Confirming a large complex structural variant using long read sequencing: a case study

Jade Forster1, 2Jane Gibson, 2Sarah Frampton, 1Mildred Iro, 1Tony Williams

1 University of Southampton , The Wessex Investigational Sciences Hub laboratory, Southampton , United Kingdom 2 Cancer Genomics Group, Cancer Sciences , Southampton, United Kingdom

Background

Here we present a case study of a 13-month-old boy referred for further investigations after presenting with bacterial meningitis (H Influenzae), despite having full immunisations. Subsequent immunological investigations identified a low to absent response to protein and conjugate polysaccharide vaccines, their identical twin brother similarly showed a comparable lack of antibody response. Extended Functional immune investigations have since taken place over the last 16 years along with genomic investigations, including exome and whole genome sequencing via the 100,000 genomes project. With little success, the patients have yet to obtain a formal diagnosis of their condition but hold a working molecular diagnosis of common immune pathway defect.

A recent GECIP review of the 100,000 genomes project whole genome sequencing data identified a novel large complex structural variant (SV); an inverted duplication triplication event on chromosome 6, thought to impact the *TNFAIP3* gene and predicted to cause defective production of NFKB1-dependent cytokines. Due to the size and complexity we are unable to characterise the SV using the short read sequencing data.

Aim

To use long read Oxford Nanopore Technologies sequencing to fully resolve, characterise and confirm this highly complex structural variant.

Methods

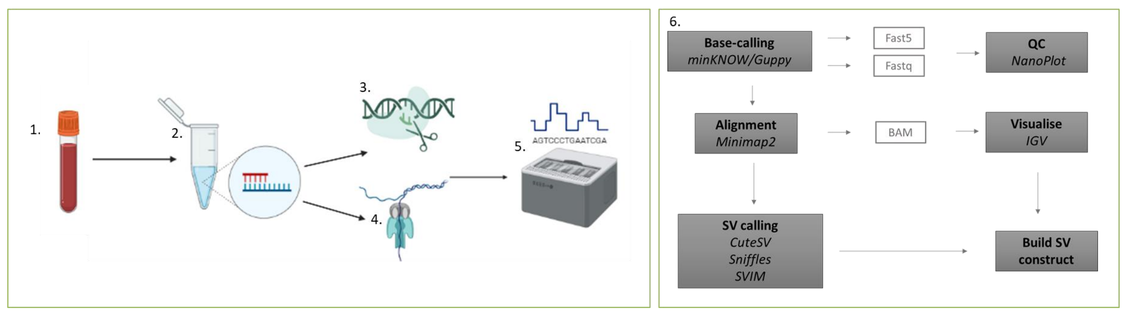

Figure 1 Experimental overview investigating complex structural variant, inverted duplication triplication event. Oxford Nanopore long read sequencing methodologies were identified as having potential to full characterise the event. Image created using BioRender.

1. A whole blood sample was taken from patient for DNA extraction using the NEB HMW monarch kit
2. DNA was sheared using a needle syringe (27G x10 passes)
3. Guide RNAs were designed around a 20kb region encompassing the *TNFAIP3* gene for Cas9 enrichment-double cut excision approach
4. Ligation preparation with adaptive sampling feature of a 35Mb region encompassing the proposed SV on chr6: 137,671,557 - 138,138,251. A FASTA ref file of our region of interest is provided to the GridION before sequencing
5. Sequencing on the GridION using R9.4.1 flow cells. For adaptive sampling enrichment of region of interest (FASTA file) was chosen, anything not aligning to this would be ejected from the pores, whilst sequencing
6. Bioinformatics pipeline for data analysis. Alignment using minimap2 used –y setting to prevent hardclipping of the reads. EPI2ME labs cas9 workflow was used for the cas9 QC. A multitude of structural variant callers were used in data analysis, such as CuteSV, Sniffles and SVIM to produce vcf file, these were also uploaded to IGV to help identify the breakpoints

Results

Cas9 enrichment

- The cas9 double cut excision approach to which we designed our guide RNA probes at either side of the *TNFAIP3* gene provided on average around 111x enrichment over our 20kb region of interest (**Figure 2**)
- Cas9 enrichment does not identify the complex structural variant over this region
- We needed to expand region to encompass the full 466kb loci
- Adaptive sampling was used as an alternative method

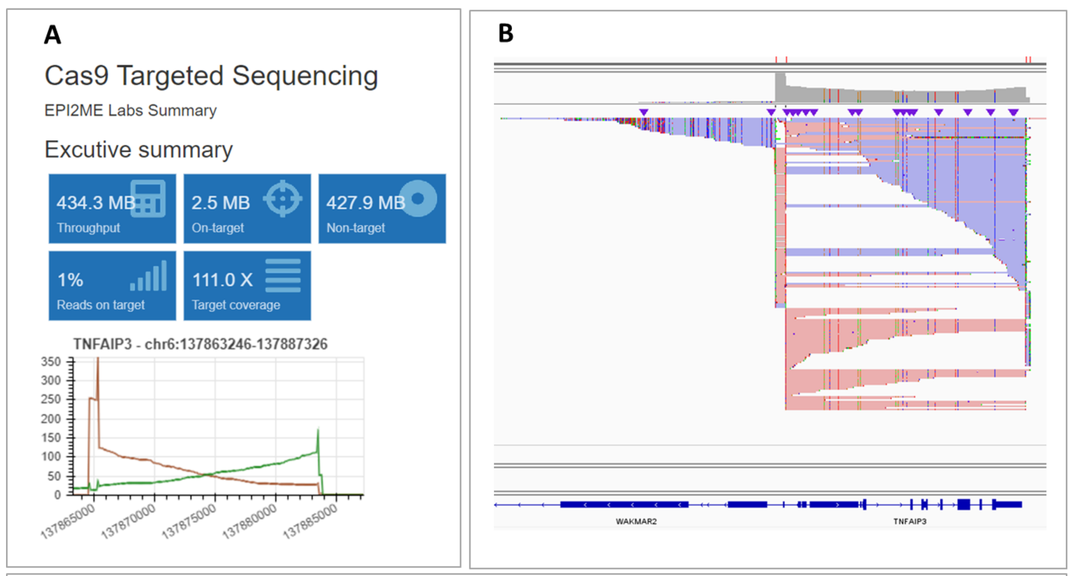

Figure 2 **Cas9 enrichment sequencing results A** Double cut excision approach of a 20kb region encompassing *TNFAIP3* gene. 1% of reads were on target, giving on average 111X target coverage. **B** IGV screenshot of aligned reads over the TNFAIP3 gene, no structural variants were called but reads indicate inversion with clipped reads and SNP fall out.

Adaptive sampling

- Adaptive sampling enriched our region of interest (35Mb) to around 25x
- No SV calling tool could identify the large (460kb) complex inverted duplication triplication event
- But the breakpoints were identified in the region (**Figure**)
- Ongoing analysis to arrange and fully characterise the structural variant into a complete construct

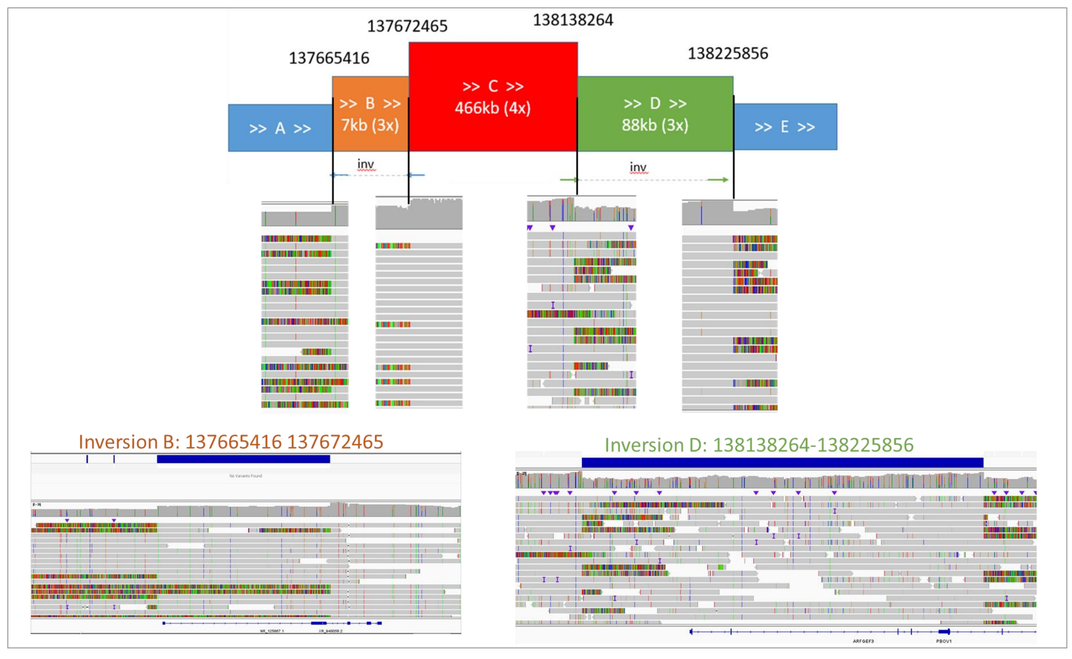

Figure 3 Breakpoints can be identified from the adaptive sampling data. IGV screenshots in relation to breakpoints. Read depth relates to breakpoints and inversion events identified by the structural variant callers CuteSV and SVIM (vcf in blue). Soft clipping is prevalent across implicated region due to the inversion duplication-triplication events. Acknowledgement to Alistair Pagnamenta for providing potential structural configuration constructs from their own long read data to assist our own investigations.

Discussion and Future work

- Oxford Nanopore Technologies long read sequencing cas9 and adaptive sampling features were used to try to deconvolute a complex structural rearrangement in twins with a common immune pathway defect
- Bioinformatic approaches are under going development to further characterise this complex inverted duplication triplication event:
  - i.e. trying to phase reads-difficulty due to the SV
- DNA methylation analysis of this region

Contact info
Dr Jade Forster (j.forster@soton.ac.uk)
@JadeForster