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

Genome Research, 2003

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. Can we find them from expression data on 19738 genes, 500+ experiments?

Other groups

41 Major Sperm Protein (MSP) genes

6 Synaptonemal Complex genes

6 Meiotic Repair genes

C. elegans Topomap

VxInsight's variant of multidimensional scaling

Widely adopted by C. elegans community following Kim et al (Science 2001).

Each gene is a point in the plane

High correlation $\Longleftrightarrow$ small distance

Find center of Rb genes

Investigate non-Rb genes near that center

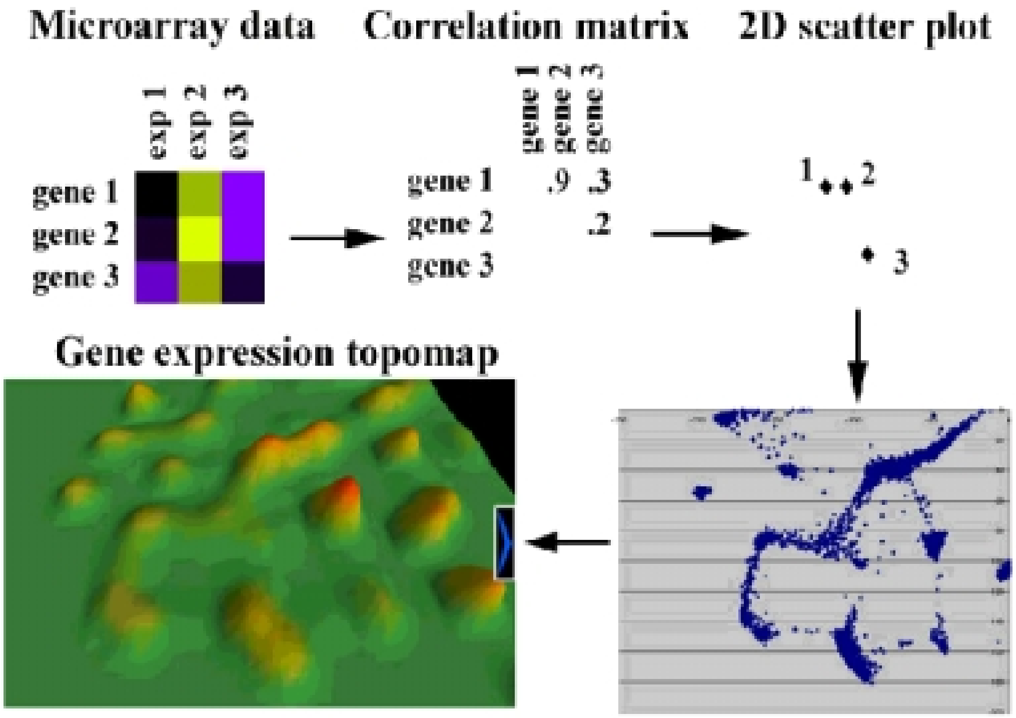
Gene recommender

Build a "cluster" around a seed group of known genes.

Mimic algorithms used to recommend movies/books

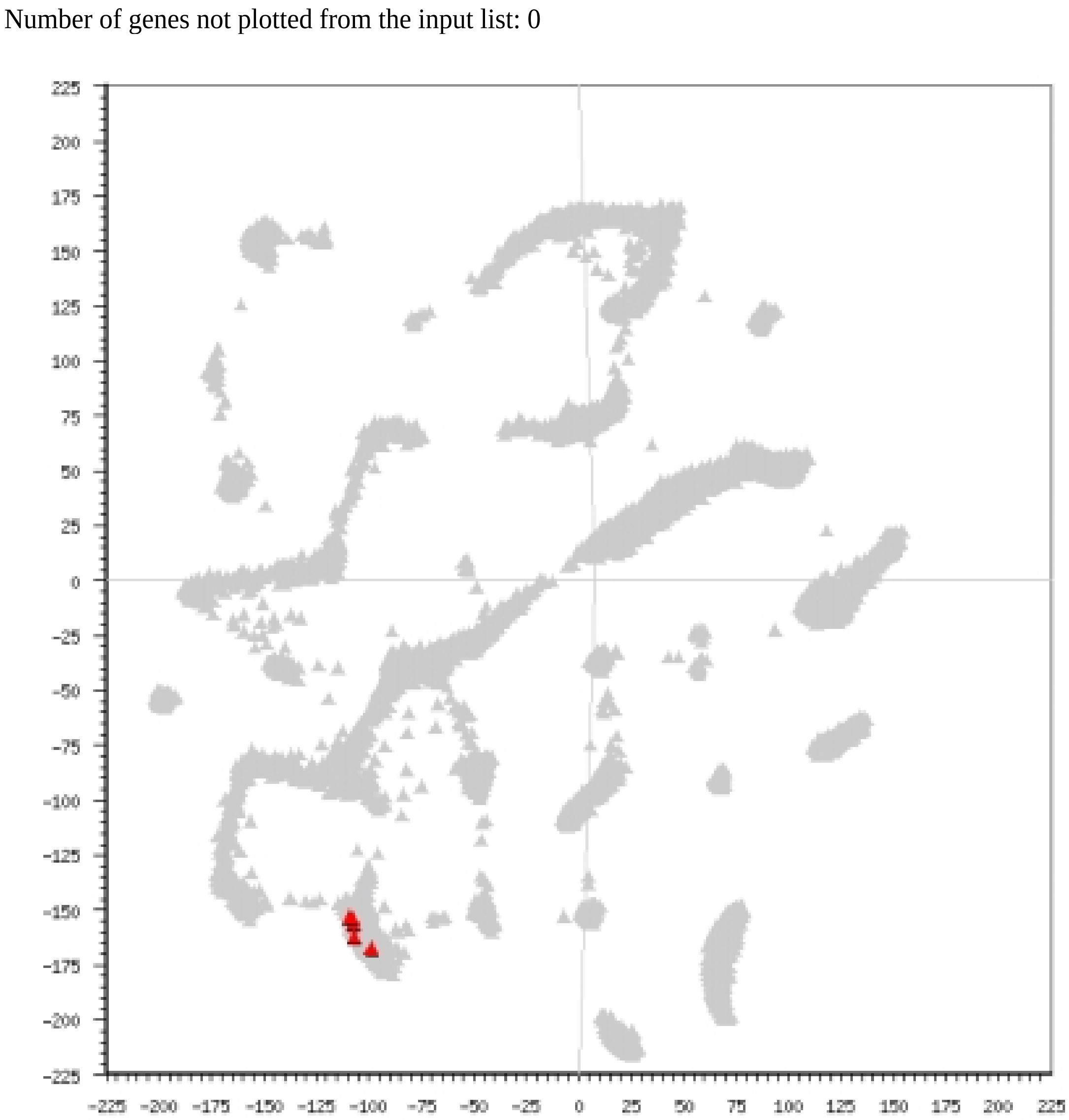
C. elegans topomap

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



5 Retinoblastoma genes

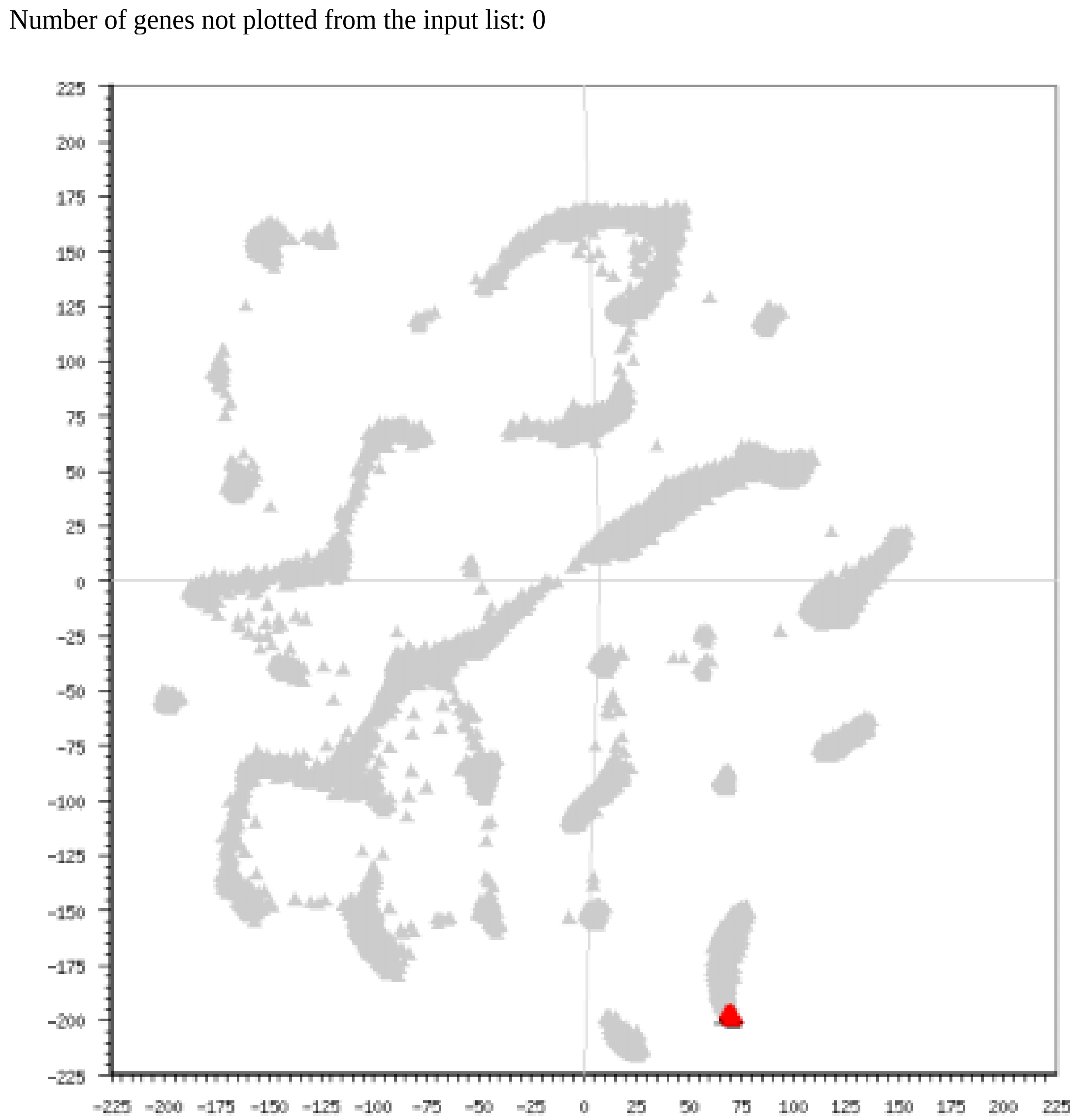
From: http://cmgm.stanford.edu/∼kimlab/topomap/



Number of genes not plotted from the input list: 0

43 MSP genes

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, Friedlander, Bergmann, Sarig, Ziv, Barkai. (Nature Genetics, 2002)

Google set labs.google.com/sets

Query: Larry, Moe, Curly

Result: Moe, Curly, Larry, Shemp, Joe

Query: John, Paul, George

Result: Paul, George, John, Ringo

Query: Donut, Bagel, Washer, Inner Tube

Result: Donut, Bagel, Bagel with Cream Cheese, Egg ala carte, Peaches, . . . , Side Pancakes

Omits Washer and Inner Tube

Query: stop, yield, one way

Result: stop, yield, one way, SLOW, VIEWPOINT AHEAD, Obey your thirst, WALK

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

Result: Check your spelling · · ·

Experiment list

553 experiments from

Eggs, larvae, dauer, adult

Heat shock and other stresses

Mutants

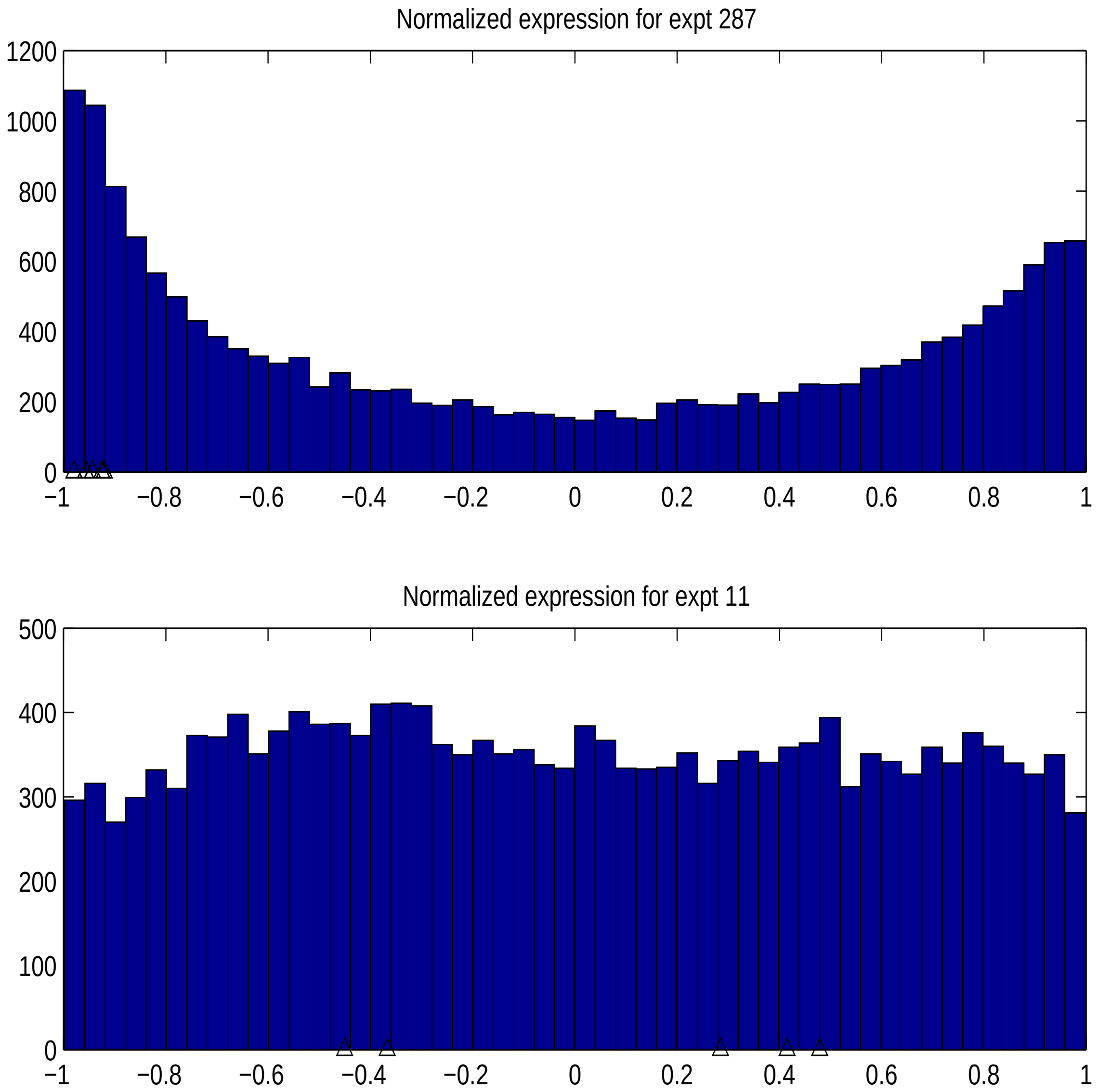
Various labs

Some experiments irrelevant to Rb, or MSP, or repair

They’ll add noise (or undesired signal)

Recommender approach uses selected expts only

Rb data for two experiments



Normalized expression for expt 287

Normalized expression for expt 11

Expression data

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

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

Raw expression data W_{ij}

Normalized expression X_{ij}

X_{ij} are ranks 1 to p for gene i scaled to $[-1, 1]$.

$$X_{ij} = 2\frac{\#\{j' \mid W_{ij'} \le W_{ij}\}}{553} - 1$$

Experiment scores

$$Z_{\text{Rb},j} = \frac{|\bar{X}_{\text{Rb},j}|}{\sqrt{V_{\text{Rb},j}}} \times \sqrt{N_{\text{Rb},j}}$$

where

$\bar{X}_{\text{Rb},j}$ = Avg X for Rb genes in expt j

$V_{\text{Rb},j}$ = Variance(X) for Rb genes in expt j

$N_{\text{Rb},j}$ = # nonmissing Rb vals for expt j

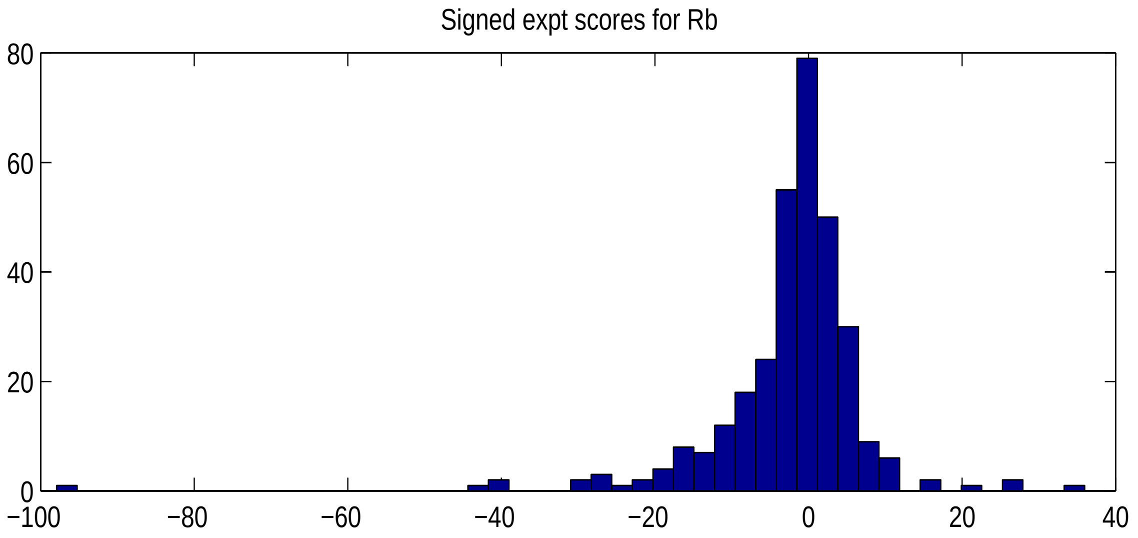
Examples

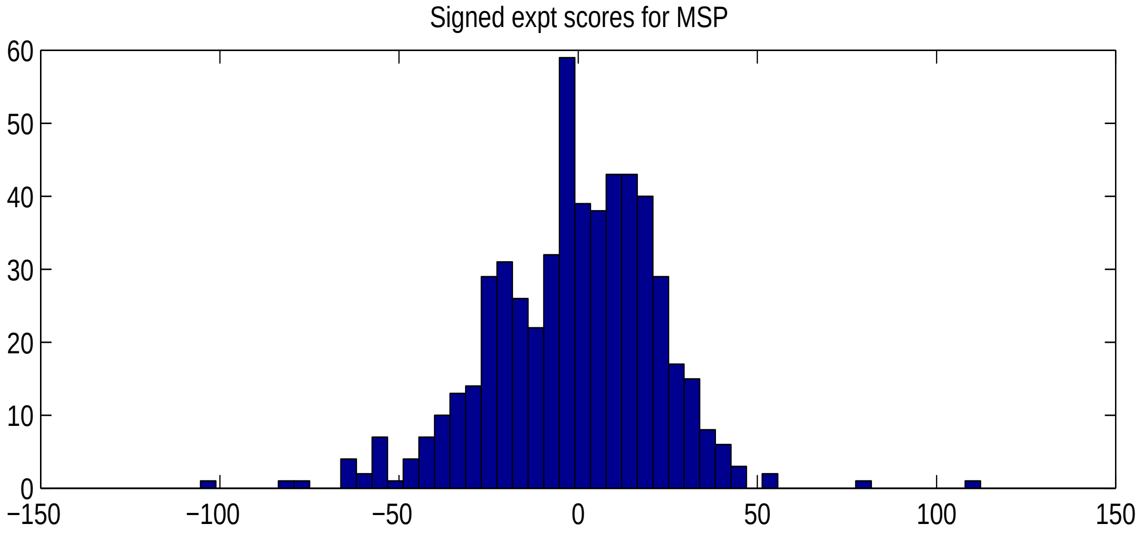
Expt 287 scored 97.8

Expt 11 scored 0.41

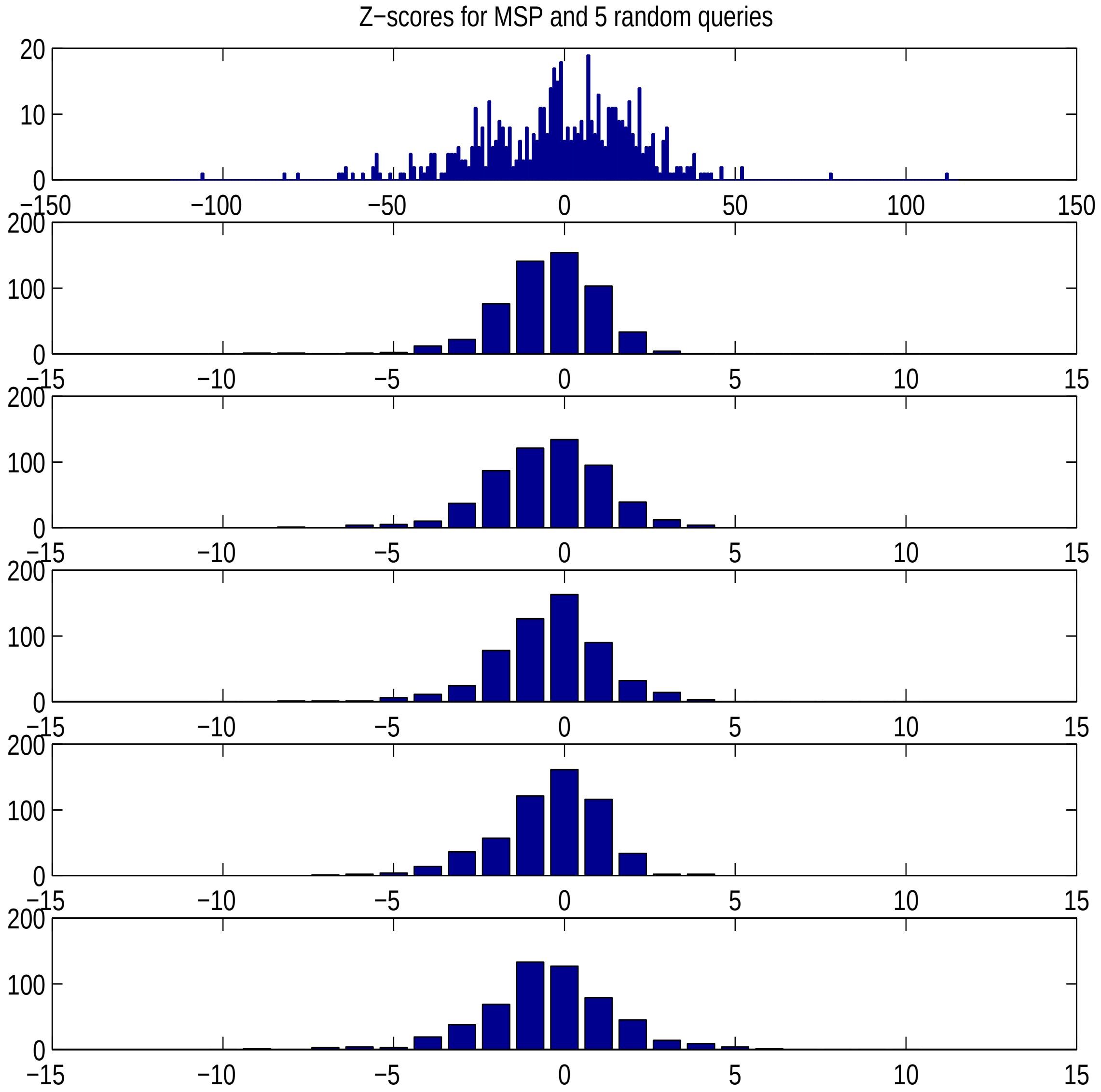
Roughly: $Z \sim |N(0,1)|$ if “nothing going on”

Signed experiment scores

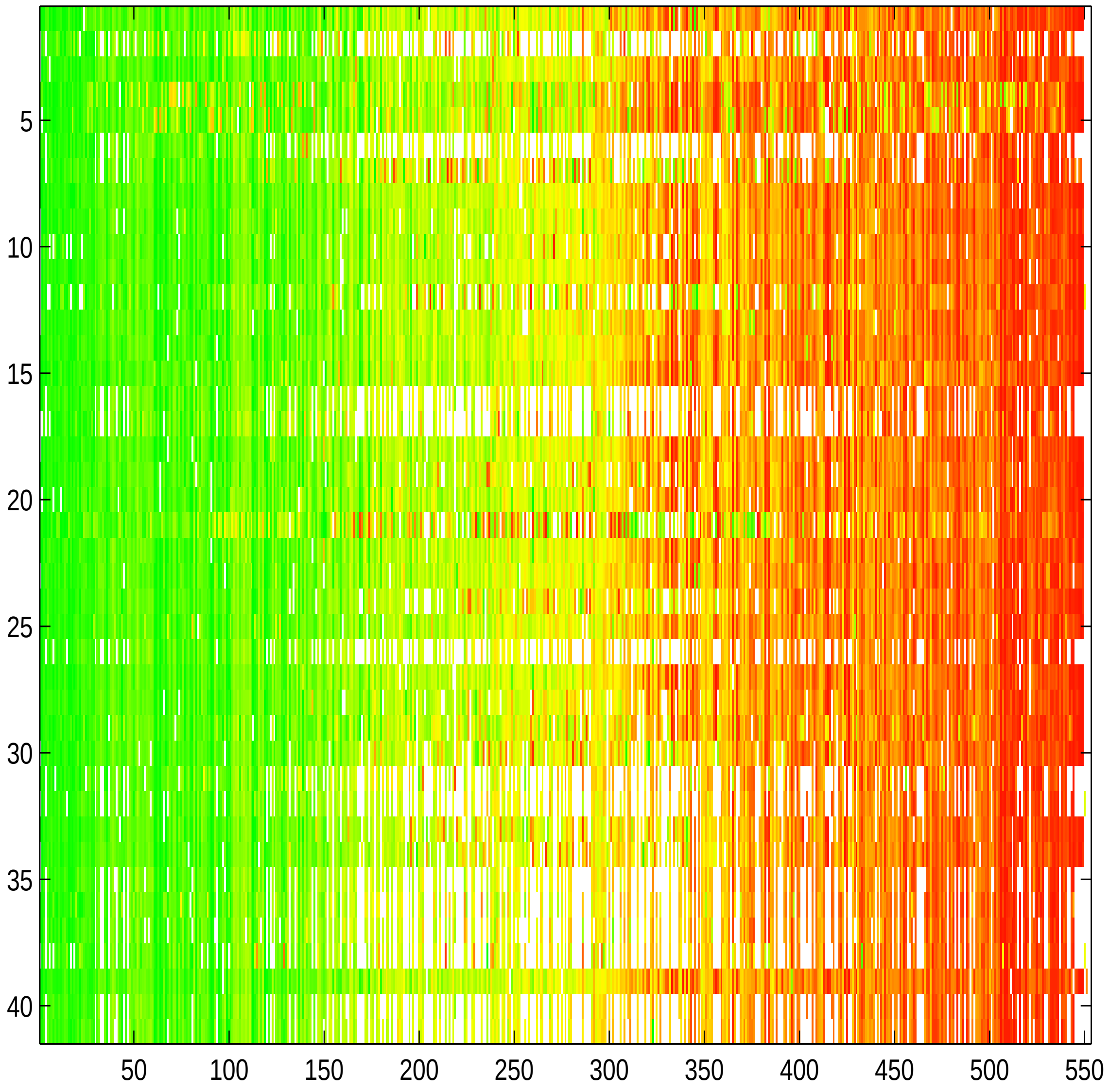




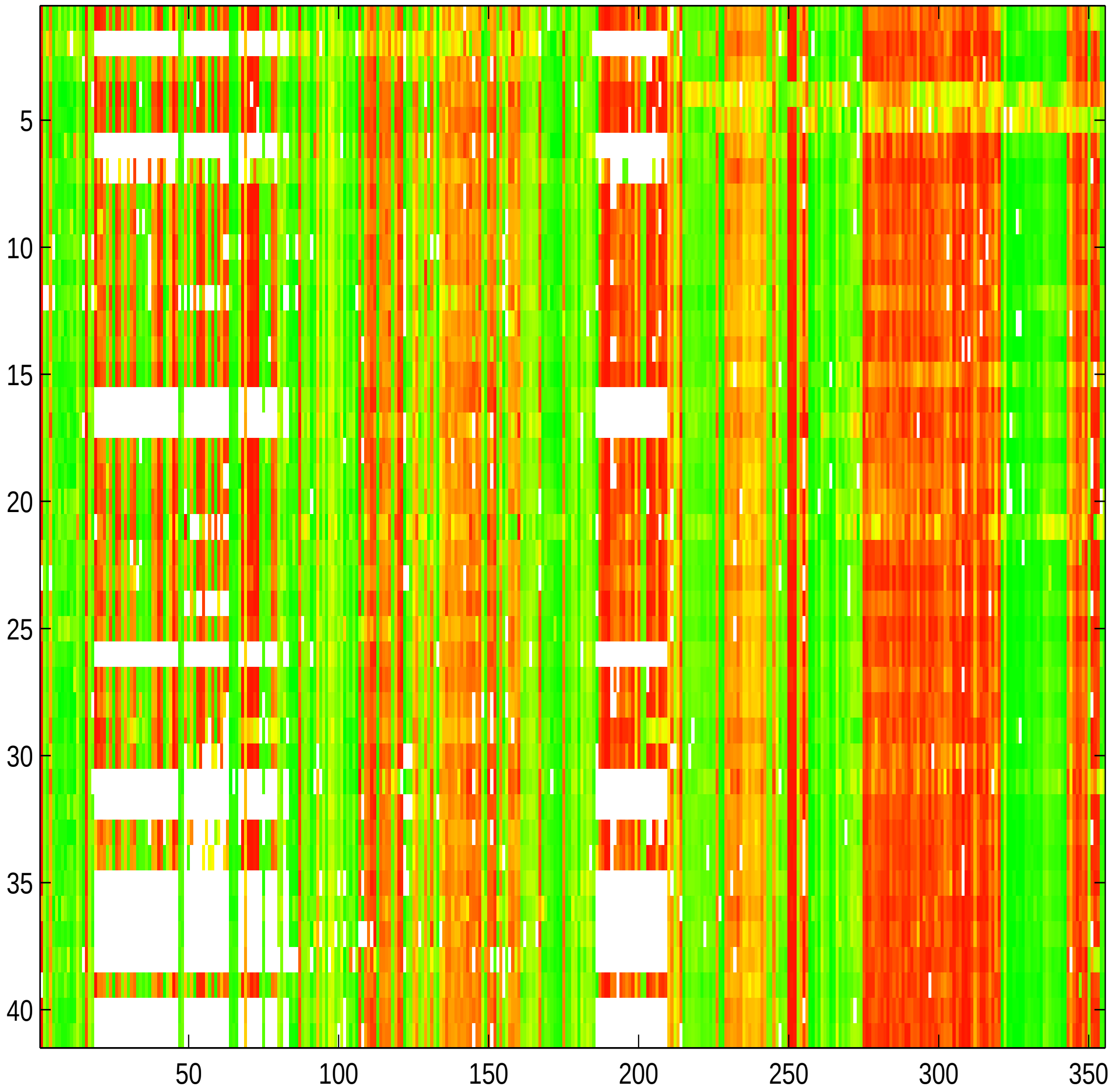
Rb and random 5 gene queries



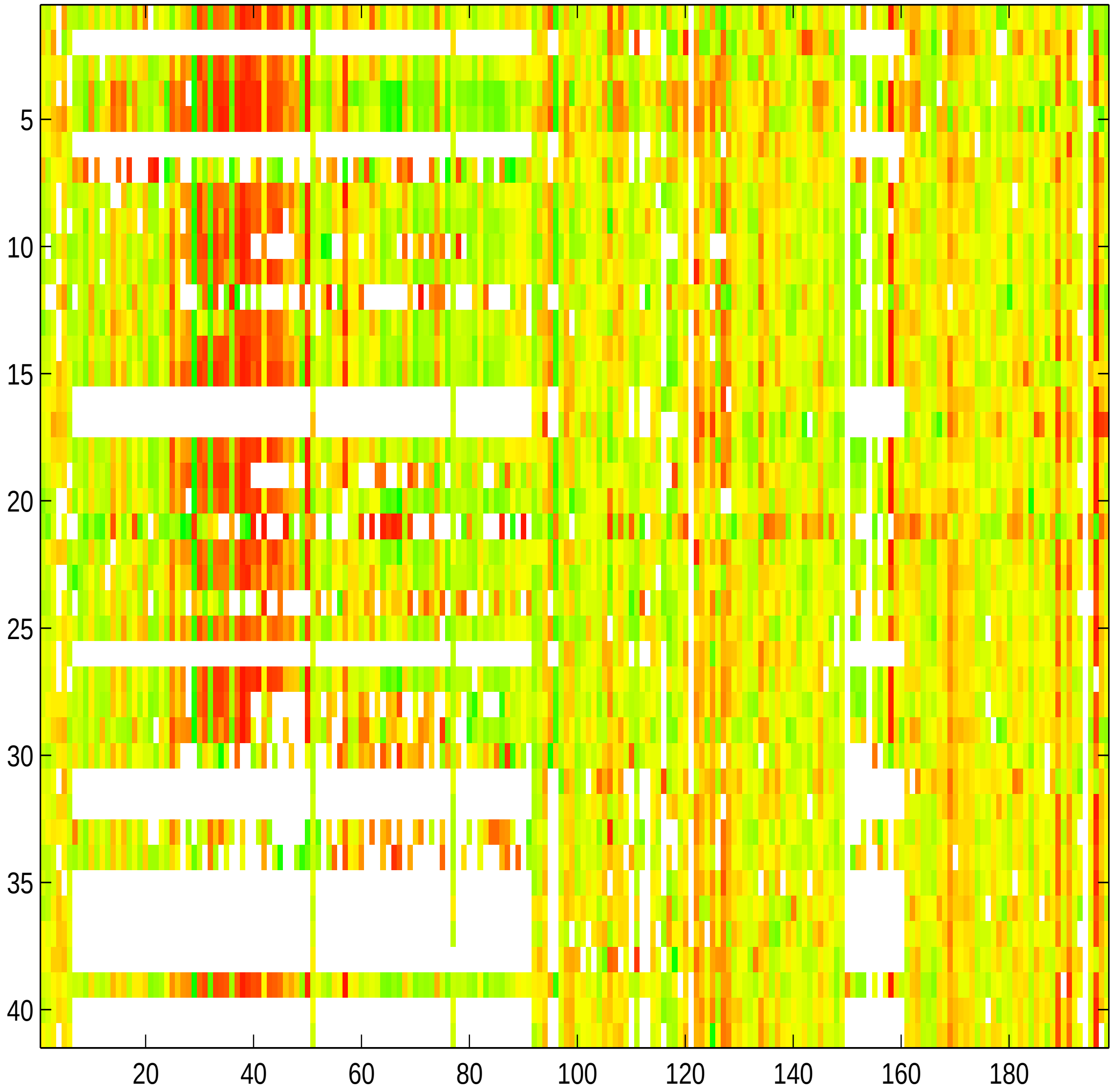
MSP genes, experiments ordered by MSP score



MSP genes, expts with high abs score



MSP genes, expts with low abs score



Gene scores

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

Notes:

Sum only includes high scoring expts E_{Rb}

Score translates into a correlation with Rb in E_{Rb}

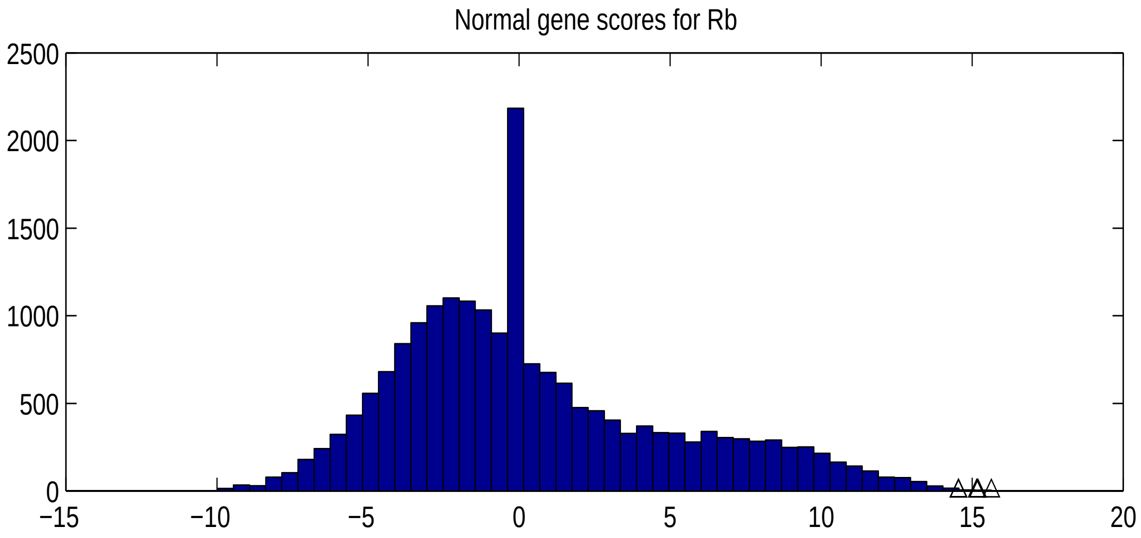
Score translates into a distance from Rb in E_{Rb}

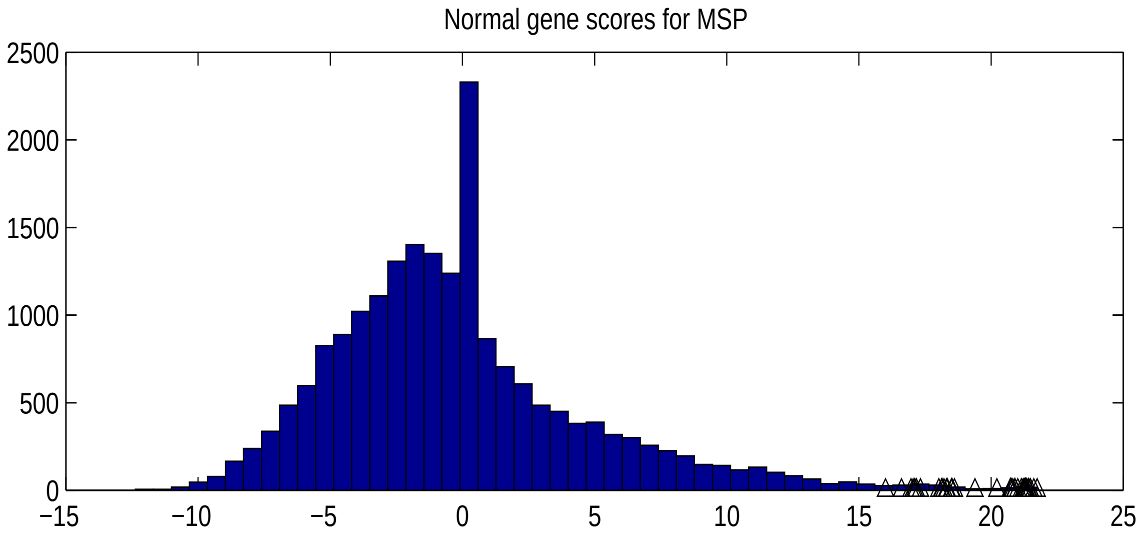
Usage

We rank genes by “Rb-ness”

Explore candidates

Some gene scores





ORF	Score	Z	#E
○ T23G7.1	0.247	10.533	162
● K07A1.12	0.244	15.632	320
○ K12D12.1	0.240	15.370	320
● C32F10.2	0.238	15.188	320
○ R06C7.8	0.237	15.048	317
● C53A5.3	0.237	15.152	320
○ B0464.6	0.233	14.720	315
○ R06F6.1	0.233	14.825	319
○ T16G12.5	0.231	14.691	315
○ F55A3.7	0.230	14.690	318
○ K06H7.1	0.229	14.558	318
● ZK637.7	0.227	14.546	320
● F44B9.6	0.227	14.543	320
○ F35G12.8	0.227	14.519	317

Picking the threshold

We use experiments j with $Z_{\text{Rb}j} > Z^*$

Z^* too large $\longrightarrow$ too few expts to score genes

Z^* too small $\longrightarrow$ include noisy expts

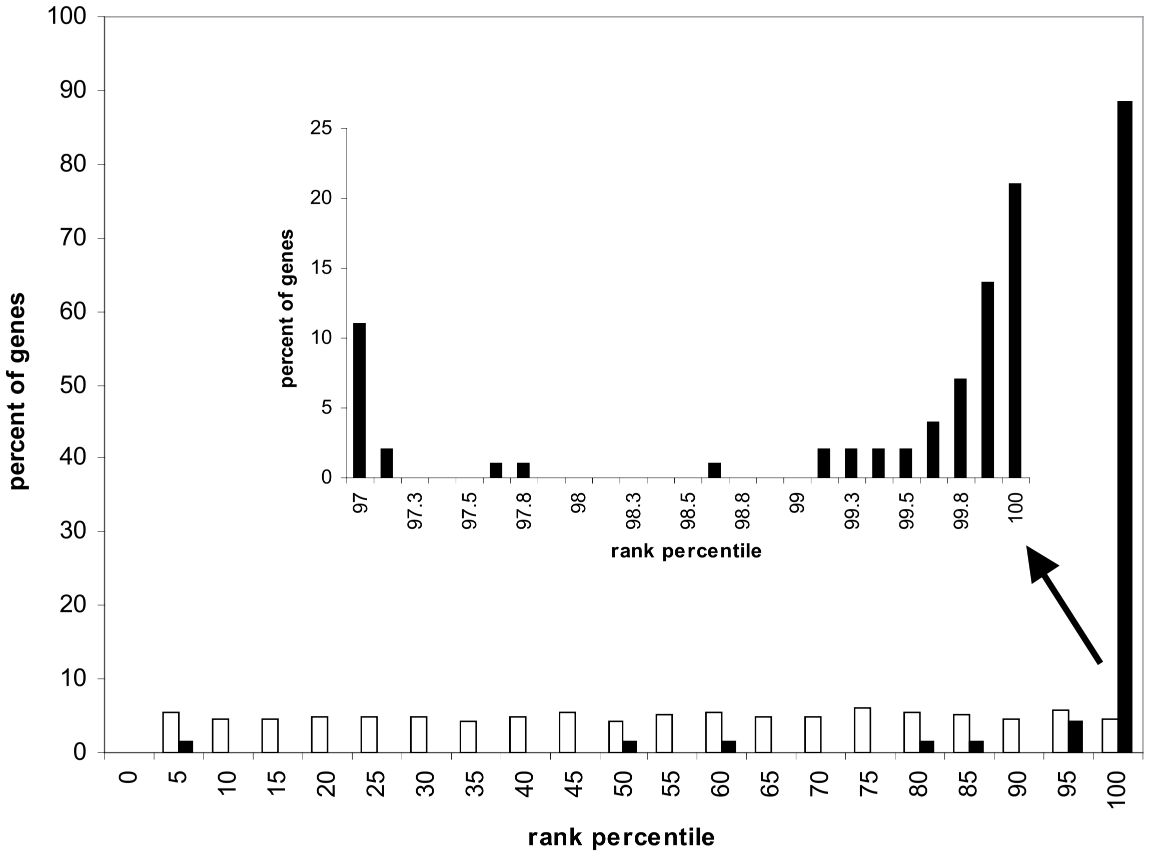
We pick Z^* to bring query group to top of list

Minimize # non-group genes scoring better than half of gp genes

This is circular, but . . .

1. ranks changed little using leave-one-out methods
2. queries with 5 or more random genes did not have high scoring experiments

Leave-one-out results



Recommender more precise than Topomap

Group sizes at 50% recall

Query	Size	Reco	Topo
Retinoblastoma	5	6	138
Recombination/repair	6	57	1271
Synaptonemal	6	4	246
MSP	43	32	225

The Rb query has 5 genes.

To get at least 3 Rb query genes, requires the top 6 genes in the gene recommender ordering.

For the topomap ordering it takes the top 138 genes.

Similar improvements at other recall levels.

Is topomap comparison fair?

Yes and No

Topomap had no free params

Topomap is not optimized per query

. . . because it takes $O(n^2)$ work

Topomap was de facto standard

Interpreting output

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

No p-value can confirm relevance for genes

. . . or movies or web docs

Confirmation

1. Literature
   - Found known MSP genes not in our list
   - Top 2 Rb candidates: dpl-1 & K12D12.1 similar to mammalian genes that interact with Rb
   - Four new genes involved in cell cycle and chromatin regulation; shared function with Rb
2. RNAi knockouts
   - Tried top 50 ranked genes
   - wrm-1 embryonic lethality, supressed by loss of lin-35
   - JC8.6 had a synMuv phenotype

Try it!

Enter ORFs at

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