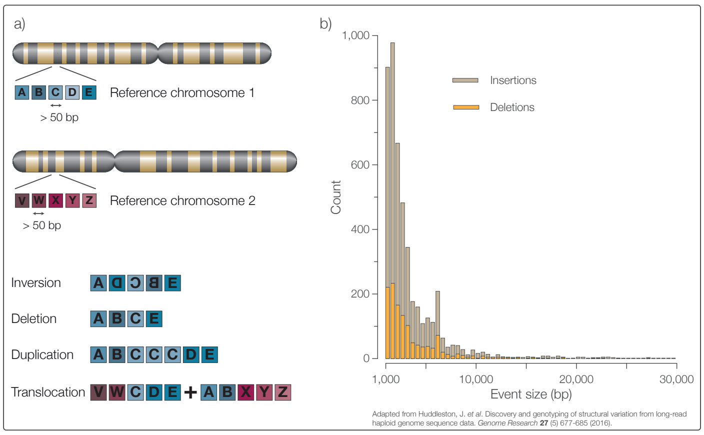
Using long nanopore reads to delineate structural variants (SVs) in the human genome

SVs, including large deletions, duplications, inversions, translocations and copy-number changes are abundant in large genomes, and require long reads for precise characterisation

Contact: **publications@nanoporetech.com** More information at: **www.nanoporetech.com** and **publications.nanoporetech.com**

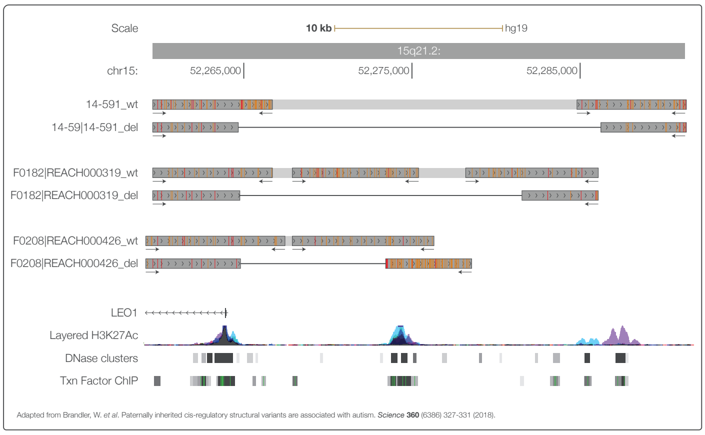


Adapted from Huddleston, J. *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research* **27** (5) 677-685 (2016).

Fig. 1 Structural variation a) classes b) variant size and frequency in the human genome

Structural variation: large inversions, deletions, duplications and translocations

Structural variation (SV) refers to inversions, insertions, deletions and translocations > 50 bp in length (Fig. 1a). SV encompasses millions of bases of DNA per human genome, can span tens of kilobases containing entire genes and their regulatory regions (Fig. 1b) and contributes substantially to genome variation. SV can alter the copy number of dosage-sensitive genes, can unmask recessive alleles and can disrupt the integrity or regulation of a gene, all of which can cause genetic disease. The study of SVs is challenging because they frequently arise in repetitive regions of the genome, and can have highly complex structures. Short-read sequencing technologies cannot span long SVs, leading to incomplete reference assemblies.



Adapted from Brandler, W. *et al.* Paternally inherited cis-regulatory structural variants are associated with autism. *Science* **360** (6386) 327-331 (2018).

Fig. 3 Confirmation of LEO1 breakpoints and parental origin with nanopore reads

Deletion of a regulatory element in autistic patients validated by long nanopore reads

To demonstrate the utility of long nanopore reads in resolving structural variants, we amplified and barcoded patient and wild-type alleles from three families with known deletions in the LEO1 locus on chromosome 15, and sequenced them on a flowcell. Deletion amplicons were approximately 10 kb in length, and the amplifiable wild-type amplicons spanned up to 20 kb. LEO1 encodes an RNA polymerase-associated protein which is expressed during foetal brain development. For the deletions, we created consensus reference haplotypes using Nanopolish and realigned reads to these references for SNP-calling with MUMmer. All three deletions, as well as the parental origin, were successfully validated by the nanopore reads (Fig. 3).

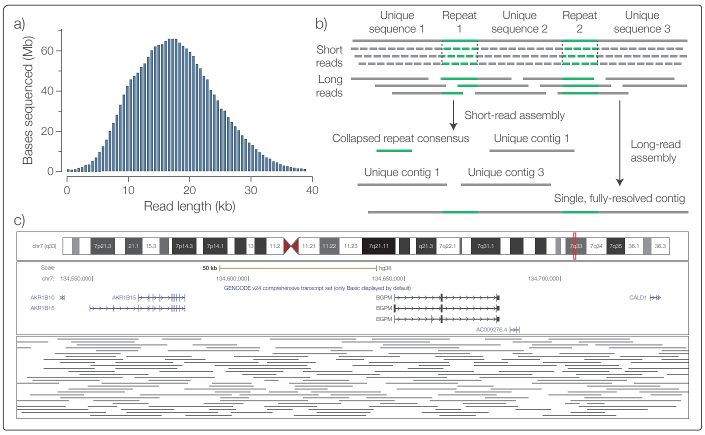
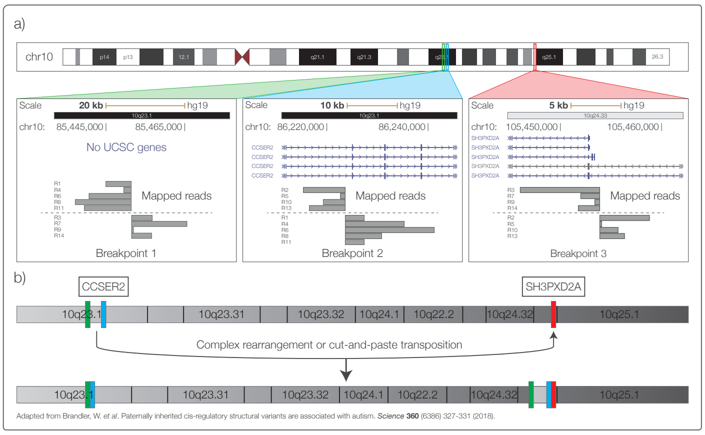


Fig. 2 Read length a) typical distribution b) assembly c) mapped long human MinION reads

Nanopore sequencing can give extremely long reads without size selection

The read length that can be obtained from nanopore sequencing is limited only by the integrity of the DNA extracted from the sample and the care taken during library preparation. The read-length distribution corresponds closely to the fragment-length distribution of the sample DNA. When starting with high-molecular weight genomic DNA, it is straightforward to obtain reads that are tens of kilobases in length (Fig. 2a). The longer the sequence read, the longer the repetitive region or SV that can be resolved, allowing the correct structure of the variant to be elucidated (Fig. 2b). Recent increases in throughput make it realistic to sequence whole human genomes on a MinION (Fig. 2c).



Adapted from Brandler, W. *et al.* Paternally inherited cis-regulatory structural variants are associated with autism. *Science* **360** (6386) 327-331 (2018).

Fig. 4 Detection of SVs by whole-genome sequencing a) mapped reads b) SV resolution

Using long-read whole-genome sequencing to resolve SVs in the human genome

One individual who participated in the autism spectrum disorder study described in Fig. 3 had been diagnosed with depression/anxiety. She appeared to have an SV in chromosome 10 which had been identified as a complex break-end by Lumpy analysis of paired-end Illumina data. The SV was not found in the individual's parents, so was taken to be *de novo*, but the precise structure was unclear. We performed whole-genome library prep using an LSK-108 kit, and sequenced the library on a FLO-MIN106 flowcell, generating approximately 24 Gb of sequence data. The long reads allowed us to fully resolve the variant, and nanopore data was phased using WhatsHap, revealing the individual's mother to be the parent of origin of the SV.