2022

Comprehensive Characterization of Structural Variations Using Long-Read Sequencing Data

Yu Chen
University Of Alabama At Birmingham

COMPREHENSIVE CHARACTERIZATION OF STRUCTURAL VARIATIONS USING LONG-READ SEQUENCING DATA

by

YU CHEN

HEMANT K. TIWARI, COMMITTEE CHAIR
ZECHEN CHONG
ROBERT R. KIMBERLY
ELLIOT J. LEFKOWITZ
ALEXANDER F. ROSENBERG

A DISSERTATION

Submitted to the graduate faculty of The University of Alabama at Birmingham,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

BIRMINGHAM, ALABAMA

2022

COMPREHENSIVE CHARACTERIZATION OF STRUCTURAL VARIATIONS USING LONG-READ SEQUENCING DATA

YU CHEN

GRADUATE BIOMEDICAL SCIENCES, GENETIC, GENOMIC AND BIOINFORMATICS THEME

ABSTRACT

Structural variants (SVs) contribute to genomic diversity and play pathogenic roles in a wide range of genetic disorders. Accurate characterization of SVs is critical for genomic research and studies of disease mechanisms. The rapid development of Third-Generation Sequencing (TGS) technologies has largely increased sequencing read length compared to Next-Generation Sequencing (NGS), bringing both great potentials and challenges in SV discovery through alignment-based and assembly-based approaches. In order to take full advantage of TGS data, I have developed a suite of bioinformatics tools focusing on comprehensive characterization of SVs.

For the alignment-based SV discovery, I have developed DeBreak to identify SVs directly from long-read alignments. With the implanted density-based clustering algorithm and breakpoint refinement method, DeBreak can accurately identify SVs with precise breakpoint locations in both simulated and real datasets. When compared to the assembly-based SV callsets, DeBreak showed highest consistency among the four tested alignment-based SV callers. For the assembly-based SV discovery, I have developed Inspector to assess and improve the quality of whole-genome *de novo* assembly results. Inspector achieved highest accuracy in reporting both small-scale and larger assembly errors among the three tested assembly evaluation tools on simulated datasets. When applied on the assemblies of a real human genome, Inspector revealed that both small-

scale and structural assembly errors are enriched in repetitive regions for most assemblers. With its error correction module, Inspector reduced number of assembly errors and improved the assembly quality after polishing with long reads. In addition, I have developed FusionSeeker to detect gene fusions caused by SVs from long-read cancer transcriptome sequencing data. FusionSeeker reports gene fusions in both exonic and intronic regions with high accuracy and can reconstruct fused transcript sequences in simulated and cancer cell line datasets. These tools will facilitate the SV analysis using long-read sequencing data in the community.

Keywords: structural variant discovery, third-generation sequencing, *de novo* assembly, gene fusion

DEDICATION

To my parents, Yang Chen and Yan Zhan.

ACKNOWLEDGMENTS

First, I would like to express my gratitude to my mentor, Dr. Zechen Chong, for his advice, support, and patience throughout my graduate school training. Pursuing PhD is never an easy process, but Dr. Chong managed to make it fun and rewarding. He has always been supportive, open-minded, and passionate about science. I have learned how to become a scientist from him and will continue to be inspired by him for many years to come.

I would like to thank my thesis committee members, Dr. Hemant Tiwari, Dr. Robert Kimberly, Dr. Elliot Lefkowitz, and Dr. Alexander Rosenberg, for their insightful suggestions and contributions to my development as a scientist.

I would like to thank our lab members and collaborators, Dr. Peng Xu, Weisheng Chen, Dr. Herbert Chen, Jun Wang, Dr. John Parant, Yong Sun, Dr. Yabing Chen, Dr. Xinyang Zhao, Dr. Zhuo Zhang, Dr. Eddy Yang, Dr. Karin Hardiman, Dr. Susan Bellis, and Dr. Amy Wang, for their contributions in the collaborative projects and insights from biological and clinical aspects.

I would like to thank my family and friends, Yanshuang Wang, Zaria Zhao, Heavenleigh Zhao, Xiaobing Liu, and Yiqing Wang, for their love and friendship during my PhD. I would like to offer my special thanks to Yixin Zhang for being a loving husband and always providing support and encouragement.

TABLE OF CONTENTS

Page

LIST OF TABLES

Tables *Page*

DEBREAK: DECIPHERING THE EXACT BREAKPOINTS OF STRUCTURAL VARIATIONS USING LONG SEQUENCING READS

ACCURATE LONG-READ *DE NOVO* ASSEMBLY EVALUATION WITH INSPECTOR

GENE FUSION DETECTION AND CHARACTERIZATION IN LONG-READ CANCER TRANSCRIPTOME SEQUENCING DATA WITH FUSIONSEEKER

LIST OF FIGURES

Figures *Page*

INTRODUCTION TO STRUCTURAL VARIANT DETECTION

DEBREAK: DECIPHERING THE EXACT BREAKPOINTS OF STRUCTURAL VARIATIONS USING LONG SEQUENCING READS

ACCURATE LONG-READ *DE NOVO* ASSEMBLY EVALUATION WITH INSPECTOR

GENE FUSION DETECTION AND CHARACTERIZATION IN LONG-READ CANCER TRANSCRIPTOME SEQUENCING DATA WITH FUSIONSEEKER

LIST OF ABBREVIATIONS

CNV	Copy Number Variation
DEL	Deletion
DNA	Deoxyribonucleic Acid
DUP	Duplication
INS	Insertion
INV	Inversion
NGS	Next Generation Sequencing
PCR	Polymerase Chain Reaction
RNA	Ribonucleic Acid
SV	Structural variant
TGS	Third Generation Sequencing
TRA	Translocation

CHAPTER 1

INTRODUCTION TO STRUCTURAL VARIANT DETECTION

Structural Variant in Genetics Studies

Structural variant (SV) is a type of genetic variant that spans at least 50 bp in length. SVs are often classified as different types, including deletion (DEL), insertion (INS), duplication (DUP), inversion (INV), translocation (TRA), and copy number variation (CNV), according to the rearrangement forms and signatures. SVs contribute largely to genetic diversity and species divergence, together with Single Nucleotide Variations (SNVs) and indels [1-3]. In the human genome, SVs involve a total of approximately 10 million base pairs, which is more than the sum of SNVs (~5Mbp) and indels (~3Mbp) [4]. In a long run of evolution, some SVs were beneficial and fixed in a specific species or in a population, some were deleterious and purged during the course of evolution, and the remaining are neutral and conserved in the populations [5, 6].

SVs play important roles in genetic disorders and cancer genomes [7]. For example, complex inverted-duplication/triplication rearrangements are proven to be associated with both MECP2 duplication syndrome and Pelizaeus-Merzbacher disease [8], and deletions in the 22q11 band can lead to DiGeorge syndrome and velocardiofacial syndrome [9]. Deleterious SVs within FCGR gene family can lead to increased risk of developing systematic lupus erythematosus [10]. SVs are prevalent in numerous cancer

types and may define cancer subtypes, such as for ovarian, pancreatic, and breast cancers [11-13]. Accurate discovery of SVs in these disease and cancer genomes is critical for causal mutation identification and for studies on diagnosis and progression of diseases.

Like SNVs, SVs are mutations that derived from various mechanisms, including DNA recombination-, DNA replication-, and DNA repair-associated processes [14, 15]. For both *in vitro* experiments on human cells and *in vivo* experiments in model organisms, studies on SV mutagenesis mechanisms are based on the analysis of DNA sequences flanking SV breakpoints, which requires the precise breakpoint positions to be reported during SV discovery [16, 17]. Thus, accurate characterization of SVs is the foundation of genomics research on their contribution to genetic diversity, genetic disease, and other phenotypic traits.

Next and Third Generation High-throughput Sequencing

In 1970s, Sanger sequencing was developed for chain-termination DNA sequencing, which is known as the first-generation sequencing. Decades later, massive parallel sequencing platforms, also known as Next Generation Sequencing (NGS), were developed and became commercially available since 2005 [18]. NGS platforms usually amplify input DNA with Polymerase Chain Reaction (PCR) and then sequence the order of nucleotides during DNA synthesis. Compared to Sanger sequencing, NGS has largely improved the sequencing throughput and therefore enlarged the application of DNA sequencing in biomedical research. Commonly used NGS platforms include HiSeq and MiSeq from Illumina and SOLiD and Ion Proton from Life Technologies. These NGS

platforms usually generate single-end sequencing reads or paired-end sequencing reads with fixed read length. The base accuracy can be up to 99.9% as the signal-to-noise ratio is largely increased by PCR, which makes NGS a great method for smaller genetic variant detection including SNVs and indels. However, the read length of NGS platforms is often limited to a few hundred base pairs to avoid PCR ambiguity, which provides less power in SV detection, *de novo* assembly, and haplotype phasing.

In the past decade, Third Generation Sequencing (TGS) platforms have been developed to generate long sequencing reads from single DNA molecules. Most frequently used TGS platforms include Single-Molecular Real-Time (SMRT) sequencing from Pacific Biosciences (PacBio) and Nanopore sequencing from Oxford Nanopore. Similar to NGS, PacBio SMRT sequencing also sequences DNA molecules through DNA synthesis [19]. However, it can sequence one single DNA molecule without the need for PCR amplification, as it increased the signal-to-noise ratio by reducing background signals with the design of its zero-mode waveguide. The read length of PacBio SMRT sequencing is therefore not limited by PCR and can reach 10-20kbp, depending on the life of DNA polymerase. Oxford Nanopore sequencing passes the DNA molecular through a nanopore structure and measures the ionic current to predict the order of nucleotides according to the different electronic charges carried by four types of nucleotides [20]. Read length of Nanopore sequencing ranges between 10-30kbp and can reach up to a few million base pairs in the ultra-long platform. Such long read lengths of both PacBio and Nanopore sequencing platforms provide greater potential in detecting SVs than short-read sequencing, as the longer reads can often span the entire SV regions and better cover the repetitive regions in the genome. As both platforms sequence single

molecules, the base accuracy of TGS ranges between 75% to 90%, which also brings challenges in read alignment and small genetic variant detection.

Recently, PacBio HiFi platform is available for generating highly accurate long reads by converting double-strand DNA fragment into a circular single-strand structure and then sequencing the circular DNA fragment in repeated passes to correct potential sequencing errors [21]. This HiFi platform allows accurate detection of both SNVs/indels and SVs in the same experiment and can further benefit genome assembly, haplotype phasing, and even correction for the current reference genomes. The wide application of TGS in clinical and genomic research is currently limited by its high sequencing cost, which will be further reduced in near future.

Two Classic SV Discovery Approaches

Typically, there are two approaches for SV discovery from high-throughput sequencing data: the alignment-based and assembly-based approach (**Fig. 1**). The alignment-based approach identifies SV candidates from read alignments directly and then filters out noise candidates, which is relatively straightforward and consumes fewer computational resources. Currently, several alignment-based SV callers are available, including PBHoney [22], Sniffles [23], pbsv, and cuteSV [24], and their performance are far from optimal, especially in detecting SVs located in repetitive regions. As alignment-based SV callers detect SVs directly from raw reads, true SV calls are often contaminated with the noise signals originated from sequencing errors or alignment errors. Current SV callers usually cluster or merge raw signals of the same SV event and then discard

candidates supported by fewer reads (Sniffles and cuteSV) or lower fraction of reads from that region (pbsv), which makes the clustering/merging methods a decisive step for SV discovery accuracy. The clustering/merging methods of current alignment-based SV callers are proven to be less effective according to our benchmark [25], especially for SVs located in repetitive regions, demonstrating a clear need for SV callers with more efficient clustering methods.

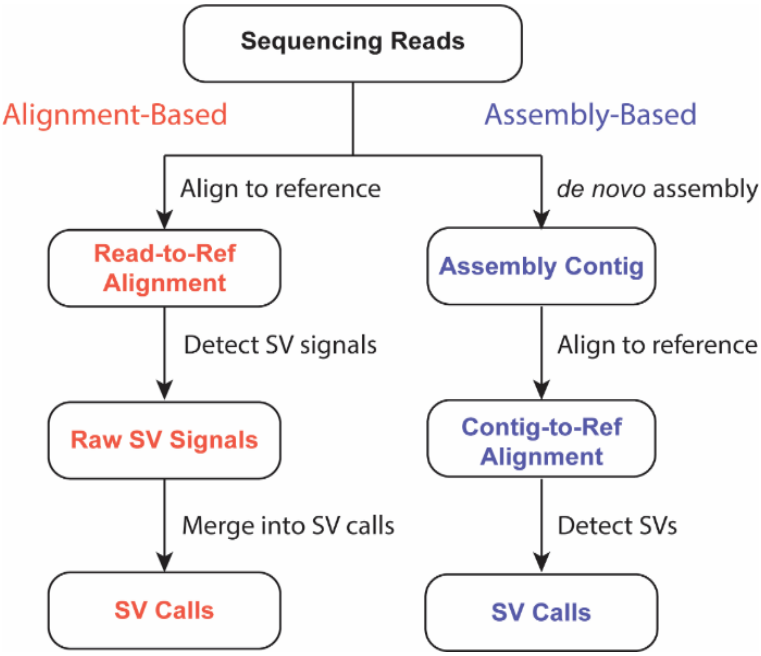

Figure 1 Two SV discovery approaches. General workflow of alignment-based and assembly-based SV discovery using high-throughput sequencing data.

In contrast, the assembly-based approach first performs local or global *de novo* assembly to generate longer and more accurate assembly contigs and then identifies SVs from the assembly results. Existing assembly-based SV callers include PAV [26], Dipcall [27], and SVIM-asm [28]. Compared to alignment-based approaches, this approach can better resolve complex SVs, SVs with ultra-large sizes, and SVs involving repetitive sequences. However, it requests much more computational resources for the assembly

process. Recently, long-read SV discovery using assembly-based approach is evolving quickly, owing to the improvement of whole-genome assembly methods, such as Canu[29], Flye [30], and hifiasm[31]. The SV discovery accuracy of this approach largely depends on the quality of assembly itself, as the mis-assemblies within assembled contigs can be falsely considered as SVs. It is thus critical to ensure the accuracy of assemblies before performing assembly-based SV calling.

Gene Fusion in Cancer Transcriptome

In cancer genome, SVs at DNA level can cause gene fusions at transcript level, which leads to disrupted gene regulation and abnormal protein functions. Gene fusions are common in many cancer types, including prostate cancer, breast cancer, and ovarian cancer [32-34]. For example, TMPRSS2-ERG fusion is observed in approximately 50% of prostate cancer patients, causing the over-expression of ERG gene fusion product [35]. Gene fusions are sometimes important drivers of tumorigenesis that can serve as diagnostic biomarkers and therapeutic targets [36]. For instance, BCR-ABL fusion is found in most Chronic Myeloid Leukemia (CML) patients and has been used as the target of standard treatment for CML [37, 38].

Previously gene fusions were often detected using short-read RNA sequencing when a read or a read pair is aligned to two distinct genes [39]. However, short-read data provides limited information about the exact fusion junction positions and sequences of fused transcripts. Recent development of long-read RNA sequencing with TGS enables full-length transcript sequencing, which is more advanced in gene fusion detection and transcript sequences reconstruction. Direct RNA sequencing from Oxford Nanopore

sequences full-length transcripts with relatively higher error rate, while Iso-Seq platform from PacBio generates highly accurate reads with fixed length ranges. Current long-read gene fusion callers include JAFFAL [40] and LongGF [41], both of which have limited performance in reporting gene fusions located in intronic regions and do not report fused transcript sequences. A gene fusion caller for both intronic and exonic regions is needed for more comprehensive detection of gene fusions and the downstream functional analysis.

Organization of the Dissertation

In this dissertation, I will introduce three novel bioinformatics tools that I have developed to facilitate accurate SV discovery and gene fusion characterization. First, I have developed an SV caller, DeBreak, to identify SVs using the alignment-based approach. I have benchmarked DeBreak and the other three alignment-based SV callers in simulated and real human datasets to assess the SV discovery and genotyping accuracy. I have compared the alignment-based and assembly-based SV callsets and applied DeBreak on a cancer dataset to demonstrate its clinical utility. Second, I have developed an assembly evaluation and correction tool, Inspector, to improve accuracy of assembly-based SV discovery by assessing and improving the quality of assembly results. Inspector identifies structural and small-scale assembly errors from read-to-contig alignment and corrects these reported errors with local *de novo* assembly. Third, I have developed a gene fusion caller, FusionSeeker, to discover gene fusion caused by SVs at transcriptome level. FusionSeeker reports gene fusions in both exonic and intronic

regions and reconstructs the accurate transcript sequences for the reported gene fusion events. Last, I will summarize the completed work and discuss future research directions.

DEBREAK: DECIPHERING THE EXACT BREAKPOINTS OF STRUCTURAL VARIATIONS USING LONG SEQUENCING READS

by

YU CHEN, AMY WANG, COURTNEY BARKLEY, YIXIN ZHANG, XINYANG ZHAO, MIN GAO, MICKY EDMONDS, ZECHEN CHONG

Submitted to *Nature Communications*

Format adapted for dissertation

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-allele 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 can also be used for supplementing whole-genome assembly-based SV discovery.

INTRODUCTION

Structural variations (SVs), or genomic rearrangements, including insertions, deletions, inversions, duplications, translocations, and complex forms of multiple events, contribute a large proportion of genetic variations in many species. In humans, SVs affect larger genomic regions in size than any other type of variants[1-5] and play a pathogenic role in a wide range of genetic disorders[6-10]. SVs are also associated with diverse phenotypes in non-human organisms[11-13]. 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 technologies[14-17]. 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.[18] resolved the complete sequence of 26,079 euchromatic structural variations, which is a sixfold increase when compared with priorwork[3] 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 sevenfold increase in SV number using multiple platforms in which the PacBio results contributed the most[4].

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, PBHoney[19] uses BLASR[20] 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, Sniffles[21] uses both within-read alignments and split-read alignments from NGMLR aligner. Sniffles can analyze both PacBio and Nanopore sequencing data and report some complex forms of SVs besides simple SVs. For both PBHoney and Sniffles, the breakpoints inferred from clusters of alignments usually are not precise, preventing experimental validation and mechanism analysis[22-25] that rely on SV junction sequences. Moreover, both tools are deficient in detecting the full spectrum of SVs. For instance, long insertions close to or longer than nearby read lengths are often missed. These issues remain a concern with a recently published alignment-based method, CuteSV[26].

Besides alignment-based methods, methods using local *de novo* assembly of mapped reads are also applied to SV discovery[4, 18, 27]. These local assembly-based methods utilize reads from haploid genomes or phased reads from diploid genomes and perform *de novo* assembly in a window. The consensus sequences generated can usually identify precise breakpoints. However, they only report insertions, deletions, and subsets of inversions. During phasing, only about two-thirds of reads in each sample can be

haplotype-partitioned. These partitions require different tools applied to the data from other platforms, preventing more generic and broader applications of SV analysis.

Whole genome *de novo* assembly can be considered the ultimate solution for SV characterization. Although researchers have made progress on this front[28-31], whole genome *de novo* assembly with long noisy reads is inherently challenging. It requires high-coverage sequencing data, and difficulties remain in dealing with long repetitive sequences, tandem repeats, as well as heterozygosity. Moreover, whole genome *de novo* assembly usually requires high-memory computing nodes and long running time, and it is difficult to evaluate accuracy.

Here, we present DeBreak (Deciphering Exact BREAKpoints), a novel algorithm for comprehensive and accurate SV discovery from long reads. DeBreak detects SV events using two different strategies, depending on whether SVs can be spanned within

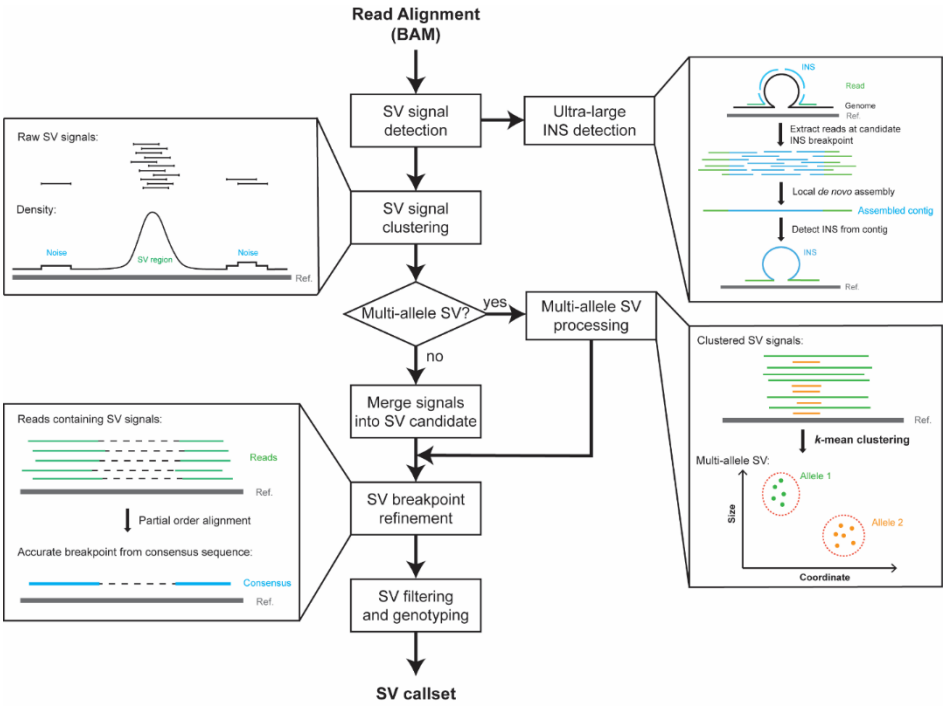

Figure 1 Workflow of DeBreak. The major steps of DeBreak SV discovery include SV signal detection, signal clustering, breakpoint refinement, and filtering and genotyping. Detailed descriptions of each step can be found in Methods.

reads (**Fig. 1**, Methods). For SVs contained within reads, DeBreak scans all read alignments for raw SV signals for each category of SV (**Fig. S1**) and then clusters these signals using a unique density-based clustering algorithm with flexible clustering window sizes (**Fig. S2**). This approach allows for accurate SV candidate identification for SVs with varying lengths and local sequence contents. In the next step, the SV breakpoint refinement with a partial order alignment (POA)[32] 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 (**Fig. S3**). DeBreak completes the analysis by integrating all identified SV events together to form a final, high-confidence SV callset.

RESULTS

1. 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 to those observed in real human samples, with Alu and LINE peaks[33] (**Fig. S4a**). PacBio-like and Nanopore-like reads were simulated based on a modified genome with pbsim[34] and Badread[35] and aligned to the human reference genome. To mimic long reads generated from different library preparation protocols, simulations were performed using three datasets with different insert sizes (**Fig. S4b**). We applied these SV callers to identify SVs and then compared the SV callsets to the ground truth to assess SV discovery accuracy for each SV caller. DeBreak achieved the highest F1 scores among the four tested SV callers in all five types of simulated SVs in PacBio datasets and similarly achieved the highest F1 scores in four SV types in Nanopore datasets (**Table 1, Fig. S5a**). Overall, DeBreak achieved accuracy of 99.23% on simulated PacBio and 98.35% on simulated Nanopore data, which was higher than the other three SV callers. Simulated read length had minor effect on DeBreak SV discovery accuracy, as DeBreak achieved similar F1 score for all types of SVs in all three replicates (**Table S1**). All evaluated SV callers have a critical parameter, “minimum number of supporting reads”, which determines the sensitivity of SV detection for these tools. In the PacBio simulation, 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. S5b**).

Table 1 SV discovery accuracy on simulated datasets

	DeBreak			Sniffles			pbsv			cuteSV		
Type	Rec	Pre	F1	Rec	Pre	F1	Rec	Pre	F1	Rec	Pre	F1
PacBio												
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
Nanopore												
DEL	98.08	**98.82**	98.45	94.83	98.72	96.74	**98.65**	98.55	**98.60**	97.87	98.27	98.07
INS	**98.17**	99.68	**98.92**	92.70	**99.73**	96.09	97.05	95.92	96.48	96.04	94.01	95.01
DUP	**97.00**	93.60	**95.27**	91.97	98.37	95.06	46.03	87.37	60.23	39.67	**99.50**	56.72
INV	91.63	98.71	**95.04**	93.67	95.18	94.41	88.20	**99.96**	93.71	**97.37**	50.00	66.07
TRA	92.83	**99.83**	**96.17**	97.17	91.38	94.18	97.83	25.26	40.15	**99.17**	50.17	66.63
Total	**97.73**	**98.96**	**98.35**	93.71	98.91	96.24	95.08	94.60	94.84	94.41	91.46	92.91

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.

As SVs tend to emerge around repeats in the genome, we performed additional simulation of repeat-associated SVs. Four SV callers were applied to identify SVs and benchmarked to assess their abilities in resolving SVs located in repeats. DeBreak achieved accuracy of 98.67% and 97.71% using PacBio and Nanopore data respectively, which was higher than the other three SV callers (**Table S2**). In both insertion and deletion detection, DeBreak achieved highest F1 score among four tested SV callers

using both PacBio and Nanopore data. The SV discovery accuracy for repeat-associated SVs was slightly lower than the previous random SVs for both insertions and deletions.

DeBreak implanted a large-insertion detection module to identify insertions longer than sequencing reads with local *de novo* assembly. To assess the improvement on maximal detectable insertion size of the large-insertion detection module, we embedded 1,000 insertions with size ranging from 5kbp to 100kbp into Chr1 and then simulated PacBio-like reads with mean read length of 15kbp. In general, insertion (INS) detection recall dropped as the insertion size increased for each SV caller, and DeBreak achieved highest recall in each size category (**Fig. S6**). When the length of insertions exceeded sequencing reads (20k-30k), DeBreak identified 70% of homozygous INS and 40% of heterozygous INS, while the other three SV callers failed to detect any events. The maximal detectable insertion size of DeBreak was approximately twice of the average read length, as the recall dropped dramatically when insertions were longer than 30kbp.

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. S7a**). From the accurate consensus sequences, DeBreak can infer more precise SV breakpoints than merely from the raw reads (**Fig. S7b**). DeBreak identified 59.81% of SVs with exact breakpoint positions and 81.33% of SVs within 1bp of the true SV breakpoint. We then down-sampled the simulated datasets to assess the effect of sequencing depth on breakpoint accuracy. Overall, the breakpoint accuracy was improved when sequencing depth increased in both PacBio and Nanopore simulations, and DeBreak was able to achieve high breakpoint accuracy starting at 20x (**Fig. S8**).

These results demonstrated that DeBreak can detect all five types of SVs in the simulated datasets with high accuracy.

2. Benchmark on real human genome

We next benchmarked SV discovery accuracy on a real human genome, HG002, from the Genome in a Bottle (GIAB) Consortium[36]. 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" regions[37]. 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 density pattern. The clustering window is larger for longer SV events and smaller for shorter SV events, improving effectiveness by merging raw SV signals into SV candidates while excluding noisy signals nearby (**Fig. S9**). For SVs located in repetitive regions, the clustering window is automatically adjusted to tolerate shifts of raw SV signals caused by repeated segments (**Fig. S10**). We then stratified the SVs into different repeat classes in both ground-truth callset and SVs reported by four SV callers in HG002 to assess the SV discovery accuracy in each repeat

type. Among the nine annotated repeat classes and non-repeat regions, DeBreak achieved highest accuracy in 9 classes using CLR data, 5 classes using HiFi data, and 10 classes using Nanopore data, suggesting higher accuracy of DeBreak in resolving repeat-associated SVs (**Fig. S11**).

Table 2 SV discovery accuracy on HG002

	DeBreak			Sniffles			pbsv			cuteSV			PAV		
	Rec	Pre	F1	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	96.60	**96.40**	96.50
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	96.18	91.32	93.69
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. Nano, Nanopore. -, Not applicable.

Previous work[38] has highlighted the functional importance of multi-allelic copy number variations (mCNVs) in gene dosage and gene expression. DeBreak can accurately identify multi-allele SVs (mSVs) in individual genomes. 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. S12**). We applied this method to identify putative mSVs in HG002. In total, we identified 802 multi-allele SVs in a single genome. The majority of mSVs (78/87, 89.66% in CLR; 49/53, 92.45% in HiFi; 74/90, 82.22% in Nanopore) in high-confidence regions had at least one allele matching with the ground-truth SV set. Multiple alternative alleles of the same SV event

were probably merged into one allele in the high-confidence SV callset, while DeBreak can report both alleles (**Fig. S13**).

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 S3**). DeBreak achieved the highest genotyping accuracy in PacBio CLR and Nanopore datasets, while cuteSV achieved slightly higher genotyping accuracy in the PacBio HiFi datasets. We performed down-sampling for PacBio CLR, HiFi and Nanopore datasets and assessed the genotyping accuracy in depth ranging from 10x to 100x. Greater sequencing depth of the input dataset increased DeBreak's genotyping accuracy, but data type had only a relatively minor impact on genotyping accuracy (**Fig. S14**).

We then assessed the SV discovery accuracy for SVs of different sizes. DeBreak achieved consistent and high accuracy for small and large SVs, especially for 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. S15**), 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. S16**). 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 with exact SV breakpoints and 63.53% of SVs with breakpoint shift within 1bp as reported in GIAB, which was higher than pbsv (41.73%

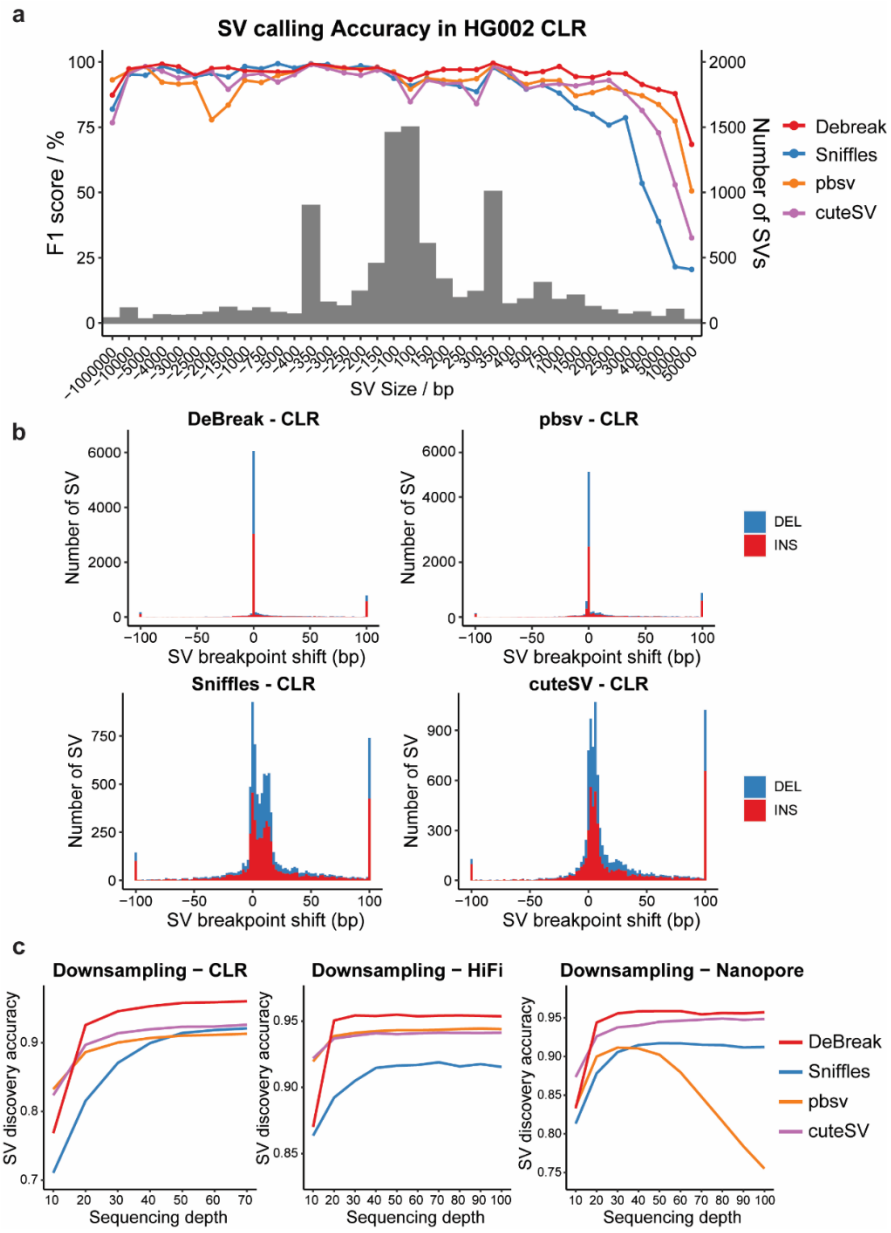

Figure 2 SV discovery in HG002. a SV discovery accuracy for insertions (positive SV size) and deletions (negative SV size) at different size ranges in CLR dataset. Bars indicates the number of SVs in each size range, and lines show the SV discovery accuracy for each SV caller. **b** SV breakpoint accuracy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

and 47.41%), Sniffles (4.99% and 13.45%), and cuteSV (5.18% and 13.87%) 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. S17**).

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 (**Fig. S18**). For datasets with depth ≥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. We further benchmarked the SV breakpoint accuracy in downsampled datasets. DeBreak reported most SVs with exact breakpoints starting at 20x in Pacbio HiFi and Nanopore datasets and at 30x for PacBio CLR data (**Fig. S19**). Taken together, these results highlight that DeBreak can accurately identify insertions and deletions, two major types of SVs, with precise breakpoints in real human genomes.

3. Comparison to assembly-based SV discovery

Currently, *de novo* assembly is used to comprehensively characterize genome-wide SVs[28-31]. To compare alignment-based with assembly-based SV discovery

approaches, we applied DeBreak, pbsv, Sniffles, 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 pipeline[28]. Overall, the assembly-based SV approach discovered a slightly higher number of SV events (22,897 to 27,187) compared with alignment-based methods. By treating these SVs as the "ground truth", we evaluated the SV discovery accuracy of four 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%), Sniffles (71.44%), and cuteSV (77.38%) (**Table S4, Table S5**). In each sample, DeBreak achieved both higher recall and precision than the other three callers (**Fig. 3a**), suggesting a higher consistency with the assembly-based SV discovery than the other three alignment-based SV callers. Among four tested SV callers, cuteSV showed highest consistency in SV genotyping with an assembly-based approach (**Table S6**). DeBreak detected a total of 3,100 mSVs in the three CLR datasets and 3,097 mSVs in the three HiFi datasets, with approximately 71% of these 6,200 and 6,194 alternative alleles validated by assembly-based approach in CLR and HiFi datasets, respectively (**Table S7**). A total of 23 and 24 mCNVs were

identified in the three CLR and HiFi datasets with further annotation of copy number variation using *k*-mer counts (**Fig. S20**).

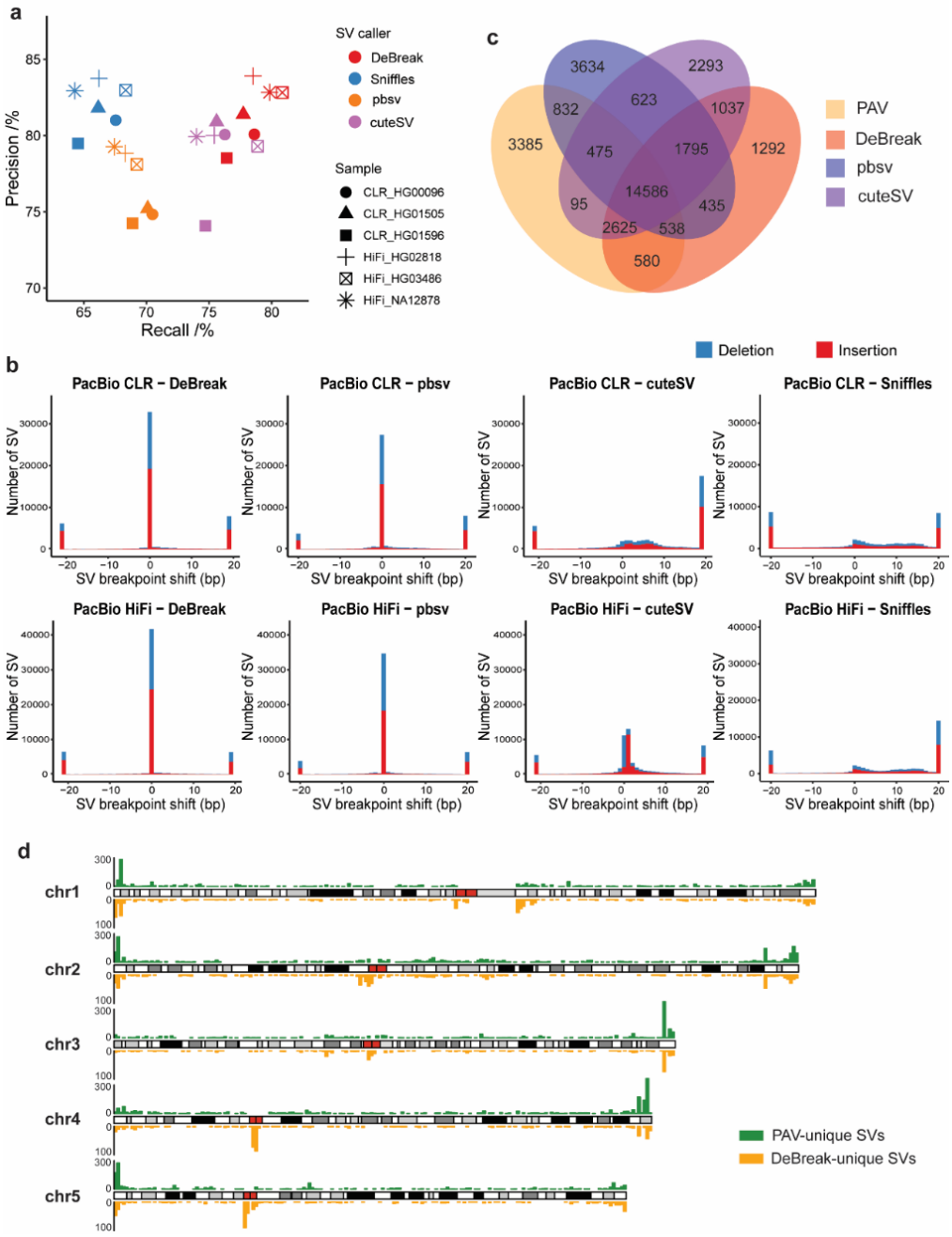

Figure 3 Alignment-based and assembly-based SV discovery. a SV discovery recall and precision of alignment-based SV callers when compared with the assembly-based SV callset. **b** SV breakpoint accuracy of DeBreak, pbsv, cuteSV, and Sniffles in PacBio CLR (top) and HiFi (bottom) datasets. **c** Venn diagram showing the overlap among 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.

As assembly-based SV calls usually have accurate SV breakpoints inferred from assembled contigs, we also compared the SV breakpoint accuracy of four SV callers. With the breakpoint-refinement module, DeBreak identified 46.83% of SVs with exact SV breakpoints and 48.12% of SVs within 1bp shift on the three PacBio CLR datasets, while pbsv reported 39.03% and 40.99%, Sniffles reported 3.02% and 6.82%, and cuteSV reported 2.75% and 7.69% of SVs with exact breakpoints and within 1bp shift, respectively (**Fig. 3b**). For the three PacBio HiFi datasets, DeBreak also achieved better breakpoint accuracy than the other three callers, as 56.98%, 47.41%, 2.99%, and 15.33% of SVs were identified with exact SV breakpoints, and 58.09%, 48.85%, 6.41%, and 35.54% of SVs were identified within 1bp shift by DeBreak, pbsv, Sniffles, 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. S21**). 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. 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. S22**). 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.

We then compared four alignment-based SV callers to assembly-based approach in the CHM13 cell line, where a complete telomere-to-telomere assembly is available[39]. An assembly-based SV callset was generated using Dipcall[40] on the CHM13 assembly. We selected the high-confidence regions of Dipcall as "ground truth". Four alignment-based SV callers were applied on the PacBio CLR (70x), HiFI (57x), and Nanopore (126x) read alignment files. For both insertion and deletion detection, DeBreak achieved highest consistency with assembly-based SV callset in all three data types (**Table S8**). All SVs in CHM13 should be homozygous, as CHM13 is a haploid cell line. We then benchmarked the genotyping accuracy of four alignment-based SV callers, with only 'GT=1/1' as correct genotype. In CHM13, DeBreak achieved highest genotyping accuracy in PacBio HiFi and Nanopore datasets, and cuteSV achieved slightly higher genotyping accuracy in PacBio CLR dataset (**Table S9**).

4. SV discovery in cancer genomes

SVs play important roles in cancer development and progression[41-43]. Unlike germline SVs, cancer genomes may contain more large-scale deletions, duplications,

inversions, translocations, and other complex SVs[44-46]. 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. The PacBio CLR dataset (72x, mean read length of 9.87kbp) of SKBR3 was downloaded and aligned to the human reference genome. 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, large proportions of SVs were identified by all four SV callers (**Fig. S23**). 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 data[47] (**Fig. S24**). 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. S25**). 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**). The cross-validation rate was 26.83%, which was higher than Sniffles (25/116, 21.55%), pbsv (5/39, 12.82%), and cuteSV (7/58, 12.07%). 6 out

of the 11 cross-validated gene fusions identified by DeBreak have been previously reported using transcriptomic data[47-50]. 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. S26**). 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. S27**). 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.

5. 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 S10**). 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 inference[22-25]; 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-allele SVs, which are functionally important; and 5) multiple functions can be applied to both healthy and unhealthy genomes.

Due to the limited availability of ground-truth SV sets, DeBreak was benchmarked for insertion and deletion discovery in HG002 and HGSVC samples, but not for duplication, inversion, or translocation. Further validation of SV discovery accuracy on these SV types would be desirable and will help improve DeBreak's performance if comprehensive high-confidence truth SV sets become more readily available. 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.

Several features of the input sequencing dataset have essential impact on SV discovery accuracy. Data type (sequencing platform) affects SV discovery accuracy and breakpoint accuracy. Based on our benchmarks and as expected, datasets with lower sequencing error rates often lead to better SV discovery accuracy and breakpoint accuracy than datasets with higher error rates at similar levels of sequencing depth. Sequencing depth also affects accuracy for SV discovery, breakpoint position, and genotyping. Sequencing read length can affect maximal size of detectable SVs, especially for insertion detection.

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. In all situations, DeBreak and other alignment-based methods consume much less computational resources than assembly-based methods. Although comprehensive evaluation and validation between alignment-based and assembly-based approaches are needed, alignment-based methods will continue to serve important roles in 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) multi-allele SV identification, 5) SV breakpoint refinement, and 6) 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 insertions and deletions can be contained within a single alignment (**Fig. S1a**). 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 (**Fig. S1b**). Insertions are inferred when there are extra sequences in the read between two adjunct alignments. Deletions are inferred when a region on reference genome is skipped between two alignments. Duplications are inferred when two alignments are overlapped on the reference genome. Inversions are inferred when two alignments have distinct orientation. Translocations are inferred when read is aligned to two distinct chromosomes with help of "SA" tag in the BAM file. 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 using a density-based clustering algorithm for insertion, deletions, duplications, and inversions (**Fig. S2**). All signals from the same chromosome with the same SV type are sorted based on coordinates. The density of SV raw signals is computed for each position on the chromosome. DeBreak scans the chromosome for density peaks above the threshold, which is automatically adjusted according to the sequencing depth of the input dataset. For each peak, the boundaries of the SV region are defined on both sides of the peak summit when the density drops to 10% of the summit height. All raw signals located within the SV region are then merged into one SV candidate. For translocation, positions of both breakpoints are clustered with fixed window of 400/800bp. The window size is determined by the standard variation of breakpoint positions. A 400bp window is used for groups of raw signals with smaller standard variation and an 800bp window is used for groups of raw signals with larger standard variation.

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 wtdbg2[51] to refine breakpoint locations. For each SV candidate, DeBreak collects all reads containing raw signals of 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. Reads containing raw signals of a particular SV event are considered as 'supporting reads' for this SV. 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 single reads. While scanning read alignments for raw SV signals, DeBreak also records positions of clipped ends of read alignments. Read alignments with at least 200bp unmapped sequences (clipped sequences) on either side are considered as 'clipped' alignments (**Fig. S3a**). After scanning through a chromosome, DeBreak identifies candidate insertion breakpoint regions with enriched clipped alignment, where at least N_{supp} reads are clipped on left side of the candidate breakpoint and another N_{supp} reads are clipped on right side of the breakpoint. It then collects these clipped reads at each candidate breakpoint region and performs local *de novo* assembly with wtdbg2 to reconstruct assembly contigs that contains full-length inserted sequence (**Fig. S3b**). 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 consistent 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. PacBio-like reads were simulated from the modified genome using pbsim (v1.0.3) with options "--data-type CLR --model_qc model_qc_clr --depth 25 --accuracy-mean 0.85". Nanopore-like reads were simulated using Badread (v0.2.0) with options "--quantity 25x --junk_reads 0 --random_reads 0 --chimeras 0 --glitches 0,0,0". 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 (v1.2) was applied to minimap2 alignment results with default settings. pbsv (v2.6.2) was run on pbmm2 alignment results with default settings, and Sniffles (v1.0.8) was run on the ngmlr alignment results with options "--genotype -s 4/5/6/7/8/9/10". A serious 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 (v1.0.11) was run on minimap2 alignment results with options "--genotype". All SVs with length ≥45bp 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) $ABS(Cor_G – Cor_C) <= 1kbp$
3) $0.5 * Size_G <= Size_C <= 2 * Size_G$

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.

2.3 Simulation of repeat-associated SVs. RepeatMasker annotation of human genome was downloaded from UCSC Table Browser. 10,000 repeats were randomly selected with size ranging from 50bp to 20kbp. Insertions were simulated by adding an additional copy of the repeat, and deletions were simulated by removing the repeat. SVs were assigned as 'homozygous and 'heterozygous' with a ratio of 1:1. PacBio-like and Nanopore-like reads were simulated using pbsim (v1.0.3) and Badread (v0.2.0) with sequencing depth of 50x and average read length of 10kbp, 15kbp, and 20kbp. Sequencing reads were aligned to human reference genome with minimap2, ngmlr, and pbmm2. DeBreak (v1.2) and pbsv (v2.6.2) were applied on read alignment files with default settings. A series of "-s" were provided for Sniffles (v1.0.8) and cuteSV (v1.0.11), and the SV callsets with highest accuracy were used for comparison. SV discovery accuracy was benchmarked using the same criterial as in section 2.2.

2.4 Simulation of ultra-large insertion. 1,000 insertions were randomly simulated and embedded into human Chromosome 1, with insertion size ranging from 5kbp to 100kbp. Insertions were assigned as 'homozygous and 'heterozygous' with a ratio of 1:2. 50x PacBio-like reads were simulated with average read length of 15kbp and then aligned to human reference genome. DeBreak (v1.2), Sniffles (v1.0.8), pbsv (v2.6.2), and cuteSV (v1.0.11) were applied on read alignment files to identify SVs with default settings. Recall for insertion detection was benchmarked at different size ranges using the same criterial as in section 2.2.

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 minimap2[52], ngmlr, and pbmm2 with default settings. DeBreak (v1.2) and cuteSV (v1.0.11) were applied to minimap2 alignment with default settings. pbsv (v2.6.2) was applied to pbmm2 alignment with default settings. Sniffles (v1.0.8) 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. Coordinates of SVs in the PAV callset were converted from hg38 to hg19 using LiftOver. SV callsets of four alignment-based SV callers and PAV were benchmarked within the high-confidence regions (HG002_SVs_Tier1_v0.6.bed) by comparing to the benchmark SV callset using the same criterial as for the simulation benchmark. Repeat types of SVs were classified with RepeatMasker (v4.1.2) using sequences of the longest allele. Shifts of SV breakpoints were 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 in HG002, we downsampled the PacBio CLR dataset to a series of depth from 10x to 70x, and downsampled PacBio HiFi and Nanopore datasets to a series of depth from 10x to 100x.

Sequencing reads were randomly selected to generate datasets of 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

4.1 SV discovery in HGSVC samples. Raw PacBio CLR or HiFi reads of sample HG00096, HG01505, HG01596, HG02818, HG03486, and NA12878 were downloaded and aligned to GRCh38 with minimap2, ngmlr, and pbmm2 under default settings. DeBreak (v1.2), pbsv (v2.6.2), Sniffles (v1.0.8), and cuteSV (v1.0.11) were applied to the alignment files to identify SVs with default settings. The merged assembly-based SV callset was downloaded from HGSVC2 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 karyoploteR[53].

4.2 SV benchmark in CHM13 cell line. Dipcall (v0.3) was applied on the Telomere-to-Telomere assembly of CHM13 with default settings to generate assembly-based callset.

SVs with size of at least 50bp were used as ground truth callset. PacBio CLR, HiFi, and Nanopore reads were downloaded and aligned to GCh38 with minimap2, ngmlr, and pbmm2. DeBreak (v1.2), Sniffles (v1.0.8), pbsv (v2.6.2), and cuteSV (v1.0.11) were applied on read alignment files with default settings. SV callsets were benchmarked only in the high-confidence regions suggest by Dipcall. Multimatch was allowed when comparing alignment-based SV callsets to the ground truth. Genotyping accuracy was benckmarked with 'GT=1/1' as correct and the remaining as incorrect genotypes.

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 callers[47]. 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)[54], and the specificity was verified with UCSC in-silico PCR (**Fig. S25**). 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.

6. Data and code 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 HGSVC2 data portal at https://www.internationalgenome.org/data-portal/data-collection/hgsvc2. T2T assembly and sequencing reads of CHM13 were downloaded from https://github.com/marbl/CHM13. The PacBio CLR and Iso-Seq data of SKBR3 cell line were downloaded from NCBI SRA under BioProject PRJNA476239. SV callsets evaluated in the paper are available at https://zenodo.org/record/7214225.

DeBreak is publicly available at https://github.com/Maggi-Chen/DeBreak under the MIT License. We used v1.2 version for SV discovery and benchmark presented in the manuscript. Key custom Python scripts used in the manuscript are available at https://github.com/Maggi-Chen/DB_code.

REFERENCES

1. Chen Y, Wang AY, Barkley C, Zhao X, Gao M, Edmonds M, Chong Z: DeBreak: Deciphering the exact breakpoints of structural variations using long sequencing reads. *Preprint from Research Square* 2022.

2. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al: Global variation in copy number in the human genome. *Nature* 2006, 444:444-454.

3. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH, et al: An integrated map of structural variation in 2,504 human genomes. *Nature* 2015, 526:75-81.

4. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al: Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* 2019, 10:1784.

5. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, et al: Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* 2019, 176:663-675 e619.

6. Stankiewicz P, Lupski JR: Structural variation in the human genome and its role in disease. *Annu Rev Med* 2010, 61:437-455.

7. Weischenfeldt J, Symmons O, Spitz F, Korbel JO: Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* 2013, 14:125-138.

8. Tanzi RE, Bird ED, Latt SA, Neve RL: The amyloid beta protein gene is not duplicated in brains from patients with Alzheimer's disease. *Science* 1987, 238:666-669.

9. Chartier-Harlin MC, Kachergus J, Roumier C, Mouroux V, Douay X, Lincoln S, Levecque C, Larvor L, Andrieux J, Hulihan M, et al: Alpha-synuclein locus duplication as a cause of familial Parkinson's disease. *Lancet* 2004, 364:1167-1169.

10. Yang L, Luquette LJ, Gehlenborg N, Xi R, Haseley PS, Hsieh CH, Zhang C, Ren X, Protopopov A, Chin L, et al: Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* 2013, 153:919-929.

11. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, et al: Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell* 2020, 182:145-161 e123.

12. Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, et al: Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 2018, 557:43-49.

13. Weissensteiner MH, Bunikis I, Catalan A, Francoijs KJ, Knief U, Heim W, Peona V, Pophaly SD, Sedlazeck FJ, Suh A, et al: Discovery and population genomics of structural variation in a songbird genus. *Nat Commun* 2020, 11:3403.

14. Merker JD, Wenger AM, Sneddon T, Grove M, Zappala Z, Fresard L, Waggott D, Utiramerur S, Hou Y, Smith KS, et al: Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med* 2018, 20:159-163.

15. Mizuguchi T, Suzuki T, Abe C, Umemura A, Tokunaga K, Kawai Y, Nakamura M, Nagasaki M, Kinoshita K, Okamura Y, et al: A 12-kb structural variation in progressive myoclonic epilepsy was newly identified by long-read whole-genome sequencing. *J Hum Genet* 2019, 64:359-368.

16. Sanchis-Juan A, Stephens J, French CE, Gleadall N, Megy K, Penkett C, Shamardina O, Stirrups K, Delon I, Dewhurst E, et al: Complex structural variants in Mendelian disorders: identification and breakpoint resolution using short- and long-read genome sequencing. *Genome Med* 2018, 10:95.

17. Sone J, Mitsuhashi S, Fujita A, Mizuguchi T, Hamanaka K, Mori K, Koike H, Hashiguchi A, Takashima H, Sugiyama H, et al: Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease. *Nat Genet* 2019, 51:1215-1221.

18. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, Antonacci F, Surti U, Sandstrom R, Boitano M, et al: Resolving the complexity of the human genome using single-molecule sequencing. *Nature* 2015, 517:608-611.

19. English AC, Salerno WJ, Reid JG: PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics* 2014, 15:180.

20. Chaisson MJ, Tesler G: Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* 2012, 13:238.

21. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC: Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* 2018, 15:461-468.

22. Liu P, Carvalho CM, Hastings PJ, Lupski JR: Mechanisms for recurrent and complex human genomic rearrangements. *Curr Opin Genet Dev* 2012, 22:211-220.

23. Dittwald P, Gambin T, Szafranski P, Li J, Amato S, Divon MY, Rodriguez Rojas LX, Elton LE, Scott DA, Schaaf CP, et al: NAHR-mediated copy-number variants in a clinical population: mechanistic insights into both genomic disorders and Mendelizing traits. *Genome Res* 2013, 23:1395-1409.

24. Beck CR, Carvalho CMB, Akdemir ZC, Sedlazeck FJ, Song X, Meng Q, Hu J, Doddapaneni H, Chong Z, Chen ES, et al: Megabase Length Hypermutation Accompanies Human Structural Variation at 17p11.2. *Cell* 2019, 176:1310-1324 e1310.

25. Carvalho CM, Lupski JR: Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* 2016, 17:224-238.

26. Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, Wang Y: Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol* 2020, 21:189.

27. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, et al: Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res* 2017, 27:677-685.

28. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al: Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 2021, 372.

29. Garg S, Fungtammasan A, Carroll A, Chou M, Schmitt A, Zhou X, Mac S, Peluso P, Hatas E, Ghurye J, et al: Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat Biotechnol* 2021, 39:309-312.

30. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, et al: Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol* 2020, 38:1044-1053.

31. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al: Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 2019, 37:1155-1162.

32. Lee C, Grasso C, Sharlow MF: Multiple sequence alignment using partial order graphs. *Bioinformatics* 2002, 18:452-464.

33. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, et al: Mapping copy number variation by population-scale genome sequencing. *Nature* 2011, 470:59-65.

34. Ono Y, Asai K, Hamada M: PBSIM: PacBio reads simulator--toward accurate genome assembly. *Bioinformatics* 2013, 29:119-121.

35. Wick RR: Badread: simulation of error-prone long reads. *The Journal of Open Source Software* 2019, 4.

36. Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, Irvine SA, Trigg L, Truty R, McLean CY, et al: An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol* 2019, 37:561-566.

37. Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, Sherry S, Koren S, Phillippy AM, Boutros PC, et al: A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol* 2020, 38:1347-1355.

38. Handsaker RE, Van Doren V, Berman JR, Genovese G, Kashin S, Boettger LM, McCarroll SA: Large multiallelic copy number variations in humans. *Nat Genet* 2015, 47:296-303.

39. Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, Taylor DJ, Shafin K, Shumate A, Xiao C, et al: A complete reference genome improves analysis of human genetic variation. *Science* 2022, 376:eabl3533.

40. Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, MacArthur D: A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods* 2018, 15:595-597.

41. Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K, Khurana E, Waszak S, Korbel JO, Haber JE, et al: Patterns of somatic structural variation in human cancer genomes. *Nature* 2020, 578:112-121.

42. Petljak M, Alexandrov LB, Brammeld JS, Price S, Wedge DC, Grossmann S, Dawson KJ, Ju YS, Iorio F, Tubio JMC, et al: Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* 2019, 176:1282-1294 e1220.

43. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, et al: The genomic complexity of primary human prostate cancer. *Nature* 2011, 470:214-220.

44. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, Pleasance ED, Lau KW, Beare D, Stebbings LA, et al: Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 2011, 144:27-40.

45. Cortes-Ciriano I, Lee JJ, Xi R, Jain D, Jung YL, Yang L, Gordenin D, Klimczak LJ, Zhang CZ, Pellman DS, et al: Publisher Correction: Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet* 2020.

46. Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, Park K, Kitabayashi N, MacDonald TY, Ghandi M, et al: Punctuated evolution of prostate cancer genomes. *Cell* 2013, 153:666-677.

47. Nattestad M, Goodwin S, Ng K, Baslan T, Sedlazeck FJ, Rescheneder P, Garvin T, Fang H, Gurtowski J, Hutton E, et al: Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res* 2018, 28:1126-1135.

48. Kim D, Salzberg SL: TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol* 2011, 12:R72.

49. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, Rye IH, Nyberg S, Wolf M, Borresen-Dale AL, Kallioniemi O: Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol* 2011, 12:R6.

50. Chen K, Navin NE, Wang Y, Schmidt HK, Wallis JW, Niu B, Fan X, Zhao H, McLellan MD, Hoadley KA, et al: BreakTrans: uncovering the genomic architecture of gene fusions. *Genome Biol* 2013, 14:R87.

51. Ruan J, Li H: Fast and accurate long-read assembly with wtdbg2. *Nat Methods* 2020, 17:155-158.

52. Li H: Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018, 34:3094-3100.

53. Gel B, Serra E: karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* 2017, 33:3088-3090.

54. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG: Primer3--new capabilities and interfaces. *Nucleic Acids Res* 2012, 40:e115.

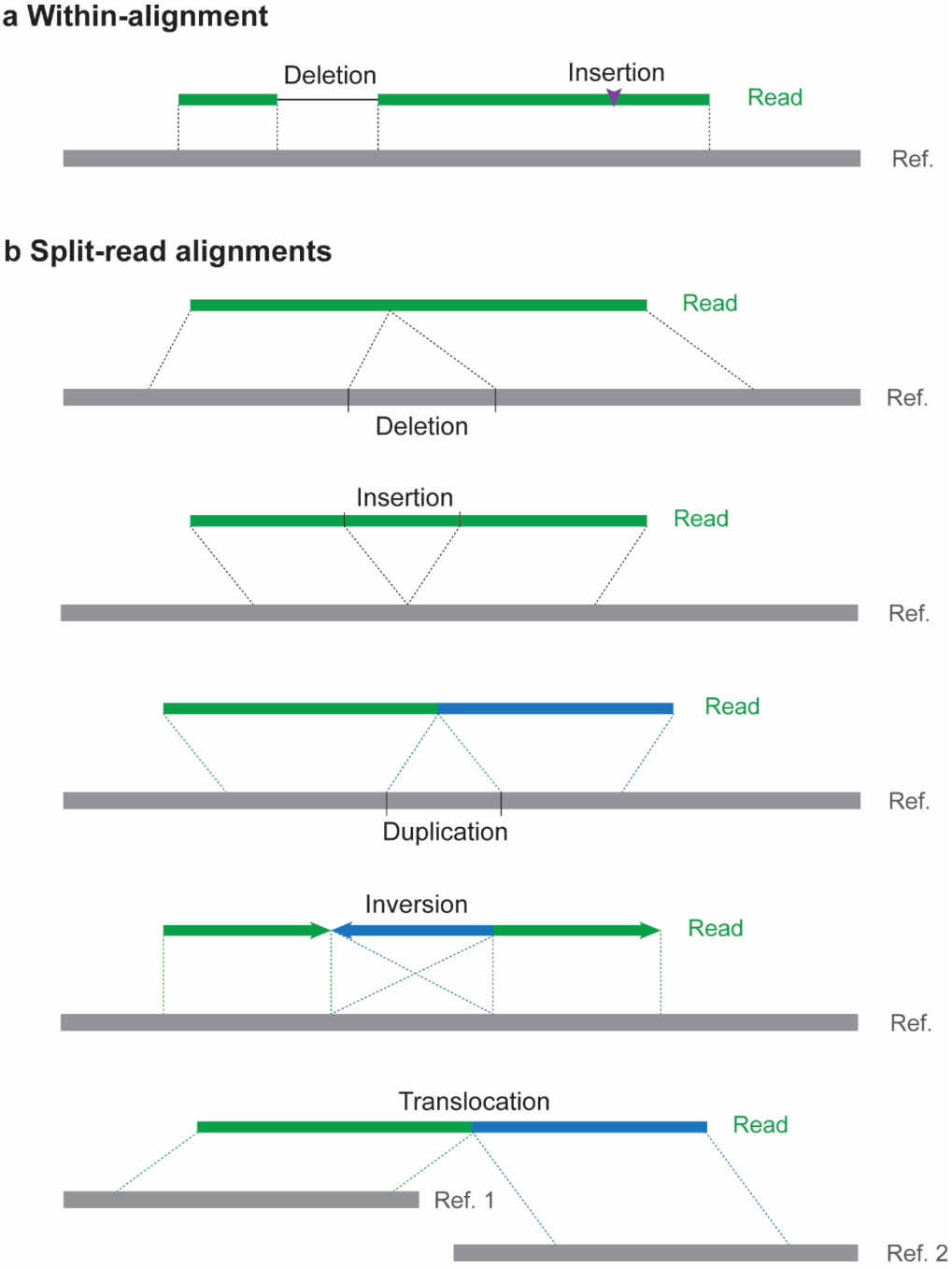

Figure S1 SV raw signal detection of DeBreak. a Deletion and insertion can be directly inferred within a single-read alignment. **b** Larger deletion and insertion, duplication, inversion, and translocation can be inferred from split-read alignments based on the location and orientation of two alignment segments. Alignments with distinct colors represent separate alignments of the same sequencing read.

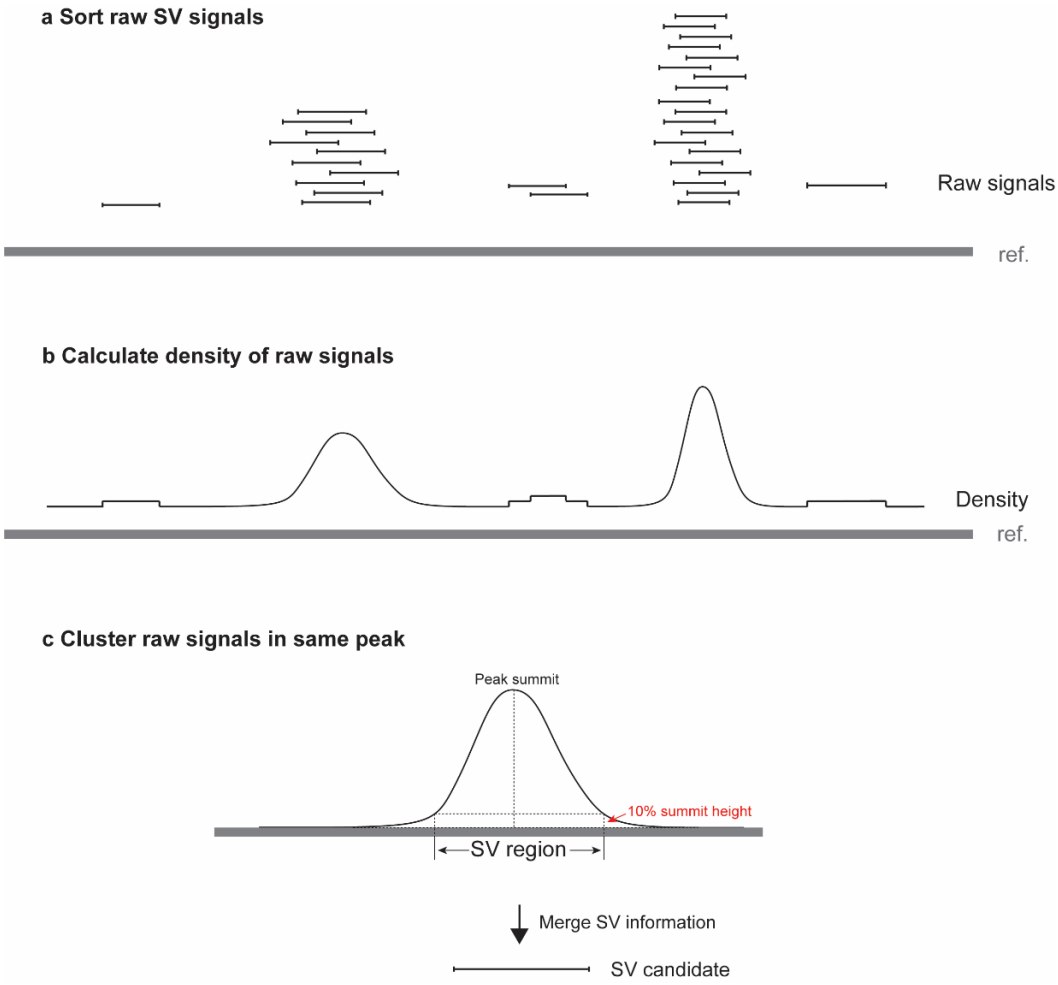

Figure S2 SV Density-based SV raw signal clustering. a SV raw signals from the same chromosome with the same SV type are sorted based on coordinates. **b** Density of raw signals is calculated for each base pair on the reference genome. DeBreak scans through the chromosome for peaks above a defined threshold. **c** For each peak, boundaries of the SV region for a SV event are determined where density drops to 10% of the peak summit height. All raw signals located within the SV region are merged into one SV candidate.

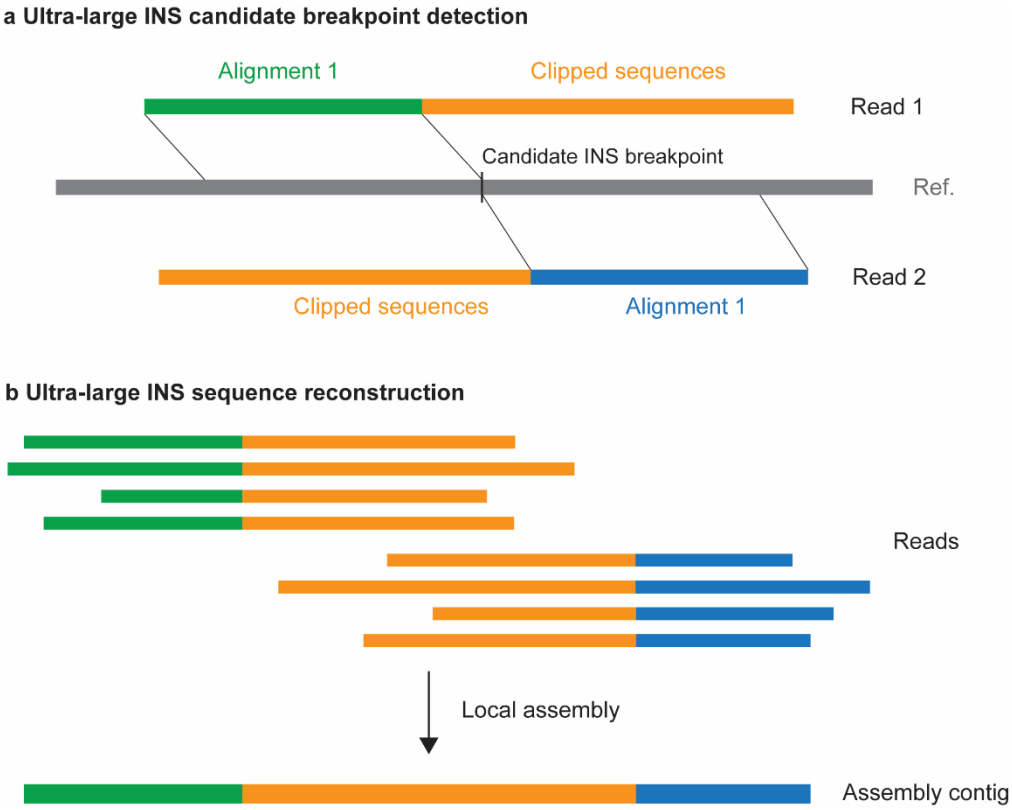

Figure S3 Ultra-large INS detection. a Example of reads with clipped alignment. Enriched "clipped" reads are required for both side of the candidate INS breakpoint for following local assembly. **b** Local *de novo* assembly using "clipped" reads. Reads aligned to both sides of the candidate INS breakpoint are collected for local assembly to generate an assembly contig that includes the full-length insertion sequence.

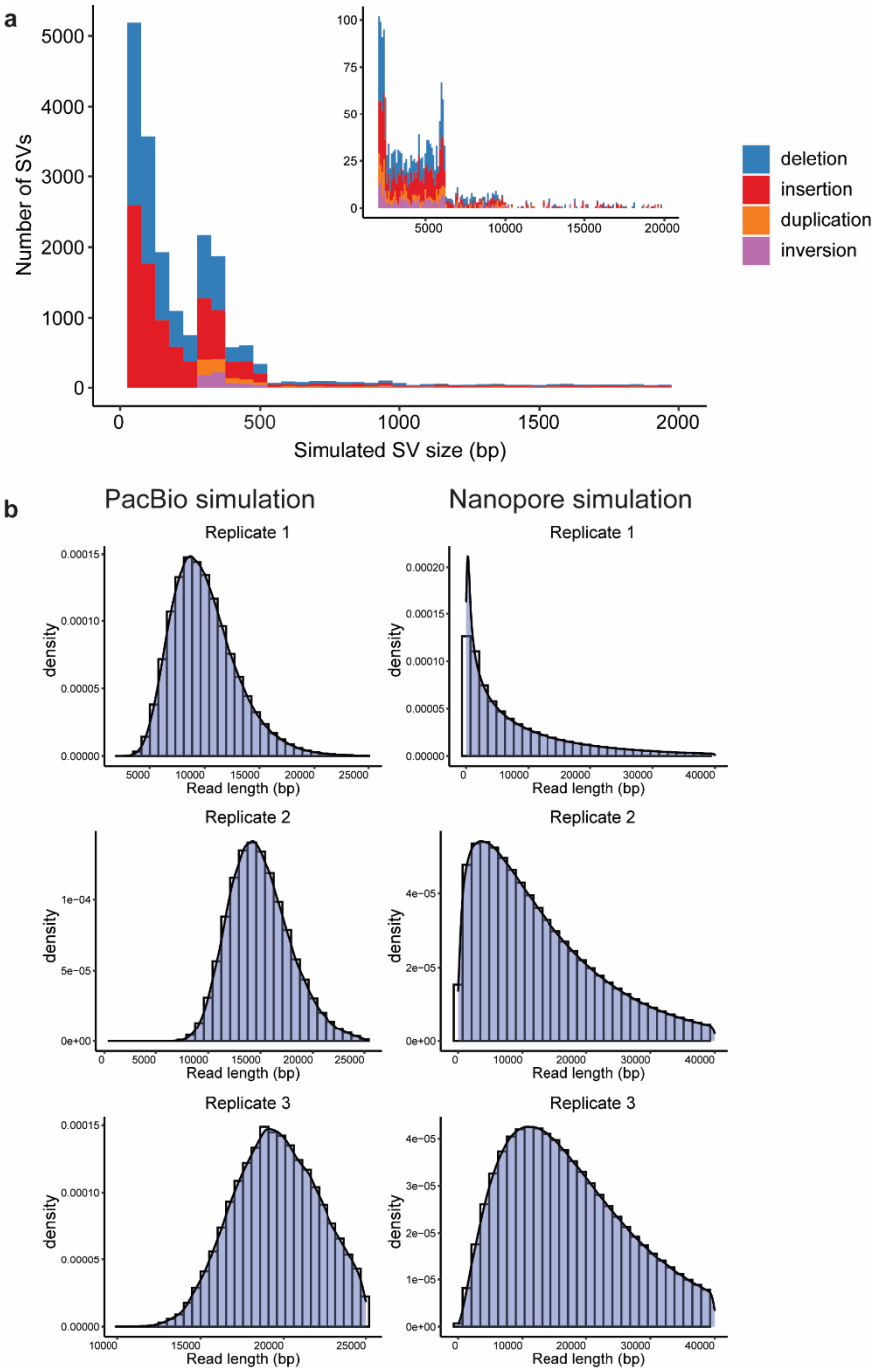

Figure S4 Characteristics of simulated datasets. a Size distribution of simulated SVs. Peaks at 300-350bp were simulated to mimic Alu elements, and peaks near 6kbp were simulated to mimic LINE mobile elements. **b** Length distributions of simulated PacBio (left) and Nanopore (right) reads in three simulated datasets.

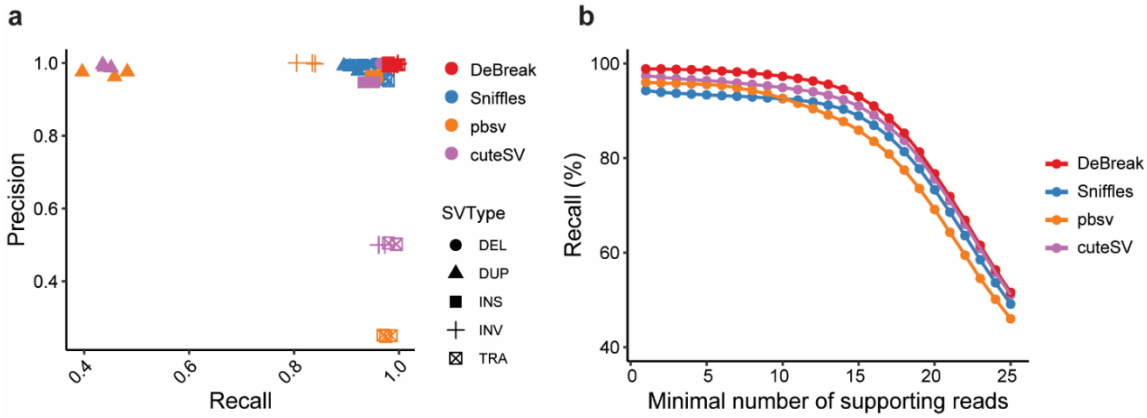

Figure S5 SV discovery accuracy in simulated PacBio datasets. a SV discovery accuracy for four tested SV callers in three simulated datasets. **b** Recall for deletion and insertion detection at different thresholds of 'minimal supporting reads' for the four SV callers in simulated datasets.

Table S1 SV discovery accuracy (F1 score) on three replicated simulated datasets

	DeBreak			Sniffles			pbsv			cuteSV		
Type	**Rep1**	**Rep2**	**Rep3**	**Rep1**	**Rep2**	**Rep3**	**Rep1**	**Rep2**	**Rep3**	**Rep1**	**Rep2**	**Rep3**
PacBio												
DEL	**99.54**	**99.53**	**99.56**	97.56	97.60	97.69	98.07	98.49	98.59	98.19	98.40	98.48
INS	**98.89**	**99.09**	**99.16**	94.95	96.25	96.20	95.58	96.13	95.69	93.96	94.96	95.01
DUP	**98.40**	**98.49**	**98.44**	94.11	94.90	94.63	64.52	62.06	56.33	60.54	61.91	60.71
INV	**99.65**	**99.05**	**99.85**	95.79	95.68	96.16	89.20	91.01	91.21	65.75	65.82	66.06
TRA	**99.50**	**99.75**	**98.99**	96.55	96.77	96.53	39.63	39.96	40.00	66.67	66.78	66.67
Total	**99.20**	**99.26**	**99.34**	96.15	96.78	96.81	94.12	94.58	94.32	92.53	93.12	93.16
Nanopore												
DEL	98.52	98.38	98.45	96.83	98.38	98.45	**98.74**	**98.59**	**98.47**	98.12	98.04	98.06
INS	**99.03**	**98.85**	**98.88**	96.02	98.85	98.88	96.63	96.45	96.36	94.97	95.13	94.94
DUP	**95.00**	**95.39**	**95.42**	94.89	95.39	95.42	62.07	57.37	61.25	55.72	57.43	57.00
INV	94.99	**94.57**	**95.55**	**95.01**	94.57	95.55	93.06	94.29	93.79	66.03	65.99	66.19
TRA	94.18	**98.48**	**95.83**	**95.15**	98.48	95.83	39.92	40.04	40.50	66.67	66.67	66.55
Total	**98.40**	**98.29**	**98.35**	96.28	98.29	98.35	94.97	94.77	94.78	92.89	92.97	92.89

The unit for the F1 score is %. The highest F1 score in each replicate is shown in bold.
Rep1, replicate 1 (10kbp). Rep2, replicate 2 (15kbp). Rep3, replicate 3 (20kbp).

Table S2 SV discovery accuracy of SV involving repeats

	Deletion			Insertion			Total		
Type	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
PacBio									
DeBreak	**99.03**	99.86	**99.44**	**97.69**	98.12	**97.91**	**98.36**	98.99	**98.67**
Sniffles	97.72	99.86	98.78	95.19	97.85	96.49	96.46	98.85	97.64
Pbsv	98.13	99.97	99.04	88.95	**98.26**	93.37	93.54	**99.15**	96.26
cuteSV	97.97	**99.97**	98.96	94.77	97.06	95.89	96.37	98.52	97.43
Nanopore									
DeBreak	98.12	**98.51**	**98.31**	**97.35**	96.89	**97.12**	**97.73**	97.69	**97.71**
Sniffles	97.65	97.98	97.81	93.01	**98.21**	95.52	95.33	**98.09**	96.68
Pbsv	**99.13**	97.25	98.18	92.60	97.95	95.20	95.87	97.58	96.72
cuteSV	98.31	97.95	98.13	93.87	97.60	95.70	96.09	97.78	96.93

Averages for three replicates (10kbp, 15kbp,20kbp). The unit for recall, precision, and F1 score is %. The highest recall, precision, and F1 score for each category are shown in bold.

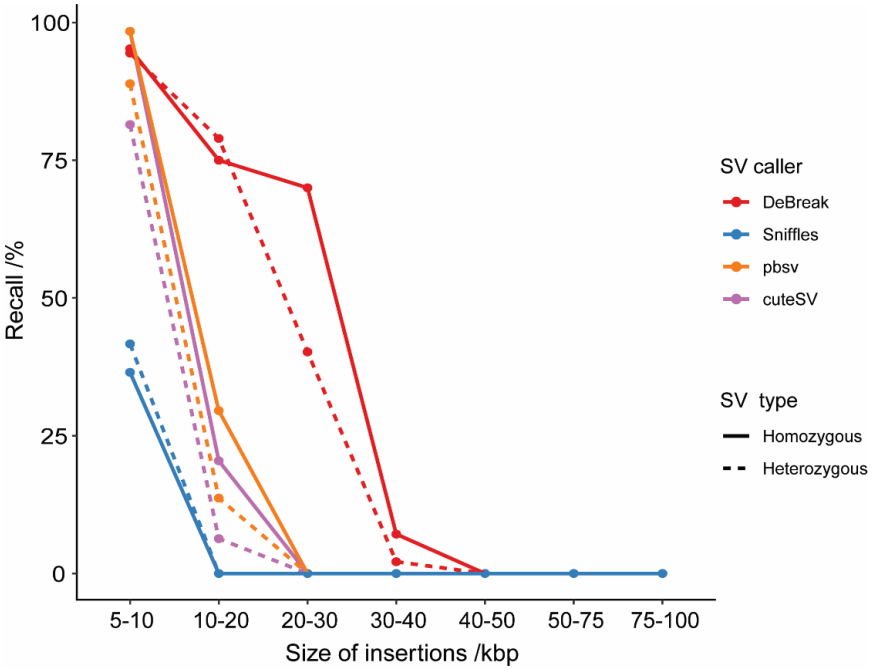

Figure S6 Large insertion detection in simulated datasets. Recall of insertion detection at different size ranges. The average length of sequencing reads was 15kbp. The maximal detectable insertion size is 10kbp for Sniffles, 20kbp for pbsv and cuteSV, and 30kbp for DeBreak.

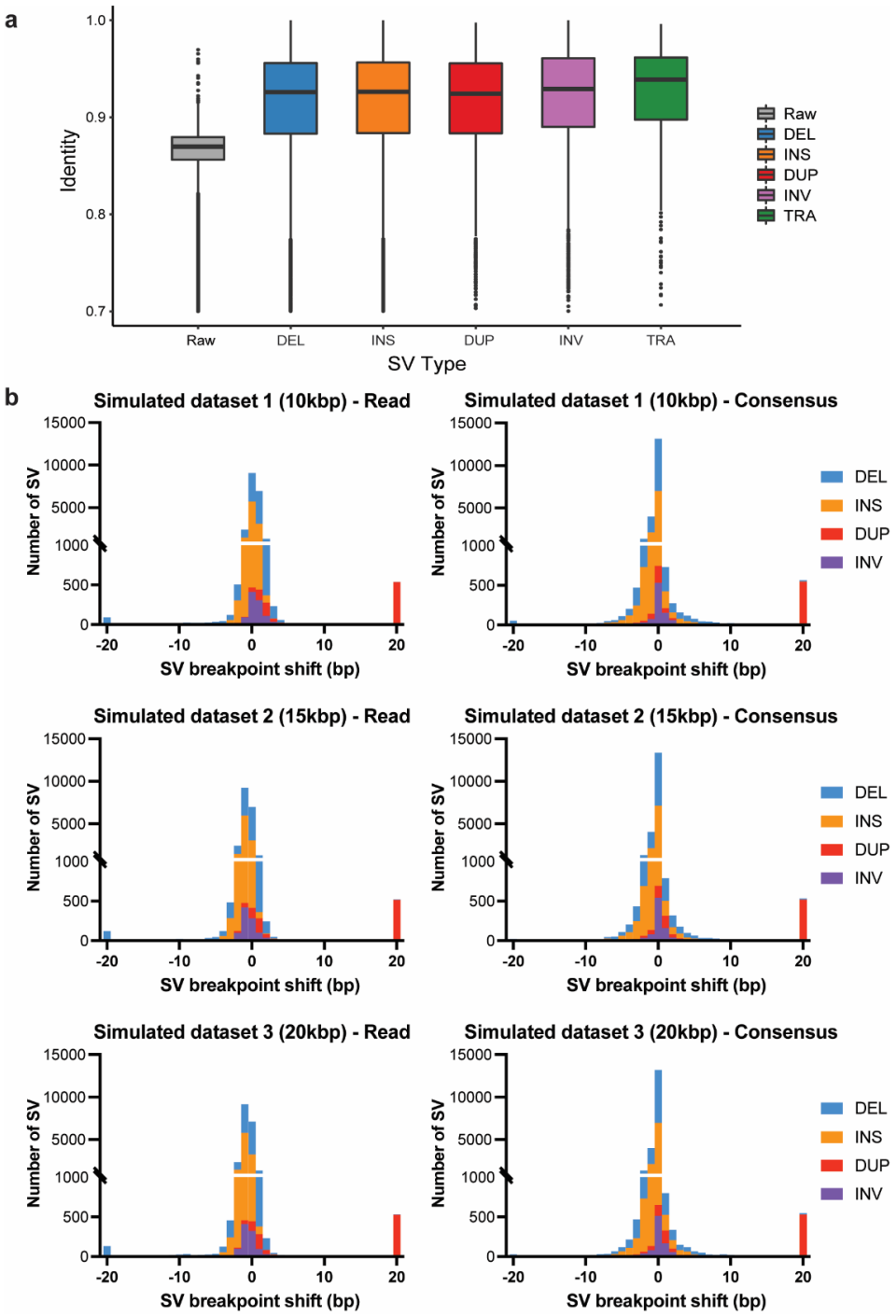

Figure S7 SV breakpoint refinement in simulated datasets. a The identity of reads/consensus sequences compared with sequences around simulated SVs for individual SV type. **b** Shift of SV breakpoints inferred from raw reads (left) and from consensus sequences (right) in three simulated datasets (top, center, and bottom). SVs with shifts more than 20bp were combined into the first and last bins.

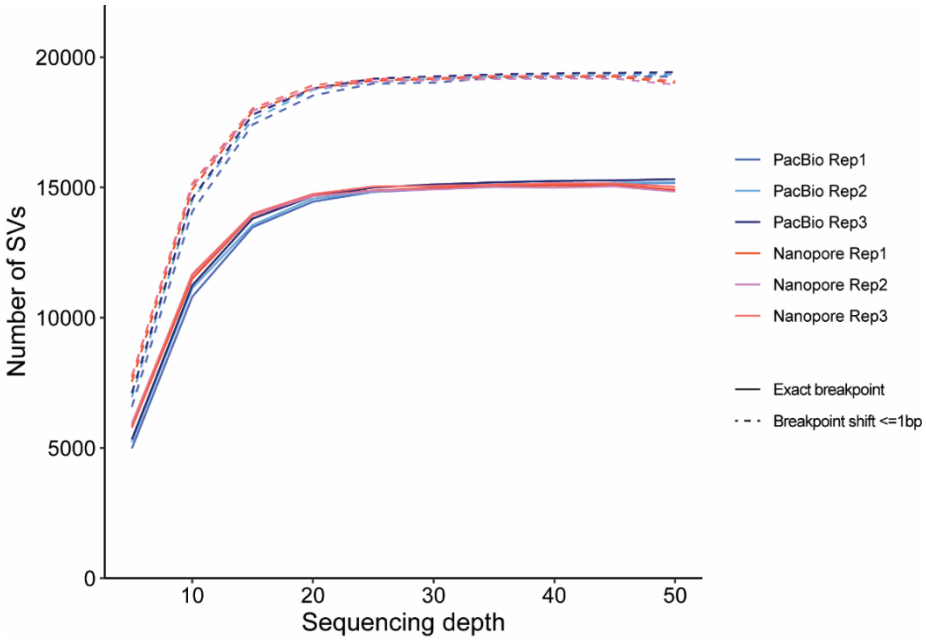

Figure S8 SV breakpoint accuracy in down-sampled simulated datasets. Number of detected SVs with exact breakpoint (solid line) and shift ≤1bp (dashed line) in three PacBio and Nanopore replicates.

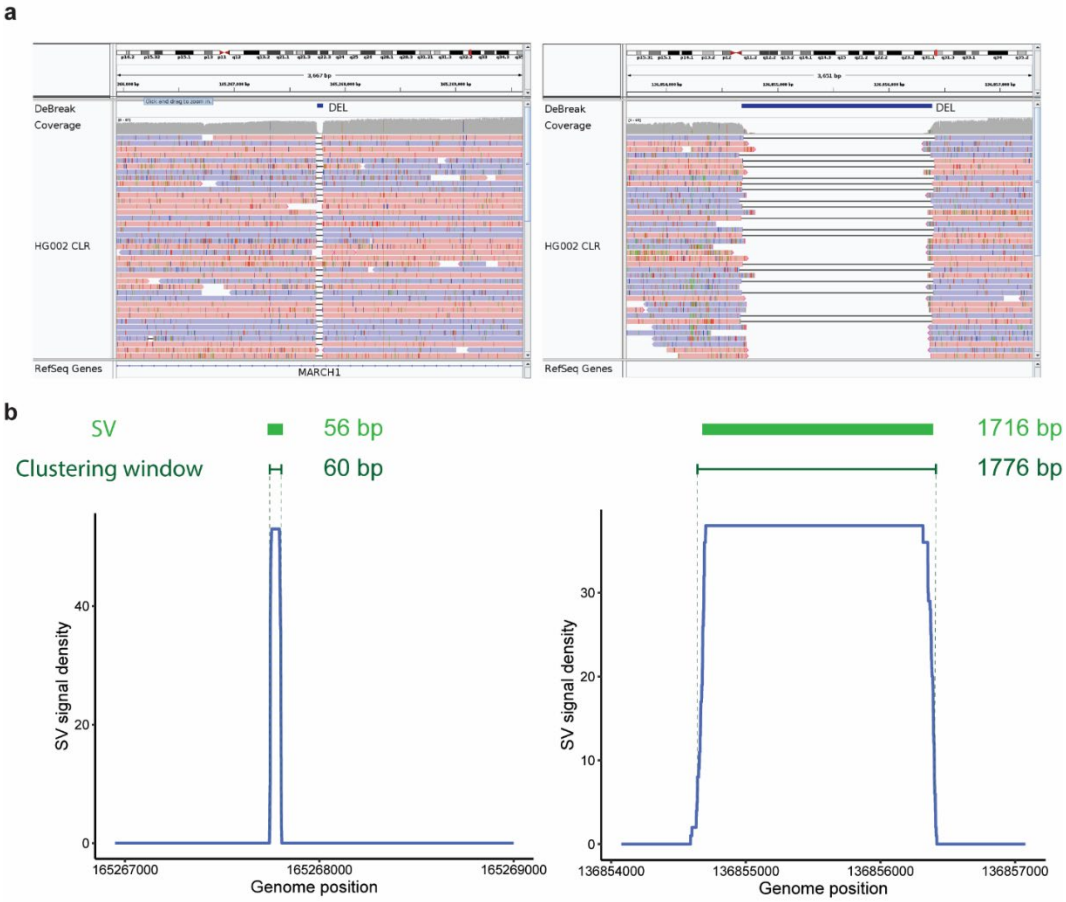

Figure S9 Adjustable clustering window size. a IGV view of read alignments flanking a small SV (left) and a large SV (right). **b** Raw SV signal density of the two SVs shown in **a**. Based on the density pattern of raw SV signals, the clustering window is smaller for shorter SVs (left) and larger for longer SVs (right).

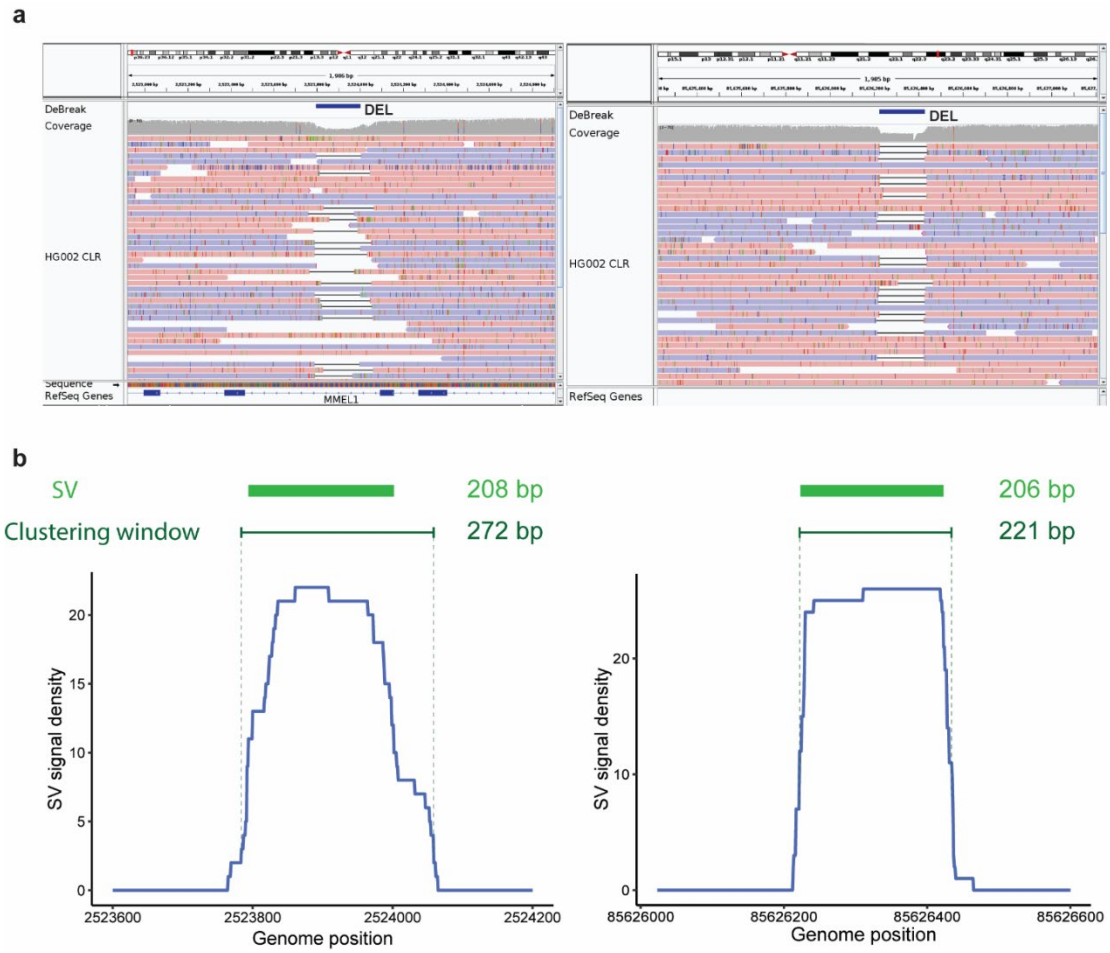

Figure S10 Adjustable clustering window for repeat regions. a IGV view of read alignments flanking SVs located within a low-complexity region (left) and a non-repeat region (right). Raw SV signals have various breakpoint positions in a low-complexity region and relatively consistent breakpoint positions in a non-repeat region. **b** Raw SV signal density of the two SVs shown in **a**. The clustering window size is larger in a low-complexity region when raw signals are diverged (left) and smaller in a non-repeat region (right) to exclude potential noise signals nearby.

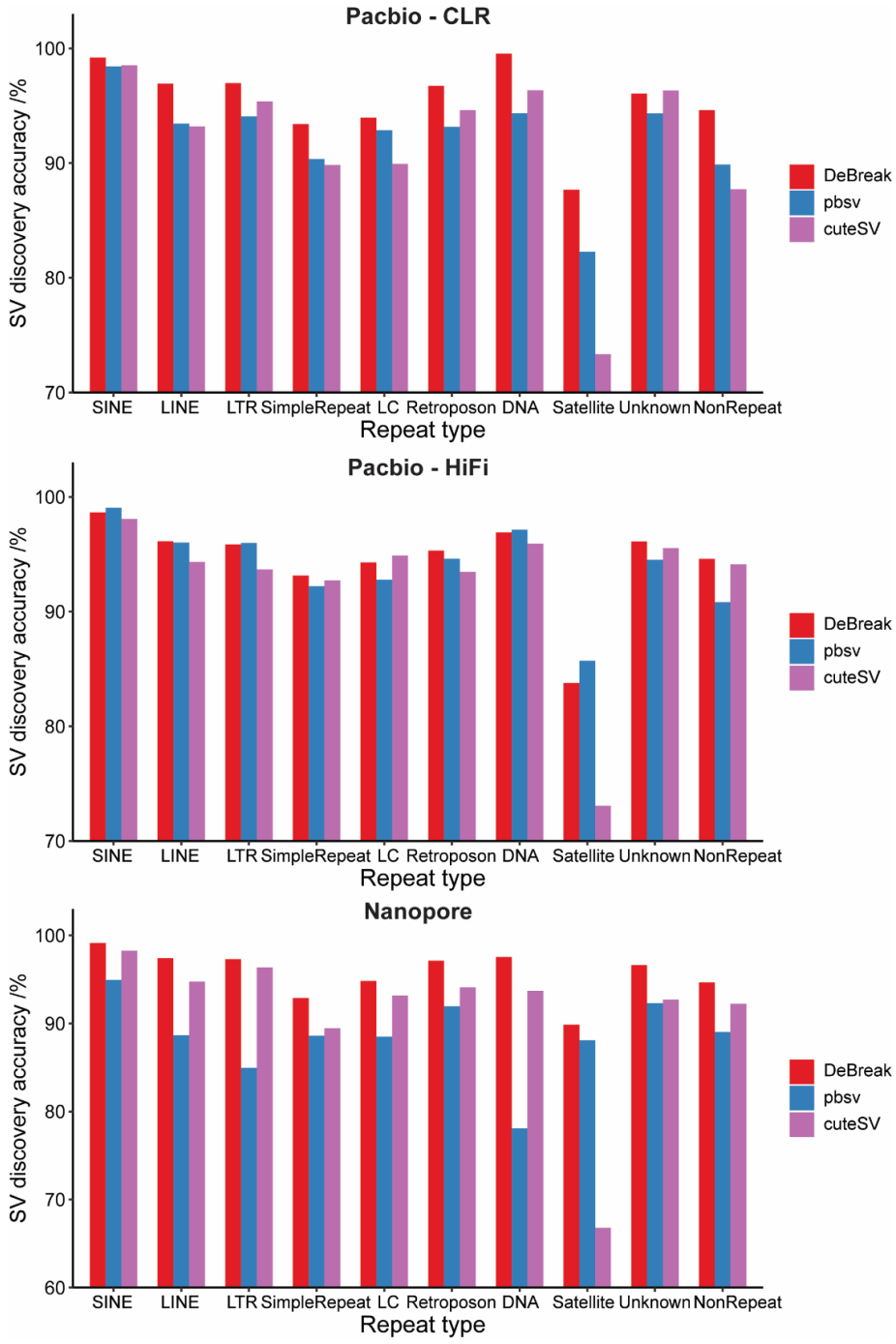

Figure S11 SV discovery accuracy in different repeat types in HG002. F1 score of SV discovery in high-confidence regions of HG002 using PacBio CLR, HiFi and Nanopore data. The repeat type was annotated with RepeatMasker using sequences of longest allele for each SV. LC, low-complexity.

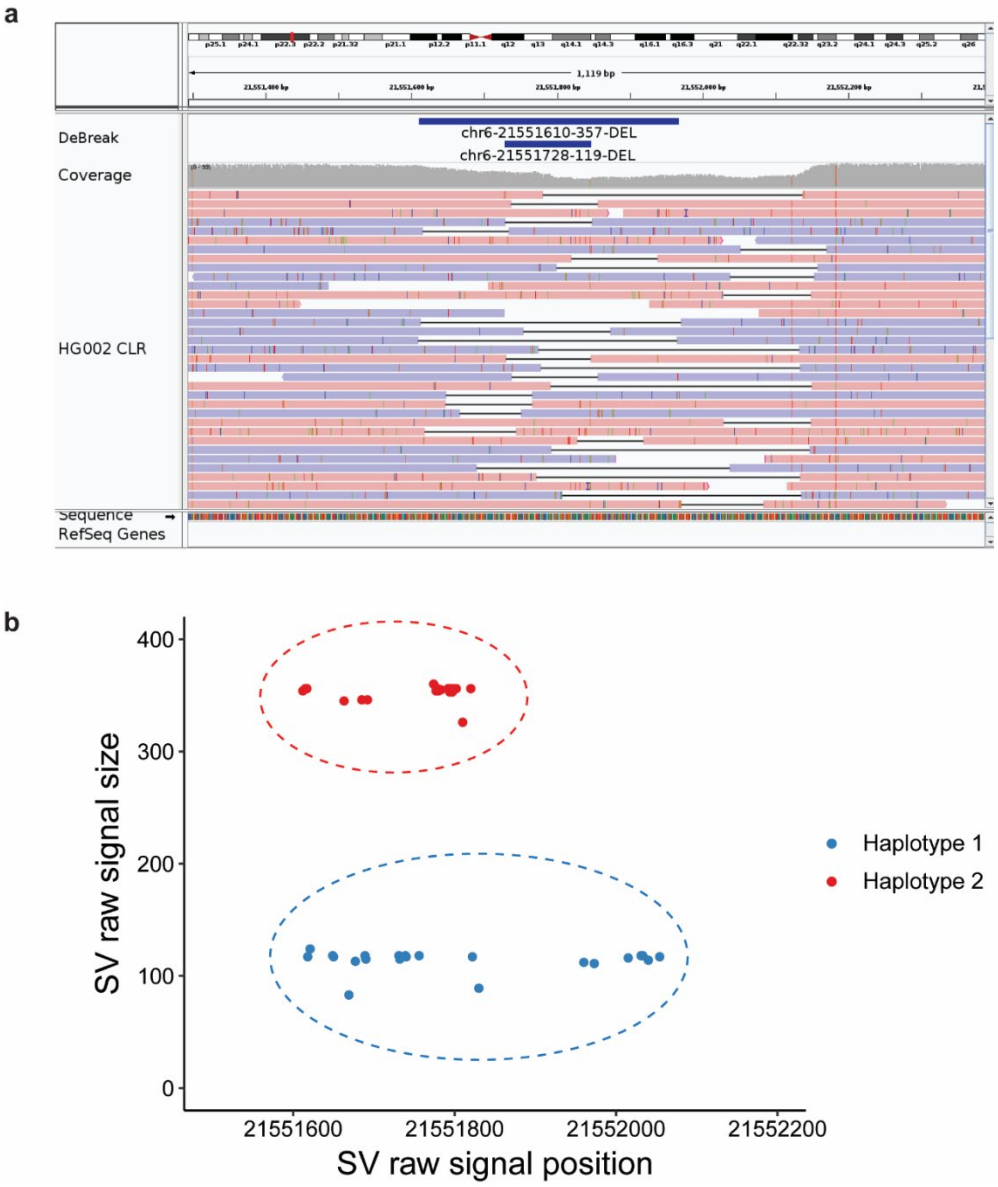

Figure S12 Example of a multi-allele SV in HG002. a IGV view of a multi-allelic SV on chromosome 6. Some reads contain ~100bp (shorter) deletion signals, and other reads contain ~300bp (longer) deletion signals. **b** k-means clustering of SV raw signals from the multi-allelic SV region shown in **a**. Based on SV size and position, SV raw signals are clustered into two groups, each representing one allele.

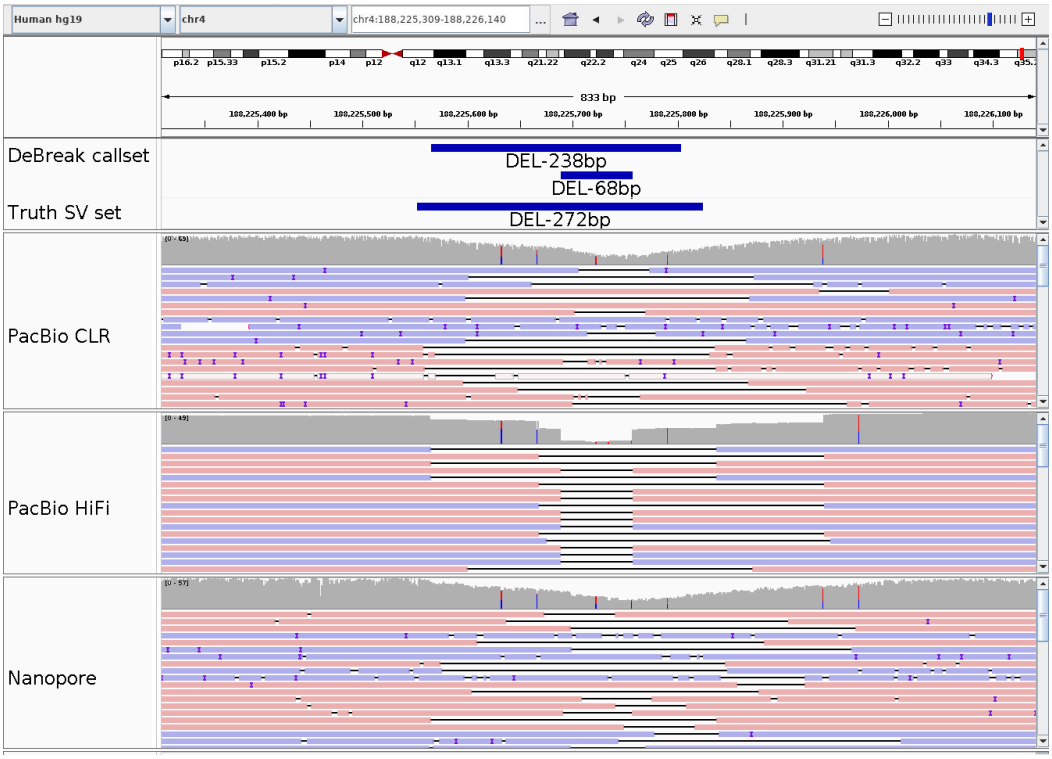

Figure S13 Example mSV in HG002 high-confidence regions. Two alternative alleles were reported by DeBreak (238bp DEL and 68bp DEL) for this mSV, and one of the two alleles matched with the truth SV set (272bp DEL). In PacBio CLR, HiFi, and Nanopore datasets, raw signals of both sizes (~250bp and ~70bp) are present in read alignments in this region.

Table S3 SV genotyping accuracy in HG002

	PacBio CLR			PacBio HiFi			Nanopore		
	DEL	**INS**	**Total**	**DEL**	**INS**	**Total**	**DEL**	**INS**	**Total**
DeBreak	**93.22**	**84.00**	**87.98**	90.46	84.70	87.21	**92.18**	86.38	**88.92**
Sniffles	47.47	38.51	42.43	52.38	45.77	48.70	61.69	56.08	58.55
pbsv	92.35	75.09	82.14	93.26	75.19	82.34	80.22	79.81	79.99
cuteSV	92.53	78.99	84.52	**91.95**	**88.56**	**90.02**	88.05	**87.45**	87.71

The genotyping accuracy is calculated as the number of SVs with the correct genotype divided by the total number of SVs reported by each SV caller. The highest genotyping accuracy in each group is shown in bold. The unit of genotyping accuracy is %.

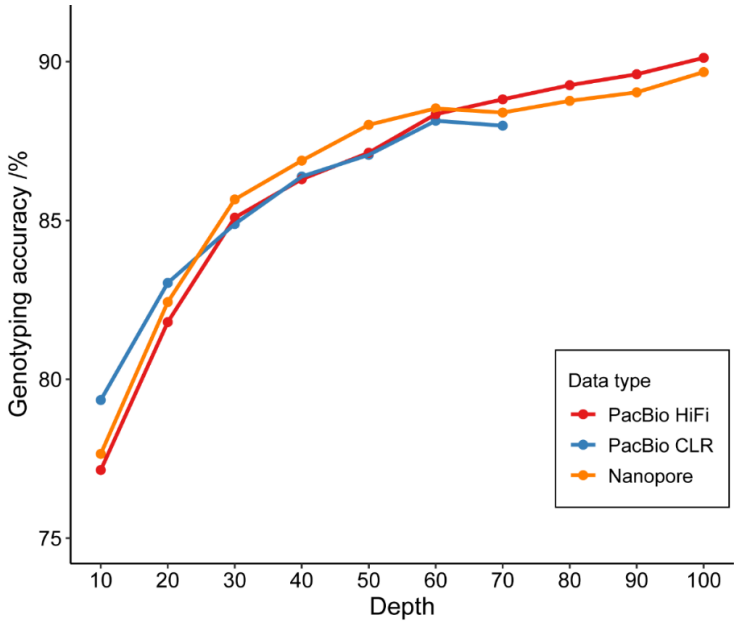

Figure S14 Genotyping accuracy in down-sampled datasets in HG002. The PacBio CLR dataset was downsampled from 10x to 70x. The PacBio HiFi and Nanopore datasets were downsampled from 10x to 100x.

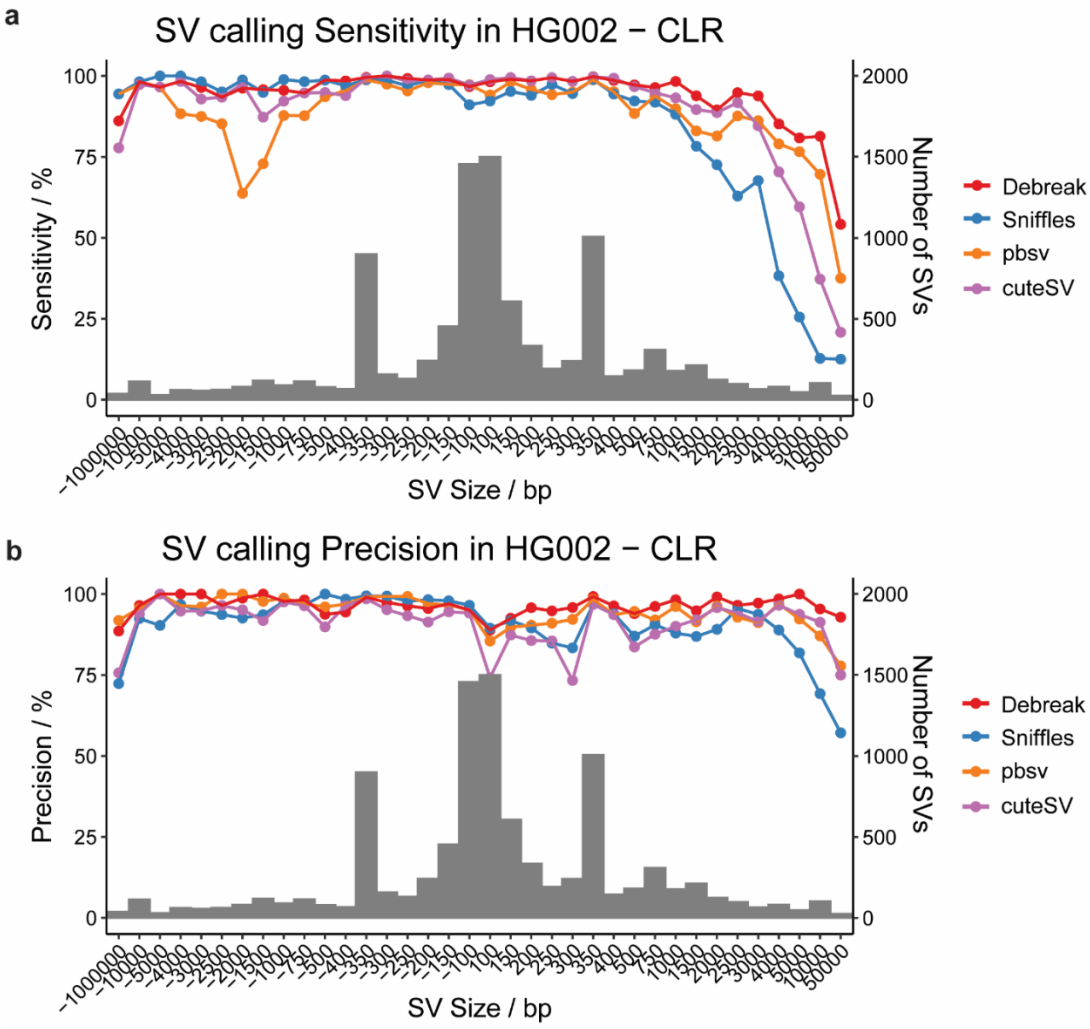

Figure S15 SV calling recall and precision for insertions (positive SV size) and deletions (negative SV size) for four tested SV callers in HG002 high-confidence regions. The bar plot indicates the number of SVs in each size range.

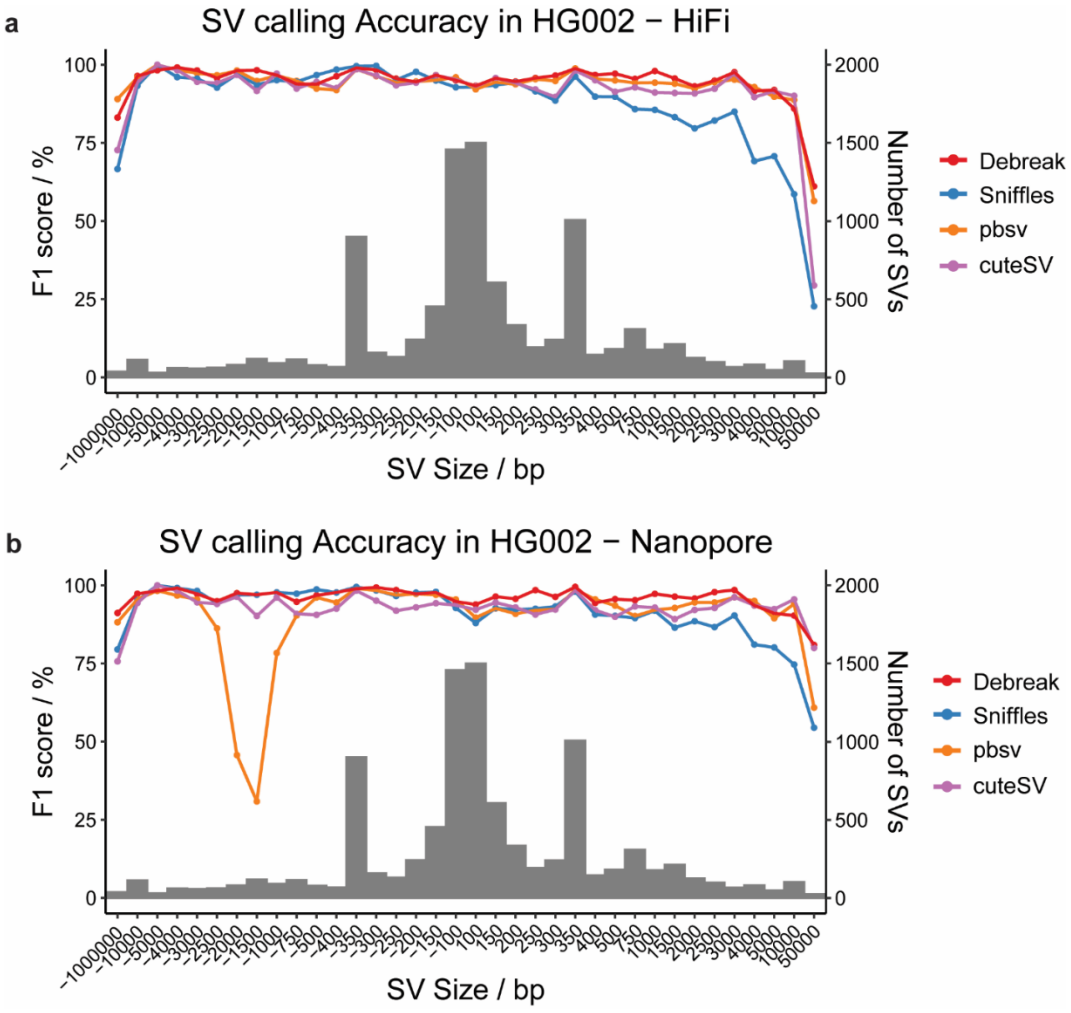

Figure S16 SV calling accuracy for insertions (positive SV size) and deletions (negative SV size) for four tested SV callers in PacBio HiFi and Nanopore datasets. The bar plot indicates the number of SVs in each size range.

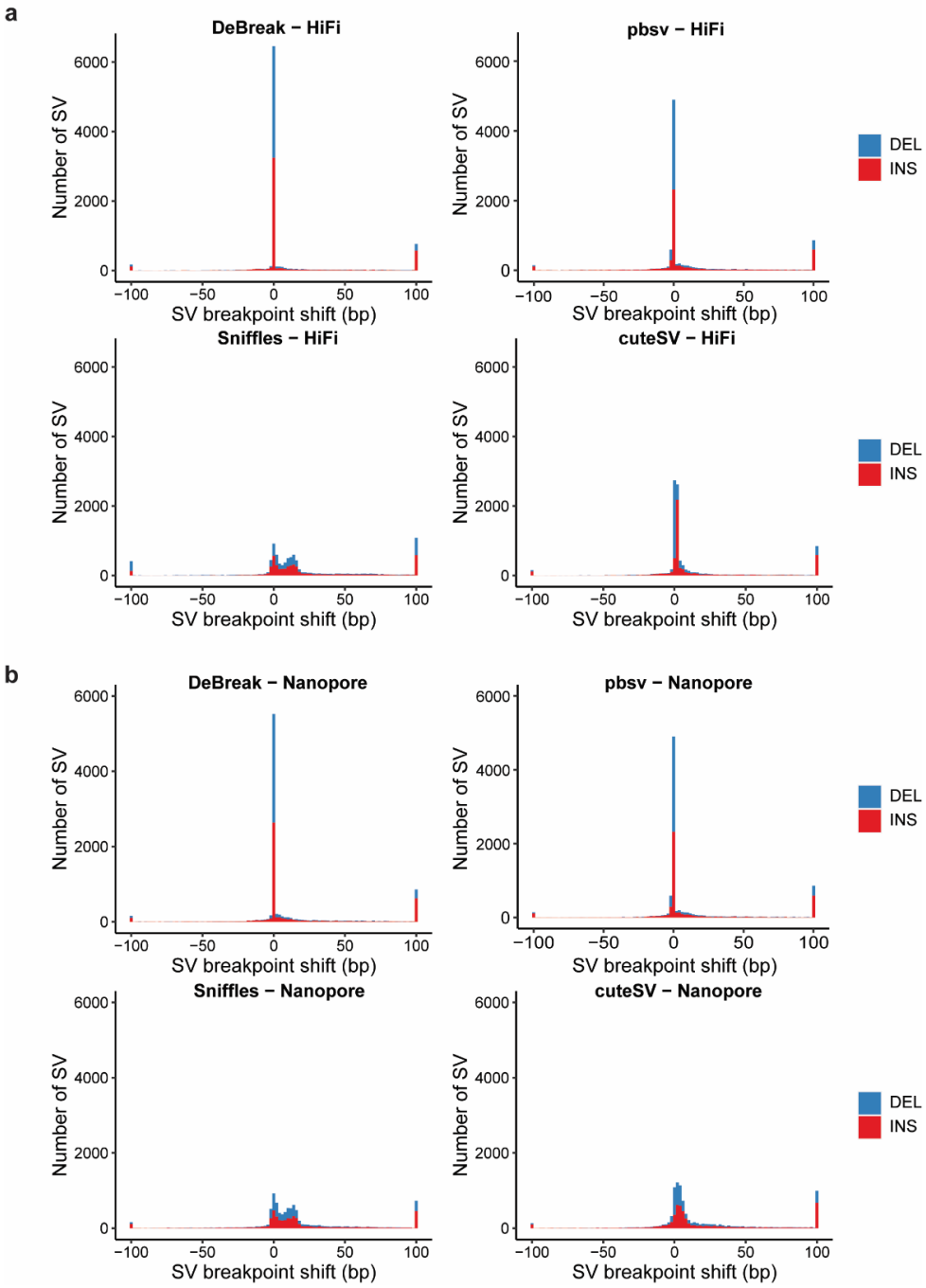

Figure S17 SV breakpoint accuracy in HG002 HiFi and Nanopore datasets. a SV breakpoint shift of four SV callers in the HiFi dataset. 64%, 57%, 5%, and 25% of SVs were identified with exact breakpoint position by DeBreak, pbsv, Sniffles, and cuteSV, respectively. **b** SV breakpoint shift of four SV callers in Nanopore dataset. 54%, 49%, 5%, and 7% of SVs were identified with exact breakpoint position by DeBreak, pbsv, Sniffles and cuteSV, respectively.

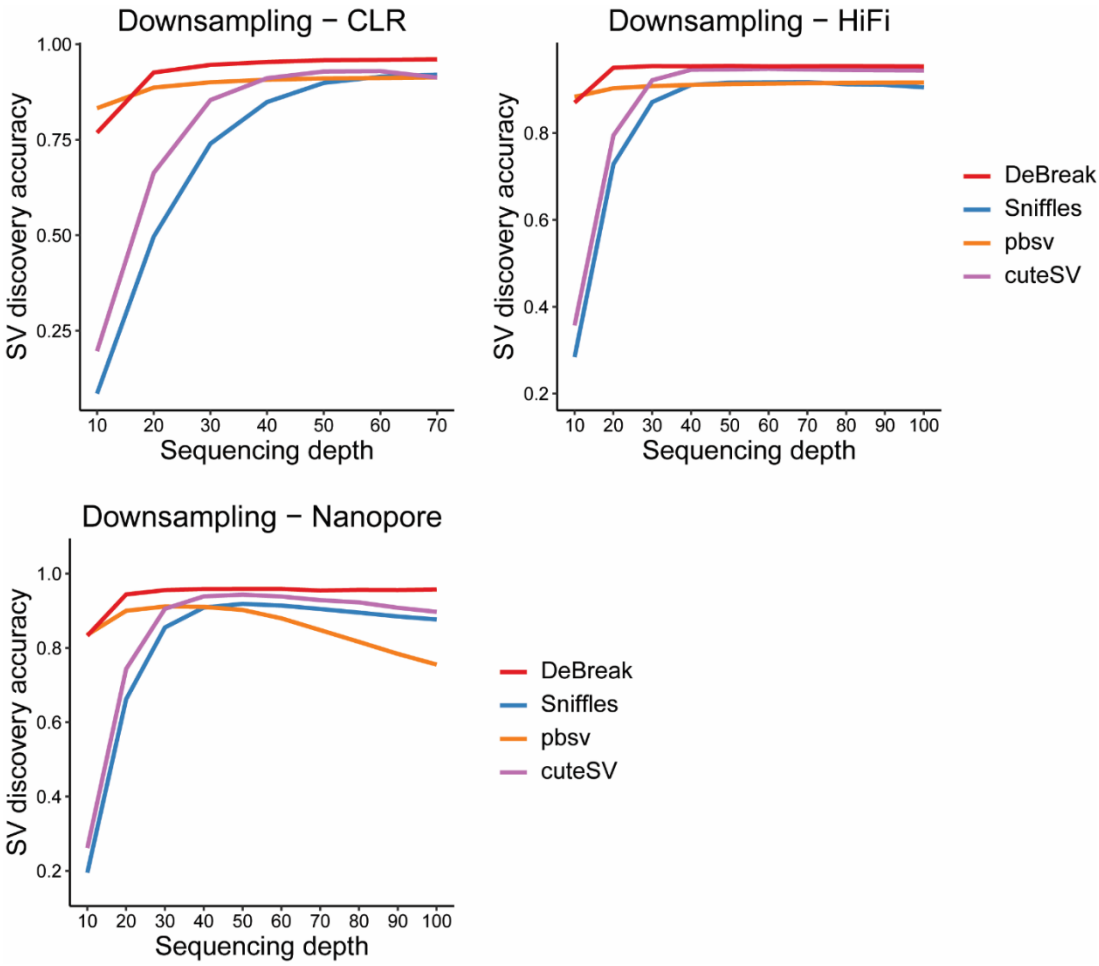

Figure S18 SV discovery accuracy in down-sampled datasets under default settings. Sniffles and cuteSV demonstrate lower accuracy at lower sequencing depths.

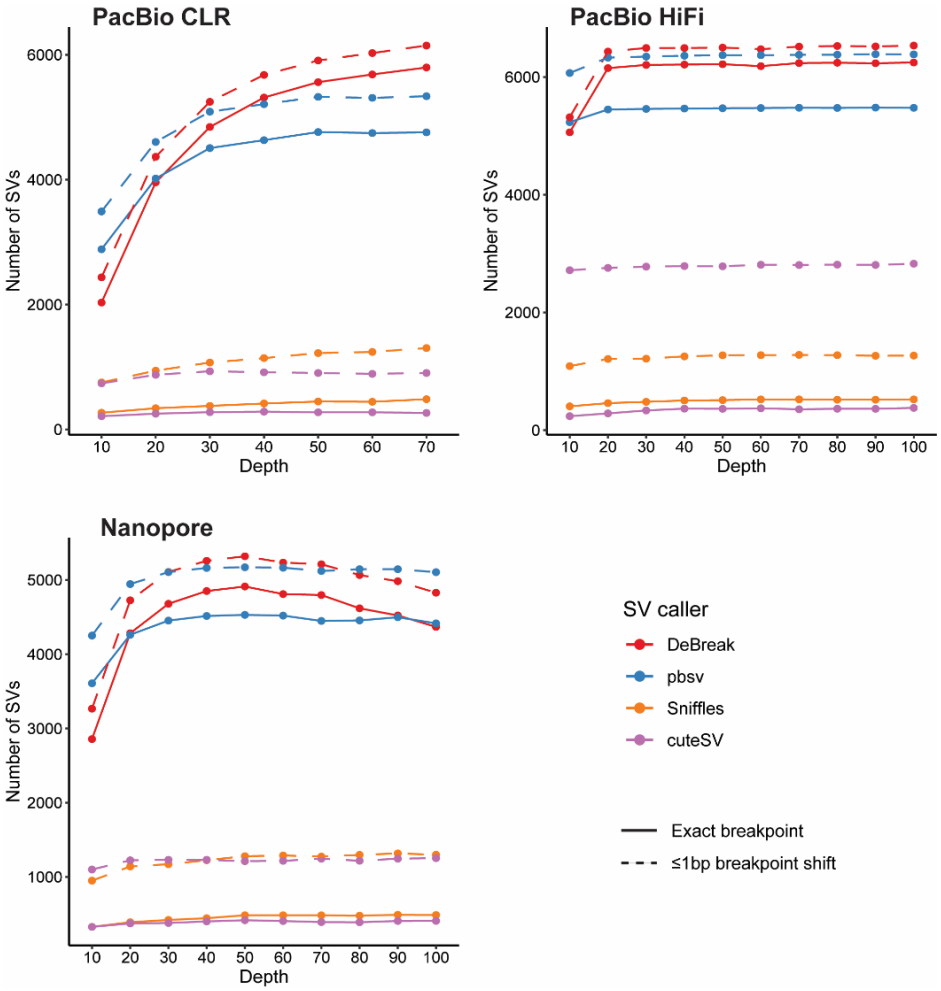

Figure S19 Breakpoint accuracy in downsampled datasets in HG002. Number of SV detected with exact breakpoint and with ≤1bp shift in breakpoints using PacBio CLR, HiFi, and Nanopore data at different sequencing depths. PacBio CLR data was downsampled from 10x to 70x. PacBio HiFi and Nanopore data were downsampled from 10x to 100x.

Table S4 SV discovery accuracy compared to assembly-based SV callsets

	DeBreak			pbsv			cuteSV			Sniffles		
Sample	**DEL**	**INS**	**Total**	**DEL**	**INS**	**Total**	**DEL**	**INS**	**Total**	**DEL**	**INS**	**Total**
CLR												
HG00096	80.52	78.59	**79.34**	76.51	69.85	72.59	78.72	77.73	78.12	78.39	69.54	73.18
HG01505	80.77	78.72	**79.52**	76.96	69.51	72.57	79.18	77.51	78.16	78.03	69.00	72.73
HG01596	80.15	75.83	**77.45**	75.46	68.72	71.47	78.28	72.10	74.39	76.76	66.80	70.69
HiFi												
HG02818	81.90	80.57	**81.12**	78.43	69.05	73.19	78.45	77.08	77.65	77.75	65.73	71.01
HG03486	82.17	81.58	**81.82**	78.56	69.30	73.40	79.71	78.65	79.09	79.14	68.22	73.01
NA12878	80.78	81.66	**81.31**	77.80	69.17	72.87	77.32	76.54	76.85	73.64	63.88	68.00

SV discovery accuracy was evaluated with the assembly-based SV callset as the ground truth. The highest accuracy (F1 score) among four tested alignment-based SV callers is shown in bold in each sample. The unit of F1 score is %.

Table S5. SV discovery recall and precision compared to assembly-based SV callsets

	DeBreak				pbsv				cuteSV				Sniffles			
	R-D	P-D	R-I	P-I	R-D	P-D	R-I	P-I	R-D	P-D	R-I	P-I	R-D	P-D	R-I	P-I
CLR																
HG00096	79.46	81.62	78.08	79.11	77.24	80.25	75.64	79.94	78.44	74.67	65.40	74.95	75.27	81.29	61.67	80.79
HG01505	79.40	82.20	76.68	80.87	77.41	81.04	74.47	80.81	78.99	75.03	64.51	75.34	75.04	81.04	59.80	82.42
HG01596	77.68	82.79	75.63	76.03	76.04	80.66	73.89	70.41	77.44	73.58	63.59	74.75	71.21	84.47	59.53	76.14
Total	**78.85**	**82.19**	**76.79**	78.64	76.90	80.65	74.66	76.79	78.30	74.43	64.50	75.01	73.86	82.18	60.33	**79.71**
HiFi																
HG02818	81.44	82.36	76.54	85.04	78.40	78.50	73.41	81.13	80.36	76.59	60.19	80.97	75.94	80.18	58.66	87.12
HG03486	83.14	81.22	79.31	83.98	81.29	78.20	77.26	80.09	81.71	75.64	60.85	80.47	78.20	79.18	60.94	86.54
NA12878	81.80	79.80	78.58	84.99	75.91	78.78	72.78	80.71	79.19	76.45	59.90	81.81	72.65	78.30	57.98	87.07
Total	**82.15**	**81.19**	**78.13**	84.65	78.71	78.46	74.57	80.62	80.50	76.21	60.33	81.04	75.79	79.28	59.26	**86.90**

SV discovery accuracy was evaluated with the assembly-based SV callset as the ground truth. The highest recall and precision among four tested alignment-based SV callers is marked in bold. The unit of recall and precision is %.
R-D, recall for deletion. P-D, precision for deletion. R-I, recall for insertion. P-I, precision for insertion.

Table S6 SV genotyping accuracy in HGSVC samples

	DeBreak	pbsv	cuteSV	Sniffles
CLR				
HG00096	72.70	67.81	**75.06**	45.88
HG01505	73.18	67.82	**75.82**	46.13
HG01596	**70.04**	65.11	68.75	39.05
HiFi				
HG02818	72.62	72.91	**74.91**	45.45
HG03486	71.75	72.54	**74.13**	45.25
NA12878	71.68	70.43	**73.52**	39.06

SV genotyping accuracy was assessed with the assembly-based SV callset as the ground truth. The highest accuracy among four tested alignment-based SV callers is shown in bold for each sample. The unit of genotyping accuracy is %.

Table S7 DeBreak mSV discovery in HGSVC samples

	mSV	Alternative allele	Validation rate	mCNV
CLR				
HG00096	1097	2194	73.38	12
HG01505	1011	2022	71.96	4
HG01596	992	1984	70.46	7
Total	3100	6200	71.98	23
HiFi				
HG02818	1031	2062	73.81	7
HG03486	1187	2374	72.11	10
NA12878	879	1758	67.86	7
Total	3097	6194	71.47	24

Alternative alleles that are also reported in assembly-based SV callset were considered as validated. mCNV is classified using k-mer counts. The unit of validation rate is %.

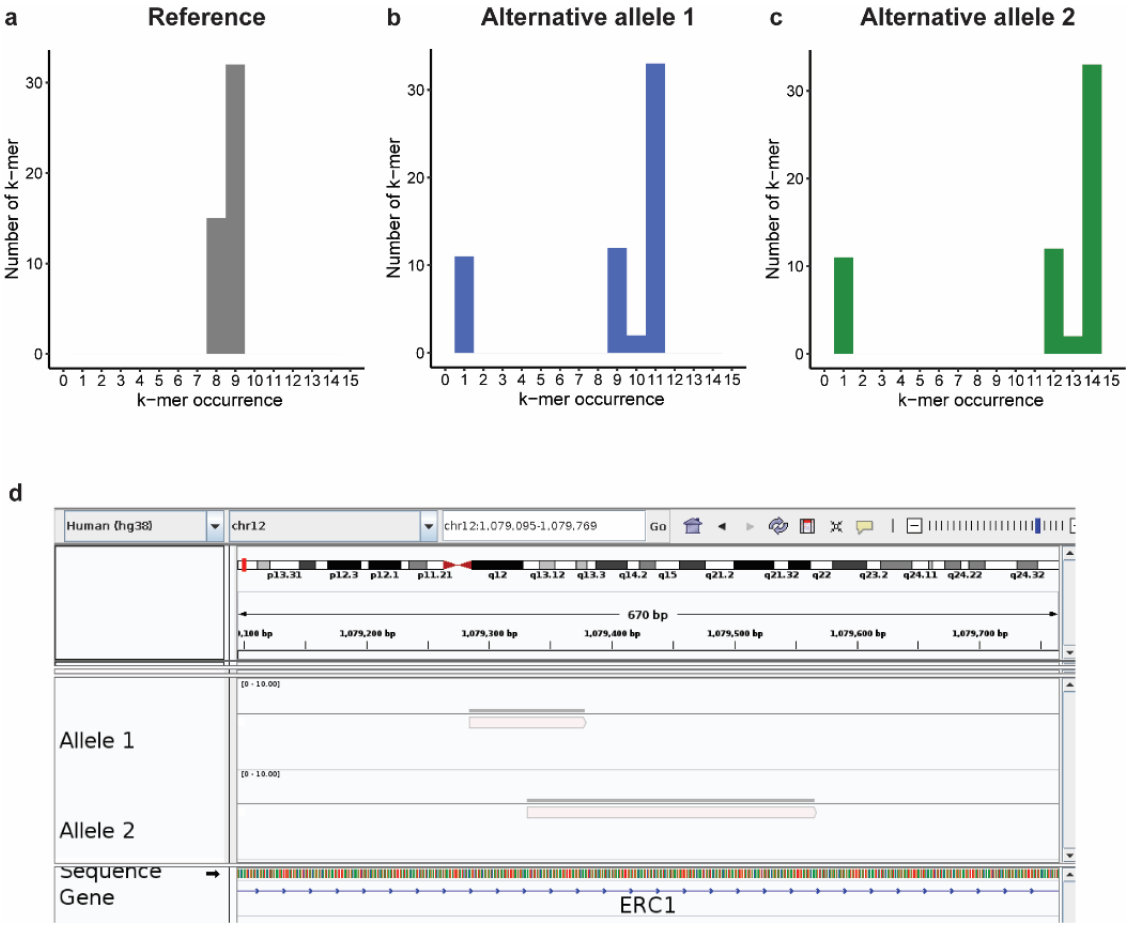

Figure S20 Example an mCNV event (chr12-1079285-94bp-INS/chr12-1079285-235bp-INS) in HG02818. a k-mer occurrence of reference sequences. A peak of k-mer at 9 indicates 9 copies of a specific repeat unit in the reference genome. **b, c** k-mer occurrence of alternative allele 1 (**b**, INS of 94bp) and alternative allele 2 (**c**, INS of 235bp). 11 and 14 copies of the repeat unit are present in the two alleles. **d** Alignment of inserted sequences of two alternative alleles. Both alleles can be fully aligned to the reference genome near the mSV breakpoint.

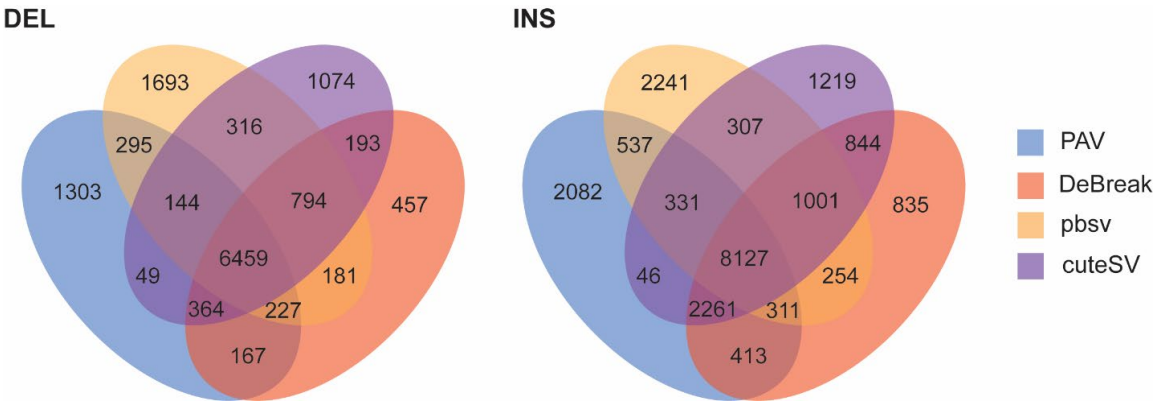

Figure S21 SV discovery consistency between alignment-based and assembly-based SV discovery approaches. Venn diagrams showing the overlap between alignment-based and assembly-based SV callsets. Numbers indicate the number of SVs in each group.

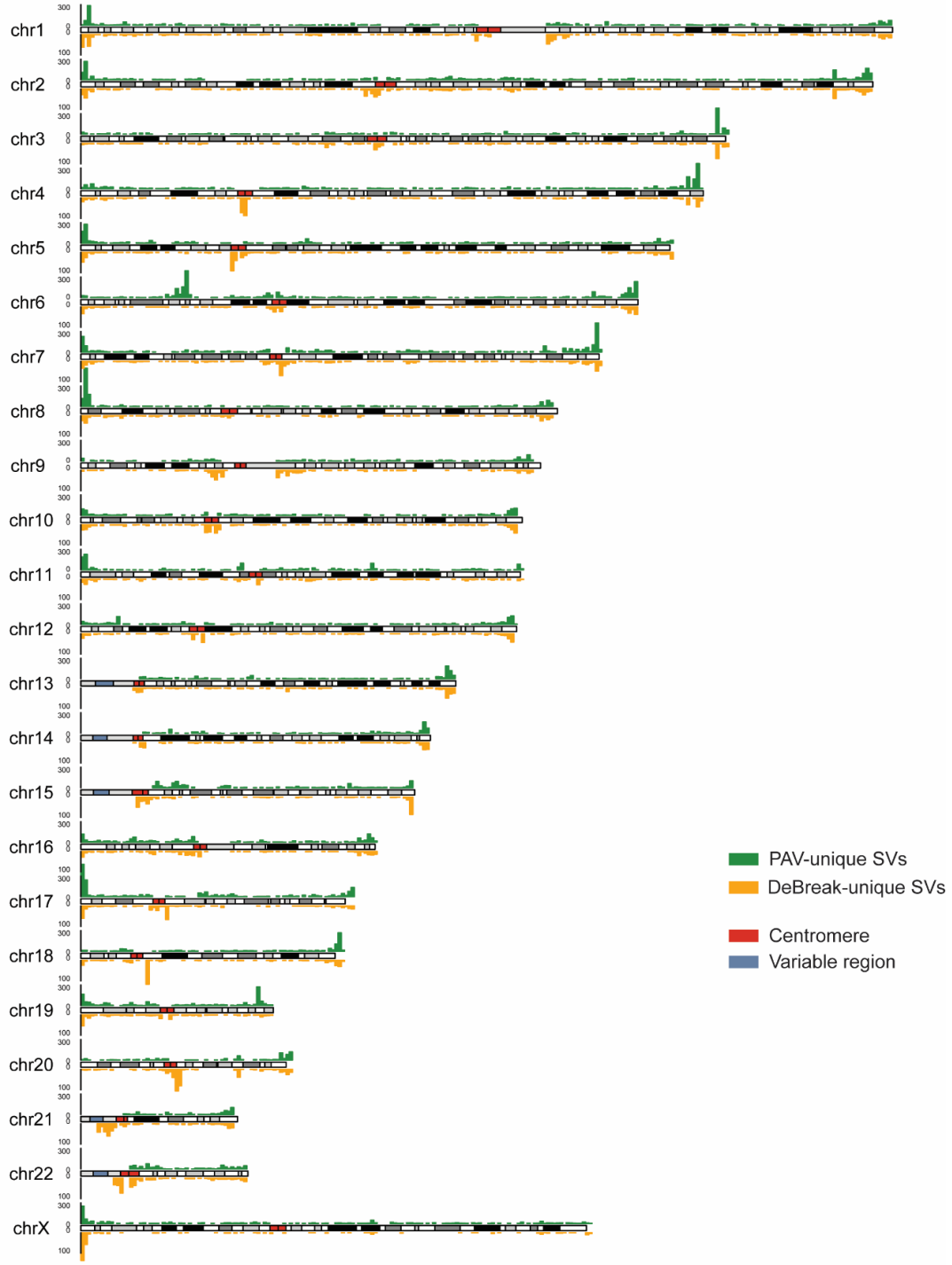

Figure S22 PAV-unique and DeBreak-unique SV distribution on the genome. The PAV-unique SVs (green) are enriched at the telomere regions. DeBreak-unique SVs (orange) are enriched at the centromere and telomere regions.

Table S8 SV discovery accuracy in CHM13

	Deletion			Insertion			Total		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
CLR									
DeBreak	82.41	88.58	**85.38**	79.29	87.19	83.05	80.45	87.72	**83.93**
Sniffles	78.66	86.01	82.17	63.03	90.42	74.28	68.86	88.49	77.45
pbsv	81.84	84.76	83.27	70.74	85.52	77.43	74.88	85.21	79.71
cuteSV	83.84	86.35	85.08	81.44	84.74	**83.06**	82.33	85.35	83.81
HiFi									
DeBreak	83.71	86.12	**84.90**	80.81	89.89	**85.11**	81.89	88.39	**85.02**
Sniffles	79.44	83.84	81.58	60.42	88.75	71.89	67.51	86.50	75.84
pbsv	80.69	83.18	81.92	65.04	88.26	74.89	70.88	86.01	77.71
cuteSV	85.93	79.68	82.69	84.03	85.07	84.55	84.74	82.97	83.84
Nanopore									
DeBreak	85.72	82.83	**84.25**	83.46	87.70	**85.53**	84.30	85.76	**85.03**
Sniffles	81.87	78.57	80.19	64.49	88.54	74.63	70.97	83.94	76.91
pbsv	84.44	54.55	66.28	68.12	86.07	76.05	74.21	69.06	71.54
cuteSV	88.52	71.71	79.23	86.09	83.24	84.64	87.00	78.41	82.48

SV discovery accuracy was evaluated with the assembly-based SV callset as the ground truth. The highest F1 score in each SV type are shown in bold. The unit of recall, precision, and F1 score is %.

Table S9 SV genotyping accuracy in CHM13

	CLR	HiFi	Nanopore
DeBreak	77.00	**86.02**	**78.74**
Sniffles	32.03	26.19	48.73
pbsv	62.65	81.31	59.33
cuteSV	**77.33**	74.36	61.50

SV genotyping accuracy was evaluated with only 'GT=1/1' as correct genotype. The highest genotyping accuracy in each data type are shown in bold. The unit of genotyping accuracy is %.

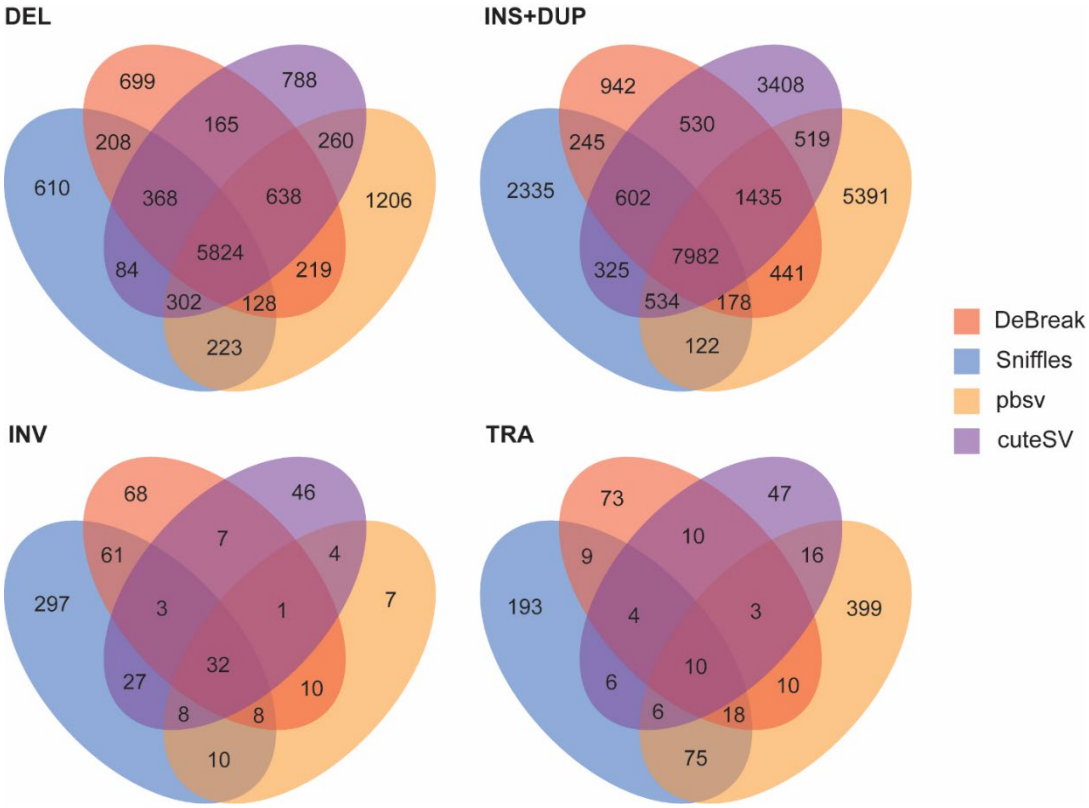

Figure S23 SV discovery in SKBR3 cell line. Venn diagrams showing the overlap between SV callsets from four SV callers. Insertions and duplications have been merged for comparison, as duplications are sometimes considered as insertions.

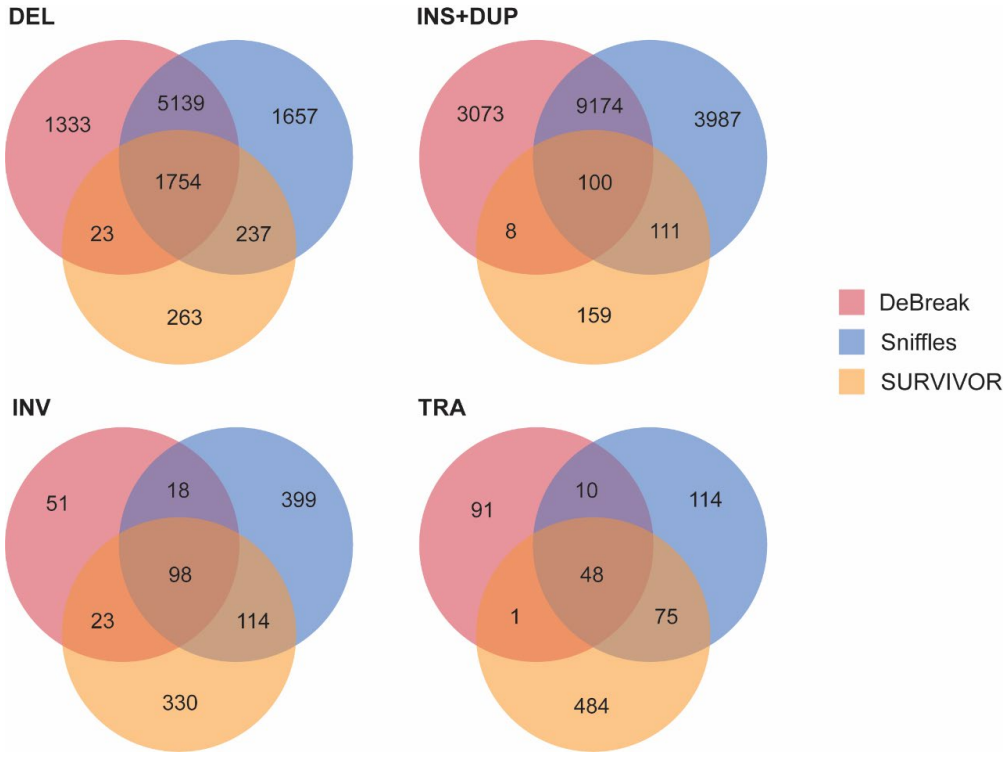

Figure S24 SV discovery in SKBR3 cell line. Venn diagrams showing the overlap between DeBreak SV callset and SV calls previously reported from short-read (SURVIVOR) and long-read data (Sniffles). Numbers indicate the number of SVs in each category.

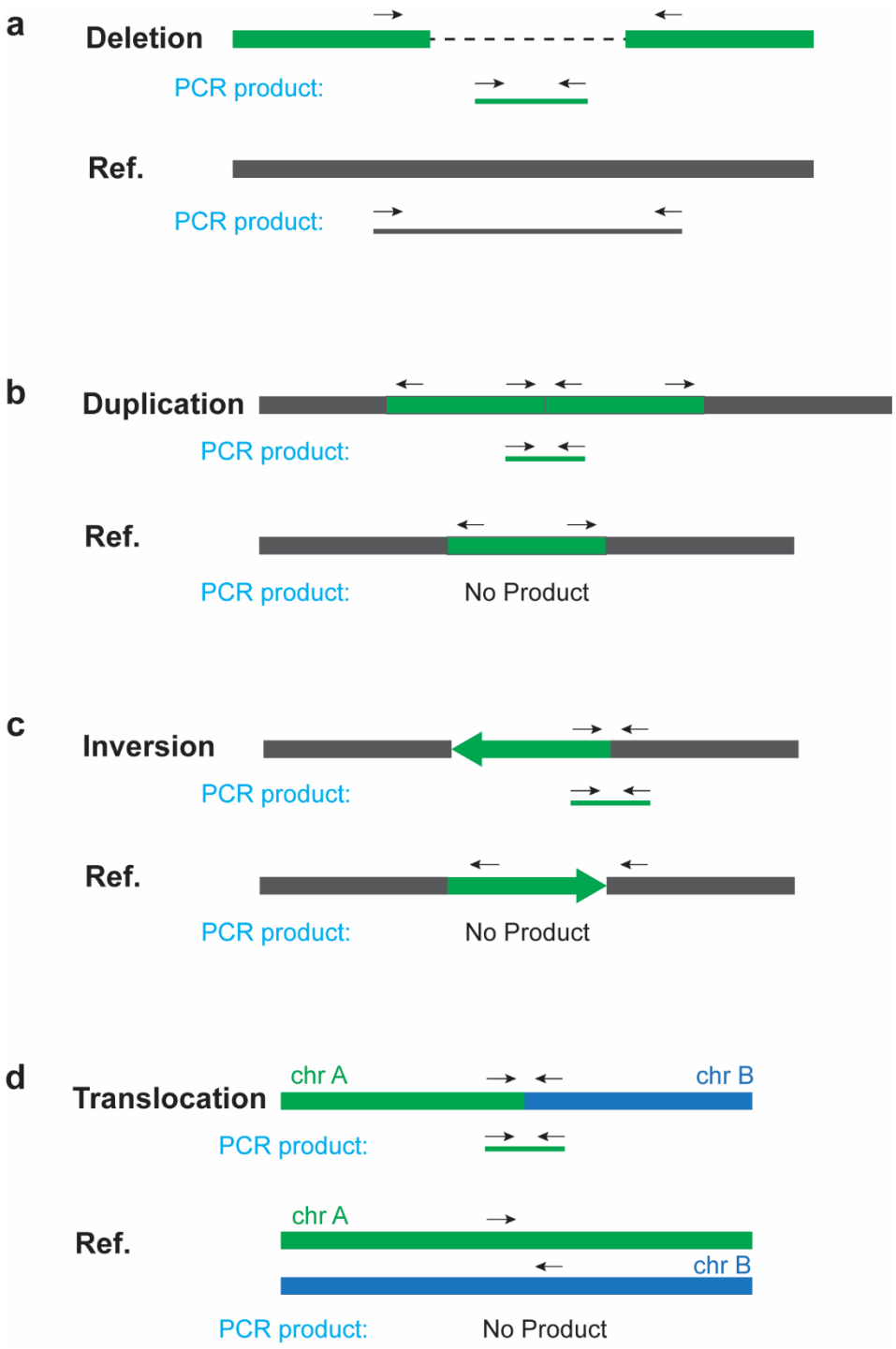

Figure S25 PCR primer for validation. PCR primer design for deletions (**a**), duplications (**b**), inversions (**c**), and translocations (**d**). For deletions, the PCR product size is much smaller when the SV is a true event than for false positive (reference allele). For duplications, inversions, and translocations, the PCR product is expected only when SV is a true event.

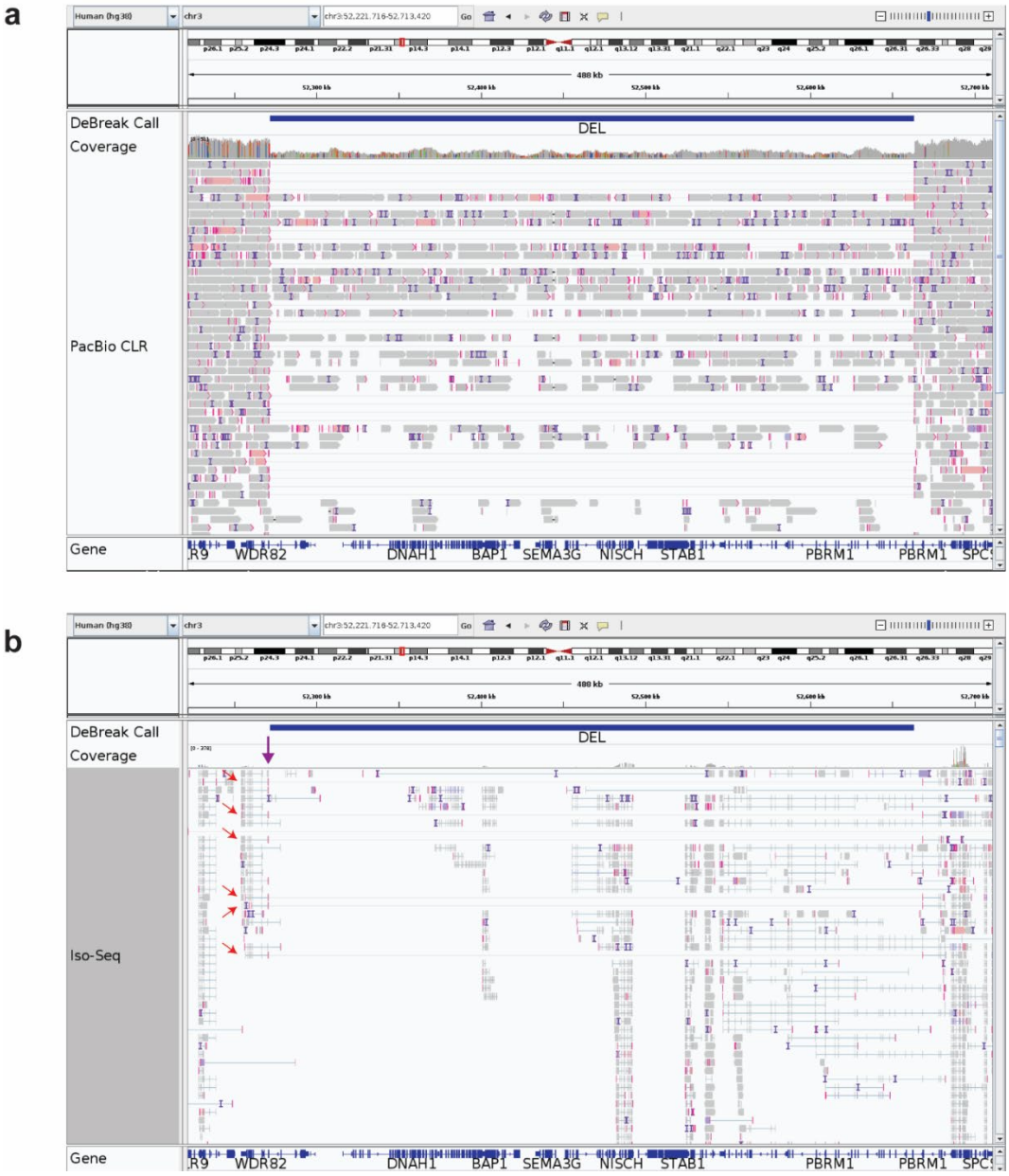

Figure S26 Gene fusion of WDR82 and PBRM1. IGV view of PacBio CLR (**a**) and Iso-Seq data (**b**) at the gene fusion junction. The 'DeBreak Call' panel shows SVs identified from PacBio data with DeBreak. Red arrows indicate IsoSeq reads that contain sequences from both WDR82 and PBRM1. The purple arrow indicates the gene fusion junction position inferred from IsoSeq reads.

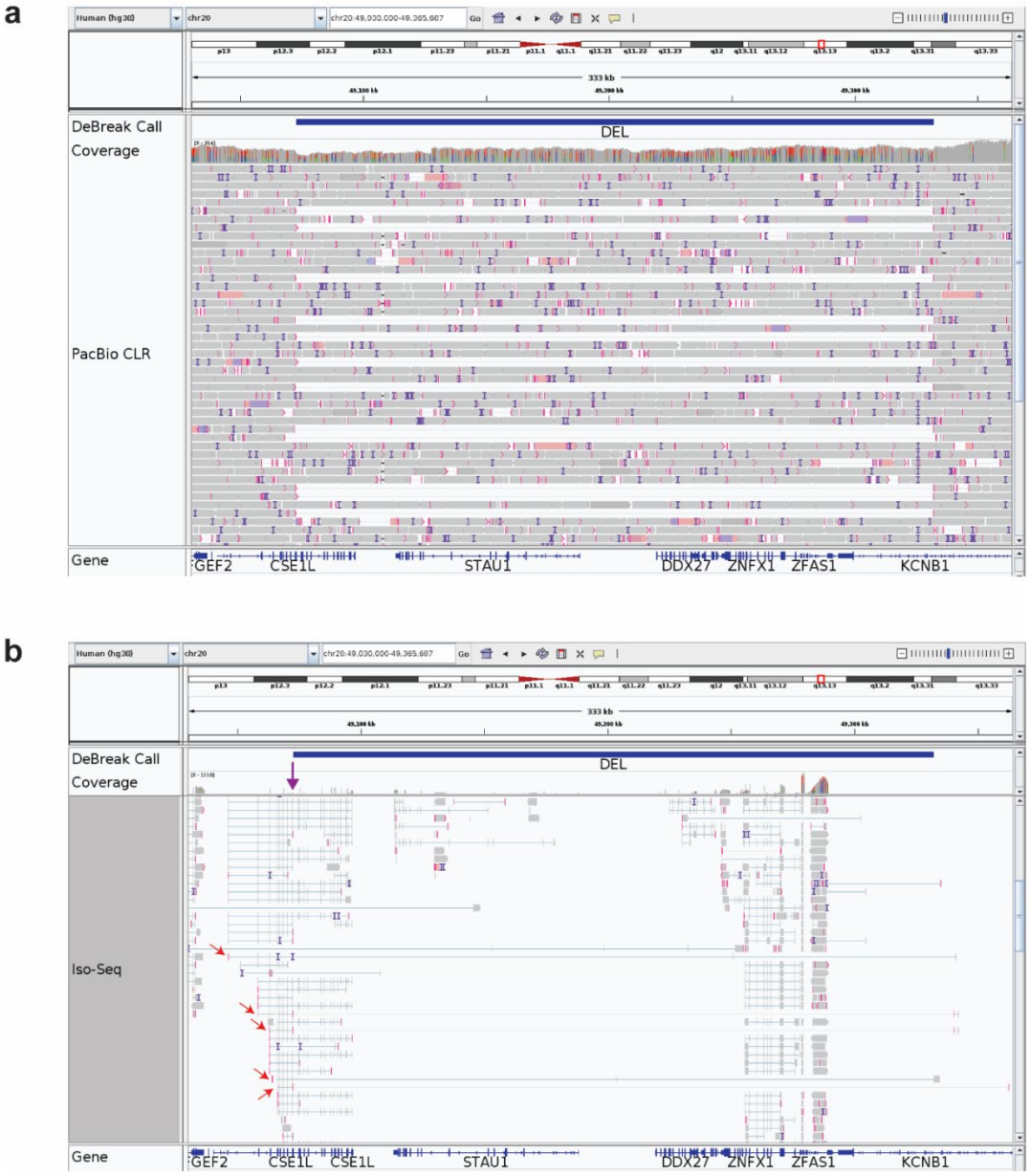

Figure S27 Gene fusion of CSE1L and KCNB1. IGV view of PacBio CLR (**a**) and Iso-Seq data (**b**) at the gene fusion junction. The 'DeBreak Call' panel shows SVs identified from PacBio data with DeBreak. Red arrows indicate the IsoSeq reads that contain sequences from both CSE1L and KCNB1. The purple arrow indicates the gene fusion junction position inferred from IsoSeq reads.

Table S10 Runtime and memory usage of SV callers

	DeBreak	Sniffles	pbsv	cuteSV
CPU x node	12 x 1	12 x 1	12 x 1	12 x 1
Wall-clock time	12:22:48	03:02:37	1-21:06:15	01:29:37
CPU time	1-19:11:22	1-08:28:26	1-22:42:03	08:10:08
Peak Memory (GB)	63.00	12.62	71.84	3.39

ACCURATE LONG-READ *DE NOVO* ASSEMBLY EVALUATION WITH INSPECTOR

by

YU CHEN, YIXIN ZHANG, AMY Y. WANG, MIN GAO, ZECHEN CHONG

Genome Biology

Format adapted and errata corrected for dissertation

ABSTRACT

Long-read *de novo* genome assembly continues to advance rapidly. However, there are a lack of effective tools to accurately evaluate the assembly results, especially for structural errors. We present Inspector, a reference-free long-read *de novo* assembly evaluator which faithfully reports types of errors and their precise locations. Notably, Inspector can correct the assembly errors based on consensus sequences derived from raw reads covering erroneous regions. Based on *in silico* and long-read assembly results from multiple long-read data and assemblers, we demonstrate that in addition to providing generic metrics, Inspector can accurately identify both large-scale and small-scale assembly errors.

INTRODUCTION

Whole genome *de novo* assembly is essential for investigating species without reference genomes and is critical for characterizing the full spectrum of genetic variants for species with a reference genome [1-9]. With the advancement of long-read sequencing technologies, long reads are becoming more accurate, much longer, and more affordable [10, 11]. Accordingly, numerous long-read whole-genome *de novo* assemblers [12-20] have been developed and are widely applied to small-scale [21-23] and consortium projects [4, 5, 24].

Despite these advancements, it is challenging to achieve high-quality assembly, even for long reads. The algorithms of assemblers differ greatly, and each assembler typically includes a wide range of parameters. Moreover, the input data may originate from individual or multiple platforms with varying read lengths. For long-read assemblers, the input may include hybrid reads, long noisy reads (PacBio raw CLR or Nanopore), HiFi reads, reads from trio samples, and other types. Additional complexity due to ploidy, genetic diversity, heterozygosity, repetitive sequences, as well as sequencing depth of sequenced genomes make *de novo* assembly even more challenging.

De novo assembly quality assessment is therefore essential both for users to obtain optimal assembly results and for developers to improve assembly algorithms. In the short-read era, Assemblathon [25, 26] guided best practices for *de novo* assembly. However, there are limited toolsets that can evaluate long-read assemblies. QUAST-LG [26, 27], an extension of QUAST [28], is able to evaluate large genome assemblies. It

accepts sequencing data from multiple platforms and can generate reports with rich assembly metrics as well as plots. However, QUAST-LG relies heavily on existing reference genomes, which limits its application in species without a reference genome or for samples that differ substantially from reference genomes. In addition, the mis-assembly evaluation of QUAST-LG is easily affected by the presence of genetic variants. Although it accepts raw reads as input, only Illumina data will be used to call structural variations (SVs) with GRIDSS [29], while long reads can only be used to report simple read statistics. Even if short reads are provided, due to the insufficiency of detecting SVs from short reads [4], it is challenging to evaluate assembly errors.

Merqury [30], inspired by KAT [31], is a reference-free toolkit for evaluating assembly quality (QV), completeness, and phasing based on the *k*-mer spectra. By comparing *k*-mers in assemblies to raw reads, Merqury can estimate base-level accuracy and completeness. Nevertheless, Merqury requires high-accuracy reads as input, such as Illumina data, which limits its application on long-read-only assembly results. While it provides base-level error estimates, Merqury cannot explicitly validate structural errors.

BUSCO [32] is a rapid and accurate method for assessing genome assembly and annotation completeness based on evolutionary ortholog genes. However, BUSCO evaluates conserved genomic regions and is not informative on the most divergent sequences in the genome.

Assembly polishing following *de novo* assembly is a typically used approach for improving assembly quality for downstream genomic analysis. Most current polishing algorithms correct assembly errors based on read-to-assembly alignment, as used in Racon [33], Pilon [34], GCpp [35], and CONSENT [36]. Other algorithms use *k*-mer

based approaches, such as POLCA [37] and ntEdit [38]. Nanopolish [39] and Medaka [40] polishing methods have been designed particularly for Oxford Nanopore data. Most polishing methods target small-scale errors for correction, while polishing performance on a larger scale remains unknown due to a lack of efficient evaluation methods. Another limitation is that, these polishing methods often require excessive computational resources for large genomes, such as mammal genomes.

We have developed Inspector to comprehensively evaluate assembly quality and identify assembly errors in haploid and diploid genomes. Instead of relying on reference genomes, Inspector evaluates assemblies with only third-generation sequencing reads, which are the most faithful representations of target genomes. By aligning sequencing reads to the contigs with minimap2 [41], Inspector generates read-to-contig alignment and performs downstream assembly evaluation (**Fig. 1**). Statistical analysis is initially performed to assess contig continuity and completeness. Structural assembly errors and small-scale assembly errors are identified from read-to-contig alignment and distinguished from genetic variants based on the ratio of error-containing reads. Inspector includes a targeted error-correction module that addresses identified errors to improve local assembly quality. The output of Inspector includes an evaluation summary report, list of structural errors, list of small-scale errors, read alignment file, and evaluation plots.

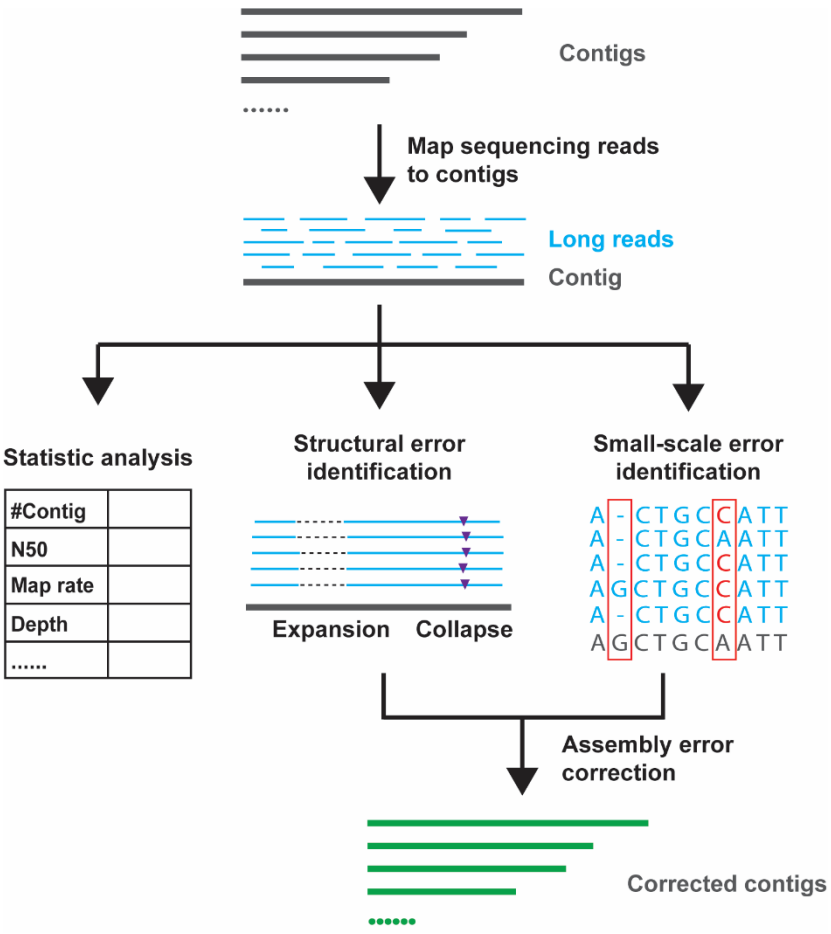

Figure 1 Inspector workflow for evaluating of *de novo* assembly results. By mapping the long reads to the contigs, besides basic statistic assembly evaluation metrics, Inspector calculates and reports precise structural errors and small-scale errors. The identified errors can also be corrected by Inspector to generate more accurate contigs.

RESULTS

1. Small-scale assembly errors and structural assembly errors

We have classified assembly errors into two groups, small-scale errors (<50bp) and structural errors (≥50bp). Small-scale errors consist of three types: base substitution, small collapse, and small expansion (**Additional file 1: Fig. S1**). Small-scale errors can be directly inferred from the pileup results of read alignments and filtered based on the number of error-supporting reads (**Methods**). We also have defined four types of structural assembly errors: expansion, collapse, haplotype switch, and inversion (**Additional file 1: Fig. S2**). Collapse and expansion are reported when part of the target genome sequence is incorrectly collapsed or expanded in the assembly. For example, collapse and expansion can occur within repetitive regions, as the presence of repeat units often forms bifurcated paths on assembly graphs, which are difficult to resolve. Haplotype switches occur at heterozygous SV breakpoints, when two haplotypes are different. The assembler fails to reconstruct either haplotype but instead generates a sequence somewhat between the two haplotypes. In these cases, reads from one haplotype will suggest a 'Collapse', and reads from the other haplotype will suggest an 'Expansion'. Inversions occur when a section of the target genome sequence is inverted in the assembly.

2. Benchmark with simulation

To benchmark the accuracy of assembly error detection of assembly evaluators, we compared Inspector with two other long-read assembly evaluators, Merqury and QUAST-LG, on the simulation dataset. We simulated a human genome from the reference genome (GRCh38) and introduced 1,000,000 single nucleotide and 20,000 structural variants. The SV size spectrum follows a geometric distribution similar to a real human genome [1-2] (**Additional file 1: Fig. S3**). A total of 2,000 structural errors and approximately 580,000 small-scale errors (base substitutions and 1bp indels) were randomly embedded into the simulated assembly (**Additional file 2: Table S1**). PacBio CLR-like reads and HiFi-like reads were simulated by PBSIM [42] and provided for Inspector to identify assembly errors. The reported assembly errors and problematic *k*-mers were compared to the ground truth to assess the accuracy of error identification for each evaluator.

Under the default settings, Inspector achieved the highest accuracy (F1 score) for assembly error detection in both haploid and diploid genomes (**Table 1**). For structural errors, Inspector correctly identified over 95% of simulated errors with both PacBio CLR and HiFi data. It achieved slightly better accuracy when working with HiFi data than CLR, as HiFi reads contain fewer sequencing errors. The precision was over 98% in both haploid and diploid simulations, although the number of SVs was approximately ten times greater than the true structural errors. For small-scale errors, the accuracy of Inspector was over 99% when working with HiFi data. The recall for small-scale error detection was lower (~ 86%) for CLR data, due to the lower signal-to-noise ratio caused by a higher sequencing error rate. In particular, the recall for base-substitution error was

higher than for small collapse or expansion, as the latter two subtypes are more susceptible to sequencing errors (**Additional file 1: Fig. S4**). Most false-negative small-scale errors exhibited a lower ratio of error-supporting reads and were filtered out by Inspector for failing to reject the null hypothesis of the binomial test. The precision of small-scale error detection was over 96% for both PacBio CLR and HiFi data, benefiting from the stringent filter implemented in Inspector. Merqury identified ~71% of the assembly errors with a precision of ~91.6% on both CLR and HiFi data. Merqury failed to detect more small collapses than base substitution and small expansions, and over 70% of Merqury-missed small-scale errors were located in repeat regions (**Additional file 1: Fig. S5**). QUAST-LG had much lower recall and precision than Inspector and Merqury, as some misassemblies were indeed caused by SVs (18% in haploid and 36% in diploid). In both haploid and diploid simulated assemblies, Inspector detected the structural assembly errors and small-scale errors with the highest accuracy among the three evaluators.

Table 1 Assembly error identification accuracy in simulated assembly

	Haploid			**Diploid**		
	Recall	**Precision**	**F1 score**	**Recall**	**Precision**	**F1 score**
Inspector structural – CLR	96.76	100.0	98.35	95.98	98.48	97.21
Inspector structural – HiFi	97.64	100.0	98.80	97.61	98.87	98.23
Inspector small-scale – CLR	86.84	99.53	92.75	86.60	96.99	91.50
Inspector small-scale – HiFi	98.99	99.65	99.32	98.91	99.62	99.26
Merqury	71.01	91.66	80.03	70.92	91.63	79.95
QUAST-LG	5.73	5.96	5.84	7.08	8.48	7.72

3. Human genome assembly evaluation

We next performed whole genome *de novo* assembly on a real human genome and evaluated the assembly results. We used an Ashkenazi Jewish sample, HG002, from Genome in a Bottle (GIAB) for the analysis. This sample has been sequenced by multiple platforms, including PacBio CLR, PacBio HiFi, Oxford Nanopore, and Illumina. There are experimentally or multiple-platform validated SNP/indel callset and SV callset at high-confidence regions publicly available for this sample [43-45], which enables the validation of identified assembly errors. We tested five the-state-of-art assemblers, Canu [15], Flye [16], wtdbg2 [17], hifiasm [20], and Shasta [18], on the PacBio CLR (~70X), HiFi (~55X), and Nanopore (~60X) dataset, if applicable. Besides Inspector, we have applied Merqury and QUAST-LG to evaluate and compare the assembly results (**Table 2**).

Inspector first estimated assembly continuity. Most assemblies contained a total of 2.7-3.0 giga base pairs, close to the reference genome, suggesting that these assemblers can reconstruct the overall structure of the target genome using long reads. Based on the maximal contig length and the N50, the sequence length of the shortest contig at 50% of the total contig lengths, Flye achieved the best continuity in the PacBio CLR and Nanopore datasets, while hifiasm outperformed the other assemblers in the HiFi dataset. Inspector then aligned the sequenced reads to contigs and identified assembly errors from read-to-contig alignments. Canu introduced the fewest structural errors as well as small-scale errors in CLR and HiFi assemblies. Hifiasm achieved results similar to Canu. Nanopore assemblies contained much more structural errors and small-scale errors than CLR and HiFi assemblies. This was likely due to the higher error rate of the

Table 2 Evaluation summary of HG002 assemblies

Assembly		Contig Continuity				Assembly Error			QUAST-LG		Merqury	Reference-based Mode		
		# Contig	Total	Max	N50	Structural	Small-scale	QV	Misassembly	MM	QV	NA50	MR (%)	Coverage (%)
CLR														
	Canu	4751	2.91	72.0	7.2	103	39.82	43.63	8341	18.84	38.51	1.32	99.15	89.41
	Flye	2168	2.82	66.6	12.0	192	30.88	43.38	4005	16.46	38.71	1.47	99.36	88.67
	Wtdbg2	2947	2.77	48.5	7.0	158	430.00	33.46	8943	29.13	29.42	0.43	97.77	86.17
HiFi														
	Canu	1376	3.37	192.2	65.3	5	1.90	54.85	47672	29.17	46.57	2.20	95.95	91.71
	Flye	2379	2.96	136.6	35.1	256	20.74	43.69	14478	17.34	48.08	2.28	97.82	90.36
	wtdbg2	1652	2.76	74.8	16.3	251	83.06	39.42	4124	14.65	42.66	1.56	99.38	86.77
	hifiasm	559	3.07	199.4	111.1	18	3.62	53.62	31143	21.47	45.88	2.53	97.37	92.03
Nanopore														
	Canu	745	2.90	101.3	33.1	1432	3845.99	24.05	14926	100.03	22.94	0.27	98.27	88.46
	Flye	584	2.87	109.9	51.7	481	316.46	34.30	7688	33.94	30.46	1.48	99.32	89.80
	wtdbg2	7959	2.97	54.2	8.2	2226	2116.76	24.91	23159	65.88	24.49	0.30	93.79	84.91
	Shasta	1258	2.80	129.3	23.3	2527	2554.72	25.74	9063	70.15	24.76	0.31	99.16	87.71

The unit of Max, N50, and NA50 is Mbp. The unit of Total is Gbp. The unit of Small-scale and MM is per Mbp. Misassembly of QUAST-LG includes both extensive and local misassembly. Mismatch of QUAST-LG includes both mismatches and indels.

Total = total number of bases. Max = length of the longest contig. MM = number of mismatches. MR = mapping ratio of assembled contigs.

Nanopore sequencing data. Flye generated the most accurate assembly among the four assemblers with Nanopore data. Note that the assemblers were tested using their default or recommended parameters. Optimized *de novo* assembly results by fine-tuning the parameters of individual assemblers may render different evaluation results.

For an overall evaluation of assembly quality, we introduce the Quality Value (QV) score. QV score is calculated based on the identified structural and small-scale errors scaled by the total base pairs of the assemblies (**Methods**). In general, PacBio HiFi assemblies demonstrated higher QV scores than CLR and Nanopore assemblies. Canu achieved the highest QV score in PacBio CLR and HiFi datasets, and Flye outperformed other assemblers in Nanopore dataset. We also evaluated all assemblies using Merqury. QV scores calculated by Merqury highly correlated with Inspector's results (**Additional file 1: Fig. S6**). QUAST-LG was also used to evaluate the assemblies. As the SVs were not excluded from the misassembly list, the total number of misassemblies was much larger than Inspector's result in all assemblies.

When the reference genome is available, Inspector can also assess the assembly synteny by aligning contigs to the reference genome. Based on the contig-to-reference alignment, Inspector computes NA50 (N50 calculated on the basis of aligned blocks instead of contig lengths), contig mapping ratio, and reference genome coverage for each assembly, reflecting the completeness of the assembly. Inspector also generates N1-N100 plots (**Additional file 1: Fig. S7**) and Dotplot (**Additional file 1: Fig. S8**) to reflect the consistency between the assembly and reference genome. NA50 and reference genome coverage in HiFi assemblies were larger than the CLR and Nanopore assemblies, which suggests that HiFi assemblies were more complete and more consistent with the reference

genome. Because the reference genome is different from the evaluated genome, these statistics may be slightly affected by genetic variants. Overall, we found that HiFi assemblies were more accurate and complete than CLR and Nanopore assemblies, suggesting that better assembly results can be achieved from long and accurate sequences.

4. Distinguish assembly errors from genetic variants

Inspector distinguishes assembly errors from genetic variants mainly from the number of reads that support the error. We identify them as "error-supporting" reads. The expected ratio of error-supporting reads is higher for assembly errors than genetic variants (**Additional file 1: Fig. S9, S10**). For small-scale errors, Inspector counts the number of reads supporting errors and contigs, and then performs binomial test to select assembly errors with significant p-values depending on the input data (**Methods**). For structural errors, a stringent filter of assembly errors is designed to sift out SVs based on the ratio of error-supporting reads and other features such as read mapping quality. We have defined the false discovery rate (FDR) of assembly error in HG002 as the errors that are actually genetic variants. We compared the identified assembly errors to the high-confidence variant callsets and computed the FDRs in each assembly. Inspector efficiently distinguished small-scale and structural (collapse and expansion) assembly errors from genetic variants, with an average FDR of 2.88% and 1.15%, respectively (**Table 3**). The FDR for Merqury and QUAST-LG were both higher than for Inspector. We also evaluated accuracy for haplotype switches and validated that over 90% of the

reported events occurred near heterozygous SV breakpoints (**Additional file 2: Table S2**).

Table 3 False discovery rate of assembly errors in HG002 assemblies

		Inspector		Merqury	QUAST-LG
		Small-scale	**Structural**		
CLR	Canu	3.57	-*	14.36	35.23
	Flye	5.77	0.00	21.93	51.65
	wtdbg2	0.94	0.00	15.33	38.37
HiFi	Canu	6.21	-*	3.61	38.96
	Flye	0.41	0.00	56.13	52.64
	wtdbg2	0.90	2.38	72.64	64.23
	hifiasm	8.85	0.00	9.99	51.63
Nanopore	Canu	1.01	0.00	3.89	23.16
	Flye	1.28	7.69	5.22	52.39
	wtdbg2	0.72	0.00	6.37	12.32
	Shasta	1.96	0.32	5.15	46.68
Mean		2.88	1.15	19.51	42.48

*Assemblies with no structural error located in the benchmark regions of HG002 are marked with '-'.

We further characterized the structural errors identified from these assemblies (**Fig. 2a**). The error patterns varied among the assemblers and among different data types. For example, Flye consistently showed a predominance of haplotype switches, suggesting a possible systematic error when assembling the heterozygous regions. In addition, Canu and wtdbg2 showed more collapses in Nanopore assemblies than PacBio CLR and HiFi assemblies. This may be due to a higher deletion error rate in Nanopore data, in contrast to a higher insertion error in PacBio data. In general, structural errors were dominated by relatively small errors, with 84.8% of structural errors shorter than 500bp (**Fig. 2b**). Collapses accounted for 88.9% of structural errors that were larger than 1kbp. Inversion errors were much rarer than the other three types and were usually large in size (493 kbp

on average). The error pattern of small-scale errors also varied among assemblers but showed more consistency within the same data type (**Additional file 1: Fig. S11**).

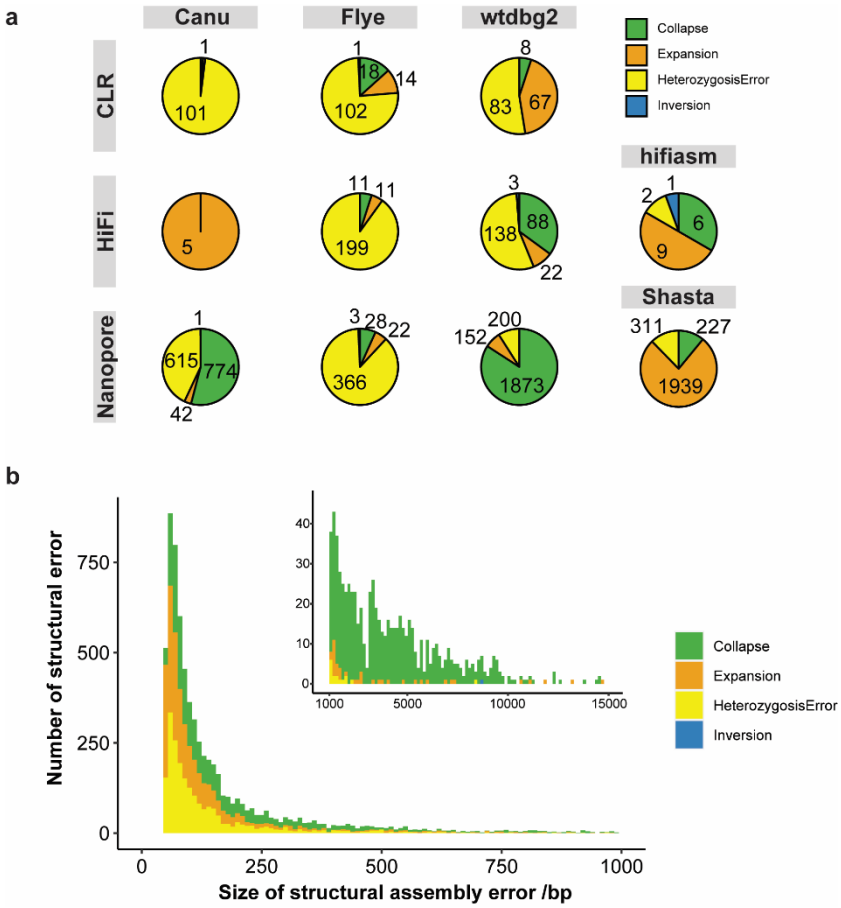

Figure 2 Characterization of structural assembly errors in HG002 assemblies. a Pie charts showing the proportion of four types of structural errors identified in Canu, Flye, wtdbg2, hifiasm, and Shasta assemblies with CLR, HiFi, and Nanopore datasets, respectively. The number of assembly error is also marked on each sector. **b** Size distribution of identified structural assembly errors in all HG002 assemblies.

To assess the effect of sequencing depth on Inspector's evaluation performance, we merged three HiFi datasets from GIAB and downsampled to a series of depths ranging from 10X to 100X. We evaluated the same assembly with these downsampled HiFi datasets. The number of assembly errors reported by Inspector was stabilized when the sequencing depth was higher than 30X (**Additional file 1: Fig. S12**), which suggests

that the sequencing depth has minor effect on Inspector's error detection, and a 30X dataset is sufficient for accurate assembly evaluation with Inspector.

5. Assembly errors are enriched in repetitive regions

Inspector reports precise locations of structural and small-scale errors, which allows us to further annotate assembly errors from each assembly result. We projected the coordinates of identified assembly errors to the reference genome and annotated these assembly errors (**Methods**). To ensure accurate repeat analysis, we used HiFi data to identify small-scale errors in all assemblies. We found that both structural errors and small-scale errors were enriched in the repetitive sequences (**Fig. 3a**). Given that approximately 55% [46] of the human genome is annotated as repetitive sequences [46], we observed a significantly higher proportion of structural (82.09%) and small-scale (73.61%) errors located in repetitive regions, suggesting that repeats remain challenging for long-read *de novo* assembly. We further examined the seven types of repetitive sequences that each account for more than 1% of the reference genome (**Additional file 1: Fig. S13**). We found that both structural and small-scale errors were enriched in simple repeats. The average percentage of structural errors located in simple repeats was 45.9%, which was a ten-fold enrichment compared to the genome baseline. Small-scale errors were also enriched in LINE, SINE, LTR, and DNA repeat elements for these assemblies as a whole. We observed an overall lower percentage of errors located in the segmental duplication and satellite regions, although some assemblies showed a higher-than-expected assembly error rate.

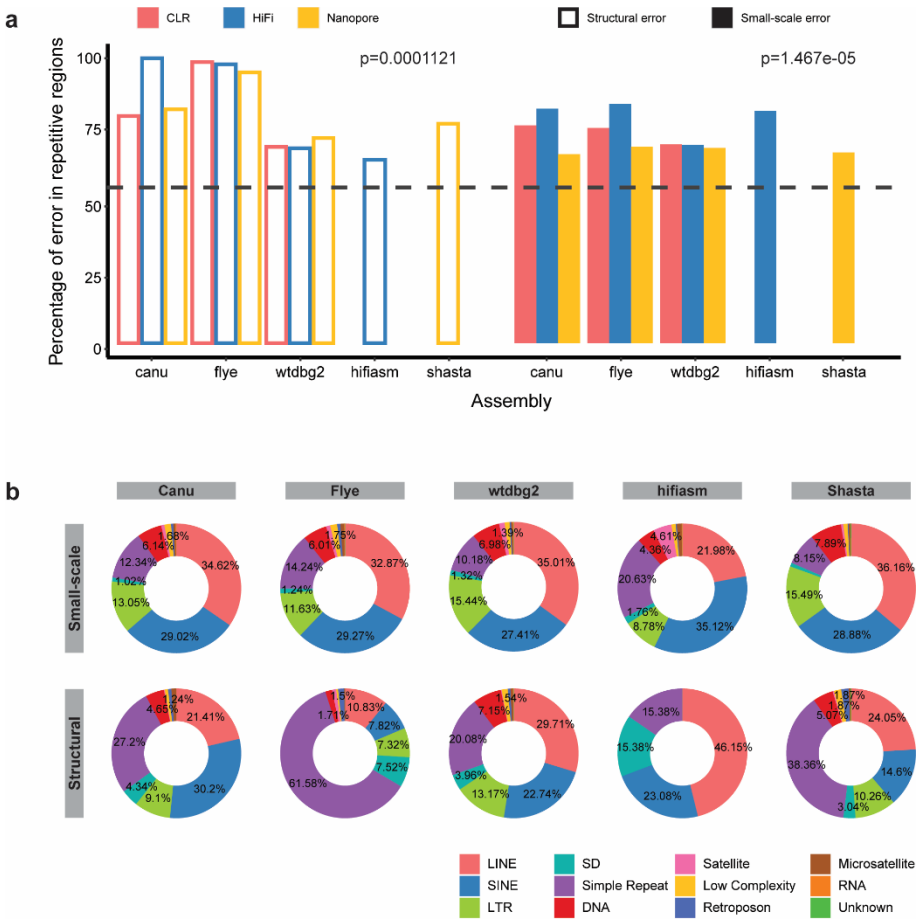

Figure 3 Enrichment of assembly errors in repetitive regions. a Proportion of assembly errors located in repetitive regions in each assembly. Dashed line indicates fraction of human reference genome annotated as repeats. P-values were calculated by one-sample t-test to compare the proportion of assembly errors with the baseline. **b** Repeat annotation of structural and small-scale errors for five assemblers.

We next characterized the repeat-associated assembly errors for the five tested assemblers. The composition of different types of repeats was relatively consistent for small-scale errors among the five assemblers tested (**Fig. 3b**), with majority of errors located in LINE, SINE, simple repeat, and LTR regions. When separating assemblies from three different data types, we observed consistent patterns in CLR assemblers and Nanopore assemblies (**Additional file 1: Fig. S14**). In the four HiFi assemblies, there was strong enrichment of simple repeats in the Flye assembly, suggesting that Flye may have worse base accuracy when resolving simple repeat regions than other genomic regions. For the structural errors, both Flye and Shasta (merely appliable to Nanopore

data) demonstrated strong enrichment in simple repeats than the other three assemblers (**Fig. 3b**). This enrichment is consistent in PacBio CLR, HiFi and Nanopore assemblies for Flye (**Additional file 1: Fig. S15**). Taken together, Inspector revealed the enrichment of assembly errors in repetitive regions and distinct repeat enrichment patterns of different assemblers, which provides guidance for further development and improvement of assemblers.

6. Assembly error correction

Equipped with the coordinates of assembly errors, Inspector includes an error-correction module for improving assembly quality, which facilitates downstream analysis. The error-correction module eliminates highly confident assembly errors (**Fig. 4a**). Small-scale errors are corrected by replacing mis-assembled bases at reported

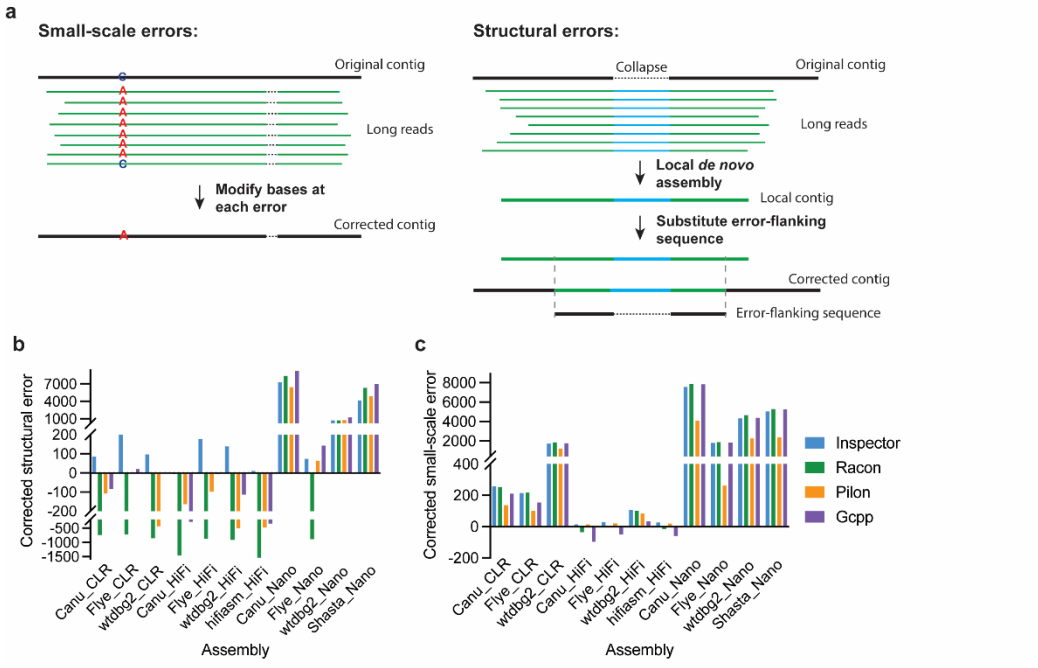

Figure 4 Improved assembly accuracy after error correction. a Methods of assembly error correction for small-scale and structural errors. **b,c** Number of corrected structural (**b**) and small-scale errors (**c**) in HG002 assembly. Negative values indicate more assembly errors after the polishing process.

locations. Structural errors are corrected by performing local *de novo* assembly around each error (**Methods**). Because the local assembly utilizes sequencing reads from only this locus (and from only one haplotype for haplotype switches), the newly generated contig can reconstruct the target genome more accurately and can therefore fix structural assembly errors.

We evaluated genome polishing performance of the Inspector error-correction module and six state-of-the-art alignment-based polishing methods, including Racon, Pilon, GCpp, Medaka, Nanopolish, and CONSENT on HG002 assemblies from Canu, Flye, wtdbg2, hifiasm, and Shasta. We used one HiFi dataset of HG002 for polishing and used another HiFi dataset to evaluate the original and polished assemblies to avoid bias (**Methods**). After polishing with HiFi reads, Inspector corrected most structural errors among four tested polishing tools in the CLR and HiFi assemblies, while GCpp corrected most structural errors in the Nanopore assemblies (**Fig. 4b**). Nevertheless, in CLR and HiFi assemblies, there were more structural errors after polishing with Racon, Pilon and GCpp, suggesting that these polishing methods can correct structural errors in lower-quality assemblies but may introduce more structural errors in relatively accurate assemblies. For small-scale errors, Inspector, Racon, and GCpp achieved higher error-correction rates than Pilon in most assemblies (**Fig. 4c**). GCpp introduced more small-scale errors in the HiFi assemblies. Based on the increased QV score after polishing, Inspector outperformed other polishing methods in CLR and HiFi assemblies, while Racon achieved the best QV score improvement in Nanopore assemblies (**Additional file 1: Fig. S16a**). Estimation of QV score with Merqury also supported that Inspector and

Racon achieved the highest assembly quality among the tested polishing methods (**Additional file 1: Fig. S16b**).

When polishing the assemblies with CLR and Nanopore reads, Racon, CONSENT and Medaka introduced new structural errors after polishing the CLR and HiFi assemblies (**Additional file 1: Fig. S17**). The number of small-scale errors in CLR and HiFi assemblies was also increased after polishing with noisy reads, especially with Nanopore reads. Inspector and Pilon reduced assembly errors or introduced the fewest errors when given noisy reads as inputs for polishing. Compared with polishing using CLR and Nanopore reads, Inspector achieved the highest error correction rate using HiFi reads for both small-scale errors (**Additional file 2: Table S3**) and structural errors (**Additional file 2: Table S4**), owing to the highest base accuracy of the HiFi dataset.

We also evaluated short-read polishing on the HG002 assemblies. Although the small-scale errors were reduced in all assemblies (**Additional file 1: Fig. S18a**), the number of structural errors increased in most assemblies after short-read polishing with Racon or Pilon (**Additional file 1: Fig. S18b**). QV scores estimated by Inspector and Merqury were both increased in CLR and Nanopore assemblies but showed minor or no improvement in HiFi assemblies (**Additional file 1: Fig. S18c**), suggesting that additional high-accuracy short-read datasets can only improve the quality of assemblies generated from noisy long reads.

In addition to the human genome, we also tested the Inspector error-correction module on the genome of Anna's hummingbird (*Calypte anna*) [47]. We performed whole-genome assembly with Canu, Flye, and wtdbg2 and corrected identified assembly errors using Inspector. The number of structural errors and small-scale errors both

dropped after Inspector error correction, with increased QV scores for all assemblies (**Additional file 1: Fig. S19**). We also compared the original and Inspector-corrected assemblies to the curated genome to validate that the structural errors in the original assemblies were accurately corrected by Inspector (**Additional file 1: Fig. S20**). Taken together, the error-correction module of Inspector can improve assembly quality by correcting both structural and small-scale errors and can achieve better error-correction efficiency than other polishing methods in more accurate assemblies.

7. Runtime and memory usage

Inspector and other assembly evaluation and polishing methods were tested on Intel Xeon E5-2680 v3 CPUs with 2.5GHz. It took 13.6 hours to evaluate a human genome assembly (Canu assembly of HG002) using 50X PacBio HiFi dataset with peak memory of 35GB (**Additional file 2: Table S5**). The error correction of this assembly took 26 minutes with peak memory of 17GB (**Additional file 2: Table S6**).

DISCUSSION

We have developed a reference-free long-read *de novo* genome assembly evaluator, Inspector, which reports exact locations, sizes, and types of assembly errors without being affected by genetic variants. In addition, Inspector improves assembly results by correcting discovered errors. These features are unique to Inspector and have not been achieved by other available assembly evaluators. We also performed detailed error analysis on different assemblers applied to different datasets. As expected, errors appear predominantly in repetitive regions. However, not all types of repeats are enriched with assembly errors. This information is important for the investigation of systematic defects in assembler algorithms. Therefore, Inspector can provide guidance for users and developers on achieving optimal assembly results.

Inspector implements multi-thread processing for read alignment, assembly error identification, and assembly error correction. For identification and correction of assembly errors, Inspector processes one contig per thread, which largely reduces runtime and memory usage. The read alignment by minimap2 is the most time-consuming step in Inspector evaluation (accounting for approximately 70% of total runtime). Therefore, the runtime of Inspector largely depends on the sequencing depth of the input dataset. The total runtime for Inspector is longer than for Merqury and QUAST, but it requires much less memory (**Additional file 2: Table S5**). For assembly error correction, the runtime of Inspector depends on the number of structural errors present in the assembly, as Inspector performs local assembly for each error. Inspector used shorter computing time and less

memory than Racon, Pilon, GCpp, and Medaka (**Additional file 2: Table S6**), benefiting from known the error positions from previous evaluation results. Nanopolish and CONSENT both required excessive computing resources for whole-genome polishing (requiring over 10 days for polishing one human genome) and thus were tested on only one contig.

Detecting assembly errors from read-to-contig alignment is a challenging problem similar to detecting genetic variants from read-to-reference alignment. Identification of small-scale error is extremely challenging with error-prone reads. The abundance of sequencing errors not only introduces ambiguity in read alignment but also reduces signal strength during error detection. To ensure high precision of assembly error detection, Inspector applies a stringent filter to exclude heterozygous variants, which will lead to a lower recall for small-scale errors in the CLR data, as shown in **Table 1**. In the real PacBio datasets, the HiFi data also reported lower QV score and more assembly errors, especially small-scale errors, than the CLR data. This is because the accurate HiFi reads are more sensitive for detecting errors. Advanced algorithms for better characterization of small-scale variants can improve the sensitivity of error detection from noisy sequencing data. When available, we will include this enhancement in future Inspector releases.

In this work, we have described our methods for benchmarking and analysis of human and Anna's hummingbird genomes. Inspector can also be applied to other species with monoploid or diploid genomes. The principles of structural error identification and binomial testing for small-scale errors are both designed with the assumption that a genome is diploid. These principles are also applicable to a haploid genome, which can be considered as an extreme case of a diploid genome with only homozygous bases.

Evaluation for species with higher ploidy levels may not be as accurate under the current version. With further development, we plan to expand the application of Inspector to species with polyploid genome in future versions.

CONCLUSIONS

This paper presents a reference-free evaluation method for *de novo* assembly. Inspector can report the precise locations and sizes for structural and small-scale assembly errors and distinguish true assembly errors from genetic variants. With its error-correction module, Inspector can improve the assembly quality by correcting the identified assembly errors. These functions exceed those achieved by existing assembly evaluators. Inspector is an accurate assembly evaluator, which can facilitate future improvement of *de novo* assembly quality.

METHODS

1. Overview of Inspector

Inspector is a tool for evaluating long-read *de novo* assembly results. As shown in **Fig. 1**, inspector consists of the following main functions: 1) standard assembly metrics; 2) structural error identification; 3) small-scale error identification; and 4) assembly error correction. Inspector also introduces a Quality Value (QV) to estimate the overall assembly quality. Given a reference genome, Inspector can assess synteny by aligning contigs to the reference genome. The detailed methods and implementation are described below.

1.1 Contig continuity and read alignment. Inspector first calculates standard assembly statistical metrics and then evaluates contig continuity based on the lengths of all contigs. Standard statistical metrics include number of contigs, total bases in the assembly, longest and second longest contig lengths, and N50, which reflect continuity of assembly results.

The statistics of read-to-contig alignments are also calculated to assess assembly quality, including read mapping rate, read splitting rate, and average alignment depth. Read mapping rate indicates the proportion of reads that can be aligned to assembled contigs. A higher read mapping rate suggests better completeness of the assembly, while a lower mapping rate suggests that parts of the genome have not been reconstructed in the assembly. The read splitting rate is the proportion of aligned reads that have split alignments. A low read splitting rate indicates better consistency between reads and assemblies and fewer large assembly errors. In contrast, a high splitting rate suggests that

there are more assembly errors which have caused the divergence between reads and assembled contigs. The average alignment depth is calculated as total length of aligned reads divided by total contig length. For good assembly, average alignment depth should be similar to sequencing depth of input reads.

1.2 Structural assembly errors. Inspector detects structural assembly errors (³50 bp) based on disagreement between reads and assembled contigs. The first step is to scan all read alignments for raw error signals of expansion (gap in read alignment), collapse (extra sequence in read), and inversion (inverted read alignment). Density-based clustering is then performed independently for each type of structural error. Instead of setting a fixed window size for clustering raw signals, Inspector's density-based clustering utilizes adjustable window size to tolerate larger shifts of raw signal positions within repetitive regions while keeping tight window size for clear genomic regions. Expansions and collapses are merged to identify haplotype switches, in which expansions overlap with collapses. To remove noise caused by sequencing errors or incorrect read alignments, Inspector filters out candidates with numbers of supporting reads below a threshold value (three by default).

To remove false-positive candidates caused by genetic variants, Inspector includes a filter based on the ratio of error-supporting read, local coverage, and read mapping quality. The ratio of error-supporting read is the fundamental criterion and computed with the number of error-supporting reads divided by the local coverage. As shown in **Additional file 1: Fig. S9**, read alignments at homozygous variants do not show inconsistency with the contig, as both haplotypes are the same as the contig

sequence. Heterozygous variant regions show an alternative allele in about 50% of reads (from one haplotype). However, at true assembly error regions, both haplotypes are different from the contig, including the haplotype switch, leading to a theoretical ratio of about 100% for error-supporting reads. The ratio of error-supporting read for assembly errors can be lower than 100% in practice due to sequencing errors or inaccurate read alignments but are still higher than heterozygous variants, as shown in **Additional file 1: Fig. S10**. The filter also discards candidates with extremely high coverage or poor average read mapping quality to ensure the reported assembly errors are confident. By default, Inspector reports coordinates on contigs for all assembly errors in BED format, which can be easily loaded to visualization tools such as IGV [48].

1.3 Small-scale assembly errors. Inspector identifies small-scale assembly errors (<50bp) to estimate the base accuracy of an assembly. Samtools [49] is used to generate pileup information for each contig based on read-to-contig alignments. Inspector then scans pileup results for candidate small-scale errors in regions that are enriched with mismatches or indels. All bases with less than 20% of reads supporting a small-scale error were excluded to remove most noise caused by sequencing errors. Similar to structural errors, a true small-scale error is expected to be supported by reads from both haplotypes (100% of reads), while mismatches or indels caused by heterozygous variants are supported by only one haplotype (50% of reads). For a given position on the assembly, each aligned read is treated as an independent experiment, containing either the same or a different base (or indel) with the base in the contig. All bases in the reads at this position follow a binomial distribution, with n being the number of reads and p being

the probability that the base is a different base from the contig. Inspector performs a one-tailed binomial test for each candidate position to distinguish small-scale errors from genetic variants. The null hypothesis of the binomial test is that the probability of a read that contains a different base against the contig is 0.5 (genetic variant at this location), and the alternative hypothesis is that the probability is higher than 0.5 (small-scale error at this location). A significant p-value from the binomial test would reject the null hypothesis and support that there is a small-scale error at the tested position. The probability of a read to support an error used in binomial test is set to 0.5 for high-accuracy HiFi data, and set to 0.4 for low-accuracy data (CLR and Nanopore), considering the sequencing error rate of 15-20%. Candidates with significant p-values (<0.01 for HiFi and <0.05 for CLR and Nanopore data) are reported as small-scale errors. Similar to structural errors, small-scale errors are also reported in BED format.

1.4 Assembly quality estimation. Structural and small-scale assembly errors are used to estimate the overall accuracy of an assembly result. Given a list of structural errors and small-scale errors of the assembly, the total bases of assembly error, N_{Err}, can be calculated as:

$$N_{Err} = N_{Exp} + N_{Col} + N_{Her} + N_{Small} + n_{Inv}$$

where N_{Exp} , N_{Col}, N_{Her}, and N_{Small} are the total bases affected by expansions, collapses, haplotype switches, and small-scale errors, while n_{Inv} is the total number of inversion errors. Since the number of total bases in an assembly, N_{asm}, is usually very large, N_{Err} can be considered as the expectation of incorrect bases. Thus, the estimated error rate, E, can be defined as:

$$E = \frac{N_{Err}}{N_{asm}} = \frac{N_{Exp} + N_{Col} + N_{Her} + N_{Small} + n_{Inv}}{N_{asm}}$$

The Phred quality score is computed as $QV = -10log_{10}E$.

1.5 Assembly error correction. Inspector includes an error-correction module to address identified structural and small-scale assembly errors. For small-scale errors, Inspector substitutes problematic bases with bases supported by the majority of reads. For structural assembly errors, Inspector collects the error-supporting reads and performs a local *de novo* assembly with Flye (v2.8.3) [16] for each assembly error. In particular, for haplotype switches, Inspector only collects reads from one haplotype to perform the local assembly. For each structural error, the local assembly uses the reads from the region around the error and from the same haplotype, which simplifies the assembly process and can therefore generate a more accurate contig than whole genome *de novo* assembly. For structural errors located within repetitive regions, Inspector collects reads only from the current repeat unit without interference from other repeat units, increasing the accuracy of local assembly at repetitive regions. Inspector aligns the new contigs from local assemblies to the original contigs and substitutes the sequences flanking each error with new sequences from the local assembly results.

1.6 Reference-based mode. To assess the synteny of an assembly with a known reference genome, Inspector includes a reference-based module to evaluate assembly quality. The module aligns contigs to the reference genome with minimap2 [41] preset parameter '-x asm5'. Statistics for contig-to-reference alignment are calculated, including contig alignment NA50, contig mapping rate, and reference genome coverage. A Dotplot is

generated based on contigs and reference alignment results. In addition, structural errors and small-scale errors are detected. Inspector reports coordinates on the reference genome and on the contig for all assembly errors. Note that assembly errors detected from contig-to-reference alignment also include genetic variants of the sequenced genome (including SVs, SNPs, and indels) and substitutions.

2. Simulation benchmark

To benchmark the evaluation accuracy of Inspector, testing used a simulated human whole genome assembly containing both structural and small-scale assembly errors. A total of 1,000,000 SNPs and 20,000 SVs (deletions and insertions) were introduced into autosomes and X chromosome of human reference genome hg38. 67% of all variants were randomly assigned as heterozygotes and 33% as homozygotes. PBSIM [42] was used to simulate 50X PacBio CLR-like and HiFi-like reads with options '--data-type CLR --model_qc model_qc_clr --length-mean 15000 --length-sd 3000 --accuracy-mean 0.85' and '--data-type CCS --model_qc model_qc_ccs --length-mean 15000 --length-sd 3000 --accuracy-mean 1.00', respectively. The mean base accuracy was 0.85 for CLR-like reads and 0.98 for HiFi-like reads according to the log file from PBSIM. Assembled contigs were simulated by splitting the simulated human genome at 'N' bases. Small fragments shorter than 10,000 bp were excluded. A total of 2,000 structural errors (900 expansions, 900 collapses, 190 haplotype switches, and 10 inversions) and about 580,000 small-scale errors (50% base substitution, 25% 1-bp expansion, and 25% 1-bp collapse) were spiked in as the ground truth. A haploid human genome was also simulated by selecting only haplotype 1 from the diploid simulation.

Inspector was applied with default settings. The reported structural and small-scale errors were compared to the ground truth to calculate recall, precision, and F1 score ($\frac{2*recall*precision}{recall+precision}$). Human reference genome hg38 was provided to QUAST-LG as the reference. Although the minimum length for structural errors was 50bp in simulated assemblies, QUAST-LG can only report the coordinates of extensive misassemblies longer than 85 bp. These extensive misassemblies were compared with a subset of ground-truth structural errors that were longer than 85bp to assess the accuracy of QUAST-LG. Since Merqury requires high-accuracy reads as input data, the simulated HiFi dataset (with sequencing error rate < 2%) was provided to Merqury to identify erroneous *k*-mers that were only present in the assembly but not in the input reads. A series of overlapping *k*-mers were merged into one single event for the benchmark.

3. Whole genome de novo assembly of HG002

Whole genome *de novo* assembly was performed for HG002 with PacBio CLR, HiFi (15-20kb), and Nanopore datasets. The expected genome size was set to 3.1G for all assemblers. Canu (v2.0) was run with options '-pacbio' for the PacBio CLR and '-pacbio-hifi' for the PacBio HiFi dataset. The Canu assembly of the Nanopore dataset was obtained from a previous publication [18]. Contigs marked with 'suggestBubble=yes' were removed from evaluation. Flye (v2.8.2) was run with options '--pacbio-raw' for the CLR, '--pacbio-hifi' for the HiFi, and '--nano-raw' for the Nanopore dataset, respectively. Wtdbg2 (v2.5) was run with options '-p 17' for the CLR and Nanopore datasets, and preset '-x ccs' for the HiFi dataset. Hifiasm (v0.13) was only applied to PacBio HiFi datasets with the default settings. The Shasta assembly of Nanopore dataset

was also obtained from a previous publication [18]. All assemblies were evaluated by Inspector with default settings. CLR assemblies were evaluated with the raw CLR dataset, HiFi assemblies were evaluated with the HiFi dataset (15-20kb), and Nanopore assemblies were evaluated with the raw Nanopore dataset.

4. Other Assembly evaluation tools

QUAST-LG (v5.0.2), a reference-based approach, and Merqury (v1.1), a *k*-mer based approach, were also used to evaluate assemblies. For QUAST-LG, GRCh38 was provided as the reference genome. QUAST-LG was run with command:

'quast-lg.py contig.fa -o output/ -r hg38.fa -m 10000 -x 86'

The number of misassemblies included both extensive and local misassemblies, and number of mismatches included both mismatches and indels.

For Merqury, a meryl database was first generated with approximately 50X Illumina paired-end reads with *k*-mer size of 21bp. Merqury was then run based on the Illumina meryl database to evaluate HG002 assemblies with default settings:

'meryl *k*=21 count output read-db.meryl allread.fa'

'merqury.sh read-db.meryl contig.fa output'

The assembly-only *k*-mers were collected from Merqury's output and the overlapping *k*-mers were merged into a single event.

5. Benchmark of assembly error in HG002

The false discovery rate of assembly errors was calculated by comparing reported assembly errors to the genetic variant callset of HG002. Coordinates of assembly errors were projected to the human reference genome based on contig-to-reference alignment. Matched base pairs between contigs and the reference genome were stored in a hash table. The corresponding reference coordinate of an assembly error can be inferred from the hash table according to its assembly coordinate. Small-scale errors were compared to the small variant callset (v4.2.1) from GIAB. Since the high-confidence SV callset is only available in 'benchmark regions' of HG002 [44], structural assembly errors located only in benchmark regions were compared to the SV callset to calculate FDRs.

Coordinates of misassemblies reported by QUAST-LG were extracted from filtered contig alignment. Misassemblies located within benchmark regions were compared to the SV callset for FDR assessment. Assembly-only *k*-mers from Merqury's output were merged and projected to the reference genome. FDR was computed by comparing the locations of *k*-mers to the merged variant callset (SVs and small variants).

6. Down-sampling of HG002

To evaluate the robustness of Inspector, three HiFi datasets (11kb, 15kb and 15-20kb) of HG002 were merged to generate a HiFi dataset with an ultra-high depth. It was then downsampled to a series of depths, ranging from 10X to 100X, by randomly selecting reads. Depth was determined as total number of base pairs in reads divided by the human genome size (3.1Gbp). Inspector was applied to identify assembly errors using default settings to validate its robustness in addressing datasets of varying depth.

7. Repeat annotation of assembly errors

Coordinates of assembly errors were projected to the human reference genome. Those assembly errors located in unaligned parts of the assembly cannot be projected to the reference genome and therefore were excluded from analysis. Repeat annotation of all assembly errors was performed by a custom Python script, which compared reference coordinates of assembly errors to the genomic repeat annotation downloaded from UCSC Genome Browser [50].

8. Polishing of HG002 assemblies

Inspector correction and other polishing methods were tested on HG002 assemblies. The error correction module of Inspector was tested with PacBio CLR (70x), PacBio HiFi (15-20kbp, 51x), and Nanopore (53x) datasets with default settings. The input datatype was specified for each dataset to enable accurate local assembly in the structural error correction process. Racon (v1.4.20) and Pilon (v1.24) were tested with PacBio CLR, PacBio HiFi, and Nanopore datasets with default settings. GCpp (v 2.0.2) was tested with downsampled raw subreads of PacBio HiFi dataset (70X). Medaka (v 1.4.3) polished HG002 assemblies with Nanopore datasets with the options "--model r941_min_high_g303 --batch 200 --bam_chunk 2000000". Nanopolish (v0.13.3) was tested with Nanopore dataset using default settings. CONSENT (v2.2.2) polished HG002 assemblies with PacBio CLR datasets with options "--windowSize 50000". Nanopolish and CONSENT were tested on only one contig (10Mbp in length) per assembly due to the excessive requirement of computational resources for whole-genome correction. The input read alignment files for Racon, Pilon, Medaka, and Nanopolish were aligned by

minimap2 and sorted by Samtools sort. The read alignment files provided to GCpp were aligned by pbmm2 and sorted by Samtools. All polishing tools were tested with only one round of the polishing process. We also polished the HG002 assemblies with Illumina dataset (downsampled to 50X) to assess the improvement of assembly quality from short reads. The original and polished assemblies were evaluated using Inspector with a merged HiFi dataset (11kbp and 15kbp, total of 58x) and using Merqury with meryl database generated from Illumina dataset.

9. Whole-genome assembly of Anna's hummingbird sample

The PacBio CLR (~70X) data of Anna's hummingbird (*Calypte anna*) was downloaded from the Vertebrate Genomes Project and used to for whole-genome *de novo* assembly with Canu, Flye, and wtdbg2 with genome size of 1.1Gbp. Inspector was run with default settings to evaluate and correct errors for the three assemblies. The curated assembly was obtained from GenomeArk as the ground truth. The uncorrected and corrected assemblies were compared to curated assembly with Mauve [51] to visualize structural errors before and after Inspector error correction.

Declarations

Availability of data and materials

Inspector is publicly available at https://github.com/ChongLab/Inspector and https://codeocean.com/capsule/9679766/tree under the MIT License. The sequencing data of HG002 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 assembly evaluation and error

correction, and PacBio CCS 11kb and 15kb were used for evaluating assemblies before and after error correction. The benchmark variant callsets used for assembly error validation were downloaded from GIAB [43, 44]. The PacBio CLR dataset and curated genome assembly of Anna's hummingbird were downloaded from GenomeArk [52].

Competing interests

The authors declare that they have no competing interests.

Funding

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.W.

Authors' contributions

Z.C. conceived and managed the project. Y.C. implemented the algorithm, collected all the datasets, and performed all the analysis. Z.C., Y.Z., A.W., and M.G. were involved in data analysis and testing of the algorithm. Y.C. prepared the figures and tables and wrote the manuscript draft, and Z.C. and A.W. revised it. All authors have read and approved the final manuscript.

Acknowledgements

We are grateful to Mr. Haoxiang Gao for the discussion and suggestion of statistical analysis of the small-scale assembly errors. We also would like to thank Dr. Miten Jain and Dr. Benedict Paten for sharing the HG002 nanopore assembly results from Canu and Shasta. We also thank Dr. Aaron Wenger, Dr. Nathanael D. Olson, and Dr. Justin M. Zook for sharing the PacBio and Nanopore raw data of HG002.

REFERENCES

1. Chen Y, Zhang Y, Wang AY, Gao M, Chong Z: Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol* 2021, 22:312.

2. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y, et al: An integrated map of structural variation in 2,504 human genomes. *Nature* 2015, 526:75-81.

3. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR: A global reference for human genetic variation. *Nature* 2015, 526:68-74.

4. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al: Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* 2019, 10:1784.

5. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al: Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 2021, 372.

6. Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H, Ramakrishnan S, Maumus F, Ciren D, et al: Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell* 2020, 182:145-161.e123.

7. Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, et al: Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 2018, 557:43-49.

8. He Y, Luo X, Zhou B, Hu T, Meng X, Audano PA, Kronenberg ZN, Eichler EE, Jin J, Guo Y, et al: Long-read assembly of the Chinese rhesus macaque genome and identification of ape-specific structural variants. *Nat Commun* 2019, 10:4233.

9. Audano PA, Sulovari A, Graves-Lindsay TA, Cantsilieris S, Sorensen M, Welch AE, Dougherty ML, Nelson BJ, Shah A, Dutcher SK, et al: Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* 2019, 176:663-675.e619.

10. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al: Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 2019, 37:1155-1162.

11. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, Dilthey AT, Fiddes IT, et al: Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* 2018, 36:338-345.

12. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al: Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 2013, 10:563-569.

13. Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM: Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol* 2015, 33:623-630.

14. Li H: Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* 2016, 32:2103-2110.

15. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM: Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* 2017, 27:722-736.

16. Kolmogorov M, Yuan J, Lin Y, Pevzner PA: Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 2019, 37:540-546.

17. Ruan J, Li H: Fast and accurate long-read assembly with wtdbg2. *Nat Methods* 2020, 17:155-158.

18. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, et al: Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol* 2020, 38:1044-1053.

19. Chen Y, Nie F, Xie S-Q, Zheng Y-F, Dai Q, Bray T, Wang Y-X, Xing J-F, Huang Z-J, Wang D-P, et al: Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat Commun* 2021, 12:60.

20. Cheng H, Concepcion GT, Feng X, Zhang H, Li H: Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* 2021, 18:170-175.

21. Loman NJ, Quick J, Simpson JT: A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* 2015, 12:733-735.

22. Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR: Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol* 2015, 16:294.

23. Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz MC, McCombie WR: Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome Res* 2015, 25:1750-1756.

24. Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, et al: Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 2020, 585:79-84.

25. Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HOK, Buffalo V, Zerbino DR, Diekhans M, et al: Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res* 2011, 21:2224-2241.

26. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, et al: Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* 2013, 2:10.

27. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A: Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* 2018, 34:i142-i150.

28. Gurevich A, Saveliev V, Vyahhi N, Tesler G: QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013, 29:1072-1075.

29. Cameron DL, Schröder J, Penington JS, Do H, Molania R, Dobrovic A, Speed TP, Papenfuss AT: GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res* 2017, 27:2050-2060.

30. Rhie A, Walenz BP, Koren S, Phillippy AM: Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* 2020, 21:245.

31. Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ: KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* 2017, 33:574-576.

32. Seppey M, Manni M, Zdobnov EM: BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* 2019, 1962:227-245.

33. Vaser R, Sovic I, Nagarajan N, Sikic M: Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 2017, 27:737-746.

34. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM: Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 2014, 9:e112963.

35. GCpp: Generate Highly Accurate Reference Contigs. [https://github.com/PacificBiosciences/gcpp]

36. Morisse P, Marchet C, Limasset A, Lecroq T, Lefebvre A: Scalable long read self-correction and assembly polishing with multiple sequence alignment. *Sci Rep* 2021, 11:761.

37. Zimin AV, Salzberg SL: The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS Comput Biol* 2020, 16:e1007981.

38. Warren RL, Coombe L, Mohamadi H, Zhang J, Jaquish B, Isabel N, Jones SJM, Bousquet J, Bohlmann J, Birol I: ntEdit: scalable genome sequence polishing. *Bioinformatics* 2019, 35:4430-4432.

39. Loman NJ, Quick J, Simpson JT: A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* 2015, 12:733-735.

40. Medaka, a tool to create consensus sequences and variant calls from nanopore sequencing data. [https://nanoporetech.github.io/medaka/]

41. Li H: Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018, 34:3094-3100.

42. Ono Y, Asai K, Hamada M: PBSIM: PacBio reads simulator—toward accurate genome assembly. *Bioinformatics* 2013, 29:119-121.

43. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M: Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* 2014, 32:246-251.

44. Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, Sherry S, Koren S, Phillippy AM, Boutros PC, et al: A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol* 2020, 38:1347-1355.

45. Zook JM, McDaniel J, Olson ND, Wagner J, Parikh H, Heaton H, Irvine SA, Trigg L, Truty R, McLean CY, et al: An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol* 2019, 37:561-566.

46. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: Initial sequencing and analysis of the human genome. *Nature* 2001, 409:860-921.

47. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al: Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 2021, 592:737-746.

48. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: Integrative genomics viewer. *Nat Biotechnol* 2011, 29:24-26.

49. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009, 25:2078-2079.

50. Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, Powell CC, Nassar LR, Maulding ND, Lee CM, et al: The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res* 2021, 49:D1046-D1057.

51. Darling AC, Mau B, Blattner FR, Perna NT: Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* 2004, 14:1394-1403.

52. GenomeArk: Vertebrate Genomes Project. [https://vgp.github.io/genomeark/]

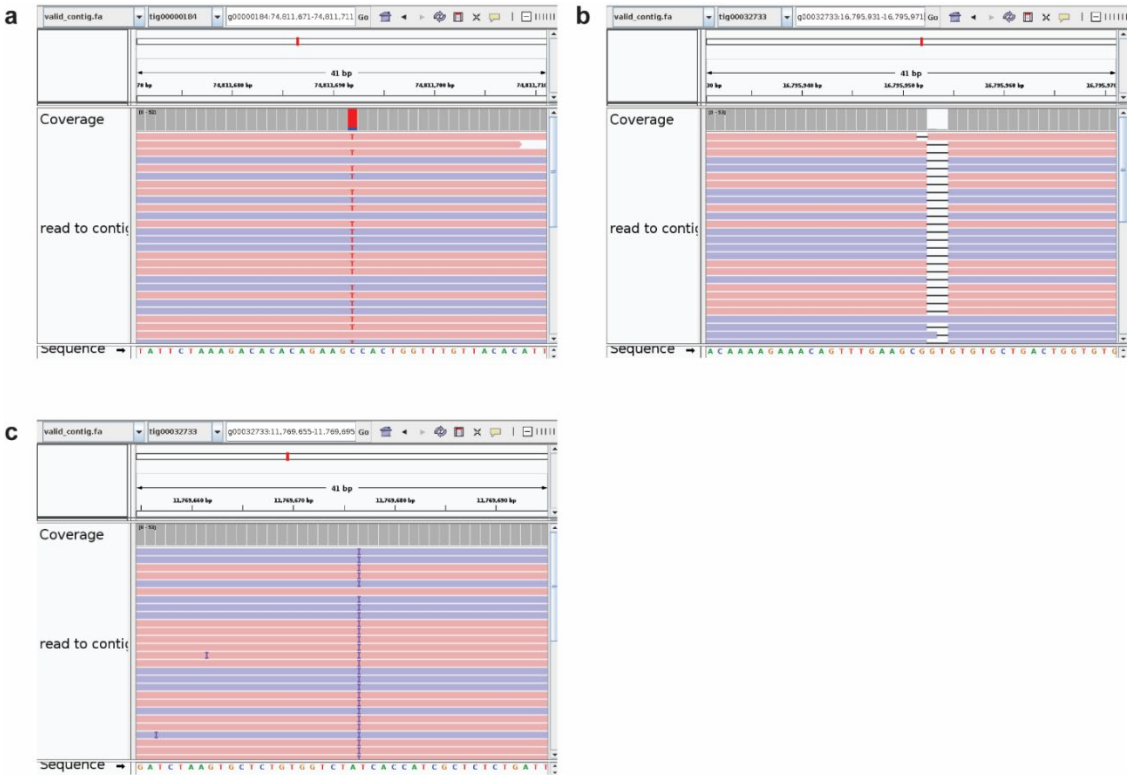

Figure S1 IGV views of examples of small-scale assembly errors. There are discrepancies between the contig and the majority of reads in base substitution (**a**), small expansion (**b**), and small collapse (**c**).

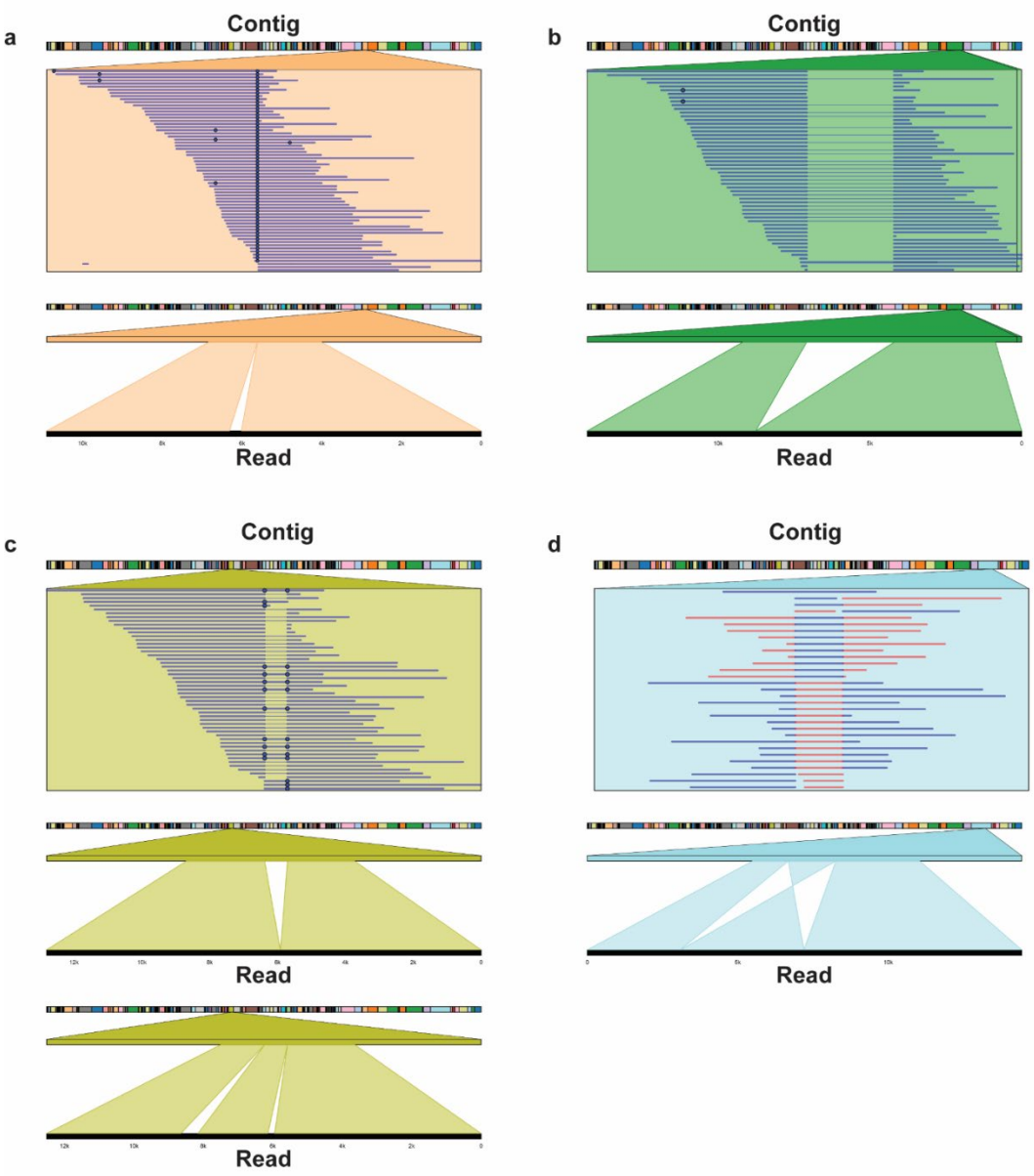

Figure S2 Examples of structural assembly errors. **a** An insertion-like pattern in read alignment representing a collapse error, as this part of sequence is collapsed in the contig. **b** A deletion-like pattern in read alignment representing an expansion error, as these sequences in contig are expanded and not present in the reads. **c** An insertion-like pattern in half of the reads and a deletion-like pattern in the other half of the reads representing a haplotype switch, as the contig is different from both haplotypes at this heterozygous region. **d** Inverted alignment within reads representing as an inversion error.

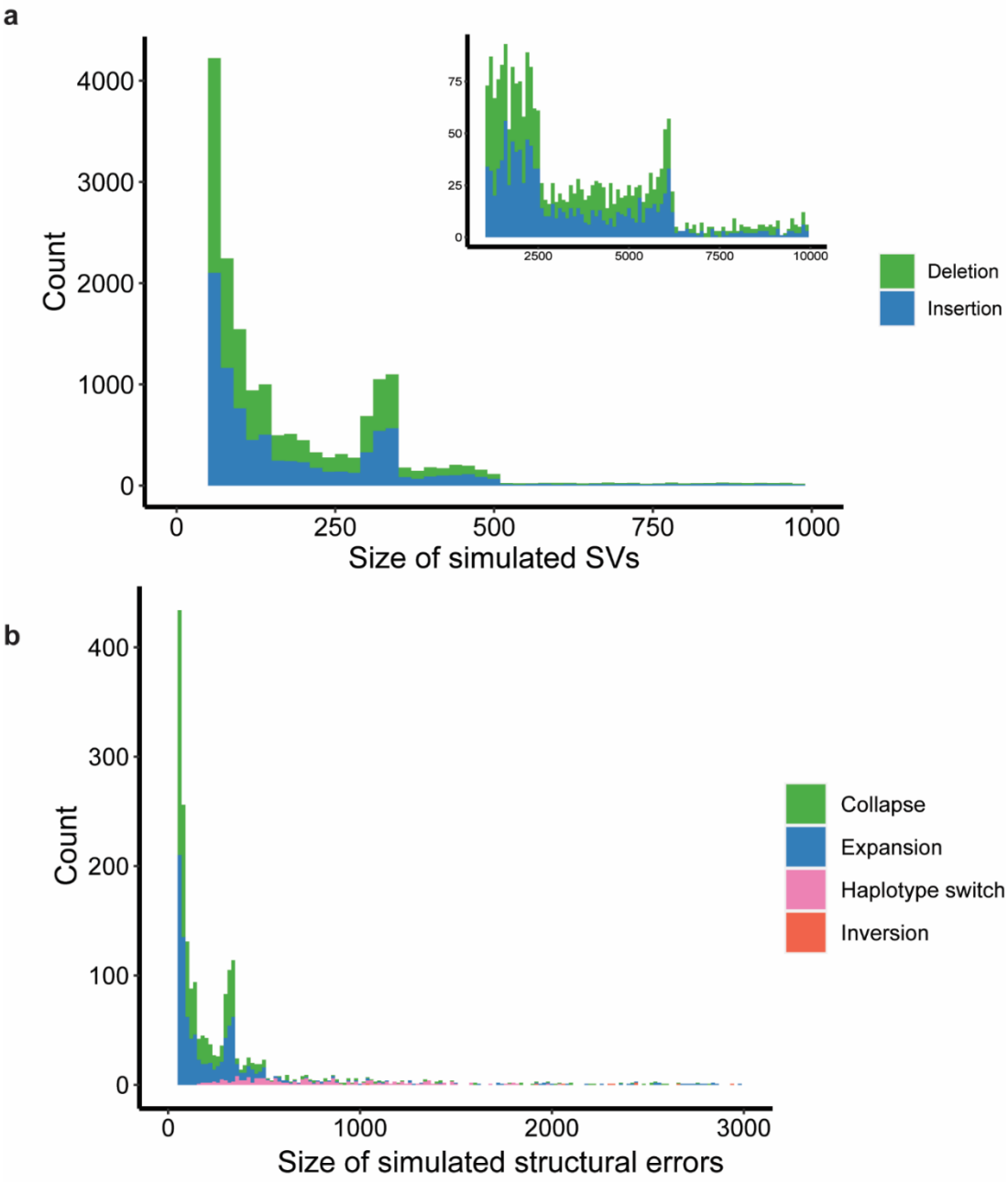

Figure S3 a Size distribution of structural variants in the simulated genome. The peak at ~350bp and ~6kbp were induced to mimic SVs caused by Alu and LINE elements. **b** Size distribution of the simulated structural assembly errors.

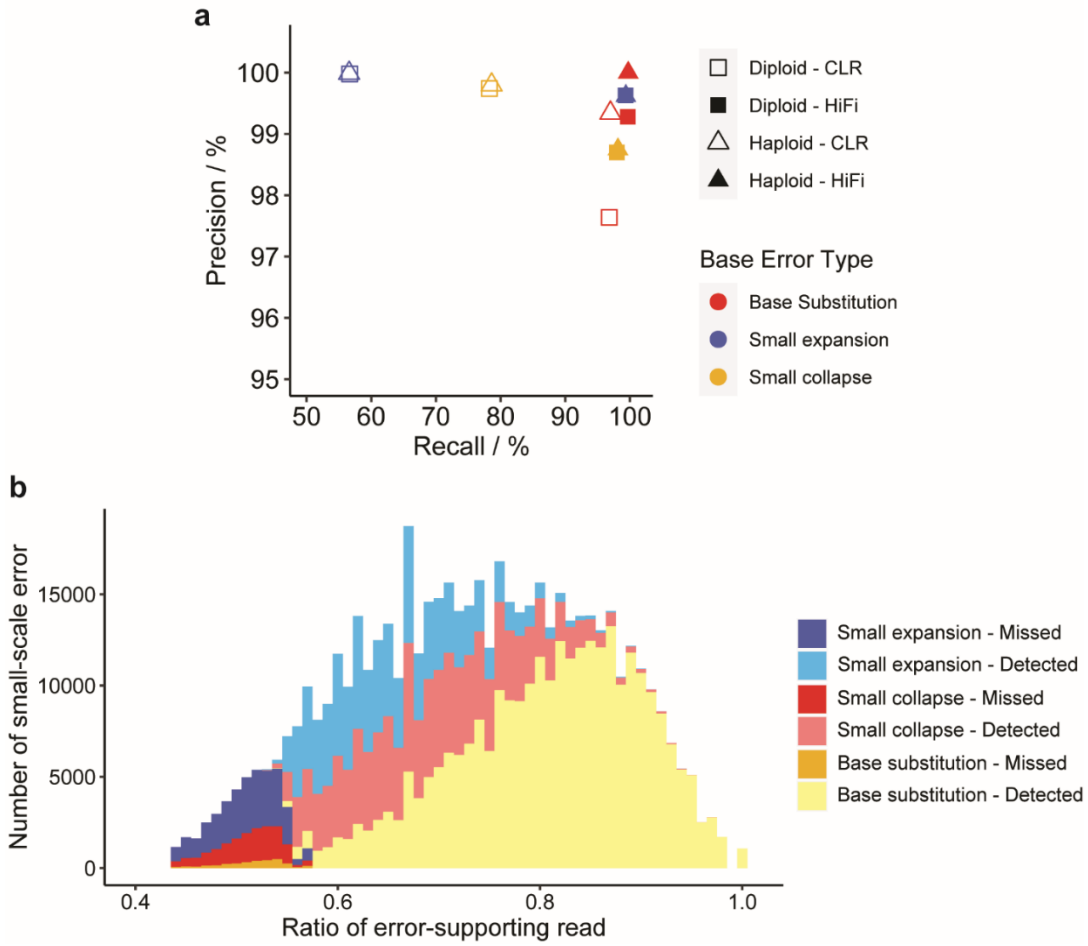

Figure S4 Small-scale error detection in the simulated dataset. **a** Recall and precision of small-scale error detection. The recall was lower for small expansion and collapse in two CLR datasets. **b** Distribution of ratio of error-supporting read of three subtypes of small-scale errors in Diploid-CLR evaluation. Missed assembly errors showed lower ratio of error-supporting read, owing to the presence of sequencing errors.

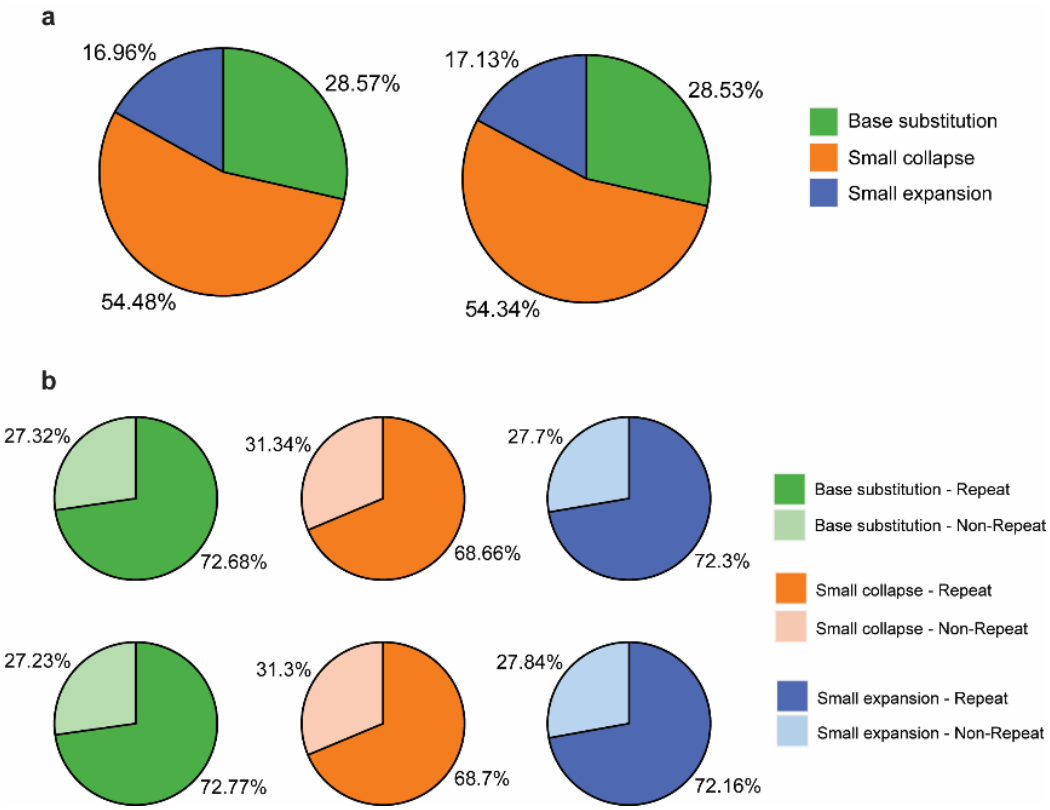

Figure S5 Small-scale error missed by Merqury but detected by Inspector in the simulated dataset. **a** Three subtypes of small-scale errors detected by Inspector but not by Merqury in haploid (left) and diploid (right) simulation. **b** Composition of Merqury-missed assembly errors located within and outside the repetitive regions for haploid (top) and diploid (bottom) simulation.

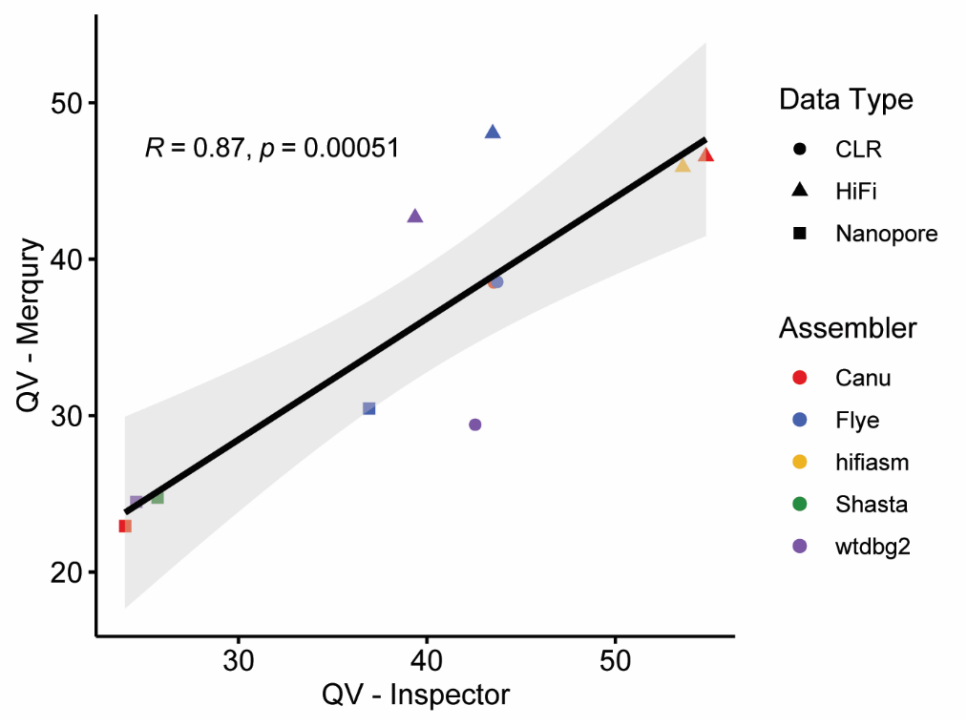

Figure S6 Correlation between QV scores computed by Inspector and Merqury in all HG002 assemblies.

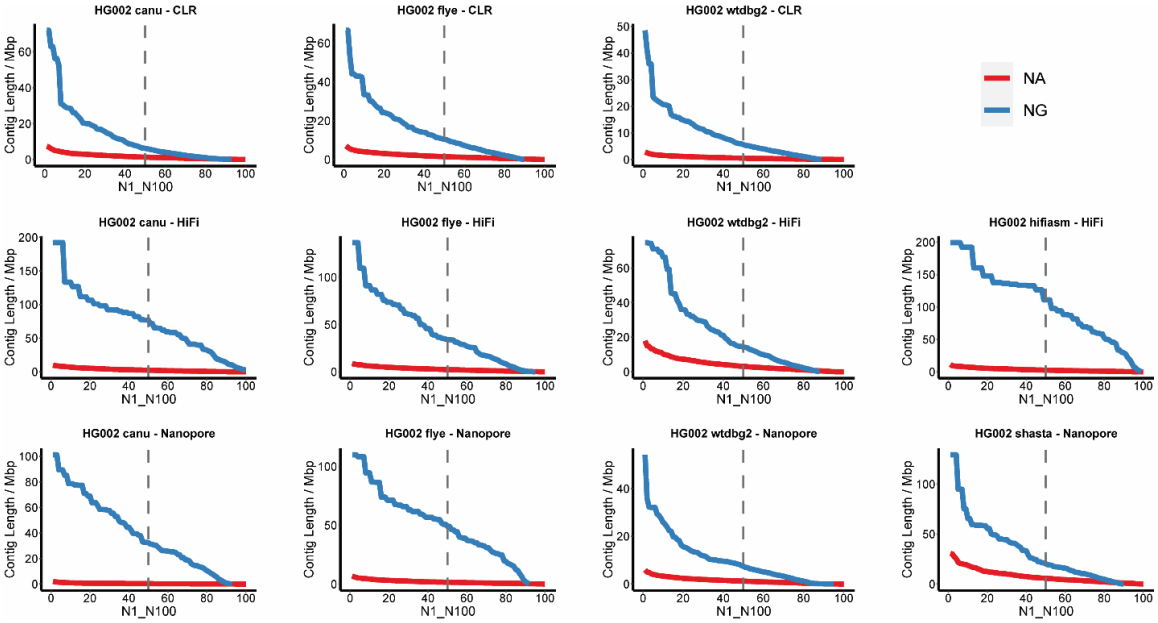

Figure S7 N1-N100 plot of HG002 assemblies. Dashed lines indicate the NA50 and NG50 at 50% of total assembly length. NAs were calculated on the basis of aligned blocks instead of the contig lengths. NGs were calculated on the basis of known or estimated genome size.

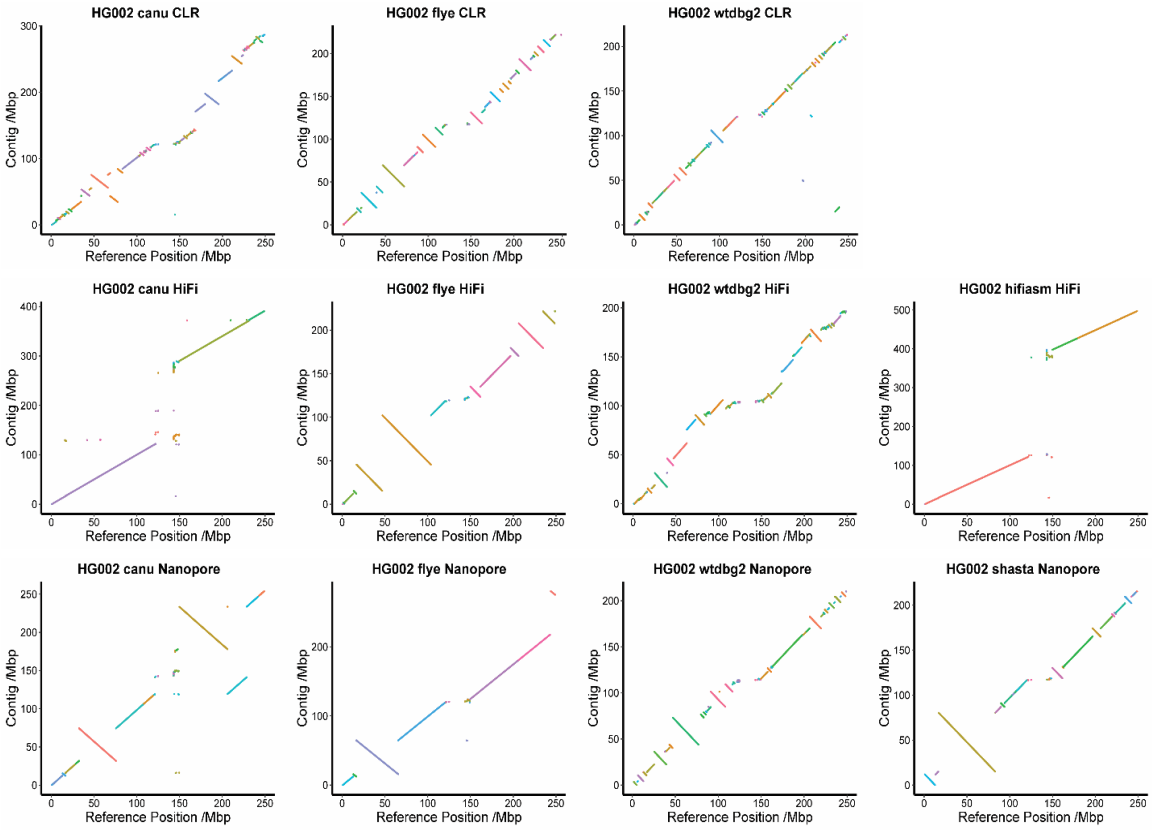

Figure S8 Dotplot of HG002 assemblies. Each dot represents the base match between contig and the reference genome. Dots from the same contig are marked with the same color.

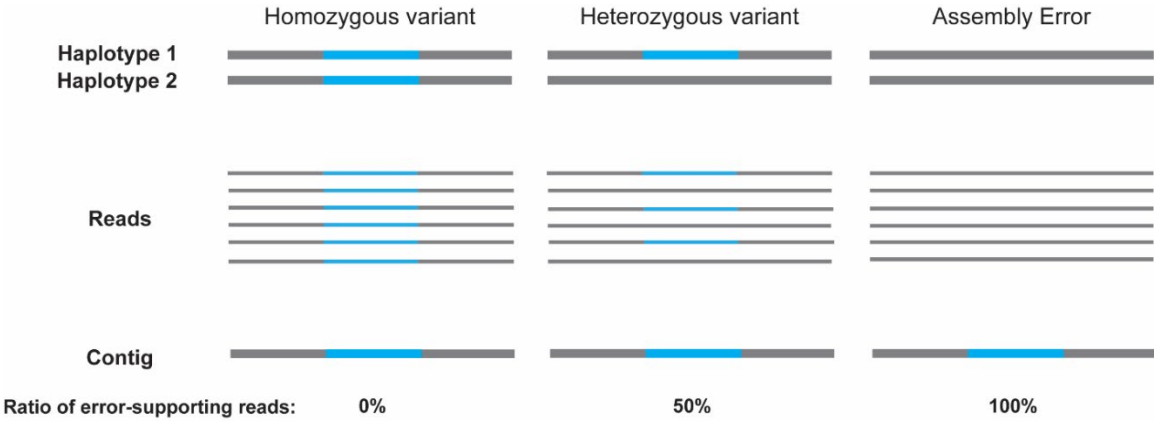

Figure S9 A theoretical interpretation of the difference between an assembly error and genetic variants in a diploid genome. Sequences differing from the reference genome are marked in blue. For a homozygous variant (left), the contig is consistent with both haplotypes, in which all reads are identical with the contig. In this case, there is no assembly error. For a heterozygous variant (middle), reads from one haplotype are different from the contig, with a ratio of error-supporting reads around 50%. This ratio is close to the frequency of a heterozygote. This is not an assembly error. Only a substantially high ratio (close to 100% theoretically) of reads supporting the error will be considered assembly error (right).

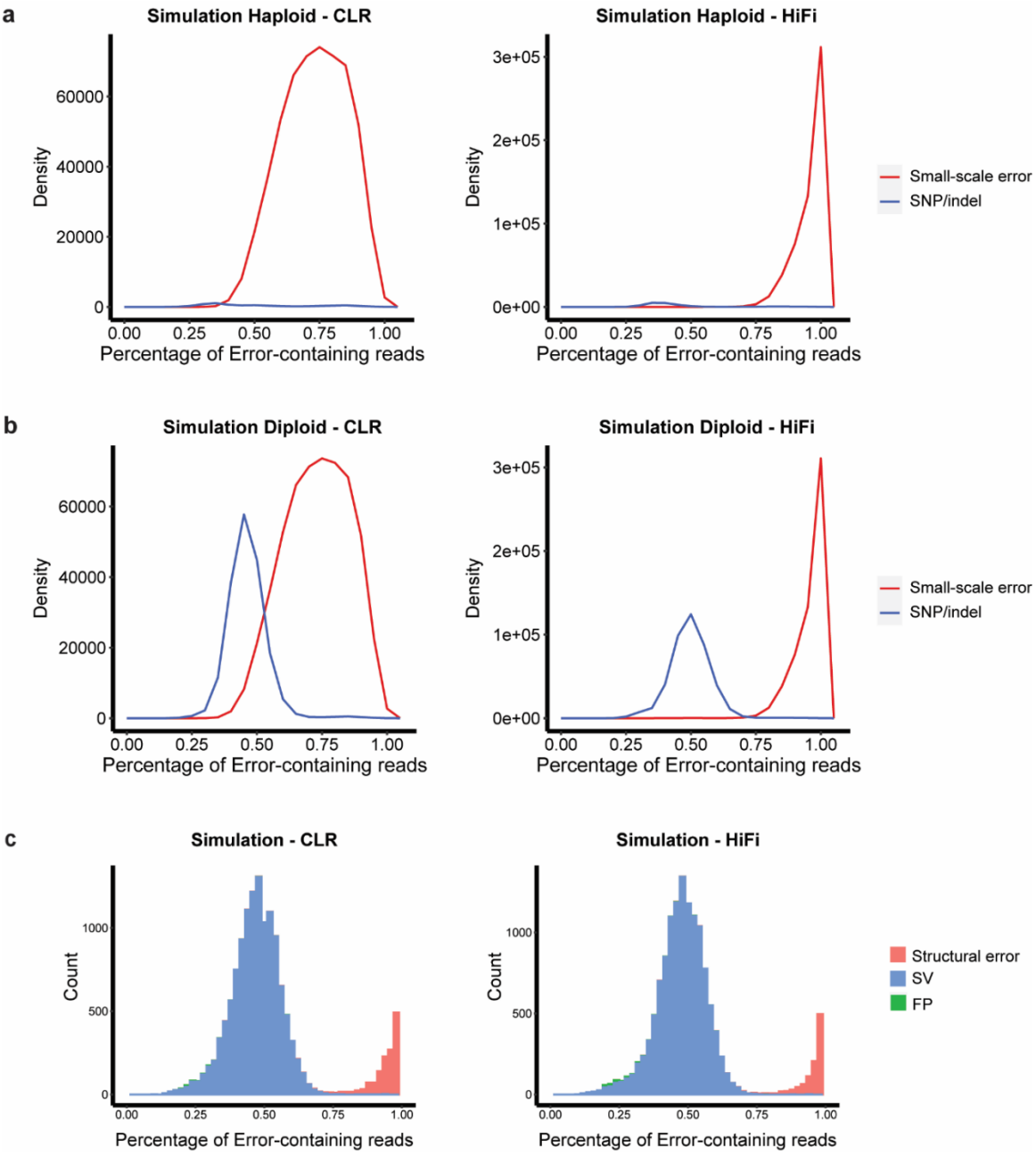

Figure S10 a,b Distribution of ratios of error-supporting reads for small-scale assembly errors and SNPs/indels in simulated haploid (**a**) and diploid (**b**) datasets. Small-scale errors are more separate from genetic variants in HiFi datasets than in CLR datasets. **c** Distribution of ratios of error-supporting reads for structural errors. Structural errors show higher ratios than SVs in both CLR and HiFi datasets.

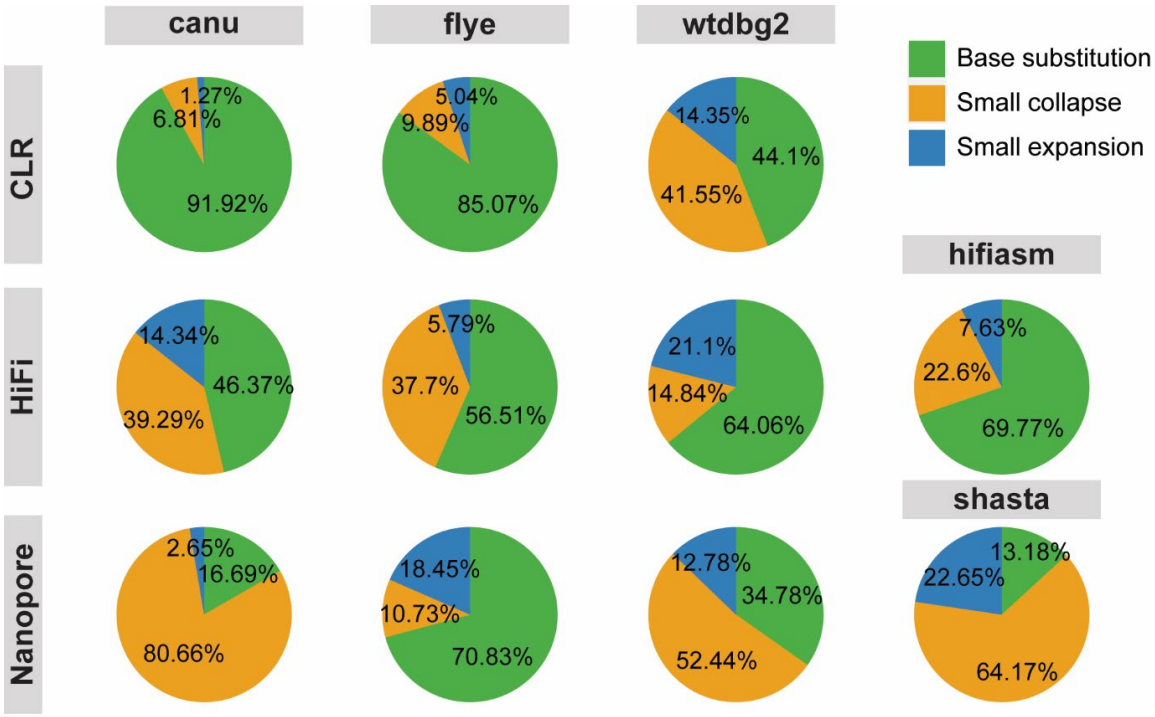

Figure S11 Pie charts of three types of small-scale errors in HG002 assemblies. The percentage of each error type in total errors is also labeled in each section.

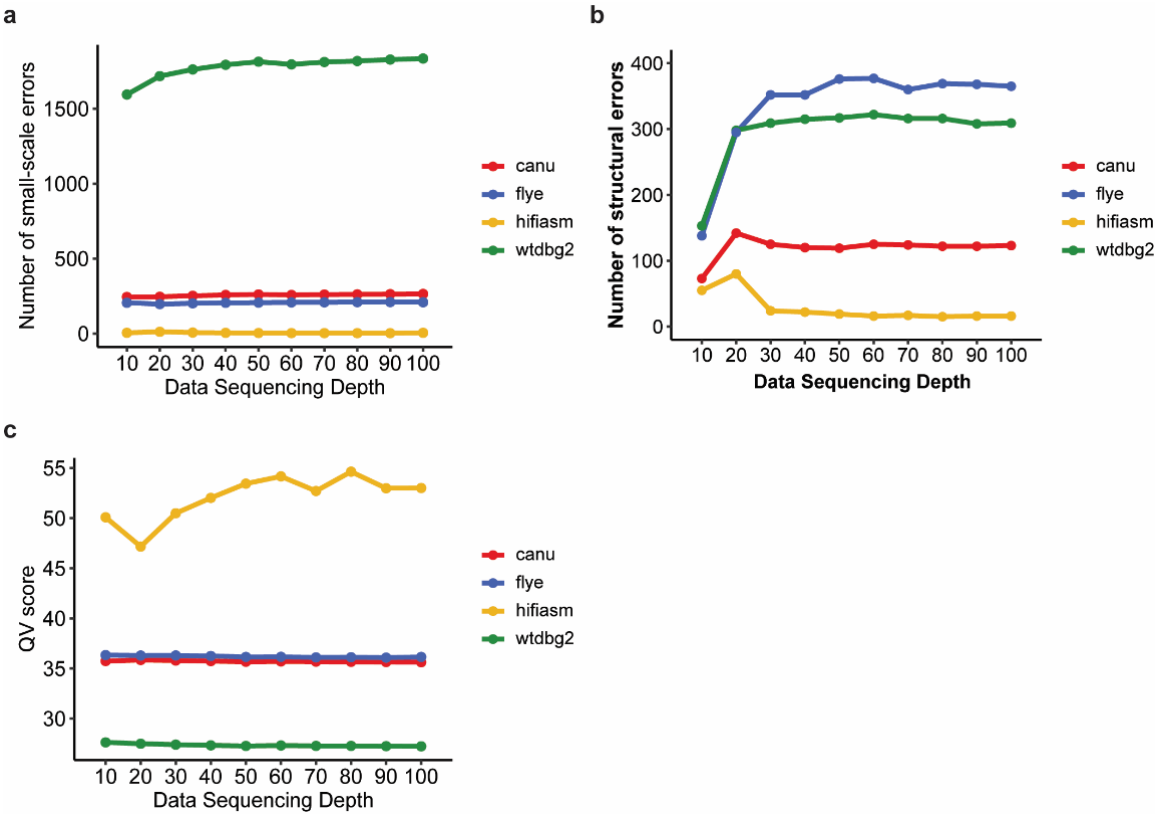

Figure S12 Inspector evaluation with down-sampled dataset. **a,b** Number of small-scale errors (**a**) and structural errors (**b**) reported from datasets with differing sequencing depth. The numbers of structural errors were fluctuant at 10-20X and stabilized after 30X. **c** QV score at different sequencing depth.

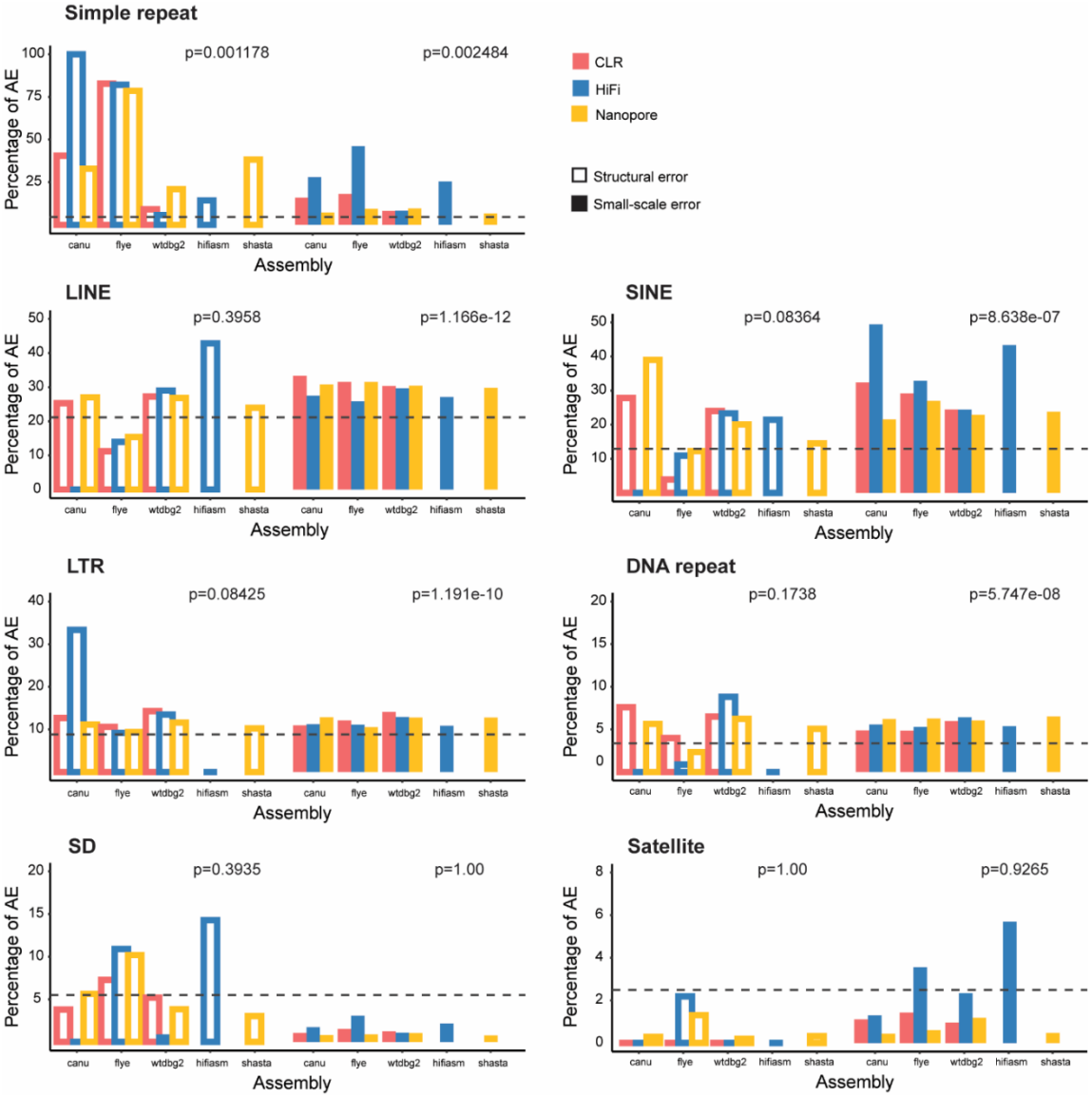

Figure S13 Proportion of assembly errors located in each type of repeat. P-value (on each panel, left: structural errors, right: small-scale errors) was calculated with one-sample t-test. The dashed line in each plot indicates the percentage of reference genome covered by that repeat type.

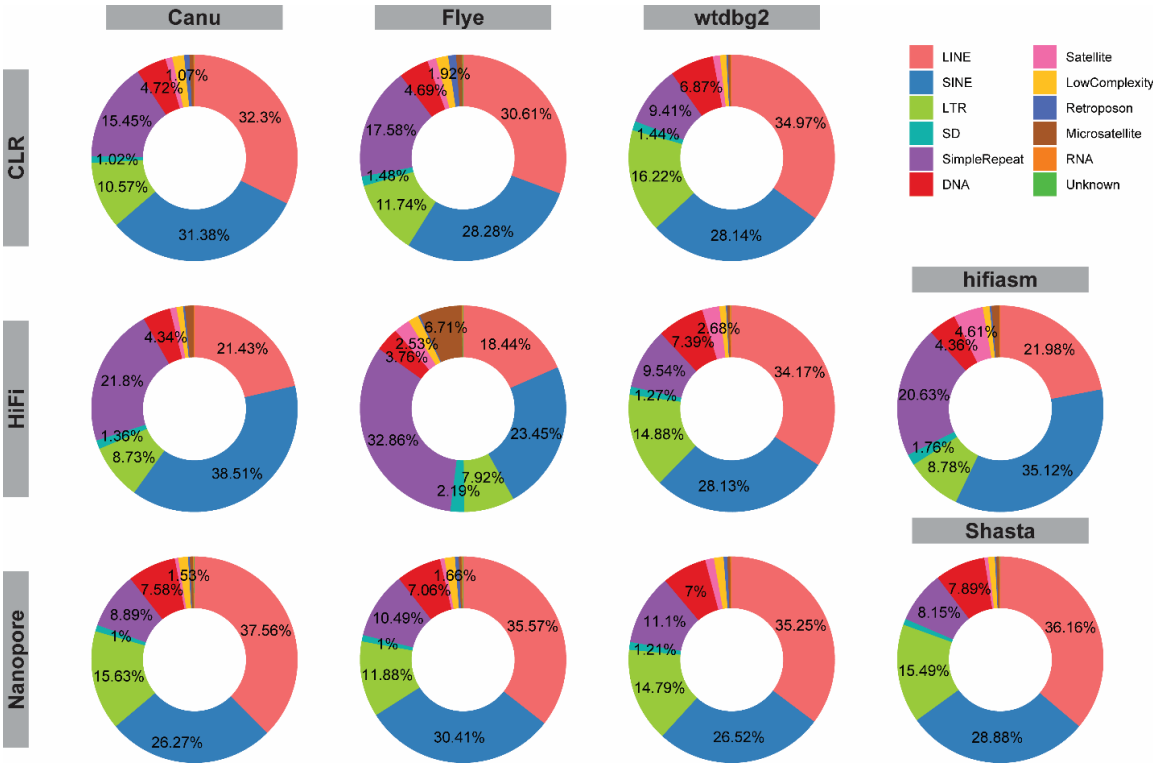

Figure S14 Repeat annotation for small-scale errors in HG002 assemblies. The proportions of sectors that are larger than 1% are marked in each plot.

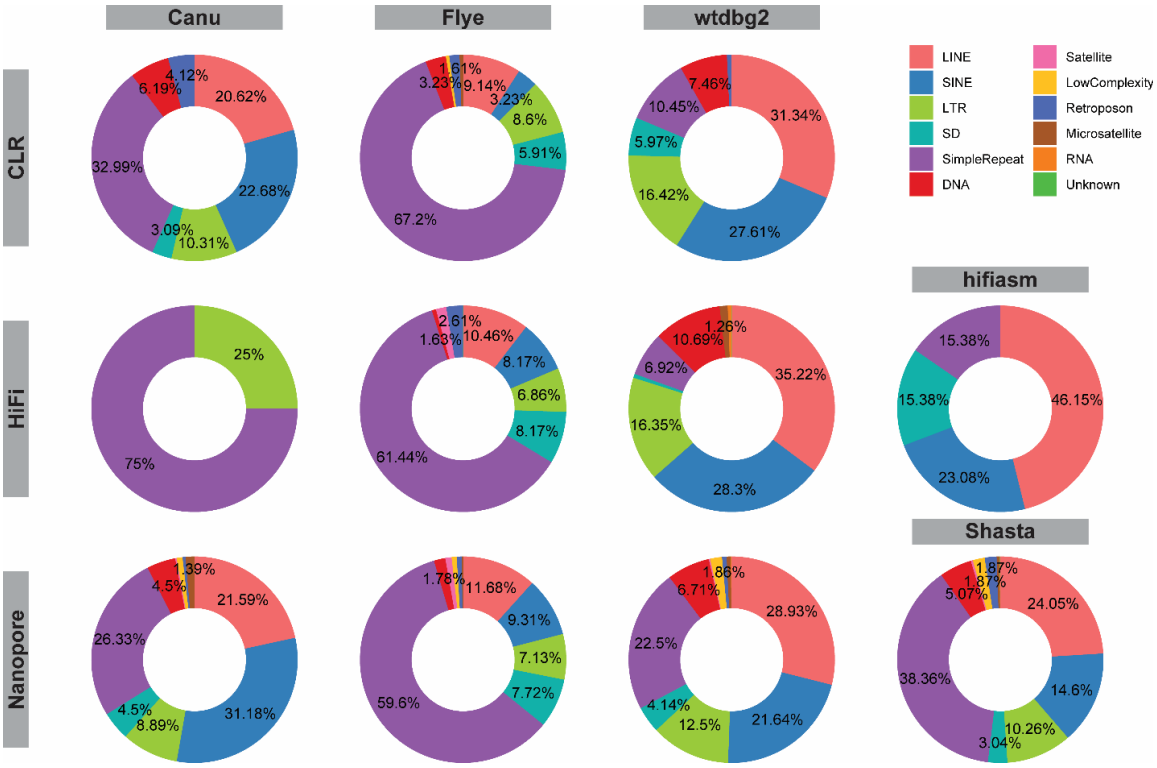

Figure S15 Repeat annotation for structural errors in HG002 assemblies. The proportions of sectors that are larger than 1% are marked in each plot.

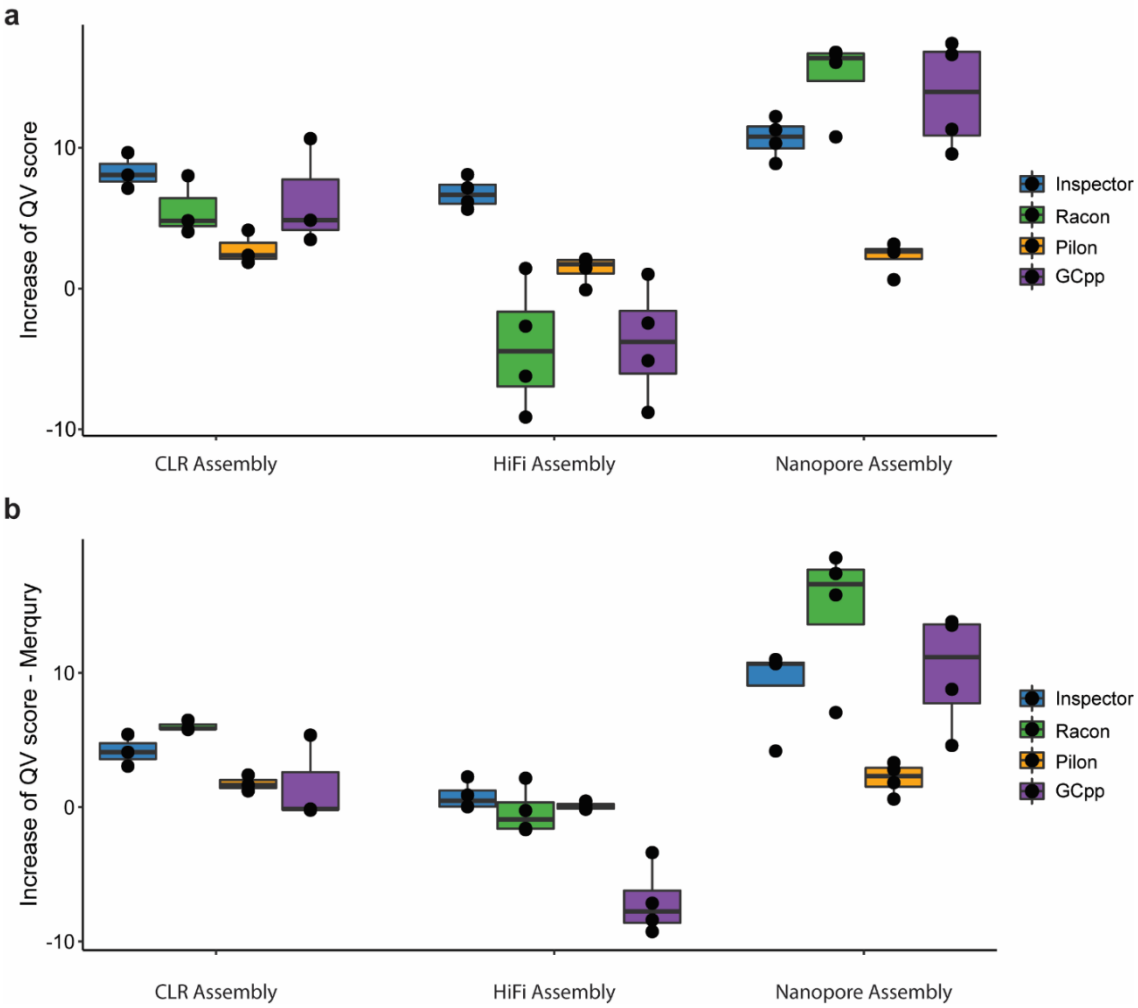

Figure S16 QV score improvement after polishing with PacBio HiFi reads. **a** QV score of polished assemblies estimated by Inspector. Inspector showed highest improvement in CLR and HiFi assemblies, and Racon showed highest improvement in Nanopore assembly. **b** QV score of polished assemblies estimated by Merqury. Inspector showed best improvement in HiFi assembly, and Racon showed best improvement in HiFi and Nanopore assemblies.

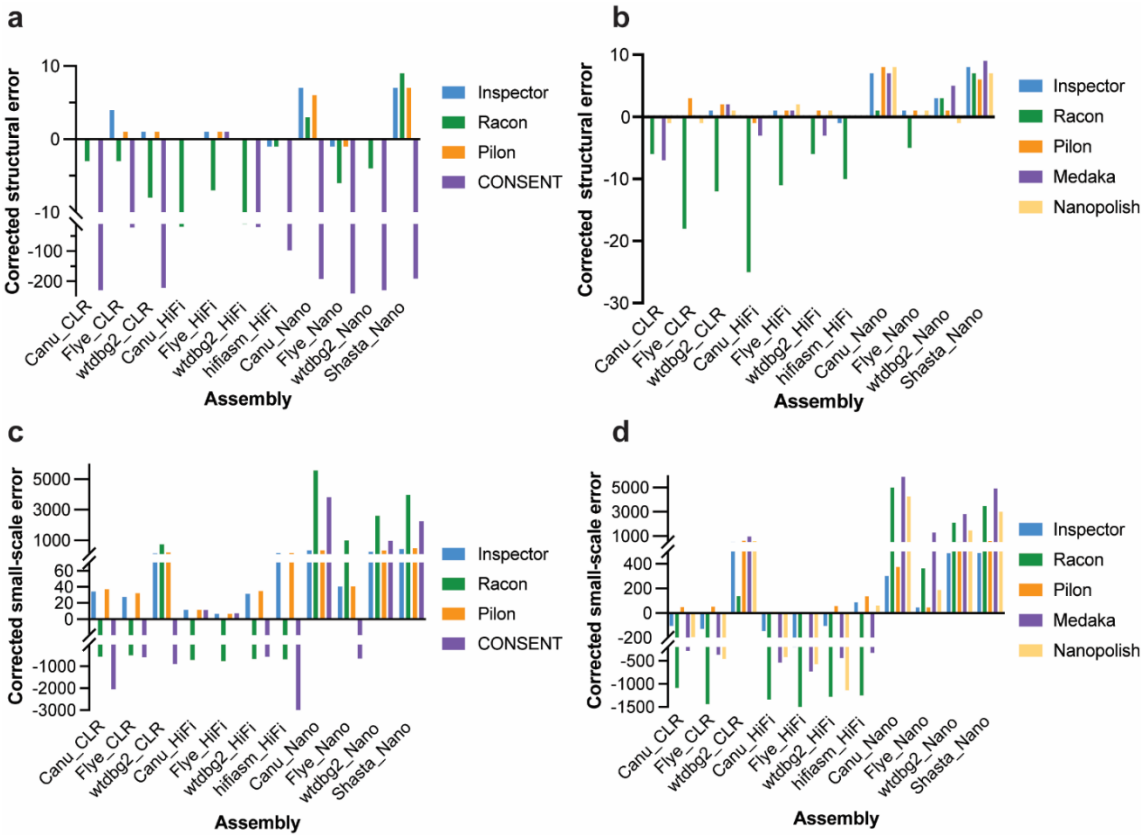

Figure S17 Assembly error correction with CLR and Nanopore data. **a,b** Number of corrected structural errors after polishing with CLR (**a**) and Nanopore (**b**) data. Inspector fixed most structural errors among tested polishing methods in 9 and 4 out of 11 assemblies in CLR and Nanopore data, respectively. **c,d** Number of corrected small-scale errors after polishing with CLR (**c**) and Nanopore (**d**) data.

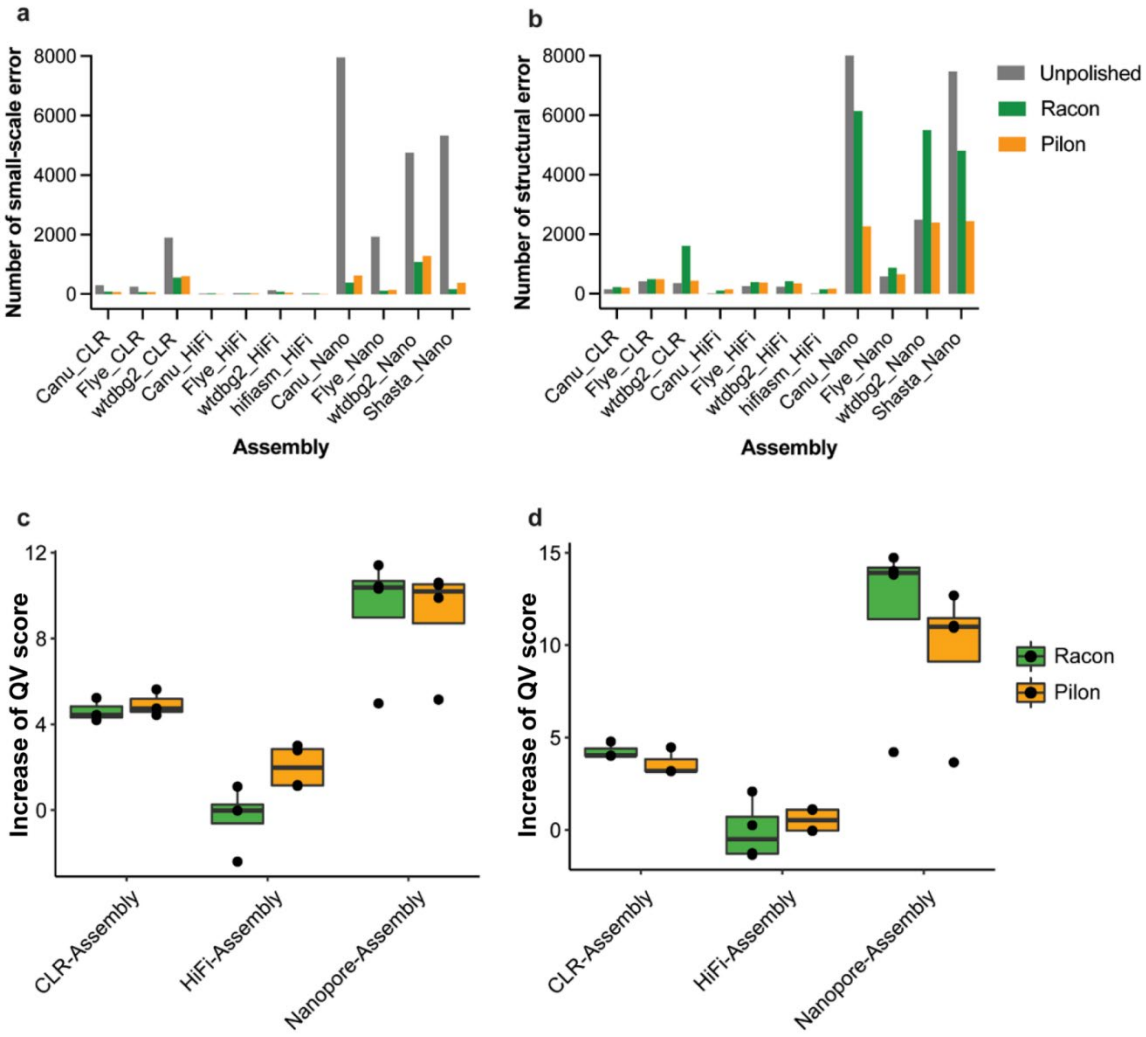

Figure S18 Polishing HG002 assemblies with Illumina dataset. **a** Number of small-scale errors in the original and polished assemblies. The number of small-scale errors was reduced after short-read polishing with both Racon and Pilon. **b** Number of structural errors before and after polishing with Illumina data. After short-read polishing, the number of structural errors increased in 9 Racon-polished and 8 Pilon-polished assemblies out of 11 total assemblies. **c,d** Improvement of QV scores after short-read polishing process estimated by Inspector (**c**) and Merqury (**d**). The QV scores of CLR and Nanopore assemblies were increased after short-read polishing, while the QV scores of HiFi assemblies showed minor improvement from short-read polishing.

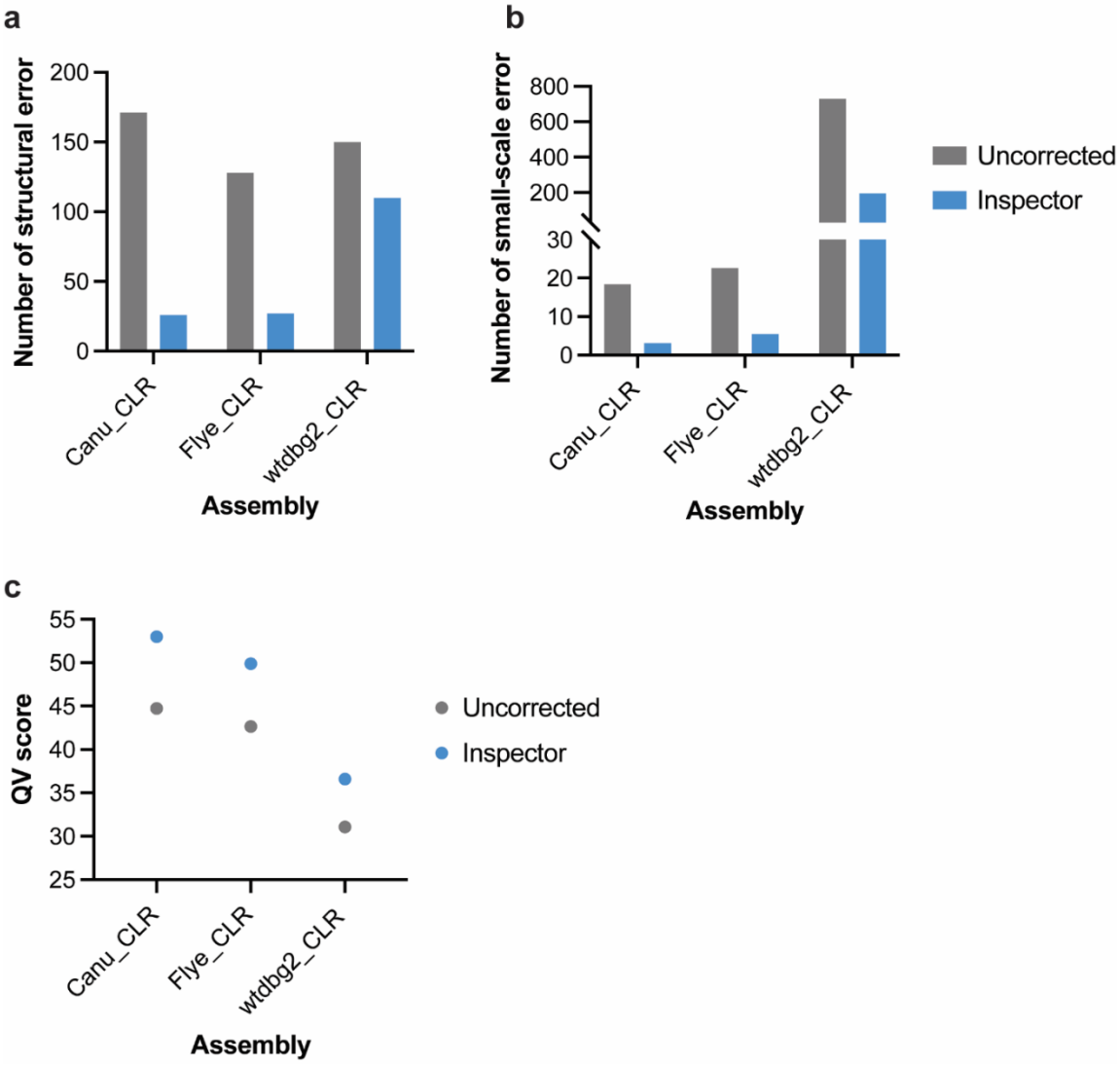

Figure S19 Inspector error correction in Anna's Hummingbird genome assemblies. **a,b** Number of structural (**a**) and small-scale (**b**) errors in the uncorrected and Inspector-corrected assemblies. Both structural and small-scale errors dropped after error correction. **c** QV score of uncorrected and Inspector-corrected assemblies. The QV score was increased in all three assemblies.

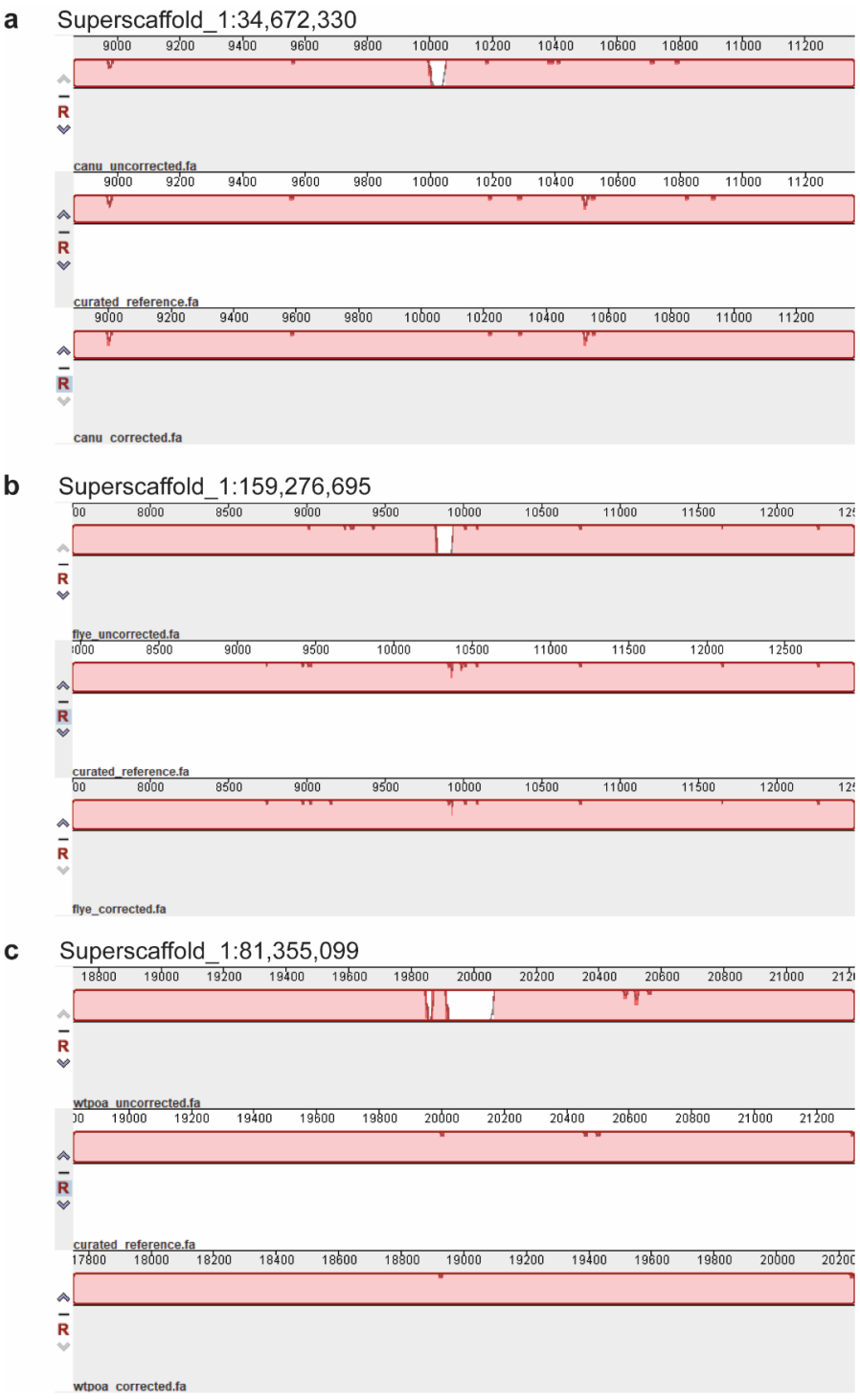

Figure S20 Example of structural errors corrected by Inspector error-correction module in Canu (**a**), Flye (**b**), and wtdbg2 (**c**) assemblies. The uncorrected contigs (top) showed inconsistency with the curated genome (middle), while the same regions in corrected contigs (bottom) were consistent with curated genome.

GENE FUSION DETECTION AND CHARACTERIZATION IN LONG-READ CANCER TRANSCRIPTOME SEQUENCING DATA WITH FUSIONSEEKER

by

YU CHEN, YIQING WANG, WEISHENG CHEN, ZHENGZHI TAN, YUWEI SONG, HUMAN GENOME STRUCTURAL VARIATION CONSORTIUM, HERBERT CHEN, ZECHEN CHONG

Cancer Research

Format adapted for dissertation

ABSTRACT

Gene fusions are prevalent in a wide array of cancer types with different frequencies. Long-read transcriptome sequencing technologies, such as PacBio Iso-Seq and Nanopore direct RNA sequencing, provide full-length transcript sequencing reads, therefore showing great potentials in gene fusion detection. In this work, we developed a novel method, FusionSeeker, to comprehensively characterize gene fusions in long-read cancer transcriptome data and reconstruct accurate fused transcripts from raw reads. FusionSeeker reports gene fusions occurred in both exonic and intronic regions, allowing comprehensive characterization of gene fusions in cancer transcriptomes. It reconstructs fused transcript sequences by correcting sequencing errors in the raw reads through partial order alignment algorithm. Using these accurate transcript sequences, FusionSeeker refines gene fusion breakpoint positions and predicts breakpoints at single basepair resolution. Overall, FusionSeeker enables users to discover gene fusions accurately using long-read data, which facilitates downstream functional analysis as well as diagnosis and targeted therapy.

INTRODUCTION

Gene fusions are recognized as important cancer-driving events for over 30 years [2]. They often play critical roles in tumorigenesis and progression and sometimes serve as therapeutic targets [3]. A large number of tools have been developed and applied to short-read cancer transcriptome sequencing data for gene fusion detection. However, it's always challenging to identify chimeric reads or discordant read pairs that represent gene fusions from short reads, especially given the innate splicing structures of isoforms. Recent development of long-read RNA sequencing technologies enables full-length transcript sequencing and may alleviate these issues, therefore showing great potential in gene fusion detection. However, only two tools, JAFFAL [4] and LongGF [5], are currently available for long-read gene fusion detection, and their performance is limited when detecting gene fusions occurred in intronic regions. Accurate sequences of the reported gene fusions also remain unknown, which limits further functional analysis of identified gene fusions.

Here, we present FusionSeeker, a long-read gene fusion caller to accurately identify gene fusion events and reconstruct their transcript sequences. FusionSeeker takes read alignment file and gene annotation file as input and outputs a list of confident gene fusions and their transcript sequences (**Fig. 1**). It first scans the read alignments for candidate fusions when a single read is aligned to two or more genes. Candidate fusions are then grouped according to these genes and clustered with the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm into gene fusion calls. The

gene fusion calls are filtered based on the number of supporting reads to remove noise signals caused by sequencing errors and incorrect read alignments. FusionSeeker then performs a partial order alignment (POA) using fusion-containing reads to generate a consensus transcript sequence for each confident gene fusion.

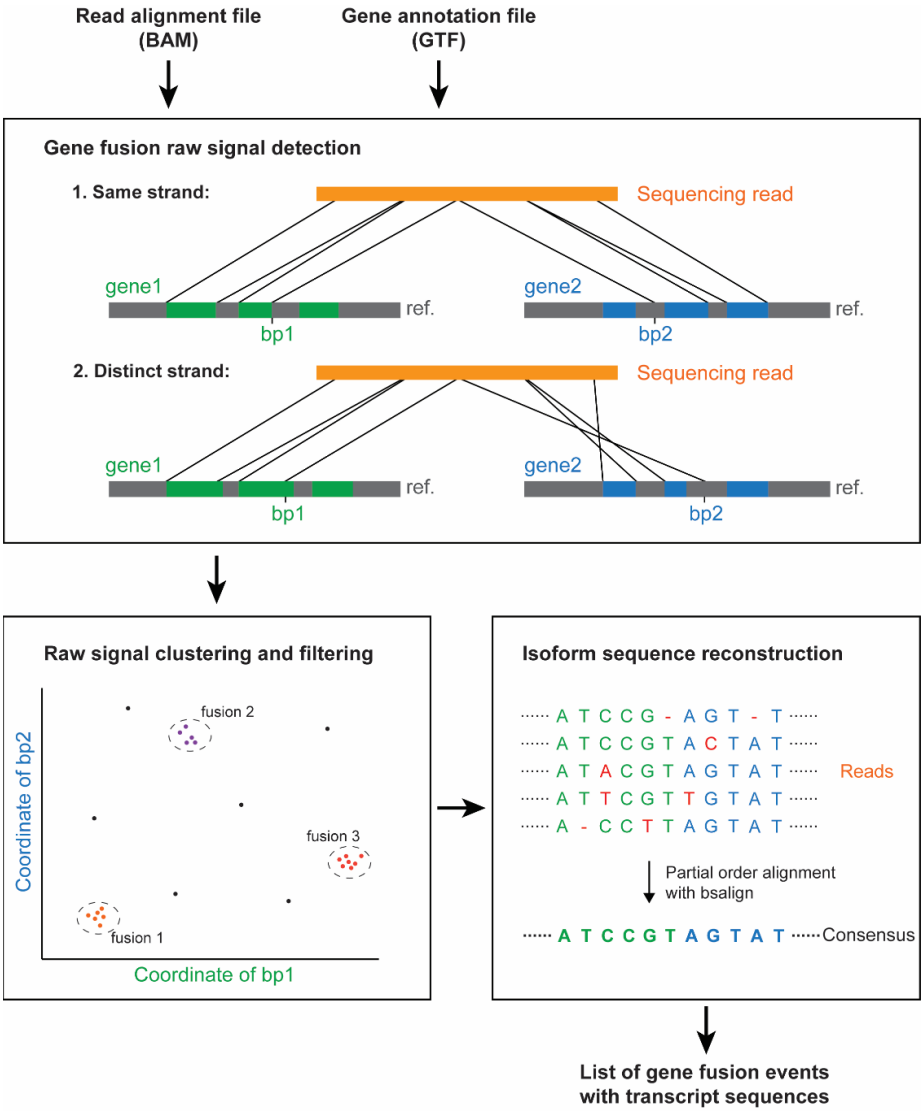

Figure 1. Workflow of FusionSeeker. FusionSeeker scans the input file of the read alignments for split read alignments and records candidate fusions of gene fusions when two segments from one read are aligned to two distinct genes. It then clusters the candidate fusions into gene fusion calls and removes noise calls supported by only a few reads. For each fusion call, FusionSeeker generates a consensus transcript sequence by performing a partial order alignment with fusion-containing reads. The final output of FusionSeeker includes a list of confident gene fusion events and corresponding transcript sequences.

MATERIALS AND METHODS

1. Gene fusion candidate detection

FusionSeeker first scans all read alignments for split-read patterns. In order to quickly annotate read alignments, FusionSeeker generates a list containing the coordinates of each gene and its exons on every chromosome based on the input genome annotation file (GTF). Input BAM file is then processed chromosome by chromosome. Reads with only one alignment are skipped to reduce computational burden. For reads with multiple alignments (with SA tags), FusionSeeker annotates each alignment and records essential information, including chromosome, alignment start and end positions, length of clipped sequences on both sides, read name, strand, mapping quality, simplified CIGAR tag, etc. As candidate fusion detection process is the most time-consuming step, FusionSeeker can process each chromosome in parallel to reduce the overall runtime. After all alignments are processed, FusionSeeker checks the alignment information from the same read and reports a candidate fusion when:

1) two breakpoints from one read are annotated to two distinct genes (Gene A and Gene B),
2) length of alignment is longer than 100bp on both genes,
3) Length of overlap between two alignments (the part of read sequences present in both alignments) is shorter than 100bp and 50% of the shorter alignment,
4) Coordinates of Gene A and Gene B do not overlap in the GTF file,
5) Gene A is not an antisense sequence of Gene B.

2. Gene fusion signal clustering and filtering

Candidate fusions are first grouped based on gene names, for instance, Gene A and Gene B. The candidate fusions from the same pair of fused genes are then clustered based on the breakpoint positions on the two genes. To achieve this, a density-based spatial clustering of applications with noise algorithm (DBSCAN) is adopted to cluster the candidate fusions, with a default of maximal distance of 20bp for high accuracy reads and 40bp for noisy reads. Next, the candidate fusions from each cluster are merged into a gene fusion call, with temporary breakpoint positions as the mean values from the candidate fusions. All gene fusion calls are then filtered based on the number of fusion-supporting reads. By default, the cutoff of minimal supporting reads N_{min} is set as $N_{min} = N_{can}\ /50000 + 3$, where N_{can} is the total number of the candidate fusions detected in the input dataset. Fusion calls supported by more than N_{min} reads are reported as confident gene fusion calls.

3. Fused transcript reconstruction and breakpoint refinement

For each gene fusion event, FusionSeeker extracts the sequences of the fusion-supporting reads from the BAM file and writes into a new FASTQ file. It then performs Partial Order Alignment (POA) for each call independently using bsalign (https://github.com/ruanjue/bsalign). All consensus sequences generated from POA are combined into a FASTA file and linked to each gene fusion call with its ID. When a reference genome is provided, FusionSeeker then aligns all the transcript sequences to the reference genome with minimap2 [6]. The precise breakpoint positions of each gene

fusion call are inferred from the transcript sequence alignment and used to replace the temporary positions inferred from the candidate fusions.

Simulation and benchmark methods can be found in **Supplementary Note 2**.

RESULTS

1. Benchmark gene fusion detection on the simulated datasets

We first benchmarked the accuracy of gene fusion detection of FusionSeeker on the simulated datasets. A total of 150 gene fusion transcripts (100 with breakpoints in exons, 50 in introns) were randomly generated and assigned to different expression levels (10x, 50x, and 100x). PacBio Iso-Seq-like and Nanopore-like reads were simulated with pbsim [7] and Badread (v0.2.0) [8] and then aligned to the reference genome. FusionSeeker and another two long-read gene fusion callers, JAFFAL and LongGF, were used to detect gene fusions from the simulated reads. We repeated the simulation for three times, and FusionSeeker consistently achieved the highest F1 score among the three tools in both Iso-Seq and Nanopore datasets (**Table 1**). In all three simulated datasets, FusionSeeker identified more true-positive events than the other two tools, with slightly more false-positive calls than LongGF (**Fig. S1**). The higher recall of FusionSeeker was mainly beneficial from its ability to detect gene fusions located in intronic regions, where FusionSeeker identified 94.67% of intronic events while JAFFAL and LongGF only reported 14.67% and 54.67%, respectively, using Iso-Seq data (**Table 1** and **Table S1**). In general, all three fusion callers achieved higher recall in detecting fusions with high and medium expression levels than fusions with low expression level (**Table S2**).

Approximately 67% of the gene fusions missed by FusionSeeker were from the low-expression-level group, and the missing was caused by the low coverage of reads.

Table 1. The accuracy of gene fusion detection on the simulated datasets

	FusionSeeker			JAFFAL			LongGF		
	Recall	Precision	F1 score	Recall	Precision	F1 score	Recall	Precision	F1 score
Iso-Seq									
Exonic	96.00	96.88	96.32	69.33	96.57	80.72	96.00	97.03	96.51
Intronic	94.67	90.00	92.28	14.67	66.67	24.04	54.67	94.87	69.36
Total	**95.56**	93.89	**94.71**	51.11	82.73	63.15	82.22	**96.14**	88.58
Nanopore									
Exonic	99.00	94.98	96.95	73.00	96.35	83.06	98.00	96.70	97.35
Intronic	99.33	78.72	87.84	18.00	53.81	26.98	56.67	91.80	70.08
Total	**99.11**	87.65	**93.03**	54.67	82.23	65.62	84.22	**95.26**	89.36

Recall, precision, and F1 score in the table are the mean values of three replicate simulation datasets. Highest recall, precision, and F1 score among the three fusion callers are marked as bold.

We then evaluated the fused transcript sequences generated by FusionSeeker. To generate high-accuracy transcript sequences, FusionSeeker performs a partial order alignment using fusion-containing reads and calculates a consensus sequence for each gene fusion event. In the simulated datasets, FusionSeeker reconstructed full-length fused transcripts for more than 99.5% of events, with average sequence identities of 99.87% and 99.14% using Iso-Seq and Nanopore reads, respectively (**Table S3**). When aligned to the reference genome, the FusionSeeker transcript sequences showed a better identity than raw reads (**Fig. S2**). Taken together, we have demonstrated that FusionSeeker can accurately identify gene fusions and report full-length fused transcript sequences in the simulated datasets.

2. Gene fusion discovery in cancer transcriptomes

We then applied the three gene fusion callers on three cancer cell lines, SKBR-3, MCF7, and HCT116. The PacBio Iso-Seq and Nanopore reads of each cell line were downloaded and aligned to the human reference genome [9-12]. In the SKBR-3 cell line, FusionSeeker identified 31 gene fusions, among which 15 events have been previously discovered and validated (**Table 2**) [12-15]. Three of the previous studies for gene fusion detection in SKBR-3 were based on short-read RNA sequencing data [13-15], except the Nattestad *et al* [12] which used the PacBio Iso-Seq dataset. Tested on this Iso-seq data, FusionSeeker showed a better consistency with Nattestad *et al.* than the other short-read results (**Table S4**). JAFFAL and LongGF identified 13 and 10 previously validated gene fusions, respectively. Comparing the gene fusion lists of three callers, 8 gene fusions were reported by all the three tools, 3 gene fusions were reported by both FusionSeeker and JAFFAL, and 3 gene fusions were reported by both JAFFAL and LongGF (**Fig. 2A**). There were 19 FusionSeeker-unique, 11 JAFFAL-unique, and 5 LongGF-unique events. We cross-validated these unique gene fusion events with long-read DNA sequencing data and considered a gene fusion as validated when at least 3 DNA sequencing reads were aligned to both genes (**Fig. S3**). 17 out of 19 (89.47%) FusionSeeker-unique gene fusions were validated by DNA sequencing, which was higher than JAFFAL (3/11, 27.27%) and LongGF (3/5, 60.00%). In particular, with further investigation we observed a 4-hop intronic gene fusion from FusionSeeker-unique calls, CSNK2A1:NCOA3:MMP24OS:TSHZ2, which was also supported by DNA sequencing data (**Fig. S4**).

Table 2. Detection of previously validated gene fusions in cancer cell lines

		FusionSeeker		JAFFAL		LongGF	
Cell line	Data type	Reported	Previously validated	Reported	Previously validated	Reported	Previously validated
SKBR-3							
	Iso-Seq	30	**15**	25	13	16	10
MCF-7							
	Iso-Seq	172	21	184	**23**	285	20
	Nanopore	61	**20**	34	18	41	**20**
HCT-116							
	Iso-Seq	3	**1**	2	**1**	2	**1**
	Nanopore	17	**1**	12	**1**	10	**1**

Reported, number of gene fusions reported by each fusion caller. Previously validated, number of previously validated gene fusions detected by each fusion caller.

In MCF-7 cell line, FusionSeeker identified 172 gene fusions in Iso-Seq dataset and 61 gene fusions in Nanopore dataset (**Table 2**), with 21 and 20 previously validated gene fusions identified using Iso-Seq and Nanopore datasets, respectively (**Table S5**). In HCT-116 cell line, FusionSeeker reported 3 and 17 gene fusions in Iso-Seq and Nanopore dataset, respectively. In particular, a previously known gene fusion, TXLNG:SYAP1, in MCF-7 cell line has two validated alternative breakpoint positions in TXLNG, with one located in the first exon and the other located in the first intron of TXLNG [15]. FusionSeeker reported both exonic and intronic breakpoints for this fusion event, while JAFFAL and LongGF only reported the exonic breakpoint and missed the intronic breakpoint (**Fig. S5**). The few previously validated events detected by JAFFAL but not by FusionSeeker were supported by ≤4 reads and therefore failed to pass the filter of FusionSeeker (**Table S6**). When comparing gene fusion callsets of the three callers, 47 and 19 gene fusions were reported by all three callers in Iso-Seq and Nanopore dataset, respectively (**Fig. 2B**). Gene fusions reported by JAFFAL or LongGF but not by

FusionSeeker were usually supported by fewer reads, with 88.35% of them supported by ≤3 reads in MCF-7 Iso-Seq dataset (**Fig. S6**). Within the 77 and 29 FusionSeeker-unique calls in MCF-7 Iso-Seq and Nanopore dataset, we designed PCR primers for 10 most confident novel events and validated 7 of them using RNA extracted from MCF-7 cell line (**Table S7**). All four events discovered in both Iso-Seq and Nanopore datasets were validated by PCR.

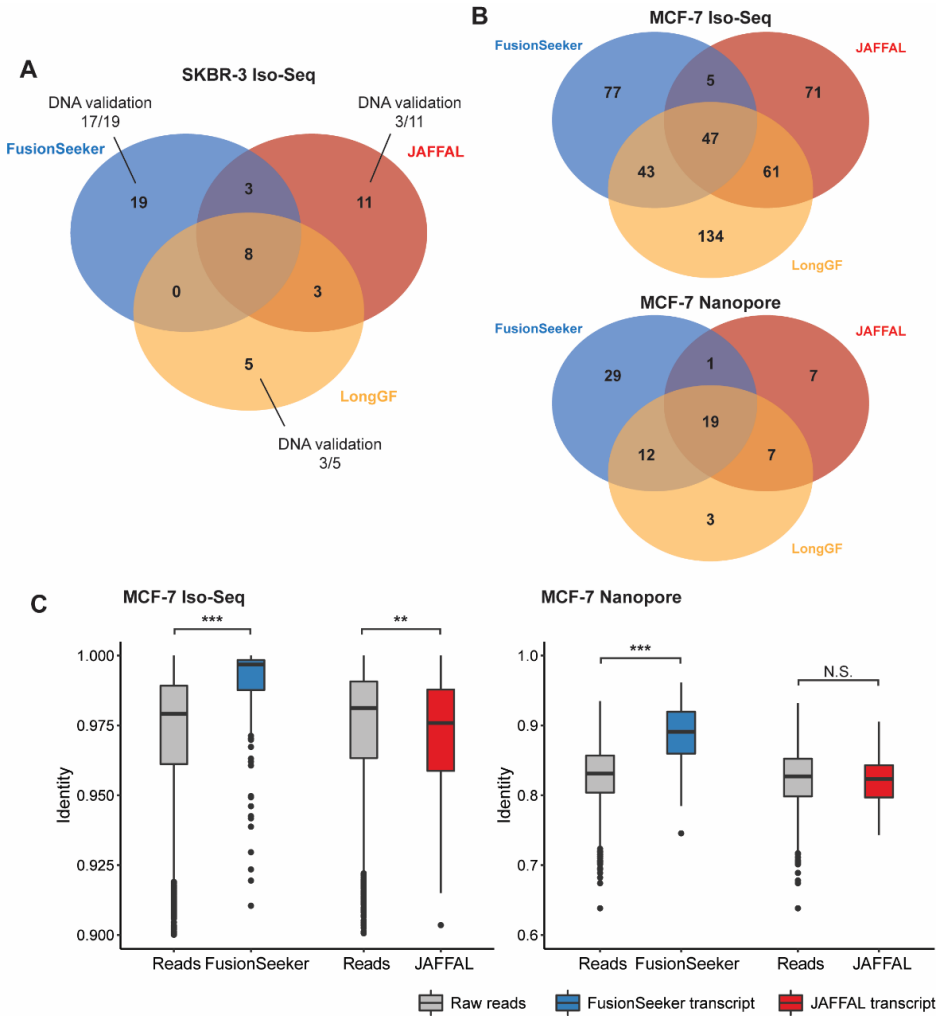

Figure 2. Gene fusion discovery in cancer cell lines. A Venn diagram of gene fusion calls by FusionSeeker, JAFFAL, and LongGF in SKBR-3 cell line. **B** Venn diagrams of gene fusion calls by the three fusion callers in MCF-7 cell line using Iso-Seq (top) and Nanopore direct RNA sequencing (bottom) data. **C** The identity of raw reads and transcript sequences reported by FusionSeeker and JAFFAL in MCF-7 Iso-Seq (left) and Nanopore (right) direct RNA sequencing dataset. **, p<0.01. ***, p<0.001. N.S., not significant. The p-values were calculated by Mann-Whitney U test.

When comparing two lists of gene fusion calls from Iso-Seq and Nanopore datasets for each caller, we observed slightly higher overlapping ratio in FusionSeeker callsets than JAFFAL and LongGF, with Jaccard index of 0.1208 for FusionSeeker, 0.0741 for JAFFAL, and 0.0584 for LongGF in MCF-7 cell line, respectively (**Fig. S7**). This overall low overlapping rate was probably caused by the evolution of the cell line or inconsistent sequencing depth on each gene in the two datasets during sequencing (**Fig. S8**). This systemic difference may need further investigation.

We then applied three fusion callers on non-cancer datasets from Human Genome Structural Variation Consortium (HGSVC) to assess the false discovery rate of three tools. In all the 12 non-cancer samples, FusionSeeker reported the fewest number of gene fusions, suggesting that FusionSeeker had lowest false discovery rates among the three tested fusion callers (**Table S8**). We have also applied FusionSeeker on a patient sample with acute myeloid leukemia (AML) to demonstrate its clinical utility [5]. FusionSeeker identified a pre-validated gene fusion between RUNX1 and RUNX1T1 and reported another 7 confident gene fusion events in the patient sample (**Table S9**).

3. Isoform sequence reconstruction with de novo assembly

We next evaluated the transcript sequences generated by FusionSeeker. Compared to the raw reads, FusionSeeker transcript sequences showed significant higher identity with reference gene sequences in both Iso-Seq and Nanopore datasets of MCF-7 cell line (**Fig. 2C**). JAFFAL also reported one of the fusion-containing reads as the transcript sequence, which showed no significant difference in identity comparing with the raw reads. In the Iso-Seq dataset of the SKBR3 and the Nanopore dataset of the HCT-116 cell

lines, FusionSeeker reported more accurate transcript sequences than the raw reads, while transcript sequences reported by JAFFAL showed no significant differences (**Fig. S9**). There was no significant difference between FusionSeeker transcript sequences and raw reads in HCT-116 Iso-Seq dataset, likely due to only three gene fusions were reported. Note that the identity calculated by comparing with the reference is an underestimation of transcript sequence accuracy, owing to the presence of genetic variants in these cell lines. These genetic variants can often be maintained in the transcript sequences (**Fig. S10**).

DISCUSSION

In this work, we presented FusionSeeker for gene fusion detection in long-read cancer transcriptome sequencing data. FusionSeeker can detect gene fusions in both exonic and intronic regions. Based on simulation and three cancer cell line data, we have demonstrated that FusionSeeker outperformed existing methods in characterizing gene fusion events. Besides, we have both orthogonally and experimentally validated many gene fusion events only detected by FusionSeeker. These novel gene fusions may be important for tumorigenesis and progression, which deserves further investigation. Since the long-read sequencing platform can almost generate full-length transcripts, FusionSeeker provides accurate full-length fusion transcripts based on an assembly approach. The full-length fusion transcripts may facilitate downstream functional and clinical research.

After candidate fusion detection, FusionSeeker used DBSCAN to cluster candidate fusions that share the same breakpoints. DBSCAN was implanted as it does not require pre-determined number of clusters, which allows FusionSeeker to report gene fusions with one or multiple breakpoints in the same gene pair. DBSCAN can also robustly exclude outliers while clustering, which is necessary in this case as there are often abundant noise signals in long-read RNA sequencing read alignments.

Data access

The source code of FusionSeeker is available at https://github.com/Maggi-Chen/FusionSeeker under MIT license, and the scripts used for benchmark in the manuscript are available at https://github.com/Maggi-Chen/FS_code. The Nanopore direct RNA sequencing data of the MCF-7 and HCT-116 cell lines are available at https://github.com/GoekeLab/sg-nex-data/ [11]. The PacBio Iso-Seq sequencing data of the MCF-7 and the HCT116 cell lines are available at SRA under the accessions SRP055913 [10] and SRP091981 [9]. The PacBio Iso-Seq and CLR sequencing data of the SKBR-3 cell line are downloaded from SRA under accession SRP150606 [12]. PacBio Iso-Seq data of HGSVC samples are available at HGSVC data portal (https://www.internationalgenome.org/data-portal/). AML patient data is downloaded from SRA under SRR12048357 [5].

Acknowledgements

The authors acknowledge Dr. Anna Sorace for providing MCF-7 cell line. This work was supported by a 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 a funding from Robert Reed Foundation to H.C.

REFERENCES

1. Chen Y, Wang Y, Chen W, Tan Z, Song Y, Human Genome Structural Variation Consortium N, Chen H, Chong Z: Gene fusion detection and characterization in long-read cancer transcriptome sequencing data with FusionSeeker. *Cancer Res* 2022.

2. Edwards PA: Fusion genes and chromosome translocations in the common epithelial cancers. *J Pathol* 2010, 220:244-254.

3. Forsythe A, Zhang W, Phillip Strauss U, Fellous M, Korei M, Keating K: A systematic review and meta-analysis of neurotrophic tyrosine receptor kinase gene fusion frequencies in solid tumors. *Ther Adv Med Oncol* 2020, 12:1758835920975613.

4. Davidson NM, Chen Y, Sadras T, Ryland GL, Blombery P, Ekert PG, Goke J, Oshlack A: JAFFAL: detecting fusion genes with long-read transcriptome sequencing. *Genome Biol* 2022, 23:10.

5. Liu Q, Hu Y, Stucky A, Fang L, Zhong JF, Wang K: LongGF: computational algorithm and software tool for fast and accurate detection of gene fusions by long-read transcriptome sequencing. *BMC Genomics* 2020, 21:793.

6. Li H: Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018, 34:3094-3100.

7. Ono Y, Asai K, Hamada M: PBSIM: PacBio reads simulator--toward accurate genome assembly. *Bioinformatics* 2013, 29:119-121.

8. Wick RR: Badread: simulation of error-prone long reads. *Journal of Open Source Software* 2019, 4:1316.

9. Centre BCR: Transcriptome dynamics of CLK dependent exon recognition and conjoined gene formation revealed with a novel small molecule inhibitor. 2017.

10. Iowa Uo: Full-length transcripts of the MCF-7 breast cancer cell line by PacBio SMRT sequencing. 2015.

11. Ying Chen NMD, Yuk Kei Wan, Harshil Patel, Fei Yao, Hwee meng Low, Christopher Hendra: A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines. BioRxiv; 2021.

12. Nattestad M, Goodwin S, Ng K, Baslan T, Sedlazeck FJ, Rescheneder P, Garvin T, Fang H, Gurtowski J, Hutton E, et al: Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res* 2018, 28:1126-1135.

13. Chen K, Navin NE, Wang Y, Schmidt HK, Wallis JW, Niu B, Fan X, Zhao H, McLellan MD, Hoadley KA, et al: BreakTrans: uncovering the genomic architecture of gene fusions. *Genome Biol* 2013, 14:R87.

14. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, Rye IH, Nyberg S, Wolf M, Borresen-Dale AL, Kallioniemi O: Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol* 2011, 12:R6.

15. Inaki K, Hillmer AM, Ukil L, Yao F, Woo XY, Vardy LA, Zawack KF, Lee CW, Ariyaratne PN, Chan YS, et al: Transcriptional consequences of genomic structural aberrations in breast cancer. *Genome Res* 2011, 21:676-687.

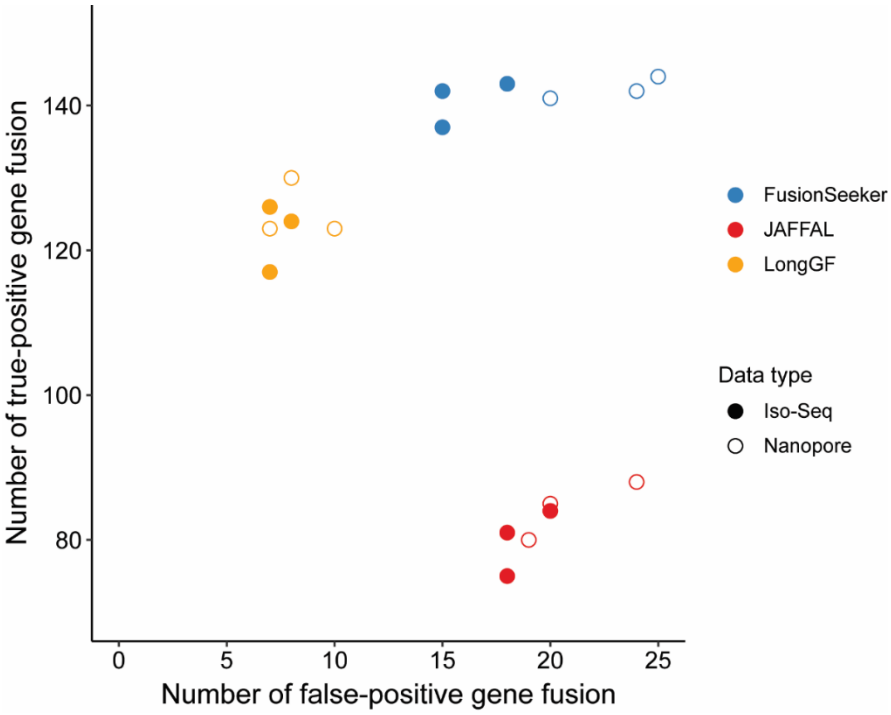

Figure S1. Number of true-positive and false-positive gene fusion calls in simulated datasets.

Table S1. Gene Fusion detection on simulated datasets

		Iso-Seq				Nanopore			
Tool	**Dataset**	**Total**	**TP**	**TP-exon**	**TP-intron**	**Total**	**TP**	**TP-exon**	**TP-intron**
FusionSeeker	Rep1	147	139	96	43	153	142	94	48
	Rep2	149	139	92	47	164	145	96	49
	Rep3	149	141	95	46	159	146	98	48
JAFFAL	Rep1	86	71	68	3	93	77	71	6
	Rep2	96	75	68	7	108	81	73	8
	Rep3	96	84	72	12	99	88	75	13
LongGF	Rep1	120	116	94	22	126	121	96	25
	Rep2	136	129	96	33	139	131	98	33
	Rep3	129	125	98	27	133	127	100	27

Total, total number of gene fusion events reported by each tool. TP, true positive events. TP-exon, true positive events located in exomes. TP-intron, true positive events located in introns.

Table S2. Recall of gene fusion discovery at different expression levels in simulation

		Iso-Seq			Nanopore		
		High	Medium	Low	High	Medium	Low
FusionSeeker							
	Rep1	96.00	97.96	90.20	98.00	100.0	100.0
	Rep2	100.0	94.12	88.64	100.0	98.04	95.45
	Rep3	100.0	100.0	91.84	100.0	100.0	100.0
	Mean	**98.67**	**97.36**	**90.22**	**99.33**	**99.35**	**98.48**
JAFFAL							
	Rep1	54.00	55.10	33.33	62.00	57.14	35.29
	Rep2	49.09	56.86	43.18	52.73	62.75	45.45
	Rep3	57.45	57.41	53.06	61.70	61.11	53.06
	Mean	53.51	56.46	43.19	58.81	60.33	44.60
LongGF							
	Rep1	82.00	85.71	64.71	84.00	87.76	70.59
	Rep2	81.82	86.27	90.91	81.82	90.20	90.91
	Rep3	80.85	81.48	87.76	80.85	81.48	91.84
	Mean	81.56	84.49	81.12	82.22	86.48	84.44

Highest mean recall in each expression level among three tested fusion callers is labeled as bold.

Table S3. Transcript reconstruction of FusionSeeker on simulated datasets

Data type	Dataset	Detected GF	Full-length transcript	Identity
Iso-Seq	Replicate 1	139	139	99.99
	Replicate 2	139	138	99.93
	Replicate 3	141	141	99.69
Nanopore	Replicate 1	142	142	99.08
	Replicate 2	145	144	99.19
	Replicate 3	146	145	99.15

Transcript sequences output by FusionSeeker were considered as 'full-length' when more than 95% of the simulated fused transcript sequence was reconstructed.
Detected GF, number of detected gene fusion in each dataset. Full-length transcript, number of gene fusion with full-length transcript sequences. Identity, identity of reconstructed transcript sequence compared to the ground-truth sequences.

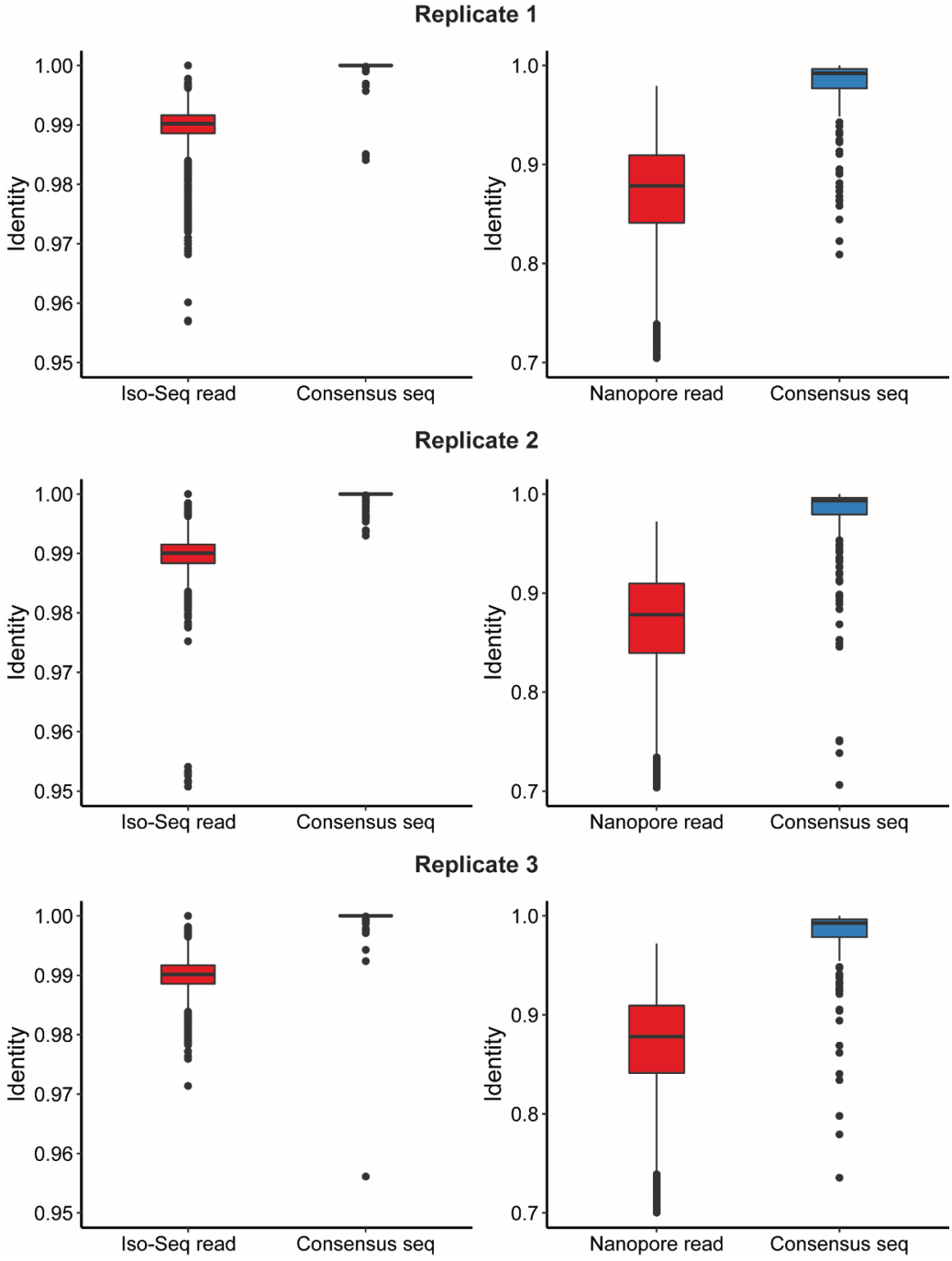

Figure S2. Identity of simulated raw reads and consensus transcript sequences of FusionSeeker. The consensus sequences generated by partial order alignment showed higher identity than raw reads in both PacBio Iso-Seq (left) and Nanopore (right) datasets.

Table S4. Detection of previously validated gene fusions in SKBR-3 cell line

	Previous validation					Tested callers		
Gene fusion	Number	Edgren	Inaki	Nattestad	Chen	FusionSeeker	JAFFAL	LongGF
TATDN1:GSDMB	3		✓	✓	✓	✓	✓	✓
RARA:PKIA	3	✓	✓		✓			
DHX35:ITCH	3	✓	✓	✓		✓	✓	✓
SUMF1:LRRFIP2	3	✓		✓	✓	✓	✓	✓
TBC1D31:ZNF704	3	✓	✓	✓		✓	✓	✓
ANKHD1:PCDH1	3	✓	✓		✓			
CYTH1:EIF3H	2	✓		✓			✓	
CCDC85C:SETD3	2	✓			✓			
NFS1:PREX1	1	✓						
ATAD5:TLK2	2		✓	✓		✓	✓	✓
PREX1:CPNE1	2		✓		✓	✓		
DEPDC1B:PDE4D	1		✓			✓	✓	✓
TAF2:COLEC10	1		✓			✓	✓	✓
TRIO:FBXL7	1		✓			✓	✓	
RPTOR:RNF213	1		✓					
PBRM1:WDR82	1		✓				✓	✓
BLOC1S6:AKAP13	1		✓					
VSTM2L:CTNNBL1	1		✓					
COL14A1:MTSS1	1		✓					
CBX3:CCDC32	1		✓					✓
RANBP10:PSKH1	1		✓					
SAMD12:MTBP	2			✓	✓	✓		
KLHDC2:SNTB1	1			✓		✓	✓	✓
PVT1:LINC00536	1			✓		✓	✓	
MECOM:LMCD1-AS1	1			✓		✓	✓	
RAD51B:SEMA6D	1			✓		✓		
TOX2:STAU1	1			✓		✓		
LINC01524:PHF20	1			✓		✓		

Gene fusions detected by FusionSeeker with option "—min_supp 3" were marked green.
Number, number of previous studies that validated this gene fusion.

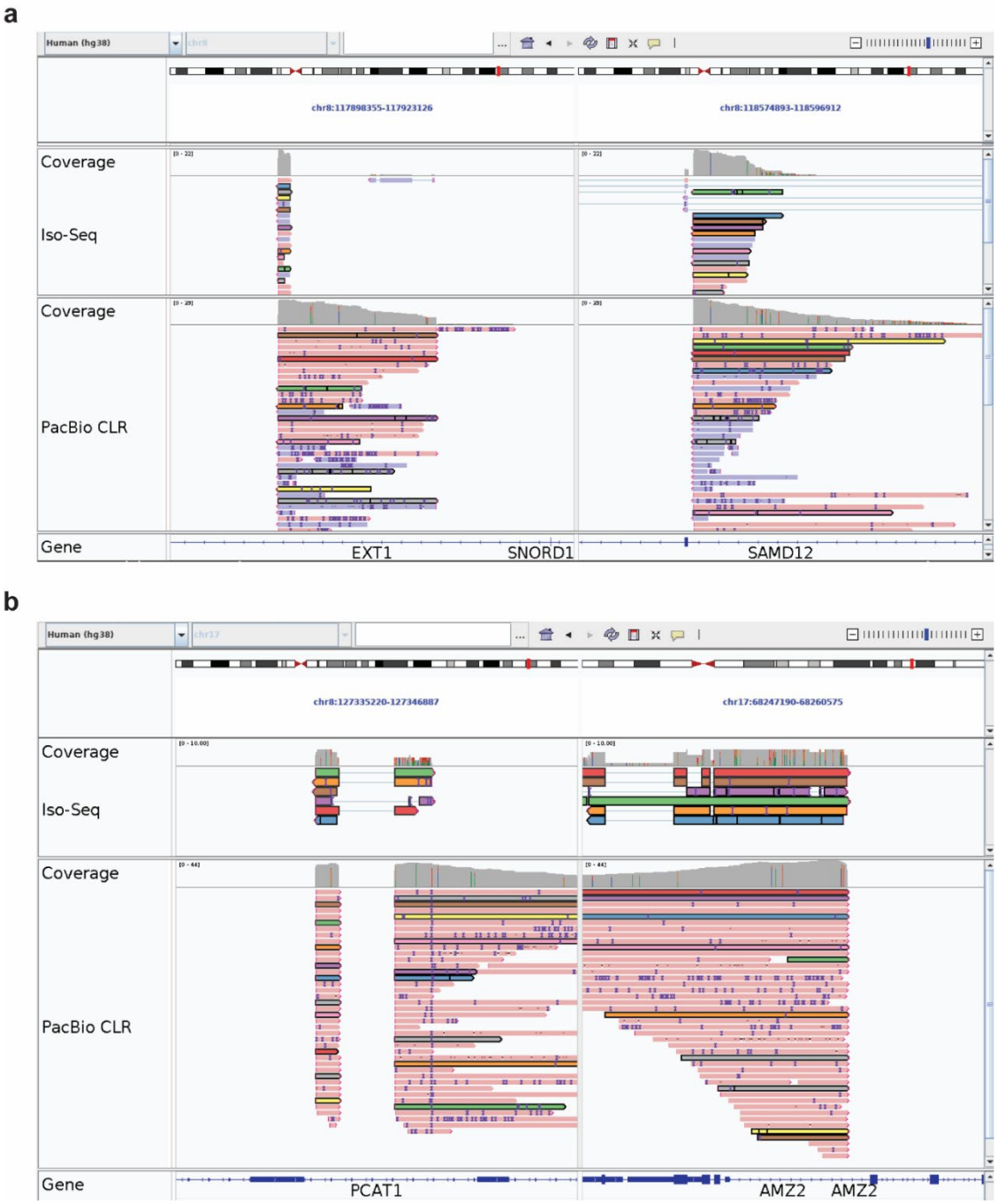

Figure S3. Examples of FusionSeeker-unique gene fusion calls supported by DNA sequencing reads in SKBR-3 cell line. IGV view of PacBio Iso-Seq sequencing (top) and PacBio CLR DNA-sequencing (bottom) read alignments at gene fusion EXT1:SAMD12 (**a**) and PCAT1:AMZ2 (**b**). Alignments of the same read were colored with same color at two breakpoints.

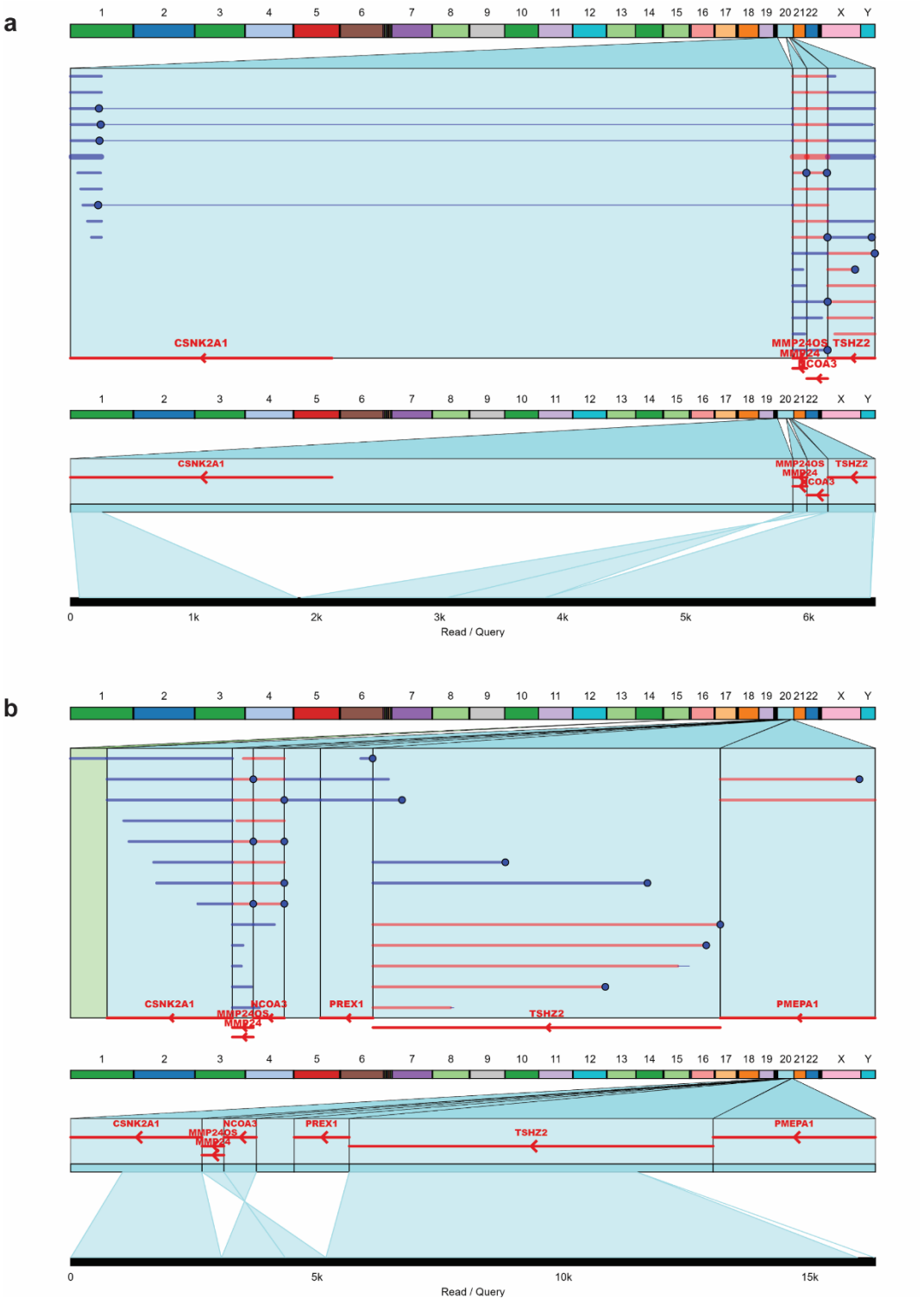

Figure S4. FusionSeeker-unique 4-hop gene fusion CSNK2A1:NCOA3:MMP24OS:TSHZ2 in SKBR-3 cell line. PacBio Iso-Seq sequencing (a) and PacBio CLR DNA-sequencing (b) read alignments covering the gene fusion. Top panel shows all read alignments. Bottom panel shows alignment of a single read from top panel.

Table S5. Detection of previously validated gene fusions in MCF-7 cell line

	PacBio Iso-Seq			Nanopore direct RNA		
Gene fusion	**FusionSeeker**	**JAFFAL**	**LongGF**	**FusionSeeker**	**JAFFAL**	**LongGF**
ARFGEF2:SULF2	✓	✓	✓	✓	✓	✓
BCAS4:BCAS3	✓	✓	✓	✓	✓	✓
RPS6KB1:DIAPH3	✓	✓	✓			
VPS35L:IQCK	✓	✓	✓			
RPS6KB1:VMP1	✓	✓	✓	✓	✓	✓
TBL1XR1:RGS17	✓	✓	✓	✓	✓	✓
TXLNG:SYAP1	✓	✓	✓	✓	✓	✓
MYO6:SENP6	✓	✓	✓	✓	✓	✓
GATAD2B:NUP210L	✓	✓	✓	✓	✓	✓
PAPOLA:AK7	✓	✓	✓	✓		✓
ESR1:CCDC170	✓	✓	✓			
SULF2:PRICKLE2		✓				
POP1:MATN2		✓				
SLC25A24:NBPF6	✓	✓	✓	✓	✓	✓
SYTL2:PICALM	✓	✓	✓	✓	✓	✓
ATP1A1:ZFP64	✓	✓	✓	✓	✓	✓
NAV1:GPR37L1	✓	✓	✓	✓		✓
BCAS3:ATXN7	✓	✓	✓	✓		
MYH9:EIF3D	✓	✓	✓			
PNPLA7:DPH7	✓	✓	✓			
RSBN1:AP4B1-AS1	✓	✓		✓	✓	
BMERB1:ABCC1	✓		✓			
RAD51C:ATXN7		✓	✓			
VAV3:AP4B1-AS1	✓					
BCAS3:AMPD1		✓				
CHEK2:XBP1		✓				
GCN1:MSI1				✓		✓
ATXN7L3:FAM171A2				✓	✓	✓
SMARCA4:CARM1				✓	✓	✓
AHCYL1:RAD51C				✓	✓	✓
DEPDC1B:ELOVL7				✓	✓	✓
MYO9B:FCHO1				✓	✓	✓
BCAS4:ZMYND8					✓	✓
PLCG1:TOP1					✓	✓

Gene fusions not detected by any long-read fusion callers were not shown.

Figure S5. Previously validated gene fusion breakpoints of TXLNG:SYAP1 in MCF-7 cell line. PacBio Iso-Seq sequencing (**a**) and Nanopore direct RNA sequencing (**b**) read alignments covering the two validated gene fusion breakpoints on TXLNG. One validated breakpoint located in first exon (black arrow) was reported by all three gene fusion callers. The other validated breakpoint located in first intron (red arrow) was only reported by FusionSeeker.

Table S6. Previously validated gene fusions missed by FusionSeeker in MCF-7 cell line

		JAFFAL		FusionSeeker		
Dataset	Gene fusion	#Spanning reads	Classification	Reported	#Candidate fusion	Pass filter
Iso-Seq						
	POP1:MATN2	4	LowConfidence	No	0	No
	BCAS3:AMPD1	2	HighConfidence	Yes	2	No
	CHEK2:XBP1	2	HighConfidence	Yes	2	No
	RAD51C:ATXN7	2	HighConfidence	Yes	2	No
	SULF2:PRICKLE2	2	HighConfidence	Yes	3	No
Nanopore						
	BCAS4:ZMYND8	4	LowConfidence	Yes	2	No
	PLCG1:TOP1	3	LowConfidence	Yes	3	No

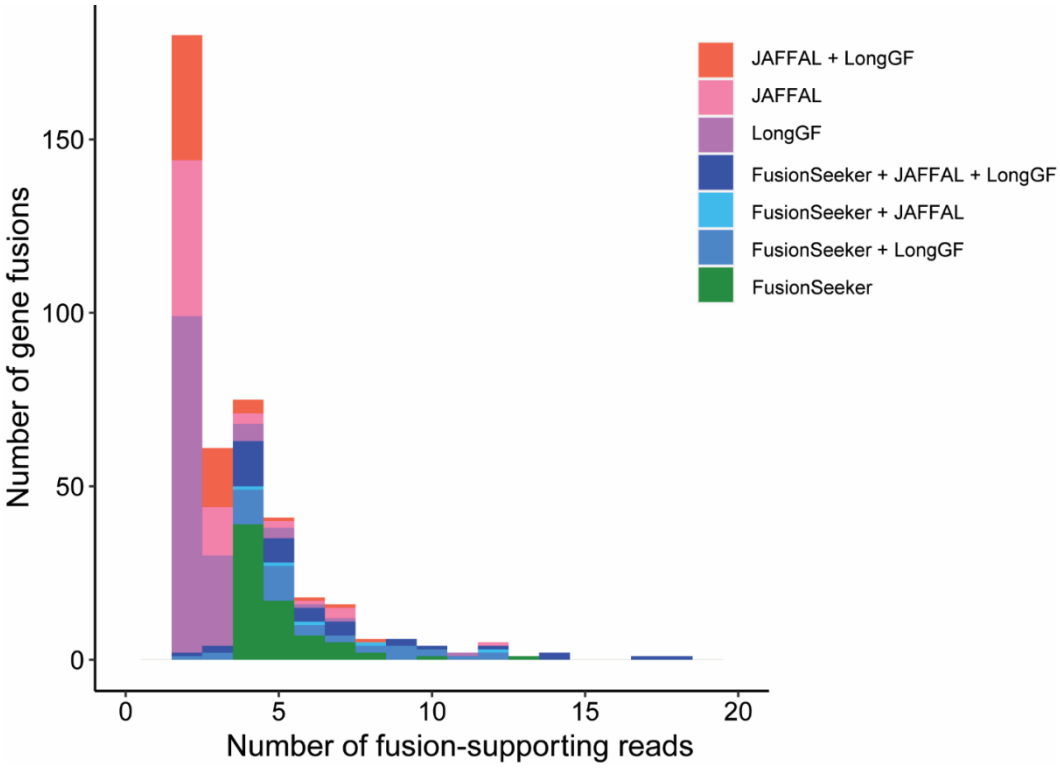

Figure S6. Number of fusion-supporting reads of gene fusion calls. Distribution of number of fusion-supporting reads of gene fusion calls reported by three callers in MCF-7 Iso-Seq dataset. More than 88% of the three groups of gene fusions not reported by FusionSeeker (red, pink, and purple) are supported by ≤ 3 reads.

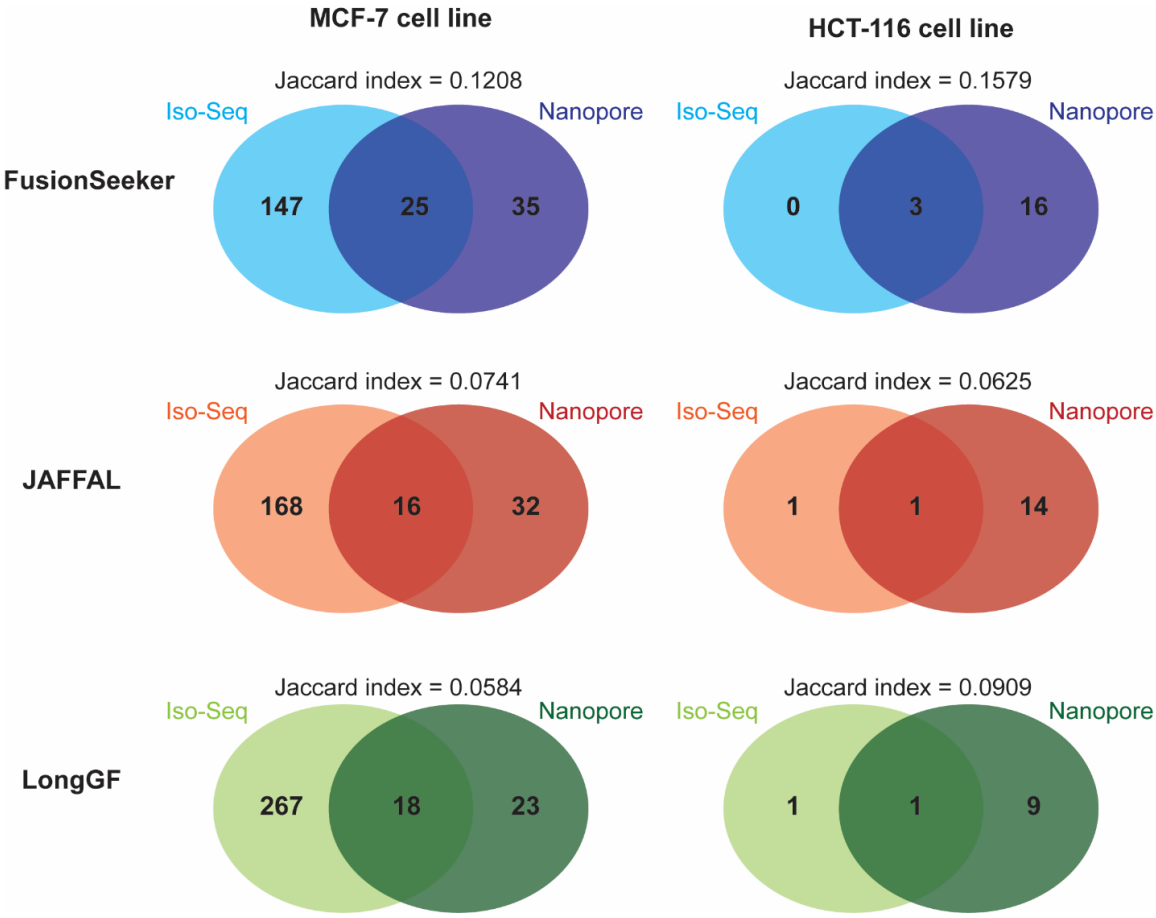

Figure S7. Overlap between gene fusion calls from Iso-Seq and Nanopore datasets. Venn diagram of number of gene fusion calls detected in Iso-Seq and Nanopore dataset of MCF-7 (left) and HCT-116 (right) cell lines. Jaccard index of two lists of gene fusions was labeled on each group.

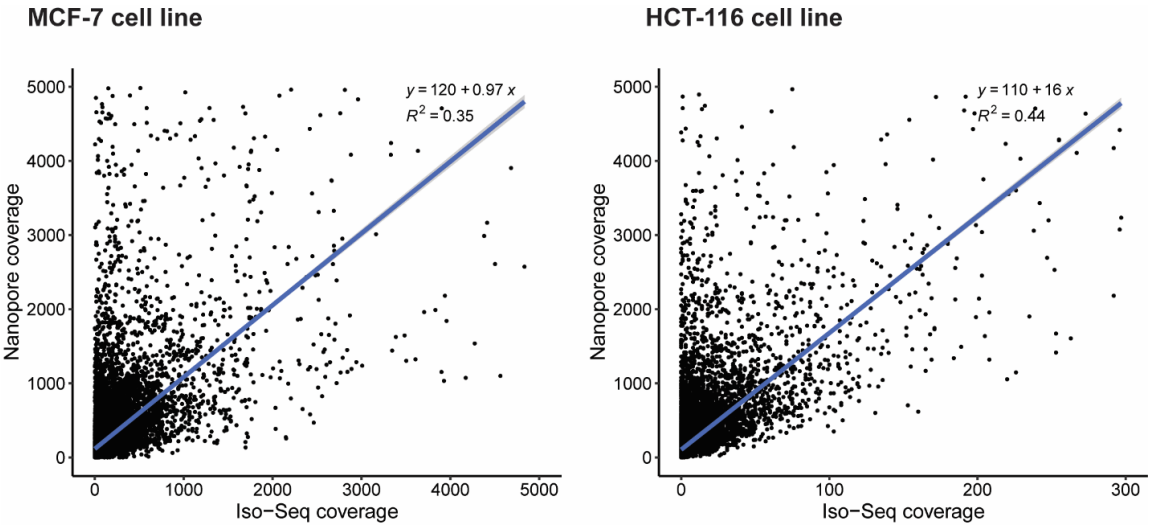

Figure S8. Correlation of coverage between Iso-Seq and Nanopore dataset. Low correlation of sequencing coverages in Iso-Seq and Nanopore datasets in MCF-7 (left) and HCT-116 (right) cell lines. Each dot represents a protein-coding gene. Regression line and R-square were calculated with linear regression model.

Table S8. Number of false-positive gene fusions in non-cancer datasets

Sample	FusionSeeker	JAFFAL	LongGF
HG00268	**18**	19	59
HG01457	**6**	21	74
HG02106	**16**	27	81
HG02666	**11**	22	60
HG03248	**17**	21	81
HG03807	**10**	20	50
HG04217	**15**	32	66
NA18989	**16**	22	65
NA19317	**16**	36	109
NA19331	**14**	34	88
NA19347	**4**	15	41
NA19384	**9**	14	40

Gene fusion calls of JAFFAL only include 'HighConfidence' gene fusions. Fewest false-positive calls were marked in bold.

Table S9. Gene fusion detected in an AML patient sample

Gene 1	Gene 2	Breakpoint 1	Breakpoint 2	Num_supp
RUNX1	RUNX1T1	Chr21: 34849656	Chr8: 92073489	40
RPS25	ARL14EPL	Chr11: 119018340	Chr5: 116052176	38
NBEAL1	RPL12	Chr2: 203190770	Chr9: 127451393	27
EEF1A1	EEF1A1P5	Chr6: 73518555	Chr9: 92073489	20
PSPHP1	KMT2C	Chr7: 55773181	Chr7: 116052176	14
PTMA	DTWD2	Chr2: 231708528	Chr5: 127451393	14
HNRNPH1	ACTB	Chr5: 179614796	Chr7: 133020454	13
SYN3	ACOT13	Chr22: 32532571	Chr6: 152367259	11

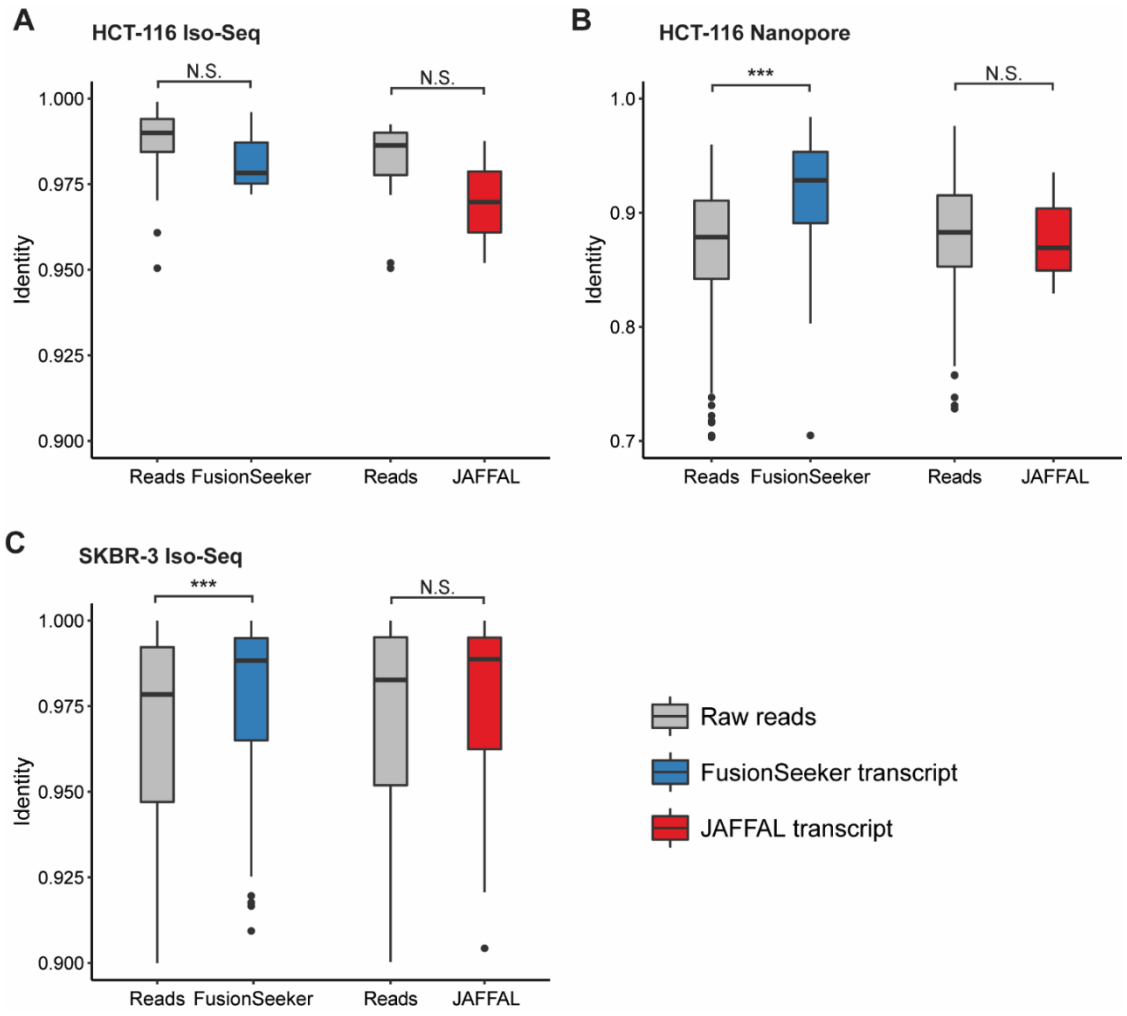

Figure S9. Identity of raw reads and transcript sequences reported by FusionSeeker and JAFFAL. ***, p<0.001. N.S., not significant. p-values were calculated by Mann-Whitney U test.

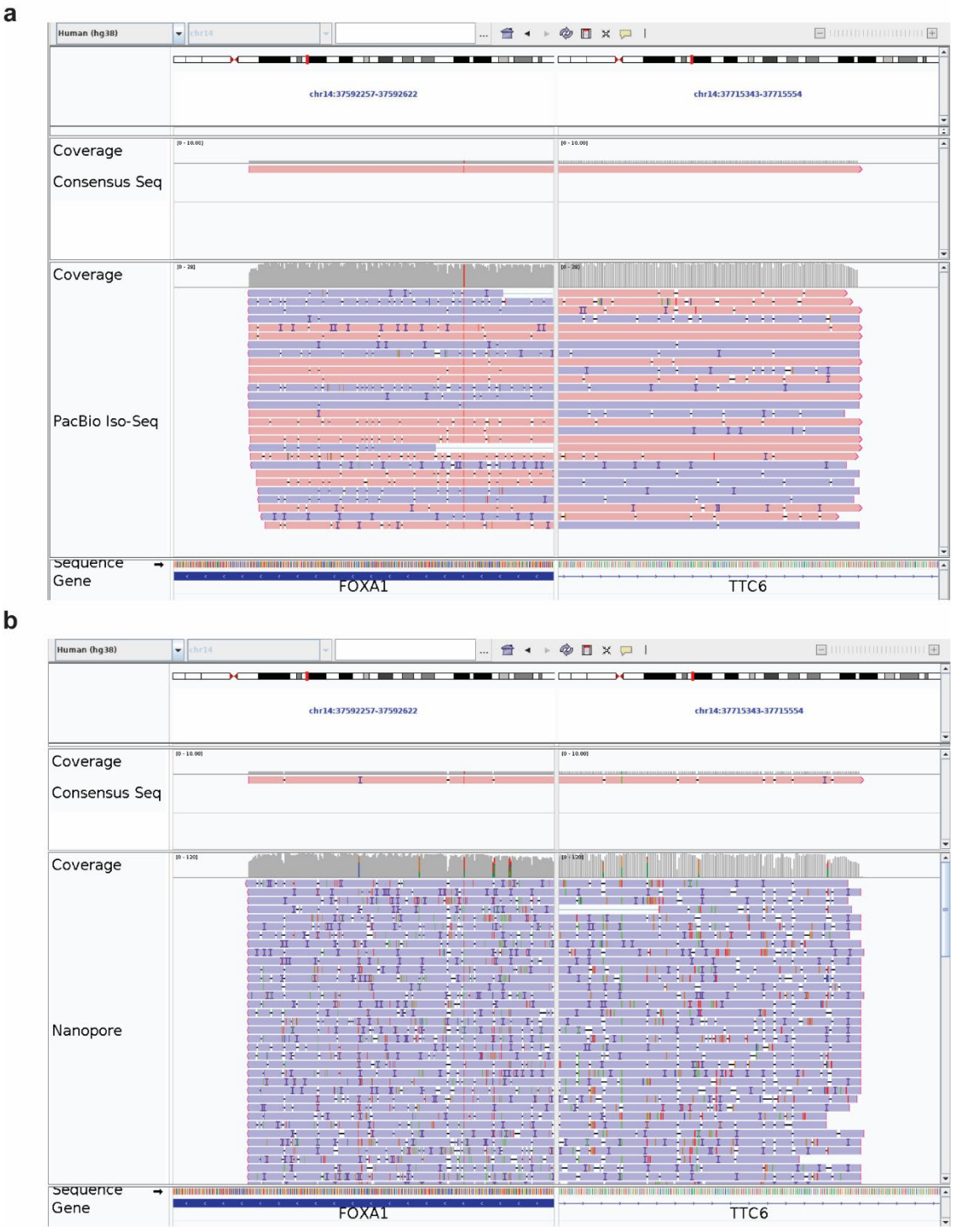

Figure S10. Examples of better base accuracy in FusionSeeker consensus transcript sequence. IGV view of gene fusion FOXA1:TTC6 in Iso-Seq (**a**) and Nanopore (**b**) datasets of MCF-7 cell line. In both datasets, there are fewer mismatches and indels in FusionSeeker consensus transcript sequences (top panel) than in fusion-containing raw reads (bottom panel). The SNP of C>T at chr14:37592537 was retained in consensus transcript sequence in both Iso-Seq and Nanopore datasets.

Supplementary Note 1. Full list of HGSVC members

The members of the Human Genome Structural Variation Consortium (HGSVC) are Haley J. Abel, Hufsah Ashraf, Peter A. Audano, Anna O. Basile, Christine Beck, Marc Jan Bonder, Harrison Brand, Marta Byrska-Bishop, Mark J.P. Chaisson, Yu Chen, Ken Chen, Zechen Chong, Nelson T. Chuang, Wayne E. Clarke, André Corvelo, Scott E. Devine, Peter Ebert, Jana Ebler, Evan E. Eichler, Uday S. Evani, Susan Fairley, Paul Flicek, Sky Gao, Mark B. Gerstein, Maryam Ghareghani, Ira M. Hall, Pille Hallast, William T. Harvey, Patrick Hasenfeld, Alex R. Hastie, Wolfram Höps, PingHsun Hsieh, Sarah Hunt, Jan O. Korbel, Sushant Kumar, Charles Lee, Alexandra P. Lewis, Chong Li, Bin Li, Yang I. Li, Jiadong Lin, Tsung-Yu Lu, Rebecca Serra Mari, Tobias Marschall, Ryan E. Mills, Zepeng Mu, Katherine M. Munson, David Porubsky, Benjamin Raeder, Tobias Rausch, Allison A. Regier, Jingwen Ren, Bernardo Rodriguez-Martin, Ashley D. Sanders, Martin Santamarina, Xinghua Shi, Chen Song, Oliver Stegle, Michael E. Talkowski, Luke J. Tallon, Jose M.C. Tubio, Aaron M. Wenger, Xiaofei Yang, Kai Ye, Feyza Yilmaz, Xuefang Zhao, Weichen Zhou, Qihui Zhu, and Michael C. Zody.

Supplementary Note 2. Supplementary methods

Simulated datasets generation

A transcriptome including all protein-coding transcripts from human GENCODE v39 (RRID:SCR_014966) and additional 150 fused transcripts was simulated. 300 protein-coding genes were randomly selected and paired as the gene fusions, among which 100 fused transcripts were generated with both breakpoints located in exons, and 50 fused transcripts were generated with one breakpoint in an intronic region. All transcripts were randomly assigned into three groups with low, medium, and high expression levels. PacBio Iso-Seq-like reads and Nanopore-like reads were simulated using pbsim (v 1.0.3) and Badread (v0.2.0) with a depth of 10x (low expression), 50x (medium expression), and 200x (high expression), respectively. The simulation process, including the fused transcript generation, was repeated for three times.

Benchmark in simulated datasets

The simulated reads were aligned to the human reference genome GRCh38 without alternative contigs using minimap2 (v2.24) with options "-x splice:hq" for PacBio Iso-Seq simulation and options "-x splice" for Nanopore simulation. FusionSeeker (v1.0.1) was applied on read alignments with options "--min_supp 5". LongGF (v0.1.2) and JAFFAL (v2.2) was applied with the default settings. All unique pairs of fusion genes reported by each tool were compared to the 150 ground-truth gene fusions to count number of true positive (TP) and false positive (FP). Recall and precision were calculated as $recall = \frac{TP}{150}$ and $precision = \frac{TP}{TP+FP}$, and F1 score was

calculated as $F1 = \frac{2*recall*precision}{recall+precision}$. Breakpoint accuracy was evaluated by measuring the distance from reported breakpoint positions from each caller to the breakpoints from ground truth. The identity of the raw reads and FusionSeeker transcript sequences were measured as $identity = 1 - \frac{N_{edit}}{N_{alignment}}$, where $N_{alignment}$ is length of read/transcript sequence aligned to the reference genome, and N_{edit} is number of mismatches (including deletions and insertions) in the alignment.

Gene fusion detection in cancer cell lines

PacBio Iso-Seq (4.16Gbp) and Nanopore direct RNA sequencing data (6.36Gbp) of MCF-7, PacBio Iso-Seq (0.21Gbp) and Nanopore direct RNA sequencing data (6.07Gbp) of HCT-116 cell lines, and PacBio Iso-Seq data (10.91Gbp) of SKBR-3 cell lines were downloaded and aligned to the human reference genome GRCh38 using minimap2 (v2.24) with options "-x splice:hq" for Iso-Seq data and options "-x splice" for Nanopore data. FusionSeeker (v1.0.1) and LongGF (v0.1.2) were applied to the Iso-Seq and Nanopore read alignment files with the default settings. JAFFAL (v2.2) was applied to the PacBio Iso-seq and Nanopore reads with default settings. All unique pairs of gene fusions located on the autosomes and sex chromosomes were used for comparison. Previously validated gene fusion events of MCF-7, SKBR-3, and HCT116 cell lines were collected from previous publication and curated by removing fusions of genes without official gene names and updating gene names to official gene symbols in Ensembl v104 annotation (RRID:SCR_002344).

Validation of FusionSeeker-unique gene fusions in MCF-7 cell line

The MCF-7 cell line was obtained from Dr. Anna Sorace's laboratory at University of Alabama at Birmingham, Birmingham, AL and incubated in DMEM in the presence of 10% FBS, 1% Na-pyruvate, 1% L-glutamine, 1% pen-strep for 48 hours prior to RNA harvesting. RNeasy Plus Mini Kit (Qiagen, 74136) was used for RNA extraction. iScript Reverse transcription supermix (Bio-Rad, 1708841) was used for RT-PCR. RNA to cDNA conversion was performed in Thermal Cycler (Bio-Rad, C1000 Touch) with the preset program. The cDNAs (100ng) were amplified by PCR with a set of primers (10uM). The primer sequences used in the PCR reaction were listed in the **Table S7**. 2% agarose gel was used in the gel electrophoresis analysis. 10uL of PCR products were loaded in each well. BON genomic Notch1 was used to serve as the positive control at 210bp.

Gene fusion detection in non-cancer datasets

PacBio Iso-Seq datasets of 12 HGSVC samples were downloaded and aligned to human reference genome GRCh38 using minimap2 (v2.24) with options "-x splice:hq". FusionSeeker (v1.0.1) and LongGF (v0.1.2) were applied to the Iso-Seq read alignment files with default settings. JAFFAL (v2.2) was applied to the PacBio Iso-seq reads with default settings, and only fusion calls with 'HighConfidence' tag were kept. All unique pairs of gene fusions located on the autosomes and sex chromosomes were used for comparison.

Gene fusion detection in patient sample

Nanopore dataset of an AML patient sample was downloaded and aligned to human reference genome GRCh38 using minimap2 (v2.24) with options "-x splice". FusionSeeker (v1.0.1) was applied to the read alignment file with default settings.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

Work Summary

In this dissertation, I introduced three bioinformatics tools, DeBreak, Inspector, and FusionSeeker, for more comprehensive characterization of SVs and showed some applications of these tools in genetics and clinical research. I present DeBreak in Chapter 2 as an alignment-based SV discovery method for efficient and accurate structural variant detection using long-read sequencing data. I have demonstrated the high SV discovery accuracy and breakpoint accuracy of DeBreak in both simulations and in real benchmark sample HG002. The higher SV discovery accuracy of DeBreak benefited from its density-based clustering methods, in which the clustering window size is adjustable according to SV size and local sequence context. DeBreak achieved single-basepair accuracy for SV breakpoint prediction as it generates highly accurate consensus sequences that contain few sequencing errors to infer precise SV breakpoints. DeBreak also doubled the maximal detectable insertion size by performing local de novo assembly for detecting ultra-large insertions. The tumor mode of DeBreak allows it to discover nearby breakpoints of complex SVs in cancer genome. These novel functions of DeBreak enable us to identify SVs with precise breakpoint locations in noisy long-read sequencing data.

In Chapter 3, I present Inspector as a reference-free evaluation method for *de novo* assembly results. Inspector utilizes raw sequencing data to evaluate assembly quality without help of a reference genome. It can identify structural and small-scale errors in the assembly contigs by distinguishing true assembly errors from inherent genetic variants. Inspector's evaluations on real assemblies of HG002 revealed distinct assembly error patterns for different assemblers and the enrichment of assembly errors in the repetitive regions in human genome for most assemblies. With its error-correction module, Inspector can improve the assembly quality by correcting the identified assembly errors, which will improve the precision of the following assembly-based SV discovery. These functions exceed those achieved by existing assembly evaluators. Inspector is an accurate assembly evaluator and correction tool, which can facilitate future improvement of *de novo* assembly quality.

In Chapter 4, I present FusionSeeker for gene fusion detection in long-read cancer transcriptome sequencing data. Unlike other existing long-read gene fusion callers, FusionSeeker can detect gene fusions located in both exonic and intronic regions and generate error-free transcript sequences for reported gene fusion events. In this chapter, I have demonstrated that FusionSeeker achieved higher accuracy in characterizing gene fusion events than other tools using both simulation data and real cancer cell line data. I have designed PCR experiments and validated 7 novel gene fusions reported by only FusionSeeker in MCF-7 cell line, which may be important for tumorigenesis and progression. By correcting sequencing errors in the raw reads, FusionSeeker reconstructs accurate full-length fusion transcripts, which will facilitate downstream functional and clinical research on these gene fusions.

Together, these three tools have promoted accurate detection of SVs from both alignment-based and assembly-based approaches and accurate detection of gene fusions at transcriptome level. This suite of bioinformatics tools can be applied to facilitate SV-related analysis in the future genetics and clinical research.

Future Research Directions

1. Applying bioinformatics in biomedical research

With the rapid development of sequencing techniques in the past few decades, sequencing-based bioinformatics research has played important roles in recent biomedical research. There is urgent need for more advanced and efficient algorithms for better utilization of sequencing data in basic genomics and clinical research. The three tools introduced in this dissertation, including DeBreak, Inspector, and FusionSeeker, have provided the community a suite of bioinformatics tools for more comprehensive characterization of structural variants, which will deepen our insights into SVs and their functions in population diversity and disease by facilitating future genomics research, clinical studies, and bioinformatics algorithm development.

In genomic research, precise discovery of SV is the foundation of accurate population-level analysis on SVs. Studies on mechanism of SV formation usually focus on flanking sequences near SV breakpoints, which relies on precise prediction of SV breakpoints[42, 43]. Recent work on telomere-to-telomere assembly of human genomes and pangenome reference project has targeted at providing more complete reference genome and comprehensive genetic profiling through *de novo* assembly[44, 45]. With these efforts, de novo assembly has become an important approach in driving future

discovery in human genomic health and disease, which could be further facilitated with assembly evaluation and improvement methods such as Inspector.

SV discovery, especially accurate prediction of SV breakpoints, is essential for disease genetics and cancer research, including studying causal mutations, disease diagnosis and progression markers, and potential treatment targets. In systemic lupus erythematosus, FCGR gene family-associated SVs are located within a complex segmental duplication region and are thus difficult to detect using routine SV discovery approaches[10]. Our efforts in both alignment- and assembly-based SV discovery of DeBreak and Inspector have enabled detection of SVs in most of PacBio HiFi datasets of SLE patients using local and whole-genome *de novo* assembly.

In bioinformatics research, our tools provide a guidance for future development of novel algorithms and pipelines. The advanced algorithms implanted in our tools, including density-based clustering, local *de novo* assembly, and parallel computing, can be applied to solve other problems. In particular, Inspector evaluation of assembly quality provides accurate positions of misassemblies, which can serve as unbiased benchmark method for further improving assembly algorithms.

2. Application on large-scale datasets

As the NGS becomes more common in genetic research, there are several publicly available large-scale NGS datasets for healthy individuals and patient samples, including HGSVC[26], PCAWG [46], and TOPMed [47]. SV analysis on such datasets provides insights on genomic evolution and disease mechanisms at population level. Large-scale

SV analysis of healthy individuals using long-read data would enable more accurate characterization of normal human genetic variations, as the foundational SV discovery is more comprehensive using long-read data. However, due to the high cost of long-read sequencing, current public long-read datasets usually contain fewer numbers of samples. For example, GIAB sequenced seven samples, including two trio families, as benchmark samples[48]. HGSVC sequenced a total of 32 samples from distinct populations using long-read platforms[26], which are much fewer compared to 3,110 samples with NGS data. Such small sample size is not sufficient to make significant conclusions in population genetics. Once the sequencing cost is reduced for TGS and more samples are sequenced, DeBreak can be applied to identify SVs for further large-scale studies.

For cancer transcriptome sequencing data, application of FusionSeeker is also limited by data availability. Currently, publicly available transcriptome sequencing data is restricted to cancer cell lines, which are more stable and less heterogenous than primary tumor biopsies[49-51]. When more patient samples are sequenced with long-read transcriptome sequencing platforms, FusionSeeker can be applied to identify gene fusions for each tumor sample. Disease-related gene fusions can then be inferred when shared by a subgroup of patients with the same cancer type.

3. Application on Non-human Species

Although the applicable organisms of DeBreak and FusionSeeker are not limited to human, most of the benchmarks in this dissertation were done based on human genomes, owing to the lack of available long-read benchmark datasets for non-human

species. Both DeBreak and FusionSeeker can be applied to other diploid or haploid non-human species with available reference genome and gene annotations. The performance of these tools may be affected when applied on other species, owing to the differences in the genomic features. Parameters for non-human species could be optimized when more sophisticated benchmarking datasets are available for non-human species.

Moreover, due to the limited availability of ground-truth SV sets, DeBreak was benchmarked for insertion and deletion discovery in HG002 and HGSVC samples, but not for duplication, inversion, or translocation. Further validation of SV discovery accuracy on these SV types would be desirable and will help improve DeBreak's performance if comprehensive high-confidence truth SV sets become more readily available.

In Chapter 3, I have shown benchmarking and analysis of human and Anna's hummingbird genomes for Inspector. When detecting assembly errors using binominal test, Inspector has an assumption of diploid genome for the input assembly. The assumption can work on any species with monoploid or diploid genomes but not for polyploid genomes. Applying current version of Inspector on polyploid genomes may lead to inaccurate identification of assembly errors and therefore inaccurate evaluation of assembly quality. I plan to expand the application of Inspector by adding statistical models for polyploid genomes in future versions.

LIST OF REFERENCES

1. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al: Global variation in copy number in the human genome. *Nature* 2006, 444:444-454.

2. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Fritz MH-Y, et al: An integrated map of structural variation in 2,504 human genomes. *Nature* 2015, 526:75-81.

3. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al: Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat Commun* 2019, 10:1784.

4. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L, et al: Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res* 2017, 27:677-685.

5. Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Paabo S, et al: A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* 2005, 437:88-93.

6. Chimpanzee Sequencing and Analysis Consortium: Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 2005, 437:69-87.

7. Carvalho CM, Lupski JR: Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* 2016, 17:224-238.

8. Beck CR, Carvalho CM, Banser L, Gambin T, Stubbolo D, Yuan B, Sperle K, McCahan SM, Henneke M, Seeman P, et al: Complex genomic rearrangements at the PLP1 locus include triplication and quadruplication. *PLoS Genet* 2015, 11:e1005050.

9. Maynard TM, Haskell GT, Lieberman JA, LaMantia AS: 22q11 DS: genomic mechanisms and gene function in DiGeorge/velocardiofacial syndrome. *Int J Dev Neurosci* 2002, 20:407-419.

10. Morris DL, Roberts AL, Witherden AS, Tarzi R, Barros P, Whittaker JC, Cook TH, Aitman TJ, Vyse TJ: Evidence for both copy number and allelic (NA1/NA2) risk at the FCGR3B locus in systemic lupus erythematosus. *Eur J Hum Genet* 2010, 18:1027-1031.

11. Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, Martincorena I, Alexandrov LB, Martin S, Wedge DC, et al: Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 2016, 534:47-54.

12. Patch AM, Christie EL, Etemadmoghadam D, Garsed DW, George J, Fereday S, Nones K, Cowin P, Alsop K, Bailey PJ, et al: Whole-genome characterization of chemoresistant ovarian cancer. *Nature* 2015, 521:489-494.

13. Waddell N, Pajic M, Patch AM, Chang DK, Kassahn KS, Bailey P, Johns AL, Miller D, Nones K, Quek K, et al: Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* 2015, 518:495-501.

14. Stankiewicz P, Lupski JR: Structural variation in the human genome and its role in disease. *Annu Rev Med* 2010, 61:437-455.

15. Weischenfeldt J, Symmons O, Spitz F, Korbel JO: Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet* 2013, 14:125-138.

16. Sakofsky CJ, Roberts SA, Malc E, Mieczkowski PA, Resnick MA, Gordenin DA, Malkova A: Break-induced replication is a source of mutation clusters underlying kataegis. *Cell Rep* 2014, 7:1640-1648.

17. Costantino L, Sotiriou SK, Rantala JK, Magin S, Mladenov E, Helleday T, Haber JE, Iliakis G, Kallioniemi OP, Halazonetis TD: Break-induced replication repair of damaged forks induces genomic duplications in human cells. *Science* 2014, 343:88-91.

18. Metzker ML: Sequencing technologies - the next generation. *Nat Rev Genet* 2010, 11:31-46.

19. Rhoads A, Au KF: PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* 2015, 13:278-289.

20. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, et al: The potential and challenges of nanopore sequencing. *Nat Biotechnol* 2008, 26:1146-1153.

21. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al: Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 2019, 37:1155-1162.

22. English AC, Salerno WJ, Reid JG: PBHoney: identifying genomic variants via long-read discordance and interrupted mapping. *BMC Bioinformatics* 2014, 15:180.

23. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC: Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* 2018, 15:461-468.

24. Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, Wang Y: Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol* 2020, 21:189.

25. Chen Y, Wang AY, Barkley C, Zhao X, Gao M, Edmonds M, Chong Z: DeBreak: Deciphering the exact breakpoints of structural variations using long sequencing reads. *Preprint from Research Square* 2022.

26. Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, et al: Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* 2021, 372.

27. Li H, Bloom JM, Farjoun Y, Fleharty M, Gauthier L, Neale B, MacArthur D: A synthetic-diploid benchmark for accurate variant-calling evaluation. *Nat Methods* 2018, 15:595-597.

28. Heller D, Vingron M: SVIM-asm: Structural variant detection from haploid and diploid genome assemblies. *Bioinformatics* 2020.

29. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM: Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* 2017, 27:722-736.

30. Kolmogorov M, Yuan J, Lin Y, Pevzner PA: Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 2019, 37:540-546.

31. Cheng H, Concepcion GT, Feng X, Zhang H, Li H: Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* 2021, 18:170-175.

32. Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, Simpson JT, Stebbings LA, Leroy C, Edkins S, Mudie LJ, et al: Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* 2009, 462:1005-1010.

33. Salzman J, Marinelli RJ, Wang PL, Green AE, Nielsen JS, Nelson BH, Drescher CW, Brown PO: ESRRA-C11orf20 is a recurrent gene fusion in serous ovarian carcinoma. *PLoS Biol* 2011, 9:e1001156.

34. Kumar-Sinha C, Tomlins SA, Chinnaiyan AM: Recurrent gene fusions in prostate cancer. *Nat Rev Cancer* 2008, 8:497-511.

35. Tomlins SA, Laxman B, Varambally S, Cao X, Yu J, Helgeson BE, Cao Q, Prensner JR, Rubin MA, Shah RB, et al: Role of the TMPRSS2-ERG gene fusion in prostate cancer. *Neoplasia* 2008, 10:177-188.

36. Latysheva NS, Oates ME, Maddox L, Flock T, Gough J, Buljan M, Weatheritt RJ, Babu MM: Molecular Principles of Gene Fusion Mediated Rewiring of Protein Interaction Networks in Cancer. *Mol Cell* 2016, 63:579-592.

37. Westbrook CA, Hooberman AL, Spino C, Dodge RK, Larson RA, Davey F, Wurster-Hill DH, Sobol RE, Schiffer C, Bloomfield CD: Clinical significance of the BCR-ABL fusion gene in adult acute lymphoblastic leukemia: a Cancer and Leukemia Group B Study (8762). *Blood* 1992, 80:2983-2990.

38. Wilda M, Fuchs U, Wossmann W, Borkhardt A: Killing of leukemic cells with a BCR/ABL fusion gene by RNA interference (RNAi). *Oncogene* 2002, 21:5716-5724.

39. Wang Q, Xia J, Jia P, Pao W, Zhao Z: Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives. *Brief Bioinform* 2013, 14:506-519.

40. Davidson NM, Chen Y, Sadras T, Ryland GL, Blombery P, Ekert PG, Goke J, Oshlack A: JAFFAL: detecting fusion genes with long-read transcriptome sequencing. *Genome Biol* 2022, 23:10.

41. Liu Q, Hu Y, Stucky A, Fang L, Zhong JF, Wang K: LongGF: computational algorithm and software tool for fast and accurate detection of gene fusions by long-read transcriptome sequencing. *BMC Genomics* 2020, 21:793.

42. Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR: The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat Genet* 2009, 41:849-853.

43. Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, Kallicki J, Kaul R, Wilson RK, Eichler EE: A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* 2010, 143:837-847.

44. Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, Taylor DJ, Shafin K, Shumate A, Xiao C, et al: A complete reference genome improves analysis of human genetic variation. *Science* 2022, 376:eabl3533.

45. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al: The complete sequence of a human genome. *Science* 2022, 376:44-53.

46. Consortium ITP-CAoWG: Pan-cancer analysis of whole genomes. *Nature* 2020, 578:82-93.

47. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM, et al: Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 2021, 590:290-299.

48. Zook JM, Hansen NF, Olson ND, Chapman L, Mullikin JC, Xiao C, Sherry S, Koren S, Phillippy AM, Boutros PC, et al: A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol* 2020, 38:1347-1355.

49. Nattestad M, Goodwin S, Ng K, Baslan T, Sedlazeck FJ, Rescheneder P, Garvin T, Fang H, Gurtowski J, Hutton E, et al: Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res* 2018, 28:1126-1135.

50. Iowa Uo: Full-length transcripts of the MCF-7 breast cancer cell line by PacBio SMRT sequencing. 2015.

51. Centre BCR: Transcriptome dynamics of CLK dependent exon recognition and conjoined gene formation revealed with a novel small molecule inhibitor. 2017.