Detection of alternative isoforms of gene fusions using long-read Nanopore RNA-seq

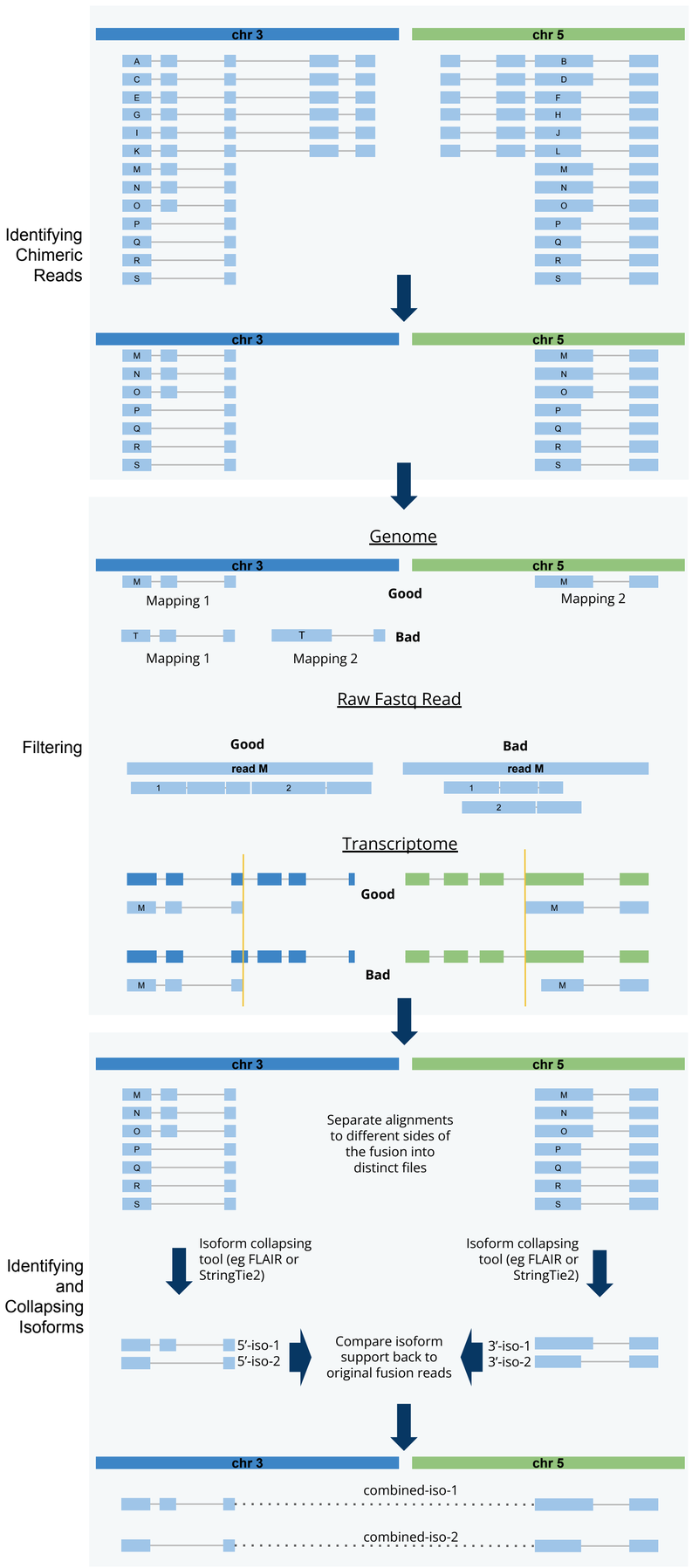
Colette Felton[1], Alison Tang[1], Catherine J Wu[2,3,4], Angela N Brooks[1]

[1]University of California Santa Cruz, Department of Biomolecular Engineering, Santa Cruz, CA, [2]Dana-Farber Cancer Institute, Department of Medical Oncology, Boston, MA, [3]Broad Institute of Harvard and MIT, , Cambridge, MA, [4] Brigham and Women's Hospital, Harvard Medical School, Department of Medicine, Boston, MA

Introduction

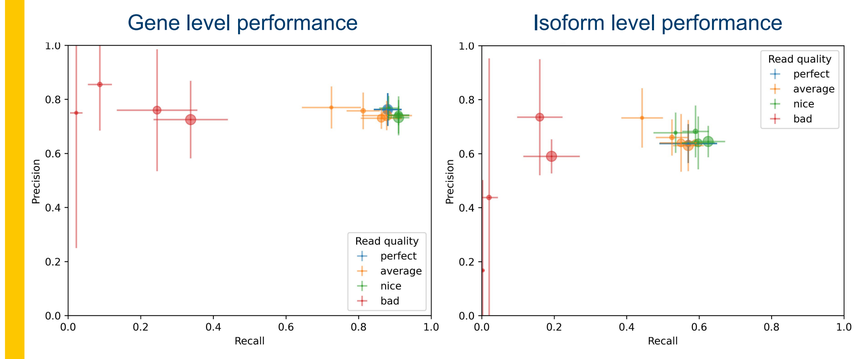
- Gene fusions are the result of chromosomal rearrangement and other structural variants
  - They can be detected in RNA-seq as chimeric transcripts
- They can be important cancer drivers and drug targets
- It is difficult to identify gene fusions from short-read (Illumina) data
  - Very few of the reads will cross the fusion breakpoint
  - Lots of noise from other mismapping reads
- Long read sequencing (nanopore, pacBio) can capture whole transcripts
  - More sequence around the fusion point allows for more confident fusion detection
  - Also allows for better alternative splicing and full-length isoform identification
    - The ability to detect alternative splicing in fusions is novel and clinically relevant
      - Alternative splicing has been shown to be a method for developing resistance to drugs targeting gene fusions
- We have developed FLAIR-Fusion, a computational tool to identify gene fusions and their full-length isoforms

How FLAIR-Fusion Works



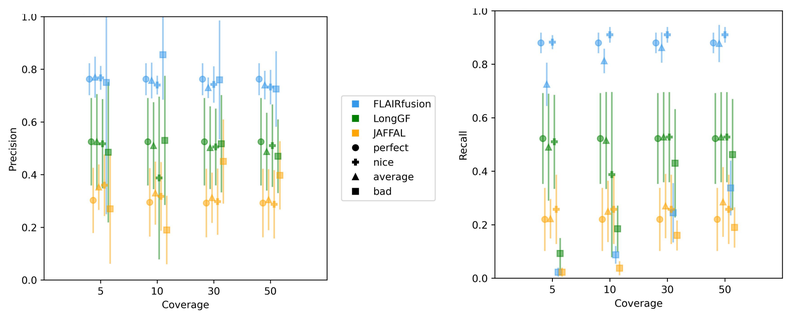
- After reads are aligned and splice sites are corrected using FLAIR-align and FLAIR-correct, reads that align to multiple loci are identified (top panel).
- Next, multiple filters are applied to separate mapping or library preparation errors from true fusions. A subset of key filters are shown: ensuring genomic distance between mappings, checking that the mappings don't include overlapping sequence, and checking that the breakpoint between the mappings is at a splice site (middle panel).
- Finally, isoforms are identified separately for each locus in a fusion and then combined to create full-length gene fusion isoforms (bottom panel).

Performance on Simulated Reads



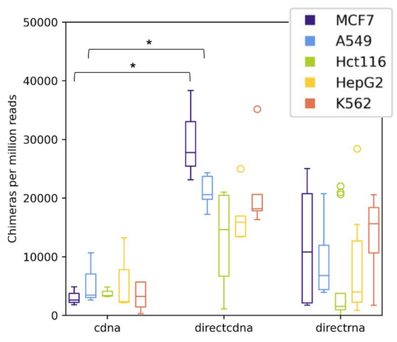
- Using Badreads, we simulated nanopore reads of different levels of quality - error-prone (very bad), average, and nice - at multiple levels of coverage.
- Simulated set of 50 human gene fusions with a background of 8,000 expressed genes
  - Each fusion has 1-10 full-length isoforms
- **Result:** FLAIR-fusion requires high coverage to detect gene fusions from low quality reads, but with high quality sequencing it has high precision and recall even at low coverage
- FLAIR-fusion also detects the full-length fusion isoforms with similarly high precision and recall

Comparison of long-read gene fusion detection tools



On this dataset, FLAIR-fusion outperforms two other long-read fusion detection tools, JAFFAL and LongGF in all cases except recall in bad quality simulated reads

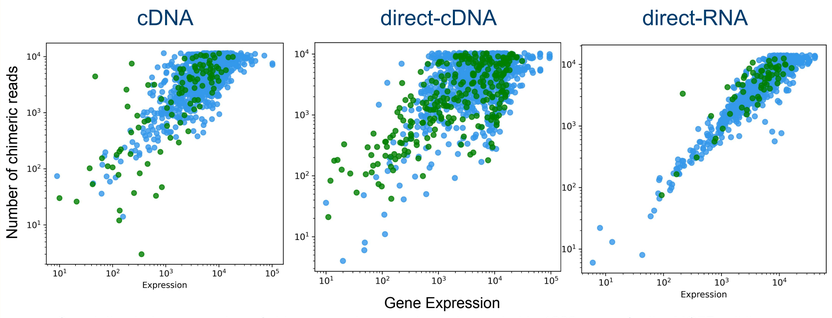
Effect of Library Prep Method on Chimeras



2-5% of all reads in most genomic and transcriptomic sequencing align chimerically (to multiple loci) - Why?

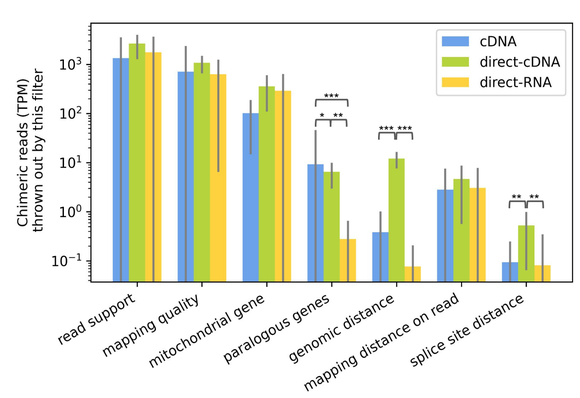
Total number of chimeras per million reads in aligned reads from nanopore sequencing is not significantly impacted by the library prep method

- This is consistent in all 5 cancer cell lines analyzed
- This indicates that most chimeras are not due to PCR artifacts, as previously believed, since direct-cDNA and direct-RNA are PCR-free methods
- 25% (SD +-10%) of chimeric reads contain central adapter sequence



Correlation between number of chimeras and expression subsampled to 1000 genes for the MCF7 cell line - each point is an expressed gene, expression is number of reads of that gene. Green is the genes involved in previously identified gene fusions

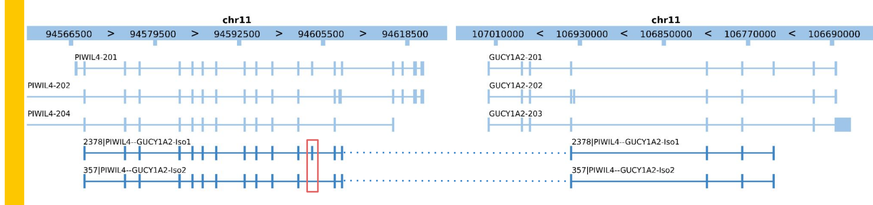
The number of chimeric reads aligned to a gene strongly correlates with total gene expression, indicating that this is a stochastic process not primarily driven by misalignment due to sequence similarity



Library prep method does lead to a different population of chimeras

- direct-RNA has less chimeras due to paralogous mappings
  - Possibly due to lower sequencing depth in direct-RNA
  - Less detection of lower-expression genes with paralogs
- direct-cDNA has more chimeras with short genomic distance between the loci

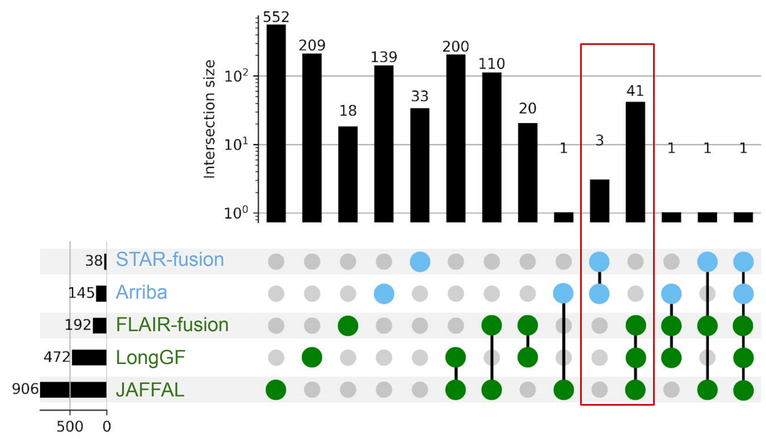
Alternative Splicing of Fusion



Alignment of the fusion isoforms of the amplicon-sequenced PIWIL4-GUCY1A2 fusion. The first number in the fusion isoform label is the number of supporting reads for that isoform. A selection of the annotated isoforms of these genes is also shown with HUGO isoform IDs from gencode 38. Note that there is an inversion between these loci.

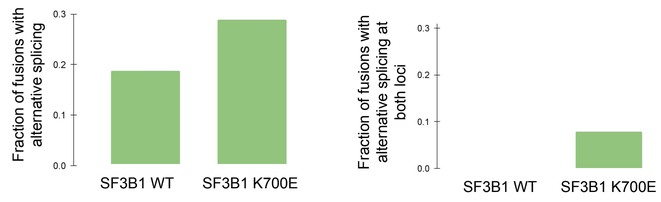
- In amplicon sequencing of the lung adenocarcinoma cell line H322, we detect two unique isoforms of the PIWIL4-GUCY1A2 fusion
- The skipped-exon isoform accounts for 13% of reads at the fusion locus and in frame, which indicates it likely leads to a biologically important amount of expressed protein
- This fusion and its breakpoint have been validated with PCR amplification and Sanger sequencing

Gene Fusions in Chronic Lymphocytic Leukemia



We used nanopore RNA sequencing to sequence chronic lymphocytic leukemia (CLL) tumor samples

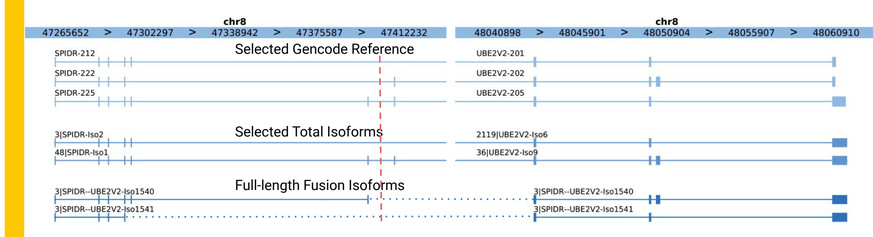
- Short-read methods have a smaller overlap of fusions detected than long-read methods
  - This is important because events detected by more than one tool are more reliable for further analysis
- Of the long-read methods, FLAIR-fusion detects the least unique fusions, indicating less false positives



SF3B1 K700E is a recurrent mutation in CLL that is associated with a poor prognosis

- This mutation disrupts SF3B1 interaction with SUGP1 in the spliceosome
- SF3B1 K700E has been shown to cause global changes in splicing, especially downregulating unproductive transcripts

There is more alternative splicing in gene fusions in the SF3B1 K700E samples, showing that splicing factor mutations can cause novel phenotypes in gene fusions



The SPIDR-UBE2V2 fusion is a novel fusion found in CLL SF3B1 K700E that has a complex alternative splicing phenotype

- Alternative splicing around the breakpoint makes this fusion appear to have two different breakpoints

The SPIDR gene is involved in DNA repair, so this fusion may impact the progression of the tumor

Acknowledgements

Thanks to everyone in the Brooks Lab, especially Alison Tang who developed FLAIR. Thanks also to Cathy Wu at the Broad Institute for the CLL samples.
This project was funded by a grant from the NIH/NHGRI