DeBreak: Deciphering the exact breakpoints of structural variations using long sequencing reads

Yu Chen

University of Alabama at Birmingham <https://orcid.org/0000-0002-2037-7337>

Amy Wang

University of Alabama at Birmingham <https://orcid.org/0000-0003-0478-2313>

Courtney Barkley

University of Alabama at Birmingham

Xinyang Zhao

University of Alabama at Birmingham

Min Gao

University of Alabama at Birmingham

Micky Edmonds

University of Alabama at Birmingham

Zechen Chong (✉ zchong@uab.edu)

University of Alabama at Birmingham <https://orcid.org/0000-0001-5750-1808>

Methods Article

Keywords:

Posted Date: January 21st, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1261915/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

DeBreak: Deciphering the exact breakpoints of structural variations using long sequencing reads

Yu Chen^{1,2}, Amy Y. Wang^{2,3}, Courtney A. Barkley¹, Xinyang Zhao⁴, Min Gao^{2,5}, Mick D. Edmonds¹, and Zechen Chong^{1,2,*}

¹Department of Genetics, Heersink School of Medicine, University of Alabama at Birmingham, Birmingham, AL 35294, USA. ²Informatics Institute, Heersink School of Medicine, University of Alabama at Birmingham, Birmingham, AL 35294, USA. ³Department of Medicine, Division of General Internal Medicine, Heersink School of Medicine, University of Alabama at Birmingham, Birmingham, AL 35294, USA. ⁴Department of Biochemistry and Molecular Genetics, Heersink School of Medicine, University of Alabama at Birmingham, Birmingham, AL 35294. ⁵Department of Medicine, Division of Cardiovascular Disease, Heersink School of Medicine, University of Alabama at Birmingham, AL 35233 Birmingham, USA.

* Correspondence: zchong@uabmc.edu

Abstract

Long-read sequencing has demonstrated great potential for characterizing all types of structural variations (SVs). However, existing algorithms have insufficient sensitivity and precision. To address these limitations, we present DeBreak, a novel method for comprehensive and accurate SV discovery. Based on alignment results, DeBreak employs a density-based approach for clustering SV candidates together with a local *de novo* assembly approach for reconstructing long insertions. A partial order alignment algorithm ensures precise SV breakpoints with single base-pair resolution, and a *k*-means clustering method can report multi-allelic SV events. DeBreak outperforms existing tools on both simulated and real long-read sequencing data from both PacBio and Nanopore platforms. An important application of DeBreak is analyzing cancer genomes for potentially tumor-driving SVs. DeBreak also demonstrates excellent performance in supplementing whole-genome assembly methods.

Introduction

In humans, structural variations (SVs), or genomic rearrangements, including deletions, insertions, inversions, duplications, translocations, and complex forms of multiple events,

compose the majority of genomic variations. At nucleotide level, SVs contribute more diversity than other types of variants¹⁻³ and play a pathogenic role in a wide range of genomic disorders⁴⁻⁸. In non-human organisms, SVs are also associated with various phenotypes⁹⁻¹¹. Therefore, comprehensive characterization of all forms of SVs is critical for fully understanding their contribution to genetic diversity, species divergence, and other phenotypic traits.

The currently available real-time long-read sequencing platforms, Pacific BioSciences (PacBio) and Oxford Nanopore, generate very long reads (>20 kbp on average) and have demonstrated superior performance over short reads on SV discovery. For example, many rare genetic diseases have been solved using long-read sequencing technologies¹²⁻¹⁵. Long-read sequencing can potentially delineate the full landscape of SVs in individual genomes. By sequencing and analyzing a haploid human genome (CHM1) using single-molecule, real-time (SMRT) DNA sequencing, Chaisson et al.¹⁶ resolved the complete sequence of 26,079 euchromatic structural variations, which is a 6-fold increase when compared with a previous study² using short read sequencing, and most of these SVs had not been reported previously. Recent work by the Human Genome Structural Variation Consortium (HGSVC) resulted in a 7-fold increase in SV number using multiple platforms in which the PacBio results contributed the most³.

Although great strides have been made, existing computational tools for SV detection using long reads remain few in number and can be further enhanced and optimized. These methods usually can only characterize a subset of SVs, and sensitivity and precision are not ideal. For example, PBHoney¹⁷ uses BLASR¹⁸ to map the PacBio subreads to collect soft-clipped reads and remap clipped tails (>200bp) to compose a “piece alignment”. However, PBHoney can only infer simple deletions, tandem duplications, inversions, and translocations and does not perform well for insertions and other types of SVs. Another alignment-based method, Sniffles¹⁹ uses both within-read alignments and split-read alignments from NGMLR aligner. Sniffles can analyze both PacBio and Nanopore sequencing data and report some

60 complex forms of SVs besides simple SVs. For both PBHoney and Sniffles, the breakpoints
61 inferred from clusters of alignments usually are not precise, preventing experimental validation
62 and mechanism analysis²⁰⁻²³ that rely on SV junction sequences. Moreover, both tools are
63 deficient in detecting the full spectrum of SVs. For instance, long insertions close to or longer
64 than nearby read lengths are often missed. These issues remain a concern with a recently
65 published alignment-based method, CuteSV²⁴.

66 Besides alignment-based methods, methods using local *de novo* assembly of mapped
67 reads are also applied to SV discovery^{3, 16, 25}. These local assembly-based methods utilize reads
68 from haploid genomes or phased reads from diploid genomes and perform *de novo* assembly in
69 a window. The consensus sequences generated can usually identify precise breakpoints.
70 However, they only report deletions, insertions, and subsets of inversions. During phasing, only
71 about two-thirds of reads in each sample can be haplotype-partitioned. These partitions require
72 different tools applied to the data from other platforms, preventing more generic and broader
73 applications of SV analysis.

74 Whole genome *de novo* assembly can be considered the ultimate solution for SV
75 characterization. Although researchers have made progress on this front²⁶⁻²⁹, whole genome *de*
76 *nov*o assembly with long noisy reads is inherently challenging. It requires high-coverage
77 sequencing data, and difficulties remain in dealing with long repetitive sequences, tandem
78 repeats, as well as heterozygosity. Moreover, whole genome *de novo* assembly usually requires
79 high-memory computing nodes and long running time, and it is difficult to evaluate accuracy.

80 Here, we present DeBreak, a novel algorithm for comprehensive and accurate SV
81 discovery from long reads. DeBreak detects SV events using two different strategies, depending
82 on whether SVs can be spanned within reads (**Fig. 1**, Methods). For SVs contained within
83 reads, DeBreak scans all read alignments for raw SV signals for each category of SV, and then
84 clusters these signals using a unique density-based clustering algorithm with flexible clustering
85 window sizes. This approach allows accurate SV candidate identification for SVs with varying

lengths and local sequence contents. In the next step, the SV breakpoint refinement with partial order alignment (POA)³⁰ algorithm can accurately infer SV breakpoints with single base-pair resolution. With its automatic sequencing-depth estimation and parameter optimization, DeBreak filters SV candidates and reports a high-confidence SV callset with genotyping information. For SVs that are too large to be spanned within reads, DeBreak first identifies candidate SV breakpoints and then performs local *de novo* assembly to reconstruct SV containing sequences. DeBreak completes the analysis by integrating all identified SV events together to form a final, confident SV callset.

Results

Benchmark on simulated dataset

To benchmark the performance of SV discovery, we first compared DeBreak with three SV callers, Sniffles, pbsv, and cuteSV, using *in silico* datasets. A total of 22,200 SVs were simulated and embedded into the human reference genome (GRCh38), serving as the ground truth. The sizes of simulated SVs follow similar distributions observed in real human samples, with Alu and LINE peaks³¹ (**Fig S1**). Long reads were simulated based on a modified genome with pbsim³² and aligned to the human reference genome. To mimic PacBio reads generated from different library preparation protocols, three sets of data with different insert sizes were simulated (**Fig S1b**). We applied these SV callers to identify SVs and compared the SV callsets to the ground truth to assess SV discovery accuracy for each SV caller. In all five types of simulated SVs, DeBreak consistently achieved the highest F1 scores among the four tested SV callers (**Table 1, Fig S2a**). All evaluated SV callers have a critical parameter—“minimum number of supporting reads”—which determines the sensitivity of SV detection for these tools.

We manually set the parameter of “minimum number of supporting reads” to a series of values for each caller and assessed the sensitivity of SV discovery at different thresholds.

Although the recall of all SV callers dropped as the number of supporting reads increased, DeBreak consistently demonstrated the highest sensitivity at each threshold (**Fig S2b**). We then investigated the accuracy of detection of SV breakpoints. By refining SV breakpoints with the partial order alignment algorithm, DeBreak reconstructed the consensus sequences flanking the SVs, which showed much higher base accuracy than the raw reads (**Fig. S3a**). From the accurate consensus sequences, DeBreak can infer more precise SV breakpoints than merely from the raw reads (**Fig S3b**). DeBreak identified 59.81% of SVs with exact breakpoint positions and 81.33% of SVs with ± 1 bp shift around the SV breakpoint. These results demonstrated that DeBreak can detect all five types of SVs in the simulated datasets with high accuracy.

Benchmark on real human genome

We next benchmarked the SV discovery accuracy on a real human genome, HG002, from the Genome in a Bottle (GIAB) Consortium³³. We aligned PacBio CLR, HiFi, and Nanopore reads of HG002 to the human reference genome and applied four SV callers, DeBreak, Sniffles, pbsv, and cuteSV, on the three datasets. The GIAB community genome provided an SV callset of 4,237 deletions and 5,440 insertions from multiple platforms in defined “high-confidence” regions³⁴. Thus, we first benchmarked the SV discovery accuracy in these high-confidence regions. In all three datasets, DeBreak achieved the highest SV discovery accuracy among the four tested SV callers, especially for insertions (**Table 2**). The higher SV discovery accuracy of DeBreak resulted from its advanced clustering algorithm, in which clustering window size is adjustable for SVs of different types, sizes, and local sequence content. Instead of setting a clustering window of fixed size, DeBreak computes the density of raw SV signals and determines the boundaries of the clustering window based on the density pattern. The clustering window is larger for longer SV events and smaller for shorter SV events, which improves effectiveness by merging raw SV signals into SV candidate while excluding noisy signals nearby

(**Fig S4**). For SVs located in repetitive regions, the clustering window is automatically adjusted to tolerate shifts of raw SV signals caused by repeated segments (**Fig S5**).

Previous work³⁵ has highlighted the functional importance of multi-allelic copy number variations (CNVs) in gene dosage and gene expression. DeBreak can accurately identify multi-allelic SVs (mSVs). After density-based clustering, raw SV signals of candidate mSVs are further clustered through a *k*-means clustering algorithm to characterize two non-reference alleles (**Fig S6**). We applied this method to identify putative mSVs in HG002. In total, we identified 802 multi-allelic SVs in a single genome.

We also benchmarked genotyping accuracy of the four tested SV callers with the high-confidence SV callset. On the three datasets, DeBreak and cuteSV performed better than pbsv and Sniffles (**Table S1**). DeBreak achieved the highest genotyping accuracy in PacBio CLR and Nanopore datasets, while cuteSV achieved slightly higher genotyping accuracy in the PacBio HiFi dataset.

We then assessed the SV discovery accuracy for SVs of different sizes. DeBreak achieved stable and high accuracy for small and large SVs, especially in detecting insertions (**Fig 2a**). Notably, for ultra-large insertions longer than the sequencing reads (>10kbp), DeBreak achieved higher accuracy, recall, and precision than the other three SV callers (**Fig S7**), benefiting from its large-insertion detection module with local *de novo* assembly. In the PacBio HiFi and Nanopore datasets, DeBreak also achieved relatively high accuracy for SVs of different sizes (**Fig S8**). We next evaluated the accuracy of SV breakpoint positions reported by DeBreak. The high sequencing error rate of long reads often causes imprecise inference of SV breakpoints. We compared SV callsets from the four tested SV callers to high-confidence benchmark SV callset to assess for shifts in breakpoint positions. With the breakpoint refinement module, DeBreak identified 59.90% of SVs that were consistent with exact SV breakpoints reported in GIAB, higher than pbsv (49.17%), Sniffles (4.99%), and cuteSV (5.18%)

using the PacBio CLR dataset (**Fig 2b**). For PacBio HiFi and Nanopore datasets, DeBreak also achieved the highest SV breakpoint accuracy among the four evaluated SV callers (**Fig S9**).

To assess the effect of sequencing depth on SV discovery accuracy, we downsampled the PacBio CLR, HiFi, and Nanopore datasets by a series of depths by randomly sampling the reads. At each depth, DeBreak and pbsv were applied with the default parameters. In contrast, a set of parameters were tested for Sniffles and cuteSV, and the SV calls with the highest F1 scores were selected for comparison. Overall, SV discovery was more accurate for datasets having higher sequencing depth (**Fig 2c**). Starting from 20x, DeBreak already achieved accuracy of over 90% for PacBio CLR, HiFi, and Nanopore datasets. For datasets with depth $\geq 20x$, DeBreak consistently identified SVs with the highest accuracy among the four tested SV callers. Note that DeBreak and pbsv automatically adapted to lower depths using default settings, while Sniffles and cuteSV both required extra effort in manually tuning parameters to optimize performance (**Fig S10**). Taken together, these results highlight that DeBreak can accurately identify different types of SVs with precise breakpoints in real human genomes.

Comparison to assembly-based SV discovery

Currently, *de novo* assembly is used to comprehensively characterize genome-wide SVs²⁶⁻²⁹. To compare alignment-based with assembly-based SV discovery approaches, we applied DeBreak, pbsv and cuteSV on six samples from the Human Genome Structural Variation Consortium (HGSVC). Three of these six samples were sequenced with the PacBio CLR platform, and the other three were sequenced with the PacBio HiFi platform. For these samples, highly accurate assembly-based SV callsets were generated by performing haplotype-resolved *de novo* assembly with phased sequencing reads and subsequent SV discovery from whole-genome assembly with the PAV pipeline²⁶. Overall, the assembly-based SV approach discovered a slightly higher number of SV events (22,897 – 27,187) compared with alignment-based methods. By treating these SVs as the “ground truth”, we evaluated the SV discovery

accuracy of three alignment-based SV callers. DeBreak identified SV with an average F1 score of 80.09% in the six samples, which was higher than pbsv (72.68%) and cuteSV (77.38%) (**Table S2**). In each sample, DeBreak achieved both higher recall and precision than pbsv and cuteSV (**Fig 3a**), suggesting higher consistency with the assembly-based SV discovery than the other two alignment-based SV callers. As assembly-based SV calls usually have accurate SV breakpoints inferred from assembled contigs, we also compared the SV breakpoint accuracy of three SV callers. With the breakpoint-refinement module, DeBreak identified 32,813 SVs with exact SV breakpoints on the three PacBio CLR datasets, while pbsv reported 27,345 and cuteSV reported 1,927 precise SV breakpoints, respectively (**Fig 3b**). For the three PacBio HiFi datasets, DeBreak also achieved better breakpoint accuracy than pbsv and cuteSV, as 41,691, 34,691, and 11,218 SVs were identified with exact SV breakpoints by DeBreak, pbsv, and cuteSV, respectively.

Approximately 82% of DeBreak SV calls overlapped with the assembly-based SV callset. There are several thousand unique SVs were reported either by DeBreak or by the assembly approach. To characterize these SVs, we performed a four-way comparison of SV callsets from DeBreak, pbsv, cuteSV, and PAV on the sample HG00096. For the SVs identified by PAV but not by DeBreak, 27.1% of deletions and 30.5% of insertions were reported by either pbsv or cuteSV (**Fig 3c, Fig S11**). In contrast, 71.9% of deletions and 71.5% of insertions reported by DeBreak but not PAV were also reported by either pbsv or cuteSV, suggesting that alignment-based SV callers identified SV calls more consistently. We further characterized the 3,385 SVs only reported by PAV and 1,292 SVs only reported by DeBreak in all six samples. Note that DeBreak reported the fewest number of unique SVs. By examining the SV locations on the genome, we found that there was strong enrichment in telomere regions for PAV-unique SVs (43.5% located near telomeres, 5.8% located near centromeres, and 46.4% located in repetitive regions) (**Fig 3d, Fig S12**). While DeBreak-unique SVs were enriched in the telomere and centromere regions (31.2% located near telomeres, 27.7% located near centromeres, and

38.7% located in repetitive regions). Although DeBreak controls read depth and minimum number of supporting reads, alignment-based SV discovery may have inaccurate read alignment in these regions. However, it is also challenging to assemble reads with abnormal coverage and ascertain phasing status of individual reads without bias. Additional efforts are needed to validate these SVs.

SV discovery in cancer genomes

SVs play important roles in cancer development and progression³⁶⁻³⁸. Unlike germline SVs, cancer genomes may contain more large-scale deletions, duplications, inversions, translocations, and other complex SVs³⁹⁻⁴¹. DeBreak includes a “tumor” mode to identify “abnormal” SVs and SVs with clustered breakpoints in cancer genomes. To assess SV discovery in cancer genomes, we applied the four SV callers to identify SVs in a breast cancer cell line, SKBR3. Under the “tumor” mode, DeBreak identified 8,249 deletions, 9,226 insertions, 3,129 duplications, 190 inversions, and 137 translocations. We compared the SV callsets from the four SV callers. As expected, a large proportion of the SVs were identified by all four SV callers (**Fig S13**). Among the four SV callers, DeBreak reported relatively fewer singleton SV calls, especially in insertion/duplication detection, suggesting high precision of DeBreak SV callsets. We also compared the SV callset of DeBreak to previously reported SV lists from long-read and short-read data⁴² (**Fig S14**). DeBreak reported 1,333 deletions, 3,073 insertion/duplications, 51 inversions, and 91 translocations that were not previously discovered. To validate potential cancer-related SVs reported only by DeBreak, we designed primers flanking breakpoints for 15 randomly selected SVs (6 deletions, 4 duplications, 3 inversions, and 2 translocations) that spanned more than 10kbp (**Fig S15**). Polymerase chain reaction (PCR) experiments validated 12 out of 15 DeBreak-novel SVs, with a validation rate of 80% (**Supp file 2**).

We further analyzed SVs in the SKBR3 breast cancer cell line by annotating breakpoints and identified 41 putative gene fusions. By cross-validating these gene fusion events with Iso-Seq data (Methods), we found 11 gene fusions that can be validated at the transcripts level (**Supp file 3**). 6 out of 11 gene fusions have been previously reported using transcriptomic data⁴²⁻⁴⁵. Therefore, SV discovery using DNA-seq data with DeBreak identified 5 novel gene fusions: WDR82-PBRM1, PDE4D-DEPDC1B, CPNE1-PHF20, CSE1L-KCNB1, and CSNK2A1-NCOA3. The fusion of WDR82 and PBRM1 was caused by a hemizygous deletion of 392kbp on chromosome 3, with the fusion junction located in the intronic region of both genes (**Fig S16**). A deletion of 259kbp on chromosome 20 caused the fusion of CSE1L and KCNB1, where the seventh exon of CSE1L was fused with the intron of KCNB1 (**Fig S17**). The gene fusion junction locations observed in the Iso-Seq reads were highly consistent with SV breakpoint positions inferred by DeBreak, suggesting that DeBreak can accurately predict SV breakpoint positions in cancer genomes. These results indicate that DeBreak can be applied to cancer genomes and identify previously unknown SVs.

Runtime and memory usage

DeBreak and other SV callers were tested on Intel Xeon E5-2680 v3 CPUs with 12 cores and 2.5GHz of frequency. It took 12.4 hours for DeBreak to identify SVs from a human genome (SKBR3 cell line) using the 67x PacBio CLR dataset with peak memory of 63 GB (**Table S3**). Due to the local assembly module and partial order alignments, DeBreak consumed more runtime and memory than Sniffles (3.0h, 13GB) and cuteSV (1.5h, 3GB). However, DeBreak was much faster and consumed less memory than pbsv (45.1h, 72GB).

Discussion

In this work we present DeBreak, a method for efficient and accurate structural variation detection from long-read sequencing data. Based on simulation data, real human genome data, and cancer cell line data, DeBreak has demonstrated excellent performance when compared with several state-of-the-art long-read SV callers. The improved performance is due to several innovative design features: 1) the density-based clustering method can accurately identify candidate SV events with a variety of sizes; 2) the partial order alignments can produce a consensus sequence for accurate breakpoint inference, which is helpful for experimental validation and mechanism inference²⁰⁻²³; 3) local *de novo* assembly facilitates discovery of long insertion events, which usually cannot be inferred within individual reads; 4) *k*-means approach can accurately identify multi-allelic SVs, which are functionally important; and 5) multiple functions can be applied to both healthy and unhealthy genomes.

Although the benchmark was based on human genomes, DeBreak can be applied to other diploid or haploid non-human long-read resequencing data. The overall workflow may be applied to polyploid genomes as well. Based on our knowledge and experience, SV discovery in polyploid genomes is challenging for any currently available tools. More sophisticated benchmarking work is needed.

We observed that SV callers reporting more accurate breakpoint positions (DeBreak and pbsv) required more computational resource than SV callers with less accurate breakpoints (Sniffles and cuteSV). During SV discovery for DeBreak, breakpoint refinement and ultra-large insertion detection were the two most time-consuming steps, accounting for approximately 45% and 32% of total runtime, respectively. When we disabled these two features, DeBreak accomplished SV detection within 2.8 hours for the same sample, similar to the runtime of Sniffles and cuteSV. The extra runtime and memory usage helped improve the quality and accuracy of the DeBreak SV callset. No matter what situation, DeBreak and other alignment-based methods consume much less computational resources. Although comprehensive

evaluation and validation between alignment-based and assembly-based approaches are needed, alignment-based methods will continue to serve important roles for SV analysis.

Methods

1. DeBreak workflow

1.1 Overall workflow of DeBreak

DeBreak detects SVs from read-to-reference alignments generated by any long-read aligner, such as minimap2, pbmm2, and ngmlr. The workflow of DeBreak includes 1) raw SV signal detection, 2) large insertion identification, 3) SV signal clustering, 4) SV breakpoint refinement, and 5) SV filtering and genotyping. The output of DeBreak is a standard VCF file containing confident SV calls.

1.2 Raw SV signal detection and clustering

Raw SV signals are detected from read-to-contig alignment. DeBreak scans all read alignments for intra-alignment and inter-alignment SV signals. Smaller indels can be contained within a single alignment. For larger indels, inversions, duplications, and translocations, DeBreak utilizes split-read information and classifies SV type based on orientation and clipping location of two segments from the same read. Once a raw SV signal meets criteria, DeBreak records features of the signal and the read supporting it, including SV type, SV coordinate, SV size, read name, and read mapping quality. As it scans through read alignments, DeBreak also estimates sequencing depth of the input dataset and automatically adjusts parameters used in the following clustering and filtering processes.

Raw SV signals are then clustered into SV candidates with a density-based clustering algorithm. All signals from the same chromosome are sorted based on coordinates. The density of SV raw signals is computed for each position on the chromosome (supp fig). DeBreak scans the chromosome for peaks above a predefined threshold. The boundary of the SV region is

defined on both side of the summit when density drops to 10% of the summit or lower than the predefined threshold. All raw signals located within one SV region are then merged into one SV candidate.

For each SV candidate, DeBreak determines whether it is a multi-allele SV based on the first quartile (Q1) and third quartile (Q3) of SV size from all raw signals. If Q3 is smaller than twice of Q1, all raw signals are merged into a single-allele SV, excluding outliers of extremely large or small size. If Q3 is larger than twice Q1, DeBreak separates raw SV signals for each allele with k -means clustering ($k=2$ for diploid genomes) and merges signals from each cluster separately as a multi-allele SV candidate. The detection and clustering of SV signals are processed separately for each chromosome, allowing DeBreak to perform multi-thread SV detection, drastically reducing runtime.

1.3 SV breakpoint refinement

After SV signal clustering, DeBreak assigns each SV candidate a breakpoint coordinate by computing the mean value of raw signals. Raw signals can be highly imprecise due to the high error rate of long-read sequencing and the presence of low-complexity regions in the genome. DeBreak implants the POA algorithm from wtdbg2⁴⁶ to refine breakpoint locations. For each SV candidate, DeBreak collects all reads containing this SV candidate and performs POA to generate accurate consensus sequences. DeBreak then aligns these consensus sequences to the reference genome with minimap2 and detects SVs from consensus sequence alignments. The breakpoint location detected from consensus sequence is used to refine the breakpoint coordinates of SV candidates. If POA fails to generate consensus sequences for an SV candidate, or the consensus sequence cannot be properly aligned back to the genome, DeBreak will keep the mean value of the raw signals as breakpoint coordinates.

1.4 Depth-based filtering and genotyping

During raw SV signal detection, DeBreak records the total length of aligned reads on each chromosome and computes the average sequencing depth. The minimum threshold of supporting reads (N_{supp}) is determined based on the average sequencing depth: $N_{supp} = \frac{Depth}{10} + 2$. SV candidates supported by at least N_{supp} reads are kept for further consideration, and the rest are discarded to remove background noise. SVs of low mapping quality are also filtered to remove false positives caused by inaccurate read alignment. For multi-allele SVs, DeBreak filters each allele independently. If only one allele passes, a single-allele SV will be reported instead. SVs are genotyped based on the ratio of SV supporting reads to the local sequencing depth at each SV location.

1.5 Large insertion detection via local assembly

DeBreak utilizes a local *de novo* assembly approach to detect ultra-large insertions that are too long to be spanned within a single read. While scanning read alignments for raw SV signals, DeBreak also records positions of clipped end of read alignments. After scanning through a chromosome, DeBreak identifies candidate insertion breakpoint regions with enriched clipped alignment. It collects all the reads clipped at each candidate breakpoint region and performs local *de novo* assembly with these reads to reconstruct the inserted sequence. DeBreak aligns assembled contigs to the reference genome with minimap2 and detects insertions from these contigs. Detected insertions are filtered out if 1) multiple contigs are assembled during local *de novo* assembly, 2) a detected insertion is located in another chromosome or too far away from the candidate insertion breakpoint, or 3) the detected insertion is smaller than 1kbp.

1.6 Duplication identification

DeBreak includes an optional duplication-rescuing module that distinguishes tandem duplications from insertion calls, as smaller tandem duplications are often treated as insertions by aligners. The inserted sequence of tandem duplication shows high similarity with the duplicated region, while insertions usually consist of novel sequence or sequences from distinct regions of the genome. For each insertion call, DeBreak collects reads supporting the SV event and extracts inserted sequence from each read. It utilizes minimap2 to re-align these inserted sequences back to the local region (1kbp flanking the insertion breakpoint) on the reference genome. If more than 50% of inserted sequences can be aligned back to the local region, DeBreak corrects the SV type to tandem duplication for this insertion call.

2. Benchmark in simulated dataset

2.1 Simulated dataset generation

Three simulated datasets with ground-truth SVs were generated for benchmarking. For each dataset, a total of 22,200 SVs (10,000 deletions, 10,000 insertions, 1,000 duplications, 1,000 inversions and 200 translocations) were randomly simulated on Chr1 to Chr22 and ChrX. The sizes of simulated SVs followed the geometric distribution as observed in real human genomes, including peaks at ~350bp and ~6000bp. These simulated SVs were assigned as heterozygotes and homozygotes with a ratio of 2:1, and heterozygous SVs were randomly assigned to two haplotypes. The human reference genome GRCh38 (autosomes and the X chromosome) were modified according to the type and size of simulated SVs to generate haplotype 1 and haplotype 2. Long reads were simulated from the modified genome using pbsim with options “--data-type CLR --model_qc model_qc_clr --depth 25 --accuracy-mean 0.85”. The depth was set to 25X for each haplotype, generating a simulated dataset of 50X when merging all reads from both haplotypes. The average read length was set to 10kbp, 15kbp, and 20kbp for three simulated datasets.

2.2 SV discovery in simulated datasets

The simulated reads were aligned to the reference genome with minimap2, ngmlr, and pbmm2 under default settings. DeBreak was applied to minimap2 alignment results with default settings. pbsv was run on pbmm2 alignment results with default settings, and Sniffles was run on the ngmlr alignment results with options “--genotype -s 4/5/6/7/8/9/10”. A series of -s (minimum number of reads supporting an SV) was tested for Sniffles, and the threshold with best accuracy was selected for comparison with other SV callers. cuteSV was run on minimap2 alignment results with options “--genotype”. All SVs with length ≥ 45 bp were selected for benchmark.

The SV callsets of DeBreak, Sniffles, pbsv, and cuteSV were compared to the ground-truth SV set to assess the recall, precision, and F1 score. An SV call (DEL, INS, DUP, and INV) is considered as true positive (TP) if all three conditions are met:

- 1) $Type_G = Type_C$
- 2) $ABS(Cor_G - Cor_C) \leq 1\text{ kbp}$
- 3) $0.5 * Size_G \leq Size_C \leq 2 * Size_G$

Where the $Type_G$, $Type_C$, Cor_G , Cor_C , $Size_G$, and $Size_C$ are the SV type, start coordinates and size of the ground truth SV call and the candidate SV call. For translocations (TRAs), the coordinates of both breakpoints on two chromosomes should be within 1kbp flanking the ground-truth breakpoints to be determined as TP.

3. Benchmark in HG002 dataset

3.1 SV discovery accuracy benchmark

Raw sequencing reads (PacBio CLR, HiFi, and Nanopore data) were downloaded and aligned to GRCh37 with minimap2⁴⁷, ngmlr, and pbmm2 with default settings. DeBreak and cuteSV were applied to minimap2 alignment with default settings. pbsv was applied to pbmm2 alignment with default settings. Sniffles was applied to ngmlr alignment with option “--genotype -

s 9/9/12” for PacBio CLR, HiFi, and Nanopore dataset, respectively. A series of minimal supporting read (-s option) were tested for Sniffles, and the callset with best performance was used for evaluation. SV callsets of four SV callers were benchmarked within the high-confidence regions (HG002_SVs_Tier1_v0.6.bed) by comparing to the benchmark SV callset using the same criteria as in simulation benchmark. Shift of SV breakpoint was also evaluated with the high-confidence SV benchmark callset. The SV coordinates in DeBreak and cuteSV callsets are 1-based, so all the breakpoint positions were transformed to 0-based to keep consistent with the benchmark callset. Genotyping accuracy of four SV callers was evaluated based on the genotype information in the benchmark callset.

3.2 Down-sampling

To evaluate SV callers at varying sequencing depths, we downsampled the PacBio CLR dataset of HG002 to a series of depth from 10X to 70X. Sequencing reads were randomly selected to generate datasets with desired depth. The depth of each down-sampled dataset was validated by the total number of bases in reads divided by human genome size (3.1Gbp). Four SV callers were first applied to downsampled datasets with default settings. In addition, to achieve the best performance of Sniffles and cuteSV, a series of min_supp (-s option) was provided to Sniffles and cuteSV at each depth, and the SV callset with the highest accuracy was selected for comparison.

4. Comparison with assembly-based SV callset

Raw PacBio CLR or HiFi reads of sample HG00096, HG01505, HG01596, HG02818, HG03486, and NA12878 were downloaded and aligned to GRCh38 with minimap2 and pbmm2 under default settings. DeBreak, pbsv and cuteSV were applied to the alignment files to identify SVs with default settings. The merged assembly-based SV callset was downloaded from HGSC2 data portal, and SVs of each sample were extracted with custom script. The comparison of SV

calls was performed for autosomes and the X chromosome. SVs located within 5Mbp of both ends of the chromosomes were classified as 'near telomere'. SVs located within 5Mbp of centromere were classified as 'near centromere'. Remaining SVs were annotated according to the repeat annotation from Table Browser. SV Distribution on the genome was plotted with karyoploteR⁴⁸.

5. SV validation in SKBR3 cell line

5.1 PCR validation of novel SVs

PCR validation was performed for SVs identified by DeBreak that were not reported previously by Sniffles and short-read SV callers⁴². Fifteen putative cancer-related SVs were randomly selected from SVs spanning more than 10kbp on the genome. Insertions were not validated due to length limitations of PCR. PCR primers were designed for each type of SV with Primer3 (v0.4.0)⁴⁹, and the specificity was verified with UCSC in-silico PCR (**Fig S15**). An SV event was validated if the PCR and following gel electrophoresis confirmed PCR product of the predicted size.

5.2 Gene fusion annotation and validation with Iso-Seq data

PacBio CLR sequencing data of SKBR3 was aligned to GRCh38 for SV discovery with DeBreak. Breakpoints of deletions, duplications, inversions, and translocations were annotated based on the Ensembl GRCh38 annotation (v104). An SV was considered to cause gene fusion when its two breakpoints were located within two different genes. Iso-Seq reads were downloaded from NCBI and aligned to GRCh38. For each gene fusion event, the total number of Iso-Seq reads aligned to both genes were counted. Gene fusion events supported by at least 3 Iso-Seq reads were considered as validated.

Data availability

PacBio CLR, HiFi, and Nanopore HG002 sequences were downloaded from GIAB at https://github.com/genome-in-a-bottle/giab_data_indexes, where PacBio 70x (CLR), PacBio CCS 15kb_20kb chemistry2 (HiFi), and Oxford Nanopore ultralong were used for SV discovery. The Tier1 benchmark SV callset and high-confidence HG002 region were obtained from https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/NIST_SVs_Integration_v0.6/. Sequencing reads and assembly-based SV callsets of HG00096, HG01505, HG01596, HG02818, HG03486, and NA12878 were downloaded from the HGSC2 data portal at <https://www.internationalgenome.org/data-portal/data-collection/hgsvc2>. The PacBio CLR and IsoSeq data of SKBR3 cell line were downloaded from NCBI SRA under BioProject PRJNA476239.

Code availability

DeBreak is publicly available at <https://github.com/Maggi-Chen/DeBreak> under the MIT License. We used v1.0.2 version for SV discovery and benchmark presented in the manuscript.

Acknowledgements

This work was supported by grant from National Institute of General Medical Sciences (1R35GM138212); the BioData Catalyst Fellowship from National Heart, Lung, and Blood Institute (a subaward from 1OT3HL147154) to Z.C.; and the Center for Clinical and Translational Science grant from the National Center for Advancing Translational Sciences (UL1TR003096) to A.Y.W.

488 **Author contributions**

489 Z.C. conceived and managed the project. Y.C. implemented the algorithm, collected all
490 datasets, and performed primary data analysis. Z.C. and M.G. were involved in data analysis
491 and testing of the algorithm, and Z.C., A.Y.M, and M.G. were involved in interpretation of the
492 results. Y.C., Z.C., and A.Y.W. wrote the manuscript. X.Z., C.A.B, and M.D.E. performed
493 experimental validation. All authors have read and approved the final manuscript.

References

1. Redon, R. et al. Global variation in copy number in the human genome. *Nature* **444**, 444-454 (2006).
2. Sudmant, P.H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75-81 (2015).
3. Chaisson, M.J.P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* **10**, 1784 (2019).
4. Stankiewicz, P. & Lupski, J.R. Structural variation in the human genome and its role in disease. *Annu Rev Med* **61**, 437-455 (2010).
5. 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 (2013).
6. Tanzi, R.E., Bird, E.D., Latt, S.A. & Neve, R.L. The amyloid beta protein gene is not duplicated in brains from patients with Alzheimer's disease. *Science* **238**, 666-669 (1987).
7. Chartier-Harlin, M.C. et al. Alpha-synuclein locus duplication as a cause of familial Parkinson's disease. *Lancet* **364**, 1167-1169 (2004).
8. Yang, L. et al. Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919-929 (2013).
9. Alonge, M. et al. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell* **182**, 145-161 e123 (2020).
10. Wang, W. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43-49 (2018).
11. Weissensteiner, M.H. et al. Discovery and population genomics of structural variation in a songbird genus. *Nat Commun* **11**, 3403 (2020).
12. Merker, J.D. et al. Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med* **20**, 159-163 (2018).
13. Mizuguchi, T. et al. A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. *J Hum Genet* **64**, 359-368 (2019).
14. Sanchis-Juan, A. et al. Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med* **10**, 95 (2018).
15. Sone, J. et al. Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. *Nat Genet* **51**, 1215-1221 (2019).
16. Chaisson, M.J. et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608-611 (2015).
17. English, A.C., Salerno, W.J. & Reid, J.G. PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics* **15**, 180 (2014).
18. Chaisson, M.J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
19. Sedlazeck, F.J. et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* **15**, 461-468 (2018).
20. Liu, P., Carvalho, C.M., Hastings, P.J. & Lupski, J.R. Mechanisms for recurrent and complex human genomic rearrangements. *Curr Opin Genet Dev* **22**, 211-220 (2012).
21. Dittwald, P. et al. NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. *Genome Res* **23**, 1395-1409 (2013).
22. Beck, C.R. et al. Megabase Length Hypermutation Accompanies Human Structural Variation at 17p11.2. *Cell* **176**, 1310-1324 e1310 (2019).

541 23. Carvalho, C.M. & Lupski, J.R. Mechanisms underlying structural variant formation in genomic
542 disorders. *Nat Rev Genet* **17**, 224-238 (2016).

543 24. Jiang, T. et al. Long-read-based human genomic structural variation detection with cuteSV.
544 *Genome Biol* **21**, 189 (2020).

545 25. Huddleston, J. et al. Discovery and genotyping of structural variation from long-read haploid
546 genome sequence data. *Genome Res* **27**, 677-685 (2017).

547 26. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural
548 variation. *Science* **372** (2021).

549 27. Garg, S. et al. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat*
550 *Biotechnol* **39**, 309-312 (2021).

551 28. Shafin, K. et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly
552 of eleven human genomes. *Nat Biotechnol* **38**, 1044-1053 (2020).

553 29. Wenger, A.M. et al. Accurate circular consensus long-read sequencing improves variant
554 detection and assembly of a human genome. *Nat Biotechnol* **37**, 1155-1162 (2019).

555 30. Lee, C., Grasso, C. & Sharlow, M.F. Multiple sequence alignment using partial order graphs.
556 *Bioinformatics* **18**, 452-464 (2002).

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

559 32. Ono, Y., Asai, K. & Hamada, M. PBSIM: PacBio reads simulator--toward accurate genome
560 assembly. *Bioinformatics* **29**, 119-121 (2013).

561 33. Zook, J.M. et al. An open resource for accurately benchmarking small variant and reference calls.
562 *Nat Biotechnol* **37**, 561-566 (2019).

563 34. Zook, J.M. et al. A robust benchmark for detection of germline large deletions and insertions.
564 *Nat Biotechnol* **38**, 1347-1355 (2020).

565 35. Handsaker, R.E. et al. Large multiallelic copy number variations in humans. *Nat Genet* **47**, 296-
566 303 (2015).

567 36. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112-
568 121 (2020).

569 37. Petljak, M. et al. Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals
570 Episodic APOBEC Mutagenesis. *Cell* **176**, 1282-1294 e1220 (2019).

571 38. Berger, M.F. et al. The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-
572 220 (2011).

573 39. Stephens, P.J. et al. Massive genomic rearrangement acquired in a single catastrophic event
574 during cancer development. *Cell* **144**, 27-40 (2011).

575 40. Cortes-Ciriano, I. et al. Publisher Correction: Comprehensive analysis of chromothripsis in 2,658
576 human cancers using whole-genome sequencing. *Nat Genet* (2020).

577 41. Baca, S.C. et al. Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666-677 (2013).

578 42. Nattestad, M. et al. Complex rearrangements and oncogene amplifications revealed by long-
579 read DNA and RNA sequencing of a breast cancer cell line. *Genome Res* **28**, 1126-1135 (2018).

580 43. Kim, D. & Salzberg, S.L. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.
581 *Genome Biol* **12**, R72 (2011).

582 44. Edgren, H. et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing.
583 *Genome Biol* **12**, R6 (2011).

584 45. Chen, K. et al. BreakTrans: uncovering the genomic architecture of gene fusions. *Genome Biol*
585 **14**, R87 (2013).

586 46. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* **17**, 155-158
587 (2020).

- 588 47. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100
589 (2018).
590 48. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes
591 displaying arbitrary data. *Bioinformatics* **33**, 3088-3090 (2017).
592 49. Untergasser, A. et al. Primer3--new capabilities and interfaces. *Nucleic Acids Res* **40**, e115
593 (2012).
594

Figure 1 Workflow of DeBreak. The major steps of DeBreak SV discovery include SV signal detection, signal clustering, breakpoint refinement, and filtering and genotyping. Detailed descriptions of each step can be found in Methods.

Table 1 SV discovery accuracy on simulated datasets

Type	DeBreak			Sniffles			pbsv			cuteSV		
	Rec	Pre	F1	Rec	Pre	F1	Rec	Pre	F1	Rec	Pre	F1
DEL	99.59	99.50	99.54	95.50	99.83	97.62	97.21	99.58	98.38	96.94	99.82	98.36
INS	98.51	99.65	99.08	92.35	99.52	95.80	95.38	96.22	95.80	94.58	94.71	94.64
DUP	98.27	97.04	97.65	90.80	98.63	94.55	44.53	97.11	60.97	44.13	99.03	61.05
INV	99.10	99.40	99.25	94.57	97.22	95.88	82.67	99.92	90.47	96.53	50.00	65.88
TRA	99.17	99.67	99.41	97.67	95.60	96.62	97.67	25.04	39.86	99.00	50.30	66.70
Total	99.02	99.45	99.23	93.85	99.47	96.58	93.36	95.34	94.34	93.50	92.39	92.94

The unit for recall, precision, and F1 score is %. The highest recall, precision, and F1 score among four tested SV callers are marked in bold.

Rec = Recall. Pre = Precision. F1 = F1 score.

Table 2 SV discovery accuracy on HG002

	DeBreak			Sniffles			pbsv			cuteSV		
	Rec	Pre	F1	Rec	Pre	F1	Rec	Pre	F1	Rec	Pre	F1
Deletion												
CLR	97.73	96.48	97.10	95.14	96.69	95.91	95.28	96.21	95.74	97.31	94.18	95.72
HiFi	98.16	95.11	96.61	97.45	91.67	94.47	96.74	94.88	95.80	97.71	93.33	95.47
Nano	98.40	95.07	96.71	96.29	94.62	95.45	97.40	81.08	88.50	98.18	89.27	93.51
Insertion												
CLR	97.15	93.36	95.22	88.38	89.58	88.98	93.22	83.43	88.05	95.40	81.67	88.00
HiFi	97.26	92.84	95.00	90.90	87.79	89.32	97.41	80.42	88.10	96.64	89.88	93.14
Nano	97.46	93.91	95.65	90.57	90.01	90.29	95.07	85.60	90.09	96.99	89.27	92.97

The unit for recall, precision, and F1 score is %. The highest recall, precision, and F1 score among four tested SV callers are marked in bold.

Rec = Recall. Pre = Precision. F1 = F1 score.

Figure 2 SV discovery in HG002. **a** SV discovery accuracy for insertions (positive SV size) and deletions (negative SV size) at different size ranges. Bars indicates the number of SVs in each size range, and lines show the SV discovery accuracy for each SV caller. **b** SV breakpoint accuracy for four tested SV callers. SVs with breakpoint shifting >100bp were included in the ± 100 bp bins. **c** SV discovery accuracy in downsampled PacBio CLR (left), HiFi (middle), and Nanopore (right) datasets for HG002.

Figure 3 Alignment-based and assembly-based SV discovery. **a** SV discovery recall and precision of alignment-based SV callers when compared with the assembly-based SV callset. **b** SV breakpoint accuracy of DeBreak, pbsv, and cuteSV in PacBio CLR (top) and HiFi (bottom) datasets. **c** Venn diagram showing the overlap among the four SV callsets. The number of SV events in each category is labeled within each section. **d** Distribution of PAV-unique and DeBreak-unique SV calls on chromosomes 1-5. Red boxes indicate positions of centromeres.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFigure.pdf](#)
- [SupplementaryFile2.xlsx](#)
- [SupplementaryFile3.xlsx](#)