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

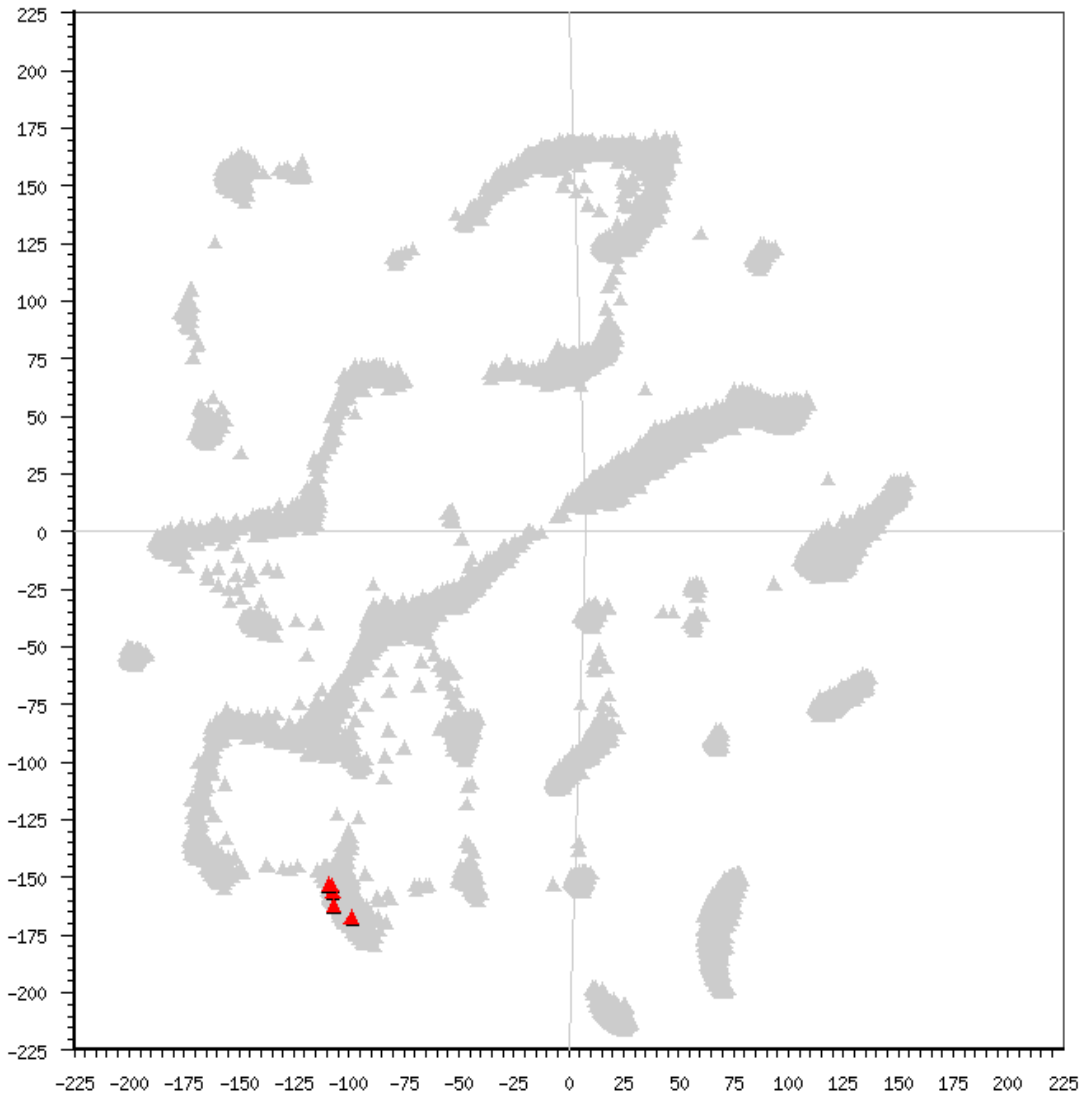
6 Meiotic Repair genes

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

Experiment list

553 experiments from

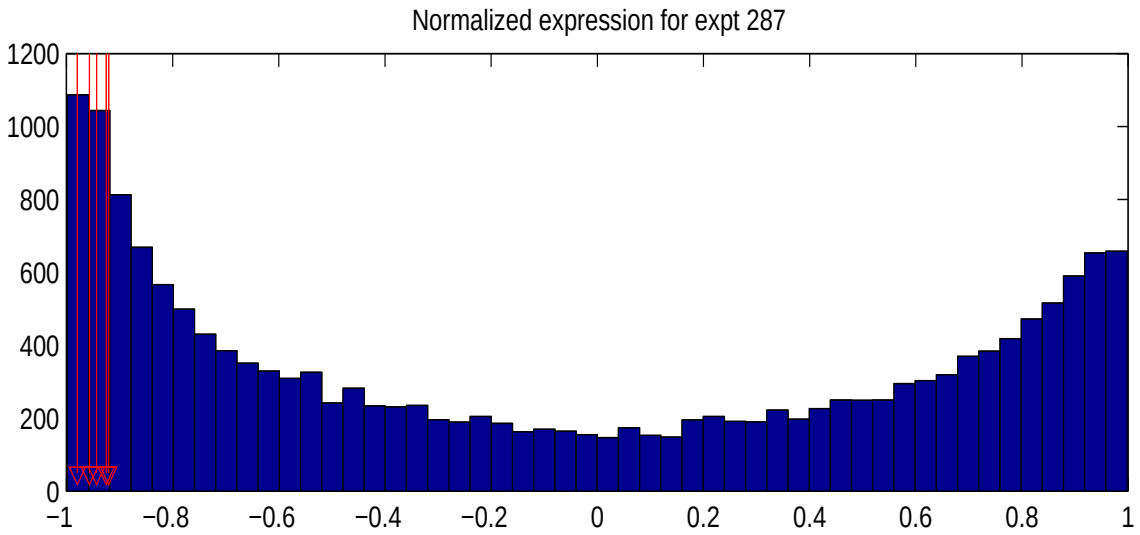
- Eggs, larvae, dauer, adult
- Heat shock and other stresses
- Mutants
- Various labs

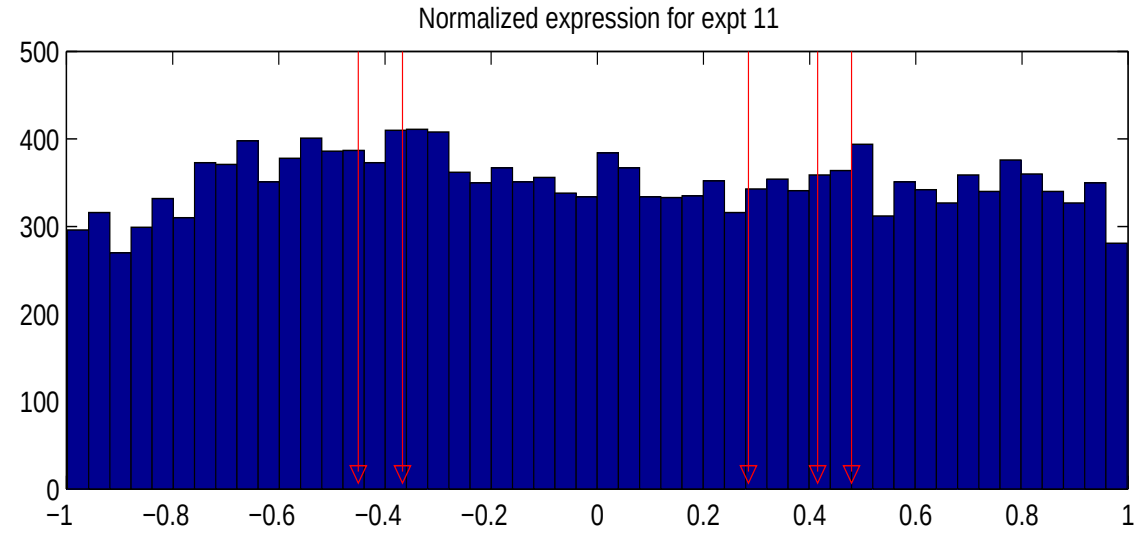
Some experiments irrelevant to Rb, or MSP, or repair

They'll add noise

Recommender approach uses selected expts only

Two experiments





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

$Z_j =$ Mean/Std-Dev of Rb values, in Expt j

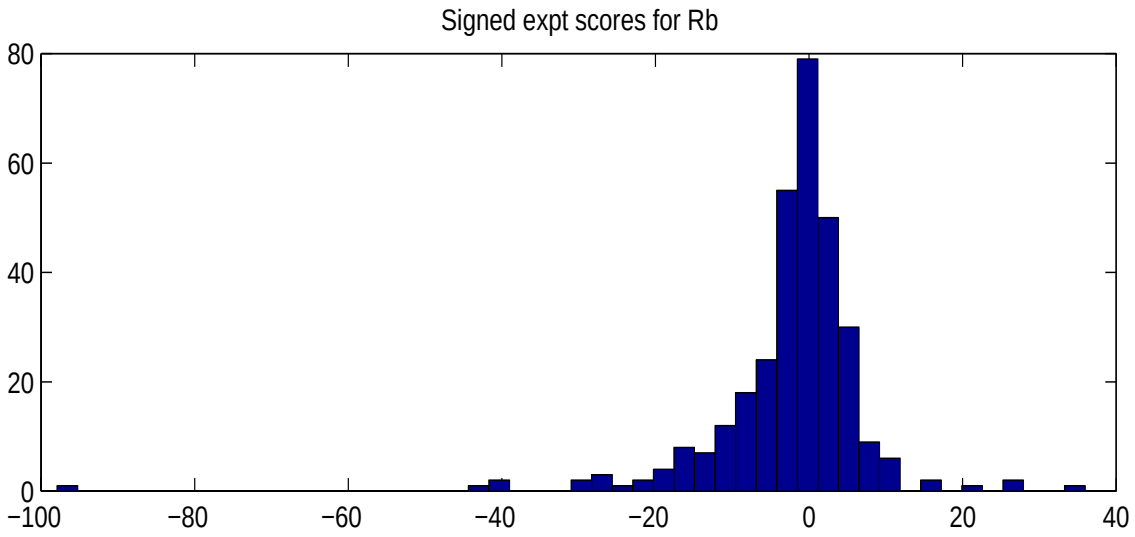
Examples

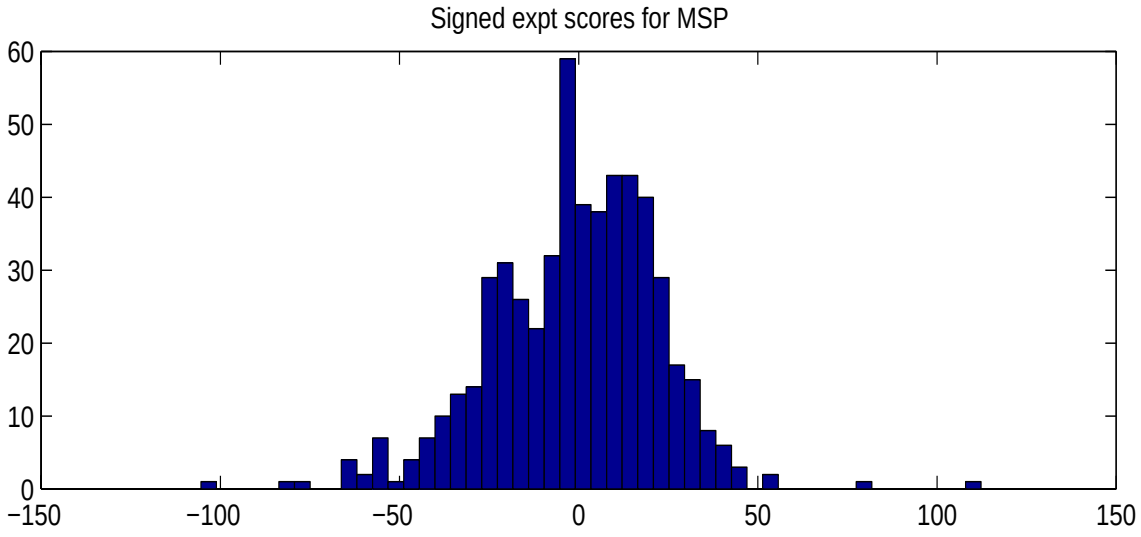
Expt 287 scored 97.8

Expt 11 scored 0.41

$Z_j \approx N(0,1)$ **for perfectly irrelevant expt**

Experiment scores





Gene scores

Gene i scores high if:

it runs **parallel** to average Rb

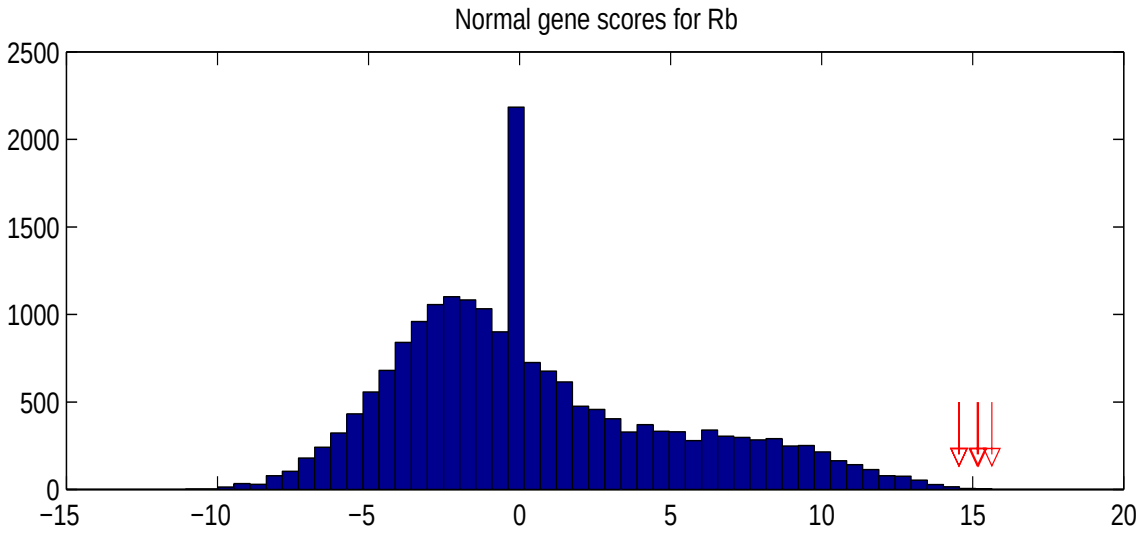
in **high scoring** experiments E_{Rb}

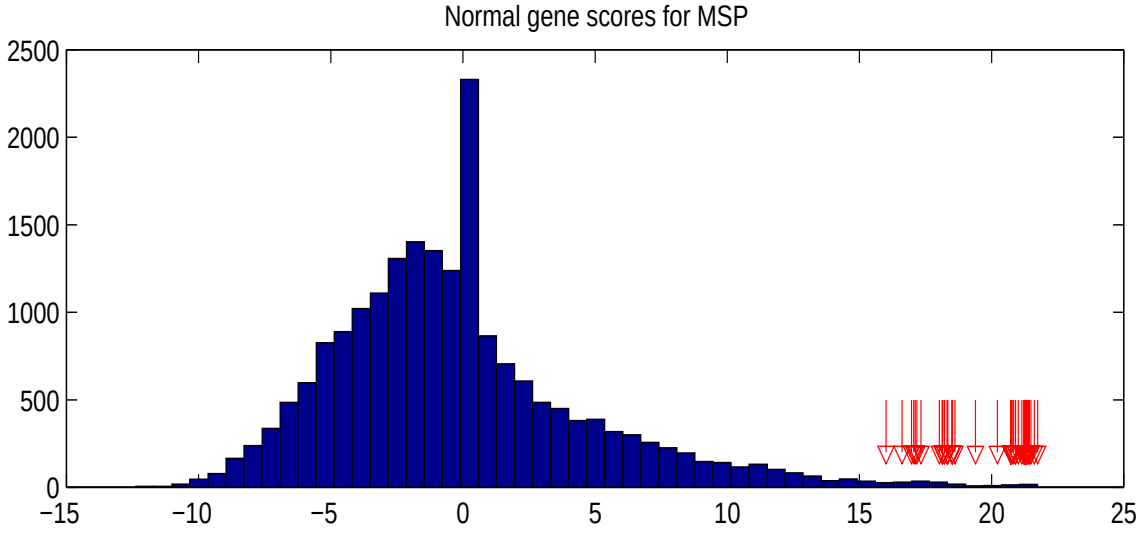
Gene score

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

Large $S_i \implies$ higher “Rb-ness”

Some gene scores





Rb query results

	dpl-1	0.247	E2F (fly)
●	lin-53	0.244	Rb (human)
	K12D12.1	0.240	Topoisomerase II
●	lin-35	0.238	Rb (human)
	Ce Bub1	0.237	Bup1p (yeast)
●	hda-1	0.237	Histone deacetylase
	B0464.6	0.233	Unknown
	R06F6.1	0.233	Histone hairpin (human)
	T16G12.5	0.231	Unknown
	F55A3.7	0.230	Spt16p (yeast)/DRE 4 (fly)
	plk-1	0.229	Polo kinase
●	lin-9	0.227	synMuv
●	lin-36	0.227	synMuv
	smc-4	0.227	In SMC family

Cell Cycle/Chromatin Known Rb interaction

How many experiments?

We use the k best experiments

Small $k \longrightarrow$ not enough to correlate

Large $k \longrightarrow$ 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

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.

In recommender list: 3rd Rb gene placed 6th

For topomap: 3rd Rb gene placed 138th

Similar improvements at other recall levels.

Interpretation

We get a list of candidates

High rank does not prove group membership:

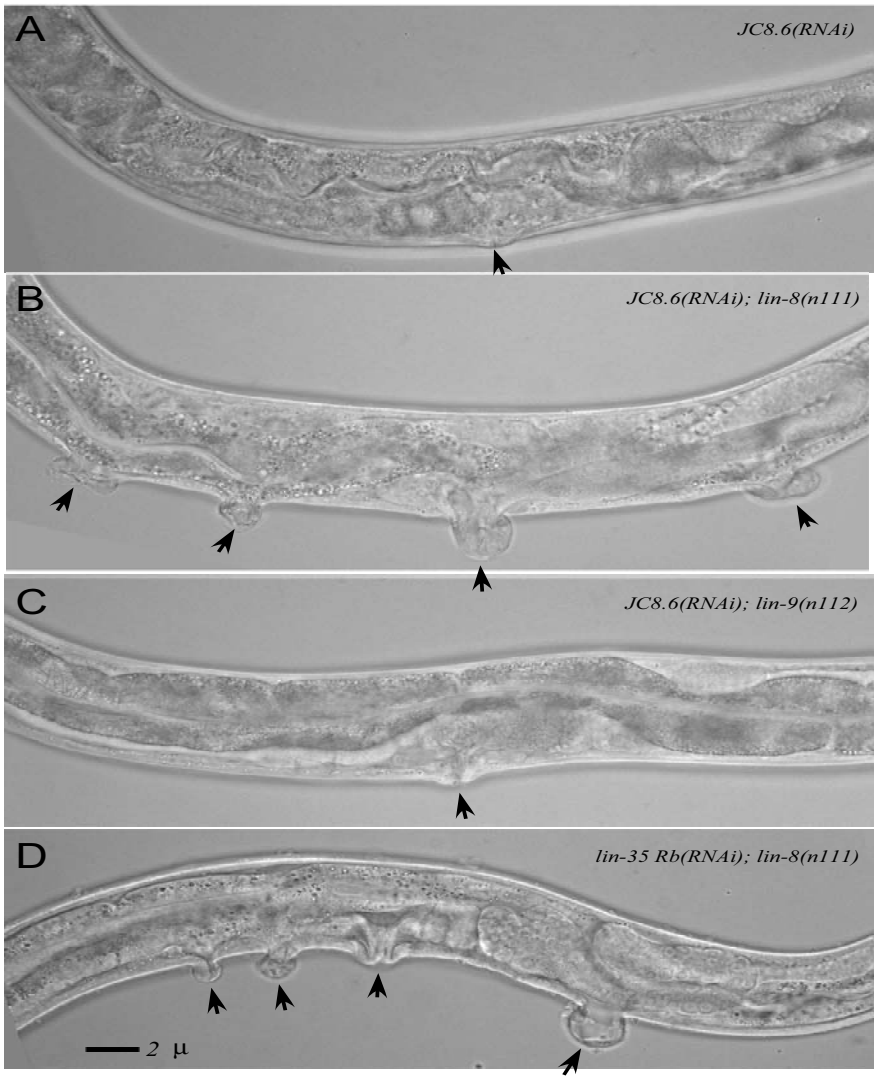
1. We might already have the whole group
2. Strong expression correlations can arise for other reasons

Similar caveats hold for document retrieval or recommenders

RNA Interference

- Tried top 50 ranked genes
- wrm-1 embryonic lethality, supressed by loss of lin-35
- JC8.6 had a synMuv phenotype

JC8.6 $\Longrightarrow$ SynMuv



Try it!

Enter ORFs at:

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

Email queries to:

owen@stat.stanford.edu