Oxford NANOPORE Technologies

WHITE PAPER

Nanopore sequencing

The application and advantages of long-read nanopore sequencing to structural variation analysis

Contents

Introduction

Genomic variation comes in many forms, from single point mutations through to much larger structural variations. To date, the majority of genetic research has focused on single nucleotide polymorphisms (SNPs); however, it is now known that structural variation accounts for a greater number of variable bases than SNPs[1,2]. Perhaps unsurprisingly then,

...structural variants have, to a greater extent than SNPs, been shown to be responsible for human evolution, genetic diversity between individuals and a rapidly increasing number of diseases or susceptibility to diseases[3].

While advances in sequencing technology over the last 40 years have increased our understanding of the genome considerably, its application to structural variation analysis has been limited by the short-read technology commonly used. This review is designed to provide an overview of structural variation and show how researchers are utilising the advantages of nanopore long-read sequencing to further enhance our understanding of genomic diversity and disease.

1

What is structural variation?

The term structural variation (SV) covers a range of genetic alterations, including copy number variation (CNV), duplications, translocations and inversions (Figure 1).

Traditionally, structural variants were defined as spanning over 1,000 base pairs; however, with the advent of higher resolution genome analysis techniques, some researchers have now revised this definition to over 50 base pairs[4]. Such structural variants have been associated with a number of diseases (including autism[5], obesity[6], schizophrenia[7] and cancer[8]) and, for this reason, have become increasingly important areas for research.

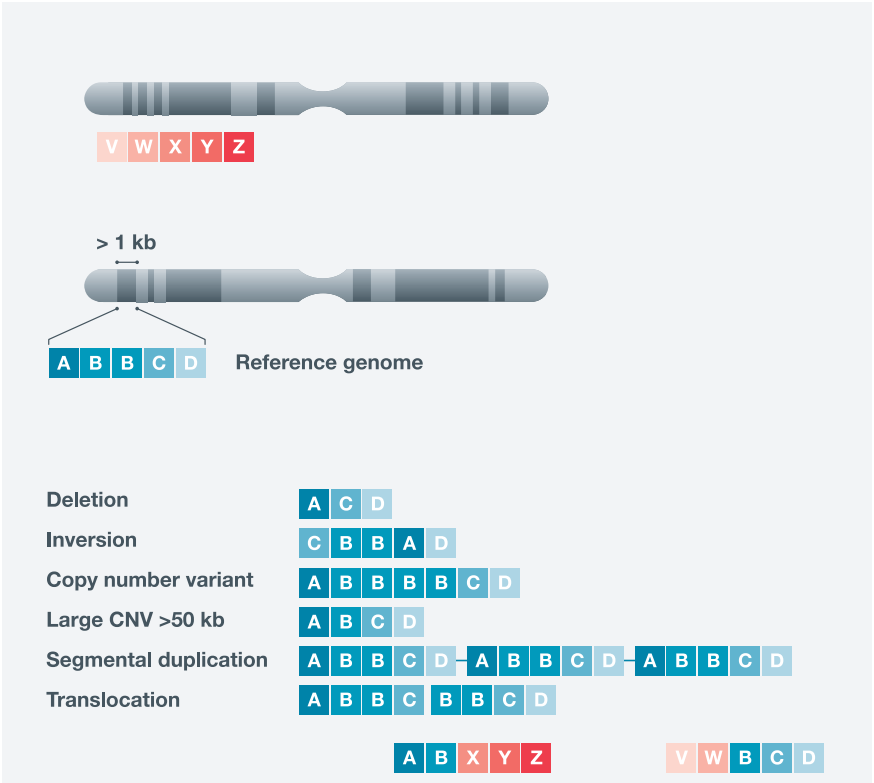


Figure 1
Schematic of common structural variants.

2

Advantages of long-read sequencing for structural variation analysis

Most high-throughput sequencing technologies require users to sequence short lengths of DNA (typically 150-300bp). Clearly, these short reads are less likely to span larger structural variation, making analysis particularly challenging, especially in repetitive genomic regions[9]. Given that repetitive regions (such as centromeres, telomeres and other repetitive elements) encompass over half of the human genome[10], this is a significant concern when mapping structural variants using short-read technology. As a consequence, most existing genome assemblies, which have been created using short reads, exhibit numerous gaps, corresponding to repetitive regions and structural variation that cannot be resolved[9,10,11,12,13].

These challenges are now being addressed by the emergence of long-read nanopore sequencing, which allows entire DNA fragments to be sequenced, regardless of their length. Users can choose to sequence long fragments of several kilobases to ultra-long reads with lengths currently approaching 1 Mb[14]. Such long reads are able to span most structural variation and repetitive regions simplifying their analysis (Figure 2).

Long-read lengths are more likely to contain the whole structural variant and/or repetitive region providing much simpler analysis and more accurate genome assemblies.

Nanopore sequencing of long and ultra-long fragments makes it possible to span most structural variation and repetitive regions, simplifying their analysis.

Figure 2
A schematic highlighting the advantages of long reads for structural variation analysis. Long-read lengths are more likely to contain the whole structural variant and/or repetitive region providing much simpler analysis and more accurate genome assemblies.

3

Application of long reads to structural variation analysis

Long-read sequencing offers advantages in all applications where accurate analysis of structural variation is required, delivering a more complete view of genetic diversity. While the study of structural variants has typically focused on genetic diseases in eukaryotes, recent studies also demonstrate that genome rearrangements can have a profound impact on prokaryotic genomes[15]. The following section will provide an overview of some of the new, ground-breaking research that has been made possible through the use of nanopore long-read sequencing.

Human health and disease

Numerous studies have associated structural variants with both common and rare human diseases, with phenotypes ranging from cognitive disabilities to predispositions to obesity, cancer and other disorders[1].

The accurate and timely detection of tumour-associated alterations, including structural variants, can be important for effective patient management. Researchers at Johns Hopkins University, USA, set out to assess the potential clinical utility of nanopore sequencing to detect a range of structural variation in tumour samples[16]. The analysis of tumour samples is often complicated by the presence of contaminating normal cells, requiring sensitive detection[16].

Nanopore sequencing using the MinION allowed accurate detection of structural variants, even in repetitive regions, with as few as 500 reads per sample.

Furthermore, results were delivered in minutes rather than the hours or days required for traditional sequencing technology[16]. In summary, the researchers stated: "*Given the speed, small footprint, and low capital cost, nanopore sequencing could become the ideal tool for the low-level detection of cancer-associated structural variants needed for molecular relapse, early detection or therapeutic monitoring*" [16].

Similarly, researchers at the Jackson Lab, USA are using the MinION™ to investigate whole-genome structural variation in cancer samples[17]. Their research in a triple-negative breast cancer (TNBC) cell line identified thousands of genome rearrangements affecting genes of relevance to their study. Long-read nanopore sequencing also allowed complex structural variation patterns to be deciphered that had been overlooked by previous short-read analysis, revealing genome-scale breakpoints at base resolution (Figure 3).

Echoing the findings of the team at Johns Hopkins University, Dr Chia-Lin Wei, Director of Genome Technologies at the Jackson Lab, noted: *"Our data supports the feasibility of using long-read sequencing for structural variation detection in cancer cohorts at real-time to trace clonal amplification, monitor drug sensitivity and the outcome of therapeutic intervention."*

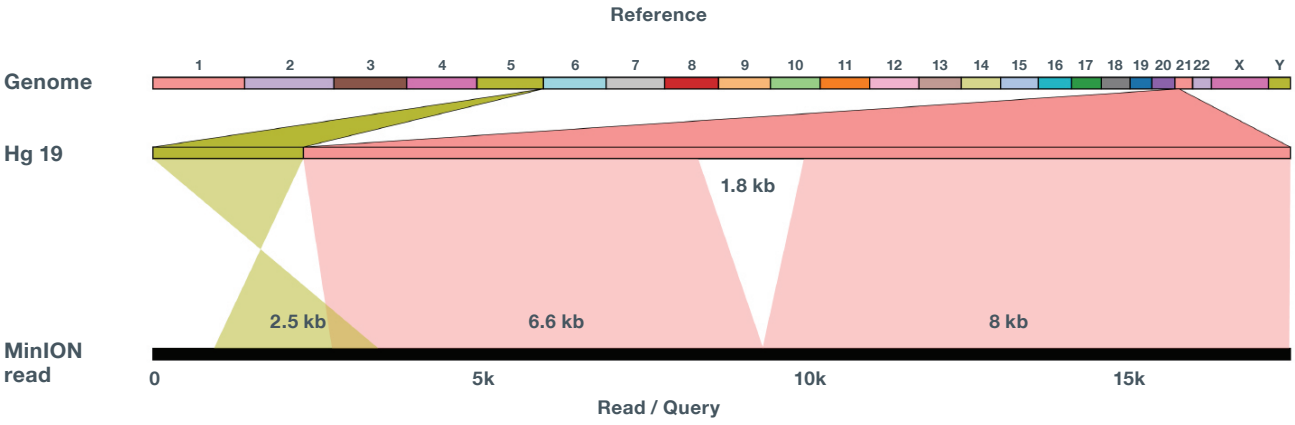


Figure 3
Nanopore sequencing identifies complex structural variation. In this example, a translocation of chromosome 5 (2.5 kb) and 21 (14.6 kb) is detected together with a deletion (1.8 kb) in chromosome 21 of a TNBC cell line when compared against hg19. Data courtesy of Dr Chia-Lin Wei, Jackson Lab, USA.

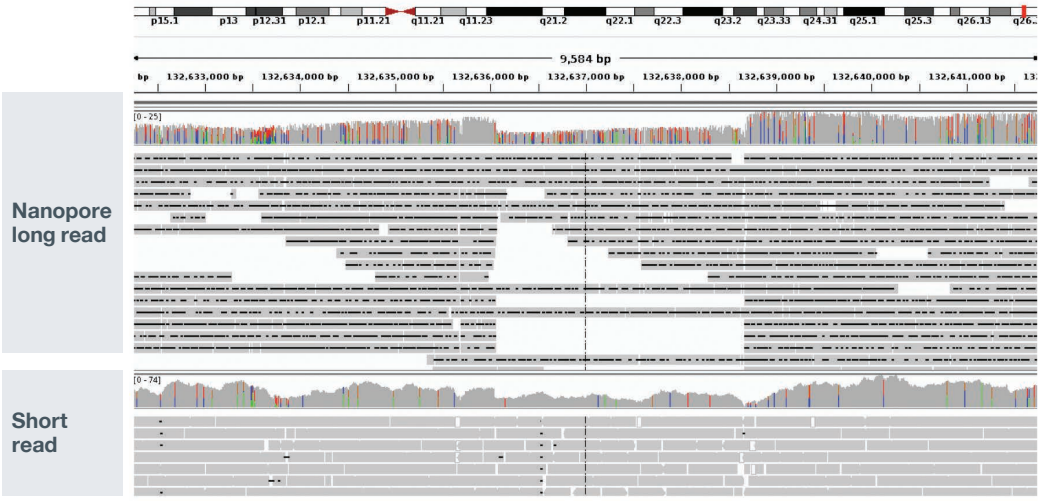


Figure 4
The long-reads generated by nanopore sequencing allow identification of large structural variations. A 2.5 kb heterozygous deletion on chromosome 10 of the well-studied human reference genome, NA12878, was clearly identified using the MinION. This deletion was not evident using short-read sequencing technology. Data courtesy of Michael Simpson, Genomics plc and WTCHG, UK.

Creation of high-quality reference genomes

The limitations of short-read sequencing for analysing structural variation and repetitive regions are well documented[11,18]. The realisation that such regions of the genome are important factors in health and evolution has led to renewed focus on the creation of higher quality reference genomes[12]. As demonstrated by researchers at Genomics plc and the Wellcome Trust Centre for Human Genetics (WTCHG), even the well-characterised human genome has large gaps in its assembly[13]. The team applied long-read sequencing using the MinION to one of the most extensively studied human genomes, NA12878. Their work identified numerous large structural variations that were not readily apparent using short-read sequencing technology (Figure 4)[13]. The data from this study is freely available to download from the WTCHG website[19].

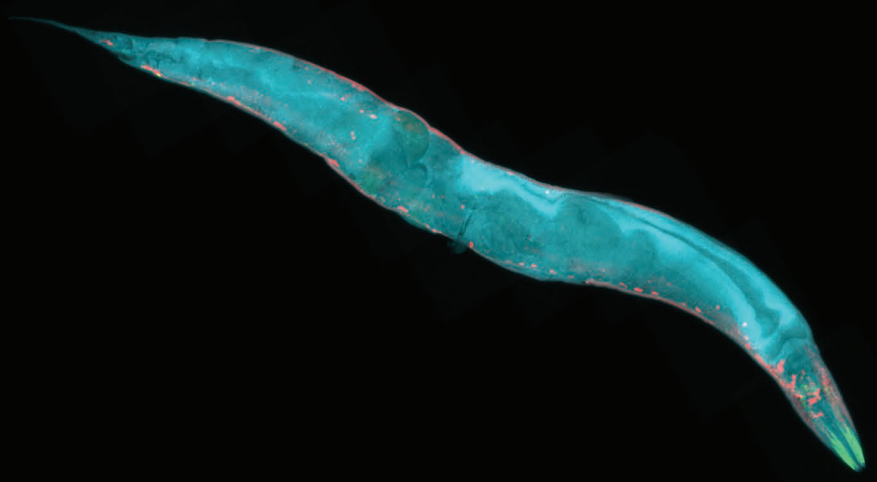
Additional nanopore sequencing data on the NA12878 genome has been made available by the Nanopore-WGS-Consortium and is accessible on www.github.com.

Researchers at the University of British Columbia, Canada have also utilised the MinION to create a *de novo* genome assembly of *Caenorhabditis elegans*[20]. The *C. elegans* genome was the first metazoan genome to be completely sequenced[21] and represents an excellent model genome for assessing new whole-genome sequencing technologies.

Metazoan genomes can prove difficult to assemble due to a high level of repetitive DNA.

For example, approximately 12% of the 100 Mb *C. elegans* genome is derived from transposable elements. Such transposons range in size from 1-3 kb and can confound genomic assemblies made using short-read sequencing technology, resulting in ambiguous mapping positions. The *de novo* assembly of MinION derived data generated a remarkably complete genome with just 145 contigs covering >99% of the reference genome. This can be contrasted with the 38,645 contigs acquired from the traditional short-read sequencing approach.

Interestingly, the short-read sequencing assembly significantly overestimated the genome size, whereas the long-read data delivered a more accurate assembly. This anomaly was also described by Jansen *et al* (2017) when analysing the eel genome using both short- and long-read approaches[22]. To assess the quality of the assembly with respect to repetitive elements, the team compared the number and position of Tc1, a 1.6 kb transposon, to the reference genome. Traditional paired-end reads are not sufficiently long to span the Tc1 transposons and cannot be unambiguously mapped, while all previously identified Tc1 elements were present in the MinION contigs and corresponded to their position in the reference genome.



4

The nanopore advantage

As discussed, long-read sequencing offers considerable advantages over traditional short-read sequencing technologies for the accurate analysis of structural variation; however, there are many additional advantages of nanopore-based long-read technology, including:

- Sensitive detection from limited starting material
- Ultra-long reads
- Fast time to results
- Low cost
- Small footprint

Sensitive detection from limited starting material

In certain applications, for example the analysis of structural variation in cancer samples, the amount of starting material can be limited. Furthermore, the sample may be a heterogeneous mix of "normal" and cancer cells which necessitates sensitive detection to identify low-frequency variants or rare events[16]. Nanopore-based sequencing can be performed with very low starting amounts of DNA.

Using the MinION, whole genome sequencing of bacterial genomes has been successfully achieved using just 10pg of starting material[23].

A number of protocols are available for the MinION, allowing optimised whole genome sequencing for a range of sample types and DNA input amounts.

To assess the sensitivity of nanopore long-read sequencing for structural variation detection, researchers at Johns Hopkins University created a 1:100 dilution of 6 structural variant amplicons in a background of intact p16 genomic sequence[16]. These 6 amplicons included 2 simple interstitial deletions, 2 translocations, 1 inversion, and 1 complex combination of an inversion and translocation. All 6 structural variants could be accurately mapped against a reference genome, demonstrating the high sensitivity of nanopore sequencing.

Ultra-long reads

Nanopore sequencing processes the fragments presented to the pore regardless of their length, allowing researchers to investigate extremely complex or large structural variants. To date, the longest fragment processed using nanopore sequencing is 950 kb[14].

An extreme example of structural variation is demonstrated by the phenomenon of chromothripsis, whereby hundreds of chromosomal rearrangements, localised to a limited number of genomic regions, can be acquired in a single catastrophic event[24]. Such events have been shown to occur in both cancer and congenital diseases[25]. Long-read sequencing is ideally suited to analysing such large regions of structural variation. Dr Wigard Kloosterman's team at UMC Utrecht utilised the MinION to sequence a genome exhibiting extremely complex germline chromothripsis[26]. Nanopore sequencing allowed the identification of all of the breakpoints which had been previously characterised by paired end sequencing of mate-pair libraries using traditional short-read technology.

Expanding their research to review genome-wide structural variation, they found similar concordance of results between the long- and short-read data; however, the long-read data also identified many additional variants which were not seen in the short-read data (Figure 5), further underlining the limitations of short-read analysis for structural variant identification. Long-read sequencing on the MinION had the added advantage of allowing the haplotype phasing to be determined for the chromothripsis sample, enabling the confirmation of chromothripsis on the paternal chromosome.

Fast time to results

For some applications, such as cancer detection and monitoring, time to result is of paramount importance. In their research, Norris and coworkers used PCR amplicons to detect specific structural variants and reported a sequencing time of just 15-33 minutes using the MinION[16]. This compares favourably with the minimum of 4 hours that would be required for short-read sequencing, which, as stated by the authors, would struggle to accurately resolve the structural variants[16].

Oxford Nanopore scientists have also successfully utilised the MinION to analyse the BCR-ABL1 fusion gene, which occurs in 90% of people with chronic myelogenous leukaemia (CML). The long read length capability of nanopore sequencing allows all fusion breakpoints to be analysed in a single assay. The researchers suggest that, in the future, the MinION could allow the testing of chronic conditions like CML to be performed in a doctor's surgery, minimising the time necessary to return results and allowing therapeutic intervention within hours of sample acquisition[27].

Unlike traditional sequencing techniques, which typically deliver data in bulk at the end of a run, nanopore sequencing also allows real-time data streaming for immediate analysis. This means that sequencing can be stopped as soon as sufficient data has been obtained, saving more time.

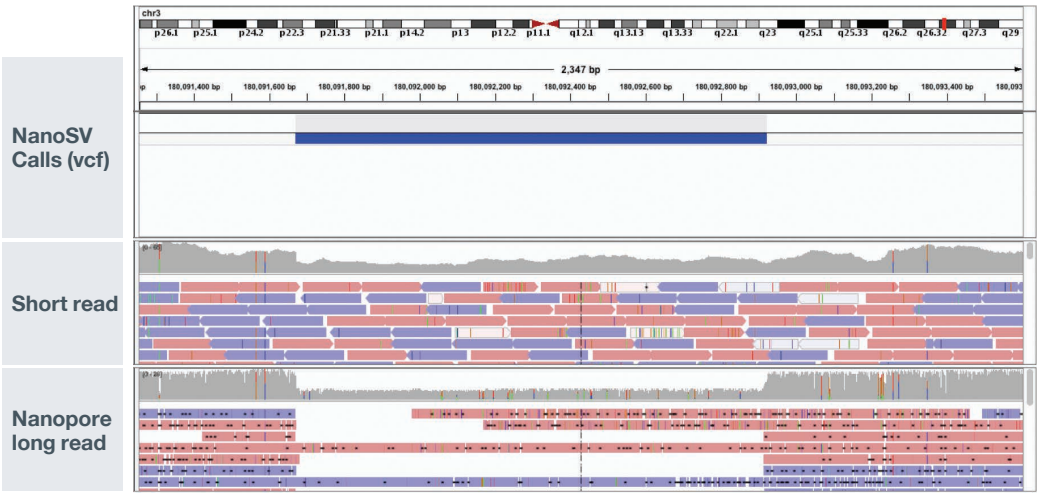


Figure 5
Nanopore long-read sequencing allowed clear identification of a large deletion, which was not evident when using short-read traditional sequencing technology. Figure courtesy of Dr Wigard Kloosterman, UMC Utrecht, The Netherlands.

Low cost

Alternative sequencing platforms typically require large capital investments (>$100k)[16], adaptations to infrastructure and calibration by trained engineers – requiring significant funds and implementation time. For this reason, the majority of sequencing instruments are confined to specialist sequencing centres[26] and core facilities which can further increase time to results. In stark contrast, at only $1,000 for a starter pack, the MinION allows researchers to implement their work immediately within their own labs. The MinION starter pack includes all of the materials required to run initial nanopore sequencing library preparation, including a MinION device, flow cells, kits and membership of the Nanopore Community.

Small footprint

Measuring and weighing about the same as a confectionery bar, the USB powered MinION is uniquely portable, making it ideal for all labs, including those where workspace is at a premium (Figure 6). In addition, researchers have transported the MinION to various remote locations in hand luggage, showing both the portability and robustness of the device[28,29]. No specialist calibration or setup is required: simply load and run.

MinION stats:
Weight: 87 g (103 g with a flow cell)
Size: W 105 mm, H 23mm, D 33 mm

Figure 6
The MinION from Oxford Nanopore is a pocket-sized, portable sequencing device that seamlessly integrates into the laboratory and remote environments.

5

Data analysis tools

Nanopore sequencing enables basecalling and onward data analysis to be performed in parallel as the experiment progresses and the length of time the sequencing continues for can be tailored by the user to the specific application based on the accumulated data. This enables efficient workflows and even real-time selective sequencing.

Customised analysis pipelines can be created by the user from a growing number of community-developed data-analysis tools. These tools work with the different file formats available within the read files, including the easily accessible .fastq files. Existing tools for the analysis of structural variation using nanopore long-read data, including nanoSV[26], Sniffles[30], TULIP[22] and LUMPY[31] (Table 1).

In addition, the Nanopore Community is actively refining and developing new tools for a range of long-read sequencing applications. The community is also a rich resource of support for users, promoting discussion and collaborative experimentation.

Tool	Link	Reference
nanoSV	https://github.com/mroosmalen/nanosv	Kloosterman (2016)[26]
Sniffles	https://github.com/fritzsedlazeck/Sniffles	Sedlazeck (2015)[30]
TULIP	https://github.com/Generade-nl/TULIP	Jansen (2017)[22]
LUMPY	https://github.com/arq5x/lumpy-sv	Layer (2014)[32]

Table 1
List of structural variation data analysis tools designed or adapted for Oxford Nanopore long reads.

For the latest information on analysis tools, visit **www.nanoporetech.com/publications**

6 Summary

Long-read nanopore sequencing has the potential to revolutionise our understanding of structural variation and its role in disease, evolution and genetic diversity.

Researchers are already realising the benefits of long reads to more accurately characterise structural variation in a range of species. The MinION combines improved analysis of structural variation with sensitive detection and rapid time to results, all delivered in a cost-effective and portable package. As stated by Dr Wigard Kloosterman: *"The valuable* [MinION] *data output combined with the small-size and low-cost investment makes nanopore sequencing a prominent platform for… sequencing in future life sciences research and diagnostics*[26]*."*

About Oxford Nanopore Technologies

Oxford Nanopore Technologies is at the forefront of genomics having introducing the world's first nanopore DNA sequencer, the MinION – a portable, real-time, long-read, low-cost device. Through the utilisation of long reads, the MinION is ideally suited to the analysis of large structural variation, offering significant advantages over traditional short-read sequencing technology.

For higher throughput requirements, the GridION™ X5 and PromethION™ are available. These compact benchtop systems utilise the same nanopore technology as the MinION, offering up to 5 and 48 flow cells respectively. Each flow cell can be used independently, with the user choosing how many are used at any one time, enabling different experiments to be run in parallel. Both the GridION X5 and PromethION are available with no capital cost – only consumables need to be purchased – delivering scalable, cost-effective analysis.

For the latest information about structural variation analysis using long-read nanopore sequencing, visit: **www.nanoporetech.com/applications.**

7 References

1. Sudmant, P.H. *et al* (2015) An integrated map of structural variation in 2,504 human genomes. Nature. 2015 Oct 1;526(7571):75-81. doi: 10.1038/nature15394
2. Conrad, D.F. *et al* (2010) Origins and functional impact of copy number variation in the human genome. Nature 464:704–12
3. Stankiewicz, P. and Lupski, J.R. (2010) Structural variation in the human genome and its role in disease. Annu Rev Med. 2010;61:437-55. doi: 10.1146/annurev-med-100708-204735
4. Eichler, E. (2015) Genome structural variation. Presentation. Available at: http://evomicsorg.wpengine.netdna-cdn.com/wp-content/oads/uploads2015/01/011815CeskyKrumlowcompgenomicse.pdf [Accessed: 05 January 2017]
5. Sebat, J. *et al* (2007) Strong association of de novo copy number mutations with autism. Science 316, 445 – 449
6. Bochukova E.G. *et al* (2010) Large, rare chromosomal deletions associated with severe early-onset obesity. Nature 463, 666–670
7. Stefansson, H. *et al* (2008) Large recurrent microdeletions associated with schizophrenia. Nature 455, 232–236
8. Diskin, S. J. (2009) Copy number variation at 1q21.1 associated with neuroblastoma. Nature 459(7249): 987–991
9. Chaisson, M.J.P. *et al* (2015) Genetic variation and the de novo assembly of human genomes. Nat Rev Genet 16(11):62740
10. Lander, E.S. *et al* (2001) Initial sequencing and analysis of the human genome. Nature 409 (6822): 860-921
11. Mac, A.C. *et al* (2016) Genome-Wide Structural Variation Detection by Genome Mapping on Nanochannel Arrays. Genetics 202(1):351-62
12. National Institute of Standards and Technology (NIST), (2016). Study Highlights Need for Better Characterized Genomes for Clinical Sequencing [online] Available at: <https://www.nist.gov/news-events/news/2016/03/study-highlights-need-better-characterized-genomes-clinical-sequencing> [Accessed: 23 February 2017]
13. Simpson, M. (2016) Sequencing human genomes with nanopore technology. Presentation. Available at: https://nanoporetech.com/events/ncm16#193976224 [Accessed: 05 January 2017]
14. Loman, N. (2017) 950kb, yeah! [Twitter] 08 March. Available from: https://twitter.com/pathogenomenick/status/839604187100934144 [Accessed: 09 March 2017] http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/?rev1
15. Periwal, V. and Scaria, V. (2015) Insights into structural variations and genome rearrangements in prokaryotic genomes Bioinformatics 31(1): 1-9 doi:10.1093/bioinformatics/btu600
16. Norris, A.L *et al* (2016) Nanopore sequencing detects structural variants in cancer. Cancer Biol Ther 17(3): 246-253
17. Wei, C.L. (2016) Structural Variation Detection in Cancer Genome by Nanopore Sequencing. Presentation, Nanopore Community Meeting 2016
18. Lu, H., Giordano, F. and Ning, Z. (2016) Oxford Nanopore MinION Sequencing and Genome Assembly. Genomics Proteomics Bioinformatics 14(5):265-79
19. Oxford Genomic Centre, (2016). Nanopore genomes [online] Available at: http://www.well.ox.ac.uk/ogc/nanopore-human-genome [Accessed: 23 February 2017]
20. Tyson, J.R. (2017) Whole genome sequencing and assembly of a Caenorhabditis elegans genome with complex genomic rearrangements using the MinION sequencing device. bioRxiv doi: https://doi.org/10.1101/099143
21. C. elegans Sequencing Consortium. (1998) Genome sequence of the nematode C. elegans: a platform for investigating biology. Science 282(5396):2012-8
22. Jansen, H.J. *et al* (2017) Rapid de novo assembly 1 of the European eel genome from nanopore sequencing reads. bioRxiv doi: http://dx.doi.org/10.1101/101907
23. Oxford Nanopore Technologies, 2017. Sequence from as little as 10pg starting material [online] Available at: https://nanoporetech.com/about-us/news/sequence-little-10pg-starting-material [Accessed: 23 February 2017]
24. Maher, C.A. and Wilson, R.K. (2012). Piecing together the shattering process. Cell 148(0): 29–32.
25. Kloosterman, W.P. and Cuppen, E. (2013) Chromothripsis in congenital disorders and cancer: similarities and differences. Curr Opin Cell Biol. 25(3):341-8
26. Kloosterman, W. (2016) Characterization of structural variations and chromothripsis in nanopore sequencing data of human genomes. Presentation. Available at: https://nanoporetech.com/events/ncm16#193982637 [Accessed 05 January 2017]
27. Oxford Nanopore Technologies, 2017. Potential for self-monitoring of chronic myelogenous leukaemia gene fusion using VolTRAX and MinION [online] Available at: https://nanoporetech.com/publications/potential-self-monitoring-chronic-myelogenous-leukaemia-gene-fusion-using-voltrax-and [Accessed: 23 February 2017]
28. Quick, J. (2015) Lab in a suitcase, and other adventures with Nanopore sequencing. Presentation. Available at: https://nanoporetech.com/publications/josh-quick-lab-suitcase-and-other-adventures-nanopore-sequencing [Accessed: 02 October 2016]
29. Edwards, A. *et al* (2016) Extreme metagenomics using nanopore DNA sequencing: a field report from Svalbard, 78 N. bioRxiv 073965; doi: https://doi.org/10.1101/073965
30. Sedlazeck, F.J. *et al* (2015) Detection of structural variants using third generation sequencing [Poster] Available at: http://schatzlab.cshl.edu/publications/posters/2015/2015.GI.Sniffles.pdf [Accessed: 23 February 2017]
31. Layer, R.M. *et al* (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biology 2014; 15:R84

Oxford Nanopore Technologies
phone +44 (0)845 034 7900
email info@nanoporetech.com
twitter @nanopore
www.nanoporetech.com

SV_W1002_v1_revA_24Mar2017