CANCER BIOLOGY & THERAPY
2016, VOL. 17, NO. 3, 246–253
http://dx.doi.org/10.1080/15384047.2016.1139236

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

OPEN ACCESS

Nanopore sequencing detects structural variants in cancer

Alexis L. Norris[a,*], Rachael E. Workman[b,*], Yunfan Fan[b], James R. Eshleman[a], and Winston Timp[b]

[a]Departments of Pathology and Oncology, The Sol Goldman Pancreatic Cancer Research Center, Johns Hopkins School of Medicine, Baltimore, MD, USA; [b]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA

ABSTRACT

Despite advances in sequencing, structural variants (SVs) remain difficult to reliably detect due to the short read length (<300 bp) of 2nd generation sequencing. Not only do the reads (or paired-end reads) need to straddle a breakpoint, but repetitive elements often lead to ambiguities in the alignment of short reads. We propose to use the long-reads (up to 20 kb) possible with 3rd generation sequencing, specifically nanopore sequencing on the MinION. Nanopore sequencing relies on a similar concept to a Coulter counter, reading the DNA sequence from the change in electrical current resulting from a DNA strand being forced through a nanometer-sized pore embedded in a membrane. Though nanopore sequencing currently has a relatively high mismatch rate that precludes base substitution and small frameshift mutation detection, its accuracy is sufficient for SV detection because of its long reads. In fact, long reads in some cases may improve SV detection efficiency.

We have tested nanopore sequencing to detect a series of well-characterized SVs, including large deletions, inversions, and translocations that inactivate the *CDKN2A/p16* and *SMAD4/DPC4* tumor suppressor genes in pancreatic cancer. Using PCR amplicon mixes, we have demonstrated that nanopore sequencing can detect large deletions, translocations and inversions at dilutions as low as 1:100, with as few as 500 reads per sample. Given the speed, small footprint, and low capital cost, nanopore sequencing could become the ideal tool for the low-level detection of cancer-associated SVs needed for molecular relapse, early detection, or therapeutic monitoring.

Abbreviations: SV, Structural Variation; TSG, Tumor Suppressor Gene; PDAC, Pancreatic Ductal Adenocarcinoma; PCR, Polymerase Chain Reaction.

ARTICLE HISTORY
Received 12 August 2015
Revised 8 December 2015
Accepted 1 January 2016

KEYWORDS
3rd generation sequencing; cancer diagnostics; Deletions; DNA sequencing; inversions; nanopore sequencing; next generation sequencing; structural variation; translocations; tumor suppressor gene

Introduction

Structural variants (SVs) are a hallmark of the genomic instability that underlies cancer, and include translocations, large deletions, amplifications, and inversions.[1-3] SVs are often driver alterations, with translocations and amplifications activating oncogenes, and deletions and inversions inactivating tumor suppressor genes (TSGs). *CDKN2A/p16* and *SMAD4/DPC4* are 2 of the most commonly deleted TSGs in human cancer, and complex SVs have been found to underlie approximately half of these deletions in pancreatic ductal adenocarcinoma (PDAC).[4-6]

The sensitive detection of tumor-specific mutations, including both small alterations such as single base substitutions and large alterations such as SVs, of circulating tumor DNA is critical for applications such as molecular relapse,[7] early detection,[8] and possibly therapeutic monitoring of cancer patients.[9] The arrival of 2nd generation sequencing has provided ample opportunity to investigate small alterations, but the large SV alterations remain under-studied because of the difficulty detecting them with the short reads (<300 bp) of 2nd generation sequencing. Not only do the paired-end reads need to straddle a breakpoint, but repetitive elements often lead to ambiguities in the alignment. Given that repetitive regions (including centromeres, telomeres, and other repetitive elements) encompass over half (56%) of the human genome, this is a significant concern when mapping SVs.[10] The long reads (up to 20 kb) generated by 3rd generation DNA sequencing strategies can easily straddle these repetitive regions, allowing for unique alignment.[11]

Until recently, 3rd generation sequencing was limited to PacBio, which requires a high capital investment, a large footprint, and technical expertise. These factors limit the utility of PacBio-based 3rd generation sequencing in clinical testing. The new 3rd generation sequencing platform, the MinION™ (Oxford Nanopore Technologies™), lacks the prohibitive factors of PacBio. The MinION instrument is the size of a large USB stick, with low (~$1k) capital cost and easy operation. Thus, nanopore sequencing on a MinION instrument may prove to be a valuable tool for clinical testing.

Nanopore sequencing, first proposed by Church et al[12] operates via a similar principle to a Coulter counter, using a measurement of the current through a hole in a membrane to characterize sample passing through the hole. In the case of

CONTACT James R. Eshleman jeshlem@jhmi.edu; Winston Timp wtimp@bme.jhu.edu
*These authors equally contributed to this work.

nanopore sequencing, the hole is nanometers in diameter, and the DNA molecule passing through the pore influences the current in a way which is characteristic of the local base sequence. The MinION device consists of 512 independently addressed measurement channels, each with 4 sensor wells. The software controlling the MinION selects the "best" sensor during a process called multiplexing, a process repeated several times throughout a sequencing run. Each sensor well has a semi-synthetic membrane containing a proprietary protein pore molecule. An electric field is applied across the membrane, allowing both current measurement and providing the motive force for driving the negatively charged DNA molecule through the pore. The DNA library is enriched via the tether at the membrane surface, and then diffuses along the membrane. When the DNA leader is within range of the pore, it is captured and driven through up to the motor protein. The DNA is driven through pore primarily via electric field acting on the charged phosphate backbone, with translocation velocity controlled by a proprietary motor protein coupled to the DNA molecule. The pore is large enough only for a single stranded DNA molecule; 5 bases are within the central constriction of the pore at a given time and have a significant influence on the current. After the forward DNA strand has completely run through, the hairpin is run, then the reverse or cDNA strand is also sequenced. The consensus of the top and bottom reads is termed a "2D read" and increases the accuracy of base calling.

Though nanopore sequencing method still has high error rate, it is rapidly improving; in our hands v7 flowcells had an average of 67.4% of the read correct, 24.2% mismatched, 7.5% insertions and 8.3% deletions,[13] but the newer v7.3 flowcells had an average of 86% correct, with 9.7% mismatch, 4.2% insertion and 4.4% deletion – a dramatic improvement. Though the high error rate currently precludes their application to detecting single base substitutions (KRAS codons 12 and 13[14]) and small frameshift mutations,[15,16] the long read length easily generated with nanopore sequencing, i.e. average of 8 kb reads,[13] enables easy detection of SV even in repetitive regions.

In this paper, we demonstrate the ability of nanopore sequencing to detect SVs that inactivate the *p16* and *SMAD4* TSGs in PDAC cancer cell lines. Our set of 10 SVs includes large deletions, translocations, inversions, and the complex combination of a translocation and inversion ("TransFlip"). These SVs were previously defined by SNP microarray and whole genome sequencing (WGS), and confirmed by PCR and Sanger sequencing across the junctions. We show proof-of-principle, using dilutions of PCR products containing these SVs into the corresponding wildtype amplicons and show the ability to detect these SVs at 1:100 dilutions.

Results

Ability to detect simple and complex SVs

To demonstrate the value of long read sequencing to detect SVs, we selected a panel of 10 well-characterized SVs in the genes *CDKN2A/p16* and *SMAD4/DPC4* identified in pancreatic cancer cell lines.[6] These 10 SVs included 2 interstitial deletions, 4 translocations, 4 inversions, and 1 combination of an inversion and translocation ("TransFlip" mutations, Table 1). Wildtype (WT) sequence (intact genomic sequence; no SV) served as a control and one SV (SV01) had a technical replicate (SV07). Using Oxford Nanopore barcodes, libraries for all 12 PCR amplicons were generated

Table 1. Details of Amplicons included in this study.

Amplicon ID	Amplicon Size Without Barcodes (bp)	TSG Deleted	SV Type	SV Left Breakpoint (hg19)	SV Right Breakpoint (hg19)	Expected alignment: Left (hg19)	Expected alignment: Center (hg19)	Expected alignment: Right (hg19)
SV01, SV07	573	*p16*	TRANS	chr9:24,353,014	chr22:36,338,191	chr9:24352894-24353014(+)		chr22:36338191-36338601(+)
SV02	579	*p16*	(WT)	chr9:21,970,115	chr9:21,970,649	chr9: 21970115-21970649(−)		
SV03	562	*p16*	INV+TRANS	chr9:21,083,362	chr9:21,083,521	chr9: 21083139-21083362(+)	chr9: 21083440-21083521(−) (81bp)	chr3:79387683-79387899(−)
SV04	573	*p16*	TRANS	chr10:132,412,941	chr9:27,096,867	chr10: 132412940-132413131(−)		chr9:27096866-27097203(+)
SV05	576	*SMAD4*	ID	chr18:48,570,319	chr18:49,191,882	chr18:48569959-48570319(+)		chr18:49191882-49192052(+)
SV06	561	*p16*	INV	chr9:24,320,470	chr9:24,323,843	chr9: 24323843-24324156(−)		chr9: 24320470-24320672(+)
SV08	559	*p16*	INV	chr9:25,968,399	chr9:25,969,868	chr9: 25968120-25968399(+)		chr9: 25969634-25969868(−)
SV09	581	*p16*	INV	chr9:25,969,504	chr9:25,972,326	chr9: 25969502-25969820(−)		chr9: 25972324-25972543(+)
SV10	584	*p16*	INV+TRANS	chr9:21,326,884	chr7:140,023,555	chr9: 21326735-21326867(+)	chr9: 21326884-21326931(−) (47bp)	chr7: 140023553-140023913(+)
SV11	578	*SMAD4*	TRANS	chr6:124,911,349	chr18:53,465,051	chr6:124911349-124911707(−)		chr18:53465049-53465220(+)
SV12	573	*SMAD4*	ID	chr18:48,434,141	chr18:49,851,882	chr18:48433731-48434141(+)		chr18:49851882-49852004(+)

Tumor Suppressor Gene (TSG), Structural Variant (SV), Translocation (TRANS), Wild-type Control (WT), Inversion (INV), Interstitial Deletion (ID).

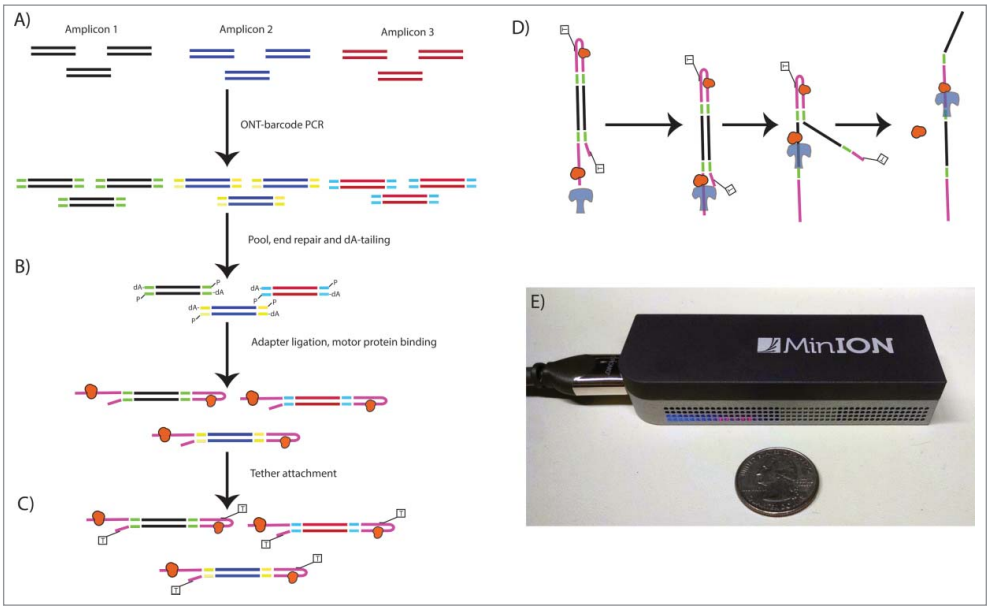

Figure 1. Nanopore Library Prep Workflow. Oxford Nanopore barcodes were incorporated into amplicons by PCR- individually for each SV- then resultant reactions were pooled (A). After NEB End Repair and dA-tailing modules (B), hairpin and leader adapters were ligated on, each containing a motor protein. Only the hairpin protein contained a his-tag, which was used to enrich for molecules containing a leader adapter and his-tag (his-tag selection step not shown). Tether attachment (C) allowed for direct attachment of the molecules to the flow cell membrane. Within the MinION flowcell (D), DNA molecules are pulled through a protein pore (blue), with motor protein (orange) affecting speed of DNA translocation through the pore. One side of the DNA molecule is read, then the hairpin, then the second side. Both reads were aligned to produce a 2D consensus read.

and multiplexed in one flowcell run (Fig. 1). The run produced a total of 3,987 2D reads from 194 of 512 channels, resulting in a 2.5 Mb yield. The average read was 640 bp long, full-length of our PCR products, and an average PHRED score of 11.50 (Fig. 2 A-B, Table 2). Importantly, nanopore reads have no discernible quality dependence with length, compared to the cycle dephasing commonly seen in 2nd generation sequencing.

All SV amplicons (12/12) mapped to their expected region(s) of hg19 (Table 3, Fig. 3), with overall mapping percentage of 99.6% and 79% of aligned reads with correctly matched bases (Table 2). Importantly, the representation of

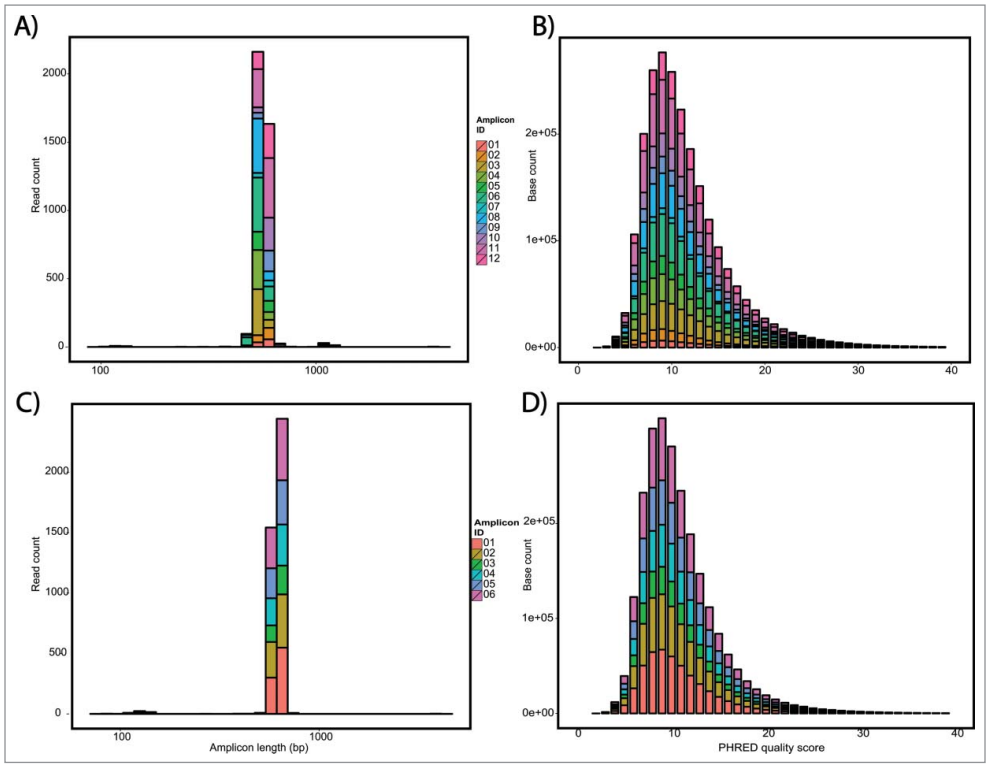

Figure 2. Nanopore sequencing QC data. QC of Flow cell 1 A) length and B) PHRED quality histograms of each of the barcodes as a stacked bar graph. Average length of 570 bp and PHRED score of 11.5. QC of flow cell 2 C) length and D) PHRED quality histograms. Average length of 573 bp and PHRED score of 10.9.

Table 2. Yield and Quality of Exp1, Limited to 2D reads.

Amplicon	Avg. Length (bp)	Yield (bp)	Yield (reads)	Quality (PHRED)	% Match	% Mismatch	% Insertion	% Deletion
SV01	533.07	53,307	100	11.52	81.3%	12.6%	2.3%	6.1%
SV02	582.07	81,490	140	10.98	79.7%	13.2%	2.4%	7.1%
SV03	555.62	228,914	412	11.86	76.3%	15.7%	2.6%	8.0%
SV04	562.60	200,285	356	11.50	75.2%	16.0%	2.5%	8.8%
SV05	596.14	134,131	225	11.31	79.0%	13.6%	2.2%	7.4%
SV06	548.78	311,156	567	11.47	81.2%	12.5%	2.2%	6.3%
SV07	560.40	44,832	80	11.53	80.9%	12.7%	2.2%	6.4%
SV08	547.34	266,554	487	11.33	78.1%	14.0%	1.8%	8.0%
SV09	610.68	123,358	202	11.17	81.4%	12.2%	2.6%	6.3%
SV10	595.92	182,353	306	11.58	78.2%	15.4%	3.0%	6.4%
SV11	578.32	419,283	725	11.55	77.6%	14.8%	2.5%	7.6%
SV12	583.76	225,914	387	12.26	76.4%	14.7%	2.5%	8.9%
Average	571.22	189,298	332	11.50	78.8%	14.0%	2.4%	7.3%

the SV amplicons seems independent of the complexity of their SV: intact genomic sequence (SV02) represented 3.5% of aligned reads, and a complex combination of a deletion, inversion, and translocation (SV10) represented 5.1% of aligned reads (Table 3). The technical replicates (SV01, SV07) had comparable results, as expected. However, some amplicons had a surprisingly low percentage of properly aligned SV structures, specifically of note is SV03 with only 16.5% correctly aligned. In this case only 68 (16.5%) reads had the full alignment of left, center and right sections. However, 313 reads (76.0%) had the left and right alignment, and 12 further reads (2.9%) had left and center alignment.

Ability to detect low frequency SVs

We next wanted to determine the sensitivity of nanopore based SV detection to low frequency or rare events, to simulate a clinical scenario. To this end, we performed a 1:100 dilution of 6 SV amplicons in a background of intact *p16* genomic sequence (SV02, Table 4). These 6 Amplicons included 2 simple interstitial deletions, 2 translocations, 1 inversion, and 1 complex combination of an inversion and translocation. The run produced a total of 4,058 2D reads from 270 of 512 channels, for a total yield of 2.6 Mb, with an average read length of 650 bp and an average PHRED score of 10.9 (Fig. 2C-D, Table 5). All 6 SV barcodes were represented in the alignment (range 9–21%) and aligned to the expected regions of hg19 (Table 4). Remarkably, even with only 378 2D reads in the case of SV04, the SV was detected, with 10 of the 378 reads supporting a chromosome 9-chromosome 10 translocation.

SV breakpoint location detection

To determine the accuracy of breakpoint location detection with this new sequencing methodology, we first employed LUMPY,[17] an established tool for breakpoint detection for both discordant paired end short-read sequencing and long, split read alignments. Using the alignment files generated from BWA above, we extracted the split reads and fed the resulting BAM file into LUMPY. The results are included in Tables 3 and 4.

For some of our samples, LUMPY detected the correct breakpoint, and only one breakpoint, (SV01), or detected no breakpoint in the WT sample (SV02), but in general the breakpoints it detected, though correct in type, lacked precision. In the duplicate sample of SV01, SV07, the same correct breakpoint was detected. Not as many pieces of evidence (as decided by LUMPY) support this breakpoint as when simply examining coverage, because LUMPY has strict map quality filters which remove some of the reads from consideration. In many cases LUMPY detects many breakpoints at slightly shifted conditions – to a max of 8 breakpoints detected in SV03. The breakpoint

Table 3. All SVs are detected by Nanopore multiplex (1:12) experiment [Exp1].

Amplicon ID	SV Type	2D reads	% total reads per barcode	2D reads aligned to hg19 (%)	Reads properly aligned* (%)	Off-target Reads	Lumpy break-points	Top Lumpy Breakpoint
SV01	TRANS	100	2.5%	91 (91.0%)	77 (77.0%)	6 (6.0%)	1	chr9:24353014/chr22:36338191 (58)
SV02	n/a (WT)	140	3.5%	139 (99.3%)	115 (82.1%)	24 (17.1%)	0	None
SV03	INV+TRANS	412	10.3%	412 (100.0%)	68 (16.5%)	1 (0.2%)	8	chr3:79387939/chr9:21083384 (132)
SV04	TRANS	356	8.9%	356 (100.0%)	303 (85.1%)	6 (1.7%)	3	chr9:27096843/chr10:132412942 (183)
SV05	ID	225	5.6%	224 (99.6%)	198 (88.0%)	7 (3.1%)	2	chr18:48570319 (154)
SV06	INV	567	14.2%	567 (100.0%)	549 (96.8%)	7 (1.2%)	4	chr9:24320456/chr9:24323864 (120)
SV07	TRANS	80	2.0%	78 (97.5%)	70 (87.5%)	0 (0.0%)	2	chr9:24353014/chr22:36338191 (52)
SV08	INV	487	12.2%	487 (100.0%)	449 (92.2%)	3 (0.6%)	4	chr9:25968397/chr9:25969868 (384)
SV09	INV	202	5.1%	202 (100.0%)	190 (94.1%)	5 (2.5%)	2	chr9:25969501/chr9:25972324 (172)
SV10	INV+TRANS	306	7.7%	301 (98.4%)	254 (83.0%)	1 (0.3%)	5	chr7:140023522/chr9:21326914 (167)
SV11	TRANS	725	18.2%	725 (100.0%)	471 (65.0%)	3 (0.4%)	3	chr6:124911349/chr18:53465049 (362)
SV12	ID	387	9.7%	387 (100.0%)	274 (70.8%)	3 (0.8%)	2	chr18:48434141 (115)
Average		332	8.3%	331 (98.8%)	252 (78.2%)	6 (2.8%)		

*To be considered properly aligned, a read must align to all expected regions (eg. Left sequence, Center sequence, and Right sequence, from Table 1).

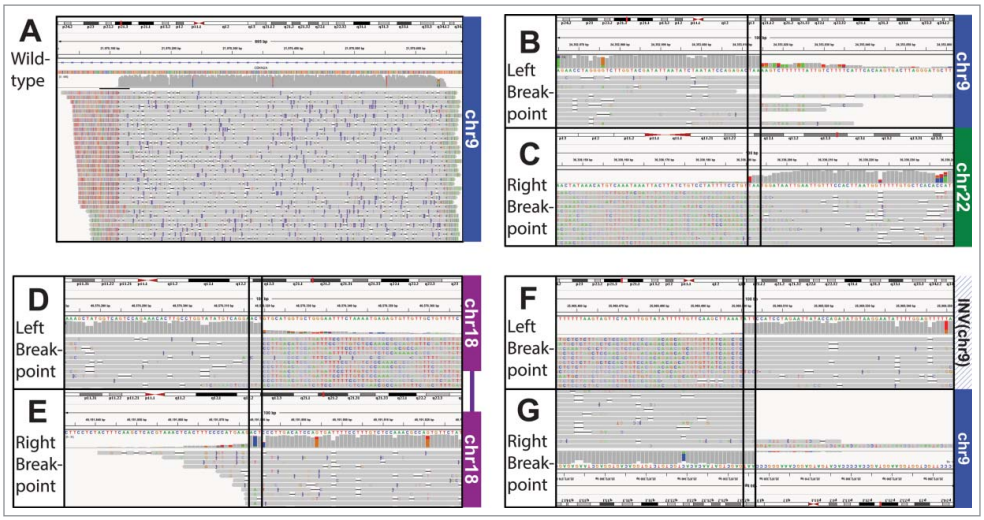

Figure 3. IGV screenshot alignment of WT (SV02). B-C) IGV Screenshot of Translocation (SV01) alignment. B) Shows the alignment to the area in chr9 and C) the alignment to the area in chr22. Note the erroneous extension of the read past the breakpoint in the bottom left. D-E) IGV Screenshot of Interstitial Deletion (SV05) alignment. The plot shows the alignment to the area upstream D) and downstream E) of the deletion in chr18. Note the erroneous extension of the read past the breakpoint in the top right. F-G) IGV Screenshot of Inversion (SV09) alignment. The plot shows the alignment to the inverted area F) and G) the area downstream of the inversion. We have flipped G) to show how the 2 parts align. Note the erroneous extension of the read past the breakpoint in the top left.

with the plurality of reads accepted by LUMPY is represented in Tables 3 and 4.

We examined the alignment more carefully to determine the cause of these artifacts in breakpoint location detection. Figs. 3 B-G give a hint as to the problem – a careful examination notes that the reads frequently align past the breakpoint, but most of the bases are mismatched in these locations. With BWA bound by the promiscuous settings used for nanopore sequencing alignment, it continues the alignment past the breakpoint. We summarized these findings in Table 6. Though in many cases the alignment termini are set to the breakpoints correctly, for SV03 in particular this happens a minority of the time. When the alignment slips past the downstream boundary of the left fragment, there is no longer sufficient sequence for the center fragment to align.

Discussion

The accurate and timely detection of tumor-associated alterations, including SVs, is important for patient management, from early detection to monitoring for molecular relapse, as well as determining or predicting chemoresponse. Detection of all tumor-associated alterations is complicated by the low tumor cellularity often present in tumor samples and biopsies due to contaminating normal cells. Tumor-associated SVs are additionally complicated when they occur in repetitive regions, which account for over half of the human genome. The ability of long-read 3rd generation sequencing methods, such as nanopore, to read through repetitive regions could make it an ideal tool for detecting tumor-associated SVs.

This work serves as proof-of-principle, showing the ability of nanopore sequencing to correctly and reliably detect SVs with only hundreds, instead of millions of reads. Furthermore, we have demonstrated the feasibility of the MinION for the detection of well-characterized patient-specific SV rearrangements using *in vitro* mixtures of PCR amplicons at 1:100 dilutions in wildtype sequence. The 4 types of SVs assessed in this study include simple interstitial deletions, translocations, inversions, and the complex combination of a translocation and an inversion ("TransFlip"). This is accomplished despite the error rate (at the single-base level) of this emerging technology, because the read length is relatively long, the human genome is known, and this level of accuracy is sufficient to correctly map hundreds to thousands of bases, even if they contain multiple point

Table 4. Results of low frequency serial dilutions of SVs 1:100 into wildtype [Exp2].

Amplicon ID	SV Type	Dilution into WT	# 2D reads	% per barcode	Reads aligned to hg19	Aligned reads mapped to WT	Aligned reads mapped to SV*	# Off-target Reads	Lumpy break-points	Top Lumpy Breakpoint
SV01	TRANS	1:100	867	21.37%	851 (98.2%)	838 (96.7%)	11 (1.3%)	3 (0.3%)	1	chr9:24353014/chr22:36338197 (7)
SV03	INV+ TRANS	1:100	760	18.73%	741 (97.5%)	685 (90.1%)	7 (0.9%)	50 (6.6%)	3	chr3:79387933/chr9:21083384 (13)
SV04	TRANS	1:100	378	9.31%	377 (99.7%)	367 (97.1%)	10 (2.6%)	1 (0.3%)	1	chr9:27096848/chr10:132412942 (8)
SV05	ID	1:100	577	14.22%	571 (99.0%)	538 (93.2%)	31 (5.4%)	3 (0.5%)	1	chr18:48570319 (25)
SV09	INV	1:100	621	15.30%	617 (99.4%)	601 (96.8%)	16 (2.6%)	1 (0.2%)	1	chr9:25969504/chr9:25972324 (14)
SV12	ID	1:100	855	21.07%	849 (99.3%)	810 (94.7%)	26 (3.0%)	14 (1.6%)	1	chr18:48434141 (11)
Mean			676	16.67%	668 (98.8%)	640 (94.8%)	17 (2.6%)	12 (1.6%)		

*To be considered properly aligned, a read must align to all expected regions (eg. Left sequence, Center sequence, and Right sequence, from Table 1).

Table 5. Yield and Quality of Experiment 2, Limited to 2D reads.

Amplicon	Avg Length (bp)	Yield (bp)	Yield (reads)	Quality (PHRED)	% Match	% Mismatch	% Insertion	% Deletion
SV01	570.85	494,925	867	10.79	80.2%	13.0%	2.6%	6.8%
SV03	573.37	435,760	760	10.83	79.2%	13.7%	2.8%	7.1%
SV04	575.37	217,491	378	10.90	80.5%	12.7%	2.6%	6.7%
SV05	571.22	329,593	577	10.85	80.0%	13.1%	2.8%	6.8%
SV09	572.65	355,617	621	10.91	80.6%	12.8%	2.7%	6.6%
SV12	573.27	490,146	855	10.93	80.1%	13.0%	2.6%	6.9%
Average	572.79	387,255	676	10.87	80.1%	13.1%	2.7%	6.8%

mutation errors. Precision of breakpoint location is still limited, but this can be solved bioinformatically, via alignment parameter optimization or breakpoint detection tailored to the idiosyncrasies of nanopore sequencing data.

The primary advantages for nanopore sequencing over 2nd generation sequencing methods for detection of SV are its (1) ability to sequence through repetitive regions, (2) speed, and (3) low cost and availability.

First, the long-read nature (up to 20 kb) of nanopore sequencing allows reading through repetitive regions. Even with long mate-pair sequencing at deep coverage, 2nd generation methods' short-read sequences prohibit accurate and efficient mapping of repetitive regions, which often house SVs. Previous work has demonstrated that long-read sequencing on its own has been able to detect novel SVs; 10% of the ~30,000 SVs detecting in a single individuals somatic genome were detected only via long-read PacBio sequencing.[18]

Second, the speed of real-time nanopore sequencing offers results in minutes, allowing for rapid diagnosis and treatment. To have 99% confidence of a variant at 1:100 in the sample, we need ~450X coverage over the region of interest.[19] In nanopore, each of the 512 channels can generate a read separately, with each read completed and analyzable in minutes. From the 2 sequencing runs we performed in this paper, generating 450 reads required 15 minutes and 33 minutes respectively. In contrast, 2nd generation sequencing generates millions of reads simultaneously, but the reads are only complete after hours or days, meaning that any analysis has to wait for completion. For example, the fastest Illumina 2nd generation sequencing run requires 4 hours to obtain 1 × 36 bp 12 M reads (MiSeq v2); and such short reads would prove challenging for SV detection. In both cases we are omitting the library preparation time, but these times are largely equivalent.

Third, the low cost (approximately $1k per device) and small size (USB stick) of the MinION nanopore sequencing instrument offer accessibility to testing in nearly any setting. In contrast, the instrumentation for 2nd generation methods require a substantial upfront investment (>$100k) and sufficient lab space for their large footprint, which are prohibitive to many research and clinical labs.

There are currently 2 limitations that restrict the utility of nanopore sequencing: (1) a relatively high mismatch and indel error rate and (2) limited yield (on the scale of Megabases or Gigabases), but both of these factors continue to improve. In our hands, error rate per read decreased from 32% to 14% over a 6 month period. Better tools for corrected basecalling,[20] alignment[21,22] and assembly tools[23] have already been generated by the community. While still insufficient for whole-genome sequencing, the MinION yield has been increasing, and yields per flow cell by other

Table 6. Alignment termini position error.

Amplicon ID	Overlapping alignments	Correct Upstream Termini (%)	Mean Upstream Error ± SD	Correct Downstream Termini (%)	Mean Downstream Error ± SD
SV01, SV07 (L)	77	3 (3.9%)	1.8 ± 5.4	53 (68.8%)	5.3 ± 14.5
SV01, SV07 (R)	85	4 (4.7%)	1.1 ± 5.5	1 (1.2%)	−1.3 ± 17.3
SV02	116	7 (6.0%)	0.1 ± 5.2	13 (11.2%)	−4.3 ± 10.1
SV03 (L)	407	6 (1.5%)	4.9 ± 8.3	16 (3.9%)	20.8 ± 13.7
SV03 (C)	83	1 (1.2%)	10.8 ± 18.4	58 (69.9%)	4.7 ± 18.7
SV03 (R)	391	51 (13.0%)	1.8 ± 17.2	3 (0.8%)	39.1 ± 24.2
SV04 (L)	317	23 (7.3%)	11.0 ± 19.3	1 (0.3%)	7.3 ± 7.0
SV04 (R)	349	2 (0.6%)	19.0 ± 23.3	196 (56.2%)	−5.7 ± 14.2
SV05 (L)	225	4 (1.8%)	5.7 ± 8.5	107 (47.6%)	2.9 ± 14.2
SV05 (R)	206	1 (0.5%)	7.3 ± 9.8	62 (30.1%)	−3.8 ± 9.7
SV06 (L)	565	0 (0.0%)	−20.9 ± 5.2	229 (40.5%)	−1.0 ± 4.5
SV06 (R)	555	4 (0.7%)	5.6 ± 13.6	297 (53.5%)	−1.6 ± 7.8
SV08 (L)	70	6 (8.6%)	2.7 ± 5.4	45 (64.3%)	2.9 ± 14.7
SV08 (R)	78	3 (3.8%)	4.4 ± 14.3	1 (1.3%)	−1.1 ± 7.9
SV09 (L)	476	125 (26.3%)	4.4 ± 9.9	116 (24.4%)	−2.1 ± 6.6
SV09 (R)	464	37 (8.0%)	0.0 ± 11.0	276 (59.5%)	7.0 ± 28.4
SV10 (L)	271	32 (11.8%)	−0.5 ± 4.7	0 (0.0%)	49.9 ± 19.4
SV10 (C)	260	0 (0.0%)	144.0 ± 26.5	0 (0.0%)	−9.9 ± 14.7
SV10 (R)	302	0 (0.0%)	28.8 ± 24.4	8 (2.6%)	15.7 ± 27.6
SV11 (L)	725	4 (0.6%)	34.7 ± 51.5	91 (12.6%)	−5.7 ± 7.9
SV11 (R)	472	8 (1.7%)	−0.1 ± 8.0	75 (15.9%)	2.4 ± 8.6
SV12 (L)	386	72 (18.7%)	5.2 ± 6.9	151 (39.1%)	4.1 ± 11.2
SV12 (R)	278	11 (4.0%)	−0.6 ± 7.9	6 (2.2%)	9.9 ± 9.9

groups have reached nearly ~200 Mb,[24] with substantially greater improvements (10-fold) in yield expected soon. Additionally, the throughput of nanopore sequencing should increase with the release of the PromethION, GridION, and subsequent systems from Oxford Nanopore.

The capacity of the MinION system is currently sufficient to sequence tumor DNA for SVs provided that a small subset of the genome is first captured. For example, deletions within the *p16/CDKN2A* locus in pancreatic cancer can span up to 10 megabases.[4,25] Given the ability for long reads, it may be the ideal tool for phasing of 2 mutations within the same gene, provided frozen tissue is available. To test for circulating tumor DNA (ctDNA) in plasma additional improvements in throughput will be required to achieve the require 1000–100,000X coverage required for minimal residual disease testing and early detection of solid tumors. This will require the PromethION, GridION or subsequent instrument.

Here we have shown the ability and reliability of nanopore sequencing, a 3rd generation sequencing method, to detect well-characterized SVs, and at low levels that simulate that seen in the clinical testing. Importantly, the SV sequences were represented equally well in the alignments of nanopore sequencing data - from simple (interstitial deletions) to complex (inversions, translocations, and TransFlips) SVs. Further development is needed on bioinformatics tools which can precisely align to and detect breakpoint locations. It will be critically important to demonstrate the ability to detect SVs from cancer:normal cell titrations of genomic DNA, as well as plasma from pre- and post-resection patients. Ongoing studies involve further dilution experiments and detection of novel (unknown) SVs directly from patient samples.

Materials and methods

Identification of SVs

Genomic DNA was extracted from previously described PDAC cancer cell lines using QIAamp DNA mini kit (Qiagen), per manufacturer's instruction.[25] Structural variants associated with *p16* and *SMAD4* deletions were identified by high density SNP microarray and WGS, and confirmed by PCR amplifying across the novel DNA:DNA junction and bidirectional Sanger sequencing.[6] Primers were designed upstream and downstream of 10 *p16* and *SMAD4* deletions associated with different SVs (Table 1), as well as *p16* wildtype sequence, to produce amplicons of 550–600 basepairs.[26] We also included a technical replicate in our design to control for technical variation (SV01 and SV07). Residual nucleotides and oligonucleotides were removed using QIAquick PCR purification kit (Qiagen), per manufacturer's instructions. PCR specificity was verified by gel electrophoresis and quantified by Qubit DNA double-stranded high sensitivity assay (dsDNA HS assay, Life Technologies).

Library preparation

Barcodes were added to the PCR amplicons with Oxford Nanopore primers complementary to the tail sequence with a sample specific barcode (Barcode Developer Kit I) using Long Range PCR kits (NEB) (Fig. 1). This allowed for multiplexing of up to 12 samples on a single flow cell. Barcoded PCR libraries were quantified with Qubit dsDNA HS Assay kit, normalized, and pooled to a final amount of 1 μg. For sequencing, the libraries were end-repaired and dA-tailed using NEB DNA Ultra modules, followed by the ligation of hairpin and Oxford Nanopore-specific leader adapters using Genomic DNA Sequencing Kit MAP-004 (Oxford Nanopore). A motor protein was bound to both the leader and hairpin adapters, and serves to ratchet each molecule through the nanopore one base at a time. Enrichment for molecules containing hairpin adapters and bound motor protein was performed using His-Tag Dynabeads® (Life Technologies).

Flowcell runs

For the first flowcell run, the 12 Amplicons were multiplexed together at equal concentrations. For the second flowcell run, *in vitro* dilutions were performed to assess the ability to detect low-level SVs, to simulate clinical samples. Specifically, the following *p16*- and *SMAD4*-associated SVs were diluted at 1:100 in wildtype *p16* sequence (SV02): an inversion (SV09), an inversion with translocation (SV03), translocations (SV01 and SV04), and simple interstitial deletions (SV05 and SV12). These dilutions were barcoded and multiplexed together at equal concentrations.

Oxford nanopore MinION™ sequencing and basecalling

The MinION Flow Cell (R7.3 chemistry) was run for 48 hours on MinKNOW software (v0.49.3.7), producing thousands of fast5 files, each file corresponding to a molecule read by the sequencer. Cloud-based basecalling software (Metrichor™, v2.29.1, Oxford Nanopore) was used to convert electrical event data from MinKNOW into basecalled files. Three basecalled reads were produced: a "1D template" and "1D complement," and "2D read." The 2D read is the consensus sequence between the template and complement reads, and a basic quality filter is applied to keep only 2D reads with a ratio of template bases to complement bases between 0.5 and 2.

Nanopore basecalling is performed by Metrichor using a hidden Markov model, similar to the process described in a simulated data set previously.[27] Briefly, each pentamer generates a specific current which, although difficult to distinguish uniquely, combined with the controlled translocation rate, allows for basecalling the best full sequence.

Alignment and SV calling of reads

Using only 2D nanopore reads which passed the quality filter, we de-mulitplexed and extracted fastq data with custom code in python (https://github.com/timp0/timp_nanoporesv). Data is available at the SRA archive with accession number SRP069199. We then aligned the nanopore long reads against the hg19 reference genome using BWA-MEM, with the –x ont2d option set for nanopore specific alignment parameters.[22] A custom python script to extract split read alignments and calculate error in alignment location is

also included in the online git repository (https://github.com/timp0/timp_nanoporesv).

Disclosure of potential conflicts of interest

Dr. Timp holds 2 patents (US2011/0226623 A1 and US2012/0040343 A1) which have been licensed by Oxford Nanopore Technologies.

Acknowledgments

We thank Oxford Nanopore for outstanding technical support. We thank Aaron Quinlan and Ryan Layer for helpful discussions.

References

1. Nowell PC, Hungerford DA. Chromosome studies on normal and leukemic human leukocytes. J Natl Cancer Inst 1960; 25:85-109; PMID:14427847
2. Dryja TP, Rapaport JM, Joyce JM, Petersen RA. Molecular detection of deletions involving band q14 of chromosome 13 in retinoblastomas. Proc Natl Acad Sci U S A 1986; 83:7391-4; PMID:2876425; http://dx.doi.org/10.1073/pnas.83.19.7391
3. Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, McGuire WL. Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene. Science 1987; 235:177-82; PMID:3798106; http://dx.doi.org/10.1126/science.3798106
4. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 2008; 321:1801-6; PMID:18772397; http://dx.doi.org/10.1126/science.1164368
5. Biankin AV, Waddell N, Kassahn KS, Gingras MC, Muthuswamy LB, Johns AL, Miller DK, Wilson PJ, Patch AM, Wu J, et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature 2012; 491:399-405; PMID:23103869; http://dx.doi.org/10.1038/nature11547
6. Norris AL, Kamiyama H, Makohon-Moore A, Pallavajjala A, Morsberger LA, Lee K, Batista D, Iacobuzio-Donahue CA, Lin MT, Klein AP, et al. TransFlip mutations produce deletions in pancreatic cancer. Genes, Chromosomes Cancer 2015; (54)8:472-481; PMID:26031834; http://dx.doi.org/10.1002/gcc.22258
7. Diehl F, Schmidt K, Choti MA, Romans K, Goodman S, Li M, Thornton K, Agrawal N, Sokoll L, Szabo SA, et al. Circulating mutant DNA to assess tumor dynamics. Nat Med 2008; 14:985-90; PMID:18670422; http://dx.doi.org/10.1038/nm.1789
8. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, Bartlett BR, Wang H, Luber B, Alani RM, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med 2014; 6:224ra24; PMID:24553385; http://dx.doi.org/10.1126/scitranslmed.3007094
9. Tie J, Kinde I, Wang Y, Wong HL, Roebert J, Christie M, Tacey M, Wong R, Singh M, Karapetis CS, et al. Circulating Tumor DNA as an Early Marker of Therapeutic Response in Patients with Metastatic Colorectal Cancer. Ann Oncol 2015; PMID:25851626; http://dx.doi.org/10.1093/annonc/mdv177
10. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. http://www.repeatmasker.org: 2013.
11. Ritz A, Bashir A, Sindi S, Hsu D, Hajirasouliha I, Raphael BJ. Characterization of structural variants with single molecule and hybrid sequencing approaches. Bioinformatics 2014; 30:3458-66; PMID:25355789; http://dx.doi.org/10.1093/bioinformatics/btu714
12. Church G, Deamer DW, Branton D, Baldarelli R, Kasianowicz J, USPTO. Characterization of individual polymer molecules based on monomer-interface interactions. 1995;
13. Timp W, Nice AM, Nelson EM, Kurz V, McKelvey K, Timp G. Think Small: Nanopores for Sensing and Synthesis. IEEE Access 2014; 2:1396-408; http://dx.doi.org/10.1109/ACCESS.2014.2369506
14. Tsiatis AC, Norris-Kirby A, Rich RG, Hafez MJ, Gocke CD, Eshleman JR, Murphy KM. Comparison of Sanger sequencing, pyrosequencing, and melting curve analysis for the detection of KRAS mutations: diagnostic and clinical implications. J Mol Diagn 2010; 12:425-32; PMID:20431034; http://dx.doi.org/10.2353/jmoldx.2010.090188
15. Eshleman JR, Markowitz SD, Donover PS, Lang EZ, Lutterbaugh JD, Li GM, Longley M, Modrich P, Veigl ML, Sedwick WD. Diverse hypermutability of multiple expressed sequence motifs present in a cancer with microsatellite instability. Oncogene 1996; 12:1425-32; PMID:8622858
16. Markowitz S, Wang J, Myeroff L, Parsons R, Sun L, Lutterbaugh J, Fan RS, Zborowska E, Kinzler KW, Vogelstein B, et al. Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. Science 1995; 268:1336-8; PMID:7761852; http://dx.doi.org/10.1126/science.7761852
17. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Gen Biol 2014; 15:R84; PMID:24970577; http://dx.doi.org/10.1186/gb-2014-15-6-r84
18. English AC, Salerno WJ, Hampton OA, Gonzaga-Jauregui C, Ambreth S, Ritter DI, Beck CR, Davis CF, Dahdouli M, Ma S, et al. Assessing structural variation in a personal genome-towards a human reference diploid genome. BMC Gen 2015; 16:286; PMID:25886820; http://dx.doi.org/10.1186/s12864-015-1479-3
19. Lin M-T, Mosier SL, Thiess M, Beierl KF, Debeljak M, Tseng L-H, Chen G, Yegnasubramanian S, Ho H, Cope L, et al. Clinical Validation of KRAS, BRAF, and EGFR Mutation Detection Using Next-Generation Sequencing. Am J Clin Pathol 2014; 141:856-66; PMID:24838331; http://dx.doi.org/10.1309/AJCPMWGWGO34EGOD
20. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Meth 2015; 12:733-5; PMID:26076426; http://dx.doi.org/10.1038/nmeth.3444
21. Sovic I, Sikic M, Wilm A, Fenlon SN, Chen S, Nagarajan N. Fast and sensitive mapping of error-prone nanopore sequencing reads with GraphMap. bioRxiv 2015:020719.
22. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997 2013;
23. Goodwin S, Gurtowski J, Ethe-Sayers S, Deshpande P, Schatz M, McCombie WR. Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. Genome Res 2015; 25:1750–6; PMID: 26447147; http://dx.doi.org/10.1101/gr.191395.115
24. Ip CLC, Loose M, Tyson JR, de Cesare M, Brown BL, Jain M, Leggett RM, Eccles DA, Zalunin V, Urban JM, et al. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Research [Internet] 2015 [cited 2015 Dec 8]; Available from: http://f1000research.com/articles/4-1075/v1
25. Norris AL, Roberts NJ, Jones S, Wheelan SJ, Papadopoulos N, Vogelstein B, Kinzler KW, Hruban RH, Klein AP, Eshleman JR. Familial and sporadic pancreatic cancer share the same molecular pathogenesis. Fam Cancer 2015; 14:95-103; PMID:25240578; http://dx.doi.org/10.1007/s10689-014-9755-y
26. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. Primer3—new capabilities and interfaces. Nucl Acid Res 2012; 40:e115-e115; PMID:22730293; http://dx.doi.org/10.1093/nar/gks596
27. Timp W, Comer J, Aksimentiev A. DNA base-calling from a nanopore using a Viterbi algorithm. Biophys J 2012; 102:L37-9; PMID:22677395; http://dx.doi.org/10.1016/j.bpj.2012.04.009