Genion, an accurate tool to detect gene fusion from long transcriptomics reads

Fatih Karaoglanoglu[1,4], Cedric Chauve[3], and Faraz Hach[2,4]

[1]School of Computing Science, Simon Fraser University; [2]Department of Urologic Sciences, University of British Columbia; [3]Department of Mathematics, Simon Fraser University; and [4]Vancouver Prostate Centre, Vancouver (BC), V6H 3Z6, Canada.

BACKGROUND

The advent of next-generation sequencing technologies empowered a wide variety of transcriptomics studies. A widely studied topic is gene fusion which is observed in many cancer types and suspected of having oncogenic properties. Gene fusions are the result of structural genomic events that bring two genes closely located and result in a fused transcript. This is different from fusion transcripts created during or after the transcription process. These chimeric transcripts are also known as read-through and trans-splicing transcripts. Gene fusion discovery with short reads is a well-studied problem, and many methods have been developed. But the sensitivity of these methods is limited by the technology, especially the short read length. Advances in long-read sequencing technologies allow the generation of long transcriptomics reads at a low cost. Transcriptomic long-read sequencing presents unique opportunities to overcome the shortcomings of short-read technologies for gene fusion detection while introducing new challenges.

APPROACH

From the mapping of transcriptomic long reads to a reference genome, Genion first identifies chains of exons. Reads with chains that contain exons from several genes provide an initial set of reads supporting potential gene fusions.

Given the chimeric read chains, Genion aims to categorize the chimeric chains into three classes: (i) random-pairings, (ii) read-throughs, and (iii) gene fusions candidates. Note that, in the presence of random-pairings created by template switching and segmentation errors, it is difficult to argue a chimeric candidate originates from trans-splicing unless it has significant expression.

METHODS

Preprocessing:

Transcriptomic long reads are mapped to genome using any splice-aware mapper. Using interval trees, Genion converts mappings to sets of segments where each segment is a pair of genomic interval and an exon from gene annotation. Note that, each genomic interval can be paired with multiple exons and a sequence on the read may create multiple genomic intervals (in the case of multi mapping).

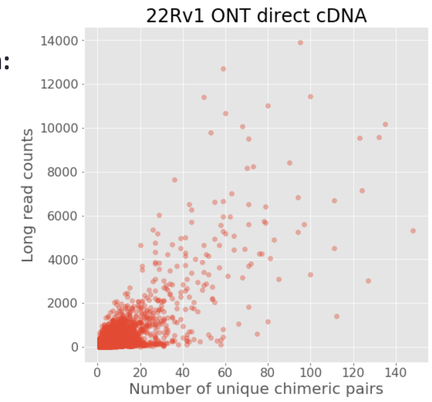
Chimeric Read identification:

In this step we aim to associate to each read an ordered list of non-overlapping unique exons, that we call an exon chain. Exon chains will be used as a coarse encoding of reads in order to cluster reads into groups likely originating from the same isoform.

Such clusters will then be used in two ways: (i) clusters associated to multi-gene exon chains will be considered as a ground set that contains reads from potential gene fusion isoforms, and (ii) clusters associated to single-gene exon chains will be used to calculate the expected expression of individual genes, an important feature to filter out false positives.

Chimeric Cluster Characterization:

In Long RNA sequencing datasets, we observed many background chimeras, we refer to these as random pairings. While majority of these random pairings simply can be filtered by counts, between two abundant genes we can observe many random pairing chimeras.



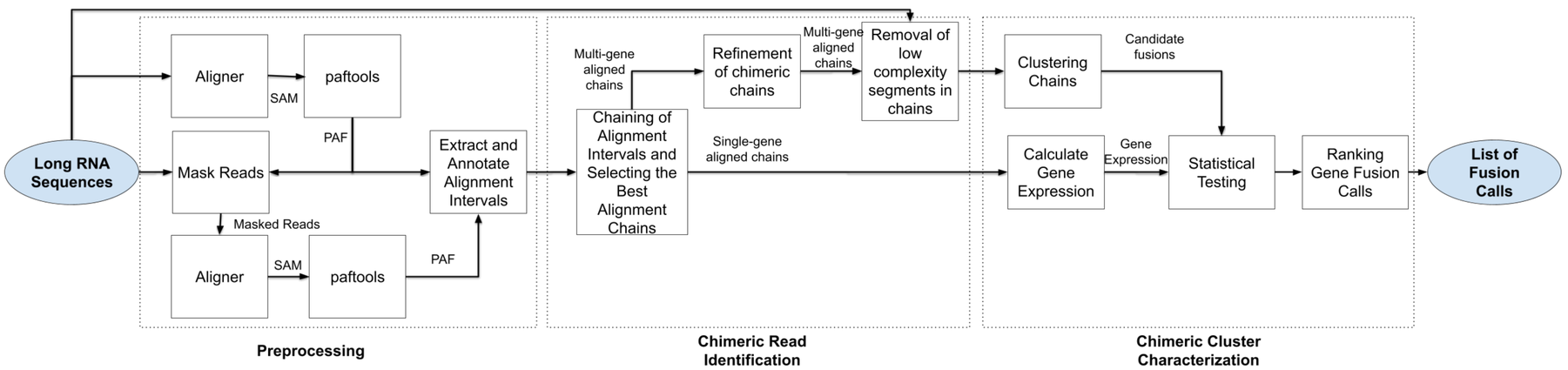
To differentiate gene fusions from random pairings, we introduce a simple statistical model. Let A and B be two unfused genes and A:B the chimera containing both A and B. We expect to sequence p_{rp}*$mean$(|A|,|B|) chimeric A:B molecules between these genes, where p_{rp} is the probability of two molecules attaching derived from global rate of chimeras.

For each chimeric candidate, we test the null hypothesis that $N_{A:B} \sim p_{rp}$*$mean$(|A|,|B|) indicating that number of chimeric reads is not significantly more than what we would expect from random pairing using the Fisher's exact test. Probability of observing $N_{A:B}$ chimeric reads between gene A and B will be:

$$Pr(x = N_{A:B}) = \frac{\binom{n}{N_{A:B}}\binom{n}{n*p_{rp}}}{\binom{n}{n*p_{rp}+N_{A:B}}}$$

Where n=$mean(E_A,E_B)$+$N_{A:B}$ and x is the random variable recording the number of observed chimeric reads.

The p-value for a one-tailed Fisher's exact test is the probability of observing at least $N_{A:B}$ chimeric reads $Pr(N_{A:B} \leq x)$. We use false discovery rate (FDR) control on the p-values computed using the Fisher's exact test. We decided to use Benjamini-Yekutieli procedure due to the dependency between the candidates (caused by shared member genes and global random pairing rate used to calculate expected number of random pairings). This procedure reports the corrected p-value for each chimeric candidate and reports if it rejects the null hypothesis. This ensures the precision to be (1-FDR); note that this is the precision of differentiating random pairings from gene fusions, not the final precision of the gene fusions called by Genion.



RESULTS

Real datasets:

We tested Genion and LongGF on three real dateset: (i) MCF-7 breast cancer cell line sequenced using PacBio (accession: PRJNA277461), (ii) 22Rv1 prostate cancer cell line sequenced in-house on an ONT MinION sequencer and ONT sequencing of NA12878 germline as a negative control.

We know 3 gene fusions validated in the lab and 13 fusions orthogonally validated using short reads + long reads. Genion called 22 gene fusions and 15 read-through events in this dataset including 3 lab validated and 8 short read validated fusions.

To our knowledge there is no validated gene fusions in 22Rv1 cell line. This cell line is known for expressing alternative isoforms of the AR gene. To the best of our knowledge, this cell line does not have any reported gene fusion and we think it is not enriched for gene fusions. Genion called only the ARHGAP15:GTDC1, HOXA5:HOXA6 and KISS1:GOLT1A gene fusions and 12 read-throughs in 22Rv1.

As expected from a negative control, Genion did not call any fusions or read-through in NA12878.

Simulation Experiment:

- Expression profile of inhouse MinION sequencing of 22Rv1 cell line
- Selected the 16 top studied gene fusions from the Cosmic fusion database and simulated fusion transcripts using the most common breakpoints
- We simulated ~ 3 million long reads and 2900 from fusion transcripts (varying between 10 and 1000 read for each)
- Genion called 15 out of 16 simulated fusions with no false positives. One fusion was filtered out due to overlap between the member genes.

Gene1	Gene2	Genion	Gene1	Gene2	Genion
TCF3	PBX1	✓	KMT2A	MLLT3	✓
JAZF1	SUZ12	✓	BCR	ABL1	✓
DNAJB1	PRKACA	✓	TMPRSS2	ERG	✓
KIAA1549	BRAF	✓	CCDC6	RET	✓
NAB2	STAT6	✗	CBFA2T3	GLIS2	✓
EWSR1	FLI1	✓	PML	RARA	✓
SS18	SSX1	✓	RUNX1	RUNX1T1	✓
COL1A1	PDGFB	✓	CRTC1	MAML2	✓

Code:

Poster:

Paper:

NSERC CRSNG