BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 6 2010, pages 722–729
doi:10.1093/bioinformatics/btq027

Sequence analysis

Advance Access publication February 9, 2010

Microindel detection in short-read sequence data

Peter Krawitz[1,2,3,†], Christian Rödelsperger[1,2,3,†], Marten Jäger[1,3], Luke Jostins[4], Sebastian Bauer[1,3] and Peter N. Robinson[1,2,3,*]

[1]Institute for Medical Genetics, Charité-Universitätsmedizin Berlin, Augustenburger Platz 1, 13353 Berlin, [2]Berlin-Brandenburg Center for Regenerative Therapies, Augustenburger Platz 1, 13353 Berlin, [3]Max Planck Institute for Molecular Genetics, Ihnestrasse 63-73, 14195 Berlin and [4]Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Several recent studies have demonstrated the effectiveness of resequencing and single nucleotide variant (SNV) detection by deep short-read sequencing platforms. While several reliable algorithms are available for automated SNV detection, the automated detection of microindels in deep short-read data presents a new bioinformatics challenge.

Results: We systematically analyzed how the short-read mapping tools MAQ, Bowtie, Burrows-Wheeler alignment tool (BWA), Novoalign and RazerS perform on simulated datasets that contain indels and evaluated how indels affect error rates in SNV detection. We implemented a simple algorithm to compute the equivalent indel region *eir*, which can be used to process the alignments produced by the mapping tools in order to perform indel calling. Using simulated data that contains indels, we demonstrate that indel detection works well on short-read data: the detection rate for microindels (<4 bp) is >90%. Our study provides insights into systematic errors in SNV detection that is based on ungapped short sequence read alignments. Gapped alignments of short sequence reads can be used to reduce this error and to detect microindels in simulated short-read data. A comparison with microindels automatically identified on the ABI Sanger and Roche 454 platform indicates that microindel detection from short sequence reads identifies both overlapping and distinct indels.

Contact: peter.krawitz@googlemail.com; peter.robinson@charite.de

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

Received on September 22, 2009; revised on January 14, 2010; accepted on January 16, 2010

1 INTRODUCTION

Microinsertions and microdeletions (‘indels’) constitute a class of genetic mutations that play an important role in human genetic disease (Ball *et al.*, 2005). The reliable detection of microinsertions and microdeletions is thus a prerequisite for current efforts to assess the medical relevance of genetic variation including small indels across the human genome. Structural variations on the order of kilobases, whose prevalence has long been underestimated because of the lack of appropriate methods of detection, have been recently shown to be responsible for more polymorphism than single nucleotide variants (SNVs) as measured by nucleotide content per genome (Korbel *et al.*, 2007; Redon *et al.*, 2006). Therefore, we hypothesized that also on the scale of only a few nucleotides, the frequency of microindels might have been underestimated. Harismendy *et al.* (2009) performed an analysis on sequences amplified by long-range PCR to compare three next-generation sequencing (NGS) platforms (Mardis, 2008), Illumina GA, Roche 454 FLX and ABI SOLiD, to the *de facto* gold standard of ABI Sanger sequencing. All three NGS platforms showed high sensitivity (>95%) in variant calling for sequence sites covered to saturation. However, for microindel detection, they only compared the results of the automated microindel detection pipeline of the Roche 454 platform to ABI Sanger sequencing. Currently only few software solutions for microindel detection in short read sequence data are available and they do not yet meet the need for unambiguous microindel positioning. In addition, the evaluation of automated microindel detection on NGS data remains difficult, as a gold standard is lacking—on the ABI Sanger platform the automated detection of heterozygous microindels remains highly error prone (Bhangale *et al.*, 2005). In the targeted sequence analyzed in Harismendy *et al.* (2009), 11 indels were identified by ABI Sanger, whereas 43 additional indels that were not found by ABI Sanger were called by Roche 454. On the other hand, there were five single-base indels in homopolymers that were called by ABI Sanger but not by Roche 454. This illustrates that the existing approaches to automated detection of microindels remain a technological challenge with presumably high false positive and negative rates.

Because of the high error rates of automated indel detection with the ABI Sanger platform and the Roche 454 technology, we were motivated to study the potential to identify microindels in short-read sequence data produced on NGS platforms. We introduce a simple indel calling algorithm that is based on the efficient mapping of short reads and which takes into account the fact that short sequence reads containing indels may often not be unequivocally aligned to the reference genome due to the surrounding sequence. Our microindel calling algorithm makes use of gapped alignments produced by efficient short-read mapping tools, such as Burrows-Wheeler alignment tool (BWA), Novoalign or RazerS, in order to call SNVs and indels. As the true distributions and frequencies of microindels in genomic sequences remain unknown due to technological shortcomings, we use a simulation approach to perform an analysis for varying microindel sizes and frequencies and study the effect on SNV

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

as well as microindel detection. Finally, we apply our microindel calling approach for short-read data to the aforementioned cross-platform-validated datasets (Harismendy *et al.*, 2009) and show that microindel detection on Illumina GA short-read sequences is feasible. Although our approach is applicable to any sort of short-read data, we focus in the following analysis on short-read data format of the Illumina GA platform, as the greatest variety of efficient mapping tools is available for this technology.

2 METHODS

2.1 Simulating short-read data

The 296 kb reference sequence used for generating sequence reads was constructed by extracting the following sequences from the human genome using the UCSC Genome Bioinformatics site (genome.ucsc.edu) build hg18 (NCBI build 36): chr11:73836950-73862566, chr21:34656259-34672486, chr21:34734911-34819450, chr2:223615214-223628218, chr3:38553978-38665289 and chr7:150268610-150312098. Serial repeats in the sequence fragments were detected with mreps 2.5 (Kolpakov *et al.*, 2003). Altogether 296 repeats of length ≥ period +9 were reported for the targeted sequence; this is comparable with the repeat frequency of randomly chosen 300 kb sequence fragments of human DNA (e.g. there are ∼260 000 repeats in 330 Mb of chr. 1). Deletions and insertions were simulated with frequencies ranging from 0.1 microindels/kb to 10 microindels/kb. The definition of microindels with respect to size varies in the literature and has often been defined in accordance with the detection limits of the technologies used in a study. We studied microindels up to a length of 5 bp for 36 bp reads and up to a length of 10 bp for 76 bp reads. Single nucleotide polymorphisms (SNPs) were simulated with a fixed frequency of 1 SNP/kb in all datasets. Microindels and SNPs were simulated as homozygous or heterozygous variants with a rate of 0.5.

In the simulated sequence, the positions of microindels were randomly chosen with the constraint that neighboring deletions may not overlap. This means that in simulated dataset with microindels of size k bp the positions of two different deletions must at least lie $k+1$ nt apart.

The datasets studied in Harismendy *et al.* (2009) show a distribution of the sequencing depth per chromosomal position that is platform specific. The datasets produced on the Genome Analyzer had a mean sequencing depth of 180-fold. We therefore simulated reads such that the read depth at each chromosomal position follows a Poisson distribution with a mean of 180. For the mapping statistics, the original position of every read with respect to the reference sequence was added to the read identifier. The quality scores for 36 bp reads were randomly picked from the corresponding experimental Illumina GA2 data (Harismendy *et al.*, 2009). Analogous quality scores of an unrelated Illumina GA2 run were used for the simulated 76 bp reads. Nucleotides were then switched to variant bases with probabilities according to their quality scores. For every combination of microindel length and frequency 10 runs were simulated. We compared short-read datasets consisting of 150 000 36 bp reads to 450 000 36 bp reads and 70 000 76 bp reads, corresponding to a mean sequencing depth of 18, 54 and 18.

2.2 Aligning short sequence reads

Short sequence read data was downloaded from ftp://ftp.jcvi.org/pub/data /NGS_cross_validation/ or simulated as described above. Short reads were mapped to the reference genome using BWA 0.4.9 (Li and Durbin, 2010), Novoalign Release 2.05.02 (Hercus, 2009) and RazerS 1.0 (Weese *et al.*, 2009) with default settings for mismatch penalty, gap opening penalty and gap extension penalty:

```
bwa aln -e 5 -t 8 <ref.fa> <reads.fastq>
novoalign -o SAM -d <ref.ndx> -f <reads.fastq>
razers -id -i 80 -rr 100 <ref.fa> <reads.fa>
```

For variant detection, Bowtie 0.11.3 (Langmead *et al.*, 2009) and MAQ 0.7.1 (Li *et al.*, 2008) were also tested with their default settings, which do not allow gapped alignments:

```
maq.pl easyrun <ref.fa> <reads.fastq>
bowtie -S -p 8 <ref.fa> <reads.fastq> <out.sam>
```

MAQ uses a spaced-seed approach to align reads. With default setting only reads that map to the reference genome with less than three mismatched bases in the first 28 bases of the read will be aligned. The ungapped alignment with the best alignment score is reported. Bowtie and BWA are based on backward search schemes with a Burrows–Wheeler transformation to efficiently align short sequencing reads against large reference sequences. Bowtie allows two mismatches or fewer within the high-quality end of each read, and it places an upper limit on the sum of the quality values at mismatched alignment positions. Novoalign finds global optimum alignments using full Needleman–Wunsch algorithm with affine gap penalties. RazerS adapts a q-gram counting technique for read filtering and maps reads using edit or Hamming distance as thresholds.

For all alignments the target sequence was used as reference sequence. When instead the whole genome was used as reference sequence, a certain proportion of reads mapped to locations outside the target region. These reads yielded higher alignment scores at wrong positions due to simulated mutations or sequencing errors.

All alignments were converted to Sequence Alignment/Map (sam) format that codes the position of an indel in the short-read in CIGAR string format. The consensus sequence was called according to the MAQ consensus model (Li *et al.*, 2008) with samtools release 0.1.7:

```
samtools pileup -vcf <ref.fa> <aln.bam>
```

The resulting raw consensus sequence was further filtered with:

```
samtools.pl varFilter -D100
```

This step also filtered out SNVs that are in a 10 bp window around a gap. For SNV detection we only considered reads with a read mapping quality of above 20 and for indel detection with a read mapping quality of above 50.

2.3 SNV and microindel calling

A coverage threshold of at least five reads covering a sequence position was used for variant and indel calling. For SNV calling approach, we used a frequency threshold as filter as described in Harismendy *et al.* (2009): a heterozygous SNV was called when 20–80% of the aligned reads showed the variant nucleotide. The position was called as a homozygous SNV if >80% of aligned reads showed the variant nucleotide.

In contrast to SNVs, the position of an indel with respect to the reference sequence is not necessarily unambiguously defined by a single coordinate, as the example in Figure 1 illustrates. The insertion of an adenosine into the local sequence motive of $C_iA_{i+1}A_{i+2}G_{i+3}$ after position i results in a mutated sequence that is identical to the sequence produced by an inserted adenosine after position $i+1$ or $i+2$. We assume both that each of these insertions has the identical biological meaning and that they are furthermore indistinguishable by mutation detection methods, so that in our example, calls of an insertion at position i, $i+1$ and $i+2$ represent one and the same mutation. An unambiguous annotation for this insertion would therefore have to list all equivalent indel positions, i.e. $+A\{i, i+1, i+2\}$. In a random sequence with all nucleotides occurring with same frequencies, the probability that the position of a single inserted nucleotide is unambiguously defined by a single coordinate is only 9/16. In genomic sequences, where homopolymers and small tandem repeats are more frequent than randomly expected, it is even less likely that an indel can be defined unambiguously by a single position. Therefore, when a read was aligned with a gap, the equivalent indel region, *eir*, was determined by computing all equivalent positions with respect to the sequence of this specific insertion or deletion. The following

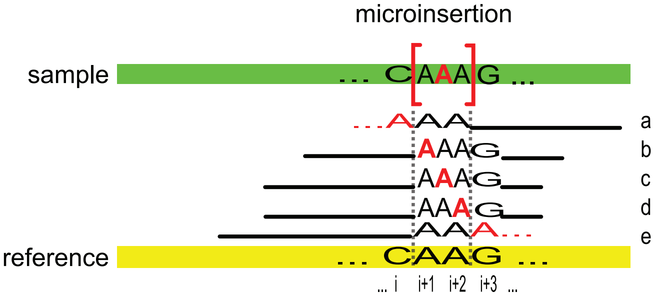

Fig. 1. The sample sequence has an adenosine triplet, compared with the doublet in the reference sequence. A short sequence read can be aligned to the reference with the insertion of an adenosine at any one of three positions (b, c, d). The position of the inserted A is not unambiguously defined by a single coordinate, but only by set of equivalent positions $eir = +A(i, i+1, i+2)$. Depending on the settings of the local alignment algorithm and on the surrounding sequence, a false alignment with mismatched nucleotides may yield a higher alignment score (a, e).

example illustrates how we proceeded with non-homopolymeric indels: if the reference sequence is r = CAGAT, then a called insertion of an AG at position 3 (i.e. following the three nucleotides CAG) leads to the same mutant sequence as a called insertion of GA at position 4: CAGAGAT. Our algorithm therefore identifies all called indel positions that lead to the identical mutated sequence (Fig. 1). To do so, we search for all positions in the reference sequence, where the insertion or deletion of the appropriate sequence pattern will lead to an identical mutated sequence. We refer to the set of all such positions as the *eir*, and consider all reads with called indels in the *eir* as equivalent for the purposes of indel calling. For the above example sequence r = CAGAT, an insertion of AG called at positions 1 and 3, as well as an insertion of GA called at positions 2 and 4, will lead to identical mutated sequences, thus the *eir* is $+AG(1-4)$. The pseudocode for our algorithm is shown in Figure 2. The frequency of an indel was computed by counting all reads that showed the indel sequence in the *eir* and dividing by the total number of reads covering the *eir*. We note that if multiple calls result in two or more distinct overlapping *eir*s, they were treated as separate for the purposes of indel calling. An indel was called if >10% of mapped reads showed the indel sequence in the *eir*.

```
Algorithm 1: Computation of eir
   Input: sequence $s$, position $i_p$, pattern $p$
   Output: eir
1  $x \leftarrow p$;                          // Extend eir to the right
2  $i_r \leftarrow i_p$;
3  $r \leftarrow s_{i_r+1}$;
4  $x' \leftarrow x_2 \ldots x_k.x_1$;
5  while $(x.r == r.x')$ do
6      $x \leftarrow x'$;
7      $i_r \leftarrow i_r + 1$;
8      $r \leftarrow s_{i_r+1}$;
9      $x' \leftarrow x_2 \ldots x_k.x_1$;
10 $x \leftarrow p$;                          // Extend eir to the left
11 $i_l \leftarrow i_p$;
12 $l \leftarrow s_{i_l-1}$;
13 $x' \leftarrow x_k.x_1 \ldots x_{k-1}$;
14 while $(l.x == x'.l)$ do
15     $x \leftarrow x'$;
16     $i_l \leftarrow i_l - 1$;
17     $l \leftarrow s_{i_l-1}$ ;
18     $x' \leftarrow x_k.x_1 \ldots x_{k-1}$;
19 eir $\leftarrow \{x, i_l, i_r\}$;
20 return eir
```

Fig. 2. An *eir* is computed from the genomic sequence s around an indel of a sequence pattern p after position i_p. i_r denotes the rightmost position of the *eir* and r the nucleotide to the right of i_r. Line 4 computes a cyclic permutation x' of the pattern in x. The '.' operator indicates a string concatenation. Lines 1–9 extend the *eir* to the right. Following the extension to the left (lines 10–18), the left and rightmost positions are returned together with the leftmost pattern.

3 RESULTS

Due to the large amount of short-read data that NGS platforms produce, efficient read mapping tools quickly narrow down the candidate regions where a read possibly maps. In this candidate region, local alignment algorithms are used to minimize the mismatched nucleotides and inserted gaps. An alignment score is finally used to report the best matching alignment. Generally two different terms contribute to the alignment score, a penalty α from mismatched bases and a penalty β in case of gap insertions. The exact values of α and β depend, on the one hand, on global settings of the alignment algorithm that is on the applied similarity matrices and on gap opening and extension penalties. Quality values of the aligned nucleotides as well as their positions in the reads can be taken into account. On the other hand, the alignment score is locally influenced by the surrounding sequence. It is crucial to understand that the optimal alignment score that is reported for a read depends on the algorithmic parameters as well as on the sequence context (Durbin *et al.*, 1999). This explains why one and the same read may be aligned to the very same starting and ending positions by two different mapping tools and yet their alignment shapes and score may differ. A read that covers a certain microindel of a certain size, may also be aligned with a gap in one sequence context, or with mismatches in another, depending on the neighboring nucleotides (Fig. 1). If the position of the inserted or deleted sequence is located at the beginning or end of the short sequence read or if the surrounding nucleotides are similar to the indel sequence an alignment with partially mismatched bases might yield a better alignment score and thus lead to a false alignment with respect to the true sample sequence.

3.1 Mapping efficiency of reads covering indels

The origin of the simulated reads was used to calculate the rates of reads that were mapped to the correct position. A read was counted as correctly mapped, if either its mapped starting or ending position agreed with its original coordinate. Due to the different algorithmic approaches, we expected different mapping efficiencies for the mapping tools we tested. By default, BWA only maps reads that show less than three mismatched bases in the seed to the reference sequence, not counting gaps. This means a read with more than three sequencing errors in the seed sequence cannot be mapped. Novoalign reports the best unique alignment, regardless of the number of mismatched bases. A read that maps to more than one position with the same score is not reported. RazerS maps all reads that can be aligned within a certain editing distance (80% identity).

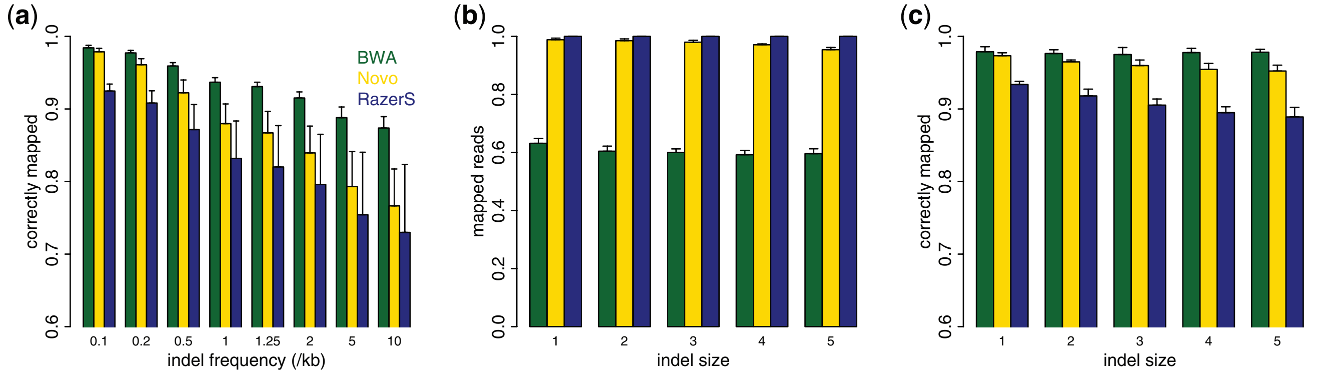

Fig. 3. (**a**) The rate of mapped reads with indels that are correctly mapped decreases with increasing indel frequency. (**b**) In contrast, the rate of all reads with indels that can be mapped shows only a weak dependency on the indel size. (**c**) Also the rate of mapped reads with indels that are *correctly* mapped, shows only a weak dependency on the indel size. As can be seen in (b and c), BWA is able to align a lower proportion of all reads with indels owing to its more stringent criteria, but the accuracy of its alignments is higher for the reads that can be aligned. The height of the barplot indicates the mean proportion of mapped or correctly mapped reads, and error bars indicate 2 SDs. The indel frequency in panels b and c was fixed to 0.2/kb.

Due to the relatively high sequencing error rate in the simulated data, BWA is only able to map ~80% of reads from regions without indels, compared with >98% by Novoalign and RazerS. Of all mapped reads that did not cover indel sites, >99 % were mapped to the correct coordinates by Novoalign and BWA and >95% by RazerS. In Figure 3a, the rate of mappable reads from regions with at least one indel that were mapped to the correct coordinate is shown. For all mapping tools the rate of correctly mapped reads decreases for an increasing indel frequency. Reads covering more than one indel are poorly mapped by all of the alignment tools tested. For an indel frequency of $f_{\text{indel}} = 0.1$/kb, the probability that two simulated indels have a smaller distance than 36 bp is ~0.4%. This probability increases to >30% in datasets with an indel frequency of $f_{\text{indel}} = 10$/kb. While many reads that cover more than one indel per site are mapped to incorrect positions by Novoalign and RazerS, these reads are not mapped at all by BWA (Figure 3b and c)

3.2 Microindels affect SNV detection error rates

Many efficient short-read alignment tools that are used for SNV detection map short reads by only allowing a certain number of mismatches. Gapped alignments—that are required for indel detection—are not yet enabled in some widely used alignment tools. The alignment tool MAQ (Li *et al.*, 2008) or Bowtie (Langmead *et al.*, 2009) for instance, will not detect microindels by their default settings and optimize their local alignment only by minimizing the number of mismatched bases. As a consequence ungapped alignments near microindels are prone to false-positive SNV calls. Indeed, Ossiwski *et al.* (2008) found that microindels are a major source of false SNV detection by MAQ. We studied how the frequency of microindels affects SNV detection and analyzed whether SNV calling profits from gapped short-read alignments.

The differences between individuals with respect to SNVs are now adequately known on a genome-wide scale from the comparison of complete diploid genome sequences (Ahn *et al.*, 2009; Bentley *et al.*, 2008; Levy *et al.*, 2007; Wheeler *et al.*, 2008). The haploid genomes of two individuals of Central European descent differ at approximately two million chromosomal positions totalling in about three million homozygous and heterozygous SNVs. In our simulated sequence data, we thus adjusted the SNP frequency between the reference sequence that was used as template for the short-read alignments and the simulated test sequence to a rate of $f_{\text{SNP}} = 1$/kb and assumed that these variants are randomly distributed. For our simulations, we took the frequencies of microindels that were reported in Harismendy *et al.* (2009) as a lower bound to the estimated microindel prevalence. The four analyzed individual samples differed in an average of nine microindels ≤ 5 bp per 88 kb, which translates to a frequency of $f_{\text{indel}} = 0.1$/kb. We arbitrarily chose $f_{\text{indel}} = 10$/kb as an upper bound for the simulations.

Our SNV calling was based on the consensus sequence produced by samtools incorporating a Bayesian model. We further filtered for heterozygous and homozygous variations as described in Harismendy *et al.* (2009): We called a nucleotide a heterozygous variant whenever at least five reads fulfilling the quality criteria covered the sequence position and the variant frequency ranged between 20% and 80%. A homozygous variant was called for a variant frequency >80%.

We compared the effect of different microindel frequencies on SNV detection by mapping tools that do or do not allow gapped alignments. A short sequence read coming from a region in the sample sequence with a microindel compared with the reference sequence may thus be aligned with mismatched bases instead of gaps, resulting in a higher rate of variant bases at indel positions. In Figure 4a and b, the rate of falsely called SNVs at microindel positions versus indel frequency and indel size is shown for 36 bp reads with a mean sequencing depth of 18. The error rate for mapping tools that perform ungapped alignments (Bowtie and MAQ) is higher for SNV detection compared with mapping tools that allow gapped alignments (BWA, Novoalign and RazerS) for all simulated datasets. For increasing indel size the error rate of indel positions that are falsely called as SNVs decreases for ungapped mapping tools, whereas there is an increasing trend for gapped mapping tools (Fig. 4b). For gapped mapping tools, the probability increases with indel size that a read that encompasses a large indel near one of its ends is falsely mismatch aligned. In contrast, ungapped mapping tools tend to not align these reads at all.

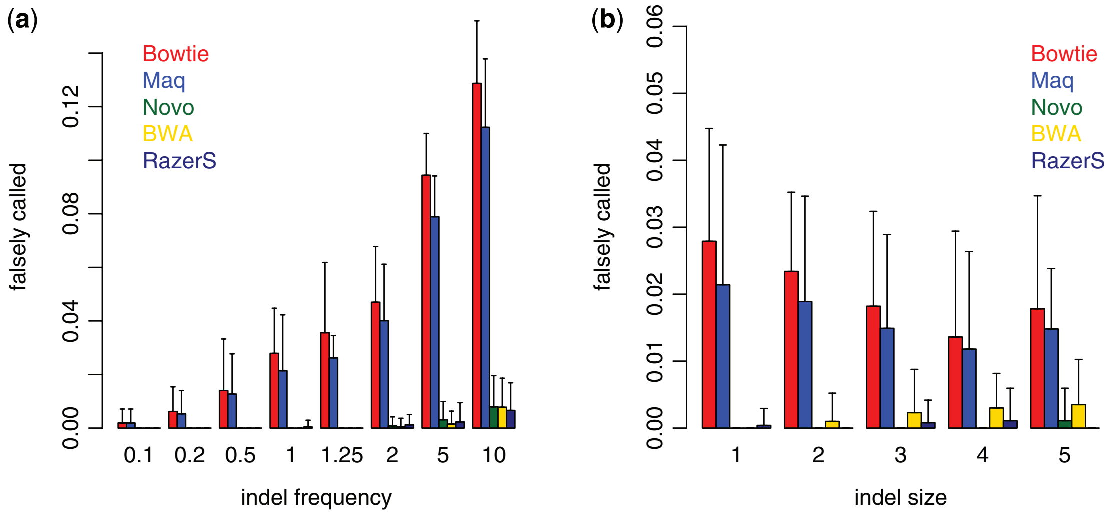

Fig. 4. Thirty-six bp short reads from simulated datasets with a mean sequencing depth of 18 with indel frequencies ranging from 0.1/kb to 10/kb and indel sizes ranging from 1 to 5 were aligned with mapping tools that perform ungapped (Bowtie and MAQ) and gapped (BWA, Novoalign and RazerS) alignments. SNVs were called from the consensus sequence based on a simple frequency threshold. (**a**) The rate of false positive SNVs that are called at simulated indel sites of size one increases with the indel frequency. SNV calling based on ungapped alignments exhibit higher error rates. (**b**) Also for increasing indel sizes the error rates are significantly lower when gapped mapping tools are used. Ungapped mapping tools show a decreasing trend in error rates for increasing indel size, as less of these reads get mapped, whereas gapped mapping tools show an increasing trend, however on a much lower level.

The Bayesian model used by samtools to generate the consensus sequence takes the quality score of the base as well as the mapping score of a read into account. When an indel is falsely aligned as mismatch, the probability increases such that the following bases are also mismatch aligned. A mismatching base in a read with a low-mapping quality will thus be less weighted in calling the consensus base. It has to be noted, that Bowtie and RazerS do not report mapping quality scores in the sam format but report a constant mapping score. Alignments produced by these mapping tools thus benefit less from the Bayesian SNV calling model.

3.3 Detection of microindels in simulated short-read data

In our second experiment, the short-read mapping tools BWA (Li and Durbin, 2010) and Novoalign (Hercus, 2009), which enable gapped alignments, were used to align the same simulated sequence data containing microindels. For each inserted or deleted sequence fragment we computed the equivalent indel region, *eir*. An indel was called if the indel frequency was >10% and at least five reads covered the *eir*. We excluded RazerS from this analysis, as its algorithmic approach that is based on editing distance is not compatible with our indel calling approach based on *eir*. RazerS tends to split up larger indels into smaller subunits of indels. Although these combinations of smaller indels may lead to the same mutated sequence, they will not be recognized as part of the *eir* (see Supplementary Material for further information).

In Figure 5, the sensitivities of indel detection are shown for varying indel frequencies and indel sizes and for datasets of different read length and sequencing depth. An indel was counted as correctly called, when the correct indel sequence was detected in the *eir* at a rate >10%. The sensitivity for detecting indels of size 1 nt depends only weakly on the indel frequency (Fig. 5a–c). In general, the sensitivity of the detection of indels of a certain size is not overly dependent on the indel frequency itself, suggesting that indel detection is quite robust over a wide range of indel frequencies. For increasing indel size, the sensitivities differ for the different mapping tools in datasets of a mean sequencing depth of 18 and 36 bp short reads. While ~90% of indels of size three are correctly detected in reads aligned by Novoalign, this rate drops to <50% in BWA alignments (Fig. 5d). The sensitivity of detecting larger indels benefits from larger read length and higher sequencing depth. The effect of increased sequencing depth and higher read lengths also outweighs the effects of changed parameter settings of the alignment tools by far (data not shown). For the datasets of 36 bp reads and a mean sequencing depth of 18, we also tested whether a more tolerant mode of indel detection might be used to increase sensitivity rates: in the 10 bp window mode, an indel was counted as correctly called, if an indel was aligned at a rate >10% at any of the 10 bp surrounding a true indel. This also means that the exact indel sequence was not necessary for an indel to be counted as correctly called. In the tolerant window mode, the sensitivity as well as the positive predictive values increase slightly for larger indel sizes (Supplementary Material). However, for indels of size one the sensitivity is higher when indels are called based on the *eir* algorithm. This can be explained as follows: When an indel has an equivalent indel region larger than size one, not all indels are usually placed at the same position by the mapping tool. In some of the cases, the frequency of indels at a single position does not suffice to be called as indel. However, when the indel frequency threshold is based on the equivalent indel region, *eir*, all such indels contribute to the indel frequency, regardless of their position in the *eir*. It should be noted that the tolerant window mode is not able to distinguish between non-equivalent indels occurring within the same window.

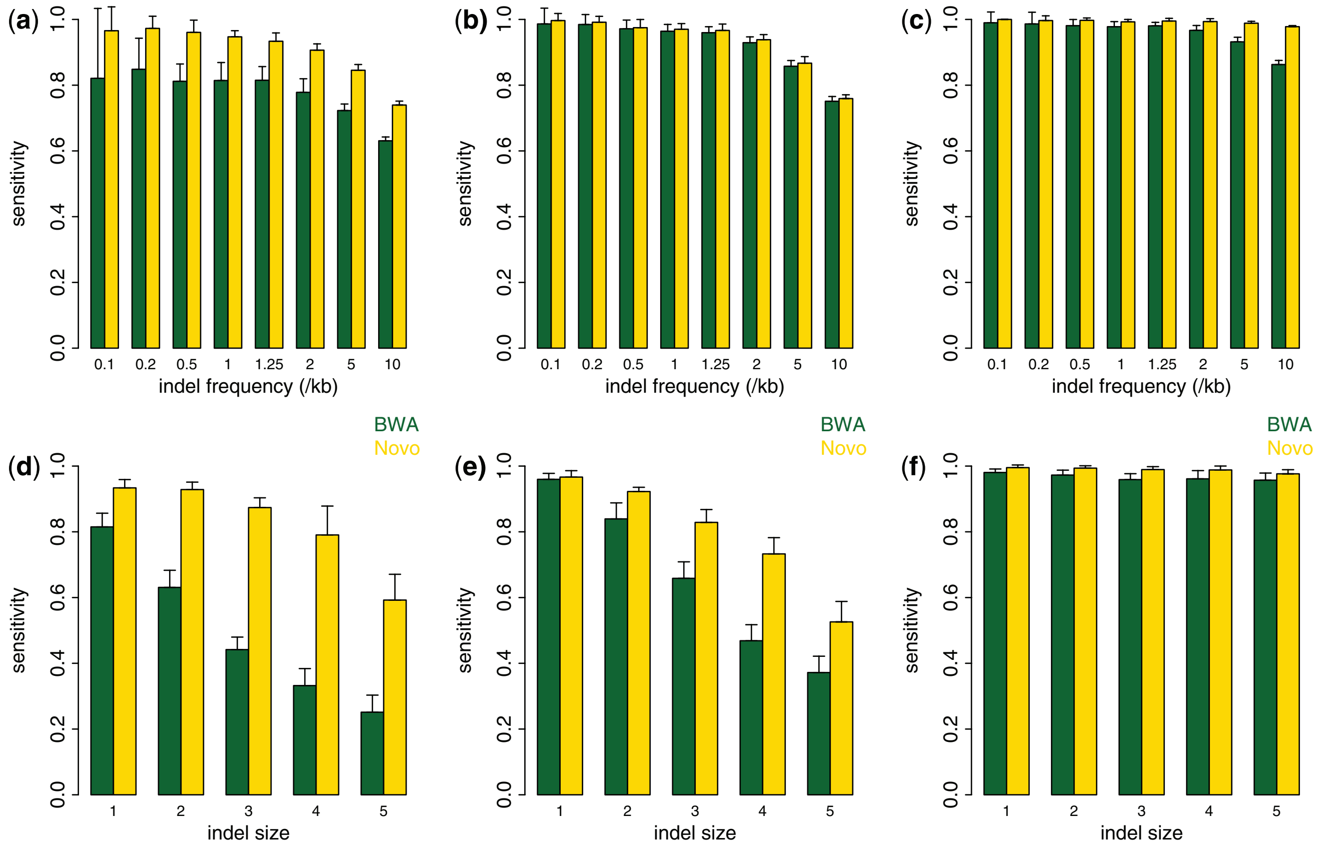

Fig. 5. Short reads containing microindels were mapped with BWA and Novoalign, which enable gapped alignments and indels were called as described in the main text. In (**a, d**) the sensitivity of indel detection was measured for datasets containing 36 bp reads and a mean sequencing depth of 18, in (**b, e**) for 36 bp reads and a mean sequencing depth of 54 and in (**c, f**) for 76 bp reads and a mean sequencing depth of 18. (a–c) The sensitivity for detecting indels of size one decreases for increasing indel frequencies. The sensitivity of indel detection based on Novoalign and especially BWA alignments benefits from higher sequencing depth and read length. (d–f) In datasets of an indel frequency of 1/kb the sensitivity of indel detection of larger indels benefits markedly from longer reads.

3.4 Microindel detection in real data

In Harismendy *et al.* (2009), sequence fragments of a total length of 88 kb were sequenced from four different individuals with ABI 3730xL Sanger, Roche 454, Illumina Genome Analyzer and ABI SOLiD technologies. Indels were automatically detected only on the ABI Sanger and the Roche 454 platform. Altogether 36 microindels of ≤5 bp length were detected, 6 by ABI Sanger and an additional 30 by Roche 454. Only 1 out of 6 microindels detected by ABI Sanger were also identified by Roche 454. In Harismendy *et al.* (2009), Illumina and ABI SOLiD short reads were not analyzed for microindels, as no detection algorithms were available at the time of analysis. To evaluate whether microindel detection is also applicable to real short-read data, we mapped the Illumina GA 36 bp short-read data of Harismendy *et al.* (2009) using BWA and Novoalign, and called indels as described. Seven out of the reported 36 microindels could be detected with our approach (Supplementary Material). One of the indels was also detected by ABI Sanger, the other six by Roche 454 (Supplementary Material). In addition a large number of new indels was called for each of the four individuals. For example, in NA17156 11 and 12 additional indels were called from short sequence reads that were aligned by BWA and Novoalign.

When we analyze indels that were called on the total of 296 kb of all four samples covered by short reads, altogether 331 indels were called based on alignments of BWA and Novoalign (Fig. 6). Of these, 138 indels were called in both alignments. We visually inspected all indels that were only called in one of the two alignments: the overwhelming majority of indels that could not be called in both alignments have a low frequency and are not called in one of the two alignments because of the frequency threshold of 10%. Only in three cases different indel sequences were called to the same intervall and were thus counted as different indels (Supplementary Material).

4 DISCUSSION

The genome-wide frequency of small base pair insertions and deletions might have been previously underestimated for a simple reason: traditional sequencing techniques are simply not overly good at automated detection of short indels, especially if they are heterozygous. In Harismendy *et al.* (2009), the ABI Sanger and Roche 454 platforms were used for automated indel detection. Only 1 out of 36 microindels (≤5 bp) was detected by both platforms, suggesting high false negative and false positive rates of both

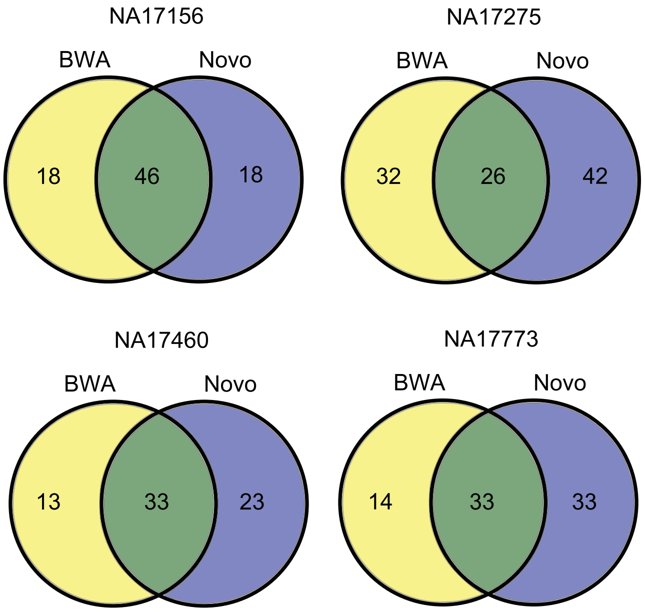

Fig. 6. Venn diagram for microindels called on altogether 296 kb based on the different mapping tools, BWA and Novoalign. In individual NA17156, 46 microindels were detected in alignments of both mapping tools, whereas 18 microindels were only detected in BWA or Novoalign. Most indels that were only detected in one alignment have a low frequency in the aligned reads.

technologies. The pyrosequencing technology of Roche 454 uses the fluorescent signal strength of incorporated nucleotides in a homopolymer to estimate its length. However the signal strength for homopolymer stretches is only linear for up to eight consecutive nucleotides, resulting in a higher error rate for larger homopolymer stretches (Margulies *et al.*, 2005). Five indels detected by ABI Sanger but not by Roche 454 and one indel detected by Roche 454 but not by ABI Sanger are flanked by such homopolymers.

Another issue is whether indels detected in the targeted sequence are actually present in the genomic DNA of the individual, or are artifacts of the long-range PCR amplification process. Especially sites with short tandem repeats exhibit higher mutation rates in PCR reactions due to DNA slippage (Lai *et al.*, 2003; Shinde *et al.*, 2003). Mutations that occurred during the sample amplification will thus be present at frequencies far below 100%. Seven of the 30 indels only detected by Roche 454 are such extensions or contractions of short tandem repeats. Interestingly four of these seven indels could also be seen in alignments of Illumina short-reads, however they were not called as indels by our approach, as their frequency did not pass the frequency threshold of 10%. This might indicate that some indels detected by Roche 454 and other NGS platforms are actually false positives, due to the sample preparation. This error rate might be reduced if DNA enrichment techniques are used that are not based on an in vitro amplification step.

We analyzed the performance of a indel calling algorithm that uses an unambiguous definition of an indel region on simulated datasets containing indels with frequencies ranging from $f_{\mathrm{indel}} = 0.1/\mathrm{kb}$ to $f_{\mathrm{indel}} = 10/\mathrm{kb}$ and demonstrated that the sensitivity and positive predictive value are almost constant over a range of two orders of magnitude.

We may now use the positive predictive value that was measured in our simulations for Novoalign to estimate the true microindel frequency in the targeted sequences that were amplified by long-range PCR in the four individuals analyzed in Harismendy *et al.* (2009): $f_{\mathrm{indel}} = (64+68+56+66)/(4\times 296)\times 0.9 = 0.19$ microindels/kb.

In the first diploid genome that was sequenced using paired-end short reads of the Illumina platform a microindel frequency of 0.033 indels/kb was reported (Bentley *et al.*, 2008). Sequencing of the diploid genome of a famous geneticist using the Roche 454 platform identified an order of magnitude more indels (Wheeler *et al.*, 2008). Therefore, we claim that the range of frequencies of indels used in our simulations are not unrealistic. Additionally, it is plausible that sequencing platform specific as well as algorithmic differences are responsible for at least part of the wide discrepancy of the indel frequencies in these two diploid genome sequences.

Compared with the frequency of SNVs, the microindel frequency seems to be at least an order of magnitude smaller. The effect of microindels on the false positive error rate of SNV detection should thus be relatively small (≤ 0.05) (Fig. 3). However, further studies on real datasets should investigate whether mapping tools that allow gapped alignments reduce false positive SNVs called from short-read data, as our simulations suggest.

We outlined that the unambiguous annotation of an indel may require more than just a single coordinate with respect to the reference depending on the sequence context and suggested the equivalent indel region, *eir*, for this purpose. Databases such as dbSNP have not yet systematically dealt with this annotation problem. For instance, there are two entries in dbSNP that correspond to one of the indels reported in Harismendy *et al.* (2009): rs72552124 and rs41312514 report an inserted guanine at the beginning and alternatively at the end of a 6 base polyguanine tract.

As demonstrated, our method is well suited for automated indel detection in short sequence reads. We also showed that the sensitivity of indel detection benefits considerably from higher sequencing depth and longer reads. This should be considered in the experimental design. For datasets with a low coverage and short reads, the sensitivity may be maximized at the expense of computing time by using more accurate mapping tools. In future work, additional methods to further increase sensitivity and positive predictive values for indel detection will be analyzed. These include the evaluation of paired-end reads and a restriction to indels called only in the center part of a short read.

5 CONCLUSION

Our study has provided insight into systematic errors in SNV detection that is based on short-read sequence alignments. False positive error rates in SNV detection can be markedly reduced by using mapping tools that enable gapped alignments. Microindel detection in short-read alignments using a simple algorithm to calculate the *equivalent indel region* was shown to be technically feasible in simulated datasets. The sensitivity of automated indel detection from short reads is comparable with automated indel detection methods on ABI Sanger or the Roche 454 platform. Continued improvements in our understanding of the technical issues of NGS platforms will allow the development of more sophisticated analysis methodologies.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers that helped in improving our work by their comments. We thank Mathew Wakefield, Nadia Chuzhanova and Marcel Schulz for helpful discussions on the manuscript.

Funding: Berlin-Brandenburg Center for Regenerative Therapies (BCRT) (Bundesministerium für Bildung und Forschung, project number 0313911); Deutsche Forschungsgemeinschaft (DFG SFB 760).

Conflict of Interest: none declared.

REFERENCES

Ahn,S.-M. *et al*. (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res.*, **19**, 1622–1629.

Ball,E.V. *et al*. (2005) Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum. Mutat.*, **26**, 205–213.

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. *et al*. (1999) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, The Edinburgh Building, Cambridge CB2 2RU, UK.

Harismendy,O. *et al*. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol.*, **10**, R32.

Hercus,C. (2009) www.novocraft.com (last accessed date November, 2009).

Kolpakov,R. *et al.* (2003) mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.*, **31**, 3672–3678.

Korbel,J.O. *et al*. (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.

Lai,Y. *et al*. (2003) The mutation process of microsatellites during the polymerase chain reaction. *J. Comput. Biol.*, **10**, 143–155.

Langmead,B. *et al*. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

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

Li,H. and Durbin,R. (2010) Fast and accurate long read alignment with Burrows-Wheeler transform. *Bioinformatics*. [Epub ahead of print, January 15, 2010]

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

Mardis,E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, **9**, 387–402.

Margulies,M. *et al*. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Ossowski,S. *et al*. (2008) Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res.*, **18**, 2024–2033.

Redon,R. *et al*. (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

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.

Weese,D. *et al*. (2009) RazerS–fast read mapping with sensitivity control. *Genome Res.*, **19**, 1646–1654.

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