MIKADO

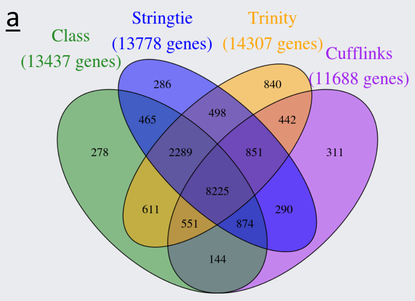
Integrating Multiple Transcript Assemblies for Improved Gene Structure Annotation

Venturini Luca, Caim Shabhonam, Mapleson Daniel, Kaithakottil Gemy George, Swarbreck David

The Rationale for Mikado

In the spirit of the RGASP challenge [1], we assessed the performance of different transcript assemblers and aligners in the three original species of the study *(C. elegans, D. melanogaster* and *H. sapiens)* and in the model organism *Arabidopsis thaliana*. We observed that each method showed considerable variation in terms of number of fully reconstructed genes, gene fusions and missing genes.

a



b

Number of methods	# loci identified as fully reconstructed	Cumulative total
4 methods	8225 (44.4%)	8225 (44.4%)
3 methods	4565 (24.6%)	12,790 (69.0%)
2 methods	2450 (13.2%)	15,240 (82.2%)
1 method only	1715 (9.2%)	16,955 (91.4%)
No method succeeded	1548 (8.3%)	18,543 (100.0%)

Fig 1. Numbers of fully reconstructed genes: in this example, we are considering assembly on *A. thaliana* following alignment with STAR. The various tools display considerable variation, with only a handful of genes reconstructed by all methods (8225, a) while 1715 genes can be reconstructed only by one of the methods but not the others (b)

a

b

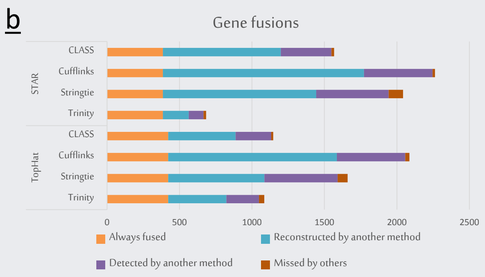


Fig 2. Missing genes and fusions: The most sensitive method, Stringtie (a), also produced a high number of spurious gene fusions compared to competitors (b), regardless of the aligner used. Trinity and Class fused a smaller number of genes, but their sensitivity degraded greatly when used in combination with TopHat2 (a).

The Mikado Algorithm

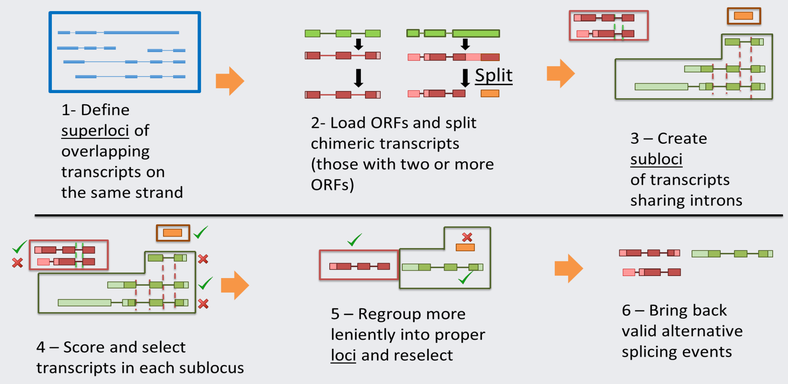


Fig 3: The philosophy of Mikado is to gather together multiple assemblies and select the best in each group according to a configurable set of predefined criteria, e.g. maximizing the ORF length and targeting a UTR/CDS ratio of 20:80 (the average ratio in the *A. thaliana* transcriptome).

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

Table 1: A total of 56 metrics are available for selection; for each of them, Mikado can be instructed to look for the assembly with the maximum possible value, the minimum one, or for the nearest to a predefined target.

Availability and Documentation

Mikado is on GitHub! Just head over to http://www.github.com/lucventurini/mikado

to download a copy of the tool. Mikado is available under the **LGPL3**.

Full documentation for the project is available at: https://mikado.readthedocs.org/

The documentation covers details on how to install the tool, tutorials, and a full documentation for the library underlining Mikado

Improved Transcript Reconstruction

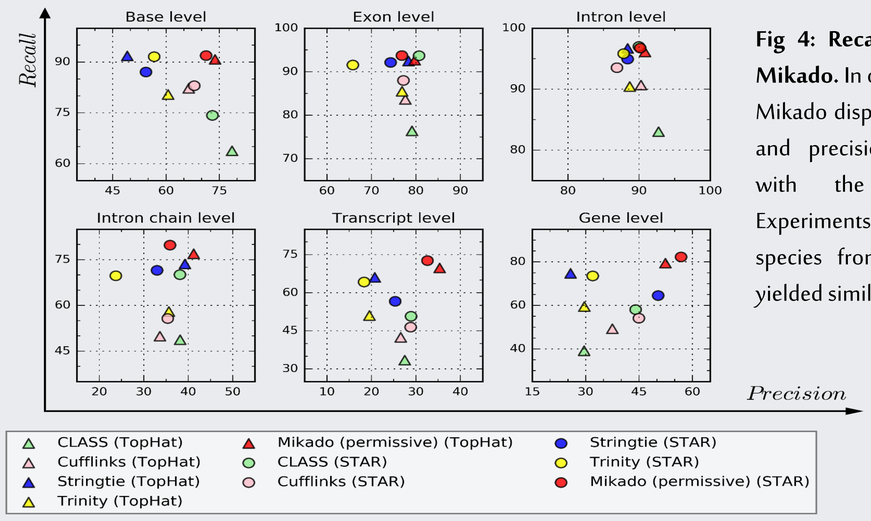


Fig 4: Recall and precision of Mikado. In our test on *A. thaliana*, Mikado displayed increased recall and precision when compared with the input methods. Experiments on three other species from the RGASP study yielded similar results.

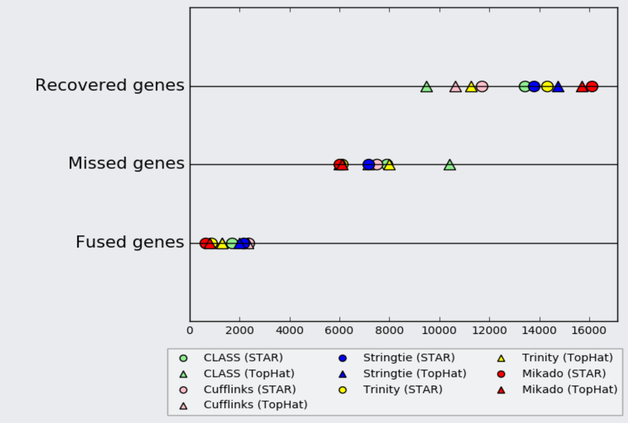


Fig 5: Removal of artefacts through Mikado.
Mikado outperforms the best individual methods both in terms of reconstruction efficiency and in avoidance of fusion events. When using STAR as aligner, the effect was particularly marked – Mikado reported 1531 less fused genes compared with Stringtie, a decrease of over 70%..

The Mikado Team

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.

References
Steijger et al., "Assessment of transcript reconstruction methods for RNA-Seq", Nature Methods 10 1177-1184 (2013)

Earlham Institute, Norwich Research Park, Norwich, Norfolk, NR4 7UG, UK
www.earlham.ac.uk

Earlham Institute
Decoding Living Systems