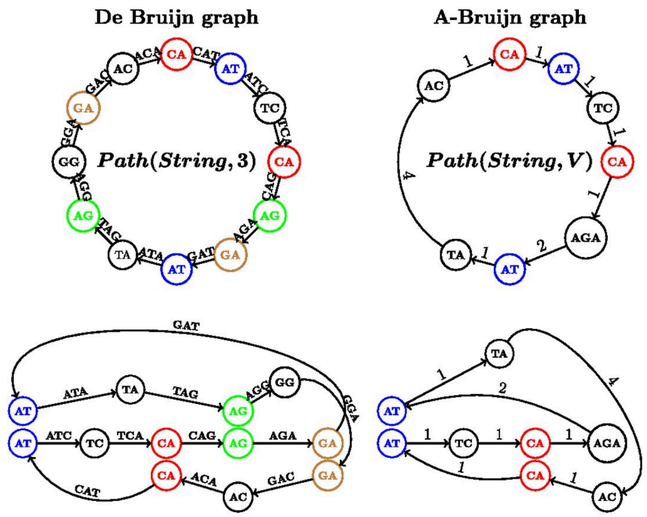
Genome Assembly and Annotation

Jason Stajich
Univ of California, Riverside

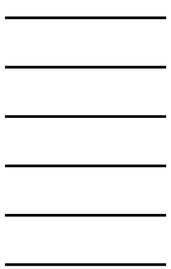
Genome Assembly

Assembly approaches

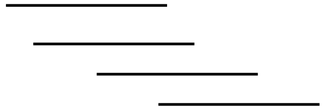
De Brujin Graphs



Raw reads

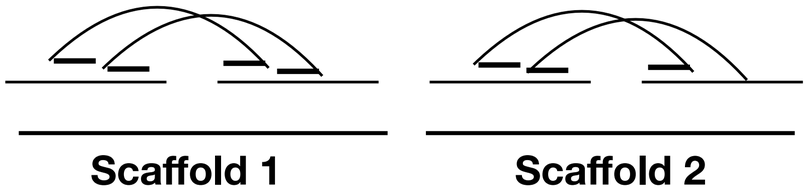


Overlap Consensus



Contigs

Scaffolds



Annotation

- Identify genes, repetitive elements, tRNA genes, etc
- One tool which combines prediction, training, and functional annotation into single package is called funannotate. Developed for Fungi but useful for many eukaryotic systems.
- https://funannotate.readthedocs.io/en/latest/ and https://github.com/nextgenusfs/funannotate/ and https://doi.org/10.5281/zenodo.1134477
- Can be installed with conda using bioconda or available as docker images
- Caveats
  Does not annotate organelles and mitochondria; relies on a host of other tools so some installations are complicated.
- Can run *de novo* repetitive element finding, but some steps maybe better outside the tools

Annotation I

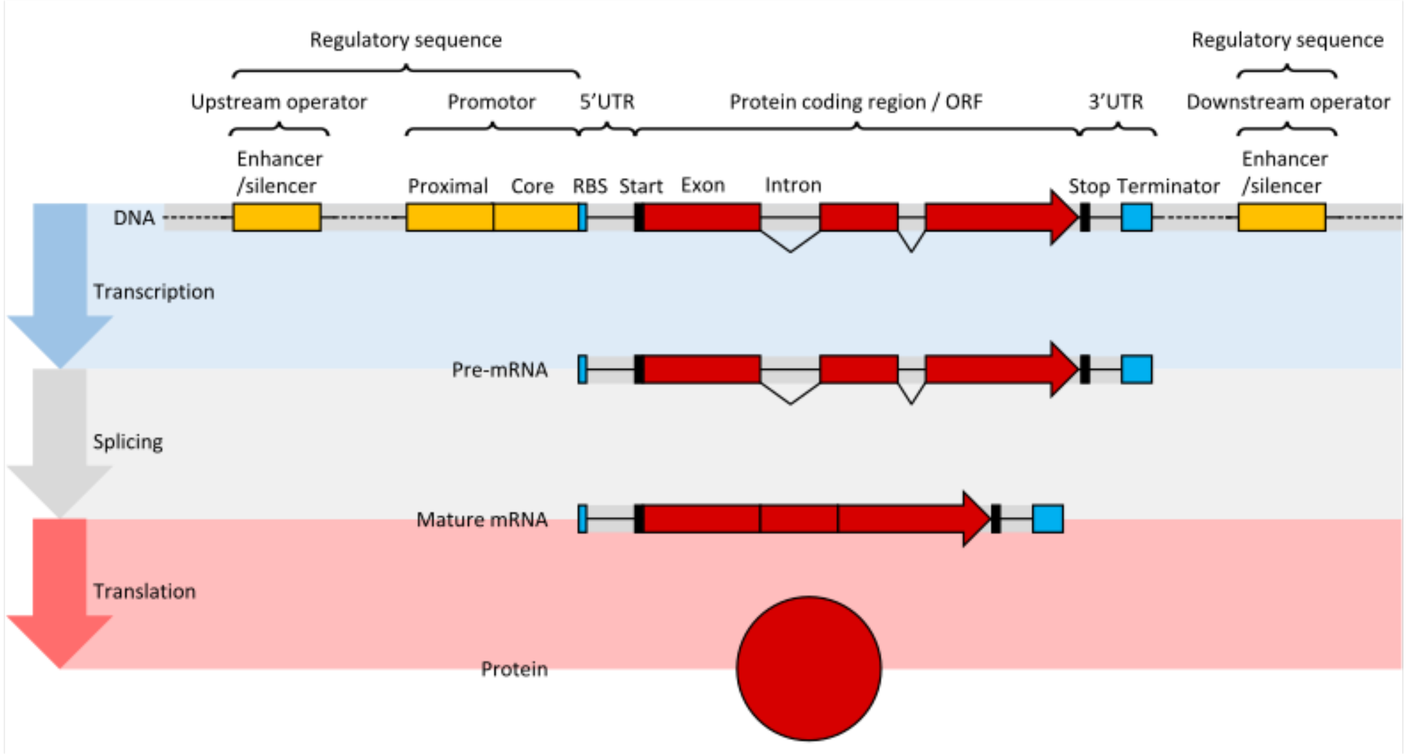
- funannotate mask - Identify repetitive sequences
  - RepeatMasker, tantan
- *funannotate train - Train gene predictors*
  - Align mRNA transcripts to genome to identify exon/gene regions exonerate;
  - Align RNA-Seq reads to genome, assemble transcripts with Trinity
  - Refinement with PASA (Program to Assemble Spliced Alignments)
  - Train *ab initio* gene predictor from these spliced models

Annotation II

- funannotate predict
  - Align protein (and mRNA Transcripts) to genome as evidence to support predictions; refine alignment with exonerate to spliced exons
  - Gene prediction (augustus, SNAP, genemark, glimmerhmm)
  - Combine predictions with EVM

Annotation III

- *funannotate update - refine gene models with mRNA*
  - Run PASA with models to extend Untranslated regions
  - Alternative Splicing



Annotation IV

- funannotate annotate - align to databases to add functional info
  - BLAST against uniprot, swissprot, MEROPs database
  - HMM searches against Pfam, CAZyDB
  - Incorporate InterPro, AntiSMASH (secondary metabolite prediction)
  - Produces annotated genome files ready to upload to NCBI