Statistical Modeling of RNA-Seq Data

Mingyao Li

Department of Biostatistics and Epidemiology
University of Pennsylvania Perelman School of Medicine

SAGES 2015, Philadelphia

Transcriptomic Variations

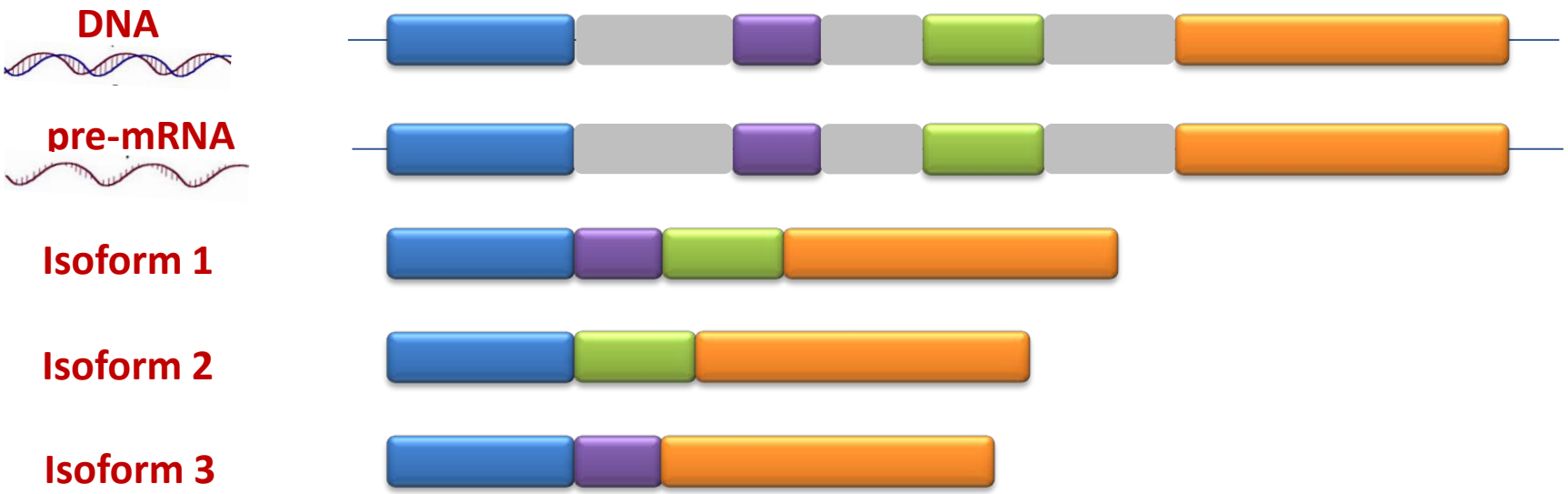
Genome

~21,000 protein coding genes

Transcriptome

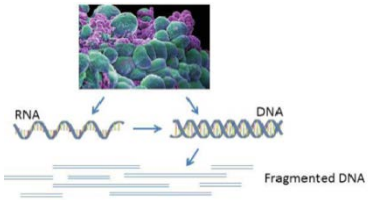
~100,000 human transcripts

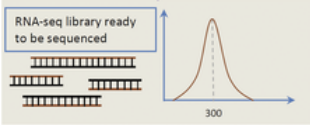
increase in complexity!

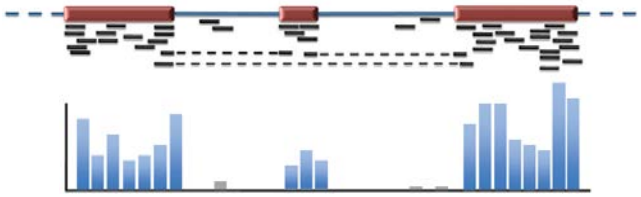


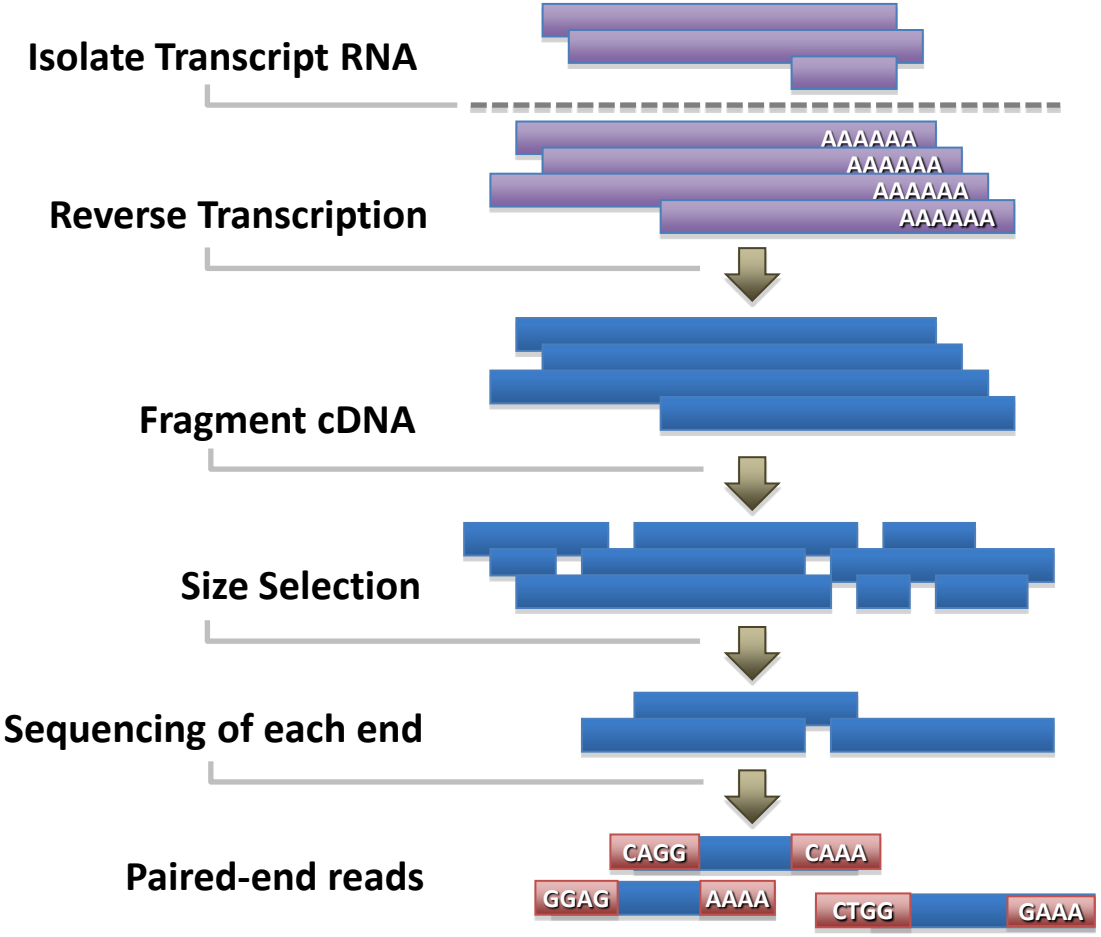
~90% of human genes are alternatively spliced

RNA Sequencing









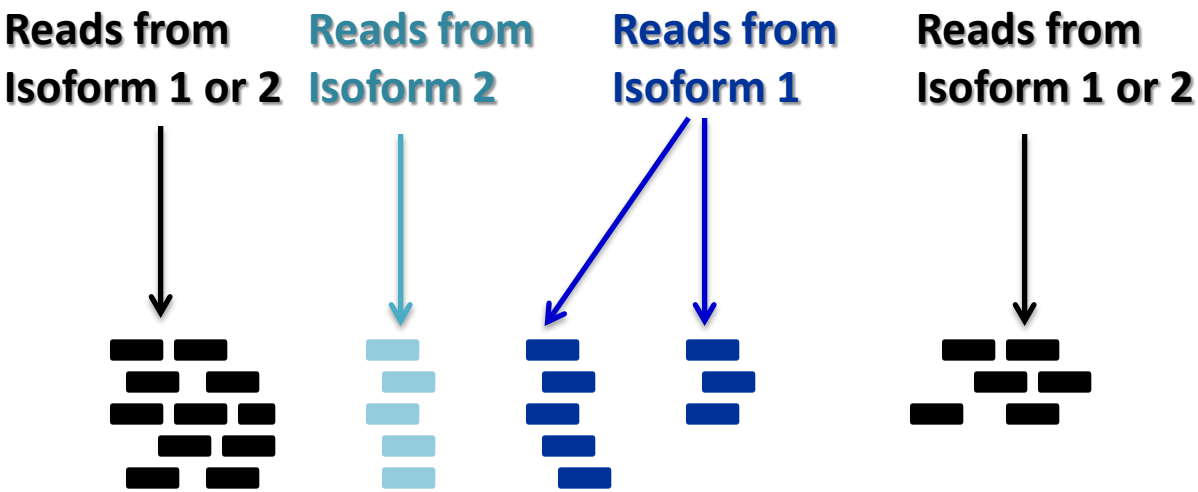
Outline of My Talk

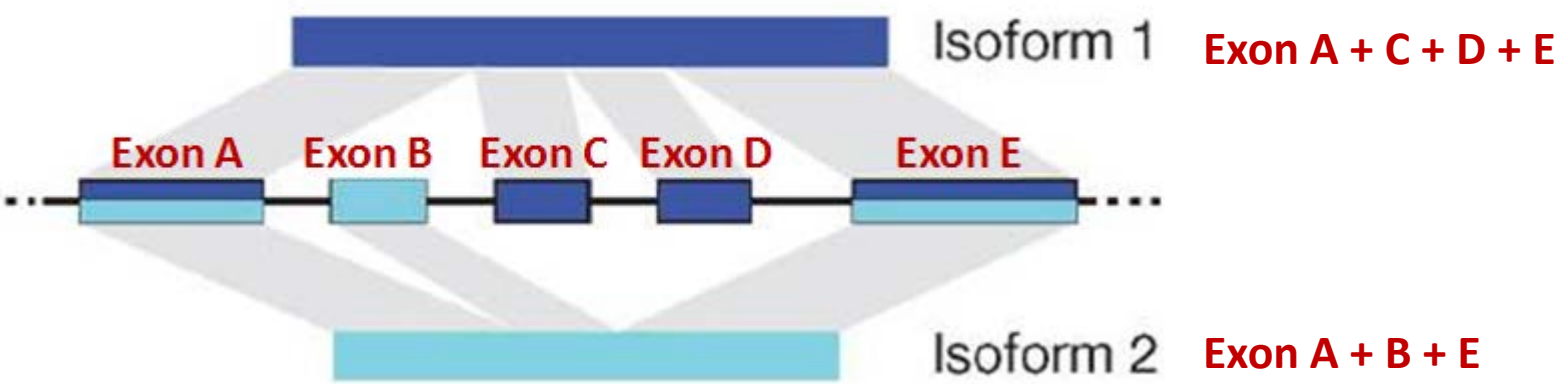
I. Isoform-specific gene expression estimation

II. Isoform-specific differential expression between conditions

III. Differential alternative splicing between conditions

Part I: Isoform-Specific Gene Expression Estimation

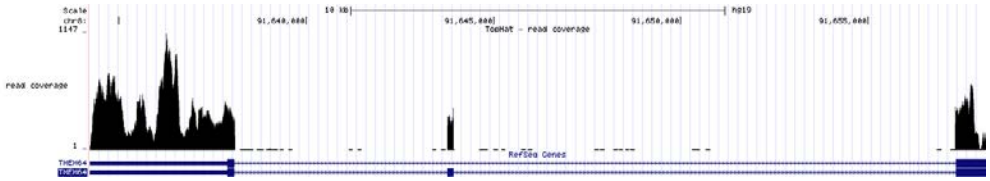




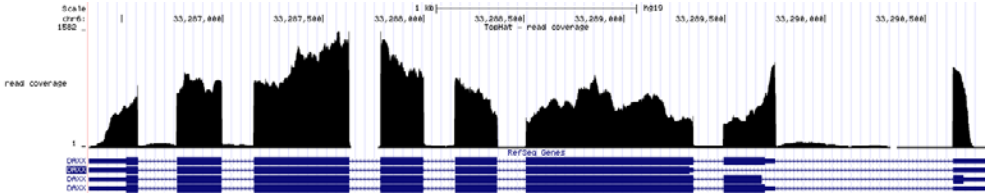
Challenge: Non-uniform Read Distribution

- Most methods assume sequencing reads are uniformly distributed along transcripts
- However, true distributions often deviate substantially from uniformity
- Appropriate modeling of non-uniformity is critical for accurate estimation of isoform expression

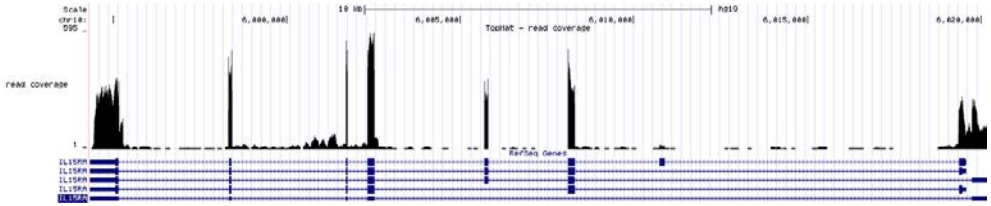
TMEM64



DAXX



IL15RA



Our Approach — PennSeq

- Existing methods often take parametric-based approaches
- Non-uniform read distributions can vary substantially from gene to gene, or even different isoforms within the same gene
- Parametric models are unlikely to capture all factors that lead to non-uniformity
- Our goal: develop a method that allows each isoform to have its own non-uniform distribution
- PennSeq does not make distributional assumptions, but rather let the data speak for themselves

PennSeq. Hu et al. Nucleic Acids Research. 2014, 42(3):e20.

Observed Likelihood

$$L(\Theta \mid \mathbf{R}) = \prod_{r \in \mathbf{R}} P(\text{read} = r, \text{start} = s)$$

$$= \prod_{r \in \mathbf{R}} \sum_{i \in \mathbf{I}} P(\text{iso.} = i) P(\text{read} = r, \text{start} = s, \text{frag. len.} = L_i(r,s) \mid \text{iso.} = i)$$

$$= \prod_{r \in \mathbf{R}} \sum_{i \in \mathbf{I}} \tilde{\theta}_i \boxed{P(\text{read} = r \mid \text{start} = s, \text{frag. len.} = L_i(r,s), \text{iso.} = i)}$$

$$\times \boxed{P(\text{start} = s, \text{frag. len.} = L_i(r,s) \mid \text{iso.} = i)}$$

$$P(\text{read} = r \mid \text{start} = s, \text{frag. len.} = L_i(r,s), \text{iso.} = i) = \prod_{j=1}^{m} q_j(x_j, y_{i,j+s-1})$$

<u>Key to the above likelihood calculation is on the modeling of</u>

$$P(\text{start} = s, \text{frag. len.} = L_i(r,s) \mid \text{iso.} = i) = h_i(r,s)$$

Read Start Distribution

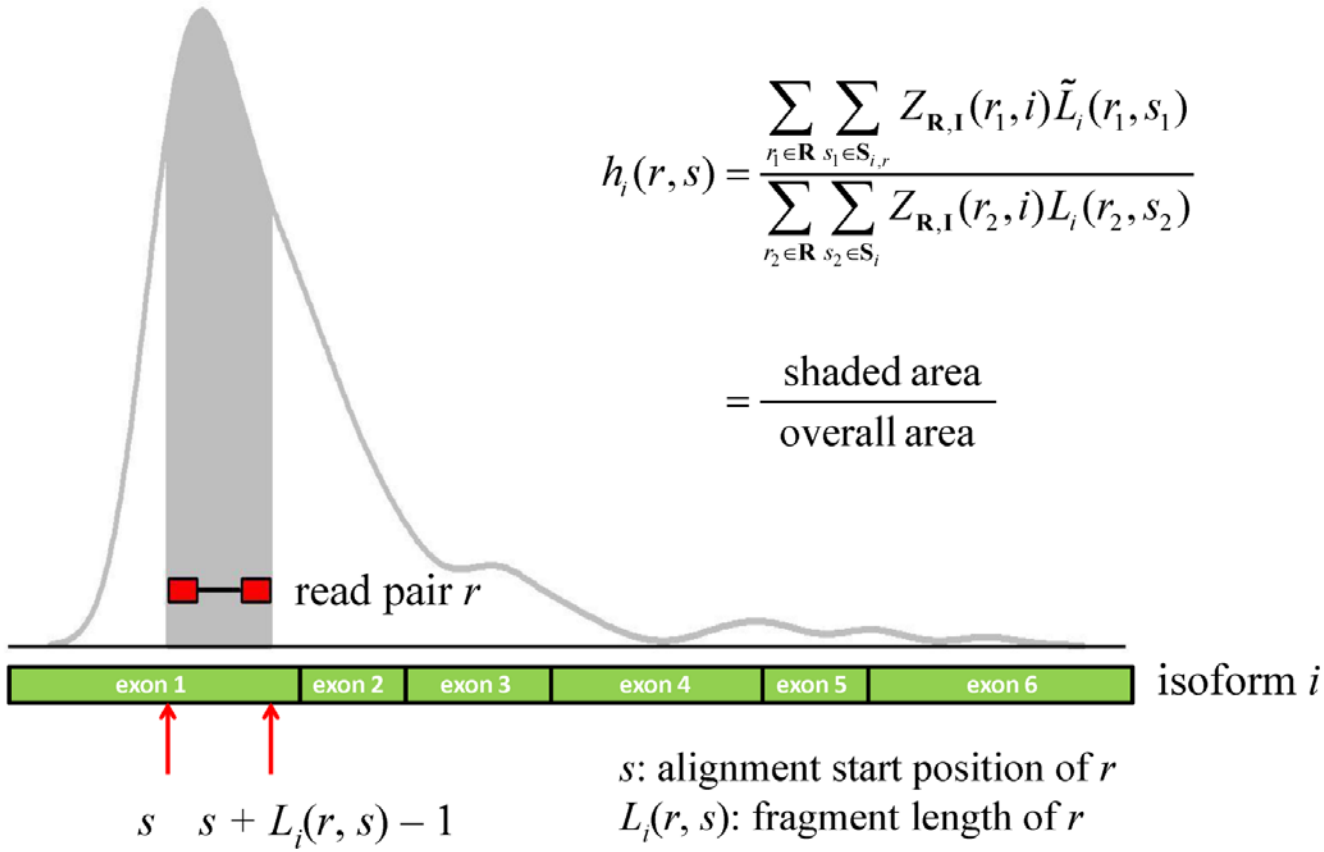
Most existing methods assumes that the read start position is <u>uniformly distributed</u>, i.e.,

$$h_i(r,s) = \frac{1}{\tilde{l}_i - L_i(r,s) + 1}$$

Our approach

$$h_i(r,s) = \frac{\sum_{r_1 \in \mathbf{R}} \sum_{s_1 \in \mathbf{S}_{i,r}} Z_{\mathbf{R},\mathbf{I}}(r_1,i)\tilde{L}_i(r_1,s_1)}{\sum_{r_2 \in \mathbf{R}} \sum_{s_2 \in \mathbf{S}_i} Z_{\mathbf{R},\mathbf{I}}(r_2,i)L_i(r_2,s_2)}$$

$$= \frac{\text{shaded area}}{\text{overall area}}$$



s: alignment start position of r
$L_i(r, s)$: fragment length of r

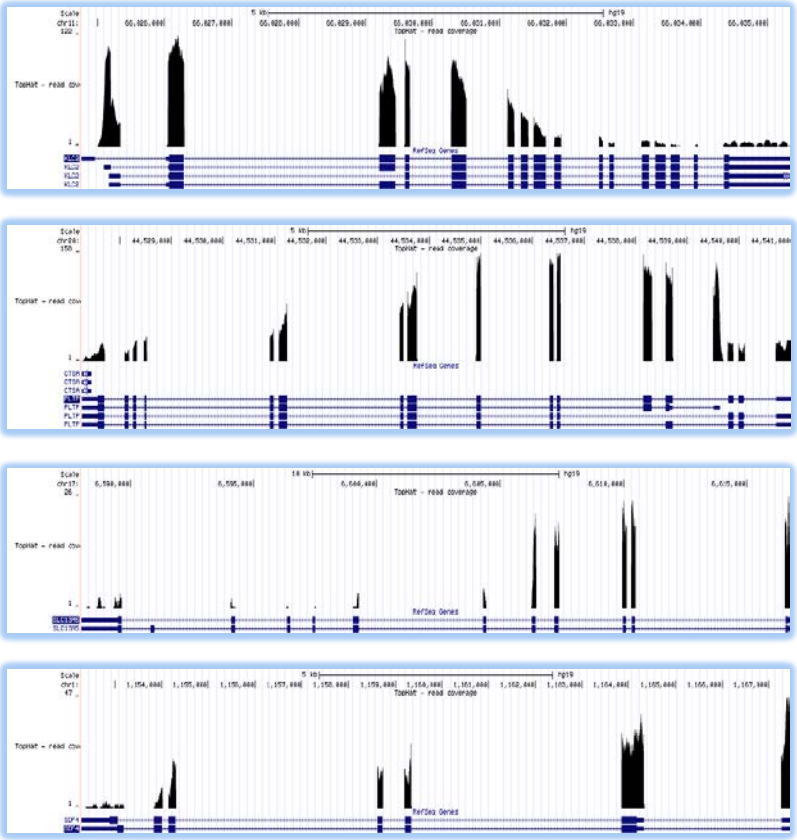
Simulation Setup

http://sammeth.net/confluence/display/SIM

- Simulate systematic bias in the abundance and distribution of produced reads by *in silico* library preparation and sequencing

- 100 million (M) paired-end reads
- Randomly selected 10M, 20M, and 60M reads to evaluate the impact of sequencing depth

Read coverage in selected genes

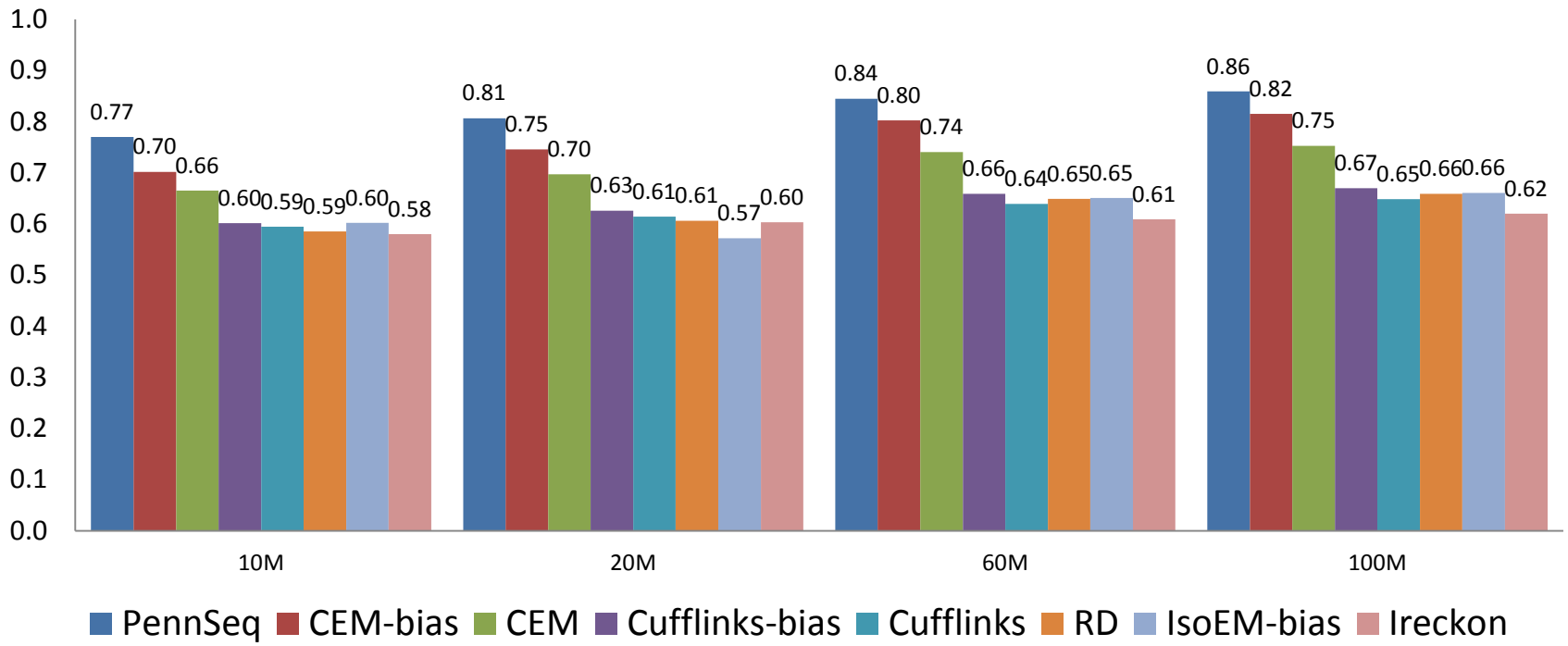


Griebel et al. (2012) Nucleic Acids Research

Comparison of Estimation Accuracy

Measure of estimation accuracy

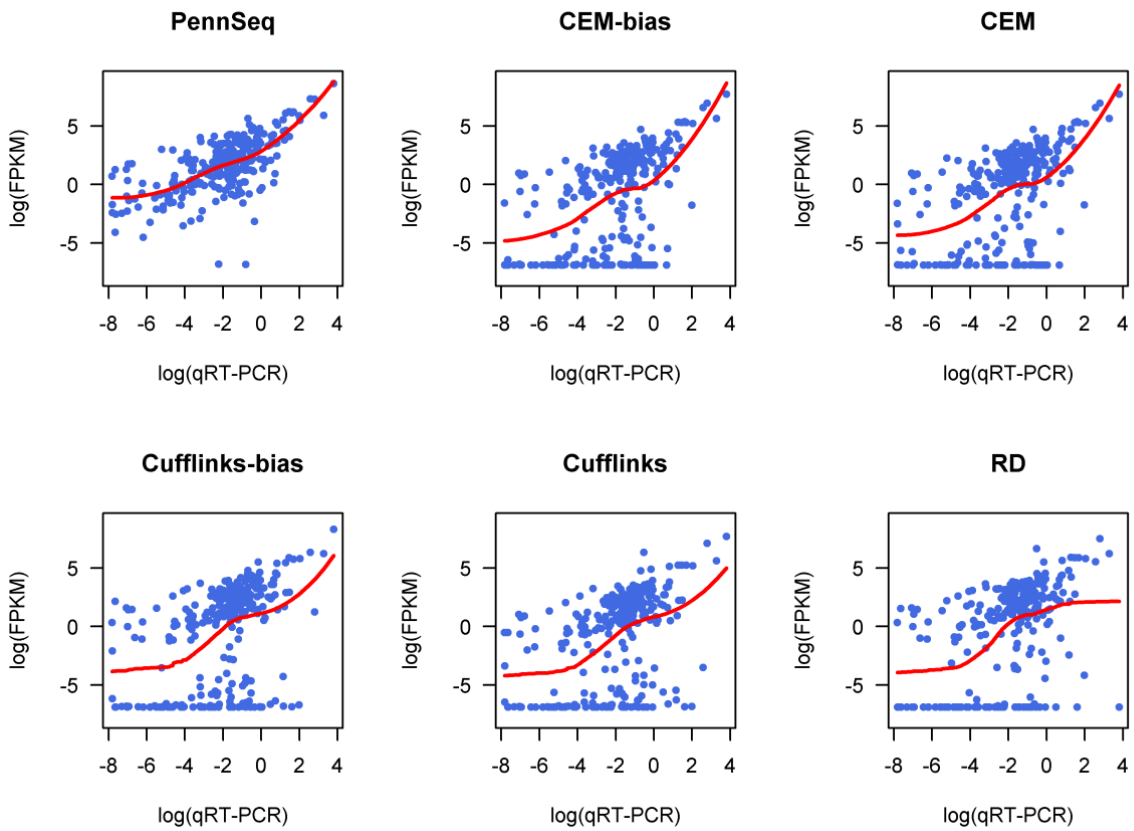
R^2 coefficient of determination (i.e. squared Pearson correlation) between estimated isoform relative abundance and true value.



Comparison based on Benchmark Data

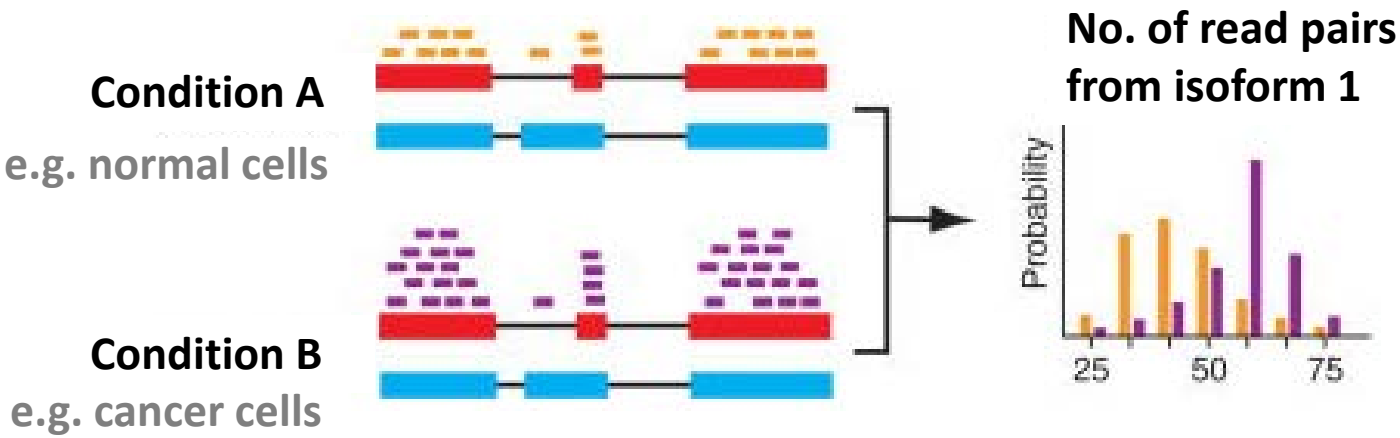
MAQC (MicroArray Quality Control data)

qRT-PCR measurements available (treated as gold standard)



Isoforms with underestimated expression levels are typically from genes with severe non-uniformity and low-to-moderate coverage.

Part II: Isoform-Specific Differential Expression



Analytical challenges

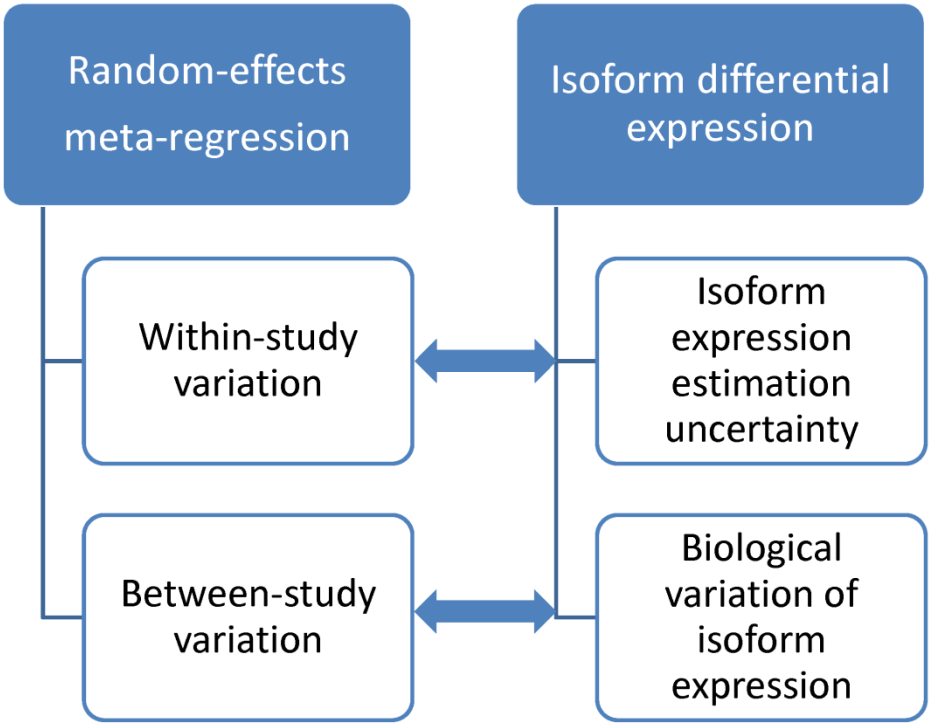
- Several sources of variation
  - Isoform expression estimation uncertainty
  - Variations across biological replicates
- Influence from covariates/confounders
  - E.g. age, gender, environment etc

Existing Methods and Limitations

- Cuffdiff, baySeq, EBSeq, NOIseq
  - Account for isoform expression estimation uncertainty
  - Cannot adjust covariates/confounders
- DESeq, DESeq2, edgeR
  - Can adjust covariates/confounders
  - Count based methods
  - Cannot model isoform expression estimation uncertainty

Our Approach — MetaDiff

Goal of random-effects meta-regression: synthesize results of multiple studies to test moderator effect



MetaDiff. Jia et al. BMC Bioinformatics: in press.

Model Setup

$$\boldsymbol{log(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 Z_i + U_i + e_i}$$

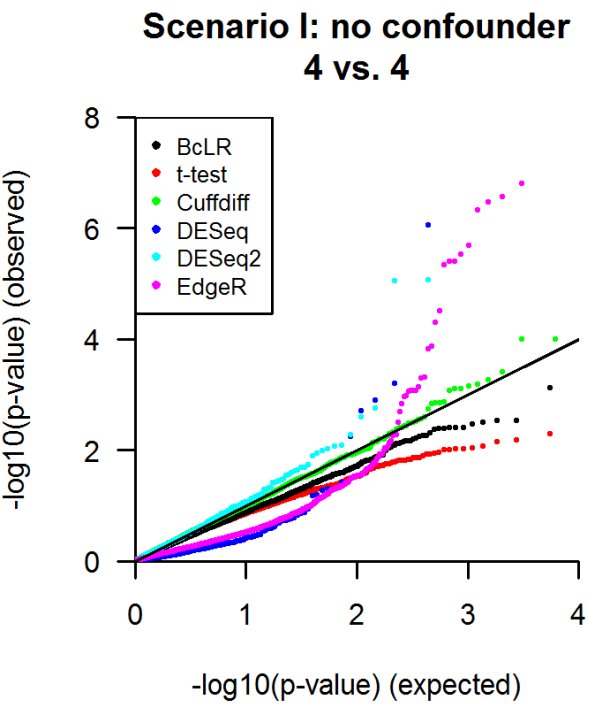
Y_i: estimated isoform expression level for subject *i*
X_i: phenotype of interest for subject *i,* e.g., disease status
Z_i: covariate/confounder variable, e.g., age, gender

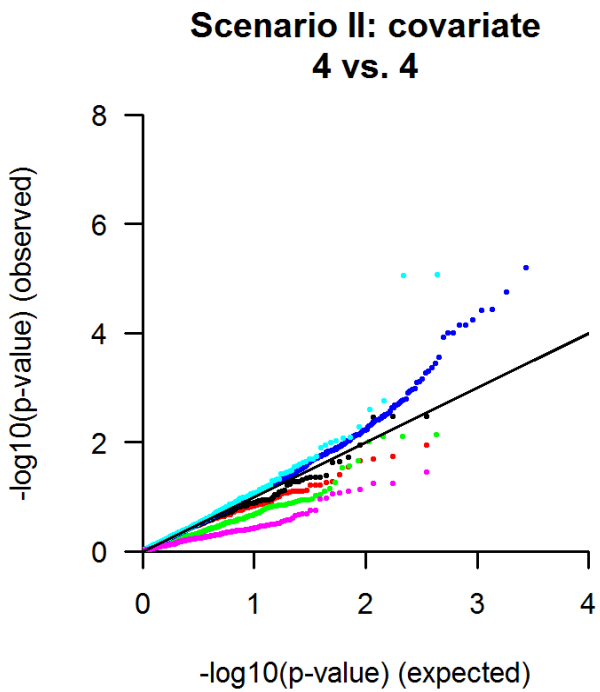
U_i: error term due to isoform expression estimation uncertainty (**within sample variation**)
$U_i \sim N(0, {\sigma_i}^2)$, where ${\sigma_i}^2$ is known

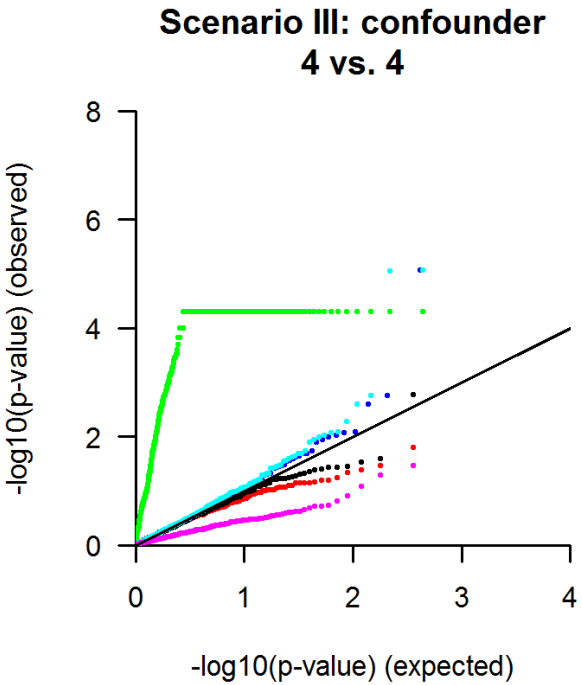
e_i: error term due to unmodeled differences between subjects (**between sample variation**)
$e_i \sim N(0, \tau^2)$, where τ^2 is unknown

Test: likelihood ratio test (BcLR) or t-test

Comparison with Other Methods



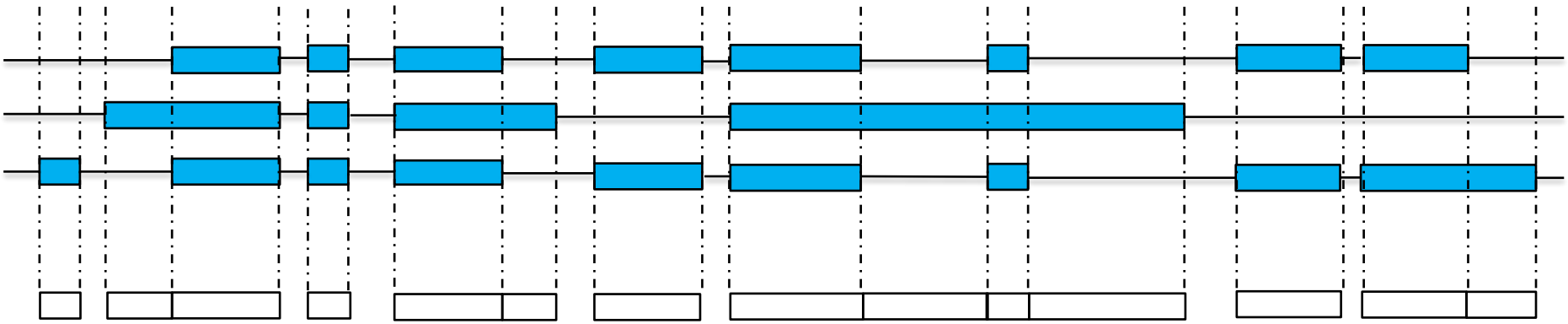




Part III: Differential Alternative Splicing

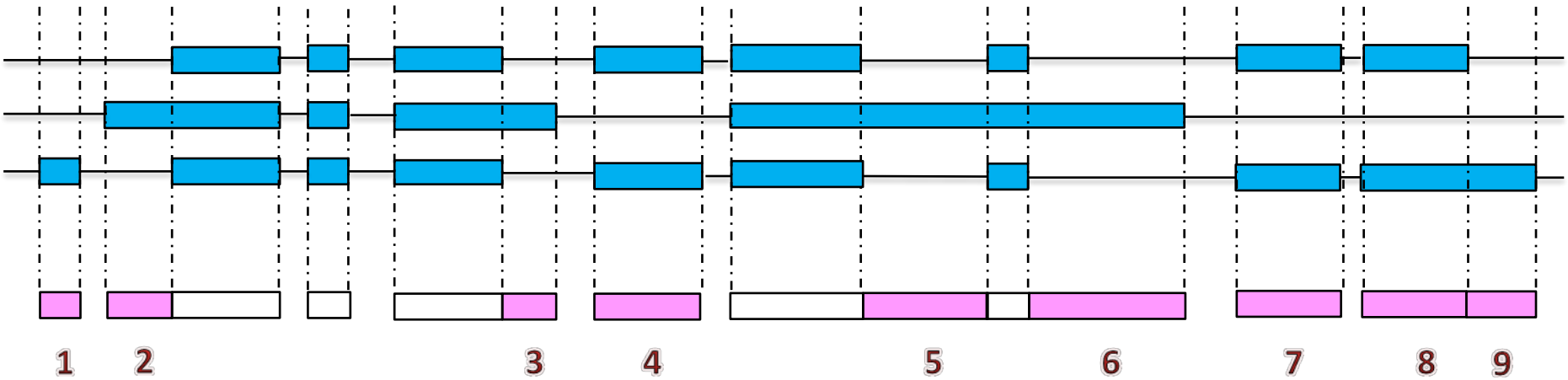
- Exon-based methods
  - Compare exon-inclusion levels (i.e., fraction of transcripts with the exon included) between conditions
  - Software: MISO, MATS/rMATS, DEXSeq
- Gene-based methods
  - Compare isoform relative abundances between conditions
  - Software: Cuffdiff, Splicing Compass, DiffSplice, IUTA

Gene Structure



14 virtual exons

Gene Structure

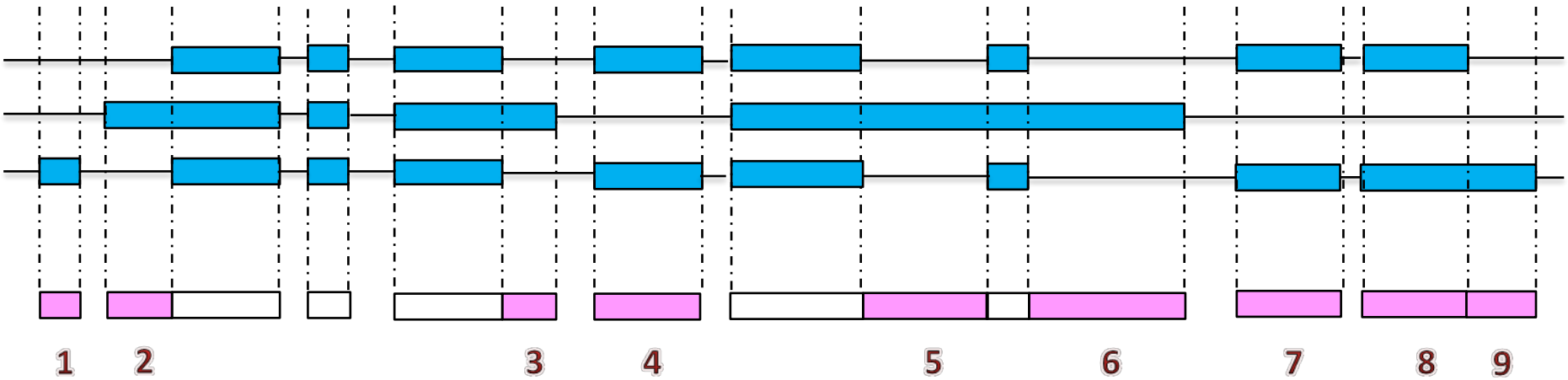


14 virtual exons

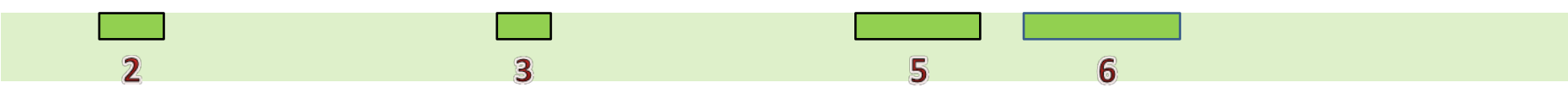
<u>Only alternatively spliced exons are informative for DAS</u>

- **9 informative for AS**
- 5 uninformative
- **DEXSeq:** test differential exon usage for all 14 virtual exons
- **rMATS:** terminal exons cannot be tested due to requirement of flanking exons

Grouping of Alternatively Spliced Exons



exon-inclusion level = θ_3



exon-inclusion level = θ_2

exon-inclusion level = $\theta_1 + \theta_3$

Our Approach — PennDiff

Stage I: quantification of AS using exon-inclusion level

- Estimate isoform relative abundances for a given gene using existing software (e.g., PennSeq, Cufflinks, RSEM etc.)
- Estimate exon-inclusion level for each alternatively spliced exon *e* in subject i : $x_{i,e} = \sum_{j \in I_e} \theta_{i,j}$

  where I_e is the set of isoforms with exon *e* included

Note: exons from the same group will have the same exon-inclusion level → only need to perform one test for each group, which will reduce the number of multiple testing

Our Approach — PennDiff

Stage II: detecting DAS between two conditions (A vs. B)

- Assume exon-inclusion level for exon group g in subject i follows a Beta distribution with mean $\mu_{i,m}$ and precision parameter φ_m
- Gaussian copula marginal regression
  - Marginal model: $h[E(X_{i,m})] = logit(\mu_{i,m}) = \beta_0 + \beta_m Z_i$, where

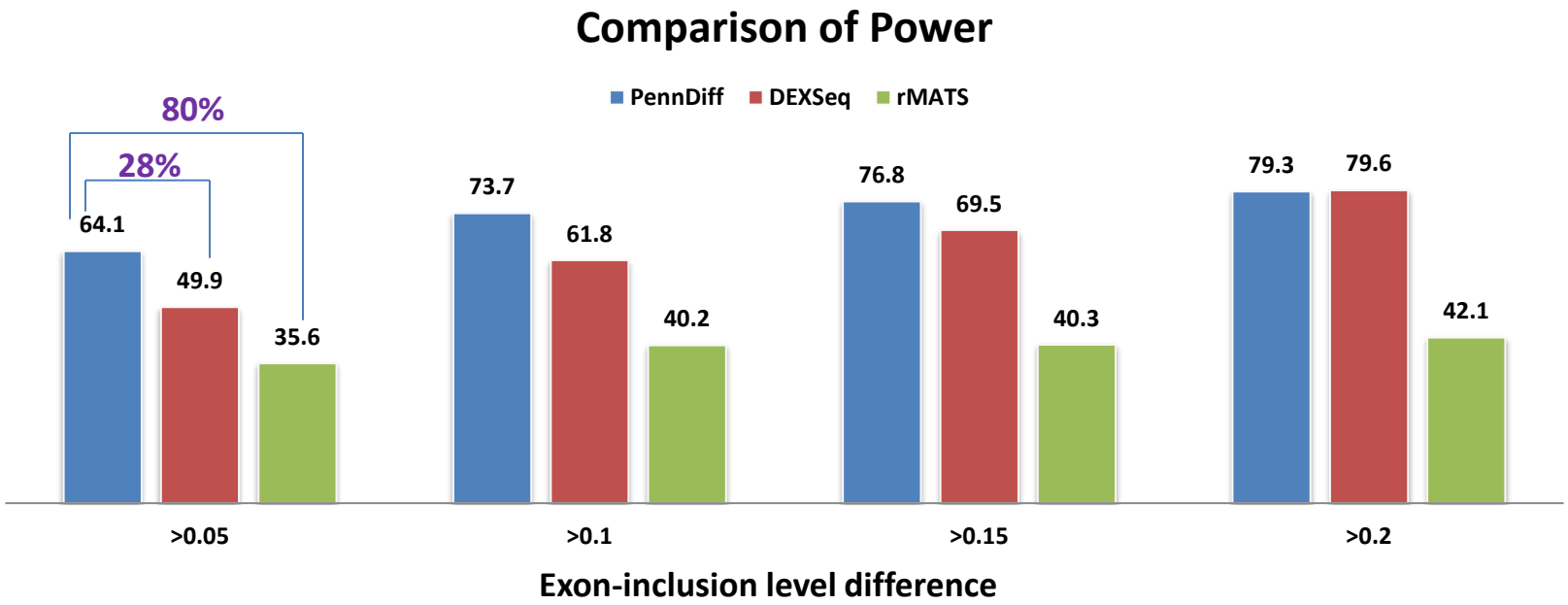
    $X_{i,m}$ is exon-inclusion level for exon group m in subject i
    Z_i is condition indicator for subject i (1 for condition A, 0 for condition B)

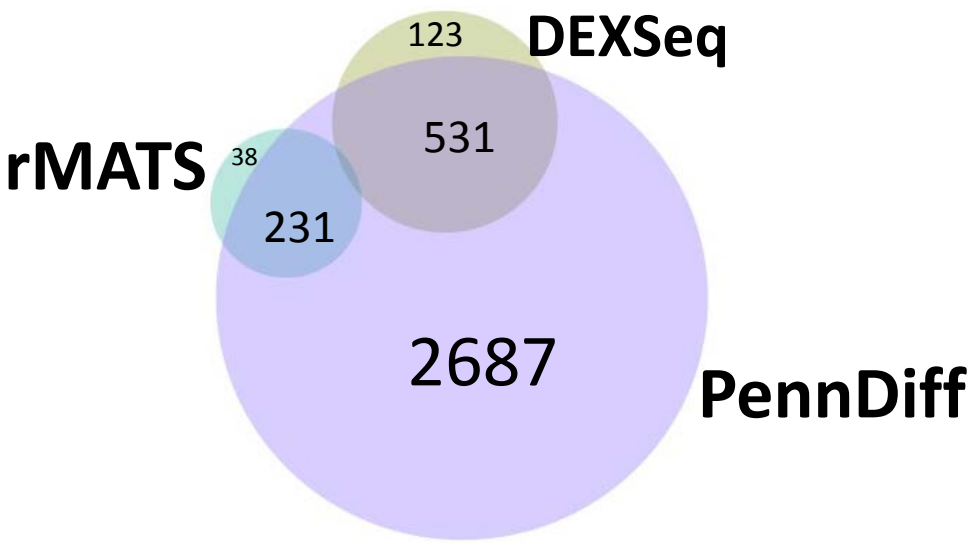
  - Joint model: $\Phi_M\{\Phi^{-1}[F(X_{i,1})], \dots, \Phi^{-1}[F(X_{i,M})] \mid \Gamma\}$
- Exon-based test: $H_0: \beta_m = 0$ for exon group m
- Gene-based test: $H_0: \beta_1 = \cdots = \beta_M = 0$ for all m exon groups

Advantage of PennDiff

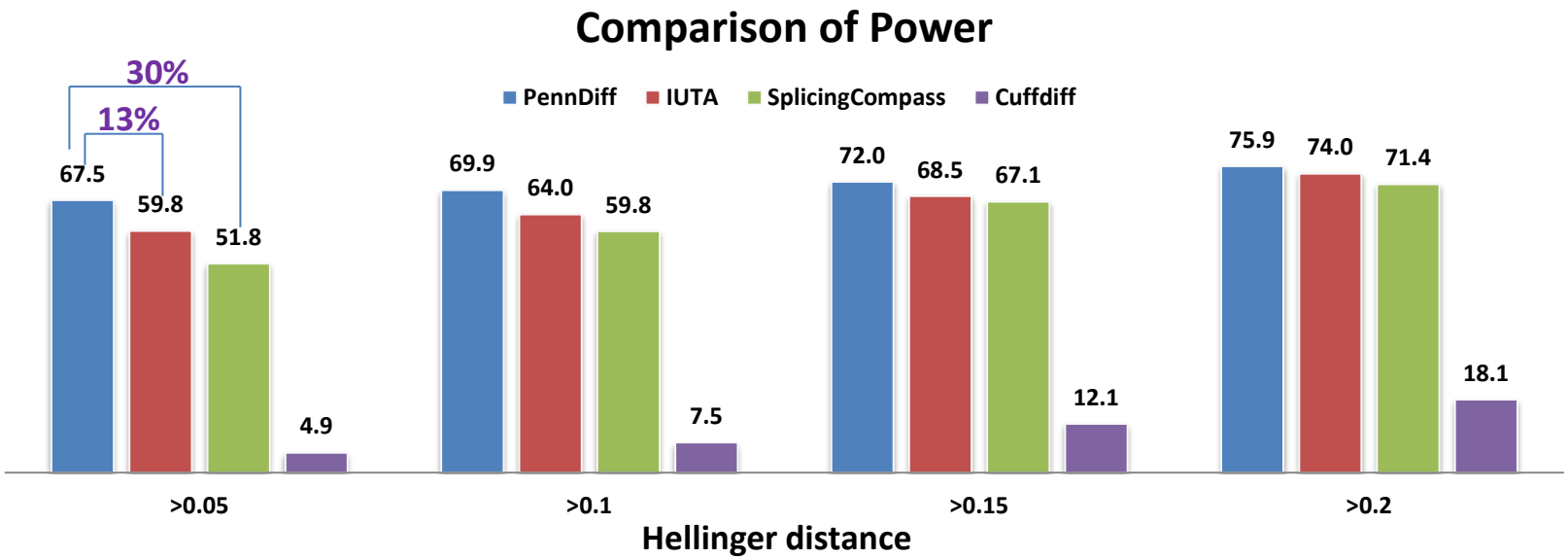
- Grouping exons avoids multiple testing for “exons” originated from the same isoform
- Utilize all available sequencing reads in exon-inclusion level estimation; this is in sharp contrast to DEXSeq, rMATS that only use exon+junction reads
- Collapsing isoforms sharing the same alternatively spliced exons reduces the impact of isoform expression estimation uncertainty and yields more accurate estimate of exon-inclusion level

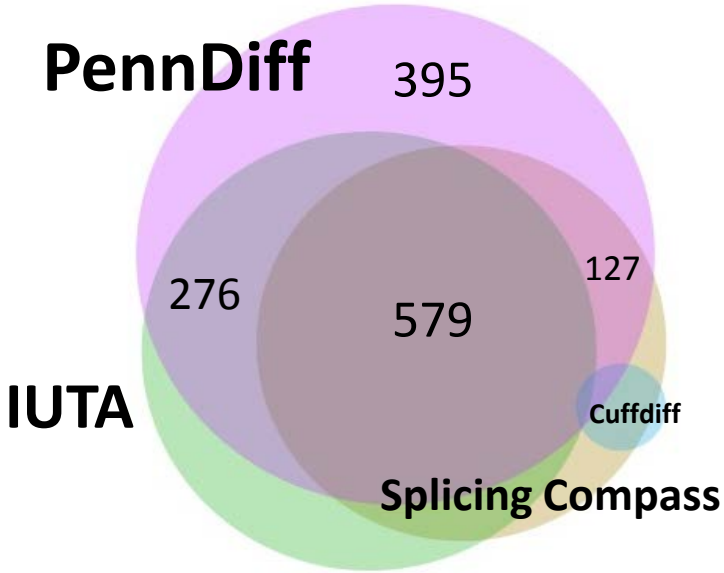
Performance of Exon-based Tests



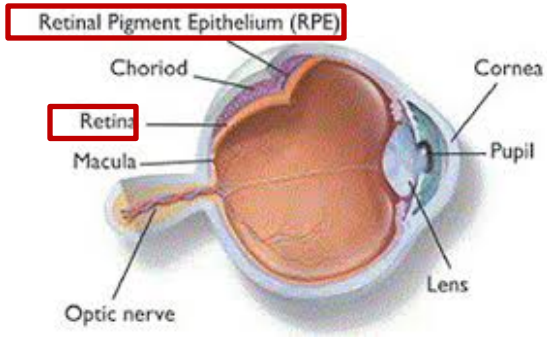


Performance of Gene-based Tests





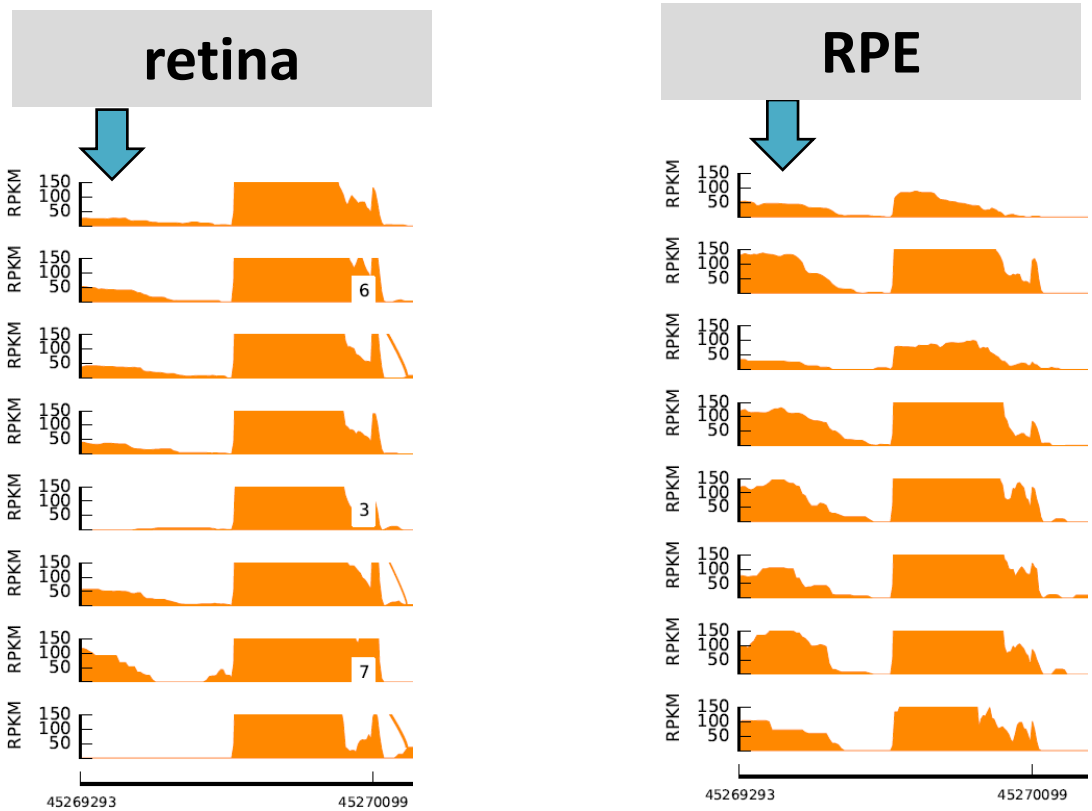
Application to Human Eyes



- RNA-Seq: 8 post-mortem human eyes
- DAS analysis between retina and retinal pigment epithelium (RPE) using PennDiff, Cuffdiff, DEXSeq, and rMATS

NELL2

DAS detected by Penndiff, but missed by other methods



Summary

- RNA-Seq is a powerful tool for studying transcriptomic variations
- Major challenge: reads are much shorter than transcripts from which they are derived from
- Proper RNA-Seq data analysis needs to consider
  - Hidden information on isoform origin
  - Sequencing bias
  - Expression estimation uncertainty
  - Biological variation

Acknowledgements

Yu Hu

Cheng Jia

Dwight Stambolian

Research supported by NIH R01GM108600