MoGUL: Detecting Common Insertions and Deletions in a Population

Seunghak Lee[1,2], Eric Xing[2], and Michael Brudno[1,3,⋆]

[1] Department of Computer Science, University of Toronto, Canada
[2] School of Computer Science, Carnegie Mellon University, USA
[3] Banting and Best Dept. of Medical Research, University of Toronto, Canada
brudno@cs.toronto.edu

Abstract. While the discovery of structural variants in the human population is ongoing, most methods for this task assume that the genome is sequenced to high coverage (e.g. 40x), and use the combined power of the many sequenced reads and mate pairs to identify the variants. In contrast, the 1000 Genomes Project hopes to sequence hundreds of human genotypes, but at low coverage (4-6x), and most of the current methods are unable to discover insertion/deletion and structural variants from this data.

In order to identify indels from multiple low-coverage individuals we have developed the MoGUL (Mixture of Genotypes Variant Locator) framework, which identifies potential locations with indels by examining mate pairs generated from all sequenced individuals simultaneously, uses a Bayesian network with appropriate priors to explicitly model each individual as homozygous or heterozygous for each locus, and computes the expected Minor Allele Frequency (MAF) for all predicted variants. We have used MoGUL to identify variants in 1000 Genomes data, as well as in simulated genotypes, and show good accuracy at predicting indels, especially for MAF > 0.06 and indel size > 20 base pairs.

1 Introduction

Next generation sequencing technologies have dramatically decreased the cost of sequencing human genomes. These technologies are enabling the 1000 Genomes Project - an ambitious undertaking to reconstruct hundreds of genotypes and understand the polymorphisms present in the human population. The resequencing of humans for the 1000 Genomes Project uses a combination of approaches, including deep sequencing of several individuals and whole-exome resequencing via DNA-capture. Simultaneously, the largest fraction of individuals will be sequenced via a low-coverage whole-genome shotgun approach, where each individual will be sequenced to ∼4-6x coverage. At this point in time it is not clear if this low coverage will be sufficient to identify a large fraction of the human variation, especially structural genomic polymorphisms.

While methods for the discovery of SNPs from read mapping have been available for some time [1], and the past two years have seen several tools developed

⋆ To whom correspondence should be addressed.

B. Berger (Ed.): RECOMB 2010, LNBI 6044, pp. 357–368, 2010.
© Springer-Verlag Berlin Heidelberg 2010

specifically for discerning SNPs from NGS data ([2,3,4]), the development of algorithms for the identification of larger, structural variants (SVs), including insertions and deletions (indels), is still a very active research area. While the identification of very small indels can be accomplished by directly analyzing the read mappings, with 36bp reads it is difficult to identify indels > 10 bases. The identification of larger indels and other rearrangements is typically accomplished via the mate pair, or paired-end mapping technique (see [5] for a review). In this approach, two reads are sequenced from the two ends of a DNA fragment (the insert). Because the size of the DNA fragment is (approximately) known, structural variants can be identified by comparing the expected insert size to the distance between the mapped reads in the reference genome: if these are significantly different (the mate pair is termed discordant), it is likely that an SV has occurred between the two mappings. The past few years have seen the development of several novel methodologies and tools for SV discovery based on the analysis of discordant mate pairs, including a formal framework for identification of structural variants [6], tools that allow for flexible clustering of mate pairs to identify SVs [7], maximum parsimony and maximum likelihood approaches for SV detection [8], as well as tools that combine paired-end mapping with careful analysis of unpaired reads to assemble SV breakpoints [9].

Previously we proposed MoDIL [10], a method for SV identification based on the analysis of all mate pairs (concordant and discordant) that span a particular genomic location. MoDIL (Mixture of Distributions Indel Locator) fits two (possibly shifted) distributions of insert sizes (corresponding to the two haploid genotypes in a diploid) to the observed mapped distances at each location in the genome. By analyzing these distributions it is possible to discover much smaller indels than with other mate pair-based approaches. MoDIL, however, cannot be directly applied to low coverage individuals, including the bulk of the 1000 Genomes data, as it requires at least 20 inserts covering a genomic locus to identify indels (it is difficult to accurately fit two distributions with fewer data points). In the 1000 Genomes data, each locus is expected to be covered, on average, by 4 mate pairs in each individual. While the total coverage from all individuals is much higher, and most polymorphisms are di-allelic (i.e. there are only two alleles at a given locus in the human population), MoDIL expects the fractions of mate pairs sampled from each haplotype to be approximately equal. In contrast, in the 1000 Genomes data the fractions are determined by the allele frequencies and will vary across the loci.

In this work we build a Bayesian approach for the discovery of indel polymorphisms from mixtures of large numbers of genotypes, such as 1000 Genomes data. Our approach, MoGUL (Mixture of Genotypes Variant Locator), builds a Bayesian network that uses priors to explicitly model each individual as homozygous or heterozygous, and computes the expected Minor Allele Frequency (MAF) at each location along the chromosome. We use MoGUL to identify variants in the 1000 Genomes data and simulated genotypes, and demonstrate that it allows for the identification of indels > 30 bases for MAF > 0.04, while indels as small as 20 bases can be identified for MAF > 0.06.

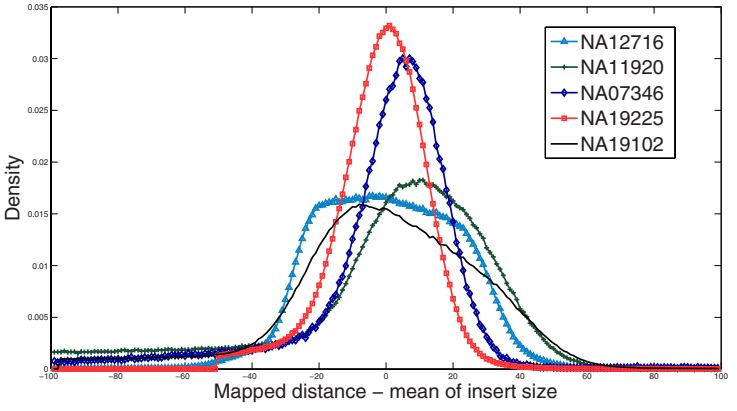

Fig. 1. Distribution of insert sizes from different individuals, shifted so that they are all centered at zero. Note the discrepancies among the individual distribution, necessitating modeling them as separate random variables. Here mean of insert sizes are set to be zero.

2 Methods

The main difficulty in identifying indels from paired-end data is differentiating mate pairs coming from a locus with an indel from those with an anomalous insert size. The insert size from each individual l follows a distribution, $p(Y_l)$ (see Figure 1), and individual mate pairs generated from the tail of the distribution are impossible to discern from mate pairs overlapping an indel. Previous methods, such as MoDIL [10] and BreakDancer [9], use support from other mate pairs, generated by the high mate pair coverage to separate these cases. While each individual in our dataset will have only a few mate pairs sampled at every genomic location, our algorithm combines the mate pairs generated from many individuals to achieve sufficient coverage. MoGUL models mate pairs as generated from either one or two unknown distributions, corresponding to the two possible alleles at this location among the human genotypes. Our algorithm does not consider tri-allelic variants, which are rare.

Our algorithm starts by mapping all of the mate pairs onto the reference genome. We use the MrFAST tool [11], which identifies mappings for every mate pair that has at most 2 mismatches in each read and has the *mapped distance* (the distance between the forward and reverse reads of the pair) closest to the expected insert size. If this mapped distance is within 3 standard deviations, only the best mapping is identified. If no such mapping is found, all possible mappings for the two reads are returned, and our algorithm considers all of them. For every genomic location we identify those mate pairs that would be affected if that location was the site of an indel. These mate pairs will have the two reads mapping on opposite sides of the genomic location, and we refer to this set of mate pairs as a cluster (see next section).

If the genomic location is the site of an indel that is polymorphic in the human population, mate pairs in the corresponding cluster may be generated from two distributions, corresponding to the two alleles (with and without the indel). Using a Bayesian network we infer the size of the indel, as well as the individuals with indels for each cluster. Because our model may identify the same indel calls from multiple clusters, a final post-processing step is used to combine these calls and to compute the log likelihood ratio between our model and the null model. For simplicity, in the following sections we will call a mate pair "discordant" if there is significant disparity between insert size and mapped distance, and "concordant" otherwise. Note that these terms are only used for convenience – we do not *a priori* assign mate pairs to these categories.

2.1 Clustering Mate Pairs

We first generate clusters with mappings of mate pairs for each genomic locus, and determine whether or not the locus contains a common indel. In this step we find a set of mate pairs $\boldsymbol{C}$ from L number of individuals, all of which overlap with a particular genomic location. Figure 2 illustrates our clustering scheme.

For each mate pair we look at one base after the left read and all mate pairs overlapping the location form a cluster $\boldsymbol{C}$. We explain how we detect indels from these clusters by example. Suppose the mate pairs in Figure 2 are from the same mate pair library with the first two mate pairs discordant and the rest concordant. In such a case, as shown in Figure 2, the first two mate pairs agree on a certain indel size (they have similar mapped distance), and the indel can be detected from the second to the fourth cluster containing the two discordant mate pairs (we merge indel calls in a post-processing step).

If we use all the clusters generated by this scheme, the number of clusters will be close to the number of mate pairs, and the algorithm will be too slow. Instead, we filter out clusters if it is very likely that there is no indel iat the corresponding location. For each individual l, we compute the likelihood that the mate pairs were generated from a cluster with no indel (p-value). If there is at least one individual with significant p-value (< 0.001) or two individuals with less significant p-value (< 0.05), the locus is deemed significant.

We define the p-value as the probability of having at least predicted size of indel ($> \gamma$) given no indels. Let $\{D_{l1}, \ldots, D_{ln}\}$ represent independent and identically distributed random variables corresponding to the mapped distances of mate pairs generated from the l-th individual with insert size distribution $p(Y_l)$, mean μ_{Y_l} and standard deviation σ_{Y_l}. Their mean follows the Gaussian distribution with mean equal to the mean of the insert size μ_{Y_l} and standard deviation of $\sigma_{Y_l}/\sqrt{n}$ according to the central limit theorem. We define the p-value for the individual l with the size of indel γ as follows:

$$\text{p-value} = \sum_{\gamma}^{\infty} P(X; 0, \sigma_{Y_l}/\sqrt{n}) = \sum_{-\infty}^{0} P(X; \gamma, \sigma_{Y_l}/\sqrt{n})$$

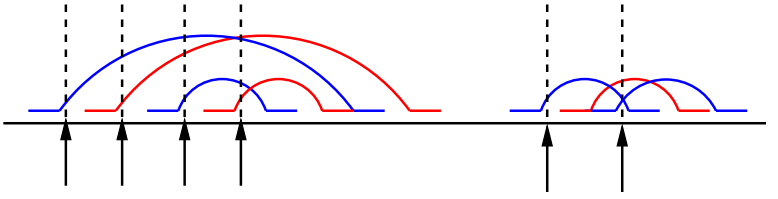

Fig. 2. This figure shows how to generate clusters with mapped mate pairs in the reference genome. Mate pairs are colored by red or blue representing different individuals. For each mate pair X_i, we generate a cluster consisting of all mate pairs overlapping a genomic location of one base after the left read of the mate pair X_i (the locations of the arrows).

Here, $X = D - \mu_{Y_l}$ is the expected size of the indel, and $P(X)$ follows the Gaussian distribution. The second equality can be proven via symmetry of Gaussian.

In computing the p-value we correct for the possibility that the cluster contains a heterozygous indel by using a shifted sample mean: $\gamma' = 2\gamma$.

2.2 Detecting Common Indels Using a Bayesian Network

The clusters from Sec. 2.1 include mate pairs generated from many individuals, all of which have unique distributions of insert sizes (see Figure 1). We define the variable X_{lm} as the expected size of indel from the m-th mate pair of individual l:

$$X_{lm} = D_{lm} - \mu_{Y_l}$$

where D_{lm} is mapped distance of the m-th mate pair of the individual l and μ_{Y_l} is mean of the insert size distribution $p(Y_l)$. We will use random variabe X_{li} instead of D_{li} because it shifts the distributions for all individuals so that they are all centered at zero. Given a cluster of mate pairs as an input, we developed a Bayesian network (Figure 3) to infer the size of the indel polymorphism (if one exists), and haplotypes of individuals that contain the indel. The Bayesian network generates mate pairs $\{X_{lm}\}$, while internal states correspond to the presence/absence of indel and its heterozygosity. All random variables are defined for an input cluster, rather than the whole individual genome.

We model the individual l with random variable Z_l:

$$Z_l = \begin{cases} 0 \text{ if individual } l \text{ has no indel} \\ 1 \text{ if individual } l \text{ has an indel.} \end{cases}$$

We use the random variable Q_{lm} to model the two copies of chromosomes (alleles) in individual l. Note that subscript l refers to individual l and m denotes m-th mate pair generated from this individual:

$$Q_{lm} = \begin{cases} 0 \text{ if } Z_l = 1 \text{ and chromosome contains no indel} \\ 1 \text{ if } Z_l = 1 \text{ and chromosome contains an indel} \\ 2 \text{ if } Z_l = 0. \end{cases}$$

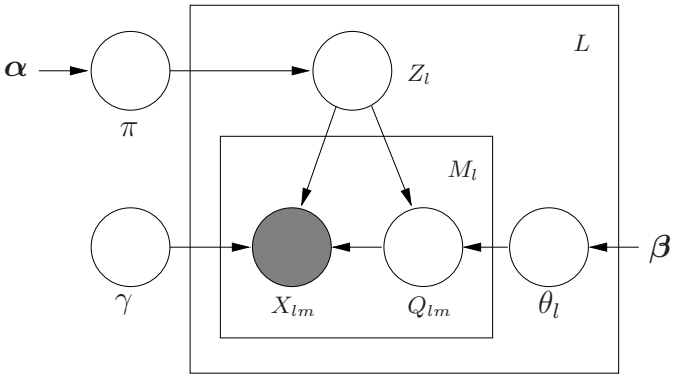

Fig. 3. Bayesian network for detecting common indels at a particular locus in the genome. Here L represents the number of individuals and M_l is the number of mate pairs from individual l. The random variable Z_l determines whether individual l has an indel or not. If individual l has an indel ($Z_l = 1$), Q_{lm} generates a mate pair X_{lm} and θ_l controls the heterozygosity of Z_l. Mate pair, X_{lm}, is generated from distribution of insert sizes with zero mean or with shifted mean of γ if X_{lm} has an indel. If individual l has no indel ($Z_l = 0$), mate pairs $\{X_{lm}\}_{m=1}^{M_l}$ are generated from $p(Y_l)$ with zero mean. Priors π and θ_l are controlled by $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parameters.

As shown in Figure 3 we can generate X_{lm} given Z_l, Q_{lm} and size of indel γ. For example, if $Q_{lm} = 1$, X_{lm} is generated from $p(Y_l)$ with an indel size of γ. If $\{Z_l = 0 \cup Q_{lm} = 0\}$ then X_{lm} is generated from $p(Y_l)$ with no indel. For simplicity we omit the $p(Y_l)$s in Figure 3. To avoid overfitting problems we applied Bayesian priors π and θ_l to Z_l and Q_{lm}, respectively.

We smooth the distribution of $p(Y_l)$, and define a new probability distribution of insert sizes, $q(X_l)$, for individual l as follows:

$$q(X_l) = \begin{cases} \sum_{k_i\sigma_{Y_l} \le y-\mu_{Y_l} < k_{i+1}\sigma_{Y_l}} p_{Y_l}(y) & \text{if } k_i\sigma_{Y_l} \le X_l < k_{i+1}\sigma_{Y_l} \\ \sum_{-k'_{j+1}\sigma_{Y_l} \le y-\mu_{Y_l} < -k'_j\sigma_{Y_l}} p_{Y_l}(y) & \text{if } -k'_{j+1}\sigma_{Y_l} \le X_l < -k'_j\sigma_{Y_l} \end{cases}$$

Here we sum $p_{Y_l}(y)$s over the intervals $[k_i\sigma_{Y_l}, k_{i+1}\sigma_{Y_l})$ for deletions and $[-k'_{j+1}\sigma_{Y_l}, -k'_j\sigma_{Y_l})$ for insertions. In our experiments, we used 10 values of k_i and k'_js ($i, j \in \{1, 2, \ldots, 10\}$, $k_1 = k'_1 = 0$). Probability distributions of the random variables in Figure 3 are defined as follows:

$$p(Z_l = z|\pi) = \pi^z(1-\pi)^{1-z}$$

where $z = 0$ if individual l has no indel and $P(Z_l = 0) = \pi$ and $P(Z_l = 1) = 1-\pi$.

$$p(Q_{lm} = q|Z_l = 1, \theta_l) = \theta_l^q(1-\theta_l)^{1-q}$$

where $q = 0$ if the chromosome contains no indel, and 1 otherwise. If $Z_1 = 0$, we do not generate mate pair X_{lm} from Q_{lm} and set $q = 2$. We generate X_{lm} from the following distribution:

$$p(X_{lm}|Z_l, Q_{lm}, \gamma) = \begin{cases} q(X_{lm}) & \text{if } \{Z_l = 0 \cup q = 0\} \\ q(X_{lm} - \gamma) & \text{if } q = 1. \end{cases}$$

The priors π and θ_l follow the beta distribution, which is the conjugate prior of binomial distributions.

To infer the states of our model, we find maximum a posteriori (MAP) solution because it is fast and deterministic. We initialize our model using heuristics (e.g. $Q_{lm} = 1$ if $X_{lm} > \sigma_l$) and random configurations, and run the model multiple times to avoid local maxima. Given current states of the model the update rules are given as follows (updated states are denoted by (*)):

$$\pi^* = \frac{u + \alpha_1 - 1}{L + \alpha_1 + \alpha_2 - 2}$$

where u is the number of individuals with no indel. In practice we use $\boldsymbol{\alpha} = \{30, 1\}$ because most variants have a small MAF [12].

$$\theta_l^* = \frac{v + \beta_1 - 1}{M_l + \beta_1 + \beta_2 - 2}$$

where v is the number of mate pairs with $Q_{lm} = 0$ in individual l, and we set $\boldsymbol{\beta} = \{5, 5\}$, favoring heterozygous indels, as these are more likely under a neutral evolutionary model.

We update the random variables γ, Z_l and Q_{lm} as follows:

$$\gamma^* = \arg\max_{\gamma} \prod_{l=1}^{L} \prod_{m=1}^{M_l} P(X_{lm}|Z_l, Q_{lm}, \gamma)$$

$$Z_l^* = \arg\max_{Z_l \in \{0,1\}} P(Z_l|\pi) \prod_{m=1}^{M_l} P(X_{lm}|Z_l, \gamma, Q_{lm}) P(Q_{lm}|Z_l, \theta_l)$$

$$Q_{lm}^* = \arg\max_{Q_{lm} \in \{0,1,2\}} P(Q_{lm}|Z_l, \theta_l) P(X_{lm}|Z_l, Q_{lm}, \gamma).$$

This algorithm is iterated, with each hidden random variable updated until the posterior probability of the model cannot be improved by the value of the threshold (e.g. $\tau = 10^{-4}$).

2.3 Merging and Assigning Confidence to Indel Calls

In the post-processing step we merge duplicated indel calls. As shown in Sec. 2.1, a single indel may be found in multiple clusters. We merge indel calls if they meet three criteria: (1) the predicted indel regions overlap, (2) the expected size of the indel is similar ($< \sigma_{\text{mix}}$) , (3) the sets of individuals for whom the indel is predicted overlap. Here σ_{mix} is the standard deviation of insert sizes from all individuals.

To assign confidence values for every cluster we compute the log likelihood ratio R between our model and null model as follows:

$$R = \sum_{l=1}^{L} \sum_{m=1}^{M_l} \log P(X_{lm}|Z_l, Q_{lm}, \gamma) - \sum_{l=1}^{L} \sum_{m=1}^{M_l} \log P(X_{lm}|Z_l, Q_{lm}, 0).$$

We discard indel calls if the log likelihood ratio is not significantly larger than a pre-specified threshold (by default, 30).

3 Results

In the sections below, we use two different approaches to validate our algorithms. First, we use simulated data to evaluate how well MoGUL performs at different variant frequencies, and then use MoGUL to perform variant discovery on one chromosome of the current 1000 Genomes dataset, that includes 124 individuals sequenced at approximately 4x.

3.1 Simulation Results

We first validate our model through simulation results. In our simulation, we sampled mate pairs from 120 individuals, with the mate pair library size of each individual l following the experimental distribution $p(Y_l)$.

We generated indels of 10-100 base pairs and implanted them in the individual genomes, varying the minor allele frequency (MAF) from 0.02 to 0.5. Figure 4 shows the heatmap for the performance of MoGUL. MoGUL works well for MAF greater than 0.06, for indels > 20 base pairs.

To investigate the precision and recall rate of MoGUL we generated 10,000 clusters with 50 individuals (100 haplotypes). 1000 clusters contained implanted indels of 20-150 base pairs, while 100 clusters contained implanted indels of 150-1000 base pairs. For each individual we sampled mate pairs with approximately 2-3x read coverage. We detected indels for these individuals using MoGUL. The recall and precision rates of our algorithm are shown in Table 1.

3.2 1000 Genome Project Pilot Dataset

In order to validate MoGUL on real data, we downloaded low coverage individuals generated by the pilot project for the 1000 Genomes project from the NCBI

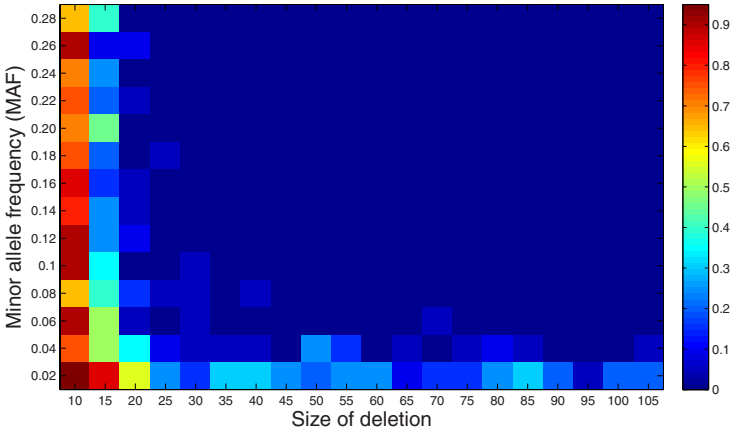

Fig. 4. Heatmap representing the performance of MoGUL. The color of each cell indicates average error rate of 20 MoGUL simulations for a given combination of deletion size (X axis) and Minor Allele Frequencies (Y axis). If the size of indel predictions by MoGUL is more than 10bp away from the true size of deletion we consider it incorrect.

Table 1. Comparison between indel calls in chromosome 20 located by our approach with the datasets generated by Mills et al. [13] (all MoGUL indels), and MoDIL [10] (only indels in NA18507, the same individual as was studied by Lee et al., was considered). For the simulation experiments we consider the indel call is correct if the difference between the true indel size and the predicted one is less than 10bp and the log likelihood ratio is greater than 10.

		Population			NA18507			Simulation	
Length	Type	MoGUL	Mills et al.	Overlap	MoGUL	MoDIL	Overlap	Recall	Precision
≥100bp	INS	6	20	0	2	1	1	0.91	1
	DEL	1009	183	57	34	13	10	0.89	1
50-100bp	INS	56	44	15	19	4	4	0.92	0.68
	DEL	486	71	42	22	6	5	0.86	0.99
20-50bp	INS	170	231	43	25	24	12	0.64	0.37
	DEL	1818	327	194	101	84	31	0.57	0.74

trace archive, aligned these to the NCBI reference genome with MrFAST [11], and predicted indels for all of these on chromosome 20. The results are summarized in Table 1. Overall, MoGUL predicted 3,545 events in any individual on chromosome 20. This is approximately 630 events per individual. We compare these indels to previously discovered variants both across the population, and for one specific individual, NA18507.

To our knowledge, the only previous study that has characterize small to medium size indels in the human populations is by Mills et al [13]. They used low coverage Sanger-style reads from 36 individuals to identify indels via the split-read mapping approach. Thus they are able to identify the exact size of the indel, while the MoGUL method infers it indirectly from the discordant mappings. Overall, the overlap between MoGUL and the indels of Mills et al was statistically significant. While exact sensitivity and specificity of the two methods is difficult to analyze, as different (and fewer) individuals were used for the Mills et al study, the size correlation of overlapping indels was very strong, and the overall error of MoGUL size estimates was small (see Figure 6).

In order to enable the direct comparison of indel discovery from single high coverage individual versus multiple low coverage individuals, we included in our dataset a down-sampled version of the NA18507 Yoruban genome which we previously analyzed using the MoDIL method. Remarkably, MoGUL was able to identify 83% of the indels $>$ 50bp (20/24) that were previously detected by MoDIL, while identifying several additional variants that were missed by MoDIL, possibly due to low coverage in the NA18507 individual specifically. Of the events 20-50bp, 40% (43/108) were recovered by MoGUL.

In Figure 5 we plot the minor allele frequency of the variants discovered by our method. The distribution agrees with the expected curve until $\sim$ MAF 0.07, but then drops rapidly – demonstrating MoGUL's inability to identify indels at low minor allele frequencies.

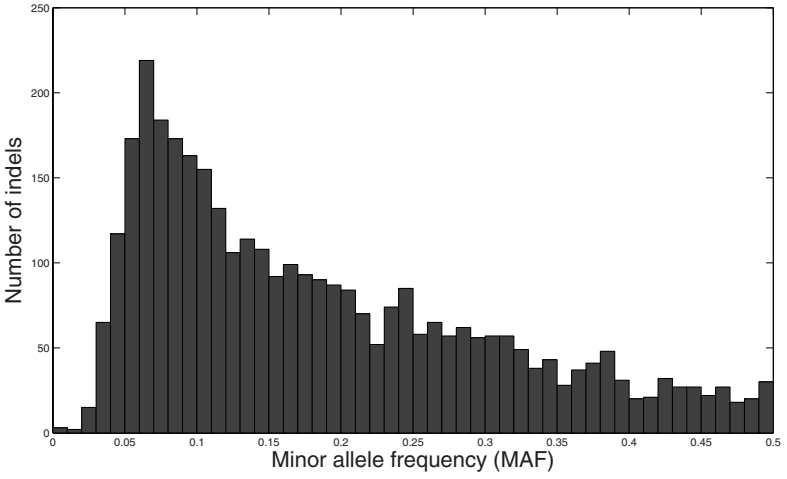

Fig. 5. Distribution of minor allele frequencies for indels in the 1000 Genomes dataset

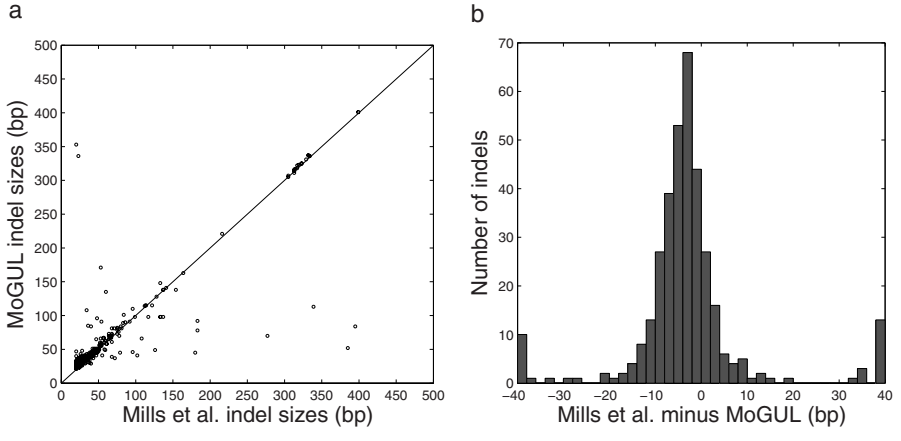

Fig. 6. (A) A scatter plot showing the lengths of overlapping indels between the Mills et al. dataset and MoGUL predictions. Overall the lengths are highly correlated. The cluster of indels of length 300 corresponds to Alu element activity. (B) The absolute error in the estimation of indel length. The predicted lengths of the indels are very close (typically within 10 bases) of the true indel size. Overall the distribution of the error follows a Gaussian, as expected from the model (see [10] for details). The outliers may indicate either false positives for either dataset or tri-allelic variants.

4 Discussion

The identification of various polymorphisms in the human population is an important step towards understanding the landscape of human genotypes. In this paper we present MoGUL: the Mixture of Genotypes Variant Locator, a tool

to identify common insertion/deletion polymorphisms from many individuals sequenced at low coverage. We validate our approach via simulated data at various allele frequencies, as well as with data from the 1000 Genomes project. MoGUL can identify indels >20 base pairs with at least 0.06 MAF, using the current low coverage data; it is expected that the coverage will double to 6-8x per individual for the final 1000 Genomes project data release, and we are hopeful that MoGUL's performance will further improve on this larger dataset. Another application of MoGUL is resequencing of biopsy tissues, where the diseased (tumourous) tissue is biopsied (and sequenced) together with the healthy surrounding tissue, leading to a mixture of several genotypes at each location.

Simultaneously, MoGUL is only capable of recapturing a small fraction of the rare variants that predominate in the human population. While capturing common genotypes is important, it is thought that rare alleles, ones with MAF $<$ 0.01, are much more likely to be evolutionarily harmful and disease related [12]. Designing methods that can find these variants from paired-end data, possibly incorporating direct information on read matches, as in the Pindel tool [14], is an important avenue for further research.

Acknowledgments

We would like to thank Lisa Brooks of the NIH for permission to use 1000 genomes data, and to Can Alkan and Fereydoun Hormozdiari for providing us with the MrFAST mapping tool.

References

1. Marth, G.T., et al.: A general approach to single-nucleotide polymorphism discovery. Nature Genetics 23, 452–456 (1999)
2. Li, H., Ruan, J., Durbin, R.: Mapping short dna sequencing reads and calling variants using mapping quality scores. Genome Research 18, 1851–1858 (2008)
3. Li, R., et al.: Snp detection for massively parallel whole-genome resequencing. Genome Research 19, 1124–1132 (2009)
4. Hoberman, R., et al.: A probabilistic approach for SNP discovery in high-throughput human resequencing data. Genome Research 19, 1542–1552 (2009)
5. Medvedev, P., Stanciu, M., Brudno, M.: Computational methods for discovering structural variation with high throughput sequencing. Nature Methods 6, S13–S20 (2009)
6. Lee, S., Cheran, E., Brudno, M.: A robust framework for detecting structural variations in a genome. Bioinformatics 24, i59–i67 (2008)
7. Korbel, J., et al.: Paired-end mapping reveals extensive structural variation in the human genome. Science 318, 420–426 (2007)
8. Hormozdiari, F., Alkan, C., Eichler, E.E., Sahinalp, S.C.: Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome Research 19, 1270–1278 (2009)
9. Chen, K., et al.: Breakdancer: an algorithm for high-resolution mapping of genomic structural variation. Nature Methods 6, 677–681 (2009)

10. Lee, S., Hormozdiari, F., Alkan, C., Brudno, M.: MoDIL: Detecting INDEL Variation with Mixtures of Distributions. Nature Methods 6, 473–474 (2009)
11. Alkan, C., et al.: Personalized copy number and segmental duplication maps using next-generation sequencing. Nature Genetics 41, 1061–1067 (2009)
12. Kryukov, G., Shpunt, A., Stamatoyannopoulos, J., Sunyaev, S.: Power of deep, all-exon resequencing for discovery of human trait genes. Proceedings of the National Academy of Sciences 106, 3871–3876 (2009)
13. Mills, R., et al.: An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Research 16, 1182–1190 (2006)
14. Ye, K., Schulz, M.H., Long, Q., Apweiler, R., Ning, Z.: Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25, 2865–2871 (2009)