



清華大學

Tsinghua University

Isoform Abundance Inference Provides a More Accurate Estimation of Gene Expression Levels in RNA-Seq

Xi Wang, Zhengpeng Wu, Xuegong Zhang

*MOE Key Laboratory of Bioinformatics and
Bioinformatics Division, TNLIST /
Department of Automation, Tsinghua University
Beijing 100084, China*

Gene expression

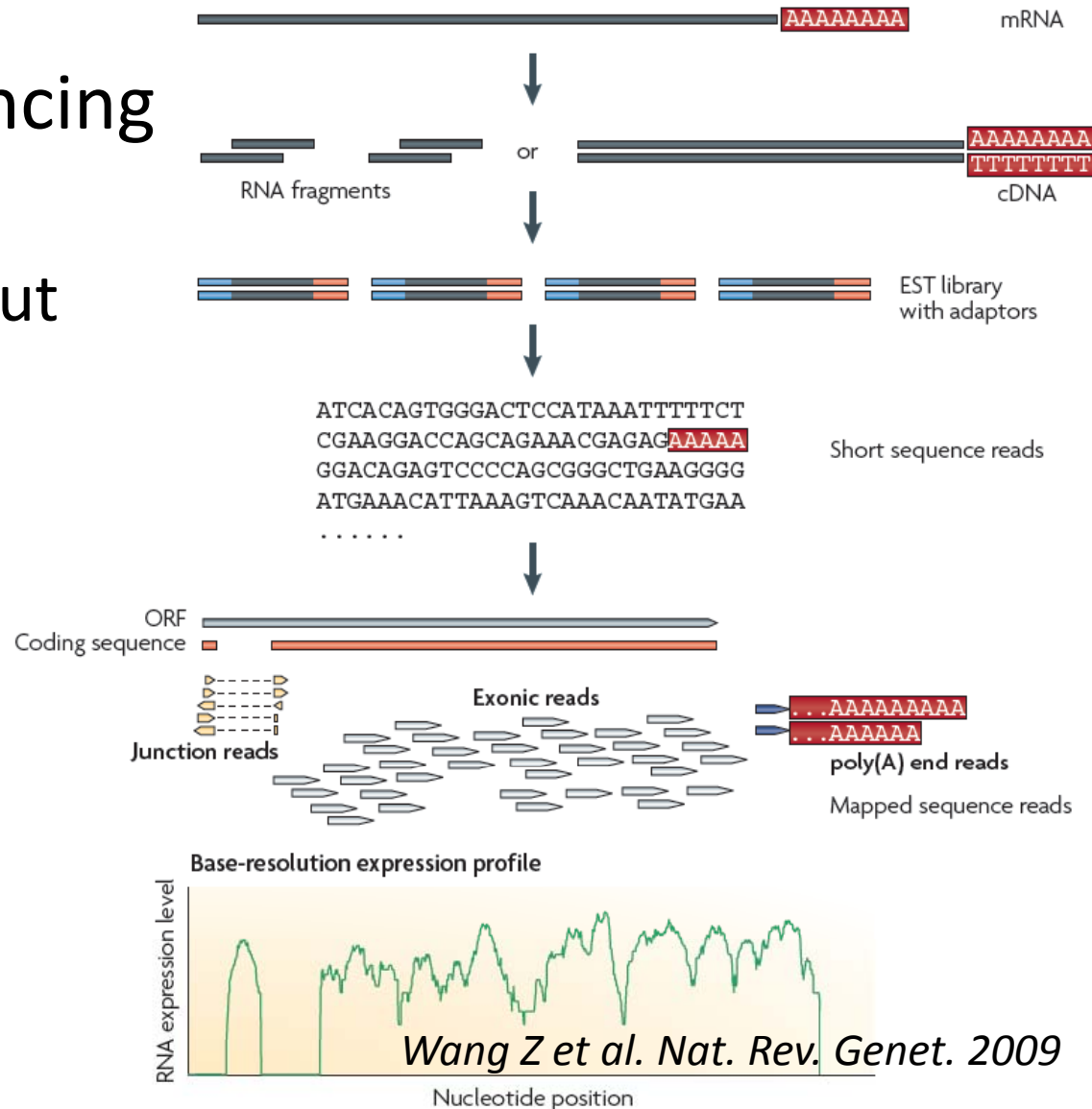
- Importance
 - To understand functional elements of the genome
 - To identify gene expression differences among diseased and healthy tissues
 - To reveal the molecular constituents of cells/tissues
 - To understand development and diseases
 - To study the mechanisms and the evolution of gene regulation
 -

Gene expression

- At the RNA level
 - Mature RNA: transcription + (alternative) splicing
 - Measurement:
 - Northern blotting (1977)
 - Real time PCR (1993)
 - **Microarray** (1995): high-throughput
 - SAGE (1995): digital
 - MPSS (2000): digital
 - **RNA-seq** (2008): high-throughput + digital

RNA-seq

- Next-Gen Sequencing
 - Lower cost
 - Higher throughput
- RNA-seq
 - To quantitatively measure transcriptome
 - To study gene expression at the RNA level



RPKM method

- RPKM: *Reads Per Kilo-base per Million reads*
(Mortazavi A *et al*, *Nat Methods*, 2008)
- Normalized against the region length and the sequencing depth
- Read density (averaged coverage) of a region
 - *How about a region with non-uniform distributed reads?*
 - A generalized question

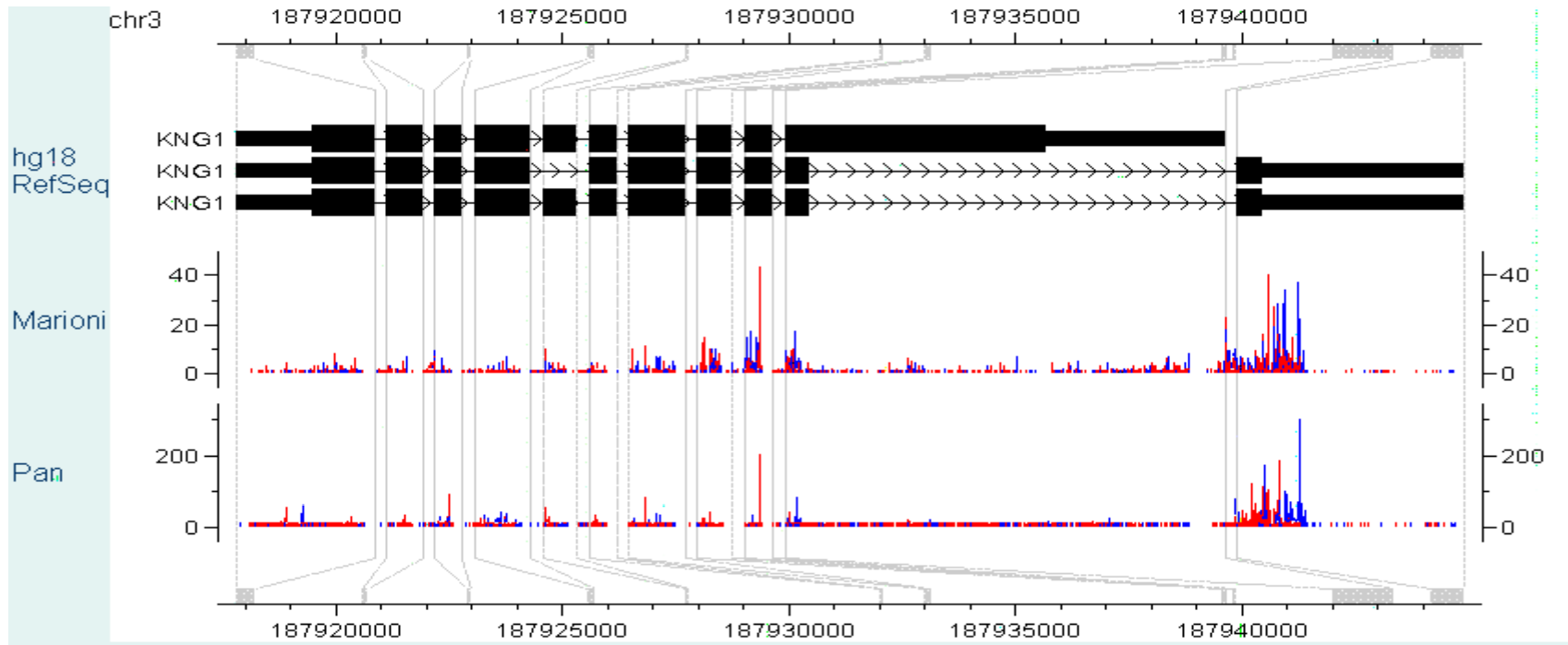
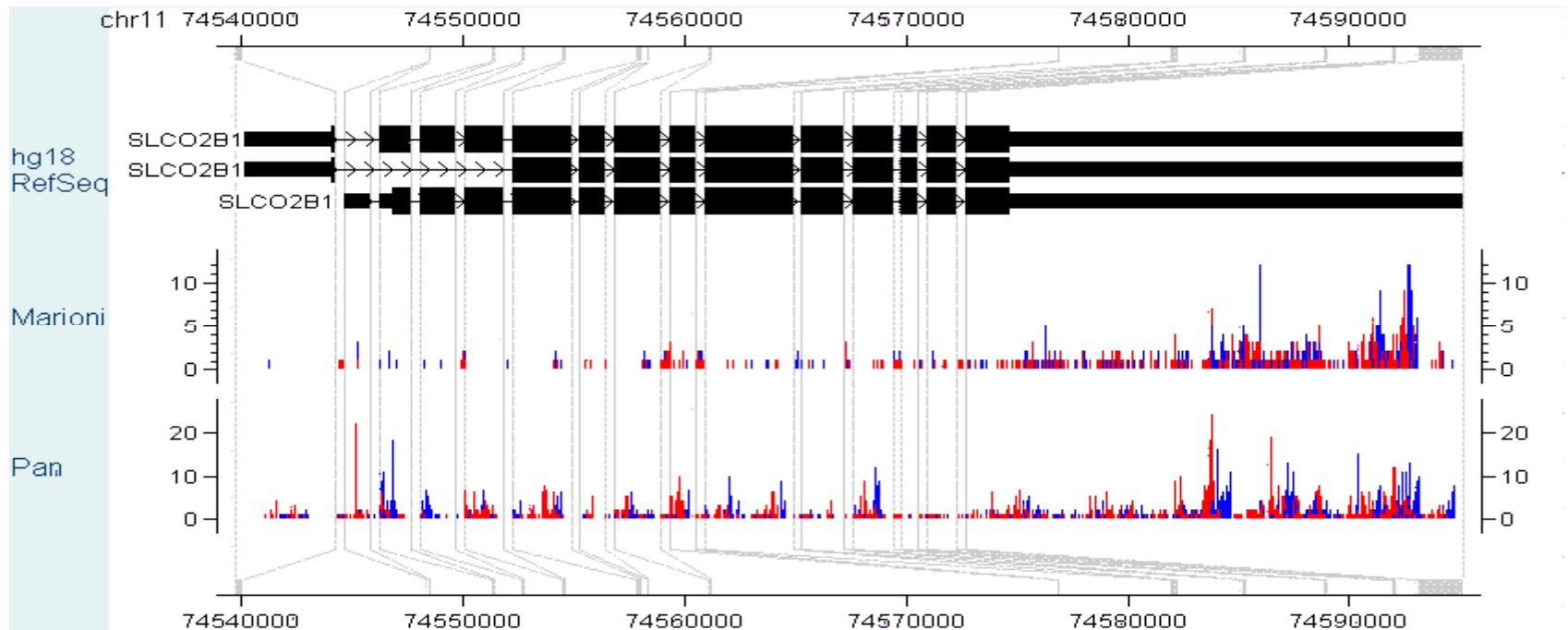
RPKM in multi-isoform genes

- The RPKM computation under-estimates gene expression levels in multi-isoform genes
 - Average among the whole exonic bases
 - Shown in the cufflinks paper
(Trapnell C *et al*, *Nat Biotechnol*, 2010)

Proposition 3. *The totally projective normalization method is correct only for single isoform genes. If a gene has two or more isoforms the expression is underestimated.*

Proof: The effective length of the gene is overestimated, hence the expression level is underestimated. To see this, first note that the length of some transcript in a gene is less than the total number of exonic bases among all transcripts. Then, if a_1, \dots, a_n are real numbers all greater than zero and b_1, \dots, b_n are not all equal, we have

$$(40) \quad \frac{\sum_{i=1}^n a_i b_i}{\sum_{i=1}^n a_i} < \max_i(b_i),$$



The challenge

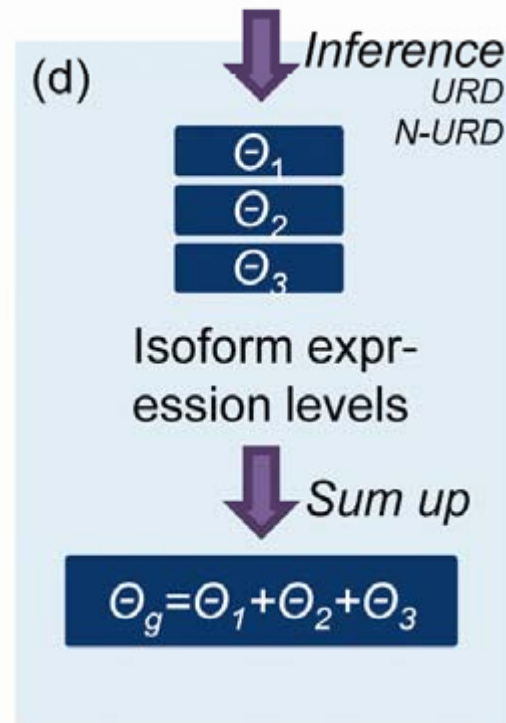
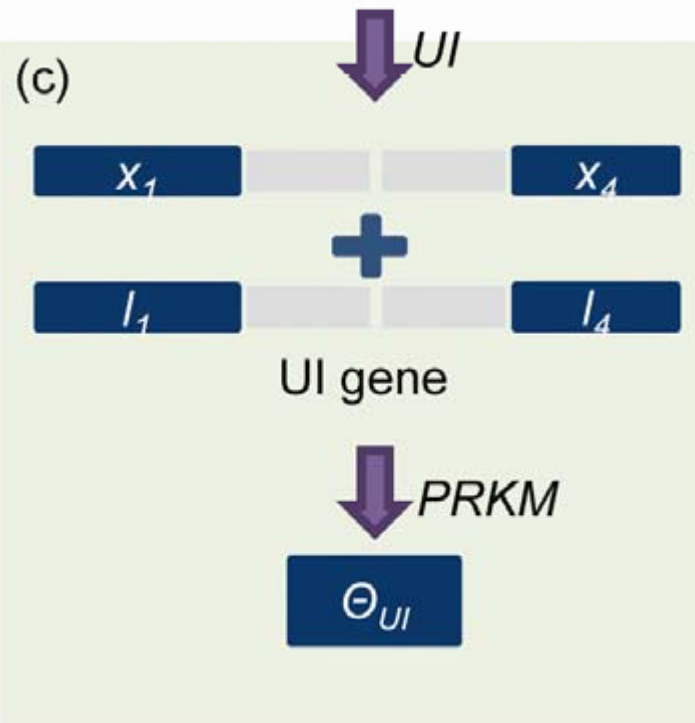
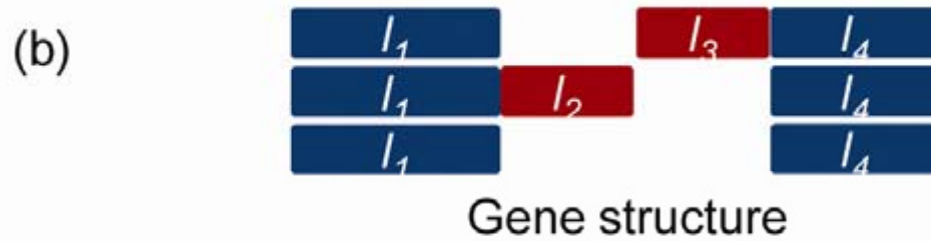
- RPKM computation doesn't always work
- Alternative splicing
 - Cannot treat every exonic base equally
- Sequencing biases
 - Non-uniform read distribution
 - Some systemic biases cannot be averaged out

Comparison with microarray

- Probe-sets for expression array
 - Usually, only fragments at 3' ends of genes
- Unlike exon-arrays and tiling arrays, expression arrays take a gene as a whole
- In computation, not taking into account alternative splicing
(Li & Wong, *Proc. Natl. Acad. Sci*, 2001)
- So, no such issues in expression arrays

Towards the unbiased estimation

- Two strategies
 - UI-based (Bullard JH et al, *BMC bioinformatics*, 2010)
 - Based on the RPKM method with the union-intersection bases in a gene
 - Avoiding the effect of gene structure
 - Isoform level-based
 - Summing over isoform expression levels in a gene
 - Requiring isoform expression inference at first
 - Being able to further consider sequencing biases



Two isoform-based methods: *isoform expression inference*

- The problem

- n exons with length (l_1, l_2, \dots, l_n)
- m isoforms with expression $\Theta = (\theta_1, \theta_2, \dots, \theta_m)$
- Given reads counts $X = (x_1, x_2, \dots, x_n)$
- Estimate expressions $\hat{\Theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_m)$

- URD model (Jiang H *et al*, *Bioinformatics*, 2009)

- Log-likelihood function

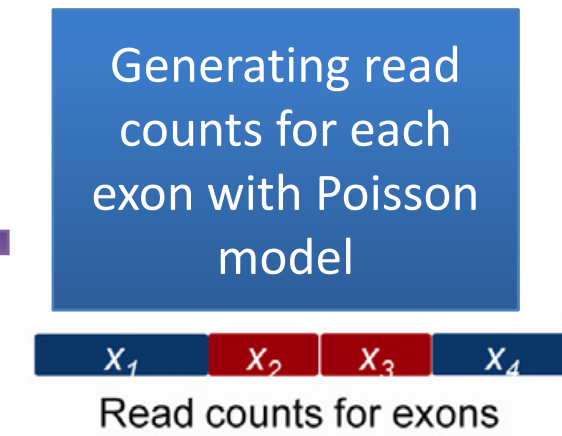
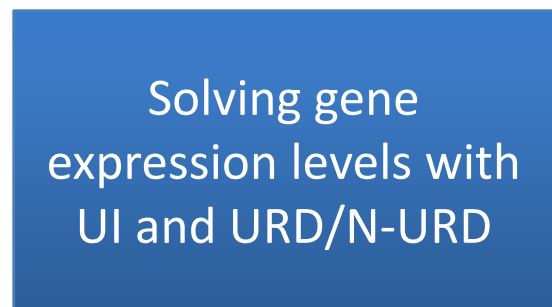
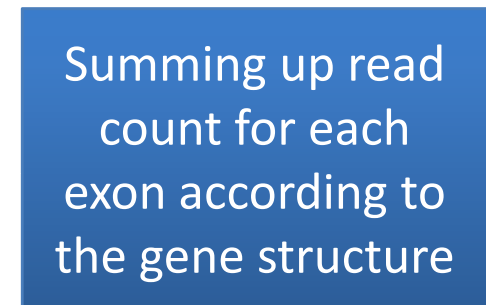
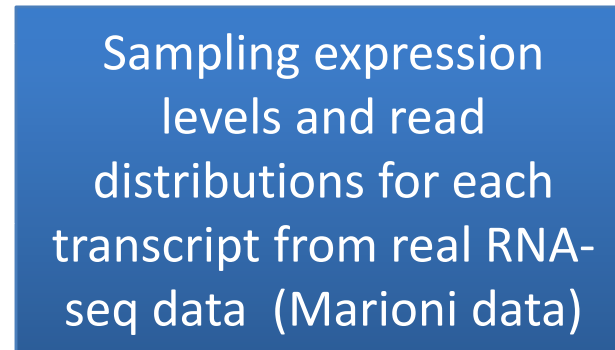
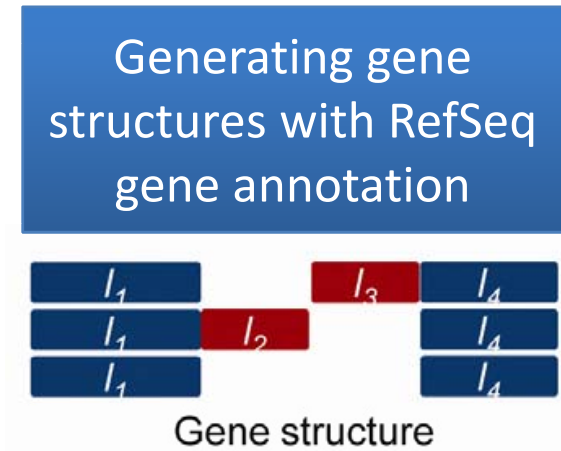
$$\log(\mathcal{L}(\Theta | x_1, x_2, \dots, x_n)) = -w \sum_{j=1}^n \sum_{i=1}^m l_j a_{ij} \theta_i + \sum_{j=1}^n x_j \log \left(l_j w \sum_{i=1}^m a_{ij} \theta_i \right) - \sum_{j=1}^n \log(x_j!)$$

- N-URD model (Wu Z *et al*, *Bioinformatics*, 2010)

- Log-likelihood function

$$\log(\mathcal{L}(\Theta | x_1, x_2, \dots, x_n)) = -w \sum_{j=1}^n \sum_{i=1}^m l_j b_{ij} \theta_i + \sum_{j=1}^n x_j \log \left(l_j w \sum_{i=1}^m b_{ij} \theta_i \right) - \sum_{j=1}^n \log(x_j!)$$

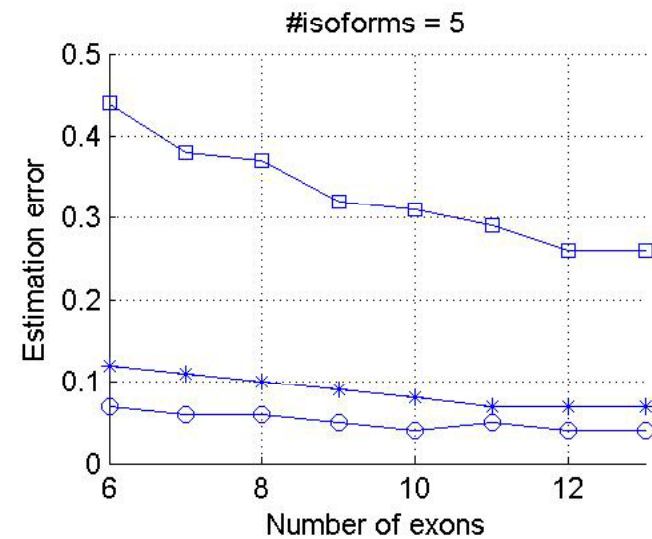
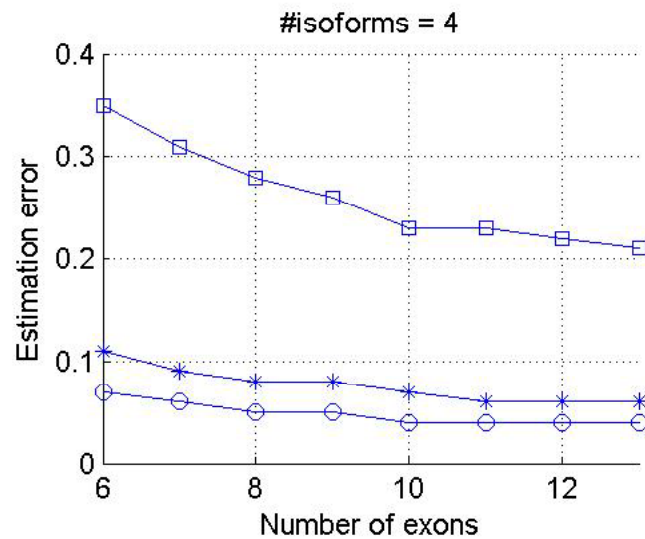
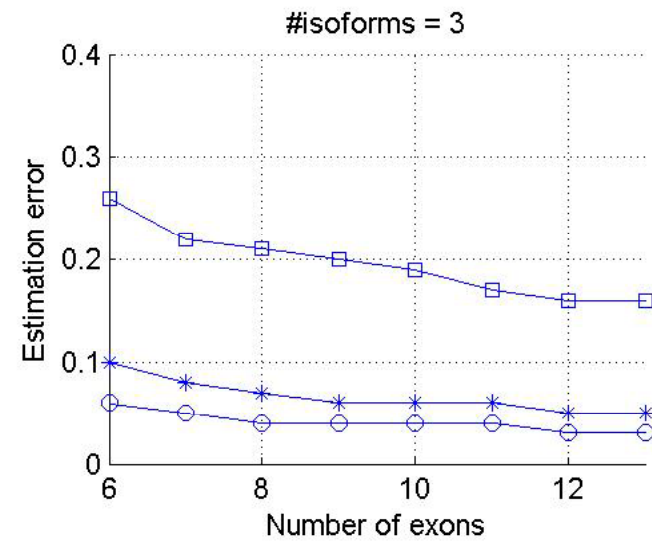
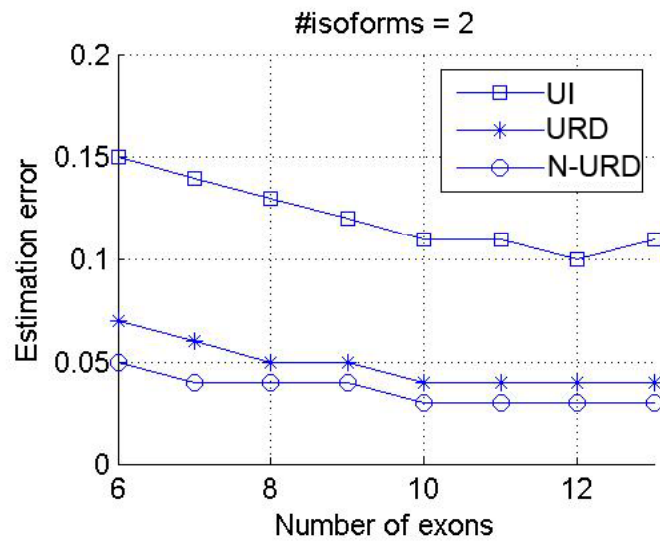
Simulation design



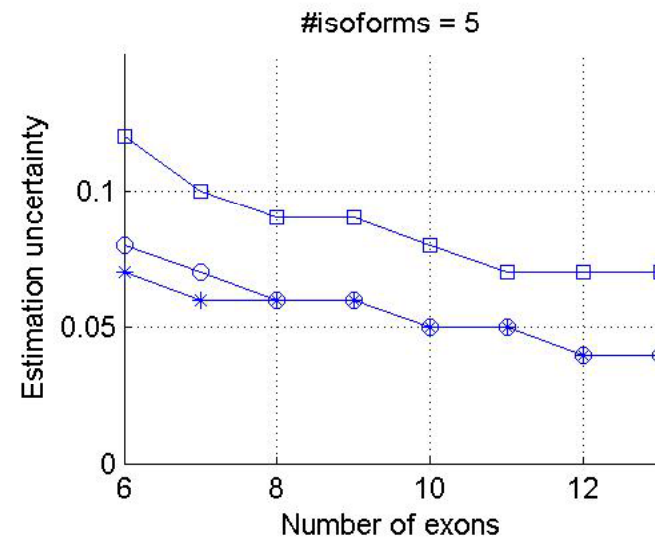
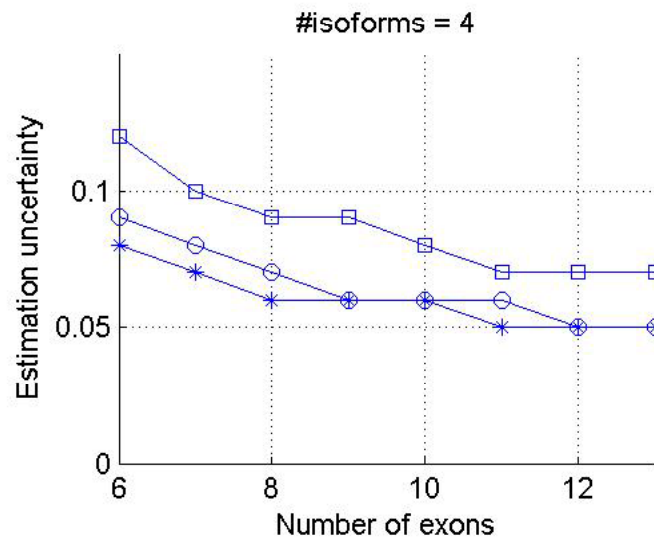
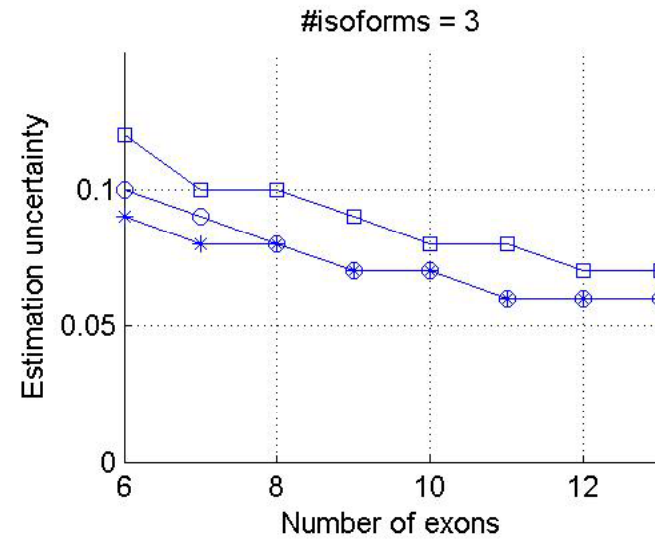
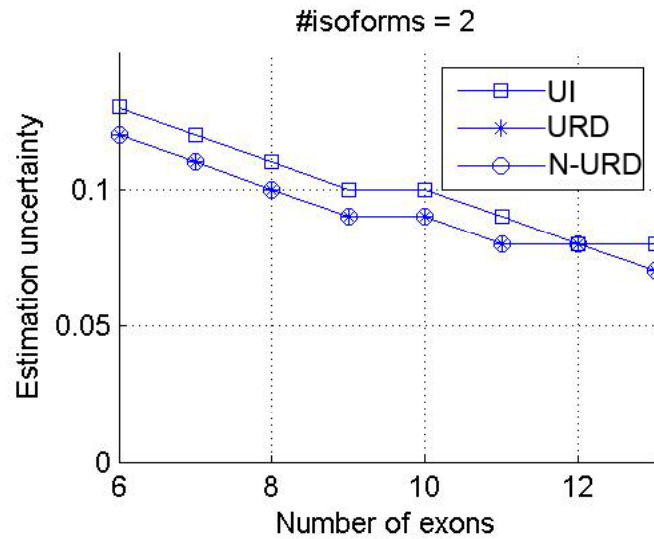
Evaluation of estimation

- Estimation error
 - The distance between true & estimated values
 - Biological meaning
 - Modeling the gene expression using transcription factors' binding affinities
 - Identification of DEGs
- Estimation uncertainty
 - Measured by 95% confidence interval
 - Biological meaning
 - Identification of DEGs -> detection power

Comparison: estimation error



Comparison: estimation uncertainty

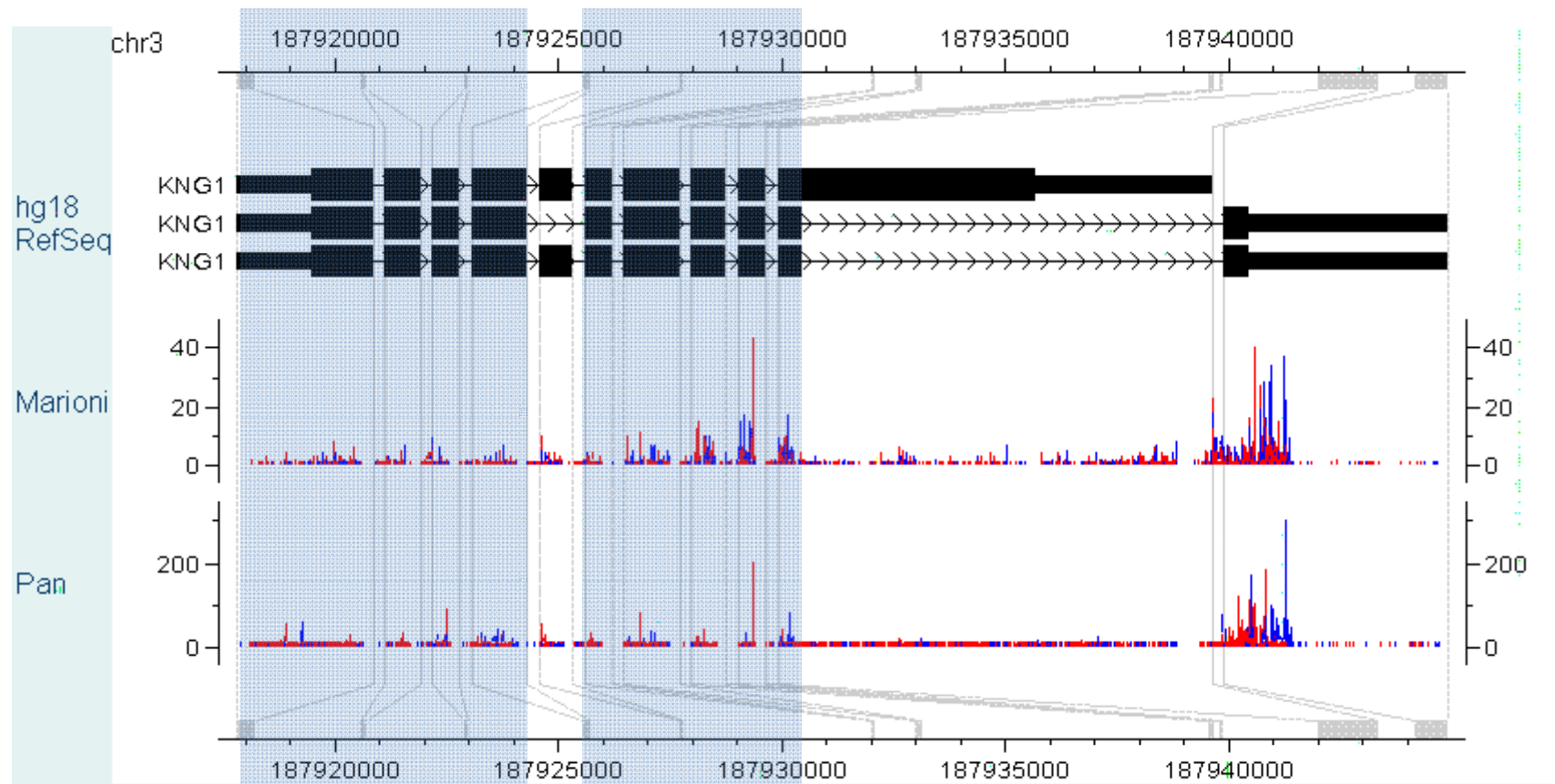


Comparison results

- Estimation error:
 - URD reduces 53–76% error from UI
 - N-URD reduces 67–87% error from UI
 - *t*-test: the difference is statistically significant
- Estimation uncertainty
 - URD and N-URD both have smaller uncertainties
 - URD and N-URD are very close
 - *t*-test: the difference between the two strategies is statistically significant

Illustration (I)

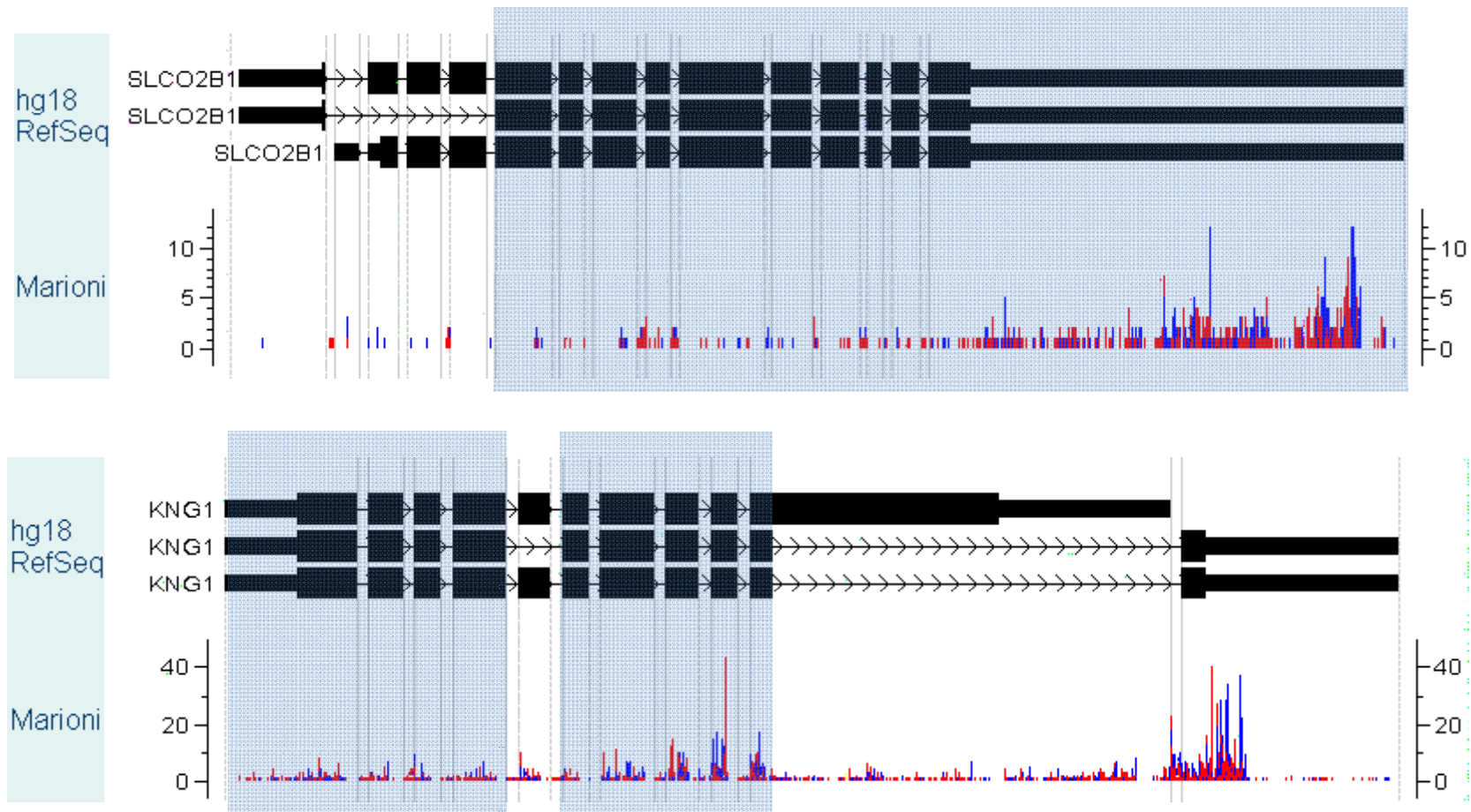
- Lose information



Sequencing preference can be averaged out, if the region measured is long enough

Illustration (II)

- Suffering from sequencing biases



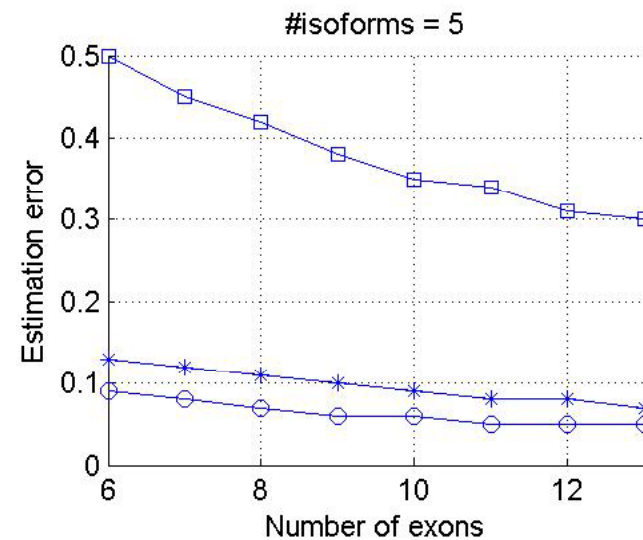
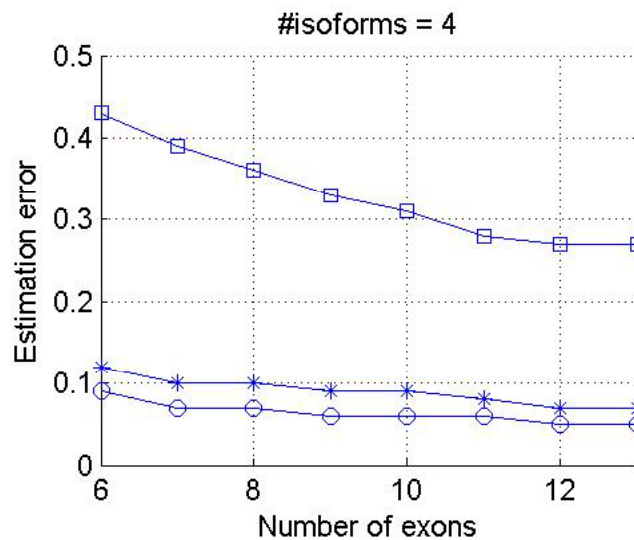
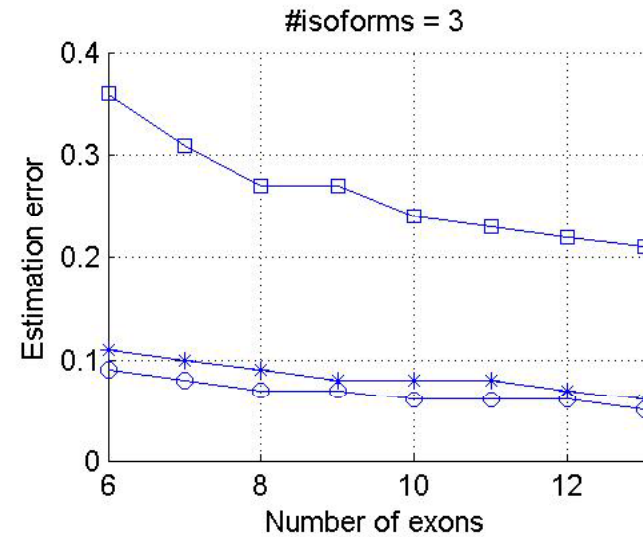
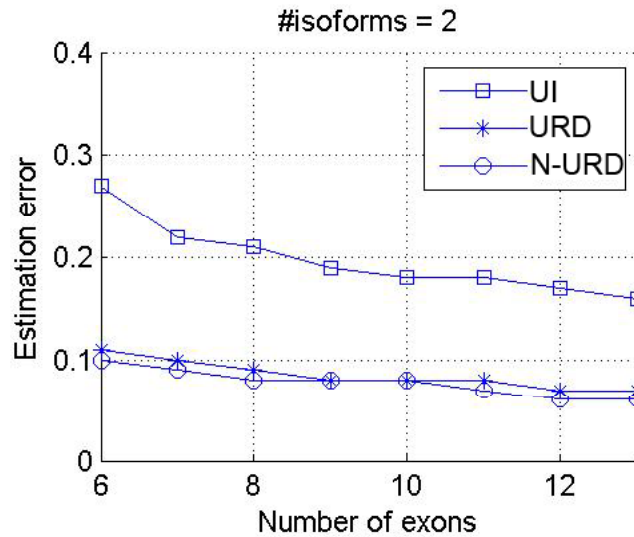
Position-depend bias may cause the gene expression estimation biased between genes

Incomplete gene annotation

- Both strategies depend on known gene structure
- How about genes with undiscovered isoforms?
- Experiment design
 - Randomly removed one isoform from the gene structure when estimating expression levels

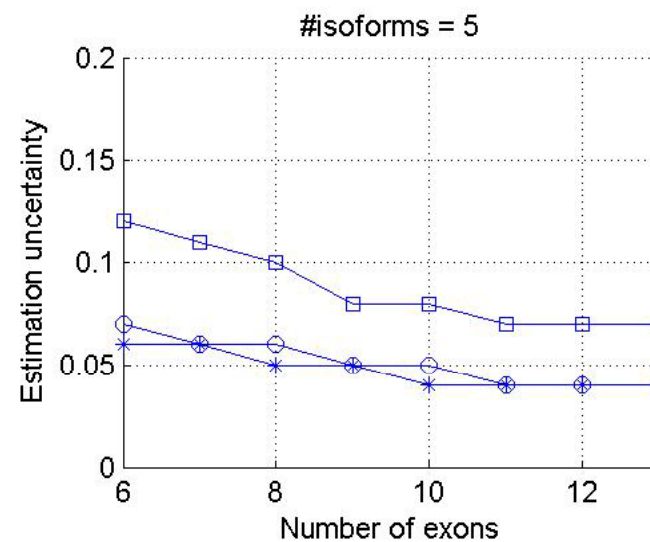
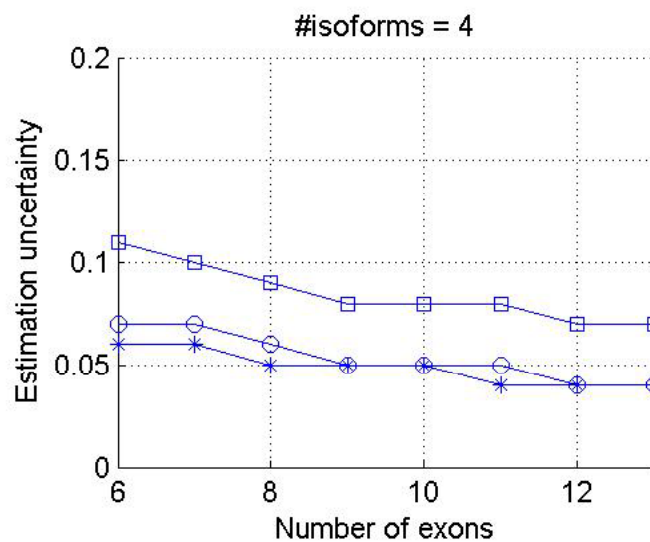
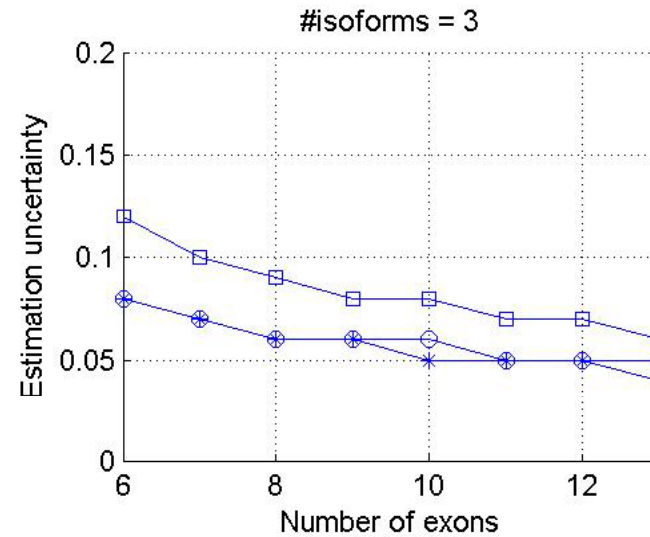
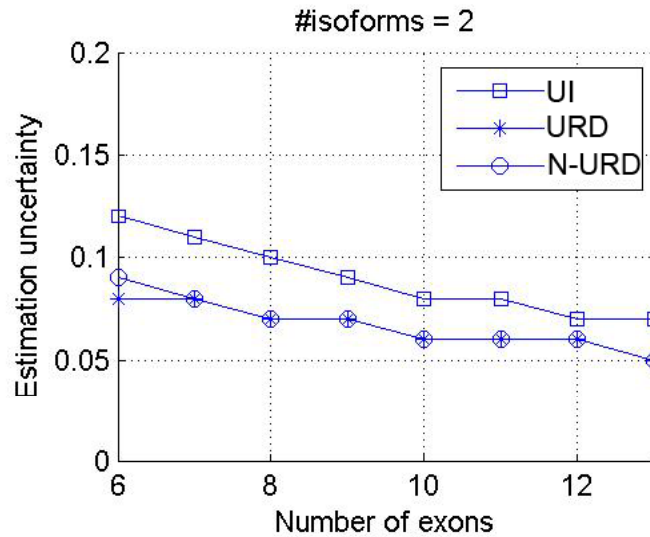
Comparison: estimation error

Incomplete gene annotation



Comparison: estimation uncertainty

Incomplete gene annotation



Real RNA-seq replicate data

- Marioni *et al* (*Genome Res*, 2008)
 - ~120m reads from human liver and kidney with technical replicates
- To assess the estimation consistency
- Isoform-based methods achieve smaller deviation

	Kidney	Liver
UI	0.0168	0.0206
URD	0.0151	0.0181
N-URD	0.0151	0.0181

Discussion

- Downstream analysis
 - Current functional knowledge is mostly at the gene level
 - Isoforms of a gene may have various functions
 - *To construct isoform-based functional annotation*
- Upstream analysis
 - Transcription followed by splicing
 - *Estimation of gene expression levels is important to study gene activation: mechanism of transcription*
 - Does splicing feed back to affect gene activation?
 - *Differentially expressed genes*
 - *Differentially spliced genes*



清華大學

Tsinghua University

Acknowledgements



Prof. Xuegong Zhang



Zhengpeng Wu



清華大學

Tsinghua University

Thanks for your attention!

Questions?