A Gene Recommender for *C. elegans*

Art B. Owen  
Department of Statistics  
Stanford University  
owen@stat.stanford.edu

With:  
Josh Stuart, Kathy Mach,  
Anne Villeneuve, Stuart Kim

http://pmgm2.stanford.edu/~kimlab/cassettes
A specific problem

These genes are involved in the Retinoblastoma complex in C. elegans:

lin-9  lin-35  lin-36  lin-53  hda-1

Questions

1. Are there more?

2. If so, which ones?

3. How to find them with expression data?

Other groups

41 Major Sperm Protein (MSP) genes
6 Synaptonemal Complex genes
6 Meiotic Repair genes

Art Owen, Stanford University
Rb genes in topomap

Kim et al. (Science 2001)

From: http://cmgm.stanford.edu/~kimlab/topomap/

Number of genes not plotted from the input list: 0
Recommenders

For movies

1. Start with a list of movies
2. Find viewers who rated them highly
3. Find other movies those viewers liked

For genes

1. Start with a list of genes
2. Find experiments where they’re co-expressed
3. Find other genes with similar profiles in those expts

Similar and independent: Ihmels et al. (Nature Genetics, 2002) for Yeast

Art Owen, Stanford University
Experiment list

553 experiments from

- Eggs, larvae, dauer, adult
- Heat shock and other stresses
- Mutants
- Various labs

Some experiments irrelevant to Rb, or MSP, or repair
They’ll add noise
Recommender approach uses selected expts only

Art Owen, Stanford University
Two experiments

Normalized expression for expt 287

Normalized expression for expt 11

Art Owen, Stanford University
Expression data

Genes $i = 1, \ldots, n = 19738$

Experiments $j = 1, \ldots, p = 553$

Expression for each gene ranked 1 to 553

Ranks were then scaled from $-1$ to 1

Reduces effects of outliers

(Formulas in article)
Experiment scores

Expt \( j \) scores high if:

- Rb genes **cluster** in expt \( j \)
- cluster is at an **extreme** value

\[ Z_j = \text{Mean/Std-Dev of Rb values, in Expt } j \]

**Examples**

Expt 287 scored 97.8
Expt 11 scored 0.41

\[ Z_j \approx N(0, 1) \text{ for perfectly irrelevant expt} \]
Experiment scores

Signed expt scores for Rb

Signed expt scores for MSP

Art Owen, Stanford University
Gene scores

Gene $i$ scores high if:

- it runs **parallel** to average Rb
- in **high scoring** experiments $E_{Rb}$

**Gene score**

$$S_i \propto \sum_{j \in E_{Rb}} X_{ij} \bar{X}_{Rb,j}$$

Large $S_i$ $\implies$ higher “Rb-ness”
Some gene scores

Normal gene scores for Rb

Normal gene scores for MSP

Art Owen, Stanford University
### Rb query results

<table>
<thead>
<tr>
<th>Gene/Protein</th>
<th>Score</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>dpl-1</td>
<td>0.247</td>
<td>E2F (fly)</td>
</tr>
<tr>
<td>li-53</td>
<td>0.244</td>
<td>Rb (human)</td>
</tr>
<tr>
<td>K12D12.1</td>
<td>0.240</td>
<td>Topoisomerase II</td>
</tr>
<tr>
<td>li-35</td>
<td>0.238</td>
<td>Rb (human)</td>
</tr>
<tr>
<td>Ce Bub1</td>
<td>0.237</td>
<td>Bup1p (yeast)</td>
</tr>
<tr>
<td>hda-1</td>
<td>0.237</td>
<td>Histone deacetylase</td>
</tr>
<tr>
<td>B0464.6</td>
<td>0.233</td>
<td>Unknown</td>
</tr>
<tr>
<td>R06F6.1</td>
<td>0.233</td>
<td>Histone hairpin (human)</td>
</tr>
<tr>
<td>T16G12.5</td>
<td>0.231</td>
<td>Unknown</td>
</tr>
<tr>
<td>F55A3.7</td>
<td>0.230</td>
<td>Spt16p (yeast)/DRE 4 (fly)</td>
</tr>
<tr>
<td>plk-1</td>
<td>0.229</td>
<td>Polo kinase</td>
</tr>
<tr>
<td>li-9</td>
<td>0.227</td>
<td>synMuv</td>
</tr>
<tr>
<td>li-36</td>
<td>0.227</td>
<td>synMuv</td>
</tr>
<tr>
<td>smc-4</td>
<td>0.227</td>
<td>In SMC family</td>
</tr>
</tbody>
</table>

**Cell Cycle/Chromatin Known Rb interaction**
How many experiments?

We use the $k$ best experiments

Small $k \rightarrow$ not enough to correlate

Large $k \rightarrow$ use irrelevant expts

We minimize # non-Rb genes beating at least half of Rb genes

This is circular, but . . .

ranks changed little using leave-one-out methods
Recommender more precise than Topomap

Group sizes at 50% recall

<table>
<thead>
<tr>
<th>Query</th>
<th>Size</th>
<th>Reco</th>
<th>Topo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retinoblastoma</td>
<td>5</td>
<td>6</td>
<td>138</td>
</tr>
<tr>
<td>Recombination/repair</td>
<td>6</td>
<td>57</td>
<td>1271</td>
</tr>
<tr>
<td>Synaptonemal</td>
<td>6</td>
<td>4</td>
<td>246</td>
</tr>
<tr>
<td>MSP</td>
<td>43</td>
<td>32</td>
<td>225</td>
</tr>
</tbody>
</table>

The Rb query has 5 genes.

In recommender list: 3rd Rb gene placed 6th

For topomap: 3rd Rb gene placed 138th

Similar improvements at other recall levels.

Art Owen, Stanford University
Interpretation

We get a list of candidates

High rank does not prove group membership:

1. We might already have the whole group

2. Strong expression correlations can arise for other reasons

Similar caveats hold for document retrieval or recommenders
RNA Interference

- Tried top 50 ranked genes
- wrm-1 embryonic lethality, suppressed by loss of lin-35
- JC8.6 had a synMuv phenotype
JC8.6  $\rightarrow\rightarrow$  SynMuv
Try it!

Enter ORFs at:

http://pmgm2.stanford.edu/~kimlab/cassettes

Email queries to:

owen@stat.stanford.edu