Identifying DE isoforms in an RNA-seq experiment

Survival-supervised latent Dirichlet allocation for genomics

Christina Kendziorski
Department of Biostatistics and Medical Informatics
University of Wisconsin-Madison

http://www.biostat.wisc.edu/~kendzior/

RNA-Seq: Advantages and Opportunities

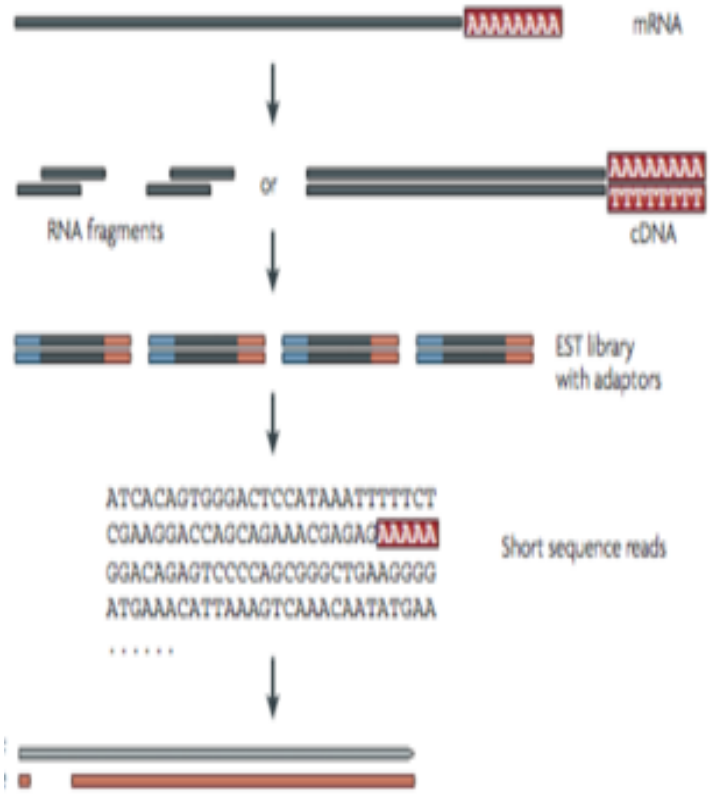
- Advantages
  - Low background noise
  - High resolution
  - Large dynamic range
  - Allele specific expression

- Opportunities
  - Splice junction identification
  - Novel transcript detection
  - Identification of DE genes and isoforms

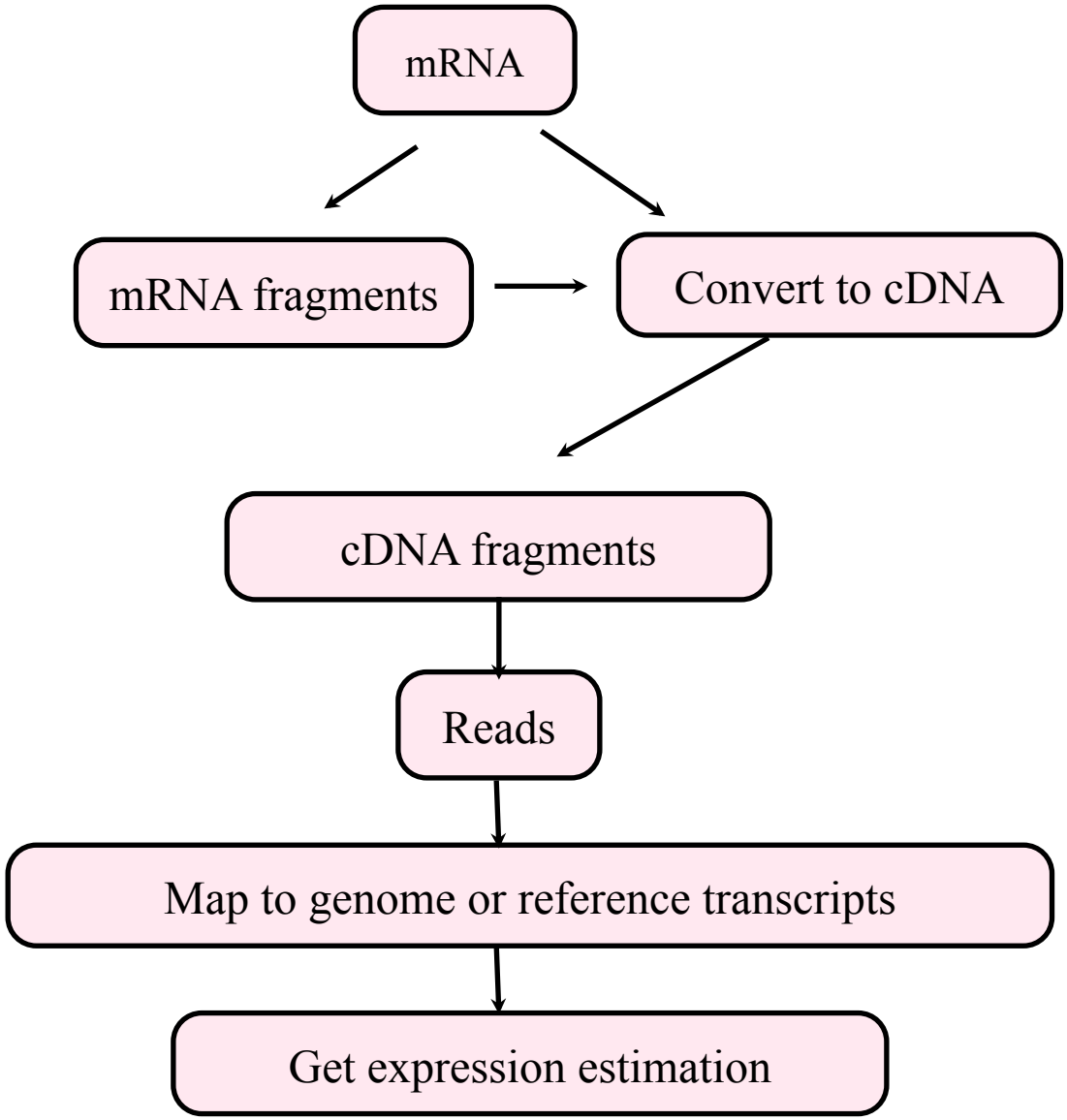
Outline

- Brief overview of RNA-seq data collection steps
- Motivation for isoform DE
- Methods for identifying DE genes do not work well
  - uncertainty in isoform expression estimation
  - isoform composition
- Quick fixes work pretty well…
  but not if there are outliers
- EBSeq for identifying DE isoforms and genes

RNA-Seq: Data Collection

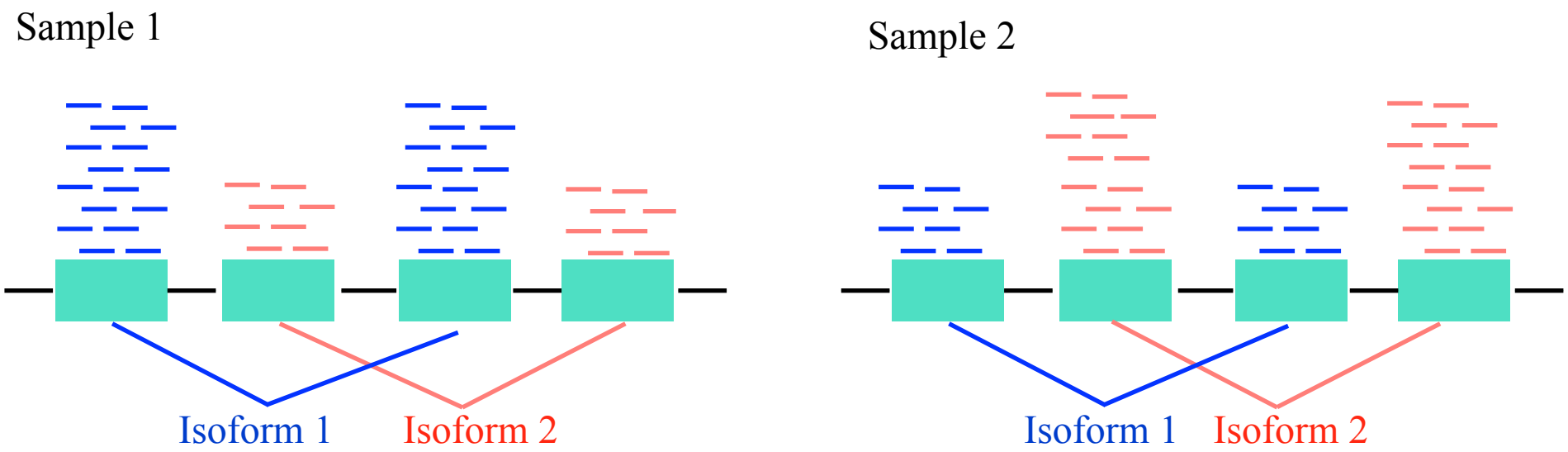


Wang, Gerstein, Snyder (2009)

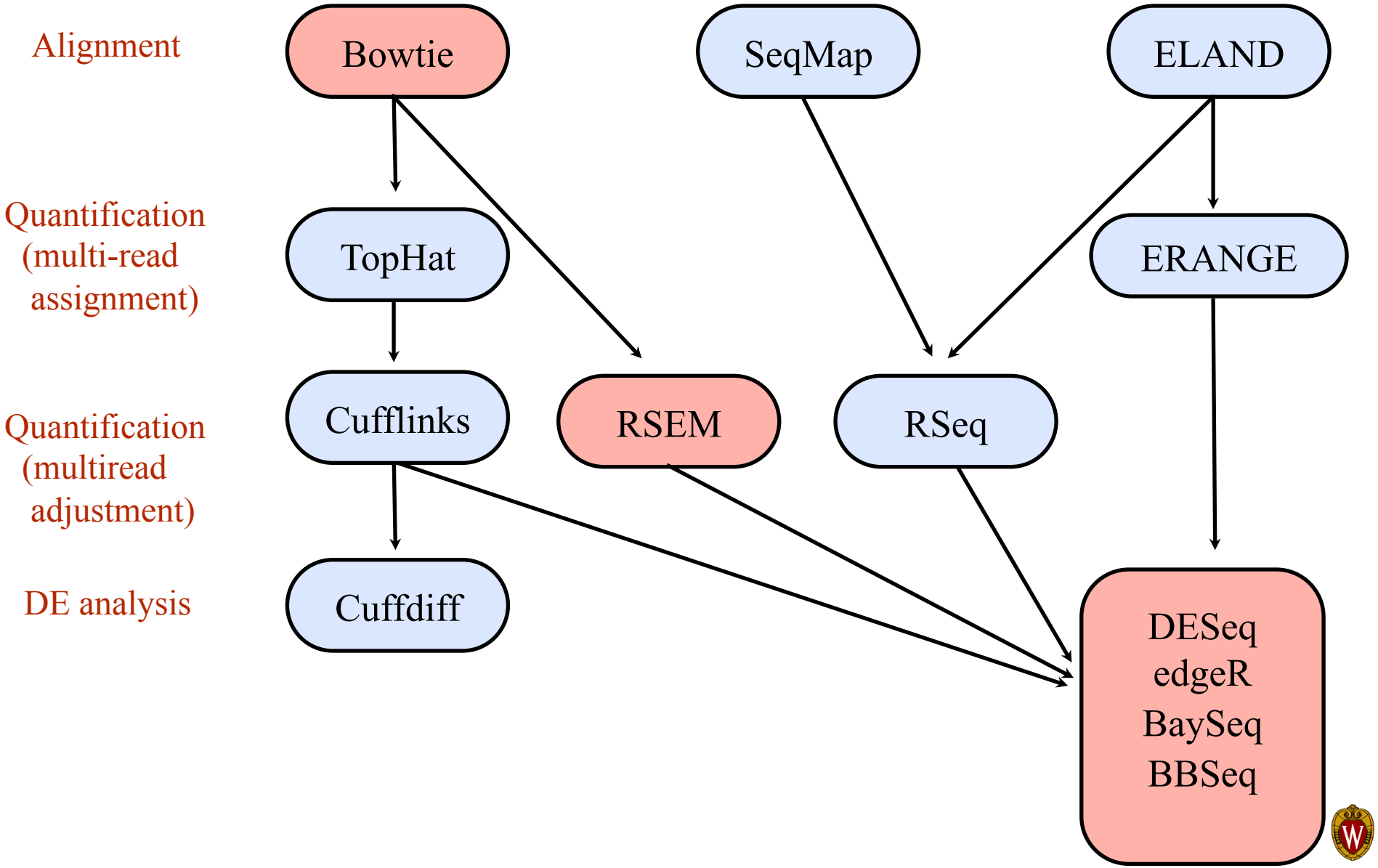


Isoform expression has important implications

- Alternative splicing is active in over 90% of human genes (Wang *et al*., Nature, 2008).
- AS variants from the same gene often have different biological functions
- A gene may be EE with DE isoforms

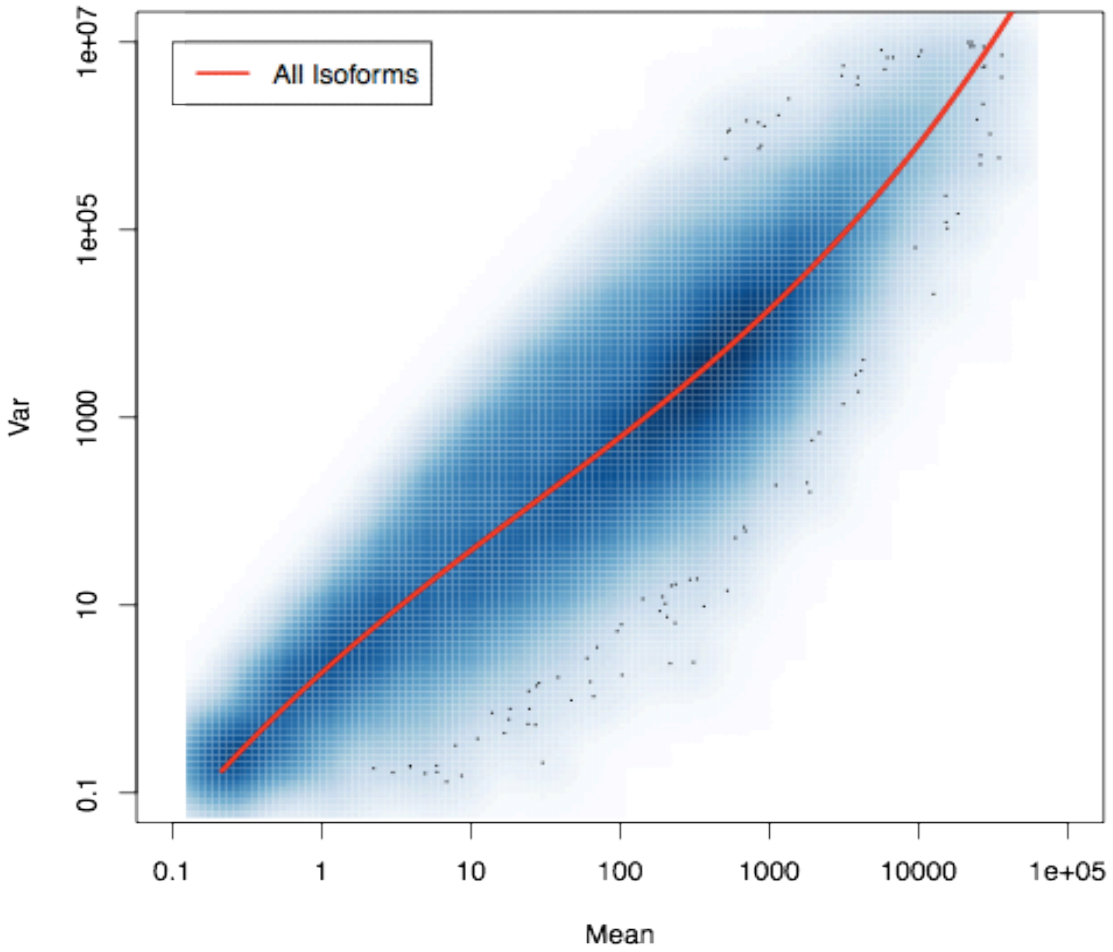


RNA-Seq: Methods



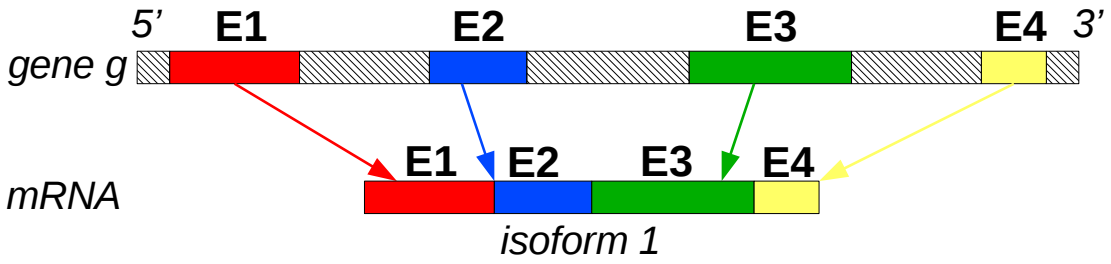
RNA-Seq: Methods for identifying DE genes

Most methods assume $X_{gs} \sim NB(r_g, q_g)$

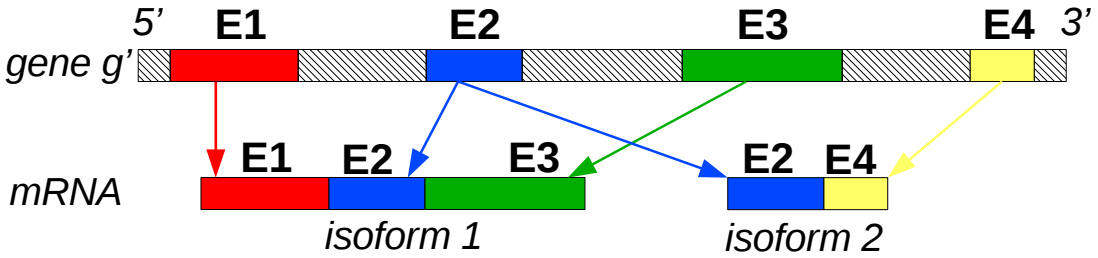


Mean-variance relationship changes with isoform complexity

N_g represents the number of isoforms of gene g

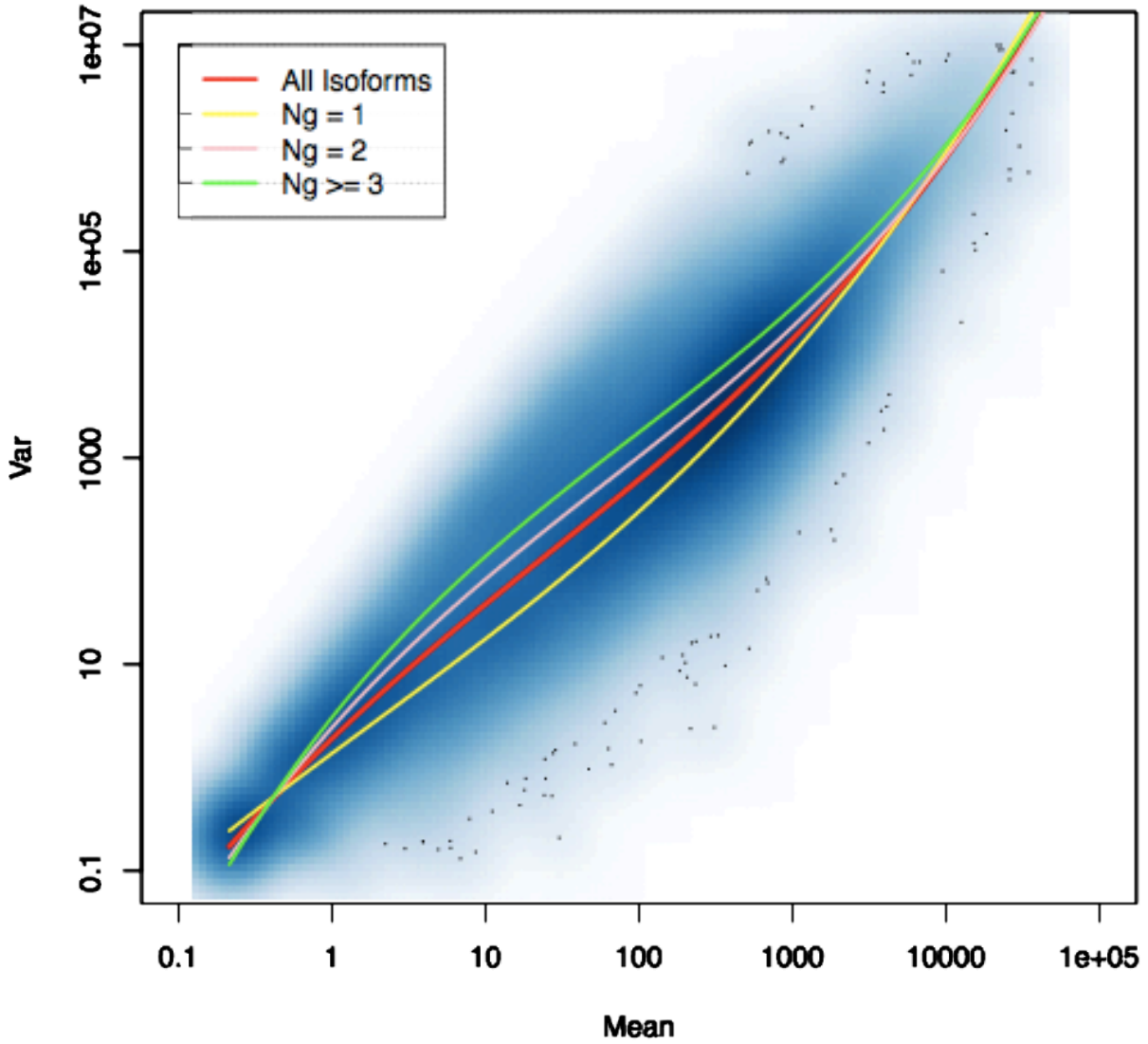


$N_g = 1$

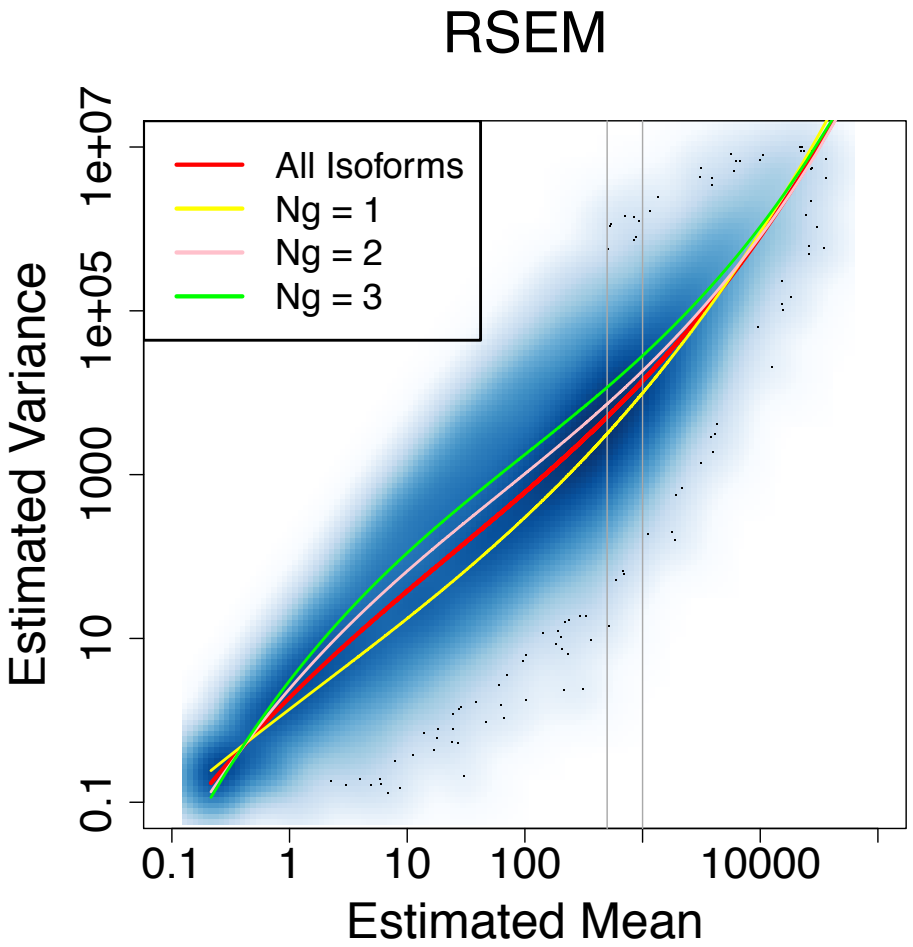


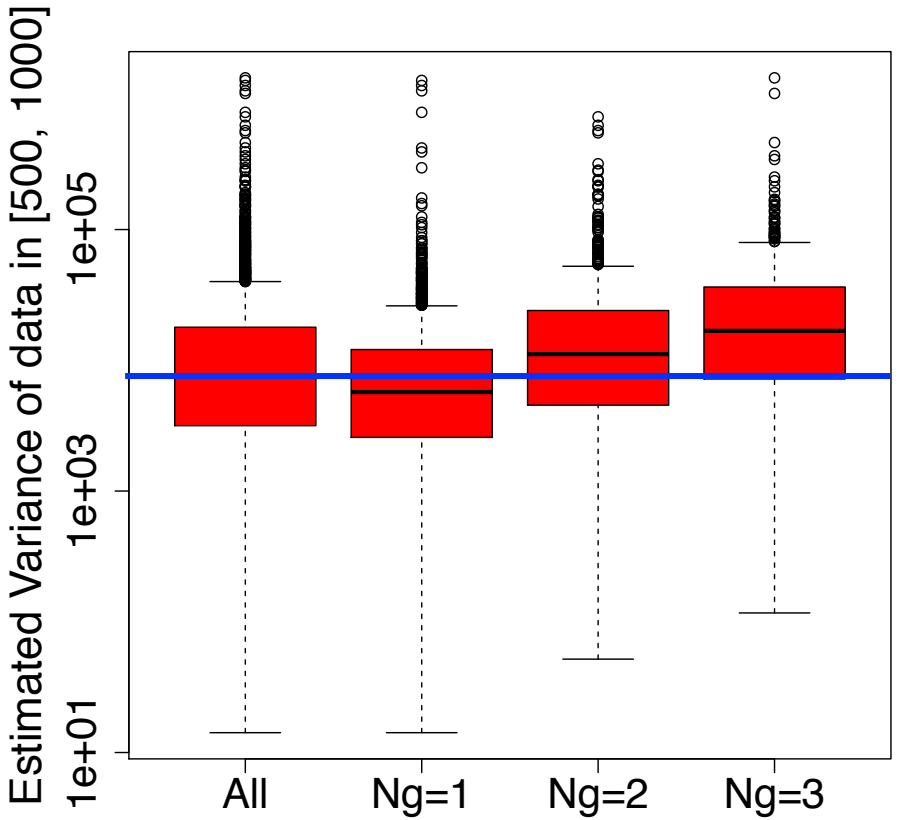
$N_{g'} = 2$

Mean-var relationship changes with N_g

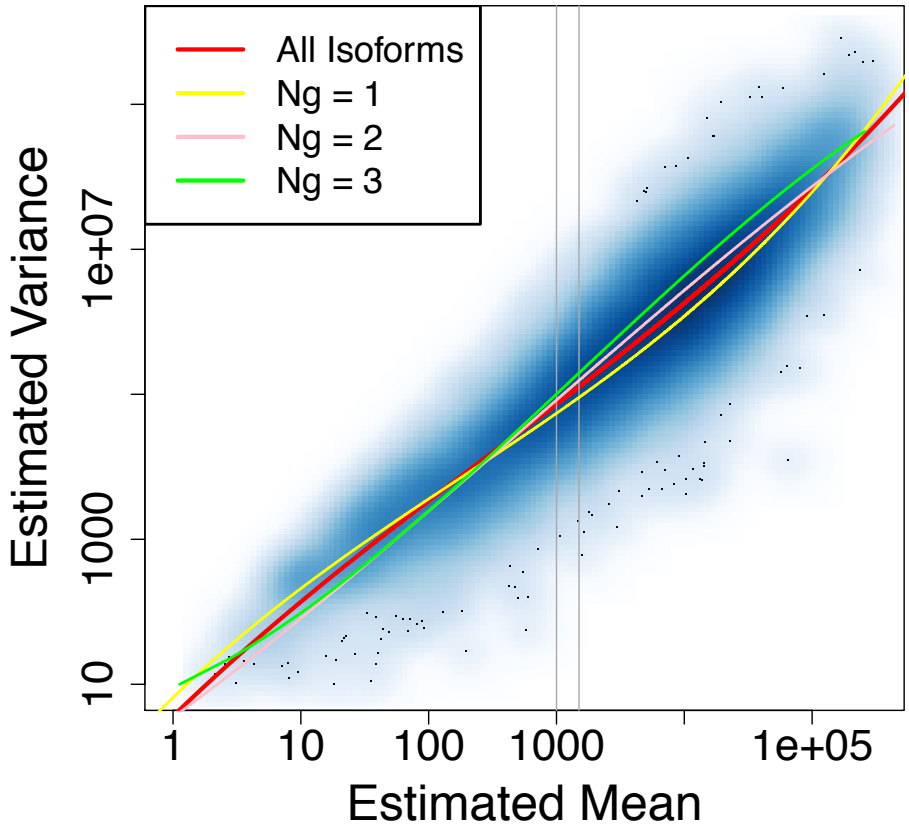


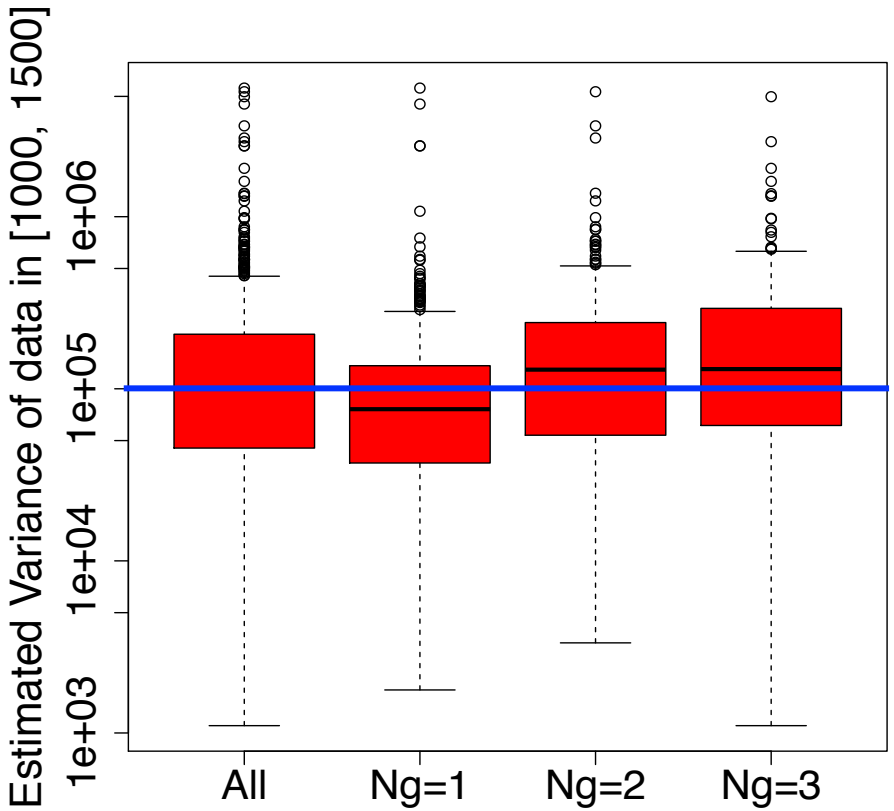
RSEM processed Gould lab data



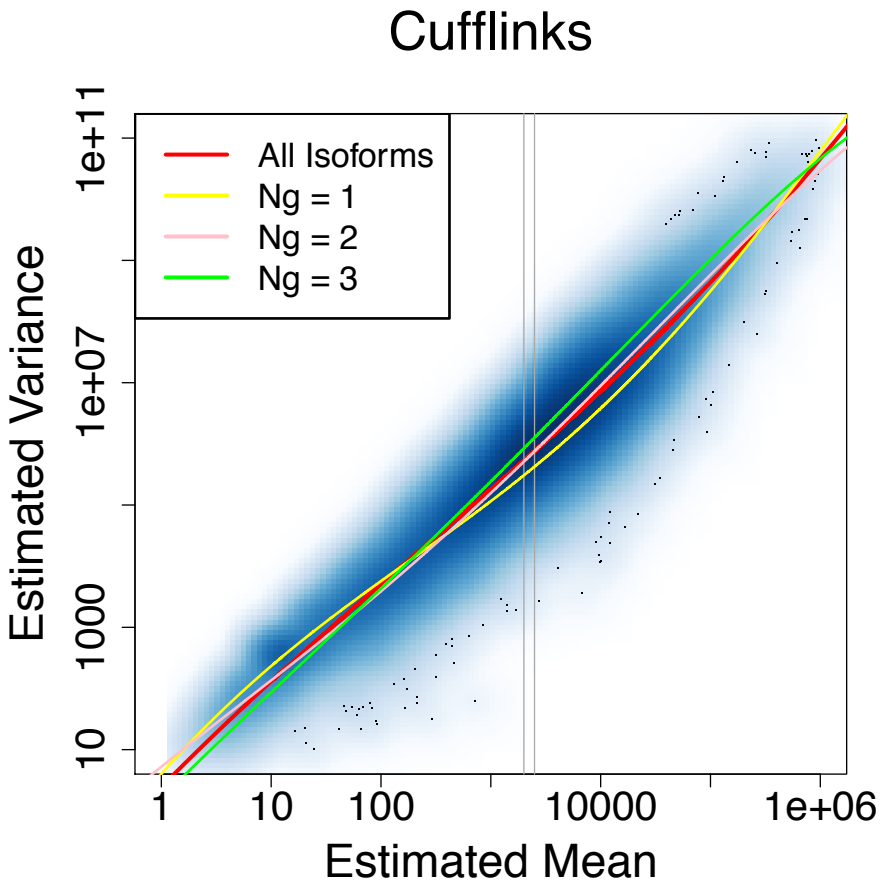


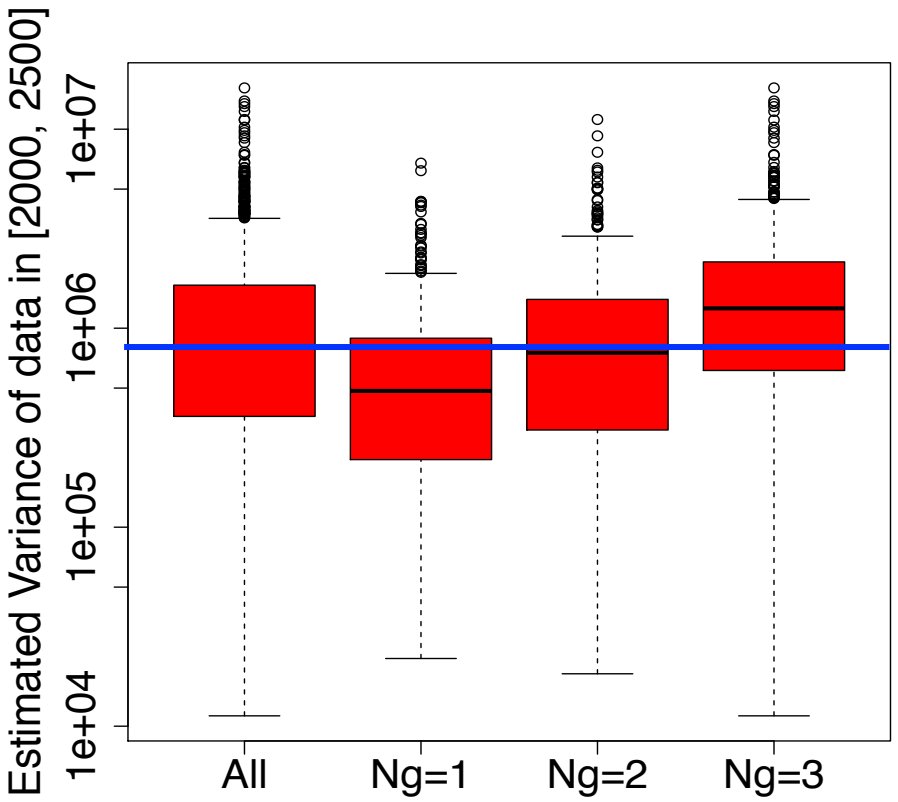
Cufflinks processed Gould lab data



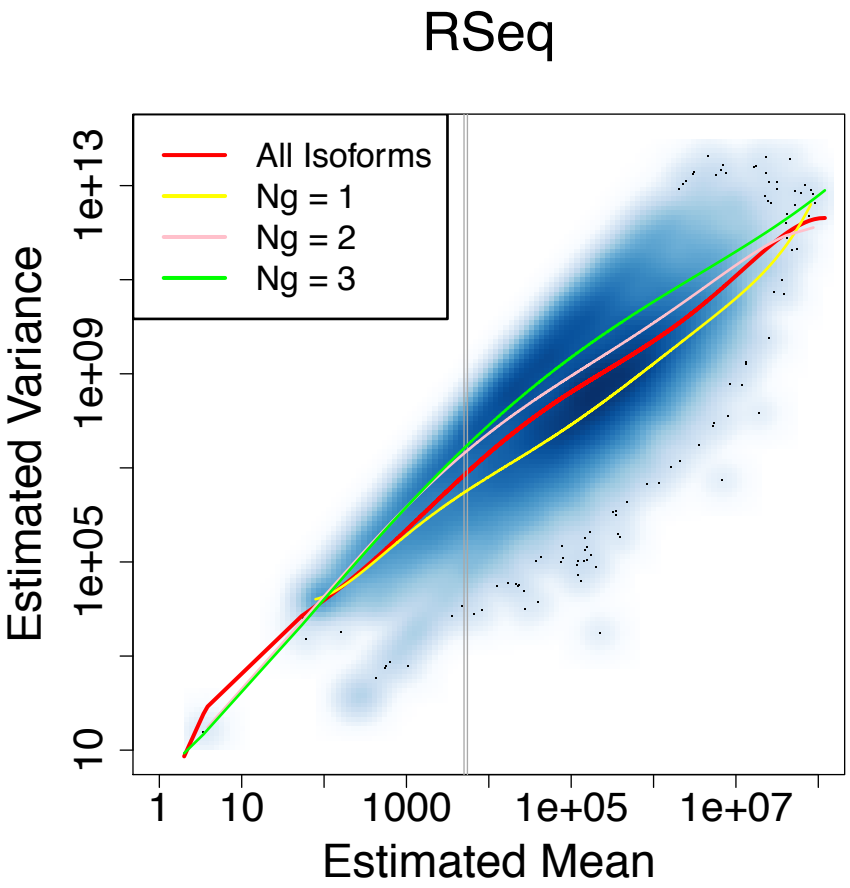


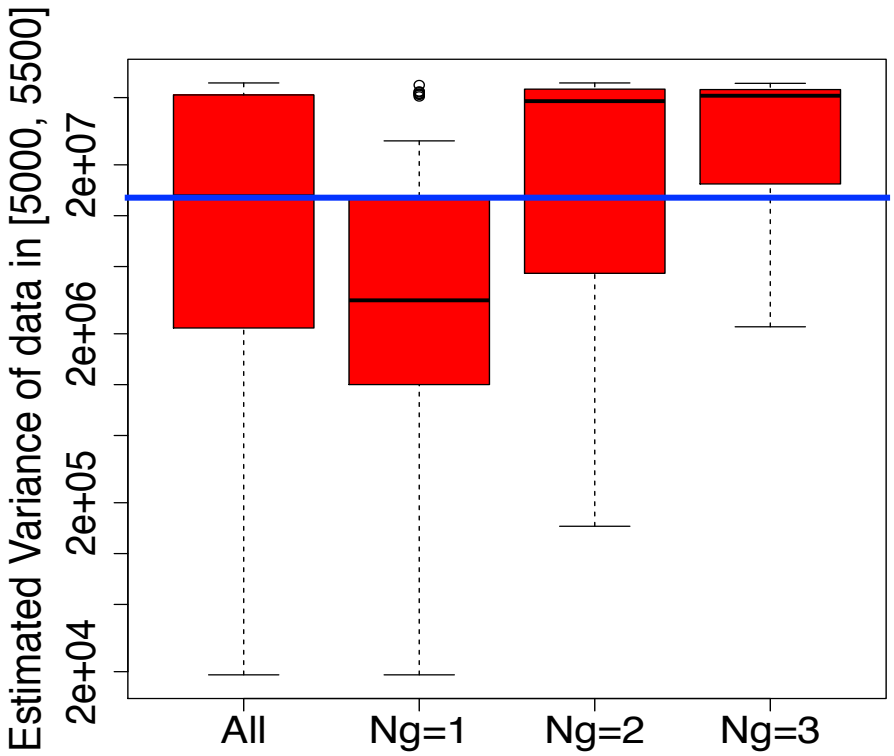
Cufflinks processed Hsu lab data





RSeq processed MAQC brain data





Expression levels are also isoform-class specific

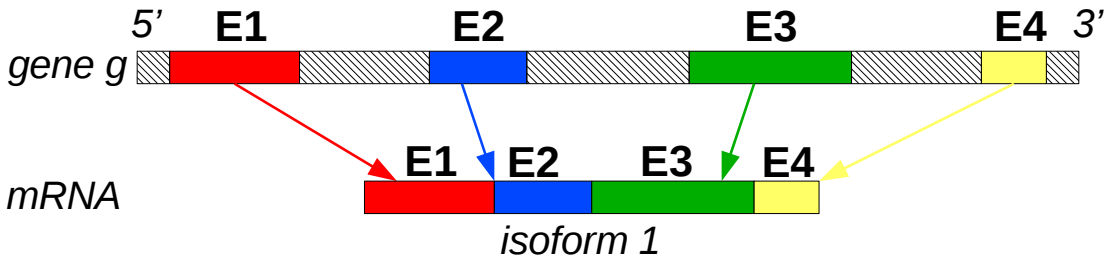
b_{gi} : presence/absence of 5’ and 3’ most exons (E1 and E4)

b_{gi}=1 3’ no 5’ (AD)

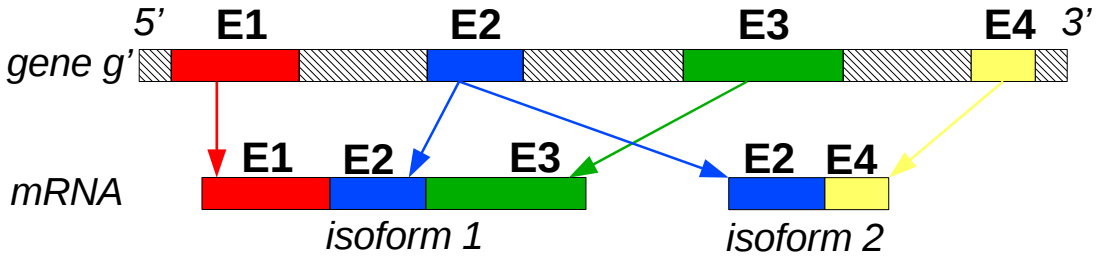
b_{gi}= 0 neither 5’ nor 3’ (AD and AA)

b_{gi}=2 5’ no 3’ (AA)

b_{gi}= 3 both 5’ and 3’



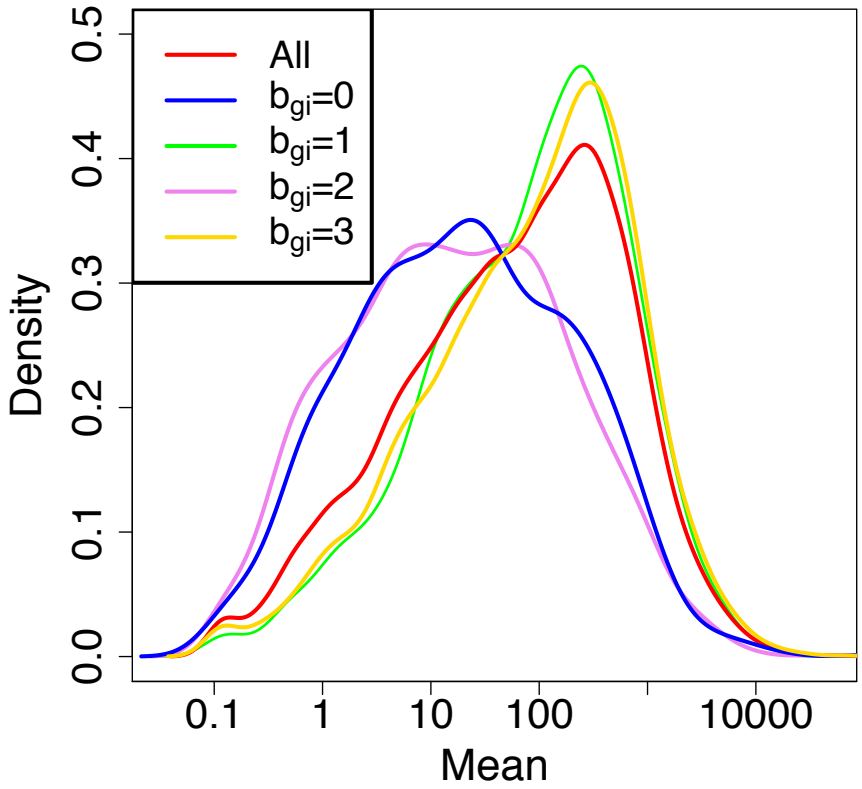
$N_g = 1, b_{gi} = 3$



$N_{g'} = 2, b_{g'i1} = 2$

$b_{g'i2} = 1$

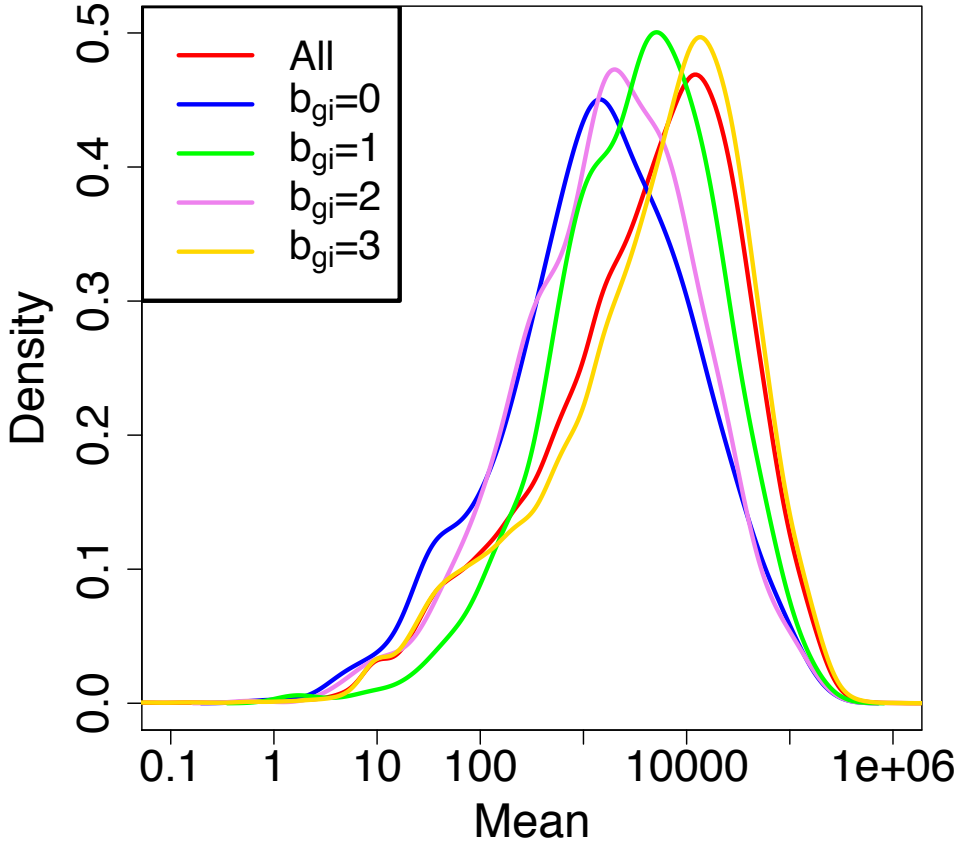
Means change with b_{gi} (RSEM processed Gould lab data)



All isoforms
Isoforms with 5' and 3'
Isoforms with AD
Isoforms with AA
Isoforms with AD and AA

Oligo-dT primed

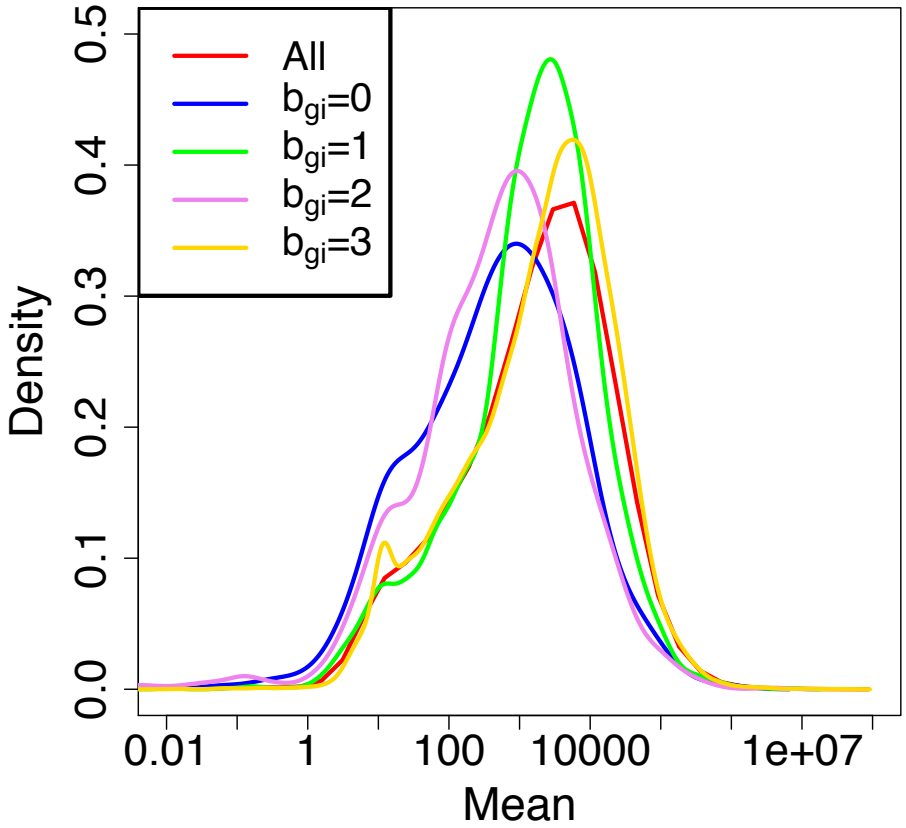
Cufflinks processed Gould lab data



All isoforms
Isoforms with 5' and 3'
Isoforms with AD
Isoforms with AA
Isoforms with AD and AA

Oligo-dT primed

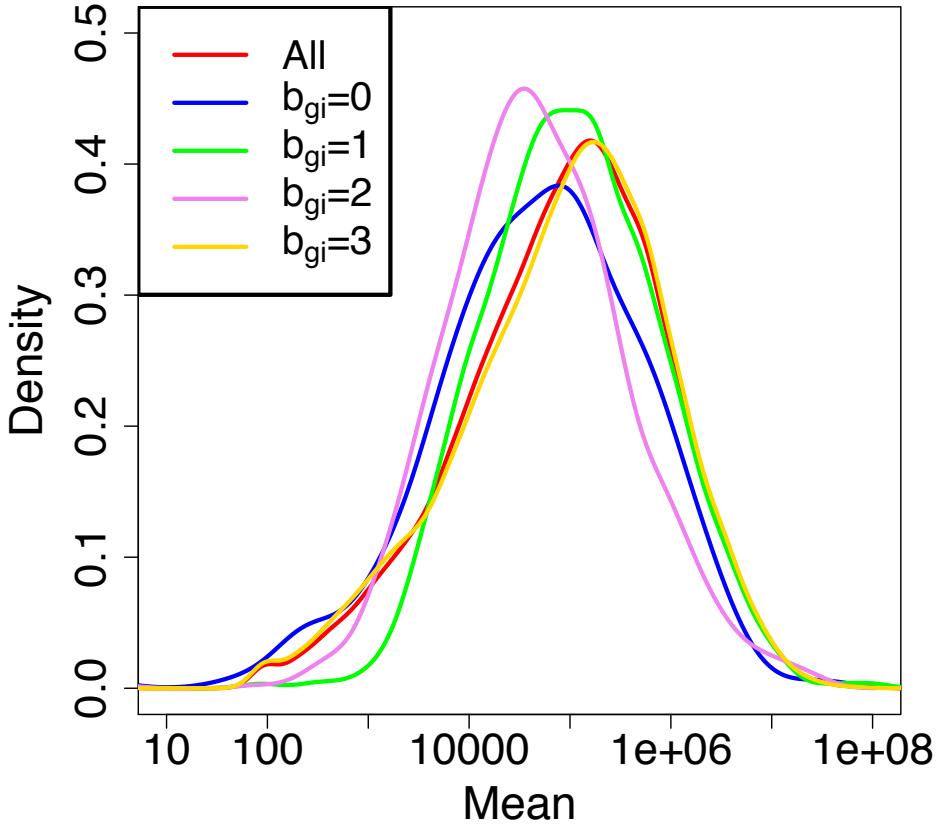
Cufflinks processed Hsu lab data



All isoforms
Isoforms with 5' and 3'
Isoforms with AD
Isoforms with AA
Isoforms with AD and AA

Random primed (?)

RSeq processed MAQC brain data



All isoforms
Isoforms with 5' and 3'
Isoforms with AD
Isoforms with AA
Isoforms with AD and AA

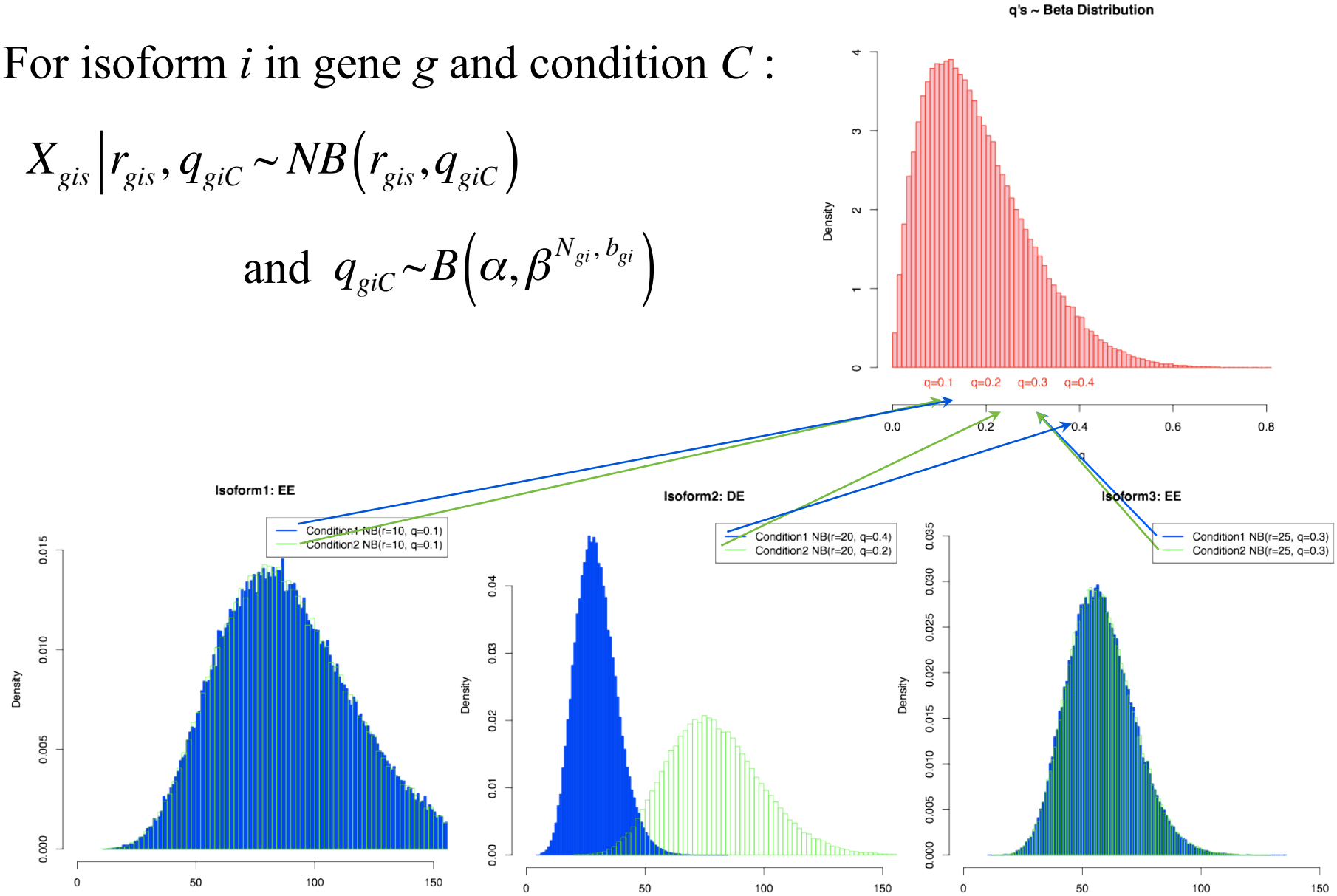
Random primed

EBSeq: An empirical Bayes NB-Beta Model

For isoform i in gene g and condition C :

$$X_{gis} \,|\, r_{gis}, q_{giC} \sim NB\left(r_{gis}, q_{giC}\right)$$

and $$q_{giC} \sim B\left(\alpha, \beta^{N_{gi}, b_{gi}}\right)$$



EBSeq

s : Sample

g : Gene

i : Isoform

l_s : Library size parameter

$X_{gi,s}$: Expression of isoform i in gene g and sample s

$r_{gi,0}$: Isoform specific parameter shared by all samples

p_0 : The prior probability of being EE

p_1 : The prior probability of being DE

$$X_{gi,s} \mid r_{gi,s}, q_{gi}^{C} \sim NB(r_{gi,s}, q_{gi}^{C}) \equiv NB\left(\mu_{gi,s} = \frac{r_{gi,s}(1-q_{gi}^{C})}{q_{gi}^{C}}, \sigma_{gi,s}^{2} = \frac{r_{gi,s}(1-q_{gi}^{C})}{(q_{gi}^{C})^{2}}\right)$$

$$q_{gi}^{C} \mid \alpha, \beta^{N_{gi}, b_{gi}} \sim Beta(\alpha, \beta^{N_{gi}, b_{gi}}) \text{ and } r_{gi,s} = l_s \bullet r_{gi,0}$$

The isoform is EE if $q_{gi}^{C1} = q_{gi}^{C2}$ and DE if $q_{gi}^{C1} \neq q_{gi}^{C2}$; then $X_{gi} \sim p_0 f_0(X_{gi}) + p_1 f_1(X_{gi})$ where

EE: $$f_0(X_{gi}) = \int \prod_{X_{gi,s} \in X_{gi}} P(X_{gi,s} \mid r_{gi,s}, q) P(q \mid \alpha, \beta^{N_{gi}, b_{gi}}) dq$$

DE: $$f_1(X_{gi}) = \int \prod_{X_{gi,s} \in X_{gi}^{C1}} P(X_{gi,s} \mid r_{gi,s}, q) P(q \mid \alpha, \beta^{N_{gi}, b_{gi}}) dq \int \prod_{X_{gi,s} \in X_{gi}^{C2}} P(X_{gi,s} \mid r_{gi,s}, q) P(q \mid \alpha, \beta^{N_{gi}, b_{gi}}) dq$$

Of primary interest is $$P(DE \mid X_{gi}) = \frac{p_1 f_1(X_{gi})}{p_0 f_0(X_{gi}) + p_1 f_1(X_{gi})}$$

The gene level model is similar but with β shared by all the genes.

CK October 2011

Gene Level Simulation

- As in Robinson and Smith (2007), we assume :

$$X_{gi,s} \sim NB\left(\mu_{gi,s} = l_s \mu_{gi}^C,\ \sigma_{gi,s}^2 = l_s \mu_{gi}^C (1 + \mu_{gi}^C \phi_{gi}) \right)$$

  here, μ_{gi}^C and ϕ_{gi} are sampled (ϕ_{gi} within N_g group).

- 10 % DE where $\mu_{gi}^{C2} = \Delta \mu_{gi}^{C1}$; 4 replicates in each condition.

- DESeq, edgeR, baySeq and BBSeq are applied to all of the isoforms at once, and within each N_g group.

- Results are averaged across 100 simulations, with thresholds chosen to control FDR at 5%.

Results from simulation (gene-level)

	Power	FDR
baySeq	0.71	0
BBSeq	0.7	0.02
DESeq	0.91	0.22
edgeR	0.89	0.15
EBSeq	0.79	0.05

Results from simulation (isoform-level)

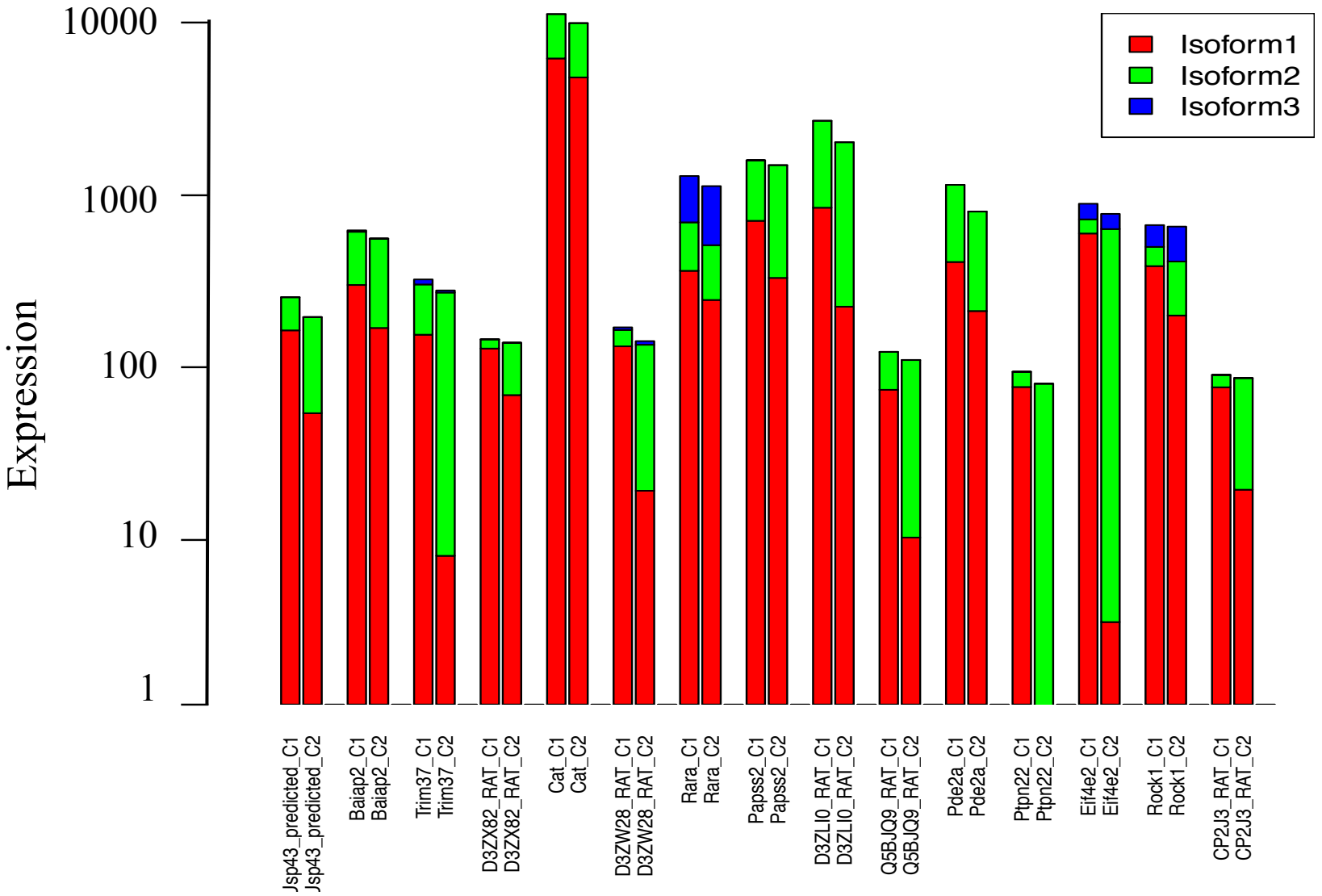
	Ng=1 Power	Ng=1 FDR	Ng=2 Power	Ng=2 FDR	Ng=3 Power	Ng=3 FDR
baySeq	0.64	0	0.62	0	0.55	0.01
baySeq Each	0.67	0	0.63	0	0.50	0.01
BBSeq	0.62	0.01	0.61	0.04	0.56	0.04
BBSeq Each	0.62	0.04	0.62	0.03	0.53	0.04
DESeq	0.78	0.02	0.86	0.24	0.89	0.29
DESeq Each	0.80	0.08	0.77	0.07	0.74	0.07
edgeR	0.79	0.02	0.86	0.18	0.88	0.24
edgeR Each	0.80	0.09	0.76	0.06	0.72	0.07
EBSeq	0.70	0.05	0.73	0.07	0.70	0.08

Results from simulation with outliers (isoform-level)

As before, but with a single value *x* redefined as 10**x*

	Ng=1 Power	Ng=1 FDR	Ng=2 Power	Ng=2 FDR	Ng=3 Power	Ng=3 FDR
baySeq	0.5	0	0.52	0.01	0.43	0.02
baySeq Each	0.61	0.03	0.41	0.01	0.43	0.03
BBSeq	0.62	0.01	0.59	0.02	0.52	0.02
BBSeq Each	0.61	0.02	0.45	0.02	0.51	0.03
DESeq	0.73	0.44	0.82	0.47	0.83	0.47
DESeq Each	0.76	0.47	0.72	0.43	0.66	0.41
edgeR	0.77	0.28	0.83	0.35	0.84	0.36
edgeR Each	0.79	0.41	0.73	0.27	0.69	0.34
EBSeq	0.71	0.04	0.73	0.08	0.69	0.08

DE isoforms in EE genes – Gould lab data



Summary

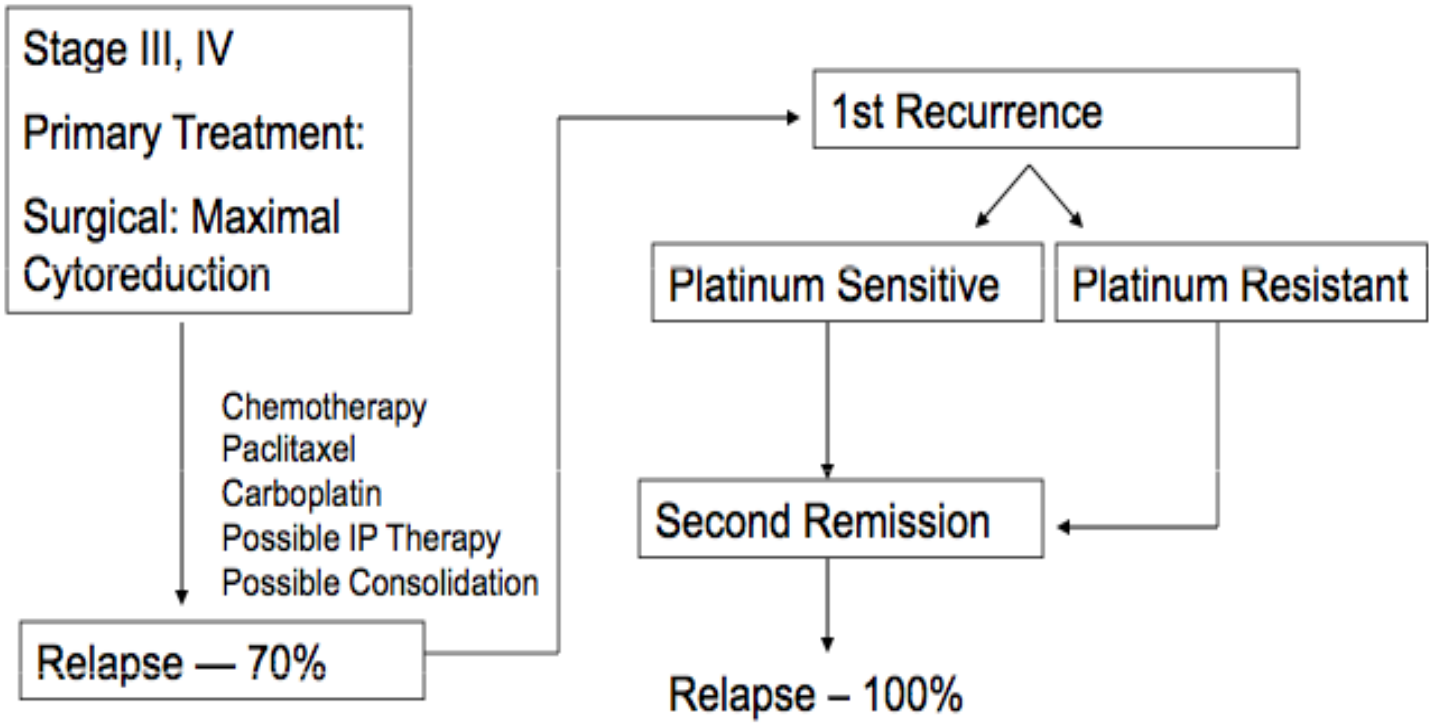
- Methods for identifying DE genes do not work well when applied directly to isoforms as they do not accommodate uncertainty in isoform expression estimation and other structure.
- Applying within N_g group works well unless there are outliers.
- EBSeq identifies both DE genes and isoforms, accommodates uncertainty and some biases, and is fairly robust to outliers; ….can be used without mixing over b_{gi} .

Using LDA to tell the story of cancer

joint work with John Dawson

Ovarian Cancer Overview

- 5th leading cause of death among American women
- ~22,000 new cases in 2010 with ~14,000 deaths
- 5 year survival < 50%.
- Protocol for secondary treatment not clear



The Cancer Genome Atlas Project

Clinical

mRNAs

miRNAs

Methylation

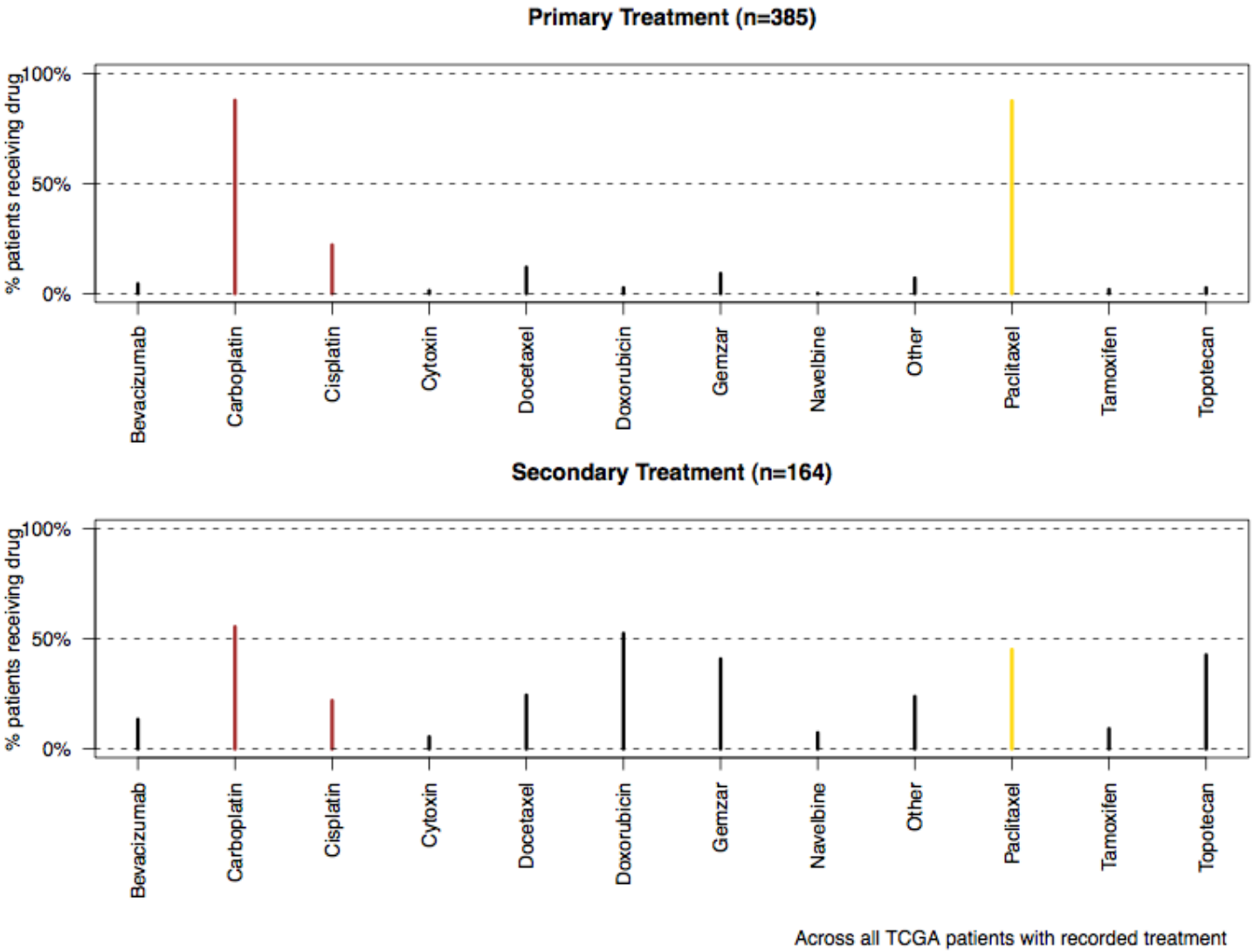
CNVs

SNPs

…to understand the molecular basis of cancer and thereby improve our ability to diagnose, treat, and prevent this disease.

…started in 2005 with ovary, lung, and brain

Can we guide secondary treatment ?



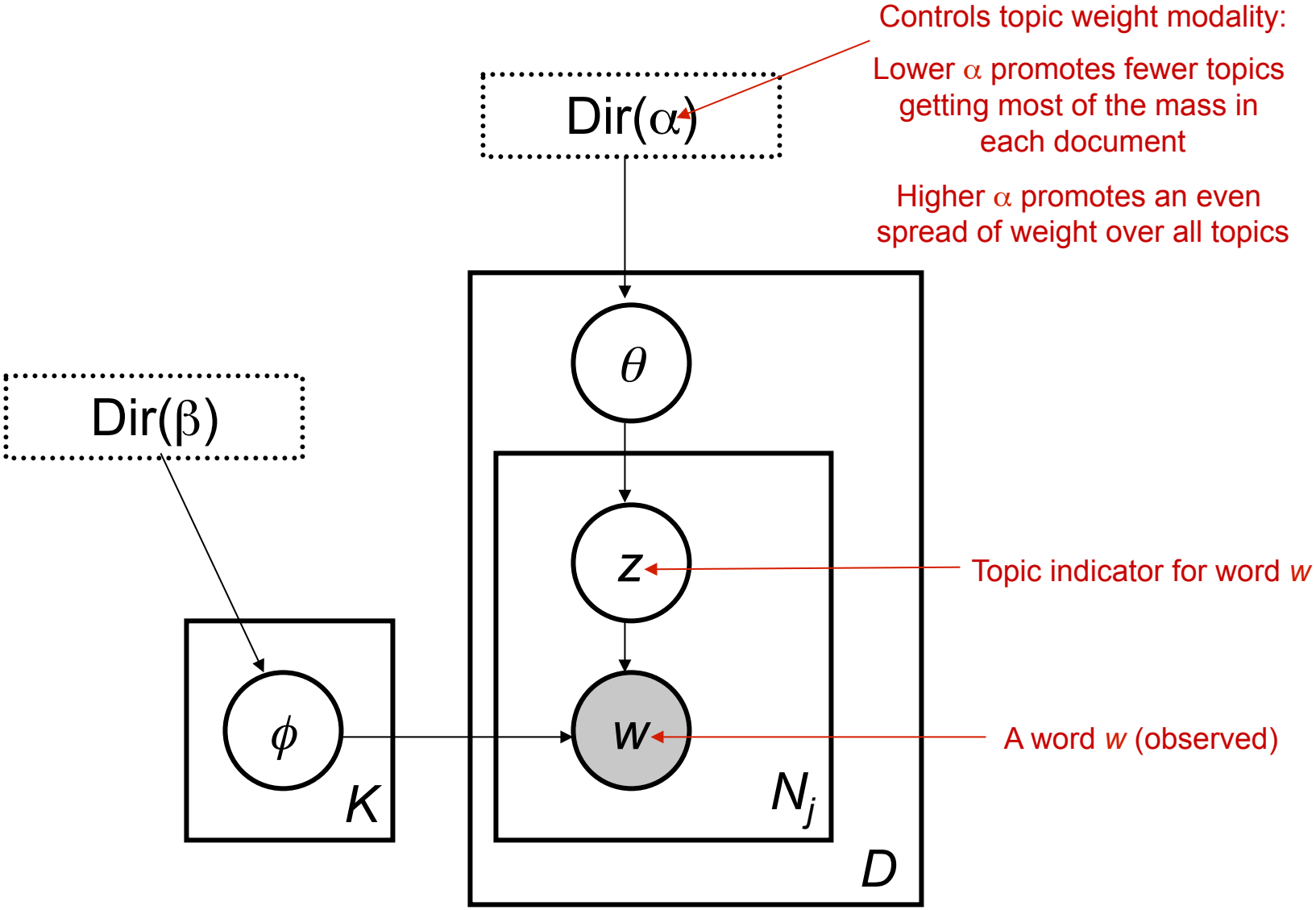
Latent Dirichlet allocation model (LDA)

- LDA: latent Dirichlet allocation model by Blei, Ng and Jordan (2003)
  - Bag-of-words or topic model
- Developed for the soft classification of documents
- General idea:
  - Discover the 'topics' (distributions across words)
  - Estimate the document-specific topic distributions
  - Group the documents based on their topic distributions

LDA Notation

- D documents, indexed by j, N_j words per document
- Vocabulary of size W
  - Usually this is the unique set of all words over the documents
- K latent (unknown) topics, indexed by k
  - Each topic is a distribution over the W words, given by ϕ_k
  - Each document is a distribution over the topics, given by θ_j
  - A document's topic weights govern how its words will arise
- Goal: Provide posterior inference on ϕ_k and θ_j

LDA Plate Diagram



LDA in Action

- LDA has been used with great utility on a variety of data sets:
  - Text document classification
  - Email spam identification
  - Image processing
  - Finding communities in social networks
  - Modeling manuscripts in *Science* from 1980-2002
  - Sequence based applications in genetics/genomics

5 topics from LDA model fit to *Science* manuscripts 1980-2002

computer	chemistry	cortex	orbit	infection
methods	synthesis	stimulus	dust	immune
number	oxidation	vision	jupiter	aids
advantage	reaction	neuron	system	infected
principle	product	recordings	solar	viral
design	organic	visual	gas	hiv
access	conditions	stimuli	atmospheric	vaccine
processing	molecule	motor	mars	antibodies

Blei & Lafferty 2007

Diary of a patient - entries relevant to cancer

11/18 - Diagnosed with ovarian cancer

11/27 - Surgery today

12/3 - Closed on new house

1/5 - Finished first course of platinum and taxol; very tired

1/6 - Leaving for weekend at lake

3/8 - Finished with chemo !

3/12 - dog broke her leg

7/30 - CA-125 up, it's back…..

8/05 - Started Doxil

Suppose we could add to that diary…

2/18 - HRG overexpressed

2/24 - HPC1 turned off

3/11 - TP53 hyper-methylated

3/29 - CDH1 lost

3/31 - ZNF604P overexpressed

4/17 - PDZD7 turned off

5/12 - KAI1 turned off

6/30 - BCL2 overexpressed

8/05 - CCL2 overexpressed

Turning patients into documents (three kinds of words)

- Drug words:
  - Drugs given to a patient as adjuvant or primary-recurrence treatment
  - Different words for different drugs at different treatment stages
  - e.g., D1-carboplatin or D2-topotecan
- Gene words:
  - Selected ~1000 genes in cancer related KEGG pathways
  - For each gene, partition patients into tenths based on expression
  - middle 40%-ile gets no words; highest 80,90,100 %-iles get 8,9,10 words with –HI tagged. Similar for lowest 10, 20, 30%.
- Methylation words:
  - as gene words

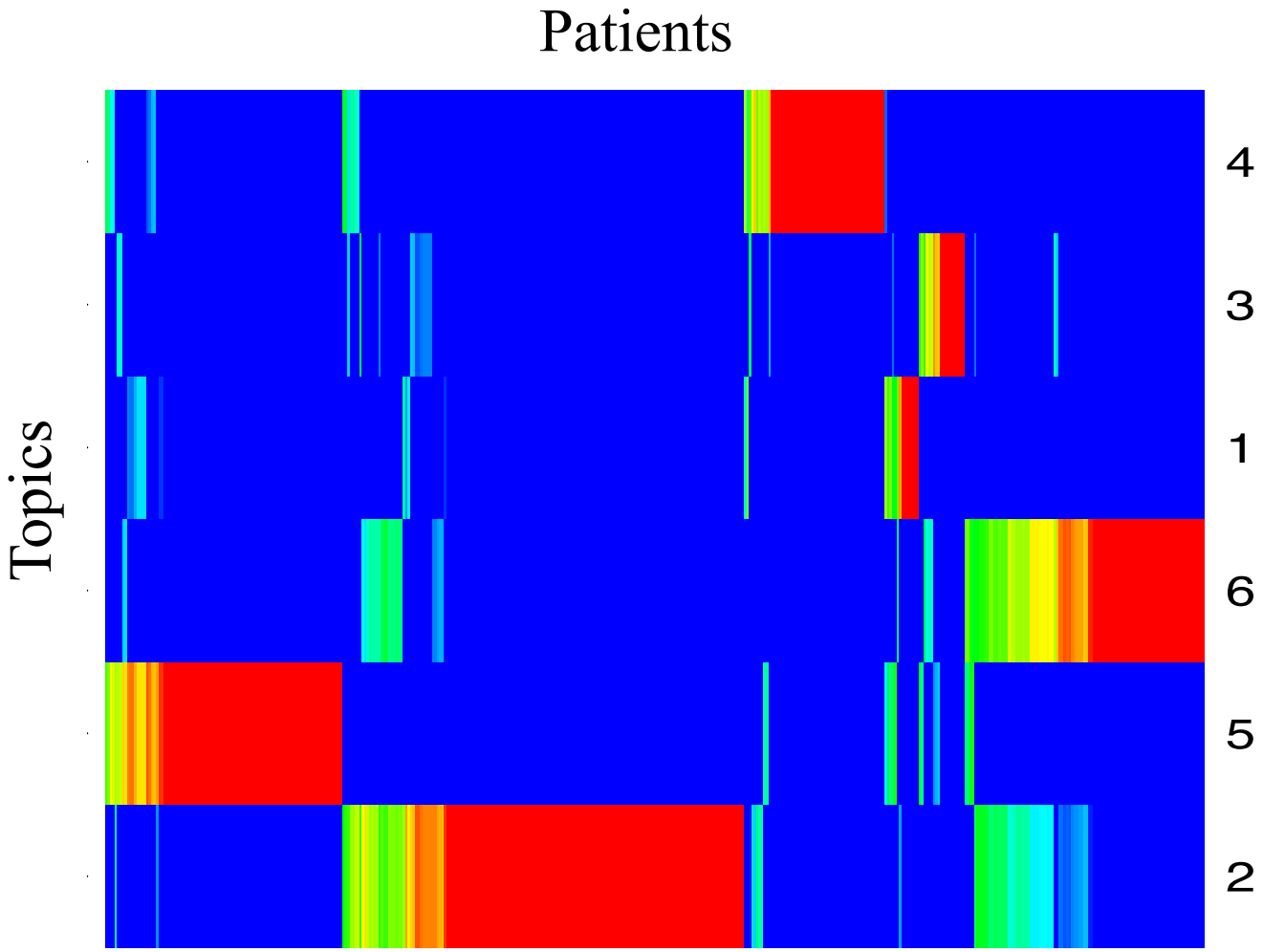
Example document

Consider a patient who:

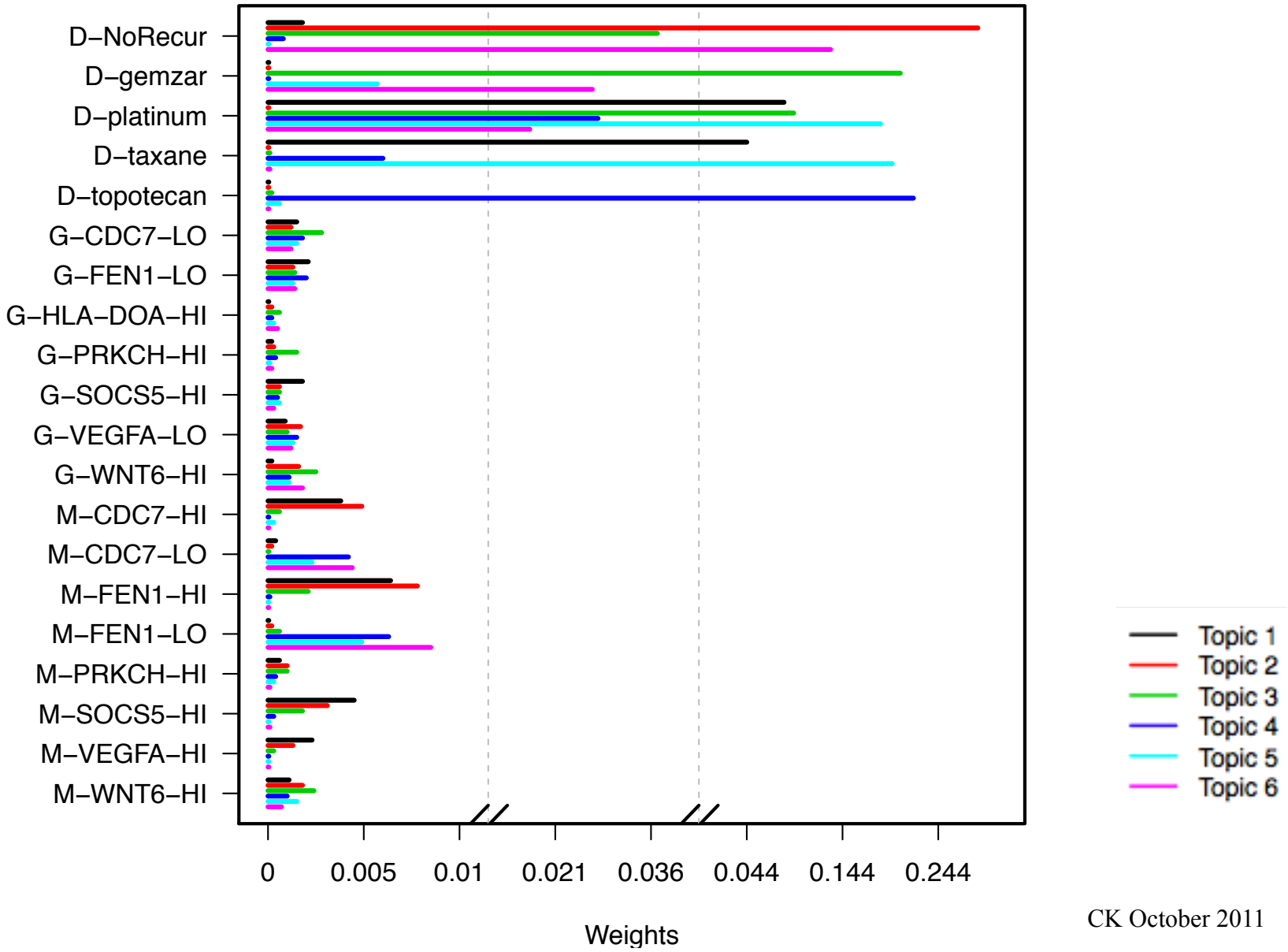
- had the mid-range mRNA expression levels for all genes but two,
- those two being APC (lowest 10%-ile) and MYC (10-20%-ile),
- had high methylation for p16 and MAPK (upper 80% and 90%-iles)
- received carboplatin and paclitaxel as adjuvant therapy,
- recurred after eight months, received topotecan, died two years later.

D_j = (G-APC-LOW, …, G-APC-LOW, …. , G-MYC-LOW, …, G-MYC-LOW,
10 *9*
M-p16-HI, …, M-p16-HI, … , M-MAPK-HI, … , M-MAPK-HI,
8 *9*
D-topotecan, …, D-topotecan
n

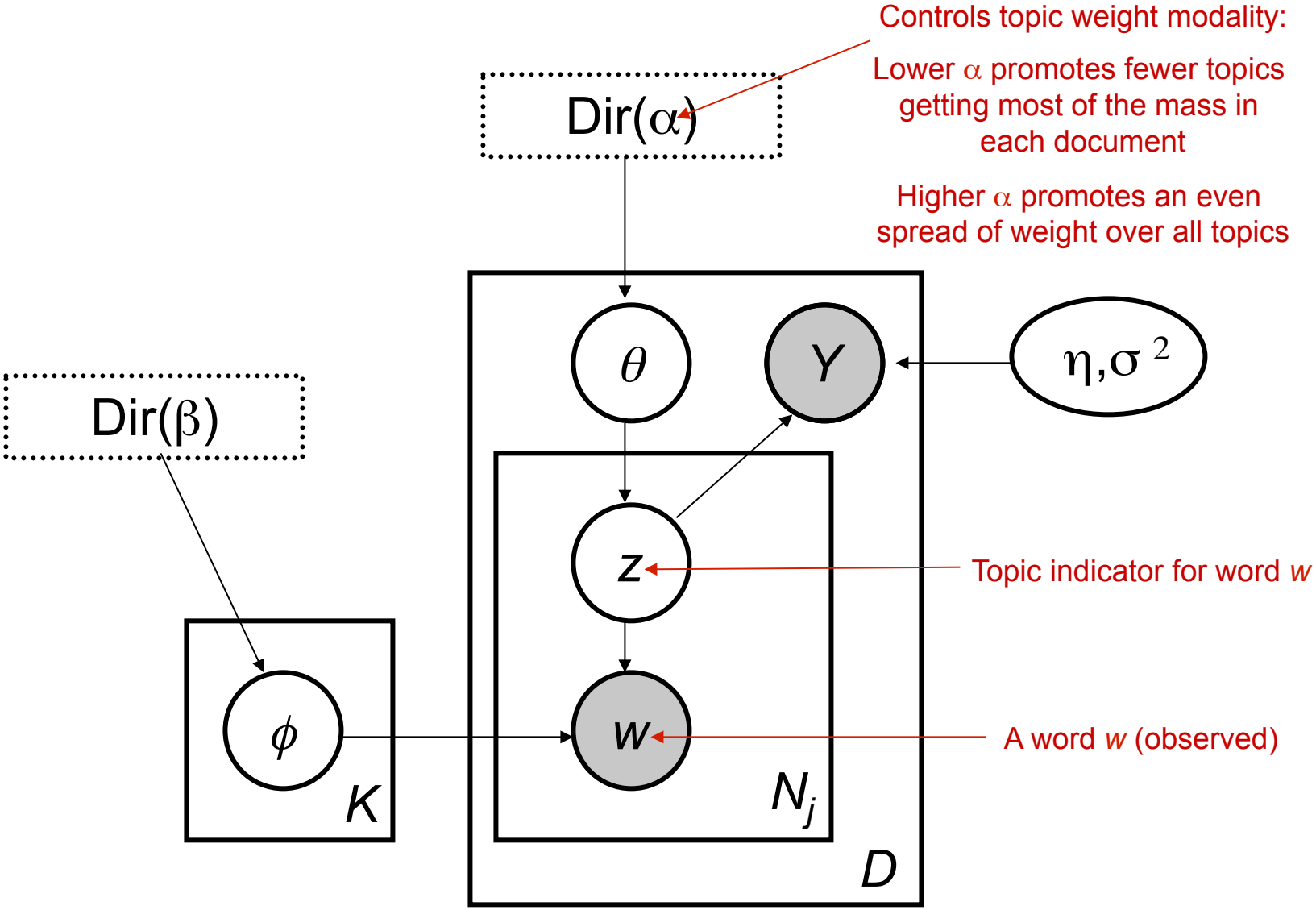
LDA using mRNA, Methylation, and drug words



Topic-specific distributions



Supervised LDA plate diagram



Blei and McAuliffe 2009

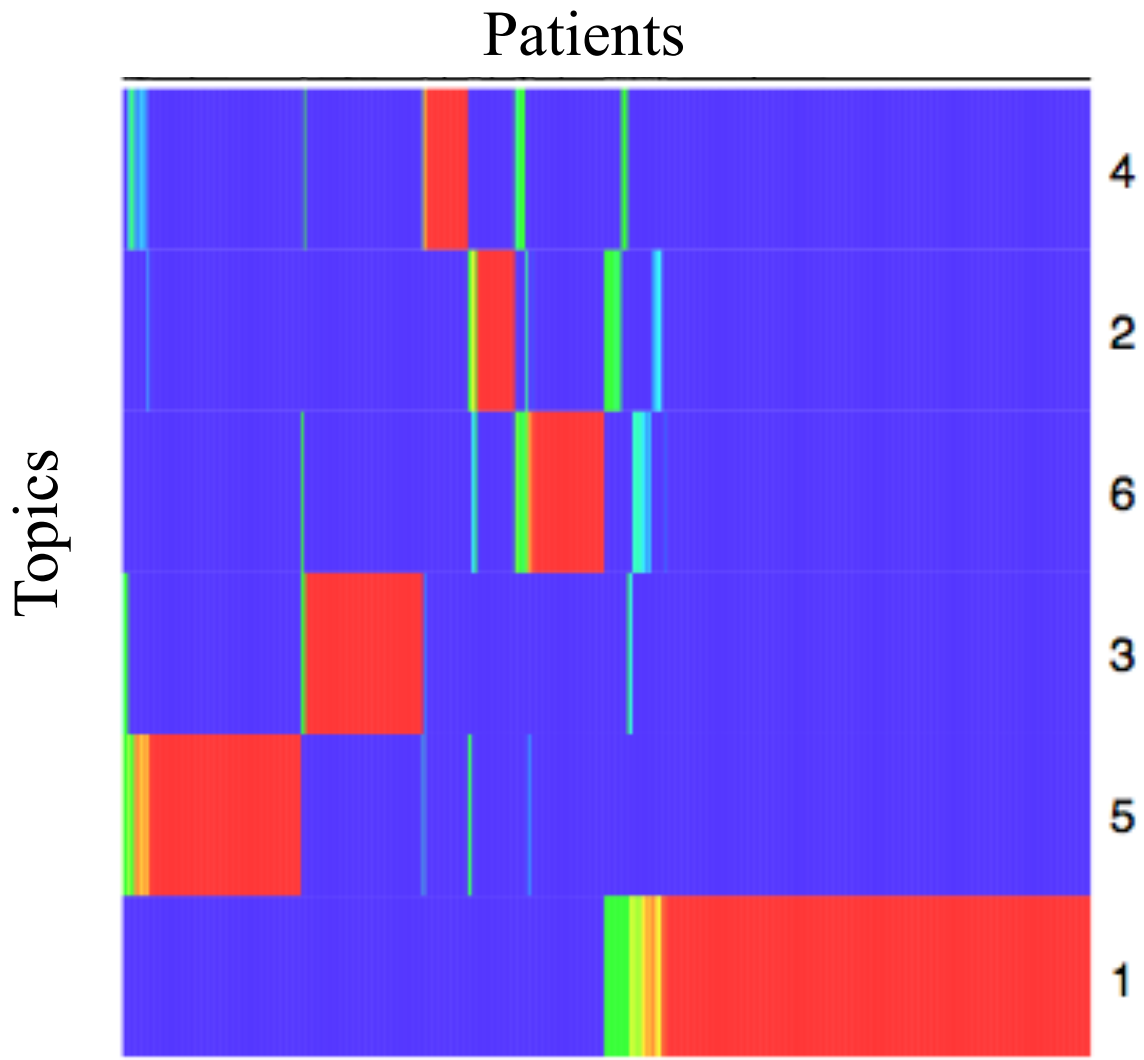
Survival-supervised LDA

- Some notation: J documents, K topics, W vocab words, N_j words in doc j
- Given model parameters $\pi = \{\alpha, \{\beta_k\}, \eta\}$; for doc j:

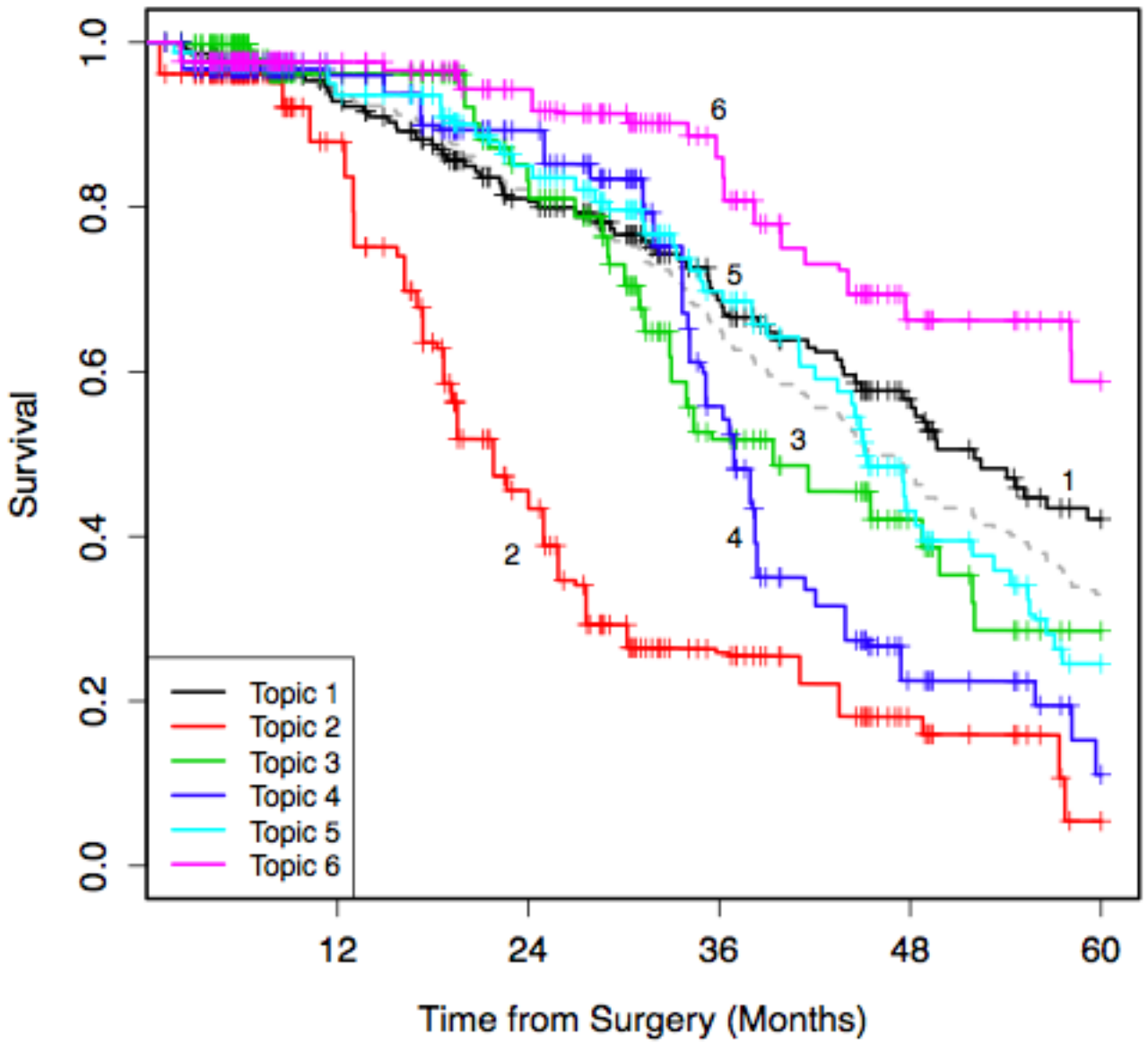
1) Draw $\theta_j \sim \text{Dir}(\alpha)$
2) For each word w_n, $n = 1, \ldots, N_j$:
   i. Draw topic $z_n \sim \text{Discrete}(\theta_j)$
   ii. Draw a word $w_n \sim \text{Discrete}(\beta_{z_n})$
3) Draw $Y_j \sim S(z_n, \eta)$

- Idea: Estimate the $\{\theta_j\}$ and the $\{\beta_k\}$ using the $\{w_{nj}\}$ and survival

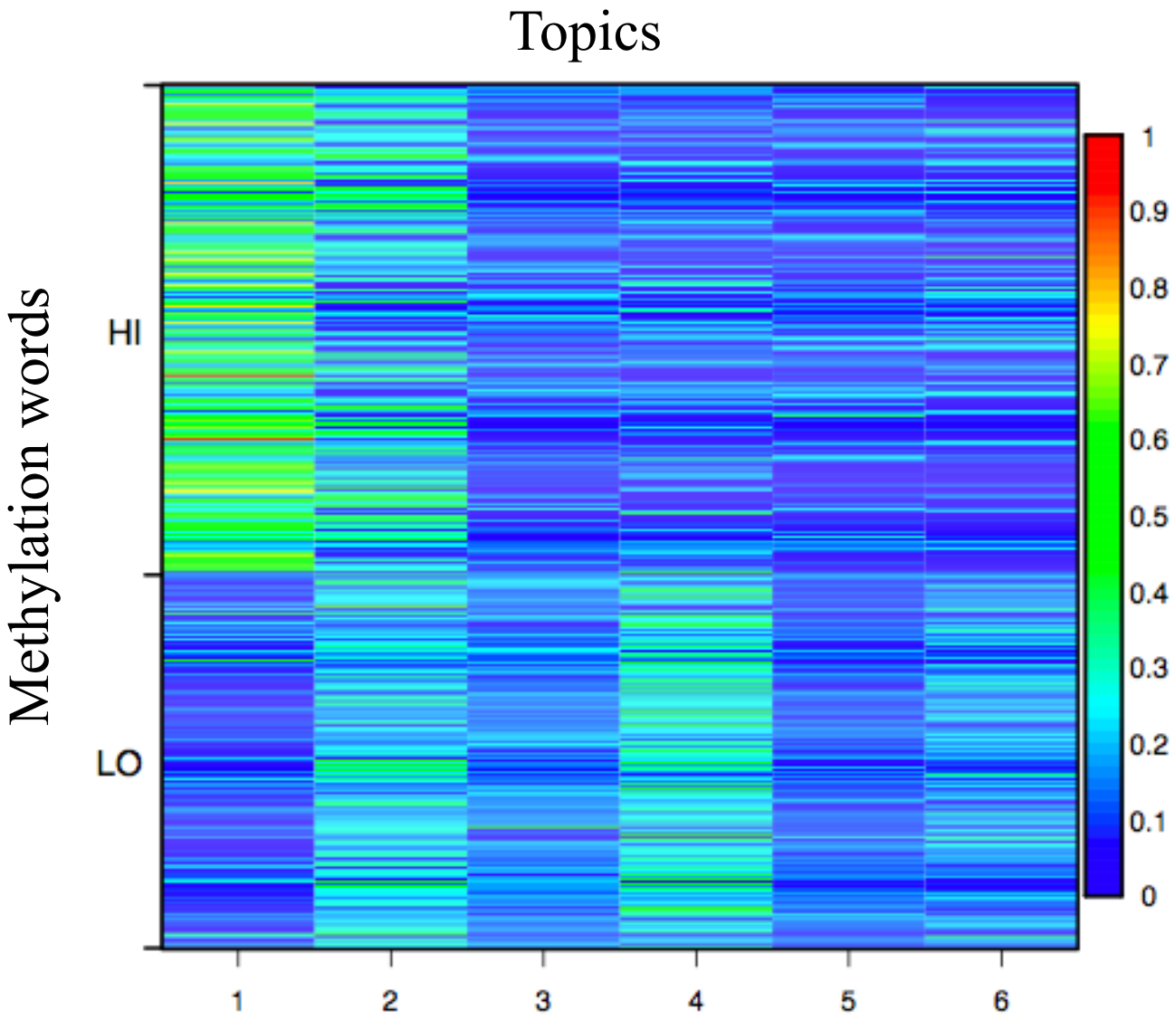
survLDA: Patient-specific distributions over topics



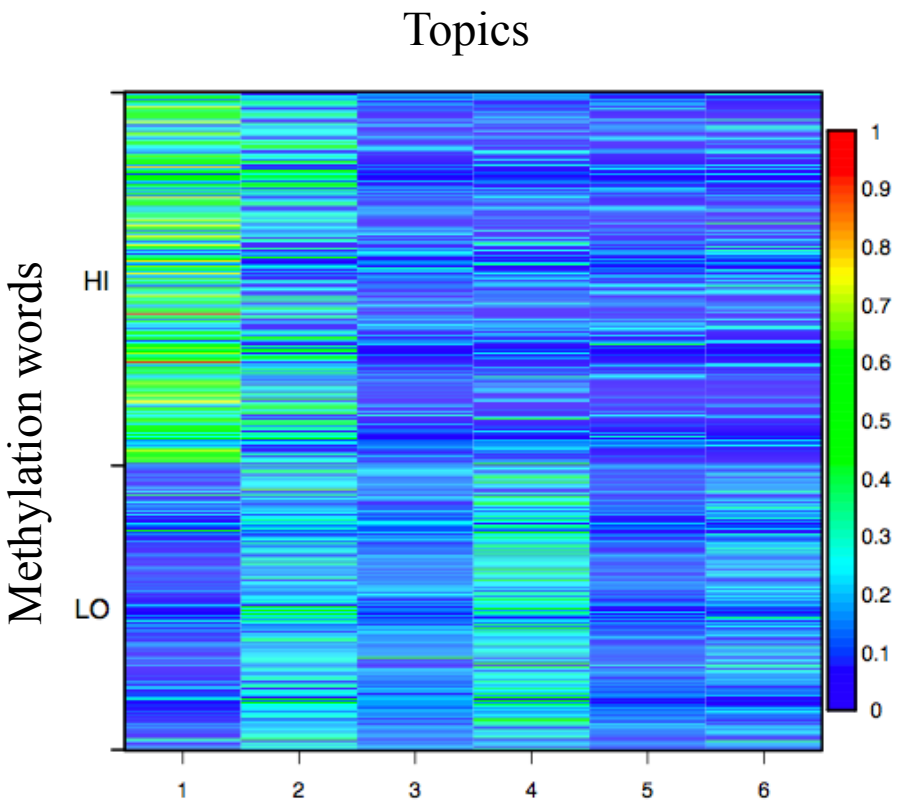
survLDA on Drugs, mRNA and Methylations

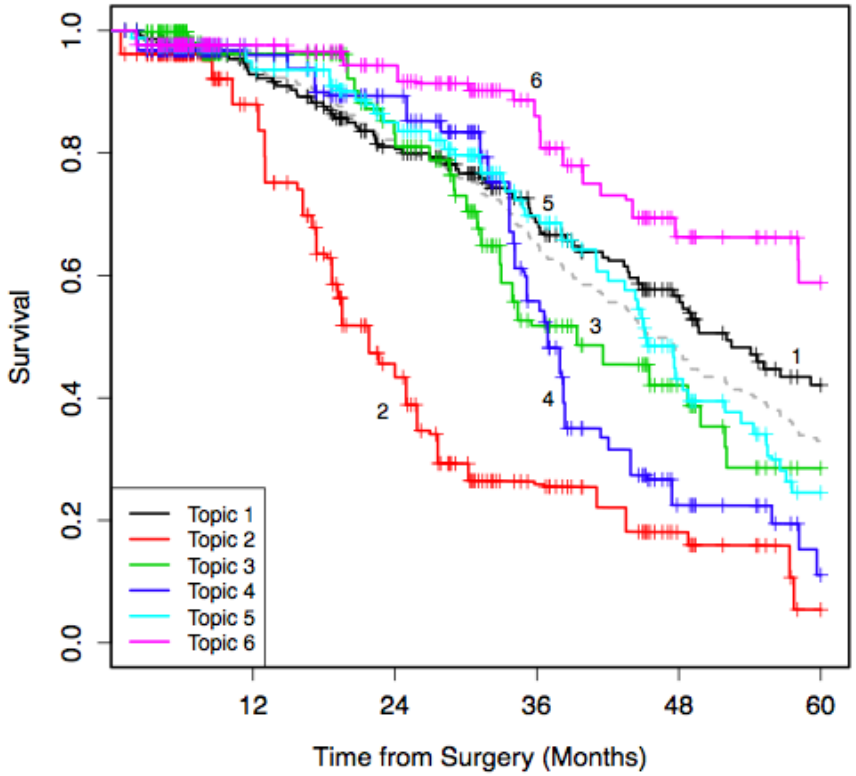


survLDA: Topic-distributions over methylation words



survLDA on Drugs, mRNA and Methylations





Summary

- Our goal in the TCGA ovary project is to derive genomic based signature useful for guiding ovarian cancer treatment at the time of first recurrence.
- Using LDA and survival-supervised LDA to integrate data (methylation, expression, CNV, SNP, LOH, clinical information) for improved biological discovery and prediction.
- Currently evaluating many methods for document creation
- Improvements are observed with adjustments on Dirichlet priors
- Framework allows for correlated topics and/or documents

Acknowledgements

Michael Gould PhD
James Thomson PhD, DVM

Janet Rader MD
William Bradley MD

John Dawson
Shuyun Ye
Ning Leng
Oleg Moskvin PhD
Kevin Eng PhD