WHITEPAPER

STRUCTURAL VARIATION IN THE HUMAN GENOME

THE LEADER IN LONG-READ SEQUENCING

PACBIO®

HUMAN BIOMEDICAL RESEARCH

The past quarter century has brought tremendous progress in the detection of single nucleotide variants (SNVs), but intermediate-sized (50 bp to 50 kb) structural variants (SV) remain a challenge. Such variants are too small to detect with cytogenetic methods, but too large to reliably discover with short-read DNA sequencing. **Recent high-quality genome assemblies using PacBio long-read sequencing have revealed that each human genome has approximately 20,000 structural variants, spanning 10 million base pairs, more than twice the number of bases affected by SNVs[1,2,3,4].**

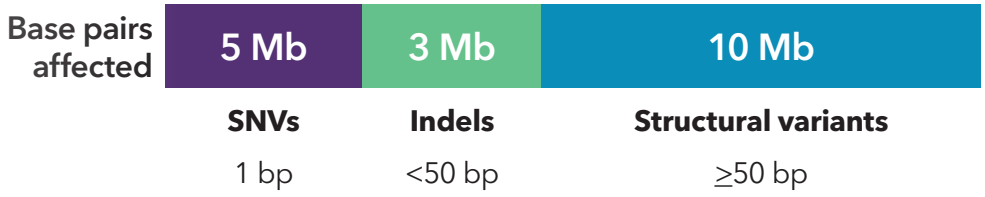


Figure 1. Variation between two human genomes, by number of base pairs impacted.
SNVs = Single nucleotide variants; indels = insertions and deletions[4].

Structural variants of all types are known to cause Mendelian disease and contribute to complex disease. All of these variants can be most robustly detected by PacBio Single Molecule, Real-Time (SMRT®) Sequencing.

Genomics studies have shown that the insertions, deletions, duplications, translocations, inversions, and tandem repeat expansions in the structural variant size range can cause Mendelian disease, including Carney Complex[5], Potocki-Lupski syndrome[6], and Smith-Magenis syndrome[7]. However, robust detection of structural variants remains a hurdle. **More than half of rare Mendelian disease patients go undiagnosed, in part because current diagnostic tools cannot reliably detect the entire size range of genomic variation[8].**

While these disorders are individually rare, they collectively affect ~350 million people globally[9], and the lengthy diagnostic odysseys these patients face has a significant economic impact on society.

Structural variants also contribute to complex disorders, including autism[10,11], cancer[12,13], Alzheimer's disease[14], and schizophrenia[15,16]. Genome-wide association studies (GWAS) have identified genomic loci associated with a wide range of complex diseases, but the precise contributing variant is often

Structural Variant	Disease Examples	PacBio Advantage
INSERTION 	Charcot-Marie Tooth disease, Tay-Sachs disease	- Base pair resolution of breakpoints - Complete inserted sequence
DELETION 	Williams syndrome, Duchenne muscular dystrophy, Smith-Magenis syndrome, Carney Complex	- Base pair resolution of break points - High sensitivity even in repeats
INTERSPERSED DUPLICATION 	APP in Alzheimer's disease, Potocki-Lupski syndrome, Prader-Willi syndrome, Angelman syndrome	- Precise copy number - Base pair resolution of the duplicated sequence - Genomic context of additional copies
TRANSLOCATION 	Down syndrome, XX male syndrome (SRY), schizophrenia (chr 11) , Burkitt's Lymphoma	- Detection of balanced events - Complete sequence information - Unambiguous resolution of genomic context
INVERSION 	Hemophilia A, Hunter Syndrome, Emery-Dreifuss muscular dystrophy	- Detection of balanced events - Continuous sequence information - Base pair resolution of break points
TANDEM DUPLICATION 	FMR1 in Fragile-X, Huntington's disease, Spinocerebellar ataxia	- Complete repeat sequence, including interruptions - Quantitation of repeat expansions

Table 1. Structural variants of all types are known to cause Mendelian disease and contribute to complex disease. *All of these variants can be most robustly detected by PacBio SMRT Sequencing.*

(a)

	Deletions (count)	Deletions (bp)	Insertions (count)	Insertions (bp)
Tandem Repeat	3,963	2,119,904	5,180	2,502,514
Alu	1,137	356,552	1,002	329,314
L1	214	1,036,570	130	596,590
SVA	28	50,579	21	35,814
Unannotated	2,781	660,970	6,399	1,191,048
Total	8,123	4,224,575	12,732	4,655,280

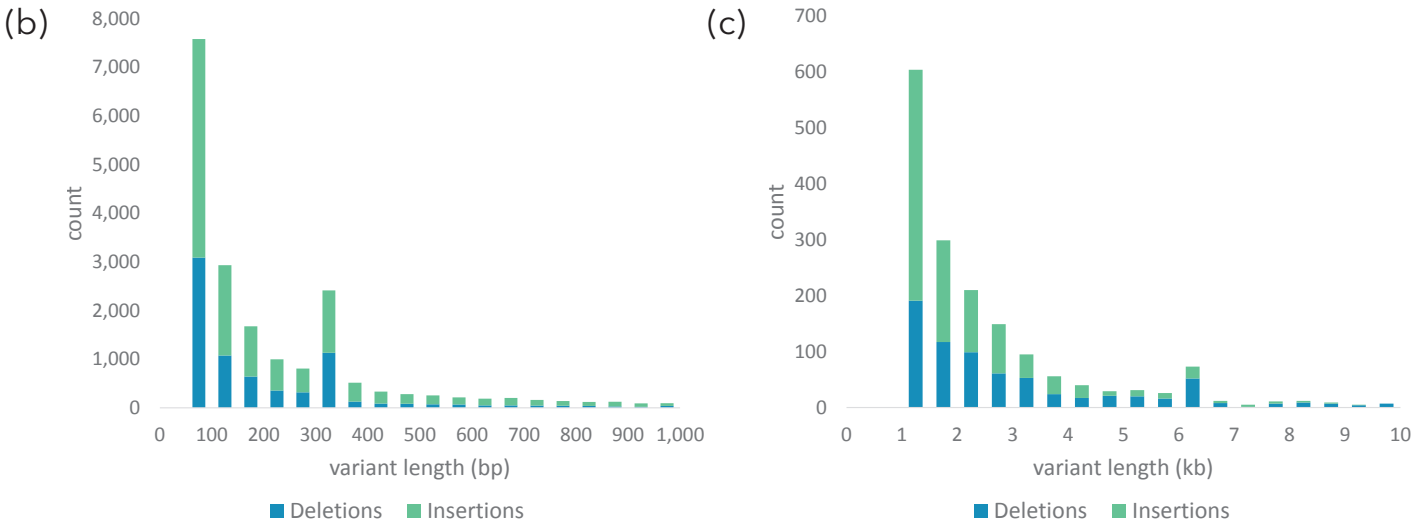


Figure 2. *SMRT Link Structural Variant Calling powered by pbsv identifies structural variants using PacBio long reads. From 10-fold PacBio coverage of the human HG00733, pbsv calls around 20,000 structural variants spanning nearly 10 Mb. SMRT Link provides summary reports on the HG00733 call set that show (a) the variant count by type and repeat annotation; (b) the size distribution for variants shorter than 1,000 bp; and (c) the size distribution for variants 1,000 bp long or greater.*

unknown. In some cases, an associated SNP found at a risk locus is in fact a red herring in linkage disequilibrium with an overlooked causative structural variant that cannot be robustly detected with short reads. A prime example of this is the complement component 4 (*C4*) gene finding in schizophrenia. Strong support for disease association was found for the MHC region on chromosome 6p21.3 across numerous schizophrenia study cohorts. However, none of the SNVs initially investigated could be linked to the disease. Instead, the combination of two structural variants located 5 kb apart in the C4 gene predisposes patients to psychosis[17]. Incomplete information about variants at risk loci has blocked progress in understanding the underlying biology of many complex diseases.

Finally, population genomics studies have shown that structural variants affect numerous phenotypic traits, adding to the human diversity that makes each of us unique. For example, structural variants contribute to height[18], fertility[19], starch digestion efficiency[20], and drug response[21].

The ability to reliably detect rare structural variants in individual genomes and to comprehensively catalog common structural variants in populations are critical gaps in the current genomics

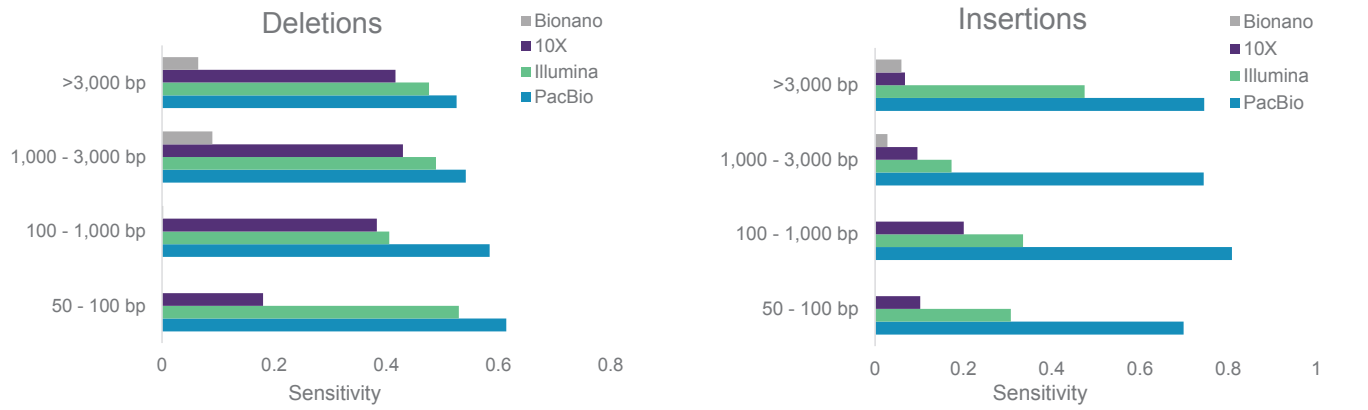


Figure 3. *PacBio outperforms other sequencing technologies in the robust detection of structural variants across the complete size spectrum relevant to understanding human genomic diversity. Recall of consensus insertion and deletion structural variants in the HG002* Genome in a Bottle *integration set v0.3. In this analysis, the reference call set is limited to insertions and deletions which can be detected by at least two orthogonal technologies.*

toolbox[22]. Recent *de novo* assemblies of human genomes have demonstrated the superior power of PacBio SMRT Sequencing to fill this technology gap and sensitively identify structural variants.

Technologies for SV Discovery and Genotyping

Over the past several decades, progress in the systematic discovery of new types of SVs has been tightly linked to technological advances. Karyotyping, FISH, and microarrays were the first technologies to yield insights into SVs, offering 5 Mb, 100 kb, and 50 kb resolution respectively. Today, DNA sequencing delivers base pair resolution of SVs at high throughput, with sensitivity limited by read length and sequencing bias. PacBio reads show no context bias, and PacBio is the current gold standard for producing high quality genome assemblies and detecting structural variants in human genome[1,2,3,4,23,24]. **With highly accurate, continuous long reads and bias-free coverage, PacBio reveals 80% more structural variants than short-read sequencing with a validation rate of 97%** [4,17]**.** In addition, the SMRT Analysis integrated tool pbsv makes it simple to generate a high confidence list of structural variants from PacBio data, as shown in Figure 2. In contrast, short-read sequencing (e.g. Illumina, Ion Torrent) is unable to resolve repetitive and highly homologous regions of the human genome that are enriched for structural variants[25]. These regions are also among the most highly-polymorphic in human populations (e.g. MHC region[26]) and include both genes and pseudogenes (e.g. CYP2D6, CYP2D7[21]) that significantly impact health and disease. These shortcomings extend to linked-read and synthetic long-read technologies that are built on the short-read platforms (e.g. 10X Genomics, TruSeq). Linked reads and synthetic long reads lack the continuous sequence information needed to reliably detect insertions, are prone to spurious false positive deletions, and are hampered by the well-documented context bias of

PacBio outperforms other sequencing technologies in the robust detection of the complete size spectrum of structural variants relevant to understanding human genomic diversity.

Sensitivity and Reliability of Structural Variant Calling Platforms

PacBio has the highest sensitivity

	Deletions			Insertions		
	Counts	FDR	Sensitivity	Counts	FDR	Sensitivity
PacBio (30-fold)[28]	8,737	3%	95%	12,378	3%	93%
PacBio (10-fold)[28]	6,798	3-10%	83%	11,252	3-10%	83%
ONT (30-fold)[29]	28,791	65%	93%	3,900	65%	11%
10X Genomics (30-fold)	3,166	Not reported	39%	Not reported	N/A	0%
Illumina[30]	1,910	2-4%	24%	1,090	1-4%	9%
BioNano[31,32]	522	3%	6%	769	2%	6%

Oxford Nanopore has the highest false discovery rate

Other technologies struggle with poor sensitivity

Table 2. *PacBio has the highest sensitivity and specificity for structural variants greater than 50 bp, even at low coverage. Other technologies struggle with poor sensitivity and/or high false positive rates. PacBio: NGM-LR (GRCh37/hg19) & PBHoney; Illumina: 1000genomes.org; 10X: http://www.slideshare.net/GenomeInABottle/sept2016-sv-10x; ONT: https://github.com/nanopore-wgs-consortium/NA12878, 12/5/2016 release*

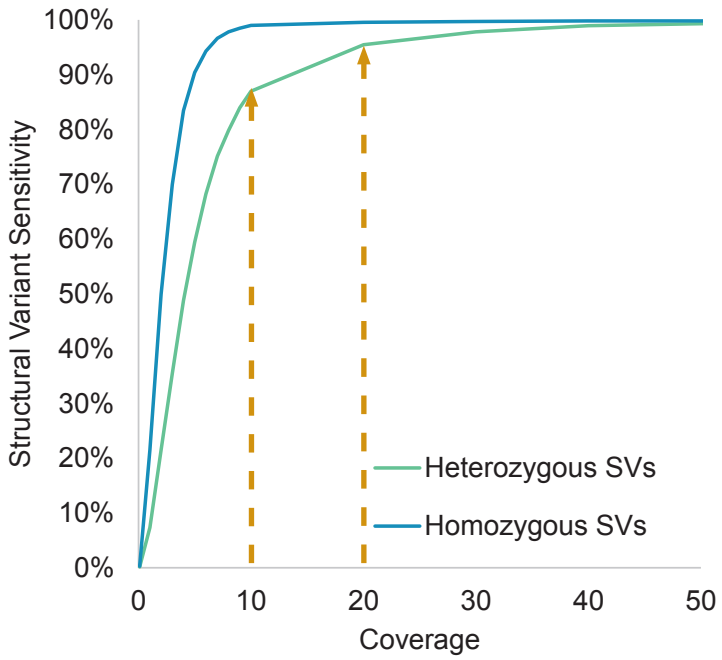


Figure 4. *Low-fold coverage with PacBio sensitively recalls structural variants from a high-coverage call set. Data is from sequencing of HG00733 on the Sequel® System, with variants called using pbsv[33]. 100% detection is defined as recapitulation of the call set generated with 70-fold coverage. Recall is 87% and 95% at 10- and 20-fold coverage, respectively.*

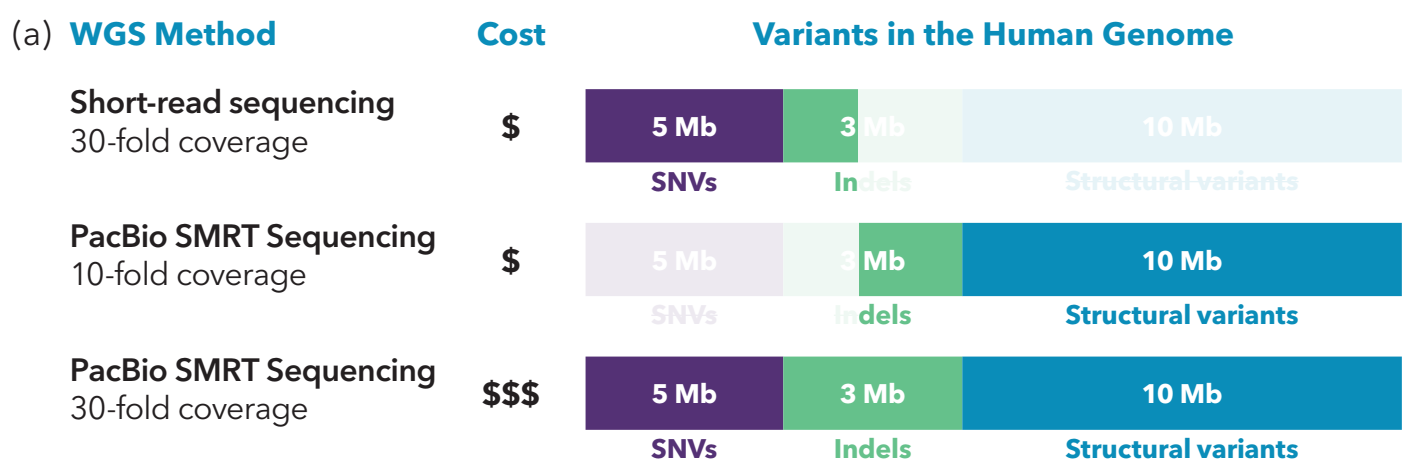
the underlying sequencer[27]. The Oxford Nanopore Technologies (ONT) MinION produces continuous reads that span several kilobases, but due to systematic errors the technology is currently unable to capture simple repeats reliably and suffers from an extremely high false positive rate for deletion calling.

Structural Variant Detection with Low-Coverage PacBio long-read sequencing

While *de novo* assembly is the most comprehensive way to identify variants in a genome, the sequencing and bioinformatics resources required are significant. However, a full assembly is not required for robust variant discovery. Low, unbiased coverage of the genome with long-read sequencing reveals most of the structural variants uncovered with *de novo* assembly, but at a price point that enables characterization of large cohorts for disease-associated SV discovery or cataloguing common SVs in different ethnic populations. The low-coverage approach can also make it economical to more effectively screen an individual genome for a causative structural variant in patients with rare, undiagnosed disease.

As shown in Figure 4, sequencing a genome to a 10-fold coverage with PacBio identifies ~85% of the deletions and insertions called by *de novo* assembly. The low-coverage method was validated by down sampling 70-fold coverage dataset of HG00733 collected on the Sequel System and comparing structural variants called with pbsv to the comprehensive set detected using all the data. Not only does low-coverage compare well to the gold standard method, it reveals thousands more high confidence SVs than can be found using short reads. The reason such robust detection is possible with low coverage is that sampling is truly uniform and errors are random, so that even two reads supporting a structural variant enables a high confidence call. In addition, the majority of SVs in the genome are less than 10 kb, often mediated by short interspersed elements (SINEs) and long interspersed elements (LINEs), which is far shorter than the median read lengths produced by SMRT Sequencing.

(a)



(b)

Application	Study Design
Disease Research • Find causative mutations in undiagnosed cases • Find novel disease genes in disease cohorts	≥10 samples at 10-fold PacBio coverage
Population Genetics • Identify polymorphic structural variants present at ≥1% frequency in the population	≥250 samples at 5-fold PacBio coverage
Clinical Genetics • Find causative mutations in patients with genetic disease	1 sample at 30-fold PacBio coverage

Figure 5. *Low-coverage SMRT Sequencing for structural variant discovery. (a) Low-coverage PacBio SMRT Sequencing is a cost-effective complement to short-read sequencing to obtain the most comprehensive view of human genetic variation. (b) The optimal balance of cohort size and sequencing coverage depends on the specific goals of a study. For many studies, low-coverage sequencing is the most effective approach.*

Conclusions

Structural variants play a key role in human disease, evolution and genetic diversity. Without access to the full breadth of variation present in genomes and exomes, future precision medicine efforts will be limited in their ability to connect genotypes to phenotypes and distinguish common from rare or potentially disease-linked variants. New long-read sequencing approaches are needed to meet this challenge, as short-read sequencing technologies only detect 20% of the SVs present in the human genome[27].

PacBio long-read sequencing is the only available technology that can access the full range of genomic variants, from SNVs to multi-kilobase SVs, in a fully integrated solution with haplotype-resolved sequences. To fully characterize the large cohorts required to understand complex disease or catalog population-specific common variants, low-coverage genome sequencing with the PacBio Sequel System provides an affordable and effective alternative to *de novo* assembly.

References

1. Shi, L. et al. (2016) Long-read sequencing and *de novo* assembly of a Chinese genome. *Nature Communications*. 7, 12065.
2. Seo, J. et al. (2016) *De novo* assembly and phasing of a Korean human genome. *Nature*. 538, 243–247.
3. Pendleton, M. et al. (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nature Methods*. 12(8), 780–786.
4. Huddleston, J. et. al. (2017) Discovery and genotyping of structural variation from long-read haploid genome sequencing data. *Genome Research*. 27, 677-685.
5. Merker, J. et. al. (2017) Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genetics in Medicine*. doi:10.1038/gim.2017.86
6. Wang, M. et al. (2015) PacBio-LITS: a large-insert targeted sequencing method for characterization of human disease-associated chromosomal structural variations. *BMC Genomics* 16, 214.
7. Fonseca, C. et.al. (2016) Combined next generation sequencing techniques untangle the genetic structure of complex nonrecurrent deletions in subjects with Smith-Magenis syndrome and reveal a strong bias to paternally deleted chromosomes. *ASHG Annual Meeting 2016 Poster Talk,* PgmNr 164.
8. Biesecker, L. et. al. (2011). Exome sequencing: the expert view. *Genome Biology*. 12, 128.
9. Global Genes. RARE Diseases: Facts and Statistics. Retrieved from https://globalgenes.org/rare-diseases-facts-statistics/
10. Dennis, M. et al. (2017) The evolution and population diversity of human-specific segmental duplications. *Nature Ecology & Evolution*. 0069. doi: 10.1038/s41559-016-0069.
11. Hoischen, A. et al. (2014) Prioritization of neurodevelopmental disease genes by discovery of new mutations. Nature Neuroscience. 17, 764–772.
12. Vogelstein B, et. al. (2013) Cancer Genome Landscapes. *Science*. 339, 6127.
13. Zack, T. et. al. (2013) Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*. 45(10), 1134-1143.
14. Roses, A. (2016) Polyallelic structural variants can provide accurate, highly informative genetic markers focused on diagnosis and therapeutic targets: Accuracy vs. Precision. *Clinical Pharmacology & Therapeutics*. 99(2), 169-71.
15. Klar, A. (2004) A genetic mechanism implicates chromosome 11 in schizophrenia and bipolar diseases. *Genetics* 167, 1833–1840.
16. CNV and Schizophrenia Working Groups of the Psychiatric Genomics Consortium; Psychosis Endophenotypes International Consortium.(2017) Contribution of copy number variants to schizophrenia from a genome-wide study of 41,321 subjects. *Nature Genetics*. Jan;49(1), 27-35.
17. Sekar, A. et al. (2016) Schizophrenia risk from complex variation of complement component 4. *Nature*. 530(7589), 177-83.
18. Hardy, J. et al. (2005) Evidence suggesting that Homo neanderthalensis contributed the H2 MAPT haplotype to Homo sapiens. *Biochemical Society Transactions*. 33(Pt 4):582–585.
19. Stefansson, H. et al. (2005) A common inversion under selection in Europeans. *Nature Genetics*. 37(2), 129–137.
20. Perry ,G. et al. (2007) Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*. 39(10), 1256–1260.
21. Qiao, W. et. al. (2016) Long-read single molecule real-time full gene sequencing of cytochrome P450-2D6. *Human Mutation*. 37(3), 315-23.
22. Huddleston, J. et al. (2016) An incomplete understanding of human genetic variation. *Genetics*. 202(4), 1251-4.
23. Zook, J. et. al. (2016) Data Descriptor: Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*. 3, 160025.
24. Chaisson, M. et. al. (2015) Resolving the complexity of the human genome using single molecule sequencing. *Nature*. 517, 608-611.
25. Mandelker, D. et. al. (2016). Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. *Genetics in Medicine*. 18, 1282–1289.
26. Norman, P., et. al. (2017) Sequences of 95 human MHC haplotypes reveal extreme coding variation in genes other than highly polymorphic HLA class I and II. *Genome Research*. 27, 813-823.
27. Schatz, M. (2017) "Personalized Phased Diploid Genomes of the EN-TEx Samples". Advances in Genome Biology and Technology Conference (AGBT), Hollywood, FL.
28. Wenger, A. "Identifying structural variants in NA12878 from low-fold coverage sequencing on the PacBio Sequel System." Web blog post. PacBio Blog. PacBio, 19 Oct 2016, Web. 10 May 2017.
29. Kloosterman, W. (2016) "Characterization of structural variations and chromothripsis in a nanopore sequencing of human genomes". Nanopore Community Meeting 2016.
30. Sudmant, P. et. al. (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*. 526, 75–81.
31. Mak, A. et al. (2016) Genome-wide structural variation detection by genome mapping on nanochannel arrays. *Genetics*. 202(1), 351-62.
32. Hastie, A. et. al. (2017) Rapid automated large structural variation detection in a diploid genome by nanochannel based next-generation mapping. http://dx.doi.org/10.1101/102764.
33. "The Human Genome Structural Variation Consortium". The 1000 Genomes Project. IGSR: The International Genome Sample Resource. PacBio Single molecule real time (SMRT) Sequel System high-coverage dataset for HG00733. March 2017.

THE LEADER IN LONG-READ SEQUENCING

PN: WP100-100517