Berghia genome annotation

# Genome Filtering

## Run Purgedups

Input(s): Berghia\_Apr2021\_hirise.fasta (assembly file from Dovetail)

Output(s): Berghia\_Apr2021\_hirise\_purged.fasta

*Minimap2 version 2.18-r1015*

*Purge\_dups version 1.2.5*

Split fasta file into smaller sequences:

split\_fa Berghia\_Apr2021\_hirise.fasta > Berghia\_Apr2021\_hirise.fasta.split

Map split sequences to themselves:

minimap2 -xasm5 -DP Berghia\_Apr2021\_hirise.fasta.split Berghia\_Apr2021\_hirise.fasta.split -t 8 | gzip -c - > Berghia\_Apr2021\_hirise.fasta.split.self.paf.gz

Remove duplicates:

purge\_dups -2 -T cutoffs -c PB.base.cov Berghia\_Apr2021\_hirise.fasta.split.self.paf.gz > dups.bed 2> purge\_dups.log

get\_seqs -e dups.bed Berghia\_Apr2021\_hirise.fasta

mv purged.fa Berghia\_Apr2021\_hirise\_purged.fasta

## Run Blobtoolkit analyses

Input(s): Berghia\_Apr2021\_hirise\_purged.fasta, XDOVE\_2019\_pacbioreads\_merged.fastq.gz,

Output(s): Berghia\_Apr2021\_purged blobtools directory

*blobtools version 2.3.3*

*BUSCO version 5.1.2*

*Blastn version 2.11.0+*

*Minimap2 version 2.18-r1015*

Create blobtools directory for Berghia:

blobtools create --fasta Berghia\_Apr2021\_hirise\_purged.fasta --meta Berghia\_Apr2021\_purged.yaml --taxid 1287507 --taxdump taxdump Berghia\_Apr2021\_purged

BUSCO - Metazoa odb10:

busco -m genome -i Berghia\_Apr2021\_hirise\_purged -l metazoa\_odb10 -o berghia\_genome\_Apr2021\_purged\_busco.v4.0.5 -c 20

Coverage

minimap2 -ax sr -t 16 Berghia\_Apr2021\_hirise\_purged.fasta XDOVE\_2019\_pacbioreads\_merged.fastq.gz | samtools sort -@16 -O BAM -o Berghia\_Apr2021\_reallypurged\_pacbiomerged.bam

BLASTn runs

export BLASTDB=$BLASTDB:/ocean/projects/bio210009p/shared/data/databases/nt

blastn -db nt -query Berghia\_Apr2021\_hirise\_purged.fasta -outfmt "6 qseqid staxids bitscore std" -max\_target\_seqs 10 -max\_hsps 1 -evalue 1e-25 -num\_threads 30 -out Berghia\_Apr2021\_purged.ncbi.blastn.out

Add coverage, busco, and blast hits:

blobtools add \

--hits Berghia\_Apr2021\_reallypurged.ncbi.blastn.out\

--hits Berghia\_Apr2021\_hirise\_reallypurged.fasta.diamond.blastx.out \

--taxrule bestsumorder \

--taxdump /ocean/projects/bio210009p/shared/data/databases/taxdump \

--busco berghia\_genome\_Apr2021\_purged\_busco.v4.0.5/run\_metazoa\_odb10/full\_table.tsv

--cov Berghia\_Apr2021\_reallypurged\_pacbiomerged.bam \

Berghia\_Apr2021\_purged

--threads 8

## Filter genome

Input(s): Berghia\_Apr2021\_purged blobtools directory

Output(s): Berghia\_Apr2021\_hirise\_purged.filtered.fasta

Filtered genome by length and bacteria:

blobtools filter --param length--Min=200000 --param bestsumorder\_phylum--Keys=no-hit,Bacteria-undef,Tenericutes --fasta Berghia\_Apr2021\_hirise\_purged.fasta --summary STDOUT Berghia\_Apr2021\_purged

RECOMMENDED - Simplify scaffold names:

NOTE: The main reason for this is that BRAKER2 (Augustus specifically I think) has issues with scaffold names that have symbols in them or are too long. I fixed this much later in the process (which is why this file is not used in the next few steps), but this command would have removed the extra (and unnecessary) symbols.

sed 's/;HRSCAF=[0-9]\*\_[0-9]\*//g' Berghia\_Apr2021\_hirise\_purged.filtered.fasta | sed 's/\_//g' > Berghia\_Apr2021\_hirise\_purged.filtered.edited.fasta

# Repeat Masking

Input(s): Berghia\_Apr2021\_hirise\_purged.filtered.fasta, Bsteph\_ref.fa.no\_tes.fa (reference proteome for Berghia from transcriptome assembly)

Output(s): Repeatmasked genome (\*.softmasked, \*.hardmasked)

## Make Repeat Library (RepeatModeler)

*RepeatModeler Version 2.0.1*

*BLAST 2.10.0+*

[*https://blaxter-lab-documentation.readthedocs.io/en/latest/filter-repeatmodeler-library.html*](https://blaxter-lab-documentation.readthedocs.io/en/latest/filter-repeatmodeler-library.html)

Compile initial repeat library (with Classifications):

RepeatModeler -database berghia -pa 2 -LTRStruct

Blast proteome against RepeatMasker TE database:

blastp -query ../Berghia\_alltissues\_onerep\_trinity291\_transdecoder\_cdhit95\_noaliens\_fulltranscripts\_novectors\_nocontaminants.fasta.transdecoder.pep -db /ocean/projects/bio210009p/shared/tools/miniconda3/envs/repeatmodeler/share/RepeatMasker/Libraries/RepeatPeps.lib -outfmt '6 qseqid staxids bitscore std sscinames sskingdoms stitle' -max\_target\_seqs 25 -culling\_limit 2 -num\_threads 8 -evalue 1e-5 -out Bsteph\_ref.pep.vs.RepeatPeps.25cul2.1e5.blastp.out

Remove TEs from proteome:

fastaqual\_select.pl -f ../Berghia\_alltissues\_onerep\_trinity291\_transdecoder\_cdhit95\_noaliens\_fulltranscripts\_novectors\_nocontaminants.fasta.transdecoder.cds -e <(awk '{print $1}' Bsteph\_re..pep.vs.RepeatPeps.25cul2.1e5.blastp.out | sort | uniq) > Bsteph\_ref.fa.no\_tes.fa

Blast proteome against RepeatModeler library:

makeblastdb -in Bsteph\_ref.fa.no\_tes.fa -dbtype nucl

blastn -task megablast -query consensi.fa.classified -db Bsteph\_ref.fa.no\_tes.fa -outfmt '6 qseqid staxids bitscore std sscinames sskingdoms stitle' -max\_target\_seqs 25 -culling\_limit 2 -num\_threads 8 -evalue 1e-10 -out repeatmodeller\_lib.v.Bsteph\_ref.fa.no\_tes.25cul2.1e10.megablast.out

Remove hits from RepeatModeler library:

fastaqual\_select.pl -f consensi.fa.classified -e <(awk '{print $1}' repeatmodeller\_lib.v.Bsteph\_ref.fa.no\_tes.25cul2.1e25.megablast.out | sort | uniq) > consensi.fa.classified.filtered\_for\_CDS\_repeats.fa

## Repeat Mask Genome (RepeatModeler)

*RepeatMasker version 4.1.2-p1*

Hardmask for STAR aligner:

RepeatMasker Berghia\_Apr2021\_hirise\_purged.filtered.fasta -e ncbi -lib RM\_24730.MonJul121509252021/consensi.fa.classified.filtered\_for\_CDS\_repeats.fa -pa 20

mv Berghia\_Apr2021\_hirise\_purged.filtered.fasta.masked Berghia\_Apr2021\_hirise\_purged.filtered.fasta.hardmasked

mv Berghia\_Apr2021\_hirise\_purged.filtered.fasta.out Berghia\_Apr2021\_hirise\_purged.filtered.fasta\_hard.out

mv Berghia\_Apr2021\_hirise\_purged.filtered.fasta.cat Berghia\_Apr2021\_hirise\_purged.filtered.fasta\_hard.cat

mv Berghia\_Apr2021\_hirise\_purged.filtered.fasta.tbl Berghia\_Apr2021\_hirise\_purged.filtered.fasta\_hard.tbl

Softmask for Augustus/BRAKER:

RepeatMasker Berghia\_Apr2021\_hirise\_purged.filtered.fasta -e ncbi -lib RM\_24730.MonJul121509252021/consensi.fa.classified.filtered\_for\_CDS\_repeats.fa -xsmall -pa 20

mv Berghia\_Apr2021\_hirise\_purged.filtered.fasta.masked Berghia\_Apr2021\_hirise\_purged.filtered.fasta.softmasked

mv Berghia\_Apr2021\_hirise\_purged.filtered.fasta.out Berghia\_Apr2021\_hirise\_purged.filtered.fasta\_soft.out

mv Berghia\_Apr2021\_hirise\_purged.filtered.fasta.cat Berghia\_Apr2021\_hirise\_purged.filtered.fasta\_soft.cat

mv Berghia\_Apr2021\_hirise\_purged.filtered.fasta.tbl Berghia\_Apr2021\_hirise\_purged.filtered.fasta\_soft.tbl

# Alignment of RNA-seq reads (Illumina)

Input(s): Berghia\_Apr2021\_hirise\_purged.filtered.fasta.hardmasked, combined Illumina short reads from all stages and tissues listed in paper

Output(s): Berghia\_Apr2021\_masked\_starmappedAligned\_sort.bam and Berghia\_Apr2021\_masked\_starmappedAligned\_sort.bam.bai

*STAR version 2.7.9a*

*HiSat2 version 2.2.1*

## Align RNA-seq data

### STAR

STAR runThreadN 8 --runMode genomeGenerate --genomeSAindexNbases 13 --genomeDir star\_index\_repeatmodeler\_masked --genomeFastaFiles Berghia\_Apr2021\_hirise\_purged.filtered.fasta.hardmasked --twopassMode --outSAMtype BAM SortedByCoordinate

STAR --runThreadN 8 --genomeDir star\_index\_repeatmodeler\_masked --readFilesIn Berghia\_alltissues\_allstages\_withbrain\_R1.fq.gz Berghia\_alltissues\_allstages\_withbrain\_R2.fq.gz --outFileNamePrefix Berghia\_Apr2021\_repeatmodeler\_masked\_starmapped --readFilesCommand zcat

samtools view -bS Berghia\_Apr2021\_repeatmodeler\_masked\_starmappedAligned.out.sam > Berghia\_Apr2021\_repeatmodeler\_masked\_starmappedAligned.out.bam

samtools sort Berghia\_Apr2021\_repeatmodeler\_masked\_starmappedAligned.out.bam -o Berghia\_Apr2021\_repeatmodeler\_masked\_starmappedAligned\_sort.bam

samtools index Berghia\_Apr2021\_repeatmodeler\_masked\_starmappedAligned\_sort.bam

### HiSat2

hisat2-build -p 8 Berghia\_Apr2021\_hirise\_purged.filtered.fasta.hardmasked hisat2\_index\_repeatmodeler\_masked

hisat2 -p 8 -x hisat2\_index\_repeatmodeler\_masked -1 Berghia\_alltissues\_allstages\_withbrain\_R1.fq.gz -2 Berghia\_alltissues\_allstages\_withbrain\_R2.fq.gz -S Berghia\_Apr2021\_repeatmodeler\_masked\_hisat2mapped.sam

samtools view -bS Berghia\_Apr2021\_repeatmodeler\_masked\_starmappedAligned.out.sam > Berghia\_Apr2021\_repeatmodeler\_masked\_starmappedAligned.out.bam

samtools sort Berghia\_Apr2021\_repeatmodeler\_masked\_starmappedAligned.out.bam -o Berghia\_Apr2021\_repeatmodeler\_masked\_starmappedAligned\_sort.bam

samtools index Berghia\_Apr2021\_repeatmodeler\_masked\_starmappedAligned\_sort.bam

# Alignment of IsoSeq reads (PacBio)

Input(s): Berghia\_Apr2021\_hirise\_purged.filtered.fasta.hardmasked, combined IsoSeq long reads from all stages and tissues listed in paper

Output(s): Berghia\_Apr2021\_masked\_starmappedAligned\_sort.bam and Berghia\_Apr2021\_masked\_starmappedAligned\_sort.bam.bai

*minimap2 version 2.18-r1015*

## Align RNA-seq data

minimap2 -t 30 -ax splice -uf --secondary=no -C5 -O6,24 -B4 \

Berghia\_Apr2021\_hirise\_purged.filtered.fasta \

lyons\_alltissues\_flnc.fasta \

> berghia\_flnc\_isoforms.sam \

2> berghia\_flnc\_isoforms.sam.log

samtools view -S -b berghia\_flnc\_isoforms.sam > berghia\_flnc\_isoforms.bam

samtools sort berghia\_flnc\_isoforms.bam -o berghia\_flnc\_isoforms.sorted.bam

samtools index berghia\_flnc\_isoforms.sorted.bam

# BRAKER2 Annotation

Input(s): Berghia\_Apr2021\_hirise\_purged.filtered.fasta\_edited.softmasked,

Output(s): Annotation gtf and fasta files for masked genome

*BRAKER2 version 2.1.6*

## Initial Annotation

braker.pl --species=Bsteph --genome=Berghia\_Apr2021\_hirise\_purged.filtered.fasta\_edited.softmasked \

--prot\_seq mollusca\_odb10\_with\_berghiabuscos.fasta \

--bam=Berghia\_Apr2021\_repeatmodeler\_masked\_starmappedtwopassAligned.sortedByCoord.out.bam,Berghia\_Apr2021\_repeatmodeler\_masked\_hisat2mapped\_sort\_shortheaders.bam,berghia\_flnc\_isoforms\_bestlocus.sorted.bam \

--etpmode --softmasking -cores 12 \

--useexisting --gff3

## Filter Annotations

Input(s): BRAKER2 files augustus.hints.gtf and hintsfile.gff

Output(s): Filtered annotation files

*BRAKER2 version 2.1.6*

### Filter sequences

From BRAKER/scripts/prediction\_analysis:

python selectSupportedSubsets.py --anySupport augustus.anysupport.gtf --fullSupport augustus.fullsupport.gtf --noSupport augustus.nosupport.gtf augustus.hints.gtf hintsfile.gff

sh filter\_seqs.sh

filter\_seqs.sh:

#pull out the transcripts from the filtered gtf file

awk '{print $10}' augustus.anysupport.gtf > anysupport.genes

#filter so that you get a file containing unique names

cat anysupport.genes | uniq > anysupport.genes.uniq

#remove the " and ; from each name

sed -i 's/"//g' anysupport.genes.uniq

sed -i 's/;//g' anysupport.genes.uniq

#perl one liner that uses the id file you just created to pull out corresponding sequences in the fasta file produced by braker

perl -ne 'if(/^>(\S+)/){$c=$i{$1}}$c?print:chomp;$i{$\_}=1 if @ARGV' anysupport.genes.uniq augustus.hints.aa > augustus.hints.anysupport.aa

### BUSCO Assessment

busco -m proteins -i augustus.hints.aa -l metazoa\_odb10 -o berghia\_genome\_annotation\_Sept2021\_busco.proteins.v4.0.5 -c 10

busco -m proteins -i augustus.hints.aa -l mollusca\_odb10 -o berghia\_genome\_annotation\_Sept\_busco.mollusca.proteins.v4.0.5 -c 10

busco -m proteins -i augustus.hints.anysupport.aa -l metazoa\_odb10 -o berghia\_genome\_annotation\_Sept2021\_anysupport\_busco.proteins.v4.0.5

-c 10

busco -m proteins -i augustus.hints.anysupport.aa -l mollusca\_odb10 -o berghia\_genome\_annotation\_Sept2021\_anysupport\_busco.mollusca.proteins.v4.0.5 -c 10

# Functional Annotations

## BLASTP

Input(s): augustus.hints.anysupport.aa and blast databases

Output(s): Blast output files

*Protein-Protein BLAST 2.11.0+*

*genomeGTFtools (https://github.com/wrf/genomeGTFtools)*

### BLASTP runs

blastp -query augustus.hints.anysupport.aa -db uniprot\_sprot.fasta -num\_threads 8 -max\_target\_seqs 1 -outfmt 6 -evalue 1e-3 > blastp.uniprot-sprot.outfmt6

blastp -query augustus.hints.anysupport.aa -db refseq\_protein -num\_threads 8 -max\_target\_seqs 1 -outfmt 6 -evalue 1e-3 > blastp.refseq.outfmt6

### BLASTP output to genome annotation

blast2genomegff.py -b blastp.uniprot-sprot.outfmt6 -d uniprot\_sprot.fasta -g augustus.anysupport.gff3 -x -p blastp > berghia\_sprot.genome.gff

### Create BLASTP table with gene names for use in R

1. Downloaded blastp results to my computer
2. Pulled out UniProt and RefSeq IDs to search for metadata
   1. awk '{ print $2 }' blastp/blastp\_augustus\_RMiso.uniprot-sprot.outfmt6 | awk 'BEGIN { FS = "|" } ; { print $2 }' | sort | uniq > uniprot\_ids\_filtered.txt
   2. awk '{ print $2 }' blastp/blastp\_augustus\_RMiso.refseq.outfmt6 | sort | uniq > refseq\_ids\_filtered.txt
3. Used download\_uniprot.pl and download\_uniprot\_refseq.pl to download uniprot hits for IDs where uniprot metadata exists
   1. perl download\_uniprot.pl uniprot\_ids\_filtered.txt > uniprot\_data.txt
   2. perl download\_uniprot\_refseq.pl refseq\_ids\_filtered.txt > uniprot\_data\_refseq.txt
4. Run first part of R-script Bs\_genome\_blasthits.R to get the list of RefSeq annotations that are missing
5. Split the nouniprothits output file for batch entrez downloads
   1. split -l 2000 Bs\_protein-nouniprothits-info\_unique.txt Bs\_protein-nouniprothits-info\_unique.
6. Upload files to Batch Entrez (https://www.ncbi.nlm.nih.gov/sites/batchentrez) and download then combine summary text files
   1. cat protein\_result\_[A-Z].txt > protein\_result\_allRefSeq.txt
7. Process batch entrez summary files to get a useable table for refseq results
   1. bash process\_batch\_entrez.sh
8. Run second part of Bs\_genome\_blasthits.R to get the full blast list
   1. **Output file:** Bs\_protein-blasthit-ALL-info.txt

## InterProScan

Input(s): augustus.hints.anysupport.aa

Output(s): Interproscan annotation files - \*.gff3, \*.json, \*.xml, \*.tsv

*InterProScan v.5.52-86.0*

*CDD-3.18,Coils-2.2.1,Gene3D-4.3.0,Hamap-2020\_05,MobiDBLite-2.0,PANTHER-15.0,Pfam-33.1,Phobius-1.01,PIRSF-3.10,PIRSR-2021\_02,PRINTS-42.0,SFLD-4,SignalP\_EUK-4.1,SMART-7.1,SUPERFAMILY-1.75,TIGRFAM-15.0,TMHMM-2.0c*

*MAKER script - iprscan2gff3*

### InterProScan Run

interproscan.sh -i augustus.hints.anysupport.aa -b berghia\_RM\_iso\_anysupport\_ipscan\_2021\_11 -goterms -dp -t p -appl CDD-3.18,Coils-2.2.1,Gene3D-4.3.0,Hamap-2020\_05,MobiDBLite-2.0,PANTHER-15.0,Pfam-33.1,Phobius-1.01,PIRSF-3.10,PIRSR-2021\_02,PRINTS-42.0,SFLD-4,SignalP\_EUK-4.1,SMART-7.1,SUPERFAMILY-1.75,TIGRFAM-15.0,TMHMM-2.0c

### InterProScan output to genome annotation (from MAKER scripts)

iprscan2gff3 berghia\_RM\_iso\_anysupport\_ipscan\_2021\_11.tsv braker\_annotations.gff3 > braker\_iprscan\_annotations.gff