Bayesian Statistics for Genetics
Lecture 8: Hierarchical Model

June, 2024

Kidney cancer example

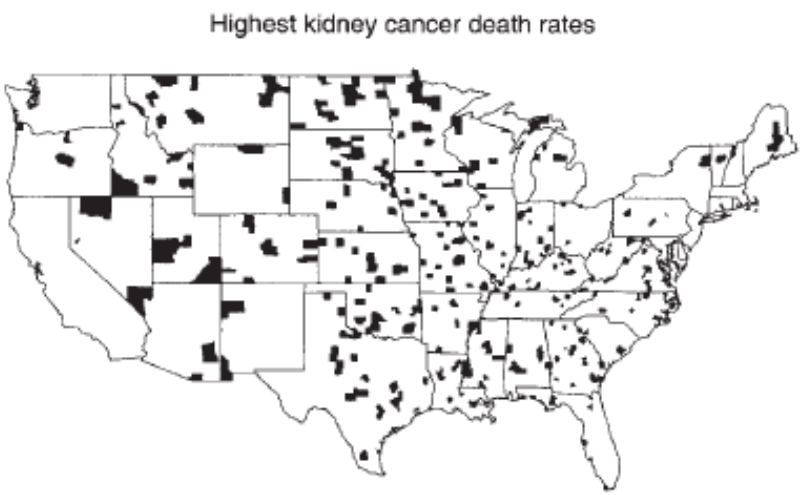


Figure 2.7 *The counties of the United States with the highest 10% age-standardized death rates for cancer of kidney/ureter for U.S. white males, 1980–1989. Why are most of the shaded counties in the middle of the country? See Section 2.8 for discussion.*

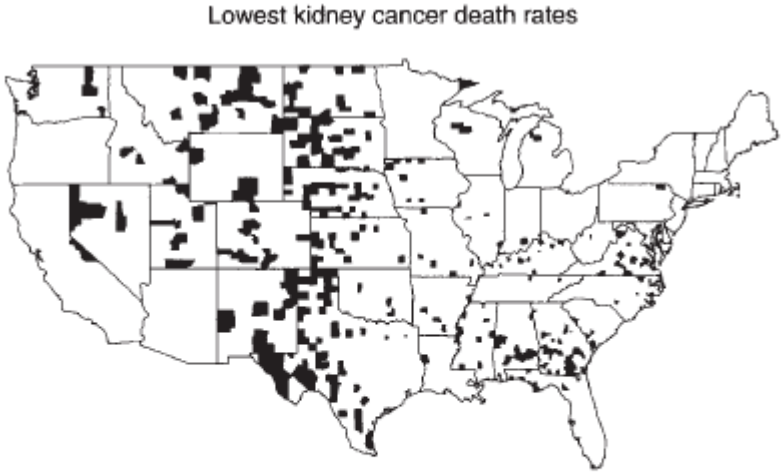
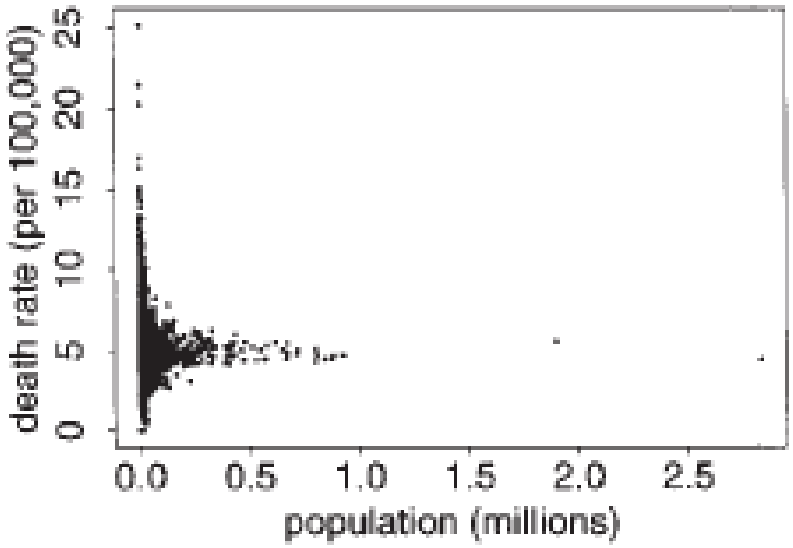
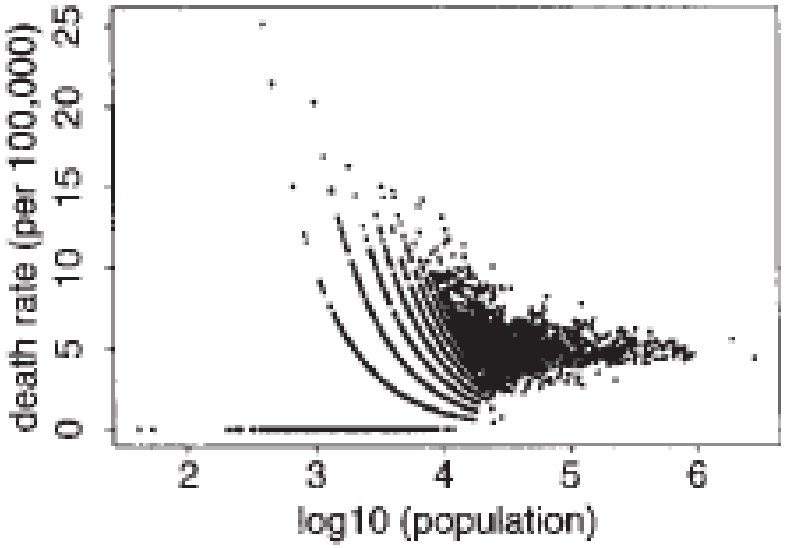


Figure 2.8 *The counties of the United States with the lowest 10% age-standardized death rates for cancer of kidney/ureter for U.S. white males, 1980–1989. Surprisingly, the pattern is somewhat similar to the map of the highest rates, shown in Figure 2.7.*

Kidney cancer death rate versus population size





Typical high-throughput data

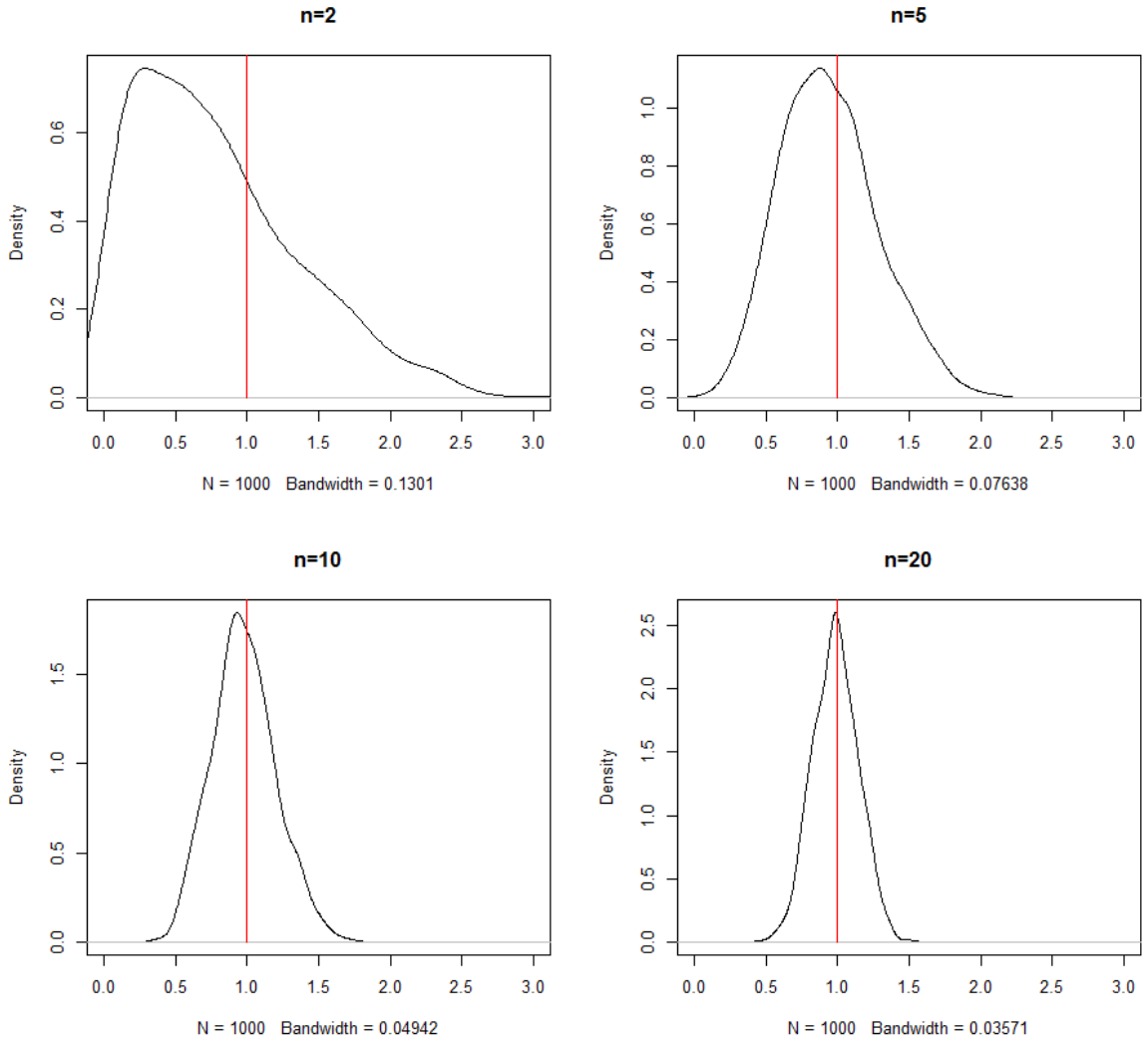
- n: sample size
- m: number of features (genes, proteins)
- m>>n
- Normally, the number of samples are limited.

Detection of DE genes

- A classical problem in gene expression study: detect differentially expressed (DE) genes.
- DE genes: genes from various samples are expressed differentially in different cell types, tissues, developmental stages or diseases.
- Many applications: RNA-seq, ChIP-seq, ATAC-seq, …

Typically the number of replicates is rather low.

The problem


n=2
Density
0.0
0.2
0.4
0.6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
N = 1000 Bandwidth = 0.1301
n=5
Density
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
N = 1000 Bandwidth = 0.07638
n=10
Density
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
N = 1000 Bandwidth = 0.04942
n=20
Density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
N = 1000 Bandwidth = 0.03571

Methods for detecting DE genes

- Fold change
- Classical *t*-test
- SAM (Significance Analysis of Microarray)

  Add a constant to the denominator of the t-statistics

  Tusher et al. 2001.
- Model-based methods (Li and Wong 2001):
  - LIMMA (Linear Models for Microarray Data)

  Use Bayesian hierarchical model in multiple regression setting (Smyth 2004).

An motivating example

To estimate the probability of tumor in a population of female F344 laboratory rats that receive a certain dose of the drug.

0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/47	15/46	9/24

From Bayesian Data Analysis Gelman et al.

The probability model

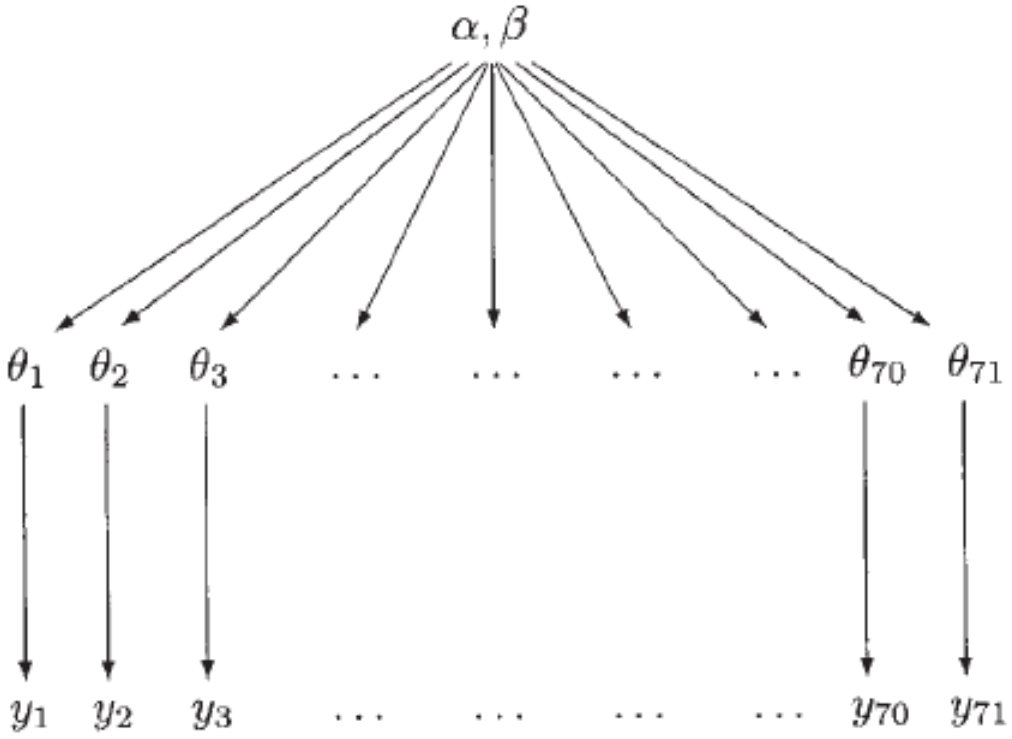
- $(n_i, y_i), i = 1, \ldots, n$
- $y_i \sim Binom(n_i, \theta_i)$

Choose from the following models

- **Separate:** assume data from each experiment follow its own Binomial distribution: θ_i's distinct.
- **Pooled:** assume data from all experiments follow the same Binomial distribution: θ_i's identical.
- **Hierarchical:** something in between. But how?

Hierarchical model

- When you don't have much, borrow.

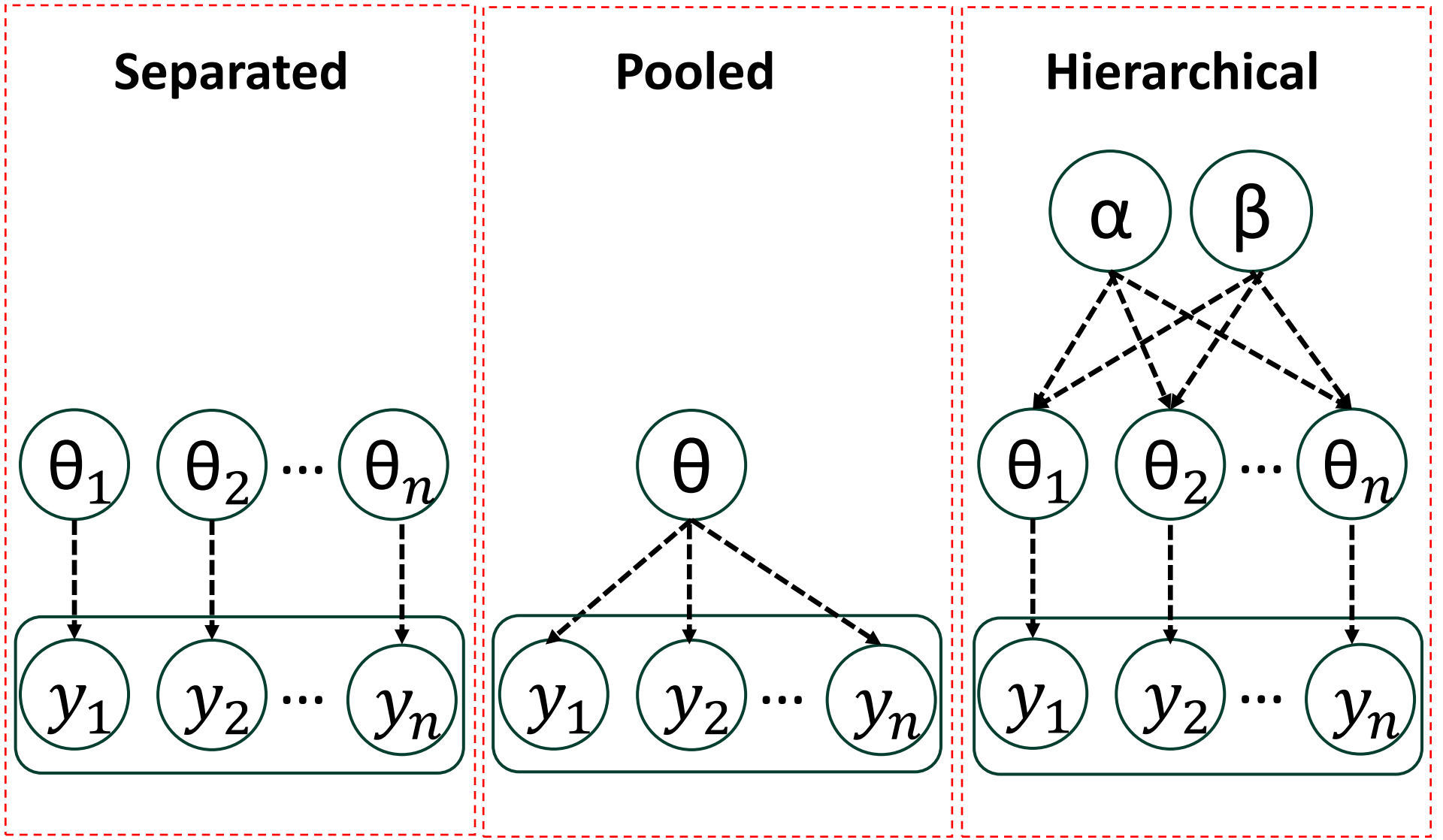


From Bayesian Data Analysis Gelman et al.

Hierarchical model

- Each experiment follow its own Binomial distribution. But we assume all the θ_i's are sampled from a common distribution—Hierarchical distribution.

$$y_i \sim Binom(n_i, \theta_i)$$
$$\theta_i \sim Beta(\alpha, \beta)$$


Separated
Pooled
Hierarchical
α
β
θ
θ1
θ2
θn
y1
y2
yn

Inference of hierarchical model

- Data and parameters: $\alpha, \beta, \theta_j, y_j, j = 1, \ldots, J$

- Joint distribution:

$$P(\alpha, \beta, \theta_j | y_j) \propto p(\alpha, \beta) p(\theta | \alpha, \beta) p(y | \theta, \alpha, \beta)$$

$$\propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_j^{\alpha-1} (1-\theta_j)^{\beta-1} \prod_{j=1}^{J} \theta_j^{y_j} (1-\theta_j)^{n_j - y_j}$$

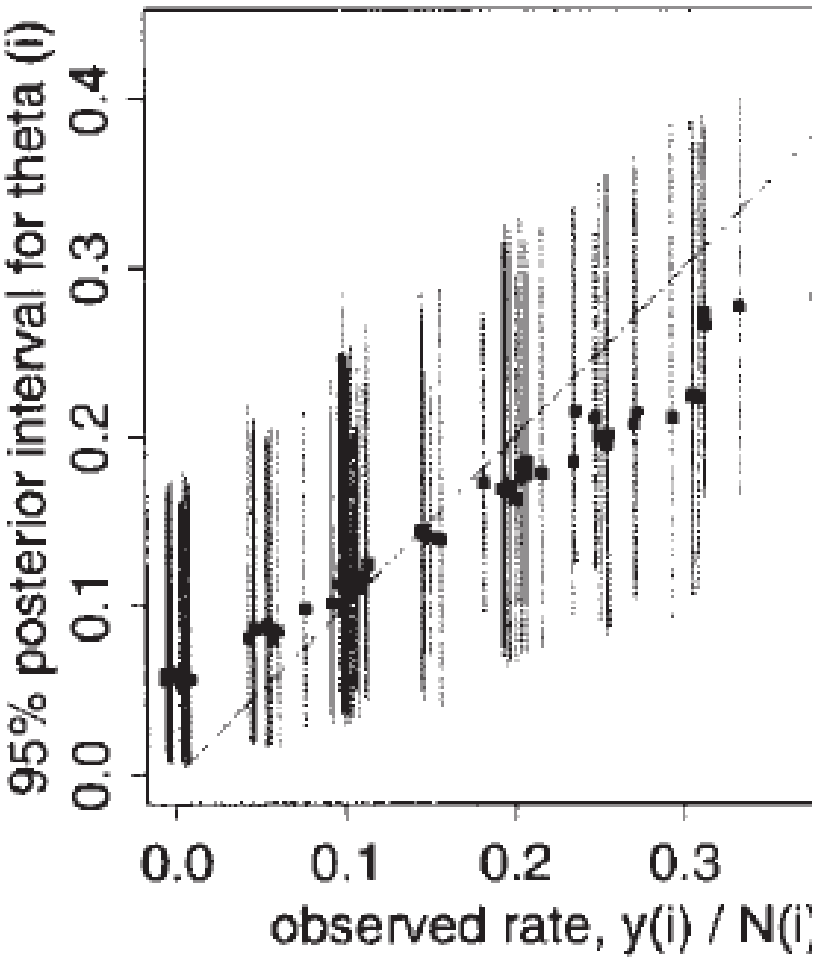
- Posterior distribution for hyperparameters α, β:

$$P(\alpha, \beta \,|y) \propto p(\alpha, \beta) \prod_{j=1}^{J} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}$$

- Posterior distribution for hyperparameters θ_j:

$$p(\theta_j | \alpha, \beta, y) = \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)} \theta_j^{\alpha+y_j-1} (1-\theta_j)^{\beta+n_j-y_j-1}$$

Impact of hierarchical model



From Bayesian Data Analysis Gelman et al.

For microarray data

X_1

X_2

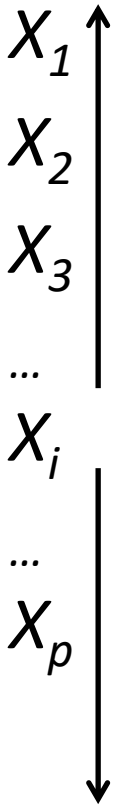
X_3

…

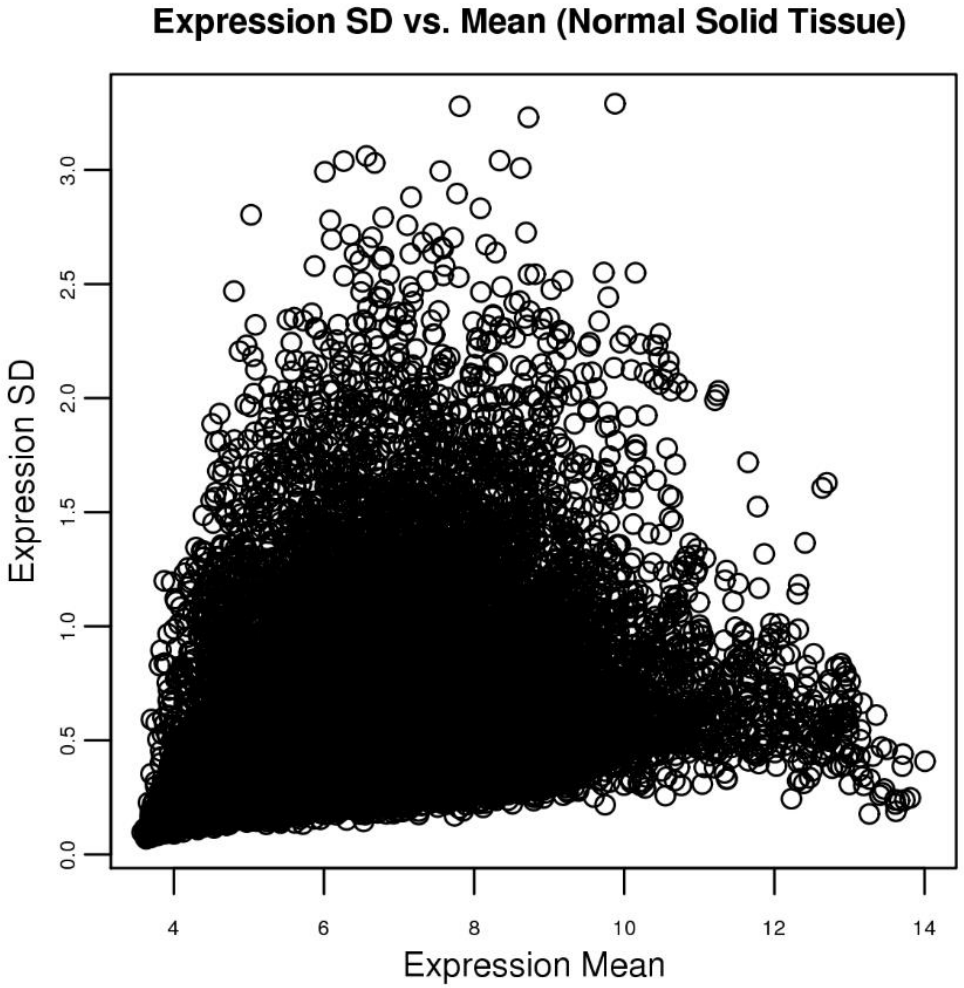
X_i

…

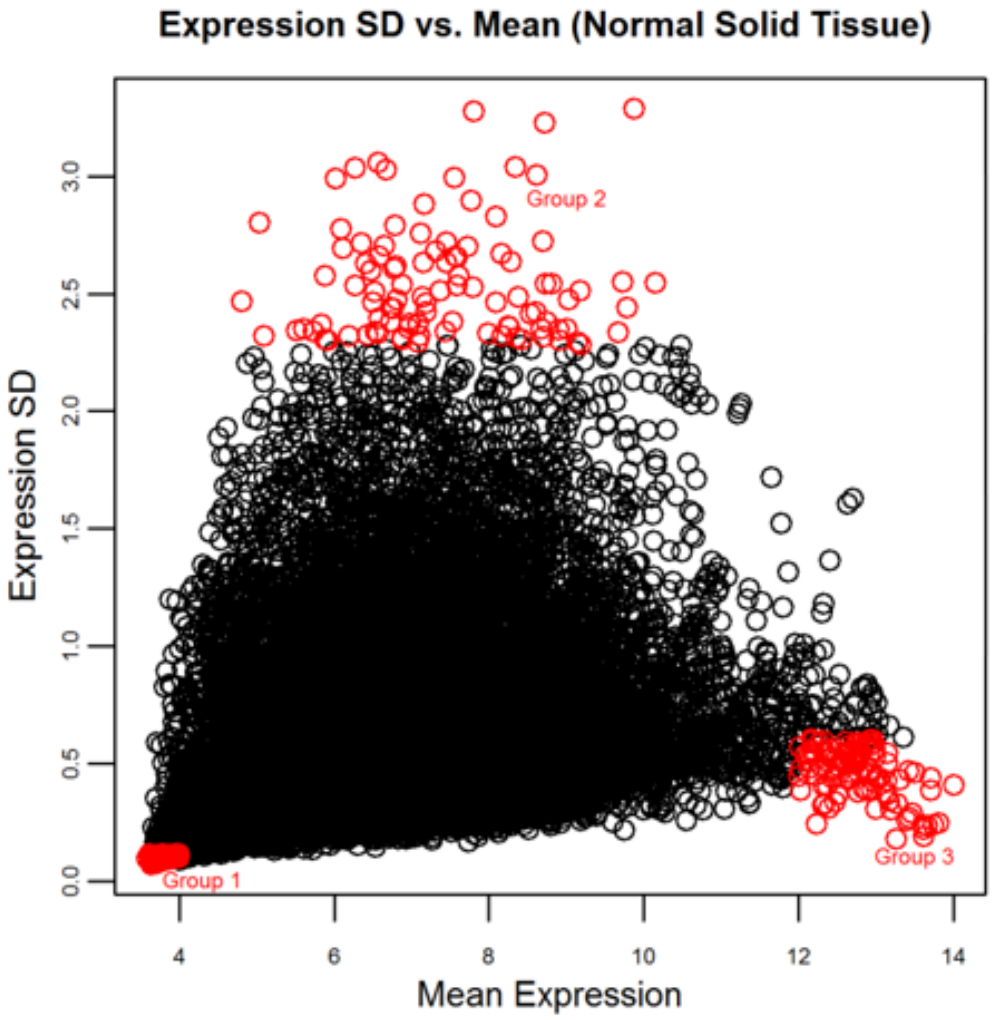
X_p


X_1
X_2
X_3
…
X_i
…
X_p

Std dev vs mean

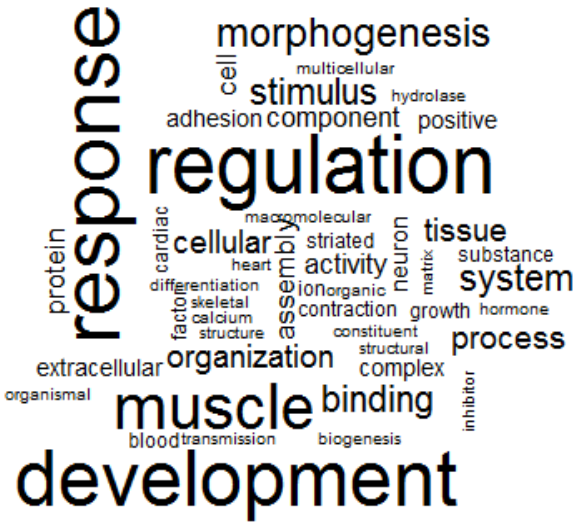


Std dev vs mean

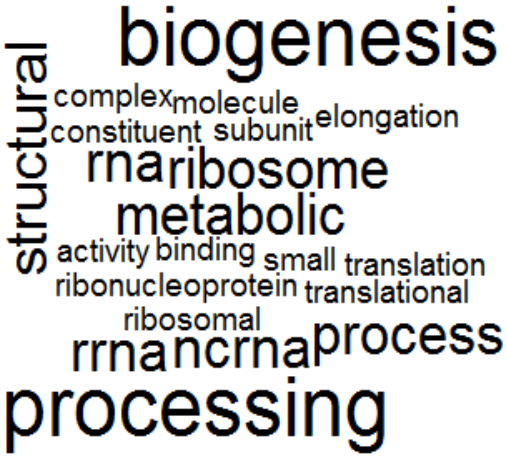


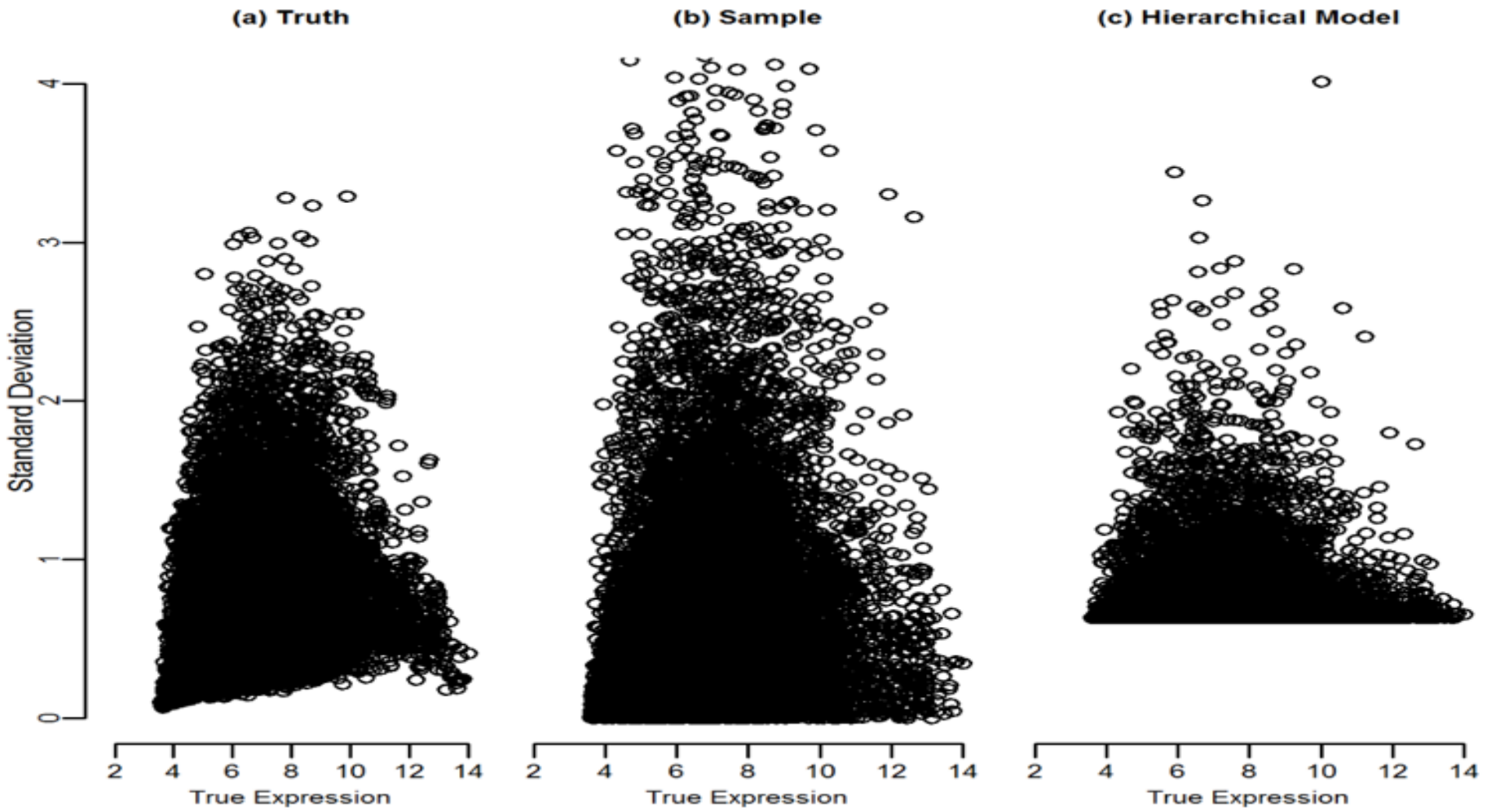
Diverse functions

Group 2



Group 3

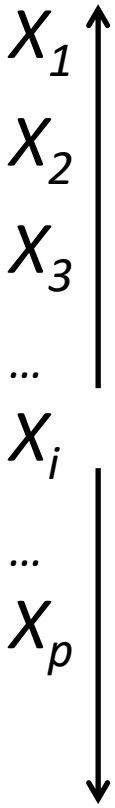



(a) Truth
(b) Sample
(c) Hierarchical Model
Standard Deviation
0
1
2
3
4
2
4
6
8
10
12
14
True Expression

Drawbacks of hierarchical models

- Restrict to current dataset.
- May overcorrect, especially at the lower end.
- Inflated variance means much less discovery power—conservative.

Informative prior derived from historical data


X_1
X_2
X_3
…
X_i
…
X_p

But why not this way?

X_1

X_2

X_3

…

X_i $X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i},\ldots$

…

X_p

A microarray compendium

CORRESPONDENCE

A global map of human gene expression

To the Editor:
Although there is only one human genome sequence, different genes are expressed

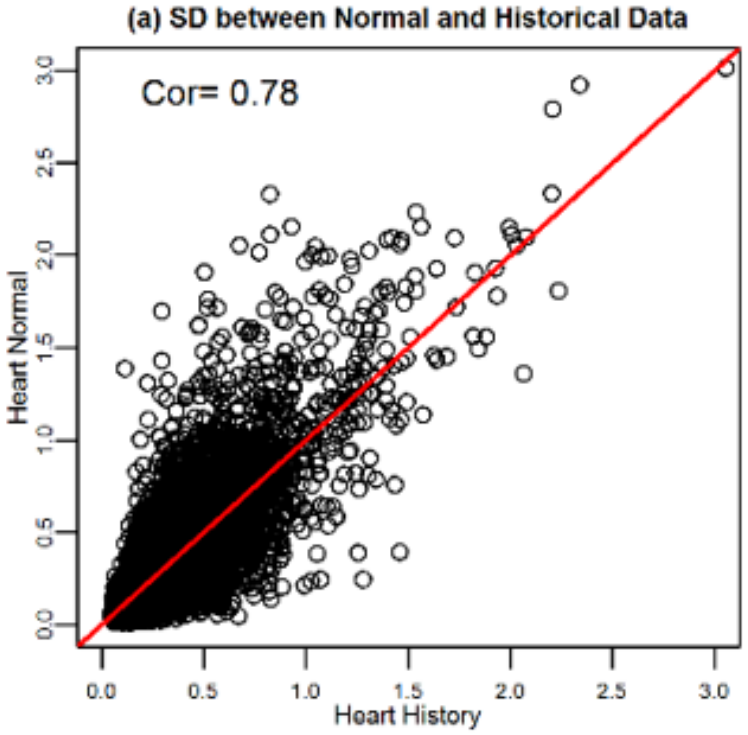
- 5,372 samples
- 206 different studies
- From 163 different labs

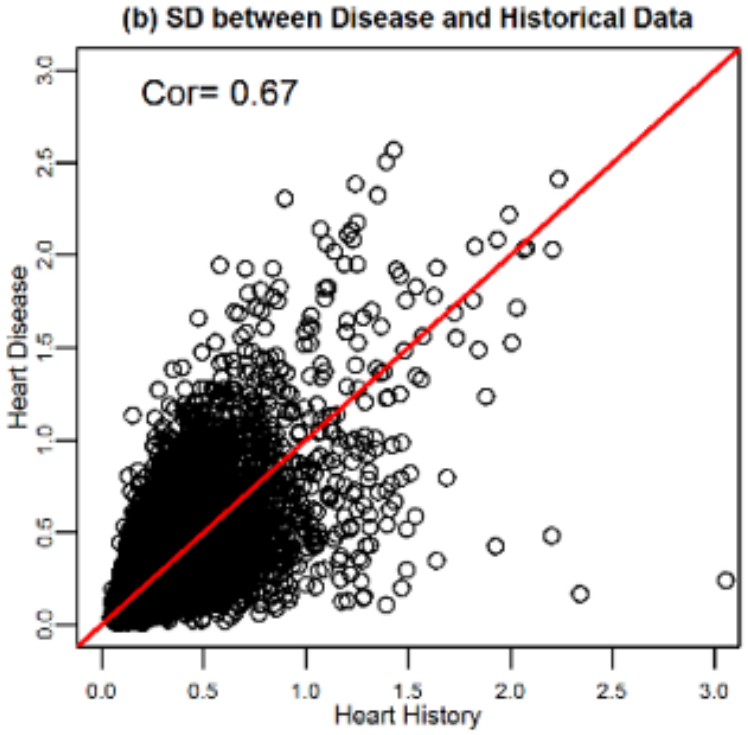
Lukk et al. 2010.

The global gene expression map

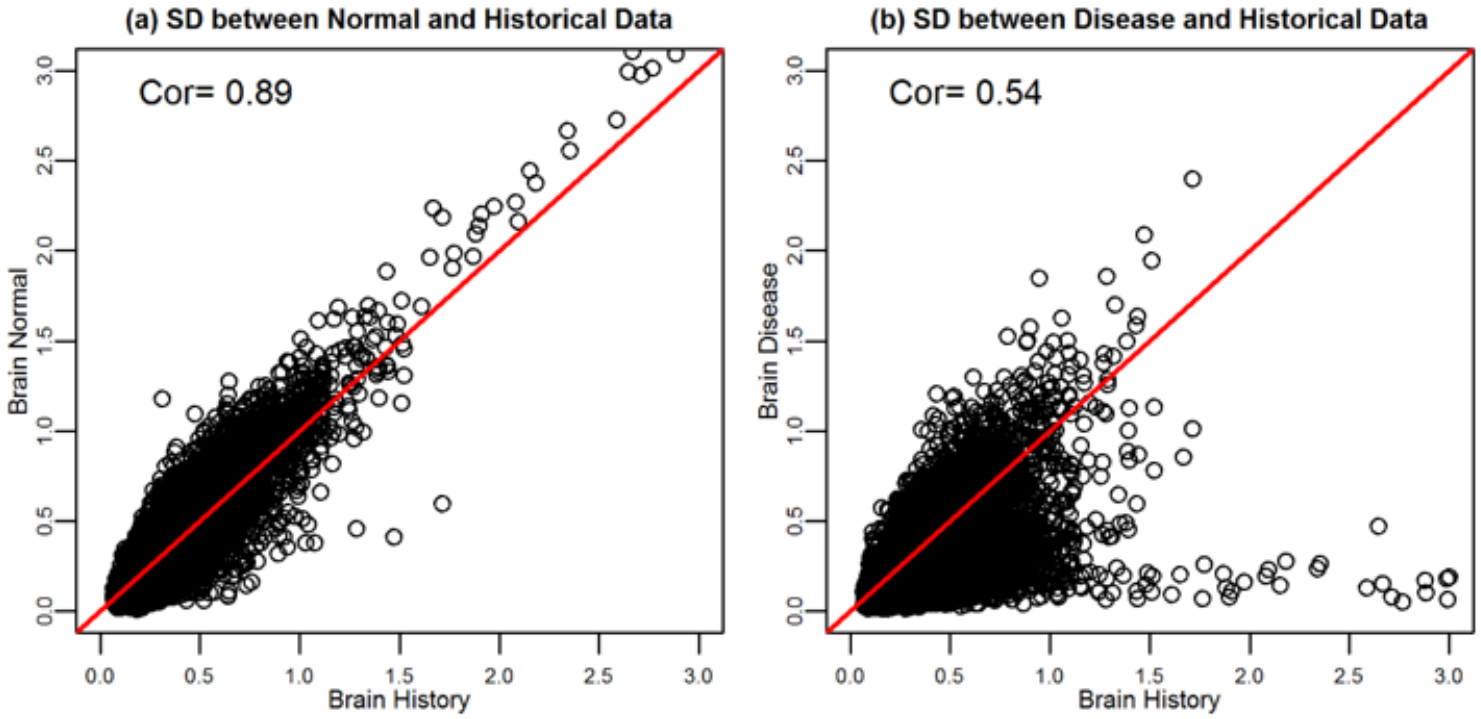
4 meta groups		15 meta groups	
Group	# of samples	Group	# of samples
cell line	1259	blood neoplasm cell line	166
		non neoplastic cell line	262
		solid tissue neoplasm cell line	831
disease	765	blood non neoplastic disease	388
		solid tissue non neoplastic disease	377
neoplasm	2315	breast cancer	672
		germ cell neoplasm	71
		leukemia	567
		nervous system neoplasm	112
		non breast carcinoma	288
		non leukemic blood neoplasm	334
		other neoplasm	167
		sarcoma	104
normal	1033	normal blood	467
		normal solid tissue	566

Standard deviations from different studies (heart)





Standard deviations from different studies (brain)



Simple shrinkage with historical information

	Sample variance	historical variance	Weight W	Adjusted Variance
Gene 1	Var1	Var_hist1	Var_hist1/(Var_hist1 + Var1)	(1-W)*Var1+W*Var_hist1
Gene 2	Var2	Var_hist2	Var_hist1/(Var_hist1 + Var1)	(1-W)*Var2+W*Var_hist2
……	……	……	……	……
……	……	……	……	……
Gene 1000	Var1000	Var_hist1000	Var_hist1000/(Var_hist1000 + Var1000)	(1-W)*Var1000+W*Var_hist1000

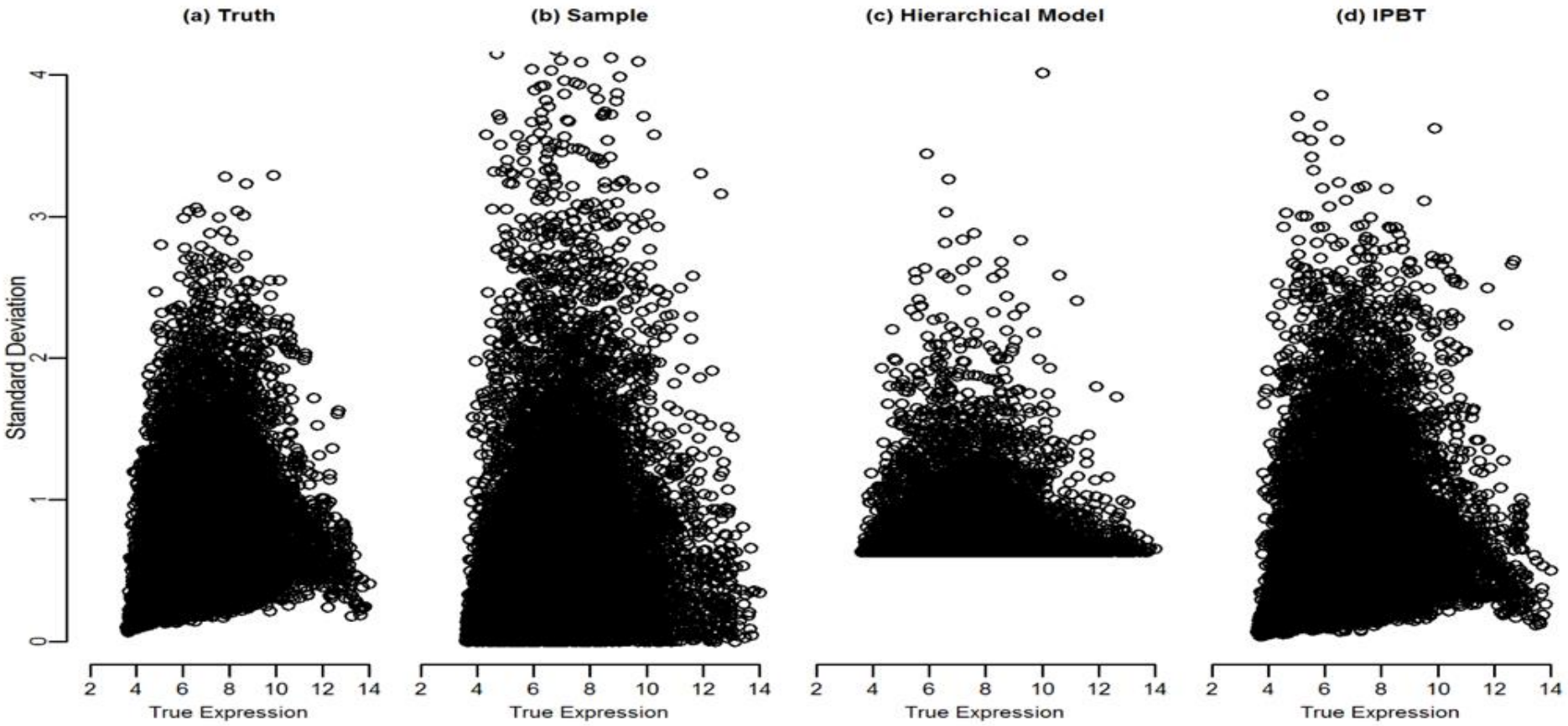
Combine historical information by simply doing a weighted average between historical variance and sample variance.

The weight is decided by the relative value of historical variance and sample variance

Informative Prior Bayesian Test (IPBT)

- Use historical data to build gene-specific, informative priors.
- Conduct Bayesian inference on σ_i, the standard deviation of gene *i*.
- Either calculate a Bayes factor or test statistics of an adjusted *t*-test and rank genes based on that.

Compare variance estimates



Methods compared

- Classical Student's *t*-test
- SAM
- Limma
- IPBT
- Z test

Simulation study

- 1,000 genes, each has a unique distribution

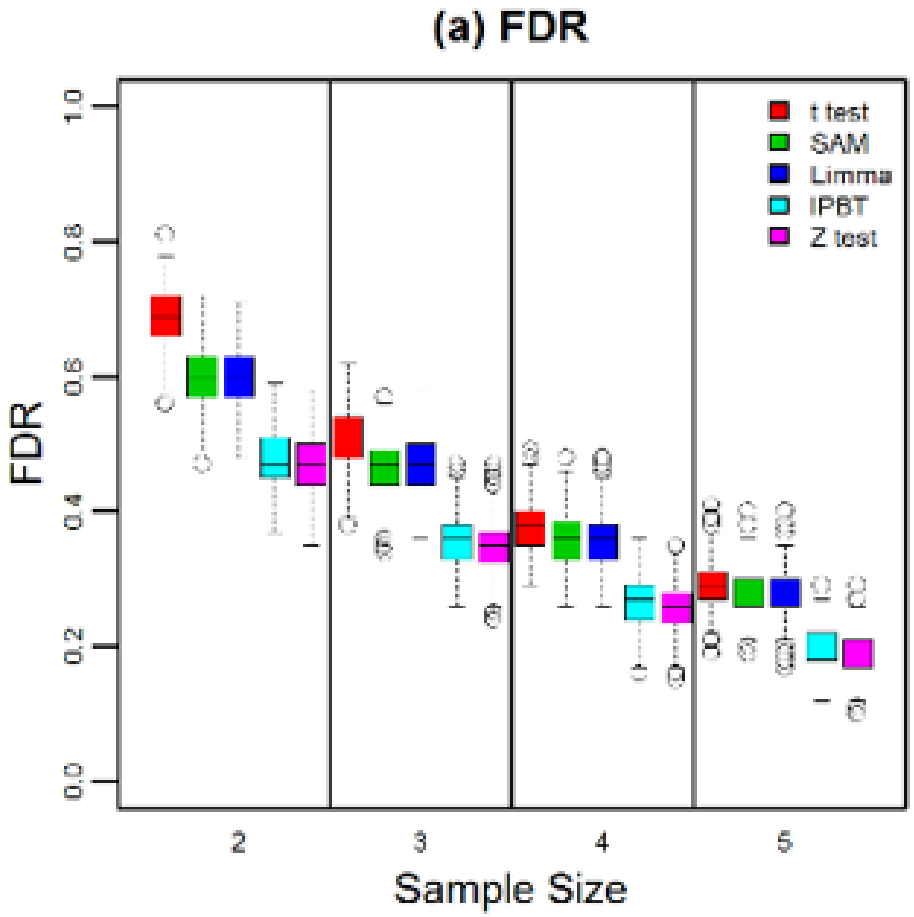
$$N(\mu_i, \sigma_i^2).$$

- 10% differentially expressed.
- All controls are sampled from $N(\mu_i, \sigma_i^2)$.
- 10% of treatment sampled from $N(\mu_i + 2\sigma_i, \sigma_i^2)$.
- 50 “historical datasets”.

Simulation Study

- Equal sample size for treatment/control is assumed, with k= 2, 3, 4, 5.
- Simulated runs were repeated 500 times for each setting. Each time calculate the False discovery rate (FDR) for each method.
- The boxplot for the 500 FDRs are plotted for each method.

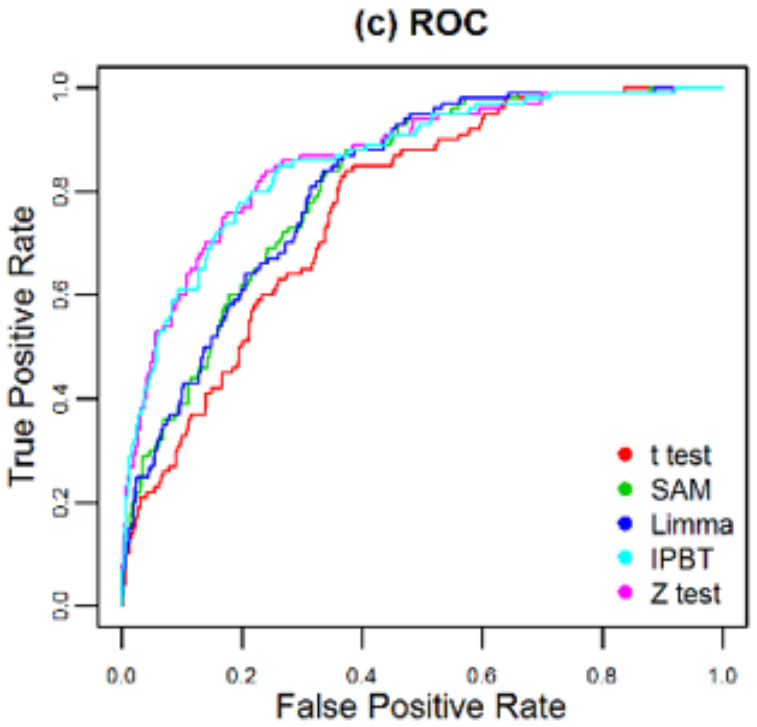
FDR boxplot


(a) FDR
t test
SAM
Limma
IPBT
Z test
FDR
0.0
0.2
0.4
0.6
0.8
1.0
2
3
4
5
Sample Size

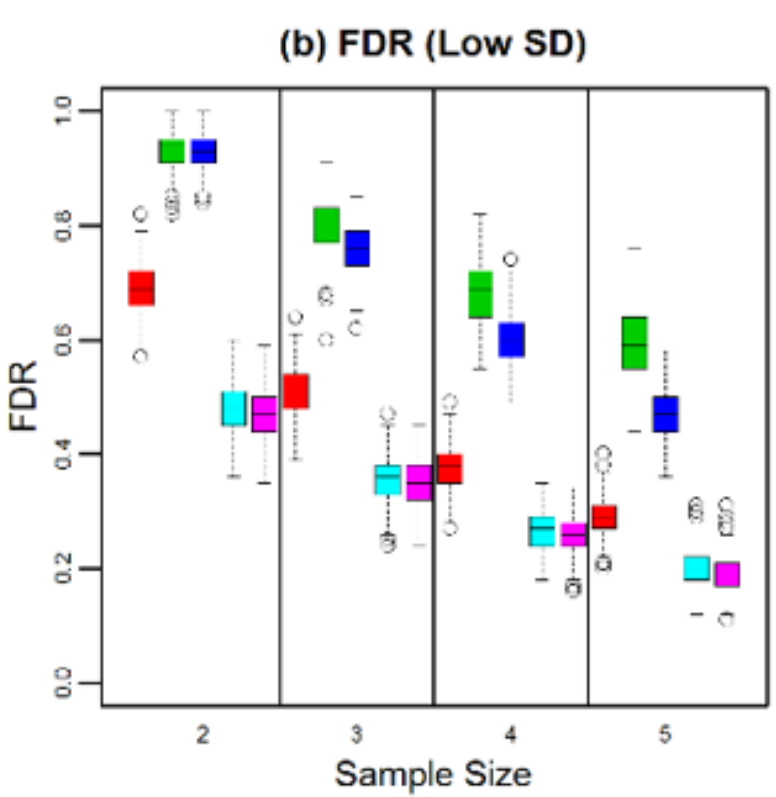
ROC curve for one simulation

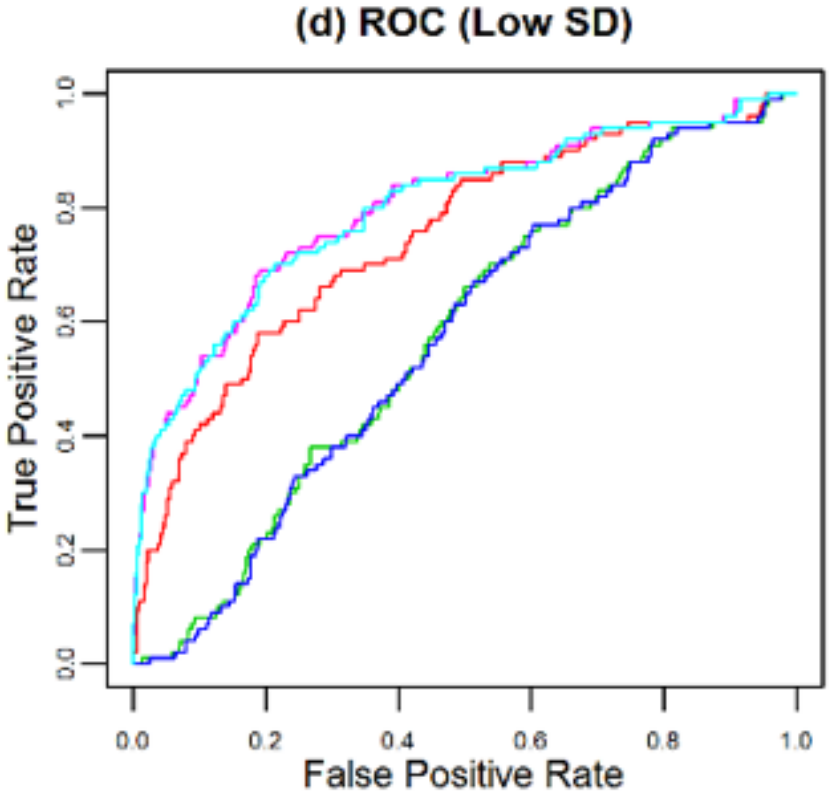
AUC for each method

Method	Random	Low Var
student's t-test	0.770	0.747
SAM	0.814	0.573
Limma	0.813	0.570
IPBT	**0.861**	**0.798**
Z test	0.864	0.800

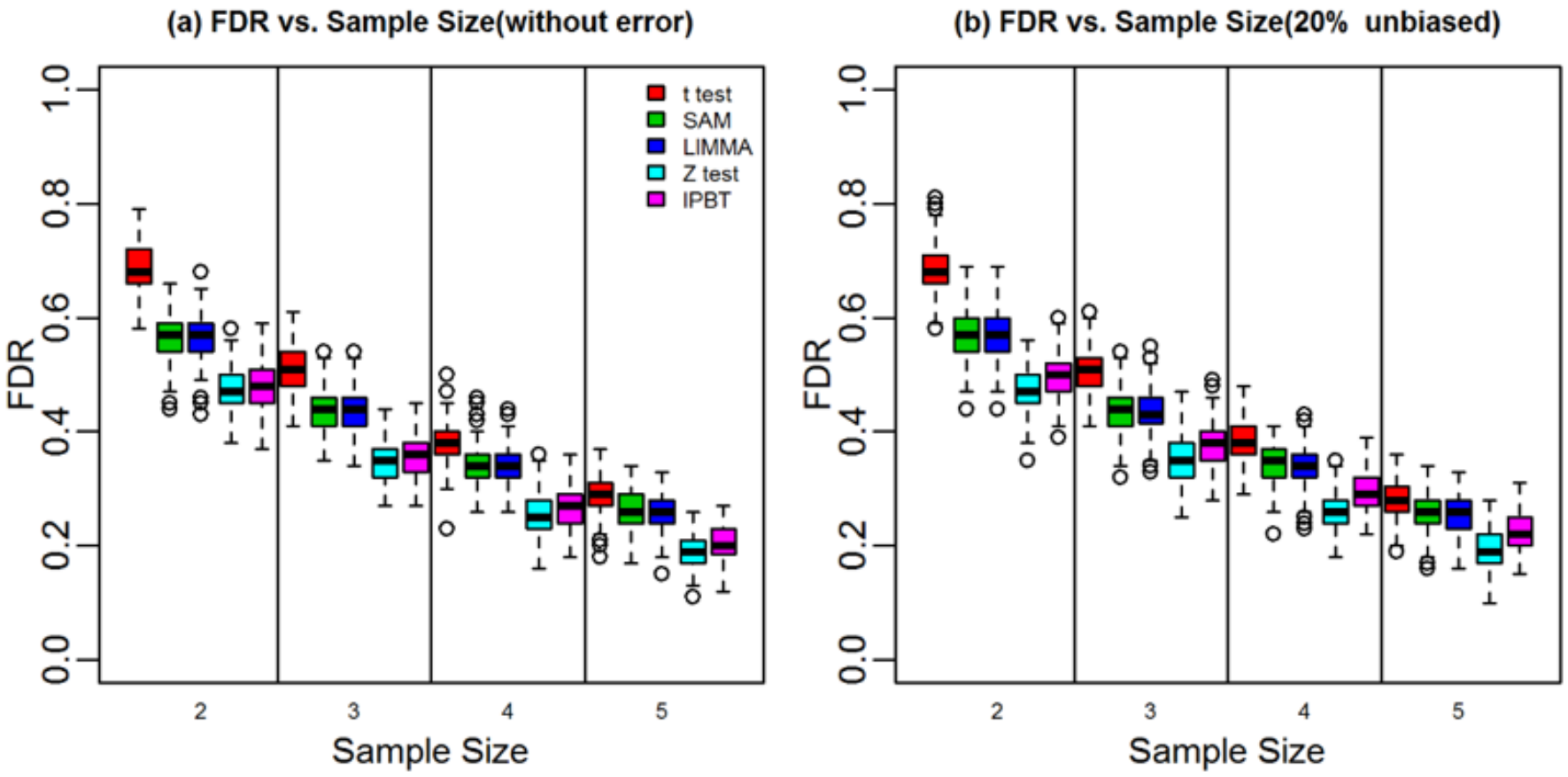


Low variance DE gene detection





When historical data is noisy



Real data analysis

- All the real data analysis used a global gene expression map of microarray data(U133A) from Lukk et al. (2010)
- All the microarray data are preprocessed (including normalization and summarization etc.) by robust multiarray analysis (RMA, Irizarry, Hobbs et al. 2003

Real data analysis

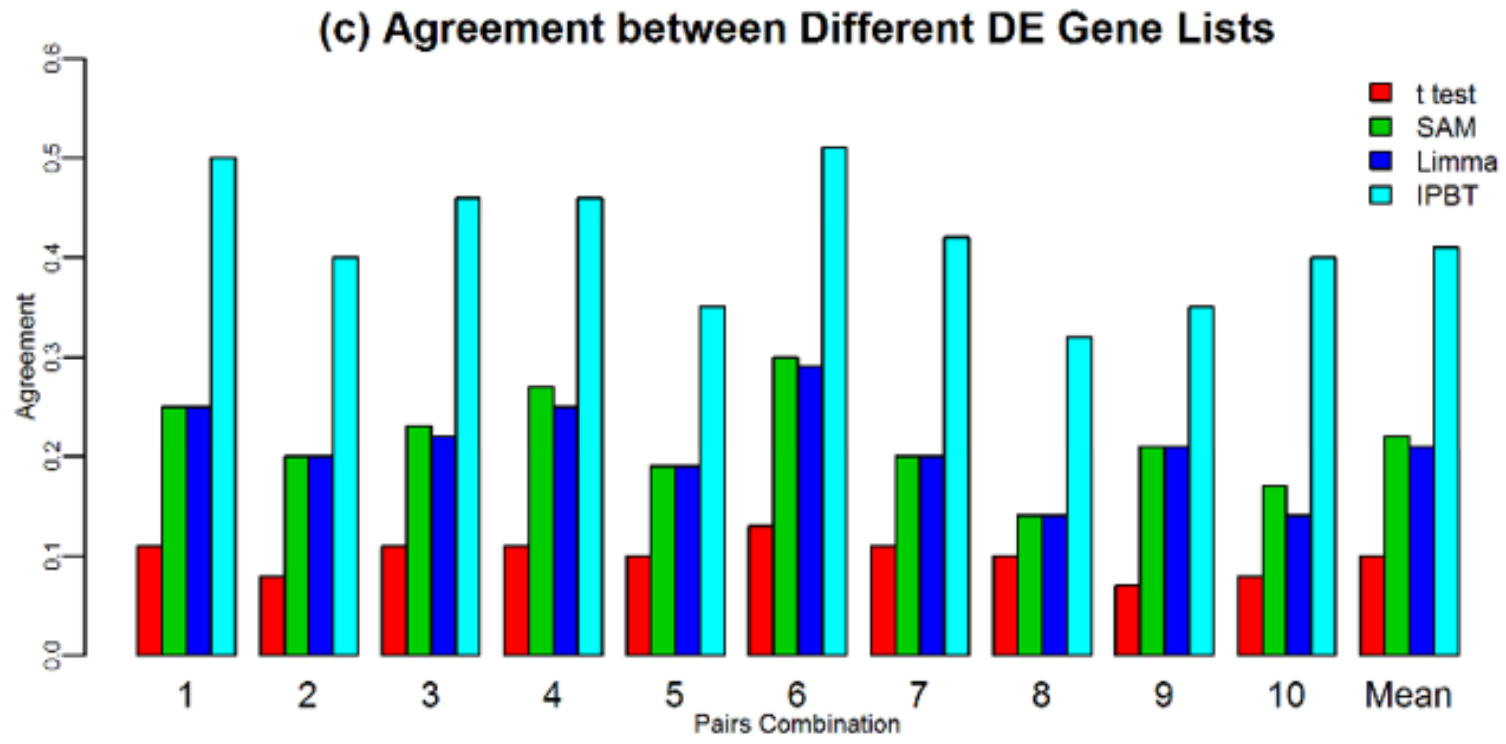
We conduct two real data analysis

- (1) Latin Square hgu133a spike-in experiment
- (2) Brain and heart data from the global gene expression map of microarray data

Real data (heart)

- Data on heart tissue
  - 36 normal (from 2 different studies)
  - 51 disease (from 4 different studies)
- Randomly select two samples from heart disease and normal samples, respectively, as the control and treatment data.
- The remainder 34 normal sample used to form historical data.
- Conduct tests and identify top 1000 DE genes.
- Repeat the sampling and testing procedures 5 times.
- Assess the agreement between every pair of the five DE gene lists.

Agreement evaluation using heart data



Summary

- Gene-specific properties such as variance can be captured by exploiting existing data that are public-available.
- Utilizing historical data in detecting differentially expressed genes is a better alternative than classical hierarchical model.
- Using informative prior can overcome difficulties faced in low-sample size inference problems.
- It is possible to reduce the number of replicates.

Reference

Bioinformatics, 2015, 1–8
doi: 10.1093/bioinformatics/btv631
Advance Access Publication Date: 30 October 2015
Original Paper

Gene expression

Bayesian inference with historical data-based informative priors improves detection of differentially expressed genes

Ben Li[1], Zhaonan Sun[2], Qing He[1], Yu Zhu[2,*] and Zhaohui S. Qin[1,3,*]

[1]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA, [2]Department of Statistics, Purdue University, West Lafayette, IN 47906, USA and [3]Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 30322, USA

*To whom correspondence should be addressed.
Associate Editor: Jonathan Wren

Received on April 15, 2015; revised on September 30, 2015; accepted on October 26, 2015

Li et al. *Bioinformatics* 2015.

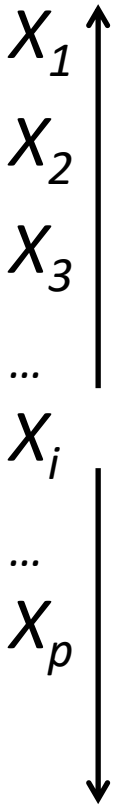
Partial utilization of the historical data

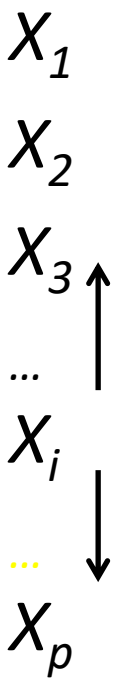
Limitations

- IPBT assumes that historical data is “similar” to the current data.
- Both historical data and current data have to come from the same platform.

Exchangeable

- A key assumption in hierarchical model
- Assume some kind of homogeneity among the features (genes in our context).
- However, this is often unrealistic
  - Genes are supposed to perform different functions hence have different properties.
- What can we do?
  - Overkill to borrow strength from all 25,000 genes
  - Just need a small subset


X_1
X_2
X_3
…
X_i
…
X_p


X_1
X_2
X_3
…
X_i
…
X_p

Two strategies

- Decompose genes into groups, such that genes in the same group are homogeneous. Apply hierarchical model within each group separately.
- For each individual gene, identify some of its "neighbors", and run hierarchical model among these neighboring genes.

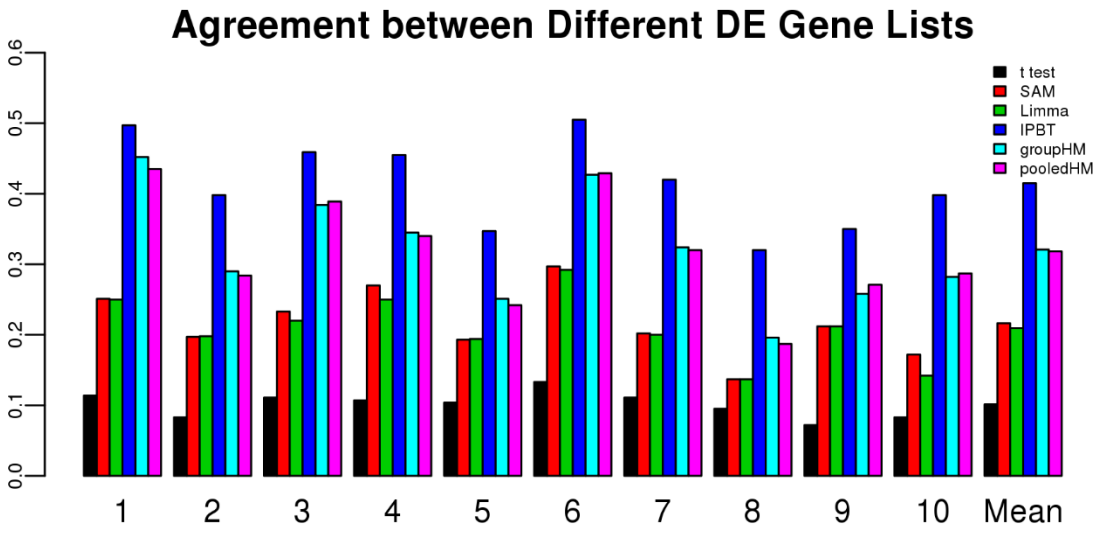
How to define groups?

- Use historical data
- Rank all genes using the variances estimated from historical data.

Real data (heart)

- Data on heart tissue
  - 36 normal (from 2 different studies)
  - 51 disease (from 4 different studies)
- Randomly select two samples from heart disease and normal samples, respectively, as the control and treatment data.
- The remainder 34 normal sample used to form historical data.
- Conduct tests and identify top 1000 DE genes.
- Repeat the sampling and testing procedures 5 times.
- Assess the agreement between every pair of the five results.

Real data (heart)



Summary

- Utilize historical data, but only a small part of them.
- The adaptiveHM Can be applied across platforms, e.g., use microarray historical data in RNA-seq analysis.
- Borrow strength both vertically and horizontally.

Stat Biosci
DOI 10.1007/s12561-016-9156-x

CrossMark

Improving Hierarchical Models Using Historical Data with Applications in High-Throughput Genomics Data Analysis

Ben Li[1] · Yunxiao Li[1] · Zhaohui S. Qin[1,2]