RNA-Seq hypothesis testing

Brian J. Knaus
USDA Forest Service
Pacific Northwest Research Station
Hypothesis Testing

Software:
- EdgerR (Bioconductor)
- NBPSeg (CRAN)
- Genecounter (in prep.)
- Cufflinks (with bowtie and tophat)
- CLC Bio
Coos Bay
P.m. var. menziesii

Sample#1
6,036,336
reads

Sample#2
8,029,643
reads

Yakima
P.m. var. glauca

Sample#1
4,900,955
reads

Sample#2
4,580,341
reads

Experimental design
* Each sample is a pool of six seedlings grown in a common environment.
<table>
<thead>
<tr>
<th>Gene</th>
<th>cb_a</th>
<th>cb_b</th>
<th>yk_a</th>
<th>yk_b</th>
</tr>
</thead>
<tbody>
<tr>
<td>isotig18613_gene=isogroup07808_length=677_numContigs=1</td>
<td>17</td>
<td>18</td>
<td>139</td>
<td>159</td>
</tr>
<tr>
<td>isotig01880_gene=isogroup00225_length=652_numContigs=4</td>
<td>11</td>
<td>10</td>
<td>162</td>
<td>56</td>
</tr>
<tr>
<td>isotig07160_gene=isogroup01638_length=3698_numContigs=4</td>
<td>31</td>
<td>81</td>
<td>276</td>
<td>226</td>
</tr>
<tr>
<td>isotig06362_gene=isogroup01321_length=1396_numContigs=4</td>
<td>32</td>
<td>31</td>
<td>149</td>
<td>91</td>
</tr>
<tr>
<td>isotig06005_gene=isogroup01197_length=1204_numContigs=4</td>
<td>52</td>
<td>68</td>
<td>169</td>
<td>198</td>
</tr>
<tr>
<td>isotig06363_gene=isogroup01321_length=1470_numContigs=4</td>
<td>21</td>
<td>27</td>
<td>73</td>
<td>100</td>
</tr>
<tr>
<td>contig29123_gene=isogroup00629_length=686</td>
<td>30</td>
<td>15</td>
<td>75</td>
<td>161</td>
</tr>
<tr>
<td>isotig30058_gene=isogroup19254_length=1101_numContigs=1</td>
<td>31</td>
<td>36</td>
<td>75</td>
<td>400</td>
</tr>
<tr>
<td>contig50604_gene=isogroup01657_length=1247</td>
<td>272</td>
<td>405</td>
<td>1153</td>
<td>724</td>
</tr>
<tr>
<td>contig21101_gene=isogroup01657_length=559</td>
<td>47</td>
<td>96</td>
<td>264</td>
<td>165</td>
</tr>
<tr>
<td>isotig05419_gene=isogroup01011_length=1938_numContigs=4</td>
<td>32</td>
<td>49</td>
<td>103</td>
<td>126</td>
</tr>
<tr>
<td>contig03433_gene=isogroup00629_length=496</td>
<td>21</td>
<td>10</td>
<td>55</td>
<td>71</td>
</tr>
<tr>
<td>isotig05877_gene=isogroup01156_length=2570_numContigs=4</td>
<td>91</td>
<td>70</td>
<td>154</td>
<td>762</td>
</tr>
<tr>
<td>isotig04630_gene=isogroup00778_length=2155_numContigs=4</td>
<td>44</td>
<td>14</td>
<td>214</td>
<td>97</td>
</tr>
<tr>
<td>isotig29494_gene=isogroup18690_length=2738_numContigs=1</td>
<td>433</td>
<td>511</td>
<td>48</td>
<td>93</td>
</tr>
<tr>
<td>contig46195_gene=isogroup00001_length=493</td>
<td>10</td>
<td>17</td>
<td>17</td>
<td>260</td>
</tr>
<tr>
<td>contig48481_gene=isogroup00006_length=378</td>
<td>33</td>
<td>16</td>
<td>54</td>
<td>274</td>
</tr>
<tr>
<td>isotig30667_gene=isogroup19863_length=1653_numContigs=1</td>
<td>160</td>
<td>74</td>
<td>361</td>
<td>517</td>
</tr>
<tr>
<td>isotig33299_gene=isogroup22495_length=1270_numContigs=1</td>
<td>13</td>
<td>14</td>
<td>21</td>
<td>194</td>
</tr>
<tr>
<td>isotig24557_gene=isogroup13753_length=1317_numContigs=1</td>
<td>442</td>
<td>547</td>
<td>111</td>
<td>65</td>
</tr>
</tbody>
</table>
Hypothesis Testing

Requirements:
• Table of read counts

Model:
• $Y \sim \text{Negbinom}(\mu, \varphi)$

Issues:
• Normalization for unequal library size
• Estimate of dispersion parameter ($\varphi$)
• Estimate of rate parameter ($\mu$)
• Test for differential expression (exact test)
• Correct for multiple comparisons
Normalization

Comparison of libraries assumes equal library size.

Proposed methods:

- **RPKM** – reads per thousand bases of exon per million mapped reads.
- **FPKM** – fragments per thousand bases of exon per million mapped reads.
- Quantile normalization.
- Upper quantile normalization.
- Thinning to the minimum library size.
'arab' dataset

![Bar chart showing categories mock1, mock2, mock3, hrcc1, hrcc2, hrcc3 with varying heights.](image-url)
Dispersion estimation

Model:
- \( Y \sim \text{Negbinom}(\mu, \phi) \)

Phi models biological variability among samples.

Proposed methods:
- Common dispersion among all genes.
- Moderated tagwise dispersion.
- NBP
$Y \sim \text{Pois}(\lambda)$


http://www.bepress.com/sagmb/vol10/iss1/art24/
Multiple comparisons

In RNA-Seq a hypothesis test is made for every gene/transcript:

• Tens of thousands of hypothesis
• Great big fishing trip, bound to catch something, but what?

Corrections for multiple comparisons:

• Bonferroni, type I & II error.
• Sequential Bonferroni, type I & II error.
• False discovery rate, type I error.
FC plot using tagwise dispersion

Yakima
210 Significant
235 Non-significant

Coos Bay
To the computer...