
Earlham
Institute
Decoding Living Systems

PORTCULLIS

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

Dr Daniel Mapleson
Analysis Pipelines Project Leader

Earlham Institute

Outline

- Splice junction (SJ) detection is the first step in detecting alternative splicing (AS) events. Also, an accurate set of SJs is useful for transcript reconstruction and gene modelling.
- In 2013, the RGASP consortium established that reducing the number of SJ errors is an ongoing challenge for RNAseq mappers
- In this talk:
    - Brief analysis of junction-level variation between RNAseq mappers across different datasets
    - Portcullis - Post-alignment filtering false positive junctions directly from BAM files, and comparison to similar tools
    - Effect of portcullis filtered junctions on downstream tasks

Creating simulated RNAseq datasets

… because real datasets contain an unknown number of genuine junctions

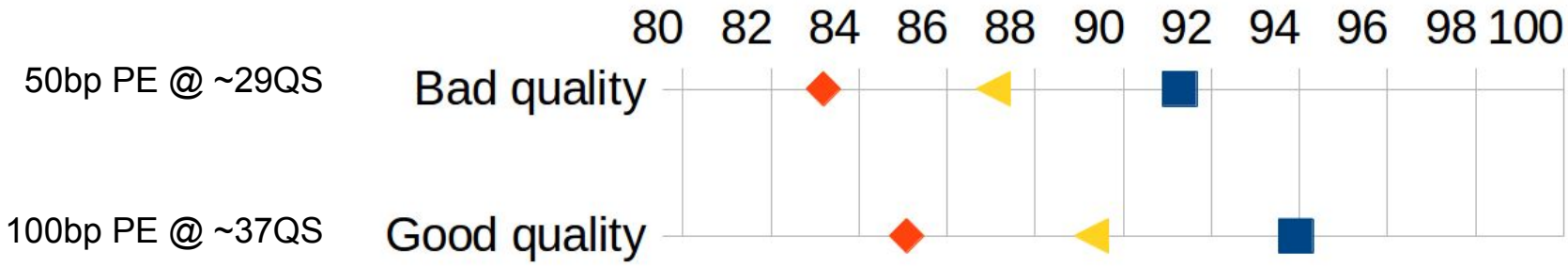
- We need to analyse a range of conditions to reflect real-life scenarios:
  - Expression levels, sequencing depth and quality, genomic (and intronic) properties
- We used SPANKI to generate unstranded simulated PE reads (FastQs), perfect alignments (BAM) and a complete set of true junctions with realistic expression levels and error profiles

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

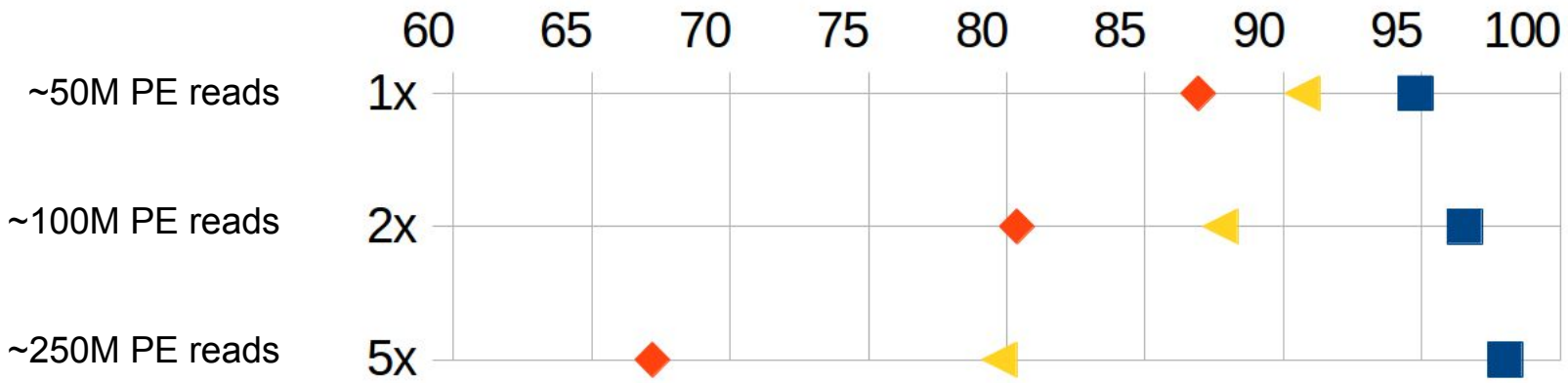
Some dataset properties have consistent effects across aligners

STAR alignments on variants of the Human simulated dataset

Read length and quality:

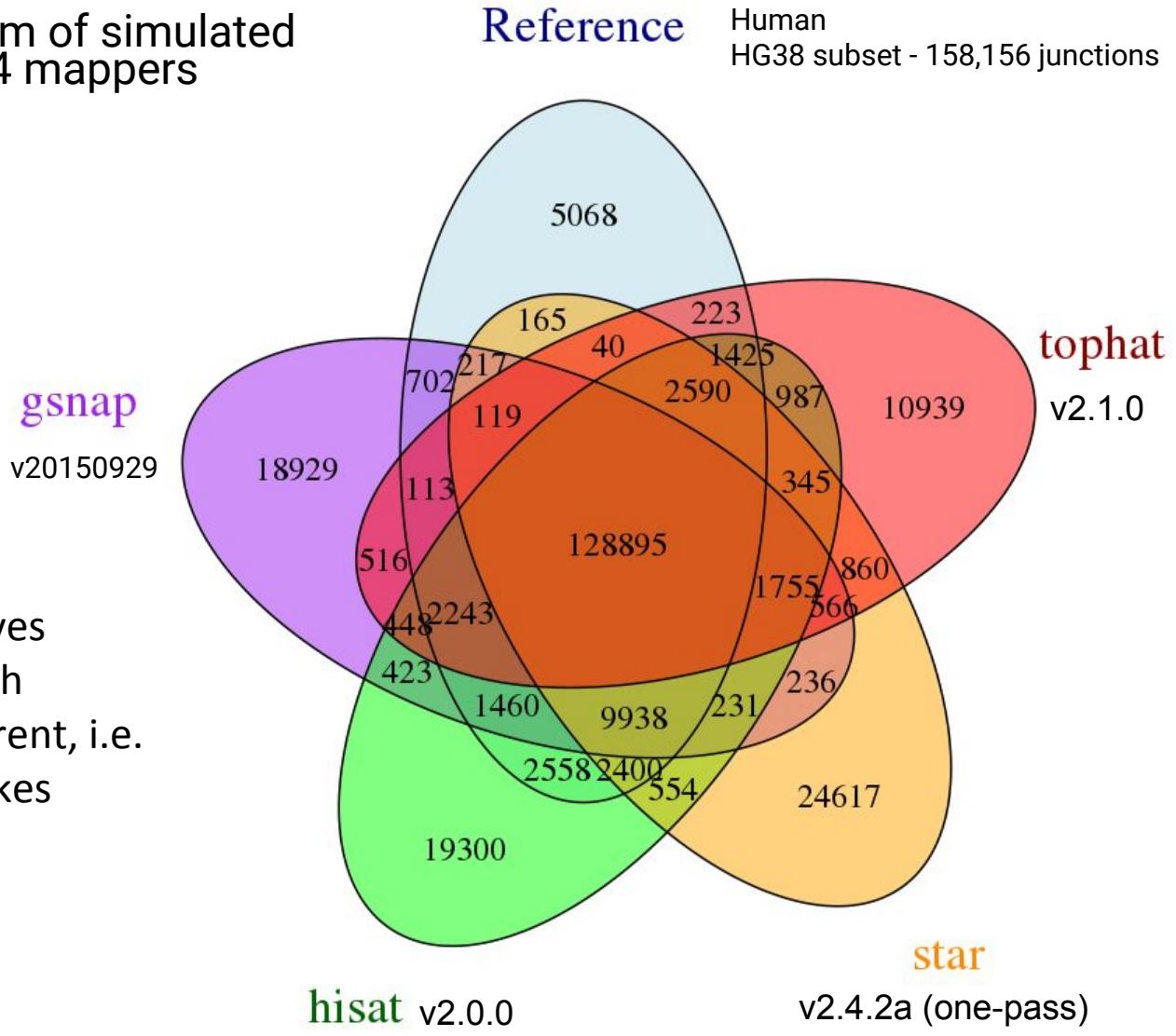


Depth:



False positives are mostly mapper specific

5-way Venn diagram of simulated human data from 4 mappers

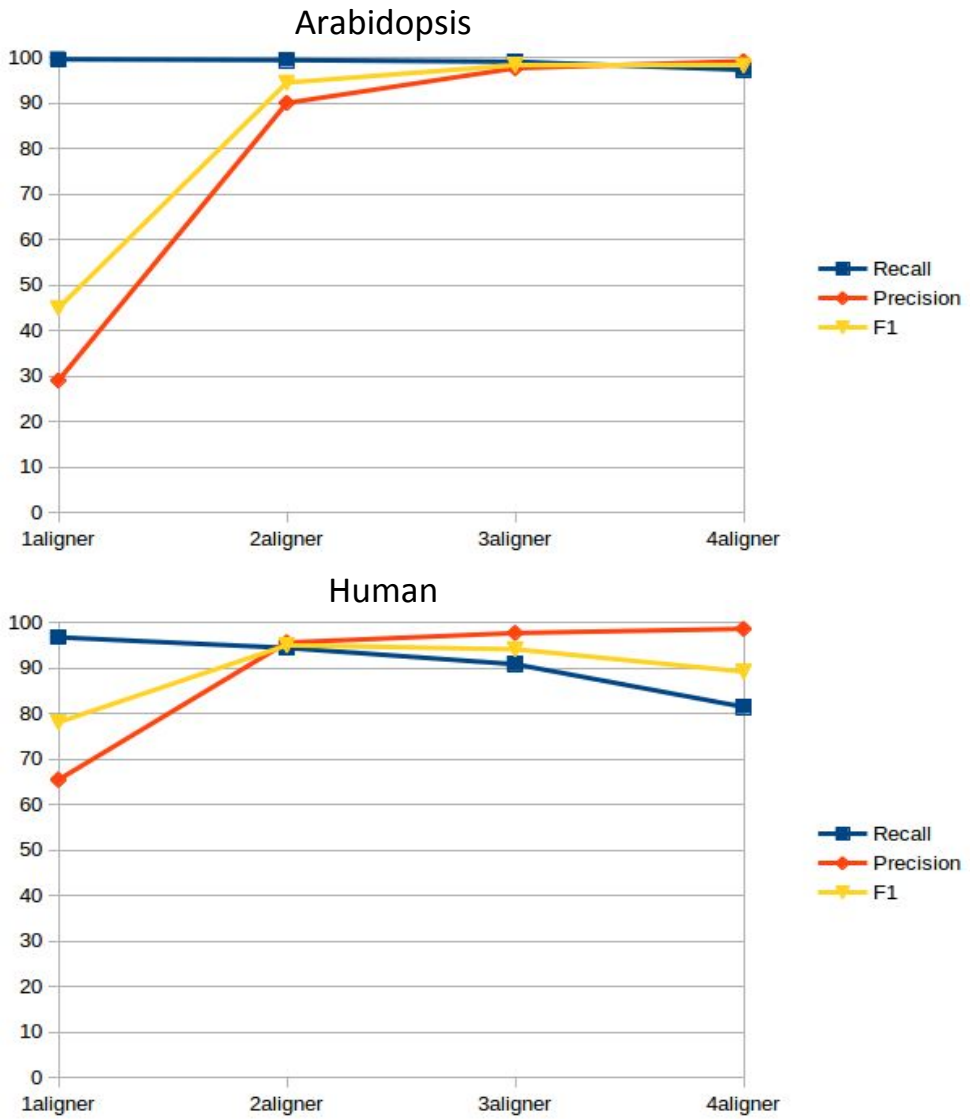


Most false positives generated by each mapper are different, i.e. each mapper makes different errors

Improved SJ accuracy by finding consensus between mappers

But at a price...

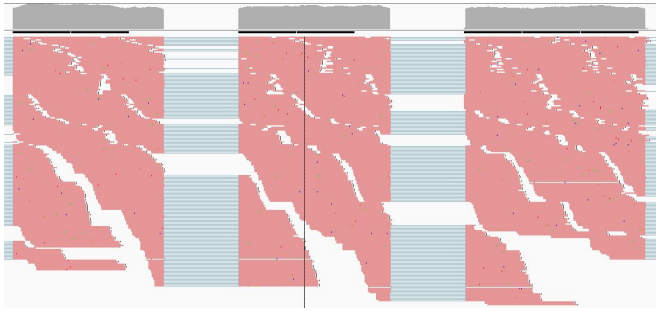
- No way to know what level of agreement will give best results
- Computationally expensive
- No single BAM file to take forward for downstream analysis
- Only works with aligners with good sensitivity (fortunately most do)



Junction Features

A few RNAseq-mapper-derived and genomic features useful for junction validation

- Supporting split reads - depth



- Shannon Entropy - better gauge than depth

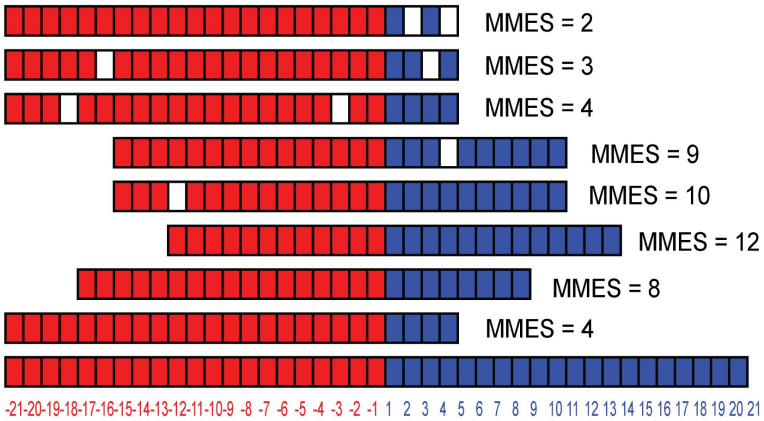
$$H(X) = -\sum_{i=1}^{n} P(x_i)\log_2 P(x_i)$$

Where:

  - X = distribution of number of reads starting at each position in left junction anchor
  - x_i = number of reads starting at position *i*
  - n = total number of reads in junction

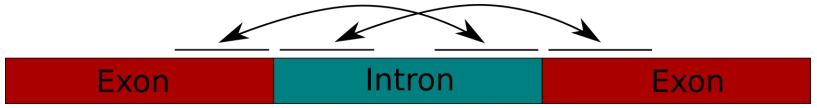
- Portcullis calculates over 30 metrics
  - Most derived, or adapted, from literature
  - A few are novel to portcullis

- MaxMMES - Maximum of the Minimum Match on Either Side - Level of coverage



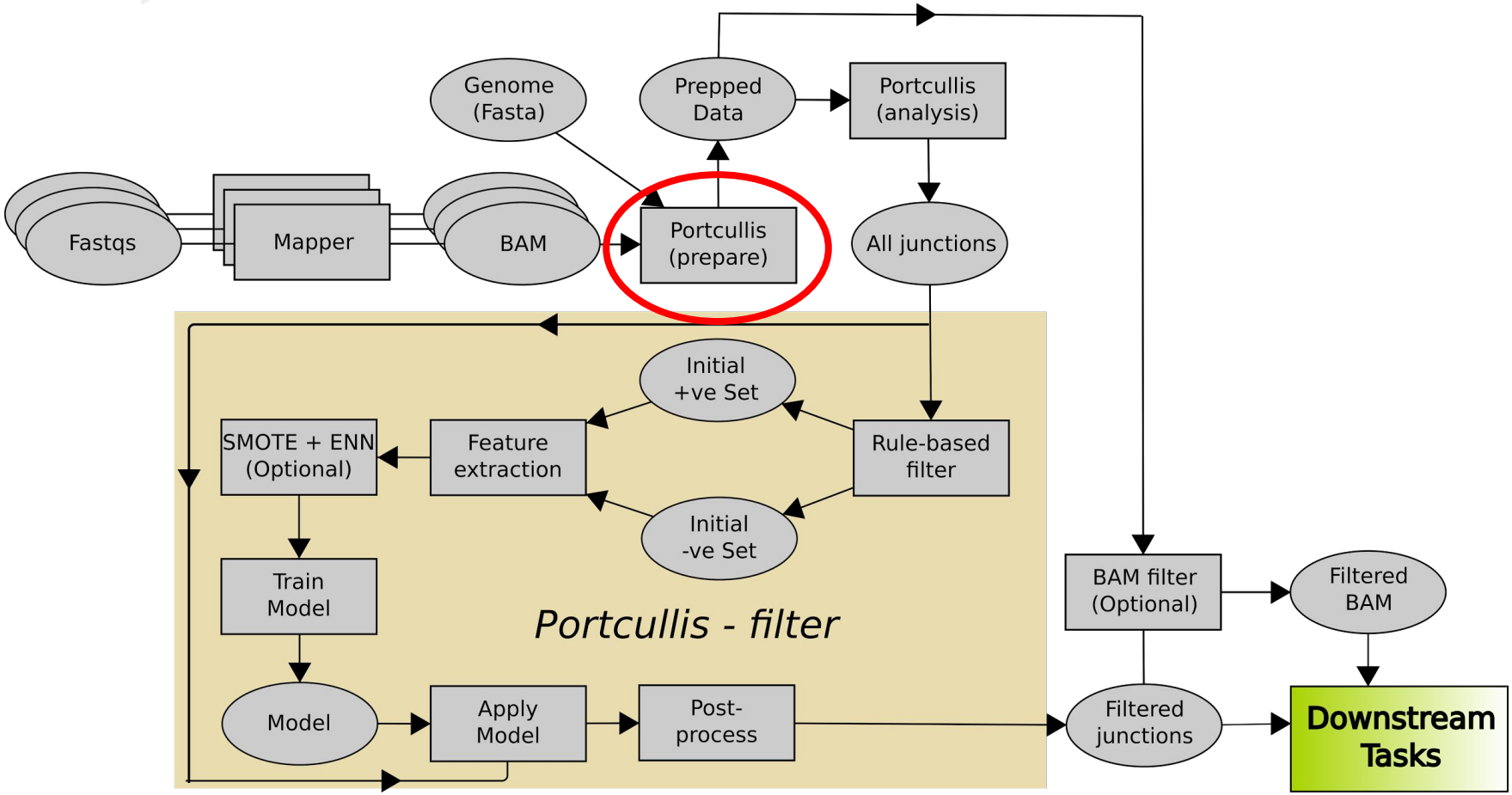
Wang, L., Xi, Y., Yu, J., Dong, L., Yen, L., & Li, W. (2010). A statistical method for the detection of alternative splicing using RNA-seq. PloS one, 5(1), e8529.

- Hamming distances - genomic feature - potential repeat region detection



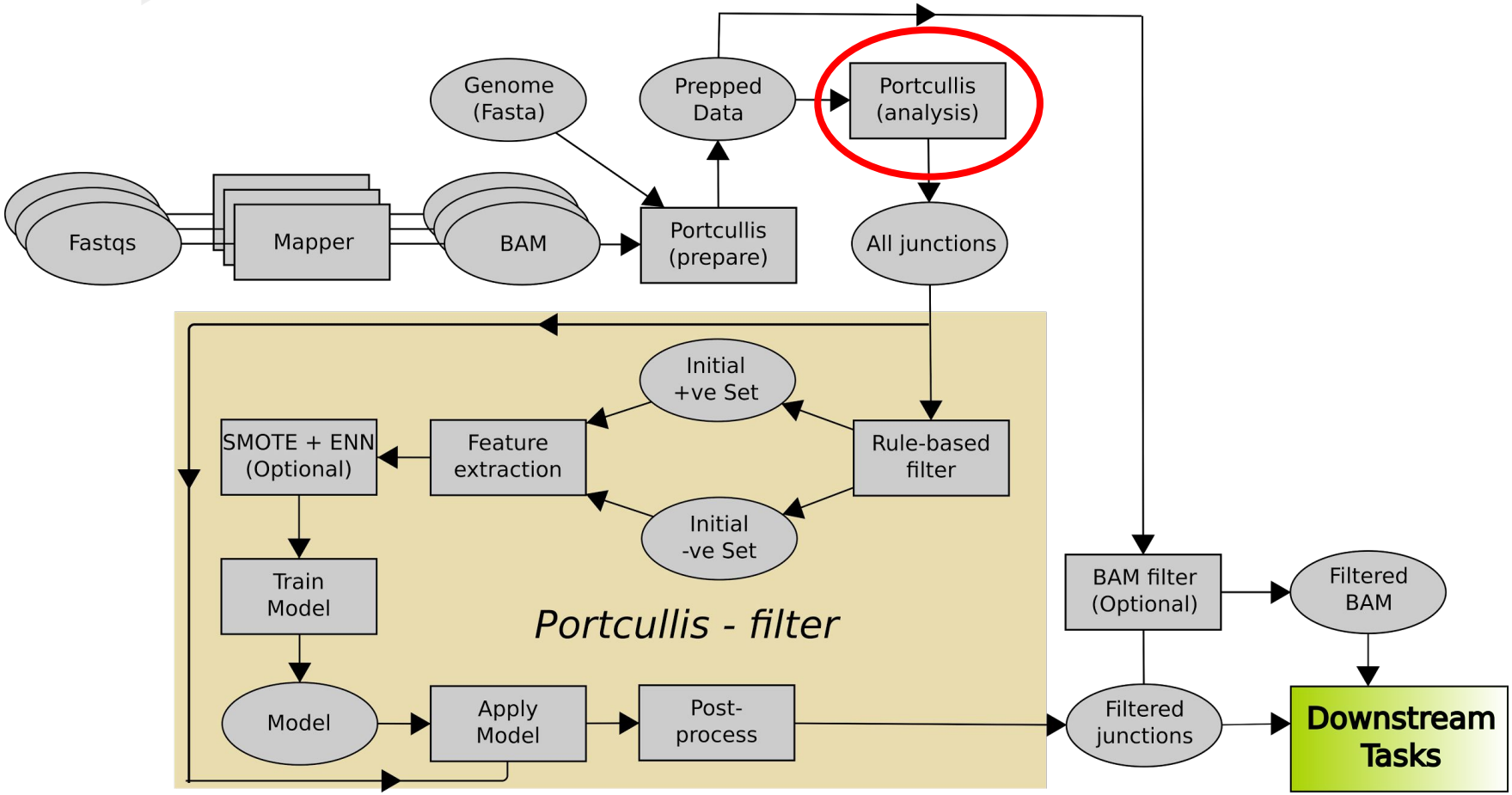
Portcullis Pipeline

Data preparation - BAM merging and indexing



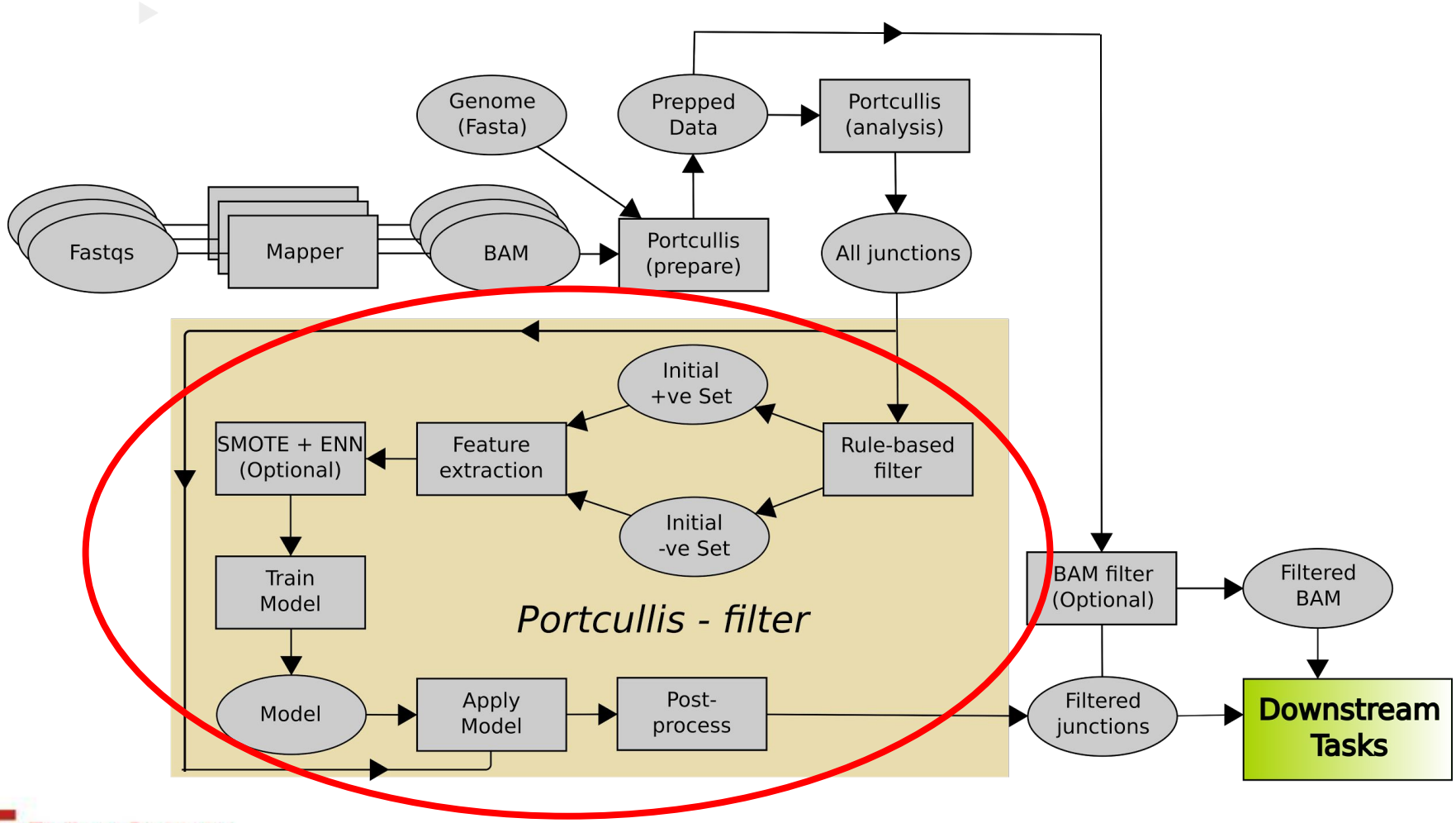
Portcullis Pipeline

Junction Analysis - Calculate values for junction metrics



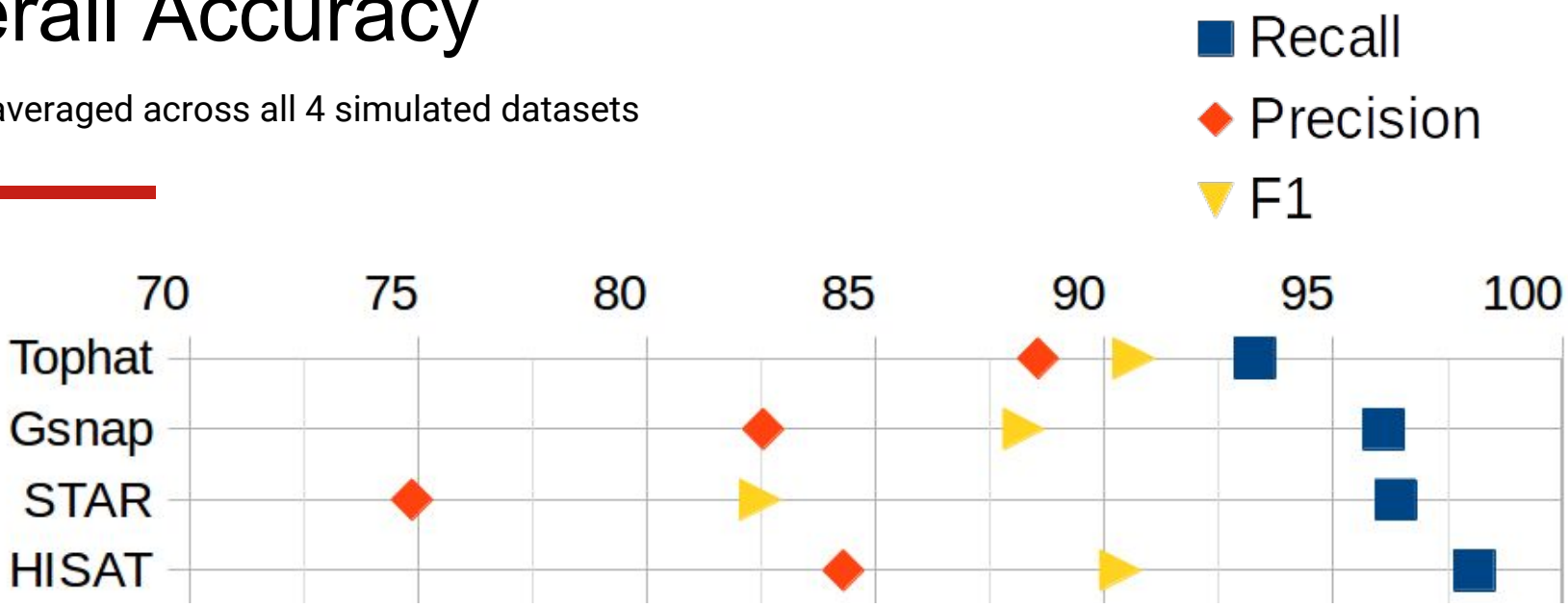
Portcullis Pipeline

Adaptive machine learning filtering - learns each datasets separately



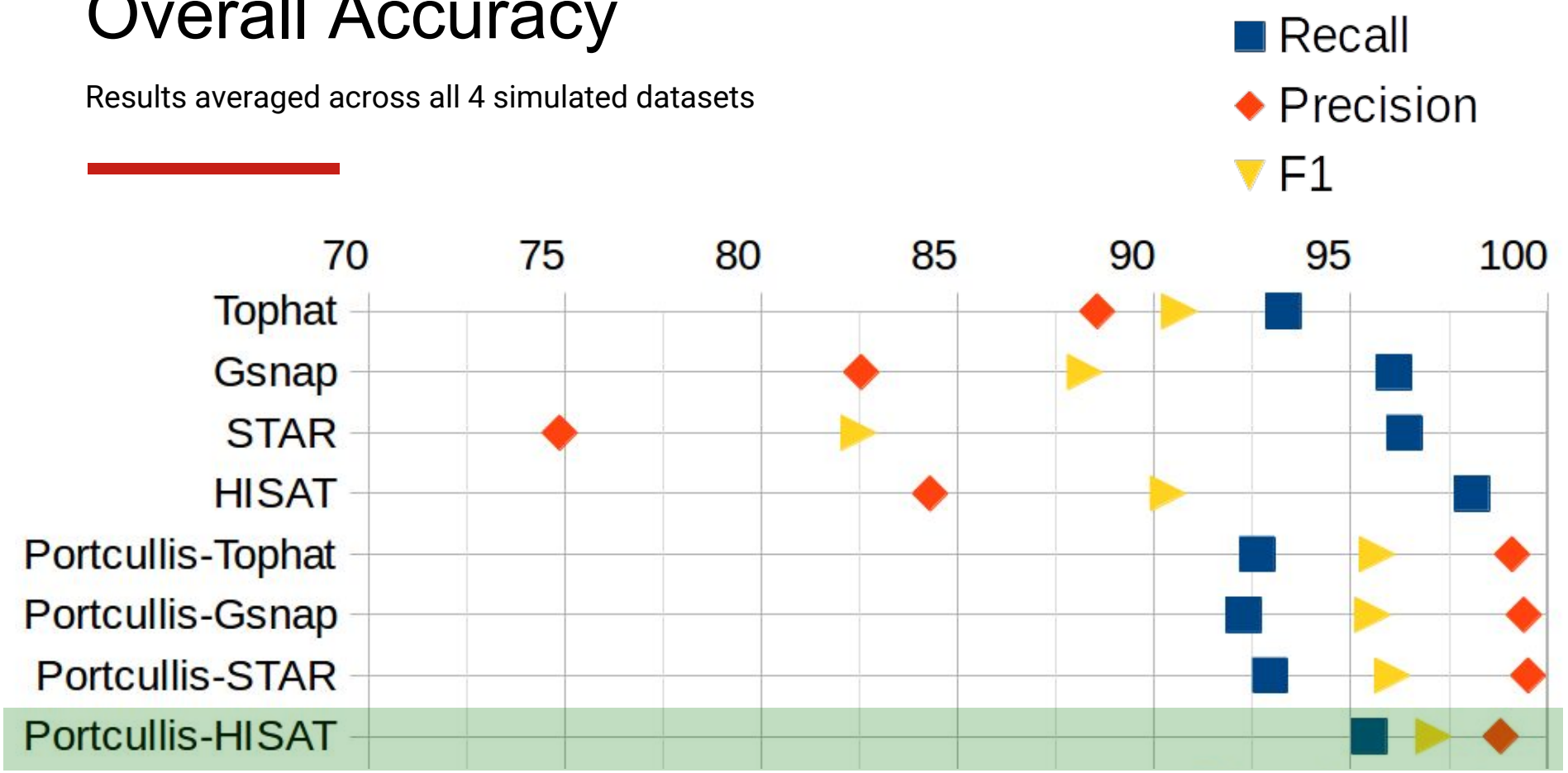
Overall Accuracy

Results averaged across all 4 simulated datasets



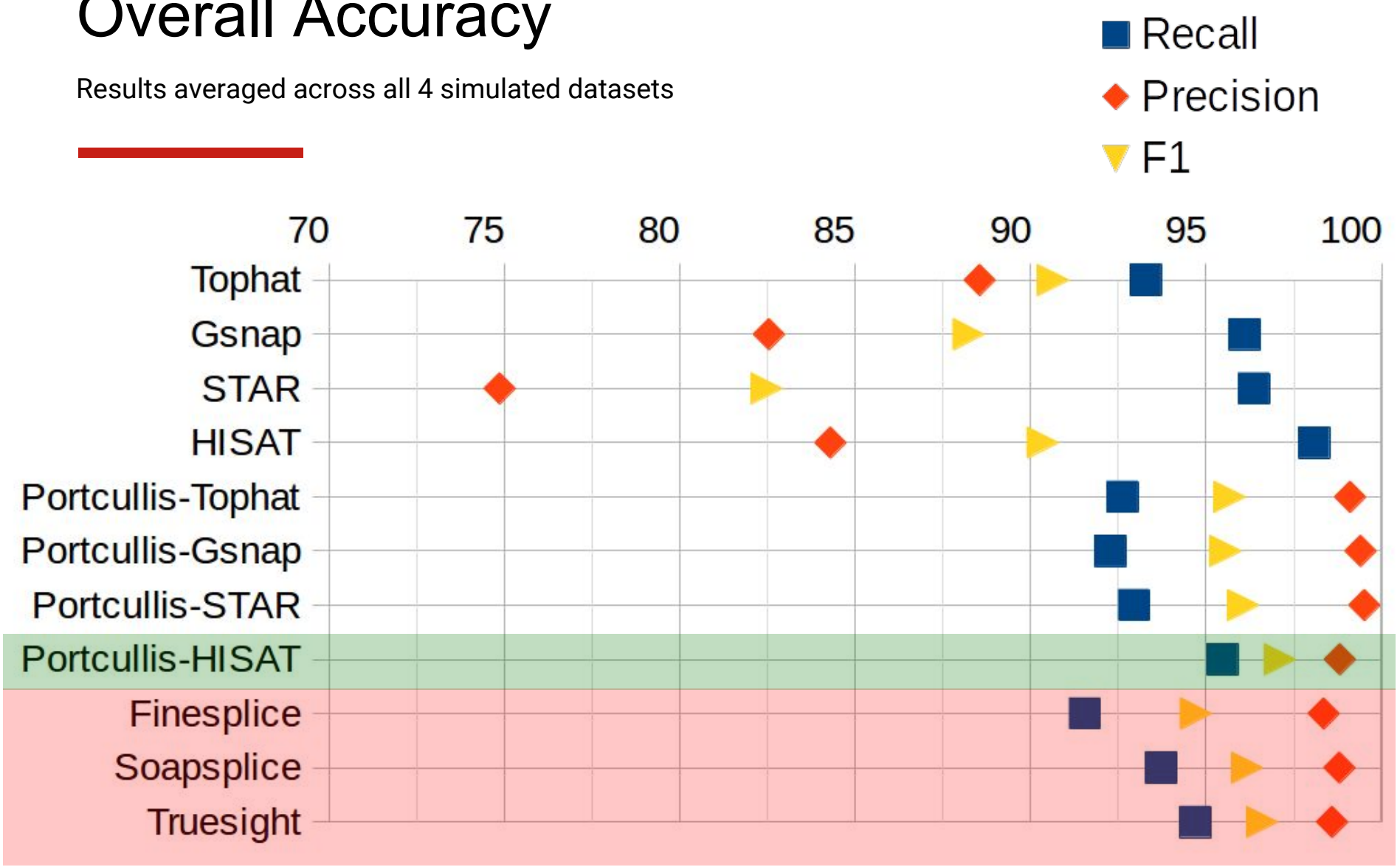
Overall Accuracy

Results averaged across all 4 simulated datasets



Overall Accuracy

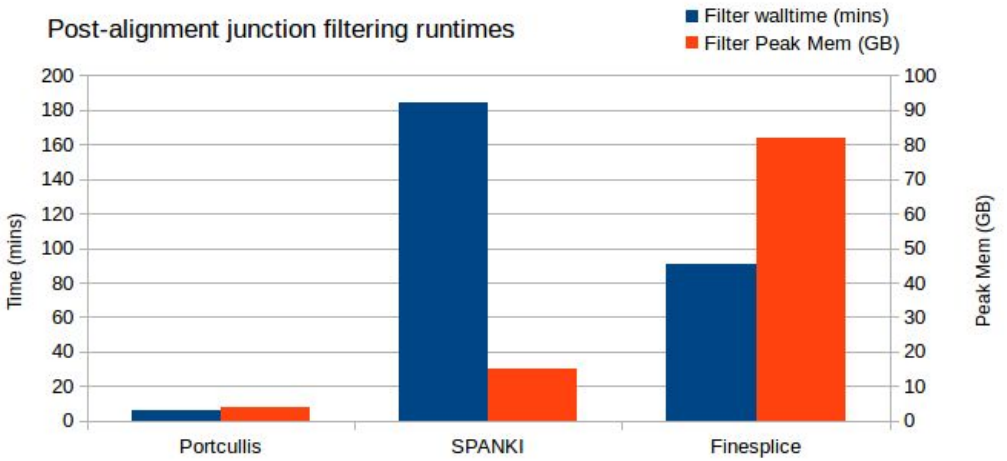
Results averaged across all 4 simulated datasets



Runtime Performance

Massive improvement over post-alignment competitors

Human
Data - 4 threads used where available

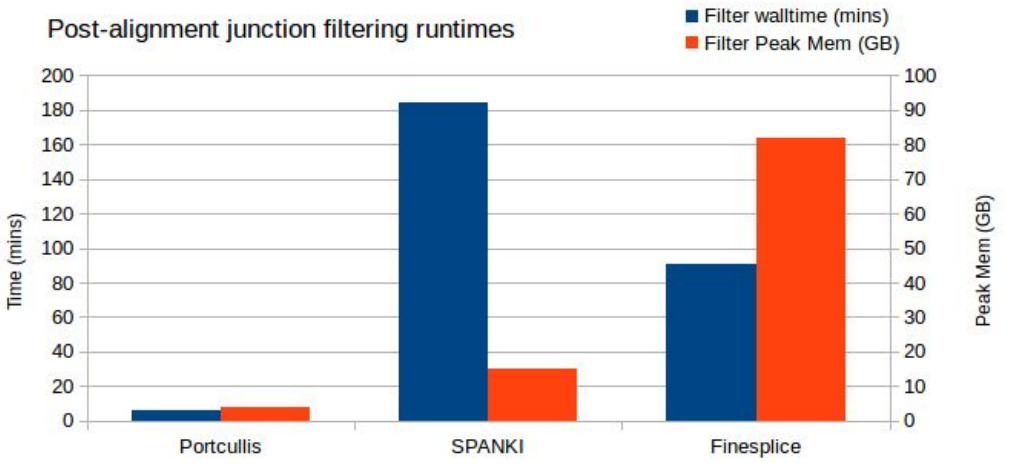


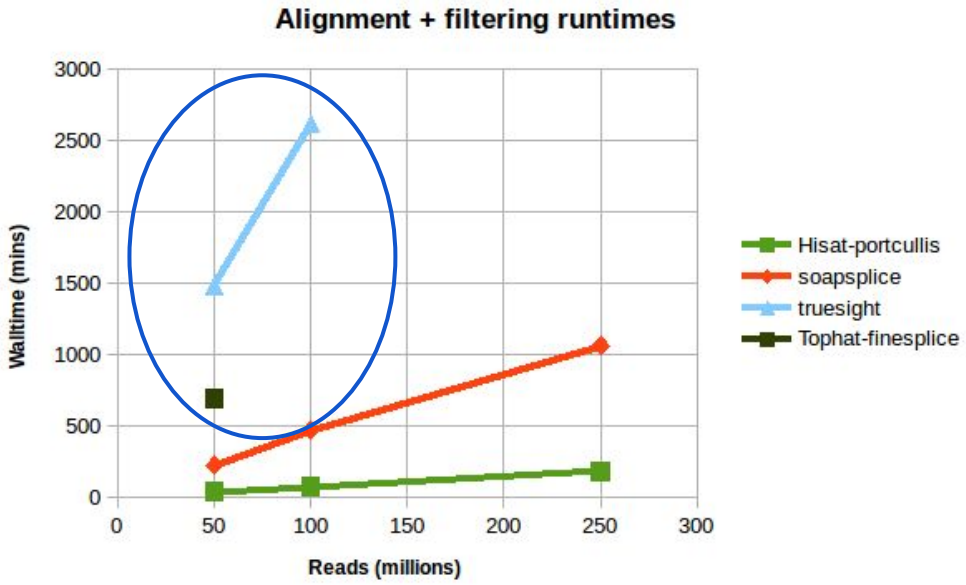
Runtime Performance

Some competitors are slow and require too much memory to be practical

Human
Data - 4 threads used where available

- We did not have enough memory to run Finesplice and Truesight for all cases (>100GB required)



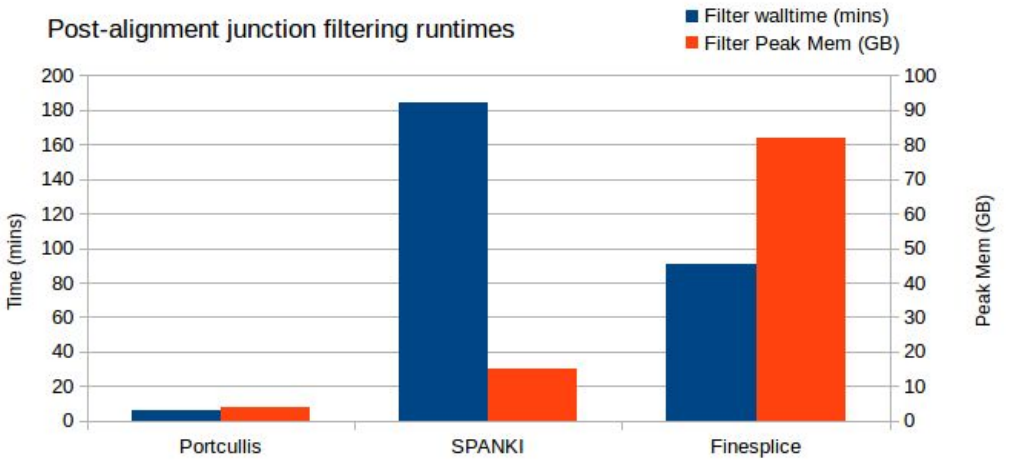


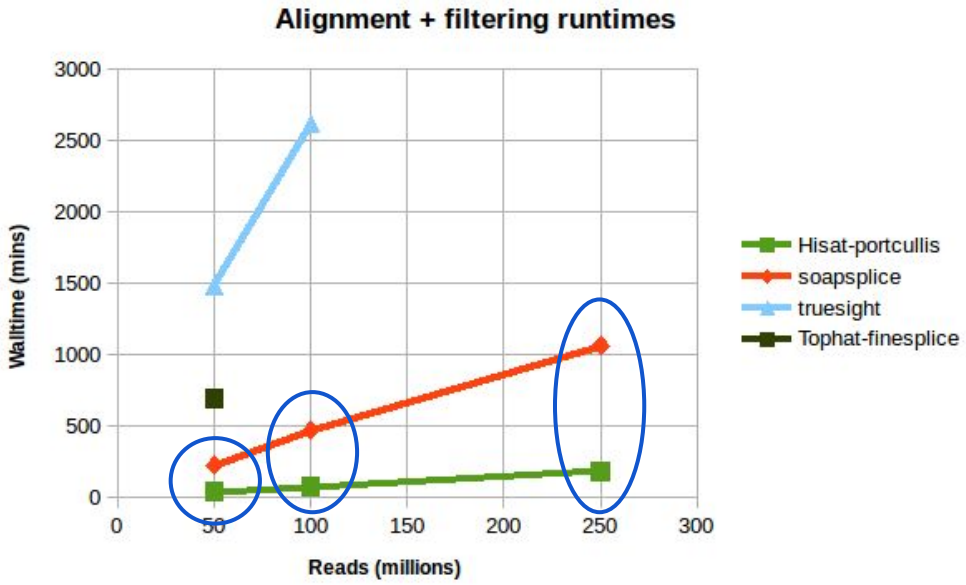
Runtime Performance

5X improvement over soapsplice when portcullis is coupled with HISAT

Human
Data - 4 threads used where available

- We did not have enough memory to run Finesplice and Truesight for all cases (>100GB required)
- Soapsplice runtimes and memory usage are ~5X slower than hisat-portcullis, also we couldn't run it on arabidopsis





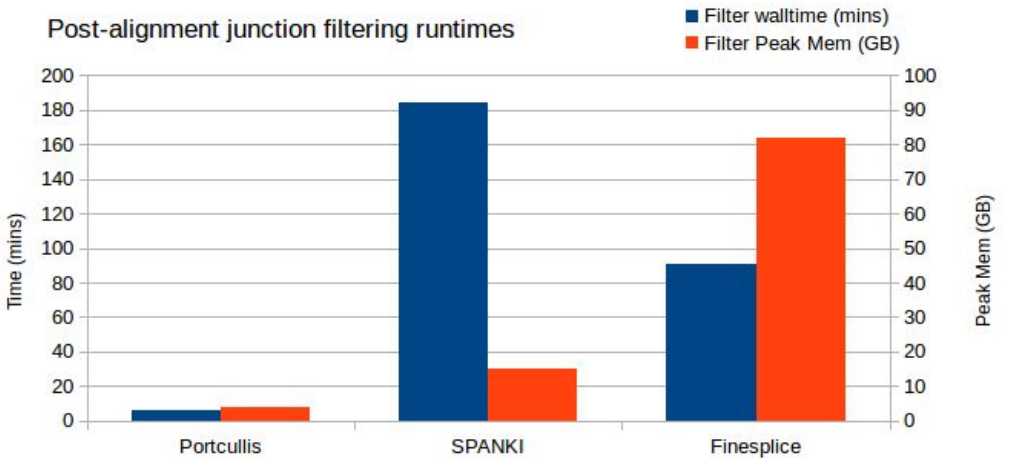
Earlham Institute

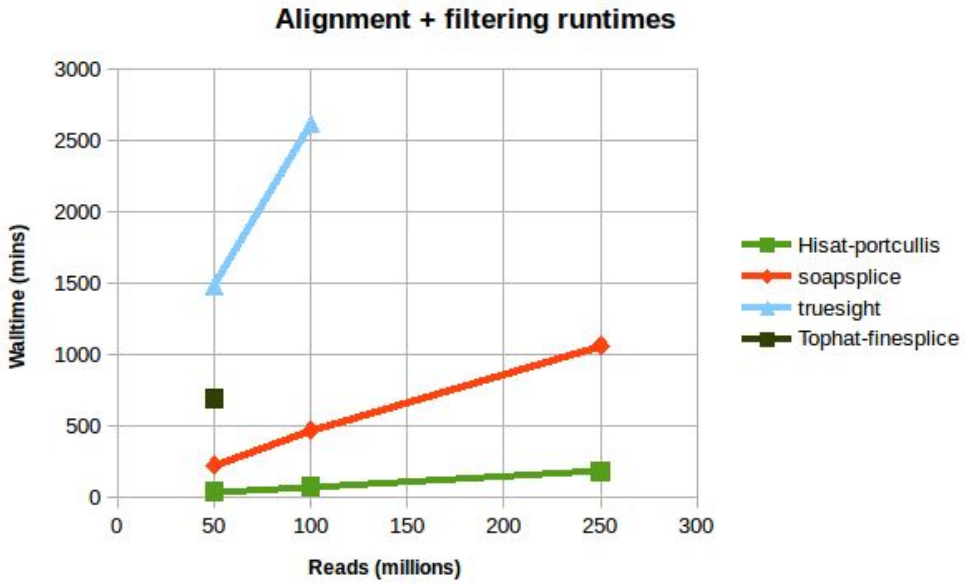
Runtime Performance

Portcullis has coped fine with every dataset we've given it so far

Human
Data - 4 threads used where available

- We did not have enough memory to run Finesplice and Truesight for all cases (>100GB required)
- Soapsplice runtimes and memory usage are ~5X slower than hisat-portcullis, also we couldn't run it on arabidopsis
- Portcullis copes with a fragmented wheat genome. Using 10 threads, processed a 170 million read RNA seq library in < 60mins, using < 20GB RAM

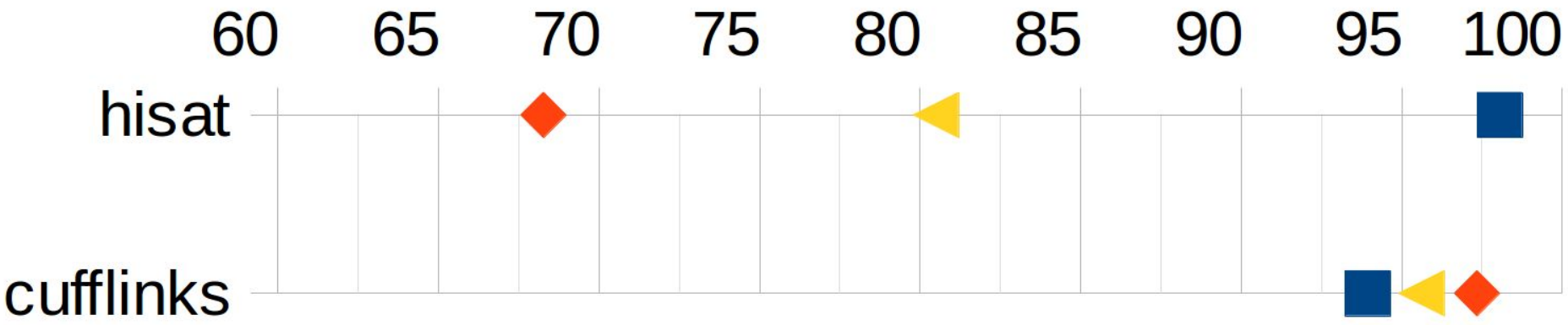




Downstream applications

Transcript reconstruction and gene modelling

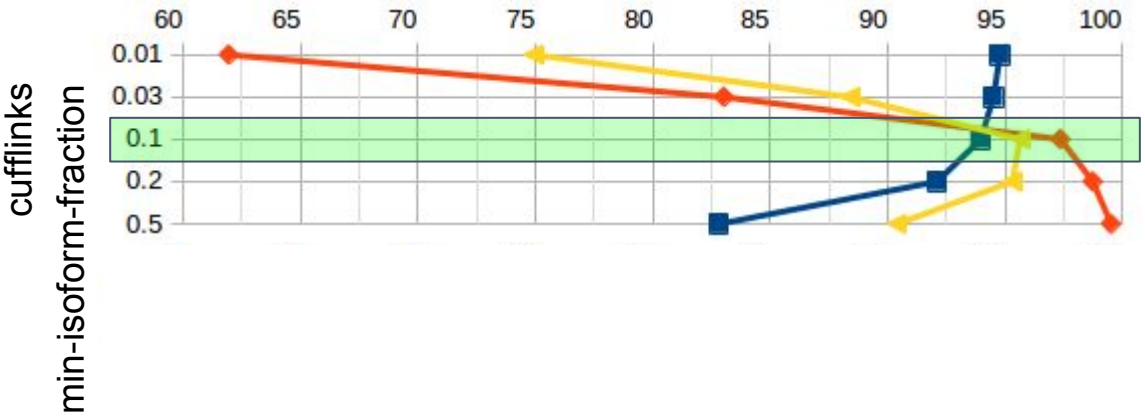
Junction level accuracy on hisat-cufflinks assemblies of the 250M read human dataset



Downstream applications

Transcript reconstruction and gene modelling

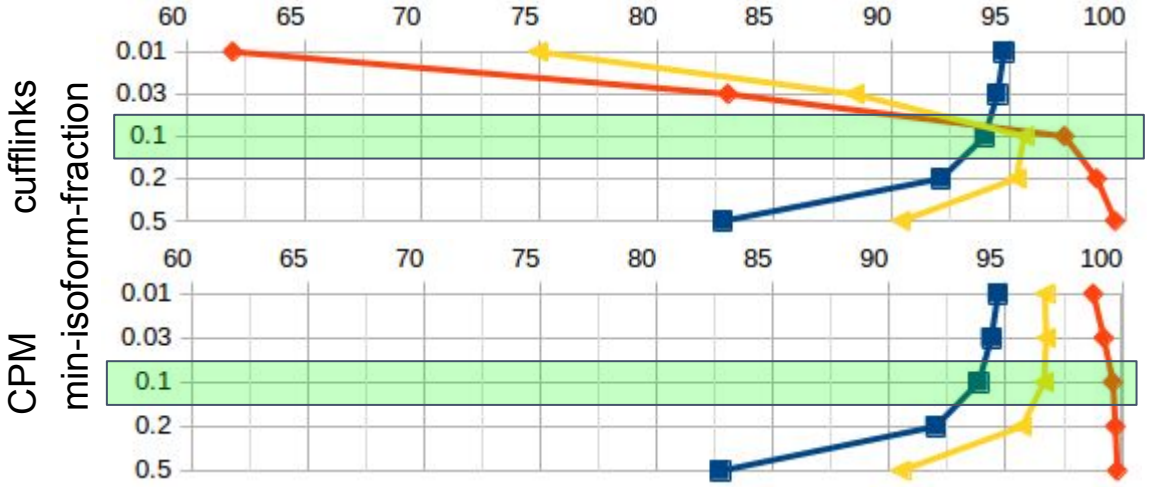
Junction level accuracy on hisat-cufflinks assemblies of the 250M read human dataset



Downstream applications

Transcript reconstruction and gene modelling

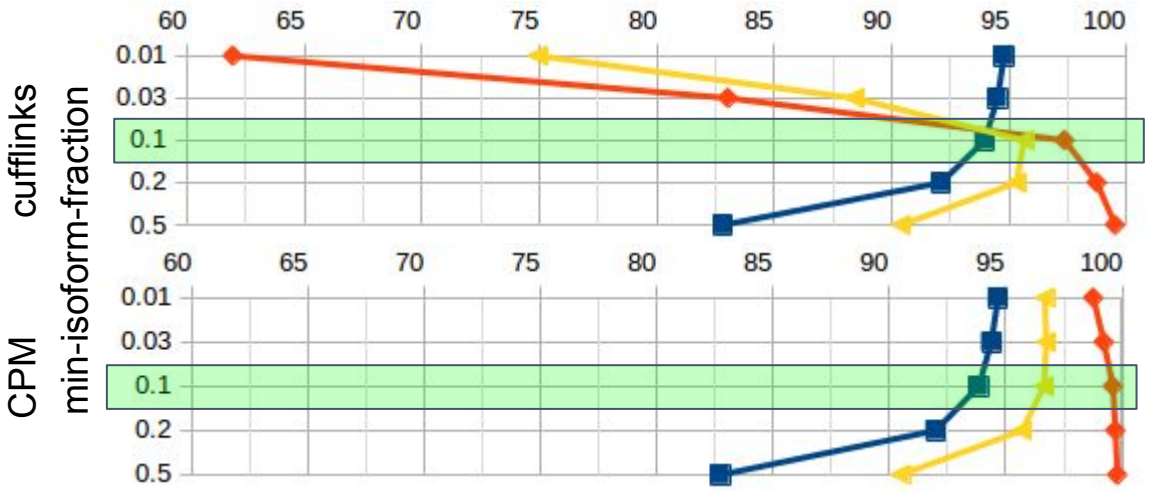
Junction level accuracy on hisat-cufflinks assemblies of the 250M read human dataset



Downstream applications

Transcript reconstruction and gene modelling

Junction level accuracy on hisat-cufflinks assemblies of the 250M read human dataset

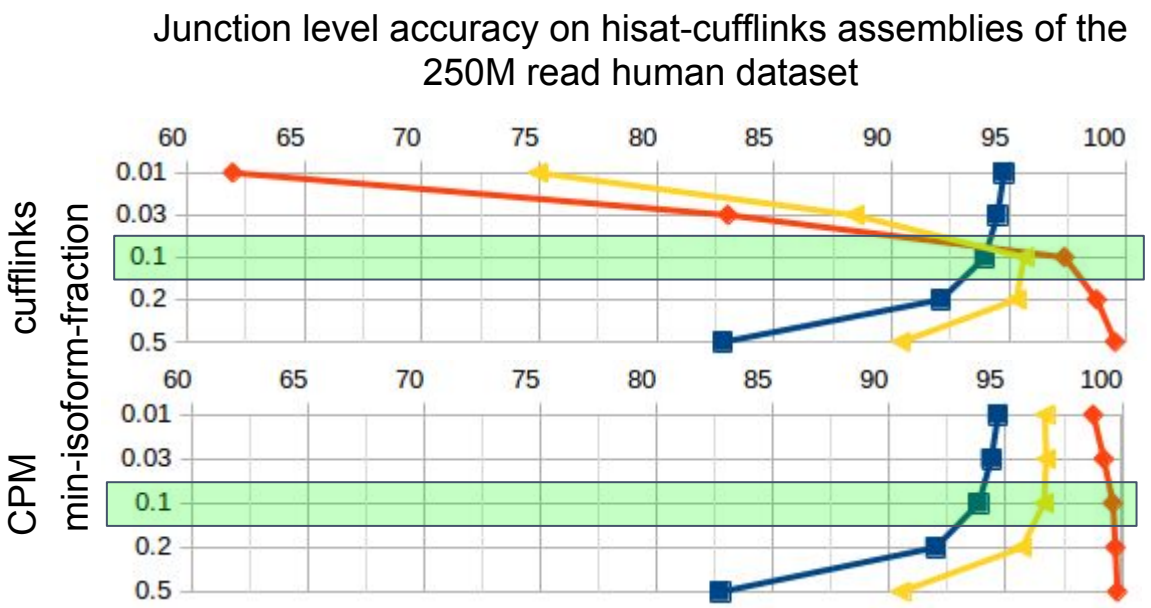


Cufflinks (min-isoform-fraction: 0.1):

- Junction-level precision: 97.35%
- Transcripts with invalid intron-chains: **12.2%** (2344)

Downstream applications

Transcript reconstruction and gene modelling



Cufflinks (min-isoform-fraction: 0.1):

- Junction-level precision: 97.35%
- Transcripts with invalid intron-chains: **12.2%** (2344)

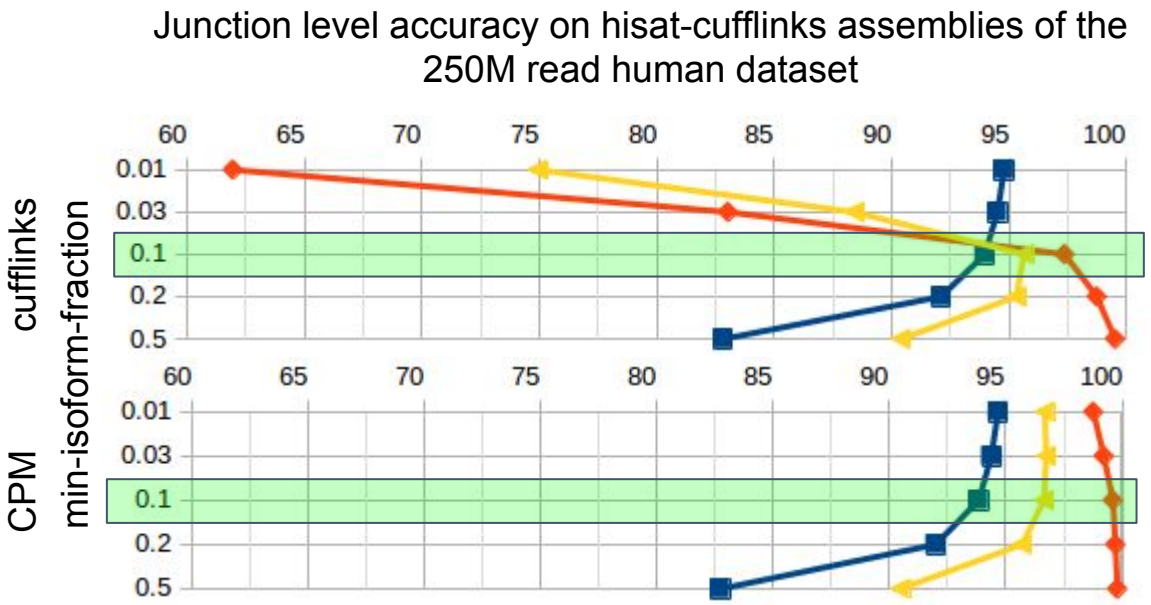
Portcullis intersected:

- Junction-level precision: 99.87% (up 2.5%)
- Transcripts with invalid intron-chains: **2.1%** (361 - down 10.1%)

(Loss of only 73 (~1%) valid transcripts)

Downstream applications

Transcript reconstruction and gene modelling



Cufflinks (min-isoform-fraction: 0.1):

- Junction-level precision: 97.35%
- Transcripts with invalid intron-chains: **12.2%** (2344)

Portcullis intersected:

- Junction-level precision: 99.87% (up 2.5%)
- Transcripts with invalid intron-chains: **2.1%** (361 - down 10.1%)

(Loss of only 73 (~1%) valid transcripts)

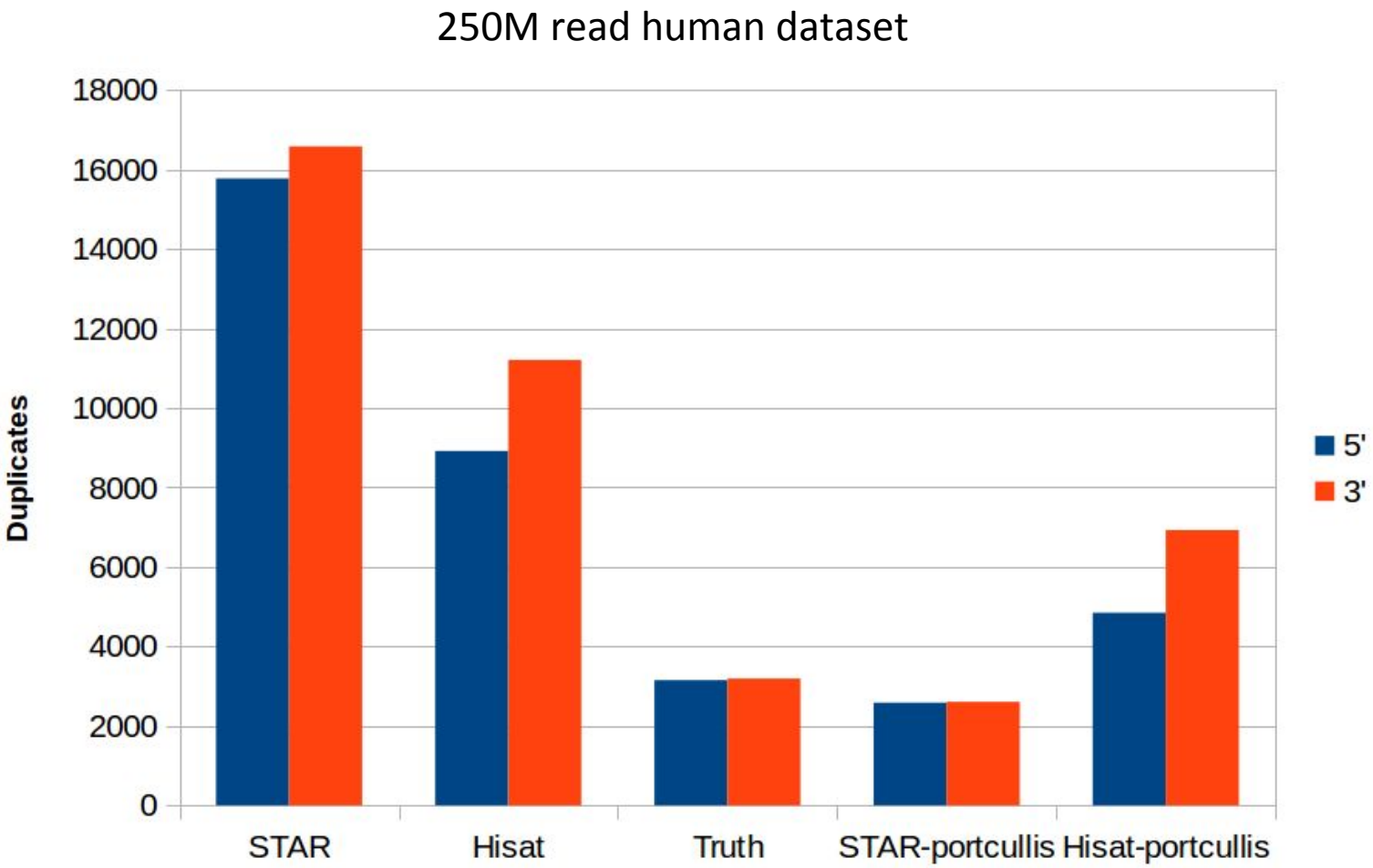
- Portcullis provides useful information that can be leveraged to filter invalid transcripts or inform gene modellers

MIKADO

Earlham Institute

Downstream applications

Alternative splicing analysis



Summary

Fast, robust and accurate splice junction prediction from RNAseq data

- RNAseq mappers produce large numbers of FP junctions, especially in high coverage datasets, and, generally, each mapper produces a different set of FPs
- Portcullis significantly reduces FP junctions from any RNAseq mapper, with a tolerable increase in FNs
- Portcullis is much faster, requires less resources, is more flexible, useful and reliable than the competition
- Portcullis can have a positive impact on downstream tasks such as transcript assembly, gene modelling and alternative splicing analysis
- For more information...

https://github.com/maplesond/portcullis

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

Acknowledgements

Earlham Institute

David Swarbreck

Luca Venturini

Gemy Kaithakottil

Shabonham Caim

Sarah Bastkowski


Earlham Institute