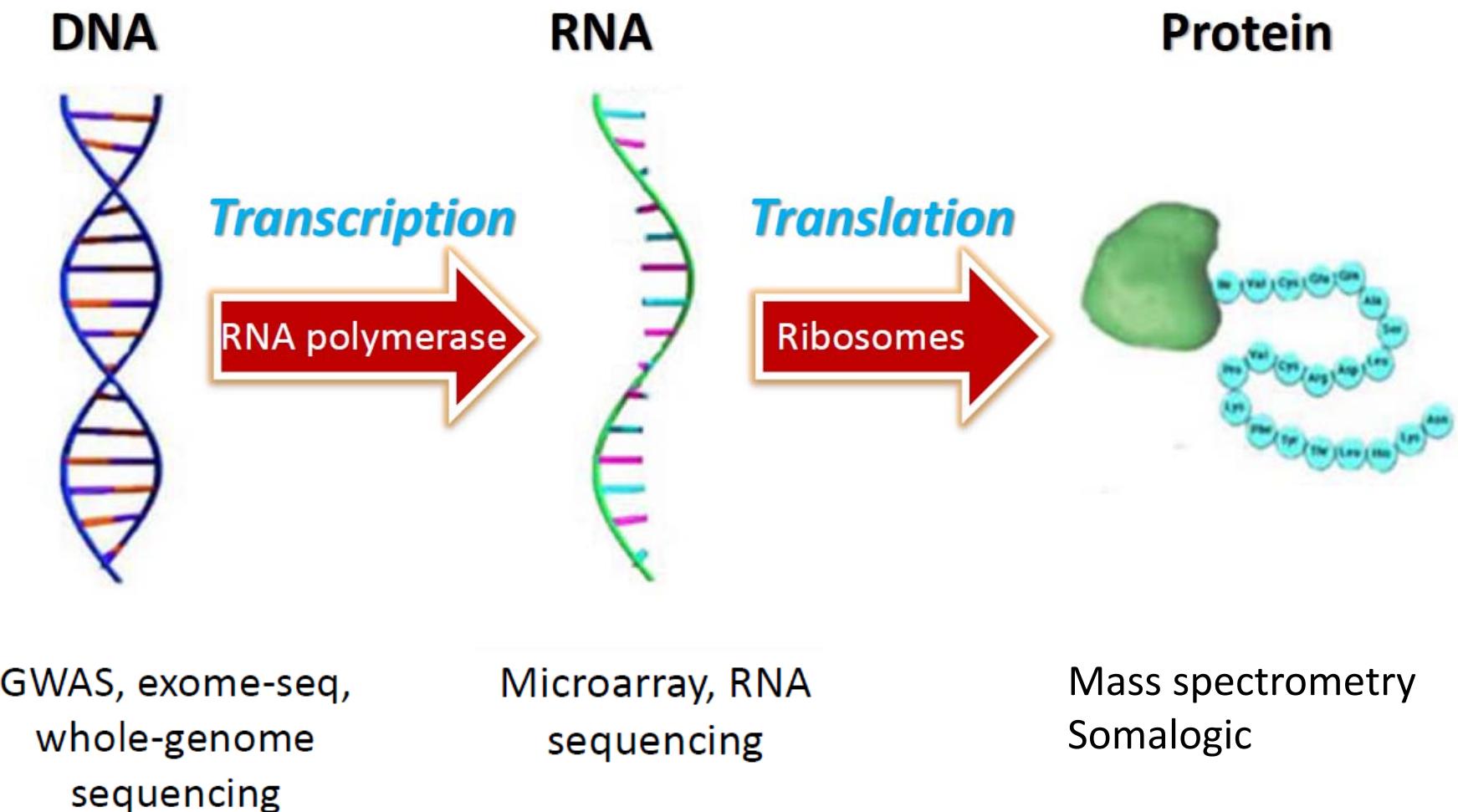


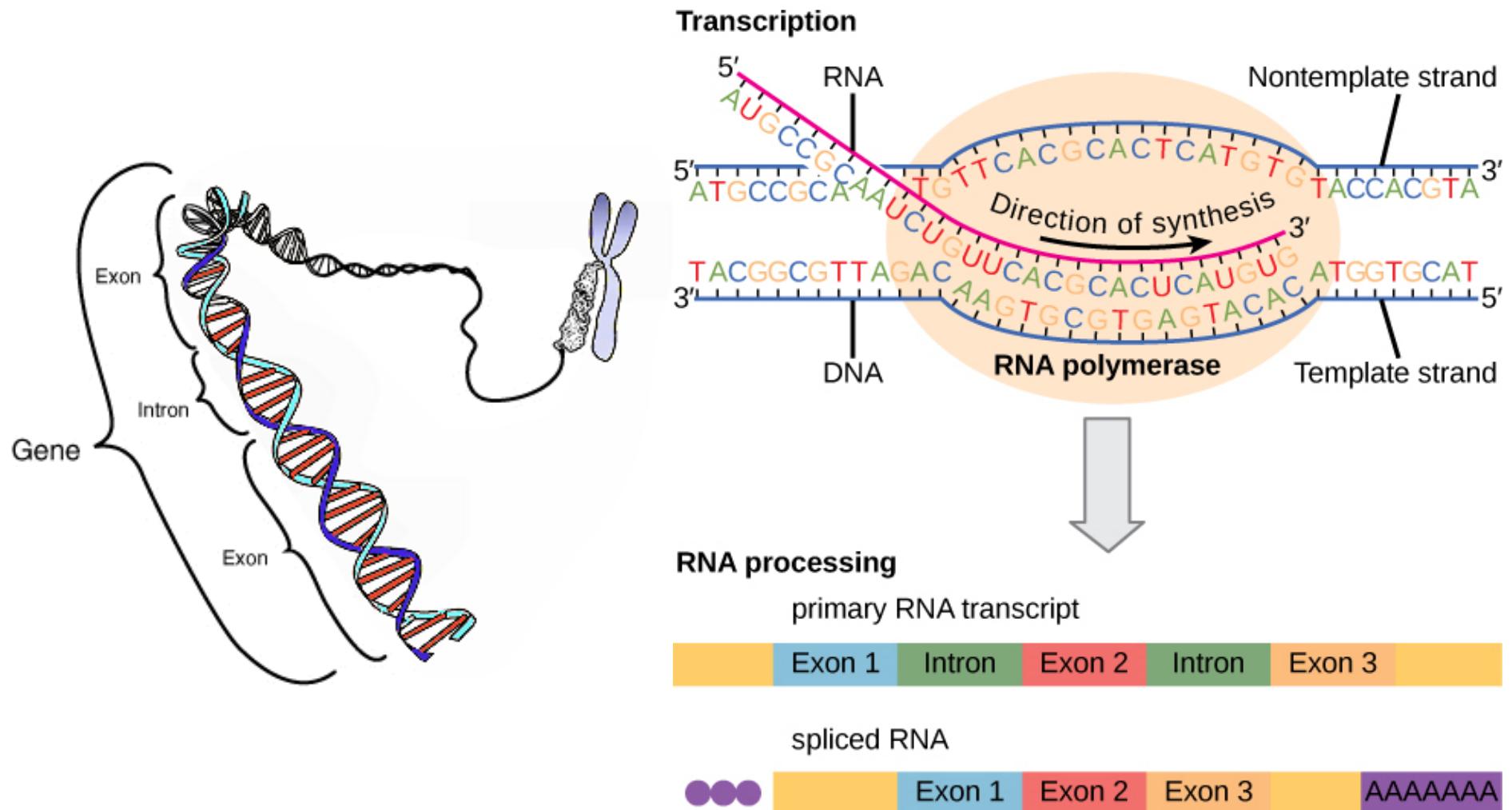
RNA-Seq in human diseases

2019 Dragon Star Bioinformatics Course (Day 5)

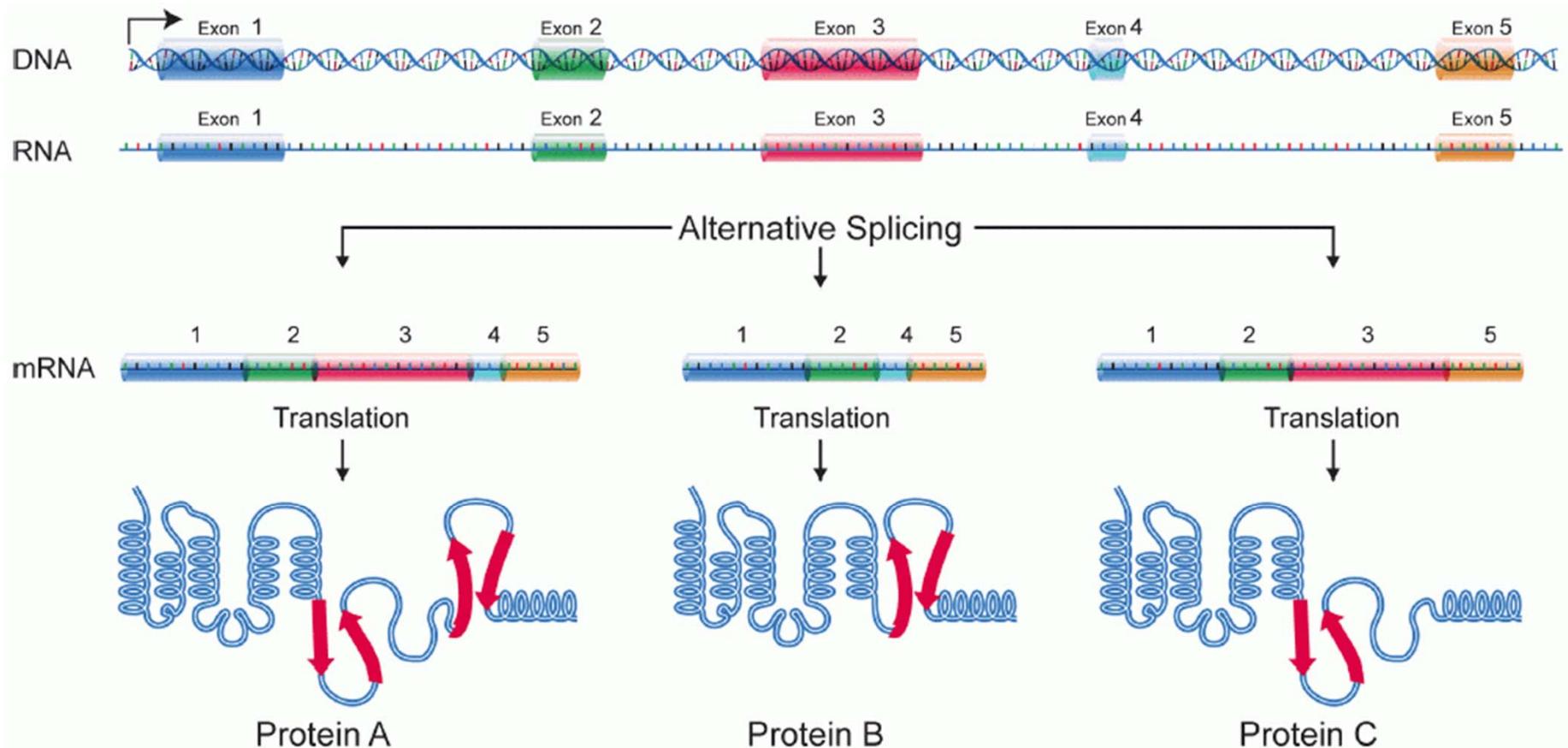
Central Dogma of Biology



Gene Transcription



Alternative Splicing



~90% of human genes are alternatively spliced

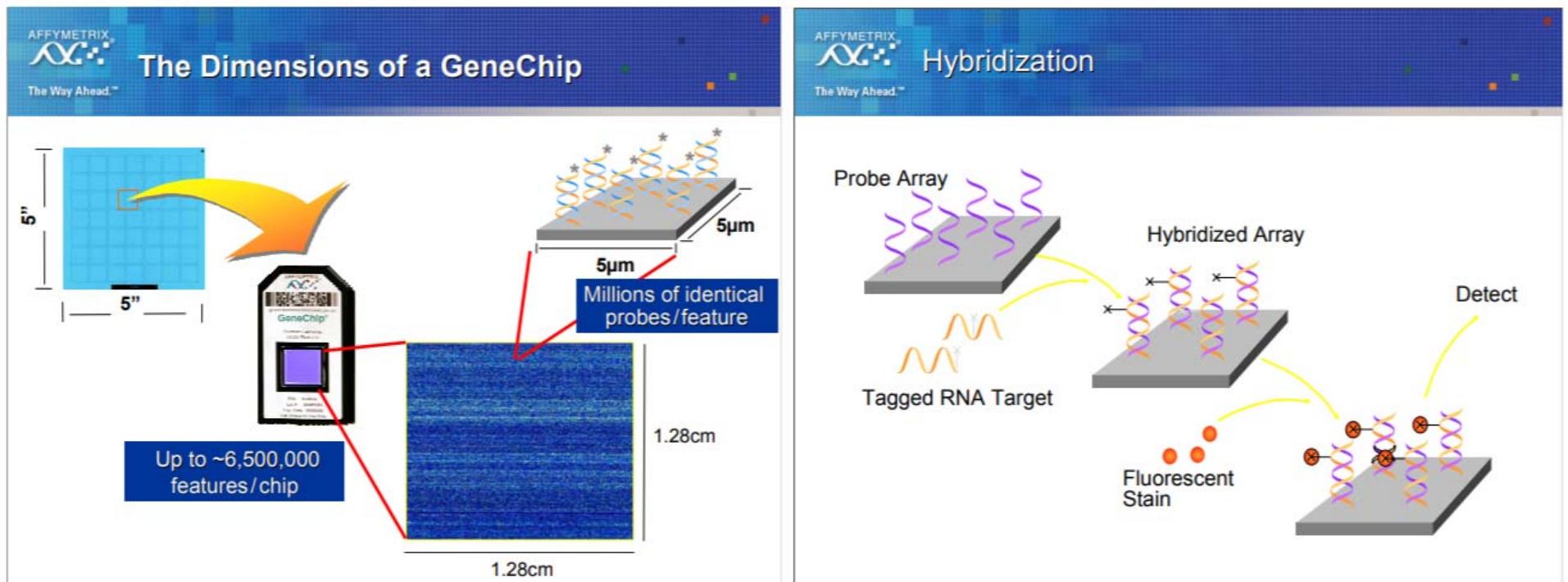
Alternatively splicing substantially increased transcriptome complexity.

What can a transcriptome tell us?

- In humans and other organisms, nearly every cell contains the same genes, but different cells show different patterns of gene expression.
- These differences are responsible for the many different properties and behaviors of various cells and tissues, both in health and disease.
- By comparing transcriptomes of different cell types, we can gain a deeper understanding of
 - What constitutes a specific cell type
 - How that type of cell normally functions
 - How changes in the normal level of gene activity may reflect or contribute to disease

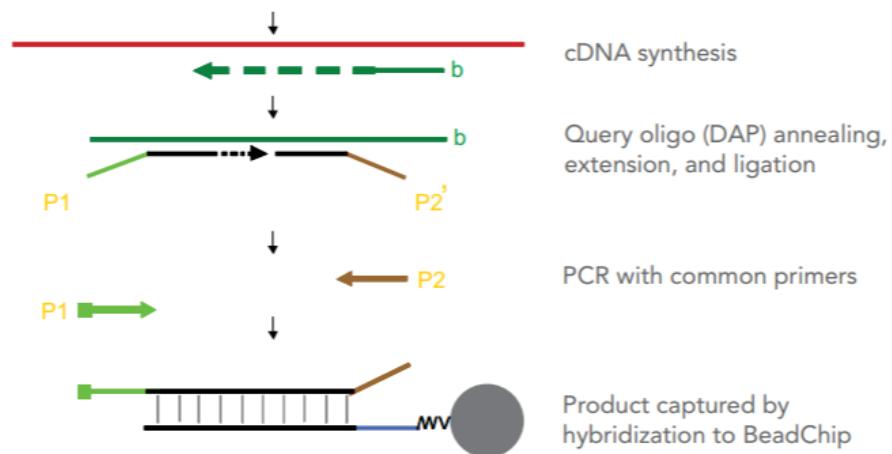
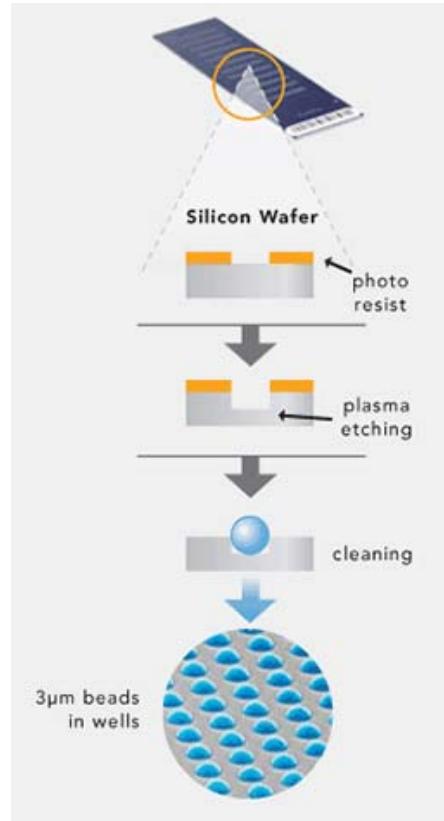
Technologies for Transcriptomics: expression microarrays

- Affymetrix GeneChip is widely used before RNA-Seq is available



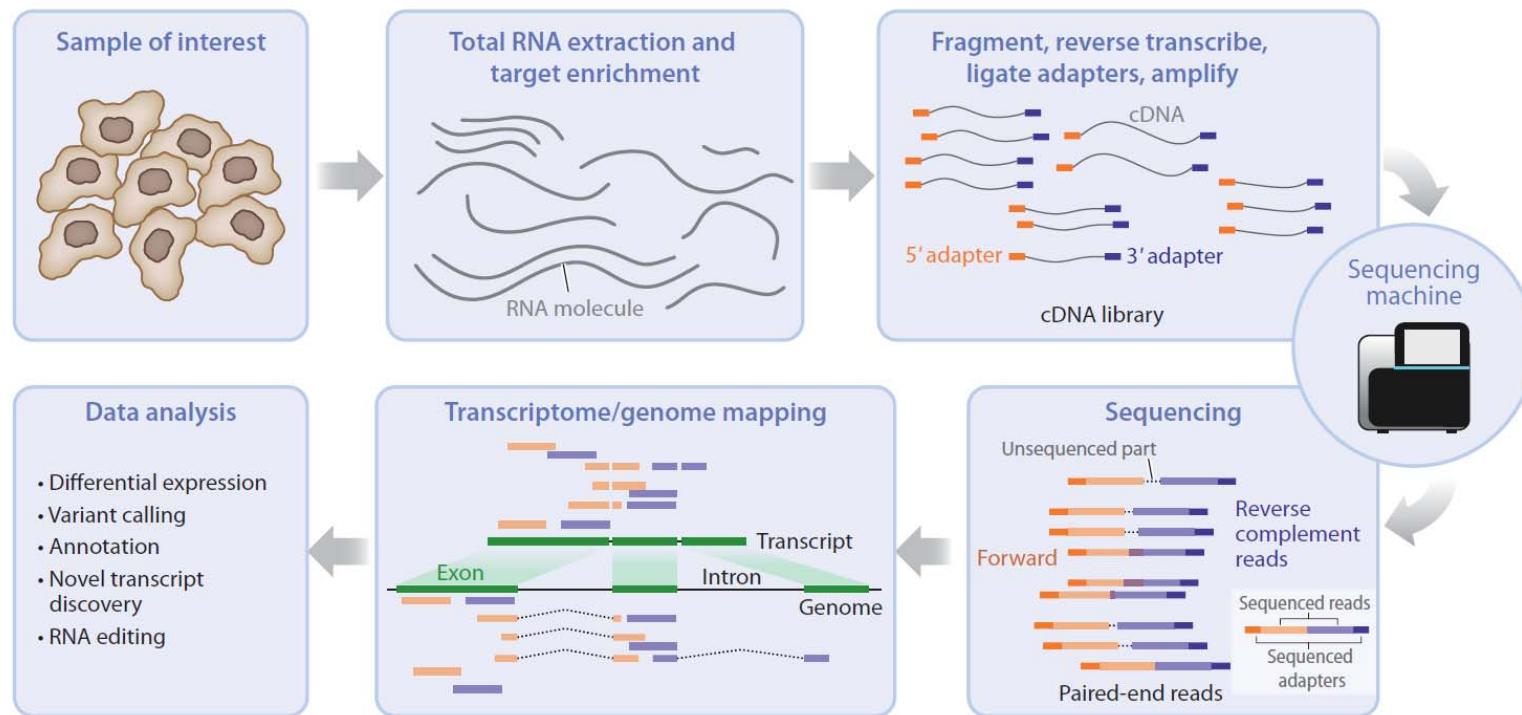
Technologies for Transcriptomics: expression microarrays

- Illumina BeadChip can also be used for gene expression analysis

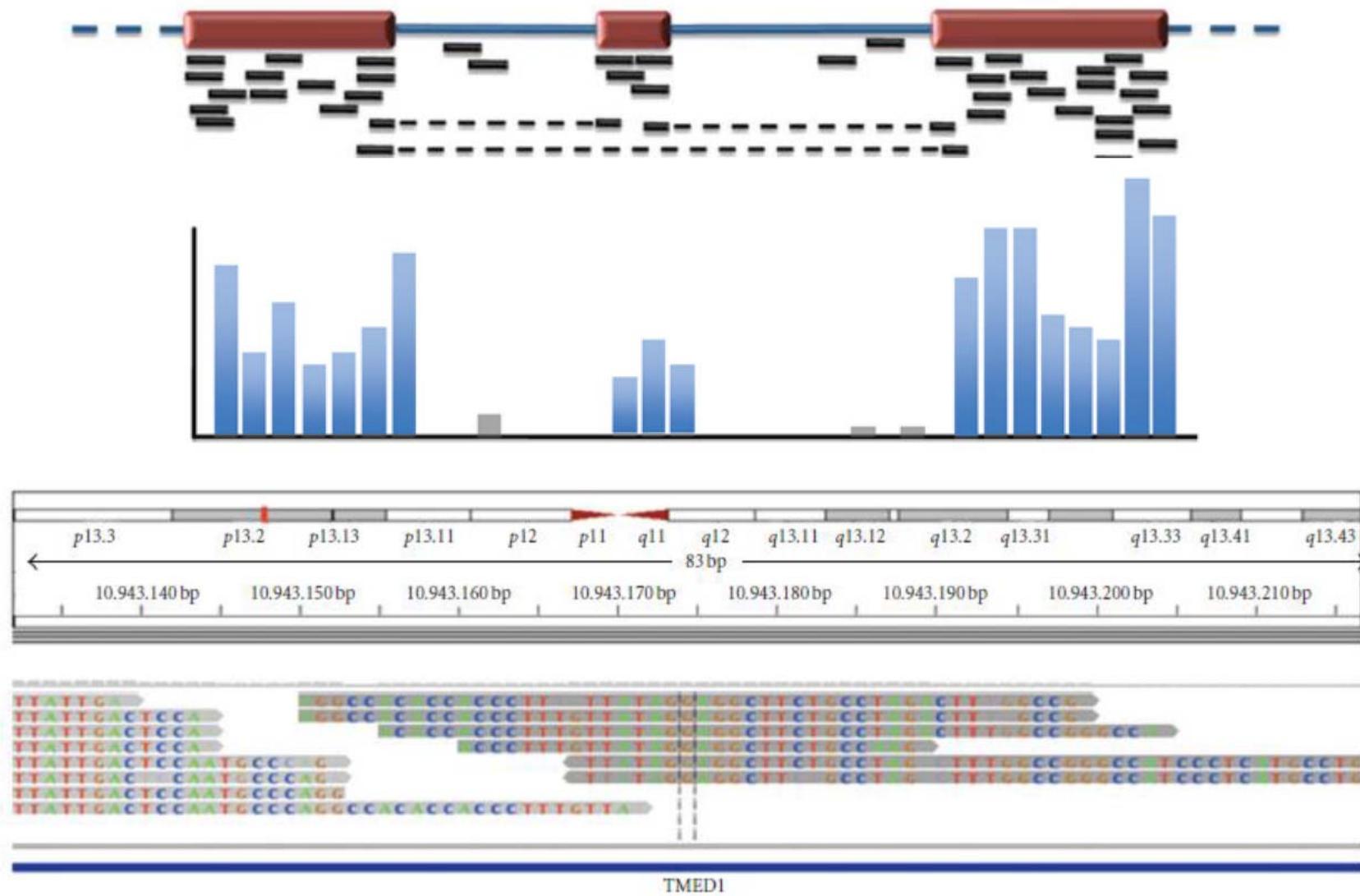


Technologies for Transcriptomics: RNA Sequencing (RNA-Seq)

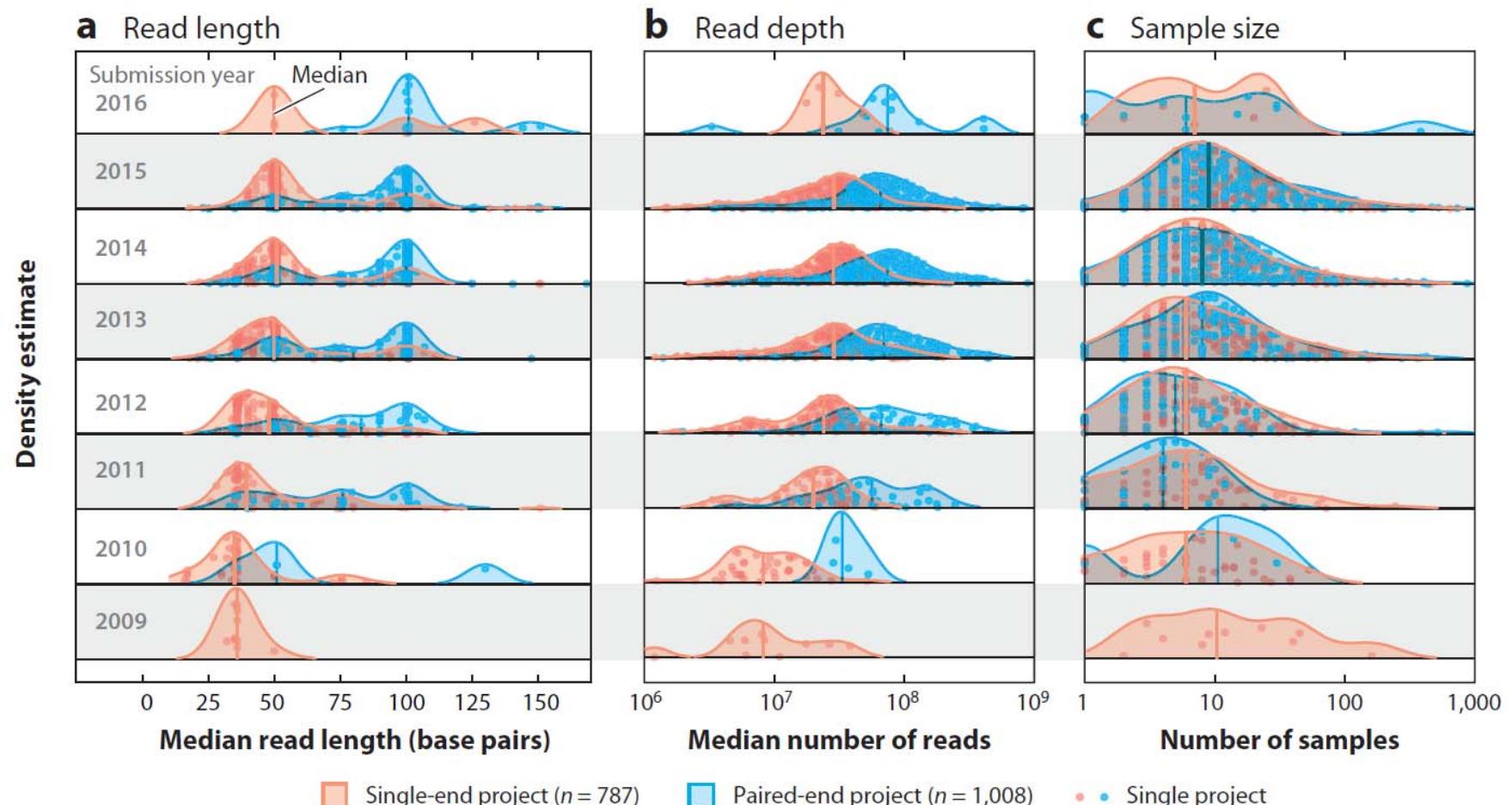
- RNA-Seq gradually replaced expression microarrays as the most widely used methods for gene expression analysis



Information in RNA-seq Data



Evolution of experimental design over the past several years



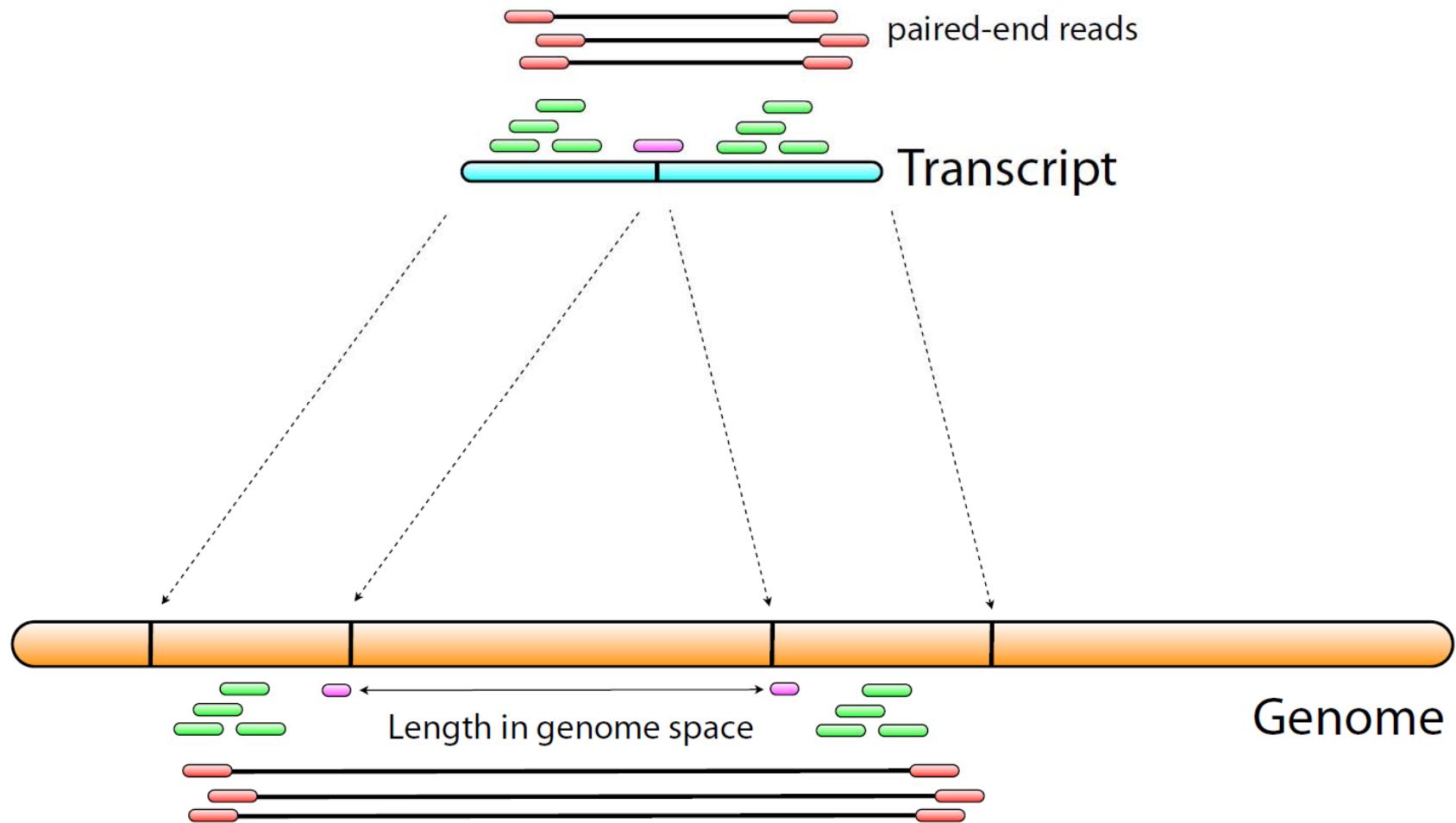
Advantages of RNA-Seq over Microarrays

- Microarrays measure only genes corresponding to predetermined probes on a microarray while RNA-seq measures any expressed transcripts in a sample.
- With RNA-Seq, there is no need to identify probes prior to measurement or to build a microarray.
- RNA-Seq provides count data which may be closer, at least in principle, to the amount of mRNA produced by a gene than the fluorescence measures produced with microarray technology.

Advantages of RNA-Seq over Microarrays

- RNA-Seq provides information about transcript sequence in addition to information about transcript abundance.
- Thus, with RNA-Seq, it is possible to separately measure the expression of different transcripts (i.e., isoforms) that would be difficult to separately measure with microarray technology due to cross hybridization.
- Sequence information also permits the identification of allele specific expression, single nucleotide polymorphisms (SNPs), and other forms of sequence variation such as RNA editing.

RNA-Seq Alignment

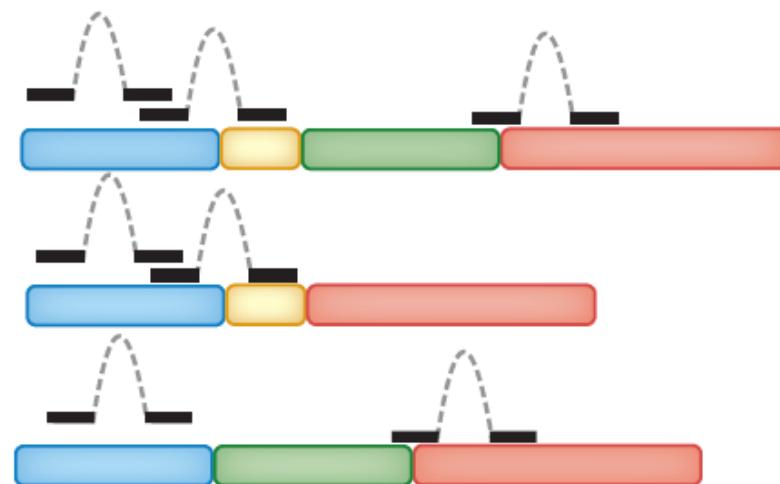


Splice-aware RNA-Seq alignment methods

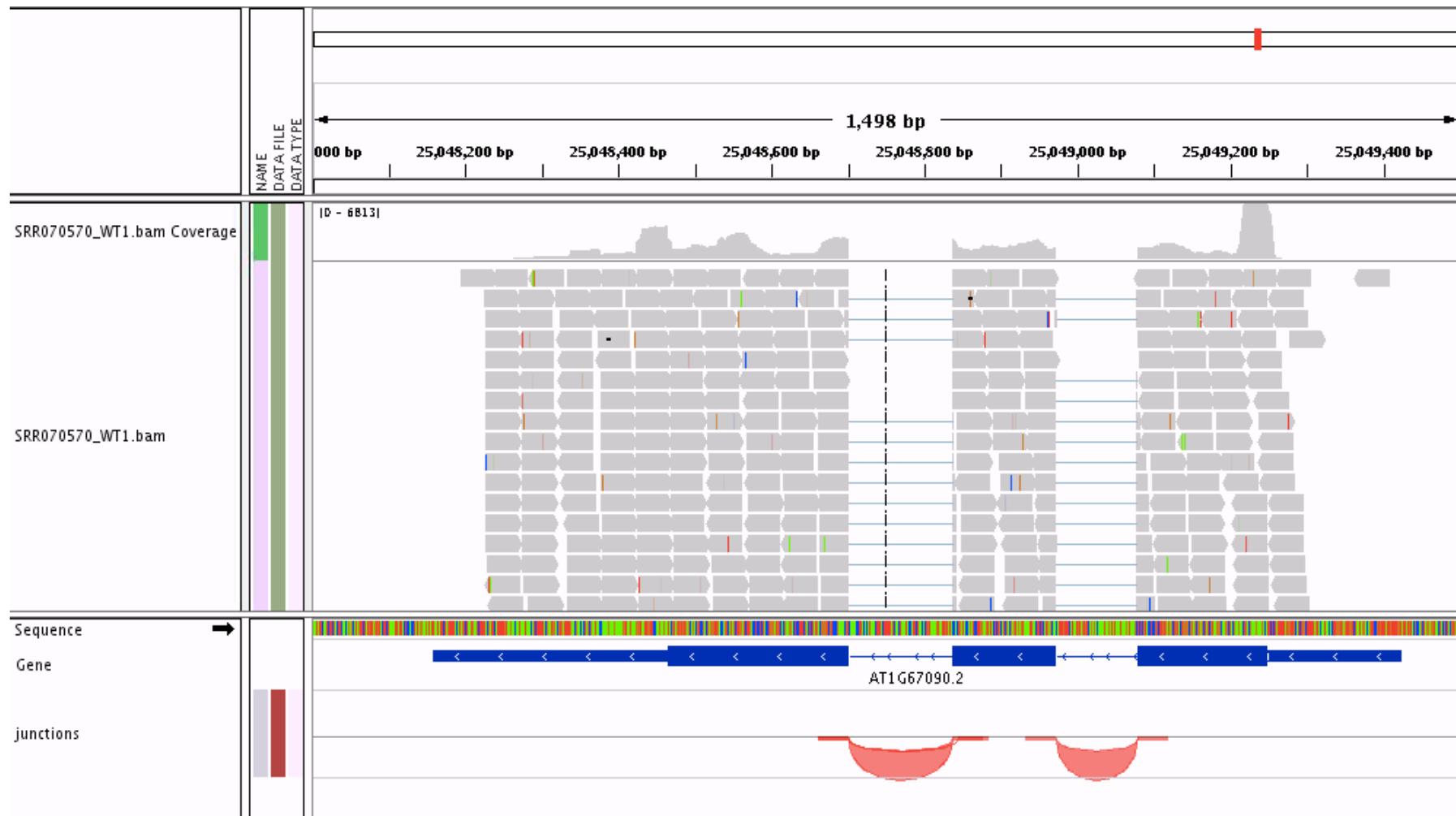
a Spliced alignment against genome



b Unspliced alignment against transcriptome

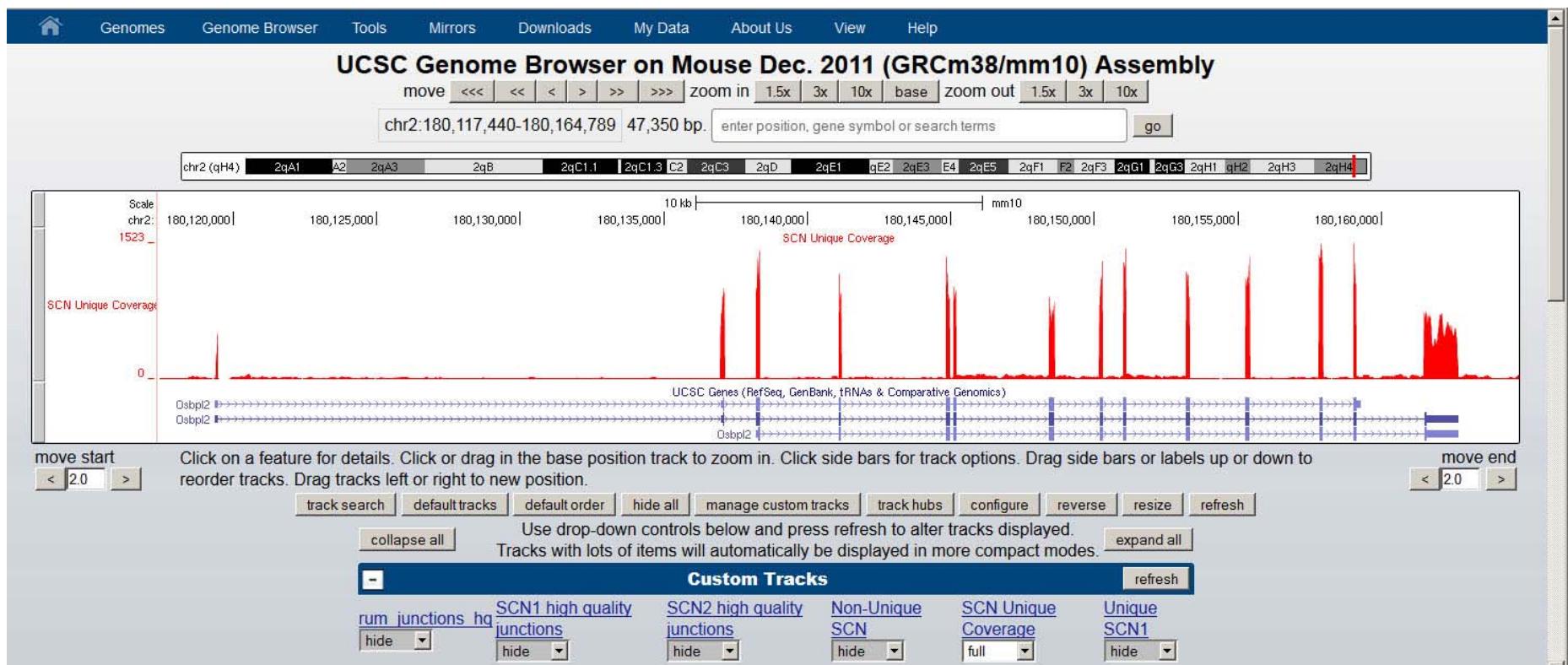


How does the Aligned File Look Like?



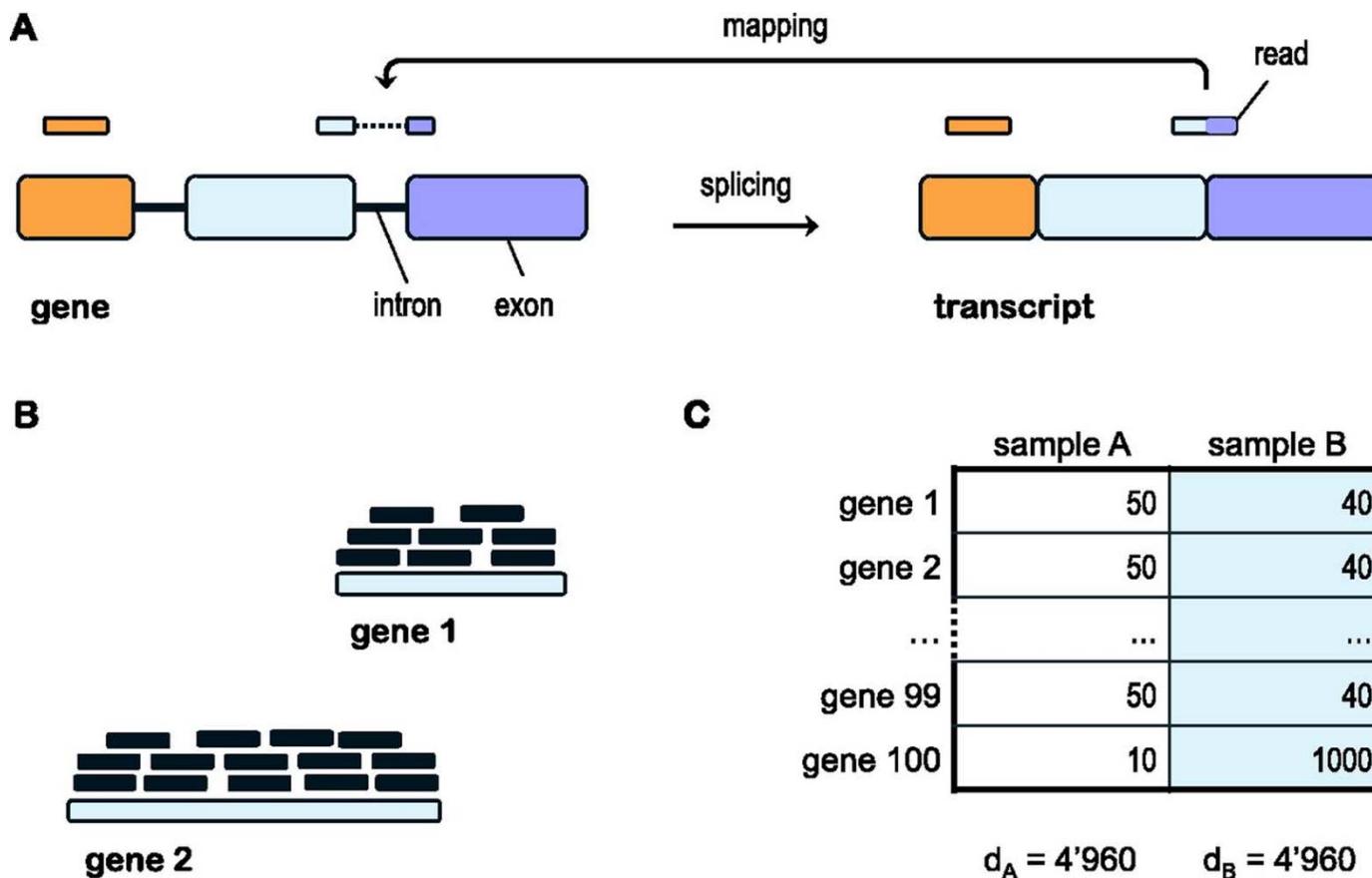
You can visualize the coverage data in IGV by opening an indexed BAM file

How does the Aligned File Look Like?



You can visualize the coverage data in Genome Browser
(<https://genome.ucsc.edu/goldenPath/help/bam.html>)

Typical RNA-Seq considerations: alignment, count normalization, sequencing biases



Selected RNA-Seq Alignment Programs

Short-read RNA-Seq

- TopHat2
- HISAT2
- MapSplice
- SOAPSplice
- SpliceMap
- RUM
- GSNAP
- STAR

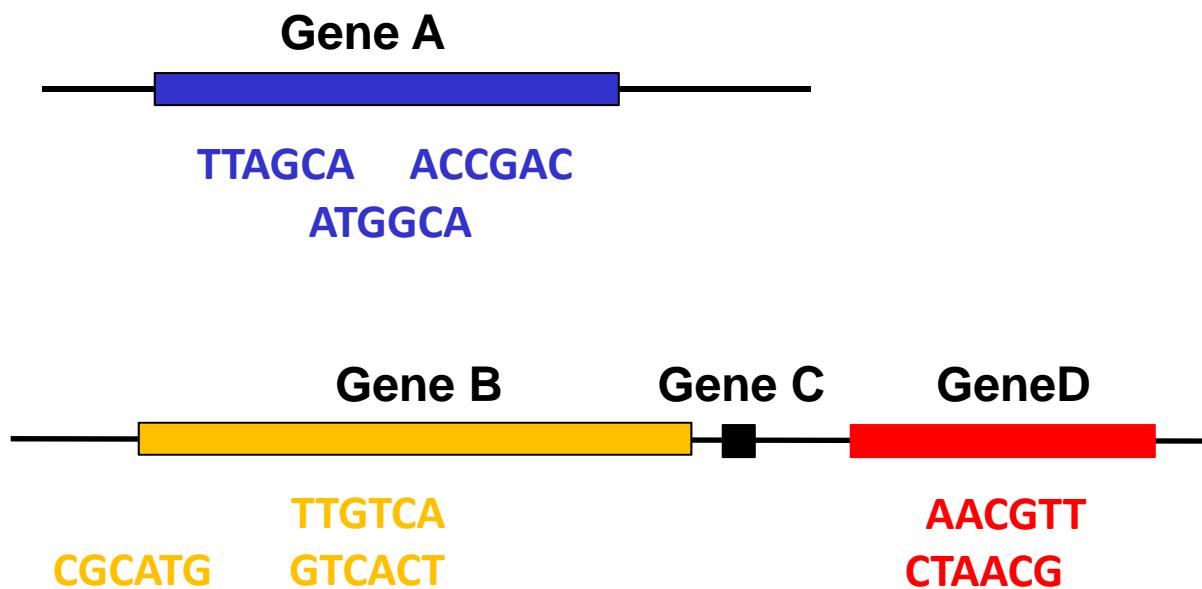
Long-read RNA-Seq

- GMAP
- STAR
- BBMap
- minimap2

Note: STAR+RSEM is widely used today in various RNA-Seq studies

In general, you can just use STAR for both alignment and quantification (counting)

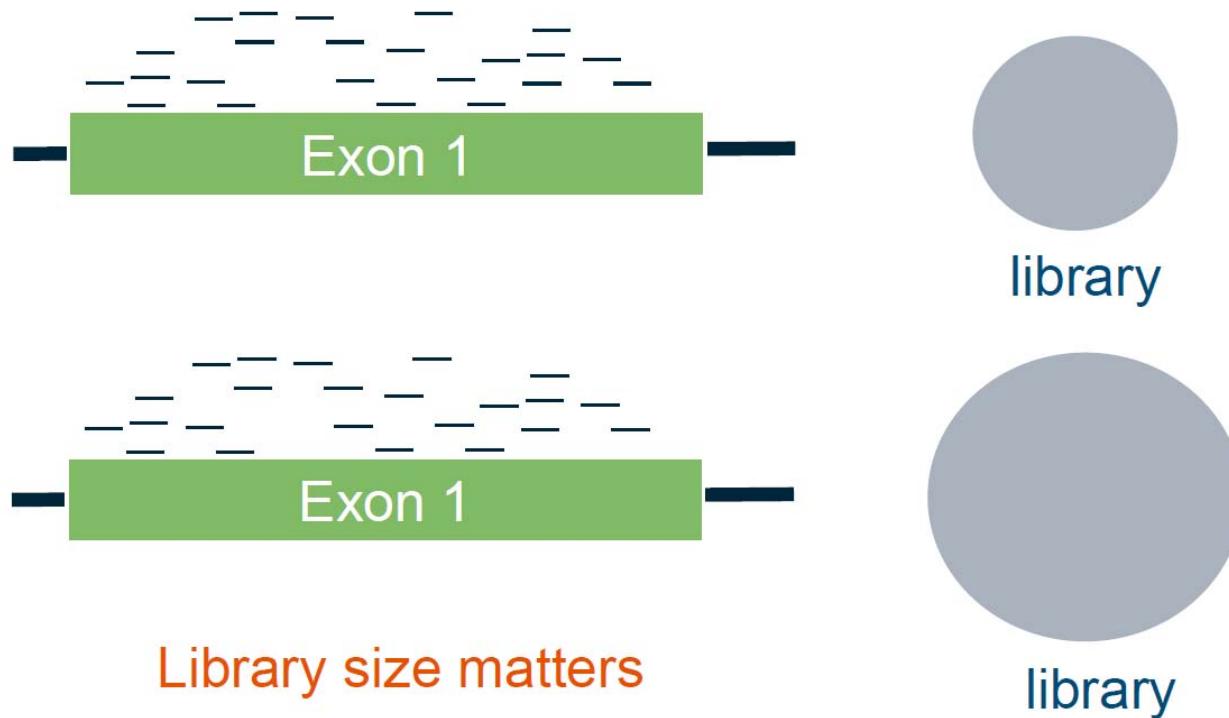
Quantification of Gene Expression



Gene ID	Sample1
A	3
B	3
C	0
D	2

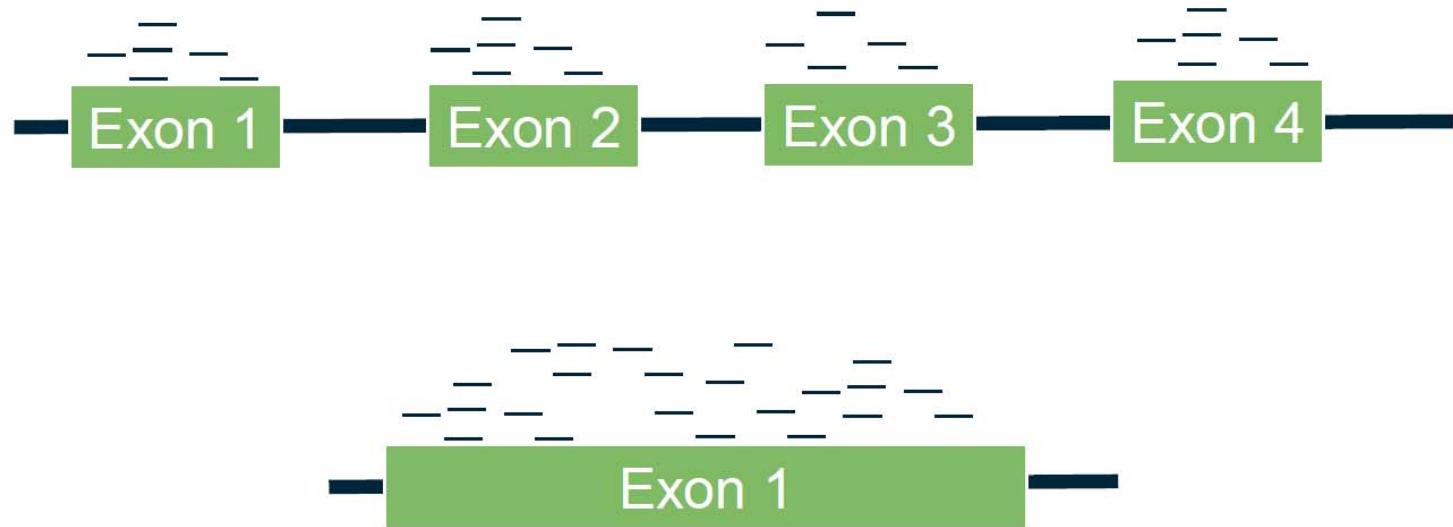
For a given gene, the number of reads aligned to the gene measures its expression level.

Normalization is Important



Larger library has more reads

Normalization is Important



Transcript/gene length matters

Larger transcript/gene has more reads

Normalization: RPKM

Reads Per Kilobase of transcript per Million mapped reads



$$RPKM_i = \frac{X_i}{\left(\frac{\tilde{l}_i}{10^3}\right) \left(\frac{N}{10^6}\right)} = \frac{X_i}{\tilde{l}_i N} \cdot 10^9$$

adjust gene length difference
 \tilde{l}_i is gene length

adjust library size difference
 N is total library size

RPKM is used for single-end data

Normalization: FPKM

Fragments Per Kilobase of transcript per Million mapped reads

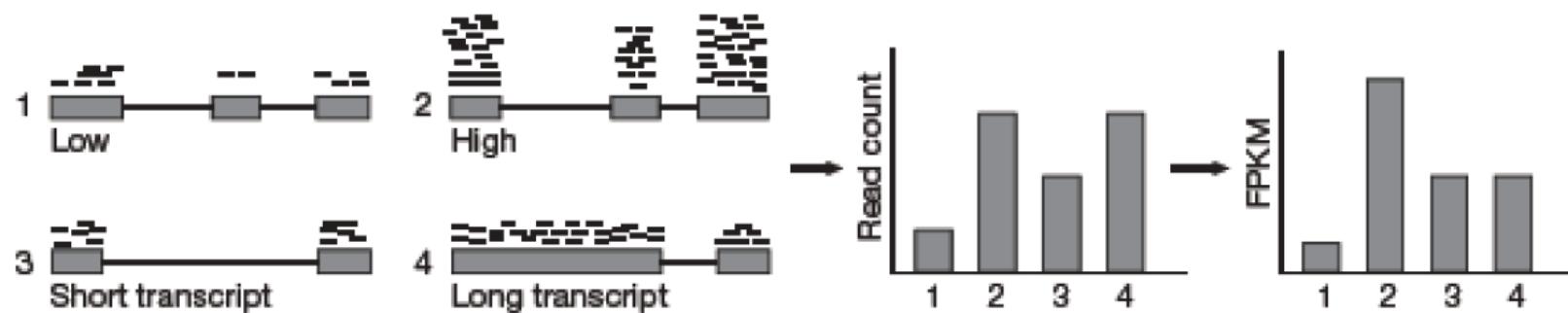
FPKM is analogous to RPKM

Sequencing fragments



FPKM is used in paired-end data

Different picture emerges from raw counts and RPKM/FPKM values



Normalization: TPM

Transcripts per million (TPM) is a measurement of the proportion of transcripts in your pool of RNA.

$$\text{TPM}_i = \frac{X_i}{\tilde{l}_i} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\tilde{l}_j}} \right) \cdot 10^6$$

RPKM, FPKM or TPM?

While TPM and RPKM/FPKM normalization methods both account for sequencing depth and gene length, RPKM/FPKM measures are not recommended. **The reason is that the normalized count values output by the RPKM/FPKM method are not comparable between samples.**

gene	sampleA	sampleB
MOV10	5.5	5.5
ABCD	73.4	21.8
...
Total RPKM	1,000,000	1,500,000

Sample A has a greater proportion of counts associated with MOV10 than sample B, although the RPKMs are the same.

RPKM, FPKM or TPM?

TPM (recommended)

In contrast to RPKM/FPKM, TPM-normalized counts normalize for both sequencing depth and gene length, but have the same total TPM-normalized counts per sample. Therefore, the **normalized count values are comparable both between and within samples**.

Relationship btw TPM and FPKM

$$\begin{aligned} \text{TPM}_i &= \frac{X_i}{\tilde{l}_i} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\tilde{l}_j}} \right) \cdot 10^6 \propto \frac{X_i}{\tilde{l}_i \cdot N} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\tilde{l}_j \cdot N}} \right) \\ &\propto \frac{X_i}{\tilde{l}_i \cdot N} \cdot 10^9 \end{aligned}$$

If you have FPKM, you can easily compute TPM:

$$\text{TPM}_i = \left(\frac{\text{FPKM}_i}{\sum_j \text{FPKM}_j} \right) \cdot 10^6$$

Differential Gene Expression

Transcript	Group 1			Group 2		
1	14	18	10	47	13	24
2	10	3	15	1	11	5
3	1	0	10	80	21	34
4	0	0	0	0	2	0
5	4	3	3	5	33	29
.
.
.
53256	47	29	11	71	278	339
Total	22910173	30701031	18897029	20546299	28491272	27082148

Two groups of samples (3 vs 3)

Differential Gene Expression Analysis

- To determine if gene 1 is DE, we should examine whether the proportion of reads aligned to gene 1 tends to be different for samples in group 1 than for samples in group 2.

14 out of 22910173

47 out of 20546299

18 out of 30701031

vs. 13 out of 28491272

10 out of 18897029

24 out of 27082148

Poisson Approximation to Binomial

- Let n be the total number of reads, and θ be the relative abundance of a gene, then the read count for the gene $Y \sim \text{Binomial}(n, \theta)$.
- When n is large, the distribution of Y can be approximated by $\text{Poisson}(\lambda = n\theta)$.
- Thus, we may choose to model the count for group i , gene j , and sample k as $Y_{ijk} \sim \text{Poisson}(n_{ik}\theta_{ij})$, where n_{ik} is the total number of reads for group i sample k .

Problem with Poisson: Over-dispersion

- Recall that $Y \sim \text{Poisson}(\lambda)$ implies

$$E(Y) = \lambda \text{ and } \text{Var}(Y) = \lambda.$$

- From the fit of the generalized linear model, we can estimate count means and variances and assess whether the Poisson mean-variance relationship holds.
- When the actual counts are more variable than we would expect based on the Poisson assumption, the data are said to be over-dispersed.

Overdispersion in Real RNA-Seq Data

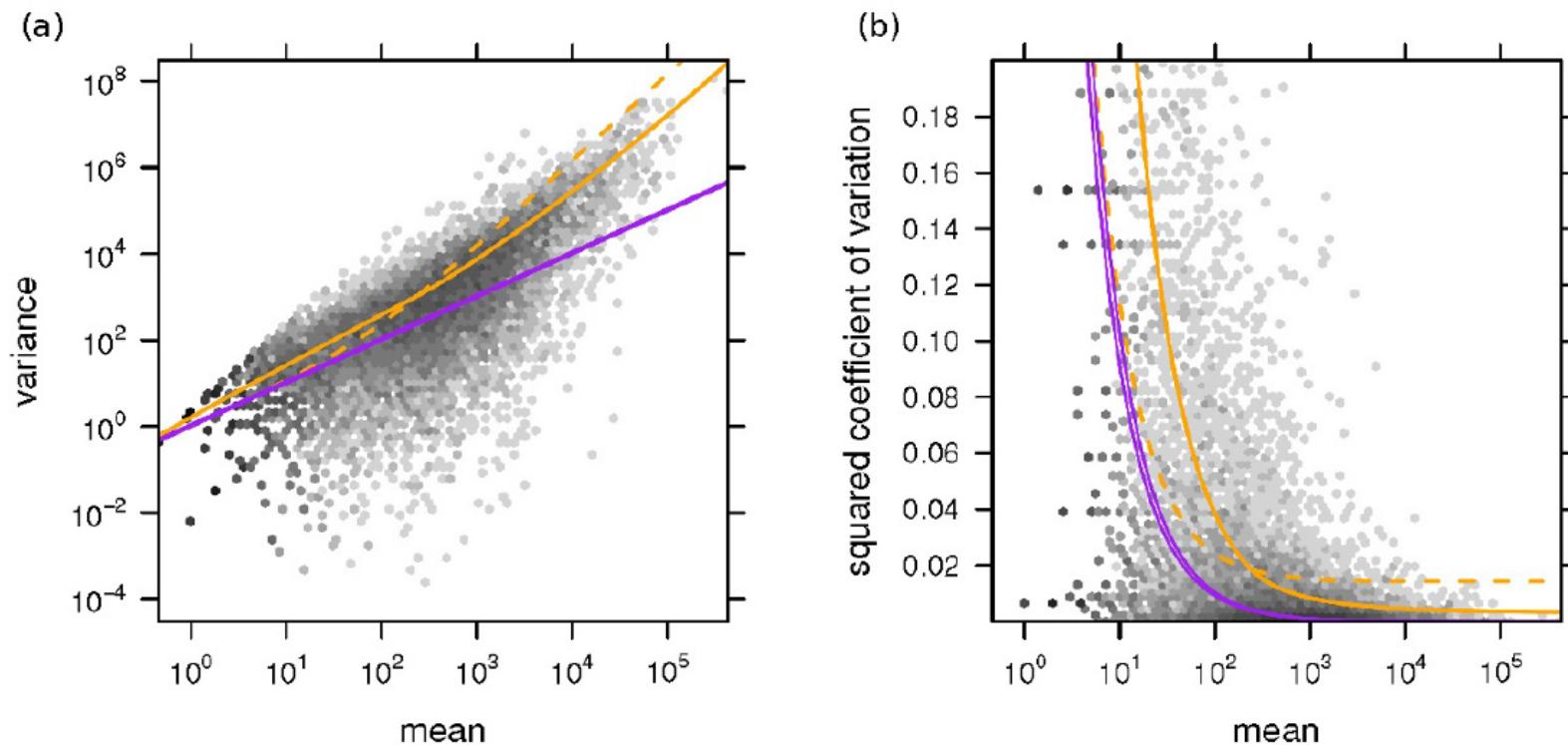


Figure 1 Dependence of the variance on the mean for condition A in the fly RNA-Seq data.

The purple lines show the variance implied by the Poisson distribution.

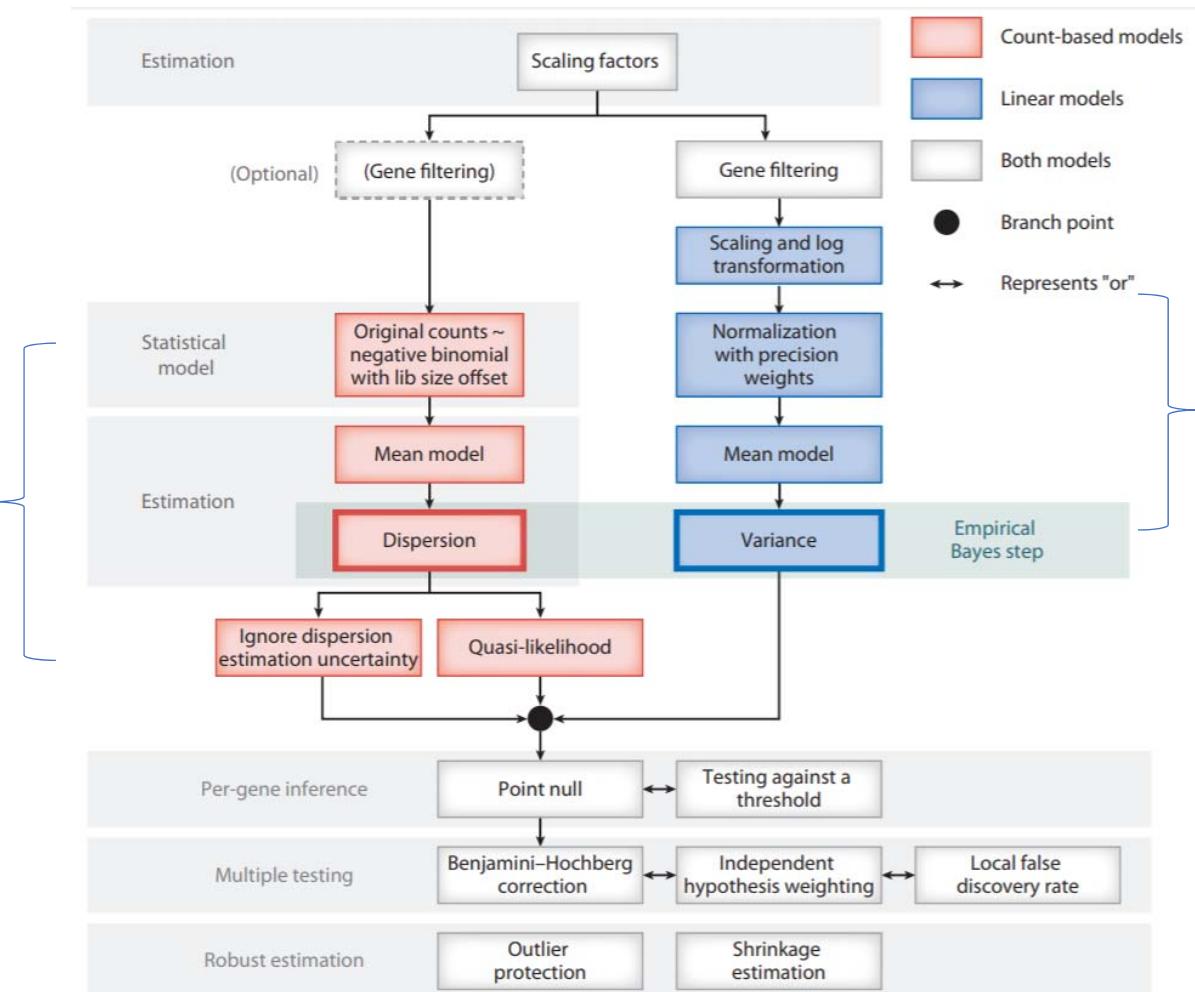
Estimating Overdispersion

- If $E(Y)=\lambda$ and $\text{Var}(Y)=\varphi\lambda$, φ is the dispersion parameter.
- When $\varphi>1$, the data are overdispersed.
- φ can be estimated by $-2 \log \Lambda / (N-p)$, where N is the number of observations, p is the number of free parameters in the model, and Λ is the likelihood ratio comparing to the previous Poisson model with the saturated Poisson model that estimates $E(Y_{ijk})$ by Y_{ijk} .
- Asymptotically the residual deviance should be χ^2 distributed with mean equal to the degrees of freedom. So divide it by its degrees of freedom, you should get about 1 if the data is not overdispersed.

Accounting for Overdispersion

DESeq
edgeR

voom



Accounting for Overdispersion: Linear models from microarray studies

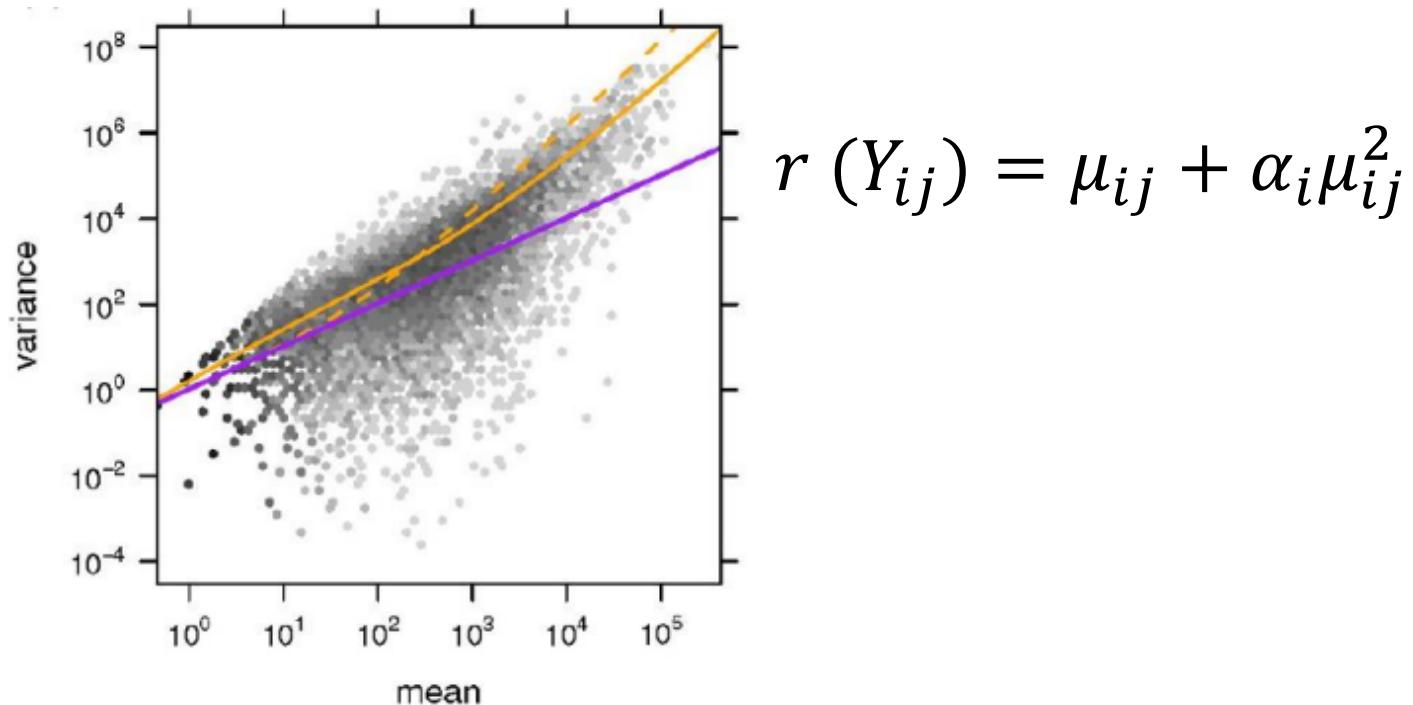
- A large number of methods were developed for microarrays (for example, limma), but they are not optimal for RNA-Seq analysis
- A modified version of limma, ‘limma voom’, estimate mean–variance relationship through lowess fit and used to estimate gene-wise variances. For each gene, the inverse of the variance is then used as weight in the ‘limma’ framework.

Accounting for Overdispersion: count models using negative binomial (NB)

- A number of algorithms and software tools are developed to account for over-dispersion in RNA-Seq.
- It is generally accepted that NB-based methods performs better than Poisson-based counterparts
- edgeR and DESeq, which are among the best performers and most widely used, are both based on the NB model, but implement different strategies for dispersion estimation.

DESeq2: Empirical Bayes Shrinkage for Dispersion Estimation

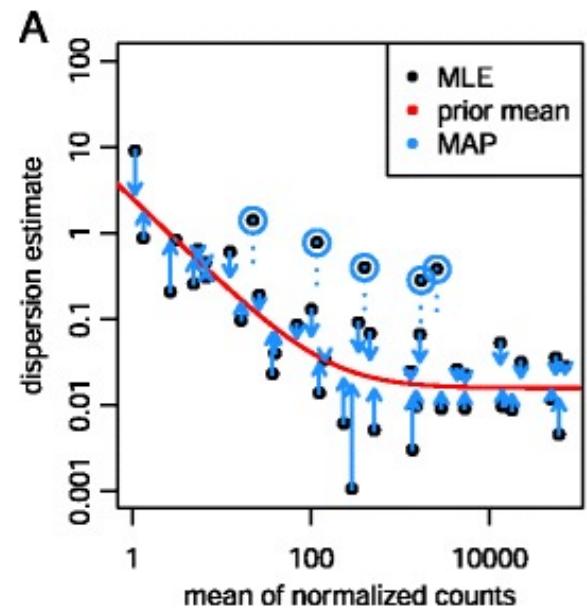
- Evidence of overdispersion on RNA-seq
- Dispersion parameter α_i in sample i



DESeq2: Empirical Bayes Shrinkage for Dispersion Estimation

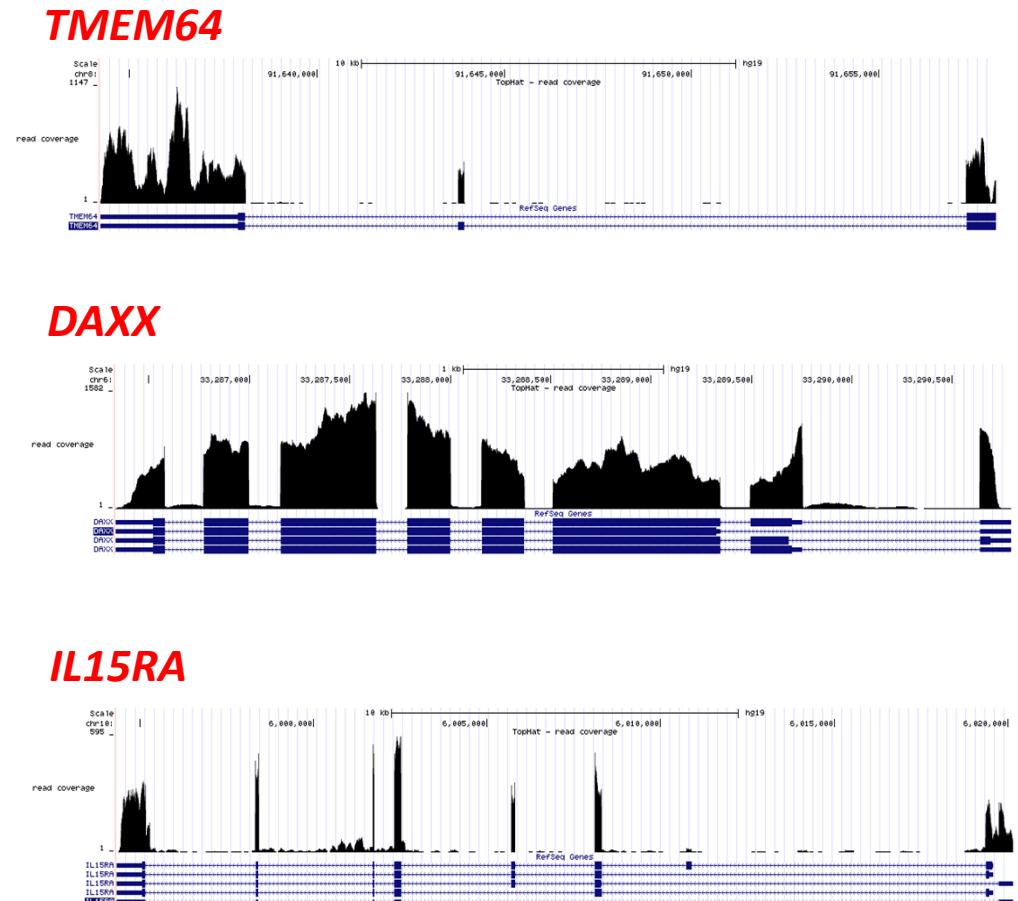
- Dispersion estimation in 3 steps:

- Estimate gene-wise dispersion α_i^{gw}
- Fit dispersion trend (prior)
- Shrink towards final estimation(MAP= maximum *a posteriori*)



Additional challenges: data biases

- Most methods assume sequencing reads are uniformly distributed along transcripts, i.e., each position is equally likely to be sequenced.
- However, true distributions often deviate substantially from uniformity.
- Bias correction is critical for accurate estimation of isoform expression.

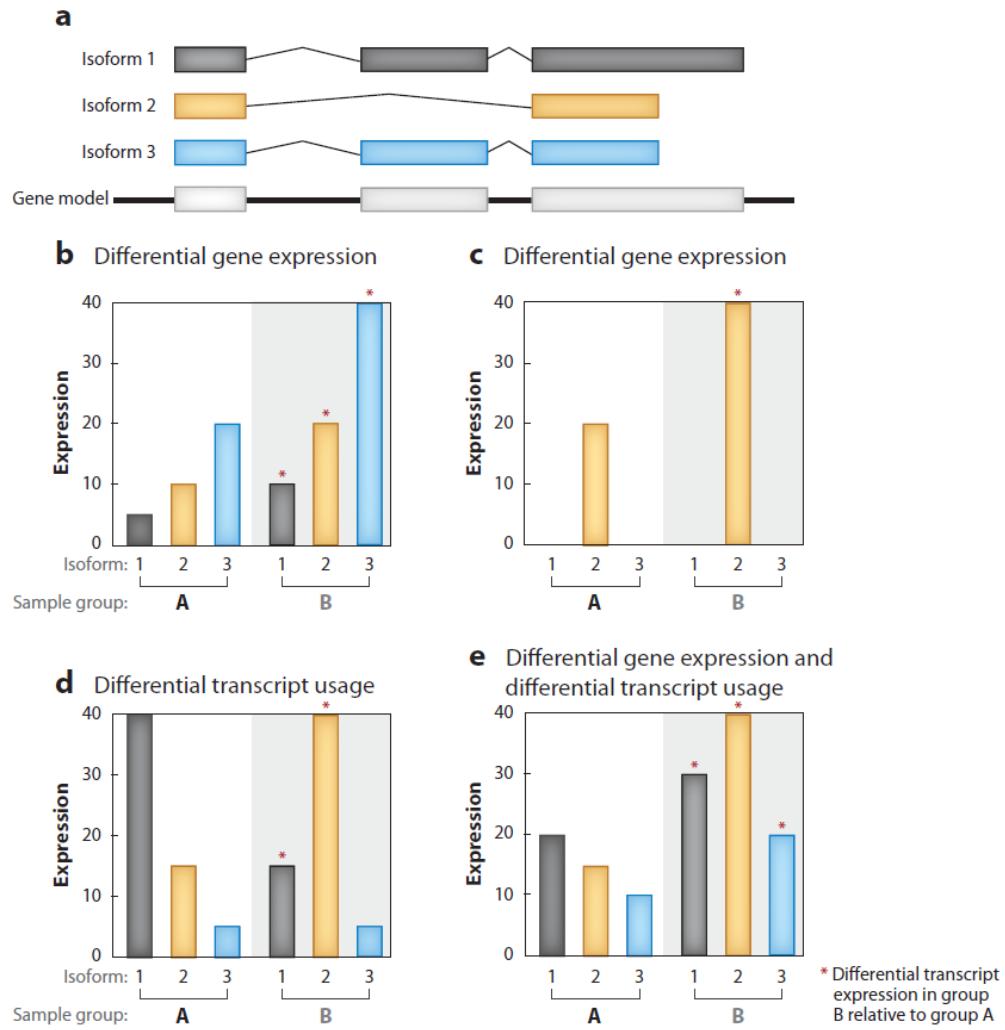


Genes selected from a human adipose study

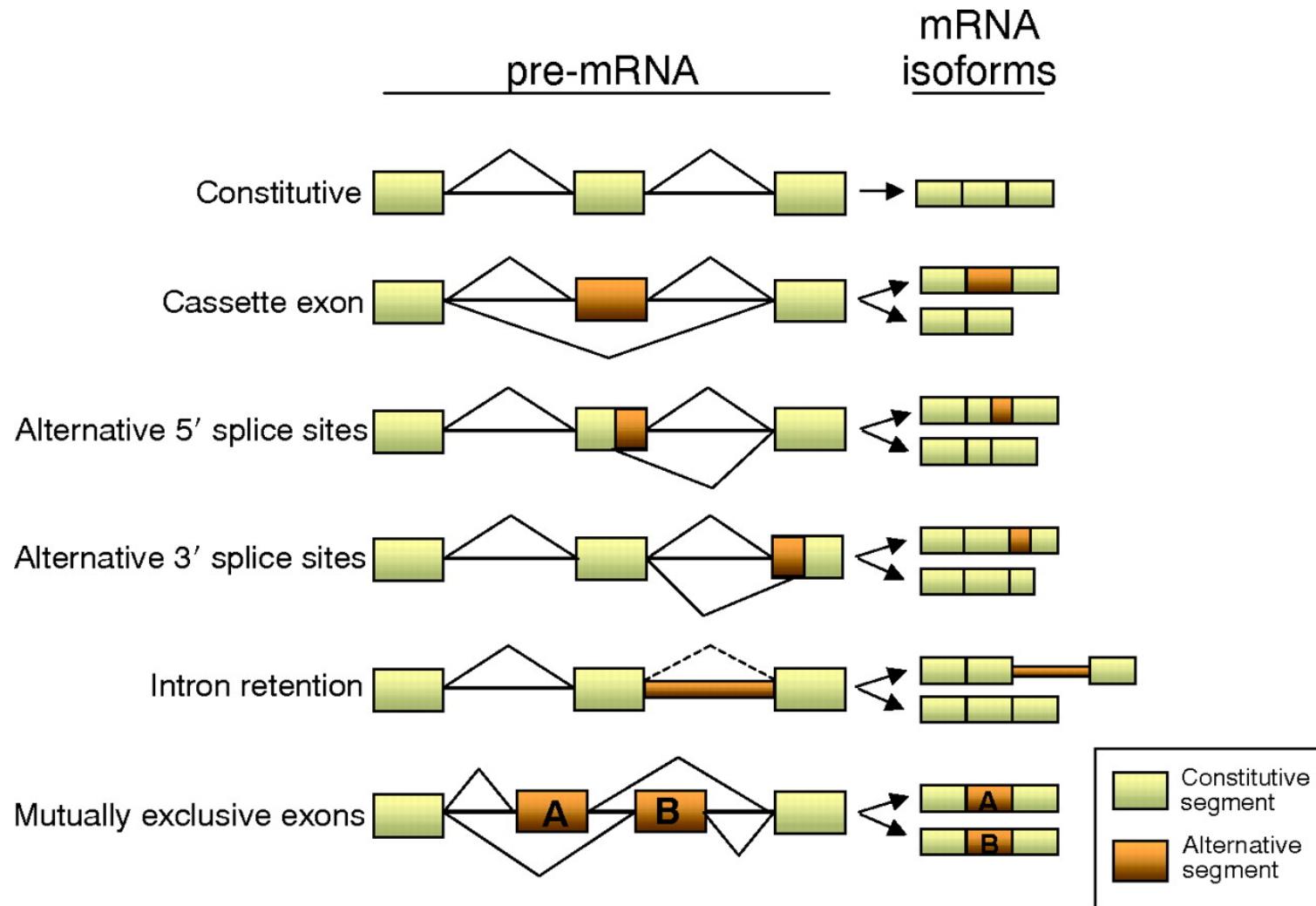
Available Methods for Bias Correction

- Li et al. (2010): bias correction caused by local sequence difference
- Li & Dewey (2011): model empirical read distribution in the transcriptome; implemented in program **RSEM**
- Roberts et al. (2011): bias correction of both sequence and positional bias; implemented in program **Cufflinks**
- Nicolae et al. (2011): bias correction using a reweighting scheme; implemented in program **IsoEM**
- Wan et al. (2012): parametric modeling of non-uniformity caused by RNA degradation; implemented in program **RD**
- Li et al. (2012): bias correction using a quasi-multinomial model; implemented in program **CEM**
- Hu et al. (2014): Non-parametric method that allows each isoform to have its own non-uniform distribution; implemented in **PennSeq**

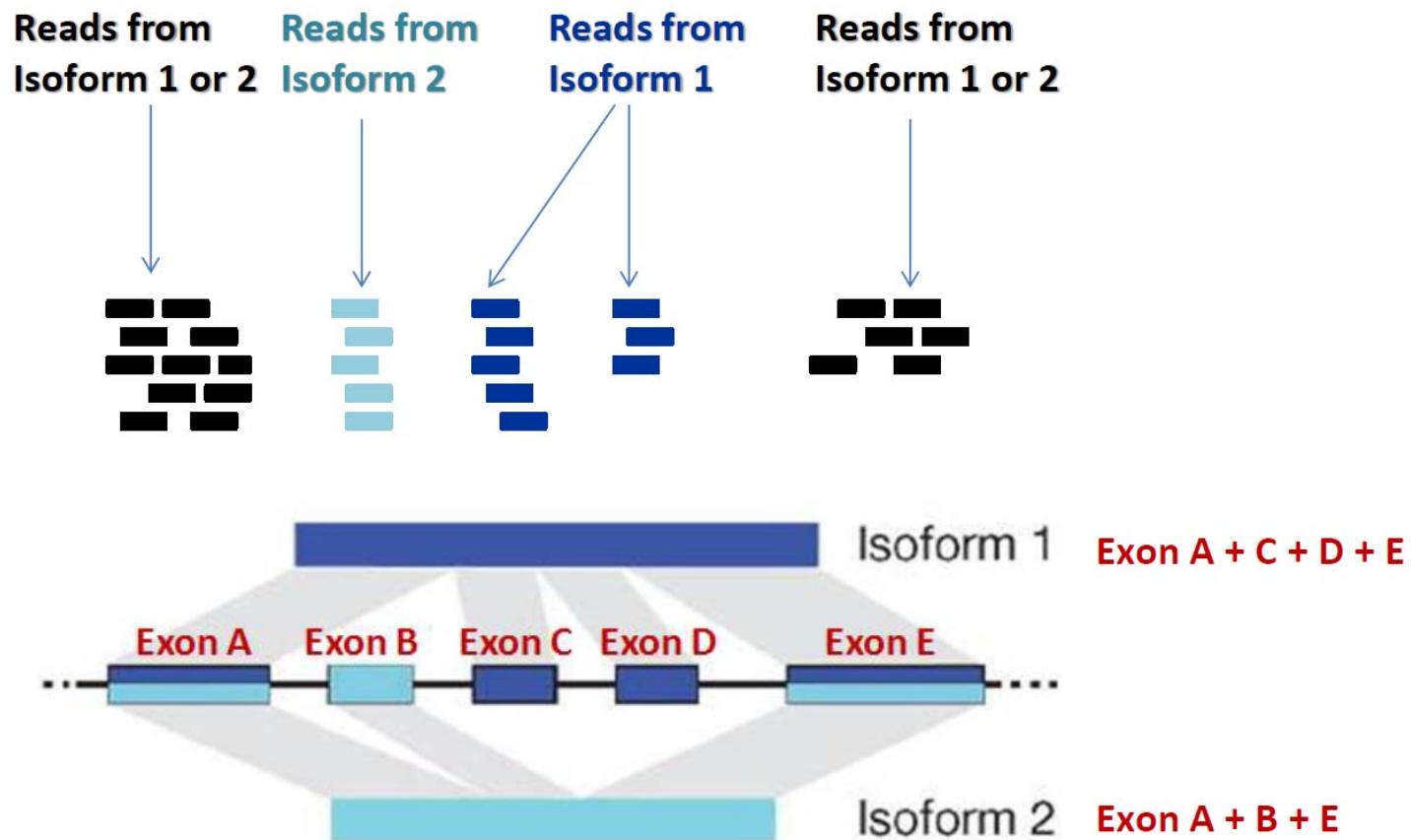
Differential expression vs differential splicing



Types of Alternative Splicing



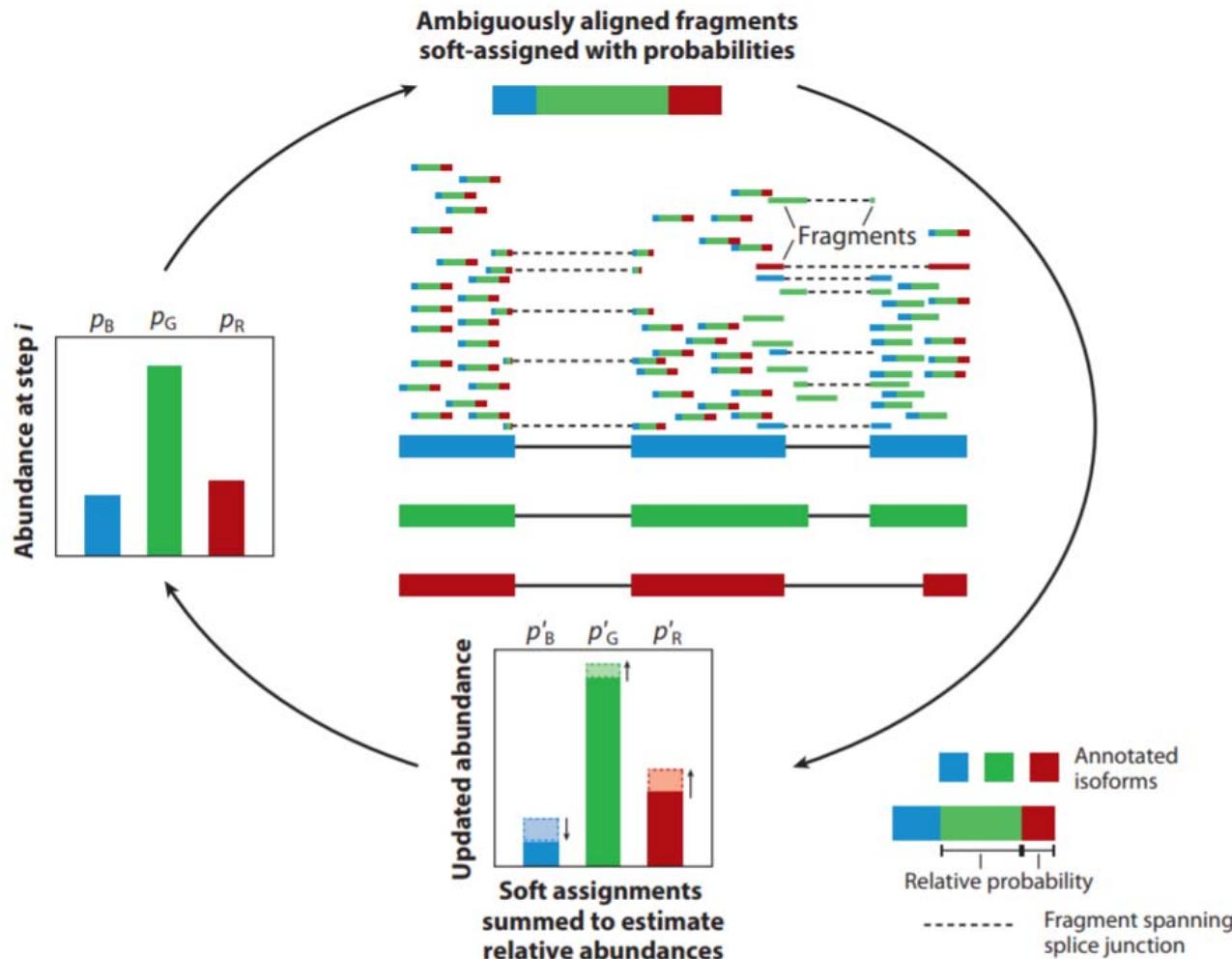
assignment of reads to different isoforms



Isoform	Fraction
I_1	θ_1
I_2	θ_2

We can use mapped reads to learn the isoform mixture $\{\theta\}$

Iterative assignment of reads to different isoforms by expectation maximization in RSEM



Tools for differential splicing analysis

rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data

Shihao Shen^{a,1}, Juw Won Park^{a,1}, Zhi-xiang Lu^a, Lan Lin^a, Michael D. Henry^{b,c}, Ying Nian Wu^d, Qing Zhou^d, and Yi Xing^{a,2}

**A new view of transcriptome complexity
and regulation through the lens of local
splicing variations**

Jorge Vaquero-Garcia^{1,2†}, Alejandro Barrera^{1,2†}, Matthew R Gazzara^{1,3†},
Juan Gonzalez-Vallinas^{1,2}, Nicholas F Lahens⁴, John B Hogenesch⁴,
Kristen W Lynch^{1,3}, Yoseph Barash^{1,2*}

PennDiff: detecting differential alternative splicing and transcription by RNA sequencing

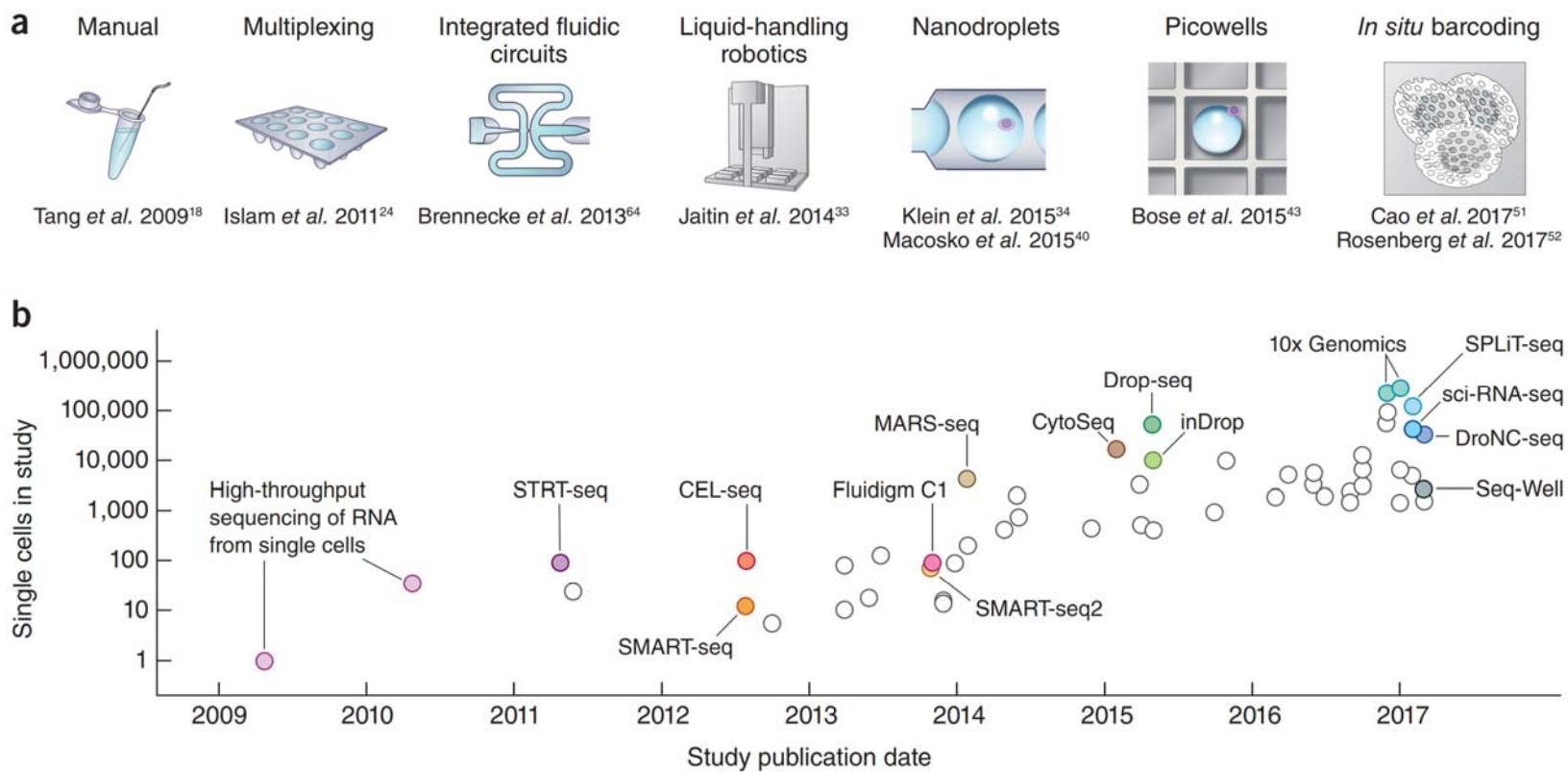
Yu Hu¹, Jennie Lin², Jian Hu¹, Gang Hu³, Kui Wang³, Hanrui Zhang⁴,
Muredach P. Reilly⁴ and Mingyao L¹

**Detecting differential usage of exons from
RNA-seq data**

Simon Anders,^{1,2} Alejandro Reyes,¹ and Wolfgang Huber

European Molecular Biology Laboratory, 69111 Heidelberg, Germany

A decade of single-cell RNA-seq

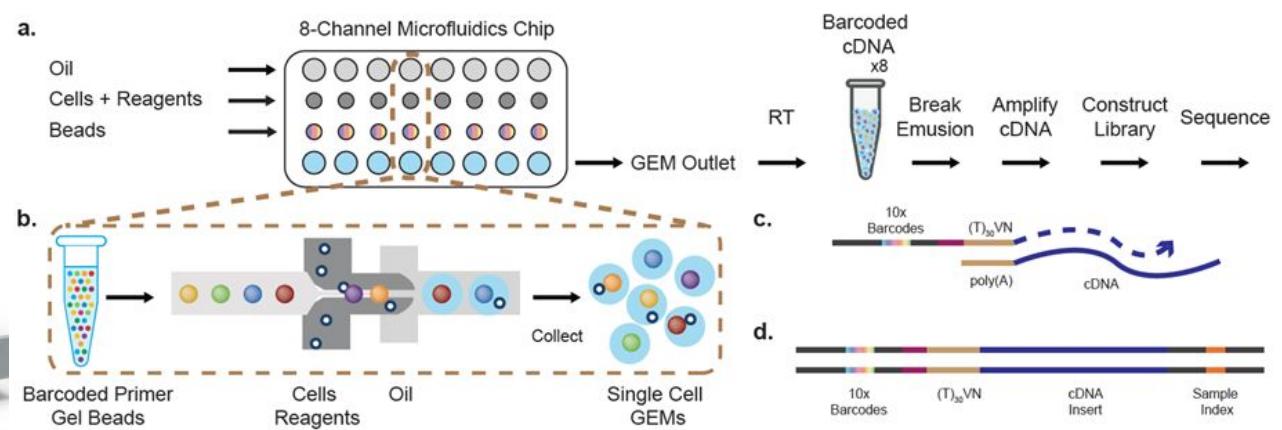


Comparison of different scRNA-Seq technologies

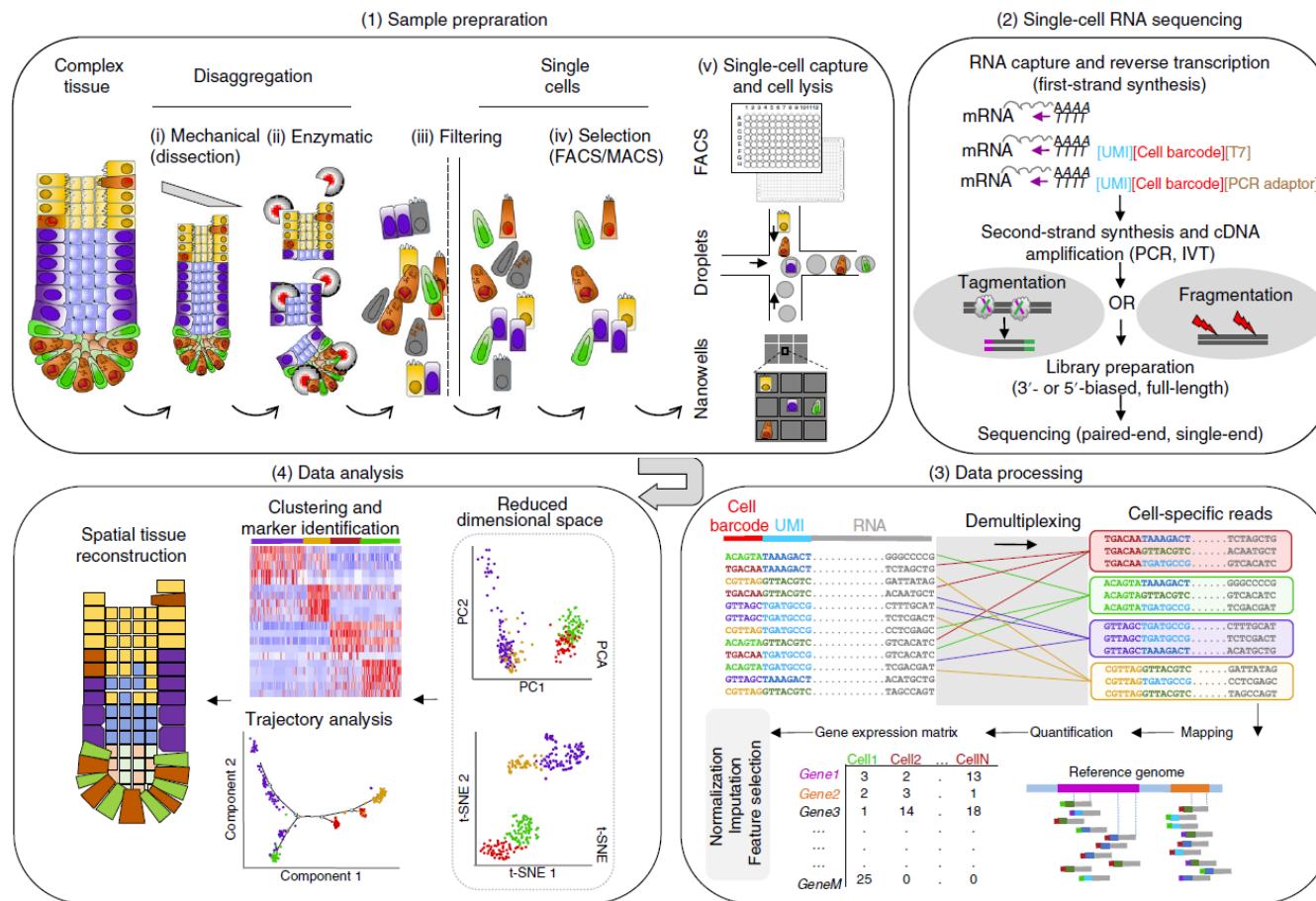
	SMART-seq2	CEL-seq2	STRT-seq	Quartz-seq2	MARS-seq	Drop-seq	inDrop	Chromium	Seq-Well	sci-RNA-seq	SPLiT-seq
Single-cell isolation	FACS, microfluidics	FACS, microfluidics	FACS, microfluidics, nanowells	FACS	FACS	Droplet	Droplet	Droplet	Nanowells	Not needed	Not needed
Second strand synthesis	TSO	RNase H and DNA pol I	TSO	PolyA tailing and primer ligation	RNase H and DNA pol I	TSO	RNase H and DNA pol I	TSO	TSO	RNase H and DNA pol I	TSO
Full-length cDNA synthesis?	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes
Barcode addition	Library PCR with barcoded primers	Barcoded RT primers	Barcoded TSOs	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers and library PCR with barcoded primers	Ligation of barcoded RT primers
Pooling before library?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Library amplification	PCR	In vitro transcription	PCR	PCR	In vitro transcription	PCR	In vitro transcription	PCR	PCR	PCR	PCR
Gene coverage	Full-length	3'	5'	3'	3'	3'	3'	3'	3'	3'	3'
Number of cells per assay	10 ²	10 ²	10 ²	10 ²	10 ²	10 ³	10 ³	10 ³	10 ³	10 ⁴	10 ⁴

Chromium system from 10x Genomics

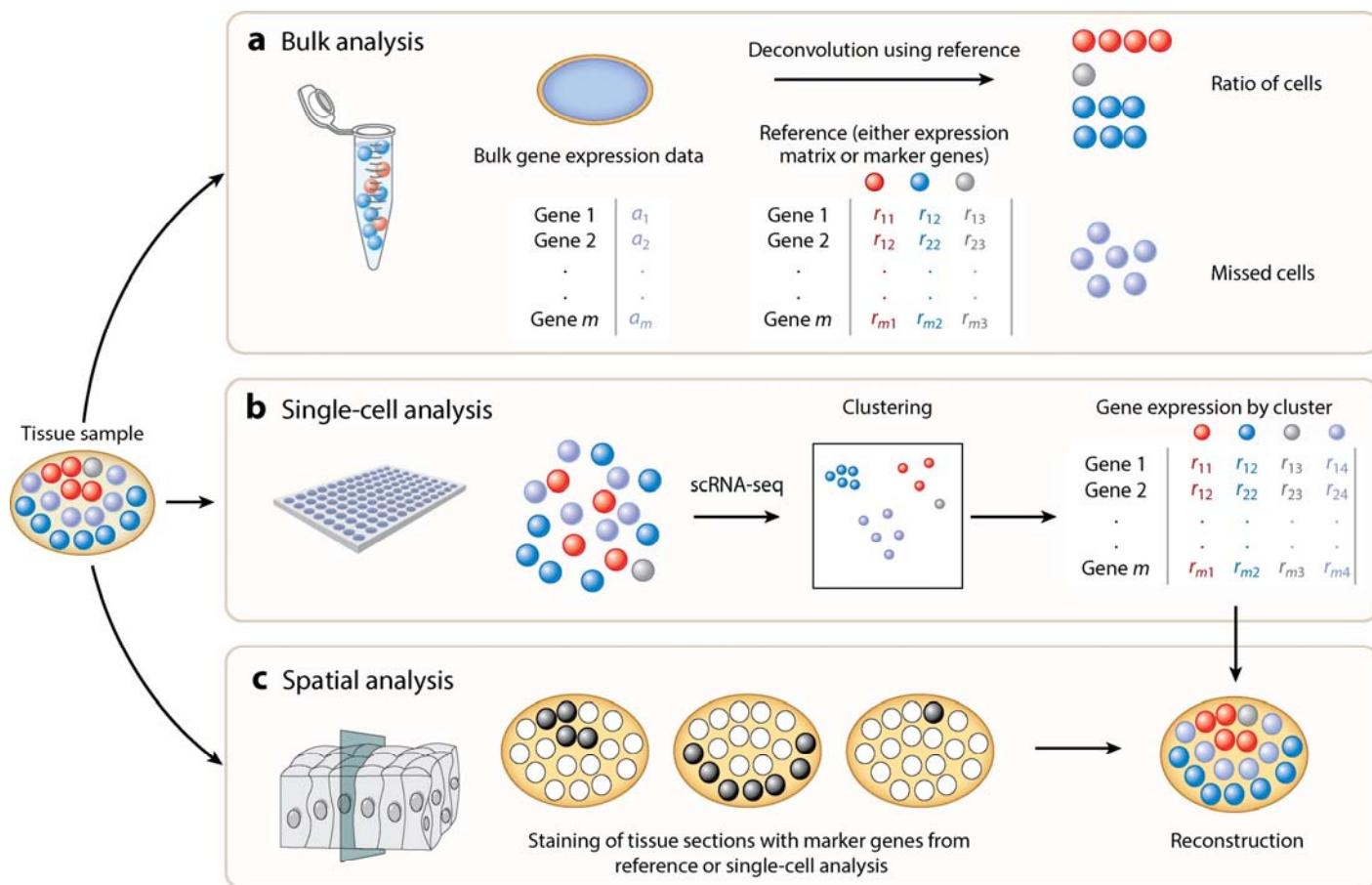
- Offers high-throughput profiling of 3' end of RNAs of single cells with high capture efficiency.
- Enables analysis of rare cell types in a sufficiently heterogeneous biological space.
- This technology encapsulates 500 to 20,000 cells or nuclei per library together with micro-beads into nano-droplets.
- Each bead is loaded with adapters containing one of 750,000 different barcodes for the single cell RNA-seq library preps.



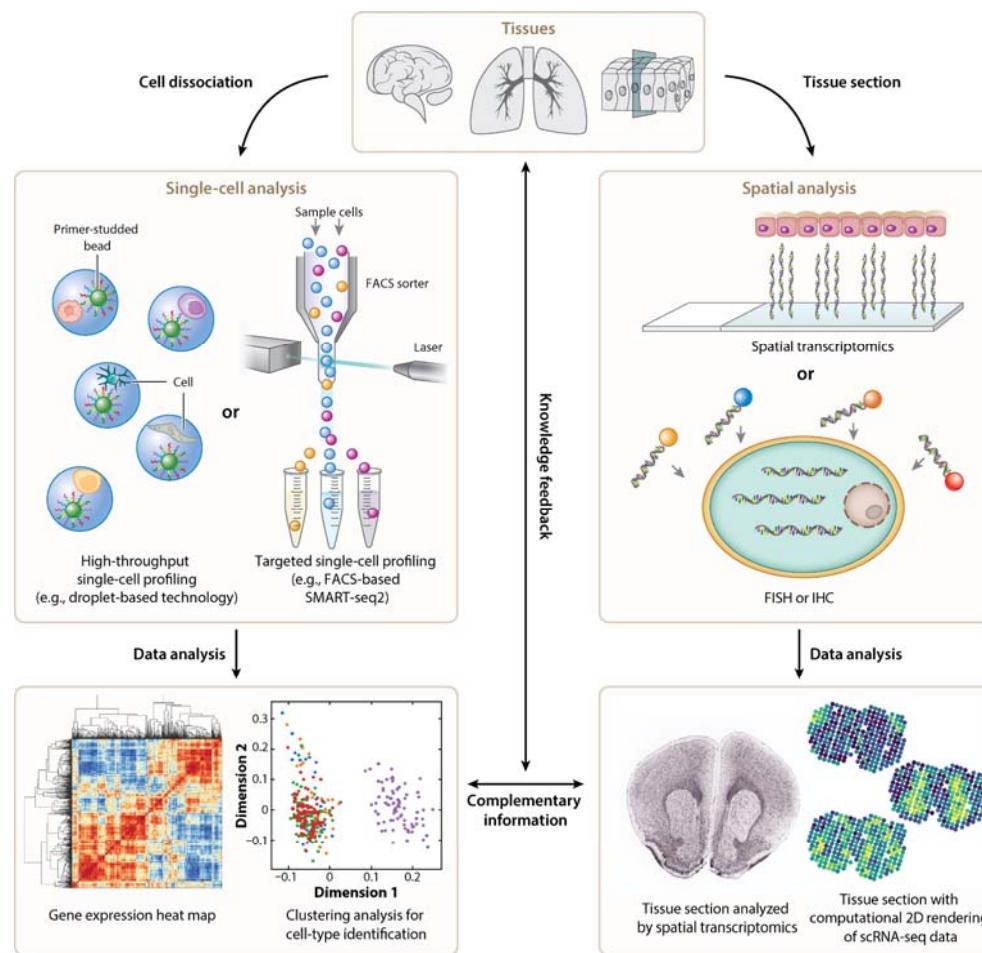
Single cell RNA-Seq: typical procedure



Extension to spatial analysis



Robust deconvolution of cell types: combining scRNA-Seq with spatial information



Commercial solutions will be available soon

← Tweet



10x Genomics
@10xGenomics

Excited to be getting closer to the launch of Visium Spatial Gene Expression Solution, our new product that allows scientists to measure all gene activity within a tissue sample & map where activity occurs. Learn the types of discoveries possible. bit.ly/2JMGVON #Spatial



1:17 | 944 views

4:07 PM · Jul 18, 2019 · Sprout Social