Transcriptome analysis using long nanopore reads

wf-transcriptomes – a cDNA and RNA sequencing data analysis workflow that leverages long nanopore reads, providing a detailed view of the transcriptome.

Contact: **neil.horner@nanoporetech.com**
More information at: **github.com/epi2me-labs/wf-transcriptomes**
Data used in this analysis is available to download from **labs.epi2me.io/lc2024-datasets**

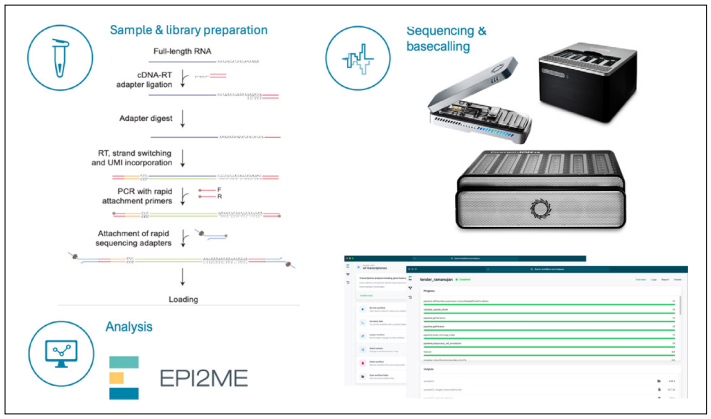

Fig. 1 Overview of library prep, sequencing, and data analysis

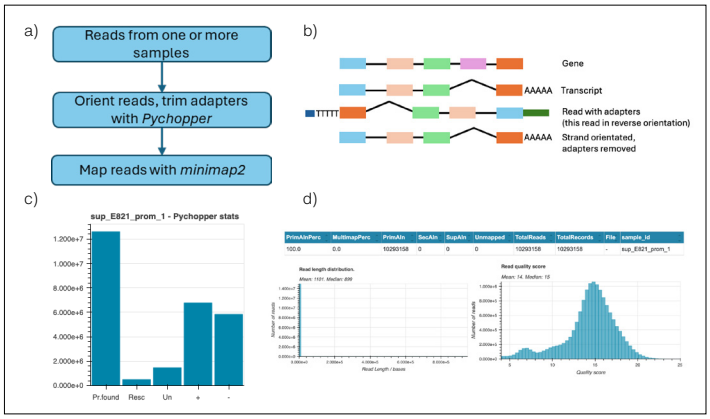

Fig. 2 a) Pre-processing overview b) Reads are stranded, orientated, and trimmed c) Pychopper results, detailing number of full-length reads and unclassified reads as well as the original strandedness d) Read and mapping summaries

Initial insights into your cDNA or RNA reads

Transcriptomic analysis using traditional sequencing technologies relies on the use of short reads. These represent short fragments of transcripts and miss important information, resulting in an incomplete view of the transcriptome. Long nanopore reads are well suited to transcriptomic analysis as they can cover full-length transcripts, allowing more accurate calling of isoforms. Long nanopore reads also lend themselves to the identification of recombination events, which can be useful in discovering gene fusions. Here we introduce wf-transcriptomes, which is a simple-to-use workflow to map your nanopore sequencing data, build and identify transcripts, gain a view of isoform diversity, and identify potential fusion genes, as well as to characterise differential gene expression and differential transcript usage. To highlight some of the features of the workflow, we have analysed a Universal Human Reference RNA (UHRR) sample that consists of ten different cell lines.

Initial preprocessing of reads

The minimum required inputs to the workflow are FASTQ reads and a reference genome sequence plus annotation. Reads after basecalling will be in either mRNA sense or antisense orientation unless generated using the direct RNA kit. The preprocessing step orients reads, such that they are all positive sense, by identifying the adapters at each end of the read. The identified adapters are removed, and any chimeric reads are split into multiple subreads (Fig. 2b). This preprocessing is done with pychopper. The workflow report includes the pychopper summary chart (Fig. 2c) that details the number of full length reads identified, the identified strand, the number of unusable and rescued reads (reads originating from chimeric reads). The workflow report also contains a QC section showing the read length and quality distribution plots as well as the mapping statistics, providing a useful first glance at your data that may help to identify any issues encountered during library preparation or sequencing (Fig. 2d).

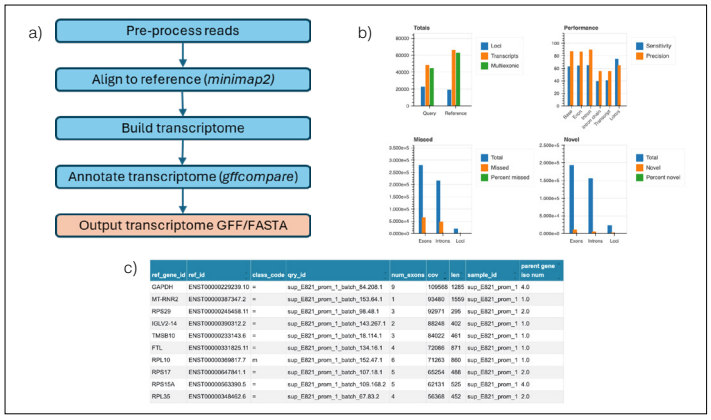

Fig. 3 Isoform identification. a) Overview of isoform detection b) *gffcompare* results, summarising sample transcriptome c) Table detailing each identified isoform

Identification of isoforms

Per-sample transcriptomes are made using stringtie. This is annotated with gffcompare and transcriptome FASTA and associated GFF annotation files are output (Fig. 3a). Amongst other data, the report contains an isoform table that details gene, transcript isoform, coverage, and gffcompare class (for example, '=' is a complete match to a reference transcript; Fig. 3c). The annotation summary section (Fig. 3b) details the gffcompare comparison of the workflow and reference transcriptomes. It shows total transcripts found and the numbers of novel and missed isoforms, as well as sensitivity and precision, giving an idea of the amount of overlap between the sample transcriptome and the reference annotation.

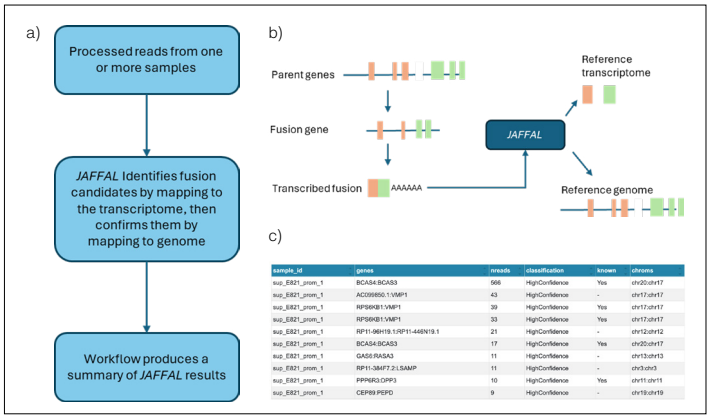

Fig. 4 Fusion gene identification. a) Gene fusion identification overview b) Schematic of a gene fusion read c) Gene fusion results table

Identification of gene fusions

Gene fusions are the combination of two genes by recombination. Uncovering identities of gene fusions in RNA cancer research samples provides valuable insights and potentially allows cancer subtypes to be identified. JAFFAL initially aligns reads to a reference transcriptome identifying reads that map to two genes. Candidates are then validated by mapping to the reference genome, followed by the grouping of fusions by the analysis of breakpoints and then ranking by confidence. Results are displayed in a table in the report, listing potential gene fusions and their gene partners, as well as the number of supporting reads and JAFFAL classification. BCAS4:BCAS3 is the most highly represented fusion (Fig 4c), which likely originates from one of the breast cancer cell lines from the UHRR sample.