MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions

To the Editor: Human genetic variation comes in a wide range of sizes, from single-nucleotide polymorphisms and very small insertions and deletions (indels) to 'structural' variants, in which large segments of the genome are inserted, deleted, inverted or duplicated. Recently several methods for the identification of both small-size indels (<10 base pairs (bp))[1] and larger ones (>50 bp)[2,3] from high-throughput sequencing have been developed. There should also be a large amount of 'medium-sized' variation: insertions and deletions of 10–50 nucleotides. Here we describe MoDIL, mixture of distributions indel locator, the first method to identify 20–50-bp indels from high-throughput sequencing data. MoDIL is available at http://compbio.cs.toronto.edu/modil/.

Most sequencing techniques allow for the generation of mate pairs (two reads at an approximately known distance (insert size)). Mate pairs are used to locate structural variants by comparing the distance between the mapped locations of the read pairs on a reference genome (mapped distance) and the known insert size. A large deviation implies a structural variant, whereas smaller differences are likely due to variance in clone sizes and hence are not considered by most methods[2–4]. This limits the resolution of the structural variation detection approaches to indels >50 bp. Although deviation by several mate pairs from the expected insert size is likely, in MoDIL we used the intuition that deviation of many mate pairs (even by a small number of nucleotides) is indicative of an insertion or deletion.

The MoDIL algorithm (summarized in **Fig. 1a**,**b**) compares the distribution of insert sizes in the sequenced library (we call it $p(Y)$) to the distribution of the observed mapped distances at a particular genomic location (i). Given this location, the cluster C_i includes all of the mate pairs that overlap i. In clusters from loci without indel polymorphisms, the distribution of the observed mapped distances $p(C_i)$ will be distributed identically to $p(Y)$. If there has been a homozygous insertion or deletion at this location, the distribution $p(C_i)$ will shift (**Fig. 1a**). If the observed cluster is the site of a heterozygous indel, approximately half of the observed mate pairs will be generated from the shifted distribution, and the other half will come from the original, unshifted $p(Y)$ (**Fig. 1b**). MoDIL represents the random variable of the expected size of indel (mean of insert size minus the mapped distance) with two random variables, one for each haplotype. Given a cluster, MoDIL identifies the two distributions, $\{D_1,D_2\}$, with the fixed shape of $p(Y)$ and arbitrary means that best fits the observed data using the Kolmogorov-Smirnov test. To find the means of the two distributions, MoDIL uses the expectation-maximization algorithm and appropriate Bayesian priors to prevent over-fitting.

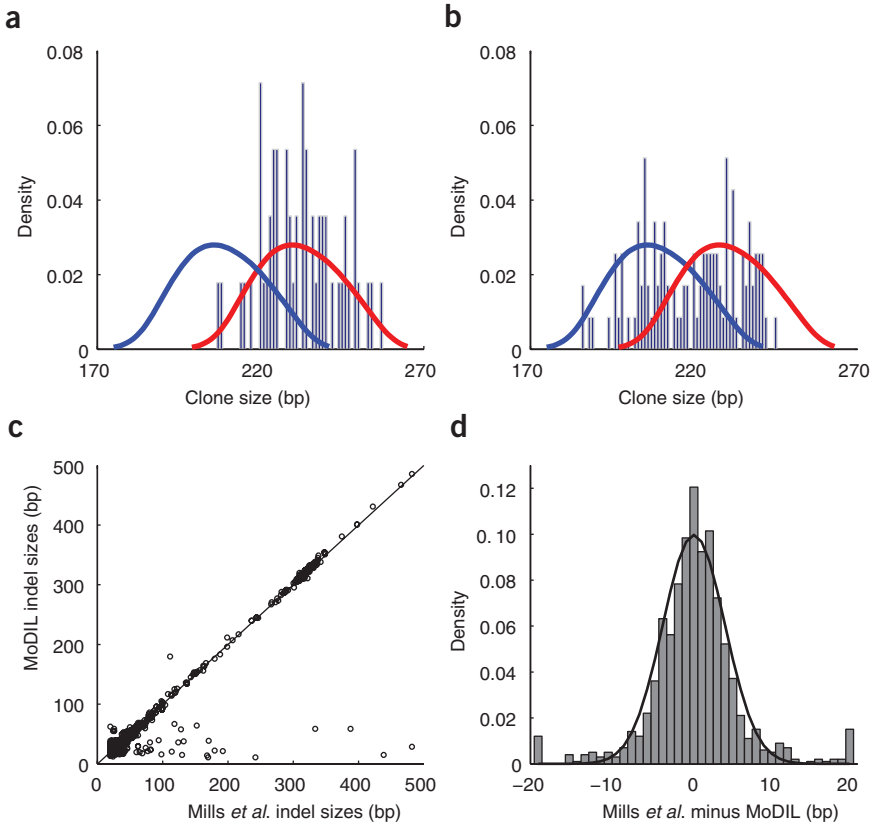

Figure 1 | Summary of the MoDIL algorithm and accuracy of indel size estimation. (**a**,**b**) MoDIL algorithm for example homozygous (**a**) and heterozygous (**b**) deletions with the observed distribution of mapped distances within a cluster $p(C)$ (gray). A homozygous deletion of ~24 bp: the shift of $p(Y)$ (blue) that best matches the observed distribution $p(C)$ (red, centered at 232 bp; **a**). A heterozygous deletion of ~22 bp: mapped distances are generated from two distributions, with means 230 bp and 208 bp (red and blue, respectively; **b**). (**c**) Size correlation of overlapping indels (20–500 bp) between Mills data[5] and MoDIL indel calls (square of Pearson's correlation coefficient, $r^2 = 0.96$). (**d**) Difference between the known indel sizes and MoDIL predictions, superimposed on a Gaussian distribution with $\sigma = 4$ (the expected s.d. for a cluster with 20 mate pairs).

Table 1 | Evaluation of MoDIL on various datasets

			Overlap with known indels[7]			Simulation	
Size	Type	MoDIL	Total	Found	FNR	Recall	Precision
≥20 bp	Insertion	1,336	78	75	0.04	0.85	0.90
	Deletion	3,799	196	187	0.05	0.91	0.89
15–19 bp	Insertion	1,601	119	84	0.29	0.61	0.65
	Deletion	5,333	178	126	0.29	0.78	0.45
10–14 bp	Insertion	936	370	130	0.65	0.44	0.37
	Deletion	3,682	593	227	0.62	0.54	0.27

Number of insertions and deletions of each size identified by MoDIL from Illumina data[6] and the number of previously known indels[7] (total) overlapped by MoDIL predictions (found). We considered indels discovered in ref. 7 but not by us to be false negatives, and the ratio of these as a function of all indels in ref. 7 the false negative rate (FNR). Using a simulated dataset (simulation), we computed the fraction of true indels discovered by MoDIL (recall) and the fraction of predicted indels that were real (precision).

For each distribution $D_{k \in \{1,2\}}$ the size of the indel event can be estimated with high confidence: its expected size follows a Gaussian distribution with mean μ and standard deviation σ where

$$\mu = \mu_{p(Y)} - \mu_{D_k} \text{ and } \sigma = \sigma_{p(Y)} / \sqrt{n}$$

with n being the number of mate pairs in the distribution, regardless of the shape of $p(Y)$. MoDIL thus uses higher clone coverage to locate progressively shorter indel variation. For proof and thorough description of the algorithms, see **Supplementary Note**.

To evaluate our method, we conducted simulation experiments by implanting known human indels[5] into chromosome 1 and simulating mate-pair data. We used this simulated dataset (51×10^6 mate pairs) to predict indels in the chromosome. MoDIL achieved both precision and recall ≥ 0.85 for indels that were ≥20 base pairs (**Table 1**). We compared MoDIL to tools[1,2] for structural and indel variation discovery using this simulated data (**Supplementary Note**), and no other tool we evaluated identified 15–40-bp indels.

We also applied MoDIL to Illumina whole-genome shotgun-sequencing data[6]. The 3.5×10^9 reads provided 40-fold read and 120-fold clone coverage of the National Center for Biotechnology Information (NCBI) reference human genome. The reads had been mapped[2], with observed insert size $\mu = 208$ bp and $\sigma = 13$ bp. We required each cluster to have at least 20 mate pairs, and used MoDIL to identify 3,981 insertions and 13,147 deletions in the sequenced individual genome relative to the NCBI reference genome. The sizes were 6–118 nucleotides for insertions and 6–66,361 nucleotides for deletions (a full list of predicted indels is available at http://compbio.cs.toronto.edu/modil/).

The genome of the same individual was previously sequenced to 0.3-fold coverage using Sanger sequencing[7], allowing for discovery of a small fraction of the short indels in the genome. We estimated the false negative rate of our approach by computing the fraction of these known indels that were missed by our method, but had 20 overlapping clones in our dataset. The sensitivity of our approach varied widely depending on the indel size, but was >95% for indels ≥20 base pairs (**Table 1**).

Because MoDIL does not observe the indels directly the predicted indel size is an approximation of the true size. To verify the accuracy of MoDIL indel size estimates, we compared them to the sizes of overlapping indels from the Mills dataset[5]. The sizes were extremely highly correlated with a large number of indels of ~300–350 bp owing to *Alu* mobile elements (**Fig. 1c**). As expected, the difference between the true size of an indel and our predicted size followed a Gaussian distribution (**Fig. 1d**) with a mean of zero and variance inversely proportional to the number of mate pairs in the cluster. Together, these results indicate that MoDIL accurately recovered smaller variants than was previously possible using high clone coverage of short-read sequencing technologies.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We acknowledge Canadian Institutes of Health Research Catalyst grant to M.B. for funding, and Q. Morris and A. Valouev for useful discussions.

Seunghak Lee[1], Fereydoun Hormozdiari[2], Can Alkan[3] & Michael Brudno[1,4]

[1]Department of Computer Science, University of Toronto, Toronto, Canada. [2]School of Computing Science, Simon Fraser University, Burnaby, Canada. [3]Department of Genome Sciences, University of Washington and the Howard Hughes Medical Institute, Seattle, Washington, USA. [4]Banting and Best Department of Medical Research, University of Toronto, Toronto, Canada. e-mail: brudno@cs.toronto.edu

PUBLISHED ONLINE 31 MAY 2009; DOI:10.1038/NMETH.F.256

1. Li, H., Ruan, J. & Durbin, R. *Genome Res.* **18**, 1851–1858 (2008)
2. Hormozdiari, F. *et al. Genome Res.* (in the press).
3. Korbel, J.O. *et al. Genome Biol.* **10**, R23 (2009).
4. Lee, S., Cheran, E. & Brudno, M. *Bioinformatics* **24**, i59–i67 (2008).
5. Mills, R.E. *et al. Genome Res.* **16**, 1182–1190 (2006).
6. Bentley, D.R. *et al. Nature* **456**, 53–59 (2008).
7. Kidd, J.M. *et al. Nature* **453**, 56–64 (2008).