{10}(0.5) & \text{if } t < 0.5 \end{cases}$$

$F(t)$ is effectively the log posterior probability. For both the rice and wheat TILLING-by-sequencing experiment, the predictions of CAMBa and the other methods are compared against the corresponding set of confirmed mutations.

To investigate the relationship between the number of predictions and the rates of true positives and false negatives, we ran the methods with different parameter settings. This was especially useful in the case of VarScan and the Poisson method, which yielded unreasonable numbers of predictions at their default settings (e.g., the Poisson method at a pvalue of 0.01, recommended by its authors, yielded thousands of predictions). So, for a range of values of CAMBa's $F(t)$, we generated a set of predictions for CAMBa and then set the threshold for the other methods to return the same number of predictions. Given an objective number of actual mutations in a data set, this approach makes it easy to estimate the false negatives and false positives directly. As a good estimate of the actual mutations for each data set, we use the above determined estimates (based on external knowledge) of 107 mutations for wheat and 75 for rice.

In Tables 1 and 2 we show the overlap between the predictions of each method and the tested and confirmed sets of mutation candidates, as well as the overlap between the predictions of the two methods. On the rice data set, CAMBa dominates the other methods clearly. Increasing the threshold causes the number of predictions for rice to increase at a greater rate than for wheat, for all three methods. This is likely due to the lower quality of the rice data, thus higher amount of noise, as well as the variable sequencing coverage across pools, as we show below. Interestingly, neither VarScan nor CRISP generated any predictions on this data. At the threshold as determined above of 75 mutations in rice, CAMBa predicted correctly 10 out of the 11 confirmed mutations. These results are strong evidence that given lower quality data CAMBa can utilize the overlapping pools experimental setup to its advantage better than the other methods could.

On the wheat data set, we see greater overlap between the predictions of CAMBa and the other methods. This is consistent with the fact that the wheat data set is of higher quality, and thus mutations are easier to identify. At the threshold as determined above of 107 mutations in wheat, CAMBa predicted correctly 36 out of the 39 confirmed mutations. (Looking closer in the sequence data for the 3 false negatives, we found out that one of them is due to strand-specific bias which resulted in wrong frequencies of base changes, and another was due to under-sequencing; the third showed up in the table at a much higher threshold.)

It is straight forward to estimate the false positive rate and false negative rate using these tables. To do so, first we assume that the confirmed mutations have been randomly chosen from the set of all mutations. Thus, we scale by 107/39 the true positives, false positives, and false negatives. For CAMBa, in wheat this gives $TP = 99$, $FP = 8$, and $FN = 8$, for a false negative rate of $7.5{\times}10^{-2}$ i.e. sensitivity of 92.5%, and a false positive rate of $2{\times}10^{-3}$ i.e. specificity of 99.8%. Similarly, in rice, CAMBa has sensitivity of 93.4% and specificity of 99.96%. Interestingly, even the much lower read data quality of the rice data (as given by the Phred quality scores above) does not seem to affect CAMBa's performance.

By comparison, on wheat, CRISP with default parameters yielded (in the best of three scenarios, see above, when run on all 20 pools) 118 predictions of which 32 overlapped with the 39 confirmed mutations. This works out to sensitivity of 82% and specificity of 99.6% in wheat (CRISP yielded no predictions on the rice data set).

To further evaluate the reason for CAMBa's advantage over CRISP, we also ran CAMBa in a "naive mode" by removing the information that pools overlap (i.e. one dimensional pooling). We did this by giving CAMBa a bi-dimensional arrangement of only 1 row and 12 column-pools (i.e. we used only the column pools). CAMBa predicted 103 mutations of which 30 overlapped with the confirmed 39. When repeated with the 8 row-pools only there were 99 predictions of which 26 overlapped the 39 confirmed ones. We could not run CAMBa in "naive mode" on all 20 libraries because they were not independent. CRISP, on the other hand, predicted 147 mutations, of which 30 overlapped with the 39 confirmed ones when ran on the row-pools only, and predicted 104, including 25 of the 39 confirmed, when ran on the column-pools only. CRISP appears to have an advantage if we compare its run on all 20 pools versus CAMBa on either the row- or column-pools only, but such a comparison is not objective and the advantage is in part due to CAMBa working on less data. Neither method shows consistend advantage when compared on the same number of pools (8 or 12) in the one dimensional case. Overall, this is strong evidence that using an overlapping experimental pool design imparts better advantage. We could not compare CAMBa to CRISP on the data sets used in the CRISP paper [7] as the sequence data could not be made available to us.

Coverage Variability over Libraries

CAMBa shows small performance advantage over the other methods on the wheat experiment, which has consistently high coverage levels across all libraries and between genes. In contrast, CAMBa has a very clear advantage in the rice experiment, and here we investigate the difference.

In Fig. 3 we show the coverage variance across libraries for all genes in both experiments, wheat in gray, and rice in black. It is apparent that the black lines are overall higher on the plot than the gray lines, especially the line for HLP1. To test the hypothesis that coverage variability gives CAMBa an advantage over the other methods, we performed two computational studies. In the first one, we modify the rice data set to exclude the TILLING sequences for gene HLP1 which has both the lowest mean coverage and the highest coverage variability across libraries in this experiment. On this modified data set, all methods perform comparably, as shown in Table 3.

In the second study, we gradually increase the coverage level variance across libraries in the wheat experiment by selectively discarding base reads. We set the new coverage level of each library on a given gene to be the coverage ratio of that library to the library with the highest coverage on that gene, raised to the scaling factor s, multiplied by the coverage of the highest coverage library for that gene. To reach the desired coverage level for a given gene on a given library, we discard each base call with fixed probability. CAMBa gains a solid advantage over the Outlier method between $s = 2$ and $s = 3$. At these levels the coverage still is pretty high overall. The predictions with added coverage variance at $s = 3$ are shown in Table 4. This level of coverage variance would correspond to a level of 0.7 in Fig. 3. We conclude that the likeliest reason for the advantage we see in CAMBa's performance is due to its insensitivity to coverage variability in the data, an effect of both its explicit use of coverage in the model, and the greater signal sensitivity imparted by the overlapping pool design. While still possible that, on the average, low coverage may be the culprit, the second study above makes that unlikely.

We note that while the rice experiment uses the mutagen MNU and the wheat experiment uses EMS, the choice of mutagen does not seem to have a significant effect on either the overall mutagenesis rate or the proportion of canonical versus non-canonical mutations [14].

Conclusions

We demonstrated that our probabilistic method, which explicitly takes into account the bi-dimensional, overlapping pools experimental setup, and sequence coverage at each position for each library, can effectively discover rare mutations in large populations, as well as the individuals that carry them. It also has a performance advantage over other methods for detecting mutations from high-throughput sequencing of a TILLING population when there is significant coverage variability over libraries or lower quality data. More generally, it follows from our experiments that accounting for sequencing coverage variability can improve mutation detection in overlapping DNA pools. It would be interesting to work out the relationship between coverage depth and pool size. Likewise, we demonstrated that an overlapping pooling scheme, beyond offering carrier identification, also yields increased sensitivity of mutation detection when the data is less than ideal. This work implies a possible association between the amount of pool overlap (i.e. pool design code efficiency, or dimensionality of an experimental setup) and detection sensitivity, which deserves closer attention, especially for experiments on larger populations.

There are several directions in which our tool can be improved. We can add to our model an explicit account for position dependence of the mutations. Also, we can extend the model to allow multiple mutations at any given position (because of prior estimates of such events, we suspect that those improvements together will yield less than 10% improvement). We plan to continue using and improving CAMBa in Comai's TILLING laboratory. As other technological issues like higher throughput and sequence tagging get introduced into our pipeline, the issues of coverage sufficiency and higher multidimensional TILLING will be addressed.

We note that properly accounting for coverage variability may improve results in other genomics problems benefiting from 2nd generation sequencing, like sequence mapping, genome assembly, and motif finding.

Authors contributions

The TILLING-by-Sequencing technology was developed by LC. VM, LC,and VF concieved the computational framework, VM and VF designed the studies. VM developed the software and run the experiments. VM and VF wrote the paper, LC helped in editing.

Acknowledgments

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

References

1. McCallum C, et al: **Targeting Induced Local Lesions IN Genomes (TILLING) for Plant Functional Genomics**. *Plant Physiology* 2000, **123**:439–42.
2. Slade A, et al: **A reverse genetic, nontransgenic approach to wheat crop improvement by TILLING**. *Nature Biotechnology* 2004, **23**:75–81.
3. Shendure J, Ji H: **Next-generation DNA sequencing**. *Nature Biotechnology* 2008, **26**:1135–45.
4. Harismendy O, Frazer KA: **Method for improving sequence coverage uniformity**. *BioTechniques***46**:229–31.
5. Rigola D, et al: **High-Throughput Detection of Induced Mutations and Natural Variation Using KeyPointTM Technology**. *PLOS One* 2009, **4**(3):e4761.
6. Koboldt DC, et al: **VarScan: variant detection in massively parallel sequencing of individual and pooled samples**. *Bioinformatics* 2009, **25**:2283–85.
7. Bansal V: **A statistical method for the detection of variants from next-generation resequencing of DNA pools**. *Bioinformatics* 2010, **26**:i318–24.
8. Druley TE, et al: **Quantification of rare allelic variants from pooled genomic DNA**. *Nature Methods* 2009, **6**:263–65.
9. Li Hea: **Mapping short DNA sequencing reads and calling variants using mapping quality scores**. *Genome Research* 2008, **18**:1851–1858.
10. Marth G, et al: **A general approach to single-nucleotide polymorphism discovery**. *Nature Genetics* 1999, **23**:452.
11. Chen K, et al: **PolyScan: An automatic indel and SNP detection approach to the analysis of human resequencing data**. *Genome Research* 2007, **17**:659–66.
12. Stephens M, et al: **Automating sequence-based detection and genotyping of SNPs from diploid samples**. *Nature Genetics* 2006, **38**:375–81.
13. Prabhu S, Pe'er I: **Overlapping pools for high-throughput targeted resequencing**. *Genome Research* 2009, **19**:1254–61.
14. Till B, et al: **Discovery of chemically induced mutations in rice by TILLING**. *BMC Plant Biology* 2007, **7**:19.
15. Uauy C, et al: **A modified TILLING approach to detect induced mutations in tetraploid and hexaploid wheat**. *BMC Plant Biology* 2009, **9**:1.
16. Porreca GJ, et al: **Multiplex amplification of large sets of human exons**. *Nature Methods* 2007, **4**:907–9.

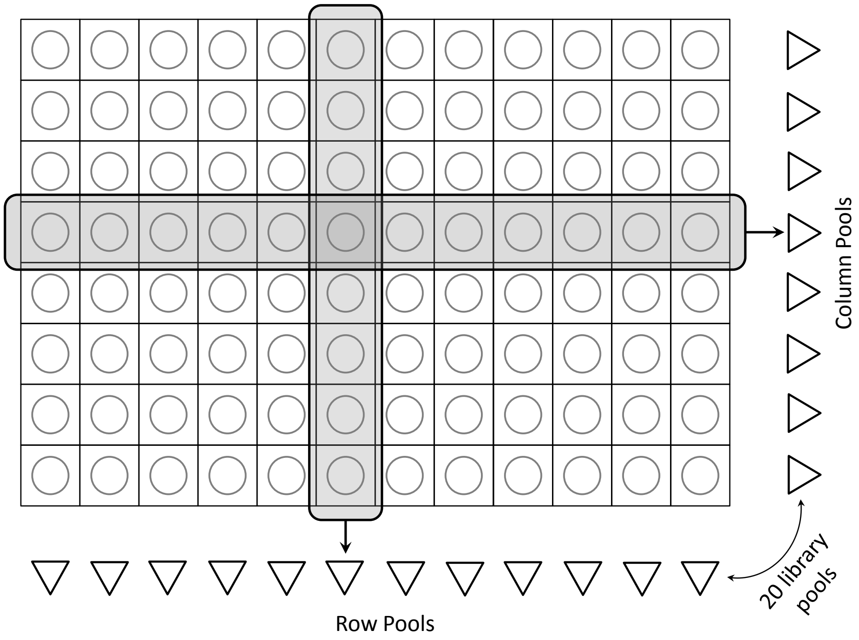

Figure 1: There are 96 wells and 20 pools (12 column- and 8 row-pools) in our bi-dimensional pooling scheme. Thus, each individual is present in two pools.

Figures

Figure 1 - Bi-dimensional arrangement of the overlapping pools experiments

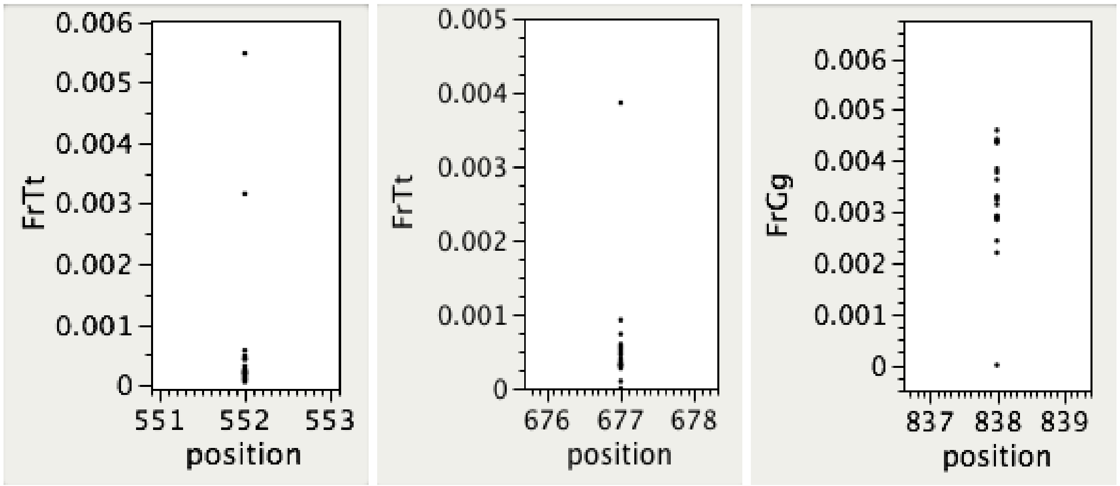

Figure 2: Three mutations ordered, left to right, by increasing difficulty to identify visually. Left and middle, $C \rightarrow T$ mutations at positions 552 and 677, respectively, in wheat genes APHYC and AVRN, resp. Right, an $A \rightarrow G$ mutation at position 838 in rice gene OsRDR2.Each dot in the plots is a library pool, and on the y-axis is the frequency of the base to which the reference has been mutated.

Figure 2 - Example base positions with mutations in the data of varying difficulty for identification

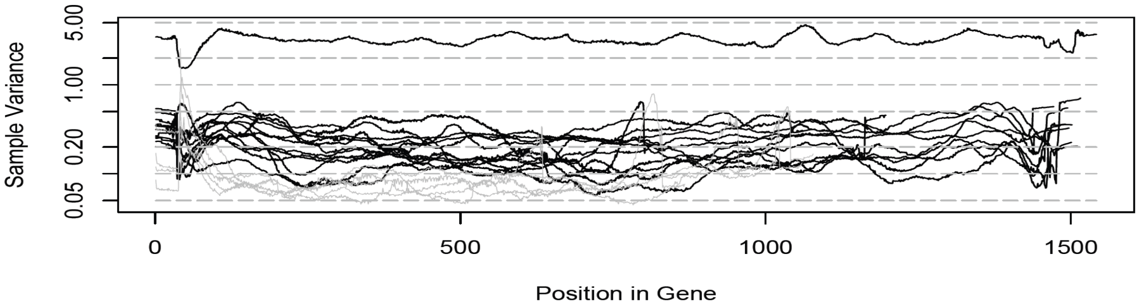

Figure 3: Normalized variance of coverage levels across libraries in TILLING genes in rice (black) and wheat (gray). HLP1 is on top.

Figure 3 - Variance in coverage across libraries in the data

Predictions	CAMBa		Outlier		Poisson	
	$F(t)$	Confirmed	z-score	Confirmed	pvalue	Confirmed
308	0	11	3.92	11	3.77e-08	5
131	1	11	4.85	10	1.58e-11	2
90	1.595	11	5.54	8	2.86e-13	2
75	2	10	5.87	8	1.23e-14	2
54	3	10	6.36	6	4.96e-16	1
46	4	9	6.78	6	2.06e-17	1
40	5	7	7.17	5	1.20e-18	1

Table 1: Overlap of predictions by CAMBa, Outlier method, and Poisson outlier method, with the 11 confirmed mutations for Rice. We underscore the line associated with the suggested number of predictions. VarScan returned no mutations at any threshold. CRISP did not return any predictions with default parameters.

Table 1 - Performance comparison of CAMBa to other methods on the Rice data set

Predictions	CAMBa		Outlier		Poisson		VarScan	
	$F(t)$	Confirmed	z threshold	Confirmed	pvalue	Confirmed	pvalue	Confirmed
310	-10	37	1.83	37	1.42e-02	35	9.51e-07	19
172	-5	36	2.27	36	7.18e-03	35	2.51e-09	16
107	0	36	3.07	36	1.15e-03	34	5.71e-13	12
92	5	33	4.93	32	4.63e-06	32	3.52e-14	10
81	10	31	7.01	32	9.26e-08	29	5.04e-15	9
59	15	21	10.2	23	1.21e-11	22	5.90e-19	7

Table 2: Overlap of Predictions by CAMBa, Outlier method, the Poisson outlier method, and VarScan with the 39 confirmed mutations for Tetraploid Wheat.We underscore the line associated with the suggested number of predictions.

Table 2 - Performance comparison of CAMBa to other methods on the Wheat data set

Predictions	CAMBa		Outlier	
	$F(t)$	Confirmed	z-score	Confirmed
252	0	11	3.47	11
73	2	10	4.53	10
40	5	7	6.03	7
24	10	6	7.58	4

Table 3: Excluding HLP1 from the rice data lowers the coverage variance across libraries and CAMBa performs comparably.

Table 3 - Effects of lowered variance in the data on CAMBa's performance

Predictions	CAMBa		Outlier	
	$F(t)$	Confirmed	z-score	Confirmed
104	0	34	3.86	30
65	5	27	6.57	24
43	10	21	9.41	17
18	15	10	17.1	9

Table 4: When coverage variance is artificially increased in wheat, CAMBa has the advantage.

Table 4 - Effects of increased variance in the data on CAMBa's performance