

Genome Annotation

Presented by Suzy Strickler
BCBC
Boyce Thompson Institute
Ithaca, NY

Objectives:

1. Understand steps involved in genome annotation
2. Demonstrate the use of data and tools that can be used in genome annotation
3. Learn how to QC genome assemblies and annotations
4. Understand how to derive functional predictions for genes

Goals of genome annotation

1. Predict, categorize, and mask repetitive elements
2. Determine gene structures as accurately as possible
3. Predict putative functions of predicted genes
4. Associate gene ontology terms, domains, etc for downstream analyses

Annotation Workflow Overview

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

1. Assembly QC

- Assembly quality (total length, N50, etc)

	<i>S. lycopersicoides</i>	<i>S. pennellii</i> v2	<i>S. lycopersicum</i> 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%)

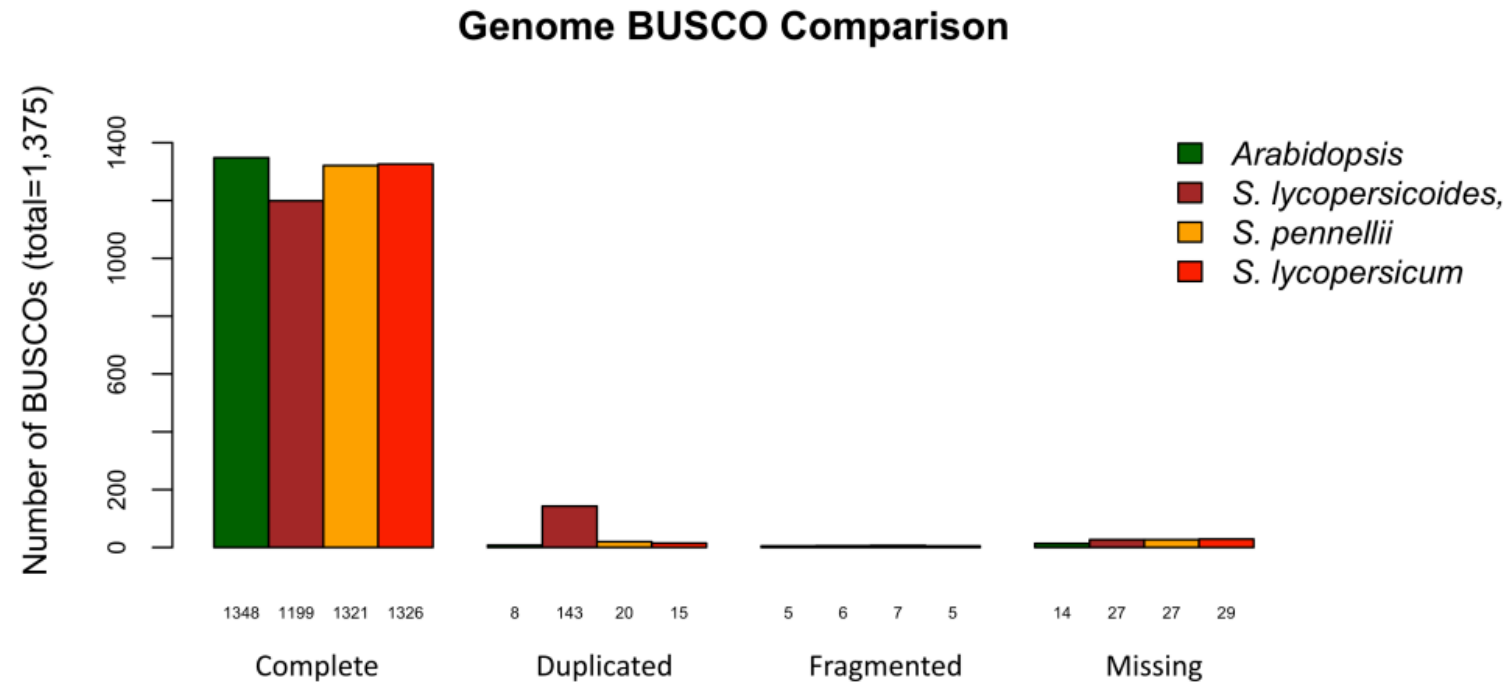
1. Assembly QC

- 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

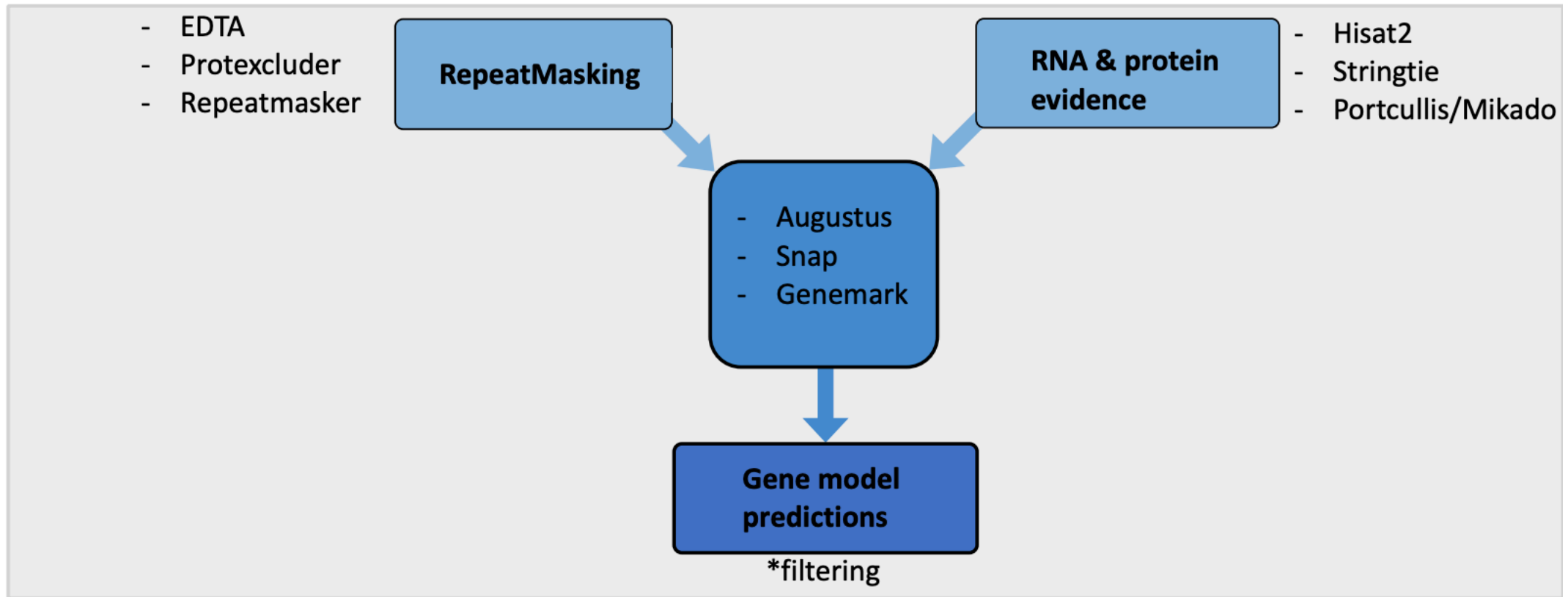
Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies

[Michael J. Roach](#) , [Simon A. Schmidt](#) & [Anthony R. Borneman](#)

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

10k Accesses | 136 Citations | 12 Altmetric | [Metrics](#)

2. Structural Annotation



Annotation pipelines:

Maker

Braker

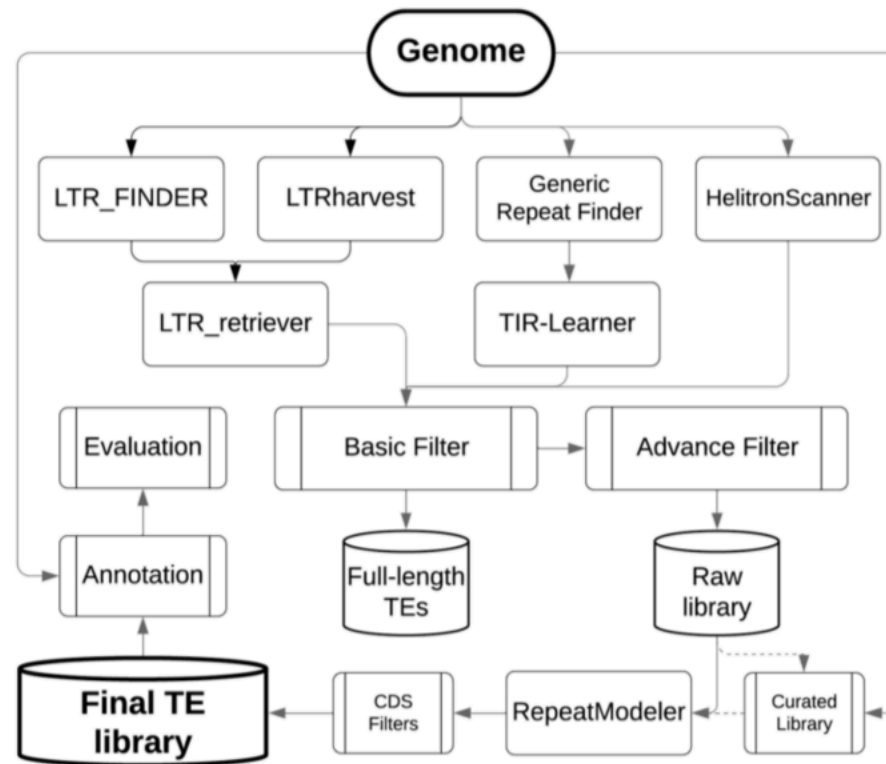
2. Structural Annotation: Repeat Masking

- 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/replib/>
 - 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* Prediction

- 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

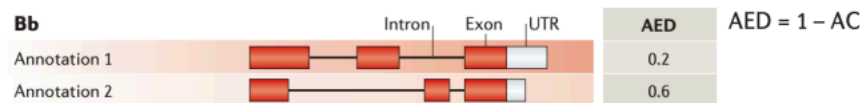
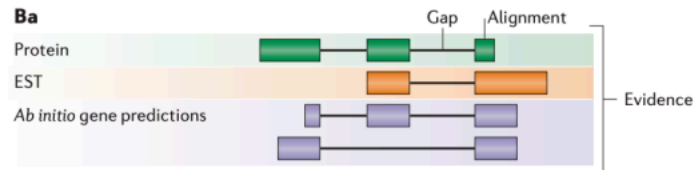
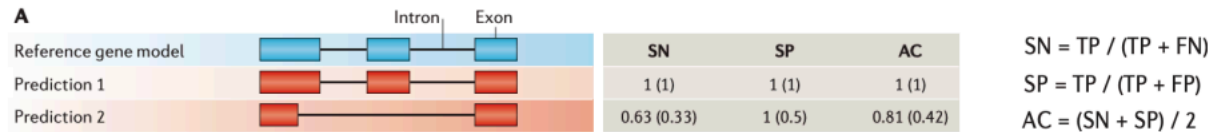
- Why/What? Tools to align RNA and protein evidence to genome, usually output to gff3 or bam
- Tools:
 - Hisat2 - align RNA-seq <http://daehwankimlab.github.io/hisat2/>
 - Gmap - align mRNA <https://academic.oup.com/bioinformatics/article/21/9/1859/409207>
 - Mikado/Portcullis - RNA-seq clean-up <https://mikado.readthedocs.io/en/stable/>
 - 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 WC: Postprocessing, Cleanup, and QC

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

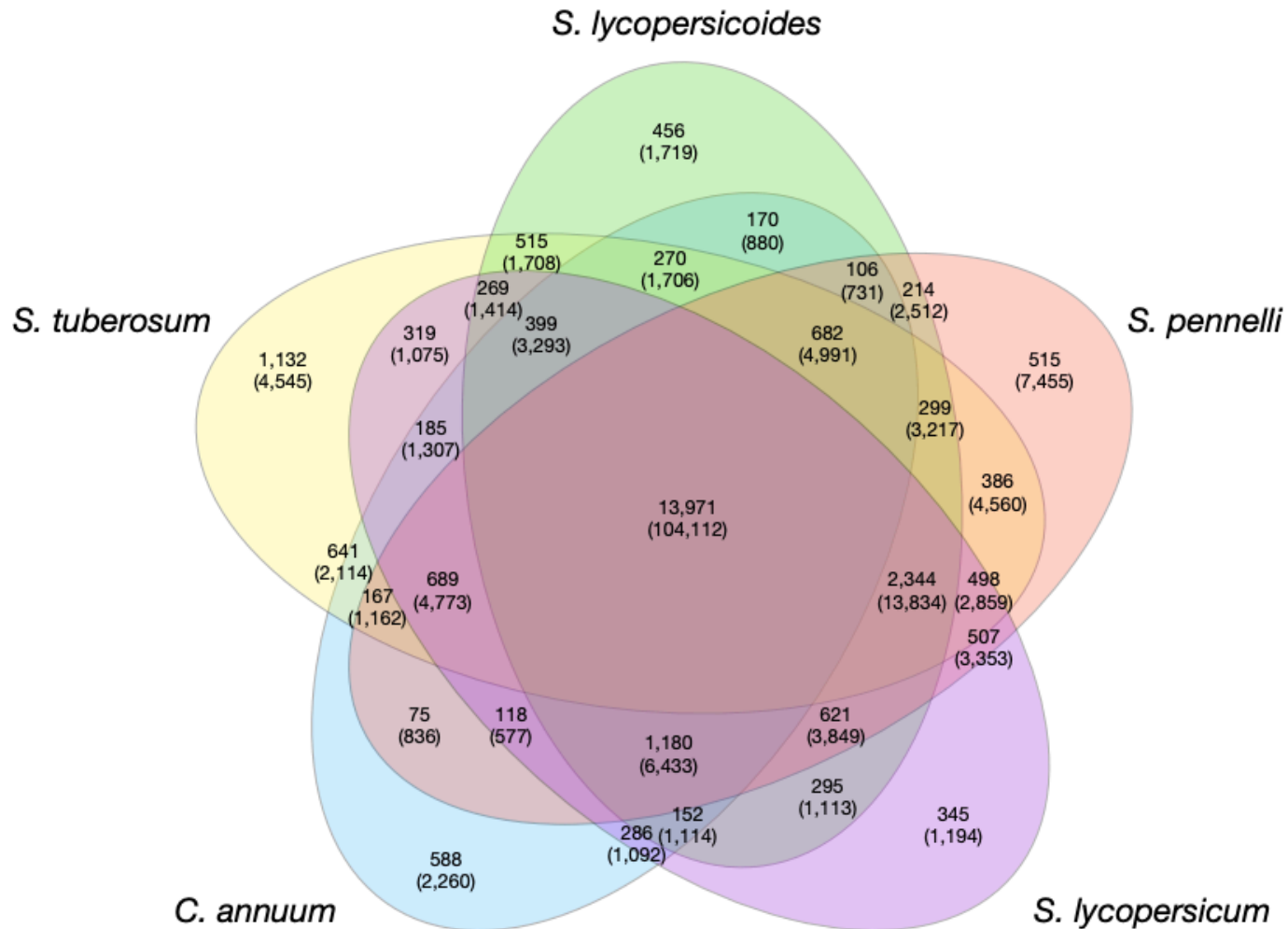


Yandell and Ence 2012

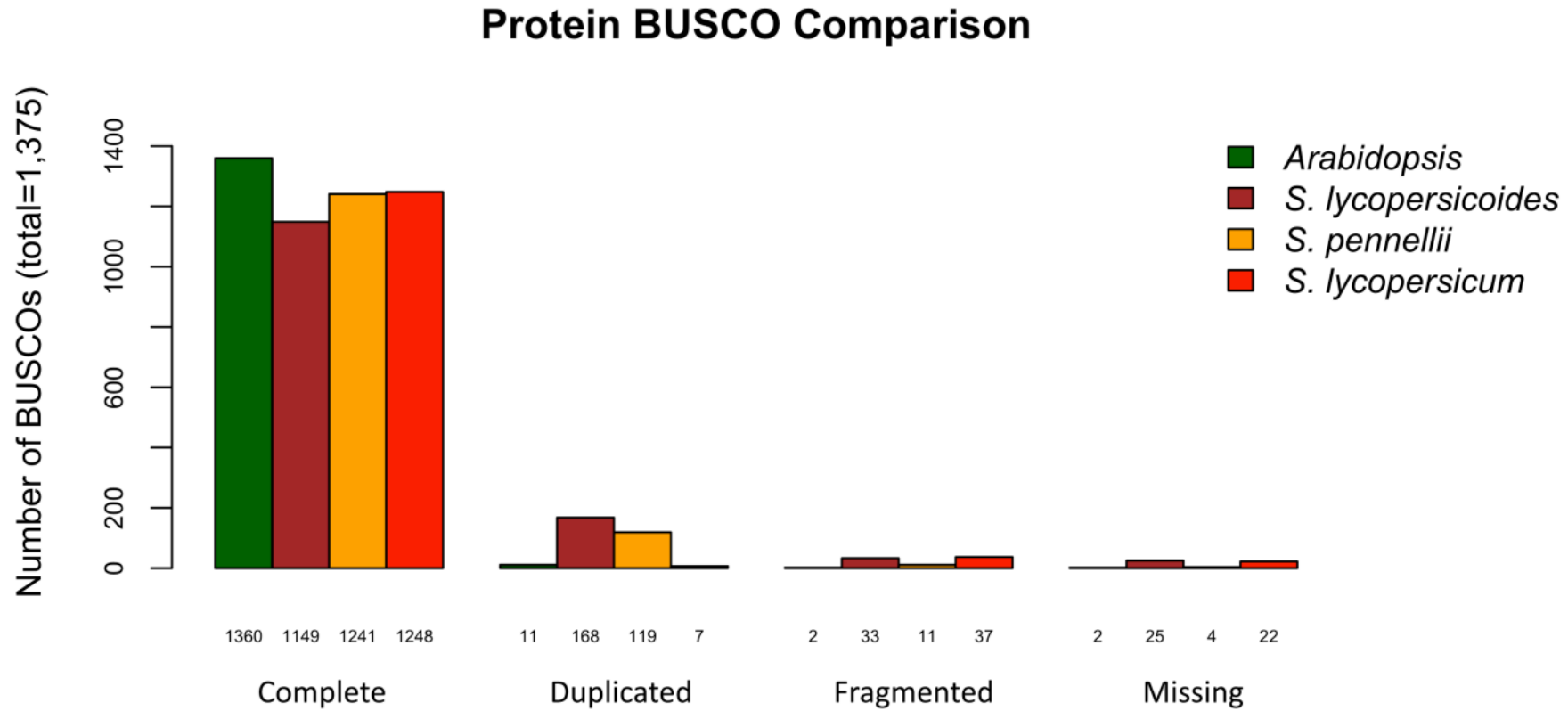
3. Annotation QC: Comparison to a related species

	<i>S. lycopersicoides</i> v1.1	<i>S. pennellii</i> v2	<i>S. lycopersicum</i> 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



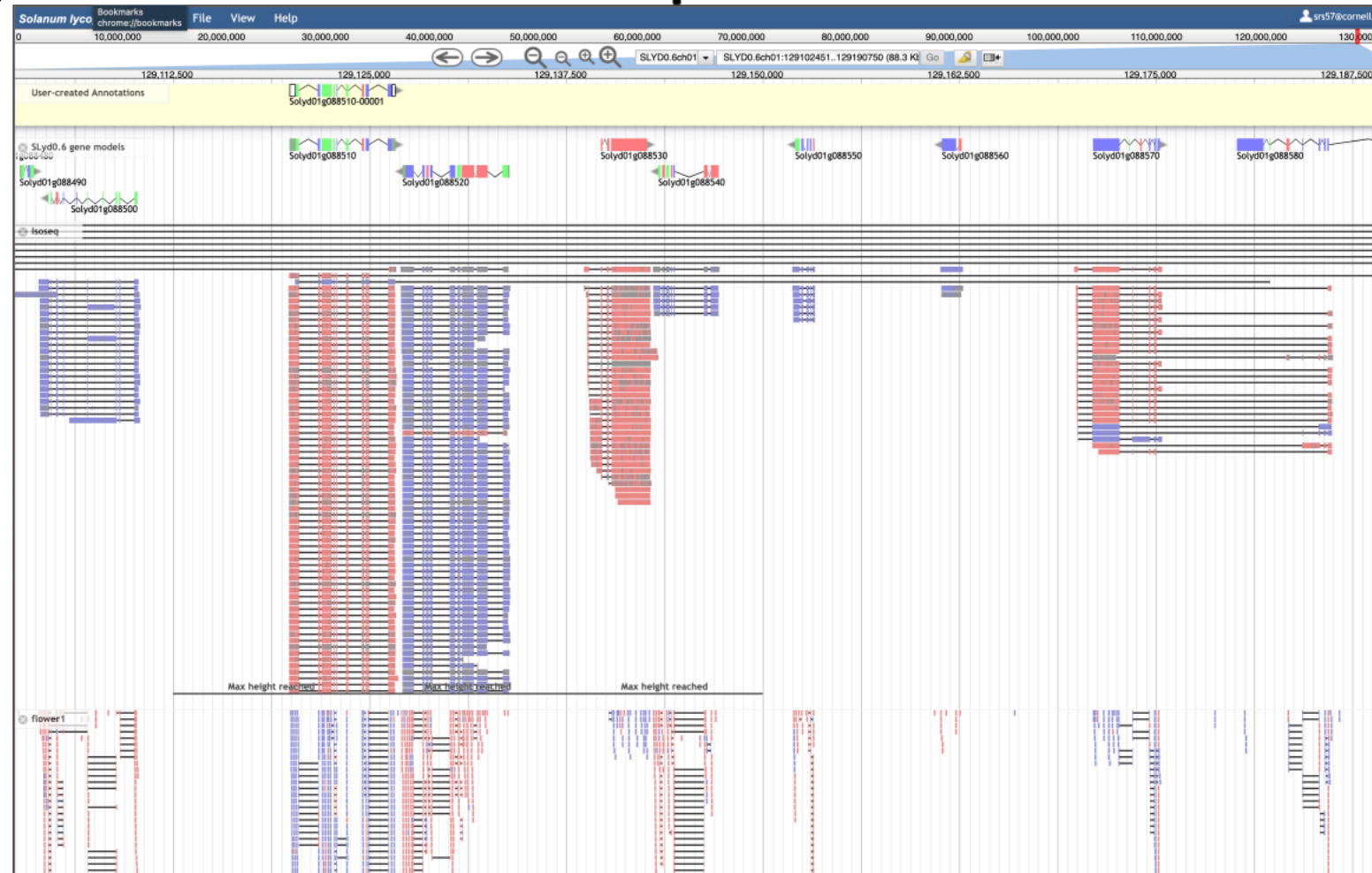
3. Annotation QC: BUSCO



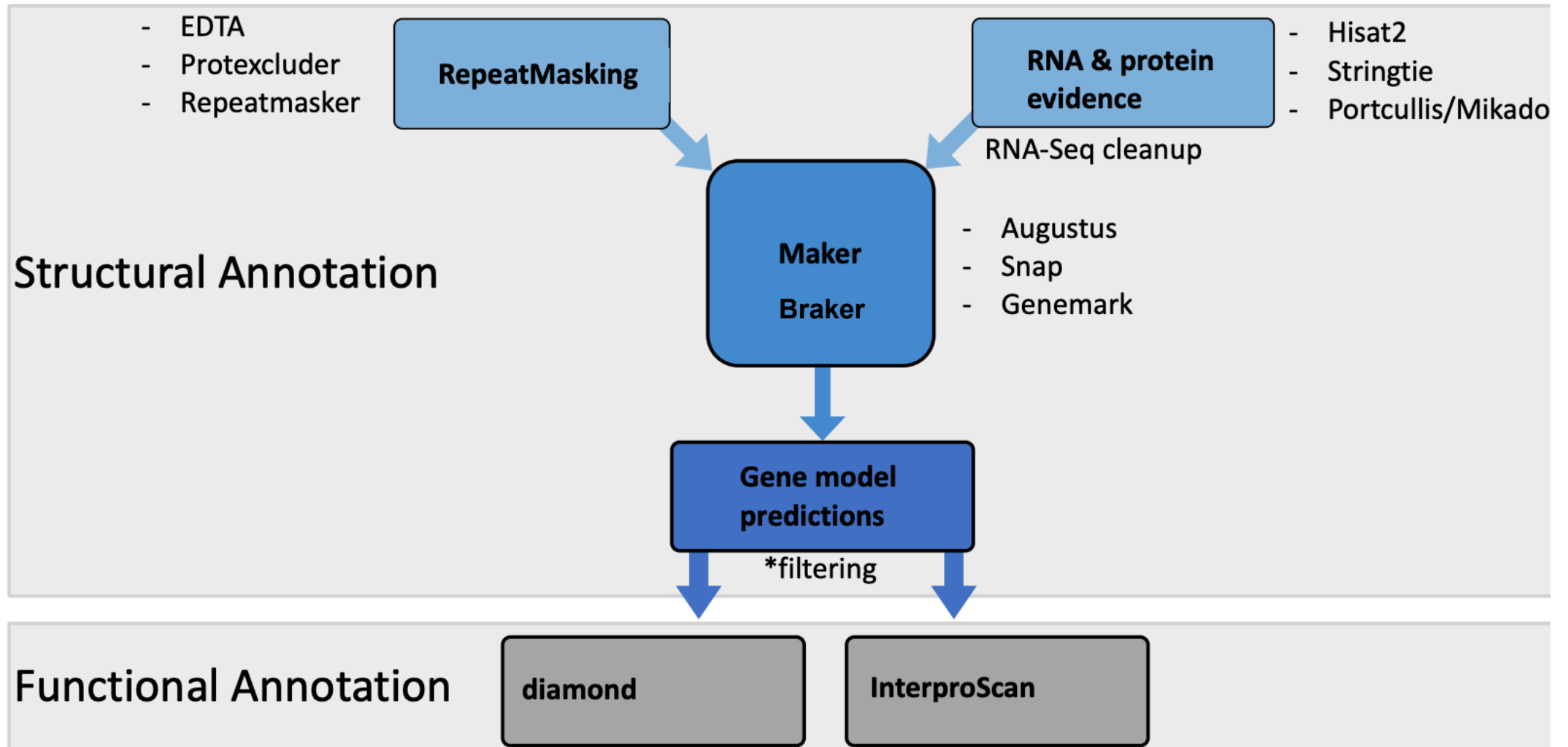
3. Annotation QC: Post-processing, Clean-up, and QC

- 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



4. Functional Annotation

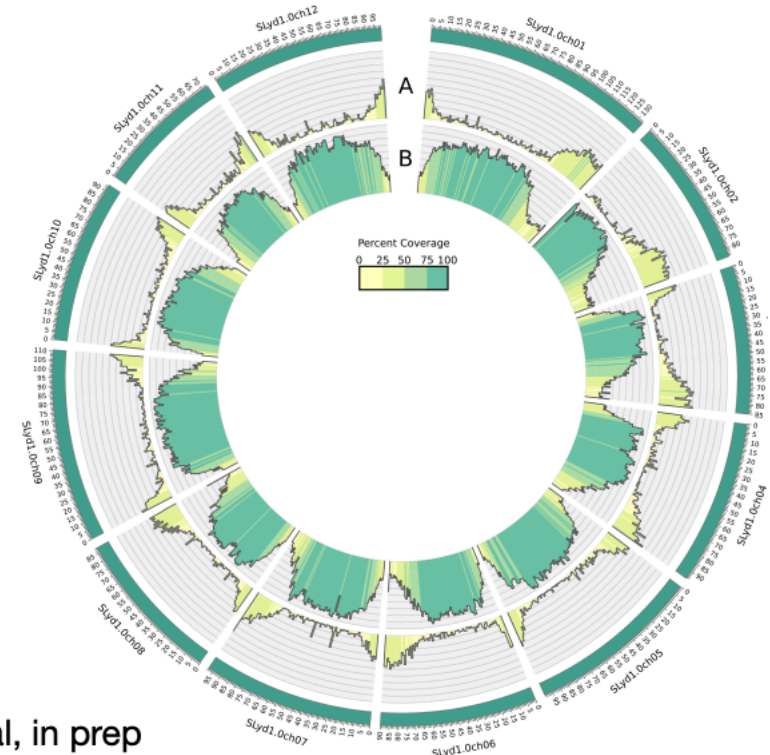
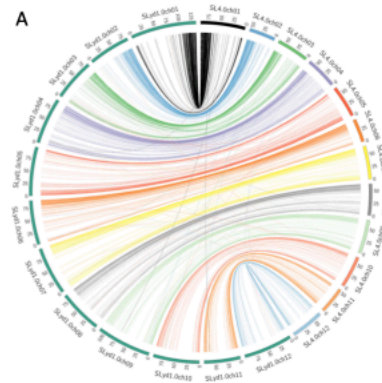


4. Functional Annotation: Tools

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

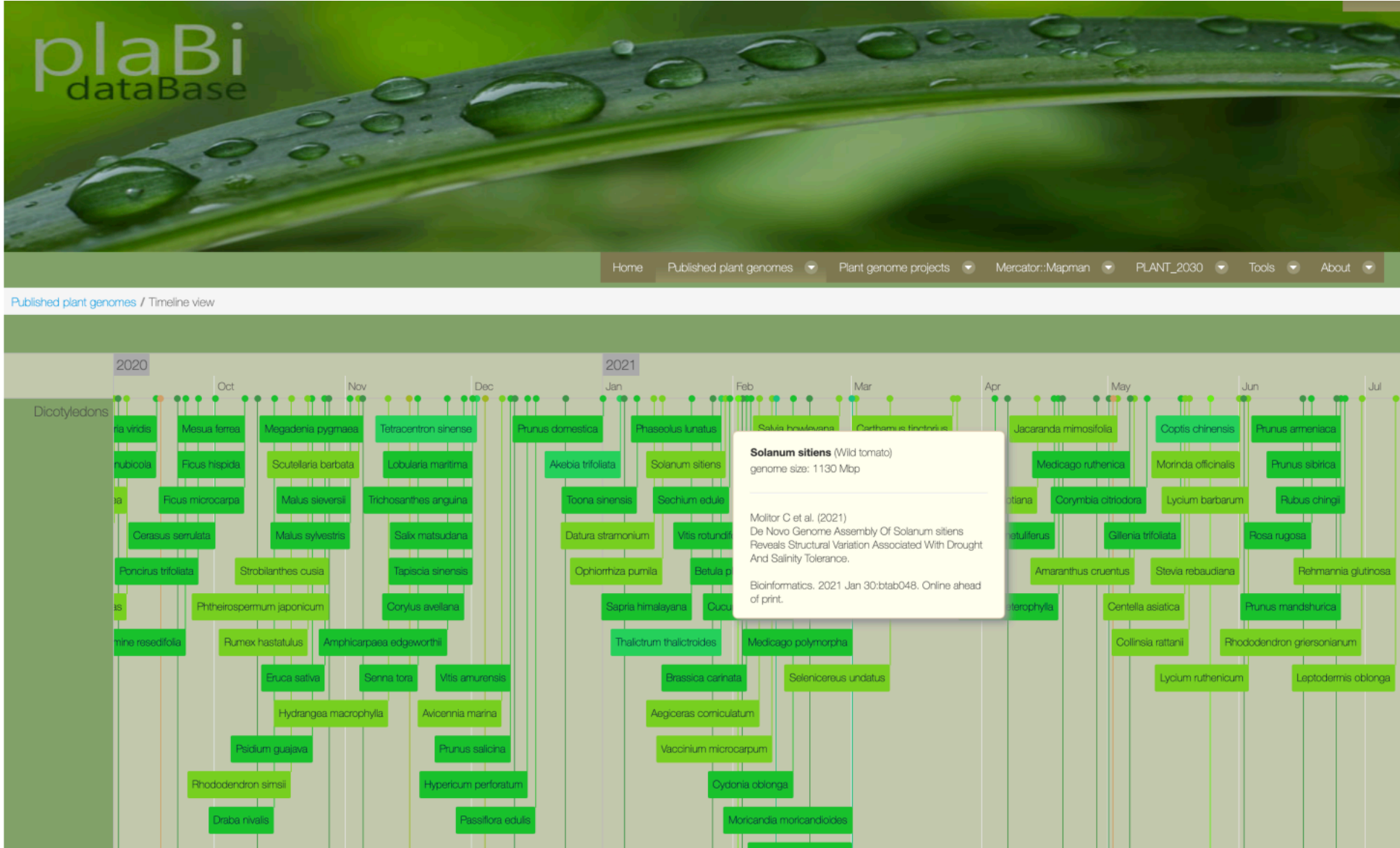
Publishing your plant genome

- 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

Publishing your plant genome



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])
- All this stuff plus some output files in
/home/user/work/data/iplant/home/shared/
Botany2020NMGWorkshop/

All scripts are on GitHub

```
$ cd ~/
```

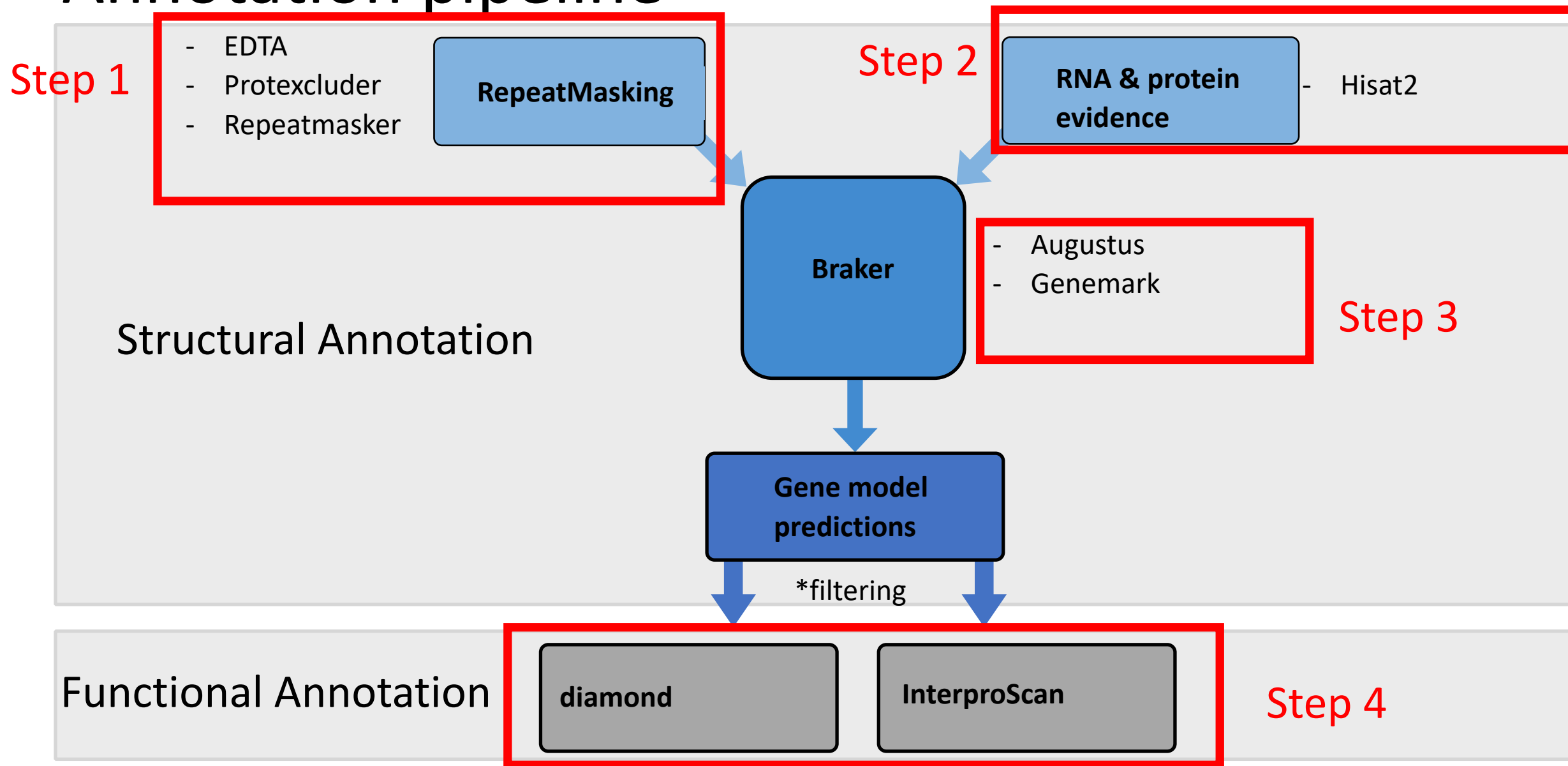
```
$ git clone https://github.com/bcbc-group/Botany2021NMGWorkshop.git
```

```
$ 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°

Braker

- Augustus
- Genemark

Structural Annotation

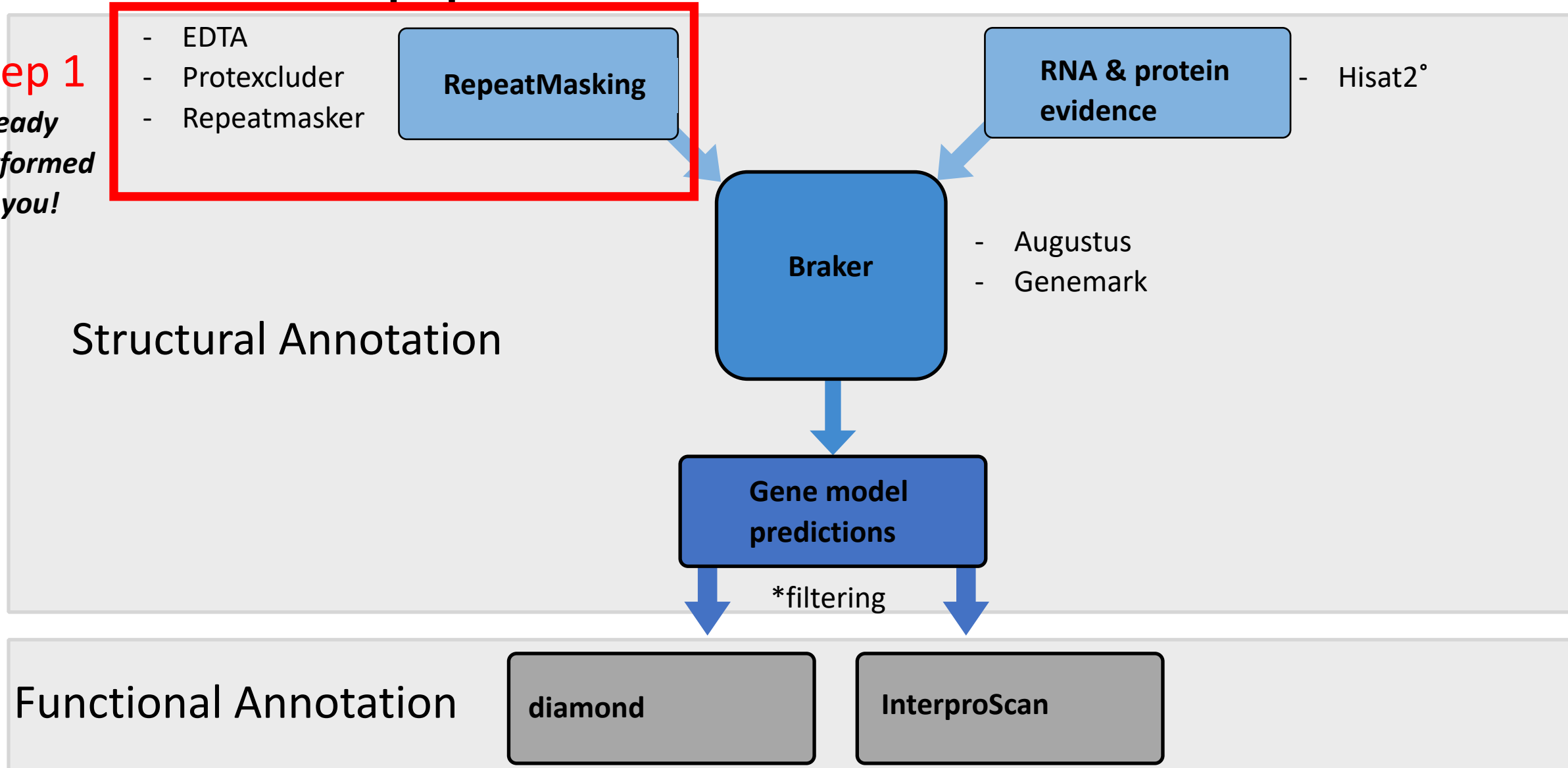
**Gene model
predictions**

*filtering

Functional Annotation

diamond

InterproScan

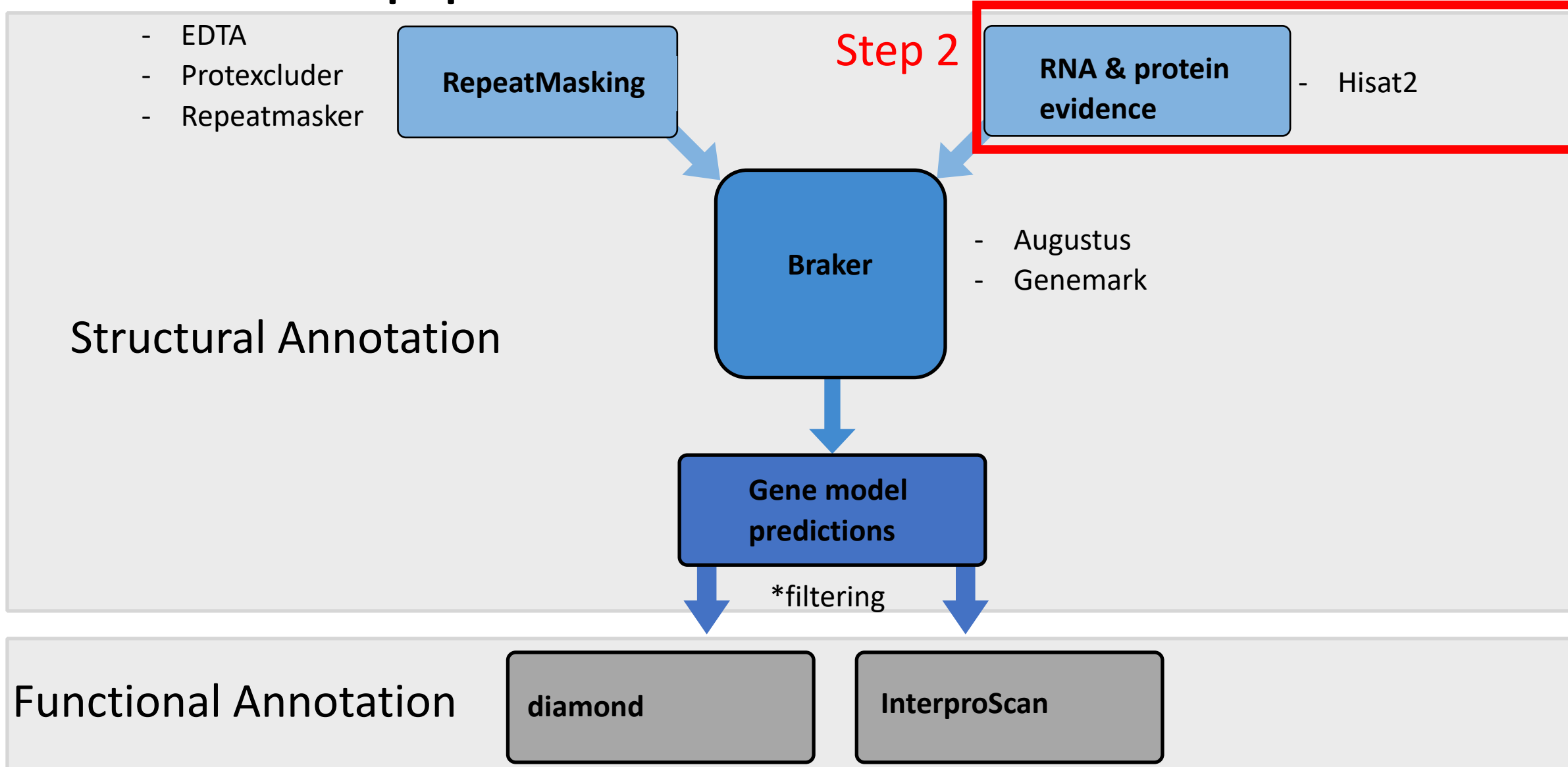


Step 1: Repeat Masking

https://github.com/bcbc-group/Botany2022NMGWorkshop/blob/main/5.Annotation/1_repeatmasking.sh

*this has already been performed to conserve time

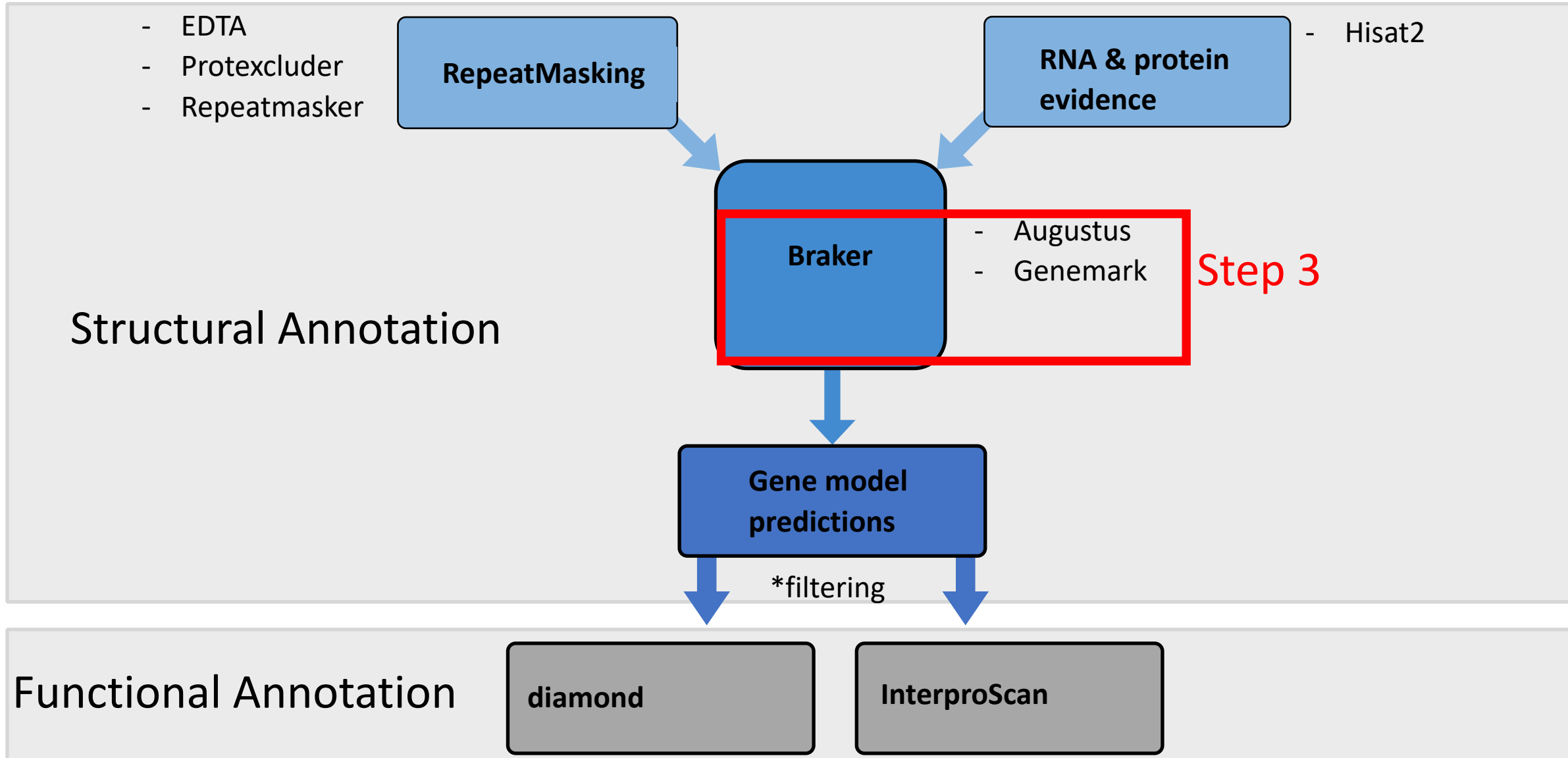
Annotation pipeline



Step 2: RNA-Seq read mapping

https://github.com/bcbc-group/Botany2022NMGWorkshop/blob/main/5.Annotation/2_hisat_pe_annot.sh

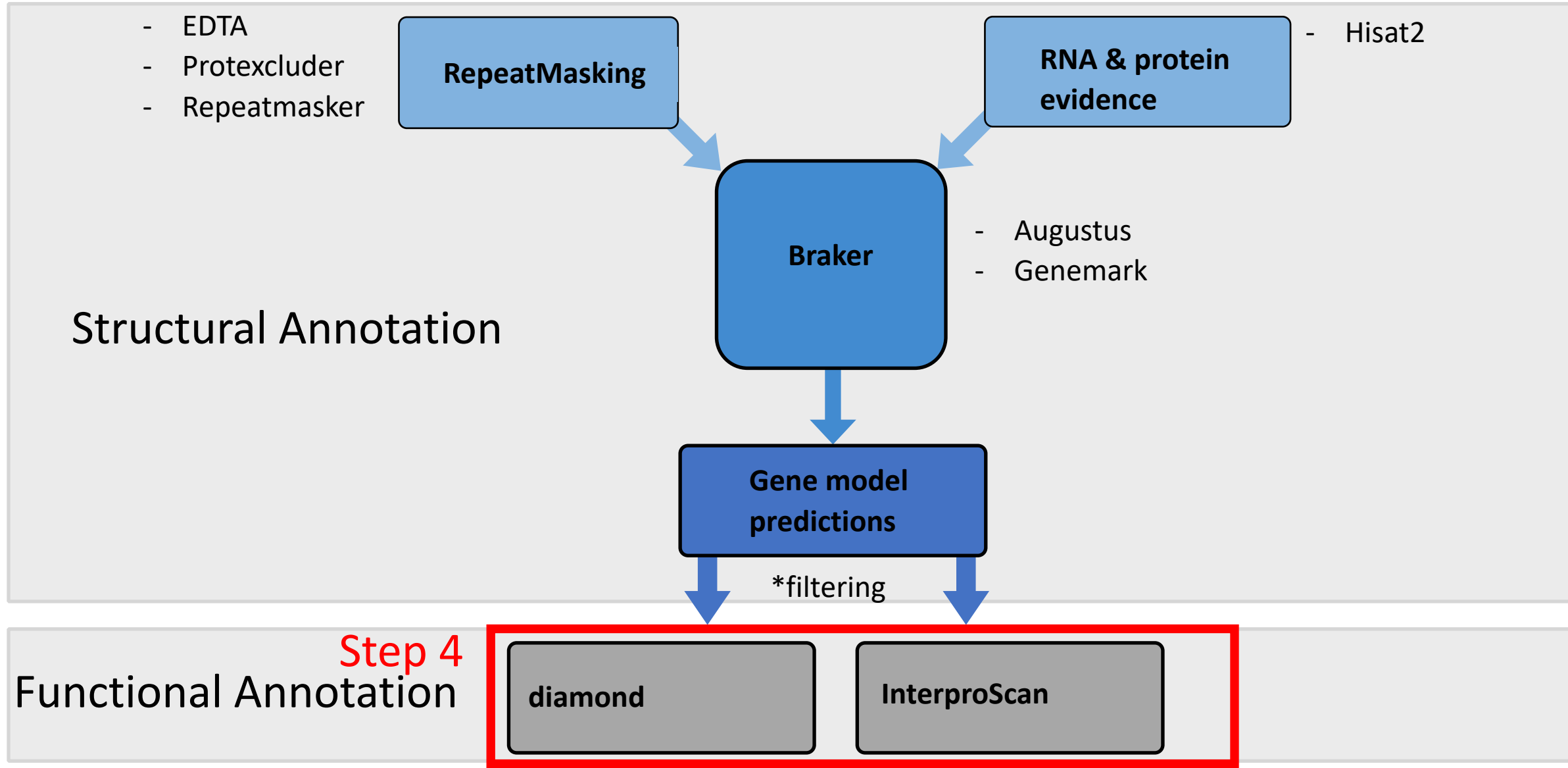
Annotation pipeline



Step 3: Running Braker

- https://github.com/bcbc-group/Botany2022NMGWorkshop/blob/main/5.Annotation/3_braker.sh

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 4: 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

Maker

- If you are interested in running maker check out: https://github.com/bcbc-group/Botany2021NMGWorkshop/tree/main/5.Annotation/maker_scripts