Bayesian Models that combine microRNA and gene expression data for breast cancer

Work completed with Colin Campbell and Yiming Ying
at Bristol University, UK

Phaedra Agius
Tartu University, Estonia

Contents

- Motivational data
- Existing data fusion methods
- Two new models – JMM and CORR
- Results
  - Toy data
  - S. Cervisiae
  - Protein folding problem
  - Breast cancer data
- Conclusion

Two data types

Two breast cancer datasets for 78 breast tumors

- Gene expression data (microarray) – denote as E
  Dimensions: G genes by D patients

- microRNA data – denote as C
  Dimensions: H microRNA by D patients

Question: how can we build build a model that uses **both datasets simultaneously** to **cluster** the samples D?

Data merging approaches

Classification using multiple kernels in Bayesian methods
Semi-definite programming

Not so much work on data fusion for unsupervised learning

Natural suggestions:

•Clump the two datasets together
Objection! What if one dataset is much larger than the other?

•Cluster with each dataset separately and then combine the clusterings somehow …
messy and ill-defined.

Issues with our data

(E) Gene expression data

G x D

G=22,000

Two main problems

- varying size
- correspondence between them
  (microRNA regulate gene expression)

➢How can we build a model that does not accord undue influence to E because of its size? (JMM)

➢How can we build a model that also accounts for the correspondence between the two data types? (CORR)

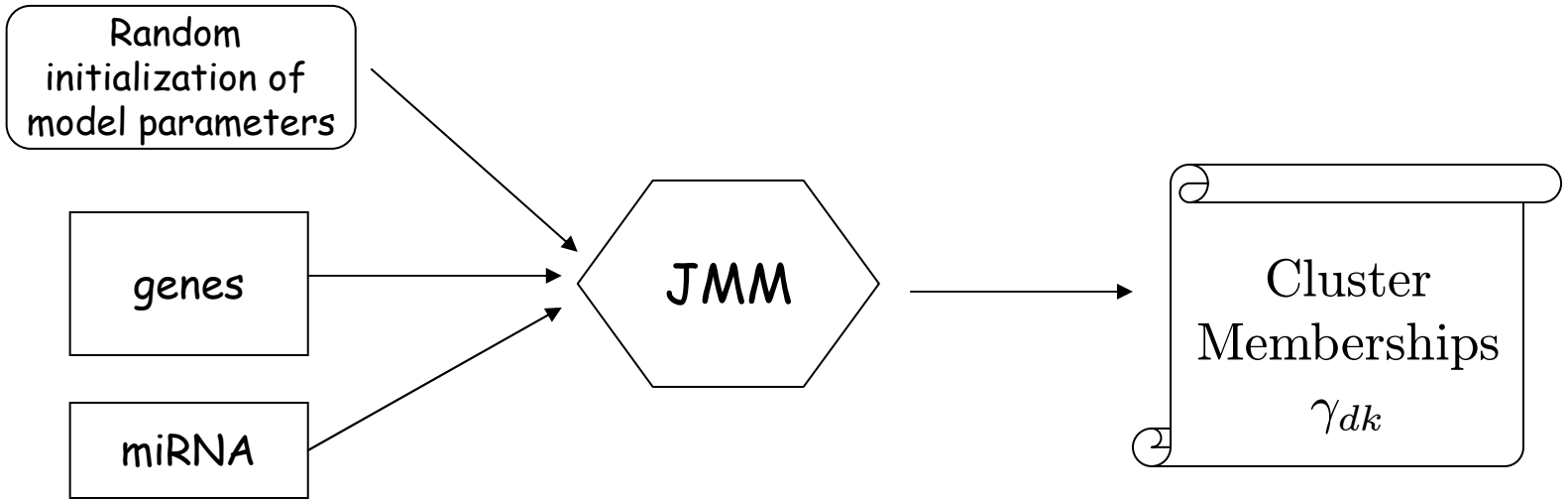
(C) microRNA
H x D … H=137

Joint Mixture Model (JMM)

A model that equally mixes the two datasets regardless of their size.

$$P(\theta_d|\alpha)P(C_d|\tilde{\mu},\tilde{\sigma},\theta_d)P(E_d|\mu,\sigma,\theta_d)$$

Two disparate datasets allowed to share a common prior distribution and latent variables



Computational details - JMM

The overall joint distribution

$$\prod_d p(C_d, E_d, z_d, y_d, \theta_d|\Theta)$$

where

$$p(C_d, E_d, z_d, y_d, \theta_d|\Theta) = p(\theta_d|\alpha) \prod_h p(z_{dh}|\theta_d) N(C_{dh}|\tilde{\mu}_{hz_{dh}}, \tilde{\sigma}_{hz_{dh}})$$
$$\text{x} \prod_g p(y_{dg}|\theta_d) N(E_{dg}|\mu_{gy_{dg}}, \sigma_{gy_{dg}})$$

$$z_{dh}, y_{dg} \in \{1, \ldots K\} \text{ where } K \text{ is the number of clusters}$$

Ideal Goal: Compute posterior distribution and learn model parameters
Computationally expensive!!!

Altered Goal: Use variational inference to minimize the KL-divergence between the variational distribution of the latent variables and the posterior distribution

Result: a set of variational EM-type updates for variational and model parameters. Iterative procedure pursued until convergence of the KL-divergence

CORRespondence Model

Inspired by correspondence **L**atent **D**irichlet **A**llocation [Blei et al] which was proposed for the joint modeling of images and their corresponding caption words

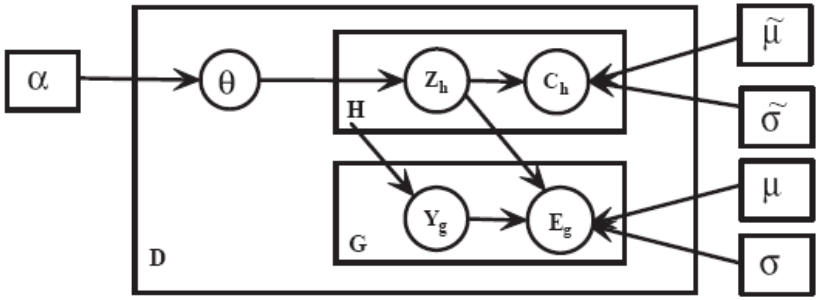
Motivation is that microRNA drives gene expression.

Our model assumes a dependency of E on C.

$$P(\theta_d|\alpha)P(C_d|\tilde{\mu},\tilde{\sigma},\theta_d)P(E_d|\mu,\sigma,\theta_d,C_d)$$

Computational details and algorithmic updates are similar to JMM.
But computational cost involves H times more iterations.

CORR model - details



For a given data index d for both E ($G \times D$ matrix) and C ($H \times D$ matrix)

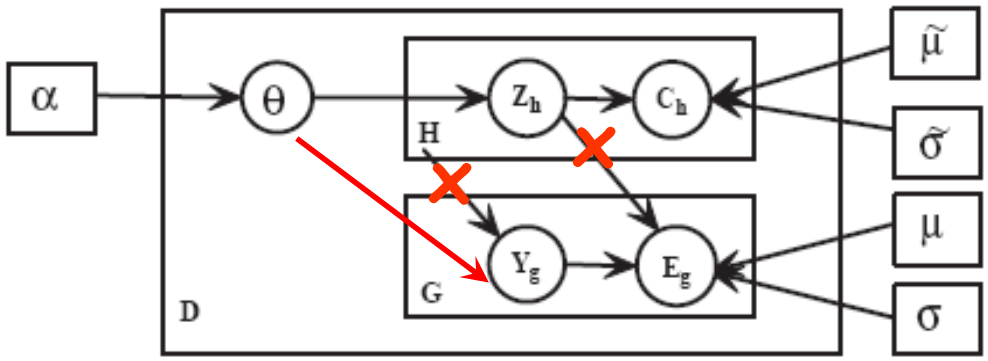
1. Prior distributions: $\theta_d \sim \mathrm{Dir}_K(\alpha)$
2. Choose C_d:
   (a) Choose process for C_{hd}: $z_{dh} \sim \mathrm{Multi}(\theta_d)$
   (b) Sample $C_{hd} \sim \mathcal{N}(C_{hd}|\tilde{\mu}_{hz_{dh}}, \tilde{\sigma}_{hz_{dh}})$ where $\mathcal{N}(C_{hd}|\tilde{\mu}, \tilde{\sigma}^2)$ denotes a normal distribution with mean $\tilde{\mu}$ and variance $\tilde{\sigma}^2$.
3. Choose E_d:
   (a) Sample gene correspondence: $y_{dg} \sim \mathrm{Uniform}(1, \ldots, H)$
   (b) Sample $E_{gd} \sim \mathcal{N}(E_{gd}|\mu, \sigma, z, y_{dg}) = \mathcal{N}(E_{gd}|\mu_{gz_{dh}}, \sigma^2_{gz_{dh}}, y_{dg} = h)$

CORR versus JMM

- Can be used to cluster samples in two related datasets, taking into account their correspondence
- Resulting clustering heavily depends on C

- Can be used to cluster samples in any two datasets
- Can be used to cluster any number of datasets
- Resulting clustering depends equally on C and E

For JMM



Results

- Toy Data
- S. Cerevisiae
- Protein fold classification
- Breast cancer data

For all the results presented,
the data was normalized across the samples.

Toy data

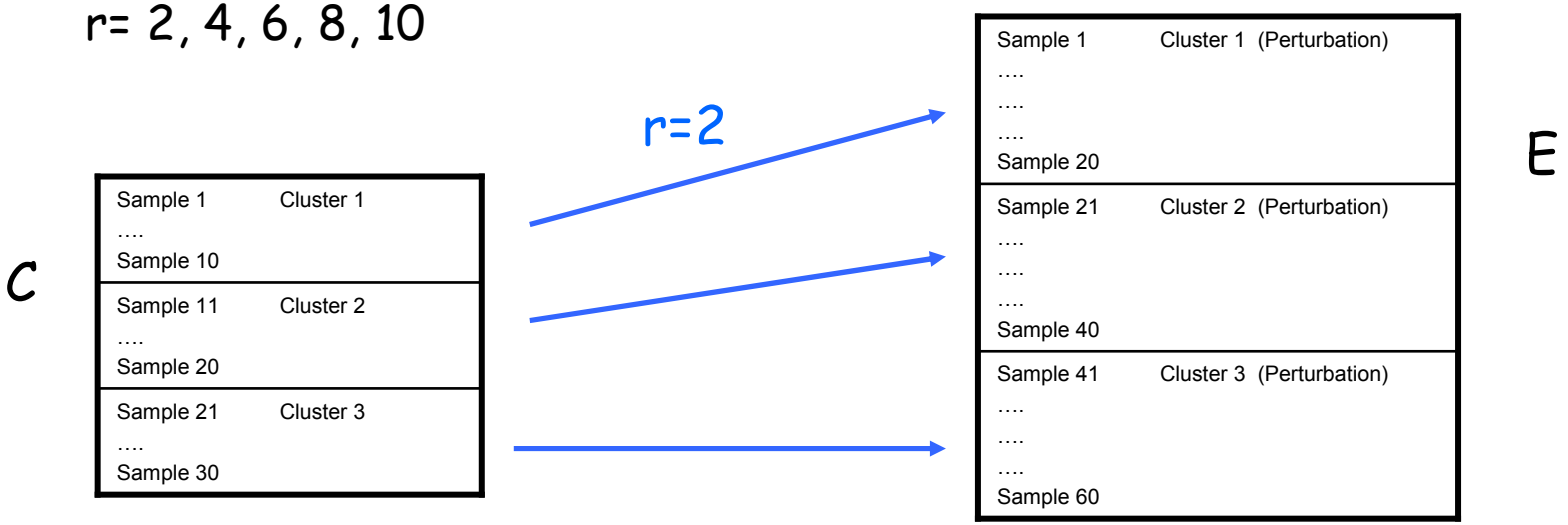
Artificial data generated for C and E

For C: Data generated for 3 clusters, 10 samples per cluster

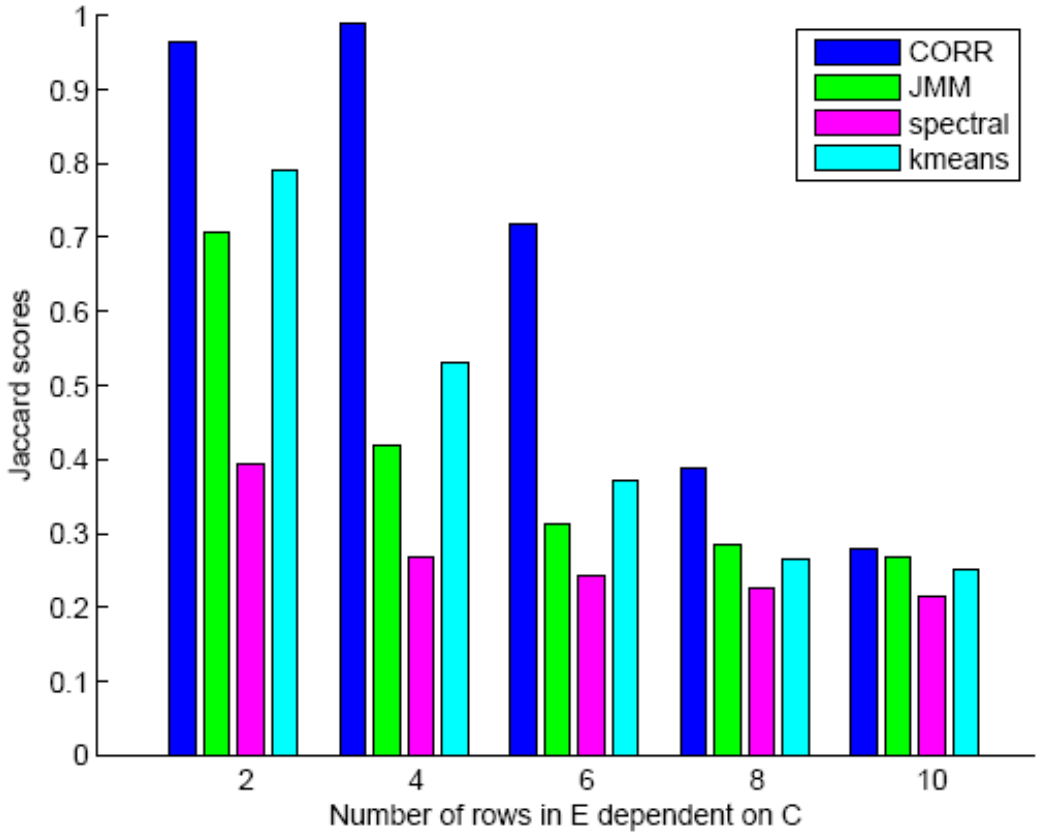
For E, (#rows=1): Data generated for 3 clusters such that each row in E is correlated to rows in C with each feature perturbed by a small Gaussian random deviate addition

For E, (#rows=r): Data generated for 3 clusters such that every r rows in E …

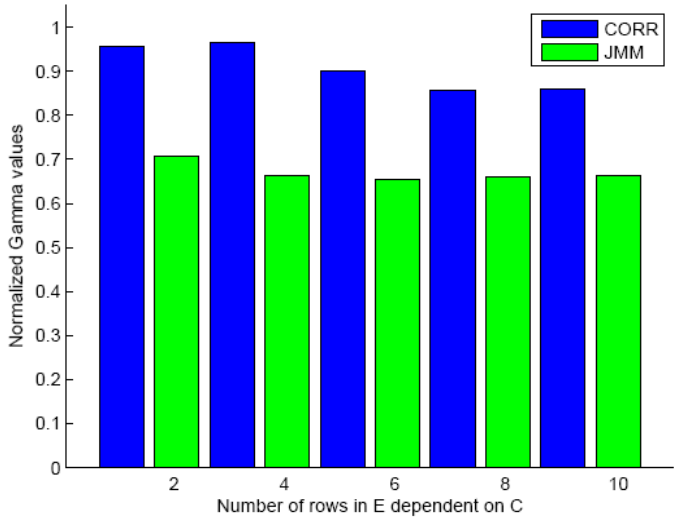
r= 2, 4, 6, 8, 10



Toy data - results



Clustering accuracy



Cluster membership

A second artificial example, using real data

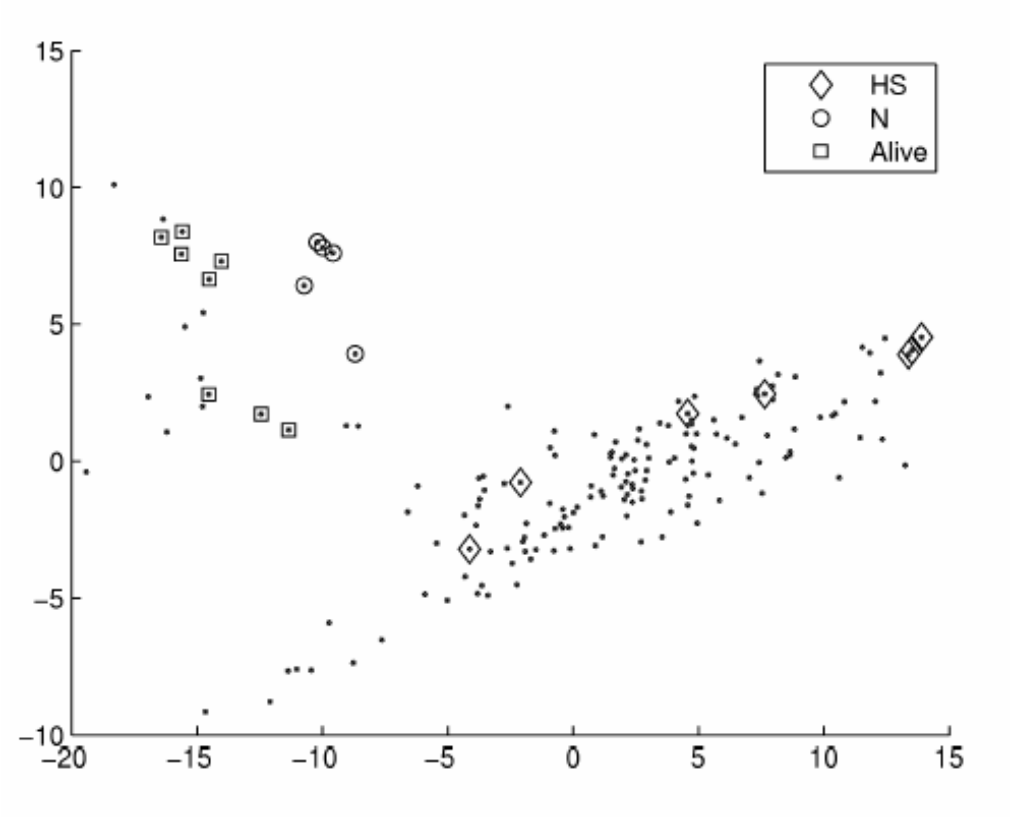
S. Cerevisiae data used by [Middendorf et al]
Authors identified a strong regulating factor USV1 believed to influence up to 305 genes.

This is an extreme example:
C consists of just one gene, USV1
E consists of the 305 other gene expressions
Therefore E is now 305 times bigger!!!

Data: *S. Cerevisiae* subjected to a series of experimental conditions.
We pick three experimental conditions:
7 heatshock experiments over various time intervals
5 nitrogen depletion experiments over various time intervals
7 stationary phase experiments, growth under normal conditions over a period of 1 to 28 days (used as time-zero references)

Goal: Use JMM or CORR to correctly group the samples into the 3 classes

S. Cerevisiae [Middendorf et al]



← PCA plot shows nice separation

Results

Both JMM and CORR correctly classified the experiments (30 random initializations, highest log-likelihood selected)

To compare, we amalgamated USV1 expression with the 305 genes and ran k-means and spectral k-means 100 times each.

Kmeans = 62/100, spectral=81/100

Protein fold classification

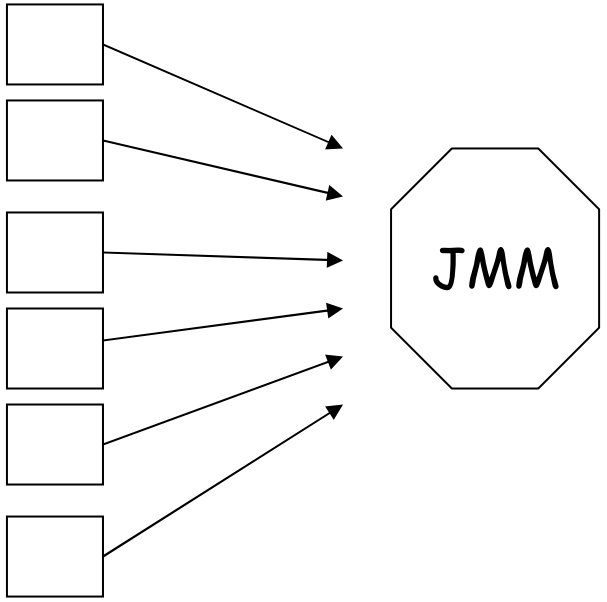
With reference to the protein fold classification dataset derived by [Girolami and Zhong]

698 proteins, classifiable into four main protein fold groups:
α, β, α + β, α/β (Known labels, can validate)

Protein descriptive factors:

1. Amino acid composition
2. Hydrophobicity profile
3. Polarity
4. Polarizability
5. Secondary structure
6. Van der Waals interaction

This time we used JMM on 6 datasets!



Protein fold classification - Results

Process (cluster) memberships turned out to be occasionally ambiguous for this data. Therefore, thresholds of 0.4 and 0.5 were set so that only proteins that exceeded the threshold were considered for the Jaccard score.

For comparison, we used two other algorithms:

Latent Process Decomposition (LPD, [Rogers et al])

Chinese Restaurant Clustering models (CRC, [Qin])

These algorithms only take one dataset at a time. Options are:

- amalgamate the 6 datasets
- choose the single best performing dataset

Threshold		Amalgamated	Amalgamated	Single best	Single best
	JMM	**LPD**	**CRC**	**LPD**	**CRC**
0.4	0.48	0.32	0.18	0.29 (Van der Waals)	0.18 (hydrophobicity)
0.5	0.59	0.33	0.12	N/A	0.19 (amino acid composition)

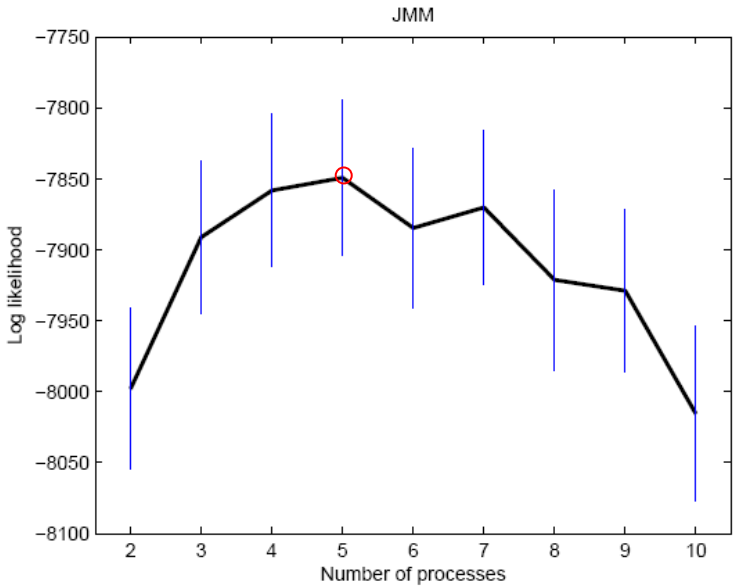
Breast cancer data

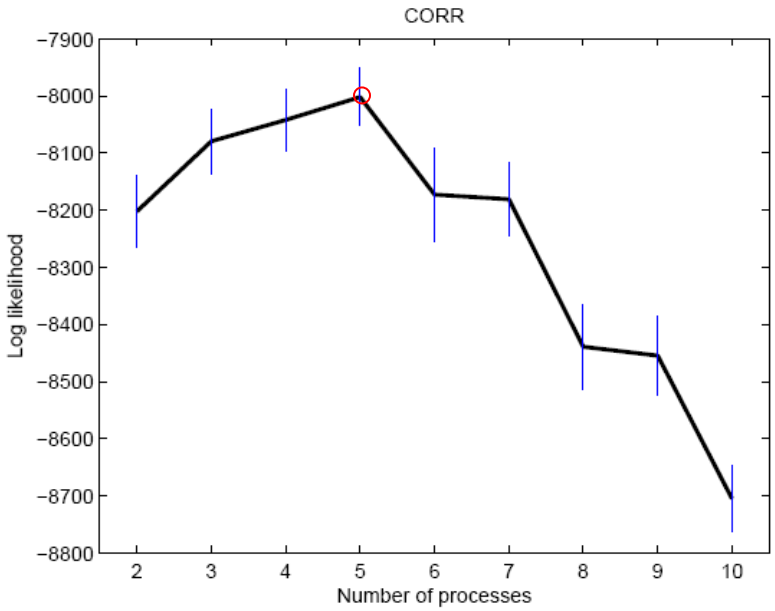
78 samples
22,000 genes; top 600 most variant genes only
133 miRNA

Data is normalized

Optimal number of clusters

- 10 fold cross-validation
- Model used to estimate the log-likelihood of the hold-out data
- Number of clusters ranged from 2 to 10
- Optimal number of clusters occurs when the log-likelihood peeks





In both cases, the optimal number of clusters (processes) is 5

Kaplan Meier plots

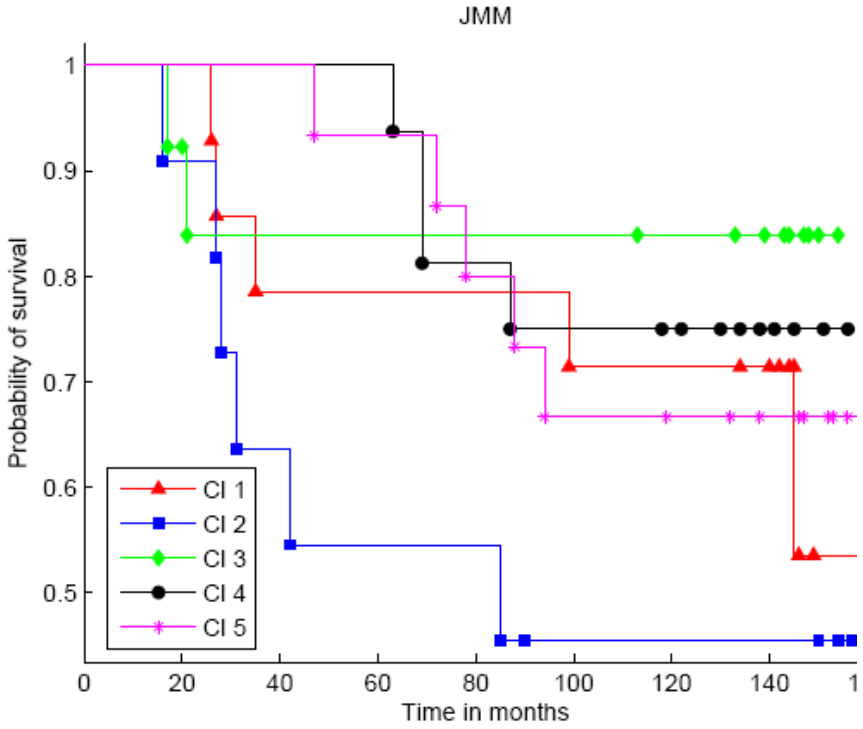
Visual display of the survival function

For every patient in the dataset, let $t_1, t_2, \ldots t_N$ denote the time of death in months.
Let $S(t)$ denote the probability that a cancer patient survives beyond time t.
Then the maximum likelihood estimate of $S(t)$ is

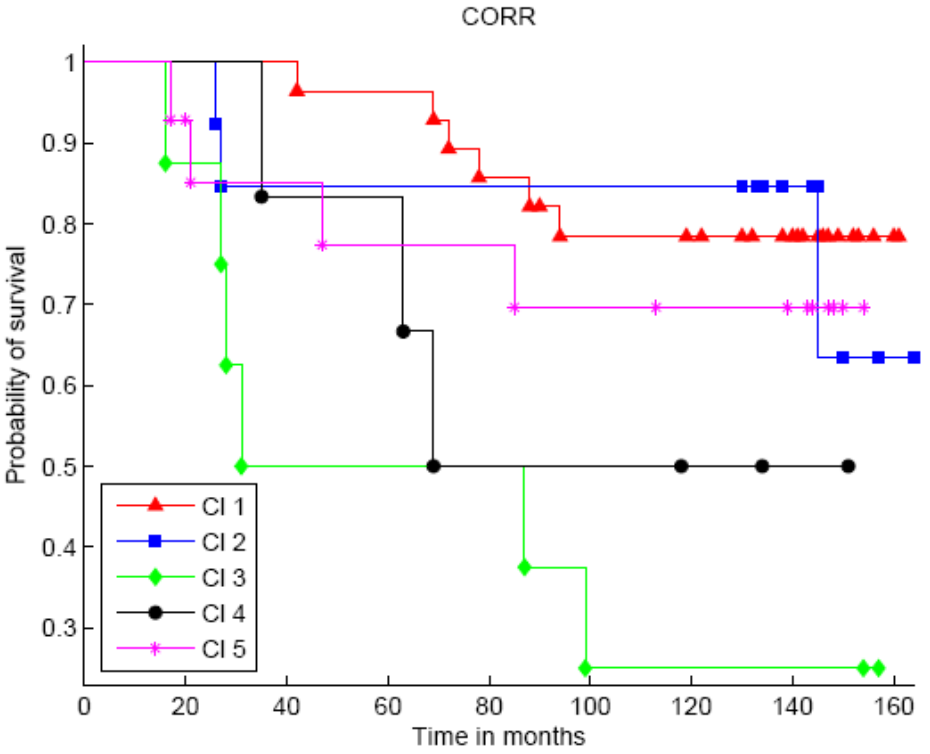
$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

where n_i is the number of survivors prior to time t and d_i is the number of deaths at time t_i.

Kaplan Meier plots



One aggressive subtype (Cl 2)



Two aggressive subtypes (3 & 4)

Using Mann Whitney scores to find abnormally expressed genes

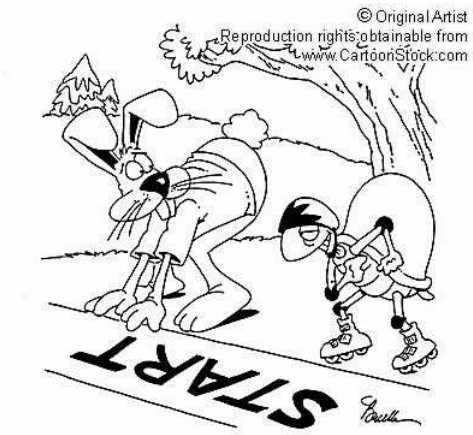
Mann Whitney (MW) - rank based score.

Null hypothesis: probability of an observation from one population exceeding that from a second population is 0.5.
(Assumption: the population distributions are the same)

Classic example: Aesop is dissatisfied with his experiment where one tortoise beats one hare. Try more races! Race 6 tortoises versus 6 hares.
The finishing line is crossed in this order:

1	2	3	4	5	6	7	8	9	10	11	12
T	H	H	H	H	H	H	T	T	T	T	T	T	H

Take each tortoise and count number of hares it beats.
We get 6,1,1,1,1,1. Set U=6+1+1+1+1+1=11.



Mann Whitney genes

ordered gene expressions … 10 11 12

Gene X THHHHHHTTTTTTH ------- T,H are clusters

From U to a statistically interpretable *z*-score:

$$z = \frac{(U - m_U)}{\sigma_U}$$

$$m_U = \frac{n_1 \cdot n_2}{2}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

In our context, MW may be applied pairwise to find abnormally expressed genes within one subtype relative to the 4 other subtypes found by the CORR model.

Top 20 abnormally expressed genes

Least aggressive

Most aggressive

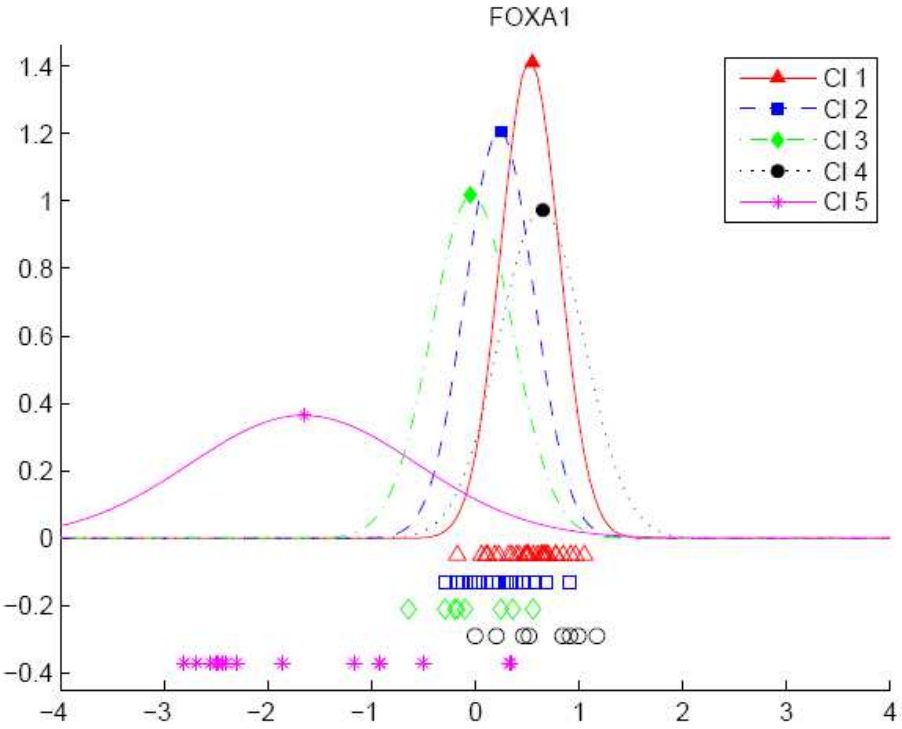
Cl 1	Cl 2	Cl 5	Cl 4	Cl 3
UBE2C	*COL11A1*	**GATA3**	*CTGF*	*GSDML*
CDC20	*TIMP3*	**FOXC1**	*RARRES1*	*ORMDL3*
POSTN	*AEBP1*	*STARD10*	*C1S*	**ERBB2**
CYBRD1	*COL10A1*	**MLPH**	*PRKACB*	*STARD3*
OGN	*PLAU*	*TOB1*	*FBLN2*	*FGFR4*
ADH1B	*MFAP5*	**AGR2**	*TNC*	**ESR1**
ADH1A	*COL12A1*	**FBP1**	*ACTA2*	*PERLD1*
CYP4X1	*MMP11*	*GPR160*	*CR598488*	*CTXN1*
COL10A1	*FN1*	*C10orf116*	*COL6A1*	*DQ582071*
TIMP3	*SULF1*	*BCAS1*	*SPON1*	**GRB7**
TK1	*COL8A1*	*DEGS2*	*ASS1*	*RAP1GAP*
SH3BGRL	*POSTN*	**XBP1**	*FLNA*	*C1S*
SUSD3	*NBL1*	*CRYAB*	*PKIB*	*U79293*
MIA	*DCN*	*EEF1A2*	*SBEM*	*PRSS8*
CPA3	*OGN*	*SLC39A6*	*abParts*	*C17orf37*
PPP1R3C	*GJB2*	*KRT19*	*FLJ42258*	*MFAP2*
SFRP1	*THBS2*	*GALNT6*	*CRISPLD2*	**TFF1**
ATP1B1	*ACTA2*	**FOXA1**	*BAMBI*	**CA12**
SLC40A1	*TBC1D9*	**GABRP**	*SYT13*	*TBC1D9*
CILP	*LOXL2*	*NPNT*	*IGHA2*	*CAPS*

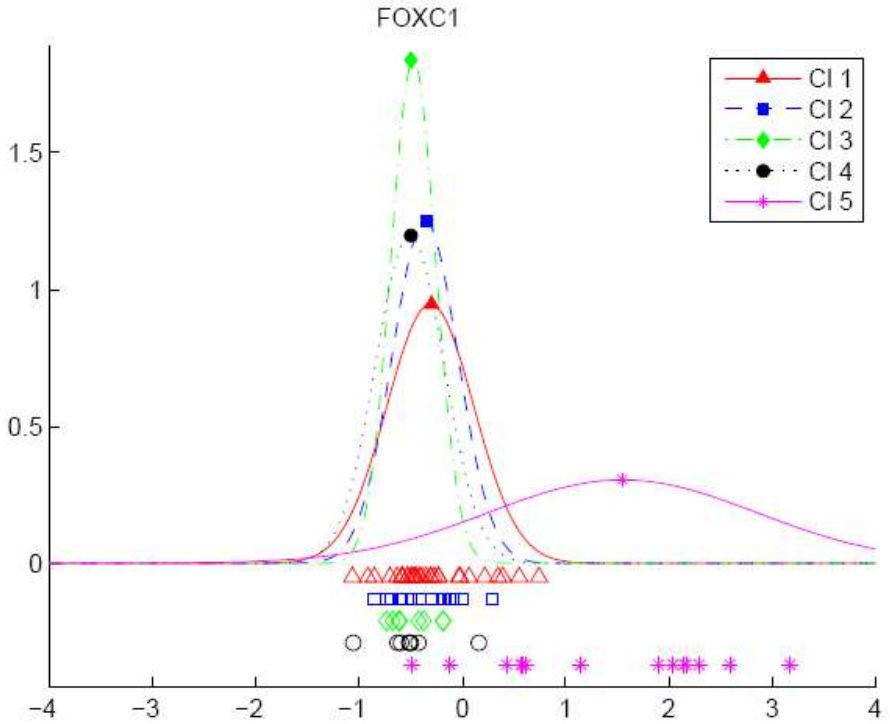
X box binding protein is believed to be regulated by FOXA1 [Carroll et al]

GATA3 has associated co-expression with XBP1 and ESR1 [Lacroix et al]

ERBB2 and GRB7 elevated in some breast cancer subtypes [Sorlie et al]

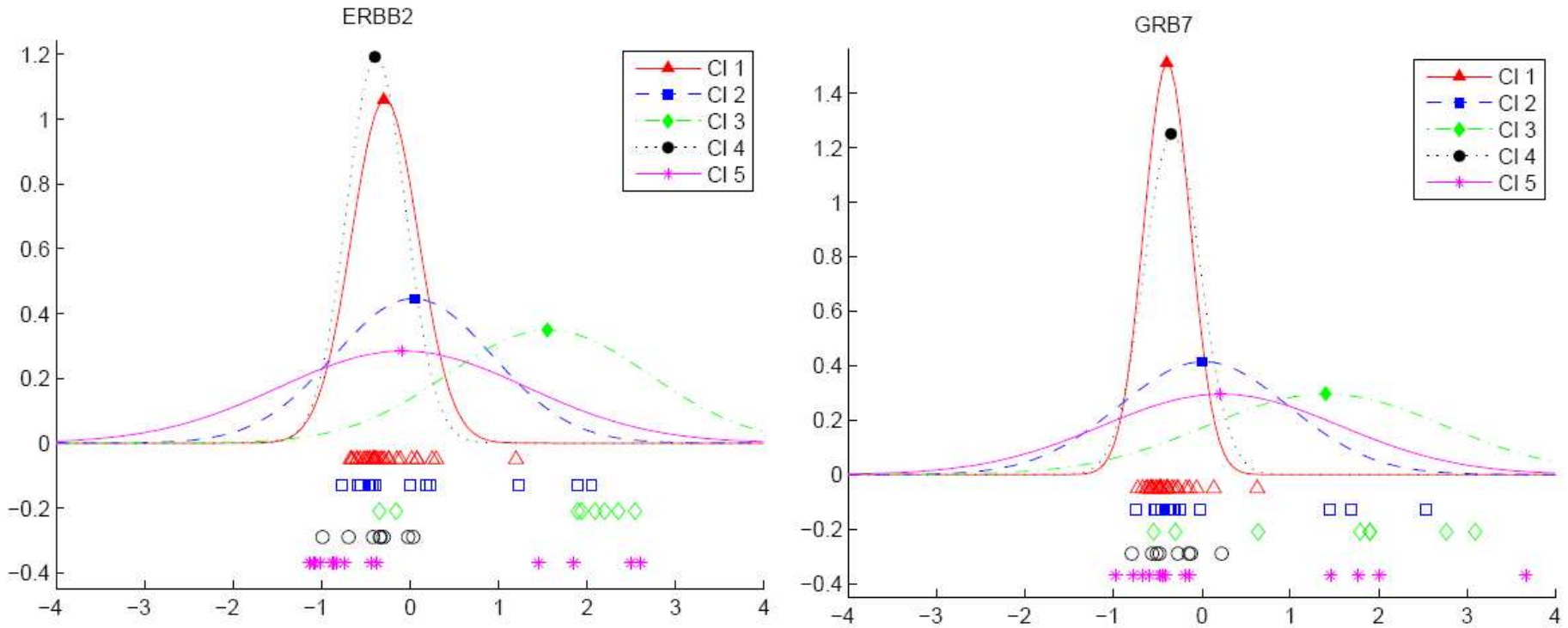
Interesting gene profiles (CORR)





In Cl5, while FOXA1 under-expresses, FOXC1 over-expresses.

More gene profiles



ERBB2 and GRB7 over-express in subtype Cl3

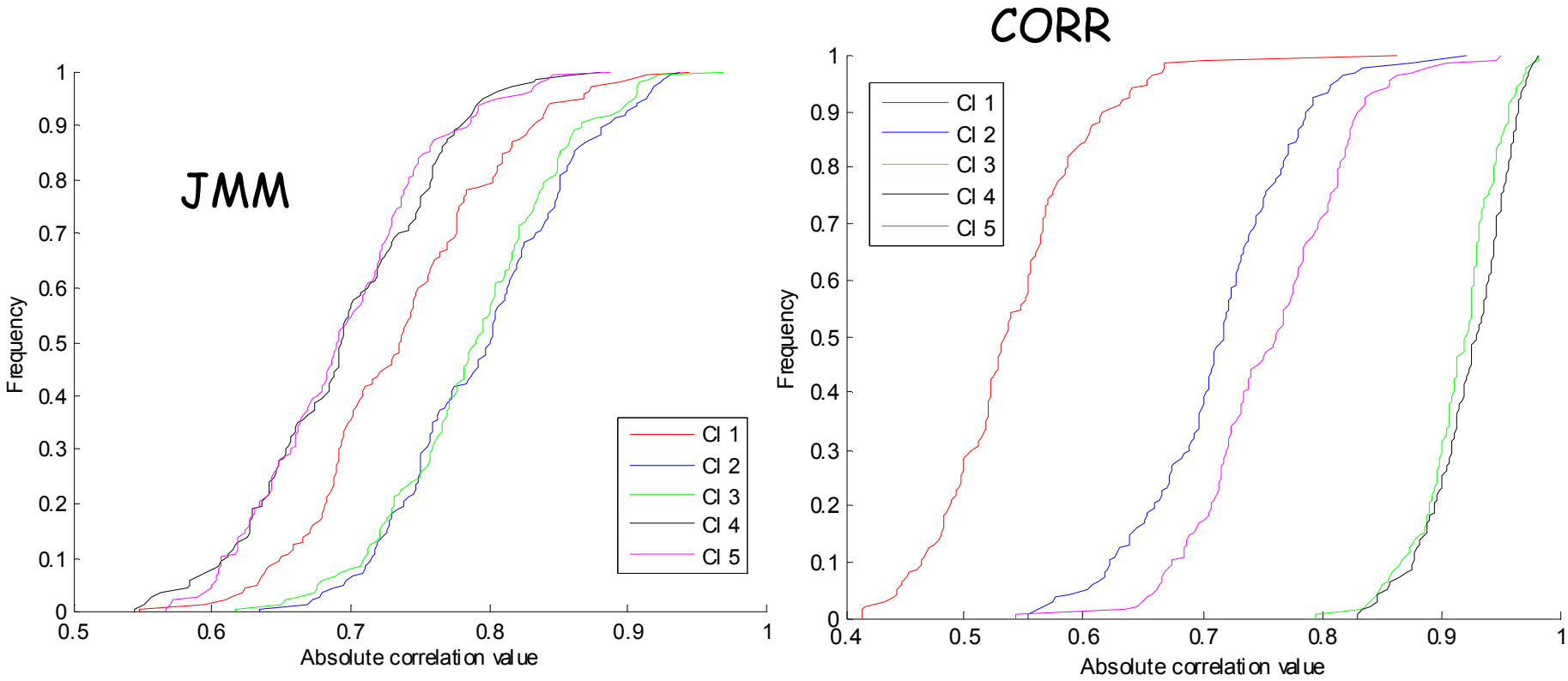
Top 10 abnormally expressed microRNA

C1		C2		C3		C4		C5	
miR-505	0.38	miR-137	0.26	miR-152	1.13	miR-30b	0.66	miR-199a	0.62
miR-181c	0.37	miR-133a	0.19	miR-342	0.99	miR-15b	0.63	miR-99a	0.57
miR-142-5p	0.36	miR-9	0.19	miR-29a	0.98	miR-15a	0.60	miR-199b	0.555
miR-185	0.31	miR-9	0.18	miR-331	0.96	miR-30c	0.57	miR-199a	0.547
miR-203	0.31	miR-18a	0.08	miR-214	0.95	miR-195	0.55	miR-214	0.474
miR-200a	0.30	miR-128b	0.07	miR-199b	0.94	miR-16	0.49	miR-100	0.471
miR-183	0.29	miR-138	0.06	miR-126	0.90	miR-21	0.49	miR-130a	0.453
miR-509	0.29	miR-211	0.03	miR-145	0.89	miR-20a	0.45	miR-382	0.429
miR-107	0.29	miR-335	0.03	miR-24	0.89	miR-30a-3p	0.45	miR-125b	0.42
miR-93	0.29	miR-429	0.02	miR-27a	0.88	miR-210	0.44	let-7b	0.40

Some published work referencing microRNA and breast cancer

- miR-126, miR-335 [Tavazoie et al]
- miR-145, miR-21 [Sempere et al]

Absolute miRNA-gene correlations in cancer subtypes



Largest absolute correlation between miRNA-gene pairs chosen.

Stronger correlations observable for the CORR model in the aggressive subtypes Cl3 and Cl4. Not for JMM (Cl2=aggressive, Cl3=least aggressive).

Conclusion

- More data, other cancer types
- Properly investigate the top MW genes and miRNA
- Can we extend the corr model to more than one dataset? In what biological context?
- Can we use the CORR model to directly infer the correspondence between two datasets?
- Are the correlations truly meaningful? (our dataset was rather small)

- The proposed models can handle missing values – have not yet tried this
- The models were presented using only continuous data. But they could equally well use discrete data using multinomial or Poisson distributions – eg. a CORR model for motif counts and gene expression

References

J Carroll et al. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FOXA1. *Cell*, 133:33-43, 2005.

Tavazoie SF et al. Endogenous human microRNAs that suppress breast cancer metastasis.
Nature, 2008 Jan 10; 451(7175): 147-52.

Sempere LF et al. Altered microRNA expression confined to specific epithelial cell subpopulations in breast cancer. *Cancer Research*, 2007 Dec 125; 67(24):11612-20.

Girolami M and Zhong M. Data integration for classification problems employing gaussian process priors. In *Twentieth Annual Conference on Neural Information Processing Systems*, 2007.

Rogers S, Girolami M, Campbell C, Breitling R. The latent process decomposition of cDNA microarray datasets. *IEEE/ACM Transations on Computational Biology and Bioinformatics*, 2:143-156, 2005.

Qin Z S. Clustering microarray gene expression data using weighted chinese restaurant process. *Bioinformatics*, 22: 1988-1997, 2006.