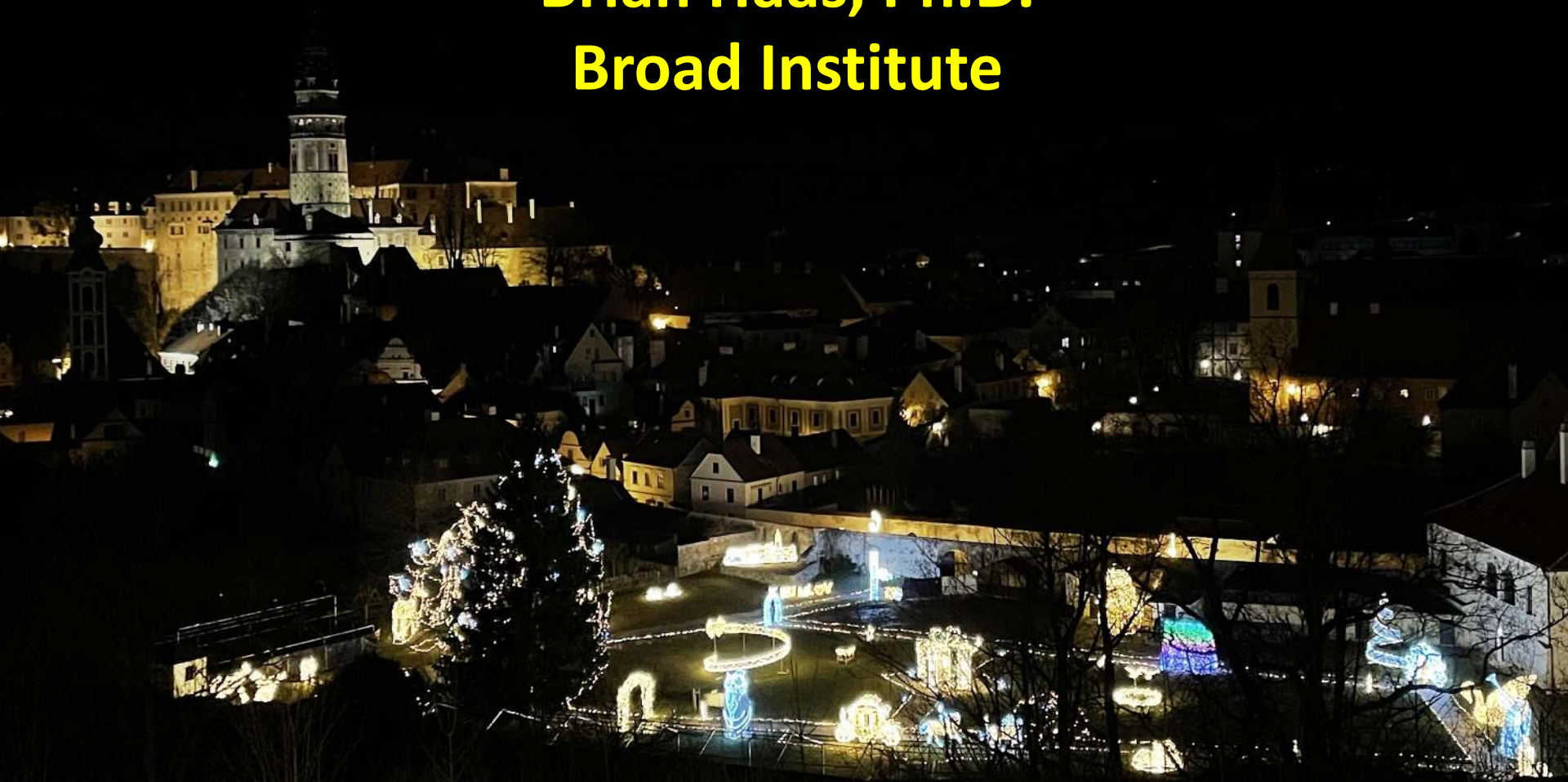


Transcriptomics

Brian Haas, Ph.D.
Broad Institute



Workshop on Genomics, Cesky Krumlov, January 2025

Intro to Brian Haas



Education and Career History



BS,MS Molecular Bio
DNA Repair
SUNY Albany, New York
1991-1999



The Institute for Genomic Research
Rockville, Maryland, USA
(1999-2007)
Bioinformatics Analyst & Engineer
MS. Computer Science / Johns Hopkins

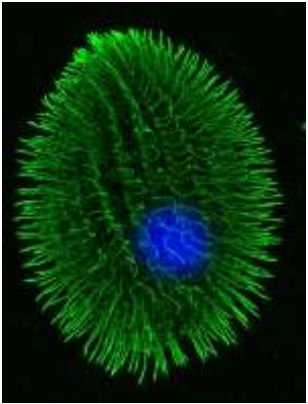
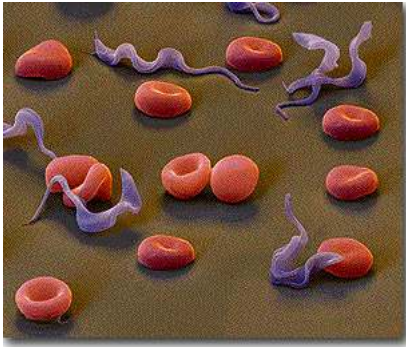
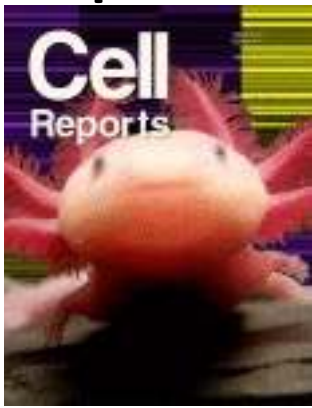
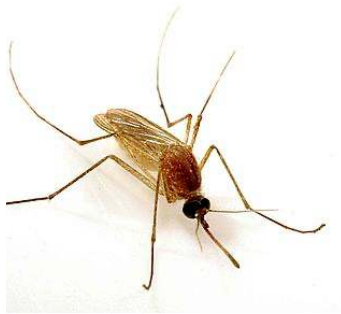


Cambridge, Massachusetts, USA

2007-current

Computational Biologist / Manager /
Principal Computational Scientist
Ph.D. Bioinformatics / Boston University

Annotation and Analysis for Diverse Genomes and Transcriptomes



My Favorite Activity – Bioinformatics Tool

Development and Applications



NAR, 2003



Bioinformatics, 2004



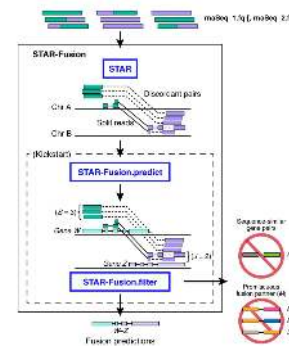
EvidenceModeler
Genome Biology, 2008



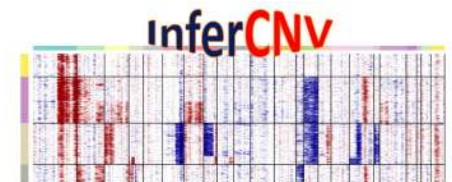
Chimera Slayer
Genome Research, 2011



Nature Biotech, 2011
Nature Protocols, 2013



STAR-Fusion
Genome Biology, 2019



FusionInspector
Cell Reports Methods, 2023

My Favorite Activity – Bioinformatics Tool

Development and Application



NAR, 2003



Bioinformatics, 2004



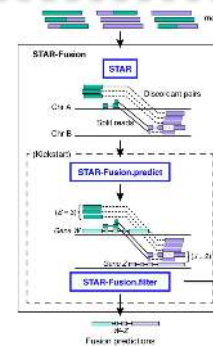
EvidenceModeler
Genome Biology, 2008



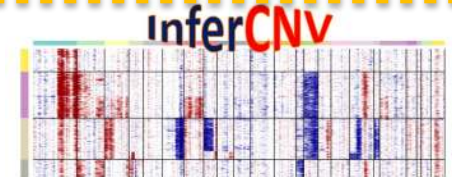
Chimera Slayer
Genome Research, 2011



Nature Biotech, 2011
Nature Protocols, 2013

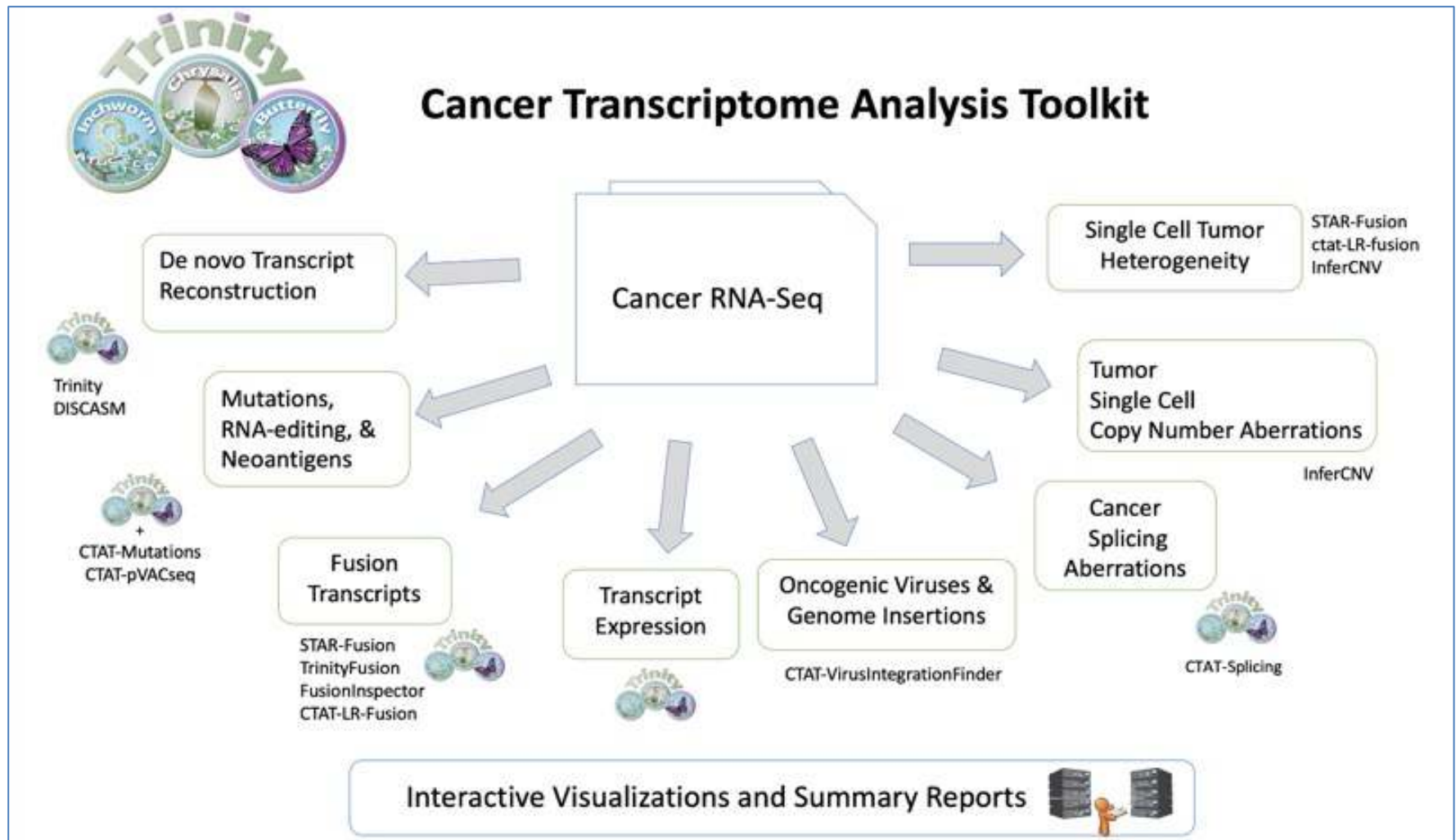


STAR-Fusion
Genome Biology, 2019



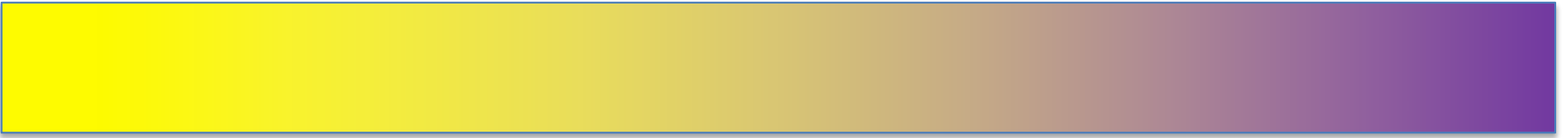
FusionInspector
Cell Reports Methods, 2023

Latest developments focused on cancer transcriptomics:



Overview of Trinity CTAT. Given cancer RNA-seq as input, Trinity CTAT provides modules for exploring characteristics of the cancer transcriptome (and cancer genome) including both genome-guided and genome-free analyses, targeting bulk or single-cell transcriptomes. Interactive visualizations and reports are provided to facilitate downstream analysis and for clinical review.

Transcriptomics Lecture Outline



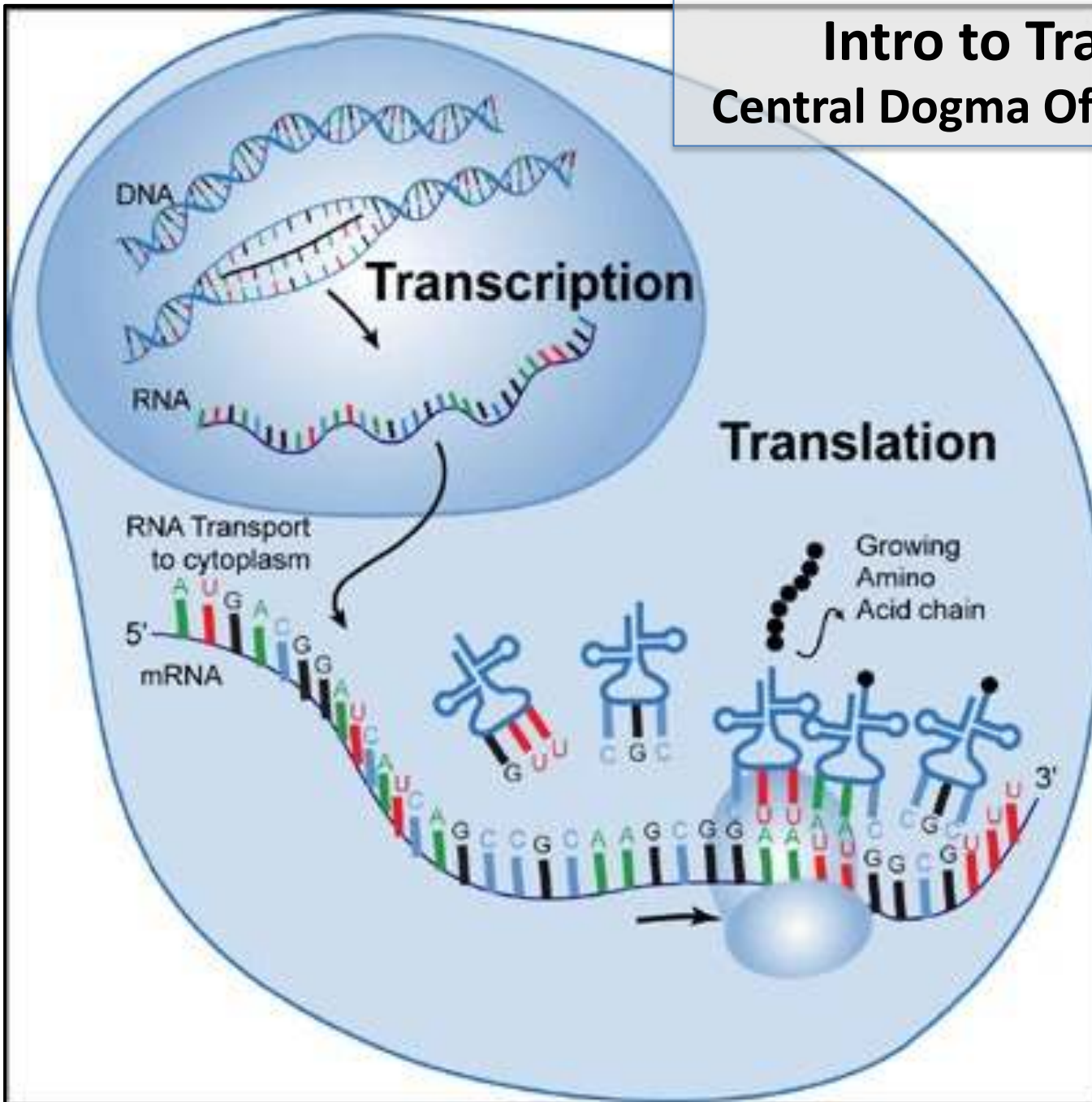
1. Intro to transcriptomics
2. Transcript reconstruction methods
3. Genome-free transcriptomics (eg. for non-model orgs)
4. Expression quantification
5. Differential expression (brief – more details in Rachel's workshop!)
6. Latest advancements in long read isoform sequencing
7. Overview of single cell transcriptomics
8. Overview of spatial transcriptomics
9. Applications in Cancer Transcriptomics



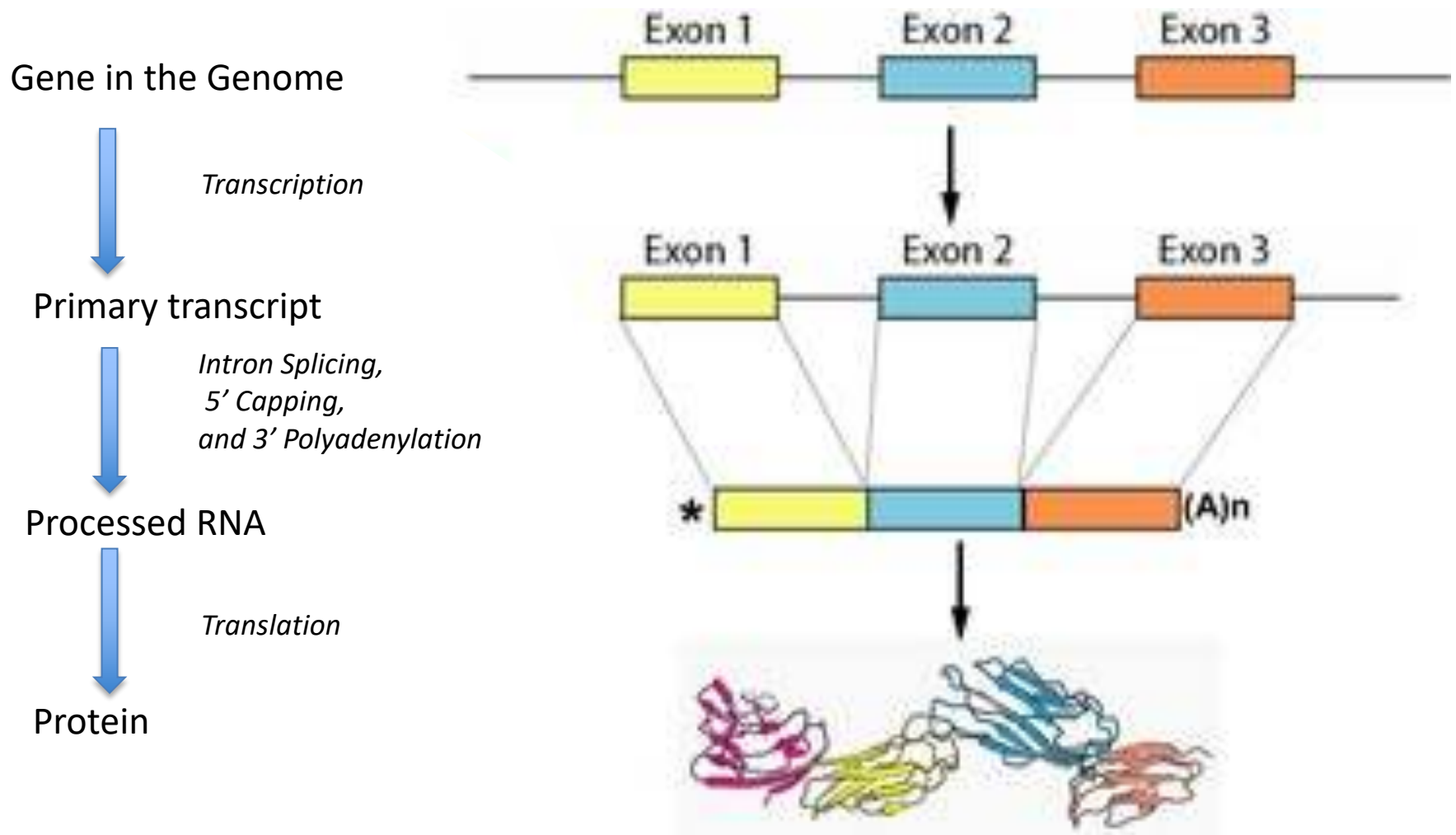
Part 1. Intro to Transcriptomics

Intro to Transcriptomics

Central Dogma Of Molecular Biology

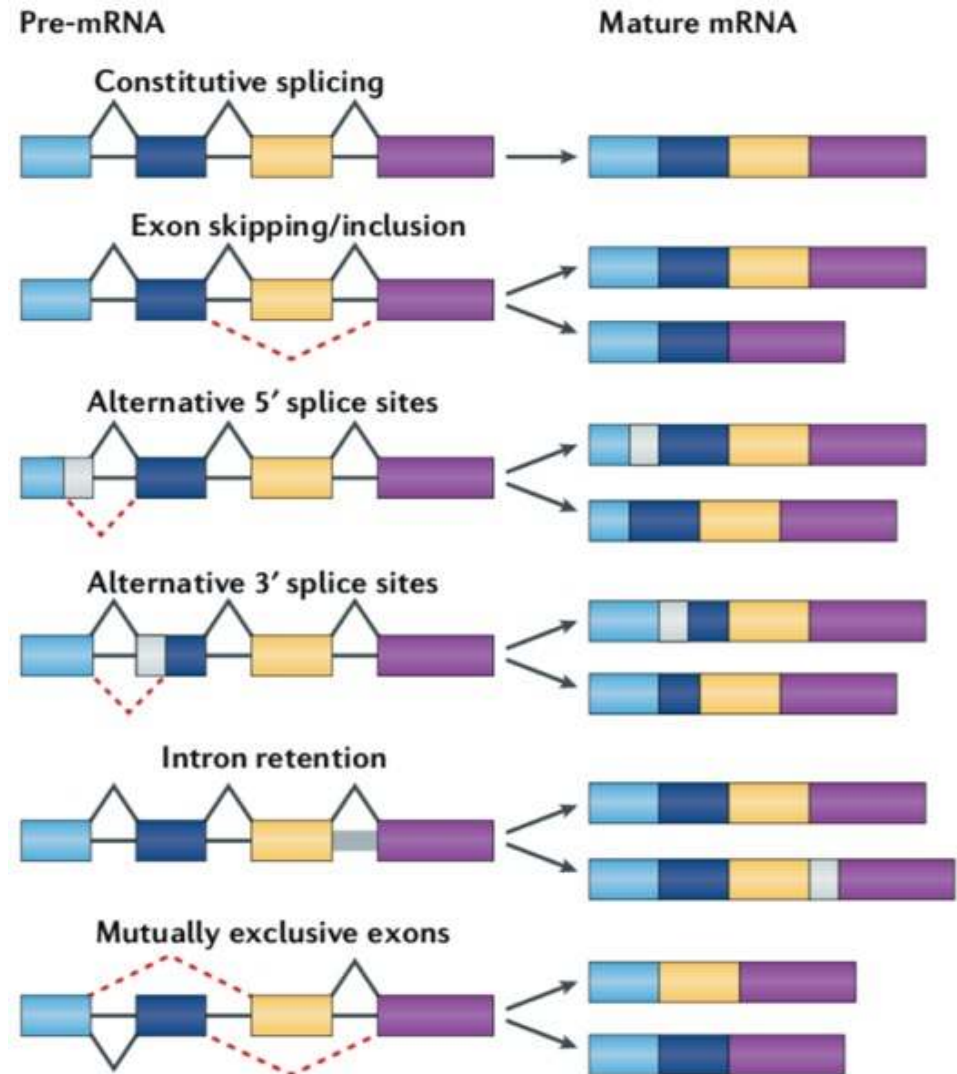


Primary mRNA molecules Often Undergo Splicing in Eukaryotes



Alternative Splicing – Multiple Products from Single Genes

- Core regulatory process – diversifies the function of genes.
- Generates mRNAs that differ in coding sequence and UTRs. Effects:
 - Protein isoforms
 - Translation efficiency
 - Stability
 - Localization
 - Reading frame changes
- Estimated 90-95% of human genes undergo alternative splicing



Think of genes as protosentences

Gene: A catalytically $\frac{\text{active}}{\text{inactive}}$ kinase $\frac{\text{with}}{\text{without}}$ a NLS

Think of genes as protosentences

Gene: A catalytically $\frac{\text{active}}{\text{inactive}}$ kinase $\frac{\text{with}}{\text{without}}$ a NLS

Alternative splicing

A catalytically $\frac{\text{active}}{\text{inactive}}$ kinase $\frac{\text{with}}{\text{without}}$ a NLS

A catalytically $\frac{\text{active}}{\text{inactive}}$ kinase $\frac{\text{with}}{\text{without}}$ a NLS

A catalytically $\frac{\text{active}}{\text{inactive}}$ kinase $\frac{\text{with}}{\text{without}}$ a NLS

A catalytically $\frac{\text{active}}{\text{inactive}}$ kinase $\frac{\text{with}}{\text{without}}$ a NLS

Fully formed sentences $\approx$ mature mRNA

Gene: A catalytically $\frac{\text{active}}{\text{inactive}}$ kinase $\frac{\text{with}}{\text{without}}$ a NLS

Alternative splicing

A catalytically $\frac{\text{active}}{\text{inactive}}$ kinase $\frac{\text{with}}{\text{without}}$ a NLS

A catalytically $\frac{\text{active}}{\text{inactive}}$ kinase $\frac{\text{with}}{\text{without}}$ a NLS

A catalytically $\frac{\text{active}}{\text{inactive}}$ kinase $\frac{\text{with}}{\text{without}}$ a NLS

A catalytically $\frac{\text{active}}{\text{inactive}}$ kinase $\frac{\text{with}}{\text{without}}$ a NLS

Transcripts

A catalytically active kinase with a NLS

A catalytically active kinase without a NLS

A catalytically inactive kinase with a NLS

A catalytically inactive kinase without a NLS

RNA isoform sequencing provides structural insight

Gene: A catalytically $\frac{\text{active}}{\text{inactive}}$ kinase $\frac{\text{with}}{\text{without}}$ a NLS

Alternative splicing

A catalytically $\frac{\text{active}}{\text{inactive}}$ kinase $\frac{\text{with}}{\text{without}}$ a NLS

A catalytically $\frac{\text{active}}{\text{inactive}}$ kinase $\frac{\text{with}}{\text{without}}$ a NLS

A catalytically $\frac{\text{active}}{\text{inactive}}$ kinase $\frac{\text{with}}{\text{without}}$ a NLS

A catalytically $\frac{\text{active}}{\text{inactive}}$ kinase $\frac{\text{with}}{\text{without}}$ a NLS

Transcripts

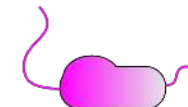
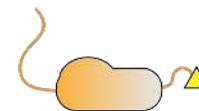
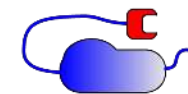
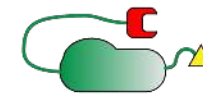
A catalytically active kinase with a NLS

A catalytically active kinase without a NLS

A catalytically inactive kinase with a NLS

A catalytically inactive kinase without a NLS

Proteins



Cellular function

kinase
nuclear targets

kinase
cytoplasmic targets

competitive inhibitor
nuclear targets

competitive inhibitor
cytoplasmic targets

Biological Investigations Empowered by Transcriptomics

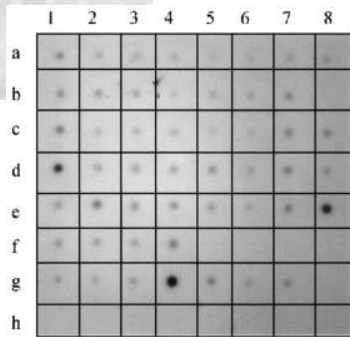
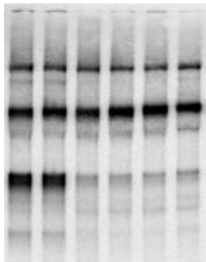


Extract RNA,
... some protocol for processing, ...

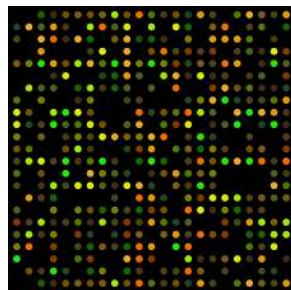


Analysis Method
(pick your favorite)

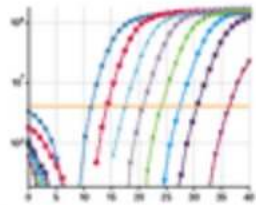
Northern



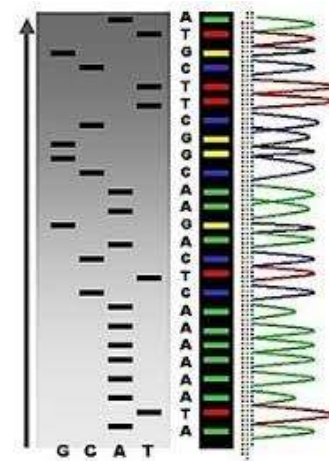
Dot Blot



Microarray



qRT-PCR



Sanger Sequencing



Other...



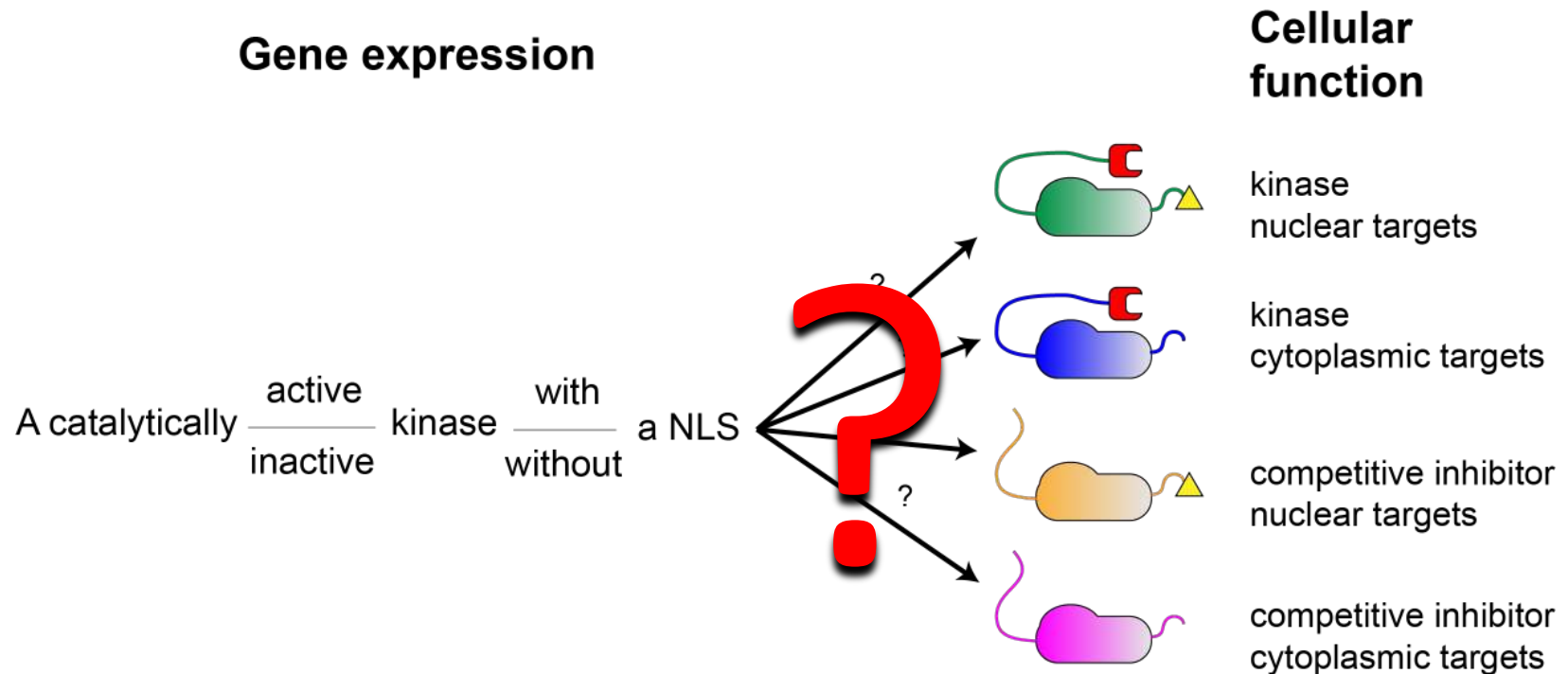
Minion



MinION Mk1: portable, real time biological analyses

MinION

Gene expression analyses ignore isoform variation



Need to resolve isoforms for deeper insights into cellular functions

Historical Timeline to Modern Transcriptomics (from 1970)

Reverse Transcription (1970)

Northern Blot
Sanger Sequencing
(1977)

Expressed Sequence Tags (1992)

cDNA microarrays (1995)

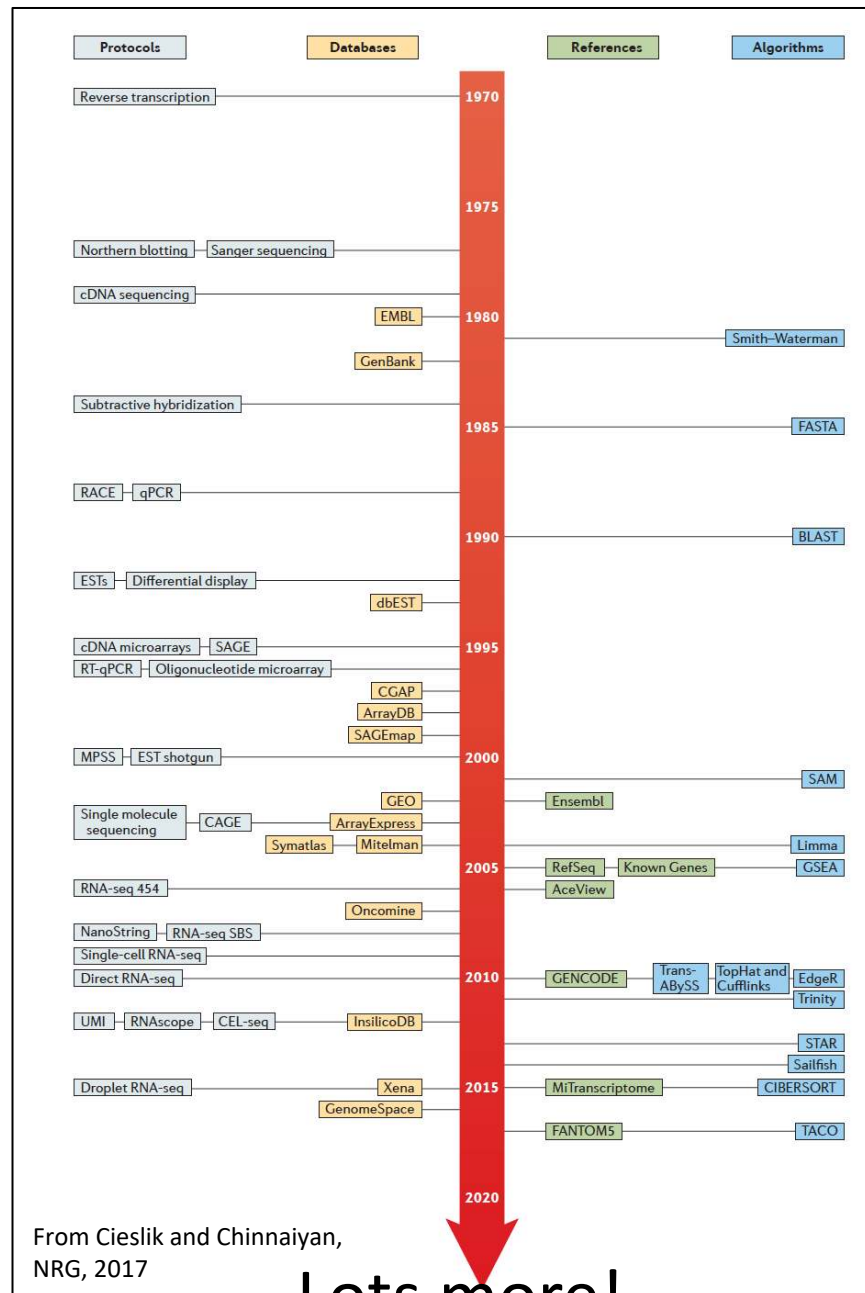
RNA-Seq (2006-2008)

PacBio IsoSeq (2014)

Droplet single cell RNA-Seq (2015)

Direct RNA Seq Nanopore (2018)

SlideSeq-v2 (2021)



Note: Just a small sampling of what's available.

Smith Waterman (1981)

BLAST (1990)

SAMtools (2009)

Tophat/Cufflinks (2010)



STAR (2013)

StringTie (2015)

Kallisto (2016)

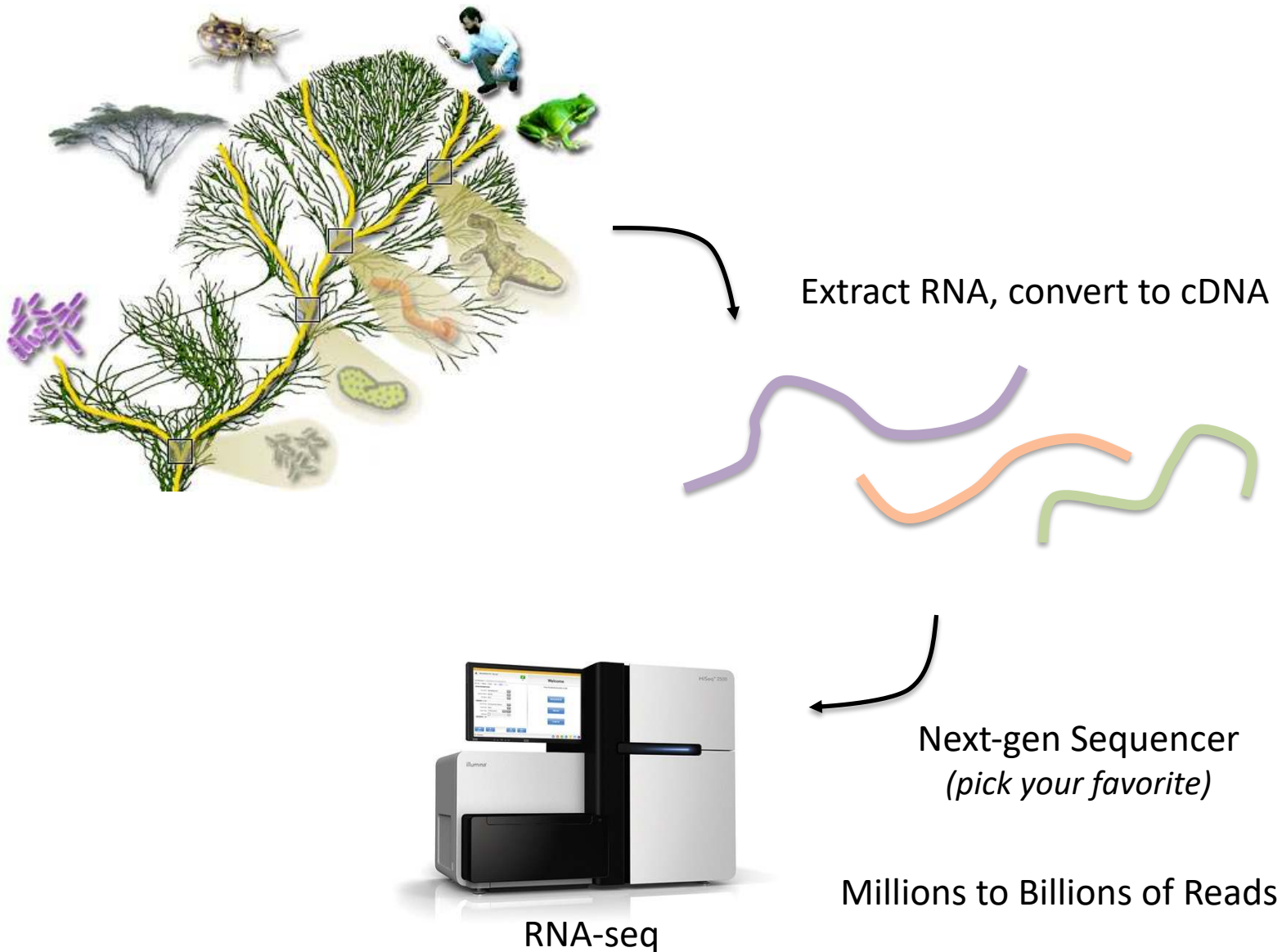
Salmon (2017)

minimap2 (2018)

Seurat-v2 (2021)

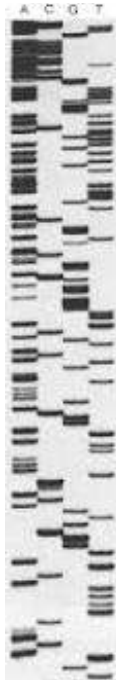
Lots more!

Modern Transcriptome Studies Empowered by RNA-seq



Personal Reflections...

Circa 1995



Generating RNA-Seq: *How to Choose?*

Platform	iSeq Project Firefly 2018	MiniSeq	MiniSeq	Next Seq 550	HiSeq 2500 RR	HiSeq 2500 V3	HiSeq 2500 V4	HiSeq 4000	HiSeq X	Nova Seq S1 2018	Nova Seq S2	Nova Seq S4	5500 XL	318 HiQ 520	Ion 530	Ion Proton P1	PGM HiQ 540	RS P6-C4	Sequel	R&D end 2018	Smidg ION RnD	Mini ION R9.5	Grid ION X5	Prome thION RnD	Prome thION theor etical	QiaGen Gene Reader	BGI SEQ 500	BGI SEQ 50	#
Reads: (M)	4	25	25	400	600	3000	4000	5000	6000	3300	6600	20000	1400	3-5	15-20	165	60-80	5.5	38.5	--	--	--	--	--	--	400	1600	1600	--
Read length: (paired-end*)	150*	150*	300*	150*	100*	100*	125*	150*	150*	150*	150*	150*	60	200 400	200 400	200	200	15K	12K	32K	--	--	--	--	--	--	100*	50	--
Run time: (d)	0.54	1	2	1.2	1.125	11	6	3.5	3	1.66	1.66	1.66	7	0.37	0.16	--	0.16	4.3	--	--	--	2	2	2	--	--	1	0.4	--
Yield: (Gb)	1	7.5	15	120	120	600	1000	1500	1800	1000	2000	6000	180	1.5	7	10	12	12	5	150	4	8	40	2400	11000	80	200	8	--
Rate: (Gb/d)	1.85	7.5	7.5	100	106.6	55	166	400	600	600	1200	3600	30	5.5	50	--	93.75	2.8	--	--	--	4	20	1200	5500	--	200	20	--
Reagents: (\$K)	0.1	1.75	1	5	6.145	23.47	29.9	--	--	--	--	--	10.5	0.6	--	1	1.2	2.4	--	1	--	0.5	1.5	--	--	0.5	--	--	--
per-Gb: (\$)	100	233	66	50	51.2	39.1	31.7	20.5	7.08	18	15	5.8	58.33	--	--	100	--	200	80	6.6	--	62.5	37.5	20	4.3	--	--	--	--
hg-30x: (\$)	12000	28000	8000	5000	6144	4692	3804	2460	849.6	1800	1564	700	7000	--	--	12000	--	24000	9600	1000	--	7500	4500	2400	500	--	600	--	--
Machine: (\$)	30K	49.5K	99K	250K	740K	690K	690K	900K	1M	999K	999K	999K	595K	50K	65K	243K	242K	695K	350K	350K	--	--	125K	75K	75K	--	200K	--	--

#Page maintained by <http://twitter.com/albertvilella> <http://tinyurl.com/ngslytics> #Editable version: <http://tinyurl.com/ngsspecsshared>

#curl "https://docs.google.com/spreadsheets/d/1GMMfhYLK0-q8Xklo3YxlWaZA5vVMuhU1kg41g4xLkXc/export?gid=4&format=csv" | grep -v '^#' | grep -v '^\$' | column -t -s, | less -S

Stats circa 2018

For current, see: <https://tinyurl.com/wbgcs65>



Illumina

PacBio

ONT

*Not all shown at scale

Maybe something fast and portable?



Oxford Nanopore Technology (ONT) Minion



A Plethora of Biological Sequence Analyses Enabled by RNA-Seq

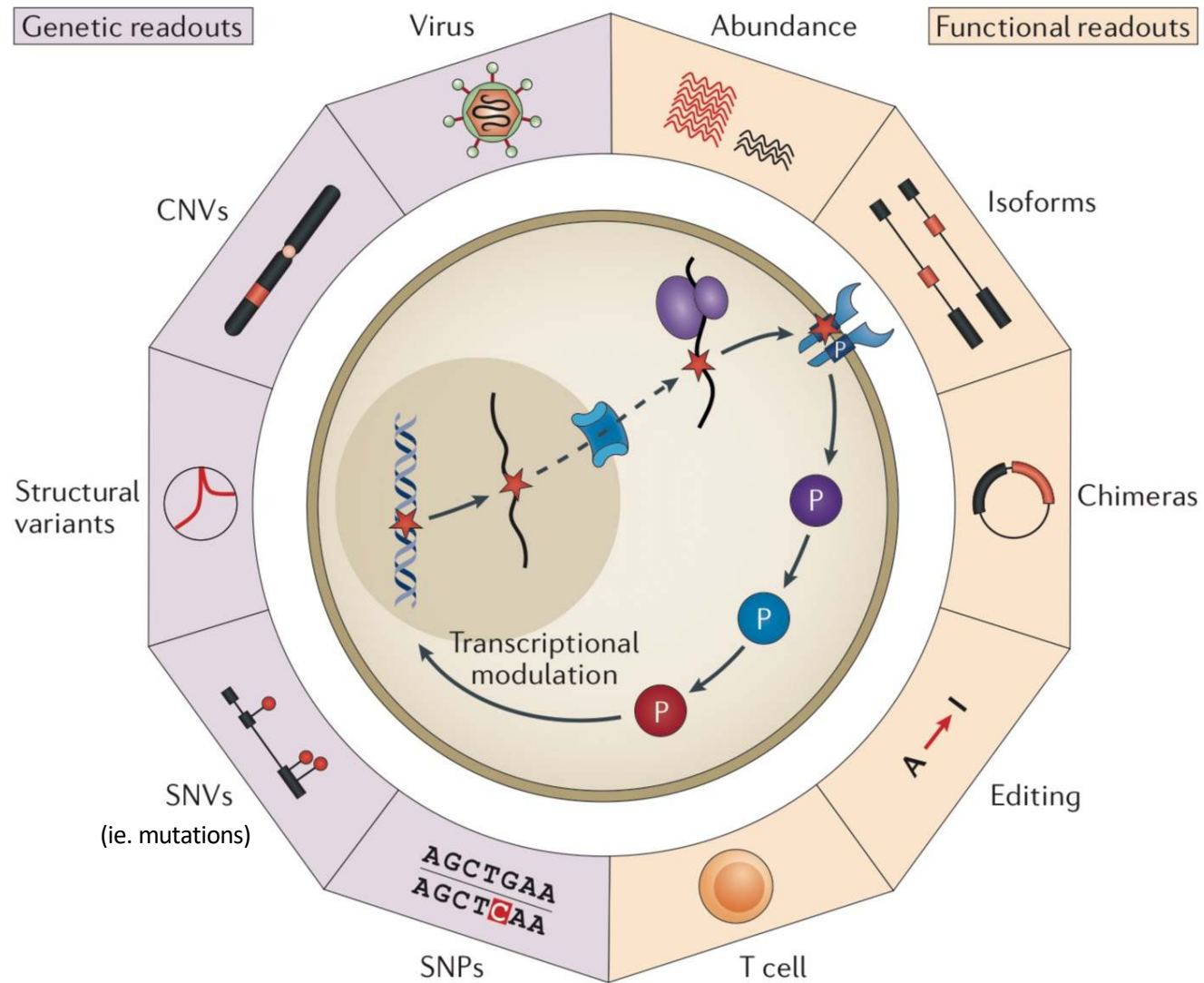
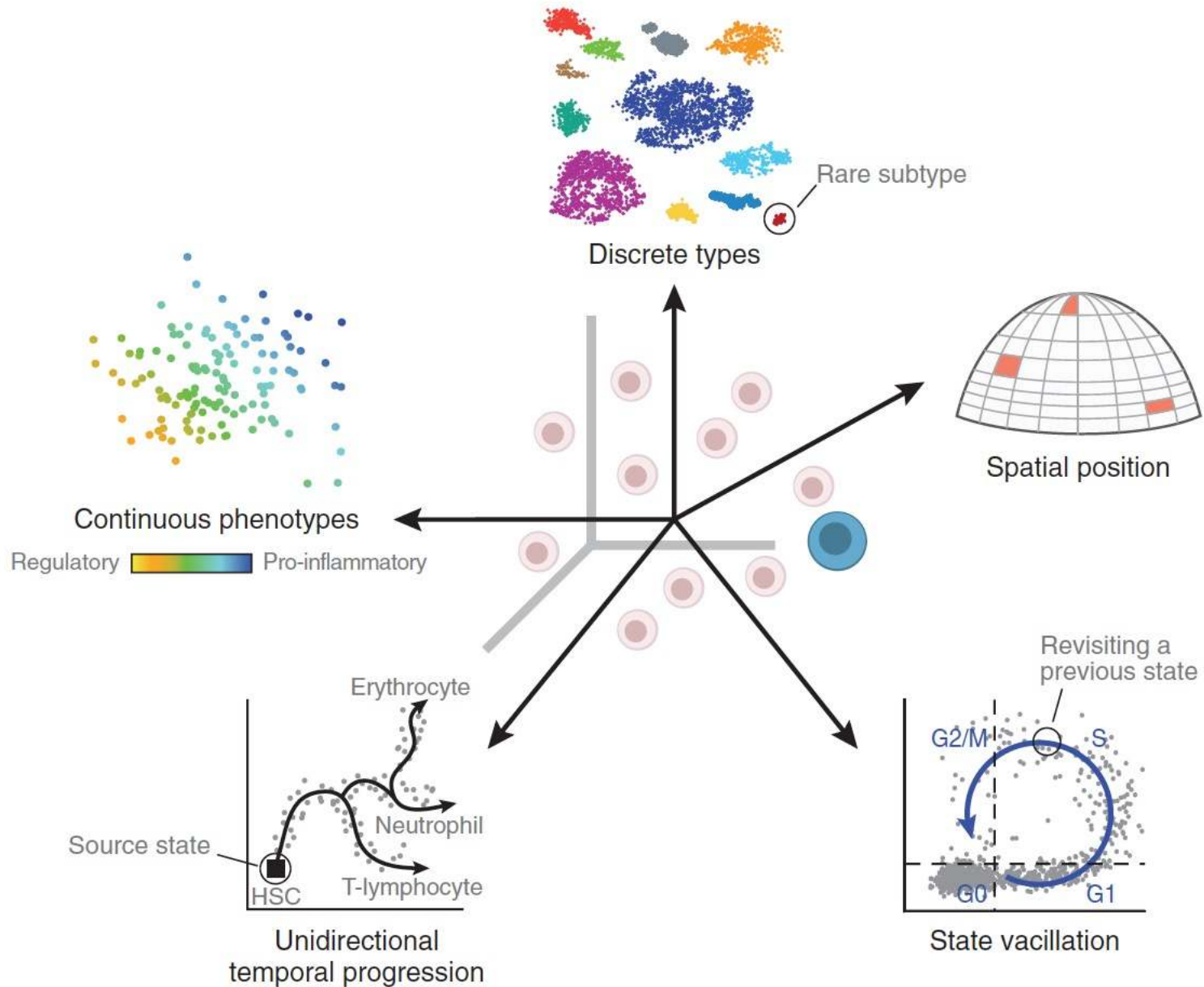


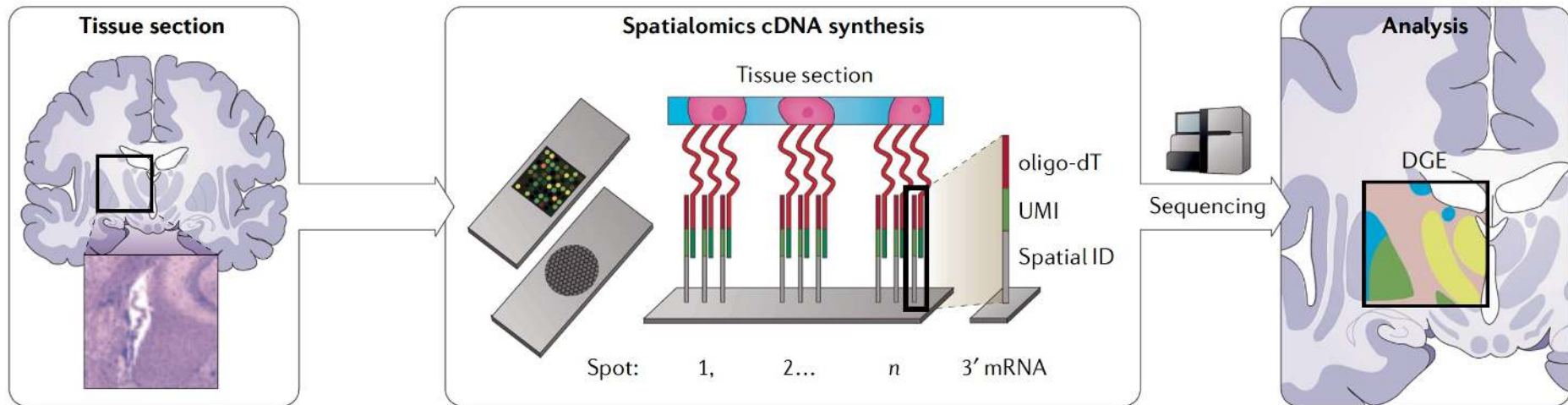
Figure 2 | **Transcriptome profiling for genetic causes and functional phenotypic readouts.**

RNA-Seq is Empowering Discovery at Single Cell Resolution



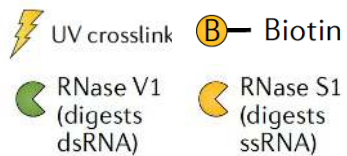
Spatial Transcriptomics

Spatial Encoding



A Myriad of Other Specialized RNA-seq -based Applications

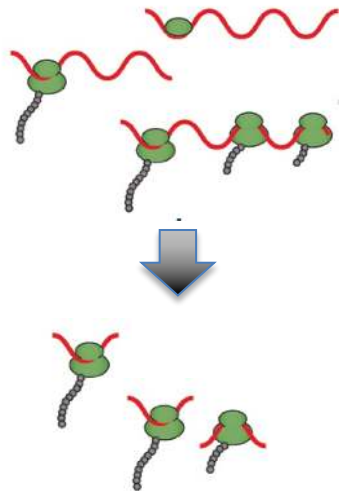
RNA-Sequencing as your lens towards biological discovery



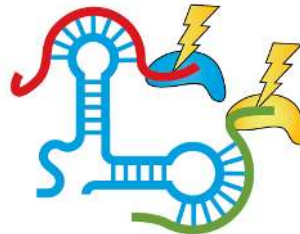
Adapted from "RNA sequencing: the teenage years"
Rory Stark, Marta Grzelak & James Hadfield
Nature Reviews Genetics volume 20, pages631–656(2019)

A Myriad of Other Specialized RNA-seq -based Applications

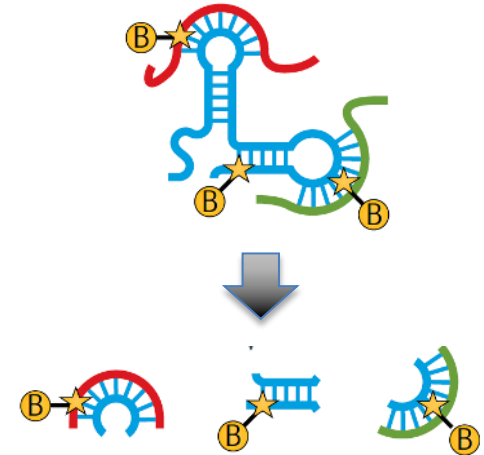
Ribosomal profiling



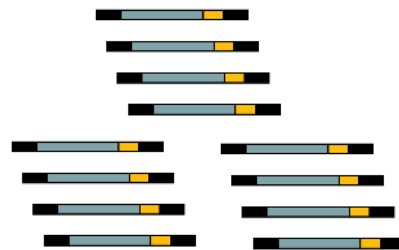
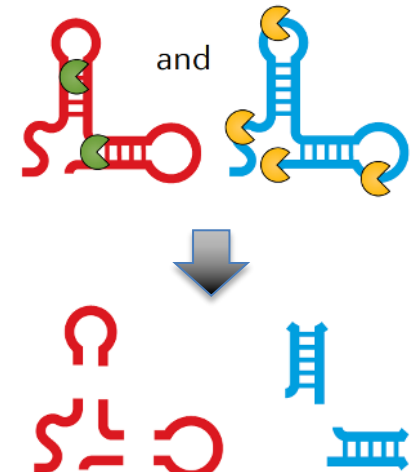
RNA-Protein Interactions



RNA-RNA interactions



RNA Structuromics

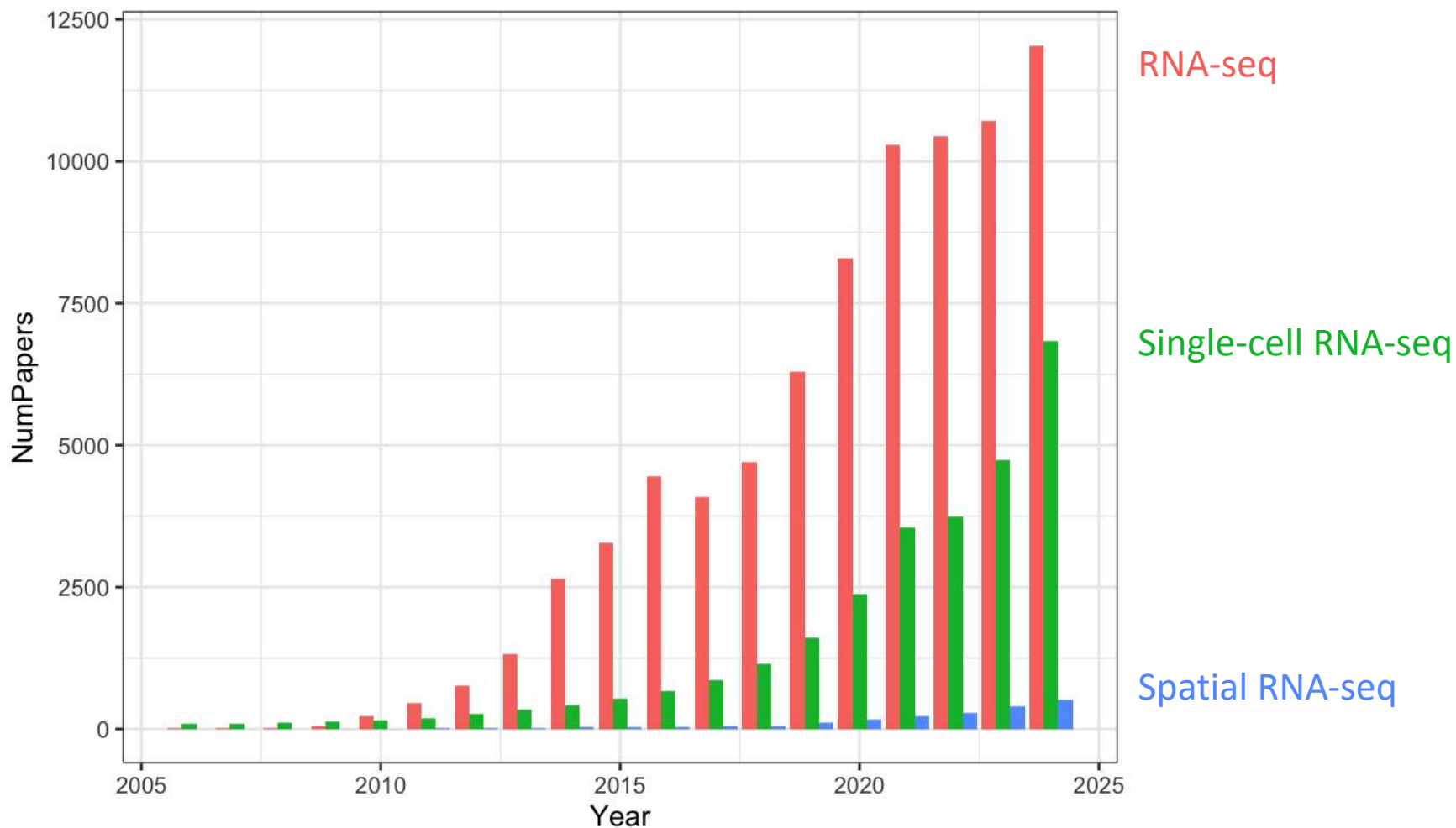


⚡ UV crosslink B — Biotin
 🍷 RNase V1 (digests dsRNA) 🍷 RNase S1 (digests ssRNA)

Adapted from "RNA sequencing: the teenage years"
 Rory Stark, Marta Grzelak & James Hadfield
 Nature Reviews Genetics volume 20, pages631–656(2019)

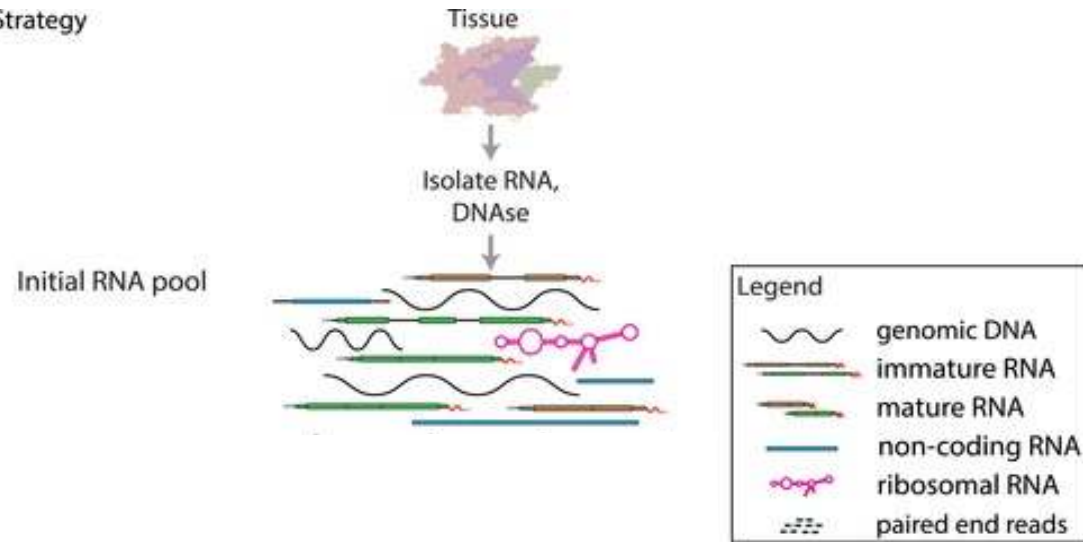
RNA-seq Publication Trend

Paper Counts from PubMed



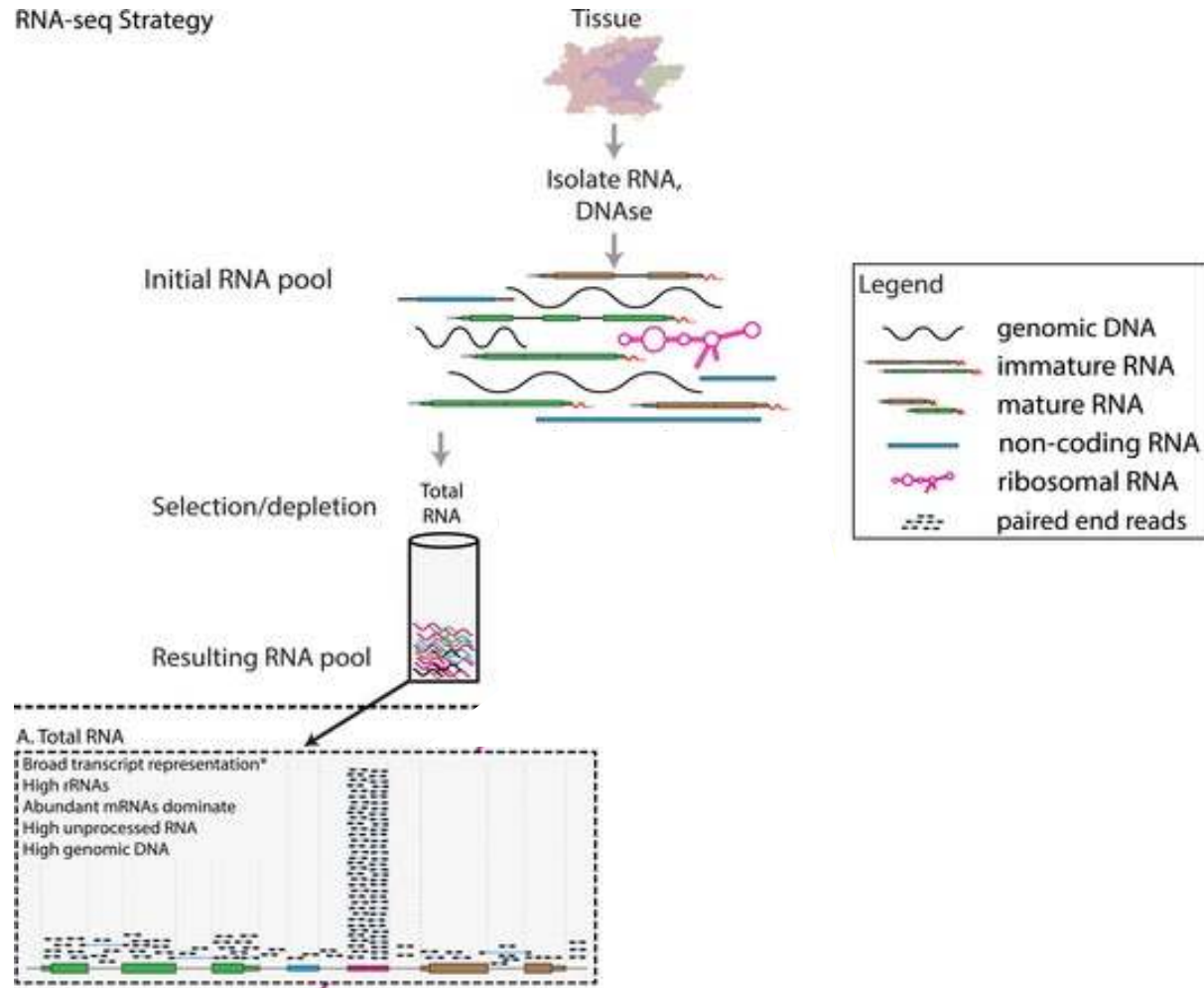
RNA-seq library enrichment strategies that influence interpretation and analysis.

RNA-seq Strategy



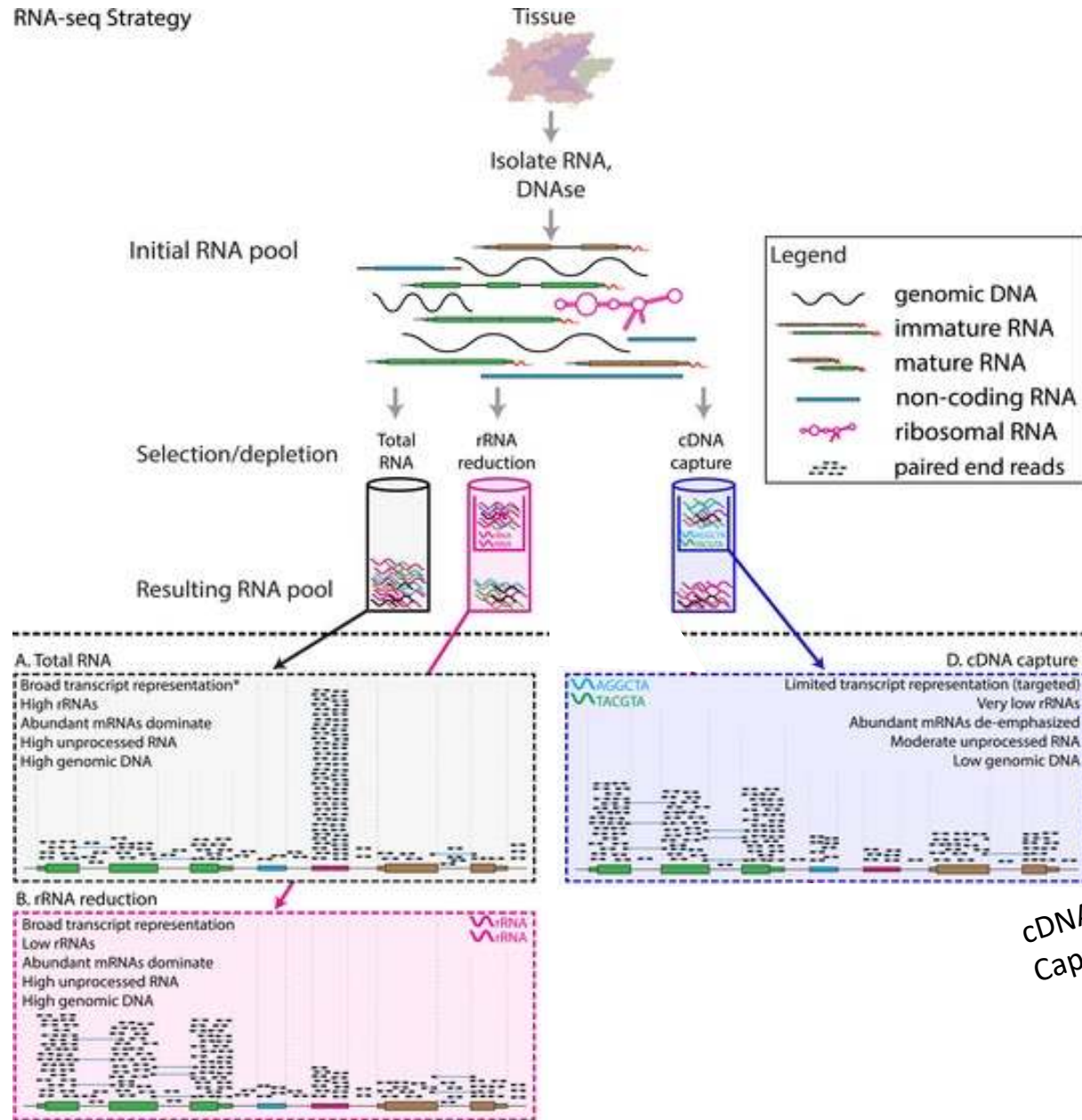
RNA-seq library enrichment strategies that influence interpretation and analysis.

RNA-seq Strategy



RNA-seq library enrichment strategies that influence interpretation and analysis.

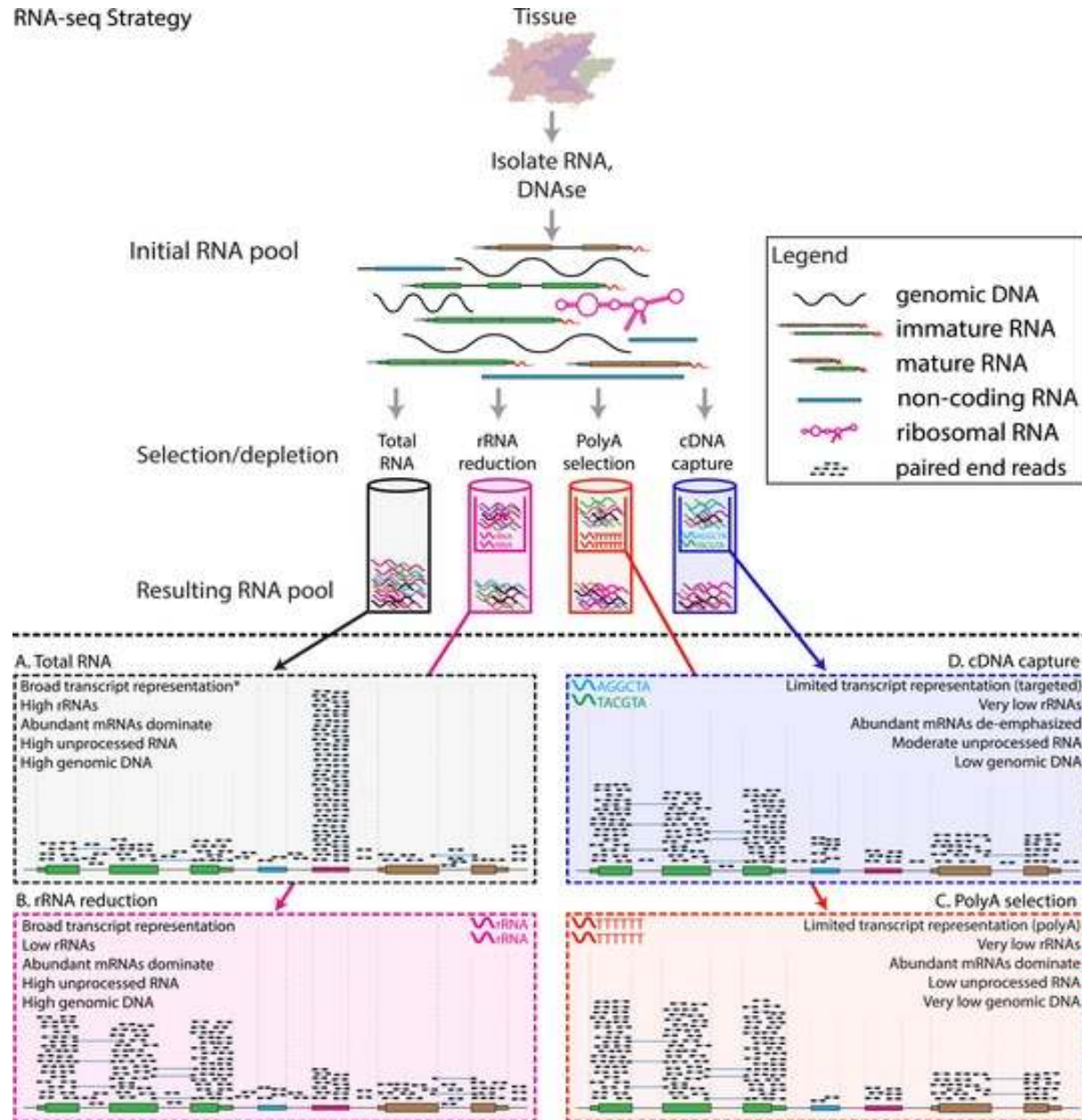
RNA-seq Strategy



Expected Alignments

RNA-seq library enrichment strategies that influence interpretation and analysis.

RNA-seq Strategy



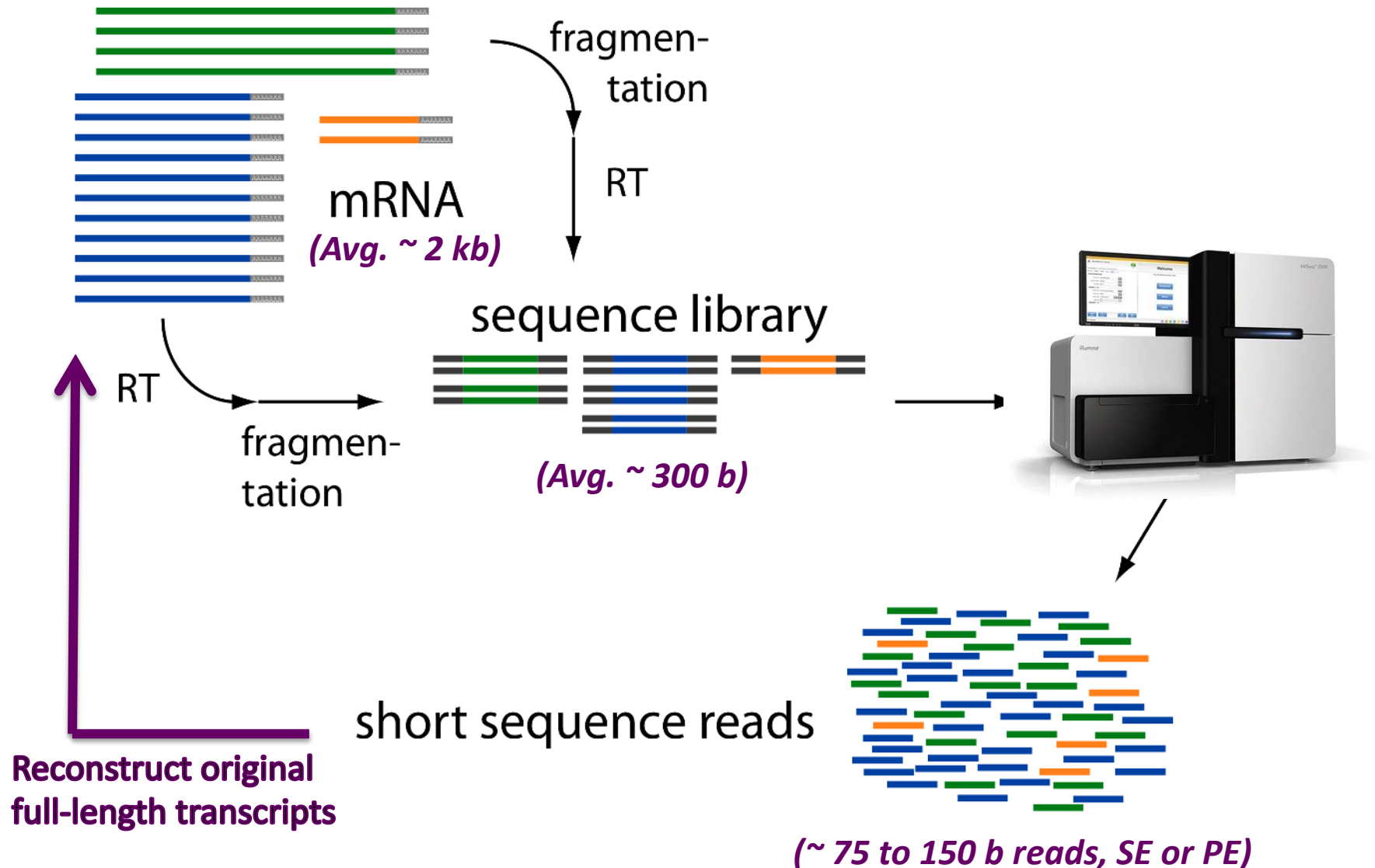
Expected Alignments

AAAAA

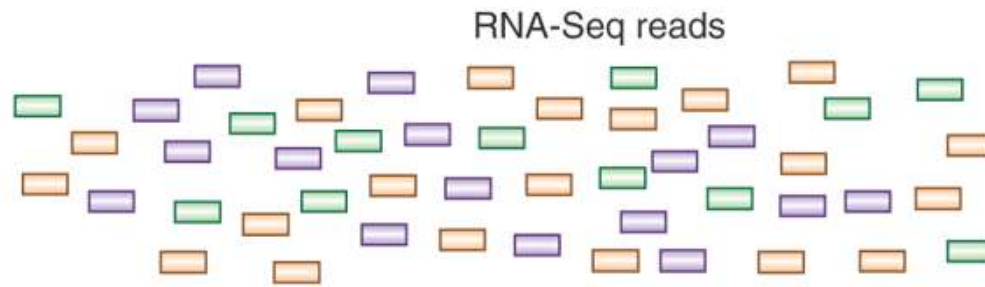


Part 2. Transcript Reconstruction Methods

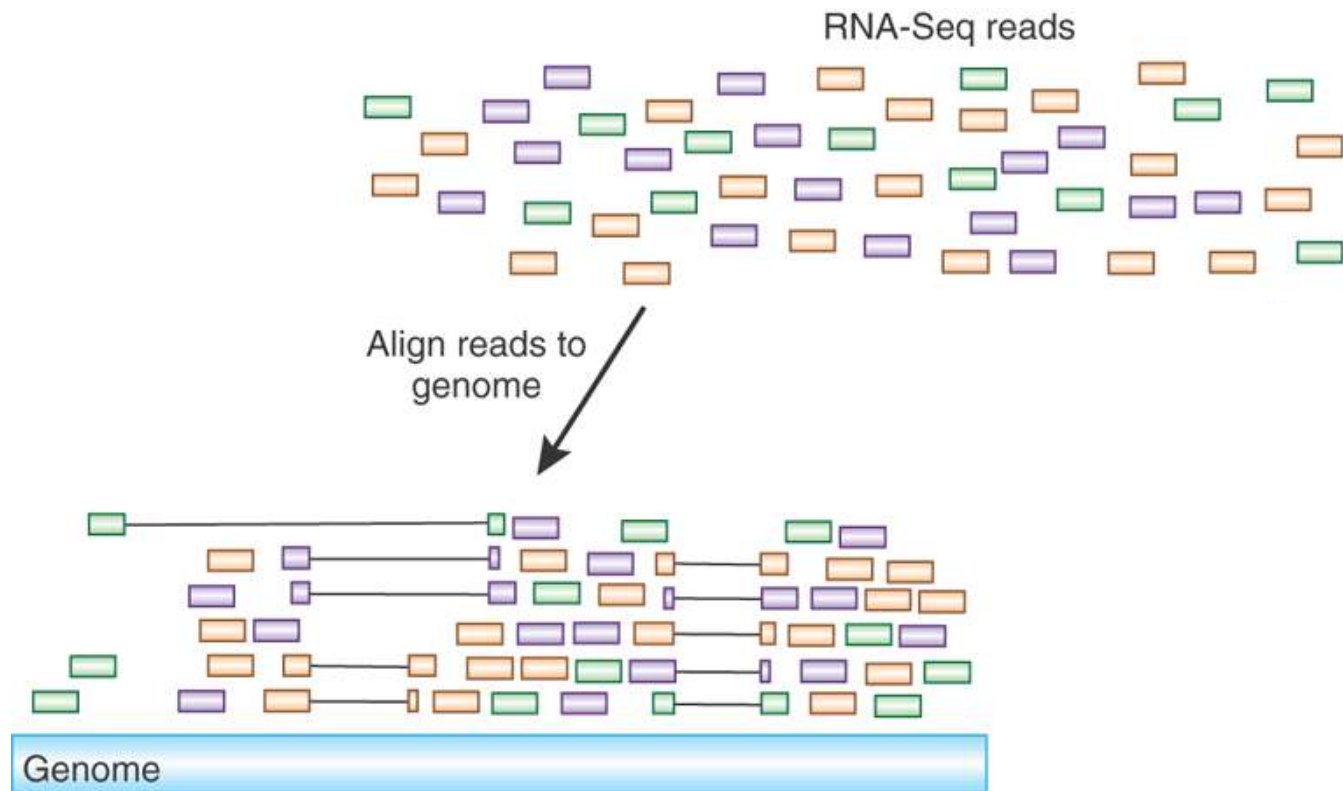
RNA-Seq Challenge: Transcript Reconstruction



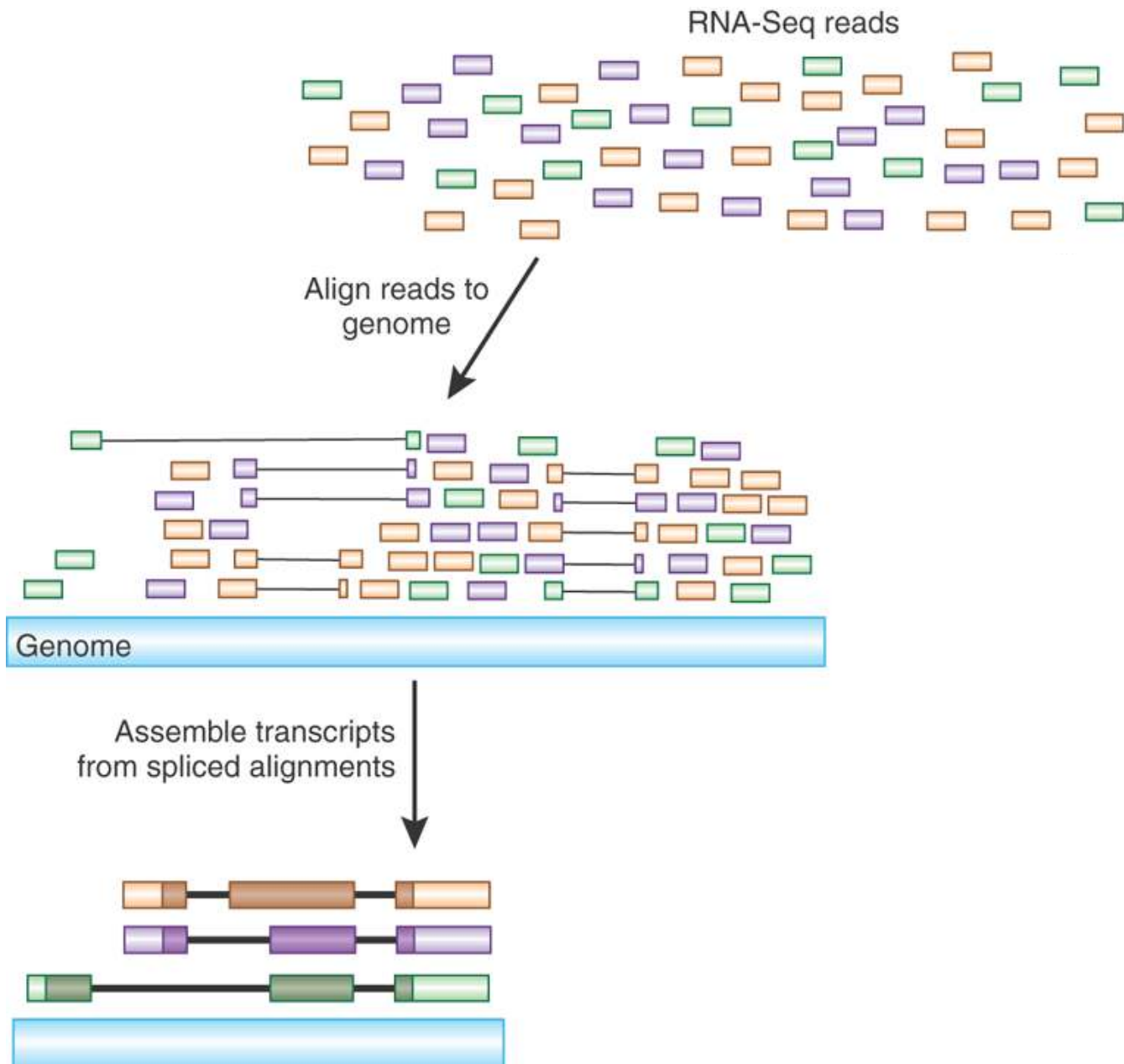
Transcript Reconstruction from (short) RNA-Seq Reads



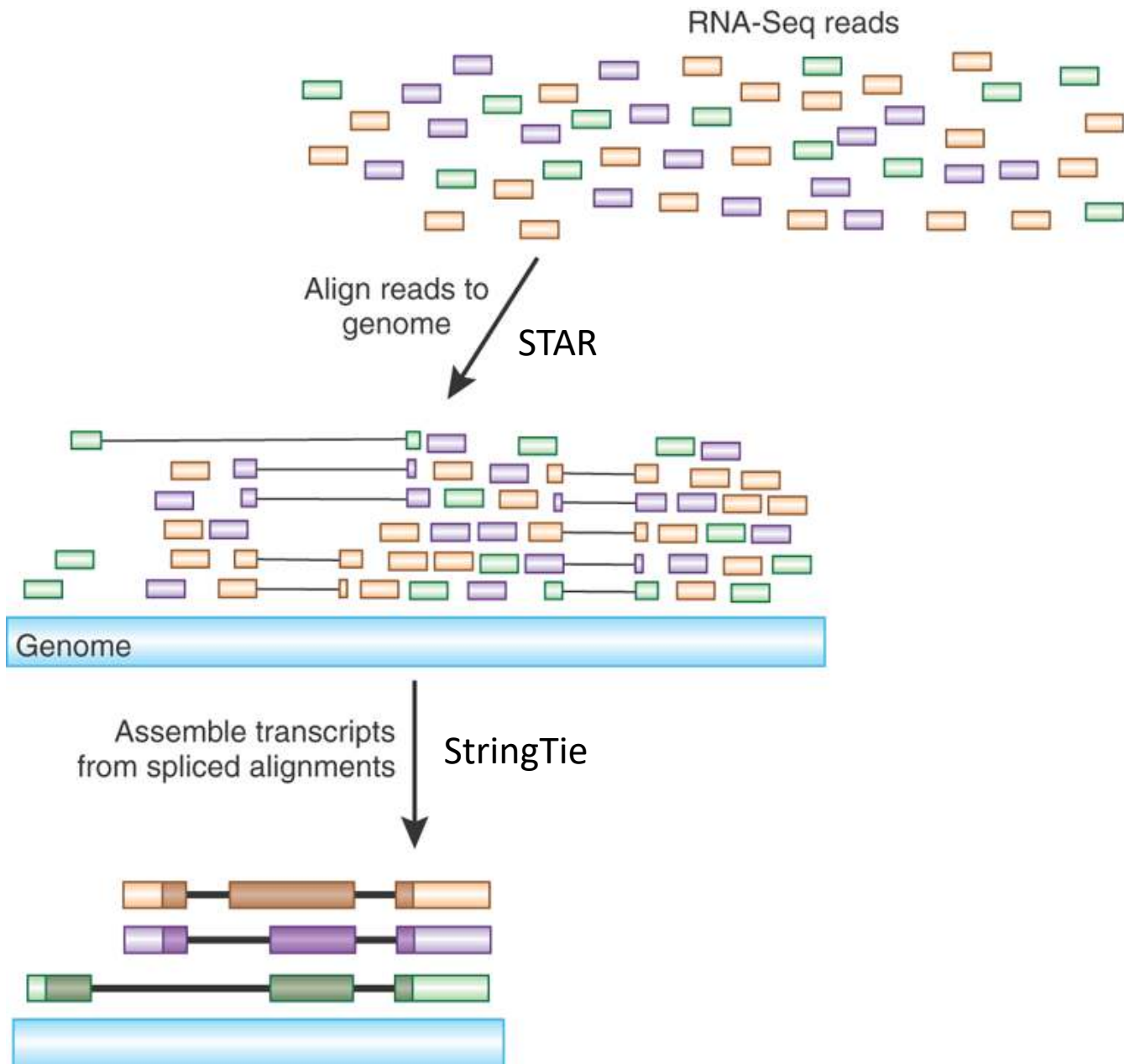
Transcript Reconstruction from (short) RNA-Seq Reads



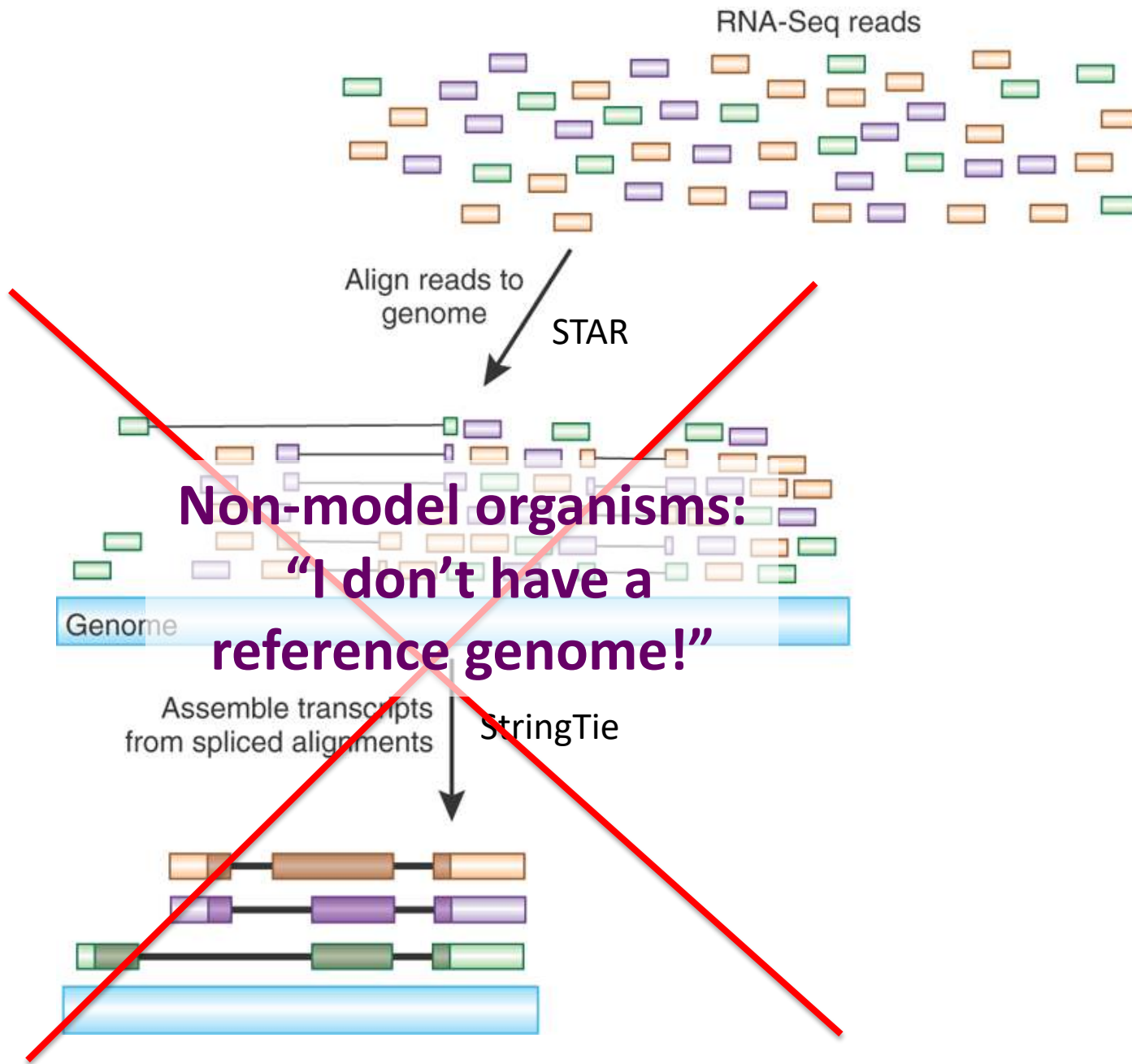
Transcript Reconstruction from (short) RNA-Seq Reads



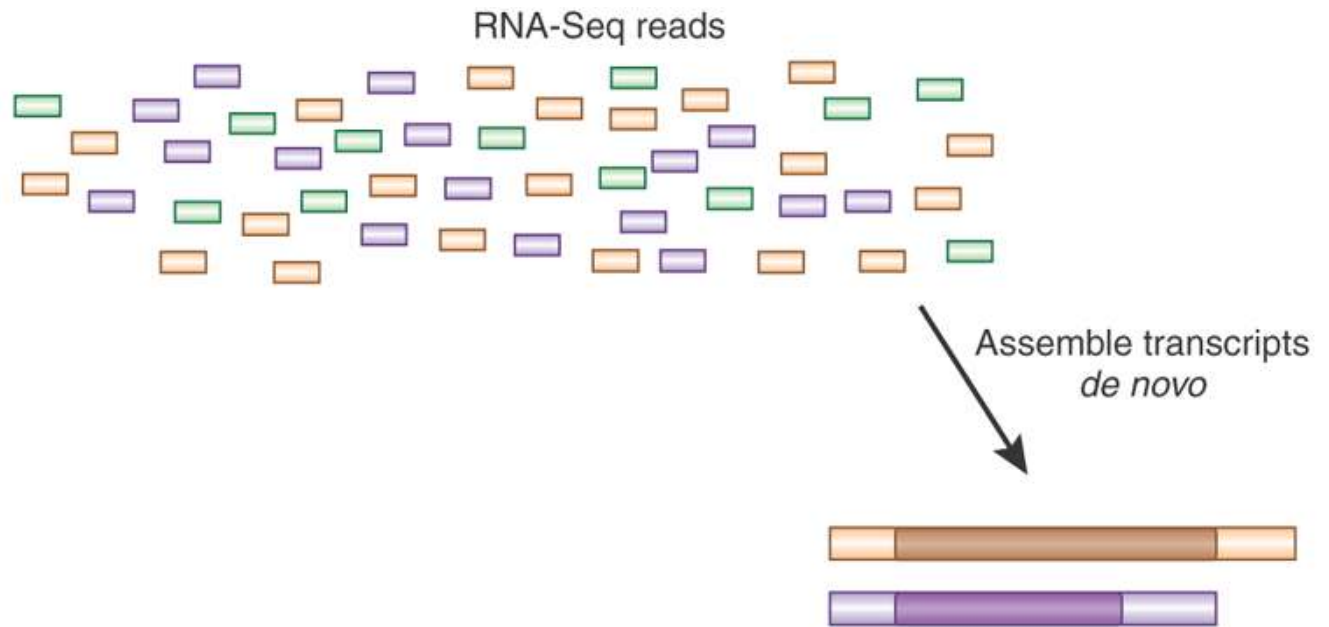
Transcript Reconstruction from (short) RNA-Seq Reads



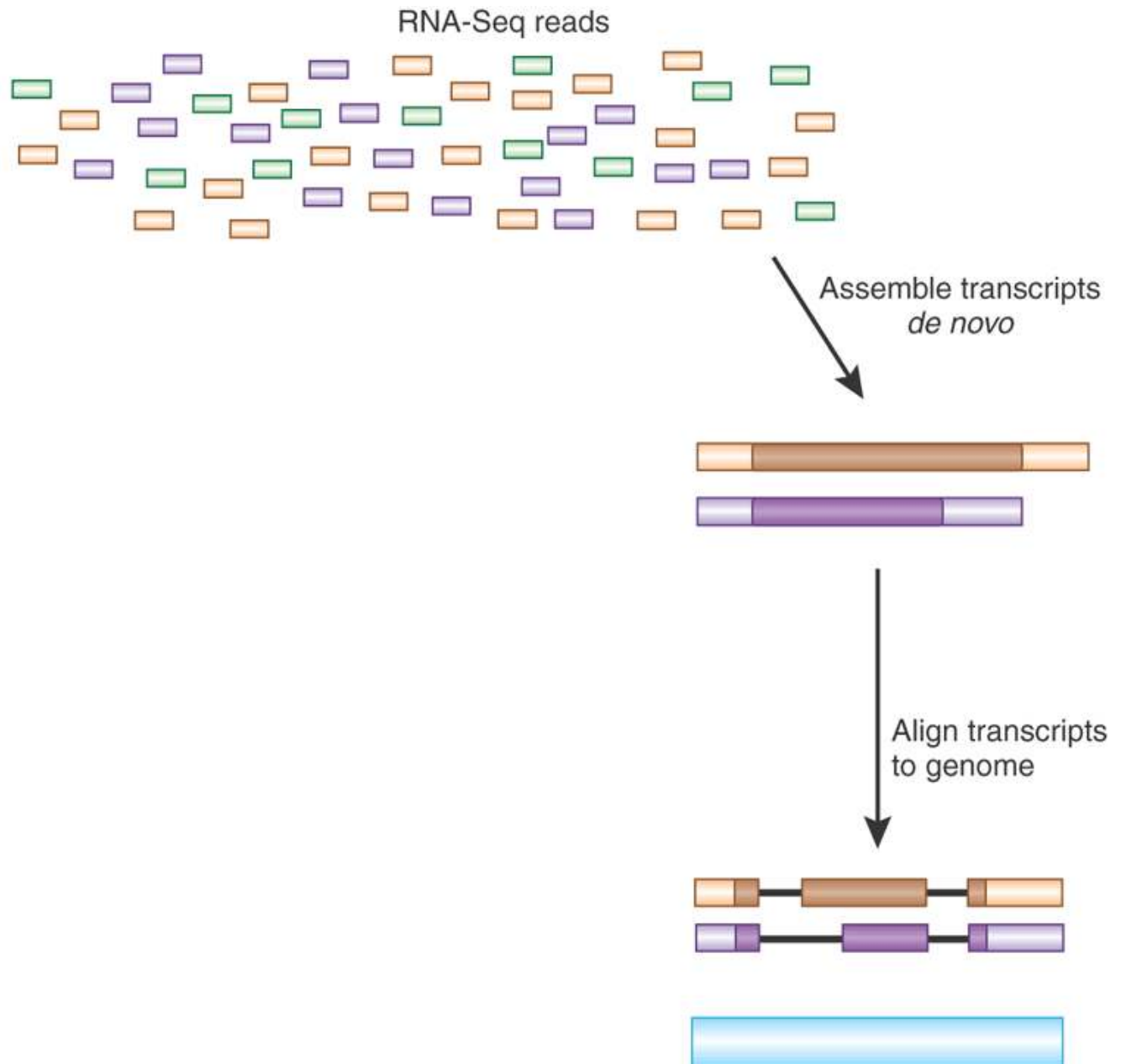
Transcript Reconstruction from (short) RNA-Seq Reads



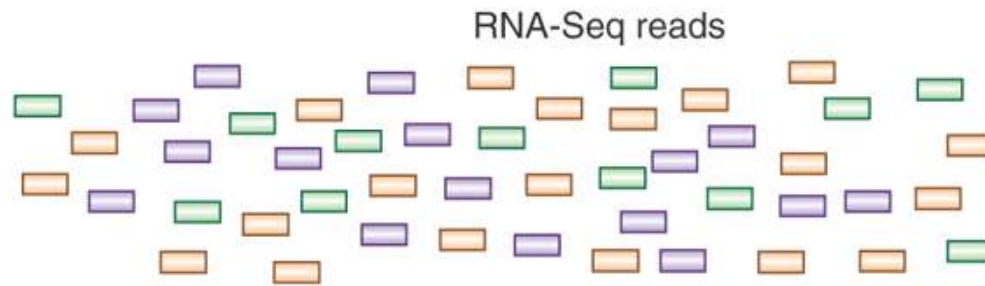
Transcript Reconstruction from (short) RNA-Seq Reads



Transcript Reconstruction from (short) RNA-Seq Reads



Transcript Reconstruction from (short) RNA-Seq Reads



End-to-end **Transcriptome**-based RNA-Seq Analysis Software Package

NATURE PROTOCOLS | PROTOCOL

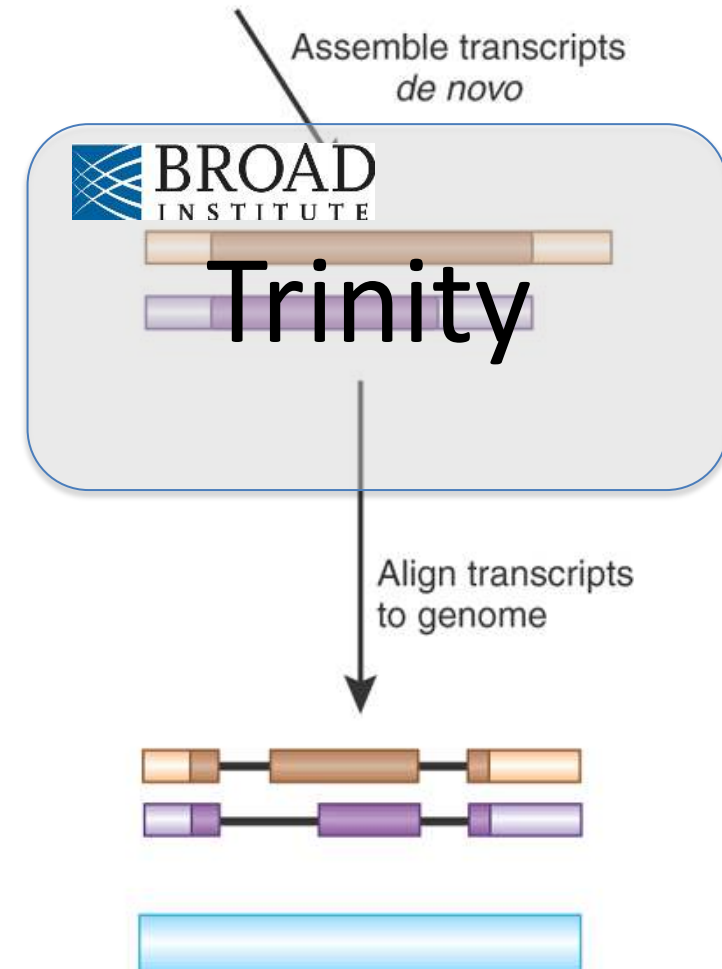
De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis

Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey, Robert Henschel, Richard D LeDuc, Nir Friedman & Aviv Regev

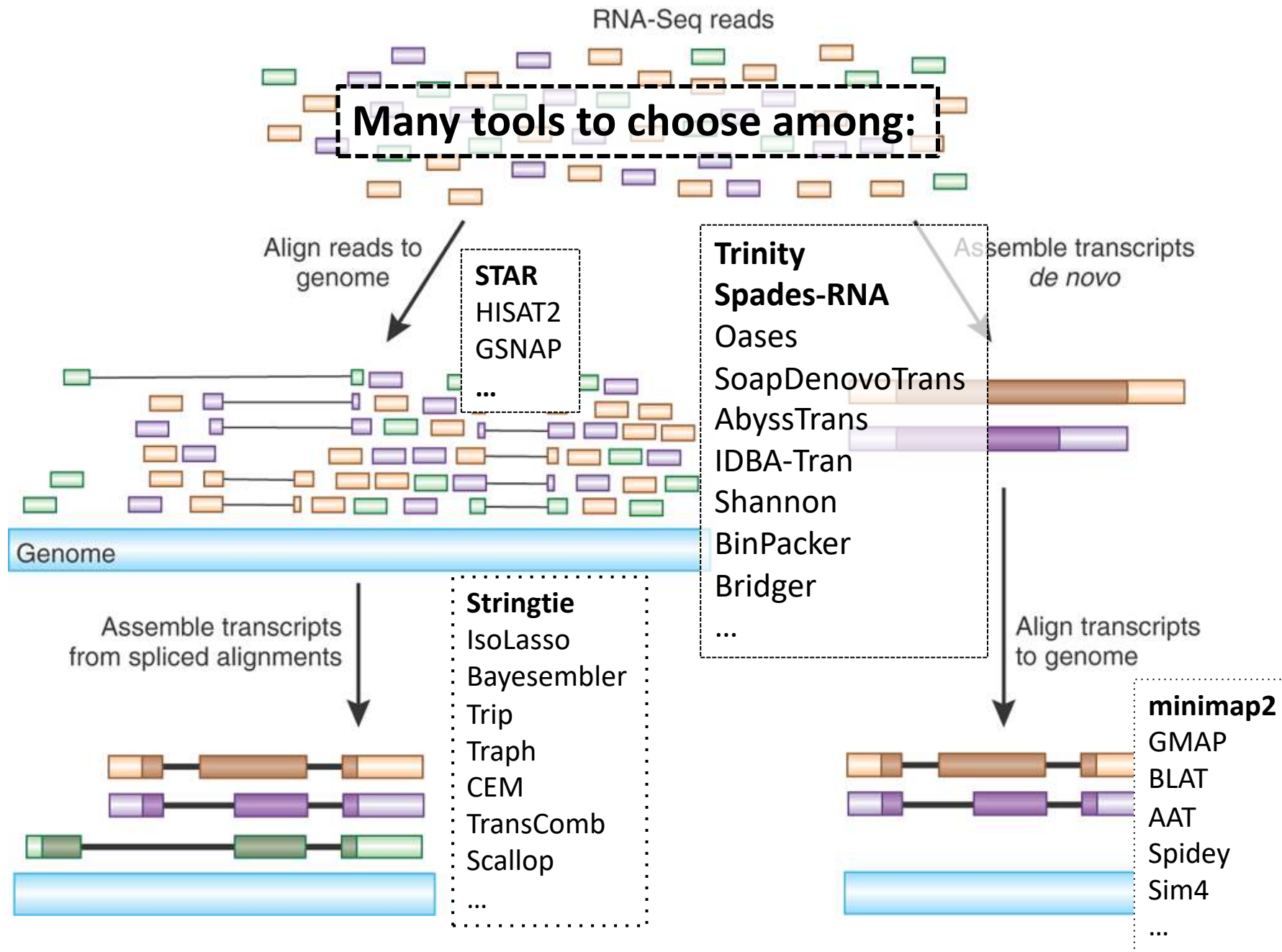
[Affiliations](#) | [Contributions](#) | [Corresponding authors](#)

Nature Protocols **8**, 1494–1512 (2013) | doi:10.1038/nprot.2013.084

Published online 11 July 2013



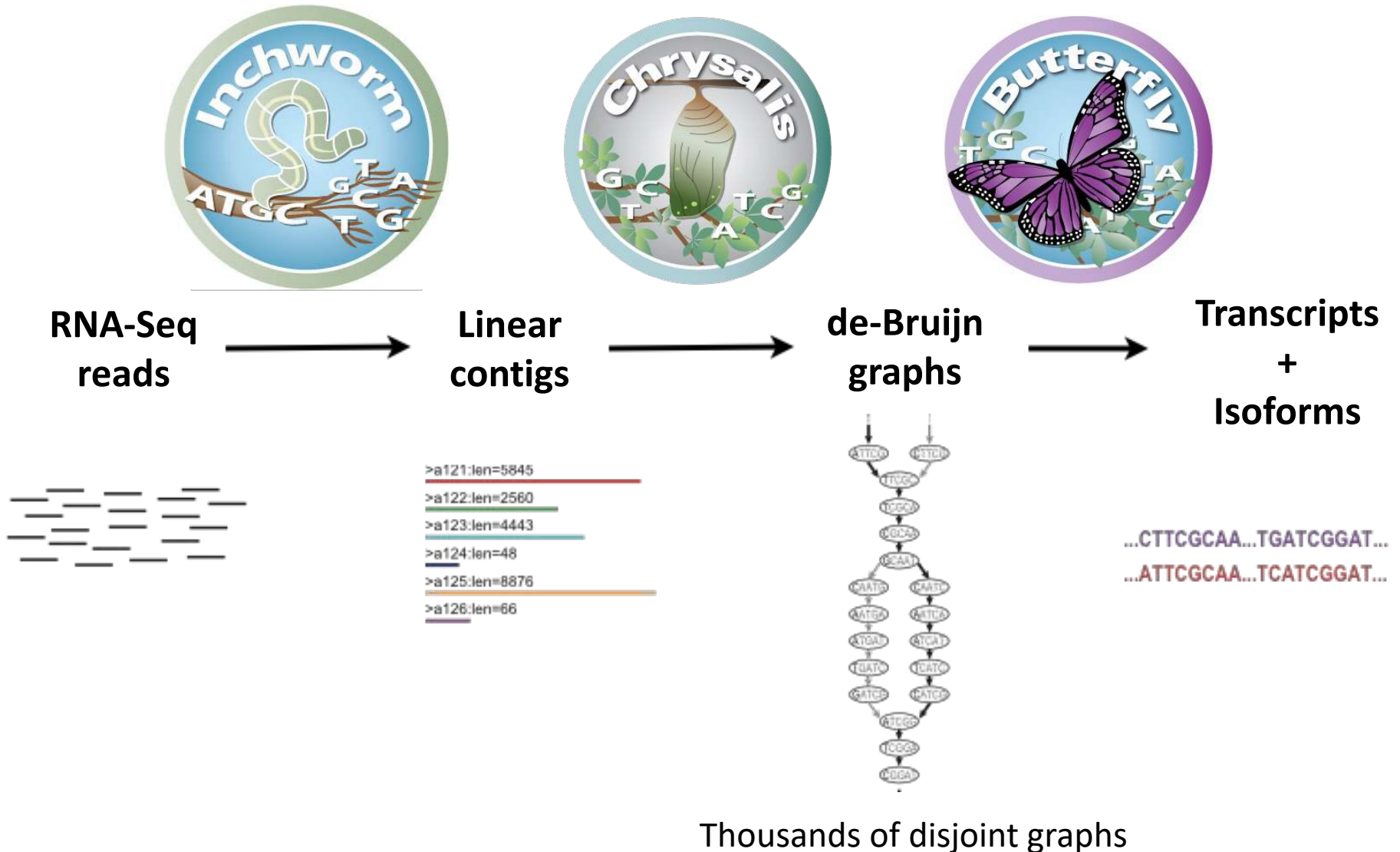
Transcript Reconstruction from (short) RNA-Seq Reads





Part 3. Trinity for Genome-free transcriptomics (eg. for non-model orgs)

Trinity – How it works:



Trinity – How it works:



Younger
me



Manfred
Grabherr



Moran
Yassour

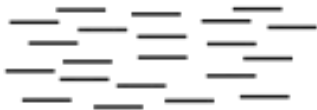


RNA-Seq
reads

Linear
contigs

de-Bruijn
graphs

Transcripts
+
Isoforms



```
>a121:len=5845
>a122:len=2560
>a123:len=4443
>a124:len=48
>a125:len=8876
>a126:len=66
```



...CTTCGCAA...TGATCGGAT...
...ATTCGCAA...TCATCGGAT...

Thousands of disjoint graphs

Trinity – How it works:



RNA-Seq
reads

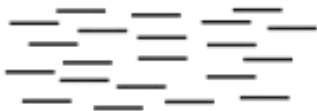


Linear
contigs



de-Bruijn
graphs

Transcripts
+
Isoforms



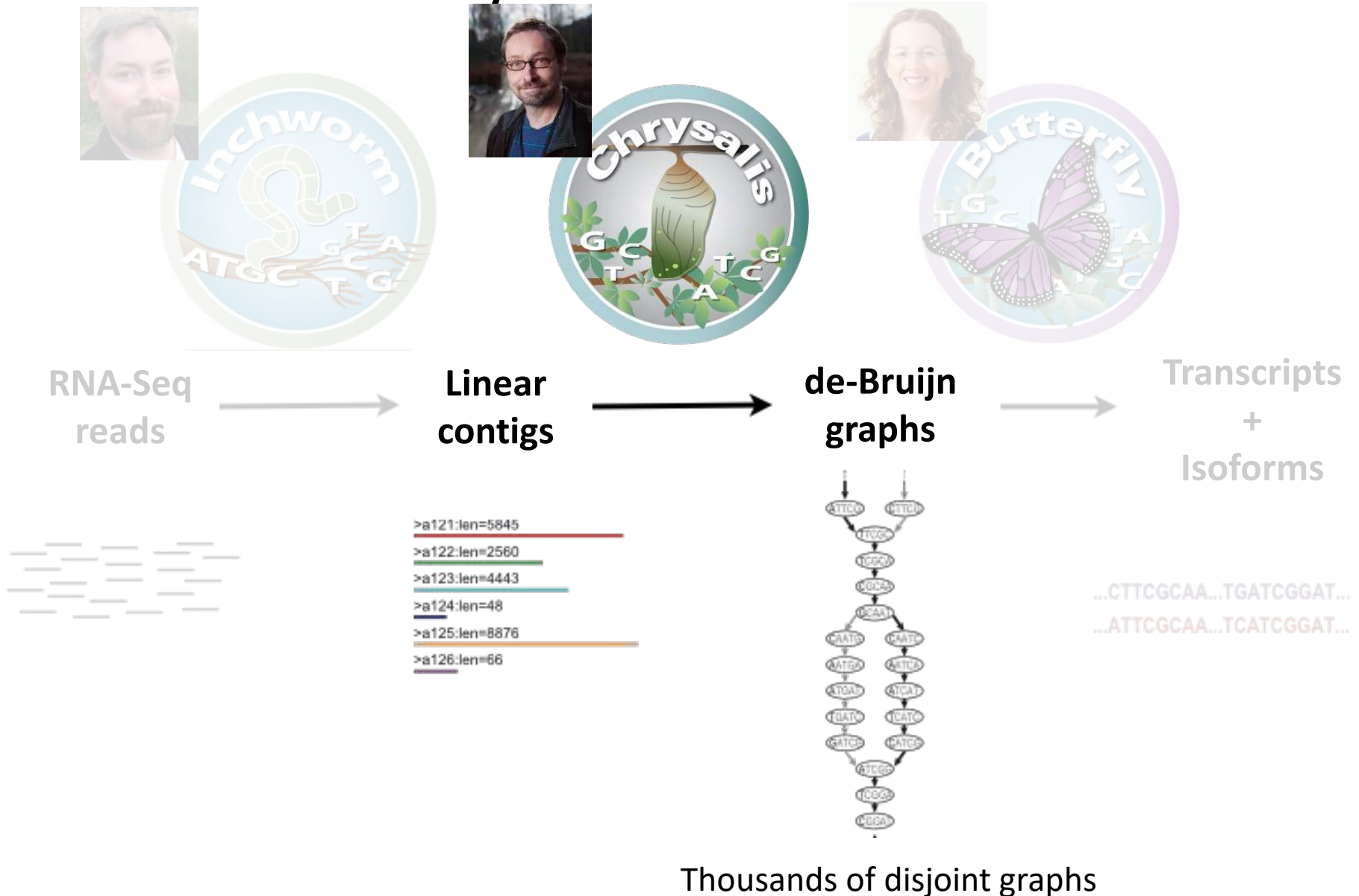
```
>a121:len=5845
>a122:len=2560
>a123:len=4443
>a124:len=48
>a125:len=8876
>a126:len=66
```



```
...CTTCGCAA...TGATCGGAT...
...ATTCGCAA...TCATCGGAT...
```

Thousands of disjoint graphs

Trinity – How it works:





(continued)

Linear contigs

```

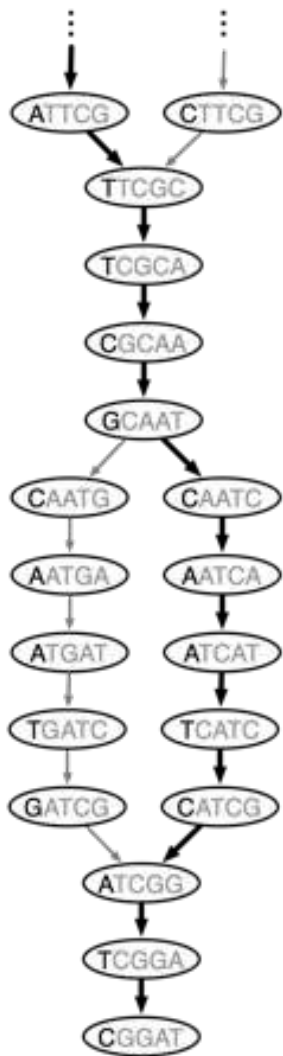
graph TD
    s((s)) --> ATTC((ATTC))
    s --> HTTC((HTTC))
    ATTC --> TTGC((TTGC))
    HTTC --> TTGC
    TTGC --> TCGC((TCGC))
    TCGC --> TCCG((TCCG))
    TCCG --> TCCAT((TCCAT))
    TCCAT --> TCCATG((TCCATG))
    TCCAT --> TCCATC((TCCATC))
    TCCATG --> ATTCG((ATTCG))
    TCCATC --> ATTC((ATTC))
    ATTCG --> ATCG((ATCG))
    ATTC --> ATCG
    ATCG --> TATG((TATG))
    ATTC --> TATC((TATC))
    TATG --> GATG((GATG))
    TATC --> TATCG((TATCG))
    GATG --> ATCGG((ATCGG))
    TATCG --> ATCGG
    ATCGG --> TCGG((TCGG))
    TCGG --> t((t))
  
```

Transcripts + Isoforms

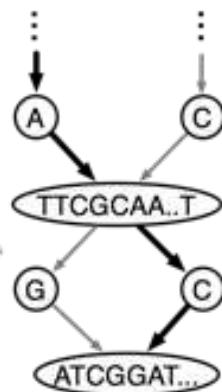
...ATTCGCAA...TCATCGGAT...

Thousands of disjoint graphs

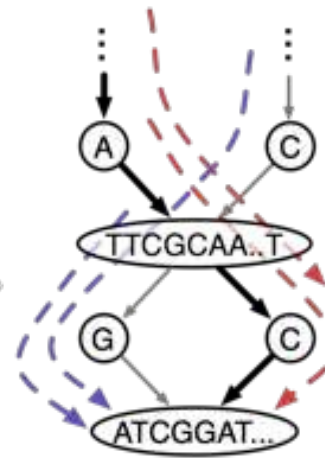
Butterfly



de Bruijn
graph



compact
graph



compact
graph with
reads

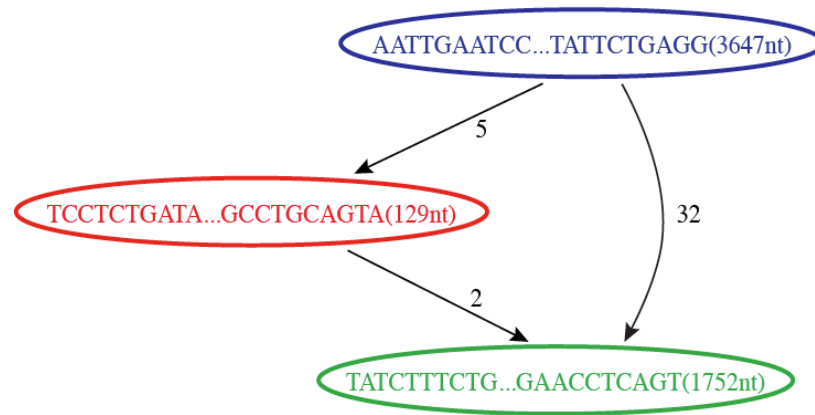


..CTTCGCAA..TGATCGGAT..
..ATTGCAA..TCATCGGAT..

sequences
(isoforms and paralogs)

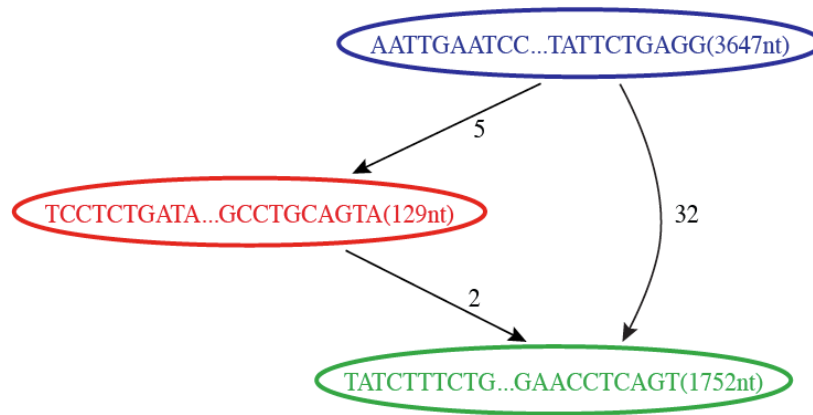
Butterfly Example 1: Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted
Sequence Graph



Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted
Sequence Graph

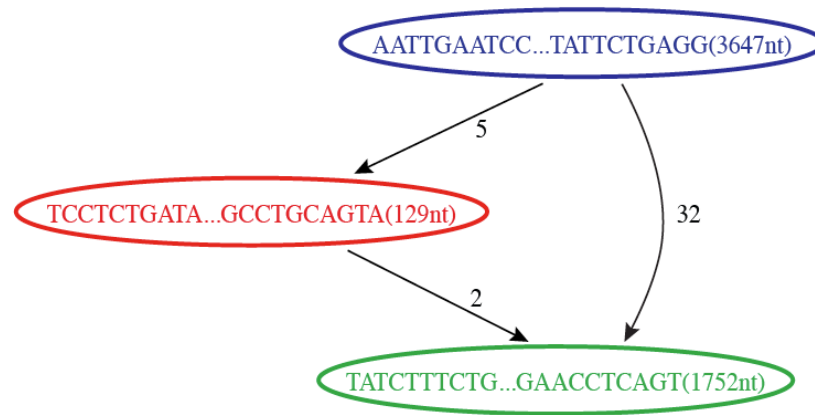


Reconstructed Transcripts



Reconstruction of Alternatively Spliced Transcripts

Butterfly's Compacted
Sequence Graph

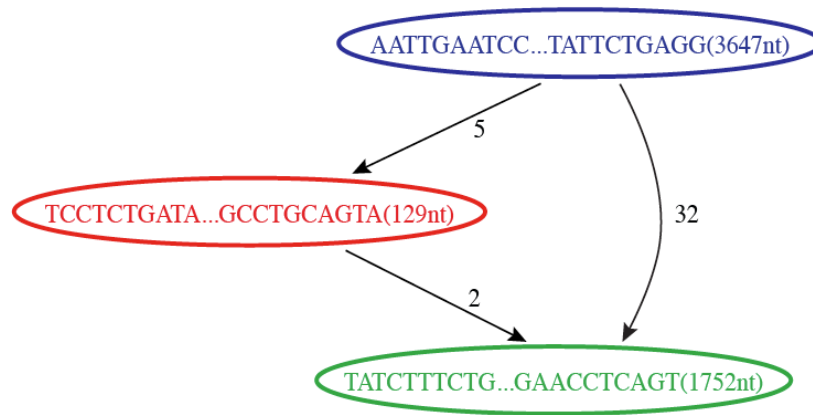


Reconstructed Transcripts



Reconstruction of Alternatively Spliced Transcripts

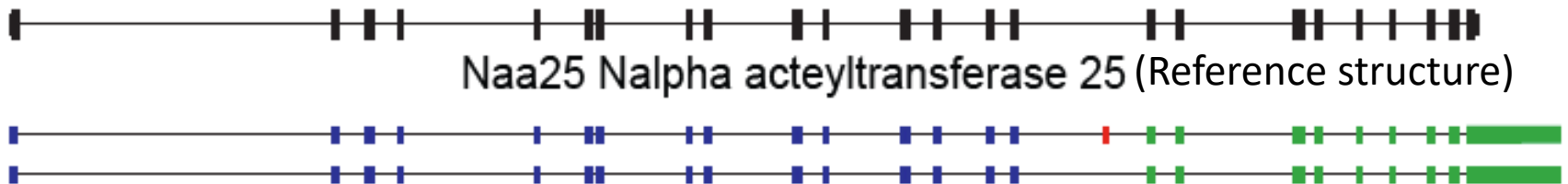
Butterfly's Compacted
Sequence Graph



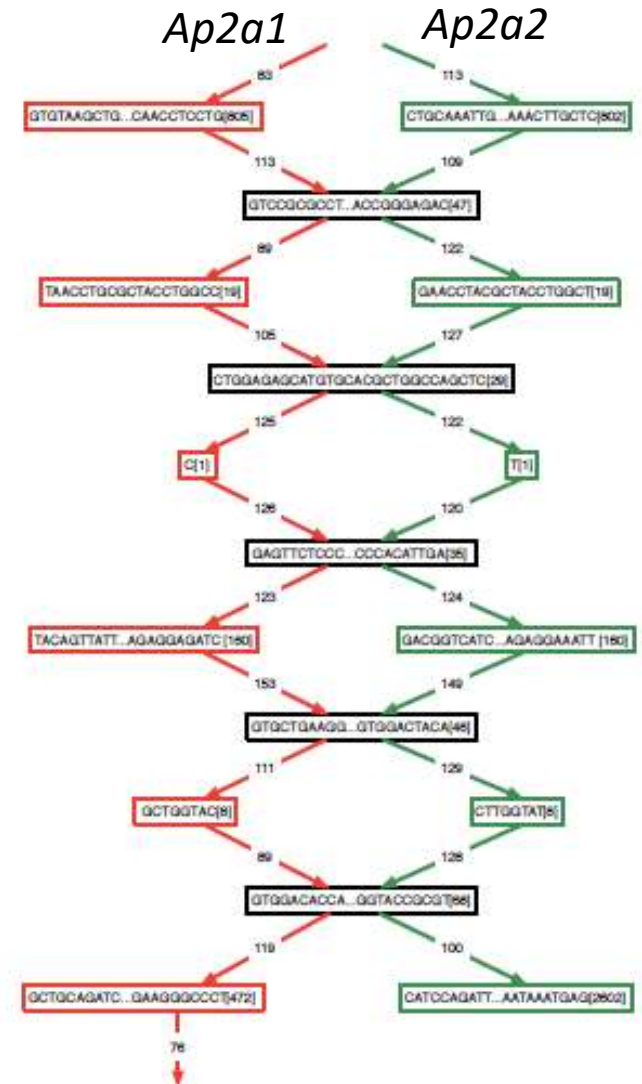
Reconstructed Transcripts



Aligned to Mouse Genome



Butterfly Example 2: Teasing Apart Transcripts of Paralogous Genes



Teasing Apart Transcripts of Paralogous Genes

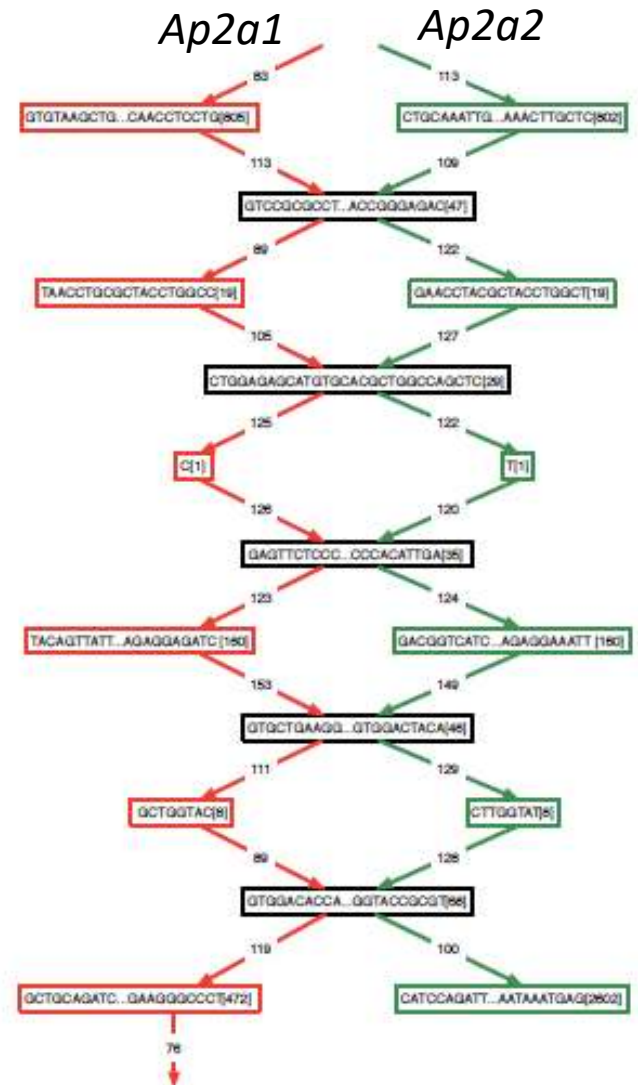
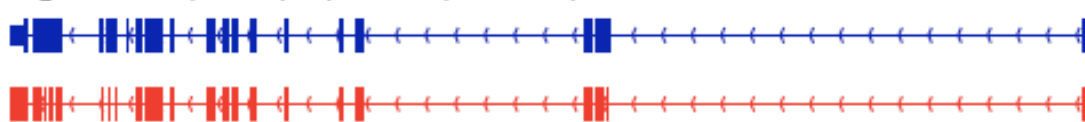
chr7:148,744,197-148,821,437

NM_007459; Ap2a2 adaptor protein complex AP-2, alpha 2 subunit



chr7:52,150,889-52,189,508

NM_001077264; Ap2a1 adaptor protein complex AP-2, alpha 1 subunit



Strand-specific RNA-Seq is Preferred

Computationally: fewer confounding graph structures in de novo assembly:

ex. Forward != reverse complement

(GGAA != TTCC)

Biologically: separate sense vs. antisense transcription

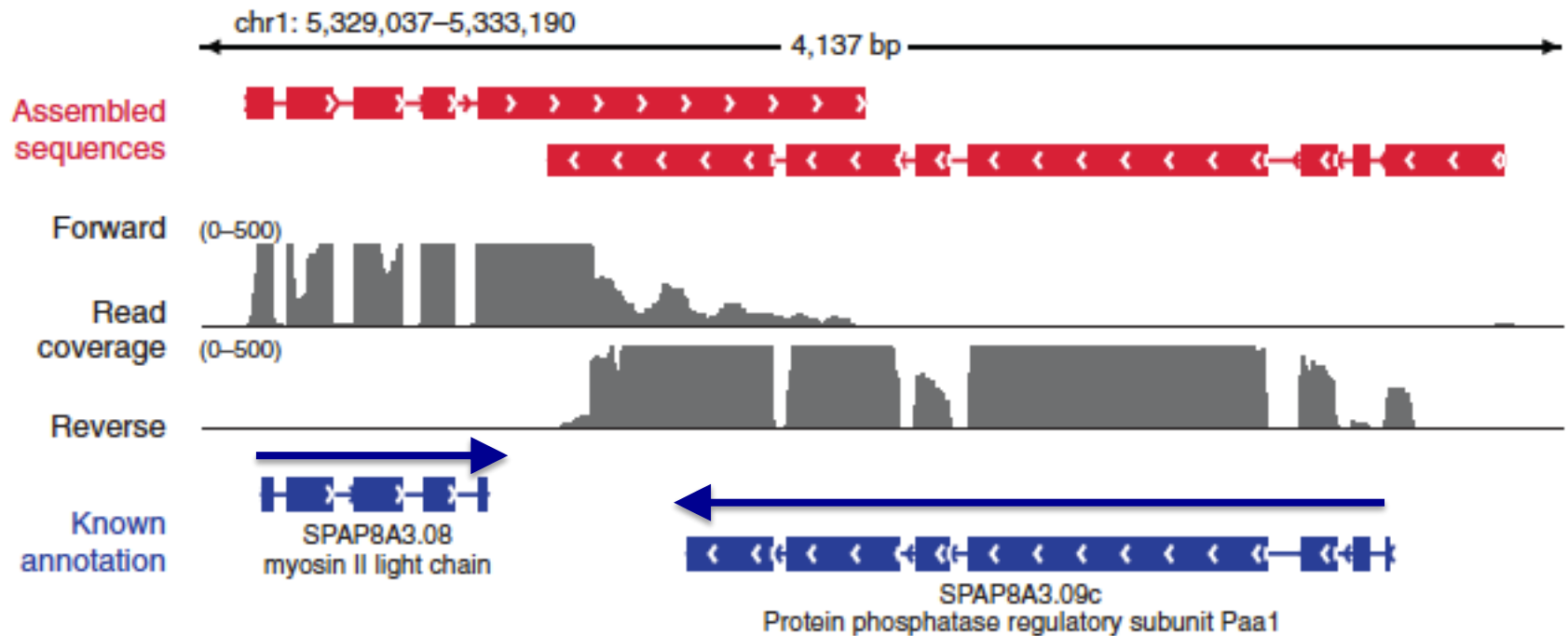
Illumina TruSeq Stranded mRNA Kit:



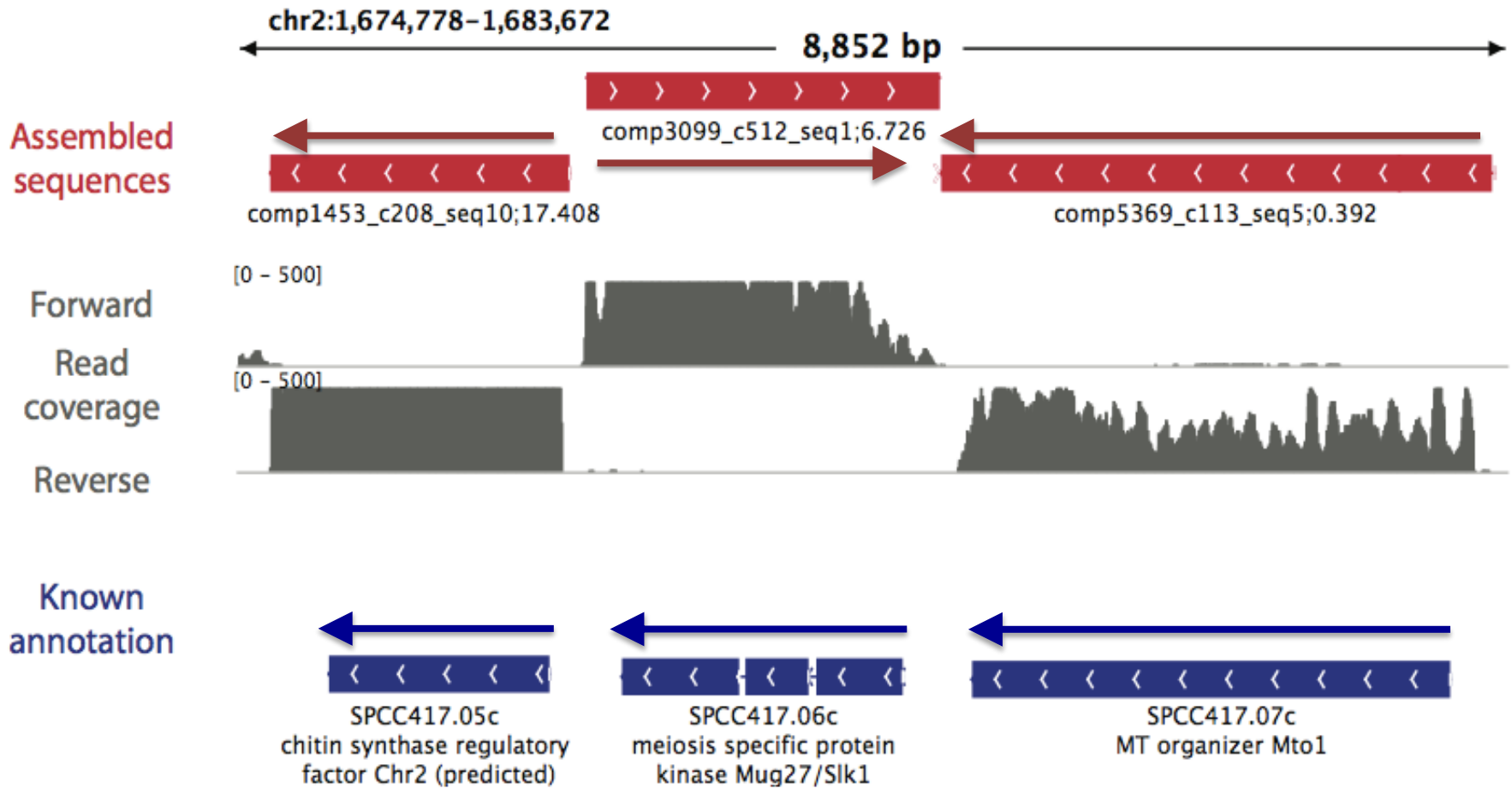
Overlapping UTRs from Opposite Strands



Schizosacharomyces pombe
(fission yeast)



Antisense-dominated Transcription



Trinity is a Highly Effective and Popular RNA-Seq Assembler



Nature Biotechnology, 2011

Thousands of routine users.

>15k literature citations

Freely Available, Well-supported,
Open Source Software



<http://trinityrnaseq.github.io>

Transcriptome Assembly is Just the End of the Beginning...

NATURE PROTOCOLS | PROTOCOL

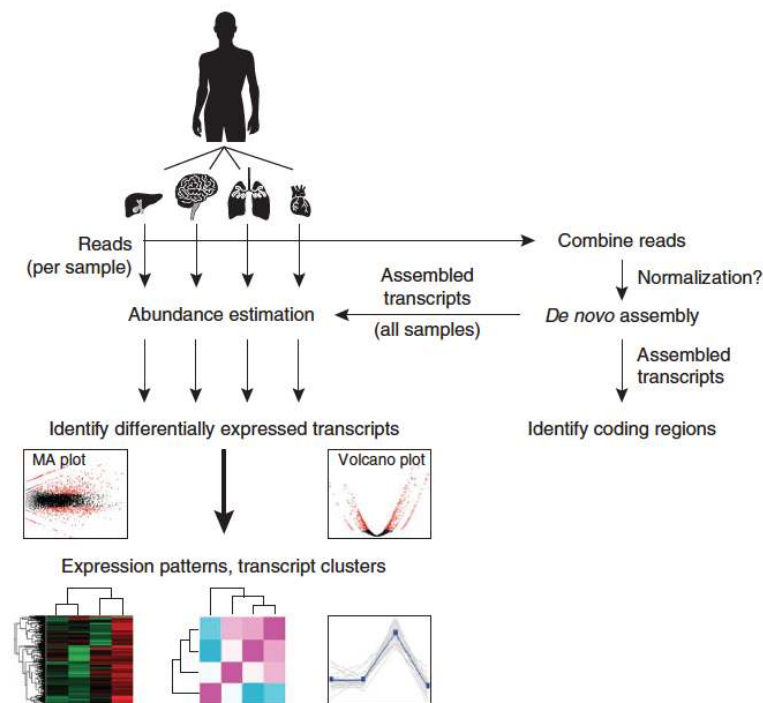
De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis

Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey, Robert Henschel, Richard D LeDuc, Nir Friedman & Aviv Regev

[Affiliations](#) | [Contributions](#) | [Corresponding authors](#)

Nature Protocols **8**, 1494–1512 (2013) | doi:10.1038/nprot.2013.084

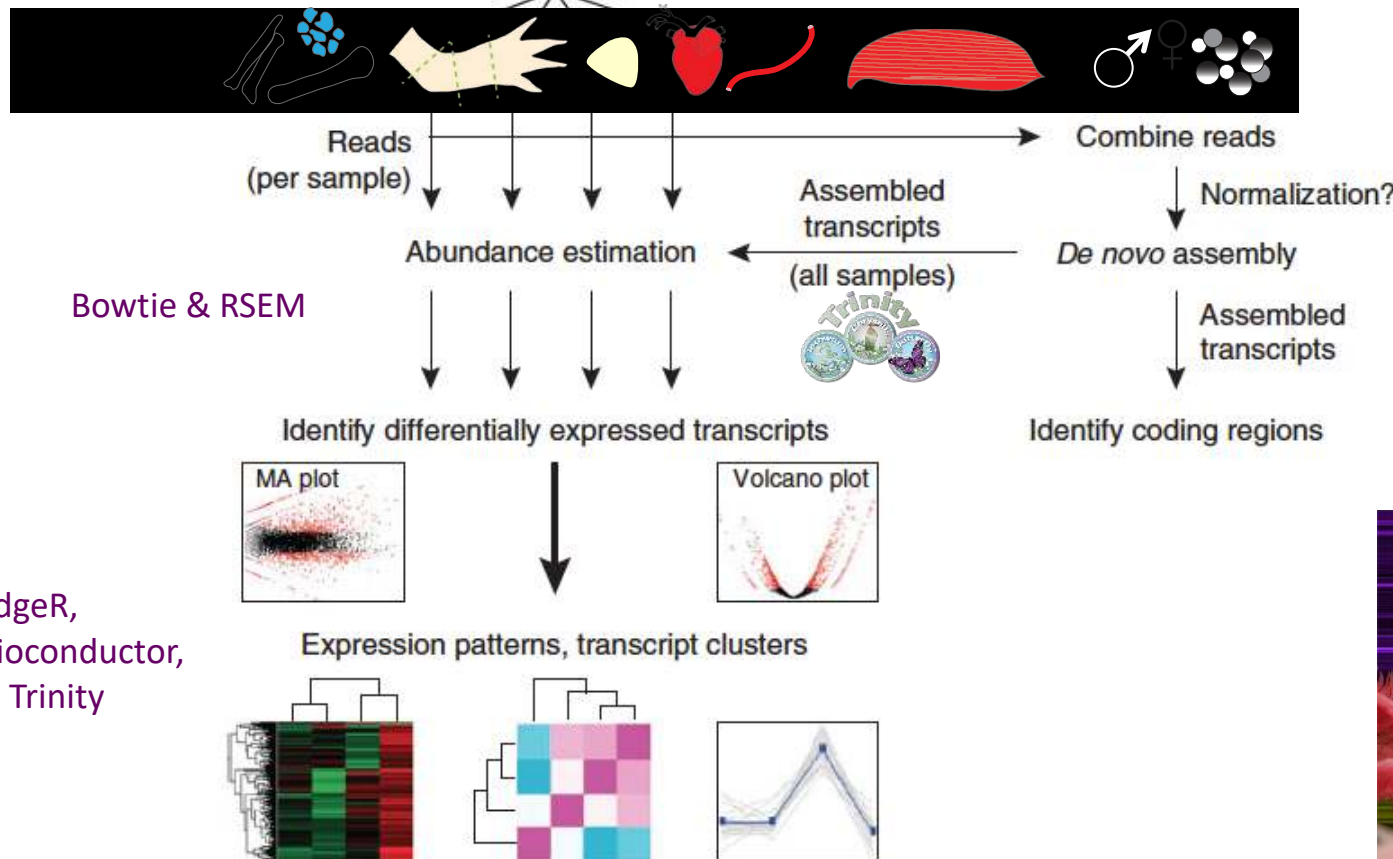
Published online 11 July 2013



Framework for De novo Transcriptome Assembly and Analysis



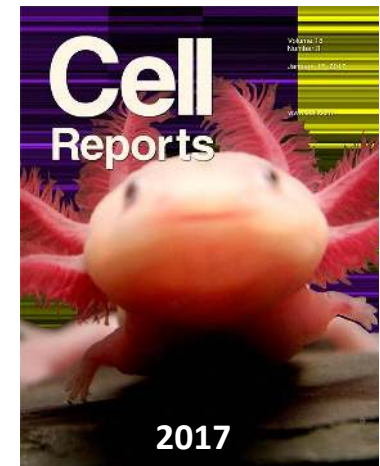
Applied to Axolotl for studies of limb regeneration



1.3 Billion
Total Reads

86 Million
Normalized Reads

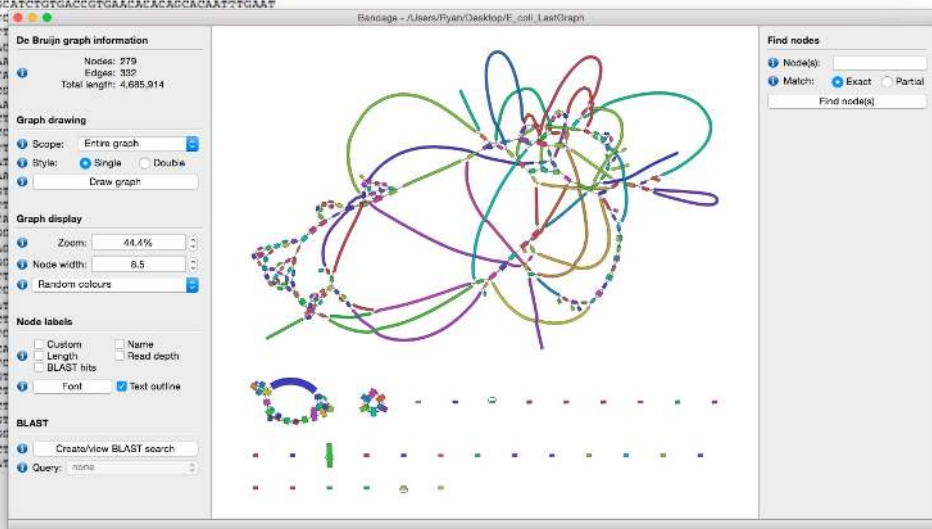
EdgeR,
Bioconductor,
& Trinity



```
>comp0 c0 seq1 len=5528 path=[1:0-3646 10775:3647-3775 3648:3776-5527]
```

Can visualize using Bandage

<https://rrwick.github.io/Bandage/>



IGV

← → ↻ www.broadinstitute.org/igv/ ☆ a



Integrative Genomics Viewer

- Home
- Downloads
- Documents
 - Hosted Genomes
 - FAQ
 - IGV User Guide
 - File Formats
 - Release Notes
 - Credits
- Contact

Search website

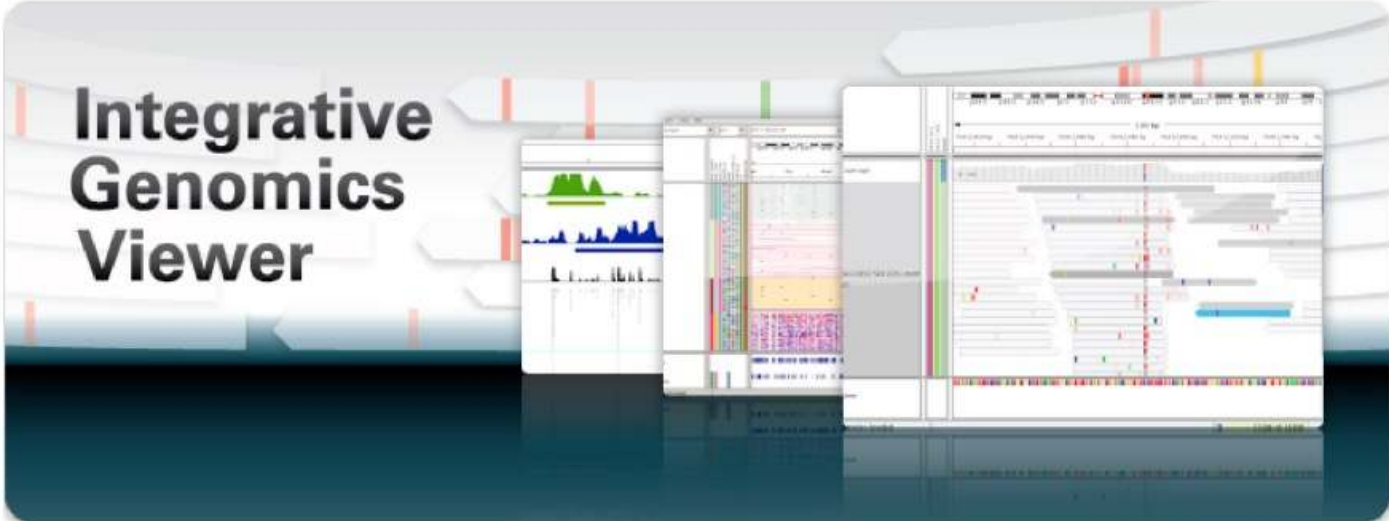
search

[Broad Home](#)
[Cancer Program](#)



© 2012 Broad Institute

Home



Integrative Genomics Viewer

What's New

 **July 3, 2012.** Soybean (Glycine max) and Rat (rn5) genomes have been updated.

April 20, 2012. IGV 2.1 has been released. See the [release notes](#) for more details.

April 19, 2012. See our new [IGV paper](#) in Briefings in Bioinformatics.

Citing IGV

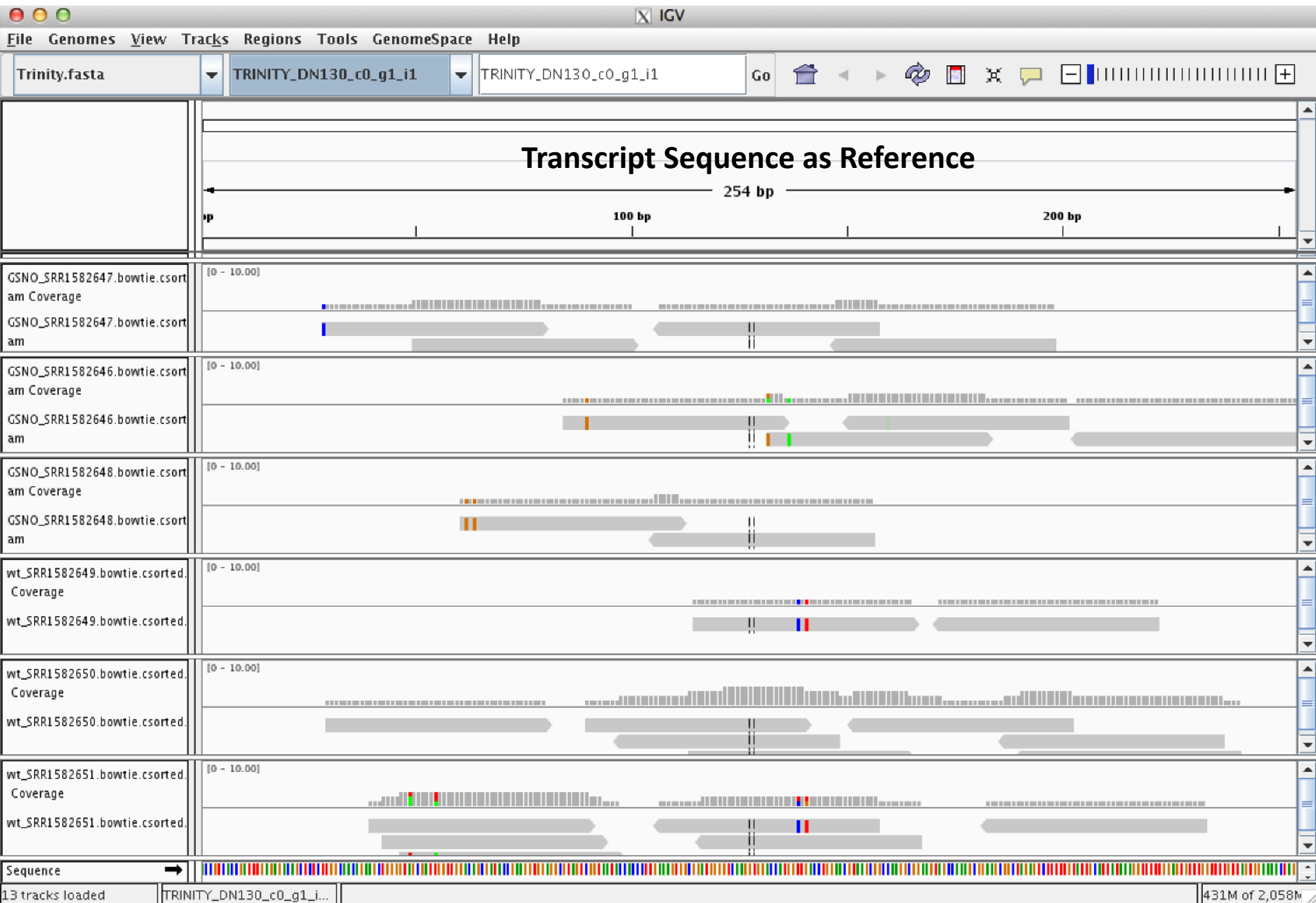
To cite your use of IGV in your publication:

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. [Integrative Genomics Viewer](#). *Nature Biotechnology* 29, 24–26 (2011), or

Helga Thorvaldsdottir, James T. Robinson, Jill P. Mesirov. [Integrative Genomics Viewer \(IGV\): high-performance genomics data visualization and exploration](#).

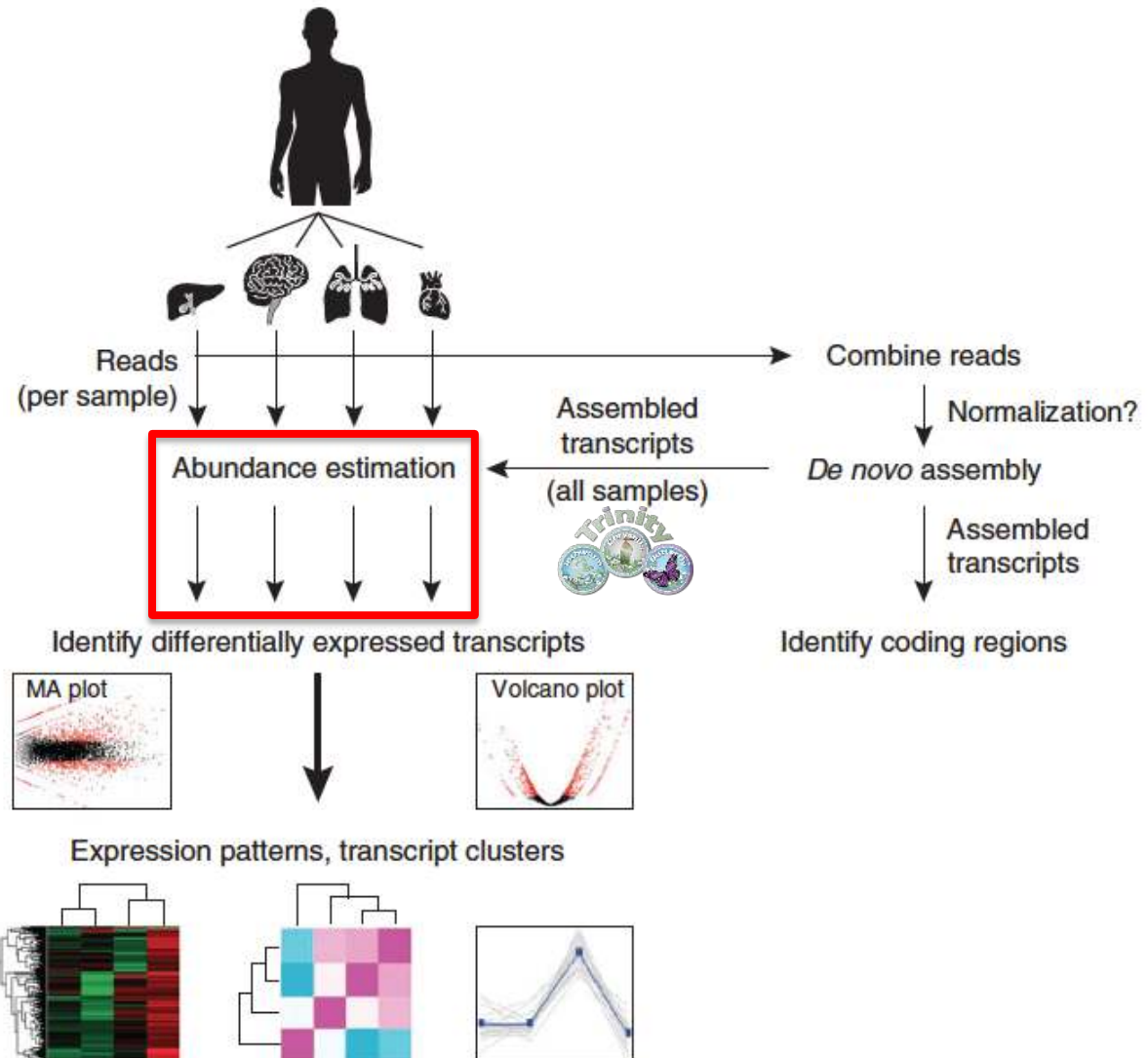
Overview

Can Examine Transcript Read Support Using IGV



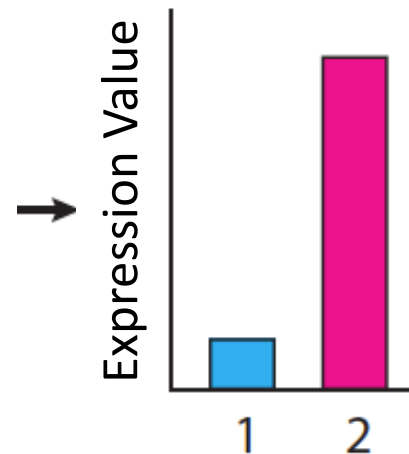
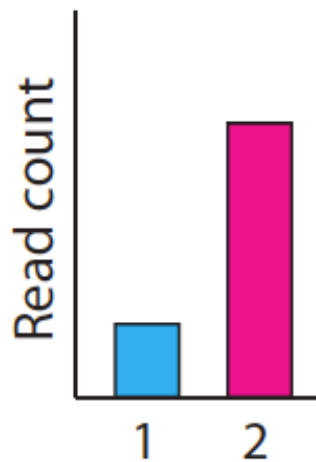
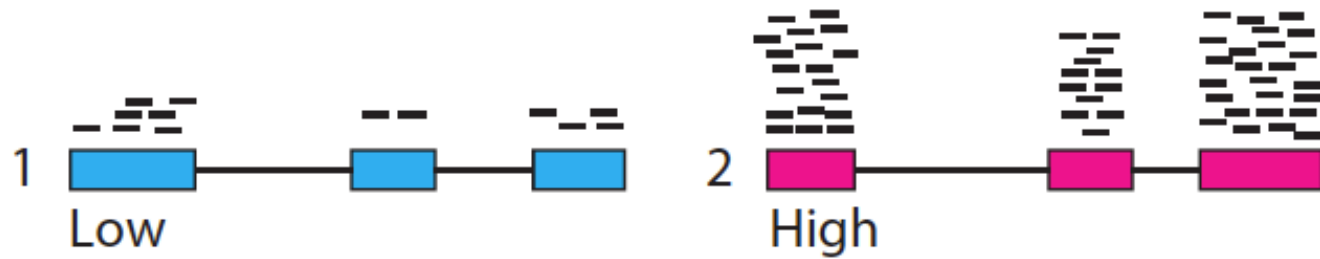


Part 4. Expression Quantification

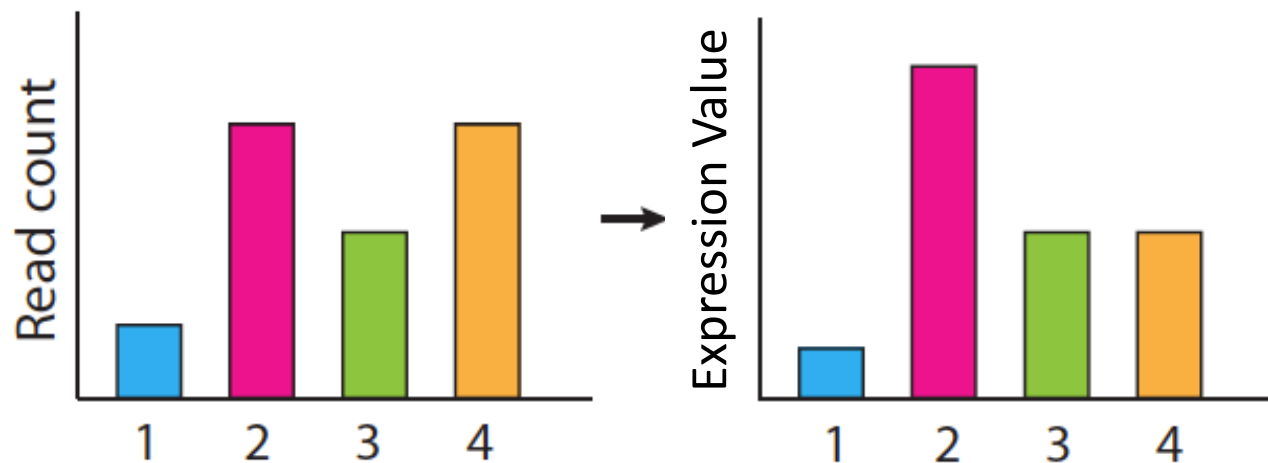
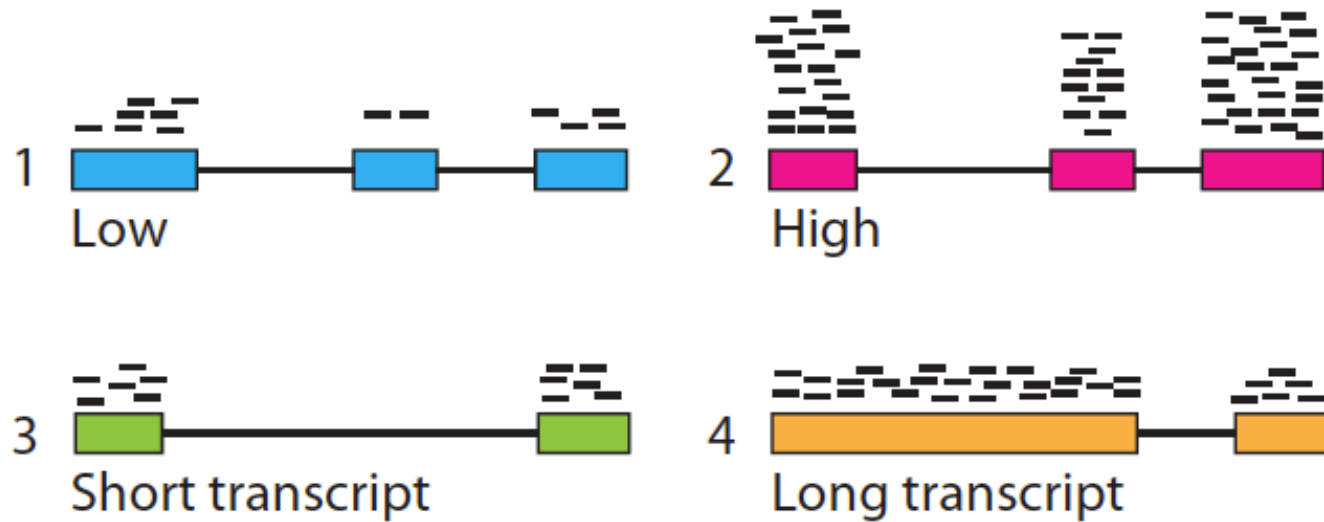


Bioconductor,
& Trinity

Calculating expression of genes and transcripts



Calculating expression of genes and transcripts



Normalized Expression Values

- Transcript-mapped read counts are normalized for both length of the transcript and total depth of sequencing.
- Reported as: Number of RNA-Seq **F**ragments
Per **K**ilobase of transcript
per total **M**illion fragments mapped
FPKM

RPKM (reads per kb per M) used with Single-end RNA-Seq reads
FPKM used with Paired-end RNA-Seq reads.

Transcripts per Million (TPM)

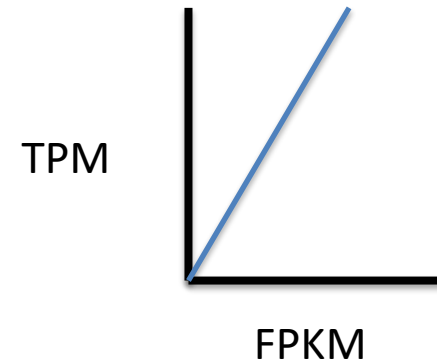
$$TPM_i = \frac{FPKM_i}{\sum_j FPKM} * 1e6$$

Preferred metric for measuring expression

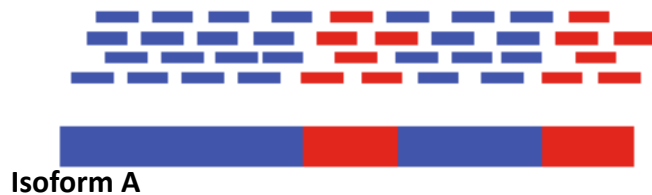
- Better reflects transcript concentration in the sample.
- Nicely sums to 1 million

Linear relationship between TPM and FPKM values.

Both are valid metrics, but best to be consistent.



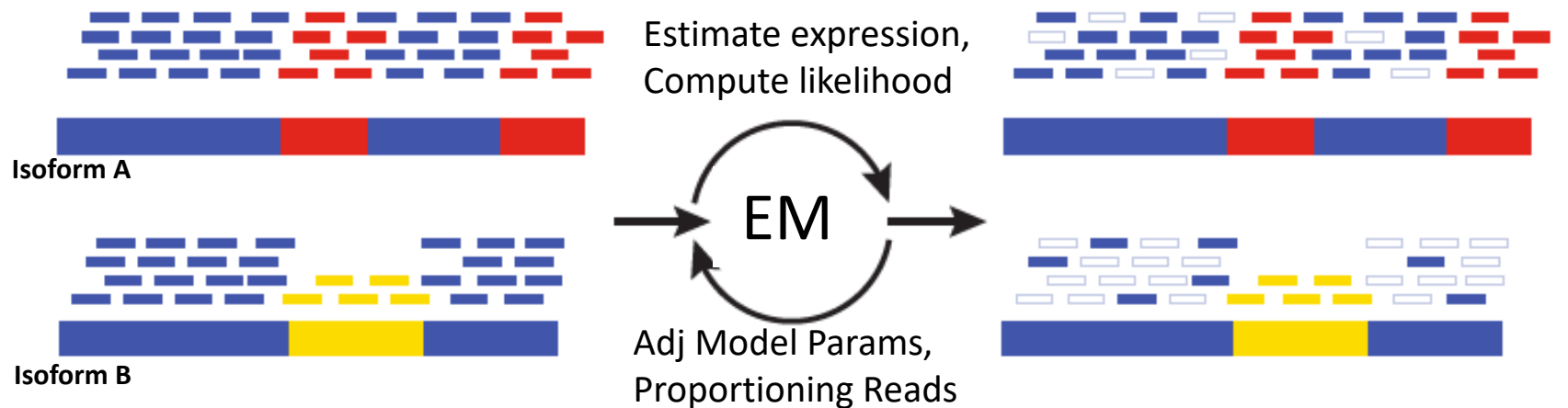
Multiply-mapped Reads Confound Abundance Estimation



Blue = multiply-mapped reads

Red, Yellow = uniquely-mapped reads

Multiply-mapped Reads Confound Abundance Estimation

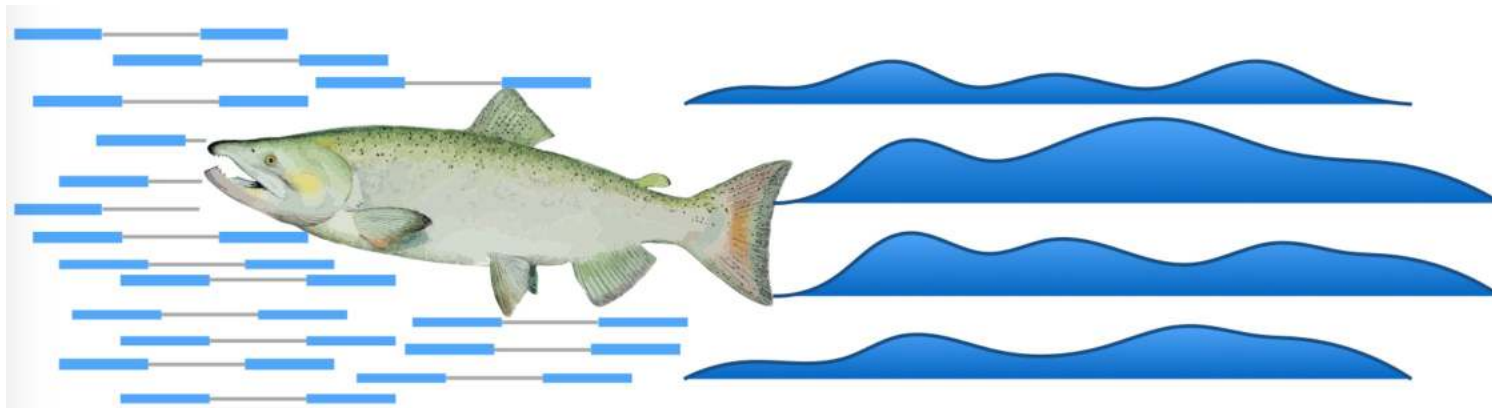


Blue = multiply-mapped reads
Red, Yellow = uniquely-mapped reads

Use Expectation Maximization (EM) to find the most likely assignment of reads to transcripts.

Performed by:

- RSEM (genome-free)
- Kallisto, Salmon (alignment-free)



Salmon —Don't count . . . quantify!

Uses a suffix array
instead of the
de Bruijn graph

 nature|methods

 Altmetric: 210 Citations: 42 [More detail >>](#)

Brief Communication

Salmon provides fast and bias-aware quantification of transcript expression

Rob Patro , Geet Duggal, Michael I Love, Rafael A Irizarry & Carl Kingsford 

Nature Methods **14**, 417–419 (2017)
doi:10.1038/nmeth.4197
[Download Citation](#)

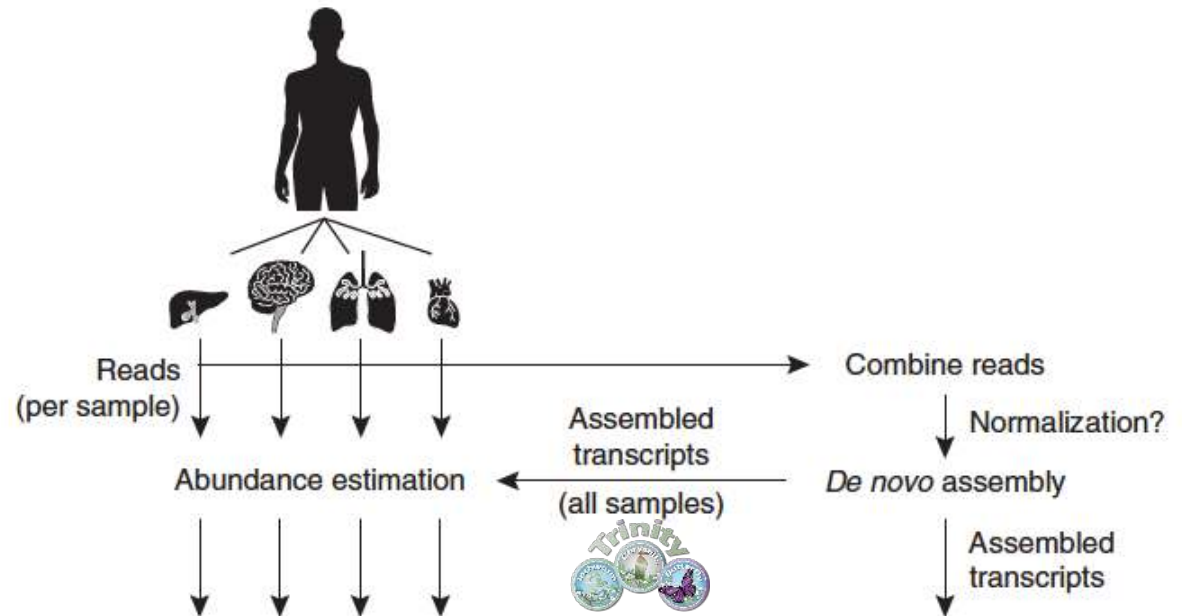
Received: 29 August 2016
Accepted: 22 January 2017
Published online: 06 March 2017

<https://combine-lab.github.io/salmon/>

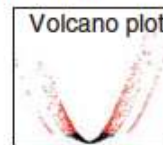
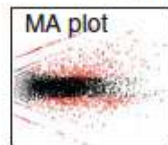


Part 5. Differential Expression

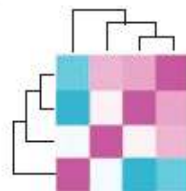
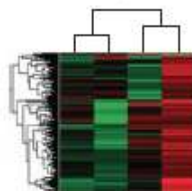
More specifics in
Rachel Steward's
afternoon workshop



Identify differentially expressed transcripts



Expression patterns, transcript clusters



Bioconductor,
& Trinity

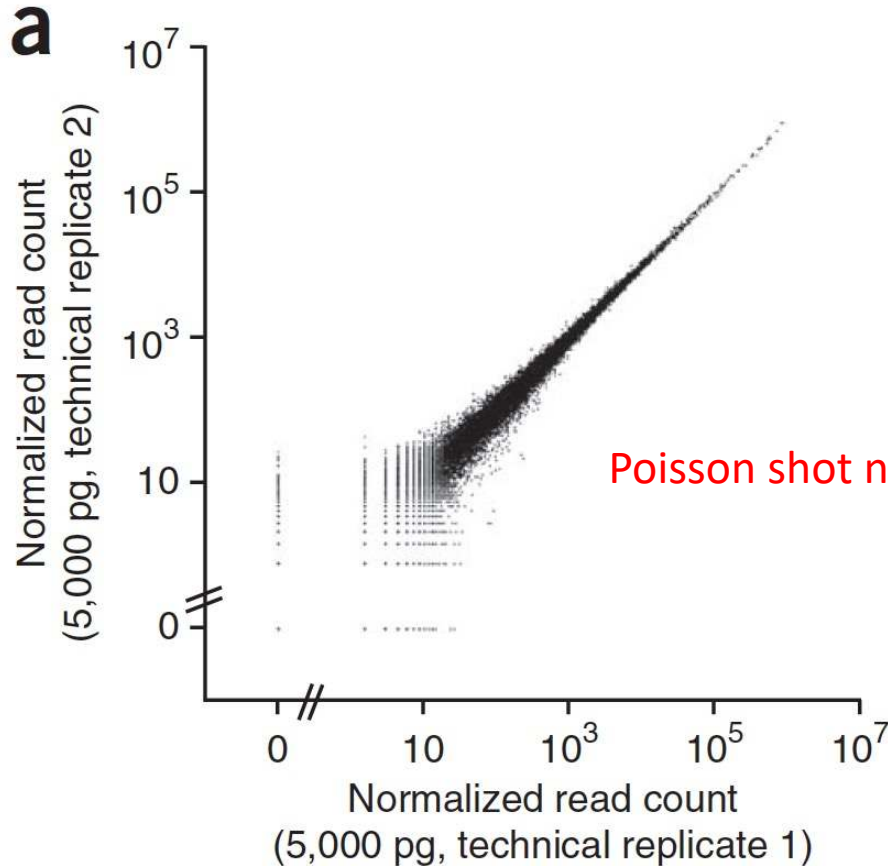
Differential Expression Analysis Involves

- Counting reads mapped to features
- Statistical significance testing

Beware of small counts leading to notable fold changes

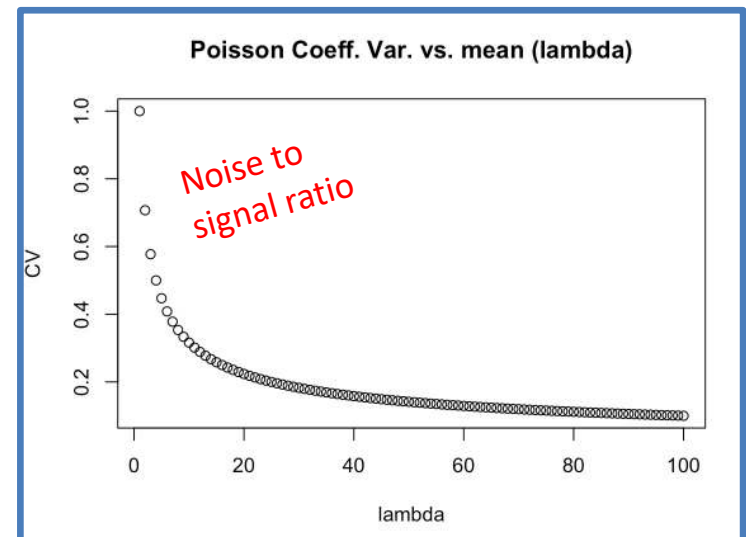
	Sample_A	Sample_B	Fold_Change	Significant?
Gene A	1	2	2-fold	No
Gene B	100	200	2-fold	Yes

Variation Observed Between Technical Replicates



Variation observed is well described by models of random sampling (Poisson Distribution)

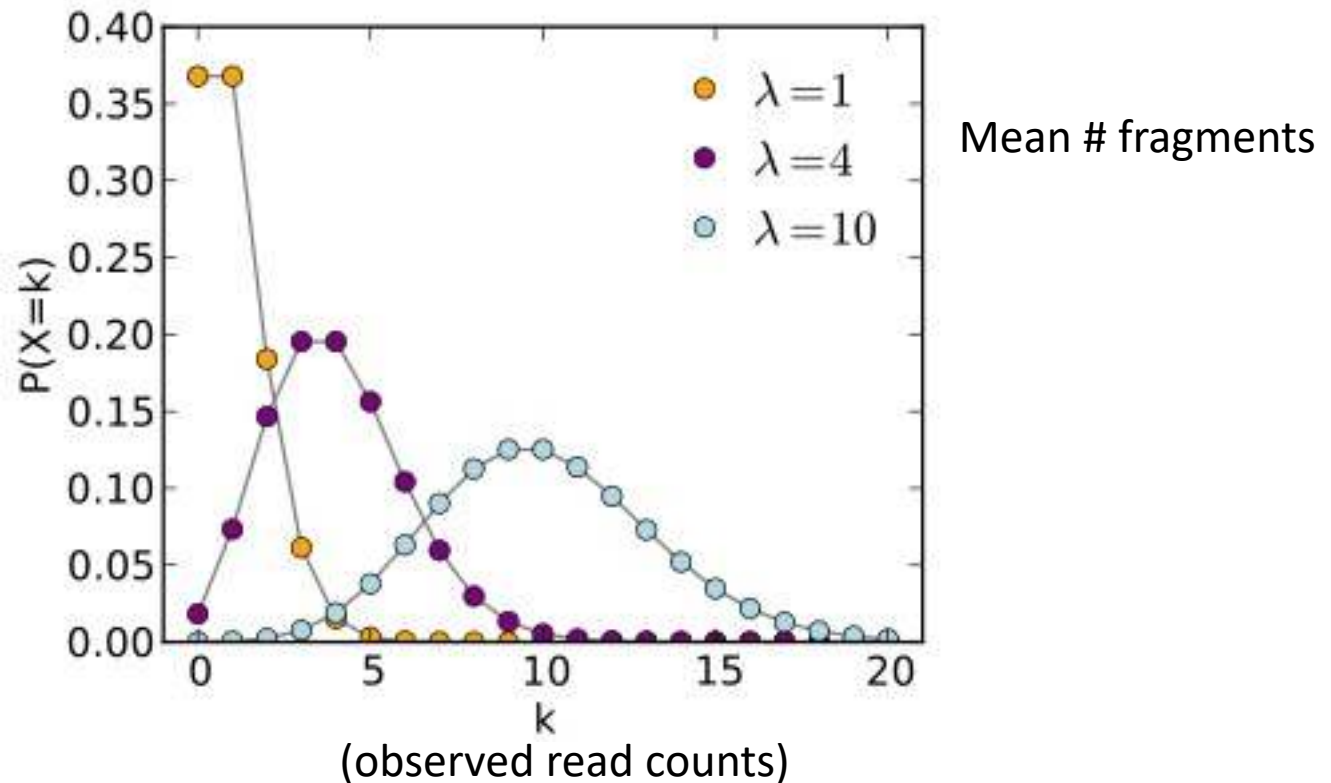
Poisson shot noise is high for small counts.



* plot from Brennecke, et al. Nature Methods, 2013

Observed RNA-Seq Counts Result from Random Sampling of the Population of Reads

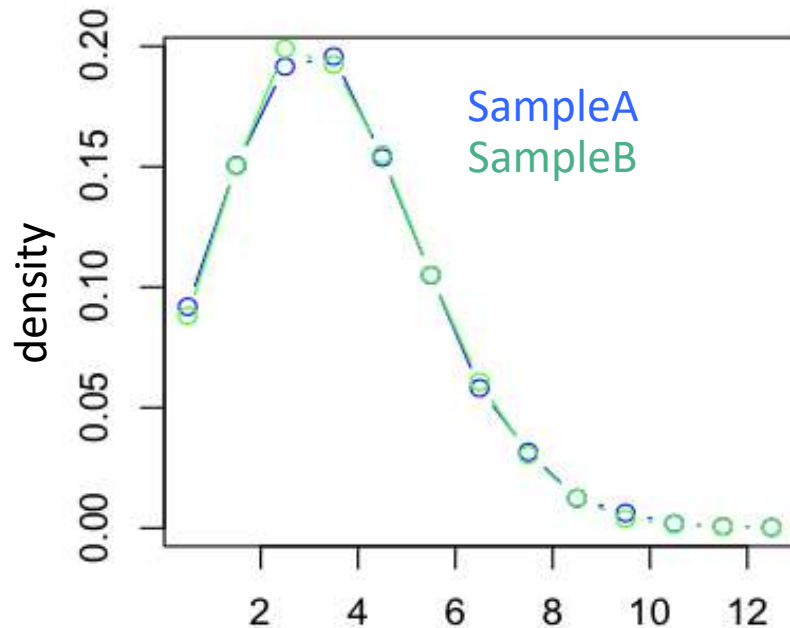
Technical variation in RNA-Seq counts per feature is well modeled by the Poisson distribution



Example: One gene*not* differentially expressed

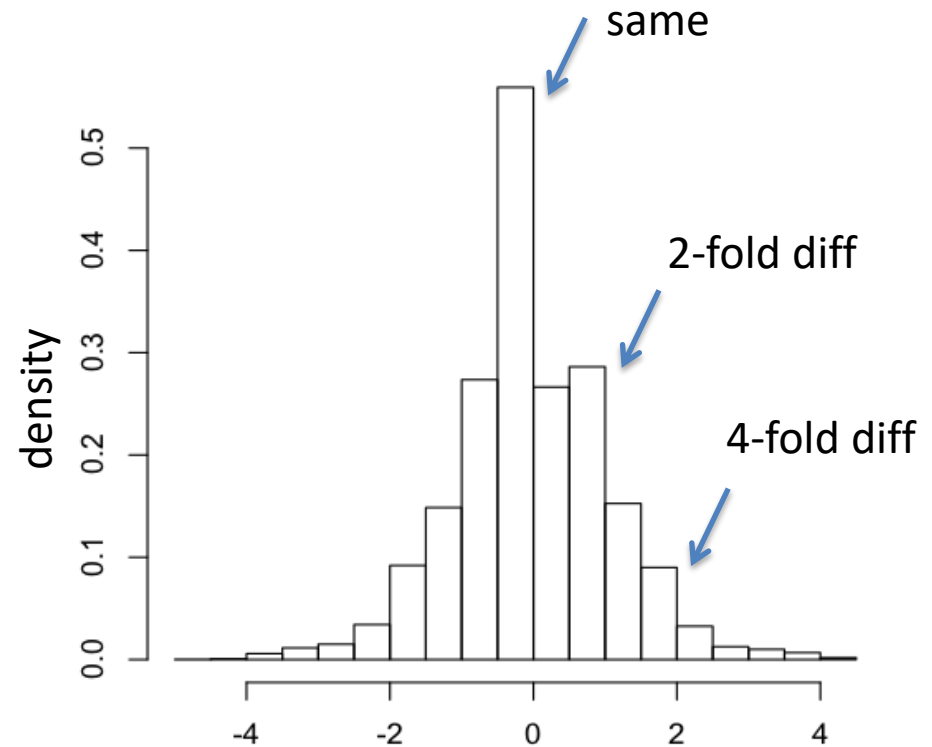
Example: $\text{SampleA}(\text{gene}) = \text{SampleB}(\text{gene}) = 4$ reads

Distribution of observed counts for single gene
(under Poisson model)



(k) number of reads observed
for a single gene

Dist. of $\log_2(\text{fold change})$ values



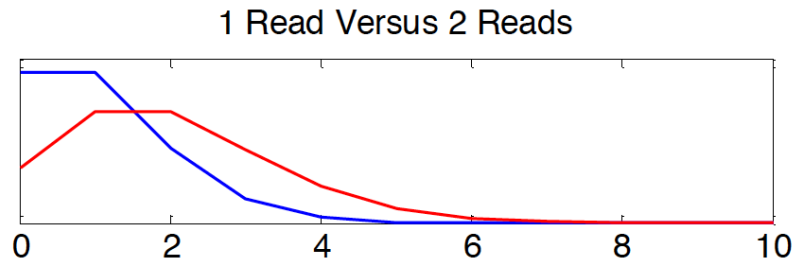
$x = \log_2(\text{SampleA}/\text{SampleB})$

Sequencing Depth Matters

Poisson distributions for counts based on **2-fold** expression differences

Shallow sequencing

$P(x=k)$



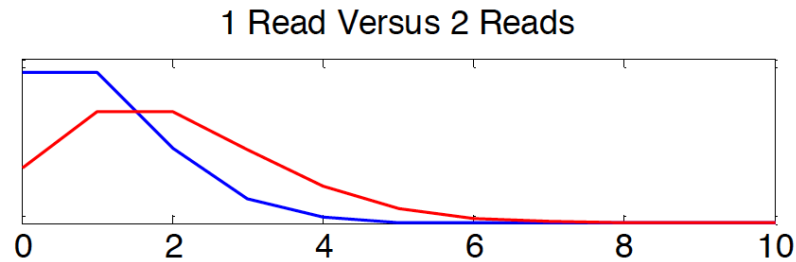
No confidence in 2-fold difference. Likely observed by chance.

Sequencing Depth Matters

Poisson distributions for counts based on **2-fold** expression differences

Shallow sequencing

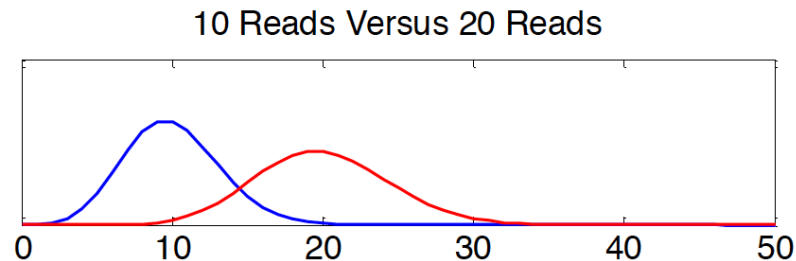
$P(x=k)$



No confidence in 2-fold difference. Likely observed by chance.

Deeper sequencing

$P(x=k)$

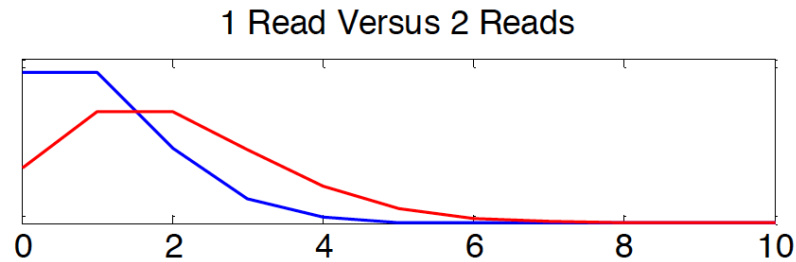


Sequencing Depth Matters

Poisson distributions for counts based on **2-fold** expression differences

Shallow sequencing

$P(x=k)$



No confidence in 2-fold difference. Likely observed by chance.

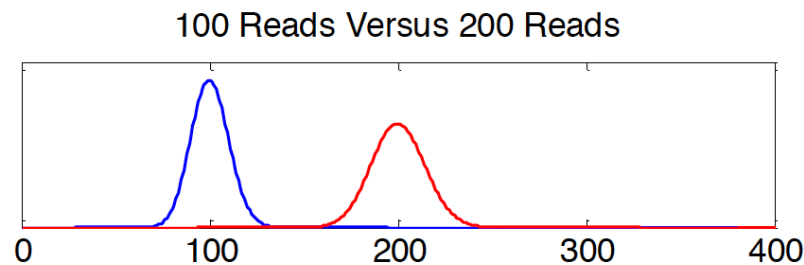
Deeper sequencing

$P(x=k)$



Deepest sequencing

$P(x=k)$



High confidence in 2-fold difference. Unlikely observed by chance.

Observed Read Count (k)

Greater Depth = More Statistical Power

Example: Single gene, reads sampled at different sequencing depths

Reads per sample	Sample A Number of reads	Sample B Number of reads	P-value (Fishers Exact Test)
100,000	1	2	1
1,000,000	10	20	0.099
10,000,000	100	200	8.0e-09

Technical vs. Biological Replicates

RNA-Seq Technical replicates aren't essential

(Technical variation is well-modeled by the Poisson distribution)

“We find that the Illumina sequencing data are highly replicable, with relatively little technical variation, and thus, for many purposes, it may suffice **to sequence each mRNA sample only once**” *Marioni et al., Genome Research, 2008*

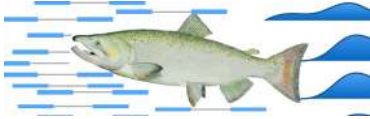
However, biological replicates ***ARE*** essential

$\text{total_variance} = \text{technical_variance} + \text{biological_variance}$

(Total variance well-modeled by negative binomial distribution)

“... **at least six biological replicates should be used**, rising to at least 12 when it is important to identify SDE genes for all fold changes.” *Schurch et al., RNA, 2016*

DE analysis requires a counts matrix



Transcript_ID	Sample Type A, 3 Bio replicates			Sample Type B, 3 Bio replicates		
TR24 c0_g1_i1	90.00	67.00	85.00	36.00	35.00	34.00
TR2779 c0_g1_i1	186.00	137.00	217.00	147.00	186.00	197.00
TR127 c1_g1_i1	9.00	23.00	16.00	2.00	0.00	1.00
TR2107 c1_g1_i1	59.00	65.00	47.00	6.00	6.00	7.00
TR2011 c5_g1_i1	11.00	4.00	4.00	8.00	5.00	7.00
TR4163 c0_g1_i1	368.00	422.00	425.00	172.00	216.00	210.00
TR5055 c0_g2_i1	36.00	17.00	27.00	4.00	7.00	3.00
TR1449 c0_g1_i1	196.00	230.00	207.00	66.00	113.00	91.00
TR1982 c2_g1_i1	7.00	7.00	6.00	4.00	3.00	8.00
TR1859 c3_g1_i1	0.00	0.00	1.00	0.00	0.00	0.00
TR1492 c0_g1_i2	1895.00	1906.00	1921.00	1104.00	1263.00	1319.00
TR1122 c0_g1_i1	2.00	3.00	0.00	3.00	0.00	0.00
TR2278 c0_g1_i1	497.00	610.00	598.00	333.00	406.00	413.00
TR4084 c0_g1_i1	95.00	148.00	86.00	77.00	111.00	127.00
TR4761 c0_g1_i1	2089.00	1746.00	1875.00	155.00	174.00	165.00
TR3638 c0_g1_i1	647.00	676.00	712.00	117.00	184.00	174.00
TR2090 c0_g1_i1	0.00	0.00	0.00	22.00	0.00	0.02
TR3854 c0_g1_i1	1878.00	1734.00	1864.00	1775.00	2173.00	2151.00
TR131 c0_g1_i1	32.00	28.00	31.00	1001.00	1233.00	1208.00
TR5075 c0_g1_i1	13.00	22.00	21.00	6.00	8.00	10.00
TR2182 c3_g2_i6	1.44	2.70	3.84	3.35	0.00	0.00
TR3788 c0_g1_i1	17.00	30.00	22.00	91.00	132.00	125.00
TR4859 c0_g1_i1	6.00	12.00	8.00	4.00	1.00	3.00
TR2487 c0_g1_i1	386.00	383.00	424.00	689.00	866.00	806.00
TR2122 c0_g2_i2	145.00	135.00	136.00	155.00	157.00	201.00
TR4277 c0_g1_i1	4466.00	4701.00	4284.00	118.00	134.00	164.00
TR4669 c0_g2_i1	0.00	0.00	0.00	209.00	0.00	217.50
TR3091 c0_g1_i1	22.00	17.00	19.00	250.00	308.00	284.00

Typical output from DE analysis

Transcript_id	logFC	logCPM	PValue	FDR
TRINITY_DN876_c0_g1_i1	-7.15049572793027	10.6197708379285	0	0
TRINITY_DN6470_c0_g1_i1	-7.26777912190146	7.03987604865422	1.687485656951e-287	6.46813252309319e-284
TRINITY_DN5186_c0_g1_i1	-7.85623682454322	9.18570464327063	1.17049180235068e-278	2.99099671894011e-275
TRINITY_DN768_c0_g1_i1	7.72884741150304	9.7514619195169	4.32504881419265e-272	8.28895605240022e-269
TRINITY_DN70_c0_g1_i1	-12.7646078189688	7.86482982471445	3.92853491279431e-253	6.02322972829624e-250
TRINITY_DN1587_c0_g1_i1	-5.89392061881667	9.07366563894607	6.32919557933429e-243	8.08660221852944e-240
TRINITY_DN3236_c0_g1_i1	-7.27029815068473	8.02209568234202	3.64955175271959e-235	3.99678053376405e-232
TRINITY_DN4631_c0_g1_i1	-7.45310693639574	6.91664918183241	4.30540921272851e-229	4.1256583780971e-226
TRINITY_DN5082_c0_g5_i1	-5.33154406167545	10.6977538760467	2.74243356676259e-225	2.33594396920022e-222
TRINITY_DN1789_c0_g3_i1	10.2032564835076	7.32607652700285	1.44273728647186e-213	1.10600240380933e-210
TRINITY_DN4204_c0_g1_i1	4.81030233739325	9.88844409410644	9.27180216086162e-205	6.46160321501501e-202
TRINITY_DN799_c0_g1_i1	-4.22044475626154	6.9937398638711	1.24746518421083e-197	7.96922341846683e-195
TRINITY_DN196_c0_g2_i1	4.60597918494257	9.86878463857276	1.9819997623131e-192	1.16877001368402e-189
TRINITY_DN5041_c0_g1_i1	-4.27126549355785	9.70894399883	1.8930437900069e-185	1.03657669244235e-182
TRINITY_DN1619_c0_g1_i1	-4.47156415953777	9.22535948721718	1.76766063029526e-181	9.03392426122899e-179
TRINITY_DN899_c0_g1_i1	-4.90914328409143	7.93768691394594	1.11054513767547e-180	5.32089939088761e-178
TRINITY_DN324_c0_g2_i1	4.87160837667488	6.84850312231775	2.20092562166991e-179	9.92487989160089e-177
TRINITY_DN3241_c0_g1_i1	-4.77760618069256	7.94111259715689	1.60585457735621e-173	6.83915621667372e-171
TRINITY_DN4379_c0_g1_i1	3.85133572453294	7.23712813663389	3.48140532848425e-164	1.4046554341137e-161
TRINITY_DN1919_c0_g1_i1	4.05998814332136	6.95937301668582	1.8588621194715e-161	7.12501850393425e-159
TRINITY_DN2504_c0_g1_i1	-6.92417817059644	6.20370039359785	2.42022459856956e-160	8.83497227268296e-158

...



Up vs. Down regulated



Avg. expression level



Significance

Tools for DE analysis with RNA-Seq



edgeR

ShrinkSeq

DESeq

baySeq

Vsf

Limma/Voom

mmdiff

cuffdiff

ROTS

TSPM

DESeq2

EBSeq

NBPSeq

SAMseq

NoiSeq

Sleuth

*(italicized not in R/Bioconductor
but stand-alone)*

See: <http://www.biomedcentral.com/1471-2105/14/91>

A comparison of methods for differential expression analysis of RNA-seq data

Soneson & Delorenzi, 2013



Part 6. Latest advancements in long read isoform sequencing

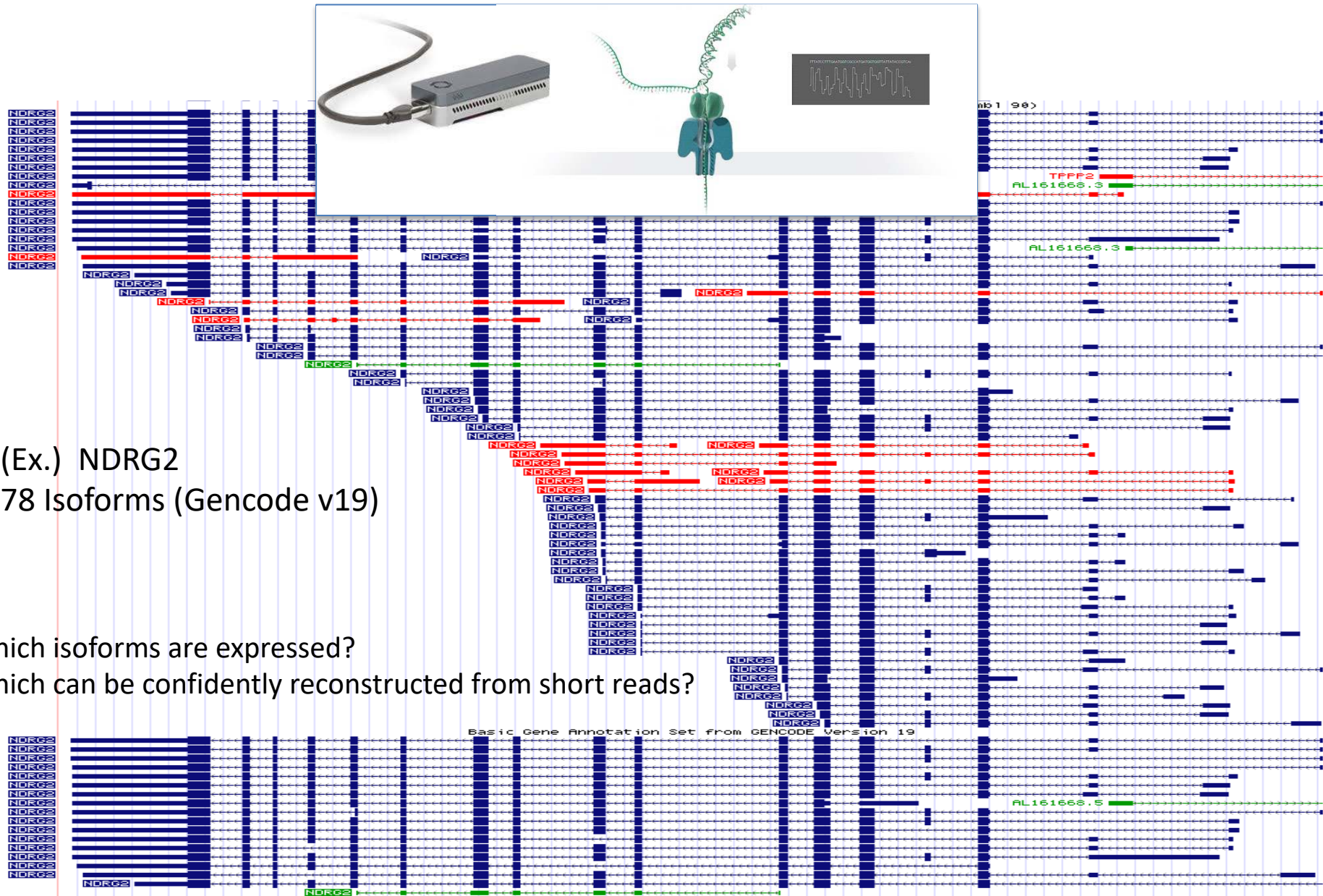
Some transcripts can be challenging to reconstruct from short reads

- Complex alternative splicing (many isoforms)
- Very long RNAs (ex. Titin – up to 36 kb)
- Transcripts containing repetitive sequences

Transcript Reconstruction or Expression Analysis can be Quite Difficult at Complex Loci



Too complex... don't guess from short reads, use long reads.



Long Read Isoform Sequencing via PacBio MAS-Iso-Seq (Kinnex)

Editorial | Published: 12 January 2023

Method of the Year 2022: long-read sequencing

The variables on RNA molecules: concert or cacophony? Answers in long-read sequencing

Inflection point for LR transcriptomics



Aziz Al'Khafaji

PacBio
Revio
2023

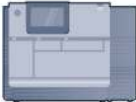


MAS-seq
(commercially
Kinnex)



40-120 million
cDNA reads

8M
<1%

Long read					
PacBio SMRT sequencing 2011	ONT MinION 2015	PacBio Sequel 2015	ONT GridION 2017	ONT PromethION 2018	PacBio Sequel II 2021
					
Throughput : 75k Error rate : 10%	1-10 million 5-10%	500k 10%	10-30 million 5-10%	30-150 million 5-10%	4M <1%

Info on error rates for long reads – impressive!!

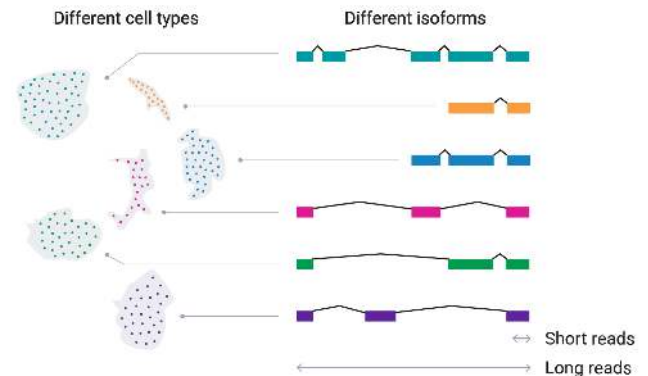
<https://nanoporetech.com/accuracy>

<https://www.pacb.com/technology/hifi-sequencing/>

99% 99.9%

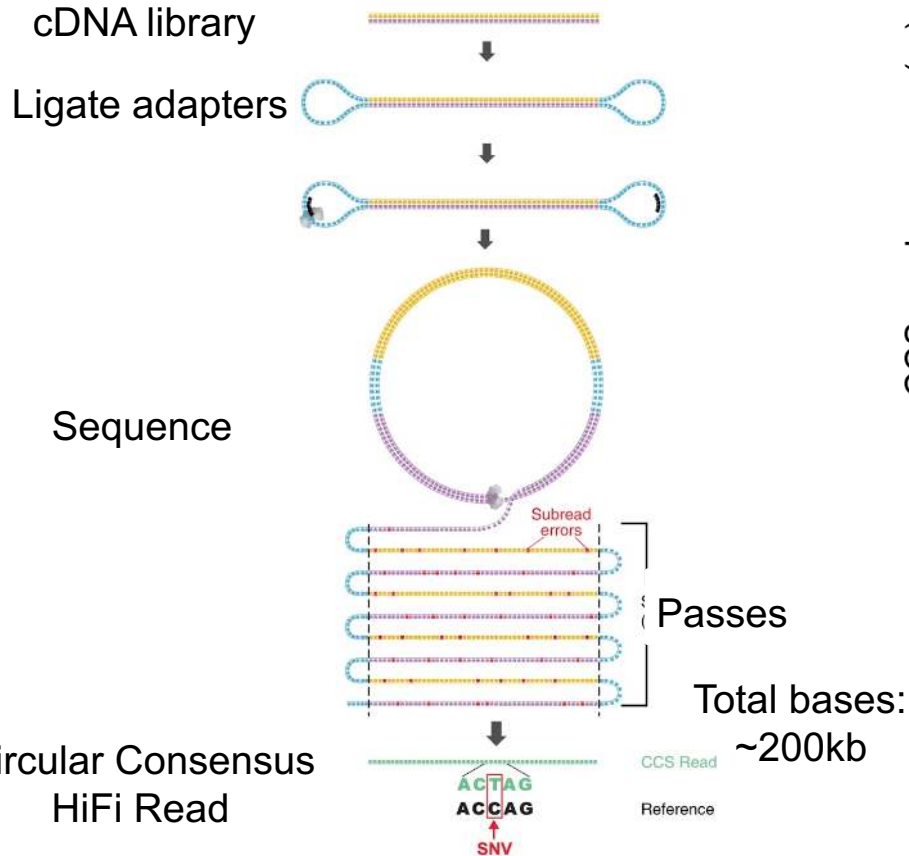
Q20 Q30

Long reads for Single Cell Transcriptomes!!

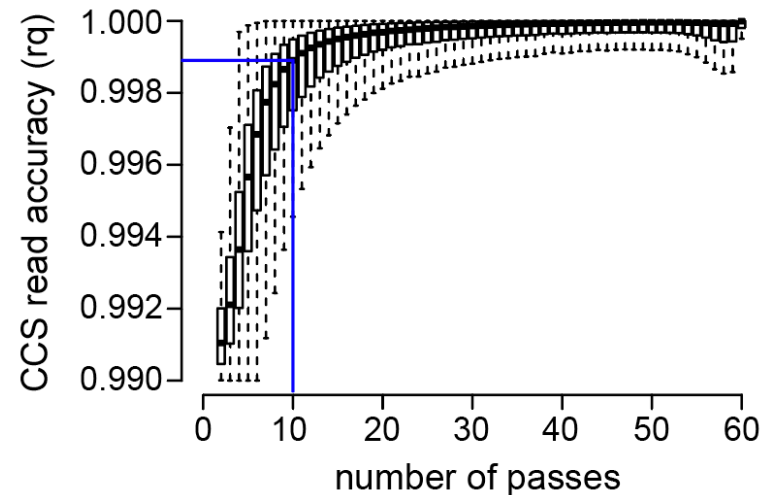


Standard isoform sequencing is inefficient on the PacBio platform

PacBio HiFi Sequencing



CCS read accuracy ~ # passes

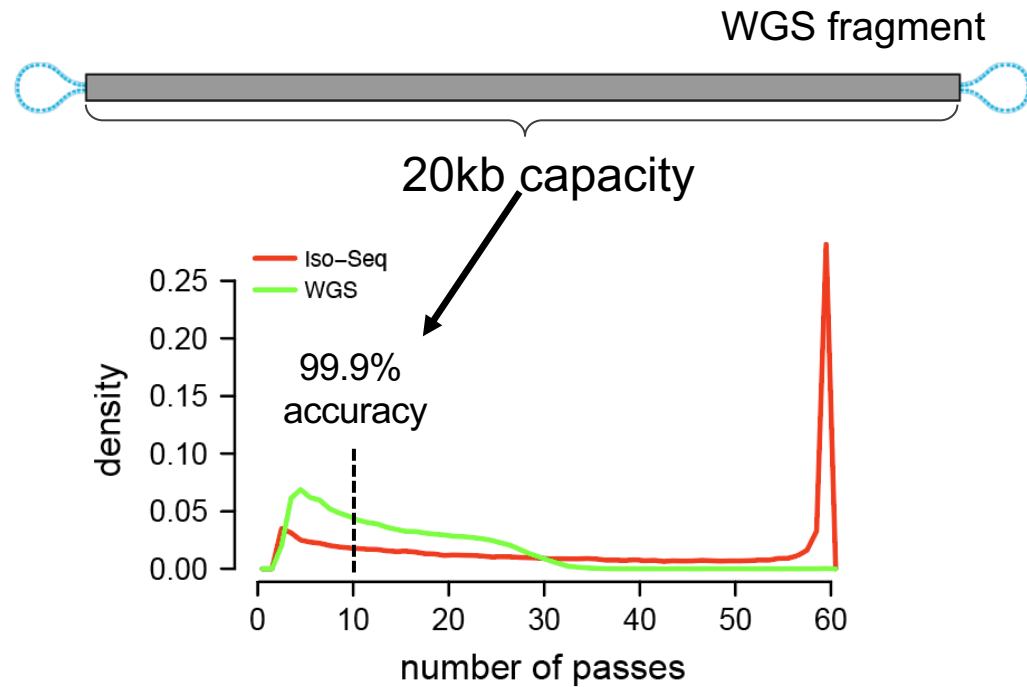
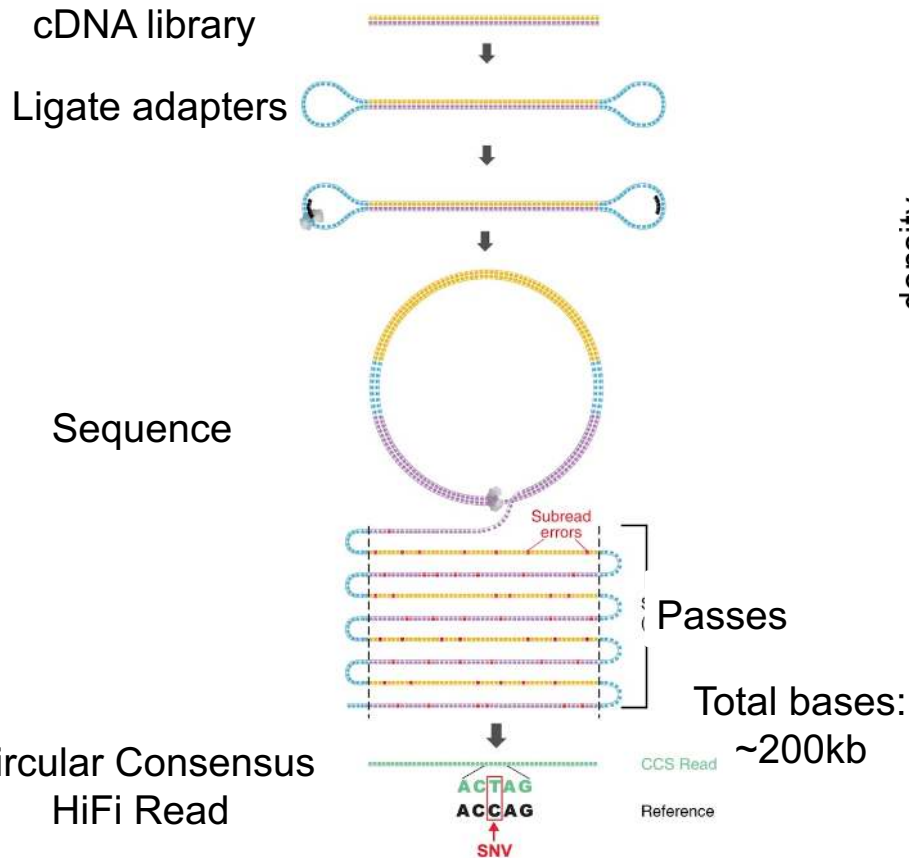


Base calling accuracy increases with the number of consensus reads.
~Q30 (99.9%) @ 10 passes.

200kb total = 20kb / pass

Standard isoform sequencing is inefficient on the PacBio platform

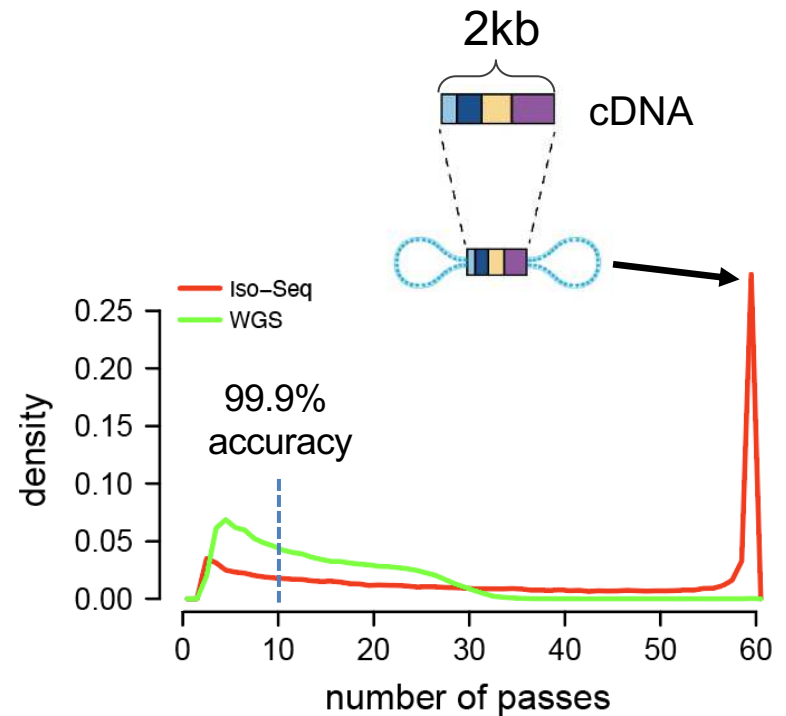
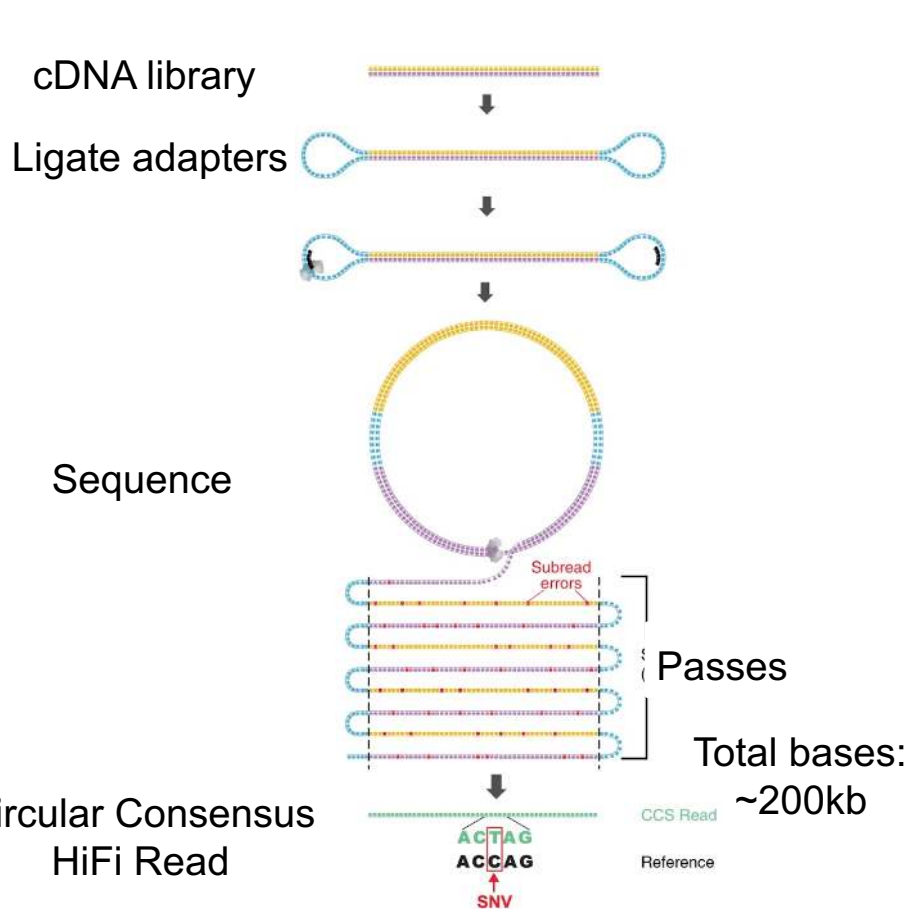
PacBio HiFi Sequencing



HiFi for WGS involves 20kb segments

Standard isoform sequencing is inefficient on the PacBio platform

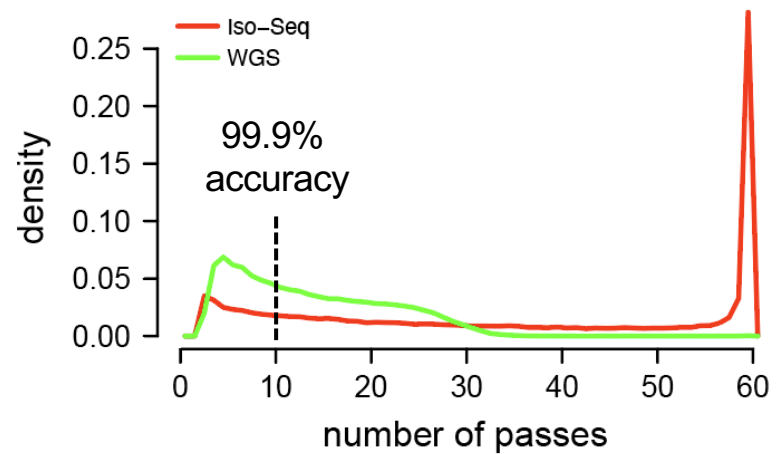
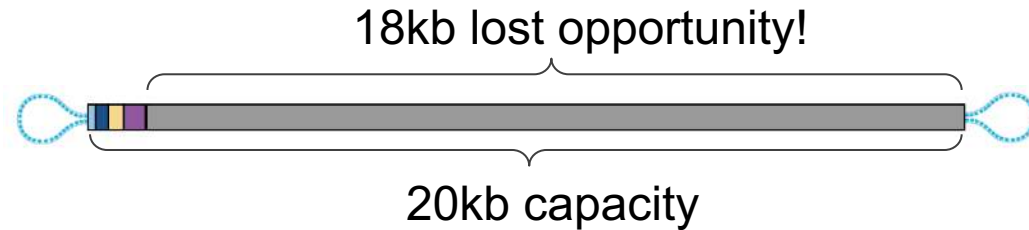
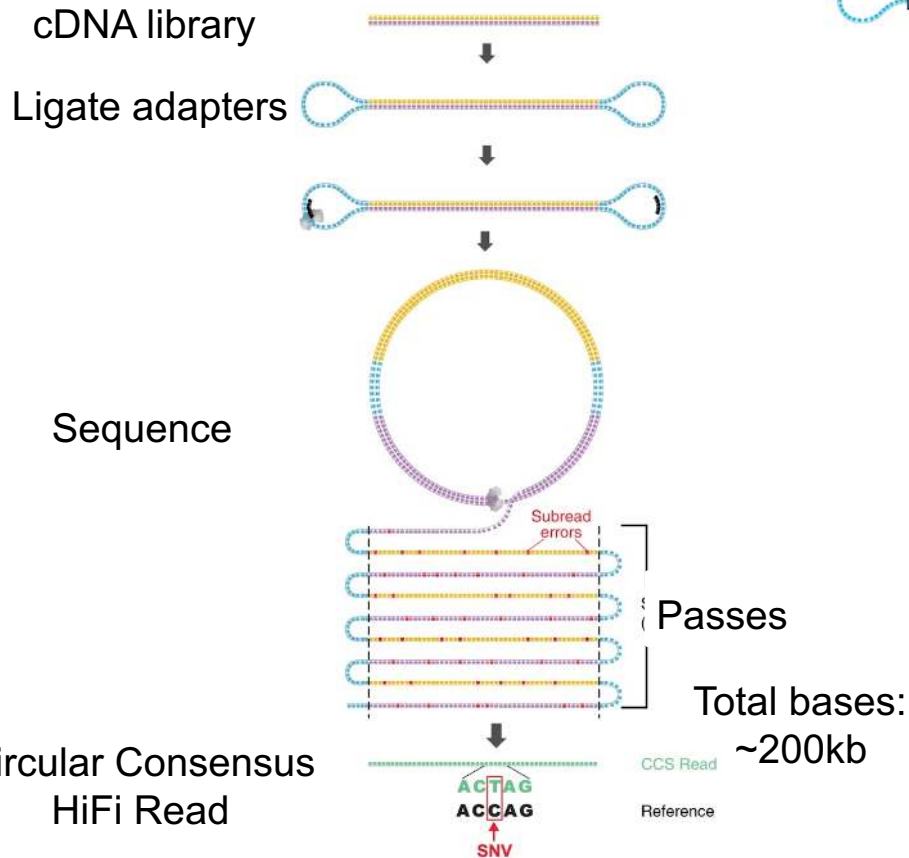
PacBio HiFi Sequencing



Most transcripts are <5kb and get >60 passes. Wasted sequencing potential!

Standard isoform sequencing is inefficient on the PacBio platform

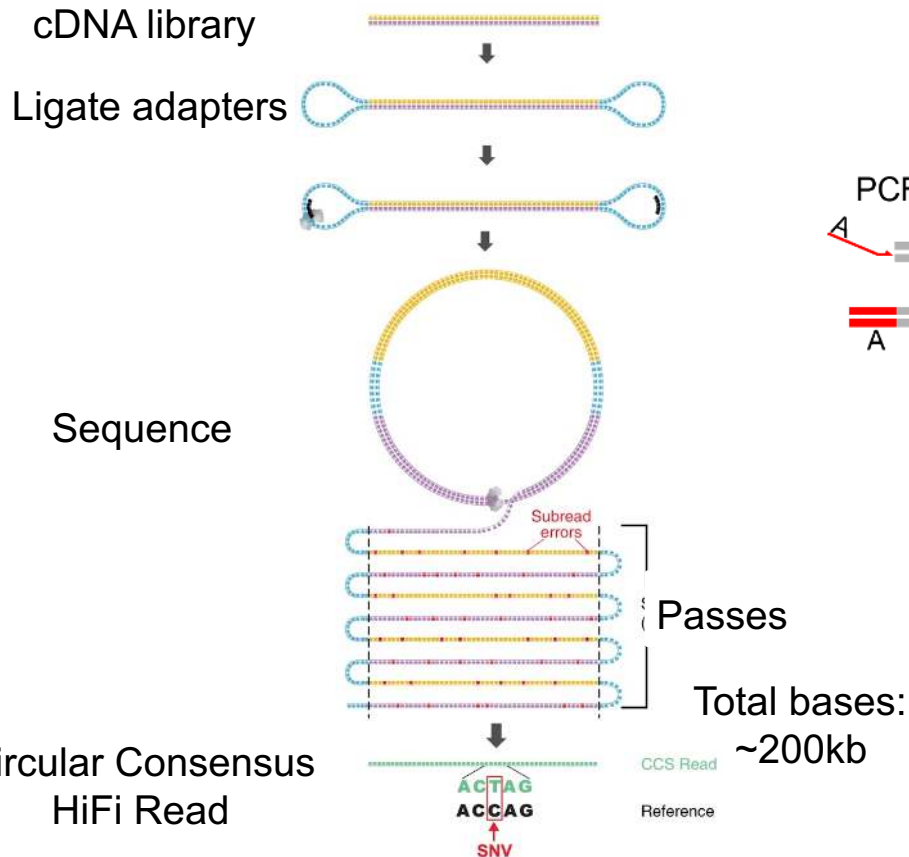
PacBio HiFi Sequencing



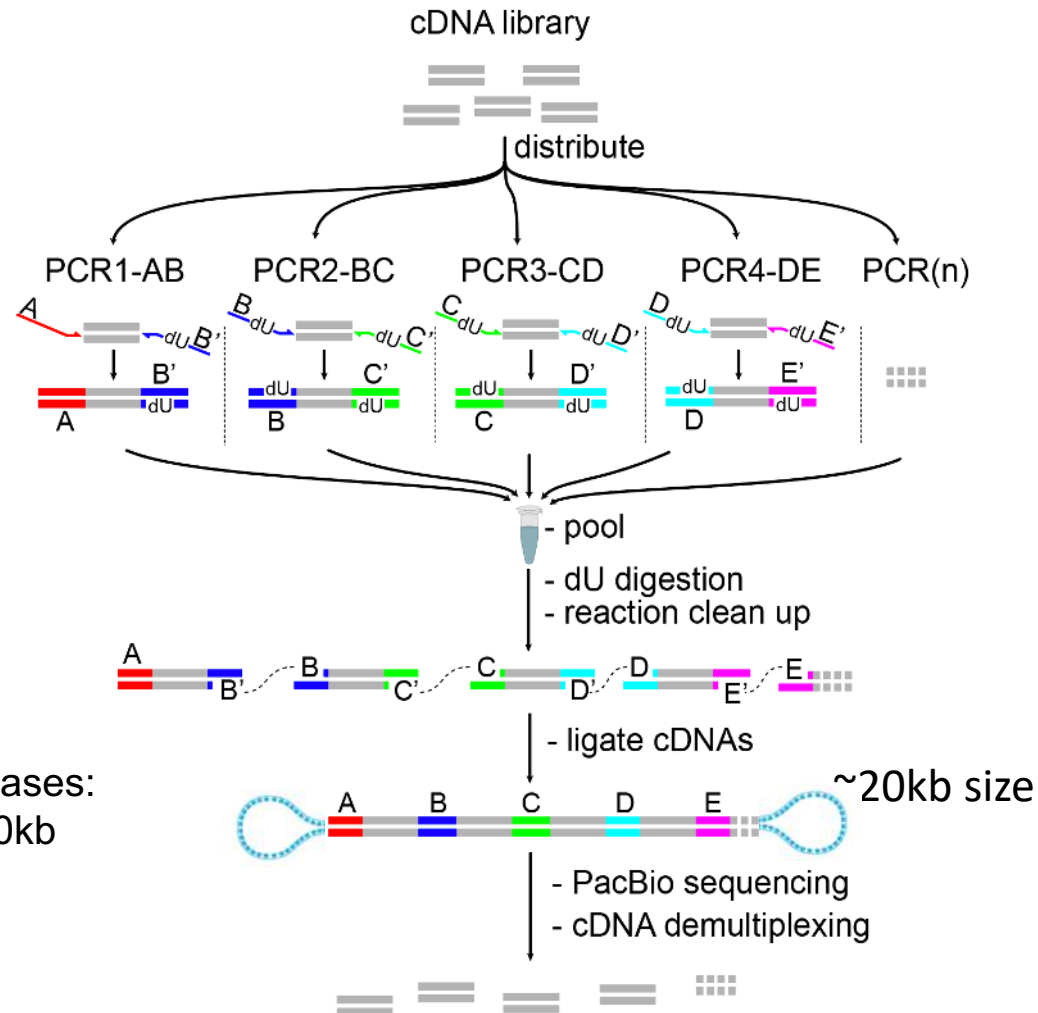
Of the 20kb segment, RNAs only use ~2kb

Standard isoform sequencing is inefficient on the PacBio platform

PacBio HiFi Sequencing



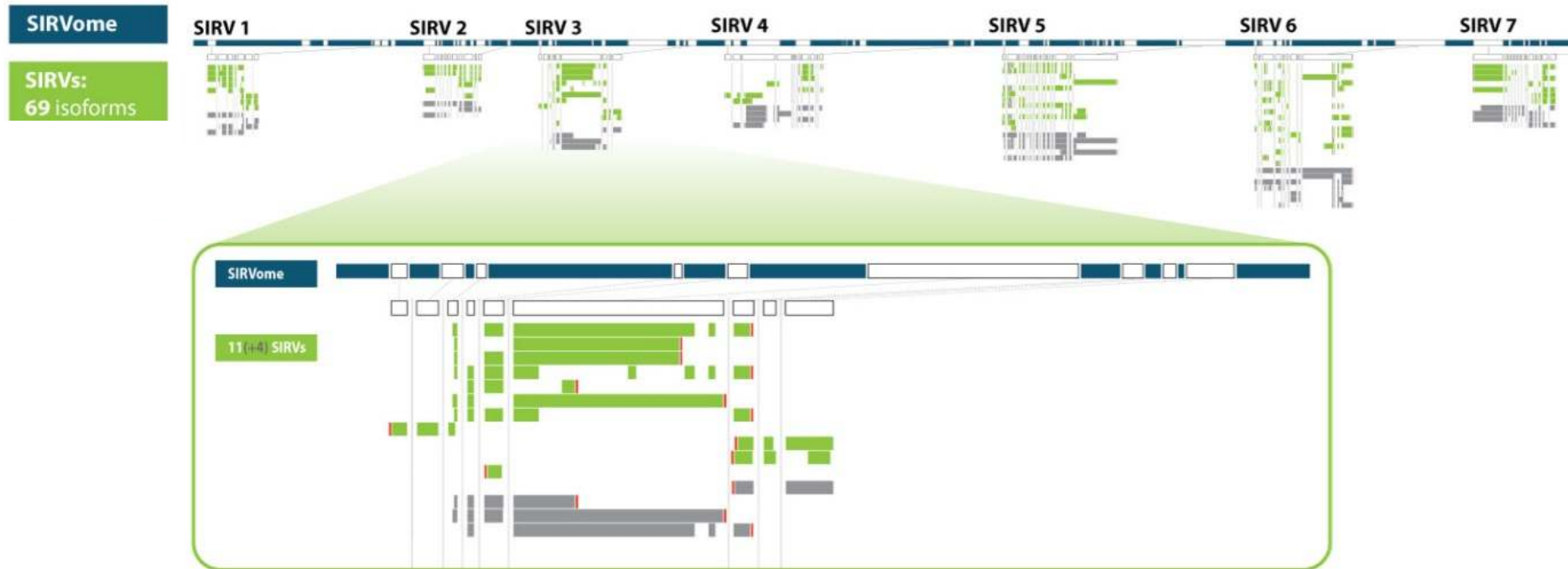
Multiplexed Array Sequencing (MAS-Seq)



>15-fold increase in throughput

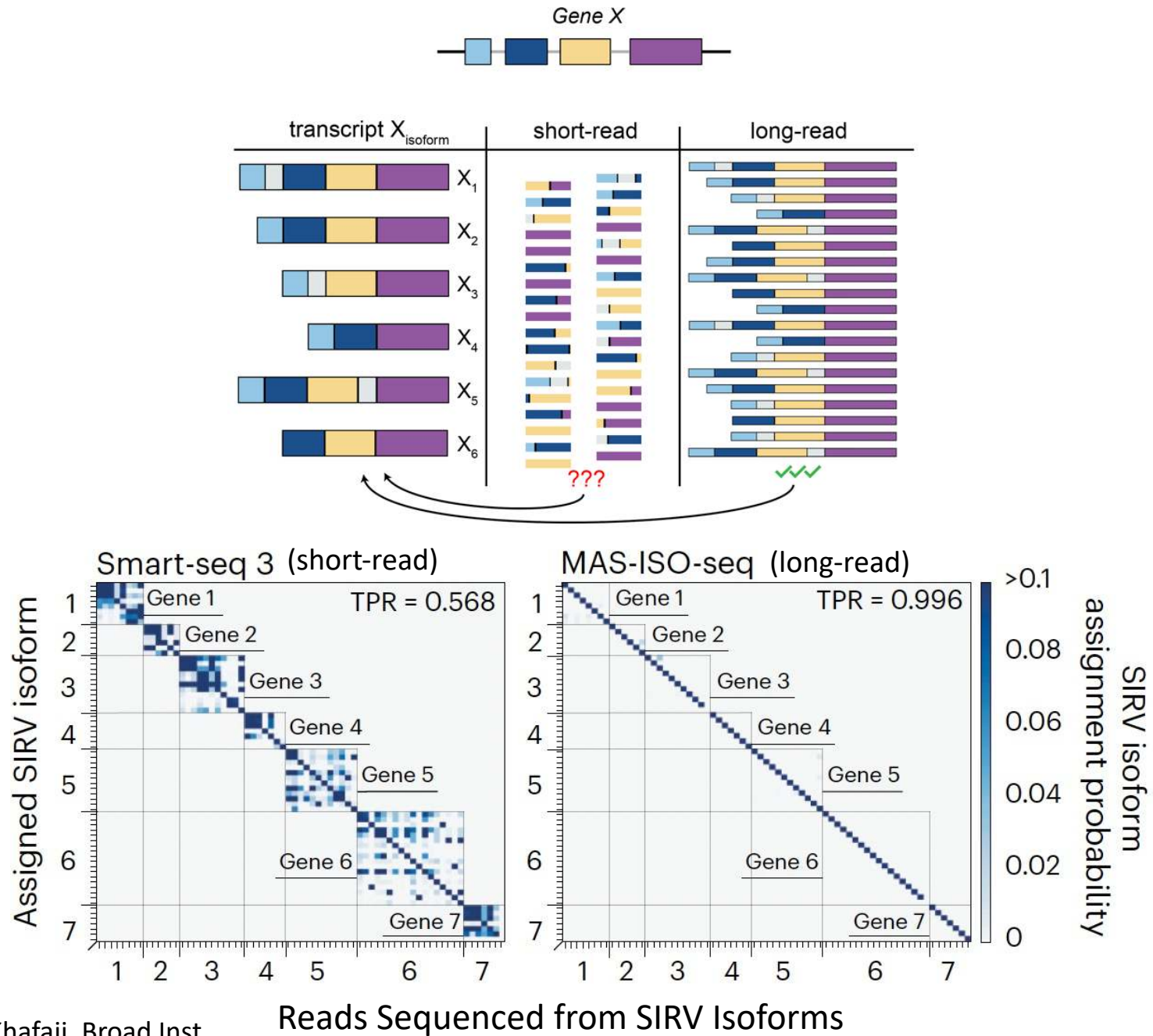
Technical validation using RNA isoform standards

SIRVs (Spike-in RNA Variant Control Mixes) are synthetic gene isoforms



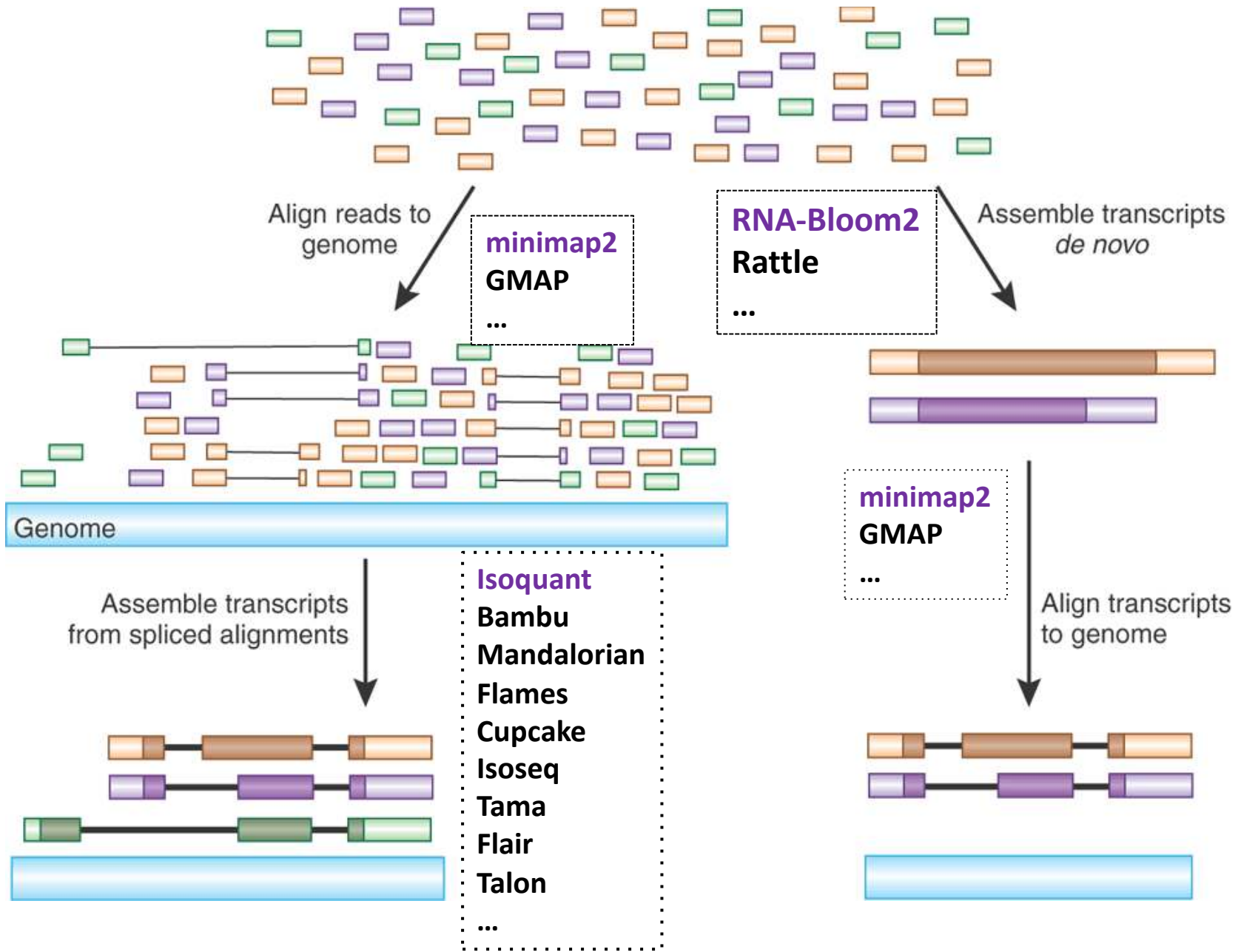
SIRVS serve as truth dataset to evaluate MAS-seq's ability to accurately identify RNA isoforms.

Long-read sequencing accurately identify RNA isoform standards



Transcript Reconstruction from (Long) RNA-Seq Reads

RNA-seq Long Reads (*not drawn to scale*)



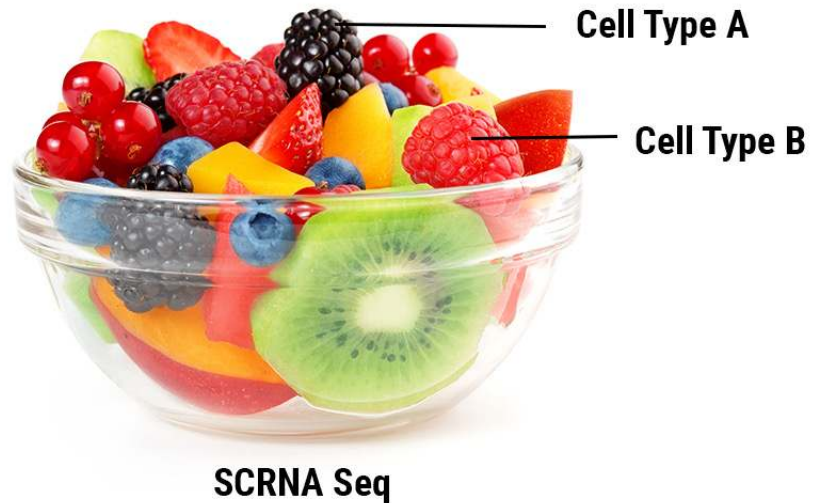


Part 7. Overview of Single Cell Transcriptomics

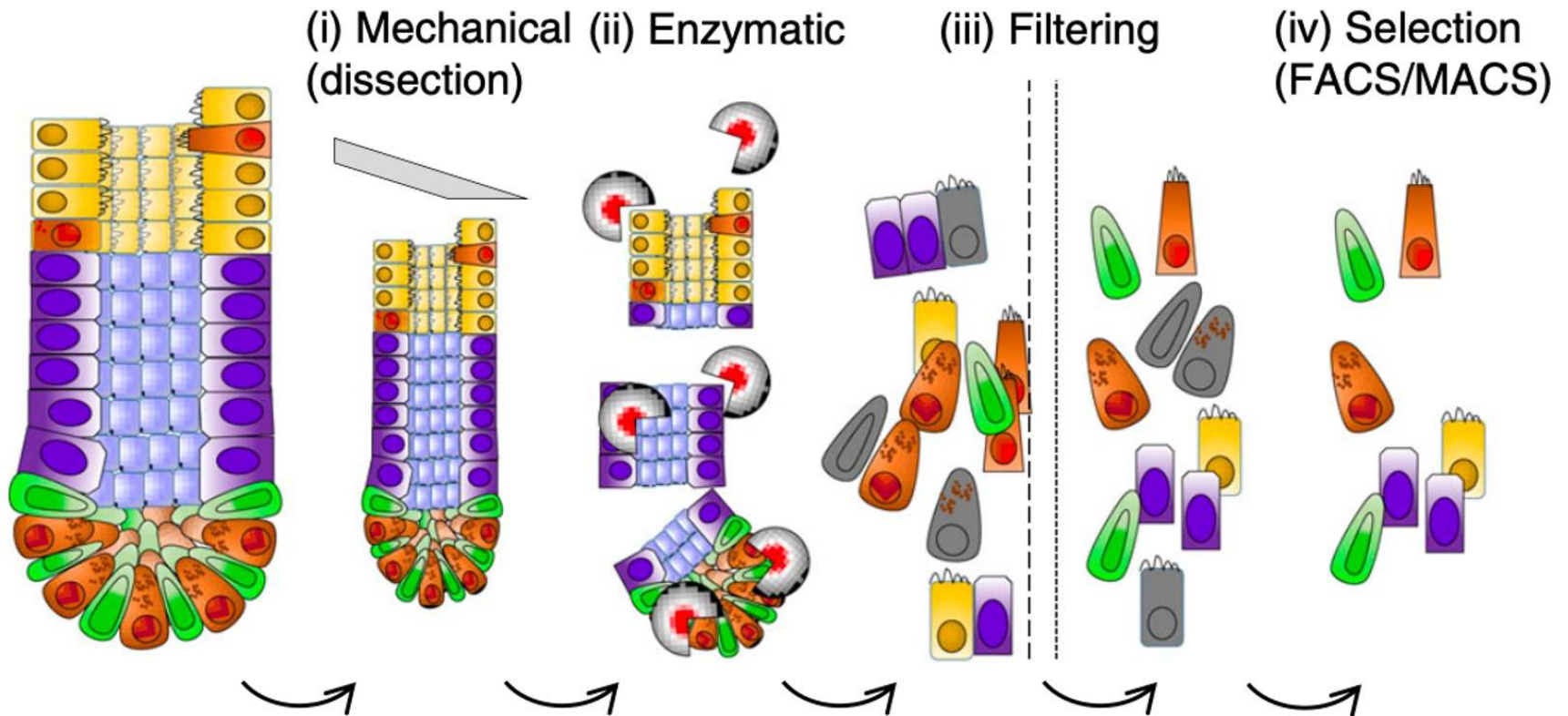
The Quintessential “Fruit Smoothie Metaphor” for Bulk RNA-seq



vs.



Step 1: Break down tissue to single cells (or nuclei)

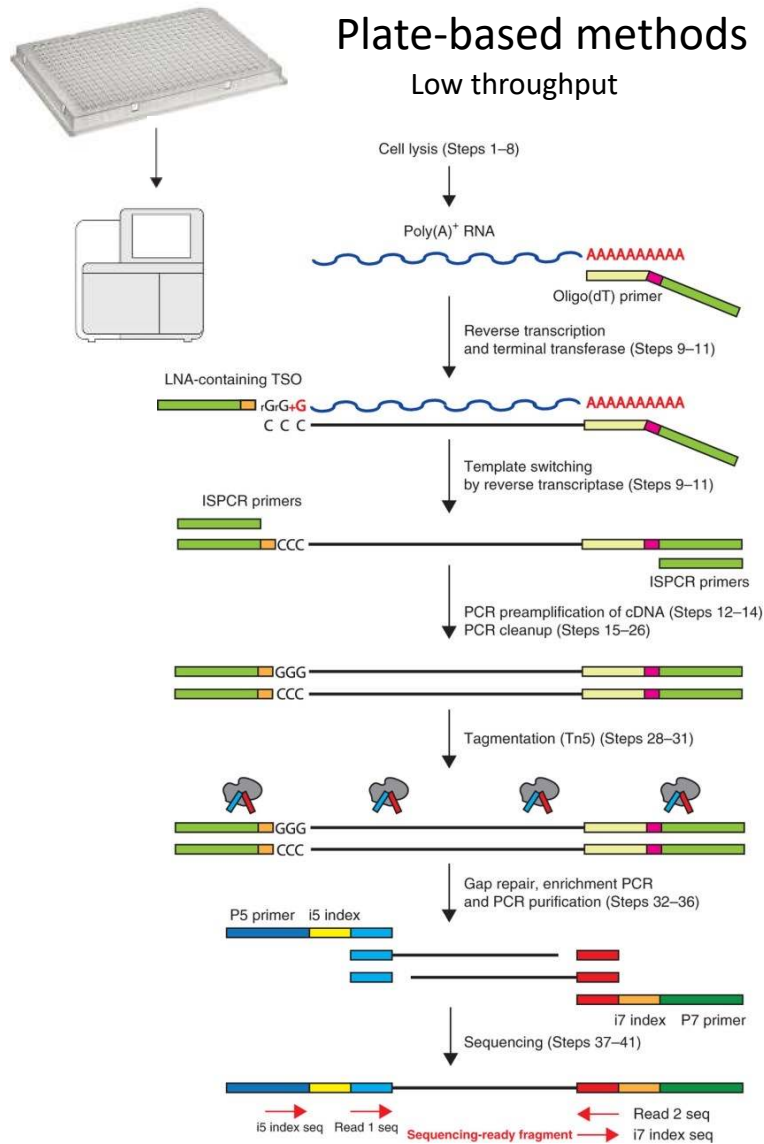


Can also extract and sequence nuclei instead of whole cells – popular in neurobiology

Examples of Different Popular Classes of Single Cell Sequencing

Plate-based methods

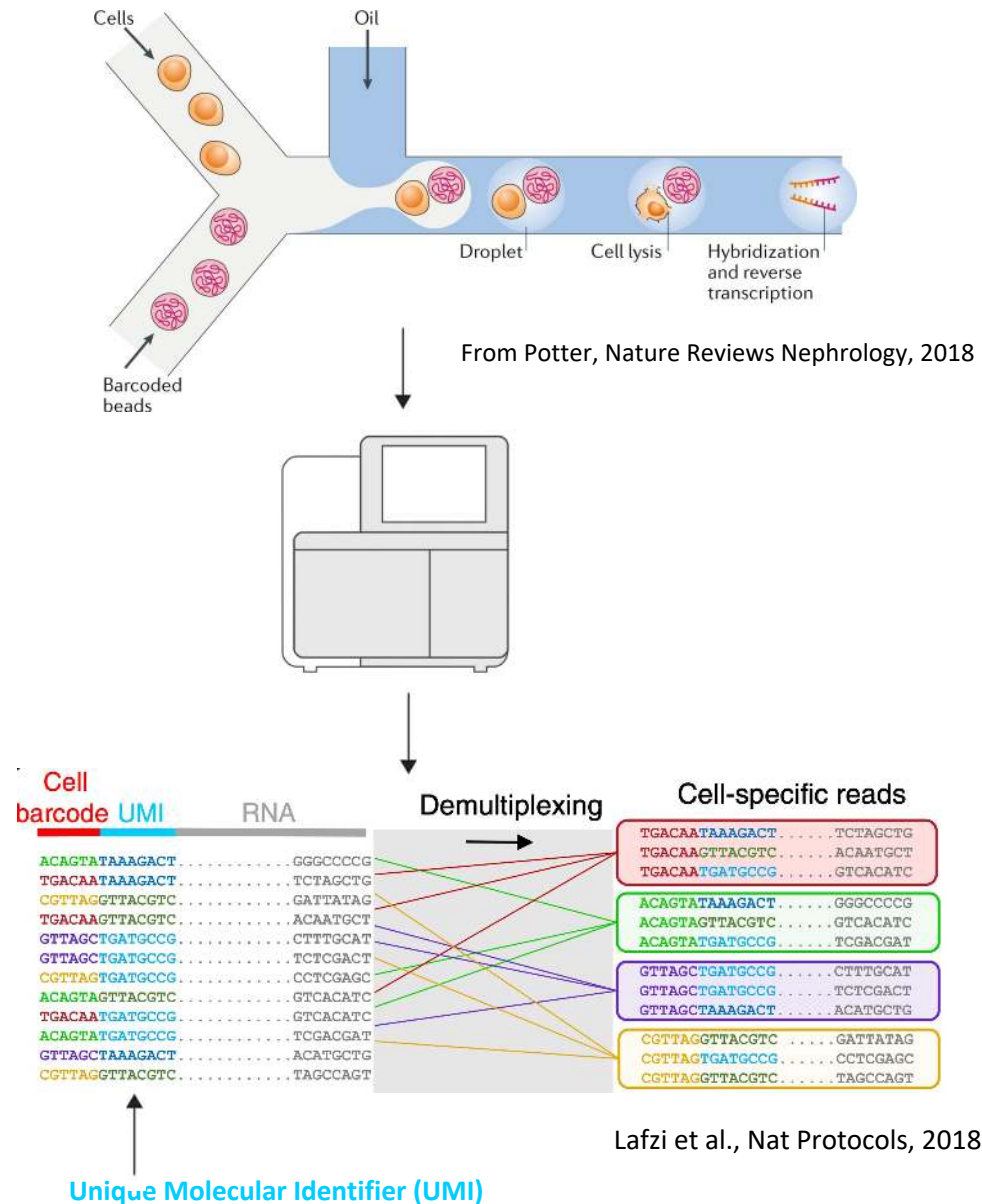
Low throughput



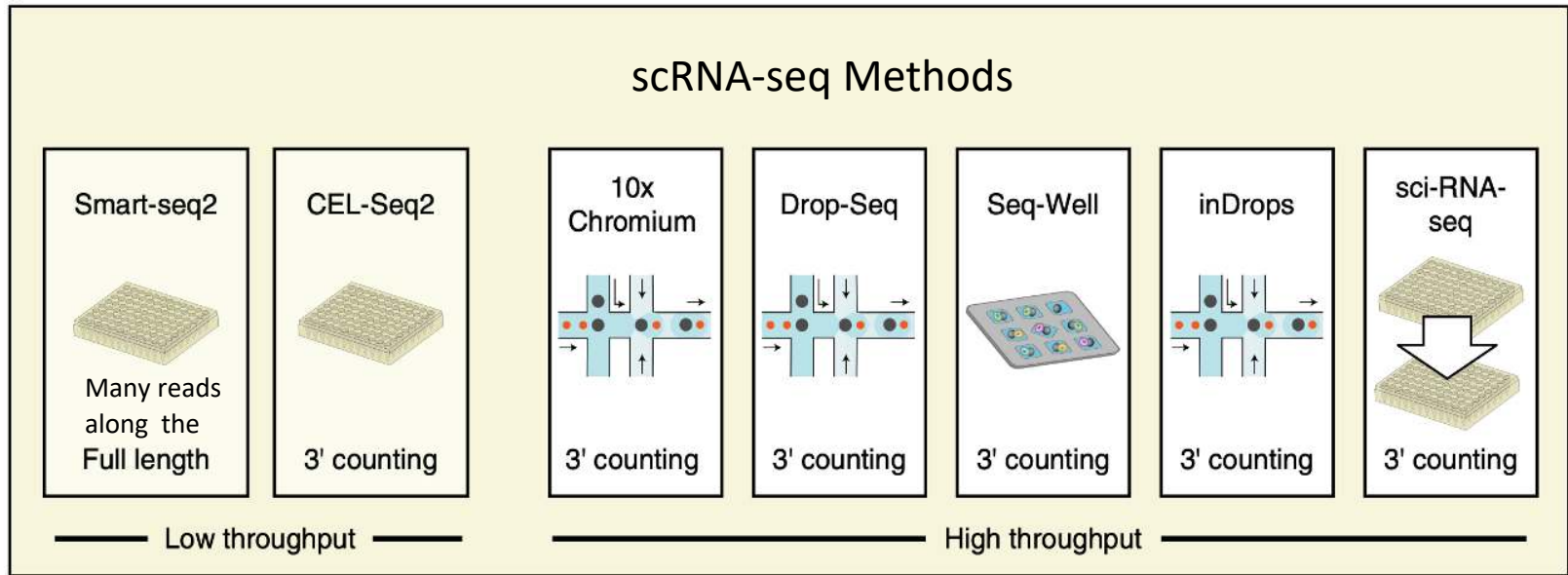
Smart-seq2 Method: Get reads covering the full length of the RNA molecule.

Picelli et al., Nature Protocols, 2014

Droplet-based methods



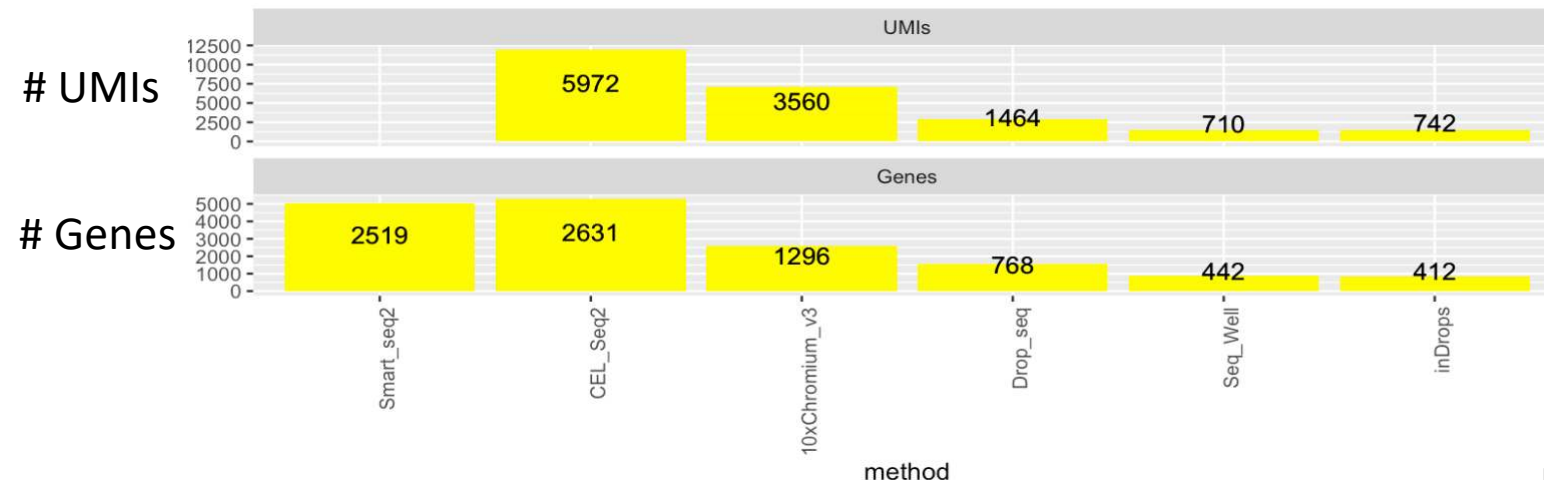
Single Cell Transcriptome Sequencing Methods



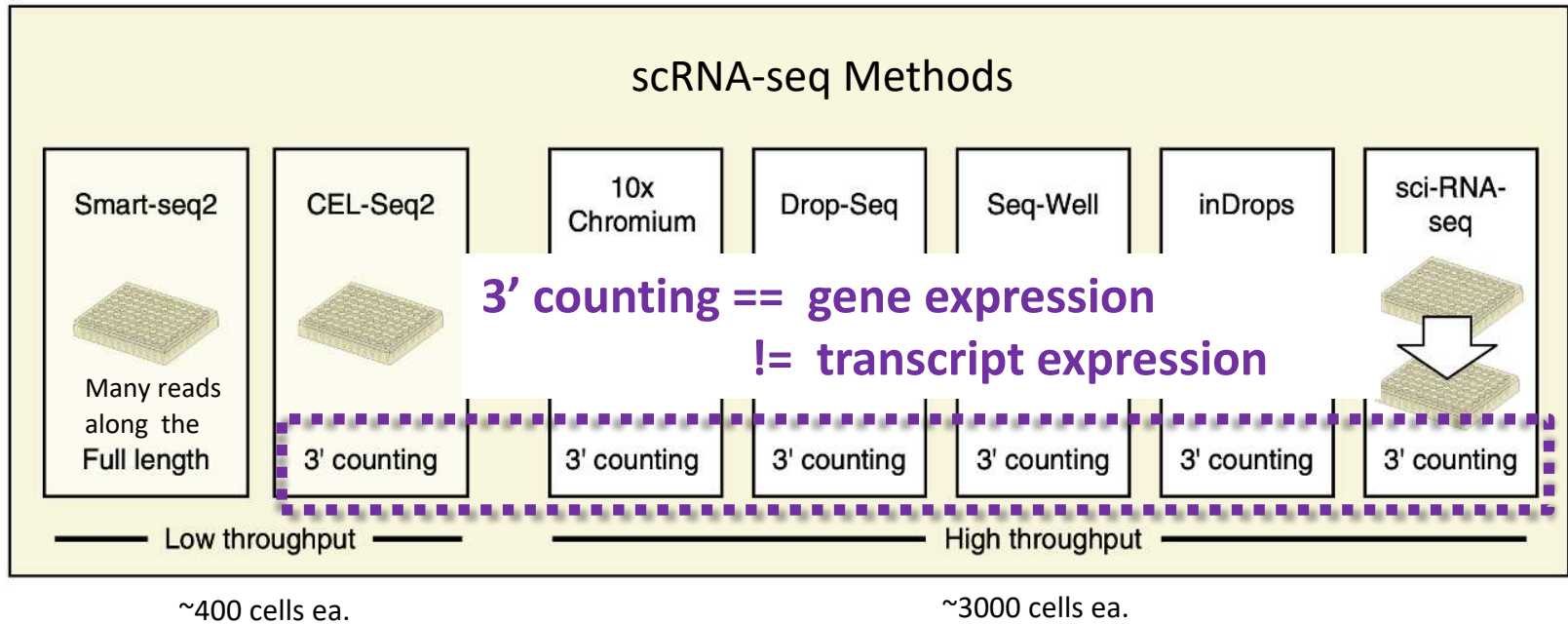
~400 cells ea.

~3000 cells ea.

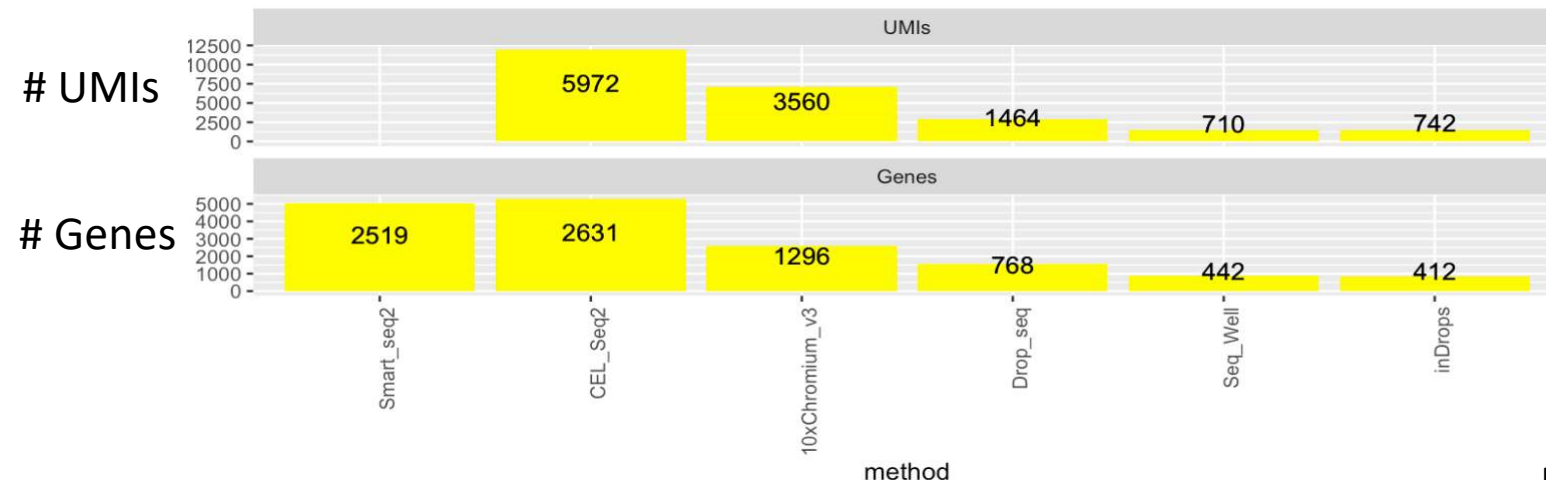
Averaged counts of UMIs and Genes per cell by method



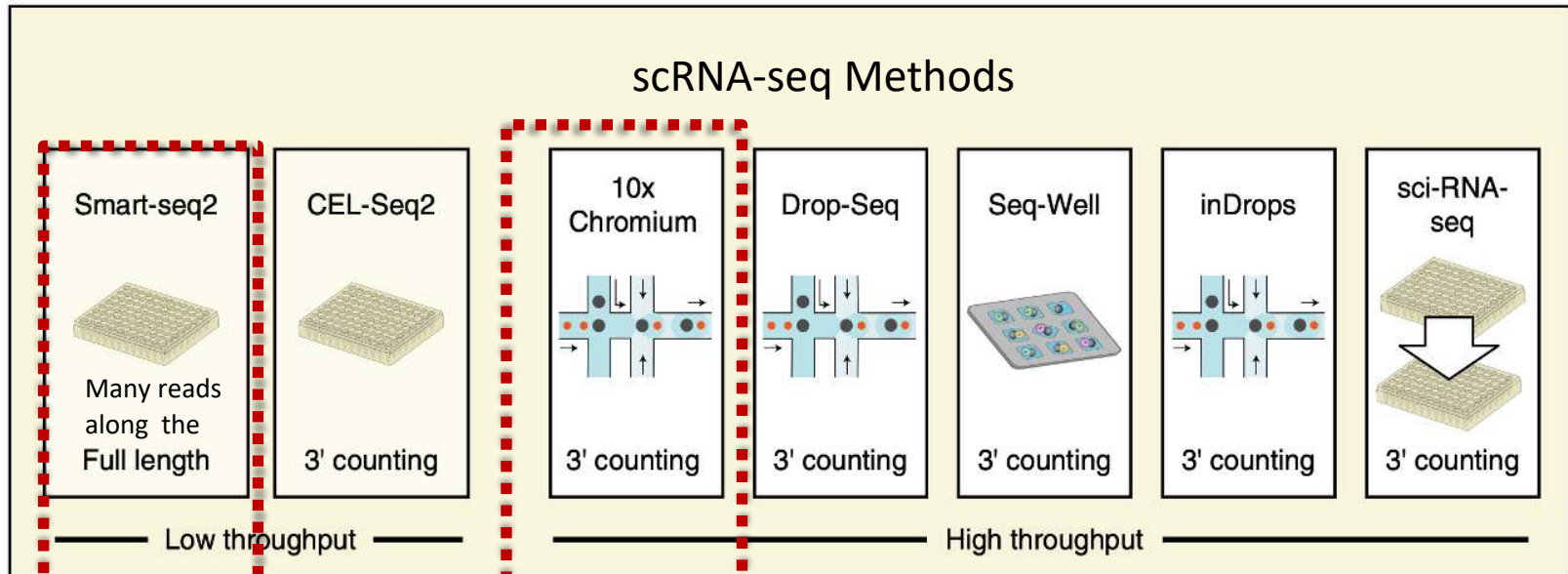
Single Cell Transcriptome Sequencing Methods



Averaged counts of UMIs and Genes per cell by method



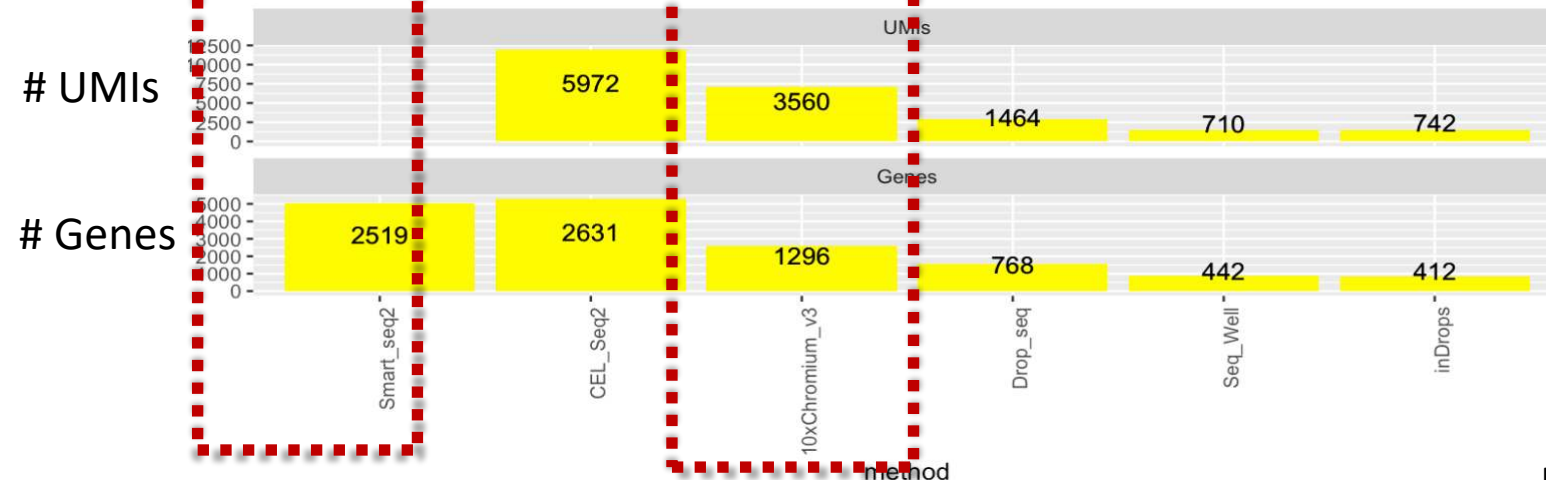
Single Cell Transcriptome Sequencing Methods



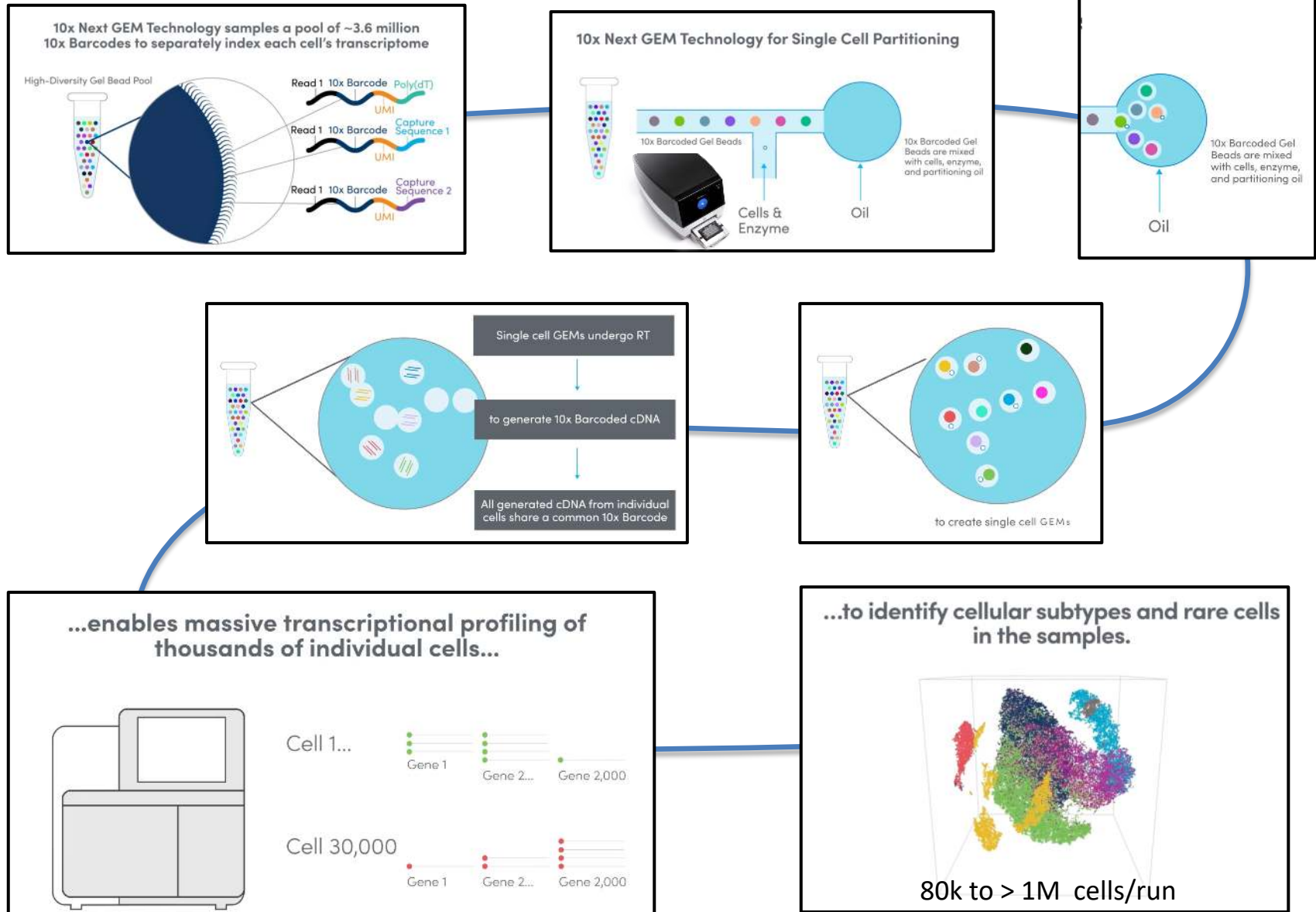
~400 cells ea.

~3000 cells ea.

Averaged counts of UMIs and Genes per cell by method

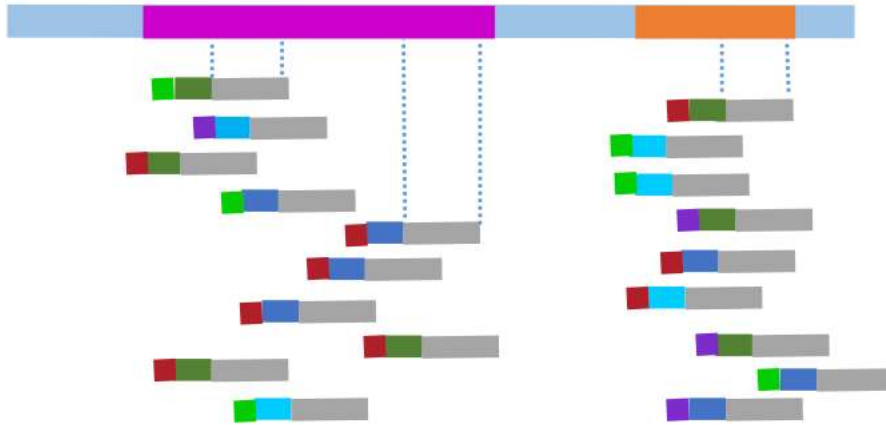


10x Genomics Chromium Single Cell Transcriptome Sequencing



Analysis Workflow for Single Cell Transcriptomics

Reference genome

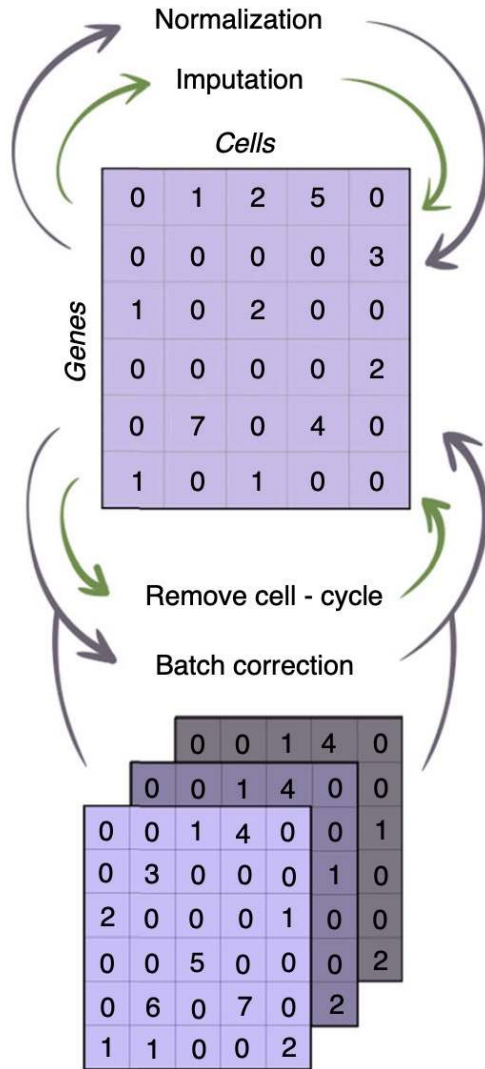


- Align reads to the reference genome
- Collapse PCR duplicates (by UMIs)

	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

- Build a {Gene X Cell} UMI counts matrix

Single Cell Transcriptomics Data Processing Workflow



Gene 'count' matrices for single cell data tend to be very large and very sparse

eg. 25k genes x 100k cells

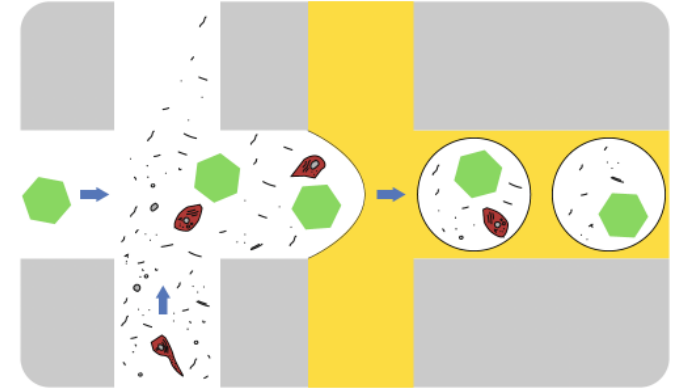
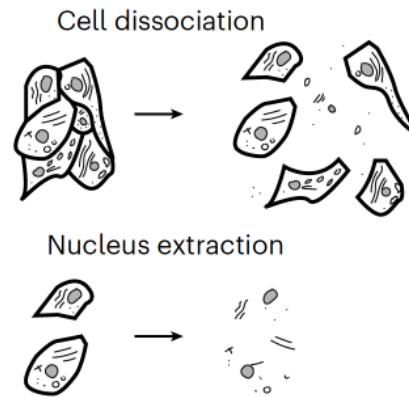
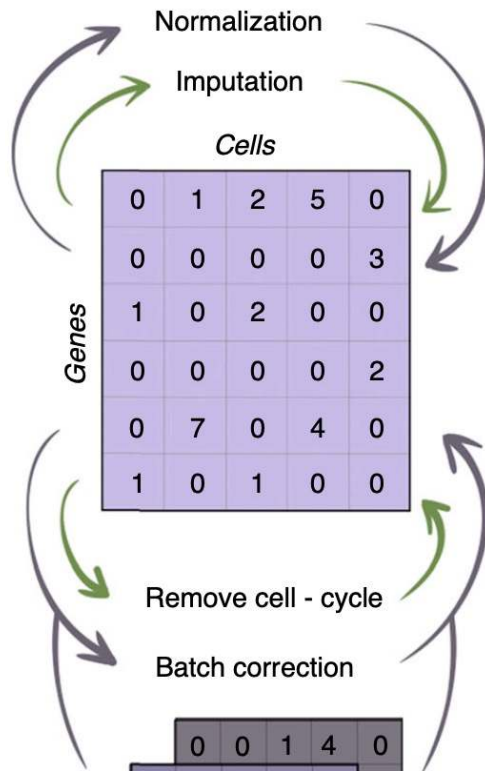
(almost all zeros – no reads detected)

Various processing needed:

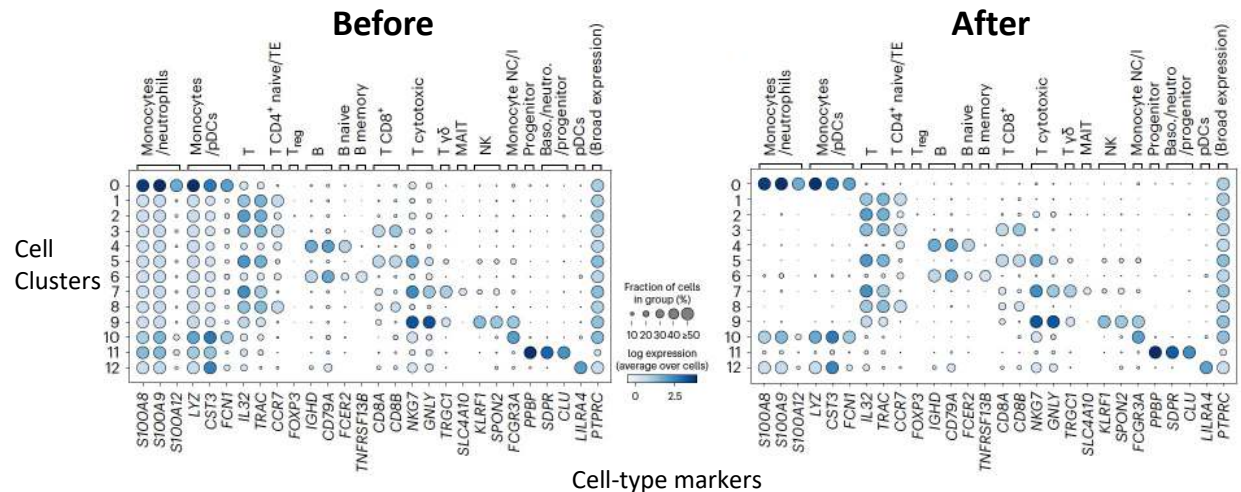
- Which cells are 'good' cells? vs dying/stressed cells, doublets, or empty droplets?
- possibly remove confounding cell cycle signatures from expression data.
- Multiple experiments/replicates - batch correction?

In Silico Removal of Ambient RNA (by Cellbender)

Phenomenology of ambient RNA

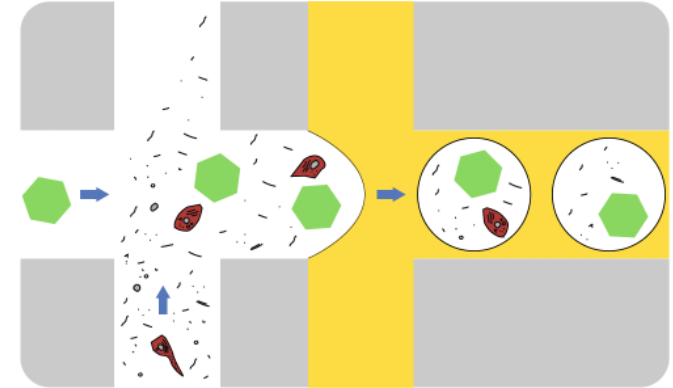
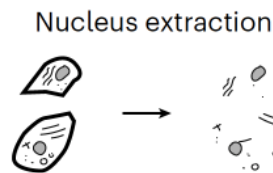
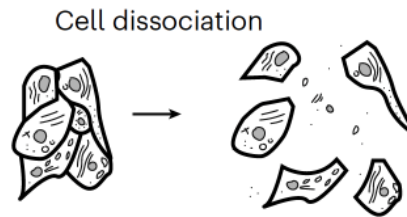
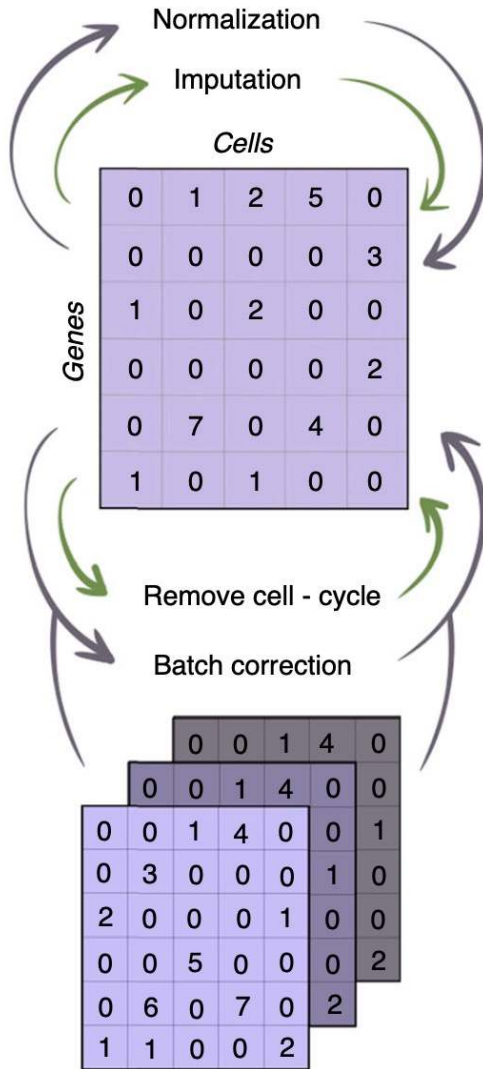


Cell Markers and Read Quantities by Cell Type



In Silico Removal of Ambient RNA (by Cellbender)

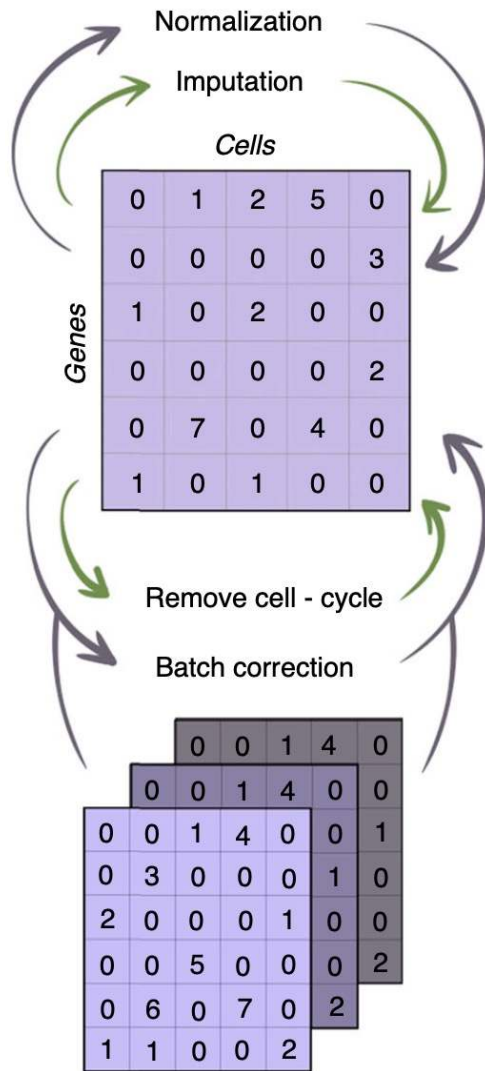
Phenomenology of ambient RNA



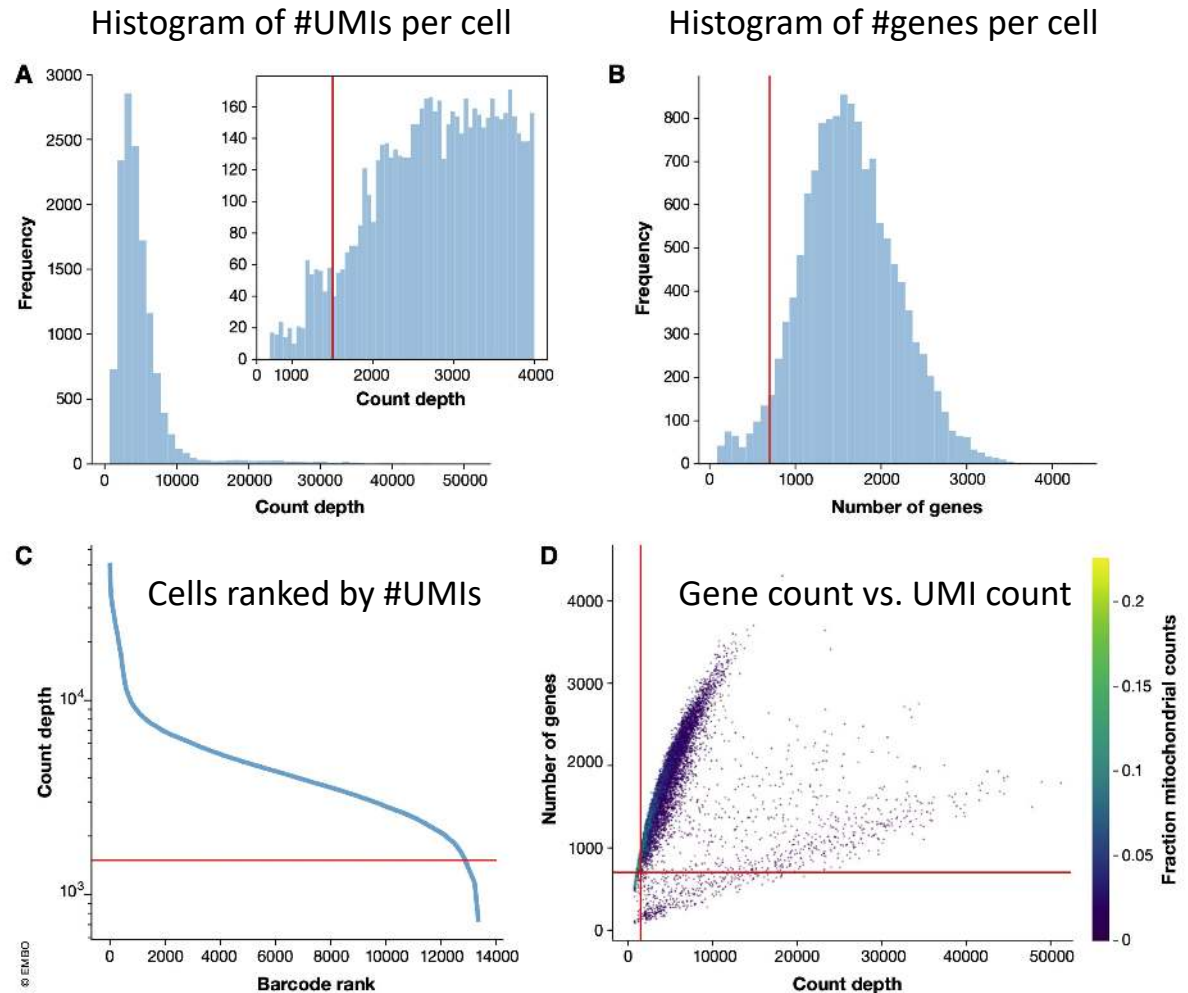
CellBender

<https://github.com/broadinstitute/CellBender>

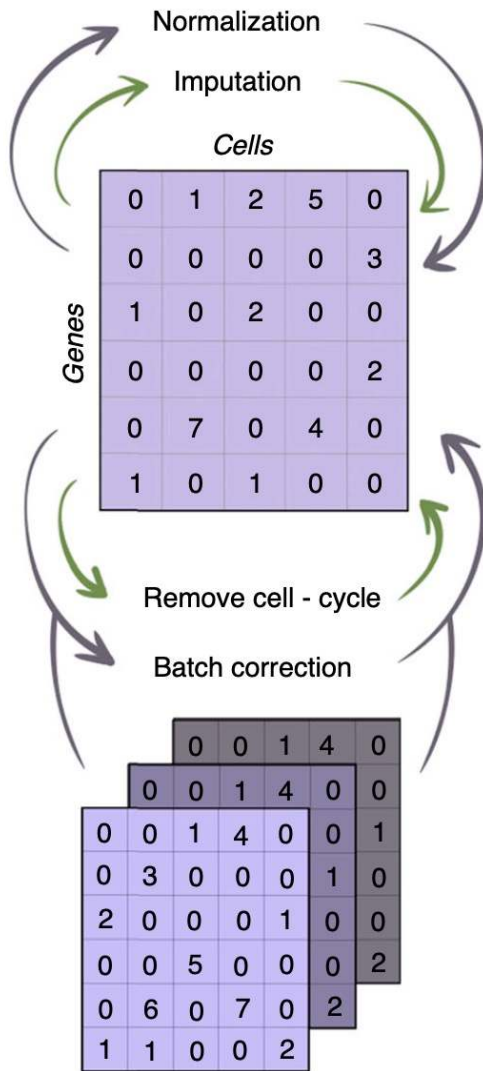
Metrics for Filtering Cells – Keep the Good Ones



Filter cells based on #genes, #UMIs, and %Mito RNA



Batch Correction for Single Cell Transcriptomes



Plot your cells and paint by batch to examine this.
Batch correction methods are available

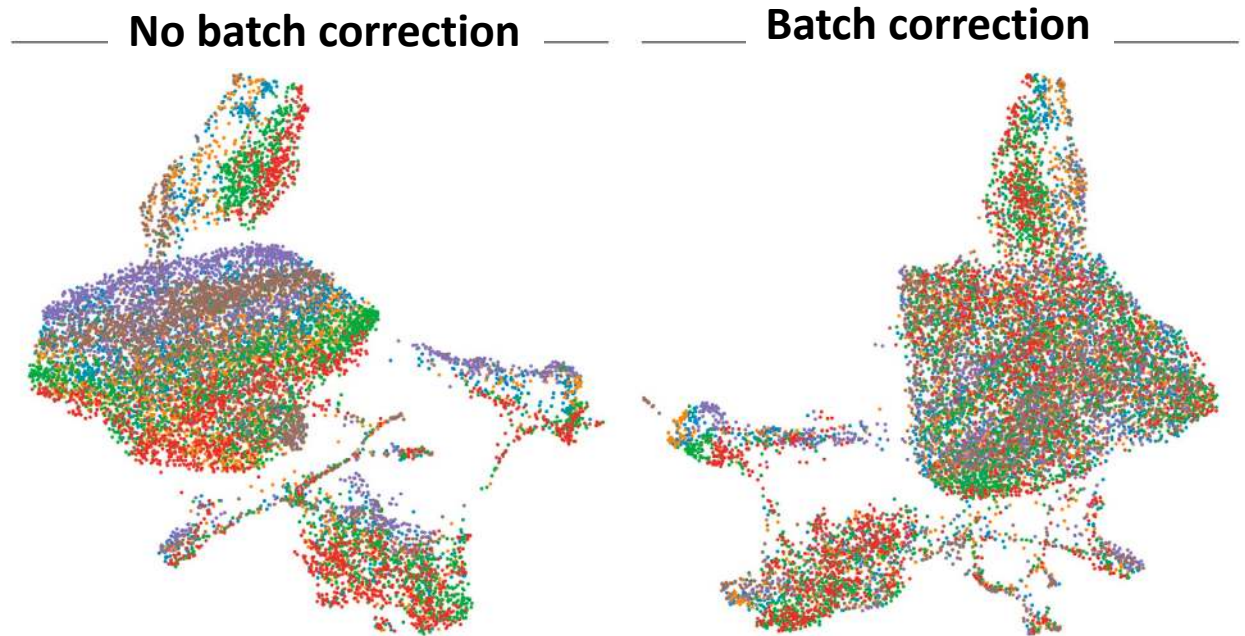
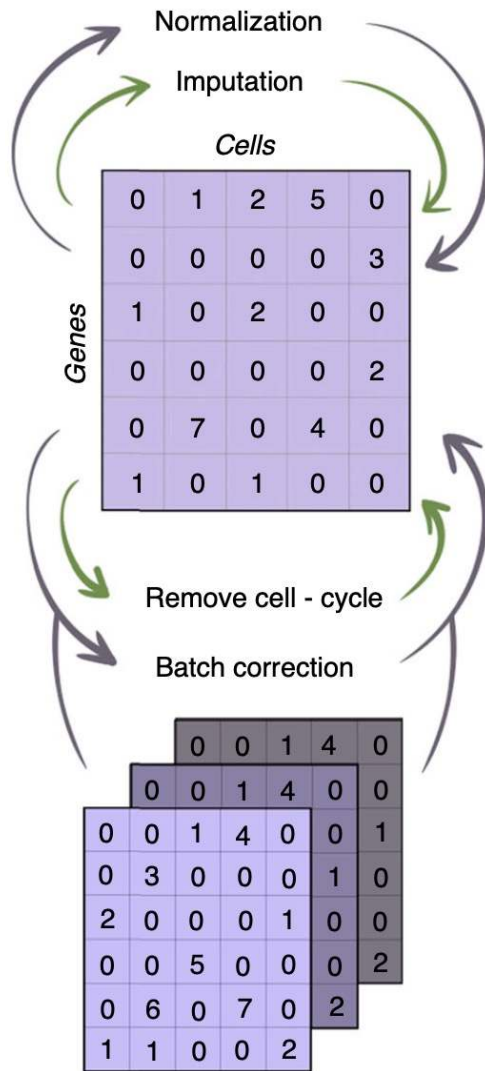


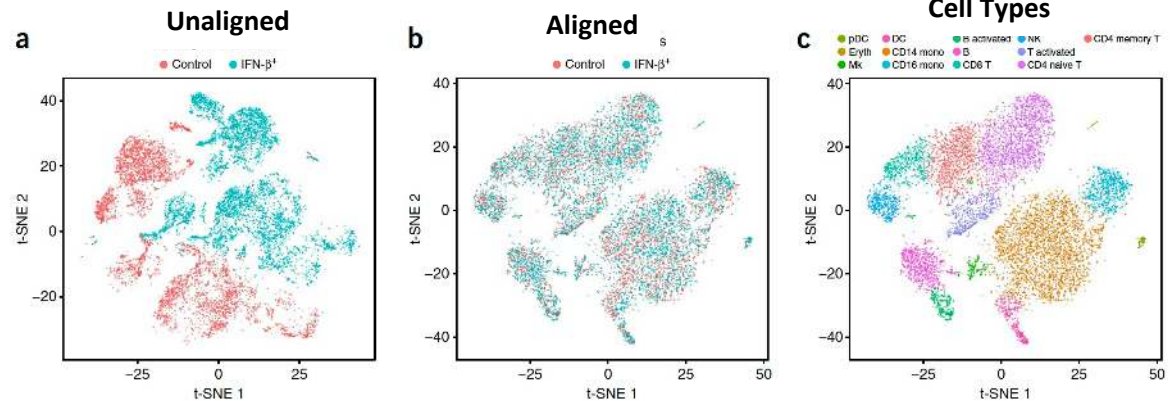
Figure 3. UMAP visualization before and after batch correction.

Cells are coloured by sample of origin. Separation of batches is clearly visible before batch correction and less visible afterwards. Batch correction was performed using ComBat on mouse intestinal epithelium data from Haber *et al* (2017).

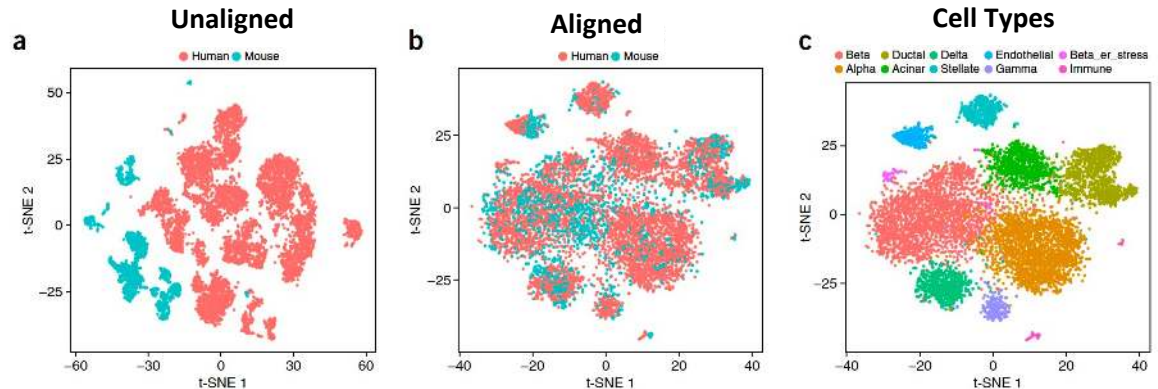
Integrating scRNA-seq data sets based on common sources of variation



Peripheral blood mononuclear cells (PBMCs) +/- stimulation



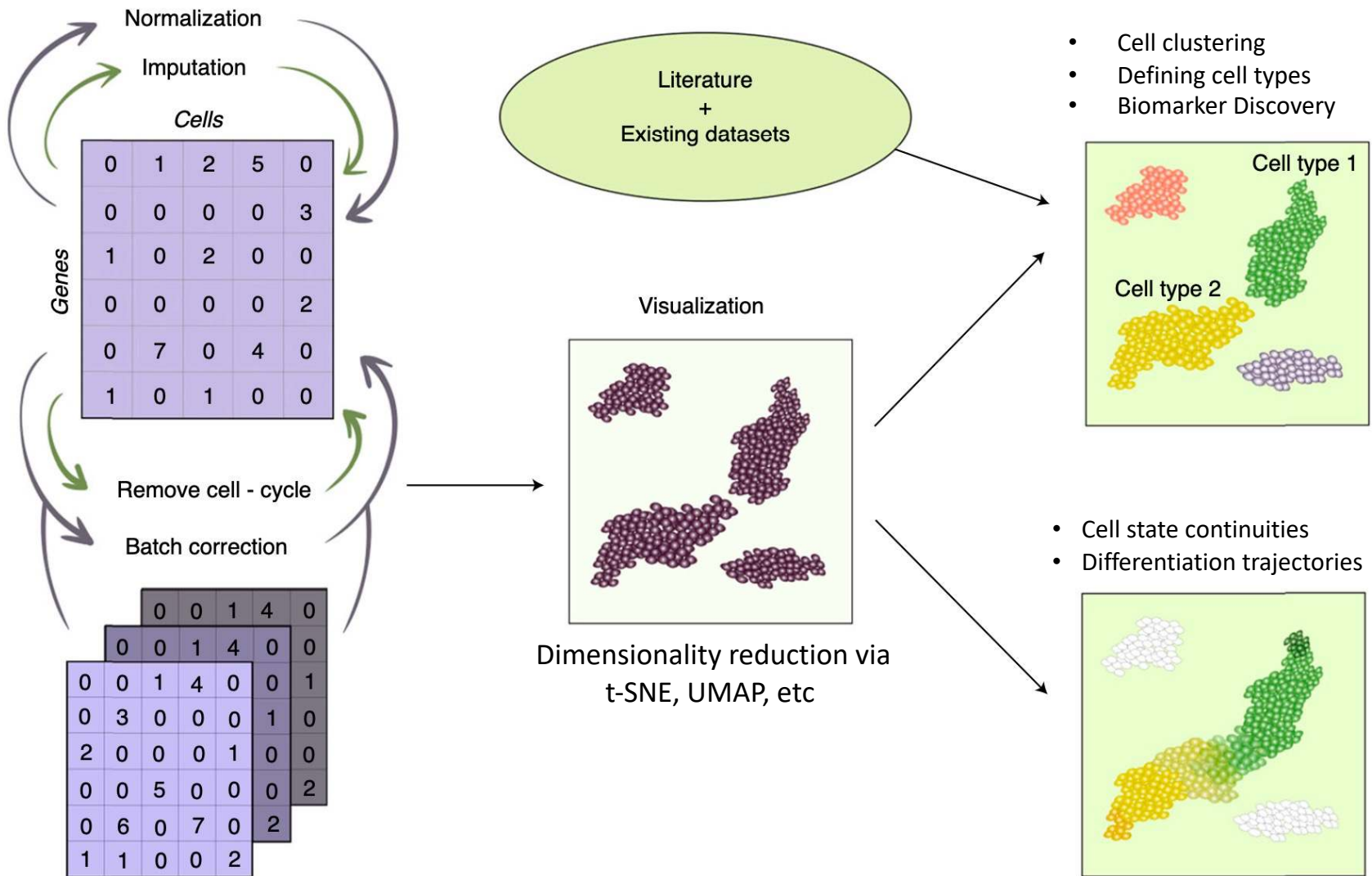
Mouse and human pancreas islet cells



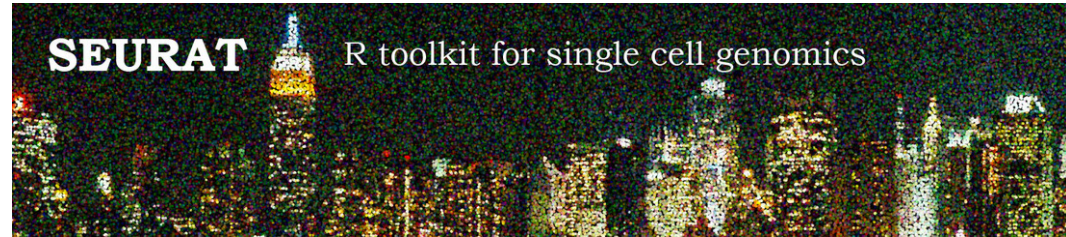
Aligned using Seurat via canonical correlation analysis (CCA)

Butler et al., Nature Biotech, 2018

Finally, Single Cell Data Exploration and Biological Discovery



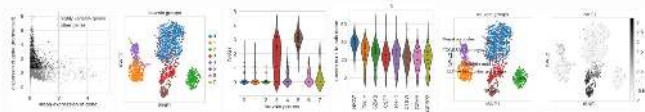
Popular Software Packages for Single Cell Transcriptome Studies



Tutorials

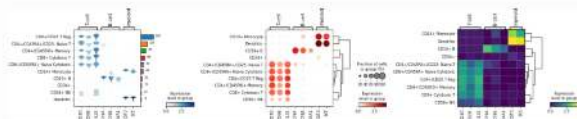
Clustering

For getting started, we recommend Scanpy's reimplementation [→ tutorial: pbmc3k](#) of Seurat's [^cite_satija15] clustering tutorial for 3k PBMCs from 10x Genomics, containing preprocessing, clustering and the identification of cell types via known marker genes.



Visualization

This tutorial shows how to visually explore genes using scanpy. [→ tutorial: plotting/core](#)



Trajectory inference

Get started with the following example for hematopoiesis for data of [^cite_paul15]:

[→ tutorial: paga-paul15](#)



F. Alexander Wolf, Philipp Angerer & Fabian J. Theis,
Genome Biology, 2018;
Isaac Virshup: lead developer since 2019

Vignettes ▾ Extensions FAQ News Reference Archive

Introductory Vignettes

PBMC 3K guided tutorial

Data visualization vignette

SCTransform, v2 regularization

Using Seurat with multi-modal data

Seurat v5 Command Cheat Sheet

Data Integration

Introduction to scRNA-seq integration

Integrative analysis in Seurat v5

Mapping and annotating query datasets

Multi-assay data

Dictionary Learning for cross-modality integration

Weighted Nearest Neighbor Analysis

Integrating scRNA-seq and scATAC-seq data

Multimodal reference mapping

Mixscape Vignette

Massively scalable analysis

Sketch-based analysis in Seurat v5

Sketch integration using a 1 million cell dataset from Parse Biosciences

Map COVID PBMC datasets to a healthy reference

BPCells Interaction

Spatial analysis

Analysis of spatial datasets (Imaging-based)

Analysis of spatial datasets (Sequencing-based)

Other

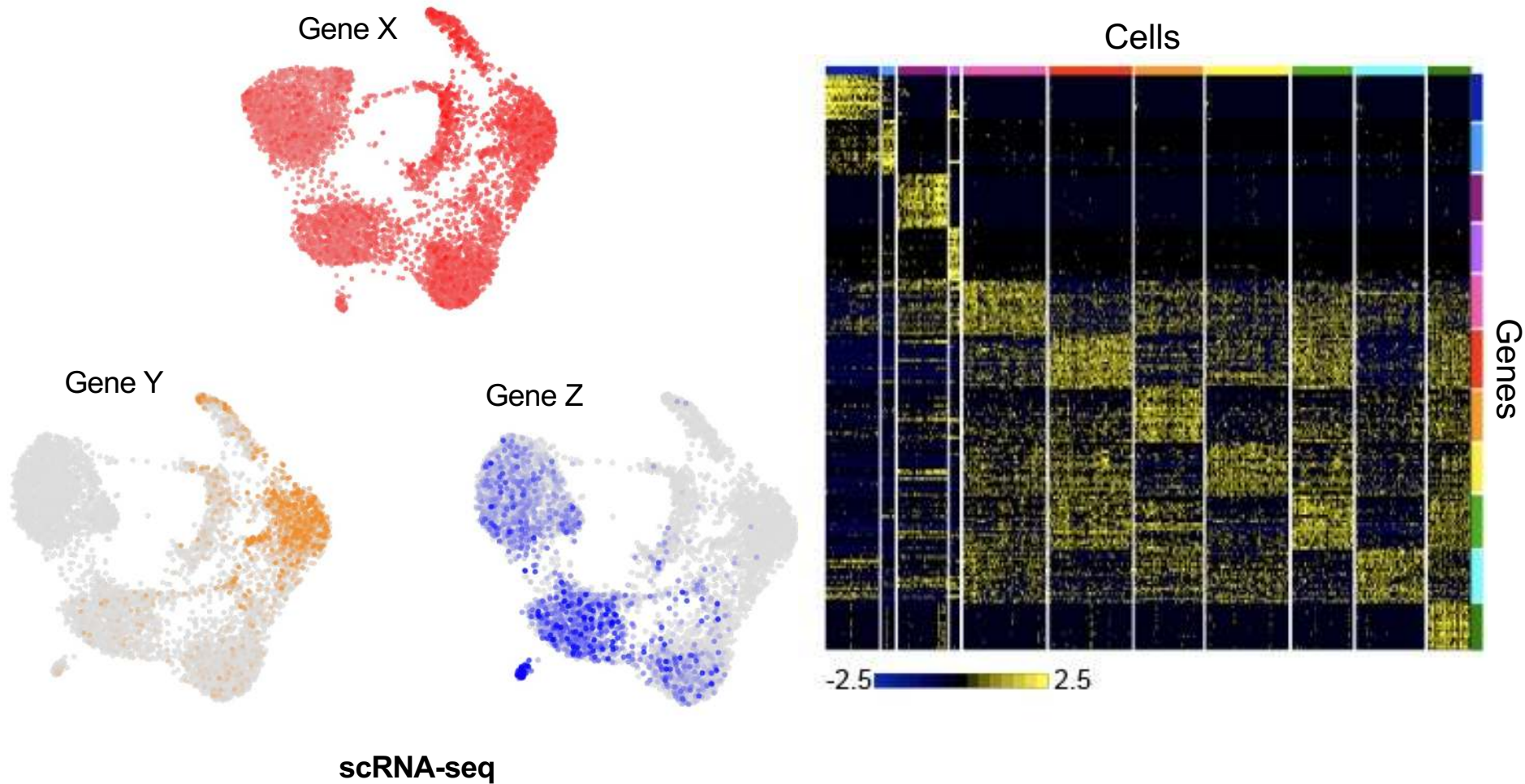
Cell-cycle scoring and regression

Differential expression testing

Demultiplexing with hashtag oligos (HTOs)

From
Rahul Satija's
lab

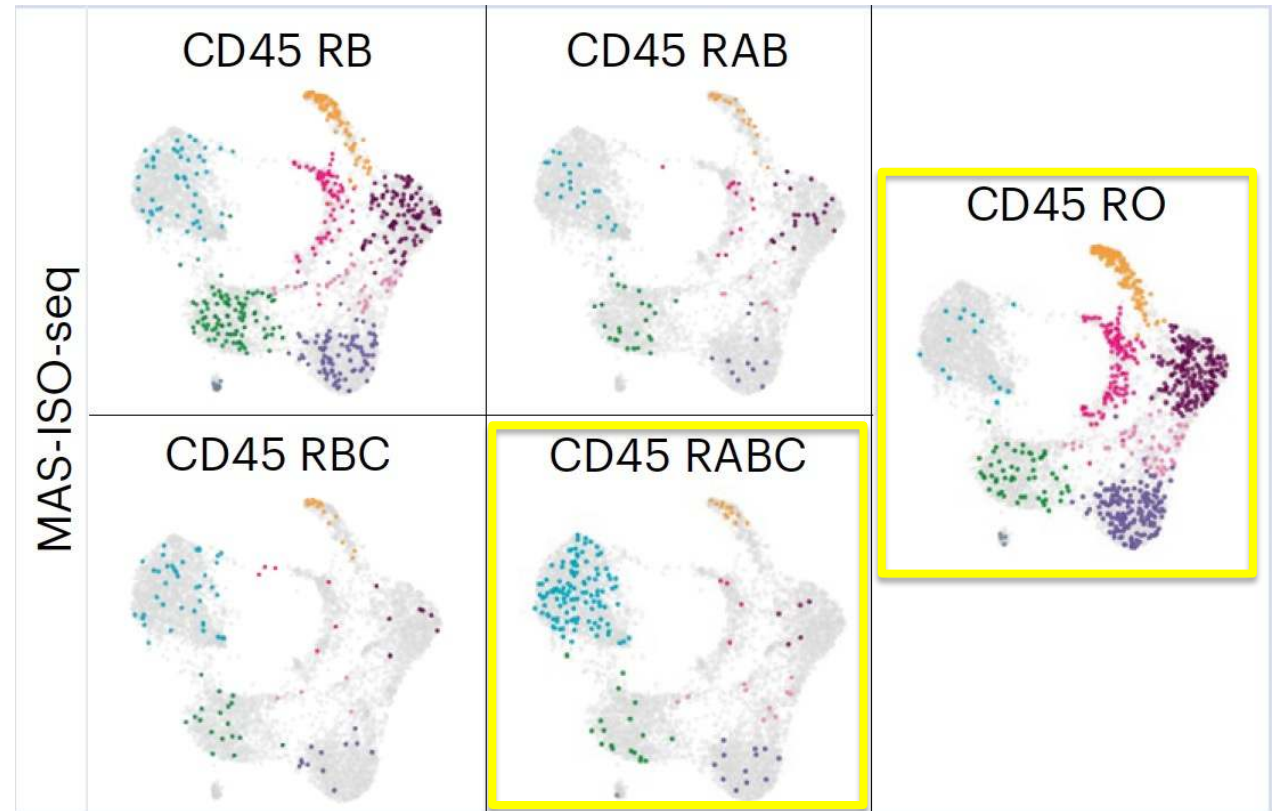
Gene expression $\neq$ transcript expression



But – long isoform reads to the rescue!!

Long read scRNA-seq (scMAS-Iso-seq) of tumor infiltrating CD8 T cells

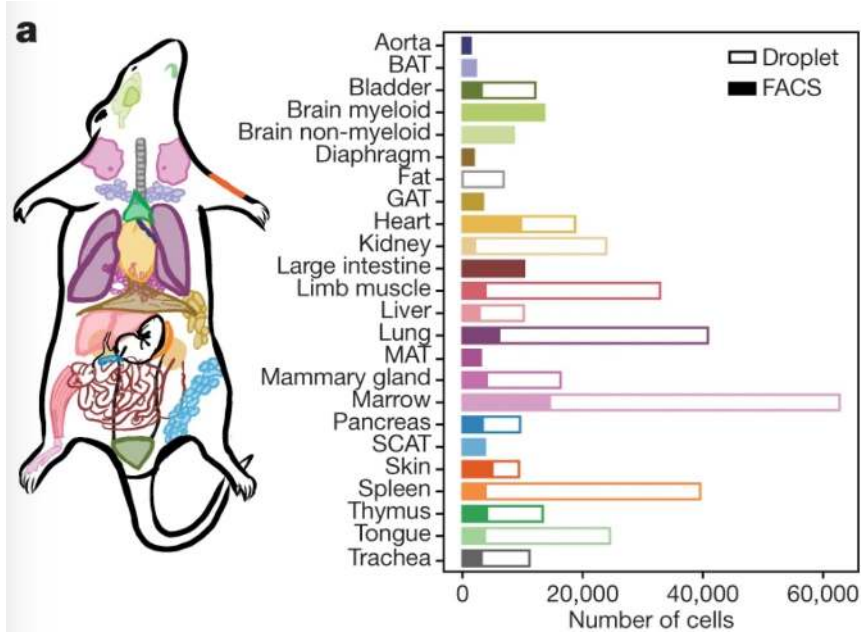
CD45 T-cell Marker Isoform expression resolved via long reads



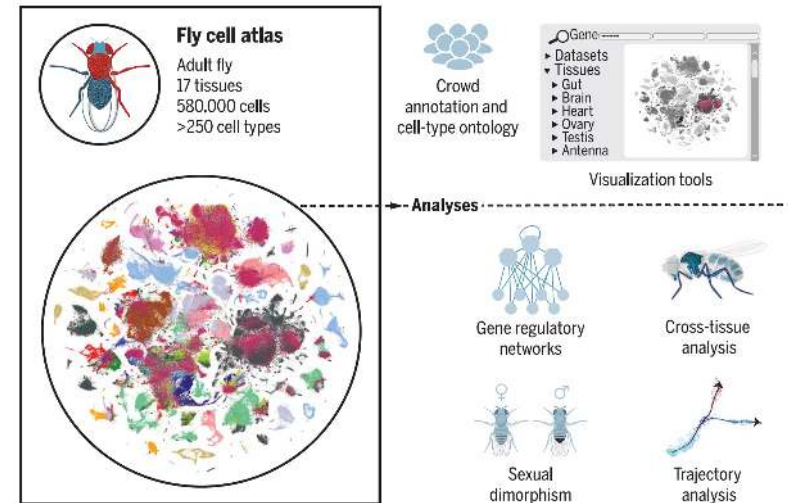
Perform MAS-Iso-seq on the 10x sc libraries to get long isoform reads at single cell resolution

Cataloguing Cell Types and Building Cell Atlases

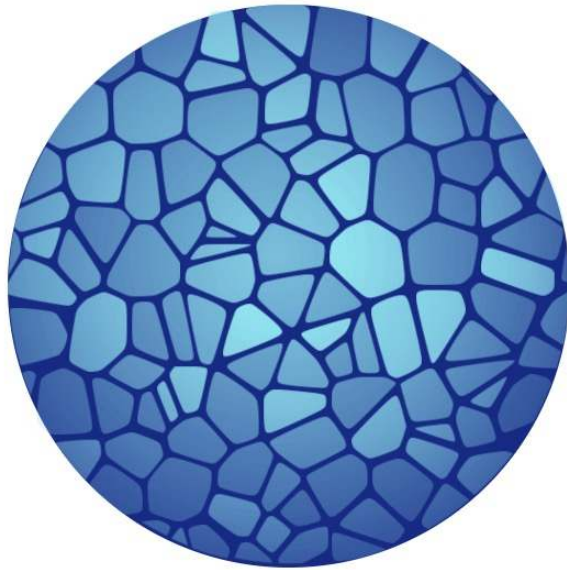
Tabula Muris



Tabula Drosophila



Tabula Drosophilae. In this single-cell atlas of the adult fruit fly, 580,000 cells were sequenced and >250 cell types were annotated. They are from 15 individually dissected sexed tissues as well as the entire head and body. All data are freely available for visualization and download, with featured analyses shown at the bottom right.

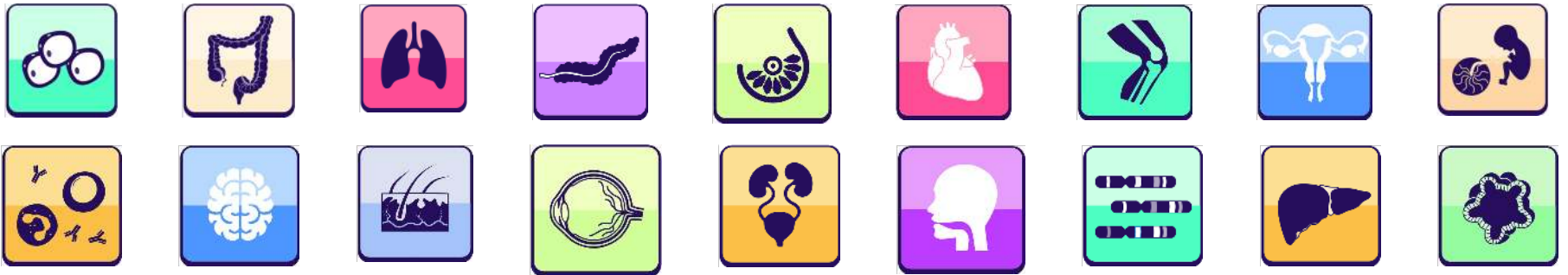


HUMAN CELL ATLAS

Characterize the ~37 trillion cells in the human body

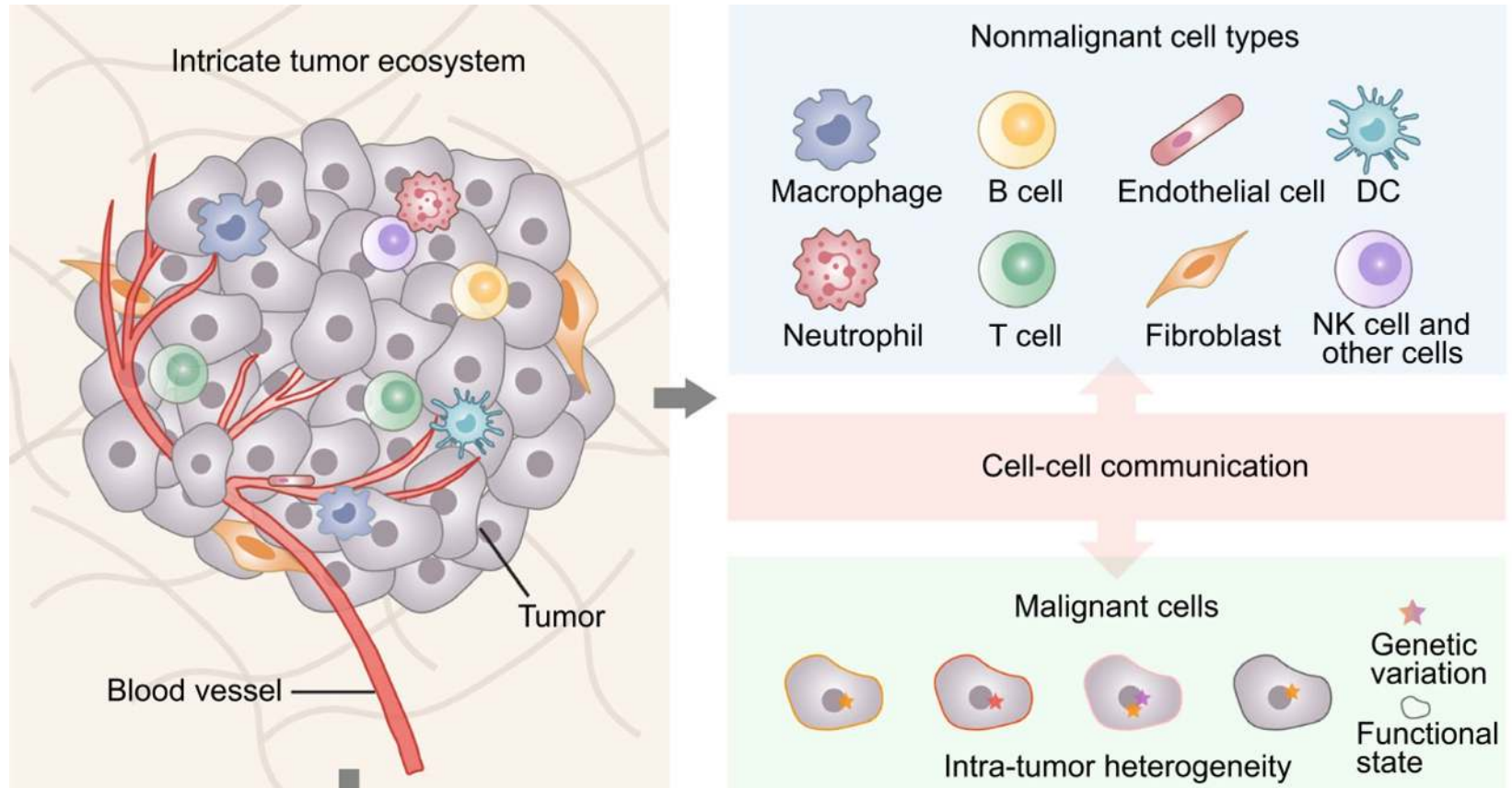
HCA is a global initiative of > 3k members

Initially targeting 18 biological networks of organs and tissues

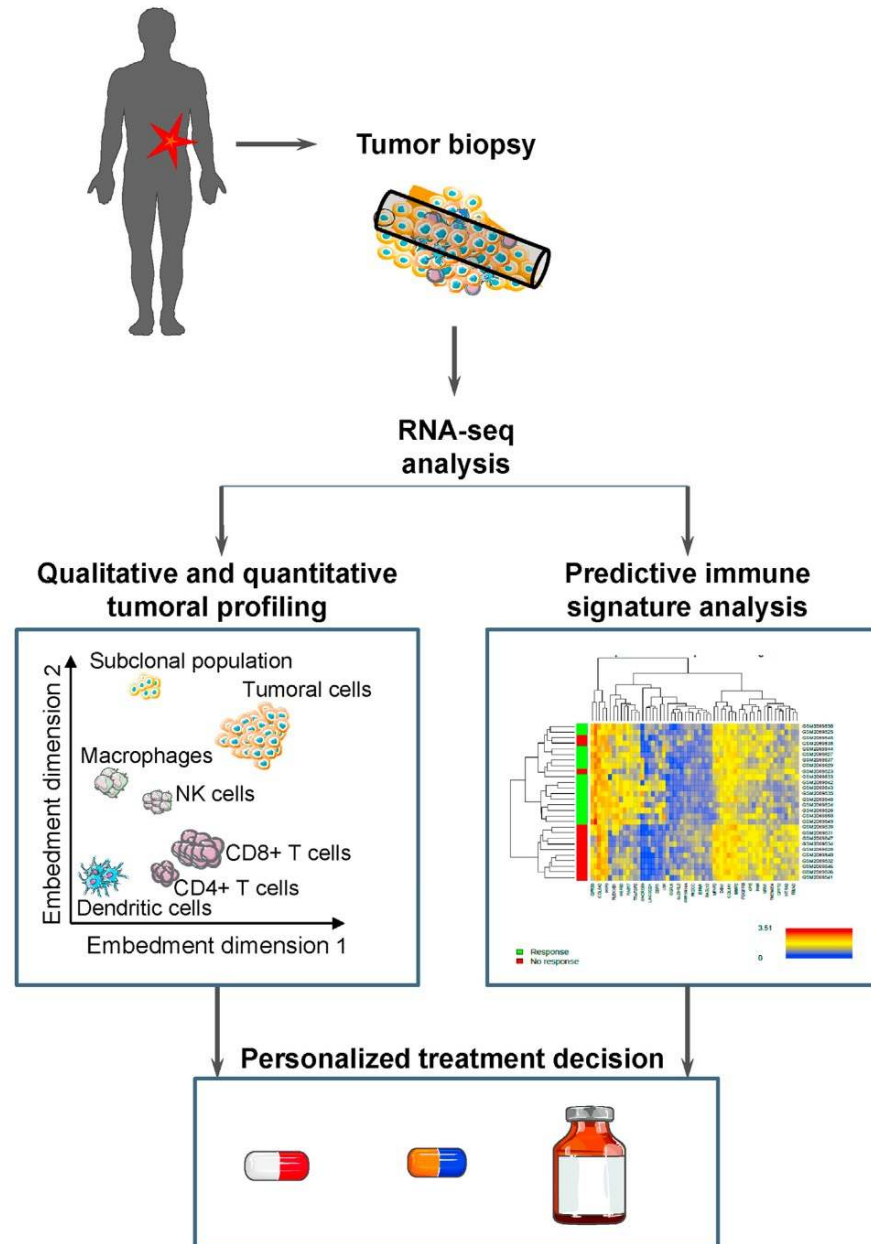


<https://www.humancellatlas.org/>

Single cell analysis is revolutionizing cancer research



Clinical Application for Tumor Single Cell Transcriptomics

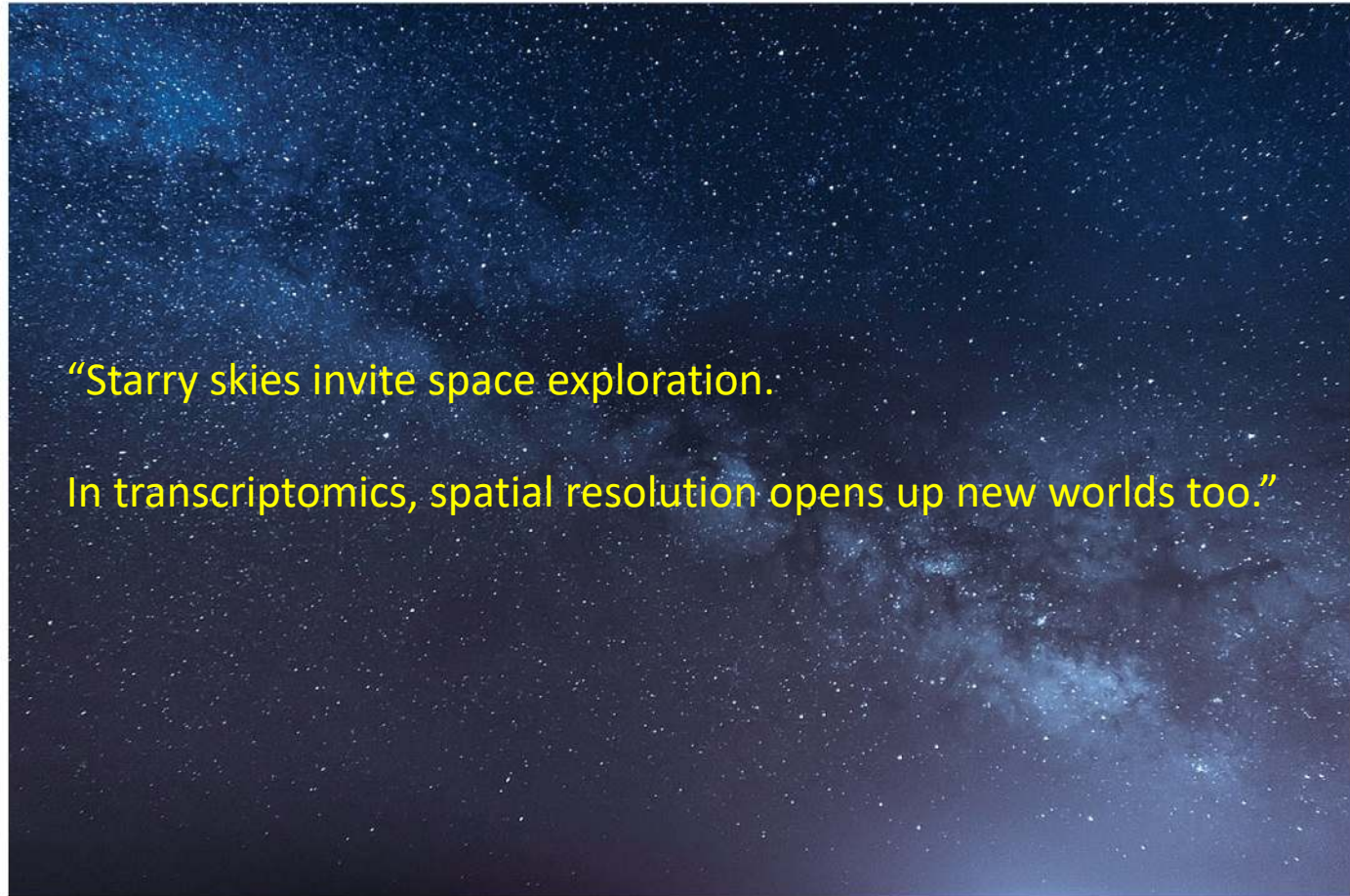




Part 8. Overview of Spatial Transcriptomics

Method of the Year: spatially resolved transcriptomics

Nature Methods has crowned spatially resolved transcriptomics Method of the Year 2020.



“Starry skies invite space exploration.

In transcriptomics, spatial resolution opens up new worlds too.”

Starry skies invite space exploration. In transcriptomics, spatial resolution opens up new worlds too.

Credit: bjdlsx/Getty Images

Method of the Year: spatially resolved transcriptomics

Nature Methods has crowned spatially resolved transcriptomics Method of the Year 2020.



Starry skies invite space exploration. In transcriptomics, spatial resolution opens up new worlds too.

Credit: bjdlsx/Getty Images

Single Cells vs. Spatial Transcriptomics



Car parts ~ single cells

Vs.



Car ~ tissue

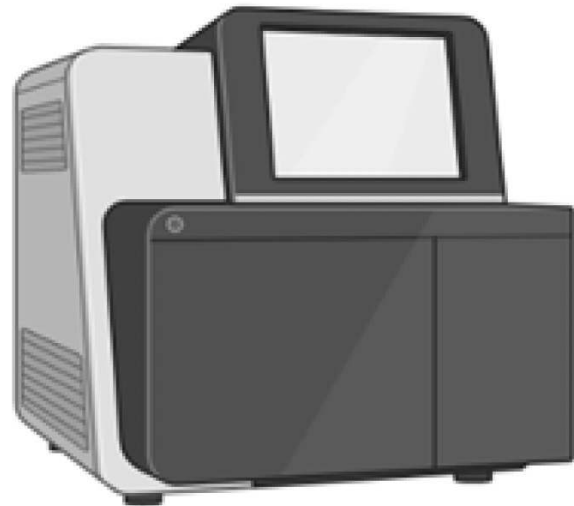
Classes of Spatial Transcriptomics

Imaging Readout



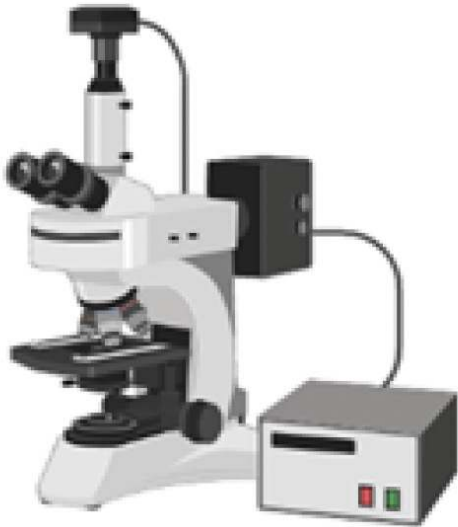
Based on In Situ Hybridization (ISH)
and fluorescent tags

Sequencing Readout



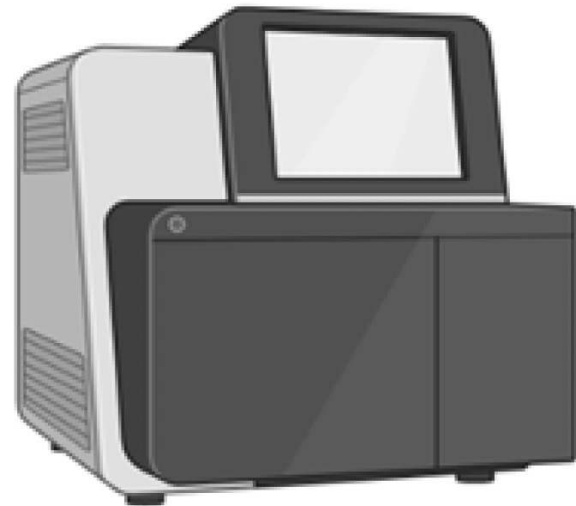
Classes of Spatial Transcriptomics

Imaging Readout



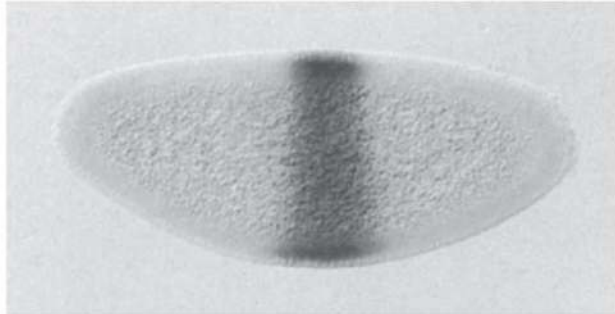
Based on In Situ Hybridization (ISH)
and fluorescent tags

Sequencing Readout



Single Molecule Fish (smFISH) Methods for Visualizing RNA Molecules at Sub-cellular Resolution

a Long probe, many labels



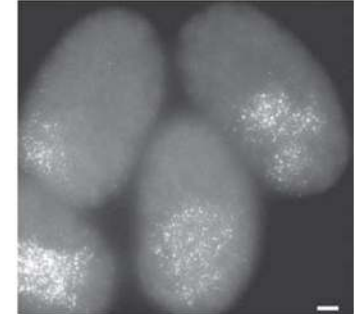
Target: hunchback RNA in *Drosophila* embryo

b Shorter probes, fewer labels



Target: single transcripts in mammalian cells

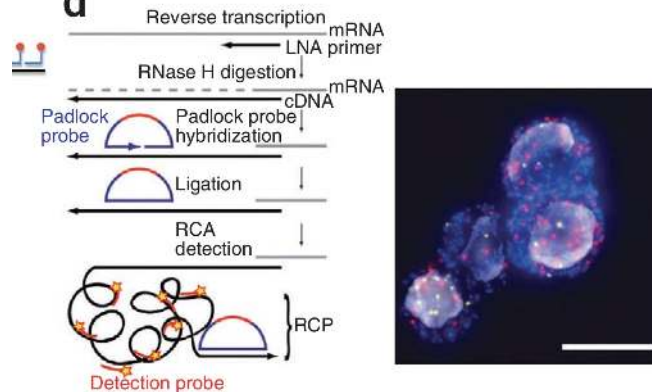
c Many probes, single label ea.



Target: end-1 gene in *C.elegans* embryos

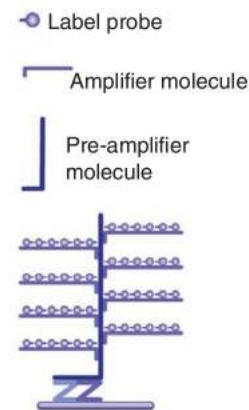
Rolling circle amplification (RCA) of 'padlock probes'.

d Labels hyb to RCA product.

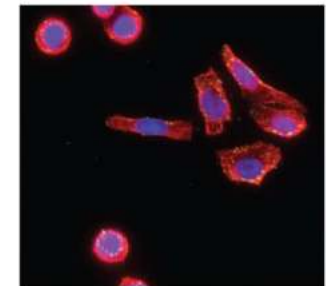


TARGET: ERBB2 (aka. HER2) in human fibroblasts

e



Branched oligo sets that amplify labeling



Target: ERBB2 (green) and 18SrRNA (red)

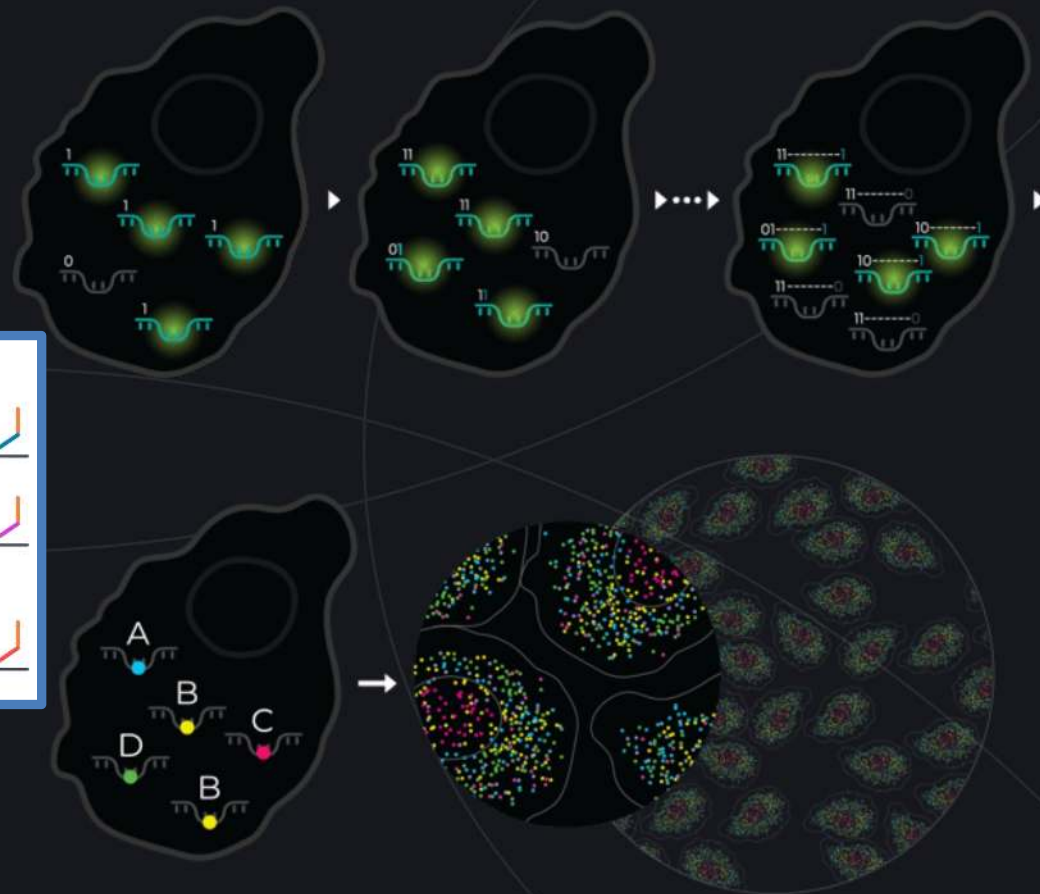
Multiplexed Error-Robust Fluorescence in situ Hybridization

MERFISH is a massively multiplexed single-molecule imaging technology for spatially resolved transcriptomics capable of simultaneously measuring the copy number and spatial distribution of hundreds to tens of thousands of RNA species in individual cells.

COMBINATORIAL LABELING

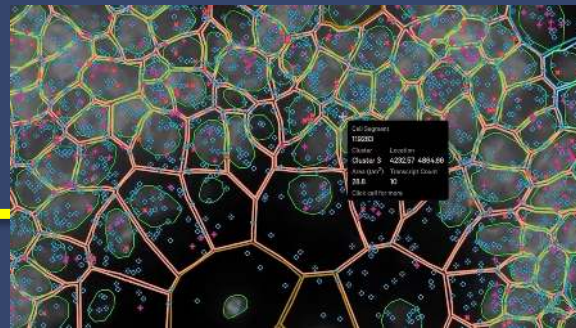
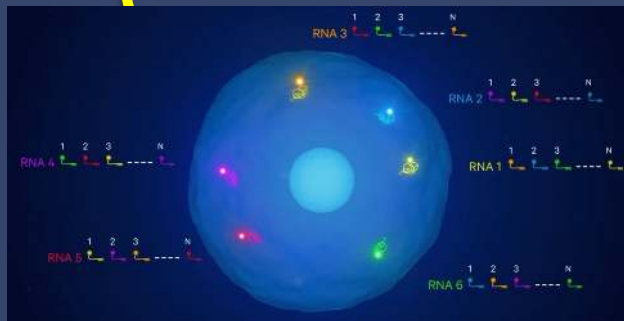
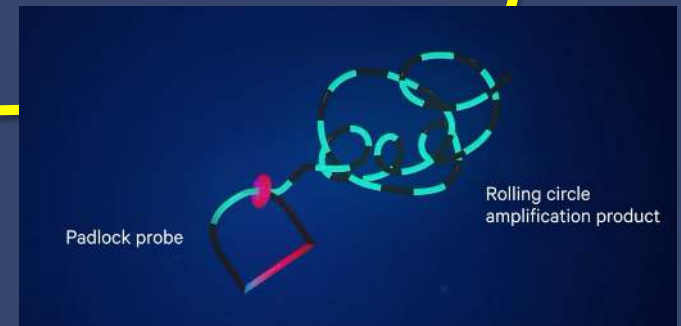
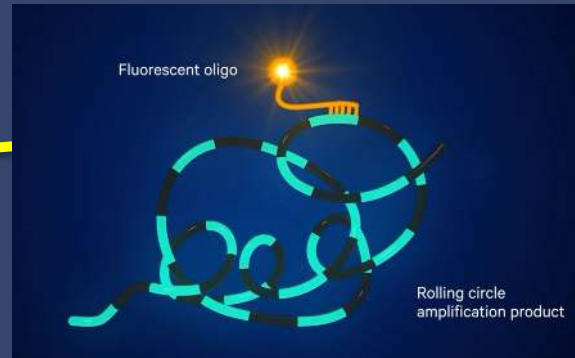
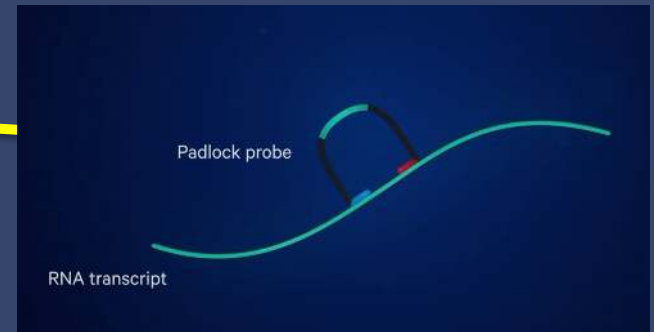
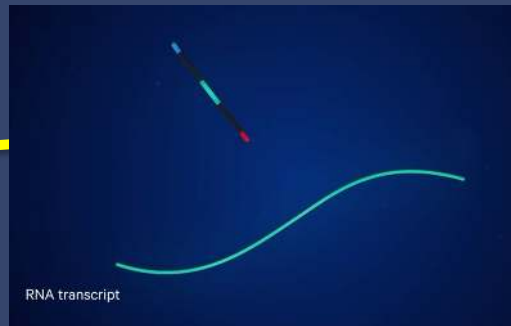
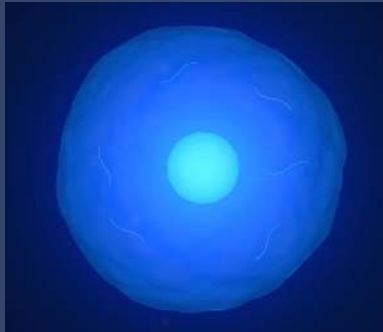
• SEQUENTIAL IMAGING

• ERROR ROBUST BARCODING



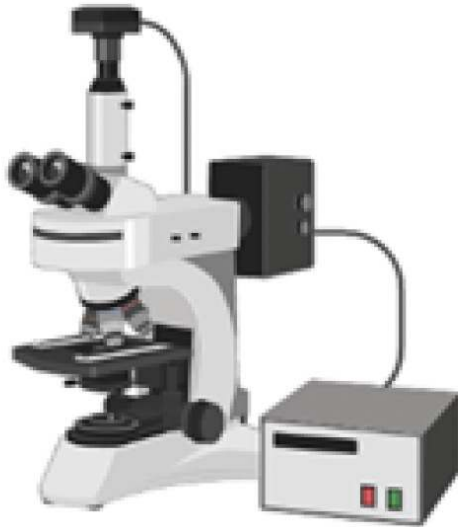
Transcripts	Assigned Barcode	Encoding Probes
Transcript 1	1011000...1	
Transcript 2	0110100...1	
⋮		
Transcript N	0000111...1	

10X Genomics Xenium – 100s to 1000s of Targeted RNAs visualized at subcellular resolution



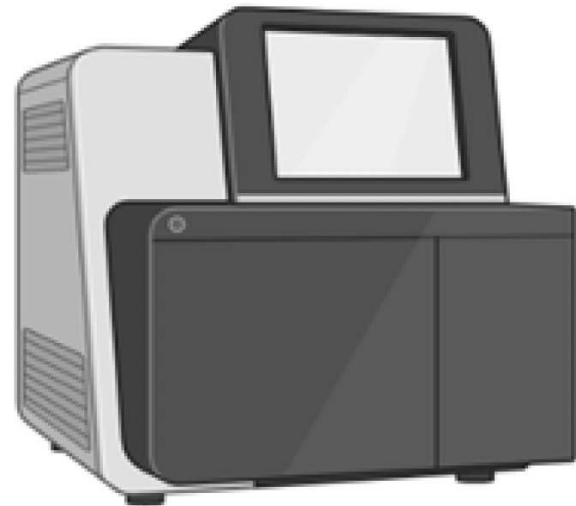
Classes of Spatial Transcriptomics

Imaging Readout

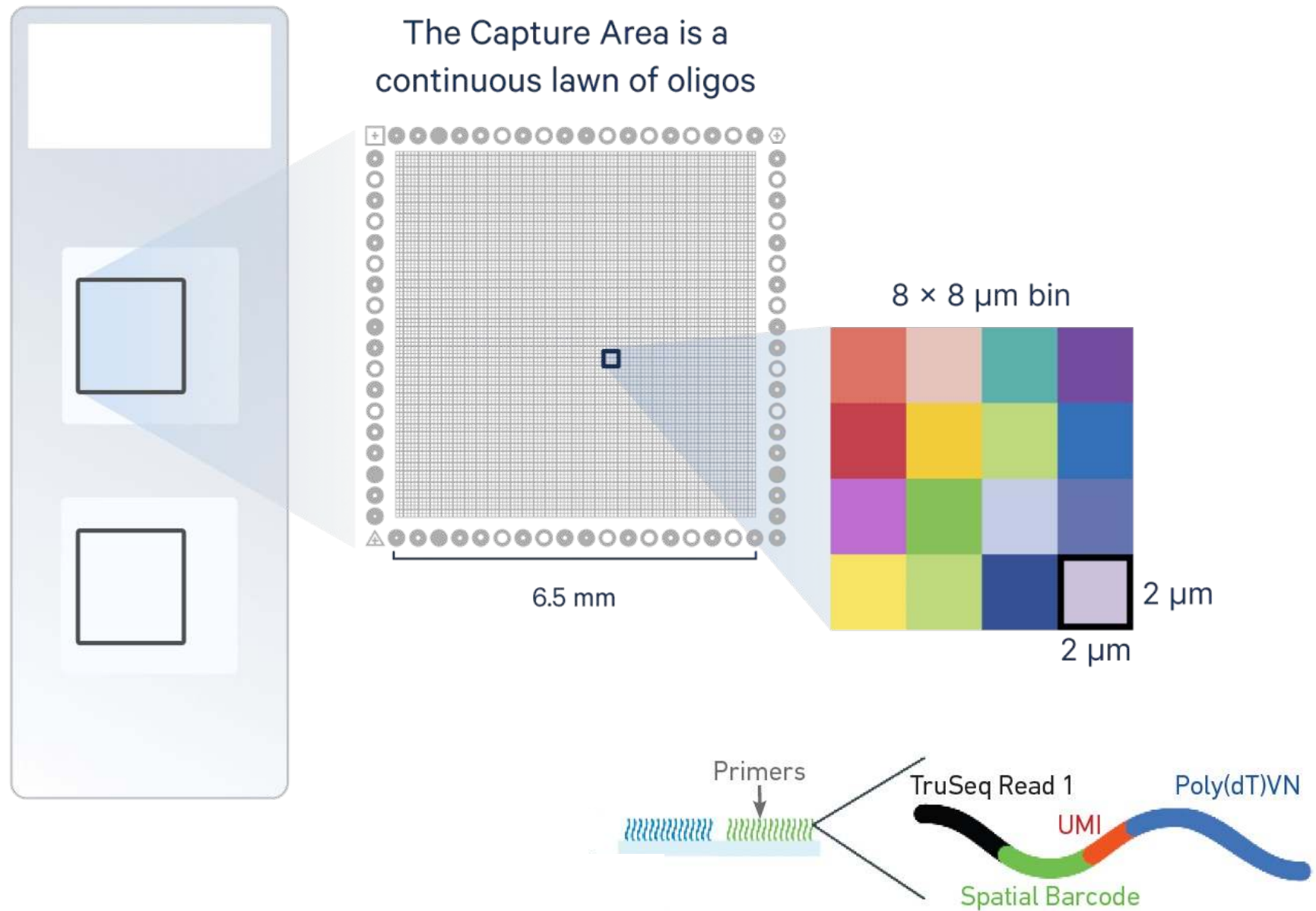


Based on In Situ Hybridization (ISH)
and fluorescent tags

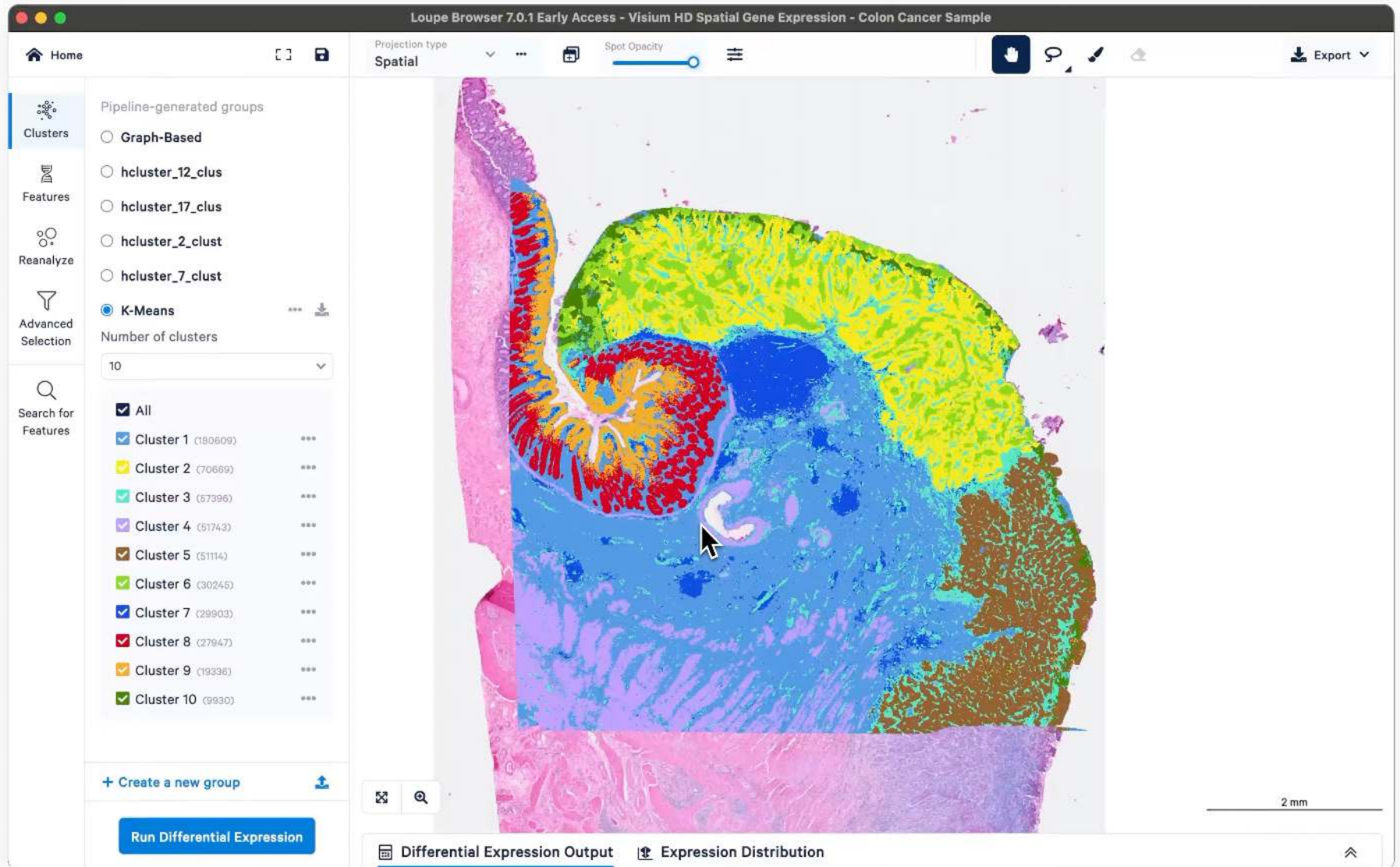
Sequencing Readout



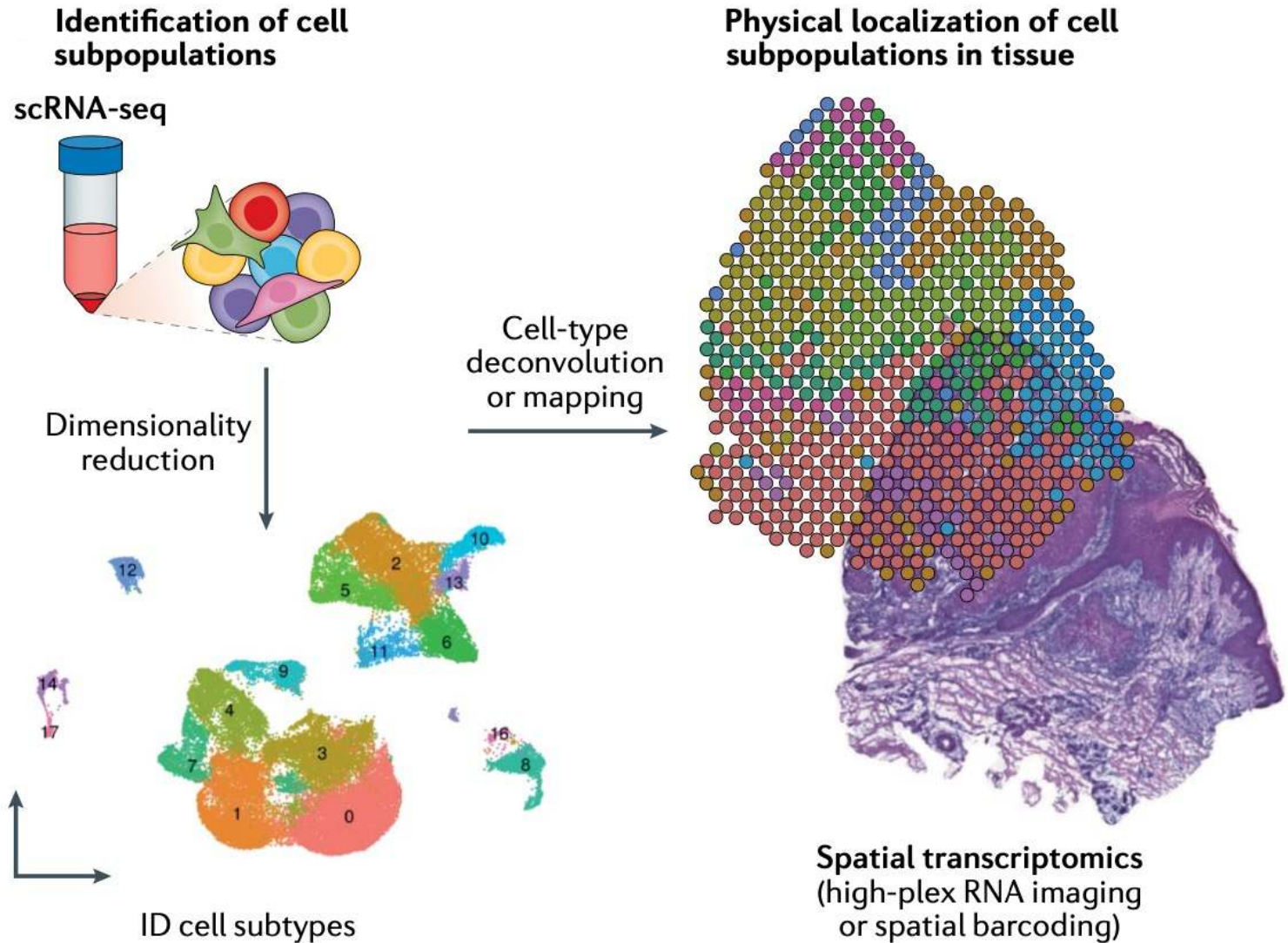
Spatial RNA-seq – 10X Visium HD



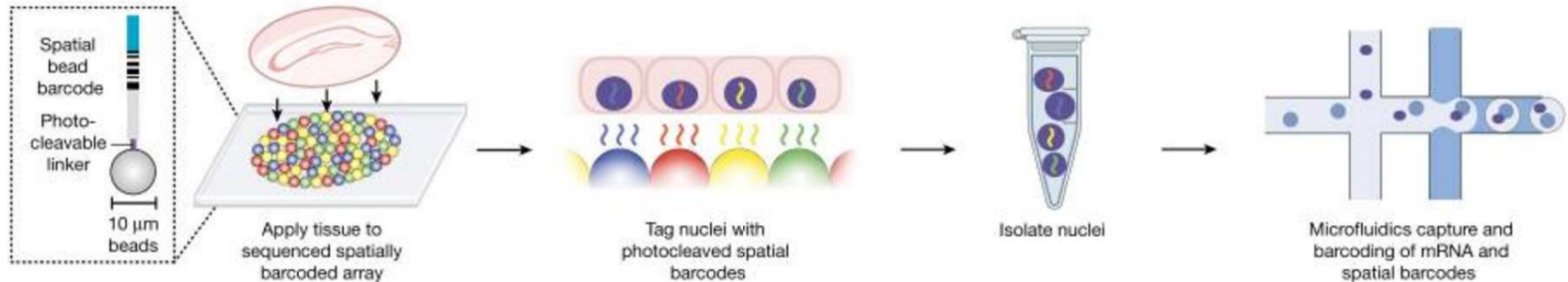
Spatial RNA-seq – 10X Visium HD



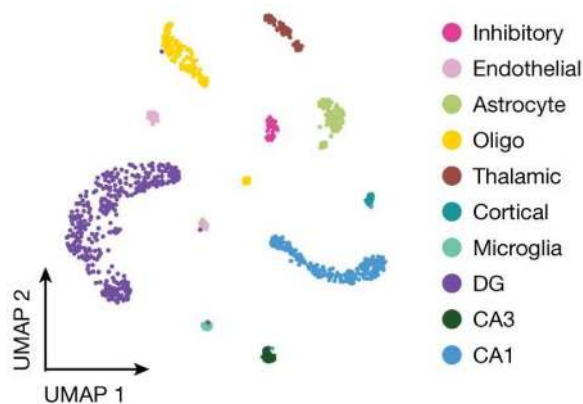
Integration of Single Cell and Spatial Transcriptomes



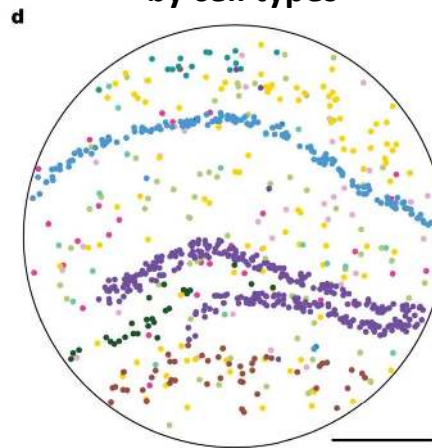
Slide-Tags: integrated single nuclei and spatial transcriptomics



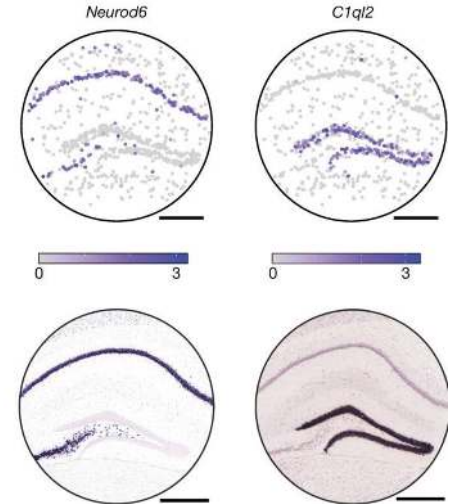
UMAP and cell types,
expression-based clustering



Cells plotted according to
spatial coordinates, colored
by cell types



Spatial Expression of Marker Genes

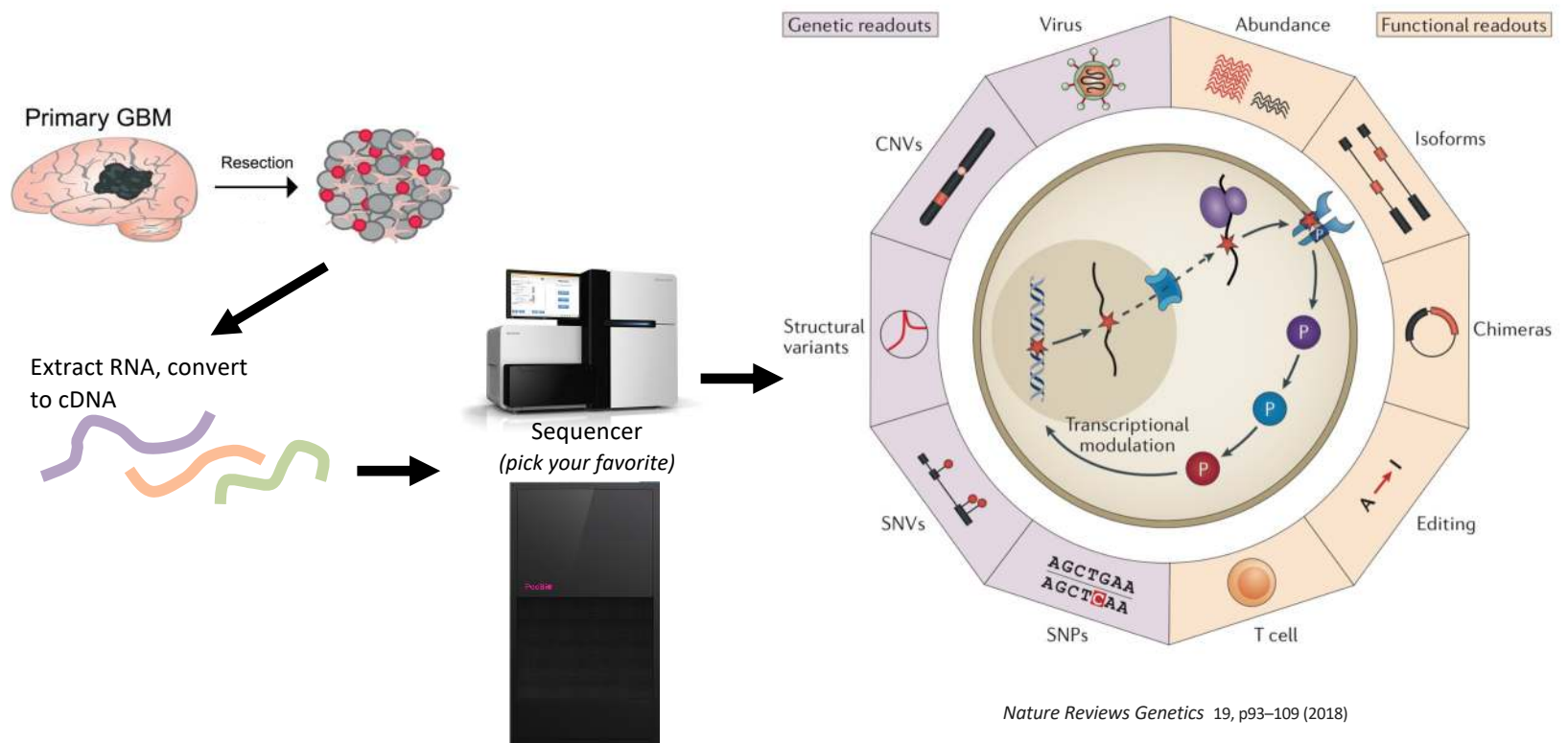


In situ hybridization
(Allen Mouse Brain Atlas)

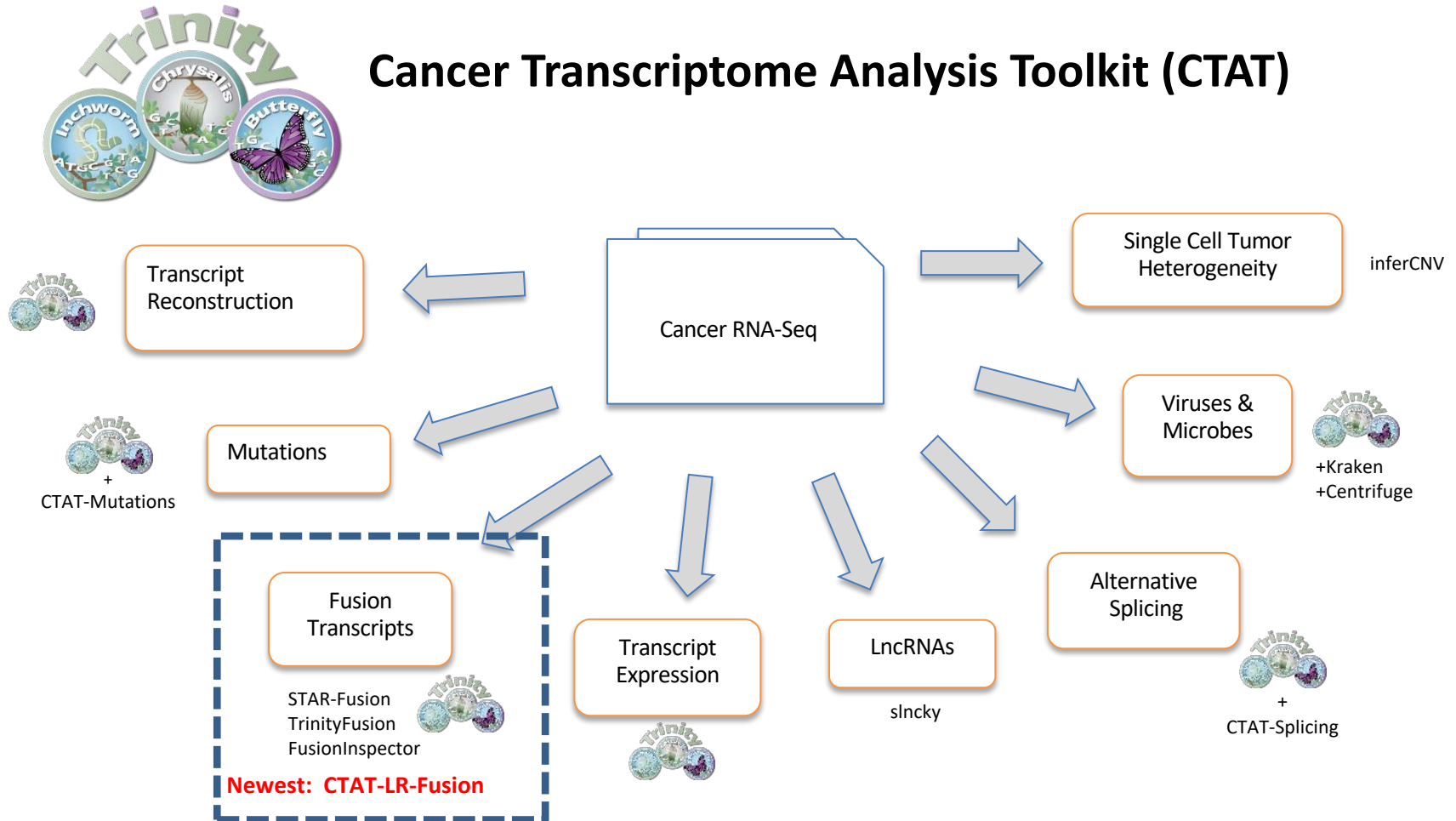


Part 9. Applications in Cancer Transcriptomics

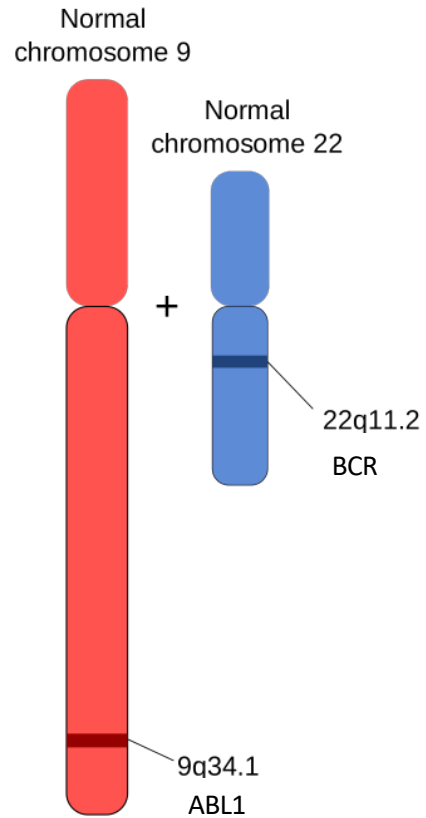
RNA-Seq Empowers Transcriptome Studies of Cancer



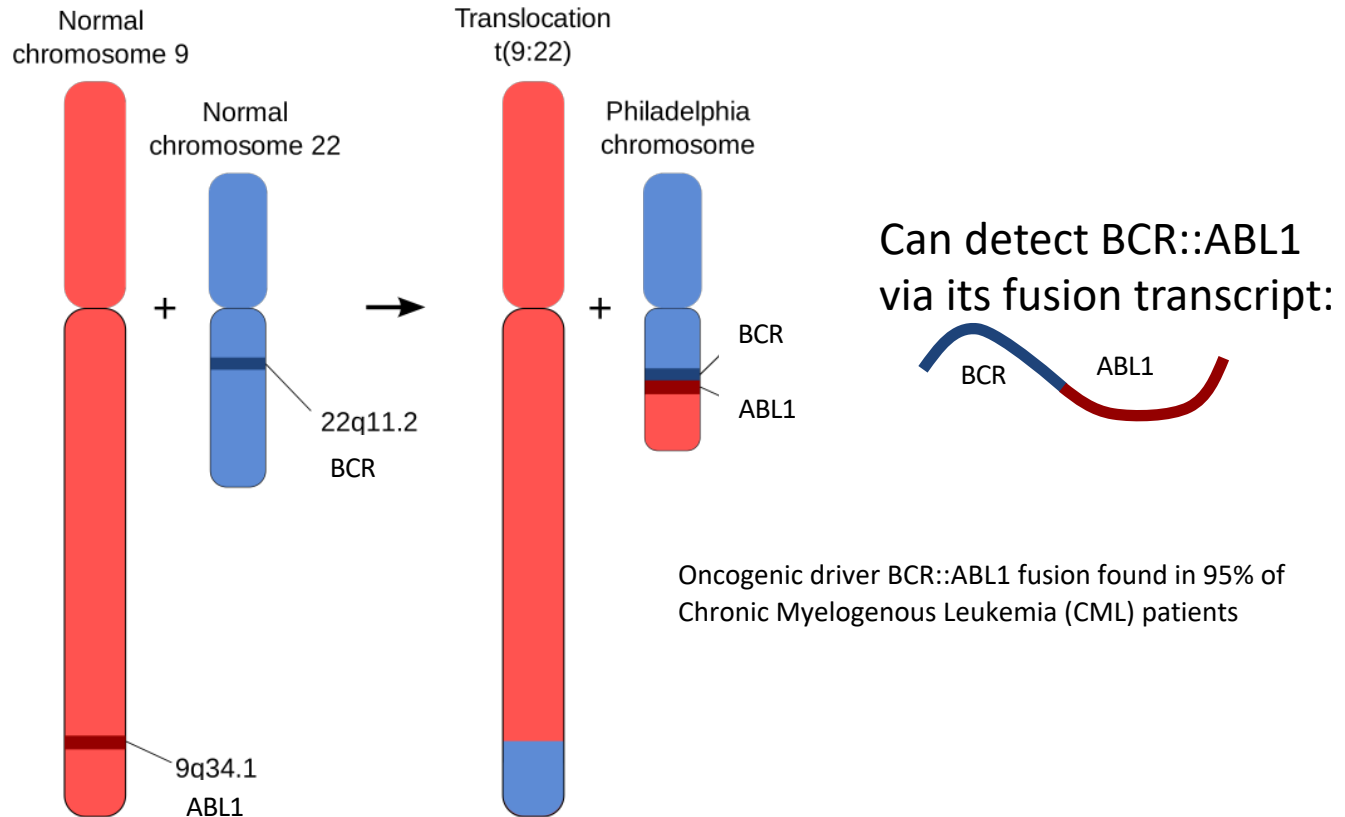
Cancer Transcriptome Analysis Toolkit (CTAT)



Chromosomal Translocations Can Lead to Oncogenic Fusion Transcripts



Chromosomal Translocations Can Lead to Oncogenic Fusion Transcripts



Diagnostics and Therapeutics Involving Oncogenic Fusion Transcripts in Cancer

BCR-ABL1 (Philadelphia chromosome)

- Chronic Myelogenous Leukemia (CML) cases (95% of cases)
- Treatable with tyrosine kinase inhibitors

SS18—SSX

- Synovial sarcoma (~100% of cases)

TMPRSS2-ERG

- Prostate cancers (50% of cases)

EML4-ALK

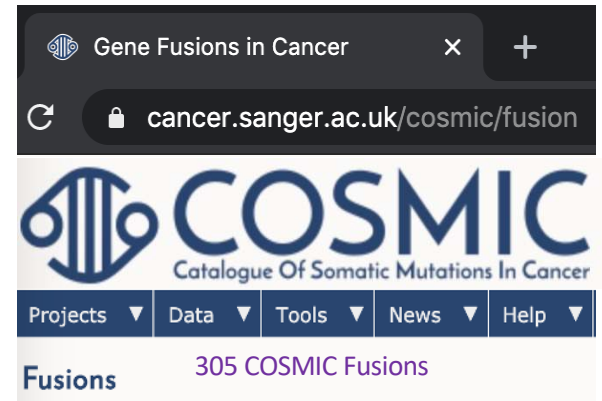
- Non small cell lung carcinoma (4% of cases)
- anaplastic lymphoma kinase (ALK) inhibitors improve patient outcome

DNAJB1-PRKACA

- Fibrolamellar hepatocellular carcinoma (FL-HCC), 100% of cases, but a rare cancer.

FGFR3-TACC3

- ~8% of glioblastoma patients



Diagnostics and Therapeutics Involving Oncogenic Fusion Transcripts in Cancer

BCR-ABL1 (Philadelphia chromosome)

- Chronic Myelogenous Leukemia (CML) cases (95% of cases)
- **Treatable with tyrosine kinase inhibitors**

SS18—SSX

- Synovial sarcoma (~100% of cases)

TMPRSS2-ERG

- Prostate cancers (50% of cases)

EML4-ALK

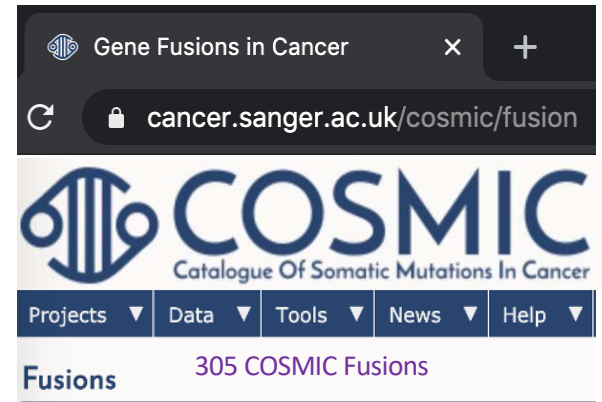
- Non small cell lung carcinoma (4% of cases)
- **anaplastic lymphoma kinase (ALK) inhibitors improve patient outcome**

DNAJB1-PRKACA

- Fibrolamellar hepatocellular carcinoma (FL-HCC), 100% of cases, but a rare cancer.

FGFR3-TACC3

- ~8% of glioblastoma patients



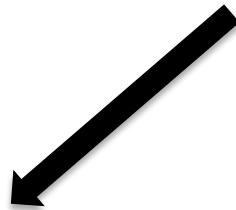
General Approaches to Fusion Transcript Discovery

Paired-end Illumina RNA-Seq

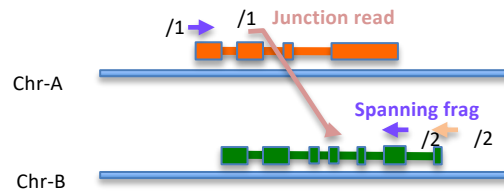


General Approaches to Fusion Transcript Discovery

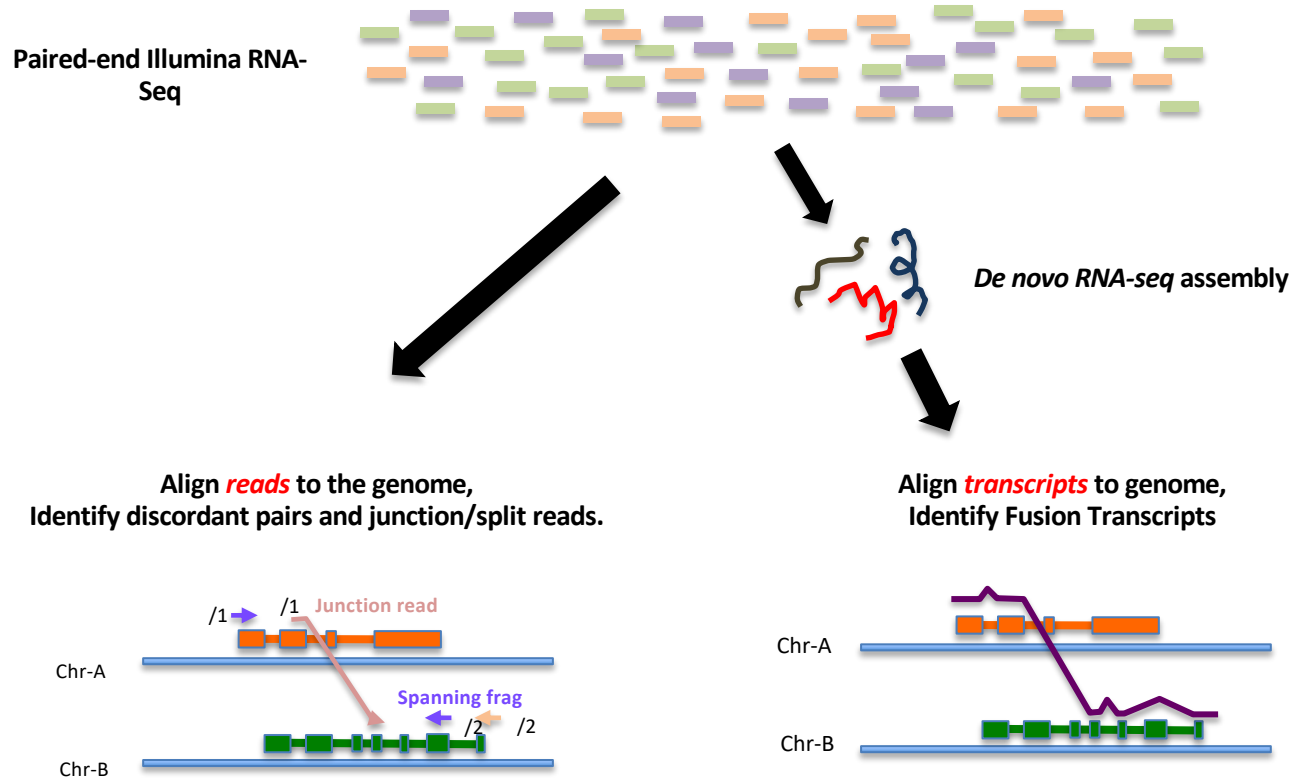
Paired-end Illumina RNA-Seq



Align **reads** to the genome,
Identify discordant pairs and junction/split reads.



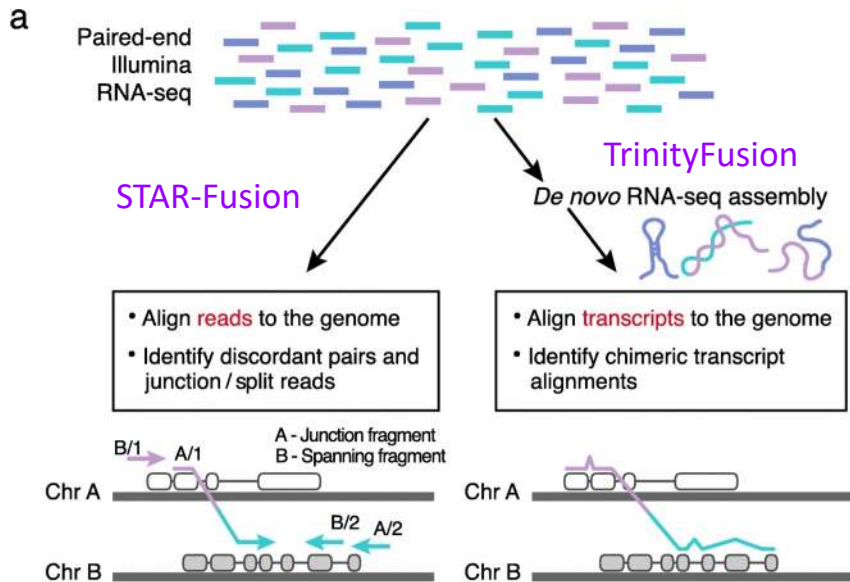
General Approaches to Fusion Transcript Discovery



Our Prior Work on Fusion Detection, Benchmarking, and Analysis via Illumina RNA-seq

Accuracy assessment of fusion transcript detection via read-mapping and de novo
fusion transcript assembly-based methods

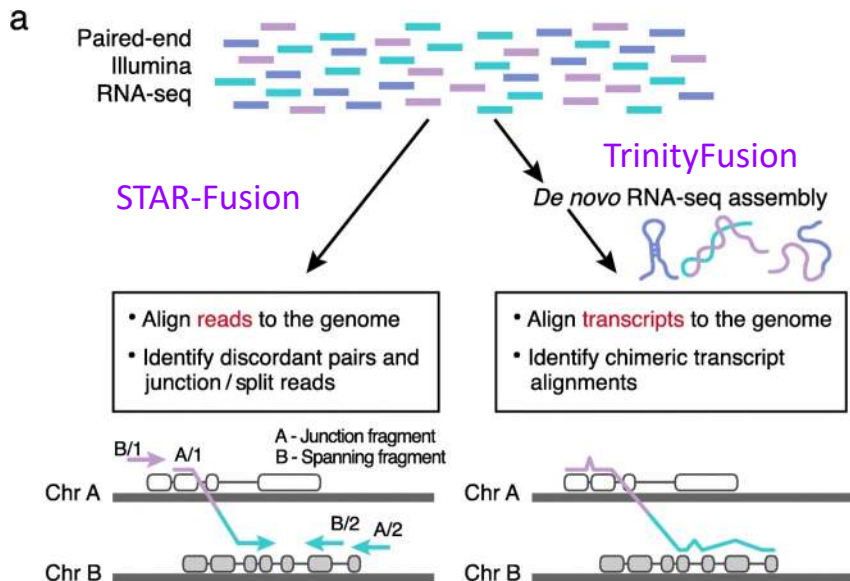
[Genome Biology volume 20, Article number: 213 \(2019\)](#)



Our Prior Work on Fusion Detection, Benchmarking, and Analysis via Illumina RNA-seq

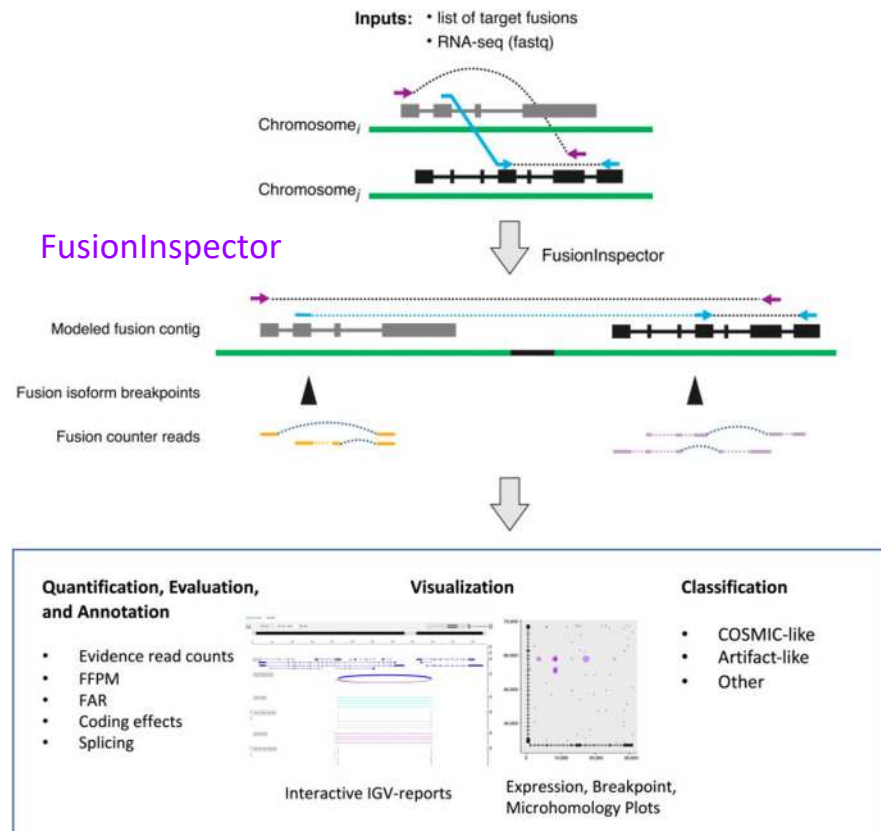
Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods

[Genome Biology volume 20, Article number: 213 \(2019\)](#)



Targeted in silico characterization of fusion transcripts in tumor and normal tissues via FusionInspector

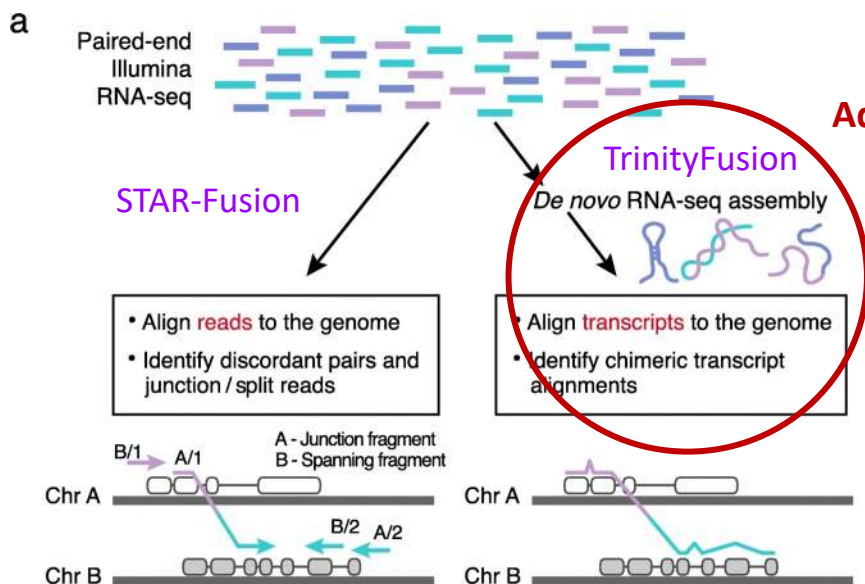
[Cell Rep Methods. 2023 May 8;3\(5\):100467.](#)



Adapting TrinityFusion and FusionInspector to Long Read Fusion Detection

Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods

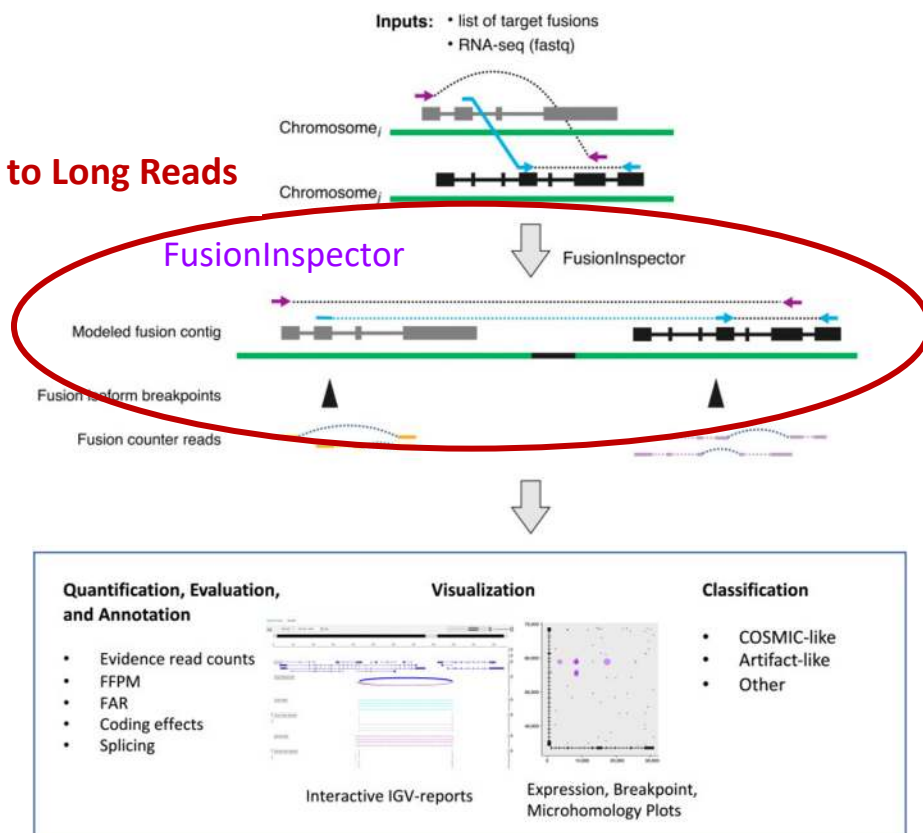
[Genome Biology volume 20, Article number: 213 \(2019\)](#)



Adapt to Long Reads

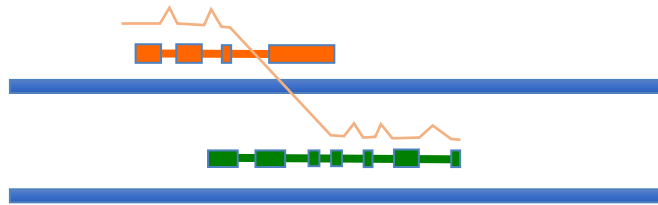
Targeted in silico characterization of fusion transcripts in tumor and normal tissues via FusionInspector

[Cell Rep Methods. 2023 May 8;3\(5\):100467.](#)



New Addition to our **Cancer Transcriptome Analysis Toolkit**: **CTAT-LR-fusion**
(*borrowes general approach from TrinityFusion and FusionInspector, adapted for LR*)

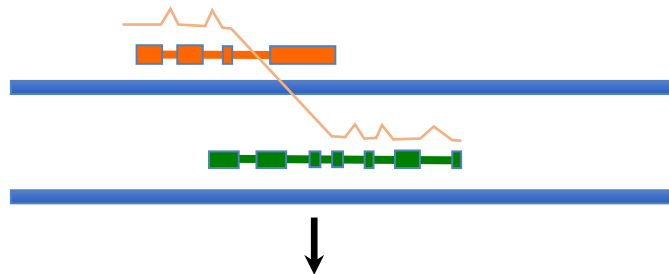
(1) Quickly Identify Fusion Candidate (ctat-minimap2) *4x faster*



*TrinityFusion-
like*

New Addition to our **Cancer Transcriptome Analysis Toolkit**: **CTAT-LR-fusion**
(borrows general approach from *TrinityFusion* and *FusionInspector*, adapted for LR)

- (1) Quickly Identify Fusion Candidate (ctat-minimap2) *4x faster*



TrinityFusion-like

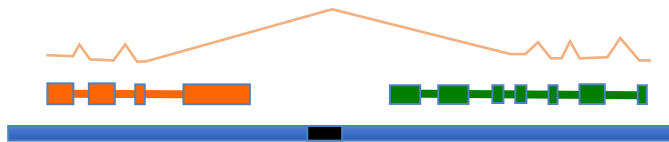
- (2)

Make mini-fusion contigs



FusionInspector-like

Rigorous minimap2-alignment, capture precise breakpoints



CTAT-LR-Fusion Interactive Reports for Visualization and Analysis



Single Cell MAS-Iso-seq Applied to T-cell Enriched Melanoma Patient Sample

Long and Short scRNA-seq

~ 20M PacBio MAS-Iso-seq reads

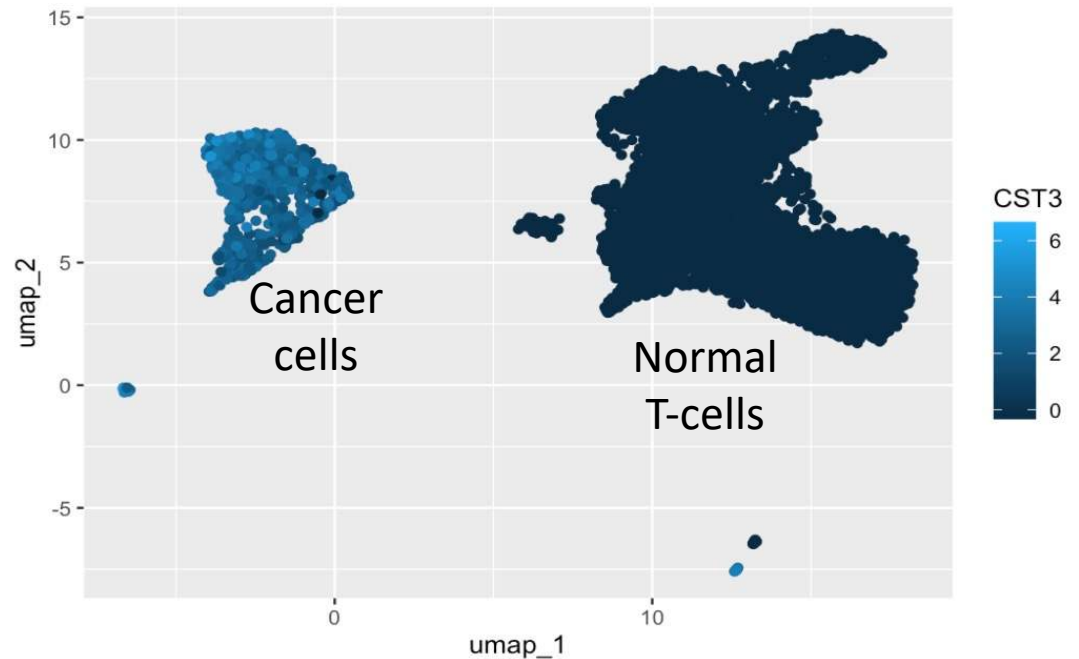
~ 200M Illumina 10x 3' reads

~ 7k Total cells (10% cancer cells)

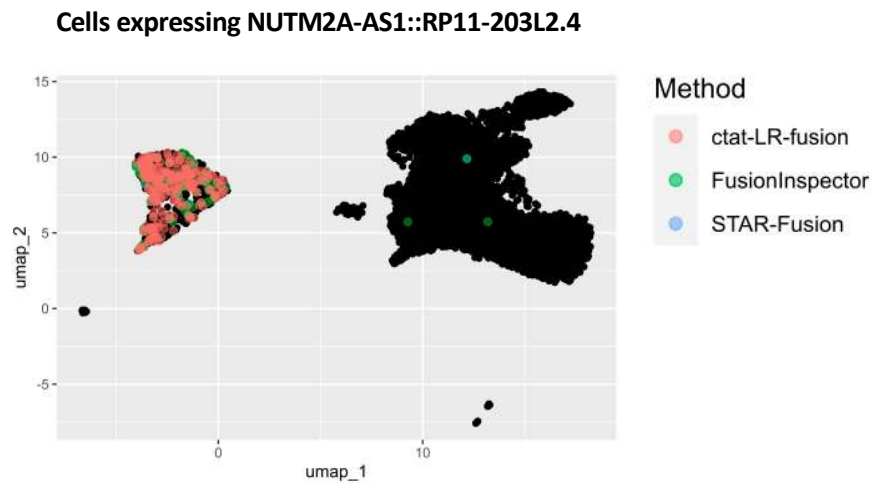
Brief Communication | [Published: 08 June 2023](#)

High-throughput RNA isoform sequencing using programmed cDNA concatenation

[Aziz M. Al'Khafaji](#) ✉, [Jonathan T. Smith](#), [Kiran V. Garimella](#) ✉, [Mehrtash Babadi](#) ✉, [Victoria Popic](#) ✉, [Moshe Sade-Feldman](#), [Michael Gatzen](#), [Siranush Sarkizova](#), [Marc A. Schwartz](#), [Emily M. Blaum](#), [Allyson Day](#), [Maura Costello](#), [Tera Bowers](#), [Stacey Gabriel](#), [Eric Banks](#), [Anthony A. Philippakis](#), [Genevieve M. Boland](#), [Paul C. Blainey](#) ✉ & [Nir Hacohen](#) ✉

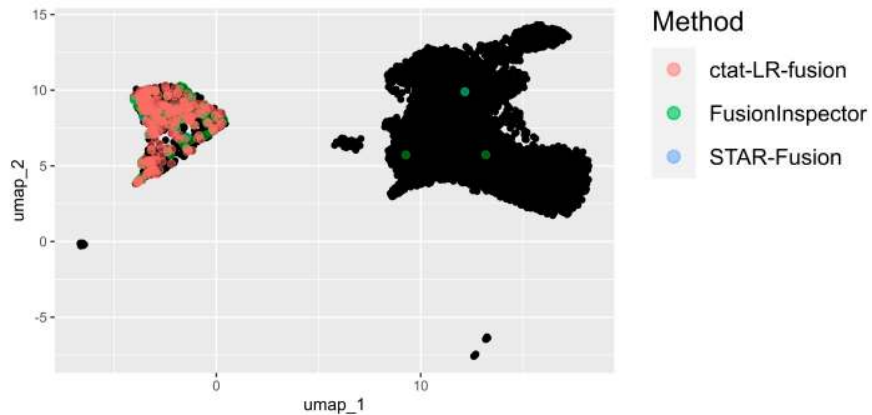


Single Tumor-specific Fusion Transcript Detected: NUTM2A-AS1 (Oncogene) :: RP11-203L2.4

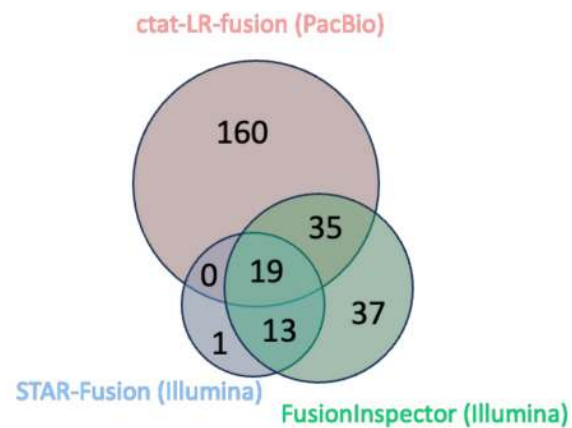


Single Tumor-specific Fusion Transcript Detected: NUTM2A-AS1 (Oncogene) :: RP11-203L2.4

Cells expressing NUTM2A-AS1::RP11-203L2.4

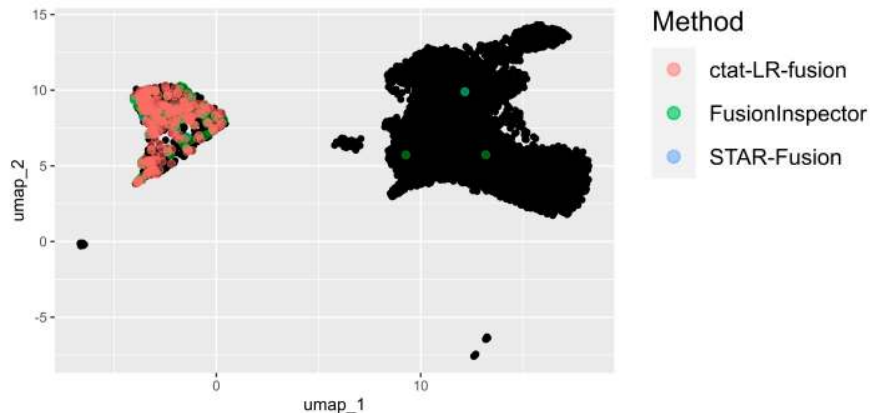


Cells Identified with NUTM2A-AS1::RP11-203L2.4

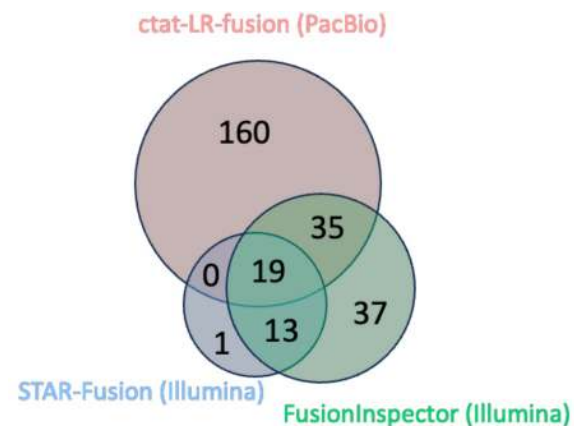


Single Tumor-specific Fusion Transcript Detected: NUTM2A-AS1 (Oncogene) :: RP11-203L2.4

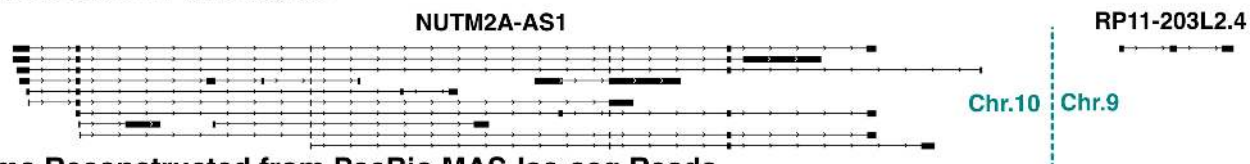
Cells expressing NUTM2A-AS1::RP11-203L2.4



Cells Identified with NUTM2A-AS1::RP11-203L2.4



Reference Isoform Structures



Isoforms Reconstructed from PacBio MAS-Iso-seq Reads



8 full-length fusion isoforms

Fusion Isoform Junctions Detected from Short Illumina Reads



5 sets of isoform breakpoints

Detection of Fusion Transcripts in High Grade Serous Ovarian Cancer via Long Read Isoform Sequencing

Detection of isoforms and genomic alterations by high-throughput full-length single-cell RNA sequencing in ovarian cancer

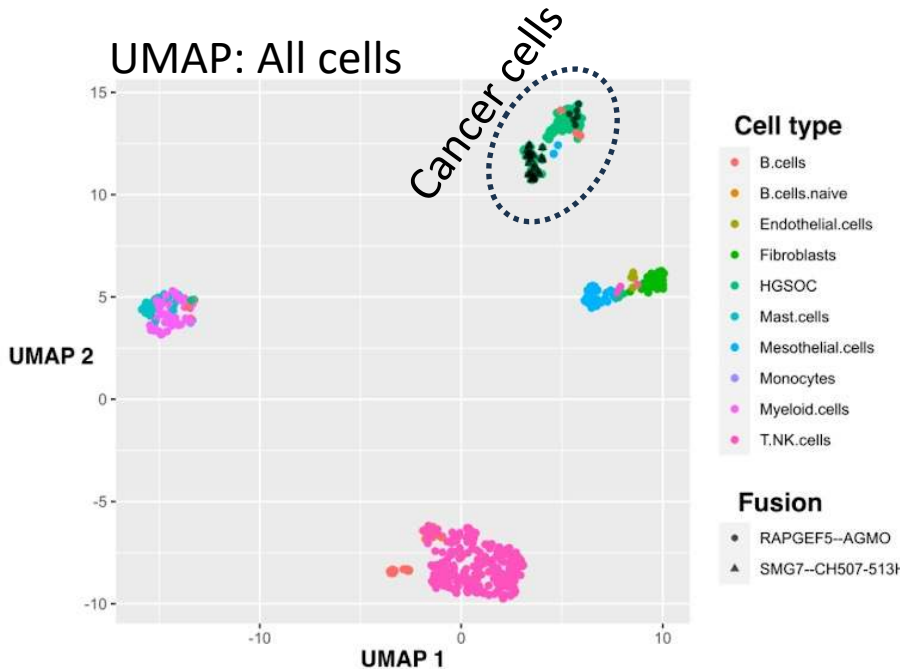
[Arthur Dondi](#), [Ulrike Lischetti](#), [Francis Jacob](#), [Franziska Singer](#), [Nico Borgsmüller](#), [Ricardo Coelho](#),
[Tumor Profiler Consortium](#), [Viola Heinzlmann-Schwarz](#), [Christian Beisel](#) & [Niko Beerenwinkel](#)

Nature Communications 14, Article number: 7780 (2023) | [Cite this article](#)

Patient 1:

- 54M PacBio Isoform reads
- 35M Illumina 10x 3' reads
- ~500 cells (20% cancer)
- 4 cancer-specific fusions detected

UMAP: All cells



Detection of Fusion Transcripts in High Grade Serous Ovarian Cancer via Long Read Isoform Sequencing

Detection of isoforms and genomic alterations by high-throughput full-length single-cell RNA sequencing in ovarian cancer

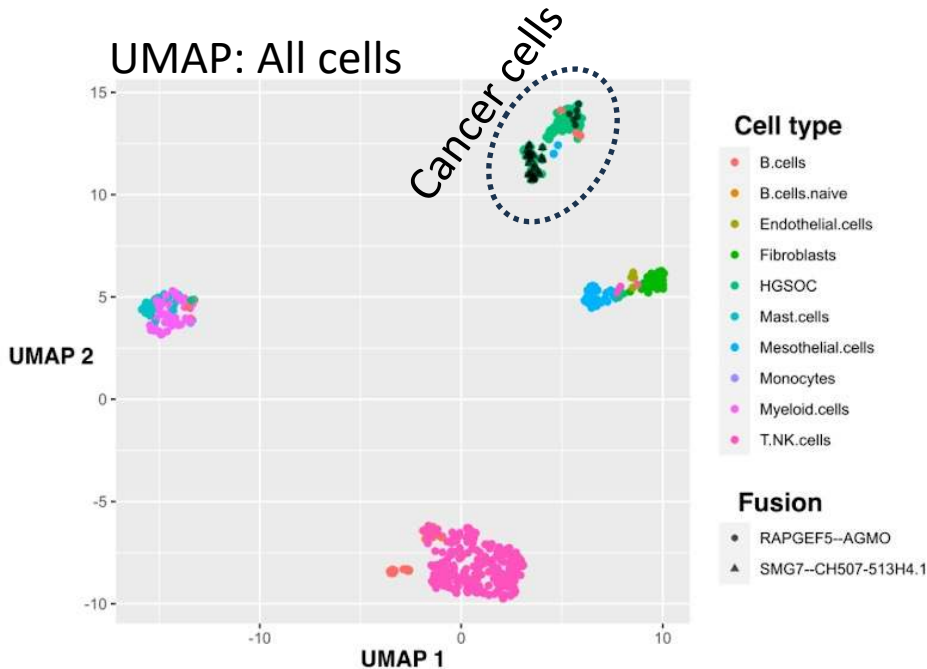
[Arthur Dondi](#), [Ulrike Lischetti](#), [Francis Jacob](#), [Franziska Singer](#), [Nico Borgsmüller](#), [Ricardo Coelho](#),
[Tumor Profiler Consortium](#), [Viola Heinzlmann-Schwarz](#), [Christian Beisel](#) & [Niko Beerenwinkel](#)

Nature Communications 14, Article number: 7780 (2023) | [Cite this article](#)

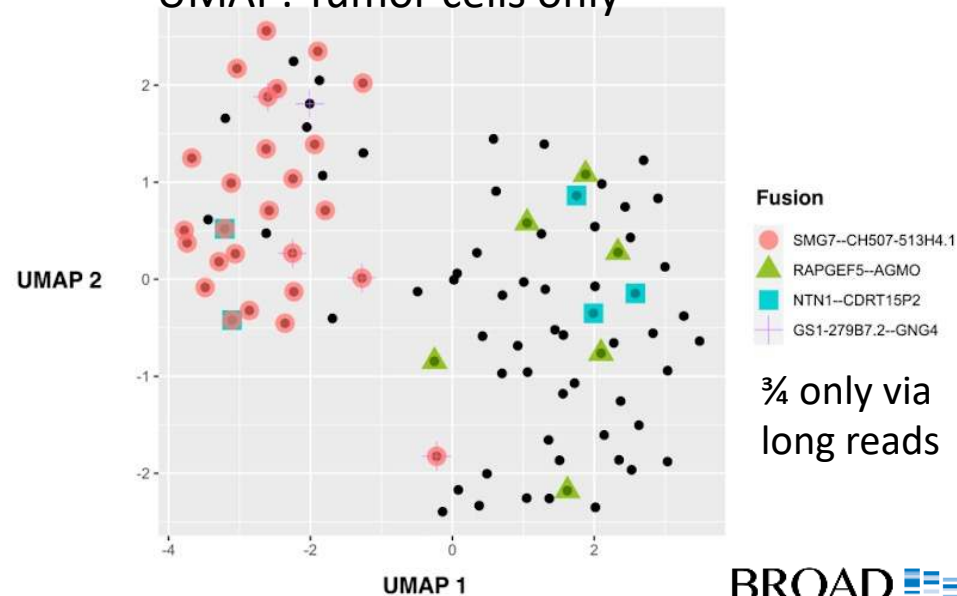
Patient 1:

- 54M PacBio Isoform reads
- 35M Illumina 10x 3' reads
- ~500 cells (20% cancer)
- 4 cancer-specific fusions detected

UMAP: All cells



UMAP: Tumor cells only



In Summary

- Many applications for RNA-seq, technology continues to evolve.
- Analysis can involve reference genomes or be genome-free via de novo transcriptome assembly – Trinity can help.
- Quantification involves counting reads and considering read-mapping uncertainty
- Long reads now available for applications previously limited to short reads, involve far less read mapping uncertainty, and enable isoform rather than gene expression analyses.
- Single cell and spatial transcriptomics studies are revolutionizing our understanding of tissue complexity, diversity of cell types, and cellular interactions - particularly in studies of cancer.
- Massive resources being built - whole organism cell atlases and high-resolution spatial maps, and new software tools and algorithms developed for leveraging long reads in bulk, single cell, and spatial studies.