Post alignment filtering of splice junction defining reads

Overview

High-throughput transcriptome sequencing (RNA-Seq) provides a powerful tool to study alternative splicing and facilitate gene structure prediction. The accurate identification of intron-spanning reads represents a critical step in reconstructing transcripts especially in intron rich species. While recent years have seen a plethora of tools developed for mapping transcript reads to a reference genome the alignment of short NGS reads remains an extremely challenging task. Splice aware aligners show considerable variation in alignment yield, tolerance for mismatches or low coverage and novel junction identification (Engström et al. 2013). With high-throughput sequencing generating hundreds of millions of reads per sample even low false positive rates can lead to thousands of false positive splice junctions being identified. The incorrect identification of splice junctions negatively impacts transcript reconstruction and gene annotation, inflating numbers of alternatively spliced transcripts, providing support for incorrect gene merges, and leading to incorrect calls for intergenic and anti-sense transcripts. A number of metrics have been suggested to identify true splice junctions from false positive splice junctions. While a subset of these have been applied within individual alignment programs or incorporated into splicing analysis toolkits (Sturgill et al. 2013) tools for post alignment assessment and filtering of junctions and junction containing reads remain limited. We propose XXXX a command line utility for assessing and filtering spliced RNA-Seq reads using qualitative and quantitative filtering strategies that can be customized to individual needs and provides the required output files for transcript reconstruction tools such as cufflinks and augustus.

http://www.nature.com/nmeth/journal/v10/n12/pdf/nmeth.2722.pdf
http://www.biomedcentral.com/content/pdf/1471-2105-14-320.pdf

Tool outline

Unix command line tool

Input = bam file/s, genomic reference fasta file
Output = bam file containing filtered reads (spliced and unspliced), bed/GFF file of filtered spliced reads and associated junctions, txt output (for each spliced read and junction giving full details of each of the assessed metrics. It would also be useful to output the reads/junctions failing filter as separate bam,gff and bed files.

Each metric (below) would be a separate parameter to allow custom filtering, but we would also want a parameter giving some predefined filtering i.e. a combination of metric parameters that we set as low, medium and high stringency (the default being the low stringency filter)

It would be useful to allow multiple bam files as input, these could represent different samples or the output from different aligners. The logic in the proposed tool would either assess all samples collectively (i.e. as if a single bam had been provided) or independently (i.e. output a bam/bed etc for each input bam file). There are some additional options around this that would have use such as outputting the intersection or union of all junctions found in multiple bams (e.g. give me only junctions shared in all tissues) or those found in more than a specific number of input bams (e.g. junctions shared among at least two methods). Where multiple bams are provided we would need to distinguish between A) bams relating to different samples (where read names wont overlap) and B) bams relating to different alignments of the same sample i.e. output from different aligners (where read names may overlap), for B) you would want a different logic for how you deal with these when assessing collectively.

The tool would need to extract spliced reads (the cigar containing operation N), and from these define the left and right alignment blocks i.e. for a read spanning a single intron the Lstart, Lend, Rstart, Rend with the junction defined by Lend-Rstart you would also associate to this read specific info required for the metrics below e.g. edit distance, MAPQ, if the same read is mapped to multiple position etc. From the extracted spliced reads you would define the set of unique junctions and associate these to the specific reads that support them. These "junction assemblies" would subsequently have the information for the junction level metrics associated to them e.g. maxMMES, hamming3' etc. Based on the input parameters a set of junctions passing filter are retrieved these together with the reads that support them would be reported in the output files. The simplest way of applying the individual filters would be additively though there is value for example for allowing more stringent criteria for non-canonical splicing junctions and less stringent for canonical splicing.

The filtered bam file can be used as input for cufflinks or similar transcript assembly tools, the junction assemblies

Reported Metrics

1. **nReads = number of reads supporting junction**
2. **Donor and Aceptor motif (e.g. GT-AG)**
3. **Intron size**
4. **maxMinAnchor = maximum of the minimum anchor size of all reads associated to junction**
5. **diffAnchor = difference between min L/R anchor and max L/R anchor with the smaller of the two values reported.**
6. **Entropy = Entropy calculation described in Graveley et al., 2011**
7. **nDistinctAnchors = number of distinct L/R Anchors with the smaller of the two values reported.**
8. **nNonredundantReads = number of non-redundant supporting reads. Non-redundant is defined as a read that is not mapped to exactly the same assembly position.**
9. **nUniq = number of uniquely mapping reads supporting junction**

10. **nUp = number of upstream supporting reads (i.e. non spliced reads overlapping the left anchor)**
11. **nDown = number of downstream supporting reads (i.e. non spliced reads overlapping the right anchor)**
12. **maxMMES = Minimal Match on Either Side of Exon junction, as defined in Wang L, Xi Y, Yu J, Dong L, Yen L, et al. (2010) A Statistical Method for the Detection of Alternative Splicing Using RNA-Seq. PLoS ONE 5(1): e8529. doi:10.1371/journal.pone.0008529**
13. **hamming3 = Edit distance between the 5-prime intron sequence and the 3-prime exon anchor (Sturgill et al. *BMC Bioinformatics* 2013, 14:320 doi:10.1186/1471-2105-14-320)**
14. **hamming5 = Edit distance between the 3-prime intron sequence and the 5-prime exon anchor (Sturgill et al. *BMC Bioinformatics* 2013, 14:320 doi:10.1186/1471-2105-14-320)**
15. **covScore = Coverage score as described in Li et al. Nucleic Acids Research Volume 41, Issue 4Pp. e51**
16. **UniqueJunction = True/False is junction unique i.e. are there other junctions sharing acceptor and donor sites.**
17. **PrimaryJunction = True/False Is Junction the primary junction i.e. of junctions sharing either acceptor and donor sites does this junction have the greatest coverage (nReads).**

Metrics 4,6,12,13,14 are implemented in spankijunc
Metric 6 described in http://www.nature.com/nature/journal/v471/n7339/extref/nature09715-s1.pdf (section Splice Junction Discovery and Validation)
Metrics 8,9,10,11 are described in http://www.stanford.edu/group/wonglab/doc/RNA-seq-talk-JSM2010.pdf (section Parameters for junction quality) and http://www.stanford.edu/group/wonglab/SpliceMap/filters.html
Metric 15 is described in http://nar.oxfordjournals.org/content/41/4/e51.long (section RNA-seq mapping derived features)

Input Parameters

Input files (bams, fasta)
Each of the above metrics would have an associated parameter i.e. nReads = min number of supporting reads required for a junction to be reported, maxMMES = min maxMMES score for a junction to be reported etc