Appendix 3. Transcript annotation and seasonal gene expression in kidney tissue of the Australian pygmy bluetongue skink, *Tiliqua adelaidensis*

Carmel Maher

2022

Contents

1   Analysis Workflow

Transcript annotation and seasonal gene expression in kidney tissue of the Australian pygmy bluetongue skink, *Tiliqua adelaidensis*

This document outlines the pipeline methods for assembly of poly-a selected transcripts extracted from kidney tissues of three male and five female Australian pygmy bluetongue skinks collected in South Australia during spring (September) and autumn (March or April). Long read transcripts were sequenced from one individual "G6" using Pacific Biosciences' IsoSeq Sequel platform, and short reads for all eight individuals were sequenced on an Illumina HiSeq. The wet-lab library preparation up to sequencing is outlined in Chapter 3.

Sample ID	Experimental factors
G1	September Female
G2	September Female
G3	March/April Female
G4	March/April Male
G5	September Male
G6	September Male
G7	March/April Female
G8	March/April Female

There are three main sections to this document, which are referenced in three separate chapters of the thesis:

Assembly (Chapter 4): Assembly, cleaning and clustering of transcript isoforms from **both** sequencing formats to form a pseudo-reference for *T. adelaidensis* and to compare these methods.

Annotation (Chapter 5): Annotation and further analysis of the full length transcript isoforms derived from the long read analysis.

Gene expression (Chapter 6): Gene expression analysis using the short read data aligned to the long read pseudo-reference produced in Chapter 4, to assess seasonal variation of gene expression.

Analysis was done using a combination of computing and HPC machines with specifications as below:

- Long read cleaning, read frame prediction and blast searches were run on the Australian National eResearch Collaboration Tools and Resources project (NECTAR) cloud, using an 'm1.xlarge flavour' allocation with 8 AMD Opteron 63xx class CPUs, 32GB RAM, 10GB root disk, and 240GB of secondary disk storage.
- The ANGEL program which required a higher memory allocation was run on a Dell PowerEdge server with 40 cores and 512G RAM.
- Short read QC, trimming, initial trinity assembly, and alignments were completed using Flinders University's high-performance computer "Deep thought" with 16 x 256Gb (4TB RAM), 1024 AMD x86 CPU cores at 2.0GHz across 14 standard compute and 2 data science nodes, and 100 TiB of usable storage via the Dell EMC Network Storage reference architecture.
- Downstream gene expression analysis and manual manipulation of annotation files was completed on a local Intel core i7 machine using RStudio or in-house bash scripts applied in Git Bash for Windows.

All larger scripts included in this document have also been uploaded to Github separately and are linked here at the relevant steps.
All scripts and commands in this pipeline depend on specific directory structure, and may show inconsistencies in this structure between steps where files have been transferred between computing resources. The input and output files of note are highlighted in the text between steps to bridge these gaps.

2 Assembly (Chapter 4)

A genome from a closely related species was unavailable for these data. Gene expression analysis of short read data was initially intended to be carried out using a *de-novo* assembly of this same data as a pseudo-reference for expression counts.

When long read data became available, both datasets have been assembled *de-novo* and these assemblies compared. The long read assembly then provided the reference for expression counts in Chapter 6.

This section is divided into the two methods of assembly; using the long read IsoSeq data, and using the short read HiSeq data. Both these sections contain titles referencing the main program in use at each step.

2.1 *Long read data*

De-novo assembly of long-read IsoSeq data sequenced from poly-a selected RNA extracted from a single kidney from september male G6. These sequences were be used to annotate genes found in kidney tissue of T. adelaidensis and create a set of reference transcripts for gene expression analysis conducted in later chapters.

Sequences were downloaded from Pacific Biosciences & checksumms verified.

2.1.1 SMRT Tools - IsoSeq3

SMRT Tools provides various programs for appropriate treatment of Pacific Biosciences sequencing outputs. IsoSeq3 was used to collapse circular sequences, initial polishing, and to characterise full length transcript reads. This analysis was conducted in an environment on the eRSA NECTAR research cloud.

A primer reference file was created manually for reference in the Isoseq3 script using the primers listed in the Clontech SMARTer cDNA Library prep kit which was used during the synthesis of cDNA for these samples.

primers.fasta:

```
>primer_5p
AAGCAGTGGTATCAACGCAGAGTACATGGG
>primer_3p
GTACTCTGCGTTGATACCACTGCTT
```

A single script was used to complete the Pacific Biosciences Isoseq3 pipeline isoseq3.sh:

```bash
#!/bin/bash
#
# This script is used to characterise full length transcripts de novo using
↪    smrttools and bioconda as specified by PacBio workflow
# see: https://github.com/PacificBiosciences/IsoSeq3/blob/master/README_v3.0.md and
↪    follow links for bioconda references
#
# The paths in the script assume that a specific directory structure has been set
↪    up.
# This directory structure is for the Nectar cloud VM

# open a screen
# usage from /mnt
```

```bash
# bash isoseq3.sh <ctrl-a ctrl-d>
#
# Carmel Maher
# April 2019


#---------Installed Programs---------

# smrttools                 -> smrttools-release_6.0.0.47835
  # help:
  #
  ↪  /pacbio/smrtlink/install/smrtlink-release_6.0.0.47841/bundles/smrttools/install/
  ↪  smrttools-release_6.0.0.47835/smrtcmds/bin/isoseq3 -h
  #
  ↪  /pacbio/smrtlink/install/smrtlink-release_6.0.0.47841/bundles/smrttools/install/
  ↪  smrttools-release_6.0.0.47835/smrtcmds/bin/isoseq3 --version
  # Version: isoseq3 3.0.0 (commit v3.0.0-7-gcc6cddd)

# python 2.7          -> to run miniconda - automatically installed on nectar
# miniconda           -> recommended python 2.7 compatible version by PacBio
↪  workflow. Miniconda2 | VER: 4.6.14
# ccs                      -> installed through conda, inside pbccs package | VER:
↪  pbccs-3.4.1
# lima                     -> installed through conda | VER: lima-1.9.0


#------adjust these for your run------

#as per the local installation, the smrttools program directories are (in /mnt).
↪   Notable directories listed in installation:
SMRT_ROOT=/mnt/pacbio/smrtlink

# Isoseq3 executeable location (from /mnt)
Iso3_DIR=$SMRT_ROOT/install/smrtlink-release_6.0.0.47841/bundles/smrttools/install/
↪   smrttools-release_6.0.0.47835/smrtcmds/bin

DATA=/mnt/data
MOVIE="MAH6260A1_m54196_190204_223227"
THREADS=0     #0 for the Iseseq3 tools =autodetection of threads and is the default


#-----------------------------------

function error_exit
{
    # Exit function due to fatal error
    # Accepts 1 arg:
    # string - descriptive error message

    echo "${PROGNAME}: ${1:-"Unknown error"}" 1>&2
    exit 1
}
```

```
# go to the working data directory
cd $DATA || error_exit "$LINENO: directory error"

# Generate consensus sequnces (ccs file) from raw subread data
# NOTE: this automatically runs with default autodetection of threads = 8
ccs $MOVIE.subreads.bam $MOVIE.ccs.bam --noPolish --minPasses 1 || error_exit
↪   "$LINENO: consensus sequence ccs error"

echo " - - - consensus sequnces generated"

# The primers.fasta as per the recommended Clontech SMARTer cDNA library prep - no
↪   barcodes used for this sample
# primers.fasta has been uploaded to /mnt/data and does not need to be created
↪   here:
      # primers.fasta
      # >primer_5p
      # AAGCAGTGGTATCAACGCAGAGTACATGGG
      # >primer_3p
      # GTACTCTGCGTTGATACCACTGCTT

# Remove primers and demultiplex:
lima --isoseq --dump-clips --no-pbi $MOVIE.ccs.bam primers.fasta demux.bam ||
↪   error_exit "$LINENO: Primer demultiplex lima error"

echo " - - - primers removed and demultiplexed"

                  #!#!#!#!#
                  # CHECK #  Note: A search using Git bash on the output files for any
                  ↪   of the primers listed for Isoseq by Pacific biosciences returns
                  ↪   no results after this script
                  #!#!#!#!#

# *********************
# From here on, execute the following steps for each output BAM file. -- only one
↪   in this case.
# isoseq3 tool usage: isoseq3 <tool>
# due to nectar permissions and installation location, usage: $Iso3_DIR/isoseq3
↪   <tool>
# *********************

### cluster Tool: Cluster CCS reads and generate unpolished transcripts.
# recommended to give this step as many cores as possible
# Usage
# isoseq3 cluster [options] input output
# Example
# isoseq3 cluster <two types of potential inputfile> demux.bam <OR>
↪   $MOVIE.consensusreadset.xml  unpolished.bam

# Cluster consensus sequences to generate unpolished transcripts:
# note the demux.bam is named after the headers used in primers.fasta
$Iso3_DIR/isoseq3 cluster --verbose -j $THREADS demux.primer_5p--primer_3p.bam
↪   unpolished.bam || error_exit "$LINENO: Isoseq3 cluster error"
```

```
    echo " - - - css reads clustered"

    ### polish Tool: Polish transcripts using subreads.
    # Usage
    # isoseq3 polish [options] input_1 input_2 output

    # Polish transcripts using subreads:
    $Iso3_DIR/isoseq3 polish --verbose -j $THREADS unpolished.bam $MOVIE.subreads.bam
    ↪ polished.bam || error_exit "$LINENO: Isoseq3 polish error"

    echo " - - - transcripts polished"

    ### summarize Tool: Create a .csv-format barcode overview from transcripts.
    # Usage
    # isoseq3 summarize [options] input output
    $Iso3_DIR/isoseq3 summarize --verbose polished.bam summary.csv || error_exit
    ↪ "$LINENO: Isoseq3 summary error"

    echo " - - - summary file created"
    echo " - - - done"
```

All of the above output was moved into a folder ~/data/2019-04-30_isoseq3_out

The output from this script, the polished.hq.fasta.gz file was unzipped and then used as the input for the program COGENT.

2.1.2 Cogent

Transcript clustering and collapse of redundant isoforms. Line by line instructions for running Cogent for clustering of transcript isoforms in a screen on the NECTAR machine as below. Includes instructions for Cupcake and minimap to collapse redundant transcript isoforms.

Some notes and commentary are based on personal communication with Tessa Bradford & Terry Bertozzi at the SA Museum, as well as the Cogent GitHub Tutorial page

Environments were used for installations (most often achieved with mini conda) on the NEXTAR machine. Activate screen, activate the environment, make Cogent output directory, and go to analysis working directory

```
screen -s cogent
conda activate anaCogent
mkdir /mnt/IsoSeq-analysis/data/Cogent
cd /mnt/IsoSeq-analysis/data/Cogent
```

From the analysis directory the input data directory is ../2019-04-30_out/polished.hq.fasta. The first output will be in this original data directory. Check paths (During installation, all paths were put into .profile)

2.1.2.0.1 Family Finding Running Family Finding for a small dataset

Create a k-mer profile of the input and calculate pairwise distances
Note: default K-mer sixe =30

```
python /mnt/Prog/Cogent/Cogent/run_mash.py --cpus=7
↪  /mnt/IsoSeq-analysis/data/2019-04-30_out/polished.hq.fasta
```

The output will be <fasta_filename>.k<sketch_size>.dist
Output written to: /mnt/IsoSeq-analysis/data/2019-04-30_out/polished.hq.fasta.s1000k30.dist

Process the distance file and create the partitions for each gene family

```
process_kmer_to_graph.py ../2019-04-30_out/polished.hq.fasta
↪  ../2019-04-30_out/polished.hq.fasta.s1000k30.dist ./hq hq
```

This generates an output log file hq.partition.txt and for each partition (isoform set), a subdirectory called hq/<partition_number> which contains the subset of fasta sequences belonging to that isoform set. Note that sequences that don't belong to any partition (ones with no similarity with other sequences) will be "unassigned" and noted in the partition log file.

2.1.2.0.2 Coding Genome Reconstruction Each isoform set family must be reconstructed individually for coding region.
Reconstructed contigs will contain the whole coding region. Reconstructed file is called cogent2.renamed.fasta

The script for an individual folder is:

```
reconstruct_contig.py -S T.adelaidensis hq/hq_0
```

The following script was written to loop this procedure and process hundreds of isoform set folders with sequences requiring coding region identification reconstructContig_1.sh:

```bash
#!/bin/bash
#
## this script is for coding region reconstruction of Cogent gene families as these
↪   must be completed individually.
# This script assumes you are following Running-Cogent.txt and using the same
↪   directory structure
# This script to be used after the process_kmer_to_graph.py step

# script in /mnt/IsoSeq-analysis/src
# usage from /mnt/IsoSeq-analysis/data/Cogent/hq
## bash ../../../src/reconstructContig.sh

# Carmel Maher
# August 2019

#Terry's error exit
function error_exit
{
    # Exit function due to fatal error
    # Accepts 1 arg:
    # string - descriptive error message
    echo "${PROGNAME}: ${1:-"Unknown error"}" 1>&2
    exit 1
}

cd ~/mnt/IsoSeq-analysis/data/Cogent/hq

for d in */ ; do

  DIR=${d%*}
  echo $DIR

  reconstruct_contig.py -S T.adelaidensis ./$DIR || error_exit "$LINENO: Error
  ↪   "$DIR""

done
echo compete
```

An error message prompted further checking of folder completion as below

Number of gene families:

```
ls ./hq/*/cogent2.renamed.fasta | wc -l
4454
```

Finding failed reconstruction jobs (pers. comm. Terry Bertozzi, SA Museum)

Number of hq gene folders:

```
ls hq | wc -l #number of folders
4459
```

Number of /hq folders completed:

```
ls ./hq/*/COGENT.DONE | wc -l
4454
```

This matches the total gene family count so all outputs have been recorded correctly, but does not match the total /hq folder count = 5 folders not completed.

The following will list the folders that have not finished:

```
comm -3 <(find /mnt/IsoSeq-analysis/data/Cogent/hq -iname 'COGENT.DONE' -printf
 ↪  '%h\n' | sort -u) <(find /mnt/IsoSeq-analysis/data/Cogent/hq -maxdepth 1
 ↪  -mindepth 1 -type d | sort) | sed -e 's/^.*hq\///'

hq_2605
hq_2738
hq_320
hq_6391
hq_9030
```

Some testing was required to create a modified script with a loop for failed folders, in order to include 5 folders which failed to run with kmer length 30 informed by this issue.
Due to how the loop functions inputting kmer size of 27 on execution will actually run all folders at the desired kmer size 30.

```
bash /mnt/IsoSeq-analysis/src/TBreconstructContig-edit.sh
 ↪  /mnt/IsoSeq-analysis/data/Cogent/hq 27 T.adelaidensis
```

The modified script TBreconstructContig-edit.sh:

```
#!/bin/bash

# This script runs reconstruct_contig.py on every subdirectory contained in the top
 ↪  directory. reruns with a kmer parameter of +3 (to a max of 99) for failed
 ↪  directories.
# This script assumes directories and files exist as up until the Coding Genome
 ↪  Reconstruction step in the Running Cogent GitHub wiki
#
# usage:  bash <path/script> <full directory containing hq_folders> <starting kmer
 ↪  value> <Species Name>
# e.g.        bash /mnt/IsoSeq-analysis/src/TBreconstructContig-edit.sh
 ↪  /mnt/IsoSeq-analysis/data/Cogent/hq 27 T.adelaidensis

# DIRNAME=$1
# species=$3

kmerSize=$2

find $1 -mindepth 1 -maxdepth 1 -type d | while read line; do #lists off all the
 ↪  /hq_* folders within the input directory and passes them through the script one
 ↪  by one (by line)
  FILE=$line/cogent2.fa
```

```
    if [ ! -f $FILE ]; then
        echo "failed no cogent2.fa Increase K-mer size" #Note all jobs will "fail" on
        ↪    first attempt unless you are running this in a previously reconstructed
        ↪    directory. In the latter case only the failed directories will continue
        ↪    to rerun.
        while [ ! -f "$FILE" ];do
            kmerSize=$((kmerSize + 3))
            rerunCMD="reconstruct_contig.py -k ${kmerSize} -S $3 $line"
            ↪    #reconstruction script increasing kmer parameter by 3 -note because
            ↪    all directories fail initially the input kmer value should be 3 below
            ↪    the desired actual value.
            echo ${rerunCMD}
            eval ${rerunCMD}
            if [ -f $FILE ]; then
                touch $1/hq_kmerSize.txt #This text file will give a list of the used
                ↪    kmerSize per hq_* folder
                echo "SUCCESS! The increased Kmer size $kmerSize was successful for
                ↪    $line." >> $1/hq_kmerSize.txt
            fi
            if [ ${kmerSize} -gt 99 ];then
                cp $line/in.fa $line/cogent2.fa # if a folder reruns until 99kmer
                ↪    parameter length is reached the input is simply copied into
                ↪    cogent2.fa
                touch $1/failed-jobs.txt #This text file will remain empty or give a
                ↪    list of all failed folders
                echo "Failed to succeed reconstruction with largest Kmer on ${line}.
                ↪    Copied input transcripts to cogent2.fa as output." >>
                ↪    $1/failed-jobs.txt
            fi
        done
        kmerSize=$2
    fi
done

if [ -s $1/failed-jobs.txt ]; then
    echo "failed-jobs.txt not empty. Some jobs failed. Please re-run them."
else
    echo "failed-jobs.txt empty or doesnt exist. All jobs completed."
fi
```

Re-check output.

Number of completed gene families:

```
ls ./hq/*/COGENT.DONE | wc -l
4459
```

OR:

```
ls ./hq/*/cogent2.renamed.fasta | wc -l
4459
```

Finding failed reconstruction jobs.

Number of items in ./hq:

```
ls hq | wc -l
4460
```

The added 1 is the text file reporting the successful kmer size for each "gene" family folder. Therefore all have completed.

List the names of folders that have not finished:

```
comm -3 <(find /mnt/IsoSeq-analysis/data/Cogent/hq -iname 'COGENT.DONE' -printf
↪   '%h\n' | sort -u) <(find /mnt/IsoSeq-analysis/data/Cogent/hq -maxdepth 1
↪   -mindepth 1 -type d | sort) | sed -e 's/^.*hq\///'
```

(this command listed no folders, indicating the script has completed all folders successfully)

See failed-jobs.txt and/or hq_kmerSize.txt. Note: all completed at kmer size 30 this time, and filed-jobs.txt does not exist as none failed.

2.1.2.0.3   Using Cogent to collapse redundant transcripts in absence of genome   Creating the "fake genome"

List and number of unassigned sequences
Make sure you are in the correct Cogent directory and can see the file hq.partition.txt:

```
tail -n 1 hq.partition.txt |tr ',' '\n' > unassigned.list
```

Create the unassigned.list file in the same directory:

```
tail -n 1 hq.partition.txt |tr ',' '\n' | wc -l
3193
```

Note: this number did not change with the addition of the five failed reconstruction folders. i.e. they failed completely and did not end up in the unassigned dataset.

The unassigned hq sequences were made into a fasta file.

```
export PATH=$PATH:/mnt/Prog/cDNA_Cupcake/sequence
get_seqs_from_list.py ../2019-04-30_out/polished.hq.fasta unassigned.list >
↪   unassigned.fasta
```

Unassigned sequences were concatenated with the Cogent reconstructed contigs by putting the reconstructed genes plus the unassigned single hq isoforms into a single fasta file:

```
mkdir collected
cd collected
cat ../hq/*/cogent2.renamed.fasta ../unassigned.fasta > cogent_fake_genome.fasta
```

Redundant isoforms were collapsed:

Create a SAM alignment with minimap2 (This could also be done with GMAP).
Obtain a final set of unique (non-redundant) transcript Isoforms that can be used as a reference gene set.
As there is natural 5' degradation in RNA some sequences will represent identical isoforms which may not all be identified in clustering.
These parameters are default or suggested in the Cogent tutorial page.

Create an aligned, sorted SAM file:

```
export PATH=$PATH:/mnt/Prog/minimap2
minimap2 -ax splice -t 30  -uf --secondary=no cogent_fake_genome.fasta
↪ ../../2019-04-30_out/polished.hq.fasta > hq.fasta.sam
```

Output:

```
[M::mm_idx_gen::2.392*1.00] collected minimizers
[M::mm_idx_gen::3.526*1.32] sorted minimizers
[M::main::3.528*1.32] loaded/built the index for 9960 target sequence(s)
[M::mm_mapopt_update::3.708*1.30] mid_occ = 31
[M::mm_idx_stat] kmer size: 15; skip: 5; is_hpc: 0; #seq: 9960
[M::mm_idx_stat::3.810*1.29] distinct minimizers: 6267388 (82.12% are singletons);
↪ average occurrences: 1.346; average spacing: 2.846
[M::worker_pipeline::13.901*5.65] mapped 25117 sequences
[M::main] Version: 2.11-r797
[M::main] CMD:
↪ /mnt/Prog/pacbio/smrtlink/install/smrtlink-release_6.0.0.47841/bundles/smrttools/
↪ install/smrttools-release_6.0.0.47835/private/thirdparty/minimap2/minimap2_2.11/binwrap/../../../..
↪ private/thirdparty/minimap2/minimap2_2.11/bin/minimap2 -ax splice -t 30 -uf
↪ --secondary=no cogent_fake_genome.fasta ../../2019-04-30_out/polished.hq.fasta
[M::main] Real time: 13.942 sec; CPU: 78.558 sec
```

Then follow the collapse tutorial from Cupcake
There is another cupcake page with information but the example scripts are provided on the cogent tutorial page.

Sort the SAM file:

```
sort -k 3,3 -k 4,4n hq.fasta.sam > hq.fasta.sorted.sam
```

Colapse identical isoforms to obtain a list of full length, unique, hq isoforms to use as reference transcripts:

```
cd /mnt/Prog/cDNA_Cupcake/sequence
which collapse_isoforms_by_sam.py
/home/ubuntu/miniconda2/envs/anaCogent/bin/collapse_isoforms_by_sam.py

cd /mnt/IsoSeq-analysis/data/Cogent/collected
```

Usage is:

```
# usage: collapse_isoforms_by_sam.py [-h] [--input INPUT] [--fq] -s SAM -o
#                                    PREFIX [-c MIN_ALN_COVERAGE]
```

```
#                                  [-i MIN_ALN_IDENTITY]
#                                  [--max_fuzzy_junction MAX_FUZZY_JUNCTION]
#                                  [--flnc_coverage FLNC_COVERAGE]
#                                  [--dun-merge-5-shorter]
```

Collapse:

```
collapse_isoforms_by_sam.py --input ../../2019-04-30_out/polished.hq.fasta -s
↪   hq.fasta.sorted.sam -c 0.94 -i 0.85 --dun-merge-5-shorter -o hq.fasta.no5merge
```

These parameters taken from pers. comm. Tessa Bradford, SA Museum.
Output is the name 'stem' the collapsed.rep.fa is for annotation.
The output files are <-o>.collapsed.gff, <-o>.collapsed.rep.fq, and <-o>.collapsed.group.txt i.e. hq.fasta.no5merge.collapsed.g
hq.fasta.no5merge.collapsed.rep.fq, and hq.fasta.no5merge.collapsed.group.txt

The naming system for the post-collapse isoform is PB.<loci_index>.<isoform_index>

More terminal output from the above has been excluded. This is an example:

```
Ignored IDs written to: hq.fasta.no5merge.ignored_ids.txt
Output written to:
hq.fasta.no5merge.collapsed.gff
hq.fasta.no5merge.collapsed.group.txt
hq.fasta.no5merge.collapsed.rep.fa
Namespace(allow_extra_5exon=False, flnc_coverage=-1, fq=False,
↪   input='../../2019-04-30_out/polished.hq.fasta', max_3_diff=100,
↪   max_5_diff=1000, max_fuzzy_junction=5, min_aln_coverage=0.94,
↪   min_aln_identity=0.85, prefix='hq.fasta.no5merge', sam='hq.fasta.sorted.sam')
```

Check outputs:

```
ls

cogent_fake_genome.fasta                      hq.fasta.no5merge.collapsed.rep.fa
hq.fasta.no5merge.collapsed.gff               hq.fasta.no5merge.ignored_ids.txt
hq.fasta.no5merge.collapsed.gff.unfuzzy       hq.fasta.sam
hq.fasta.no5merge.collapsed.group.txt         hq.fasta.sorted.sam
hq.fasta.no5merge.collapsed.group.txt.unfuzzy
```

The file hq.fasta.no5merge.collapsed.group.txt names the isoforms as PB.<loci_index>.<isoform_index>
and lists the collapsed identical isoforms.
Each 'locus' consists of a strand-specific locus with isoforms that overlap by at least 1 bp. So PB.11.1 and
PB.11.2 means this locus has two isoforms.

Count the 'loci' (transcripts) found:

```
wc -l hq.fasta.no5merge.collapsed.group.txt
15729 hq.fasta.no5merge.collapsed.group.txt
```

Write file statistics:

```
get_abundance_post_collapse.py hq.fasta.no5merge.collapsed
↪  /mnt/IsoSeq-analysis/data/2019-04-30_out/polished.cluster_report.csv

WARNING: isoseq3 format detected. Output `length` column will be `NA`.

Read stat file written to hq.fasta.no5merge.collapsed.read_stat.txt
Abundance file written to hq.fasta.no5merge.collapsed.abundance.txt
```

Note: this is true, hq.fasta.no5merge.collapsed.read_stat.txt has no length information and a whole column of NA values, however the headers of hq.fasta.no5merge.collapsed.rep.fa do have length information. This is how IsoSeq3 format data is processed and does not affect other aspects of this script.

Get count information from the abundance and group .txt files

Add the headers:

```
echo -e "pbid\tcount_fl" > output.collapsed.abundance.txt
```

Add the PB.loci_index.isoform_index information and number of isoforms:

```
paste <(awk '{print $1}' hq.fasta.no5merge.collapsed.group.txt) <(awk -F , '{print
↪  NF}' hq.fasta.no5merge.collapsed.group.txt) >> output.collapsed.abundance.txt
```

All these commands except reconstructContig_1.sh and TBreconstructContig.edit.sh are provided in the one document Running-Cogent-2.sh on Github.

The longest transcript for each set of isoforms is in hq.fasta.no5merge.collapsed.rep.fa. This file is used in the following step and also provides the full transcripts to be sub-set as a reference based on further clustering in Chapter 5.

2.1.3 ANGEL

ANGEL: Robust Open Reading Frame prediction

Line by line instructions for running ANGEL in order to predict open reading frame and peptide sequences of transcript Isoforms. Also outputs 5' and 3' untranslated regions and scores transcript completeness.

This analysis was tested on an environment on the eRSA NECTAR research cloud, final data run was completed by Terry Bertozzi on a Dell PowerEdge server with 40 cores and 512G RAM due to memory limitations of the NECTAR allocation.

All steps and script parameters are accurate here, however, this analysis was run on a different HPC machine due to memory constraints. Directories in these commands are consistent with input data and where output was then moved back to, and serve as an example only. All results were immediately placed back in ~IsoSeq-analysis/data/ANGEL

```
#The python dependencies for ANGEL are:

numpy
Biopython
scikit-learn
CD-HIT version 4.8.1
conda install -n anaCogent scikit-learn

mkdir /mnt/IsoSeq-analysis/data/ANGEL
cd /mnt/IsoSeq-analysis/data/ANGEL
```

2.1.3.0.1 Dumb ORF prediction dumb_predict.py takes as input a FASTA file. It outputs the longest ORF in all frames - use to create a top training dataset.

Usage:

```
dumb_predict.py <fasta_filename> <output_prefix>
        [--min_aa_length MIN_AA_LENGTH]
        [--use_firstORF]
        [--use_rev_strand] [--cpus CPUS]
```

By default, only the forward strand is used. This is especially true for PacBio transcriptome sequencing output.

```
dumb_predict.py
→  /mnt/IsoSeq-analysis/data/Cogent/collected/hq.fasta.no5merge.collapsed.rep.fa
→  pygmy.dumb --min_aa_length 100 --cpus 24

# ------
13335  finished      7611  clusters
```

2.1.3.0.2 Creating a non-redundant training dataset ANGEL classifier training:

```
angel_make_training_set.py pygmy.dumb.final pygmy.dumb.final.training --random
→  --cpus 24

angel_train.py pygmy.dumb.final.training.cds pygmy.dumb.final.training.utr
→  pygmy.dumb.final.classifier.pickle --cpus 12
```

2.1.3.0.3 Robust ORF prediction Based on both the ANGEL Classifier training and the dumb ORF prediction.

Sequences are output as fasta files with whether they are complete, 5' partial, 3' partial, or internal in the header information, which also includes the aa length.
Preidctions are tagged as confident, likely, or suspicious, and dumb ORF predictions as dumb.
The proportion of each isoform that is untranslated, as well as the position of the cds sequence is also output.

```
angel_predict.py ../Cogent/collected/hq.fasta.no5merge.collapsed.rep.fa
→  pygmy.dumb.final.classifier.pickle pygmy --output_mode=best
→  --min_angel_aa_length 100 --min_dumb_aa_length 100 --cpus 48
```

Output is written to:

- pygmy.ANGEL.cds
- pygmy.ANGEL.pep
- pygmy.ANGEL.utr

These outputs are used further in Chapter 5 for additional clustering and BLASTx searches.

Note: The reverse strand options within ANGEL are not necessary on Isoseq data.

All these commands are provided in the one document Running_ANGEL.sh on Github.

2.1.4 CD-HIT

Clustering of transcript isoforms. Notes and output recorded for CD-HIT analysis of peptide sequences output from ANGEL ORF prediction. This analysis was conducted on the CD-HIT web server.

CD-HIT was run by uploading pygmy.ANGEL.pep to the server and selecting a cut-off of 0.99. CD-HIT notes and recorded output are in CDHit notes.pep99.bash and below.

```
cd-hit - run on pygmy.ANGEL.pep

separate run cut-off at .99

***
Your job 1598936109 is finished.
Program you ran: cd-hit
You input file is pygmy.ANGEL.pep and we named it as 1598936109.fas.0
Summary information for 1598936109.fas.0 included in 1598936109.fas.0.stat
You required 1 runs for sequence clustering
    1. Fasta file for representative sequences at 99% identity is 1598936109.fas.1
        Summary information for 1598936109.fas.1 included in 1598936109.fas.1.stat
        Corresponding cluster file is1598936109.fas.1.clstr
        Sorted cluster file by size is 1598936109.fas.1.clstr.sorted
Generated shell script is run-1598936109.sh

faa_stat.pl 1598936109.fas.0
faa_stat.pl 1598936109.fas.1
/data5/data/NGS-ann-project/apps/cd-hit/clstr_sort_by.pl no <
↪   1598936109.fas.1.clstr > 1598936109.fas.1.clstr.sorted
/data5/data/NGS-ann-project/apps/cd-hit/clstr_list.pl 1598936109.fas.1.clstr
↪   1598936109.clstr.dump
gnuplot1.pl < 1598936109.fas.1.clstr > 1598936109.fas.1.clstr.1; gnuplot2.pl
↪   1598936109.fas.1.clstr.1 1598936109.fas.1.clstr.1.png
/data5/data/NGS-ann-project/apps/cd-hit/clstr_list_sort.pl 1598936109.clstr.dump
↪   1598936109.clstr_no.dump
/data5/data/NGS-ann-project/apps/cd-hit/clstr_list_sort.pl 1598936109.clstr.dump
↪   1598936109.clstr_len.dump len
/data5/data/NGS-ann-project/apps/cd-hit/clstr_list_sort.pl 1598936109.clstr.dump
↪   1598936109.clstr_des.dump des
```

The generated CD-HIT script is uploaded to github as run-1598936109.sh

The full output metadata file is uploaded to github as 1598936109.out and the summary data provided below:

```
================================================================
Program: CD-HIT, V4.7 (+OpenMP), Sep 06 2018, 09:36:42
Command: /data5/data/NGS-ann-project/apps/cd-hit/cd-hit -i
         1598936109.fas.0 -d 0 -o 1598936109.fas.1 -c 0.99 -n 5
         -G 1 -g 1 -b 20 -l 10 -s 0.0 -aL 0.0 -aS 0.0 -T 4 -M
         32000

Started: Mon Aug 31 21:55:20 2020
================================================================
                        Output
----------------------------------------------------------------
```

```
    total seq: 13882
    longest and shortest : 2513 and 99
    Total letters: 4952779
    Sequences have been sorted


    ...



    13882  finished       9907  clusters
```

The output 1598936109.fas.1.clstr.sorted contains the sorted cluster groups generated by CD-Hit and indicates the longest representative transcript for each cluster.

2.1.5   Reading CD-HIT output in R

dplyr was used to manipulate the CD-Hit output to create a list of unique transcript IDs of the longest representative transcript for each cluster. This list file is required in the following step to subset from the full length transcript file which contains the untranslated regions.

Load libraries and set the working directory

Import 1598936109.fas.1.clstr.sorted, the CD-Hit output file with all transcript IDs assigned to a sorted cluster (note: ./ReferenceClusters/ has been continually included due to setwd errors).

```
clstr <- read.csv("./ReferenceClusters/1598936109.fas.1.clstr.sorted",
    sep = "\t", row.names = NULL, header = FALSE, stringsAsFactors = FALSE)
head(clstr)
```

Initial file had a blank line with ">Cluster #" in between new clusters. Extend the >Cluster # down the column so that all transcripts are directly associated with a cluster number in their row.

```
clstr2 <- clstr
n = nrow(clstr)
x = 0
numbers_only <- function(x) !grepl("\\D", x)
for (row in c(1:n)) {
    if (numbers_only(clstr2[row, 1]) == TRUE) {
        clstr2[row, 1] <- x
    } else {
        NULL
    }
    x <- clstr2[row, 1]
}
head(clstr2, 20)
```

Count the number of transcript members for each cluster.

```
clstr.sums <- data.frame(dplyr::count(clstr2, V1))
write.csv(clstr.sums, "~/@Uni/@Flinders University
↪    PhD/RWorkingDir/Workflow-readthedown/ReferenceClusters/clstr.sums.csv",
    quote = FALSE, row.names = FALSE, col.names = FALSE)
head(clstr.sums)
```

```
##                V1  n
## 1    >Cluster 0 19
## 2    >Cluster 1 16
## 3   >Cluster 10 12
## 4  >Cluster 100  6
## 5 >Cluster 1000  3
## 6 >Cluster 1001  3
```

Remove the additional blank row without a transcript name from between each cluster group.

```
switch <- clstr.sums[1, 2]
clstr3 <- cbind(clstr2[1], clstr)
clstr3[c((switch - 5):(switch + 5)), ]
```

```
clstr4 <- clstr2[-which(clstr2$V2 == ""), ]
clstr4[c(1:5, (switch - 5):(switch + 5)), ]
```

Separate CD-Hit information into columns by delimiters:

- Remove > symbol from transcript names
- Move number of amino acids into a new column by separating "aa"
- Move statistics of % match into a new column byt separating "..."
- Give the columns headers

```
clstr5 <- clstr4
clstr5[] <- lapply(clstr5, gsub, pattern = ">", replacement = "")
clstr5.2 <- data.frame(str_split_fixed(clstr5$V2, "aa, ", 2))
clstr5.3 <- data.frame(str_split_fixed(clstr5.2$X2, "... ", 2))
clstr6 <- cbind(clstr5[1], clstr5.2[1], clstr5.3[1:2])
colnames(clstr6) <- c("cluster", "aa", "TranscriptID", "stat")
head(clstr6)
```

Filter based on rows that have a "*" in the stat column. These indicate the representative sequences for each cluster determined by CD-Hit and will act as the main reference.

```
clstr7 <- filter(clstr6, stat == "*")
head(clstr7, 50)
```

Pull out the column "TranscriptID" and export into a txt file so it can be used as a list to extract these sequences.

```
clstrTranscriptID <- pull(clstr7, var = TranscriptID)

write.csv(clstrTranscriptID, "~/@Uni/@Flinders University
↪   PhD/RWorkingDir/Workflow-readthedown/ReferenceClusters/clstrTranscriptID.csv",
    quote = FALSE, row.names = FALSE, col.names = FALSE)
head(clstrTranscriptID)
```

```
## [1] PB.6088.11|e88f16|path1:164-2792(+)|transcript/22345|m.11362
## [2] PB.2919.1|6c1fb9|path4:13-2574(+)|transcript/10624|m.5342
```

```
## [3] PB.535.2|136fa8|path0:1-2775(+)|transcript/4312|m.913
## [4] PB.4689.2|afa3de|path19:22-2096(+)|transcript/11491|m.8666
## [5] PB.6357.1|f27f14|path2:4-2108(+)|transcript/12511|m.11846
## [6] PB.715.4|1b393a|path6:6-8031(+)|transcript/8|m.1216
## 13882 Levels: PB.10.1|004815|path1:5-1713(+)|transcript/17534|m.3 ...
```

Pull out the first segment before the first | from column "TranscriptID" and export into a txt file. this reduced the ID to the initial PB.<loci_index>.<isoform_index> identifier.

```
WholeTranscriptID <- word(clstrTranscriptID, 1, sep = "\\|")
write.csv(WholeTranscriptID, "~/@Uni/@Flinders University
↪    PhD/RWorkingDir/Workflow-readthedown/ReferenceClusters/WholeTranscriptID.csv",
     quote = FALSE, row.names = FALSE, col.names = FALSE)
head(WholeTranscriptID)
```

```
## [1] "PB.6088.11" "PB.2919.1"  "PB.535.2"   "PB.4689.2"  "PB.6357.1"
## [6] "PB.715.4"
```

Note there will be "duplicates" if the name is truncated all the way back to the PB.<loci_index>.<isoform_index> identifier due to multiple isoforms that are retained. A longer, but still truncated version of this ID (excluding the information appended by ANGEL), was required to subset from the hq.fasta.no5merge.collapsed.rep.fa file.

The clstrTranscriptID.csv is now the full list of representative transcripts based on clustering of the translated proteins in predicted read frame. This file was then used in NECTAR to extract a .fasta of all the reference transcripts.

2.1.6 Seqtk

Seqtk was used on the NECTAR machine to manipulate the fasta files and subset based on list files of transcript names.

Installation:

```
#Carmel July 2020

#Used on NECTAR research cloud

#Relies on hard-coded paths & filenames
#run line by line in shell

----------------

cd /mnt/Prog
conda install -c bioconda seqtk

#___
The following packages will be downloaded:

    package                    |             build
    ---------------------------|-----------------
    certifi-2020.6.20          |   py37hc8dfbb8_0         151 KB  conda-forge
    conda-4.8.3                |   py37hc8dfbb8_1         3.0 MB  conda-forge
```

```
         openssl-1.1.1g              |           h516909a_1        2.1 MB  conda-forge
         python_abi-3.7              |              1_cp37m          4 KB  conda-forge
         seqtk-1.3                   |           hed695b0_2         39 KB  bioconda
         ------------------------------------------------------------
                                                  Total:          5.3 MB

    The following NEW packages will be INSTALLED:

      python_abi          conda-forge/linux-64::python_abi-3.7-1_cp37m
      seqtk               bioconda/linux-64::seqtk-1.3-hed695b0_2
    #___

    Conda packages are located in /mnt/Prog/miniconda3/bin/ if you dont set a path
    ----------------
```

Manually convert clstrTranscriptID.csv to an clstrTranscriptID.lst file.

Seqtk was used to subset the pygmy.ANGEL.cds fasta file based on the R cluster list outputs: Moved files into /mnt/IsoSeq-analysis/data/Seqtk

```
    screen          #to be safe
    cd /mnt/IsoSeq-analysis/data/Seqtk
    /mnt/Prog/miniconda3/bin/seqtk subseq ../ANGEL/pygmy.ANGEL.cds
    ↪  clstrTranscriptID.lst > clstrTranscriptID.fa
```

check:

```
    wc -l clstrTranscriptID.lst
    # 9907
    wc -l clstrTranscriptID.fa
    # 19814

    # 19814/2 = 9907. So it is correct
```

The hq.fasta.no5merge.collapsed.rep.fa file created in the program ANGEL in Chapter 4 is a fasta file where sequence information ran over multiple "lines" within the file.

Seqtk was used to convert this file into a .fasta with sequence on a single line.

```
    17/11/2020
    # just converting hq.fasta.no5merge.collapsed.rep.fa into a fasta with sequence all
    ↪   on one line
    # command seq is basic fasta/q conversion

    cd /mnt/IsoSeq-analysis/data/Seqtk
    /mnt/Prog/miniconda3/bin/seqtk seq
    ↪  ../Cogent/collected/hq.fasta.no5merge.collapsed.rep.fa >
    ↪  hq.fasta.no5merge.collapsed.rep.1L.fa
```

Checked that the file still has the correct number of sequences:

```
$ grep -c ">" hq.fasta.no5merge.collapsed.rep.fa
15729


grep -c ">" hq.fasta.no5merge.collapsed.rep.1L.fa
15729
```

Seqtk was also used to search for transcripts which returned BLAST results for some genes of interest as data exploration. The reference list files for this were extracted from the BLASTx output were created using find and sort functions in notepad++ and excel, and then applied to the pygmy.ANGEL.cds fasta file to extract sequences.

```
#Usage from /mnt/IsoSeq-analysis/data/ANGEL

cd /mnt/IsoSeq-analysis/data/Seqtk
/mnt/Prog/miniconda3/bin/seqtk subseq ../ANGEL/pygmy.ANGEL.cds AQP.lst > AQP.fa

/mnt/Prog/miniconda3/bin/seqtk subseq ../ANGEL/pygmy.ANGEL.cds HSP.lst > HSP.fa

/mnt/Prog/miniconda3/bin/seqtk subseq ../ANGEL/pygmy.ANGEL.cds SLC.lst > SLC.fa

/mnt/Prog/miniconda3/bin/seqtk subseq ../ANGEL/pygmy.ANGEL.cds MAPK.lst > MAPK.fa
```

2.1.7 Subsetting full length transcripts which represent each cluster

Creation of a subset of the longest representative transcripts for each cluster as per CD-HIT and the previous two to form the reference for counts in the following Chapter 6.

As only the peptide coding sequence of transcripts was used in the BLASTx search and the CD-HIT clustering in order to cluster similarly expressed transcripts, the untranslated regions of transcripts are absent. Our goal is to include these untranslated regions in the reference to be used for gene expression counts.
Rather than attempting to piece pygmy.ANGEL.cds and pygmy.ANGEL.utr back together, a truncated version of the transcript names was used to pull the full length unaltered transcript out of the earlier hq.fasta.no5merge.collapsed.rep.fa file.

Initial checks of clstrTranscriptID.csv were carried out. Files are continually checked for transcript ID or sequence counts to ensure manual manipulation did not remove any data that was not being targeted. Exploration of discrepancies below identified a number of transcripts that with clustering of multiple isoforms, have seeded more than one cluster. This means that there were duplicates in the list file once coding region identifiers were removed and sequences had to be extracted manually and not by applying the clstrTranscriptID list to hq.fasta.no5merge.collapsed.rep.1L.fa in seqtk.

The following was done on a local Intel core i7 machine using RStudio or in-house bash scripts applied in Git Bash for Windows. Some later steps were carried out on the NECTAR allocation.

All files copied to "~/IsoSeq-analysis/data/Seqtk/reference COPIES 2020-11" and contents checked

The following file contains all of the below:

```
Carmel@CarmelASUS MINGW64 ~
cd "E:/@DATA/@@PBT_KidneyIsoSeq_Working
↪  Data_C.Maher/IsoSeq-analysis/data/Seqtk/reference COPIES 2020-11"

2020-11-17
```

```
cd "C:\Users\Carmel\Desktop\reference COPIES 2020-11"


-----


19814 clstrTranscriptID.fa                          /2 = 9907
9907 clstrTranscriptID.lst

hq.fasta.no5merge.collapsed.rep.fa
hq.fasta.no5merge.collapsed.rep.1L.fa
pygmy.ANGEL.cds
pygmy.ANGEL.pep
pygmy.ANGEL.utr
```

Sort and filter clstrTranscriptID.lst by unique values to make sure there are no dupicates

```
sort clstrTranscriptID.lst | uniq > clstrUniq.txt
wc -l clstrUniq.txt
9907 clstrUniq.txt
```

List duplicates if there are any found (there were not)

```
sort clstrTranscriptID.lst | uniq -d > clstrUniqD.txt
wc -l clstrUniqD.txt
0 clstrUniqD.txt
```

```
Carmel@CarmelASUS MINGW64 /e/@DATA/@@PBT_KidneyIsoSeq_Working
↪  Data_C.Maher/IsoSeq-analysis/data/Seqtk/reference COPIES 2020-11
```

---> 1. Isoseq HQ output file: hq.fasta.no5merge.collapsed.rep.fa

header format

```
head -n 2 hq.fasta.no5merge.collapsed.rep.fa
>PB.1.1|000643|path0:1-1879(+)|transcript/14742 transcript/14742
↪  full_length_coverage=3;length=1887;num_subreads=52
ATAGAGTACAAGTACTAGAGTTTTAGTGTGGTCGAAGAGAAATCCAAGATCCACCATGGG
```

no of sequences

```
grep -c ">" hq.fasta.no5merge.collapsed.rep.fa
15729
```

---> 2. Angel CDS output from Isoseq HQ file: pygmy.ANGEL.cds

header format

```
head -n 2 pygmy.ANGEL.cds
>PB.2.1|002537|path0:1-1624(+)|transcript/18304|m.1 type:likely-NA len:135 strand:+
↪  pos:282-686

↪  ATGTCCTACTGCAGGCAAGAAGGAAAGGACAGAATTATATTTGTGACAAAAGAGGACCATGAAACTCCAAGTAGTGCTGAGTTAGTAGCGGAT(
```

Note Sequence is on a single line

no of sequences

```
grep -c ">" pygmy.ANGEL.cds
13882
```

---> Angel CDS file clusterd with CD-HIT: 1598936109.fas.1.clstr

header format

```
cd cd-hit\ 99
head 1598936109.fas.1.clstr
>Cluster 0
0       2513aa, >PB.715.4|1b393a|path6:6-8031(+)|transcript/8|m.1216... *
1       2250aa, >PB.715.5|1b393a|path6:278-8030(+)|transcript/13|m.1217... at
↪  99.56%
2       2369aa, >PB.715.6|1b393a|path6:524-8029(+)|transcript/20|m.1218... at
↪  99.58%
3       2148aa, >PB.715.7|1b393a|path6:1187-8030(+)|transcript/49|m.1219... at
↪  99.53%
4       1913aa, >PB.715.9|1b393a|path6:1893-8028(+)|transcript/118|m.1221... at
↪  99.79%
5       1688aa, >PB.715.11|1b393a|path6:2596-8054(+)|transcript/279|m.1223... at
↪  99.59%
6       1452aa, >PB.715.14|1b393a|path6:3304-8029(+)|transcript/591|m.1226... at
↪  100.00%
7       1333aa, >PB.715.15|1b393a|path6:3629-8031(+)|transcript/935|m.1227... at
↪  99.77%
8       1314aa, >PB.715.16|1b393a|path6:3916-8078(+)|transcript/962|m.1228... at
↪  100.00%
```

File contains truncated headers from the Angel CDS file, clustered acoording to similarity. Truncation appears to be at the first whitespace no of sequences

```
grep -c ', >' 1598936109.fas.1.clstr      #search pattern changed to avoid the ">"
↪  in front of each cluster label eg >Cluster 0
13882
```

Sequence number matches Angel CDS file

no of clusters

```
grep -c '>C' 1598936109.fas.1.clstr
9907
```

```
grep -c '>C' 1598936109.fas.1.clstr.sorted
9907
```

both match

```
tail 1598936109.fas.1.clstr
>Cluster 9902
0       99aa, >PB.8346.1|transcript/22300:1-1143(+)|transcript/22300|m.14112... *
>Cluster 9903
0       99aa, >PB.8372.1|transcript/22508:1-1103(+)|transcript/22508|m.14137... *
>Cluster 9904
0       99aa, >PB.8489.1|transcript/23315:1-855(+)|transcript/23315|m.14244... *
>Cluster 9905
0       99aa, >PB.8706.1|transcript/25054:1-372(+)|transcript/25054|m.14469... *
>Cluster 9906
0       99aa, >PB.9028.1|transcript/4242:1-2955(+)|transcript/4242|m.14776... *
```

```
tail 1598936109.fas.1.clstr.sorted
>Cluster 9902
0       99aa, >PB.8346.1|transcript/22300:1-1143(+)|transcript/22300|m.14112... *
>Cluster 9903
0       99aa, >PB.8372.1|transcript/22508:1-1103(+)|transcript/22508|m.14137... *
>Cluster 9904
0       99aa, >PB.8489.1|transcript/23315:1-855(+)|transcript/23315|m.14244... *
>Cluster 9905
0       99aa, >PB.8706.1|transcript/25054:1-372(+)|transcript/25054|m.14469... *
>Cluster 9906
0       99aa, >PB.9028.1|transcript/4242:1-2955(+)|transcript/4242|m.14776... *
```

Cluster numbers also match both files:Cluster 9906 + Cluster 0 = 9907

Check if the header changed from Isoseq HQ after processing with Angel?

```
cd ../
grep '>PB.715.4|1b393a|path6:6-8031(+)|transcript/8'
↪  hq.fasta.no5merge.collapsed.rep.fa
>PB.715.4|1b393a|path6:6-8031(+)|transcript/8 transcript/8
↪  full_length_coverage=43;length=7837;num_subreads=60
```

note: the "|m.xxxxx" suffix in the ANGEL output file is added at the first whitespace

---> **3. The longest seed sequence from the CD-HIT output: clstrUniq.txt**

header format

```
head clstrUniq.txt
PB.10.1|004815|path1:5-1713(+)|transcript/17534|m.3
PB.100.1|049515|path5:1-1563(+)|transcript/19243|m.132
PB.1000.1|26d4b2|path0:1-2889(+)|transcript/4521|m.1797
PB.1000.2|26d4b2|path0:1287-2889(+)|transcript/18232|m.1798
PB.1002.1|26ea71|path2:1-2673(+)|transcript/6271|m.1799
PB.1002.2|26ea71|path2:11-2691(+)|transcript/8089|m.1800
PB.1003.1|26fc93|path2:1-3987(+)|transcript/1197|m.1801
PB.1005.1|2718e2|path0:1-5294(+)|transcript/258|m.1804
PB.1006.1|272202|path0:1-2713(+)|transcript/5814|m.1805
PB.1007.1|272202|path1:1-3370(+)|transcript/2554|m.1806
```

Prefix ">" and suffix "... *" have been removed in R Studio

no of strings

```
grep -c 'PB.' clstrUniq.txt
9907
```

matches number of clusters above

strip the "|m.xxxx" from the strings in clstrUniq.txt

```
sed 's/|m.*//' clstrUniq.txt > new_cluster.txt
```

check

```
head new_cluster.txt
PB.10.1|004815|path1:5-1713(+)|transcript/17534
PB.100.1|049515|path5:1-1563(+)|transcript/19243
PB.1000.1|26d4b2|path0:1-2889(+)|transcript/4521
PB.1000.2|26d4b2|path0:1287-2889(+)|transcript/18232
PB.1002.1|26ea71|path2:1-2673(+)|transcript/6271
PB.1002.2|26ea71|path2:11-2691(+)|transcript/8089
PB.1003.1|26fc93|path2:1-3987(+)|transcript/1197
PB.1005.1|2718e2|path0:1-5294(+)|transcript/258
PB.1006.1|272202|path0:1-2713(+)|transcript/5814
PB.1007.1|272202|path1:1-3370(+)|transcript/2554
```

```
grep -c 'PB.' new_cluster.txt
9907
```

How many of thes strings are in the original ONE LINE isoseq file?

```
grep -c -Ff new_cluster.txt hq.fasta.no5merge.collapsed.rep.1L.fa
9813
```

94 strings are missing

---> **4. find the missing strings**

convert the isoform hq file into the same format as new_cluster.txt

```
grep '>' hq.fasta.no5merge.collapsed.rep.1L.fa| sed s'/ tra.*//' | sed s'/>//' >
↪  hq_headers
head hq_headers
PB.1.1|000643|path0:1-1879(+)|transcript/14742
PB.2.1|002537|path0:1-1624(+)|transcript/18304
PB.3.2|00361c|path0:43-2240(+)|transcript/10564
PB.4.1|004079|path1:6-745(+)|transcript/23781
PB.4.2|004079|path1:6-612(+)|transcript/24311
PB.5.1|004079|path2:209-642(+)|transcript/24981
PB.6.1|004079|path3:273-616(+)|transcript/25193
PB.7.1|004079|path5:6-664(+)|transcript/24242
PB.8.1|004079|path6:1-607(+)|transcript/24477
PB.9.1|004079|path7:1-725(+)|transcript/23873
```

```
cat hq_headers |wc -l
15729
```

Checking for the same number of matches

```
grep -Ff new_cluster.txt hq_headers > matching_strings
cat matching_strings |wc -l
9813
```

Checking for duplicates

```
cat hq_headers | sort |uniq -cd
0

cat new_cluster.txt | sort |uniq -cd
      2 PB.1206.1|2e8181|path5:1-4617(+)|transcript/551
      2 PB.1261.1|30ba06|path3:633-1922(+)|transcript/21324
      2 PB.1570.2|3bd5a8|path2:20-1934(+)|transcript/13977
      2 PB.1719.1|404d72|path0:1-3429(+)|transcript/2407
      2 PB.1764.3|429969|path8:44-2062(+)|transcript/12739
      2 PB.1813.1|446aed|path0:1-2871(+)|transcript/4320
      2 PB.1865.1|46bd4e|path0:1-2536(+)|transcript/6726
      2 PB.1869.1|46d663|path2:1-1320(+)|transcript/21375
      2 PB.220.1|08884f|path0:1-4120(+)|transcript/991
      2 PB.2360.1|5843a6|path32:1-2062(+)|transcript/11326
      2 PB.2541.3|5e93d1|path0:5-2495(+)|transcript/7381
      2 PB.2560.1|5f09e2|path0:1-1196(+)|transcript/21957
      2 PB.2618.1|612418|path1:1-2013(+)|transcript/12200
      2 PB.2713.3|65abaa|path0:37-2495(+)|transcript/7693
      2 PB.2892.1|6b359a|path12:1-4597(+)|transcript/526
      2 PB.3037.2|702e9b|path5:2-3018(+)|transcript/3923
      2 PB.3243.1|784f79|path0:1-2683(+)|transcript/5903
      2 PB.3259.1|792211|path3:1-5233(+)|transcript/268
      2 PB.3286.3|7a2645|path3:103-3087(+)|transcript/3780
      2 PB.3405.1|7f2f96|path0:5-2081(+)|transcript/11903
      3 PB.3430.1|8055ae|path8:1-3589(+)|transcript/2043
      2 PB.3461.1|81f072|path5:1-2492(+)|transcript/7285
```

```
2 PB.351.4|0cccb8|path5:27-4293(+)|transcript/775
2 PB.3562.1|864d1b|path0:1-3492(+)|transcript/2174
2 PB.3615.1|883e51|path3:1-4044(+)|transcript/1347
2 PB.3836.3|8ff0e9|path0:4-3972(+)|transcript/946
2 PB.3934.1|93b547|path0:1-3353(+)|transcript/2676
2 PB.4042.2|97fba8|path2:34-2096(+)|transcript/13603
2 PB.4219.2|9d8e09|path3:9-3566(+)|transcript/2045
2 PB.4442.1|a6620e|path0:1-2906(+)|transcript/4555
2 PB.4458.1|a6e2a5|path0:1-3673(+)|transcript/1739
2 PB.4822.1|b4950f|path1:1-3315(+)|transcript/2177
2 PB.4822.2|b4950f|path1:1240-3315(+)|transcript/10982
2 PB.4919.1|b7da5b|path0:1-3327(+)|transcript/2652
2 PB.4923.1|b7fed9|path0:1-2394(+)|transcript/8643
2 PB.5080.1|bde542|path3:1-3060(+)|transcript/3777
2 PB.5096.2|be8c26|path4:70-2754(+)|transcript/5782
2 PB.5149.1|c15864|path0:1-1628(+)|transcript/17934
2 PB.5308.1|c8b5b5|path2:35-3805(+)|transcript/1487
2 PB.5395.1|cbd93f|path0:1-1699(+)|transcript/17107
2 PB.5483.2|cf1828|path1:23-2344(+)|transcript/9126
2 PB.5619.1|d4d2ea|path0:1-2677(+)|transcript/6673
2 PB.5663.1|d5e71b|path0:1-5101(+)|transcript/303
2 PB.5939.1|e23f0e|path0:1-2960(+)|transcript/3907
2 PB.5939.2|e23f0e|path0:5-4542(+)|transcript/622
2 PB.613.1|16d05c|path0:1-1558(+)|transcript/18673
2 PB.6133.1|ea032e|path9:1-2052(+)|transcript/12270
2 PB.6276.1|ee992b|path0:1-1487(+)|transcript/19762
2 PB.6283.1|eee48e|path1:1-3454(+)|transcript/2331
2 PB.6286.1|ef0b10|path12:1-2744(+)|transcript/5222
2 PB.6303.1|efc514|path0:1-3222(+)|transcript/2889
2 PB.6378.3|f31894|path9:80-2754(+)|transcript/5975
2 PB.6385.1|f32c31|path0:1-1500(+)|transcript/12840
2 PB.6426.1|f4ded6|path0:1-1796(+)|transcript/16785
2 PB.6576.2|facceb|path2:1-3314(+)|transcript/3157
2 PB.659.1|19437d|path1:1-3483(+)|transcript/2210
2 PB.6763.1|transcript/10186:1-2230(+)|transcript/10186
2 PB.6877.1|transcript/1099:1-4113(+)|transcript/1099
2 PB.6966.1|transcript/11636:1-2103(+)|transcript/11636
2 PB.713.1|1b043a|path0:1-965(+)|transcript/23154
2 PB.7227.1|transcript/138:1-5869(+)|transcript/138
2 PB.7287.1|transcript/14227:1-1898(+)|transcript/14227
3 PB.7342.1|transcript/14710:1-1879(+)|transcript/14710
2 PB.7349.1|transcript/14730:1-1903(+)|transcript/14730
2 PB.7353.1|transcript/14772:1-1860(+)|transcript/14772
2 PB.74.1|02ebe4|path0:1-2606(+)|transcript/6537
2 PB.7501.1|transcript/1596:1-3759(+)|transcript/1596
2 PB.7528.1|transcript/16151:1-1797(+)|transcript/16151
2 PB.7678.1|transcript/1728:1-3672(+)|transcript/1728
2 PB.7707.1|transcript/17496:1-1683(+)|transcript/17496
2 PB.7868.1|transcript/188:1-5558(+)|transcript/188
2 PB.7973.1|transcript/19628:1-1476(+)|transcript/19628
2 PB.8082.1|transcript/20433:1-1412(+)|transcript/20433
2 PB.816.1|1e334e|path4:1-2040(+)|transcript/12316
2 PB.8325.1|transcript/22158:1-1191(+)|transcript/22158
```

```
2 PB.834.2|1ec273|path2:1291-2750(+)|transcript/19945
2 PB.8389.1|transcript/22625:1-1060(+)|transcript/22625
2 PB.8709.1|transcript/252:1-5299(+)|transcript/252
2 PB.8883.1|transcript/3393:1-3152(+)|transcript/3393
2 PB.8886.1|transcript/3402:1-3146(+)|transcript/3402
2 PB.8920.1|transcript/3588:1-3076(+)|transcript/3588
2 PB.9068.1|transcript/4442:1-2918(+)|transcript/4442
2 PB.9162.1|transcript/4983:1-2815(+)|transcript/4983
2 PB.9209.1|transcript/5238:1-2771(+)|transcript/5238
2 PB.924.1|22d3cf|path3:1-1494(+)|transcript/19025
2 PB.9277.1|transcript/5650:1-2707(+)|transcript/5650
2 PB.944.1|241290|path4:2-2629(+)|transcript/6557
2 PB.9535.1|transcript/7343:1-2508(+)|transcript/7343
2 PB.9581.1|transcript/7647:1-2463(+)|transcript/7647
2 PB.9683.1|transcript/8355:1-2402(+)|transcript/8355
2 PB.9766.1|transcript/9024:1-2324(+)|transcript/9024
2 PB.9784.1|transcript/915:1-4197(+)|transcript/915
```

During the clustering it appears that several of the sequences have been made seed sequences (denoted with a *) for more than one cluster.

The counts above equate to how many strings which is the discrepancy noted in section 3 above.

```
cat new_cluster.txt | sort |uniq -cd | cut -d' ' -f7 | awk '{sum+=$1-1}
↪  END{printf("%d\n",sum)}'
94
```

Both hq.fasta.no5merge.collapsed.rep.fa and hq.fasta.no5merge.collapsed.rep.1L.fa. were checked in this way but the single line file is required in later steps so only those commands have been included.

Final pulling out of transcripts sequences into a .fasta on a single line, also removes the skipped lines that grep outputs as '--'.
note: NOT using seqtk anymore:

```
grep -Ff new_cluster.txt -A 1 hq.fasta.no5merge.collapsed.rep.1L.fa | grep -v ^-- >
↪  reference_transcripts.1Lv.fasta

grep -c ">" reference_transcripts.1Lv.fasta
9813
```

---> **next step**

Persisting poly-a tails needed to be trimmed from the following file:

```
reference_transcripts.1Lv.fasta
```

Note: The Trim polya command in bbduk leaves many tails intact due to single non 'A' bases and does not facilitate hamming distance command at the same time (?), however specifying Ktrim=r literal or "trimming twice" (poly-a command and THEN ktrim) sequences become very short even with different K values.

Terry Bertozzi provided perl script as below to manually trim:

```
# Terry Bertozzi
#command line to remove poly A tails

#for use on /mnt/IsoSeq-analysis/data/Seqtk/reference_transcripts.1Lv.fasta

cd /mnt/IsoSeq-analysis/data/Seqtk

perl -pe 's/(A{5,}.{0,2})+$//gm' reference_transcripts.1L.fasta >
↪  reference_transcripts.1L.clean.fasta

perl -pe 's/(A{5,}.{0,2})+$//gm' reference_transcripts.1Lv.fasta >
↪  reference_transcripts.1Lv.clean.fasta

grep -c ">" reference_transcripts.1Lv.clean.fasta
9813

head reference_transcripts.1Lv.clean.fasta
#also looks good.
```

reference_transcripts.1Lv.clean.fasta - IS THE REFERENCE

––––––––––––––––––––––––––––––––

2.2 *Short Read data*

De-novo assembly of short-read Transcripts:
Short read transcripts used for gene expression analysis in Chater 6 were initially intended for de-novo assembly and analysis due to a lack of closely related genomic references for *T. adelaidensis*.

This section outlines the methods for assembly of short read datasets in the absence of a reference or genome for the purposes of comparing these final transcript datasets to the assembly generated from the long reads as above.

Other notes: Initial plans for Chapter 6 did not include the availability of long-read sequencing to supplement a reference transcript list. Therefore, a pooled transcript assembly was created by concatenating reads from all eight samples in order to use this larger dataset as a pseudo-reference for other programs.
This longer assembly is included in the BUSCO analysis below to illustrate assembly completeness and duplication.

Sequences were downloaded from the Australian Genomics Research Facility server & checksumms verified. This analysis was conducted on the Flinders University 'Deep Thought' HPC.

2.2.1 FASTQC quality assessment

Quality of sequencing output was quantified using the following script and FastQC v0.11.8 (note there are two commands because there were two sequencing run data folders AGRF_CAGRF13871_HCGYYBCXY and AGRF_CAGRF20022_CDNJBANXX).

The FastQC Reports were then combined for visualisation using NGSReports The full FastQC output summary .html file of all eight samples PRIOR to trimming is available here.
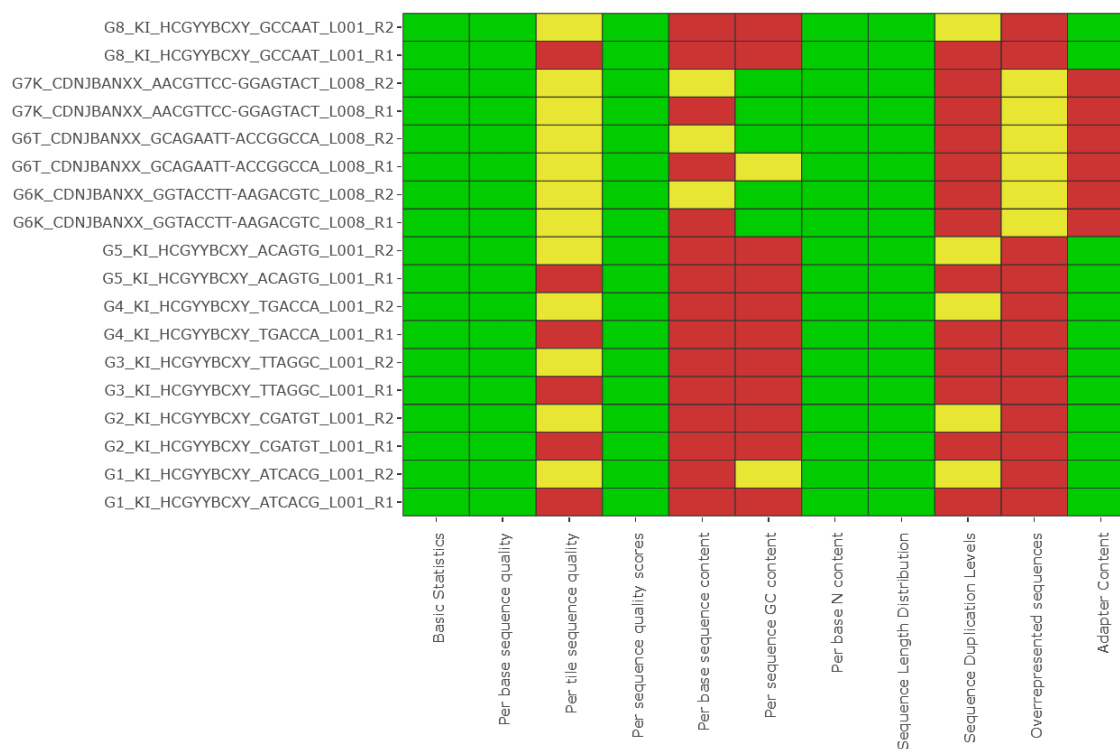
Figure 1: FastQC Summary of scores for paired reads of eight samples prior to any adapter or quality trimming

Samples G6 and G7 were run on a second sequencing run, and have different overall lengths. This is the main source of variation in sequencing scores.

2.2.2 BBduk adapter trimming

Trimming to remove Illumina TruSeq adapters and SMARTer PCR tags as well as trimming based on quality scores was done using bbduk, in in bbmap v38.46 tools and the following script removeAdapters2.sh.

```bash
!/bin/bash

# November 2019
# Carmel Maher
# This script is to submit a BBduk script to cut adapters and quality on raw reads
↪   through Flinders' DeepThought HPC Slurm
# This script therefore assumes submission to Flinders Deepthought with an assumed
↪   directory structure and use of available modules


#------------------------------------
# Required Modules:
module add bbmap #/38.46
module add fastqc #/0.11.8

#adapters location :ref=/cm/shared/apps/bbmap/38.46/resources/adapters.fa
```

```
THREADS=8

#----------------------------------
# Terry's error exit

function error_exit
{
    # Exit function due to fatal error
    # Accepts 1 arg:
    # string - descriptive error message

    echo "${PROGNAME}: ${1:-"Unknown error"}" 1>&2
    exit 1
}

# in = /scratch/user/mahe0050/DE-analysis/0_rawData/
# out = /scratch/user/mahe0050/DE-analysis/1_trimmedData

#----------------------------------

#change directories into raw sequencing directory 1
cd /scratch/user/mahe0050/DE-analysis/0_rawData/AGRF_CAGRF13871_HCGYYBCXY ||
↪   error_exit "$LINENO: directory error 1"
mkdir ../../1_trimmedData/temp

for file in *R1.fastq.gz
do
  echo $file
  FILESTEM=${file%_*}
  echo $FILESTEM
  FILESTEM2=`echo $file | cut -d "_" -f1,2 --output-delimiter="_"`
  echo $FILESTEM2

  #remove Illumina adapters from all paired data using bbduk
  bbduk.sh in=$file in2=$FILESTEM"_R2.fastq.gz"
  ↪   out=../../1_trimmedData/temp/$FILESTEM2"_R1_Iclean.fq.gz"
  ↪   out2=../../1_trimmedData/temp/$FILESTEM2"_R2_Iclean.fq.gz"
  ↪   outs=../../1_trimmedData/temp/$FILESTEM2"_singletons_Iclean.fq.gz"
  ↪   literal=GATCGGAAGAGCACA,AGATCGGAAGAGCGT ktrim=r k=15 mink=15 hdist=1
  ↪   threads=$THREADS || error_exit "$LINENO: Error removing Illumina adapters
  ↪   from R1 or R2 - directory 1"

  #remove SMARTer PCR tags from all paired data using bbduk
  bbduk.sh in=../../1_trimmedData/temp/$FILESTEM2"_R1_Iclean.fq.gz"
  ↪   in2=../../1_trimmedData/temp/$FILESTEM2"_R2_Iclean.fq.gz"
  ↪   out=../../1_trimmedData/$FILESTEM2"_R1_clean.fq.gz"
  ↪   out2=../../1_trimmedData/$FILESTEM2"_R2_clean.fq.gz"
  ↪   outs=../../1_trimmedData/$FILESTEM2"_singletons_clean.fq.gz"
  ↪   literal=AAGCAGTGGTATCAACGCAGAGTACNNNNN,GTACTCTGCGTTGATACCACTGCTT ktrim=r k=15
  ↪   mink=15 hdist=1 qtrim=r trimq=10 minlength=30 threads=$THREADS || error_exit
  ↪   "$LINENO: Error removing SMARTer adapters from R1 or R2 - directory 1"
```

```bash
    #remove SMARTer PCR tags from the singleton file using bbduk
    bbduk.sh in=../../1_trimmedData/temp/$FILESTEM2"_singletons_Iclean.fq.gz"
    ↪    out=../../1_trimmedData/$FILESTEM2"_singletons1.fq.gz"
    ↪    literal=AAGCAGTGGTATCAACGCAGAGTACNNNNN,GTACTCTGCGTTGATACCACTGCTT ktrim=r k=15
    ↪    mink=15 hdist=1 qtrim=r trimq=20 minlength=30 threads=$THREADS || error_exit
    ↪    "$LINENO: Error removing SMARTer adapters from singletons - directory 1"

    #concatenate the singleton files
    cat ../../1_trimmedData/$FILESTEM2"_singletons1.fq.gz" >>
    ↪    ../../1_trimmedData/$FILESTEM2"_singletons_clean.fq.gz" || error_exit
    ↪    "$LINENO: Error concatenating singletons - directory 1"


    #clean up: i.e. remove temporary directories
    rm ../../1_trimmedData/$FILESTEM2"_singletons1.fq.gz" || error_exit "$LINENO:
    ↪    Error deleting singleton intermediate - directory 1"

done


#change directories into raw sequencing directory 2 and do it again for these
↪    samples (output to same trimmed directory for further analysis together)
cd /scratch/user/mahe0050/DE-analysis/0_rawData/AGRF_CAGRF20022_CDNJBANXX ||
↪    error_exit "$LINENO: directory error 2"

for file in *R1.fastq.gz
do
  echo $file
  FILESTEM=${file%_*}
  echo $FILESTEM
  FILESTEM2=`echo $file | cut -d "_" -f1,2 --output-delimiter="_"`
  echo $FILESTEM2

#(repeat of the above but edit directories & names: same output folder as all
↪    samples have unique names)
    #remove Illumina adapters from all paired data using bbduk
    bbduk.sh in=$file in2=$FILESTEM"_R2.fastq.gz"
    ↪    out=../../1_trimmedData/temp/$FILESTEM2"_R1_Iclean.fq.gz"
    ↪    out2=../../1_trimmedData/temp/$FILESTEM2"_R2_Iclean.fq.gz"
    ↪    outs=../../1_trimmedData/temp/$FILESTEM2"_singletons_Iclean.fq.gz"
    ↪    literal=GATCGGAAGAGCACA,AGATCGGAAGAGCGT ktrim=r k=15 mink=15 hdist=1
    ↪    threads=$THREADS || error_exit "$LINENO: Error removing Illumina adapters
    ↪    from R1 or R2 - directory 2"

    #remove SMARTer PCR tags from all paired data using bbduk
    bbduk.sh in=../../1_trimmedData/temp/$FILESTEM2"_R1_Iclean.fq.gz"
    ↪    in2=../../1_trimmedData/temp/$FILESTEM2"_R2_Iclean.fq.gz"
    ↪    out=../../1_trimmedData/$FILESTEM2"_R1_clean.fq.gz"
    ↪    out2=../../1_trimmedData/$FILESTEM2"_R2_clean.fq.gz"
    ↪    outs=../../1_trimmedData/$FILESTEM2"_singletons_clean.fq.gz"
    ↪    literal=AAGCAGTGGTATCAACGCAGAGTACNNNNN,GTACTCTGCGTTGATACCACTGCTT ktrim=r k=15
    ↪    mink=15 hdist=1 qtrim=r trimq=10 minlength=30 threads=$THREADS || error_exit
    ↪    "$LINENO: Error removing SMARTer adapters from R1 or R2 - directory 2"
```

```
    #remove SMARTer PCR tags from the singleton file using bbduk
    bbduk.sh in=../../1_trimmedData/temp/$FILESTEM2"_singletons_Iclean.fq.gz"
    ↪   out=../../1_trimmedData/$FILESTEM2"_singletons1.fq.gz"
    ↪   literal=AAGCAGTGGTATCAACGCAGAGTACNNNNN,GTACTCTGCGTTGATACCACTGCTT ktrim=r k=15
    ↪   mink=15 hdist=1 qtrim=r trimq=20 minlength=30 threads=$THREADS || error_exit
    ↪   "$LINENO: Error removing SMARTer adapters from singletons - directory 2"

    #concatenate the singleton files
    cat ../../1_trimmedData/$FILESTEM2"_singletons1.fq.gz" >>
    ↪   ../../1_trimmedData/$FILESTEM2"_singletons_clean.fq.gz" || error_exit
    ↪   "$LINENO: Error concatenating singletons - directory 2"


    #clean up: i.e. remove temporary directories
    rm ../../1_trimmedData/$FILESTEM2"_singletons1.fq.gz" || error_exit "$LINENO:
    ↪   Error deleting singleton intermediate - directory 2"
done

cd /scratch/user/mahe0050/DE-analysis/1_trimmedData
rm -rf temp || error_exit "$LINENO: Error removing temp directory"
```

Quality of trimmed data was assessed again using FastQC v0.11.8.

```
    #run cleaned data through fastqc
    mkdir -p /scratch/user/mahe0050/DE-analysis/1_trimmedData/FastQC-clean

    for file in *R1_clean.fq.gz
    do
      echo $file
      FILESTEM2=`echo $file | cut -d "_" -f1,2 --output-delimiter="_"`
      echo "FILESTEM2" $FILESTEM2

      fastqc --noextract --threads $THREADS -o ./FastQC-clean
      ↪   ./$FILESTEM2"_R1_clean.fq.gz" ./$FILESTEM2"_R2_clean.fq.gz"
      ↪   ./$FILESTEM2"_singletons_clean.fq.gz" || error_exit "$LINENO: Error with
      ↪   FastQC at "$FILESTEM2""
    done


    #primers check
    #zcat G1_KI_HCGYYBCXY_ATCACG_L001_R1.fastq.gz | grep "AGATCGGAAGAGCAC" -m10
    #zcat G1_KI_HCGYYBCXY_ATCACG_L001_R1.fastq.gz | grep "GATCGGAAGAGCACA" -m10
    #zcat G1_KI_HCGYYBCXY_ATCACG_L001_R2.fastq.gz | grep "AGATCGGAAGAGCGT" -m10
```

The output html file of all eight samples FastQC reports AFTER trimming is available here.
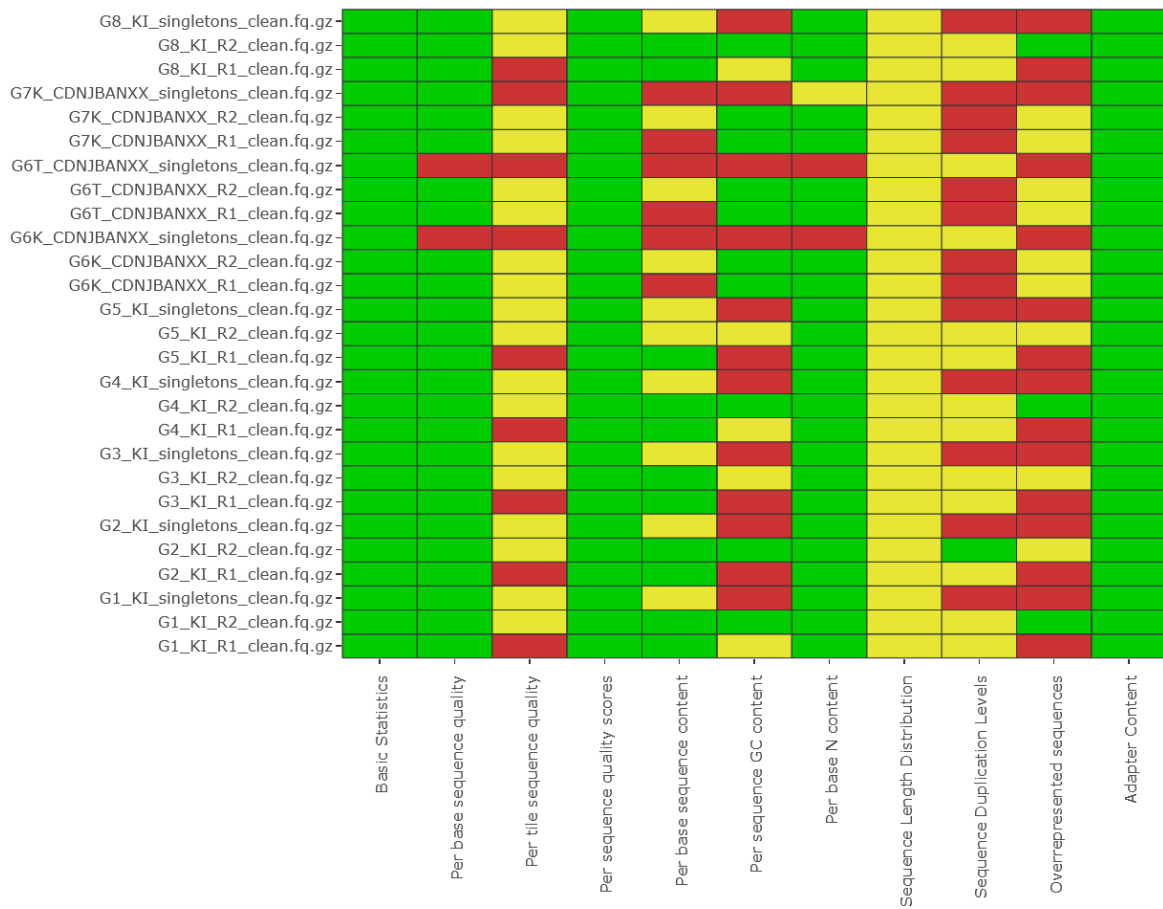
Figure 2: FastQC Summary of scores for paired reads of eight samples and resulting 'singleton' files for trimmed reads with no pair, after adapter and quality trimming in BBduk

2.2.3 De-novo assembly with TRINITY

De-novo assembly of transcripts.
Short reads were assembled *de novo* using the program TRINITY v2.8.4.

Sequencing run AGRF_CAGRF13871_HCGYYBCXY involved the synthesis of a non-stranded cDNA library prior to sequencing, while sequencing run AGRF_CAGRF20022_CDNJBANXX was completed with a different kit. All samples were treated as non stranded for consistency. Singleton reads (where a paired read was eliminated in quality trimming) were included for all TRINITY assembly runs.

Assembly was run on each sample individually as well as on a pooled (concatenated) file of all reads.

The script trinity2019.sh

Single samples were run through Trinity as per the linked script above using the following:

```
THREADS=16
MEM="64G"

cd /scratch/user/mahe0050/DE-analysis/1_trimmedData
```

To run trinity separately on each sample

```
mkdir -p ./trinity-sep
mkdir -p ./trinity-sep/temp

for file in *R1_clean.fq.gz
do
  FILESTEM=${file%_*}
  #this FILESTEM only cuts to _clean, the _R1 is included
  FILESTEM=${FILESTEM/R1/}
  #removes R1 from FILESTEM (FILESTEM ends in _  therefore not needed in "text"
  ↪  names)
```

concatenate relevant singleton file to R1 as library is not stranded?

```
cat ./$file ./$FILESTEM"singletons_clean.fq.gz" >
↪  ./trinity-sep/$FILESTEM"concat_R1.fq.gz" || error_exit "$LINENO: Error
↪  concatenating R1 and singletons"
```

Run TRINITY to assemble transcripts from a single paired sample file. Note trinity requires output directory with "trinity" in name.

```
Trinity --seqType fq --max_memory $MEM --bflyCalculateCPU --left
↪  ./trinity-sep/$FILESTEM"concat_R1.fq.gz" --right ./$FILESTEM"R2_clean.fq.gz"
↪  --CPU $THREADS --output ./trinity-sep/$FILESTEM"trinity"/trinity ||
↪  error_exit "$LINENO: Error running trinity-sep at $FILESTEM"
```

clean up

```
rm -rf ./trinity-sep/$FILESTEM"concat_R1.fq.gz" || error_exit "$LINENO: Error
↪  removing trinity-sep concat at $FILESTEM"

done
echo "Trinity separate samples complete"
```

All kidney samples were run through Trinity together for a single large assembly as per the linked script above above using the following:

Sample files to be concatenated and named appropriately

```
# create temp & working dir
mkdir -p ./trinity-all
mkdir -p ./trinity-all/temp
```

concatenate relevant sample files for one transcriptome - add singletons to R1 as its not stranded

```
cat ./*_R1_clean.fq.gz ./*_singletons_clean.fq.gz >>
→  ./trinity-all/temp/PBTKI_cat_R1.fq.gz || error_exit "$LINENO: Error
→  concatenating R1 and singletons"
cat ./*_R2_clean.fq.gz >> ./trinity-all/temp/PBTKI_cat_R2.fq.gz || error_exit
→  "$LINENO: Error concatenating R2"
```

Run TRINITY

```
Trinity --seqType fq --max_memory $MEM --bflyCalculateCPU --left
→  ./trinity-all/temp/PBTKI_cat_R1.fq.gz --right
→  ./trinity-all/temp/PBTKI_cat_R2.fq.gz --CPU $THREADS --output
→  ./trinity-all/trinity || error_exit "$LINENO: Error running Trinity-all"
```

clean up

```
rm -rf ./trinity-all/temp || error_exit "$LINENO: Error removing trinity-all temp
→  directory"
echo "Trinity concatenated kidney samples complete"
```

TRINITY outputs each samples assembly as Trinity.fasta, within a /trinity output folder, within the folder assigned with a sample name. In order to compare different Trinity.fasta results in one directory the following script was used to copy and rename these final files and place them in a common folder TrinityName.sh.

This renaming was done using a script to avoid any errors in manual handling of files and it is these moved files which are used later in a BUSCO comparison of assemblies.

2.3 *Summary / Quantification*

2.3.1 Long read sequences data summary

15,729 total transcript isoforms retained in initial cleaning and clustering of redundant isoforms.

13,882 sequences in a predicted open reading frame.

9,907 clusters identified based on translated proteins of predicted open reading frame.

9,813 full length transcripts were subset into a reference file as the longest representatives of one or more transcript clusters.

2.3.2 Short read sequences data summary

Total Transcript count of each TRINITY generated dataset was simply conducted using shell commands. Total transcript count for long read files are included in the BUSCO usage file below as headers required manipulation for input into BUSCO and totals were checked at this step.

Sample ID	Experimental factors	File name	Total transcripts/sequences assembled
G1	September Female	G1_KI_trinity	46538
G2	September Female	G2_KI_trinity	43283
G3	March/April Female	G3_KI_trinity	35991
G4	March/April Male	G4_KI_trinity	39239
G5	September Male	G5_KI_trinity	37379
G6	September Male	G6K_CDNJBANXX_trinity	214799
G7	March/April Female	G7K_CDNJBANXX_trinity	266539
G8	March/April Female	G8_KI_trinity	43050
All Samples combined	n/a	trinity-all	373287

2.4 BUSCO

Completeness of assemblies were assessed using Benchmarking Universal Single-Copy Orthologs BUSCO version 5.0 (BUSCO) and the dataset vertebrata_odb10.

Moving finalised assembly files into the BUSCO Directory:
TRINITY automatically names the final outputs "trinity.fasta" within folders which have the sample ID. Therefore for all these files to be moved into one directory for BUSCO to reference, these sequence IDs from the folders must be appended to the file names.

Conda Version 4.8.3 has been used for all analyses up to this point. Updated conda as version 4.8.4 or higher is recommended for BUSCO. This analysis was conducted on eRSA's NECTAR research cloud

Working directory of full (single line) hq transcript Isoseq dataset, and trimmed, non-redundant file of transcripts identified as cluster representatives.

```
/mnt/IsoSeq-analysis/data/Seqtk/hq.fasta.no5merge.collapsed.rep.1L.fa
/mnt/IsoSeq-analysis/data/Seqtk/reference_transcripts.1Lv.clean.fasta
```

Working directory of assembled (using TRINITY) short read transcript files

```
/mnt/DE-analysis/2_alignedData/trinity
```

concatenated short reads all assembled using trinity into largest transcript file /mnt/DE-analysis/2_alignedData/trinity/trinit
all/trinity

```
/mnt/DE-analysis/2_alignedData/trinity/trinity-all/trinity
# Then under
# /trinity-sep/<sample_ID_trinity>/trinity

  # Sample IDs folders:
  #   G1_KI_trinity
  #   G2_KI_trinity
  #   G3_KI_trinity
  #   G4_KI_trinity
  #   G5_KI_trinity
  #   G6K_CDNJBANXX_trinity
  #   G7K_CDNJBANXX_trinity
  #   G8_KI_trinity
```

This was done using the script output-BUSCO_Name2020.sh which also moved the long-read assemblies into the correct path:

```
#!/bin/bash
#
#
# This script is used to rename and copy trinity output files with their input file
↪    "filestem" into the /BUSCO directory so all files have similar naming
↪    convention for input into the BUSCO script.
#
# The paths in the script assume that a specific directory structure has been set
↪    up.
#
# Modules required: none
#
# usage:  bash Output-BUSCO_Name2020.sh
# location:            cd /scratch/user/mahe0050/BUSCO/

# Carmel Maher
# August 2020



#------adjust these for your run-----

#----------------------------------


function error_exit
{
    # Exit function due to fatal error
    # Accepts 1 arg:
    # string - descriptive error message
```

```bash
        echo "${PROGNAME}: ${1:-"Unknown error"}" 1>&2
        exit 1
}


# go to the working directory - working from /clean
cd /scratch/user/mahe0050/DE-analysis/1_trimmedData || error_exit "$LINENO:
↪   Directory Error 1"

for file in *R1_clean.fq.gz
  do

        FILESTEM=${file%_*}
        #this FILESTEM only cuts to _clean, the _R1 is included in stem
        FILESTEM=${FILESTEM/R1/}
        #removes R1 from FILESTEM (FILESTEM ends in _ therefore not needed in "text"
        ↪   names)

        cp -i /scratch/user/mahe0050/DE-analysis/2_alignedData/trinity-sep/
        ↪   $FILESTEM"trinity"/trinity/Trinity.fasta
        ↪   /scratch/user/mahe0050/BUSCO/$FILESTEM"t_assembled".fasta || error_exit
        ↪   "$LINENO: Error copying separate trinity outputs"

            # -> all files should end up in /scratch/user/mahe0050/BUSCO named
            ↪   G#_XX.t_assembled.fasta

  done

  echo "Trinity-sep done"

cp -i /scratch/user/mahe0050/DE-analysis/2_alignedData/trinity-all/trinity/
↪   Trinity.fasta /scratch/user/mahe0050/BUSCO/trinity-all_assembled.fasta ||
↪   error_exit "$LINENO: Error copying concatenated trinity-all output"

        # -> one ANGEL file should end up in /scratch/user/mahe0050/BUSCO named
        ↪   trinity-all_assembled.fasta

  echo "Trinity-all done"

cp -i /scratch/user/mahe0050/IsoSeq-analysis/data/Cogent/collected/
↪   hq.fasta.no5merge.collapsed.rep.fa
↪   /scratch/user/mahe0050/BUSCO/hq.fasta.no5merge.collapsed.rep_assembled.fasta
↪   || error_exit "$LINENO: Error copying Cogent output"

        # -> one Cogent file should end up in /scratch/user/mahe0050/BUSCO named
        ↪   hq.fasta.no5merge.collapsed.rep_assembled.fasta

  echo "Cogent done"

cp -i /scratch/user/mahe0050/IsoSeq-analysis/data/ANGEL/pygmy.ANGEL.cds
↪   /scratch/user/mahe0050/BUSCO/pygmy.ANGEL_assembled.fasta || error_exit
↪   "$LINENO: Error copying ANGEL output"
```

```
        # -> one ANGEL file should end up in /scratch/user/mahe0050/BUSCO named
    ↪    pygmy.ANGEL_assembled.fasta


    echo "ANGEL done"



# all file sizes checked after run = matches
# Trinity outputs removed from scratch 20/12/2020 to save space
```

Editing .fasta headers so that BUSCO can input the files

This BUSCO_all_v5.0.sh document contains notes and single lines run to run BUSCO 5.0 installed using the conda package on all short read and long read assembled transcript files. It outlines how to format the input files in preparation for BUSCO and the usage of the BUSCO script.

Sample format compatibility

```
cd /mnt/IsoSeq-analysis/BUSCO/

# NOTE the Isoseq files have symbols in fasta headers before transcript IDs which
↪    need to be removed for BUSCO
```

Header formats:

```
head /mnt/IsoSeq-analysis/data/Seqtk/hq.fasta.no5merge.collapsed.rep.1L.fa
>PB.4.2|004079|path1:6-612(+)|transcript/24311 transcript/24311
↪    full_length_coverage=2;length=640;num_subreads=60

head /mnt/IsoSeq-analysis/data/Seqtk/reference_transcripts.1Lv.clean.fasta
>PB.11.2|004815|path2:1047-3035(+)|transcript/13401 transcript/13401
↪    full_length_coverage=2;length=1995;num_subreads=8

head /mnt/IsoSeq-analysis/data/ANGEL/pygmy.ANGEL.cds
>PB.11.2|004815|path2:1047-3035(+)|transcript/13401|m.5 type:dumb-complete len:311
↪    strand:+ pos:214-1146
```

Total number of sequences:

```
grep -c ">" /mnt/IsoSeq-analysis/data/Seqtk/hq.fasta.no5merge.collapsed.rep.1L.fa
15729
grep -c ">" /mnt/IsoSeq-analysis/data/Seqtk/reference_transcripts.1Lv.clean.fasta
9813
grep -c ">" /mnt/IsoSeq-analysis/data/ANGEL/pygmy.ANGEL.cds
13882
```

after sed to remove all "/" resulted in an error at metaeuk headers "ValueError: could not convert string to float: '+'", after sed to remove all "+" same error.

Goal is a %score of alignments to BUSCOs and which transcripts match is no explored further. Therefore all headers were grossly simplified. The files created here are ONLY used for BUSCO.

```
cut -d '|' -f1
→  /mnt/IsoSeq-analysis/data/Seqtk/reference_transcripts.1Lv.clean.fasta >
→  reference_transcripts.1Lv.clean_.fasta

cut -d '|' -f1 /mnt/IsoSeq-analysis/data/ANGEL/pygmy.ANGEL.cds > pygmy.ANGEL_.cds

cut -d '|' -f1
→  /mnt/IsoSeq-analysis/data/Seqtk/hq.fasta.no5merge.collapsed.rep.1L.fa >
→  hq.fasta.no5merge.collapsed.rep.1L_.fa
```

Total number of sequences retained as counted above

```
grep -c ">" reference_transcripts.1Lv.clean_.fasta
9813
grep -c ">" pygmy.ANGEL_.cds
13882
grep -c ">" hq.fasta.no5merge.collapsed.rep.1L_.fa
15729
```

Two of these files will have "duplicate" headers in this simplified format. Numbers were simply appended as below so that all values are unique. These files are for use in BUSCO *only*:

```
awk '/^>/{$0=$0"_"(++i)}1'  pygmy.ANGEL_.cds > pygmy.ANGEL__.cds
awk '/^>/{$0=$0"_"(++i)}1'  hq.fasta.no5merge.collapsed.rep.1L_.fa >
→  hq.fasta.no5merge.collapsed.rep.1L__.fa

grep -c ">" pygmy.ANGEL__.cds
13882
grep -c ">" hq.fasta.no5merge.collapsed.rep.1L__.fa
15729
```

BUSCO Installation & usage:

The script run-conda-BUSCO.bash was used to run the program BUSCO on all samples on the NECTAR machine.

```
cd /mnt/IsoSeq-analysis/BUSCO

#Create a conda environment for the BUSCO Install

conda create -n conda-BUSCO -c bioconda -c conda-forge busco=5.0.0
conda activate conda-BUSCO
```

The basic repeated usage of BUSCO for all samples listed below will be:

```
busco -i [SEQUENCE_FILE] -l [LINEAGE] -o [OUTPUT_NAME] -m transcriptome


# note: E-value cut-off for BLAST searches Default: 1e-03

# busco --list-datasets
```

```
    #check $LINEAGE in script file
    #     - vertebrata_odb10
    #     may also --auto-lineage-euk

    # script currently set to threads (-c) = 12
```

Run BUSCO -> usage

```
    # location of usage doesnt matter as paths are specified within script.
    # use from within /mnt/IsoSeq-analysis/BUSCO/ in case of log files.
    # Usage:
            conda activate conda-BUSCO
            screen
            bash run-conda-BUSCO.bash 2>&1 | tee BUSCO-out.txt
```

run-conda-BUSCO.bash:

```
    #!/bin/bash
    #
    #
    # This script is used to assess transcriptome outputs against BUSCO lineages
    # The paths in the script assume that a specific directory structure has been set
    ↪   up.
    # Analysis run on eRSA NECTAR machine
    #
    #Usage:
    #           conda activate conda-BUSCO
    #        screen
    #           bash run-conda-BUSCO.bash 2>&1 | tee BUSCO-out.txt

    # Carmel Maher
    # February 2021


    # See notes file associated with usage: BUSCO_all_v5.0.sh
    # Inputs include:
    #    Short-read Trinity outputs for all 8 Kidney samples assembled from short
    ↪   reads
    #    Short-read Trinity output for one file of all 8 kidney sequencing
    ↪   concatenated and assembled into a single transcript file
    #    Isoseq3 long-read de novo assembly full list of hq transcripts
    #    Isoseq3 long-read de novo assembly list of non-redundant representative
    ↪   transcripts of peptide clusters, with additional poly-a tail trimming


    #------adjust these for your run-----

    LINEAGE="vertebrata_odb10"

    #----------------------------------

    function error_exit
```

```
{
    echo "${PROGNAME}: ${1:-"Unknown error"}" 1>&2
    exit 1
}

# (commented out because it failed after "done" - rerun with # )

#go to the directory containing trimmed files to pull ID names
cd /mnt/DE-analysis/1_trimmedData

for file in *R1_clean.fq.gz
  do
      FILESTEM=${file%_*}
      #this FILESTEM only cuts to _clean, the _R1 is included in stem
      FILESTEM=${FILESTEM/R1/}
      #removes R1 from FILESTEM (FILESTEM ends in _ therefore not needed in "text"
      ↪   names)

      mkdir /mnt/IsoSeq-analysis/BUSCO/$FILESTEM"BUSCOout" || error_exit "$LINENO:
      ↪   Error creating trinity-sep output directory at $FILESTEM"

      cd /mnt/IsoSeq-analysis/BUSCO/

      busco -i
      ↪   /mnt/DE-analysis/2_alignedData/trinity/trinity-sep/$FILESTEM"trinity"/
      ↪   trinity/Trinity.fasta -l $LINEAGE -f -c 12 -o $FILESTEM"BUSCOout" -m
      ↪   transcriptome || error_exit "$LINENO: Error running BUSCO at $FILESTEM"

      cd /mnt/DE-analysis/1_trimmedData
  done

#

cd /mnt/IsoSeq-analysis/BUSCO/

mkdir ./trinity-all_BUSCOout  || error_exit "$LINENO: directory error at
↪   trinity-all"

busco -i /mnt/DE-analysis/2_alignedData/trinity/trinity-all/trinity/Trinity.fasta
↪   -l $LINEAGE -f -c 12 -o trinity-all_BUSCOout -m transcriptome || error_exit
↪   "$LINENO: Error running BUSCO at trinity-all"

#

cd /mnt/IsoSeq-analysis/BUSCO/

mkdir ./reference-transcripts_BUSCOout || error_exit "$LINENO: directory error at
↪   reference-transcripts"

busco -i reference_transcripts.1Lv.clean_.fasta -l $LINEAGE -f -c 12 -o
↪   reference-transcripts_BUSCOout -m transcriptome || error_exit "$LINENO: Error
↪   running BUSCO at reference-transcripts"
```

```
          #

          mkdir ./ANGEL.cds_BUSCOout || error_exit "$LINENO: directory error at
          ↪   reference-transcripts"

          busco -i pygmy.ANGEL__.cds -l $LINEAGE -f -c 12 -o ANGEL.cds_BUSCOout -m
          ↪   transcriptome || error_exit "$LINENO: Error running BUSCO at ANGEL.cds"

          #

          mkdir ./hq-fasta_BUSCOout || error_exit "$LINENO: directory error at hq-fasta"

          busco -i hq.fasta.no5merge.collapsed.rep.1L__.fa -l $LINEAGE -f -c 12 -o
          ↪   hq-fasta_BUSCOout -m transcriptome || error_exit "$LINENO: Error running BUSCO
          ↪   at hq-fasta"

          #

          echo "all BUSCOs complete"
```

Visualisation

BUSCO outputs a file named short_summary.specific.vertebrata_odb10.<-o>.txt. These are used to plot
a summary figure of all samples.

generate plot.py and r script of all sample BUSCO summaries

```
          mkdir BUSCO_summaries

          cp G1_KI_BUSCOout/short_summary.*_odb10.G1_KI_BUSCOout.txt BUSCO_summaries/.
          cp G2_KI_BUSCOout/short_summary.*_odb10.G2_KI_BUSCOout.txt BUSCO_summaries/.
          cp G3_KI_BUSCOout/short_summary.*_odb10.G3_KI_BUSCOout.txt BUSCO_summaries/.
          cp G4_KI_BUSCOout/short_summary.*_odb10.G4_KI_BUSCOout.txt BUSCO_summaries/.
          cp G5_KI_BUSCOout/short_summary.*_odb10.G5_KI_BUSCOout.txt BUSCO_summaries/.
          cp G6K_CDNJBANXX_BUSCOout/short_summary.*_odb10.G6K_CDNJBANXX_BUSCOout.txt
          ↪   BUSCO_summaries/.
          cp G7K_CDNJBANXX_BUSCOout/short_summary.*_odb10.G7K_CDNJBANXX_BUSCOout.txt
          ↪   BUSCO_summaries/.
          cp G8_KI_BUSCOout/short_summary.*_odb10.G8_KI_BUSCOout.txt BUSCO_summaries/.
          cp reference-transcripts_BUSCOout/
          ↪   short_summary.*_odb10.reference-transcripts_BUSCOout.txt BUSCO_summaries/.
          cp ANGEL.cds_BUSCOout/short_summary.*_odb10.ANGEL.cds_BUSCOout.txt
          ↪   BUSCO_summaries/.
          cp hq-fasta_BUSCOout/short_summary.*_odb10.hq-fasta_BUSCOout.txt BUSCO_summaries/.
          cp trinity-all_BUSCOout/short_summary.*_odb10.trinity-all_BUSCOout.txt
          ↪   BUSCO_summaries/.

          #(The path below is where the config files for conda-BUSCO environment are located)

          python3 /mnt/Prog/miniconda3/envs/conda-BUSCO/bin/generate_plot.py -wd
          ↪   /mnt/IsoSeq-analysis/BUSCO/BUSCO_summaries
```

Short Read summaries:

```
mkdir BUSCO_SR_summaries

cp G1_KI_BUSCOout/short_summary.*_odb10.G1_KI_BUSCOout.txt BUSCO_SR_summaries/.
cp G2_KI_BUSCOout/short_summary.*_odb10.G2_KI_BUSCOout.txt BUSCO_SR_summaries/.
cp G3_KI_BUSCOout/short_summary.*_odb10.G3_KI_BUSCOout.txt BUSCO_SR_summaries/.
cp G4_KI_BUSCOout/short_summary.*_odb10.G4_KI_BUSCOout.txt BUSCO_SR_summaries/.
cp G5_KI_BUSCOout/short_summary.*_odb10.G5_KI_BUSCOout.txt BUSCO_SR_summaries/.
cp G6K_CDNJBANXX_BUSCOout/short_summary.*_odb10.G6K_CDNJBANXX_SR_BUSCOout.txt
↪   BUSCO_summaries/.
cp G7K_CDNJBANXX_BUSCOout/short_summary.*_odb10.G7K_CDNJBANXX_SR_BUSCOout.txt
↪   BUSCO_summaries/.
cp G8_KI_BUSCOout/short_summary.*_odb10.G8_KI_BUSCOout.txt BUSCO_SR_summaries/.
cp trinity-all_BUSCOout/short_summary.*_odb10.trinity-all_BUSCOout.txt
↪   BUSCO_SR_summaries/.

python3 /mnt/Prog/miniconda3/envs/conda-BUSCO/bin/generate_plot.py -wd
↪   /mnt/IsoSeq-analysis/BUSCO/BUSCO_SR_summaries
```

Long Read summaries:

```
mkdir BUSCO_LR_summaries

cp reference-transcripts_BUSCOout/
↪   short_summary.*_odb10.reference-transcripts_BUSCOout.txt BUSCO_LR_summaries/.
cp ANGEL.cds_BUSCOout/short_summary.*_odb10.ANGEL.cds_BUSCOout.txt
↪   BUSCO_LR_summaries/.
cp hq-fasta_BUSCOout/short_summary.*_odb10.hq-fasta_BUSCOout.txt
↪   BUSCO_LR_summaries/.

python3 /mnt/Prog/miniconda3/envs/conda-BUSCO/bin/generate_plot.py -wd
↪   /mnt/IsoSeq-analysis/BUSCO/BUSCO_LR_summaries
```

These files for each sample were also moved into an R working directory to allow the generated figure to be edited.
The R script produced by the python command above was edited to allow for local directory structure and streamline the labelling and order of sample names.
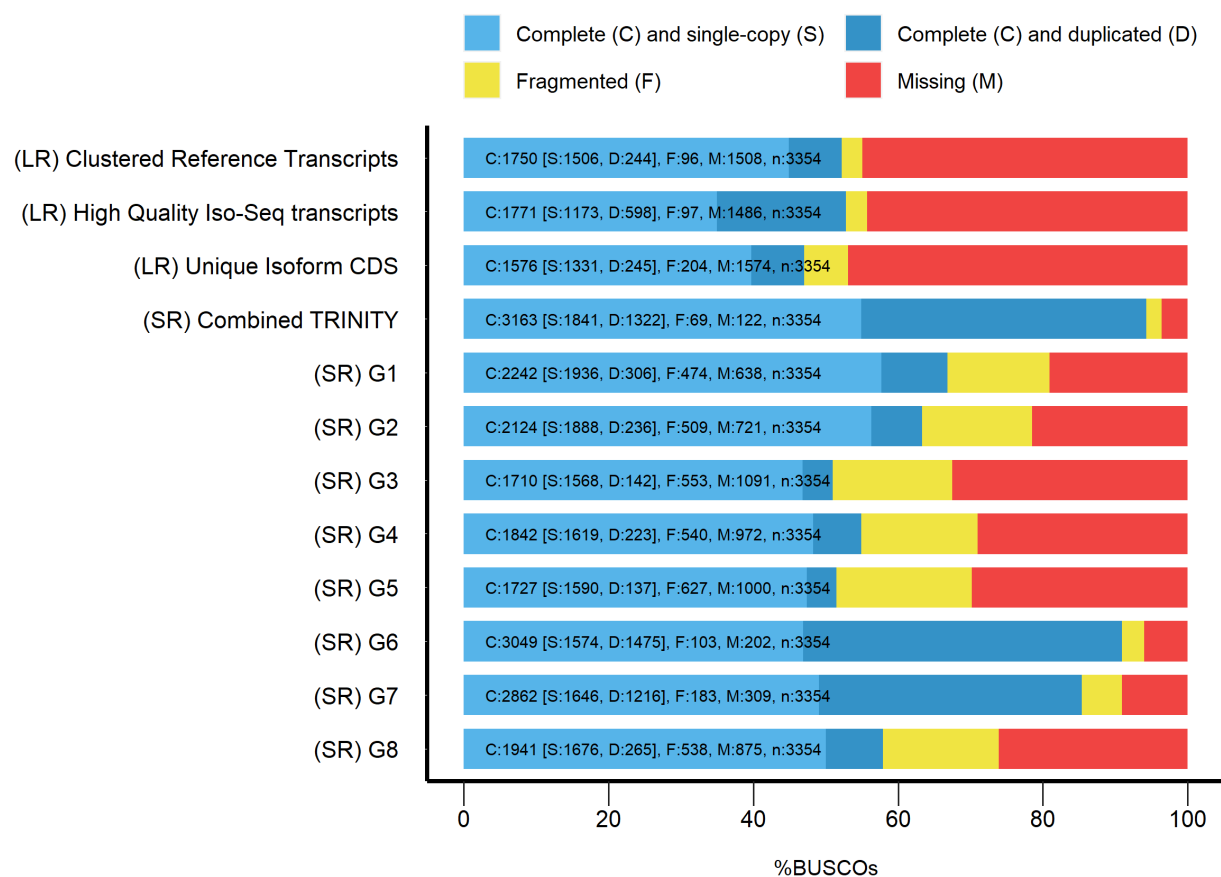
Figure 3: Busco Summary Statistics (BUSCO v5.0.0 run on database vertebrata_odb10. LR indicates transcript databases at different assembly steps, assembled from long read sequencing data. SR indicates short read sequencing sample data assembled with TRINITY)

3 Annotation (Chapter 5)

Completed on the long read data from Chapter 4

- pygmy.ANGEL.cds
- pygmy.ANGEL.pep
- pygmy.ANGEL.utr

3.1 BLASTx

BLASTx searches were conducted on the dataset at this stage to include isoforms that are collapsed later on.

BLASTx was run on the pygmy.ANGEL.cds file as it contains the predicted open reading frame sequence of nucleotides for translateable proteins. This was completed on the NECTAR machine.

```
# Package: blast 2.9.0, build Mar 11 2019 15:20:05
# Uniprot sprot & trembl databases downloaded: 28/5/19     (note: trembl database
↪   not indexed or used)
# Anolis .pep fasta files downloaded 12/11/19
        Anolis_carolinensis.AnoCar2.0.pep.all.fa
        Anolis_carolinensis.AnoCar2.0.pep.abinitio.fa

# To format a database: #
# makeblastdb -in mydb.fsa -dbtype nucl -parse_seqids
```

UniProt SwissProt database:

```
/mnt/Prog/blast/ncbi-blast-2.9.0+/bin/makeblastdb -in uniprot_sprot.fasta
↪   -parse_seqids -blastdb_version 5 -title "sprot" -dbtype prot -out sprot
```

Anolis carolinensis proteins database:

```
#AnoCar2.0.pep.all
/mnt/Prog/blast/ncbi-blast-2.9.0+/bin/makeblastdb -in
↪   Anolis_carolinensis.AnoCar2.0.pep.all.fa -parse_seqids -blastdb_version 5 -title
↪   "sprot" -dbtype prot -out AnoCar2.0.pep.all
```

Anolis Carolinensis database *ab initio*:

```
#AnoCar2.0.pep.abinitio
/mnt/Prog/blast/ncbi-blast-2.9.0+/bin/makeblastdb -in
↪   Anolis_carolinensis.AnoCar2.0.pep.abinitio.fa -parse_seqids -blastdb_version 5
↪   -title "sprot" -dbtype prot -out AnoCar2.0.pep.abinitio
```

Various searches were run in a screen to examine the effects of parameters and databases. Final searches are as listed below.

Searches included all databases listed above and variations of the parameters: -max_target_seqs 5 or -max_target_seqs 1 -max_hsps 5 or -max_hsps 1 -evalue 0.00001 or -evalue 1e-10

The NCBI Blast results most often referenced in the thesis are final parameters blastx -fmt6 -maxtarget1 -maxhsps1 -eval0.00001 against the UniProt SwissProt database.

```
# uniprot sprot
/mnt/Prog/blast/ncbi-blast-2.9.0+/bin/blastx -db
↪  /mnt/Prog/blast/blastdb/sprot/sprot -query
↪  /mnt/IsoSeq-analysis/data/ANGEL/pygmy.ANGEL.cds -out
↪  pygmy.ANGEL_blastx_sprot-maxtarg1-maxhsp1.out -outfmt 6 -max_target_seqs 1
↪  -max_hsps 1 -evalue 0.00001 -num_threads 3
```

```
#AnoCar2.0.pep.all
/mnt/Prog/blast/ncbi-blast-2.9.0+/bin/blastx -db
↪  /mnt/Prog/blast/blastdb/AcarProt/AnoCar2.0.pep.all/AnoCar2.0.pep.all -query
↪  /mnt/IsoSeq-analysis/data/ANGEL/pygmy.ANGEL.cds -out
↪  pygmy.ANGEL_blastx_AnoCar.pep.all-maxtarg1-maxhsp1.out -outfmt 6
↪  -max_target_seqs 1 -max_hsps 1 -evalue 0.00001 -num_threads 3
```

All searches except the below were output in tab separated -outfmt 6. Due to import requirements an -outfmt 5 .xml file was required for results to be visualised in the program BLAST2GO (-max_target_seqs 5 -max_hsps 5 -evalue 0.00001).

```
/mnt/Prog/blast/ncbi-blast-2.9.0+/bin/blastx -db
↪  /mnt/Prog/blast/blastdb/AcarProt/AnoCar2.0.pep.all/AnoCar2.0.pep.all -query
↪  /mnt/IsoSeq-analysis/data/ANGEL/pygmy.ANGEL.cds -out
↪  pygmy.ANGEL_blastx_AnoCar.pep.all-BLAST2GO -outfmt 5 -max_target_seqs 5
↪  -max_hsps 5 -evalue 0.00001 -num_threads 4

/mnt/Prog/blast/ncbi-blast-2.9.0+/bin/blastx -db
↪  /mnt/Prog/blast/blastdb/sprot/sprot -query
↪  /mnt/IsoSeq-analysis/data/ANGEL/pygmy.ANGEL.cds -out
↪  pygmy.ANGEL_blastx_sprot-BLAST2GO -outfmt 5 -max_target_seqs 5 -max_hsps 5
↪  -evalue 0.00001 -num_threads 4
```

The above and other trial searches are all listed in the file IsoSeq analysis post-ANGEL blast notes.sh

3.2 Gene Name IDs assigned to Transcript ID

The ANGEL predicted open reading frame transcripts (Section 2.1.3) were searched against the UniProt swiss prot database using BLASTx with parameters: -max_target_seqs 1 -max_hsps 1 -evalue 0.00001) (Section 3.1 above).
The BLASTx results are uploaded here.

Protein IDs from the BLASTx results were uploaded to the Uni Prot Retrieve/ID mapping tool and mapped to Gene names. This produced a list of unique UniProt swiss prot identifiers and their corresponding gene name ID.

3.2.1 Full transcript putative gene reference file

Using dplyr the BLAST results and these gene ID mappings were joined and filtered to create a reference file of transcript IDs assigned both a Uni Prot Swiss prot database protein ID, and a gene name.

The BLASTx results were imported:

```
setwd("~/@Uni/@Flinders University
↪  PhD/RWorkingDir/Workflow-readthedown/ReferenceClusters")

Unisprot.BLASTx.out <- read.csv("pygmy.ANGEL_blastx_sprot-maxtarg1-maxhsp1.out",
    sep = "\t", header = FALSE)
head(Unisprot.BLASTx.out)
```

```
##                                                             V1      V2      V3  V4  V5
## 1     PB.2.1|002537|path0:1-1624(+)|transcript/18304|m.1 Q6PBC3 95.000 100    5
## 2    PB.3.2|00361c|path0:43-2240(+)|transcript/10564|m.2 P55209 88.995 209   22
## 3     PB.11.1|004815|path2:1-3039(+)|transcript/3426|m.4 P36633 61.892 719 258
## 4 PB.11.2|004815|path2:1047-3035(+)|transcript/13401|m.5 Q8JZQ5 67.010 194   63
## 5    PB.12.1|004815|path5:1-1665(+)|transcript/17167|m.6 Q8JZQ5 61.712 222   80
## 6     PB.13.1|004c10|path2:1-2811(+)|transcript/4993|m.7 Q9H063 83.984 256   38
##   V6  V7   V8  V9 V10       V11 V12
## 1  0   1  300   1 100  2.49e-69 207
## 2  1   1  627 184 391  5.86e-61 198
## 3  8  46 2175  32 743  0.00e+00 855
## 4  1 196  777 513 705  3.11e-92 246
## 5  3   1  666 532 748  3.96e-77 249
## 6  1   1  759   1 256 4.61e-156 437
```

```
nrow(Unisprot.BLASTx.out)
```

```
## [1] 12894
```

Column Headers Renamed to match BLAST output format 6:

```
Unisprot.BLASTx.out <- Unisprot.BLASTx.out %>%
    rename(TranscriptID = V1, UniprotID = V2, pident = V3, length = V4,
        mismatch = V5, gapopen = V6, qstart = V7, qend = V8,
        sstart = V9, send = V10, evalue = V11, bitscore = V12)

colnames(Unisprot.BLASTx.out)
```

```
## [1] "TranscriptID" "UniprotID"     "pident"         "length"        "mismatch"
## [6] "gapopen"      "qstart"        "qend"           "sstart"        "send"
## [11] "evalue"      "bitscore"
```

head(Unisprot.BLASTx.out)

```
##                                                   TranscriptID UniprotID pident
## 1      PB.2.1|002537|path0:1-1624(+)|transcript/18304|m.1     Q6PBC3 95.000
## 2     PB.3.2|00361c|path0:43-2240(+)|transcript/10564|m.2     P55209 88.995
## 3      PB.11.1|004815|path2:1-3039(+)|transcript/3426|m.4     P36633 61.892
## 4 PB.11.2|004815|path2:1047-3035(+)|transcript/13401|m.5     Q8JZQ5 67.010
## 5     PB.12.1|004815|path5:1-1665(+)|transcript/17167|m.6     Q8JZQ5 61.712
## 6      PB.13.1|004c10|path2:1-2811(+)|transcript/4993|m.7     Q9H063 83.984
##    length mismatch gapopen qstart qend sstart send      evalue bitscore
## 1     100        5       0      1  300      1  100    2.49e-69      207
## 2     209       22       1      1  627    184  391    5.86e-61      198
## 3     719      258       8     46 2175     32  743    0.00e+00      855
## 4     194       63       1    196  777    513  705    3.11e-92      246
## 5     222       80       3      1  666    532  748    3.96e-77      249
## 6     256       38       1      1  759      1  256   4.61e-156      437
```

tail(Unisprot.BLASTx.out)

```
##                                                   TranscriptID UniprotID
## 12889    PB.9915.1|transcript/993:1-4136(+)|transcript/993|m.15721     Q99PI4
## 12890 PB.9916.1|transcript/9931:1-2240(+)|transcript/9931|m.15722     Q9C0E8
## 12891 PB.9918.1|transcript/9940:1-2240(+)|transcript/9940|m.15723     P11413
## 12892 PB.9919.1|transcript/9966:1-2239(+)|transcript/9966|m.15724     O08688
## 12893 PB.9921.1|transcript/9979:1-2232(+)|transcript/9979|m.15725     A2VDM0
## 12894 PB.9923.1|transcript/9981:1-2263(+)|transcript/9981|m.15727     Q2HJ89
##       pident length mismatch gapopen qstart qend sstart send      evalue bitscore
## 12889 70.083    361      103       2      1 1083    493  848    0.00e+00      530
## 12890 73.729    236       57       3      1  708      1  231   2.53e-118      349
## 12891 85.156    512       76       0    127 1662      4  515    0.00e+00      935
## 12892 78.193    642      137       2     70 1992      1  640    0.00e+00     1070
## 12893 46.512    172       87       1    139  639     14  185    2.57e-48      159
## 12894 76.534    554      125       2     58 1704     30  583    0.00e+00      852
```

Gene ID mapping list for Uniprot IDs were imported:

```
BLASTx.to.gene <- read.csv("~/@Uni/@Flinders University
↪   PhD/RWorkingDir/Workflow-readthedown/ReferenceClusters/pygmy.ANGEL_blastx_sprot-maxtarg1-maxhsp1==G
↪   Name IDs.txt",
    sep = "\t", header = TRUE)
```

head(BLASTx.to.gene)

```
##      From      To
## 1 Q6PBC3 chchd4
## 2 P55209 NAP1L1
## 3 P36633   Aoc1
```

```
## 4 Q8JZQ5   Aoc1
## 5 Q9H063   MAF1
## 6 A1XBS5 FAM92A
```

Column Headers were renamed:

```
BLASTx.to.gene <- BLASTx.to.gene %>%
    rename(UniprotID = From, Gene = To)
colnames(BLASTx.to.gene)
```

```
## [1] "UniprotID" "Gene"
```

Although The Uni Prot database returns a non-redundant, unique list, conventions of gene and protein names of different species means there is some duplication when treated as case sensitive. Here all protein IDs will be made upper case, and all gene IDs will be made lower case.

```
BLASTx.to.gene <- BLASTx.to.gene %>%
    mutate(UniprotID = toupper(UniprotID))

BLASTx.to.gene <- BLASTx.to.gene %>%
    mutate(Gene = tolower(Gene))

head(BLASTx.to.gene)
```

```
##   UniprotID   Gene
## 1    Q6PBC3 chchd4
## 2    P55209 nap1l1
## 3    P36633   aoc1
## 4    Q8JZQ5   aoc1
## 5    Q9H063   maf1
## 6    A1XBS5 fam92a
```

```
nrow(BLASTx.to.gene)
```

```
## [1] 7221
```

Note that nrow = 7221, now the duplicated rows will be filtered out:

```
BLASTx.to.gene %>%
    count(UniprotID) %>%
    filter(n > 1)
```

```
##   UniprotID n
## 1    P32969 4
```

Uni prot ID P32969 corresponds to gene IDs: RPL9, RPL9P7, RPL9P8, and RPL9P9. As these data are not specifically assigned to any one of these 4 genes as they all have the same protein ID, and to avoid duplication of transcripts on joining, this will be simplified to one entry of RPL9.

```
BLASTx.to.gene <- BLASTx.to.gene %>%
    distinct(UniprotID, .keep_all = TRUE)
nrow(BLASTx.to.gene)
```

```
## [1] 7218
```

```
BLASTx.to.gene %>%
    group_by(Gene) %>%
    summarize(n = n())
```

```
## # A tibble: 6,343 x 2
##     Gene        n
##     <chr>    <int>
##  1 aaas         1
##  2 aadacl4      1
##  3 aadat        2
##  4 aagab        1
##  5 aamp         1
##  6 aanat        1
##  7 aar2         1
##  8 aars1        1
##  9 aarsd1       1
## 10 aass         1
## # ... with 6,333 more rows
```

There are now 7218 Unique Uniprot IDs listed and after making case consistent, 6343 unique Gene IDs listed. Multiple Protein IDs map to the same Gene ID. This may be due to data sourced from a variety of species' and particular transcripts returning a BLAST hit for gene orthologues, or gene haplotypes. These putative Gene IDs are for reference only, and protein IDs will not be removed from transcript data, so no information will be lost.

Assign gene ID to the BLAST result by joining based on Uni Prot ID

```
BLASTx.gene.join <- left_join(Unisprot.BLASTx.out, BLASTx.to.gene,
    by = "UniprotID")
head(BLASTx.gene.join)
```

```
##                                               TranscriptID UniprotID pident
## 1      PB.2.1|002537|path0:1-1624(+)|transcript/18304|m.1    Q6PBC3 95.000
## 2     PB.3.2|00361c|path0:43-2240(+)|transcript/10564|m.2    P55209 88.995
## 3      PB.11.1|004815|path2:1-3039(+)|transcript/3426|m.4    P36633 61.892
## 4 PB.11.2|004815|path2:1047-3035(+)|transcript/13401|m.5    Q8JZQ5 67.010
## 5     PB.12.1|004815|path5:1-1665(+)|transcript/17167|m.6    Q8JZQ5 61.712
## 6      PB.13.1|004c10|path2:1-2811(+)|transcript/4993|m.7    Q9H063 83.984
##    length mismatch gapopen qstart qend sstart send     evalue bitscore   Gene
## 1     100        5       0      1  300      1  100   2.49e-69      207 chchd4
## 2     209       22       1      1  627    184  391   5.86e-61      198 nap1l1
## 3     719      258       8     46 2175     32  743   0.00e+00      855   aoc1
## 4     194       63       1    196  777    513  705   3.11e-92      246   aoc1
## 5     222       80       3      1  666    532  748   3.96e-77      249   aoc1
## 6     256       38       1      1  759      1  256  4.61e-156      437   maf1
```

```
BLASTx.gene.join <- relocate(BLASTx.gene.join, Gene, .before = pident)
head(BLASTx.gene.join)
```

```
##                                              TranscriptID UniprotID   Gene
## 1      PB.2.1|002537|path0:1-1624(+)|transcript/18304|m.1    Q6PBC3 chchd4
## 2     PB.3.2|00361c|path0:43-2240(+)|transcript/10564|m.2    P55209 nap1l1
## 3      PB.11.1|004815|path2:1-3039(+)|transcript/3426|m.4    P36633   aoc1
## 4 PB.11.2|004815|path2:1047-3035(+)|transcript/13401|m.5    Q8JZQ5   aoc1
## 5     PB.12.1|004815|path5:1-1665(+)|transcript/17167|m.6    Q8JZQ5   aoc1
## 6      PB.13.1|004c10|path2:1-2811(+)|transcript/4993|m.7    Q9H063   maf1
##    pident length mismatch gapopen qstart qend sstart send    evalue bitscore
## 1 95.000    100        5       0      1  300      1  100  2.49e-69      207
## 2 88.995    209       22       1      1  627    184  391  5.86e-61      198
## 3 61.892    719      258       8     46 2175     32  743  0.00e+00      855
## 4 67.010    194       63       1    196  777    513  705  3.11e-92      246
## 5 61.712    222       80       3      1  666    532  748  3.96e-77      249
## 6 83.984    256       38       1      1  759      1  256 4.61e-156      437
```

```
tail(BLASTx.gene.join)
```

```
##                                               TranscriptID UniprotID
## 12889   PB.9915.1|transcript/993:1-4136(+)|transcript/993|m.15721    Q99PI4
## 12890 PB.9916.1|transcript/9931:1-2240(+)|transcript/9931|m.15722    Q9C0E8
## 12891 PB.9918.1|transcript/9940:1-2240(+)|transcript/9940|m.15723    P11413
## 12892 PB.9919.1|transcript/9966:1-2239(+)|transcript/9966|m.15724    O08688
## 12893 PB.9921.1|transcript/9979:1-2232(+)|transcript/9979|m.15725    A2VDM0
## 12894 PB.9923.1|transcript/9981:1-2263(+)|transcript/9981|m.15727    Q2HJ89
##         Gene pident length mismatch gapopen qstart qend sstart send     evalue
## 12889  lpin3 70.083    361      103       2      1 1083    493  848   0.00e+00
## 12890   lnpk 73.729    236       57       3      1  708      1  231 2.53e-118
## 12891   g6pd 85.156    512       76       0    127 1662      4  515   0.00e+00
## 12892  capn5 78.193    642      137       2     70 1992      1  640   0.00e+00
## 12893   tifa 46.512    172       87       1    139  639     14  185   2.57e-48
## 12894 fkbp10 76.534    554      125       2     58 1704     30  583   0.00e+00
##       bitscore
## 12889      530
## 12890      349
## 12891      935
## 12892     1070
## 12893      159
## 12894      852
```

3.2.1.1 Manipulation of 'BLASTx.gene.join_filt_clustered' to match reference files

- The fasta file which contains *only* protein coding regions (and sequences which were part of the input for the BLASTx search) has the characters "|m.*" at the end of transcript IDs
- The fasta file containing the full length of these transcripts corresponding to the clustered protein coding regions (above), and which was used as the reference file for Kallisto does not have this name extension

In order to be able to filter this file based on either naming convention and compare to BLASTx outputs, genes of interest lists, and gene expression results: Here a corresponding column with the "|m.*" removed is added

Duplicate the ID column

```
BLASTx.gene.join_split <- cbind(BLASTx.gene.join, replicate(1,
    BLASTx.gene.join$TranscriptID))
BLASTx.gene.join_split <- rename(BLASTx.gene.join_split, TranscriptID_2 = "replicate(1,
↪  BLASTx.gene.join$TranscriptID)")
names(BLASTx.gene.join_split)
```

```
##  [1] "TranscriptID"    "UniprotID"      "Gene"           "pident"
##  [5] "length"          "mismatch"       "gapopen"        "qstart"
##  [9] "qend"            "sstart"         "send"           "evalue"
## [13] "bitscore"        "TranscriptID_2"
```

```
BLASTx.gene.join_split <- relocate(BLASTx.gene.join_split, "TranscriptID_2",
    .before = "UniprotID")
head(BLASTx.gene.join_split)
```

```
##                                                    TranscriptID
## 1     PB.2.1|002537|path0:1-1624(+)|transcript/18304|m.1
## 2     PB.3.2|00361c|path0:43-2240(+)|transcript/10564|m.2
## 3      PB.11.1|004815|path2:1-3039(+)|transcript/3426|m.4
## 4 PB.11.2|004815|path2:1047-3035(+)|transcript/13401|m.5
## 5     PB.12.1|004815|path5:1-1665(+)|transcript/17167|m.6
## 6      PB.13.1|004c10|path2:1-2811(+)|transcript/4993|m.7
##                                                    TranscriptID_2 UniprotID    Gene
## 1     PB.2.1|002537|path0:1-1624(+)|transcript/18304|m.1    Q6PBC3 chchd4
## 2     PB.3.2|00361c|path0:43-2240(+)|transcript/10564|m.2    P55209 nap1l1
## 3      PB.11.1|004815|path2:1-3039(+)|transcript/3426|m.4    P36633   aoc1
## 4 PB.11.2|004815|path2:1047-3035(+)|transcript/13401|m.5    Q8JZQ5   aoc1
## 5     PB.12.1|004815|path5:1-1665(+)|transcript/17167|m.6    Q8JZQ5   aoc1
## 6      PB.13.1|004c10|path2:1-2811(+)|transcript/4993|m.7    Q9H063   maf1
##    pident length mismatch gapopen qstart qend sstart send      evalue bitscore
## 1 95.000    100        5       0      1  300      1  100   2.49e-69      207
## 2 88.995    209       22       1      1  627    184  391   5.86e-61      198
## 3 61.892    719      258       8     46 2175     32  743   0.00e+00      855
## 4 67.010    194       63       1    196  777    513  705   3.11e-92      246
## 5 61.712    222       80       3      1  666    532  748   3.96e-77      249
## 6 83.984    256       38       1      1  759      1  256 4.61e-156      437
```

Split the Transcript ID column based on the last "|" and return as a data frame

```
BLASTx.gene.join_split2 <-
↪  lapply(stri_split_regex(stri_reverse(BLASTx.gene.join_split$TranscriptID_2),
    pattern = "[|\\s]+", n = 2), stri_reverse)
BLASTx.gene.join_split2 <- setNames(data.table::transpose(BLASTx.gene.join_split2)[2:1],
    c("output1", "output2"))
BLASTx.gene.join_split2 <- as.data.frame(c(list(input =
↪  BLASTx.gene.join_split$TranscriptID_2),
```

```
    BLASTx.gene.join_split2))
head(BLASTx.gene.join_split2)
```

```
##                                                              input
## 1       PB.2.1|002537|path0:1-1624(+)|transcript/18304|m.1
## 2      PB.3.2|00361c|path0:43-2240(+)|transcript/10564|m.2
## 3       PB.11.1|004815|path2:1-3039(+)|transcript/3426|m.4
## 4 PB.11.2|004815|path2:1047-3035(+)|transcript/13401|m.5
## 5      PB.12.1|004815|path5:1-1665(+)|transcript/17167|m.6
## 6       PB.13.1|004c10|path2:1-2811(+)|transcript/4993|m.7
##                                                      output1 output2
## 1       PB.2.1|002537|path0:1-1624(+)|transcript/18304     m.1
## 2      PB.3.2|00361c|path0:43-2240(+)|transcript/10564     m.2
## 3       PB.11.1|004815|path2:1-3039(+)|transcript/3426     m.4
## 4 PB.11.2|004815|path2:1047-3035(+)|transcript/13401     m.5
## 5      PB.12.1|004815|path5:1-1665(+)|transcript/17167     m.6
## 6       PB.13.1|004c10|path2:1-2811(+)|transcript/4993     m.7
```

Rename Columns and Join the truncated columns back into the original dataframe based on TranscriptID_2

```
names(BLASTx.gene.join_split2)
```

```
## [1] "input"     "output1" "output2"
```

```
colnames(BLASTx.gene.join_split2) <- c("TranscriptID_2", "TranscriptID_3",
    "TranscriptID_4")
head(BLASTx.gene.join_split2)
```

```
##                                                     TranscriptID_2
## 1       PB.2.1|002537|path0:1-1624(+)|transcript/18304|m.1
## 2      PB.3.2|00361c|path0:43-2240(+)|transcript/10564|m.2
## 3       PB.11.1|004815|path2:1-3039(+)|transcript/3426|m.4
## 4 PB.11.2|004815|path2:1047-3035(+)|transcript/13401|m.5
## 5      PB.12.1|004815|path5:1-1665(+)|transcript/17167|m.6
## 6       PB.13.1|004c10|path2:1-2811(+)|transcript/4993|m.7
##                                                     TranscriptID_3 TranscriptID_4
## 1       PB.2.1|002537|path0:1-1624(+)|transcript/18304          m.1
## 2      PB.3.2|00361c|path0:43-2240(+)|transcript/10564          m.2
## 3       PB.11.1|004815|path2:1-3039(+)|transcript/3426          m.4
## 4 PB.11.2|004815|path2:1047-3035(+)|transcript/13401          m.5
## 5      PB.12.1|004815|path5:1-1665(+)|transcript/17167          m.6
## 6       PB.13.1|004c10|path2:1-2811(+)|transcript/4993          m.7
```

```
BLASTx.gene.join_split3 <- full_join(BLASTx.gene.join_split,
    BLASTx.gene.join_split2, by = "TranscriptID_2")
head(BLASTx.gene.join_split3)
```

```
##                                                      TranscriptID
## 1       PB.2.1|002537|path0:1-1624(+)|transcript/18304|m.1
```

```
## 2      PB.3.2|00361c|path0:43-2240(+)|transcript/10564|m.2
## 3       PB.11.1|004815|path2:1-3039(+)|transcript/3426|m.4
## 4 PB.11.2|004815|path2:1047-3035(+)|transcript/13401|m.5
## 5      PB.12.1|004815|path5:1-1665(+)|transcript/17167|m.6
## 6       PB.13.1|004c10|path2:1-2811(+)|transcript/4993|m.7
##                                       TranscriptID_2 UniprotID   Gene
## 1      PB.2.1|002537|path0:1-1624(+)|transcript/18304|m.1   Q6PBC3 chchd4
## 2      PB.3.2|00361c|path0:43-2240(+)|transcript/10564|m.2   P55209 nap1l1
## 3       PB.11.1|004815|path2:1-3039(+)|transcript/3426|m.4   P36633   aoc1
## 4 PB.11.2|004815|path2:1047-3035(+)|transcript/13401|m.5   Q8JZQ5   aoc1
## 5      PB.12.1|004815|path5:1-1665(+)|transcript/17167|m.6   Q8JZQ5   aoc1
## 6       PB.13.1|004c10|path2:1-2811(+)|transcript/4993|m.7   Q9H063   maf1
##    pident length mismatch gapopen qstart qend sstart send    evalue bitscore
## 1 95.000    100        5       0      1  300      1  100 2.49e-69      207
## 2 88.995    209       22       1      1  627    184  391 5.86e-61      198
## 3 61.892    719      258       8     46 2175     32  743 0.00e+00      855
## 4 67.010    194       63       1    196  777    513  705 3.11e-92      246
## 5 61.712    222       80       3      1  666    532  748 3.96e-77      249
## 6 83.984    256       38       1      1  759      1  256 4.61e-156     437
##                                       TranscriptID_3 TranscriptID_4
## 1      PB.2.1|002537|path0:1-1624(+)|transcript/18304          m.1
## 2      PB.3.2|00361c|path0:43-2240(+)|transcript/10564          m.2
## 3       PB.11.1|004815|path2:1-3039(+)|transcript/3426          m.4
## 4 PB.11.2|004815|path2:1047-3035(+)|transcript/13401          m.5
## 5      PB.12.1|004815|path5:1-1665(+)|transcript/17167          m.6
## 6       PB.13.1|004c10|path2:1-2811(+)|transcript/4993          m.7
```

Check: (they all have the same numbers of rows and no data was lost)

```
nrow(BLASTx.gene.join)
```

```
## [1] 12894
```

```
nrow(BLASTx.gene.join_split)
```

```
## [1] 12894
```

```
nrow(BLASTx.gene.join_split2)
```

```
## [1] 12894
```

```
nrow(BLASTx.gene.join_split3)
```

```
## [1] 12894
```

Remove the now unneeded duplicate 'TranscriptID_2' column and 'TranscriptID_4' column Reorder columns so that TranscriptID_3 appears on the left next to the full TranscriptID

```
BLASTx.gene.join_split4 <- subset(BLASTx.gene.join_split3, select = -c(TranscriptID_2,
    TranscriptID_4))
BLASTx.gene.join_split4 <- relocate(BLASTx.gene.join_split4,
    TranscriptID_3, .before = UniprotID)
head(BLASTx.gene.join_split4)
```

```
##                                           TranscriptID
## 1    PB.2.1|002537|path0:1-1624(+)|transcript/18304|m.1
## 2    PB.3.2|00361c|path0:43-2240(+)|transcript/10564|m.2
## 3    PB.11.1|004815|path2:1-3039(+)|transcript/3426|m.4
## 4 PB.11.2|004815|path2:1047-3035(+)|transcript/13401|m.5
## 5    PB.12.1|004815|path5:1-1665(+)|transcript/17167|m.6
## 6    PB.13.1|004c10|path2:1-2811(+)|transcript/4993|m.7
##                                        TranscriptID_3 UniprotID   Gene pident
## 1    PB.2.1|002537|path0:1-1624(+)|transcript/18304     Q6PBC3 chchd4 95.000
## 2    PB.3.2|00361c|path0:43-2240(+)|transcript/10564     P55209 nap1l1 88.995
## 3    PB.11.1|004815|path2:1-3039(+)|transcript/3426     P36633   aoc1 61.892
## 4 PB.11.2|004815|path2:1047-3035(+)|transcript/13401     Q8JZQ5   aoc1 67.010
## 5    PB.12.1|004815|path5:1-1665(+)|transcript/17167     Q8JZQ5   aoc1 61.712
## 6    PB.13.1|004c10|path2:1-2811(+)|transcript/4993     Q9H063   maf1 83.984
##   length mismatch gapopen qstart qend sstart send   evalue bitscore
## 1    100        5       0      1  300      1  100 2.49e-69      207
## 2    209       22       1      1  627    184  391 5.86e-61      198
## 3    719      258       8     46 2175     32  743 0.00e+00      855
## 4    194       63       1    196  777    513  705 3.11e-92      246
## 5    222       80       3      1  666    532  748 3.96e-77      249
## 6    256       38       1      1  759      1  256 4.61e-156     437
```

Reorder the rows so that the Data is arranged by Gene in alphabetical order (and so that transcripts
coresponding to the same gene are listed together).

```
BLASTx.gene.join_split4 <- BLASTx.gene.join_split4 %>%
    arrange(Gene)
head(BLASTx.gene.join_split4)
```

```
##                                              TranscriptID
## 1 PB.7010.1|transcript/12012:1-2076(+)|transcript/12012|m.12781
## 2      PB.2837.1|6a2a91|path0:1-1784(+)|transcript/15955|m.5151
## 3      PB.2838.1|6a2a91|path1:1-2161(+)|transcript/11047|m.5152
## 4      PB.2778.1|67914a|path1:1-1875(+)|transcript/14733|m.5057
## 5     PB.2778.2|67914a|path1:22-1698(+)|transcript/16857|m.5058
## 6      PB.6589.1|fb4549|path0:1-2510(+)|transcript/7231|m.12232
##                                           TranscriptID_3 UniprotID    Gene
## 1 PB.7010.1|transcript/12012:1-2076(+)|transcript/12012     Q9NRG9    aaas
## 2      PB.2837.1|6a2a91|path0:1-1784(+)|transcript/15955     Q5VUY2 aadacl4
## 3      PB.2838.1|6a2a91|path1:1-2161(+)|transcript/11047     Q5VUY2 aadacl4
## 4      PB.2778.1|67914a|path1:1-1875(+)|transcript/14733     Q5E9N4   aadat
## 5     PB.2778.2|67914a|path1:22-1698(+)|transcript/16857     Q5E9N4   aadat
## 6      PB.6589.1|fb4549|path0:1-2510(+)|transcript/7231     Q8N5Z0   aadat
##     pident length mismatch gapopen qstart qend sstart send   evalue bitscore
## 1 66.048     539      162       2    127 1737      1  520 0.00e+00      684
## 2 44.477     344      188       3      1 1026     63  405 8.03e-98      298
```

```
## 3 42.820    383    217    2    79 1224    24  405 6.29e-108    327
## 4 68.235    425    135    0     1 1275     1  425 0.00e+00     647
## 5 68.235    425    135    0     1 1275     1  425 0.00e+00     647
## 6 58.824    425    174    1     1 1272     1  425 0.00e+00     543
```

```r
nrow(BLASTx.gene.join_split4)  #confirming that number of rows remains consistent. The
↪    initial BLASTx result has 12894 rows, and 12894 rows remain.
```

```
## [1] 12894
```

The aforementioned 6343 genes are all present in this dataset, with the additional record representing rows with no Gene ID assigned to Protein ID.

```r
BLASTx.gene.join_split4 %>%
    group_by(Gene) %>%
    summarize(n = n())
```

```
## # A tibble: 6,344 x 2
##     Gene         n
##     <chr>    <int>
##  1 aaas         1
##  2 aadacl4      2
##  3 aadat        9
##  4 aagab        3
##  5 aamp         2
##  6 aanat        2
##  7 aar2         2
##  8 aars1        1
##  9 aarsd1       1
## 10 aass         1
## # ... with 6,334 more rows
```

Write the output to .csv

```r
write.csv(BLASTx.gene.join_split4, "~/@Uni/@Flinders University
↪    PhD/RWorkingDir/Workflow-readthedown/ReferenceClusters/BLASTx.gene.join_split4.csv",
    quote = FALSE, row.names = FALSE, col.names = TRUE)
```

Remove all of the rows where Gene = NA

```r
BLASTx.gene.join_filt <- BLASTx.gene.join_split4 %>%
    filter(!is.na(Gene))
head(BLASTx.gene.join_filt)
```

```
##                                                       TranscriptID
## 1 PB.7010.1|transcript/12012:1-2076(+)|transcript/12012|m.12781
## 2     PB.2837.1|6a2a91|path0:1-1784(+)|transcript/15955|m.5151
## 3     PB.2838.1|6a2a91|path1:1-2161(+)|transcript/11047|m.5152
## 4     PB.2778.1|67914a|path1:1-1875(+)|transcript/14733|m.5057
## 5     PB.2778.2|67914a|path1:22-1698(+)|transcript/16857|m.5058
```

```
## 6       PB.6589.1|fb4549|path0:1-2510(+)|transcript/7231|m.12232
##                                          TranscriptID_3 UniprotID    Gene
## 1 PB.7010.1|transcript/12012:1-2076(+)|transcript/12012   Q9NRG9    aaas
## 2     PB.2837.1|6a2a91|path0:1-1784(+)|transcript/15955   Q5VUY2 aadacl4
## 3     PB.2838.1|6a2a91|path1:1-2161(+)|transcript/11047   Q5VUY2 aadacl4
## 4     PB.2778.1|67914a|path1:1-1875(+)|transcript/14733   Q5E9N4   aadat
## 5    PB.2778.2|67914a|path1:22-1698(+)|transcript/16857   Q5E9N4   aadat
## 6       PB.6589.1|fb4549|path0:1-2510(+)|transcript/7231   Q8N5Z0   aadat
##   pident length mismatch gapopen qstart qend sstart send   evalue bitscore
## 1 66.048    539      162       2    127 1737      1  520 0.00e+00      684
## 2 44.477    344      188       3      1 1026     63  405 8.03e-98      298
## 3 42.820    383      217       2     79 1224     24  405 6.29e-108     327
## 4 68.235    425      135       0      1 1275      1  425 0.00e+00      647
## 5 68.235    425      135       0      1 1275      1  425 0.00e+00      647
## 6 58.824    425      174       1      1 1272      1  425 0.00e+00      543
```

```
tail(BLASTx.gene.join_filt)
```

```
##                                                        TranscriptID
## 12527    PB.2713.1|65abaa|path0:1-2495(+)|transcript/7761|m.4957
## 12528 PB.2713.2|65abaa|path0:37-2312(+)|transcript/10678|m.4958
## 12529  PB.2713.3|65abaa|path0:37-2495(+)|transcript/7693|m.4959
## 12530  PB.2713.3|65abaa|path0:37-2495(+)|transcript/7693|m.4960
## 12531    PB.2714.1|65abaa|path1:1-2323(+)|transcript/9134|m.4961
## 12532  PB.2714.2|65abaa|path1:4-2323(+)|transcript/10134|m.4962
##                                                   TranscriptID_3 UniprotID Gene pident
## 12527    PB.2713.1|65abaa|path0:1-2495(+)|transcript/7761   Q04584  zyx 64.286
## 12528 PB.2713.2|65abaa|path0:37-2312(+)|transcript/10678   Q04584  zyx 81.985
## 12529  PB.2713.3|65abaa|path0:37-2495(+)|transcript/7693   Q04584  zyx 92.857
## 12530  PB.2713.3|65abaa|path0:37-2495(+)|transcript/7693   Q04584  zyx 66.055
## 12531    PB.2714.1|65abaa|path1:1-2323(+)|transcript/9134   Q04584  zyx 81.724
## 12532  PB.2714.2|65abaa|path1:4-2323(+)|transcript/10134   Q04584  zyx 81.724
##       length mismatch gapopen qstart qend sstart send   evalue bitscore
## 12527     84       24       1    514  765    285  362 7.93e-29      107
## 12528    272       43       1    832 1647    258  523 1.66e-163     479
## 12529    126        9       0      3  380    375  500 7.10e-85      260
## 12530    109       31       1    931 1257    258  360 3.96e-37      144
## 12531    290       47       1   1090 1959    258  541 6.38e-175     512
## 12532    290       47       1    988 1857    258  541 2.28e-175     511
```

```
nrow(BLASTx.gene.join_filt)
```

```
## [1] 12532
```

See above, 12532 transcripts remain which were mapped to one of the 6343 Gene IDs.

Write the output to .csv

```
write.csv(BLASTx.gene.join_filt, "~/@Uni/@Flinders University
↪   PhD/RWorkingDir/Workflow-readthedown/ReferenceClusters/BLASTx.gene.join_filt.csv",
    quote = FALSE, row.names = FALSE, col.names = TRUE)
```

3.2.2 Putative genes for clustered transcripts used as the gene expression analysis reference

The above list of transcripts represent all transcript isoforms assembled for *T. adelaidensis* with a predicted open reading frame determined by ANGEL (Section 2.1.3). After further clustering using translated protein sequences in CD-HIT (Section 2.1.4-7), a smaller reference set of transcripts were used as the reference for gene expresion analysis (Section 4.1).

The above list will be filtered based on the same list of cluster representatives used in Section 2.1 to create a smaller list of all BLAST results relating to the transcripts used as gene expression references.

Import the list of unique clustered transcripts that were used as the reference for gene expression analysis. Note that in this case, data after the ANGEL step are being sub-set, whereas in Section 2.1.7 full length fasta sequences of corresponding transcripts were desired. As noted in section 2.1 a few full length transcripts give rise to more than one transcript isoform listed here, and thus may be assigned more than one putative gene ID.

ClstrTranscriptID is already loaded:

```
head(clstrTranscriptID)
```

```
## [1] PB.6088.11|e88f16|path1:164-2792(+)|transcript/22345|m.11362
## [2] PB.2919.1|6c1fb9|path4:13-2574(+)|transcript/10624|m.5342
## [3] PB.535.2|136fa8|path0:1-2775(+)|transcript/4312|m.913
## [4] PB.4689.2|afa3de|path19:22-2096(+)|transcript/11491|m.8666
## [5] PB.6357.1|f27f14|path2:4-2108(+)|transcript/12511|m.11846
## [6] PB.715.4|1b393a|path6:6-8031(+)|transcript/8|m.1216
## 13882 Levels: PB.10.1|004815|path1:5-1713(+)|transcript/17534|m.3 ...
```

Filter the BLASTx results with gene name based on the transcript references used for the expression analysis (i.e. the clustered transcript file).

```
BLASTx.gene.join_filt_clustered <- BLASTx.gene.join_filt %>%
    filter(TranscriptID %in% clstrTranscriptID)
head(BLASTx.gene.join_filt_clustered)
```

```
##                                                      TranscriptID
## 1 PB.7010.1|transcript/12012:1-2076(+)|transcript/12012|m.12781
## 2       PB.2838.1|6a2a91|path1:1-2161(+)|transcript/11047|m.5152
## 3       PB.2778.1|67914a|path1:1-1875(+)|transcript/14733|m.5057
## 4       PB.6589.1|fb4549|path0:1-2510(+)|transcript/7231|m.12232
## 5       PB.6589.2|fb4549|path0:46-2510(+)|transcript/8322|m.12233
## 6     PB.6590.3|fb4549|path10:52-2691(+)|transcript/5755|m.12236
##                                            TranscriptID_3 UniprotID    Gene
## 1 PB.7010.1|transcript/12012:1-2076(+)|transcript/12012    Q9NRG9    aaas
## 2       PB.2838.1|6a2a91|path1:1-2161(+)|transcript/11047    Q5VUY2 aadacl4
## 3       PB.2778.1|67914a|path1:1-1875(+)|transcript/14733    Q5E9N4   aadat
## 4         PB.6589.1|fb4549|path0:1-2510(+)|transcript/7231    Q8N5Z0   aadat
## 5       PB.6589.2|fb4549|path0:46-2510(+)|transcript/8322    Q8N5Z0   aadat
## 6     PB.6590.3|fb4549|path10:52-2691(+)|transcript/5755    Q5E9N4   aadat
##   pident length mismatch gapopen qstart qend sstart send    evalue bitscore
## 1 66.048    539      162       2    127 1737      1  520  0.00e+00      684
## 2 42.820    383      217       2     79 1224     24  405 6.29e-108      327
## 3 68.235    425      135       0      1 1275      1  425  0.00e+00      647
## 4 58.824    425      174       1      1 1272      1  425  0.00e+00      543
```

```
## 5 56.000    425     162      2      1 1200     1  425 1.02e-177      504
## 6 60.984    305     119      0      7  921   121  425 8.15e-144      414
```

```
nrow(BLASTx.gene.join_filt_clustered)
```

```
## [1] 8861
```

Note: of the initial 13882 sequences in the predicted open read frame set submitted for a BLASTx search, only 12602 query sequences returned a match, and not all protein ID matches returned a Gene ID match. Coupled with some transcripts that may have been identified to putative gene ID being removed at the CD-HIT clustering step, the total remaining number of transcripts with a coresponding Protein ID and Gene ID here is expected to be less than the clustered reference transcript dataset containing 9813 unique full-length transcripts.
Ninety four full length transcripts produced more than one isoform in predicted open read frame which was retained at the clustering step, so there is also the possibility that duplicate entries here apply to a single full length 'master' transcript.

Here 8861 transcripts in predicted open read frame have been assigned both a protein ID and putative gene ID, from a total list of 9813 corresponding full length transcripts which were used as a reference in later gene expression analysis.

3.2.3   Presence of identified "genes of interest" in the *T. adelaidensis* transcript set

In order to generate a list of potential "genes of interest" and narrow down the focus of analysis, an NCBI gene database search was conducted aiming for genes identified in reptiles associated with renal function, water homeostasis and heat regulation.

The final NCBI Gene database search was accessed on on June 14, 2020 including the following terms:
- (Sauria[Organism]) AND (Aquaporin OR Bile OR Dehydration OR Diffus* OR Excretion OR Filt* OR Fluid OR (Heat AND Stress) OR Heat Shock OR Heat Stress OR Hibernation OR Homeostasis OR Ion Channel OR Ion Transport OR Kidney OR Membrane And (Potential OR Permeability OR Pore) OR Metabolic OR Osmo* OR Permeability OR Ph Balance OR Renal OR Solute OR Stress OR Temperature OR Thermal OR Uric OR Water OR Water Retention OR Water Permeability OR Water Transport)

These full search results were exported and refined in excel with care taken to manually preserve gene symbol data fields. Note: summary counts of genes IDs assigned to transcripts was initially conducted in excel (Appendix 4). Manual manipulation of the exported database data was conducted to summarise genes identified, and to create a list of unique genes returned by the search term. Duplicate gene entries were prioritised by taxa and a list of 993 genes of particular interest to this study were identified.

Excel was used to create initial summary statistics of the taxa and genes returned, as well as record information retained at each step of manual filtering (Appendix 4).

Excel is not ideal for the manipulation of lists of gene names this represents preliminary data exploration, so great care was taken to ensure gene IDs integrity was maintained and all further filtering was conducted in R below. The following analysis was conducted in to subset the sequenced *T. adelaidensis* transcript lists based on putative gene annotation.

Import the reference "genes of interest" list built from an NCBI database search (Chapter 5). This imported list is after initial duplicate filtering in excel as well as adding a column for gene sumbol/ID in all lower case. It contains 993 unique gene symbols for the most favoured taxa (as per the list in Appendix 4).

```r
Sauria_2020.06.14 <- read.csv("~/@Uni/@Flinders University
↪  PhD/RWorkingDir/Workflow-readthedown/ReferenceClusters/Returned organisms 2020-06-15
↪  Sauria-RM-dup-loc FULL.csv",
    sep = ",", header = TRUE)
nrow(Sauria_2020.06.14)
```

```
## [1] 993
```

```r
head(Sauria_2020.06.14)
```

```
##   tax_id Org_name.new_name          Org_name    GeneID CurrentID Status
## 1   9031                 J      Gallus gallus    416811         0   live
## 2  38654                 I  Alligator sinensis 102374204         0   live
## 3   9031                 J      Gallus gallus    418791         0   live
## 4   9031                 J      Gallus gallus    423767         0   live
## 5  28377                 A Anolis carolinensis 100565298         0   live
## 6   9031                 J      Gallus gallus    373916         0   live
##   Symbol Symbol.CAPS      Aliases
## 1   AACS        aacs
## 2  ABCA5       abca5
## 3  ABCC4       abcc4
## 4  ABCG2       abcg2
## 5   ace2        ace2
## 6   ACO1        aco1 IREB1, IREBP
##                                                        description
## 1                                       acetoacetyl-CoA synthetase
## 2                             ATP binding cassette subfamily A member 5
## 3                             ATP binding cassette subfamily C member 4
## 4 ATP binding cassette subfamily G member 2 (Junior blood group)
## 5                             angiotensin I converting enzyme 2
## 6                                             aconitase 1, soluble
##
## 1
## 2                                                                                    ATP-bindin
## 3                                         multidrug resistance-associated protein 4|AT
## 4 broad substrate specificity ATP-binding cassette transporter ABCG2|ATP-binding cassette sub-family
## 5
## 6              cytoplasmic aconitate hydratase|IRE-BP 1|IRE-binding protein 1|Iron responsive
##   map_location chromosome genomic_nucleotide_accession.version
## 1           15         15                          NC_006102.5
## 2           Un                                                 
## 3            1          1                          NC_006088.5
## 4            6          6                          NC_006093.5
## 5            3          3                          NC_014778.1
## 6            Z          Z                          NC_006127.5
##   start_position_on_the_genomic_accession end_position_on_the_genomic_accession
## 1                                 4746614                               4783923
## 2                                      NA                                    NA
## 3                               147470494                             147695961
## 4                                18374408                              18392256
## 5                               121501128                             121544841
## 6                                71167742                              71204524
```

```
##   orientation exon_count OMIM Column1
## 1       minus         18   NA      NA
## 2                      NA   NA      NA
## 3        plus         37   NA      NA
## 4        plus         18   NA      NA
## 5       minus         20   NA      NA
## 6        plus         21   NA      NA
```

The "Gene Symbol" column output by the NCBI database matches the format of "gene names" output by
the conversion of Uni Prot Protein IDs generated

Check Gene ID title case for non-unique entries

```
Sauria_2020.06.14 %>%
    count(Symbol.CAPS) %>%
    filter(n > 1)
```

```
## [1] Symbol.CAPS n
## <0 rows> (or 0-length row.names)
```

Rename column Symbol.CAPS to Gene so column can be compared to the Gene column in BLASTx.gene.join_filt,
make this column characters to allow joins

```
Sauria_2020.06.14 <- rename(Sauria_2020.06.14, Gene = "Symbol.CAPS")
Sauria_2020.06.14$Gene <- as.character(Sauria_2020.06.14$Gene)
head(Sauria_2020.06.14)
```

```
##   tax_id Org_name.new_name            Org_name    GeneID CurrentID Status
## 1   9031                 J        Gallus gallus    416811         0   live
## 2  38654                 I  Alligator sinensis 102374204         0   live
## 3   9031                 J        Gallus gallus    418791         0   live
## 4   9031                 J        Gallus gallus    423767         0   live
## 5  28377                 A Anolis carolinensis 100565298         0   live
## 6   9031                 J        Gallus gallus    373916         0   live
##   Symbol  Gene      Aliases
## 1   AACS  aacs
## 2  ABCA5 abca5
## 3  ABCC4 abcc4
## 4  ABCG2 abcg2
## 5   ace2  ace2
## 6   ACO1  aco1 IREB1, IREBP
##                                                          description
## 1                                  acetoacetyl-CoA synthetase
## 2                     ATP binding cassette subfamily A member 5
## 3                     ATP binding cassette subfamily C member 4
## 4 ATP binding cassette subfamily G member 2 (Junior blood group)
## 5                          angiotensin I converting enzyme 2
## 6                                  aconitase 1, soluble
##
## 1
## 2                                                            ATP-bindi
## 3                               multidrug resistance-associated protein 4|AT
```

```
## 4 broad substrate specificity ATP-binding cassette transporter ABCG2|ATP-binding cassette sub-family
## 5
## 6                     cytoplasmic aconitate hydratase|IRE-BP 1|IRE-binding protein 1|Iron responsive
##   map_location chromosome genomic_nucleotide_accession.version
## 1           15                                         NC_006102.5
## 2           Un
## 3            1                                         NC_006088.5
## 4            6                                         NC_006093.5
## 5            3                                         NC_014778.1
## 6            Z                                         NC_006127.5
##   start_position_on_the_genomic_accession end_position_on_the_genomic_accession
## 1                                 4746614                               4783923
## 2                                      NA                                    NA
## 3                               147470494                             147695961
## 4                                18374408                              18392256
## 5                               121501128                             121544841
## 6                                71167742                              71204524
##   orientation exon_count OMIM Column1
## 1       minus         18   NA      NA
## 2                      NA   NA      NA
## 3        plus         37   NA      NA
## 4        plus         18   NA      NA
## 5       minus         20   NA      NA
## 6        plus         21   NA      NA
```

Create a larger summary file of the transcripts which have a BLASTx result. All BLASTx results will be retained, but there is no need to retain information for 'genes of interest' which were not identified in this dataset. This will be used later to identify information on transcripts which may be identified in the gene expression analysis.

```
BLASTx.gene.join_Database.Full.PBT.Summary <- left_join(BLASTx.gene.join_filt,
    Sauria_2020.06.14, by = "Gene")
nrow(BLASTx.gene.join_Database.Full.PBT.Summary)
```

```
## [1] 12532
```

```
head(BLASTx.gene.join_Database.Full.PBT.Summary)
```

```
##                                                          TranscriptID
## 1 PB.7010.1|transcript/12012:1-2076(+)|transcript/12012|m.12781
## 2      PB.2837.1|6a2a91|path0:1-1784(+)|transcript/15955|m.5151
## 3      PB.2838.1|6a2a91|path1:1-2161(+)|transcript/11047|m.5152
## 4      PB.2778.1|67914a|path1:1-1875(+)|transcript/14733|m.5057
## 5      PB.2778.2|67914a|path1:22-1698(+)|transcript/16857|m.5058
## 6      PB.6589.1|fb4549|path0:1-2510(+)|transcript/7231|m.12232
##                                         TranscriptID_3 UniprotID    Gene
## 1 PB.7010.1|transcript/12012:1-2076(+)|transcript/12012    Q9NRG9    aaas
## 2      PB.2837.1|6a2a91|path0:1-1784(+)|transcript/15955    Q5VUY2 aadacl4
## 3      PB.2838.1|6a2a91|path1:1-2161(+)|transcript/11047    Q5VUY2 aadacl4
## 4      PB.2778.1|67914a|path1:1-1875(+)|transcript/14733    Q5E9N4   aadat
## 5      PB.2778.2|67914a|path1:22-1698(+)|transcript/16857    Q5E9N4   aadat
## 6      PB.6589.1|fb4549|path0:1-2510(+)|transcript/7231    Q8N5Z0   aadat
```

```
##    pident length mismatch gapopen qstart qend sstart send    evalue bitscore
## 1 66.048    539      162       2    127 1737      1  520  0.00e+00      684
## 2 44.477    344      188       3      1 1026     63  405  8.03e-98      298
## 3 42.820    383      217       2     79 1224     24  405 6.29e-108      327
## 4 68.235    425      135       0      1 1275      1  425  0.00e+00      647
## 5 68.235    425      135       0      1 1275      1  425  0.00e+00      647
## 6 58.824    425      174       1      1 1272      1  425  0.00e+00      543
##    tax_id Org_name.new_name Org_name GeneID CurrentID Status Symbol Aliases
## 1     NA              <NA>     <NA>     NA        NA   <NA>   <NA>    <NA>
## 2     NA              <NA>     <NA>     NA        NA   <NA>   <NA>    <NA>
## 3     NA              <NA>     <NA>     NA        NA   <NA>   <NA>    <NA>
## 4     NA              <NA>     <NA>     NA        NA   <NA>   <NA>    <NA>
## 5     NA              <NA>     <NA>     NA        NA   <NA>   <NA>    <NA>
## 6     NA              <NA>     <NA>     NA        NA   <NA>   <NA>    <NA>
##    description other_designations map_location chromosome
## 1        <NA>              <NA>         <NA>       <NA>
## 2        <NA>              <NA>         <NA>       <NA>
## 3        <NA>              <NA>         <NA>       <NA>
## 4        <NA>              <NA>         <NA>       <NA>
## 5        <NA>              <NA>         <NA>       <NA>
## 6        <NA>              <NA>         <NA>       <NA>
##    genomic_nucleotide_accession.version start_position_on_the_genomic_accession
## 1                                 <NA>                                      NA
## 2                                 <NA>                                      NA
## 3                                 <NA>                                      NA
## 4                                 <NA>                                      NA
## 5                                 <NA>                                      NA
## 6                                 <NA>                                      NA
##    end_position_on_the_genomic_accession orientation exon_count OMIM Column1
## 1                                    NA        <NA>         NA   NA      NA
## 2                                    NA        <NA>         NA   NA      NA
## 3                                    NA        <NA>         NA   NA      NA
## 4                                    NA        <NA>         NA   NA      NA
## 5                                    NA        <NA>         NA   NA      NA
## 6                                    NA        <NA>         NA   NA      NA
```

Export this faile as a .csv

```r
write.csv(BLASTx.gene.join_Database.Full.PBT.Summary, "~/@Uni/@Flinders University
↪  PhD/RWorkingDir/Workflow-readthedown/ReferenceClusters/BLASTx.gene.join_Database.Full.PBT.Summary.cs
    quote = FALSE, row.names = FALSE, col.names = TRUE)
```

The number of transcripts which returned a BLASTx result for each gene was already counted in a previous section. Make this a dataframe.

```r
Gene_transcript_count <- BLASTx.gene.join_split4 %>%
    group_by(Gene) %>%
    summarize(n = n())
head(Gene_transcript_count)
```

```
## # A tibble: 6 x 2
##   Gene          n
```

```
##   <chr>   <int>
## 1 aaas       1
## 2 aadacl4    2
## 3 aadat      9
## 4 aagab      3
## 5 aamp       2
## 6 aanat      2
```

Filter the genes of interest list based on Gene IDs which match to a BLASTx result from the full transcript list

```
Sauria_2020.06.14_Transcripts.Present.Full <- semi_join(Sauria_2020.06.14,
    BLASTx.gene.join_filt, by = "Gene")
head(Sauria_2020.06.14_Transcripts.Present.Full)
```

```
##   tax_id Org_name.new_name      Org_name GeneID CurrentID Status Symbol   Gene
## 1   9031              J Gallus gallus 418791         0   live  ABCC4  abcc4
## 2   9031              J Gallus gallus 373916         0   live   ACO1   aco1
## 3   9031              J Gallus gallus 420090         0   live ACSBG2 acsbg2
## 4   9031              J Gallus gallus 421534         0   live  ACTA1  acta1
## 5   9031              J Gallus gallus 373918         0   live  ACTN1  actn1
## 6   9031              J Gallus gallus 419194         0   live    ADA    ada
##       Aliases                                         description
## 1              ATP binding cassette subfamily C member 4
## 2 IREB1, IREBP                          aconitase 1, soluble
## 3              acyl-CoA synthetase bubblegum family member 2
## 4                        actin alpha 1, skeletal muscle
## 5                                      actinin, alpha 1
## 6                                     adenosine deaminase
##
## 1                                   multidrug resistance-associated protein 4|ATP-binding cassette
## 2 cytoplasmic aconitate hydratase|IRE-BP 1|IRE-binding protein 1|Iron responsive element binding prot
## 3                                                                                           long-cl
## 4
## 5                                      alpha-actinin-1|F-actin cross-linking prot
## 6
##   map_location chromosome genomic_nucleotide_accession.version
## 1                     1                             NC_006088.5
## 2                     Z                             NC_006127.5
## 3                    28                             NC_006115.5
## 4                     3                             NC_006090.5
## 5                     5                             NC_006092.5
## 6                    20                             NC_006107.5
##   start_position_on_the_genomic_accession end_position_on_the_genomic_accession
## 1                               147470494                             147695961
## 2                                71167742                              71204524
## 3                                 1746737                               1763012
## 4                                40088178                              40090517
## 5                                28427294                              28508906
## 6                                 5355077                               5367713
##   orientation exon_count OMIM Column1
## 1        plus         37   NA      NA
## 2        plus         21   NA      NA
```

```
## 3        plus        18   NA        NA
## 4       minus         7   NA        NA
## 5        plus        23   NA        NA
## 6        plus        11   NA        NA
```

```
nrow(Sauria_2020.06.14_Transcripts.Present.Full)
```

```
## [1] 393
```

Add the counts of number of transcripts matching each 'gene of interest' as a new column to the genes of interest table, reorder columns.

```
Sauria_2020.06.14_Transcripts.Present.Count <-
→   left_join(Sauria_2020.06.14_Transcripts.Present.Full,
    Gene_transcript_count, by = "Gene")
Sauria_2020.06.14_Transcripts.Present.Count <-
→   rename(Sauria_2020.06.14_Transcripts.Present.Count,
    Transcript_Count = "n")

Sauria_2020.06.14_Transcripts.Present.Count <-
→   relocate(Sauria_2020.06.14_Transcripts.Present.Count,
    description, .before = Aliases)
Sauria_2020.06.14_Transcripts.Present.Count <-
→   relocate(Sauria_2020.06.14_Transcripts.Present.Count,
    Transcript_Count, .before = Aliases)
Sauria_2020.06.14_Transcripts.Present.Count <-
→   select(Sauria_2020.06.14_Transcripts.Present.Count,
    -Org_name.new_name)  #this column was used as alphabetical factors in excel to sort
    →   the organisms by category so that the desired filtering based on taxa group could
    →   be achieved when duplicate gene database results were removed. It is no longer
    →   needed
head(Sauria_2020.06.14_Transcripts.Present.Count)
```

```
##   tax_id      Org_name GeneID CurrentID Status Symbol   Gene
## 1   9031 Gallus gallus 418791         0   live  ABCC4  abcc4
## 2   9031 Gallus gallus 373916         0   live   ACO1   aco1
## 3   9031 Gallus gallus 420090         0   live ACSBG2 acsbg2
## 4   9031 Gallus gallus 421534         0   live  ACTA1  acta1
## 5   9031 Gallus gallus 373918         0   live  ACTN1  actn1
## 6   9031 Gallus gallus 419194         0   live    ADA    ada
##                                  description Transcript_Count     Aliases
## 1       ATP binding cassette subfamily C member 4                2
## 2                          aconitase 1, soluble                4 IREB1, IREBP
## 3 acyl-CoA synthetase bubblegum family member 2                3
## 4               actin alpha 1, skeletal muscle                2
## 5                              actinin, alpha 1                2
## 6                           adenosine deaminase                1
##
## 1                                   multidrug resistance-associated protein 4|ATP-binding cassette
## 2 cytoplasmic aconitate hydratase|IRE-BP 1|IRE-binding protein 1|Iron responsive element binding pro
## 3                                                                                             long-cl
## 4
```

```
## 5                                         alpha-actinin-1|F-actin cross-linking pro
## 6
##   map_location chromosome genomic_nucleotide_accession.version
## 1                      1                           NC_006088.5
## 2                      Z                           NC_006127.5
## 3                     28                           NC_006115.5
## 4                      3                           NC_006090.5
## 5                      5                           NC_006092.5
## 6                     20                           NC_006107.5
##   start_position_on_the_genomic_accession end_position_on_the_genomic_accession
## 1                               147470494                             147695961
## 2                                71167742                              71204524
## 3                                 1746737                               1763012
## 4                                40088178                              40090517
## 5                                28427294                              28508906
## 6                                 5355077                               5367713
##   orientation exon_count OMIM Column1
## 1        plus         37   NA      NA
## 2        plus         21   NA      NA
## 3        plus         18   NA      NA
## 4       minus          7   NA      NA
## 5        plus         23   NA      NA
## 6        plus         11   NA      NA
```

Export this faile as a .csv

```
write.csv(Sauria_2020.06.14_Transcripts.Present.Full, "~/@Uni/@Flinders University
↪    PhD/RWorkingDir/Workflow-readthedown/ReferenceClusters/Sauria_2020.06.14_Transcripts.Present.Count.
     quote = FALSE, row.names = FALSE, col.names = TRUE)
```

Filter the genes of interest list based on Gene IDs which match to a BLASTx result from the clustered
transcript list used as a reference for gene expression analysis

```
Sauria_2020.06.14_Transcripts.Present.Clustered <- semi_join(Sauria_2020.06.14,
    BLASTx.gene.join_filt_clustered, by = "Gene")
nrow(Sauria_2020.06.14_Transcripts.Present.Clustered)
```

```
## [1] 392
```

Note: only a single gene that was both identified through the BLASTx search *and* identified as a 'gene of
interest' in the database search was lost in this filtering step.

filter the Full transcript list based on Gene symbol identified in NCBI database search

```
nrow(BLASTx.gene.join_filt)
```

```
## [1] 12532
```

```
BLASTx.gene.join_Database.Present <- semi_join(BLASTx.gene.join_filt,
    Sauria_2020.06.14, by = "Gene")
nrow(BLASTx.gene.join_Database.Present)
```

```
## [1] 955
```

Of the 12532 transcripts which returned a BLASTx result for their predicted coding region, 955 are represented in the 'genes of interest' dataset created from an NCBI gene database search of terms relevant to renal function and water homeostasis.

Export this faile as a .csv

```
write.csv(BLASTx.gene.join_Database.Present, "~/@Uni/@Flinders University
↪    PhD/RWorkingDir/Workflow-readthedown/ReferenceClusters/BLASTx.gene.join_Database.Present.csv",
     quote = FALSE, row.names = FALSE, col.names = TRUE)
```

3.3 BLAST 2 GO

BLAST2GO (Within OmicsBox – Bioinformatics Made Easy, BioBam Bioinformatics, March 3, 2019) was used to perform a gene ontology analysis of the above BLASTx results.

Götz S., Garcia-Gomez JM., Terol J., Williams TD., Nagaraj SH., Nueda MJ., Robles M., Talon M., Dopazo J. and Conesa A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic acids research, 36(10), 3420-35.

For import into the program a BLASTx search with the output set to .xml was performed using the fasta file containing the predicted open reading frame compared to the *Anolis carolinensis* proteins database (-max_target_seqs 5 -max_hsps 5 -evalue 0.00001).

Comparison to GO databases and visualisation was all completed within a Windows x64 v1.2.4 build of the OmicsBox program and figures were exported as below.
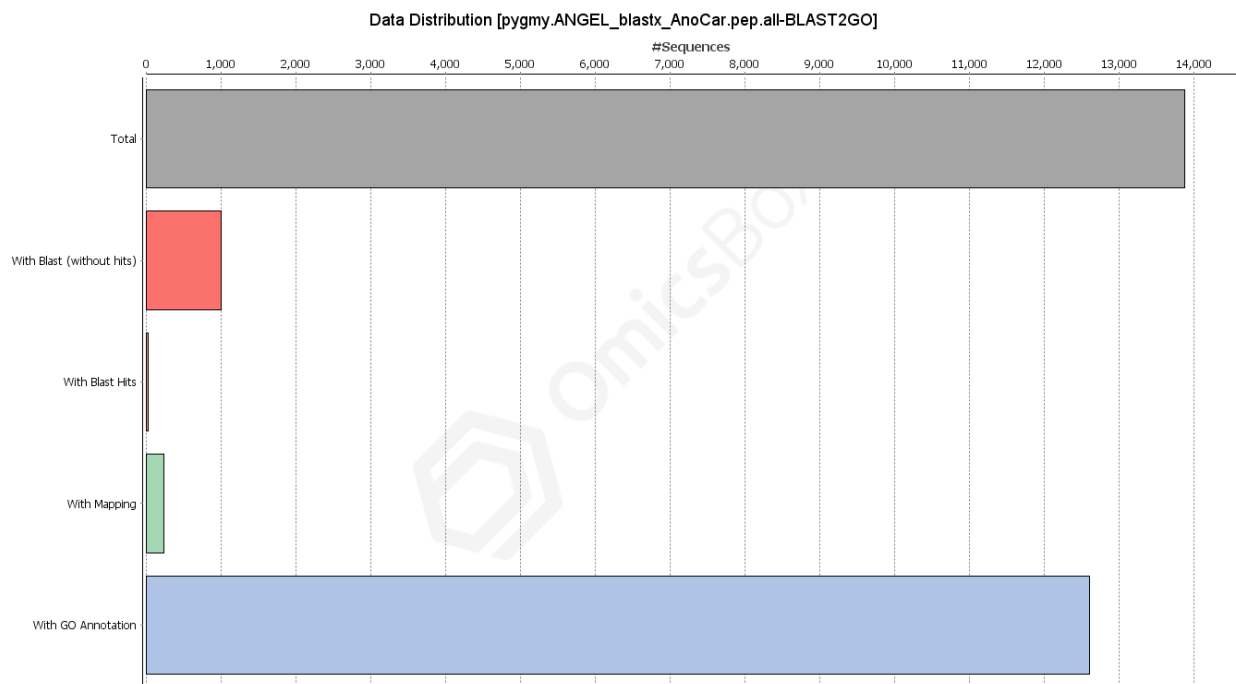

Figure 4: Summary data distribution. pygmy.ANGEL contains 13,882 unique sequences representing predicted open reading frame of transcripts: 12,602 of these sequences produced >1 BLASTx hit with Mapping and GO Annotation, 243 produced a BLASTx hit and mapping only, 32 produced a BLAST hit only, and 1005 did not produce BLASTx hits
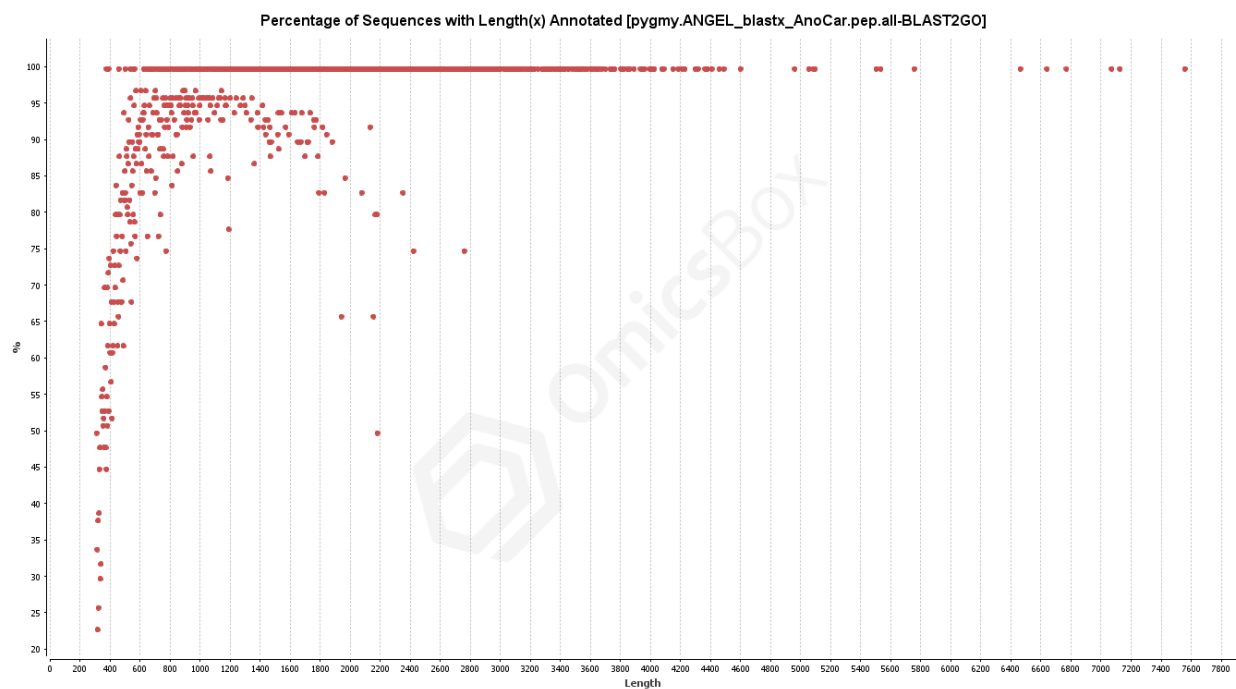
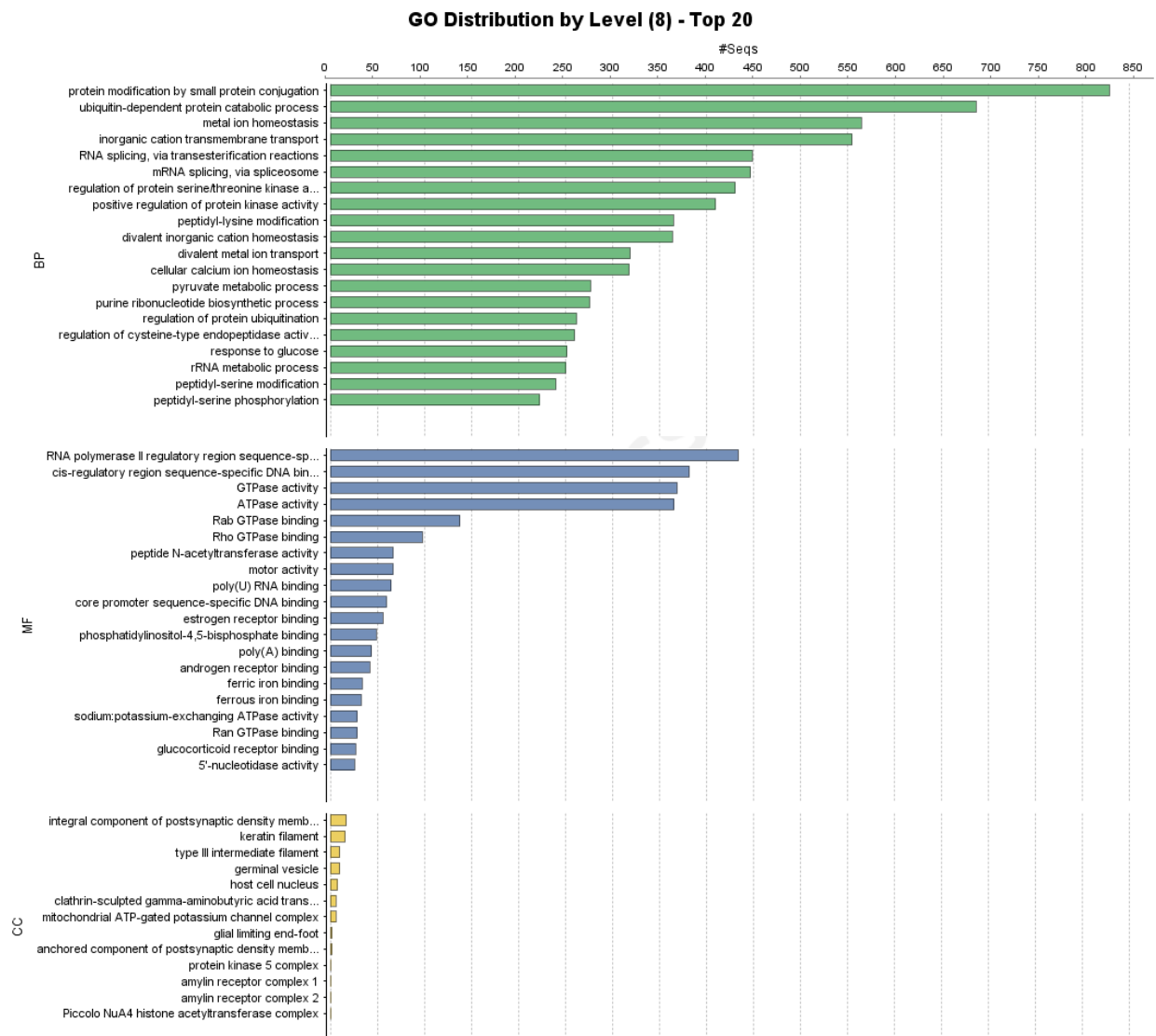Figure 5: Percentage of sequences with length (x) that were annotated

Figure 6: Top 20 annotation results (at level 8) for each category BP - Biological Process, MF - Molecular Function, and CC - Celular Component

Graph Level 2 Pie Chart of #Seqs [Biological Process]

reproductive process: (1,378 / 2%)

reproduction: (1,385 / 2%)

locomotion: (1,476 / 2%)

cell population proliferation: (1,638 / 2%)

immune system process: (2,559 / 3%)

multi-organism process: (2,703 / 3%)

signaling: (4,152 / 5%)

negative regulation of biological process: (4,494 / 5%)

positive regulation of biological process: (5,124 / 6%)

developmental process: (5,336 / 6%)

localization: (5,670 / 6%)

multicellular organismal process: (5,694 / 6%)

cellular component organization or biogenesis: (5,713 / 6%)

cellular process: (11,301 / 12%)

metabolic process: (9,266 / 10%)

biological regulation: (8,523 / 9%)

regulation of biological process: (7,922 / 9%)

response to stimulus: (7,012 / 8%)

Figure 7: Pie chart of proportion of sequences annotated to Level 2 of the Biological Process category

Figure 8: Wordcloud of all GO terms annotated based on _T. adelaidensis transcript BLASTx results against the Anolis genome, all GO levels combined

Some key words from this wordcloud were used to supplement terms included in the NCBI Gene database search described in section 3.2.3 and outlined in Appendix 4.

3.4   Summary

15,729 Total transcript isoforms retained in initial cleaning and clustering of redundant isoforms.

13,882 sequences in a predicted open reading frame.

9,907 clusters identified based on translated proteins of predicted open reading frame.

9,813 full length transcripts have been subset into a reference file as the longest representatives of one or more transcript clusters.

12,602 of the 13,882 sequences in a predicted open reading frame were successfully annotated to a BLASTx result, protein ID, and gene ontology category.

4   Gene expression analysis (Chapter 6)

4.0.0.0.1   Long reads (reference index) - Previous analysis:

- Isoseq3 pipe
- Collapsing of isoforms & generating fasta of longest transcripts in predicted ORFs (Cogent/ANGEL)
- BLASTx search of coding sequence file for the longest transcripts in predicted ORFs
- Grouping of transcript clusters/putative gene families (CD-HIT)
- Longest representative transcript collapsed into a reference .fasta (R & Shell scripts)
- As this longest transcript was taken from earlier in the pipeline, persisting poly-a tails were trimmed from the reference .fasta file (Perl script)

4.0.0.0.2   Short reads - Previous analysis:

- FASTQC quality assessment
- BBduk trimming
- FASTQC quality assessment (note: TRINITY assembly not used for counts)

4.0.0.0.3   Both datasets are combined in this analysis:

- Transcript counts per million estimated using Kallisto
  - Clean short reads (above) counted against reference index generated from long read .fasta (above)
- Kallisto output imported into R for analysis using EdgeR

4.1   Kallisto

Short reads have already been trimmed using BBDuk as per Chapter 4. These cleaned sequences were imported directly into Kallisto.

As TRINITY was run on the Flinders' HPC Deep Thought that is where trimming was completed, Kallisto however was run on the eRSA NECTAR cloud machine and files were moved accordingly, keeping directories relative locations intact.

The file reference_transcripts.1Lv.clean.fasta created in the subsetting section of Chapter 4's methods was used as the index.
All of this was achieved using the following kallisto-clstr-clean.sh script summarised below.

create the index the below contains the representative transcript for clustered isoforms with (most) of the poly-a tails trimmed:

```
kallisto index -i reference_transcripts.1Lv.clean.idx
→   /mnt/IsoSeq-analysis/data/Seqtk/reference_transcripts.1Lv.clean.fasta
```

Output:

```
# [build] loading fasta file
→   /mnt/IsoSeq-analysis/data/Seqtk/reference_transcripts.1Lv.clean.fasta
# [build] k-mer length: 31
# [build] counting k-mers ... done.
# [build] building target de Bruijn graph ...  done
# [build] creating equivalence classes ...  done
# [build] target de Bruijn graph has 59186 contigs and contains 17770125 k-mers
```

Run kallisto:

```
    KallistoDIR=/mnt/DE-analysis/2_alignedData/kallisto-clstr-clean
    ReadDIR=/mnt/DE-analysis/1_trimmedData #contains trimmed R1, R2, and singletons

    # need to be in the input directory for for file
    cd $ReadDIR

    for file in *R1_clean.fq.gz
    do
      FILESTEM=${file%_*}
      #this FILESTEM only cuts to _clean, the _R1 is included
      FILESTEM=${FILESTEM/R1/}
      #removes R1 from FILESTEM (FILESTEM ends in _ therefore not needed in "text"
      ↪  names)

    echo $FILESTEM

    kallisto quant -i $KallistoDIR/reference_transcripts.1Lv.clean.idx -o
    ↪  $KallistoDIR/$FILESTEM"kallisto-out" -b, --bootstrap-samples=100 --threads=8
    ↪  --pseudobam $file $FILESTEM"R2_clean.fq.gz" || error_exit "$LINENO: kallisto
    ↪  error at $FILESTEM"
```

The full kallisto terminal output is here kallisto-clstr-clean-log.txt.

4.2 EdgeR Expression Analysis

Preliminary data exploration for these data has been included twice for different groupings. The first analysis below including all original eight samples, and other group exploration is listed at the end of this document.

Samples G6 and G7 have ultimately been removed due to batch effects caused by sequencing runs.

4.3 Six *T. adelaidensis* individuals collected between two seasonal periods.

(These methods exclude samples G6 and G7).

This document separates samples by Seasonal group factors AND Sex Group Factors separately to explore the effects on the data.

Import the data:

```
# create paths
setwd("~/@Uni/@Flinders University PhD/RWorkingDir/R-kallisto-clstr-6&7/kallisto-clstr/")
DIR <- ("~/@Uni/@Flinders University
↪  PhD/RWorkingDir/R-kallisto-clstr-6&7/kallisto-clstr/")

head(DIR)
```

```
## [1] "~/@Uni/@Flinders University PhD/RWorkingDir/R-kallisto-clstr-6&7/kallisto-clstr/"
```

```
paths <- list.dirs(path = DIR, full.names = FALSE, recursive = FALSE)
# the working dir only includes kallisto output for the
# kidneys at this time
head(paths)
```

```
## [1] "G1_SepF" "G2_SepF" "G3_MarF" "G4_MarM" "G5_SepM" "G8_MarF"
```

```
Kcaught <- catchKallisto(paths, verbose = TRUE)
```

```
## Reading G1_SepF, 9813 transcripts, 100 bootstraps
## Reading G2_SepF, 9813 transcripts, 100 bootstraps
## Reading G3_MarF, 9813 transcripts, 100 bootstraps
## Reading G4_MarM, 9813 transcripts, 100 bootstraps
## Reading G5_SepM, 9813 transcripts, 100 bootstraps
## Reading G8_MarF, 9813 transcripts, 100 bootstraps
```

```
setwd("~/@Uni/@Flinders University PhD/RWorkingDir/R-kallisto-clstr-6&7")
# View(Kcaught)
```

Set the group factors. Input Sample files are listed in number order G1-G8 excluding G6 and G7. Group factors are applied to the DGElist object and are the point of comparison for expression analyses.

4.3.0.1 Season of collection: 1 = March, 2 = Sep. For the purposes of consistent labelling per group factors some April collections are referred to in the March group. Accurate collection information and dates are outlined in the methods chapter.

```
# these factors correlate to season of collection. 1 =
# March, 2 = Sep.
season_group <- factor(c(2, 2, 1, 1, 2, 1))

# these factors correlate to individuals sex. 1 = Female, 2
# = Male
sex_group <- factor(c(1, 1, 1, 2, 2, 1))
```

4.3.0.2 Individuals sex: 1 = Female, 2 = Male.

4.3.1 Group Factor: Season

Create the EdgeR DGE list for use in subsequent analyses

```
dge_Season <- DGEList(counts = Kcaught$counts/Kcaught$annotation$Overdispersion,
    genes = Kcaught$annotation, group = season_group)
# View(dge_Season) names(dge_Season)
dge_Season
```

```
## An object of class "DGEList"
## $counts
##                                                              G1_SepF      G2_SepF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304           1.790000e+02 118.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564          8.410000e+02 746.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534          3.682740e-07   1.363036
## PB.11.1|004815|path2:1-3039(+)|transcript/3426           2.639676e+01  28.388835
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401       0.000000e+00   0.000000
##                                                              G3_MarF      G4_MarM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304            49.0000000 113.0000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564          645.0000000 997.0000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534            0.4894863   0.9870162
## PB.11.1|004815|path2:1-3039(+)|transcript/3426            10.4093586   4.4945206
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401         0.0000000   0.0000000
##                                                             G5_SepM  G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304             77.00000   147.000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564           415.00000  1577.000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534             0.00000     0.000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426            29.50226    12.422
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401          0.00000     0.000
## 9808 more rows ...
##
## $samples
##         group lib.size norm.factors
## G1_SepF     2  5796184            1
## G2_SepF     2  4480604            1
## G3_MarF     1  3867272            1
## G4_MarM     1  4517543            1
## G5_SepM     2  3518940            1
## G8_MarF     1  4920453            1
##
## $genes
##                                                     Length EffectiveLength
## PB.2.1|002537|path0:1-1624(+)|transcript/18304        1530        1197.281
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564       2065        1732.281
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       1612        1279.281
## PB.11.1|004815|path2:1-3039(+)|transcript/3426        2999        2666.281
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    1965        1632.281
##                                                     Overdispersion
## PB.2.1|002537|path0:1-1624(+)|transcript/18304            1.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564          1.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534          3.253130
## PB.11.1|004815|path2:1-3039(+)|transcript/3426           1.288037
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401       1.000000
## 9808 more rows ...
```

Obtain Counts Per Million to standardise count data comparison

```
myTPM_Season <- dge_Season$counts
# head(myTPM_Season)
```

4.3.2 Initial Data Exploration

Which values in myCPM are greater than 0.5? This produces a logical matrix with TRUEs and FALSEs. TRUE values are samples with > 0.5 counts in that sample per million

```
thresh_Season <- myTPM_Season > 0.5
# This produces a logical matrix with TRUEs and FALSEs
head(thresh_Season)
```

```
##                                               G1_SepF G2_SepF G3_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304    TRUE    TRUE    TRUE
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564    TRUE    TRUE    TRUE
## PB.10.1|004815|path1:5-1713(+)|transcript/17534   FALSE    TRUE   FALSE
## PB.11.1|004815|path2:1-3039(+)|transcript/3426     TRUE    TRUE    TRUE
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 FALSE   FALSE   FALSE
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993     TRUE    TRUE    TRUE
##                                               G4_MarM G5_SepM G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304    TRUE    TRUE    TRUE
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564    TRUE    TRUE    TRUE
## PB.10.1|004815|path1:5-1713(+)|transcript/17534    TRUE   FALSE   FALSE
## PB.11.1|004815|path2:1-3039(+)|transcript/3426     TRUE    TRUE    TRUE
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 FALSE   FALSE   FALSE
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993     TRUE    TRUE    TRUE
```

Insert colour palette for following visualisations consistency

```
library(RColorBrewer)
myPalette <- c("#999999", "#F0E442", "#56B4E9", "#009E73", "#D55E00",
    "#CC79A7")
```

Plot CPM distribution with logged CPM

```r
# export the plot as png
png("plotDensities_Season.png", width = 600)
unfilteredExpr_Season <- cpm(dge_Season, log = T)
plotDensities(unfilteredExpr_Season, col = myPalette, legend = TRUE)
```
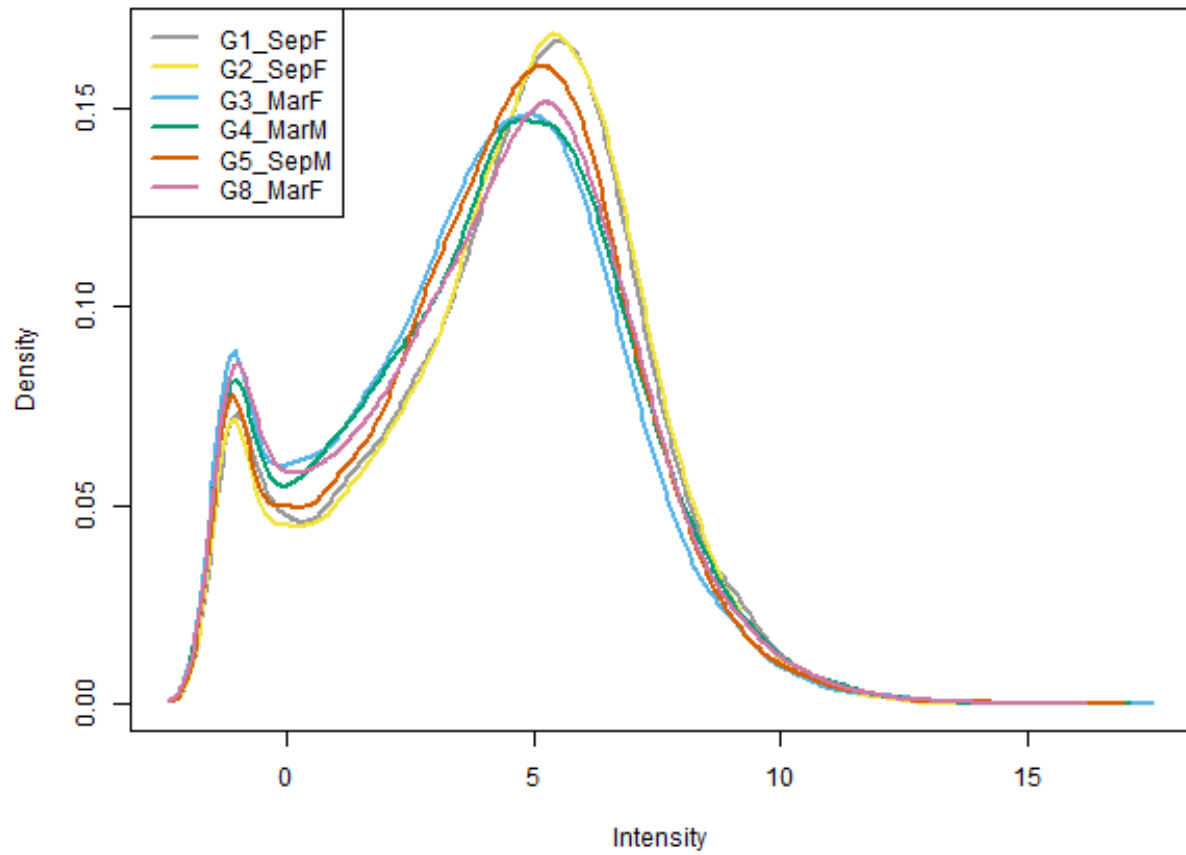

Figure 9: Log Counts per million Densities, for poly-a selected mRNA expression data from six *T. adelaidensis* kidney samples

Visualise library sizes per sample

```
# export the plot as png
png("LibSize_Season.png", width = 600)
# The names argument tells the barplot to use the sample
# names on the x-axis The last argument rotates the axis
# names
barplot(dge_Season$samples$lib.size, names = colnames(dge_Season),
    las = 2, col = myPalette, legend = TRUE)
# title('Library size per sample')
```
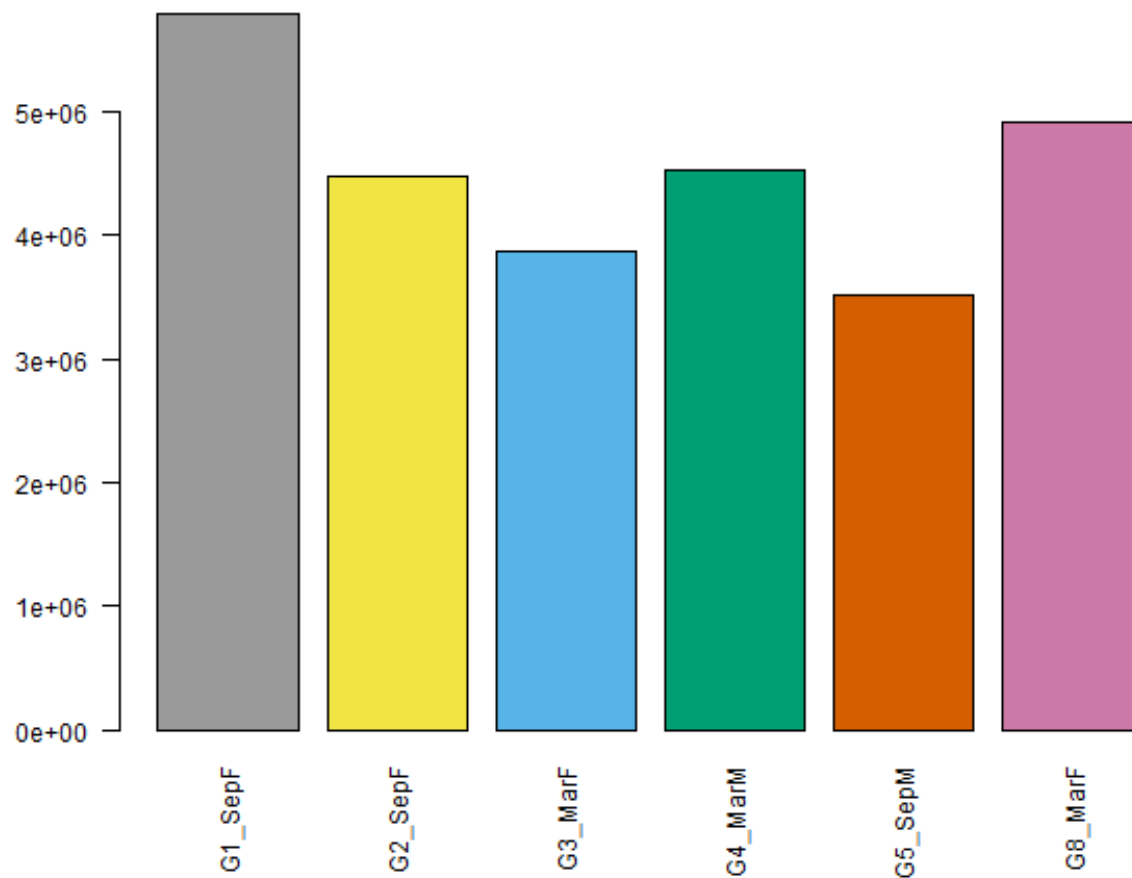

Figure 10: Library size of poly-a selected mRNA sequenced from kidney tissue of six *T. adelaidensis* individuals

Visualise total estimated transcript counts per sample

```
# export the plot as png
png("boxplot_Season.png", width = 600)
boxplot(dge_Season$counts, col = myPalette, las = 2, legend = TRUE)
# title('Boxplot of transcript total estimated counts')
```
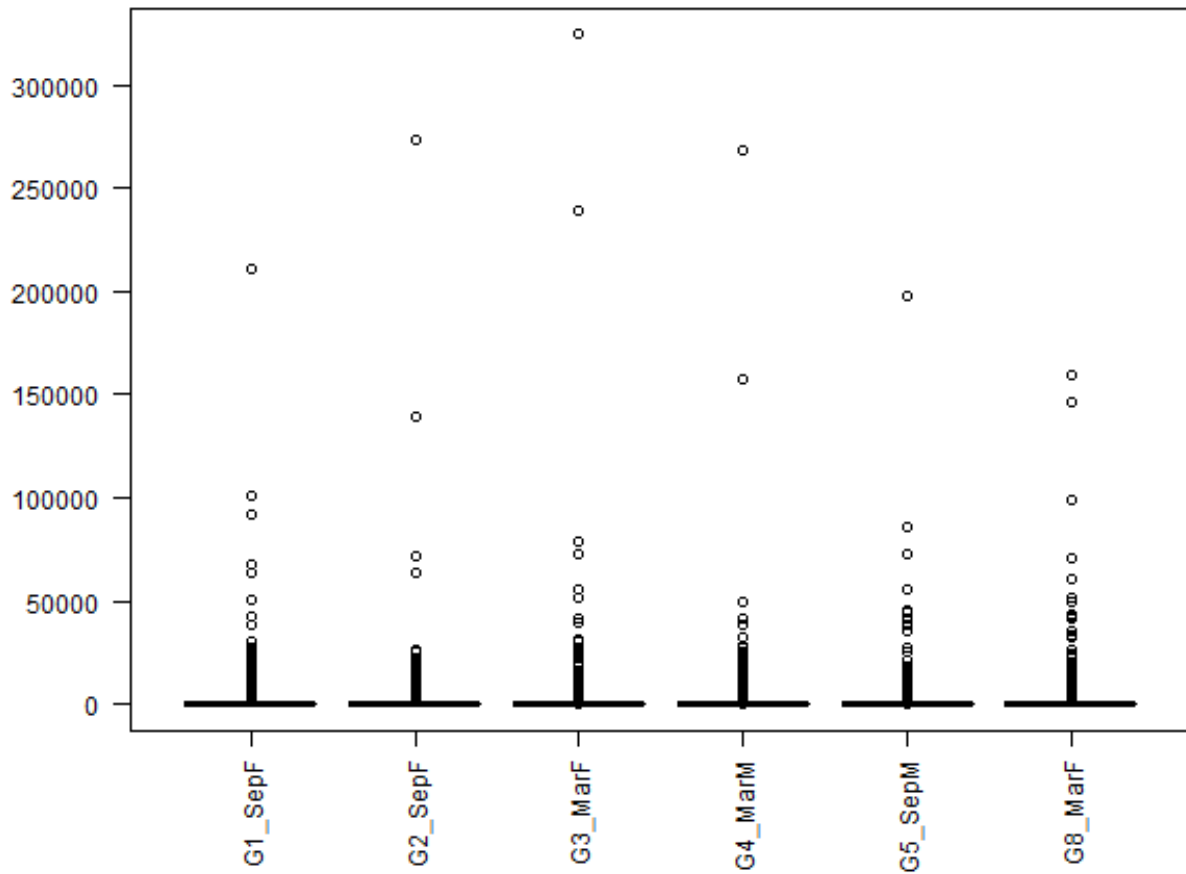

Figure 11: Total estimated counts for poly-a selected mRNA expression data from six *T. adelaidensis* kidney samples

Visualise logged total estimated transcript counts per sample

```r
# Get log2 counts per million
logcounts_Season <- cpm(dge_Season, log = TRUE)
# Check distributions of samples using boxplots export the
# plot as png
png("boxplot-logcounts_Season.png", width = 600)
boxplot(logcounts_Season, xlab = "", ylab = "Log2 counts per million",
    las = 2, col = myPalette, legend = TRUE)
# Let's add a blue horizontal line that corresponds to the
# median logCPM
abline(h = median(logcounts_Season), col = myPalette, legend = TRUE)
# title('Boxplots of log Counts Per Million')
```
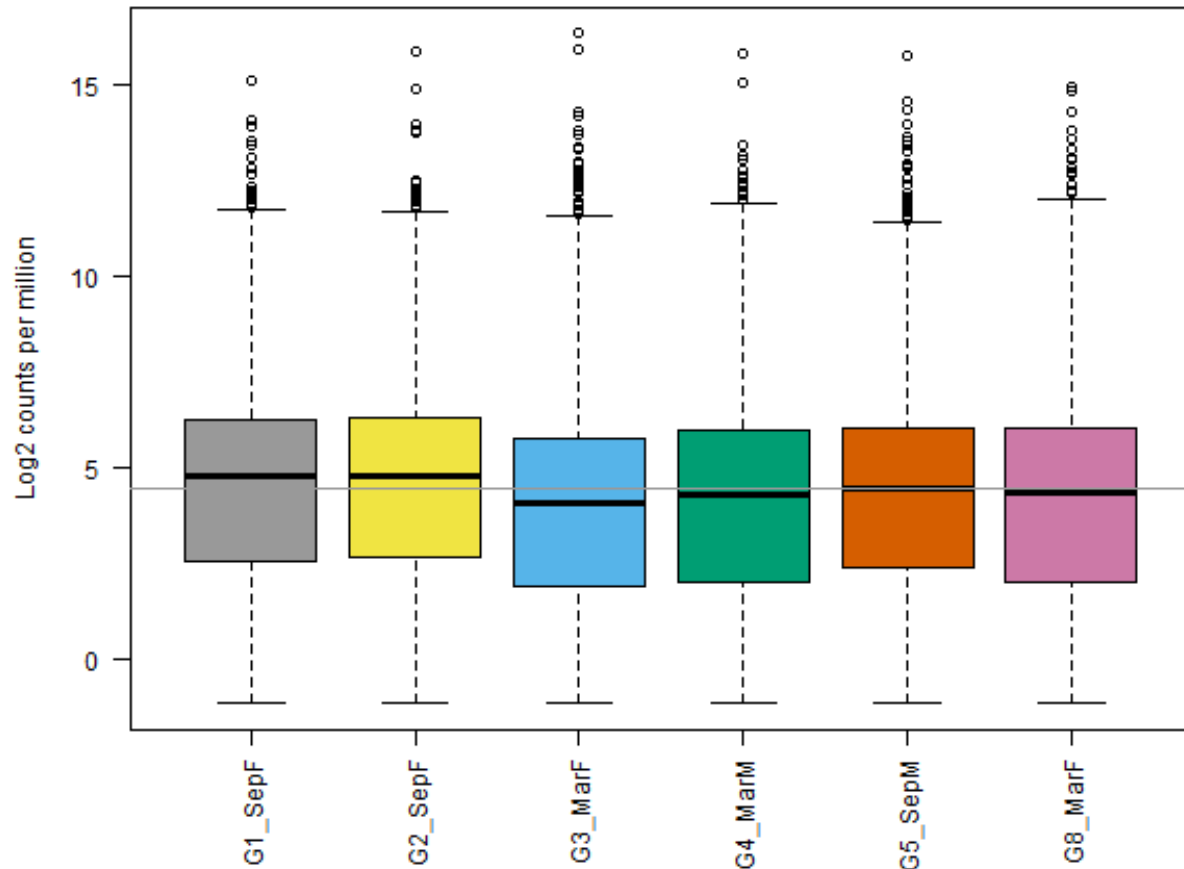

Figure 12: Log2 of estimated counts per million, for poly-a selected mRNA expression data from six *T. adelaidensis* kidney samples

Visualise sample variation using an MDS plot

```
# export the plot as png
png("MDS_Season.png", width = 600)
plotMDS(dge_Season, col = myPalette)
```
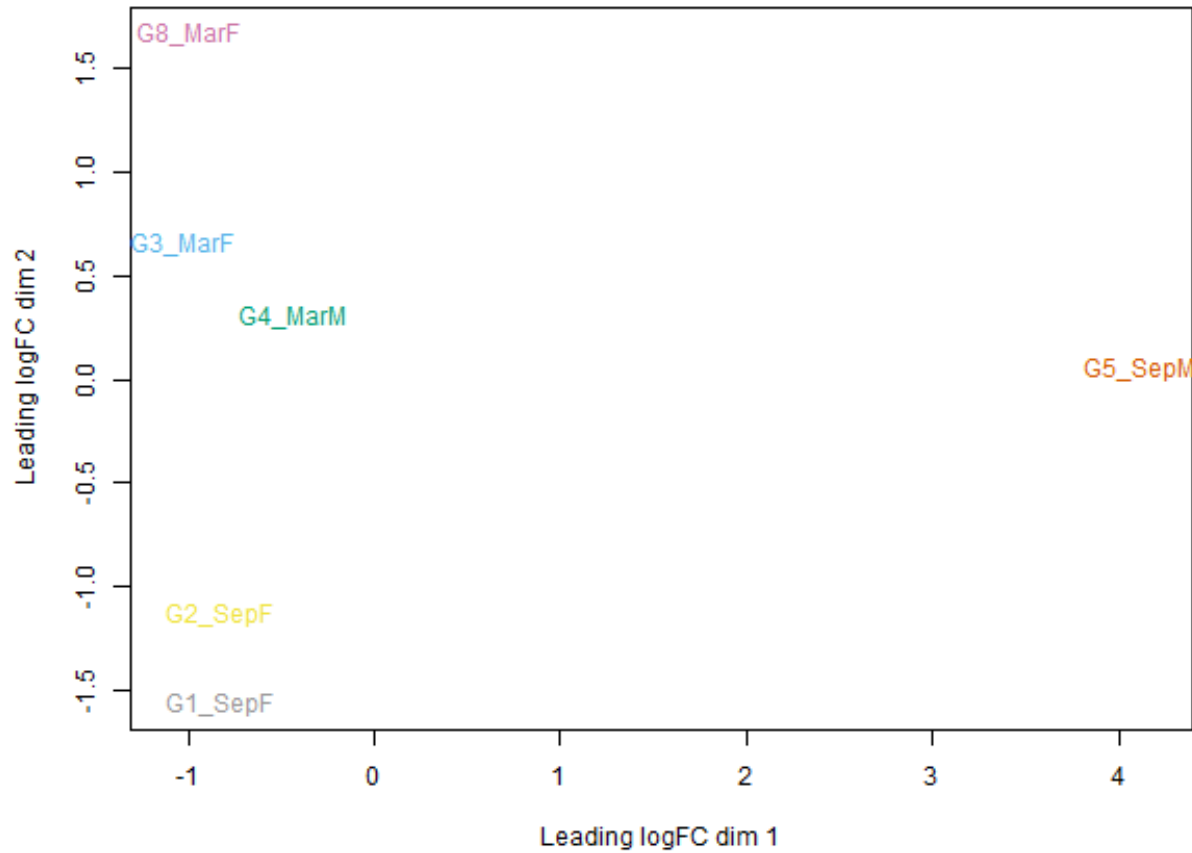

Figure 13: MDS plot of variation among six *T. adelaidensis* individuals

Prepare data for a HeatMap (based off of [this document](#))

Estimate the variance for each row in the logcounts matrix

```
var_genes_season <- apply(logcounts_Season, 1, var)
head(var_genes_season)
```

```
##      PB.2.1|002537|path0:1-1624(+)|transcript/18304
##                                        2.129658e-01
##      PB.3.2|00361c|path0:43-2240(+)|transcript/10564
##                                        2.533823e-01
##      PB.10.1|004815|path1:5-1713(+)|transcript/17534
##                                        1.112634e-01
##       PB.11.1|004815|path2:1-3039(+)|transcript/3426
##                                        8.949062e-01
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401
##                                        2.129924e-30
##       PB.13.1|004c10|path2:1-2811(+)|transcript/4993
##                                        6.973540e-01
```

Transcript IDs for the top 250 most variable transcripts

```
select_var_decT_season <- names(sort(var_genes_season, decreasing = TRUE))[1:250]
head(select_var_decT_season)
```

```
## [1] "PB.4788.1|b34d56|path10:2-1859(+)|transcript/15227"
## [2] "PB.5298.2|c8653e|path1:2-1500(+)|transcript/19961"
## [3] "PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195"
## [4] "PB.5264.3|c67b04|path4:35-1046(+)|transcript/22702"
## [5] "PB.2493.6|5d0eeb|path5:27-5305(+)|transcript/20823"
## [6] "PB.966.1|257403|path2:1-1611(+)|transcript/18188"
```

Subset logcounts matrix

```
highly_variable_lcpm_season <- logcounts_Season[select_var_decT_season,
    ]
dim(highly_variable_lcpm_season)
```

```
## [1] 250   6
```

```
head(highly_variable_lcpm_season)
```

```
##                                                       G1_SepF      G2_SepF
## PB.4788.1|b34d56|path10:2-1859(+)|transcript/15227  0.7541505   0.82792179
## PB.5298.2|c8653e|path1:2-1500(+)|transcript/19961   8.4777062  -1.17531272
## PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195   2.6269095   2.65902383
## PB.5264.3|c67b04|path4:35-1046(+)|transcript/22702  7.5221514  -1.17531272
## PB.2493.6|5d0eeb|path5:27-5305(+)|transcript/20823  0.1566818  -0.65153689
## PB.966.1|257403|path2:1-1611(+)|transcript/18188   -0.1472913   0.03140609
##                                                       G3_MarF      G4_MarM
```

```
## PB.4788.1|b34d56|path10:2-1859(+)|transcript/15227 -0.04654642 -0.5650195
## PB.5298.2|c8653e|path1:2-1500(+)|transcript/19961   0.96325415  0.2590954
## PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195    2.63110803 10.8339962
## PB.5264.3|c67b04|path4:35-1046(+)|transcript/22702 -0.52485253 -0.5919392
## PB.2493.6|5d0eeb|path5:27-5305(+)|transcript/20823 -1.17531272  4.0999879
## PB.966.1|257403|path2:1-1611(+)|transcript/18188   -0.07249537 -0.1405300
##                                                          G5_SepM     G8_MarF
## PB.4788.1|b34d56|path10:2-1859(+)|transcript/15227 11.56898532 -0.63723533
## PB.5298.2|c8653e|path1:2-1500(+)|transcript/19961   7.49680086  7.98334543
## PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195  10.93486157  2.02845409
## PB.5264.3|c67b04|path4:35-1046(+)|transcript/22702 -0.01304149  8.04184470
## PB.2493.6|5d0eeb|path5:27-5305(+)|transcript/20823  9.85798545 -0.62606240
## PB.966.1|257403|path2:1-1611(+)|transcript/18188   10.40291931 -0.01877115
```

Reorder columns to cluster Season factor groups, and sexes

```
col.order_season <- c("G1_SepF", "G2_SepF", "G5_SepM", "G3_MarF",
    "G8_MarF", "G4_MarM")
highly_variable_lcpm_season <- highly_variable_lcpm_season[,
    col.order_season]
```

```
head(highly_variable_lcpm_season)
```

```
##                                                          G1_SepF     G2_SepF
## PB.4788.1|b34d56|path10:2-1859(+)|transcript/15227  0.7541505  0.82792179
## PB.5298.2|c8653e|path1:2-1500(+)|transcript/19961   8.4777062 -1.17531272
## PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195   2.6269095  2.65902383
## PB.5264.3|c67b04|path4:35-1046(+)|transcript/22702  7.5221514 -1.17531272
## PB.2493.6|5d0eeb|path5:27-5305(+)|transcript/20823  0.1566818 -0.65153689
## PB.966.1|257403|path2:1-1611(+)|transcript/18188   -0.1472913  0.03140609
##                                                          G5_SepM     G3_MarF
## PB.4788.1|b34d56|path10:2-1859(+)|transcript/15227 11.56898532 -0.04654642
## PB.5298.2|c8653e|path1:2-1500(+)|transcript/19961   7.49680086  0.96325415
## PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195  10.93486157  2.63110803
## PB.5264.3|c67b04|path4:35-1046(+)|transcript/22702 -0.01304149 -0.52485253
## PB.2493.6|5d0eeb|path5:27-5305(+)|transcript/20823  9.85798545 -1.17531272
## PB.966.1|257403|path2:1-1611(+)|transcript/18188   10.40291931 -0.07249537
##                                                          G8_MarF     G4_MarM
## PB.4788.1|b34d56|path10:2-1859(+)|transcript/15227 -0.63723533 -0.5650195
## PB.5298.2|c8653e|path1:2-1500(+)|transcript/19961   7.98334543  0.2590954
## PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195   2.02845409 10.8339962
## PB.5264.3|c67b04|path4:35-1046(+)|transcript/22702  8.04184470 -0.5919392
## PB.2493.6|5d0eeb|path5:27-5305(+)|transcript/20823 -0.62606240  4.0999879
## PB.966.1|257403|path2:1-1611(+)|transcript/18188   -0.01877115 -0.1405300
```

Colour pallete

```
mypalette2 <- brewer.pal(11, "RdYlBu")
morecols <- colorRampPalette(mypalette2)
```

Colour vector for season_group factor variable: note - factor groups changed to match column reordering in highly_variable_lcpm March=1 September=2

```r
Reordered_season_group <- factor(c(2, 2, 2, 1, 1, 1))
col.cell_season <- c("chartreuse3", "orange")[Reordered_season_group]
```

Create the HeatMap

```
png(file = "High_var_genes.heatmap-season.png", width = 7.5 *
    300, height = 5 * 300, res = 300, pointsize = 8)
# 5 x 300 pixels height = 5*300, res = 300,# 300 pixels per
# inch
heatmap.2(highly_variable_lcpm_season, col = rev(morecols(50)),
    Colv = FALSE, trace = "none", main = "Top 250 most variable genes across samples",
    ColSideColors = col.cell_season, scale = "row", margins = c(8,
        4), srtCol = 45, labRow = "", )
```
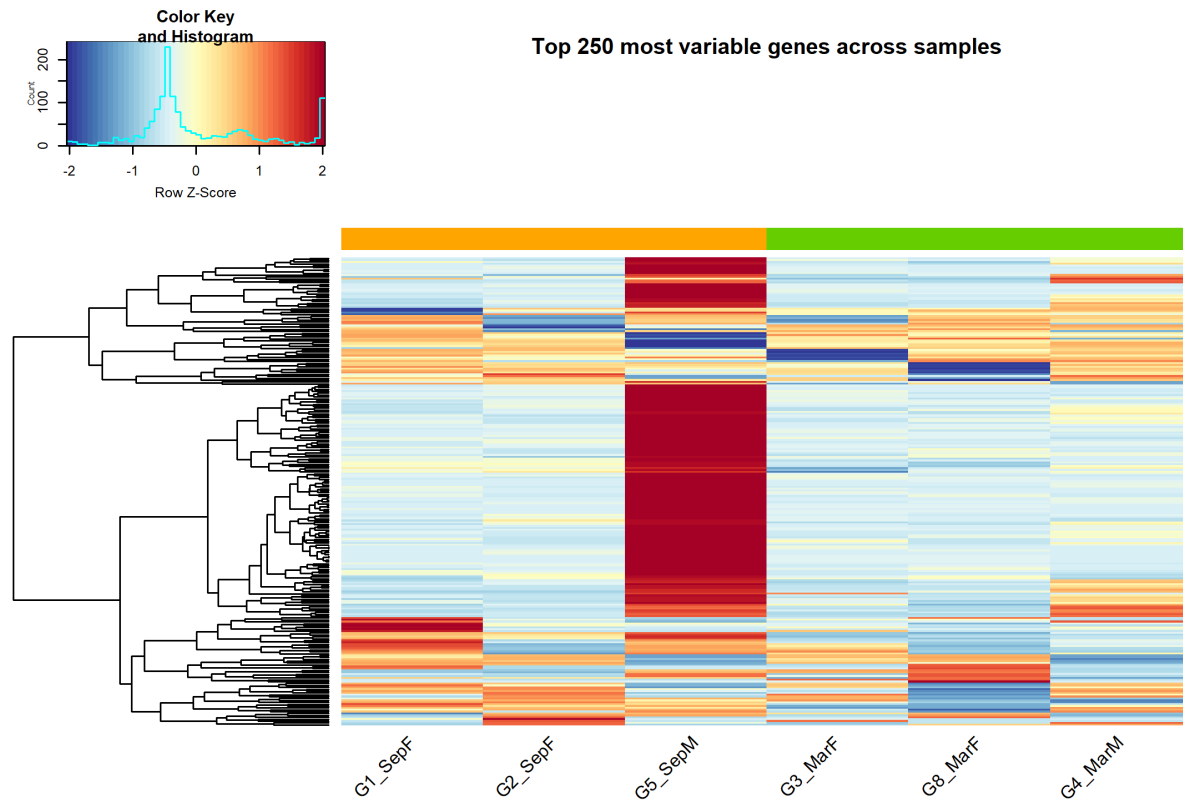

Figure 14: Top 250 most variable transcripts among six *T. adelaidensis* individuals, based on total counts per million values

Transcript IDs for the bottom 250 most variable transcripts = i.e the least variable transcripts

```
select_var_decF_season <- names(sort(var_genes_season, decreasing = FALSE))[1:250]
head(select_var_decF_season)
```

```
## [1] "PB.11.2|004815|path2:1047-3035(+)|transcript/13401"
## [2] "PB.28.1|016cdc|path0:1-842(+)|transcript/23336"
## [3] "PB.32.1|017c97|path0:1-1679(+)|transcript/17216"
## [4] "PB.40.1|019024|path0:1-4793(+)|transcript/8273"
## [5] "PB.130.4|056771|path1:13-5910(+)|transcript/2903"
## [6] "PB.159.1|068d5c|path0:1-2199(+)|transcript/10689"
```

Subset logcounts matrix

```
low_variable_lcpm_season <- logcounts_Season[select_var_decF_season,
    ]
dim(low_variable_lcpm_season)
```

```
## [1] 250    6
```

```
head(low_variable_lcpm_season)
```

```
##                                                         G1_SepF   G2_SepF
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    -1.175313 -1.175313
## PB.28.1|016cdc|path0:1-842(+)|transcript/23336        -1.175313 -1.175313
## PB.32.1|017c97|path0:1-1679(+)|transcript/17216       -1.175313 -1.175313
## PB.40.1|019024|path0:1-4793(+)|transcript/8273        -1.175313 -1.175313
## PB.130.4|056771|path1:13-5910(+)|transcript/2903      -1.175313 -1.175313
## PB.159.1|068d5c|path0:1-2199(+)|transcript/10689      -1.175313 -1.175313
##                                                         G3_MarF   G4_MarM
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    -1.175313 -1.175313
## PB.28.1|016cdc|path0:1-842(+)|transcript/23336        -1.175313 -1.175313
## PB.32.1|017c97|path0:1-1679(+)|transcript/17216       -1.175313 -1.175313
## PB.40.1|019024|path0:1-4793(+)|transcript/8273        -1.175313 -1.175313
## PB.130.4|056771|path1:13-5910(+)|transcript/2903      -1.175313 -1.175313
## PB.159.1|068d5c|path0:1-2199(+)|transcript/10689      -1.175313 -1.175313
##                                                         G5_SepM   G8_MarF
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    -1.175313 -1.175313
## PB.28.1|016cdc|path0:1-842(+)|transcript/23336        -1.175313 -1.175313
## PB.32.1|017c97|path0:1-1679(+)|transcript/17216       -1.175313 -1.175313
## PB.40.1|019024|path0:1-4793(+)|transcript/8273        -1.175313 -1.175313
## PB.130.4|056771|path1:13-5910(+)|transcript/2903      -1.175313 -1.175313
## PB.159.1|068d5c|path0:1-2199(+)|transcript/10689      -1.175313 -1.175313
```

Reorder columns to cluster Season factor groups, and sexes

```
col.order_season <- c("G1_SepF", "G2_SepF", "G5_SepM", "G3_MarF",
    "G8_MarF", "G4_MarM")
low_variable_lcpm_season <- low_variable_lcpm_season[, col.order_season]
```

```r
head(low_variable_lcpm_season)
```

```
##                                                          G1_SepF    G2_SepF
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 -1.175313 -1.175313
## PB.28.1|016cdc|path0:1-842(+)|transcript/23336       -1.175313 -1.175313
## PB.32.1|017c97|path0:1-1679(+)|transcript/17216      -1.175313 -1.175313
## PB.40.1|019024|path0:1-4793(+)|transcript/8273       -1.175313 -1.175313
## PB.130.4|056771|path1:13-5910(+)|transcript/2903     -1.175313 -1.175313
## PB.159.1|068d5c|path0:1-2199(+)|transcript/10689     -1.175313 -1.175313
##                                                          G5_SepM    G3_MarF
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 -1.175313 -1.175313
## PB.28.1|016cdc|path0:1-842(+)|transcript/23336       -1.175313 -1.175313
## PB.32.1|017c97|path0:1-1679(+)|transcript/17216      -1.175313 -1.175313
## PB.40.1|019024|path0:1-4793(+)|transcript/8273       -1.175313 -1.175313
## PB.130.4|056771|path1:13-5910(+)|transcript/2903     -1.175313 -1.175313
## PB.159.1|068d5c|path0:1-2199(+)|transcript/10689     -1.175313 -1.175313
##                                                          G8_MarF    G4_MarM
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 -1.175313 -1.175313
## PB.28.1|016cdc|path0:1-842(+)|transcript/23336       -1.175313 -1.175313
## PB.32.1|017c97|path0:1-1679(+)|transcript/17216      -1.175313 -1.175313
## PB.40.1|019024|path0:1-4793(+)|transcript/8273       -1.175313 -1.175313
## PB.130.4|056771|path1:13-5910(+)|transcript/2903     -1.175313 -1.175313
## PB.159.1|068d5c|path0:1-2199(+)|transcript/10689     -1.175313 -1.175313
```

Colour pallete

```r
mypalette2 <- brewer.pal(11, "RdYlBu")
morecols <- colorRampPalette(mypalette2)
```

Colour vector for season_group factor variable: note - factor groups changed to match column reordering in low_variable_lcpm March=1 September=2

```r
Reordered_season_group <- factor(c(2, 2, 2, 1, 1, 1))
col.cell_season <- c("chartreuse3", "orange")[Reordered_season_group]
```

Create the HeatMap

```
png(file = "Low_var_genes.heatmap-season.png", width = 7.5 *
    300, height = 5 * 300, res = 300, pointsize = 8)
# 5 x 300 pixels height = 5*300, res = 300,# 300 pixels per
# inch
heatmap.2(low_variable_lcpm_season, col = rev(morecols(50)),
    Colv = FALSE, trace = "none", main = "Bottom 250 most variable genes across samples",
    ColSideColors = col.cell_season, scale = "row", margins = c(8,
        4), srtCol = 45, labRow = "", )
```
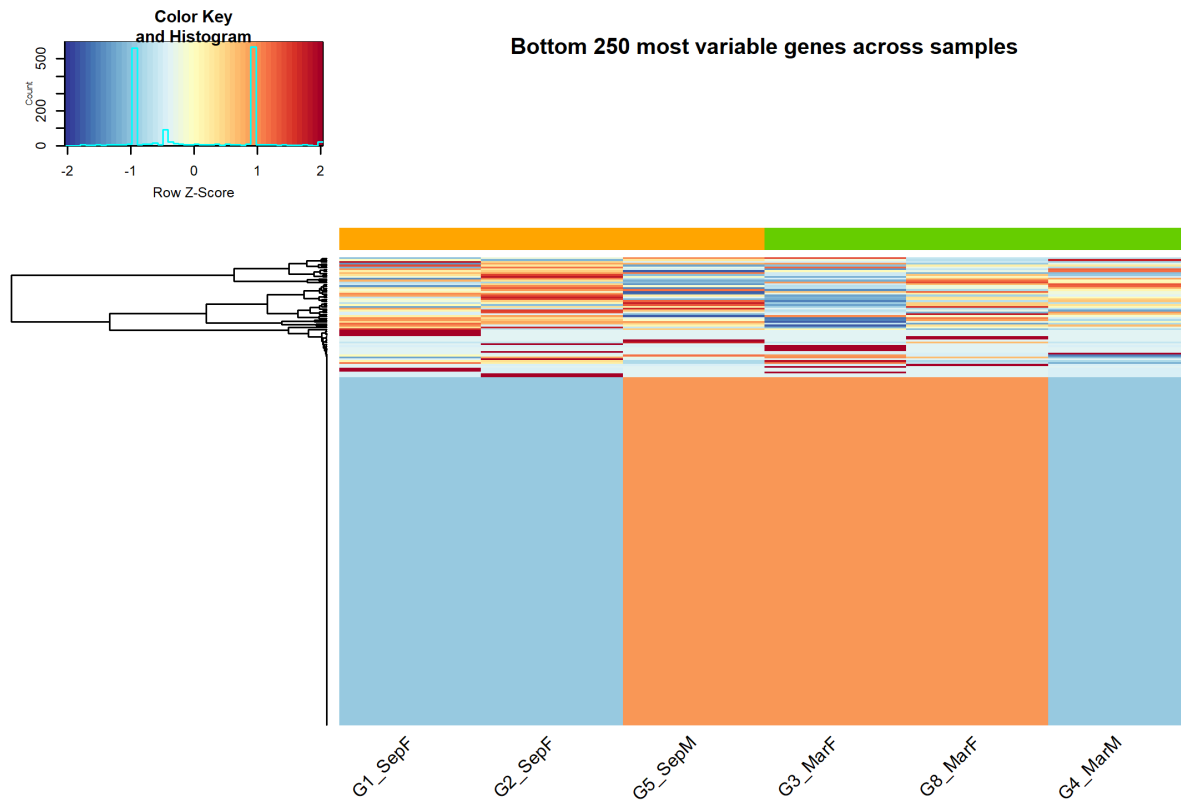

Figure 15: Bottom 250 most variable transcripts among six *T. adelaidensis* individuals, based on total counts per million values

4.3.3 Calculations

```
# Data with actual counts, not TPM = 'dge'
head(dge_Season)
```

```
## An object of class "DGEList"
## $counts
##                                                       G1_SepF      G2_SepF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      1.790000e+02 118.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     8.410000e+02 746.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534     3.682740e-07   1.363036
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      2.639676e+01  28.388835
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 0.000000e+00   0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993      7.179928e+01  33.065458
##                                                       G3_MarF      G4_MarM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304       49.0000000 113.0000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     645.0000000 997.0000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       0.4894863   0.9870162
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       10.4093586   4.4945206
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    0.0000000   0.0000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993       10.3920012  12.2814560
##                                                       G5_SepM      G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304       77.00000   147.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     415.00000  1577.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       0.00000     0.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       29.50226    12.42200
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    0.00000     0.00000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993       10.39200    25.50764
##
## $samples
##         group lib.size norm.factors
## G1_SepF     2  5796184            1
## G2_SepF     2  4480604            1
## G3_MarF     1  3867272            1
## G4_MarM     1  4517543            1
## G5_SepM     2  3518940            1
## G8_MarF     1  4920453            1
##
## $genes
##                                                     Length EffectiveLength
## PB.2.1|002537|path0:1-1624(+)|transcript/18304        1530         1197.281
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564       2065         1732.281
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       1612         1279.281
## PB.11.1|004815|path2:1-3039(+)|transcript/3426        2999         2666.281
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    1965         1632.281
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993        2714         2381.281
##                                                     Overdispersion
## PB.2.1|002537|path0:1-1624(+)|transcript/18304            1.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564           1.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534           3.253130
## PB.11.1|004815|path2:1-3039(+)|transcript/3426            1.288037
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401        1.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993            1.058506
```

```
head(myTPM_Season)
```

```
##                                                      G1_SepF      G2_SepF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304     1.790000e+02 118.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564    8.410000e+02 746.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534    3.682740e-07   1.363036
## PB.11.1|004815|path2:1-3039(+)|transcript/3426     2.639676e+01  28.388835
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 0.000000e+00   0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993     7.179928e+01  33.065458
##                                                        G3_MarF      G4_MarM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304       49.0000000 113.0000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     645.0000000 997.0000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       0.4894863   0.9870162
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       10.4093586   4.4945206
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    0.0000000   0.0000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993       10.3920012  12.2814560
##                                                       G5_SepM     G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304       77.00000   147.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     415.00000  1577.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       0.00000     0.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       29.50226    12.42200
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    0.00000     0.00000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993       10.39200    25.50764
```

Filter out genes with less than 5 reads in 2 samples for each transcript

```
filter_Season <- apply(dge_Season, 1, function(x) length(x[x >
    5]) >= 2)
filtered_Season <- dge_Season[filter_Season, ]
head(filtered_Season)
```

```
## An object of class "DGEList"
## $counts
##                                                    G1_SepF    G2_SepF    G3_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304    179.00000  118.00000   49.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564   841.00000  746.00000  645.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426     26.39676   28.38883   10.40936
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993     71.79928   33.06546   10.39200
## PB.14.1|006283|path0:1-1955(+)|transcript/13821   708.75635  442.39668  326.26755
## PB.15.1|006e23|path0:1-3614(+)|transcript/1774    624.00000  527.00000  351.00000
##                                                     G4_MarM    G5_SepM     G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304    113.000000   77.00000   147.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564   997.000000  415.00000  1577.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      4.494521   29.50226    12.42200
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993     12.281456   10.39200    25.50764
## PB.14.1|006283|path0:1-1955(+)|transcript/13821   508.756184  306.91270   552.99585
## PB.15.1|006e23|path0:1-3614(+)|transcript/1774    336.000000  281.00000   719.00000
##
## $samples
##         group lib.size norm.factors
## G1_SepF     2  5796184            1
## G2_SepF     2  4480604            1
```

```
## G3_MarF       1  3867272               1
## G4_MarM       1  4517543               1
## G5_SepM       2  3518940               1
## G8_MarF       1  4920453               1
##
## $genes
##                                                   Length EffectiveLength
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      1530         1197.281
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     2065         1732.281
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      2999         2666.281
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993      2714         2381.281
## PB.14.1|006283|path0:1-1955(+)|transcript/13821     1920         1587.281
## PB.15.1|006e23|path0:1-3614(+)|transcript/1774      3583         3250.281
##                                                   Overdispersion
## PB.2.1|002537|path0:1-1624(+)|transcript/18304           1.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564          1.000000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426           1.288037
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993           1.058506
## PB.14.1|006283|path0:1-1955(+)|transcript/13821          1.084999
## PB.15.1|006e23|path0:1-3614(+)|transcript/1774           1.000000
```

Run calculations

```
dge_Season <- calcNormFactors(dge_Season)
dge_Season <- estimateCommonDisp(dge_Season)
dge_Season <- estimateTagwiseDisp(dge_Season)
```

4.3.4 Differentially expressed genes:

View the top ten 'differentially expressed genes'

```
results_Season <- exactTest(dge_Season)
topTags(results_Season)
```

```
## Comparison of groups:  2-1
##                                                     Length EffectiveLength
## PB.3849.1|906dad|path19:1-2010(+)|transcript/12836    1940        1607.2807
## PB.499.11|124cf9|path4:2108-4109(+)|transcript/7418   2000        1667.2807
## PB.7745.1|transcript/17825:1-1637(+)|transcript/17825 1608        1275.2807
## PB.8897.1|transcript/3462:1-3118(+)|transcript/3462   3086        2753.2807
## PB.2869.1|6b0af0|path2:1-1117(+)|transcript/22387     1086         753.2807
## PB.7443.1|transcript/15548:1-1797(+)|transcript/15548 1750        1417.2807
## PB.2003.2|4ab8e8|path0:1458-4018(+)|transcript/6743   2535        2202.2807
## PB.499.1|124cf9|path4:1-2080(+)|transcript/11611      2071        1738.2807
## PB.3885.3|918572|path1:1406-3716(+)|transcript/9242   2278        1945.2807
## PB.4788.1|b34d56|path10:2-1859(+)|transcript/15227    1771        1438.2807
##                                                     Overdispersion      logFC
## PB.3849.1|906dad|path19:1-2010(+)|transcript/12836        3.166184   4.937224
## PB.499.11|124cf9|path4:2108-4109(+)|transcript/7418       1.621394   2.891897
## PB.7745.1|transcript/17825:1-1637(+)|transcript/17825     1.000000   5.024827
## PB.8897.1|transcript/3462:1-3118(+)|transcript/3462       1.044347   2.898051
## PB.2869.1|6b0af0|path2:1-1117(+)|transcript/22387         1.242289   3.088832
```

```
## PB.7443.1|transcript/15548:1-1797(+)|transcript/15548       1.000000  4.336311
## PB.2003.2|4ab8e8|path0:1458-4018(+)|transcript/6743         1.179403  2.473788
## PB.499.1|124cf9|path4:1-2080(+)|transcript/11611            1.358840  2.448778
## PB.3885.3|918572|path1:1406-3716(+)|transcript/9242         5.786371 -5.384475
## PB.4788.1|b34d56|path10:2-1859(+)|transcript/15227          1.141272 11.520001
##                                                             logCPM       PValue
## PB.3849.1|906dad|path19:1-2010(+)|transcript/12836         5.413096 3.967872e-09
## PB.499.11|124cf9|path4:2108-4109(+)|transcript/7418        5.501158 5.425738e-08
## PB.7745.1|transcript/17825:1-1637(+)|transcript/17825      3.551103 2.893073e-07
## PB.8897.1|transcript/3462:1-3118(+)|transcript/3462        5.042973 8.245973e-07
## PB.2869.1|6b0af0|path2:1-1117(+)|transcript/22387          4.962496 8.724447e-07
## PB.7443.1|transcript/15548:1-1797(+)|transcript/15548      3.466335 2.370305e-06
## PB.2003.2|4ab8e8|path0:1458-4018(+)|transcript/6743        5.788082 3.024335e-06
## PB.499.1|124cf9|path4:1-2080(+)|transcript/11611           5.654586 3.673456e-06
## PB.3885.3|918572|path1:1406-3716(+)|transcript/9242        1.510773 3.716870e-06
## PB.4788.1|b34d56|path10:2-1859(+)|transcript/15227         9.130702 3.880790e-06
##                                                                   FDR
## PB.3849.1|906dad|path19:1-2010(+)|transcript/12836         3.893673e-05
## PB.499.11|124cf9|path4:2108-4109(+)|transcript/7418        2.662138e-04
## PB.7745.1|transcript/17825:1-1637(+)|transcript/17825      9.463241e-04
## PB.8897.1|transcript/3462:1-3118(+)|transcript/3462        1.712260e-03
## PB.2869.1|6b0af0|path2:1-1117(+)|transcript/22387          1.712260e-03
## PB.7443.1|transcript/15548:1-1797(+)|transcript/15548      3.488516e-03
## PB.2003.2|4ab8e8|path0:1458-4018(+)|transcript/6743        3.488516e-03
## PB.499.1|124cf9|path4:1-2080(+)|transcript/11611           3.488516e-03
## PB.3885.3|918572|path1:1406-3716(+)|transcript/9242        3.488516e-03
## PB.4788.1|b34d56|path10:2-1859(+)|transcript/15227         3.488516e-03
```

Output the top 25 table of values based on calculated PValue

```
tab_Season <- topTags(results_Season, n = 25, sort.by = "PValue")
head(tab_Season)
```

```
## Comparison of groups:  2-1
##                                                             Length EffectiveLength
## PB.3849.1|906dad|path19:1-2010(+)|transcript/12836           1940        1607.2807
## PB.499.11|124cf9|path4:2108-4109(+)|transcript/7418          2000        1667.2807
## PB.7745.1|transcript/17825:1-1637(+)|transcript/17825        1608        1275.2807
## PB.8897.1|transcript/3462:1-3118(+)|transcript/3462          3086        2753.2807
## PB.2869.1|6b0af0|path2:1-1117(+)|transcript/22387            1086         753.2807
## PB.7443.1|transcript/15548:1-1797(+)|transcript/15548        1750        1417.2807
##                                                            Overdispersion     logFC
## PB.3849.1|906dad|path19:1-2010(+)|transcript/12836               3.166184 4.937224
## PB.499.11|124cf9|path4:2108-4109(+)|transcript/7418             1.621394 2.891897
## PB.7745.1|transcript/17825:1-1637(+)|transcript/17825           1.000000 5.024827
## PB.8897.1|transcript/3462:1-3118(+)|transcript/3462             1.044347 2.898051
## PB.2869.1|6b0af0|path2:1-1117(+)|transcript/22387               1.242289 3.088832
## PB.7443.1|transcript/15548:1-1797(+)|transcript/15548           1.000000 4.336311
##                                                             logCPM       PValue
## PB.3849.1|906dad|path19:1-2010(+)|transcript/12836         5.413096 3.967872e-09
## PB.499.11|124cf9|path4:2108-4109(+)|transcript/7418        5.501158 5.425738e-08
## PB.7745.1|transcript/17825:1-1637(+)|transcript/17825      3.551103 2.893073e-07
## PB.8897.1|transcript/3462:1-3118(+)|transcript/3462        5.042973 8.245973e-07
```

```
## PB.2869.1|6b0af0|path2:1-1117(+)|transcript/22387      4.962496 8.724447e-07
## PB.7443.1|transcript/15548:1-1797(+)|transcript/15548 3.466335 2.370305e-06
##                                                             FDR
## PB.3849.1|906dad|path19:1-2010(+)|transcript/12836     3.893673e-05
## PB.499.11|124cf9|path4:2108-4109(+)|transcript/7418    2.662138e-04
## PB.7745.1|transcript/17825:1-1637(+)|transcript/17825 9.463241e-04
## PB.8897.1|transcript/3462:1-3118(+)|transcript/3462    1.712260e-03
## PB.2869.1|6b0af0|path2:1-1117(+)|transcript/22387      1.712260e-03
## PB.7443.1|transcript/15548:1-1797(+)|transcript/15548 3.488516e-03
```

```r
write.table(tab_Season, file = "Kgenelist-Top25_Season.txt")
```

Output the entire table of values

```r
tab_Season_all <- topTags(results_Season, n = Inf, sort.by = "PValue")
write.table(tab_Season_all, file = "Kgenelist_Season.txt")
```

Output summary expression data

```r
dim(results_Season)
```

```
## [1] 9813    3
```

```r
summary(de_Season <- decideTests(results_Season))
```

```
##           2-1
## Down       25
## NotSig 9631
## Up        157
```

```r
summary_Season <- (de_Season <- decideTests(results_Season))
head(summary_Season)
```

```
## TestResults matrix
##                                                       2-1
## PB.2.1|002537|path0:1-1624(+)|transcript/18304          0
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564         0
## PB.10.1|004815|path1:5-1713(+)|transcript/17534         0
## PB.11.1|004815|path2:1-3039(+)|transcript/3426          0
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401      0
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993          0
```

```r
write.table(summary_Season, file = "summary_Season.txt")
```

```r
detags_Season <- rownames(dge_Season)[as.logical(de_Season)]
head(detags_Season)
```

```
## [1] "PB.147.1|0620ba|path0:1-1756(+)|transcript/23387"
## [2] "PB.193.1|07b09f|path0:1-2724(+)|transcript/21409"
## [3] "PB.193.7|07b09f|path0:5-2523(+)|transcript/22910"
## [4] "PB.222.2|0889ce|path13:1-894(+)|transcript/23126"
## [5] "PB.225.1|089810|path3:1-2255(+)|transcript/24270"
## [6] "PB.329.1|0bd3d9|path3:1-656(+)|transcript/24223"
```

```
head(results_Season)
```

```
## An object of class "DGEExact"
## $table
##                                                          logFC      logCPM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      -0.01384282   4.6123326
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     -0.97428387   7.5877885
## PB.10.1|004815|path1:5-1713(+)|transcript/17534     -0.37656880  -0.9010547
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       1.38669548   2.1987804
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401   0.00000000  -1.2129277
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993       0.79643615   2.5138513
##                                                          PValue
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      0.98619206
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     0.05750860
## PB.10.1|004815|path1:5-1713(+)|transcript/17534     1.00000000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      0.06492398
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401  1.00000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993      0.26397635
##
## $comparison
## [1] "1" "2"
##
## $genes
##                                                          Length EffectiveLength
## PB.2.1|002537|path0:1-1624(+)|transcript/18304            1530        1197.281
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564           2065        1732.281
## PB.10.1|004815|path1:5-1713(+)|transcript/17534           1612        1279.281
## PB.11.1|004815|path2:1-3039(+)|transcript/3426            2999        2666.281
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401        1965        1632.281
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993            2714        2381.281
##                                                          Overdispersion
## PB.2.1|002537|path0:1-1624(+)|transcript/18304              1.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564            1.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534            3.253130
## PB.11.1|004815|path2:1-3039(+)|transcript/3426             1.288037
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401         1.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993             1.058506
```

Filter and sort this summary to only include upregulated and downregulated transcripts

```
summary_Season.sort <- read.csv("./summary_Season.txt", sep = " ",
    row.names = NULL, header = TRUE, stringsAsFactors = FALSE)
summary_Season.sort <- summary_Season.sort %>%
    rename(TranscriptID_3 = row.names, Regulation = X2.1)  #the name 'TranscriptID_3' is
↪    used for consistency and joining as seen in the next step
```

fiter out the transcripts which are significantly upregulated when compared by season

```
summary_Season.sort_UP <- filter(summary_Season.sort, Regulation ==
    "1")  #upregulated transcripts denoted as +1
head(summary_Season.sort_UP)
```

```
##                                      TranscriptID_3 Regulation
## 1 PB.147.1|0620ba|path0:1-1756(+)|transcript/23387          1
## 2 PB.193.1|07b09f|path0:1-2724(+)|transcript/21409          1
## 3 PB.193.7|07b09f|path0:5-2523(+)|transcript/22910          1
## 4 PB.222.2|0889ce|path13:1-894(+)|transcript/23126          1
## 5 PB.225.1|089810|path3:1-2255(+)|transcript/24270          1
## 6  PB.329.1|0bd3d9|path3:1-656(+)|transcript/24223          1
```

```
nrow(summary_Season.sort_UP)
```

```
## [1] 157
```

Extract the full information about the putative identiry of these transcripts from the Summary file created in Section 3.2.3 above

```
summary_Season.sort_UP_info <- semi_join(BLASTx.gene.join_Database.Full.PBT.Summary,
    summary_Season.sort_UP, by = "TranscriptID_3")
# head (summary_Season.sort_UP_info)
nrow(summary_Season.sort_UP_info)
```

```
## [1] 93
```

```
write.csv(summary_Season.sort_UP_info, "~/@Uni/@Flinders University
↪  PhD/RWorkingDir/Workflow-readthedown/ReferenceClusters/summary_Season.sort_UP_info.csv",
    quote = FALSE, row.names = FALSE, col.names = TRUE)  #write the file to a .csv
```

(note: although there are 157 transcripts significantly upregulated, only 93 of these returned a BLASTx result for their corresponding putative coding region)

Do the same for the Seasonally deownregulated transcripts

```
summary_Season.sort_DOWN <- filter(summary_Season.sort, Regulation ==
    "-1")  #downregulated transcripts denoted as -1
head(summary_Season.sort_DOWN)
```

```
##                                      TranscriptID_3 Regulation
## 1  PB.879.1|206df6|path7:41-2171(+)|transcript/12507         -1
## 2   PB.1436.1|365ff0|path0:1-3058(+)|transcript/6006         -1
## 3   PB.2872.1|6b0af0|path5:1-947(+)|transcript/23097         -1
## 4   PB.3308.1|7b2e7d|path0:1-2337(+)|transcript/9156         -1
## 5 PB.3445.2|8111f0|path13:1-2049(+)|transcript/12812         -1
## 6 PB.3445.5|8111f0|path13:4-2098(+)|transcript/14262         -1
```

```
nrow(summary_Season.sort_DOWN)
```

[1] 25

Extract the full information about the putative identiry of these transcripts from the Summary file created
in Section 3.2.3 above

```
summary_Season.sort_DOWN_info <- semi_join(BLASTx.gene.join_Database.Full.PBT.Summary,
    summary_Season.sort_DOWN, by = "TranscriptID_3")
# head (summary_Season.sort_DOWN_info)
nrow(summary_Season.sort_DOWN_info)
```

[1] 20

```
write.csv(summary_Season.sort_DOWN_info, "~/@Uni/@Flinders University
↪  PhD/RWorkingDir/Workflow-readthedown/ReferenceClusters/summary_Season.sort_DOWN_info.csv",
    quote = FALSE, row.names = FALSE, col.names = TRUE)  #write the file to a .csv
```

(note: although there are 25 transcripts significantly downregulated, only 20 of these returned a BLASTx
result for their corresponding putative coding region)

Visualise the logged Fold Change against Average logged counts per million, in expression of transcripts between season 1 (March/April) and 2 (September) using plotsmear: red indicates PValue < 0.05

```
png("plotsmear_Season.png", width = 600)
plotSmear(results_Season, de.tags = detags_Season)
```
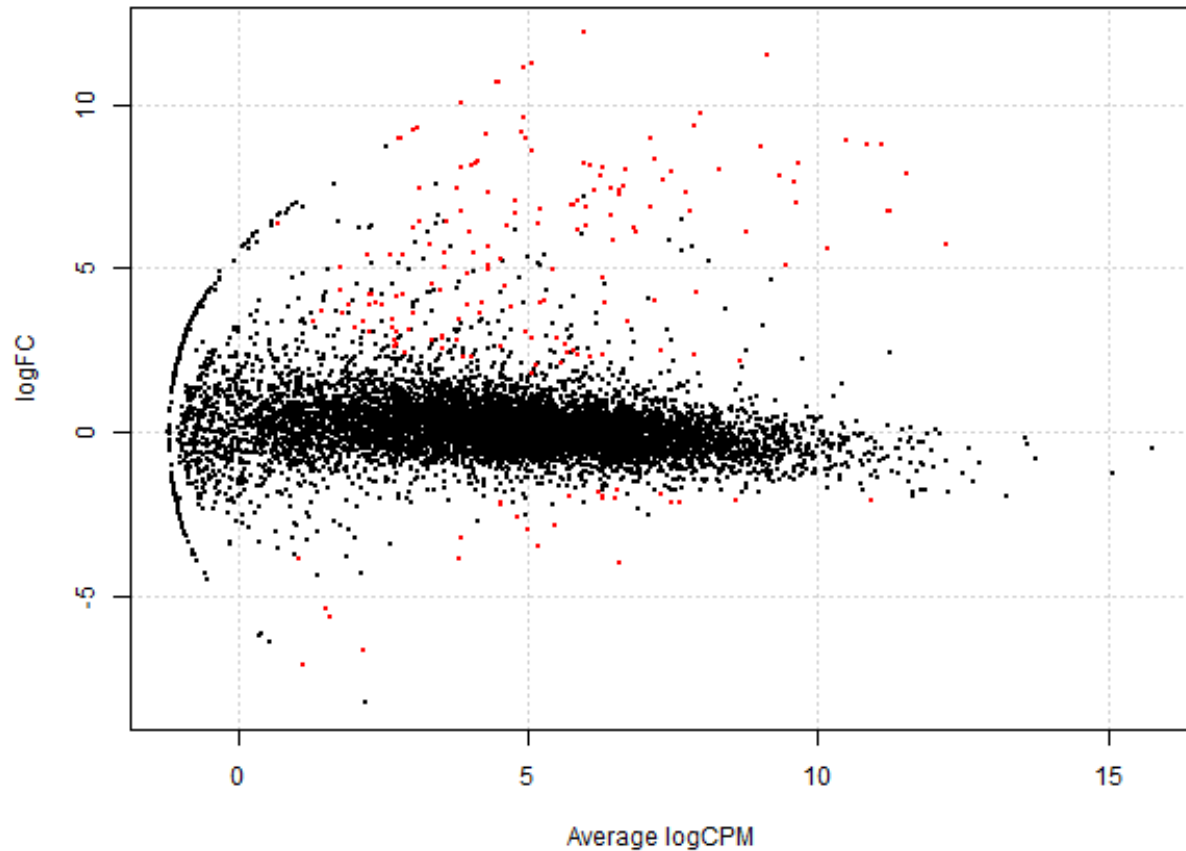

Figure 16: Log Fold Change vs average Log counts per million for expression level comparing September to March season variables among six *T. adelaidensis* individuals. Red dots indicate transcripts with a significant change over season, with p-values <0.05.

Visualise differential expression of transcripts between season 1 (March/April) and 2 (September) with negative Logged PValue against Logged Fold Change in expression: Red indicates a transcript with PValue < 0.05, and Log Fold change > 2

```r
volcanoData_Season <- cbind(results_Season$table$logFC,
↪   -log10(results_Season$table$PValue))
colnames(volcanoData_Season) <- c("logFC", "negLogPval")
DEGs_Season <- results_Season$table$PValue < 0.05 & abs(results_Season$table$logFC) >
    2

point.col_Season <- ifelse(DEGs_Season, "red", "black")
png("volcanoData_Season.png", width = 600)
plot(volcanoData_Season, pch = 16, col = point.col_Season, cex = 0.5)
# pch is shape 1-26 are vectors, cex is size
```
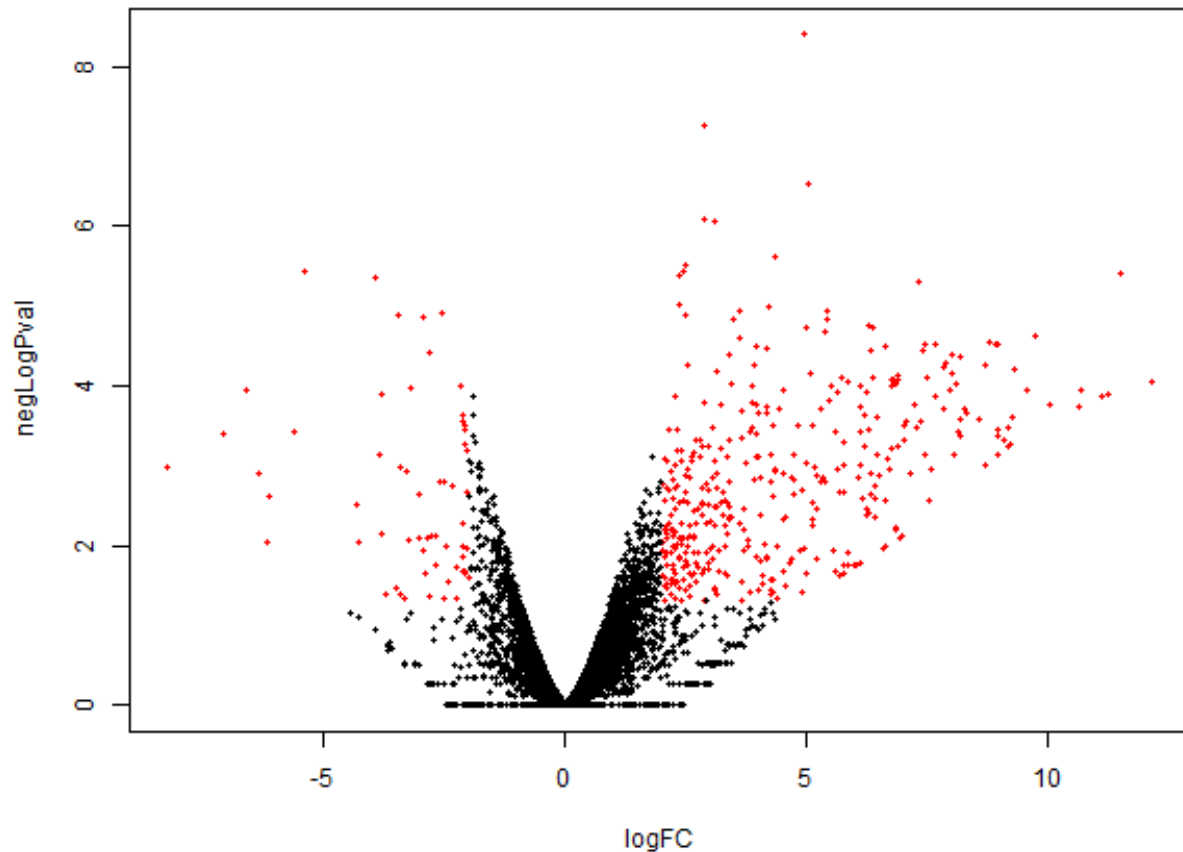

Figure 17: Negative Log of the p-value against Log Fold Change in expression level comparing September to March season variables among six *T. adelaidensis* individuals. Red dots indicate transcripts with a significant change over season, with p-values <0.05 and *also* a log fold change >2.

4.3.5 Group Factor: Sex

Create the EdgeR DGE list for use in subsequent analyses

```
dge_Sex <- DGEList(counts = Kcaught$counts/Kcaught$annotation$Overdispersion,
    genes = Kcaught$annotation, group = sex_group)
# View(dge_Sex) names(dge_Sex)
dge_Sex
```

```
## An object of class "DGEList"
## $counts
##                                                       G1_SepF      G2_SepF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304       1.790000e+02 118.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564      8.410000e+02 746.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534      3.682740e-07   1.363036
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       2.639676e+01  28.388835
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401   0.000000e+00   0.000000
##                                                        G3_MarF      G4_MarM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304         49.0000000 113.0000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564       645.0000000 997.0000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534         0.4894863   0.9870162
## PB.11.1|004815|path2:1-3039(+)|transcript/3426         10.4093586   4.4945206
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401      0.0000000   0.0000000
##                                                       G5_SepM  G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304        77.00000   147.000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564      415.00000  1577.000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534        0.00000     0.000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426        29.50226    12.422
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401     0.00000     0.000
## 9808 more rows ...
##
## $samples
##           group lib.size norm.factors
## G1_SepF      1   5796184            1
## G2_SepF      1   4480604            1
## G3_MarF      1   3867272            1
## G4_MarM      2   4517543            1
## G5_SepM      2   3518940            1
## G8_MarF      1   4920453            1
##
## $genes
##                                                      Length EffectiveLength
## PB.2.1|002537|path0:1-1624(+)|transcript/18304         1530        1197.281
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564        2065        1732.281
## PB.10.1|004815|path1:5-1713(+)|transcript/17534        1612        1279.281
## PB.11.1|004815|path2:1-3039(+)|transcript/3426         2999        2666.281
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401     1965        1632.281
##                                                      Overdispersion
## PB.2.1|002537|path0:1-1624(+)|transcript/18304             1.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564            1.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534            3.253130
```

```
## PB.11.1|004815|path2:1-3039(+)|transcript/3426          1.288037
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401       1.000000
## 9808 more rows ...
```

Obtain Counts Per Million to standardise count data comparison

```
myTPM_Sex <- dge_Sex$counts
# head(myTPM_Sex)
```

4.3.6 Initial Data Exploration

Which values in myCPM are greater than 0.5? This produces a logical matrix with TRUEs and FALSEs. TRUE values are samples with > 0.5 counts in that sample per million

```
thresh_Sex <- myTPM_Sex > 0.5
# This produces a logical matrix with TRUEs and FALSEs
head(thresh_Sex)
```

```
##                                                   G1_SepF G2_SepF G3_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304       TRUE    TRUE    TRUE
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564      TRUE    TRUE    TRUE
## PB.10.1|004815|path1:5-1713(+)|transcript/17534     FALSE    TRUE   FALSE
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       TRUE    TRUE    TRUE
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401  FALSE   FALSE   FALSE
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993       TRUE    TRUE    TRUE
##                                                   G4_MarM G5_SepM G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304       TRUE    TRUE    TRUE
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564      TRUE    TRUE    TRUE
## PB.10.1|004815|path1:5-1713(+)|transcript/17534      TRUE   FALSE   FALSE
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       TRUE    TRUE    TRUE
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401  FALSE   FALSE   FALSE
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993       TRUE    TRUE    TRUE
```

Raw Counts per million and library sizes will be consistent with the Season comparison above. There is no need to plot them again however scripts are included to create separate R objects for downstream application.

CPM distribution with logged CPM

```
# export the plot as png
png("plotDensities_Sex.png", width = 600)
unfilteredExpr_Sex <- cpm(dge_Sex, log = T)
plotDensities(unfilteredExpr_Sex, col = myPalette, legend = TRUE)
```

Library sizes per sample

```
# export the plot as png
png("LibSize_Sex.png", width = 600)
# The names argument tells the barplot to use the sample
# names on the x-axis The last argument rotates the axis
# names
barplot(dge_Sex$samples$lib.size, names = colnames(dge_Sex),
    las = 2, col = myPalette, legend = TRUE)
```

Total estimated transcript counts per sample

```
# export the plot as png
png("boxplot_Sex.png", width = 600)
boxplot(dge_Sex$counts, col = myPalette, las = 2, legend = TRUE)
# title('Boxplot of transcript total estimated counts')
```

Logged total estimated transcript counts per sample

```
# Get log2 counts per million
logcounts_Sex <- cpm(dge_Sex, log = TRUE)
# Check distributions of samples using boxplots export the
# plot as png
png("boxplot-logcounts_Sex.png", width = 600)
boxplot(logcounts_Sex, xlab = "", ylab = "Log2 counts per million",
    las = 2, col = myPalette, legend = TRUE)
# Let's add a blue horizontal line that corresponds to the
# median logCPM
abline(h = median(logcounts_Sex), col = myPalette, legend = TRUE)
# title('Boxplots of log Counts Per Million')
```

4.3.7 Calculations

```
# Data with actual counts, not TPM = 'dge'
head(dge_Sex)
```

```
## An object of class "DGEList"
## $counts
##                                                       G1_SepF      G2_SepF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304        1.790000e+02 118.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564       8.410000e+02 746.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       3.682740e-07   1.363036
## PB.11.1|004815|path2:1-3039(+)|transcript/3426        2.639676e+01  28.388835
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 0.000000e+00   0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993        7.179928e+01  33.065458
##                                                        G3_MarF      G4_MarM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304         49.0000000 113.0000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564       645.0000000 997.0000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534         0.4894863   0.9870162
## PB.11.1|004815|path2:1-3039(+)|transcript/3426         10.4093586   4.4945206
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401     0.0000000   0.0000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993         10.3920012  12.2814560
##                                                        G5_SepM      G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304         77.00000   147.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564       415.00000  1577.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534         0.00000     0.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426         29.50226    12.42200
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401     0.00000     0.00000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993         10.39200    25.50764
##
```

```
## $samples
##          group lib.size norm.factors
## G1_SepF     1  5796184            1
## G2_SepF     1  4480604            1
## G3_MarF     1  3867272            1
## G4_MarM     2  4517543            1
## G5_SepM     2  3518940            1
## G8_MarF     1  4920453            1
##
## $genes
##                                                  Length EffectiveLength
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      1530         1197.281
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     2065         1732.281
## PB.10.1|004815|path1:5-1713(+)|transcript/17534     1612         1279.281
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      2999         2666.281
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401  1965         1632.281
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993      2714         2381.281
##                                                  Overdispersion
## PB.2.1|002537|path0:1-1624(+)|transcript/18304          1.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564         1.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534         3.253130
## PB.11.1|004815|path2:1-3039(+)|transcript/3426          1.288037
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401      1.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993          1.058506
```

```r
head(myTPM_Sex)
```

```
##                                                      G1_SepF      G2_SepF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304     1.790000e+02 118.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564    8.410000e+02 746.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534    3.682740e-07   1.363036
## PB.11.1|004815|path2:1-3039(+)|transcript/3426     2.639676e+01  28.388835
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 0.000000e+00   0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993     7.179928e+01  33.065458
##                                                      G3_MarF      G4_MarM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      49.0000000 113.0000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564    645.0000000 997.0000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534      0.4894863   0.9870162
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      10.4093586   4.4945206
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401   0.0000000   0.0000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993      10.3920012  12.2814560
##                                                      G5_SepM      G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      77.00000   147.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564    415.00000  1577.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534      0.00000     0.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      29.50226    12.42200
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401   0.00000     0.00000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993      10.39200    25.50764
```

Filter out genes with less than 5 reads in 2 samples for each transcript

```
filter_Sex <- apply(dge_Sex, 1, function(x) length(x[x > 5]) >=
    2)
filtered_Sex <- dge_Sex[filter_Sex, ]
head(filtered_Sex)
```

```
## An object of class "DGEList"
## $counts
##                                                    G1_SepF    G2_SepF    G3_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304  179.00000 118.00000   49.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564 841.00000 746.00000  645.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426   26.39676  28.38883   10.40936
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993   71.79928  33.06546   10.39200
## PB.14.1|006283|path0:1-1955(+)|transcript/13821 708.75635 442.39668  326.26755
## PB.15.1|006e23|path0:1-3614(+)|transcript/1774  624.00000 527.00000  351.00000
##                                                    G4_MarM    G5_SepM     G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304  113.000000   77.00000   147.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564 997.000000  415.00000  1577.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426    4.494521   29.50226    12.42200
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993   12.281456   10.39200    25.50764
## PB.14.1|006283|path0:1-1955(+)|transcript/13821 508.756184  306.91270   552.99585
## PB.15.1|006e23|path0:1-3614(+)|transcript/1774  336.000000  281.00000   719.00000
##
## $samples
##         group lib.size norm.factors
## G1_SepF     1  5796184            1
## G2_SepF     1  4480604            1
## G3_MarF     1  3867272            1
## G4_MarM     2  4517543            1
## G5_SepM     2  3518940            1
## G8_MarF     1  4920453            1
##
## $genes
##                                                 Length EffectiveLength
## PB.2.1|002537|path0:1-1624(+)|transcript/18304    1530        1197.281
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564   2065        1732.281
## PB.11.1|004815|path2:1-3039(+)|transcript/3426    2999        2666.281
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993    2714        2381.281
## PB.14.1|006283|path0:1-1955(+)|transcript/13821   1920        1587.281
## PB.15.1|006e23|path0:1-3614(+)|transcript/1774    3583        3250.281
##                                                 Overdispersion
## PB.2.1|002537|path0:1-1624(+)|transcript/18304        1.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564       1.000000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426        1.288037
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993        1.058506
## PB.14.1|006283|path0:1-1955(+)|transcript/13821       1.084999
## PB.15.1|006e23|path0:1-3614(+)|transcript/1774        1.000000
```

Run calculations

```
dge_Sex <- calcNormFactors(dge_Sex)
dge_Sex <- estimateCommonDisp(dge_Sex)
dge_Sex <- estimateTagwiseDisp(dge_Sex)
```

4.3.8 Differentially expressed genes:

View the top ten 'differentially expressed genes'

```
results_Sex <- exactTest(dge_Sex)
topTags(results_Sex)
```

```
## Comparison of groups:  2-1
##                                                   Length EffectiveLength
## PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195   1136        803.2807
## PB.2770.1|6751e7|path1:1-1210(+)|transcript/22021   1103        770.2807
## PB.602.1|167374|path2:1-1363(+)|transcript/21155    1304        971.2807
## PB.725.1|1b70a4|path2:1-1745(+)|transcript/16546    1705       1372.2807
## PB.979.1|25edbb|path0:1-3100(+)|transcript/3575     3066       2733.2807
## PB.5826.1|dd0edc|path1:1-2280(+)|transcript/9146    2280       1947.2807
## PB.637.1|1826ff|path3:1-1155(+)|transcript/22860     905        572.3308
## PB.4152.1|9a8a98|path2:1-992(+)|transcript/22901     896        563.3308
## PB.2493.6|5d0eeb|path5:27-5305(+)|transcript/20823  1337       1004.2807
## PB.2180.1|51f17e|path0:1-673(+)|transcript/24169     469        146.6201
##                                                   Overdispersion     logFC
## PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195       1.229384  8.651305
## PB.2770.1|6751e7|path1:1-1210(+)|transcript/22021       2.615079  9.331293
## PB.602.1|167374|path2:1-1363(+)|transcript/21155        1.000000 10.079098
## PB.725.1|1b70a4|path2:1-1745(+)|transcript/16546        1.000000  6.074782
## PB.979.1|25edbb|path0:1-3100(+)|transcript/3575         1.000000  9.038182
## PB.5826.1|dd0edc|path1:1-2280(+)|transcript/9146        1.000000  9.027936
## PB.637.1|1826ff|path3:1-1155(+)|transcript/22860        1.039333  8.125067
## PB.4152.1|9a8a98|path2:1-992(+)|transcript/22901        1.539932  8.299550
## PB.2493.6|5d0eeb|path5:27-5305(+)|transcript/20823      6.372376 10.820135
## PB.2180.1|51f17e|path0:1-673(+)|transcript/24169        1.076723  6.617776
##                                                      logCPM       PValue
## PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195  9.417486 8.355261e-45
## PB.2770.1|6751e7|path1:1-1210(+)|transcript/22021  5.478547 1.527083e-39
## PB.602.1|167374|path2:1-1363(+)|transcript/21155   4.803360 1.529268e-34
## PB.725.1|1b70a4|path2:1-1745(+)|transcript/16546   6.697083 6.072447e-32
## PB.979.1|25edbb|path0:1-3100(+)|transcript/3575    5.015148 1.451210e-26
## PB.5826.1|dd0edc|path1:1-2280(+)|transcript/9146   3.754487 3.194292e-22
## PB.637.1|1826ff|path3:1-1155(+)|transcript/22860   7.398882 4.491435e-22
## PB.4152.1|9a8a98|path2:1-992(+)|transcript/22901   7.068196 2.267222e-19
## PB.2493.6|5d0eeb|path5:27-5305(+)|transcript/20823 7.443993 3.231998e-17
## PB.2180.1|51f17e|path0:1-673(+)|transcript/24169   6.798846 8.075235e-17
##                                                             FDR
## PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195  8.199017e-41
## PB.2770.1|6751e7|path1:1-1210(+)|transcript/22021  7.492632e-36
## PB.602.1|167374|path2:1-1363(+)|transcript/21155   5.002236e-31
## PB.725.1|1b70a4|path2:1-1745(+)|transcript/16546   1.489723e-28
## PB.979.1|25edbb|path0:1-3100(+)|transcript/3575    2.848144e-23
## PB.5826.1|dd0edc|path1:1-2280(+)|transcript/9146   5.224265e-19
## PB.637.1|1826ff|path3:1-1155(+)|transcript/22860   6.296350e-19
## PB.4152.1|9a8a98|path2:1-992(+)|transcript/22901   2.781031e-16
## PB.2493.6|5d0eeb|path5:27-5305(+)|transcript/20823 3.523955e-14
## PB.2180.1|51f17e|path0:1-673(+)|transcript/24169   7.924228e-14
```

Output the top 25 table of values based on calculated PValue

```
tab_Sex <- topTags(results_Sex, n = 25, sort.by = "PValue")
head(tab_Sex)
```

```
## Comparison of groups:  2-1
##                                                    Length EffectiveLength
## PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195   1136        803.2807
## PB.2770.1|6751e7|path1:1-1210(+)|transcript/22021   1103        770.2807
## PB.602.1|167374|path2:1-1363(+)|transcript/21155    1304        971.2807
## PB.725.1|1b70a4|path2:1-1745(+)|transcript/16546    1705       1372.2807
## PB.979.1|25edbb|path0:1-3100(+)|transcript/3575     3066       2733.2807
## PB.5826.1|dd0edc|path1:1-2280(+)|transcript/9146    2280       1947.2807
##                                                    Overdispersion      logFC
## PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195        1.229384   8.651305
## PB.2770.1|6751e7|path1:1-1210(+)|transcript/22021        2.615079   9.331293
## PB.602.1|167374|path2:1-1363(+)|transcript/21155         1.000000  10.079098
## PB.725.1|1b70a4|path2:1-1745(+)|transcript/16546         1.000000   6.074782
## PB.979.1|25edbb|path0:1-3100(+)|transcript/3575          1.000000   9.038182
## PB.5826.1|dd0edc|path1:1-2280(+)|transcript/9146         1.000000   9.027936
##                                                       logCPM       PValue
## PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195   9.417486 8.355261e-45
## PB.2770.1|6751e7|path1:1-1210(+)|transcript/22021   5.478547 1.527083e-39
## PB.602.1|167374|path2:1-1363(+)|transcript/21155    4.803360 1.529268e-34
## PB.725.1|1b70a4|path2:1-1745(+)|transcript/16546    6.697083 6.072447e-32
## PB.979.1|25edbb|path0:1-3100(+)|transcript/3575     5.015148 1.451210e-26
## PB.5826.1|dd0edc|path1:1-2280(+)|transcript/9146    3.754487 3.194292e-22
##                                                            FDR
## PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195   8.199017e-41
## PB.2770.1|6751e7|path1:1-1210(+)|transcript/22021   7.492632e-36
## PB.602.1|167374|path2:1-1363(+)|transcript/21155    5.002236e-31
## PB.725.1|1b70a4|path2:1-1745(+)|transcript/16546    1.489723e-28
## PB.979.1|25edbb|path0:1-3100(+)|transcript/3575     2.848144e-23
## PB.5826.1|dd0edc|path1:1-2280(+)|transcript/9146    5.224265e-19
```

```
write.table(tab_Sex, file = "Kgenelist-Top25_Sex.txt")
```

Output the entire table of values

```
tab_Sex_all <- topTags(results_Sex, n = Inf, sort.by = "PValue")
write.table(tab_Sex_all, file = "Kgenelist_Sex.txt")
```

Output summary expression data

```
dim(results_Sex)
```

```
## [1] 9813    3
```

```
summary(de_Sex <- decideTests(results_Sex))
```

```
##           2-1
## Down       25
## NotSig 9467
## Up        321
```

```
summary_Sex <- (de_Sex <- decideTests(results_Sex))

write.table(summary_Sex, file = "summary_Sex.txt")

detags_Sex <- rownames(dge_Sex)[as.logical(de_Sex)]
head(detags_Sex)
```

```
## [1] "PB.48.2|01e359|path4:4-3376(+)|transcript/2493"
## [2] "PB.147.1|0620ba|path0:1-1756(+)|transcript/23387"
## [3] "PB.147.3|0620ba|path0:1-1628(+)|transcript/24333"
## [4] "PB.193.1|07b09f|path0:1-2724(+)|transcript/21409"
## [5] "PB.193.2|07b09f|path0:3-2528(+)|transcript/11304"
## [6] "PB.193.7|07b09f|path0:5-2523(+)|transcript/22910"
```

```
head(results_Sex)
```

```
## An object of class "DGEExact"
## $table
##                                                      logFC      logCPM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304    0.1061405   4.6125270
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564  -0.1232896   7.5877828
## PB.10.1|004815|path1:5-1713(+)|transcript/17534   0.3743314  -0.9013964
## PB.11.1|004815|path2:1-3039(+)|transcript/3426    0.4425501   2.1973071
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401  0.0000000 -1.2129277
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993   -1.0490470   2.5164463
##                                                       PValue
## PB.2.1|002537|path0:1-1624(+)|transcript/18304    0.8132187
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564   0.8622934
## PB.10.1|004815|path1:5-1713(+)|transcript/17534   1.0000000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426    0.5699489
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401  1.0000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993    0.2018282
##
## $comparison
## [1] "1" "2"
##
## $genes
##                                                     Length EffectiveLength
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      1530        1197.281
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     2065        1732.281
## PB.10.1|004815|path1:5-1713(+)|transcript/17534     1612        1279.281
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      2999        2666.281
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401  1965        1632.281
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993      2714        2381.281
##                                                     Overdispersion
## PB.2.1|002537|path0:1-1624(+)|transcript/18304        1.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564       1.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       3.253130
## PB.11.1|004815|path2:1-3039(+)|transcript/3426        1.288037
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    1.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993        1.058506
```

Filter and sort this summary to only include upregulated and downregulated transcripts

```
summary_Sex.sort <- read.csv("./summary_Sex.txt", sep = " ",
    row.names = NULL, header = TRUE, stringsAsFactors = FALSE)
summary_Sex.sort <- summary_Sex.sort %>%
    rename(TranscriptID_3 = row.names, Regulation = X2.1)  #the name 'TranscriptID_3' is
→   used for consistency and joining as seen in the next step
```

fiter out the transcripts which are significantly upregulated when compared by sex

```
summary_Sex.sort_UP <- filter(summary_Sex.sort, Regulation ==
    "1")  #upregulated transcripts denoted as +1
head(summary_Sex.sort_UP)
```

```
##                                   TranscriptID_3 Regulation
## 1   PB.48.2|01e359|path4:4-3376(+)|transcript/2493          1
## 2 PB.147.1|0620ba|path0:1-1756(+)|transcript/23387          1
## 3 PB.147.3|0620ba|path0:1-1628(+)|transcript/24333          1
## 4 PB.193.1|07b09f|path0:1-2724(+)|transcript/21409          1
## 5 PB.193.2|07b09f|path0:3-2528(+)|transcript/11304          1
## 6 PB.193.7|07b09f|path0:5-2523(+)|transcript/22910          1
```

```
nrow(summary_Sex.sort_UP)
```

```
## [1] 321
```

Extract the full information about the putative identiry of these transcripts from the Summary file created in Section 3.2.3 above

```
summary_Sex.sort_UP_info <- semi_join(BLASTx.gene.join_Database.Full.PBT.Summary,
    summary_Sex.sort_UP, by = "TranscriptID_3")
# head (summary_Sex.sort_UP_info)
nrow(summary_Sex.sort_UP_info)
```

```
## [1] 183
```

```
write.csv(summary_Sex.sort_UP_info, "~/@Uni/@Flinders University
→   PhD/RWorkingDir/Workflow-readthedown/ReferenceClusters/summary_Sex.sort_UP_info.csv",
    quote = FALSE, row.names = FALSE, col.names = TRUE)  #write the file to a .csv
```

(note: although there are 157 transcripts significantly upregulated, only 93 of these returned a BLASTx result for their corresponding putative coding region)

Do the same for the Seasonally deownregulated transcripts

```
summary_Sex.sort_DOWN <- filter(summary_Sex.sort, Regulation ==
    "-1")  #downregulated transcripts denoted as -1
head(summary_Sex.sort_DOWN)
```

```
##                                TranscriptID_3 Regulation
## 1  PB.687.3|1a3601|path0:17-1696(+)|transcript/17363         -1
## 2    PB.953.1|24d6a9|path0:1-2789(+)|transcript/5211         -1
## 3   PB.1210.1|2ea8f7|path0:1-2553(+)|transcript/7797         -1
## 4 PB.1762.2|4279e1|path0:18-2114(+)|transcript/11272         -1
## 5   PB.1866.1|46c41d|path5:1-2443(+)|transcript/8811         -1
## 6 PB.1924.6|491418|path6:46-1849(+)|transcript/18901         -1
```

```r
nrow(summary_Sex.sort_DOWN)
```

```
## [1] 25
```

Extract the full information about the putative identiry of these transcripts from the Summary file created
in Section 3.2.3 above

```r
summary_Sex.sort_DOWN_info <- semi_join(BLASTx.gene.join_Database.Full.PBT.Summary,
    summary_Sex.sort_DOWN, by = "TranscriptID_3")
# head (summary_Sex.sort_DOWN_info)
nrow(summary_Sex.sort_DOWN_info)
```

```
## [1] 21
```

```r
write.csv(summary_Sex.sort_DOWN_info, "~/@Uni/@Flinders University
↪  PhD/RWorkingDir/Workflow-readthedown/ReferenceClusters/summary_Sex.sort_DOWN_info.csv",
    quote = FALSE, row.names = FALSE, col.names = TRUE)  #write the file to a .csv
```

(note: although there are 25 transcripts significantly downregulated, only 20 of these returned a BLASTx
result for their corresponding putative coding region)

Visualise the logged Fold Change against Average logged counts per million, in expression of transcripts between sex 1 (Female) and 2 (Male) using plotsmear: red indicates PValue < 0.05

```
png("plotsmear_Sex.png", width = 600)
plotSmear(results_Sex, de.tags = detags_Sex)
```
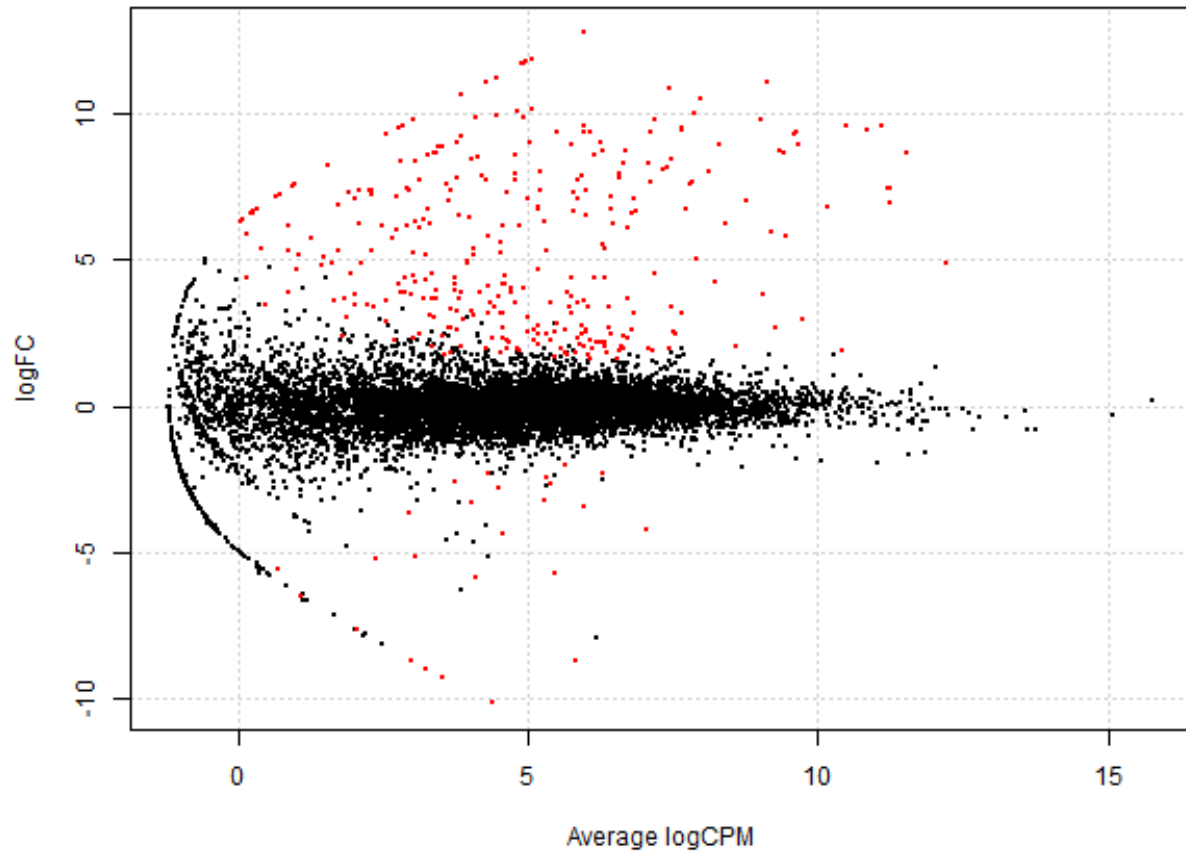

Figure 18: Log Fold Change vs average Log counts per million for expression level comparing Male to Female sex variables among six *T. adelaidensis* individuals. Red dots indicate transcripts with a significant change over season, with p-values <0.05.

Visualise differential expression of transcripts between sex 1 (Female) and 2 (Male) with negative Logged PValue against Logged Fold Change in expression: Red indicates a transcript with PValue < 0.05, and Log Fold change > 2

```
volcanoData_Sex <- cbind(results_Sex$table$logFC, -log10(results_Sex$table$PValue))
colnames(volcanoData_Sex) <- c("logFC", "negLogPval")
DEGs_Sex <- results_Sex$table$PValue < 0.05 & abs(results_Sex$table$logFC) >
    2

point.col_Sex <- ifelse(DEGs_Sex, "red", "black")
png("volcanoData_Sex.png", width = 600)
plot(volcanoData_Sex, pch = 16, col = point.col_Sex, cex = 0.5)
# pch is shape 1-26 are vectors, cex is size
```
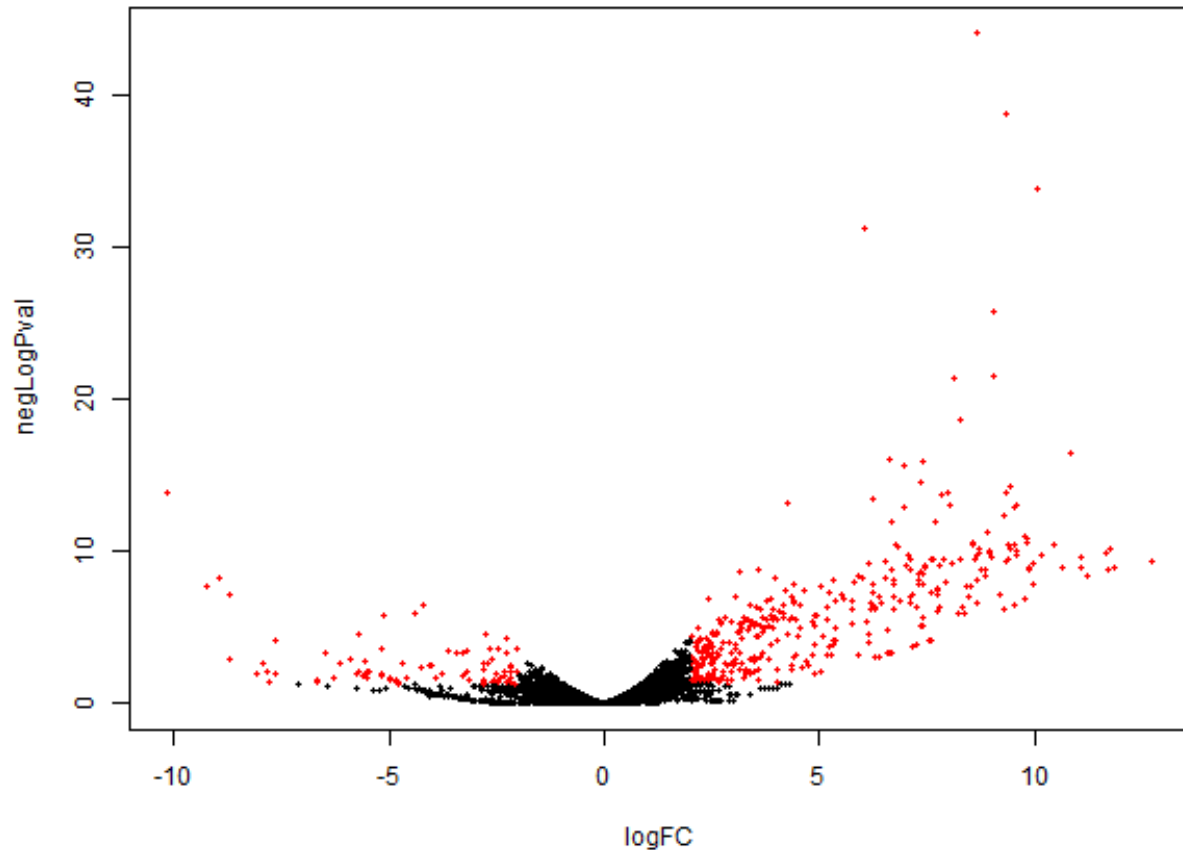

Figure 19: Negative Log of the p-value against Log Fold Change in expression level comparing Male to Female sex variables among six *T. adelaidensis* individuals. Red dots indicate transcripts with a significant change over season, with p-values <0.05 and *also* a log fold change >2.

4.4 Sequencing batch effects

Note G1, G2, G3, G4, G5 and G8 were all sequenced together on an Illumina Hiseq and library prep was done with the same kit. G6 and G7 were prepared in a separate batch using a different kit and sequenced on an Illumina Hiseq with fewer samples, and have a much higher read depth. These 2 represent a September Male, and March Female, respectively.

The removal of G6 and G7 from the dataset results in two groups of 3 per season, each with 2 females and 1 male.

4.5 Eight *T. adelaidensis* individuals collected between two seasonal periods.

Exploration of all Eight samples by each group factor:

Import the data:

```
# create paths
setwd("~/@Uni/@Flinders University PhD/RWorkingDir/R-kallisto-clstr/kallisto-clstr/")
DIR8 <- ("~/@Uni/@Flinders University PhD/RWorkingDir/R-kallisto-clstr/kallisto-clstr/")
head(DIR8)
```

```
## [1] "~/@Uni/@Flinders University PhD/RWorkingDir/R-kallisto-clstr/kallisto-clstr/"
```

```
paths8 <- list.dirs(path = DIR8, full.names = FALSE, recursive = FALSE)
# the working dir only includes kallisto output for the
# kidneys at this time View(paths)
head(paths8)
```

```
## [1] "G1_SepF" "G2_SepF" "G3_MarF" "G4_MarM" "G5_SepM" "G6_SepM"
```

```
Kcaught8 <- catchKallisto(paths8, verbose = TRUE)
```

```
## Reading G1_SepF, 9813 transcripts, 100 bootstraps
## Reading G2_SepF, 9813 transcripts, 100 bootstraps
## Reading G3_MarF, 9813 transcripts, 100 bootstraps
## Reading G4_MarM, 9813 transcripts, 100 bootstraps
## Reading G5_SepM, 9813 transcripts, 100 bootstraps
## Reading G6_SepM, 9813 transcripts, 100 bootstraps
## Reading G7_MarF, 9813 transcripts, 100 bootstraps
## Reading G8_MarF, 9813 transcripts, 100 bootstraps
```

```
setwd("~/@Uni/@Flinders University PhD/RWorkingDir/R-kallisto-clstr")
# View(Kcaught)
```

Set the group factors. Input Sample files are listed G1-G8 Group factors are applied to the DGElist object and are the point of comparison for expression analyses.

4.5.0.1 Season of collection: 1 = March, 2 = Sep. For the purposes of consistent labelling per group factors some April collections are referred to in the March group. Accurate collection information and dates are outlined in the methods chapter.

```
# these factors correlate to season of collection. 1 =
# March, 2 = Sep.
season8_group <- factor(c(2, 2, 1, 1, 2, 2, 1, 1))

# these factors correlate to individuals sex. 1 = Female, 2
# = Male
sex8_group <- factor(c(1, 1, 1, 2, 2, 2, 1, 1))
```

4.5.0.2 Individuals sex: 1 = Female, 2 = Male.

4.5.1 Group Factor: Season

Create the EdgeR DGE list

```
dge_Season8 <- DGEList(counts = Kcaught8$counts/Kcaught8$annotation$Overdispersion,
    genes = Kcaught8$annotation, group = season8_group)
# View(dge_Season8)
names(dge_Season8)
```

```
## [1] "counts"  "samples" "genes"
```

```
dge_Season8
```

```
## An object of class "DGEList"
## $counts
##                                                         G1_SepF      G2_SepF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304         1.790000e+02 118.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564        8.410000e+02 746.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534        4.158820e-07   1.539241
## PB.11.1|004815|path2:1-3039(+)|transcript/3426         2.860775e+01  30.766683
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401     0.000000e+00   0.000000
##                                                         G3_MarF      G4_MarM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304          49.0000000 113.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564        645.0000000 997.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534          0.5527639   1.114611
## PB.11.1|004815|path2:1-3039(+)|transcript/3426          11.2812462   4.870982
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401       0.0000000   0.000000
##                                                         G5_SepM      G6_SepM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304          77.00000   793.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564        415.00000  9814.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534          0.00000   136.63002
## PB.11.1|004815|path2:1-3039(+)|transcript/3426          31.97337  7657.54555
```

```
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    0.00000    47.04729
##                                                         G7_MarF    G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304         564.00000  147.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564      12448.00000 1577.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534        139.72439    0.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426        3505.65982   13.46247
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401      34.68487    0.00000
## 9808 more rows ...
##
## $samples
##           group lib.size norm.factors
## G1_SepF     2   5568916            1
## G2_SepF     2   4333716            1
## G3_MarF     1   3738942            1
## G4_MarM     1   4341981            1
## G5_SepM     2   3387520            1
## G6_SepM     2  41354721            1
## G7_MarF     1  43500763            1
## G8_MarF     1   4747833            1
##
## $genes
##                                                     Length EffectiveLength
## PB.2.1|002537|path0:1-1624(+)|transcript/18304        1530         1197.281
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564       2065         1732.281
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       1612         1279.281
## PB.11.1|004815|path2:1-3039(+)|transcript/3426        2999         2666.281
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    1965         1632.281
##                                                     Overdispersion
## PB.2.1|002537|path0:1-1624(+)|transcript/18304            1.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564           1.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534           2.880729
## PB.11.1|004815|path2:1-3039(+)|transcript/3426            1.188489
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401        3.028806
## 9808 more rows ...
```

Obtain Counts Per Million

```
myTPM_Season8 <- dge_Season8$counts
# head(myTPM_Season8)
```

4.5.2 Initial Data Exploration

```
# Which values in myCPM are greater than 0.5?
thresh_Season8 <- myTPM_Season8 > 0.5
# This produces a logical matrix with TRUEs and FALSEs
head(thresh_Season8)
```

```
##                                                   G1_SepF G2_SepF G3_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      TRUE    TRUE    TRUE
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     TRUE    TRUE    TRUE
## PB.10.1|004815|path1:5-1713(+)|transcript/17534     FALSE   TRUE    TRUE
```

```
## PB.11.1|004815|path2:1-3039(+)|transcript/3426             TRUE      TRUE      TRUE
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401        FALSE     FALSE     FALSE
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993             TRUE      TRUE      TRUE
##                                                          G4_MarM G5_SepM G6_SepM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304             TRUE      TRUE      TRUE
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564            TRUE      TRUE      TRUE
## PB.10.1|004815|path1:5-1713(+)|transcript/17534            TRUE     FALSE      TRUE
## PB.11.1|004815|path2:1-3039(+)|transcript/3426             TRUE      TRUE      TRUE
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401        FALSE     FALSE      TRUE
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993             TRUE      TRUE      TRUE
##                                                          G7_MarF G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304             TRUE      TRUE
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564            TRUE      TRUE
## PB.10.1|004815|path1:5-1713(+)|transcript/17534            TRUE     FALSE
## PB.11.1|004815|path2:1-3039(+)|transcript/3426             TRUE      TRUE
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401         TRUE     FALSE
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993             TRUE      TRUE
```

Insert colour pallate for following visualisations consistency

```r
myPalette8 <- c("#999999", "#F0E442", "#56B4E9", "#009E73", "#D55E00",
    "#0072B2", "#E69F00", "#CC79A7")
```

Plot CPM distribution with logged CPM

```
# export the plot as png
png("plotDensities_Season8.png", width = 600)
unfilteredExpr_Season8 <- cpm(dge_Season8, log = T)
plotDensities(unfilteredExpr_Season8, col = myPalette8, legend = TRUE)
```
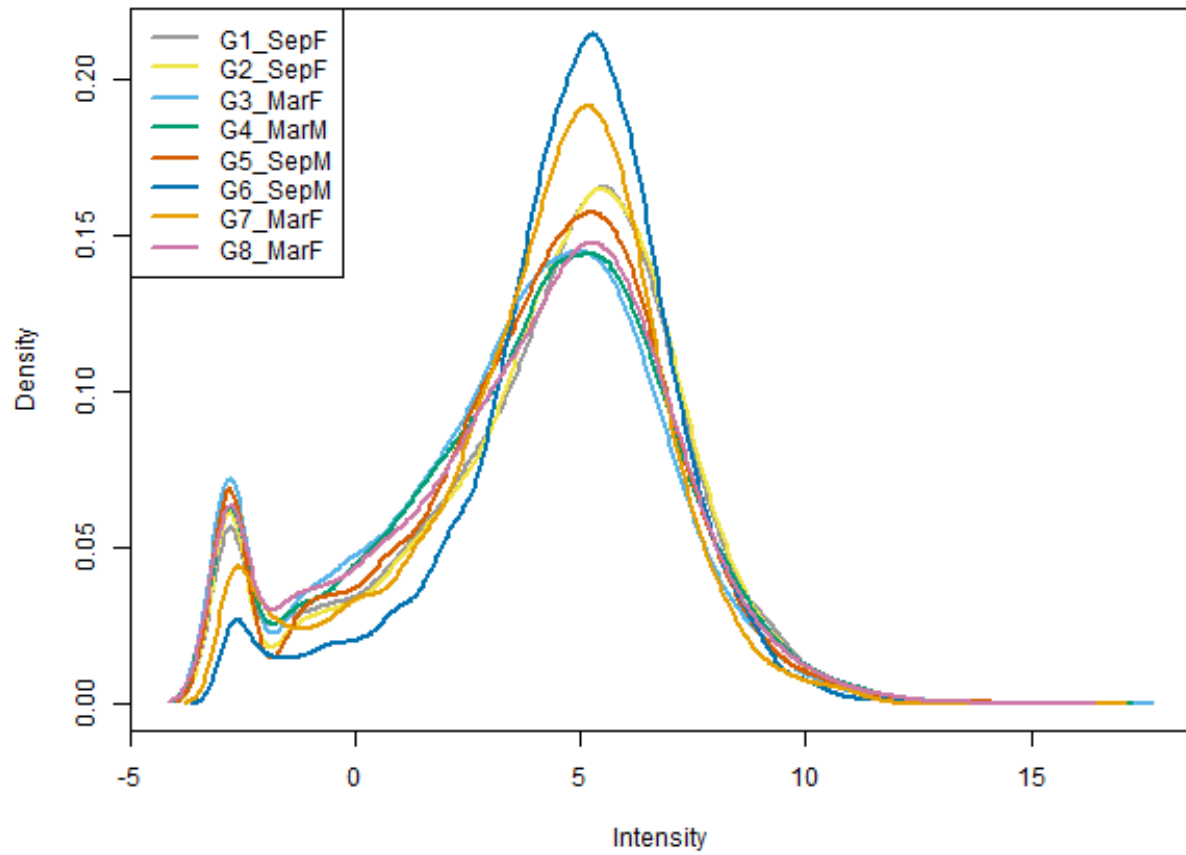

Figure 20: Log Counts per million Densities, for poly-a selected mRNA expression data from eight *T. adelaidensis* kidney samples

Visualise library sizes per sample

```r
# The names argument tells the barplot to use the sample
# names on the x-axis The las argument rotates the axis
# names export the plot as png
png("Barplot.LibSize_Season8.png", width = 600)
barplot(dge_Season8$samples$lib.size, names = colnames(dge_Season8),
    las = 2, col = myPalette8)
# title('Barplot of library sizes')
```
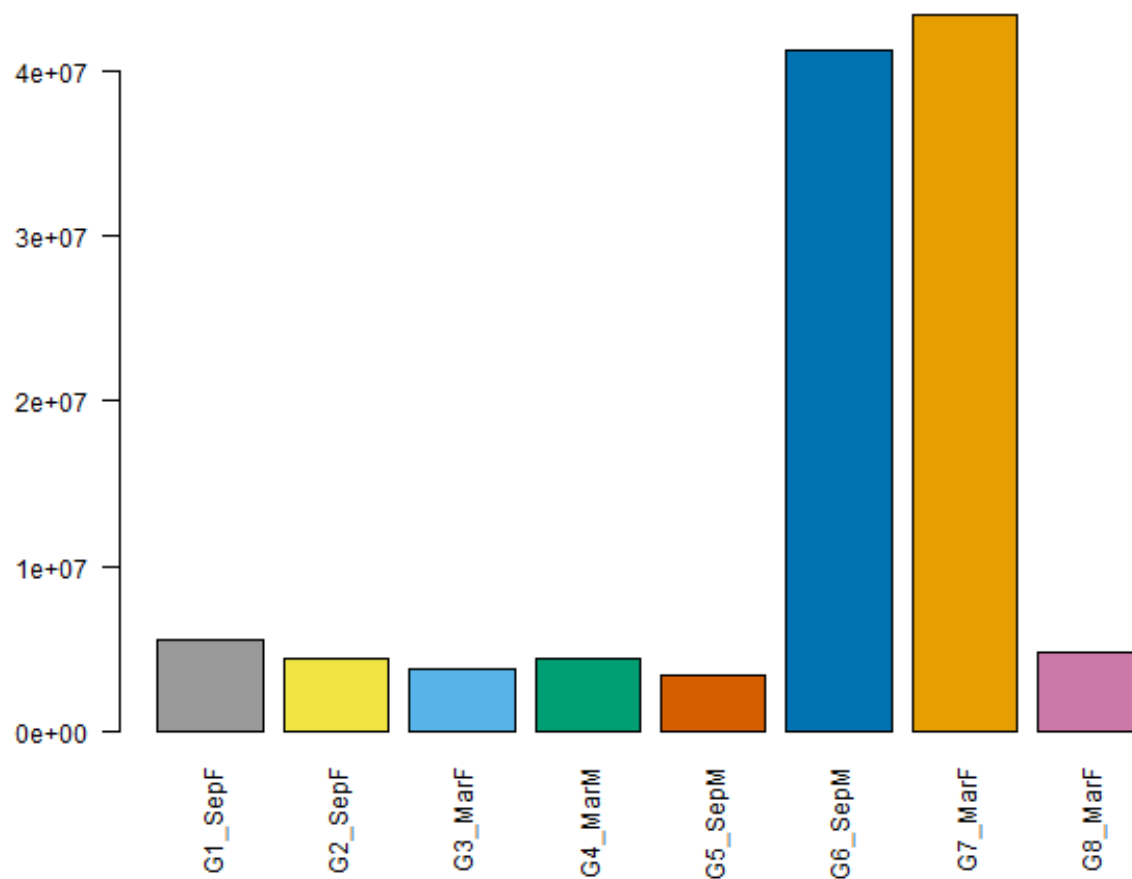

Figure 21: Library size of poly-a selected mRNA sequenced from kidney tissue of eight *T. adelaidensis* individuals

Visualise transcript total estimated counts per sample

```
# export the plot as png
png("Boxplot.Counts_Season8.png", width = 600)
boxplot(dge_Season8$counts, col = myPalette8, las = 2, legend = TRUE)
# title('Boxplot of total unfiltered Counts')
```
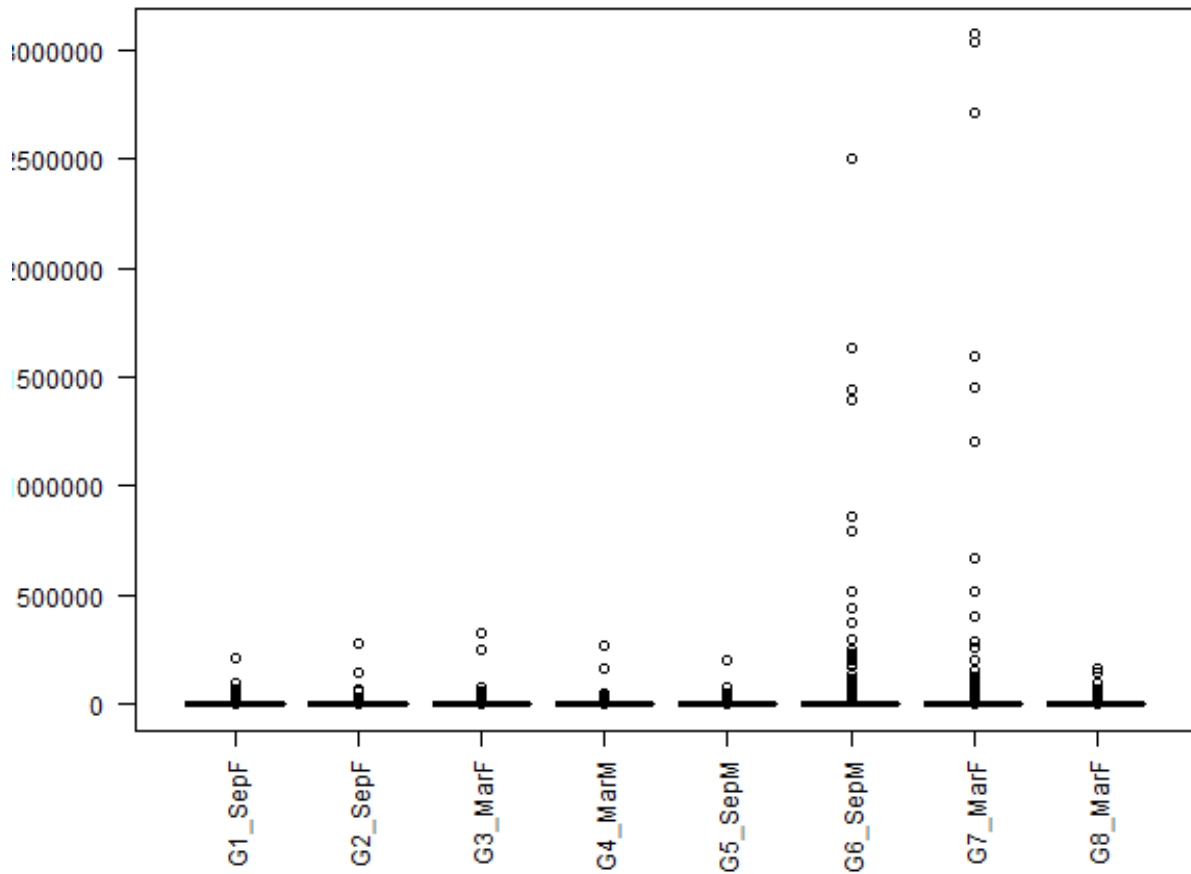

Figure 22: Total estimated counts, for poly-a selected mRNA expression data from eight *T. adelaidensis* kidney samples

```
# Get log2 counts per million
logcounts_Season8 <- cpm(dge_Season8, log = TRUE)
# Check distributions of samples using boxplots export the
# plot as png
png("boxplot.CPM_Season8.png", width = 600)
boxplot(logcounts_Season8, xlab = "", ylab = "Log2 counts per million",
    las = 2, col = myPalette8, legend = TRUE)
# Let's add a blue horizontal line that corresponds to the
# median logCPM
abline(h = median(logcounts_Season8), col = myPalette8, legend = TRUE)
# title('Boxplots of log Counts Per Million')
```
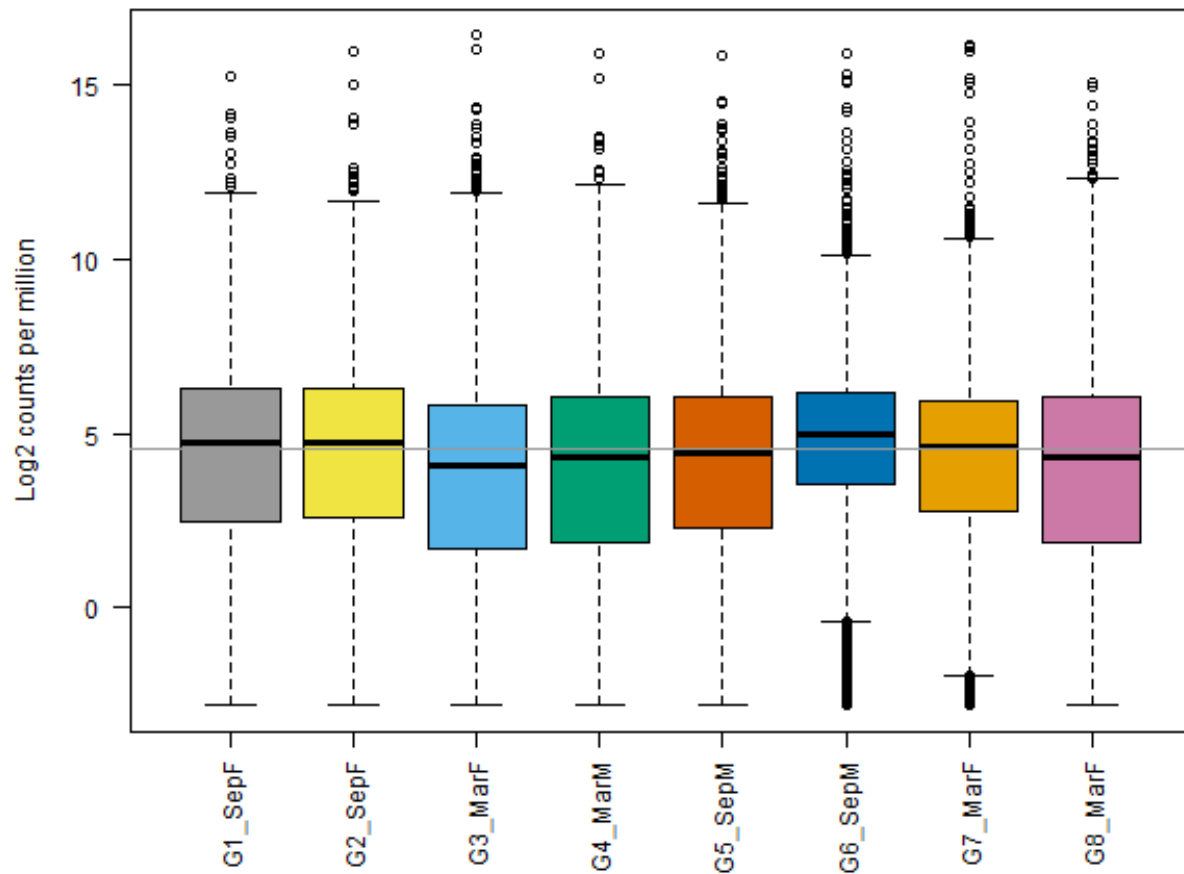

Figure 23: Log2 of estimated counts per million, for poly-a selected mRNA expression data from eight *T. adelaidensis* kidney samples

Visualise sample variation using an MDS plot

```
# export the plot as png
png("MDS_Season8.png", width = 600)
plotMDS(dge_Season8, col = myPalette8)
# title('MDS plot')
```
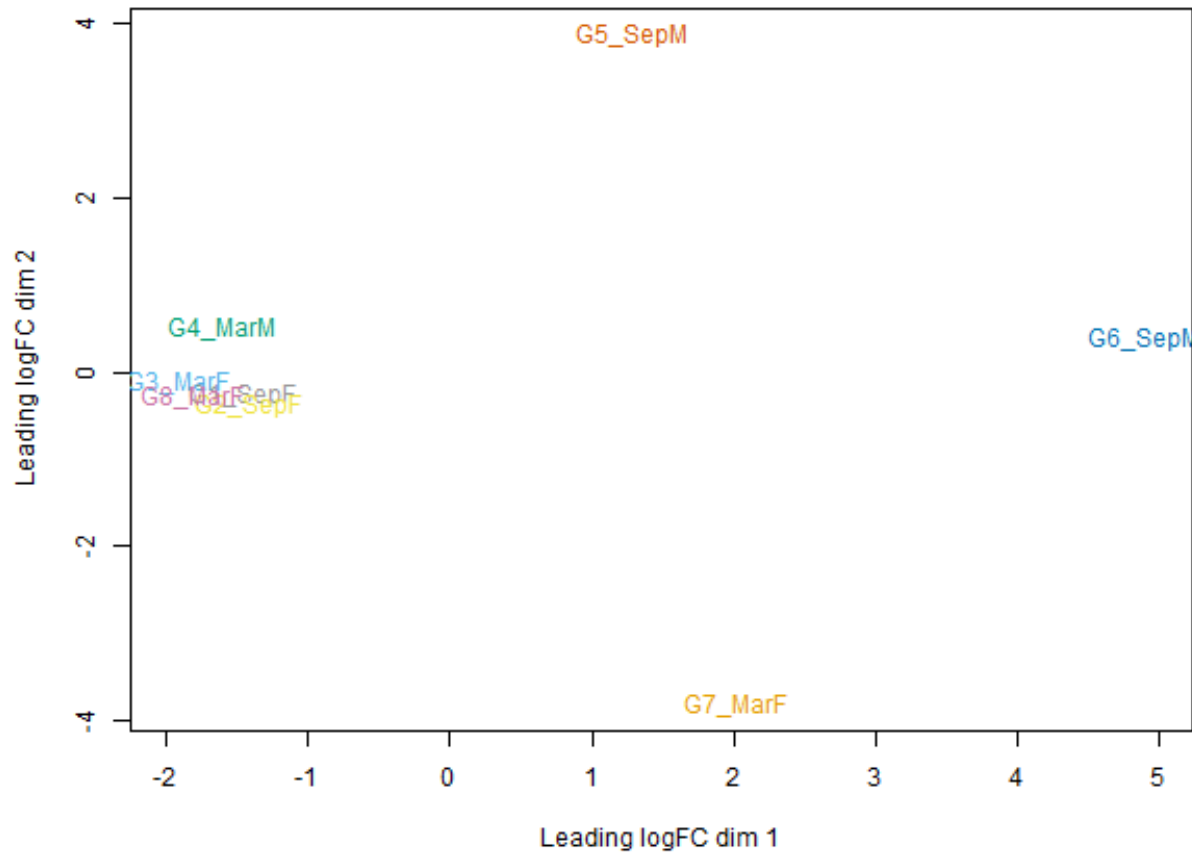

Figure 24: MDS plot of variation among eight *T. adelaidensis* individuals

4.5.3 Calculations

```
# Data with actual counts, not TPM = 'dge'
head(dge_Season8)
```

```
## An object of class "DGEList"
## $counts
##                                                        G1_SepF      G2_SepF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      1.790000e+02 118.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     8.410000e+02 746.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534     4.158820e-07   1.539241
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      2.860775e+01  30.766683
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 0.000000e+00   0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993      7.600000e+01  35.000000
##                                                         G3_MarF      G4_MarM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304       49.0000000 113.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     645.0000000 997.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       0.5527639   1.114611
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       11.2812462   4.870982
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    0.0000000   0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993       11.0000000  13.000000
##                                                        G5_SepM      G6_SepM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304       77.00000   793.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     415.00000  9814.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       0.00000   136.63002
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       31.97337  7657.54555
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    0.00000    47.04729
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993       11.00000  2959.00000
##                                                         G7_MarF      G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304        564.00000   147.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     12448.00000  1577.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       139.72439     0.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       3505.65982    13.46247
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401     34.68487     0.00000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993       4547.00000    27.00000
##
## $samples
##         group lib.size norm.factors
## G1_SepF     2  5568916            1
## G2_SepF     2  4333716            1
## G3_MarF     1  3738942            1
## G4_MarM     1  4341981            1
## G5_SepM     2  3387520            1
## G6_SepM     2 41354721            1
## G7_MarF     1 43500763            1
## G8_MarF     1  4747833            1
##
## $genes
##                                                   Length EffectiveLength
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      1530         1197.281
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     2065         1732.281
## PB.10.1|004815|path1:5-1713(+)|transcript/17534     1612         1279.281
```

```
## PB.11.1|004815|path2:1-3039(+)|transcript/3426            2999          2666.281
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401        1965          1632.281
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993            2714          2381.281
##                                                           Overdispersion
## PB.2.1|002537|path0:1-1624(+)|transcript/18304                 1.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564                1.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534               2.880729
## PB.11.1|004815|path2:1-3039(+)|transcript/3426               1.188489
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401           3.028806
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993               1.000000
```

```r
head(myTPM_Season8)
```

```
##                                                                G1_SepF      G2_SepF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304              1.790000e+02 118.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564            8.410000e+02 746.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534            4.158820e-07   1.539241
## PB.11.1|004815|path2:1-3039(+)|transcript/3426             2.860775e+01  30.766683
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401         0.000000e+00   0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993             7.600000e+01  35.000000
##                                                                G3_MarF      G4_MarM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304              49.0000000 113.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564           645.0000000 997.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534             0.5527639   1.114611
## PB.11.1|004815|path2:1-3039(+)|transcript/3426             11.2812462   4.870982
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401          0.0000000   0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993             11.0000000  13.000000
##                                                                G5_SepM      G6_SepM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304              77.00000   793.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564           415.00000  9814.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534             0.00000   136.63002
## PB.11.1|004815|path2:1-3039(+)|transcript/3426             31.97337  7657.54555
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401          0.00000    47.04729
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993             11.00000  2959.00000
##                                                                G7_MarF      G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304             564.00000   147.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564          12448.00000  1577.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534           139.72439     0.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426           3505.65982    13.46247
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401         34.68487     0.00000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993           4547.00000    27.00000
```

filter out genes with less than 5 reads in 2 samples for each transcript

```r
filter_Season8 <- apply(dge_Season8, 1, function(x) length(x[x >
    5]) >= 2)
filtered_Season8 <- dge_Season8[filter_Season8, ]
head(filtered_Season8)
```

```
## An object of class "DGEList"
## $counts
##                                                               G1_SepF      G2_SepF
```

```
## PB.2.1|002537|path0:1-1624(+)|transcript/18304        1.790000e+02 118.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564        8.410000e+02 746.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534        4.158820e-07   1.539241
## PB.11.1|004815|path2:1-3039(+)|transcript/3426         2.860775e+01  30.766683
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 0.000000e+00   0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993         7.600000e+01  35.000000
##                                                         G3_MarF     G4_MarM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304         49.0000000 113.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564       645.0000000 997.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534         0.5527639   1.114611
## PB.11.1|004815|path2:1-3039(+)|transcript/3426         11.2812462   4.870982
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401     0.0000000   0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993         11.0000000  13.000000
##                                                         G5_SepM     G6_SepM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304         77.00000   793.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564       415.00000  9814.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534         0.00000   136.63002
## PB.11.1|004815|path2:1-3039(+)|transcript/3426         31.97337  7657.54555
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401     0.00000    47.04729
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993         11.00000  2959.00000
##                                                         G7_MarF     G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304        564.00000   147.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     12448.00000  1577.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       139.72439     0.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       3505.65982    13.46247
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    34.68487     0.00000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993       4547.00000    27.00000
##
## $samples
##          group lib.size norm.factors
## G1_SepF     2  5568916            1
## G2_SepF     2  4333716            1
## G3_MarF     1  3738942            1
## G4_MarM     1  4341981            1
## G5_SepM     2  3387520            1
## G6_SepM     2 41354721            1
## G7_MarF     1 43500763            1
## G8_MarF     1  4747833            1
##
## $genes
##                                                         Length EffectiveLength
## PB.2.1|002537|path0:1-1624(+)|transcript/18304           1530         1197.281
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564          2065         1732.281
## PB.10.1|004815|path1:5-1713(+)|transcript/17534          1612         1279.281
## PB.11.1|004815|path2:1-3039(+)|transcript/3426           2999         2666.281
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401       1965         1632.281
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993           2714         2381.281
##                                                         Overdispersion
## PB.2.1|002537|path0:1-1624(+)|transcript/18304               1.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564              1.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534              2.880729
## PB.11.1|004815|path2:1-3039(+)|transcript/3426               1.188489
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401           3.028806
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993               1.000000
```
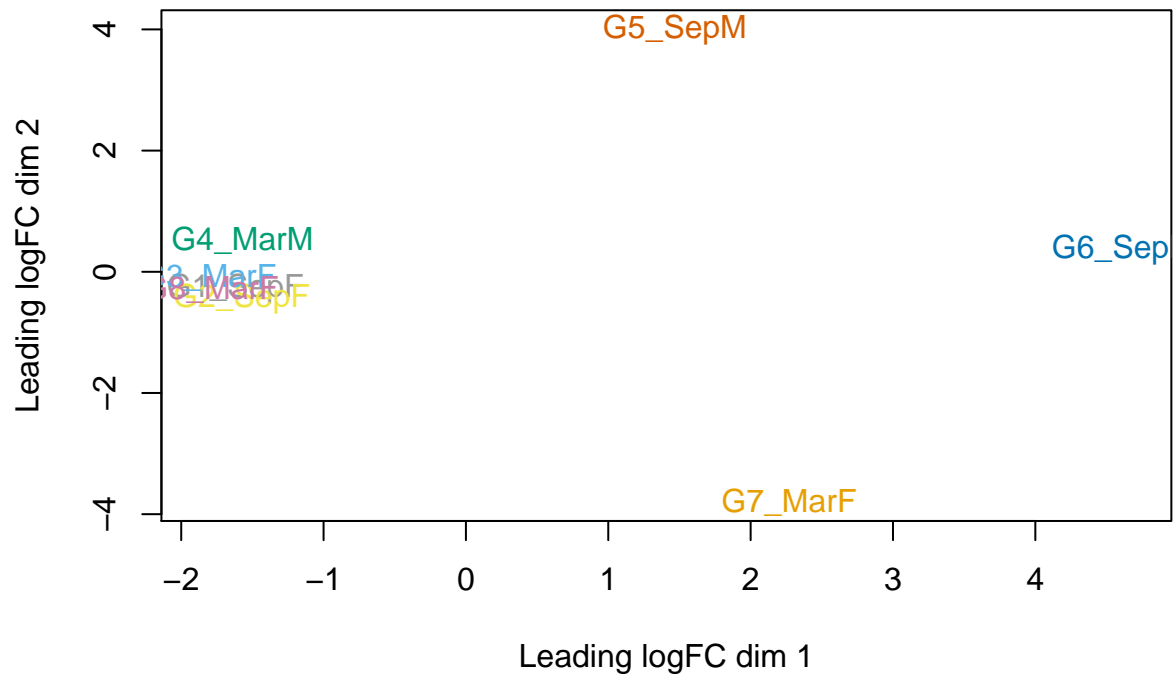
Run calculations

```
dge_Season8 <- calcNormFactors(dge_Season8)
dge_Season8 <- estimateCommonDisp(dge_Season8)
dge_Season8 <- estimateTagwiseDisp(dge_Season8)
```

Note that filtering did not change the appearance of the MDS plot

```
plotMDS(dge_Season8, col = myPalette8)
```


4.5.4 Differentially expressed genes:

View the top ten 'differentially expressed genes'

```
results_Season8 <- exactTest(dge_Season8)
topTags(results_Season8)
```

```
## Comparison of groups:  2-1
##                                                      Length EffectiveLength
## PB.499.11|124cf9|path4:2108-4109(+)|transcript/7418    2000       1667.2807
## PB.4788.1|b34d56|path10:2-1859(+)|transcript/15227     1771       1438.2807
## PB.147.1|0620ba|path0:1-1756(+)|transcript/23387        701        368.6542
## PB.3353.1|7cf053|path5:1-7730(+)|transcript/10         7701       7368.2807
```

```
## PB.621.2|172e3f|path6:3-1224(+)|transcript/21583         1227       894.2807
## PB.6088.12|e88f16|path1:164-4329(+)|transcript/5912      2488      2155.2807
## PB.1449.1|36bd5c|path2:1-1532(+)|transcript/18562        1341      1008.2807
## PB.985.1|261589|path8:1-2545(+)|transcript/6665          2510      2177.2807
## PB.6065.1|e74bdb|path2:1-2471(+)|transcript/15776        1754      1421.2807
## PB.1854.1|466cde|path0:1-1088(+)|transcript/22568         908       575.3308
##                                                       Overdispersion       logFC
## PB.499.11|124cf9|path4:2108-4109(+)|transcript/7418        1.856008   3.444092
## PB.4788.1|b34d56|path10:2-1859(+)|transcript/15227         1.293367  11.758742
## PB.147.1|0620ba|path0:1-1756(+)|transcript/23387           1.055680  14.177786
## PB.3353.1|7cf053|path5:1-7730(+)|transcript/10             2.287311   7.525344
## PB.621.2|172e3f|path6:3-1224(+)|transcript/21583           1.000000  14.179990
## PB.6088.12|e88f16|path1:164-4329(+)|transcript/5912        1.577810   8.743147
## PB.1449.1|36bd5c|path2:1-1532(+)|transcript/18562          2.304372   7.882813
## PB.985.1|261589|path8:1-2545(+)|transcript/6665            1.000000  12.154915
## PB.6065.1|e74bdb|path2:1-2471(+)|transcript/15776          1.360280   9.388137
## PB.1854.1|466cde|path0:1-1088(+)|transcript/22568          1.000000   8.436555
##                                                        logCPM       PValue
## PB.499.11|124cf9|path4:2108-4109(+)|transcript/7418   5.702989 3.114653e-08
## PB.4788.1|b34d56|path10:2-1859(+)|transcript/15227    8.735038 1.079896e-07
## PB.147.1|0620ba|path0:1-1756(+)|transcript/23387      6.322527 2.504299e-07
## PB.3353.1|7cf053|path5:1-7730(+)|transcript/10        6.704730 6.235719e-07
## PB.621.2|172e3f|path6:3-1224(+)|transcript/21583      6.324077 8.295346e-07
## PB.6088.12|e88f16|path1:164-4329(+)|transcript/5912  10.781344 9.758992e-07
## PB.1449.1|36bd5c|path2:1-1532(+)|transcript/18562     6.902254 1.430707e-06
## PB.985.1|261589|path8:1-2545(+)|transcript/6665       6.383915 1.518649e-06
## PB.6065.1|e74bdb|path2:1-2471(+)|transcript/15776     5.982169 1.530897e-06
## PB.1854.1|466cde|path0:1-1088(+)|transcript/22568     6.356782 1.552334e-06
##                                                              FDR
## PB.499.11|124cf9|path4:2108-4109(+)|transcript/7418   0.0003056409
## PB.4788.1|b34d56|path10:2-1859(+)|transcript/15227    0.0005298509
## PB.147.1|0620ba|path0:1-1756(+)|transcript/23387      0.0008191561
## PB.3353.1|7cf053|path5:1-7730(+)|transcript/10        0.0011003949
## PB.621.2|172e3f|path6:3-1224(+)|transcript/21583      0.0011003949
## PB.6088.12|e88f16|path1:164-4329(+)|transcript/5912  0.0011003949
## PB.1449.1|36bd5c|path2:1-1532(+)|transcript/18562     0.0011003949
## PB.985.1|261589|path8:1-2545(+)|transcript/6665       0.0011003949
## PB.6065.1|e74bdb|path2:1-2471(+)|transcript/15776     0.0011003949
## PB.1854.1|466cde|path0:1-1088(+)|transcript/22568     0.0011003949
```

Output summary expression data

```
dim(results_Season8)
```

```
## [1] 9813    3
```

```
summary(de_Season8 <- decideTests(results_Season8))
```

```
##           2-1
## Down        6
## NotSig 9620
## Up        187
```

```r
Summary_Season8 <- (de_Season8 <- decideTests(results_Season8))

write.table(Summary_Season8, file = "summary_Season8.txt")

detags_Season8 <- rownames(dge_Season8)[as.logical(de_Season8)]
```

Visualise with a plotsmear

```
detags_Season8 <- rownames(dge_Season8)[as.logical(de_Season8)]
# export the plot as png
png("plotSmear_Season8.png", width = 600)
plotSmear(results_Season8, de.tags = detags_Season8)
```
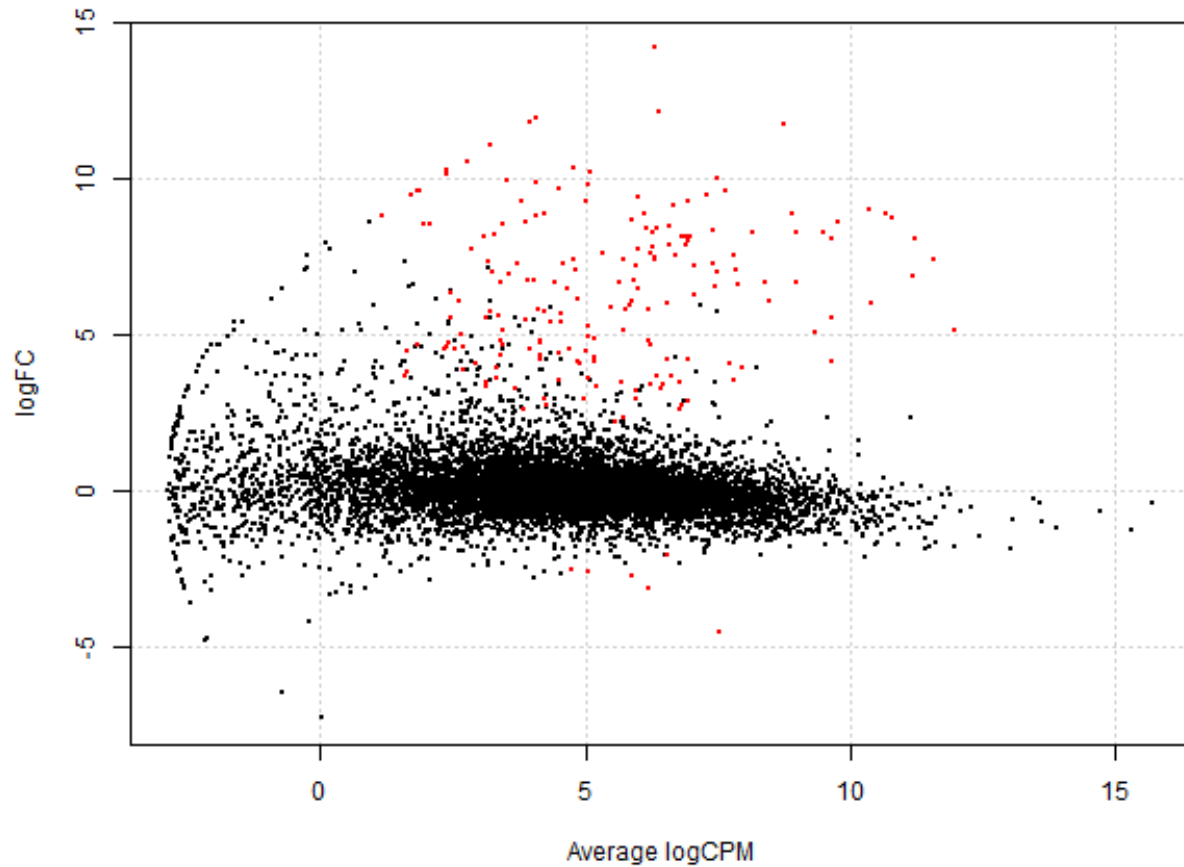

Figure 25: Log Fold Change vs average Log counts per million for expression level comparing Season 2 (September collection) to Season 1 (April/March collection) among eight *T. adelaidensis* individuals. Red dots indicate transcripts with a significant change over season, with p-values <0.05.

Visualise with a volcano plot

```r
# export the plot as png
png("volcanoData_Season8.png", width = 600)
volcanoData_Season8 <- cbind(results_Season8$table$logFC,
↪    -log10(results_Season8$table$PValue))
colnames(volcanoData_Season8) <- c("logFC", "negLogPval")
DEGs_Season8 <- results_Season8$table$PValue < 0.05 & abs(results_Season8$table$logFC) >
    2
point.col_Season8 <- ifelse(DEGs_Season8, "red", "black")
plot(volcanoData_Season8, pch = 16, col = point.col_Season8,
    cex = 0.5)
```
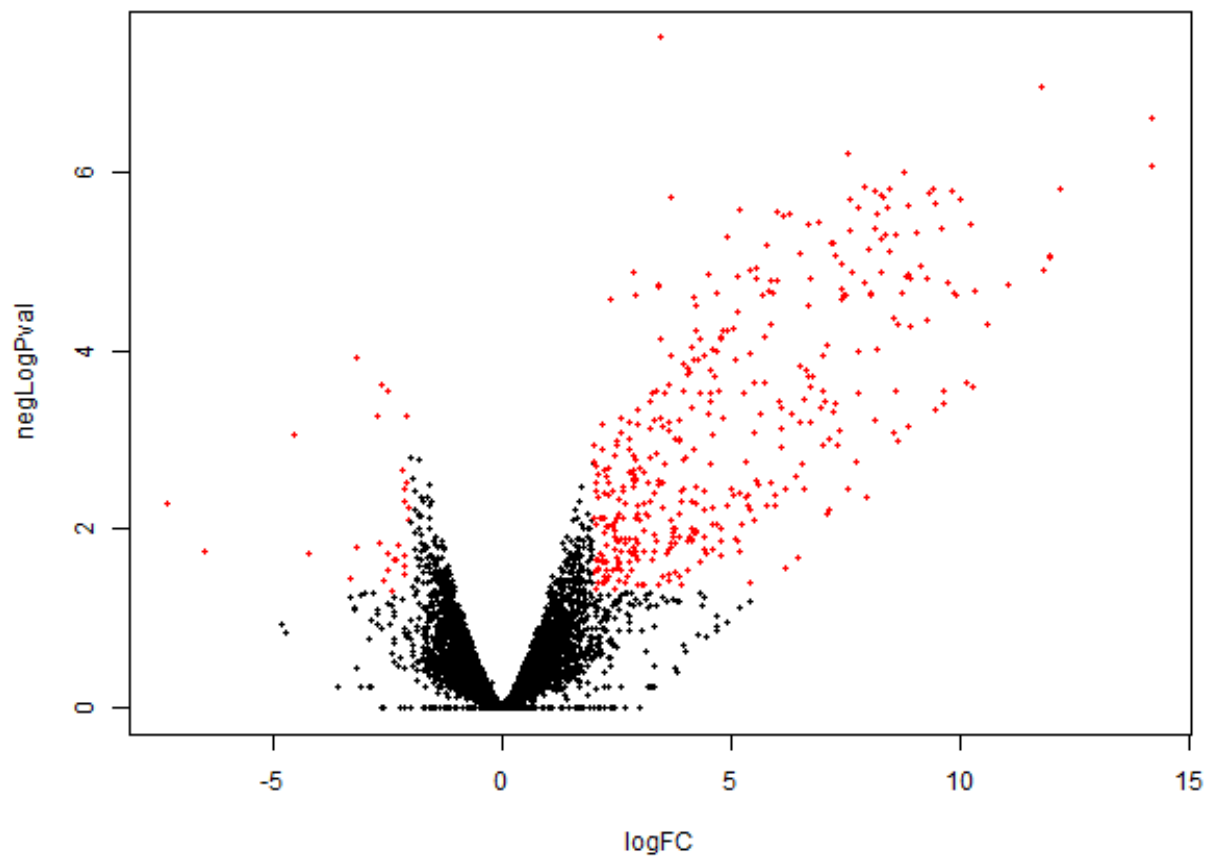

Figure 26: Negative Log of the p-value against Log Fold Change in expression level comparing Season 2 (September collection) to Season 1 (April/March collection) among eight *T. adelaidensis* individuals. Red dots indicate transcripts with a significant change over season, with p-values <0.05 and *also* a log fold change >2.

4.5.5 Group Factor: Sex

Create the EdgeR DGE list

```
dge_Sex8 <- DGEList(counts = Kcaught8$counts/Kcaught8$annotation$Overdispersion,
    genes = Kcaught8$annotation, group = sex8_group)
# View(dge_Sex8)
names(dge_Sex8)
```

```
## [1] "counts"  "samples" "genes"
```

```
dge_Sex8
```

```
## An object of class "DGEList"
## $counts
##                                                        G1_SepF      G2_SepF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304     1.790000e+02 118.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564    8.410000e+02 746.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534    4.158820e-07   1.539241
## PB.11.1|004815|path2:1-3039(+)|transcript/3426     2.860775e+01  30.766683
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 0.000000e+00   0.000000
##                                                        G3_MarF      G4_MarM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      49.0000000 113.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564    645.0000000 997.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534      0.5527639   1.114611
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      11.2812462   4.870982
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401   0.0000000   0.000000
##                                                        G5_SepM      G6_SepM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      77.00000   793.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564    415.00000  9814.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534      0.00000   136.63002
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      31.97337  7657.54555
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401   0.00000    47.04729
##                                                        G7_MarF      G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      564.00000   147.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564    12448.00000 1577.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534     139.72439     0.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426     3505.65982    13.46247
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401   34.68487     0.00000
## 9808 more rows ...
##
## $samples
##           group lib.size norm.factors
## G1_SepF       1  5568916            1
## G2_SepF       1  4333716            1
## G3_MarF       1  3738942            1
## G4_MarM       2  4341981            1
## G5_SepM       2  3387520            1
## G6_SepM       2 41354721            1
```

```
## G7_MarF     1 43500763            1
## G8_MarF     1  4747833            1
##
## $genes
##                                                     Length EffectiveLength
## PB.2.1|002537|path0:1-1624(+)|transcript/18304       1530         1197.281
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564      2065         1732.281
## PB.10.1|004815|path1:5-1713(+)|transcript/17534      1612         1279.281
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       2999         2666.281
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401   1965         1632.281
##                                                     Overdispersion
## PB.2.1|002537|path0:1-1624(+)|transcript/18304            1.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564           1.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534           2.880729
## PB.11.1|004815|path2:1-3039(+)|transcript/3426            1.188489
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401        3.028806
## 9808 more rows ...
```

```r
# Obtain Counts Per Million
myTPM_Sex8 <- dge_Sex8$counts
# Have a look at the output
head(myTPM_Sex8)
```

```
##                                                          G1_SepF       G2_SepF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304       1.790000e+02    118.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564      8.410000e+02    746.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534      4.158820e-07      1.539241
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       2.860775e+01     30.766683
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401   0.000000e+00      0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993       7.600000e+01     35.000000
##                                                          G3_MarF      G4_MarM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304        49.0000000   113.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564      645.0000000   997.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534        0.5527639     1.114611
## PB.11.1|004815|path2:1-3039(+)|transcript/3426        11.2812462     4.870982
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401     0.0000000     0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993        11.0000000    13.000000
##                                                          G5_SepM      G6_SepM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304          77.00000    793.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564        415.00000   9814.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534          0.00000    136.63002
## PB.11.1|004815|path2:1-3039(+)|transcript/3426          31.97337   7657.54555
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401       0.00000     47.04729
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993          11.00000   2959.00000
##                                                          G7_MarF      G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304         564.00000    147.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564      12448.00000   1577.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534        139.72439      0.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426        3505.65982     13.46247
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401      34.68487      0.00000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993        4547.00000     27.00000
```

4.5.6 Initial Data Exploration

```r
# Which values in myCPM are greater than 0.5?
thresh_Sex8 <- myTPM_Sex8 > 0.5
# This produces a logical matrix with TRUEs and FALSEs
head(thresh_Sex8)
```

```
##                                               G1_SepF G2_SepF G3_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304    TRUE    TRUE    TRUE
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564   TRUE    TRUE    TRUE
## PB.10.1|004815|path1:5-1713(+)|transcript/17534  FALSE    TRUE    TRUE
## PB.11.1|004815|path2:1-3039(+)|transcript/3426    TRUE    TRUE    TRUE
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 FALSE  FALSE   FALSE
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993    TRUE    TRUE    TRUE
##                                               G4_MarM G5_SepM G6_SepM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304    TRUE    TRUE    TRUE
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564   TRUE    TRUE    TRUE
## PB.10.1|004815|path1:5-1713(+)|transcript/17534   TRUE   FALSE    TRUE
## PB.11.1|004815|path2:1-3039(+)|transcript/3426    TRUE    TRUE    TRUE
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 FALSE  FALSE    TRUE
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993    TRUE    TRUE    TRUE
##                                               G7_MarF G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304    TRUE    TRUE
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564   TRUE    TRUE
## PB.10.1|004815|path1:5-1713(+)|transcript/17534   TRUE   FALSE
## PB.11.1|004815|path2:1-3039(+)|transcript/3426    TRUE    TRUE
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 TRUE   FALSE
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993    TRUE    TRUE
```

Raw Counts per million and library sizes will be consistent with the Season comparison above. There is no need to plot them again however scripts are included to create separate R objects for downstream application.

CPM distribution with logged CPM

```r
# export the plot as png
png("plotDensities_Sex8.png", width = 600)
unfilteredExpr_Sex8 <- cpm(dge_Sex8, log = T)
plotDensities(unfilteredExpr_Sex8, col = myPalette8, legend = TRUE)
```

Library sizes per sample

```r
# The names argument tells the barplot to use the sample
# names on the x-axis The las argument rotates the axis
# names export the plot as png
png("Barplot.LibSize_Sex8.png", width = 600)
barplot(dge_Sex8$samples$lib.size, names = colnames(dge_Sex8),
    las = 2, col = myPalette8)
# title('Barplot of library sizes')
```

Transcript total estimated counts per sample

```
# export the plot as png
png("Boxplot.Counts_Sex8.png", width = 600)
boxplot(dge_Sex8$counts, col = myPalette8, las = 2, legend = TRUE)
# title('Boxplot of total unfiltered Counts')
```

Logged total estimated transcript counts per sample

```
# Get log2 counts per million
logcounts_Sex8 <- cpm(dge_Sex8, log = TRUE)
# Check distributions of samples using boxplots export the
# plot as png
png("boxplot.CPM_Sex8.png", width = 600)
boxplot(logcounts_Sex8, xlab = "", ylab = "Log2 counts per million",
    las = 2, col = myPalette8, legend = TRUE)
# Let's add a blue horizontal line that corresponds to the
# median logCPM
abline(h = median(logcounts_Sex8), col = myPalette8, legend = TRUE)
# title('Boxplots of log Counts Per Million')
```

4.5.7 Calculations

```
# Data with actual counts, not TPM = 'dge'
head(dge_Sex8)
```

```
## An object of class "DGEList"
## $counts
##                                                       G1_SepF     G2_SepF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      1.790000e+02 118.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     8.410000e+02 746.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534     4.158820e-07   1.539241
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      2.860775e+01  30.766683
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401  0.000000e+00   0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993      7.600000e+01  35.000000
##                                                       G3_MarF     G4_MarM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304       49.0000000 113.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     645.0000000 997.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       0.5527639   1.114611
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       11.2812462   4.870982
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    0.0000000   0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993       11.0000000  13.000000
##                                                       G5_SepM     G6_SepM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304       77.00000   793.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     415.00000  9814.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       0.00000   136.63002
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       31.97337  7657.54555
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    0.00000    47.04729
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993       11.00000  2959.00000
##                                                       G7_MarF     G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      564.00000   147.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564   12448.00000  1577.00000
```

```
## PB.10.1|004815|path1:5-1713(+)|transcript/17534        139.72439     0.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426        3505.65982    13.46247
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401      34.68487     0.00000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993        4547.00000    27.00000
##
## $samples
##         group lib.size norm.factors
## G1_SepF     1  5568916            1
## G2_SepF     1  4333716            1
## G3_MarF     1  3738942            1
## G4_MarM     2  4341981            1
## G5_SepM     2  3387520            1
## G6_SepM     2 41354721            1
## G7_MarF     1 43500763            1
## G8_MarF     1  4747833            1
##
## $genes
##                                                      Length EffectiveLength
## PB.2.1|002537|path0:1-1624(+)|transcript/18304         1530        1197.281
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564        2065        1732.281
## PB.10.1|004815|path1:5-1713(+)|transcript/17534        1612        1279.281
## PB.11.1|004815|path2:1-3039(+)|transcript/3426         2999        2666.281
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401     1965        1632.281
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993         2714        2381.281
##                                                      Overdispersion
## PB.2.1|002537|path0:1-1624(+)|transcript/18304             1.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564            1.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534            2.880729
## PB.11.1|004815|path2:1-3039(+)|transcript/3426             1.188489
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401         3.028806
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993             1.000000
```

`head`(myTPM_Sex8)

```
##                                                          G1_SepF      G2_SepF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      1.790000e+02 118.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     8.410000e+02 746.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534     4.158820e-07   1.539241
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      2.860775e+01  30.766683
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401  0.000000e+00   0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993      7.600000e+01  35.000000
##                                                        G3_MarF      G4_MarM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      49.0000000 113.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564    645.0000000 997.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534      0.5527639   1.114611
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      11.2812462   4.870982
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401   0.0000000   0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993      11.0000000  13.000000
##                                                       G5_SepM     G6_SepM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      77.00000   793.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564    415.00000  9814.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534      0.00000   136.63002
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      31.97337  7657.54555
```

```
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    0.00000    47.04729
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993     11.00000 2959.00000
##                                                          G7_MarF      G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304        564.00000   147.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     12448.00000 1577.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534        139.72439     0.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426        3505.65982    13.46247
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401      34.68487     0.00000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993        4547.00000    27.00000
```

filter out genes with less than 5 reads in 2 samples for each transcript

```
filter_Sex8 <- apply(dge_Sex8, 1, function(x) length(x[x > 5]) >=
    2)
filtered_Sex8 <- dge_Sex8[filter_Sex8, ]
head(filtered_Sex8)
```

```
## An object of class "DGEList"
## $counts
##                                                          G1_SepF      G2_SepF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      1.790000e+02 118.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     8.410000e+02 746.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534     4.158820e-07   1.539241
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      2.860775e+01  30.766683
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 0.000000e+00   0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993      7.600000e+01  35.000000
##                                                          G3_MarF      G4_MarM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304        49.0000000 113.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564      645.0000000 997.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534        0.5527639   1.114611
## PB.11.1|004815|path2:1-3039(+)|transcript/3426        11.2812462   4.870982
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401     0.0000000   0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993        11.0000000  13.000000
##                                                          G5_SepM      G6_SepM
## PB.2.1|002537|path0:1-1624(+)|transcript/18304         77.00000   793.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564       415.00000  9814.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534         0.00000   136.63002
## PB.11.1|004815|path2:1-3039(+)|transcript/3426         31.97337  7657.54555
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401      0.00000    47.04729
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993         11.00000  2959.00000
##                                                          G7_MarF      G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304        564.00000   147.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     12448.00000  1577.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       139.72439     0.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       3505.65982    13.46247
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401     34.68487     0.00000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993       4547.00000    27.00000
##
## $samples
##          group lib.size norm.factors
## G1_SepF      1  5568916            1
## G2_SepF      1  4333716            1
## G3_MarF      1  3738942            1
```

```
## G4_MarM      2  4341981             1
## G5_SepM      2  3387520             1
## G6_SepM      2 41354721             1
## G7_MarF      1 43500763             1
## G8_MarF      1  4747833             1
##
## $genes
##                                                  Length EffectiveLength
## PB.2.1|002537|path0:1-1624(+)|transcript/18304     1530         1197.281
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564    2065         1732.281
## PB.10.1|004815|path1:5-1713(+)|transcript/17534    1612         1279.281
## PB.11.1|004815|path2:1-3039(+)|transcript/3426     2999         2666.281
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 1965         1632.281
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993     2714         2381.281
##                                                  Overdispersion
## PB.2.1|002537|path0:1-1624(+)|transcript/18304         1.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564        1.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534        2.880729
## PB.11.1|004815|path2:1-3039(+)|transcript/3426         1.188489
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401     3.028806
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993         1.000000
```

Run calculations

```
dge_Sex8 <- calcNormFactors(dge_Sex8)
dge_Sex8 <- estimateCommonDisp(dge_Sex8)
dge_Sex8 <- estimateTagwiseDisp(dge_Sex8)
```

4.5.8   Differentially expressed genes:

View the top ten 'differentially expressed genes'

```
results_Sex8 <- exactTest(dge_Sex8)
topTags(results_Sex8)
```

```
## Comparison of groups:  2-1
##                                                  Length EffectiveLength
## PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195   1136        803.2807
## PB.2770.1|6751e7|path1:1-1210(+)|transcript/22021   1103        770.2807
## PB.637.1|1826ff|path3:1-1155(+)|transcript/22860     905        572.3308
## PB.979.1|25edbb|path0:1-3100(+)|transcript/3575     3066       2733.2807
## PB.602.1|167374|path2:1-1363(+)|transcript/21155    1304        971.2807
## PB.3895.1|91ba83|path6:1-1594(+)|transcript/18563   1376       1043.2807
## PB.4152.1|9a8a98|path2:1-992(+)|transcript/22901     896        563.3308
## PB.5140.3|c12866|path2:10-2402(+)|transcript/8294   2359       2026.2807
## PB.2493.6|5d0eeb|path5:27-5305(+)|transcript/20823  1337       1004.2807
## PB.419.2|0f691f|path1:48-1473(+)|transcript/20262   1261        928.2807
##                                                  Overdispersion      logFC
## PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195       1.381179   8.978789
## PB.2770.1|6751e7|path1:1-1210(+)|transcript/22021       3.333104  10.106744
## PB.637.1|1826ff|path3:1-1155(+)|transcript/22860        1.020264   8.208384
## PB.979.1|25edbb|path0:1-3100(+)|transcript/3575         1.000000   9.770416
```

```
## PB.602.1|167374|path2:1-1363(+)|transcript/21155          1.150674  8.495981
## PB.3895.1|91ba83|path6:1-1594(+)|transcript/18563         3.245530  9.669816
## PB.4152.1|9a8a98|path2:1-992(+)|transcript/22901          1.733808  8.200739
## PB.5140.3|c12866|path2:10-2402(+)|transcript/8294         1.000000  8.957352
## PB.2493.6|5d0eeb|path5:27-5305(+)|transcript/20823        6.388942 10.964074
## PB.419.2|0f691f|path1:48-1473(+)|transcript/20262         3.986170  9.411064
##                                                   logCPM        PValue
## PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195 9.608930 6.213295e-32
## PB.2770.1|6751e7|path1:1-1210(+)|transcript/22021 5.693299 6.559852e-32
## PB.637.1|1826ff|path3:1-1155(+)|transcript/22860  7.447866 5.678994e-20
## PB.979.1|25edbb|path0:1-3100(+)|transcript/3575   4.970157 7.054220e-19
## PB.602.1|167374|path2:1-1363(+)|transcript/21155  4.468639 1.521816e-17
## PB.3895.1|91ba83|path6:1-1594(+)|transcript/18563 7.501882 5.359428e-17
## PB.4152.1|9a8a98|path2:1-992(+)|transcript/22901  6.700558 8.047662e-17
## PB.5140.3|c12866|path2:10-2402(+)|transcript/8294 5.885019 3.358521e-16
## PB.2493.6|5d0eeb|path5:27-5305(+)|transcript/20823 7.178685 1.527940e-15
## PB.419.2|0f691f|path1:48-1473(+)|transcript/20262 6.238564 2.419542e-15
##                                                            FDR
## PB.2769.1|6751e7|path0:4-1327(+)|transcript/21195 3.218591e-28
## PB.2770.1|6751e7|path1:1-1210(+)|transcript/22021 3.218591e-28
## PB.637.1|1826ff|path3:1-1155(+)|transcript/22860  1.857599e-16
## PB.979.1|25edbb|path0:1-3100(+)|transcript/3575   1.730577e-15
## PB.602.1|167374|path2:1-1363(+)|transcript/21155  2.986716e-14
## PB.3895.1|91ba83|path6:1-1594(+)|transcript/18563 8.765344e-14
## PB.4152.1|9a8a98|path2:1-992(+)|transcript/22901  1.128167e-13
## PB.5140.3|c12866|path2:10-2402(+)|transcript/8294 4.119646e-13
## PB.2493.6|5d0eeb|path5:27-5305(+)|transcript/20823 1.665964e-12
## PB.419.2|0f691f|path1:48-1473(+)|transcript/20262 2.374296e-12
```

Output summary expression data

```
dim(results_Sex8)
```

```
## [1] 9813    3
```

```
summary(de_Sex8 <- decideTests(results_Sex8))
```

```
##           2-1
## Down        3
## NotSig   9520
## Up        290
```

```
Summary_Sex8 <- (de_Sex8 <- decideTests(results_Sex8))
```

```
write.table(Summary_Sex8, file = "summary_Sex8.txt")
```

```
detags_Sex8 <- rownames(dge_Sex8)[as.logical(de_Sex8)]
```

Visualise with a plotsmear

```
# export the plot as png
png("plotSmear_Sex8.png", width = 600)
plotSmear(results_Sex8, de.tags = detags_Sex8)
```
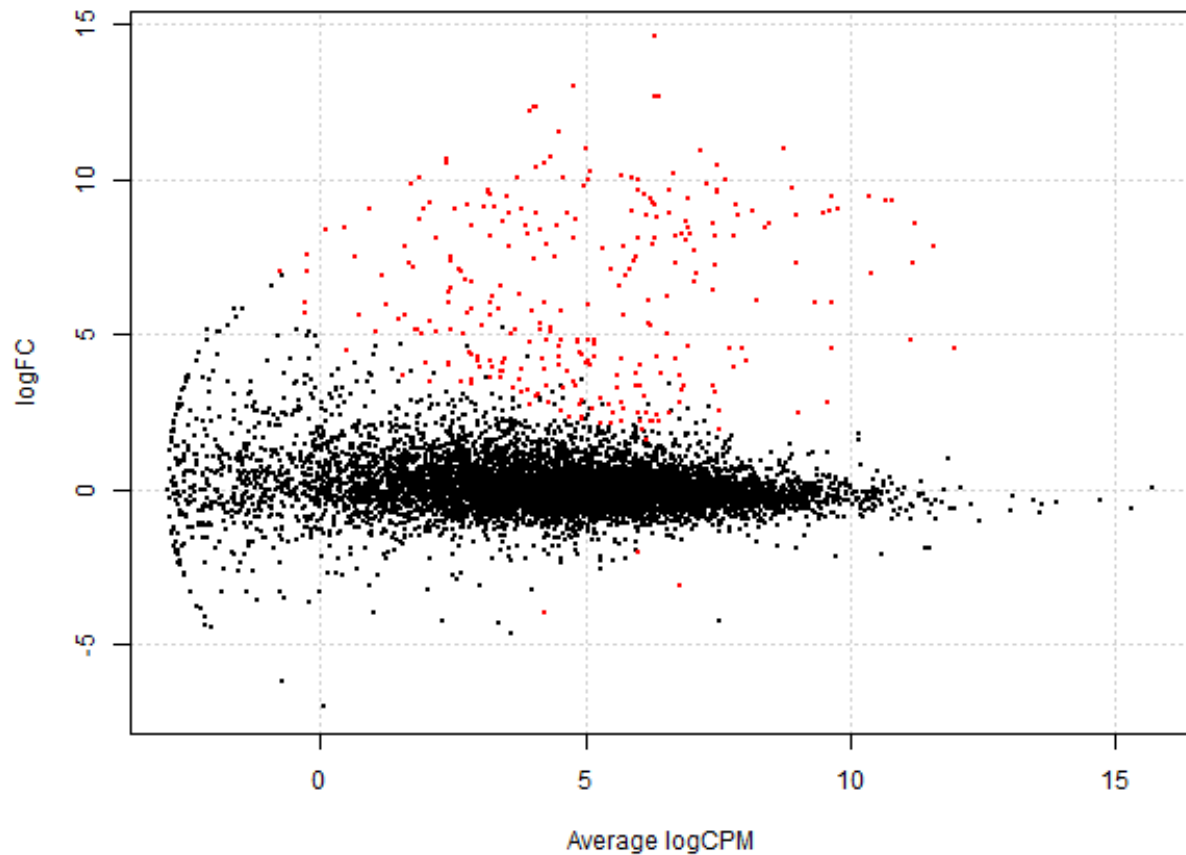

Figure 27: Log Fold Change vs average Log counts per million for expression level comparing Male to Female as variables among eight *T. adelaidensis* individuals. Red dots indicate transcripts with a significant change over season, with p-values <0.05.

Visualise with a volcano plot

```
# export the plot as png
png("volcanoData_Sex8.png", width = 600)
volcanoData_Sex8 <- cbind(results_Sex8$table$logFC, -log10(results_Sex8$table$PValue))
colnames(volcanoData_Sex8) <- c("logFC", "negLogPval")
DEGs_Sex8 <- results_Sex8$table$PValue < 0.05 & abs(results_Sex8$table$logFC) >
    2
point.col_Sex8 <- ifelse(DEGs_Sex8, "red", "black")
plot(volcanoData_Sex8, pch = 16, col = point.col_Sex8, cex = 0.5)
```
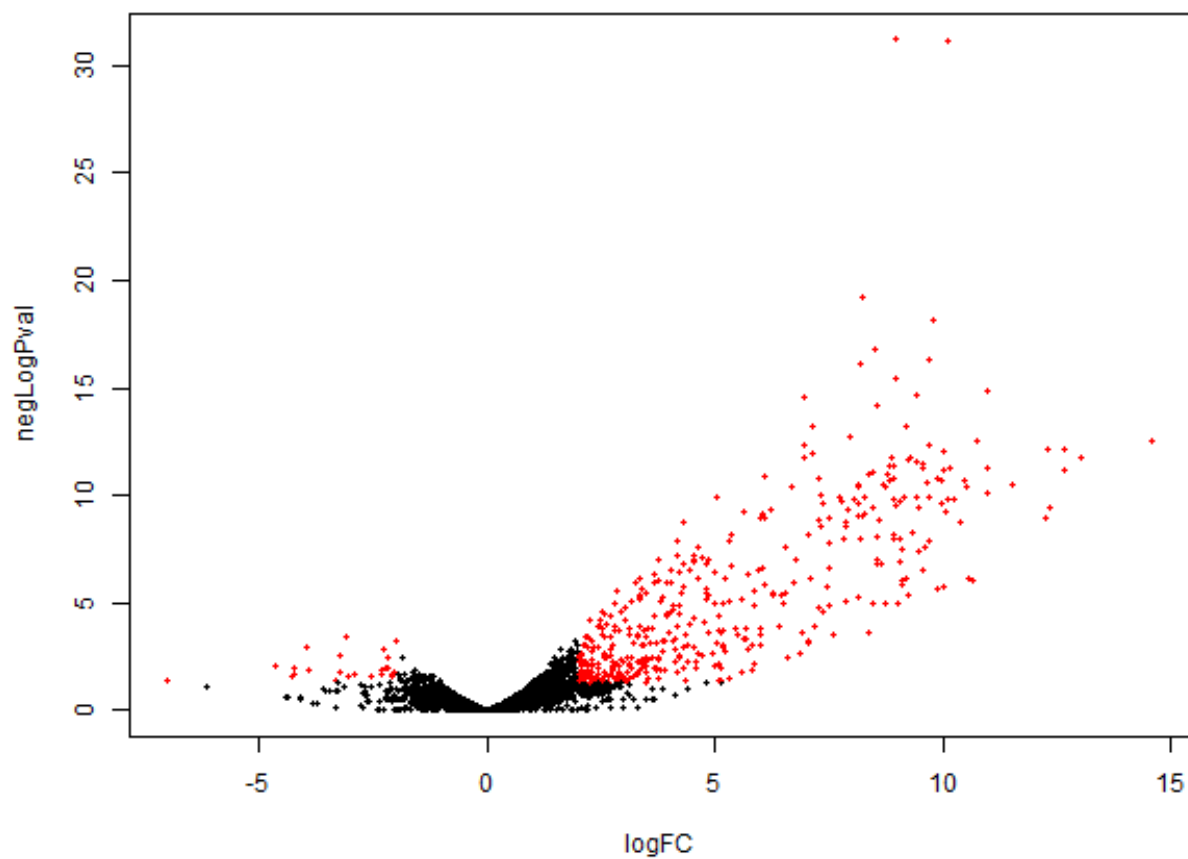

Figure 28: Negative Log of the p-value against Log Fold Change in expression level comparing Male to Female as variables among eight *T. adelaidensis* individuals. Red dots indicate transcripts with a significant change over season, with p-values <0.05 and *also* a log fold change >2.

4.6 Four *T. adelaidensis* individuals collected between two seasonal periods.

Exploration of 2:2 Female samples by Season group factor:

Skink G5 represents a September Male, and is also markedly different from the remaining 5 samples on the MDS plot. There is a chance G5 is responsible for a large amount of the observed variation in expression. These Four females represent two from September, and two from March/April, and were sequenced on the same run. This section is purely exploratory and to determine what, if any, results are consistent or markedly different on the removal of sample G5 (and the other male G4).

Import the data:

```
# create paths
setwd("~/@Uni/@Flinders University PhD/RWorkingDir/R-kallisto-clstr-4/kallisto-clstr/")
DIR4 <- ("~/@Uni/@Flinders University
↪  PhD/RWorkingDir/R-kallisto-clstr-4/kallisto-clstr/")
head(DIR4)
```

```
## [1] "~/@Uni/@Flinders University PhD/RWorkingDir/R-kallisto-clstr-4/kallisto-clstr/"
```

```
paths4F <- list.dirs(path = DIR4, full.names = FALSE, recursive = FALSE)
# the working dir only includes kallisto output for the
# kidneys at this time View(paths4F)

Kcaught4F <- catchKallisto(paths4F, verbose = TRUE)
```

```
## Reading G1_SepF, 9813 transcripts, 100 bootstraps
## Reading G2_SepF, 9813 transcripts, 100 bootstraps
## Reading G3_MarF, 9813 transcripts, 100 bootstraps
## Reading G8_MarF, 9813 transcripts, 100 bootstraps
```

```
setwd("~/@Uni/@Flinders University PhD/RWorkingDir/R-kallisto-clstr-4")
# View(Kcaught4F)
```

Set the group factors. Input Sample files are listed G1-G8 Group factors are applied to the DGElist object and are the point of comparison for expression analyses.

4.6.0.1 Season of collection: 1 = March, 2 = Sep. For the purposes of consistent labelling per group factors some April collections are referred to in the March group. Accurate collection information and dates are outlined in the methods chapter.

```
# these factors correlate to season of collection. 1 =
# March, 2 = Sep.
Season4F_group <- factor(c(2, 2, 1, 1))
```

4.6.0.2 Individuals sex: 1 = Female, 2 = Male.

4.6.1 Group Factor: Season

Create the EdgeR DGE list

```
dge_Season4F <- DGEList(counts = Kcaught4F$counts/Kcaught4F$annotation$Overdispersion,
    genes = Kcaught4F$annotation, group = Season4F_group)
# View(dge_Season4F)
names(dge_Season4F)
```

```
## [1] "counts"  "samples" "genes"
```

```
dge_Season4F
```

```
## An object of class "DGEList"
## $counts
##                                                      G1_SepF      G2_SepF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304     1.790000e+02 118.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564    8.410000e+02 746.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534    3.001053e-07   1.110734
## PB.11.1|004815|path2:1-3039(+)|transcript/3426     2.746754e+01  29.540420
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 0.000000e+00   0.000000
##                                                      G3_MarF   G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      49.0000000  147.0000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564    645.0000000 1577.0000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534      0.3988809    0.0000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      10.8316113   12.9259
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401   0.0000000    0.0000
## 9808 more rows ...
##
## $samples
##          group lib.size norm.factors
## G1_SepF      2  5788550            1
## G2_SepF      2  4478056            1
## G3_MarF      1  3874896            1
## G8_MarF      1  4915432            1
##
## $genes
##                                                    Length EffectiveLength
## PB.2.1|002537|path0:1-1624(+)|transcript/18304       1530         1197.281
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564      2065         1732.281
## PB.10.1|004815|path1:5-1713(+)|transcript/17534      1612         1279.281
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       2999         2666.281
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401   1965         1632.281
##                                                    Overdispersion
## PB.2.1|002537|path0:1-1624(+)|transcript/18304           1.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564          1.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534          3.992075
## PB.11.1|004815|path2:1-3039(+)|transcript/3426           1.237825
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401       1.000000
## 9808 more rows ...
```

Obtain Counts Per Million

```
myTPM_Season4F <- dge_Season4F$counts
# head(myTPM_Season4F)
```

4.6.2 Initial Data Exploration

```
# Which values in myCPM are greater than 0.5?
thresh_Season4F <- myTPM_Season4F > 0.5
# This produces a logical matrix with TRUEs and FALSEs
head(thresh_Season4F)
```

```
##                                                    G1_SepF G2_SepF G3_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304       TRUE    TRUE    TRUE
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564      TRUE    TRUE    TRUE
## PB.10.1|004815|path1:5-1713(+)|transcript/17534     FALSE    TRUE   FALSE
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       TRUE    TRUE    TRUE
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401  FALSE   FALSE   FALSE
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993       TRUE    TRUE    TRUE
##                                                    G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304        TRUE
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564       TRUE
## PB.10.1|004815|path1:5-1713(+)|transcript/17534      FALSE
## PB.11.1|004815|path2:1-3039(+)|transcript/3426        TRUE
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401   FALSE
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993        TRUE
```

Insert colour pallate for following visualisations consistency

```
myPalette4 <- c("#999999", "#F0E442", "#56B4E9", "#CC79A7")
```

Plot CPM distribution with logged CPM

```
# export the plot as png
png("plotDensities_Season4F.png", width = 600)
unfilteredExpr_Season4F <- cpm(dge_Season4F, log = T)
plotDensities(unfilteredExpr_Season4F, col = myPalette4, legend = TRUE)
```
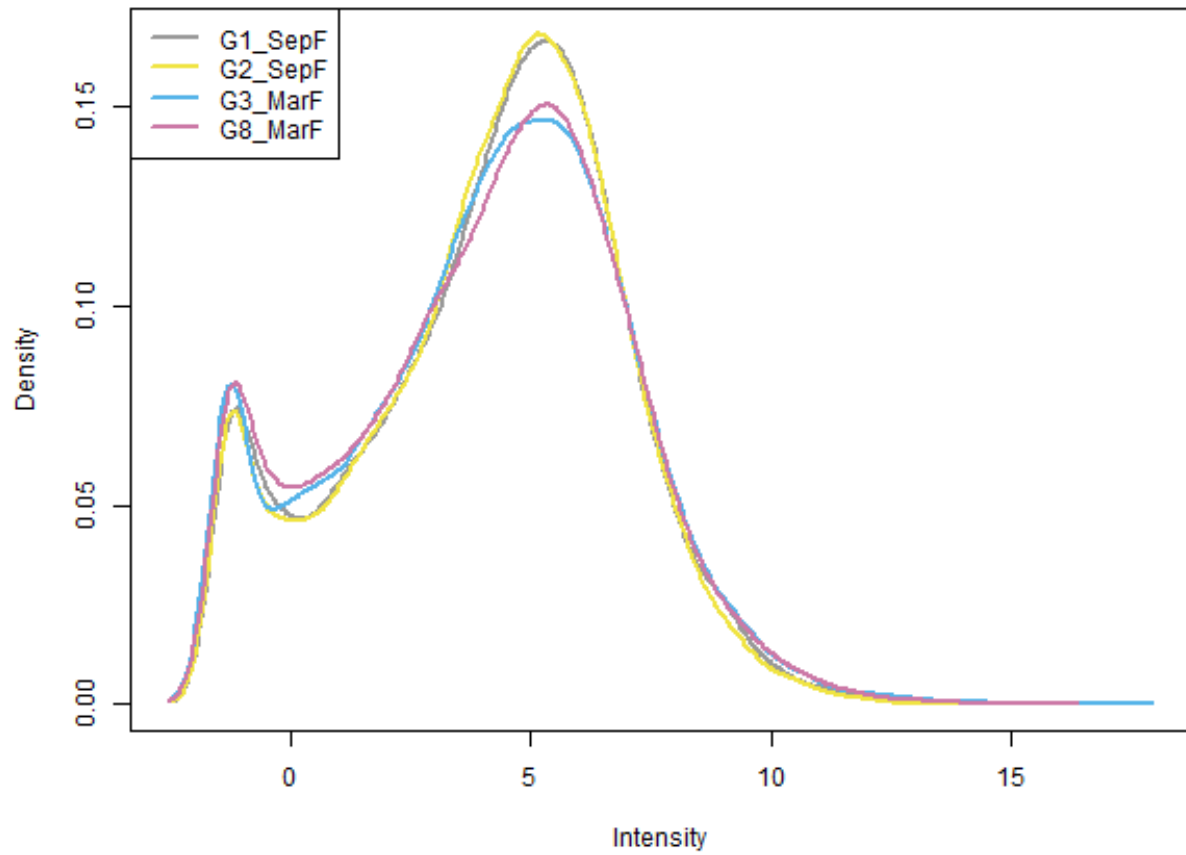

Figure 29: Log Counts per million Densities, for poly-a selected mRNA expression data from eight *T. adelaidensis* kidney samples

Visualise library sizes per sample

```
# The names argument tells the barplot to use the sample
# names on the x-axis The las argument rotates the axis
# names export the plot as png
png("Barplot.LibSize_Season4F.png", width = 600)
barplot(dge_Season4F$samples$lib.size, names = colnames(dge_Season4F),
    las = 2, col = myPalette4)
# title('Barplot of library sizes')
```
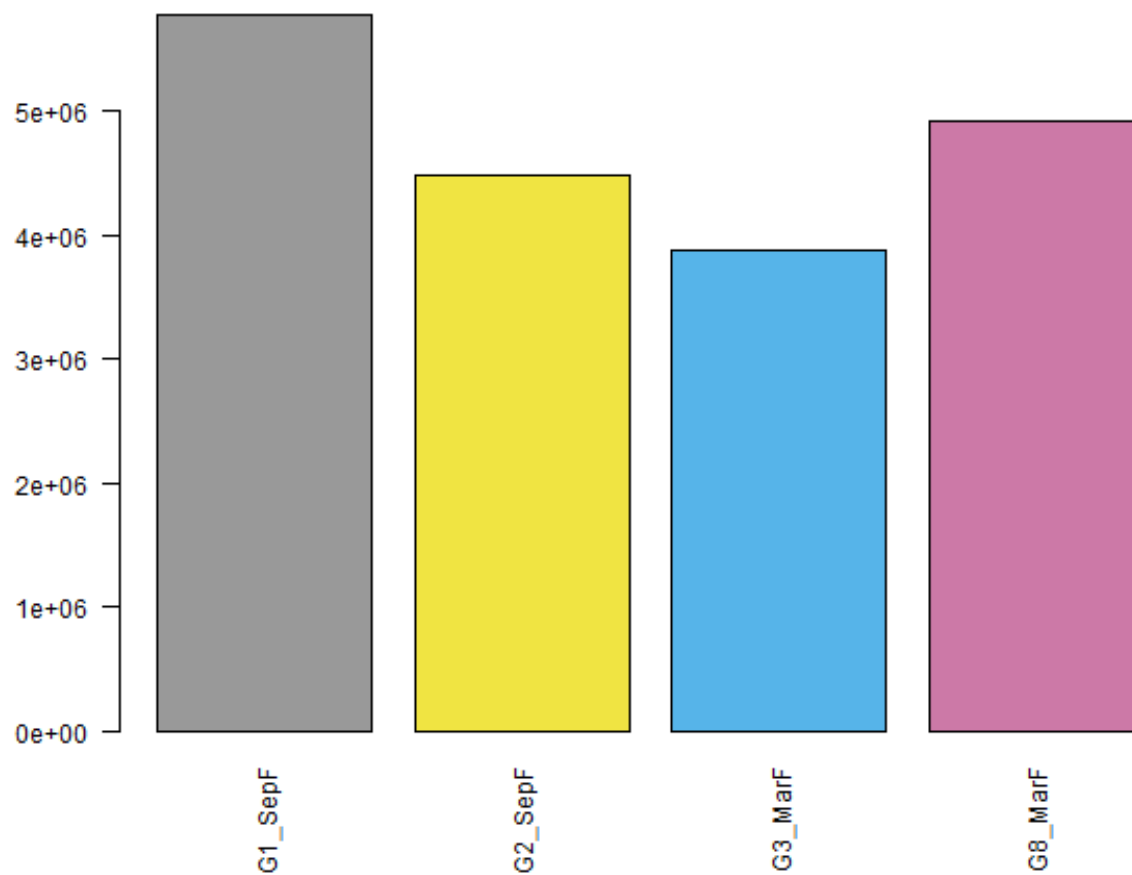

Figure 30: Library size of poly-a selected mRNA sequenced from kidney tissue of eight *T. adelaidensis* individuals

Visualise transcript total estimated counts per sample

```
# export the plot as png
png("Boxplot.Counts_Season4F.png", width = 600)
boxplot(dge_Season4F$counts, col = myPalette4, las = 2, legend = TRUE)
# title('Boxplot of total unfiltered Counts')
```
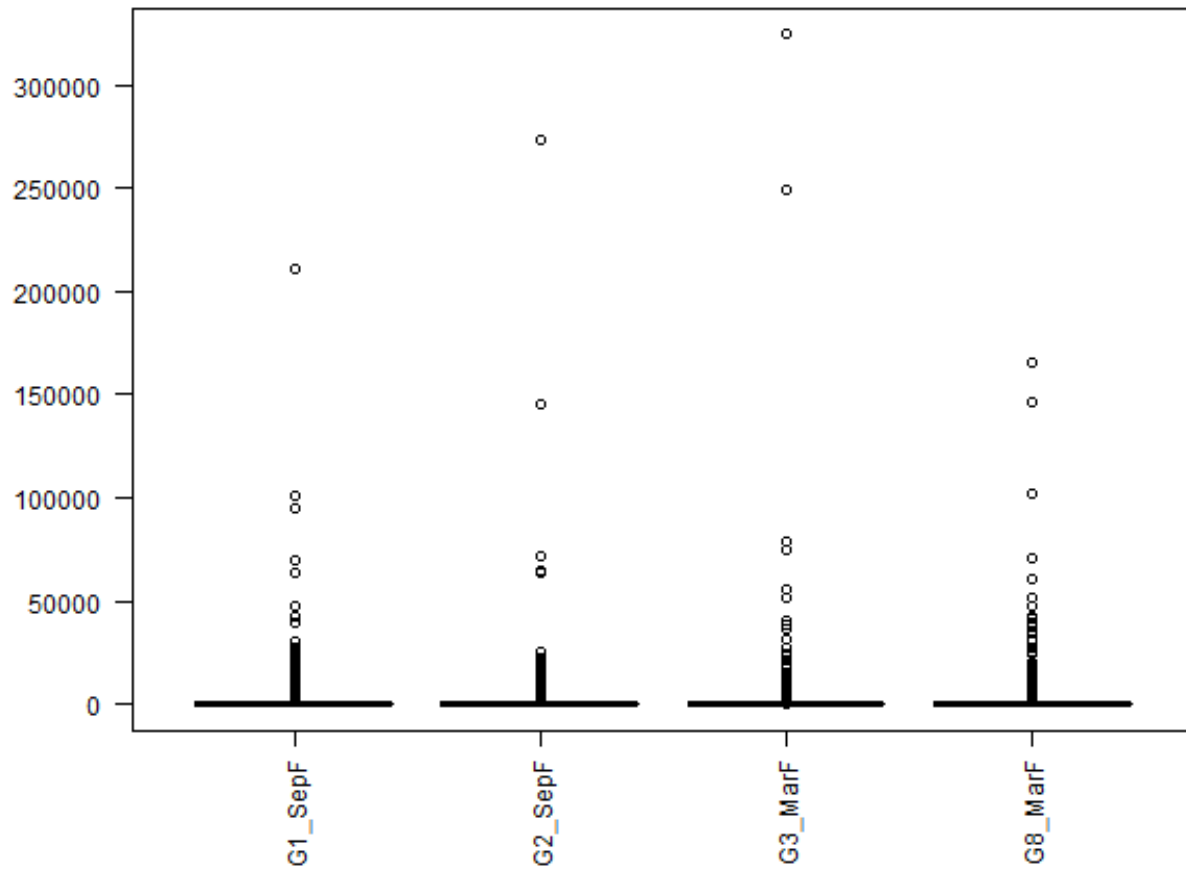

Figure 31: Total estimated counts, for poly-a selected mRNA expression data from eight *T. adelaidensis* kidney samples

```
# Get log2 counts per million
logcounts_Season4F <- cpm(dge_Season4F, log = TRUE)
# Check distributions of samples using boxplots export the
# plot as png
png("boxplot.CPM_Season4F.png", width = 600)
boxplot(logcounts_Season4F, xlab = "", ylab = "Log2 counts per million",
    las = 2, col = myPalette4, legend = TRUE)
# Let's add a blue horizontal line that corresponds to the
# median logCPM
abline(h = median(logcounts_Season4F), col = myPalette4, legend = TRUE)
# title('Boxplots of log Counts Per Million')
```
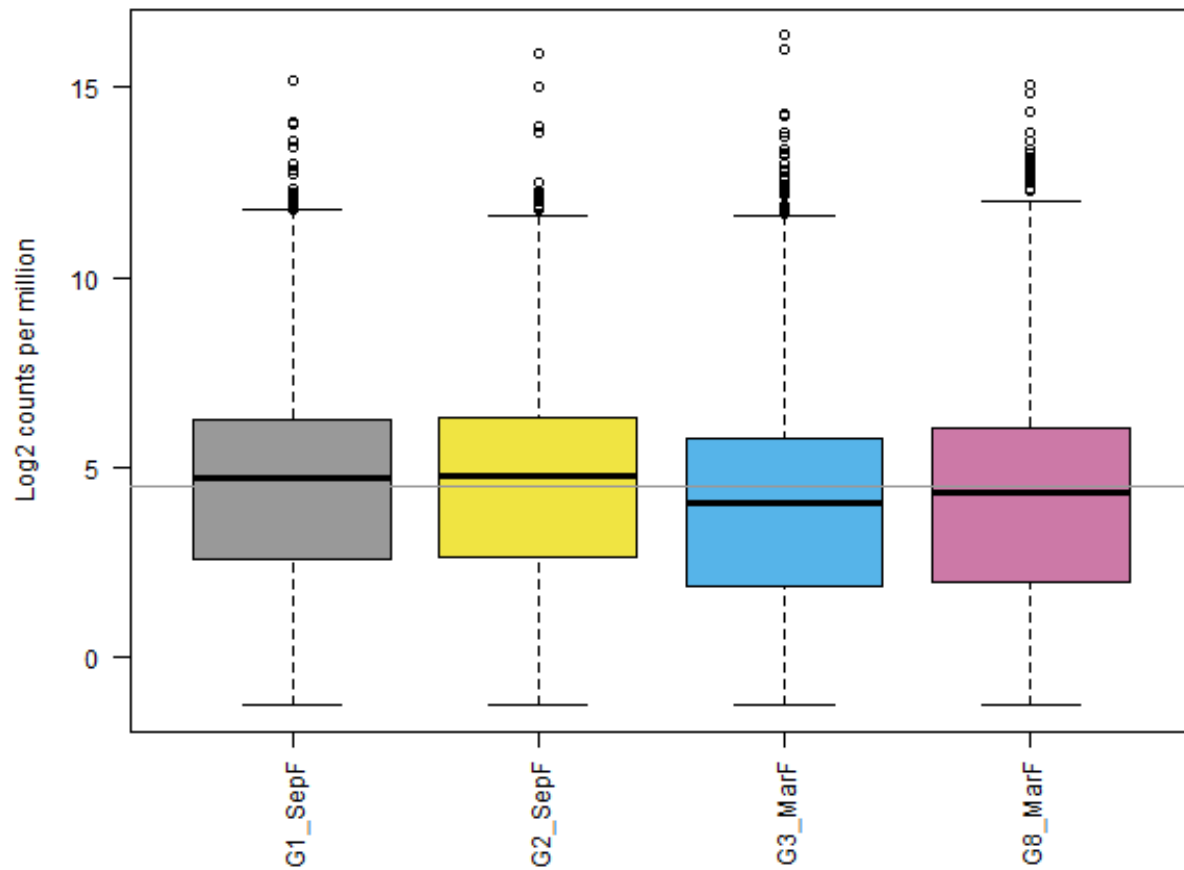

Figure 32: Log2 of estimated counts per million, for poly-a selected mRNA expression data from eight *T. adelaidensis* kidney samples

Visualise sample variation using an MDS plot

```
# export the plot as png
png("MDS_Season4F.png", width = 600)
plotMDS(dge_Season4F, col = myPalette4)
# title('MDS plot')
```
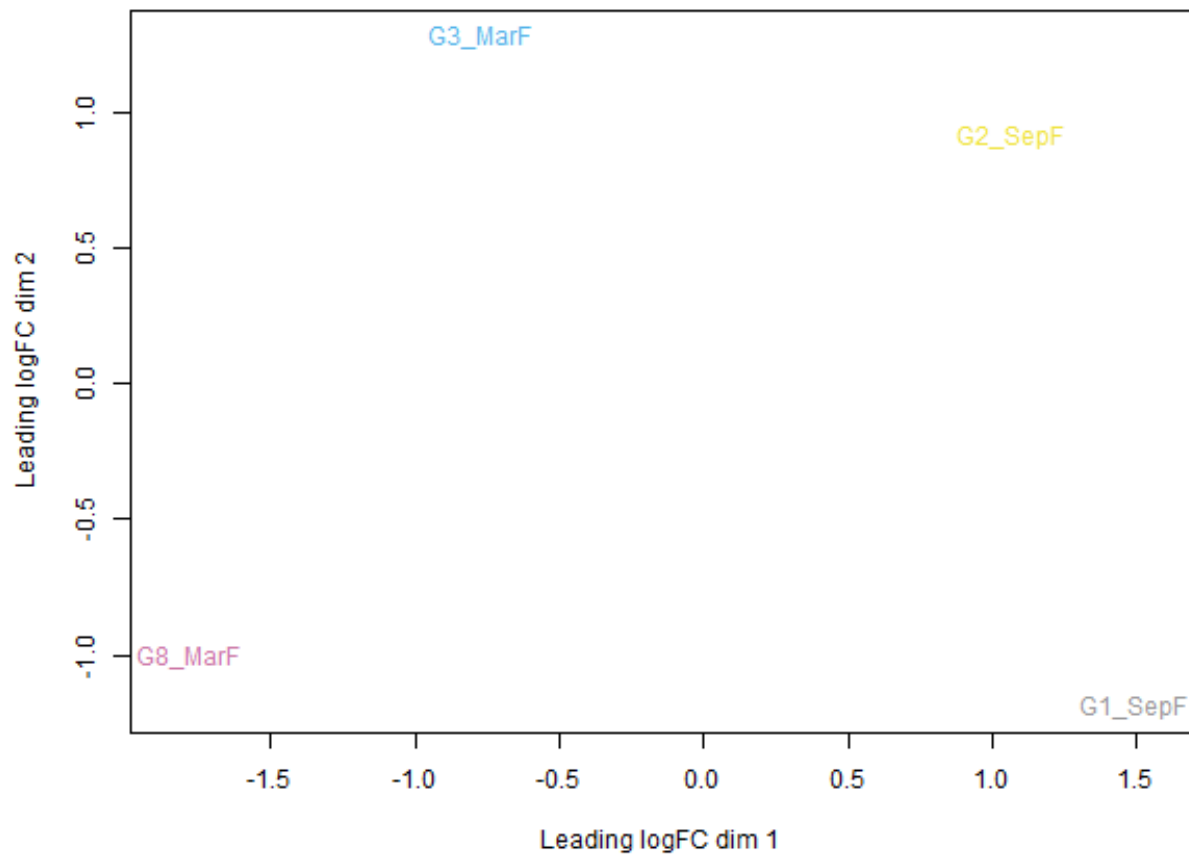

Figure 33: MDS plot of variation among eight *T. adelaidensis* individuals

4.6.3 Calculations

```
# Data with actual counts, not TPM = 'dge'
head(dge_Season4F)
```

```
## An object of class "DGEList"
## $counts
##                                                        G1_SepF      G2_SepF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304        1.790000e+02 118.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564       8.410000e+02 746.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       3.001053e-07   1.110734
## PB.11.1|004815|path2:1-3039(+)|transcript/3426        2.746754e+01  29.540420
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    0.000000e+00   0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993        7.189450e+01  33.109308
##                                                        G3_MarF      G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304         49.0000000  147.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564       645.0000000 1577.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534         0.3988809    0.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426         10.8316113   12.92590
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401      0.0000000    0.00000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993         10.4057825   25.54147
##
## $samples
##          group lib.size norm.factors
## G1_SepF      2  5788550            1
## G2_SepF      2  4478056            1
## G3_MarF      1  3874896            1
## G8_MarF      1  4915432            1
##
## $genes
##                                                        Length EffectiveLength
## PB.2.1|002537|path0:1-1624(+)|transcript/18304           1530        1197.281
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564          2065        1732.281
## PB.10.1|004815|path1:5-1713(+)|transcript/17534          1612        1279.281
## PB.11.1|004815|path2:1-3039(+)|transcript/3426           2999        2666.281
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401       1965        1632.281
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993           2714        2381.281
##                                                        Overdispersion
## PB.2.1|002537|path0:1-1624(+)|transcript/18304              1.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564             1.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534             3.992075
## PB.11.1|004815|path2:1-3039(+)|transcript/3426              1.237825
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401          1.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993              1.057105
```

```
head(myTPM_Season4F)
```

```
##                                                        G1_SepF      G2_SepF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304        1.790000e+02 118.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564       8.410000e+02 746.000000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       3.001053e-07   1.110734
```

```
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       2.746754e+01  29.540420
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401 0.000000e+00   0.000000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993       7.189450e+01  33.109308
##                                                          G3_MarF     G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304       49.0000000  147.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     645.0000000 1577.00000
## PB.10.1|004815|path1:5-1713(+)|transcript/17534       0.3988809    0.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426       10.8316113   12.92590
## PB.11.2|004815|path2:1047-3035(+)|transcript/13401    0.0000000    0.00000
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993       10.4057825   25.54147
```

filter out genes with less than 5 reads in 2 samples for each transcript

```
filter_Season4F <- apply(dge_Season4F, 1, function(x) length(x[x >
    5]) >= 2)
filtered_Season4F <- dge_Season4F[filter_Season4F, ]
head(filtered_Season4F)
```

```
## An object of class "DGEList"
## $counts
##                                                   G1_SepF    G2_SepF    G3_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304  179.00000 118.00000   49.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564 841.00000 746.00000  645.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426   27.46754  29.54042   10.83161
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993   71.89450  33.10931   10.40578
## PB.14.1|006283|path0:1-1955(+)|transcript/13821 693.03920 432.58624  319.03235
## PB.15.1|006e23|path0:1-3614(+)|transcript/1774  624.00000 527.00000  351.00000
##                                                    G8_MarF
## PB.2.1|002537|path0:1-1624(+)|transcript/18304     147.00000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564   1577.00000
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      12.92590
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993      25.54147
## PB.14.1|006283|path0:1-1955(+)|transcript/13821    540.73280
## PB.15.1|006e23|path0:1-3614(+)|transcript/1774     719.00000
##
## $samples
##         group lib.size norm.factors
## G1_SepF     2  5788550            1
## G2_SepF     2  4478056            1
## G3_MarF     1  3874896            1
## G8_MarF     1  4915432            1
##
## $genes
##                                                   Length EffectiveLength
## PB.2.1|002537|path0:1-1624(+)|transcript/18304      1530        1197.281
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564     2065        1732.281
## PB.11.1|004815|path2:1-3039(+)|transcript/3426      2999        2666.281
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993      2714        2381.281
## PB.14.1|006283|path0:1-1955(+)|transcript/13821     1920        1587.281
## PB.15.1|006e23|path0:1-3614(+)|transcript/1774      3583        3250.281
##                                                   Overdispersion
## PB.2.1|002537|path0:1-1624(+)|transcript/18304          1.000000
## PB.3.2|00361c|path0:43-2240(+)|transcript/10564         1.000000
```

```
## PB.11.1|004815|path2:1-3039(+)|transcript/3426          1.237825
## PB.13.1|004c10|path2:1-2811(+)|transcript/4993          1.057105
## PB.14.1|006283|path0:1-1955(+)|transcript/13821         1.109605
## PB.15.1|006e23|path0:1-3614(+)|transcript/1774          1.000000
```

Run calculations

```
dge_Season4F <- calcNormFactors(dge_Season4F)
dge_Season4F <- estimateCommonDisp(dge_Season4F)
dge_Season4F <- estimateTagwiseDisp(dge_Season4F)
```

Note that at this smaller scale filtering *does* change the appearance of the MDS plot

```
plotMDS(dge_Season4F, col = myPalette4)
```
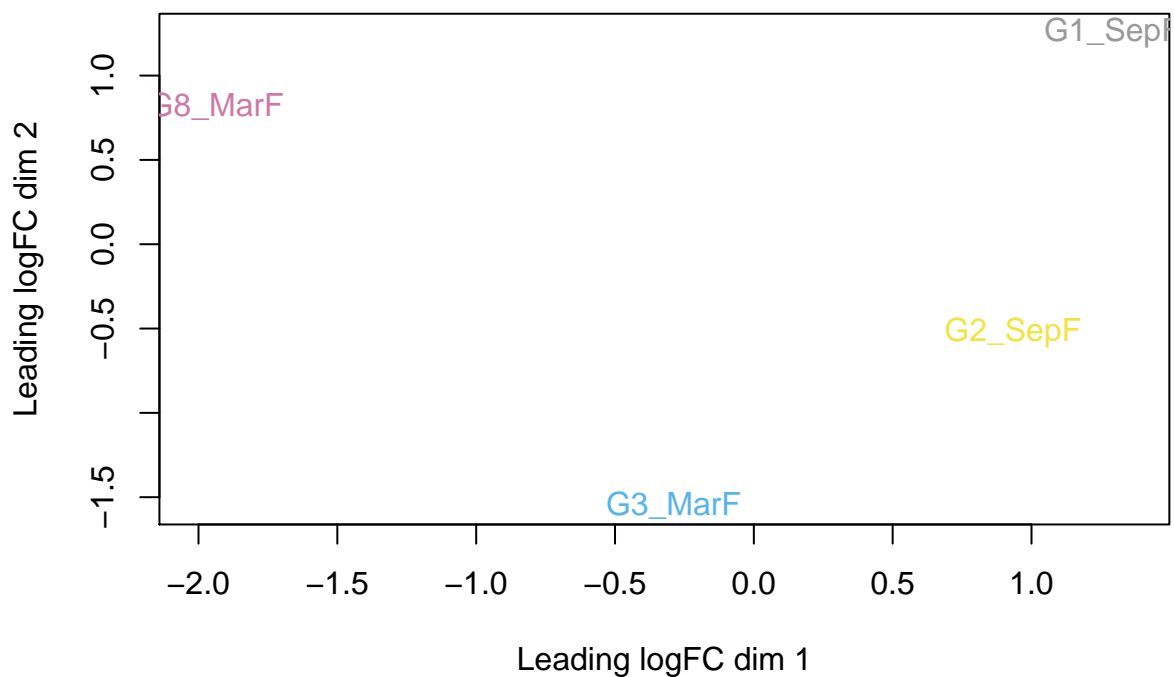

4.6.4 Differentially expressed genes:

View the top ten 'differentially expressed genes'

```
results_Season4F <- exactTest(dge_Season4F)
topTags(results_Season4F)
```

```
## Comparison of groups:  2-1
##                                                     Length EffectiveLength
## PB.4138.1|9a0bd1|path6:1-1549(+)|transcript/20634     1416       1083.2807
## PB.5802.6|dbfbde|path3:6-2293(+)|transcript/8938      2289       1956.2807
## PB.373.1|0da2d2|path16:1-2158(+)|transcript/14315     1803       1470.2807
## PB.4931.3|b8571c|path20:38-3608(+)|transcript/1966    3487       3154.2807
## PB.5398.2|cbf34d|path5:582-2207(+)|transcript/17939   1601       1268.2807
## PB.2869.1|6b0af0|path2:1-1117(+)|transcript/22387     1086        753.2807
## PB.9559.1|transcript/7471:1-2422(+)|transcript/7471   2391       2058.2807
## PB.2872.1|6b0af0|path5:1-947(+)|transcript/23097       916        583.3308
## PB.3849.1|906dad|path19:1-2010(+)|transcript/12836    1940       1607.2807
## PB.3850.1|906dad|path20:1-2023(+)|transcript/12300    1942       1609.2807
##                                                     Overdispersion      logFC
## PB.4138.1|9a0bd1|path6:1-1549(+)|transcript/20634         2.781267  -7.654449
## PB.5802.6|dbfbde|path3:6-2293(+)|transcript/8938          6.536090   6.079678
## PB.373.1|0da2d2|path16:1-2158(+)|transcript/14315         1.795692  -5.228715
## PB.4931.3|b8571c|path20:38-3608(+)|transcript/1966        2.701620  -7.141122
## PB.5398.2|cbf34d|path5:582-2207(+)|transcript/17939       1.901357   7.531794
## PB.2869.1|6b0af0|path2:1-1117(+)|transcript/22387         1.191076   3.998713
## PB.9559.1|transcript/7471:1-2422(+)|transcript/7471       1.000000   3.940997
## PB.2872.1|6b0af0|path5:1-947(+)|transcript/23097          1.163320  -4.625304
## PB.3849.1|906dad|path19:1-2010(+)|transcript/12836        3.008824   5.646231
## PB.3850.1|906dad|path20:1-2023(+)|transcript/12300        1.043555   4.659062
##                                                       logCPM       PValue
## PB.4138.1|9a0bd1|path6:1-1549(+)|transcript/20634    4.999647 2.789293e-22
## PB.5802.6|dbfbde|path3:6-2293(+)|transcript/8938     4.795953 1.169574e-19
## PB.373.1|0da2d2|path16:1-2158(+)|transcript/14315    4.485746 1.786410e-16
## PB.4931.3|b8571c|path20:38-3608(+)|transcript/1966   3.581631 4.445670e-15
## PB.5398.2|cbf34d|path5:582-2207(+)|transcript/17939  5.217766 2.082183e-13
## PB.2869.1|6b0af0|path2:1-1117(+)|transcript/22387    5.282699 7.190815e-13
## PB.9559.1|transcript/7471:1-2422(+)|transcript/7471  4.940297 7.630769e-13
## PB.2872.1|6b0af0|path5:1-947(+)|transcript/23097     4.240645 3.988756e-11
## PB.3849.1|906dad|path19:1-2010(+)|transcript/12836   5.911588 6.308377e-11
## PB.3850.1|906dad|path20:1-2023(+)|transcript/12300   3.877477 7.702300e-11
##                                                                  FDR
## PB.4138.1|9a0bd1|path6:1-1549(+)|transcript/20634    2.737133e-18
## PB.5802.6|dbfbde|path3:6-2293(+)|transcript/8938     5.738514e-16
## PB.373.1|0da2d2|path16:1-2158(+)|transcript/14315    5.843346e-13
## PB.4931.3|b8571c|path20:38-3608(+)|transcript/1966   1.090634e-11
## PB.5398.2|cbf34d|path5:582-2207(+)|transcript/17939  4.086493e-10
## PB.2869.1|6b0af0|path2:1-1117(+)|transcript/22387    1.069725e-09
## PB.9559.1|transcript/7471:1-2422(+)|transcript/7471  1.069725e-09
## PB.2872.1|6b0af0|path5:1-947(+)|transcript/23097     4.892708e-08
## PB.3849.1|906dad|path19:1-2010(+)|transcript/12836   6.878234e-08
## PB.3850.1|906dad|path20:1-2023(+)|transcript/12300   7.558267e-08
```

Output summary expression data

```
dim(results_Season4F)
```

```
## [1] 9813    3
```

```
summary(de_Season4F <- decideTests(results_Season4F))
```

```
##            2-1
## Down       62
## NotSig   9621
## Up        130
```

```
Summary_Season4F <- (de_Season4F <- decideTests(results_Season4F))

write.table(Summary_Season4F, file = "summary_Season4F.txt")

detags_Season4F <- rownames(dge_Season4F)[as.logical(de_Season4F)]
```

Visualise with a plotsmear

```
detags_Season4F <- rownames(dge_Season4F)[as.logical(de_Season4F)]
# export the plot as png
png("plotSmear_Season4F.png", width = 600)
plotSmear(results_Season4F, de.tags = detags_Season4F)
```
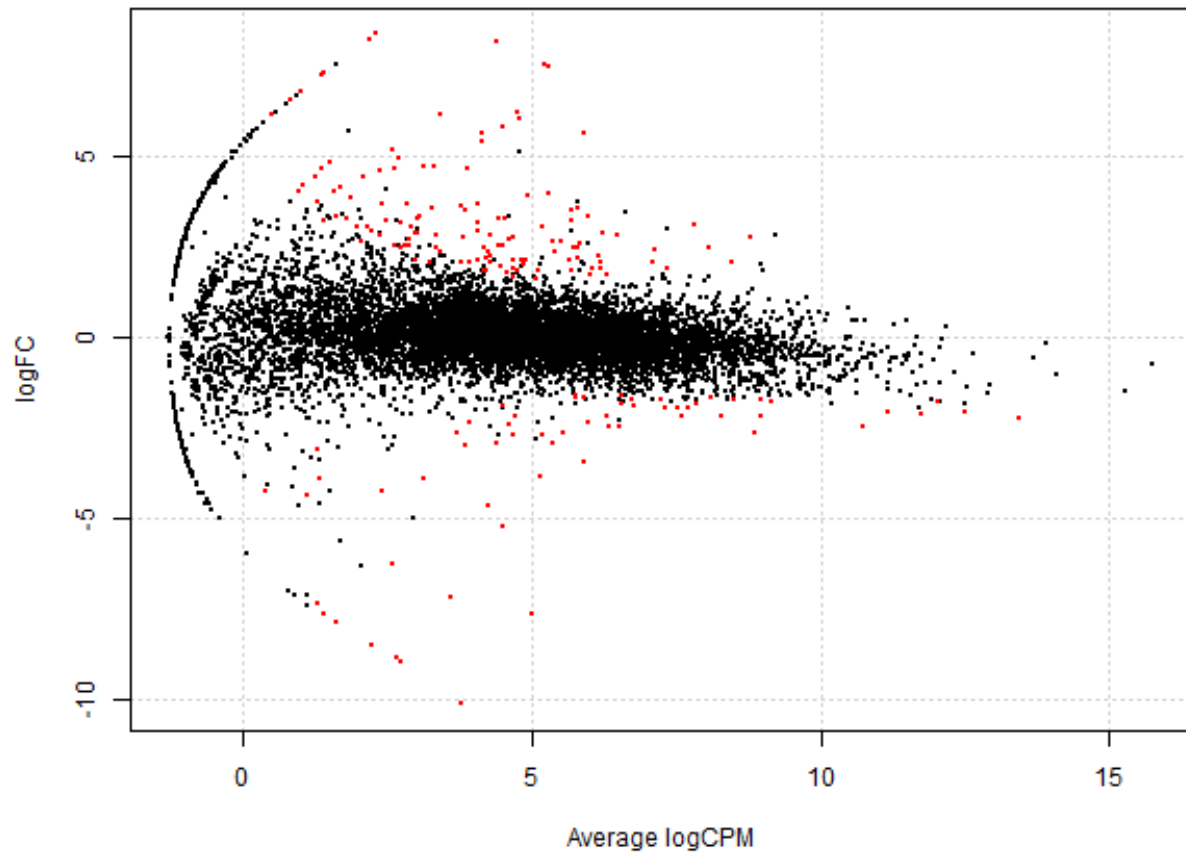

Figure 34: Log Fold Change vs average Log counts per million for expression level comparing Season 2 (September collection) to Season 1 (April/March collection) among eight *T. adelaidensis* individuals. Red dots indicate transcripts with a significant change over season, with p-values <0.05.

Visualise with a volcano plot

```
# export the plot as png
png("volcanoData_Season4F.png", width = 600)
volcanoData_Season4F <- cbind(results_Season4F$table$logFC,
↪    -log10(results_Season4F$table$PValue))
colnames(volcanoData_Season4F) <- c("logFC", "negLogPval")
DEGs_Season4F <- results_Season4F$table$PValue < 0.05 & abs(results_Season4F$table$logFC)
↪    >
    2
point.col_Season4F <- ifelse(DEGs_Season4F, "red", "black")
plot(volcanoData_Season4F, pch = 16, col = point.col_Season4F,
    cex = 0.5)
```
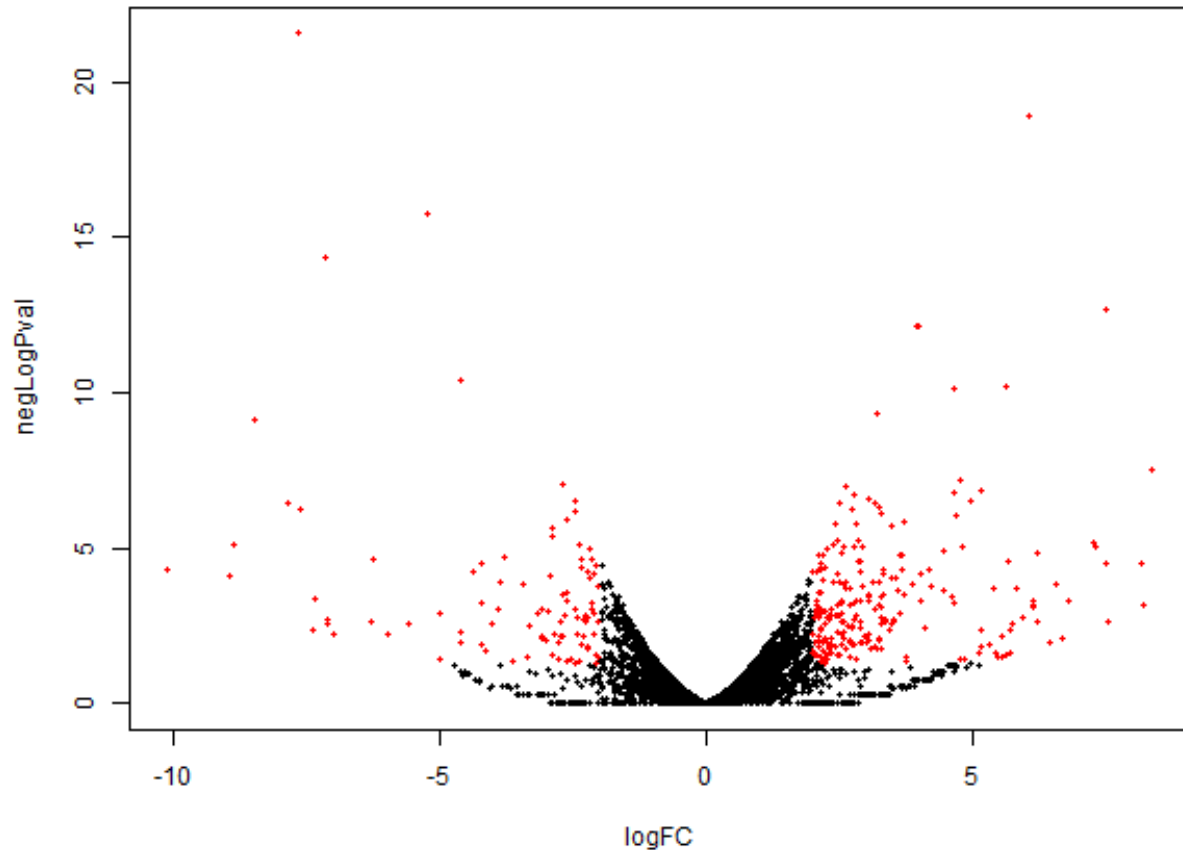

Figure 35: Negative Log of the p-value against Log Fold Change in expression level comparing Season 2 (September collection) to Season 1 (April/March collection) among eight *T. adelaidensis* individuals. Red dots indicate transcripts with a significant change over season, with p-values <0.05 and *also* a log fold change >2.