Computational Systems Biology
Deep Learning in the Life Sciences

6.802 6.874 20.390 20.490 HST.506

David Gifford
Lecture 4
February 23, 2017

Model Regularization
Models of RNA Expression Data

http://mit6874.github.io

Overall goal for today

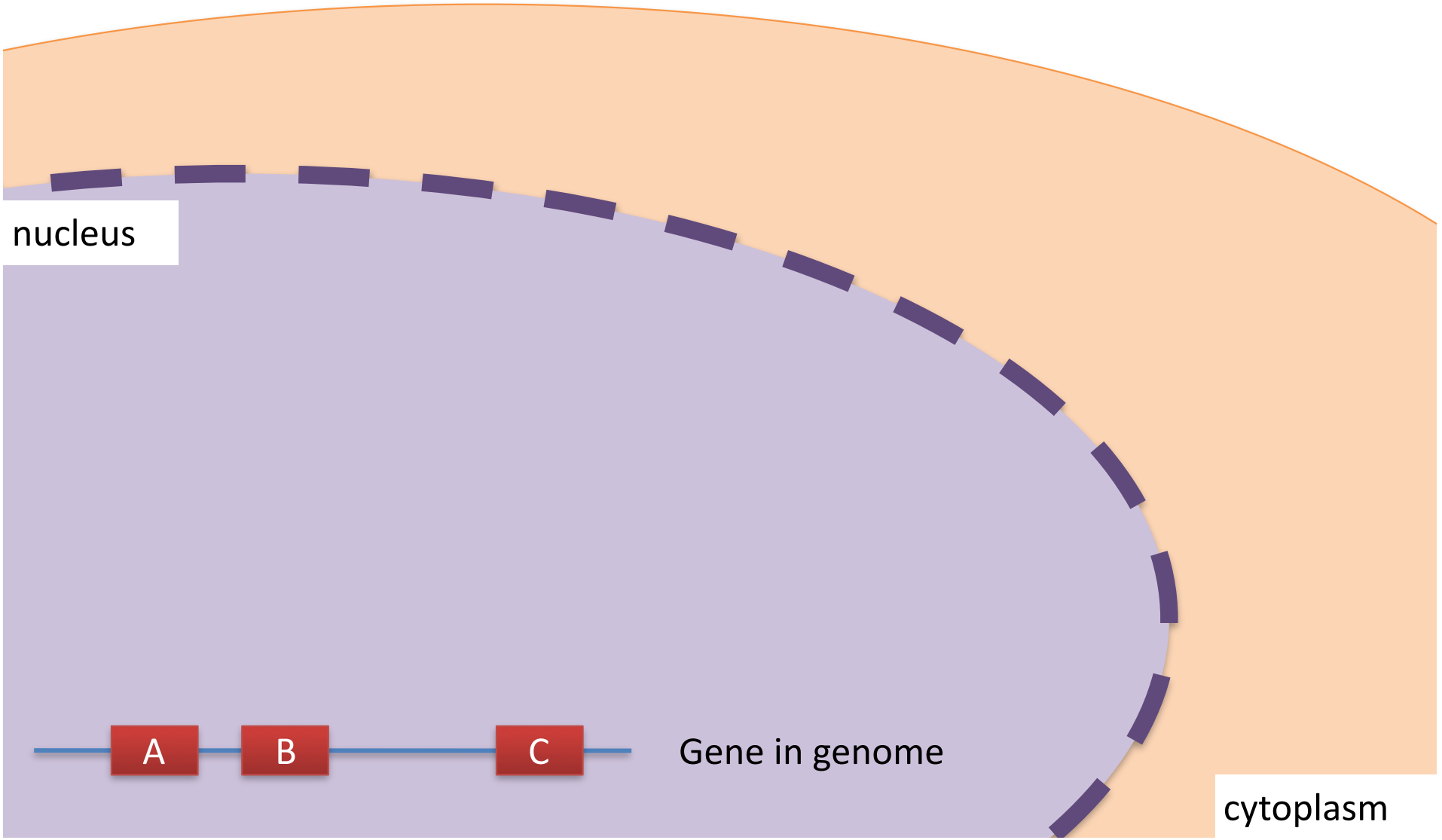
- Understanding overdispersion in count data
- Understanding isoform estimation by expectation maximization
- Understand how to regularize the number and magnitude of model parameters to avoid overfitting

Today's lecture

- Isoform estimation
  - Constraints and expectation maximization to determine maximum likelihood solutions
- Determining the significance of differential expression
  - Overdispersion caused by mixtures
  - Normalizing data and estimating negative binomial parameters
- Regularization controls model complexity
  - L1 norm (sparsity), L2 norm - (magnitude)
  - Properties of a generalizable model

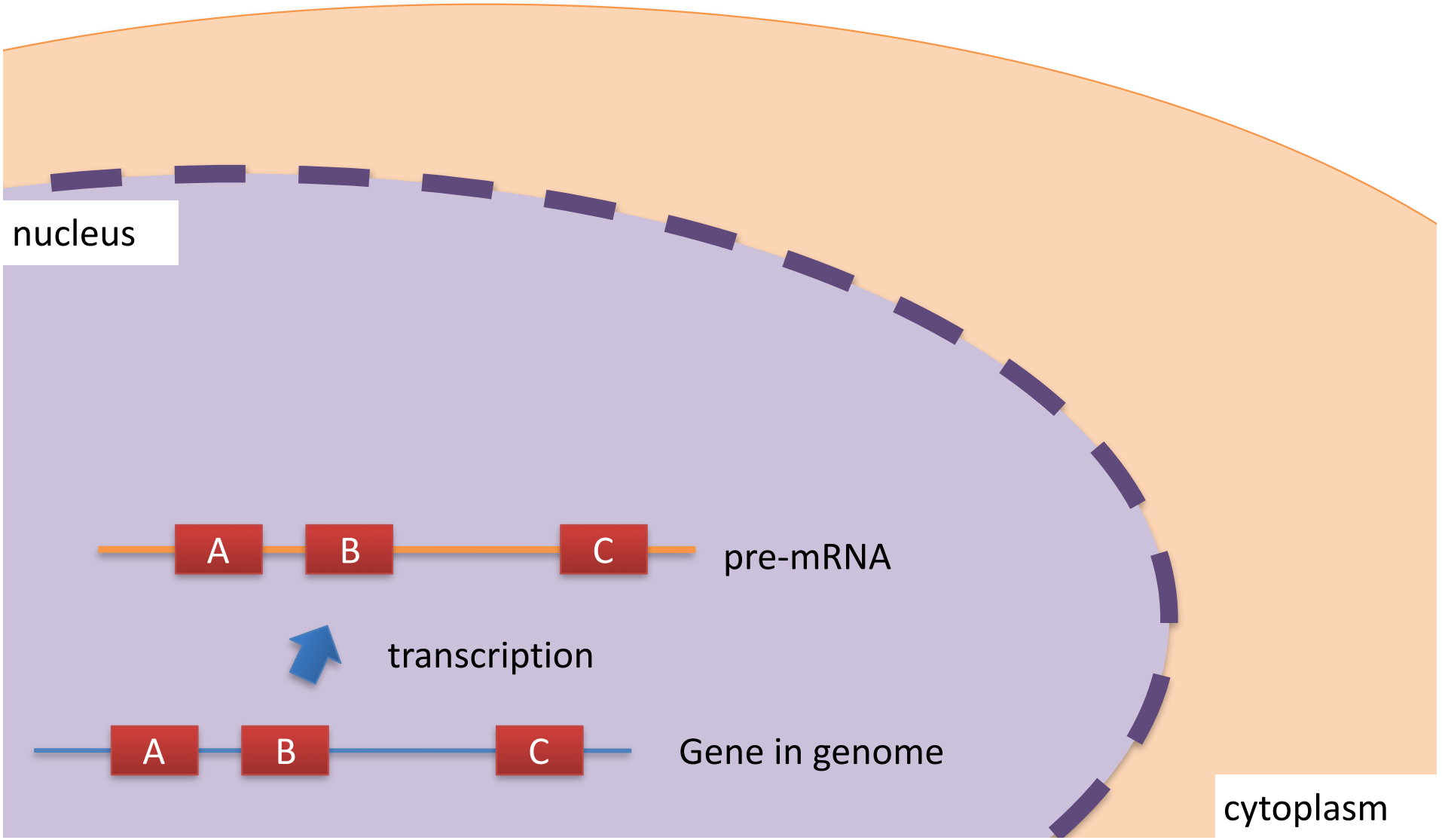
Isoform Estimation (Part I)

Transcription



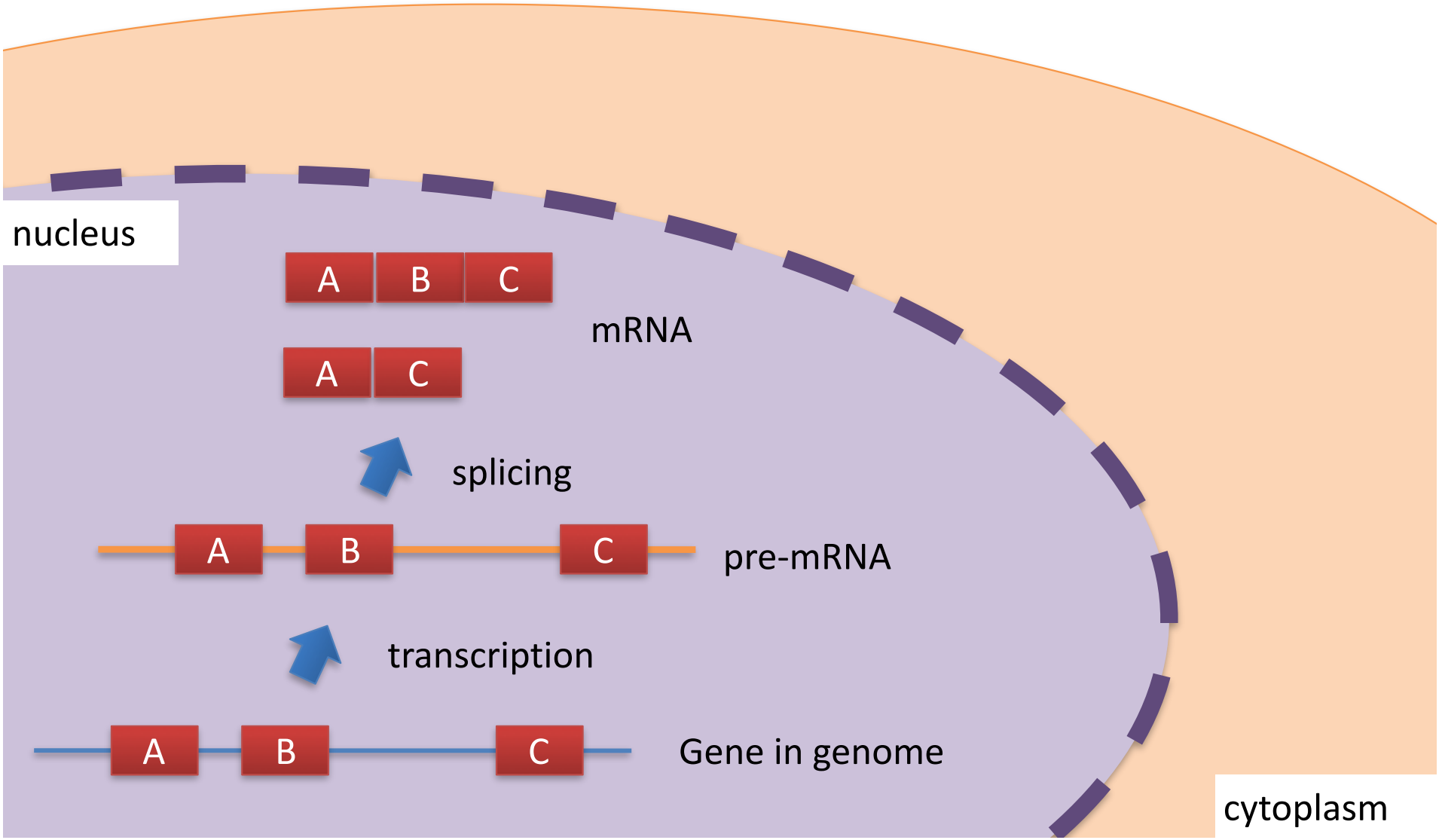
Slide courtesy Cole Trapnell

Transcription



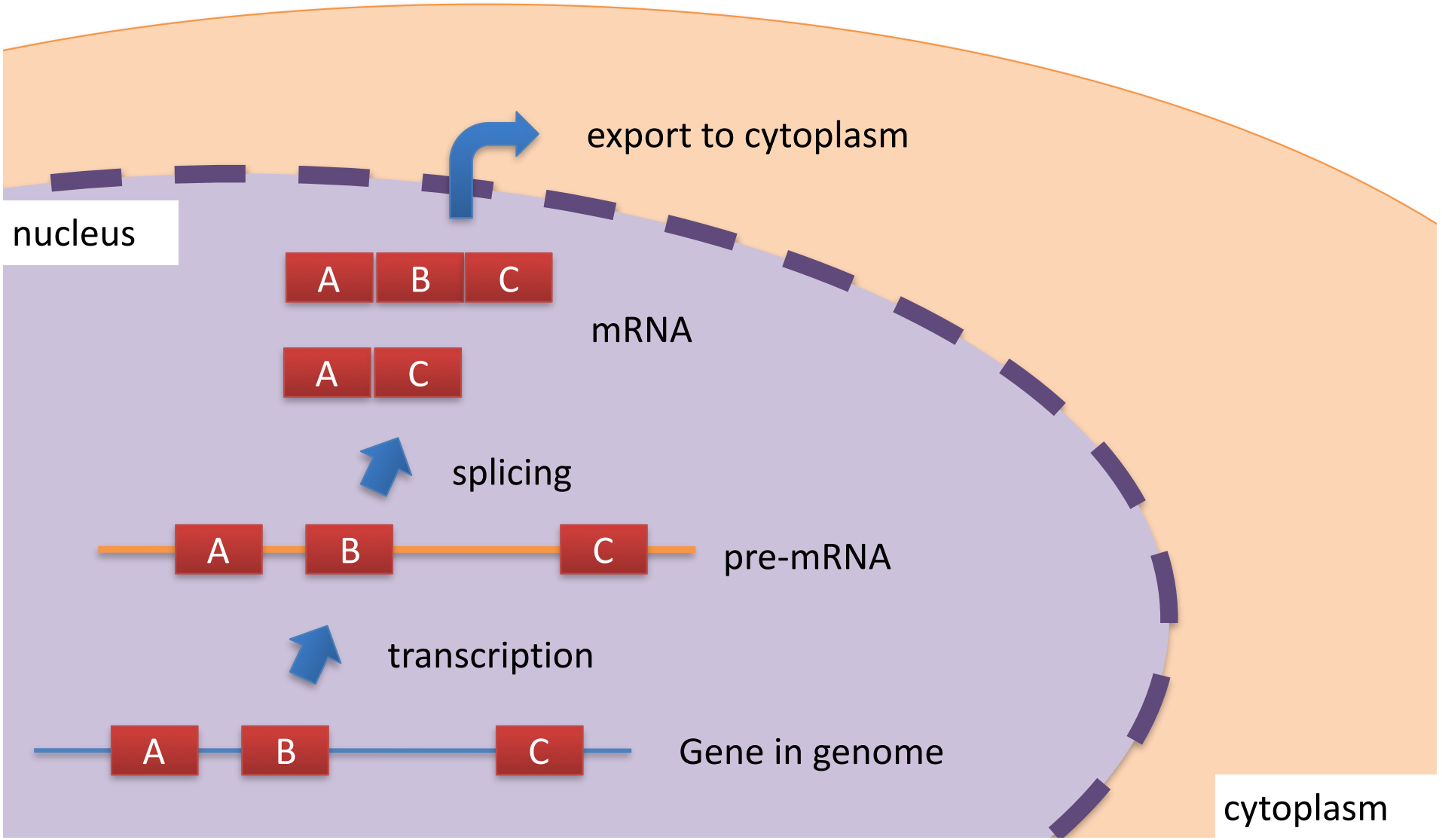
Slide courtesy Cole Trapnell

Transcription



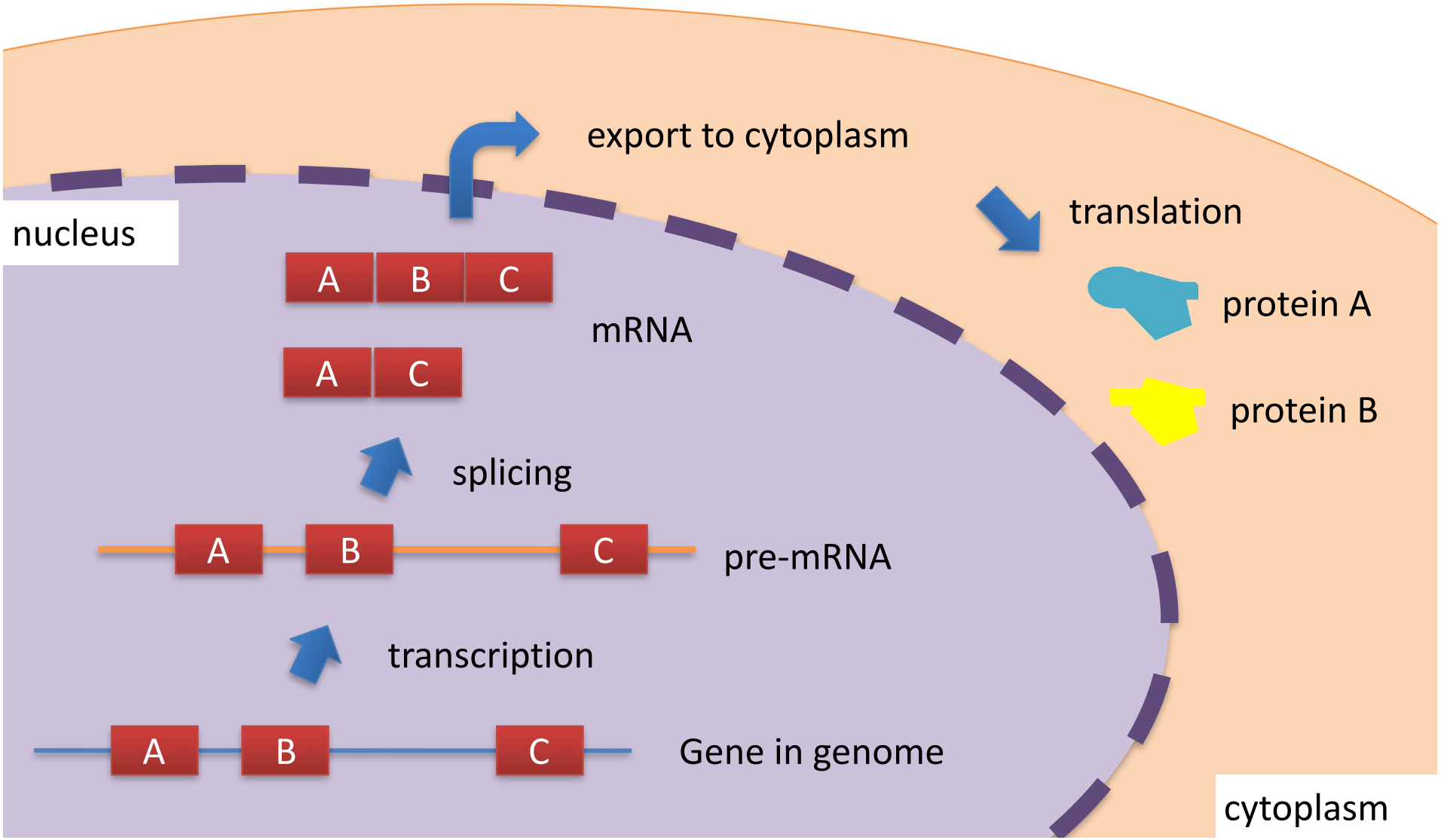
Slide courtesy Cole Trapnell

Transcription



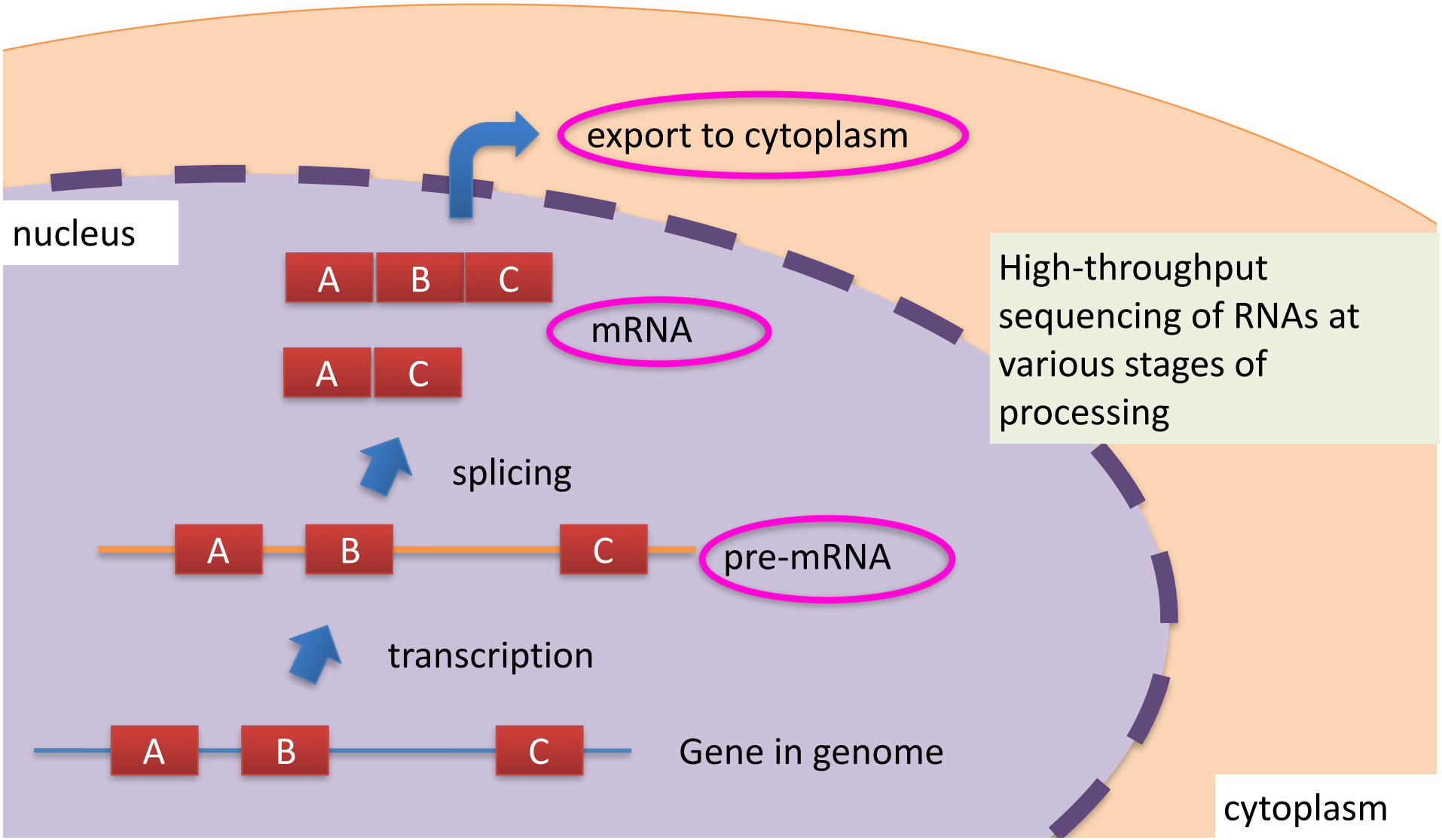
Slide courtesy Cole Trapnell

Transcription



Slide courtesy Cole Trapnell

RNA-Seq

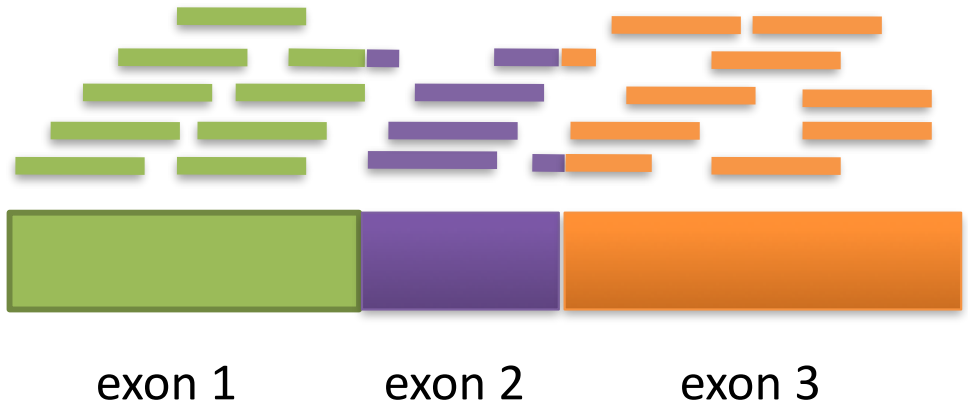


Slide courtesy Cole Trapnell

RNA-Seq

- Better resolution than tiling arrays
- No upfront design (as with exon or junction arrays)
- Vastly cheaper than traditional sequencing
- **Simultaneous discovery and expression assay**

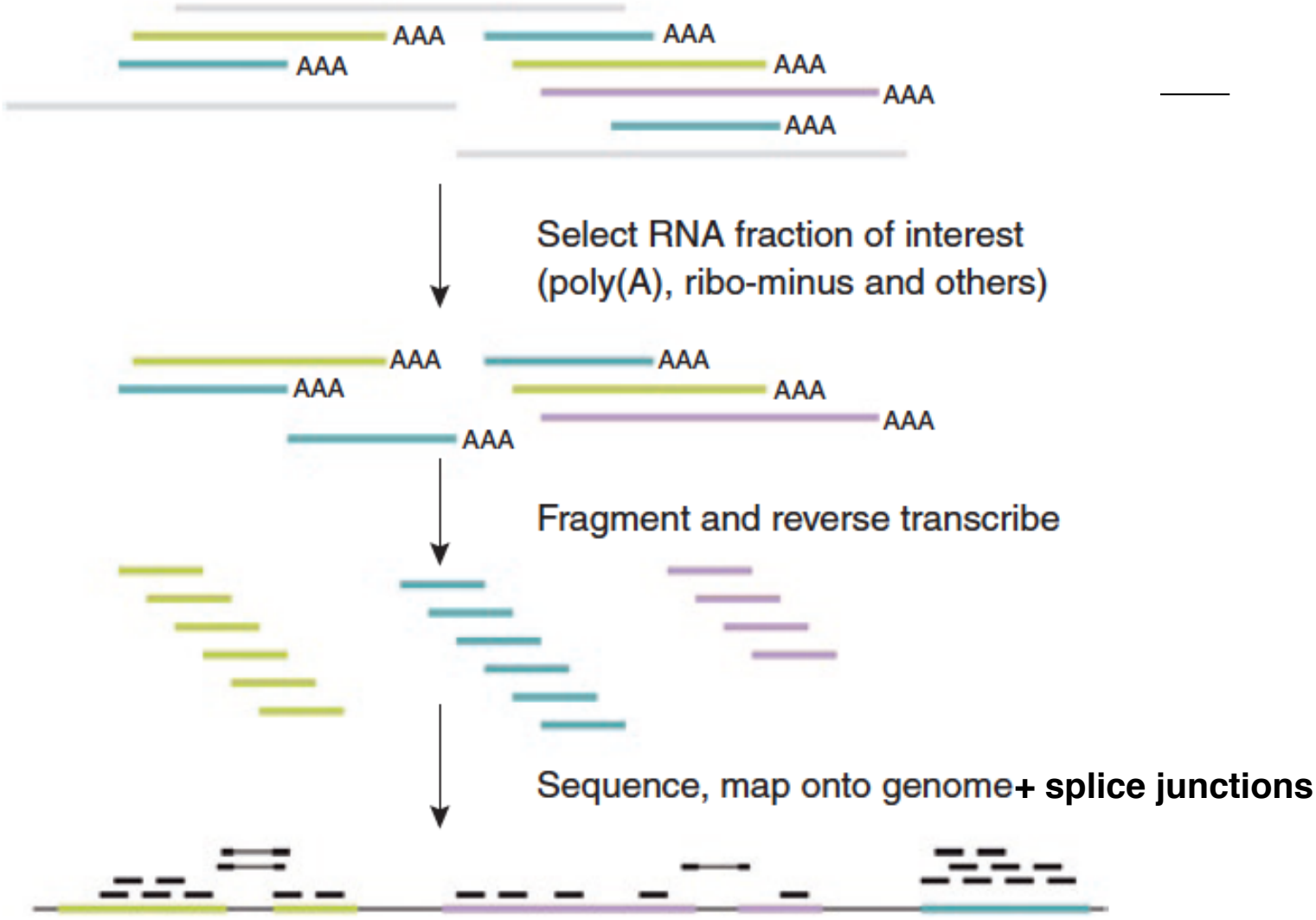
Short sequencing reads,
randomly sampled from a transcript



Slide courtesy Cole Trapnell

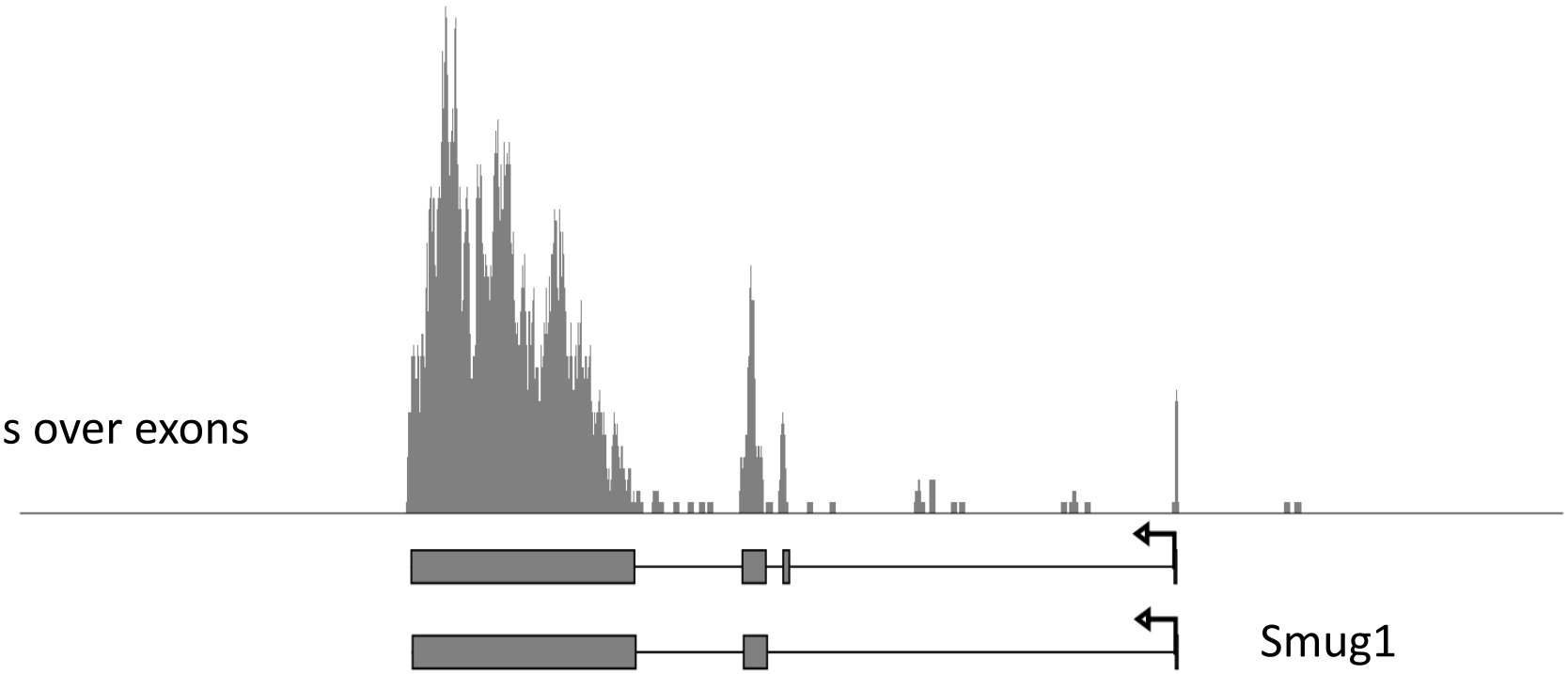
RNA-Seq: millions of short reads from fragmented mRNAs

Extract RNA from cells/tissue

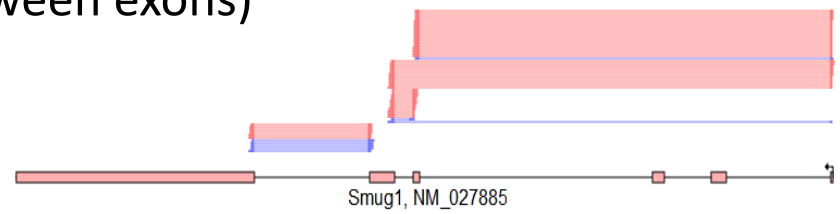


Pepke et. al. *Nature Methods* 2009

Reads over exons

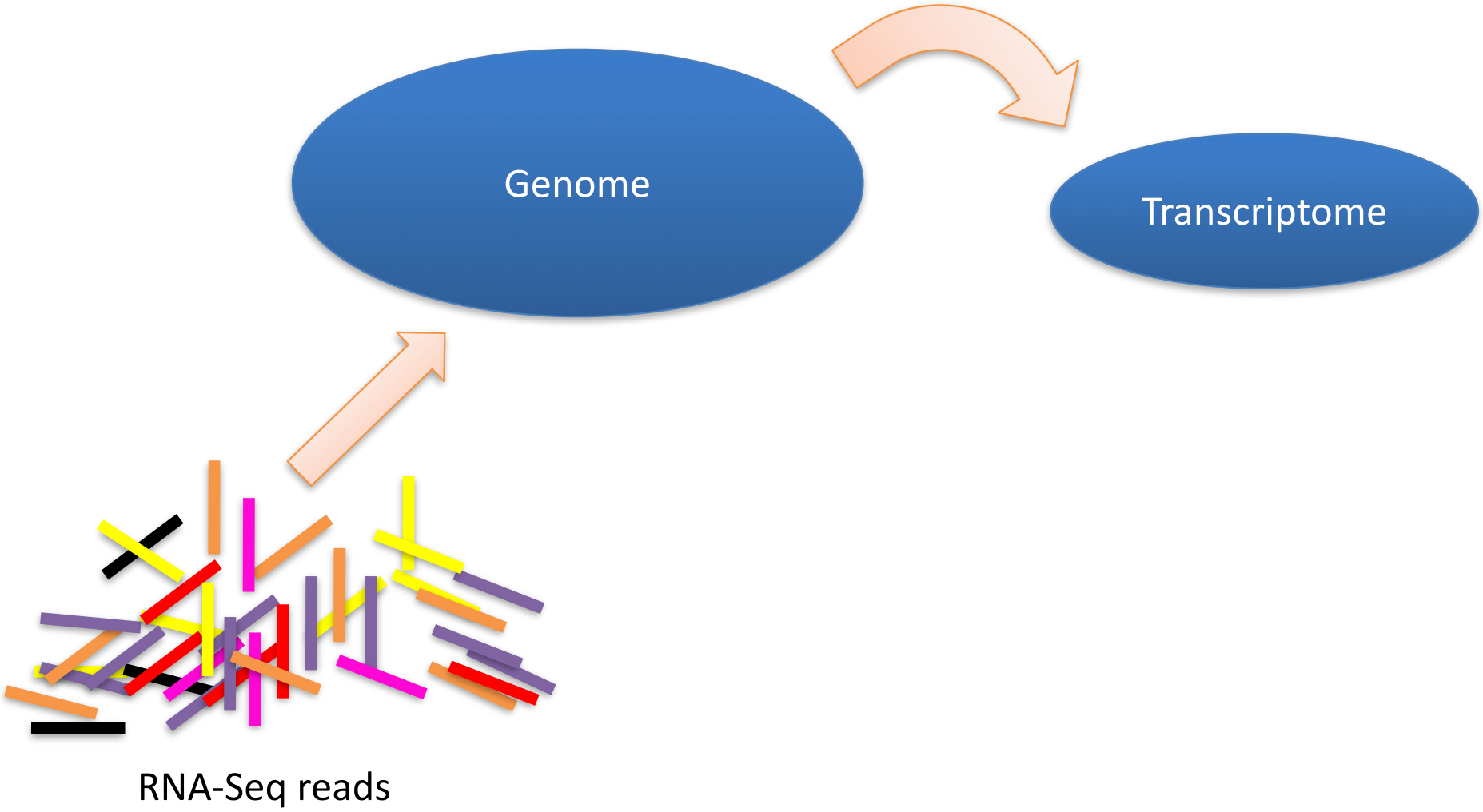


Junction reads (split between exons)



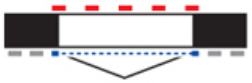
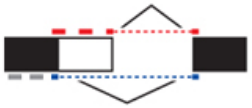
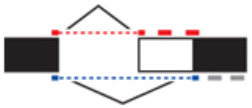
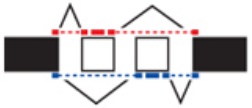
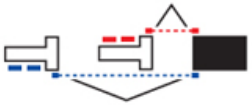
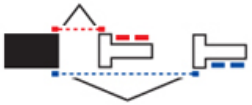
Mapping RNA-Seq reads

Goal: **identify** all transcripts and estimate relative amounts from RNA-Seq data



Slide courtesy Cole Trapnell

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

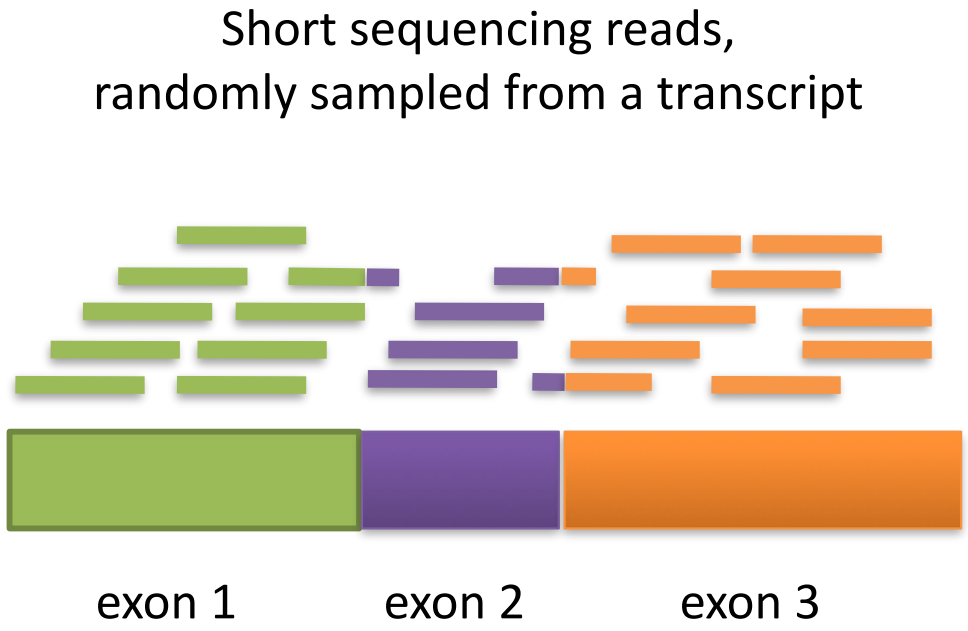
Constitutive exon or region — Body read — Junction read — pA Polyadenylation site

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

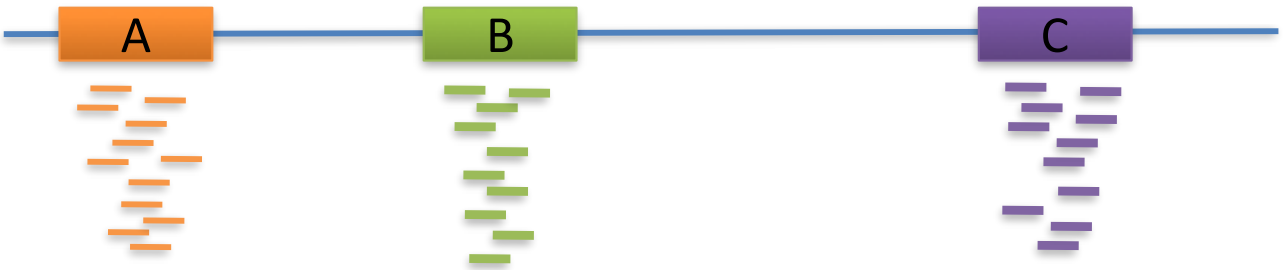
ET Wang *et al. Nature* 000, 1-7 (2008) doi:10.1038/nature07509

Analysis ideas

- Assemble reads into transcripts – problem is ambiguity in assembly graph
- Map reads to genome and identify possible isoforms using constraints
- Common RNA-seq expression metric is Reads per killobase per million reads (RPKM)
- Goal is to quantify isoforms and determine significance of differential expression

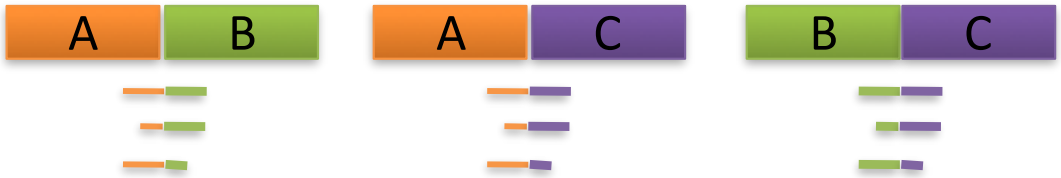


Aligned reads reveal isoform possibilities



identify candidate exons via genomic mapping

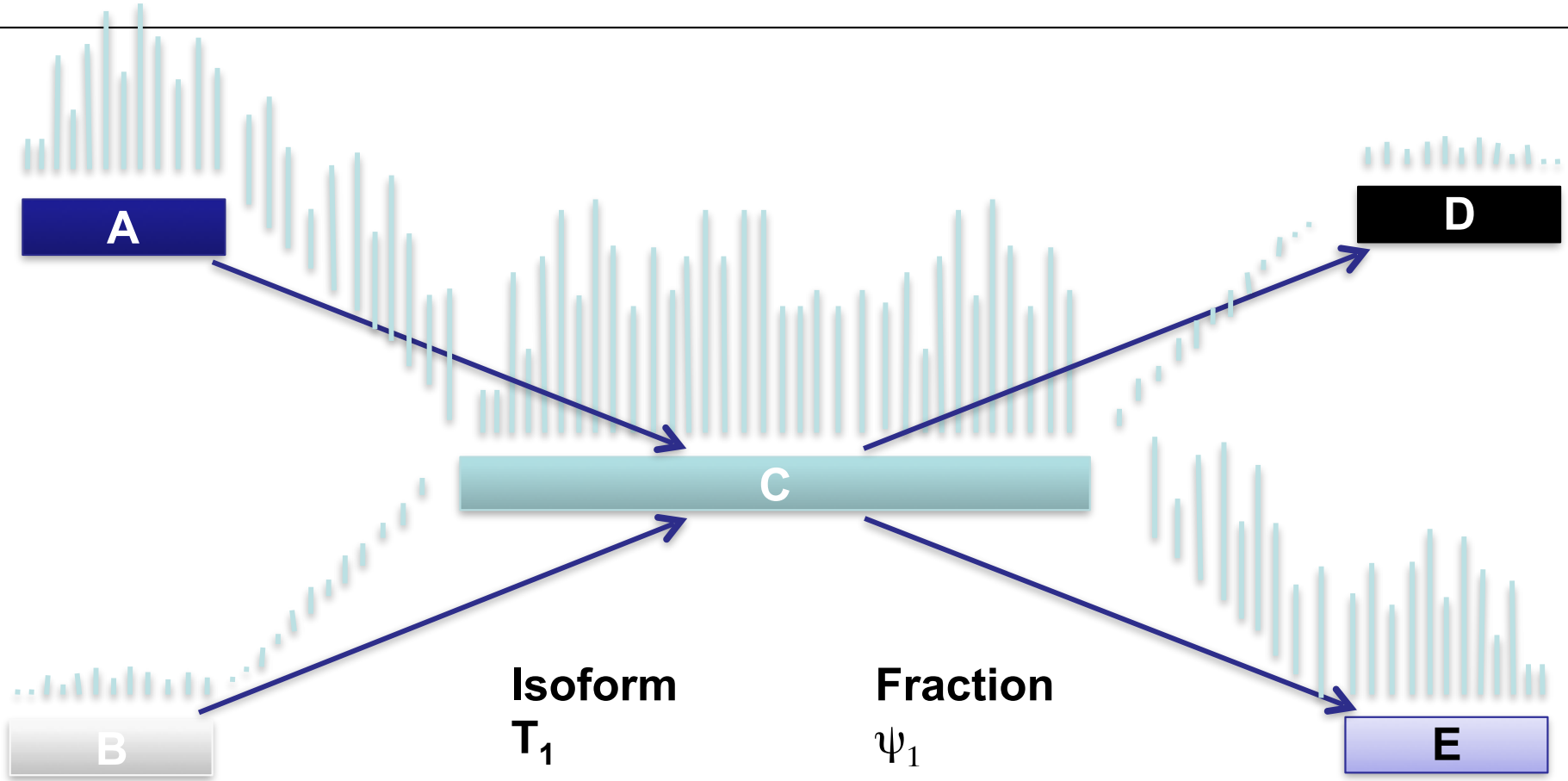
Generate possible pairings of exons



Align reads to possible junctions

Slide courtesy Cole Trapnell

We can use mapped reads to learn the isoform mixture ψ



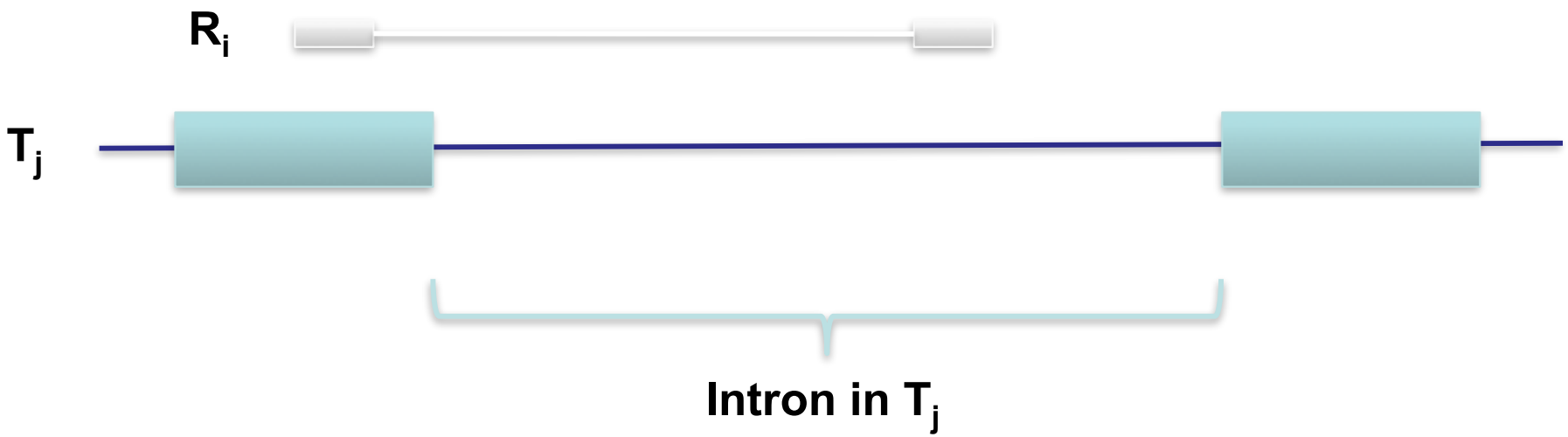
Isoform	Fraction
T_1	ψ_1
T_2	ψ_2
T_3	ψ_3
T_4	ψ_4

Slide courtesy Cole Trapnell

$P(R_i \mid T=T_j)$ – Excluded reads

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

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



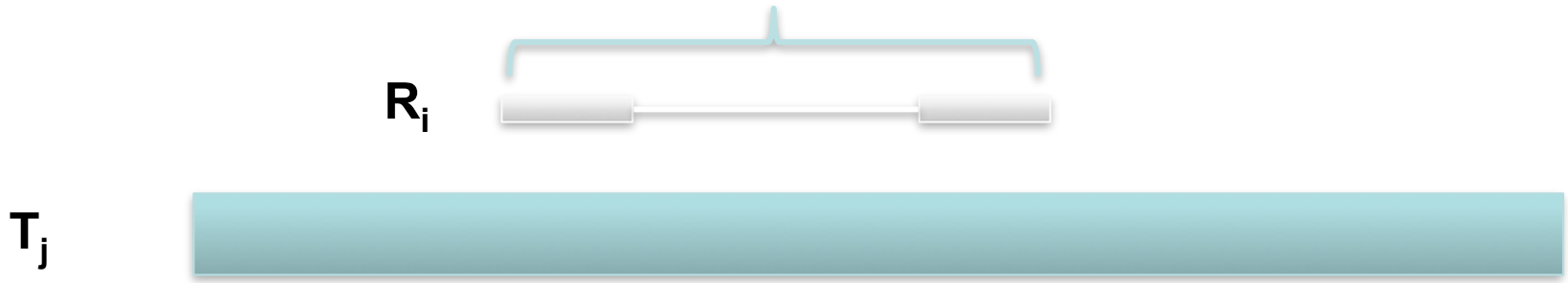
Slide courtesy Cole Trapnell

$P(R_i \mid T=T_j)$ – 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 $R_1,\ldots,R_m$

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

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

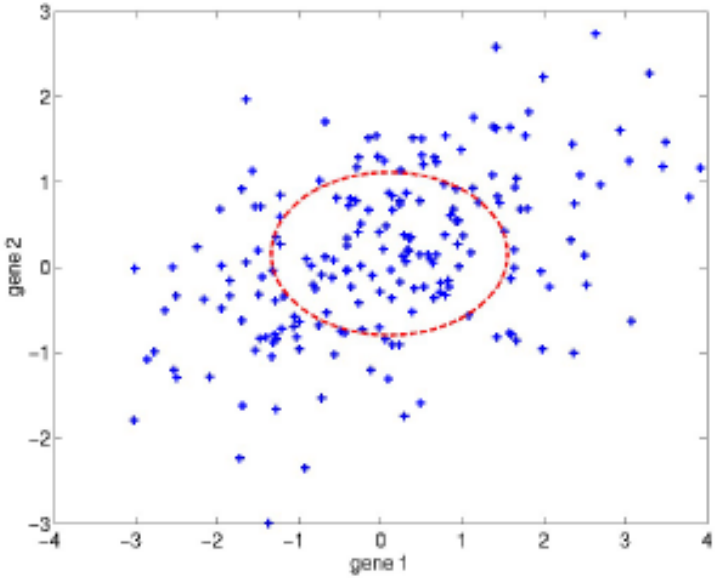
$$\Psi = \underset{\Psi}{\operatorname{argmax}} L(\Psi \mid R)$$

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

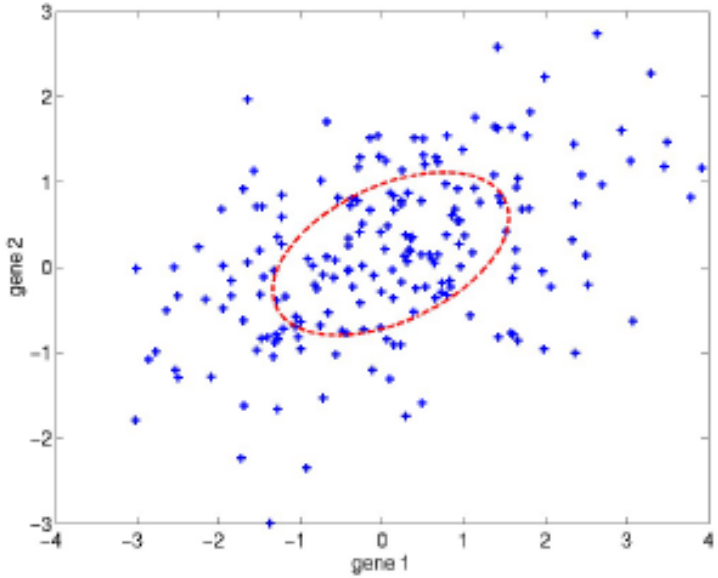
Understanding overdispersion (Part II)

Statistical tests: example

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



null hypothesis



alternative hypothesis

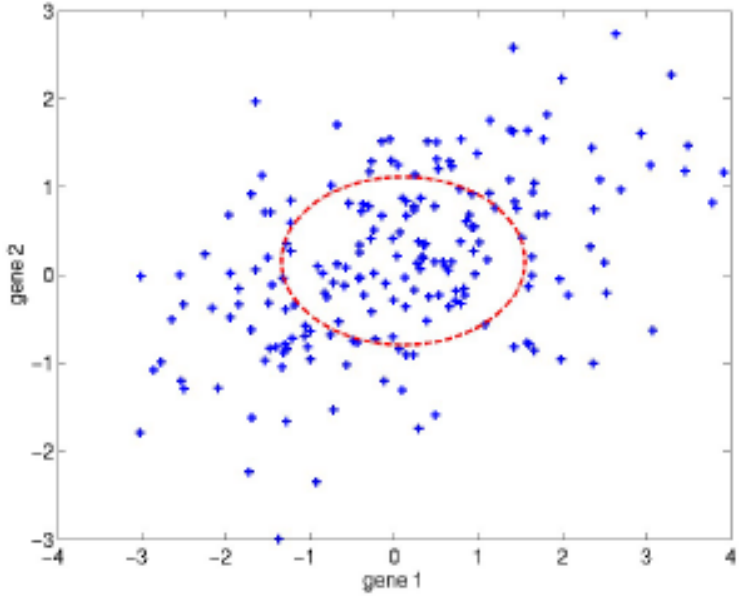
- We need to find a way of testing whether this difference is significant

Degrees of freedom

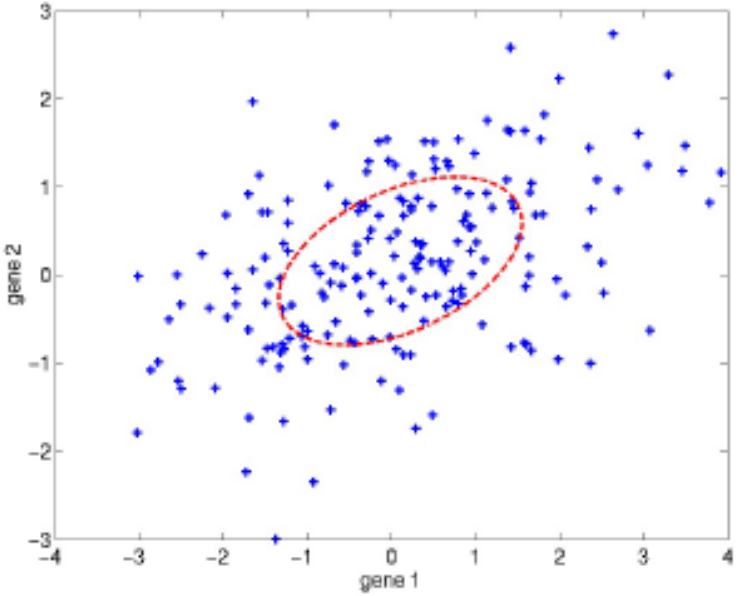
- How many degrees of freedom do we have in the two models?

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

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



H_0



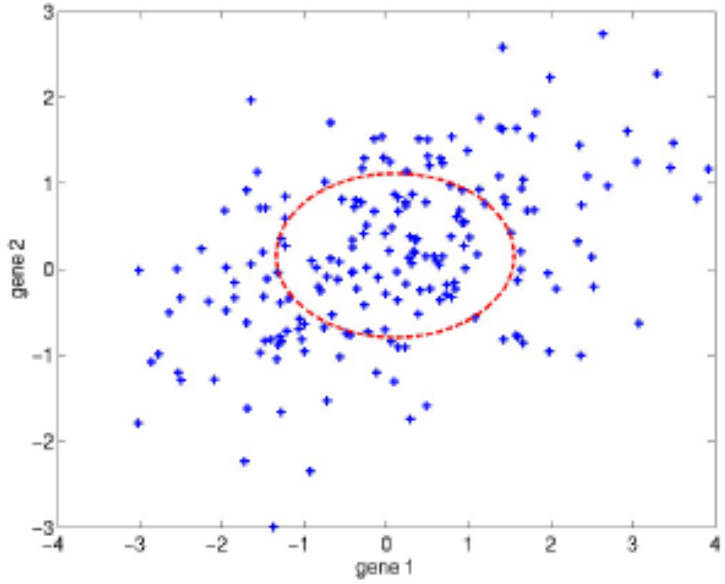
H_1

Degrees of freedom

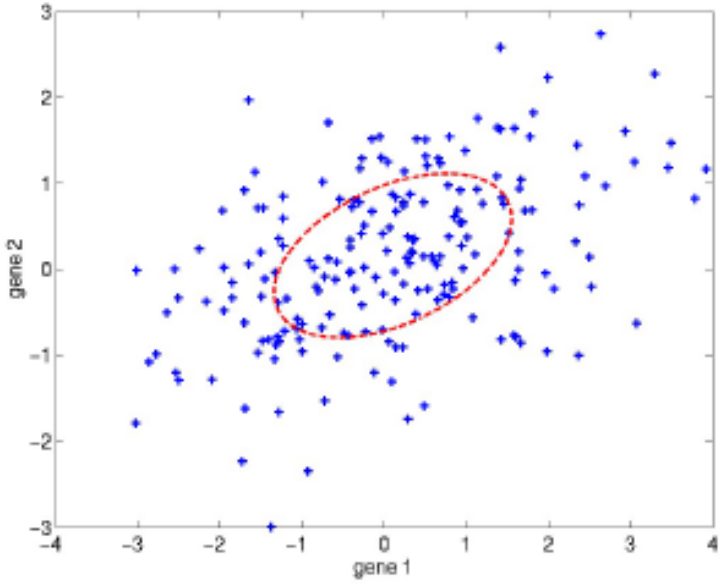
- How many degrees of freedom do we have in the two models?

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

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



H_0



H_1

- The observed data overwhelmingly supports H_1

Test statistic

- Likelihood ratio statistic

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

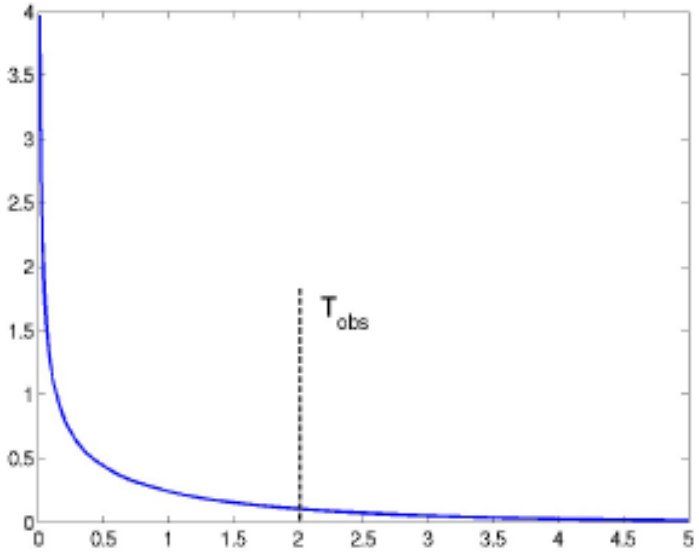
  Larger values of T imply that the model corresponding to the null hypothesis H_0 is much less able to account for the observed data

- 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

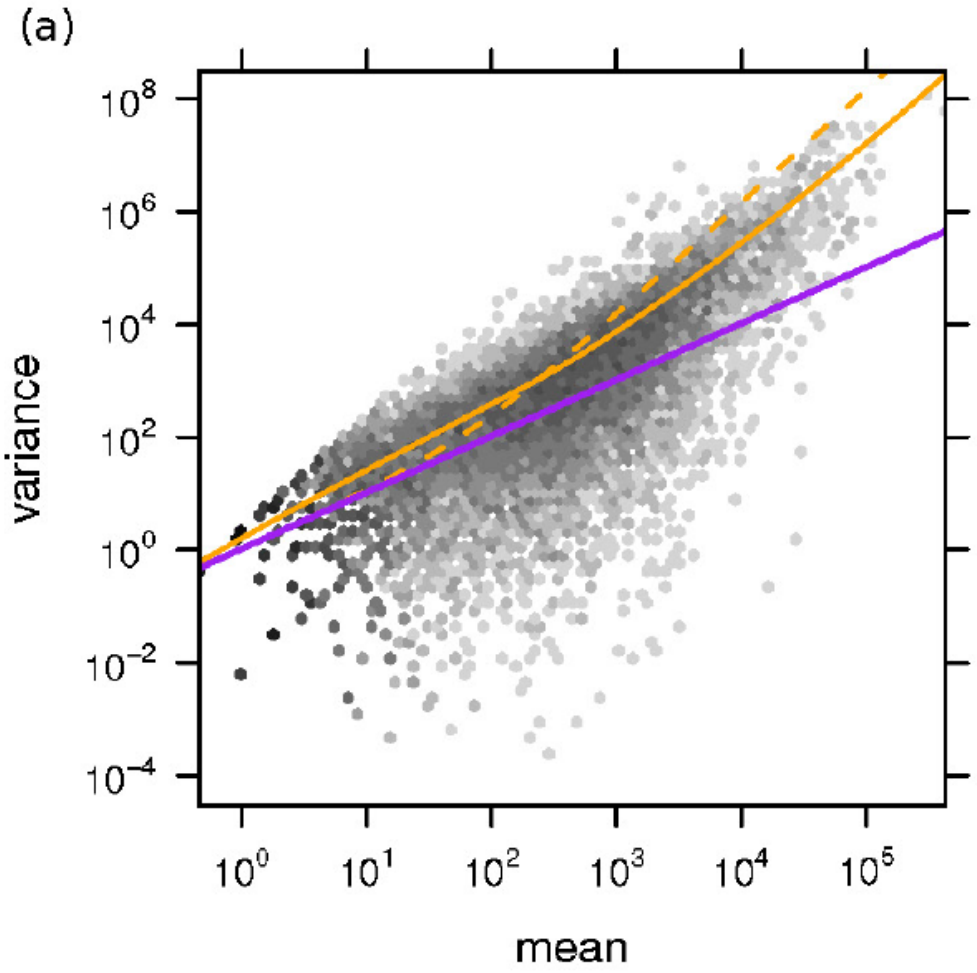
- For the likelihood ratio statistic, the sampling distribution is χ^2 with degrees of freedom equal to the difference in the number of free parameters in the two hypotheses



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

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

Count data is overdispersed for a Poisson Model (variance increases more than mean)



Orange Line – DESeq
Dashed Orange – edgeR
Purple - Poission

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

The negative binomial distribution is parameterized by its mean and variance

$NB(k|r,p)$ is the probability of k successes until an experiment is stopped after r failures and a success probability of p

$$Poisson(n|\lambda) = \frac{e^{-\lambda}\lambda^n}{n!} \tag{1}$$

$$\mu_{Poisson} = \sigma^2_{Poisson} = \lambda \tag{2}$$

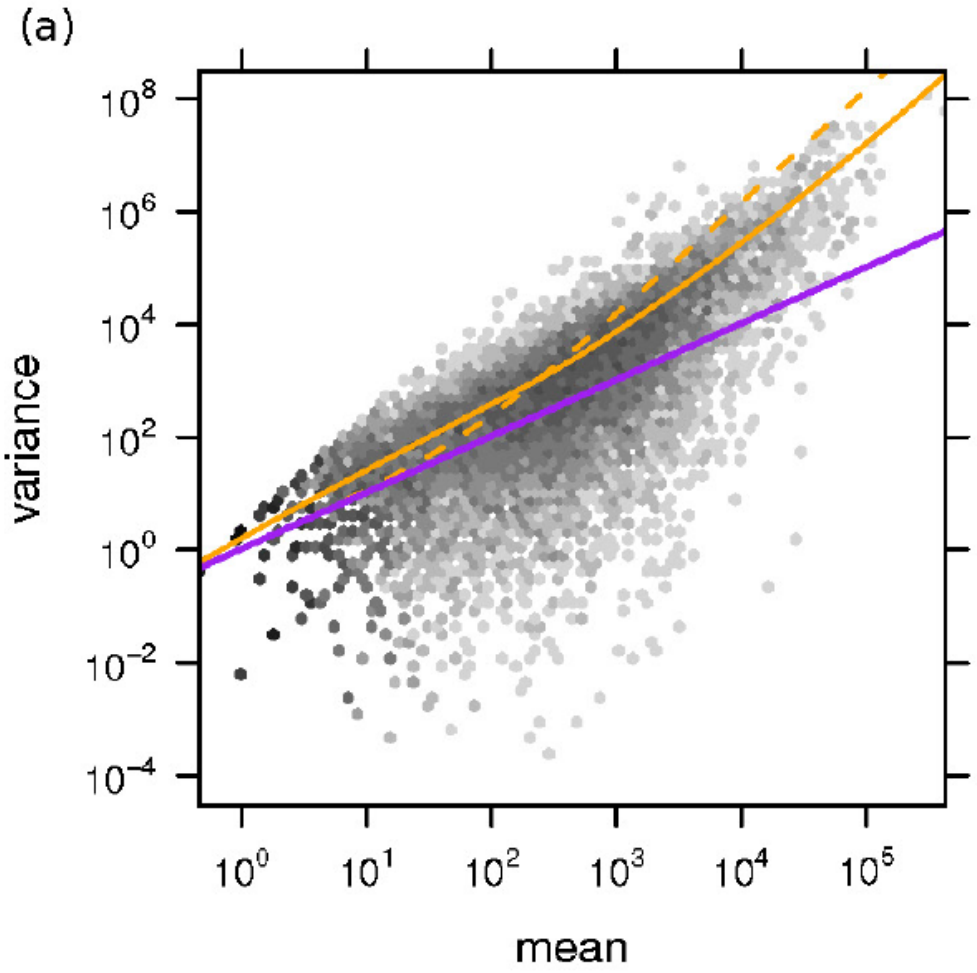
$$NB(k|r,p) = \int_0^\infty Poisson(k|\lambda) Gamma(\lambda|r, \frac{1-p}{p})\ d\lambda \tag{3}$$

$$NB(k|r,p) = \binom{k+r-1}{k}(1-p)^r\ p^k \tag{4}$$

$$\mu_{NB} = \frac{pr}{1-p} \tag{5}$$

$$\sigma^2_{NB} = \frac{pr}{(1-p)^2} \tag{6}$$

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



Orange Line – DESeq
Dashed Orange – edgeR
Purple - Poission

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

Scaling RNA-seq data (DESeq)

- i gene or isoform
- j sample (experiment)
- m number of samples
- K_{ij} number of counts for isoform i in experiment j
- s_j sampling depth for experiment j (scale factor)

$$s_j = \underset{i}{median} \frac{K_{ij}}{\left(\prod_{v=1}^{m} K_{iv}\right)^{1/m}}$$

Model for RNA-seq data (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

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

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

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

Significance of differential expression using test statistics

- Hypothesis H0 (null) – Condition A and B identically express isoform i with random noise added
- Hypothesis H1 – Condition A and B differentially express isoform
- Degrees of freedom (dof) is the number of free parameters in H1 minus the number of free parameters in H0; in this case degrees of freedom is 4 – 2 = 2 (H1 has an extra mean and variance).
- Likelihood ratio test defines a test statistic that follows the Chi Squared distribution

$$T_i = 2\log\frac{P(K_{iA}\,|\,H1)P(K_{iB}\,|\,H1)}{P(K_{iA},K_{iB}\,|\,H0)}$$

$$P(H0) \approx 1 - ChiSquaredCDF(T_i\,|\,dof)$$

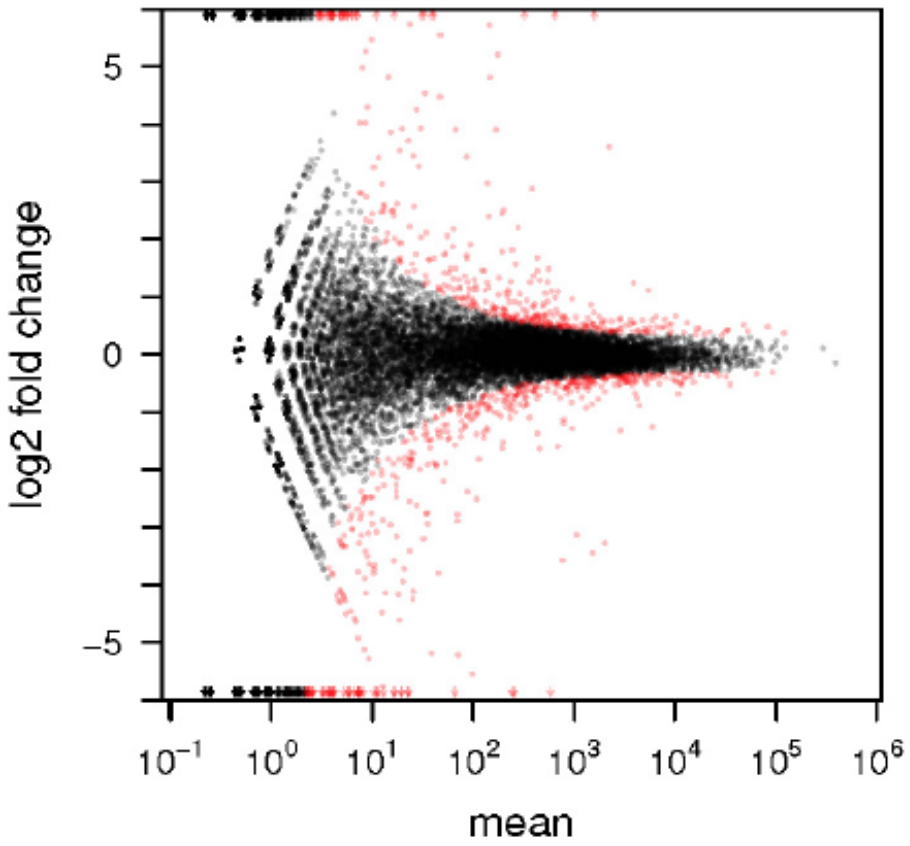


Figure 3 Testing for differential expression between conditions *A* and *B*: Scatter plot of $\log_2$ ratio (fold change) versus mean. The red colour marks genes detected as differentially expressed at 10% false discovery rate when Benjamini-Hochberg multiple testing adjustment is used. The symbols at 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 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

Model regularization controls overfitting (Part III)

Regularization is an important aspect of learning a stable model that generalizes well

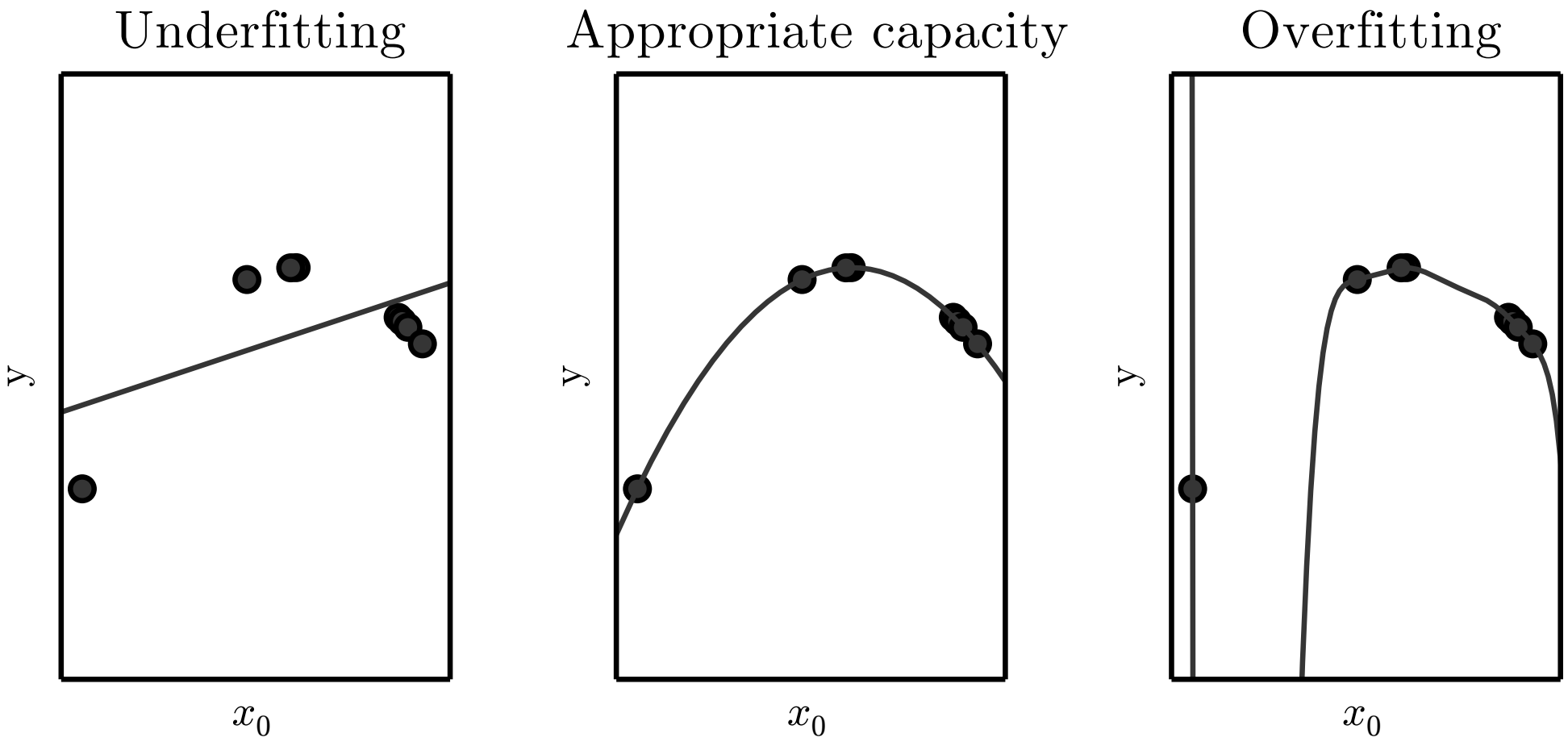


Figure 5.2

Regularizing regression

$$\mathrm{Err}_{\mathrm{ridge}}(w, w_0) = \sum_{i=1}^{n} \left(w \cdot x^{(i)} + w_0 - y^{(i)} \right)^2 + \lambda \|w\|_2^2$$

$$z_j^{(i)} = x_j^{(i)} - \bar{x}_j \qquad y_c^{(i)} = y^{(i)} - \bar{y} \qquad \|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

$$\begin{aligned} E_{\mathrm{ridge}}(W) &= (Y_c - ZW)^T (Y_c - ZW) + \lambda W^T W \\ \nabla_W E_{\mathrm{ridge}}(W) &= Z^T (ZW - Y_c) + \lambda W = 0 \\ W_{\mathrm{ridge}} &= (Z^T Z + \lambda I)^{-1} Z^T Y_c \ . \end{aligned}$$

Penalizing Weights

- L1 Norm attempts to drive weights to 0 (make weight vector sparse) (p = 1)
- L2 Norm attempts to minimize magnitude of weights (p = 2)

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

Weight Decay

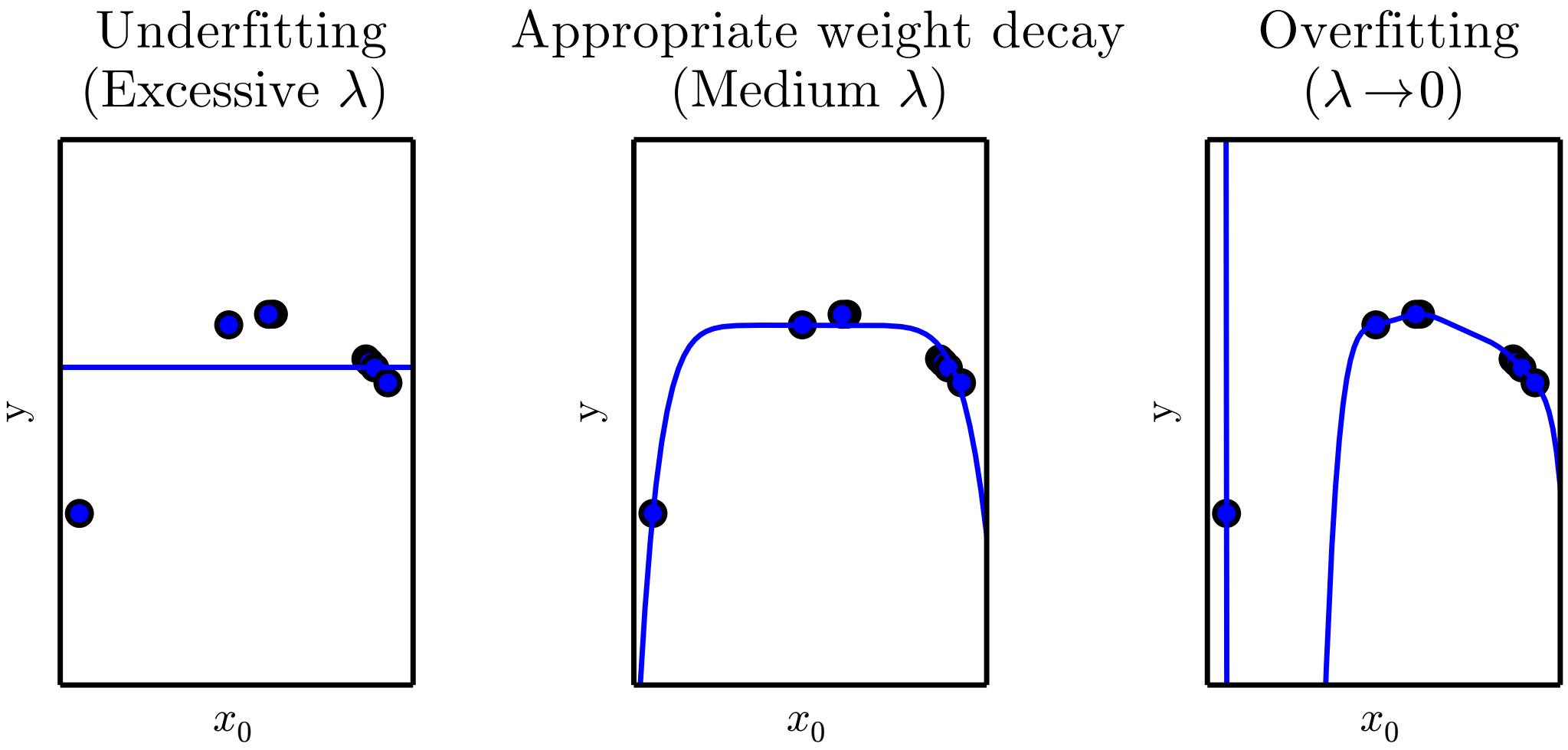


Figure 5.5

Getting the right “fit” to the training data is important

- Underfitting - high error on training set
- Overfitting - large gap between training and test error
- Correct fit is important so the model generalizes to new examples

Model capacity

- Capacity - ability to fit a wide range of functions (hypothesis space)
- Vapnik-Chervonenkis dimension - size of largest unique training set of examples a binary classifier can label arbitrarily is a measure of capacity

Generalization and Capacity

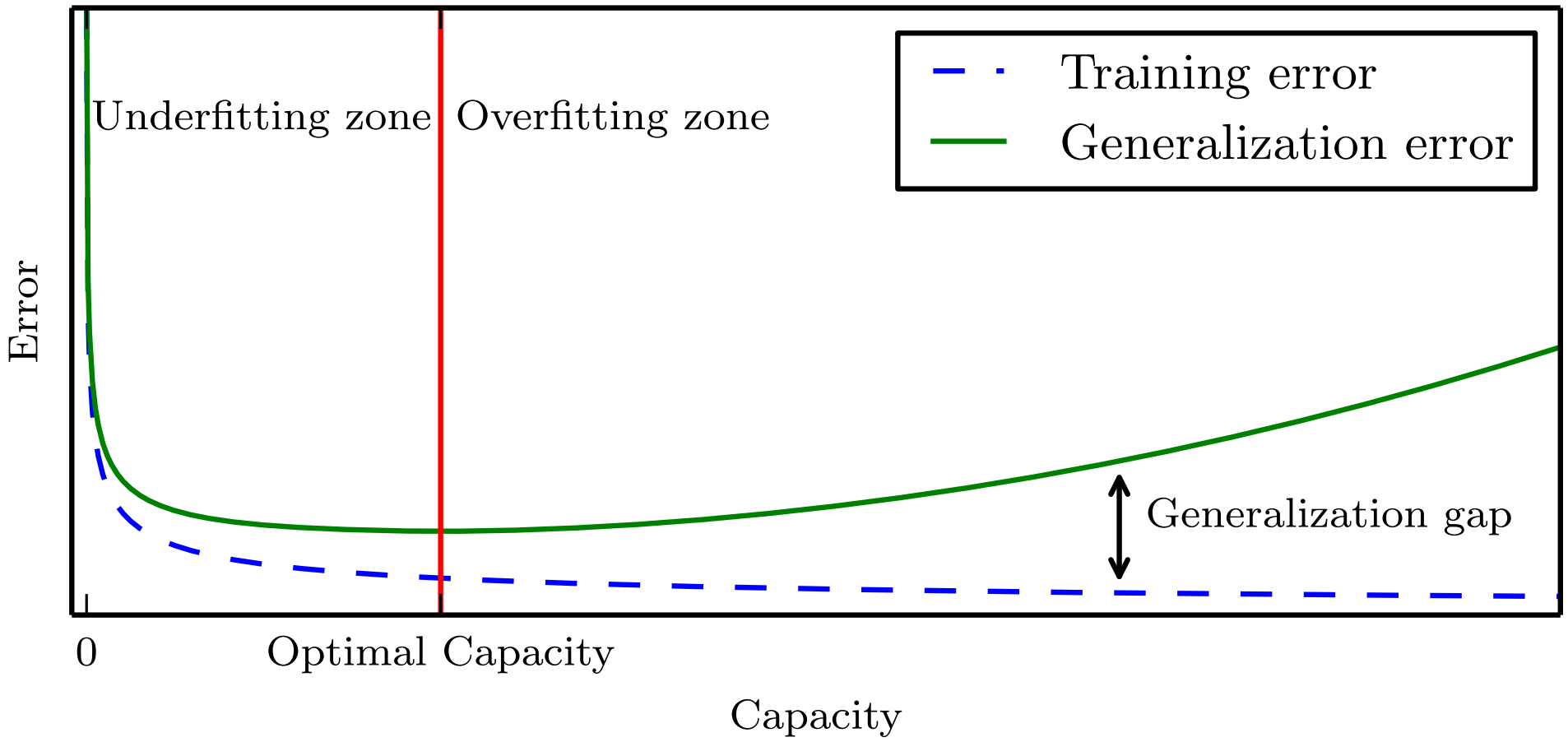


Figure 5.3

The capacity of non-parametric models is defined by the size of their training set

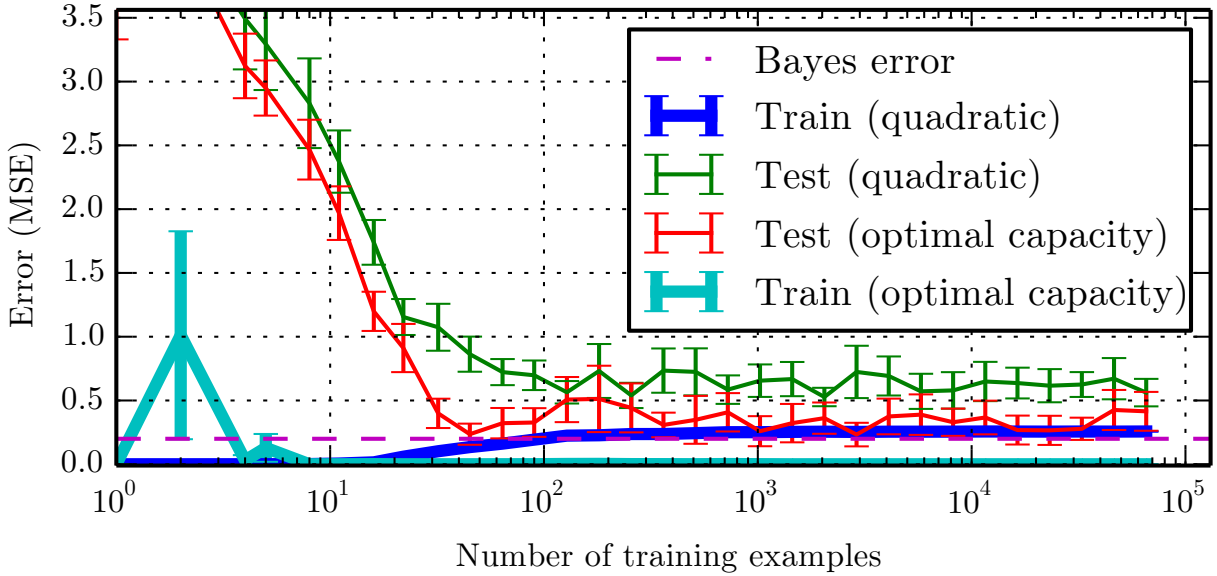
- k-nearest neighbor (KNN) regression computes its output based upon the k “nearest” training examples
- Often the best method, and certainly a baseline to beat

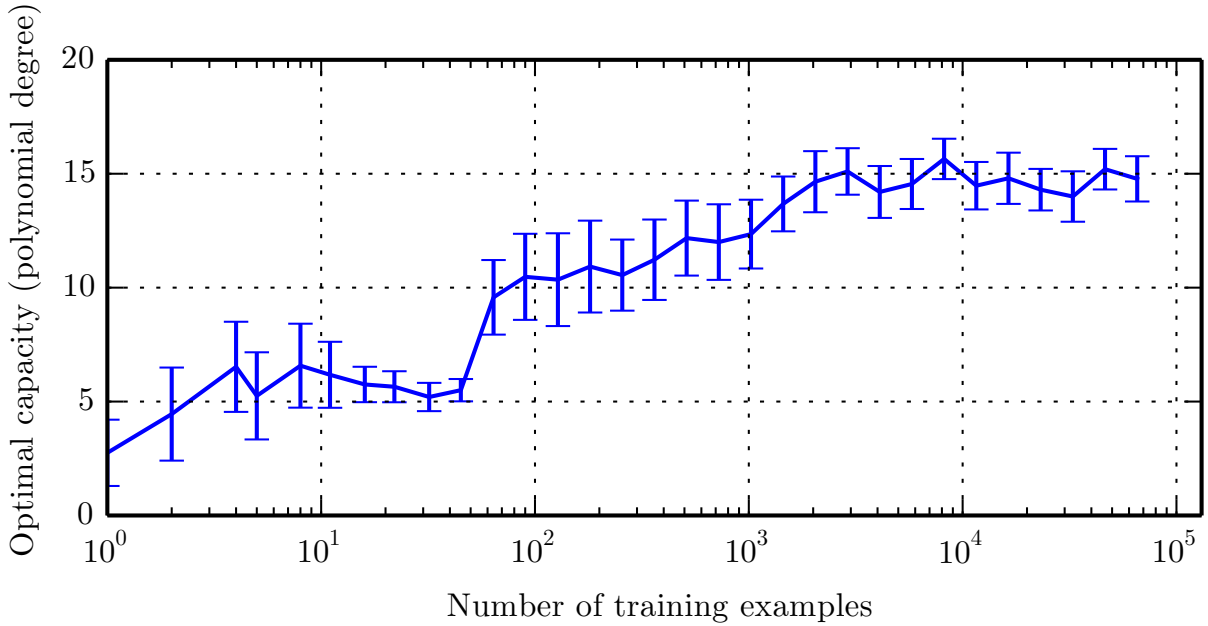
Bayes error is residual noise from confounders and observation noise

- Bayes error is the error made by an Oracle given the training set
- For example, if there is an unobserved variable that affects the labels the Oracle's predictions will be noisy

Sufficient training data are necessary to generalize well

Figure 5.4





Machine Learning has limits

- The *no free lunch* theorem: You can only obtain generalization from finitely many training examples if the algorithm searches a limited hypothesis space.
- We desire a learning algorithm to perform *generalization* and to be *stable*. It generalize if the training error on a data set will converge to the expected error. It is stable if small perturbations in the data result in only small perturbations in the output hypothesis.

Today's lecture

- Isoform estimation
  - Constraints and expectation maximization to determine maximum likelihood solutions
- Determining the significance of differential expression
  - Overdispersion caused by mixtures
  - Normalizing data and estimating negative binomial parameters
- Regularization controls model complexity
  - L1 norm (sparsity), L2 norm - (magnitude)
  - Properties of a generalizable model

FIN - Thank You