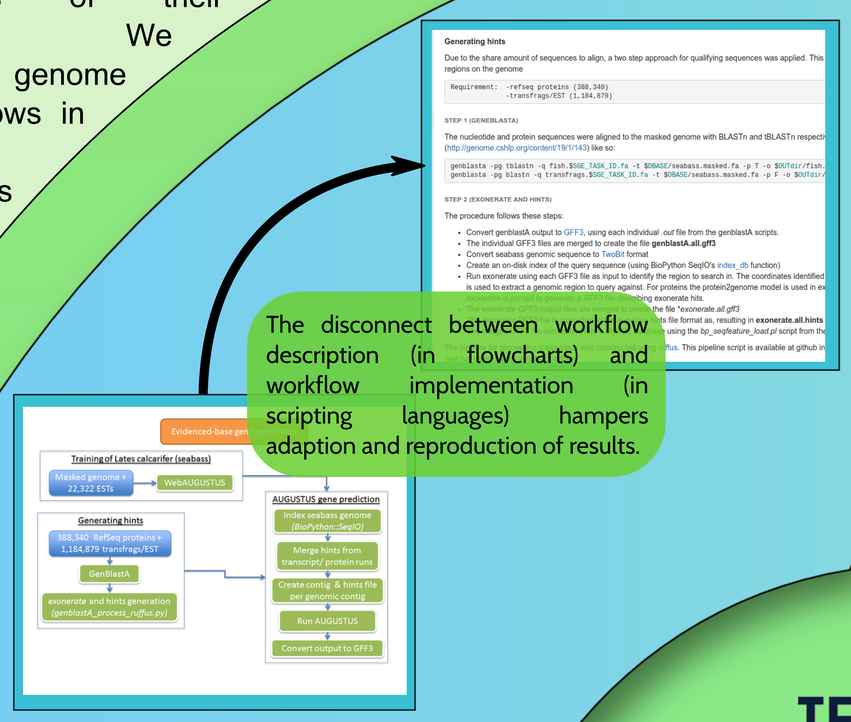
An Extensible Genome Annotation Workbench based on the Galaxy Platform

Peter van Heusden[1]*, Shubha Vij[2], Laszlo Orban[2] and Alan Christoffels[1]*
*corresponding authors: pvh@sanbi.ac.za and alan@sanbi.ac.za
[1] South African National Bioinformatics Institute, UWC ZA
[2] Temasek Life Sciences Laboratory, SG

Introduction

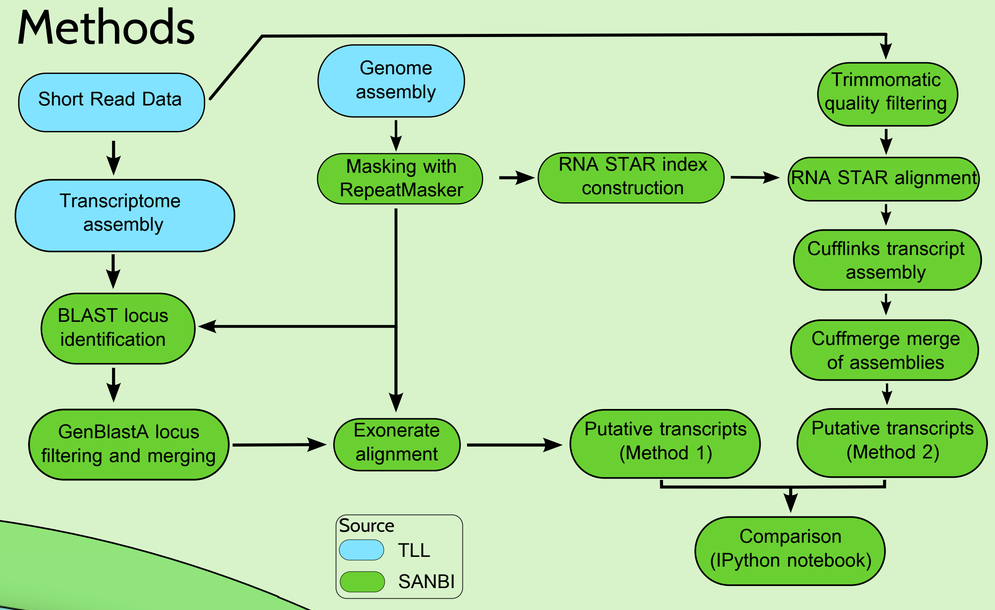
Falling costs of genetic sequencing have allowed sequencing and annotation of the genomes of non-model organism. In annotating non-model genomes, mRNA-seq data has great potential to improve annotation quality. For example the Asian seabass (*L. calcarifer*) genome annotation effort drew on previously assembled mRNA-seq data provided by the Temasek Life Sciences Laboratory (TLL, Singapore). At the South African National Bioinformatics Institute (SANBI) we undertook gene annotation on the Asian seabass genome using a pipeline built out of custom scripts. Simultaneously, a team at Saint Petersburg State University undertook the same task using MAKER2. Comparing the results of this annotation highlighted the impact of tool and parameter choice in gene prediction.

The Galaxy[1] framework allows workflows to be constructed in a high level workflow language that hides the system-specific details of their implementation. We implemented genome annotation workflows in Galaxy, demonstrating its suitability for constructing an annotation

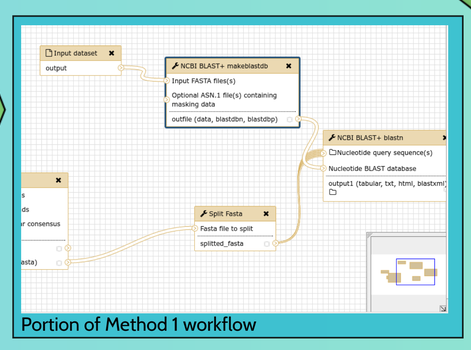


The disconnect between workflow description (in flowcharts) and workflow implementation (in scripting languages) hampers adaption and reproduction of results.

Methods

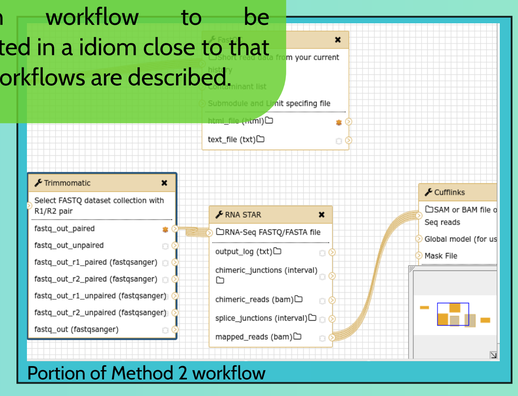


Four datasets were obtained from TLL: unmerged assembled[2] transcript data, comprising 1,184,879 contigs totalling 890 Mb, the assembled L. calcarifer genome[3] (668 Mb, N50 1,191,366) and two sets of Illumina HiSeq RNAseq reads (486 million reads after filtering for quality and length). Analysis was conducted using Galaxy workflows, an IPython notebook and custom scripts.



Portion of Method 1 workflow

Galaxy workflows allow the annotation workflow to be implemented in a idiom close to that in which workflows are described.

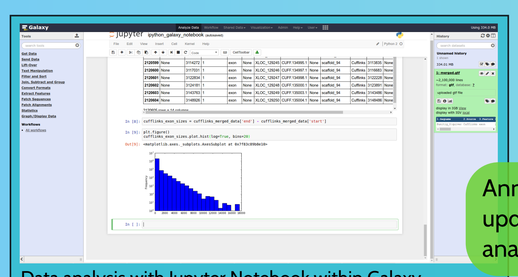


Portion of Method 2 workflow

TEMASEK LIFESCIENCES LABORATORY

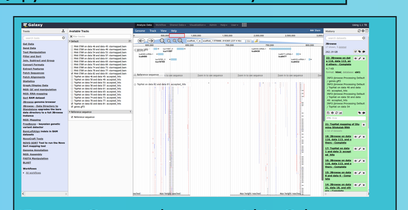
SANBI South African National Bioinformatics Institute

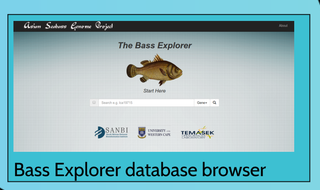
UNIVERSITY of the WESTERN CAPE



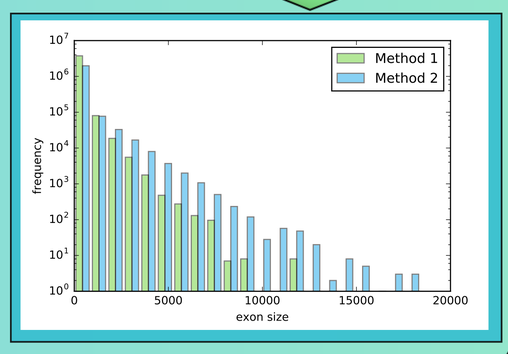
Data analysis with Jupyter Notebook within Galaxy

Annotation can be updated, visualised and analysed within Galaxy

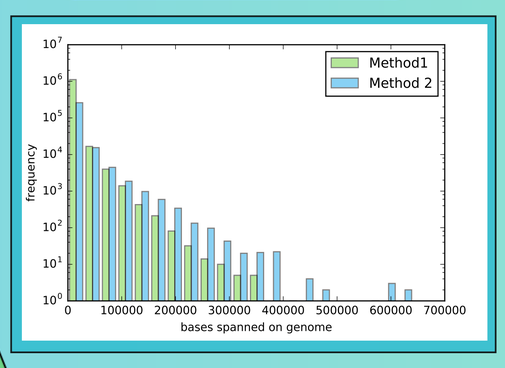

JBrowse Genome browser within Galaxy



Bass Explorer database browser



Method 2, the faster method, generates some artificially long isoforms but 96% of exons predicted overlap with those predicted by Method 1.



Results and Discussion

Method	Max Parallelism	Run time
1. Align assembled transcripts	40 way data split	12 days
2. Map reads & assemble transcripts	14 way data split	5 days

The tools in the two transcript-alignment and reconstruction workflows make use of different degrees of multithreading, but in both workflows extra parallelism was achieved by using Galaxy dataset collections. Even with parallelisation the BLAST, GenBlastA and Exonerate steps of the workflow, the RNA STAR and Cufflinks based Method 2 was more than twice as fast as Method 1.

The transcripts used in Method 1 were assembled independently in different libraries, resulting in overlapping transcript alignments and a substantially higher number of aligned transcripts compared to Method 2, which merged transcript alignments from multiple libraries into final transcripts. Method2 predicted significantly (according to Mann-Whitney test) larger exons and on-genome isoforms, with 460 spanning more than 200Kb. We suspect that these suspiciously large isoforms are artifacts possibly caused by matches to unmasked repeats. Despite these suspicious artifacts, 2,040,178 out of 2,120,605 (96%) exons predicted by Method 2 overlapped with exons predicted by Method 1, suggesting that Method 2 is worth further investigation.

Method	No. of transcripts	No. of exons	Exons per transcript	Exon size mean / stddev	Isoform size mean / stddev
1. Align assembled transcripts	1,137,181	3,853,944	3.4	210 / 291	3718 / 9624
2. Map reads & assemble transcripts	284,104	2,120,605	7.5	295 / 587	11308 / 21862

Conclusion

- Repeatable genome analysis workflows allows for reuse of methods and reproducibility of results
- Galaxy workflows allow workflow construction in a flowchart-like idiom similar to the way in which workflows are documented
- We demonstrate the construction of two workflows as part of a larger annotation workbench and their use in the annotation of the L. calcarifer genome
- Exposing results through Jupyter notebooks and export to browsers such as JBrowse and the (SANBI-authored) Bass Explorer allows results to be examined seamlessly within Galaxy

Future work

- We intend to implement further workflows to provide a complete Galaxy-based eukaryotic genome annotation workbench
- We will enhance tools that export to an annotation database browser modelled on the Bass Explorer

Bibliography and Acknowledgements

1. **The Galaxy Team.** *The Galaxy Project: Online bioinformatics analysis for everyone.* [cited 2016 Mar 29]. Available from: https://www.galaxyproject.org/
2. **Thevasagayam NM, et al.** *Transcriptome Survey of a Marine Food Fish: Asian Seabass (Lates calcarifer).* J Mar Sci Eng. 2015 Jun 2;3(2):382–400.
3. **Vij S, et al.** *Chromosomal-level assembly of the Asian seabass genome using long sequence reads and multi-layered scaffolding.* PLoS Genetics (in press)

Special thanks to the Galaxy Intergalactic Utilities Commission (IUC) members, especially Eric Rasche (for JBrowse wrapper) and Björn Grüning (for Interactive Environments) and SANBI software developers Thoba Lose (for Bass Explorer) and Ziphozakhe Mashologu (for RNA STAR index builder). This work was supported by the South African Research Chairs Initiative of the Department of Science and Technology and National Research Foundation of SA and the National Research Foundation, Prime Minister's Office, Singapore under its Competitive Research Program (CRP Award No. NRF-CRP7-2010-01).