Computational Systems Biology
Deep Learning in the Life Sciences

6.802 6.874 20.390 20.490 HST.506

David Gifford

Lecture 7

February 27, 2019

The transcriptome and differential expression

http://mit6874.github.io

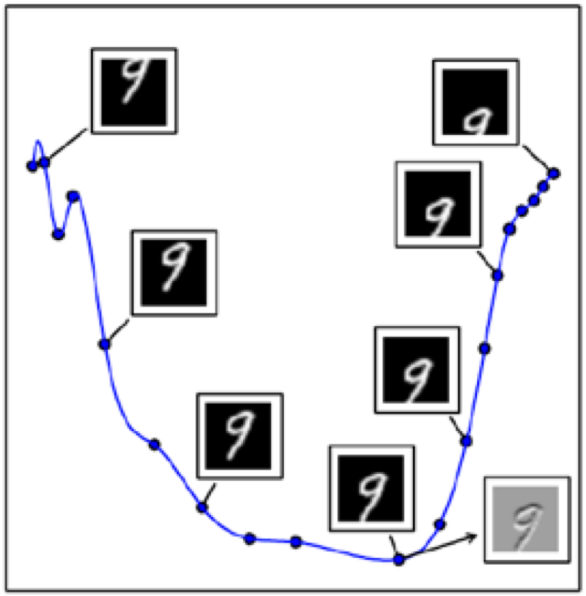
What's on tap today!

- Recap of manifolds, KL divergence, t-SNE gradients
- The transcriptome
  - Exon splicing and isoform expression
- Differential expression detection
  - Embedded models and significance testing
  - Multiple hypothesis correction
  - Gene set enrichment analysis
- Exon splicing code

1. Manifolds, KL Divergence, KL gradients

What is a manifold mapping?

Neighborhoods in high dimensional space are preserved in low dimensional space



KL Divergence is always positive

$$-\sum_{i=1}^{n} p_i \log_2 p_i \leq -\sum_{i=1}^{n} p_i \log_2 q_i$$

Gibbs Inequality

$$D_{\mathrm{KL}}(P\|Q) \equiv \sum_{i=1}^{n} p_i \log_2 \frac{p_i}{q_i} \geq 0.$$

We can use gradient methods to find an embedding

- Cost function

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$$

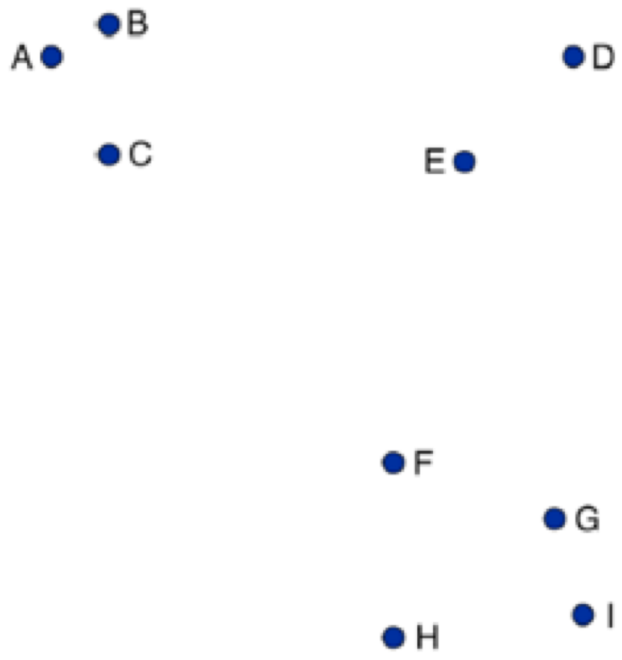
  - Large p_{ij} modeled by small q_{ij}: Large penalty
  - Small p_{ij} modeled by large q_{ij}: Small penalty
  - t-SNE mainly preserves local similarity structure of the data

- Gradient

$$\frac{\partial C}{\partial y_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(1 + ||y_i - y_j||^2)^{-1}(y_i - y_j)$$

The overall gradient on y_i is the sum of gradients from all other points

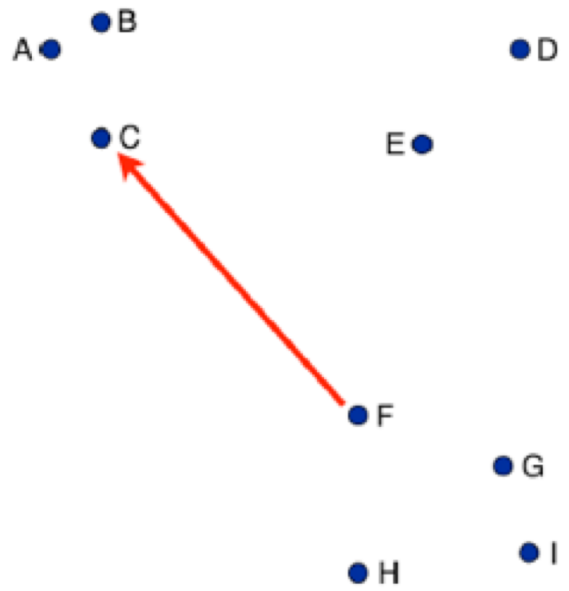
$$\frac{\partial C}{\partial y_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(1 + ||y_i - y_j||^2)^{-1}(y_i - y_j)$$



Gradient between two points is proportional to their displacement

- Displacement

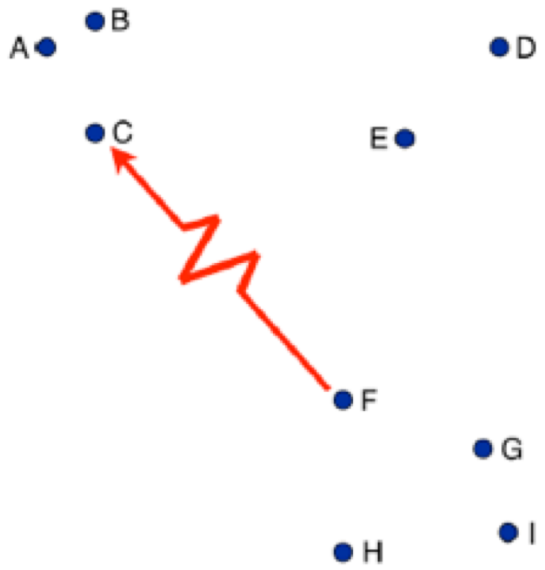
$$(y_i - y_j)$$



We can interpret a pair-wise gradient as a spring

- Exertion / Compression

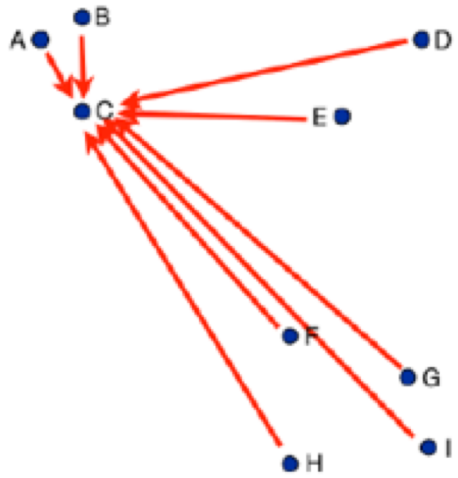
$$(p_{ij} - q_{ij})(1 + ||y_i - y_j||^2)^{-1}$$



We sum all of the gradients for a given point to update its location

- N-Body, summation

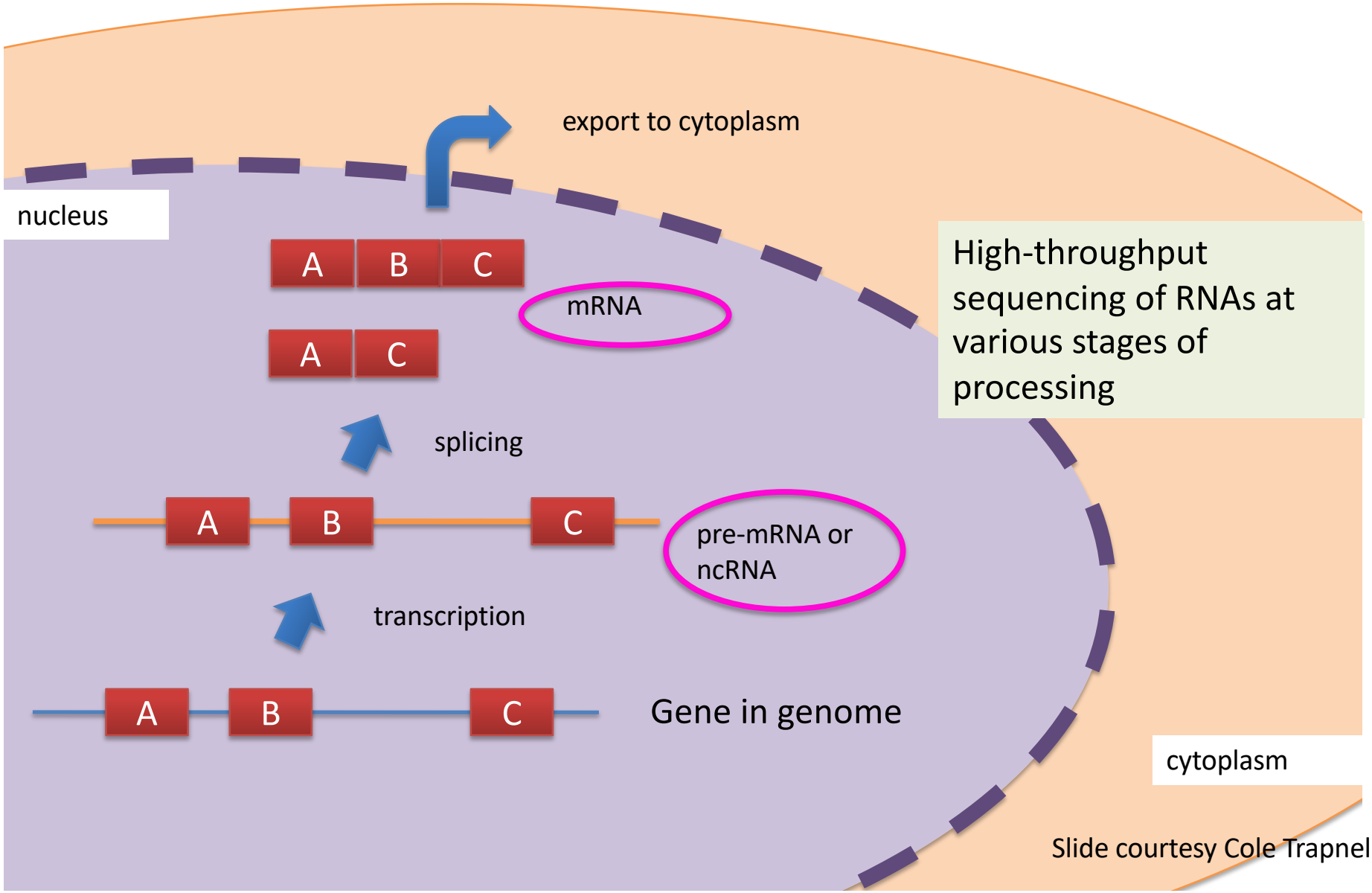
$$\frac{\partial C}{\partial y_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(1 + ||y_i - y_j||^2)^{-1}(y_i - y_j)$$



Reduce Complexity from $O(N^2)$ to $O(N \log N)$ via Barnes Hut (tree-based) algorithm

2. RNA-seq data has ~3,000 – 20,000 gene expression levels per sample

RNA-Seq characterizes RNA molecules

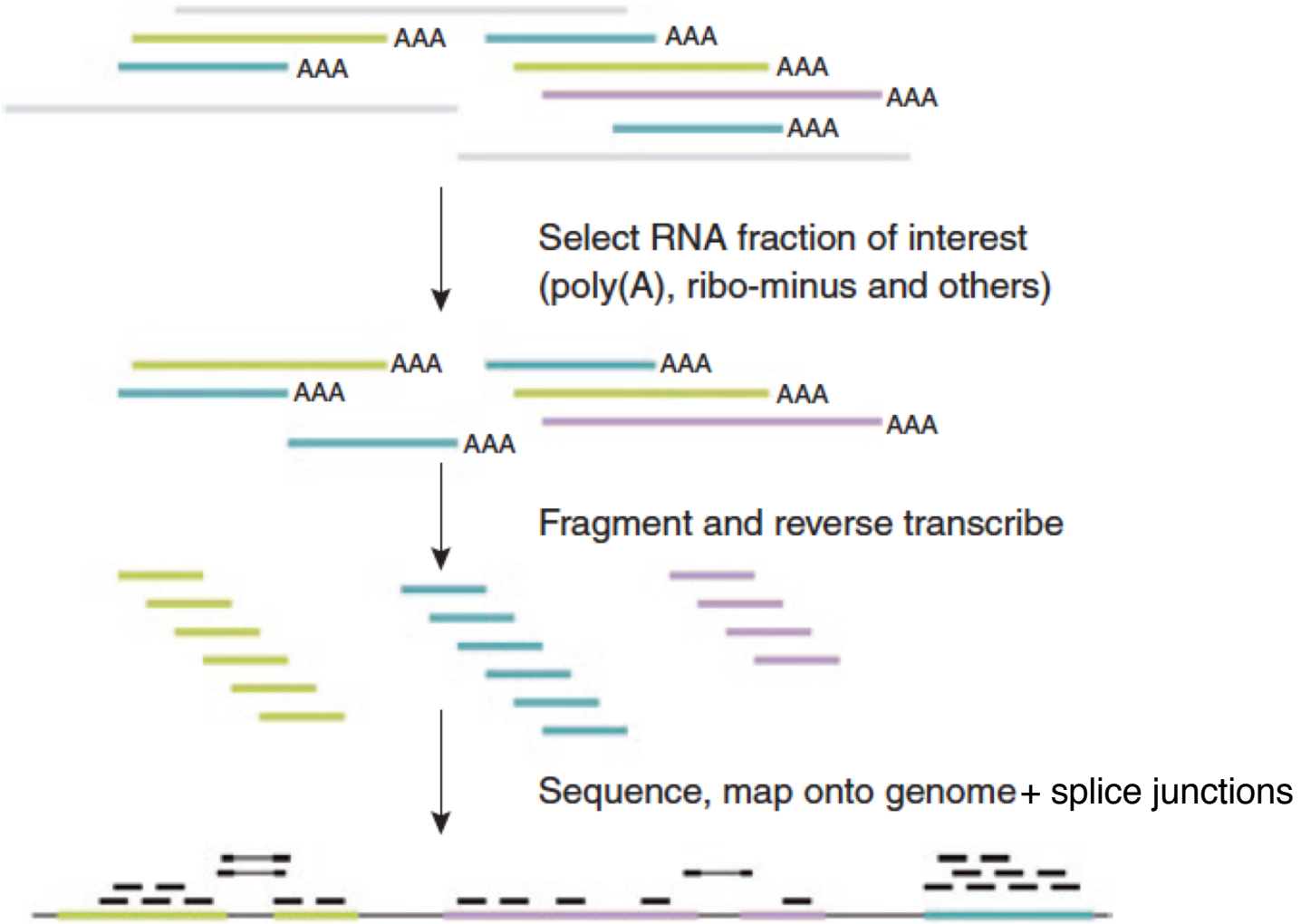


High-throughput sequencing of RNAs at various stages of processing

Slide courtesy Cole Trapnell

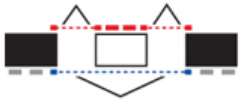
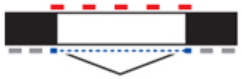
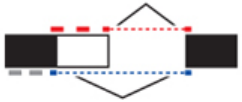
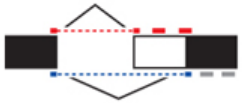
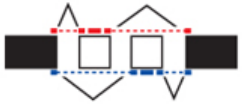
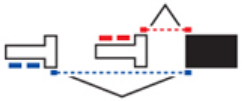
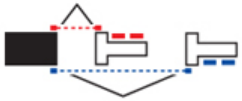
RNA-Seq: millions of short reads from fragmented mRNAs

Extract RNA from cells/tissue



Pepke et. al. *Nature Methods* 2009

Pervasive tissue-specific regulation of alternative mRNA isoforms.

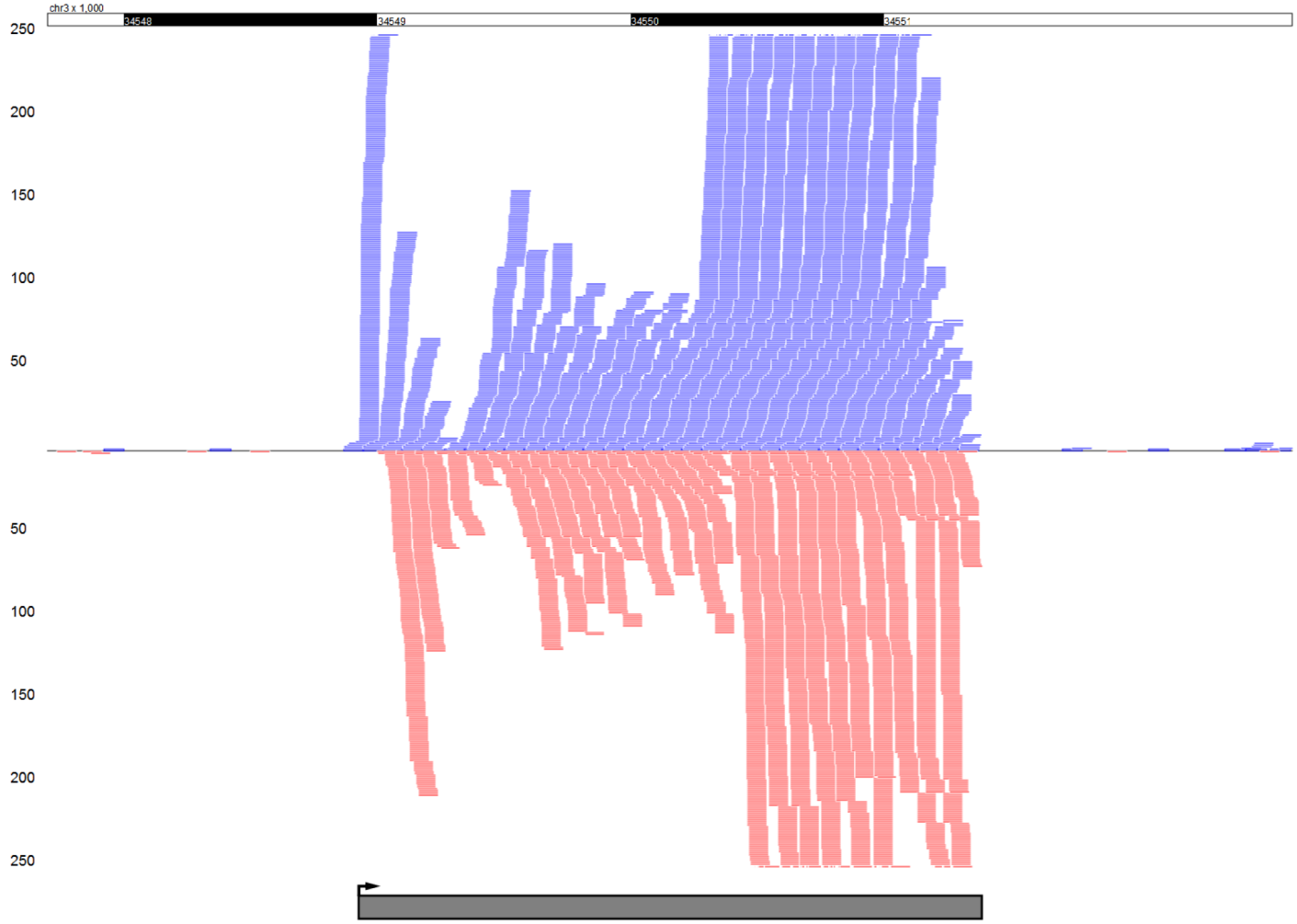
Alternative transcript events		Total events ($\times 10^3$)	Number detected ($\times 10^3$)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Skipped exon		37	35	10,436	6,822	65	72
Retained intron		1	1	167	96	57	71
Alternative 5′ splice site (A5SS)		15	15	2,168	1,386	64	72
Alternative 3′ splice site (A3SS)		17	16	4,181	2,655	64	74
Mutually exclusive exon (MXE)		4	4	167	95	57	66
Alternative first exon (AFE)		14	13	10,281	5,311	52	63
Alternative last exon (ALE)		9	8	5,246	2,491	47	52
Tandem 3′ UTRs		7	7	5,136	3,801	74	80
Total		105	100	37,782	22,657	60	68

Constitutive exon or region — Body read — Junction read — pA Polyadenylation site

Alternative exon or extension — Inclusive/extended isoform — Exclusive isoform — Both isoforms

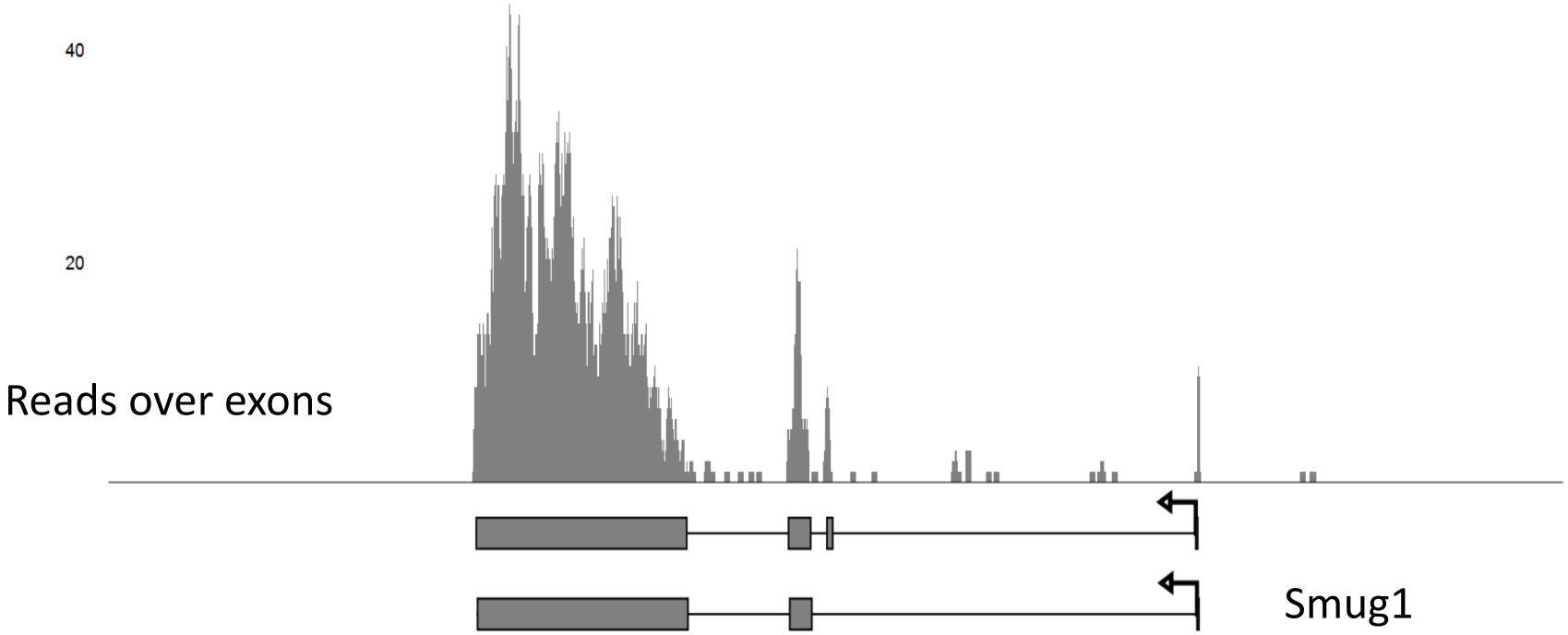
ET Wang *et al. Nature* **000**, 1-7 (2008)
doi:10.1038/nature07509

One measure of expression is Reads Per Kilobase of gene per Million reads (RPKM)

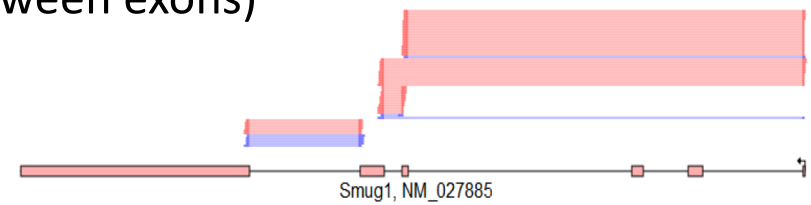


Sox2

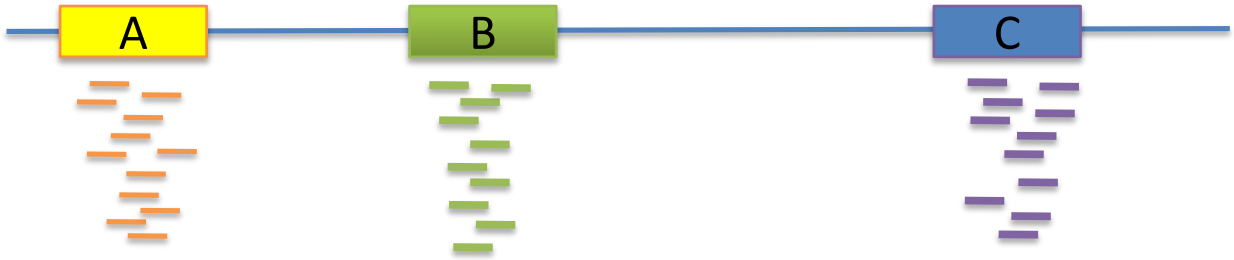
RNA-seq reads map to exons and across exons



Junction reads (split between exons)

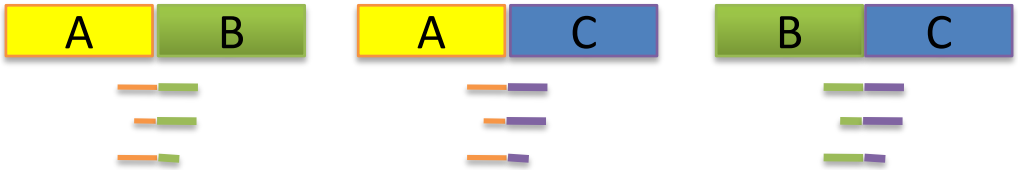


Aligned reads reveal isoform possibilities



identify candidate exons via genomic mapping

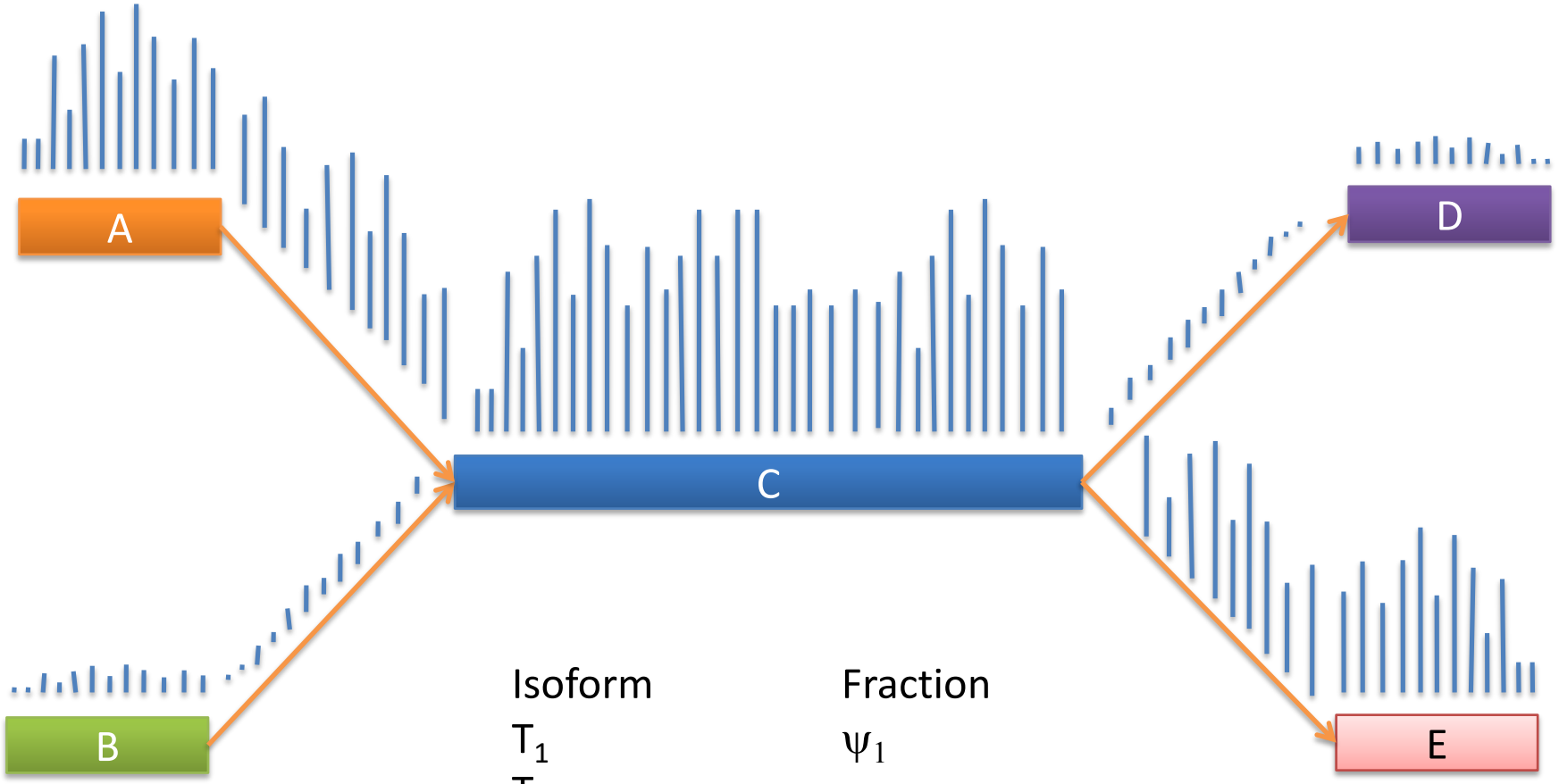
Generate possible pairings of exons



Align reads to possible junctions

Slide courtesy Cole Trapnell

We can use mapped reads to learn the isoform mixture ψ



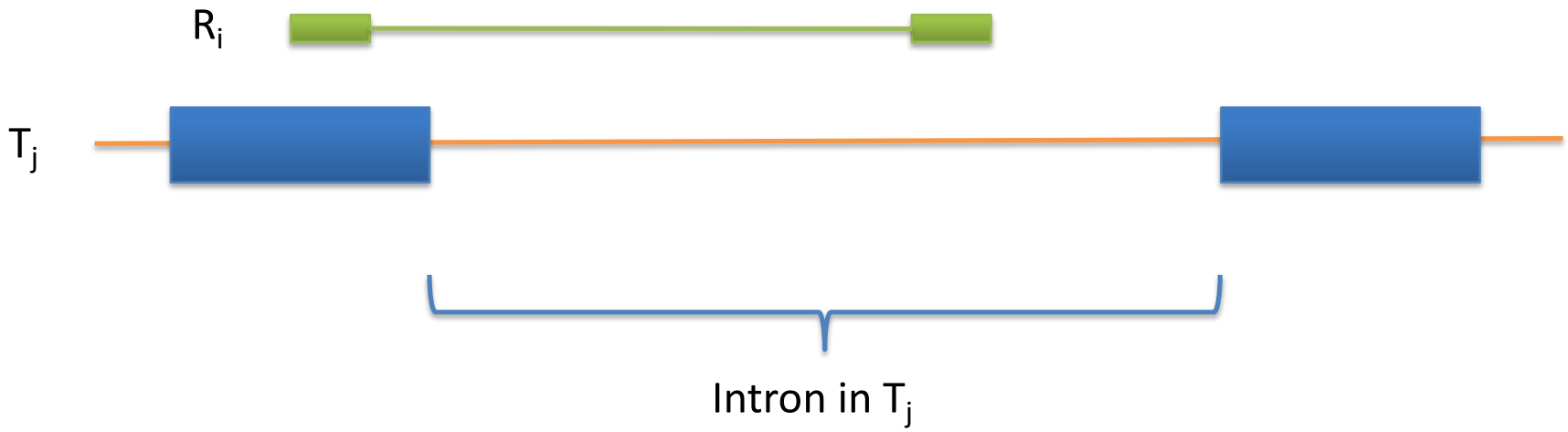
Isoform	Fraction
T_1	ψ_1
T_2	ψ_2
T_3	ψ_3
T_4	ψ_4

Slide courtesy Cole Trapnell

$P(R_i \mid T=T_j)$ – Excluded reads

If a read pair R_i is structurally incompatible with transcript T_j, then

$$P(R = R_i \mid T = T_j) = 0$$



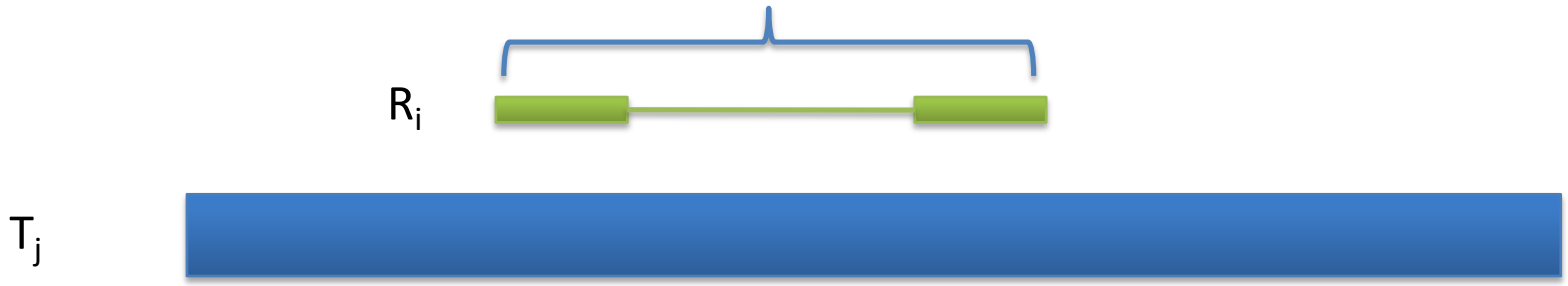
Slide courtesy Cole Trapnell

$P(R_i \mid T=T_j)$ – Paired end reads

Assume our library fragments have a length distribution described by a probability density F. Thus, the probability of observing a particular paired alignment to a transcript:

$$P(R = R_i \mid T = T_j) = \frac{F(l_j(R_j))}{l_j}$$

Implied fragment length $l_j(R_i)$



Slide courtesy Cole Trapnell

Estimating Isoform Expression

- Find expression abundances $\psi_1,\ldots,\psi_n$ for a set of isoforms $T_1,\ldots,T_n$
- Observations are the set of reads $R_1,\ldots,R_m$

$$P(R \mid \Psi) = \prod_{i=0}^{m} \sum_{j=0}^{n} \Psi_j P(R = R_i \mid T = T_j)$$

$$L(\Psi \mid R) \propto P(R \mid \Psi) P(\Psi)$$

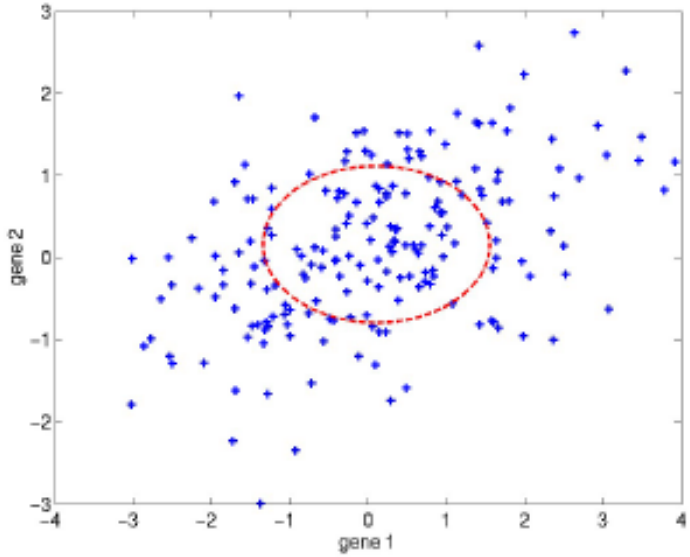
$$\Psi = \underset{\Psi}{\operatorname{argmax}} L(\Psi \mid R)$$

- Can estimate mRNA expression of each isoform using total number of reads that map to a gene and ψ

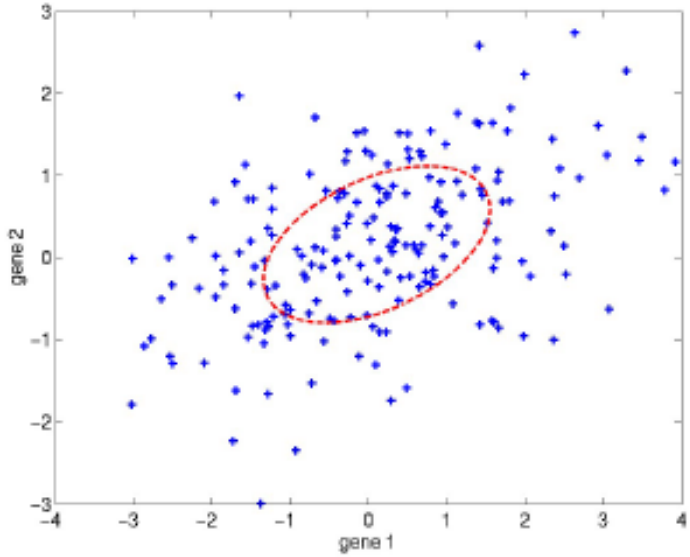
3. The significance of differential expression

Statistical tests: example

- The alternative hypothesis H_1 is more expressive in terms of explaining the observed data



null hypothesis



alternative hypothesis

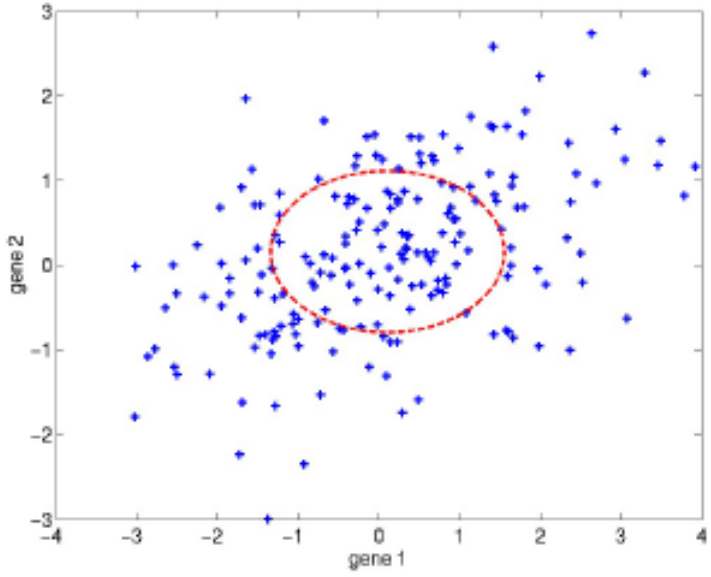
- We need to find a way of testing whether this difference is significant

Degrees of freedom

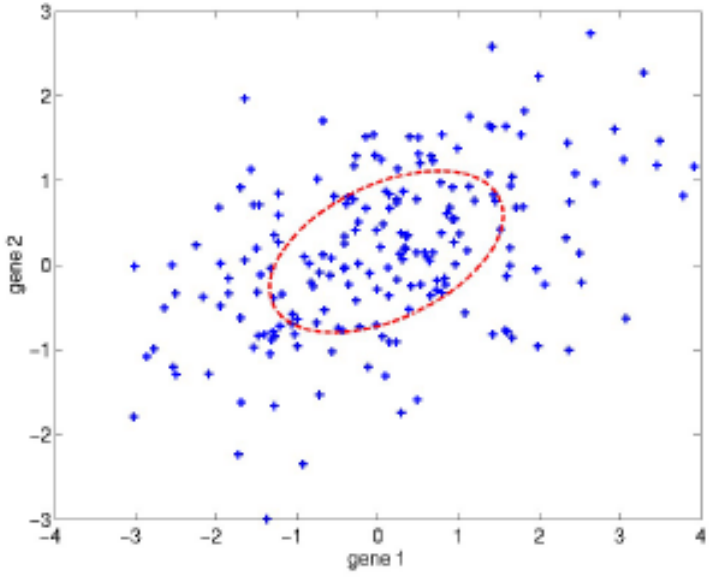
- How many degrees of freedom do we have in the two models?

$$H_0: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right)$$

$$H_1: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$



H_0



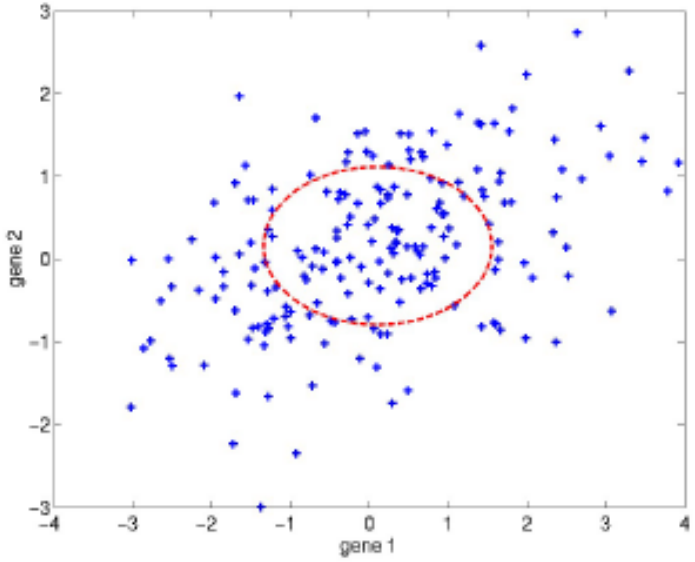
H_1

Degrees of freedom

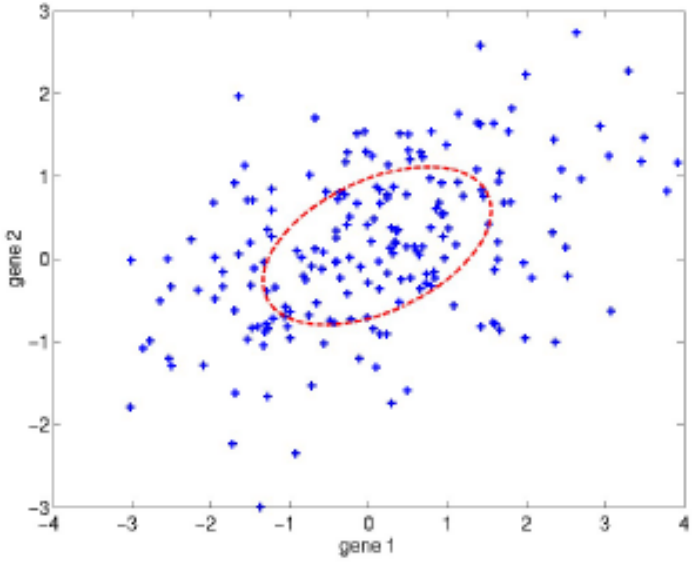
- How many degrees of freedom do we have in the two models?

$$H_0: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \right)$$

$$H_1: \quad \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$



H_0



H_1

- The observed data overwhelmingly supports H_1

Test statistic

- Likelihood ratio statistic

$$T(X^{(1)}, \ldots, X^{(n)}) = 2 \log \frac{P(X^{(1)}, \ldots, X^{(n)} | \hat{H}_1)}{P(X^{(1)}, \ldots, X^{(n)} | \hat{H}_0)} \qquad (1)$$

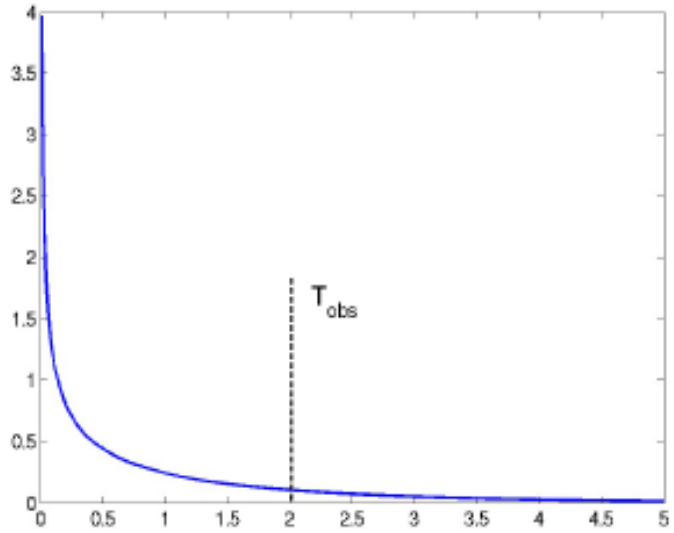
  Larger values of T imply that the model corresponding to the null hypothesis H_0 is much less able to account for the observed data

- To evaluate the P-value, we also need to know the sampling distribution for the test statistic

  In other words, we need to know how the test statistic $T(X^{(1)}, \ldots, X^{(n)})$ varies if the null hypothesis H_0 is correct

Test statistic cont'd

- For the likelihood ratio statistic, the sampling distribution is χ^2 with degrees of freedom equal to the difference in the number of free parameters in the two hypotheses



- Once we know the sampling distribution, we can compute the P-value

$$p = Prob(\, T(X^{(1)}, \ldots, X^{(n)}) \geq T_{obs} \,|\, H_0 \,) \tag{2}$$

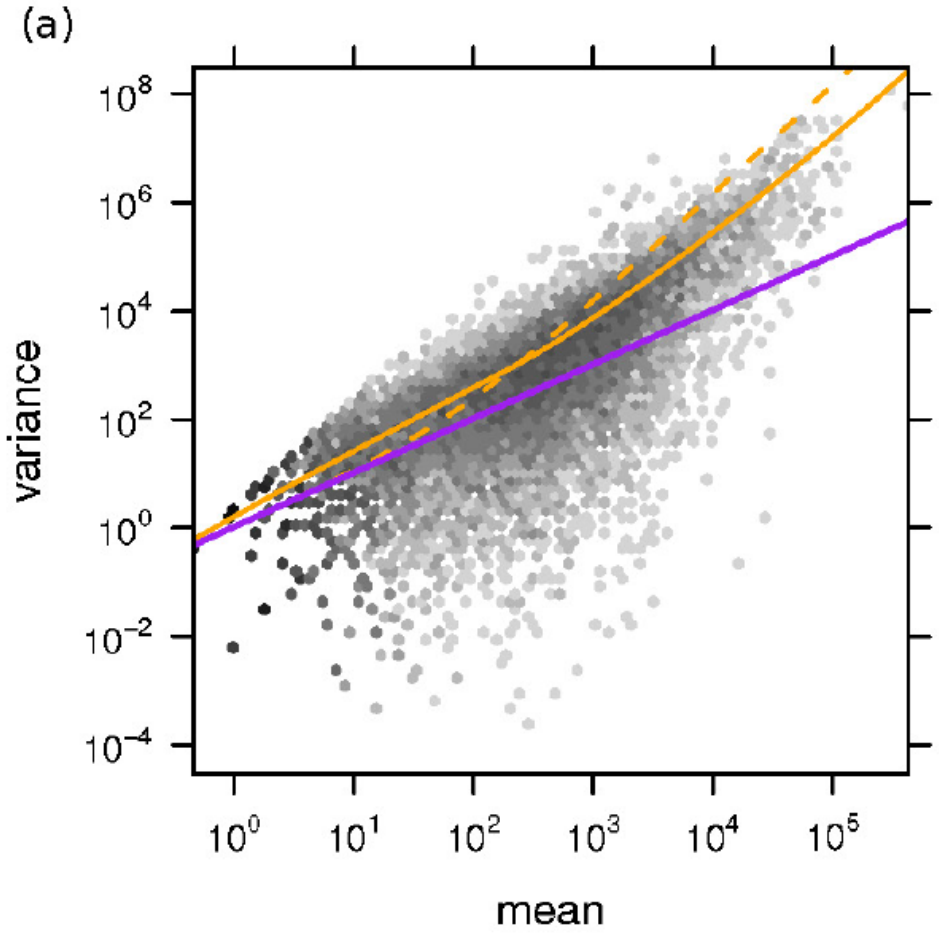
What is the right distribution for modeling read counts?

$$\lambda = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$f(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Poission?

Read count data is overdispersed for a Poission
Use a Negative Binomial instead



Orange Line – DESeq
Dashed Orange – edgeR
Purple - Poission

$$\sigma_{ij}^2 = \mu_{ij} + s_j^2 v_p\left(q_{ip(j)}\right)$$

Anders and Huber *Genome Biology* 2010, **11**:R106
http://genomebiology.com/2010/11/10/R106

A Negative Binomial distribution is better (DESeq)

- i gene or isoform p condition
- j sample (experiment) p(j) condition of sample j
- m number of samples
- K_{ij} number of counts for isoform i in experiment j
- q_{ip} Average scaled expression for gene i condition p

$$q_{ip} = \frac{1}{\#\text{ of replicates}} \sum_{\text{j in replicates}} \frac{K_{ij}}{s_j}$$

$$\mu_{ij} = q_{ip(j)} s_j \qquad \sigma_{ij}^2 = \mu_{ij} + s_j^2 v_p\left(q_{ip(j)}\right)$$

$$K_{ij} \sim NB\left(\mu_{ij}, \sigma_{ij}^2\right)$$

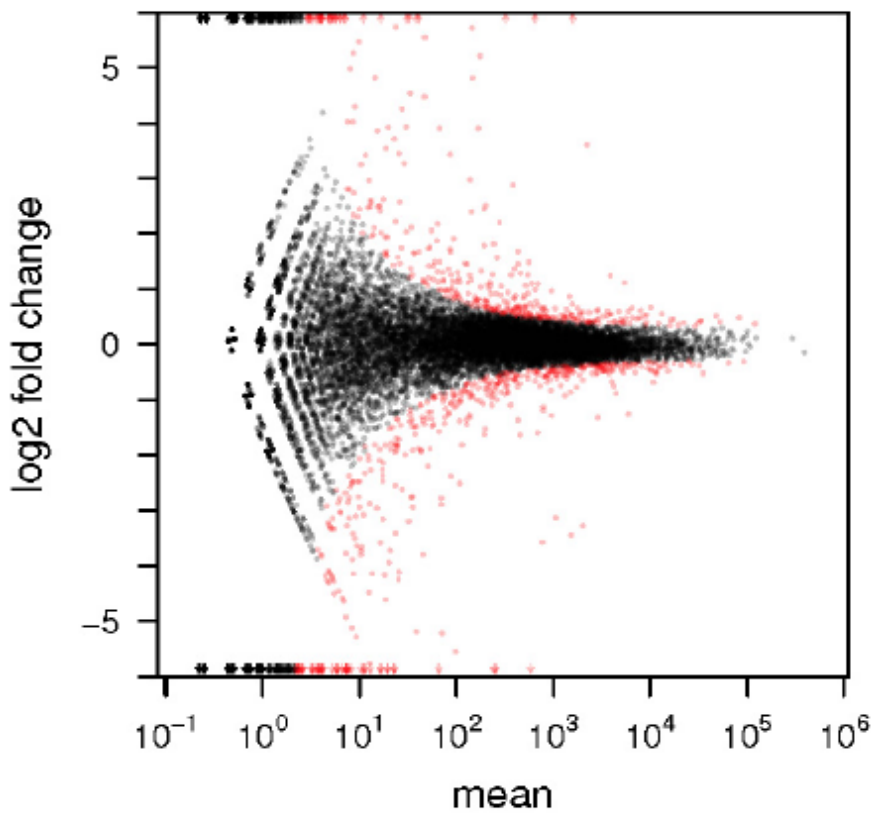


Figure 3 Testing for differential expression between conditions *A* and *B*: Scatter plot of $\log_2$ ratio (fold change) versus mean. The red colour marks genes detected as differentially expressed at 10% false discovery rate when Benjamini-Hochberg multiple testing adjustment is used. The symbols at the upper and lower plot border indicate genes with very large or infinite log fold change. The corresponding volcano plot is shown in Supplementary Figure S8 in Additional file 2.

Hypergeometric test for gene set overlap significance

N – total # of genes	1000
n1 - # of genes in set A	20
n2 - # of genes in set B	30
k - # of genes in both A and B	3

$$P(k) = \frac{\binom{n1}{k}\binom{N-n1}{n2-k}}{\binom{N}{n2}}$$

0.017

$$P(x \geq k) = \sum_{i=k}^{\min(n1,n2)} P(i)$$

0.020

Bonferroni correction

- Total number of rejections of null hypothesis over all N tests denoted by R.

$$\Pr(R>0) \tilde{=} N\alpha$$

- Need to set α' = Pr(R>0) to required significance level **over all tests**. Referred to as the **experimentwise error rate**.
- With 100 tests, to achieve overall experimentwise significance level of α'=0.05:

$$0.05 = 100\alpha$$

$$\rightarrow \alpha = 0.0005$$

- **Pointwise** significance level of 0.05%.

Example - Genome wide association screens

- Risch & Merikangas (1996).
- 100,000 genes.
- Observe 10 SNPs in each gene.
- 1 million tests of null hypothesis of no association.
- To achieve experimentwise significance level of 5%, require pointwise p-value less than 5×10^{-8}

Bonferroni correction - problems

- Assumes each test of the null hypothesis to be **independent**.
- If not true, Bonferroni correction to significance level is **conservative**.
- Loss of power to reject null hypothesis.
- Example: genome-wide association screen across linked SNPs – correlation between tests due to LD between loci.

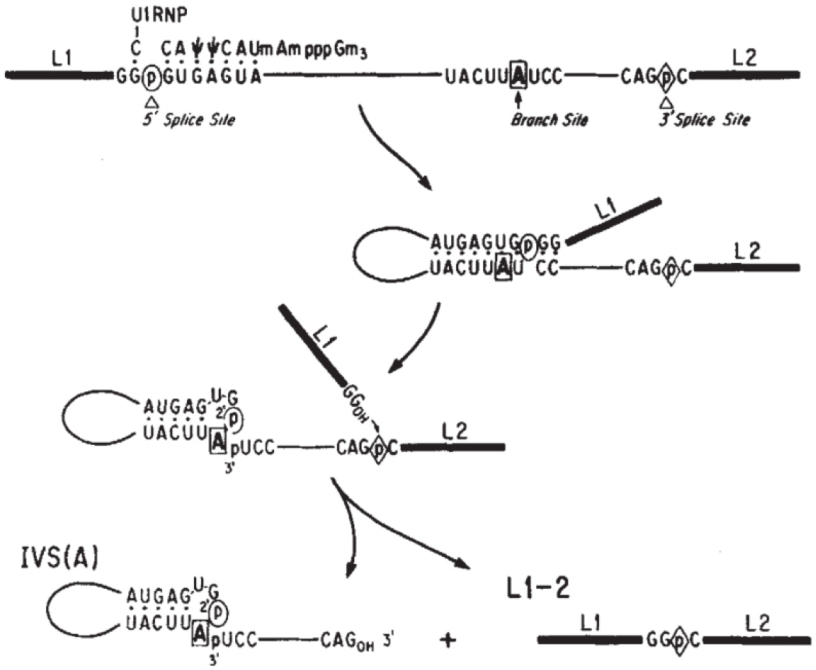
Benjamini Hochberg

- Select False Discovery Rate α
- Number of tests is *m*
- Sort p-values $P_{(k)}$ in ascending order (most significant first)
- Assumes tests are uncorrelated or positively correlated

1. For a given α, find the largest k such that $P_{(k)} \leq \frac{k}{m}\alpha$.
2. Reject the null hypothesis (i.e., declare discoveries) for all $H_{(i)}$ for $i = 1, \ldots, k$.

4. How can we predict splice isoforms from sequence?

RNA SPLICING



[Konarska, Nature, (1985)]

The spliceosome, catalyzed by small nuclear ribonucleoproteins (snRNPs) binds the 5' splice site, facilitating 5' intron base pairing with the downstream branch sequence, forming a lariat.

The 3' end of the exon is cut and joined to the branch site by a hydroxyl (OH) group at the 3' end of the exon that attacks the phosphodiester bond at the 3' splice site. The exons are covalently bound, and the lariat containing the intron is released.

RNA SPLICING:

MACHINE LEARNING HISTORY AND STATE OF THE ART

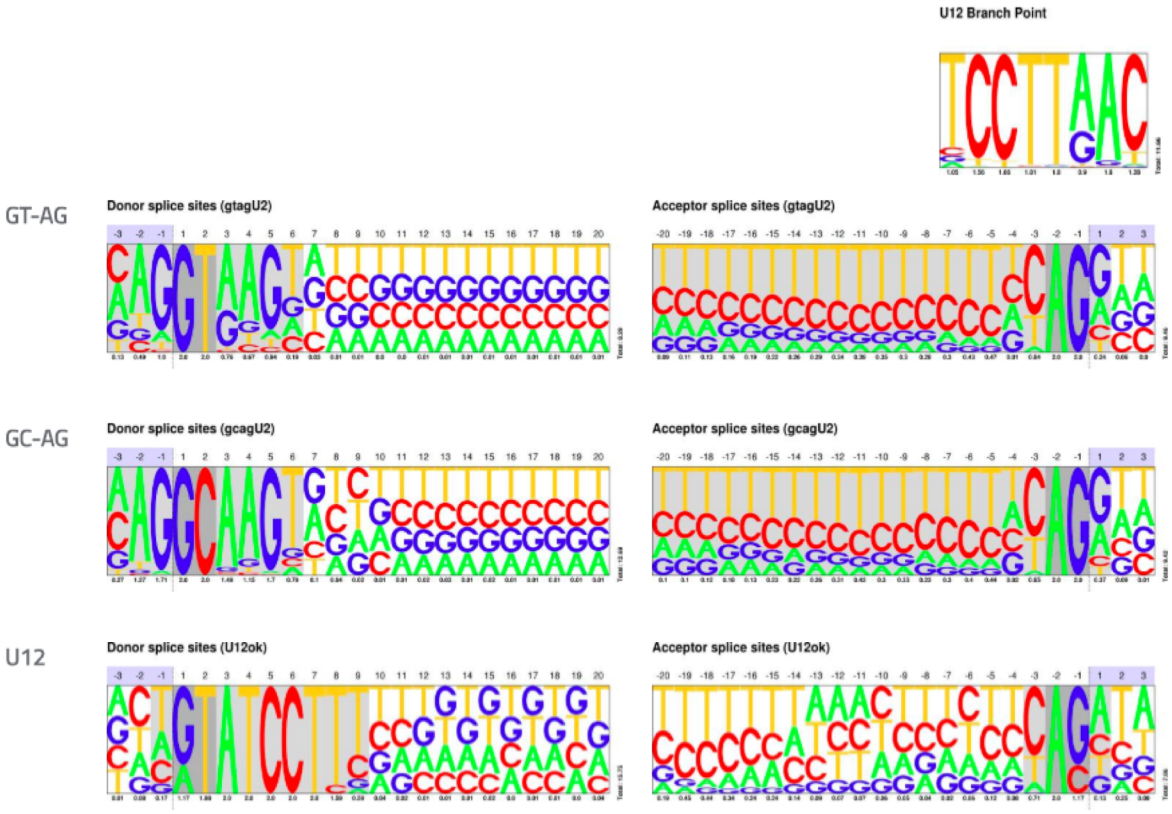
1. PWM Models
2. Hidden Markov Models
3. Maximum Entropy Models
4. Hybrid Networks

Computational Model:

PWMs

Abril, Castelo, Guigó, (2005)

The simplest mechanism for summarizing observed spice site data into a machine learning model. The PWM matrix stores at each location a nucleotide frequency, which may be convolved with a novel sequence to identify potential splice sites.



RNA SPLICING:

MACHINE LEARNING HISTORY AND STATE OF THE ART

1. PWM Models
2. Hidden Markov Models
3. Maximum Entropy Models
4. Hybrid Networks

Computational Model:

HIDDEN MARKOV MODEL

HMM (Marji & Garg, 2013)

Emits state transitions moving sequentially down a DNA sequence to predict state switching between intron and exon states.

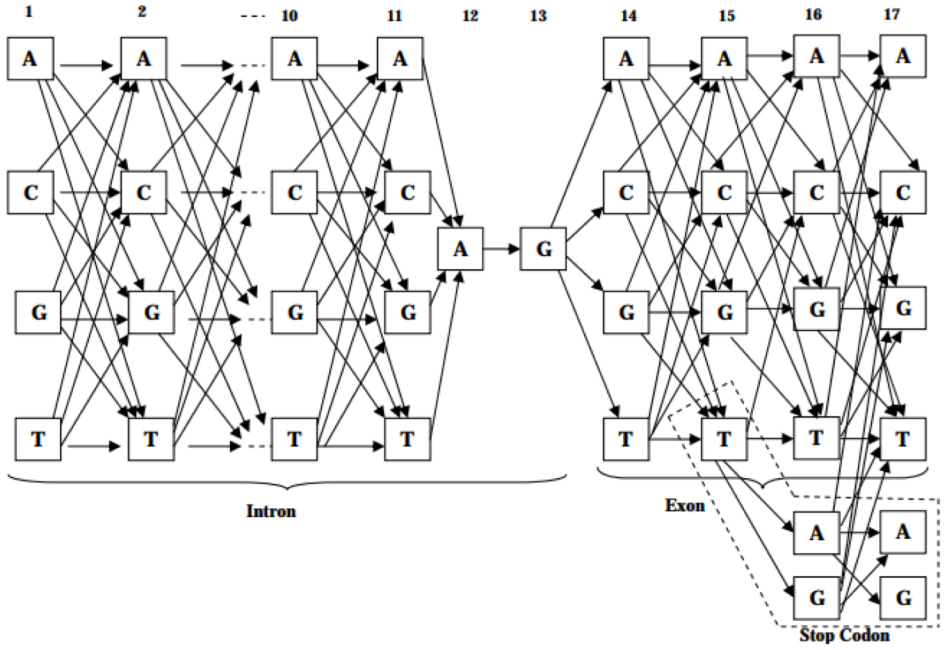


Fig. (3). The Acceptor HMM for 3' splice site.

RNA SPLICING:

MACHINE LEARNING HISTORY AND STATE OF THE ART

1. PWM Models
2. Hidden Markov Models
3. Maximum Entropy Models
4. Hybrid Networks

Computational Model:

MAXIMUM ENTROPY

MAXENT (Yeo & Burge, 2003)

Creates a maximum entropy score, allowing higher-order dependencies than in a simple, single-state Markov model. An improvement over previous models, in 2003.

TABLE 4. TOP 20 RANKED CONSTRAINTS FOR me2x5 FOR 5′ss[a]

Rank	ΔH_i	*Sign*
1	..Ggt..G.	−
2	...gt.AG.	+
3	.AGgt....	+
4	C..gt...C	+
5	...gtAA..	−
6	..GgtT...	+
7	..GgtC...	+
8	..GgtA...	−
9	...gtTA..	−
10	..Tgt..T.	−
11	..Tgt..A.	−
12	.G.gt..C.	−
13	...gtC.G.	+
14	.C.gt..C.	−
15	.T.gt..C.	−
16	..Cgt..A.	−
17	..Cgt..T.	−
18	..Agt..T.	−
19	..Agt..A.	−
20	..Cgt..G.	+

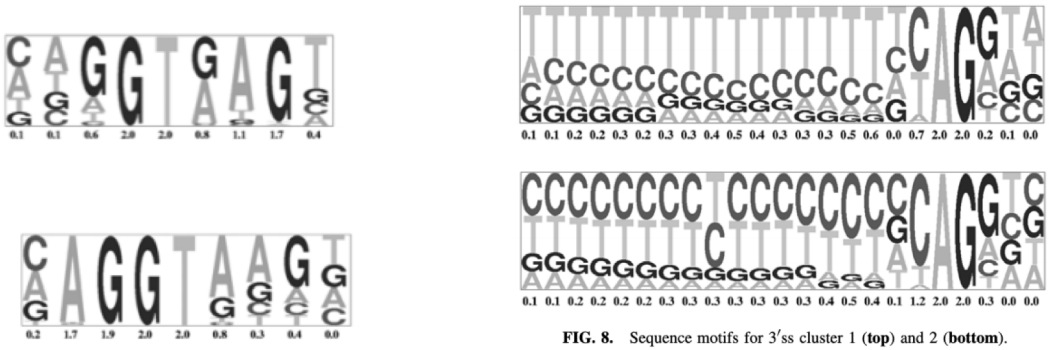


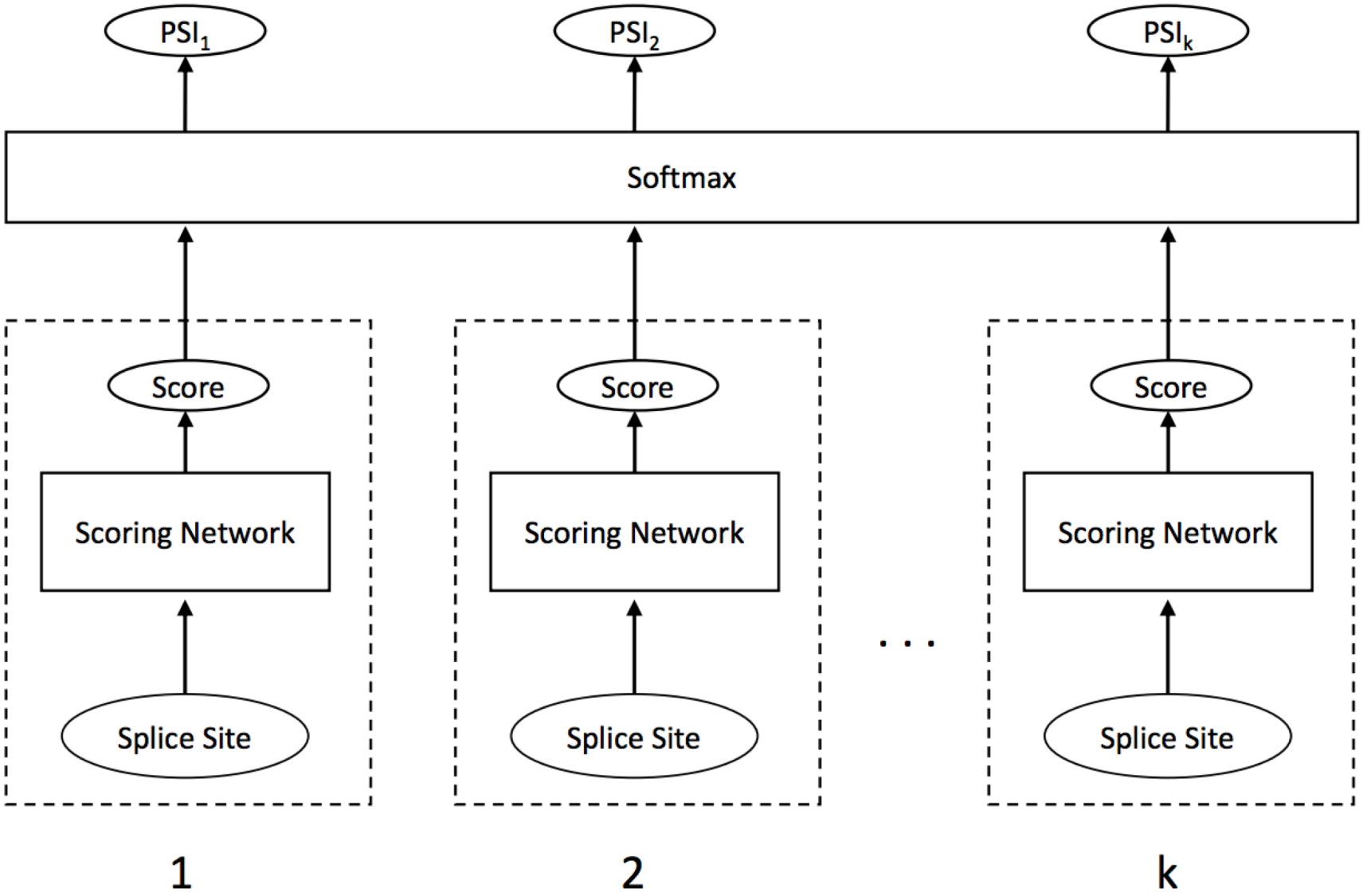
FIG. 8. Sequence motifs for 3′ss cluster 1 (**top**) and 2 (**bottom**).

RNA SPLICING:

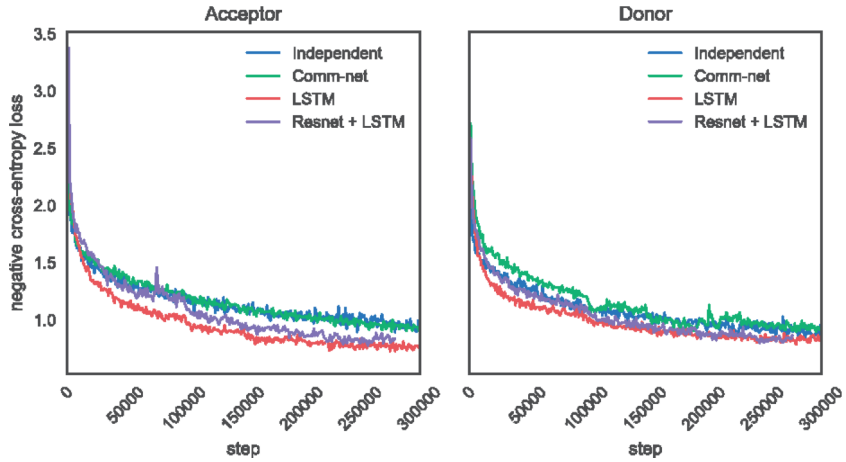
MACHINE LEARNING HISTORY AND STATE OF THE ART

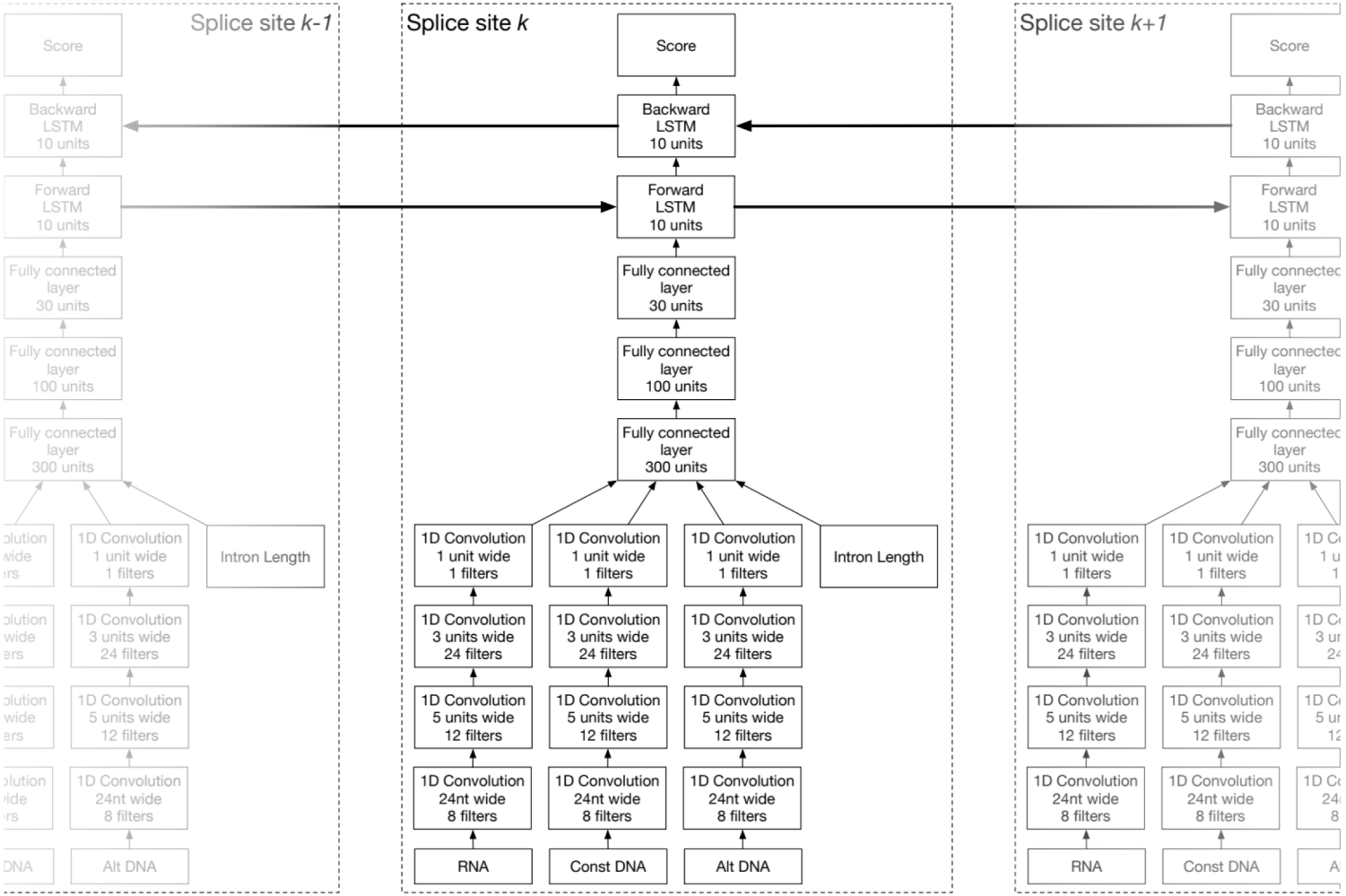
1. PWM Models
2. Hidden Markov Models
3. Maximum Entropy Models
4. Hybrid Networks

The COSSMO Model directly predicts PSI
(Bretschneider et al, 2018)



COSSMO LSTM Model
(Bretschneider et al, 2018) COSSMO uses both convolutional and LSTM layers and outperforms MAXENT scan.



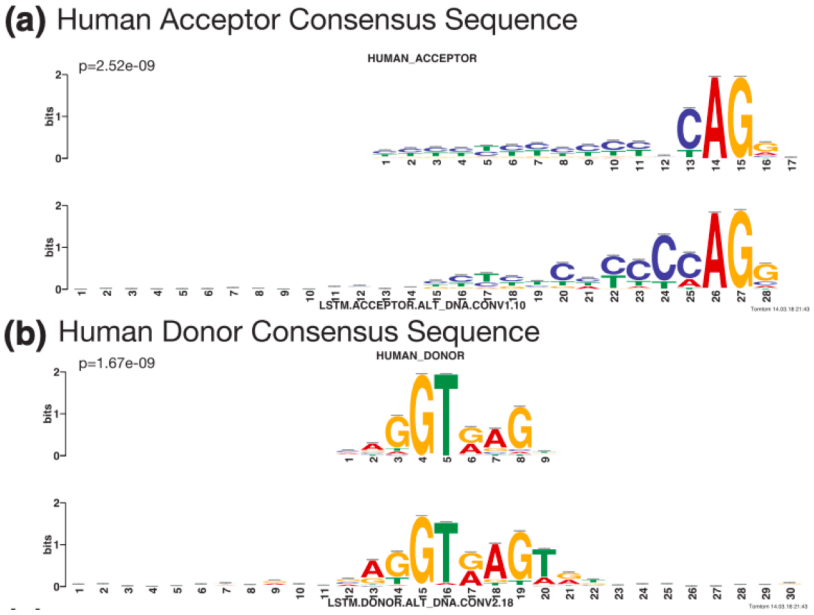


Computational Model:

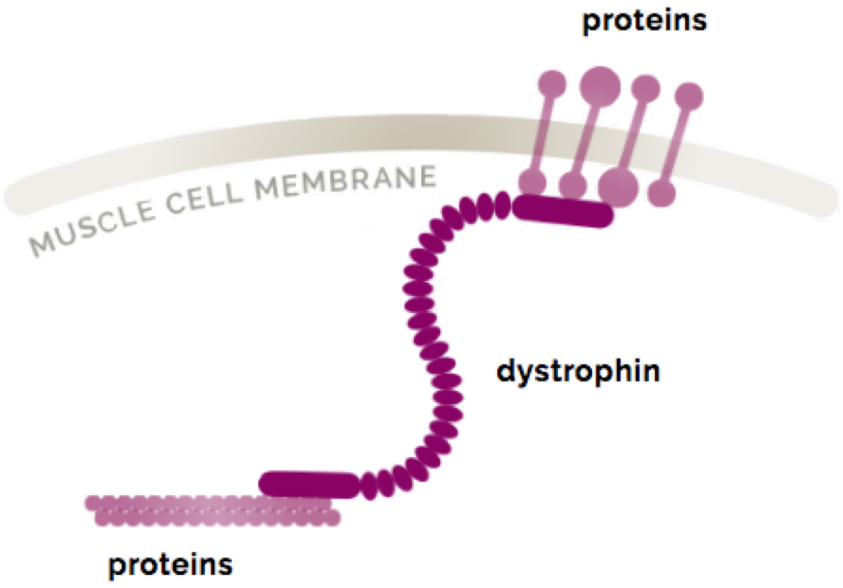
Deep Learning with “COSSMO”

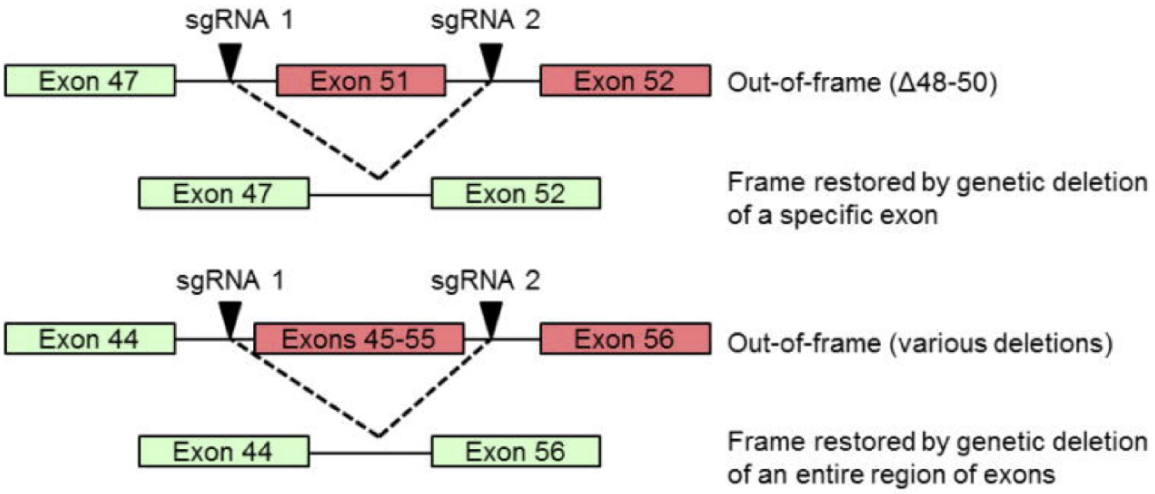
(Bretschneider et al, 2018)

COSSMO rediscovers known splicing motifs. Motifs are extracted by clustering input sequences that activate the network. Reference motifs are on the top and matching motifs learned by COSSMO are on the bottom.



Duchenne muscular dystrophy (DMD), an X-linked recessive disorder in approximately 1 in 5000 males.





https://blog.addgene.org/treating-muscular-dystrophy-with-crispr-gene-editing

FIN - Thank You