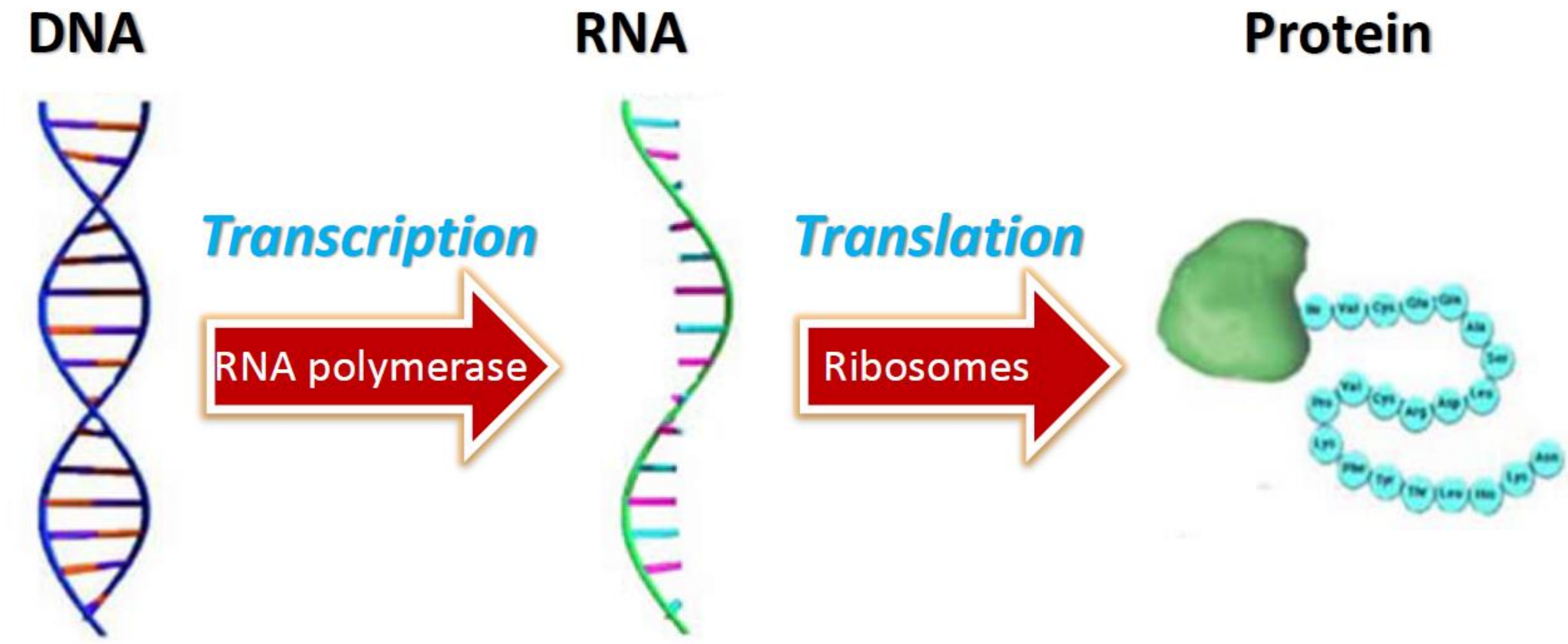
RNA-Seq in human diseases

2019 Dragon Star Bioinformatics Course (Day 5)

Central Dogma of Biology



DNA

RNA

Protein

Transcription

RNA polymerase

Translation

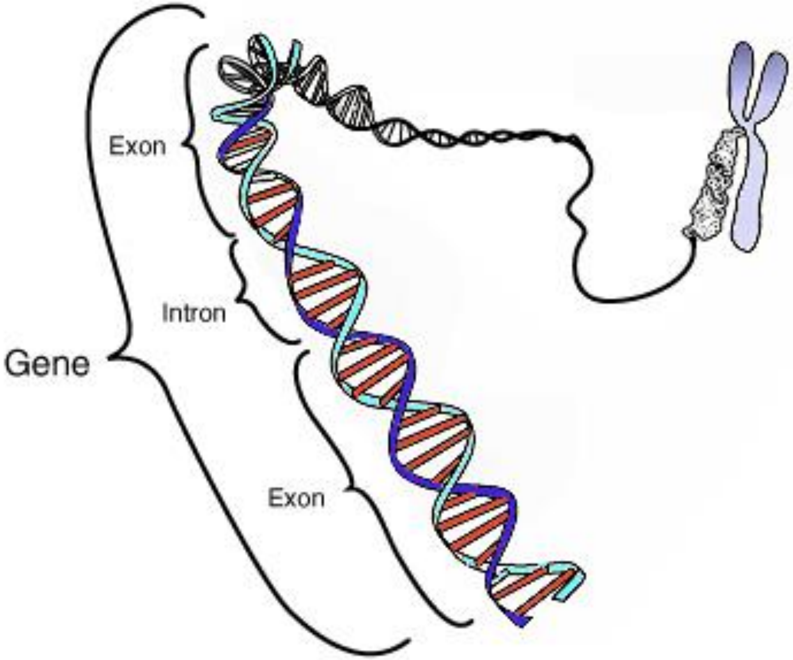
Ribosomes

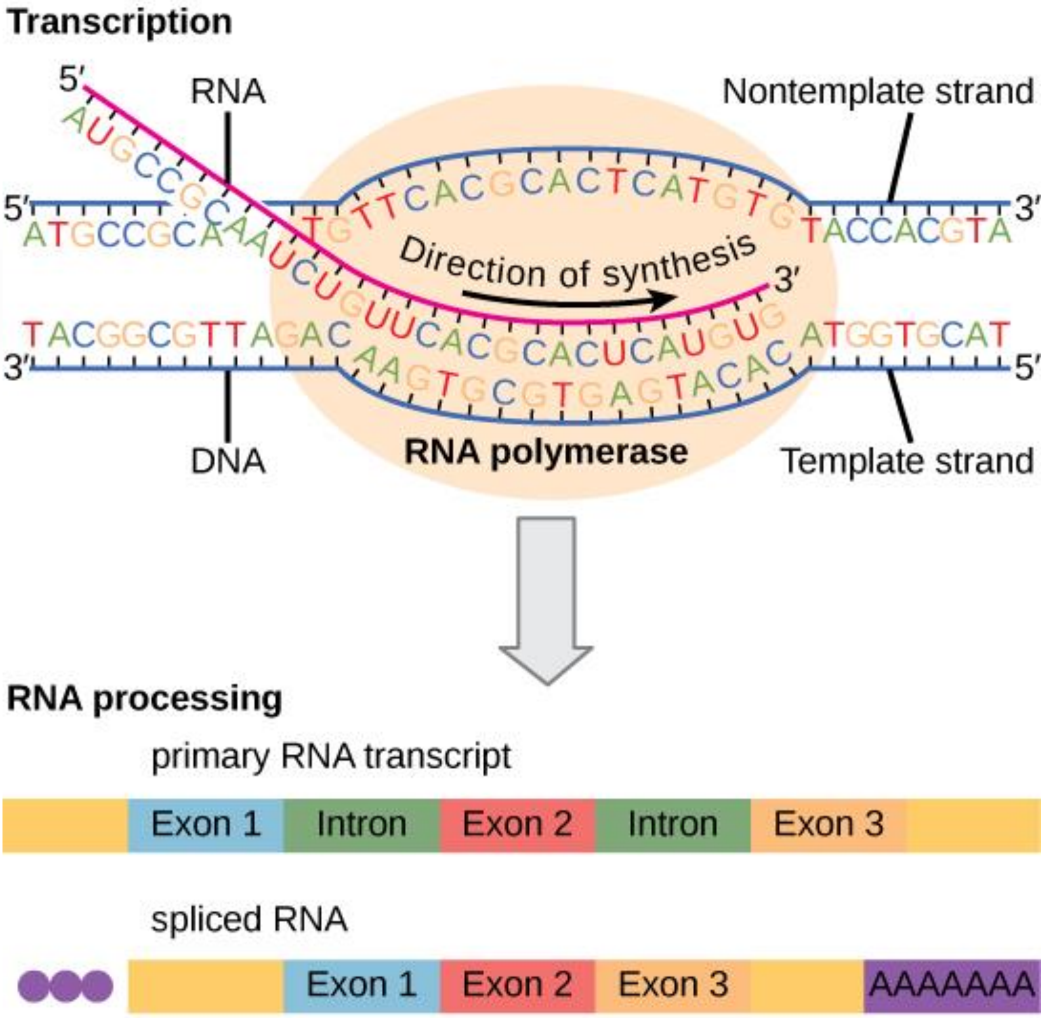
GWAS, exome-seq, whole-genome sequencing

Microarray, RNA sequencing

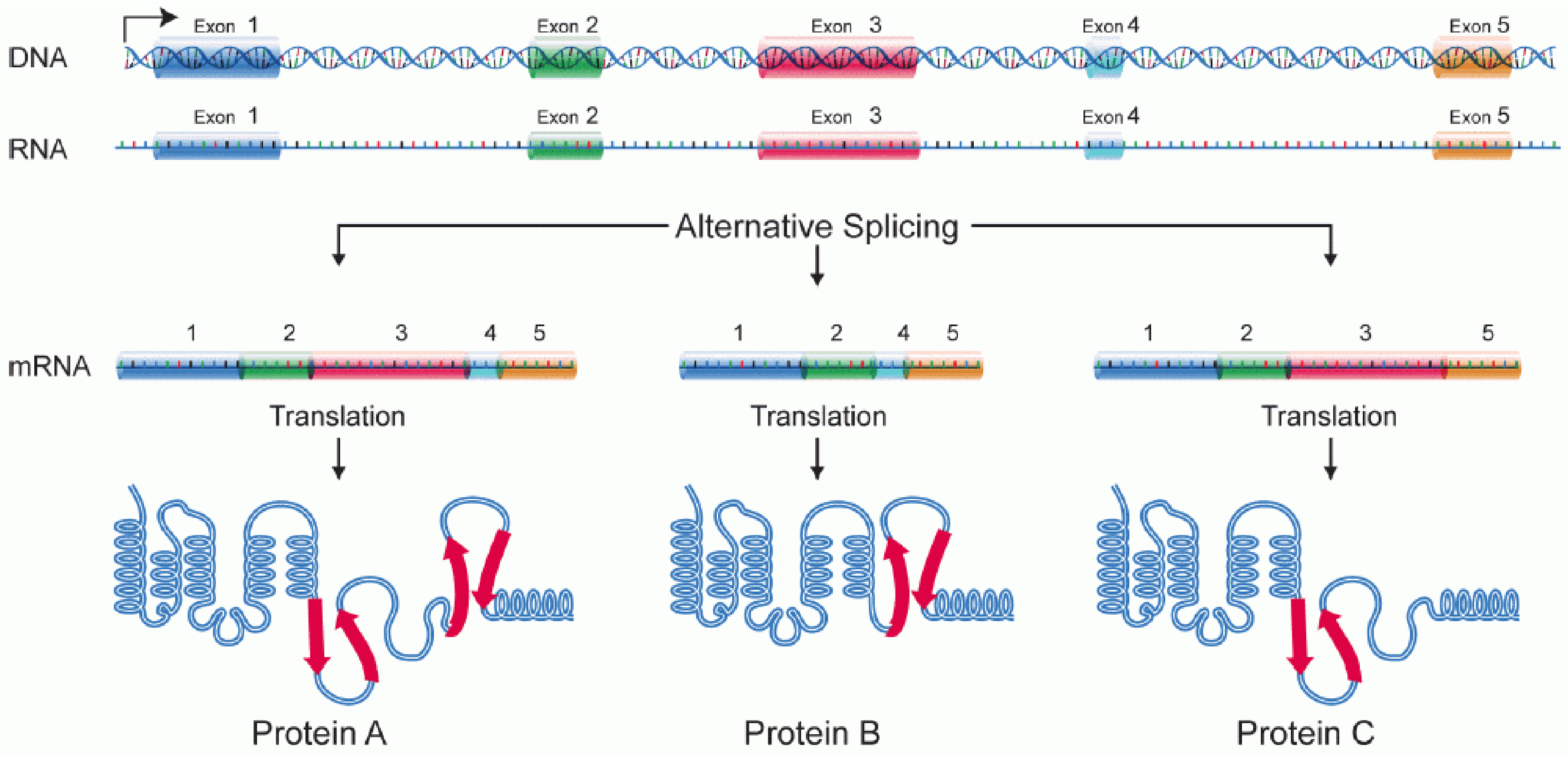
Mass spectrometry
Somalogic

Gene Transcription





Alternative Splicing



~90% of human genes are alternatively spliced

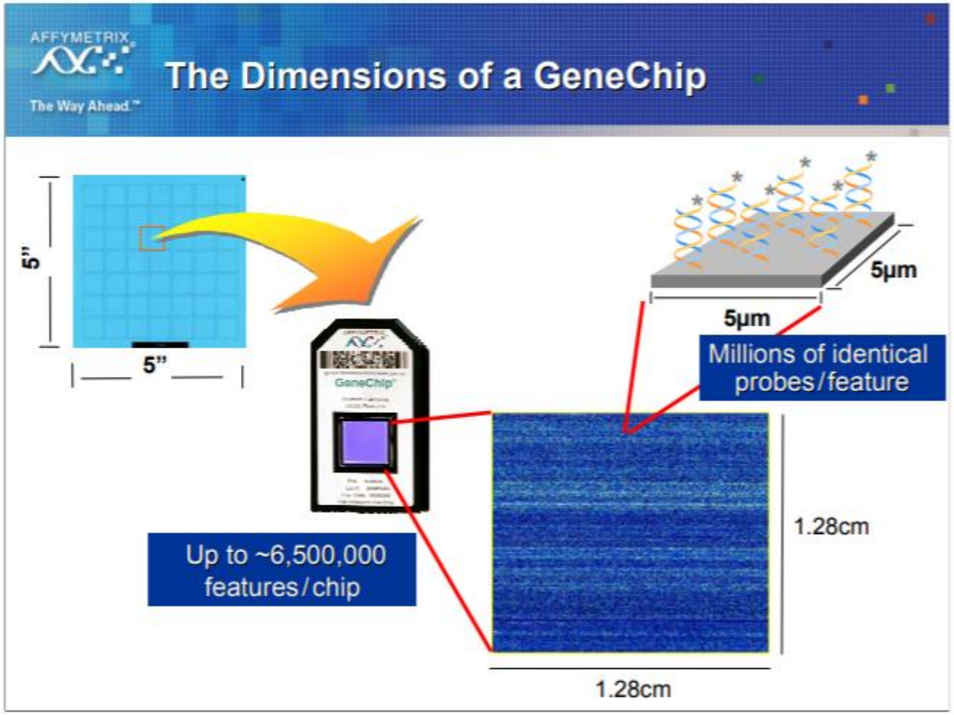
Alternatively splicing substantially increased transcriptome complexity.

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

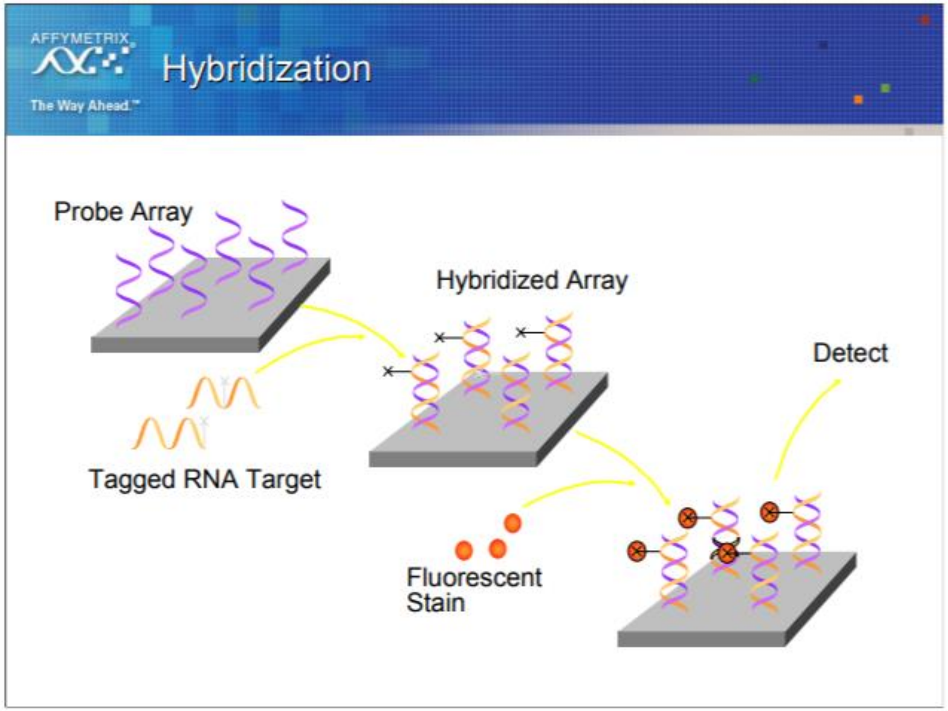


AFFYMETRIX The Way Ahead.™

The Dimensions of a GeneChip

Millions of identical probes/feature

Up to ~6,500,000 features/chip



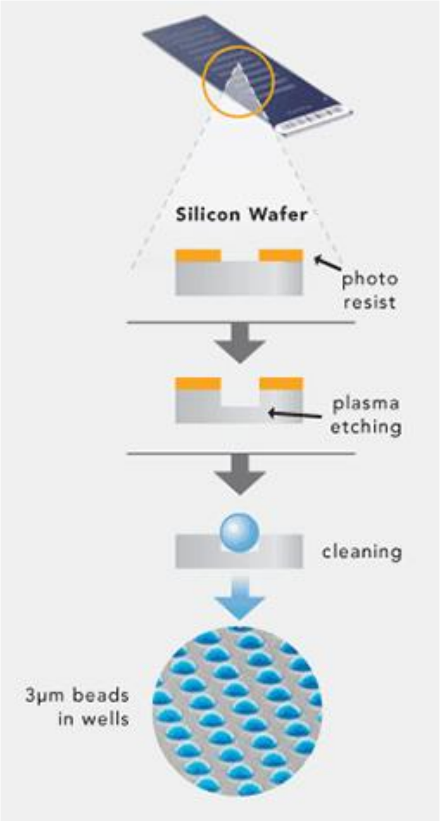
AFFYMETRIX The Way Ahead.™

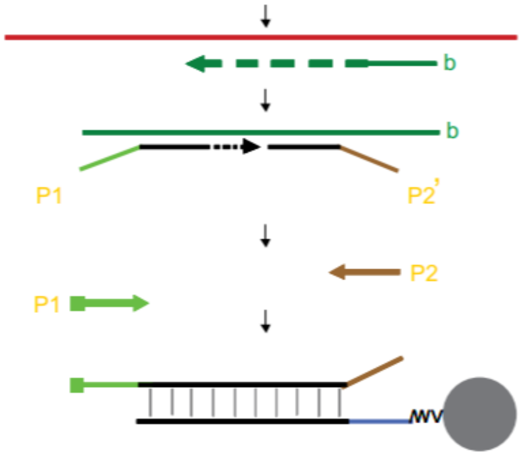
Hybridization

www.Affymetrix.com

Technologies for Transcriptomics: expression microarrays

- Illumina BeadChip can also be used for gene expression analysis





cDNA synthesis

Query oligo (DAP) annealing, extension, and ligation

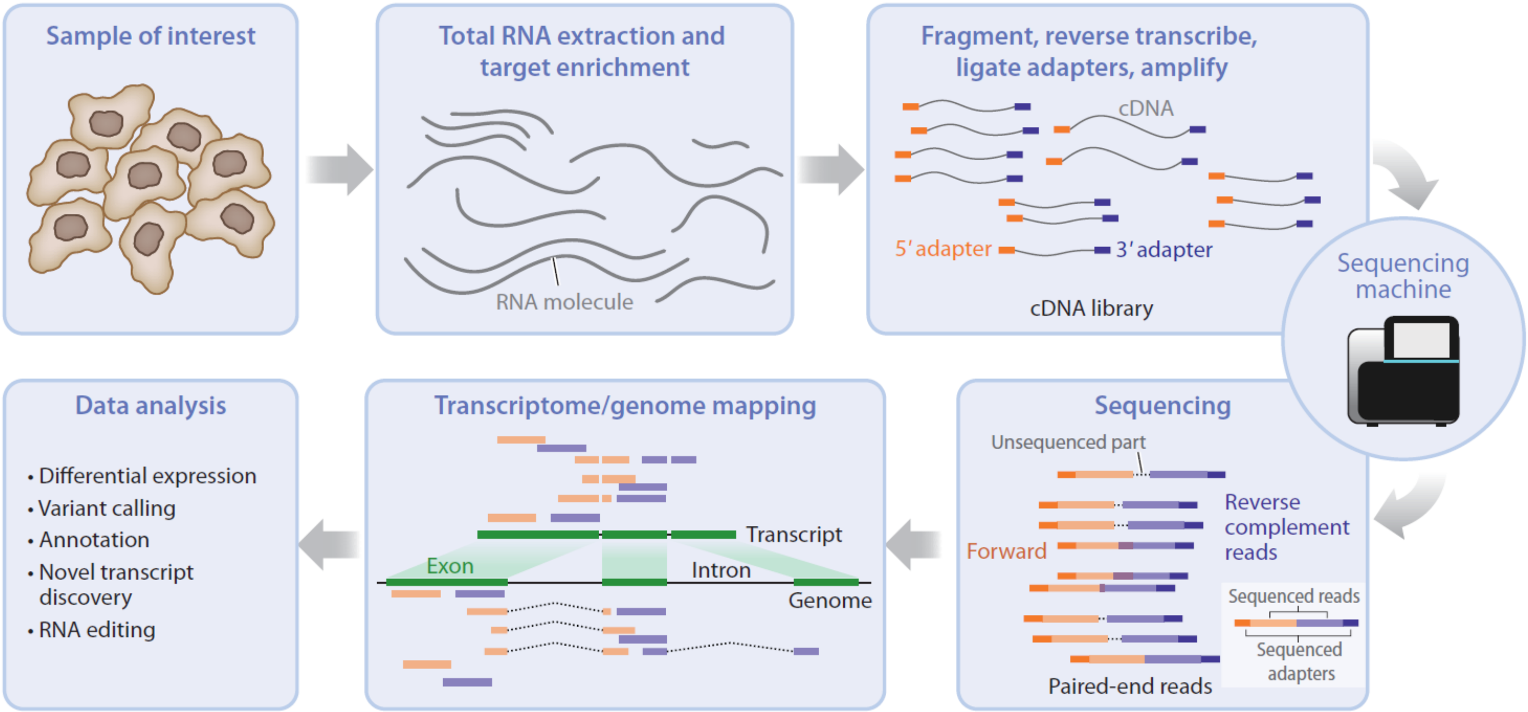
PCR with common primers

Product captured by hybridization to BeadChip

www.illumina.com

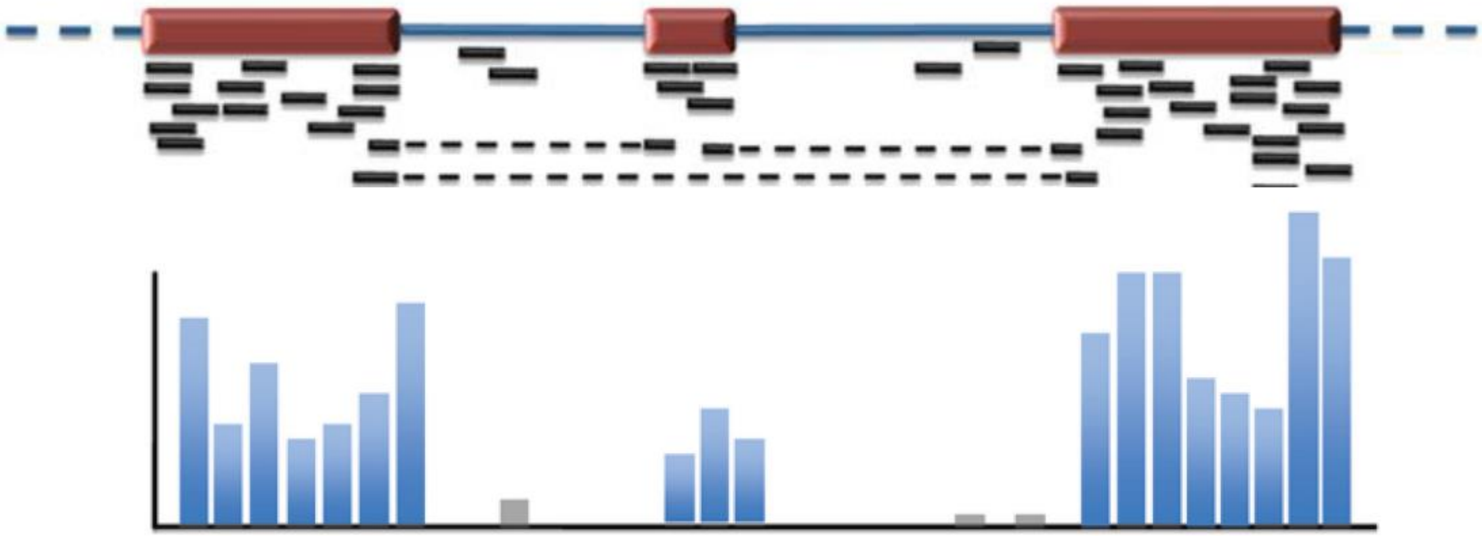
Technologies for Transcriptomics: RNA Sequencing (RNA-Seq)

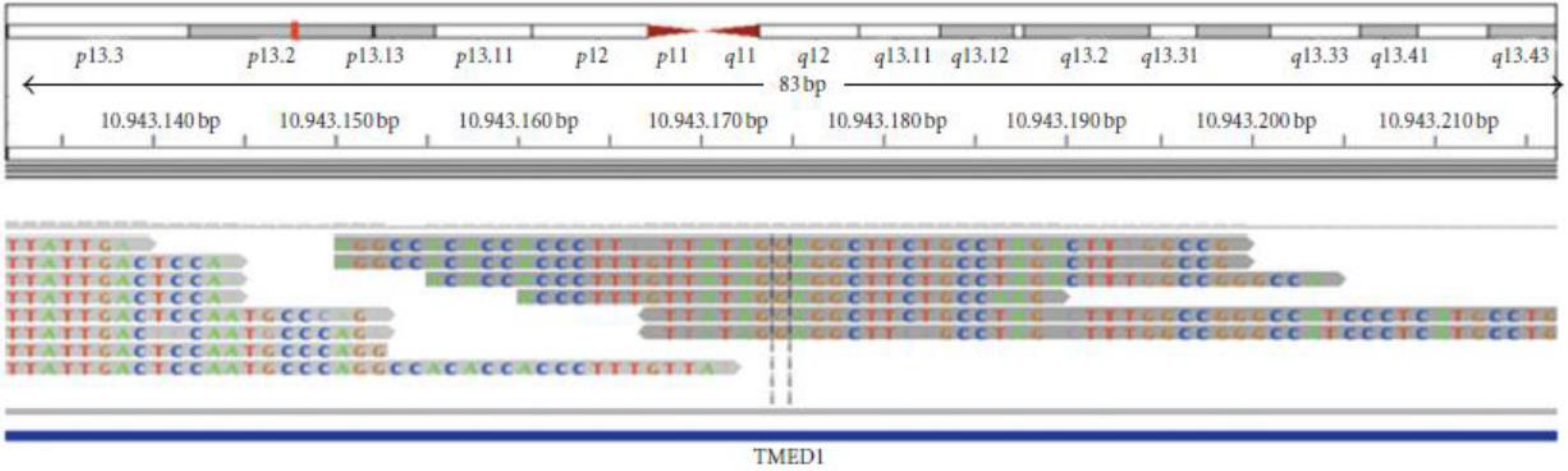
- RNA-Seq gradually replaced expression microarrays as the most widely used methods for gene expression analysis



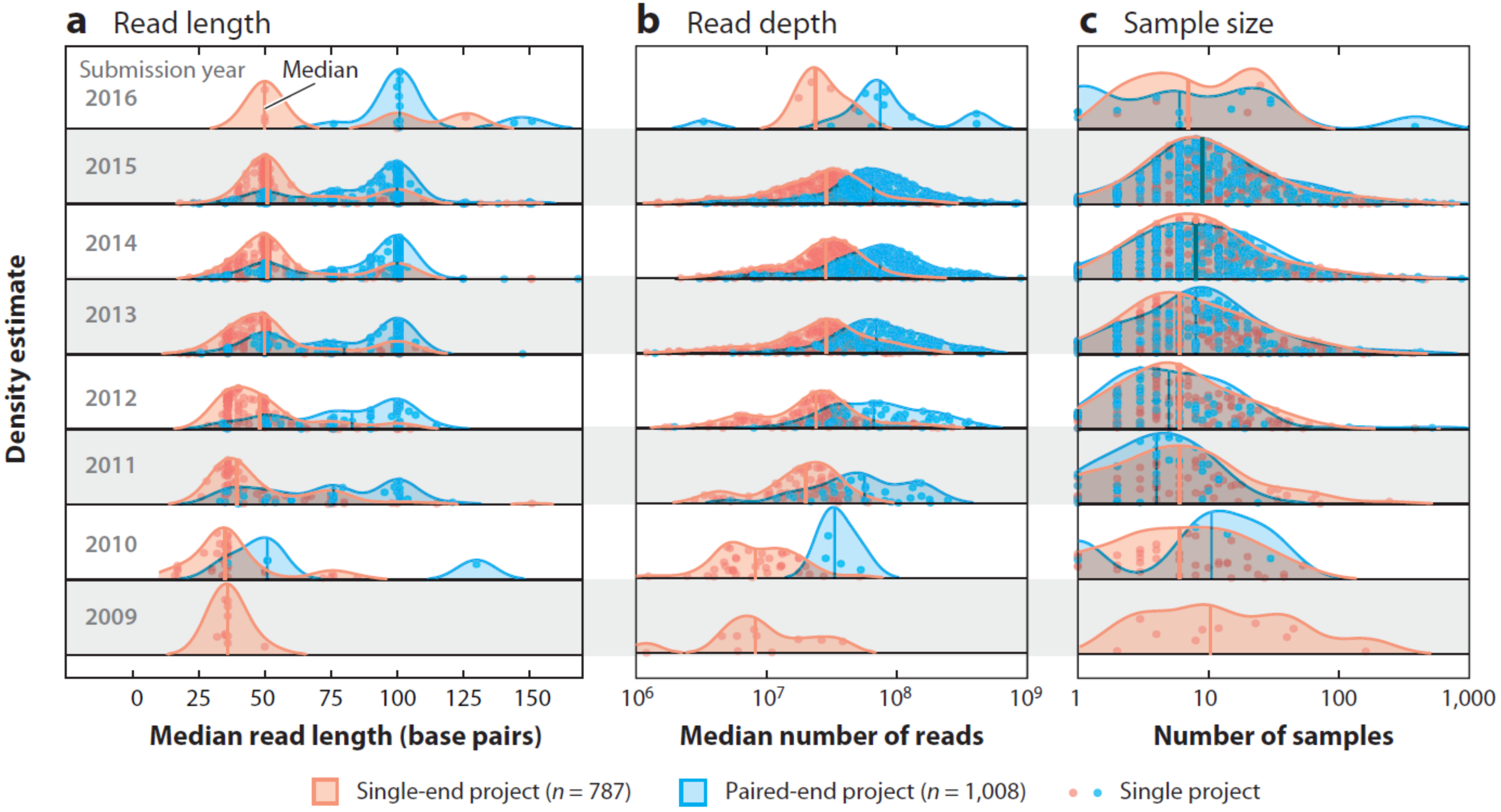
Van den Berge et al, Annual Review of Biomedical Data Science, 2019

Information in RNA-seq Data





Evolution of experimental design over the past several years



Van den Berge et al, Annual Review of Biomedical Data Science, 2019

Advantages of RNA-Seq over Microarrays

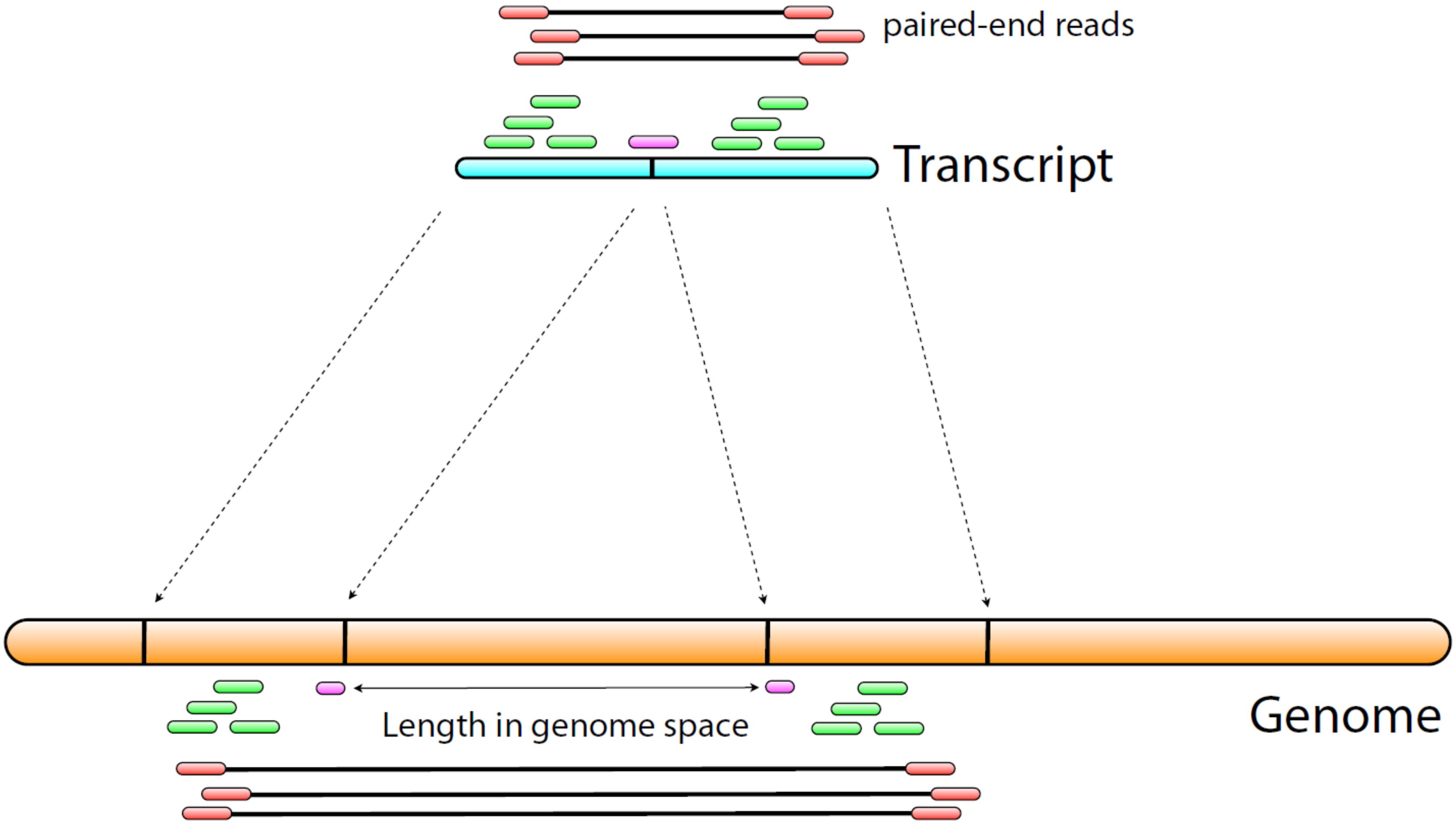
- Microarrays measure only genes corresponding to predetermined probes on a microarray

- RNA-seq
  - Measures <u>any expressed transcripts </u>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.

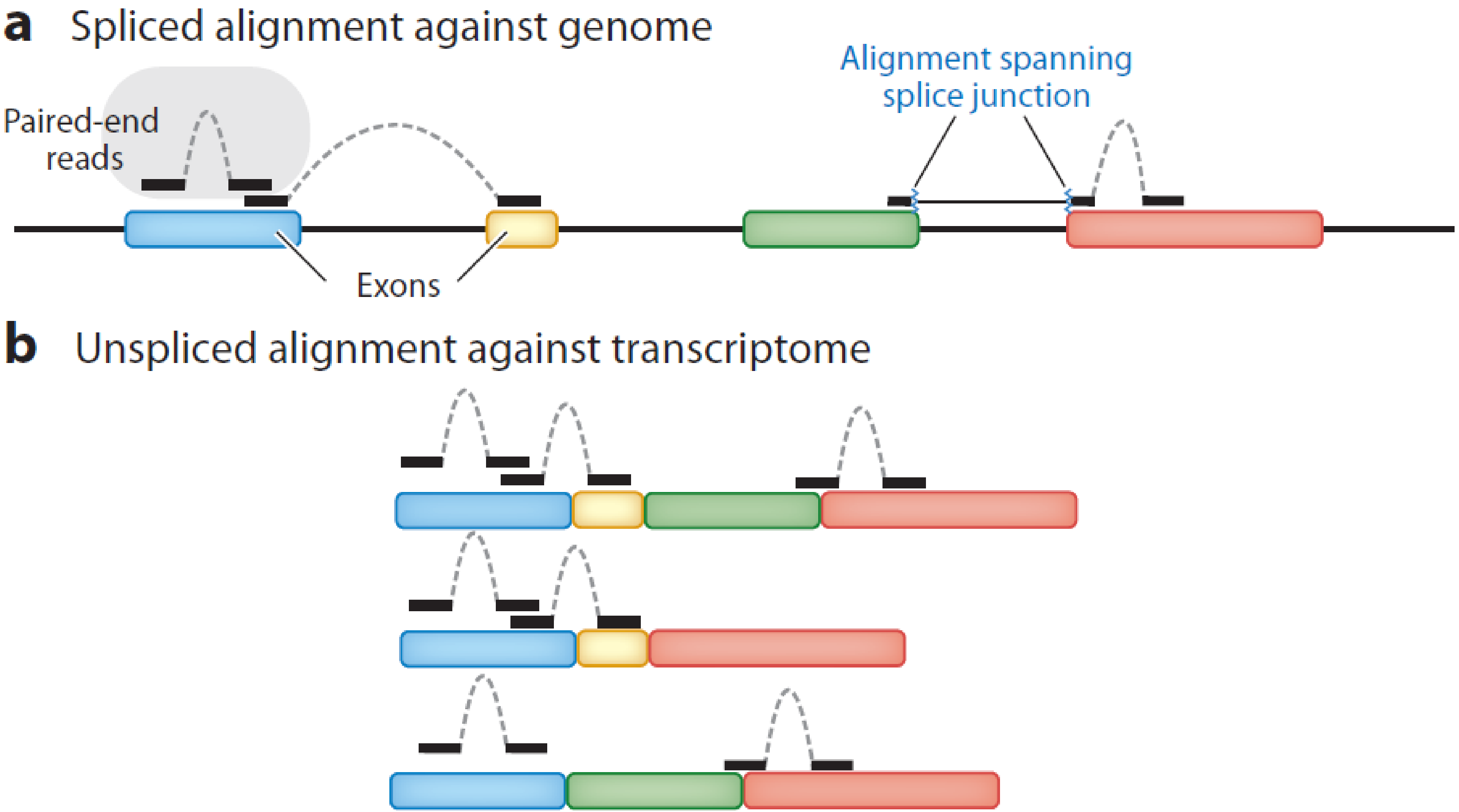
Advantages of RNA-Seq over Microarrays

- RNA-seq
  - 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

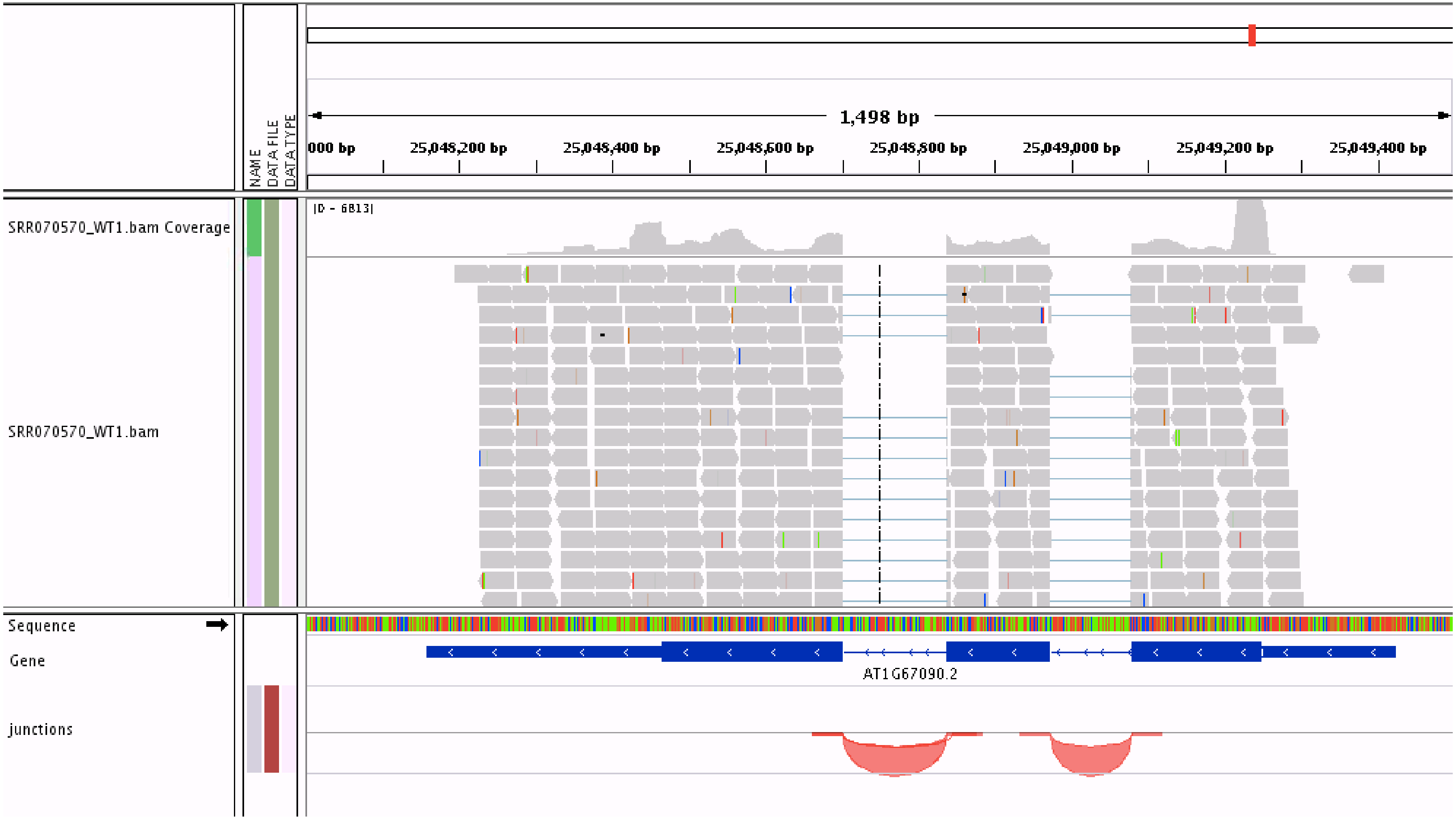


Splice-aware RNA-Seq alignment methods


a
Spliced alignment against genome
Alignment spanning splice junction
Paired-end reads
Exons
b
Unspliced alignment against transcriptome

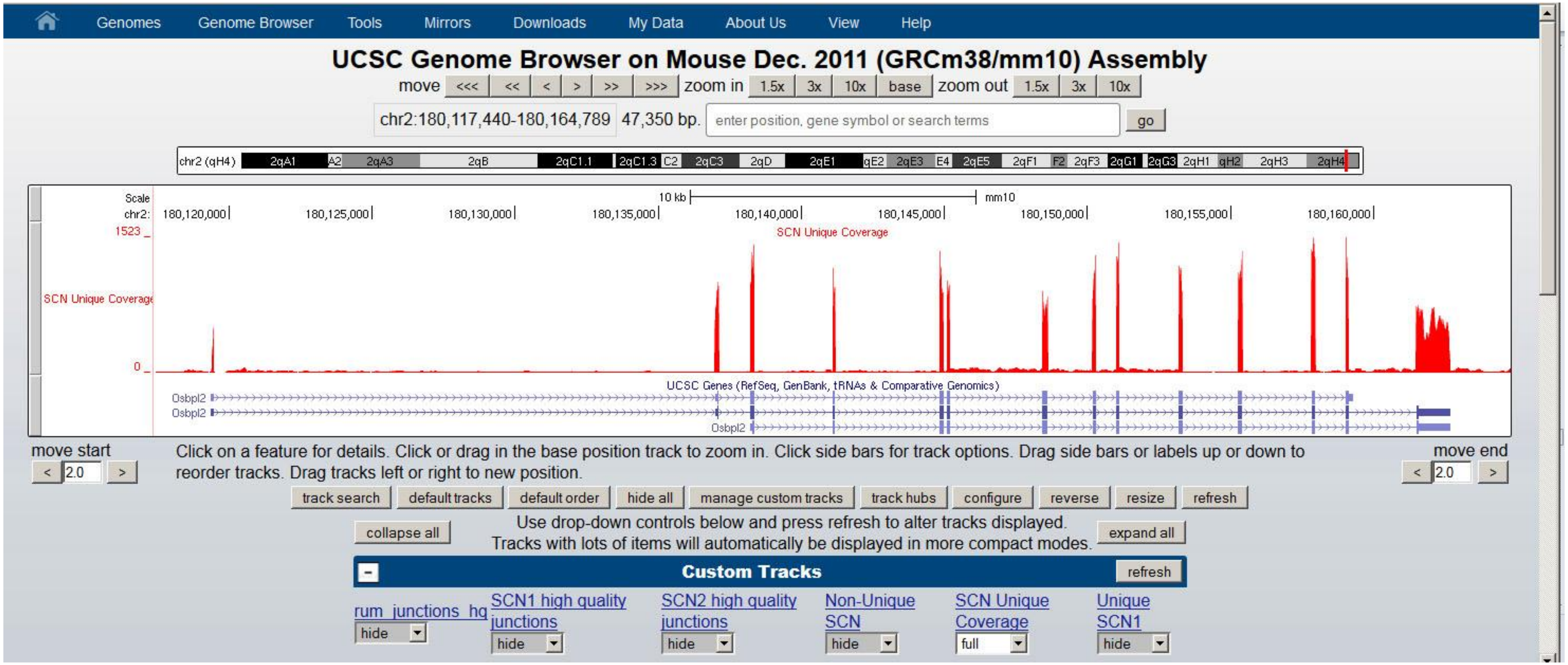
Van den Berge et al, Annual Review of Biomedical Data Science, 2019

What does the Aligned File Look Like?



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

What does the Aligned File Look Like?



UCSC Genome Browser on Mouse Dec. 2011 (GRCm38/mm10) Assembly

chr2:180,117,440-180,164,789 47,350 bp.

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

Use drop-down controls below and press refresh to alter tracks displayed. Tracks with lots of items will automatically be displayed in more compact modes.

Custom Tracks

rum_junctions_hq: hide

SCN1 high quality junctions: hide

SCN2 high quality junctions: hide

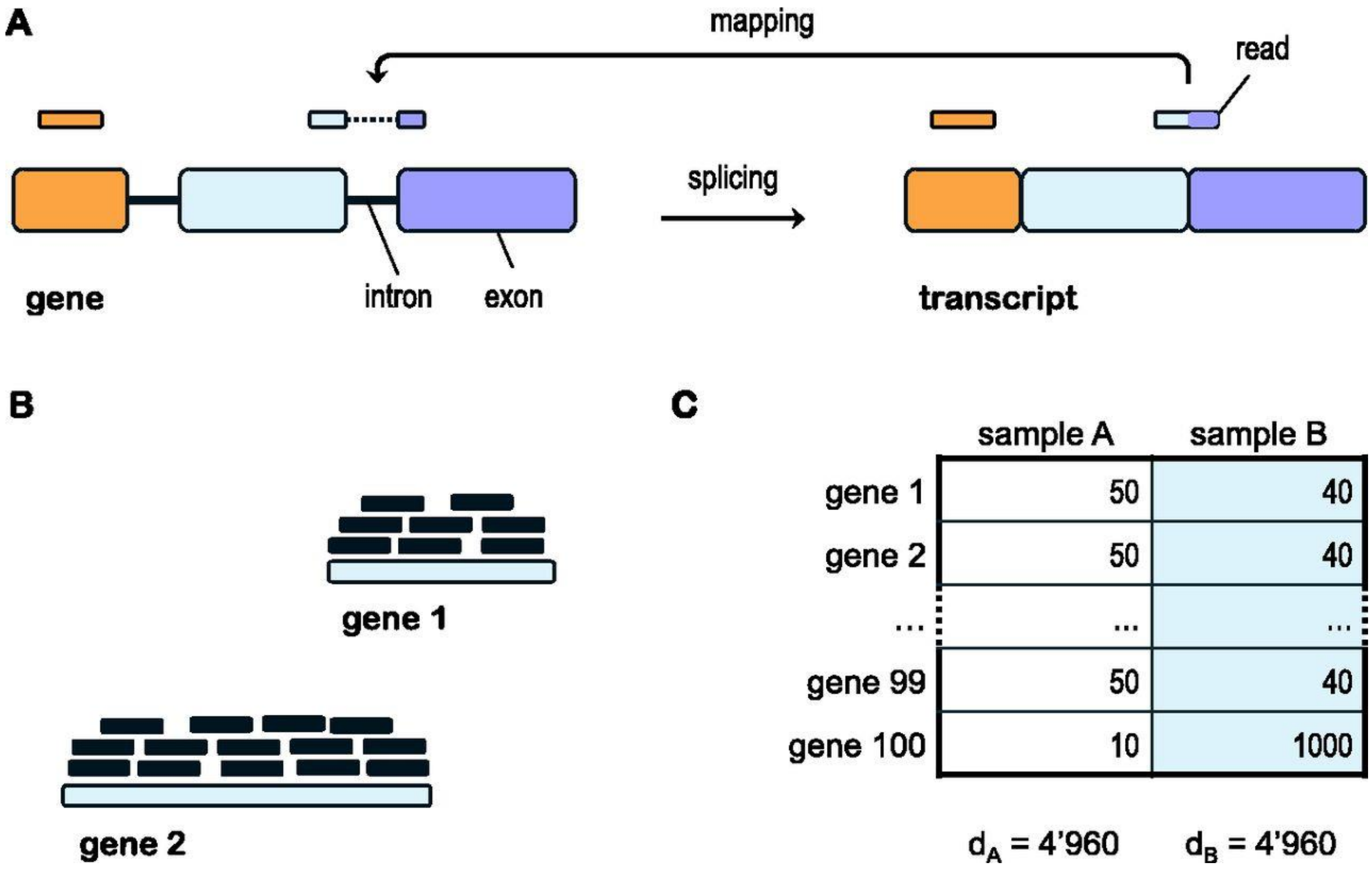
Non-Unique SCN: hide

SCN Unique Coverage: full

Unique SCN1: hide

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



	sample A	sample B
gene 1	50	40
gene 2	50	40
...	...	...
gene 99	50	40
gene 100	10	1000

d_A = 4’960 d_B = 4’960

Finotello et al, Briefings in Functional Genomics, 2015

Selected RNA-Seq Alignment Programs

Short-read RNA-Seq

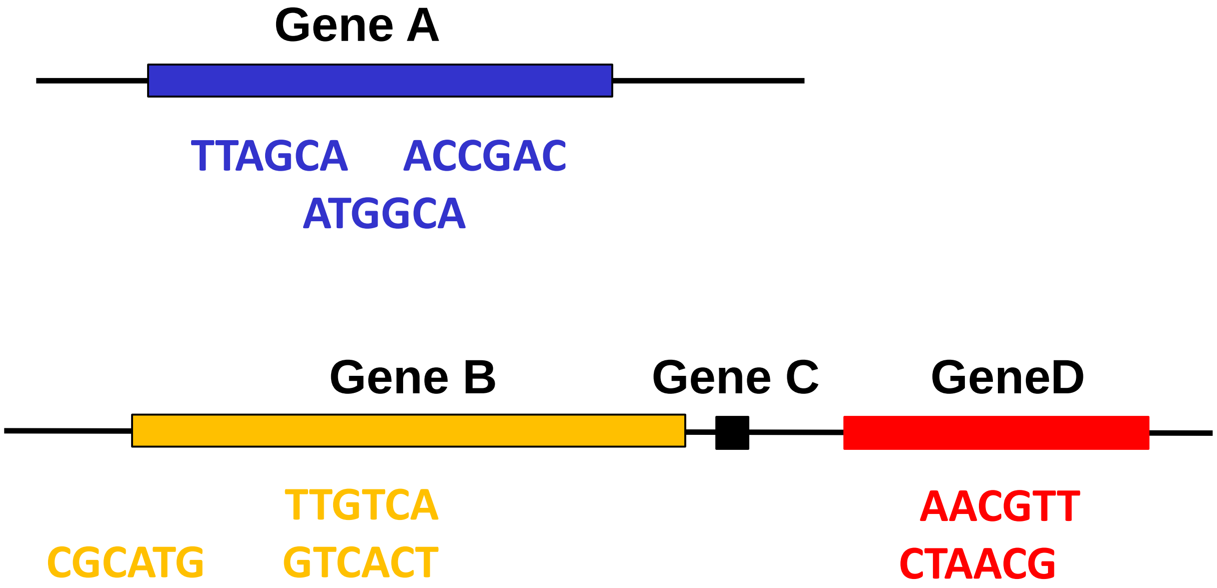
- TopHat2
- HISAT2
- MapSplice
- SOAPSplice
- SpliceMap
- RUM
- GSNAP
- STAR

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)

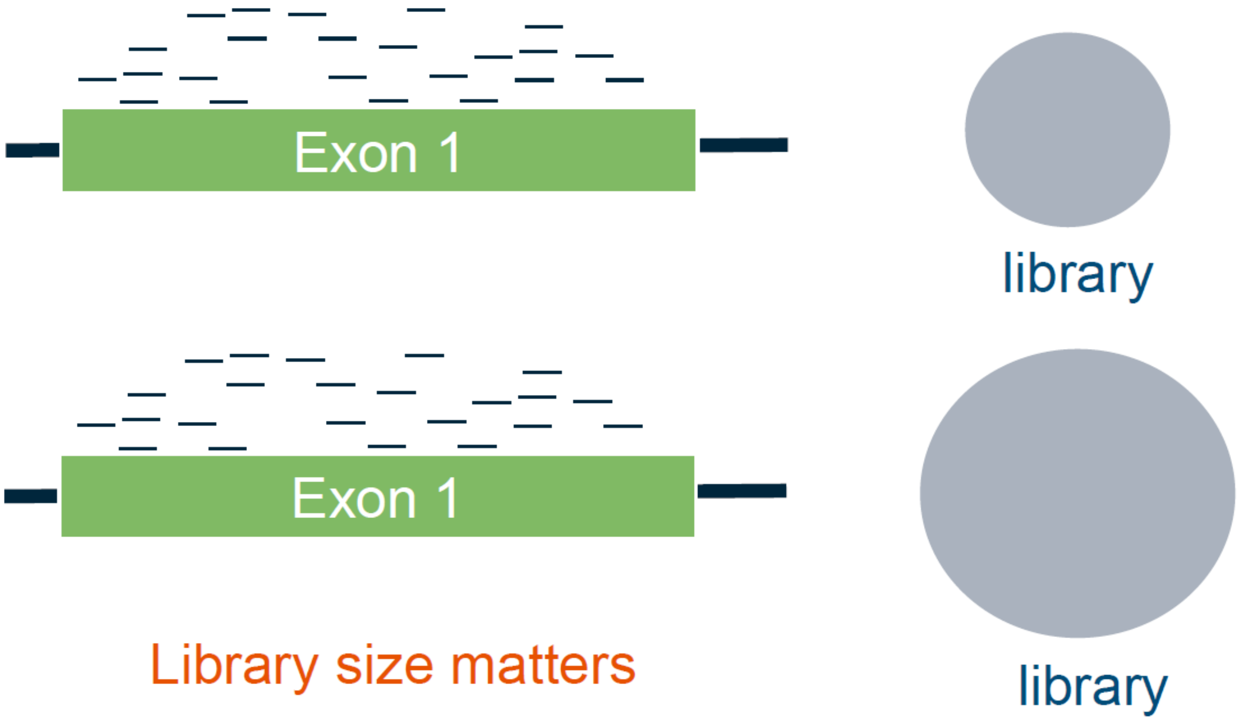
Quantification of Gene Expression



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.

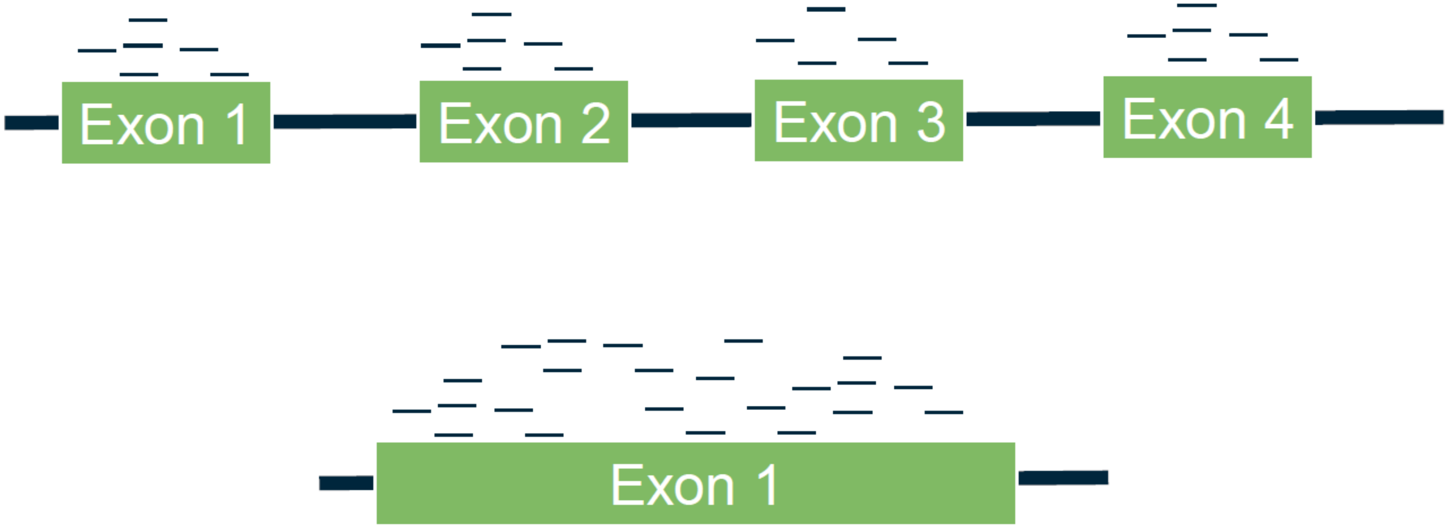
Normalization is Important



Library size matters

Larger library has more reads

Normalization is Important

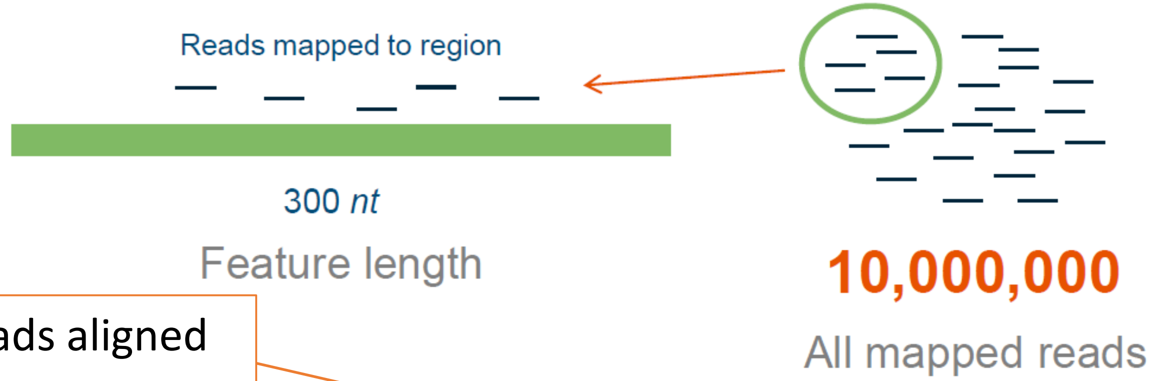


Transcript/gene length matters

Larger transcript/gene has more reads

Normalization: RPKM

Reads **P**er **K**ilobase of transcript per **M**illion mapped reads



X_i is number of reads aligned to gene i

$$RPKM_i = \frac{X_i}{\left(\frac{\widetilde{l_i}}{10^3}\right)\left(\frac{N}{10^6}\right)} = \frac{X_i}{\widetilde{l_i}N} \cdot 10^9$$

adjust gene length difference
$\widetilde{l_i}$ is gene length

adjust library size difference
N is total library size

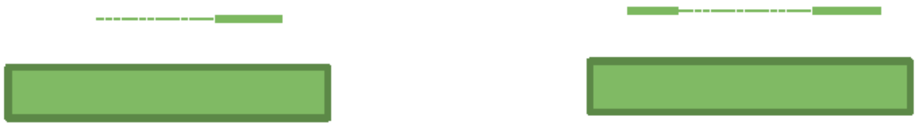
RPKM is used for single-end data

Normalization: FPKM

Fragments **P**er **K**ilobase of transcript per **M**illion mapped reads

FPKM is analogous to RPKM

Sequencing fragments



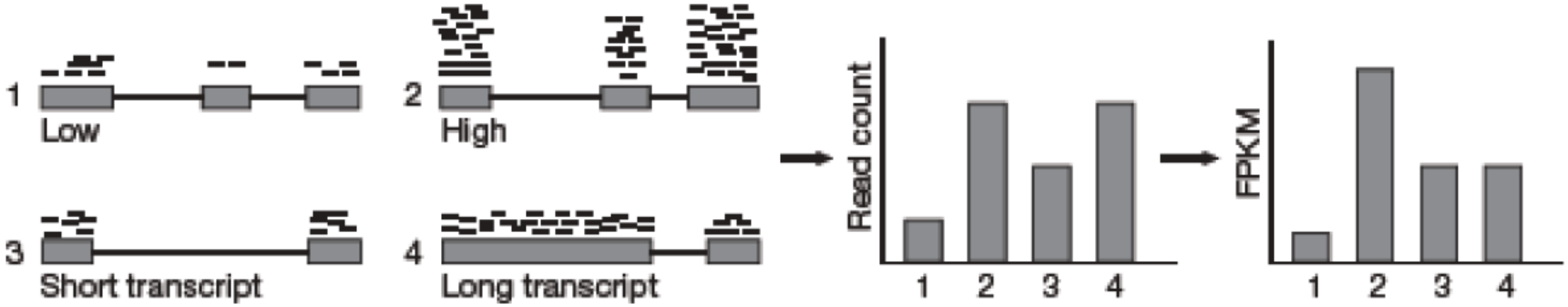
RPKM = 1

RPKM = 2

FPKM = 1

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.

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

RPKM, FPKM or TPM?

- TPM and RPKM/FPKM normalization methods both account for sequencing depth and gene length
    - But 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
  - 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

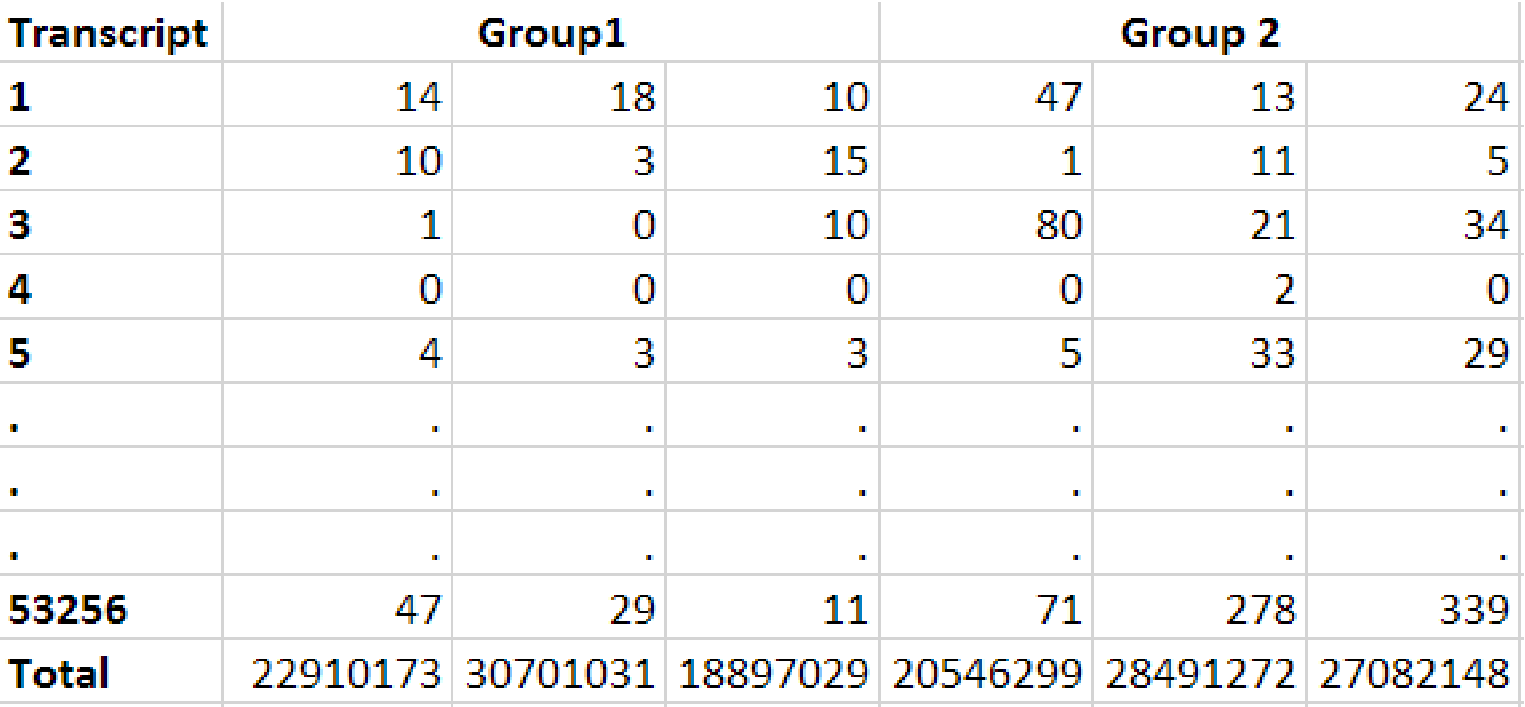
$$\mathrm{TPM}_i = \frac{X_i}{\widetilde{l}_i} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\widetilde{l}_j}} \right) \cdot 10^6 \propto \frac{X_i}{\widetilde{l}_i \cdot N} \cdot \left(\frac{1}{\sum_j \frac{X_j}{\widetilde{l}_j \cdot N}} \right)$$

$$\propto \frac{X_i}{\widetilde{l}_i \cdot N} \cdot 10^9$$

If you have FPKM, you can easily compute TPM:

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

Differential Gene Expression



Transcript	Group1			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 ~ Poisson(λ) implies

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

- From the fit of the generalized linear model, we can
  - Estimate count means and variances
  - Assess whether the Poisson mean-variance relationship holds.

- The data are said to be over-dispersed
  - When the actual counts are more variable than we would expect based on the Poisson assumption

Overdispersion in Real RNA-Seq Data

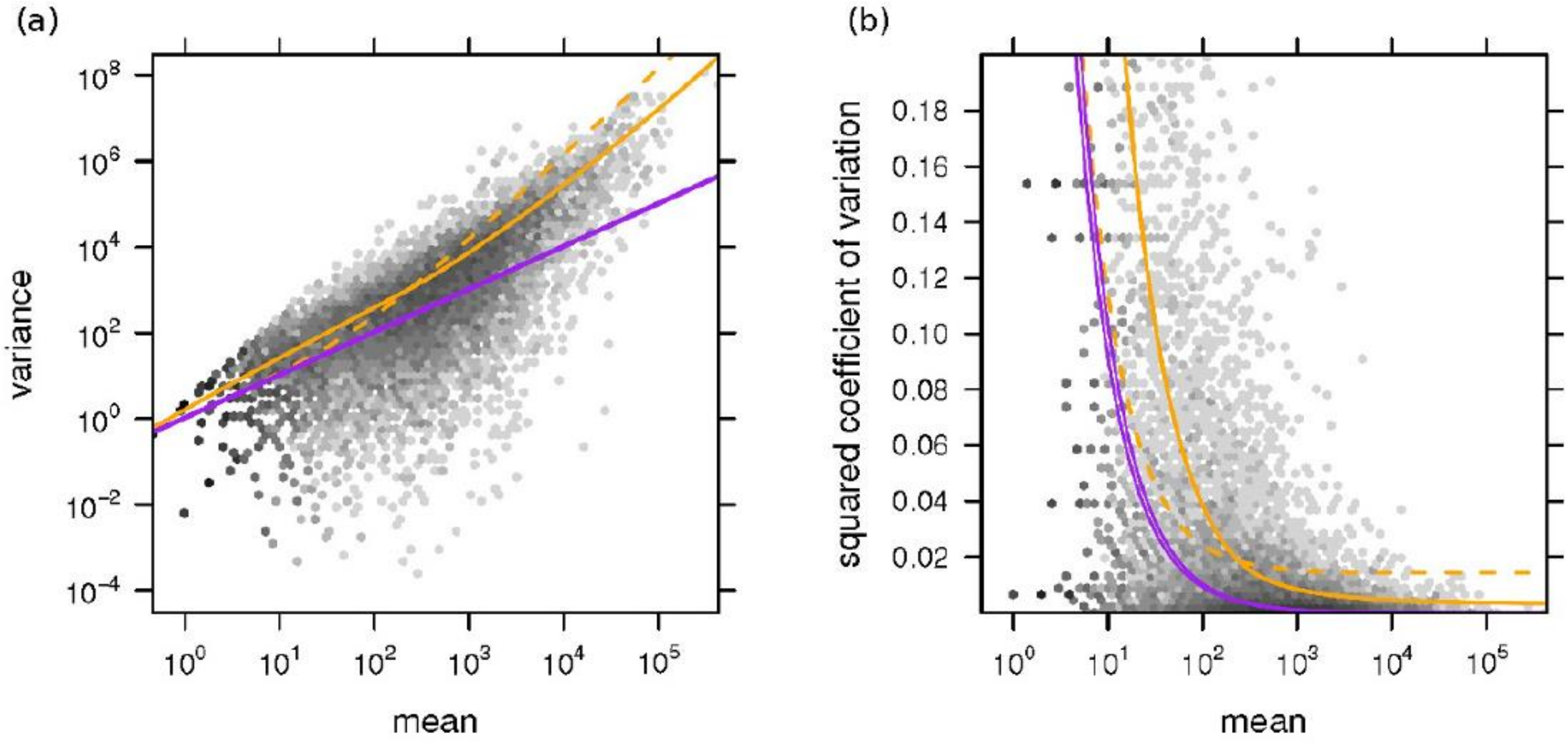
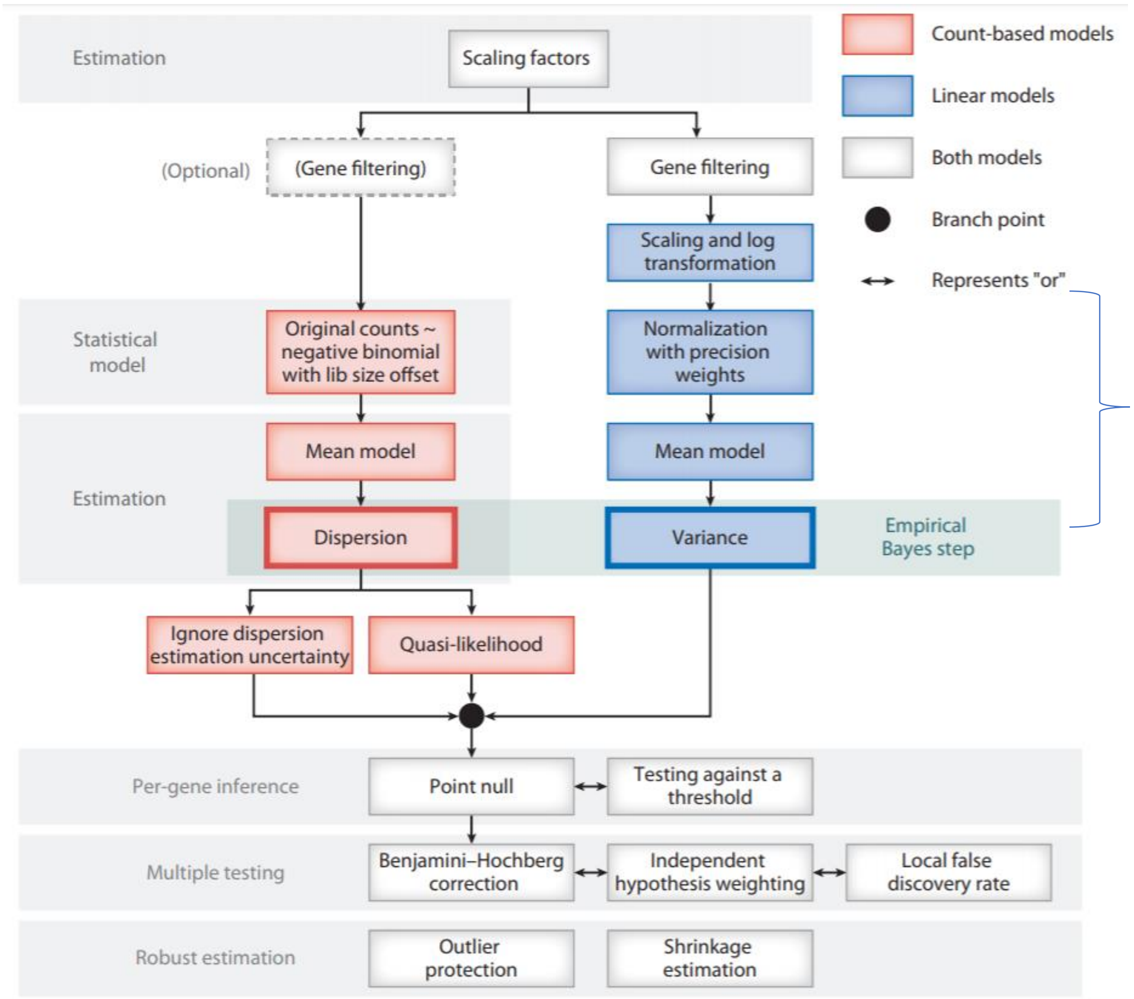


Figure 1 Dependence of the variance on the mean for condition *A* in the fly RNA-Seq data.

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

Anders and Huber, Genome Biology, 2010

Accounting for Overdispersion



DESeq
edgeR

voom

Van den Berge et al, Annual Review of Biomedical Data Science, 2019

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
  - Used to estimate gene-wise variances.
  - For each gene, the inverse of the variance is then used as weight in the ‘limma’ framework.

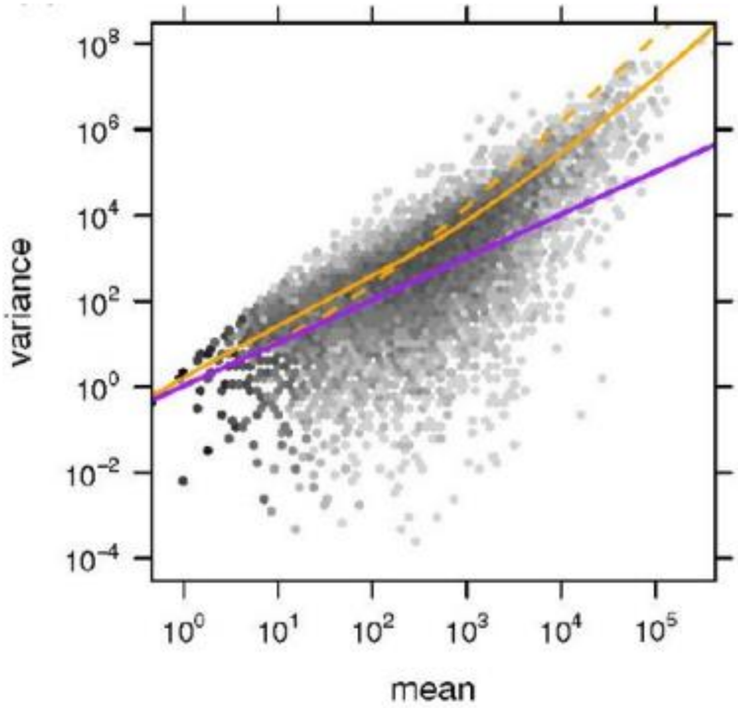
Law et al, Genome Biology, 2014

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
    - 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

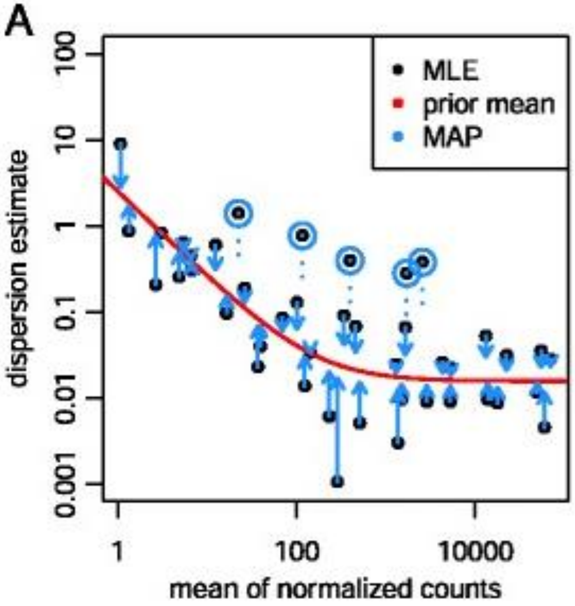
- Evidence of overdispersion on RNA-seq
- Dispersion parameter α_i in sample i



$$r\ (Y_{ij}) = \mu_{ij} + \alpha_i \mu_{ij}^2$$

DESeq2: Empirical Bayes Shrinkage for Dispersion Estimation

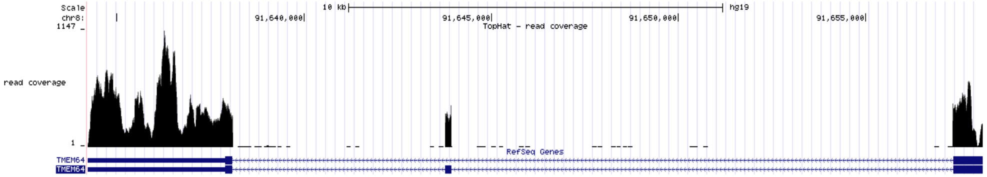
- Dispersion estimation in 3 steps:
  - Estimate gene-wise dispersion α_i^{gw}
  - Fit dispersion trend (prior)
  - Shrink towards final estimation(MAP= maximum *a posteriori*)



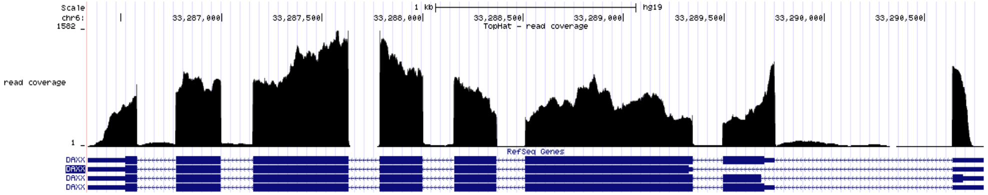
Additional challenges: data biases

- 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.

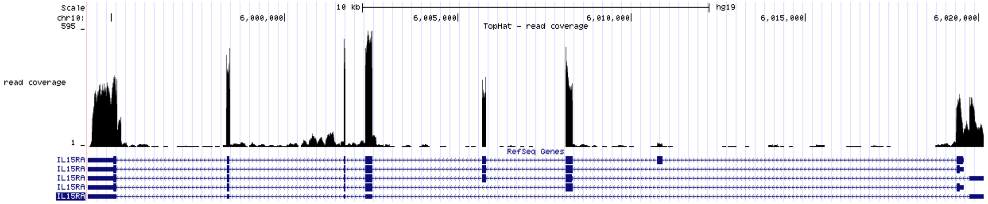
TMEM64



DAXX



IL15RA

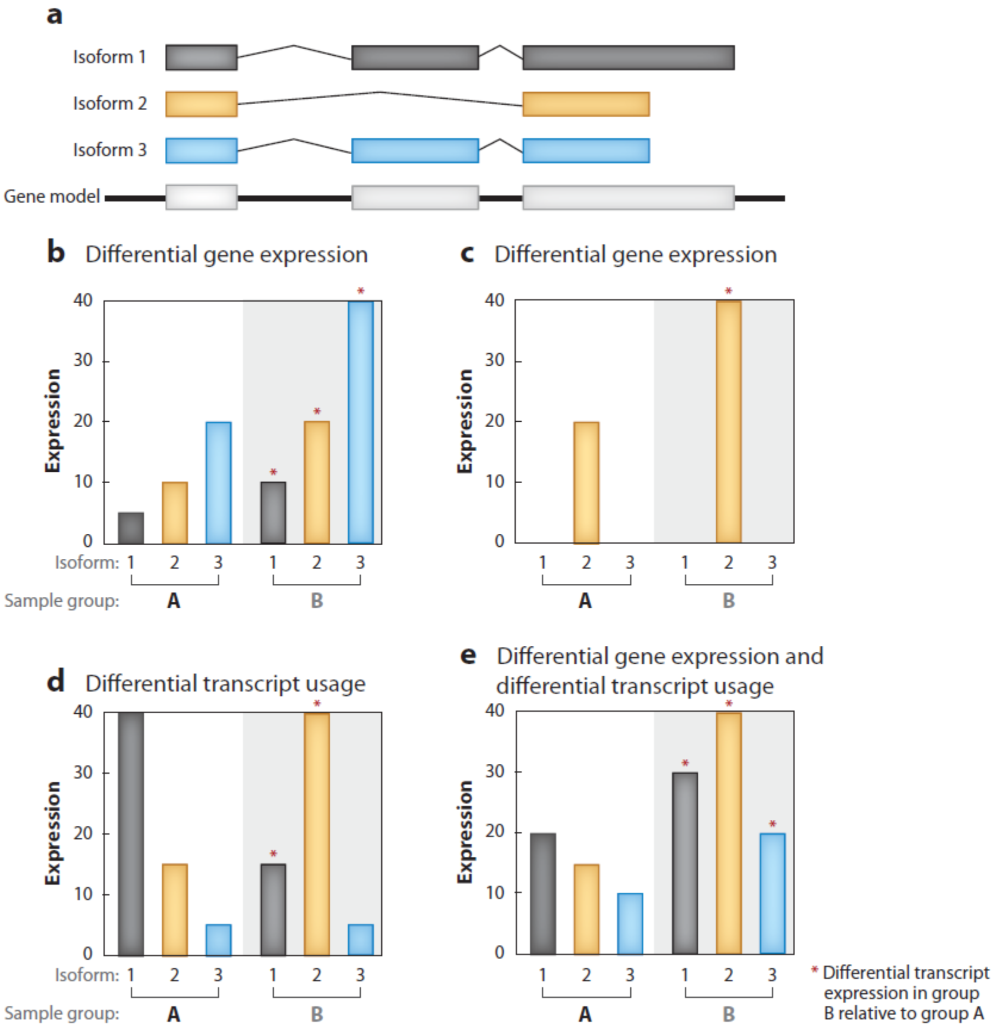


Genes selected from a human adipose study

Available Methods for Bias Correction

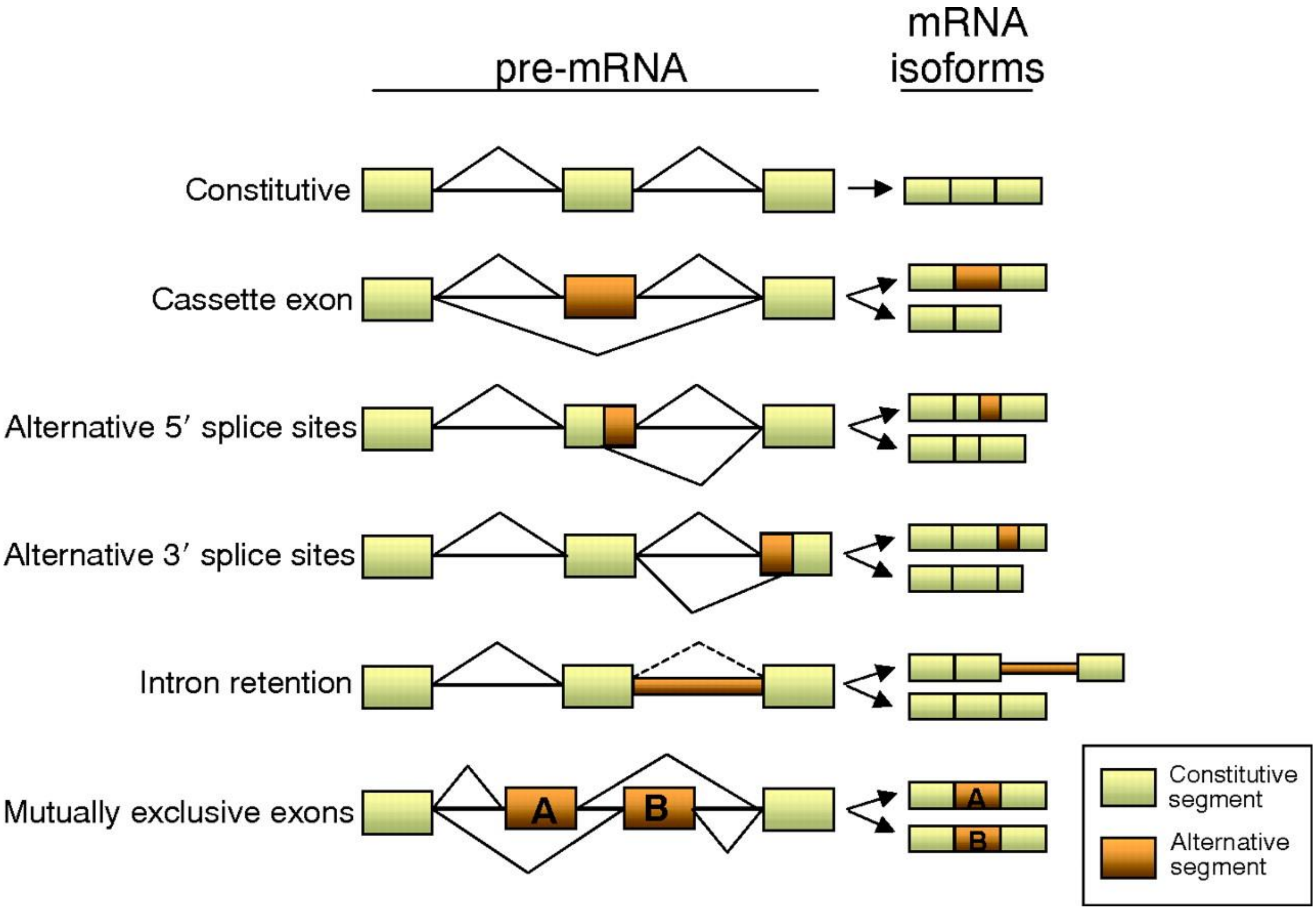
- Li et al. (2010): bias correction caused by local sequence difference
- Li & Dewey (2011): model empirical read distribution in the transcriptome; implemented in program **RSEM**
- Roberts et al. (2011): bias correction of both sequence and positional bias; implemented in program **Cufflinks**
- Nicolae 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



Chen et al, Annu Rev Biomed Data Sci, 2018

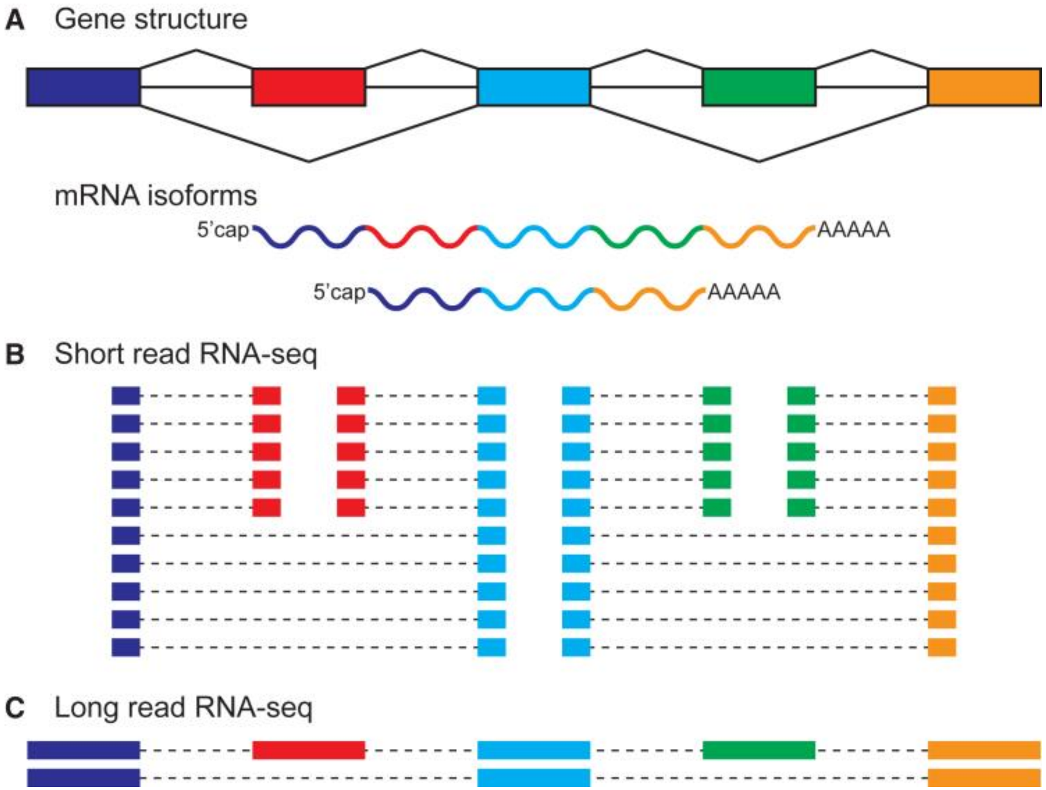
Types of Alternative Splicing



Screbrow et al. Journal of Cell Science, 2006

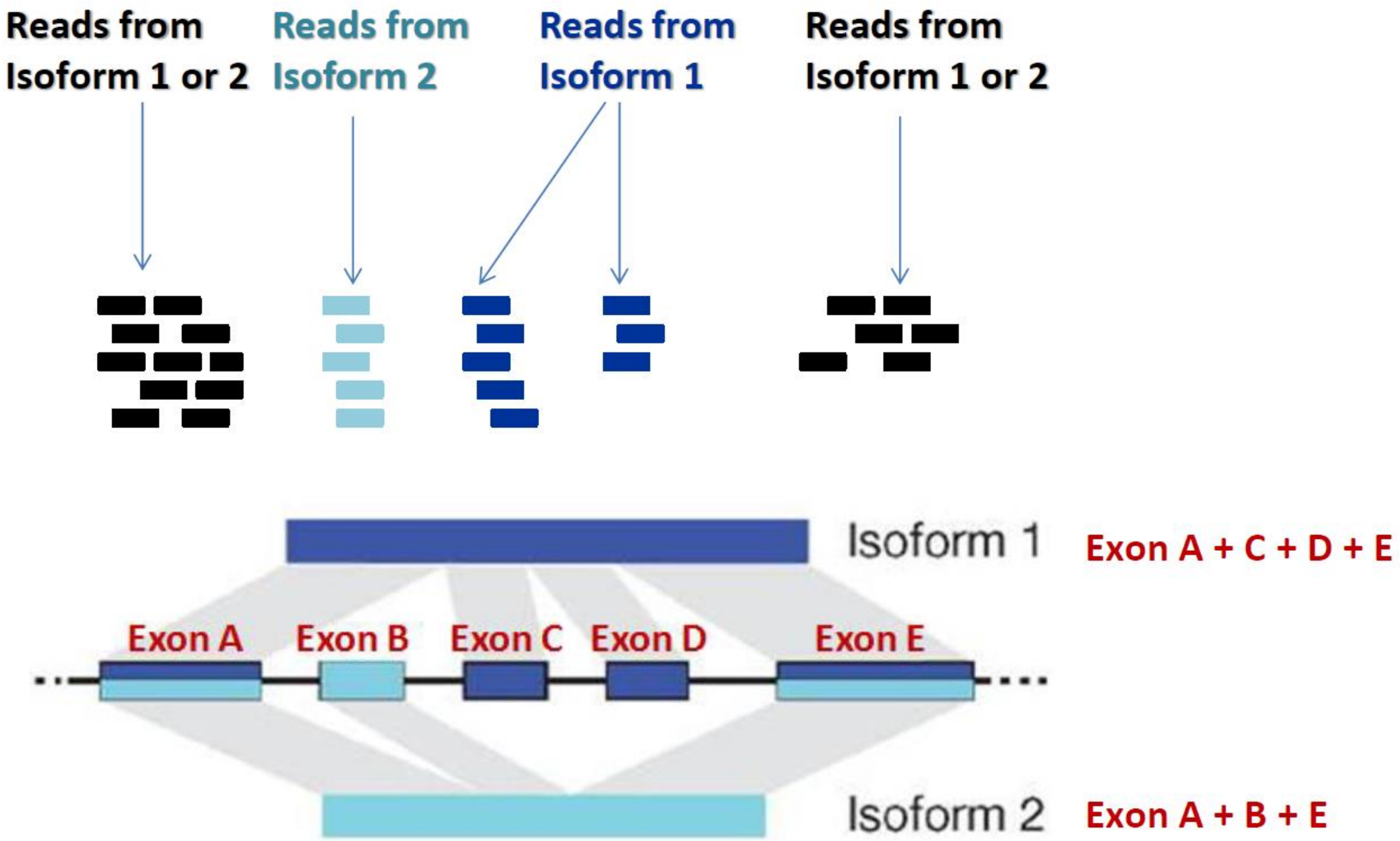
Short vs long read in RNA-Seq

- Short-read RNA-seq generates many reads, enabling the accurate quantitation of individual alternative exons, but the long-range coupling between the two alternative exons is lost.
- Long-read RNA-seq captures the long-range coupling between alternative exons and identifies the correct full-length mRNA isoforms, but the limited number of reads reduces the precision of isoform quantitation



Park et al, Am J Hum Genet, 2018

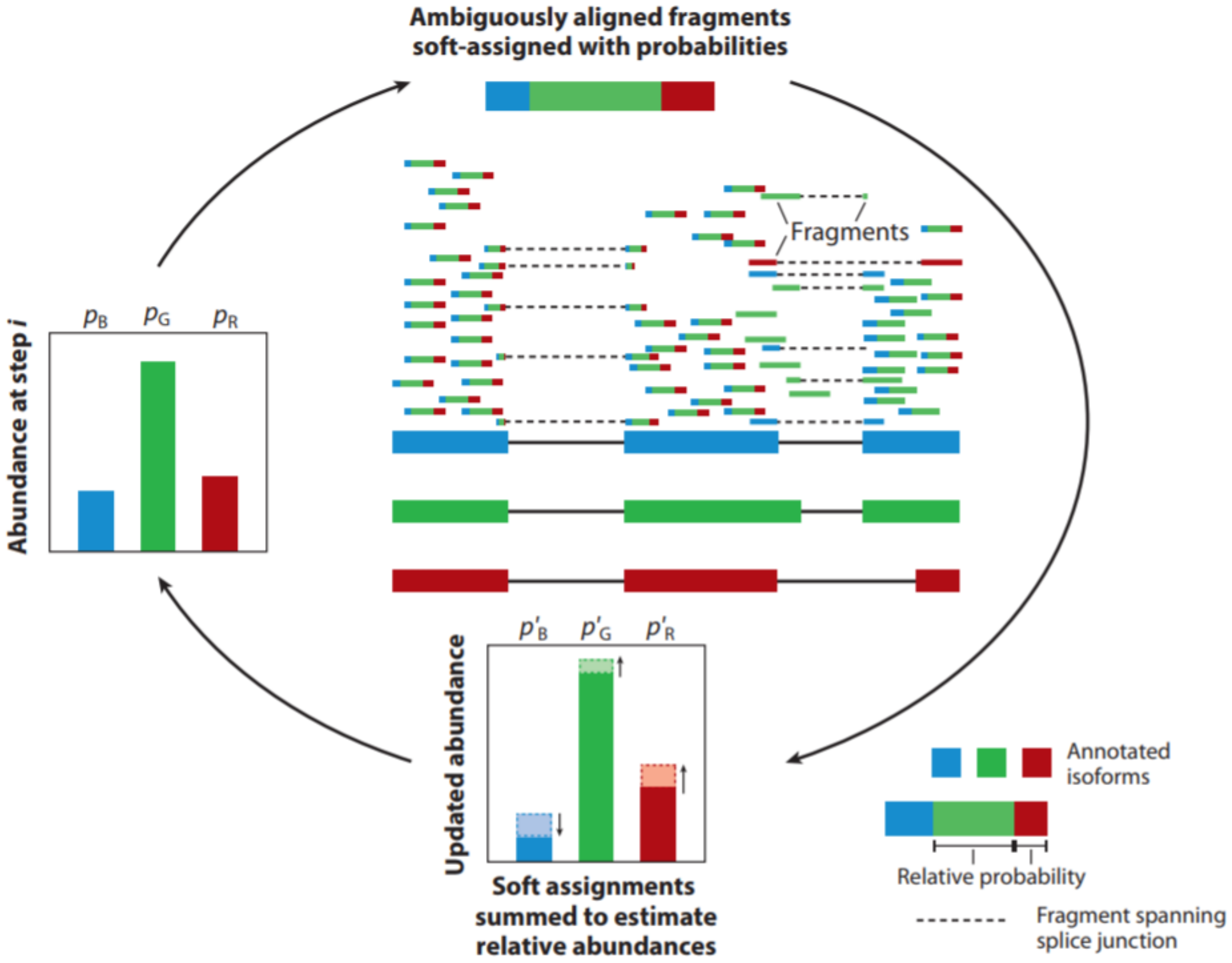
Assignment of reads to different isoforms



Isoform	Fraction
I_1	θ_1
I_2	θ_2

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

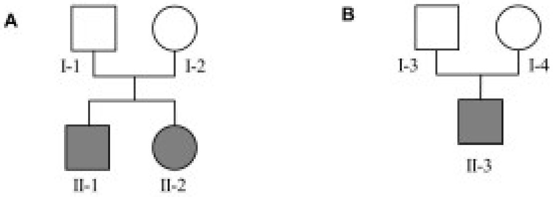
Iterative assignment of reads to different isoforms by expectation maximization in RSEM

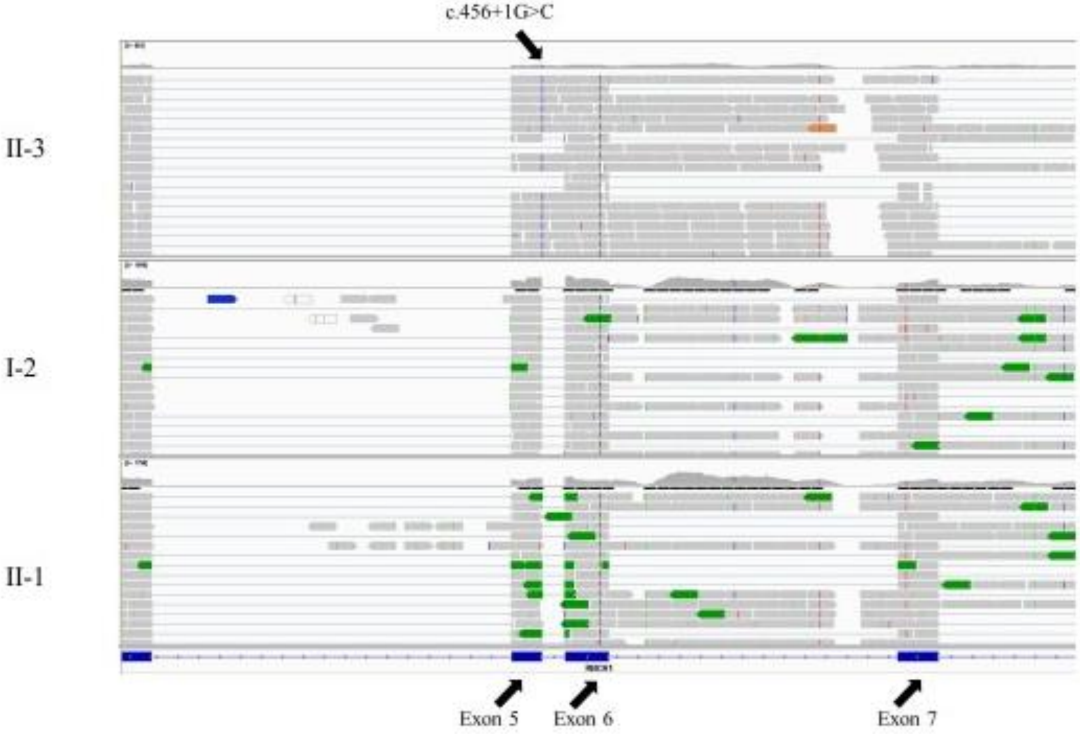

Ambiguously aligned fragments soft-assigned with probabilities
Fragments
Abundance at step i
p_B
p_G
p_R
Updated abundance
p'_B
p'_G
p'_R
Soft assignments summed to estimate relative abundances
Annotated isoforms
Relative probability
Fragment spanning splice junction

Van den Berge et al, Annual Review of Biomedical Data Science, 2019

Splicing can help find disease causal variants and new disease genes

- Whole-genome DNA/RNA sequencing identifies truncating mutations in RBCK1 in a novel Mendelian disease with neuromuscular and cardiac involvement

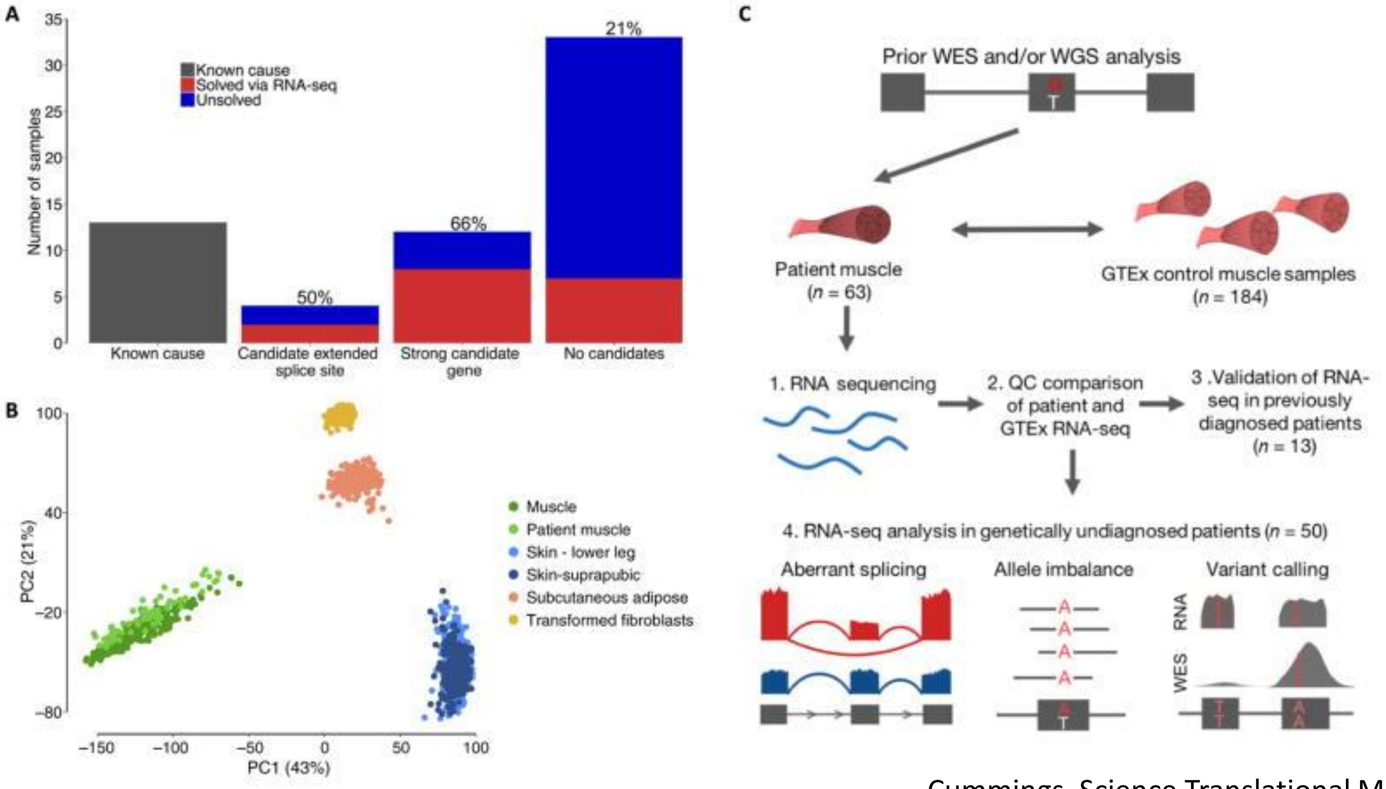




Wang, Genome Medicine, 2013

Analysis of splice variants can improve diagnostic rates from exome sequencing

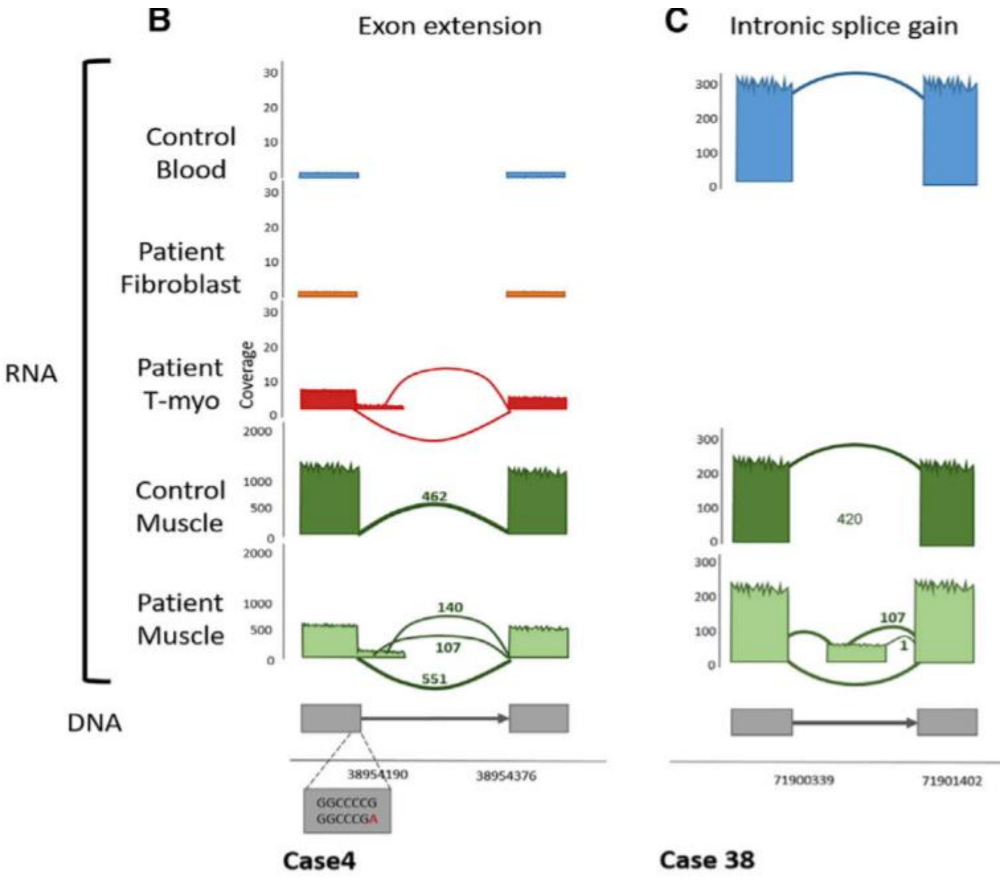
- RNA-Seq explains approximately 25% of patients clinically suggestive of having collagen VI dystrophy in whom prior genetic analysis is negative.



Cummings, Science Translational Medicine, 2017

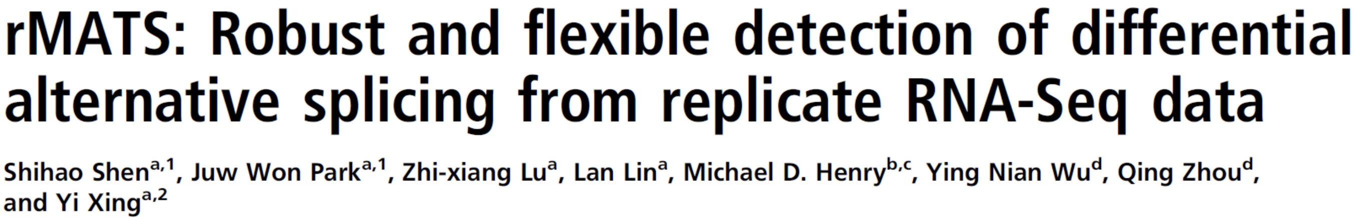
Expanding the Boundaries of RNA Sequencing as a Diagnostic Tool for Rare Mendelian Disease

- Examined a cohort of 25 exome and/or panel ‘‘negative’’ cases and provided genetic resolutior in 36% (9/25).
- Blood-based RNA-seq is no adequate for neuromuscula diagnostics, whereas myotubes generated by transdifferentiation from a individual’s fibroblasts accurately reflect the musc transcriptome



Gonorazky, Am J Hum Genet, 2019

Tools for differential splicing analysis

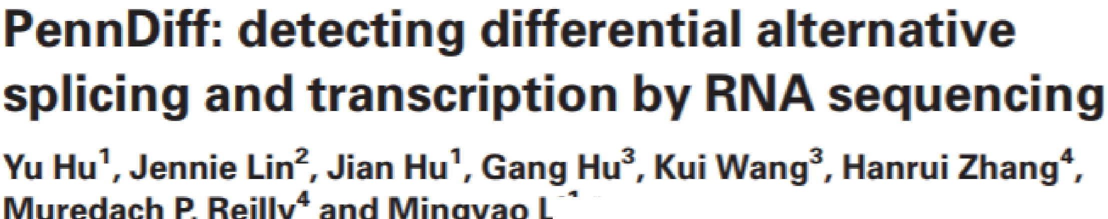


rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data

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 Lahens[4], John B Hogenesch[4], Kristen W Lynch[1,3], Yoseph Barash[1,2*]



PennDiff: detecting differential alternative splicing and transcription by RNA sequencing

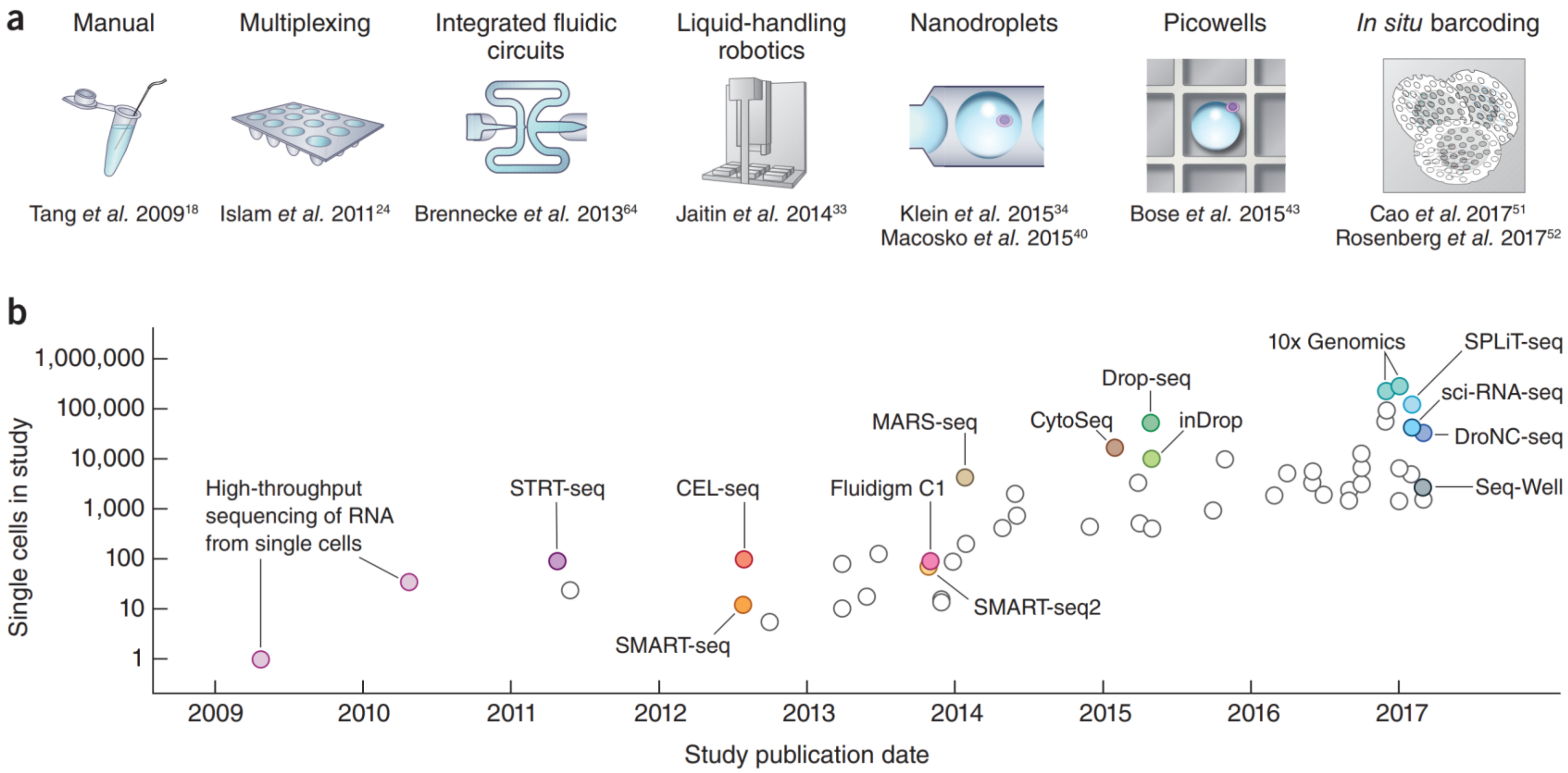
Yu Hu[1], Jennie Lin[2], Jian Hu[1], Gang Hu[3], Kui Wang[3], Hanrui Zhang[4], Muredach P. Reilly[4] and Mingyao L

Detecting differential usage of exons from RNA-seq data

Simon Anders,[1,2] Alejandro Reyes,[1] and Wolfgang Huber

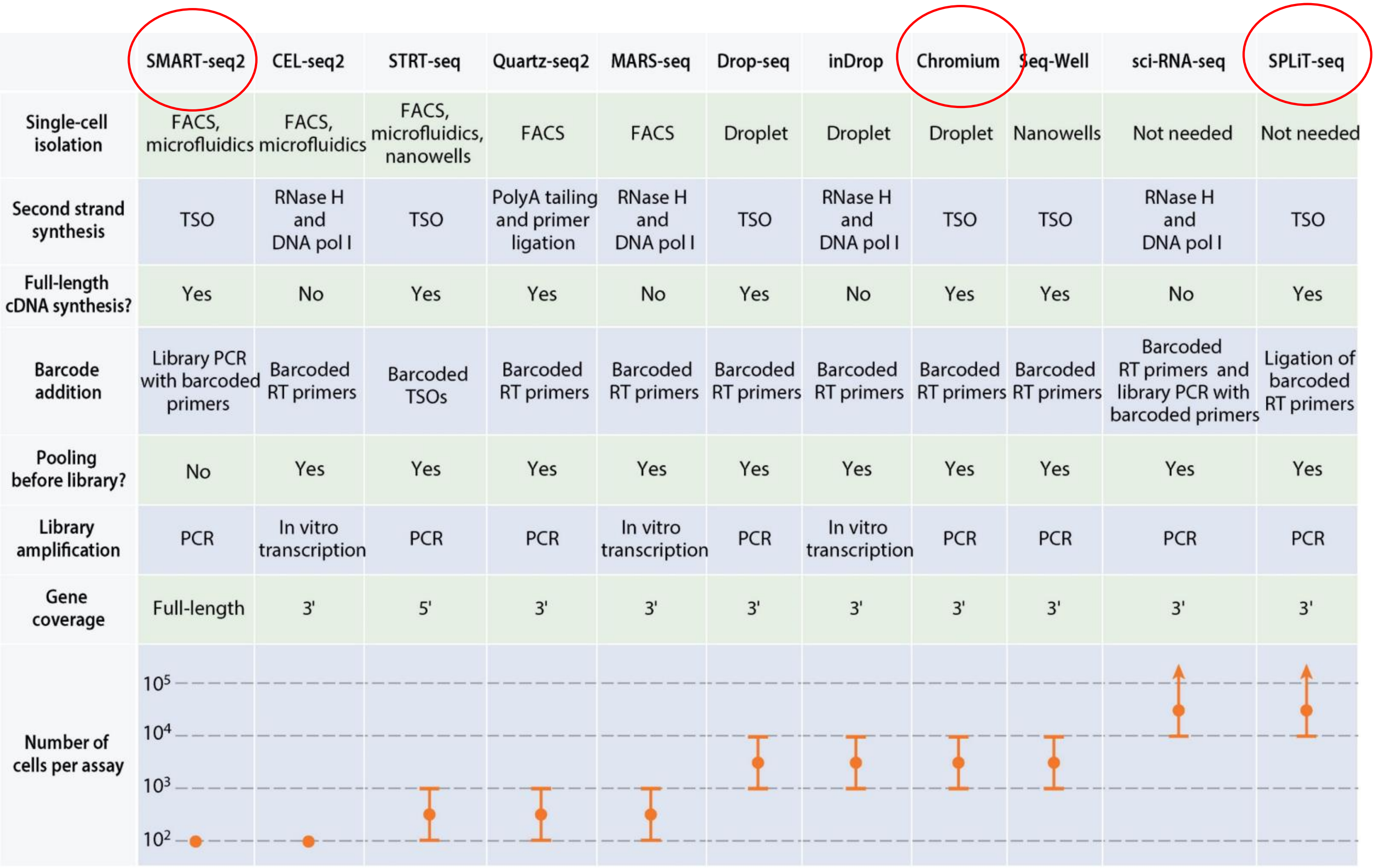
European Molecular Biology Laboratory, 69111 Heidelberg, Germany

A decade of single-cell RNA-seq



Svensson, Nature Protocols, 2018

Comparison of different scRNA-Seq technologies

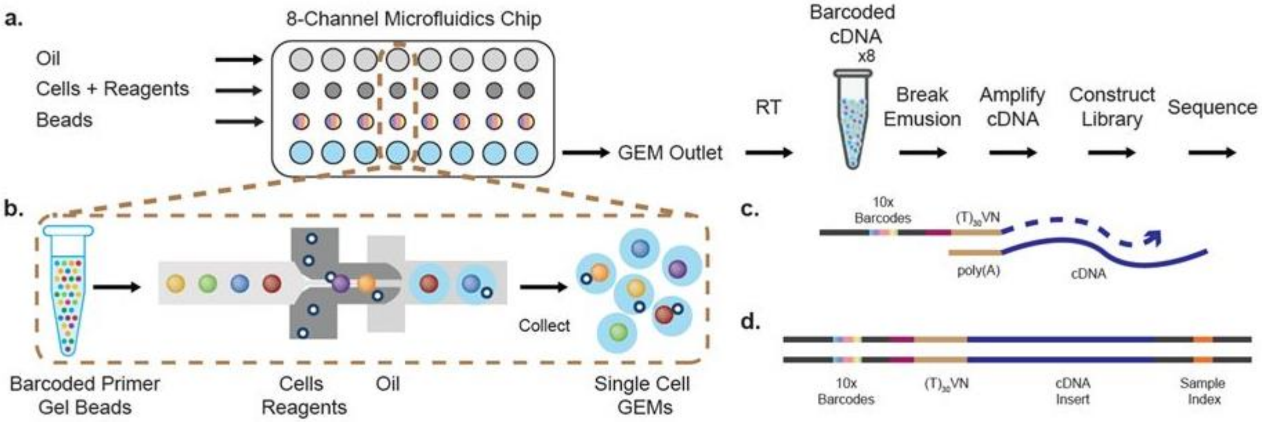


	SMART-seq2	CEL-seq2	STRT-seq	Quartz-seq2	MARS-seq	Drop-seq	inDrop	Chromium	Seq-Well	sci-RNA-seq	SPLiT-seq
Single-cell isolation	FACS, microfluidics	FACS, microfluidics	FACS, microfluidics, nanowells	FACS	FACS	Droplet	Droplet	Droplet	Nanowells	Not needed	Not needed
Second strand synthesis	TSO	RNase H 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											

Chen X, et al. 2018.
Annu. Rev. Biomed. Data Sci. 1:29–51

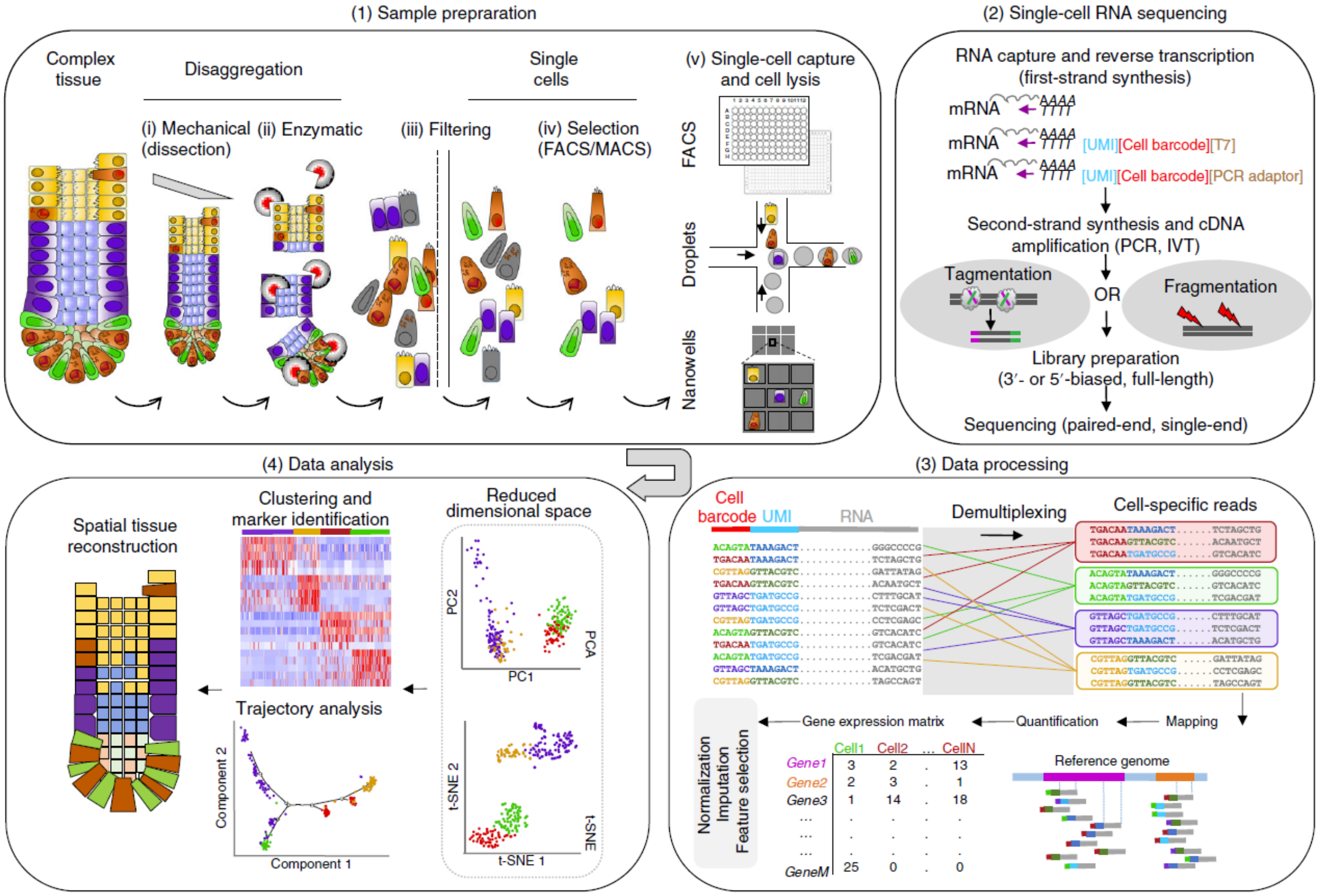
Chromium system from 10x Genomics

- 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.



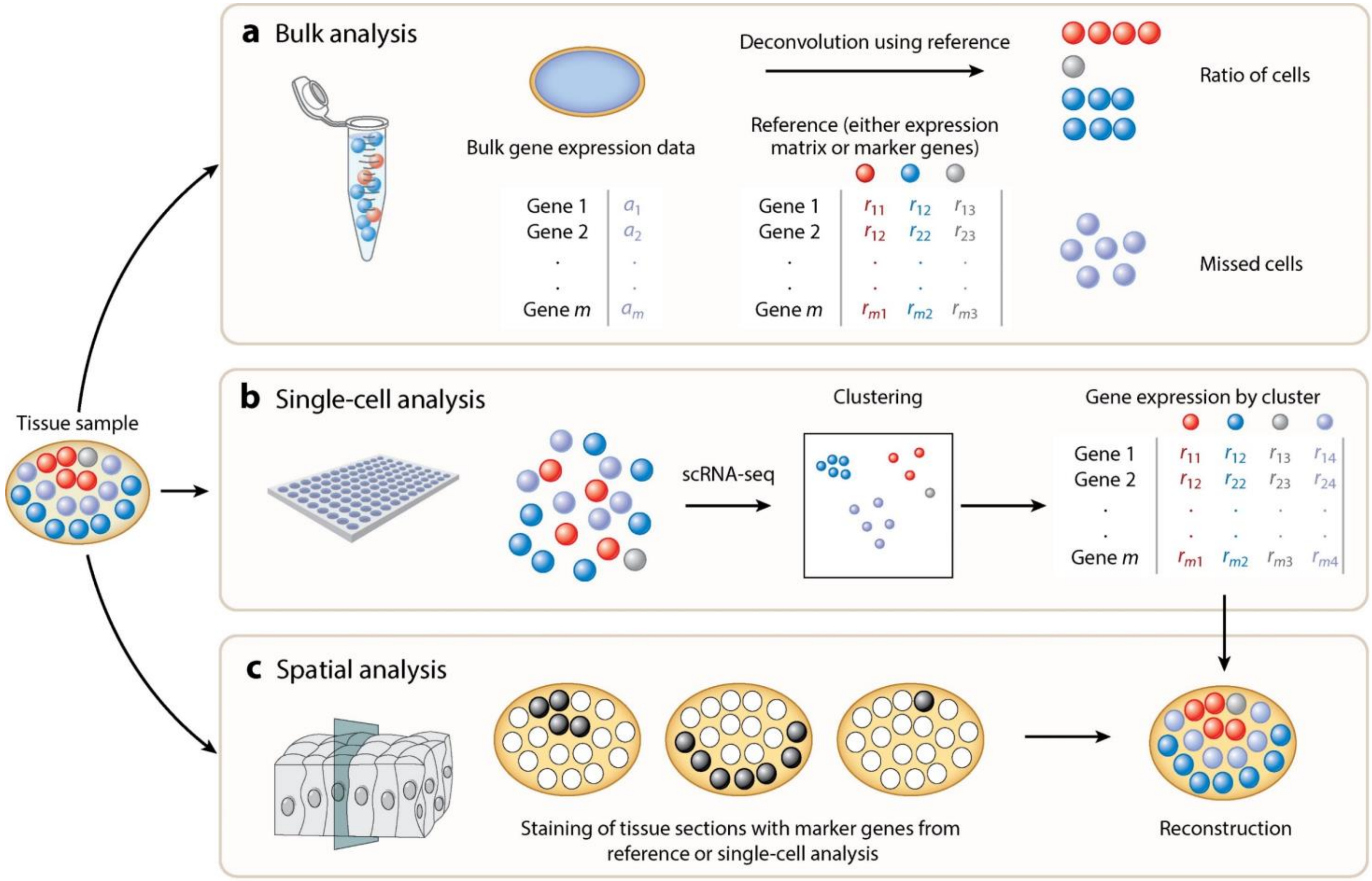
Zheng et al., Nature Communications, 2017

Single cell RNA-Seq: typical procedure


(1) Sample prepraration
Complex tissue
Disaggregation
Single cells
(v) Single-cell capture and cell lysis
(i) Mechanical (dissection)
(ii) Enzymatic
(iii) Filtering
(iv) Selection (FACS/MACS)
FACS
Droplets
Nanowells
(2) Single-cell RNA sequencing
RNA capture and reverse transcription (first-strand synthesis)
mRNA AAAA TTTT
mRNA AAAA TTTT [UMI][Cell barcode][T7]
mRNA AAAA TTTT [UMI][Cell barcode][PCR adaptor]
Second-strand synthesis and cDNA amplification (PCR, IVT)
Tagmentation
OR
Fragmentation
Library preparation (3′- or 5′-biased, full-length)
Sequencing (paired-end, single-end)
(3) Data processing
Cell barcode UMI
RNA
ACAGTATAAAGACT............GGGCCCCG
TGACAATAAAGACT............TCTAGCTG
CGTTAGGTTACGTC............GATTATAG
TGACAAGTTACGTC............ACAATGCT
GTTAGCTGATGCCG............CTTTGCAT
GTTAGCTGATGCCG............TCTCGACT
CGTTAGTGATGCCG............CCTCGAGC
ACAGTAGTTACGTC............GTCACATC
TGACAATGATGCCG............GTCACATC
ACAGTATGATGCCG............TCGACGAT
GTTAGCTAAAGACT............ACATGCTG
CGTTAGGTTACGTC............TAGCCAGT
Demultiplexing
Cell-specific reads
TGACAATAAAGACT......TCTAGCTG
TGACAAGTTACGTC......ACAATGCT
TGACAATGATGCCG......GTCACATC
ACAGTATAAAGACT......GGGCCCCG
ACAGTAGTTACGTC......GTCACATC
ACAGTATGATGCCG......TCGACGAT
GTTAGCTGATGCCG......CTTTGCAT
GTTAGCTGATGCCG......TCTCGACT
GTTAGCTAAAGACT......ACATGCTG
CGTTAGGTTACGTC ......GATTATAG
CGTTAGTGATGCCG......CCTCGAGC
CGTTAGGTTACGTC......TAGCCAGT
Mapping
Quantification
Gene expression matrix
Reference genome
Cell1 Cell2 ... CellN
Gene1 3 2 . 13
Gene2 2 3 . 1
Gene3 1 14 . 18
...
GeneM 25 0 . 0
Normalization
Imputation
Feature selection
(4) Data analysis
Reduced dimensional space
PC2
PC1
PCA
t-SNE 2
t-SNE 1
t-SNE
Clustering and marker identification
Trajectory analysis
Component 2
Component 1
Spatial tissue reconstruction

Lafzi et al., Nature Protocols, 2018

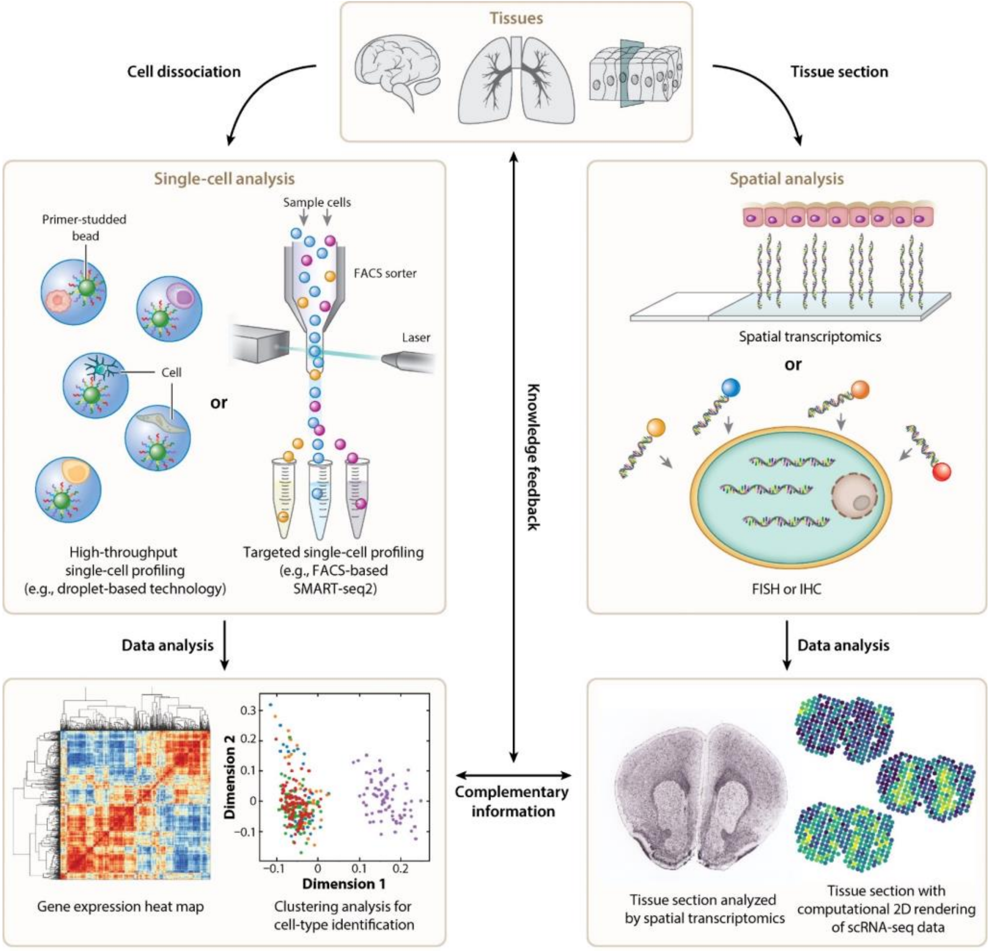
Extension to spatial analysis



Chen X, et al. 2018.
Annu. Rev. Biomed. Data Sci. 1:29–51

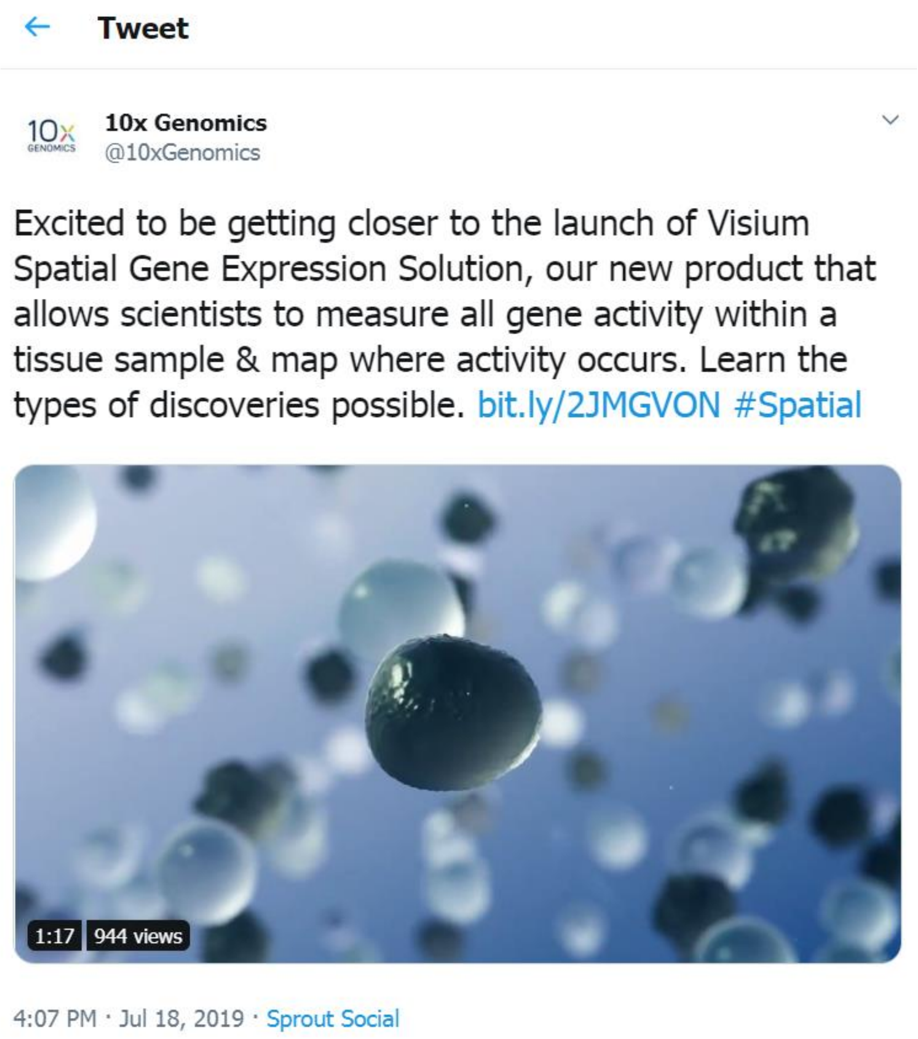
Chen et al, Annu Rev Biomed Data Sci, 2018

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


Tissues
Cell dissociation
Tissue section
Single-cell analysis
Sample cells
Primer-studded bead
FACS sorter
Laser
Cell
or
High-throughput single-cell profiling (e.g., droplet-based technology)
Targeted single-cell profiling (e.g., FACS-based SMART-seq2)
Spatial analysis
Spatial transcriptomics
or
FISH or IHC
Knowledge feedback
Data analysis
Data analysis
Gene expression heat map
0.3
0.2
0.1
0
−0.1
Dimension 2
−0.1
0
0.1
0.2
Dimension 1
Clustering analysis for cell-type identification
Complementary information
Tissue section analyzed by spatial transcriptomics
Tissue section with computational 2D rendering of scRNA-seq data

Chen X, et al. 2018.
Annu. Rev. Biomed. Data Sci. 1:29–51

Commercial solutions will be available soon



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

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