BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 15 2011, pages 2047–2053
doi:10.1093/bioinformatics/btr344

*Sequence analysis*

Advance Access publication June 7, 2011

# A probabilistic method for the detection and genotyping of small indels from population-scale sequence data

Vikas Bansal[1,*] and Ondrej Libiger[1,2]

[1]Scripps Genomic Medicine, Scripps Translational Science Institute and [2]Department of Molecular and Experimental Medicine, The Scripps Research Institute, La Jolla, CA 92037, USA

Associate Editor: Alex Bateman

## ABSTRACT

**Motivation:** High-throughput sequencing technologies have made population-scale studies of human genetic variation possible. Accurate and comprehensive detection of DNA sequence variants is crucial for the success of these studies. Small insertions and deletions represent the second most frequent class of variation in the human genome after single nucleotide polymorphisms (SNPs). Although several alignment tools for the gapped alignment of sequence reads to a reference genome are available, computational methods for discriminating indels from sequencing errors and genotyping indels directly from sequence reads are needed.

**Results:** We describe a probabilistic method for the accurate detection and genotyping of short indels from population-scale sequence data. In this approach, aligned sequence reads from a population of individuals are used to automatically account for context-specific sequencing errors associated with indels. We applied this approach to population sequence datasets from the 1000 Genomes exon pilot project generated using the Roche 454 and Illumina sequencing platforms, and were able to detect a significantly greater number of indels than reported previously. Comparison to indels identified in the 1000 Genomes pilot project demonstrated the sensitivity of our method. The consistency in the number of indels and the fraction of indels whose length is a multiple of three across different human populations and two different sequencing platforms indicated that our method has a low false discovery rate. Finally, the method represents a general approach for the detection and genotyping of small-scale DNA sequence variants for population-scale sequencing projects.

**Availability:** A program implementing this method is available at http://polymorphism.scripps.edu/~vbansal/software/piCALL/

**Contact:** vbansal@scripps.edu

**Supplementary information:** Supplementary data are available at Bioinformatics online.

Received on March 3, 2011; revised on May 27, 2011; accepted on June 1, 2011

## 1 INTRODUCTION

Small insertions and deletions (1–50 bp) represent the second most frequent class of variation in the human genome after single nucleotide polymorphisms (SNPs). Although indels exhibit greater potential to disrupt functional elements than SNPs, indels have been characterized to a significantly lesser extent, and are greatly underrepresented in public variation databases. There are two main reasons for this: (i) SNPs are more abundant in the genome and are easier to genotype using arrays and (ii) indels are more difficult to detect reliably using sequencing data. Sequencing of the first individual genome using the Sanger method by Levy *et al.* (2007) demonstrated that short indels are frequent in the human genome. Previously, Bhangale *et al.* (2005) performed a population-scale characterization of short indels in the human genome by resequencing 330 genes in 47 individuals using the Sanger method. Mills *et al.* (2006) generated a genome-wide map of indels by analyzing Sanger resequencing data. In recent years, several high-throughput DNA sequencing technologies have emerged that are able to generate gigabases of DNA sequence data in a single instrument run. These next-generation sequencing (NGS) methods offer greater potential to detect indels compared with Sanger sequencing since each allele is sequenced independently multiple times. Indeed, whole-genome shotgun sequencing of several individuals using high-throughput platforms has demonstrated the presence of several hundred thousands of short indel variants per genome (Bentley *et al.*, 2008; McKernan *et al.*, 2009; Wang *et al.*, 2008; Wheeler *et al.*, 2008).

Indels can be identified from sequence reads by performing gapped alignment of reads to a reference sequence. Several alignment tools developed for next-generation sequencing data allow gapped alignments (Homer *et al.*, 2009; Li and Durbin, 2009; Li *et al.*, 2008; Rumble *et al.*, 2009) [see Li and Homer (2010) for a survey of sequence alignment algorithms]. After reads have been aligned, the presence of multiple reads that support the same insertion/deletion event can be used to call indels. However, it can often be challenging to identify moderately long indels from short sequence reads since each read is aligned independently to the reference genome. Further, reads that span an indel event close to the ends of a read are difficult to align with gaps and can potentially be misaligned resulting in false SNPs (Krawitz *et al.*, 2010). The problem of misaligned reads can be alleviated by locally realigning reads using information from reads that are informative about the presence of an indel. Many tools for doing local realignment of short reads have recently been developed (Albers *et al.*, 2011; Homer and Nelson, 2010; McKenna *et al.*, 2010).

Once reads have been aligned to a reference sequence as accurately as possible, the next challenge is to distinguish true indels from sequencing errors and alignment artifacts. Several methods have been developed to detect single nucleotide variants using sequence data generated from individual genomes and

[*]To whom correspondence should be addressed.

successfully applied to identify millions of SNPs from whole-genome sequencing projects (Li *et al.*, 2008, 2009; McKenna *et al.*, 2010). Most of these methods utilize the base quality values for each individual base-call within a Bayesian framework to calculate the likelihood of an individual being heterozygous or homozygous for a variant allele at each site. Distinguishing true indels from sequencing errors is difficult since there are no accurate sequencing error models for indel errors. For example, base qualities for Illumina reads represent the accuracy of each individual base call but are not informative about indel errors. It is well known that insertion/deletion errors are dependent on the local sequence context, e.g. the 454/Roche sequencing platform is prone to sequencing errors in homopolymer runs (Wheeler *et al.*, 2008). The Illumina sequencing-by-synthesis technology reads individual bases in each sequencing cycle and is, therefore, less susceptible to insertion/deletion errors. However, non-sequencing artifacts such as polymerase slippage during PCR amplification of DNA molecules can result in insertion/deletion sequencing errors in homopolymer stretches as well as microsatellites (Shinde *et al.*, 2003). DNA insertion/deletion polymorphisms are known to be more frequent in such regions for primarily the same reason. Recently, Albers *et al.* (2011) developed a Bayesian method called Dindel to detect indels from short read sequence data generated using the Illumina sequencing platform. Dindel uses a probabilistic realignment model along with estimates of indel error rates from the 1000 Genomes project to distinguish indels from sequencing errors.

Accurately detecting indels from NGS data while accounting for indel errors that are dependent on the local sequence context and the sequencing platform remains difficult. In this article, we describe a probabilistic method that leverages sequence reads from a population of diploid individuals (sequenced using the same sequencing platform) to accurately detect indel variants. High-throughput sequencing technologies are increasingly being used for sequencing the genomes of populations of individuals with the goal of identifying rare and common DNA sequence variants. The availability of population-scale sequence data not only poses new challenges for variant detection methods, but also creates opportunities for the development of new approaches to variant detection. For indel detection, we reasoned that given aligned sequence reads from a population of individuals at any given site: (i) the context-specific sequencing error rate should be similar across all individuals; (ii) individuals who are homozygous for the reference allele should be informative about the average sequencing error rate; and (iii) for individuals who carry one or two copies of the variant allele, the fraction of the reads that support the variant allele should be significantly greater than the background sequencing error rate. We model the unknown sequencing error rate and the population genotypes at each position and calculate the likelihood of the population sequence data conditional on the population genotypes by integrating over the error rate. To quantify the evidence for the presence of a variant allele at a position, we compare the likelihood of the most likely genotype configuration in the presence of a variant allele with the population likelihood in the absence of a variant allele, i.e. when all individuals are homozygous for the reference allele. Positions for which this likelihood ratio is above a threshold are identified as variant sites. Additionally, information about the distribution of reads on the forward and reverse strands is used by modeling the sequencing error rates independently to further improve the accuracy of detecting variants.

Unlike previous methods for indel detection, our method does not require prior knowledge of context-specific indel error rates and is applicable to population sequence data from different sequencing platforms. To demonstrate the accuracy of our method, we utilize population sequencing datasets generated by exon sequencing in the 1000 Genomes project using the Roche 454 and Illumina sequencing platforms. Across 7 population sequencing datasets, our method identified 261 distinct indels, significantly greater than the number of indels reported in the 1000 Genomes project variant calls (Durbin *et al.*, 2010). Comparison of our indel calls to indels identified in the 1000 Genomes project demonstrated a high sensitivity of 95–100% across different populations. Analysis of the distribution of lengths of indels revealed an excess of indels whose length is a multiple of 3 ($3n$) across multiple populations and the two sequencing platforms. This increased frequency of $3n$ indels and the consistency in the number of indels detected across multiple populations and two different sequencing platforms suggest that our method has good specificity.

## 2 METHODS

Short indels can be detected by performing gapped alignments of reads to a reference sequence and various tools for doing this are available. Most alignment methods align each read independently to the reference sequence. The presence of multiple reads supporting the same insertion/deletion variant can be used to distinguish real indels from sequencing and alignment artifacts. However, reads that contain an indel (with respect to the reference sequence) near the beginning or the end of the read are typically aligned without gaps resulting in alignments with multiple mismatches. To improve the alignment of reads that span indels, one could use a realignment method to improve the initial alignment with respect to indels. Realignment of reads can make a significant difference in the subsequent detection of indels for short read sequence data but is likely to be less important for longer reads. Our objective is to utilize aligned sequence reads from a population of individuals to identify sites that harbor an insertion/deletion variant. Therefore, we assume that the reads for each individual have been aligned to the corresponding reference sequence as accurately as possible.

### 2.1 Probabilistic model for population indel detection

We consider a set of genomic loci that have been resequenced in a population of $n$ diploid individuals. Our objective is to identify positions in the sequenced region for which at least one of the $2n$ haplotypes harbors an insertion/deletion of one or more bases. For each site, we denote the reference allele by $A_0$ and the alternate allele $A_1$. If there are multiple potential variant alleles, we evaluate each allele individually. Given aligned sequence data from $n$ individuals, we denote the ordered set of population genotypes by $G=[G_1,G_2,...,G_n]$. Assuming two alleles, $A_0$ and $A_1$, we denote the three possible genotypes as $(A_0A_0),(A_0A_1)$ and $(A_1A_1)$. In the absence of a variant allele, $G_i=(A_0A_0)$ for all $i$. Let $G^0=[(A_0A_0),(A_0A_0),...,(A_0A_0)]$ represent the genotype configuration in which each individual is homozygous for the allele $A_0$. Similarly, we define $G^1=[(A_1A_1),(A_1A_1),...,(A_1A_1)]$. The aligned sequence reads for all individuals at each site represent the observed data $D=[D_1,...,D_n]$ where $D_i (1\leq i\leq n)$ is the set of aligned reads that cover the given site for individual $i$.

Given the population sequence data $D$ at each position in the genome, we can write $Pr(D)$ as

$$\sum_G Pr(D|G)Pr(G)=Pr(D|G^0)Pr(G^0)+\sum_{G'\neq G^0} Pr(D|G')Pr(G')$$

We define a likelihood ratio statistic as the ratio of the likelihood in the presence of a variant to the likelihood in the absence of a variant:

$$LR = \frac{\sum_{G' \neq G^0} Pr(D|G')Pr(G')}{Pr(D|G^0)Pr(G^0)} \tag{1}$$

In order to compute the likelihood ratio statistic, we need to define the conditional probabilities $Pr(D|G')$ and the prior probability $Pr(G')$ for each genotype configuration $G'$. If base quality values or sequencing error probabilities for each base-call are available, we can calculate the conditional genotype probabilities using these sequencing error probabilities (see e.g. Bansal *et al.*, 2010; Li *et al.*, 2008). However, for indels, such error rates are typically not available. In order to compute the conditional likelihoods, we introduce the parameter $e_{01}$ which corresponds to the probability of (incorrectly) reading the reference allele $A_0$ as the alternate allele $A_1$. One can think of this parameter as the average context-specific sequencing error rate at this particular position. For example, in a homopolymer run of 4 'T's where the alternate allele is 'TTT', this parameter would represent the probability of reading the reference allele 'TTTT' as 'TTT' due to sequencing errors. Similarly, we define the parameter, $e_{10}$ as the probability of reading the allele $A_1$ as the reference allele $A_0$ due to sequencing errors. If all individuals have been sequenced using the same sequencing platform, the context-specific sequencing error rate is expected to be similar from individual to individual.

To calculate the conditional likelihoods, we integrate over the unknown sequencing error rates as follows:

$$Pr(D|G') = \int Pr(D|G', e)\pi(e|G')de \tag{2}$$

where $e = (e_{01}, e_{10})$, $\pi(e|G')$ is the prior distribution of the sequencing error rates given the population genotype vector $G'$ and the integral is a double integral over the two variables $e_{01}$ and $e_{10}$ ($0 \leq e_{01} \leq 1$ and $0 \leq e_{10} \leq 1$). Since the sequencing error rates are independent of the presence of a variant, we can write $\pi(e|G') = \pi(e) = \pi(e_{01}) \times \pi(e_{10})$. Further, since sequence reads for each individual only affect the genotypes for that individual, we can write

$$Pr(D|G' = (G_1, G_2, \ldots, G_n), e) = \prod_{i=1}^{n} Pr(D_i|G_i, e) \tag{3}$$

Next, we describe how to calculate the conditional likelihood $Pr(D_i|G_i, e)$ for an individual.

## 2.2 Conditional likelihoods for an individual

Consider the set of reads $D_i$ covering a site in an individual $i$. We summarize the data as $r_{i0}$ and $r_{i1}$, the number of reads that support the two alleles $A_0$ and $A_1$, respectively. For a diploid individual, we consider the likelihoods for the three genotypes: $(A_0A_0), (A_0A_1)$ and $(A_1A_1)$. Assuming independence between sequencing errors from multiple reads, we can define the conditional probabilities for the three genotypes as:

$$Pr(D_i|G_i = (A_0A_0), e_{01}, e_{10}) = \binom{r_{i0}+r_{i1}}{r_{i0}}(1-e_{01})^{r_{i0}}(e_{01})^{r_{i1}} \tag{4}$$

$$Pr(D_i|G_i = (A_1A_1), e_{01}, e_{10}) = \binom{r_{i0}+r_{i1}}{r_{i1}}(1-e_{10})^{r_{i1}}(e_{10})^{r_{i0}} \tag{5}$$

$$Pr(D_i|G_i = (A_0A_1), e_{01}, e_{10}) = \binom{r_{i0}+r_{i1}}{r_{i0}}(h')^{r_{i0}}(1-h')^{r_{i1}} \tag{6}$$

where $h$ is the probability that an aligned read was sampled from the chromosome with the reference allele and $h' = h(1-e_{01}) + (1-h)e_{10}$ is the probability of observing a read with the $A_0$ allele given the genotype $G_i = (A_0A_1)$ . For SNPs, it is reasonable to assume equal probability of sampling the two chromosomes, i.e. $h = 0.5$. For indels, especially for long indels, reads that cover the insertion/deletion variant near the start or the end of the read are likely to be misaligned or not aligned. To account for this bias in favor of observing the reference allele, we set $h = 0.5 + \epsilon$ where $\epsilon$ is estimated using the length of the variant allele and the average length of sequence reads (see Supplementary Material for details).

## 2.3 Conditional data likelihoods for population of individuals

Now that we have defined the conditional likelihoods for each individual, we can calculate the probability of the population data $D$ conditional on a population genotype vector. For $G = G^0$, $G_i = (A_0A_0)$ for all individuals and from Equation (3) we have:

$$Pr(D|G^0, e_{01}, e_{10}) = \prod_{i=1}^{n}\binom{r_{i0}+r_{i1}}{r_{i0}}(1-e_{01})^{r_{i0}}(e_{01})^{r_{i1}}$$

To calculate $Pr(D|G^0)$, we need to integrate the above expression as defined in Equation (2). Since, the expression does not depend on the variable $e_{10}$, the integral reduces to

$$\prod_{i=1}^{n}\binom{r_{i0}+r_{i1}}{r_{i0}}\int_0^1 (1-e_{01})^{C_0} {e_{01}}^{C_1}\pi(e_{01})de_{01}$$

where $C_0 = \sum_i r_{i0}$ and $C_1 = \sum_i r_{i1}$.

For the error rate $e_{01}$, we choose a beta prior with parameters $\alpha$ and $\beta$. Therefore, $\pi(e_{01}) = \frac{(1-e_{01})^{\beta-1} {e_{01}}^{\alpha-1}}{B(\alpha,\beta)}$ and the integral can be written as

$$\prod_{i=1}^{n}\binom{r_{i0}+r_{i1}}{r_{i0}}\frac{\int_0^1 (1-e_{01})^{C_0+\beta-1} {e_{01}}^{C_1+\alpha-1} de_{01}}{B(\alpha,\beta)}$$

where $B$ is the $\beta$-function. The integrand in the above equation is the probability density function of the beta distribution with parameters $\alpha + C_1$ and $\beta + C_0$ scaled by the normalization constant $B(\alpha + C_1, \beta + C_0)$. Therefore, we have

$$Pr(D|G^0) = \prod_{i=1}^{n}\binom{r_{i0}+r_{i1}}{r_{i0}}\frac{B(\alpha+C_1, \beta+C_0)}{B(\alpha,\beta)} \tag{7}$$

For $G \neq G^0$, the integrand involves both parameters $e_{01}$ and $e_{10}$ and it is infeasible to evaluate the integral analytically. Therefore, we approximate it numerically by summing the value of the integrand over a 2D grid (see Supplementary material for details of the numerical integration and accuracy of the approximation).

For a given genotype $G$, the conditional likelihood $Pr(D|G)$ is calculated by integrating over the parameters $e_{01}$ and $e_{10}$. We also calculate a simple estimate for the sequencing error rate $e_{01}$ as $\frac{{C_0}'}{{C_0}' + {C_1}'}$ where ${C_0}'$ is the number of reads that support the allele $A_0$ summed over individuals with $G_i = (A_0A_0)$ and ${C_1}'$ is the number of reads supporting $A_1$.

## 2.4 Modeling strand-specific sequencing errors

Systematic sequencing errors, such as those that depend on the local sequence context, are likely to be strand specific, i.e. overrepresented on one of the two strands. Since the local sequence context on the two strands is different, we model the sequencing error rates on each strand using independent parameters. Therefore, we can rewrite Equation (2) as:

$$Pr(D|G') = \int\int Pr(D^f, D^r|G', e^f, e^r)\pi(e^f)\pi(e^r)de^f de^r$$

where $D^f$ represents the sequence reads for the population of individuals that align to the forward (+) strand and $D^r$ represents the set of sequence reads aligned to the reverse (−) strand. Also, $e^f = ((e_{01})^f, (e_{10})^f)$ represents the sequencing error rates on the forward strand and $e^r$ represents the error rates on the reverse strand. Since the sequencing error rates on one strand do not affect the conditional likelihood for the reads on the other strand, the above integral can be written as the product of two conditional likelihoods, one for each strand:

$$\int Pr(D^f|G', e^f)\pi(e^f)d(e^f) \times \int Pr(D^r|G', e^r)\pi(e^r)d(e^r)$$

Each of the two stranded conditional likelihoods can be calculated as before by using the data for the corresponding strand and integrating over the strand-specific error rates. Later, we show how the strand-specific conditional likelihoods can be used to filter out false variants for which virtually all the evidence for the presence of a variant allele is present on the reads from one strand. Li *et al.* (2010) have also demonstrated the utility of a strand-based filter to substantially reduce the number of false positive variants in analysis of mitochondrial DNA sequencing data.

## 2.5 Prior probabilities for population genotypes

Given a genotype configuration $G=(G_1,G_2,\ldots,G_n)$ for $n$ individuals, we want to calculate $Pr(G)$, the prior probability of the genotype configuration. Let $n_1$ be the number of $A_1$ alleles in the genotype $G$ and $Pr(n_1)$ be the probability of observing $n_1$ alleles of type $A_1$ in a sample of $n$ diploid individuals. Let $\theta$ be the population-scaled mutation rate. Using the allele frequency spectrum of a neutrally evolving population under the standard coalescent (Fu, 1995), we can write (see also Le and Durbin, 2011):

$$Pr(n_1)=\frac{\theta}{2}\left(\frac{1}{n_1}+\frac{1}{2n-n_1}\right)(0<n_1<2n) \tag{8}$$

Since

$$\sum_{G'}Pr(G')=\sum_{i=1}^{2n-1}Pr(n_1=i)+Pr(G=G^0)+Pr(G=G^1)=1$$

$$Pr(G=G^0)=Pr(G=G^1)=\frac{1}{2}\left(1-\theta\sum_{i=1}^{2n-1}\frac{1}{i}\right) \tag{9}$$

For a genotype configuration $G$, let $n_{00}$, $n_{01}$ and $n_{11}$ be the number of individuals with genotypes $(A_0A_0)$, $(A_0A_1)$ and $(A_1A_1)$ in $G$. Under the assumption of Hardy–Weinberg equilibrium, the probability of observing $n_{01}$ heterozygotes in a sample of $n$ diploid individuals with $n_1$ $A_1$ alleles is [see Equation (1) (Wigginton *et al.*, 2005)]

$$Pr(n_{01}\ \text{hets}|n_1)=\frac{2^{n_{01}}\binom{n}{n_{00}n_{01}n_{11}}}{\binom{2n}{n_1}}$$

Since there are $\binom{n}{n_{00}n_{01}n_{11}}$ distinct genotype vectors with $n_{01}$ heterozygotes, the probability of each such genotype vector is:

$$Pr(G)=\frac{2^{n_{01}}}{\binom{2n}{n_1}}\times\frac{\theta}{2}\left(\frac{1}{n_1}+\frac{1}{2n-n_1}\right) \tag{10}$$

for $G\neq G^0$ and $G\neq G^1$.

## 2.6 Calculating likelihood ratio statistic

For each potential variant site, we want to evaluate the likelihood ratio defined in Equation (1). Using Equations (9) and (10), we can calculate the prior probability for any genotype $G$. The conditional likelihood $Pr(D|G^0)$ for the reference genotype $G^0$ can also be calculated analytically using Equation (7). However, the numerator involves summing over the conditional likelihoods for an exponential number of possible genotypes. Further, calculating the conditional likelihood for an individual genotype is computationally expensive since it involves a numerical integration over two variables. To avoid summing over a large number of genotypes, we assume that the posterior genotype likelihood is concentrated around the most likely genotype configuration. Therefore, we can approximate the sum by $\max_{G'}Pr(D|G')Pr(G')$. This is a reasonable approximation if each sample has sufficient sequence coverage but could reduce the power to detect variants when coverage is low. In order to determine the genotype configuration $G^{max}$ for which $Pr(D|G')Pr(G')$ is maximum, we use a simple greedy algorithm where we start from an initial genotype configuration and iteratively update the genotype configuration until the genotype likelihood can be increased. Sites for which the likelihood ratio statistic is above a threshold are identified as candidate variants.

To filter out potentially false variants for which the evidence for the presence of a variant allele is strand specific, we compute the ratios $\frac{Pr(D^f|G^{max})}{Pr(D^f|G^0)}$ and $\frac{Pr(D^r|G^{max})}{Pr(D^r|G^0)}$ for the forward and reverse strands. Variants for which either of the two ratios is less than 1 are likely artifacts of strand-specific sequencing errors.

*Algorithm for population indel detection and genotyping (piCALL)*:

```
(1) For each position and each potential variant allele:
    (a) determine the allele counts for each individual $i$ for the two alleles $A_0$ and $A_1$ and the two strands
    (b) Compute an inital genotype configuration $G^{start}$ by calculating the posterior genotype likelihoods $Pr(G_i^{start}=g|D_i)$ ($g=(A_0A_0)$, $(A_0A_1)$ and $(A_1A_1)$) for each individual
    (c) set $G^{max}=G^{start}$, updates =1
    (d) while updates $>0$:
        (1) updates = 0
        (2) for $i=1$ to $n$
            • calculate $Pr(D|G)Pr(G)$ for all genotypes $G$ such that $G_j=G_j^{max}$ $(i\neq j)$ and $G_i=\{(A_0A_0),(A_0A_1),(A_1A_1)\}$
            • determine genotype $G^*$ for which $Pr(D|G)Pr(G)$ is maximum
            • if $Pr(D|G^*)Pr(G^*)>Pr(D|G^{max})Pr(G^{max})$: set $G^{max}=G^*$, updates = updates $+1$
    (e) $LLR=log10\left(\frac{Pr(D|G^{max})Pr(G^{max})}{Pr(D|G^0)Pr(G^0)}\right)$
    (f) if $LLR>=thresh$:
        • strandfilter = min$\left(\frac{Pr(D^f|G^{max})}{Pr(D^f|G^0)},\frac{Pr(D^r|G^{max})}{Pr(D^r|G^0)}\right)$
        • if strandfilter $>0.5$: output variant
```

# 3 RESULTS

## 3.1 Sequence data from 1000 Genomes project

We assessed the performance of piCALL using population sequencing data generated by the 1000 Genomes project (exon sequencing) (Durbin *et al.*, 2010). In this project, 8140 exons from 906 randomly selected genes were captured using multiple target capture technologies and subsequently sequenced at high coverage using the Roche 454 and Illumina GA sequencing platforms in 697 individuals. We utilized a subset of the sequence data from individuals of European (CEU), East Asian (CHB and CHD), and African (YRI) ancestry. For each individual, we downloaded aligned sequence reads (in BAM format) from the 1000 Genomes web site (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/data/). For samples sequenced using the 454 technology, we utilized alignments generated using the SSAHA aligner (Ning *et al.*, 2001), while for samples sequenced on the Illumina GA, we used the MAQ (Li *et al.*, 2008) alignments. The Illumina samples were sequenced using a mix of read lengths (ranging from 35 to 100 bp) and both single-end and paired-end reads. For samples sequenced using only single-end reads, we aligned the reads to the reference human genome using BWA (Li and Durbin, 2009) to allow for detection of indels. Further, to improve the alignment of reads that span insertion/deletion variants, we utilized the realignment module of the Genome Analysis Toolkit (GATK) (McKenna *et al.*, 2010) to realign the reads for each individual. The realigned BAM files for each sample were used for indel detection.

**Table 1.** Indels identified using piCALL on population sequence datasets from the 1000 Genomes Project

| Population | Platform | Individuals | No. of indels detected | | | 1000G indels |
|---|---|---|---|---|---|---|
| | | | Unfiltered | $e \leq 0.03$ | $e \leq 0.02$ | |
| CEU | Illumina | 66 | 60 (30) | 59 (30) | 58 (29) | 24 |
| CHB+CHD | Illumina | 73 | 63 (29) | 60 (29) | 60 (29) | 39 |
| YRI | Illumina | 70 | 68 (39) | 68 (39) | 68 (39) | 38 |
| CEU | Roche 454 | 40 | 166 (18) | 82 (18) | 47 (16) | 24 |
| CHB | Roche 454 | 63 | 102 (28) | 69 (28) | 60 (28) | 31 |
| CHD | Roche 454 | 78 | 68 (28) | 62 (27) | 58 (24) | 26 |
| YRI | Roche 454 | 45 | 74 (36) | 58 (36) | 57 (36) | 38 |

For each population, the number in parenthesis denotes the number of 3n indels. The column '1000G indels' is the number of indels called in the 1000 Genomes project for each population combined across the two sequencing platforms.

## 3.2 Detection of indels using piCALL

To analyze aligned reads from multiple samples using piCALL, we generated pileup files (distinct from the samtools pileup format) that represented the aligned bases covering each position. For indel detection, we only considered reads with a mapping quality of 20 or more. For each position, we determined candidate indels using the aligned reads for all samples in a population and evaluated each candidate indel using the algorithm piCALL as described in the Methods. For the beta prior on the error probabilities $e_{01}$ and $e_{10}$, we used parameters $\alpha = 1$ and $\beta = 30$. We also ran piCALL with different values for $\beta$ (25, 50) and $\alpha = 1$ and observed only minor changes in the set of the indels identified by piCALL. For the log-likelihood ratio statistic, we used a threshold of four to identify indels. Choosing a lower cutoff for the statistic is expected to increase the sensitivity and decrease the specificity of the indel calls. In the absence of a dataset with perfect knowledge of the indel variants, we used a higher threshold to increase specificity. Candidate indels for which the log-likelihood ratio statistic was above the threshold value and for which the evidence from the two strands was not conflicting (strandfilter) were retained. The number of indels that were identified in each population are reported in Table 1 along with the number of samples. For the CEU and CHB population samples sequenced using the Roche 454 platform, the number of identified indels was unusually large compared with other populations. Further analysis of the indels in these populations revealed an excess of 1 bp indels in homopolymer runs. For each indel, we calculated the mean estimate of the sequencing error rates ${e_{01}}^f$ and ${e_{01}}^r$ (see Section 2.3) using the final genotype configuration. Indels for which the sequencing error rate $e$ (average of the two strand-specific error rates) was above a threshold were removed. The number of indels in each population that passed this additional filter are also shown in Table 1 ($e \leq 0.03$ and $e \leq 0.02$). The number of indels detected for the Illumina sequenced datasets was virtually unaffected by this filter while the number of indels for the Roche 454 datasets was considerably reduced. However, the total number of indels detected across different populations and sequencing platforms was more consistent, suggesting that this filter removed false single base pair indels called from Roche 454 data.

In total, 408 indels were identified by piCALL across the seven population datasets ($e <= 0.02$). In coding regions, short insertions and deletions that cause a frameshift are likely to affect gene function. Therefore, coding indels whose length is not a multiple of 3, i.e. the length of a codon, are expected to be under purifying selection. The 204/408 (50%) indels had a length that was a multiple of 3 (see Fig. 1a for distribution of indel lengths for the two sequencing platforms). This bias in favor of $3n$ indels was consistent with previous studies of coding indels (Bhangale *et al.*, 2005; Ng *et al.*, 2009) and indicated that the set of indels did not contain many false positives. Interestingly, the proportion of $3n$ indels for the YRI population detected on both the Illumina (57%) and Roche 454 (63%) sequencing platforms was higher than in the CEU and Asian populations (Fig. 1b). The higher proportion of $3n$ coding indels in African populations is consistent with the finding of Lohmueller *et al.* (2008) that European populations harbor a greater proportion of non-synonymous SNPs as compared with African populations.

## 3.3 Comparison to indel calls in 1000 Genomes project

To assess the sensitivity of our indel calls, we compared our data with indel calls downloaded from the 1000 Genomes project web site. Since the 1000 Genomes project calls were made for each population by combining data across multiple platforms, we merged the indels identified by piCALL for each of three populations: CEU, YRI and Asian (CHB + CHD) across the two platforms. For the CEU population, 22 of the 24 indels reported by the 1000 Genomes data were also called by piCALL. Similarly, for the YRI population, 34/38 indels were identified. For the Asian samples, 33 of 39 indels were detected. We further examined the 12 indels called in the 1000 Genomes data that were missed by piCALL. Of these, three were confirmed in validation experiments, four did not validate while the remaining five indels were not evaluated (Durbin *et al.*, 2010).

We also compared indels detected from the CEU and YRI populations sequenced using the Roche 454 platform to indel calls in the trios sequenced by the 1000 Genomes project. The individuals in the YRI trio (NA19240, NA19238 and NA19239) and the CEU trio (NA12878, NA12891 and NA12892) were sequenced using the Roche 454 platform. For the YRI trio, 17 indels overlapped the genomic regions targeted in the exon population sequencing. In all, 16 of these were identified by piCALL in the YRI Roche 454 dataset. The one indel that was missed was in an exon that had low coverage in the population sequence data. Similarly, for the CEU trio, all nine indels that overlapped the sequenced exons were identified. Overall, these results demonstrated the high sensitivity of our method.

The number of indels identified by piCALL for the 1000 Genomes datasets was significantly greater than reported by Durbin *et al.* (2010). The 408 indels identified across the seven population datasets correspond to 261 distinct insertion/deletion variants. In all, 62 of these matched indels identified in the 1000 Genomes exon project. An additional 24 indels were detected in two or more populations. While it is difficult to estimate the false discovery rate without experimental validation, the increased frequency of $3n$ indels and the consistency in the number of indels detected across different populations and two different sequencing platforms suggest that our method has high specificity.

## 4 CONCLUSIONS AND FUTURE WORK

Insertions and deletions represent an important class of small-sequence variation that can be identified from high-throughput sequencing data. Hundreds of exomes have been sequenced (Li

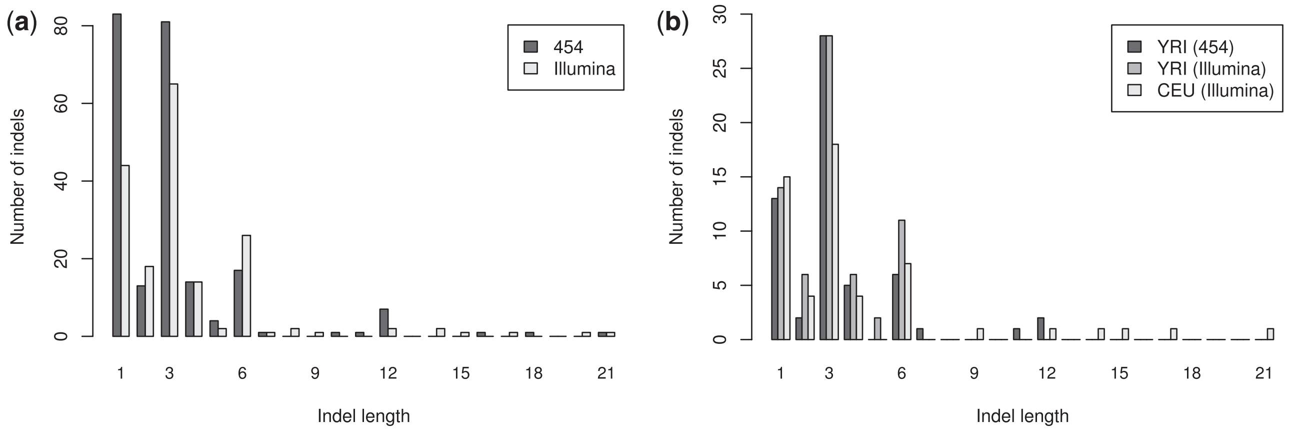

**Fig. 1.** The distribution of lengths of indels identified by piCALL on the 1000 Genomes datasets. (**a**) Distribution of indel lengths identified using samples sequenced using the Roche 454 and Illumina sequencing platform. (**b**) Comparison of indel lengths between the YRI (454 and Illumina) and CEU (Illumina only) populations.

*et al.*, 2010) and several thousand exomes are being sequenced for the comprehensive discovery of coding genetic variation and to identify rare disease susceptibility variants. In coding regions, accurate and sensitive detection of indels is important since missed indels can reduce power to detect the functional mutations while false indels increase the number of candidate mutations that could be functional. However, this requires (i) comprehensive identification of candidate indel variants from sequence reads, (ii) accurate alignment of all reads that support an insertion/deletion event and (iii) probabilistic method to discriminate true indels from indels that are artifacts of sequencing or alignment errors. We have presented a probabilistic method for the accurate detection and genotyping of indels from aligned population sequence data that automatically accounts for context-dependent sequencing errors. Application of this method to population sequence data from the 1000 Genomes project demonstrated the sensitivity and specificity of our method.

The method does not realign reads unlike indel callers such as Dindel which combine realignment with indel detection. Therefore, the accuracy of our method depends on the quality of the initial alignments. The read lengths for the Illumina sequencing platform are increasing and with longer reads, it should be easier to align reads with gaps and substantially reduce the impact of misaligned reads. A key feature of our method is that it does not require prior knowledge of the context-specific sequencing error rates, although prior information about the error rates can be easily incorporated in the model. As new sequencing platforms emerge, methods for indel detection that can automatically account for the platform-specific sequencing error profiles will be very useful.

The probabilistic model underlying piCALL uses the number of reads supporting each allele for calculating the likelihoods. It is feasible to incorporate information about base quality values and read mapping quality in the calculation of the conditional genotype likelihoods. This would further improve the accuracy of indel detection but would be even more important for calling single nucleotide variants using the same model. Also, as currently implemented, the method evaluates each potential variant allele at each site individually. Although this allows for the detection of multiallelic variants, a more accurate approach would be able handle all potential variant alleles simultaneously. In the future, we plan to extend the model and implementation of piCALL to incorporate base quality values and analyze multiple variant alleles simultaneously.

We have implemented piCALL in C for computational efficiency. The method is compatible with data from different sequencing platforms but requires all samples to be sequenced using the same sequencing platform. In addition, the method requires sequence data from a sufficient number of samples in order to accurately estimate the population genotypes. The running time of piCALL is proportional to the number of individuals and the number of candidate indels evaluated using the likelihood ratio statistic. To call indels for the seven population datasets from the 1000 Genomes project, piCALL took ~4 h on a 8 core linux machine.

*Funding*: This work was supported by the Scripps Translational Science Institute Clinical Translational Science Award (National Institutes of Health 5UL1RR025774).

*Conflict of Interest*: none declared.

## REFERENCES

Albers,C.A. *et al.* (2011) Dindel: accurate indel calls from short-read data. *Genome Res.*, **21**, 961–973 .

Bansal,V. *et al.* (2010) Accurate detection and genotyping of SNPs utilizing population sequencing data. *Genome Res.*, **20**, 537–545.

Bentley,D.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Bhangale,T.R. *et al.* (2005) Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum. Mol. Genet.*, **14**, 59–69.

Durbin,R.M. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Fu,Y.X. (1995) Statistical properties of segregating sites. *Theor. Popul. Biol.*, **48**, 172–197.

Homer,N. and Nelson,S.F. (2010) Improved variant discovery through local re-alignment of short-read next-generation sequencing data using SRMA. *Genome Biol.*, **11**, R99.

Homer,N. *et al.* (2009) BFAST: an alignment tool for large scale genome resequencing. *PLoS One*, **4**, e7767.

Krawitz,P. *et al.* (2010) Microindel detection in short-read sequence data. *Bioinformatics*, **26**, 722–729.

Le,S.Q. and Durbin,R. (2011) SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.*, **21**, 952–960.

Levy,S. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. and Homer,N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinformatics*, **11**, 473–483.

Li,H. *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,R. *et al.* (2009) SNP detection for massively parallel whole-genome resequencing. *Genome Res.*, **19**, 1124–1132.

Li,M. *et al.* (2010) Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. *Am. J. Hum. Genet.*, **87**, 237–249.

Li,Y. *et al.* (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat. Genet.*, **42**, 969–972.

Lohmueller,K.E. *et al.* (2008) Proportionally more deleterious genetic variation in European than in African populations. *Nature*, **451**, 994–997.

McKenna,A. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.

McKernan,K.J. *et al.* (2009) Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two base encoding. *Genome Res.*, **19**, 1527–1541.

Mills,R.E. *et al.* (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.*, **16**, 1182–1190.

Ng,S.B. *et al.* (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, **461**, 272–276.

Ning,Z. *et al.* (2001) SSAHA: a fast search method for large DNA databases. *Genome Res.*, **11**, 1725–1729.

Rumble,S.M. *et al.* (2009) SHRiMP: accurate mapping of short color-space reads. *PLoS Comput. Biol.*, **5**, e1000386.

Shinde,D. *et al.* (2003) Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. *Nucleic Acids Res.*, **31**, 974–980.

Wang,J. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.

Wheeler,D.A. *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature*, **452**, 872–876.

Wigginton,J.E. *et al.* (2005) A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.*, **76**, 887–893.