**Starting Point**

Transcriptome data assembled by Trinity. Three females:

1128\_SWS - Super white ZAL2M/ZAL2M

1131\_WS - White ZAL2M/ZAL2

1150\_TS - Tan ZAL2 /ZAL2

There was a potential name swap between 1131 and 1150.

in a readme from Jim: "note that 1128=1128 but 1131=1150 and 1150=1131. ie, some kind of switch occurred, probably when samples run or prepped" I have used the names that should match reality. i.e. 1128=1128, 1131=1131, 1150=1150

The files contain sequences and claim to match them to chromosome, but Dr. Yi says don't trust that information.

For now we are comparing Super White to Tan, i.e. 1128 to 1150.

How many sequences in each file?

grep ">" $1 | wc -l

(Added this as a command in ~/bin/countseqs)

1128\_SWS has 298,236 sequences

1150\_TS has 343,164 sequences

**Overview**

Find reciprocal best matches using blast search both ways across the two data sets. Look for nonsynomous mutations where alignment length is the same as the query and the result length. Blast those results against all species, looking for ones related to behavior.

**Blast**

To use blast, I added it to my biocluster account by adding this line to .pacemodules:

module load ncbi\_blast/2.2.28

To use it, I mostly followed these directions:

<http://www.blaststation.com/freestuff/en/howtoNCBIBlastMac.html>

To make databases for blast use, see makeDB.sh

**First attempt:**

9/18

blastn -query 1128\_SWS.fasta -db db\_1150 -out 1128\_search\_1150 -outfmt 10

blastn -query 1150\_TS.fasta -db db\_1128 -out 1150\_search\_1128 -outfmt 10

Stats:

Memory ~1.2GB  
Time ~5 hours

Results:

1128\_search\_1150\_v1 543MB

1150\_search\_1128\_v1 542MB

Improvements:

Make separate sh files for each search so they can run in parallel.

Include only one result per sequence, since we are just looking to match the best.

Remove some extra info from the output format.

**Second attempt:**

9/19

blastn -query 1128\_SWS.fasta -db db\_1150 -out 1128\_search\_1150\_v2 -outfmt '10 qseqid sseqid pident' -max\_target\_seqs 1

blastn -query 1150\_TS.fasta -db db\_1128 -out 1150\_search\_1128\_v2 -outfmt '10 qseqid sseqid pident' -max\_target\_seqs 1

To submit all jobs in the folder:

ls blast\*.sh | xargs -n1 msub

Stats:

Memory ~1.1GB  
Time ~1 hour

Results:

1128\_search\_1150\_v1 16MB 272,733 lines (out of 298,236)

1150\_search\_1128\_v1 16MB 266,553 lines (out of 343,164)

Further processing:

~~Rerun with more output options, like lengths~~

~~Find reciprocal~~

Filter to make sure alignment length is similar to transcript length

Remove all entries with 100% identity

Sort for lowest % identity to get the top X to look at

For top X, make sure it is protein coding RNA and look for non-synonymous changes

Blast these against other species

9/20

Added length into blast results.

blastn -query 1150\_TS.fasta -db db\_1128 -out 1150\_search\_1128\_v2 -outfmt '10 qseqid sseqid qlen slen length pident' -max\_target\_seqs 1

blastn -query 1128\_SWS.fasta -db db\_1150 -out 1128\_search\_1150\_v2 -outfmt '10 qseqid sseqid qlen slen length pident' -max\_target\_seqs 1

~~Reciprocal matches~~

~~142590 reciprocal matches~~

10/12

I noticed there were duplicate entries in the file of reciprocal matches, so I removed duplicates with:

sort -t "," -u -k1,1 reciprocal > distinctReciprocals

Distinct reciprocals:

138787

Next step is to merge alignments, because there might be more than one alignment between two transcripts. To do that I'll use BRM Parser. <http://www.bioinfo.de/isb/2007/07/0027/>

**Future Plans**

compare outputs, find sequences that match, examine differences

Other analysis:

Thomas talk 9/20

Alternate splicing

Examine the transcripts with no blast result