Transcript isoform expression and differential expression estimation with BitSeq

Peter Glaus[1], **Antti Honkela**[2], Magnus Rattray[1]

[1] Faculty of Life Science, University of Manchester, UK
[2] Helsinki Institute for Information Technology HIIT,
Department of Computer Science, University of Helsinki, Finland

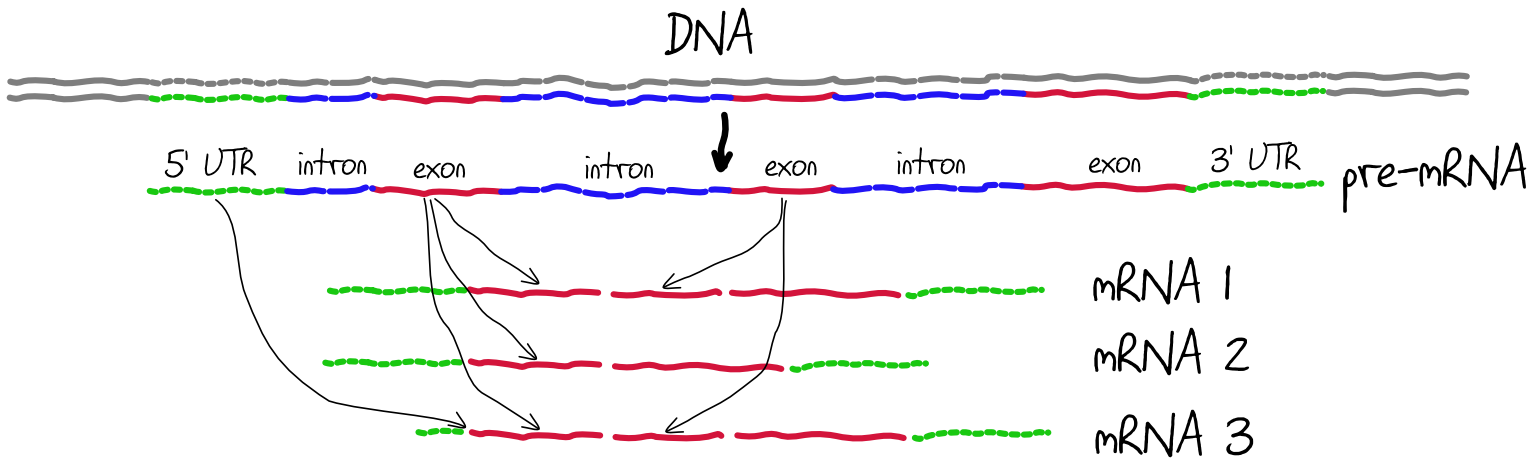
8 January 2014

RNA-seq:

- High-Throughput Sequencing of cDNA

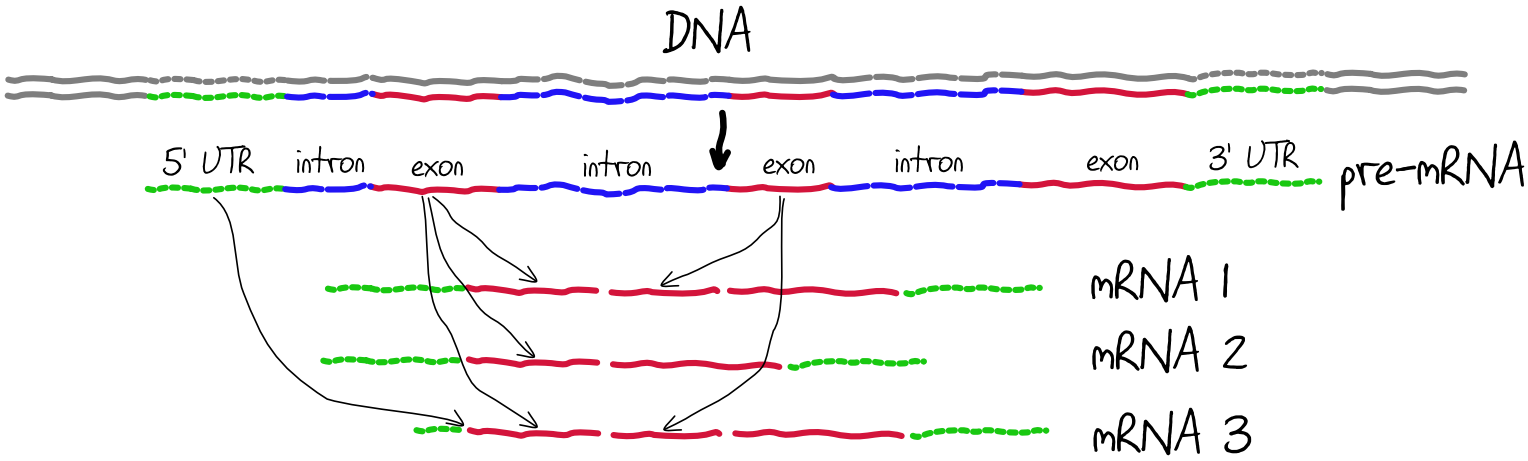
RNA-seq:

- High-Throughput Sequencing of cDNA



RNA-seq:

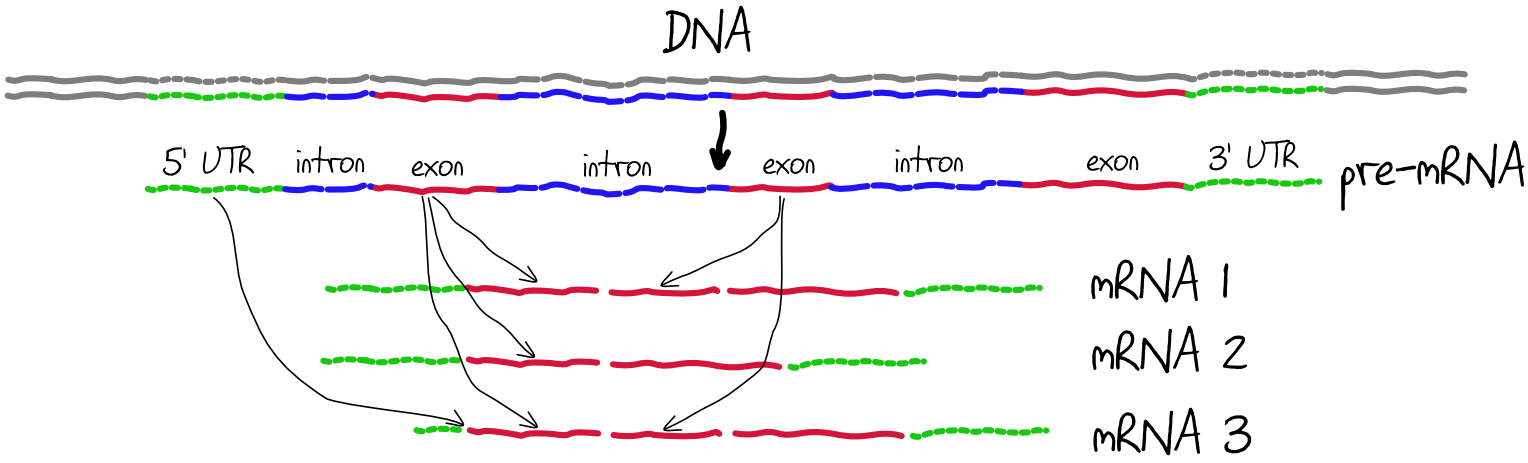
- High-Throughput Sequencing of cDNA



- mapped read count $\approx$ abundance of fragments
- abundance of fragments $\approx$ (gene expression) x (length)

RNA-seq:

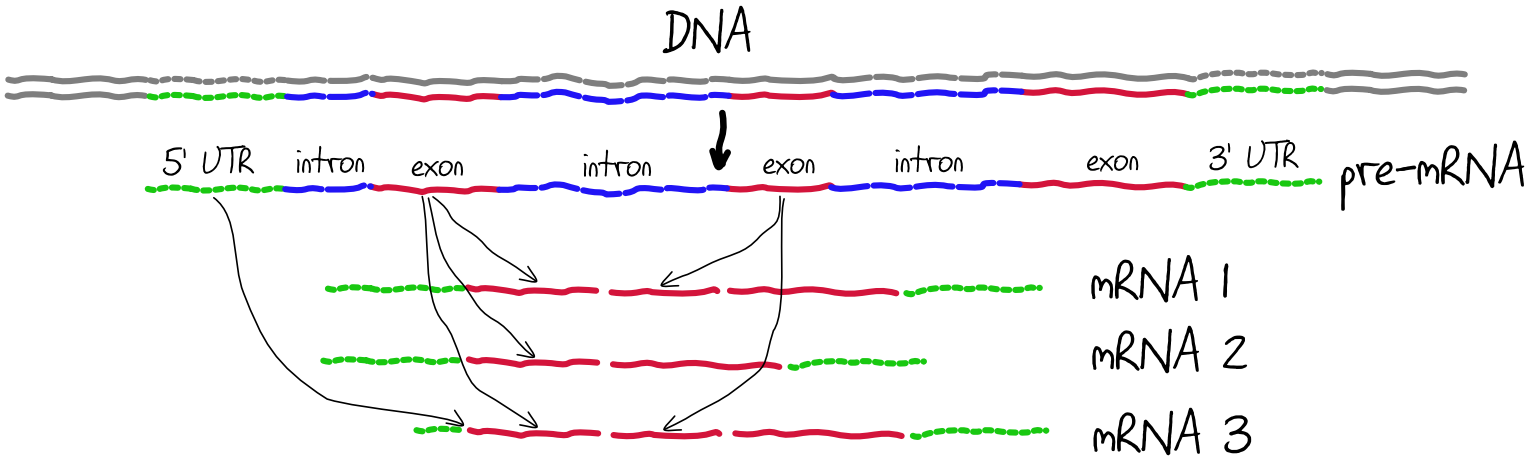
- High-Throughput Sequencing of cDNA



- mapped read count $\approx$ abundance of fragments
- abundance of fragments $\approx$ (gene expression) x (length)
- but which length? which transcripts?

RNA-seq:

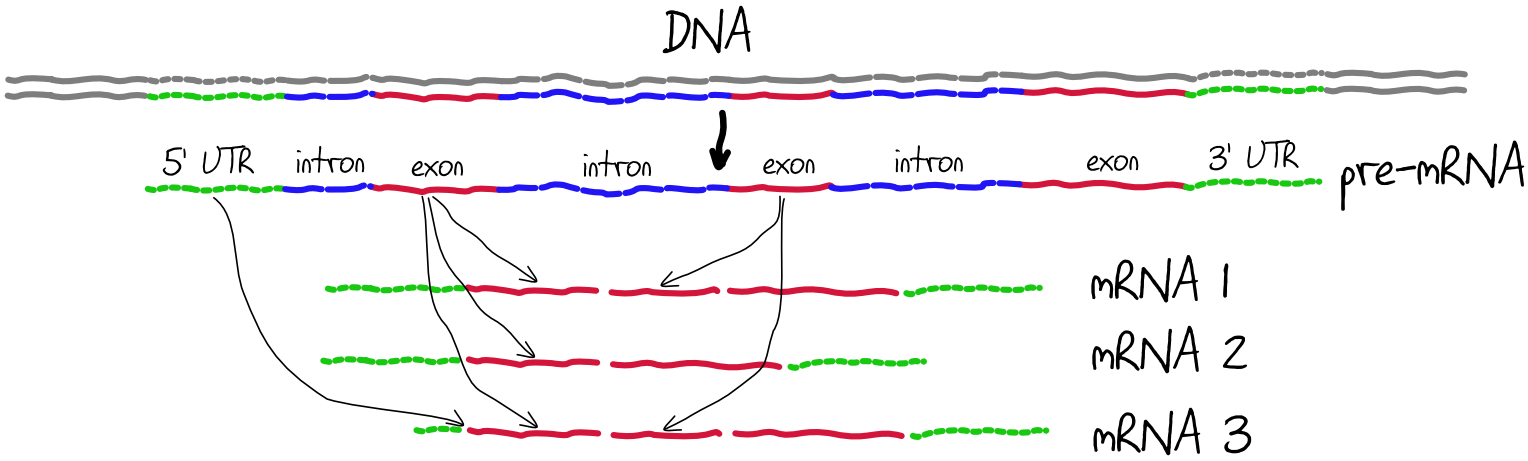
- High-Throughput Sequencing of cDNA



- mapped read count $\approx$ abundance of fragments
- abundance of fragments $\approx$ (gene expression) x (length)
- but which length? which transcripts?
- other difficulties: mismatches, varying quality of reads, non-uniform read distribution

RNA-seq:

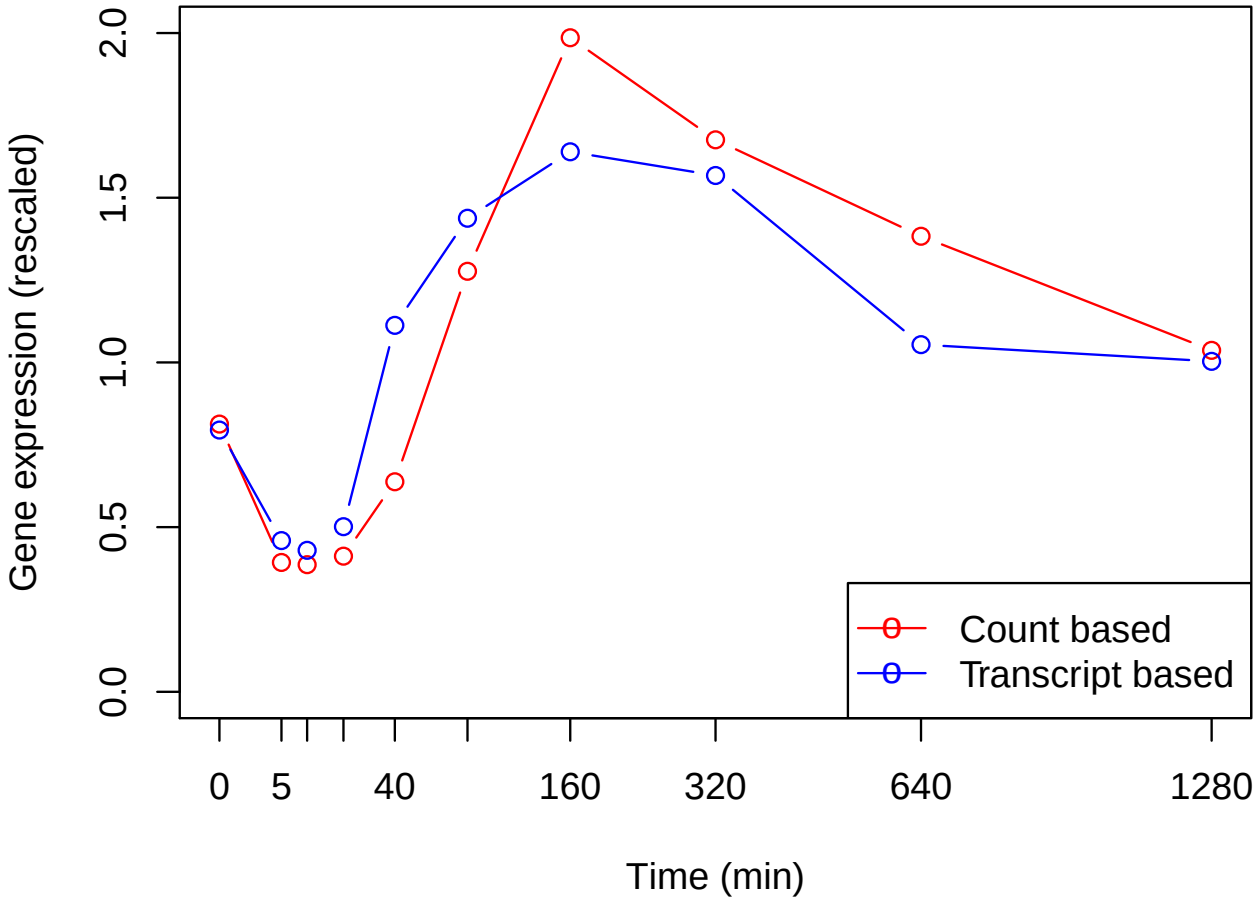
- High-Throughput Sequencing of cDNA



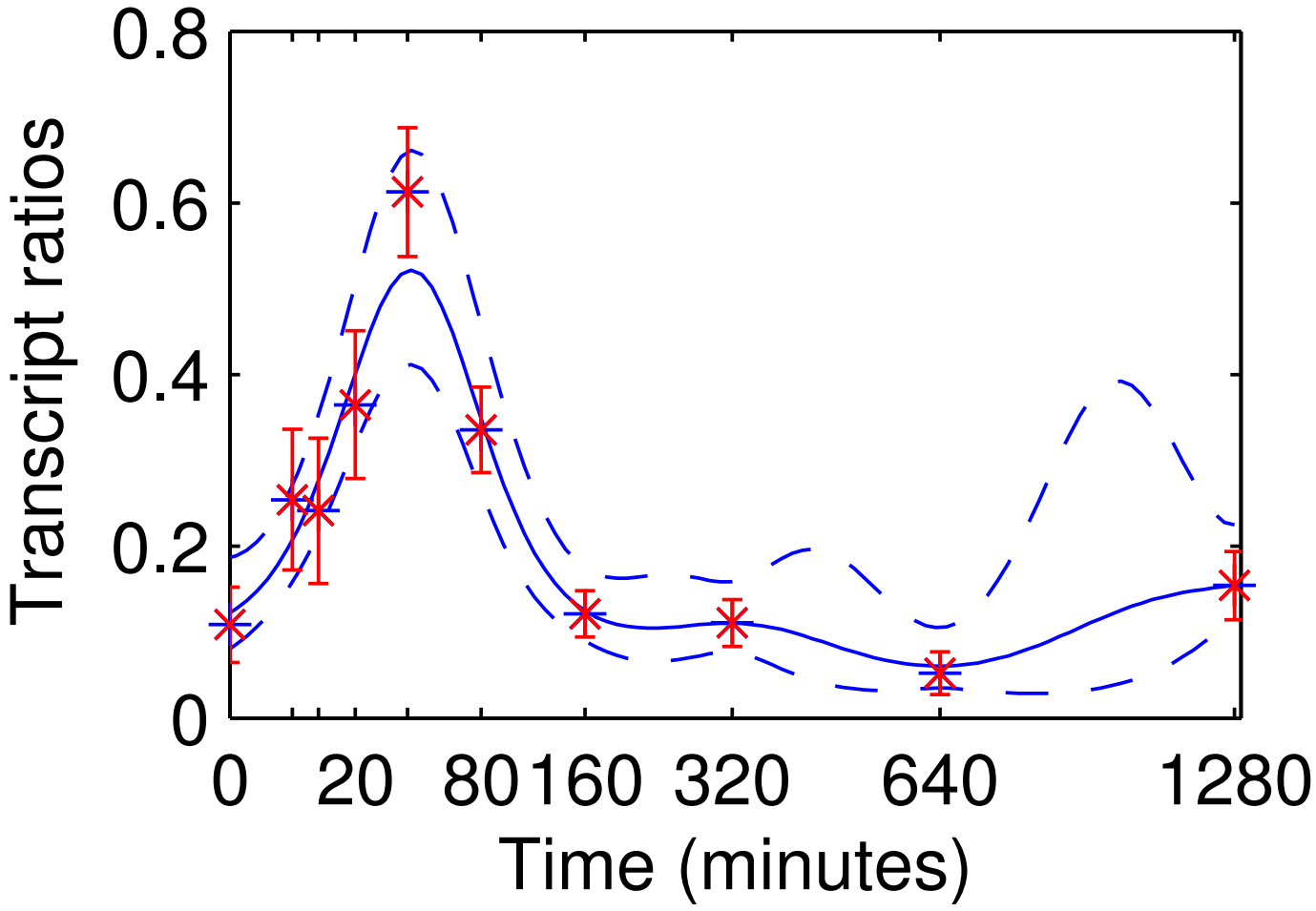
- mapped read count $\approx$ abundance of fragments
- abundance of fragments $\approx$ (gene expression) x (length)
- but which length? which transcripts?
- other difficulties: mismatches, varying quality of reads, non-uniform read distribution
- **our starting point:** reads aligned to transcriptome allowing for multiple matches (using e.g. Bowtie)

Transcripts are expressed, not genes:

gene expression $\approx$ sum over transcript expression



Transcripts are expressed, not genes:



BitSeq:

Goals

- Estimate expression of **transcripts** from RNA-seq data
- Find **differentially expressed** transcripts in multiple conditions while accounting for **biological variation**

BitSeq:

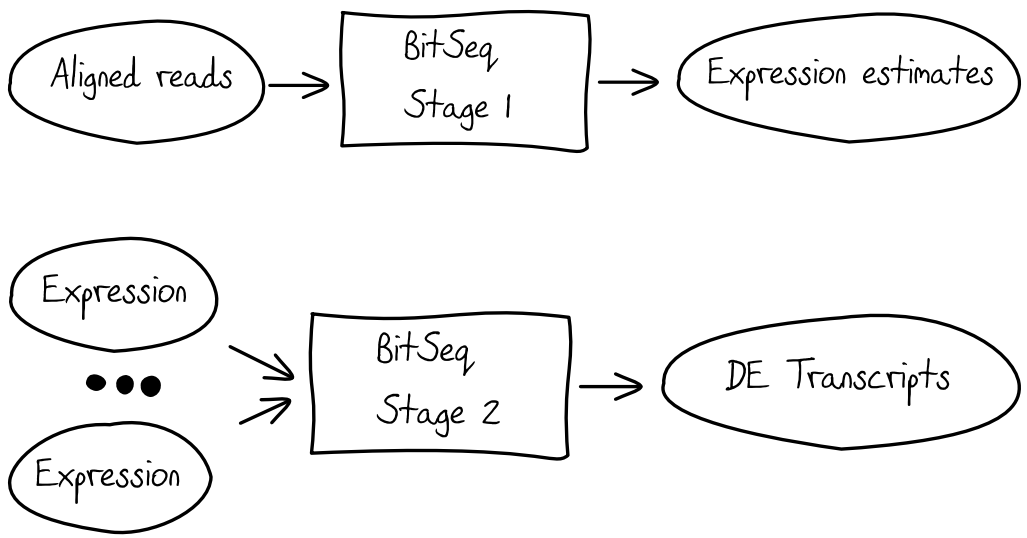
Goals

- Estimate expression of **transcripts** from RNA-seq data
- Find **differentially expressed** transcripts in multiple conditions while accounting for **biological variation**

BitSeq:

Goals

- Estimate expression of **transcripts** from RNA-seq data
- Find **differentially expressed** transcripts in multiple conditions while accounting for **biological variation**

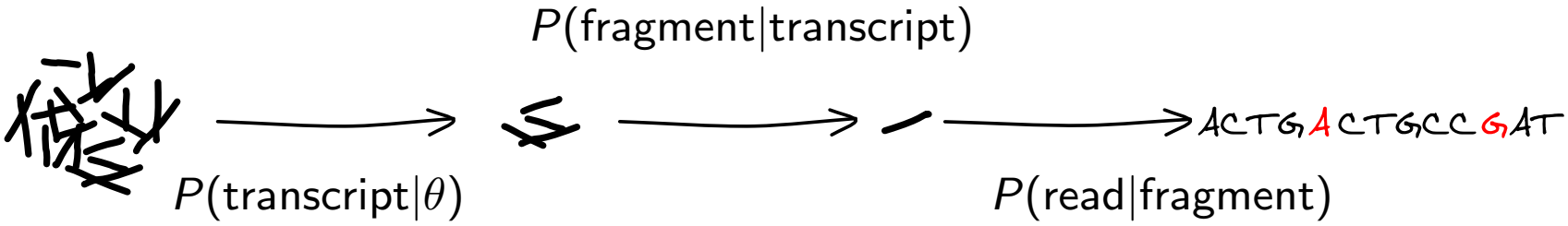


RNA-seq generative model approach:

- Unknown relative expression of transcripts' fragments θ

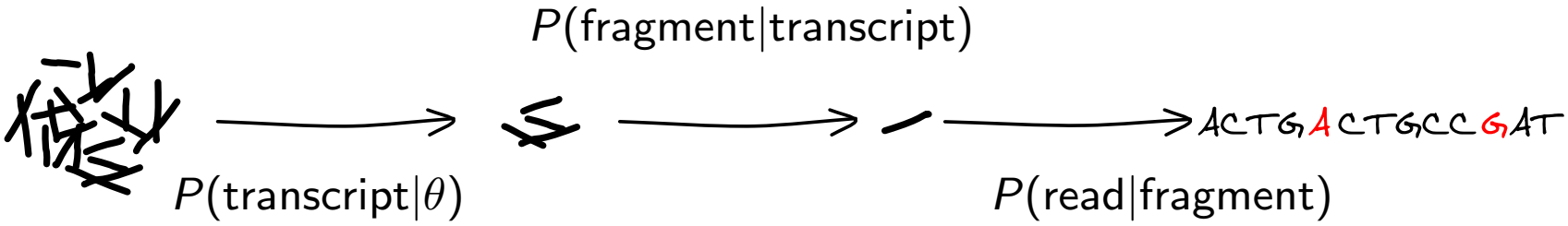
RNA-seq generative model approach:

- Unknown relative expression of transcripts' fragments θ



RNA-seq generative model approach:

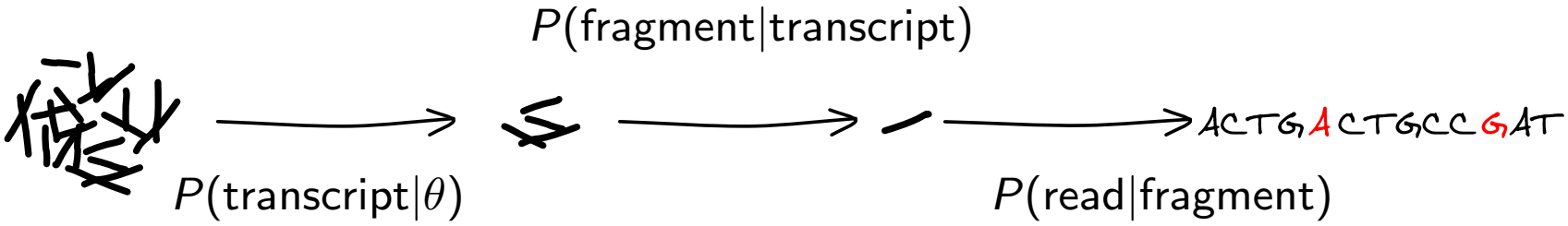
- Unknown relative expression of transcripts' fragments θ



$$P(\text{read}|\theta) = P(\text{transcript}|\theta)P(\text{fragment}|\text{transcript})P(\text{read}|\text{fragment})$$

RNA-seq generative model approach:

- Unknown relative expression of transcripts' fragments θ

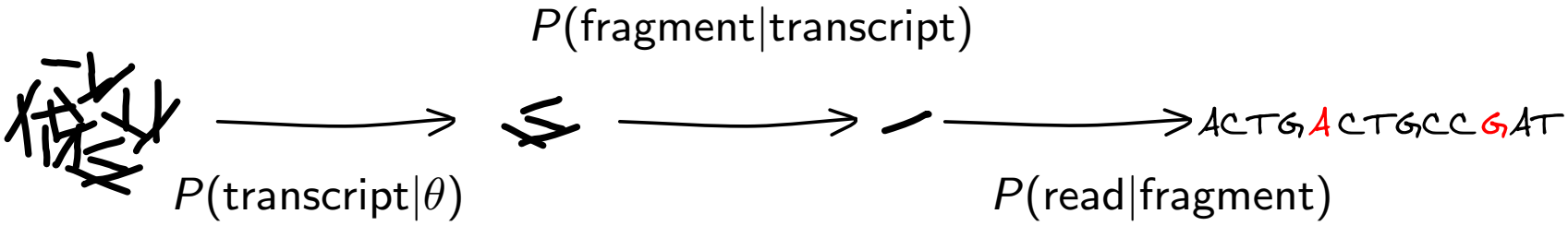


$$P(\text{read}|\theta) = P(\text{transcript}|\theta)P(\text{fragment}|\text{transcript})P(\text{read}|\text{fragment})$$

$$P(\text{Data}|\theta) = \prod_{i=1}^{n} P(\text{read}_i|\theta); \quad P(\theta|\text{Data}) = \frac{P(\text{Data}|\theta)P(\theta)}{P(\text{Data})}$$

RNA-seq generative model approach:

- Unknown relative expression of transcripts' fragments θ



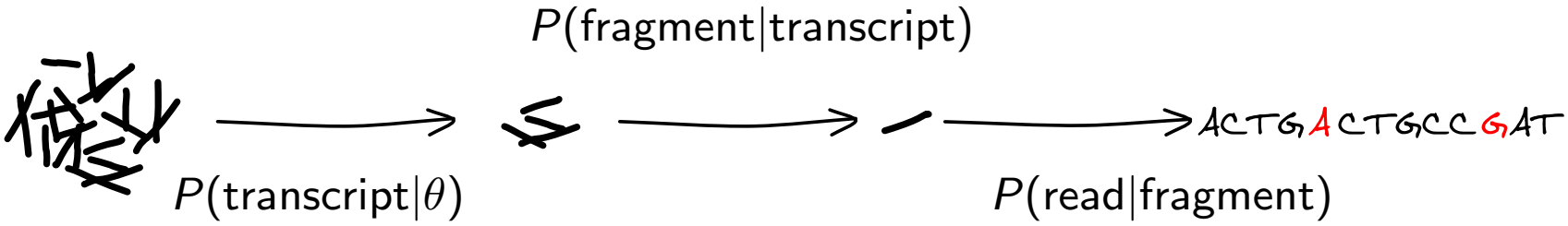
$$P(\text{read}|\theta) = P(\text{transcript}|\theta)P(\text{fragment}|\text{transcript})P(\text{read}|\text{fragment})$$

$$P(\text{Data}|\theta) = \prod_{i=1}^{n} P(\text{read}_i|\theta); \quad P(\theta|\text{Data}) = \frac{P(\text{Data}|\theta)P(\theta)}{P(\text{Data})}$$

- Bayesian inference: we use Markov Chain Monte Carlo (MCMC) algorithm to produce samples from $P(\theta|\text{Data})$

RNA-seq generative model approach:

- Unknown relative expression of transcripts' fragments θ



$$P(\text{read}|\theta) = P(\text{transcript}|\theta)P(\text{fragment}|\text{transcript})P(\text{read}|\text{fragment})$$

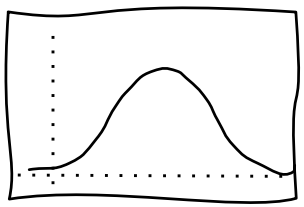
$$P(\text{Data}|\theta) = \prod_{i=1}^{n} P(\text{read}_i|\theta); \quad P(\theta|\text{Data}) = \frac{P(\text{Data}|\theta)P(\theta)}{P(\text{Data})}$$

- Bayesian inference: we use Markov Chain Monte Carlo (MCMC) algorithm to produce samples from $P(\theta|\text{Data})$
- RPKM expression units $\propto \theta/\text{transcript length}$

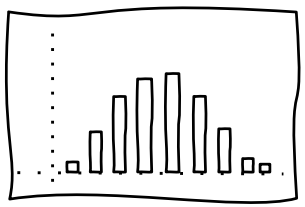
Bayesian Inference

- Represent unknowns in form of probability distribution (instead of value + confidence interval)

$\theta \sim N(\mu|\sigma)$



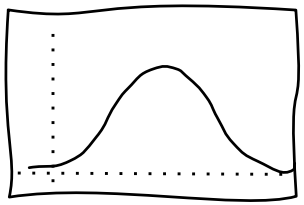
$\theta \sim S = \{s_1, \ldots, s_n\}$



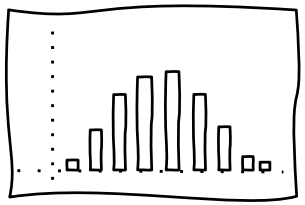
Bayesian Inference

- Represent unknowns in form of probability distribution (instead of value + confidence interval)

$\theta \sim N(\mu|\sigma)$



$\theta \sim S = \{s_1, \ldots, s_n\}$

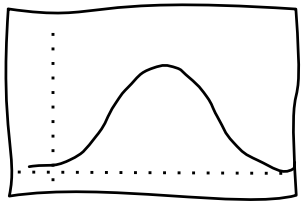


- We use probability theory (Bayes Theorem) to manipulate these distributions

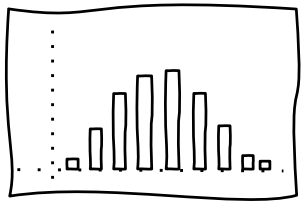
Bayesian Inference

- Represent unknowns in form of probability distribution (instead of value + confidence interval)

$\theta \sim N(\mu|\sigma)$



$\theta \sim S = \{s_1, \ldots, s_n\}$

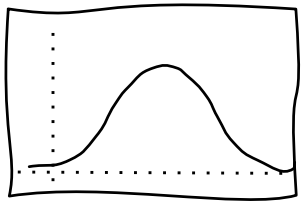


- We use probability theory (Bayes Theorem) to manipulate these distributions
- MCMC is a numerical method to generate samples from a distribution of interest

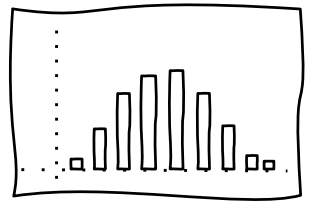
Bayesian Inference

- Represent unknowns in form of probability distribution (instead of value + confidence interval)

$\theta \sim N(\mu|\sigma)$



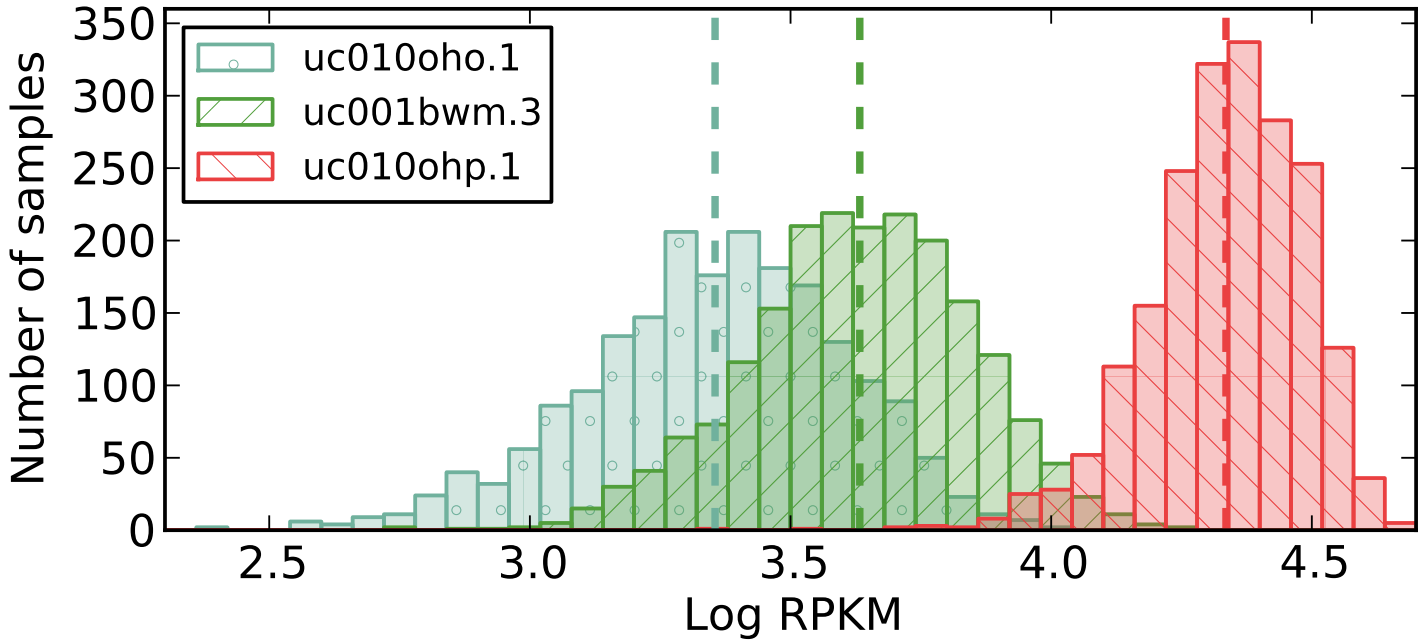
$\theta \sim S = \{s_1, \ldots, s_n\}$



- We use probability theory (Bayes Theorem) to manipulate these distributions
- MCMC is a numerical method to generate samples from a distribution of interest
- Results can be summarized by mean and standard deviation:

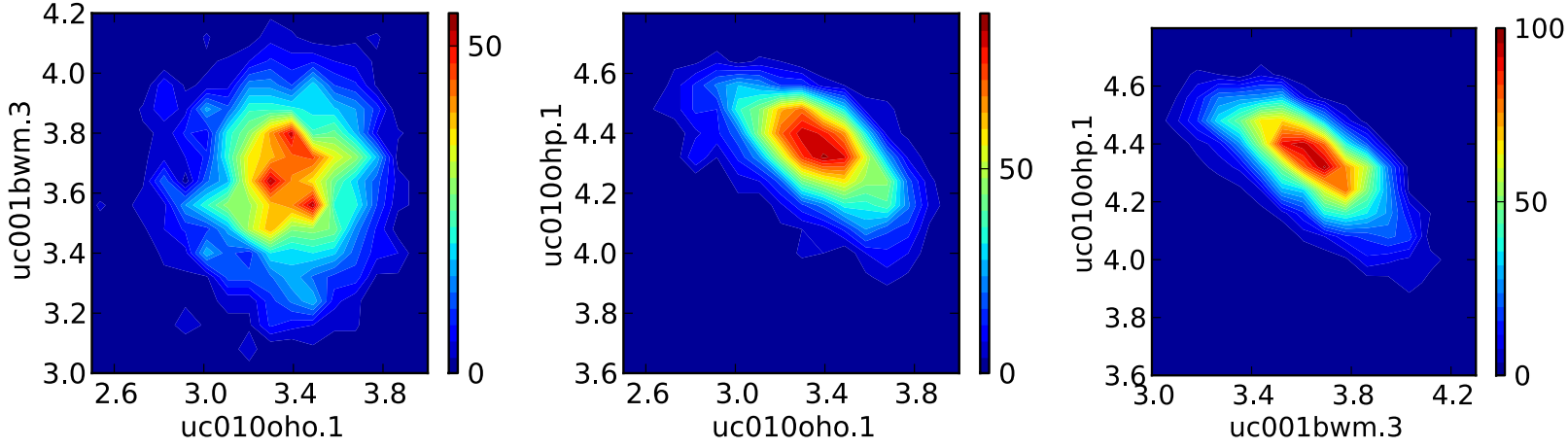
$$\mathrm{E}[\theta] = \mathrm{mean}(S); \sigma_\theta = \mathrm{stdev}(S)$$

Results:



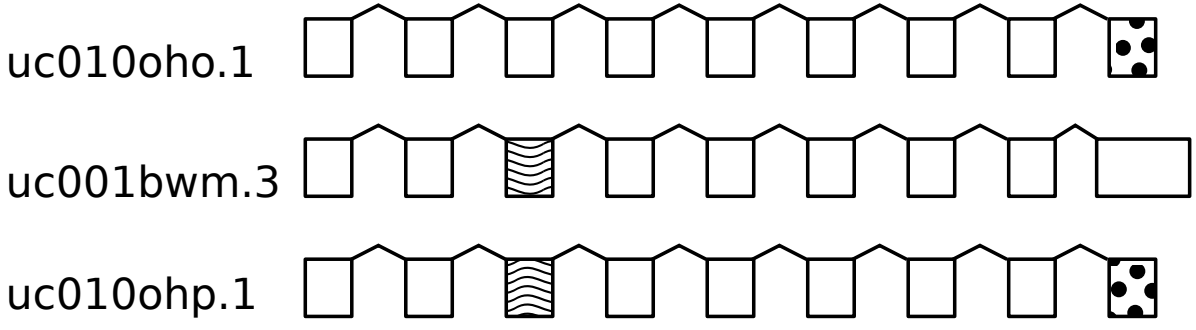
Histograms of expression MCMC samples of three transcripts of one gene.

Anticorrelation:



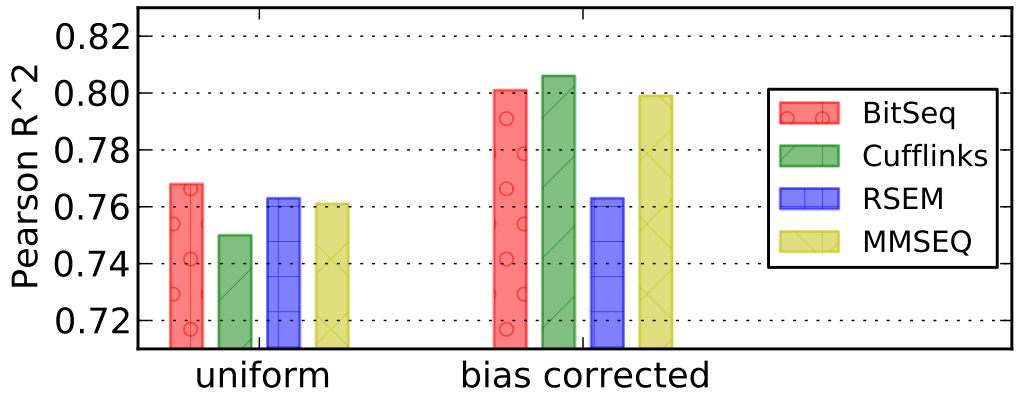
(a) (b) (c)

Density plots of expression MCMC samples of transcript pairs plotted against each other. (expression in log RPKM)



Exon model.

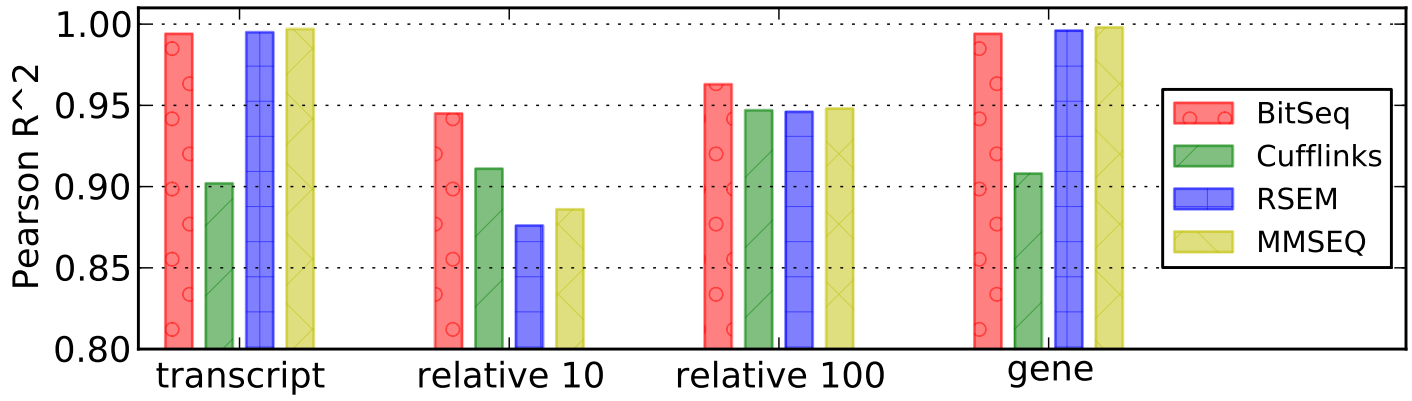
Accuracy, real data:



Comparison of expression estimation accuracy against TaqMan qRT-PCR using Pearson R^2 (893 transcripts, MAQC II)

- using uniform read distribution model and bias correction
- other methods:
  - RSEM: similar model, using Maximum Likelihood
  - MMSEQ: count based model, using Maximum Likelihood and Gibbs Sampling

Accuracy, synthetic data:



Comparison of expression estimation accuracy against ground truth using Pearson R^2 on synthetic RNA-seq data

- Transcript expression
  - (transcripts with at least 1 read)
- Relative within-gene proportion of transcripts
  - (transcripts of genes with at least 10 / 100 reads)
- Gene expression
  - (genes with at least 1 read)

Differential expression:

- For transcript m we want to know the probability of the expression in two experiments being different
- Compare the distributions represented by MCMC samples
- Probability of Positive Log Ratio — one sided Bayesian test

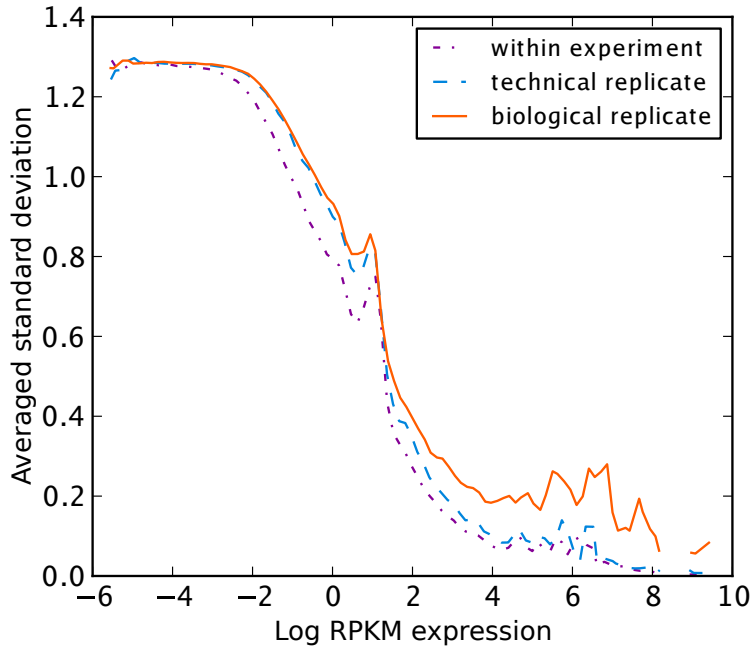
$$\mathrm{PPLR}_m = P\left(\log \frac{\theta_m^{(1)}}{\theta_m^{(2)}} > 0\right) = P(\log \theta_m^{(1)} > \log \theta_m^{(2)})$$

$$\approx \frac{1}{S} \sum_{s=1}^{S} \delta(\log \theta_m^{(1)(s)} > \log \theta_m^{(2)(s)})$$

- PPLR close to 1/0 indicates confident up/down regulation

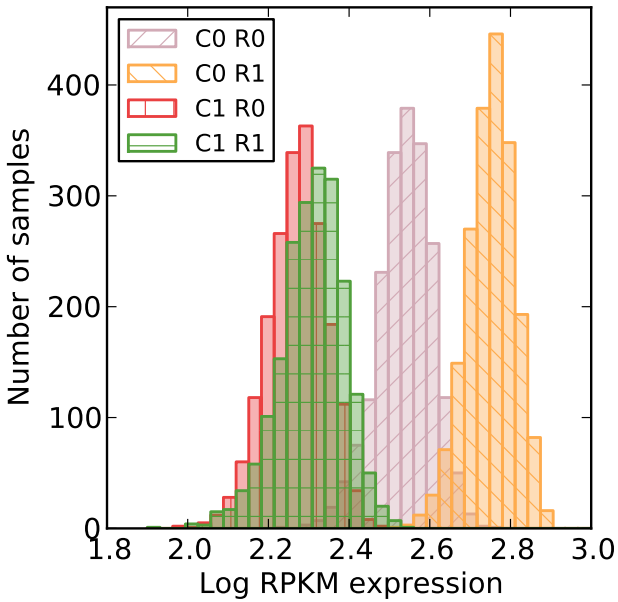
Biological variation:

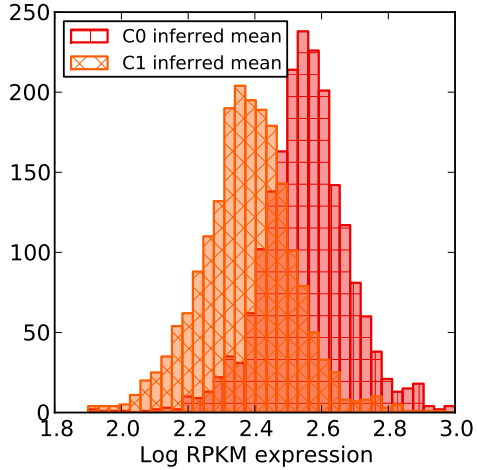
- dataset from Short Read Archive (Xu et al. 2010)
- 2 Conditions $\times$ 2 Biological replicates $\times$ Technical replicates

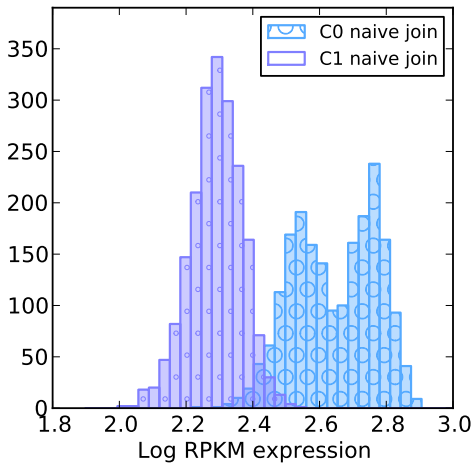


Averaged standard deviation of logged RPKM expression samples of: one MCMC run, combined MCMC samples from technical replication, combined MCMC from biological replication

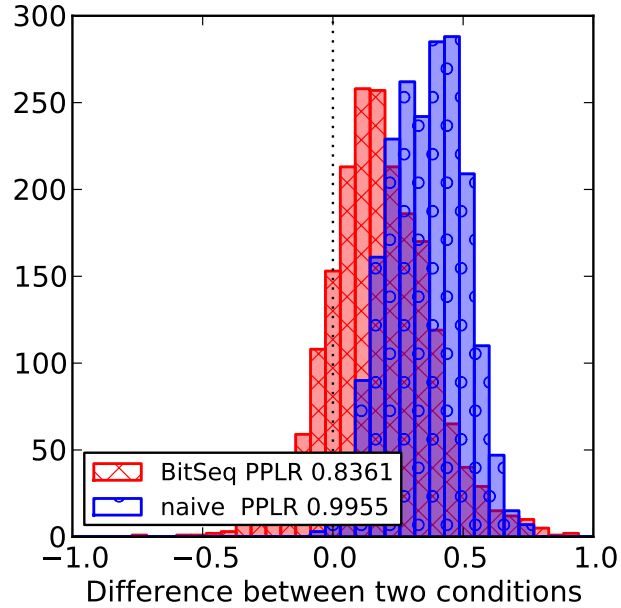
Single transcript DE analysis with biological replicates:







- initial distributions

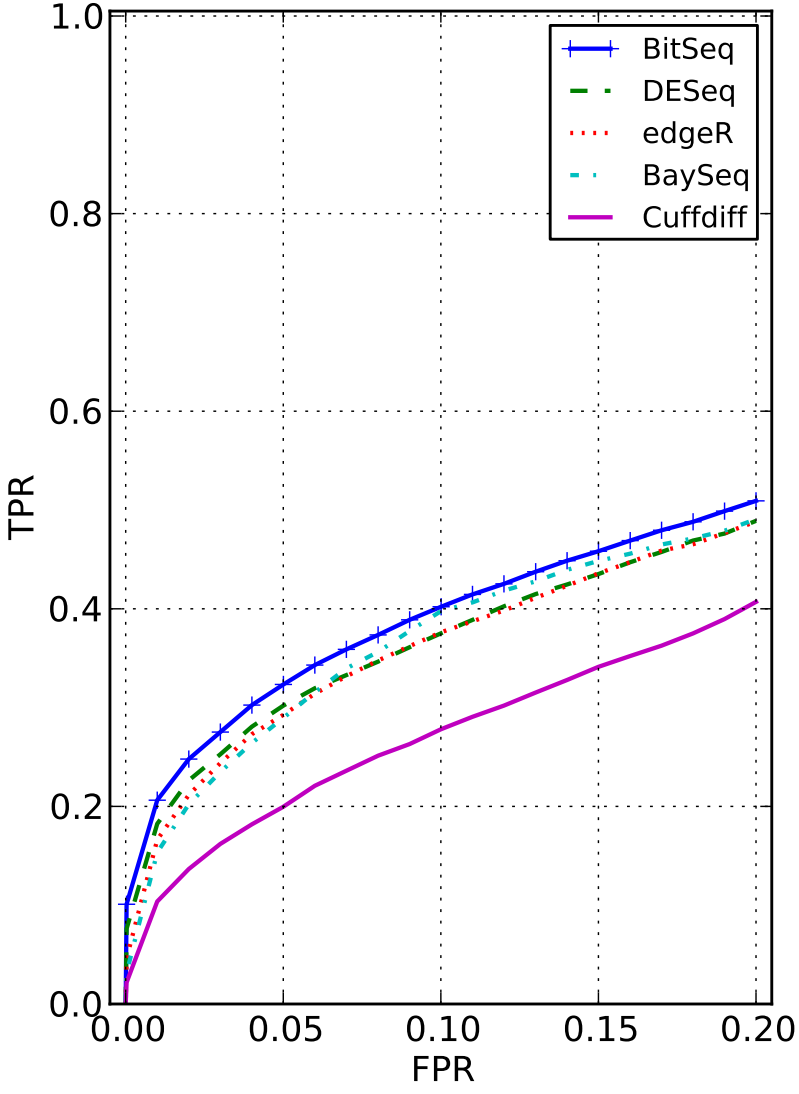


- resulting distribution of differences

- Estimating condition mean expression and differential expression in comparison with naive method of merging samples
- Simple merge of two replicates results in bimodal distribution, but PPLR 0.9955
- Our approach produces PPLR 0.8361

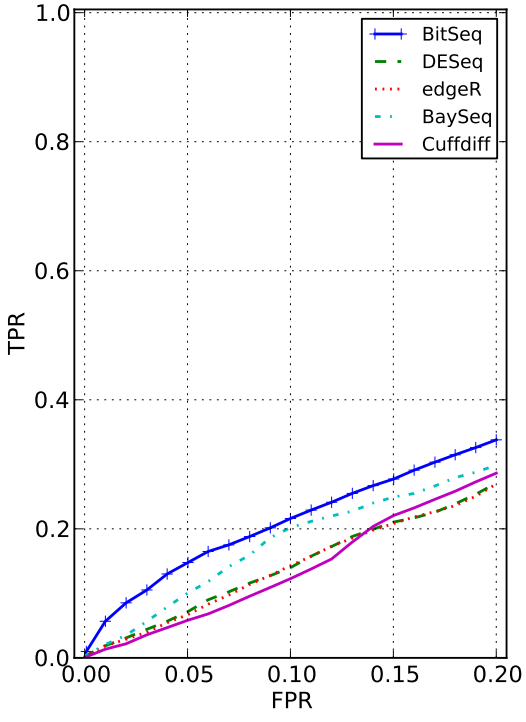
▸ DE model

Differential expression detection accuracy:

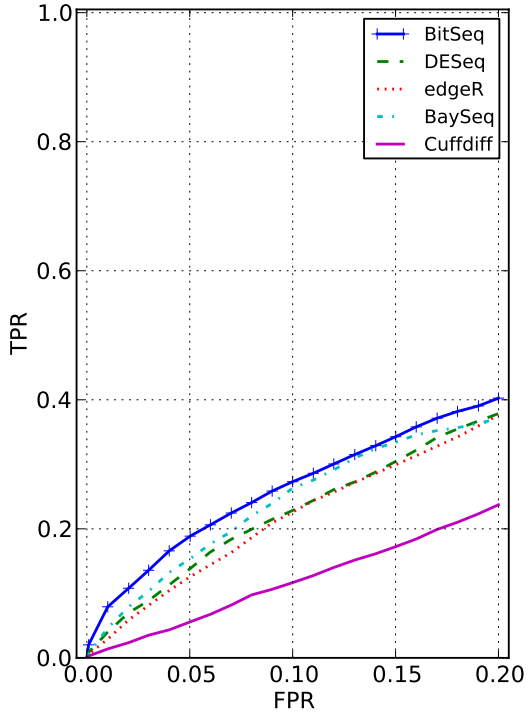


- Simulated dataset with differentially expressed transcripts
- All simulation parameters from real data
- 1/3 of transcripts differentially expressed (both up and down)
- Fold changes uniformly distributed between 1.5 and 3.5
- DESeq, edgeR, BaySeq were supplied with expression estimates from BitSeq

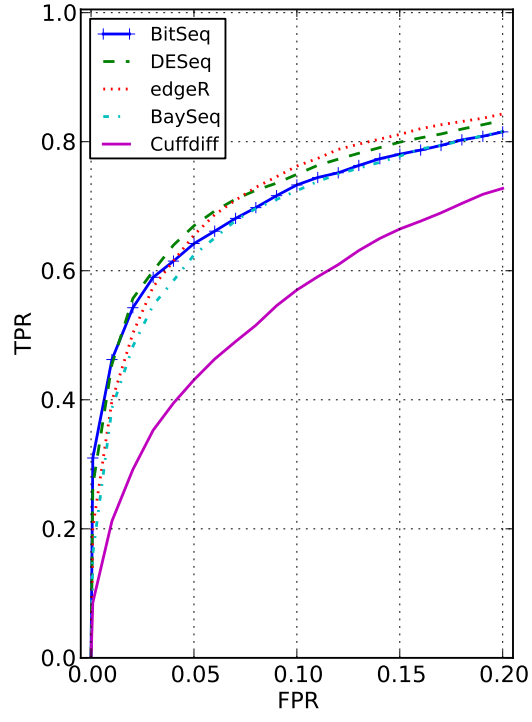
Differential expression detection accuracy split by expression level:



$1 \leq c < 3$



$3 \leq c < 19$



$19 \leq c$

Conclusion:

- Method for transcript-level expression estimation and differential expression calling
- Principled handling of:
  - read qualities, non-uniform read distribution, reads with multiple alignments, paired-end reads
- Using Bayesian methods to propagate uncertainties from read-level to DE estimates
- Accurate within-gene relative expression of transcripts
- Accounts for biological variation in differential expression

- Current/recent work:
  - Faster inference of expression values (available in BitSeq 0.7.0)

Resources:

- Papers:
  - Glaus P., Honkela A., and Rattray M. (2012) "Identifying differentially expressed transcripts from RNA-seq data with biological variation" *Bioinformatics*, **28**(13), 1721–1728.
  - Hensman J., Glaus P., Honkela A., Rattray M. (2013) "Fast approximate inference of transcript expression levels from RNA-seq data" http://arxiv.org/abs/1308.5953
- Package:
  - Bioconductor 2.10 and newer
  - standalone at `http://code.google.com/p/bitseq/`

BitSeq pipeline

1. Align reads to a reference transcriptome
   (Each transcript sequence in reference, contiguous alignments)
2. Stage 0: Pre-process alignments
   (For each sample separately; `parseAlignment`)
3. Stage 1: Estimate expression
   (For each sample separately; `estimateExpression`)
4. Stage 2: Estimate variances, condition-specific expression and probability of differential expression
   (For all samples together; `getVariance`, `estimateHyperPar`, `estimateDE`)