Bioinformatics, 2016, 1–7
doi: 10.1093/bioinformatics/btw516
Advance Access Publication Date: 8 August 2016
Original Paper

OXFORD

Sequence analysis

SRinversion: a tool for detecting short inversions by splitting and re-aligning poorly mapped and unmapped sequencing reads

Ruoyan Chen[1], Yu Lung Lau[1,2], Yan Zhang[1] and Wanling Yang[1,*]

[1]Department of Paediatrics and Adolescent Medicine, LKS Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong and [2]The University of Hong Kong-Shenzhen Hospital, Shenzhen, China

*To whom correspondence should be addressed.
Associate Editor: Bonnie Berger

Received on February 19, 2016; revised on August 1, 2016; accepted on August 2, 2016

Abstract

Motivation: Rapid development in sequencing technologies has dramatically improved our ability to detect genetic variants in human genome. However, current methods have variable sensitivities in detecting different types of genetic variants. One type of such genetic variants that is especially hard to detect is inversions. Analysis of public databases showed that few short inversions have been reported so far. Unlike reads that contain small insertions or deletions, which will be considered through gap alignment, reads carrying short inversions often have poor mapping quality or are unmapped, thus are often not further considered. As a result, the majority of short inversions might have been overlooked and require special algorithms for their detection.
Results: Here, we introduce SRinversion, a framework to analyze poorly mapped or unmapped reads by splitting and re-aligning them for the purpose of inversion detection. SRinversion is very sensitive to small inversions and can detect those less than 10 bp in size. We applied SRinversion to both simulated data and high-coverage sequencing data from the 1000 Genomes Project and compared the results with those from Pindel, BreakDancer, DELLY, Gustaf and MID. A better performance of SRinversion was achieved for both datasets for the detection of small inversions.
Availability and Implementation: SRinversion is implemented in Perl and is publicly available at http://paed.hku.hk/genome/software/SRinversion/index.html.
Contact: yangwl@hku.hk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Inversions are a type of structural variations (SVs) that may play an important role in human genetic diversity and disease susceptibility (Antonacci *et al.*, 2009; Bondeson *et al.*, 1995; Feuk *et al.*, 2005, 2006; Giglio *et al.*, 2001, 2002; Gimelli *et al.*, 2003; Koolen *et al.*, 2006; Kurotaki *et al.*, 2003; Lakich *et al.*, 1993; Osborne *et al.*, 2001; Stefansson *et al.*, 2005; Steinberg *et al.*, 2012; Visser *et al.*, 2005; Zody *et al.*, 2008). On one hand, inversions can be directly related to human diseases, such as the inversion that interrupts the coagulation factor VIII (F8) gene, causing haemophilia A (Lakich *et al.*, 1993), or the iduronate 2-sulphatase (IDS) gene that causes Hunter syndrome (Bondeson *et al.*, 1995). Inversions can also have functional impact by either inducing or being associated with other structural changes. It has been reported that inversions in parental genomes could induce disease-associated copy number variations (CNV) in the offspring (Feuk *et al.*, 2006; Gimelli *et al.*, 2003; Osborne *et al.*, 1995). Despite these examples that highlight the functional significance of inversions, our ability to detect this type of genetic variants remains limited.

Over the past decade, the rapid development and extended use of next-generation sequencing technology have widened the spectrum of SVs detected (Bentley *et al.*, 2008; Iafrate *et al.*, 2004; Korbel *et al.*, 2007; McKernan *et al.*, 2009; Stankiewicz and Lupski,

2010; The 1000 Genomes Project Consortium, 2012; Tuzun *et al.*, 2005; Wheeler *et al.*, 2008). Nevertheless, unlike other SVs such as deletions, insertions and CNVs, detection of inversions has been lagging behind. Analysis of sequencing data was reported to be able to find around 80% of known deletions, but for inversions the fraction of detection is still insignificant so far (Baker, 2012). As shown in Supplementary Figure S1, summary of all records on SVs to date from public databases such as dbVAR (Church *et al.*, 2010) and DGVa (MacDonald *et al.*, 2014) demonstrated an obvious lack of documentation of inversions, especially for small inversions (1 bp to around 100 bp). This is inconsistent with previous suggestions that small SVs are much more frequent in the genome than large SVs (Conrad *et al.*, 2008; Hurles *et al.*, 2008), suggesting that the majority of inversions might remain to be identified.

Several computational methods are available to detect SVs based on short read sequencing data, most of which have the function for inversion detection (Chen *et al.*, 2009; He *et al.*, 2016; Rausch *et al.*, 2012; Trappe *et al.*, 2014; Ye *et al.*, 2009). But only one (Trappe *et al.*, 2014) of these tools was designed specifically for detecting short inversions. Most of these tools rely on signals of inconsistent mapping directions of the reads in a read pair in paired-end sequencing (Chen *et al.*, 2009; Rausch *et al.*, 2012), which requires both reads of the pair to be mappable to the reference genome and one read to almost entirely overlap with the inversion. As a consequence, this limits the minimum length of inversions that can be detected to at least longer than the read length. Software such as Pindel (Ye *et al.*, 2009) tries to solve this problem by combining paired-end method with split-read method, which makes it possible to detect inversions that are shorter than the read length. However, by requiring at least one end of the read to be mapped to the reference genome to serve as an anchor, it loses power in detecting certain inversions such as those adjacent to other complex SVs or repeat regions, which are common for inversions as reported previously (Stankiewicz and Lupski, 2010) and in this study.

Here we introduce SRinversion, a framework that applies a split read method on next-generation sequencing (NGS) data to detect inversions smaller than 1kb. Instead of relying only on mapped reads as anchors, SRinversion also makes use of poorly mapped or unmapped reads, splits or inverts them, and then re-aligns them to the reference genome to detect inversions (Fig. 1). By considering the unmapped reads, which are overlooked for inversion detection by other tools, the sensitivity of inversion detection is improved significantly. This is because for short inversions, especially those shorter than the read length, most of the reads covering them are usually poorly mapped or unmapped.

Gustaf (Trappe *et al.*, 2014) and MID (He *et al.*, 2016) also try to use poorly mapped reads to generate split-map contigs for short inversion detection. However, the sensitivities of these methods are affected by their pre-requisites such as requiring both ends of each read be mapped to the reference genome with high confidence.

We applied SRinversion to both simulated data and real data from a parent-child trio from the 1000 Genomes Project. Pindel, BreakDancer, DELLY, Gustaf and MID were also applied on both datasets and the results were compared. A better performance for SRinversion was achieved on both datasets, especially for small inversions.

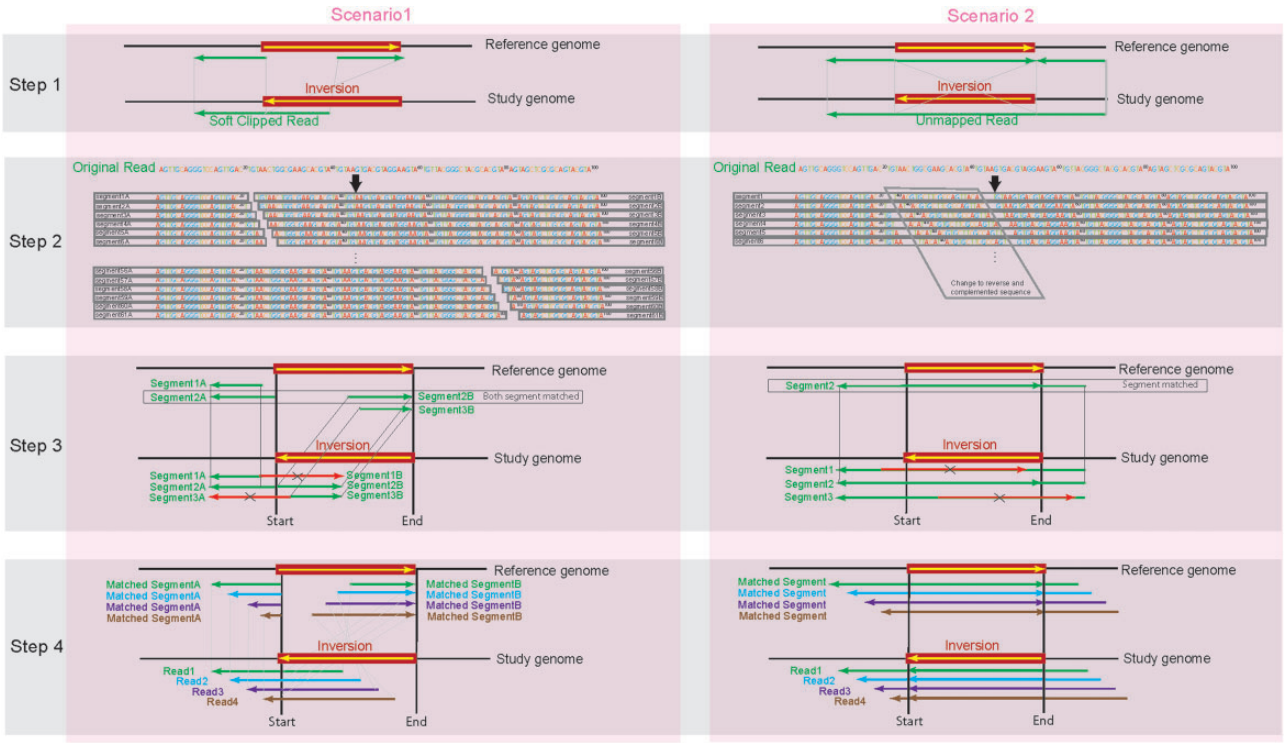

Fig. 1. SRinversion workflow by different scenarios. Step 1: Extract poorly mapped reads. In scenario 1, reads partially covering inversion region and marked as soft clipped by BWA are extracted, and in scenario 2, reads covering the whole inversion region and considered unmapped are extracted. Step 2: Split reads. In scenario 1, each extracted read will be split into two segments at different split sites, and in scenario 2, for each read, sequence within windows of different sizes at different sites will be changed into reverse complement sequence. Step 3: Re-map reads. Segment generated from step 2 are aligned back to the reference and only segment pairs or segments that are exactly matched are selected for step 4. Step 4: Inversion region detection. Each matched segment or segment pair from step 3 produces a start and end position of the inversion it covers, and by merging those positions from different reads, inversion region can be determined

2 Approach

SRinversion consists of four major steps:

1. extract unmapped reads and reads with low mapping quality
2. split or invert reads
3. re-map reads
4. detect inversion region

Particularly two scenarios are considered based on the predicted relationships between candidate reads and corresponding inversions, and different algorithms are applied in the following steps accordingly (Fig. 1).

Specifically, in scenario 1, a read is assumed to partially overlap with an inversion, and one end of the read cannot be mapped to the reference genome and hence is marked as 'clipping' in alignment by tools such as BWA (Li and Durbin, 2009). In this case, SRinversion would (1) extract the clipped reads as candidates, (2) splits each candidate read into pairs of segments at different sites along the read, (3) re-align the split segments from step (2) to the reference genome to detect the primary start and end positions of a corresponding inversion and (4) integrate position information from step (3) from all candidate reads covering the same inversion and to count the number of reads supporting the inversion.

In scenario 2, we assume that the candidate read encompasses the entire inversion, giving rise to an inverted sequence in the middle of the read, and these reads are thus most likely to be marked as 'unmapped' after the initial alignment by most of mapping algorithms. To deal with this situation, (1) all unmapped reads are extracted regardless of the status of their read pairs and (2) the middle part of these extracted reads are inverted into reverse complement sequences at different positions using a sliding window with pre-set window and step sizes. These modified reads are then realigned to reference human genome and the alignment information are integrated for inversion detection, using procedures similar to Step (3) and (4) described above for scenario 1.

SRinversion tries to include all poorly mapped and unmapped reads to increase sensitivity in inversion detection. Two key points make SRinversion better at detecting short inversions than existing SV calling methods: (1) SRinversion extracts all clipped or unmapped reads as candidate reads regardless of the mapping status of their mate pairs. Software like Pindel and Gustaf also uses clipped reads for SV calling. However, these methods require the mate pair of the clipped reads be mapped to the reference genome with high quality, which may result in exclusion of some informative reads that don't fulfill this criterion. This is especially the case when coverage or quality of the analyzed data is low or the inversion region is located in or near complex regions in human genome. SRinversion tries to solve this problem by considering every single read regardless of situations with its mate pair to ensure inclusion of all potential informative reads and tries to increase specificity by applying additional filtering steps. (2) In scenario 2, unmapped reads, either with one end or both ends unmapped, are analyzed, making it possible to detect inversions so short as to invert only the middle part of the sequences of the reads involved, which are likely to be missed by the existing methods.

3 Materials and methods

3.1 Overview of SRinversion

3.1.1 Extracting poorly mapped reads

SRinversion accepts binary aligned reads in BAM format (Li and Durbin, 2009) as input, which include all the original sequencing reads with alignment quality parameters. In scenario 1 (Fig. 1), reads with *n*bp clipped sequences on either end are obtained regardless of mapping status of their mates, where n needs to be above 5. In scenario 2 (Fig. 1), all unmapped reads, either one or both reads of a pair labeled as unmapped, are collected as candidate reads for further analysis.

3.1.2 Split or convert reads

Reads extracted from the above step are split or inverted at different positions. Specifically, in scenario 1, each candidate read is split at site i, where $i \in k, \ldots, m - k$ and m is the read length, k is the pre-set start position for splitting. In this study, k was set to 20 bp since segments smaller than 20 bp are unlikely to be mapped uniquely and properly to reference human genome. In scenario 2, a sliding window with length w bp and step size s bp is applied from position k to position $m–k–w$, where $w \in k, \ldots, m - k$ and m and k are the same as in scenario 1. Here s was set to 5 as a compromise between computing time and detection resolution. Within each window, sequences are converted into their reverse complement sequences.

3.1.3 Re-map reads

All segments generated from the steps described above are realigned to human reference genome (hg19) using the same (single ended) model from BWA (version 0.7.5a-r405), with the maximum number of alignments output for each read set to 3 as a default.

3.1.4 Inversion region detection

After realignment for each read, the segment or the segment pair that can be uniquely mapped and with mapping quality $\geq$30 is selected for inferring the boundaries of the potential inversions. In addition, in scenario 1, the orientation of each segment pair has to be opposite of each other. Afterwards, reads covering the same inversion are merged to get the depth information for further filtration. In most cases, there would be several base pair deviation on the inversion boundaries derived from different reads, likely due to mapping errors. To be more inclusive, the most extensive coordinates for each inversion are selected.

3.2 Validating the detected inversions by long reads data and PCR

Long read data were used to measure the accuracy of inversion detection on real data. Long-read sequencing data from PacBio or Illumina Moleculo platforms are available for sample NA12878, reported by Layer *et al.* (2014) and in (https://github.com/hall-lab/long-read-validation). For results from the long reads data, inversions smaller than 1Mbp were included, which were detected using parameters such as at least 1 Moleculo split read or at least 2 PacBio split reads, and a slop of 5, as suggested by the original method. In addition, PCR experiment was also conducted on a few detected inversions that are not located near repeats or repetitive sequences.

4 Results

4.1 Datasets

4.1.1 Simulated data

To evaluate the performance of SRinversion, we simulated inversions on human chromosome 21 based on reference human genome hg19. These data were used to examine how well SRinversion and the existing methods perform in detecting the inversions in the presence of SNPs, indels and sequencing errors. First, RSVSim (version 1.6.1) (Bartenhagen and Dugas, 2013) was used for generation of inversions ranging from 5 bp to 1 kb (Supplementary Table S1). The fraction of SNPs within the flanking regions of inversions was simulated as 0.

Table 1. Specificity and sensitivity for simulated datasets with different inversion length

Inversion length			5–20 bp	20–50 bp	50–100 bp	100–200 bp	200–300 bp	300–400 bp	400–500 bp	500 bp–1 kb
SRInversion	**Specificity**	**Heterozygous**	0.87	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**
		Homozygous	0.92	**1.00**	**0.97**	0.98	0.99	**0.99**	**1.00**	**1.00**
	Sensitivity	**Heterozygous**	**0.24**	**0.84**	**0.93**	0.95	0.96	0.95	0.96	0.96
		Homozygous	**0.41**	**0.86**	**0.93**	0.95	0.97	0.95	0.97	0.96
BreakDancer	**Specificity**	**Heterozygous**	0.00	0.00	0.00	0.49	0.72	0.81	0.96	0.95
		Homozygous	NA	0.00	0.12	0.58	0.83	0.85	0.96	0.98
	Sensitivity	**Heterozygous**	0.00	0.00	0.00	0.71	0.94	0.99	**1.00**	**1.00**
		Homozygous	0.00	0.00	0.01	0.82	0.99	0.98	**1.00**	**1.00**
DELLY	**Specificity**	**Heterozygous**	NA	NA	0.77	0.41	0.86	0.49	0.49	0.45
		Homozygous	NA	NA	0.74	0.47	0.95	0.53	0.58	0.53
	Sensitivity	**Heterozygous**	0.00	0.00	0.10	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**
		Homozygous	0.00	0.00	0.10	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
Pindel	**Specificity**	**Heterozygous**	**1.00**	**1.00**	0.95	**1.00**	**1.00**	**1.00**	**1.00**	0.98
		Homozygous	**1.00**	**1.00**	0.89	**1.00**	**1.00**	**0.99**	**1.00**	0.99
	Sensitivity	**Heterozygous**	0.23	0.43	0.45	0.93	0.96	0.99	0.99	0.99
		Homozygous	0.24	0.65	0.72	0.91	0.98	0.99	**1.00**	**1.00**
MID	**Specificity**	**Heterozygous**	0.85	0.75	1.00	NA	NA	NA	NA	NA
		Homozygous	0.72	0.70	NA	NA	NA	NA	NA	NA
	Sensitivity	**Heterozygous**	0.03	0.14	0.01	NA	NA	NA	NA	NA
		Homozygous	0.07	0.23	0.00	NA	NA	NA	NA	NA
Gustaf	**Specificity**	**Heterozygous**	NA	0.51	0.93	0.40	0.99	0.59	0.55	0.73
		Homozygous	NA	0.38	0.52	0.68	0.38	0.96	0.90	0.99
	Sensitivity	**Heterozygous**	0.00	0.33	0.87	0.40	0.95	0.50	0.45	0.64
		Homozygous	0.00	0.27	0.52	0.69	0.36	0.87	0.84	0.95

Method with best specificity or sensitivity in different inversion lengths are marked in bold.

001. Then ART_illumina version 2.1.8 (Huang *et al.*, 2012) was used to generate illumina pair-end reads with 101 bp read length and 250 bp mean insert size and an average coverage of 100 fold.

In order to test the effect of coverage on detection accuracy, we simulated sequencing reads with average coverage depth at 4×, 8×, 16×, 32×, 64× and 128×, respectively, using the same parameters described above. To generate heterozygous inversions, reads simulated directly from the reference genome were merged with reads with simulated inversions, each constitute half of the total reads and the coverage was the same as those of the corresponding datasets with homozygous inversions (Supplementary Fig. S2).

4.1.2 Real data

We also tested our SRinversion and the three existing methods on whole-genome paired-end Illumina sequencing dataset on a CEU HapMap trio from the 1000 Genomes Project. Only data from chromosome 1 and chromosome 21 were used due to the high computational demand running genome-wide analysis. The data have a read length of 101 bp. Data for NA12878 (the daughter) have a 128 fold average sequencing coverage with a mean insert size 417 bp. Data on NA12891 (the father) have a 69 fold sequencing coverage with a mean insert size 328 bp. And data for individual NA12892 (the mother) have a 68 fold sequencing coverage with a mean insert size 326 bp. Aligned reads were downloaded from the ftp site of the 1000 Genomes Project (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20120117_ceu_trio_b37_decoy/)

In addition to the public dataset, we also used two whole-genome paired-end sequencing datasets generated from two different platforms. These data include: (1) Illumina HiSeq X10 sequencing with 35× coverage, a mean insert size of 566 bp and a read length 150 bp and (2) Illumina MiSeq sequencing with 7× sequencing coverage, a mean insert size of 521 bp and a read length 300 bp, both for a single individual (Supplementary Table S4). For the Illumina MiSeq data, the insert size is smaller than the sum of the two reads in a pair (Supplementary Fig. S6a), allowing sequencing overlap for the paired reads. PEAR version 0.9.6 (Zhang *et al.*, 2014) was used to connect the paired reads and we compared this dataset with the original datasets described above. After connection of the paired reads, the maximum read length was increased from 301 bp to 592 bp. Since PEAR tried to select bases with higher quality scores for the overlapped regions, after connection, the base qualities around the 3' ends of the reads were also significantly improved (Supplementary Fig. S6b).

4.2 Performance on simulated inversions

In order to evaluate the performance of SRinversion, we first simulated inversions ranging from 5bp to 1kb on sequences from human chromosome 21, in the presence of SNPs, indels and sequencing errors (see Section 3 for details). Five published methods, i.e. BreakDancer version 1.1, DELLY version 0.3.3, Pindel version 0.2.5a3, Gustaf version 1.0.0 and MID were also applied to the data for comparison (see supplementary for details).

4.2.1 Sensitivity and specificity (Table 1 and Fig. 2)

Specificity was defined as percentage of detected inversions overlapped with inversion regions from simulation over all inversions detected by each method. Sensitivity was defined as percentage of inversions that are detected by each method among all inversions generated in simulation. SRinversion achieved both high sensitivity and specificity for the various inversion lengths (Table 1) on the simulated data. The performance of BreakDancer was high for inversions longer than 500 bp, while worsened quickly when the inversions became shorter (Table 1). As seen from the table, BreakDancer could not detect any inversions smaller than 50 bp, which was also the case for DELLY. Sensitivity of DELLY was high for inversions longer than 100 bp, while specificity was lower

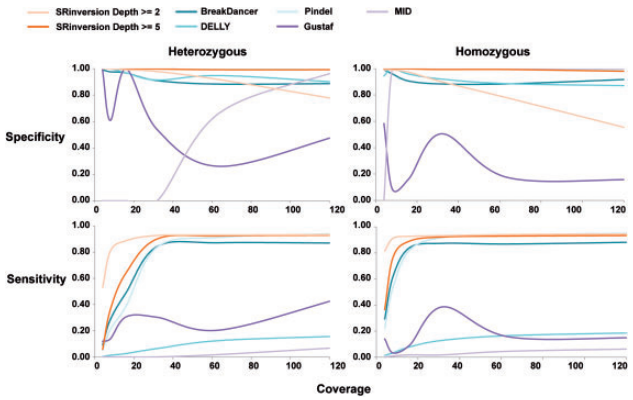

Fig. 2. Effect of coverage on sensitivity and specificity. Both heterozygous and homozygous inversions were simulated on chromosome 21 with coverage ranging from 4× to 128×. BreakDancer, DELLY, Pindel, SRinversion, Gustaf and MID were applied to compare the sensitivity and specificity under different coverage. Depth cutoff of 2× and 5× were applied for SRinversion. For MID, only inversions with length ≤100 bp were included for evaluation since the method can only detect inversions smaller than read length

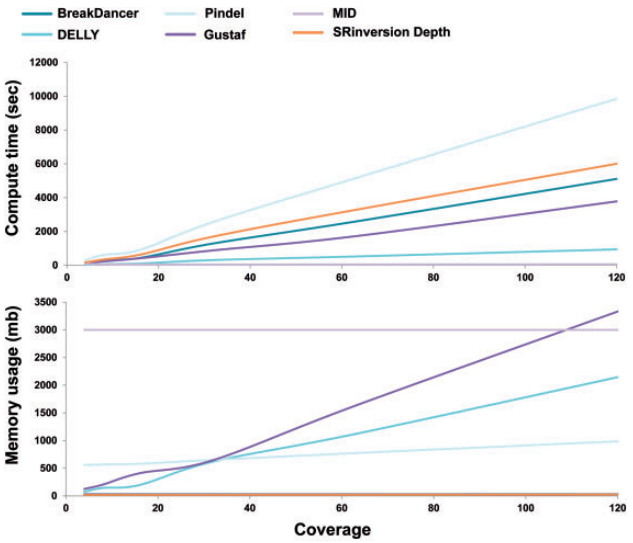

Fig. 3. Computing time and memory usage for simulated data. Computing time and memory usage of BreakDancer, DELLY, Pindel, SRinversion, Gustaf and MID were compared in different coverage of the simulated data. Here depth ≥5 cutoff was applied for SRinversion

relative to that of other methods. Pindel had the best performance among the five published methods for small inversions. Gustaf could detect inversions from 20 bp to 1 kb, but had high variation in performance among different categories. MID, as stated in the paper, could not detect inversions longer than read length, i.e. 100 bp. And for inversions smaller than 100 bp, MID showed acceptable specificity but rather low sensitivity (Table 1).

Different coverage depths were simulated in the data to test the effect of coverage on detection sensitivity and specificity (Fig. 2). In general, specificity tends to be more stable for various coverage depth. And except for DELLY and Gustaf, sensitivity of all methods converged to their respective highest levels when coverage reached 40× for heterozygous and 20× for homozygous inversions. SRinversion was tested with different cutoff for minimum number of supporting reads for one inversion to be output since depth plays important role for the performance of SRinversion. When average coverage is higher than 40×, higher depth cutoff (removing low coverage regions) increased the specificity without significantly affecting sensitivity. While for data with lower coverage overall, using higher depth as cutoff gained higher specificity but at the cost of sensitivity. These results indicate that a 5× to 10× threshold requirement would be recommended to balance specificity and sensitivity in inversion detection.

Both heterozygous and homozygous inversions were simulated for all datasets described above. Performance for detecting homozygous inversions was slightly better than for heterozygous ones for all methods and all inversion lengths. Sensitivity of detecting homozygous inversions equals that for heterozygous ones at almost half of the coverage level, indicating that the different performance between detecting heterozygous and homozygous inversions is largely an issue of the number of supporting reads, with and without reads for another allele.

4.2.2 Efficiency (Fig. 3)

Figure 3 shows the computing times and memory usage for the simulated datasets with different levels of coverage. Memory usage for SRinversion, BreakDancer and MID was little affected by the increase in coverage depth, different from that for Pindel, DELLY and Gustaf. Computing times of all the methods were approximately linear with coverage but with different slope. Among them, DELLY and MID were the fastest for all coverage levels, both of which input information for all reads from the onset, leading to reduced computing time at the cost of high memory usage when coverage was high. SRinversion was slower than BreakDancer, DELLY, Gustaf and MID, especially when coverage was high, which was probably because it processed all poorly mapped and unmapped reads to extract sequences that might harbor inversions. Pindel also scans all poorly mapped and unmapped reads and has the highest computing time demand with a high but stable memory usage.

4.3 Application to real data

4.3.1 Inversion detection on a CEU trio

SRinversion was also used on high-coverage sequencing data on chromosome 1 and chromosome 21 of a European parent-child trio (NA12878, NA12891 and NA12892) from the 1000 Genomes Project. In total, 262, 42 and 38 inversions were detected on NA12878, NA12891 and NA12892, respectively (Supplementary Table S2). The relatively smaller number of inversions detected in parents (NA12891 and NA12892) could probably be explained by the fact that a depth ≥10 cutoff was applied here to reduce false positives, which might have caused omission of certain regions that are not covered well in parents, as the coverage depth for the parents was only half of that for the daughter.

For comparison, BreakDancer, DELLY and Pindel were also used on the data from chromosome 21. Unlike for simulated data, the numbers of inversions detected by the three published methods were insignificant: 2, 2 and 4 inversions were detected by BreakDancer, DELLY and Pindel, respectively (Supplementary Table S3), all of which were also detected by SRinversion. What is more, none of the three methods detected inversions smaller than 100 bp. An inversion detected from NA12878 by BreakDancer that was reported to be 38 bp turned out to be longer than 500 bp upon further examination (Supplementary Table S3). Gustaf and MID were also tested on chromosome 21, however, both of them were not able to detect any inversions smaller than 1 kb.

4.4 Validating selected inversions by long-read data and PCR experiments

Selected inversions detected for NA12878 were compared to results from https://github.com/hall-lab/long-read-validation, which is

based on long reads data on the same individual sample. In general, 59.54% of the inversions detected by SRinversion were also reported from long read data. Specifically, 129 of the 192 inversions in chromosome 1 and 27 of the 70 inversions in chromosome 21 were detected by PacBio/Moleculo data (see Supplementary Table S2 for details). All the currently available methods for SV detection have negligible detection of the inversions on NA12878 and were not compared here.

For primer design and PCR, definitive sequence information was required and the validation process was limited to the regions not located within or near repetitive or repeat sequences (Supplementary Fig. S3). For the filtering criteria please refer to supplementary method section 2. Five inversions were selected for PCR validation. Four of the five inversion regions were validated by sequencing of the PCR amplicons, while the fifth one failed PCR amplification due to a repeat region nearby that was not detected in the filtering process.

4.5 Effect of read length on inversion detection

To test the effect of read length on inversion detection, three different types of data from a single sample were used. They are (1) short reads: whole-genome paired-end sequencing data with read length at 150 bp, depth of 35× from Illumina HiSeq X10, (2) long reads: whole-genome paired-end sequencing data with read length at 300 bp, depth of 7× from Illumina MiSeq and (3) connected long reads: connected long reads from dataset (2) (Supplementary Fig. S6a). Both SRinversion and three other methods, i.e. BreakDancer, DELLY and Pindels, were tested on the three datasets (Supplementary Fig. S5a). For dataset (3), since the two paired reads were connected to convert the paired reads into a single read, the data can no longer be processed by the three public tools that require paired-end data and insert size information as input. As shown in Supplementary Figure S5a, there are relatively more inversions detected using short reads (dataset 1) than using long reads (dataset 2) for all four programs, which might be largely due to the differences in coverage depth of the two datasets. When long reads were connected to make the read length close to 590 bp on average, SRinversion detected more inversions from the connected long reads, even with a much lower coverage depth (∼9×) than from the short reads (∼35×), indicating a significant improvement in inversion detection when longer reads are available (Supplementary Fig. S5b).

5 Discussion

Among the different types of genetic variants, such as SNVs, indels (insertions and deletions shorter than 50 bp), SVs (insertions, deletions, inversions, etc. longer than 50 bp) and complex regions (such as those with duplications, repeats and repetitive sequences), inversions shorter than 50 bp are probably the most overlooked variant type. And summary of public database of SVs does suggest inadequacy of methods in detecting inversions shorter than 50 bp (Supplementary Fig. S1).

SRinversion tries to fill the gap by focusing exclusively on detecting short inversions. By splitting sequencing reads, algorithm for scenario 1 focuses on inversions from 20 bp to 500 bp, while by inverting the sequences in the middle of the reads, algorithm for scenario 2 aims at detecting inversions smaller than 20 bp, with a resolution up to 4 bp. An important characteristic of SRinversion is that it makes use of reads with low mapping quality and unmapped reads, which are overlooked by most available methods in SV detection using NGS data. By making use of the poorly mapped reads, more inversions, especially those smaller than 50 bp that are likely to be located in the middle of sequencing reads and interrupt alignment of the reads can still be detected.

Currently available tools were developed primarily to deal with paired-end reads, which have been the dominant data type for most sequencing platforms so far. However, with the rapid development of new sequencing platforms and especially the advances of the third generation sequencing technologies, new methods like SRinversion that can deal with newer types of sequencing data as well as existing paired-end reads are useful tools in our arsenal to detect structural variants. MID also tried to generate segments from single reads with no pair-end information, which is similar to the process for scenario 2 in SRinversion. But only focusing on one of the aspects limited the length of inversions the method can detect to 30–100 bp. And applying limitations such as requiring anchors for both ends of a read and reads covering entirely the targeted inversions significantly reduced the power of the method.

SRinversion itself cannot detect inversions at nucleotide resolution, but with implementation of external assembly methods, i.e. velvet and BLAST to map the inferred inversion contigs back to reference genome, the resolution can reach 1–2 bp.

Comparison of WGS short reads with read length 150 bp and long reads with read length longer than 500 bp emphasized the potential power increase by longer read length for SRinversion. Notably more inversions were detected using connected reads with average read length longer than 500 bp, even when the coverage depth (9×) is much lower than that by the short reads (35×), suggesting that read length might be a determining factor over sequencing depth for inversion detection. Thus increasing read length whenever possible, such as by connecting two reads of a read pair when the insert size is smaller than the sum of the read length, might be beneficial for inversion detection using SRinversion. It is noted that this is based on a small sample and further studies are needed to explore the data types most suited for SV detection.

Inversions, especially smaller ones, often occur around complex regions in human genome such as short repeat regions, as found in the CEU trio in this study. As shown in this analysis, of the 262 inversions detected in chromosome 1 and chromosome 21 from NA12878, only five were located in regions without complex sequences (Supplementary Fig. S3). Thus more accurate detection of inversions might hinge upon a better understanding of these complex regions in human genome, which are highly variable among individuals and populations.

Checking the detailed location of inversions detected from the sample with three different types of data using SRinversion, we demonstrated that around 96% of the detected inversions are located around repeat regions (Supplementary Fig. S5c), which is consistent with findings on samples from the 1000 Genomes Project. This suggests a strong association between inversions and repetitive sequences in human genome, which might be resulted from the intrinsic characteristics of these regions in which inversions are more likely to occur. It is also possible that repetitive regions increase sequencing and mapping errors that may produce false positive signals.

To resolve these issues, inversion breakpoints reported by long-read sequencing data from PacBio and Moleculo platforms were used for comparison. And 59.54% of the inversions called by SRinversion were validated by these long-read data. The 40% inversions not also detected by LUMPY using PacBio/Moleculo data could be either false positives from SRinversion or false negatives from LUMPY results. SRinversion focuses on detecting inversions

smaller than 100bp, while LUMPY tries to integrate signals from different datasets and calls a variety of structural variants of all ranges (Supplementary Fig. S7). And it is apparent that more data and studies are needed to further increase the sensitivity and specificities of inversion detection.

Due to the painstaking process of recovering poorly mapped or unmapped reads, inverse partial sequences for further alignment in order to increase the sensitivity of inversion detection, SRinversion has not achieved the efficiency needed to process a large amount of whole genome sequencing data. It seems that limitations in read length, coverage depth and poor understanding of the complex regions in human genome also prevented a more powerful performance from SRinversion. At this moment, application of SRinversion on targeted sequencing data seems to be a more feasible practice. However, the algorithm provided here and the initial findings in application of SRinversion have laid the foundation for further development in this area. Complete understanding of the correlations between genotypes and human diseases will rely on our capacity to detect all variants accurately in human genome, including inversions. Thus SRinversion is a timely development in this effort.

Funding

This work has been supported by the Research Grant Council of Hong Kong (17125114 and HKU783813M). We thank University of Hong Kong Postgraduate Scholarships, The Edward & Yolanda Wong Fund and Hotung Fund for supporting postgraduate students who participated in this work.

Conflict of Interest: none declared.

References

Antonacci,F. *et al.* (2009) Characterization of six human disease-associated inversion polymorphisms. *Hum. Mol. Genet.*, **18**, 2555–2566.

Baker,M. (2012) Structural variation: the genome's hidden architecture. *Nat. Methods*, **9**, 133–137.

Bartenhagen,C. and Dugas,M. (2013) RSVSim: an R/Bioconductor package for the simulation of structural variations. *Bioinformatics*, **29**, 1679–1681.

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

Bondeson,M.L. *et al.* (1995) Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. *Hum. Mol. Genet.*, **4**, 615–621.

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

Church,D.M. *et al.* (2010) Public data archives for genomic structural variation. *Nat. Genet.*, **42**, 813–814.

Conrad,D.F. *et al.* (2008) A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.*, **38**, 75–81.

Feuk,L. *et al.* (2005) Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. *PLoS Genet.*, **1**, e56.

Feuk,L. *et al.* (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 85–97.

Giglio,S. *et al.* (2001) Olfactory receptor-gene clusters, genomic-inversion polymorphisms, and common chromosome rearrangements. *Am. J. Hum. Genet.*, **68**, 874–883.

Giglio,S. *et al.* (2002) Heterozygous submicroscopic inversions involving olfactory receptor-gene clusters mediate the recurrent t(4;8)(p16;p23) translocation. *Am. J. Hum. Genet.*, **71**, 276–285.

Gimelli,G. *et al.* (2003) Genomic inversions of human chromosome 15q11–q13 in mothers of Angelman syndrome patients with class II (BP2/3) deletions. *Hum. Mol. Genet.*, **12**, 849–858.

He,F. *et al.* (2016) Identifying micro-inversions using high-throughput sequencing reads. *BMC Genomics*, **17**, 4.

Huang,W. *et al.* (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.

Hurles,M.E. *et al.* (2008) The functional impact of structural variation in humans. *Trends Genet.*, **24**, 238–245.

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

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.

Koolen,D.A. *et al.* (2006) A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat. Genet.*, **38**, 999–1001.

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

Kurotaki,N. *et al.* (2003) Fifty microdeletions among 112 cases of Sotos syndrome: low copy repeats possibly mediate the common deletion. *Hum. Mutat.*, **22**, 378–387.

Lakich,D. *et al.* (1993) Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat. Genet.*, **5**, 236–241.

Layer,R.M. *et al.* (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, **15**, R84.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **15**, 1754–1760.

MacDonald,J.R. *et al.* (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.*, **42**, D986–D992.

Osborne,L.R. *et al.* (2001) A 1.5 million-base pair inversion polymorphism in families with Williams–Beuren syndrome. *Nat. Genet.*, **29**, 321–325.

Rausch,T. *et al.* (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **15**, i333–i339.

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

Stefansson,H. *et al.* (2005) A common inversion under selection in Europeans. *Nat. Genet.*, **37**, 129–137.

Steinberg,K.M. *et al.* (2012) Structural diversity and African origin of the 17q21.31 inversion polymorphism. *Nat. Genet.*, **44**, 872–880.

The 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

Trappe,K. *et al.* (2014) Gustaf: Detecting and correctly classifying SVs in the NGS twilight zone. *Bioinformatics*, **30**, 3484–3490.

Tuzun,E. *et al.* (2005) Fine-scale structural variation of the human genome. *Nat. Genet.*, **37**, 727–732.

Visser,R. *et al.* (2005) Identification of a 3.0-kb major recombination hotspot in patients with sotos syndrome who carry a common 1.9-Mb microdeletion. *Am. J. Hum. Genet.*, **76**, 52–67.

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

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*, **1**, 2865–2871.

Zhang,J. *et al.* (2014) PEAR: a fast and accurate Illumina paired-end read merger. *Bioinformatics*, **30**, 614–620.

Zody,M.C. *et al.* (2008) Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat. Genet.*, **40**, 1076–1083.