From scaffold to submission in a day: a new software pipeline for rapid genome annotation and analysis

Sascha Steinbiss[1], Fatima Silva[2], Brian Brunk[3], Bernardo Foth[1], Christiane Hertz-Fowler[2], Matt Berriman[1], Thomas Dan Otto[1]

[1] Wellcome Trust Sanger Institute, Hinxton, UK. Email: ss34@sanger.ac.uk

[2] University of Liverpool, Liverpool, UK.

[3] University of Pennsylvania, Philadelphia, PA, USA.

Project Website: https://github.com/satta/annot-nf
Source Code: https://github.com/satta/annot-nf
License: ISC (BSD-like)

Technological improvements have enabled genome sequencing and assembly to become efficient and accurate, but this is driving an increased need to annotate newly assembled genomes with the structure and function of genes. These annotations underpin subsequent comparative analyses to identify differences between individual species or strains, such as loss or gain of common and/or species-specific genes and functions.

While established off-the-shelf software solutions for complete genome annotation are readily available for prokaryotes, the need for an efficient eukaryote equivalent remains. Existing heavyweight eukaryotic annotation pipelines are optimized for delivering accurate protein coding gene models but usually do not address partial or non-coding genes, pseudogenes or functional annotations, nor do they generate a product that is ready to submit to public databases (a requirement for publication). The latter involves the preparation of complete annotation results (full gene sets, genomic sequences, protein sequences, functional annotation) in validated and standardized annotation formats (e.g. GFF3, EMBL, GAF). This often manual preparation can result in a substantial bottleneck influencing the total turnaround time to database submission.

We present a new full-stack software pipeline for eukaryotic genome annotation. It accepts input in various states of assembly, covering all stages from pseudochromosome contiguation and gene finding to function assignment. While built on reliable *de facto* standard components such as AUGUSTUS, SNAP (gene finding), RATT (annotation transfer), OrthoMCL (clustering) and GenomeTools (annotation handling), the pipeline includes new and improved versions of existing software such as ABACAS2 (pseudochromosome assembly) as well as bespoke software, e.g. for pseudogene identification. Special care has been taken to make the pipeline produce validated and accurate output even for highly fragmented sequence inputs, as they are common in draft genomes.

The pipeline makes extensive use of modern, state-of-the-art workflow (Nextflow) and deployment technologies (Docker) to ensure scalability, reproducibility and portability for use on powerful stand-alone PCs as well as large compute clusters (e.g. SGE, LSF, SLURM) or cloud platforms (e.g. ClusterK, DNAnexus) with a minimum of effort. In addition, we have created a web-based annotation interface to the pipeline, allowing researchers from the parasitology community to run annotation jobs and comparative analysis tasks on user-provided genomes as well as obtain and visualize the results.

We exemplify the use of the pipeline to annotate a series of new kinetoplastid parasite genomes as well as improve existing parasite annotations.