BioHackrXiv

Genome Annotation and Other Post-Assembly Workflows for the Tree of Life

Tom Brown[1], **Alice B. Dennis**[2], **and Jèssica Gómez-Garrido**[3]

1 Leibniz-Institut für Zoo- und Wildtierforschung (IZW) im Forschungsverbund Berlin e.V. Alfred-Kowalke-Straße 17 10315 Berlin Deutschland **2** University of Namur, URBE, Laboratory of Adaptive Evolution, Genomics, and Physiology, Namur, Belgium **3** CNAG

BioHackathon series :
BioHackathon Europe 2023
Barcelona, Spain, 2023
ERGA Annotation Committee

Submitted : 01 Nov 2023

Published by BioHackrXiv.org

Introduction

The European Reference Genome Atlas (ERGA (Mazzoni et al., 2023)) is an international consortium of over 700 researchers from across Europe who are committed to cataloguing eukaryotic biodiversity through the generation of high-quality reference genomes as a response to declines in biodiversity. Development and implementation of standard procedures and bioinformatic pipelines is essential for achieving ERGA's goal.

Annotation of reference genomes is necessary for many downstream analyses, but tools and workflows for genome annotation are still neither performed in a consistent manner between research groups, nor prepared for scaling up to the thousands of genomes which are to be produced as part of ERGA's efforts. As the number of sequenced genomes is rapidly increasing, there is a great need to develop standard, efficient and reproducible genome annotation workflows. In this project we will develop and implement pipelines for performing and evaluating genome annotations. As part of the Elixir BioHackathon 2023, we identified appropriate pipeline components, created environments or containers to facilitate their installation, and embedded them into robust pipelines using standard workflow managers, in this case Nextflow, Snakemake and Galaxy.

Goals

Testing pipelines on a wide range of species The majority of tools developed for the annotation of genetic sequences have been developed with model organisms in mind. We wished to determine which tools and pipelines were appropriate for which taxa and what limitations exist with established tools. Testing different pipelines and workflows on a variety of compute infrastructures A key aspect of the FAIR principles is the interoperability of metadata, data and workflows. Any pipeline produced by researchers should be written wherever possible in a way that can be deployed on a variety of compute infrastructures with minimal intervention and trouble-shooting. This BioHackathon gives a fantastic opportunity for researchers from a large number of institutes and research environments to come together and share their experiences on different compute environments. Establish a set of criteria to evaluate the efficacy of each pipeline The evaluation of genome annotation remains an outstanding question in the field, with no defined set of tools, software and criteria yet established to determine whether one annotation is necessarily better than another. In this project, we aimed to test a number of tools and establish a set of easily comparable criteria for assessing annotation quality. Establish a "minimum criteria" for annotation quality As ERGA and other Earth Biogenome Project-affiliated initiatives increase the number of published genome sequences, the compute cost required in order to evaluate and annotate these sequences will grow inordinately. We must keep in mind as researchers that Earth's resources are finite and must be aware of our environmental impact with each computational pipeline run. It is not feasible to be expected to run multiple workflows for each generated sequence in order to produce the best annotation possible. We hope to establish when an annotation can be considered "good enough" to answer

the biological questions researchers have for their genome to avoid over-computation in the future.

Genomes for Biodiversity

As part of this project, we wished to test robust annotation pipelines on genomes representing the diversity of organisms which will be produced as part of ERGA's goal to produce reference quality genomes for all eukaryotes in Europe. Included in this project were a mixture of previously published genomes, as well as assemblies produced as part of the ERGA Pilot Project (McCartney et al., 2023) or by ERGA-affiliated researchers directly contributing to the BioHackathon. Included in our analysis were genome sequences for the Cauliflower Coral Pocillopora meandrina, the Violet Copper Butterfly Helleia helle, the Lesser Trefoil Trifolium dubium (GCA_951804385.1), the sponge Phakellia ventilabrum, the Coffee-bean Snail Melampus jaumei, the Fruit Fly Drosophila melanogaster (GCA_000001215.4), and Chromosome 19 from Human Genome Release 38 (GCA_000001405.29).

Pipeline Structure(s)

The pipelines generated as part of this BioHackathon generally followed a consistent workflow, relying on a repeat-masked genome, species-specific transcript sequencing data, a database of protein sequences and software aimed to identify and predict de-novo gene structures based on the given evidences.

Our first tested pipeline (Fig. X - hereafter referred to as UNIL pipeline, developed by SJD) includes mappings from paired-end RNA-seq data specific to the species of interest and proteins from the Swissprot database (Release 2023_04 of 13-Sep-2023 (Bateman et al., 2022)). Following de-novo repeat modelling using the genome assembly as input, the Swissprot protein sequences and mapped RNA-seq reads are given as evidence data alongside the repeat-masked genome into Braker3 (Gabriel et al., 2023). The resulting gff3 file is the output used for downstream evaluation and analysis.

Our second tested pipeline (Fig. X - hereafter referred to as CNAG pipeline, developed by JGG) includes additional steps where a set of de-novo transcripts are assembled from the input RNA-seq reads and multiple evidences from Augustus (Stanke et al., 2008), GeneMark (Brůna et al., 2020) and GeneID (Alioto et al., 2018) are combined with RNA-seq and protein alignments using EvidenceModeler (EVM (Haas et al., 2008)). Furthermore isoform information is included by running PASA after initial gene models are created from EVM.

Our third tested pipeline (Fig. X - hereafter referred to as ANNOTATO, developed by PD) incorporates a number of the previous steps and allows for the addition of long-read RNA reads, for example a PacBio Iso-seq library, and uses Funannotate (Palmer & Stajich, 2020) to include to add further gene predictions alongside Braker3.

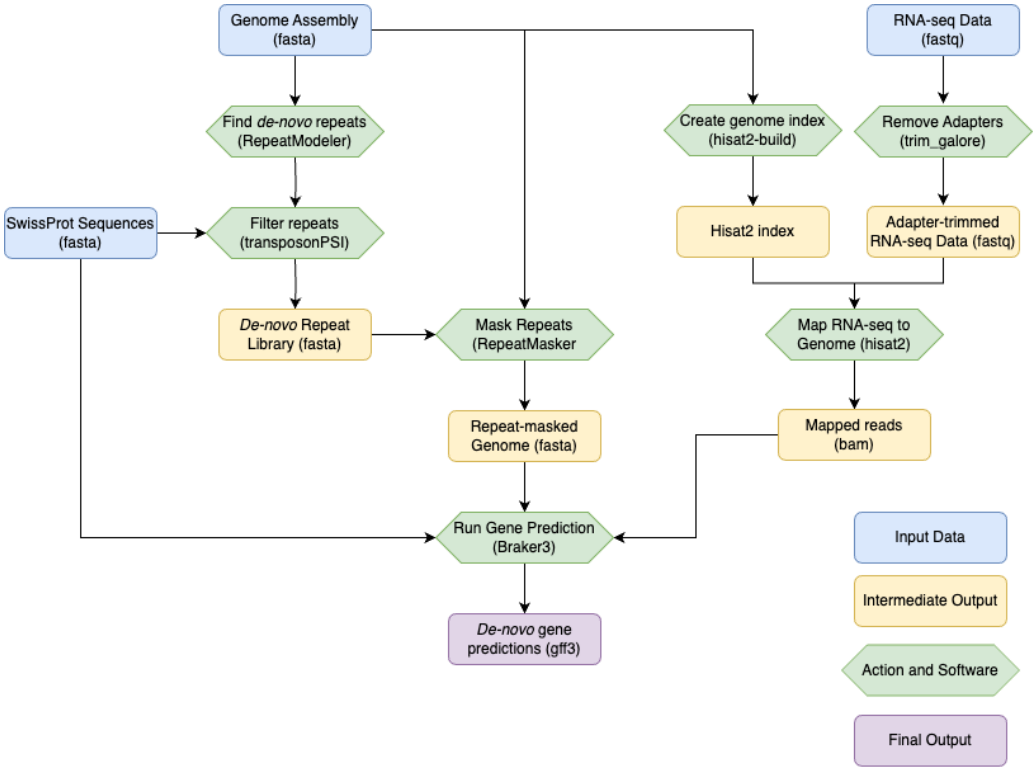

Figure 1 – Workflow for UNIL Pipeline

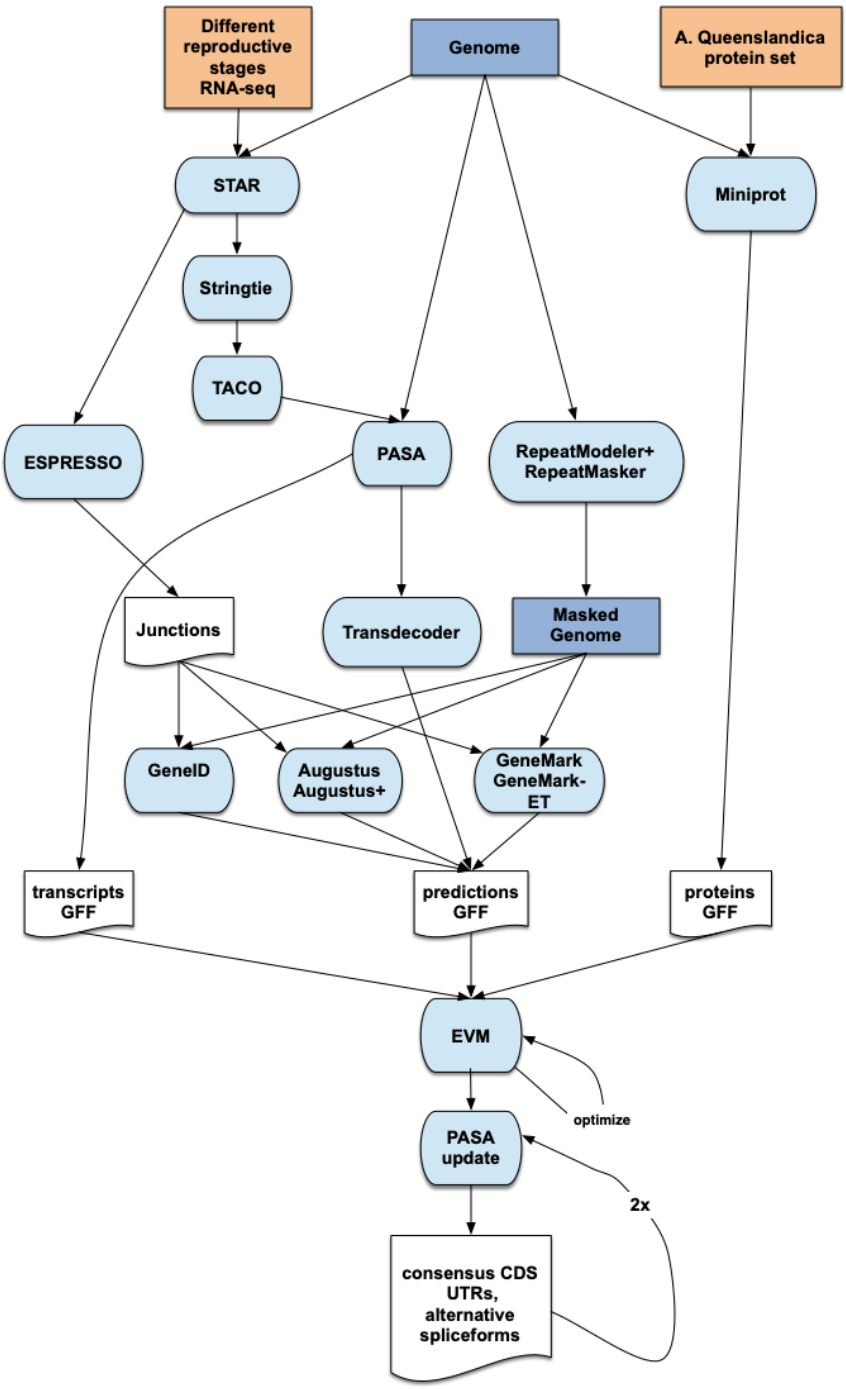

Figure 2 – Workflow for CNAG Pipeline

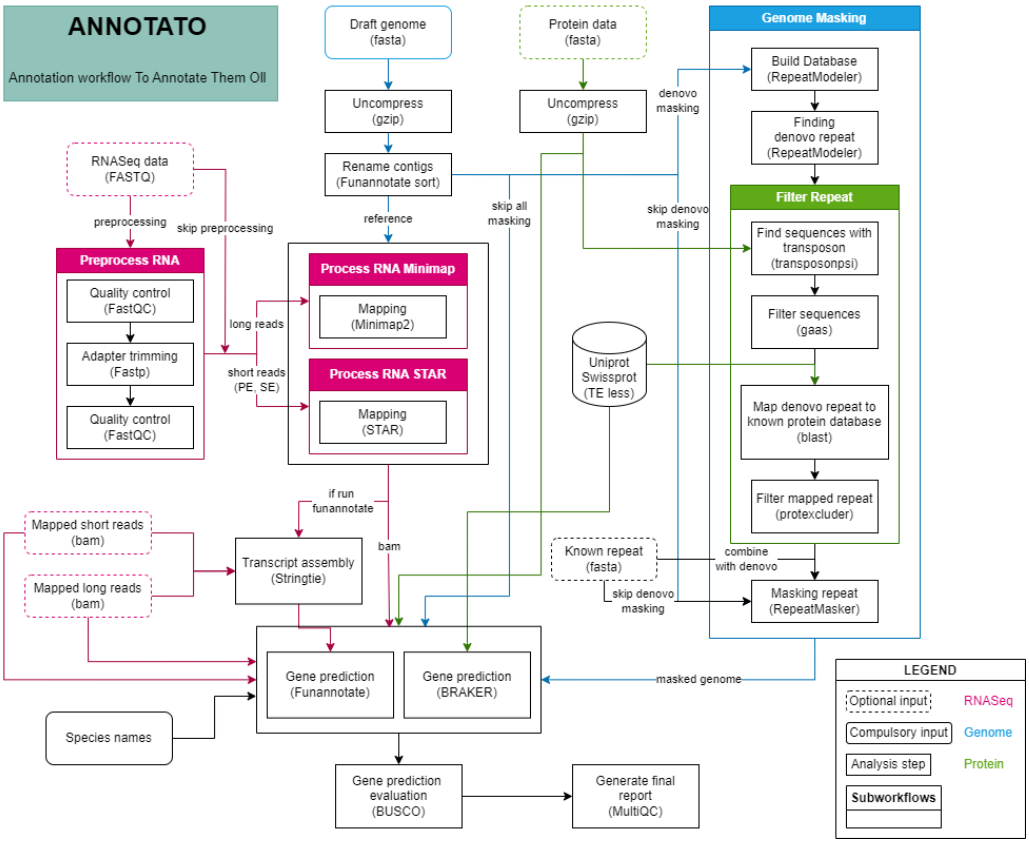

Figure 3 – Workflow for ANNOTATO Pipeline

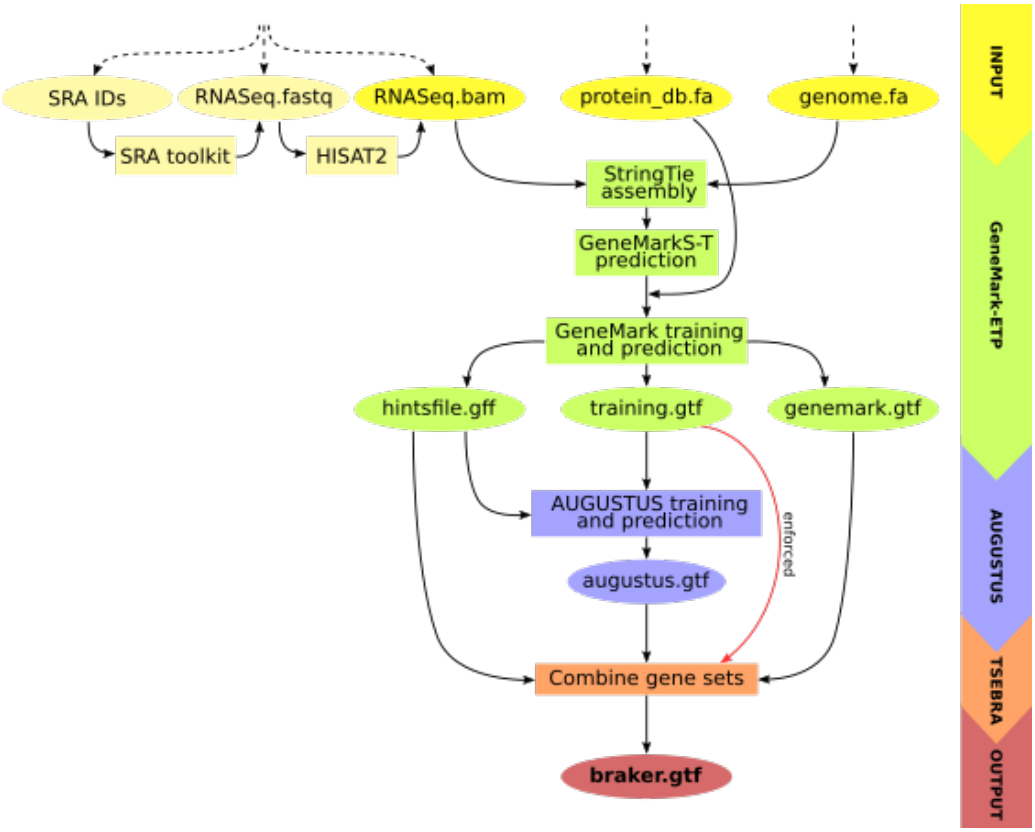

Figure 4 – Braker3 workflow

How to Evaluate Your Genome

A key outstanding question in the field of genome annotation is how to evaluate the quality of an annotation. Some key metrics to consider are the number of total genes annotated, which is itself taxon specific, the size of introns, normally correlated with the size of the genome, the number of single-exon genes, the number of distinct isoforms per gene and the completeness of the annotation regarding the expected content of the proteome which is encoded.

To evaluate the completeness of the annotation, we utilised BUSCO (Manni et al., 2021) and OMArk (Nevers et al., 2022). Both of these tools are designed to determine what proportion of protein-coding genes expected to be present in the genome from an ancestral lineage are present in the annotated sequences. OMArk also gives an indication of whether annotated sequences are potentially from contaminated sources or inconsistent with the identified lineage. The metrics of "Gene Completeness" output by both BUSCO and OMArk are key to identifying whether the annotation created is a true representation of the proteome of the species.

Acknowledgements

References

Alioto, T., Blanco, E., Parra, G., & Guigó, R. (2018). Using geneid to identify genes. *Current Protocols in Bioinformatics*, *64*(1). https://doi.org/10.1002/cpbi.56

Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., Costa Gonzales, L. J. da, Hatton-Ellis, E., Hussein, A., Ignatchenko, A., . . . Zhang,

J. (2022). UniProt : The universal protein knowledgebase in 2023. *Nucleic Acids Research*, *51*(D1), D523–D531. https://doi.org/10.1093/nar/gkac1052

Brůna, T., Lomsadze, A., & Borodovsky, M. (2020). GeneMark-EP+ : Eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics*, *2*(2). https://doi.org/10.1093/nargab/lqaa026

Gabriel, L., Brůna, T., Hoff, K. J., Ebel, M., Lomsadze, A., Borodovsky, M., & Stanke, M. (2023). BRAKER3 : Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *bioRxiv*. https://doi.org/10.1101/2023.06.10.544449

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology*, *9*(1), R7. https://doi.org/10.1186/gb-2008-9-1-r7

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update : Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654. https://doi.org/10.1093/molbev/msab199

Mazzoni, C. J., Ciofi, C., & Waterhouse, R. M. (2023). Biodiversity : An atlas of european reference genomes. *Nature*, *619*(7969), 252–252. https://doi.org/10.1038/d41586-023-02229-w

McCartney, A. M., Formenti, G., Mouton, A., DePanis, D., Marins, L. S., Leitão, H. G., Diedericks, G., Kirangwa, J., Morselli, M., Salces-Ortiz, J., Escudero, N., Iannucci, A., Natali, C., Svardal, H., Fernández, R., Pooter, T. D., Joris, G., Strazisar, M., Wood, J., … Mazzoni, C. J. (2023). The european reference genome atlas : Piloting a decentralised approach to equitable biodiversity genomics. *bioRxiv*. https://doi.org/10.1101/2023.09.25.559365

Nevers, Y., Rossier, V., Train, C. M., Altenhoff, A., Dessimoz, C., & Glover, N. (2022). Multifaceted quality assessment of gene repertoire annotation with OMArk. *bioRxiv*. https://doi.org/10.1101/2022.11.25.517970

Palmer, J. M., & Stajich, J. (2020). *Funannotate v1.8.1 : Eukaryotic genome annotation*. Zenodo. https://doi.org/10.5281/ZENODO.4054262

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, *24*(5), 637–644. https://doi.org/10.1093/bioinformatics/btn013