Bioinformatic pipeline for McCosker et al., “Molecular mechanisms underlying response to influenza in grey seals (*Halichoerus grypus*), a potential wild reservoir.”

Published in Molecular Ecology, 2025.

Access grey seal genome files: assembly (.fna) and annotation (.gtf).

```
ftp ftp.ncbi.nlm.nih.gov
cd genomes
cd refseq
cd vertebrate_mammalian
cd Halichoerus_grypus
cd GCF_012393455.1_Tufts_HGry_1.1
get GCF_012393455.1_Tufts_HGry_1.1_genomic.gtf.gz
get GCF_012393455.1_Tufts_HGry_1.1_genomic.fna.gz
quit

gunzip  GCF_012393455.1_Tufts_HGry_1.1_genomic.gtf.gz
gunzip GCF_012393455.1_Tufts_HGry_1.1_genomic.fna.gz
```

Concatenate RNAseq files for lanes 1-4 for each sample. Example code for 1 sample:

```
cat /home/kcammen/Reads/Hg266-14P_S10_L001_R1_001.fastq.gz
/home/kcammen/Reads/Hg266-14P_S10_L002_R1_001.fastq.gz
/home/kcammen/Reads/Hg266-14P_S10_L003_R1_001.fastq.gz
/home/kcammen/Reads/Hg266-14P_S10_L004_R1_001.fastq.gz > Hg266_R1.fastq.gz
```

Run FastQC on reads prior to trimming.

```
module load fastqc (v. 0.11.7)
for file in /home/cmccosker/RNAseq_Pilot/concat_reads/*.fastq.gz
do
fastqc $file --outdir=/home/cmccosker/RNAseq_Pilot/fastqc_prior_trim
done
```

Run MultiQC on pre-trim FastQC files.

Set up and access virtual environment for MultiQC.

```
module load anaconda3/5.2.0
. ~/conda.init
conda create --name cmccosker
conda activate cmccosker
```

Install MultiQC

```
pip install multiqc
```

Run MultiQC

```
module load anaconda3/5.2.0
. ~/conda.init
conda activate cmccosker

export LC_ALL=en_US.utf8
export LANG=en_US.utf8

/home/cmccosker/.local/bin/multiqc
/home/cmccosker/RNAseq_Pilot/fastqc_prior_trim/

conda deactivate
```

Trim adapters and low-quality sequence reads using Trimmomatic. Example code for 1 sample:

```
module load gcc (v. 5.2.0)
module load java (v. 14.0.1?)
module load trimmomatic (v. 0.36)

cd /home/cmccosker/RNAseq_Pilot/trimmed_reads

java -jar /opt/modules/universal/trimmomatic/0.36/trimmomatic-0.36.jar
PE /home/cmccosker/RNAseq_Pilot/concat_reads/Hg266_R1.fastq.gz
/home/cmccosker/RNAseq_Pilot/concat_reads/Hg266_R2.fastq.gz -baseout
Hg266.fastq
ILLUMINACLIP:/opt/modules/universal/trimmomatic/0.36/adapters/TruSeq3-
PE.fa:2:30:10 SLIDINGWINDOW:4:20 MINLEN:36
```

Run FastQC on reads after trimming.

```
for file in /home/cmccosker/RNAseq_Pilot/trimmed_reads/*P.fastq
do
  fastqc $file --outdir=/home/cmccosker/RNAseq_Pilot/fastqc_post_trim
done
for file in {path to fastq.gz reads}/*.fastq
do
  fastqc $file --outdir={path to & name of output directory}
done
```

Run MultiQC on post-trim FastQC files.

```
module load anaconda3/5.2.0
. ~/conda.init
conda activate cmccosker
```

```
export LC_ALL=en_US.utf8
export LANG=en_US.utf8

/home/cmccosker/.local/bin/multiqc
/home/cmccosker/RNAseq_Pilot/fastqc_post_trim/
```

de novo Transcriptome Assembly

Concatenate all left and right reads across samples.

cd /home/cmccosker/RNAseq_Pilot

mkdir trinity_input_reads

cat /home/cmccosker/RNAseq_Pilot/trimmed_reads/*_1P.fastq > /home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq

cat /home/cmccosker/RNAseq_Pilot/trimmed_reads/*_2P.fastq > /home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq

Check how many reads are in each file.

cd /home/cmccosker/RNAseq_Pilot/trinity_input_reads

wc -l all_R1_reads.fastq *[untrimmed reads]*
wc -l all_R2_reads.fastq *[untrimmed reads]*

wc -l all_R1_trim_reads.fastq
wc -l all_R2_trim_reads.fastq

Use Trinity to assemble a *de novo* transcriptome assembly.

cd /home/cmccosker/RNAseq_Pilot
mkdir trinity_denovo

module load trinity **(v. 2.2.0)**

Trinity --seqType fq --max_memory 80G --SS_lib_type RF --left /home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq --right /home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq --CPU 28 --output /home/cmccosker/RNAseq_Pilot/trinity_denovo

Run TrinityStats on trinity transcriptome assembly.

/opt/modules/centos7/trinity/2.2.0/util/TrinityStats.pl /home/cmccosker/RNAseq_Pilot/trinity_denovo/Trinity.fasta

Run transrate on trinity transcriptome assembly.

cd RNAseq_Pilot
mkdir transrate_denovo

module load transrate **(v. 1.0.3)**

transrate --assembly=/home/cmccosker/RNAseq_Pilot/trinity_denovo/Trinity.fasta --left=/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq --

right=/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq --output=/home/cmccosker/RNAseq_Pilot/transrate_denovo

Transrate indicated a low mapping rate, try Bowtie2 aligner to map reads to transcriptome assembly.

Build Bowtie2 index.

```
cd RNAseq_Pilot
mkdir bowtie2_denovo

module load Bowtie2 (v. 2.2.8, compiler gcc v. 4.1.2)

bowtie2-build -f /home/cmccosker/RNAseq_Pilot/trinity_denovo/Trinity.fasta
/home/cmccosker/RNAseq_Pilot/bowtie2_denovo/bowtie2_denovo_index
```

Align RNAseq to transcriptome with Bowtie2.

```
module load Bowtie2
module load perl

bowtie2 -x /home/cmccosker/RNAseq_Pilot/bowtie2_denovo/bowtie2_denovo_index -
1 /home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq -2
/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq -S
/home/cmccosker/RNAseq_Pilot/bowtie2_denovo/bowtie2_align_denovo
```

de novo Transcriptome Assembly w/ normalization

Run Trinity with in silico normalization

```
cd /RNAseq_Pilot
mkdir trinity_denovo_normalize

module load trinity (v. 2.2.0)

Trinity --normalize_reads --normalize_max_read_cov 200 --seqType fq --max_memory 80G --SS_lib_type RF –left /home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq --right /home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq --CPU 28 --output /home/cmccosker/RNAseq_Pilot/trinity_denovo_normalize
```

Run Trinity Stats on Trinity transcriptome assembly with normalized reads.

```
/opt/modules/centos7/trinity/2.2.0/util/TrinityStats.pl /home/cmccosker/RNAseq_Pilot/trinity_denovo_normalize/Trinity.fasta
```

Run transrate on trinity transcriptome assembly with normalized reads.

```
cd RNAseq_Pilot
mkdir transrate_denovo_normalize

module load transrate (v. 1.0.3)

transrate --assembly=/home/cmccosker/RNAseq_Pilot/trinity_denovo_normalize/Trinity.fasta --left=/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq --right=/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq --output=/home/cmccosker/RNAseq_Pilot/transrate_denovo_normalize
```

Transrate indicated a low mapping rate, try Bowtie2 aligner to map reads to transcriptome assembly.

Build Bowtie2 index.

```
cd RNAseq_Pilot
mkdir bowtie2_denovo_normalize

module load Bowtie2 (v. 2.2.8, compiler gcc v. 4.1.2)

bowtie2-build -f /home/cmccosker/RNAseq_Pilot/trinity_denovo_normalize/Trinity.fasta /home/cmccosker/RNAseq_Pilot/bowtie2_denovo_normalize/bowtie2_denovo_normalize_index
```

Align RNAseq to transcriptome with Bowtie2.

```
module load Bowtie2
module load perl

bowtie2 -x
/home/cmccosker/RNAseq_Pilot/bowtie2_denovo_normalize/bowtie2_denovo_normal
ize_index -1
/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq -2
/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq -S
/home/cmccosker/RNAseq_Pilot/bowtie2_denovo_normalize/bowtie2_align_denovo_n
ormalize.sam
```

Genome-guided Transcriptome Assembly

Create hisat2 genome index.

```
cd /home/RNAseq_Pilot/
mkdir trinity_genome-guided

cd trinity_genome-guided
mkdir hisat2-build
cd hisat2-build

module load hisat2 (v. 2.1.0)

hisat2-build -f -p 8
/home/cmccosker/RNAseq_Pilot/RefGenome/GCF_012393455.1_Tufts_HGry_1.1_genomic.fna /home/cmccosker/RNAseq_Pilot/trinity_genome-guided/hisat2-build/greyseal_hisat2_index
```

Align RNA reads to genome using hisat2 index files.

```
module load hisat2
module load perl

hisat2 --rna-strandness RF -x /home/cmccosker/RNAseq_Pilot/trinity_genome-guided/hisat2-build/greyseal_hisat2_index -1
/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq -2
/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq -S
/home/cmccosker/RNAseq_Pilot/trinity_genome-guided/hisat2_align_2.sam
```

Sort .sam file by coordinates.

```
module load samtools (v. 1.3.1)

samtools sort -o /home/cmccosker/RNAseq_Pilot/trinity_genome-guided/sorted_genome_alignment.bam
/home/cmccosker/RNAseq_Pilot/trinity_genome-guided/hisat2_align.sam
```

Use genome alignments to perform *de novo* transcriptome assembly.

```
cd RNAseq_Pilot/trinity_genome-guided
mkdir trinity_out_dir

module load trinity (v. 2.2.0)

Trinity --genome_guided_bam /home/cmccosker/RNAseq_Pilot/trinity_genome-guided/sorted_genome_alignment.bam --max_memory 80G --
genome_guided_max_intron 2500 --CPU 28 --SS_lib_type RF --output
/home/cmccosker/RNAseq_Pilot/trinity_genome-guided/trinity_out_dir
```

Run TrinityStats on genome-guided assembly.

```
/opt/modules/centos7/trinity/2.2.0/util/TrinityStats.pl
/home/cmccosker/RNAseq_Pilot/trinity_genome-guided/trinity_out_dir/Trinity-GG.fasta
```

Run transrate on genome-guided assembly.

```
cd RNAseq_Pilot
mkdir transrate_genome-guided

module load transrate (v. 1.0.3)

transrate --assembly=/home/cmccosker/RNAseq_Pilot/trinity_genome-guided/trinity_out_dir/Trinity-GG.fasta --left=/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq --right=/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq --output=/home/cmccosker/RNAseq_Pilot/transrate_genome-guided
```

Transrate had low mapping rate, try Bowtie2 aligner to map reads to transcriptome assembly.

Build Bowtie2 index.

```
cd RNAseq_Pilot
mkdir bowtie2_genome-guided

module load Bowtie2 (v. 2.2.8, compiler gcc v. 4.1.2)

bowtie2-build -f /home/cmccosker/RNAseq_Pilot/trinity_genome-guided/trinity_out_dir/Trinity-GG.fasta /home/cmccosker/RNAseq_Pilot/bowtie2_genome-guided/bowtie2_genome-guided_index
```

Align RNA sequences to genome-guided transcriptome.

```
module load Bowtie2
module load perl

bowtie2 -x /home/cmccosker/RNAseq_Pilot/bowtie2_genome-guided/bowtie2_genome-guided_index -1 /home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq -2 /home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq -S /home/cmccosker/RNAseq_Pilot/bowtie2_genome-guided/bowtie2_align_genome-guided.sam
```

Genome-guided assembly with max intron size 3000, and try max mem 100G

```
cd RNAseq_Pilot
```

mkdir trinity_genome-guided_2
module load trinity **(v. 2.2.0)**

Trinity
--genome_guided_bam /home/cmccosker/RNAseq_Pilot/trinity_genome-guided/sorted_genome_alignment.bam --max_memory 100G --genome_guided_max_intron 3000 --CPU 28 --SS_lib_type RF --output /home/cmccosker/RNAseq_Pilot/trinity_genome-guided_2

Run TrinityStats on genome-guided assembly with max intron size 3000.

/opt/modules/centos7/trinity/2.2.0/util/TrinityStats.pl /home/cmccosker/RNAseq_Pilot/trinity_genome-guided_2/Trinity-GG.fasta

Run transrate on genome-guided assembly with max intron size 3000.

cd RNAseq_Pilot
mkdir transrate_genome-guided_2

module load transrate **(v. 1.0.3)**

transrate --assembly=/home/cmccosker/RNAseq_Pilot/trinity_genome-guided_2/Trinity-GG.fasta --left=/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq --right=/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq --output=/home/cmccosker/RNAseq_Pilot/transrate_genome-guided_2

Bowtie2 alignment to genome-guided assembly with max intron size 3000.

Build Bowtie2 index.

cd RNAseq_Pilot
mkdir bowtie2_genome-guided_2

module load Bowtie2 **(v. 2.2.8, compiler gcc v. 4.1.2)**

bowtie2-build -f /home/cmccosker/RNAseq_Pilot/trinity_genome-guided_2/Trinity-GG.fasta /home/cmccosker/RNAseq_Pilot/bowtie2_genome-guided/bowtie2_genome-guided_index_2

Align reads to genome-guided transcriptome with max intron size 3000 using Bowtie2.

module load Bowtie2
module load perl

bowtie2 -x /home/cmccosker/RNAseq_Pilot/bowtie2_genome-guided_2/bowtie2_genome-guided_index_2 -1 /home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq -2

/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq -S /home/cmccosker/RNAseq_Pilot/bowtie2_genome-guided_2/bowtie2_align_genome-guided_2.sam

Merging *de novo* and Genome-Guided Assemblies

Convert fasta files for input to EvidentialGene to ensure unique IDs.

```
cd RNAseq_Pilot
mkdir evigene

module load evigene (v. 05-20-2020)
{mvapich2 v. 2.1; exonerate v. 2.2.0; cdhit v. 4.8.1; ncbi-blast v. 2.11.0}
module load perl

/opt/modules/centos7/evigene/2020-05-20/scripts/rnaseq/trformat.pl -out
/home/cmccosker/RNAseq_Pilot/evigene/txomes.fasta -in
/home/cmccosker/RNAseq_Pilot/evigene/*.fasta -log
```

Use EvidentialGene tr2aacds4 to merge assemblies and select best coding transcripts.

```
module load evigene (v. 05-20-2020)
{mvapich2 v. 2.1; exonerate v. 2.2.0; cdhit v. 4.8.1; ncbi-blast v. 2.11.0}
module load perl

tr2aacds4.pl -log -cdna
/home/cmccosker/RNAseq_Pilot/evigene_tr2aacds4/txomes.fasta -NCPU 16 -MAXMEM
32000
```

Run transrate on "okay" EvidentialGene transcriptome file.

```
cd /home/cmccosker/RNAseq_Pilot/

mkdir transrate_evigene_tr2aacds4

transrate --
assembly=/home/cmccosker/RNAseq_Pilot/evigene_tr2aacds4/txomes.okay.fasta --
left=/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq --
right=/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq --
output=/home/cmccosker/RNAseq_Pilot/transrate_evigene_tr2aacds4
```

Align reads using Bowtie2 to "okay" EvidentialGene transcriptome file.

Build Bowtie2 index.

```
cd /home/cmccosker/RNAseq_Pilot
mkdir bowtie2_evigene

module load Bowtie2 (v. 2.2.8, compiler gcc v. 4.1.2)

bowtie2-build -f /home/cmccosker/RNAseq_Pilot/evigene/txomes.okay.fasta
/home/cmccosker/RNAseq_Pilot/bowtie2_evigene/bowtie2_evigene_index
```

Align RNAseq reads to EvidentialGene transcriptome.

```
module load Bowtie2
module load perl

bowtie2 -x /home/cmccosker/RNAseq_Pilot/bowtie2_evigene/bowtie2_evigene_index -
1 /home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq -2
/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq -S
/home/cmccosker/RNAseq_Pilot/bowtie2_evigene/bowtie2_align_evigene.sam
```

Combine set of “okay” and “okalt” transcript sets from evigene_tr2aacds4

```
cat /home/cmccosker/RNAseq_Pilot/evigene_tr2aacds4/okayset/txomes.okay.tr
/home/cmccosker/RNAseq_Pilot/evigene_tr2aacds4/okayset/txomes.okalt.tr >
/home/cmccosker/RNAseq_Pilot/evigene_tr2aacds4/okay_okalt.fasta
```

Transrate on okay_okalt transcriptome

```
cd RNAseq_PIlot
mkdir transrate_evigene_okay_okalt

module load transrate

transrate --
assembly=/home/cmccosker/RNAseq_Pilot/evigene_tr2aacds4/okay_okalt.fasta --
left=/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq --
right=/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq --
output=/home/cmccosker/RNAseq_Pilot/transrate_evigene_okay_okalt
```

Align reads using Bowtie2 to okay_okalt transcriptome

Build Bowtie2 index

```
cd /home/cmccosker/RNAseq_Pilot
mkdir bowtie2_evigene_okay_okalt

module load Bowtie2 (v. 2.2.8, compiler gcc v. 4.1.2)

bowtie2-build -f /home/cmccosker/RNAseq_Pilot/evigene_tr2aacds4/okay_okalt.fasta
/home/cmccosker/RNAseq_Pilot/bowtie2_evigene_okay_okalt/bowtie2_evigene_okay_
okalt_index
```

Align reads

```
module load Bowtie2
module load perl
```

bowtie2 -x /home/cmccosker/RNAseq_Pilot/
bowtie2_evigene_okay_okalt/bowtie2_evigene_okay_okalt_index -1
/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq -2
/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq -S
/home/cmccosker/RNAseq_Pilot/bowtie2_evigene_okay_okalt/bowtie2_align_evigene_
okay_okalt.sam

Use EvidentialGene with *de novo* and genome-guided

Concatenate normalized + genome-guided

```
cd RNAseq_Pilot
mkdir evigene_norm-GG

cat /home/cmccosker/RNAseq_Pilot/trinity_denovo_normalize/Trinity.fasta
/home/cmccosker/RNAseq_Pilot/trinity_genome-guided/trinity_out_dir/Trinity-
GG.fasta > /home/cmccosker/RNAseq_Pilot/evigene_norm-GG/norm-GG.fasta
```

Run EviGene

```
cd /home/cmccosker/RNAseq_Pilot/evigene_norm-GG
module load evigene
module load perl

tr2aacds4.pl -log -cdna /home/cmccosker/RNAseq_Pilot/evigene_norm-GG/norm-
GG.fasta -NCPU 16 -MAXMEM 32000
```

Transrate

```
cd RNAseq_Pilot
mkdir transrate_evigene_norm-GG

module load transrate

transrate
--assembly=/home/cmccosker/RNAseq_Pilot/evigene_norm-GG/okayset/norm-
GG.okay.fasta --
left=/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq --
right=/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq --
output=/home/cmccosker/RNAseq_Pilot/transrate_evigene_norm-GG
```

Bowtie2

Build bowtie2 index

```
cd /home/cmccosker/RNAseq_Pilot
mkdir bowtie2_evigene_norm-GG
```

module load Bowtie2 **(v. 2.2.8, compiler gcc v. 4.1.2)**

bowtie2-build -f /home/cmccosker/RNAseq_Pilot/evigene_norm-GG/okayset/norm-GG.okay.fasta /home/cmccosker/RNAseq_Pilot/bowtie2_evigene_norm-GG/bt2_evigene_norm-GG_index

Bowtie2 Align

module load Bowtie2
module load perl

bowtie2 -x /home/cmccosker/RNAseq_Pilot/bowtie2_evigene_norm-GG/bt2_evigene_norm-GG_index -1 /home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq -2 /home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq -S /home/cmccosker/RNAseq_Pilot/bowtie2_evigene_norm-GG/bowtie2_align_evigene_norm-GG.sam

Use EvidentialGene with *de novo* and genome-guided

Concat *de novo* + genome-guided

cd RNAseq_Pilot
mkdir evigene_denovo-GG

cat /home/cmccosker/RNAseq_Pilot/trinity_denovo/Trinity.fasta /home/cmccosker/RNAseq_Pilot/trinity_genome-guided/trinity_out_dir/Trinity-GG.fasta > /home/cmccosker/RNAseq_Pilot/evigene_denovo-GG/denovo-GG.fasta

Run EviGene

cd /home/cmccosker/RNAseq_Pilot/evigene_denovo-GG/
module load evigene
module load perl

tr2aacds4.pl -log -cdna /home/cmccosker/RNAseq_Pilot/evigene_denovo-GG/denovo-GG.fasta -NCPU 16 -MAXMEM 32000

Transrate

cd RNAseq_Pilot
mkdir transrate_evigene_denovo-GG

module load transrate

transrate --assembly=/home/cmccosker/RNAseq_Pilot/evigene_denovo-GG/okayset/denovo-GG.okay.fasta --left=/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq --

right=/home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq --output=/home/cmccosker/RNAseq_Pilot/transrate_evigene_denovo-GG

Bowtie2

Build bowtie2 index

cd /home/cmccosker/RNAseq_Pilot
mkdir bowtie2_evigene_denovo-GG

module load Bowtie2 **(v. 2.2.8, compiler gcc v. 4.1.2)**

bowtie2-build -f /home/cmccosker/RNAseq_Pilot/evigene_denovo-GG/okayset/denovo-GG.okay.fasta /home/cmccosker/RNAseq_Pilot/bowtie2_evigene_denovo-GG/bt2_evigene_denovo-GG_index

Bowtie2 Align

module load Bowtie2
module load perl

bowtie2 -x /home/cmccosker/RNAseq_Pilot/bowtie2_evigene_denovo-GG/bt2_evigene_denovo-GG_index -1 /home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R1_trim_reads.fastq -2 /home/cmccosker/RNAseq_Pilot/trinity_input_reads/all_R2_trim_reads.fastq -S /home/cmccosker/RNAseq_Pilot/bowtie2_evigene_denovo-GG/bowtie2_align_evigene_denovo-GG.sam

*** evigene_denovo-GG performed better, used for futher analysis ***

Transcriptome Annotation

Copy transcriptome to new directory for annotation.

```
cd RNAseq_Pilot
mkdir annotate
cd annotate

cp /home/cmccosker/RNAseq_Pilot/evigene_denovo-GG/okayset/denovo-GG.okay.fasta /home/cmccosker/RNAseq_Pilot/annotate/txome.fasta
```

Compile BLAST database for all pinnipeds

Retrieve *genomic.fna and *genomic.gtf file for each species (retrieved 7/21/2021):

- **Gray seal** (RefSeq accession: GCF_012393455.1)
- **Harbor seal** (RefSeq accession: GCF_004348235.1)
- **Hawaiian monk seal** (RefSeq accession: GCF_002201575.1)
- **Southern elephant seal** (RefSeq accession: GCF_011800145.1)
- **Weddell seal** (RefSeq accession: GCF_000349705.1)
- **Northern fur seal** (RefSeq accession: GCF_003265705.1)
- **Steller sea lion** (RefSeq accession: GCF_004028035.1)
- **California sea lion** (RefSeq accession: GCF_009762305.2)
- **Pacific walrus** (RefSeq accession: GCF_000321225.1)

Extract transcript sequences for each species, done through Galaxy (https://usegalaxy.org/).

Input GTF feature file: GTF annotation files from NCBI
Filters: none
Filter by genomic region: No
Filter out transcripts with large introns: N/A
Replace reference sequence names: Nothing selected
Transcript merging: none
Reference Genome: From your history
Genome Reference Fasta: Gray seal genome file (GCF_012393455.1)
Reference based filters: none
Select fasta outputs: fasta file with spliced exons for each GFF transcript (-w exons.fa)
Feature File Output: none
Full GFF attribute preservation (all attributes are shown): No
Decode url encoded characters within attributes: No
Warn about duplicate transcript IDs and other potential problems with the given GFF/GTF records: No

Concatenate all exons.fasta files into a single .fasta file.

```
cat cali_genome_exons.fasta hawaiian_genome_exons.fasta Hg_genome_exons.fasta
nfur_genome_exons.fasta Pv_genome_exons.fasta selephant_genome_exons.fasta
steller_genome_exons.fasta walrus_genome_exons.fasta weddell_genome_exons.fasta
> pinniped.fasta
```

Make BLAST database using new pinniped.fasta file.

```
cd /home/cmccosker/annot_databases/pinnipedia

module load compbio

singularity run --bind $PWD:/compbio $COMPBIO_DIR/sif/blast_v_latest.sif
makeblastdb -in pinniped.fasta -dbtype nucl -out pinniped
```

BLAST entire transcriptome against pinniped database.

```
cd /home/cmccosker/RNAseq_Pilot/annotate

module load compbio

singularity run --bind $PWD:/compbio $COMPBIO_DIR/sif/blast_v_latest.sif blastn -db
/home/cmccosker/annot_databases/pinnipedia/pinniped -query txome.fasta -
max_target_seqs 1 -max_hsps 1 -evalue 0.001 -outfmt "10 std qcovs stitle" -out
txome_hg.csv
```

Retrieve transcript ID and gene ID for those that had % query coverage (QC) > 80. Paste into new file: *txome_transcript-gene_map.txt*.

Extract transcript IDs that were unannotated during blastn search by sorting results by % query coverage and pulling out transcript IDs with < 80% query coverage.

Find which transcript IDs are missing from pinniped blastn search.

Extract list of all transcript IDs from the transcriptome assembly.

```
cd RNAseq_Pilot/annotate

grep TR txome.fasta >> txome_transcriptIDs.txt
```

In Microsoft Excel: Put list of transcript IDs from pinniped blastn search in column A, list of all transcript IDs from transcriptome assembly in column B. Select list in column A, right click – Define Name, name it “Blast_List.” In column C, use formula: =ISNUMBER(MATCH(B1, Blast_list, 0)). Copy and past values for list of TRUE and FALSE, sort by TRUE/FALSE column, extract list of transcript IDs with “FALSE.” Add all transcripts with % query coverage < 80 to the list into file: *txome_pinniped_unan_transcriptID.txt*.

Trim IDs in .fasta file to get rid of extra text in sequence identifier to match transcript ID list.

```
sed -r 's/\len.+//' denovo_GG.okay.txt > denovo_GG.okay_trim.txt
```

In RStudio: Load required libraries.

```
library(Biostrings)
library(BSgenome)
```

In RStudio: Load entire transcriptome dataset into R with sequences in FASTA format.

```
denovo_GG_fasta <- readDNAStringSet("C:\\[path to file]\\denovo_GG.okay_trim.txt",
format = "fasta", seek.first.rec = TRUE, use.names = TRUE)
```

In RStudio: Read in transcript IDs

```
pinniped_miss_ID <- readLines("C:[path to
file]\\txome_pinniped_unan_transcriptID.txt")
```

In RStudio: Extract sequences

```
pinniped_miss_Seqs <- getSeq(x = denovo_GG_fasta, pinniped_miss_ID)
```

In RStudio: Write out sequences to FASTA file.

```
writeXStringSet(pinniped_miss_Seqs, "C:[path to
file]\\txome_pinnipedmiss_seqs.fasta")
```

Create swissprot database from NCBI (downloaded 7/2/2021, https://ftp.ncbi.nlm.nih.gov/blast/db/swissprot.tar.gz) and load onto computer cluster.

Unzip and extract files.

```
cd /home/cmccosker/annot_databases/swissprot/

gunzip swissprot.tar.gz

tar -xf swissprot.tar
```

Make blast database from swissprot file and load *txome_pinnipedmiss_seqs.fasta* to cluster.

```
cd /home/cmccosker/annotate

module load compbio

singularity run --bind $PWD:/compbio $COMPBIO_DIR/sif/blast_v_latest.sif
makeblastdb -in swissprot.tar -dbtype prot -parse_seqids -out swissprot
```

Conduct blastx search against swissprot database for unannotated transcripts.

```
cd RNAseq_Pilot/annotate

module load compbio

singularity run --bind $PWD:/compbio $COMPBIO_DIR/sif/blast_v_latest.sif blastx -db
/home/cmccosker/annot_databases/swissprot/swissprot -query
txome_pinnipedmiss_seqs.fasta -max_target_seqs 2 -max_hsps 1 -evalue 0.001 -
qcov_hsp_perc 80 -outfmt "10 std qcovs stitle" -out txome_pinnipedmiss_sp.csv
```

Parse through blastx results manually if manageable – use top and assign gene symbol to each transcript. Add transcript and annotation ID to transcript-to-gene map for newly annotated transcripts. Add lines for each unannotated transcript, too, using transcript ID for both gene_id and transcript_id columns.

Align Reads to Transcriptome and Quantify Expression

Prepare reference (transcriptome) index, using bowtie2 to align.

```
cd RNAseq_Pilot/RSEM
mkdir reference_index

module load compbio

cd /home/cmccosker

singularity run --bind $PWD:/compbio $COMPBIO_DIR/sif/rsem_v1_3_3.sif rsem-prepare-reference --transcript-to-gene-map /compbio/RNAseq_Pilot/RSEM/txome_transcript-gene_map.txt --bowtie2 /compbio/RNAseq_Pilot/RSEM/denovo_GG_okay_trim.fasta /compbio/RNAseq_Pilot/RSEM/reference_index/bt2_txome_index
```

Calculate expression for each sample. Example code:

```
module load compbio

singularity run --bind $PWD:/compbio $COMPBIO_DIR/sif/rsem_v1_3_3.sif rsem-calculate-expression --bowtie2 -p 8 --strandedness reverse --paired-end /home/cmccosker/RNAseq_Pilot/trimmed_reads/Hg11513*_1P.fastq /home/cmccosker/RNAseq_Pilot/trimmed_reads/Hg11513*_2P.fastq /home/cmccosker/RNAseq_Pilot/RSEM/reference_index/bt2_txome_index /home/cmccosker/RNAseq_Pilot/RSEM/Hg11513
```

Compile a single matrix of expected counts for each sample.

```
cd RNAseq_Pilot/RSEM

module load compbio

singularity run --bind $PWD:/compbio $COMPBIO_DIR/sif/rsem_v1_3_3.sif rsem-generate-data-matrix *.genes.results >txome_genes.matrix

singularity run --bind $PWD:/compbio $COMPBIO_DIR/sif/rsem_v1_3_3.sif rsem-generate-data-matrix *.isoforms.results >txome_isoforms.matrix
```

Remainder of analyses conducted in RStudio, see https://christinamccosker.github.io/hg-rna/

~Align to Genome & Quantify Expression~

BMB 502: Intro to Bioinformatics Group:
Christina McCosker, Jesus Zeno, Eleanor Glahn, Hannah Zenas, Tolu Adekeye
Performed in Galaxy: usegalaxy.org

Use FastQC (**v. 0.11.8**) to assess quality control of raw reads

Use MultiQC (**v. 1.9**) to compile & summarize FastQC reports for raw reads

Trimmomatic (**v. 0.38**) to trim adapter sequences.

Adapter sequences: TruSeq3
SLIDINGWINDOW: 4:20
MINLEN: 36

Always keep both reads: No

FastQC (**v. 0.11.8**) to assess quality control of trimmed reads

MultiQC (**v. 1.9**) to compile and summarize FastQC reports for trimmed reads

HISAT2 (**v. 2.1.0**) to align reads to the reference genome.

Reference Genome: gray seal genome (GCF_012393455.1)

Paired-end library
Input Read 1s
Input Read 2s
Strand Information: Reverse (RF)
Paired-end options: Use default values
Advanced
Spliced Alignment Options: Specify spliced alignment options
Transcriptome Assembly Reporting: Report alignments tailored to transcript assemblers including StringTie (--dta)

Samtools Sort (**v. 1.9**)

Input: BAM file from HISAT2
Sort Key: coordinate
Use default job resource parameters

StringTie (**v. 2.1.1**)

Input: Sorted BAM file
Input contains long reads?: No
Strand Information: Reverse
Use a reference to guide assembly?: Use Reference GTF/GFF3

Reference file: Use a file from history
GTF/GFF3 dataset to guide assembly: gray seal GFF annotation (GCF_012393455.1)
Use Reference Transcripts Only?: No
Output files for differential expression?: No additional output
Output coverage file?: No

StringTie Merge (**v. 2.1.1**)

Input: List of GTF files for all (n = 31) samples
Reference annotation to include in the merging: Hg Annotation GFF3 file
Galaxy default parameters:
Minimum input transcript length to include in the merge: 50
Minimum input transcript coverage to include in the merge: 0
Minimum input transcript FPKM to include in the merge: 1.0
Minimum input transcript TPM to include in the merge: 1.0
Minimum isoform fraction: 0.001
Gap between transcripts to merge together: 250
Keep merged transcripts with retained introns: No
**MSTRG.2549 in merged GTF file did not have strand (+/-) information
Galaxy suggested running in Unstranded mode, but MSTRG.2549 was only transcript w/o strand information
Straned Yes + Merged GTF file w/o MSTRG.2549 = 171,657 total counts
Stranded Reverse + Merged GTF w/o MSTRG.2549 = 7,979,658
Unstranded + Original Merged GTF (MSTRG.2549 included w/o strand info) = 7,974,348
Manually edited GTF file and tried running htseq-count on a sample (Hg1397) with MSTRG.2549 as + and -.
MSTRG.2549 + = 37
MSTRG.2549 = 1
**decided to use + and Reverse sranded mode

gffread to extract fasta sequences using StringTie merge GTF + Genome (**Cufflinks v. 2.2.1**)

Input GTF feature file: StringTie merge GTF with MSTRG.2549+
Filters: none
Filter by genomic region: No
Filter out transcripts with large introns: N/A
Replace reference sequence names: Nothing selected
Transcript merging: none
Reference Genome: From your history
Genome Reference Fasta: Gray seal genome file (GCF_012393455.1)
Reference based filters: none

Select fasta outputs: fasta file with spliced exons for each GFF transcript (-w exons.fa)
Feature File Output: none
Full GFF attribute preservation (all attributes are shown): No
Decode url encoded characters within attributes: No
Warn about duplicate transcript IDs and other potential problems with the given GFF/GTF records: No

In RStudio: Extract list of transcript IDs and gene IDs from gtf file

Load libraries

```
library(data.table)
library(rtracklayer)
library(Biostrings)
library(BSgenome)
```

Upload GTF file into RStudio

```
alltranscripts.gtf = "[path to file]\\StringTie_Merge_FinalAnnotation.gtf"
```

Create GRanges object

```
alltranscripts.gtf.gr = rtracklayer::import(alltranscripts.gtf)

alltranscripts.gtf.df = as.data.frame(alltranscripts.gtf.gr)
alltranscripts.genes.transcripts = unique(alltranscripts.gtf.df[ ,c("transcript_id", "gene_name")])

fwrite(alltranscripts.genes.transcripts, file="[path to file]\\alltranscripts_geneID_txptID_name.txt", sep="\t")
```

Open and save as excel document - GeneID column unnecessary at this point - focusing on transcript identification. Sort by gene name, extract list of unidentified transcript names.

Extract sequences for unannotated transcripts. ***Ensure transcript IDs and fasta IDs match *exactly***

Load entire FASTA file, with IDs trimmed

```
ST_fasta <- readDNAStringSet("[path to file]\\Transcript_Seqs_exons_trim.txt", format = "fasta", seek.first.rec = TRUE, use.names = TRUE)
```

Read in transcript IDs for unannotated transcripts

```
all_unannotated_transcriptID <- readLines("[path to file]\\all_unannotated_transcripts.txt")
```

Extract Sequences

```
all_unannotated_seqs <- getSeq(ST_fasta, all_unannotated_transcriptID)
```

Write out unannotated transcript sequences to FASTA file

```
writeXStringSet(all_unannotated_seqs, "[path to
file]\\gnome_unannotated_seqs.fasta")
```

Trim entire FASTA file to ensure FASTA IDs are ***exactly*** the same as transcript IDs.

```
sed -r 's/\ ge.+//' Transcript_Seqs_exons > Transcript_Seqs_exons_trim.txt
```

Conduct blastn search against pinniped database (created above) to try to identify unannotated sequences.

```
cd /home/cmccosker/RNAseq_Pilot_Genome/annotate

module load compbio

singularity run --bind $PWD:/compbio $COMPBIO_DIR/sif/blast_v_latest.sif blastn -db
/home/cmccosker/annot_databases/pinnipedia/pinniped -query
gnome_unannotated_seqs.fasta -max_target_seqs 1 -max_hsps 1 -evalue 0.001 -outfmt "10
std qcovs qcovhsp stitle" -out gnome_unannot_pinniped.csv
```

Extract transcript ID & gene ID for transcripts with >80% query coverage from blastn pinniped search. Use list of transcripts >80% QC as query list, list of OG unannoated transcripts (*all_unannotated_transcripts.txt*) and use =ISNUMBER(MATCH()) function. File saved as "*gnome_pinnipedmiss_transcripts.txt*"

Extract list of sequences for unannotated transcripts in R as above.

Conduct blastx search against swissprot database (created above) to identify unannotated sequences (*blastx_gnome_pinnipedmiss.slurm*)

```
cd /home/cmccosker/RNAseq_Pilot_Genome/annotate

module load compbio

singularity run --bind $PWD:/compbio $COMPBIO_DIR/sif/blast_v_latest.sif blastx -db
/home/cmccosker/annot_databases/swissprot/swissprot -query
gnome_unannotated_seqs.fasta -max_target_seqs 2 -max_hsps 1 -evalue 0.001 -
qcov_hsp_perc 80 -outfmt "10 std qcovs stitle" -out gnome_unannotated_sp.csv
```

Parse through blastx results manually to assign gene symbols to StringTie transcripts. Ensure there are no LOC identifiers assigned to these genes - use LOC identifier when possible to match with other genes already identified. Used =isnumber(match()) formula to check which gene

symbols were not already annotated and searched NCBI for those genes to see which gene symbol corresponds with the protein product. Add transcript ID and gene annotation to transcript-to-gene map (*gnome_transcript-gene_map.txt*). Add lines to transcript-to-gene map for unannotated transcript - use transcript ID for both gene_id and transcript_id column.

~RSEM - Align to Genome & Quantify~

Prepare reference (genome-transcripts) index using Bowtie2 to align

```
cd /home/cmccosker/

module load compbio

singularity run --bind $PWD:/compbio $COMPBIO_DIR/sif/rsem_v1_3_3.sif rsem-prepare-reference --transcript-to-gene-map /compbio/RNAseq_Pilot_Genome/RSEM/gnome_transcript-gene_map.txt --bowtie2 /compbio/RNAseq_Pilot_Genome/RSEM/Transcript_Seqs_exons_trim.fasta /compbio/RNAseq_Pilot_Genome/RSEM/reference_index/bt2_gnome_index
```

Calculate transcript expression for each sample, example code:

```
module load compbio

singularity run --bind $PWD:/compbio $COMPBIO_DIR/sif/rsem_v1_3_3.sif rsem-calculate-expression --bowtie2 -p 8 --strandedness reverse --paired-end /home/cmccosker/RNAseq_Pilot/trimmed_reads/Hg11513*_1P.fastq /home/cmccosker/RNAseq_Pilot/trimmed_reads/Hg11513*_2P.fastq /home/cmccosker/RNAseq_Pilot_Genome/RSEM/reference_index/bt2_gnome_index /home/cmccosker/RNAseq_Pilot_Genome/RSEM/Hg11513
```

Compile a single matrix of expected counts for each sample.

```
module load compbio

singularity run --bind $PWD:/compbio $COMPBIO_DIR/sif/rsem_v1_3_3.sif rsem-generate-data-matrix  /home/cmccosker/RNAseq_Pilot_Genome/RSEM/*.genes.results >gnome_genes.matrix

singularity run --bind $PWD:/compbio $COMPBIO_DIR/sif/rsem_v1_3_3.sif rsem-generate-data-matrix /home/cmccosker/RNAseq_Pilot_Genome/RSEM/*.isoforms.results >gnome_isoforms.matrix
```

Remainder of analyses conducted in RStudio, see https://christinamccosker.github.io/hg-rna/