Computational Efforts to Decipher the Splicing Code

Jing Zhang

zhang28@usc.edu

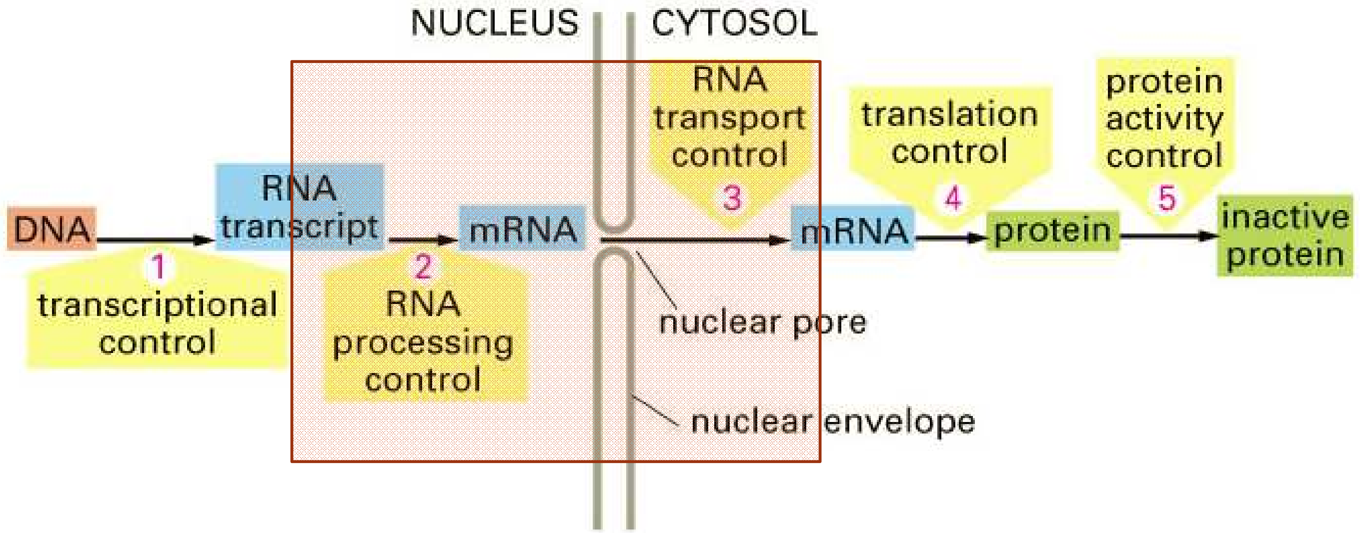
University of Southern California

Department of Electrical Engineering

Computational Biology Program, Department of Biological Science

How genes are expressed in eukaryotes?

- Gene Expression Regulation —— multi-level regulation



❖ How to quantify gene expression? —— RNA-Seq Experiments

❖ What happed during RNA processing control? —— Splicing Regulation

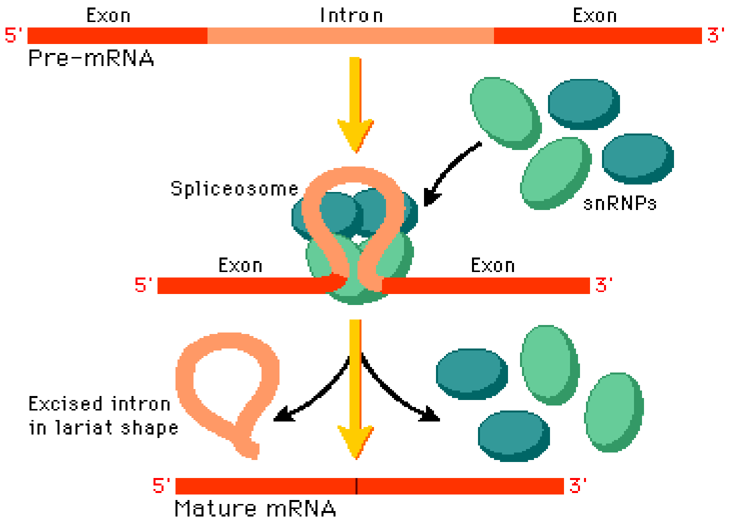
Essential Cell Biology, 2/e (@2004 Garland Science)

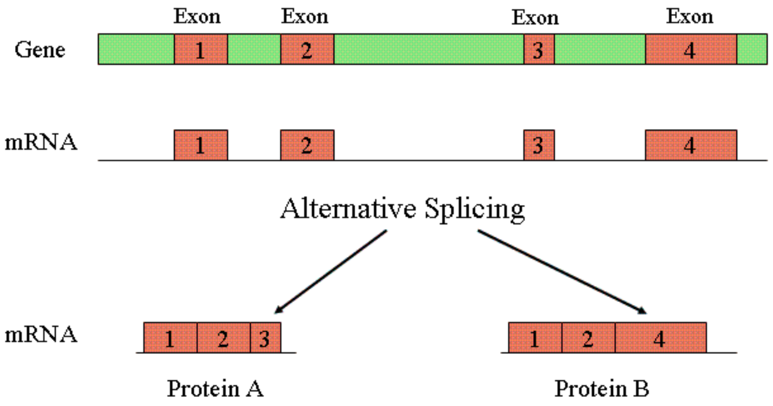
Outline

- Introduction — alternative splicing
- Part I — mRNA product quantification
  - Gene expression estimation with isoform resolution from RNA-Seq data (WemIQ)
- Part II — alternative splicing regulation
  - Part A — Context based regulation: motifs discovery via a varying coefficient regression
  - Part B — Structure based regulation: stability of mRNA secondary structures and splicing site selection
- Conclusion and future work

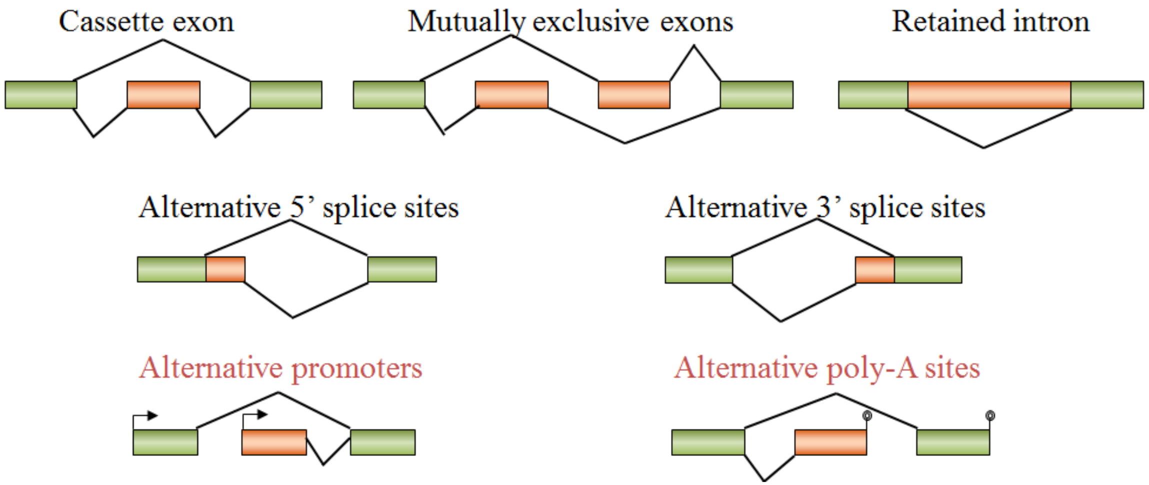
Transcriptome Diversity

- Alternative Splicing produces multiple transcript isoforms from a single gene



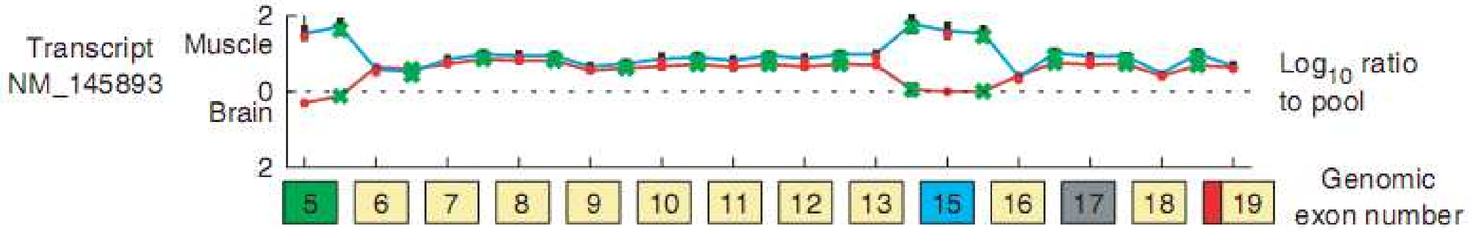


- Types of alternative splicing



Importance of alternative splicing

- **Prevalence of alternative splicing** : RNA-Seq analyses estimate that more than 90% of human genes are alternatively spliced
- **Tissue Differentiation**: transcript isoforms variations among tissues



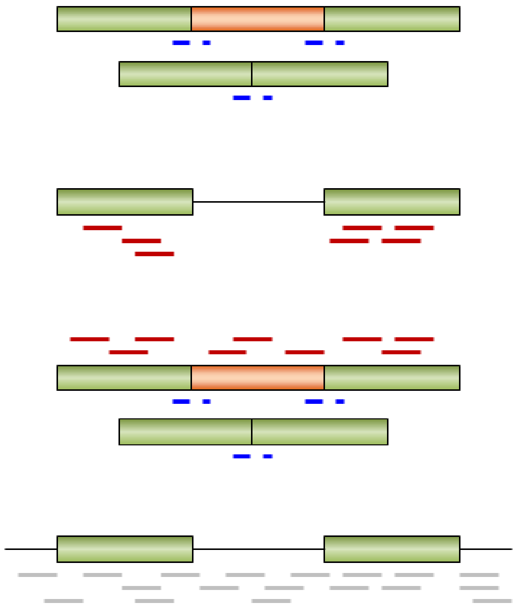
- **Diseases** : Erroneous recognition of splice sites
  - Spinal muscular atrophy: ~1 per 10,000 live births, the leading genetic cause of infant mortality, the second most common lethal autosomal recessive disorder.
  - Cystic fibrosis: 1 per 3,200-3,500 in whites (1 per 31,000 in Asian Americans), the most common lethal autosomal recessive disorder.
  - Retinitis pigmentosa: ~1 in 4000.
  - Prader-Willi syndrome: most are sporadic. 1 per 16,062 or 1 per 25,000.

Outline

- Introduction — alternative splicing
- Part I — mRNA product quantification
  - Gene expression estimation with isoform resolution from RNA-Seq data (WemIQ)
- Part II — alternative splicing regulation
  - Part A — Context based regulation: motifs discovery via a varying coefficient regression
  - Part B — Structure based regulation: stability of mRNA secondary structures and splicing site selection
- Conclusion and future work

Transcriptome Quantification

- Microarrays for mRNA expression estimation



Splice junction array

Exon tiling array

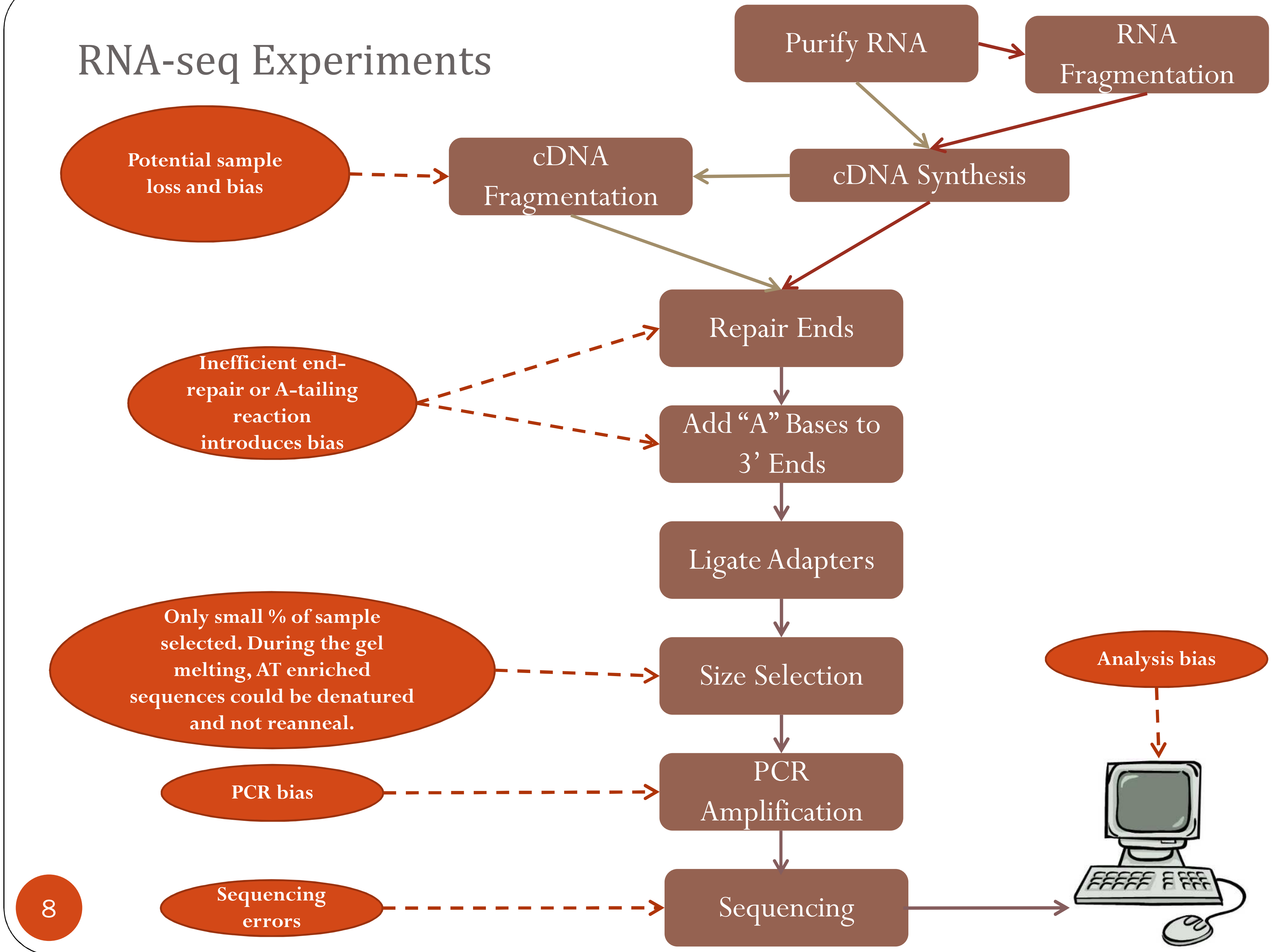
Exon array with junction and exon probes

Tiling array

- **Disadvantages**: high noise-to-signal ratio, sensitive to SNPs, rely on gene annotations/genome sequences.

- **Advantages**: can prioritize genes of interest.

RNA-seq Experiments



Purify RNA

RNA Fragmentation

cDNA Synthesis

cDNA Fragmentation

Repair Ends

Add “A” Bases to 3’ Ends

Ligate Adapters

Size Selection

PCR Amplification

Sequencing

Potential sample loss and bias

Inefficient end-repair or A-tailing reaction introduces bias

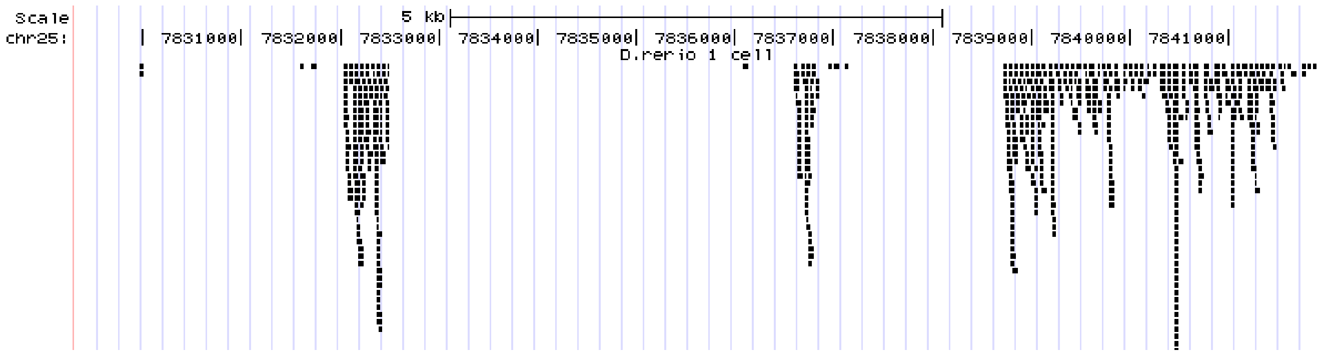
Only small % of sample selected. During the gel melting, AT enriched sequences could be denatured and not reanneal.

PCR bias

Sequencing errors

Analysis bias

RNA-seq Quantification



- **Challenge:** Remove the potential bias when estimating expression levels.
- **Focus**: Position-level read count (i.e. the number of sequence reads starting from each position of a gene/exon)
- **Common Assumptions**: Position-level read count follows a Poisson distribution with rate θ. Or equal probability of a read starting from a specific position.
  - The gene/exon length-normalized read count: maximum likelihood estimator (MLE) of θ.
  - RPKM (Reads per kilobase of exon model per million mapped reads) = length-normalized read count / total mapped reads (in million).
  - Doesn't consider the bias.

Generalized Poisson

- Probability mass function (Consul 1989):

$$\Pr(X = x) = \begin{cases} \theta(\theta + x\lambda)^{x-1} e^{-\theta - x\lambda} / x!, & x = 0, 1, 2, \ldots, \\ 0 & \textit{for } x > q \textit{ if } \lambda < 0, \end{cases}$$

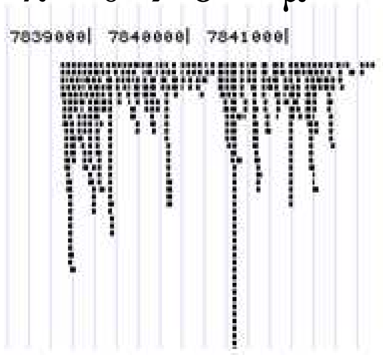
where $\theta > 0$, $\max(-1, -\theta/q) \leq \lambda \leq 1$, and q ($\geq 4$) is the largest positive integer for which $\theta + q\lambda > 0$ when $\lambda < 0$.

- θ is the average rate for the natural Poisson process (expression level)
- λ is the average rate of the effort that the subjects are making to deviate from the process, a measure of the departure from Poissonicity (bias).

$\lambda > 0 \rightarrow \sigma^2 > \mu$

$\lambda < 0 \rightarrow \sigma^2 < \mu$

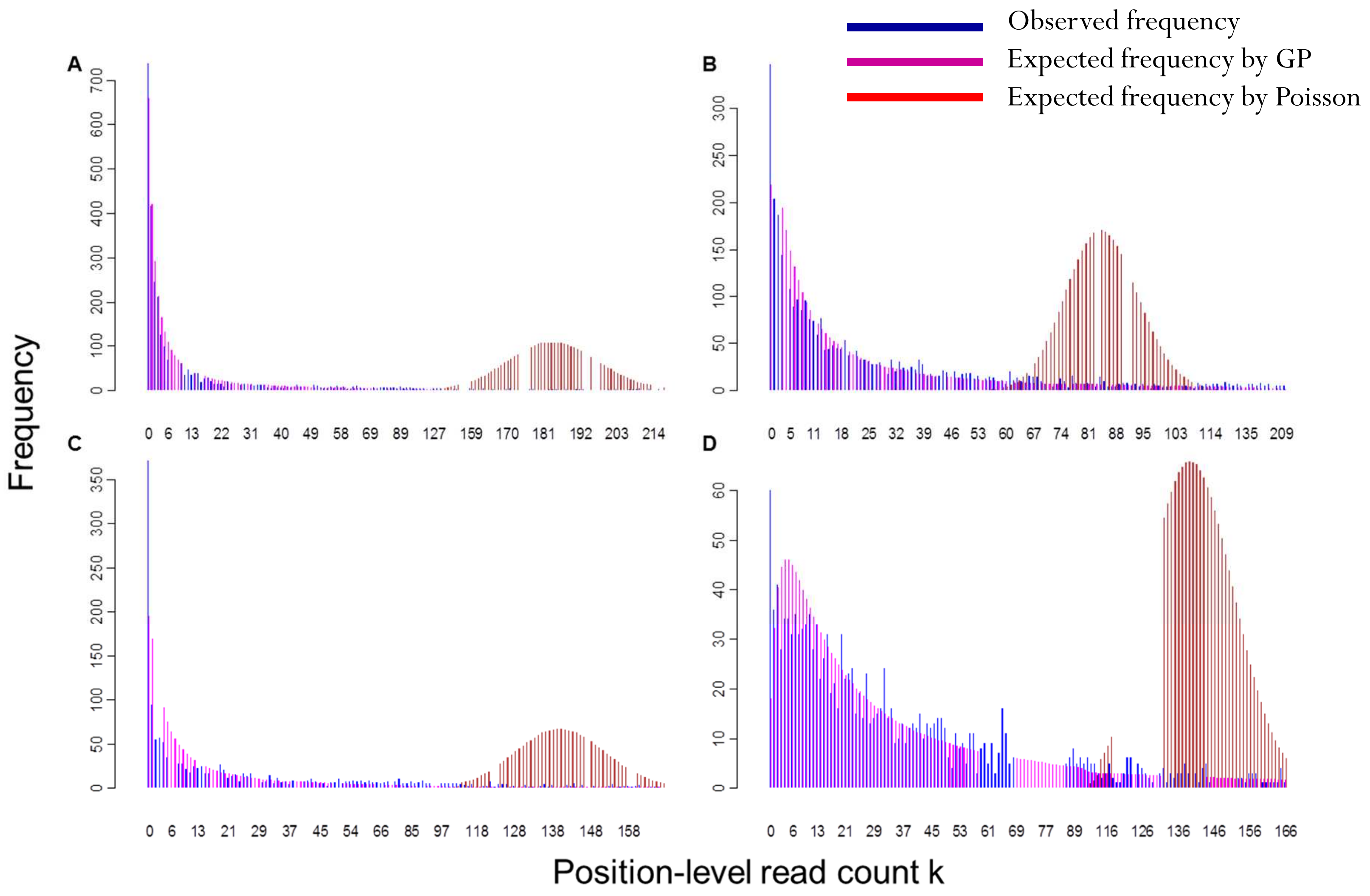
$\lambda = 0 \rightarrow \sigma^2 = \mu$



	Gene level		Exon level	
	GP	Poisson	GP	Poisson
MAQC data	85.72%	1.57%	89.62%	19.71%
Human data	77.28%	3.22%	88.78%	28.35%
Mouse data	88.57%	7.88%	91.73%	39.67%
Yeast data	93.24%	20.49%	93.21%	23.73%
MAQC-2_sep	92.93%	10.18%	92.90%	41.51%

Srivastava, S. and L. Chen*,A two-parameter generalized Poisson model to improve the analysis of RNA-seq data.?Nucleic Acids Research, 2010. 38(17): p. e170.

Examples of Poisson and Generalized Poisson fitting

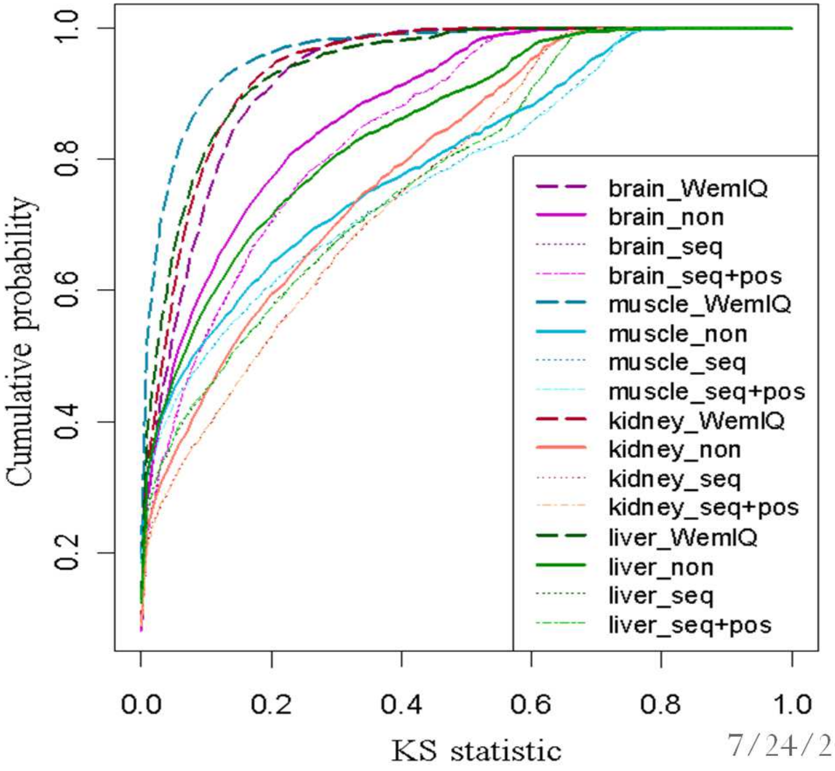


Srivastava, S. and L. Chen*,A two-parameter generalized Poisson model to improve the analysis of RNA-seq data.?Nucleic Acids Research, 2010. 38(17): p. e170.

Bias removal through Generalized Poisson model

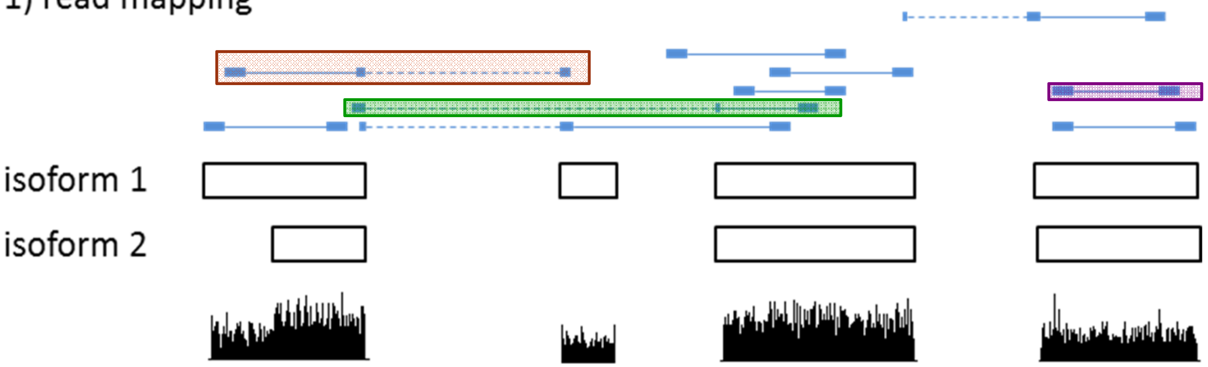
- **The heterogeneity of read counts:**
  - expression varies among genes
  - regions shared by multiple isoforms are expected to contain more reads than regions specific to a single isoform
  - inherent experimental bias of the RNA-Seq protocol
- **Goal**: remove the inherent bias while estimating the transcript isoform expression.

- Single-isoform genes
  - Data-adaptive bias correction by GP in WemIQ (WemIQ)
  - Correct the sequence-specific bias from random hexmer priming (seq)
  - Correct the bias from relative positions (pos)

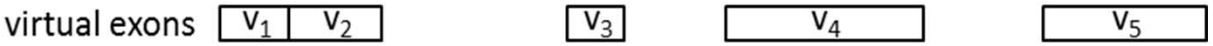


WemIQ: A EM based mRNA quantification Method

1) read mapping



2) virtual exon assignment and overdispersion parameter estimation



overdispersion λ_1 λ_2 λ_3 λ_4 λ_5

$$P(X_s = x) = \begin{cases} \theta_s(\theta_s + x\lambda_s)^{x-1} e^{-\theta_s - x\lambda_s} / x!, & x = 0,1,2,\cdots \\ 0 & x > q \text{ if } \lambda_s < 0 \end{cases}$$

3) gene and isoform expression quantification by weighted EM

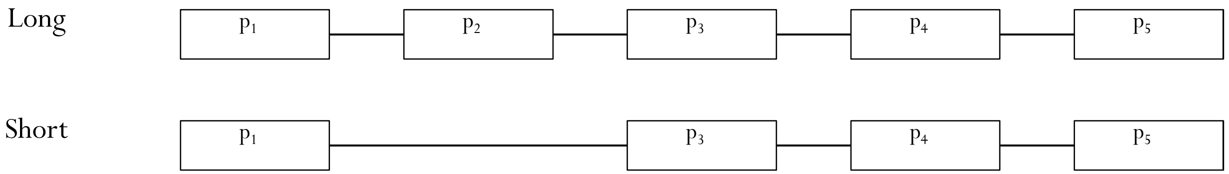
$$weighted\ \log(P\{\boldsymbol{R},\boldsymbol{\pi}|\boldsymbol{\tau}\}) = \sum_{i=1}^{n}\sum_{j=1}^{m}(1-\lambda_{S_{b_i}})I(\pi_i = j)\log\left(\frac{1}{L'_j} \times P\{l_{i,j}\} \times \tau_j\right)$$

Jing Zhang, C.-C. Jay Kuo, Liang Chen, WemIQ: a weighted-log-likelihood expectation maximization method for isoform quantification from RNA-Seq data (in Submission)

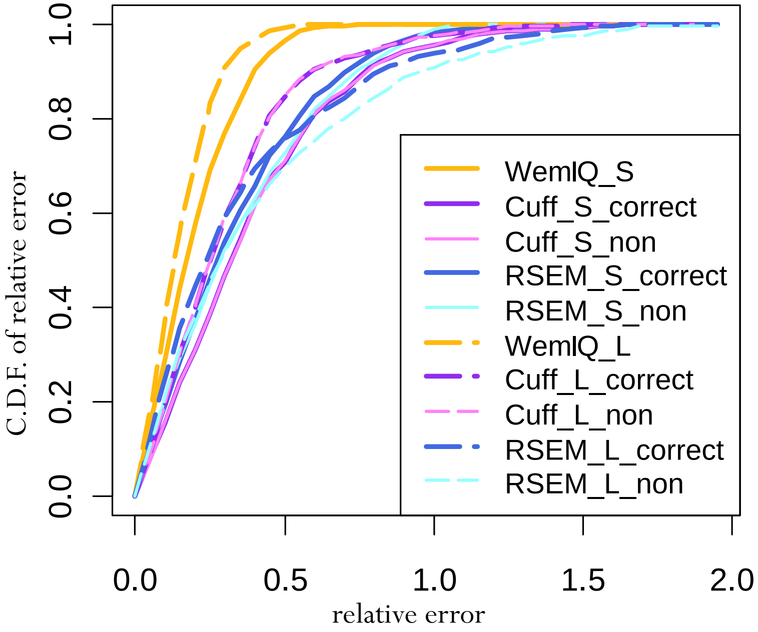
Simulation Study

- Splice Ratio Estimation in a cassette exon case

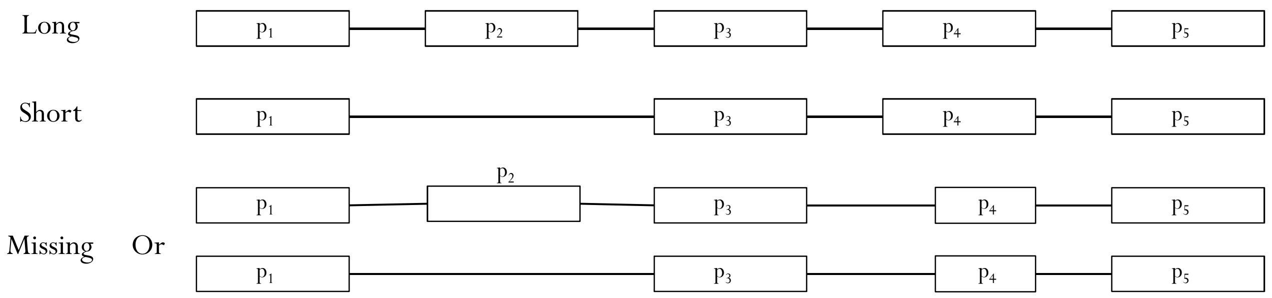
$$\left.\begin{array}{r} r \to \{2,3,4\} \\ p_i \to uniform[0.75, 0.95] \end{array}\right\} \Rightarrow \left\{\begin{array}{l} NB(r, p_i) \\ NB(1, p_i) \end{array}\right.$$

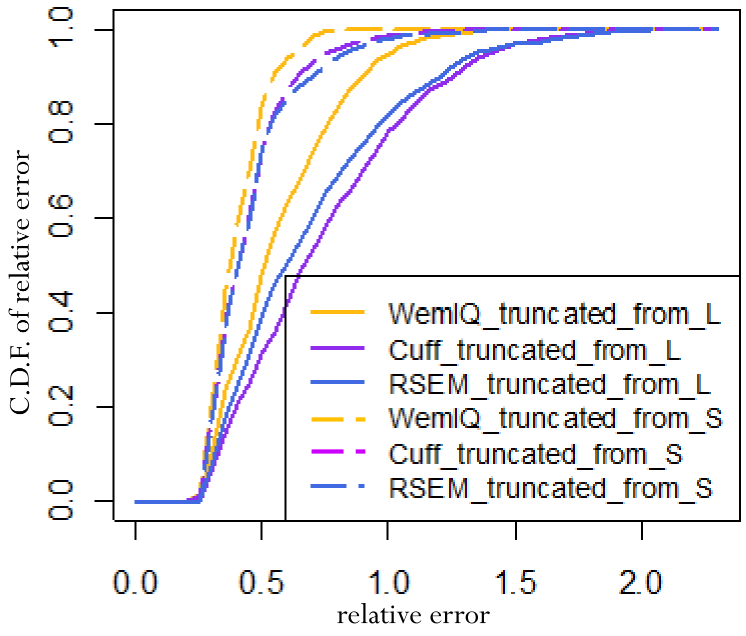


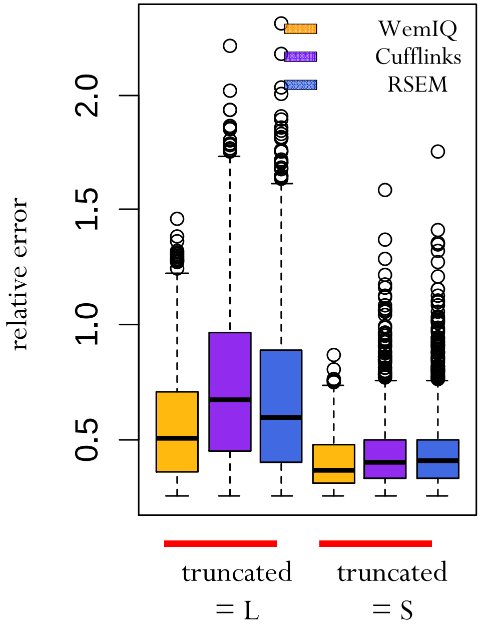
- WemIQ improves relative isoform expression
- Cufflinks has similar performance with/without bias correction
- RSEM improves its performance by using its empirical positional bias correction



Simulation Study (continued)



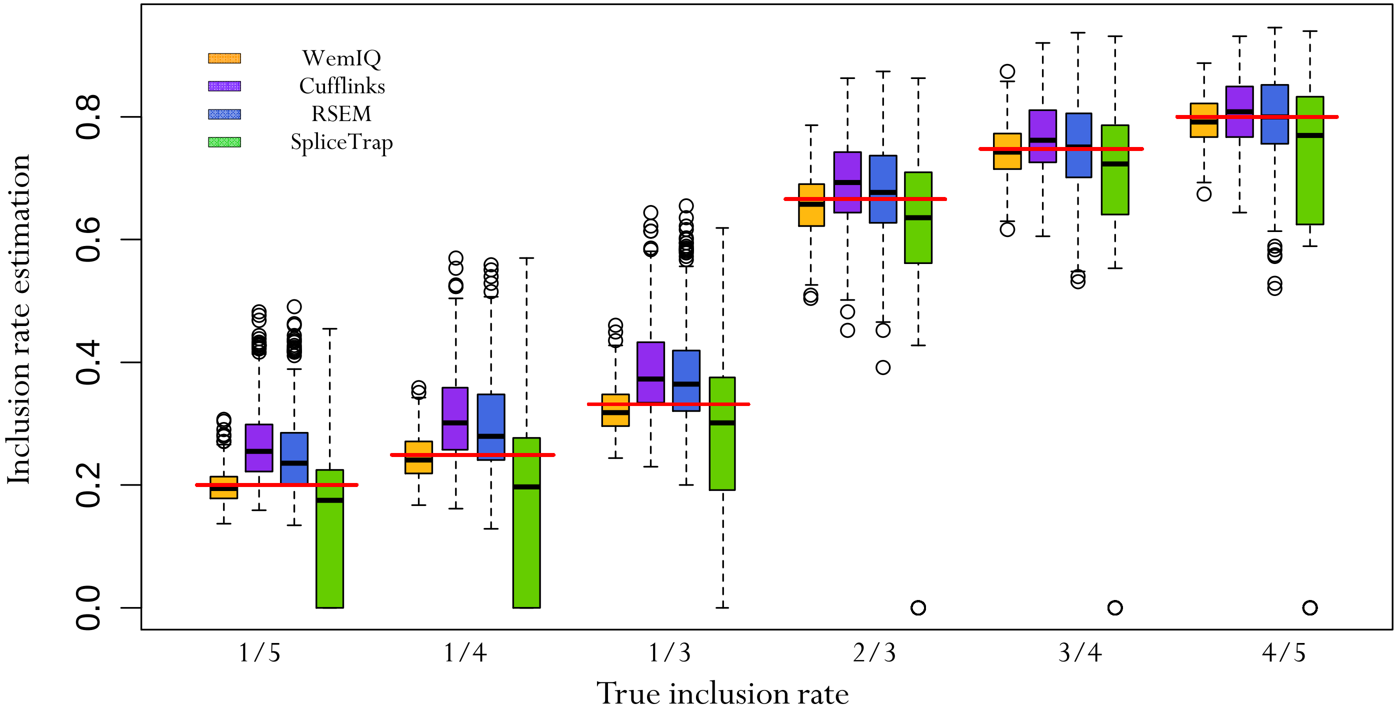




Simulation Study (exon centric measurements)

- Splice Ratio Estimation in a cassette exon case

$$\text{exon inclusion rate} = \frac{\sum \text{isoform expression with this exon}}{\sum \text{isoform expression}}$$



- 18.85%, 21.60%, and 40.85% of the exons in Cufflinks, RSEM and SpliceTrap have estimation error >0.1
- *1.35%* of the exons in WemIQ have estimation error >0.1

Real data analysis

- **qRT-PCR at the gene level**: MAQC data
  - TaqMan qRT-PCR results on approximately 1,000 genes
  - at least 75% of the qRT-PCR replicates had a detectable expression
  - Finally, 526 genes were compared across methods and platforms. The correlation of the qRT-RCR data and WemIQ was **0.739**, higher than those of Cufflinks (0.681) and RSEM (0.700)
- **Two independent RNA-Seq experiments:** two labs in GM12878 cell

Resolution	Method	A1 VS. B1	A1 VS. B2	A2 VS. B1	A2 VS. B2
Isoform	WemIQ	**0.713**	**0.817**	**0.696**	**0.798**
	Cufflinks	0.679	0.769	0.587	0.695
	RSEM	0.577	0.749	0.517	0.680
Genes	WemIQ	**0.738**	**0.835**	**0.721**	**0.816**
	Cufflinks	0.684	0.770	0.588	0.692
	RSEM	0.576	0.749	0.514	0.679

Real data analysis (continued)

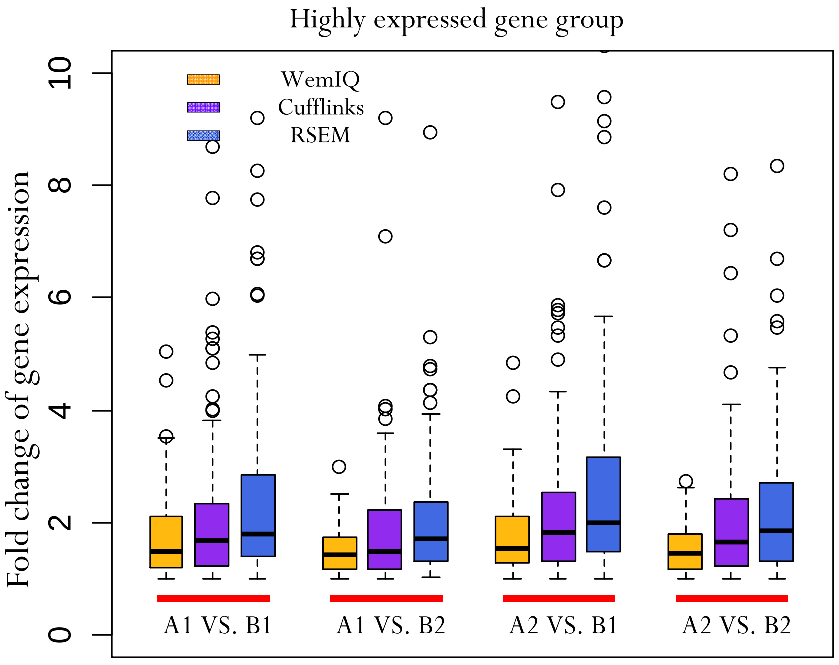
- **Results consistency on a group of highly and moderately expressed genes**

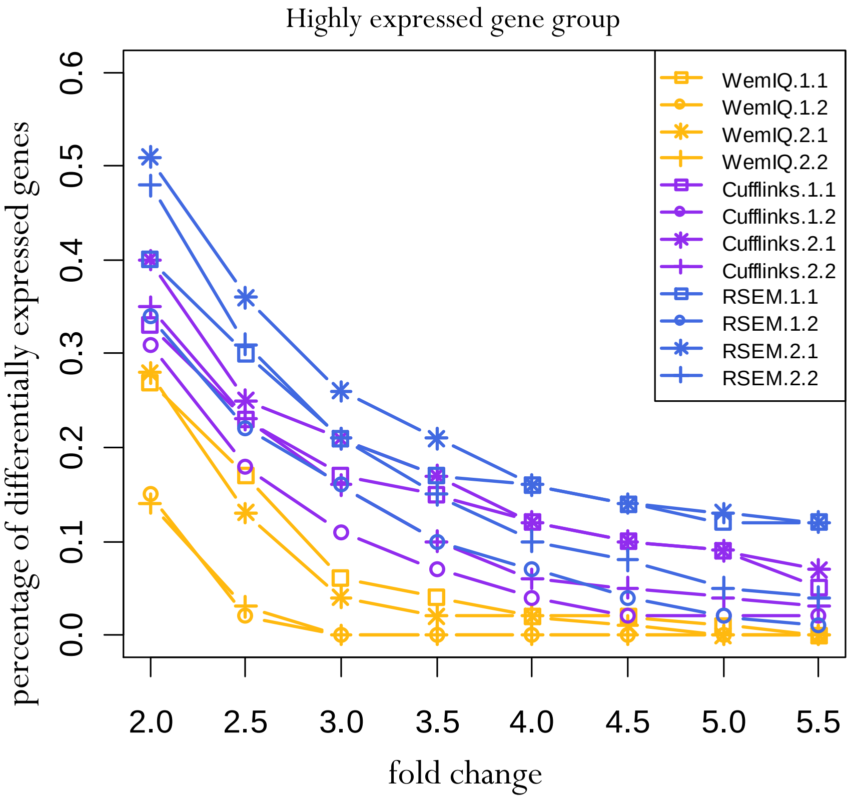
group	Resolution	Method	A1 VS. B1	A1 VS. B2	A2 VS. B1	A2 VS. B2
Highly expressed genes	Isoform	WemIQ	**0.752**	**0.844**	**0.736**	**0.823**
		Cufflinks	0.706	0.789	0.609	0.713
		RSEM	0.753	0.831	0.659	0.749
	Genes	WemIQ	**0.772**	**0.854**	**0.759**	**0.836**
		Cufflinks	0.723	0.806	0.621	0.727
		RSEM	0.758	0.839	0.661	0.753
Moderately expressed genes	Isoform	WemIQ	**0.448**	**0.618**	**0.433**	**0.606**
		Cufflinks	0.376	0.473	0.324	0.422
		RSEM	0.329	0.489	0.293	0.436
	Genes	WemIQ	0.502	**0.682**	**0.490**	**0.670**
		Cufflinks	0.429	0.483	0.390	0.448
		RSEM	**0.504**	0.664	0.436	0.596

WemIQ provides more consistent estimation across different experiments on the same tissue!

Real data analysis (continued)

- **Number of differentially expressed genes in technical replicates**





- In technical replicates, ideally there should be no differentially expressed genes
- WemIQ claims much less DE genes than Cufflinks and RSEM

WemIQ: short conclusion

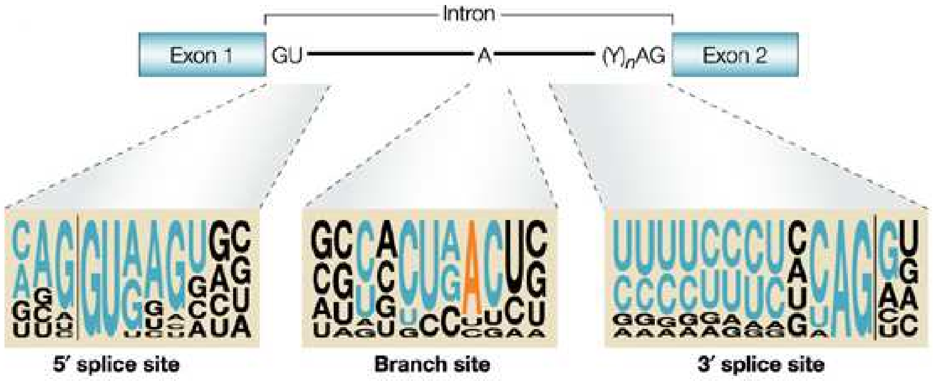
- **Characteristics of WemIQ:**
  - Use Generalized Poisson model to handle the over dispersion of the read count data from RNA-Seq
  - Estimate the biases in a data driven manner
  - Allocate the reads across isoforms through EM algorithm
  - Use bias parameter to assign different weights in EM
- **Advantage of WemIQ**
  - Simulation study shows improved isoform percentage estimation
  - Real RNA-Seq experiments provided more consistent estimates

Outline

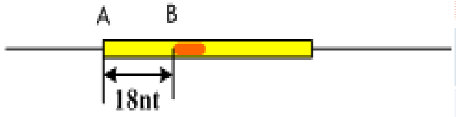
- Introduction — alternative splicing
- Part I — mRNA product quantification
  - Gene expression estimation with isoform resolution from RNA-Seq data (WemIQ)
- Part II — alternative splicing regulation
  - Part A — Context based regulation: splice factor binding sites discovery via a varying coefficient regression
  - Part B — Structure based regulation: stability of mRNA secondary structures and splicing site selection
- Conclusion and future work

How splice sites are recognized?

- **Splicing code part 1**: consensus at the junction and branch point



- **Is the consensus enough** : existence of decoy splice sites



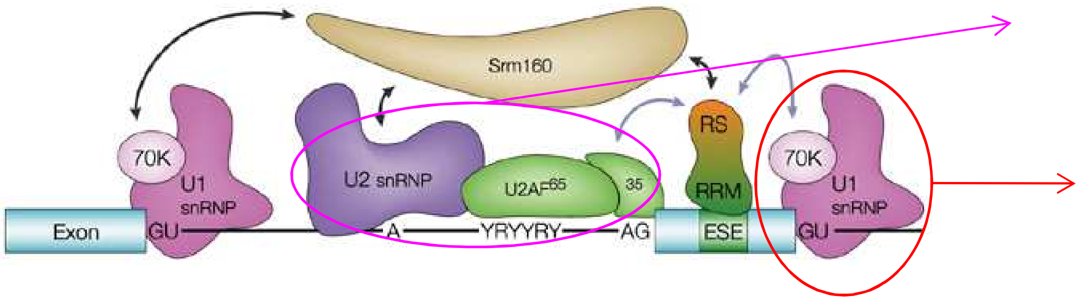
Name	position	sequence	score
A	CHR 1, 36706391	GCTTTACCATCCAGGCCTATGCTG	5.30
B	CHR 1, 36706409	CTCTCTTCCCATAGCTCCCTCCCA	10.39

- **Is branch point enough**: branch point insertion fails to promote recognition

Sun H, Chasin LA: **Multiple splicing defects in an intronic false exon.** *Mol Cell Biol* 2000, **20**(17):6414-6425.

How splice sites are recognized?

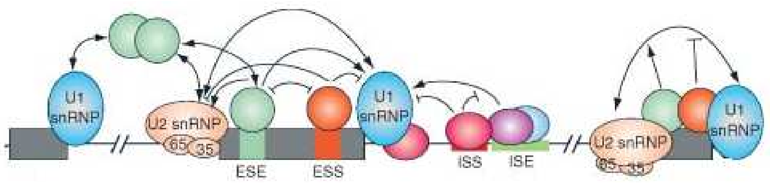
- **Splicing code part 2**: involvement of proteins (SR protein family & hnRNPs)



Indirect involvement

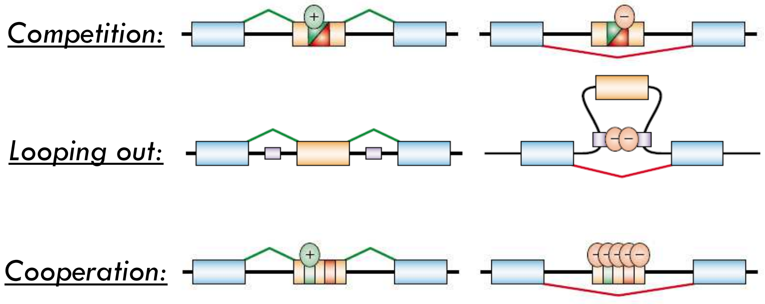
Direct involvement

- Where does proteins bind? SREs



	exonic	intronic
enhance	ESE (exonic splicing enhancer)	ISS (intronic splicing silencer)
repress	ESS (exonic splicing silencer)	ISE (intronic splicing enhancer)

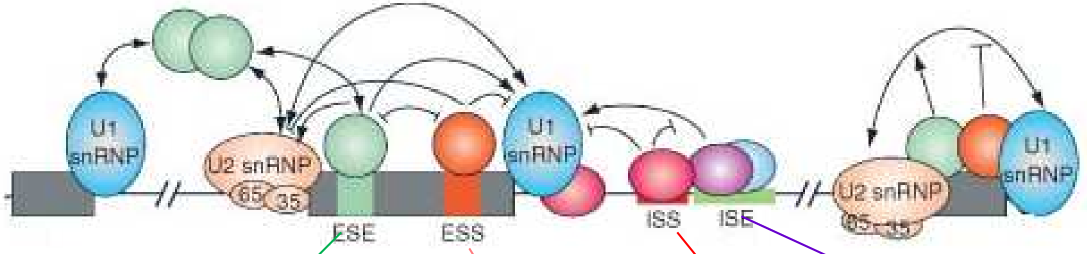
- How does SRE work?



SRE: splicing regulatory elements

Why SRE discovery is difficult?

- **Problem 1: individual SRE recognition**



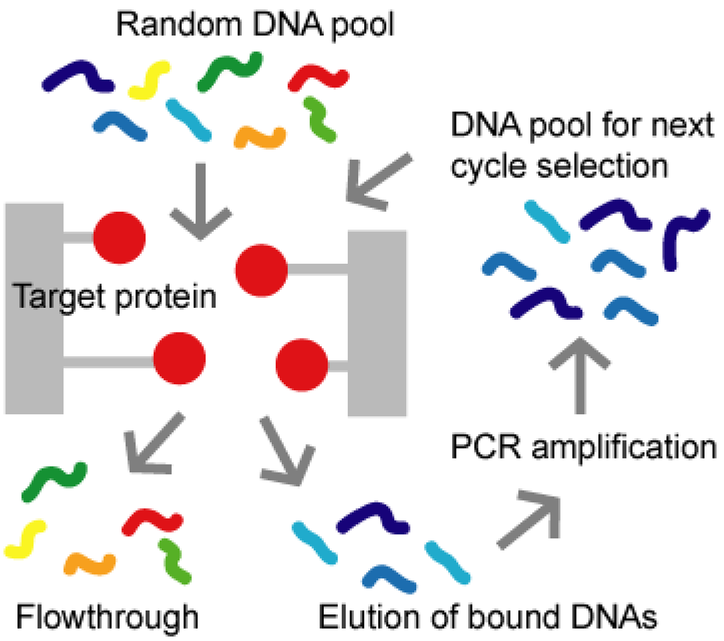
```
ATAATAATAATACAAAAATTAGCCGGGTGCAGTGGCGCATAGTCAAACAAGATTAACCTCCTGCCCATAGTCCC
CTGCACACAGCACAAGACAGGGCTAGCGACTGTCAGATGGATGAGTGCTCCACTCCTCTCTTCCCAGACCCA
CCCCACAGTGTCTGGGAAGCCACAAGAAGTCCAACCGGGCTCTTGCTCTGCCCCACTCCCTTCAGCCCCCGC
ACCTCTCTCTTCCCATAGCTCCCTCCCATGCCCCAGAGCTGCATCTAAAGCACATTGGCAAGACCTCAGATTTTT
GGACTCCCAAGTCTGTGCTGTTTCCTCACCAGACCGCAACTTCTCTCTGGTTCTAATAGTGAAATGCTATAGGAA
TGATTTTTTCCTGAAAAGATGGTGCATATTTGACCTGATTCATTGAAACTGCTTGGGTCCATCTCCAAAGCCCCTG
CTAGCTTTTAGGAGCAATGGCTTATCTGCTGATACCCTCCTGTTCCCAGCCAGGACTATCAACATGGTATCTTCCA
ATCAATTGGCTTCAAGGAATTTCACGAGTACCTGATCACTGAGGATAGCTCCCTCCCATGCCCCAGAGCTGCAT
```

- **Goals:**
- Where are they and how do they work?

- **Challenges:**
  - Length of motif: 4~7 nt
  - Tissue preference
  - Positional preference
  - Functional preference

Existing methods — experimental approaches

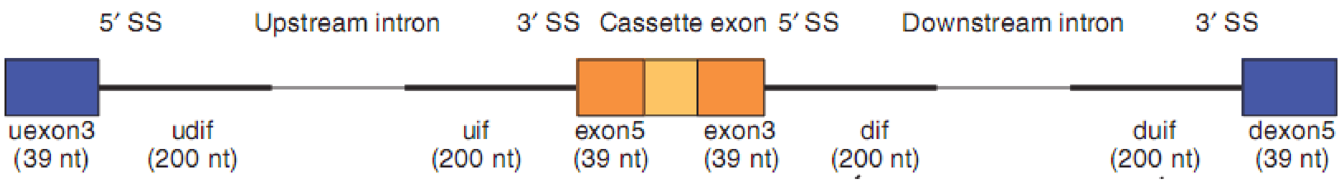
- **experimental predictions -** SELEX

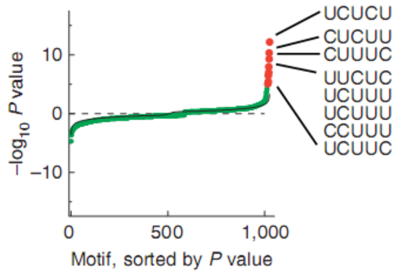


- Basic idea
  - amplify the binding sites
  - remove the non-binding ones
- Strength of SELEX
  - binding site guarantee
- Weakness of SELEX
  - very limited number of factors are know till now
  - *in vitro* binding instead of *in vivo*
  - motif discovery, not the binding site discovery

Existing methods — computational approaches

- **word count/enrichment analysis : more frequent — more functional ?**





Frequent → functional?

I. No direct link of functionality

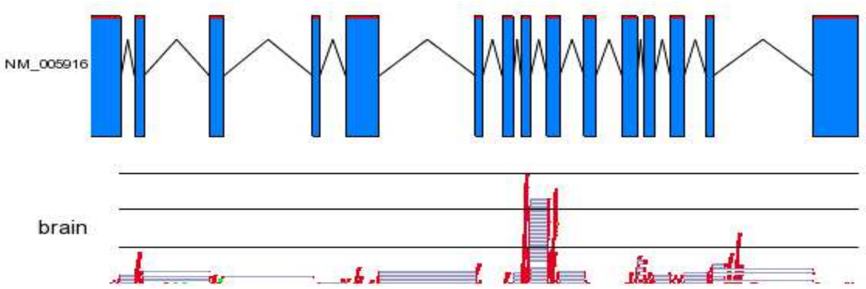
II. How to quantify its contribution to spling

Functional→ frequent ?

I. Control group selection

II. Multiple splicing factor

- **linear regressions: more correlated — more functional**



Y: exon inclusion rate

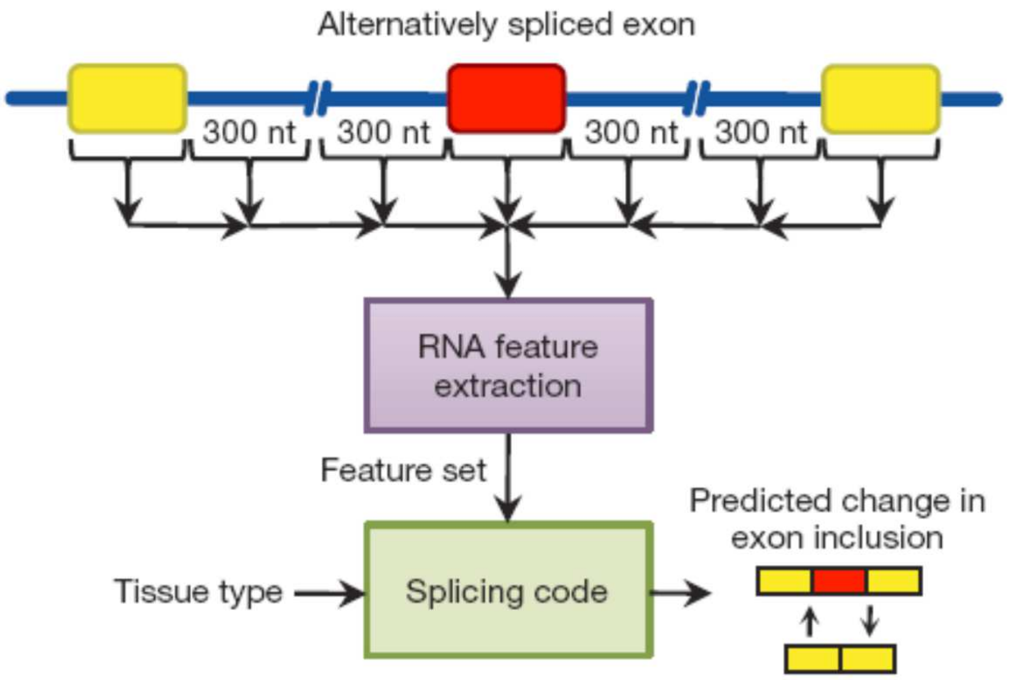
X: motif occurrence

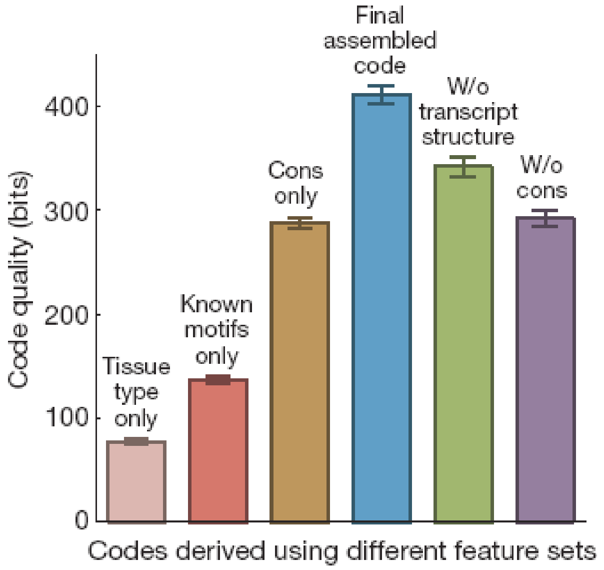
$$y = ax + b$$

Why linearity?

Motivation 1: Prior Info Helps?

- **Inspiration 1: what else besides the exon inclusion rate?**



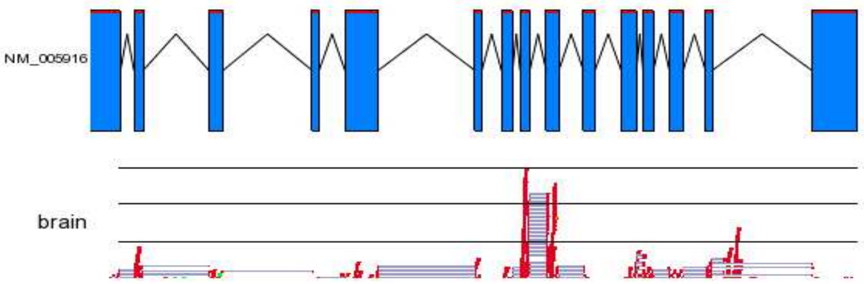


Splicing type + features => motifs

Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ, ***Deciphering the splicing code,*** Nature. 2010 May 6;465(7294):53-9.

Motivation 2: Nonlinearity? Varying Coefficient Regression

- **Inspiration 2: what else besides the exon inclusion rate?**



Y: exon inclusion rate

X: motif occurrence

$$y = ax + b$$

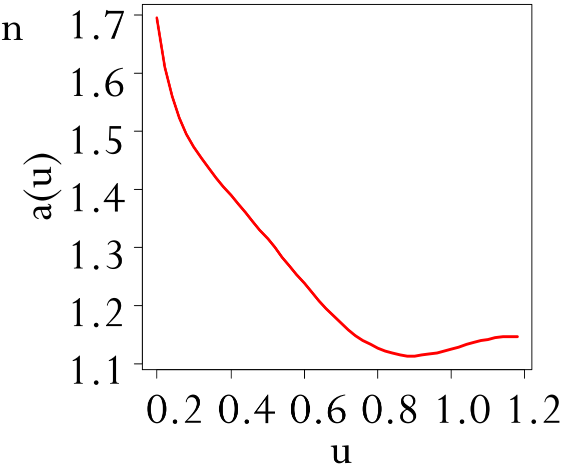
Natural extension? →

$$\boxed{y = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n + b}$$

Linearity: oversimplification of splicing regulation, no interactions ?

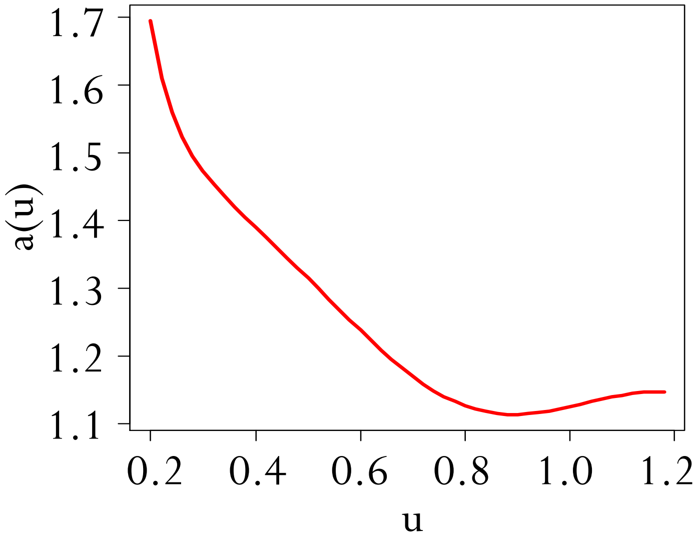
- **Model introduction :** Varying coefficient regression

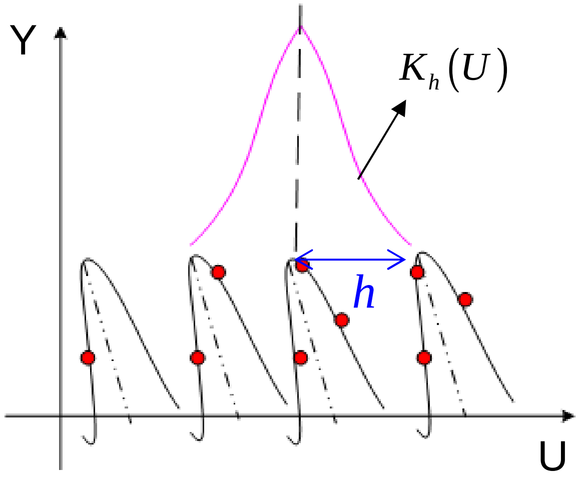
$$\boxed{\mathbf{Y}(U, \mathbf{X}) = \mathbf{X}^{\mathrm{T}} \underbrace{\mathbf{a}(U)}_{\text{varying effect}}}$$



Single SRE prediction: varying coefficient regression

- **Model introduction :** LS estimation and bandwidth selection





LS estimation:

$$\min \quad L(a,b)=\sum_{i=1}^{n}\{y_i-\mathbf{X}_\mathbf{i}^\mathrm{T}\mathbf{a}-\mathbf{X}_\mathbf{i}^\mathrm{T}\mathbf{b}(U_i-u)\}^2 K_h(U_i-u)$$

First order interpolation $a(u)=\hat{a}(u_0)+\hat{b}(u-u_0)$

h ↑ → variation ↓ ⟺ h ↓ → bias ↓

Bandwidth selection:

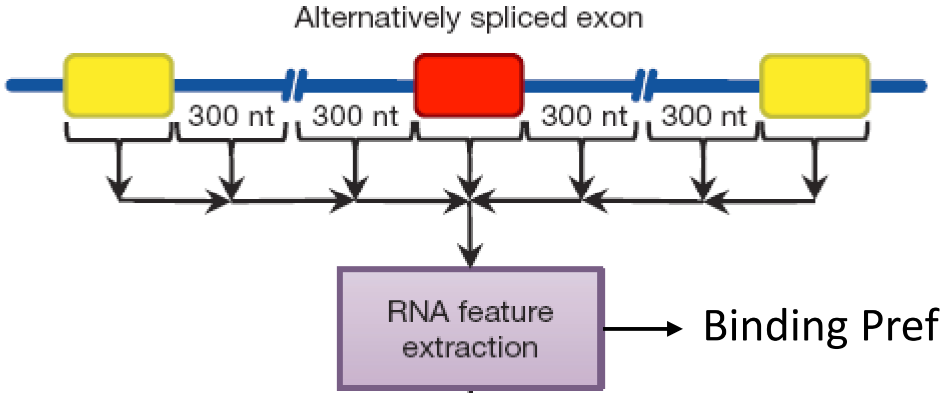
- **Theoretical calculation**
- **Leave one out cross validation**
- **AIC, BIC…**

How to predict the SREs

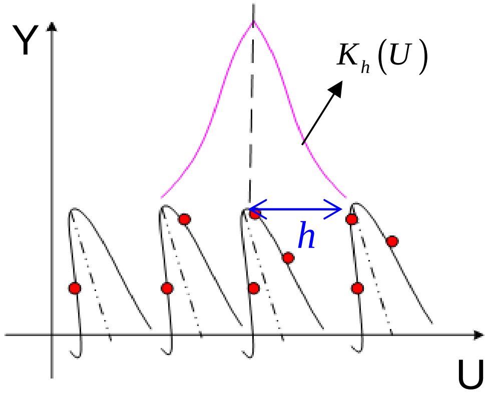
- **Model generalization** : Information integration and parametric variable

Dimension of prior information: Curse of dimensionality

How to find a neighbor in a high dimension space?



→ Binding Preference: U

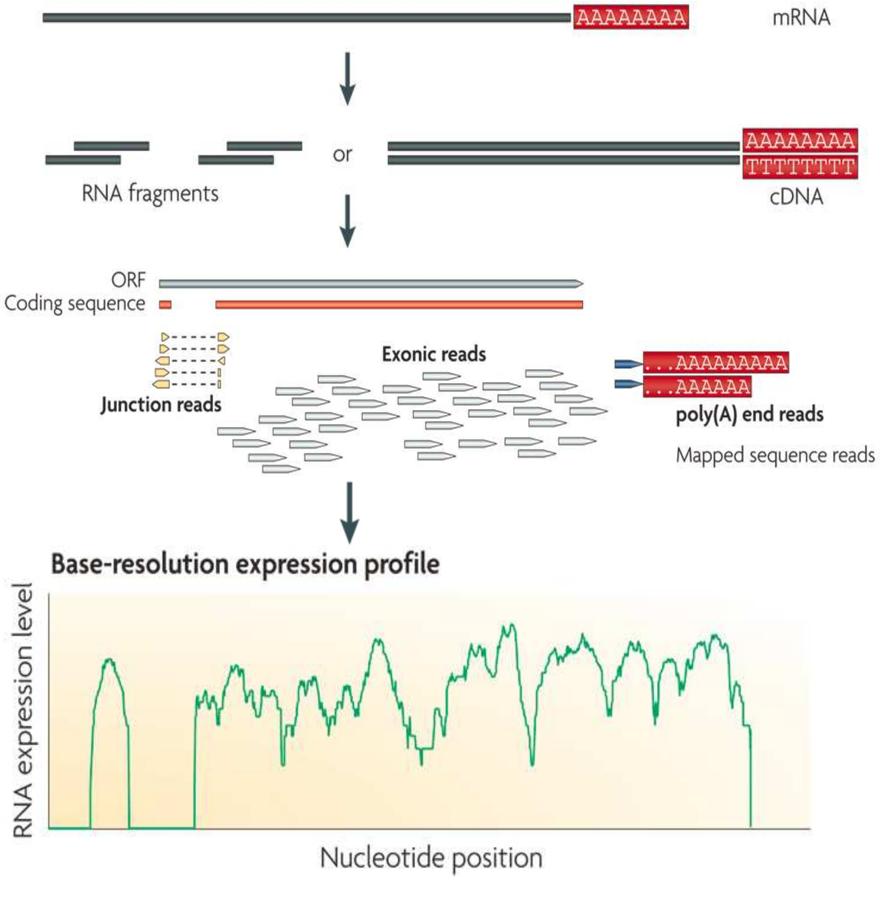


- Logistic regression
- Random forests
- Support vector machine

$$M(u,\mathbf{X},\mathbf{Z}) = \underbrace{\mathbf{X}^{T}\mathbf{a}(U)}_{\text{Non-parametric component}} + \underbrace{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}}_{\text{Parametric component}} + \varepsilon$$

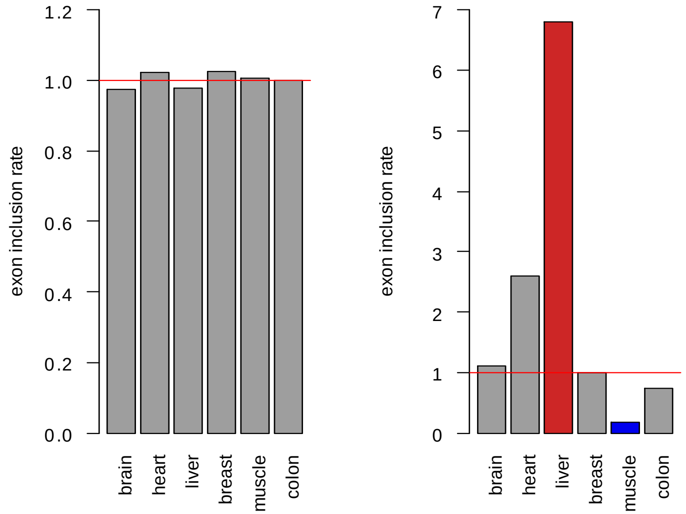
Data Preparation

- **Model implementation** : tissue specific exon inclusion rate calculation



$$e_{\underset{\downarrow\,\text{exon}}{i},j} = \frac{n^{e}_{i,j}}{n^{g}_{i,j}}, \qquad y_{i,\underset{\downarrow\,\text{tissue}}{j}} = \frac{e_{i,j}}{\frac{1}{m}\sum_{j} e_{i,j}}$$

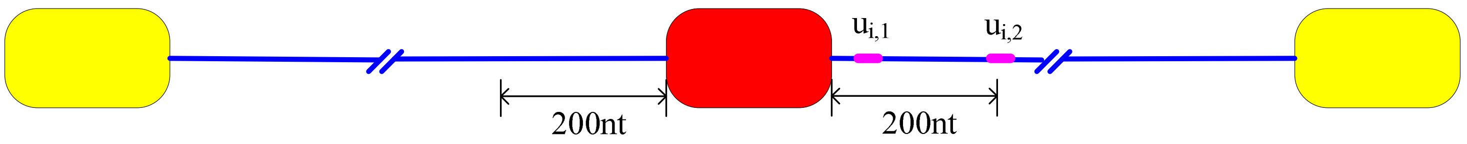
Goal: select exons with tissue difference



Wang Z, Gerstein M, Snyder M., "RNA-Seq: a revolutionary tool for transcriptomics", Nat Rev Genet. 2009 Jan;10(1):57-63. doi: 10.1038/nrg2484

Model Implementation

- **<u>Model implementation :</u>** Semi-parametric varying coefficient model



$$y_i = \sum_k a\left(u_{i,k}\right) + \beta + \varepsilon_i$$

Baseline score: Phylop Conservation score: [-1,1]

LS estimation:

$$\hat{\mathbf{a}}(u) = \left(\mathbf{I}_p, \mathbf{0}_P\right)\left(\boldsymbol{\Gamma}_\mathbf{u}^\mathbf{T}\mathbf{W}_\mathbf{u}\boldsymbol{\Gamma}_\mathbf{u}\right)^{-1}\boldsymbol{\Gamma}_\mathbf{u}^\mathbf{T}\mathbf{W}_\mathbf{u}\left(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}\right)$$

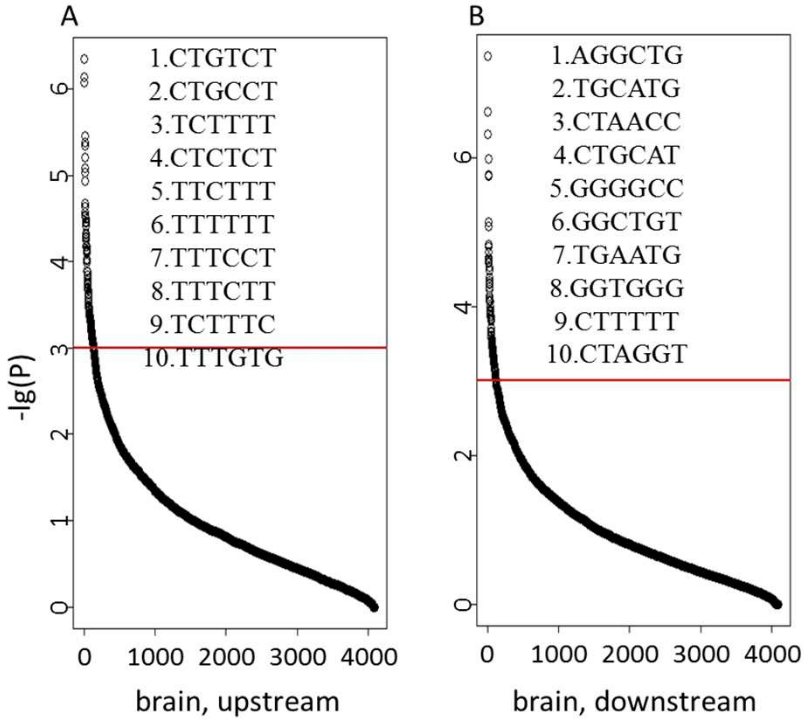
$$\hat{\boldsymbol{\beta}} = \left(\mathbf{Z}^\mathbf{T}(\mathbf{I}-\mathbf{S})^\mathbf{T}(\mathbf{I}-\mathbf{S})\mathbf{Z}\right)^\mathbf{T}\mathbf{Z}^\mathbf{T}(\mathbf{I}-\mathbf{S})^\mathbf{T}(\mathbf{I}-\mathbf{S})\mathbf{Y}$$

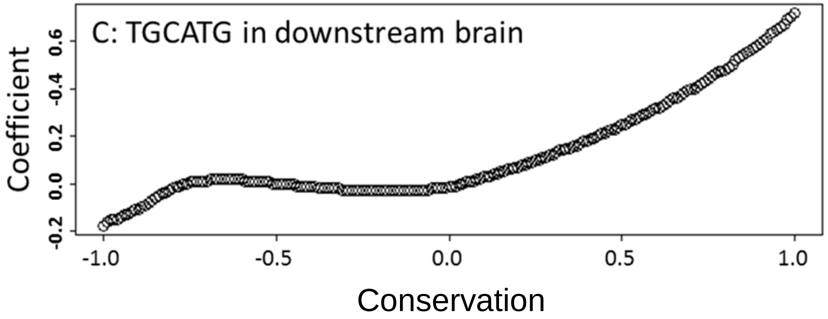
Bandwidth selection: Fivefold cross validation

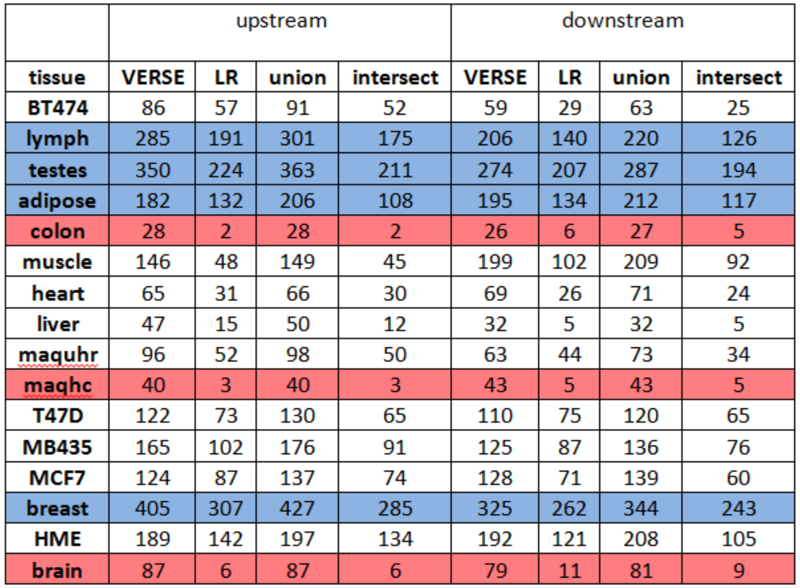
J. Zhang, C.C. Kuo, and L. Chen*, VERSE: A Varying Effect Regression for Splicing Elements Discovery.
Journal of Computational Biology, 2012.

Results

- **Individual SRE results in 16 human tissues**



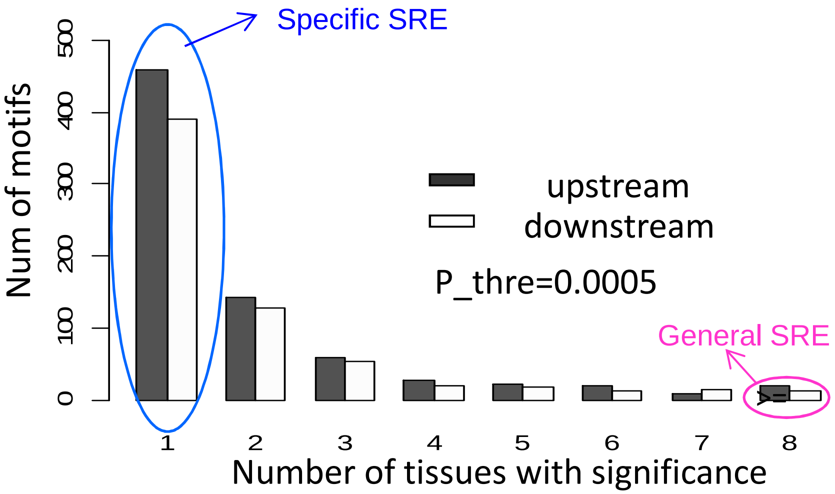




	upstream				downstream			
tissue	VERSE	LR	union	intersect	VERSE	LR	union	intersect
BT474	86	57	91	52	59	29	63	25
lymph	285	191	301	175	206	140	220	126
testes	350	224	363	211	274	207	287	194
adipose	182	132	206	108	195	134	212	117
colon	28	2	28	2	26	6	27	5
muscle	146	48	149	45	199	102	209	92
heart	65	31	66	30	69	26	71	24
liver	47	15	50	12	32	5	32	5
maquhr	96	52	98	50	63	44	73	34
maqhc	40	3	40	3	43	5	43	5
T47D	122	73	130	65	110	75	120	65
MB435	165	102	176	91	125	87	136	76
MCF7	124	87	137	74	128	71	139	60
breast	405	307	427	285	325	262	344	243
HME	189	142	197	134	192	121	208	105
brain	87	6	87	6	79	11	81	9

Results (continued)

- **Individual SRE results in 16 human tissues**



Up to ***70%*** of motifs are tissue specific

	upstream			downstream		
tissue	specific	general	specificity	specific	general	specificity
BT474	22	21	1.048	21	7	3.000
lymph	84	34	2.471	46	28	1.643
testes	92	34	2.706	67	33	2.030
adipose	34	29	1.172	47	26	1.808
colon	16	3	5.333	6	6	1.000
muscle	33	24	1.375	55	24	2.292
heart	12	15	0.800	22	14	1.571
liver	18	4	4.500	11	3	3.667
maquhr	12	29	0.414	17	20	0.850
maqhc	11	2	5.500	17	9	1.889
T47D	33	28	1.179	21	26	0.808
MB435	37	34	1.088	30	26	1.154
MCF7	31	18	1.722	38	15	2.533
breast	113	36	3.139	95	33	2.879
HME	44	28	1.571	52	24	2.167
brain	40	4	10.00	40	4	10.00

Specific SRE: identified within in only 1 tissue

General SRE: identified in at least 8 tissues

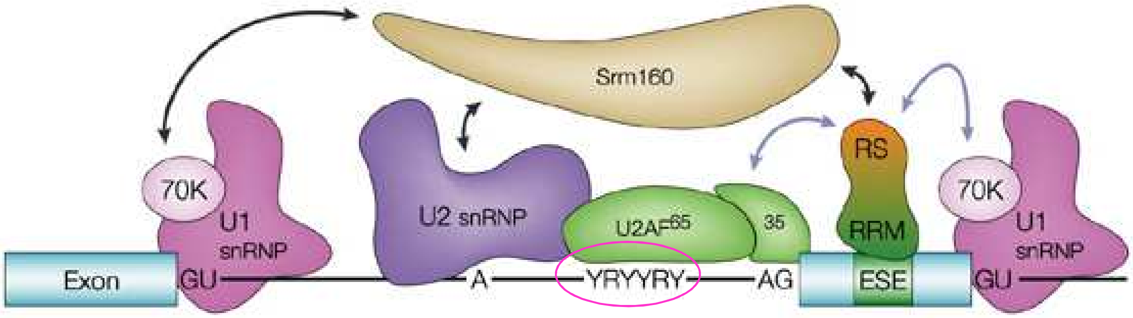
$$specificity = \frac{\#\text{specific SRE}}{\#\text{general SRE}}$$

Brain has the largest tissue specificity

MAQ-UHR has the lowest tissue specificity

Sites for the Same Splicing Factor?

- **Why clustering? : find out motifs bound to the same protein**



- Kmeans
- Hierarchical clustering
- …

Challenge: Degeneracy of motifs:

YRYYRY

Y=C or T R=A or G

TGCATG: 100

AGCATG:100

$$\begin{array}{c|cccccc} nt & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline A & 0.5 & 0 & 0 & 1 & 0 & 0 \\ G & 0 & 1 & 0 & 0 & 0 & 1 \\ C & 0 & 0 & 1 & 0 & 0 & 0 \\ T & 0.5 & 0 & 0 & 0 & 1 & 0 \end{array}$$

$$p_i(A) = \frac{n_i(A)}{n_i(A/G/C/T)}$$

Pseudo counts

$$n_p = \sqrt{\frac{1}{m}\sum_{i=1}^{m} n_i}$$

VERSE: short conclusion

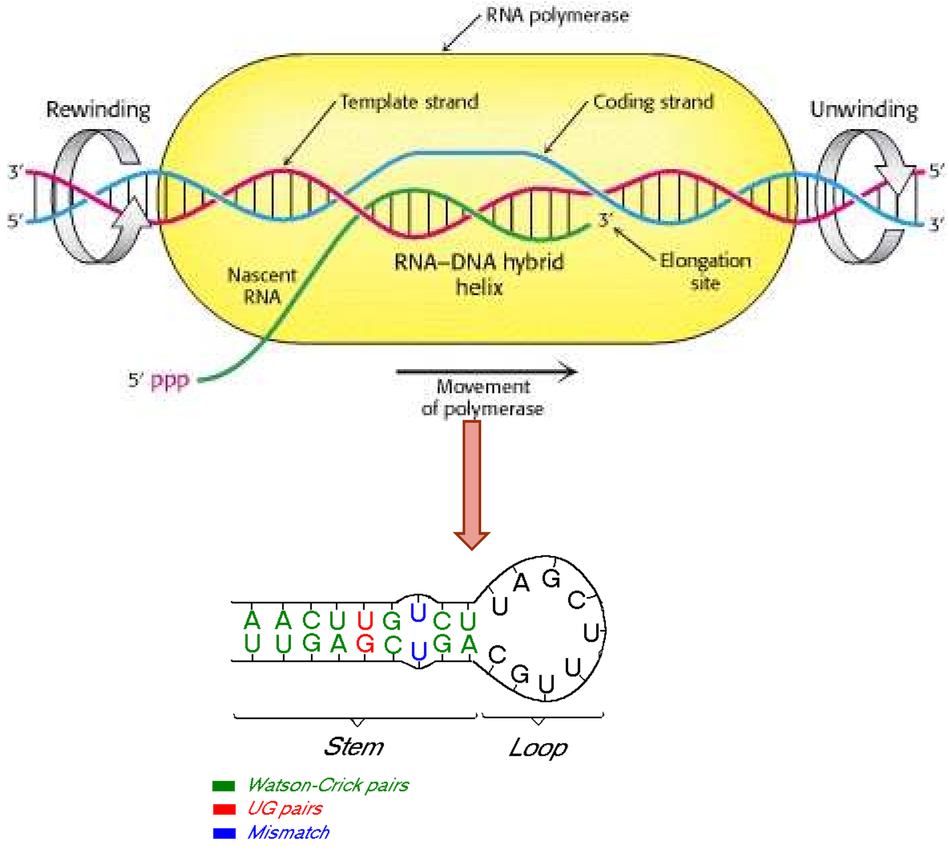
- **Characteristics of VERSE:**
  - SRE discovery by integrating multiple assisting information
  - Allows the contribution of SREs varying with different biological environments
  - A two stage clustering method to identify SREs bound by the same protein
- **Conclusion of discovered SREs**
  - Brain demonstrated unique pattern of splicing regulation at the context level
  - Conservation does take effect in some of the tissues

Outline

- Introduction — alternative splicing
- Part I — mRNA product quantification
  - Gene expression estimation with isoform resolution from RNA-Seq data (WemIQ)
- Part II — alternative splicing regulation
  - Part A — Context based regulation: motifs discovery via a varying coefficient regression
  - Part B — Structure based regulation: stability of mRNA secondary structures and splicing site selection
- Conclusion and future work

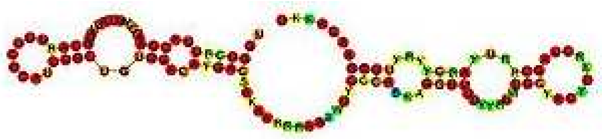
Splicing code part 3: pre-mRNA secondary structures

- **What is secondary structure?**

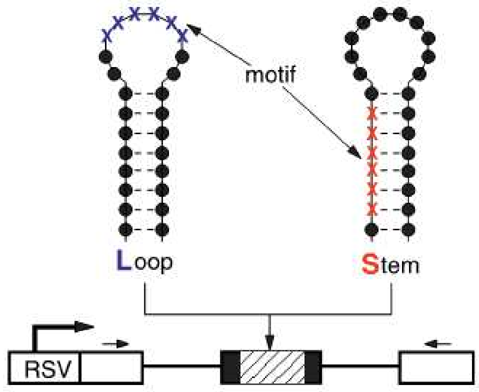


- **How to predict the topology?**

Software like RNAfold, Mfold, …



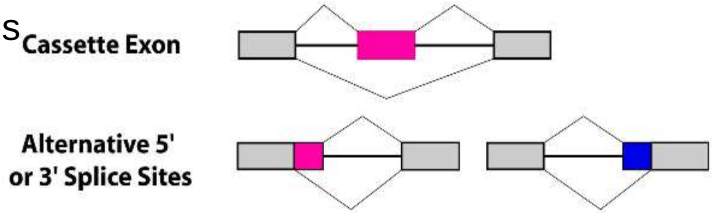
- **How does it work?**

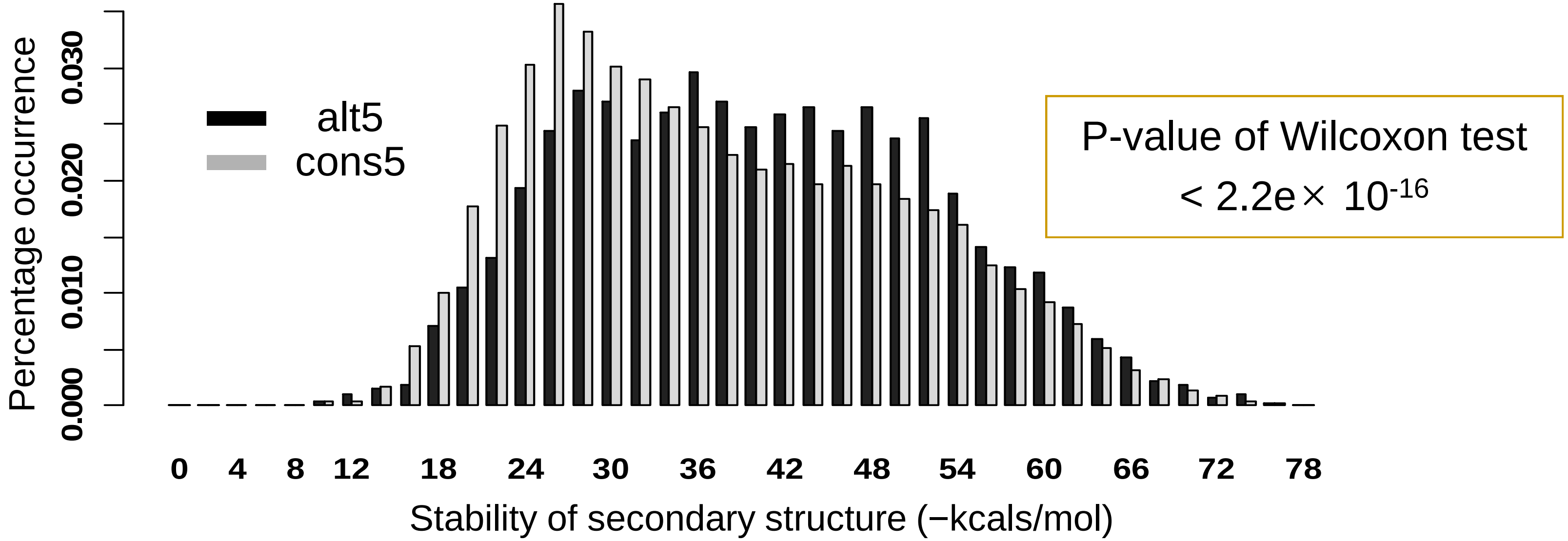


Structural stability different around splice sites

- **Alternative Spliced Sites Exhibit More Stable Structures**

- ***Data:*** UCSC hg18 for human, Eugene for mice and fruit flies
- ***Method***: RNAfold to calculate the energy
- ***Conclusion:***
  - ➢ alternatively spliced sites exhibit more stable structures
  - ➢This trend is conserved from human to mice and fruit flies





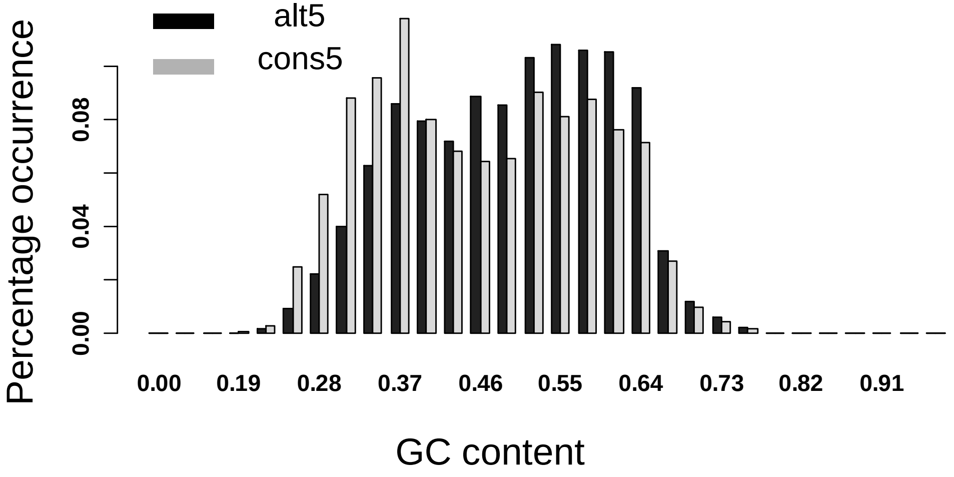
P-value of Wilcoxon test

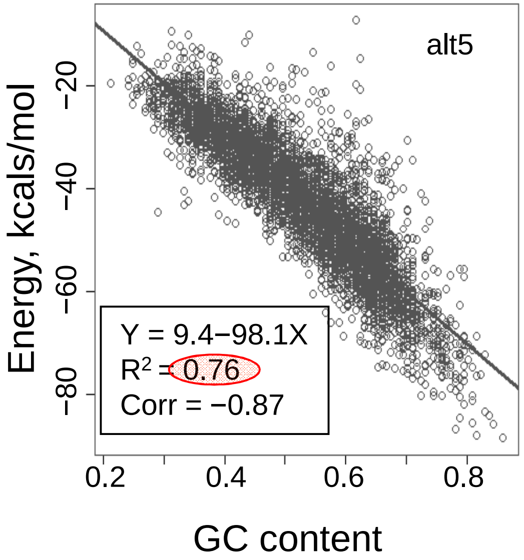
$< 2.2e\times 10^{-16}$

Zhang, J., C.C. Kuo, and L. Chen*, GC content around splice sites affects splicing through pre-mRNA secondary structures. BMC Genomics, 2011. 12: p. 90.

Explanations for Structure Difference

- **GC Content Explains the Structure Stability Difference**





Observation:

- Nearly perfect correlation between GC content and structure energy
- Regression shows similar results among all exon categories and all species

Question:

- Is GC content the only source for the structure stability difference?

Explanations for Structure Difference (continued)

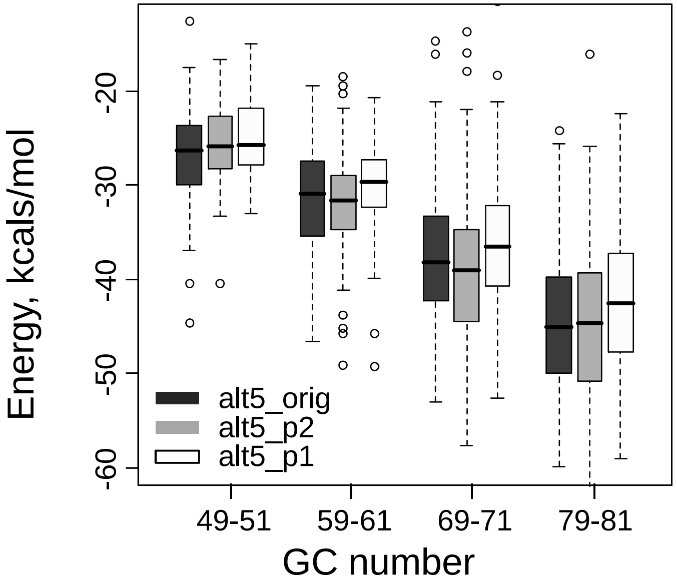
- **Neutral Selection Pressure on Nucleotide Order Effect**

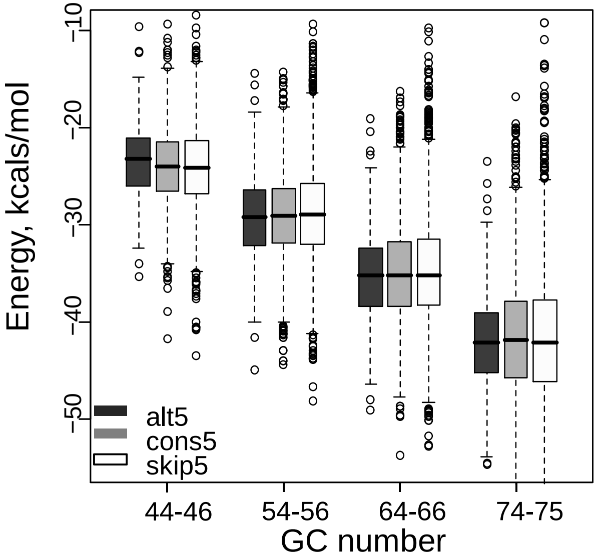
Question: factors affect stability

1. GC content – stable combination per pair
2. Context – selection to keep a thermal favorable nucleotide order

Method: Permutation study

1. Keep 1st order nucleotide frequency (p1)
2. Keep 2nd order nucleotide frequency (p2)





Method: Stability study

1. Compare different groups with similar GC

Result: GC content effect is more significant

1. Fix GC, energy is similar among groups
2. Native sequence shows similar stability with control

GC selection near Exon Junctions

- **Real Sites vs. Decoy Sites: Structure Stability is Different**

Method:

- structure and GC difference of spliced (real) and non-spliced sites (decoy)

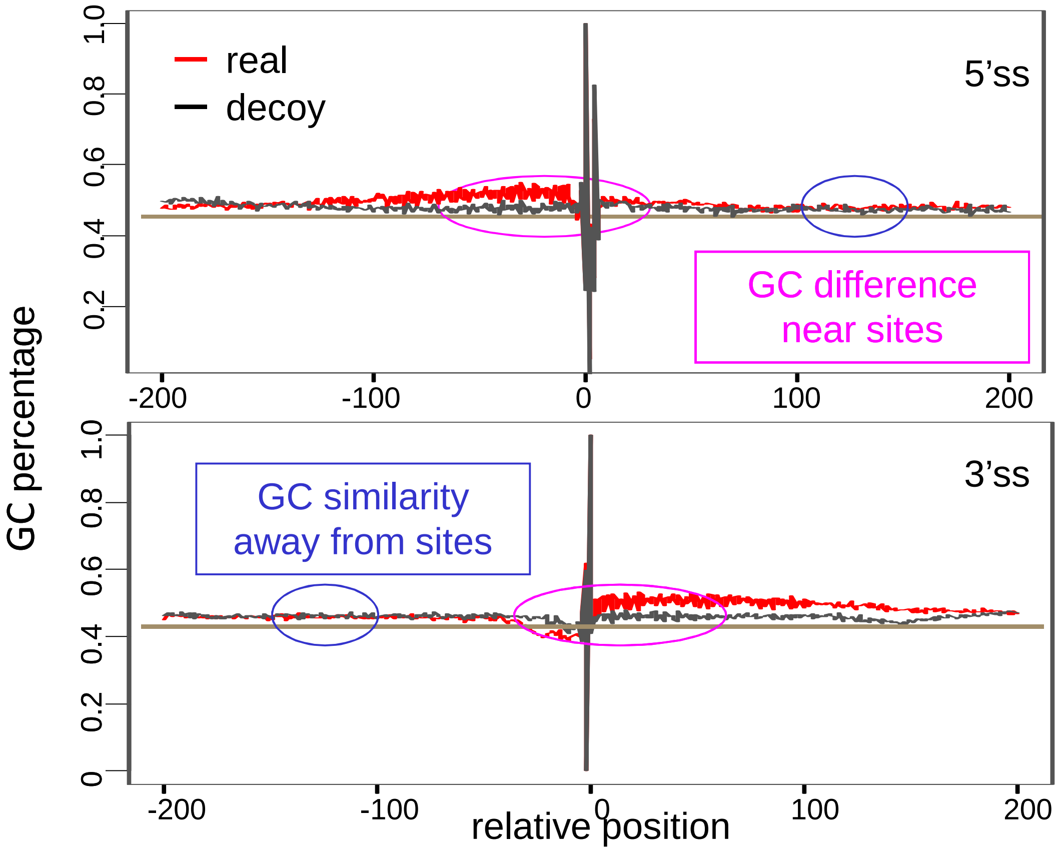
Observation:

•GC correlates perfectly with the mRNA structure stability in all sites

•GC enrichment in real sites near the consensus sequences

•Similar GC percentage far away from consensus sequence

GC difference explains the more stable structures near the real splice sites



Outline

- Introduction — alternative splicing
- Part I — mRNA product quantification
  - Gene expression estimation with isoform resolution from RNA-Seq data (WemIQ)
- Part II — alternative splicing regulation
  - Part A — Context based regulation: motifs discovery via a varying coefficient regression
  - Part B — Structure based regulation: stability of mRNA secondary structures and splicing site selection
- Conclusion and future work

Conclusion and ongoing projects

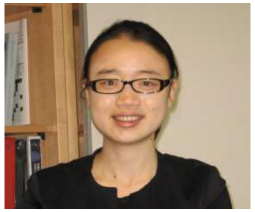
- Quantification of mRNA product at isoform level
  - Weighted EM with bias removal through GP model(WemIQ)
- Alternative splicing regulation
  - Context based: motif discovery and clustering (VERSE)
  - Structure based: structural difference around splice sites
- Ongoing projects
  - GWAS studies on Parkinson’s and Alzheimer's Disease to discover SNPS with aging effect through varying coefficient model
  - regulatory elements discovery by integrating multiple features

Acknowledgement

Liang Chen's Laboratory

Media Communications Lab

Ming Hsieh Department of Electrical Engineering



- Dr. Liang Chen and Dr. Jay Kuo
- Dr. Fengzhu Sun and Dr. Andrew Smith

- Sudeep Srivas
- Cheyu Lee