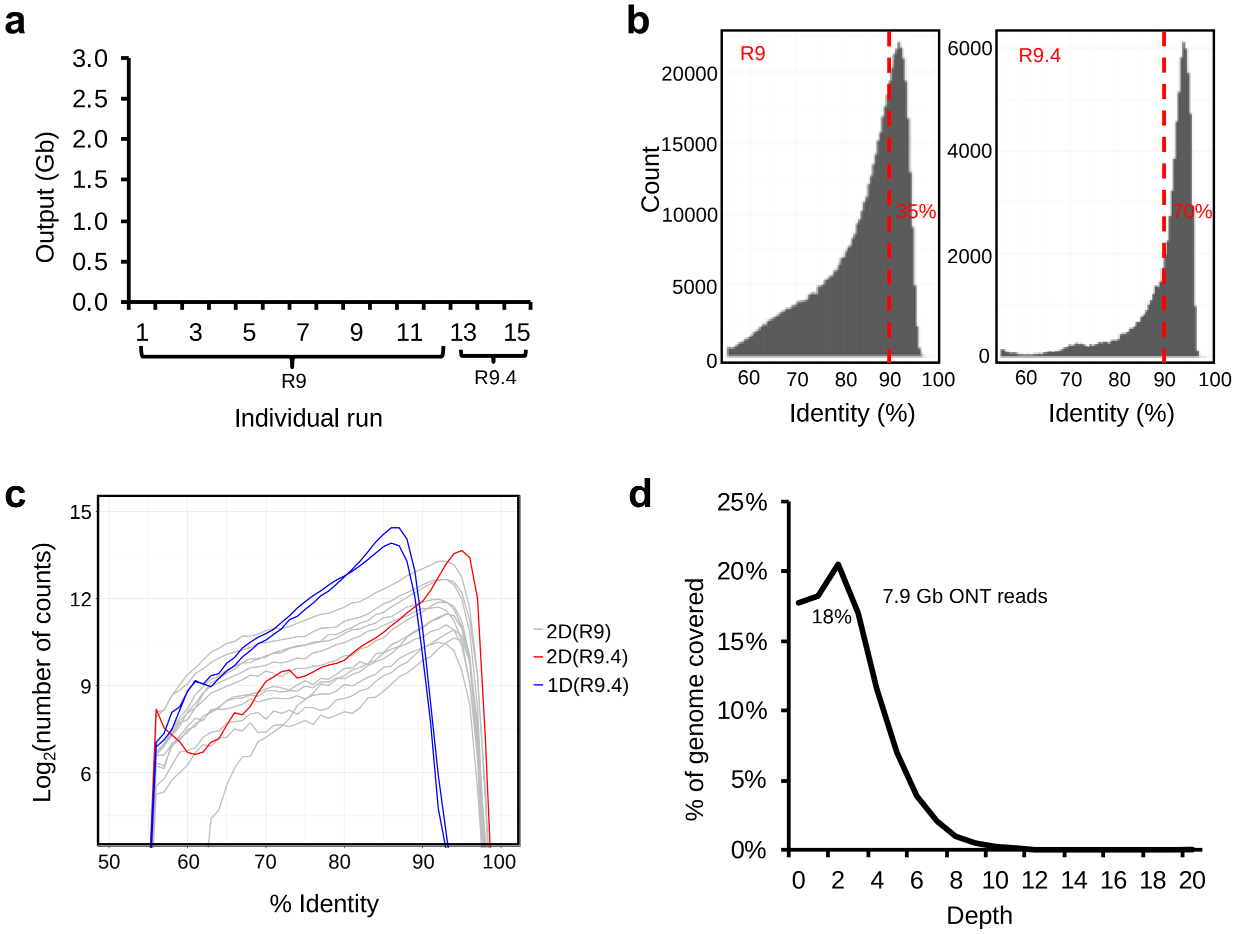
Supplementary Note

Read length, throughput and accuracy of nanopore sequencing.

To assess the utility of the nanopore platform for detecting SVs, we used it to sequence the genome of the breast cancer cell line HCC1187[22], a model for triple-negative breast cancer (TNBC) whose genome harbors extensive structural variation that has been previously characterized at the molecular level by paired-end short-read whole genome sequencing[7,27]. Thus, we can compare the results of our long-read sequencing analysis to this previous data set. Nanopore sequencing libraries were prepared from fragmented high-molecular-weight genomic DNA and subjected for sequencing on MinION instruments (see Online Methods). Sequencing of the single strand templates of the DNA fragments was performed to generate 1D data sets, while sequencing of both the template and complement strands, enabled through their covalent bridging by a hairpin adaptor, was performed to generate 2D data sets. The 2D data sets improve sequence accuracy by aligning the template and complement sequences and resolving any ambiguous base calls between them in the final read output.

With the new version of the sequencing chemistries and the control software, MinION sequencing has resulted in a higher sequencing speed (450 bases/s) and much improved sequencing throughput from R9 to R9.4 chemistry. We achieved a significant improvement in total yield from R9.4 runs, although substantial variations in flow cell quality were observed (**Supplementary Fig. SN1a** and **Supplementary Table 1**). These sequences were aligned onto the reference genome assembly using the genome alignment tool LAST[24,25] and the alignment quality was used to assess the degree of accuracy and identify base-calling errors for the different protocols and chemistries. The alignment ratio between data from R9.4 and R9 chemistry were both comparably high (average > 90%, **Supplemental Table 2**), i.e. majority of the reads from both R9 and R9.4 chemistries aligned to the hg19 reference genome. As expected, higher accuracy was observed from R9.4 data. Among all the aligned reads; 70% of the aligned reads from R9.4, compared to only 35% from R9 chemistry, exhibited > 90% identity to the reference genome (**Supplementary Fig. SN1b**) and the accuracy of the 2D reads was on average higher (94%) than that of the 1D reads (86%) (**Supplementary Fig. SN1c**, **Supplementary Fig. SN2a**).

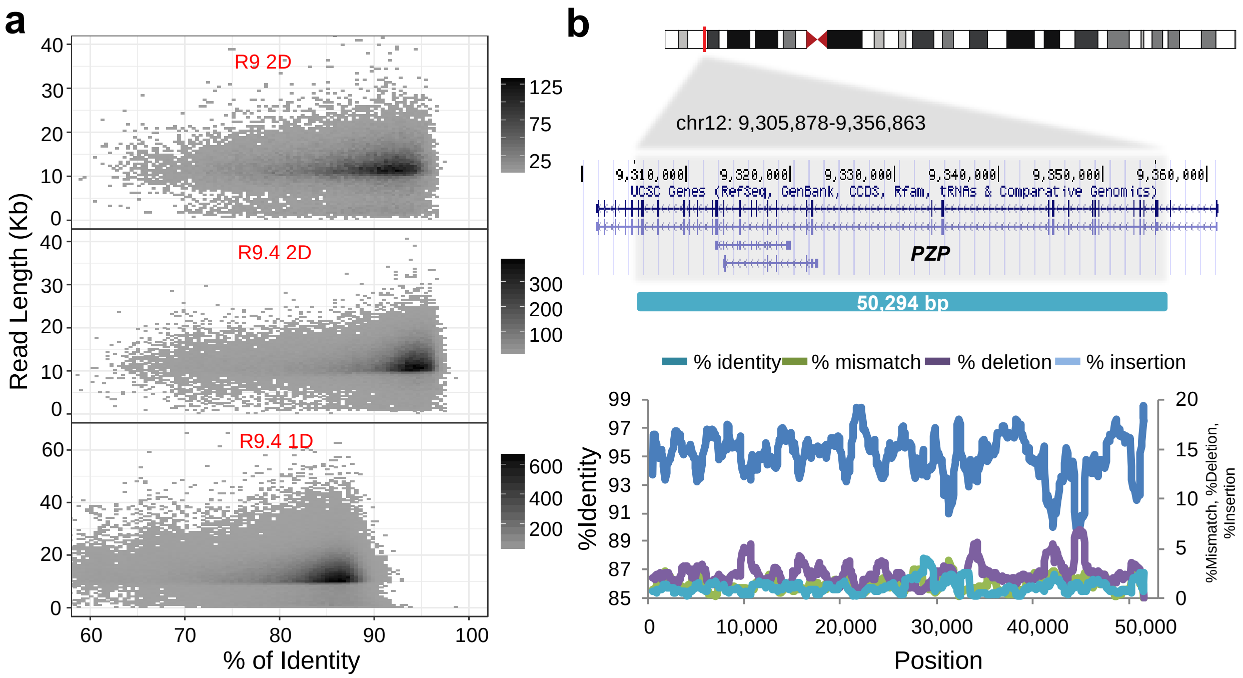
From 7.9 Gb of the aligned sequence data, we obtained 2.5X average genome coverage (based on a haploid genome size) (**Supplementary Fig. SN1d**). One of the longest 2D reads we obtained was 50 Kb (i.e. totaling over 100 Kb for sequencing template and complement strands). This read aligned to the *PZP* gene region (chr12: 9,305,878-9,356,863) at an average of 95% identity (**Supplementary Fig. SN2b**), suggesting that nanopore is capable of generating read length beyond 100 Kb. Compared to paired-end short reads at equivalent sequencing depth, nanopore long reads extended further into repetitive sequence regions of the genome and covered more of the genome (82% vs. 77%) (**Supplementary Fig. SN3a**). For example, a 14.7-Kb nanopore read extended into a region that is rich in short interspersed nuclear elements/long interspersed nuclear elements (SINEs/LINEs) (chr1: 25,732,083-25,747,923), a gap in the coverage of an ultra-high depth (185 Gb, 60X) of short-read data (**Supplementary Fig. SN3b**). To determine if nanopore sequencing has any obvious length bias, we subjected DNA templates of different sizes (3–4 Kb and 12 Kb) to sequencing. We found that the resulting read length distributions matched well with the input DNA fragment sizes (**Supplementary Fig. SN4a,b**) and the total amounts of bases sequenced were equivalent (**Supplementary Table 1**). When compared with the read length distribution from the established PacBio SMRT long-read sequencing platform, using identical preparation of 12 Kb DNA templates, nanopore exhibited less bias in read length (**Supplementary Fig. SN4c**). Taken together, the ability to generate long reads at gigabase output with high accuracy indicates that nanopore sequencing can be adopted to effectively analyze structural variation in cancer genomes.



Supplementary Fig. SN1

Performance of the nanopore long read data.

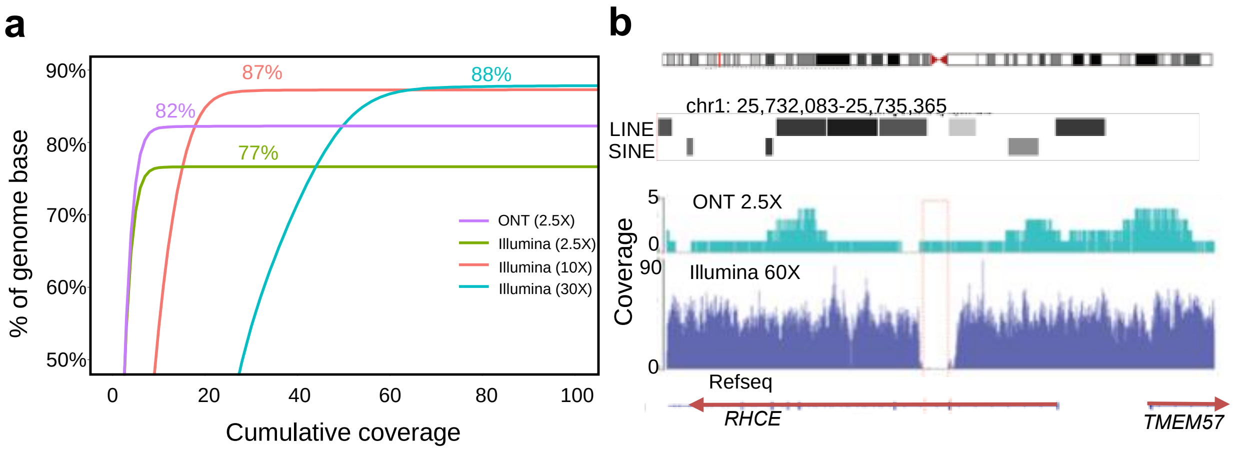
(**a**) Yield of 15 nanopore runs of different chemistry versions. (**b**) Number of read counts by the identity in R9 and R9.4 results. The red vertical dashed line indicates the identity of 90%. (**c**) Accuracy and read counts from runs with different run protocols and pore speeds. Grey, 2D and R9 (250 bases/sec). Red, 2D and R9.4 (250 bases/sec). Blue, 1D and R9.4 (450 bases/sec). (**d**) Genome coverage from the 7.9 Gb of the nanopore data generated in this study.



Supplementary Fig. SN2

Nanopore single-molecule sequencing.

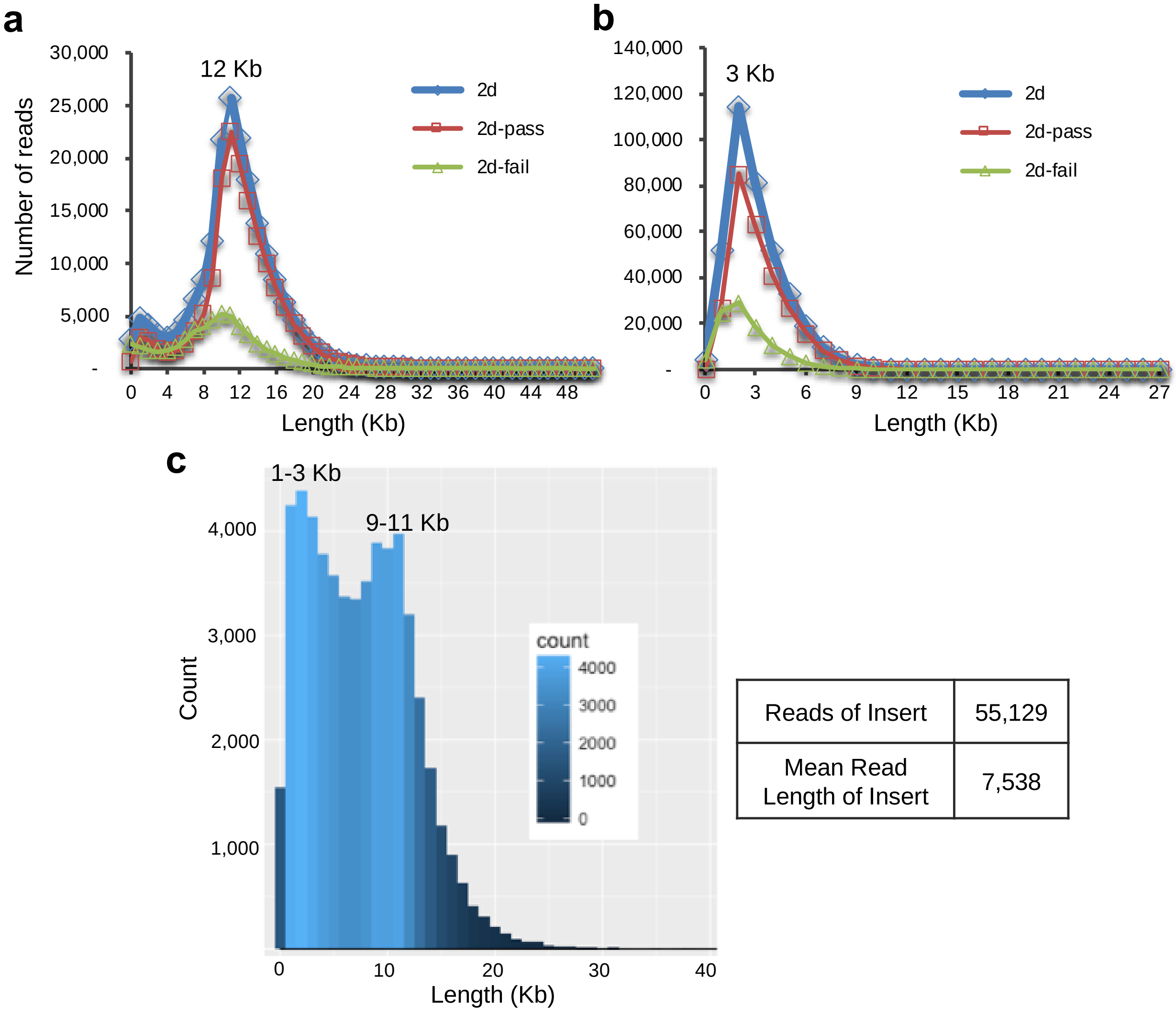
(**a**) Accuracy of nanopore reads from different versions of chemistries and run protocols. The percentages of identity among different read length distribution were shown for 2D reads in R9 (250 bases/sec, independent repeats = 13), R9.4 (250 bases/sec, independent repeats = 1) and 1D read (R9.4, 450 bases/sec, independent repeats = 2). **(b**) Example of a 50-Kb nanopore 2D read aligns to reference human genome (chr12: 9,305,878-9,356,863) with 95% average identity. Independent repeats = 1. The major error profiles are shown at the below.



Supplementary Fig. SN3

Comparison of the genome coverage from long read and short read data.

(**a**) The percentage of the genome covered by short versus long read data of different depth. (**b**) A region (chr1: 25,732,083-25,735,365) covered by a 14.7-Kb nanopore read; which remains a gap even from 185 Gb, 60X of short-read data. This region is rich in SINEs/LINEs.



Reads of Insert	55,129
Mean Read Length of Insert	7,538

Supplementary Fig. SN4

Analysis of read length bias from nanopore and PacBio reads

(**a**) Length distribution of nanopore reads from 12 Kb target template size (data from WTD03, WTD07–WTD12). (**b**) Length distribution of nanopore reads from 3 Kb target template size (data from WTD04–WTD06). (**c**) Length distribution of PacBio reads from the same 12 Kb target template size.