Integrating multiple transcript assemblies for improved gene structure annotation with Mikado

DR LUCA VENTURINI

Computational biologist

Earlham Institute

Decoding Living Systems

MIKADO

Outline

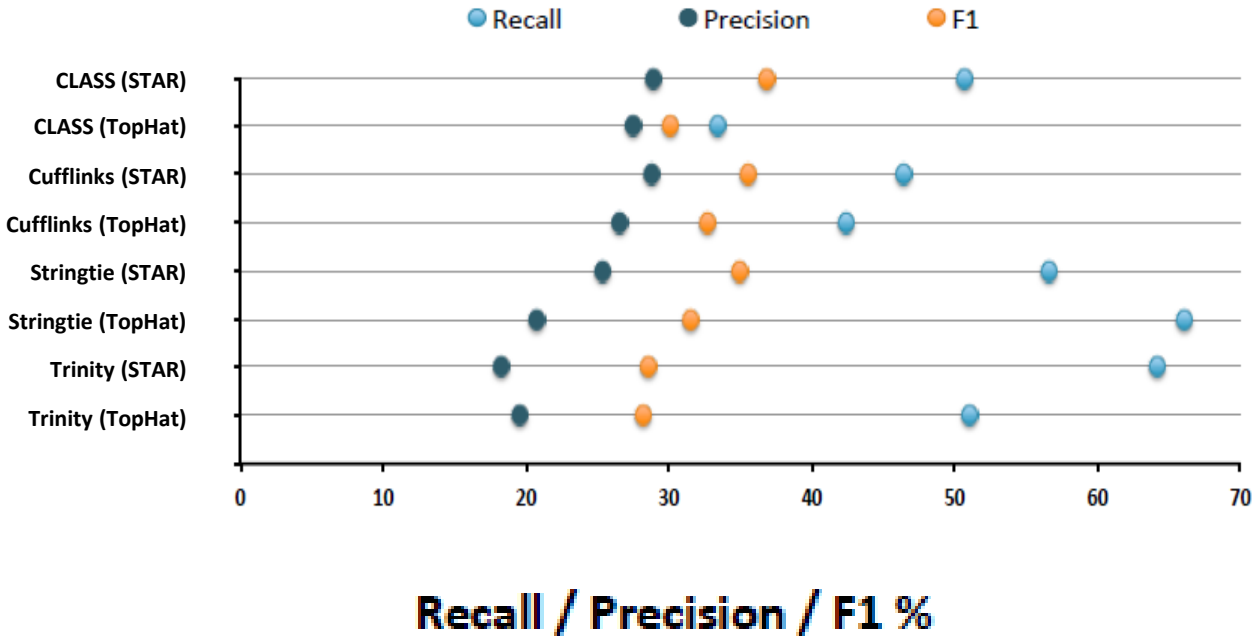
- The first step in many gene builds is to assemble RNA-Seq data to identify reliable gene models to guide *ab initio* predictors or integrator pipelines

- The RGASP challenge in 2013 demonstrated that transcript reconstruction is still an open challenge, both for *ab initio* predictors and RNA-Seq assemblers

- In this talk we will:
  - Show how RNA-Seq assemblers exhibit complementary strengths and weaknesses, making hard to choose *a priori* the best method
  - Present a novel method to integrate multiple transcript assemblies into a clean evidence-based gene build with Mikado

Varying accuracy of assemblers and aligners

Following the RGASP challenge, we decided to assess the efficiency of different combinations of aligners and assemblers on the three species analysed on the challenge (*C. elegans*, *D. melanogaster* and *H. sapiens*) and additionally the model plant organism *A. thaliana.*



Recall = proportion of transcripts correctly called

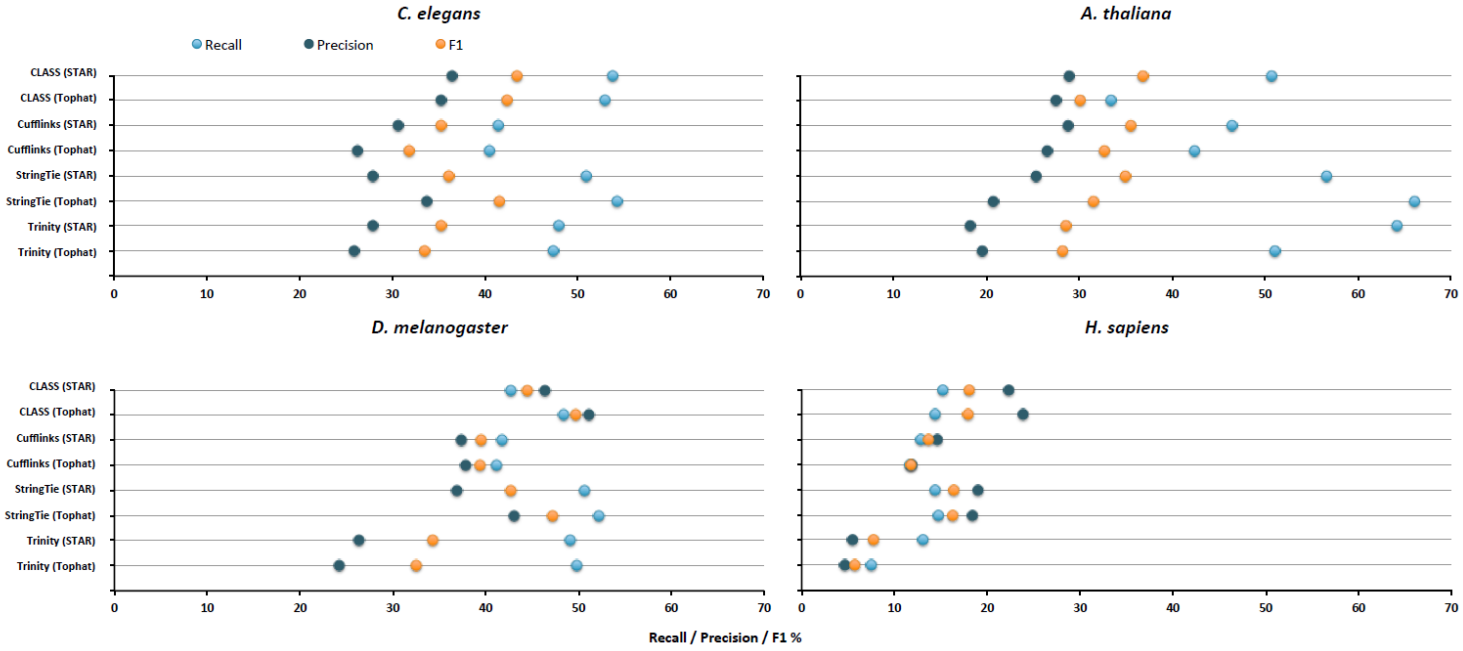
Precision = proportion of called transcripts that are correct

F1 = harmonic mean of recall and precision

Earlham Institute

Varying accuracy of assemblers and aligners

Following the RGASP challenge, we decided to assess the efficiency of different combinations of aligners and assemblers on the three species analysed on the challenge (*C. elegans*, *D. melanogaster* and *H. sapiens*) and additionally the model plant organism *A. thaliana.*



Earlham Institute

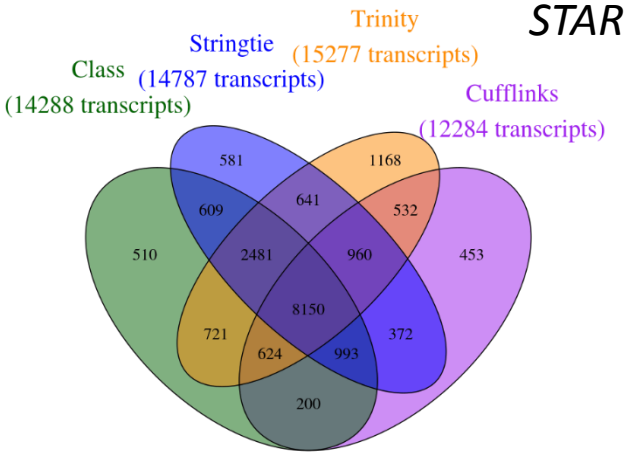
Complementarity of different assemblers

Fully reconstructed transcripts in *A. thaliana*
(from 22,577 transcripts with min 1X coverage and all splice junctions identified)

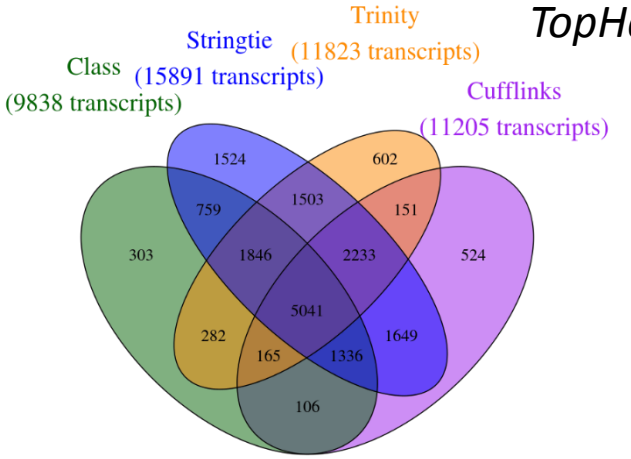
Method	Fully reconstructed transcripts	% of transcripts reconstructed
Stringtie (Tophat)	15,891	70.39%
Trinity (STAR)	15,277	67.67%
Stringtie (STAR)	14,787	65.50%
CLASS (STAR)	14,288	63.29%
Cufflinks (STAR)	12,284	54.41%
Trinity (Tophat)	11,823	52.37%
Cufflinks (Tophat)	11,205	49.63%
CLASS (Tophat)	9838	43.58%
All methods	19,855	87.94%
STAR methods	18,995	79.83%
Tophat methods	18,024	84.13%

3885 transcripts, or 17.21%, reconstructed across all 8 methods.
1771 transcripts, or 7.85%, reconstructed by only 1 of the 8 methods.

STAR



TopHat2

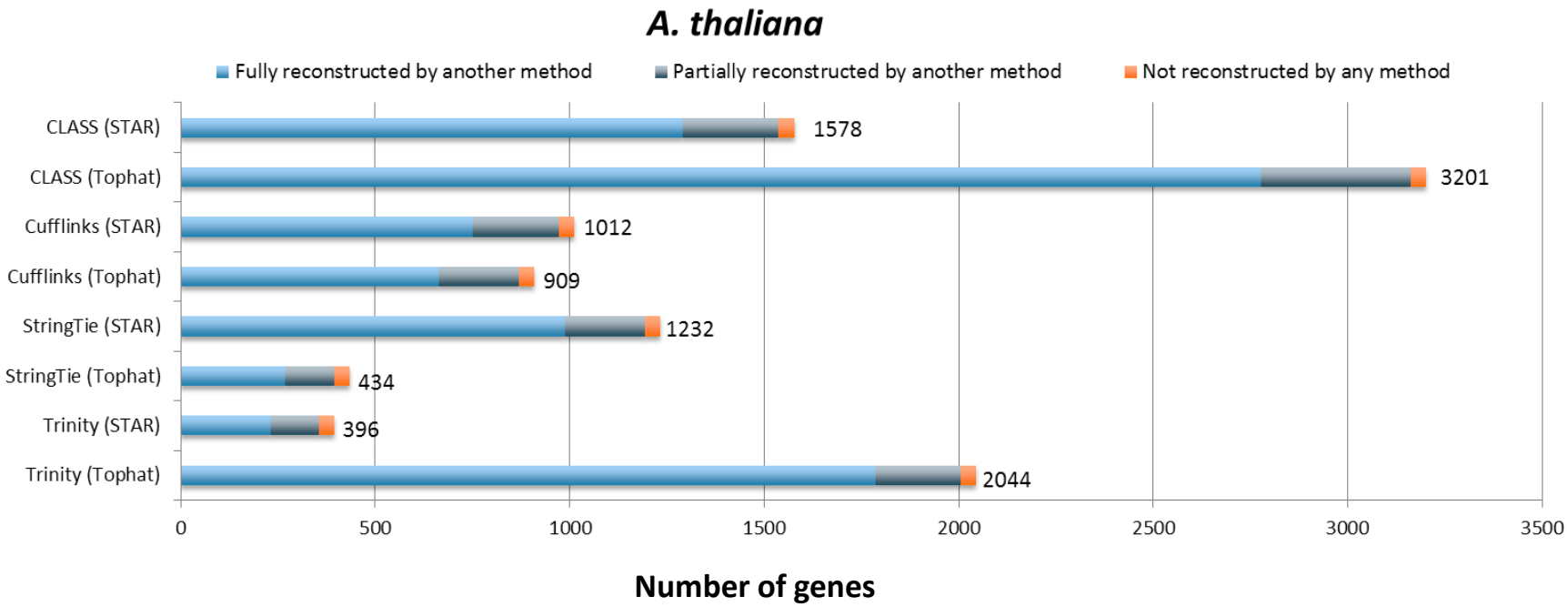


Earlham Institute

MIKADO

Complementarity of different assemblers: missing genes

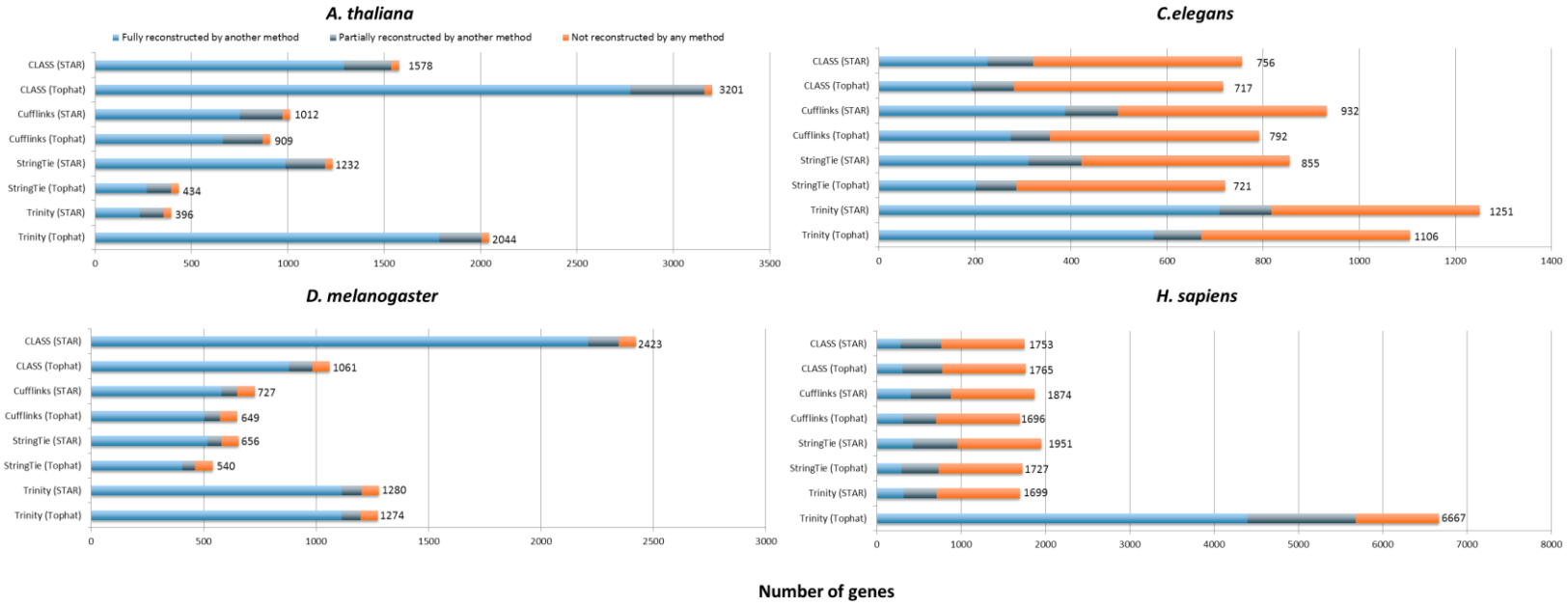
A mirroring problem is that some tools might completely fail to reconstruct the transcript at a locus, even when the available evidence is sufficient for another tool to infer the original transcript(s), or at least part thereof.



Earlham Institute

Complementarity of different assemblers: missing genes

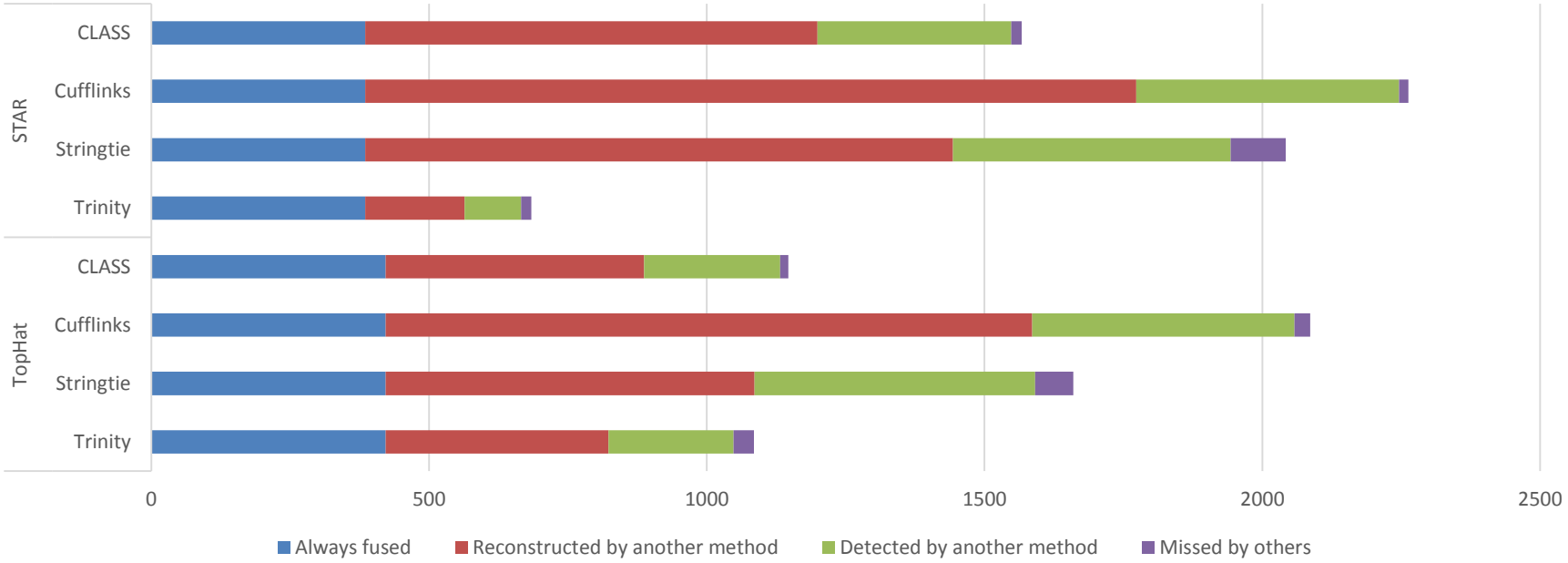
A mirroring problem is that some tools might completely fail to reconstruct the transcript at a locus, even when the available evidence is sufficient for another tool to infer the original transcript(s), or at least part thereof.



MIKADO

Artifactual fused genes

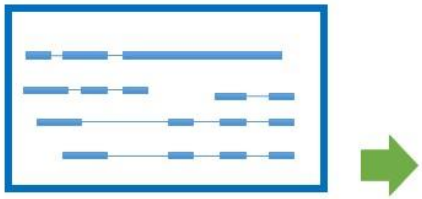
An under-appreciated problem in RNA-Seq assemblies is the tendency to merge together neighbouring genes, especially in compact genomes. Very sensitive methods such as Stringtie can be very prone to this kind of error.



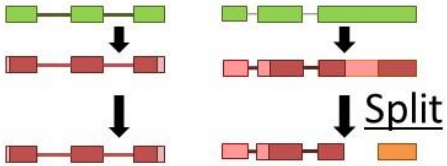
Earlham Institute

MIKADO

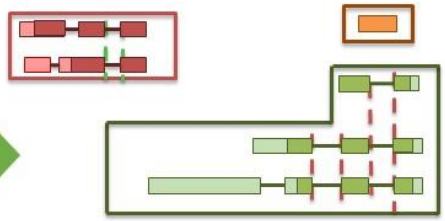
Leveraging multiple transcript assemblies



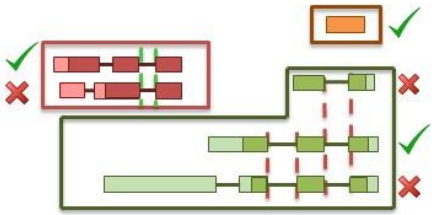
1- Define superloci of overlapping transcripts on the same strand



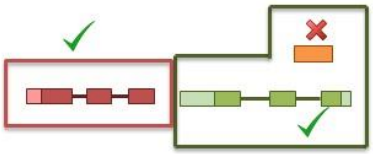
2- Load ORFs and split chimeric transcripts (those with two or more ORFs)



3 – Create subloci of transcripts sharing introns



4 – Score and select transcripts in each sublocus

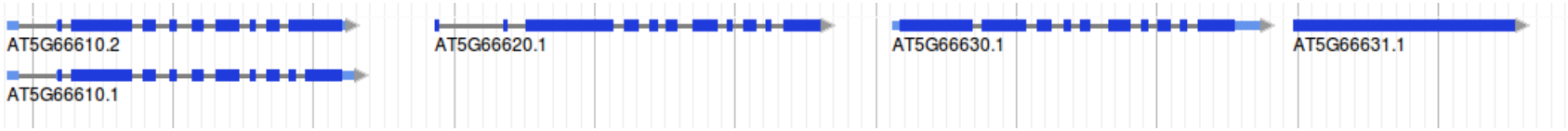


5 – Regroup more leniently into proper loci and reselect

6 – Bring back valid alternative splicing events

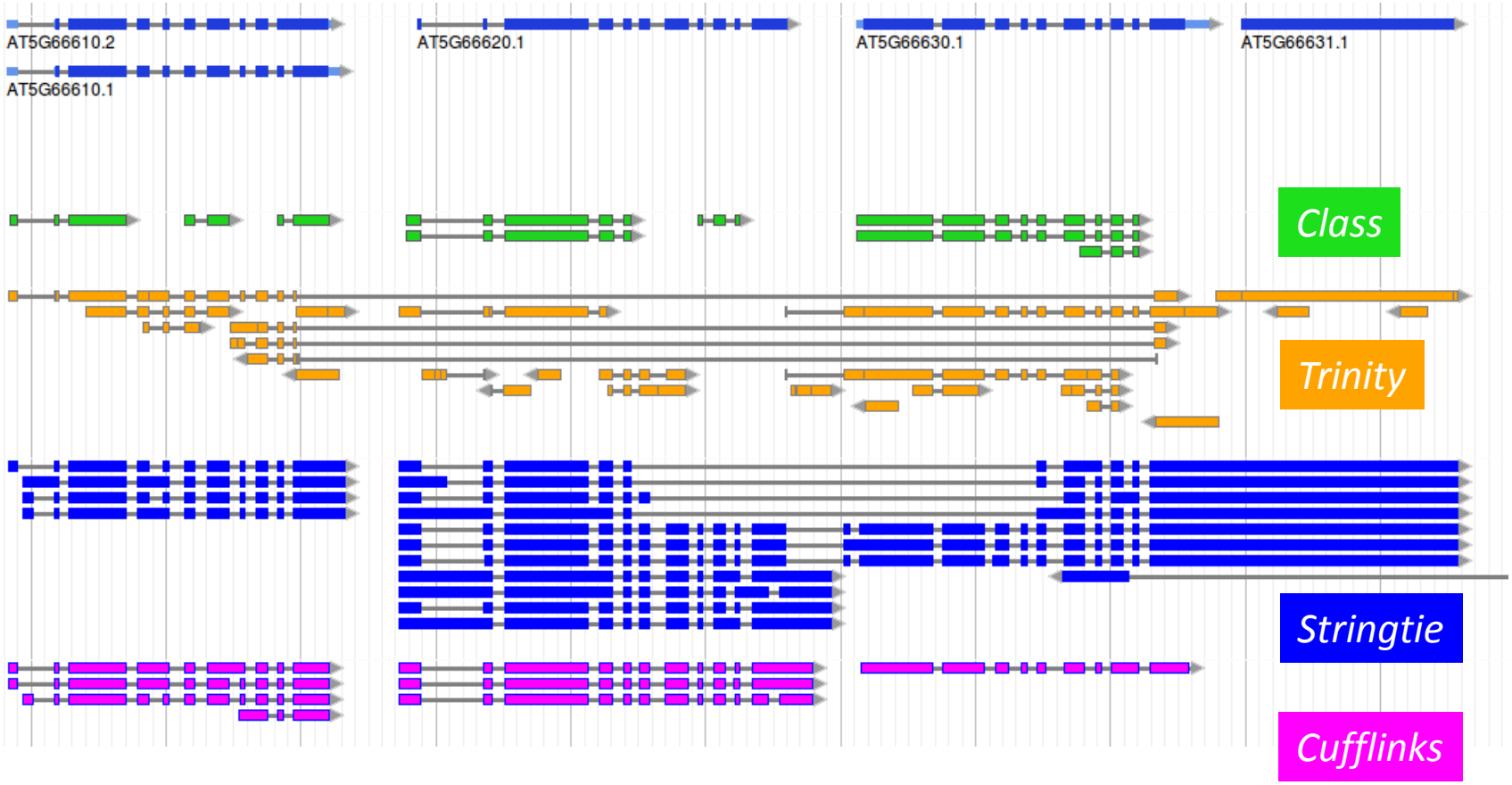
MIKADO

Solving complex regions with Mikado: an example



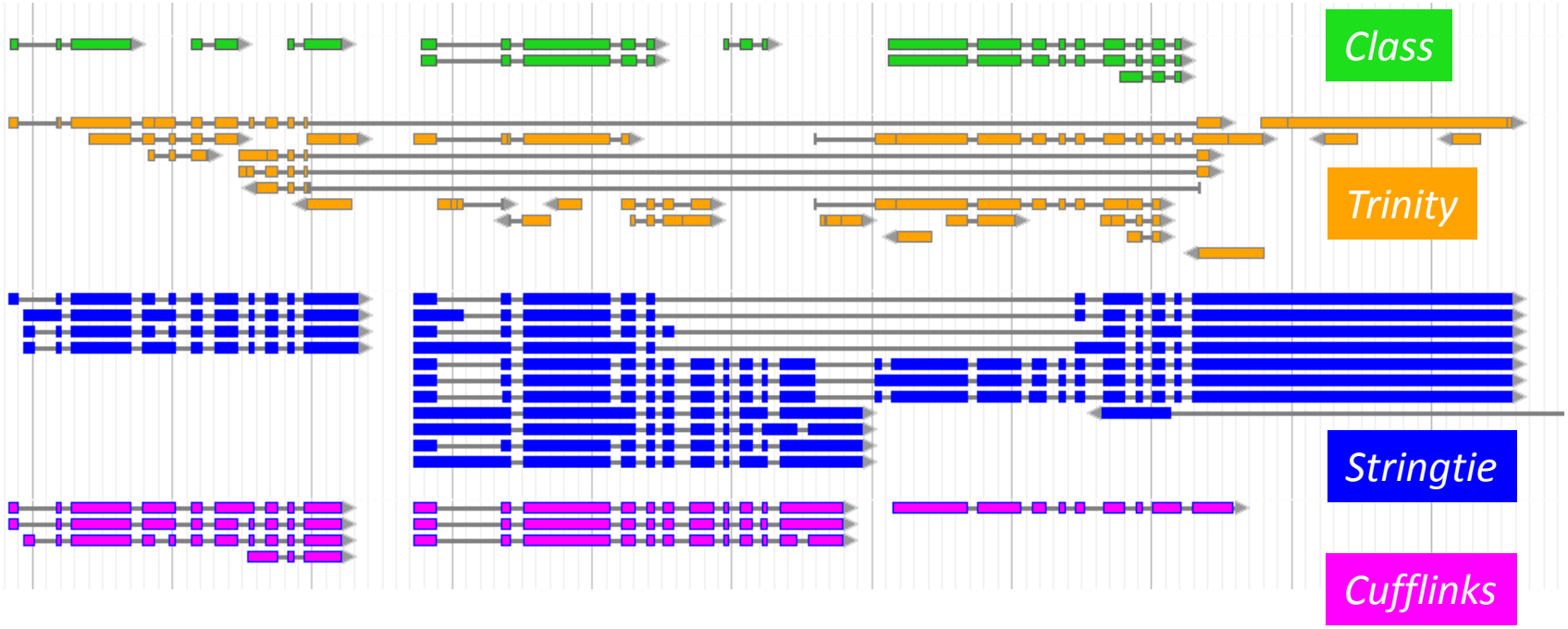
MIKADO

Solving complex regions with Mikado: an example



MIKADO

Solving complex regions with Mikado: an example



The scoring algorithm

Mikado measures each transcript along over 50 dimensions. The user can specify for each what the program should look for; a final score is computed from the sum of all metrics.

For each metric, the user can decide whether to:

- **Maximize its value**

$$s_{mt} = w_m * \left(\frac{r_{mt} - min(r_m)}{max(r_m) - min(r_m)} \right)$$

- **Minimize its value**

$$s_{mt} = w_m * \left(1 - \frac{r_{mt} - min(r_m)}{max(r_m) - min(r_m)} \right)$$

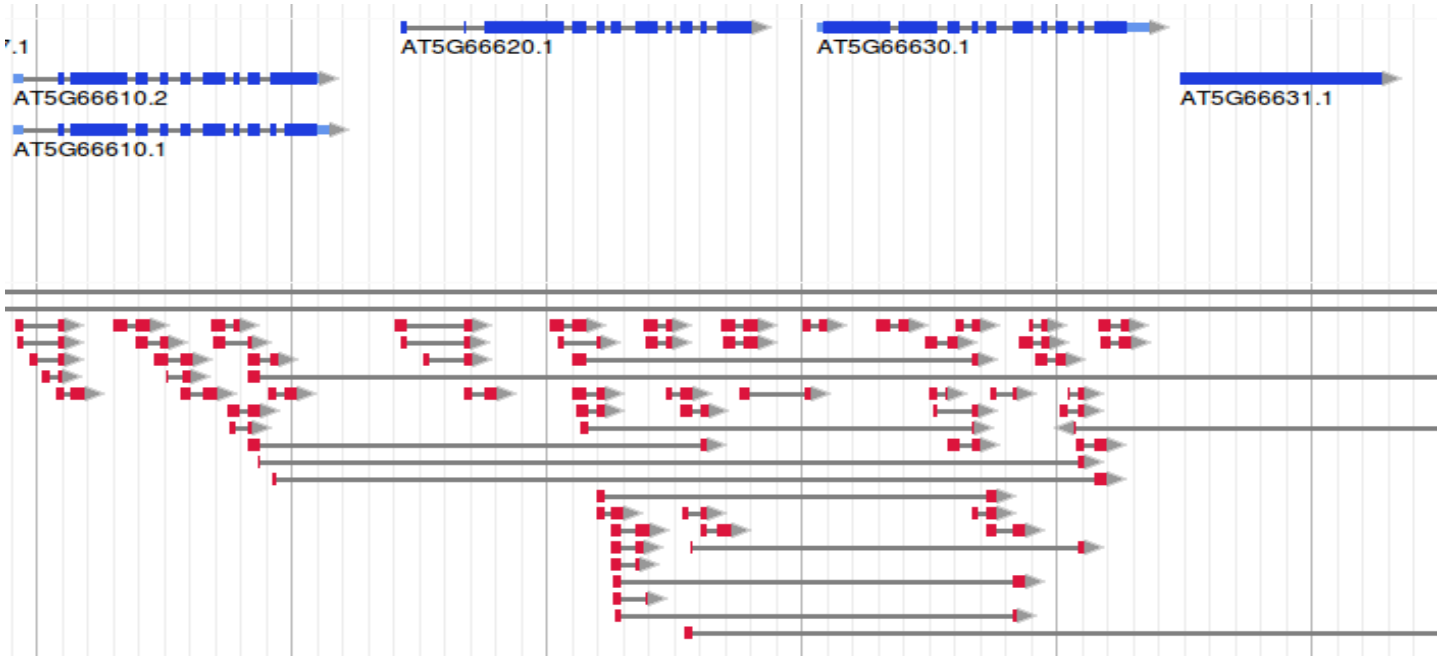
- **Look for transcripts with values most similar to a target:**

$$s_{mt} = w_m * \left(1 - \frac{|r_{mt} - v_m|}{max(|r_m - v_m|)} \right)$$

Category	Description	Count
External	Data confirmed by external programs, eg. Portcullis	7
Intron	Features related to the number of introns and their lengths	5
cDNA	basic features of any transcript such as its number of exons and its cDNA length	2
CDS	Features related to the ORF(s) assigned to the transcript	24
Locus	features of the transcript in relationship to all other transcripts in its current locus	6
UTR	features related to the UTR of the transcript	12

MIKADO

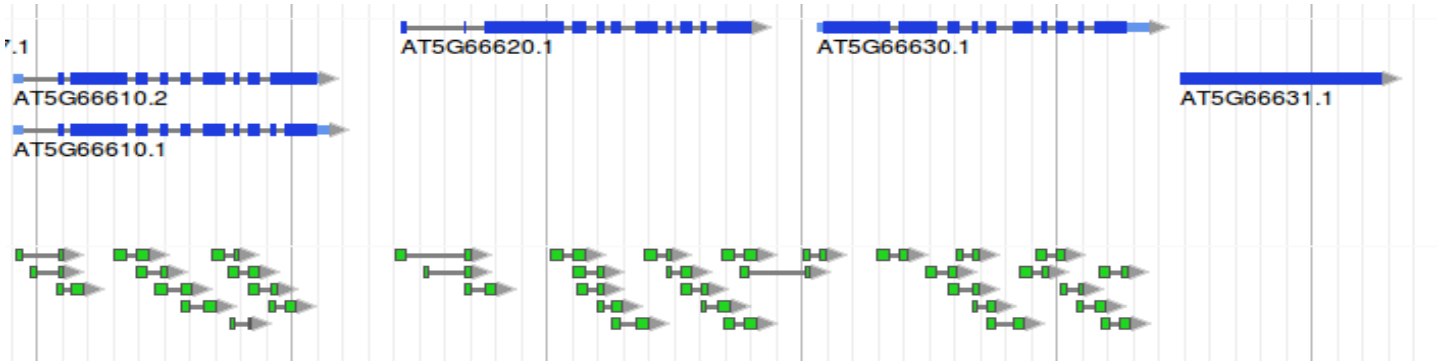
Solving complex regions with Mikado: an example



All junctions

Earlham Institute

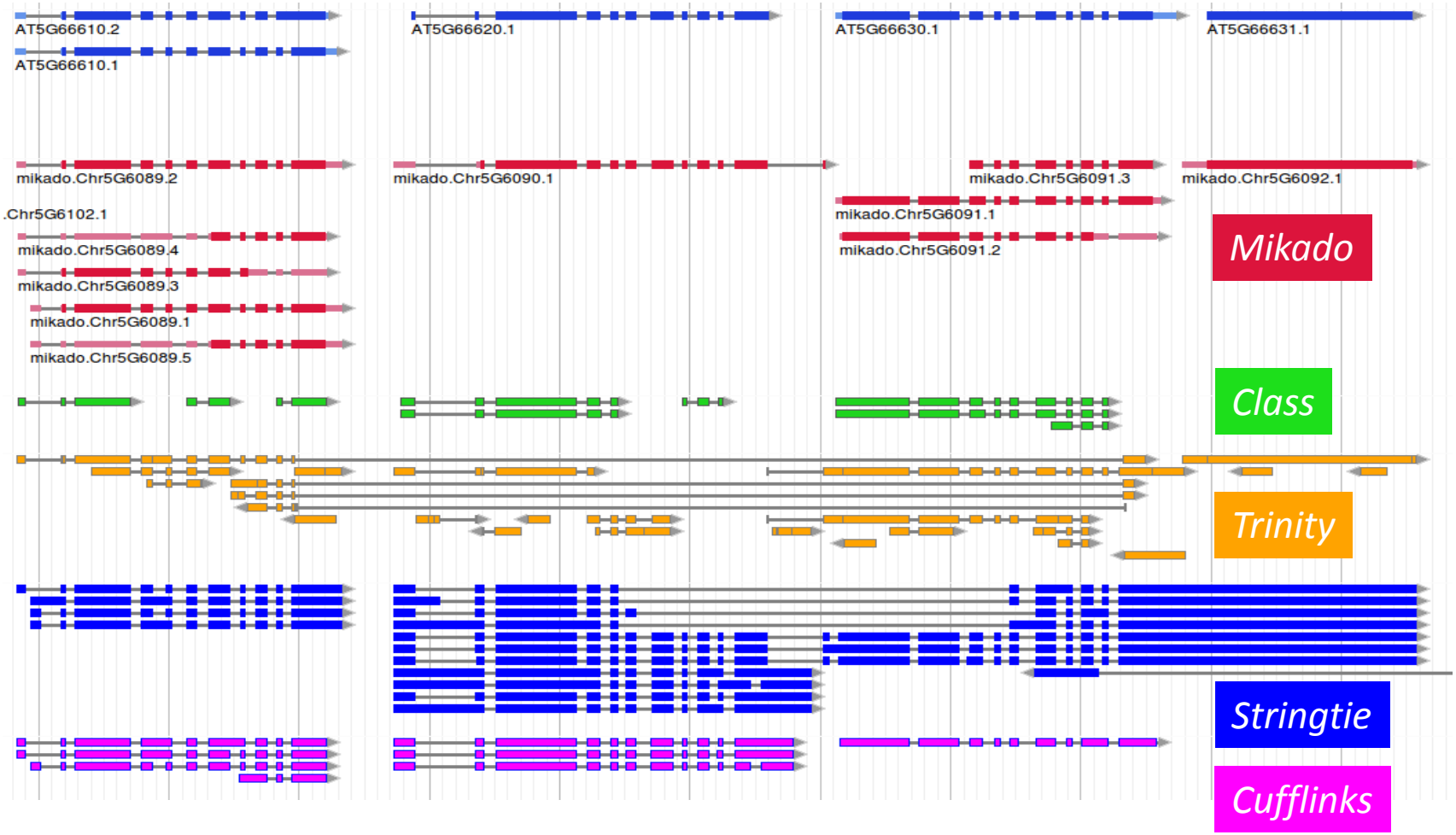
Solving complex regions with Mikado: an example



Portcullis filtered

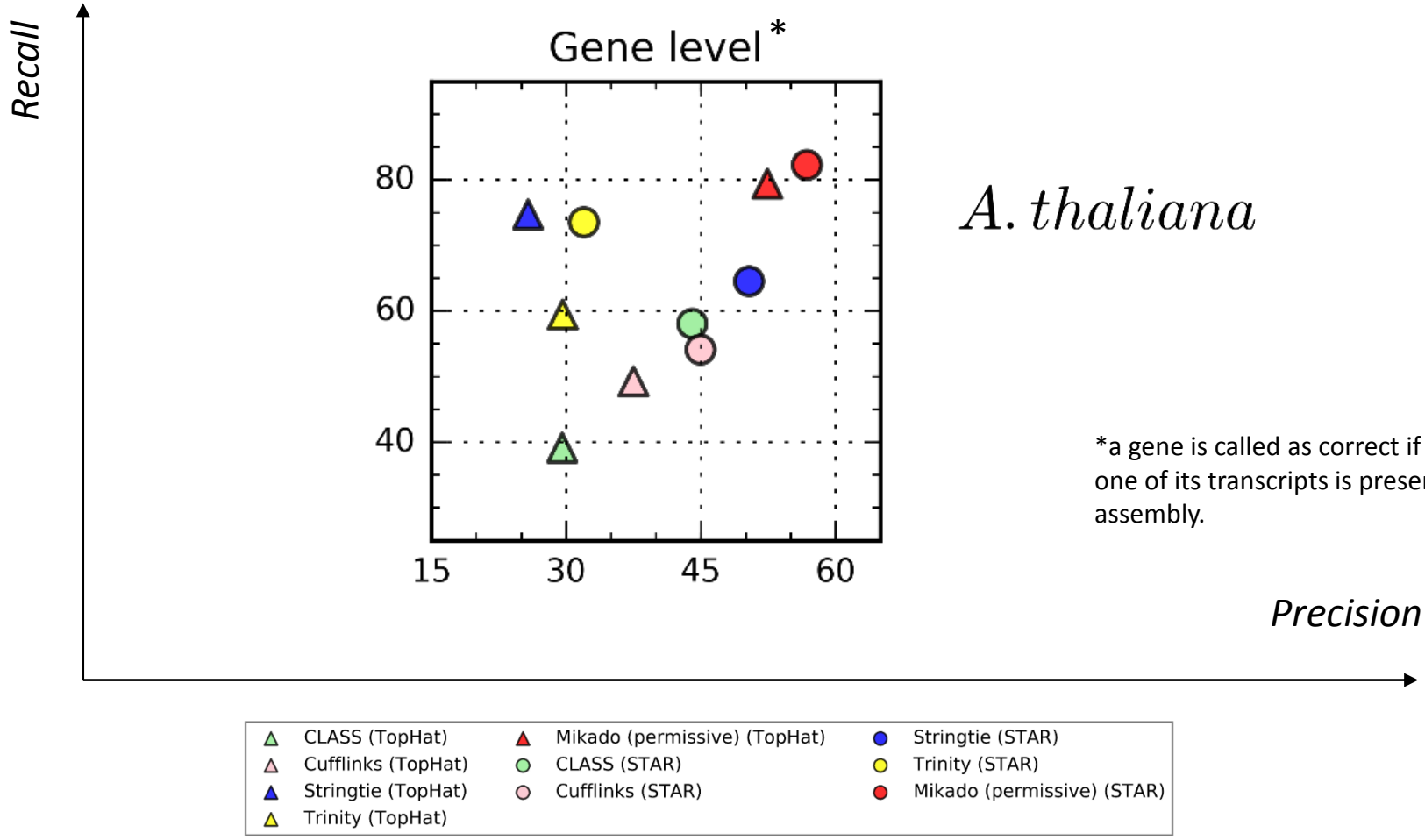
MIKADO

Solving complex regions with Mikado



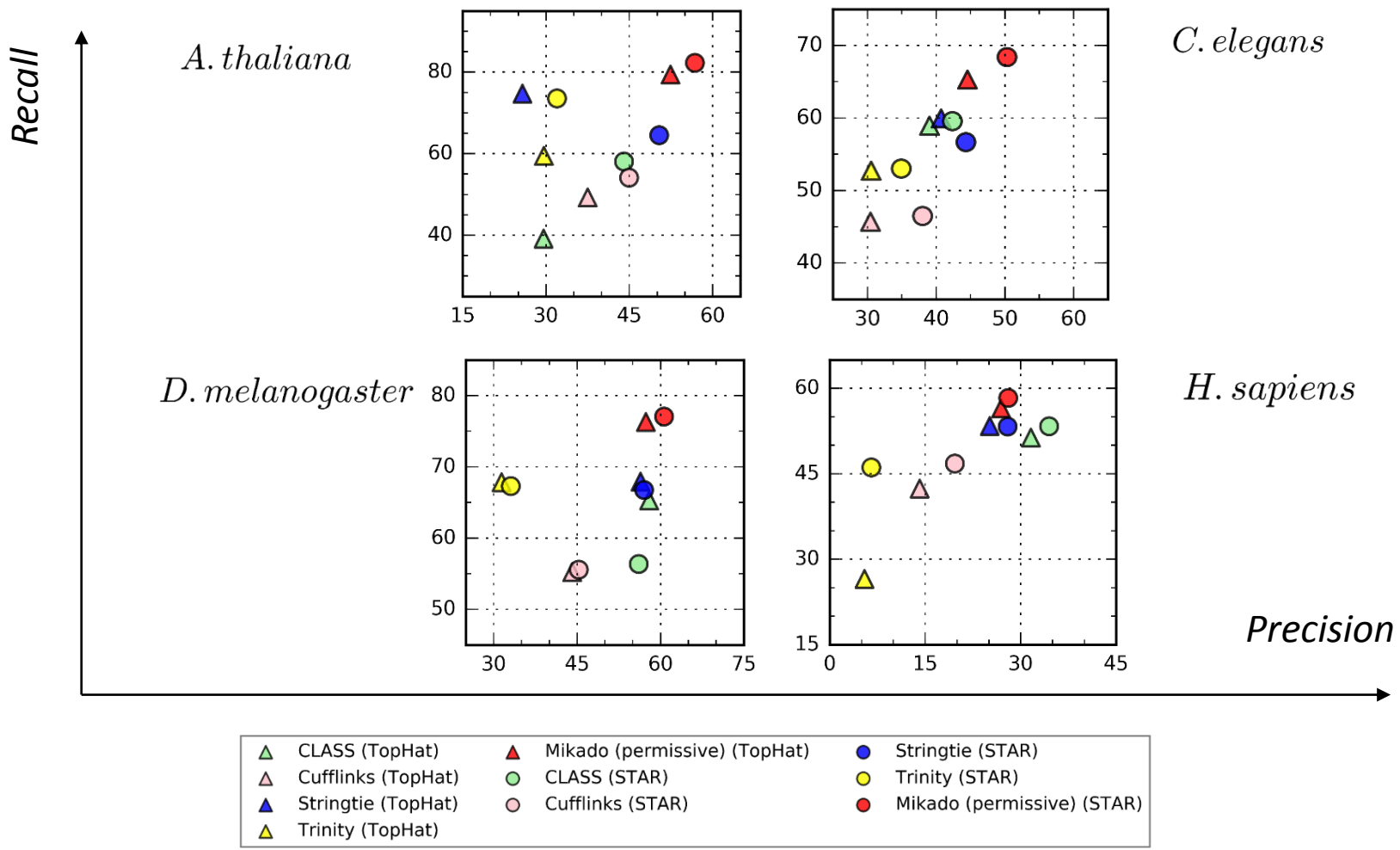
Earlham Institute

Increased precision after filtering with Mikado



Earlham Institute

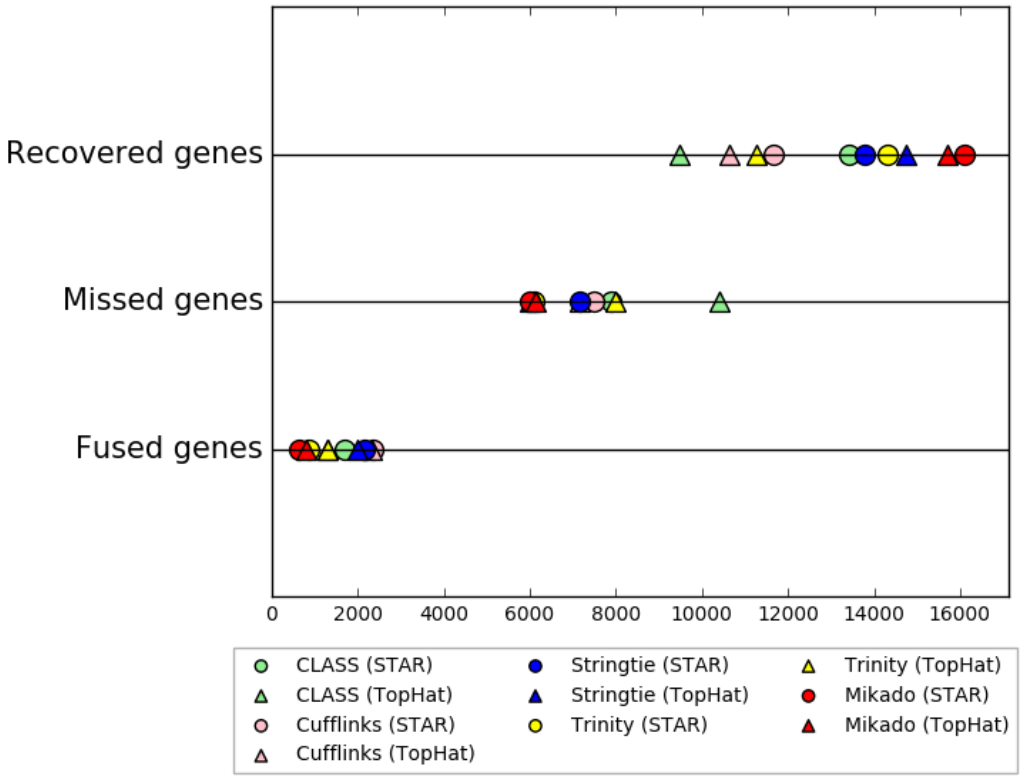
Increased precision after filtering with Mikado



Earlham Institute

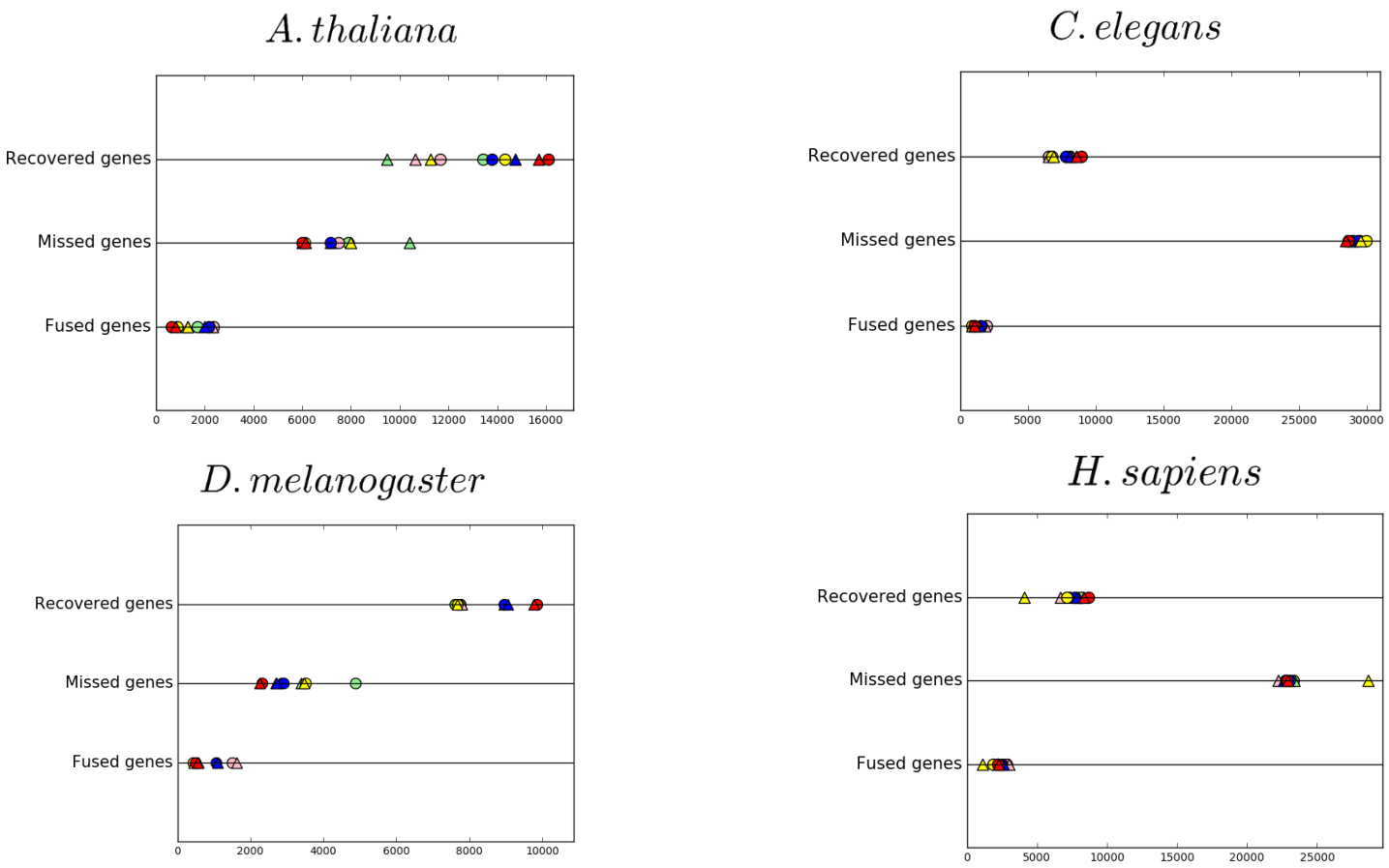
MIKADO

Mikado greatly reduces the incidence of fusions in the data



Earlham Institute

Mikado greatly reduces the incidence of fusions in the data



MIKADO

Summary

- Mikado is capable of recovering a majority of the annotated isoforms in pooled assemblies.
- The method proposed is completely transparent to the user in terms of what will be used to define acceptable transcripts.
- The software is relatively lightweight and fast – the picking operation for *A.thaliana* can be executed in ~15 minutes using approximately 3GB of RAM.
- Mikado uses standard configuration files (JSON/YAML) and file formats (GTF/GFF3/BED):
  - Extendable to other assemblers with standard output formats
- For more information, come visit me at my poster or just go to:

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

https://mikado.readthedocs.org/

MIKADO

Acknowledgements

Shabhonam Caim is the primary tester of the pipeline and his assistance has been essential in performing all the experiments presented here

Gemy George Kaithakottil verified that the pipeline was effective also on different, non-model organisms

Daniel Mapleson is the creator of Portcullis and has helped with polishing and improving the efficiency of the pipeline.

David Swarbreck is the creator of the pipeline and has shepherded the team throughout the whole development cycle.

MIKADO

MIKADO

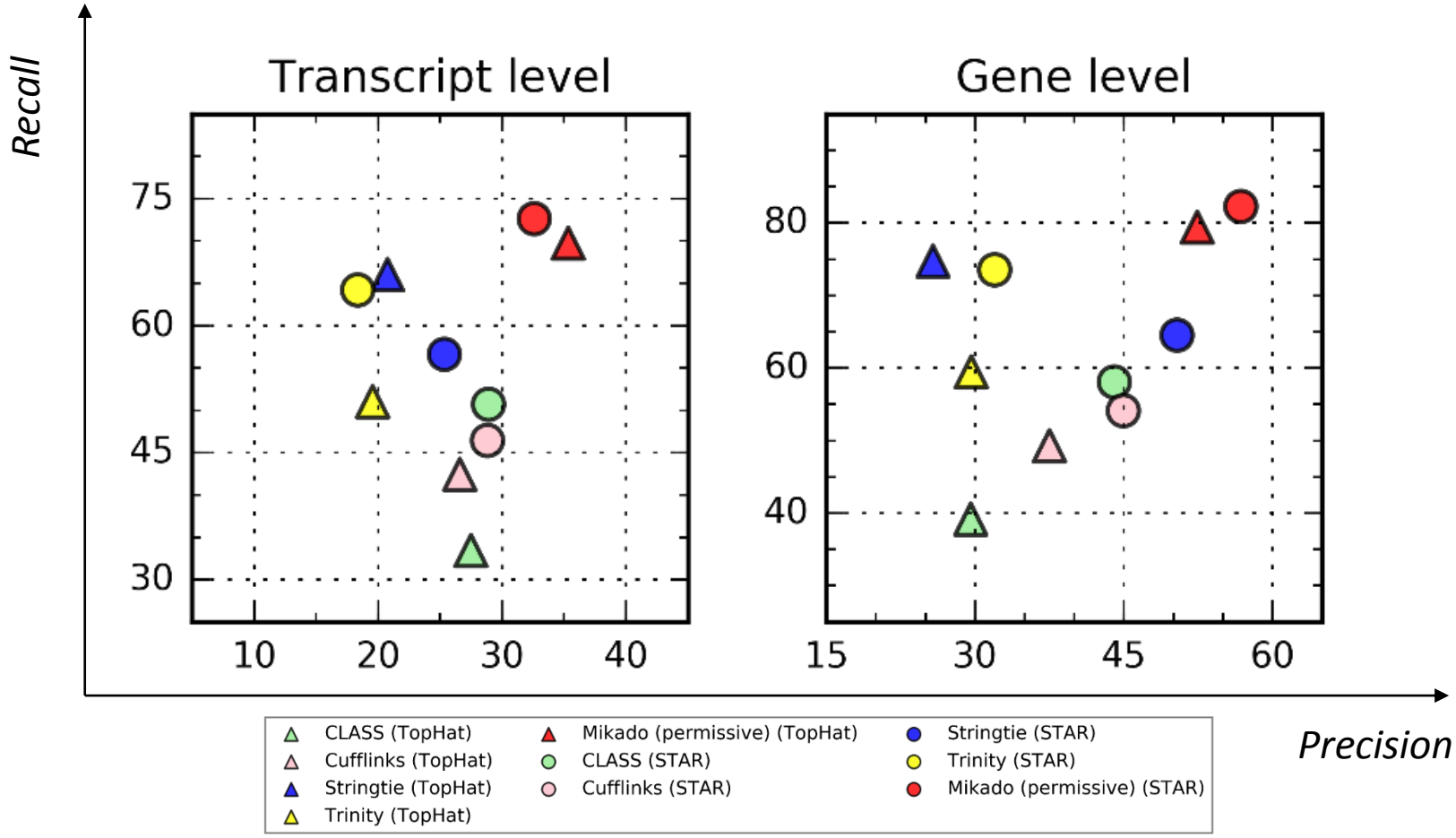
Selecting for general accuracy …

	C. elegans		*A. thaliana*		*D. melanogaster*		*H. Sapiens*		All species	
	z-score	Rank	z-score	Rank	z-score	Rank	z-score	Rank	z-score	Rank
Method										
CLASS (STAR)	7.30	1	7.61	1	-3.31	6	5.25	1	16.85	1
StringTie (Tophat)	5.51	3	0.59	4	6.63	2	3.20	3	15.93	2
CLASS (Tophat)	6.98	2	-5.56	8	9.33	1	5.00	2	15.75	3
StringTie (STAR)	-2.20	4	2.63	3	1.59	3	3.00	4	5.01	4
Cufflinks (STAR)	-2.32	5	2.70	2	-1.76	5	1.04	5	-0.34	5
Cufflinks (Tophat)	-5.36	8	-0.52	5	-1.56	4	-0.99	6	-8.44	6
Trinity (STAR)	-5.07	7	-4.19	7	-4.74	7	-3.42	7	-17.41	7
Trinity (Tophat)	-4.83	6	-3.26	6	-6.18	8	-13.08	8	-27.34	8

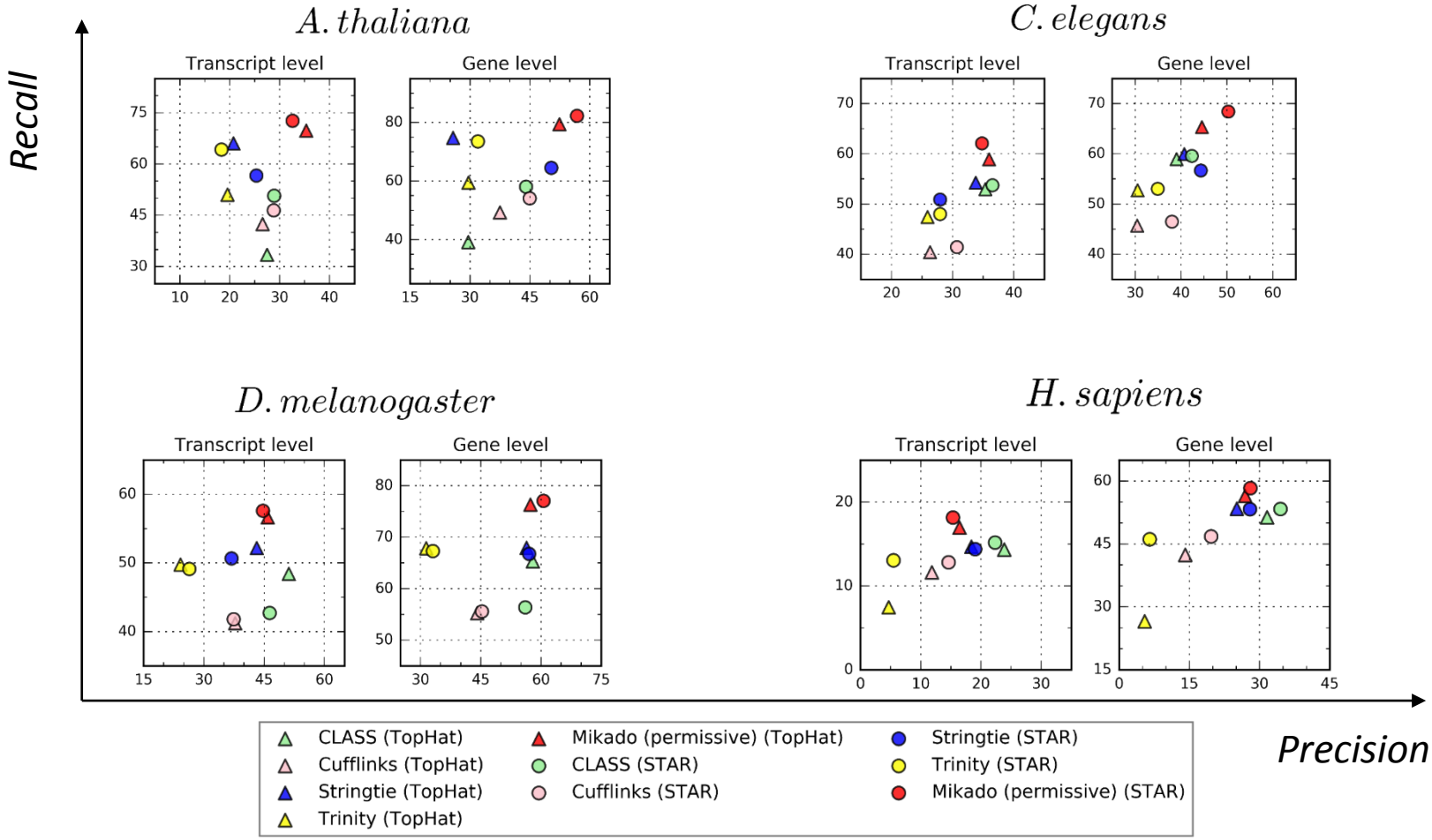
… or maximum sensitivity?

	C. elegans		*A. thaliana*		*D. melanogaster*		*H. Sapiens*		All species	
	z-score	Rank	z-score	Rank	z-score	Rank	z-score	Rank	z-score	Rank
Method										
StringTie (Tophat)	6.91	1	6.91	1	7.20	1	3.87	1	24.90	1
StringTie (STAR)	3.85	4	3.85	3	5.78	2	3.04	3	16.52	2
Trinity (STAR)	-5.22	7	5.83	2	-0.43	5	0.97	6	1.14	3
CLASS (STAR)	4.16	2	1.82	4	-8.74	8	3.59	2	0.84	4
CLASS (Tophat)	3.89	3	-11.33	8	3.24	3	2.33	4	-1.87	5
Cufflinks (STAR)	-4.05	5	-1.28	5	-3.32	6	1.54	5	-7.10	6
Cufflinks (Tophat)	-5.34	8	-4.24	7	-3.69	7	-0.52	7	-13.79	7
Trinity (Tophat)	-4.20	6	-1.57	6	-0.04	4	-14.83	8	-20.64	8

Increased precision after filtering with Mikado



Increased precision after filtering with Mikado



Earlham Institute