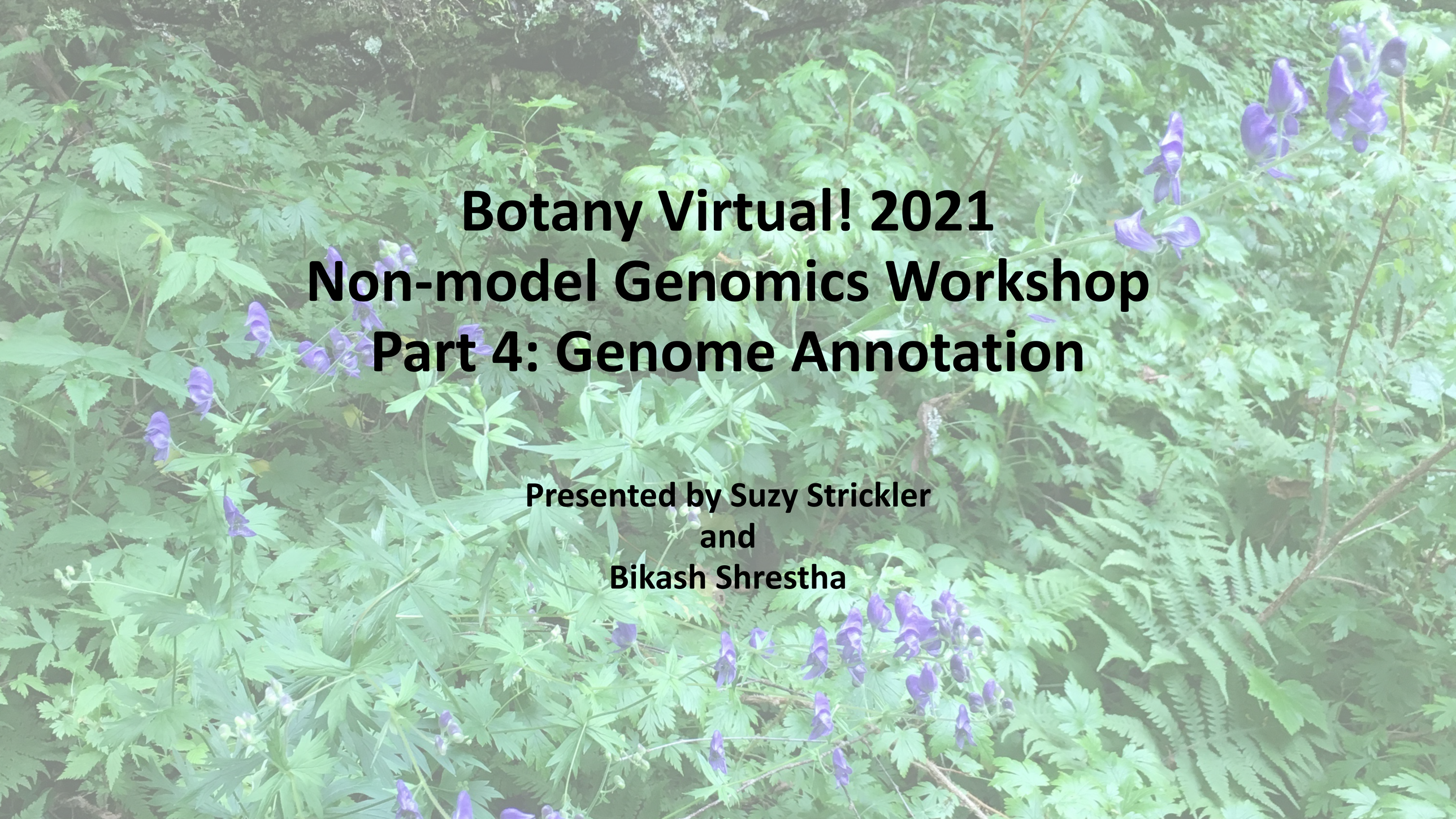


Botany Virtual! 2021
Non-model Genomics Workshop
Part 4: Genome Annotation

Presented by Suzy Strickler
and
Bikash Shrestha

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

Objectives

- Understand the steps involved in genome annotation
- Demonstrate the types of data and tools that can be used in genome annotation
- Learn how to QC genome assemblies and annotation results
- Understand how to derive functional predictions for genes

Annotation Overview

1. Assembly QC - is it good enough to annotate?
2. Structural annotation - tools, inputs, outputs
3. Annotation QC - are we capturing most of the gene models accurately?
4. Functional annotation - tools, inputs, outputs

1. Assembly QC

- Assembly quality (total length, N50, etc)

	S. lycopersicoides	*S. pennellii* v2	*S. lycopersicum* Heinz v 4.0
No. of pseudomolecules	12	12	12
longest sequence (Mbp)	133.5	109.3	90.9
Contig N50 (bp)	253,764	60,347	6,007,830
total length (Mbp)	1,152	926	782.5
expected genome size (Mbp)	1,200	942	781
Total size (bp) of unanchored contigs (% of assembly)	135,089,793 (10.5)	63,101,713 (6.4%)	9,643,250 (1.2%)

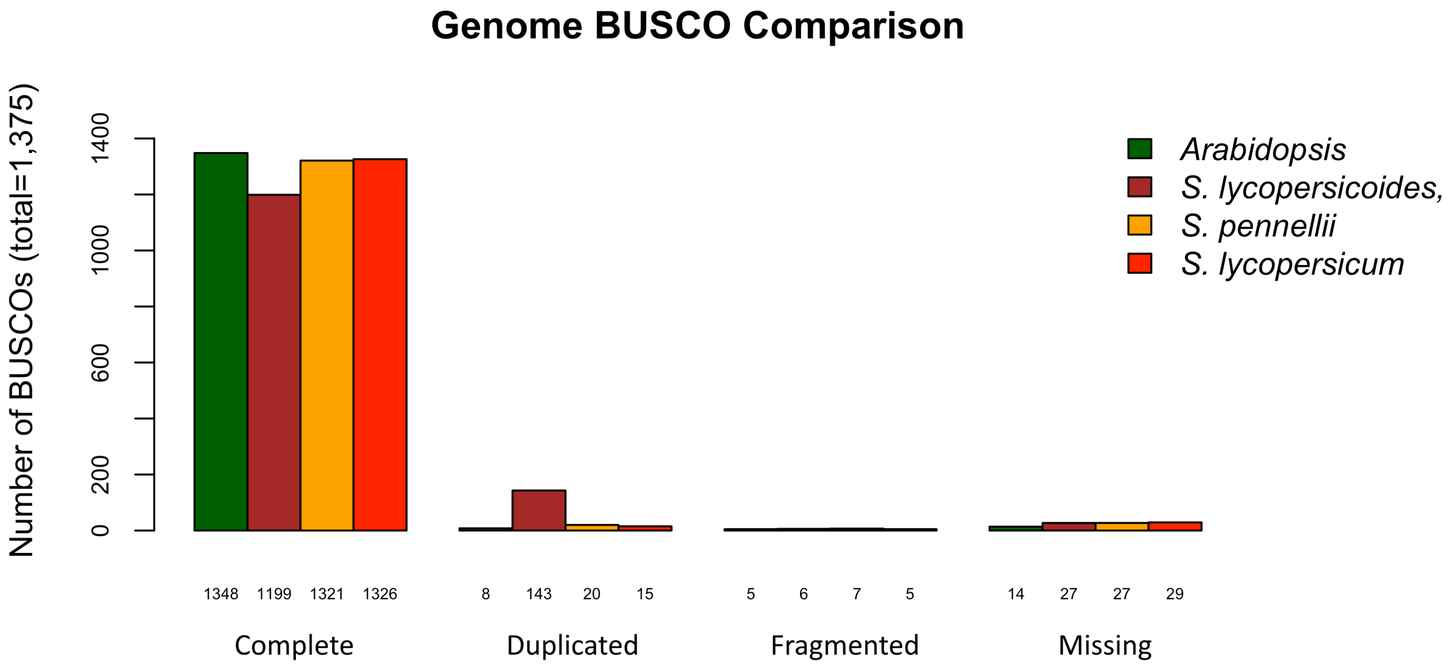
1. Assembly QC

- Assembly Errors - correction
  - Tools: Pilon - Illumina https://github.com/broadinstitute/pilon/wiki
    - Polca - part of MaSurCA package
    - Arrow - PacBio https://github.com/PacificBiosciences/GenomicConsensus
    - Nanopolish - nanopore https://github.com/jts/nanopolish
- Example: tomato Nanopore/Illumina hybrid assembly polished with Illumina reads:

Round	SNPs/Indels corrected
1	145,994
2	84,441
3	46,201

1. Assembly QC

- Assembly BUSCO metrics https://gitlab.com/ezlab/busco_biocontainer



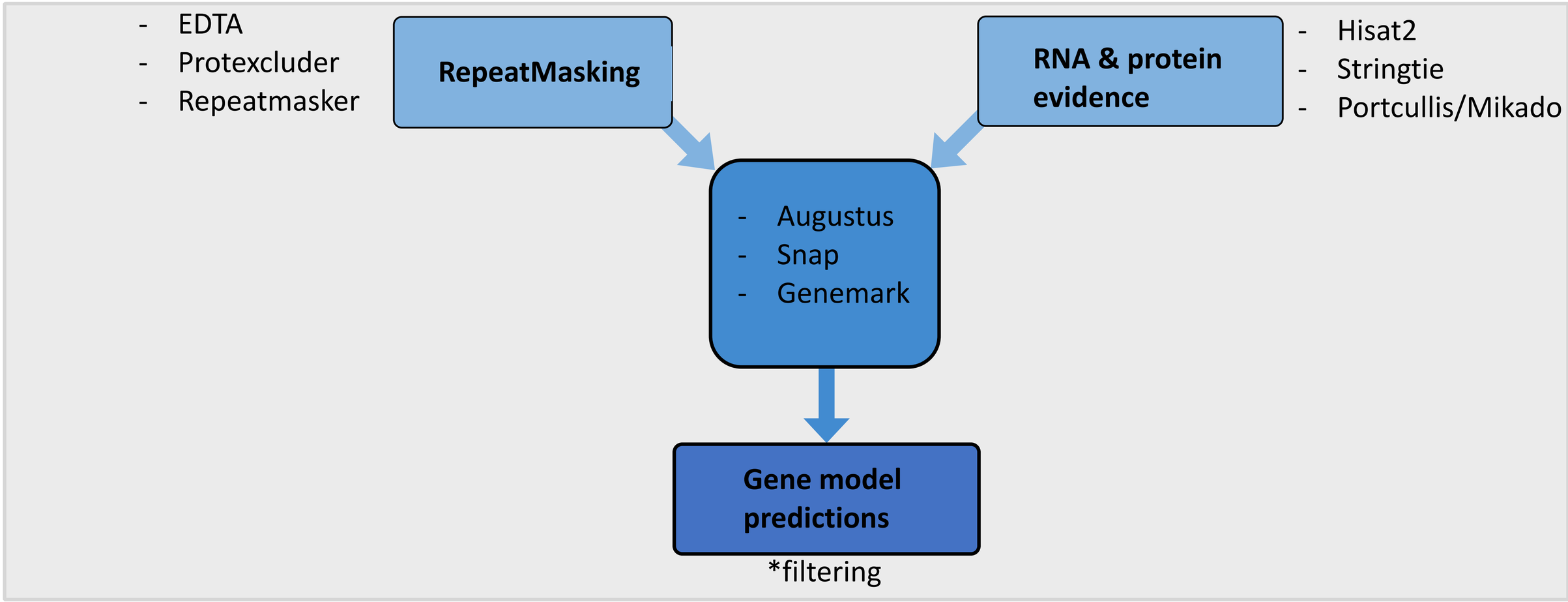
Software | Open Access | Published: 29 November 2018

Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies

Michael J. Roach, Simon A. Schmidt & Anthony R. Borneman

BMC Bioinformatics **19**, Article number: 460 (2018) | Cite this article

2. Structural Annotation



Annotation pipelines:
Maker
Braker

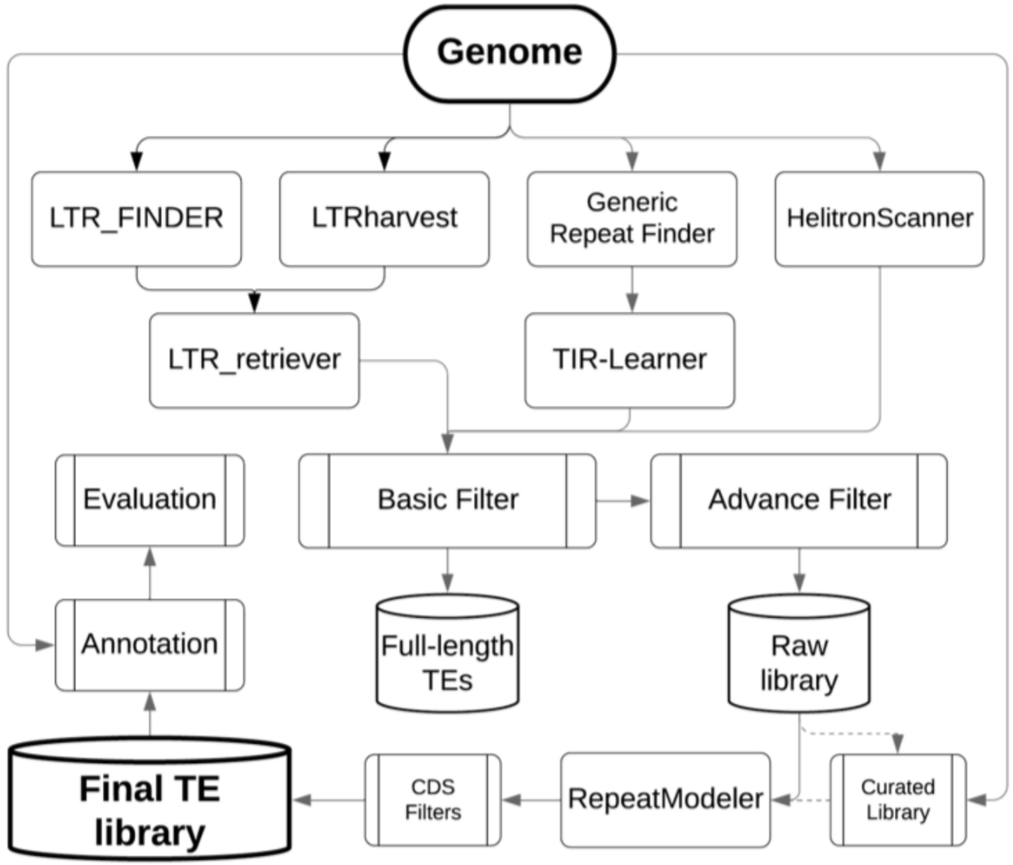
2. Structural Annotation: Repeat Masking

- Why repeat mask? Over-prediction
- Tools:
  - Repeat Modeler (Repeat Scout, RECON, LtrHarvest) - de novo https://www.repeatmasker.org/RepeatModeler/
  - EDTA - de novo and filtering https://github.com/oushujun/EDTA
  - Repbase - database https://www.girinst.org/repbase/
  - Repeatmasker - masking of genome using above output http://www.repeatmasker.org/

IMPORTANT - don’t mask domains! -> Protexcluder

2. Structural Annotation: Repeat Masking

The Extensive *de novo* TE Annotator (EDTA)



https://github.com/oushujun/EDTA

2. Structural Annotation: *Ab initio* Predictors

- Why *ab initio*? Similarity-based methods may not be applicable, propagation of errors, use statistical models to predict gene models
    - Training: "*ab initio* gene predictors use organism-specific genomic traits, such as codon frequencies and distributions of intron– exon lengths, to distinguish genes from intergenic regions and to determine intron–exon structures." -Yandell and Ence 2012
- Tools:
    - Snap - easy to train https://github.com/KorfLab/SNAP
    - Augustus - difficult to train https://github.com/Gaius-Augustus/Augustus
    - Genemark http://exon.gatech.edu/GeneMark/

2. Structural Annotation: Evidence aligners

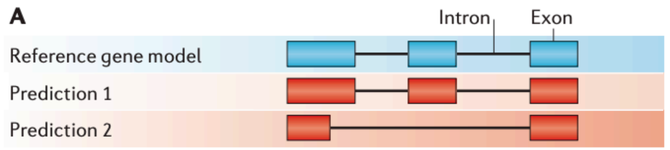
- Why/What? Tools to align RNA and protein evidence to genome, usually output to gff3 or bam
- Tools:
  - Hisat2 - align RNA-seq http://daehwankimlab.github.io/hisat2/
  - Gmap - align mRNA https://academic.oup.com/bioinformatics/article/21/9/1859/409207
  - Mikado/Portcullis - RNA-seq clean-up https://mikado.readthedocs.io/en/stable/
  - Pasa https://github.com/PASApipeline/PASApipeline/blob/master/docs/index.asciidoc

2. Structural Annotation: Pipelines

- Why/What? Uses a number of tools and inputs
- Tools:
  - Maker https://www.yandell-lab.org/software/maker.html
  - Braker https://github.com/Gaius-Augustus/BRAKER

3. Annotation QC: Postprocessing, Cleanup, and QC

- Remove
  - Transposons
  - incomplete gene models
  - Genes with no match to nr (<e-20) an FPKM <0.1 and no InterProScan domain
- Sensitivity, specificity, accuracy, AED value

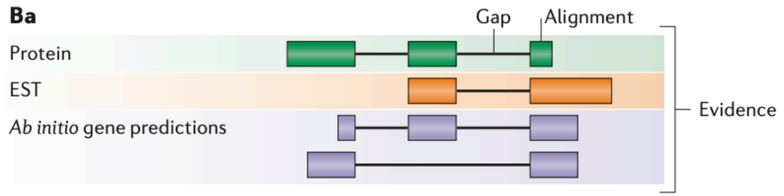


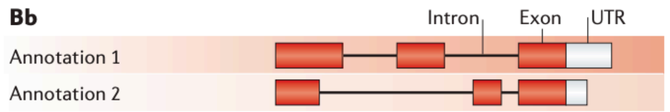
	SN	SP	AC
Prediction 1	1 (1)	1 (1)	1 (1)
Prediction 2	0.63 (0.33)	1 (0.5)	0.81 (0.42)

SN = TP / (TP + FN)

SP = TP / (TP + FP)

AC = (SN + SP) / 2





	AED
Annotation 1	0.2
Annotation 2	0.6

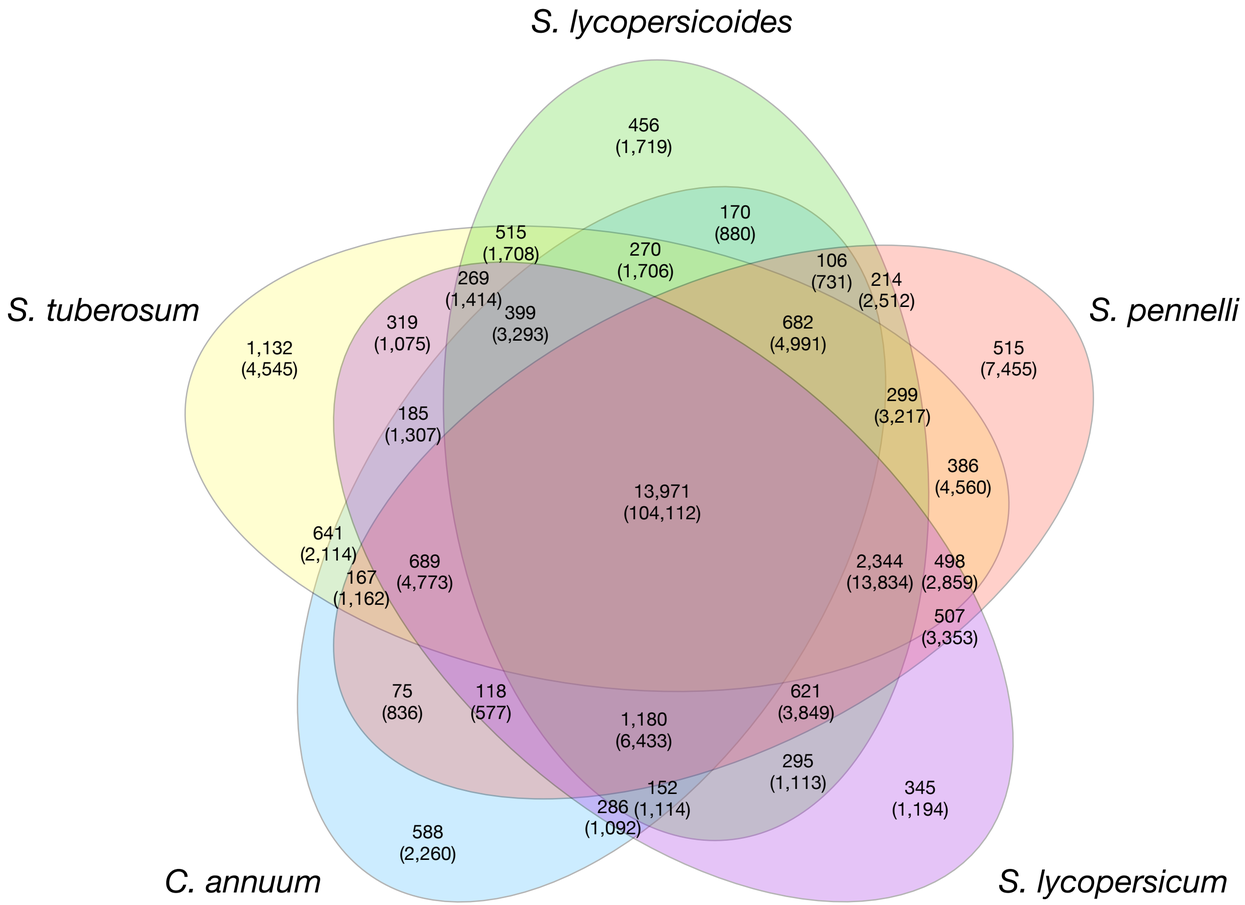
AED = 1 – AC

Yandell and Ence 2012

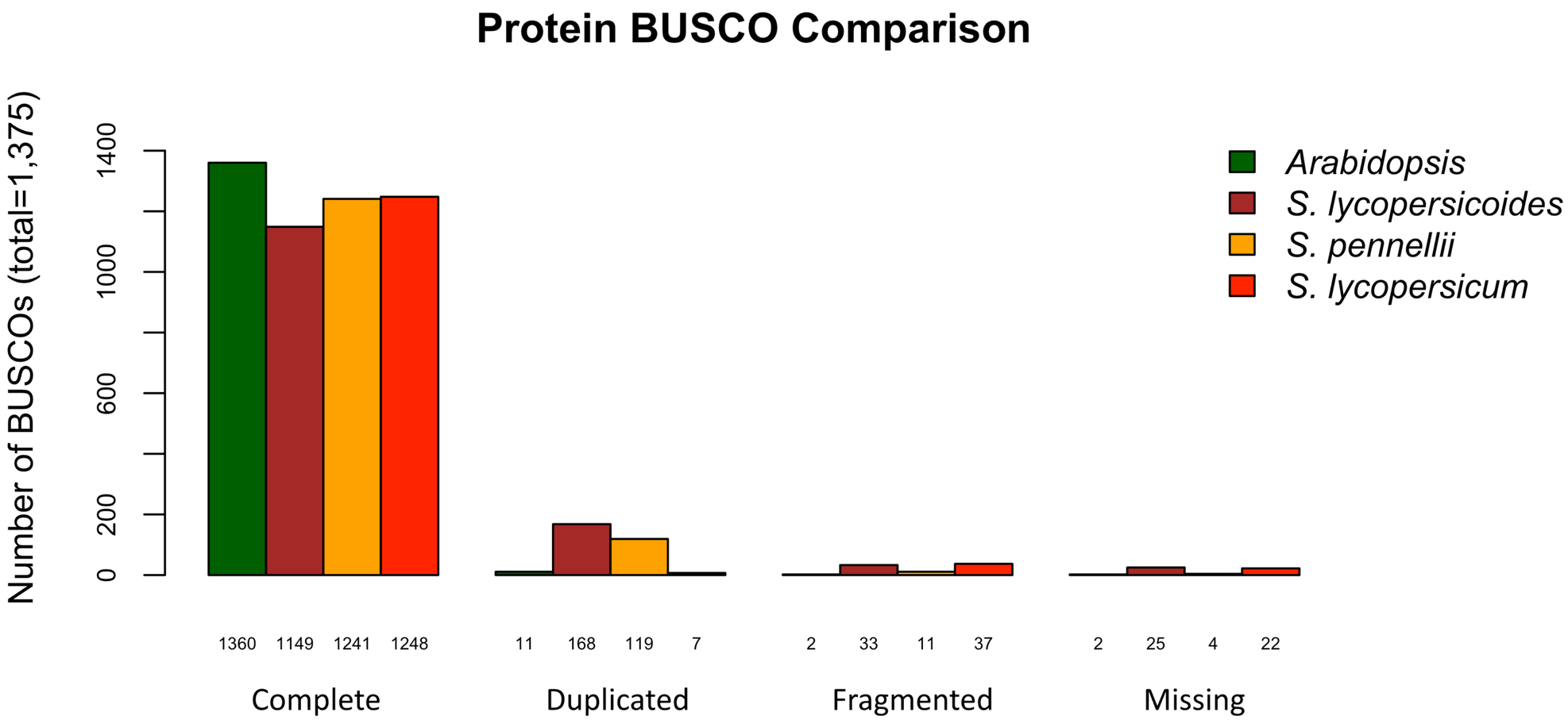
3. Annotation QC: Comparison to relative

	S. lycopersicoides v1.1	*S. pennellii* v2	*S. lycopersicum* v4.0
no. of gene models*	37,939	44,965	34,075
Average gene model length (bp)	4,388	5,962	3,571
Average CDS length (bp)*	1,232	1,549	1,027
Average exons/gene*	5.2	5.5	4.5
BUSCO	97.6%[S:87.2%,D:10.4%],F:0.4%,M:2.0%		97.5%[S:96.4%,D:1.1%],F:0.4%,M:2.1%
*calculated using the primary isoform			

3. Annotation QC: Gene families



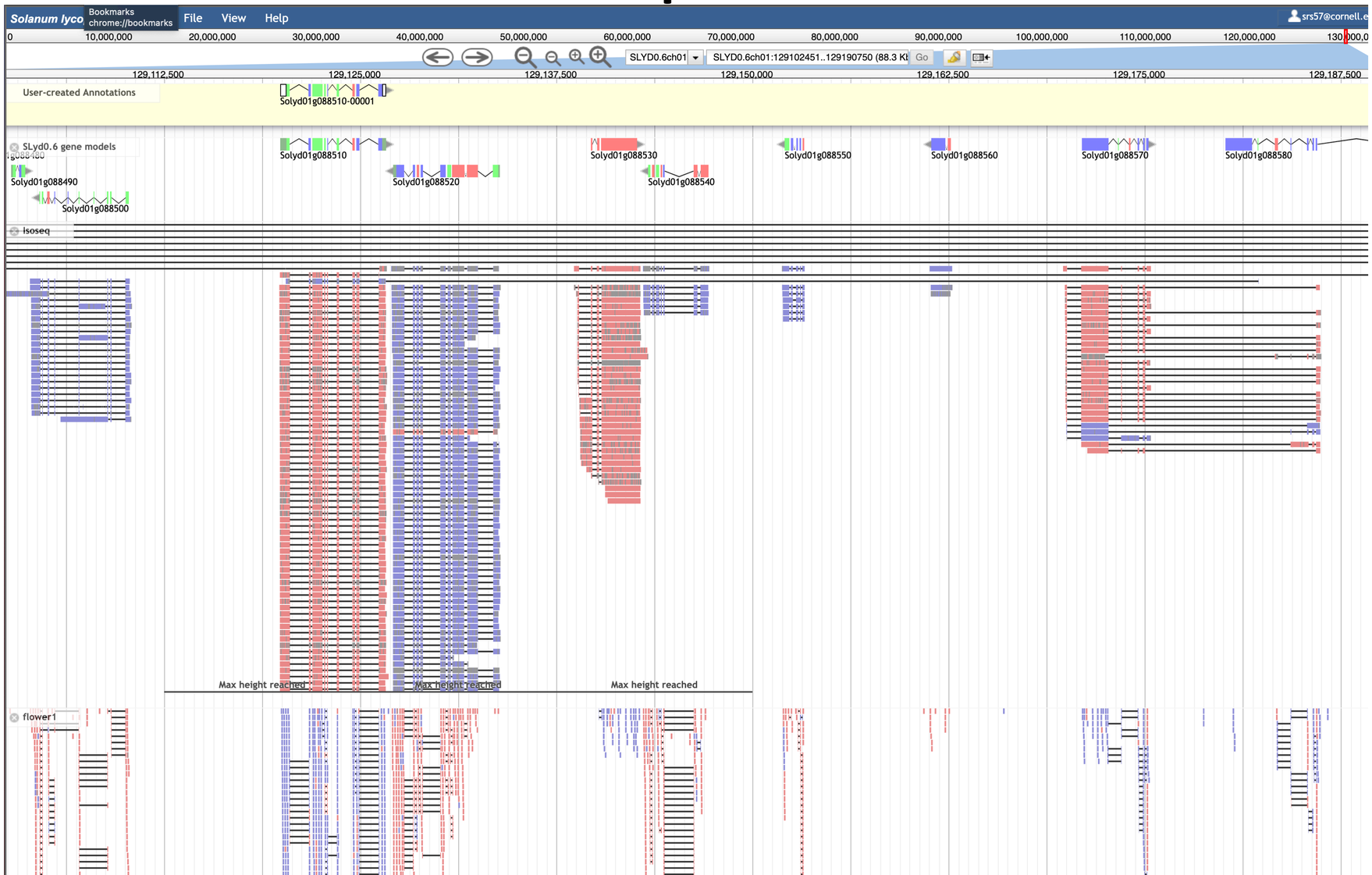
3. Annotation QC: BUSCO



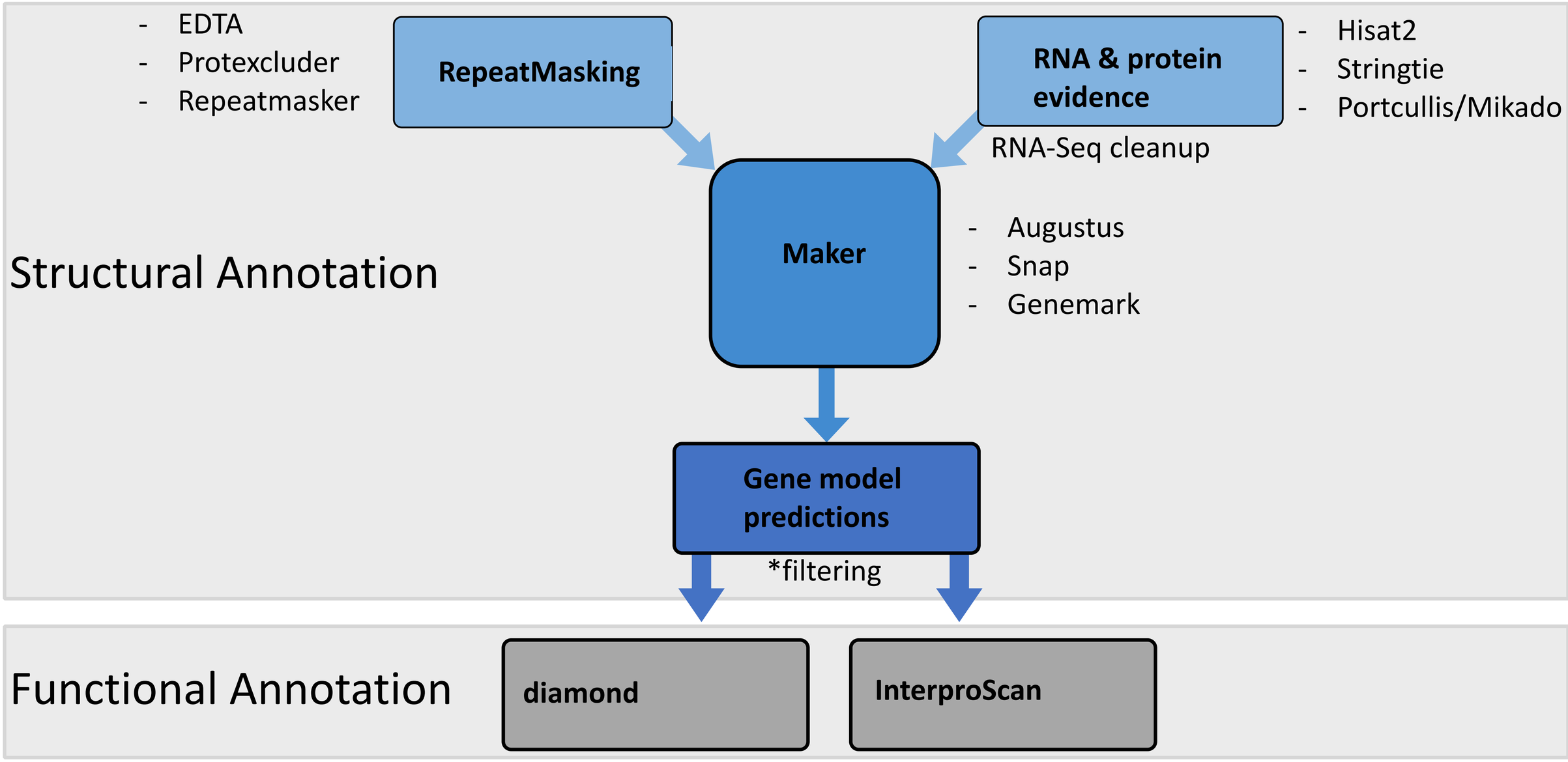
3. Annotation QC: Post-processing, Cleanup, and QC

- Change gene model names once structural annotation is completed.
  - Ex: maker-Contig3008-exonerate_est2genome-gene-0.0-mRNA-1 VS Solyd03g00650
- Versioning of genome and annotation (and keeping them in sync) – very important
- Apollo https://genomearchitect.readthedocs.io/en/latest/

3. Annotation QC: Manual curation with Apollo



Annotation pipeline

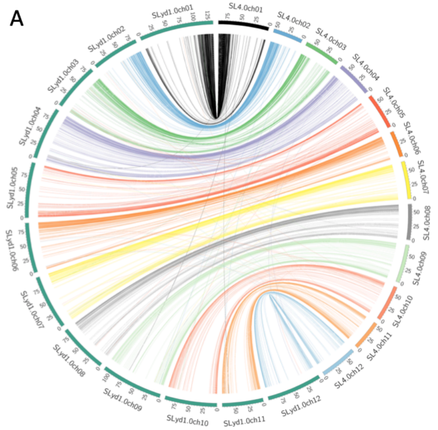


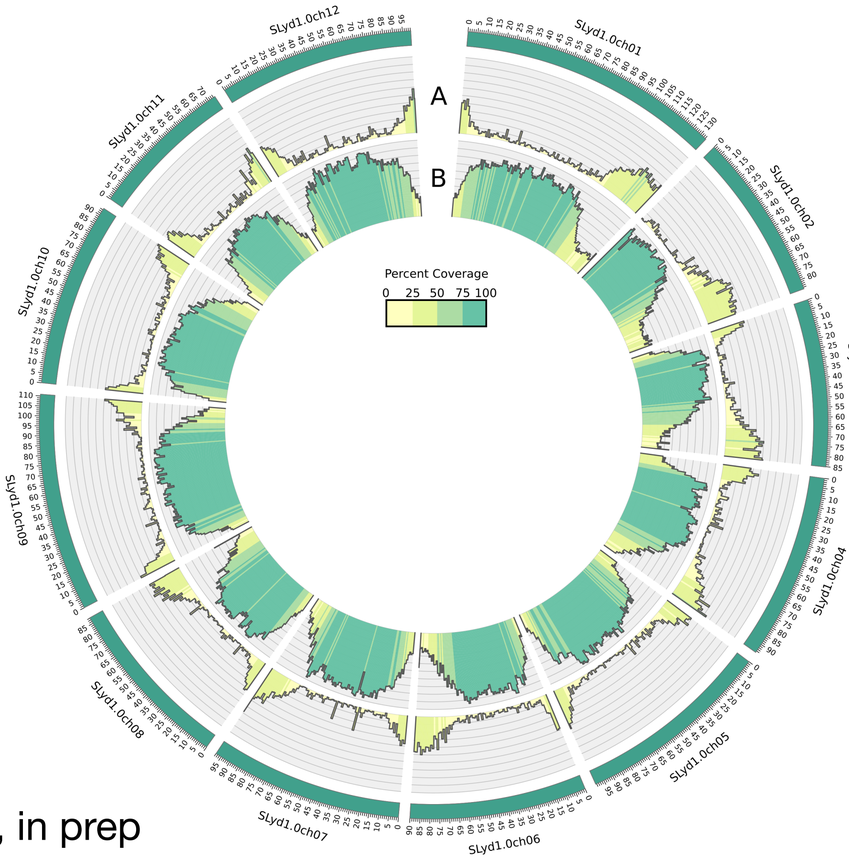
Functional annotation - tools

- Sequence searches
  - Diamond/BLAST
  - Databases: Swiss-prot, Trembl, nr, InterPro
- Domain searches
  - InterProScan
  - domains, GO terms, pathways
- Gene families
  - Orthofinder

Publishing your plant genome

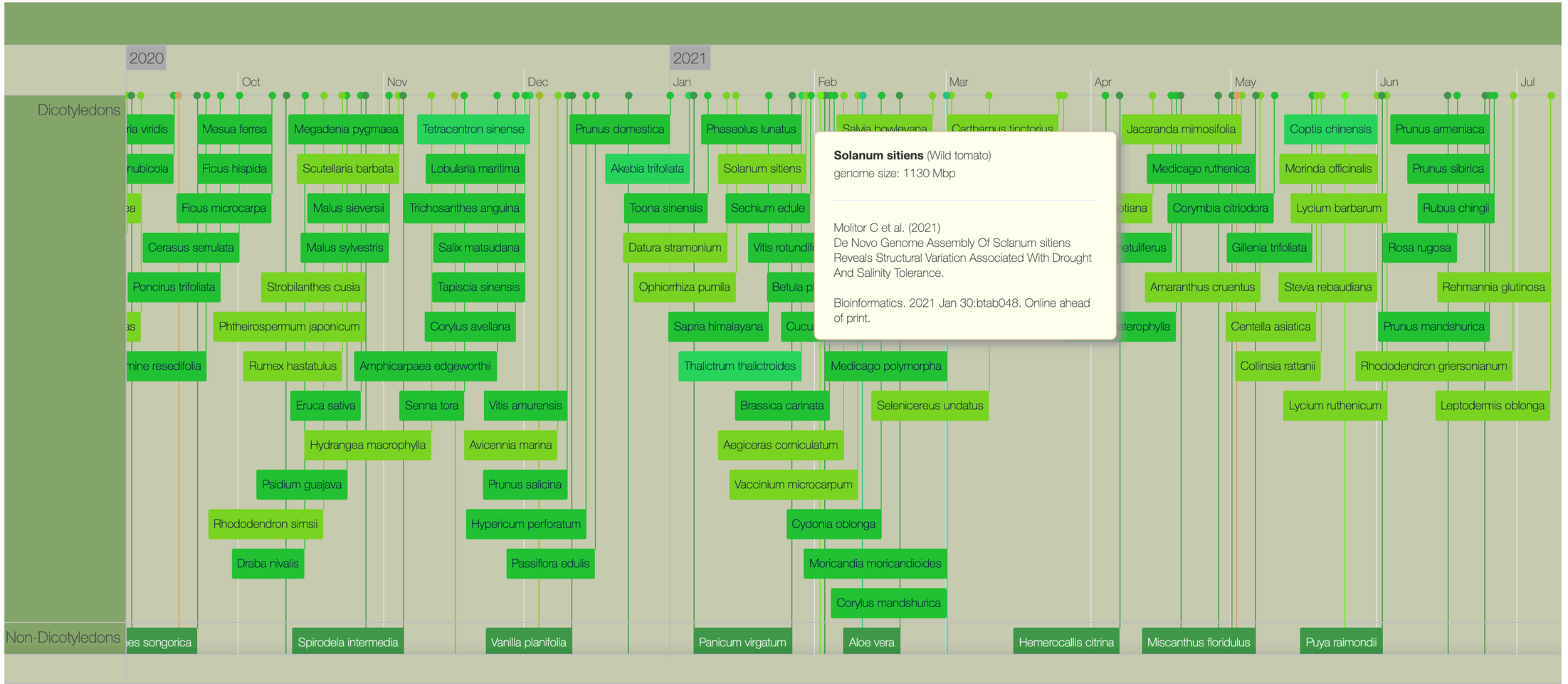
- Typical tables/figures (N50, gaps, etc, repeat content, gene families (expansion/ contraction), BUSCO, comparisons to reference)
- Circos plots
- Nice to have a biology hook
- Where to publish? https://plabipd.de/portal/sequenced-plant-genomes
- Submitting to Genbank: Project ID for publication
  - All supporting raw reads, annotation files, fasta files
- Organism-specific database
  - JBrowse
  - Apollo
  - Blast
- CyVerse/CoGe





Powell et al, in prep


plaBi
dataBase


2020
2021
Oct
Nov
Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Dicotyledons
Non-Dicotyledons
Mesua ferrea
Megadenia pygmaea
Tetracentron sinense
Prunus domestica
Phaseolus lunatus
Jacaranda mimosifolia
Coptis chinensis
Prunus armeniaca
Ficus hispida
Scutellaria barbata
Lobularia maritima
Akebia trifoliata
Solanum sitiens
Medicago ruthenica
Morinda officinalis
Prunus sibirica
Ficus microcarpa
Malus sieversii
Trichosanthes anguina
Toona sinensis
Sechium edule
Corymbia citriodora
Lycium barbarum
Rubus chingii
Cerasus serrulata
Malus sylvestris
Salix matsudana
Datura stramonium
Gillenia trifoliata
Rosa rugosa
Poncirus trifoliata
Strobilanthes cusia
Tapiscia sinensis
Ophiorrhiza pumila
Amaranthus cruentus
Stevia rebaudiana
Rehmannia glutinosa
Phtheirospermum japonicum
Corylus avellana
Sapria himalayana
Centella asiatica
Prunus mandshurica
Rumex hastatulus
Amphicarpaea edgeworthii
Thalictrum thalictroides
Medicago polymorpha
Collinsia rattanii
Rhododendron griersonianum
Eruca sativa
Senna tora
Vitis amurensis
Brassica carinata
Selenicereus undatus
Lycium ruthenicum
Leptodermis oblonga
Hydrangea macrophylla
Avicennia marina
Aegiceras corniculatum
Psidium guajava
Prunus salicina
Vaccinium microcarpum
Rhododendron simsii
Hypericum perforatum
Cydonia oblonga
Draba nivalis
Passiflora edulis
Moricandia moricandioides
Corylus mandshurica
Spirodela intermedia
Vanilla planifolia
Panicum virgatum
Aloe vera
Hemerocallis citrina
Miscanthus floridulus
Puya raimondii
Solanum sitiens (Wild tomato)
genome size: 1130 Mbp
Molitor C et al. (2021)
De Novo Genome Assembly Of Solanum sitiens Reveals Structural Variation Associated With Drought And Salinity Tolerance.
Bioinformatics. 2021 Jan 30:btab048. Online ahead of print.

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 /scratch/Botany2020NMGWorkshop/

All scripts are on GitHub

$ cd /scratch

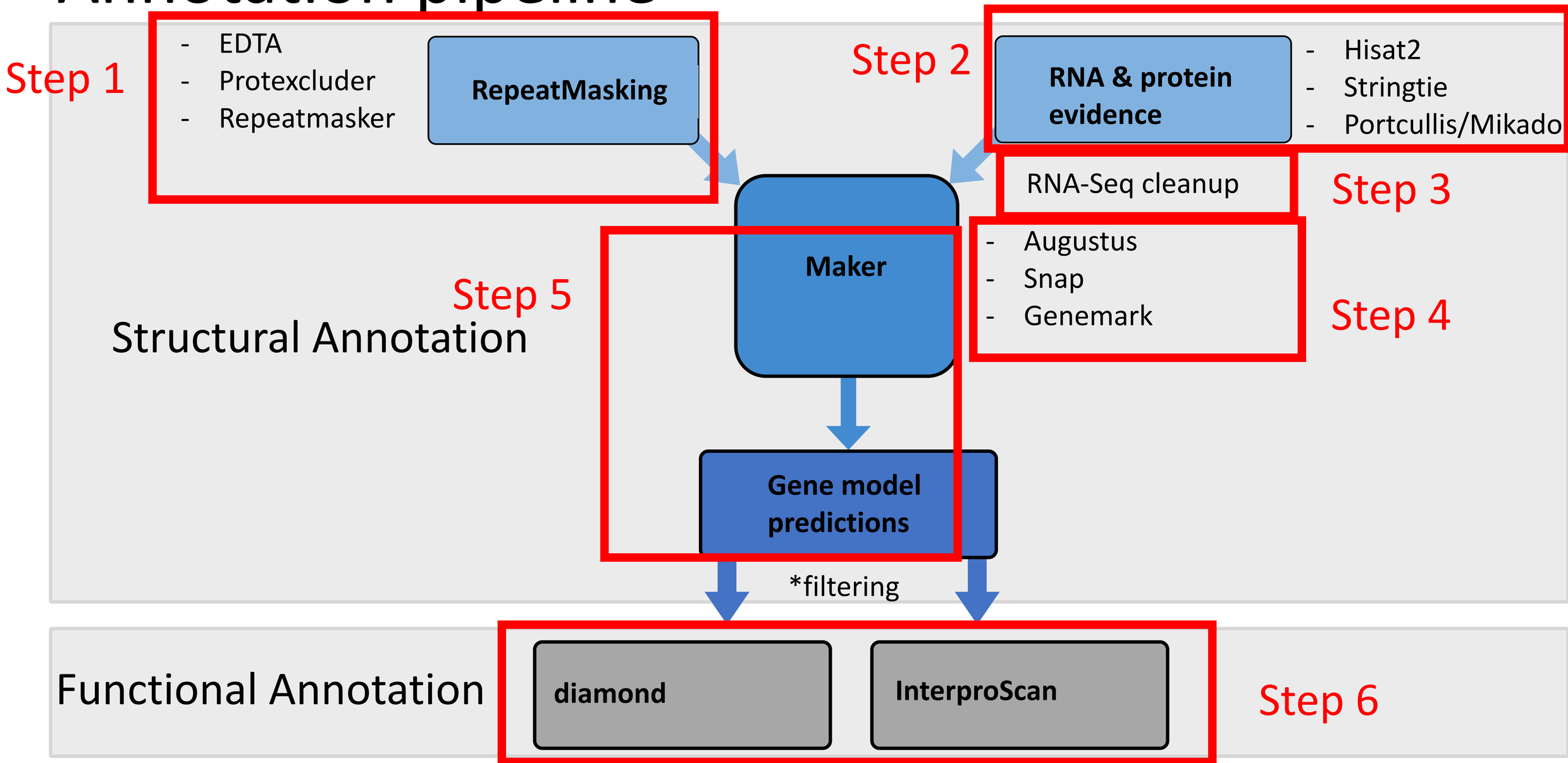
$ git clone https://github.com/bcbc-group/NMGWorkshop2021.git

$ 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



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

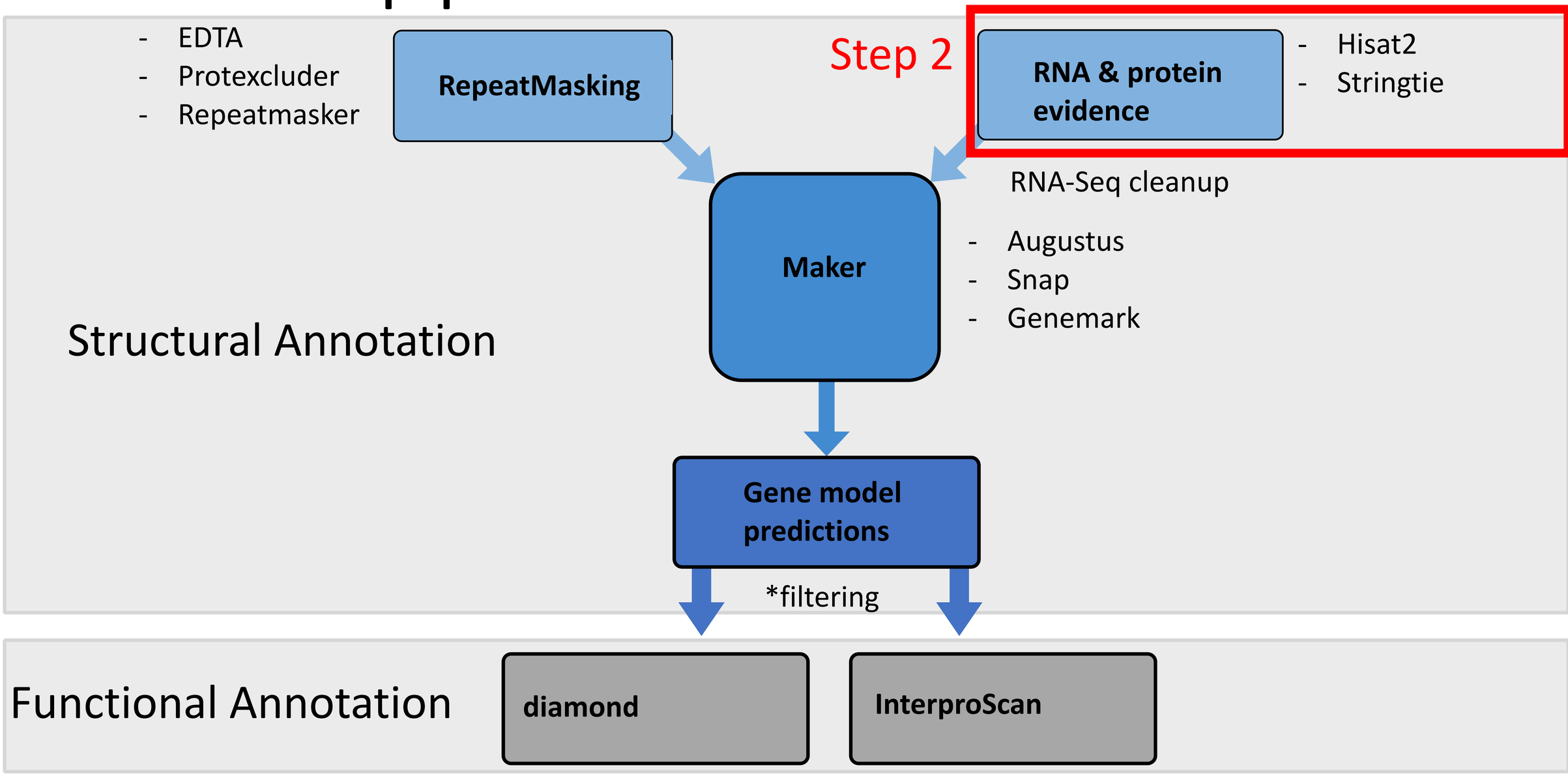
InterproScan

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

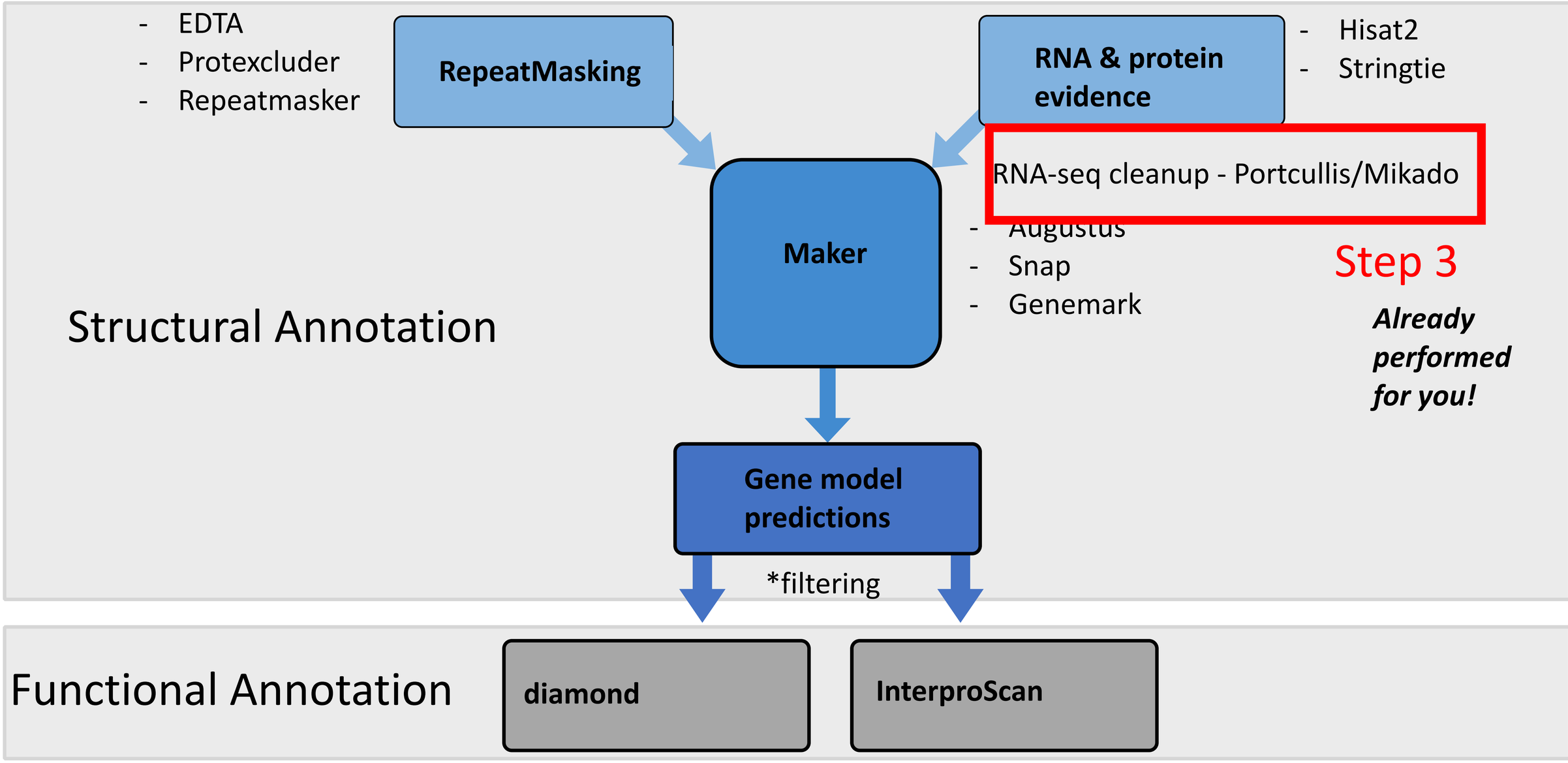
Annotation pipeline



Step 2: RNA-Seq read mapping

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

Annotation pipeline

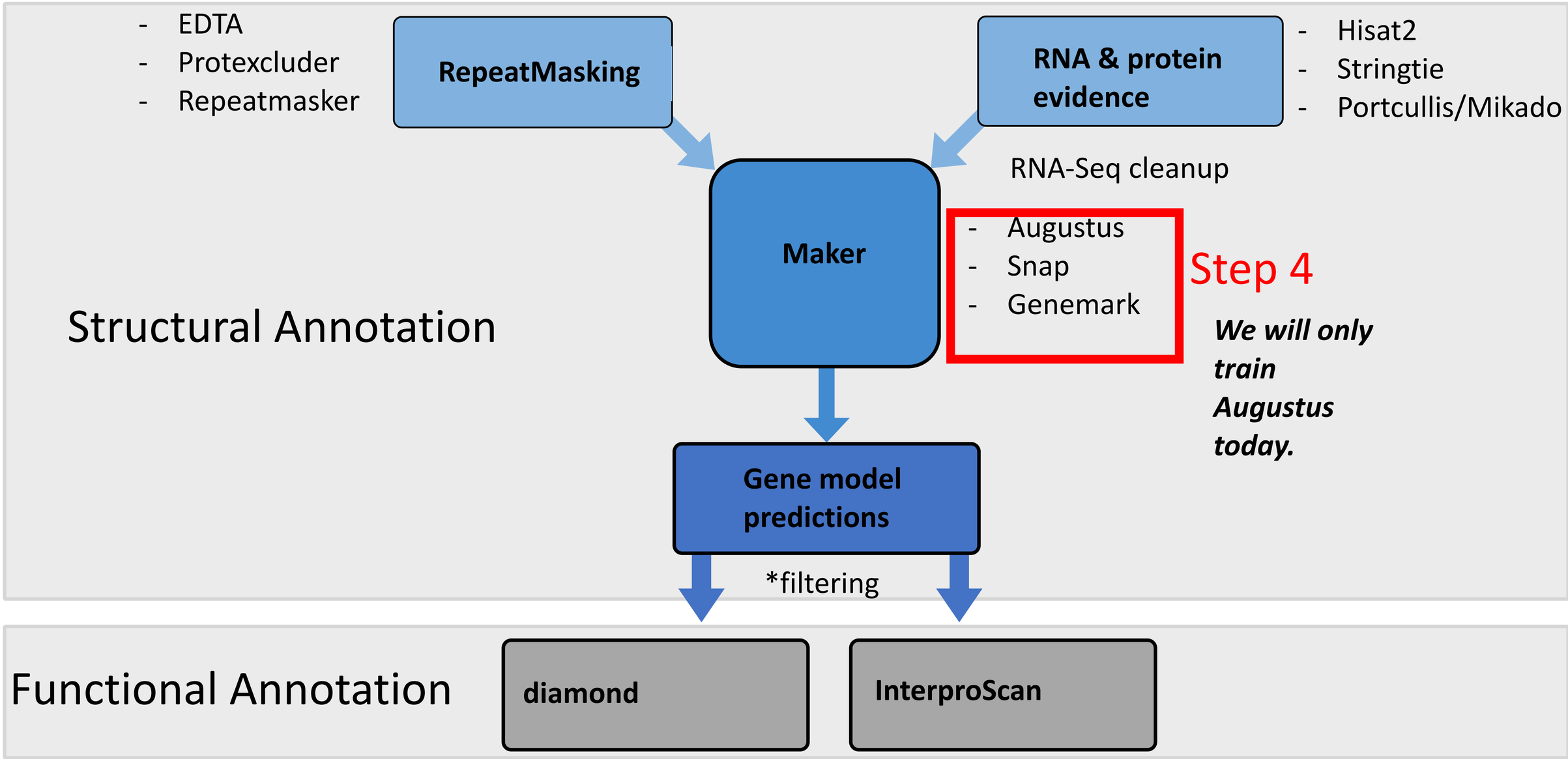


Step 3: RNA-seq cleanup

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



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

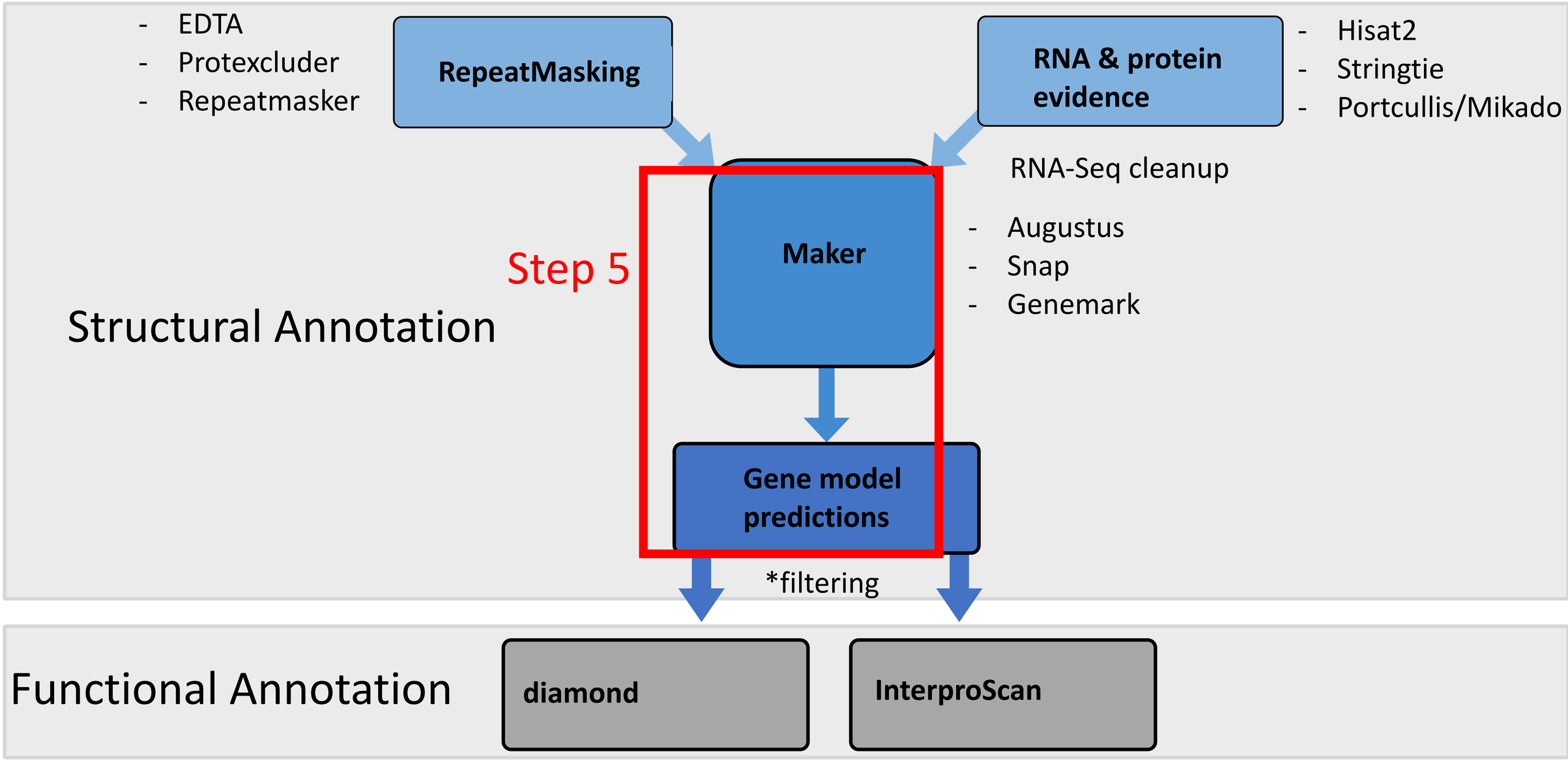
etraining --species=Ugibba genes.gb.train

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

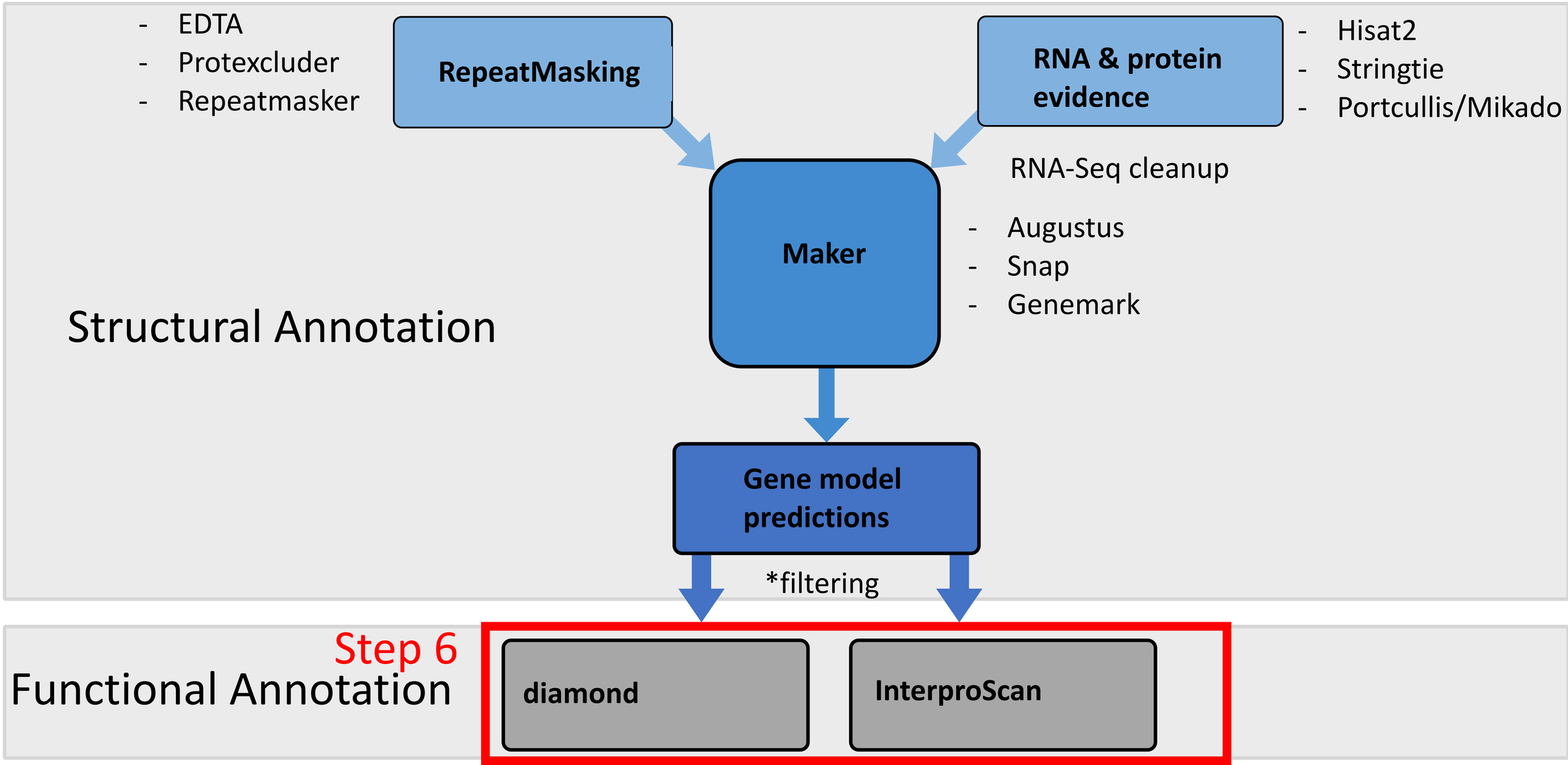


Step 5: Running maker

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

*this has already been performed to conserve time

Annotation pipeline



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