12-deg\_annot\_from\_hisat

Megan Ewing

2024-05-13

# do once  
# install.packages("R.utils")  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.3 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.3 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(R.utils)

## Loading required package: R.oo

## Warning: package 'R.oo' was built under R version 4.3.2

## Loading required package: R.methodsS3  
## R.methodsS3 v1.8.2 (2022-06-13 22:00:14 UTC) successfully loaded. See ?R.methodsS3 for help.  
## R.oo v1.26.0 (2024-01-24 05:12:50 UTC) successfully loaded. See ?R.oo for help.  
##   
## Attaching package: 'R.oo'  
##   
## The following object is masked from 'package:R.methodsS3':  
##   
## throw  
##   
## The following objects are masked from 'package:methods':  
##   
## getClasses, getMethods  
##   
## The following objects are masked from 'package:base':  
##   
## attach, detach, load, save  
##   
## R.utils v2.12.3 (2023-11-18 01:00:02 UTC) successfully loaded. See ?R.utils for help.  
##   
## Attaching package: 'R.utils'  
##   
## The following object is masked from 'package:tidyr':  
##   
## extract  
##   
## The following object is masked from 'package:utils':  
##   
## timestamp  
##   
## The following objects are masked from 'package:base':  
##   
## cat, commandArgs, getOption, isOpen, nullfile, parse, warnings

library(dplyr)

# Gene level with deseq2 pre filtering (10+reads avg of 3+samples)

## Annotating our top DEGs using Blast and Uniprot Gene Ontology (GO) ids

Read in DEG count and stats files

# for all counts: ../output/0807-logcounts\_allDEG\_8ind23r\_ToC.csv  
# for DEG stats: ../output/0807-DEGstats\_ToC\_8ind23r.tab  
  
#read in count results  
counts <- read.csv("../output/0807-logcounts\_allDEG\_8ind23r\_ToC.csv", row.names = 1)  
head(counts)

## M.C.193 M.C.216 M.C.218 M.C.226 M.C.306 M.C.329 M.C.334  
## LOC132758254 9.018234 8.991358 7.577010 7.789053 7.618000 7.903490 7.785037  
## LOC132745822 7.546969 8.093778 8.444167 8.048840 9.286832 9.820501 7.614251  
## LOC132746399 5.276290 6.314388 7.936010 7.727373 6.929499 8.016098 6.245973  
## LOC132746792 5.789978 6.525297 5.486501 5.769676 6.385635 7.139049 6.165775  
## LOC132720785 5.407278 3.874463 3.474487 4.185148 3.179462 4.353156 4.736047  
## LOC132725056 7.031642 7.395884 6.558978 7.200879 6.139855 5.858405 5.528054  
## M.C.337 M.C.339 M.C.358 M.C.360 M.C.363 M.C.373 M.C.482  
## LOC132758254 8.362197 8.102469 7.730976 7.431279 8.489877 6.870841 7.056003  
## LOC132745822 8.489700 6.665863 7.782664 7.177897 8.498897 6.770201 8.030656  
## LOC132746399 6.409642 5.553355 6.873344 7.109438 5.692989 7.077499 5.222981  
## LOC132746792 6.409642 5.398225 7.458745 5.432316 6.991711 6.799679 5.763178  
## LOC132720785 5.329371 3.402581 3.944095 4.941024 3.968640 4.973514 4.657204  
## LOC132725056 6.074927 6.692819 6.392831 6.916854 5.924841 6.473688 6.463304  
## M.C.488 M.T.18 M.T.20 M.T.22 M.T.235 M.T.245 M.T.29  
## LOC132758254 7.866917 7.002917 6.278039 6.744291 7.172415 6.699427 6.653879  
## LOC132745822 8.887287 6.953388 5.858643 6.806725 7.443312 6.929311 6.682167  
## LOC132746399 6.724682 4.798929 4.682751 5.706205 5.549657 5.462240 6.317607  
## LOC132746792 6.372353 4.325212 4.101964 5.247309 5.513920 5.108990 5.487575  
## LOC132720785 5.076312 5.817759 6.399378 6.957439 4.356273 4.640188 5.281139  
## LOC132725056 5.981913 5.119464 4.101964 5.095524 6.653936 5.296400 3.455454  
## M.T.31 M.T.32 M.T.399 M.T.43 M.T.44 M.T.500 M.T.7  
## LOC132758254 7.674504 6.960515 6.236251 6.410440 6.419648 6.607779 7.169018  
## LOC132745822 6.836827 6.504231 6.675271 7.866451 6.984149 7.760471 6.843521  
## LOC132746399 5.470048 5.218768 5.602092 6.129826 5.613907 6.159865 5.885884  
## LOC132746792 4.901269 5.370087 5.188391 5.218074 5.850376 6.092670 4.609669  
## LOC132720785 6.237069 4.898547 6.252709 7.016638 7.447696 5.158157 4.870806  
## LOC132725056 4.789809 5.049696 4.324042 4.892587 5.200124 5.506812 4.561152  
## M.T.83 M.T.8  
## LOC132758254 6.848184 7.134442  
## LOC132745822 7.864893 5.499278  
## LOC132746399 4.885269 4.420540  
## LOC132746792 5.927677 4.633276  
## LOC132720785 6.186756 4.565791  
## LOC132725056 6.689869 4.982870

# add in stats  
stats <- read.csv("../output/0807-DEGstats\_ToC\_8ind23r.tab", sep="")  
head(stats)

## baseMean log2FoldChange lfcSE stat pvalue  
## LOC132742068 68.04920 -1.4845094 0.3347656 -4.434474 9.229718e-06  
## LOC132738510 114.22423 -0.9442138 0.2535605 -3.723821 1.962304e-04  
## LOC132727777 27.05036 1.8022735 0.4081725 4.415470 1.007906e-05  
## LOC132728236 169.99708 0.5275304 0.1333260 3.956695 7.599388e-05  
## LOC132740341 163.84395 -1.0096084 0.2543228 -3.969790 7.193589e-05  
## LOC132741452 3020.53904 -0.6475836 0.1418720 -4.564563 5.005354e-06  
## padj  
## LOC132742068 0.011302182  
## LOC132738510 0.048396396  
## LOC132727777 0.011302182  
## LOC132728236 0.031379420  
## LOC132740341 0.031379420  
## LOC132741452 0.007098218

Prep to join them (could be skipped by using the join\_by( a == b), but I already wrote this chunk so..

# make column with LOC names  
  
LOCnames <- rownames(counts)  
  
counts$LOC <- LOCnames  
  
head(counts)

## M.C.193 M.C.216 M.C.218 M.C.226 M.C.306 M.C.329 M.C.334  
## LOC132758254 9.018234 8.991358 7.577010 7.789053 7.618000 7.903490 7.785037  
## LOC132745822 7.546969 8.093778 8.444167 8.048840 9.286832 9.820501 7.614251  
## LOC132746399 5.276290 6.314388 7.936010 7.727373 6.929499 8.016098 6.245973  
## LOC132746792 5.789978 6.525297 5.486501 5.769676 6.385635 7.139049 6.165775  
## LOC132720785 5.407278 3.874463 3.474487 4.185148 3.179462 4.353156 4.736047  
## LOC132725056 7.031642 7.395884 6.558978 7.200879 6.139855 5.858405 5.528054  
## M.C.337 M.C.339 M.C.358 M.C.360 M.C.363 M.C.373 M.C.482  
## LOC132758254 8.362197 8.102469 7.730976 7.431279 8.489877 6.870841 7.056003  
## LOC132745822 8.489700 6.665863 7.782664 7.177897 8.498897 6.770201 8.030656  
## LOC132746399 6.409642 5.553355 6.873344 7.109438 5.692989 7.077499 5.222981  
## LOC132746792 6.409642 5.398225 7.458745 5.432316 6.991711 6.799679 5.763178  
## LOC132720785 5.329371 3.402581 3.944095 4.941024 3.968640 4.973514 4.657204  
## LOC132725056 6.074927 6.692819 6.392831 6.916854 5.924841 6.473688 6.463304  
## M.C.488 M.T.18 M.T.20 M.T.22 M.T.235 M.T.245 M.T.29  
## LOC132758254 7.866917 7.002917 6.278039 6.744291 7.172415 6.699427 6.653879  
## LOC132745822 8.887287 6.953388 5.858643 6.806725 7.443312 6.929311 6.682167  
## LOC132746399 6.724682 4.798929 4.682751 5.706205 5.549657 5.462240 6.317607  
## LOC132746792 6.372353 4.325212 4.101964 5.247309 5.513920 5.108990 5.487575  
## LOC132720785 5.076312 5.817759 6.399378 6.957439 4.356273 4.640188 5.281139  
## LOC132725056 5.981913 5.119464 4.101964 5.095524 6.653936 5.296400 3.455454  
## M.T.31 M.T.32 M.T.399 M.T.43 M.T.44 M.T.500 M.T.7  
## LOC132758254 7.674504 6.960515 6.236251 6.410440 6.419648 6.607779 7.169018  
## LOC132745822 6.836827 6.504231 6.675271 7.866451 6.984149 7.760471 6.843521  
## LOC132746399 5.470048 5.218768 5.602092 6.129826 5.613907 6.159865 5.885884  
## LOC132746792 4.901269 5.370087 5.188391 5.218074 5.850376 6.092670 4.609669  
## LOC132720785 6.237069 4.898547 6.252709 7.016638 7.447696 5.158157 4.870806  
## LOC132725056 4.789809 5.049696 4.324042 4.892587 5.200124 5.506812 4.561152  
## M.T.83 M.T.8 LOC  
## LOC132758254 6.848184 7.134442 LOC132758254  
## LOC132745822 7.864893 5.499278 LOC132745822  
## LOC132746399 4.885269 4.420540 LOC132746399  
## LOC132746792 5.927677 4.633276 LOC132746792  
## LOC132720785 6.186756 4.565791 LOC132720785  
## LOC132725056 6.689869 4.982870 LOC132725056

LOCstat <- rownames(stats)  
  
stats$LOC <- LOCstat  
  
head(stats)

## baseMean log2FoldChange lfcSE stat pvalue  
## LOC132742068 68.04920 -1.4845094 0.3347656 -4.434474 9.229718e-06  
## LOC132738510 114.22423 -0.9442138 0.2535605 -3.723821 1.962304e-04  
## LOC132727777 27.05036 1.8022735 0.4081725 4.415470 1.007906e-05  
## LOC132728236 169.99708 0.5275304 0.1333260 3.956695 7.599388e-05  
## LOC132740341 163.84395 -1.0096084 0.2543228 -3.969790 7.193589e-05  
## LOC132741452 3020.53904 -0.6475836 0.1418720 -4.564563 5.005354e-06  
## padj LOC  
## LOC132742068 0.011302182 LOC132742068  
## LOC132738510 0.048396396 LOC132738510  
## LOC132727777 0.011302182 LOC132727777  
## LOC132728236 0.031379420 LOC132728236  
## LOC132740341 0.031379420 LOC132740341  
## LOC132741452 0.007098218 LOC132741452

Join them

# joining stats and counts  
DEG <- left\_join(x = counts, y = stats, by = "LOC")  
head(DEG)

## M.C.193 M.C.216 M.C.218 M.C.226 M.C.306 M.C.329 M.C.334 M.C.337  
## 1 9.018234 8.991358 7.577010 7.789053 7.618000 7.903490 7.785037 8.362197  
## 2 7.546969 8.093778 8.444167 8.048840 9.286832 9.820501 7.614251 8.489700  
## 3 5.276290 6.314388 7.936010 7.727373 6.929499 8.016098 6.245973 6.409642  
## 4 5.789978 6.525297 5.486501 5.769676 6.385635 7.139049 6.165775 6.409642  
## 5 5.407278 3.874463 3.474487 4.185148 3.179462 4.353156 4.736047 5.329371  
## 6 7.031642 7.395884 6.558978 7.200879 6.139855 5.858405 5.528054 6.074927  
## M.C.339 M.C.358 M.C.360 M.C.363 M.C.373 M.C.482 M.C.488 M.T.18  
## 1 8.102469 7.730976 7.431279 8.489877 6.870841 7.056003 7.866917 7.002917  
## 2 6.665863 7.782664 7.177897 8.498897 6.770201 8.030656 8.887287 6.953388  
## 3 5.553355 6.873344 7.109438 5.692989 7.077499 5.222981 6.724682 4.798929  
## 4 5.398225 7.458745 5.432316 6.991711 6.799679 5.763178 6.372353 4.325212  
## 5 3.402581 3.944095 4.941024 3.968640 4.973514 4.657204 5.076312 5.817759  
## 6 6.692819 6.392831 6.916854 5.924841 6.473688 6.463304 5.981913 5.119464  
## M.T.20 M.T.22 M.T.235 M.T.245 M.T.29 M.T.31 M.T.32 M.T.399  
## 1 6.278039 6.744291 7.172415 6.699427 6.653879 7.674504 6.960515 6.236251  
## 2 5.858643 6.806725 7.443312 6.929311 6.682167 6.836827 6.504231 6.675271  
## 3 4.682751 5.706205 5.549657 5.462240 6.317607 5.470048 5.218768 5.602092  
## 4 4.101964 5.247309 5.513920 5.108990 5.487575 4.901269 5.370087 5.188391  
## 5 6.399378 6.957439 4.356273 4.640188 5.281139 6.237069 4.898547 6.252709  
## 6 4.101964 5.095524 6.653936 5.296400 3.455454 4.789809 5.049696 4.324042  
## M.T.43 M.T.44 M.T.500 M.T.7 M.T.83 M.T.8 LOC baseMean  
## 1 6.410440 6.419648 6.607779 7.169018 6.848184 7.134442 LOC132758254 187.73392  
## 2 7.866451 6.984149 7.760471 6.843521 7.864893 5.499278 LOC132745822 225.66028  
## 3 6.129826 5.613907 6.159865 5.885884 4.885269 4.420540 LOC132746399 80.87262  
## 4 5.218074 5.850376 6.092670 4.609669 5.927677 4.633276 LOC132746792 60.55140  
## 5 7.016638 7.447696 5.158157 4.870806 6.186756 4.565791 LOC132720785 43.73189  
## 6 4.892587 5.200124 5.506812 4.561152 6.689869 4.982870 LOC132725056 65.01466  
## log2FoldChange lfcSE stat pvalue padj  
## 1 -1.188610 0.2003729 -5.931989 2.992873e-09 3.395415e-05  
## 2 -1.299604 0.2742359 -4.738998 2.147771e-06 3.929306e-03  
## 3 -1.324071 0.2742349 -4.828235 1.377486e-06 3.929306e-03  
## 4 -1.141412 0.2388462 -4.778857 1.762949e-06 3.929306e-03  
## 5 1.570396 0.3214278 4.885689 1.030675e-06 3.929306e-03  
## 6 -1.269575 0.2692982 -4.714384 2.424428e-06 3.929306e-03

Read in [annotated genome file](https://www.ncbi.nlm.nih.gov/datasets/gene/GCF_026571515.1/). Its named blast here because originally it was a blast file, but I didn’t want to go and edit all the following chunks

# read in annotationg  
blast <- read.delim("../data/ncbi\_dataset.tsv")  
head(blast)

## Accession Begin End Chromosome Orientation Name  
## 1 NC\_031332.1 1762 3330 MT plus cytochrome c oxidase subunit II  
## 2 NC\_031332.1 3397 3461 MT plus tRNA-Pro  
## 3 NC\_031332.1 3438 4685 MT plus cytochrome b  
## 4 NC\_031332.1 4686 6093 MT plus 16S ribosomal RNA  
## 5 NC\_031332.1 6094 7452 MT plus NADH dehydrogenase subunit 4  
## 6 NC\_031332.1 7453 7514 MT plus tRNA-His  
## Symbol Gene.ID Gene.Type Transcripts.accession Transcript.name  
## 1 COX2 29141288 protein-coding   
## 2 29141300 tRNA   
## 3 CYTB 29141301 protein-coding   
## 4 29141289 rRNA   
## 5 ND4 29141302 protein-coding   
## 6 29141290 tRNA   
## Protein.accession Protein.name Protein.length Locus.tag  
## 1 YP\_009305271.1 cytochrome c oxidase subunit II 522 BJM09\_gp02  
## 2 NA BJM09\_gt01  
## 3 YP\_009305272.1 cytochrome b 415 BJM09\_gp13  
## 4 NA BJM09\_gr02  
## 5 YP\_009305273.1 NADH dehydrogenase subunit 4 452 BJM09\_gp12  
## 6 NA BJM09\_gt02

Joining annotation file to list of DEGs

# join by gene name  
  
DEG\_Blast <- left\_join(x = DEG, y = blast, join\_by("LOC" == "Symbol"))  
head(DEG\_Blast)

## M.C.193 M.C.216 M.C.218 M.C.226 M.C.306 M.C.329 M.C.334 M.C.337  
## 1 9.018234 8.991358 7.577010 7.789053 7.618000 7.903490 7.785037 8.362197  
## 2 9.018234 8.991358 7.577010 7.789053 7.618000 7.903490 7.785037 8.362197  
## 3 7.546969 8.093778 8.444167 8.048840 9.286832 9.820501 7.614251 8.489700  
## 4 7.546969 8.093778 8.444167 8.048840 9.286832 9.820501 7.614251 8.489700  
## 5 5.276290 6.314388 7.936010 7.727373 6.929499 8.016098 6.245973 6.409642  
## 6 5.789978 6.525297 5.486501 5.769676 6.385635 7.139049 6.165775 6.409642  
## M.C.339 M.C.358 M.C.360 M.C.363 M.C.373 M.C.482 M.C.488 M.T.18  
## 1 8.102469 7.730976 7.431279 8.489877 6.870841 7.056003 7.866917 7.002917  
## 2 8.102469 7.730976 7.431279 8.489877 6.870841 7.056003 7.866917 7.002917  
## 3 6.665863 7.782664 7.177897 8.498897 6.770201 8.030656 8.887287 6.953388  
## 4 6.665863 7.782664 7.177897 8.498897 6.770201 8.030656 8.887287 6.953388  
## 5 5.553355 6.873344 7.109438 5.692989 7.077499 5.222981 6.724682 4.798929  
## 6 5.398225 7.458745 5.432316 6.991711 6.799679 5.763178 6.372353 4.325212  
## M.T.20 M.T.22 M.T.235 M.T.245 M.T.29 M.T.31 M.T.32 M.T.399  
## 1 6.278039 6.744291 7.172415 6.699427 6.653879 7.674504 6.960515 6.236251  
## 2 6.278039 6.744291 7.172415 6.699427 6.653879 7.674504 6.960515 6.236251  
## 3 5.858643 6.806725 7.443312 6.929311 6.682167 6.836827 6.504231 6.675271  
## 4 5.858643 6.806725 7.443312 6.929311 6.682167 6.836827 6.504231 6.675271  
## 5 4.682751 5.706205 5.549657 5.462240 6.317607 5.470048 5.218768 5.602092  
## 6 4.101964 5.247309 5.513920 5.108990 5.487575 4.901269 5.370087 5.188391  
## M.T.43 M.T.44 M.T.500 M.T.7 M.T.83 M.T.8 LOC baseMean  
## 1 6.410440 6.419648 6.607779 7.169018 6.848184 7.134442 LOC132758254 187.73392  
## 2 6.410440 6.419648 6.607779 7.169018 6.848184 7.134442 LOC132758254 187.73392  
## 3 7.866451 6.984149 7.760471 6.843521 7.864893 5.499278 LOC132745822 225.66028  
## 4 7.866451 6.984149 7.760471 6.843521 7.864893 5.499278 LOC132745822 225.66028  
## 5 6.129826 5.613907 6.159865 5.885884 4.885269 4.420540 LOC132746399 80.87262  
## 6 5.218074 5.850376 6.092670 4.609669 5.927677 4.633276 LOC132746792 60.55140  
## log2FoldChange lfcSE stat pvalue padj Accession  
## 1 -1.188610 0.2003729 -5.931989 2.992873e-09 3.395415e-05 NW\_026859905.1  
## 2 -1.188610 0.2003729 -5.931989 2.992873e-09 3.395415e-05 NW\_026859905.1  
## 3 -1.299604 0.2742359 -4.738998 2.147771e-06 3.929306e-03 NW\_026856216.1  
## 4 -1.299604 0.2742359 -4.738998 2.147771e-06 3.929306e-03 NW\_026856216.1  
## 5 -1.324071 0.2742349 -4.828235 1.377486e-06 3.929306e-03 NW\_026856403.1  
## 6 -1.141412 0.2388462 -4.778857 1.762949e-06 3.929306e-03 NW\_026856504.1  
## Begin End Chromosome Orientation  
## 1 204763 227007 plus  
## 2 204763 227007 plus  
## 3 29641 46037 minus  
## 4 29641 46037 minus  
## 5 242141 268014 plus  
## 6 70333 123083 plus  
## Name Gene.ID  
## 1 golgin subfamily A member 3-like 132758254  
## 2 golgin subfamily A member 3-like 132758254  
## 3 titin homolog 132745822  
## 4 titin homolog 132745822  
## 5 FYVE and coiled-coil domain-containing protein 1-like 132746399  
## 6 phospholipase D1-like 132746792  
## Gene.Type Transcripts.accession  
## 1 protein-coding XM\_060749812.1  
## 2 protein-coding XM\_060749813.1  
## 3 protein-coding XM\_060734808.1  
## 4 protein-coding XM\_060734809.1  
## 5 protein-coding XM\_060735521.1  
## 6 protein-coding XM\_060736043.1  
## Transcript.name Protein.accession  
## 1 golgin subfamily A member 3-like, transcript variant X1 XP\_060605795.1  
## 2 golgin subfamily A member 3-like, transcript variant X2 XP\_060605796.1  
## 3 titin homolog, transcript variant X1 XP\_060590791.1  
## 4 titin homolog, transcript variant X2 XP\_060590792.1  
## 5 FYVE and coiled-coil domain-containing protein 1-like XP\_060591504.1  
## 6 phospholipase D1-like XP\_060592026.1  
## Protein.name Protein.length  
## 1 golgin subfamily A member 3-like isoform X1 1674  
## 2 golgin subfamily A member 3-like isoform X2 1563  
## 3 titin homolog isoform X1 1350  
## 4 titin homolog isoform X2 1331  
## 5 FYVE and coiled-coil domain-containing protein 1-like 2583  
## 6 phospholipase D1-like 330  
## Locus.tag  
## 1   
## 2   
## 3   
## 4   
## 5   
## 6

# there's multiple entries for each LOC because of different isoforms, but we want just the gene level, so we are going to remove duplicates  
  
DEG\_Blast <- distinct(DEG\_Blast, LOC, .keep\_all = TRUE)  
head(DEG\_Blast)

## M.C.193 M.C.216 M.C.218 M.C.226 M.C.306 M.C.329 M.C.334 M.C.337  
## 1 9.018234 8.991358 7.577010 7.789053 7.618000 7.903490 7.785037 8.362197  
## 2 7.546969 8.093778 8.444167 8.048840 9.286832 9.820501 7.614251 8.489700  
## 3 5.276290 6.314388 7.936010 7.727373 6.929499 8.016098 6.245973 6.409642  
## 4 5.789978 6.525297 5.486501 5.769676 6.385635 7.139049 6.165775 6.409642  
## 5 5.407278 3.874463 3.474487 4.185148 3.179462 4.353156 4.736047 5.329371  
## 6 7.031642 7.395884 6.558978 7.200879 6.139855 5.858405 5.528054 6.074927  
## M.C.339 M.C.358 M.C.360 M.C.363 M.C.373 M.C.482 M.C.488 M.T.18  
## 1 8.102469 7.730976 7.431279 8.489877 6.870841 7.056003 7.866917 7.002917  
## 2 6.665863 7.782664 7.177897 8.498897 6.770201 8.030656 8.887287 6.953388  
## 3 5.553355 6.873344 7.109438 5.692989 7.077499 5.222981 6.724682 4.798929  
## 4 5.398225 7.458745 5.432316 6.991711 6.799679 5.763178 6.372353 4.325212  
## 5 3.402581 3.944095 4.941024 3.968640 4.973514 4.657204 5.076312 5.817759  
## 6 6.692819 6.392831 6.916854 5.924841 6.473688 6.463304 5.981913 5.119464  
## M.T.20 M.T.22 M.T.235 M.T.245 M.T.29 M.T.31 M.T.32 M.T.399  
## 1 6.278039 6.744291 7.172415 6.699427 6.653879 7.674504 6.960515 6.236251  
## 2 5.858643 6.806725 7.443312 6.929311 6.682167 6.836827 6.504231 6.675271  
## 3 4.682751 5.706205 5.549657 5.462240 6.317607 5.470048 5.218768 5.602092  
## 4 4.101964 5.247309 5.513920 5.108990 5.487575 4.901269 5.370087 5.188391  
## 5 6.399378 6.957439 4.356273 4.640188 5.281139 6.237069 4.898547 6.252709  
## 6 4.101964 5.095524 6.653936 5.296400 3.455454 4.789809 5.049696 4.324042  
## M.T.43 M.T.44 M.T.500 M.T.7 M.T.83 M.T.8 LOC baseMean  
## 1 6.410440 6.419648 6.607779 7.169018 6.848184 7.134442 LOC132758254 187.73392  
## 2 7.866451 6.984149 7.760471 6.843521 7.864893 5.499278 LOC132745822 225.66028  
## 3 6.129826 5.613907 6.159865 5.885884 4.885269 4.420540 LOC132746399 80.87262  
## 4 5.218074 5.850376 6.092670 4.609669 5.927677 4.633276 LOC132746792 60.55140  
## 5 7.016638 7.447696 5.158157 4.870806 6.186756 4.565791 LOC132720785 43.73189  
## 6 4.892587 5.200124 5.506812 4.561152 6.689869 4.982870 LOC132725056 65.01466  
## log2FoldChange lfcSE stat pvalue padj Accession  
## 1 -1.188610 0.2003729 -5.931989 2.992873e-09 3.395415e-05 NW\_026859905.1  
## 2 -1.299604 0.2742359 -4.738998 2.147771e-06 3.929306e-03 NW\_026856216.1  
## 3 -1.324071 0.2742349 -4.828235 1.377486e-06 3.929306e-03 NW\_026856403.1  
## 4 -1.141412 0.2388462 -4.778857 1.762949e-06 3.929306e-03 NW\_026856504.1  
## 5 1.570396 0.3214278 4.885689 1.030675e-06 3.929306e-03 NW\_026862958.1  
## 6 -1.269575 0.2692982 -4.714384 2.424428e-06 3.929306e-03 NW\_026864281.1  
## Begin End Chromosome Orientation  
## 1 204763 227007 plus  
## 2 29641 46037 minus  
## 3 242141 268014 plus  
## 4 70333 123083 plus  
## 5 82977 120119 plus  
## 6 197176 248330 minus  
## Name Gene.ID  
## 1 golgin subfamily A member 3-like 132758254  
## 2 titin homolog 132745822  
## 3 FYVE and coiled-coil domain-containing protein 1-like 132746399  
## 4 phospholipase D1-like 132746792  
## 5 malignant fibrous histiocytoma-amplified sequence 1 homolog 132720785  
## 6 serine-rich adhesin for platelets-like 132725056  
## Gene.Type Transcripts.accession  
## 1 protein-coding XM\_060749812.1  
## 2 protein-coding XM\_060734808.1  
## 3 protein-coding XM\_060735521.1  
## 4 protein-coding XM\_060736043.1  
## 5 protein-coding XM\_060704992.1  
## 6 protein-coding XM\_060710048.1  
## Transcript.name  
## 1 golgin subfamily A member 3-like, transcript variant X1  
## 2 titin homolog, transcript variant X1  
## 3 FYVE and coiled-coil domain-containing protein 1-like  
## 4 phospholipase D1-like  
## 5 malignant fibrous histiocytoma-amplified sequence 1 homolog, transcript variant X1  
## 6 serine-rich adhesin for platelets-like, transcript variant X1  
## Protein.accession  
## 1 XP\_060605795.1  
## 2 XP\_060590791.1  
## 3 XP\_060591504.1  
## 4 XP\_060592026.1  
## 5 XP\_060560975.1  
## 6 XP\_060566031.1  
## Protein.name  
## 1 golgin subfamily A member 3-like isoform X1  
## 2 titin homolog isoform X1  
## 3 FYVE and coiled-coil domain-containing protein 1-like  
## 4 phospholipase D1-like  
## 5 malignant fibrous histiocytoma-amplified sequence 1 homolog isoform X1  
## 6 serine-rich adhesin for platelets-like isoform X1  
## Protein.length Locus.tag  
## 1 1674   
## 2 1350   
## 3 2583   
## 4 330   
## 5 1231   
## 6 2552

# write\_delim(DEG\_Blast, "../output/0826-DEG\_blast\_cds\_full.tab")

the following chunk is me seeing if I could pull the SPID from the full blast output and add it to my annot DEG list. 12 SPID short after join.. not sure why

# read in blast full results  
blastfull <- read.csv("../output/0821-rphil\_blast\_cds.tab", sep="")  
  
# select for just the protein id and the saccver / spid info   
blastselect <- data.frame(protein = blastfull$protein\_id, id = blastfull$saccver)  
  
# join to the deg blast list by matching the protein ids  
blastselect\_deg <- left\_join(DEG\_Blast, blastselect, join\_by("Protein.accession" == "protein") )  
  
# removing any duplicate LOC entries that were created during the join  
blast\_id\_deg <- distinct(blastselect\_deg, LOC, .keep\_all = TRUE)

### following code chunks ignored currently (as of 8/27/25)

### if using blast instead of annotated genome file :

# make a short file of just the DEG names and other desired data  
  
# DEG\_annot <- data.frame( LOC = DEG\_Blast$LOC, geneName = DEG\_Blast$SPID, baseMean = DEG\_Blast$baseMean, log2FoldChange = DEG\_Blast$log2FoldChange, pvalue = DEG\_Blast$pvalue, padj = DEG\_Blast$padj, protein = DEG\_Blast$protein, cds = DEG\_Blast$cds )  
#   
# head(DEG\_annot)

retrieving uniprot IDs

# retrieve gene names for uniprot id lookup  
  
# this is to get just the numbers from the geneName column for uniprot  
# so sp|S8FGV1|LAC... becomes just S8FGV1  
# uniprot\_id <-DEG\_annot$geneName[!is.na(DEG\_annot$geneName)]  
# head(uniprot\_id)  
#   
# # write to table for uniprot import (or could just copy and paste but i imported a text file to uniprot just for my own santity to make sure everything was included)  
# write.table(uniprot\_id, "../output/0821-uniprot\_id\_cds\_genelevel\_8ind23r\_ToC.txt", sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)  
#   
# # and now for all the uniprot ids at the gene level that blast had  
# write.table(blast$SPID, "../output/0821-uniprot\_id\_cds\_genelevel\_8ind23r\_ToC.txt", sep = "\t", row.names = FALSE, col.names = FALSE, quote = FALSE)

GO annotations

# retrieved our unirpot id GO file from the web interface https://www.uniprot.org/  
# importing and unzipping file here  
#   
# gunzip("../data/idmapping\_2024\_05\_14.tsv.gz")  
#   
# GO\_id <- read.csv("../data/idmapping\_2024\_05\_14.tsv", sep = '\t', header = TRUE, row.names=NULL)  
# head(GO\_id)

# join GO id info to our DEG\_annot dataframe from earlier by the uniprot ids  
  
# clam\_GO\_annotations <- merge(DEG\_annot, GO\_id, by.x = "uniprot\_id", by.y = "Entry")  
# head(clam\_GO\_annotations)

# siic it looks all good so let's write to file  
  
# write.csv(clam\_GO\_annotations, "../output/clam\_GO\_annotations\_cds\_0514.csv")