Human Genomics

"Are long reads relevant in human disease?"

Structural variants (SV) in the human genome of around 1,000 base pairs in length are known to be important in many significant diseases such as autism, obesity, schizophrenia and cancer. The genomic difference between individuals caused by typical SVs is estimated to be 3–10 x higher than that caused by single nucleotide variants. Long-reads that can resolve these structural variants are therefore very relevant in many human disease, however to-date, most attention has focussed on single nucleotide variations. This is probably because the available tools to read short lengths of 300 base pairs are so accurate at reading base pairs and are also relatively fast and cheap.

"Simply put, shorter reads are less likely to cover large structural variants"

Effective long-read technologies that can adequately span these long structural variants are emerging and interest is growing in this capacity. The complexity of structural changes that can happen in human diseases including cancer are well illustrated by this publication. (ref. Kloosterman, Nature Communications 8: 1326, DOI: 10.1038/s41467-017-01343-4 (2017). Chromothripsis is believed to originate as a rare single catastrophic event, early in the cell's history. It is characterised by clustered chromosomal re-arrangements which is evident in congenital disease and a number of cancers. Although short-read technologies are excellent tools for defining the individual base pair changes of single nucleotide variations (SNV), they are not so great at detecting genetic code breakpoints and phasing of complex structural variations (SV) which can better benefit from long-read technologies. The Kloosterman group found that long-read technologies identified 32% more breakpoint junctions compared to short-read Illumina sequencing. The long-read DNA sequencing technologies managed to put together contigs that were 241kb to 1,217kb long spanning 3 to 5 chromothriptic segments. Animal genomes are large (around 3 billion bases) in comparison to bacterial (around 5 million bases) and therefore require very large amounts of sequencing data produced to cover the genome enough times

RevoluGen's latest workflow for Fire Monkey v6 covers the horse genome 7 x (21 Gb sequenced) with a single flow cell (basically saturating the yield capacity of ONT's MinION flow cell). It is anticipated that if the same experiment was repeated on a PromethION flow cell that has yield capacity up to 130 Gb more yield would be produced at a high N5. For this type of coverage Fire Monkey-generated N50 values are at 56kb whereas for similar samples Genomic Tip's N50 value is only around 20-30kb. Of course, ONT sequencing alone does not guarantee the base read specificity needed. RevouGen considers that it is possible to extract DNA from the sample with Fire Monkey, then use the gold standard sequencing technology (i.e. Illumina) on the Fire Monkey extract. Once the first sequencing pass is over, a researcher could then decide which samples they wanted to go back and perform long-read sequencing upon. Fire Monkey extraction is fecund and usually produces enough DNA yield for both Illumina and ONT runs. The practical benefit of this would be to remove the need to repeat a sample collection and to preserve precious rare samples. In this ideal scenario Fire Monkey could be sequencing-agnostic. The main issue would be the availability of an automated Fire Monkey protocol and device and the overall price differential between the Fire Monkey kits and the other short-read NAIP kits.

Long structural variants are 3 to 10 x more importance than single nucleotide variants in human disease

Structural variation in the human genome (ref Lars Feuk, Andrew R. Carson & Stephen W. Scherer Nature Reviews Genetics volume 7, pages 85–97 (2006))

1. Structural variants in the human genome include cytogenetically detectable and sub microscopic deletions, duplications, large-scale copy-number variants, inversions and translocations.
2. The ability to detect and characterize structural variants in the 1-kb to 3-Mb size range in a robust manner across the genome has not been possible until recently.
3. New developments in genome-scanning technologies and computational methodologies, and the availability of a reference sequence for comparison, have made possible the large-scale discovery of structural variants.
4. Many studies are revealing that the total content of structural variants in the human genome could equal or exceed that of SNPs.
5. Structural variants often coincide with low-copy repeat DNA (also called segmental duplications), as these highly related sequences are more likely to undergo non-allelic recombination and subsequent rearrangement.
6. Structural variation in the genome can directly or indirectly influence gene dosage through different mechanisms, and therefore influence phenotypic variation and disease.