PORTCULLIS

Fast, reliable and accurate splice junction prediction from mapped RNAseq data

Daniel Mapleson, Luca Venturini and David Swarbreck

Introduction

Correct identification of splice junctions (SJs) is critical step for alternative splicing (AS) analysis. The primary route for gathering a set of splice junctions from RNAseq datasets is via RNAseq mappers, however the RGASP consortium established that RNAseq mappers tend to generate many false junction predictions. One reason for this is that many RNAseq mappers are optimised to produce the best read placement possible in the shortest time and sequencing errors close to splice junctions can easily trigger misalignments that produce invalid junctions. Accurate junction prediction is a secondary consideration. Some mappers such as Tophat and STAR do produce a set of filtered junctions by post-processing the alignment data, however their approach is relatively unsophisticated. Our results show that while few genuine junctions are lost, many invalid junctions remain. There is a wealth of information present in the mapped RNAseq data and the genomic loci surrounding the predicted splice junctions that can be exploited for more effective SJ filtering. Some tools exist to do this, although they are either impractically slow, will only work with BAMs produce by a specific aligner, or redundantly perform the alignment task themselves. In this poster, we introduce Portcullis a tool for rapidly and accurately filtering invalid splice junctions from any given BAM file. Furthermore, producing a set of accurate junctions is useful for many downstream applications that would otherwise be biased, such as alternative splicing analysis, transcript reconstruction and gene modelling.

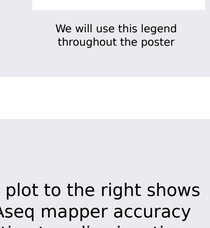
Generating simulated data

We generated 4 core simulated datasets for our experiments, using SPANKI. Each dataset contains paired end fastq files containing the reads, a BAM file with perfect alignments and a BED file containing a definitive set of junctions. Each simulated dataset is derived from a real dataset and replicates the error and expression profiles. In addition, we generated several human datasets with varying depth and quality.

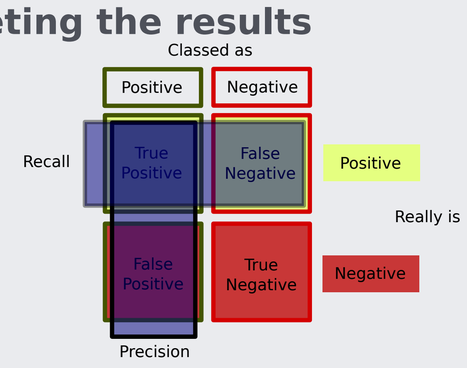
Properties of simulated dataset	Arabidopsis	Drosophila	Human	Mouse
Original accession	PRJEB7093	SRA009364	PRJEB4208	?
# reads (M)	93	47	46	12
Max read length (bp)	100	76	50	76
# splice junctions	109,989 (86% of ref)	29,275 (51% of ref)	158,156 (48% of ref)	96971 (33% of ref)
Mean Quality in error model	37	37	29	33

Interpreting the results

$$Recall = TP/(TP+FN)$$
$$Precision = TP/(TP+FP)$$
$$F_1 = 2TP/(2TP+FN+FP)$$

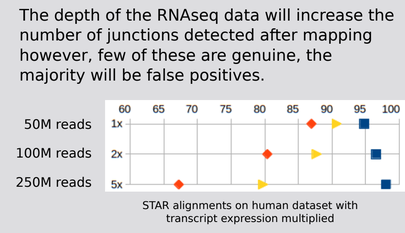


We will use this legend throughout the poster



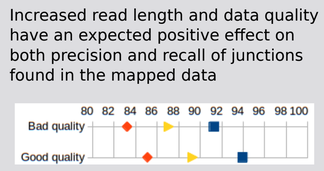
Junction level errors in mapped RNAseq data

The depth of the RNAseq data will increase the number of junctions detected after mapping however, few of these are genuine, the majority will be false positives.



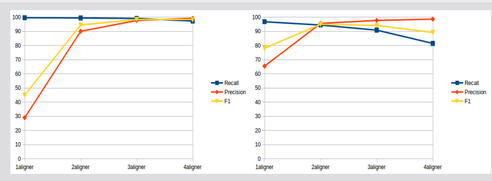
STAR alignments on human dataset with transcript expression multiplied

Increased read length and data quality have an expected positive effect on both precision and recall of junctions found in the mapped data



STAR alignments on human dataset
Good = 100bp @ ~37QS; Bad = 50bp PE @ ~29QS

This 5-way Venn diagram of simulated human data shows that while recall is good for all mappers, precision is relatively poor and each aligner has a large set of unique false positives. We can exploit this observation to generate a set of high-confidence junctions by seeking a consensus between aligners.

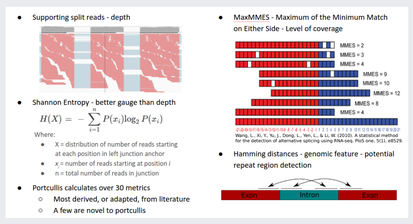


Results when varying degree of consensus when merging aligners. Arabidopsis (left), Human (right). Optimal results with the Arabidopsis dataset is achieved with 3 aligners agreeing, the human dataset with 2.

Finding a consensus between aligners results in very high accuracy, although there are several downsides to this approach, such as computational efficiency and dependence on aligner quality.

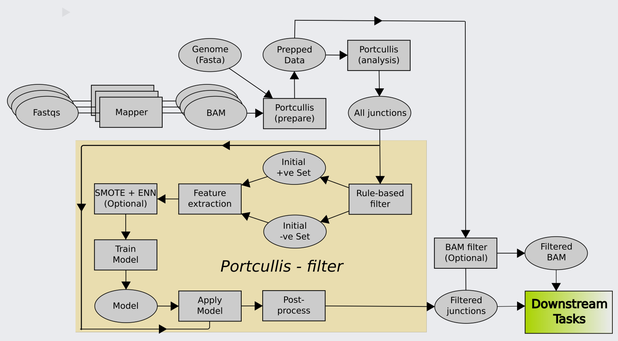
Junction Analysis

By analysing the genomic loci and mapped reads around splice junctions we are able to extract useful information that allows us to discriminate between genuine and invalid junctions. The simplest metric is the number of raw split reads supporting a junction, although it turns out there are much better features to use.



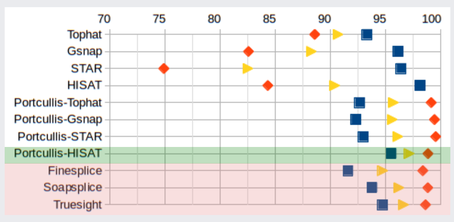
Pipeline

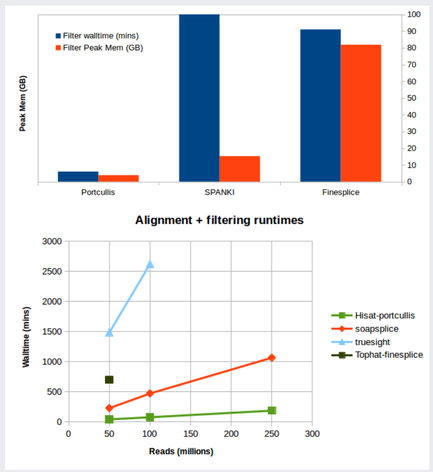
A genome file in FastA and one or more mapped RNAseq datasets in BAM format are passed into portcullis as input. The alignments are prepared and the analysed and all potential junctions (those tagged with an 'N' refskip cigar op) are extracted and analysed with respect to their context, such as other supporting alignments and genomic region. Rule-based filtering of these metrics generally produces sub-optimal results, although it allows us to determine subsets of positive and negative junctions with high confidence. We train and learner on these datasets then apply it to the full set of junctions to assign a confidence score, which we then filter based on a simple user-defined cutoff.



Results

The plot to the right shows RNAseq mapper accuracy relative to splice junction prediction tools. Despite relatively low depth of the datasets all prediction tools outperform RNAseq mappers in terms of F1, with relatively minor decreases in recall. We expect this situation to be even more stark with realistic, higher depth, datasets. Portcullis outperforms finesplice using any mapper, and is comparable with soapsplice and truesight, beating them when HISAT is used as a mapper.

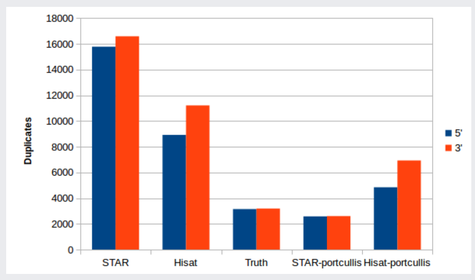




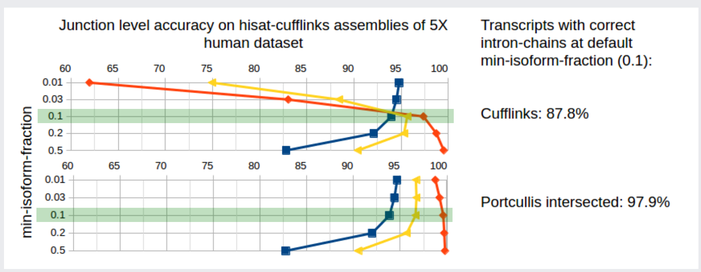
Portcullis drastically outperforms all other competitors in terms of runtime. Finesplice suffers from high runtimes, high memory usage and can only run on BAMs generated by tophat2. Truesight is even slower and also requires large amounts of memory. We did not have the resources to run finesplice and truesight on all our human datasets. Soapslice is ~5X slower than portcullis when HISAT is used as an aligner. In addition, SOAPsplice does not produce a BAM file and appears to be unreliable in some situations (we could not make it run on our Arabisopsis dataset).
Portcullis also produces richer output detailing results for all metrics. It also outputs multiple formats (BED, GFF and TSV) therefore requiring less effort to use downstream.

Downstream effects

Filtered junction from portcullis can improve results downstream. The plot on the right shows the number of duplicated 5' and 3' junction sites for various runs. These would be indicative of alt 5' and 3' splicing events. STAR and Hisat results are shown with and without filtering, and then the truth for the datasets is shown on the far right. Portcullis filtering brings the aligner results much closer to the true number.



The plot below shows the effect of modifying the min-isoform fraction setting in cufflinks. Junction-level recall increases when lowering the value, the opposite effect is seen with precision. By interesecting cufflinks junctions with portcullis filtered junctions we can retain most of the recall, whilst significantly increasing precision. The effect is more significant when looking at transcript-level precision. Even at the default level of 0.1, 12% of cufflinks transcripts contain invalid intron chains. By filtering transcripts based on portcullis junctions we can reduce that to 2%, with the loss of only 73 (or ~1% of) valid transcripts.



In a similar way, junction information can be exploited by gene modellers such as Mikado to guide selection of transcripts.

MIKADO
https://github.com/lucventurini/mikado.git

Availability

Contact: daniel.mapleson@earlham.ac.uk

GitHub

https://github.com/maplesond/portcullis

Read the Docs

http://portcullis.readthedocs.io/en/latest/

GPL v3 Free Software
Free as in Freedom

Summary

- RNAseq mappers produce BAMs containing large numbers of false junctions
- By finding consensus between multiple mappers an accurate set of junctions can be extracted
- Portcullis can get similar and more reliable results with only running a single mapper
- Portucullis is fast, and we think it's the only practical and reliable tool available for doing accurate splice junction analysis and prediction from large RNAseq datasets
- Portcullis can positively impact downstream tasks such as alternative splicing analysis, transcript reconstruction and gene modelling

Credits

David Swarbreck | Luca Venturini

Shabonham Caim | Sarah Bastkowski | Gemy Kaithakottil

Computing infrastructure for Science

NBI CiS

Earlham Institute, Norwich Research Park, Norwich, Norfolk, NR4 7UH, UK

wwwearlham.ac.uk

BBSRC bioscience for the future
EXCELLENCE WITH IMPACT COMMENDATION WINNER
norwich research park
NATIONAL INSTITUTES OF BIOSCIENCE

Earlham Institute
Decoding Living Systems