BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 5 2012, pages 619–627
doi:10.1093/bioinformatics/bts019

Genome analysis

Advance Access publication January 11, 2012

Detecting genomic indel variants with exact breakpoints in single- and paired-end sequencing data using SplazerS

Anne-Katrin Emde[1,2,*], Marcel H. Schulz[3], David Weese[1], Ruping Sun[2], Martin Vingron[2], Vera M. Kalscheuer[2], Stefan A. Haas[2] and Knut Reinert[1]

[1]Department of Computer Science, Freie Universität Berlin, Takustrasse 9, [2]Max-Planck-Institute for Molecular Genetics, Berlin, Germany, Ihnestrasse 63-73, 14195 Berlin, Germany and [3]Ray and Stephanie Lane Center for Computational Biology, Carnegie Mellon University, 7401 Gates-Hillman Complex, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA

Associate Editor: Alex Bateman

ABSTRACT

Motivation: The reliable detection of genomic variation in resequencing data is still a major challenge, especially for variants larger than a few base pairs. Sequencing reads crossing boundaries of structural variation carry the potential for their identification, but are difficult to map.

Results: Here we present a method for 'split' read mapping, where prefix and suffix match of a read may be interrupted by a longer gap in the read-to-reference alignment. We use this method to accurately detect medium-sized insertions and long deletions with precise breakpoints in genomic resequencing data. Compared with alternative split mapping methods, SplazerS significantly improves sensitivity for detecting large indel events, especially in variant-rich regions. Our method is robust in the presence of sequencing errors as well as alignment errors due to genomic mutations/divergence, and can be used on reads of variable lengths. Our analysis shows that SplazerS is a versatile tool applicable to unanchored or single-end as well as anchored paired-end reads. In addition, application of SplazerS to targeted resequencing data led to the interesting discovery of a complete, possibly functional gene retrocopy variant.

Availability: SplazerS is available from http://www.seqan.de/projects/ splazers.

Contact: emde@inf.fu-berlin.de

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

Received on August 30, 2011; revised on December 26, 2011; accepted on January 4, 2012

1 INTRODUCTION

Next-generation sequencing (NGS) technologies have had huge impact on the study of molecular biology (Metzker, 2010): ultra high-throughput NGS technologies provide excellent means for analyzing RNA content of a cell (Wang *et al.*, 2009), resequencing whole genomes (Bentley *et al.*, 2008; McKernan *et al.*, 2009; Wheeler *et al.*, 2008) and detecting, for example, disease-causing variants (Chen *et al.*, 2008) or investigating the epigenetic state of a cell (Barski *et al.*, 2007). There is a long list of biological applications and, accordingly, a high demand for specialized bioinformatics tools for the analyses. Due to the tremendous data yield, this constitutes a great challenge to computational biology. One of the most fundamental tasks is read mapping, i.e. determining the origin of the sequenced reads in a reference genome. All further analysis is based on the mapped reads, e.g. genomic variants are identified or transcript expression is quantified. Reads containing more than a few variant base pairs with respect to a reference will be hard to map, and sequencing errors further complicate this task.

While a large part of human genomic variation is due to single nucleotide polymorphisms (SNPs), recent years have shown that insertion/deletion (indel) variants dramatically contribute to genomic diversity (Mullaney *et al.*, 2010). Furthermore, they have been linked to many diseases (Stenson *et al.*, 2009), and especially large indels have been shown to have strong structural and functional impact (Stankiewicz and Lupski, 2010).

The first high-throughput method for genome-wide detection of large deletions and duplications was array comparative genomic hybridization (aCGH) where sample and reference DNA are competitively hybridized on a probe array (Pinkel and Albertson, 2005). Given steadily decreasing costs, sequencing approaches are likely to gradually replace aCGH methods, since they carry the potential to provide base pair resolution and identify novel insertions while generating less measurement noise.

Sequencing-based indel detection methods are based on different strategies often utilizing paired-end or mate-pair sequencing data. One conclusion from the 1000 Genomes Project so far is, that a comprehensive indel detection method needs to make use of different detection strategies as no strategy alone is sufficient (Durbin *et al.*, 2010). We will briefly introduce the different strategies and point to their strengths and weaknesses.

Most prominently, strategies based on paired-end or mate-pair data make use of the approximate distance and relative orientation of read pairs. Shifts in the mapped distance or changes in relative orientation indicate indel events or also more complex structural variation. Examples for popular methods that address indel detection based on abnormal insert sizes are PEMer (Korbel *et al.*, 2009), BreakDancer (Chen *et al.*, 2009), MoDIL (Lee *et al.*, 2009) and SVDetect (Zeitouni *et al.*, 2010). A drawback of these read pair methods is that they require tight insert size distributions for accurate discovery of small- to medium-sized indels and for confident localization of breakpoints (Medvedev *et al.*, 2009). So-called *anchored split read mapping* addresses this shortcoming by

*To whom correspondence should be addressed.

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at MPI Molec Genetics on March 28, 2013

directly taking advantage of reads crossing a breakpoint. In split read mapping, the 5′ and 3′ match of a read may be interrupted by a longer gap in the read-to-reference alignment enabling detection of the exact breakpoint and size of the indel. Ye *et al.* (2009) first applied this approach to short NGS read data. Their tool, Pindel, builds on the availability of paired-end reads, split mapping the unmapped but anchored end within a genomic region defined by its confidently mapped paired end. Pindel is based on a pattern growth algorithm, combining unique 5′ and 3′ matches into a split read match.

Detecting large indels in single-end or unanchored reads is more challenging, due to the short read length and to the repetitiveness of genomic DNA (especially in higher organisms such as human). Read depth methods (Xie and Tammi, 2009; Yoon *et al.*, 2009) work on single-end as well as paired-end data and are able to detect very large deletions and duplications, i.e. changes in copy number, by comparing the observed number of mapped reads in a genomic region to the expected one. These methods provide no exact breakpoints due to low resolution and suffer from artifactual read depth fluctuations.

However, as read lengths increase with advances in sequencing technology, single reads provide the potential for identifying large indels through split-read alignment. In this work, we will use single-end and unanchored reads as short as 76 bp to predict large deletions up to several kilobases in size with high confidence.

Similar to Pindel, our tool SplazerS employs a split read approach. However, it uses a more sensitive alignment method for prefix and suffix matches, allowing for mismatches and small gaps, thereby making it more robust in the presence of small genomic variation and sequencing errors. Especially, read suffixes require higher error tolerance, as sequencing error rates tend to increase toward the 3' ends of reads. Most notably, SplazerS implements a fast anchored as well as a more general unanchored/single-end alignment mode, making it applicable to paired-end as well as single-end sequencing data.

Another tool for split read alignment is GSNAP (Wu and Nacu, 2010), which indexes every third 12mer in the genome and then maps reads one after the other by searching for exact 12mer matches. It does not provide support for anchored split read mapping or for subsequent indel detection. Also BWA (Li and Durbin, 2009), a popular multi-purpose read mapping tool based on a Burrows-Wheeler index, provides split read mapping functionality to a certain extent. Further methods such as SpliceMap (Au *et al.*, 2010), MapSplice (Wang *et al.*, 2010) or SplitSeek (Ameur *et al.*, 2010) follow similar approaches but are geared toward splice junction discovery in RNA-Seq data. Typical drawbacks of these methods for indel detection are a lack of functionality for insertion detection, or their requirement for donor/acceptor splicing patterns.

Our method is novel in that it supports both Hamming (ungapped) and edit (gapped) alignment in the 5′ and 3′ matches, making it applicable to reads from different sequencing technologies. Also longer and variable read lengths as expected from upcoming sequencing technologies (Eid *et al.*, 2009) can be handled. It does not require the prefix or suffix of minimum length to be unique, but attempts to align the entire read while identifying the best split position. It can furthermore report multiple and also suboptimal matches if they exist. SplazerS can be used to split-align short reads to candidate regions as in anchored split read mapping, and is also able to directly align single-end reads to an entire genome where stricter parameter settings may be desired. For fine-tuning of small indel detection, the mapping results can be used in conjunction with arbitrary indel detection methods supporting the SAM format, such as Dindel (Albers *et al.*, 2010) or GATK (McKenna *et al.*, 2010). Split-mapped reads can also be combined with SAM output from other mapping steps, for example 'normally' mapped reads where small indels were already allowed.

We apply our method to real paired-end data and to simulated and real single-end data and demonstrate its high precision and high sensitivity, even in challenging, variant-rich regions. We additionally use paired-end data in single-end mode to show how prediction accuracy can be increased even further when considering unanchored reads. In addition, we will show how application of SplazerS led to the interesting discovery of a retrocopy of *PQBP1*, a gene which has been shown to be involved in X-linked intellectual disability (Kalscheuer *et al.*, 2003; Lenski *et al.*, 2004).

2 METHODS

We first want to give some basic notation: we have a set of reads R where $r \in R$ and a reference sequence g, where r and g are sequences over the alphabet {A,C,G,T,N}. Furthermore, the operator $|\cdot|$ denotes the length of a sequence. Given an alignment $a_{r_{k,l}}^{g_{i,j}}$ of the subsequence of read r starting in position k and ending in position l to the subsequence of g starting in i and ending in j, we define an error function $d(a_{r_{k,l}}^{g_{i,j}})$ that returns the number of errors in the alignment, i.e. the sum of the number of gaps and mismatches. Any character aligned with an 'N' is counted as an alignment error. The error rate is then given by $\epsilon = d(a_{r_{k,l}}^{g_{i,j}})/(l-k+1)$, i.e. the number of alignment errors divided by the length of the read subsequence. For the sake of clarity, we will drop the superscript and refer to alignments as $a_{r_{k,l}}$ instead of $a_{r_{k,l}}^{g_{i,j}}$.

2.1 Definition of *split read alignment*

We want to identify collinear *split read alignments*, i.e. alignments that are split into a prefix (5′) and suffix (3′) match that lie within a collinear genomic subsequence W.l.o.g. we assume our reads to only match to the forward strand, and define a split read alignment of read r as an alignment where the following holds:

(1) a (non-empty) prefix p of r aligns to $g_{i,j}$

(2) a (non-empty) suffix s of r aligns to $g_{k,l}$

(3) $|p|+|s| = |r|$ and $j+1 < k$ (read spans a deletion) or $|p|+|s| < |r|$ and $j+1 = k$ (read spans an insertion)

where $1 \le i \le j < k \le l \le |g|$.

A *valid* split read alignment (or split read *match*) is one that additionally fulfills the following criteria:

(1) $|p| \ge m$ and $|s| \ge m$

(2) $d(a_{p_{1,m}}) \le e_p$ and $d(a_{s_{|s|-m+1,|s|}}) \le e_s$

(3) $\frac{d(a_p)+d(a_s)}{|p|+|s|} \le \epsilon$

(4) $k-j-1 \le \delta$

The first condition ensures that prefix and suffix have at least a certain minimum match length m. The second condition ensures that the number of errors in the minimum length prefix (suffix) match is at most a certain maximum number e_p (e_s). Condition 3 guarantees that the sum of the number of errors in the prefix and suffix match divided by the combined length of prefix and suffix lies within the allowed error rate ϵ. The maximum gap length δ puts a constraint on the distance of prefix and suffix match. An example of a valid and an invalid split read match is given in Figure 1. In some cases, it may be desirable to set $e_s > e_p$ as error rates tend to increase toward the 3' ends of reads. Note that m, e_s and e_p are given as integer numbers, independent of read length. However, the total error rate ϵ is dependent on

Downloaded from http://bioinformatics.oxfordjournals.org/ at MPI Molec Genetics on March 28, 2013

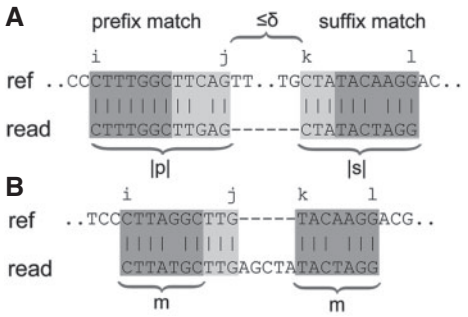

Fig. 1. Two examples of split read alignments. Given parameters $m=7$, $e_p=e_s=1$ and $\epsilon=0.1$. (**A**) is a valid alignment spanning a deletion and (**B**) spans an insertion but is not a valid match as the error rate condition is violated.

the read length, adjusting the allowed numbers of errors for reads of variable length.

For *valid anchored* split read matches, we further restrict the genomic region that the read is allowed to map to. This region is defined by the location of the confidently mapped paired end and the expected insert size. To find such valid split read alignments, we use the following algorithm which we have implemented in SplazerS—in reference to the related general purpose read mapper RazerS (Weese *et al.*, 2009)—using the SeqAn library (Döring *et al.*, 2008).

2.2 Mapping algorithm

The input to our algorithm is a set of reads (either in Fasta/Fastq format or, for anchored split mapping, in single-chromosome SAM format) and a reference sequence (in Fasta format). The output file will contain the successfully split-mapped reads in GFF or SAM format. Figure 2 outlines the main steps of the algorithm. A filtering method identifies potential prefix/suffix matches, i.e. match candidates (filtering phase, Fig. 2A). Possible match candidates are verified with a seed-and-extend alignment method (match verification, Fig. 2B) and, if successful, combined into split read matches (match combination, Fig.2C). The following describes these steps in detail.

2.2.1 Double-swift filtering identifies potential matches Our filtering phase is based on counting q-gram matches (Burkhardt *et al.*, 1999), i.e. identifying short subsequences of length q that is shared between read and reference. We build two indices: the *left* index containing all q-grams of all read prefixes of length m, and the *right* index containing all q-grams of all read suffixes of length m. Employing the SWIFT filtering algorithm (Rasmussen *et al.*, 2005), we start scanning the reference sequence with the right index. Whenever a potential suffix match region for a read r is encountered, i.e. a certain minimum number of matching q-grams have been observed within a defined region of the genome, the left index is 'dragged' behind, up to the current position of the right index. Potential prefix matches within the allowed distance, as defined by parameter δ, are recorded in a queue. If there is at least one potential prefix match for r within the allowed distance, this triggers the verification of the potential suffix match of r.

2.2.2 Prefix and suffix matches are verified separately Only if the suffix match is verified positively, the potential prefix matches are also subjected to verification. In order to avoid having to verify a potential prefix match several times, the filtering queue keeps track of the verification status of each potential prefix match. We verify potential prefix/suffix matches using either a simple scanning of diagonals in the alignment matrix for ungapped alignment, or using Myers' Bit Vector algorithm (Myers, 1999) in the case of edit distance. Verification is done only for the prefix/suffix of length m. A positively verified prefix/suffix match is then extended using either ungapped

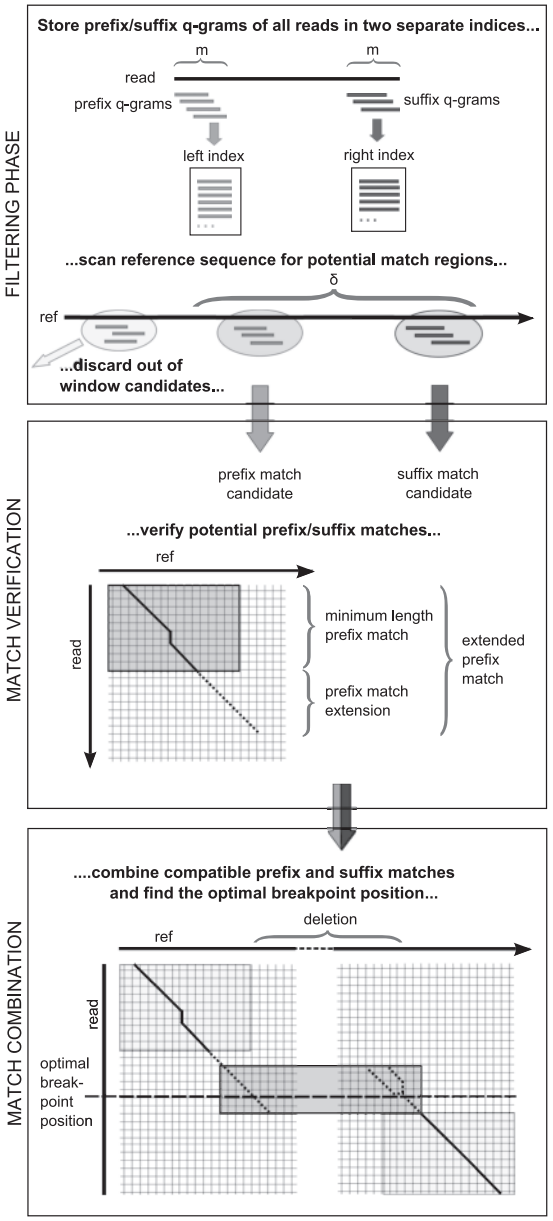

Fig. 2. Overview of the SplazerS algorithm.

or gapped X-drop extension, extending the match as far as possible within the allowed overall error rate ϵ.

2.2.3 Prefix and suffix matches are combined to identify optimal breakpoint location Whenever a prefix match a_p and a suffix match a_s are to be combined into a split read match, we first check whether the basic criteria defined in the previous section can be fulfilled. We make use of the fact that a_p and a_s have been extended as far as possible, collecting the maximum total number of errors allowed on the whole read. Therefore, if extended prefix and suffix match would indicate a deletion but together do not cover the entire read sequence, i.e. $j+1<k$ and $|p|+|s|<|r|$, then prefix p and suffix s are too short to be combined into a valid split read match. Also, if prefix and suffix match are too close to each other, i.e. $l-i<2m$ (in the case of edit distance: $l-i<2m-e_p-e_s$), then prefix and suffix cannot both fulfill the minimum match length and therefore cannot constitute a valid

Downloaded from http://bioinformatics.oxfordjournals.org/ at MPI Molec Genetics on March 28, 2013

split read match. By convention, the gap is placed to the leftmost position that yields the lowest sum of errors in the prefix and suffix match. Again, this is achieved by a simple scanning of diagonals in the case of Hamming distance. For edit distance, a banded alignment matrix is computed to identify the optimal combination of prefix and suffix match and thereby the optimal gap placement.

2.3 Match scoring and ranking

Each match with a middle gap of length ≥ 1 receives a score $sc = |p| + |s| - 2(d(a_p) + d(a_s)) - c(r)$, while matches without a middle gap, i.e. matches that map 'normally' and do not indicate an indel event, receive score $sc = |p| + |s| - 2(d(a_p) + d(a_s))$. The parameter $c(r)$ puts a penalty on the existence of a gap independent of its length and is set to $\lfloor 0.03 \times |r| \rfloor$ by default. It can be set individually, depending on the error-proneness of the reads. Matches with the same score are ranked according to the length of the middle gap (the shorter the better). By default, a match is tagged as unique if it is the single match with highest score. Uniqueness can be extended to a score range, i.e. discarding reads that have more than a certain number of matches within a range of the highest observed score. Multiple and also suboptimal matches can be reported.

2.4 Choice of mapping parameters

There are several parameters that influence our method. In most cases, the choice of the error rate ϵ and therefore also e_p and e_s is rather straightforward, depending mainly on the error pattern specific to the sequencing technology used and/or on the relatedness of sample and reference genome. The choice of parameters m and δ is mainly a matter of trade-off between sensitivity and specificity, as well as runtime. In order to aid the user in choosing sensible parameters, SplazerS provides in its verbose mode an estimate of the number of random matches one expects in a random sequence using the chosen parameter setting. To calculate these values, we use binomial statistics (see Supplementary Material S1).

In the following results, we usually set $\epsilon = 0.05$, $m = 16$ and $e_p = e_s = 1$ unless stated otherwise. The distance parameter δ will be set individually to values between 5000 and 50 000.

2.5 Indel detection

Once we have split-mapped a set of reads, we use all unique best matches to identify indel locations. For this purpose, we use the SeqAn tool snpStore (http://www.seqan.de/projects/snpStore). The indel calling procedure is mainly based on two thresholds: a minimum number and a minimum percentage of spanning reads that are required to support the indel candidate. Each pair of reference position and indel length observed in the mapped reads is considered an indel candidate. For reads that contradict an indel, a minimum overlap with the indel candidate position can be required. Figure 3 illustrates the concept. We usually require at least three indel-supporting reads, and a percentage of at least 25–50% of support from spanning reads where we consider only reads overlapping >5% of their total length.

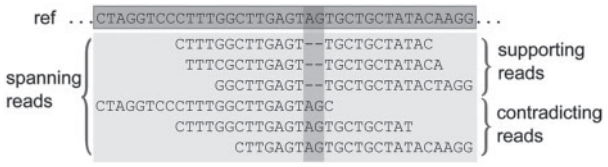

Fig. 3. An indel candidate with six spanning reads. Three reads, i.e. 50% of spanning reads, support the 2 bp deletion. The other three reads contradict it. If reads overlapping <5 bp are not considered contradicting, then the percentage of support is 60% (three out of five reads, as one contradicting read is not considered).

2.6 Evaluation data

All single-end-only evaluation analyses were conducted using NCBI build 36 of the human genome as reference sequence. The paired-end evaluation analysis uses NCBI build 37, as we use already mapped reads from the 1000 Genomes project. Furthermore, we use the variation databases dbSNP (Sherry *et al.*, 2001) (version 130) and the Database of Genomic Variants (Iafrate *et al.*, 2004) (DGV indels version10). Details on used datasets and program calls are given in Supplementary Materials S2 and S4.

2.6.1 1000 Genomes Project data We downloaded two files of Illumina reads for HapMap individual NA12878, available from the 1000 Genomes project page: one containing reads mapped/assigned to chromosome 22 and the other containing unmapped reads that could not be assigned to a chromosome. We extracted a subset of 76 bp reads accounting for $\sim 20\times$ coverage and used all unmapped but anchored reads ($\sim 1M$) as input for Pindel (version 2.2) and SplazerS in paired-end mode. Additionally, we used a total of $\sim 40M$ unmapped 76 bp reads as input for SplazerS in single-end mode. For a fair comparison, we use the same cutoff as Pindel and thus require at least three indel-supporting reads. We furthermore require the indel-supporting reads to constitute at least 50% of reads spanning the putative indel coordinate. Again similar to Pindel's settings, we set SplazerS' maximum distance parameter such that deletions up to 8 kb can be detected. Details on program calls are given in the Supplementary Material S2. We compare the indel prediction results with two reference sets: (i) from the 1000 Genomes project (Durbin *et al.*, 2010) and (ii) from a recent Sanger sequencing study (Mills *et al.*, 2011b).

2.6.2 Simulated single-end data For the simulation of single-end reads (details given in Supplementary Material S4.1), we first generate a manipulated reference sequence by randomly choosing 1000 known indels from a reference set (dbSNP+DGV) and implanting them into human chromosome 21. Our sampling procedure does not represent a realistic indel distribution, but gives us sufficient sample size for testing different indel size ranges, in particular medium- to large-sized indels. Furthermore, we add single base substitutions at a rate of 0.001 to simulate SNPs. We then generate single-end reads from the manipulated chromosome, using the Mason read simulator (Holtgrewe, 2010) with typical Illumina sequencing error settings (position-specific error probabilities increasing from 5' to 3' end). We repeated this simulation procedure with different read lengths: 100, 125 and 150 bp. Simulated coverages are 5, 10 and 30×. After mapping the set of simulated reads onto the whole human reference genome with a 'normal' ungapped mapping approach [RazerS (Weese *et al.*, 2009) with 5% error rate], we retrieve all unmapped reads for subsequent split mapping. The unmapped reads are split-mapped with SplazerS and, for comparison, with GSNAP (version 2010-07-27) and BWA (version 0.5.8a). For all tools, only one gap of at most length $\delta = 5000$ was allowed on each read. For BWA, we tested the 'log-scaled gap penalty for long deletions' feature, but achieved better results with the *n*-difference mode which we consequently use. We then use the same indel detection method for all three tools (snpStore, see Section 2.5), as it can detect indels of all size ranges. Only unique best hits are used for indel calling for GSNAP, and only hits with mapping quality >0 are used for BWA. We set the indel detection method very sensitively: at least two reads and 25% of spanning reads are required to support the indel. Parametrization details and program calls are given in Supplementary Material S4.2.

2.6.3 Real single-end data We used our method in a large-scale targeted resequencing study of 248 male patients with X-linked intellectual disability collected by the EURO-MRX consortium (Kalscheuer,V.M. et al., manuscript in preparation). X chromosome exons were targeted by solution hybridization selection (SureSelect, Agilent) (Johnston *et al.*, 2010). Purified, captured DNA was then PCR amplified and sequenced on the Solexa Genome Analyzer GAIIx, yielding single-end reads of length 76 bp. After mapping onto the human reference genome using edit distance, all unmapped reads

Downloaded from http://bioinformatics.oxfordjournals.org/ at MPI Molec Genetics on March 28, 2013

were retrieved for mapping with SplazerS. This constituted a total of almost 1.5 billion unmapped reads (on average ~6M per patient). With $m = 23$, split mapping was rather strict. Indel detection was done using different post-processing scripts. However, the same basic method was used, requiring that at least three reads and 50% of spanning reads support the indel.

2.7 Evaluation methods

2.7.1 Indel comparison Comparing a predicted indel with a reference indel is not trivial: indels, especially if located in a tandem repeat, can be placed at distances >50 bp while still constituting the same basic indel event. For indel rectification, we adopt the computation of the *extended indel region*, as defined by Krawitz (Krawitz *et al.*, 2010), which accounts for repeats and gives a window of possible locations for each indel. Predicted indel size is allowed to vary by 10% of reference indel size in all real datasets, but has to be exactly the same as implanted indel size in the simulation experiments.

2.7.2 Sensitivity and PPV We will use the measures *sensitivity* and positive predictive value (PPV) in the evaluation of the simulated single-end data.

$$\text{Sensitivity} = \frac{\text{TP}}{|\text{I}_\text{m}|} \qquad \text{and} \qquad \text{PPV} = \frac{\text{TP}}{|\text{I}_\text{p}|}$$

where true positives (TP) are computed by comparing the set of predicted indels I_p with the set of implanted indels I_m using the indel comparison method of the previous subsection.

3 RESULTS

First, we will evaluate our method's ability to accurately identify and locate genomic indel variants in paired-end data, comparing SplazerS with Pindel. Then we will investigate whether our method is also able to accurately detect indels in the simulated single-end datasets, comparing results with GSNAP and BWA. Finally, we will demonstrate our method's capabilities on real large-scale single-end data from 248 patients, providing PCR validation for an especially interesting scenario of predicted indels.

3.1 Paired-end sequencing data: anchored indel detection proves high sensitivity

The results for comparing indels predicted by SplazerS and Pindel on anchored paired-end reads are summarized in Table 1. Since we use previously mapped reads from the 1000 Genomes Project webpage, a large part of the chromosome is already covered by reads mapped also with small indels. Thus, our detected indels do not constitute the whole set of chromosome 22 indels, but rather additional ones discovered through split mapping.

Using anchored reads only, the SplazerS approach yielded a total of 392 indel calls, 301 (76.8%) of which are contained in at least one of the two reference sets [1000G (Durbin *et al.*, 2010) and Mills (Mills *et al.*, 2011b)]. Pindel called only 209 indels of which 142 (67.9%) are in one of the reference sets. The Pindel and SplazerS call sets overlap in 130 indels (62.2% of Pindel call set). Of the 79 indels unique to Pindel, 44 (55.7%) correspond to an indel in the reference set. Of the 262 indels unique to SplazerS, 195 (74.4%) are in the reference set. These results do not only prove a significantly higher sensitivity for SplazerS, but also indicate higher specificity. Adding SplazerS' unanchored split mapping results, the SplazerS set of called indels increases from 392 to 534 (last column in Table 1). Of the additional 142 indels, 90 (63.4%) are contained in one of the reference sets. This indicates a slight decrease in indel calling specificity, but at the gain of much improved sensitivity: >30% additional indels are attributed to single-end split mapping.

Table 1. Number of detected indels on 1000 Genomes Project dataset for NA12878

		Pindel	SplazerS PE	SplazerS PE + SE
Small indels	Deletions	88	183	233
	Overlap 1000G	55	127	161
	Overlap Mills	56	112	142
	Insertions	62	105	145
	Overlap 1000G	35	58	82
	Overlap Mills	42	83	111
Medium indels	Deletions	25	60	83
	Overlap 1000G	4	17	19
	Overlap Mills	8	24	32
	Insertions	24	28	40
	Overlap 1000G	0	0	0
	Overlap Mills	12	21	26
Large deletions	Deletions	10	13	27
	Overlap 1000G	2	3	3
	Overlap Mills	0	1	1
	SV Deletions	0	3	6
	Overlap 1000G	0	3	5
	Overlap Mills	0	0	0

Pindel and SplazerS PE use anchored reads only, SplazerS PE + SE additionally uses unanchored reads. Small indels are ≤10 bp. Medium indels are >10 bp, ≤50 bp. Large deletions are >50 bp, ≤1000 bp. Large SV deletions are >1 kb, ≤5 kb.

In summary, SplazerS was able to recover 391 known indels, while Pindel could only recover 142. In particular, SplazerS is more sensitive and robust in variant-rich regions, as revealed by additional analyses on a high coverage 100 bp Illumina read dataset (Supplementary Material S3 and Fig. S1). Our analyses suggest that Pindel's sensitivity is between 50% and 70% of SplazerS' sensitivity on anchored reads only (Table 1 and Supplementary Table S2). Applying the edit distance feature exhibited a further increase in sensitivity of ~2.6% (Supplementary Table S2).

3.2 Single-end reads: simulations demonstrate high accuracy of split read approach

On the simulated single-end read dataset, we tested the indel prediction accuracy of different tools for varying read lengths (100–150 bp) and coverages (5–30×). Figure 4 shows the results in terms of sensitivity. Corresponding p measurements are given in Supplementary Figure S2.

Table 2 shows the results in terms of sensitivity and PPV for 125 bp reads at ~30× coverage, dividing indels into different size ranges: three small indel classes for insertions/deletions of 1–3 bp, 4–9 bp and 10–50 bp in size; and three large structural variant classes for sizes 51–500 bp, 501 bp–1 kb and >1 kb.

Figure 4 shows how sensitivity increases with coverage and with read length. In all settings, SplazerS is the most sensitive and most precise tool. At 30×, it already recovers >90% of implanted indels on the 100 bp reads; for the 150 bp reads sensitivity reaches >95%. GSNAP is always a few percentage points behind SplazerS, with

Downloaded from http://bioinformatics.oxfordjournals.org/ at MPI Molec Genetics on March 28, 2013

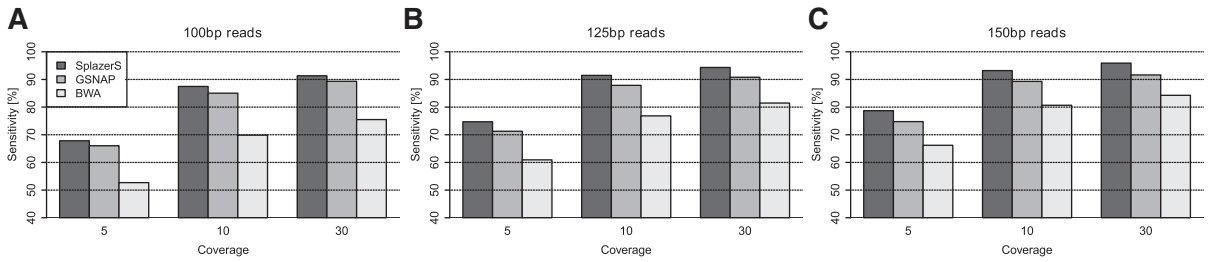

Fig. 4. Sensitivity of indel detection for increasing coverage and increasing read length. (**A**) Read length 100 bp. (**B**) Read length 125 bp. (**C**) Read length 150 bp. Note the axis scaling.

SplazerS' lead being more pronounced on longer reads and higher coverage (1.8% lead on 100 bp at 5× compared with 4.3% on 150 bp at 30×). Due to its exact 12mer matching approach, GSNAP systematically fails to detect certain indels, even in high coverage data. BWA is less sensitive and less precise than the other tools. Table 2 shows that BWA's sensitivity mainly suffers from missed deletions >50 bp. With increased read length, it can also detect larger indels and thus its overall sensitivity increases.

For indels <10 bp, BWA is the most sensitive tool, but with the lowest PPV. For indels ≥10 bp, SplazerS has the highest sensitivity. Furthermore, it maintains the highest PPV in all indel categories. Most notably, SplazerS achieves the highest sensitivities in the SV deletion categories. Both GSNAP and SplazerS exhibit a 'temporary' drop in sensitivity for the smallest SV deletion class. The low sensitivity is due to low complexity and repeat sequences where reads are often ambiguously mappable or even wrongly mapped without a middle gap. Nevertheless, SplazerS maintains a high PPV and higher sensitivity than GSNAP for these difficult-to-map indels. In order to investigate whether our results are robust with respect to different indel calling programs, we conducted an additional analysis replacing our in-house tool snpStore with Dindel (Albers *et al.*, 2010). This analysis exhibited the same relative sensitivity results for SplazerS, GSNAP and BWA, and furthermore demonstrated that SplazerS was again the most accurate tool in terms of sensitivity as well as PPV (Supplementary Table S4).

Table 2. SN and PPV results of simulations at 30× coverage with read length of 125 bp, for different indel size categories

		SplazerS		GSNAP		BWA	
		SN	PPV	SN	PPV	SN	PPV
Ins	10–30 bp	99.35	99.33	95.87	98.63	95.21	96.83
	4–9 bp	98.29	98.21	97.44	96.62	100.0	76.39
	1–3 bp	98.78	99.65	98.60	99.47	98.94	93.31
Del	1–3 bp	96.03	99.80	96.00	98.74	96.99	92.30
	4–9 bp	94.54	100.0	92.26	100.0	94.54	87.17
	10–50 bp	92.73	99.52	85.18	98.12	87.51	93.71
SV Del	51–500 bp	70.98	98.38	61.13	89.37	0	NA
	0.5–1 kb	94.69	100.0	92.47	97.56	0	NA
	1–5 kb	100.00	94.44	96.67	90.61	0	NA

Each category has at least 50 representatives. SN, sensitivity; PPV, positive predictive value.

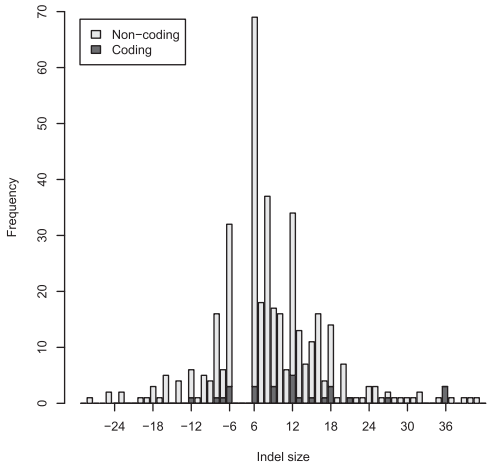

Fig. 5. Histogram of indels of sizes >5 bp and ≤40 bp. Indels are more abundant in non-coding sequences. The majority of indels in coding sequences are multiples of three, i.e. codon-length.

3.3 Application to real single-end data demonstrates versatility

Using SplazerS in the targeted exon resequencing study, on average 67 indels were predicted per patient. The overlap of predicted indels with dbSNP and DGV was between 38.89% and 71.79% per patient, with mean overlap of 54.99%. Of the total set of indels predicted in at least one patient, 39.02% were present in dbSNP or DGV.

Figure 5 shows the size distribution of all indels >5 bp (the majority of smaller indels were predicted with a different method using edit distance alignments). As expected, the majority of indels is located in non-coding sequences (417 out of 456). Non-coding indels occur mostly in tandem repeat regions in units of 2, 3 or 4. Of the 39 coding indels, 29 (74.35%) are multiples of 3, usually having lesser impact on the protein level.

Large deletions ≥100 bp are rather rare (61 in total). On average, three large deletions were predicted per patient. Interestingly, we observed a strong variance between patients. Closer inspection revealed that locations of large deletions often cluster on the chromosome. Figure 6A visualizes these clusters in 1 Mb bins over the X chromosome. Surprisingly, deletions do not only co-localize, but often their boundaries also coincide exactly with annotated exon–intron boundaries. Figure 6B shows one such case where five predicted deletions span all introns of the *PQBP1* gene. These 'intron deletions' strongly suggest the presence of a retrocopy of this gene.

Downloaded from http://bioinformatics.oxfordjournals.org/ at MPI Molec Genetics on March 28, 2013

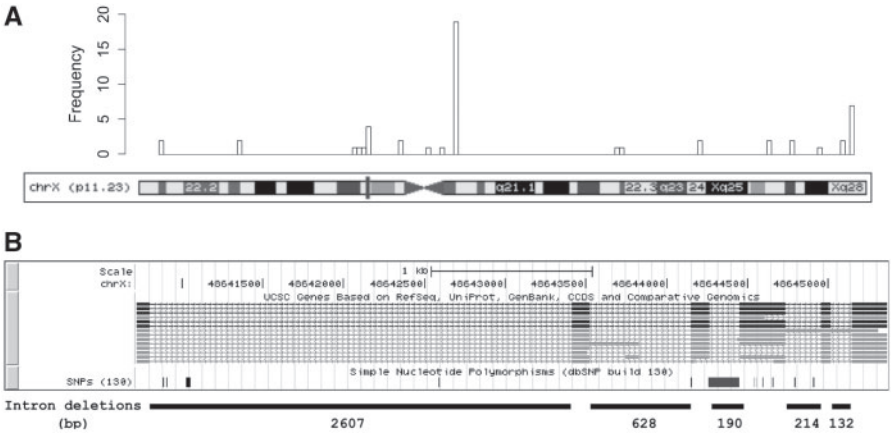

Fig. 6. (**A**) Histogram of predicted deletions ≥100 bp over their genomic coordinate on chromosome X. Clusters of large deletions are often due to retroposed genes, where spliced introns are missing. (**B**) A screenshot of the UCSC Genome Browser shows five large deletions that coincide exactly with the introns of the *PQBP1* gene. The existence of this complete retrocopy of *PQBP1* was confirmed by PCR.

PCR experiments with several primer pairs confirmed a complete, possibly functional retrocopy of *PQBP1*. This finding is particularly interesting, as it has been shown previously that mutations in *PQBP1* cause X-linked intellectual disability (Kalscheuer *et al.*, 2003; Lenski *et al.*, 2004). Additional (partial) retrocopies were predicted for *FAM104B*, *MSN*, *MPP1*, *EIF1AX*, *RBMX* and *OPHN1*. In the OPHN1 gene, large deletions spanned 19 introns, causing the large peak close to the centromer in Figure 6A.

3.4 Running times and memory

SplazerS' high sensitivity comes at the price of increased running time compared with index-based heuristic mappers. Nevertheless, it is applicable to large-scale datasets such as the one in the previous section. In particular, the parameter m can be used to achieve a significant speedup without compromising sensitivity for high-coverage datasets and longer read lengths (Table 3 and Supplementary Table S2). The observed memory increase with m is explained by a larger value of q being used for q-gram index construction. Running time disadvantage nearly disappears when mapping onto a smaller reference sequences, as demonstrated by mapping the simulation reads onto chromosome 21 only. Mapping was performed on a computing cluster, splitting up the reads into batches of 100K. Future developments will include parallelization of SplazerS.

Table 3. Running time and memory measurements for 100 000 simulated 125 bp reads

	BWA	GSNAP	SplazerS $m=16$	SplazerS $m=20$
chr21				
index	49.8 s	12.2 s	–	–
time	44.4 s	49.7 s	143.5 s	52.9 s
space	154 Mb	122 Mb	185 Mb	1.2 Gb
genome				
index	94.0 m	16.9 m	–	–
time	10.8 m	18.2 m	193.2 m	57.4 m
space	3.7 Gb	4.6 Gb	3.5 Gb	5.6 Gb

SplazerS runs are shown for different minimum match lengths (m). BWA and GSNAP require an additional preprocessing step for index construction.

4 DISCUSSION

The availability of various kinds of sequencing data, i.e. derived from different sequencing technologies and protocols, has fueled the development of various computational tools for indel identification [for reviews, see Alkan *et al.* (2011) and Medvedev *et al.* (2009)]. Often the overlap of predicted indels between different computational methods is low (Mills *et al.*, 2011a), indicating that none of the methods is fully comprehensive and a satisfactory solution is yet to be found. Therefore, a comprehensive method is likely to use an integrative approach, combining the strengths of different methods.

The split read mapping approach has its strength in the potential to exactly predict indel size and location. Early next-generation technologies yielded very short read lengths (36 bp) where only anchored split mapping was feasible, with the mapping search space greatly reduced by a confidently mapped paired end (Ye *et al.*, 2009). However, advances in sequencing technology have led to single-end reads long enough to reliably predict long deletions and also medium-sized insertions without anchoring. SplazerS supports both anchored paired-end split mapping as well as unanchored single-end split mapping, which is a unique feature among split read mapping tools. By adding unmapped paired-end reads and treating them as single-end data, a significant increase in sensitivity could be demonstrated.

In our comparisons with Pindel (Ye *et al.*, 2009), which to our knowledge is the most widely used paired-end split read indel detection method, and with state-of-the-art single-end split read aligners, SplazerS showed highest PPV and highest sensitivity, especially in variant-rich regions. This strength may prove especially

Downloaded from http://bioinformatics.oxfordjournals.org/ at MPI Molec Genetics on March 28, 2013

valuable when mapping reads from highly mutated or diverged genomes to a related reference, as is the case in cancer genome sequencing (Stratton, 2011), or in evolutionary genetics studies such as Green *et al.* (2010) where Neanderthal DNA was mapped onto a human and a chimpanzee reference sequence.

On real data, a large overlap with annotated variation was obtained when using SplazerS. In addition, predicted indels followed the expected pattern of indel size distributions for coding as well as non-coding sequences (Durbin *et al.*, 2010; Ng *et al.*, 2009). Furthermore, the percentage of coding, in-frame indels is in agreement with the previous studies (Ng *et al.*, 2009).

Altogether we show that our method is versatile as it is applicable to anchored paired-end as well as unanchored and single-end data, and is not constrained to short read lengths. Even more, its sensitivity proved to further increase with read length, making its application to upcoming longer reads promising. While not explicitly tested, SplazerS' edit-distance feature will also allow application to 454 sequencing reads (Wheeler *et al.*, 2008).

In this work, we used the mapping results of SplazerS in conjunction with a simple indel detection method. Previously, it has been shown that realignment of reads at indel candidate positions can significantly improve indel prediction accuracy (Albers *et al.*, 2010; Homer and Nelson, 2010). We expect that SplazerS' support for SAM output format will make it easy to integrate with other indel detection tools, which may further improve accuracy of split read indel detection.

In contrast to Pindel, inversions are not yet handled by SplazerS. We are currently investigating strategies to generalize the split-read approach to detect complex structural variants including interchromosomal translocations which neither Pindel nor SplazerS can handle. However, the successful detection of retrocopy events suggests that SplazerS may also be applicable for spliced mapping of RNA-seq data.

ACKNOWLEDGEMENTS

A.-K.E. would like to thank Birte Kehr for helpful discussions and Ole Schulz-Trieglaff and Illumina for sharing the indel reference set for NA18507 used in the Supplementary Material.

Funding: European Union's Seventh Framework Program under grant agreement number 241995, project GENCODYS; International Max Planck Research School for Computational Biology and Scientific Computing.

Conflict of Interest: none declared.

REFERENCES

Albers,C.A. *et al.* (2010) Dindel: accurate indel calls from short-read data. *Genome Res.*, **21**, 961–973.

Alkan,C. *et al.* (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**, 363–376.

Ameur,A. *et al.* (2010) Global and unbiased detection of splice junctions from RNA-seq data. *Genome Biol.*, **11**, R34.

Au,K.F. *et al.* (2010) Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res.*, **38**, 4570–4578.

Barski,A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

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

Burkhardt,S. *et al.* (1999) Q-gram based database searching using a suffix array (Quasar). In *Proceedings of the Third Annual International Conference on Computational Molecular Biology*, ACM press, pp. 77–83.

Chen,W. *et al.* (2008) Mapping translocation breakpoints by next-generation sequencing. *Genome Res.*, **18**, 1143–1149.

Chen,K. *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.

Döring,A. *et al.* (2008) SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinf.*, **9**, 11.

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

Eid,J. *et al.* (2009) Real-time dna sequencing from single polymerase molecules. *Science*, **323**, 133–138.

Green,R.E. *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, **328**, 710–722.

Holtgrewe,M. (2010) Mason – a read simulator for second generation sequencing data. *Technical Report TR-B-10-06*, Institut für Mathematik und Informatik, Freie Universität Berlin.

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.

Iafrate,A.J. *et al.* (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949–951.

Johnston,J.J. *et al.* (2010) Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. *Am. J. Hum. Genet.*, **86**, 743–748.

Kalscheuer,V.M. *et al.* (2003) Mutations in the polyglutamine binding protein 1 gene cause X-linked mental retardation. *Nat. Genet.*, **35**, 313–315.

Korbel,J.O. *et al.* (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol.*, **10**, R23.

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

Lee,S. *et al.* (2009) MoDIL: detecting small indels from clone-end sequencing with mixtures of distributions. *Nat. Methods*, **6**, 473–474.

Lenski,C. *et al.* (2004) Novel truncating mutations in the polyglutamine tract binding protein 1 gene (PQBP1) cause Renpenning syndrome and X-linked mental retardation in another family with microcephaly. *Am. J. Hum. Genet.*, **74**, 777–780.

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

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.

Medvedev,P. *et al.* (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods*, **6** (11 Suppl.), S13–S20.

Metzker,M.L. (2010) Sequencing technologies - the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

Mills,R.E. *et al.* (2011a) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.

Mills,R.E. *et al.* (2011b) Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.*, **21**, 830–839.

Mullaney,J.M. *et al.* (2010) Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.*, **19**, R131–R136.

Myers,G. (1999) A fast bit-vector algorithm for approximate string matching based on dynamic programming. *J. ACM*, **46**, 395–415.

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

Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.*, **37** (Suppl.), S11–S17.

Rasmussen,K. *et al.* (2005) Efficient q-gram filters for finding all epsilon-matches over a given length. In *Proceedings of the Nineth Conference on Computational Molecular Biology*, Springer, pp. 189–203.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Stankiewicz,P. and Lupski,J.R. (2010) Structural variation in the human genome and its role in disease. *Annu. Rev. Med.*, **61**, 437–455.

Stenson,P.D. *et al.* (2009) The human gene mutation database: 2008 update. *Genome Med.*, **1**, 13.

Stratton,M.R. (2011) Exploring the genomes of cancer cells: progress and promise. *Science*, **331**, 1553–1558.

Downloaded from http://bioinformatics.oxfordjournals.org/ at MPI Molec Genetics on March 28, 2013

Wang,Z. *et al.* (2009) RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.

Wang,K. *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, **38**, e178.

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.

Wu,T.D. and Nacu,S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.

Xie,C. and Tammi,M.T. (2009) CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics*, **10**, 80.

Ye,K. *et al.* (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.

Yoon,S. *et al.* (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592.

Zeitouni,B. *et al.* (2010) SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*, **26**, 1895–1896.

Downloaded from http://bioinformatics.oxfordjournals.org/ at MPI Molec Genetics on March 28, 2013