LinkedSV: Detection of mosaic structural variants from linked-read exome and genome sequencing data

Li Fang[1], Charlly Kao[2], Michael V Gonzalez[2], Fernanda A Mafra[2], Renata Pellegrino da Silva[2], Mingyao Li[3], Sören Wenzel[4], Katharina Wimmer[4], Hakon Hakonarson[2,5], Kai Wang[1,6*]

[1] Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, PA 19104, USA

[2] Center for Applied Genomics, Children's Hospital of Philadelphia, PA 19104, USA

[3] Department of Biostatistics, University of Pennsylvania, Philadelphia, PA 19104, USA

[4] Section for Human Genetics, Department of Medical Genetics, Molecular and Clinical Pharmacology, Medical University Innsbruck, Innsbruck, Austria

[5] Department of Pediatrics, University of Pennsylvania, Philadelphia, PA 19104, USA

[6] Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

* Email: wangk@email.chop.edu

Abstract

Reliable detection of structural variants (SVs) from short-read sequencing remains challenging, partly due to the presence of repetitive DNA elements that are longer than typical short reads (~100-150bp). Linked-read sequencing provides long-range information from short-read sequencing data by linking reads originating from the same DNA molecule based on barcoding, and thus has the potential to improve the sensitivity of SV detection and accuracy of breakpoint identification for certain classes of SVs. We present LinkedSV (https://github.com/WGLab/LinkedSV), a novel SV detection algorithm which combines two types of statistical evidence. Simulation studies demonstrated that LinkedSV outperformed multiple existing computational tools, and it worked particularly well on exome sequencing data and on SVs with low variant allele frequencies due to somatic mosaicism. We further demonstrated two clinical cases where LinkedSV successfully identified disease causal SVs from linked-read exome sequencing data, yet other computational methods failed, suggesting that a fraction of negative cases from clinical exome sequencing may be due to hidden SVs undetectable by traditional methods. Finally, comparative analysis of a human genome deeply sequenced by PacBio long-read sequencing (103X) and linked-read sequencing (37X) demonstrated unique advantages of linked-read data to identify large SVs that are missed from high-coverage long-read data. In summary, our results support the use of linked-read sequencing to detect hidden SVs

missed by conventional short-read sequencing approaches, and may help resolve negative cases from clinical genome or exome sequencing.

Introduction

Genomic structural variants (SVs) have been implicated in a variety of phenotypic diversity and human diseases[1]. Several approaches such as split-reads [2,3], discordant read-pairs [3,4], and assembly-based methods [5,6] have been developed for SV discovery from short reads. However, reliable detection of SVs from these approaches remains challenging. The split-reads and discordant read-pairs approaches require the breakpoint-spanning reads/read-pairs being sequenced and confidently mapped. Genomic rearrangements are often mediated by repeats and thus breakpoint junctions of SVs are highly likely to reside in repetitive regions [7-9]. Therefore, the breakpoint-spanning reads/read-pairs may be multi-mapped and have low mapping qualities. It is also difficult to perform assembly at repeat regions. Long-read sequencing such as SMRT sequencing and Nanopore sequencing are better for SV detection [10,11], but their application is limited by the higher cost and per-base error rate.

Linked-read sequencing technology developed by 10X Genomics combines the throughput and accuracy of short-read sequencing with the long-range information. In this approach, nanogram amounts of high-molecular weight (HMW) DNA molecules are dispersed into more than 1 million

droplet partitions with different barcodes by a microfluidic system [12]. Thus, only a small number of HMW DNA molecules (~10) are loaded per partition [13]. The HMW DNA molecules can be up to several hundred kilobases in size and have a length-weighted mean DNA molecule length of about 50 kb. Within an individual droplet partition, HMW DNA molecules are primed and amplified by primers with a partition-specific barcode. The barcoded DNA molecules are released from the droplets and sequenced by standard Illumina paired-end sequencing [12]. The sequenced short reads derived from the same HMW DNA molecule can be linked together, providing long-range information for mapping, phasing and SV calling. In addition, linked-read whole exome sequencing (WES) has also been developed [12], which provides a more cost-effective way for clinical genetic testing as regular WES may require help from additional techniques to resolve certain regions.

In linked-read sequencing data, barcode similarities between any two nearby genome locations are very high, because the reads tend to originate from the same sets of HMW DNA molecules. In contrast, barcode similarities between any two distant genome locations are very low, because the reads of the two genome locations originate from two different sets of HMW DNA molecules and it is highly unlikely that two different sets of HMW DNA molecules share multiple barcodes. Thus, the presence of multiple shared barcodes between two distant locations indicates that the two distant locations are close to each other in the alternative genome [14]. A few pipelines and software tools have adopted this principle to call SVs from linked-read sequencing data, such as Longranger

[12], GROC-SVs [14], NAIBR [15]. Longranger is the official pipeline developed by 10X Genomics. Longranger bins the genome into 10 kb windows and finds the barcodes of high mapping quality reads within each window. A binomial test is used to find all pairs of regions that are distant and share more barcodes than what would be expected by chance. A probabilistic approach is used to clean up this initial candidate list [12]. GROC-SVs uses a similar method to find candidate SV loci but performed assembly to identify precise breakpoint locations. GROC-SVS also provides functionality to interpret complex SVs [14]. NAIBR detects structural variants using a probabilistic model that incorporates signals from both linked-reads and paired-end reads and into a unified model [15].

However, SV detection from linked-read datasets is still in the early stage. The available SV callers face challenges if we want to detect: i) SVs from targeted region sequencing (e.g. Whole-exome sequencing); ii) somatic SVs in cancer or somatic mosaic SVs that have low allele frequencies; iii) SVs of which the exact breakpoints have no coverage or are located in repeat regions. In this study, we introduce LinkedSV, a novel open source SV caller for linked-read sequencing, which aims to address all the above challenges. LinkedSV detects candidate breakpoints using two types of evidence and quantifies the evidence using a novel probabilistic model. We evaluated the performance of LinkedSV on both whole-genome and whole-exome sequencing data sets. In each case, LinkedSV outperformed other existing tools, including Longranger, GROC-SVs and NAIBR.

Results

Illustration of two types of evidence near SV breakpoints

Two types of evidence may be introduced while a genomic rearrangement happens: 1) reads from one HMW DNA molecule which spans the breakpoint being mapped to two genomic locations and 2) reads from two distant genome locations that get mapped to adjacent positions. Both types of evidence can be used for SV detection.

First, we describe the signals of type 1 evidence. After reads mapping, the original HMW DNA molecules can be computationally reconstructed from the sequenced short reads using their barcodes and mapping positions. In order to distinguish them from the physical DNA molecules, we use “fragments” to refer to the computationally reconstructed DNA molecules. A fragment has a left-most mapping position, which we call L-endpoint, and a right-most mapping position, which we call R-endpoint. As a result of genomic rearrangement, reads from one breakpoint-spanning HMW DNA molecule would be mapped to two different genome loci on the reference genome. This split-molecule event has two consequences: 1) observing two separate fragments sharing the same barcode; 2) each of the two fragment has one endpoint close to the true breakpoints. Therefore, in a typical linked-read WGS data set, multiple split-molecule events could be captured and we would usually observe: multiple share barcodes between two distant genome loci and multiple fragment endpoints near the breakpoints.

To illustrate this, Figure 1a shows the split-molecule events of a deletion, where breakpoints 1 and 2 are marked by red arrows. Multiple fragment endpoints are enriched near the two breakpoints of a large deletion, which is typically longer than 5 kb for whole-genome sequencing (WGS) data sets. Figure 1b shows the patterns of "enriched fragment endpoints" that are introduced by different types of SVs. As an example, Figure 1c shows the number of fragment endpoints in a 5-kb sliding window near two deletion breakpoints, based on a 35X coverage linked-read WGS data generated from the NA12878 genome (genome of a female individual extensively sequenced by multiple platforms). At the breakpoints, the number of fragment endpoints in the 5-kb sliding window is more than 100 and is five times more than normal regions, forming "peaks" in the figure.

Since the fragments can be paired according to their barcodes, we can also observe fragment endpoints of this deletion in a two-dimensional view. As shown in Figure 1d, each dot indicates two endpoints from a pair of fragments which share the same barcode. The x-value of the dot is the position of the first fragment's R-endpoint and the y-value of the dot is the position of the second fragment's L-endpoint. The bottom panel and right panel in Figure 1d shows number of dots that are projected to the x-axis and y-axis. Similar with the one-dimensional plot (Figure 1c), a peak is formed near each breakpoint, which is marked by the red arrow. The background noise of the two-dimensional plot is cleaner than the one-dimensional plot since the fragments that do not share barcodes are excluded. Therefore, the two-dimensional plot is more useful when the variant allele frequency (VAF) is very low and there are only a few supporting fragments.

Next, we describe the signals of type 2 evidence. The barcodes between two nearby genome locations is highly similar because the two locations are spanned by almost the same set of input HMW DNA molecules. However, due to the genome rearrangement, the reads mapped to the left side and right side of a breakpoint may originate from different locations of the alternative genome and thus have different barcodes (Figure 1e). Dropped barcode similarity between two nearby loci therefore indicates a SV breakpoint. LinkedSV detects this type of evidence by a twin-window method, which uses two adjacent sliding windows to scan the genome and find regions where the barcode similarity between the two nearby window regions are significantly decreased. Figure 1f illustrates an inversion breakpoint detected by LinkedSV from the NA12878 genome. The change of barcode similarity was plotted and a peak was formed at the breakpoint. After searching for the two types of evidence, LinkedSV combines the candidate SV regions and quantifies the evidence using a novel probabilistic model. The breakpoints are further refined using short-read information, including discordant read pairs and split-reads.

Performance evaluation on simulated WGS data

To assess LinkedSV's performance, we simulated a 35X linked-read WGS data set with 1175 SVs inserted using LRSIM[13] (see Methods for details). The inserted SVs includes 351 deletions, 386 duplications, 353 inversions, and 85 translocations. The breakpoints of the simulated SVs were

designed to be located in repeat regions, since we found that LinkedSV and other available SV callers performed very well when the breakpoints were located in non-repeat regions, and thus we set to test the performances of all the SV callers under more challenging situations. This makes sense because SV breakpoints are more likely to be in repeat regions [7-9], and because these situations represent those that are impractical to be addressed by conventional short-read sequencing approaches.

The simulated reads were aligned to the reference genome using the Longranger [12] package provided by 10X Genomics. The Longranger pipeline internally uses the Lariat aligner [16], which was designed for the alignment of linked reads. SV calling was performed using LinkedSV as well as three other available SV callers designed for linked-read sequencing: Longranger, GROC-SVs [14] and NAIBR [15]. We used recalls, precisions and F1 scores to evaluate the four SV callers on this data set. As shown in Figure 2a, LinkedSV has the highest recall and F1 score among all methods. NABIR has a good recall (0.81), but relatively low precision (0.42). GROC-SVs has a good precision but its recall is lower than LinkedSV, so we further analyzed the false negative calls of GROC-SVs to understand this underlying reasons. A major portion of the false negative calls by GROC-SVs represents duplications that are smaller than twice the fragment length. For large duplications, the reads of the alternative allele are separated by a large gap so that we can observe two sets of fragments with the same set of barcodes, which indicate a SV (Supplemental Figure 5a). If the duplication is not large enough, the reads will be probably clustered into one fragment

(Supplemental Figure 5b). Even in this case, we can observe enriched fragment endpoints near the duplication breakpoints in LinkedSV. As an example, Figure 2b shows the endpoint signal of a missed duplication call by GROC-SVs. The supporting fragments of this duplication is shown in Figure 2c. We also evaluated the breakpoint precision of LinkedSV. Most of breakpoints predicted by fragment endpoints are within 20 bp (Figure 2d) and refined breakpoints using discordant read-pairs and split-reads have base-pair resolution (Figure 2e).

Performance evaluation on WGS data with somatic SVs with low variant allele frequencies

Somatic SVs are commonly found in cancer genomes [17-19]. However, due to the high heterogeneity of genomic alteration in cancer genomes, somatic SVs often have low allele frequencies (as opposed to ~50% in a germline genome) and thus are more difficult to detect by SV callers designed for germline SVs. We simulated two WGS data sets with VAF of 10% and 20%, respectively. Recalls, precisions and F1 values of the four linked-read SV callers were evaluated on both data sets (Figure 3a, Figure 3b). When the VAF was 20%, the recall of LinkedSV (0.803) was much higher than that of Longranger (0.238), GROC-SVs (0.402) and NAIBR (0.679). The precisions of the four SV callers ranged from 0.87-0.95. The F1 score of LinkedSV (0.855) was also the highest among all the SV callers. When the VAF was 10%, LinkedSV still had a recall of 0.761, which was 72% higher than the second best SV caller NAIBR. Longranger and GROC-SVs

almost completely failed to detect the SVs. These observations confirmed that other SV callers were mainly designed for germline genomes and had substantial difficulty in detecting SVs with somatic mosaicism. However, due to the combination of barcode overlapping and enriched fragment endpoints in our statistical model (see Methods for details), LinkedSV was able to achieve a good performance even when VAF was very low. We manually checked the barcode overlapping evidence of some SV calls using the Loupe software developed by 10X Genomics. Figure 3c shows an inversion that was missed by Longranger, and NAIBR but detected by LinkedSV (at VAF of 10%). Although the variant frequency is low, the overlapped barcodes between the two inversion breakpoints can be clearly visualized (in the black circles) in the figure. Figure 3d shows the supporting fragments of the inversion detected by LinkedSV. Each horizontal line represent two fragments that share the same barcode and support the SV. These results suggest that the manufacturer-provided software tool has limitations for SV detection, despite its strong functionality in visualization.

Performance evaluation on simulated whole-exome sequencing data

Compared with WGS, whole-exome sequencing (WES) is currently widely used in clinical settings to identify disease causal variants on patients with suspected genetic diseases, partly due to the slightly lower cost of WES. Since the WES only covers a small portion of regions in the whole

genome, it is far more challenging to detect SVs from WES data, especially when the SV breakpoints are not in the capture regions. However, by combining linked-read sequencing with WES capture platforms, it is possible to alleviate this problem, and significantly improve the sensitivity of SV detection using WES.

To evaluate SV detection on linked-read WES data, we simulated a 40X coverage linked-read WES data set with 1160 heterozygous SVs (see Methods for details). 44.3% of the breakpoints are not in exon regions. SV calling was performed using LinkedSV, Longranger, GROC-SVs and NAIBR. As shown in Figure 4a, LinkedSV has the highest recall (0.79) and highest F1 score (0.86). NAIBR has the highest precision (0.97) but its recall is lower (0.69). GROC-SVs has a good precision (0.90) but the recall (0.61) is modest.

We analyzed false negative calls of the second best SV caller NAIBR. NAIBR tends to miss some SV events that have shared barcodes but lack short-read support. For example, Figure 4b shows a deletion between chr1:172545561-173504265. Both breakpoints are located outside of capture regions. Breakpoint 1 (chr1:172545561) is 768 bp away from the nearest capture region and breakpoint 2 (chr1: 173504265) is 392 bp away from the nearest capture region. Unfortunately, no discordant read pairs that support the deletion can be found. However, shared barcodes between the two breakpoints were clearly indicated by the loupe software (Figure 4b). In addition, LinkedSV also detected 28 pairs of fragments that share the same barcodes and support the SV. These fragments were plotted in Figure 4c. Although no short-read support was found, the SV type

could be determined using the pattern of "enriched fragment endpoints" shown in Figure 1b. In this SV event, R-endpoints were highly enriched for the first set of fragments and L-endpoints were highly enriched for second set of fragments. Thus, the SV type was predicted as "deletion".

Detection of *F8* inversion from clinical WES data

We also tested the performance of LinkedSV on several clinical samples with linked-read WES data. First, we applied LinkedSV on a WES sample of a male individual with Hemophilia A. Previous experiments had shown that the patient had type I inversion of the *F8* gene, where the two breakpoints resided in intronic/intergenic regions, thus the inversion and its breakpoints cannot be inferred from conventional WES. The *F8* gene is located in Xq28. The intron 22 of *F8* gene contains a GC-rich sequence (named int22h-1) that is duplicated at two positions towards the Xq-telomere (int22h-2 and int22h-3). Int22h-2 has the same direction with int22h-1 while int22h-3 has the inverted direction. The type I inversion is induced by the recombination between int22h-1 and int22h-3 [20,21] (Figure 5a). BLAST alignment of int22h-1 and int22h-3 showed that the two sequences had 99.88% identity. Since the breakpoints were located in two segmental duplications with nearly identical sequences, the inversion is undetectable by conventional short-read sequencing. Two popular short-read SV callers Delly [3] and Lumpy [22] failed to detect the inversion

from the linked-read WES data (results were shown in Supplemental Table 1 and Supplemental Table 2).

Longranger, GROC-SVs, NAIBR and LinkedSV were also used to detect SVs from this sample. None of these methods detected this inversion (results were shown in Supplemental Table 3, Supplemental Table 4, and Supplemental Table 5), although the overlapping barcodes can be visualized using the Loupe software (Figure 5b). However, LinkedSV successfully detected this inversion by combining two types of evidence. As described above, barcode similarity between two nearby regions are very high but drops suddenly at the breakpoints. Figure 5c shows the suddenly drop of barcode similarity at the two breakpoints. Each dot in the figure represents the reciprocal of the barcode similarity between its left 40 kb window and right 40 kb window thus the Y value of the dots are reversely related to the barcode similarity and positively related to the probability of being a breakpoint. The barcode similarities are lowest at the two breakpoints and thus form two peaks in the figure (marked by red arrow). In addition, LinkedSV also identified the supporting fragments of the SV using type 1 evidence (Figure 5d). The predicted breakpoint positions are consistent with the genomic positions of int22h-1 and int22h-3.

Detection of mosaic *NF1* deletion from clinical WES data

Another linked-read WES sample was from an individual who was clinically diagnosed with Neurofibromatosis type 1. Neurofibromatosis type 1 is caused by mutations in the *NF1* gene on chromosome 17q11.2, which encodes neurofibromin, a GTPase activating protein that has a role in the regulation of RAS signaling. Since standard genetic testing techniques including cDNA sequencing and multiplex ligation-dependent probe amplification (MLPA) revealed no constitutional or mosaic pathogenic mutation in this patient, we hypothesized that this patient may carry an SV affecting the *NF1* gene that escapes the detection by the applied standard techniques. To evaluate LinkedSV, we utilized the 10X Genomics Chromium platform to generate linked-read WES data to confirm and resolve the mutation. SV detection was conducted using four SV callers: Longranger, GROC-SVs, NAIBR, and LinkedSV. Longranger detected overlapped barcodes between exon 54 of the *NF1* gene and intron 3 of *RAB11FIP4*. However, the SV type was unknown and no supporting read pairs or split-reads were found. GROC-SVs and NAIBR failed to detect this SV (Supplemental Table 6). As shown in Figure 6a, LinkedSV detected 16 fragment pairs that may support a deletion spanning the region of chr17:29684175-29822527. In addition, a discordant read pair spanning the two breakpoints were found (Figure 6b), which gave further evidence supporting the deletion. The breakpoints were estimated from this discordant read pair and thus the resolution is a few hundred base pairs. In Figure 6, each colored line represent a “reconstructed” fragment, and ~13% of the fragments belong to the variant allele, indicating the somatic mosaicism

of this deletion. The right breakpoint was within an AluJr sequence masked by repeat masker, which may explain why the deletion was difficult to be detected by conventional methods.

In comparison, the clinical lab used massive parallel sequencing (TruSightCancer panel on a MiSeq platform (Illumina)) and successfully revealed in exon 54 a transition of *NF1* sequences into a non-*NF1* derived sequence. This sequence transition at *NF1* position c.7886_7887 was present in 8% of the reads covering this site in germline DNA of the patient. Analysis of the reads displaying the aberrant sequence in exon 54 showed that the non-*NF1* derived sequence was part of an Alu element that matched best a sequence in intron 3 of the *RAB11FIP4* gene located 138 kb downstream of *NF1* exon 54. These results suggested a low-level (~8%) mosaicism of a deletion encompassing the region intervening between *NF1*:c.7886 and *RAB11FIP4*:c.337-22216, so that the true deletion spans chr17:29684367-29822453, which is very similar to our estimated breakpoints from LinkedSV above. In summary, our analysis on two clinical samples with *F8* inversion and *NF1* deletion demonstrated the unique advantage of linked-read sequencing in confirming and resolving structural variants in repetitive regions and challenging situations.

Detection of SVs from a deeply sequenced human genome by long-read sequencing

We previously reported the *de novo* genome assembly of a Chinese individual (HX1) [23]. This genome was sequenced deeply at 103X coverage using PacBio long-read sequencing. Recently,

the developers of SMRT-SV [10,24] reported the SV calls of HX1 detected from the PacBio data. Recently, we have also generated a 37X linked-read WGS dataset on HX1. Therefore, in the current study, we detected SVs from the linked-read data using LinkedSV and compared the SV calls detected by LinkedSV and SMRT-SV. The SMRT-SV call set has 17 large deletions (≥10kb), all of which were detected by LinkedSV. In addition, LinkedSV detected another 46 large deletions, which were missed by SMRT-SV. To validate these deletion calls, we mapped the PacBio reads of HX1 to GRCh38 reference genome using minimap2[25], and manually examined all the SV affected regions in both PacBio data and linked-read data, using the Integrative Genomics Viewer (IGV) [26] and the Loupe software tool. We classify a deletion as a true deletion if there are decreased read depth in the deletion region and clear boundaries at the breakpoints. After the manual inspection, we found that among the 46 deletions that are only detected by LinkedSV, 34 of them have clear evidence of deletion in the WGS data, 10 of them are complex SV events that need to be fully resolved, and 2 of them are false positive events. Figure 7a-c showed an example of a deletion that were detected by LinkedSV but missed by SMRT-SV. This is a 45 kb heterozygous deletion located in chr2:110395971-110441346. A deletion pattern was clearly indicated by the Loupe software tool (Figure 7a). However, neither decreased read depth nor clipped reads can be observed in the alignments of raw PacBio reads, as shown in IGV (Figure 7b). The deletion pattern became clear when we examined the alignments of error-corrected reads (Figure 7c). This

indicates that the relatively high sequencing errors (12-15%) in the long reads may affect read mapping and subsequent structural variant calling.

We also compared the duplication calls of LinkedSV and SMRT-SV and manually examined discordant SV calls. LinkedSV reported 6 large duplications (≥ 10kb), 5 of which were not reported by SMRT-SV. Figure 7d-f showed the evidence of a 61 kb duplication call (chr19:27338390-27399298), which was only reported by LinkedSV. A two-fold increase of read depth could be observed in the duplication region (Figure 7d), and the breakpoints were also clearly indicated in the alignments of PacBio long reads, as shown in IGV (Figure 7e,f). SMRT-SV reported 194 large duplications (≥ 10 kb). Unexpectedly, 193 of the duplication calls were not detected by LinkedSV. In addition, none of these duplications could be detected by Sniffles[11], another widely used long-read SV caller. After comparing with the segmental duplication database [27], we found that 182 of the 193 duplication calls (94.3%) were located in large segmental duplication regions. Both long reads and linked reads could not be reliably mapped in these regions. As an example, we plotted the read depth distribution in the region around a 25 kb duplication call of SMRT-SV. Neither long reads (Supplemental Figure 7A) nor linked reads (Supplemental Figure 7B) had mapping quality >20 in the duplication region. Therefore, SV detection in the super-large segmental duplication regions is still very challenging. In summary, our comparative analysis demonstrated unique advantages of linked-read WGS in resolving SVs that may be failed by even long-read sequencing platform with very deep coverage.

Methods

Breakpoint detection from type 1 evidence

First, LinkedSV reconstructs the original long DNA fragments from the reads using mapping positions and barcode information. All mapped reads are partitioned according to the barcode and sorted by mapping position. We define gap distance as the distance between two nearest reads with the same barcode. Two nearby reads are considered from the same long DNA fragment if they have the same barcode and their gap distance is less than a certain distance *G*. *G* is determined using two steps. First, we use *G* = 50 kb (the same as Zheng et.al [12]) to group the reads into fragments. This value is suitable for detection of large SVs. However, it may be too large for detection of SVs that are smaller than 50 kb. Therefore, we calculate the empirical distribution of intra-fragment gap distance, which is the distance of two nearby reads that are grouped in one fragment. We assign *G* as the 99th percentile of the empirical distribution of intra-fragment gap distance. *G* is usually between 8~15kb, depending on the data set. Fragments with a gap distance larger than *G* will be separated to two fragments.

In non-SV regions, all the reads from the same HMW DNA molecule would be reconstructed into a single DNA fragment. The reads from the breakpoint-spanning HMW DNA molecule will be mapped to two different positions in the genome. As illustrated in the Result section, this split-

molecule event has two consequences: 1) observing two fragments sharing the same barcode; 2) each of the two fragment has one endpoint close to the breakpoints. Therefore, we could observe enriched fragment endpoints near the breakpoints, in both one-dimensional view (Figure 1c) and two-dimensional view (Figure 1d). The type of the endpoints (L-endpoint or R-endpoint) that enriched near the breakpoints depends on the type of SV (Figure 1b). The two-dimensional view has less background noise because the fragments that do not share barcodes and thus do not support the SVs are excluded. Therefore, we detect the enriched endpoints in the two-dimensional view.

We now describe how we use detect the type 1 evidence of deletion calls, but the method can be applied to other types of SVs. We define *fragment pair* to be two fragments sharing the same barcode. Let b_1, b_2 be the positions of the two breakpoint candidates (assuming $b_1 < b_2$). Let n be the number of fragment pairs that may support the SV between b_1 and b_2. Let F_{i1}, F_{i2} denote the ith fragment pair that support the SV. Let $B(F)$ denote the barcode of fragment F. Therefore, we have:

$$B(F_{i1}) = B(F_{i2}), i = 1, 2, 3, \ldots, n \quad (1)$$

Let $L(F)$ denote the L-endpoint position (i.e., left-most position) of fragment F, $R(F)$ denote the R-endpoint position (i.e., right-most position) of fragment F. Since this is a deletion and $b_1 < b_2$, $R(F_{i1})$ is the position on F_{i1} that is closest to b_1 and $L(F_{i2})$ is the position on F_{i2} that is closest to b_2 (Supplemental Figure 6a). The distance between the fragment endpoint and its corresponding

breakpoint should be within gap distance distribution (explained in Supplemental Figure 4). Therefore, we have:

$$b_1 - G \leq R(F_{i1}) \leq b_1,\ b_2 \leq L(F_{i2}) \leq b_2 + G \quad (2)$$

As described above, G is the 99th percentile of the empirical distribution of intra-fragment gap distance.

If we regard ($R(F_{i1})$, $L(F_{i2})$) as a point in a two-dimensional plane, according to equation (2), (($R(F_{i1})$, $L(F_{i2})$) is restricted in a square region with a side length of G and the point (b_1, b_2) being a vertex (Supplemental Figure 6b).

We used a graph-based method to fast group the points into clusters and find square regions where the numbers of points were more than expected. First, every possible pair of endpoints ($R(F_1)$, $L(F_2)$) meeting $B(F_1) = B(F_2)$ formed a point in the two-dimensional plane. Each point indicated a pair of fragments that share the same barcode. For example, if 10 fragments share the same barcode, C_{10}^2 pairs of endpoints will be generated. A point/pair of endpoints may or may not support an SV because there are two possible reasons for observing two fragments sharing the same barcode: 1) the two fragments originated from two different HMW DNA molecules but were dispersed into the same droplet partition and received the same barcode; 2) the two fragments originated from the same HMW DNA molecule but the reads were reconstructed into two fragments due to an SV. The points are sparsely distributed in the two-dimensional plane and it is highly unlikely to observe

multiple points in a specific region. Next, a k-d tree (k = 2) was constructed, of which each node stores the (X, Y) coordinates of one point. K-d tree is binary tree that enable fast query of nearby nodes. Therefore, we could quickly find all pairs of points within a certain distance. Any two points (x_1, y_1) and (x_2, y_2) were grouped into one cluster if $|x_1 - x_2| < G$ and $|y_1 - y_2| < G$. For each cluster, if the number of points in the cluster was more than a user-defined threshold (default: 5), it was considered as a potential region of enriched fragment endpoints. If the points in the cluster were not within a G x G square region, we used a G x G moving square to find a square region which contained the most of points. The predicted breakpoints were the X and Y coordinates of the right-bottom vertex of the square. The points in the square region were subjected to a statistical test describe below.

Quantification of type 1 evidence

Let n be the number of points in the square region. Each point corresponds to a pair of fragment F_{i1}, F_{i2}, (i = 1, 2, 3, …, n) that may support the SV. Let b_1 and b_2 be the coordinates of the predicted breakpoint. Equation (1) and (2) hold for all the fragment pairs F_{i1}, F_{i2} (i = 1, 2, 3, …, n). We then test the null hypothesis that there is no SV between b_1 and b_2.

First, we test the hypothesis that the n fragment pairs F_{i1}, F_{i2} have originated from different DNA molecules, but coincidently received the same barcode. Here we define two fragments F_a and F_b

as an *independent fragment pair* if F_a and F_b share the same barcode but have originated from different DNA molecules. Thus, $R(F_a)$ and $L(F_b)$ are independent variables. All the fragment pairs that do not support SVs are independent fragment pairs. It is reasonable to assume the generation of HMW DNA molecules from chromosomal DNA is a random process thus both $R(F_a)$ and $L(F_b)$ are uniformly distributed across the chromosome. Therefore, the point $((R(F_a), L(F_b))$ is equal likely to be in any place in the two-dimensional plane. Technically, we connect all the chromosomes in a head-to-tail order so that both intra-chromosomal events and inter-chromosomal can be analyzed at the same time. Observing at least n independent fragment pairs meeting equation (2) is equivalent to the event that observing at least n points $((R(F_{i1}), L(F_{i2}))$ located in a squared region with an area of G^2 on the two-dimensional plane. The probability of this event is:

$$p_1 = \sum_{j=n}^{N} \text{Binomial_pmf}\,(\text{n}, N_{ifp}, \frac{G^2}{L^2}) \quad (3)$$

where Binomial_pmf is the probability mass function of binomial distribution; L is the total length of the genome (also the side length of the two-dimensional plane); N_{ifp} is the total number of independent fragment pairs.

Since we are doing multiple hypothesis testing in the data set, the probability need to be adjusted.

$$p_{adjusted1} = p_1 \frac{G^2}{L^2}$$

We reject the hypotheis if $p_{adjusted1} < p_{threshold}$. $p_{threshold}$ is 10^{-5} by default.

Next, we test the hypothesis that fragment pairs F_{i1}, F_{i2} (i = 1, 2, 3, …, n) have originated from the same DNA molecule, but no reads were sequenced in the gap between $R(F_{i1})$ and $L(F_{i2})$. Let g_i denote the length of the gap between F_{i1} and F_{i2}, $\bar{g}$ denote the mean of g_i, we have:

$$g_i = L(F_{i2}) - R(F_{i1}) \quad (4)$$

$$\bar{g} = \frac{1}{n}\sum_{i=0}^{n} g_i \quad (5)$$

If $\bar{g}$ is too large such that the probability of no reads being generated is smaller than a threshold, we can reject this hypothesis.

Similar to the model described by 10X Genomics[12], we assume the read generation on a DNA molecule is a Poisson process with constant rate λ across the genome. Let r be the number of reads generated in a region of length g, then $r \sim \mathrm{Pois}(\lambda g)$. Let $P_{gap}(g)$ denote the probability of no read being generated in length g, we have:

$$P_{gap}(g) = P(r = 0|\lambda g) = \frac{e^{-\lambda g}(\lambda g)^0}{0!} = e^{-\lambda g} \quad (6)$$

Therefore, the gap length g_i follows Exponential distribution: $g_i \sim \mathrm{Exp}\ (\lambda)$. Recalling that 1) the Exponential distribution with rate parameter λ is a Gamma distribution with shape parameter 1 and rate parameter λ; 2) the sum of n independent random variables from Gamma (1, λ) is a Gamma random variable from Gamma (n, λ), we have:

$$\sum_{i=0}^{n} g_i \sim Gamma\ (n, \lambda) \quad (7)$$

$$\bar{g} = \frac{\sum_{i=0}^{n} g_i}{\mathrm{n}} \sim Gamma\ (n, n\lambda)\ (8)$$

Therefore, the probability that observing n gap regions with mean length equal to or larger than $\bar{g}$ is:

p_2 = 1 - Gamma_cdf (n, $n\lambda$) (9)

where Gamma_cdf is the cumulative distribution function of Gamma distribution.

Since we are doing multiple hypothesis testing in the data set, the probability need to be adjusted.

$p_{adjusted2} = p_2 \frac{N_{rp}}{n}$ (10)

where N_{rp} is the total number of read pairs.

We reject the hypothesis if $p_{adjusted2} < p_{threshold}$. $p_{threshold}$ is set as 10^{-5} by default. If both $p_{adjusted1}$ and $p_{adjusted2}$ are less than $p_{threshold}$, we accept the hypothesis that the SV is true. For each candidate SV, we report a confidence score for type 1 evidence as:

Confidence score1 = $-\log_{10}(\max(p_{adjusted1}, p_{adjusted2}))$ (11)

Breakpoint detection from type 2 evidence

Barcode similarity between two nearby regions is very high because the reads originate from almost the same set of HMW DNA molecules. However, at the SV breakpoint, the aligned reads from the left side and right side may have originated from different locations in the alternative genome. Thus, the barcode similarity between the left side and right side of the breakpoint are dramatically reduced (as described in the Result section and shown in Figure 1e-f). To detect this, LinkedSV uses two adjacent sliding windows (twin windows, moving 100 bp) to scan the genome and calculate the barcode similarity between the twin windows. The window length can be specified by user. By default, it is *G* for WGS data sets and 40 kb for WES data sets.

The barcode similarity can be simply calculated as the fraction of shared barcodes. This method is suitable for WGS, where the coverage is continuous and uniform. But it does not perform well for WES, where the numbers of reads in the sliding windows vary a lot due to capture bias and the length of capture regions. Therefore, we use a model that considering the variation of sequencing depth and capture region positions. The barcode similarity is calculated as:

$$S = \frac{x}{m_1^a m_2^b} n e^{-\alpha d} \quad (12)$$

where:

m_1 is the number of barcodes in window 1,

m_2 is the number of barcodes in window 2,

x is the number of barcodes in both windows,

d is the weight distance between reads of the left window and the right window,

n is a constant representing the characteristic of the library,

α is a parameter of fragment length distribution,

a and b are two parameters between 0 and 1,

n, α, *a* and *b* are estimated from the data using regression. Detailed explanation of this model is in supplemental note 1.

Next, we calculate the empirical distribution of barcode similarity. Regions where the barcode similarity less than a threshold (5th percentile of the empirical distribution by default) were regarded as breakpoint candidates. If a set of consecutive regions have barcode similarity lower than the threshold, we only retain the region that has the lowest barcode similarity. If the barcode similarity of a breakpoint candidate is S_0, the empirical *p*-value is calculated as:

$$\text{empirical } p = \frac{number\ of\ twin\ windows\ with\ S \leq S_0}{total\ number\ of\ twin\ windows} \quad (13)$$

The confidence score of type 2 evidence is:

$$\text{Confidence score2} = -\log_{10}(\text{empirical } p) \quad (14)$$

Combination of both types of evidence

Type 1 evidence gives pairs of endpoints that indicate two genomic positions are joined in the alternative genome. Type 2 evidence gives genomic positions where the barcodes suddenly changed, regardless of which genomic position can be joined. Therefore, type 1 and type 2 evidence are independent. The candidate breakpoints detected from type 2 evidence were searched against the candidate breakpoint pairs detected from type 1 evidence so that the calls were merged. The combined confidence score is:

Combined score = Confidence score1 + Confidence score2a + Confidence score2b

Where Confidence score1 is the confidence score calculated from type 1 evidence (equation 11); Confidence score2a and Confidence score2b are the confidence scores of the two breakpoints calculated from type 2 evidence (equation 14).

Refining breakpoints using discordant read-pairs and split-reads

We search for discordant read-pairs and clipped reads that are within 10 kb to the predicted breakpoint pairs by the above approach. We use a graph-based approach that is similar to DELLY[3] to cluster the discordant read-pairs. We define the supporting split-reads as the clipped reads that can be mapped to the both breakpoints, and the map direction matches the SV type. If both

discordant read-pairs and split-reads are found to support the SV, we use the breakpoints inferred by split-reads as the final breakpoint position.

Generation of simulated linked-read WGS data set

The linked reads were simulated by LRSIM, which can generate linked-reads from a given FASTA file containing the genome sequences. We generated a diploid FASTA file based on hg19 reference genome with SNPs and SVs inserted. The purpose of inserting SNPs was to mimic real data. The generation of the diploid FASTA file is described below. First, we inserted SNPs to hg19 using vcf2diploid [28]. The inserted SNPs were from the gold standard SNP call set (v.3.3.2) of NA12878 genome [29]. The vcf2diploid software generated two FASTA files, each of which was a pseudohaplotype (paternal or maternal) with the phased SNPs inserted. Next, we insert SVs into the paternal FASTA file using our custom script. The breakpoints were located in the repetitive regions in hg19 and the distance between the two breakpoints were in the range of 50 kb to 1 Mb. In total, we simulated 351 deletions, 386 duplications, 353 inversions and 85 translocations, all of which were in the paternal copy and were heterozygous SVs. We then concatenate the paternal and maternal FAST file into a single FASTA file and simulated linked-reads using LRSIM. To mimic real data, the barcode sequences and molecule length distribution used for simulation were

from the NA12878 whole-genome data set released by 10X Genomics. The number of read pairs was set to 360 million so that a 35X coverage data set was generated.

Generation of WGS data set with low variant allele frequencies

In cancer samples or mosaic samples, the total DNA is a mix of a small portion of variant alleles and a large portion of normal alleles. To simulate the WGS data sets with low variant frequencies, we used the same paternal and maternal FASTA file described above but the combined FASTA file contained multiple copies of the normal allele (the maternal FASTA) and only one copy of the variant allele (the paternal FASTA). For example, to simulate a WGS data set with VAF of 20%, four copies of the maternal FASTA and one copy of the paternal FASTA were combined. The linked reads were simulated using LRSIM with the same parameters and a 35X coverage data set was generated.

Generation of simulated linked-read WES data set

To generate the linked-read WES data set, we first generate a 100X linked-read WGS data set and then down-sample it to be a WES data set. Generation of the simulated linked-read WGS data set with SNPs and SVs inserted was similar to the method described above. In total, we inserted 1160

heterozygous SVs. The SV breakpoints were randomly selected from regions that were within 2000 bp of an exon. Among the 2320 breakpoints (two breakpoints per each SV), 1028 breakpoints (44.3%) were in intronic or intergenic regions. The SV sizes are in the range of 50 kb to 1 Mb. The number of inserted SVs in the simulated WES data set was slightly smaller than that in the simulated WGS data set because the SV breakpoints were designed to reside within 2000 bp of an exon. The simulated reads were generated using LRSIM and were mapped to hg19 reference genome using the Longranger pipeline (default settings). The phased bam generated by Longranger was down-sampled to be a simulated WES data set. To mimic real WES data set, we used the coverage distribution of the linked-read WES data set of NA12878 genome (released by 10X Genomics) to guide the down-sampling process. We bin the genome into 10 bp windows and calculate number of reads mapped to each window (left mapping positions were used) in NA12878 linked-read WES data. The simulated WES data set was generated by sampling reads from the 100X WGS data according to number of reads mapped to the same 10 bp window in the NA12878 WES. In this process, if the one read is retained, the paired read will also be retained.

Error correction of PacBio reads

PacBio sequencing data of the HX1 genome was downloaded from NCBI SRA database with the accession SRX1424851. Raw PacBio reads (subreads) were extracted from the HDF5 files using

pbh5tools according to the manufacturer's instructions. Error correction of PacBio reads was performed using Canu[30] (version 1.6) with the parameters for PacBio reads (canu -correct genomeSize=3.1G -pacbio-raw path/to/pacbio_reads.fasta).

Mapping of PacBio data to the human genome

Raw PacBio reads or error-corrected PacBio reads were mapped to human reference genome GRCh38 (without ALT loci) using Minimap2[25] (version: 2.15-r905) with parameters optimized for PacBio reads (-ax map-ont). The output SAM file was converted to BAM file and sorted using SAMtools[31].

Comparison of the call sets of LinkedSV and SMRT-SV on the HX1 genome.

The SV calls of SMRT-SV were downloaded from NCBI dbVar database under accession nstd162. We only compared large SV calls that are at least 10 kb, because linked-read sequencing and LinkedSV had limited power to detect smaller SVs. Two deletion/duplications calls were considered the same if they had at least 50% reciprocal overlap (the overlapped region was more than 50% of both calls). This criterion was chosen to follow what was done by a previous study [32].

Discussion

In this study, we present LinkedSV, a novel open source algorithm for structural variant detection from linked-read sequencing data. We assessed the performance of LinkedSV on three simulated data sets and two real data sets. By incorporating the two types of evidence as outlined below, LinkedSV outperforms all existing linked-read SV callers including Longranger, GROC-SVs and NAIBR on both WGS and WES data sets.

Type 1 evidence gives information about which two genomic positions are connected in the alternative genome. It has two observations: 1) fragments with shared barcodes between two genomic locations and 2) enriched fragment endpoints near breakpoints. Current existing linked-read SV callers only use the first observation to detect SVs while LinkedSV incorporates both observations in the statistical model and is therefore more sensitive and can detect SVs with lower allele frequencies, such as somatic SVs in cancer genomes and mosaic structural variations.

Type 2 evidence gives information about which genomic position is “interrupted” with the observation that the reads on the left side and right side of a genomic position have different barcodes and should be derived from different HMW DNA molecules. LinkedSV is the only SV caller that use type 2 evidence to detect breakpoints. Type 2 evidence is independent to type 1 evidence, and gives additional confidence to identify the breakpoints. In addition, type 2 evidence can be detected locally, which means we can detect a weird genomic location without looking at

the barcodes of the other genomic locations. This is particularly useful in two situations: 1) novel sequence insertions where there is only one breakpoint; 2) only one breakpoint is detectable and the other breakpoint located in a region where there is little coverage within 50 kb, which is often the case in target region sequencing.

In recent years, WES has been widely used to identify disease causal variants for patients with suspected genetic diseases in clinical settings. Identification of SVs from WES data sets are more challenging because the SV breakpoints may not be in the capture regions and thus there would be little coverage at the breakpoints. Linked-read sequencing increases the chance of resolving such type of SVs by providing long-range information. As well as there are a few capture regions nearby, the fragments can still be reconstructed and type 1 and type 2 evidence can still be observed. Our statistical models for both type 1 evidence and type 2 evidence were designed to handle both WGS and WES data sets. GROC-SVs uses a local-assembly method to verify the SV call, which requires sufficient coverage at the breakpoints. By using these two types of evidence, LinkedSV can be less relied on short-read information (e.g. pair-end reads and split-reads). We demonstrated that LinkedSV has better recall and balanced accuracy (F1 score) on the simulated WES data set and can detect SVs even when the breakpoints were not located in capture regions and have no short-read support. In addition, LinkedSV is also the only SV caller that clearly detected the *F8* intron 22 inversion and *NF1* deletion from the clinical WES data sets.

Linked-read sequencing has several advantages over traditional short-read sequencing on the purpose of SV detection. First, the human genome is highly repetitive. Previous studies have shown that SVs are closely related to repeats and many SVs are directly mediated by homologous recombination between repeats [33]. In traditional short-read sequencing, if the breakpoint falls in a repeat region, the supporting reads would be multi-mapped and thus the SV cannot be confidently identified. However, this type of SVs are detectable by Linked-read sequencing as well as the HMW DNA molecules span the repeat region. We can observe type 1 and type 2 evidence in the non-repeat region nearby. In our benchmarking, LinkedSV resolved 95% SVs of which breakpoints located in repeat regions. Secondly, SVs are undetectable from traditional short-read sequencing if there is little coverage at the breakpoints, which is often the case in WES data sets. As described above, this type of SV can also be resolved by linked-read sequencing and LinkedSV. Third, linked-read sequencing requires less coverage for detection of SVs with low variant allele frequencies. In linked-read sequencing data, short read pairs are sparsely and randomly distributed along the HMW DNA molecule. In a typical linked-read WGS data set, the average distance between two read pairs derived from the same HMW DNA molecule is about 1000 bp and each HMW DNA molecule only has a short-read coverage of about 0.2X. Therefore, there are about 150 HMW DNA molecules (reconstructed fragments) covering a genomic location of 30X depth. A SV of 10% VAF will have 15 supporting fragment pairs in a 30X depth location in linked reads WGS data set, which is sufficient to be detected by LinkedSV. However, a SV of 10% VAF will

only 3 supporting read pairs in a 30X depth location in traditional short read WGS, which makes the detection more challenging.

Linked-read sequencing also has several advantages over long-read sequencing in terms of SV detection. The fragment length of linked-reads (typically ~50 kb) is longer than the read length of regular long-read sequencing (typically 20-30 kb). Therefore, linked-read sequencing has unique advantages for detection of large SVs. In our study, LinkedSV detected many large SVs that were missed in the long-read SV call set. We also showed that the sequencing error (13-15%) of long-read sequencing technologies had a negative effect on reads mapping and subsequent SV calling. In terms of library preparation, Linked-read sequencing only requires 1 ng input DNA, which is two orders of magnitude smaller than what is needed by long-read sequencing. Therefore, disease samples of very low DNA amount can be easily sequenced by linked-read sequencing. In addition, SNPs, indels and SVs can be detected from linked-read sequencing simultaneously.

The linked-read technology provides strong evidence to detect large SVs, but it provides little additional evidence to detect small SVs. Therefore, LinkedSV has limited power to detect SVs that are shorter than 10 kb. However, large SVs are more likely to be harmful and to cause diseases. Therefore, we still expect a strong ability of linked-read technology to resolve harmful and disease causal SVs. Like the existing linked-read SV callers, LinkedSV currently does not handle novel sequence insertions. As future work, we plan to detect novel sequence insertions using type 2

evidence, since this type of SV also cause a decrease of barcode similarity between nearby regions and can be detected by the twin-window method.

In summary, we present LinkedSV, a novel SV caller for linked-read sequencing. LinkedSV outperformed current existing SV callers, especially for identifying SVs with low allele frequency or identifying SVs from target region sequencing such as linked-read WES. We expect LinkedSV will facilitate the detection of structural variants from linked-read sequencing data and help solve negative cases of conventional short-read sequencing.

Competing Interests

The authors declare no competing interests.

Acknowledgments

The authors would like to thank members of the Center for Applied Genomics for generating the linked read exome sequencing data on the patient with *F8* inversion and *NF1* deletion. We would like to thank the authors of the simulation software LRSIM to provide tools that facilitated our benchmarking study.

References

1 Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* **14**, 125-138, doi:10.1038/nrg3373 (2013).

2 Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. 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, doi:10.1093/bioinformatics/btp394 (2009).

3 Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339, doi:10.1093/bioinformatics/bts378 (2012).

4 Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**, 677-681, doi:10.1038/nmeth.1363 (2009).

5 Chong, Z. *et al.* novoBreak: local assembly for breakpoint detection in cancer genomes. *Nat Methods* **14**, 65-67, doi:10.1038/nmeth.4084 (2017).

6 Wala, J. A. *et al.* SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res* **28**, 581-591, doi:10.1101/gr.221028.117 (2018).

7 Carvalho, C. M. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* **17**, 224-238, doi:10.1038/nrg.2015.25 (2016).

8 Payer, L. M. *et al.* Structural variants caused by Alu insertions are associated with risks for many human diseases. *Proc Natl Acad Sci U S A* **114**, E3984-E3992, doi:10.1073/pnas.1704117114 (2017).

9 Sharp, A. J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* **77**, 78-88, doi:10.1086/431652 (2005).

10 Chaisson, M. J. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608-611, doi:10.1038/nature13907 (2015).

11 Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods*, doi:10.1038/s41592-018-0001-7 (2018).

12 Zheng, G. X. *et al.* Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* **34**, 303-311, doi:10.1038/nbt.3432 (2016).

13 Luo, R., Sedlazeck, F. J., Darby, C. A., Kelly, S. M. & Schatz, M. C. LRSim: A Linked-Reads Simulator Generating Insights for Better Genome Partitioning. *Comput Struct Biotechnol J* **15**, 478-484, doi:10.1016/j.csbj.2017.10.002 (2017).

14 Spies, N. *et al.* Genome-wide reconstruction of complex structural variants using read clouds. *Nat Methods*, doi:10.1038/nmeth.4366 (2017).

15 Elyanow, R., Wu, H. T. & Raphael, B. J. Identifying structural variants using linked-read sequencing data. *Bioinformatics*, doi:10.1093/bioinformatics/btx712 (2017).

16 Bishara, A. *et al.* Read clouds uncover variation in complex regions of the human genome. *Genome Res* **25**, 1570-1580, doi:10.1101/gr.191189.115 (2015).

17 Campbell, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* **40**, 722-729, doi:10.1038/ng.128 (2008).

18 Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer* **7**, 233-245, doi:10.1038/nrc2091 (2007).

19 Stephens, P. J. *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* **462**, 1005-1010, doi:10.1038/nature08645 (2009).

20 Lakich, D., Kazazian, H. H., Jr., Antonarakis, S. E. & Gitschier, J. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. *Nat Genet* **5**, 236-241, doi:10.1038/ng1193-236 (1993).

21 De Brasi, C. D. & Bowen, D. J. Molecular characteristics of the intron 22 homologs of the coagulation factor VIII gene: an update. *J Thromb Haemost* **6**, 1822-1824, doi:10.1111/j.1538-7836.2008.03094.x (2008).

22 Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**, R84, doi:10.1186/gb-2014-15-6-r84 (2014).

23 Shi, L. *et al.* Long-read sequencing and de novo assembly of a Chinese genome. *Nat Commun* **7**, 12065, doi:10.1038/ncomms12065 (2016).

24 Huddleston, J. *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res* **27**, 677-685, doi:10.1101/gr.214007.116 (2017).

25 Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100, doi:10.1093/bioinformatics/bty191 (2018).

26 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).

27 She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927-930, doi:10.1038/nature03062 (2004).

28 Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**, 522, doi:10.1038/msb.2011.54 (2011).

29 Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* **32**, 246-251, doi:10.1038/nbt.2835 (2014).

30 Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722-736, doi:10.1101/gr.215087.116 (2017).

31 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

32 Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods* **12**, 780-786, doi:10.1038/nmeth.3454 (2015).

33 Startek, M. *et al*. Genome-wide analyses of LINE-LINE-mediated nonallelic homologous recombination. *Nucleic Acids Res* **43**, 2188-2198, doi:10.1093/nar/gku1394 (2015).

Figures

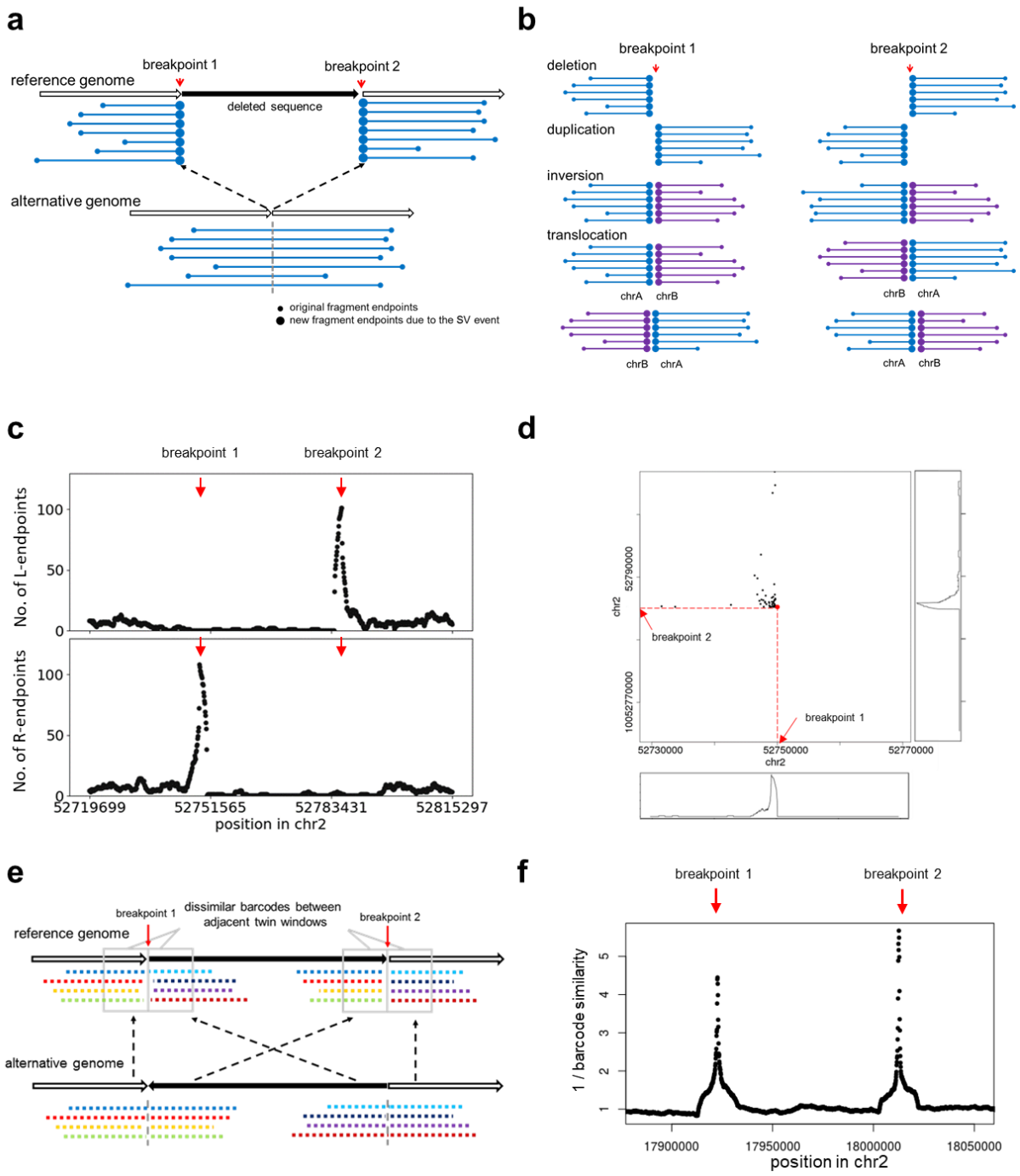


Figure 1

Two types of evidence near SV breakpoints. a) Type 1 evidence. Reads from HMW DNA molecules that span the breakpoints of a deletion are mapped to two genomic locations, resulting in two sets of observed fragments and two sets of newly introduced fragment endpoints (large dots). **b**) The patterns of enriched fragment endpoints indicate the SV types. **c**) Enriched fragment endpoints detected near two breakpoints of a deletion on NA12878 genome. L-endpoints and R-endpoints are plotted separately. The breakpoint positions are marked by red arrows. **d**) Two-dimensional view of enriched endpoints near the two breakpoints of the deletion. Each dot indicates a pair of fragments which share the same barcode and thus may support the SV. The x-value of the dot is the position of the first fragment's R-endpoint and the y-value of the dot is the position of the second fragment's L-endpoint. The background of the 2D plot is cleaner than the 1D plot (panel c) since the fragments that do not share barcodes are excluded. **e**) Type 2 evidence. Reads from two breakpoints of an inversion being mapped to nearby positions (in the grey rectangles), resulting in decreased barcode similarity between the two nearby positions. **f**) Decreased barcode similarity near the breakpoints of an inversion on NA12878 genome. The reciprocal of barcode similarity is shown in the figure. The peaks indicate the positions of the breakpoint.

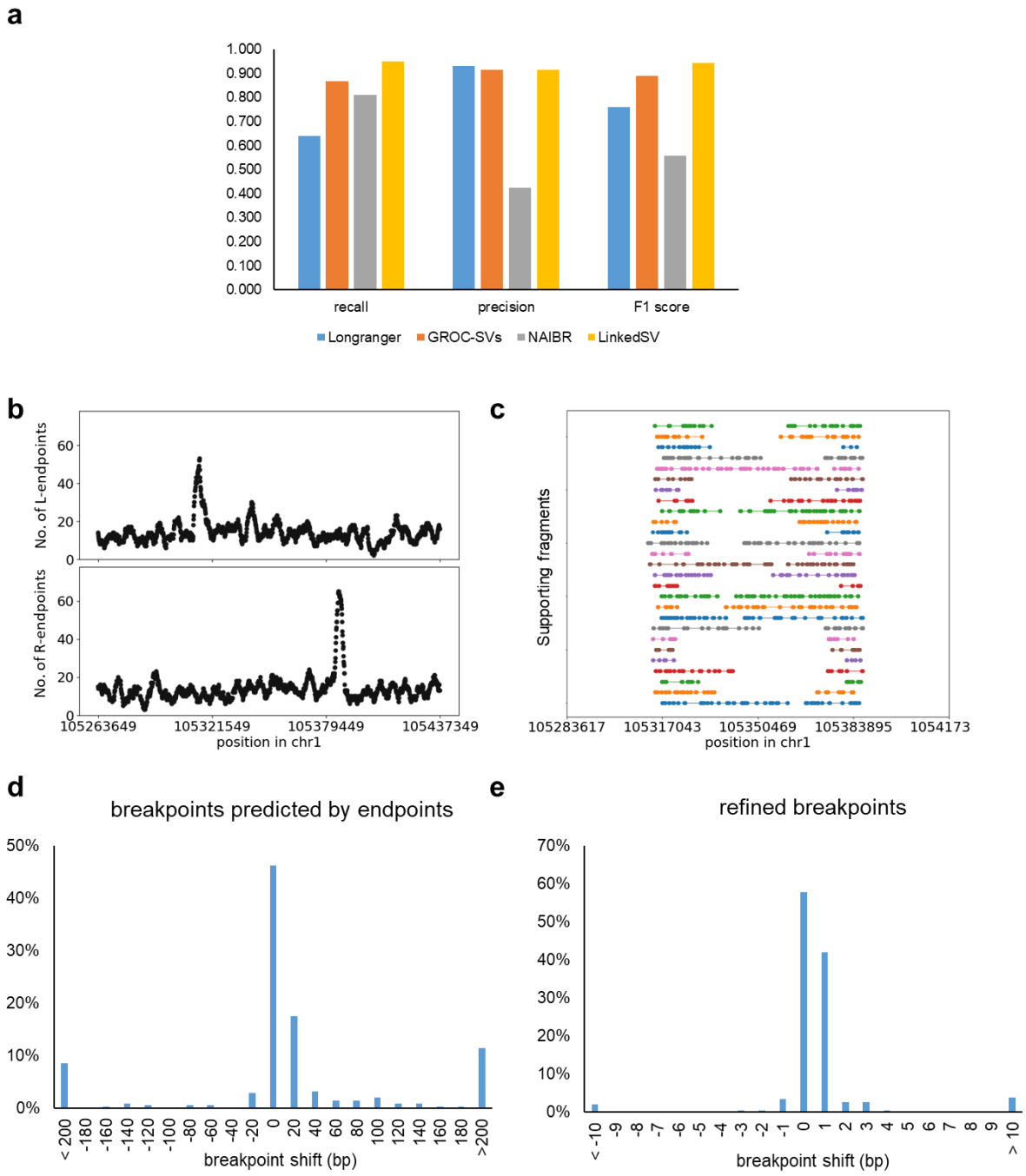


Figure 2

Performance of LinkedSV on the simulated WGS data set. a) Recalls, precisions and F1 scores of four linked-read SV callers on the simulated WGS data set. **b**) Fragment endpoint signals of a small duplication that was missed by GROC-SVs. The peaks indicate the approximate breakpoint

positions. **c**) Supporting fragments of the duplications. Horizontal lines represent linked reads with the same barcode; dots represent reads; colors indicate barcodes. **d**) Precision of breakpoints predicted by LinkedSV without checking short-read information. **e**) Precision of LinkedSV refined breakpoints using discordant read-pairs and split-reads.

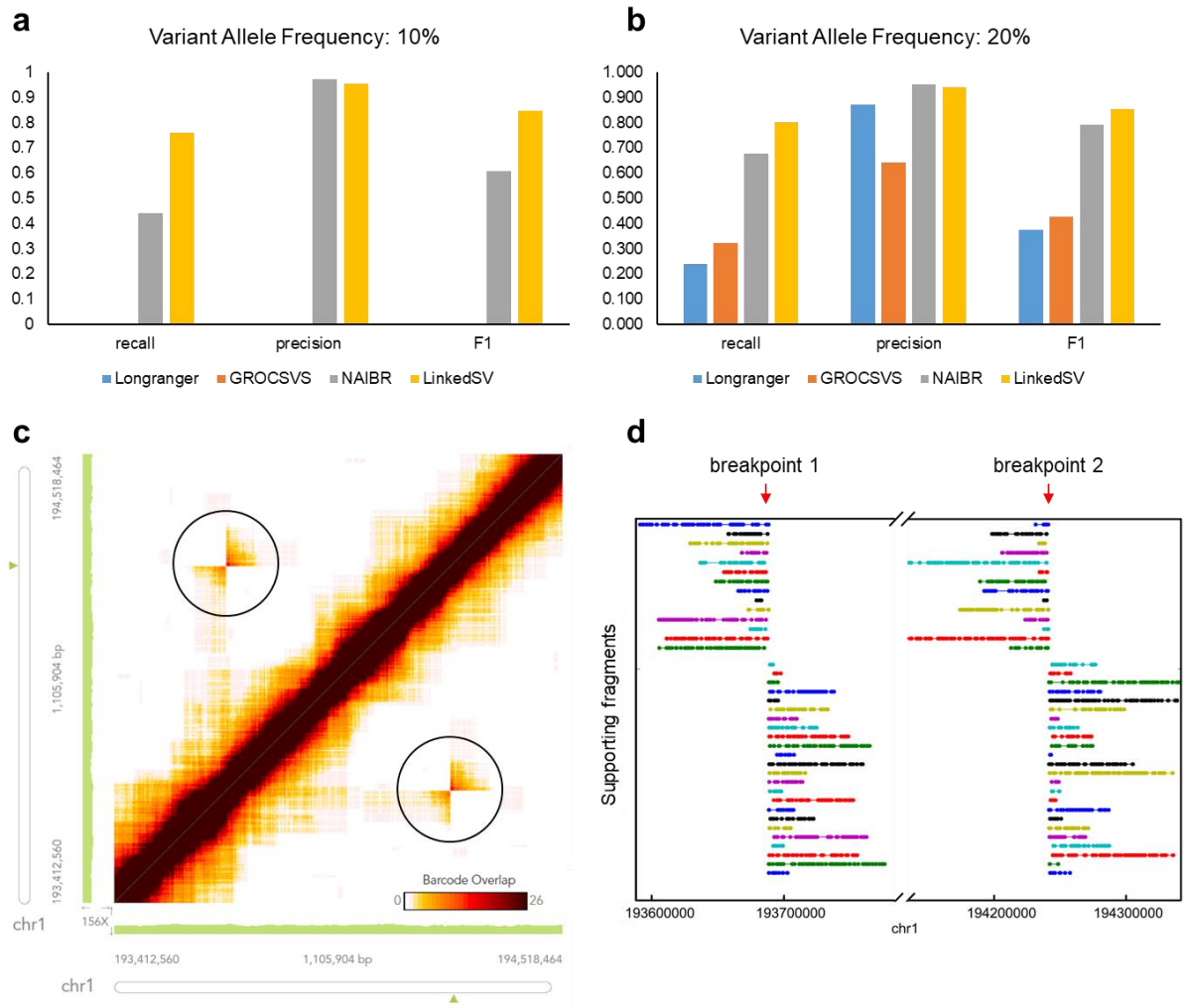


Figure 3

Performance of LinkedSV on the simulated WGS data with low variant allele frequencies. a, b) Recalls, precisions and F1 scores of four linked-read SV callers on the simulated WGS data set with VAF of 10% and 20%. **c**) Heap map of overlapping barcodes in chr1:193412560-194518464 (hg19 coordinates) showing an inversion that was missed by Longranger, and NAIBR (VAF = 10%). The overlapping barcodes between the two inversion breakpoints can be clearly visualized (in the black circles). The heat map was plotted by the Loupe software (10X Genomics). Dots represent overlapping barcodes. **d**) Supporting fragments of the inversion detected by LinkedSV. Horizontal lines represent linked reads with the same barcode; dots represent reads; colors indicate barcodes. Predicted breakpoint positions are marked by red arrows.

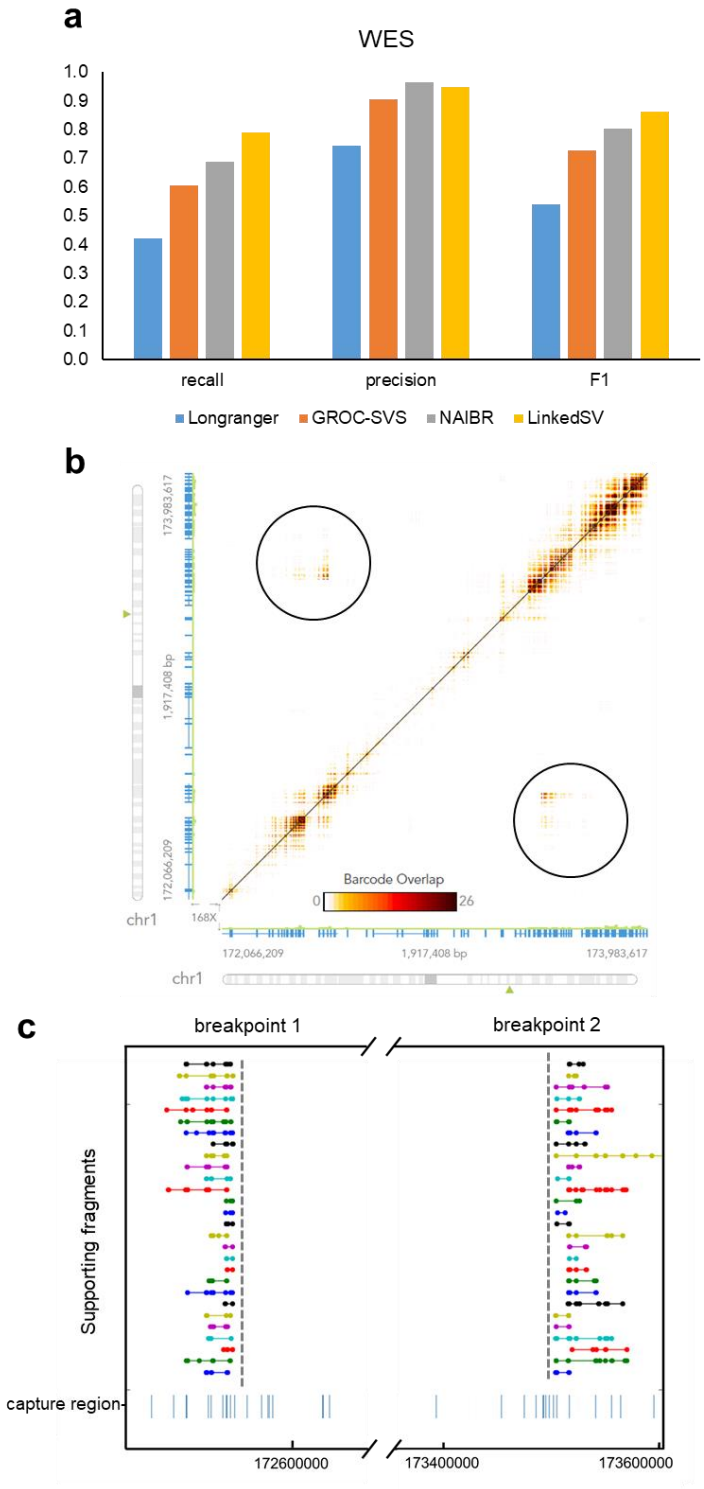

a
WES
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
recall
precision
F1
Longranger
GROC-SVS
NAIBR
LinkedSV
b
173,983,617
1,917,408 bp
172,066,209
chr1
168X
Barcode Overlap
0
26
172,066,209
1,917,408 bp
173,983,617
chr1
c
breakpoint 1
breakpoint 2
Supporting fragments
capture region
172600000
173400000
173600000

Figure 4

Performance of LinkedSV on the simulated WES data set. a) Recalls, precisions and F1 scores of four linked-read SV callers on the simulated WES data set. **b**) Heat map showing a deletion that was missed by NAIBR. The overlapping barcodes between the two breakpoints can be clearly visualized (in the black circles). The heat map was plotted by the Loupe software. Dots represent overlapping barcodes. **c**) Supporting fragments of the deletion detected by LinkedSV. Horizontal lines represent linked reads with the same barcode; dots represent reads; colors indicate barcodes. Predicted breakpoint positions are marked by red arrows. Capture regions were shown as vertical bars in the bottom.

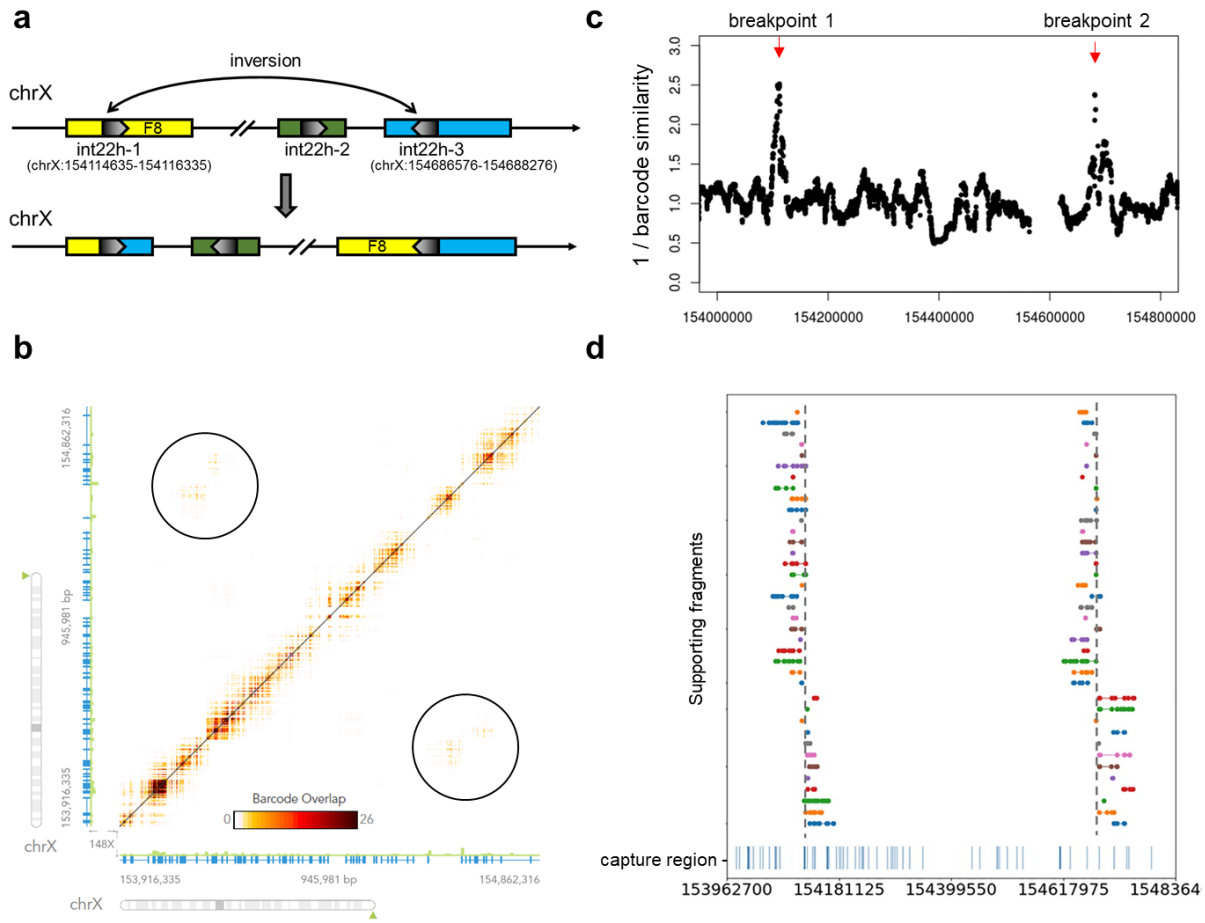


Figure 5

Detection of *F8* inversion from clinical exome sequencing data. (**a**) Illustration of type I inversion of *F8* gene. A portion of intron 22 has three copies in chrX (int22h-1, int22h-2, int22h-3). The inversion is induced by the homologous recombination between two inverted copies int22h-1 and int22h-3. Int22h-1 is located in intron 22 of *F8* gene and int22h-3 is located in the intergenic regions. (**b**) Heat map of overlapping barcodes in chrX:153916335-154862316 (hg19 coordinates), plotted by the Loupe software tool. Black circles indicate overlapping barcodes near the inversion breakpoints. Dots represent overlapping barcodes. (**c**) Decreased barcode similarity at breakpoints detected by the twin window method of LinkedSV. Window size = 40 kb (**d**)

Supporting fragments detected by LinkedSV. Horizontal lines represent linked reads with the same barcode; dots represent reads; colors indicate barcodes. Dashed vertical black lines represent breakpoints. Capture regions were shown as vertical bars in the bottom.

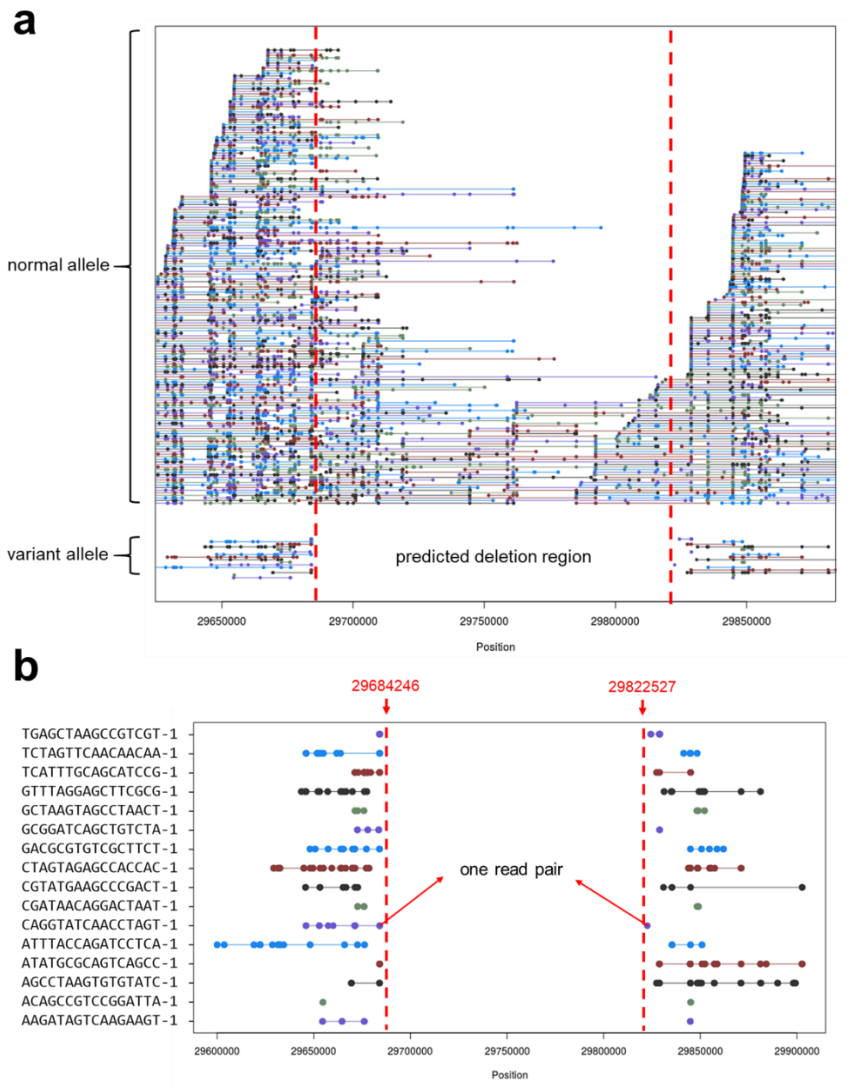


Figure 6

Detection of *NF1* deletion from clinical exome sequencing data. (**a**) Plot of linked-reads for *NF1* WES sample spanning chr17:29645000-29855000. In the normal allele (top), there are 71 fragments crossing over the left breakpoint and 38 fragments crossing over the right breakpoint. In the variant allele (bottom), the linked reads are separated by a large gap. Horizontal lines represent linked reads with the same barcode; dots represent reads; colors indicate barcodes. Dashed vertical red lines represent breakpoints. (**b**) Zoom-in plot of supporting fragments for the deletion. One read pair was found to support the deletion.

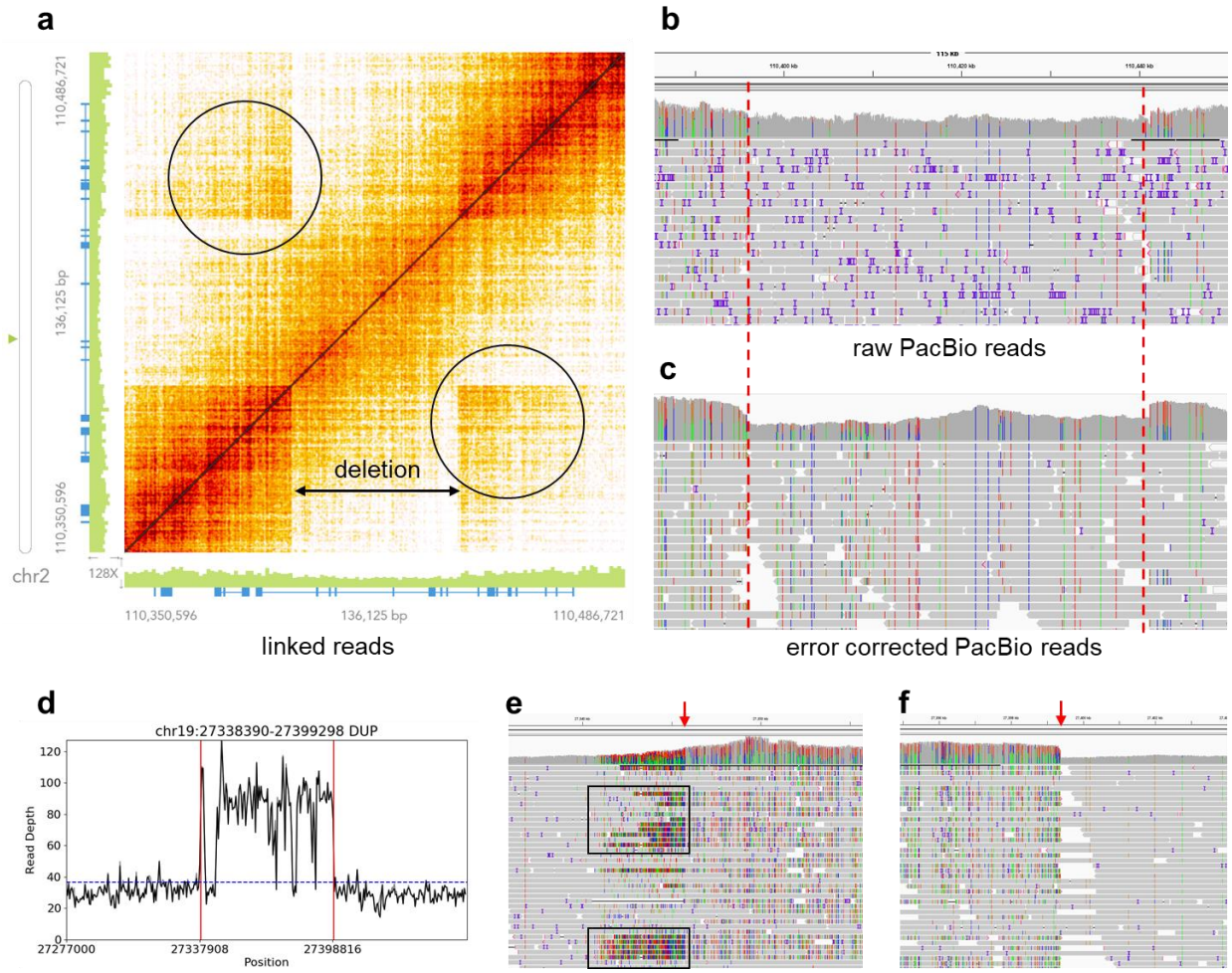
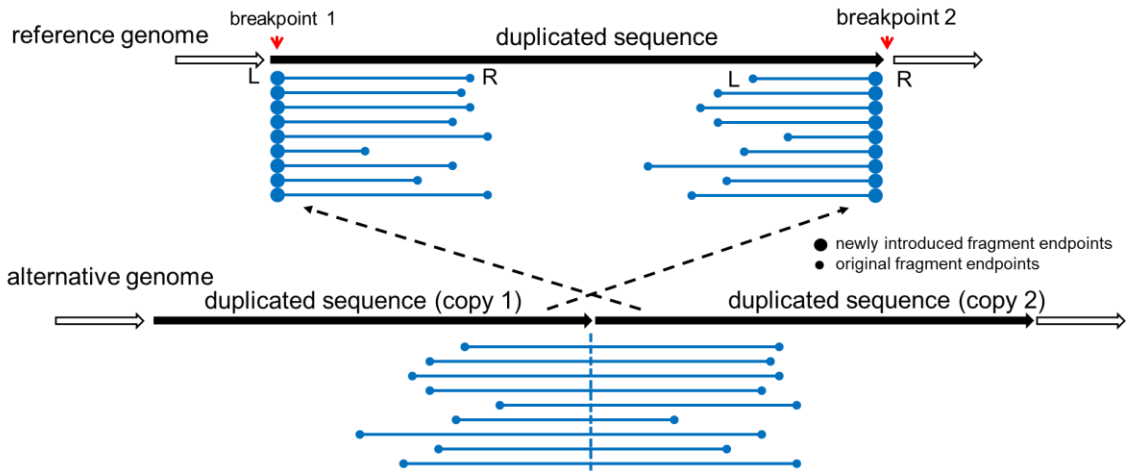


Figure 7

Structural variants detected from linked-read WGS data of the HX1 genome. a) Heatmap of overlapping barcodes for a 45 kb deletion on chromosome 2 (chr2:110360596-110476721, hg38 coordinates), plotted by the Loupe software tool. Black circles indicate overlapping barcodes near the breakpoints. The deletion was not detected by SMRT-SV from PacBio long reads; **b**) Aligned PacBio reads near the deletion region, as shown in IGV. The deletion boundaries were not clear. **c**) Error-corrected PacBio reads near the deletion region, as shown in IGV. The deletion boundaries were very clear (vertical red lines). **d**) Read depth distribution near a 61 kb duplication region on

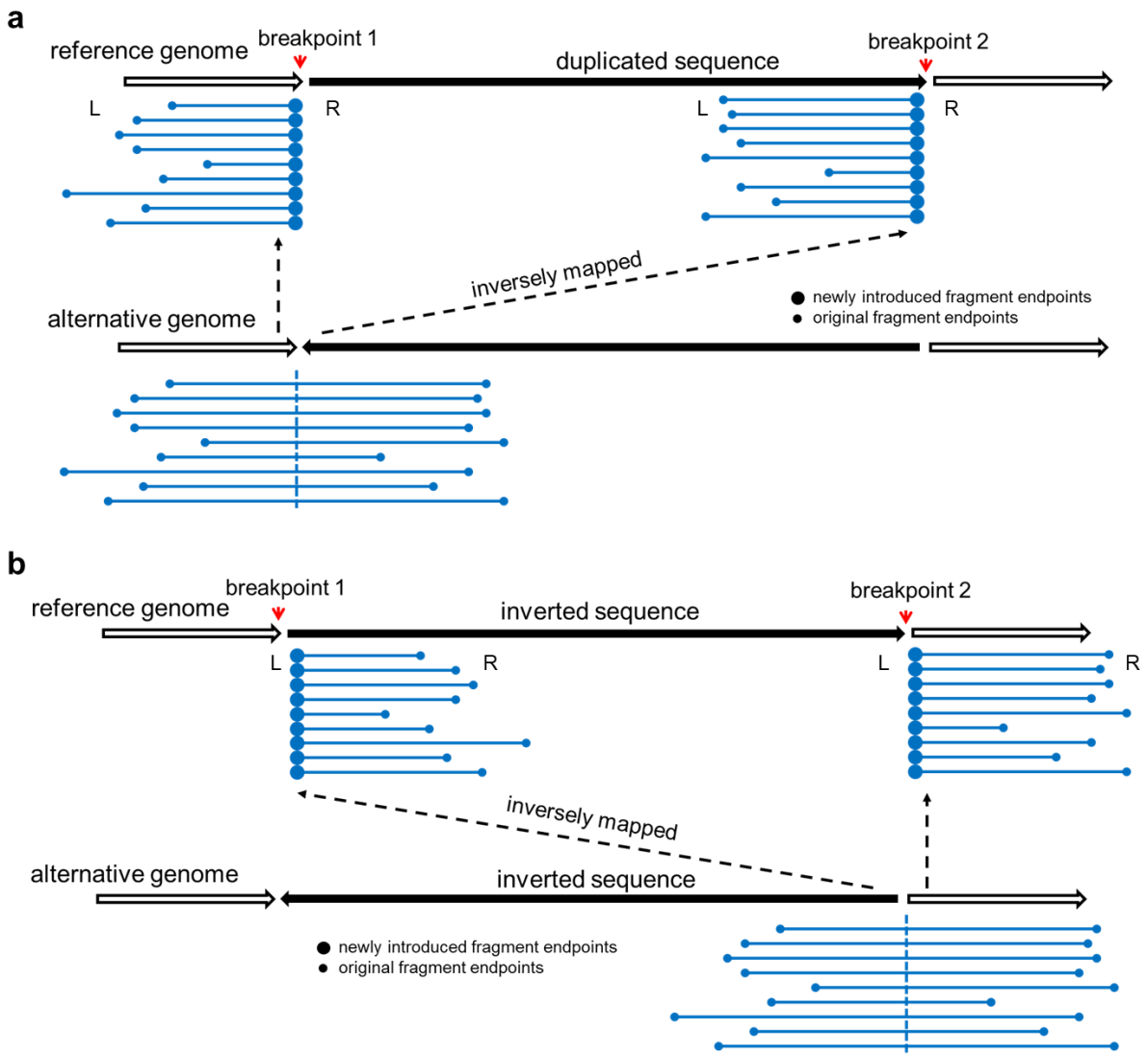
chromosome 19 (chr19:27338390-27399298, hg38 coordinates). The calculation was based on the bam file of linked-reads. Only reads with mapping quality >= 20 were counted. The dotted blue line showed the average depth across the whole genome. The predicted breakpoints were indicated by vertical red lines. The duplication was not detected by SMRT-SV using PacBio long reads; **e**, **f**) Aligned PacBio raw reads near the two breakpoints of the duplication, as shown in IGV. Increased alignment mismatches due to the SV were observed in **e**) (black rectangles). A clear duplication breakpoint was observed in **f**).

Supplementary Figures



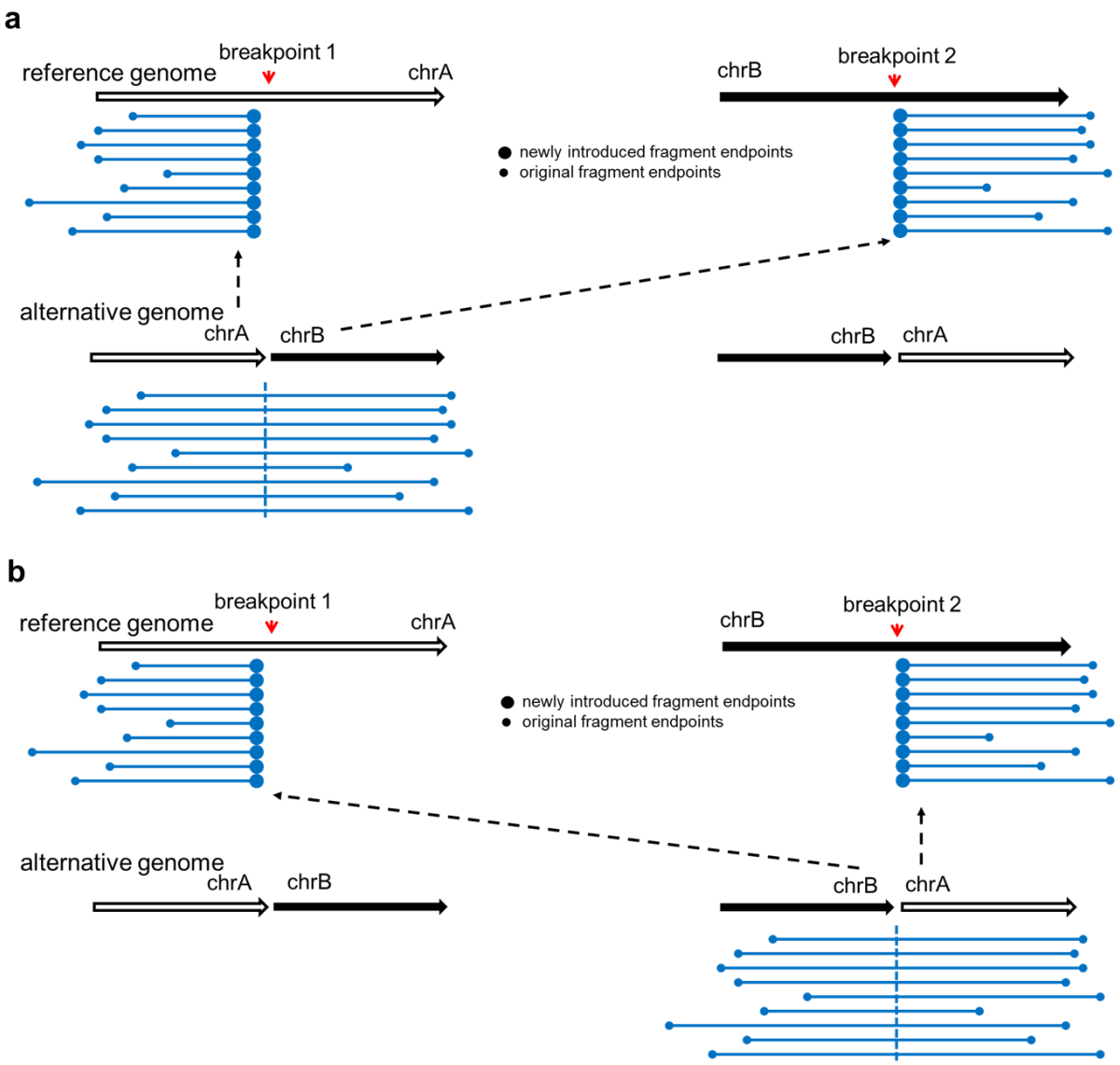
Supplemental Figure 1

The pattern of enriched fragment endpoints for tandem duplications. L-endpoints and R-endpoints are enriched near breakpoint 1 and breakpoint 2, respectively. Breakpoints are marked by red arrows.



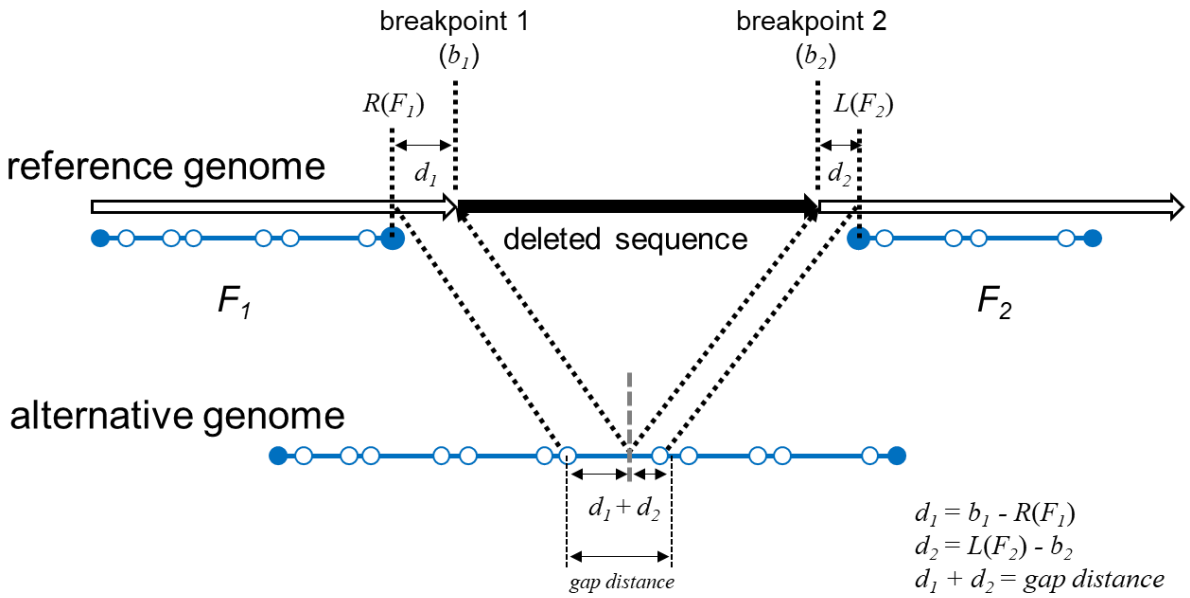
Supplemental Figure 2

The pattern of enriched fragment endpoints for inversions. The endpoints enriched near the two breakpoints are of the same type. **a**) Enriched fragment endpoints are both R-endpoints. **b**) Enriched fragment endpoints are both L-endpoints.



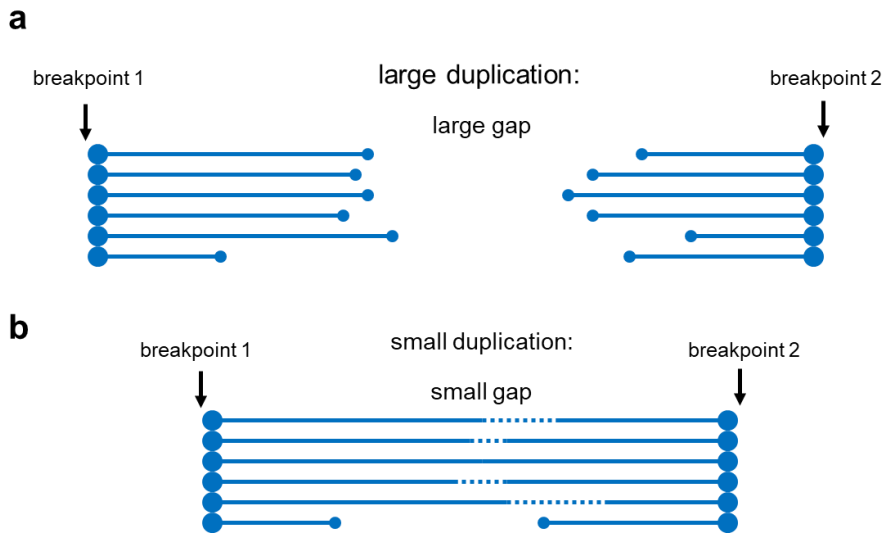
Supplemental Figure 3

The pattern of enriched fragment endpoints for interchromosomal translocations. The enriched endpoints can be either L-endpoints or R-endpoints, depending on how the two chromosomes are joined.



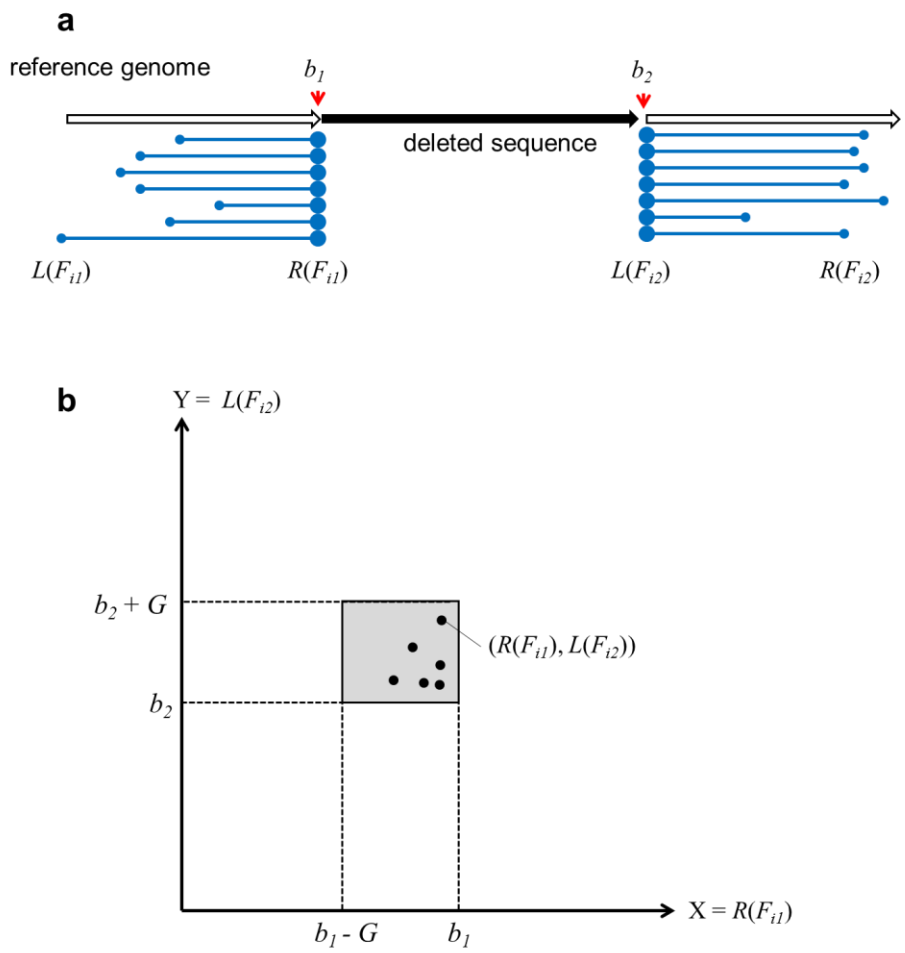
Supplemental Figure 4

Explanation of the enrichment of fragment endpoints near the breakpoints, using a deletion as an example. In this event, reads from a HMW DNA molecule that spans the breakpoints of a deletion were mapped to two genomic locations, resulting in two observed fragments (denoted by F_1 and F_2). b_1 and b_2 denote the positions of breakpoint 1 and 2. $R(F_1)$ denotes the right endpoint of F_1 and $L(F_2)$ denotes the left endpoint of F_2. d_1 and d_2 denote the distances between the fragment endpoints ($R(F_1)$, $L(F_2)$) and the corresponding breakpoints (b_1, b_2), respectively. In the original fragment of the alternative genome, $d_1 + d_2$ equals to the distance between two adjacent reads (i.e. gap distance). Therefore, gap distance is the upper limit of d_1 and d_2. Solid blue dots are endpoints of the fragments. Hollow circles are the short reads in the fragments.



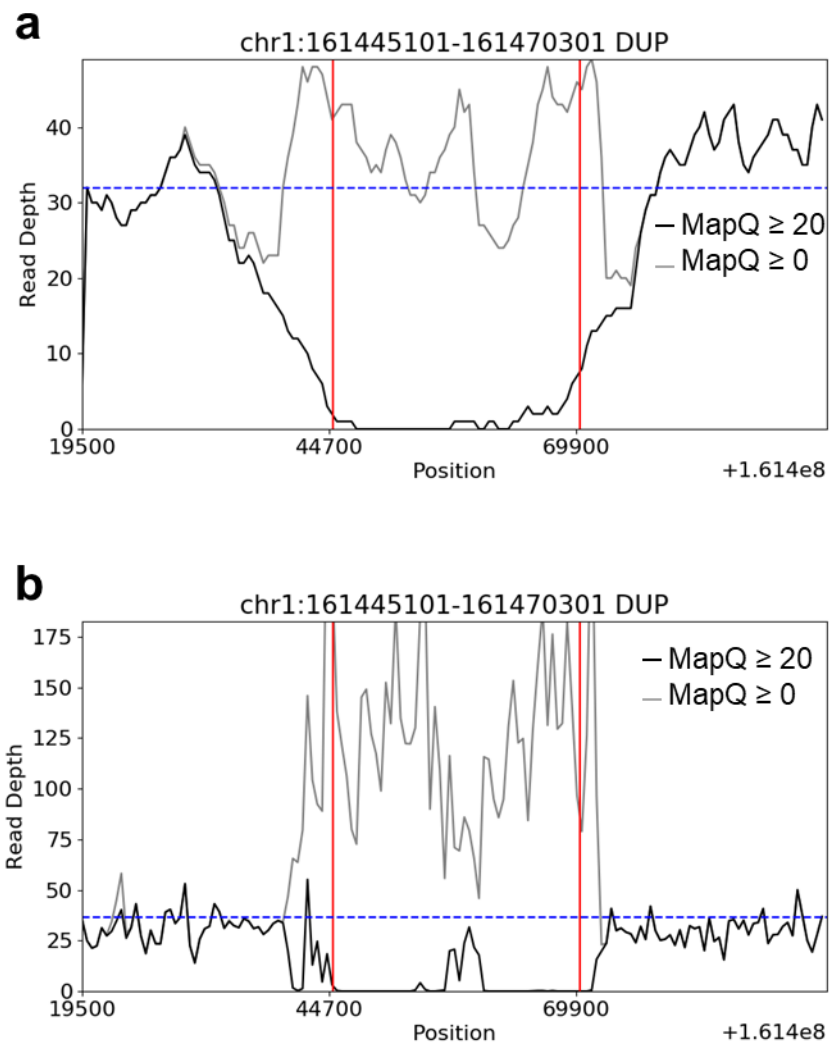
Supplemental Figure 5

a) For large duplications, the reads of the alternative allele are separated by a large gap so that we can observe two sets of fragments with the same set of barcodes, which indicate an SV. **b**) If the duplication is not large enough, the reads will be probably clustered into one fragment.



Supplemental Figure 6

Detection of type 1 evidence. We use deletion as an example but the method is also suitable for the other SV types. **a**) R-endpoints ($R(F_{i1})$) and L-endpoints ($L(F_{i2})$) are enriched near the deletion breakpoints b_1 and b_2, respectively. **b**) Two-dimensional plot of $R(F_{i1})$ and $L(F_{i2})$. $R(F_{i1})$ and $L(F_{i2})$ are restricted in the grey square according to equation (2). The background noise of the two-dimensional plot is cleaner than the one-dimensional plot since the fragments that do not share barcodes are excluded.



Supplemental Figure 7

Read depth distribution of PacBio long reads (**a**) and linked reads (**b**) near a duplication reported by SMRT-SV. The black lines showed the read depth of reads with mapping quality ≥ 20 while the grey lines showed the read depth of reads with mapping quality ≥ 0 (e.g. all reads). The read depth was calculated using SAMtools. The dotted blue line showed the average depth across the whole genome. The large space between the black line and the grey line indicates poor mapping qualities in the predicted duplication region.

886

Supplementary Tables

Supplemental Table 1

SV calls detected by Delly on the *F8* inversion sample (10 kb upstream/downstream of the inversion).

Chrom	Start	End	SV type
chrX	154002063	154002446	INV
chrX	154004508	154004709	INV
chrX	154005670	154005855	INV
chrX	154011709	154012062	DUP
chrX	154012155	154012297	INV
chrX	154014432	154014542	INV
chrX	154014535	154014759	INV
chrX	154017795	154017948	INV
chrX	154020109	154020407	INV
chrX	154045287	154045525	INV
chrX	154045633	154045781	INV
chrX	154057649	154057833	INV
chrX	154057785	154057923	INV
chrX	154083952	154084185	INV
chrX	154091092	154091440	INV
chrX	154115196	154115584	INV
chrX	154115667	154115957	INV
chrX	154124101	154124495	DUP
chrX	154129769	154129919	INV
chrX	154130086	154130222	INV
chrX	154132781	154133165	DUP
chrX	154146261	154146462	INV
chrX	154146565	154146775	INV
chrX	154152621	154152772	INV
chrX	154157646	154157941	DUP
chrX	154158133	154158450	DUP
chrX	154158358	154158630	INV
chrX	154159826	154160143	INV
chrX	154175838	154176055	INV
chrX	154176073	154176451	INV
chrX	154194502	154194721	INV
chrX	154208905	154209847	DUP
chrX	154209557	154209679	INV
chrX	154212475	154212843	INV

chrX	154212596	154212733	INV
chrX	154227431	154227667	INV
chrX	154272329	154272476	INV
chrX	154272335	154272592	DUP
chrX	154275546	154275664	INV
chrX	154275671	154275934	INV
chrX	154290114	154290323	INV
chrX	154344170	154344430	INV
chrX	154387813	154429972	INV
chrX	154387928	154429779	INV
chrX	154391126	154391291	INV
chrX	154426121	154426264	INV
chrX	154428351	154428890	DUP
chrX	154464614	154464716	INV
chrX	154508570	154508865	INV
chrX	154540689	154540872	INV
chrX	154609922	154610441	DUP
chrX	154610132	154610342	INV
chrX	154611498	154611783	INV
chrX	154611510	154688681	DEL
chrX	154612459	154612925	DUP
chrX	154613017	154613356	INV
chrX	154652935	154653319	DUP
chrX	154687101	154687535	DUP
chrX	154687312	154687480	INV
chrX	154688521	154688851	INV
chrX	154688714	154688969	INV
chrX	154689667	154689984	INV
chrX	154735308	154735475	INV
chrX	154755098	154755225	INV

Supplemental Table 2

SV calls detected by Lumpy on the *F8* inversion sample (all calls in chrX).

Chrom	Start	End	SV type
chrX	1413852	1414637	DEL
chrX	17520630	17520931	DEL
chrX	28749622	28749739	DUP
chrX	31189715	31189837	DUP
chrX	38145097	38145387	DUP
chrX	38145261	38145866	DUP

chrX	38145731	38145851	DEL
chrX	38145513	38145876	DEL
chrX	38145369	38145943	DUP
chrX	38145283	38145943	DUP
chrX	38145180	38145975	DUP
chrX	38145091	38145976	DUP
chrX	38145835	38145920	DEL
chrX	38145729	38145981	DUP
chrX	38145874	38145914	DEL
chrX	38145321	38145988	DUP
chrX	38145471	38145995	DUP
chrX	38145929	38146016	DUP
chrX	38145971	38146003	DEL
chrX	38145885	38146041	DUP
chrX	38145642	38146055	DUP
chrX	38146077	38146257	DUP
chrX	39932139	39932332	DUP
chrX	41788764	41788966	DUP
chrX	47308661	47308783	DUP
chrX	63411862	63412017	DUP
chrX	64063531	64063690	DUP
chrX	76890054	76890207	DUP
chrX	101912168	101912314	DUP
chrX	142794887	142795158	DUP
chrX	149745015	149745169	DUP

Supplemental Table 3

SVs on chrX detected by Longranger on the *F8* inversion sample.

Chrom 1	Position 1	Chrom 2	Position 2	SV type	Distance to int22h-1	Distance to int22h-3
chrX	3735199	chrX	3855112	Unknown	-150380286	-150832314
chrX	7810783	chrX	8151613	Unknown	-146304702	-146535813
chrX	7810783	chrX	8113116	Unknown	-146304702	-146574310
chrX	134855892	chrX	134985727	Unknown	-19259593	-19701699
chrX	152225317	chrX	152352113	Unknown	-1890168	-2335313
chrX	154091033	chrX	154660067	Unknown	-24452	-27359
chrX	154131339	chrX	154735755	Unknown	15854	48329

Supplemental Table 4

SVs detected by GROC-SVs on the *F8* inversion sample.

Chrom 1	Position 1	Chrom 2	Position 2	Orientation
chr1	12889258	chr1	12938883	+-
chr1	12919205	chr1	12853022	--
chr1	21754487	chr1	21794667	+-
chr1	148026036	chr1	144622101	++
chr1	223725671	chr1	223797843	+-
chr2	98162357	chr2	97860318	--
chr2	149687145	chr2	149790581	+-
chr2	234053740	chr2	234002965	-+
chr3	129809651	chr3	129762840	-+
chr4	9452594	chr4	9485059	+-
chr4	69681327	chr4	69893173	--
chr5	155137378	chr5	155188725	+-
chr5	180430554	chr5	180375038	-+
chr6	29909733	chr6	29843849	--
chr6	29913575	chr6	29844437	++
chr6	160956444	chr6	160877743	--
chr7	100550750	chr7	100609409	--
chr7	100555813	chr7	100610572	++
chr11	1162747	chr11	1212758	+-
chr11	5809264	chr11	5777102	-+
chr12	7239875	chr12	7189849	-+
chr12	11545335	chr12	11503243	-+
chr12	18018173	chr12	17922871	--
chr12	109372790	chr12	109423610	+-
chr12	132926655	chr16	86452753	+-
chr12	133041064	chr2	231869678	+-
chr13	114325993	chr13	114425990	+-
chr14	24436900	chr14	24474868	+-
chr15	20740860	chr15	23406226	+-
chr15	22743596	chr15	23572115	+-
chr15	23407964	chr15	20739466	+-
chr15	23573198	chr15	22742499	+-
chr15	28597075	chr15	28806596	++
chr15	28804860	chr15	28595402	--

chr15	83003956	chr15	82934463	--
chr15	83014625	chr15	82936703	++
chr15	84960511	chr15	84859997	++
chr16	14988609	chr16	15031370	--
chr16	70009791	chr16	74426101	-+
chr16	86453332	chr12	132926283	+-
chr17	36297237	chr17	36337601	+-
chr20	1600172	chr20	1559471	-+
chr20	56771729	chr12	2828954	--
chr20	56771963	chr12	2830313	++
chr22	18666166	chr22	18737108	--
chr22	18737933	chr22	18686213	++

Supplemental Table 5

SVs on chrX detected by NAIBR on the *F8* inversion sample.

Chrom 1	Position 1	Chrom 2	Position 2	Orientation	Distance to int22h-1	Distance to int22h-3
chrX	1399792	chrX	1400312	+-	-152715693	-153287114
chrX	26179787	chrX	26212951	++	-127935698	-128474475
chrX	49162006	chrX	49180528	++	-104953479	-105506898
chrX	49208629	chrX	49209218	+-	-104906856	-105478208
chrX	49218203	chrX	49218751	+-	-104897282	-105468675
chrX	52830446	chrX	52830826	+-	-101285039	-101856600
chrX	57147470	chrX	57162963	++	-96968015	-97524463
chrX	129651865	chrX	129652250	+-	-24463620	-25035176
chrX	140140043	chrX	140140427	+-	-13975442	-14546999
chrX	154387911	chrX	154430070	--	272426	-257356
chrX	155245001	chrX	155245507	+-	1129516	558081
chrX	155251108	chrX	155252593	+-	1135623	565167

Supplemental Table 6

SVs on chr17 detected by NAIBR on the *NF1* sample.

Chrom 1	Position 1	Chrom 2	Position 2	Orientation	Distance to breakpoint 1	Distance to breakpoint 2
chr17	10359246	chr17	10409257	--	-19325000	-19413270
chr17	18296487	chr17	18296856	+-	-11387759	-11525671
chr17	18314761	chr17	18315123	+-	-11369485	-11507404
chr17	18327938	chr17	18328330	+-	-11356308	-11494197
chr17	18343774	chr17	18344042	+-	-11340472	-11478485
chr17	18390522	chr17	18391011	+-	-11293724	-11431516
chr17	27963239	chr17	27964080	+-	-1721007	-1858447
chr17	34502325	chr17	34502718	+-	4818079	4680191
chr17	39241002	chr17	39296331	-+	9556756	9473804
chr17	39241014	chr17	39274415	--	9556768	9451888
chr17	39254370	chr17	39279917	-+	9570124	9457390
chr17	39382979	chr17	39411518	-+	9698733	9588991
chr17	39421637	chr17	39427775	+-	9737391	9605248
chr17	39421677	chr17	39432260	+-	9737431	9609733
chr17	39502590	chr17	39521159	--	9818344	9698632
chr17	40653142	chr17	40697232	++	10968896	10874705
chr17	41008335	chr17	41026393	++	11324089	11203866
chr17	43616656	chr17	43616962	+-	13932410	13794435
chr17	58073089	chr17	58073319	+-	28388843	28250792
chr17	58078460	chr17	58078688	+-	28394214	28256161
chr17	60343141	chr17	60343406	+-	30658895	30520879
chr17	60359848	chr17	60360239	+-	30675602	30537712
chr17	62213073	chr17	62214273	+-	32528827	32391746
chr17	62885755	chr17	62886258	+-	33201509	33063731
chr17	64794832	chr17	64795449	+-	35110586	34972922
chr17	78287778	chr17	78289508	+-	48603532	48466981
chr17	80105167	chr17	80106015	+-	50420921	50283488

Supplementary Notes

Supplementary Note 1: Explanation of the model for detecting type 2 evidence.

Barcode similarity between two nearby regions very high because the reads originate from almost the same set of HMW DNA molecules. However, the barcode similarity between the left side and right side of the breakpoint are dramatically reduced. We call this evidence as type 2 evidence. To detect type 2 evidence, LinkedSV uses two adjacent sliding windows (window 1 and window 2) to scan the genome and calculate the barcode similarity between the window 1 and window 2.

In WES data sets, the numbers of reads in the sliding windows vary a lot due to capture bias and the length of capture regions. To detect type 2 evidence from both WGS and WES data sets, our model considers the variation of sequencing depth and capture regions. The barcode similarity is calculated as:

$$S = \frac{x}{m_1^a m_2^b} n e^{-\alpha d} \quad (1),$$

where:

m_1 is the number of barcodes in window 1,

m_2 is the number of barcodes in window 2,

x is the number of barcodes in both windows,

d is the weight distance between reads of the left window and the right window,

n is a constant representing the characteristic of the library,

α is a parameter of fragment length distribution,

a and b are two parameters between 0 and 1,

n, α, *a* and *b* are estimated from the data using regression.

Suppose there are *n* different HMW DNA molecules span both window 1 and window 2, each of which has a different barcode and generates a number of read pairs in the library. The read pairs in the library may or may not be sequenced. We assume the *n* HMW DNA molecules have the same rate to generate read pairs in the library, so the *n* HMW DNA molecules have the same chance to be sequenced (have at least 1 read). Let m_1 be the number of HMW DNA molecules sequenced in window 1, m_2 be the number of HMW DNA molecules sequenced in window 2. m_1 and m_2 can be different due to the bias of target enrichment and the total length of target regions in each window. Let *X* be the number of HMW DNA molecules sequenced in both window 1 and window 2. *X* follows the hypergeometric distribution:

$$P(X = x \mid m_1, m_2, n) = \frac{C_{m_1}^{x} C_{n-m_1}^{m_2-x}}{C_n^{m_2}} \quad (2).$$

The expectation of X is:

$E(X) = \frac{m_1 m_2}{n}$ (3).

However, the length of sliding windows may be as long as 40 kb and not all the *n* HMW DNA molecules are long enough to span both windows. In addition, the capture regions in window 1 and window 2 may be close to each other or far away from each other. Therefore, we need to adjust *n* to be approximately $ne^{-\alpha d}$. *d* is calculated using the following equation:

$d = w_2 - w_1$ (4),

where w_1 is the mean mapping position of all reads in window 1 and w_2 is the mean mapping position of all reads in window 2. The larger *d*, the smaller number of HMW DNA molecules can span a region of length *d*. We choose exponential distribution because the length of HMW DNA molecules follows exponential distribution and thus the number of HMW DNA molecules longer than *d* also follows exponential distribution.

m_1 also need to be adjusted because not all HMW DNA molecules being sequenced in window 1 span both windows. We adjust m_1 to be approximately m_1^a and similarly adjust m_2 to be approximately m_2^b.

After the adjustment, the expectation of *X* is:

$E(X) = \frac{m_1^a m_2^b}{ne^{-\alpha d}}$ (5).

We define barcode similarity as:

$$S = \frac{x}{E(X)} = \frac{x}{m_1^a m_2^b} n e^{-\alpha d} \quad (6),$$

where x is the number of shared barcodes between window 1 and 2, $E(X)$ is the expected number of shared barcodes between window 1 and 2.

Take the log of both sides equation (6), we have:

$$\log(S) = \log(x) - a log(m_1) - b log(m_2) + \log(n) - \alpha d \quad (7).$$

Assuming most regions in the genome do not have breakpoints, we can replace S with 1 and estimate a, b, n, α from the data using linear regression.