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 QC genome assemblies
- QC annotation results

Changing docker file storage location
You likely did this with Adrian

```
#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

$ sudo service docker start
```

Download InterProScan

```
#Go to VM
$ cd /scratch
$ wget ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/5/5.51-85.0/interproscan-5.51-85.0-64-bit.tar.gz

$ tar -pxvzf interproscan-5.51-85.0-*-bit.tar.gz

$ python3 initial_setup.py
```

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

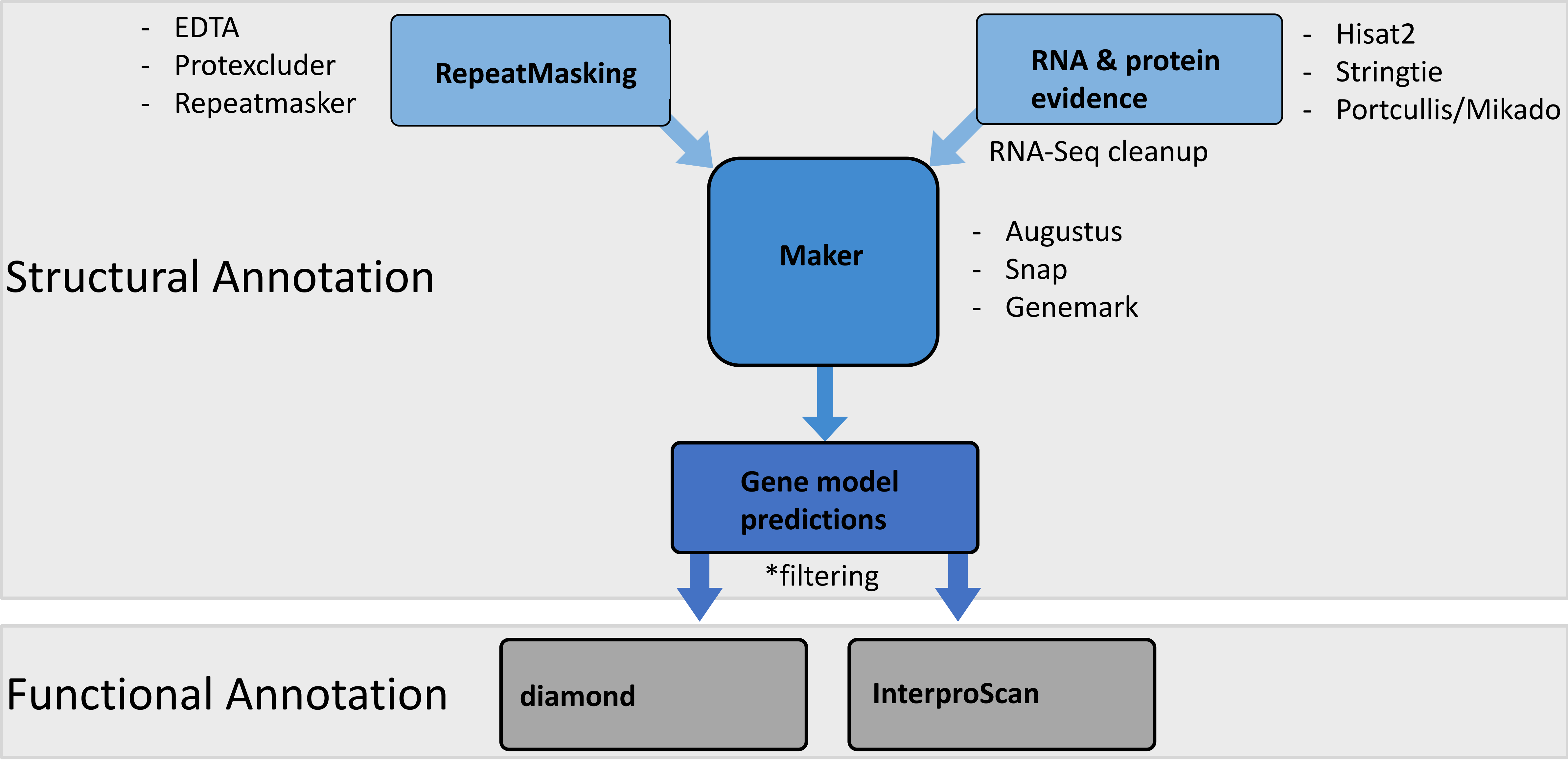
Pre-annotation QC

- 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
- Protexcluder
- Repeatmasker

RepeatMasking

RNA & protein evidence

- Hisat2
- Stringtie
- Portcullis/Mikado

RNA-Seq cleanup

Structural Annotation

Maker

- Augustus
- Snap
- Genemark

Gene model predictions

*filtering

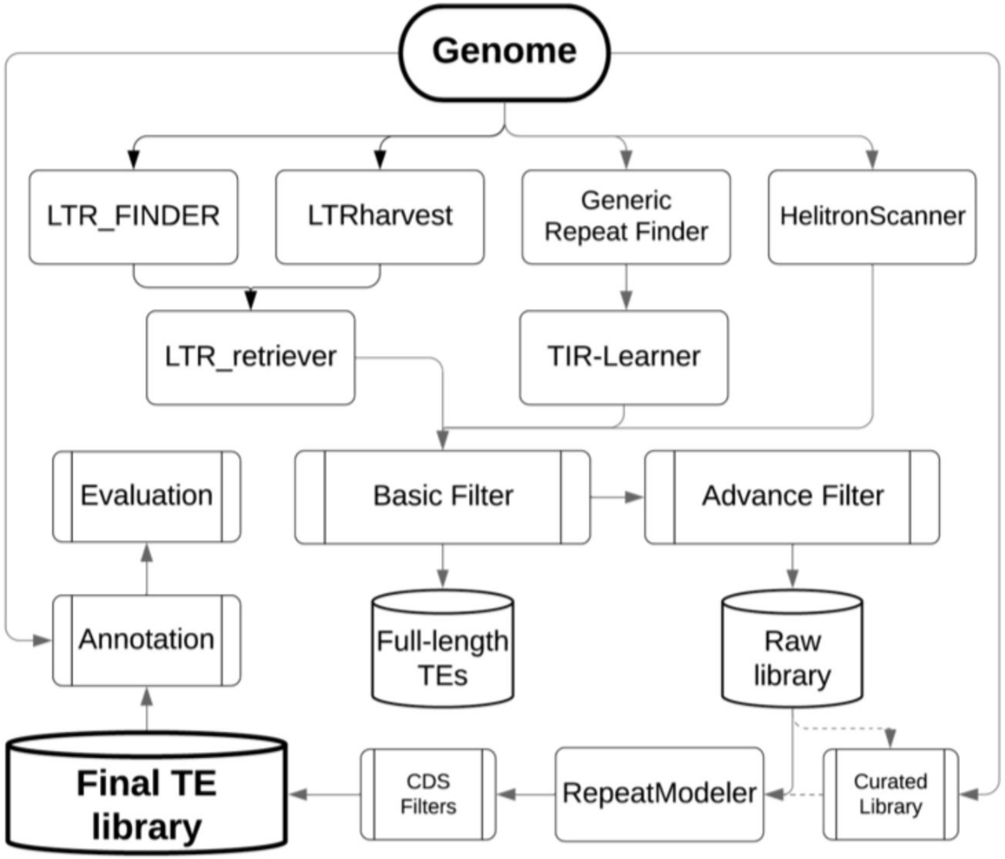
Functional Annotation

diamond

InterproScan

EDTA

The Extensive *de novo* TE Annotator (EDTA)



Tools for functional annotation

- BLAST
- Diamond
- InterProScan
- Mercator
- Databases: Swiss-prot, Trembl, nr, InterPro

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]](https://www.ncbi.nlm.nih.gov/sra/SRX2368915[accn])
- Proteins: uniprot_sprot_plants.fasta

- All this stuff plus some output files in /scratch/Botany2020NMGWorkshop/

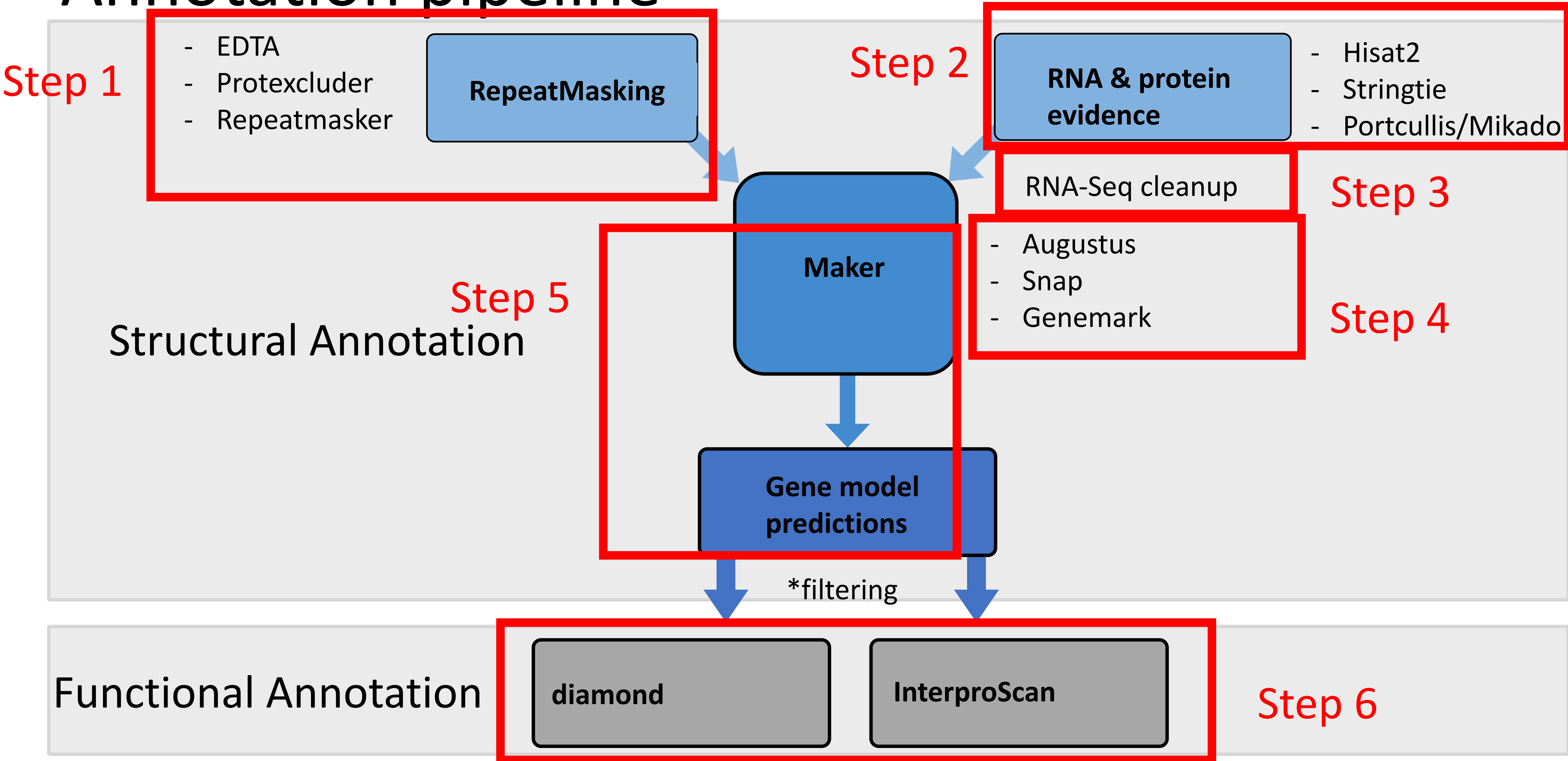
All scripts are on GitHub

```
$ cd /scratch
$ git clone https://github.com/bcbc-group/NMGWorkshop2021.git
$ cd /scratch/NMGWorkshop2021/5.Annotation/scripts
```

QC of FLYE *U. gibba* assembly

- Size = 85,700,758 bp
- N50 = 4,134,757 bp
- BUSCO = 93.6% complete

Annotation pipeline



Step 1

- EDTA
- Protexcluder
- Repeatmasker

RepeatMasking

Step 2

RNA & protein evidence

- Hisat2
- Stringtie
- Portcullis/Mikado

RNA-Seq cleanup

Step 3

Maker

- Augustus
- Snap
- Genemark

Step 4

Step 5

Structural Annotation

Gene model predictions

*filtering

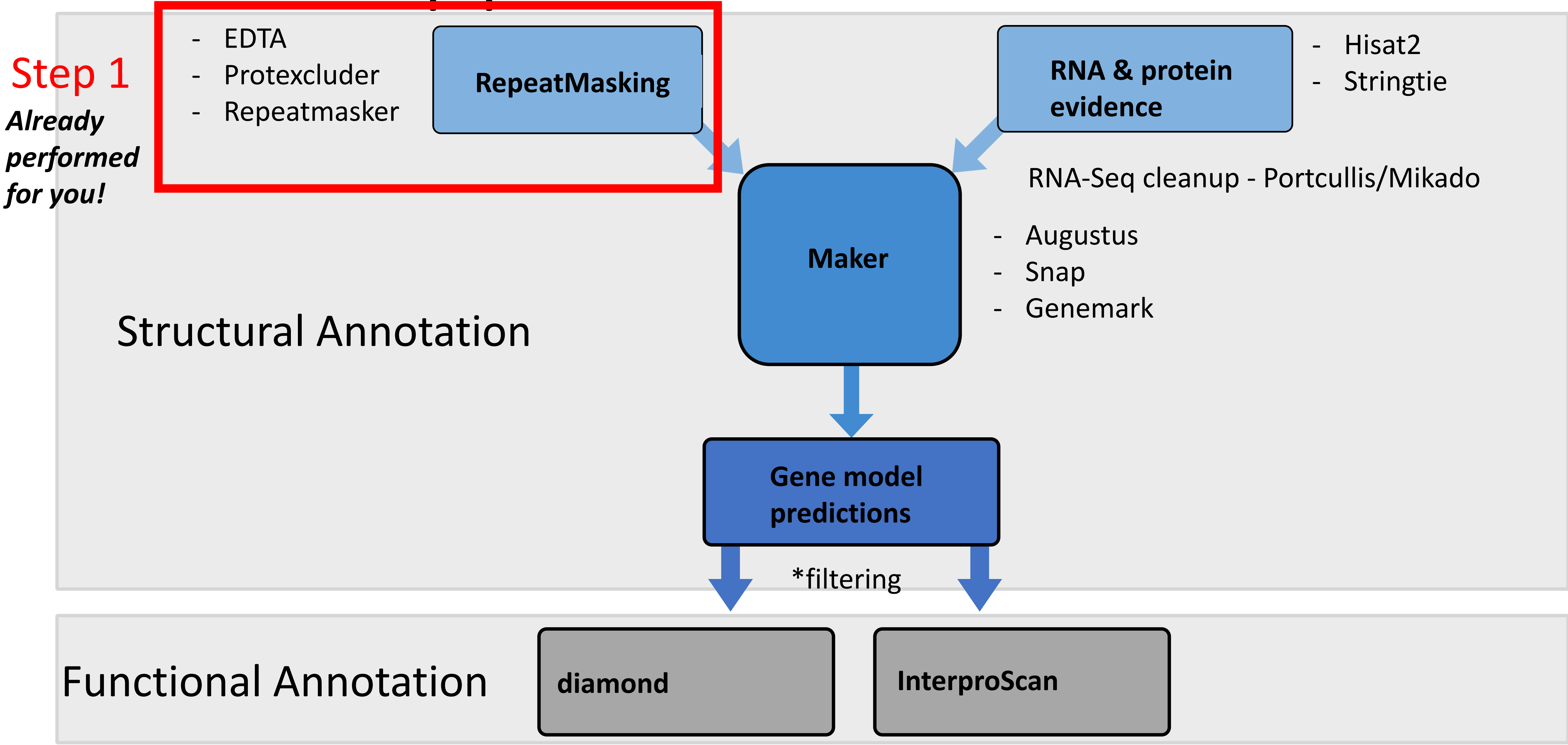
Functional Annotation

diamond

InterproScan

Step 6

Annotation pipeline



Step 1

Already performed for you!

- EDTA
- Protexcluder
- Repeatmasker

RepeatMasking

RNA & protein evidence

- Hisat2
- Stringtie

RNA-Seq cleanup - Portcullis/Mikado

Maker

- Augustus
- Snap
- Genemark

Structural Annotation

Gene model predictions

*filtering

Functional Annotation

diamond

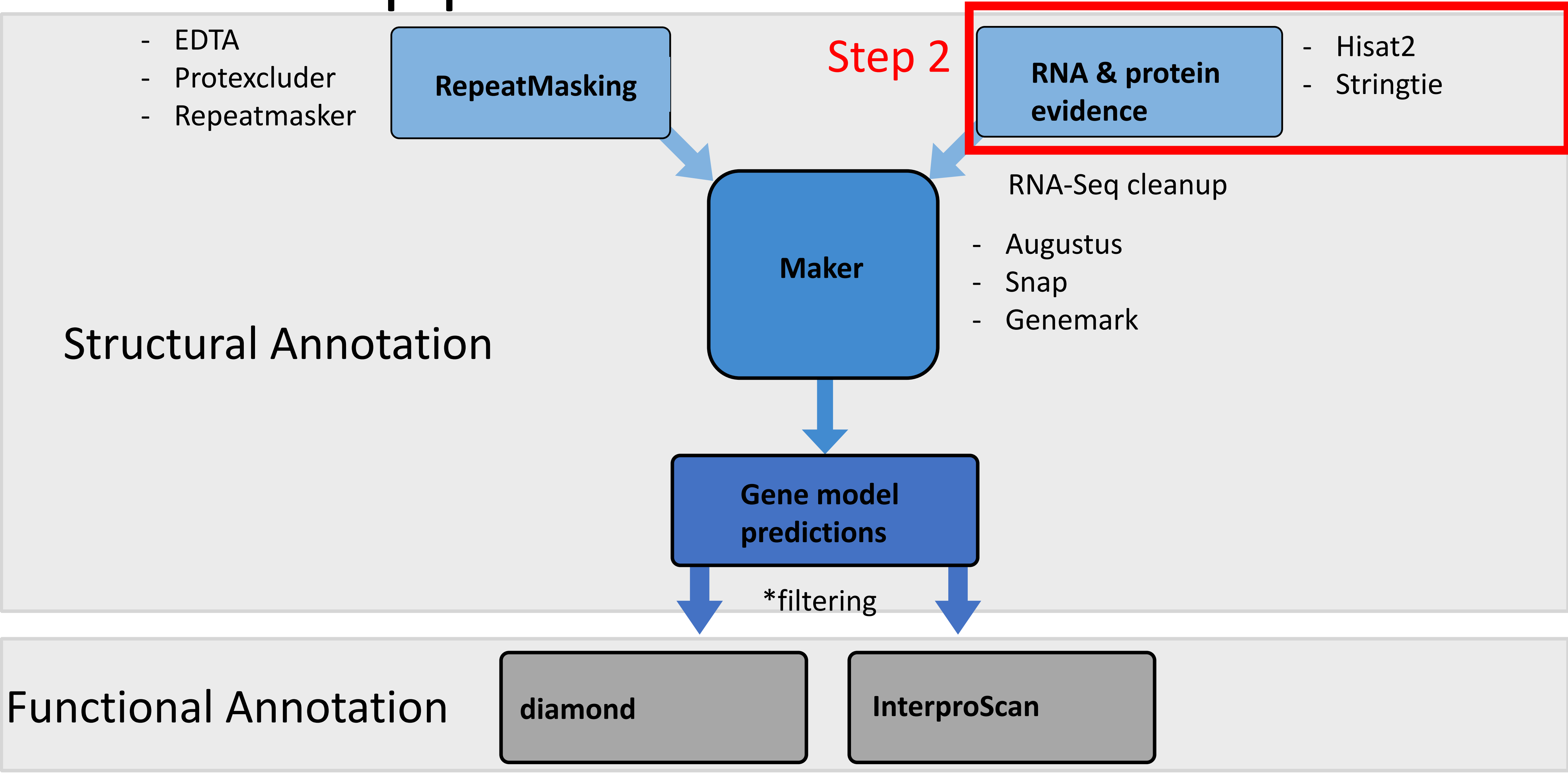
InterproScan

Step 1: Repeat Masking

https://github.com/bcbc-group/NMGWorkshop2021/blob/main/5.Annotation/scripts/1_repeatmasking.sh

*this has already been performed to conserve time

Annotation pipeline



- EDTA
- Protexcluder
- Repeatmasker

RepeatMasking

Step 2

RNA & protein evidence

- Hisat2
- Stringtie

RNA-Seq cleanup

Maker

- Augustus
- Snap
- Genemark

Structural Annotation

Gene model predictions

*filtering

Functional Annotation

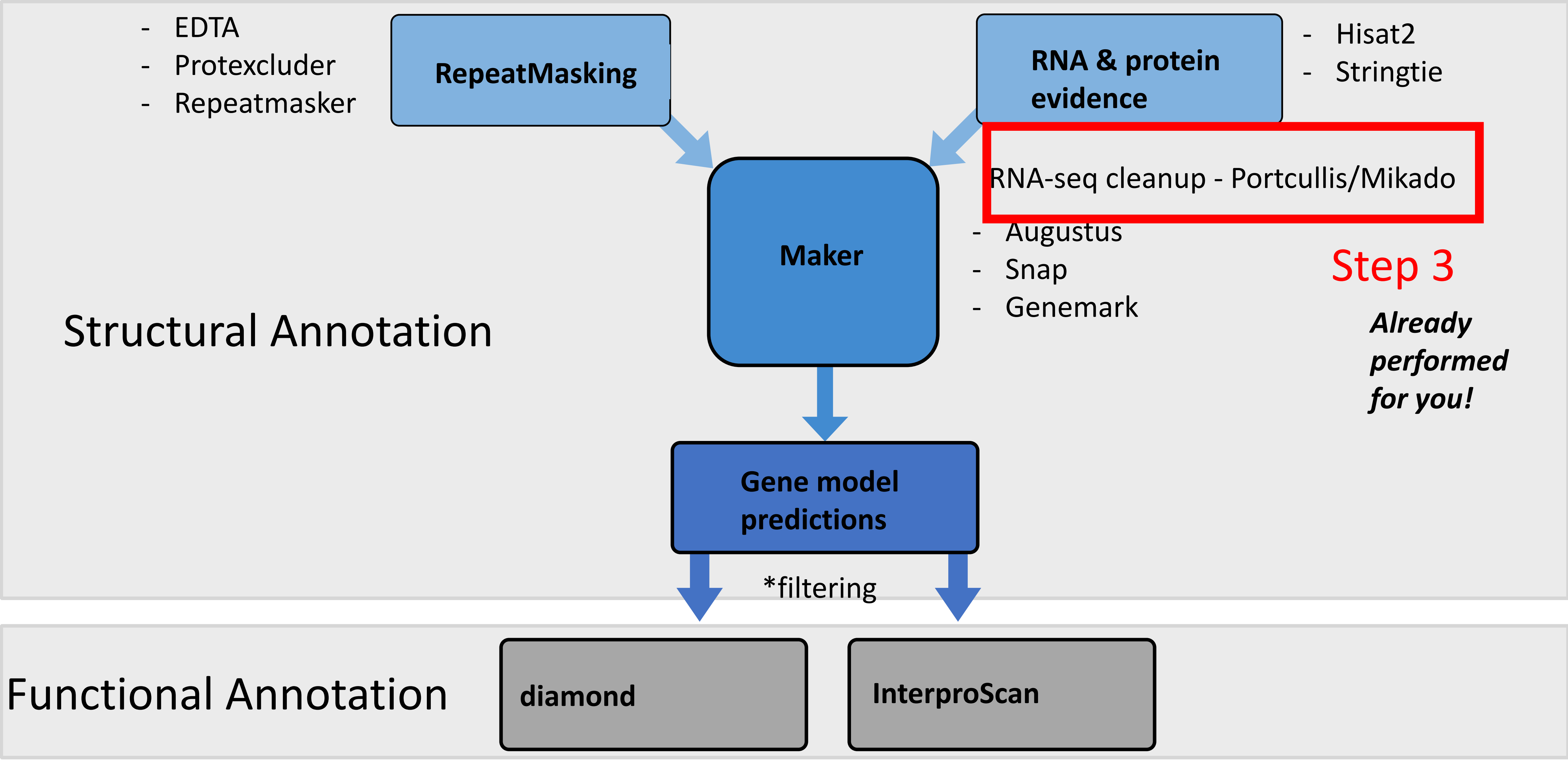
diamond

InterproScan

Step 2: RNA-Seq read mapping

https://github.com/bcbc-group/NMGWorkshop2021/blob/main/5.Annotation/scripts/2_hisat_pe_annot.sh

Annotation pipeline



- EDTA
- Protexcluder
- Repeatmasker

RepeatMasking

RNA & protein evidence

- Hisat2
- Stringtie

RNA-seq cleanup - Portcullis/Mikado

Maker

- Augustus
- Snap
- Genemark

Step 3

Already performed for you!

Structural Annotation

Gene model predictions

*filtering

Functional Annotation

diamond

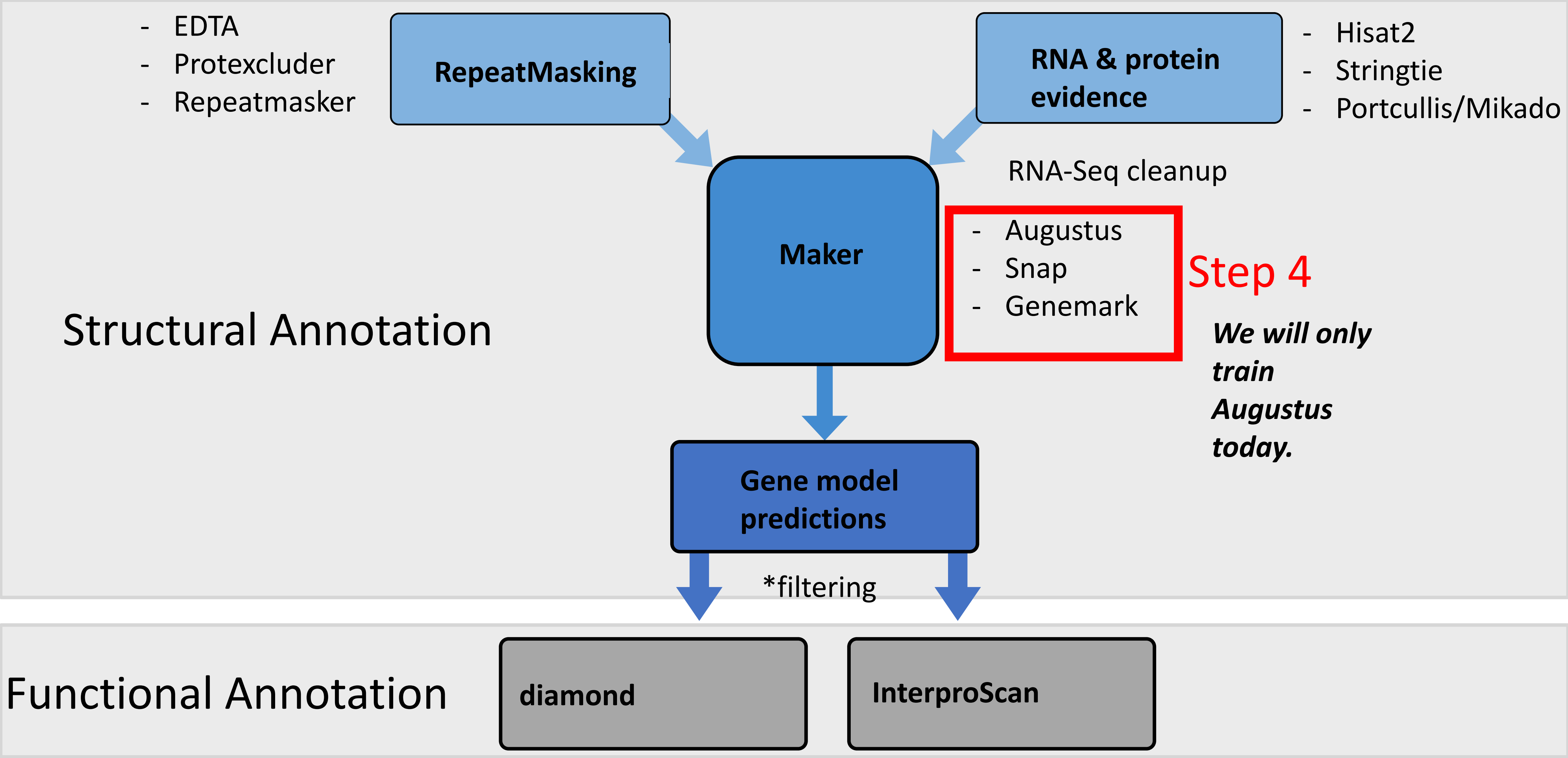
InterproScan

Step 3: RNA-seq cleanup

https://github.com/bcbc-group/NMGWorkshop2021/blob/main/5.Annotation/scripts/3_rnaseq_cleanup.sh

*this has already been performed to conserve time

Annotation pipeline



We will only train Augustus today.

Step 4: Training augustus and snap

- https://vcru.wisc.edu/simonlab/bioinformatics/programs/augustus/docs/tutorial2015/training.html

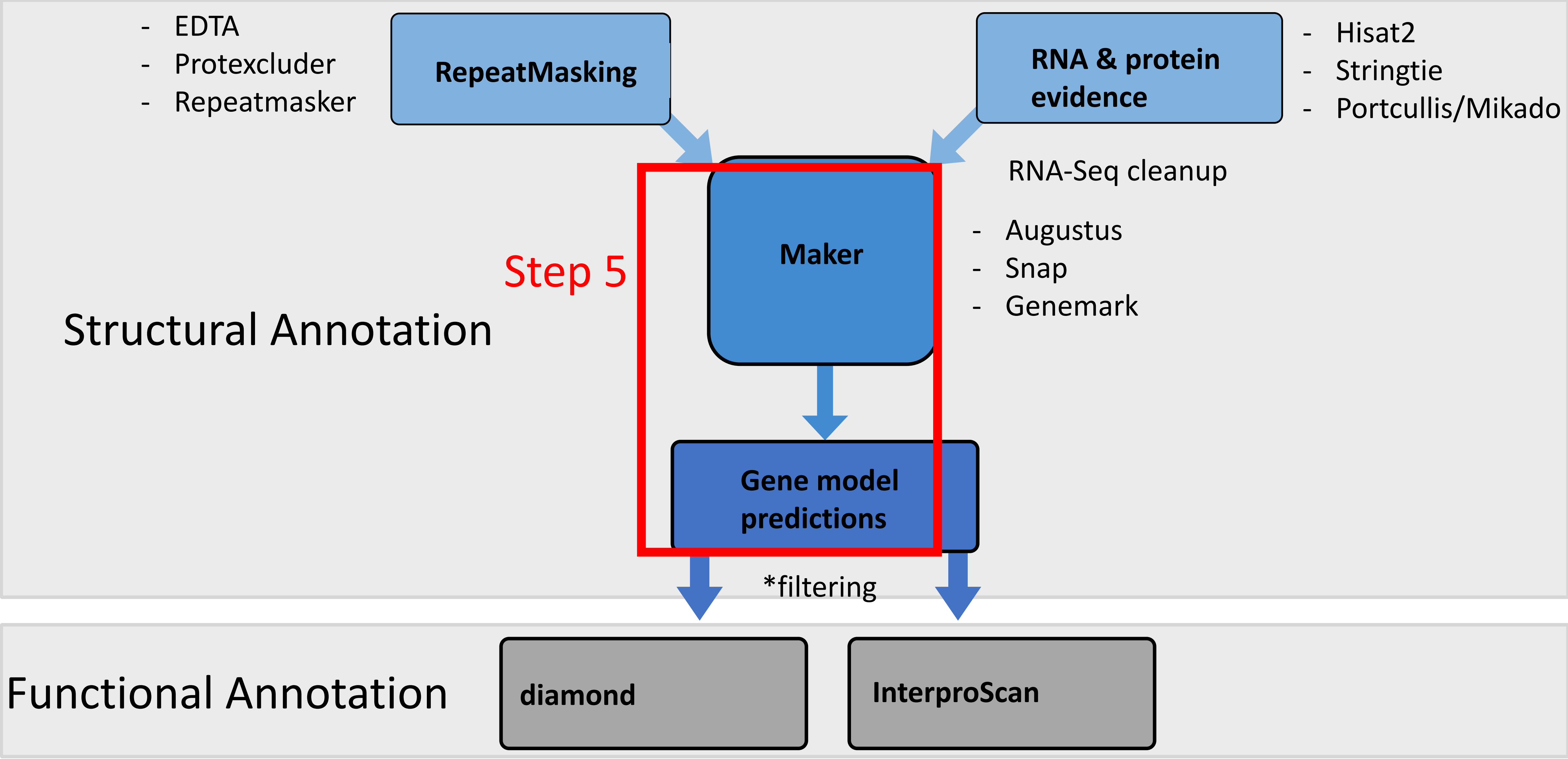
- https://github.com/bcbc-group/NMGWorkshop2021/blob/main/5.Annotation/scripts/4_training.sh

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



- EDTA
- Protexcluder
- Repeatmasker

RepeatMasking

RNA & protein evidence

- Hisat2
- Stringtie
- Portcullis/Mikado

RNA-Seq cleanup

Step 5

Maker

- Augustus
- Snap
- Genemark

Structural Annotation

Gene model predictions

*filtering

Functional Annotation

diamond

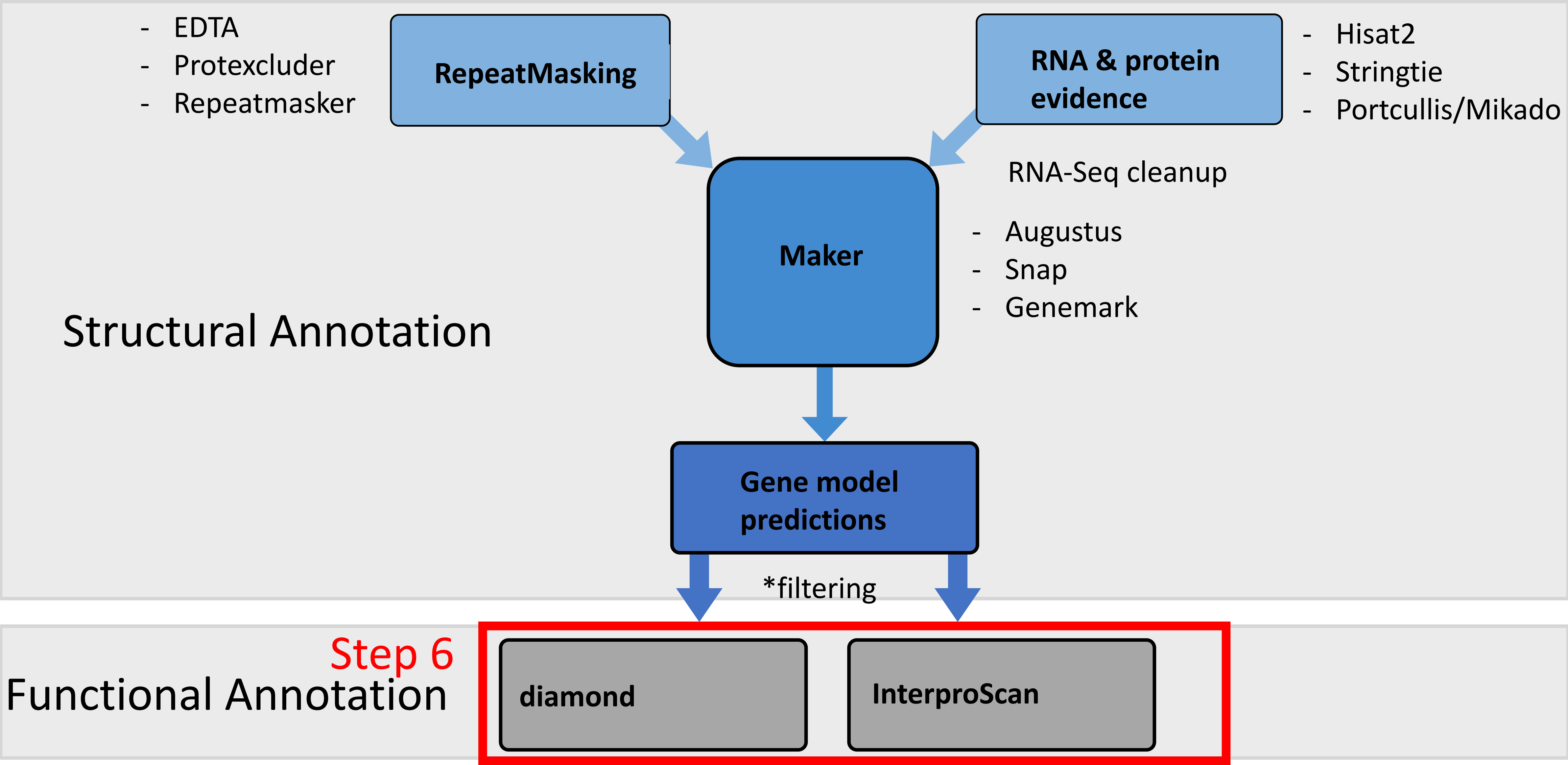
InterproScan

Step 5: Running maker

https://github.com/bcbc-group/NMGWorkshop2021/blob/main/5.Annotation/scripts/5_maker.sh

*this has already been performed to conserve time

Annotation pipeline



Postprocessing, Cleanup, and QC

- Remove Transposons
- complete genes only
- match to nr, e-20
- FPKM > 0.1
- AED value
- 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