# Genome Annotation

Presented by Suzy Strickler

# 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 annotation results

# We need these files

cd /scratch/Botany2020NMGWorkshop/

mkdir annotation2

cd annotation2

git clone https://github.com/bcbc-group/Botany2020NMGWorkshop.git

iget -rPT /iplant/home/shared/Botany2020NMGWorkshop/annotation/2transfer/contig_15.fasta

iget -rPT /iplant/home/shared/Botany2020NMGWorkshop/annotation/2transfer/SRR5046448_contig15_1.fastq

iget -rPT /iplant/home/shared/Botany2020NMGWorkshop/annotation/2transfer/SRR5046448_contig15_2.fastq

iget –rPT /iplant/home/shared/Botany2020NMGWorkshop/annotation/2transfer/mikado.loci.gff3

iget –rPT /iplant/home/shared/Botany2020NMGWorkshop/annotation/2transfer/scipio.yaml

iget –rPT /iplant/home/shared/Botany2020NMGWorkshop/annotation/2transfer/contig_15_masked.fasta

Iget –rPT /iplant/home/shared/Botany2020NMGWorkshop/annotation/2transfer/contig_15.maker.output/maker1.all.gff

iget -rPT /iplant/home/shared/Botany2020NMGWorkshop/embryophyta_odb9

# Changing docker file storage location

#stop docker

**sudo service docker stop**

#edit daemon.json

**emacs /etc/docker/daemon.json**

#and add:

```
{
 "graph": "/scratch/docker"
}
```

#copy current dir to new one

**sudo rsync -aP /var/lib/docker/ /scratch/docker**

#rename old docker dir, do no delete until you test config works

**sudo mv /var/lib/docker /var/lib/docker.old**

# 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

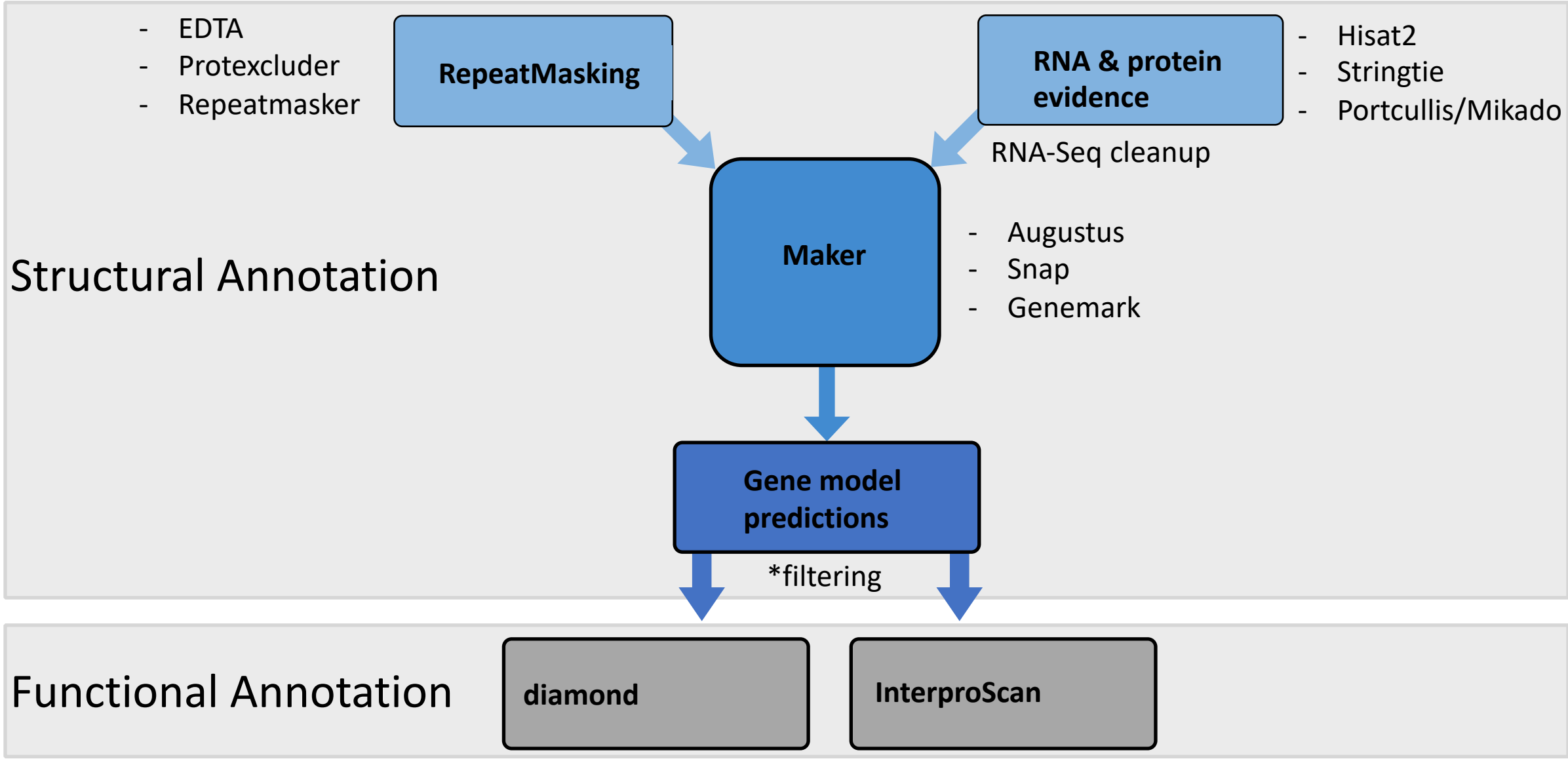
# Pre-annotation QC

- Assembly quality (total length, N50, etc)
- Errors - correction
- BUSCO metrics of genome

# Tools for structural annotation

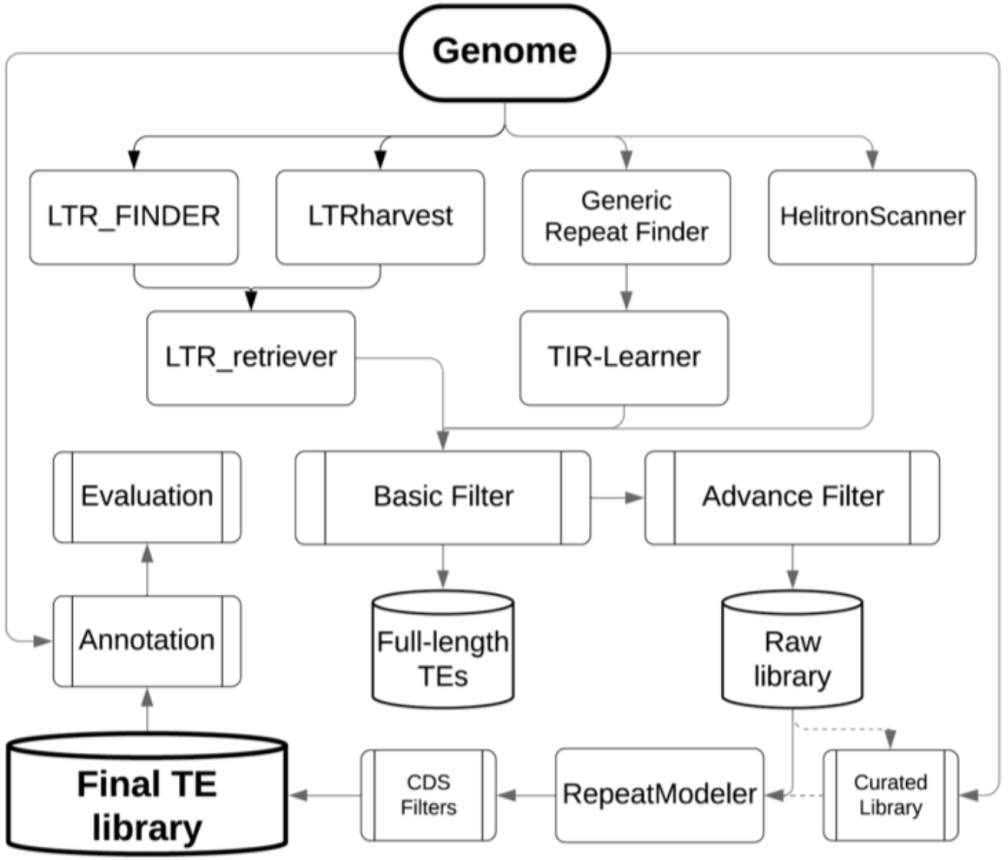
- EDTA https://github.com/oushujun/EDTA
- Repeatmasker http://www.repeatmasker.org/
- Braker https://github.com/Gaius-Augustus/BRAKER
- Augustus https://github.com/Gaius-Augustus/Augustus
- Snap https://github.com/KorfLab/SNAP
- Genemark http://exon.gatech.edu/GeneMark/
- Maker https://www.yandell-lab.org/software/maker.html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

# Annotation pipeline



# EDTA

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



# Tools for functional annotation

- BLAST
- Diamond
- InterProScan
- Mercator

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

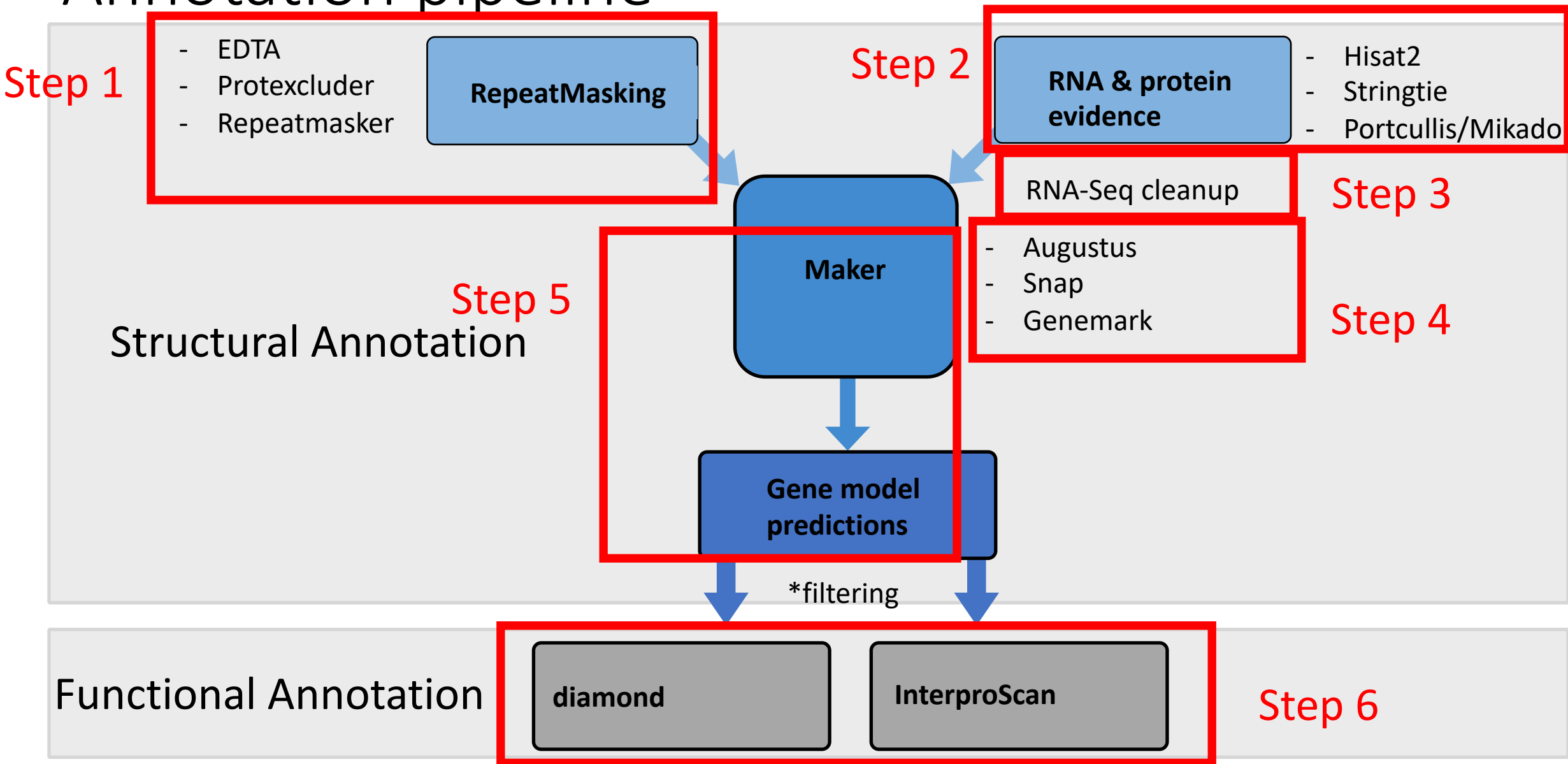
- Genome file: Ugibba_FLYE_assembly.fasta.PolcaCorrected.fa.cat.all.gz
- RNA-seq from shoots and traps: https://www.ncbi.nlm.nih.gov/sra/SRX2368915[accn]
- Proteins: uniprot_sprot_plants.fasta

- All this stuff plus some output files in /iplant/home/shared/Botany2020NMGWorkshop/annotation/

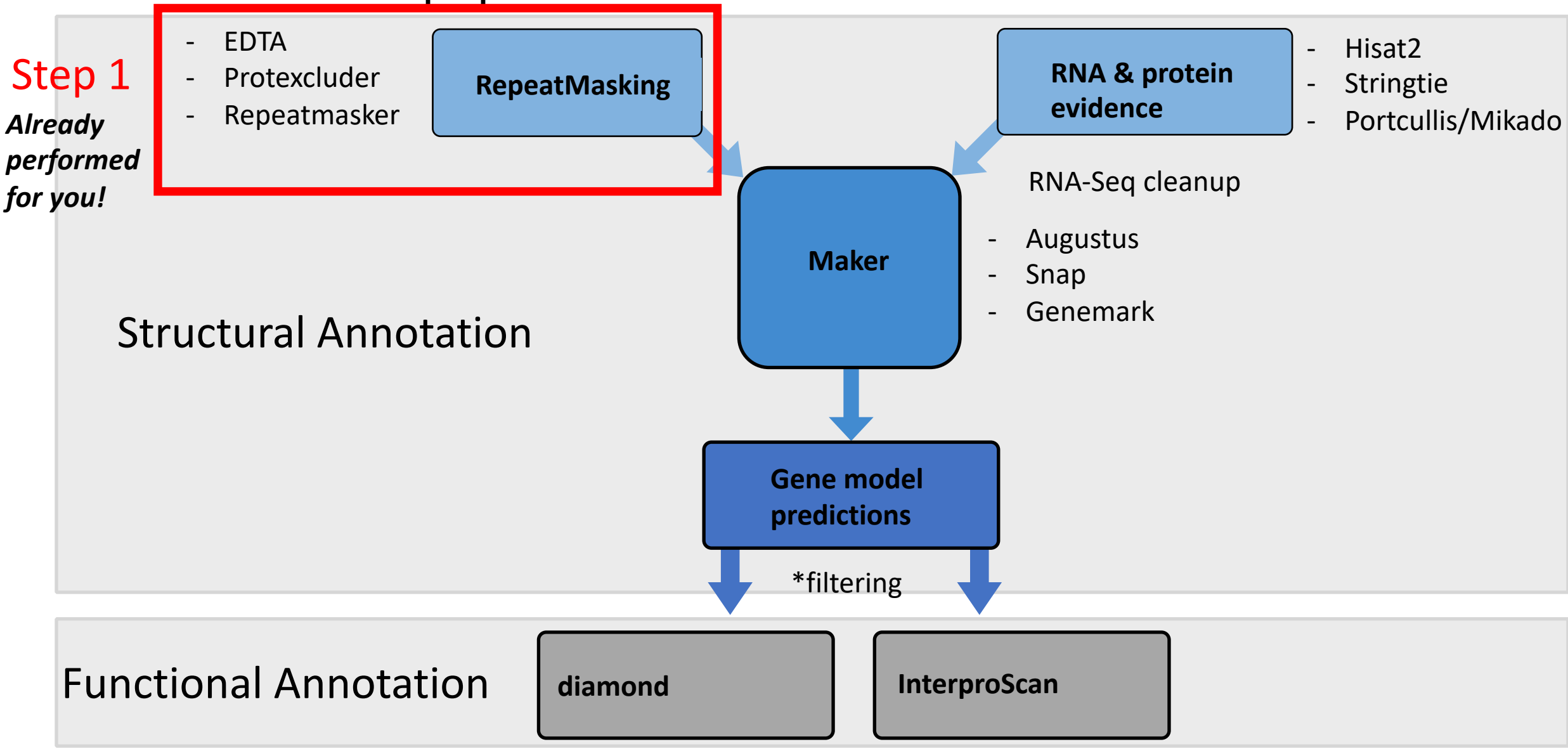
# QC of FLYE *U. gibba* assembly

- Size =
- N50 =
- BUSCO = 93.6% complete

# Annotation pipeline



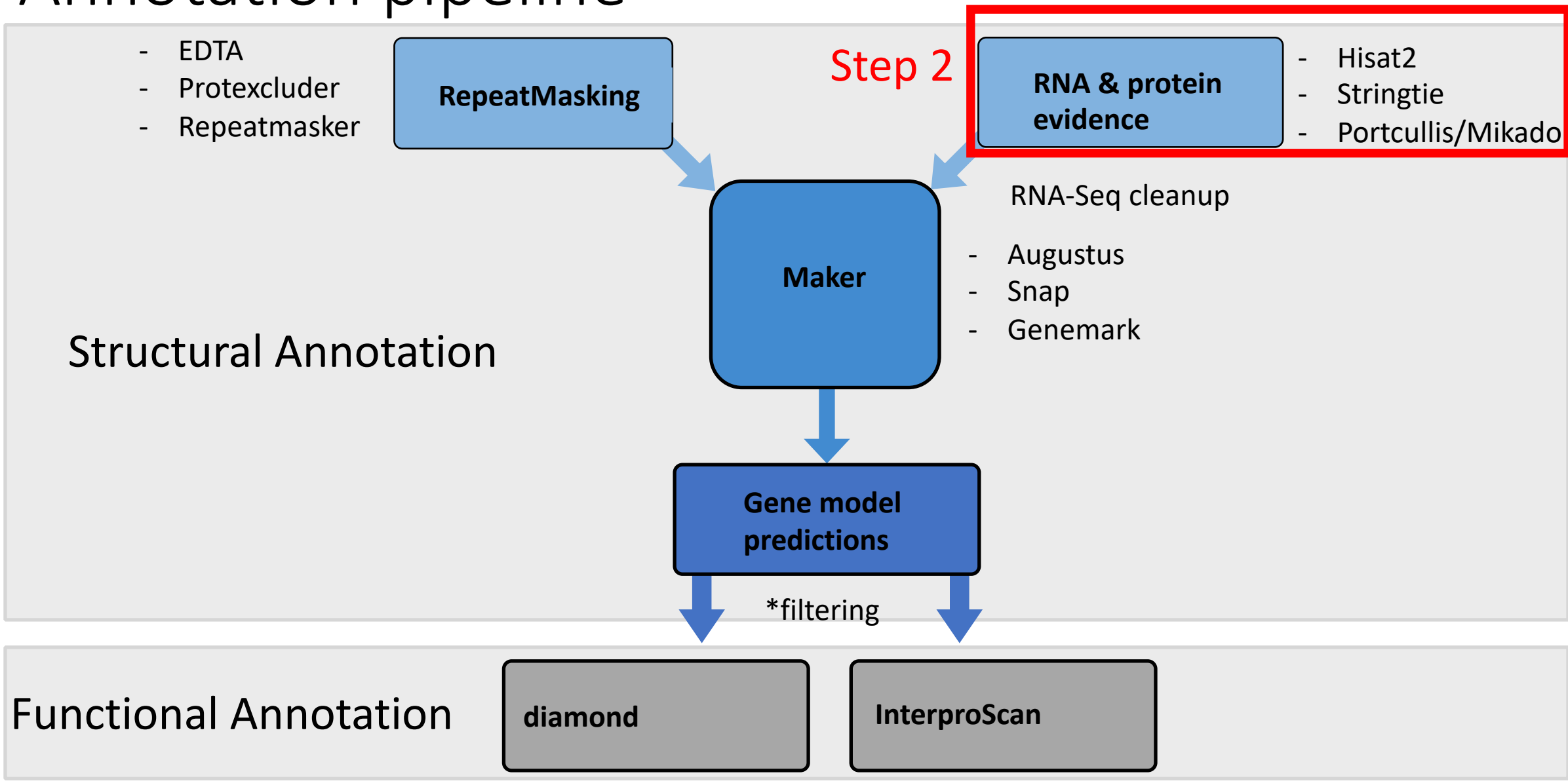
# Annotation pipeline



# Step 1: Repeat Masking

- https://github.com/bcbc-group/Botany2020NMGWorkshop/blob/master/annotation/1_repeatmasking.sh

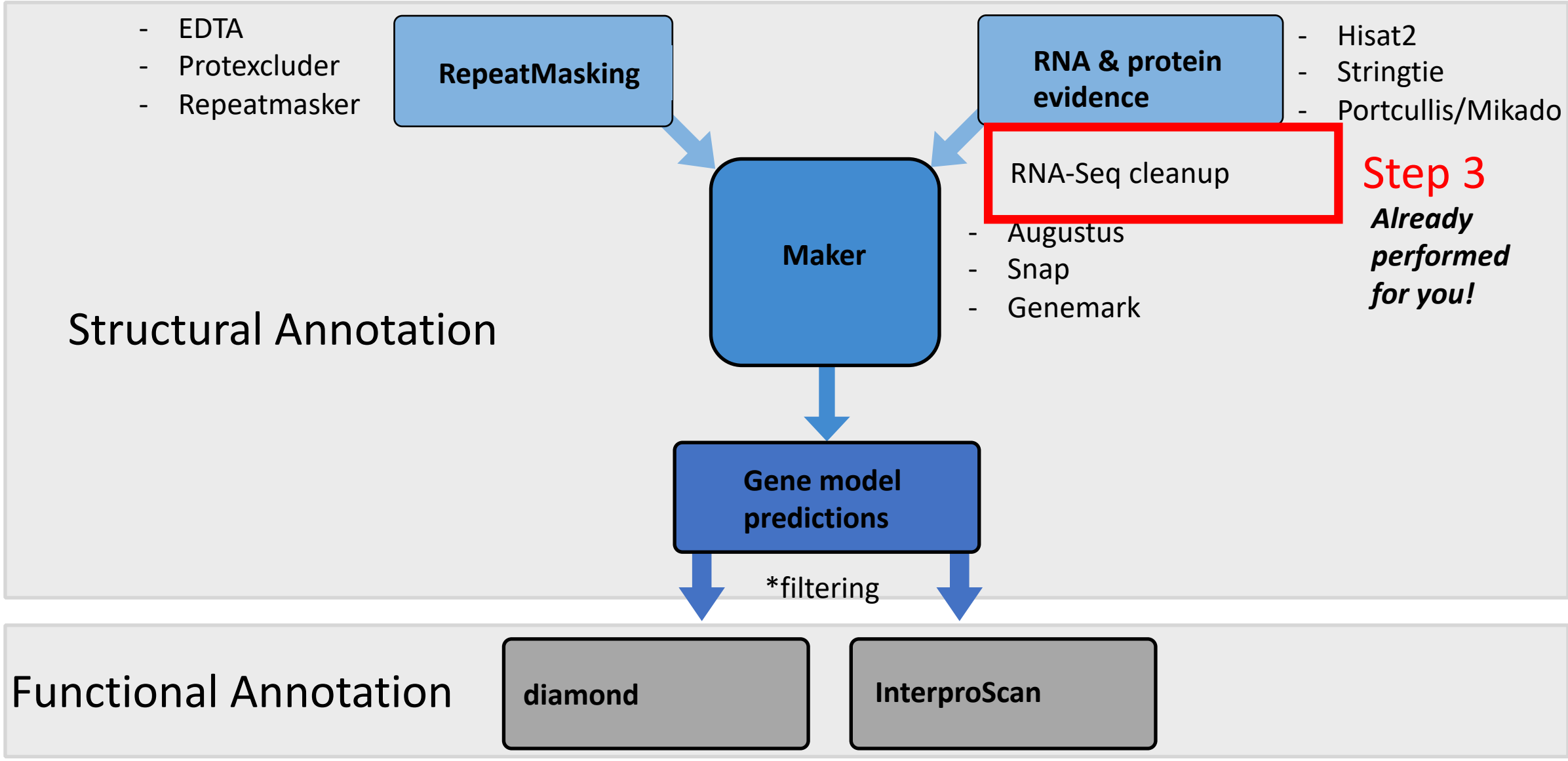
# Annotation pipeline



# Step 2: RNA-Seq read mapping

- https://github.com/bcbc-group/Botany2020NMGWorkshop/blob/master/annotation/2_hisat_pe_annot.sh

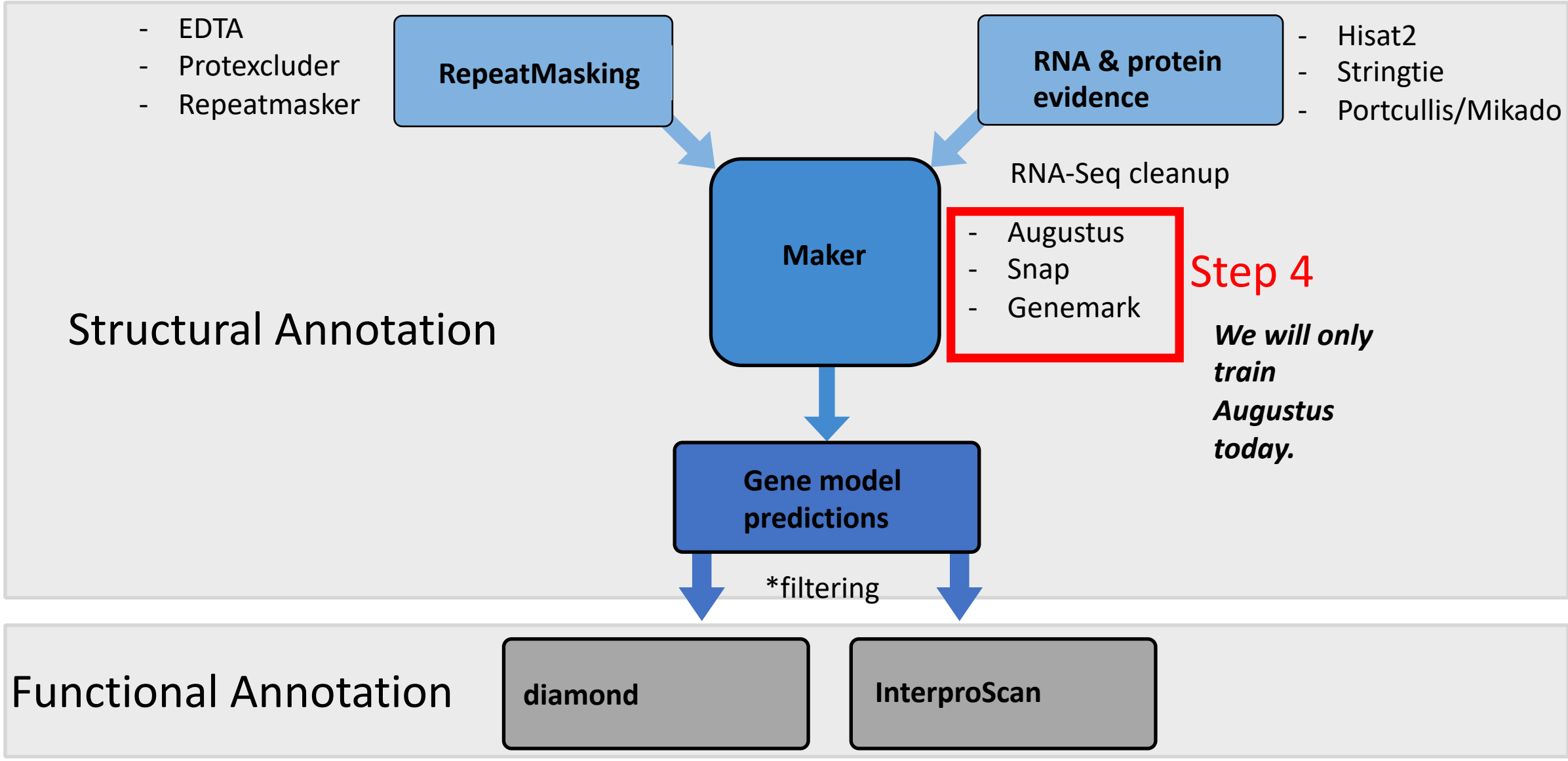
# Annotation pipeline



# Step 3: RNA-seq cleanup

- https://github.com/bcbc-group/Botany2020NMGWorkshop/blob/master/annotation/3_rnaseq_cleanup.sh

# Annotation pipeline



# Step 4: 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.gb 200

grep -c LOCUS genes.gb*

sudo chown srs57 /opt/augustus/config/species/

/opt/augustus-3.2.2/scripts/new_species.pl --species=Ugibba

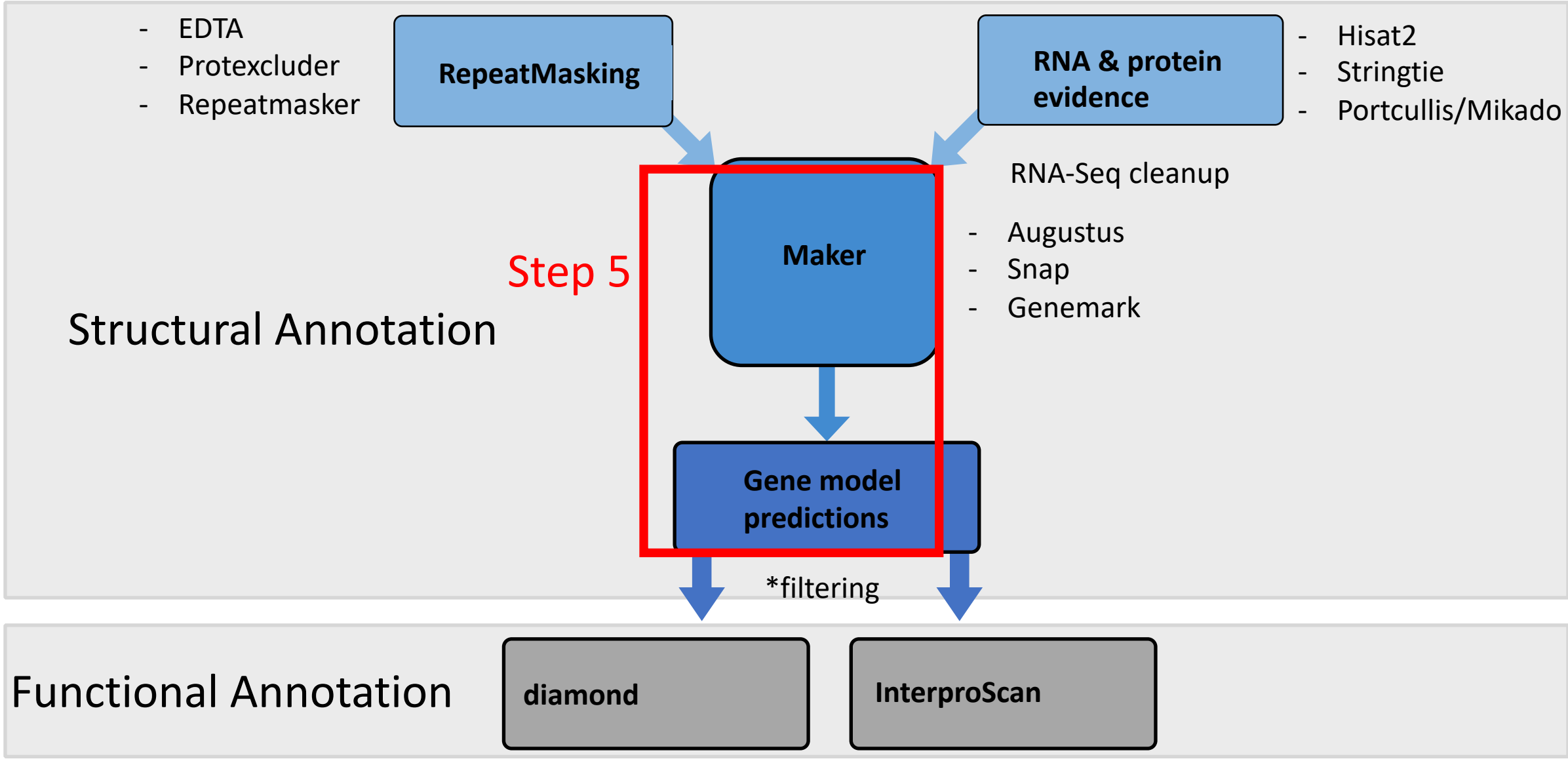
etraining --species=Ugibba genes.gb.train

ls -ort $AUGUSTUS_CONFIG_PATH/species/Ugibba

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

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

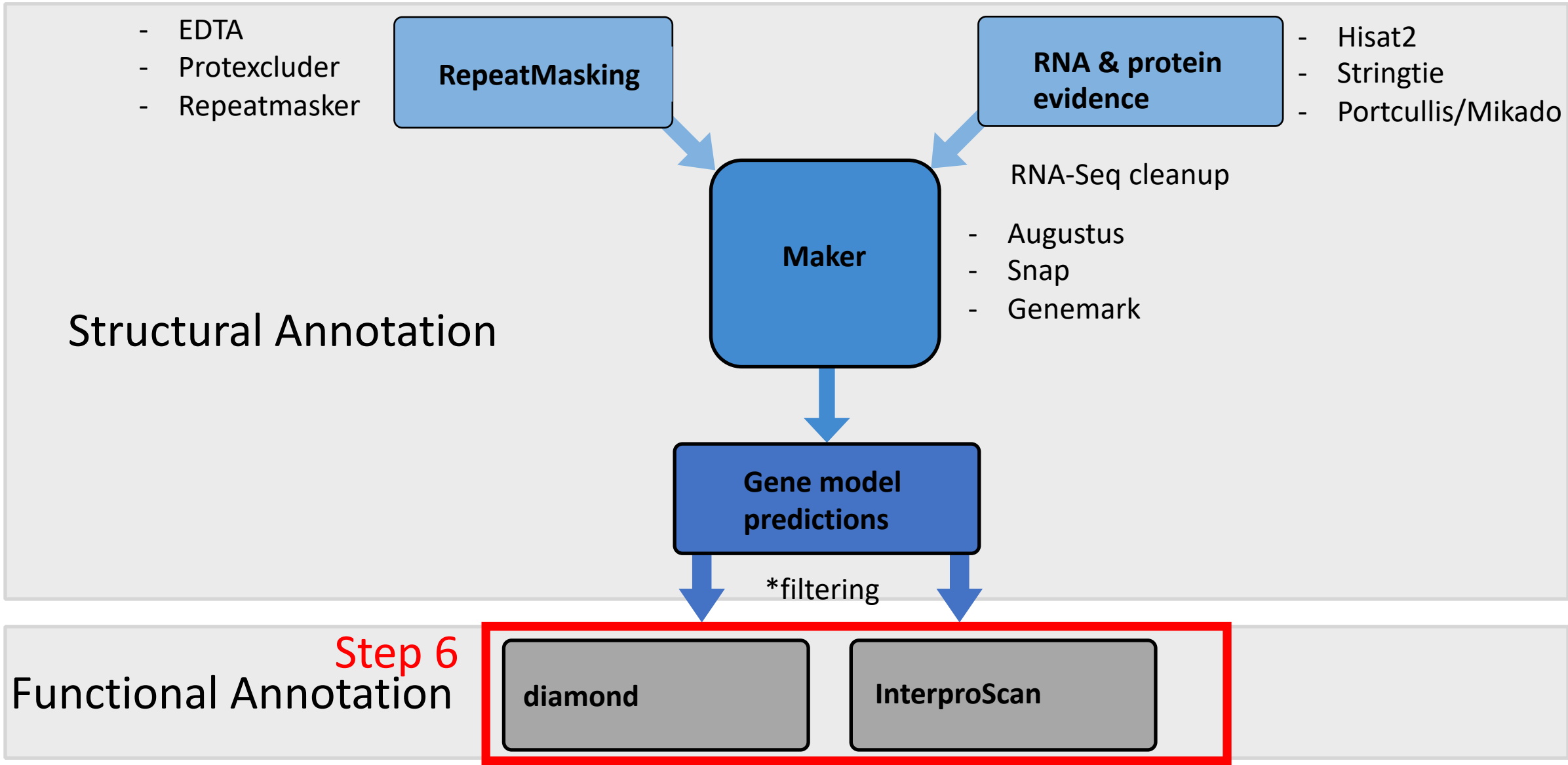
# Annotation pipeline



# Step 5: Running maker

- https://github.com/bcbc-group/Botany2020NMGWorkshop/blob/master/annotation/5_maker.sh

# Annotation pipeline



# Postprocessing, Cleanup, and QC

- Remove Transposons
- complete genes only
- match to nr, e-20
- FPKM > 0.1
- AED > 1
- InterProScan domain
- Comparison to relative, length and number of genes
- Gene families
- BUSCO
- Change gene model names once structural annotation is completed.
- Versioning –very important
- Apollo

# Step 6: Functional annotation

- https://github.com/bcbc-group/Botany2020NMGWorkshop/blob/master/annotation/6_function_annot.sh

- Maker also has several scripts for postprocessing files under: /opt/maker/bin