Process:

1. Map to X. laevis scaffolds using BWA
2. Use samtools to turn mapped data into pileups
3. Compute windowed t-tests every 50 bp based on DEPTH OF COVERAGE to compute mean and sd, but only in regions where both samples have reads above a threshold (>1.0 bp avg read depth in this case)
4. Cut out the actual sequences of these regions and do blast against human RNA (tblastx, e-val <1e-2) and CDD (rpsblast, default e-val cutoff)
5. Do GoMiner analysis of human RNA results
6. Take top hit for each region (only), then go into R and run median polish to summarize both average and stddev (on linear values, not log2)
7. Lookup gene names in human_rna.annot
8. Send table to JW/ME

(very important note: Human blast hits are only E-val<.01 - this may not be stringent enough to "call" a gene from only a 50 bp window.)

Interesting observations

1. Although there are about the same number of total reads from the two samples (42088208 and 46035732), and about the same number of total reads mapped (35886490 and 37215270 for cnt/mut), the number of bp hit in the genome is radically different: 63.4 MB for cnt, 97.4 MB for mut. So mut is cranking out more different transcripts. This observation is a natural consequence of this method - normal RNA-seq analysis methods don’t ask this question (though they do ask how many genes are observed, which is roughly equivalent when you have an annotated genome).

Some validation between my list and Jon’s:

These in my table:

<table>
<thead>
<tr>
<th>Name</th>
<th>cntmea n</th>
<th>mutmea n</th>
<th>cnts d</th>
<th>muts d</th>
<th>log2 fold change</th>
<th>T statistic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALG11</td>
<td>1.65</td>
<td>10.76</td>
<td>0.48</td>
<td>1.60</td>
<td>2.70</td>
<td>16.14</td>
<td>Homo sapiens asparagine-linked glycosylation 11 homolog (S. cerevisiae, alpha-1,2-mannosyltransferase) (ALG11), mRNA</td>
</tr>
<tr>
<td>CKB</td>
<td>13.10</td>
<td>4.84</td>
<td>4.23</td>
<td>1.36</td>
<td>1.44</td>
<td>-3.24</td>
<td>Homo sapiens creatine kinase, brain (CKB), mRNA</td>
</tr>
</tbody>
</table>

Correspond to these in Jon’s table (first page):

<table>
<thead>
<tr>
<th>geneID</th>
<th>Ctrl_F3</th>
<th>Ctrl_F5</th>
<th>RFX2KD_F3</th>
<th>RFX2KD_F5</th>
<th>Max Cov</th>
<th>log2 Fold Change</th>
<th>Adj P-value</th>
<th>z-score</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>alg11NM_001122888</td>
<td>89</td>
<td>147</td>
<td>200</td>
<td>259</td>
<td>7.47</td>
<td>1.09</td>
<td>2.96E-08</td>
<td>-8.18</td>
<td>asparagine-linked glycosylation 11, alpha-1,2-mannosyltransferase homolog (alg11), mRNA</td>
</tr>
<tr>
<td>ckbNM_001086894</td>
<td>1265</td>
<td>1184</td>
<td>227</td>
<td>196</td>
<td>26.99</td>
<td>-2.403</td>
<td>1.34E-60</td>
<td>38.38</td>
<td>creatine kinase, brain (ckb), mRNA</td>
</tr>
</tbody>
</table>

So at least direction of change agrees...

Same results in IGV

Here are these regions from IGV, showing mut & control pileups and the BLAST results for both the human and X laevis mRNA’s to the 20101210 scaffolds:

ALG11:
CKB:

Note that for both human and x laevis, CKB spans many scaffolds; I’m only showing two that have great homology btwm human & laevis. Bp 544 to 841 of the laevis version are on scaffold 463681 while bp 842 to 1405 are on scaffold 432496: