Precise characterization of somatic structural variations and mobile element insertions from paired long-read sequencing data with nanomonsv

Yuichi Shiraishi[1,*], Junji Koya[2], Kenichi Chiba[1], Yuki Saito[2,3], Ai Okada[1], Keisuke Kataoka[2]

[1] Division of Genome Analysis Platform Development, National Cancer Center Research Institute, Tokyo, Japan
[2] Division of Molecular Oncology, National Cancer Center Research Institute, Tokyo, Japan
[3] Department of Gastroenterology, Keio University School of Medicine, Tokyo, Japan

[*] Corresponding author:
Yuichi Shiraishi, Ph.D.
5-1-1, Tsukiji, Chuo-ku, Tokyo, 104-0045, Japan
E-mail: yuishira@ncc.go.jp

Abstract

We introduce our novel software, nanomonsv, for detecting somatic structural variations (SVs) using tumor and matched control long-read sequencing data with a single-base resolution. Using paired long-read sequencing data from three cancer cell-lines and their matched lymphoblastoid lines, we demonstrate that our approach can identify not only somatic SVs that can be captured with short-read technologies but also novel ones especially those whose breakpoints are located in repeat regions. In addition, we have developed a workflow for classifying mobile element insertions while elucidating their in-depth properties such as 5′ truncations, internal inversion as well as source sites in the case of LINE1 transductions. Finally, we identify complex SVs probably caused by replication mechanisms or telomere crisis by examining the co-occurrence of multiple somatic SVs in common supporting reads. In summary, our approaches applied to cancer long-read sequencing data can reveal various features of somatic SVs and will lead to further understanding of mutational processes and functional consequences of somatic SVs.

Introduction

Structural variations (SVs) have been known to play essential roles in cancer development. The advances in high-throughput sequencing technologies have enabled us to perform genome-wide somatic SV screening, and a number of cancer-driving ones have been identified[1–3]. On

the other hand, millions of repetitive elements are widely distributed across the human genome, where current standard short-reads are difficult to be unambiguously aligned. The categories of repeat sequences include satellite DNA typically found in centromeres and heterochromatin as well as transposable elements such as LINE1 (long interspersed nuclear elements 1), Alu, and SINE/VNTR/Alu (SVA) sequences. According to several computational predictions, these repeat sequences may comprise from half to two-thirds of the human genome[4,5]. Since the majority of the current sequencing data is collected using short-read Illumina sequencing technologies, several classes of SVs, especially those whose breakpoints are located in these repeat regions, have been hard to detect[6,7]. As such, although a large number of whole genome studies regarding somatic SVs have been conducted, it is plausible to assume that we still have only a limited landscape of SVs in human cancer.

Recently, long-read sequencing technologies, which surely solve some of the problems related to short-read technologies such as the ambiguous alignments on repeat regions, attract lots of attention with the hope of improving the performance of SV detection[8,9]. Several studies have demonstrated the effectiveness of long-read data developing software packages for detecting SVs from long-read sequencing data[10–14]. Meanwhile, for the identification of somatic ones, it has been typical to utilize matched control data usually collected from noncancerous parts of the same patients. However, there are still few approaches that are specifically designed for somatic SV detection using tumor and matched control long-read sequencing data. One naive approach is to perform existing algorithms for both tumor and control sequencing data individually, and take the subtraction of the set of SVs found in the tumor from that in the matched control. However, this approach can generate many false positives such as germline SVs that pass the threshold in the tumor and narrowly miss it in the matched control (e.g., because of low sequencing depths). In fact, in the context of somatic variant detection, it has been widely accepted that jointly utilizing tumor and matched control sequencing data is highly important[15,16]. In any case, due to the lack of appropriate software, the effectiveness of long-read technologies for detecting somatic SVs has not been thoroughly investigated.

Another important point that long-read technologies could address is to characterize the detailed structure of somatic long insertions, especially mobile element insertions (MEIs) represented by LINE1 retrotransposition[17,18]. Among the millions of LINE1 elements across the human genome, about one hundred are thought to be still active. They can somatically produce their RNA intermediates, which are inserted into distant genomic sites with some modifications (5′ truncation, internal inversion, and 3′ transduction). Besides, LINE1 can also help the somatic displacement of other categorical elements such as Alu, SVA, and processed pseudogenes. Short-read sequencing data can, in principle, detect the existence of these insertion events and a few hundred nucleotides from the edge of inserted sequences, and several successful studies characterized their roles in cancer[19,20]. Nevertheless, the algorithms to identify insertions from short-read sequencing tend to be complicated, and a relatively straightforward method using a long-read platform may accomplish higher sensitivity. Even more importantly, whereas short-read technology can only identify the edge of inserted sequences, long-read technologies may be able to extract the full-length inserted nucleotides, which can help elucidate the various properties of MEI events.

The other issue is the identification of complex SVs. SVs are typically characterized by “junctions” of two breakpoints in the genome: genomic coordinates and directions of the first and

second breakpoints often with inserted sequences between them. However, there are many documented mutational processes that cause a cluster of SVs (in the sense of "breakpoint junctions"). One simple case is the reciprocal inversion, in which a segment of a chromosomal region is invertedly inserted in the same position. Other examples in cancer genomes include chromoplexy[21,22], chained rearrangements caused by simultaneous double-strand breaks and cross-connections among them, and chromothripsis[23], in which a single catastrophic event fragments one or multiple chromosomal regions into small genomic segments and rejoins parts of them in an uncoordinated fashion. Some replication-based mechanisms, such as microhomology-mediated break-induced replication[24,25], which is mainly referred to in the germline context, are known to induce multiple breakpoint junctions. Characterizing these clusters of SVs and reconstructing the rearranged chromosomes are of great importance to understand the influence on genomic functions such as transcription and investigate the prevailing mutational processes. This goal can be partially accomplished by short-read sequencing data gathering significantly close SVs and narrowing down possible coupling of genomic segments consistent to breakpoint directions and copy number changes[26,27]. However, although these studies carefully argued for validity with several circumstantial evidence, complex SVs reconstructed from short-reads are only predictions in the end. Not surprisingly, concurrency of multiple breakpoint junctions cannot be proved because they are mostly too far apart for short-reads to phase, and reconstructed derivative chromosomes are very often ambiguous (cannot be limited to one pattern).

In this paper, we introduce our approach, nanomonsv (https://github.com/friend1ws/nanomonsv), that can identify somatic SVs with single-nucleotide resolution jointly using both tumor and control long-read sequencing data with Oxford Nanopore platform. With this software and newly collected paired long-read sequencing data from three cell-lines, we evaluate the effectiveness of long-read data for somatic SV detection. First, by the comprehensive comparison with short-read sequencing technologies, we demonstrate that the novel pipeline applied to long-read sequencing data can capture not only most of the SVs that can be identified using short-read sequencing platform but also additional ones specifically found by long-read sequencing data. Second, we provide a workflow for classifying full-length inserted sequences obtained by the nanomonsv approach into various types of MEIs, and show that long-read platform can clarify various properties of MEIs such as 5′ truncations, internal inversion, target site duplications and source sites in the case of LINE1 transductions. Finally, we show that long-read sequencing data can be efficiently used to recapitulate the complex somatic SV structures just using the firm evidence of the co-occurrence of multiple SVs.

Results

Method summary

Prior to performing nanomonsv, we assume that both tumor and control sequence files are already aligned to the reference genomes with minimap2[28]. After the alignment, the procedures of nanomonsv are divided into the following four steps (see Figure 1). A more detailed

description is provided in the Method section. For deletions and insertions, we focus on those whose sizes are equal or larger than 100bp.

(1) Parsing: the reads putatively supporting SVs are extracted from both tumor and matched control BAM files using CIGAR string and supplementary alignment information.
(2) Clustering: the reads from a tumor sample that presumably span the same SVs are clustered, and the possible ranges of breakpoints are inferred for each possible SV. We subsequently remove putative SVs with less than three supporting reads or median of <= 40 mapping qualities. Also, if there exist the apparent supporting reads in the matched control sample, these are also removed.
(3) Refinement: Extract the portions of the supporting reads around the breakpoints, and perform multiple sequence alignment using MAFFT[29] to generate the consensus sequence for each candidate SV. Then, aligning the consensus sequence to those around the possible breakpoint regions in the reference genome using a modified Smith-Waterman algorithm (which allows one-time jump from one genomic region to the other, see Supplementary Figure 1), we identify the exact breakpoint positions and the non-template insertions inside them.
(4) Validation: From the breakpoint determined in the previous step, we generate the "putative SV segment sequence." Then we collect the reads around the breakpoint of putative SVs and check whether the putative SV segment sequence exists (then the read is set as a "variant supporting read") or not (then the read is classified to a "reference read") in each read of the tumor and matched control. Finally, candidate SVs with (A) >=3 variants supporting reads in the tumor, and (B) no variant supporting reads in the matched control sample, are kept as the final SVs.

The first two steps are rather straightforward, and the last two steps are the key features of this pipeline. The "refinement" step plays an essential role in determining the single-nucleotide resolution breakpoint as well as error-corrected inserted sequences by multiple alignments. Particularly, polishing inserted sequences at this stage will be highly helpful for classifying long insertion events. Furthermore, identification of the breakpoint will facilitate the next "validation" step by creating an unambiguous putative SV segment sequence. The last "validation" step is vital for thoroughly confirming that the candidate SV is truly tumor-specific. More specifically, by performing the alignment of the putative SV segment sequence to each read close to putative breakpoints, precise detection of variant supporting reads, especially those that are partially covering the breakpoints and not counted in the parse step, will become possible.

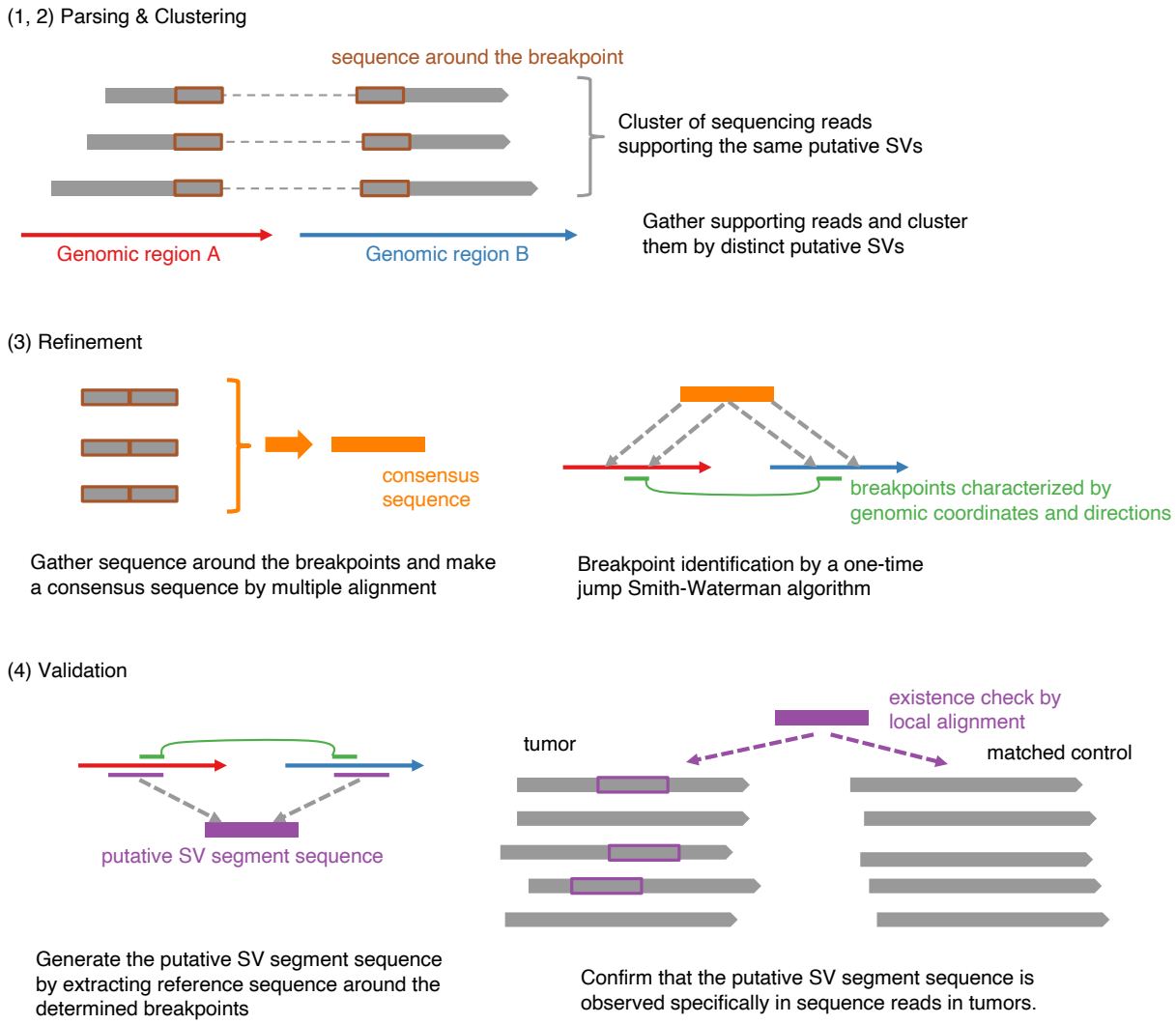

Figure1: Workflow of somatic SV detection in nanomonsv. Nanomonsv consists of four steps. The first step (parsing) is parsing all the sequencing reads potentially supporting the SVs. The second step (clustering) is clustering these supporting reads per candidate SVs. Then, in the third step (refinement), the portions of supporting reads around the breakpoints for each nominated SV are aggregated into a consensus sequence, which is used for the identification of the exact coordinate of SV breakpoints by a one-time jump Smith-Waterman algorithm. Finally, in the fourth step (validation), the putative SV segment sequence for each candidate SV is extracted and checked whether this element is only observed in tumor samples or not.

Description of sequencing data and alignment to reference genome

For evaluation, we chose three cancer cell-lines (COLO829, H2009, and HCC1954) and three lymphoblastoid cell-lines (COLO829BL, BL2009, and HCC1954BL) from the same patients as their matched controls. COLO829 (from a metastatic cutaneous melanoma patient) and COLO829BL (from a lymphoblastoid line of the same patient) have been often used as a benchmark in many previous studies[30–32]. Although this cell-line has been known to have

hypermutated nature for somatic single nucleotide variants as well as double nucleotides ones, the number of somatic SVs seems to be relatively low. H2009 (from metastatic lung adenocarcinoma) has many long insertions mainly by high LINE1 activity and has been used in studies investigating the mechanism of MEIs[19,20]. HCC1954 (from ductal breast carcinoma) and HCC1954BL also have been frequently used as a benchmark (TCGA mutation calling benchmark 4, https://gdc.cancer.gov/) and seems to have a relatively large number of somatic SVs. Although these cell-lines have been used in many studies, there have been few efforts to characterize exhaustive and accurate lists of somatic SVs from these cell-lines.

Long read and whole-genome sequencing were conducted using GridION and PromethION. The total outputs were 59.13 to 156.30 Gbps, and the median sequence lengths ranged from 3,689 to 7,997 bp (see Table 1, Supplementary Figure 2). These data were aligned by minimap2 to the human reference genome, and 93.65 to 94.48 % reads were mapped with high quality (>=40 mapping quality).

To compare the result of SVs called from long-read sequencing data with a short-read platform, we also performed sequencing of these six cell-lines using Illumina Novaseq 6000 platform. The total amounts of yield after polymerase chain reaction (PCR) duplication removal were 205.76 Gbps to 484.26 Gbps. Then, for detecting somatic SVs of general classes (deletions, duplications, insertions, inversions, and translocations), we performed four algorithms, manta[33], SvABA[34], GRIDSS[35] and our in-house pipeline GenomonSV (https://github.com/Genomon-Project/GenomonSV) used in the previous studies[3,36]. Also, we incorporated the result of TraFiC-mem[19], which is specially designed for detecting somatic MEIs where necessary.

Cell-line	Long-read yield (Gbp)	Long-read total read count	Long-read median read length	Long-read max read length	Short-read yield (Gbp)
COLO829	67.17	5,176,983	7,997	185,650	250.28
COLO829BL	59.13	6,253,574	5,691	124,349	393.24
H2009	114.91	10,319,362	6,342	238,152	484.26
BL2009	156.30	15,684,323	5,195	240,066	319.82
HCC1954	145.58	11,285,481	7,523	250,253	291.86
HCC1954BL	126.34	17,608,439	3,689	220,506	205.76

Table 1: Summary statistics of long-read (Nanopore) and short-read (Illumina) data from six cell-lines.

Comparison with short-read sequencing data

Applying nanomonsv to three cell-line long-read sequencing data, we identified 52, 749, and 727 structural variations for COLO829, H2009, and HCC1954, respectively (Figure 2a, Supplementary Data 1). As expected, H2009 had many long insertions, most of which were LINE1 retrotranspositions (shown in the next section). On the other hand, HCC1954 had all the types of SVs evenly except long insertions. For the evaluation of precision, we performed the PCR on 128 randomly selected somatic SVs, and at least 109 (85.2%) showed tumor sample-specific bands with predicted product sizes (Supplementary Figure 3). The evaluation of recall is

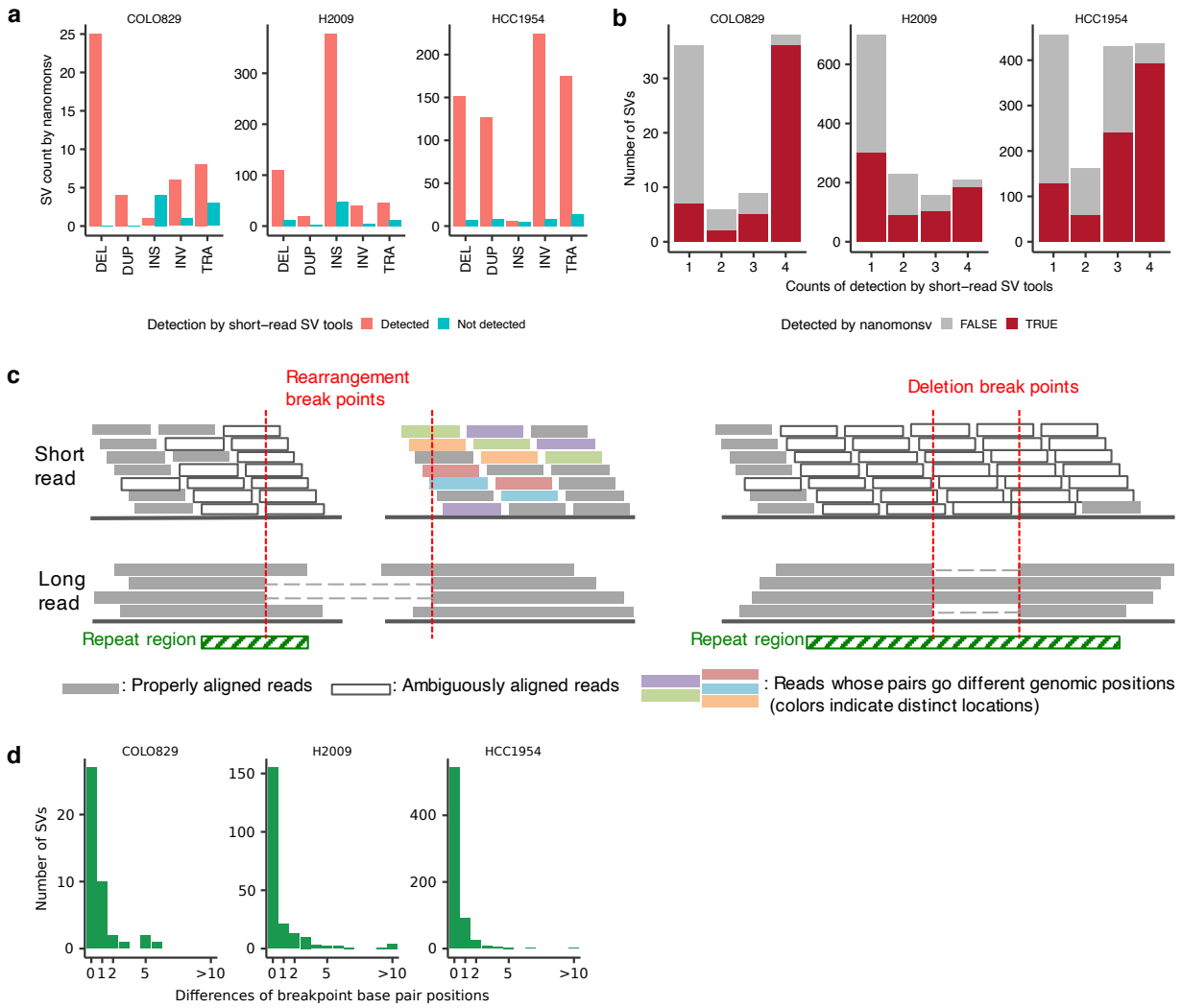

Figure 2: Overview of somatic SVs identified by nanomonsv and their comparison with short-read platforms. (a) The number of somatic SVs detected by nanomonsv grouped by the type of SVs and whether they are identified by the short-read analysis. DEL, DUP, INS, INV, and TRA stand for deletion, duplication, insertion, inversion, and translocation, respectively. (b) The number of somatic SVs called by the short-read platform stratified by how often these SVs are called by four software programs (manta, SvABA, GRIDDS, and GenomonSV), and the fraction of overlap with the result from the long-read sequence for each stratification. (c) Schematics depicting alignments of short and long reads around the breakpoints of SVs that can only be identified by long-read technologies. In the first example, one of the breakpoints is located at the repeat or low complexity regions, which will be blurred by many short-reads coming from other regions with mostly the same sequences (left). In the second example, the relatively short SV are completely covered by repeat regions and also buried by short-reads in fact originating from other genome wide regions with the same sequence classes (right). (d) Histogram of the number of SVs according to the deviations of breakpoint positions from a short-read platform.

somewhat difficult since some of the calls by short-read based methods probably include a certain amount of false positives, and the list of all true SVs including those in ambiguous regions and with low variant allele frequencies is far from comprehensive. Hence, we resorted to comparing with SVs commonly detected by all the four algorithms (manta, SvABA, GRIDSS, and GenomonSV) in the short-read platform, that are considered to be “true” somatic SVs with a high degree of accuracy. Among the total of 685 SVs by all the four algorithms, nanomonsv

applied to Nanopore sequencing data identified 611 SVs (89.2%) (Figure 2b), suggesting the high sensitivity of nanomonsv on long-read sequencing data even for relatively shallow coverage compared to short-read sequencing data. Certainly, even the list of SVs called by less than four algorithms may include several true positives. However, SVs detected by less number of short-read based tools had smaller overlap with long-read results, implying nanomonsv tend to capture many of reliable SVs.

For COLO829, H2009 and HCC1954, 8, 81, and 42, respectively (5.8% to 15.4%), were newly detected by long-read sequencing data (not identified by any of the four algorithms nor TraFiC-mem applied to high coverage Illumina short-read sequencing data). These long read-specific SVs could be divided into two classes. The first class was long insertions. Please note insertions identified without full length inserted nucleotides in the short-read platform were not included in long read-specific insertions (e.g., manta can detect insertion events with the fractions of inserted sequences from both edges).

The second class was those either of whose two breakpoints is located in repeat or low-complexity regions (Figure 2c, Supplementary Figure 4). For instance, the somatic translocation connecting chromosome 3 and 6 (chr3:26,390,428 - chr6:26,193,811) in COLO829 was missed by Illumina sequence data probably because the short-read alignment was highly ambiguous around the breakpoint of chromosome 3 (overlapping with LINE1 annotation). Another example is the somatic translocation spanning chromosome 5 and 8 (chr5:11,288 - chr8:105,285,651) in HCC1954 where the breakpoint of chromosome 5 is located near a deep subtelomeric region and annotated as simple repeat of "(TAACCC)n." Furthermore, there were several insertions and deletions that are completely contained within repeat regions.

Because of the breakpoint refinement step, nanomonsv infers single-nucleotide resolution breakpoints and non-template sequences within them (if any). To evaluate the precision of these inferences, we compared the inferred breakpoints by nanomonsv with those obtained from Illumina data. The error ratio of Illumina sequence data is much lower, and the breakpoint inference is more reliable compared to Nanopore data. The breakpoint positions by nanomonsv on Nanopore data are mostly (94.9%) within two bp of those detected by Illumina data (Figure 2d). Therefore, even though Nanopore long-read sequencing data has a considerably higher ratio of base mismatch and indels, fairly accurate identification of breakpoint positions is possible by error correction and careful examination of supporting reads and sequences around putative breakpoints.

Ninety-nine somatic SVs were those affecting known cancer-related genes[37]. These included important cancer genes such as the 12 kb deletion of *PTEN* in COLO829[30] and 5kb deletion of *STK11* in H2009 though these were also identified by the short-read platform. However, eight among them were overlooked with the short-read platform, which included two translocations affecting *EPHA3* in H2009, whose effects on tumorigenesis have been investigated in several cancers including lung carcinoma[38]. These SVs could not be identified by short-read data, probably because the breakpoints located in *EPHA3* were around a LINE1 sequence. In fact, these two translocations constituted complex SVs (discussed in the later section in detail).

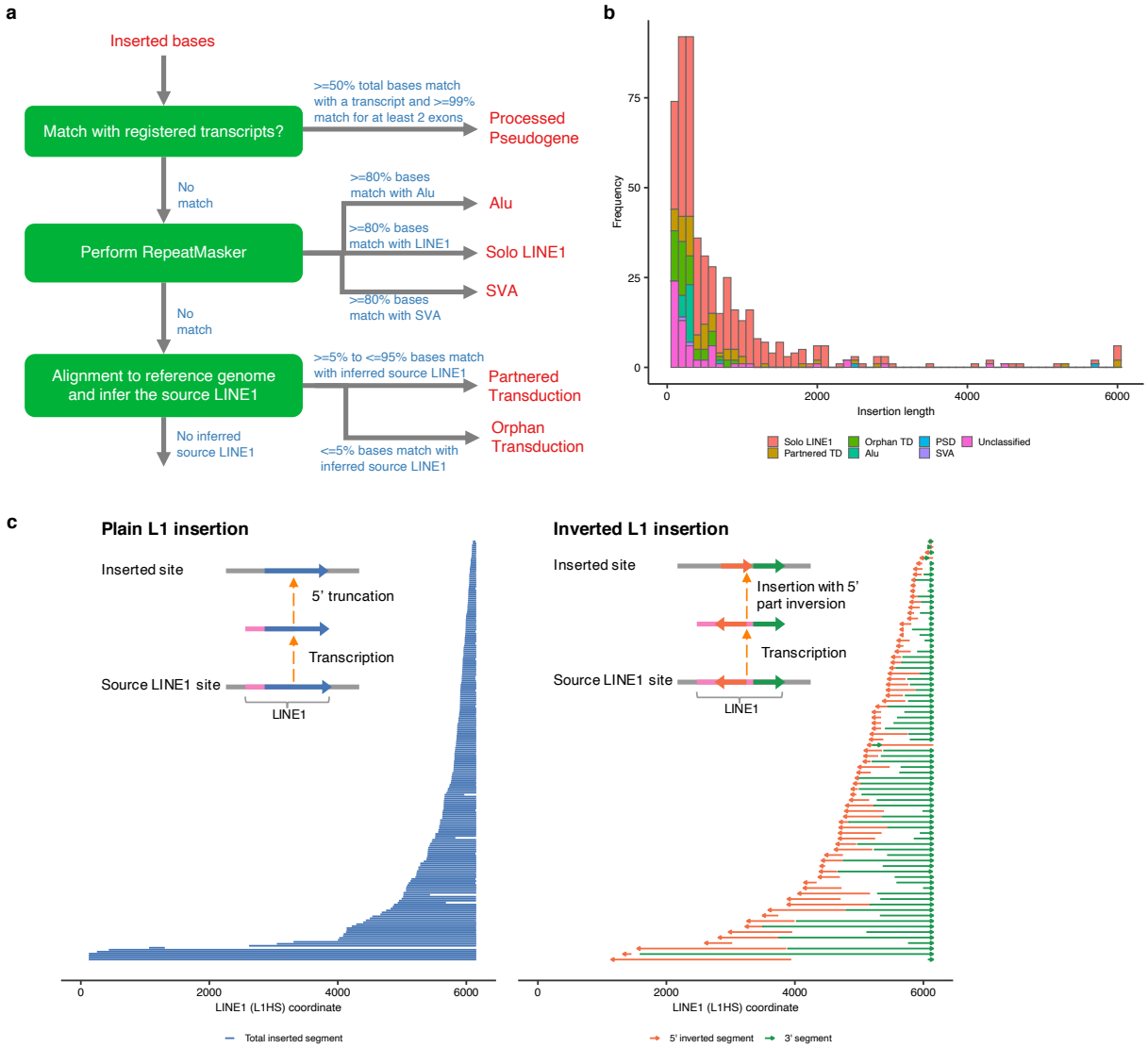

Figure 3: Classification and structure of inserted sequences between somatic SV breakpoints. (a) The chart for classifying inserted sequences used in this study. (b) The size and classification distribution (histogram in bins of 100bp) of inserted sequences. PDS, Partnered TD, and Orphan TD are processed pseudogene, partnered transduction, and orphan transduction, respectively. (c) Diagram showing the position of each solo LINE1 insertion sequence without (left) and with (right) 5′ inversion within the human-specific LINE1 sequence (L1Hs). The horizontal lines or arrows in the same vertical position show single solo LINE1 insertion events. They mostly start from the middle (by 5′ truncations) but usually end at 3′ end of LINE1.

Characterization of mobile element insertions

We identified a total of 517 long insertions, among which 500 were from H2009. For long insertions, our approach can identify complete inserted sequences as well as inserted position. There are many possible causes of insertions such as tandem duplication, MEIs, viral sequence integration, and processed pseudogene. To systematically characterize the inserted sequences, especially focusing on MEIs in this study, we have developed a pipeline for classifying the

inserted sequences based on comparison with transcriptome, annotation by repeat sequence information, and re-alignment to the reference genome (see Figure 3a).

First, if the inserted sequence significantly matched with a transcript, the insertions were classified into processed pseudogene[39,40], which are copies of mRNAs integrated into the genome by reverse transcriptase activity of LINE1 elements. We identified two processed pseudogenes in H2009. One is full-length *IBTK* (ENST00000306270) coding sequence insertion into chromosome 9. The other is the portion of *CARNMT1* (ENST00000376834), from the middle of exon 3 to the halfway of exon 8 with an internal inversion, integrated into chromosome 6 (See Supplementary Figure 5). Although the existence of these pseudogene insertions had been identified by the short-read platform using the same cell-line[39], a detailed structure of the entire inserted sequence such as the position of the inversion breakpoint could be first confirmed in this study.

Second, when three major mobile elements (LINE1, Alu, and SVA) covered most of the inserted sequence (>=80% by RepeatMaseker, http://www.repeatmasker.org), the inserted sequences were categorized into each class. We identified 323 LINE1, 23 Alu, and 2 SVA insertions in three cell-lines, respectively (Figure 3b). The LINE1 insertions are known to be frequently accompanied by inversion at the 5′ end, the mechanism of which can be explained by "twin priming"[41]. In fact, by investigating the matched position and direction of inserted sequences with LINE1, the 5′ inversions were observed in 79 (24.46%) of LINE1 insertion. Often, 5′ inversions were accompanied by the elimination of internal LINE1 sequences, which may occur during the integration process (Figure 3c). We also observed other complex structural changes. One simple example was 1,100 bp insertion at chromosome 14, which was a direct concatenation of 160 bp 5′ end and 900 bp 3′ end LINE1 sequence without a 5′ inversion. The internal part of the LINE1 sequence may be removed in the integration process of the full LINE1 sequence. These diversities of insertion structures produces deviations between inferred insert sequence lengths from short-read sequence data and those directly obtained from long-read sequence data (See Supplementary Figure 6) because accurate inference of the insert nucleotide length from short-read data is only possible when structures of insertions are relatively simple (such as full-length insertion, simple 5′ truncation or 5′ inversion without any removals of internal bases).

Next, the insertions not categorized at this stage were aligned to human genome sequences to explore the possibility of LINE1 transductions, in which immediate 3′ flanking sequences of LINE1 elements are mobilized through the continued transcription beyond the 3′ edge of LINE1 and its retrotransposition. Transposed sequences can be both parts of LINE1 elements and their downstream sequences (partnered transductions) or only downstream ones (orphan transductions). When there existed LINE1 elements upstream of the aligned site of inserted sequences, these LINE1 elements were inferred to be the source of transduction. As possible LINE1 source elements, we first extracted 5,228 full-length recent primate-specific LINE1 elements from the human reference genome (reference putative LINE1 source elements). In addition, since it is known that there are several active non-reference LINE1 source elements, which are not included in the reference genome but can be detected as polymorphic insertions, we also included 652 and 2610 full-length LINE1 insertions identified in 1000 genomes Phase 3[42] and gnomAD v2.1[43], respectively. Furthermore, when many inserted sequences were aligned to the same genomic locations, we searched for the germline LINE1

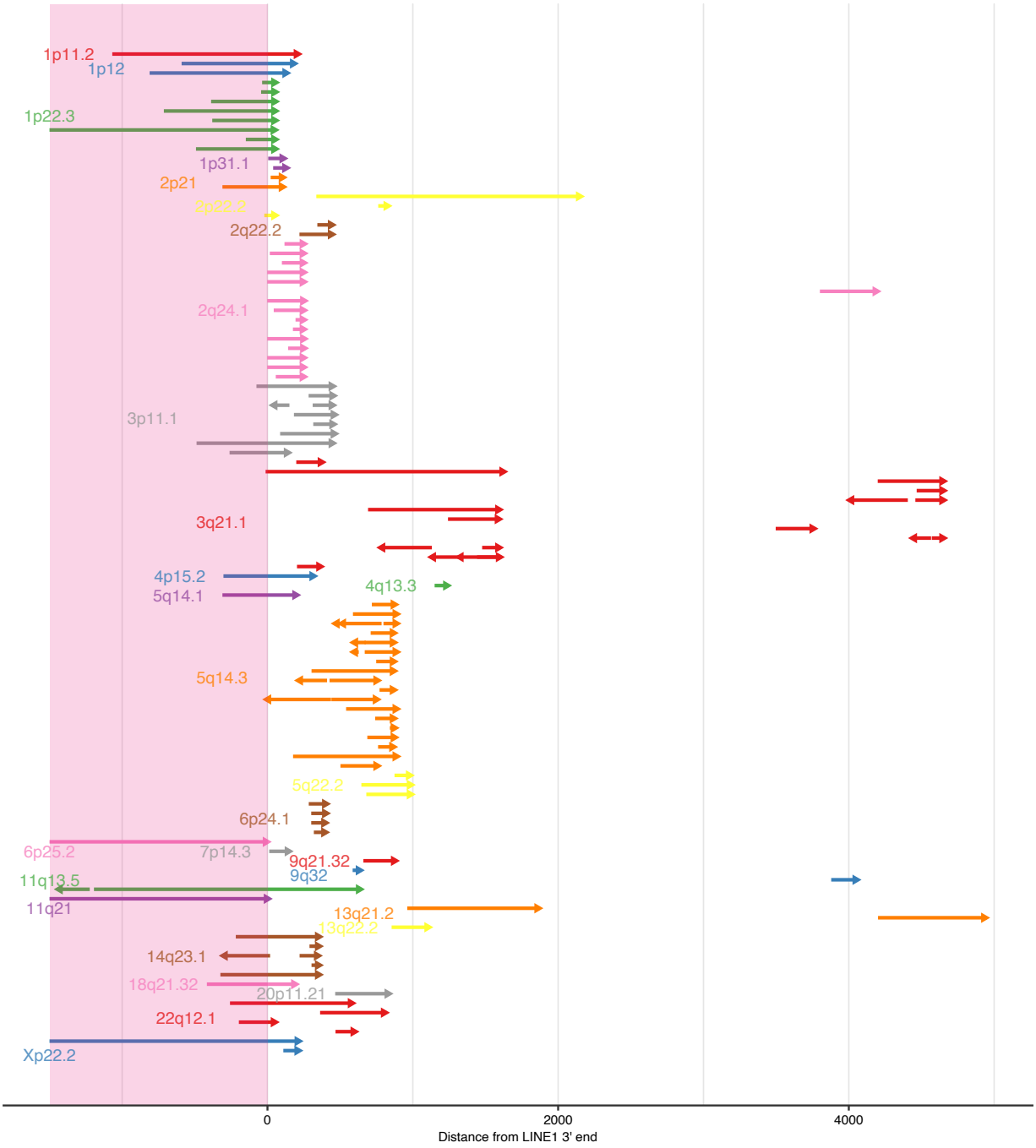

Figure 4: A comprehensive picture of L1 transductions identified in H2009. Horizontal arrows in each vertical positions show distinct LINE1 transduction events of each corresponding LINE1 source site (distinguished by color and labeled by cytoband). Arrows starting before the position of 3′ end (within LINE1 sequences shaded by light pink) are classified as partnered transductions, whereas those starting past LINE1 sites are orphan transductions. Multiple arrows in one line show some structural changes in the inserted sequences (most typically internal inversions depicted by two outwardly directed arrows).

insertion near those positions from the normal sequence data and manually curated the putative rare germline LINE1 insertions that were considered as the source of LINE1 transduction.

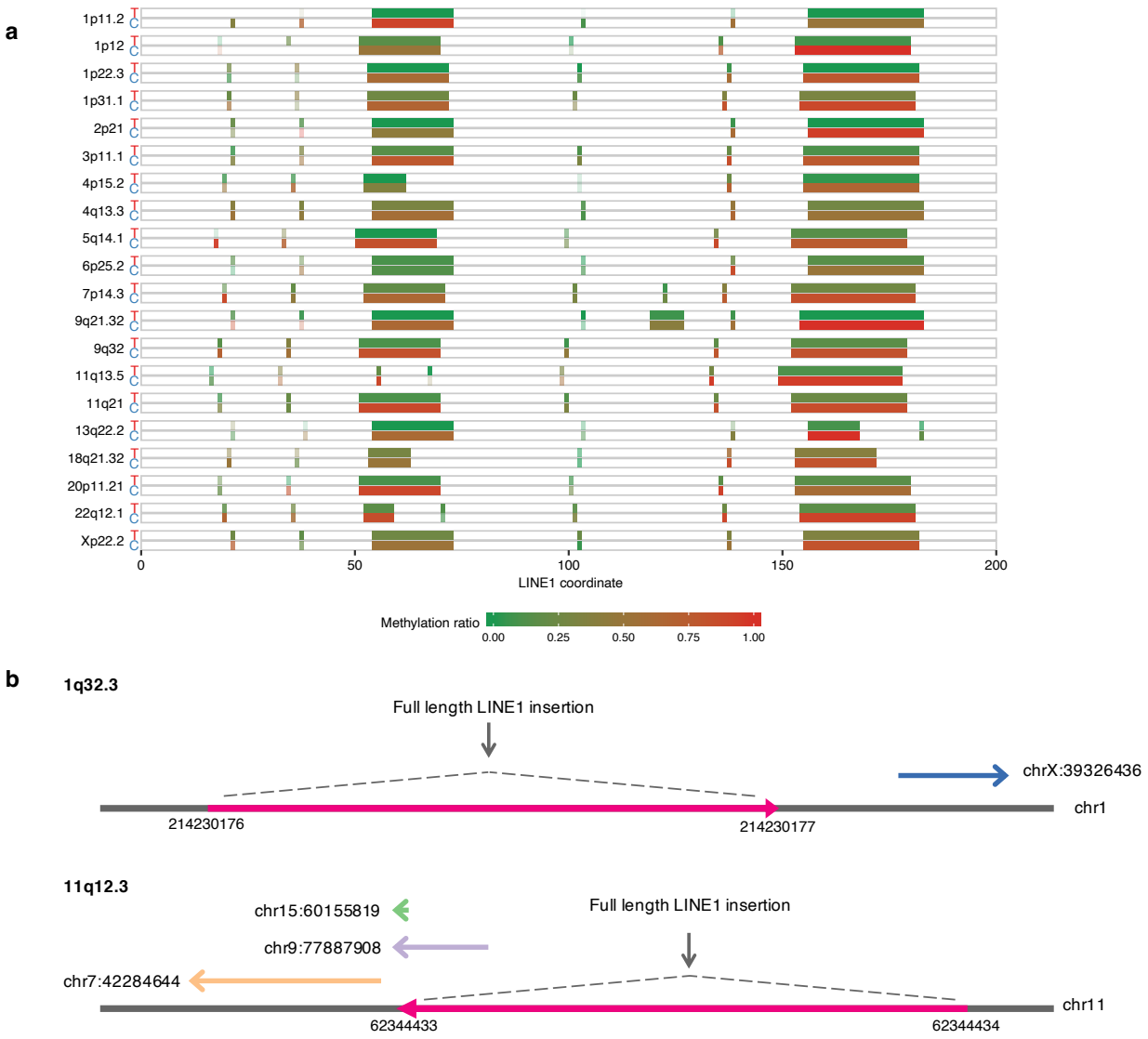

Figure 5: A comprehensive picture of L1 transductions identified in H2009 (continued). (a) Methylation status of promoters of somatic LINE1 source elements for H2009. For each LINE1 source site (labeled by cytoband), the upper and lower boxes represent the tumor (T) and matched control (C) methylation states. After the detection of methylated bases for each CpG site using nanopolish, the ratios of methylations are calculated. Contrasting density is determined by the depth of sequence covering each site. (b) Examples of nested LINE1 insertion identified in H2009. Two full-length LINE1 insertion sites became the new active sources of LINE1 transductions. The novel source site at 1q32.3 generated one orphan LINE1 transduction. The second novel source site at 11q12.3 eventually produced two partnered transductions and one orphan transduction.

We identified 107 somatic transduction events (57 were partnered and 50 were orphan transduction) from 29 putative LINE1 source elements (Figure 4). Among 20 LINE1 sources from the reference genome, 17 belonged to the Human-specific LINE1 (L1HS) subfamily, whereas the remaining three elements were the L1PA2, the second youngest primate-specific subfamily. The fact that 9 non-reference LINE1 source elements (4 from 1000 genome Phase 3, 3 from gnomAD and 2 from manual curation of matched control of H2009 cell-line) could be identified corroborates the importance of population- and individual-specific hot LINE1 elements[44]. Several transductions included the 5′ inversions, implying that the same mechanism

as solo L1 insertion such as twine priming functions during reverse transcription. For each LINE1 source element, 3′ end positions of the inserted sequences tended to concentrate at the close genomic positions. This may be because these 3′ end positions are probably the location where the transcription is terminated, and the positions with a potency of transcription termination may be scattered because they require some characteristic sequences. It has been suggested that the localized hypo-methylation of the LINE1 promoter region drives the somatic activation as source elements[19]. To confirm this, we quantified the methylation level using nanopolish[45] on raw signal-level data of Nanopore sequence data. For all the 20 reference LINE1 source elements, the methylation ratios were lower for the tumor sample compared to the matched control (Figure 5a), which is consistent with the proposed hypothesis. We also identified two examples of nested LINE1 transduction[19], where somatically inserted LINE1 elements themselves became the source of next LINE1 transduction (Figure 5b). All these results indicate that long-read sequence data has a great potential for characterizing the various mechanisms of genomic insertions.

Phasing of structural variations

Here, we provide a simple approach for phasing the co-occurrence of multiple breakpoint junctions to elucidate the form of complex SVs using long-read sequencing data (Figure 6a). First, we identified the genomic segment sandwiched by two breakpoints that were supported by the same read. Here, we confirmed the consistency between the sizes of genomic segments and positional relationships of two breakpoints in the supporting reads, and we termed the consequent segments as Genomic Segment Sandwiched by SVs (GSSSVs) and two SVs are phased at this point. Next, when there were SVs shared by distinct GSSSVs, these GSSSVs, which initially included only one genomic region, were connected to generate new GSSSVs including two genomic regions and three SVs were phased up here. We repeated these procedures greedily until when there was no shared SVs between distinct GSSSVs.

By using this approach, 8, 73, and 301 SVs were phased to any other SVs and grouped into 3, 31, and 106 clusters for COLO829, H2009, and HCC1954, respectively (Figure 6b). The majority of clusters consisted of a few SVs. Many of them constituted characteristic complex SVs whose properties had been discussed in previous studies (Figure 6c). One example identified in COLO829 was the combination of two inversions and one deletion spanning 126 kb genomic region in 9p21.1, which can probably be explained by replication-based mechanisms[24,25]. Another instance from HCC1954 was a “chain of templated insertion,” in which several genomic segments across the genomes were inserted between a rearrangement. We also identified the co-occurrence of a 3p11.1 genomic segment insertion and 2,581 bp deletion in the genomic region in 18q21.32, which disrupted cancer-related genes *EPHA3* in H2009. The inserted 3p11.1 genomic segment was, in fact, LINE1 source element inducing at least eight transductions by the analysis of the previous section. Therefore, this is probably a LINE1-mediated deletion[20]. Finally, we found clusters with an enormous number of (>= 10) SVs in HCC1954. The SVs in these clusters were concentrated on chromosomes 5 and 8 especially near telomeres, and many fragmented genomic segments were rejoined in an unorganized way (Figure 6d, Supplementary Figure 7). These are hallmarks of chromothripsis induced after

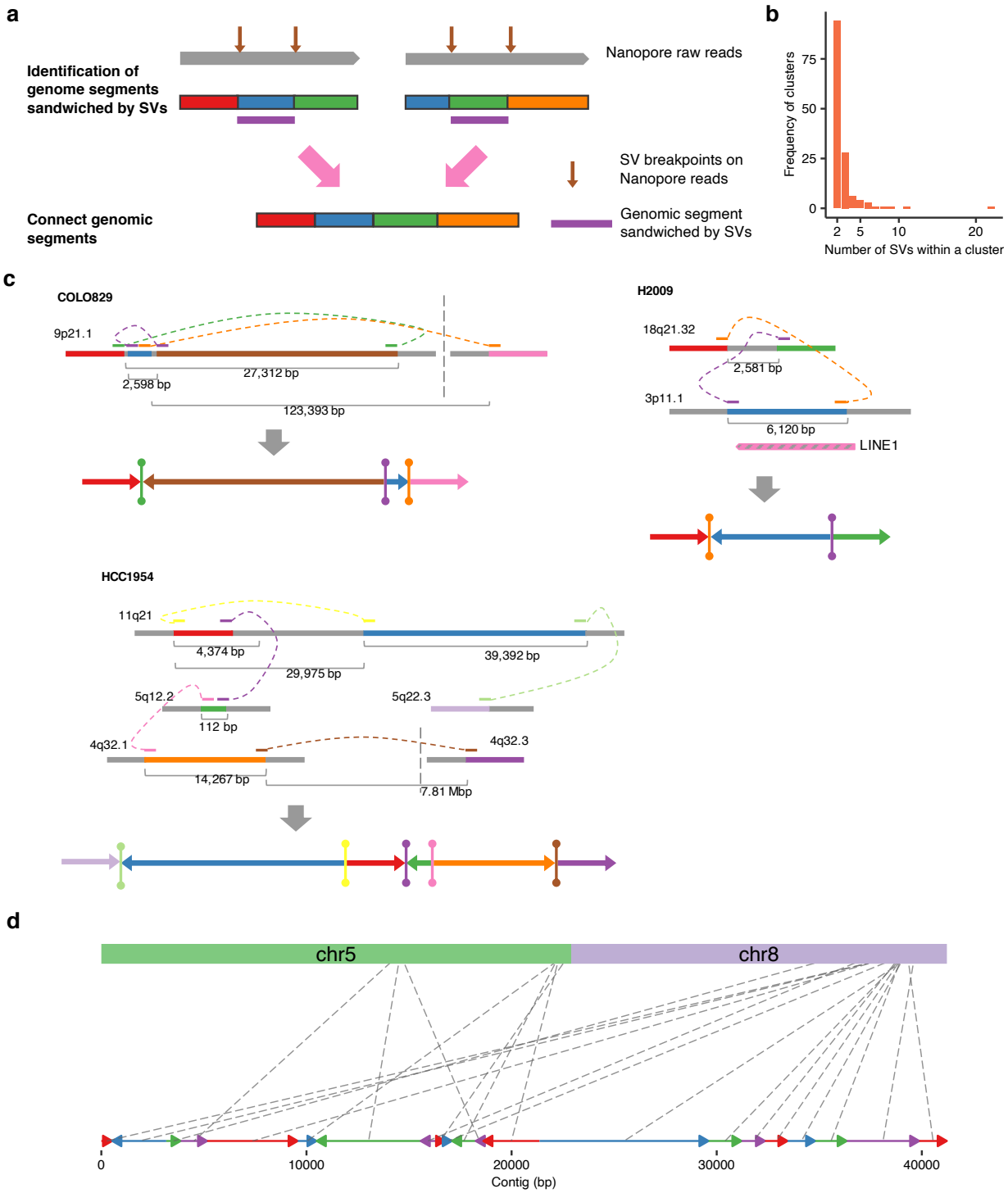

Figure 6: An approach for phasing multiple SVs and their summary of results. (a) A diagram for phasing SVs. (b) The distribution of the number of SVs within each cluster by the above procedure. (c) Examples of complex SVs for an aberration by replication-based mechanism (COLO829), a LINE1 associated deletion (H2009), and a chain of templated insertion (HCC1954). (d) Highly complex SV clusters possibly by telomere crisis. The original genomic position of each sequence segments in the contig (distinguished by colors) is linked by the dashed arrow.

telomere crisis[46,47]. Although these characteristic complex SVs have been thought to be generated in a single event, it is difficult to confirm this especially using short-read sequence

data. Long-read sequencing data can give more credible evidence of conjunction. First and foremost, the fact that several breakpoint junctions constituting these complex SVs were phased by the same supporting reads indicates at least a derivative chromosome having them exists. It is still possible that these individual junctions occurred progressively and converged to the phased form by chance. In that case, there may exist reads supporting a portion of junctions of the cluster and do not support the other junctions even though genomic positions of them are covered. But we could not find these reads at least for the examples listed above.

Discussion

We could demonstrate that the proposed approach could identify somatic structural variations from a pair of a tumor and its matched control long-read sequencing data. The proposed method is simple and did not include many heuristic filtering steps such as the panel of normal technique, where detecting putative germline or artifactual variants using several control samples with relatively relaxed criteria and gather them as a “blacklist variant set” and then, candidate of somatic variants that matched with any of these lists are removed. Although the panel of normal technique[48,49] has been successfully used in many projects using short-read sequencing data[50], this has some caveats; We need to prepare many control sequencing data from the same platform, and there is a risk that true somatic variants could be filtered out because of tumor contamination in the control samples. It seemed that our pipeline attained sufficient precision without this heuristics, demonstrating the power of long-read sequencing data. However, if in the case of no matched control, the heuristics of panel-of-normal may work efficiently.

We could determine the breakpoints of SVs with a single-nucleotide resolution with non-templated sequence insertions to some extent. Currently, most sophisticated algorithms on short-read platforms support single-nucleotide resolution detection by the use of split-read evidence or local assembly. However, there had been few evaluations on the resolution of breakpoints of SVs using even much noisy long-read sequencing data. This will ease us to identify micro-homology and non-templated sequence insertions, which could provide us with valuable information on the mutational mechanisms of SVs[25,51]. Additionally, for comparing and annotating with SVs registered in a public database, single-nucleotide resolution characterization is highly preferable. For single-base resolution breakpoint identification, we combined (1) the generation of refined contig sequences by multiple alignments of supporting reads and (2) the identification of breakpoints by a custom Smith-Waterman algorithm allowing one-time jump from one region around a breakpoint to the other. Although these approaches were accurately detected breakpoints in most cases, we feel this part could be improved even more. One possibility for improvement would be to use other error correction techniques such as those using auxiliary short-read alignment[52].

Our pipeline could successfully recapitulate the entire structures of several long insertions (~ 6000 bp). Nevertheless, considering the statistics about the sequence length (e.g., median of 5,000 ~ 8,000 bp), the sensitivity of long insertion, especially for those with relatively long sequences such as full LINE1 insertion, may be relatively low. On top of LINE1 insertion, viral integration into the cancer genome is fairly frequent in cancers such as human papillomavirus (~8,000 bp) in multiple cancers[53], Hepatitis B virus (~3300 bp) in liver cancers[54]

and Human T-cell leukemia virus type I (~9,000 bp) in adult T-cell leukemia/lymphoma[36]. Although current data would be sufficient to detect and characterize the detailed structure of these insertion events to some extent, slightly longer sequence data would be desired. We believe that future chemistry updates or kit improvements will provide stable solutions to these problems.

Reconstruction of the entire structure of complexly rearranged cancer genomes is a common goal for the cancer genome researcher. Although there have been many sophisticated approaches that combine somatic SVs, copy number changes with mathematical algorithms[27,55], their power to characterize complex SVs were limited due to short-read data. As we demonstrated that even a simple approach that investigates the co-occurrence of SVs in the same supporting reads could derive the complicated relationships among breakpoints, long-read will definitely refine the complex structural variation landscape. Still, just the information on co-occurrence is not sufficient for the reconstruction of the entire cancer genome. Overall, our novel approach and analysis demonstrated that long-read sequencing technologies are useful for capturing more precise landscapes and mechanisms of somatic SVs.

Method

Whole genome sequencing using Oxford Nanopore Technologies and Illumina Novaseq 6000

The cell-lines used in this study (COLO829, COLO829BL, H2009, BL2009, HCC1954, and HCC1954BL) were obtained from ATCC (American Type Culture Collection). For Nanopore sequencing data, high-molecular-weight genomic DNAs were extracted from these cell-lines with QIAGEN Genomic-tip 500/G (QIAGEN). Libraries were then prepared using the Ligation Sequencing Kit 1D and sequenced on the PromethION platform with R9.4.1 flow cells (Oxford Nanopore Technologies), to generate fast5 files. Then, these fast5 files were basecalled and converted to fastq files using Guppy 3.4.5. Then, these were aligned by minimap2 with "-ax map-ont -t 8 -p 0.1" option to the human reference genome provided at the Genomic Data Commons website (GRCh38.d1.vd1). For Illumina short-read sequencing data, we performed Illumina Novaseq 6000 with a standard 150 bp paired-end read protocol, and these were aligned by BWA-MEM[56] version 0.1.17 to the same human reference genome and were sorted by the genomic coordinates and remove PCR duplicates via biobambam (https://github.com/gt1/biobambam) version 0.0.191 as previously described[57].

Detailed algorithm of nanomonsv

In nanomonsv, SVs are divided into three categories according to how these SVs are supported by each read. We denote the SVs that are supported by single alignment with insertion or deletion ('I' or 'D' of CIGAR strings) as "I-type" and "D-type," respectively. On the other hand, SVs that are represented by multiple alignments (primary alignment and one or more supplementary alignments), and thus each alignment is accompanied by soft clipping ('S' of

CIGAR strings) are denoted as “S-type.” Please note that the procedure of each step becomes slightly different depending on these types.

Parsing step

Parsing I-type and D-type SV supporting reads

For putative I-type and D-type SV supporting reads, we parse a CIGAR string of each alignment of the input BAM file to collect the information such as chromosome, indel start, and end positions, putative indel size, and read IDs and organized as BED file. Then, the records are sorted by the genomic coordinate and bgzip’ed and tabix’ed (http://www.htslib.org/).

Parsing S-type SV supporting reads

In order for gathering S-type supporting reads, we search for multiple “consecutive alignment” of a single read, in which query end of one alignment is in close proximity (within 50 bp) to the query start of the next alignment, and thus the corresponding genomic coordinates become the breakpoint of putative SVs. First, by parsing the input BAM file, query start and end positions and target (genomic) start and end positions, as well as alignment directions, are collected for each read ID and alignment (primary and supplementary alignments not including secondary alignments). Then, for each read ID, we find the “consecutive alignment,” and the possible ranges (± 30 bp margin from the corresponding genomic coordinates) of the two genomic breakpoints of putative SVs, breakpoint direction and read ID are recorded and organized as BEDPE format. Then, these records are sorted by genomic coordinates and bgzip’ed and tabix’ed.

Clustering step

For each two S-type SV supporting reads, when both the possible ranges of the breakpoints overlap, they are merged so as to support the same SV. This procedure is repeated until there is no pair of supporting reads to be merged. For I-type and D-type SV, when the possible ranges of the indels overlap and the size of indel is about the same (within 20%), the two supporting reads are merged. Then, for each cluster, we remove those having less than three supporting reads or median of <= 40 mapping qualities. Also, if there exist the apparent supporting reads for the putative SVs in the matched control sample, these SVs are removed.

Refinement step

Consensus sequence generation

First, for candidate SVs, we extract the part of supporting reads around the breakpoint. For D-type and S-type SVs, 300 bp sequences before and after the position corresponding to SV breakpoints within the supporting reads are extracted. For I-type SVs, the entire inserted sequences as well as 300 bp from both ends are derived for the supporting reads. Next, for each SV, we perform multiple alignment using MAFFT[29] to generate the consensus sequence.

Then, to determine the consensus sequence, the most nucleotides at each position of multiple alignment are taken (skip that position in the case of '-').

SV breakpoint coordinate determination

A one-time jump Smith-Waterman (OJ-SW) algorithm, where one query sequence is compared with the two target sequences (starting from the target sequence 1 and switched to the target sequence 2 at some point, see Supplementary Figure 1) is used to determine the coordinates of breakpoints and inserted sequences within them. For each D-type and S-type SV, the two sequences around the regions where the possible locations of the first and the second breakpoints are extracted from the human reference genome sequence and are used as the target sequences 1 and 2 for the OJ-SW algorithm, respectively. The consensus sequence generated in the above step is set as the query sequence. After performing the OJ-SW algorithm, the two genomic coordinates corresponding to where the jump from the target sequence 1 to 2 occurred are determined to be the SV breakpoints, and the skipped query sequence by the leap is set as the inserted sequence between the breakpoints. For each I-type SV, the sequences around the putative insertion start and end positions within the produced consensus sequence are set as the target sequence 1 and 2 for the OJ-SW algorithm, respectively. For the query sequence, the sequences around the region where the insertion is considered to be located are extracted from the reference genome. Then, after performing the OJ-SW algorithm, the position where the jump occurred within the query sequence is set to be the exact coordinate of insertion, and the skipped sequences are set to be the deleted nucleotides. In addition, the points where the jump happened in the target sequences are set as the start and end of inserted bases.

Validation

For each SV candidate, constitute the putative SV segment sequence by concatenating 200 bp sequences from both the breakpoints. When there is an inserted sequence, two SV segment sequences are prepared: For each breakpoint, we prepare 200 bp sequences from one breakpoint joined to the 200 bp sequences in the opposite direction including the inserted sequence and subsequence nucleotides after the other breakpoint (if the size of insertion is below 200bp). Therefore, putative SV segment sequences are 400 bp in size. Then, after collecting the Nanopore reads spanning the SV breakpoint from both tumor and matched control samples, local alignments of SV segments sequences to each read are performed using the SSW library[58] with default score settings (match, mismatch, gap opening, and gap extension scores are 2, -2, -3, and -1, respectively). Here, in order for each Nanopore read to be an "SV variant read" (the read containing the SV segment sequence), we request that the alignment score (we adopt the larger score between the two SV segment sequences in case there is an inserted sequence) is equal or more than 560. Then, we count the number of SV variant reads for tumor and matched control samples and keep the SVs whose SV variant reads equal or more than 3 for the tumor and zero for the control samples.

Classification of the inserted sequences

Check for processed pseudogene

First, the inserted sequences of insertion SVs are aligned to the human reference genome using minimap2 with the Nanopore 2D cDNA-seq option, "-ax splice." Then, for each alignment, the intersection with the exonic regions by comprehensive gene annotation set from GENCODE version 31 is investigated. If there exists a gene in which more than one exon matches >= 95% and matched >= 50% in total, then this insertion is determined to be a processed pseudogene.

Check for Alu, Solo LINE1, SVA insertion

For the remaining insertions, we perform the RepeatMasker with "-species human" option. Then, the portions of bases annotated as LINE1 ("LINE/L1"), Alu ("SINE/Alu") or SVA ("Retroposon/SVA") among the total nucleotides subtracted by the parts annotated as poly-A or poly-T ("(T)n" or "(A)n") are calculated. When either of them is equal or more than 80%, then the insertion is classified into those categories.

Check for partnered or orphan transduction

First, we created a database of possible sources of LINE1 transduction. For LINE1 included and annotated in the human reference genome, we downloaded RepeatMasker file (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/rmsk.txt.gz) and selected the records whose family is L1, whose subfamily is among those of recent primate-specific ones (L1HS, L1PA2, L1PA3, L1PA4, and L1PA5), and whose size is equal or larger than 5,800, resulting in 5,228 records. Then, for those not included in the reference genome (and thus the polymorphism of LINE1 insertion), we obtained 1000 genomes Phase 3 SV file (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/integrated_sv_map/ALL.wgs.mergedSV.v8.20130502.svs.genotypes.vcf.gz) and filtered them by "bcftools filter" command (https://github.com/samtools/bcftools) with "INFO/SVLEN > 5800 && INFO/SVTYPE == 'LINE'" option, remaining 652 records. Also, we extracted gnomAD v2.1 SV file (https://storage.googleapis.com/gnomad-public/papers/2019-sv/gnomad_v2.1_sv.controls_only.sites.vcf.gz) and selected near full-length LINE1 polymorphisms by "bcftools filter" command with the "ALT == '<INS:ME:LINE1>' && INFO/SVLEN >= 5800" option. Since these 1000 genomes and gnomaAD SV files are based on the hg37 reference genome, we converted the coordinates using liftOver[59] to the hg38 coordinate system. Then, all the records were merged into one bed file and bgzip'ed and tabix'ed.

For each inserted sequence, alignment is performed using BWA-MEM[56]. Then, we checked whether the primary alignment has >= 30 mapping quality and any records of possible LINE1 source databases constructed above within 5,000 bp. If these requirements are not met, then the inserted sequence is classified into "Other." When these are satisfied, we set the proximal record as the corresponding LINE1 source element for the transduction, and we extract all the supplementary alignment that is within 5,000 bp of the primary alignment for possible inversion. Then, by the portion of bases annotated as LINE1 by RepeatMasker, the

insertion is classified into Orphan transduction if the ratio is below 0.01, or Partnered transduction otherwise.

Structural variation detection from short-read sequencing data

GenomonSV

GenomonSV (https://github.com/Genomon-Project/GenomonSV) version 0.7.2 was used. First, “GenomonSV parse” command was performed for both tumor and matched control BAM files. Then, “GenomonSV filt” was performed on the tumor data with the options “--min_junc_num 2”, “--min_overhang_size 30”, and “--max_control_variant_read_pair 10” with specifying the matched control BAM file for the “--matched_control_bam” option. Then we performed additional filtering with sv_utils filter, custom software for post-processing GenomonSV results (https://github.com/friend1ws/sv_utils), with “--min_tumor_allele_freq 0.07”, “--max_control_variant_read_pair 1”, “--control_depth_thres 10”, and “--inversion_size_thres 1000” options.

Manta

We used manta (https://github.com/Illumina/manta) version 1.6.0. First, we performed configManta.py with the default options and runWorkflow.py for each tumor and matched control pair with “-m local,” and “-j 8” options. Then, we extracted records tagged with “PASS” in the FILTER columns using “bcftools view” command.

SvABA

SvABA (https://github.com/walaj/svaba) version 1.1.0 was used. First, we performed “svaba run” command for each tumor and matched control data using “-p 8”, “-v 1 -A” options. Then, we performed filtering by “bcftools view” command with the “-f PASS” option and “bcftools filter” command with the “'FORMAT/AD[0:0]<=1&&FORMAT/AD[1:0]>=2&&FORMAT/DP[0:0]>=10&&FORMAT/DP[1:0]>=10” option.

GRIDSS

We used GRIDSS (https://github.com/PapenfussLab/gridss) version 2.8.0. First, “gridss.sh” was performed on tumor and control pairs with “-j gridss-2.8.0-gridss-jar-with-dependencies.jar”, “-t 8”, and “--picardoptions VALIDATION_STRINGENCY=LENIENT” options. Then “gridss_somatic_filter.R” was performed with the default option. Then we used the “bcftools view” command with “-i INFO/MATEID[0]!=’’ and “-f PASS” options.

TraFic-mem

First, since TraFic-mem currently only supports GRCh37 based BAM files, we aligned the short-reads to the GRCh37 human reference genome. Then, we performed TraFic-mem using Docker

image mobilegenomes/trafic:multispecies with default options. Then, we converted the coordinates to GRCh38.

Merge the results

Even for the identical SV, there are often slight deviations in inferred breakpoint coordinates across the software. Therefore, when SVs called by different software share the two breakpoints in close proximity (<= 10 bp), we deemed them as the same SV. GenomonSV, manta, and GRIDSS on Illumina sequencing data mostly produced equivalent coordinates of breakpoints whereas SvABA (at least the version we used) seemed not to provide non-exact breakpoint positions especially when the breakpoints share microhomology. Therefore, for the comparison of breakpoint coordinates, we did not use the results of SvABA.

PCR validation

To generate primer sequences for PCR validation for each somatic SV, we first prepared the sequence template by concatenating 800 bp nucleotides from the first breakpoint, the inserted sequence, and 800 bp nucleotides from the second breakpoint. Then, the Python bindings of Primer3[60] are performed, setting the sequence target as 25 bp nucleotides from the first breakpoint, the inserted sequence, and 25 bp nucleotides from the second breakpoint. Here, we created five pairs of primer sequences for each primer product size range of 201 to 300, 301 to 400, 401 to 500, …, 1501 to 1600. Next, we performed GenomeTester[61] to remove pairs of primer sequences that have too many binding sites (more than 10 for left or right primers) and too many alternative PCR products (more than two for insertion and deletion and more than one for other types of SVs). Finally, for each somatic SV for validation, we selected one primer pair that has a smaller product size, less number of primer binding sites, and alternative PCR products.

All PCR reactions were performed in a total of 20 uL volume using 10 uL of Go Taq Master Mix (Promega), 1 uL of each primer (Final 0.5 nM), 1 mL of gDNA (20 ng), and 8 uL of double-distilled water. The PCR samples were denatured at 95°C for 2 min, subjected to 40 cycles of amplification (95°C for 30 sec, 55°C for 30 sec and 72°C for (product size (bp) / 1,000) min and followed by a final extension step at 72°C. A list of primers is provided in Supplementary Data 2. PCR products were resolved by agarose gel electrophoresis. Representative PCR products were purified using QIAquick Gel Extraction Kit (Qiagen) according to the manufacturers' recommended protocols. Finally, the purified samples were subjected to direct capillary sequencing (eurofin). All sequence data were analyzed using ApE (https://jorgensen.biology.utah.edu/wayned/ape/) and the Chromas Lite viewer (Technlysium Pty., Ltd.).

Methylation analysis

To quantify the amount of methylation, we used nanopolish version 0.11.1 (https://github.com/jts/nanopolish). First, we performed the "nanopolish index" command from the original fast5 file to generate the index that associates read IDs and their signal-devel data.

Then, we executed the “nanopolish call-methylation” command to make the TSV file summarizing the log-likelihood ratio for methylation for each read ID and genomic position. Then, using the script provided on the software website, we obtained the ratios of methylation at each genomic position.

Data availability

The raw Oxford Nanopore sequence data and Illumina short-read sequence data used in this study will be available through the public sequence repository service.

Acknowledgment

This work is supported by Grand-in-Aid from the Japan Agency for Medical Research and Development (Project for Cancer Research and Therapeutic Evolution, 19cm0106538h0002). The authors thank Kana Shimizu and Taiki Yamada for fruitful discussions. The authors also thank Raúl Nicolás Mateos Ramos for reading the manuscript.

Author Contributions

Y.Shiraishi and KK designed the study. JK, Y.Saito, and KK contributed to data acquisition. Y.Shiraishi designed and implemented nanomonsv. KC and AO provided computational assistance. JK performed wet-lab validation of somatic SVs. Y.Shiraishi analyzed and interpreted data. Y.Shiraishi and JK generated figures and tables. Y.Shiraishi wrote the manuscript with the help of JK, Y.Saito, and KK. All authors participated in the discussion and interpretation of the data and result.

Supplementary Information

Supplementary Figure Legends

Supplementary Figure 1: Smith-Waterman algorithm with one-time jump to determine the SV breakpoints. A schematic of a one-time jump Smith-Waterman algorithm used to determine the exact breakpoints of structural variations. The first part of the consensus sequence is aligned to the genomic region A and the latter part is aligned to the genomic region B. This is basically the same as the standard Smith-Waterman algorithm except that one-time-only jump is allowed from genomic region A to B during the procedure, and the position where the jump occurred is determined to be the inferred breakpoint. When there are several inserted nucleotides, several bases of consensus sequences are also skipped during the jump.

Supplementary Figure 2: Distribution of Nanopore read length for samples used in this study. Here, we just counted the lengths of primary alignment reads (those without the

secondary alignment (0x100) nor the supplementary alignment (0x800) sam flag bits). The bin widths of histograms are 2000bp.

Supplementary Figure 3: Some somatic SVs validated by PCR for H2009. PCRs were performed on tumor (T) and matched control (C) DNAs. The bottom keys correspond to the SV_ID in Supplementary Data 2. Bands for the target SVs were pointed by red arrows. For insertions, tumor-specific bands as well as common bands for tumor and control DNAs, which are shorter because of the lack of inserted sequences, could be observed.

Supplementary Figure 4: Examples of somatic SVs identified specifically by long-read and its validation by Sanger sequencing. (a) The somatic translocation, chr3:26,390,428 - chr6:26,193,811, in COLO829. The breakpoint at chromosome 3 is located in LINE1 sequences. (b) The somatic translocation, chr5:11,288 - chr8:105,285,651, with inserted sequence spanning chromosome 5 and 8 (chr5:11,288 - chr8:105,285,651) in HCC1954 in which the breakpoint of chromosome 5 is located near deep subtelomeric region and annotated as simple repeat of "(TAACCC)n."

Supplementary Figure 5: Example of processed pseudogene insertion identified in H2009. A CARNMT1 pseudogene was somatically inserted into chromosome 6. Features frequently seen in LINE1 retrotransposition such as a target site duplication, an internal inversion, polyA tail, and 5′ truncation could be observed.

Supplementary Figure 6: Comparison between the estimated sizes of solo-L1 inserted sequences from Illumina and Nanopore data. Each point represents an insertion detected by both TraFic-mem on Illumina and nanomonsv on Nanopore sequence data. Insertions are stratified by the existence of inversion, Plain (without any inversions), Inverted (with one 5′ inversion), and "Other" (with multiple inversions). Complex LINE1 insertions (Inverted and Other) tend to have different insert size estimation.

Supplementary Figure 7: Several examples of massive clustered SVs in HCC1954. (a-d) In each cluster, the lower part shows how each genomic segment is joined through the SVs. Each arrow shows the genomic segments which are aligned to chromosomal locations connected by dashed lines with corresponding directions.

Supplementary Data

Supplementary Data 1: List of somatic SVs identified by nanomonsv.
Supplementary Data 2: List of somatic SV and primers for PCR validation.

Reference

1. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes.

Nature **578**, 102–111 (2020).

2. Quigley, D. A. *et al.* Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. *Cell* **174**, 758–769.e9 (2018).
3. Kataoka, K. *et al.* Aberrant PD-L1 expression through 3'-UTR disruption in multiple cancers. *Nature* **534**, 402–406 (2016).
4. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013--2015. (2015).
5. de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* **7**, e1002384 (2011).
6. Sedlazeck, F. J., Lee, H., Darby, C. A. & Schatz, M. C. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* **19**, 329–346 (2018).
7. Sherman, R. M. & Salzberg, S. L. Pan-genomics in the human genome era. *Nat. Rev. Genet.* **21**, 243–254 (2020).
8. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
9. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).
10. Sedlazeck, F. J. *et al.* Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods* **15**, 461–468 (2018).
11. Gong, L. *et al.* Picky comprehensively detects high-resolution structural variants in nanopore long reads. *Nat. Methods* **15**, 455–460 (2018).
12. Cretu Stancu, M. *et al.* Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat. Commun.* **8**, 1326 (2017).
13. Heller, D. & Vingron, M. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35**, 2907–2915 (2019).

14. Sakamoto, Y., Xu, L., Seki, M., Yokoyama, T. T. & Kasahara, M. Long read sequencing reveals a novel class of structural aberrations in cancers: identification and characterization of cancerous local amplifications. *bioRxiv* (2019).
15. Roth, A. *et al.* JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics* **28**, 907–913 (2012).
16. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
17. Burns, K. H. Transposable elements in cancer. *Nat. Rev. Cancer* **17**, 415–424 (2017).
18. Scott, E. C. & Devine, S. E. The Role of Somatic L1 Retrotransposition in Human Cancers. *Viruses* **9**, (2017).
19. Tubio, J. M. C. *et al.* Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, (2014).
20. Rodriguez-Martin, B. *et al.* Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat. Genet.* **52**, 306–319 (2020).
21. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214–220 (2011).
22. Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
23. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
24. Hastings, P. J., Ira, G. & Lupski, J. R. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.* **5**, e1000327 (2009).
25. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016).
26. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**,

112–121 (2020).

27. Hadi, K. *et al.* Novel patterns of complex structural variation revealed across thousands of cancer genome graphs. *bioRxiv* 836296 (2019) doi:10.1101/836296.
28. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
29. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
30. Pleasance, E. D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**, 191–196 (2010).
31. Craig, D. W. *et al.* A somatic reference standard for cancer genome sequencing. *Sci. Rep.* **6**, 24607 (2016).
32. Arora, K. *et al.* Deep whole-genome sequencing of 3 cancer cell lines on 2 sequencing platforms. *Sci. Rep.* **9**, 19123 (2019).
33. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
34. Wala, J. A. *et al.* SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).
35. Cameron, D. L. *et al.* GRIDSS: sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Res.* **27**, 2050–2060 (2017).
36. Kataoka, K. *et al.* Integrated molecular analysis of adult T cell leukemia/lymphoma. *Nat. Genet.* **47**, 1304–1315 (2015).
37. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
38. Zhuang, G. *et al.* Effects of cancer-associated EPHA3 mutations on lung cancer. *J. Natl. Cancer Inst.* **104**, 1182–1197 (2012).
39. Cooke, S. L. *et al.* Processed pseudogenes acquired somatically during cancer

development. *Nat. Commun.* **5**, 3644 (2014).

40. Kazazian, H. H., Jr. Processed pseudogene insertions in somatic cells. *Mob. DNA* **5**, 20 (2014).
41. Ostertag, E. M. & Kazazian, H. H., Jr. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.* **11**, 2059–2065 (2001).
42. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
43. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020).
44. Scott, E. C. *et al.* A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res.* **26**, 745–755 (2016).
45. Simpson, J. T. *et al.* Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).
46. Maciejowski, J., Li, Y., Bosco, N., Campbell, P. J. & de Lange, T. Chromothripsis and Kataegis Induced by Telomere Crisis. *Cell* **163**, 1641–1654 (2015).
47. Maciejowski, J. & de Lange, T. Telomeres in cancer: tumour suppression and genome instability. *Nat. Rev. Mol. Cell Biol.* **18**, 175–186 (2017).
48. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
49. Shiraishi, Y. *et al.* An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res.* **41**, e89 (2013).
50. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
51. Yi, K. & Ju, Y. S. Patterns and mechanisms of structural variations in human cancer. *Exp. Mol. Med.* **50**, 98 (2018).
52. Morisse, P., Lecroq, T. & Lefebvre, A. Long-read error correction: a survey and qualitative

comparison. *bioRxiv* 2020.03.06.977975 (2020) doi:10.1101/2020.03.06.977975.

53. Tang, K.-W., Alaei-Mahabadi, B., Samuelsson, T., Lindh, M. & Larsson, E. The landscape of viral expression and host gene fusion and adaptation in human cancer. *Nat. Commun.* **4**, 2513 (2013).
54. Fujimoto, A. *et al.* Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.* **48**, 500–509 (2016).
55. Aganezov, S. & Raphael, B. J. Reconstruction of clone- and haplotype-specific cancer genome karyotypes from bulk tumor samples. *Cancer Biology* 9431 (2019) doi:10.1101/560839.
56. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio. GN]* (2013).
57. Shiraishi, Y. *et al.* A comprehensive characterization of cis-acting splicing-associated variants in human cancer. *Genome Res.* **28**, 1111–1125 (2018).
58. Zhao, M., Lee, W.-P., Garrison, E. P. & Marth, G. T. SSW library: an SIMD Smith-Waterman C/C++ library for use in genomic applications. *PLoS One* **8**, e82138 (2013).
59. Kuhn, R. M., Haussler, D. & Kent, W. J. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).
60. Untergasser, A. *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012).
61. Andreson, R., Reppo, E., Kaplinski, L. & Remm, M. GENOMEMASKER package for designing unique genomic PCR primers. *BMC Bioinformatics* **7**, 172 (2006).