Lecture 8
Understanding Transcription
RNA-seq analysis

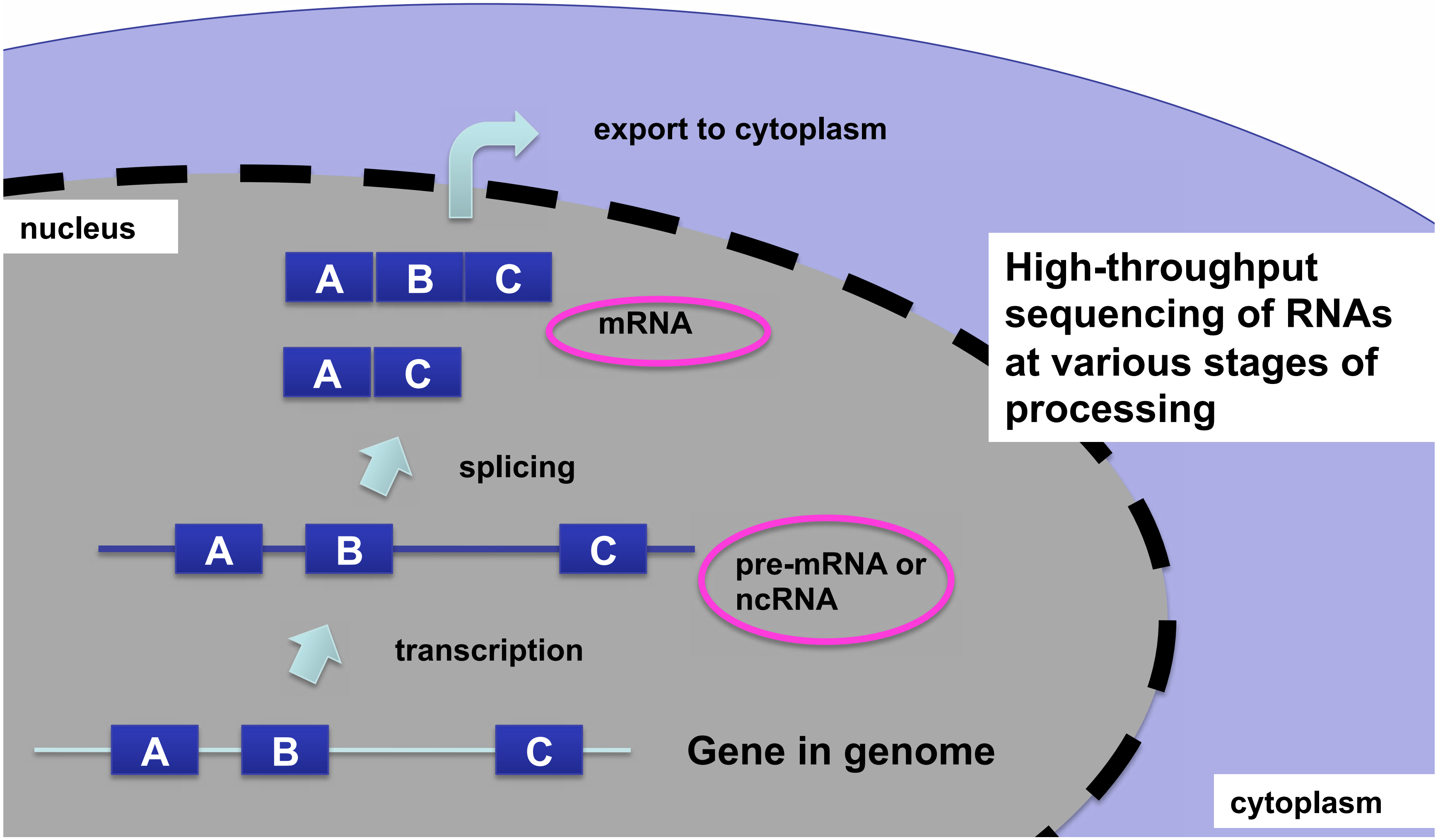
Foundations of Computational Systems Biology

David K. Gifford

Lecture 8 – RNA-seq Analysis

- RNA-seq principles
  - – How can we characterize mRNA isoform expression using high-throughput sequencing?
- Differential expression and PCA
  - – What genes are differentially expressed, and how can we characterize expressed genes?
- Single cell RNA-seq
  - – What are the benefits and challenges of working with single cells for RNA-seq?

RNA-Seq characterizes RNA molecules

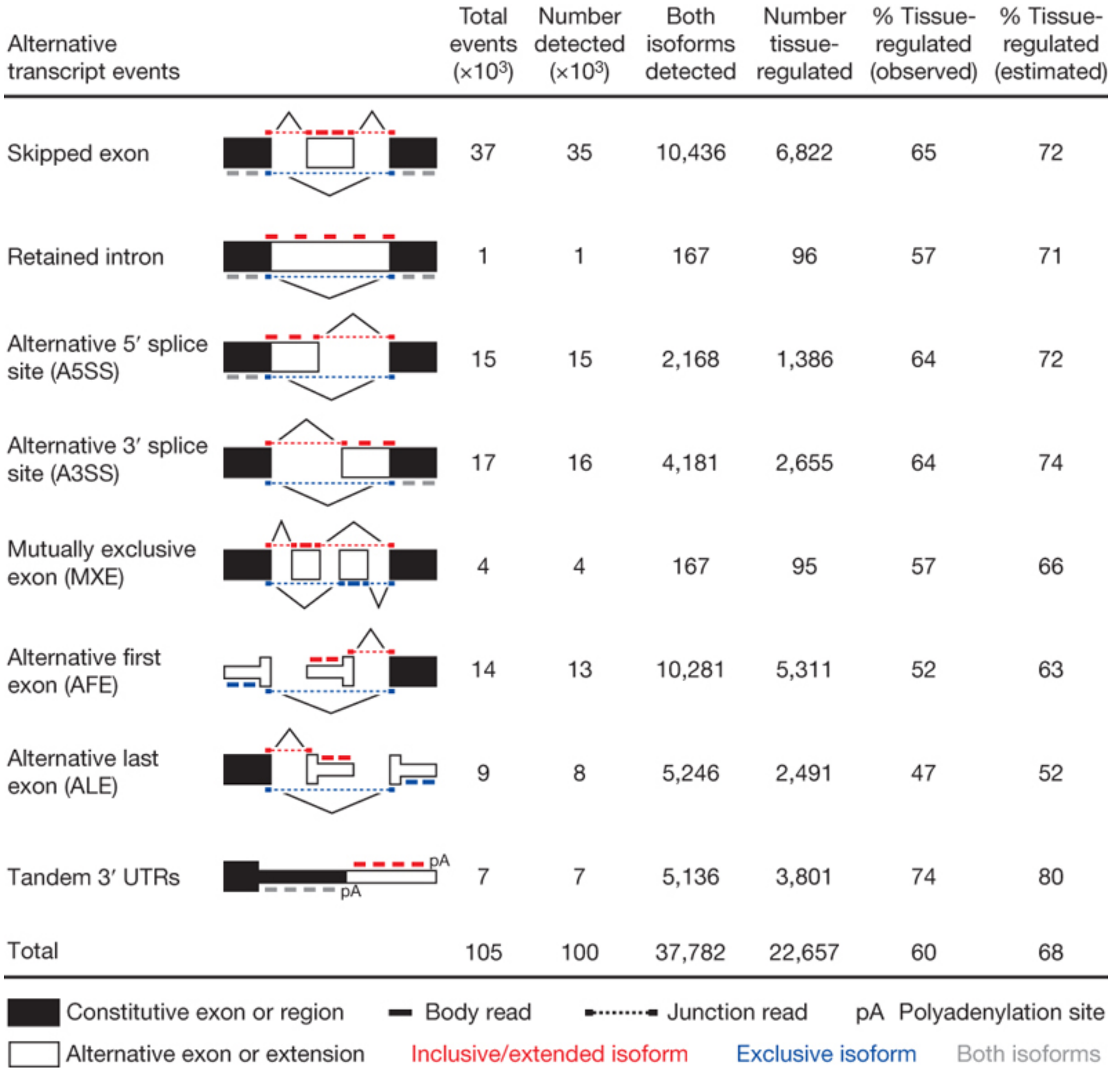


High-throughput sequencing of RNAs at various stages of processing

Courtesy of Cole Trapnell. Used with permission.

Slide courtesy Cole Trapnel

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

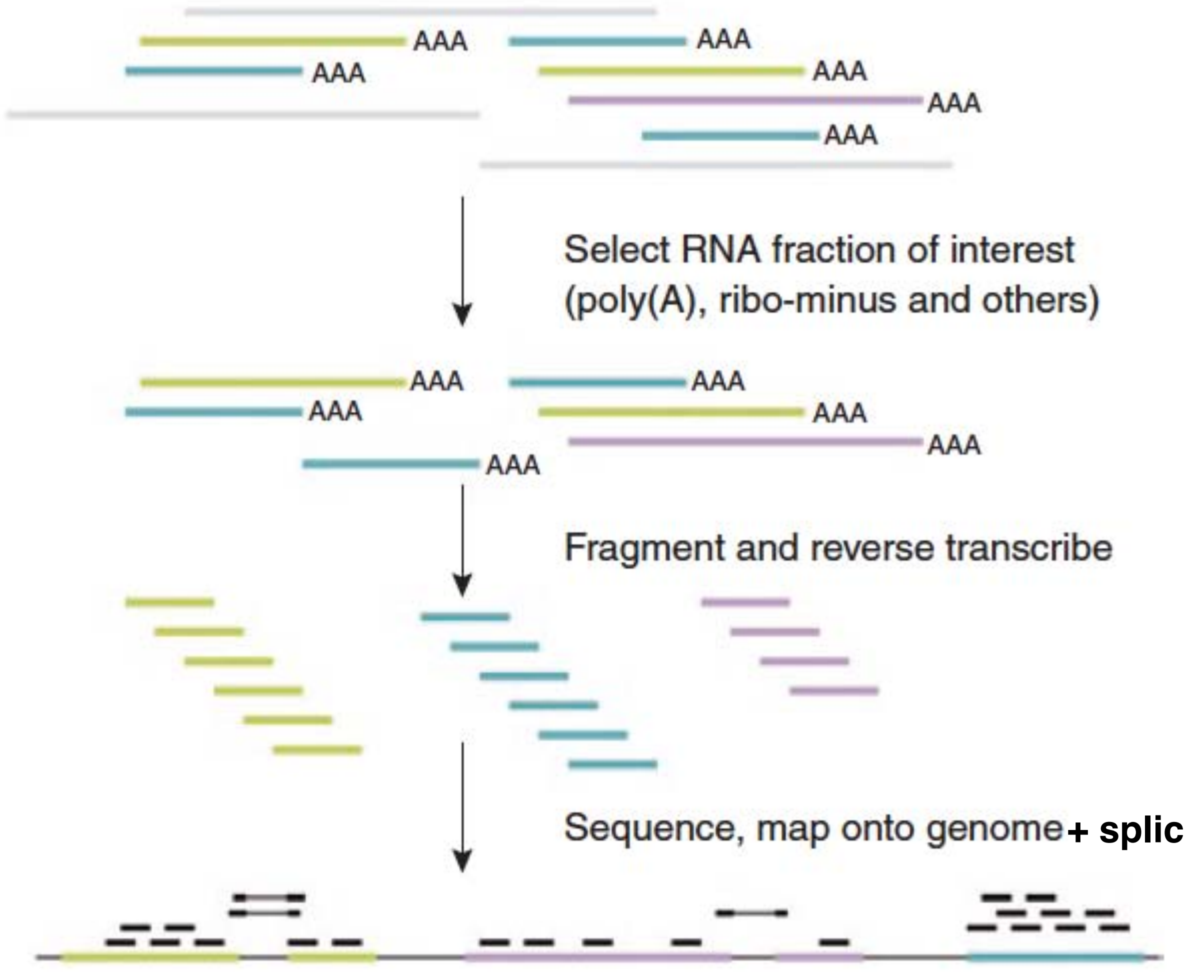
□ Alternative exon or extension Inclusive/extended isoform Exclusive isoform Both isoforms

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

Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Wang, Eric T., Rickard Sandberg, et al. "Alternative Isoform Regulation in Human Tissue Transcriptomes." *Nature* 456, no. 7221 (2008): 470-6.

RNA-Seq: millions of short reads from fragmented mRNA

Extract RNA from cells/tissue



Select RNA fraction of interest (poly(A), ribo-minus and others)

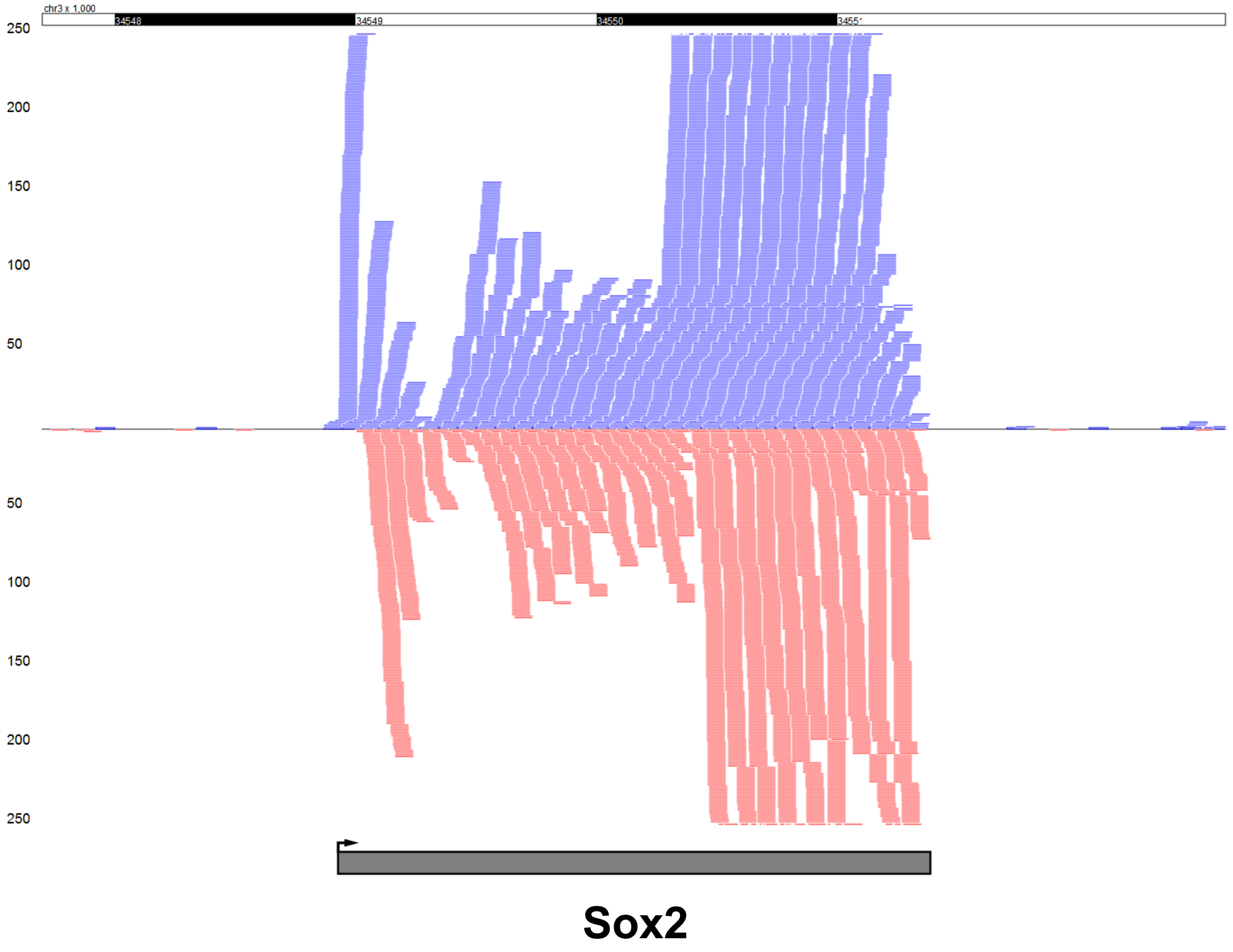
Fragment and reverse transcribe

Sequence, map onto genome + splice junctions

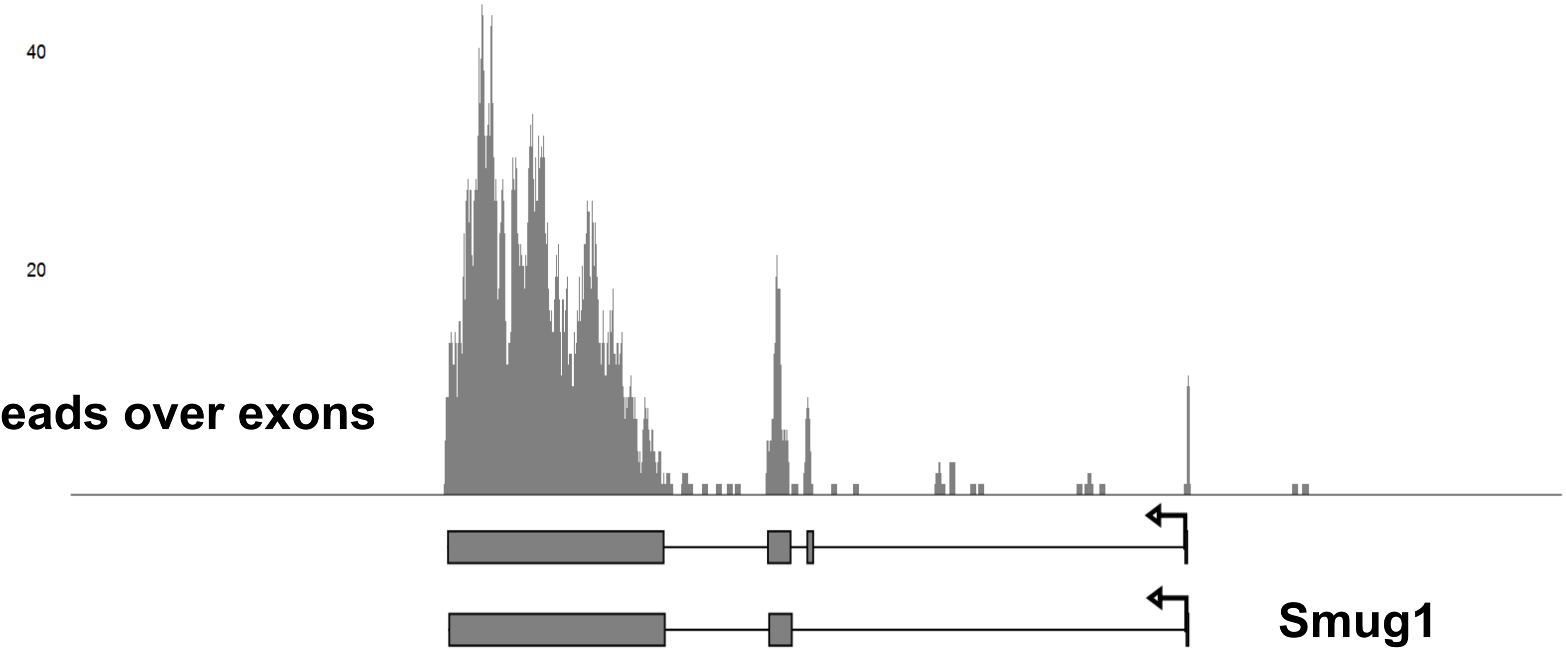
Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Pepke, Shirley, Barbara Wold, et al. "Computation for ChIP-seq and RNA-seq Studies." *Nature Methods* 6 (2009): S22-32.

Pepke et. al. *Nature Methods* 2009

Mapping RNA-seq reads to a reference genome reveals expression



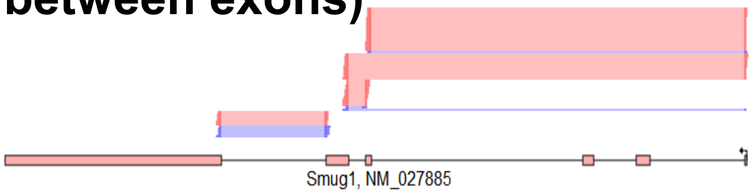
RNA-seq reads map to exons and across exons



Reads over exons

Smug1

Junction reads (split between exons)

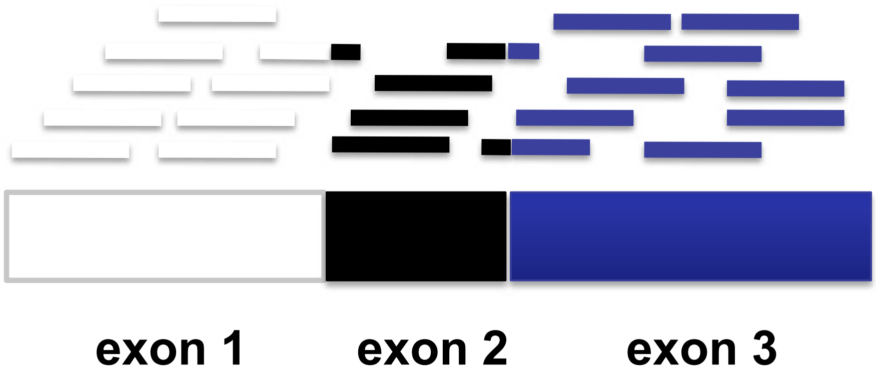


Two major approaches to RNA-seq analysis

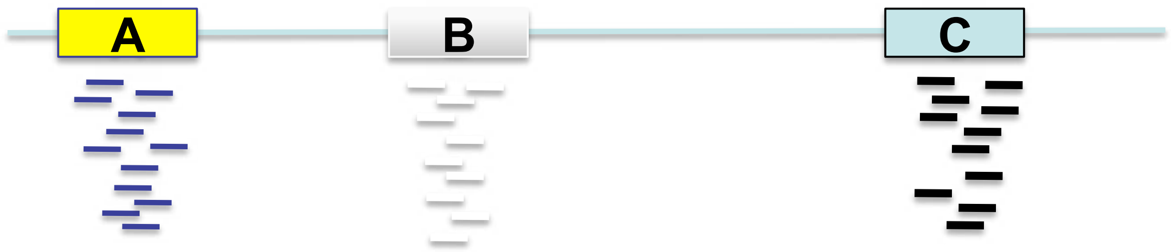
1. Assemble reads into transcripts. Typical issues with coverage and correctness.
2. Map reads to reference genome and identify isoforms using constraints

- Goal is to quantify isoforms and determine significance of differential expression
- Common RNA-seq expression metrics are Reads per killobase per million reads (RPKM) or Fragments per killobase per million (FPKM)

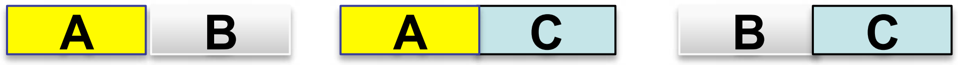
Short sequencing reads, randomly sampled from a transcript



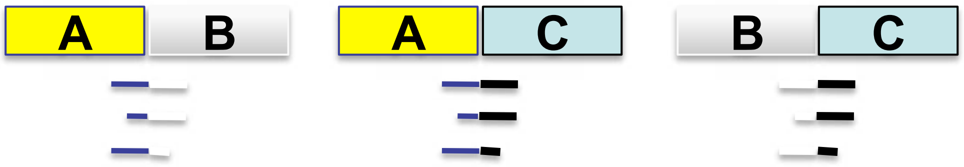
Aligned reads reveal isoform possibilities



identify candidate exons via genomic mapping



Generate possible pairings of exons

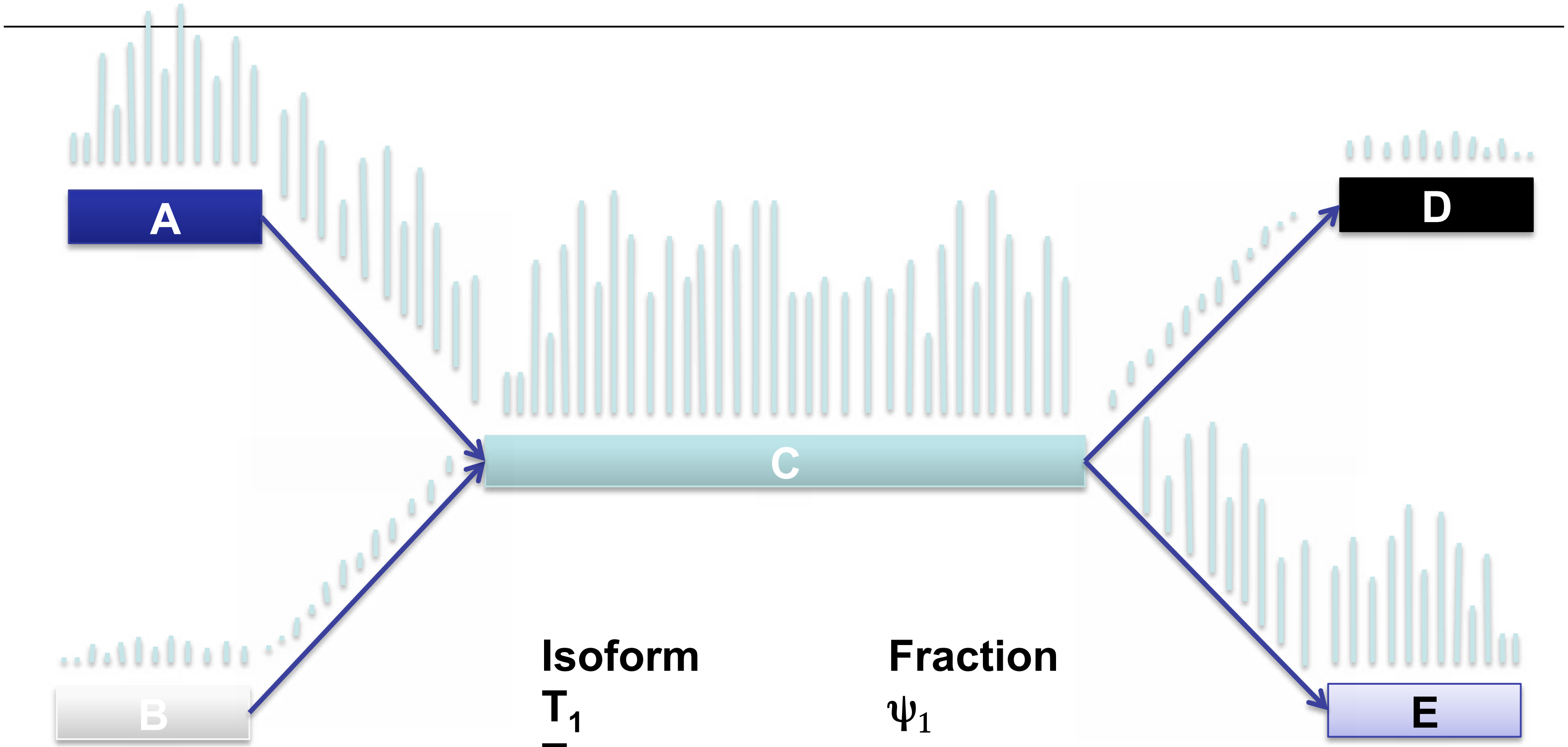


Align reads to possible junctions

Courtesy of Cole Trapnell. Used with permission.

Slide courtesy Cole Trapnell

We can use mapped reads to learn the isoform mixture ψ

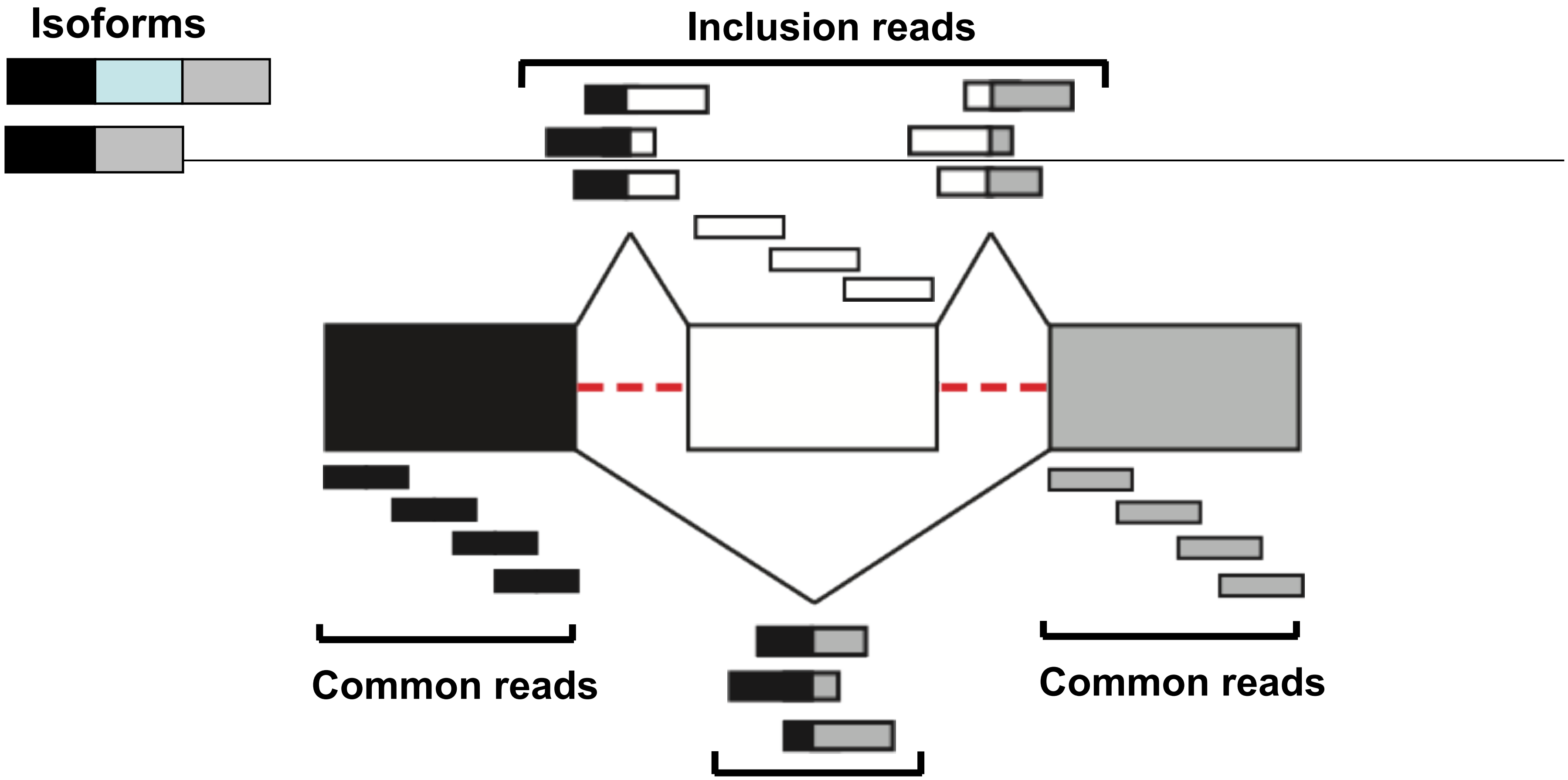


Isoform	Fraction
T_1	ψ_1
T_2	ψ_2
T_3	ψ_3
T_4	ψ_4

Courtesy of Cole Trapnell. Used with permission.

Slide courtesy Cole Trapnell

Detecting alternative splicing from mRNA-Seq data



Isoforms

Inclusion reads

Common reads

Common reads

Exclusion reads

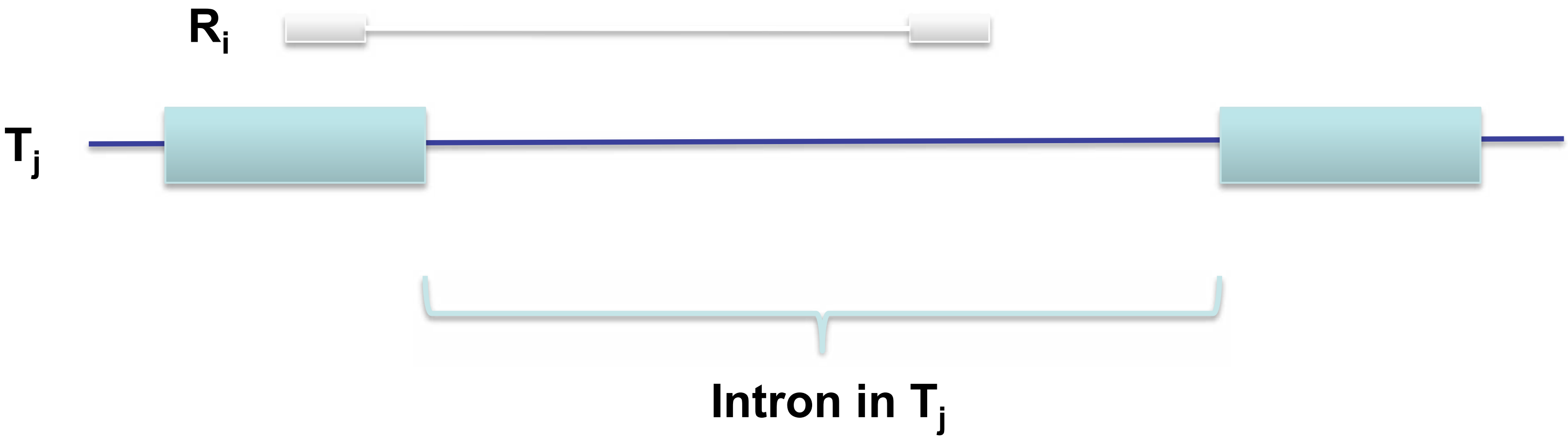
Given a set of reads, estimate:

Ψ = **Distribution of isoforms**

$P(R_i \mid T{=}T_j)$ – Excluded reads

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

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



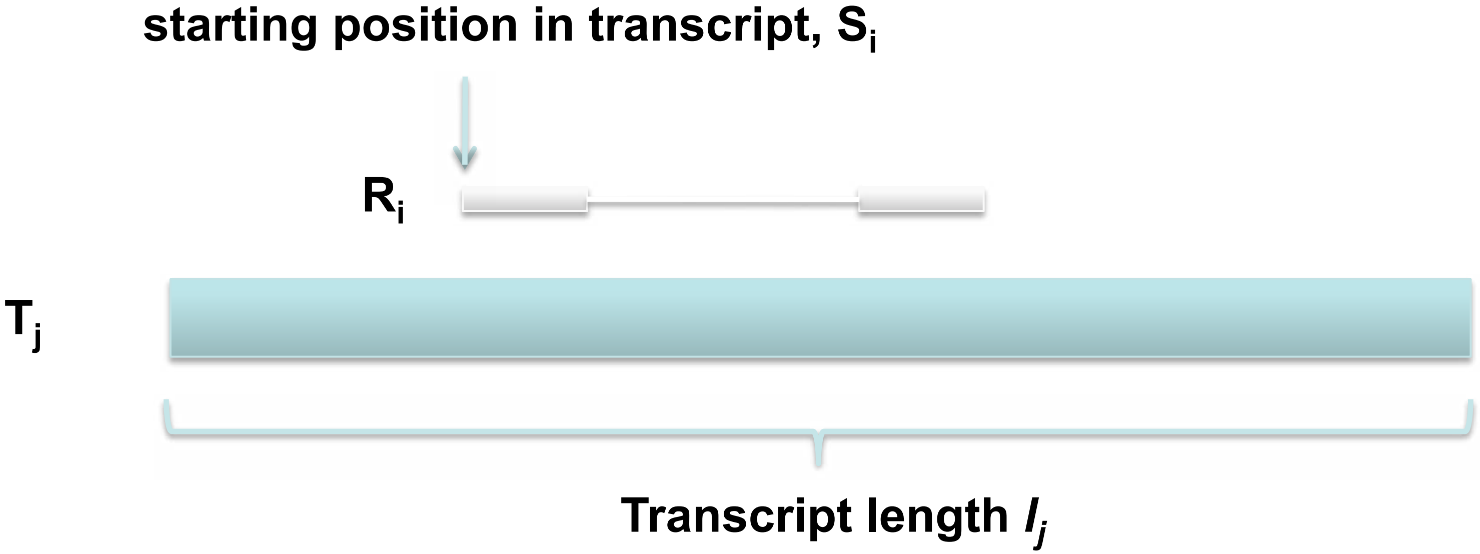
Courtesy of Cole Trapnell. Used with permission.

Slide courtesy Cole Trapnell

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

Cufflinks assumes that fragmentation is roughly uniform. The probability of observing a fragment starting at a specific position *S_i* in a transcript of length *l_j* is:

$$P(S = S_i \mid T = T_j) = \frac{1}{l_j}$$



Courtesy of Cole Trapnell. Used with permission.

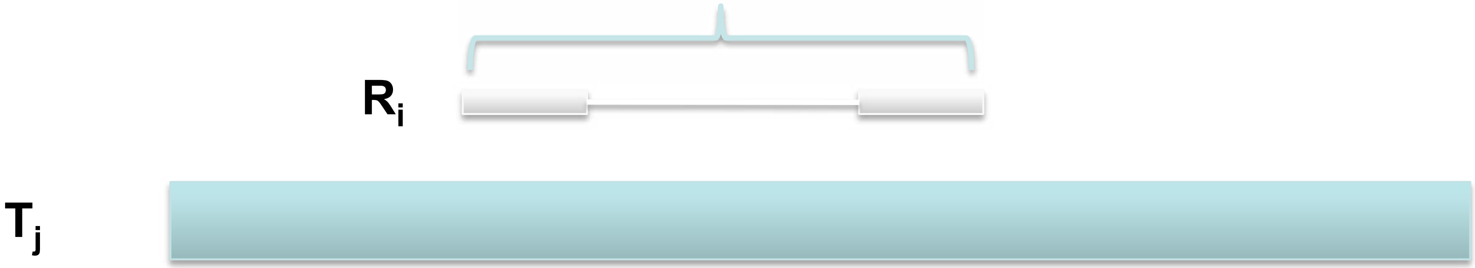
Slide courtesy Cole Trapnell

P(R_i | 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 $l_j(R_i)$



Courtesy of Cole Trapnell. Used with permission.

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 $R_1,\ldots,R_m$

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

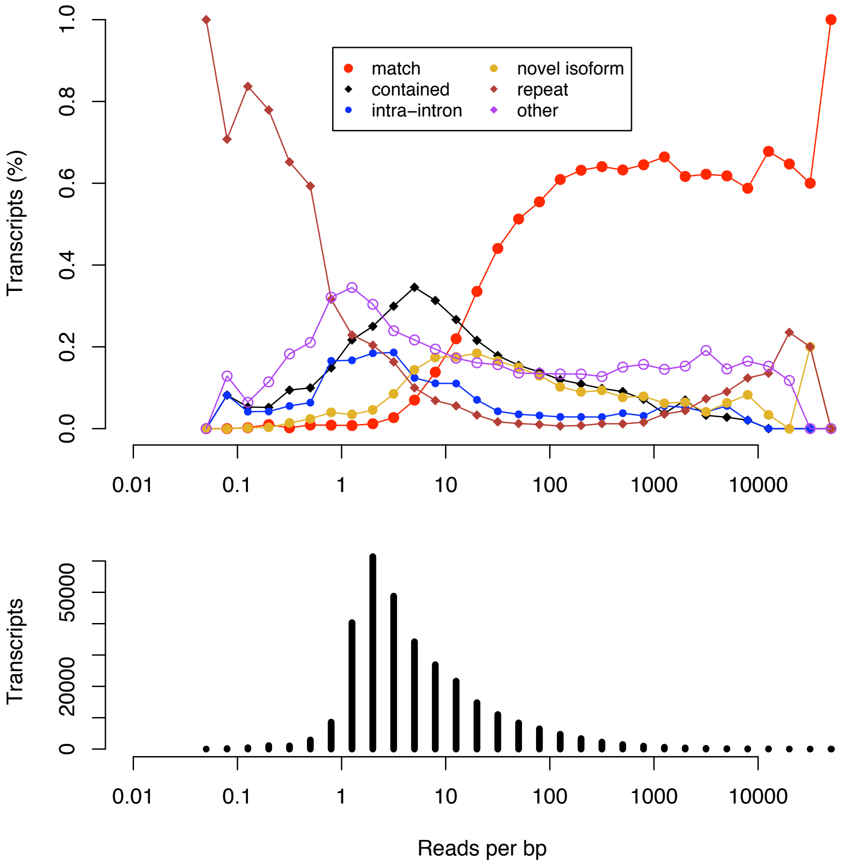
$$L(\Psi \mid R) \propto P(R \mid \Psi) P(\Psi)$$

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

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

Case study: myogenesis

Transcript categories, by coverage

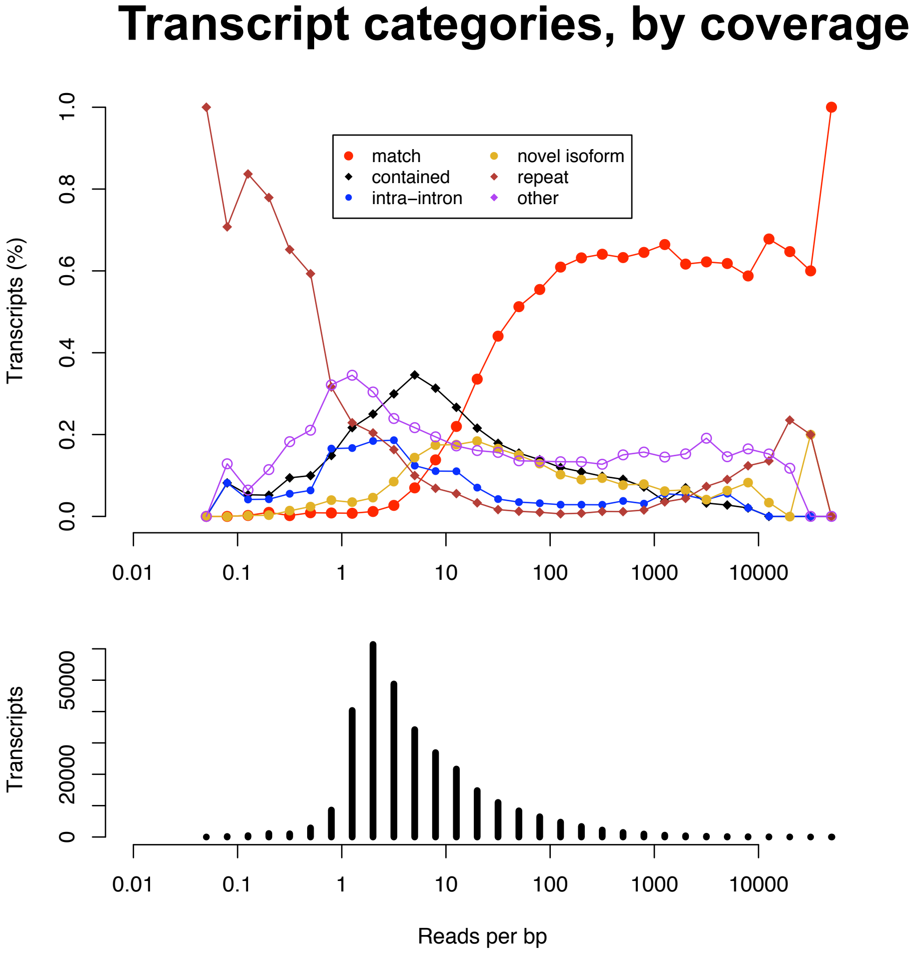


Courtesy of Cole Trapnell. Used with permission.

- Cufflinks identified 116,839 distinct transcribed fragments (transfrags)
- Nearly **70%** of the reads in 14,241 matching transcripts
- Tracked 8,134 transfrags across all time points, **5,845 complete matches** to UCSC/ Ensembl/VEGA
- Tracked **643** new isoforms of known genes across all points

Slide courtesy Cole Trapnell

Case study: myogenesis



Courtesy of Cole Trapnell. Used with permission.

- ~25% of transcripts have light sequence coverage, and are fragments of full transcripts

- Intronic reads, repeats, and other artifacts are numerous, but account for less than 5% of the assembled reads.

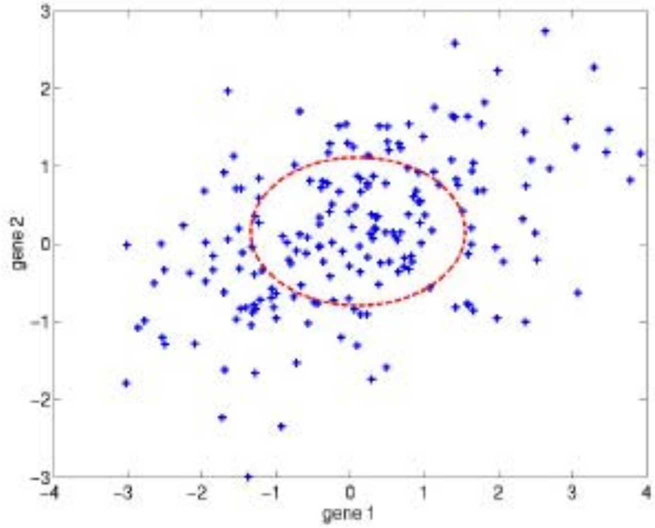
Slide courtesy Cole Trapnell

Lecture 8 – RNA-seq Analysis

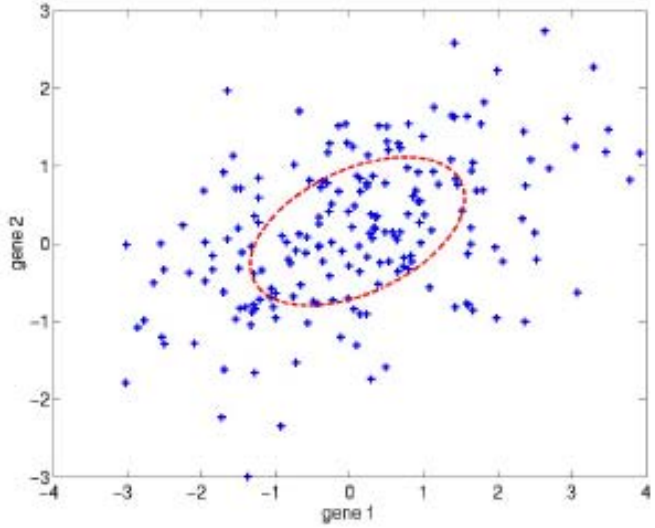
- RNA-seq principles
  - How can we characterize mRNA isoform expression using high-throughput sequencing?
- Differential expression and PCA
  - What genes are differentially expressed, and how can we characterize expressed genes?
- Single cell RNA-seq
  - What are the benefits and challenges of working with single cells for RNA-seq?

Statistical tests: example

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



null hypothesis



alternative hypothesis

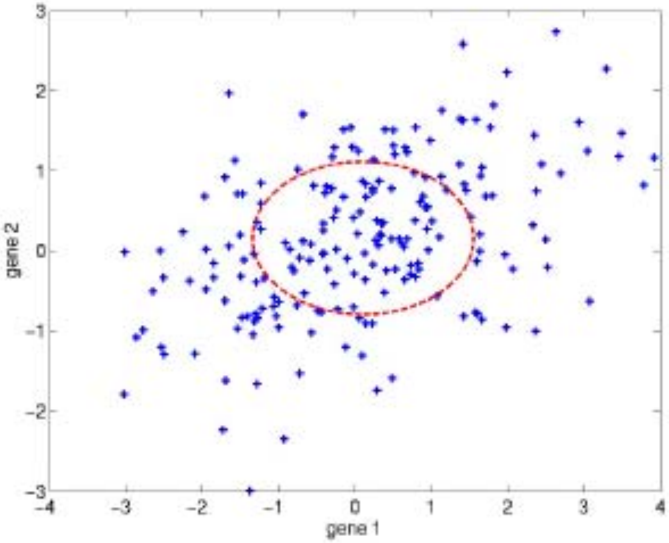
- We need to find a way of testing whether this difference is significant

Degrees of freedom

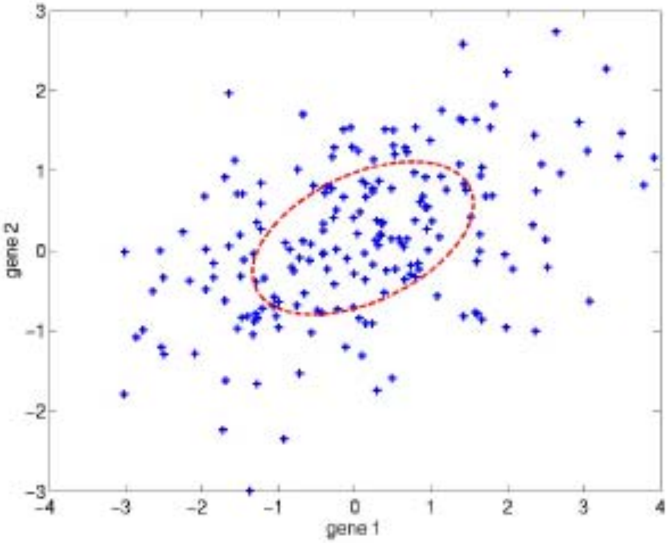
- How many degrees of freedom do we have in the two models?

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

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



H_0



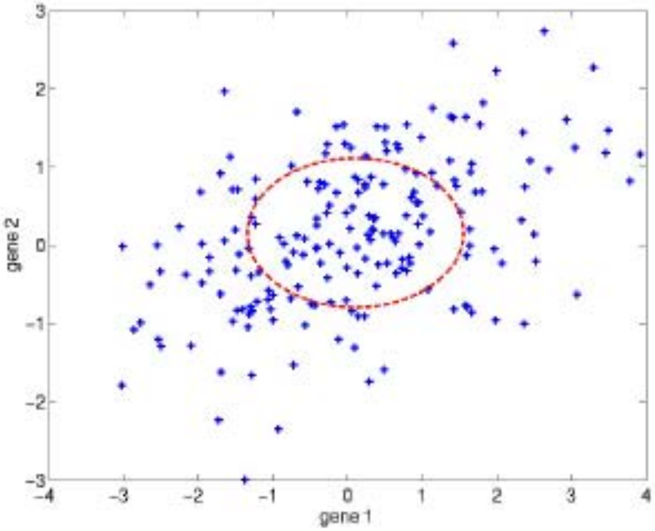
H_1

Degrees of freedom

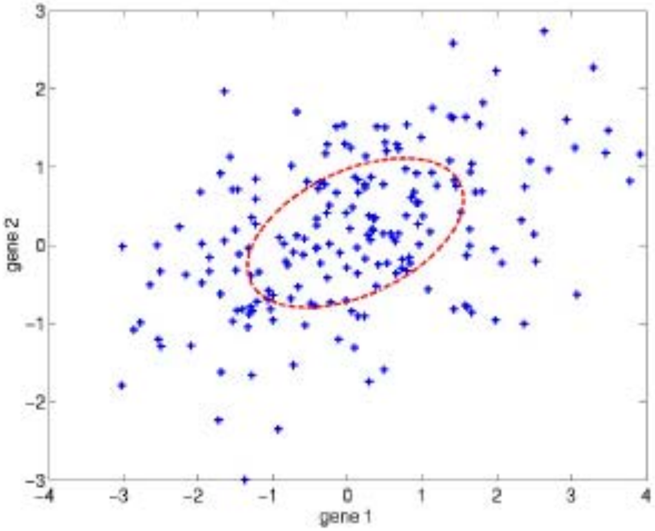
- How many degrees of freedom do we have in the two models?

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

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



H_0



H_1

- The observed data overwhelmingly supports H_1

Test statistic

- Likelihood ratio statistic

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

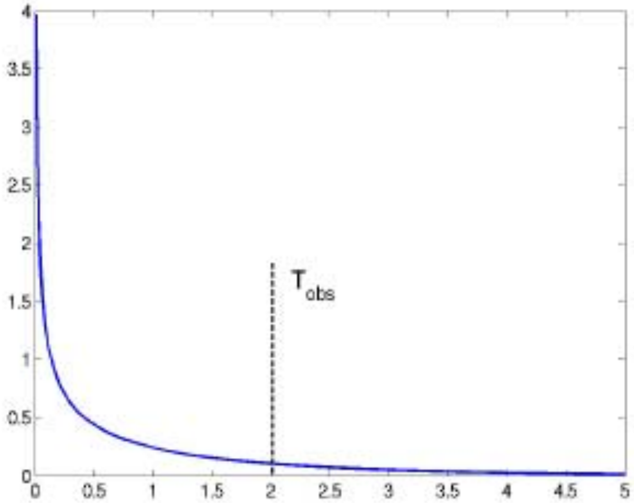
  Larger values of T imply that the model corresponding to the null hypothesis H_0 is much less able to account for the observed data

- 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

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

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)

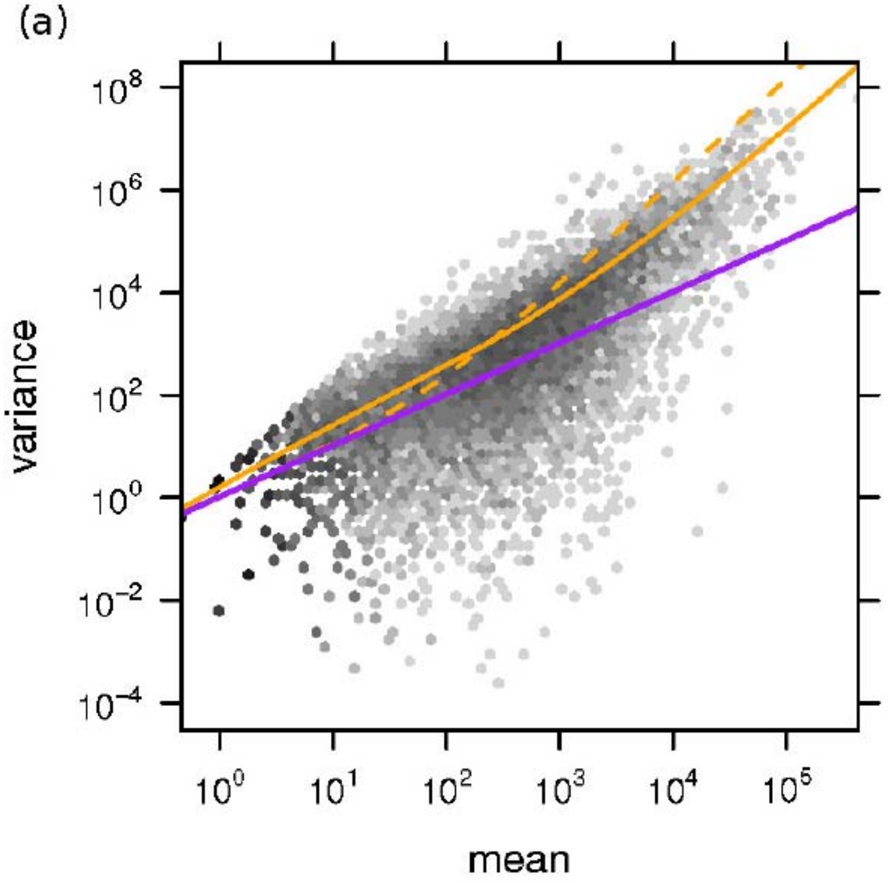
- i gene or isoform p condition
- j sample (experiment) p(j) condition of sample j
- m number of samples
- K_{ij} number of counts for isoform i in experiment j
- q_{ip} Average scaled expression for gene i condition p

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

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

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

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



Orange Line – DESeq
Dashed Orange – edgeR
Purple - Poission

Courtesy of the authors. License: CC-BY.
Source: Anders, Simon, and Wolfgang Huber. "Differential Expression Analysis for Sequence Count Data." *Genome Biology* 11, no. 10 (2010): R106.

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

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

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

$$P\left(H0\right) \approx 1 - ChiSquaredCDF\left(T_i \mid dof\right)$$

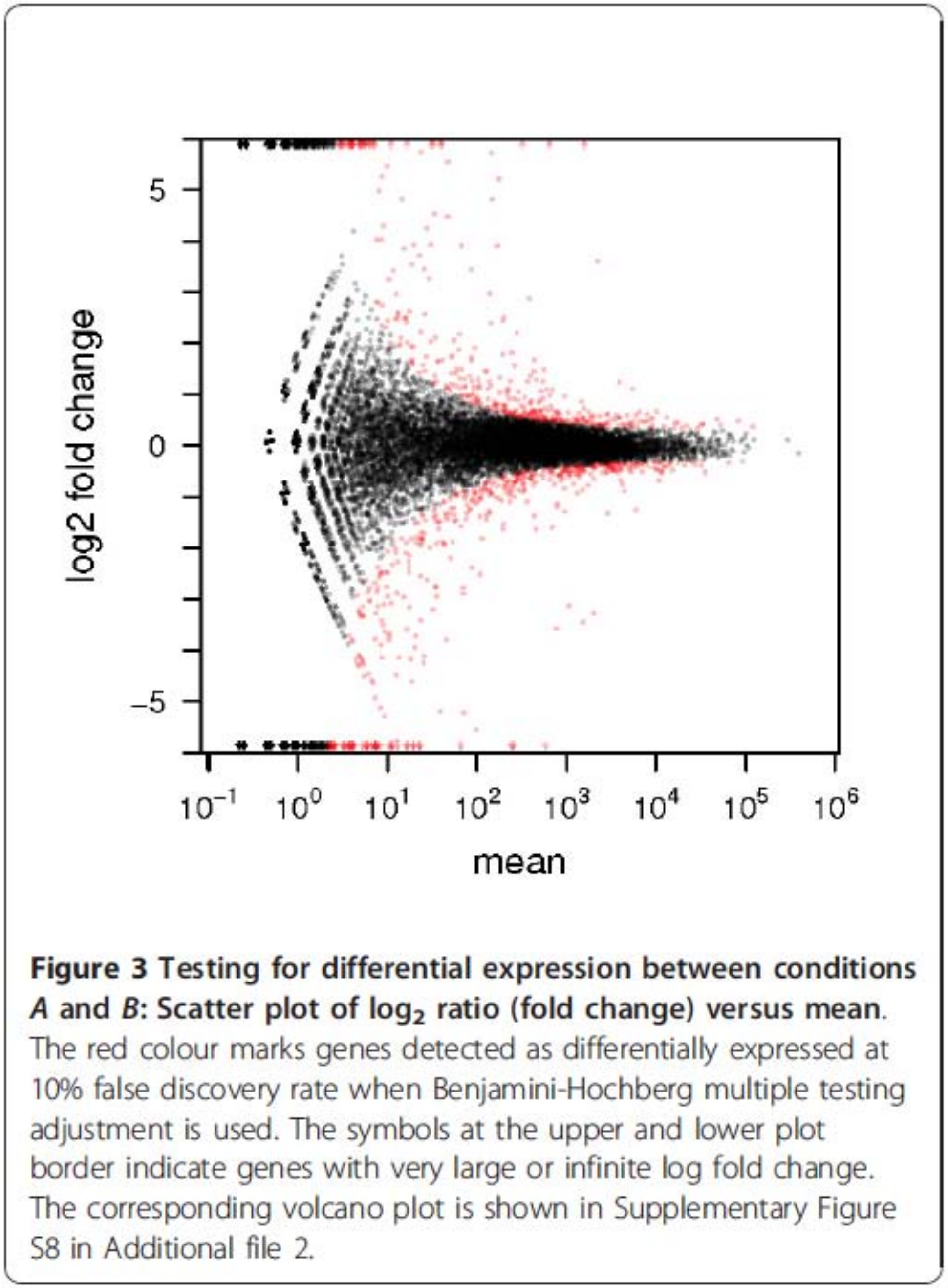


Figure 3 Testing for differential expression between conditions *A* and *B*: Scatter plot of $\log_2$ ratio (fold change) versus mean. The red colour marks genes detected as differentially expressed at 10% false discovery rate when Benjamini-Hochberg multiple testing adjustment is used. The symbols at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

Source: Anders, Simon, and Wolfgang Huber. "Differential Expression Analysis for Sequence Count Data." *Genome Biology* 11, no. 10 (2010): R106.

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

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

$$P(k) = \frac{\begin{pmatrix} n1 \\ k \end{pmatrix}\begin{pmatrix} N - n1 \\ n2 - k \end{pmatrix}}{\begin{pmatrix} N \\ n2 \end{pmatrix}}$$

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

0.017

0.020

Principle Component Analysis (PCA)

- How can we discover vector components that describe our data?
  1. To discover hidden factors that explain the data
  2. Similar to cluster centroids
  3. To reduce the dimensionality of our data

Multi-Variate Gaussian Review

- Recall multi-variate Gaussians:

$$\begin{aligned} Z_i &\sim N(0,1) && (5)\\ X &= AZ + \mu && (6)\\ \Sigma &= E[(X-\mu)(X-\mu)^T] && (7)\\ &= E[(AZ)(AZ)^T] && (8)\\ &= E[AZZ^TA^T] && (9)\\ &= AE[ZZ^T]A^T && (10)\\ &= AA^T && (11) \end{aligned}$$

- A multivariate Gaussian model

$$\begin{aligned} p(x|\theta) &= \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{ -\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) \} && (12)\\ X &\sim N(\mu,\Sigma) && (13) \end{aligned}$$

where μ is the mean vector and Σ is the covariance matrix

Principle Component Analysis (PCA)

- Consider the variance of X projected onto vector v

$$\begin{aligned}
Var(v^T X) &= E[(v^T X)^2] - E[v^T X]^2 && (14)\\
&= v^T E[XX^T]v - v^T E[X]E[X^T]v && (15)\\
&= v^T (E[XX^T] - E[X]E[X^T])v && (16)\\
&= v^T \Sigma v && (17)
\end{aligned}$$

- We would like to pick v_i to maximize the variance with the constraint $v_i^T v_i = 1$. Each v_i will be orthogonal to all of the other v_i
- The v_i are called the eigenvectors of Σ and λ_i^2 are the eigenvalues:

$$\begin{aligned}
\Sigma v_i &= \lambda_i^2 v_i && (18)\\
v_i^T \Sigma v_i &= v_i^T \lambda_i^2 v_i && (19)\\
v_i^T \Sigma v_i &= \lambda_i^2 v_i^T v_i && (20)\\
v_i^T \Sigma v_i &= \lambda_i^2 && (21)
\end{aligned}$$

Principle Component Analysis (PCA)

- How do we find the eigenvectors v_i?
- We use singular value decomposition to decompose Σ into an orthogonal rotation matrix U and a diagonal scaling matrix S:

$$\begin{aligned} \Sigma &= USU^T & (22) \\ \Sigma U &= (USU^T)U & (23) \\ &= US & (24) \end{aligned}$$

- The columns of U are the v_i, and S is the diagonal matrix of eigenvalues λ_i^2

Principle Component Analysis (PCA)

- How do we interpret eigenvectors and eigenvalues with respect to our orginal transform A?

$$X = AZ + \mu \tag{25}$$

- A is:

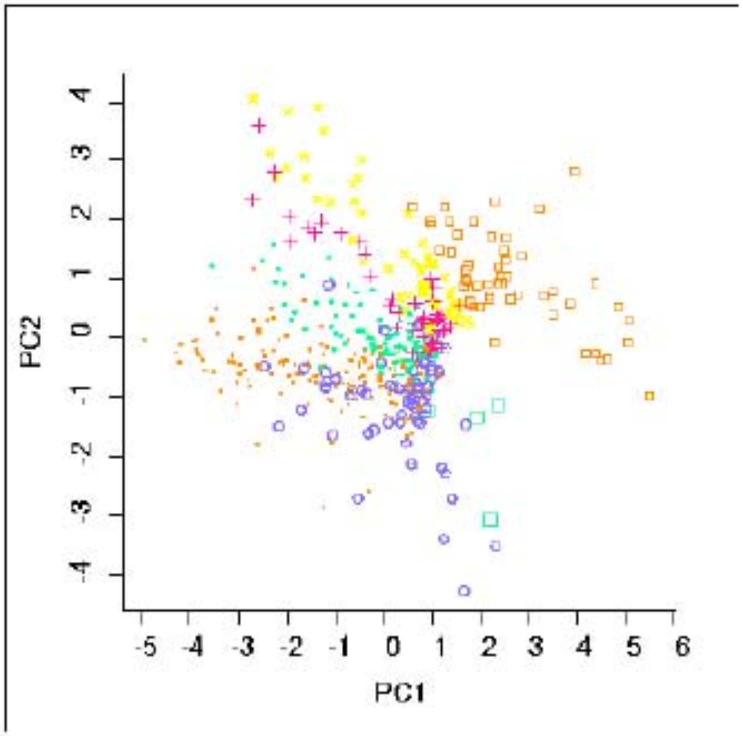
$$A = US^{1/2} \tag{26}$$
$$\Sigma = AA^T \tag{27}$$
$$\Sigma = USU^T \tag{28}$$

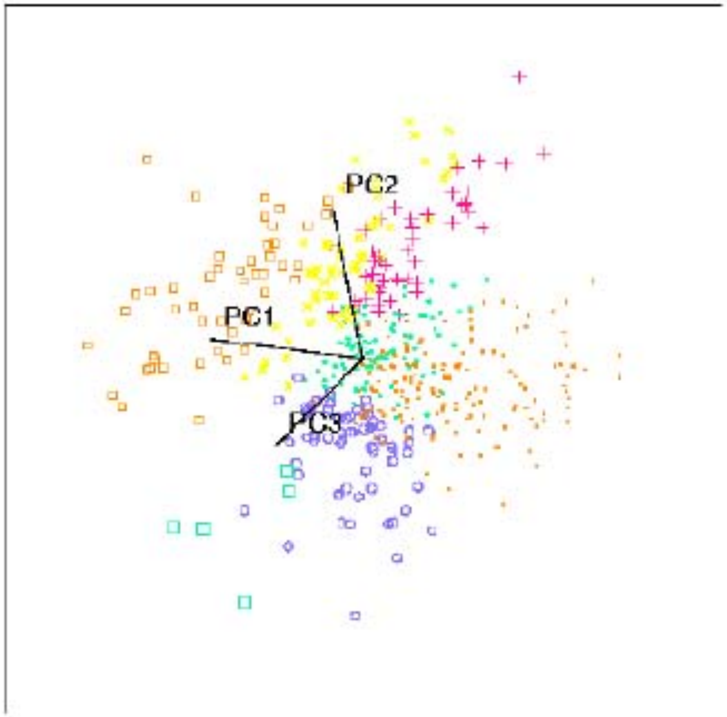
- Thus, the transformation A scales by $S^{1/2}$ and rotates by U independent Gaussians to make X

$$Z_i \sim N(0,1) \tag{29}$$
$$X = US^{1/2}Z + \mu \tag{30}$$

Example PCA Analysis

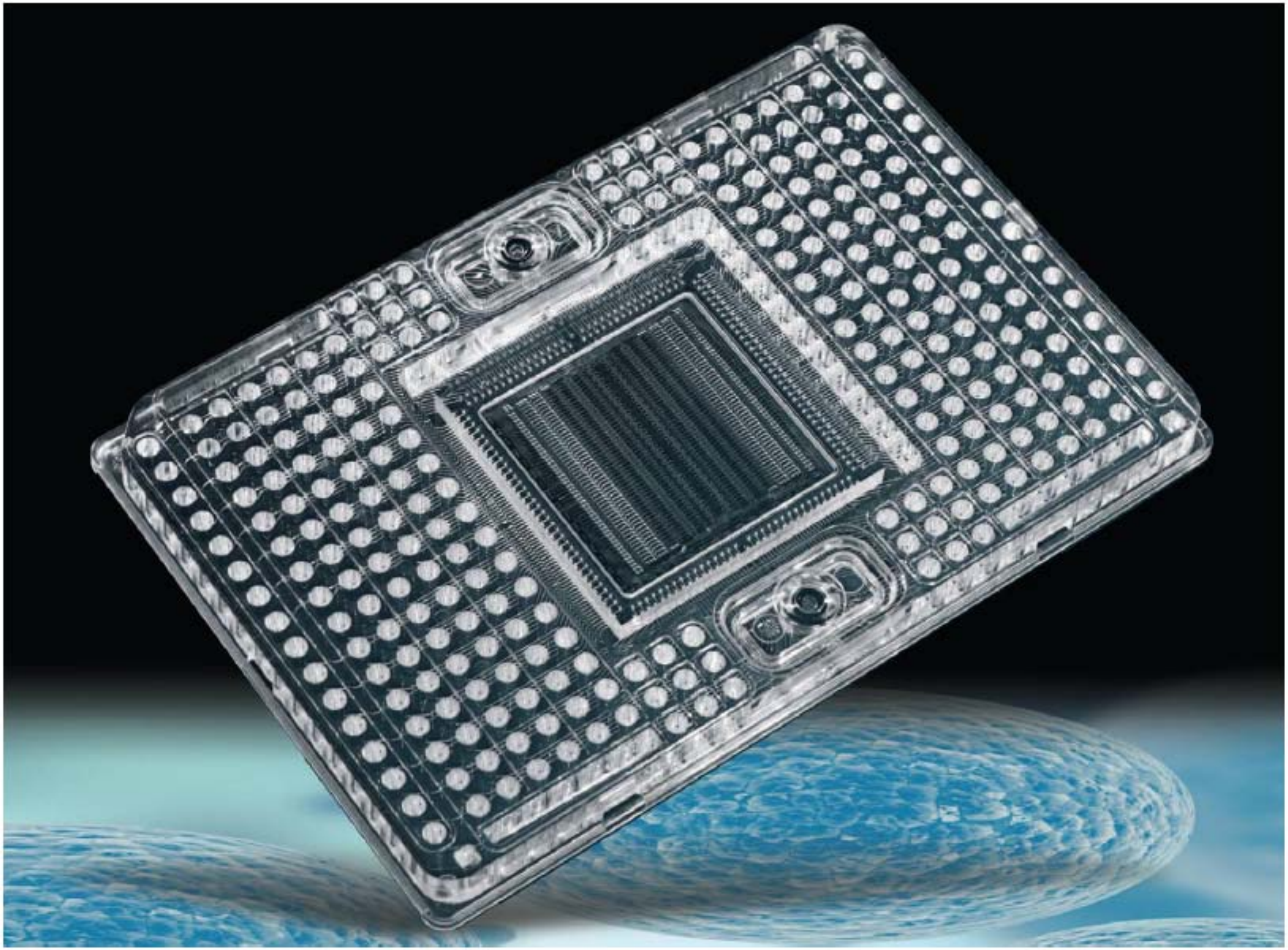
477 sporulation genes classified into seven patterns resovled by PCA





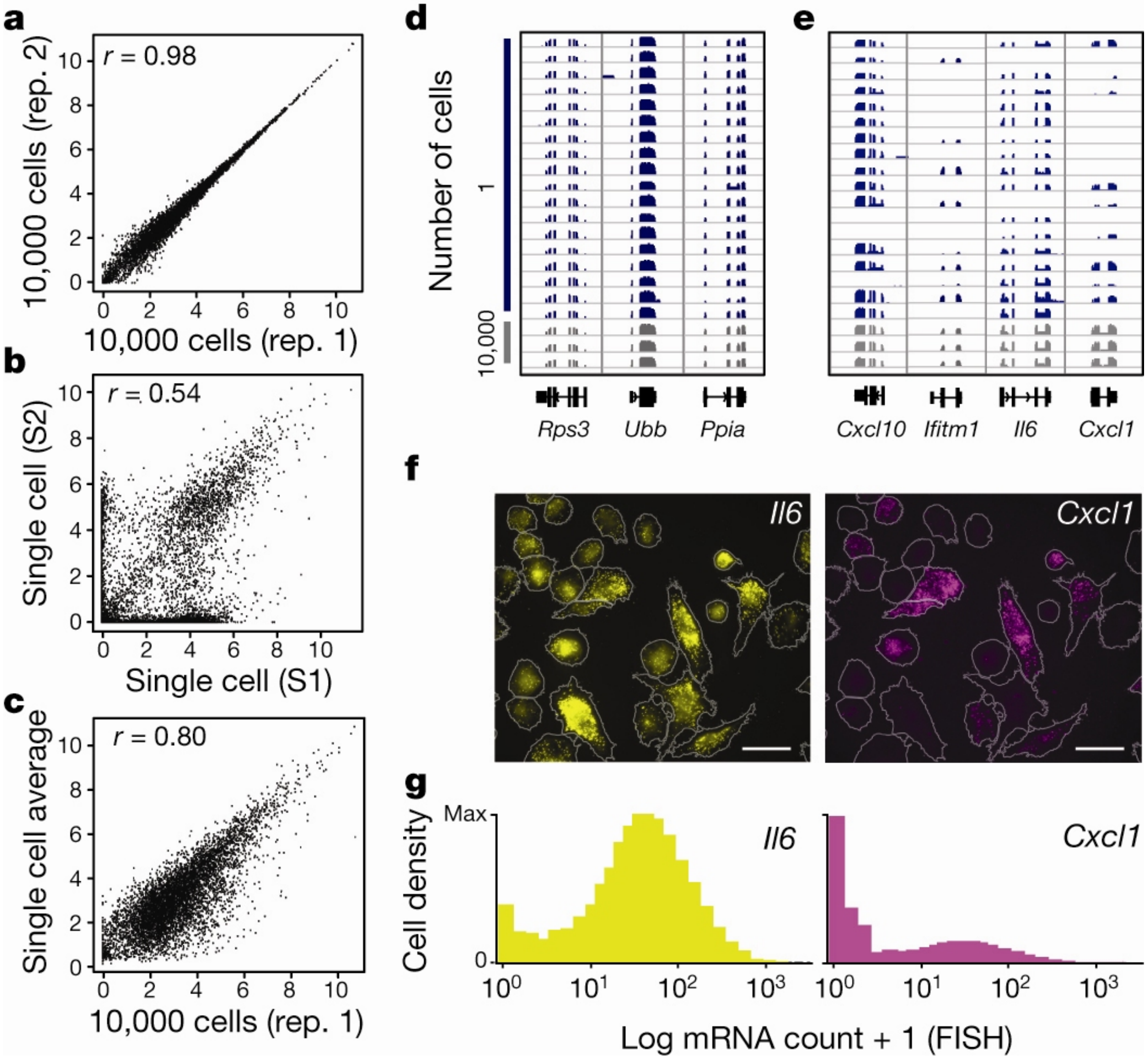
Lecture 8 – RNA-seq Analysis

- RNA-seq principles
  - How can we characterize mRNA isoform expression using high-throughput sequencing?
- Differential expression and PCA
  - What genes are differentially expressed, and how can we characterize expressed genes?
- Single cell RNA-seq
  - What are the benefits and challenges of working with single cells for RNA-seq?



Courtesy of Fluidigm Corporation. Used with permission.

Single-cell RNA-Seq of LPS-stimulated bone-marrow-derived dendritic cells reveals extensive transcriptome heterogeneity.

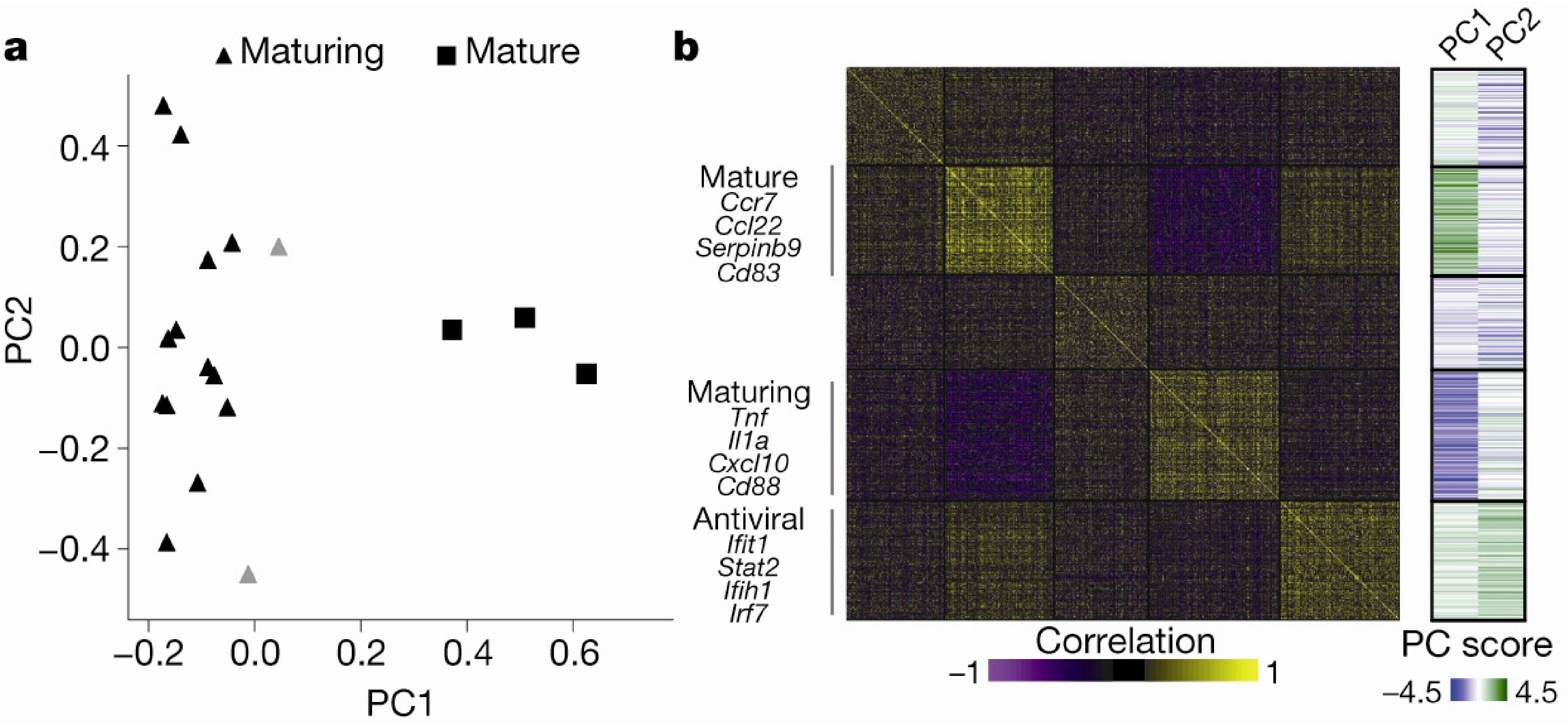


AK Shalek *et al. Nature* 000, 1-5 (2012) doi:10.1038/nature12172

Courtesy of Macmillan Publishers Limited. Used with permission.

Source: Shalek, Alex K., Rahul Satija, et al. "Single-cell Transcriptomics Reveals Bimodality in Expression and Splicing in Immune Cells." *Nature* (2013).

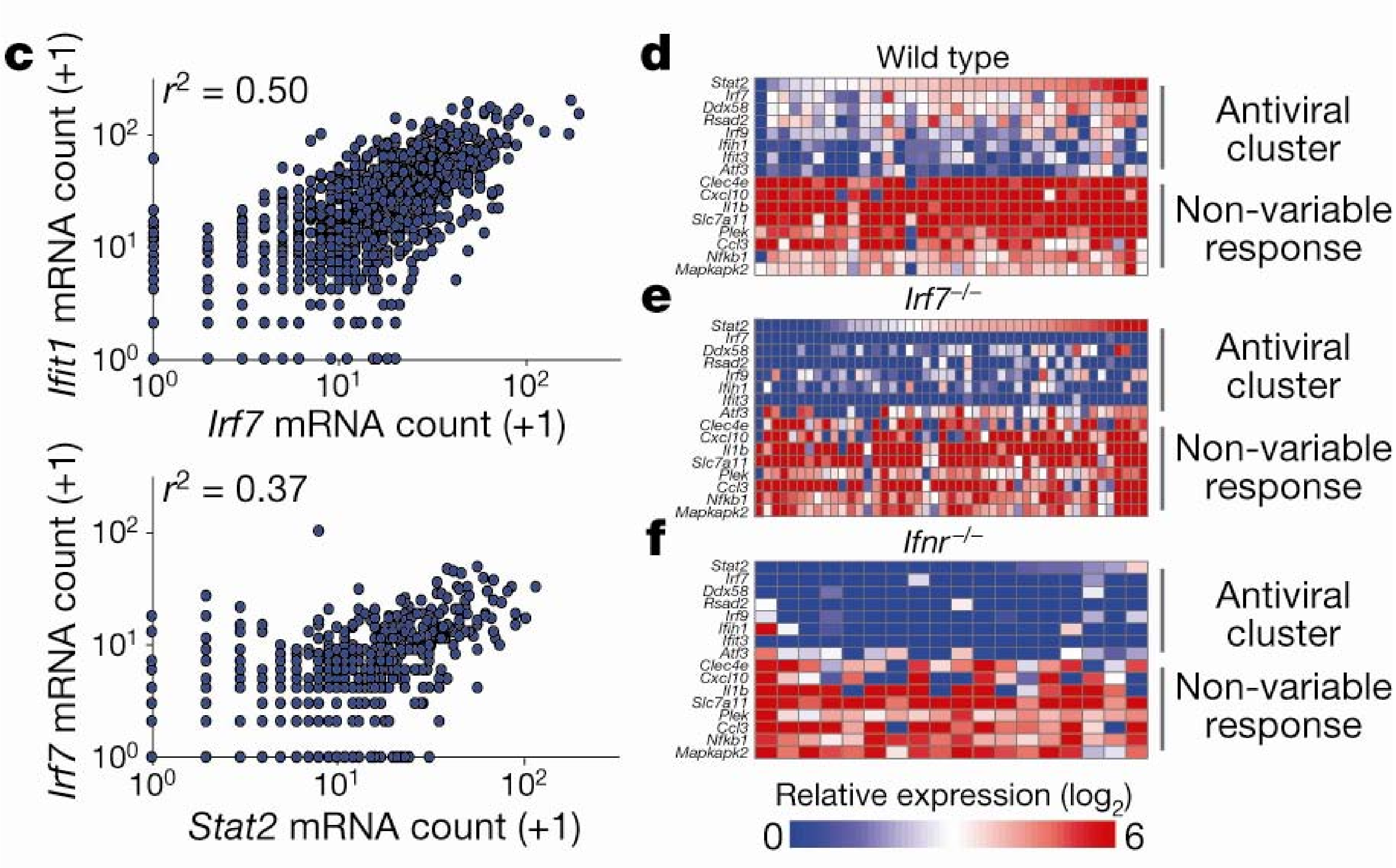
Analysis of co-variation in single-cell mRNA expression levels reveals distinct maturity states and an antiviral cell circuit.



AK Shalek *et al. Nature* 000, 1-5 (2012) doi:10.1038/nature12172

Courtesy of Macmillan Publishers Limited. Used with permission.
Source: Shalek, Alex K., Rahul Satija, et al. "Single-cell Transcriptomics Reveals Bimodality in Expression and Splicing in Immune Cells." *Nature* (2013).

Analysis of co-variation in single-cell mRNA expression levels reveals distinct maturity states and an antiviral cell circuit.

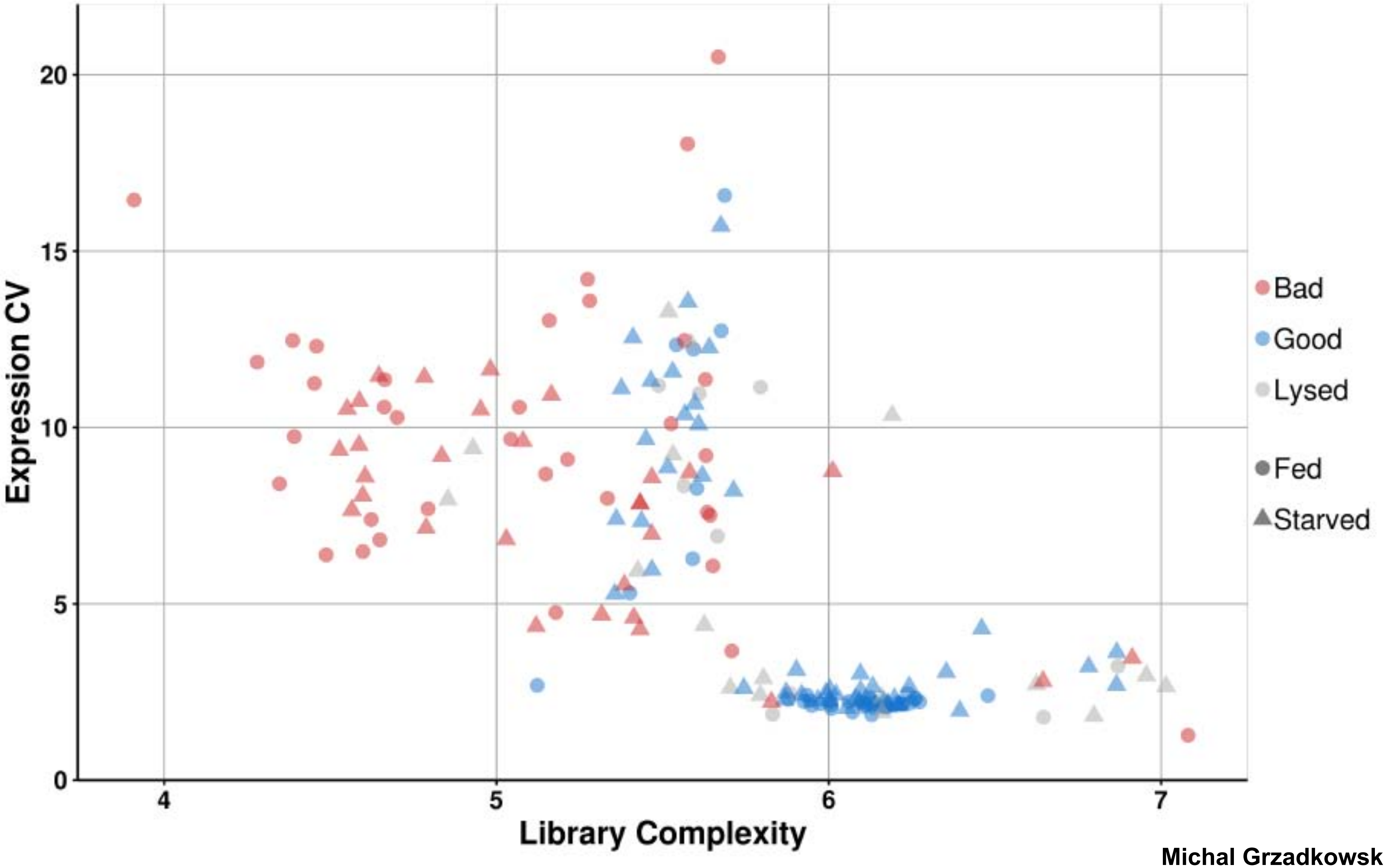


AK Shalek *et al. Nature* 000, 1-5 (2012) doi:10.1038/nature12172

Courtesy of Macmillan Publishers Limited. Used with permission.

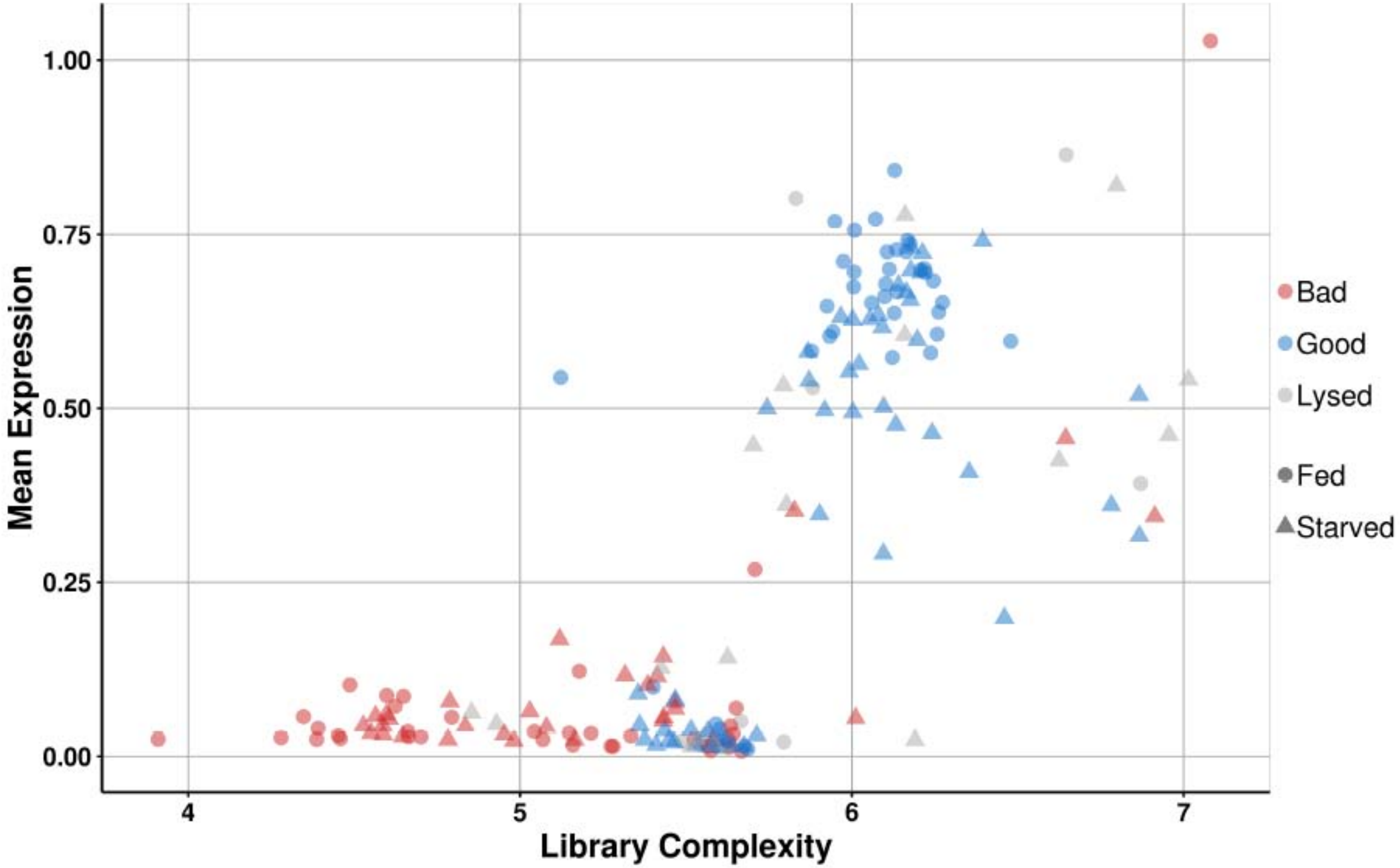
Source: Shalek, Alex K., Rahul Satija, et al. "Single-cell Transcriptomics Reveals Bimodality in Expression and Splicing in Immune Cells." *Nature* (2013).

RNA-seq library complexity can help qualify cells for analysis



Michal Grzadkowski

RNA-seq library complexity can help qualify cells for analysis



Michal Grzadkowski

FIN