# Genome Annotation

Presented by Suzy Strickler

# Download some files

```
cd /scratch

iget –rPT /iplant/home/shared/Botany2020NMGWorkshop/Annotation

cd annotation

iget –rPT
 /iplant/home/shared/Botany2020NMGWorkshop/embryophyta_odb9
```

# Objectives

- Understand steps involved in genome annotation
- Demonstrate types of data and tools that can be used in genome annotation
- Learn how to postprocess genome assemblies
- QC assembly results

# Goals of genome annotation

- Predict, categorize, and mask repetitive elements
- Determine gene structures as accurately as possible
- Predict possible functions of predicted genes
- Associate GO terms, domains, etc for downstream analyses

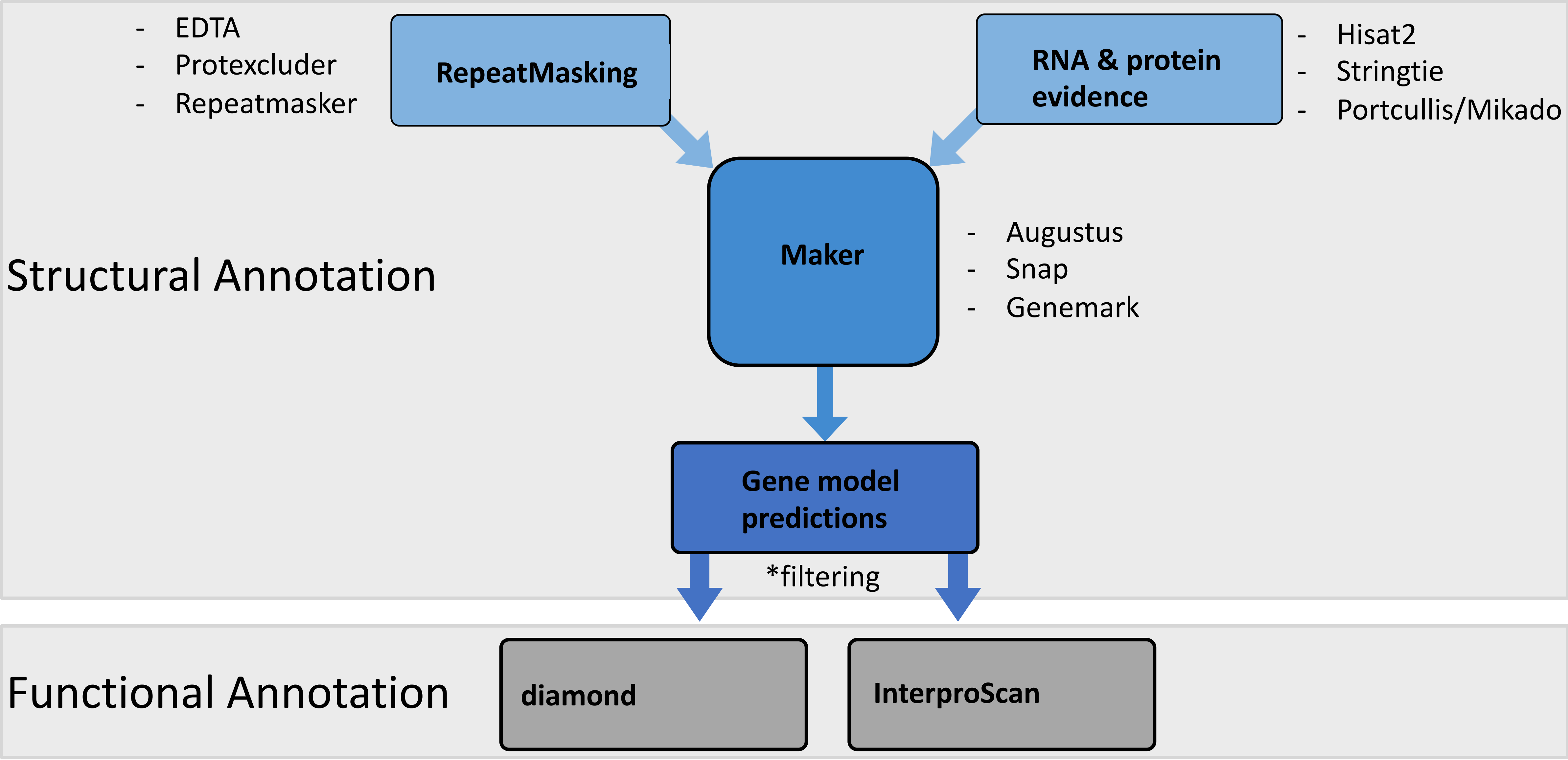
# Pre-annotation QC

- Assembly quality
- Errors - correction
- BUSCO metrics of genome

# Tools for structural annotation

- EDTA https://github.com/oushujun/EDTA
- Repeatmasker http://www.repeatmasker.org/
- Braker https://github.com/Gaius-Augustus/BRAKER
- Augustus https://github.com/Gaius-Augustus/Augustus
- Snap https://github.com/KorfLab/SNAP
- Genemark http://exon.gatech.edu/GeneMark/
- Maker https://www.yandell-lab.org/software/maker.html
- Apollo https://genomearchitect.readthedocs.io/en/latest/
- BUSCO https://gitlab.com/ezlab/busco_biocontainer
- Comparison to relative

# Annotation pipeline



Structural Annotation

- EDTA
- Protexcluder
- Repeatmasker

**RepeatMasking**

**RNA & protein evidence**

- Hisat2
- Stringtie
- Portcullis/Mikado

**Maker**

- Augustus
- Snap
- Genemark

**Gene model predictions**

*filtering

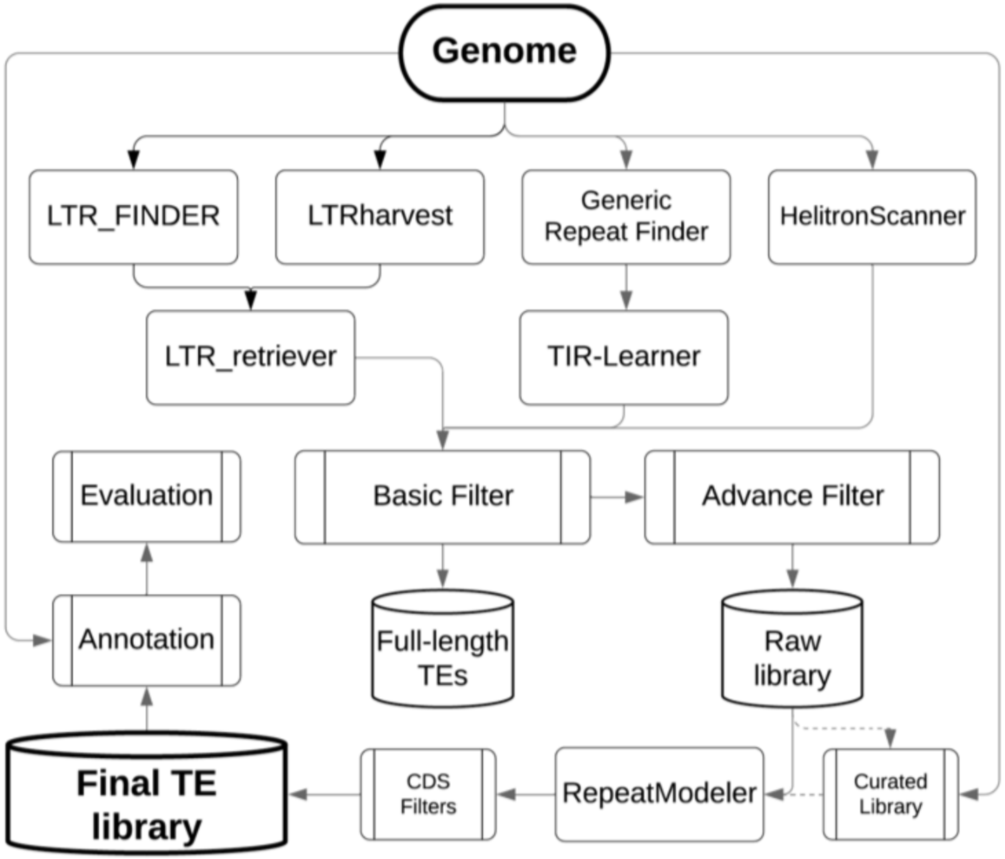
Functional Annotation

**diamond**

**InterproScan**

# EDTA

## The Extensive *de novo* TE Annotator (EDTA)



# Tools for functional annotation

- BLAST
- Diamond
- InterProScan
- Mercator

# Let’s annotate our *U. gibba* FLYE assembly!

- Genome file: Ugibba_FLYE_assembly.fasta.PolcaCorrected.fa.cat.all.gz
- RNA-seq from stem: /iplant/home/shared/Botany2020NMGWorkshop/raw_data/Ugibba/transcriptome/Ugibba_stemR1.fastq.gz
- Proteins: uniprot_sprot_plants.fasta

- All this stuff plus some output files in /iplant/home/shared/Botany2020NMGWorkshop/Annotation/Ugibba_FLYE_assembly.fasta.PolcaCorrected.fa.maker.output/Ugibba_FLYE_assembly.fasta.PolcaCorrected.fa.all.gff

# Masking

- https://github.com/bcbc-group/Botany2020NMGWorkshop/blob/master/annotation/repeatmasking.sh

# Read mapping

- https://github.com/bcbc-group/Botany2020NMGWorkshop/blob/master/annotation/hisat_se_annot.sh

# RNA-seq cleanup

- https://github.com/bcbc-group/Botany2020NMGWorkshop/blob/master/annotation/rnaseq_cleanup.sh

# Training augustus and snap

- https://vcru.wisc.edu/simonlab/bioinformatics/programs/augustus/docs/tutorial2015/training.html
- https://vcru.wisc.edu/simonlab/bioinformatics/programs/augustus/docs/tutorial2015/scipio.html
- https://github.com/bcbc-group/Botany2020NMGWorkshop/blob/master/annotation/training.sh

# Your turn to train augustus!

```
/opt/augustus-3.2.2/scripts/randomSplit.pl genes.raw.gb 100  #normally you would use gene.gb here, but this dataset is sparse

grep -c LOCUS genes.raw.gb*

sudo chown srs57 /opt/augustus/config/species/
/opt/augustus-3.2.2/scripts/new_species.pl --species=Ugibba

etraining --species=Ugibba genes.raw.gb.train

ls -ort $AUGUSTUS_CONFIG_PATH/species/Ugibba

augustus --species=Ugibba genes.raw.gb.test | tee firsttest.out
```

- These commands are also in https://github.com/bcbc-group/Botany2020NMGWorkshop/blob/master/annotation/training.sh

# Running maker

- https://github.com/bcbc-group/Botany2020NMGWorkshop/blob/master/annotation/maker.sh

# Postprocessing, Cleanup, and QC

- Remove Transposons
- Renaming
- complete genes only
- match to nr, e-20
- FPKM > 0.1
- AED > 1
- IPS domain
- Comparison to relative, length and number of genes
- Gene families
- BUSCO

# Functional annotation

- https://github.com/bcbc-group/Botany2020NMGWorkshop/blob/master/annotation/function_annot.sh

- Maker has several scripts for postprocessing files