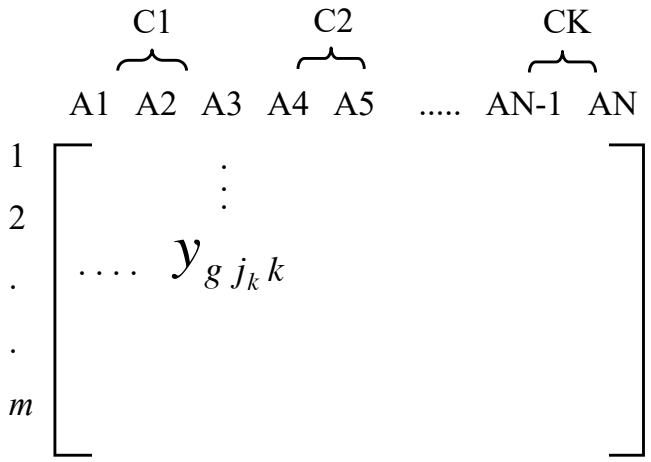
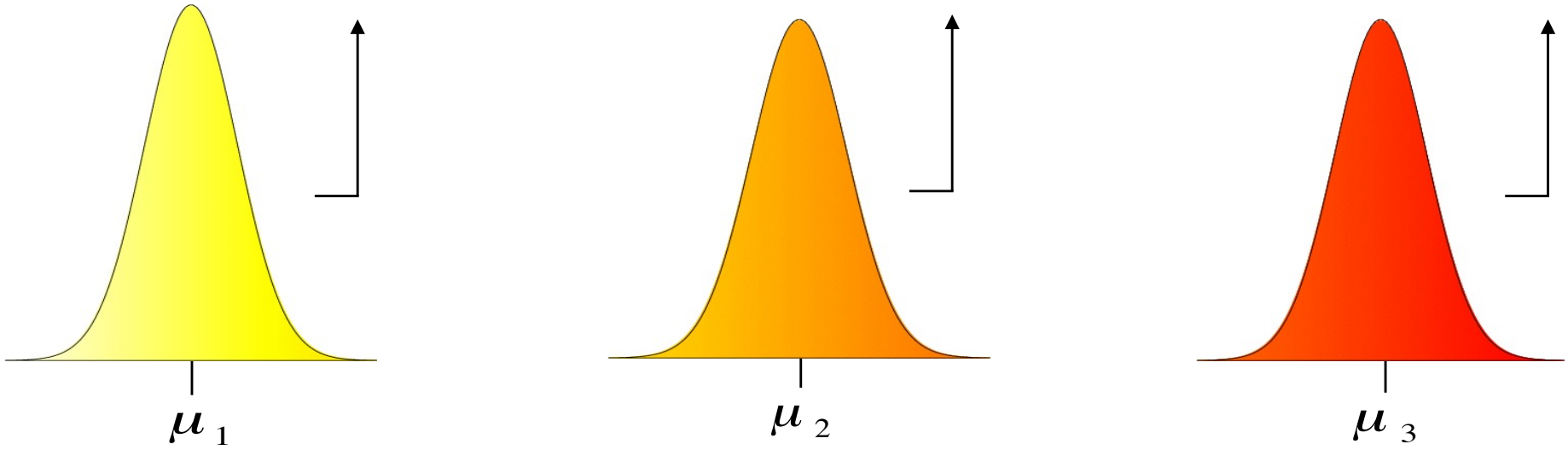
Data Structure

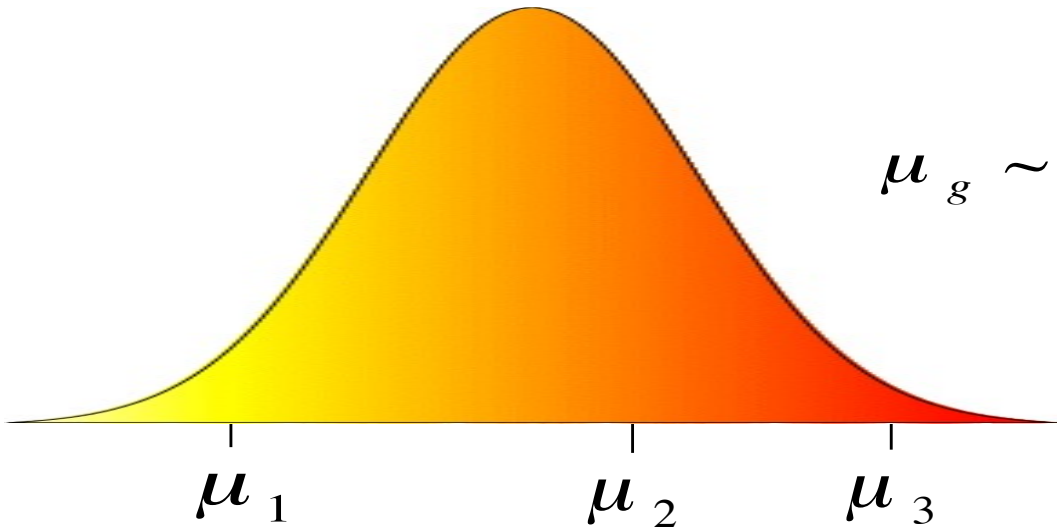


- $g = 1,2, \ldots, m$ genes (transcripts)
- $k = 1,2,..., K$ conditions
- $j = 1,2,... j_k$ replicates

Hierarchical Model for Expression Data

$$l(y_{1j})|\mu_1 \sim f(\cdot|\mu_1) \qquad l(y_{2j})|\mu_2 \sim f(\cdot|\mu_2) \qquad l(y_{3j})|\mu_3 \sim f(\cdot|\mu_3)$$





$$\mu_g \sim f(\cdot)$$

Hierarchical Model for Expression Data (Two conditions)

- Let $y = [y_{c1}, y_{c2}]$ denote data (one gene) in conditions C1 and C2 (not denoting gene or replicates here).
- Two patterns of expression:

$$\text{P0 (EE)}: \quad \mu_{c1} = \mu_{c2}$$

$$\text{P1 (DE)}: \quad \mu_{c1} \neq \mu_{c2}$$

- For P0, $y \sim \int f(y|\mu) f(\mu) d\mu \equiv f_0(y)$
- For P1, $y \sim \int f(y|\mu_{c1}, \mu_{c2}) f(\mu_{c1}, \mu_{c2}) d\mu_{c1} d\mu_{c2}$

$$\equiv \underbrace{\int f(y_{c1}|\mu_{c1}) f(\mu_{c1}) d\mu_{c1}}_{f_0(y_{c1})} \underbrace{\int f(y_{c2}|\mu_{c2}) f(\mu_{c2}) d\mu_{c2}}_{f_0(y_{c2})} \equiv f_1(y)$$

Hierarchical Mixture Model for Expression Data

- Two conditions:

$$y \sim p_0 f_0(y) + p_1 f_1(y) \quad => \quad p(P1|y) = \frac{p_1 f(y|P1)}{p_0 f(y|P0) + p_1 f(y|P1)}$$

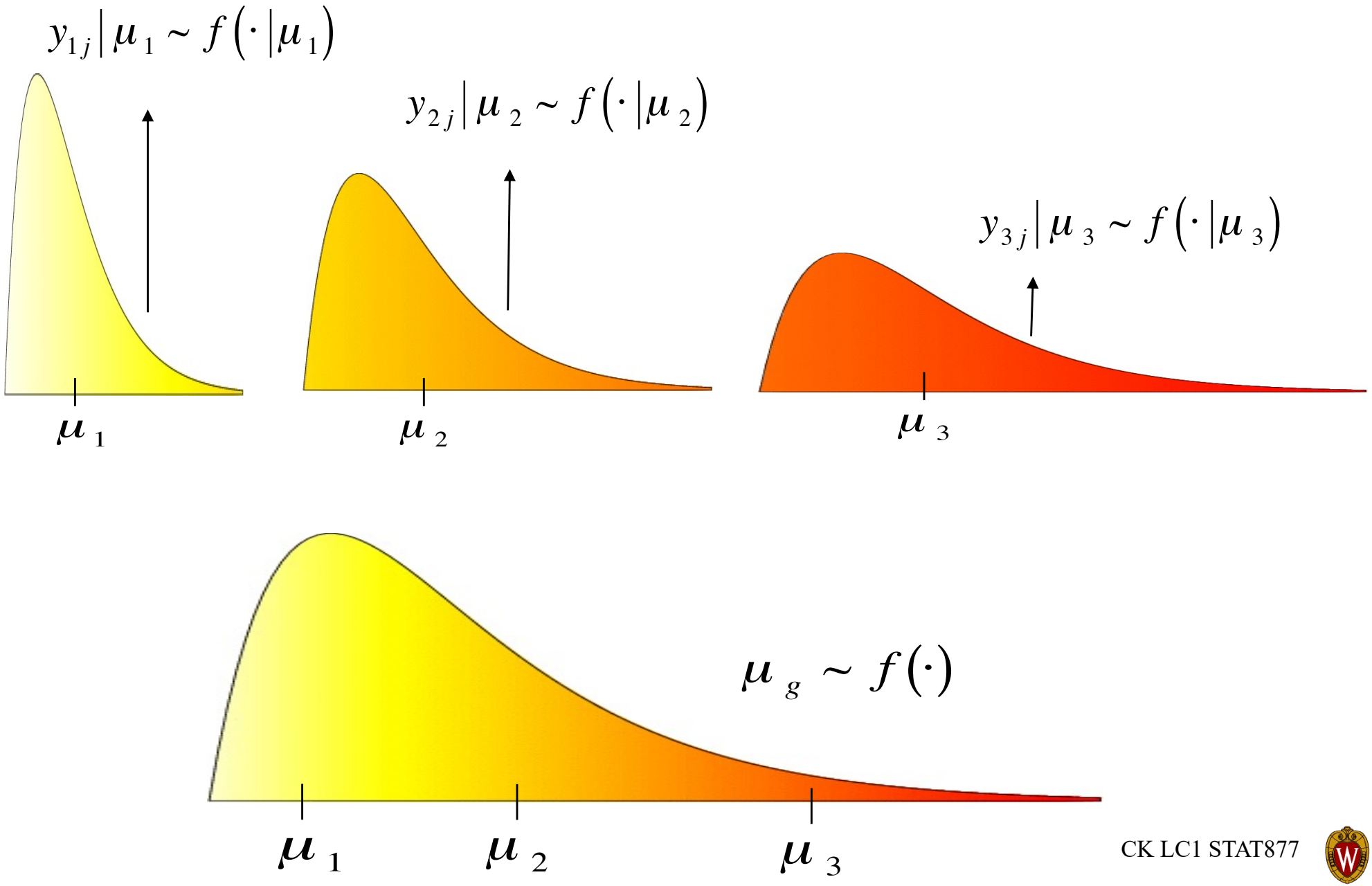
- Multiple conditions:

$$y \sim \sum_{k=1}^{K} p_k f_k(y) \quad => \quad p(Pk'|y) = \frac{p_{k'} f(y|Pk')}{\sum_{k \neq k'} p_k f(y|Pk)}$$

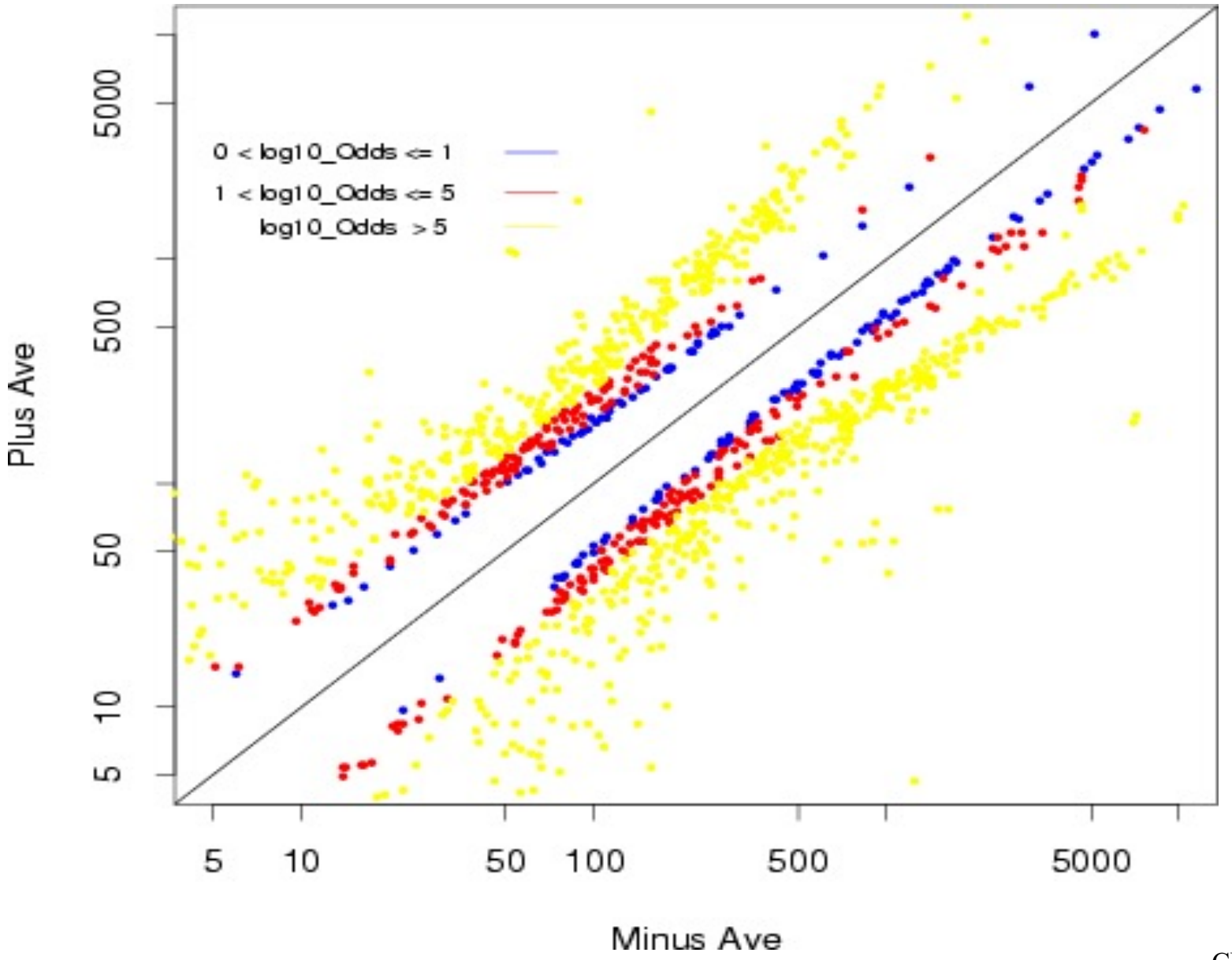
- Parameter estimates via EM (see my notes)

- Posterior probabilities can be used to determine FDR based threshold.

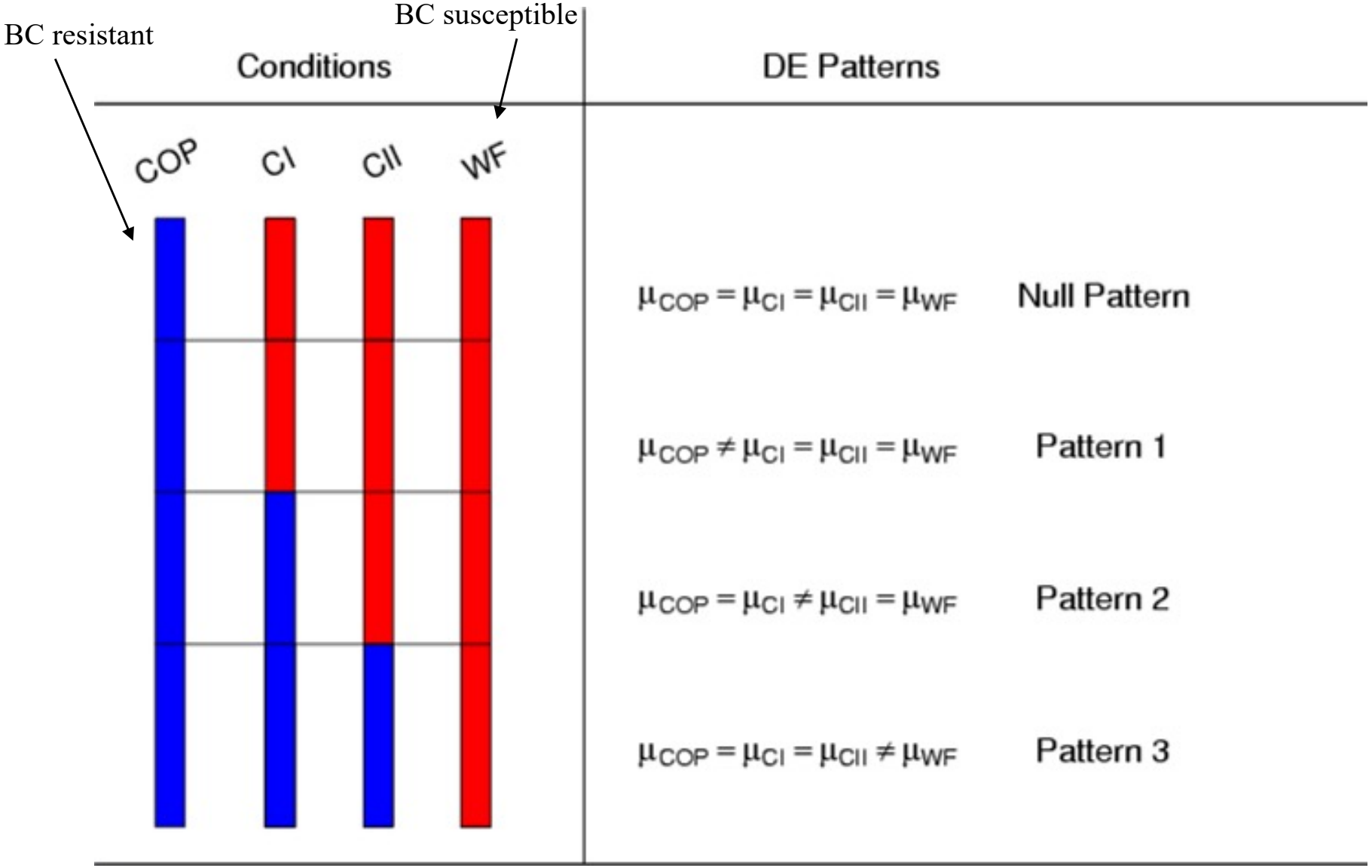
Hierarchical Model for Expression Data (One condition)



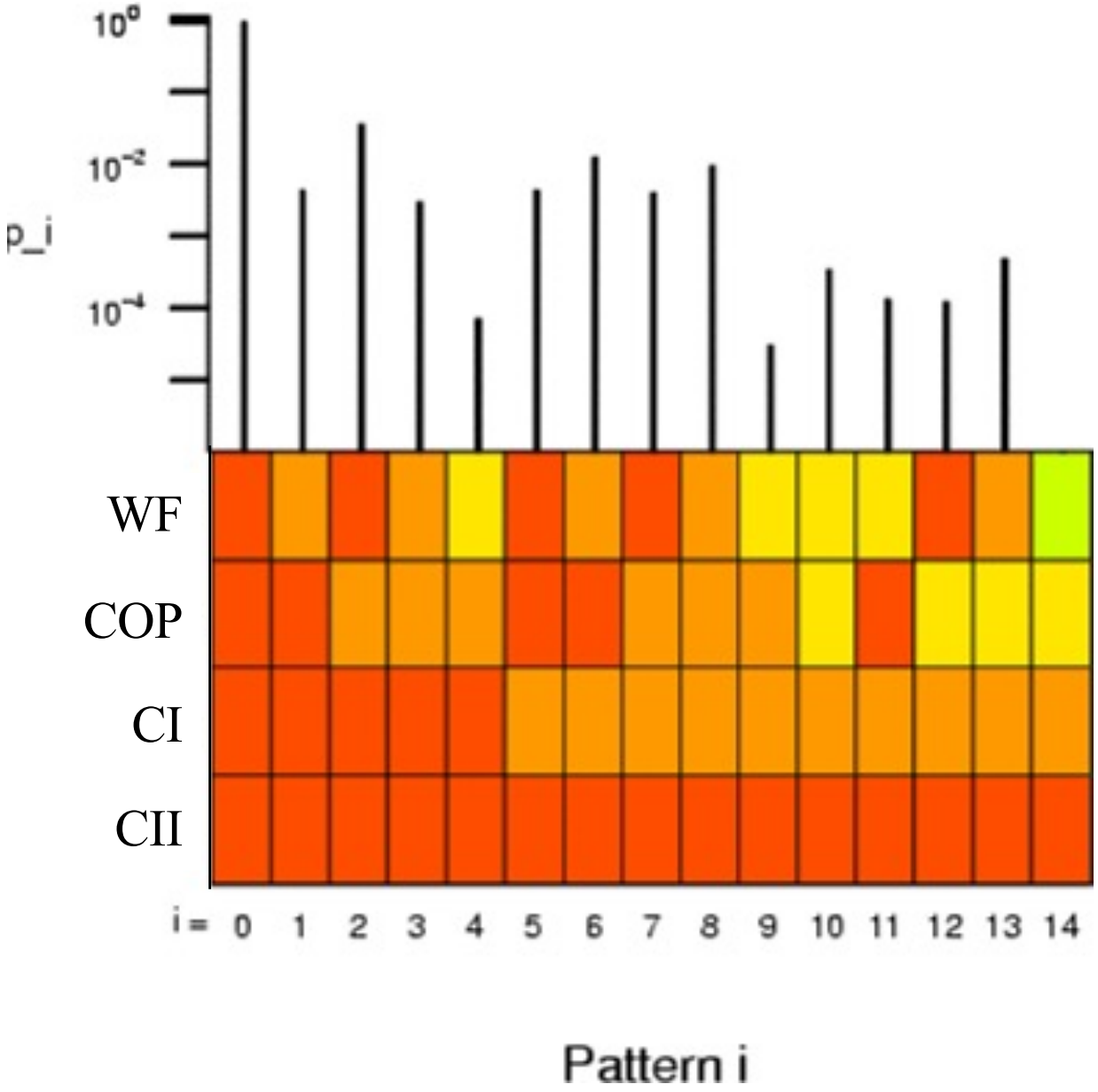
Odds plot: SCD knockout vs. SV129 (Attie lab)



WF, COP, and 2 Congenics (Gould lab)



Overall results on 4 group comparison (15 patterns)

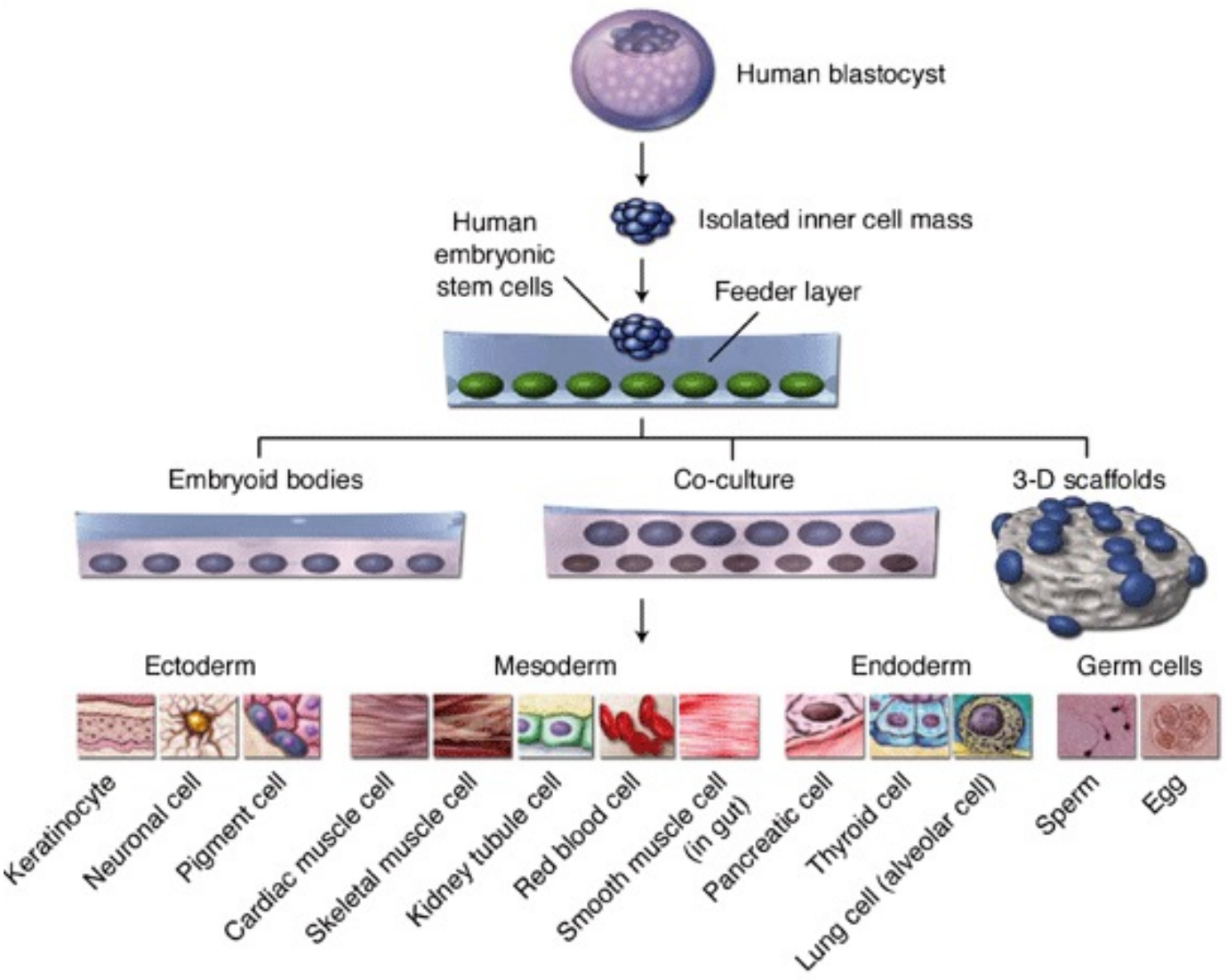


Comments on EBarrays

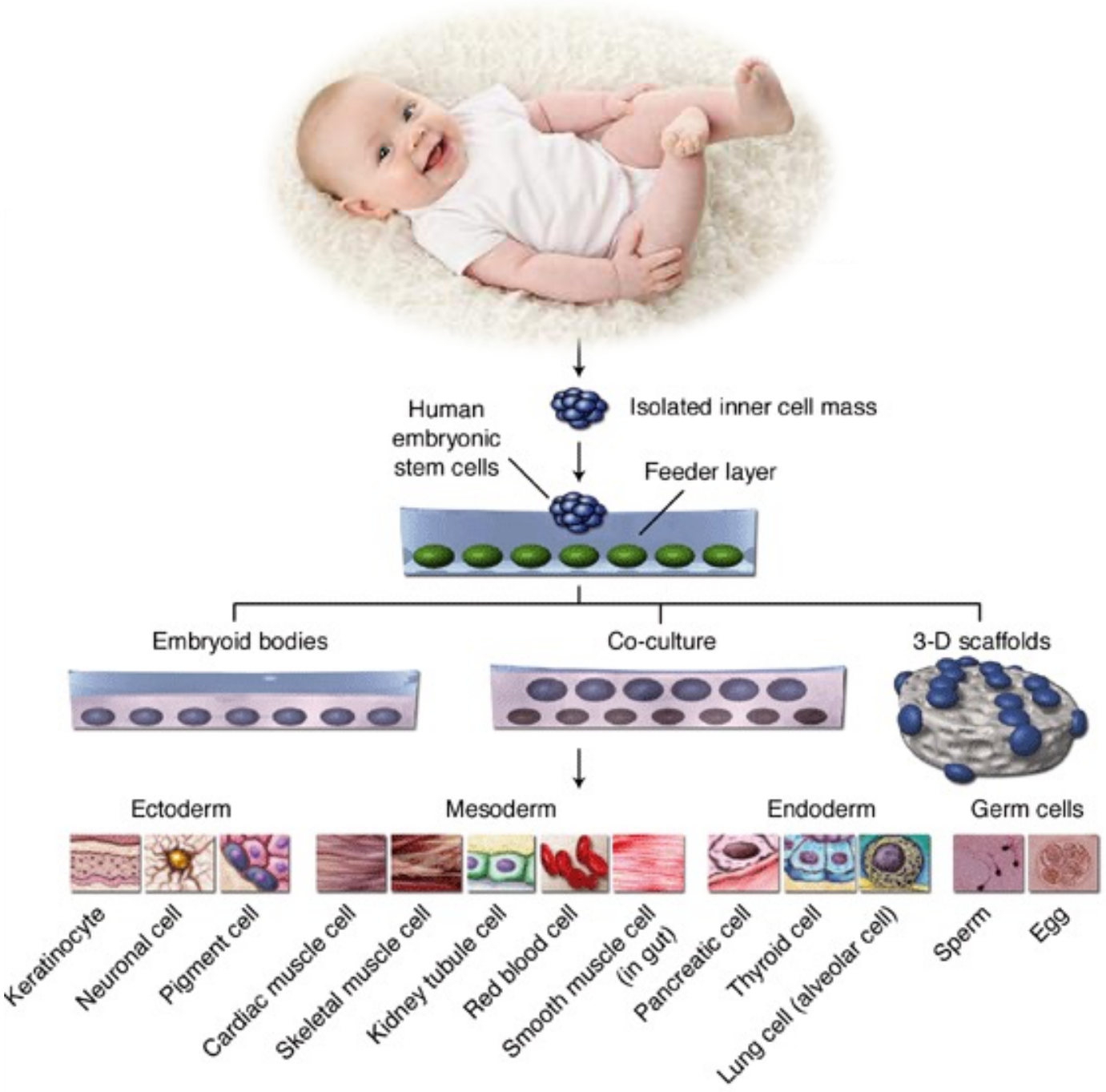
- Hierarchical model is used to identify patterns of expression. The model accounts for the measurement error process and for natural fluctuations in absolute expression levels.
- Posterior probabilities of expression patterns are calculated for every transcript.
- Multiple conditions are handled in the same way as two conditions (no extra work required!).
- Threshold can be adjusted to target a specific FDR.
- LNNMV model relaxes CCV assumption
- In Bioconductor, with families GG, LNN, and LNNMV.

Moving on to RNA-seq data analysis

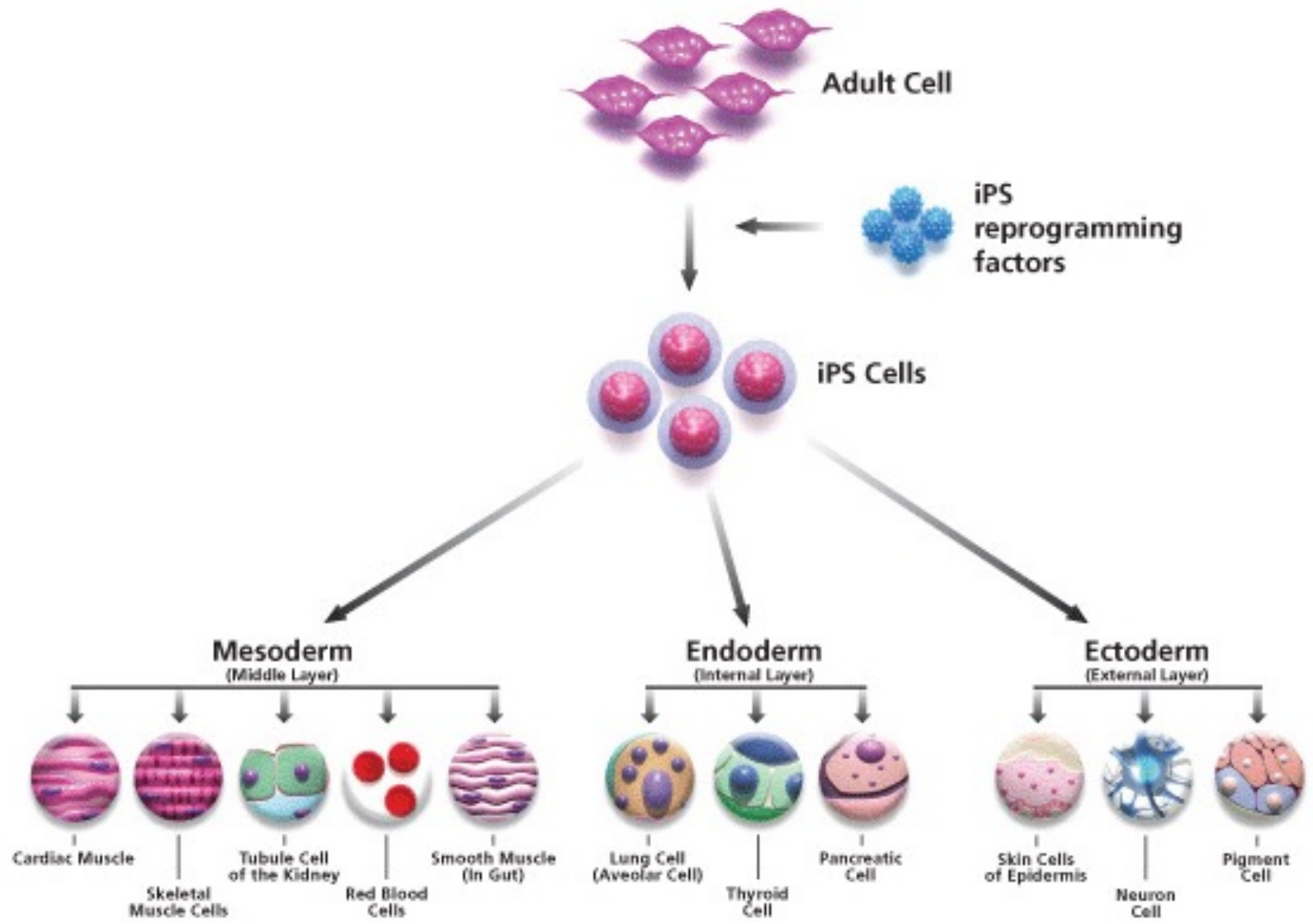
The Thomson lab isolated the first hESC in 1998



Thomson *et al.*, Science, 1998


Human embryonic stem cells
Isolated inner cell mass
Feeder layer
Embryoid bodies
Co-culture
3-D scaffolds
Ectoderm
Mesoderm
Endoderm
Germ cells
Keratinocyte
Neuronal cell
Pigment cell
Cardiac muscle cell
Skeletal muscle cell
Kidney tubule cell
Red blood cell
Smooth muscle cell (in gut)
Pancreatic cell
Thyroid cell
Lung cell (alveolar cell)
Sperm
Egg

The Thomson lab developed iPSC lines in 2007



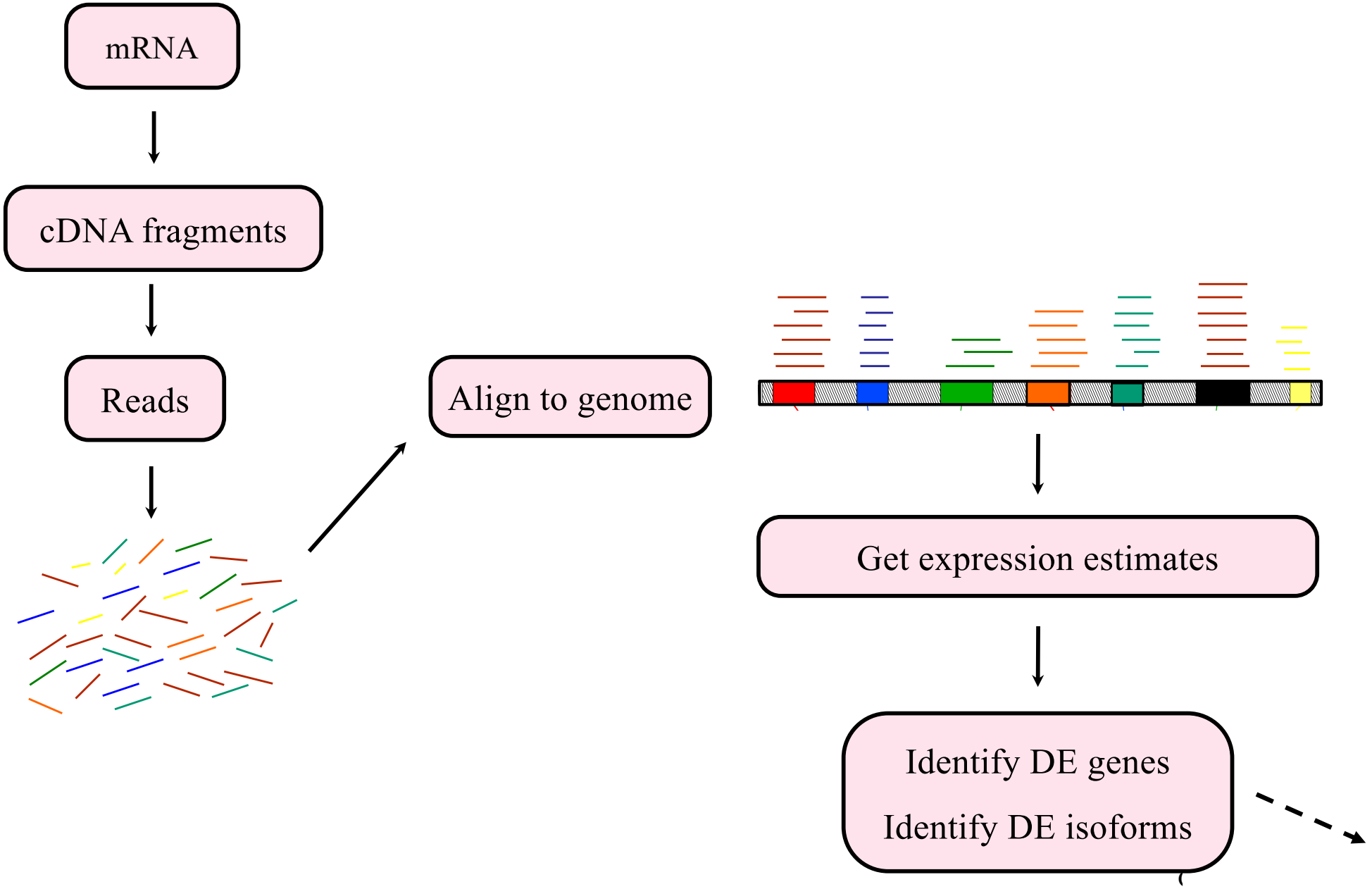
Yu *et al*., Science, 2007

Characterize differences between hESCs and iPSCs

- Mutations
- Methylation
- Protein
  - .
  - .
  - .
- Gene expression
- Isoform expression
  - 4 hESCs vs. 4 iPSCs
  - Illumina Genome Analyzer II ->Bowtie -> RSEM
  - Repeated in triplicate

Can we accurately identify differentially expressed isoforms in an RNA-seq experiment ?

RNA-Seq: Data Collection



RNA-Seq: Data Analysis Tools

Alignment

Transcriptome Assembly

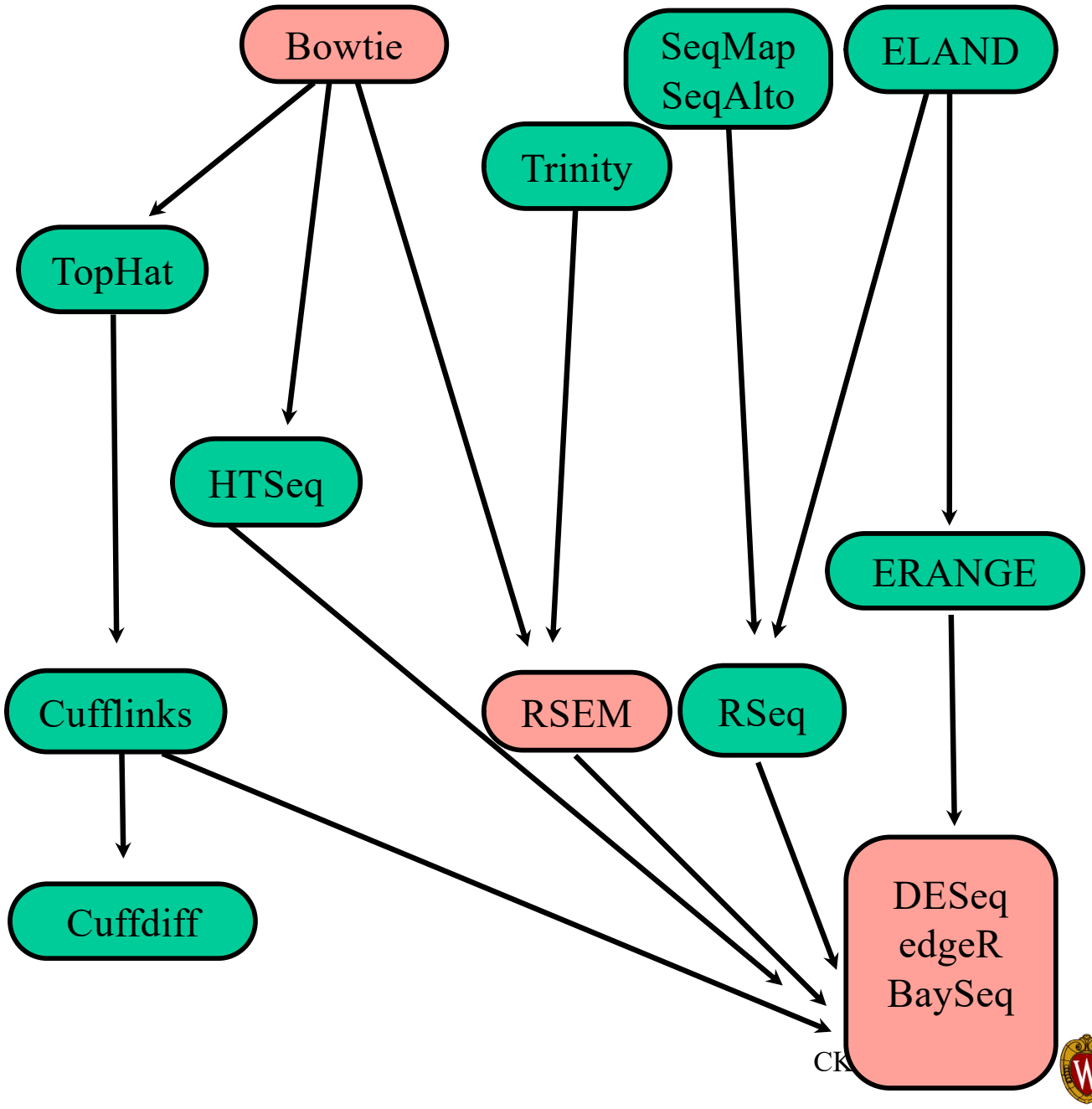
De novo Transcript Detection

Quantification (Gene and Exon)
No Multi-read assignment

Quantification (Gene)
With Multi-read assignment

Quantification (Gene and Isoform)
With Multi-read assignment

DE analysis



CK

RNA-Seq: Methods for identifying DE genes

Each method assumes:

$$X_{gs} \sim NB(\mu_{gs}, \sigma^2_{gs});$$
$$E(X_{gs}) = \mu_{gs};$$
$$Var(X_{gs}) = \sigma^2_{gs}.$$

Condition C; Sample s; Gene g;
Counts X_{gs}; Normalization Factor l_s.

- edgeR: Robinson *et al*. 2007

$$Var(X_{gs}) = \mu_{gs}(1 + \mu_{gs}\phi_g); \quad \mu_{gs} = l_s \mu_{g0C};$$
Test $H_0 : \mu_{g0C1} = \mu_{g0C2}$

- DESeq: Simon Anders and Wolfgang Huber 2010

$$Var(X_{gs}) = \mu_{gs} + l_s^2 v_{gC}$$
$$\mu_{gs} = l_s \mu_{g0C}; \quad v_{gC} = v_g(\mu_{g0C})$$
Test $H_0 : \mu_{g0C1} = \mu_{g0C2}$

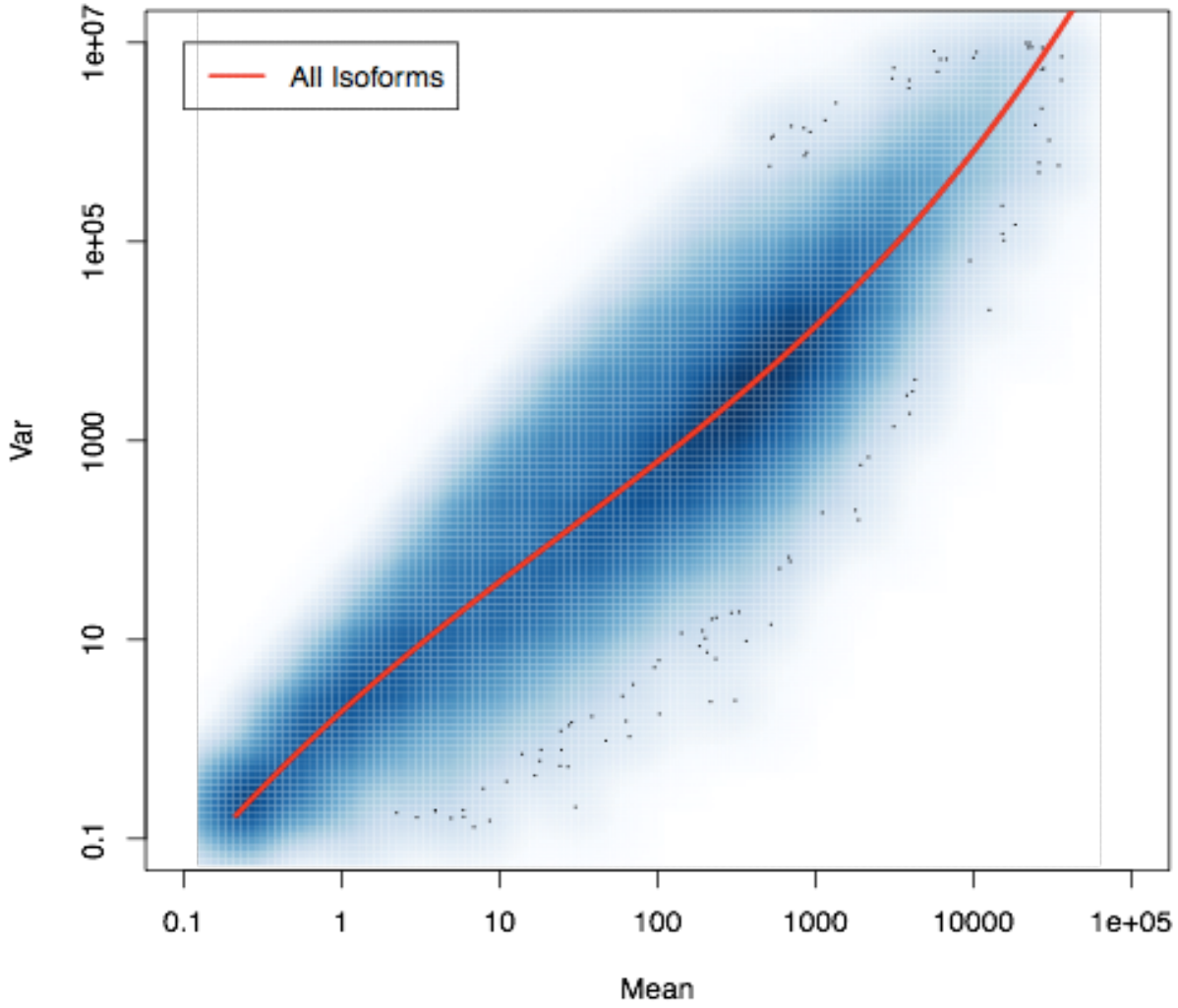
- BaySeq: Thomas J. Hardcastle *et al*. 2010

$$Var(X_{gs}) = \mu_{gs}(1 + \mu_{gs}\phi_g)$$

  – derive an empirical prior from the data

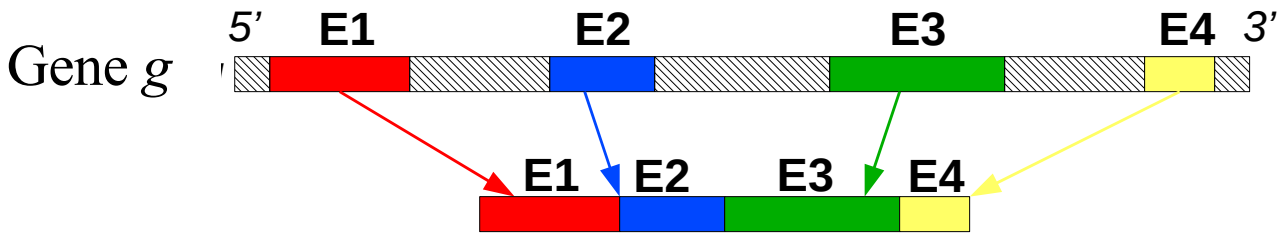
RNA-Seq: Methods for identifying DE genes

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

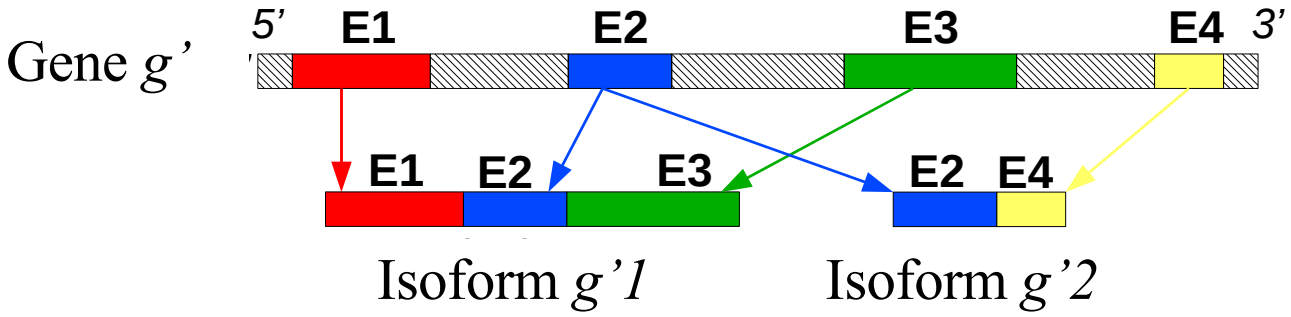


Mean-variance relationship changes with isoform complexity

Isoforms

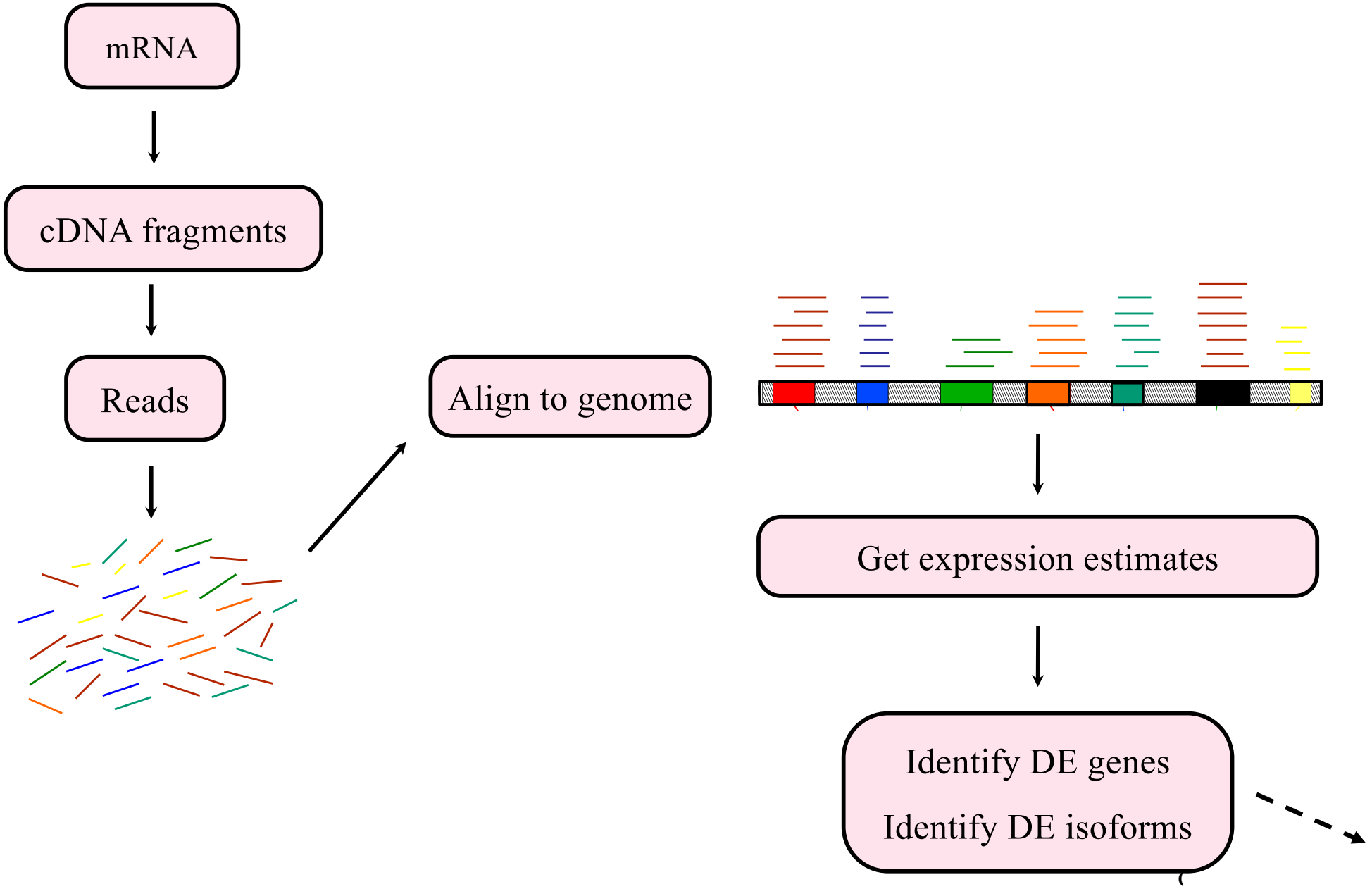


$N_g = 1$

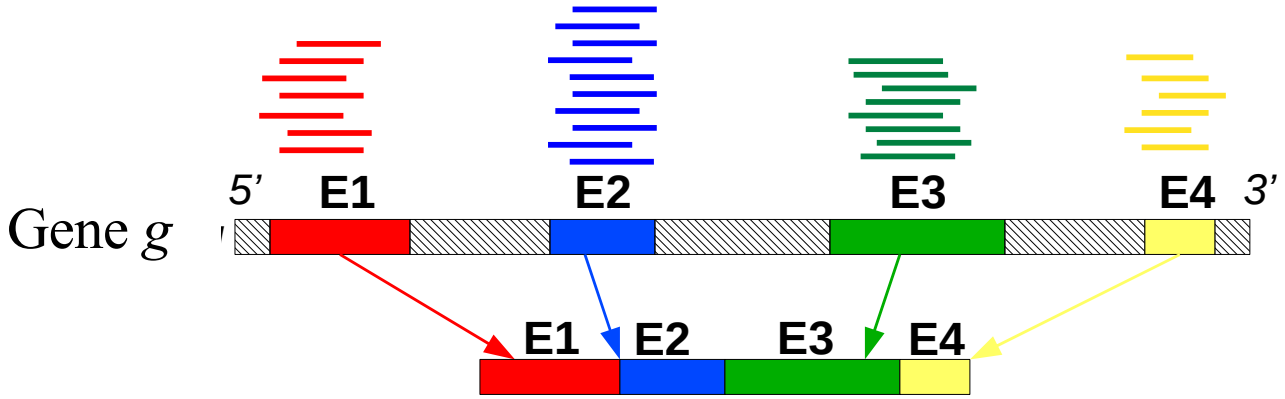


$N_{g'} = 2$

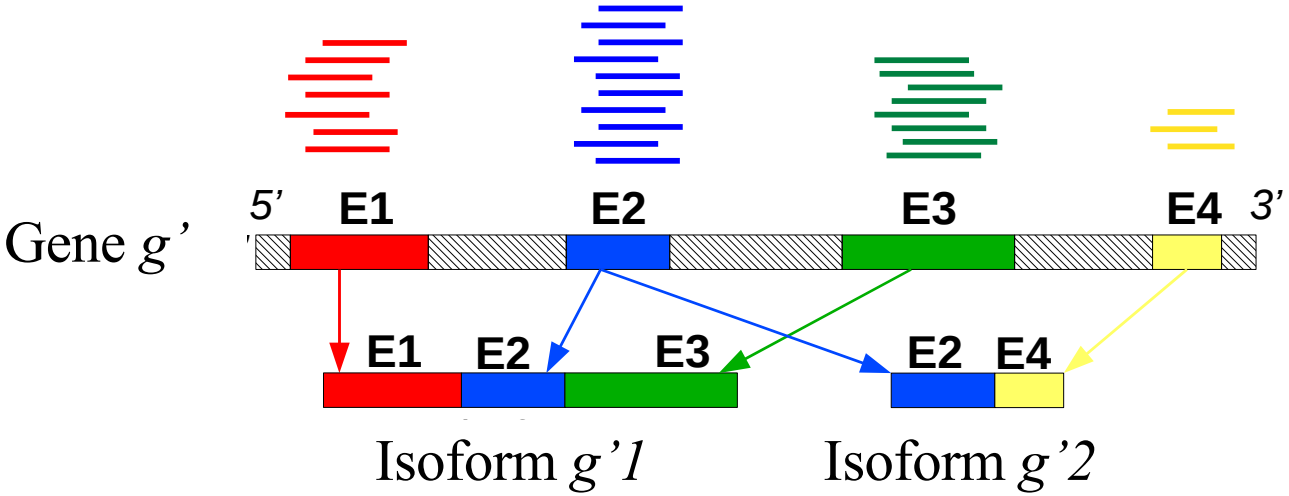
RNA-Seq: Data Collection



Estimating isoform expression



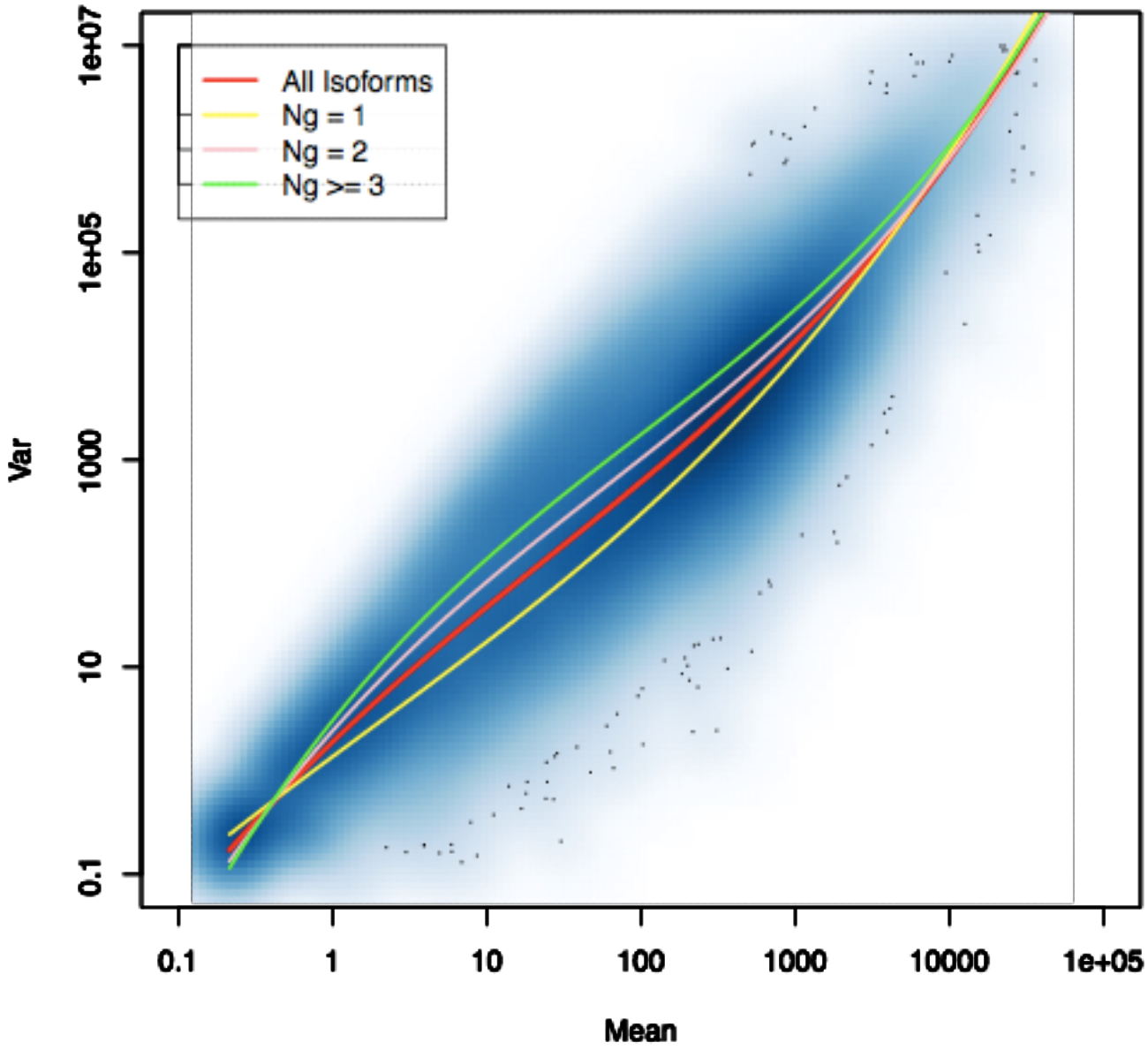
$N_g = 1$



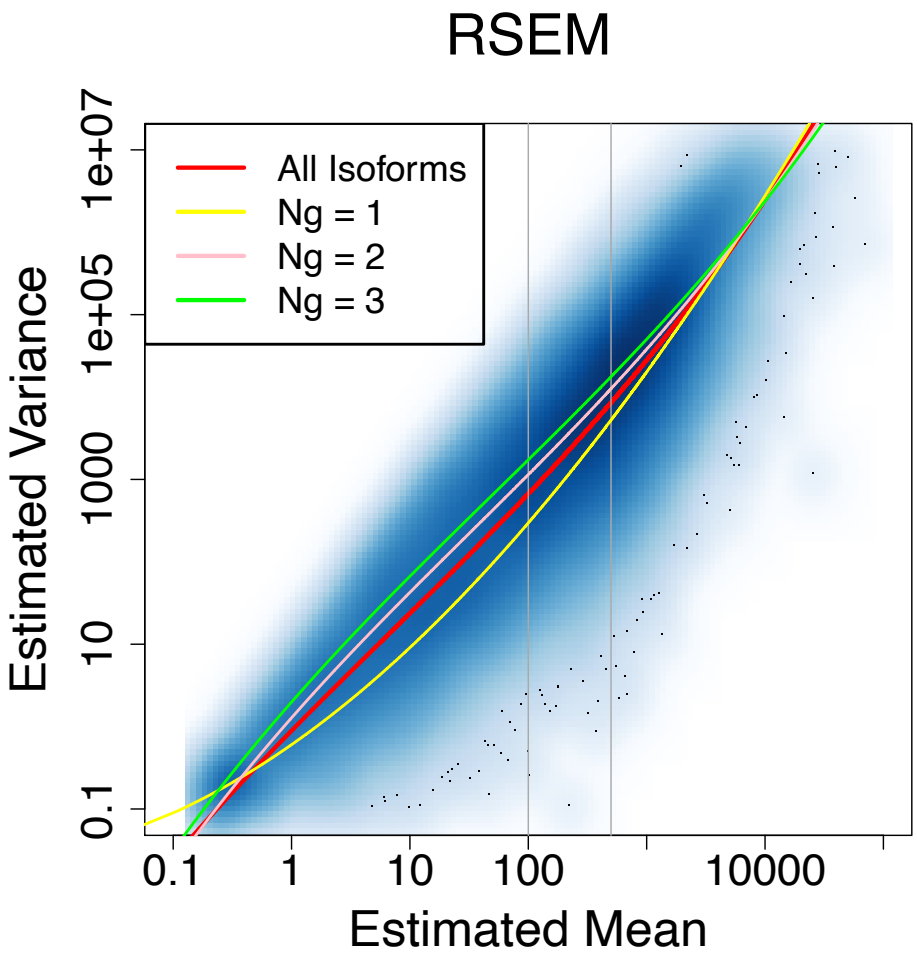
$N_{g'} = 2$

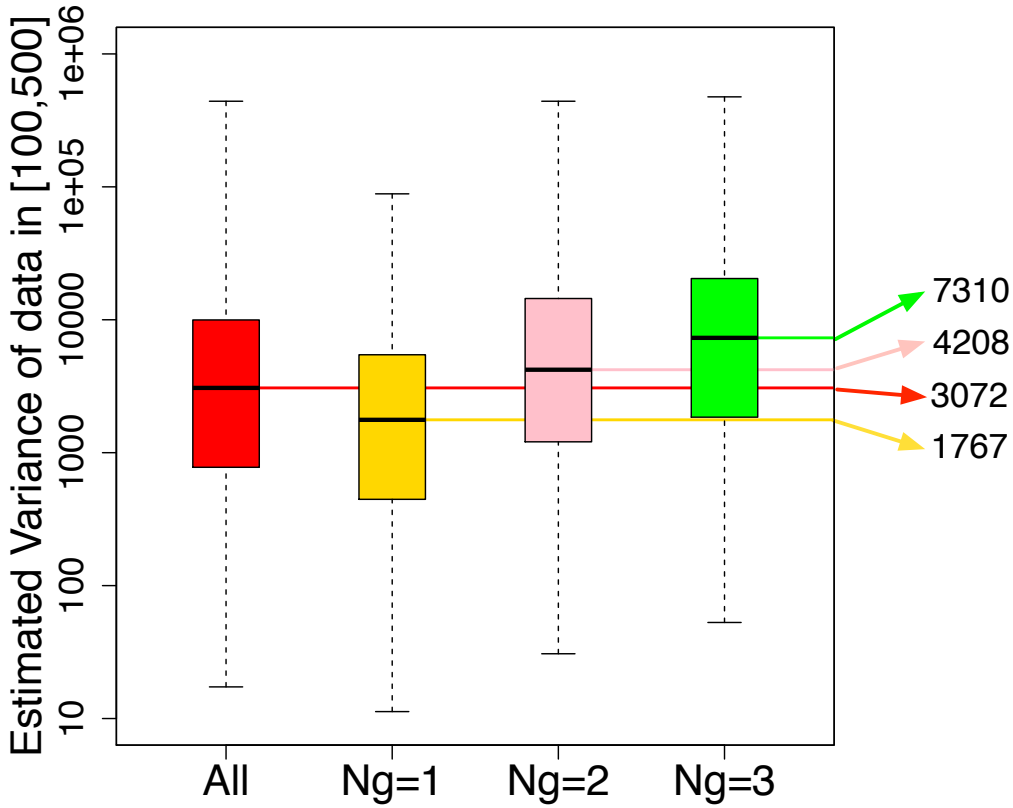
Expression estimates for complex isoforms
(those from N_g>1 genes)
have increased variability

Mean-var relationship changes with N_g

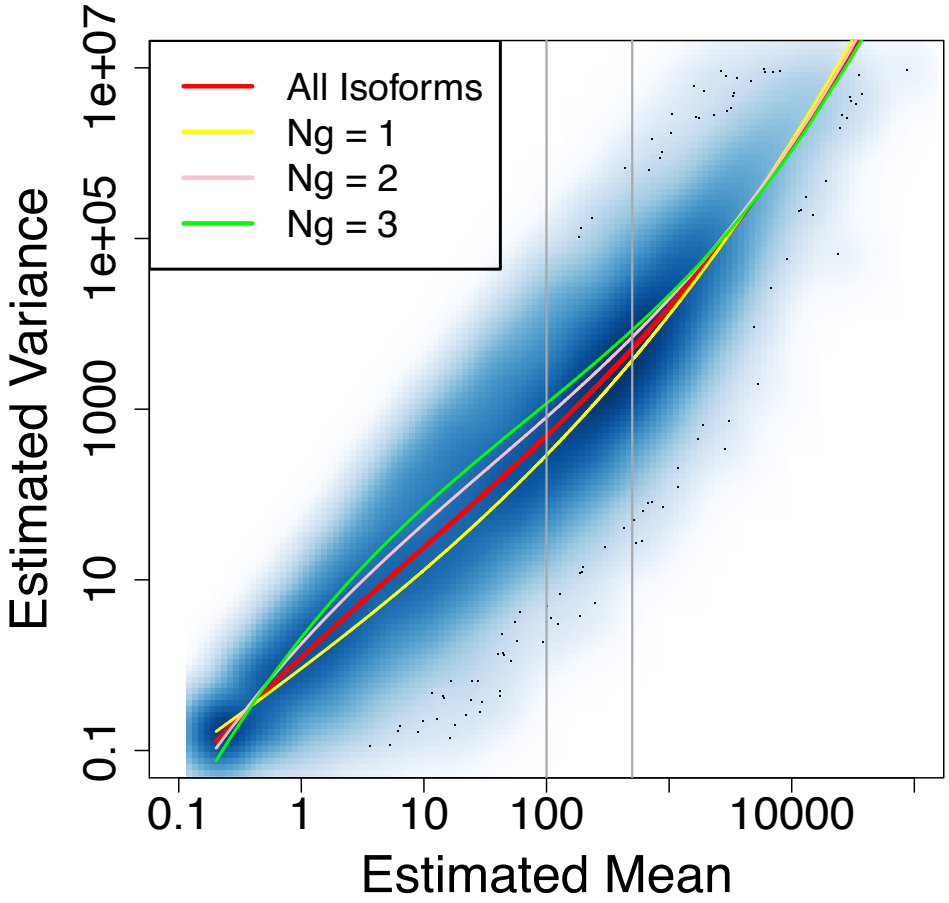


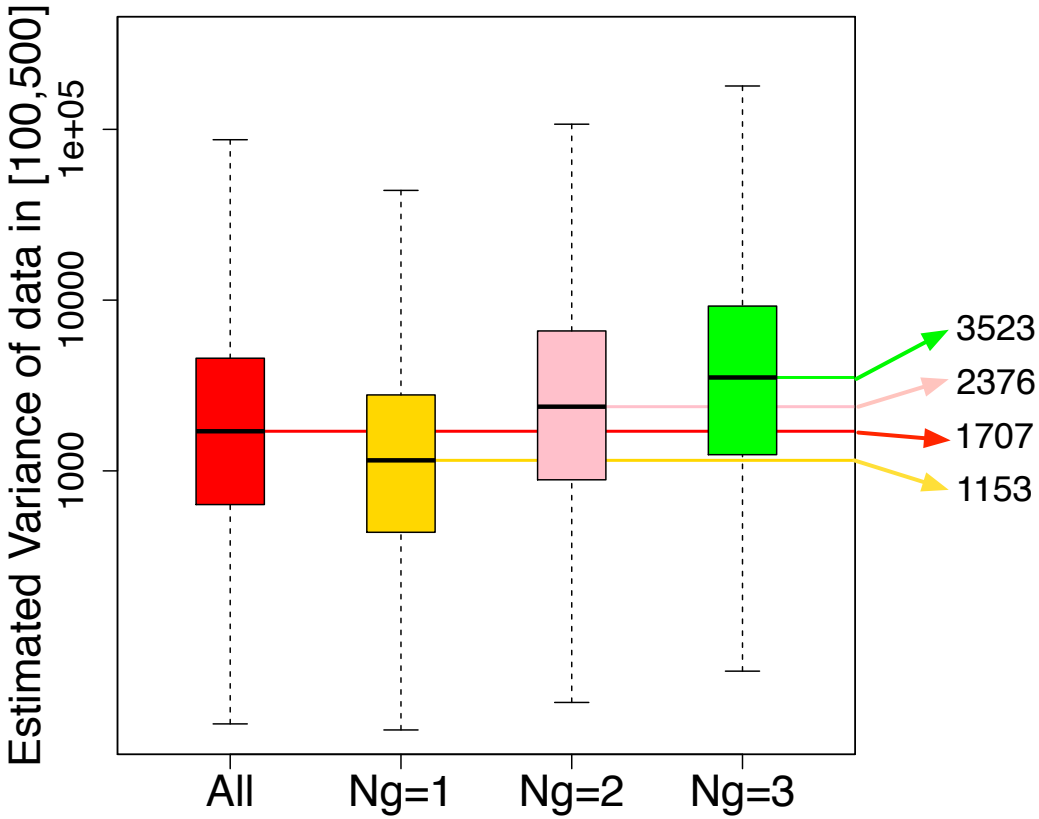
RSEM processed Thomson lab data



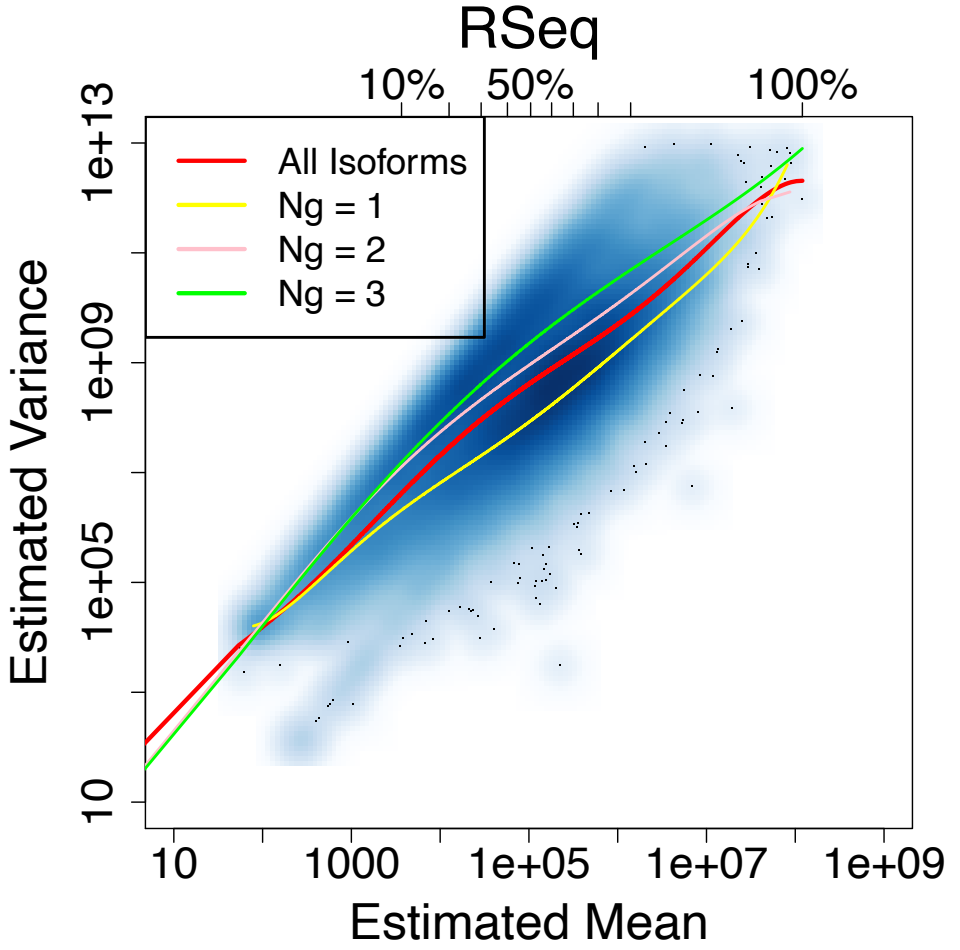


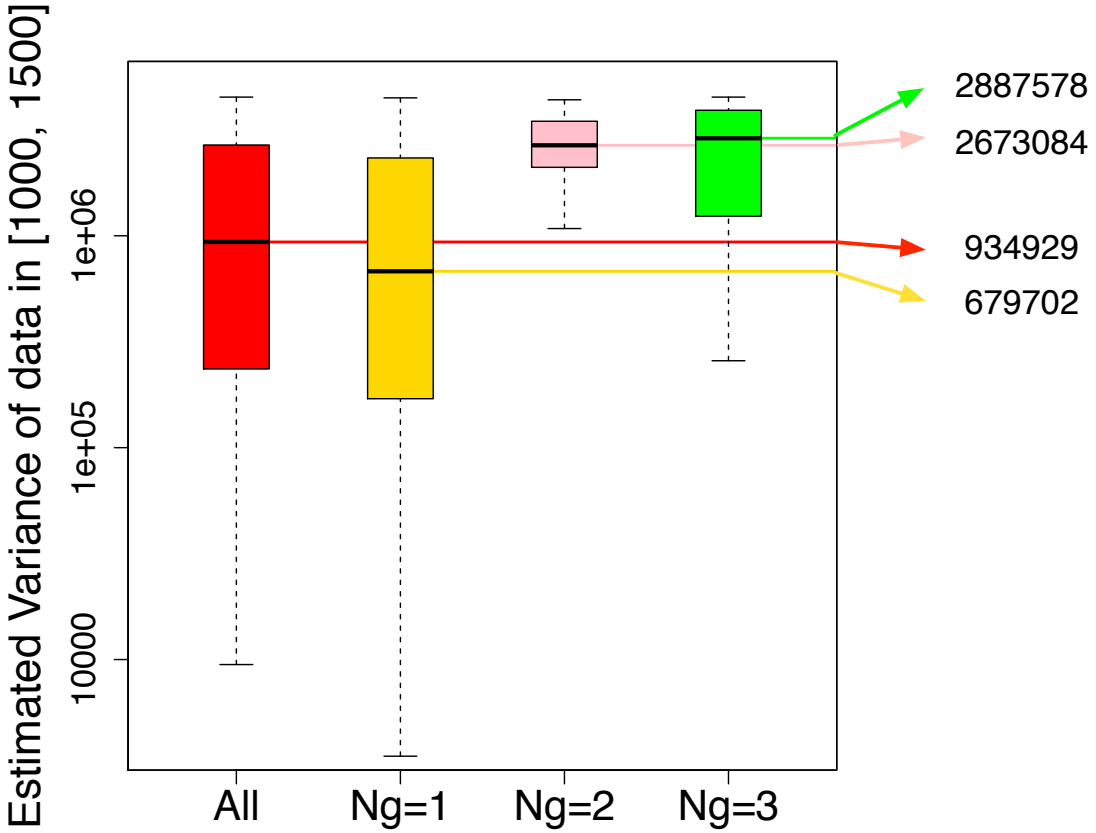
RSEM processed Gould lab data



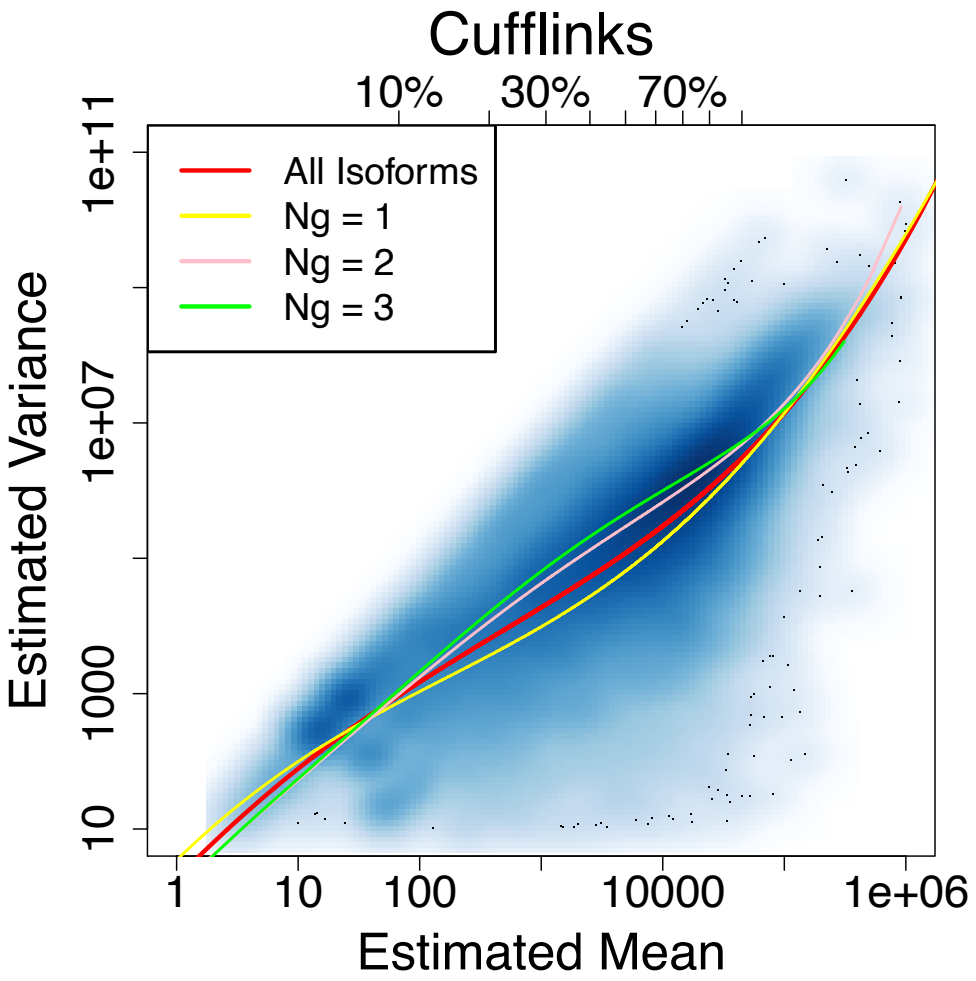


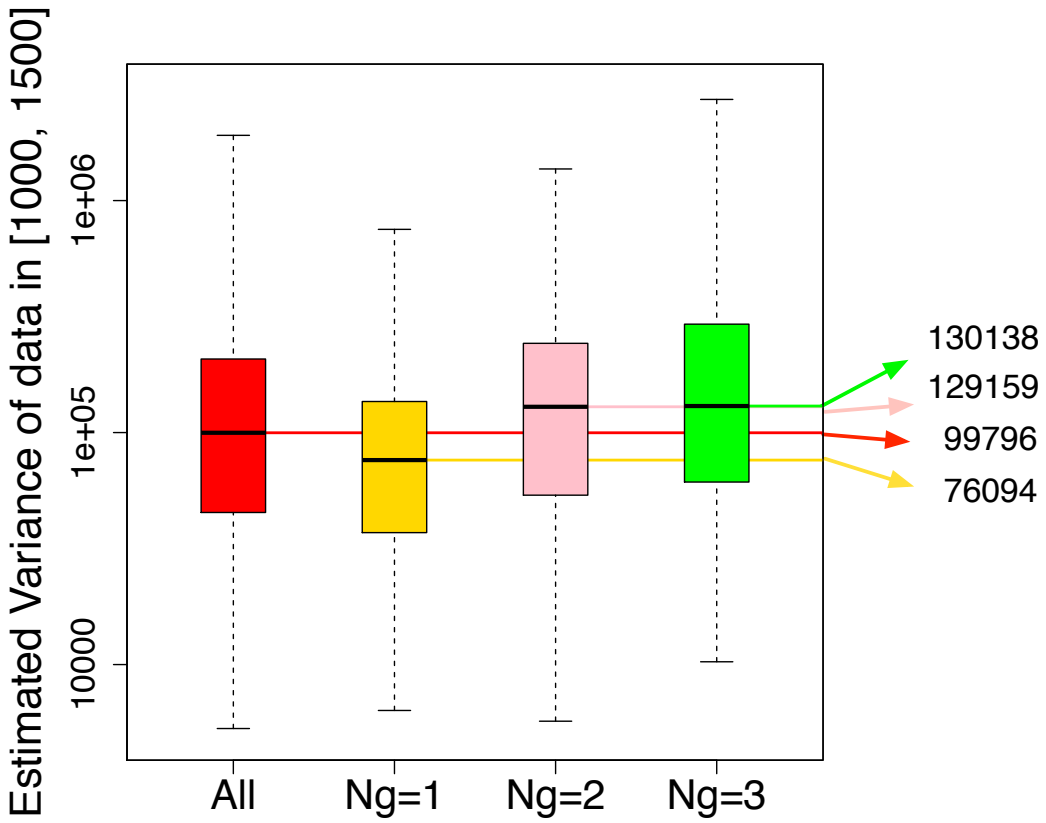
RSeq processed MAQC brain data





Cufflinks processed Wold lab data



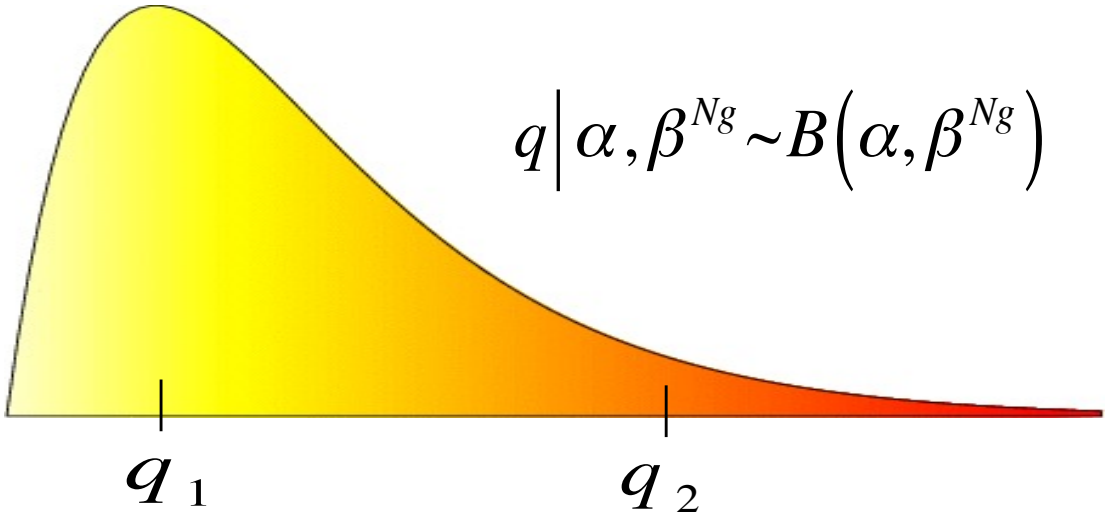


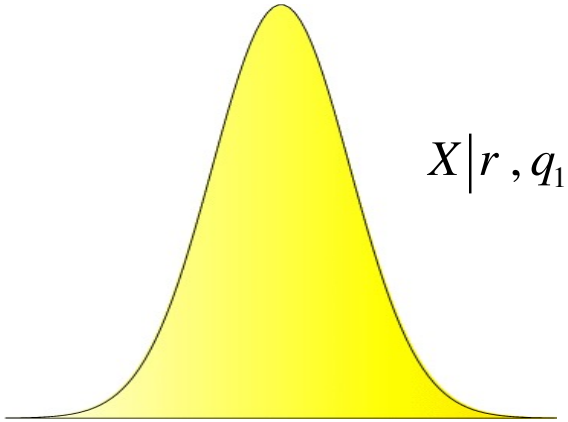
EBSeq: An Empirical Bayes Method for Identifying Differentially Expressed Genes and Isoforms in an RNA-seq experiment

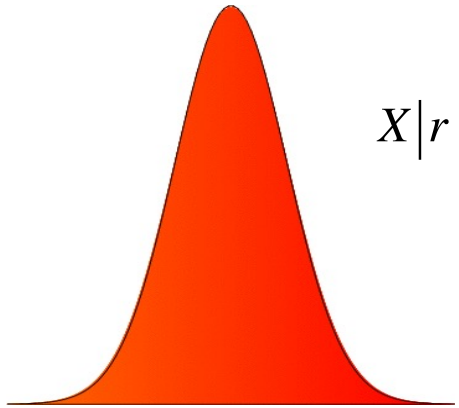
Leng *et al*., *Bioinformatics* 2013

EBSeq: An empirical Bayes NB-Beta Model

$$X|r,q \sim NB(r,q) \text{ and } q|\alpha,\beta^{Ng} \sim B(\alpha,\beta^{Ng})$$







EBSeq

X : Expression of isoform i in gene g and sample s

p_0, p_1 : The prior probability of being EE, DE

Prior depends on Ng

$$X \mid r, q^C \sim NB(r, q^C) \equiv NB\left(\mu = \frac{r(1-q^C)}{q^C}, \sigma^2_{gi,s} = \frac{r\,(1-q^C)}{(q^C)^2}\right); \; q^C \mid \alpha, \beta^{Ng} \sim Beta(\alpha, \beta^{Ng})$$

The isoform is EE if $q^{C1} = q^{C2}$ and DE if $q^{C1} \neq q^{C2}$

Then $X \sim p_0 f_0(X) + p_1 f_1(X)$ where

$$BNB\left(\mu = \frac{r\,\beta^{Ng}}{\alpha - 1}\right)$$

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

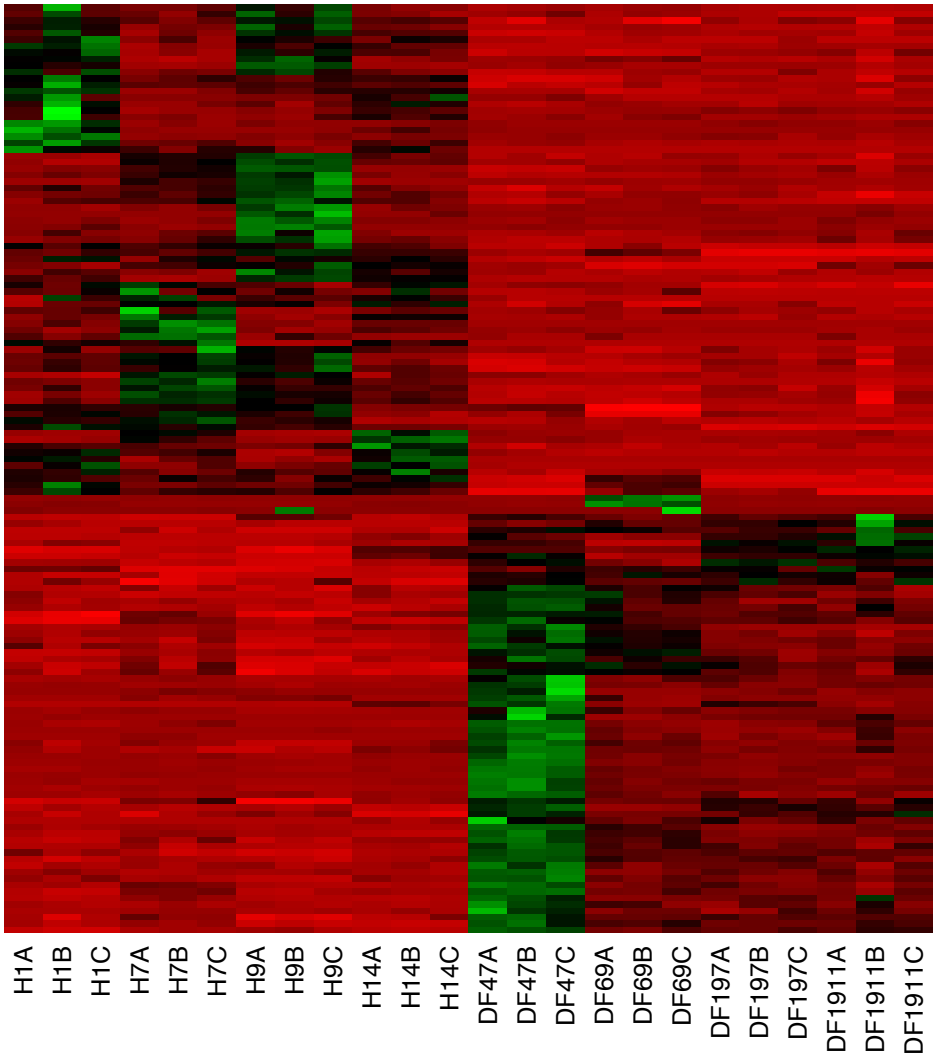
$$f_0\left(X^{C1}\right) f_0\left(X^{C2}\right)$$

DE: $$f_1(X) = \int \prod_{X_{gi,s} \in X^{C1}_{gi}} P(X \mid r, q) P(q \mid \alpha, \beta^{Ng}) dq \int \prod_{X_{gi,s} \in X^{C2}_{gi}} P(X \mid r, q) P(q \mid \alpha, \beta^{Ng}) dq$$

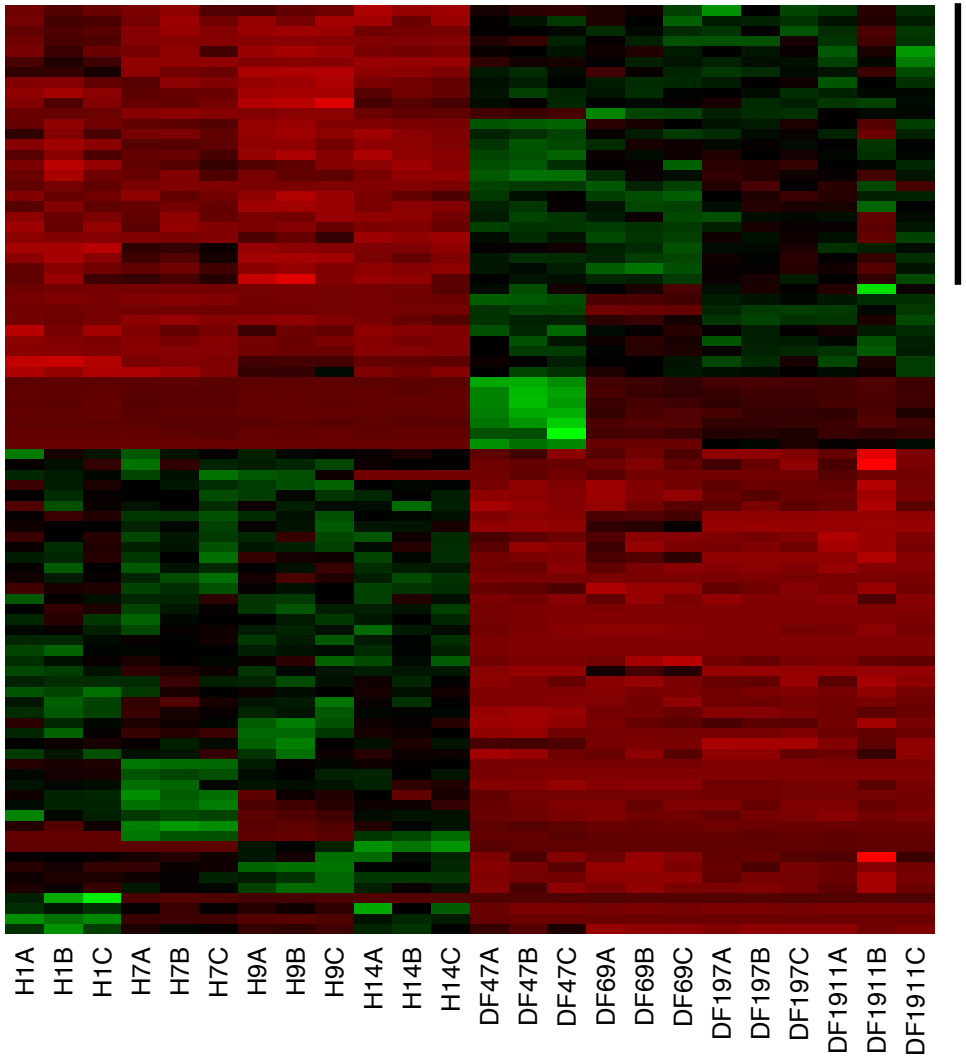
Of primary interest is $P(DE \mid X_{gi}) = \dfrac{p_1 f_1(X)}{p_0 f_0(X) + p_1 f_1(X)}$;

$$P(EE \mid X_{gi}) = \frac{p_0 f_0(X)}{p_0 f_0(X) + p_1 f_1(X)}$$

Identification of potential outliers (ESCs vs. iPSCs)



144 identified by edgeR

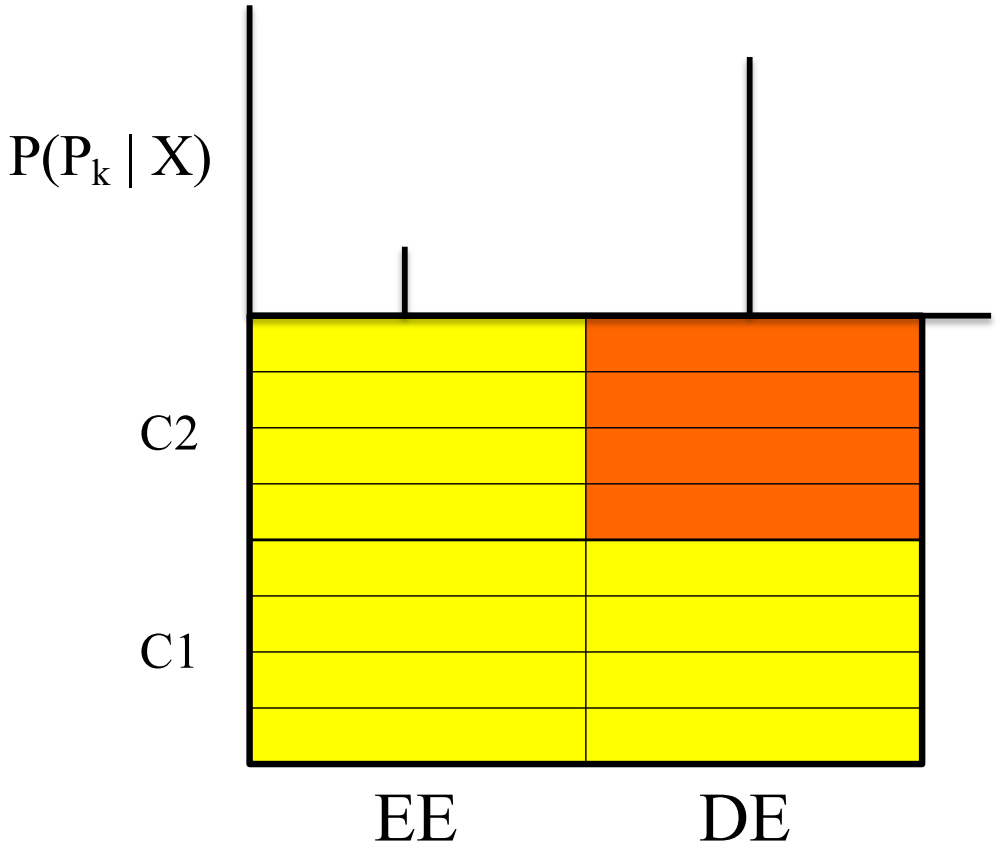


90 from EBSeq

Recall EBSeq for two conditions

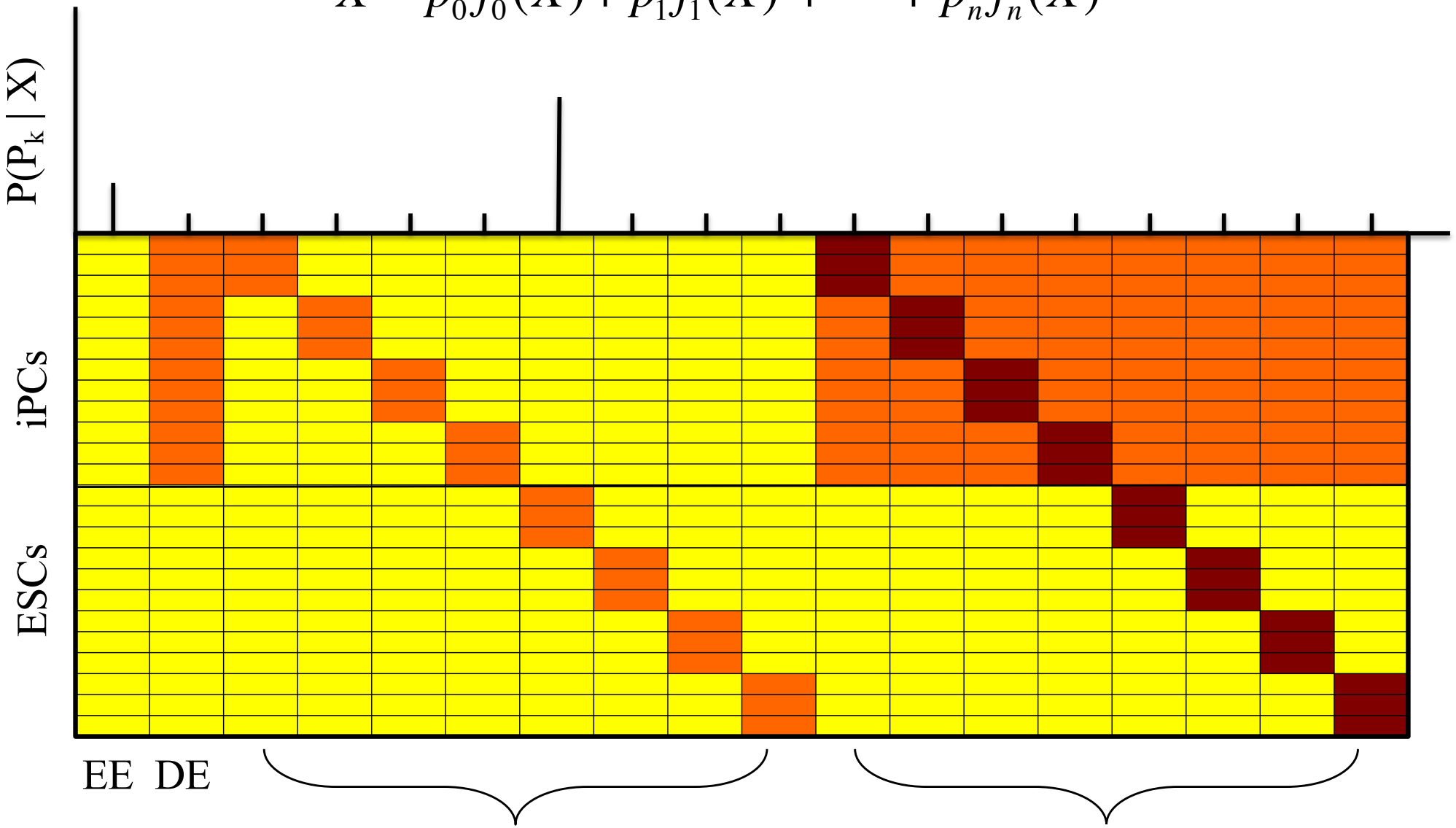
$$\text{EE}: \; q^{C1} = q^{C2} \;\; vs. \;\; \text{DE}: \; q^{C1} \neq q^{C2}$$

$$X \sim p_0 f_0(X) + p_1 f_1(X) \longrightarrow \text{P}\left(\text{DE} \middle| X\right) \text{ and } \text{P}\left(\text{EE} \middle| X\right)$$

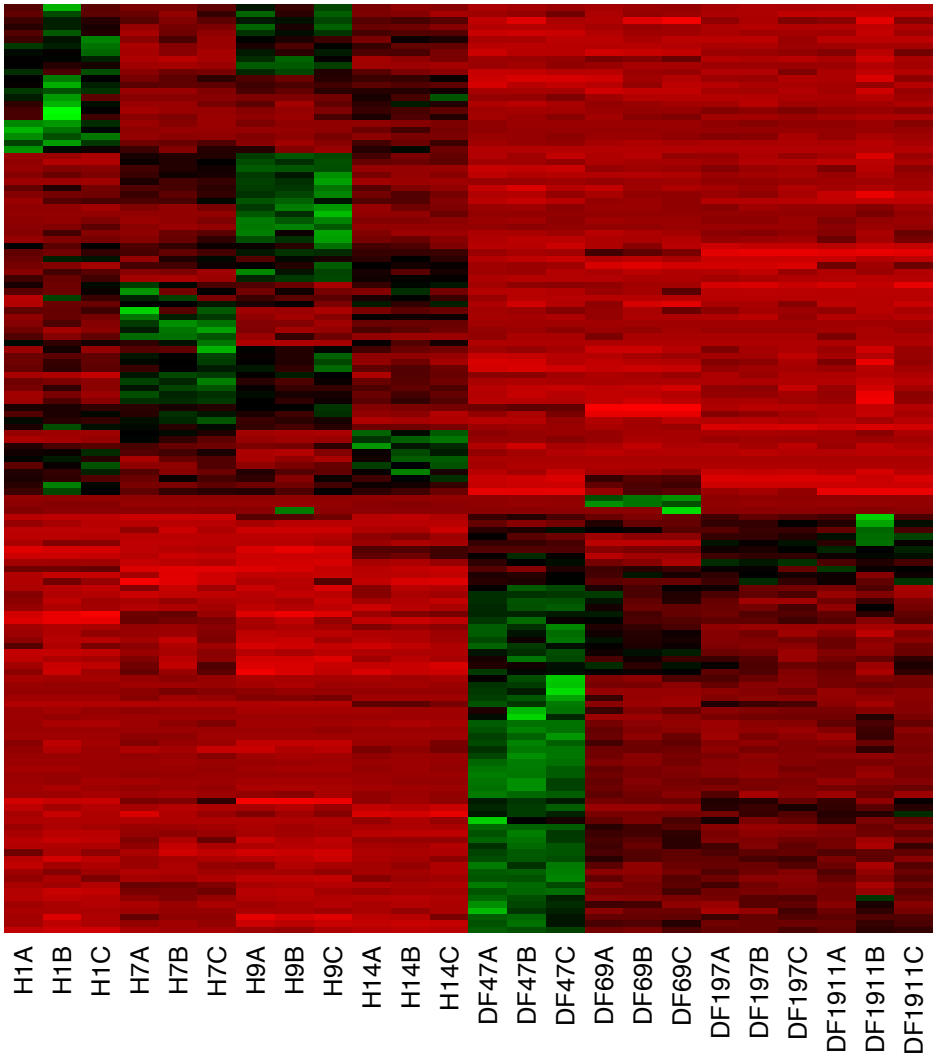


EBSeq for multiple conditions

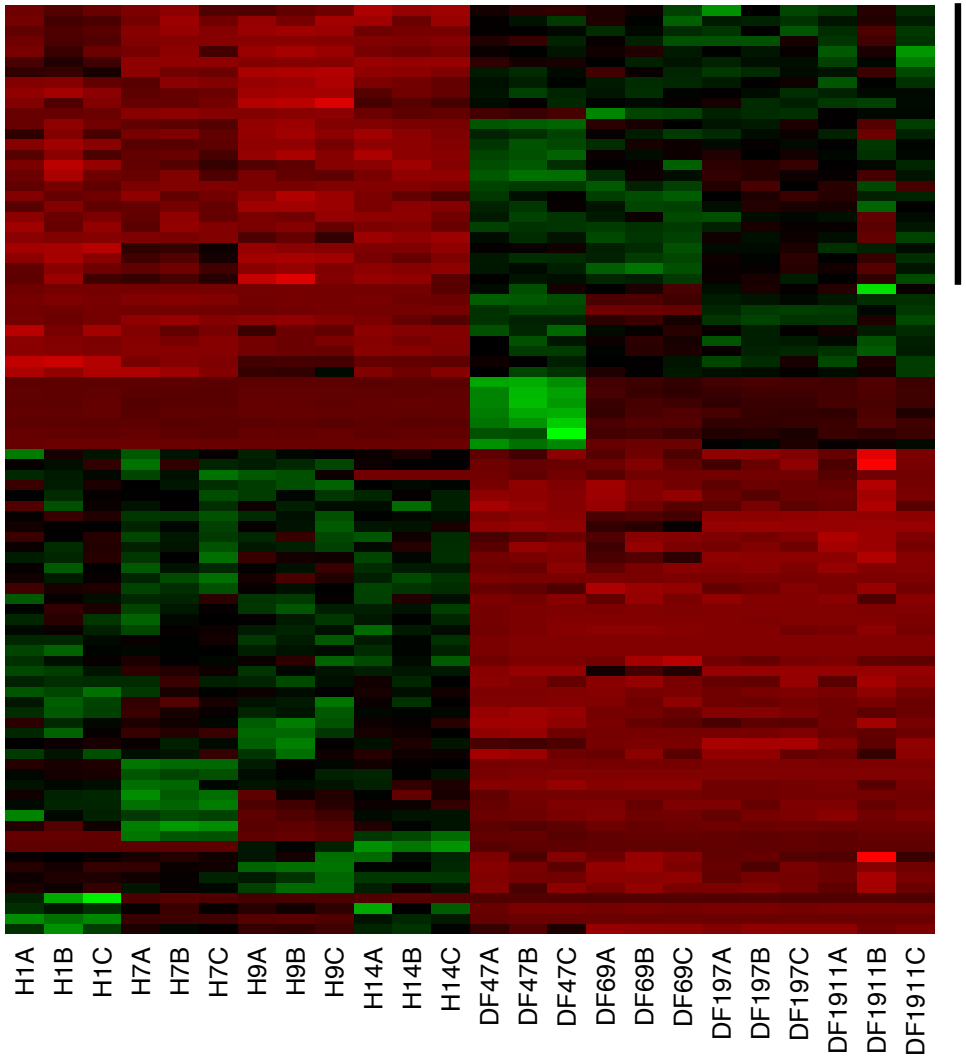
$$X \sim p_0 f_0(X) + p_1 f_1(X) + \cdots + p_n f_n(X)$$



Identification of potential outliers (ESCs vs. iPSCs)

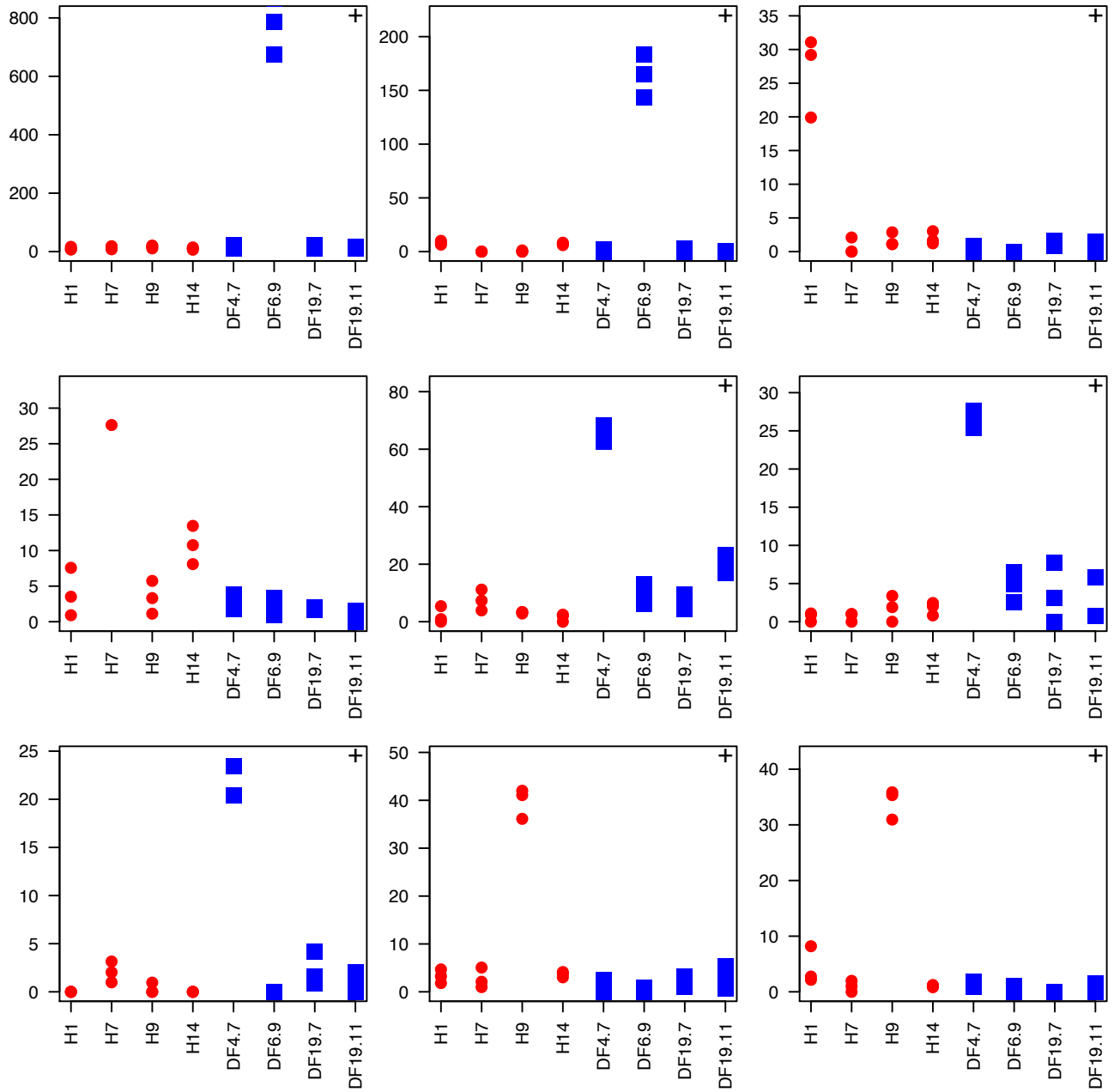


144 identified by edgeR

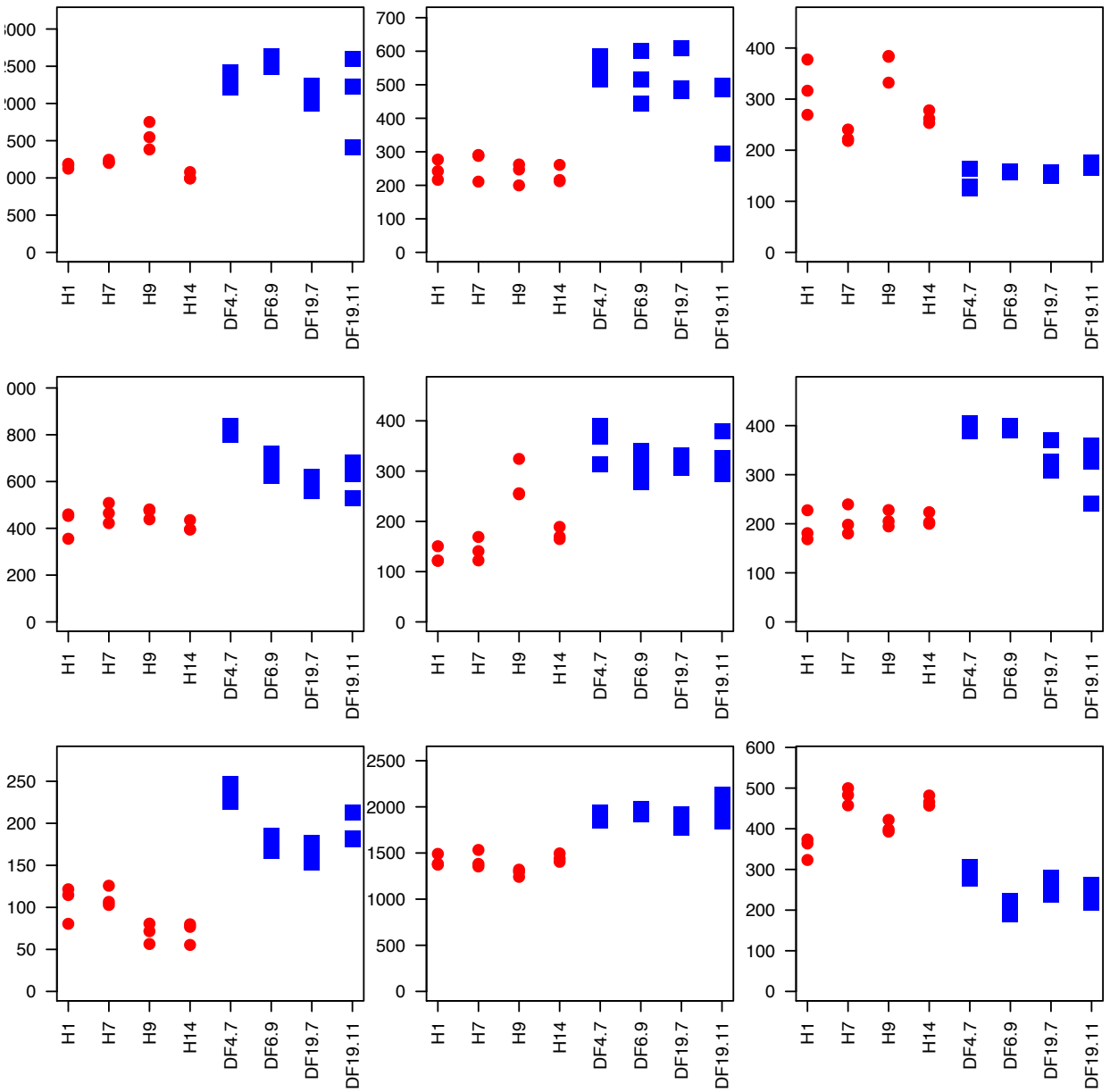


90 from EBSeq

9 genes DE by edgeR; EE by EBSeq (ESCs vs. iPSCs)



9 genes DE by EBSeq; EE by edgeR (ESCs vs. iPSCs)



Summary of EBSeq

- Methods for identifying DE genes in an RNA-seq experiment do not work well for isoform inference as they do not accommodate uncertainty in isoform expression estimation.
- EBSeq identifies both DE isoforms and genes, accommodates uncertainty, and is robust to outliers.
- EBSeq can be used with more than two conditions, and to quantify EE.
- The approach is in BioConductor, Galaxy, and a GUI. Details are in Leng *et al*., *Bioinformatics*, 2013.
- EBSeq-HMM is in *Bioinformatics*, 2015.