09:00-12:00 Genome annotation

- 09:00-09:30 Introduction to the study system and infrastructure
    - Making sure you have access to Fox
    - Submit the first set of jobs
        i. Repeat mask
        ii. Mapping protein sets
        iii. *Ab initio* gene prediction
- 09:30-11:55 Introduction to genome annotation
    - Work through the rest of the programs
        i. EvidenceModeler
        ii. BUSCO
        iii. Functional annotation
- 11:55-12:00 Summary

12:00-13:00 Lunch

13:00-14:00 Comparative genomics

- Introduction to comparative genomics
- Setting up OrthoFinder

14:00-14:15 Break

14:15-16:00

- Running OrthoFinder on proteins and CDS

Genome annotation and comparative genomics

Part 1, morning

Teachers: Bram Danneels, Helle Tessand Baalsrud, José Cerca, Ole K. Tørresen

Oslo Bioinformatics Workshop Week 2023

11th December

Learning outcomes

After attending the workshop learners should:

1. Know how to use some of the more popular tools for creating genome annotations

2. Know how to validate annotated genes

3. Know some comparative genomics analyses

The infrastructure

- We will do computations on Fox, computing cluster at UiO
- Need to create an account and apply for access at ec146
- We will go through this together

https://shorturl.at/swMZ2

What is a biodiversity genomics project?

- Produce genomes for all kinds of different species (DToL, ERGA, EBP)
- Vertebrate Genomes Project is targeted (all vertebrates), maybe not a biodiversity genomics project

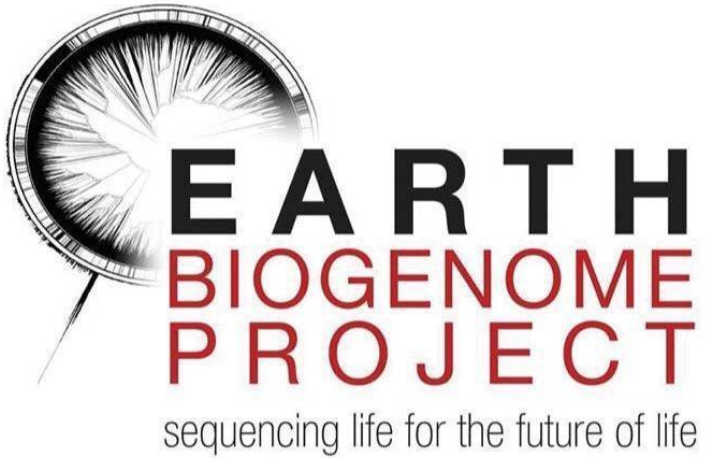
EARTH BIOGENOME PROJECT
sequencing life for the future of life

Darwin TREE of LIFE

What is the Earth Biogenome Project?

- Better understanding of biology and evolution
- Conserve, protect and restore biodiversity
- Create new benefits for society and human welfare

-> sequence all eukaryotes



EARTH BIOGENOME NOR PROJECT

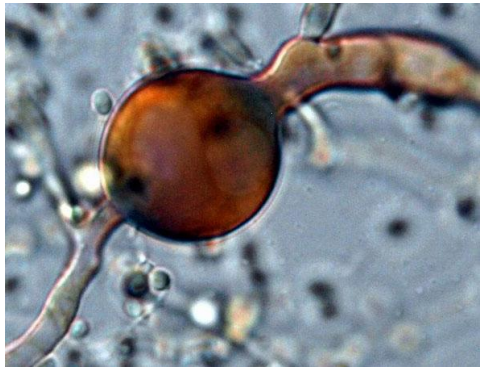
Funded by the Research Council of Norway

- Phase 1 2021-2024 (30 MNOK)
  - Do 100-150 species
  - Norwegian, marine and arctic species
  - Coordination with ERGA, DToL, VGP, EBP and other projects (e.g. https://goat.genomehubs.org/)

- Preparation for 2 phase has begun

The study system

- For genome annotation we will use *Umbelopsis ramanniana*, an abundant soil fungus
- Genome is about 25 Mbp
- Number of genes is about 10000



Courtesy of Alena Kubátová

Start some jobs

- Go to https://shorturl.at/swMZ2
- Start reading Introduction to the infrastructure and study system
- Go through and submit the first three jobs:
  - Repeat masking
  - Mapping proteins
  - *Ab initio* gene prediction (wait until repeat masking is done, couple of minutes)

We will continue after all have done this.

09:00-12:00 Genome annotation

- 09:00-09:30 Introduction to the study system and infrastructure
    - Making sure you have access to Fox
    - Submit the first set of jobs
        i. Repeat mask
        ii. Mapping protein sets
        iii. *Ab initio* gene prediction
- 09:30-11:55 Introduction to genome annotation
    - Work through the rest of the programs
        i. EvidenceModeler
        ii. BUSCO
        iii. Functional annotation
- 11:55-12:00 Summary

12:00-13:00 Lunch

13:00-14:00 Comparative genomics

- Introduction to comparative genomics
- Setting up OrthoFinder

14:00-14:15 Break

14:15-16:00

- Running OrthoFinder on proteins and CDS
-
-

Genome annotation

What?

- The process of finding functional elements in a genome

Structural annotation

- Find where the genes are

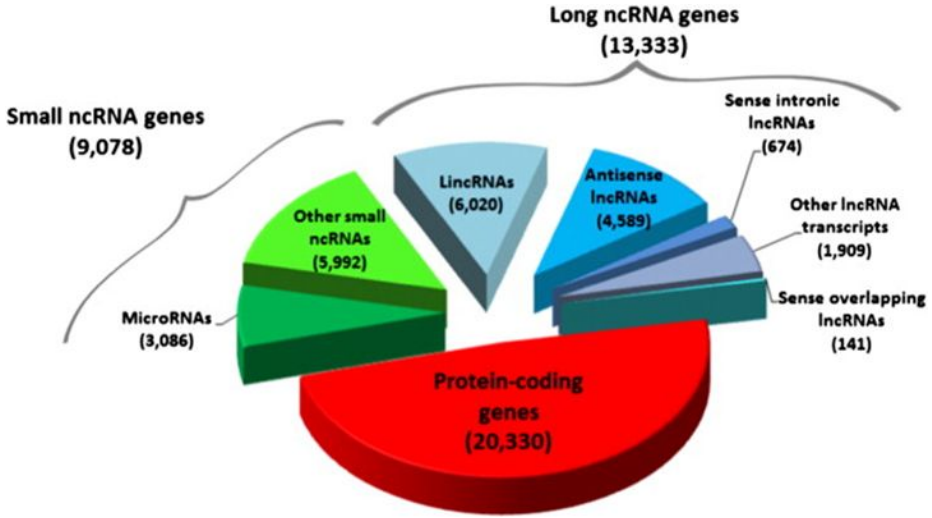
Functional annotation

- Assign functions to the genes

Genome annotation

What are we looking for?

- Genes
  - Protein coding
  - Non-coding genes
  - rRNA/tRNA
- Repeats
- Regulatory elements
- Telomeres/centromeres
- …



Hrdlickova et al (2012); *BBA*

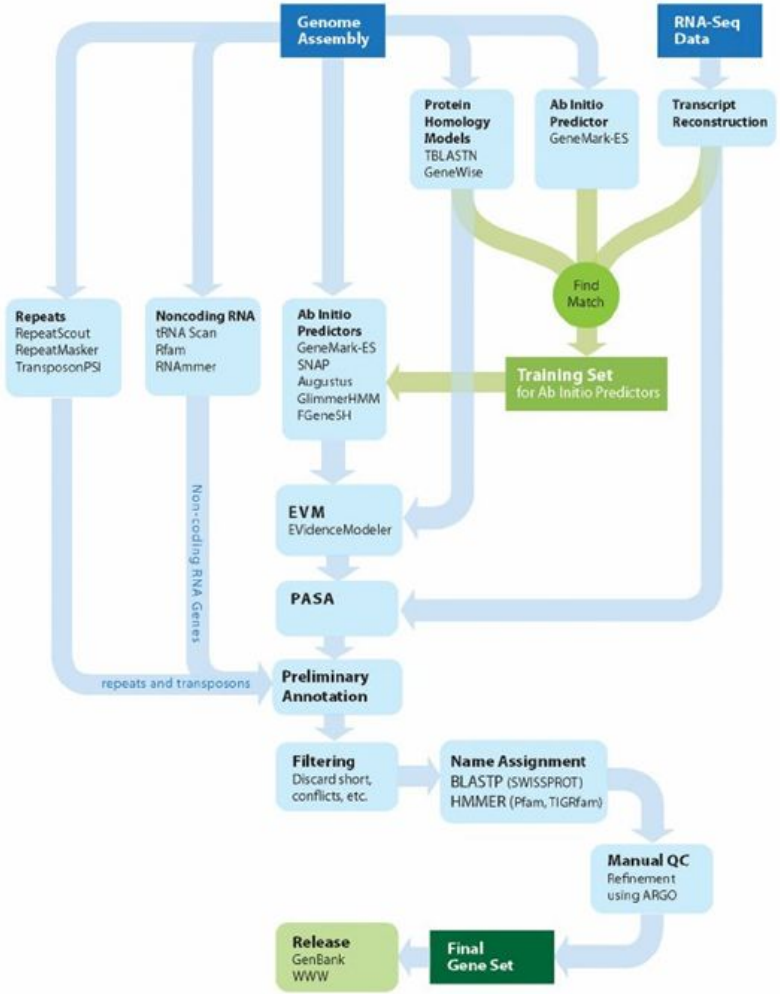
Genome annotation

Many tools & pipelines exist

E.g. Broad Institute pipeline

General workflow:

- Repeat Masking
- Gene prediction
  - *ab initio*
  - from protein and/or RNA-seq data
  - combination
- Functional gene prediction

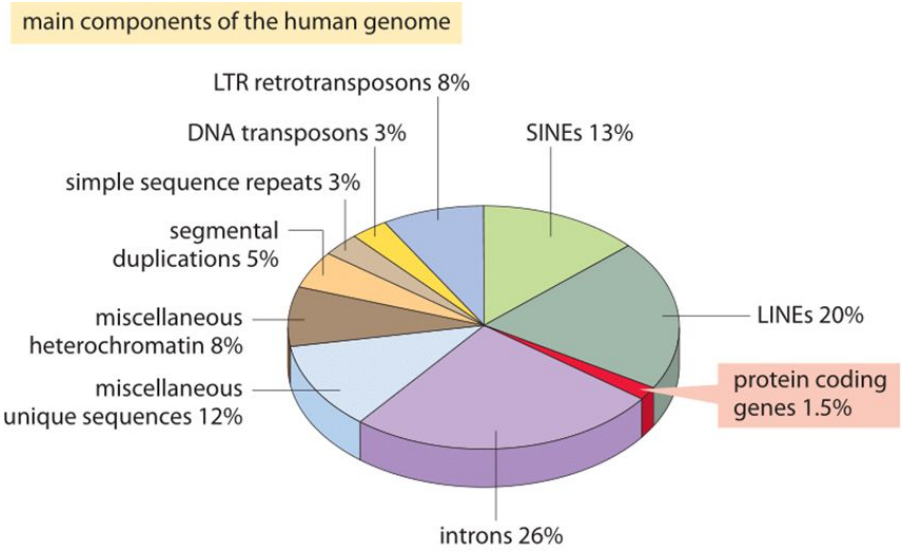


Haas et al. (2011); *Mycology*

Repeat Masking

Important first step in genome annotation

- Prevent annotation of spurious genes (e.g. transposons)
- Reduce search space



Gregory (2001); *Nature rev. in Genetics*

Gene prediction

What information can be used to find genes?

- Inherent sequence information
  - start/stop codon, GC-content, codon usage
  - *ab initio* gene prediction

- External information
  - Expression data - RNA-seq
  - Homology data - Protein

Ab initio gene prediction

Use sequence information to prediction where genes are

Needs to be “trained” to find out what a good gene looks like

- Unsupervised learning (self-training)

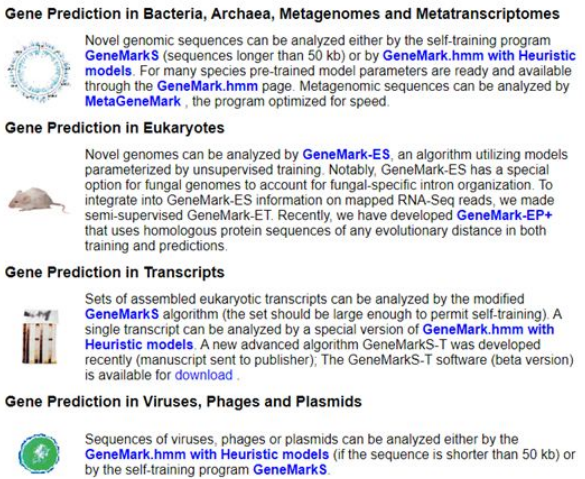
  Use general parameters

  → predict genes

  → train model on predicted genes

  → predict using new model parameters

  → …



Gene Prediction in Bacteria, Archaea, Metagenomes and Metatranscriptomes

Novel genomic sequences can be analyzed either by the self-training program **GeneMarkS** (sequences longer than 50 kb) or by **GeneMark.hmm with Heuristic models**. For many species pre-trained model parameters are ready and available through the **GeneMark.hmm** page. Metagenomic sequences can be analyzed by **MetaGeneMark** , the program optimized for speed.

Gene Prediction in Eukaryotes

Novel genomes can be analyzed by **GeneMark-ES**, an algorithm utilizing models parameterized by unsupervised training. Notably, GeneMark-ES has a special option for fungal genomes to account for fungal-specific intron organization. To integrate into GeneMark-ES information on mapped RNA-Seq reads, we made semi-supervised GeneMark-ET. Recently, we have developed **GeneMark-EP+** that uses homologous protein sequences of any evolutionary distance in both training and predictions.

Gene Prediction in Transcripts

Sets of assembled eukaryotic transcripts can be analyzed by the modified **GeneMarkS** algorithm (the set should be large enough to permit self-training). A single transcript can be analyzed by a special version of **GeneMark.hmm with Heuristic models**. A new advanced algorithm GeneMarkS-T was developed recently (manuscript sent to publisher); The GeneMarkS-T software (beta version) is available for download .

Gene Prediction in Viruses, Phages and Plasmids

Sequences of viruses, phages or plasmids can be analyzed either by the **GeneMark.hmm with Heuristic models** (if the sequence is shorter than 50 kb) or by the self-training program **GeneMarkS**.

Example: Genemark (http://exon.gatech.edu/GeneMark/)

Ab initio gene prediction

Use sequence information to prediction where genes are

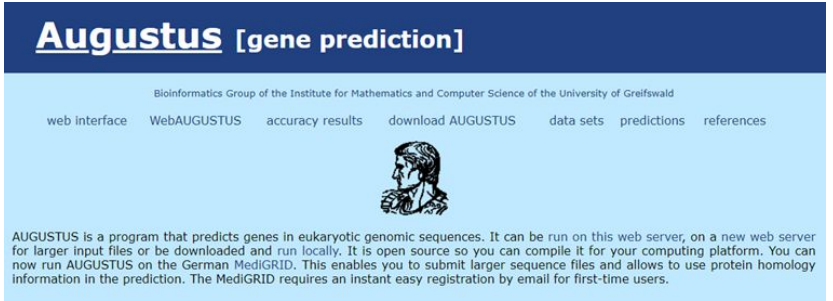
Needs to be “trained” to find out what a good gene looks like

- Supervised learning (learning by example)

  Provide a set of example genes

  → train model on predicted genes

  → predict using trained model



Example: Augustus (https://bioinf.uni-greifswald.de/augustus/)

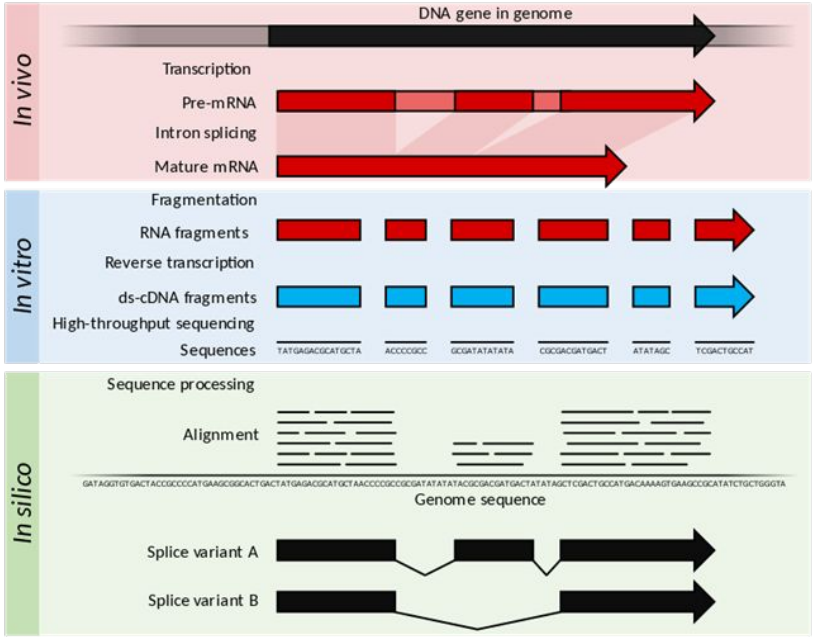
Using extrinsic evidence

We can use experimental evidence to help predict genes

→ Creating “hints”

RNA-seq:

- Gene location
- Intron/Exon boundaries
- Splice variations



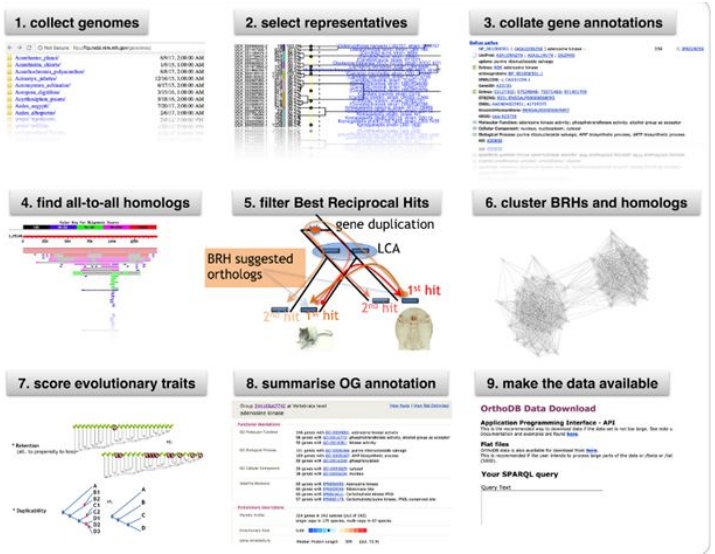
Lowe *et al*. (2017); *PLoS Computational Biology*

Using extrinsic evidence

We can use experimental evidence to help predict genes

→ Creating “hints”

OrthoDB



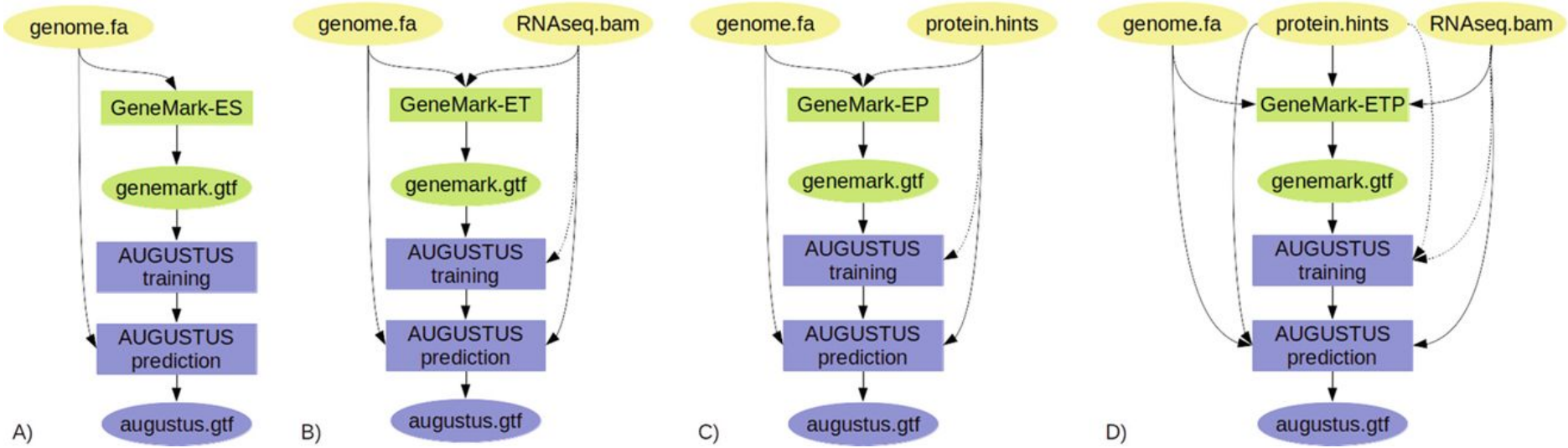
Protein data

- Location of genes
- Some splicing information
- Often data from

Orthodb (https://www.ezlab.org/orthodb.html)

Using extrinsic evidence

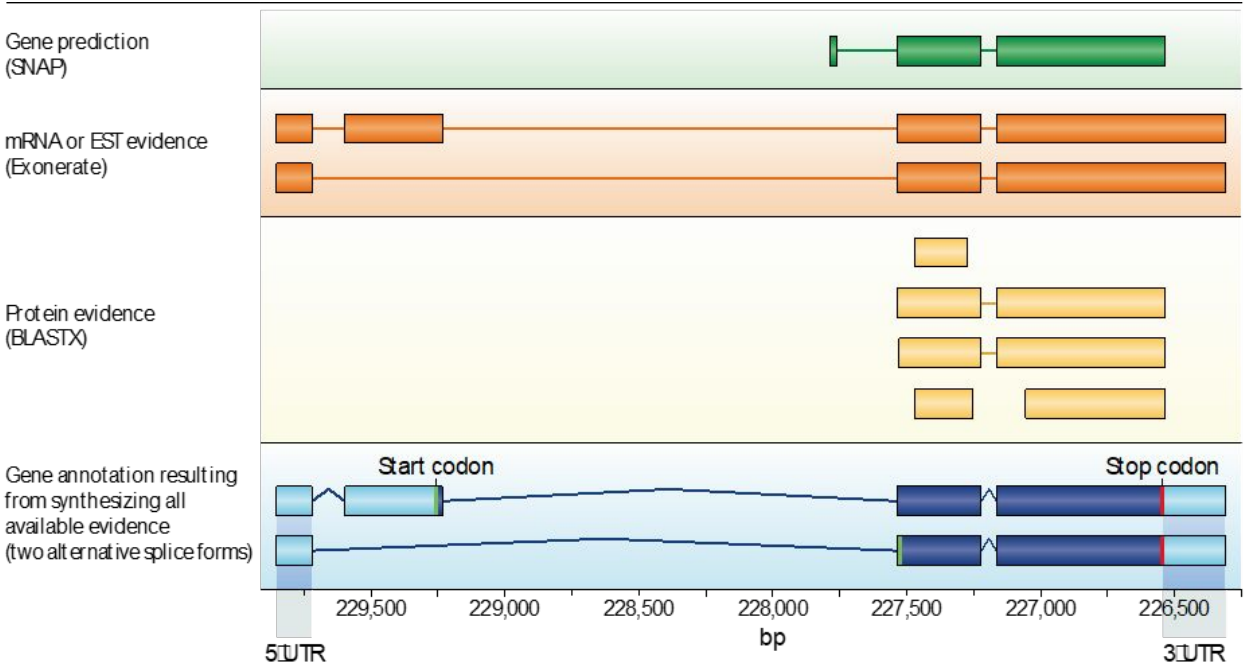
Combined gene prediction - Using different data for training



Braker2 pipelines; Hoff *et al*. (2019); *Gene prediction*

Using extrinsic evidence

Combined gene prediction - Combining predictions



Yandell and Ence (2012); *Nature reviews genetics*

Functional prediction

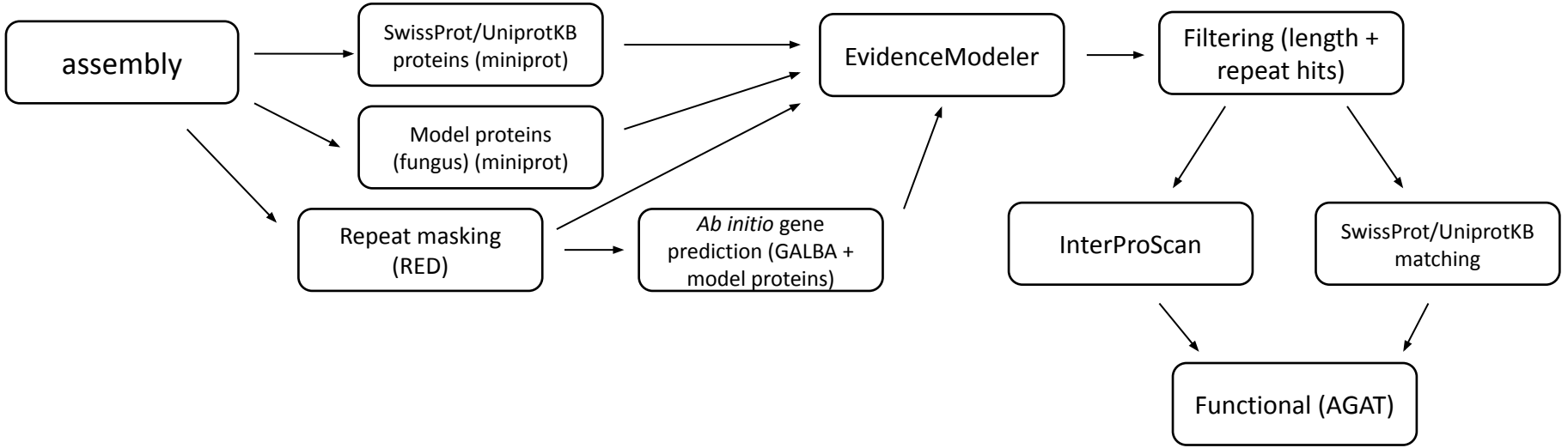
Gene functions are generally applied based on similarity:

- Detection of protein domains
  - PFAM, CDD

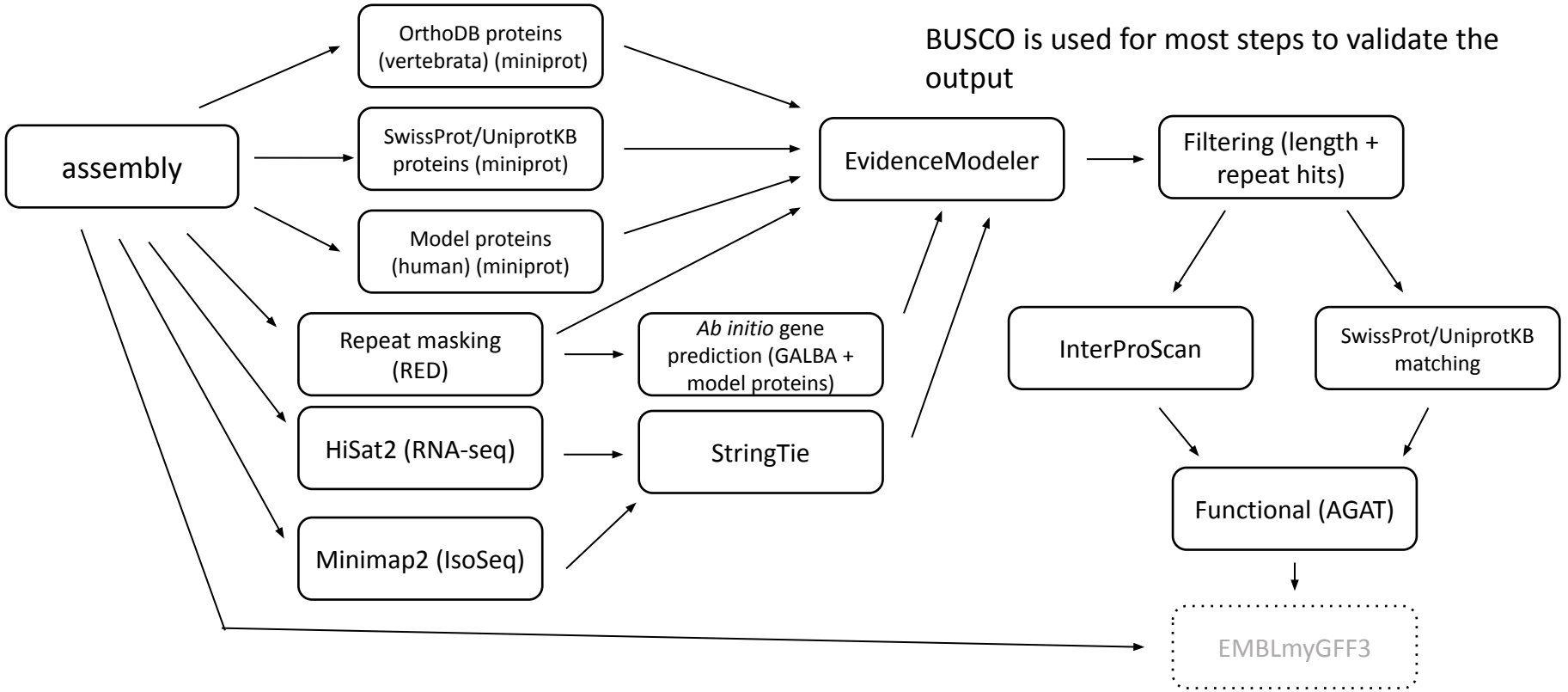
- Sequence similarity to known proteins
  - Define function based on homology principle

The annotation pipeline in this workshop

BUSCO is used for most steps to validate the output



The EBP-Nor annotation pipeline



09:00-12:00 Genome annotation

- 09:00-09:30 Introduction to the study system and infrastructure
    - Making sure you have access to Fox
    - Submit the first set of jobs
        i. Repeat mask
        ii. Mapping protein sets
        iii. *Ab initio* gene prediction
- 09:30-11:55 Introduction to genome annotation
    - Work through the rest of the programs
        i. EvidenceModeler
        ii. BUSCO
        iii. Functional annotation
- 11:55-12:00 Summary

12:00-13:00 Lunch

13:00-14:00 Comparative genomics

- Introduction to comparative genomics
- Setting up OrthoFinder

14:00-14:15 Break

14:15-16:00

- Running OrthoFinder on proteins and CDS

Summary

- You have now annotated a genome
- You should be able to validate the annotated genes
- You might have gotten around 6600 genes annotated, but only about 73-80 % complete BUSCO genes. Why?
- The other annotation had around 10600 genes and 90+ % complete BUSCO genes. What might the difference be?

Lots of possibilities: Genome annotation pipelines

- MAKER
  - Used quite a lot. Not been updated for years. I find it a bit cumbersome and not so flexible
- Funannoate
  - Originally made for fungi, but also used for other eukaryotes
  - Really flexible, can use lot of different data and processes a lot of different data
- Ensembl
  - In-house pipeline at Ensembl
  - https://www.youtube.com/watch?v=6dOcFUAKtu0
- NCBI
  - In-house pipeline at NCBI
  - Sometimes chooses genomes based on characteristics (N50, RNA-seq datasets available, etc)

Lots of possibilities: *Ab initio* gene predictors

- AUGUSTUS
  - Used quite a lot
  - Actively developed
  - Open source
- GeneMark
  - Used quite a lot
  - Actively developed
  - Restrictive license (just changed!)

Lots of possibilities: Others

- BRAKER1
  - Wrapper for AUGUSTUS to utilize RNA-seq data to train AUGUSTUS
  - Also trains GeneMark-ET
  - Relatively easy to use, almost all you need
- BRAKER2
  - Wrapper for AUGUSTUS to utilize protein data to train AUGUSTUS
  - Uses ProtHint to create hints for training (OrthoDB)
  - Also trains GeneMark-EP
  - Relatively easy to use, almost all you need
  - Ensembl uses it for rapid annotation of species without RNA-seq data
- BRAKER3 - update to BRAKER2
  - https://github.com/Gaius-Augustus/BRAKER

Lots of possibilities: Others continued

- Miniprot
  - Aligns protein data from a close relative and creates gene structures
  - Actively developed (by Heng Li)
  - Cannot find new genes
- MetaEuk
  - Aligns proteins from lots of different sources and creates gene structures
  - Actively developed
  - Cannot find new genes
- HiSat2 + StringTie
  - Map RNA-seq data to a genome
  - Assembles RNA-seq alignments to transcripts