RNAseq and Annotation

Isheng Jason Tsai

Introduction to NGS Data and Analysis
Lecture 6 [v2020]

Lecture outline

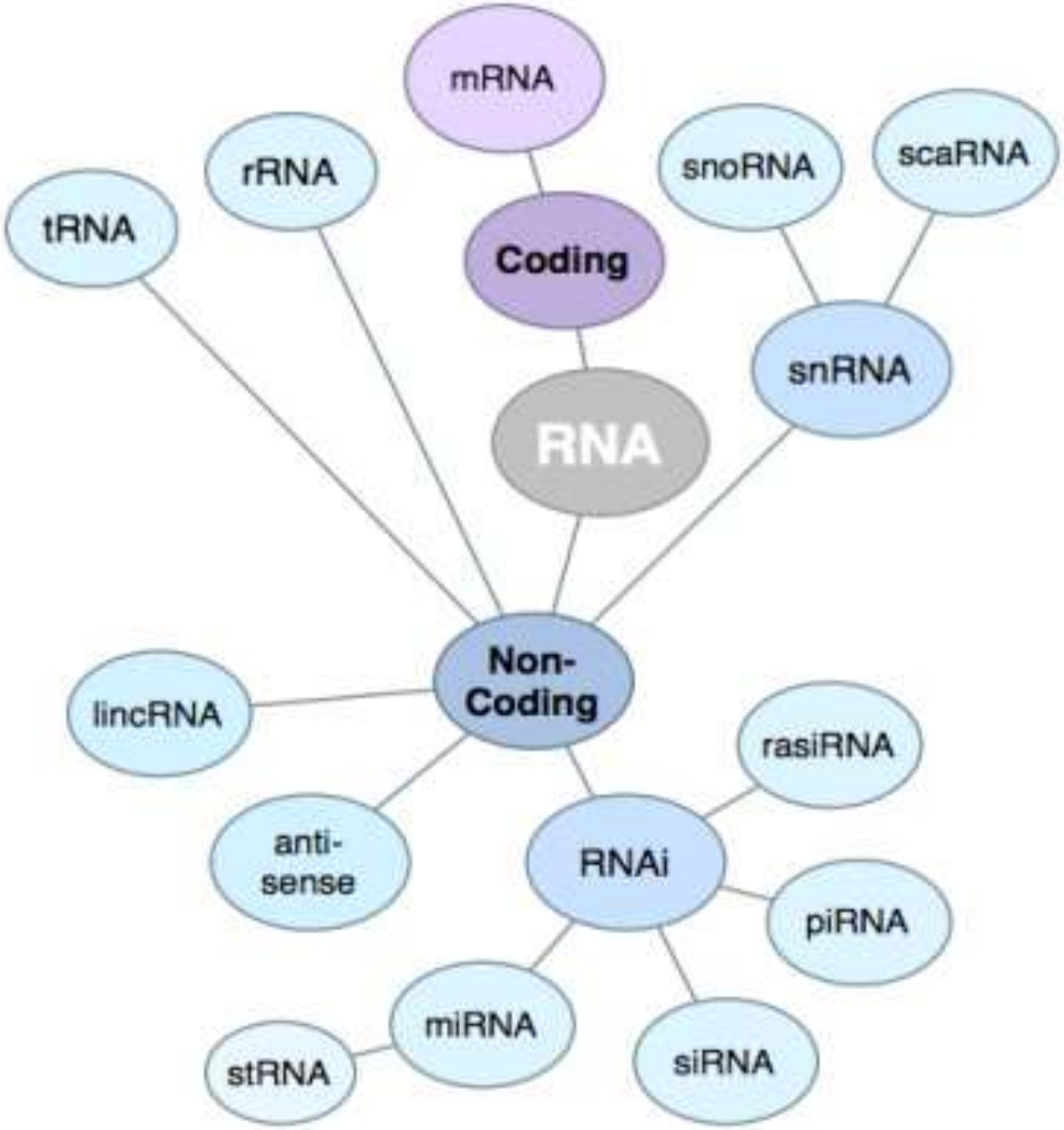
What I will cover today

- mRNAseq (for **bulk** RNAseq)
- mapping
- assembly / reconstruction
- annotation
- experimental design
- differential expression
- single cell genomics

What I won't cover due to time constraints but equally important

- *pseudoalignment* (kallisto, salmon etc)
- noncoding RNAs
- long read mapping

Types of RNA



http://finchtalk.blogspot.com/2009/05/small-rnas-get-smaller.html

Type	Percent of total RNA by mass	Molecules per cell	Average size (kb)	Total weight picograms/cell	Notes	Reference
rRNAs	80 to 90	3–10 × 10^6 (ribosomes)	6.9	10 to 30		Blobel and Potter (1967), Wolf and Schlessinger (1977), Duncan and Hershey (1983)
tRNA	10 to 15	3–10 × 10^7	<0.1	1.5 to 5	About 10 tRNA molecules /ribosome	Waldron and Lacroute (1975)
mRNA	3 to 7	3–10 × 10^5	1.7	0.25 to 0.9		Hastie and Bishop (1976), Carter et al. (2005)
hnRNA (pre-mRNA)	0.06 to 0.2	1–10 × 10^3	10*	0.004 to 0.03	Estimated at 2–4% of mRNA by weight	Mortazavi et al. (2008), Menet et al. (2012)
Circular RNA	0.002 to 0.03	3–20 × 10^3	~0.5	0.0007 to 0.005	Estimated at 0.1–0.2% of mRNA**	Salzman et al. (2012), Guo et al. (2014)
snRNA	0.02 to 0.3	1–5 × 10^5	0.1–0.2	0.008 to 0.04		Kiss and Filipowicz (1992), Castle et al. (2010)
snoRNA	0.04 to 0.2	2–3 × 10^5	0.2	0.02 to 0.03		Kiss and Filipowicz (1992), Cooper (2000), Castle et al. (2010)
miRNA	0.003 to 0.02	1–3 × 10^5	0.02	0.001 to 0.003	About 10^5 molecules per 10 pg total RNA	Bissels et al. (2009)
7SL	0.01 to 0.2	3–20 × 10^4	0.3	0.005 to 0.03	About 1–2 SRP molecules/100 ribosomes	Raue et al. (2007), Castle et al. (2010)
Xist	0.0003 to 0.02	0.1–2 × 10^3	2.8	0.0001 to 0.003		Buzin et al. (1994), Castle et al. (2010)
Other lncRNA	0.03 to 0.2	3–50 × 10^3	1	0.002 to 0.03	Estimated at 1–4% of mRNA by weight	Mortazavi et al. (2008), Ramsköld et al. (2009), Menet et al. (2012)

The size for the average unspliced pre-mRNA is 17 kb; however, most pre-mRNAs are partially spliced at any given time, and the average size of hnRNA is estimated at 10 kb (Salditt-Georgieff et al., 1976).

***Based on the finding that 1–2% of all mRNA species generate circular RNA, which is present at 10% of the level of the parental mRNA.*

Palazzo and Lee (2015) Front. Genet.,

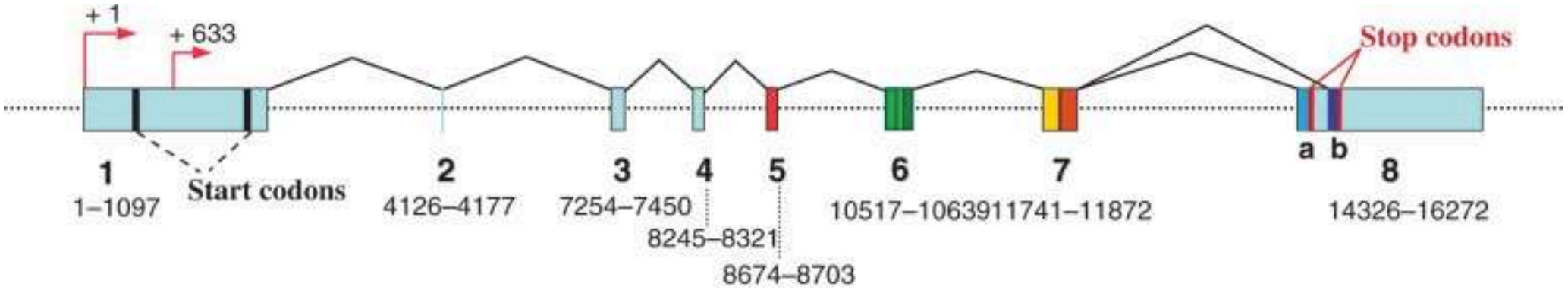
Type	Percent of total RNA by mass	Molecules per cell	Average size (kb)	Total weight picograms/cell	Notes	Reference
rRNAs	80 to 90	3–10 × 10^6 (ribosomes)	6.9	10 to 30		Blobel and Potter (1967), Wolf and Schlessinger (1977), Duncan and Hershey (1983)
tRNA	10 to 15	3–10 × 10^7	<0.1	1.5 to 5	About 10 tRNA molecules /ribosome	Waldron and Lacroute (1975)
mRNA	3 to 7	3–10 × 10^5	1.7	0.25 to 0.9		Hastie and Bishop (1976), Carter et al. (2005)
hnRNA (pre-mRNA)	0.06 to 0.2	1–10 × 10^3	10*	0.004 to 0.03	Estimated at 2–4% of mRNA by weight	Mortazavi et al. (2008), Menet et al. (2012)
Circular RNA	0.002 to 0.03	3–20 × 10^3	~0.5	0.0007 to 0.005	Estimated at 0.1–0.2% of mRNA**	Salzman et al. (2012), Guo et al. (2014)
snRNA	0.02 to 0.3	1–5 × 10^5	0.1–0.2	0.008 to 0.04		Kiss and Filipowicz (1992), Castle et al. (2010)
snoRNA	0.04 to 0.2	2–3 × 10^5	0.2	0.02 to 0.03		Kiss and Filipowicz (1992), Cooper (2000), Castle et al. (2010)
miRNA	0.003 to 0.02	1–3 × 10^5	0.02	0.001 to 0.003	About 10^5 molecules per 10 pg total RNA	Bissels et al. (2009)
7SL	0.01 to 0.2	3–20 × 10^4	0.3	0.005 to 0.03	About 1–2 SRP molecules/100 ribosomes	Raue et al. (2007), Castle et al. (2010)
Xist	0.0003 to 0.02	0.1–2 × 10^3	2.8	0.0001 to 0.003		Buzin et al. (1994), Castle et al. (2010)
Other lncRNA	0.03 to 0.2	3–50 × 10^3	1	0.002 to 0.03	Estimated at 1–4% of mRNA by weight	Mortazavi et al. (2008), Ramsköld et al. (2009), Menet et al. (2012)

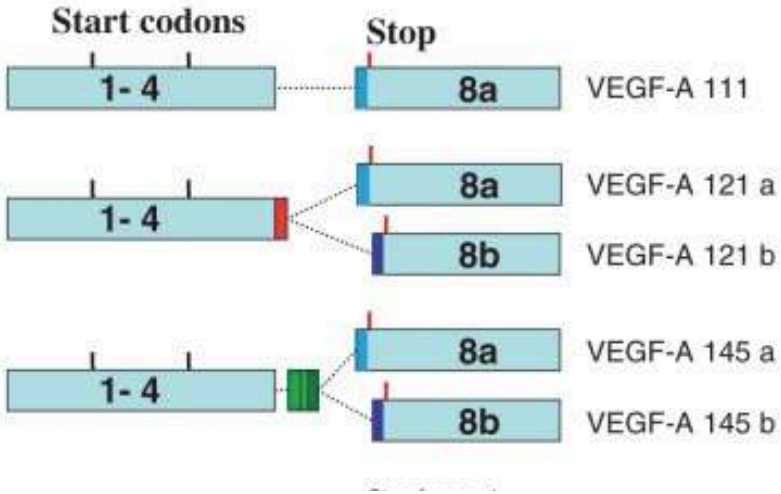
The size for the average unspliced pre-mRNA is 17 kb; however, most pre-mRNAs are partially spliced at any given time, and the average size of hnRNA is estimated at 10 kb (Salditt-Georgieff et al., 1976).

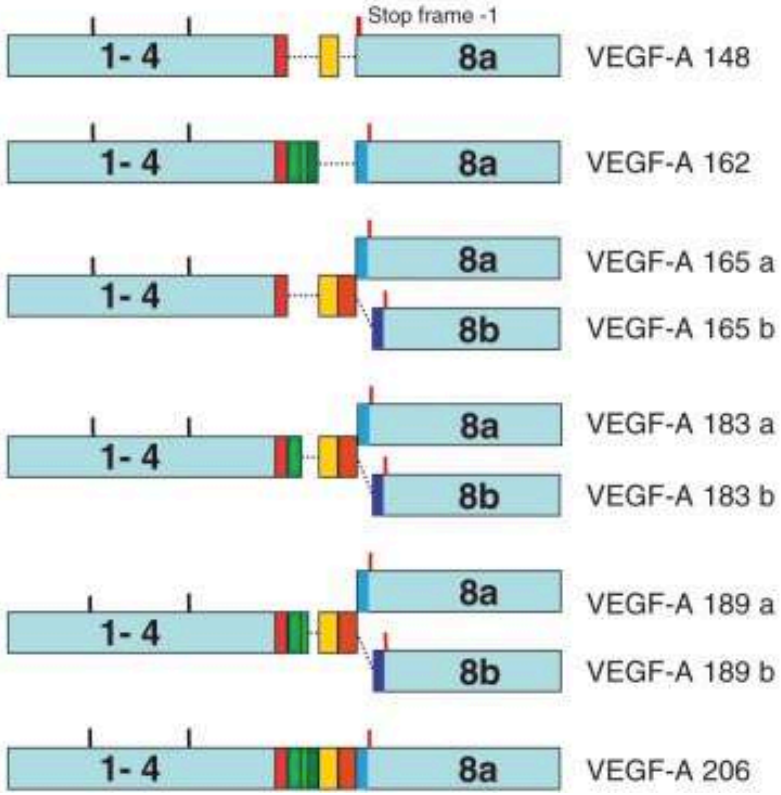
***Based on the finding that 1–2% of all mRNA species generate circular RNA, which is present at 10% of the level of the parental mRNA.*

Palazzo and Lee (2015) Front. Genet.,

Gene and isoforms







Arcondeguy et al (2015)

Typical RNAseq Workflow

1. Experiment / Generate data
2. Map or Assemble
3. Count / Differential expression
4. Analysis

Conesa *et al. Genome Biology* (2016) 17:13
DOI 10.1186/s13059-016-0881-8

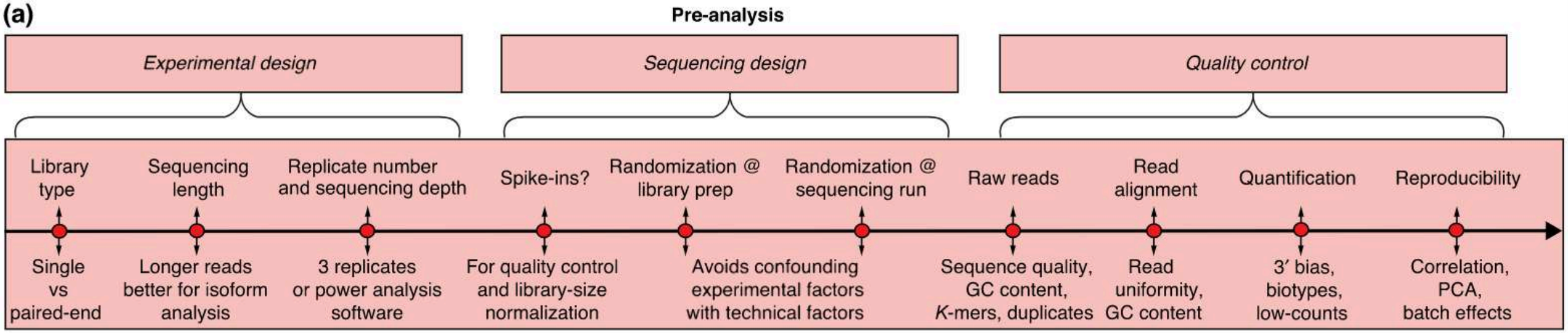
Genome Biology

REVIEW Open Access

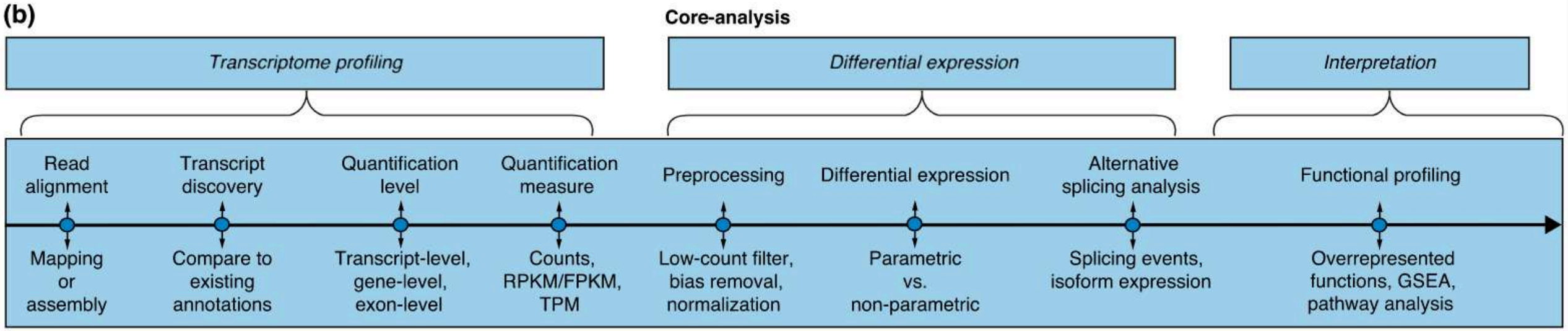
A survey of best practices for RNA-seq data analysis

Ana Conesa[1,2*], Pedro Madrigal[3,4*], Sonia Tarazona[2,5], David Gomez-Cabrero[6,7,8,9], Alejandra Cervera[10], Andrew McPherson[11], Michał Wojciech Szcześniak[12], Daniel J. Gaffney[3], Laura L. Elo[13], Xuegong Zhang[14,15] and Ali Mortazavi[16,17*]

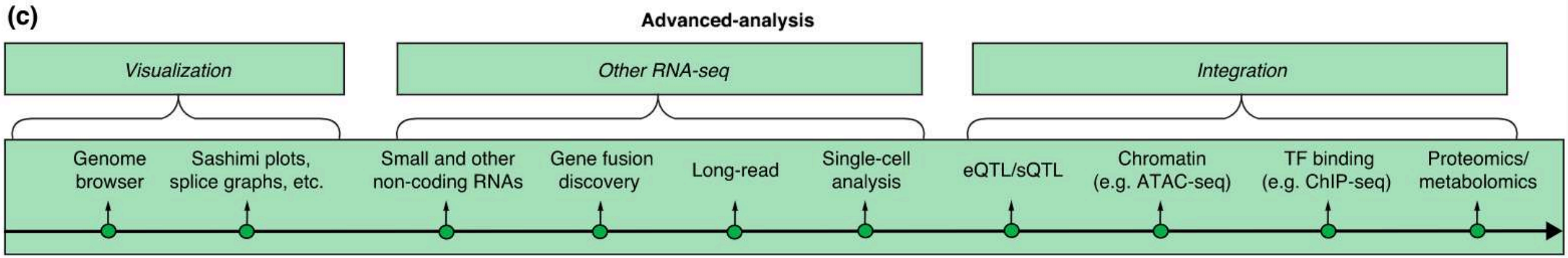
Pre-analysis



Core-analysis

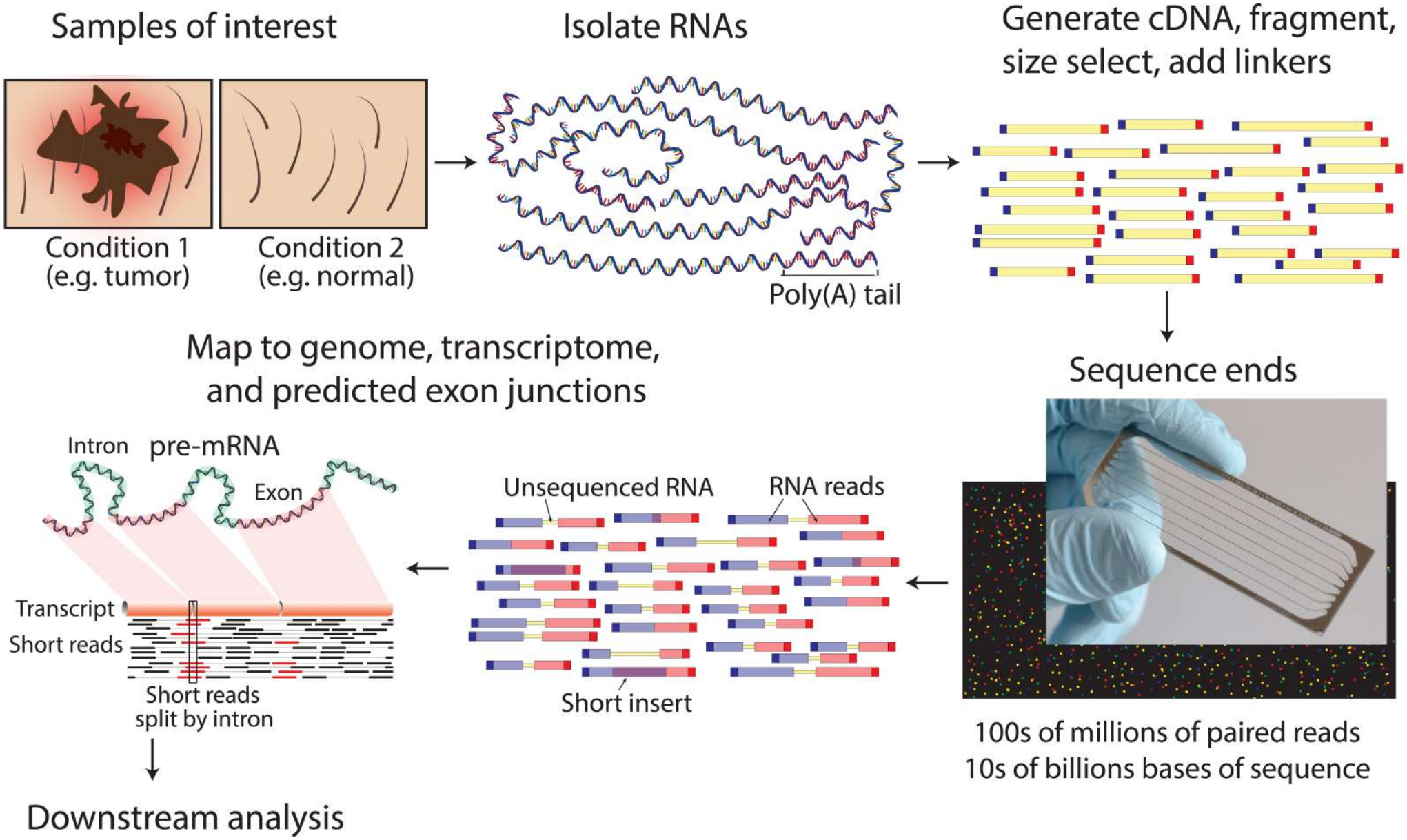


Advanced-analysis (not covered in this lecture but should be mentioned)



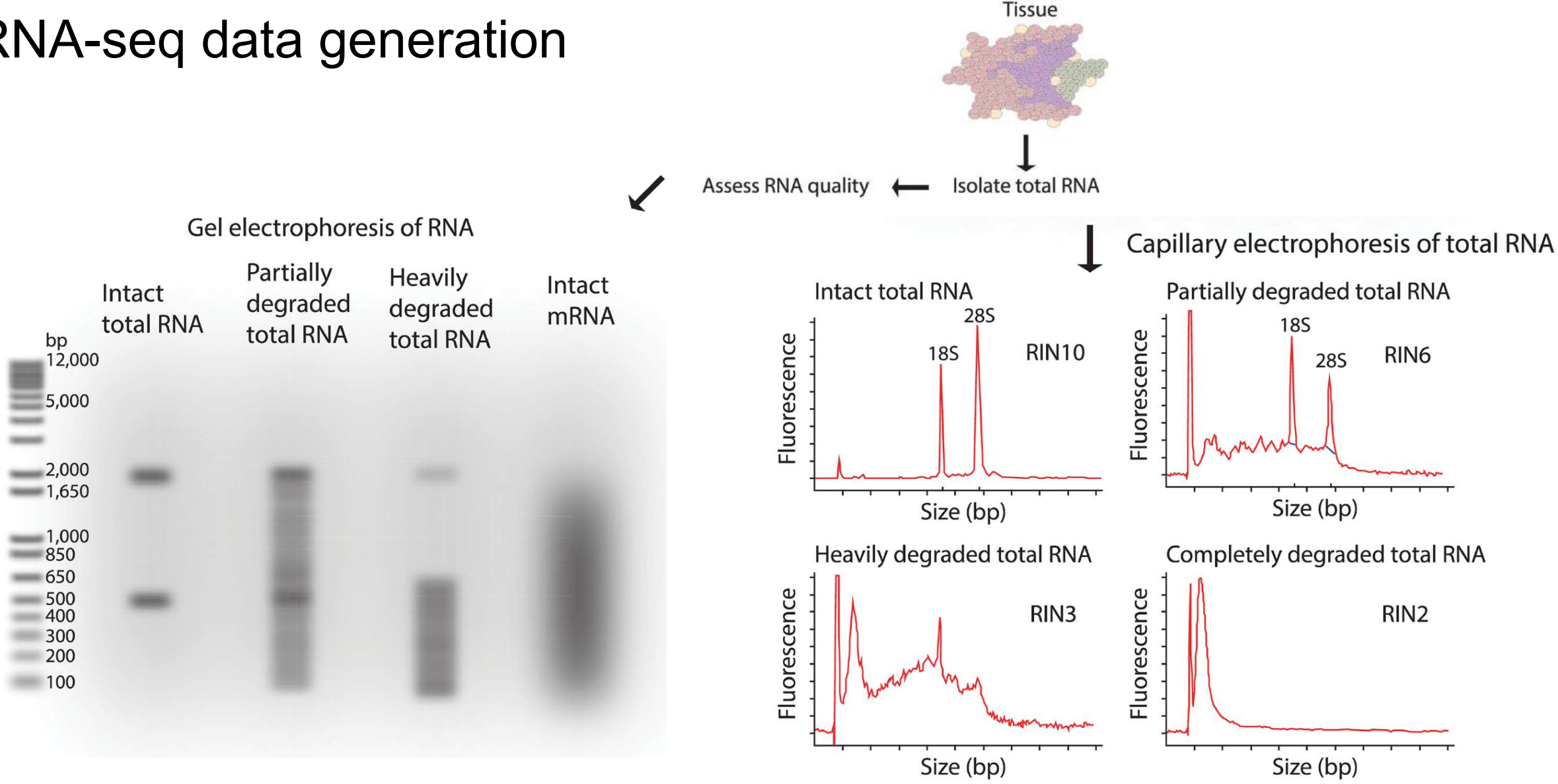
1. Generate data

RNA-seq data generation



Griffith et al (2015) Plos Computational Biology

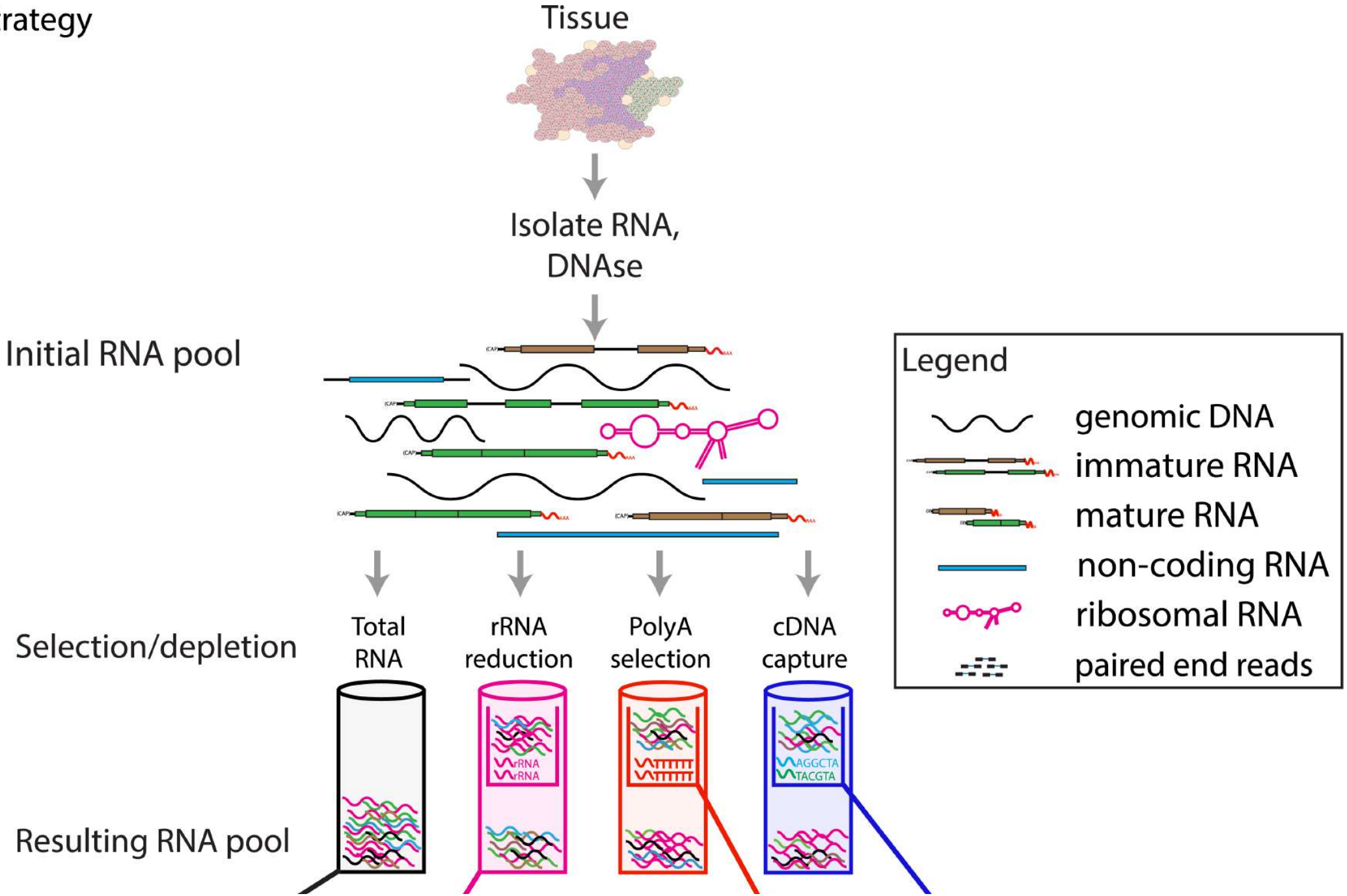
RNA-seq data generation

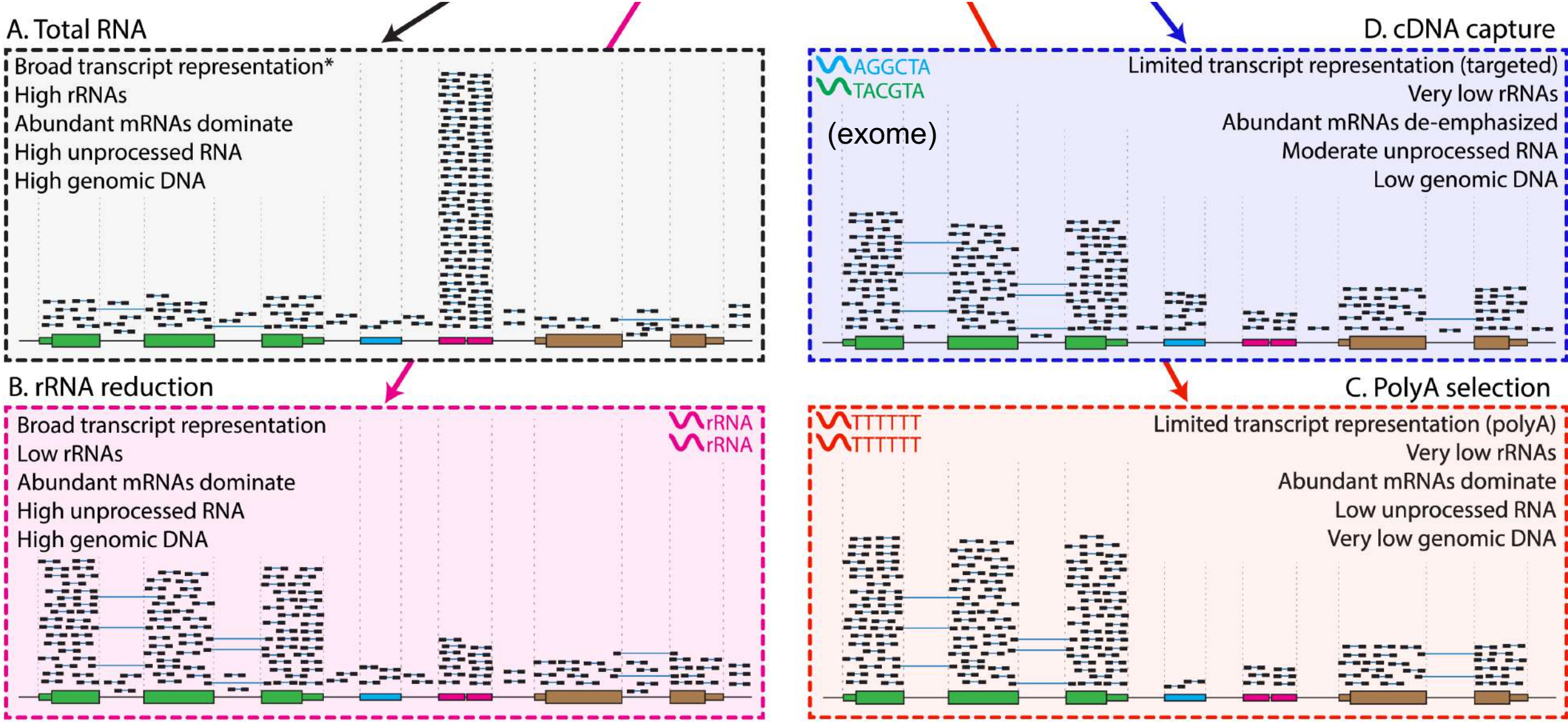


RIN = 28S:18S ratio

Griffith et al (2015) Plos Computational Biology

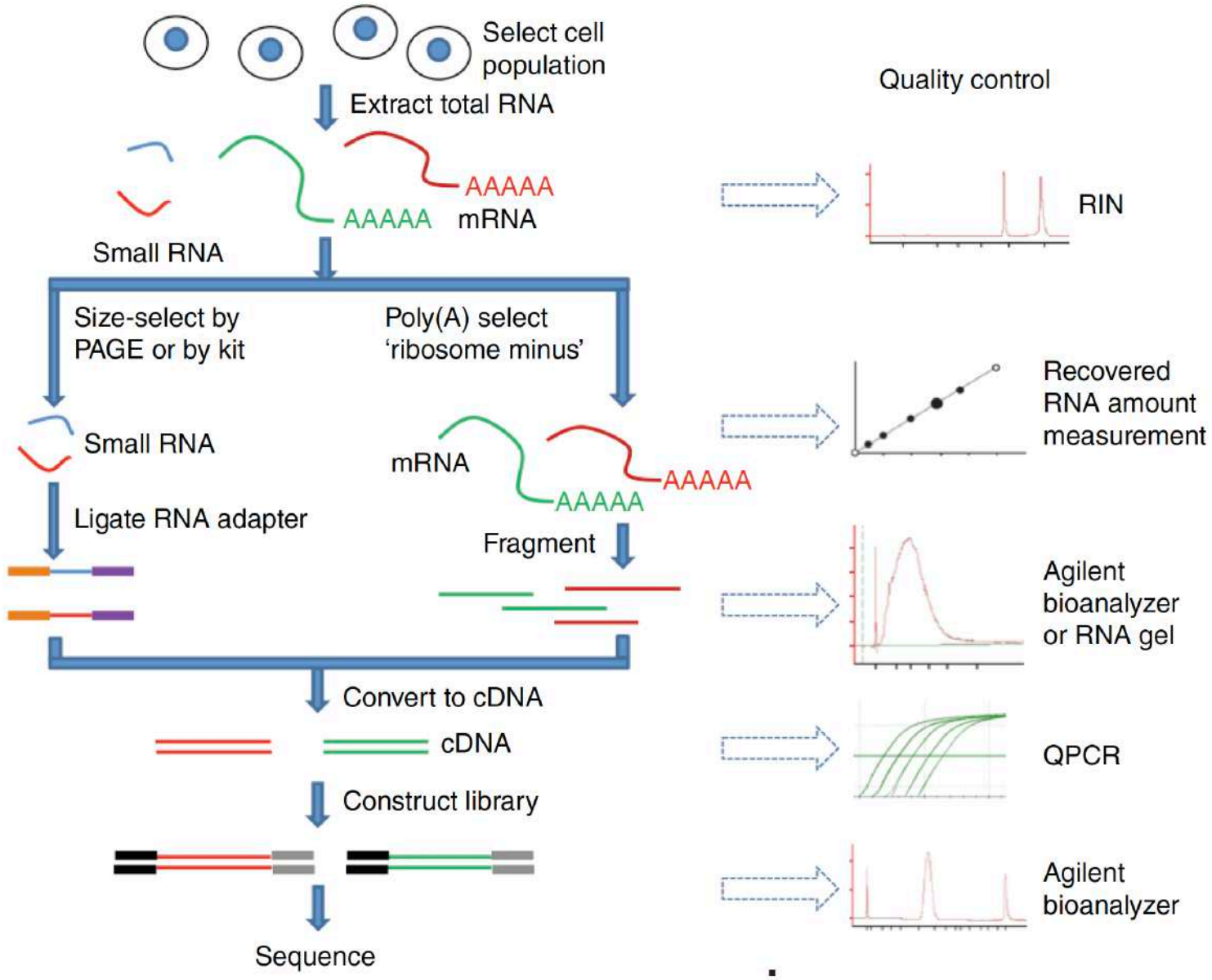
RNA-seq Strategy





Expected Alignments

General RNA library preparation workflow



i) RNA extraction and measuring its integrity,

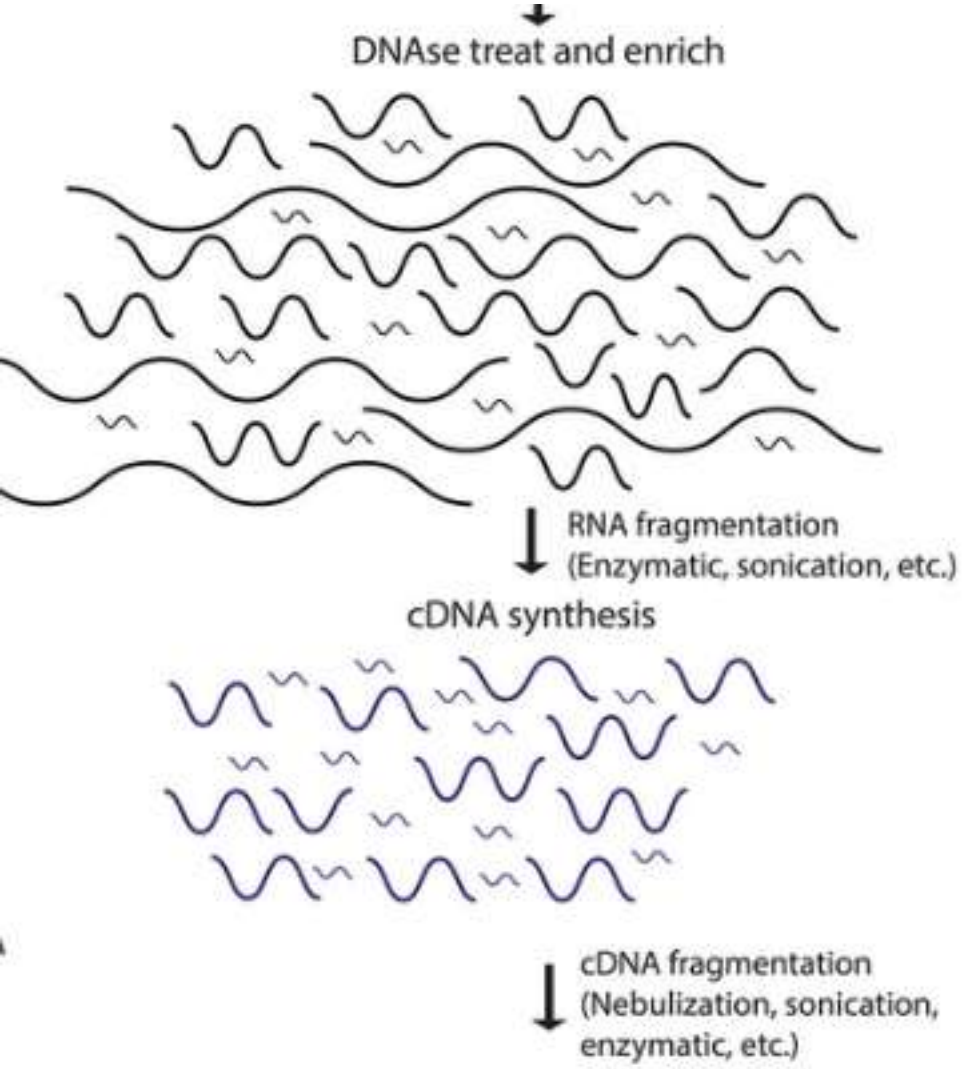
ii) rRNA is depleted (either using poly(A)-selection or rRNA depletion)

iii) the remaining RNA molecules are fragmented, ideally achieving a uniform size distribution.

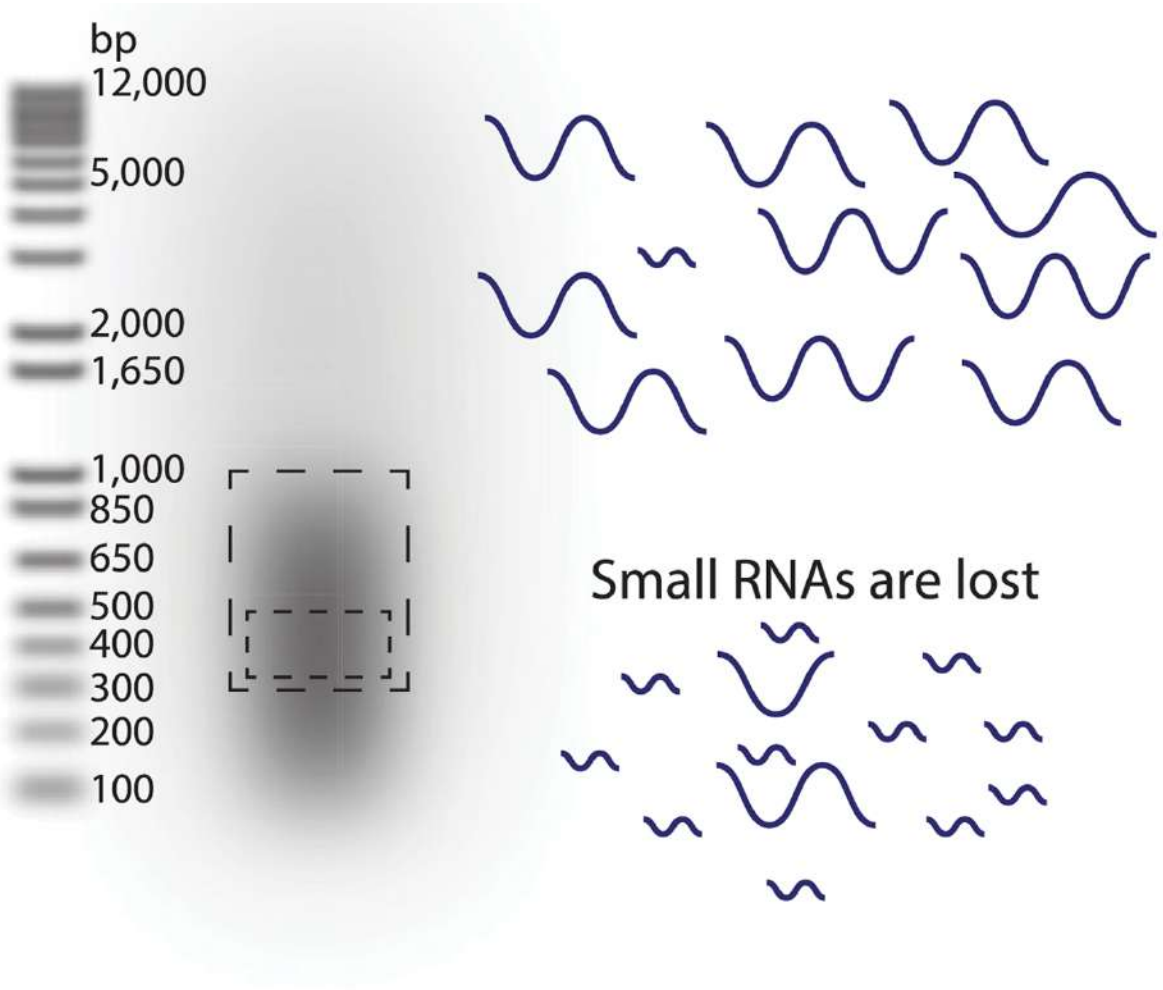
iv) Double-stranded cDNA is synthesized and the adapters for sequencing are added to construct the final library whose fragment size distribution should be unimodal and well-defined.

RNA-seq data generation ; cDNA

Total RNA

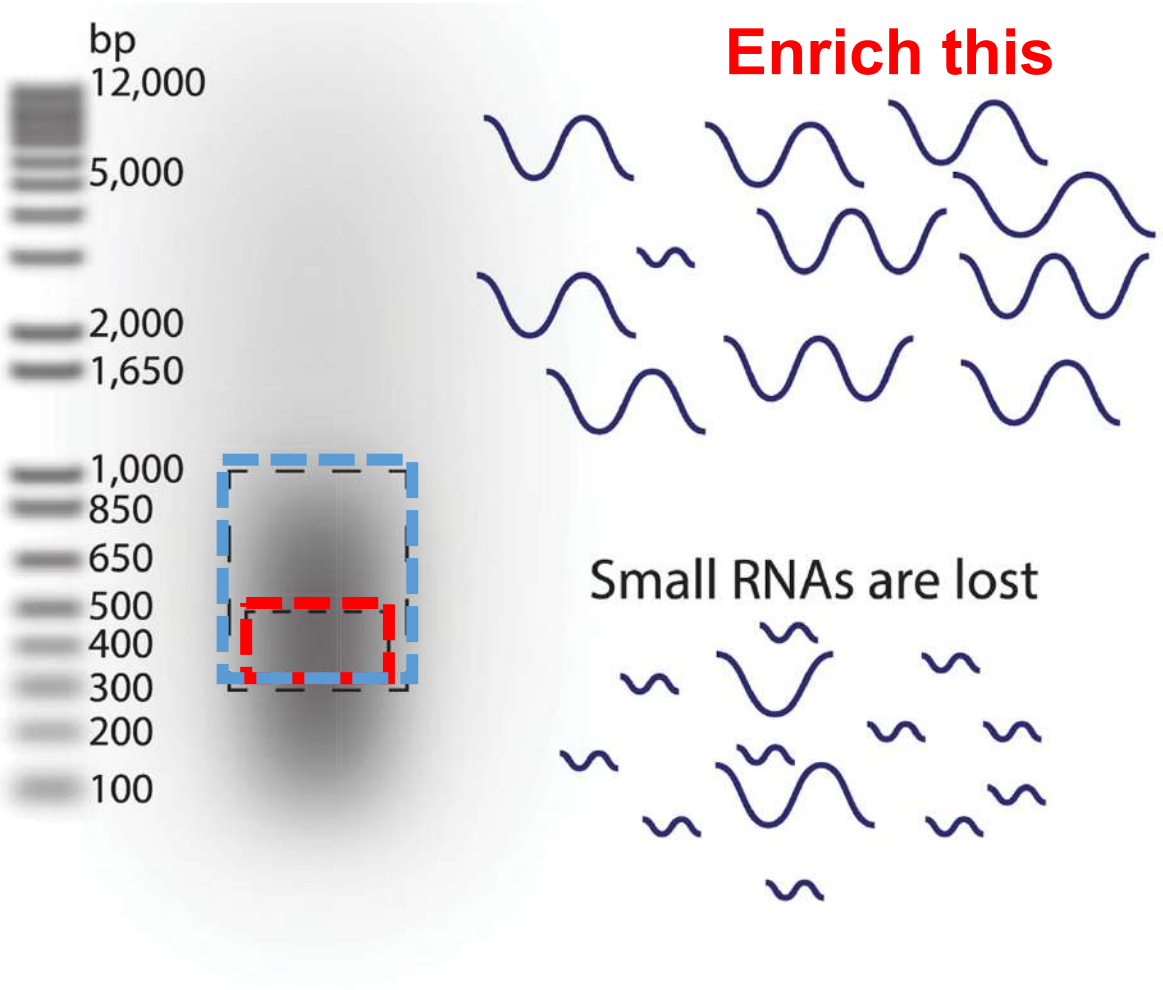


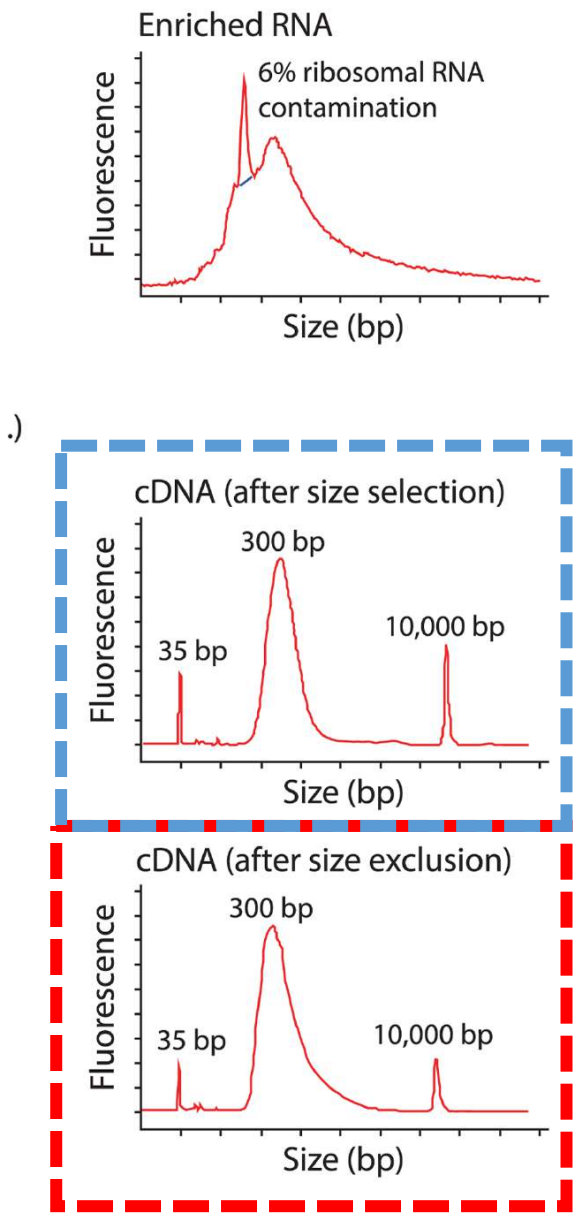
Size selection



Griffith et al (2015) Plos Computational Biology

RNA-seq data generation





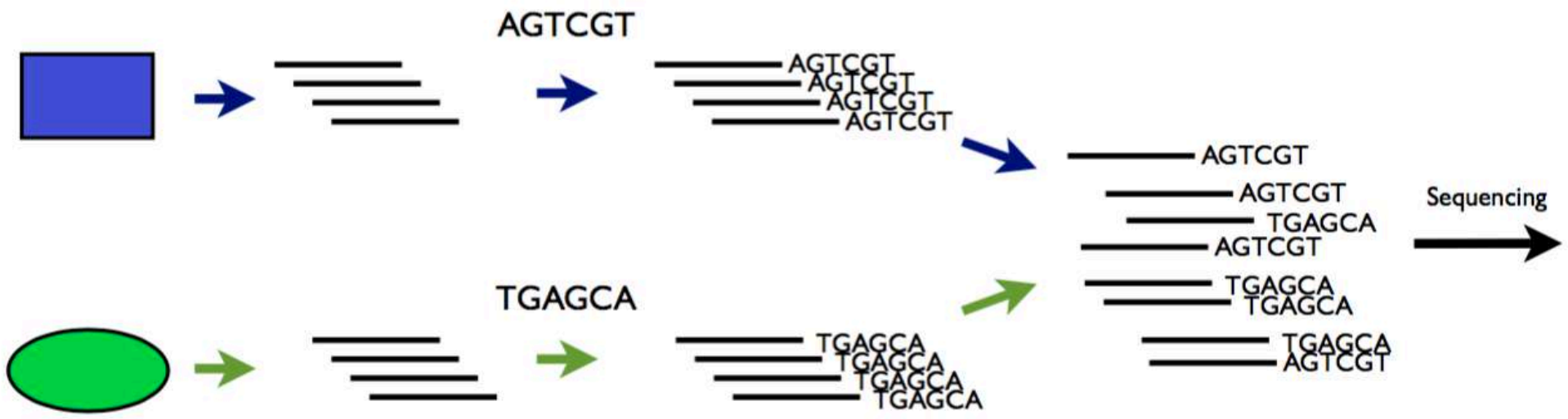
Column selection

Gel selection

Griffith et al (2015) Plos Computational Biology

Enriched cDNA -> Added sequencing adaptors -> Sequencing

High coverage of Illumina allows multiplexing:
(Use of 4-6 nucleotides to identify different samples in the same run)



1.1 Data type

Fastq file

Sequence name

Seq 1

@D00368:375:HT3TKBCX2:1:1107:1135:2137 1:N:0:GTGGCC
CTCAGCCTTAGTGCTCAAGAAACGGGAGGGAGAGTCGATGT
CGGCGTTCATGGCCGACATCTGGATGAAGCGGTCGATATCGT
+
DDDDDHIIIIIIIIIIIIIIIIIIIHHIIIHHIIHIIIIIIIIIHHHIHHIIHIIIIIIIIIIHHIIIIIH
IHHHIIIIIIHIIHIIIHIIIHHIIGIICFHIIIIIIIIIIHIHIHIIIIIIIIIIHIIIIIIIHEDGH

Sequence

+

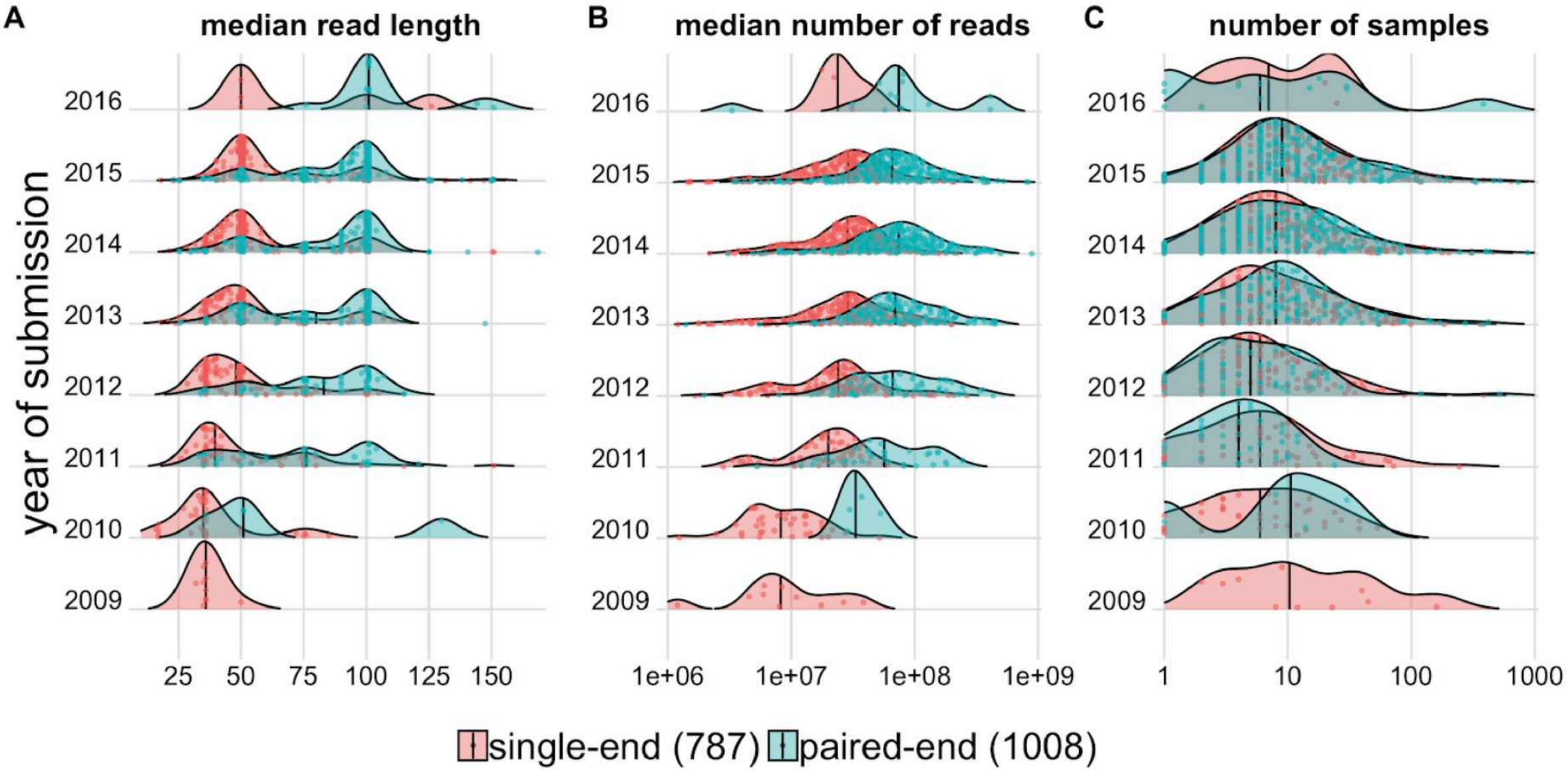
Quality score

Seq 2

@D00368:375:HT3TKBCX2:1:1107:1266:2124 1:N:0:GTGGCC
GTCTGACCTTGTCCACGCGAGATCGAGGACCGCCAGCGGCT
GATCTTGCAATGGCAGTTCCGACTGCGGTCCTTGCGTTGGA
+
DDDDDI?EEE?FHHIIIDHDDC?1F@EH<EGHIHIHIIGDH?<GHHH
HHHIIGHH?GHHE?GECEHHIHHCHHEEHHHHIICGE/?E<

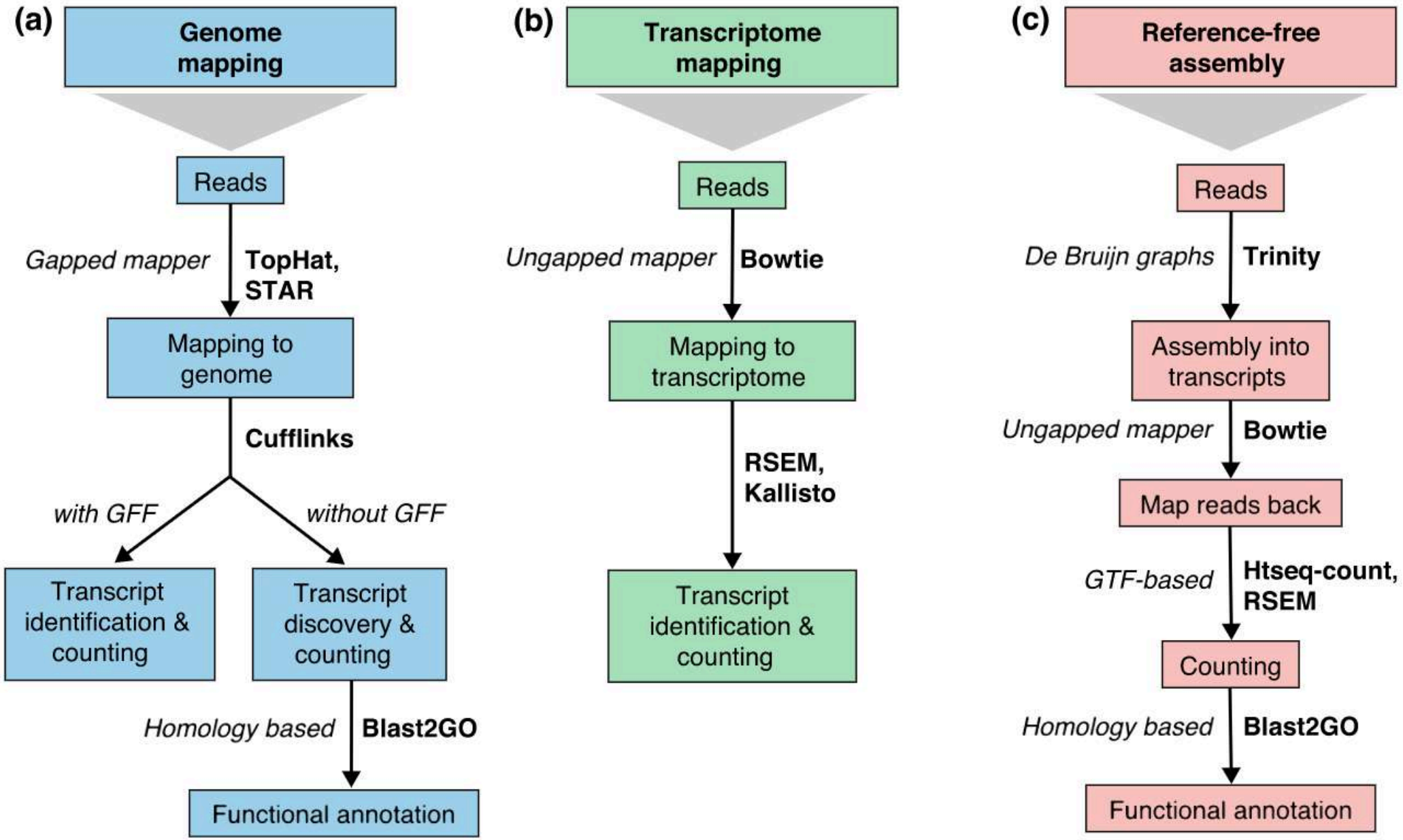
.
.
.
.

Evolution of RNAseq over time (from SRA)



Berge *et al.,* Annual Review of Biomedical Data Science (2019)

Read mapping and transcript identification strategies

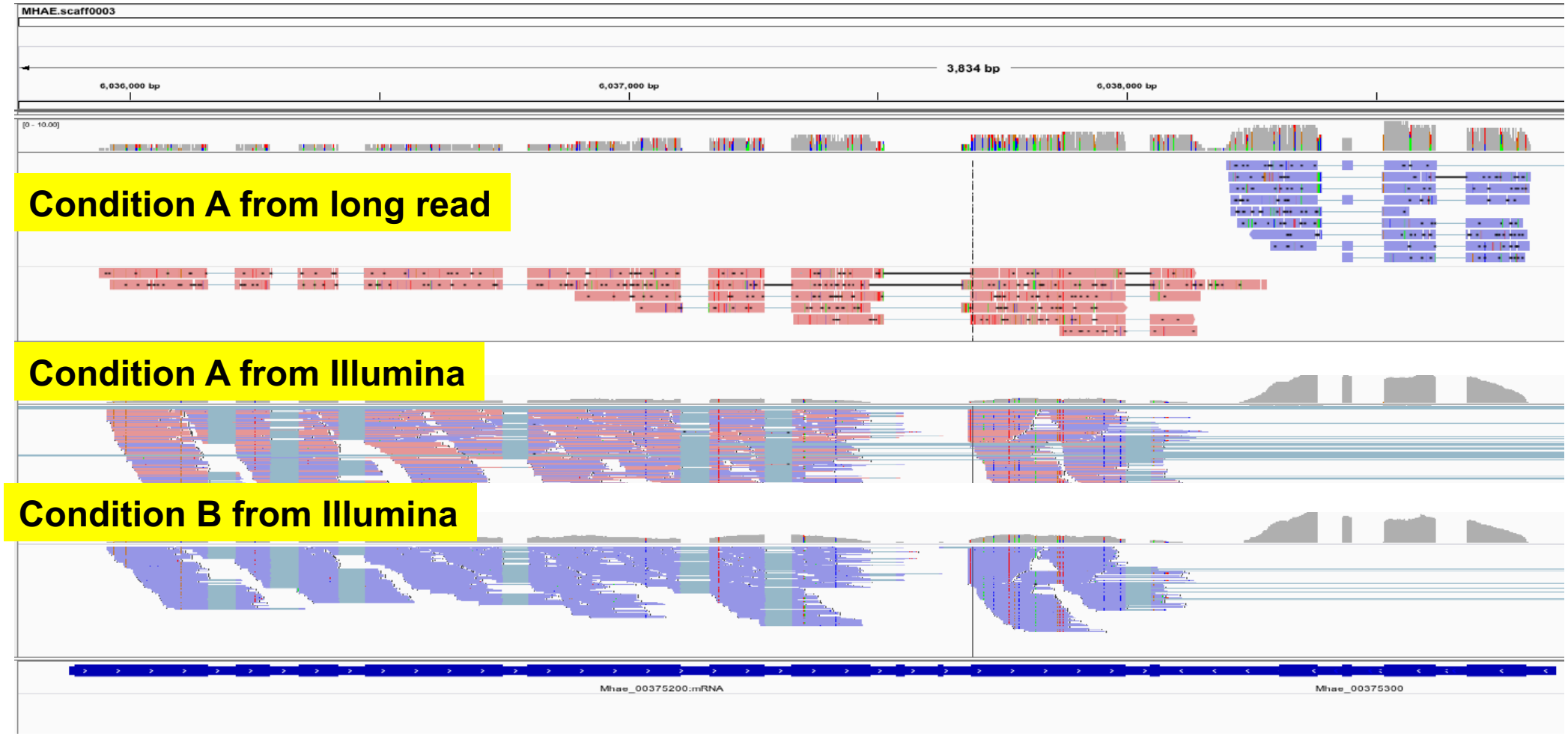


Read visualisation

Load reference, annotation and bam into a program

- Artemis http://www.sanger.ac.uk/science/tools/artemis
- IGV http://software.broadinstitute.org/software/igv/

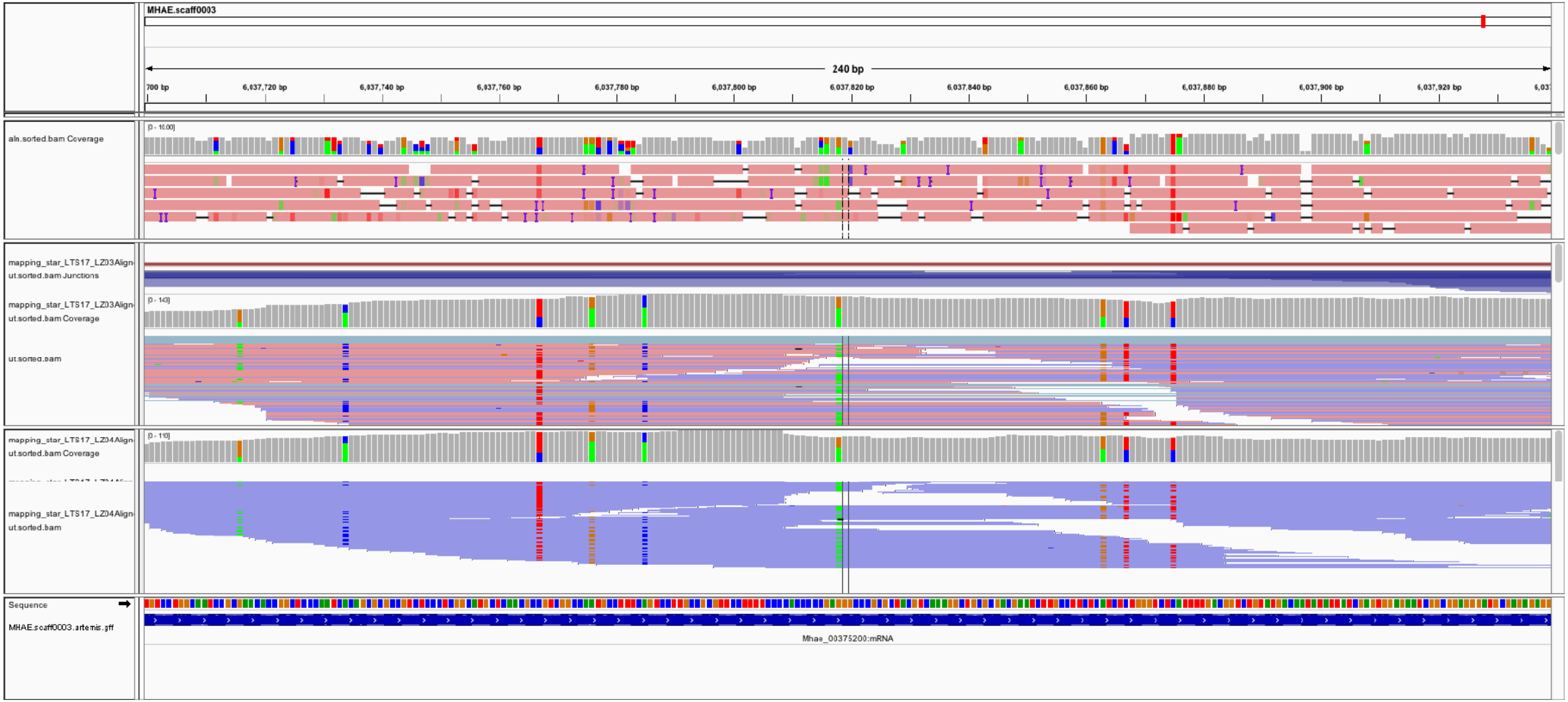
Scenario 1



Annotation: two genes of two orientations

Hueimien Ke

Long reads have more errors

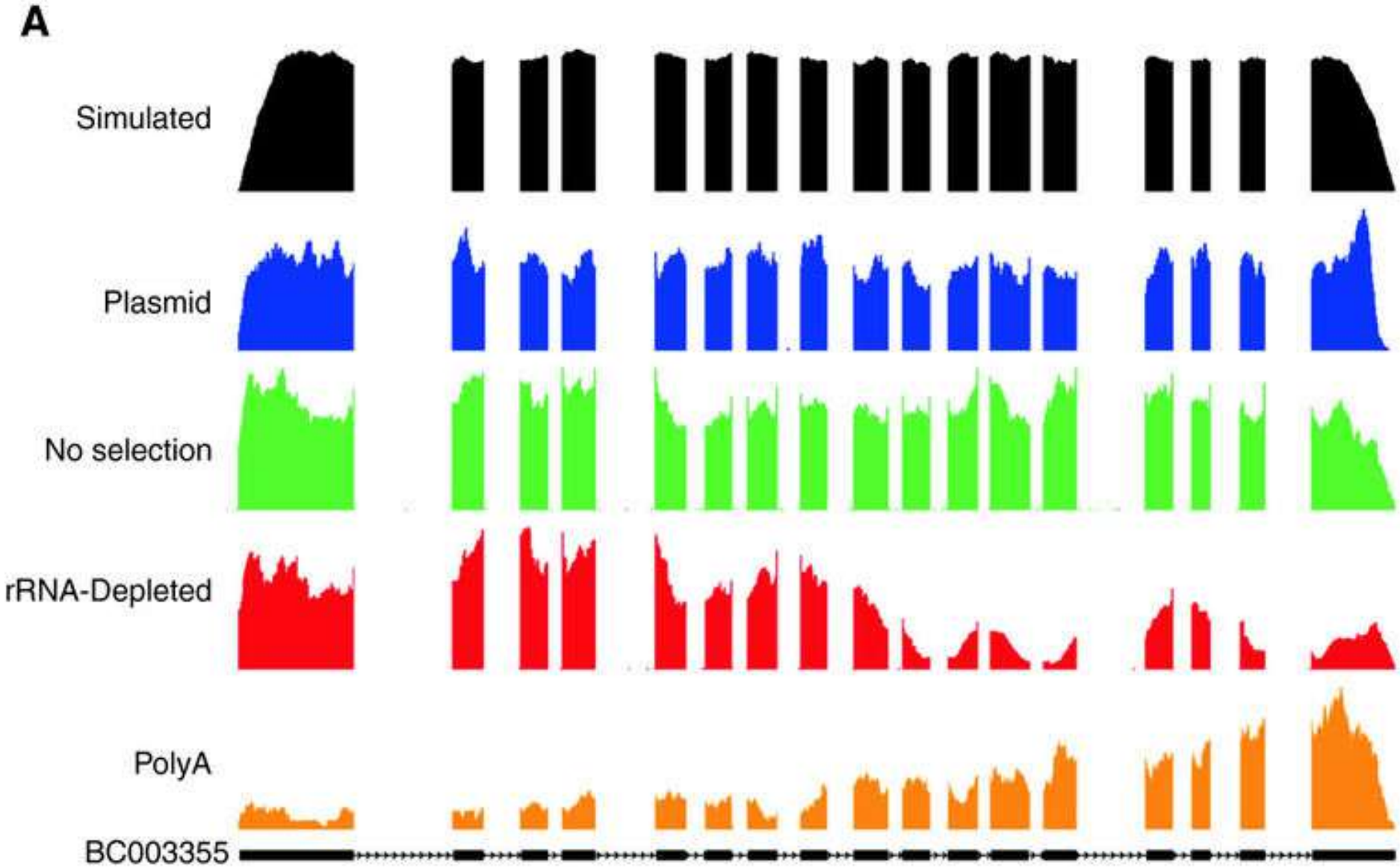


Hueimien Ke

Wrong annotation (wrong gene fusion / wrong exons)

Hueimien Ke

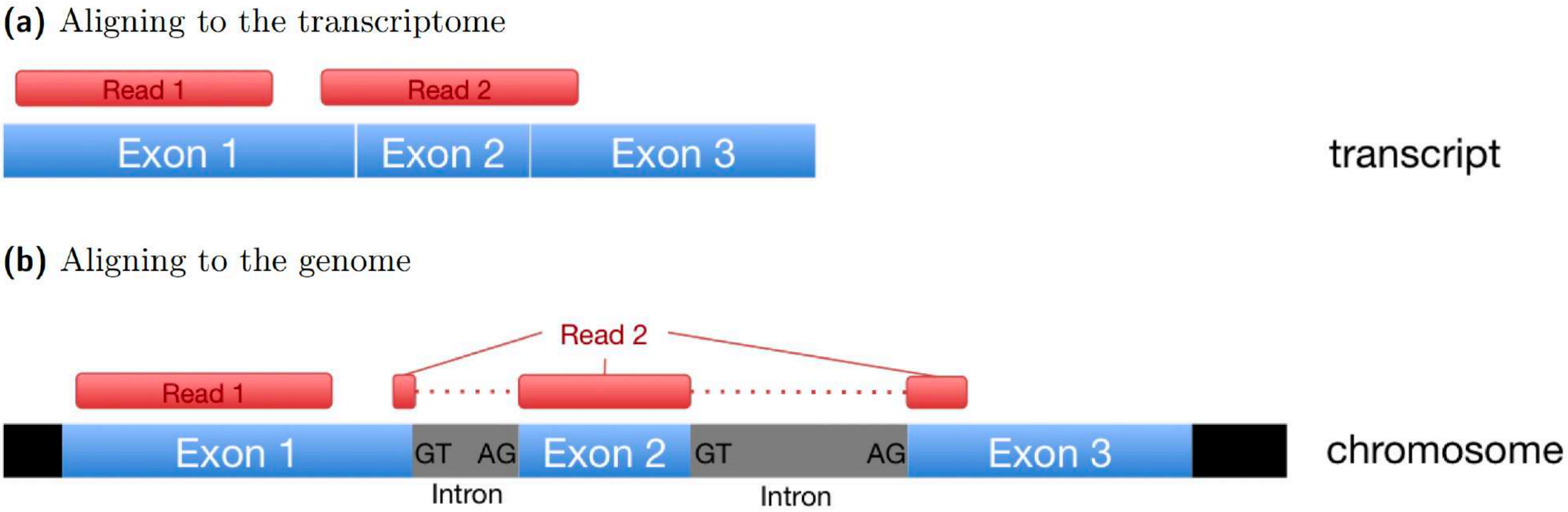
Library enrichment result in sequencing bias



https://doi.org/10.1186/gb-2014-15-6-r86

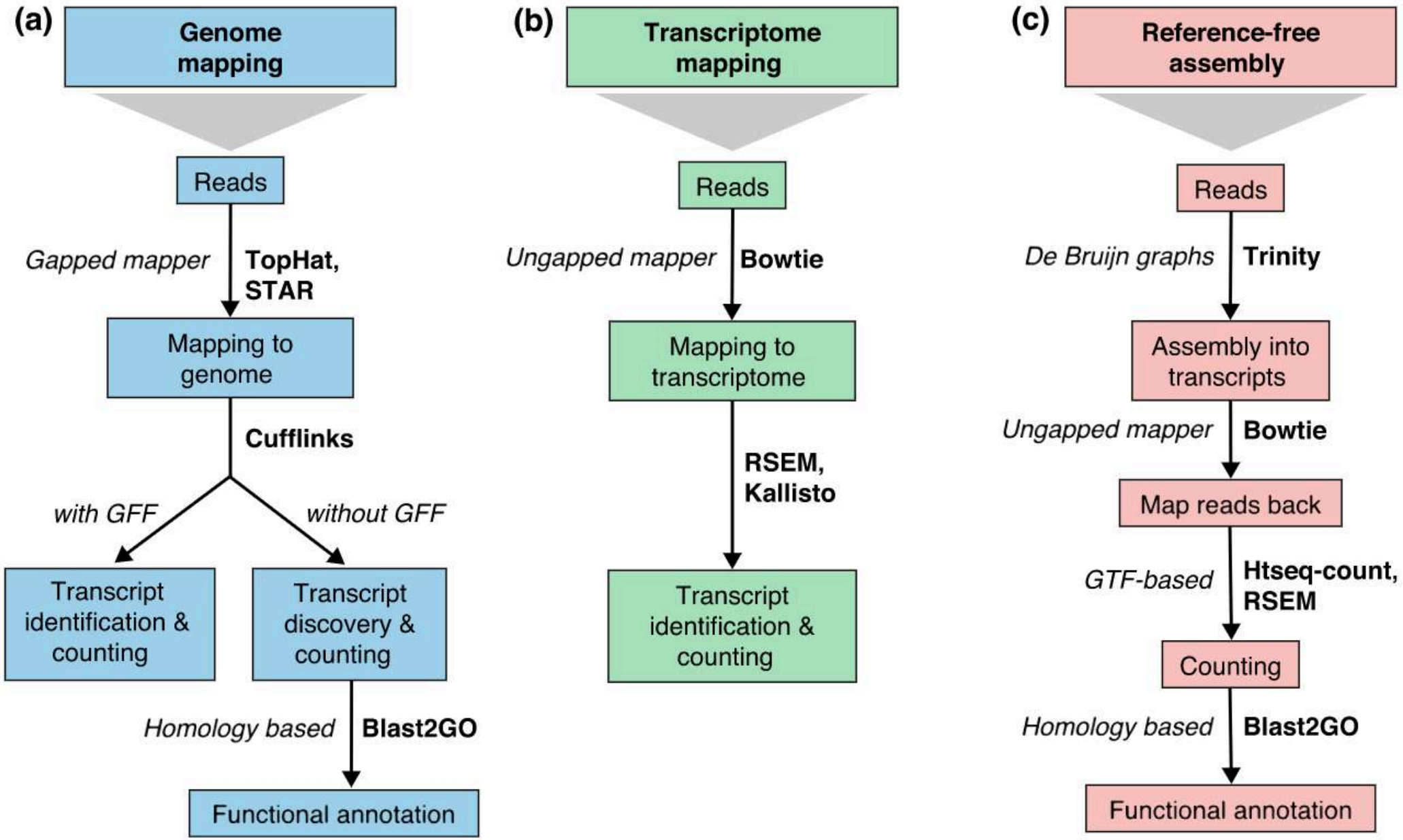
2. Mapping

Read mapping and transcript identification strategies



Dündar *et al* (2018)

Read mapping and transcript identification strategies



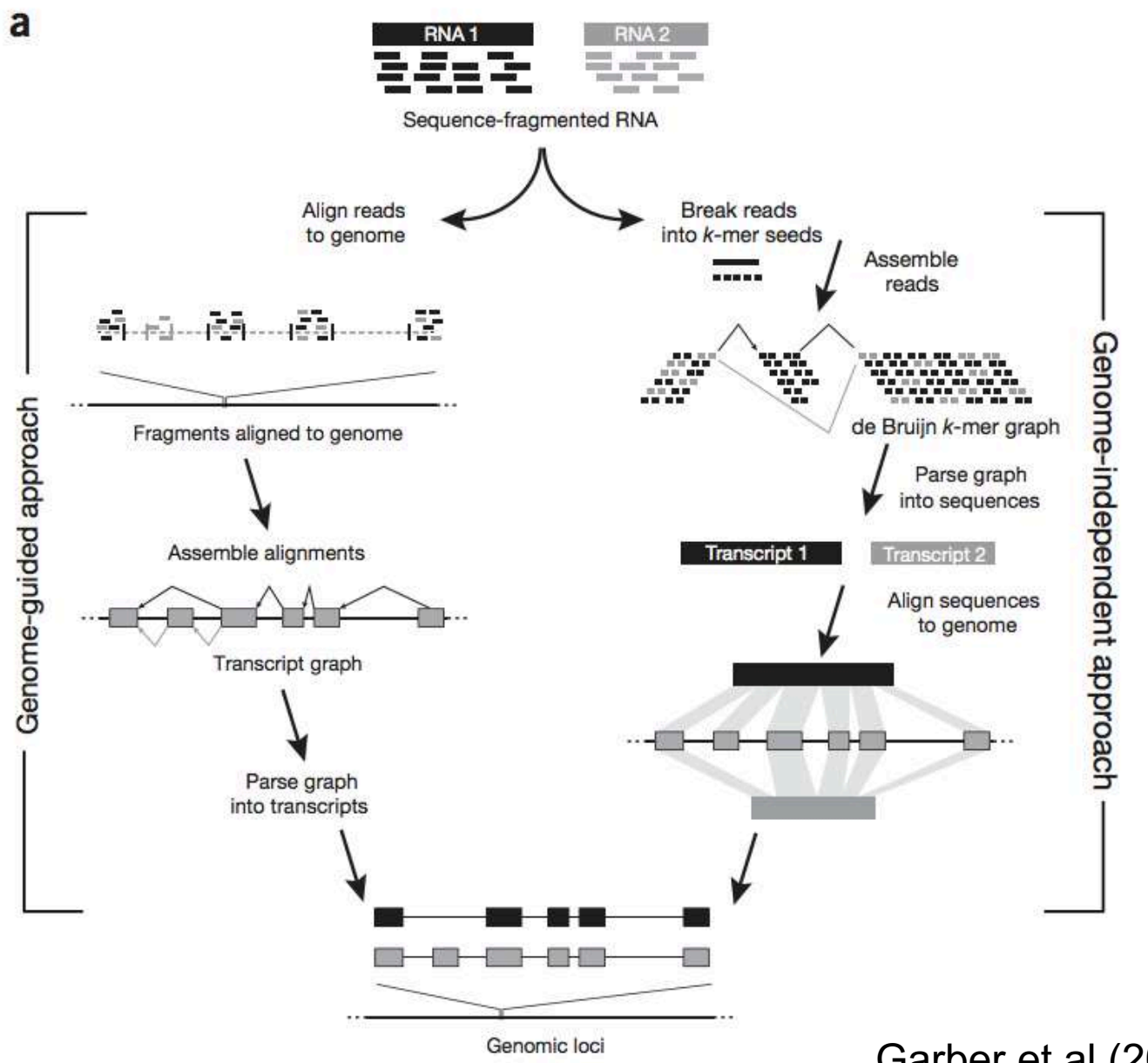
General workflow for RNAseq to produce annotation

Options:

- Align and then assemble
- Assemble and then align

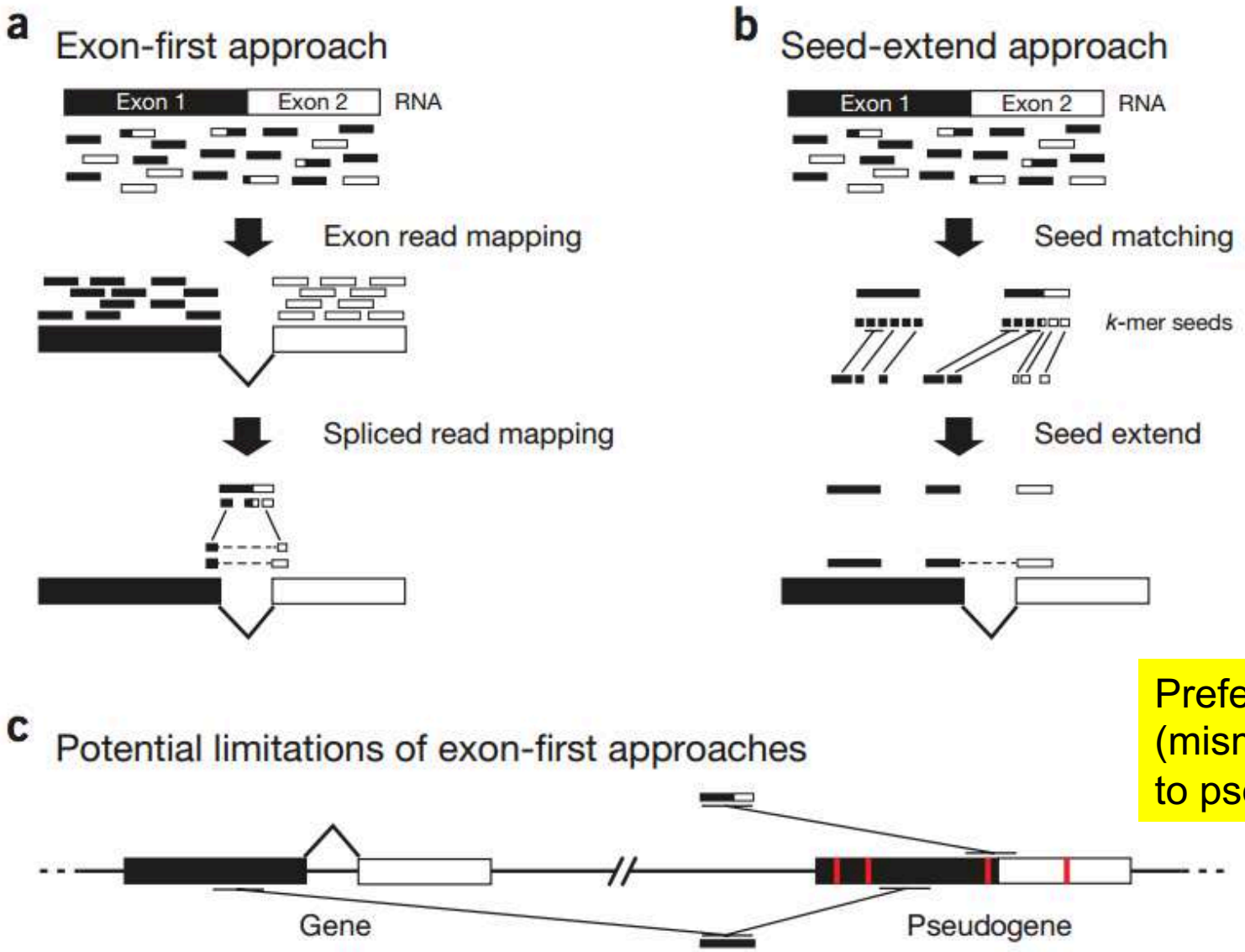
Align to

- Genome
- Transcriptome (if no genome)



Garber et al (2010)

Strategies for gapped alignments of RNAseq reads



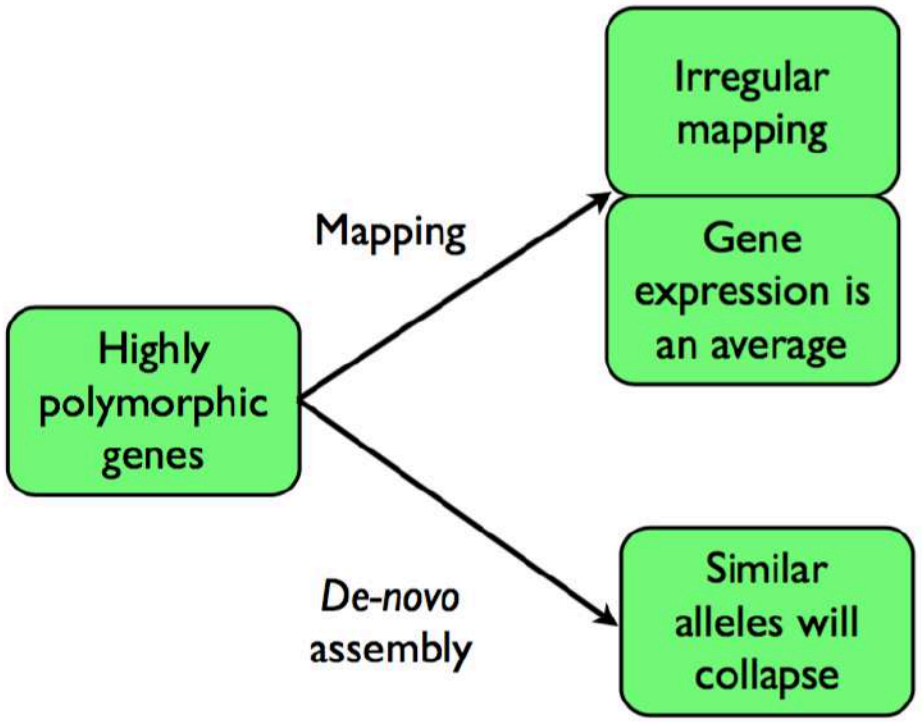
Preferential alignment (mismatch rather than split) to pseudogene

Garber et al (2010)

A potential mapping problem

High heterozygosity/Polyploid problem:

mRNA from species with a high heterozygosity or a polyploid genome can produce highly polymorphic reads for the same gene.



Reference Gene I

ATGCGCGCTAGACGACATGACGACA

CACTTGACGACATGACG **Gene I A**

CTTGACGACATGACGAC
CCCTTGACGACATGACG **Gene I B**
CGCCCTTGACGACATGA

Expression Gene I = A + B

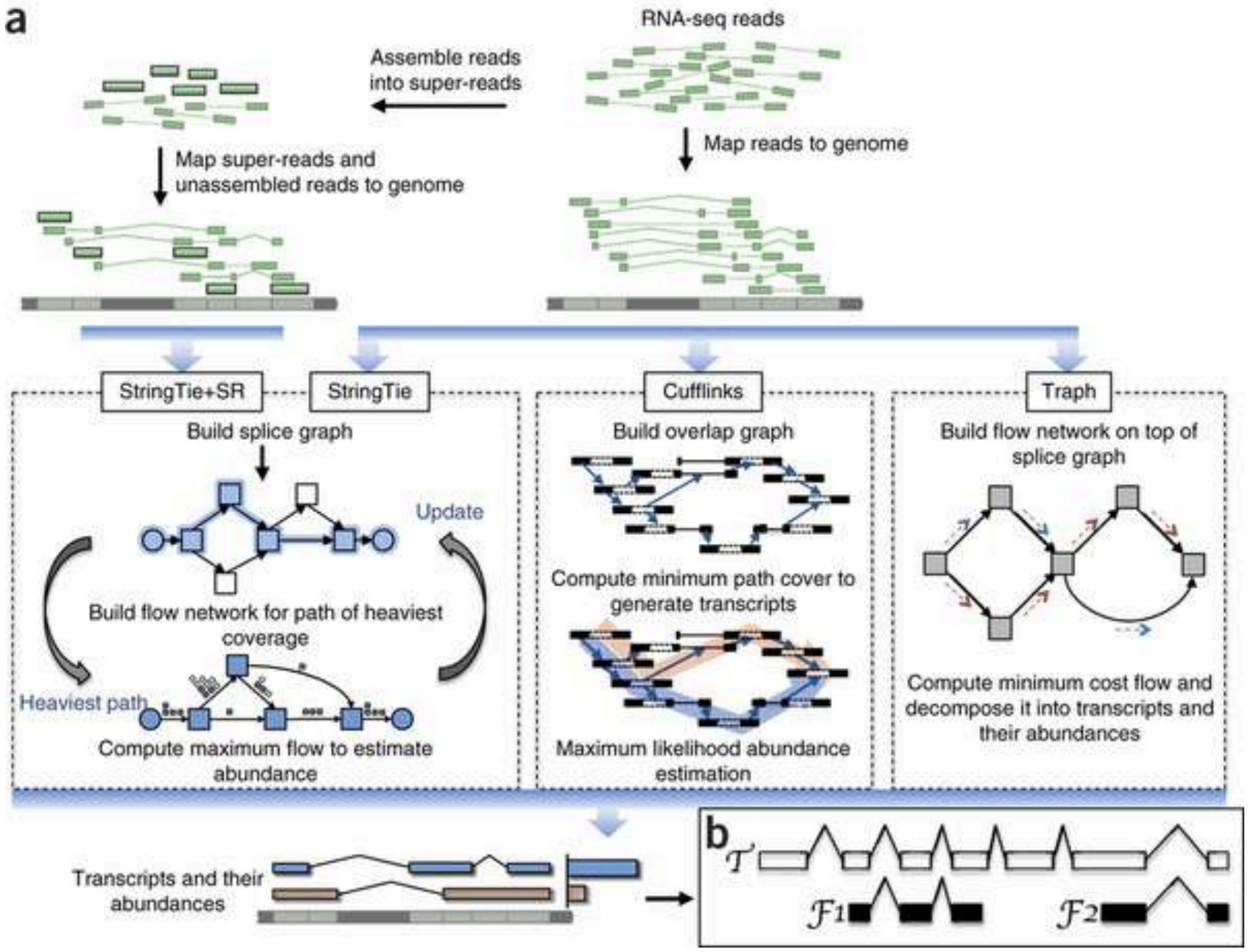
CACTTGACGACATGACG **Gene I A**

CTTGACGACATGACGAC
CCCTTGACGACATGACG **Gene I B**
CGCCCTTGACGACATGA

CGCCCTTGACGACATGACGACA

Collapsed consensus Gene A + Gene B

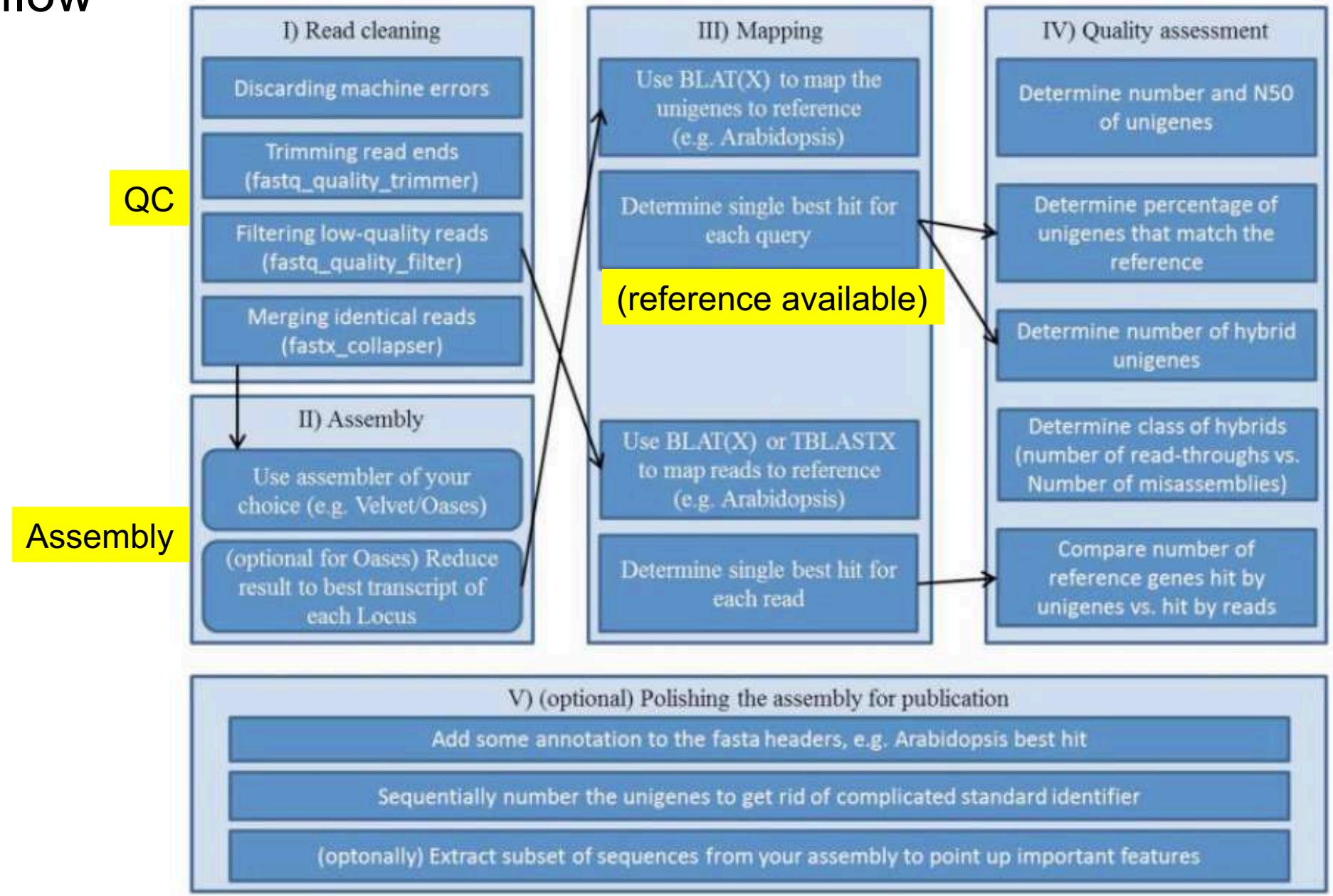
Transcript reconstruction: Cufflinks and StringTie



Pertea et al (2015)

De novo assembly of transcriptomes

Workflow



Schliesky S, Gowik U, Weber AP, Bräutigam A. (2012) **RNA-Seq Assembly – Are We There Yet?** *Front Plant Sci doi:* *10.3389/fpls.2012.00220*

Transcriptome assembly benchmarks

Table 1: Benchmarking of Trans-**ABySS**, **IDBA**-tran, **SOAP**denovo-Trans, **Trinity**, and **SPAdes** on *M. musculus* RNA-seq dataset (accession number SRX648736, 11 million Illumina 100 bp long paired-end reads). The annotated transcriptome of *M. musculus* consists of 38924 genes and 94545 isoforms. The best values for each metric are highlighted with bold.

Assembler	ABySS	IDBA	SOAP	Trinity	SPAdes
Assembled transcripts	63871	38304	61564	47717	48876
Unaligned transcripts	232	98	273	160	817
Misassemblies	156	272	**35**	247	456
Database coverage, %	17.7	16.9	17.1	**18.4**	17.9
Duplication ratio	1.09	**1.004**	1.013	1.155	1.015
50%-assembled genes	6368	6562	6383	6695	**6972**
95%-assembled genes	1763	1572	1804	2251	**2391**
50%-assembled isoforms	6984	6795	6592	**7461**	7140
95%-assembled isoforms	1815	1572	1818	**2388**	2391

Bushmanova *et al* (2019) GigaScience

What is the best protocol for RNAseq analysis?

(Quick answer: no quick answer)

What analysis combinations should we do?

ARTICLE

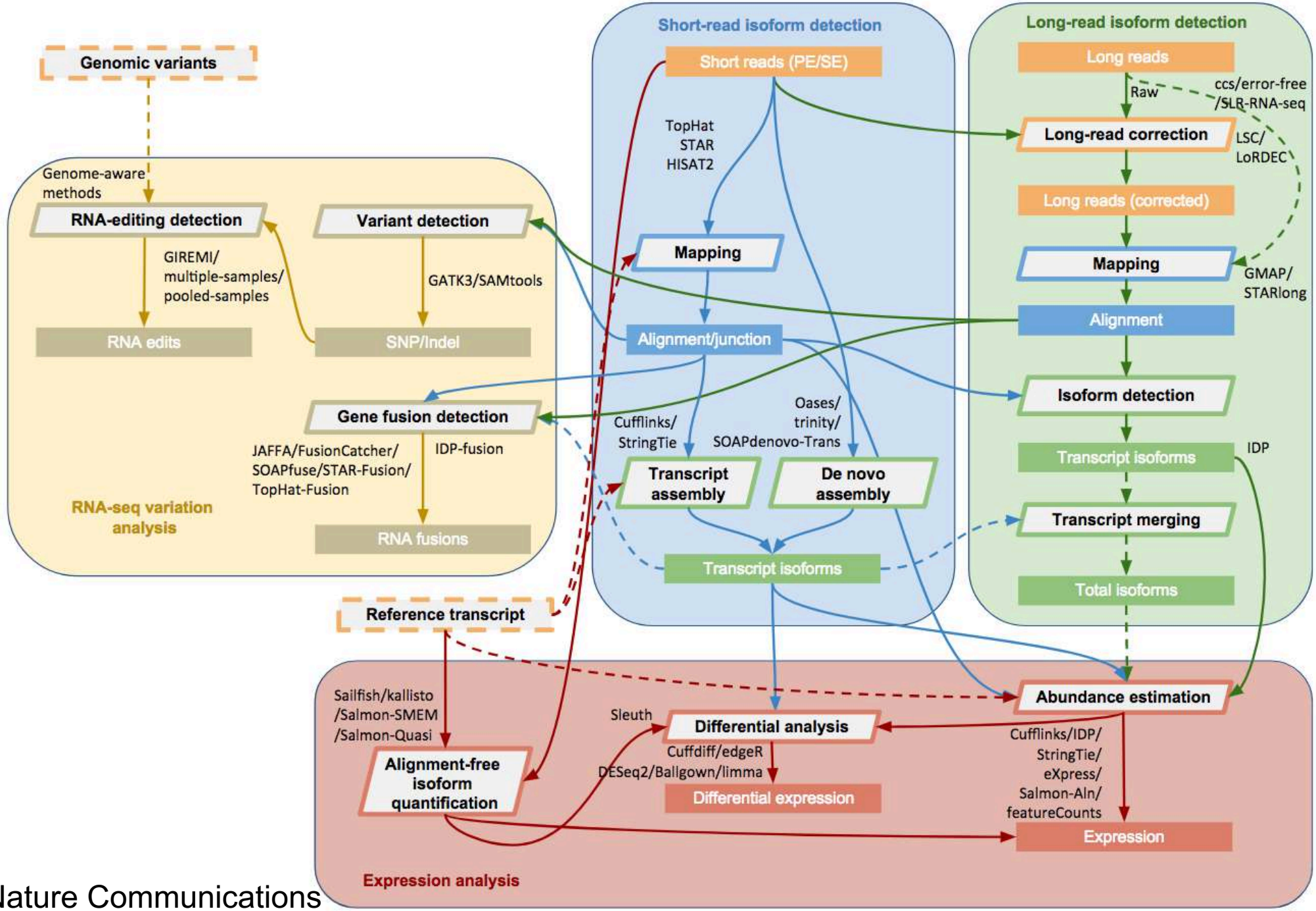
DOI: 10.1038/s41467-017-00050-4 OPEN

Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis

Sayed Mohammad Ebrahim Sahraeian[1], Marghoob Mohiyuddin[1], Robert Sebra[2], Hagen Tilgner[3], Pegah T. Afshar[4], Kin Fai Au[5], Narges Bani Asadi[1], Mark B. Gerstein[6], Wing Hung Wong[7], Michael P. Snyder[3], Eric Schadt[2] & Hugo Y.K. Lam[1]

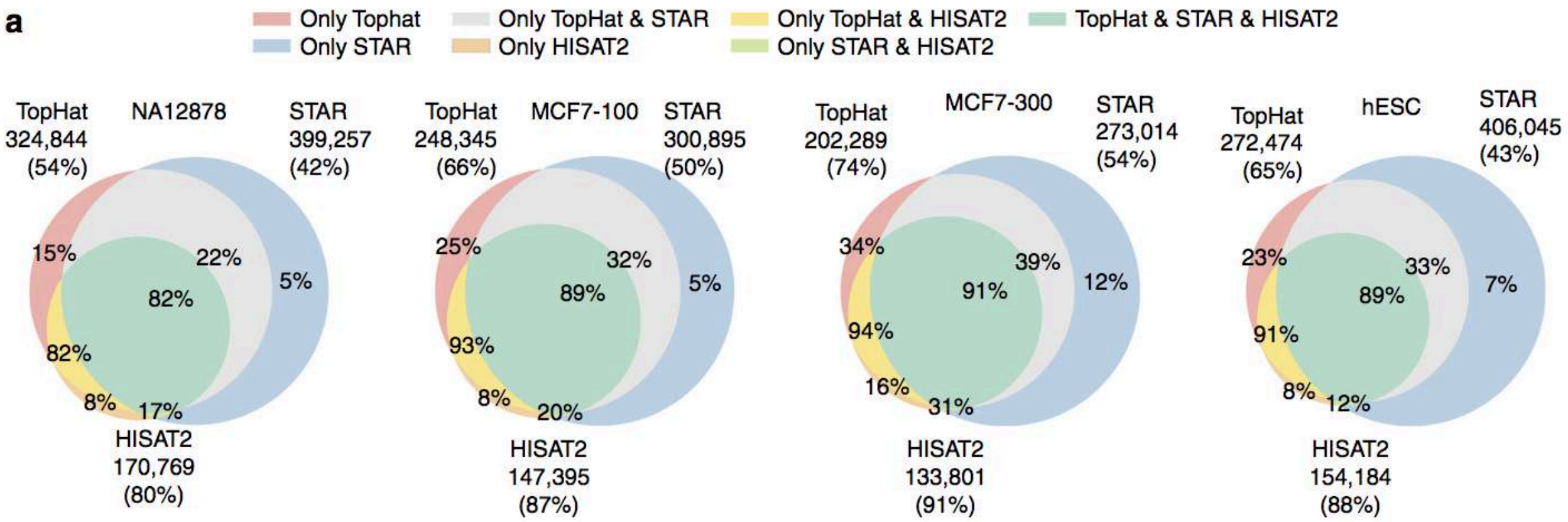
… Here we conduct an extensive study analysing a broad spectrum of RNA-seq workflows. Surpassing the expression analysis scope, our work also includes **assessment of RNA variant-calling, RNA editing and RNA fusion detection techniques**. Specifically, **we examine both short- and long-read RNA-seq technologies, 39 analysis tools resulting in ~120 combinations**, and ~490 analyses involving 15 samples with a variety of germline, cancer and stem cell data sets.

Sahraeian *et al* (2017) Nature Communications



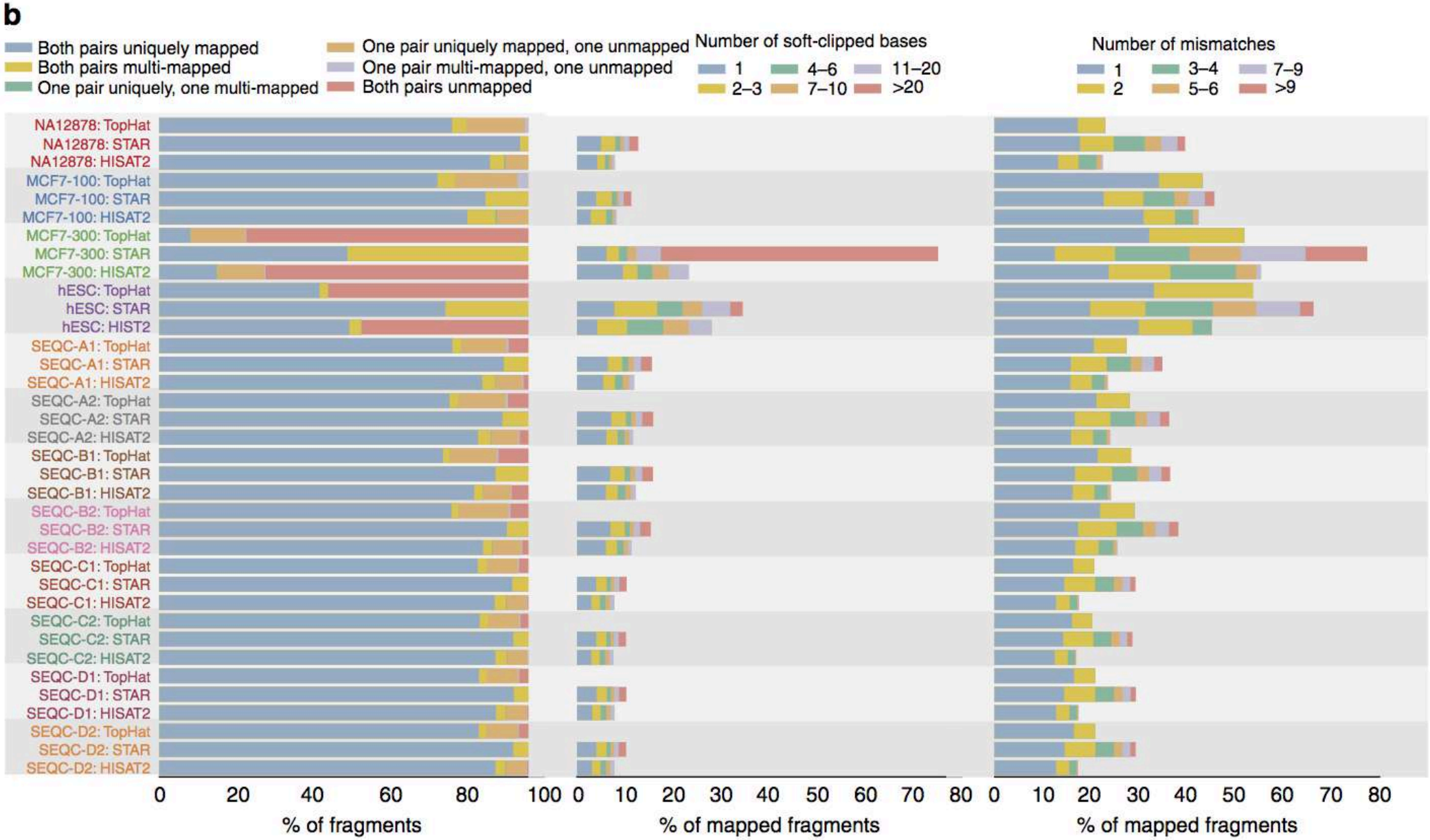
Sahraeian *et al* (2017) Nature Communications

Can be ~10% difference in mapping



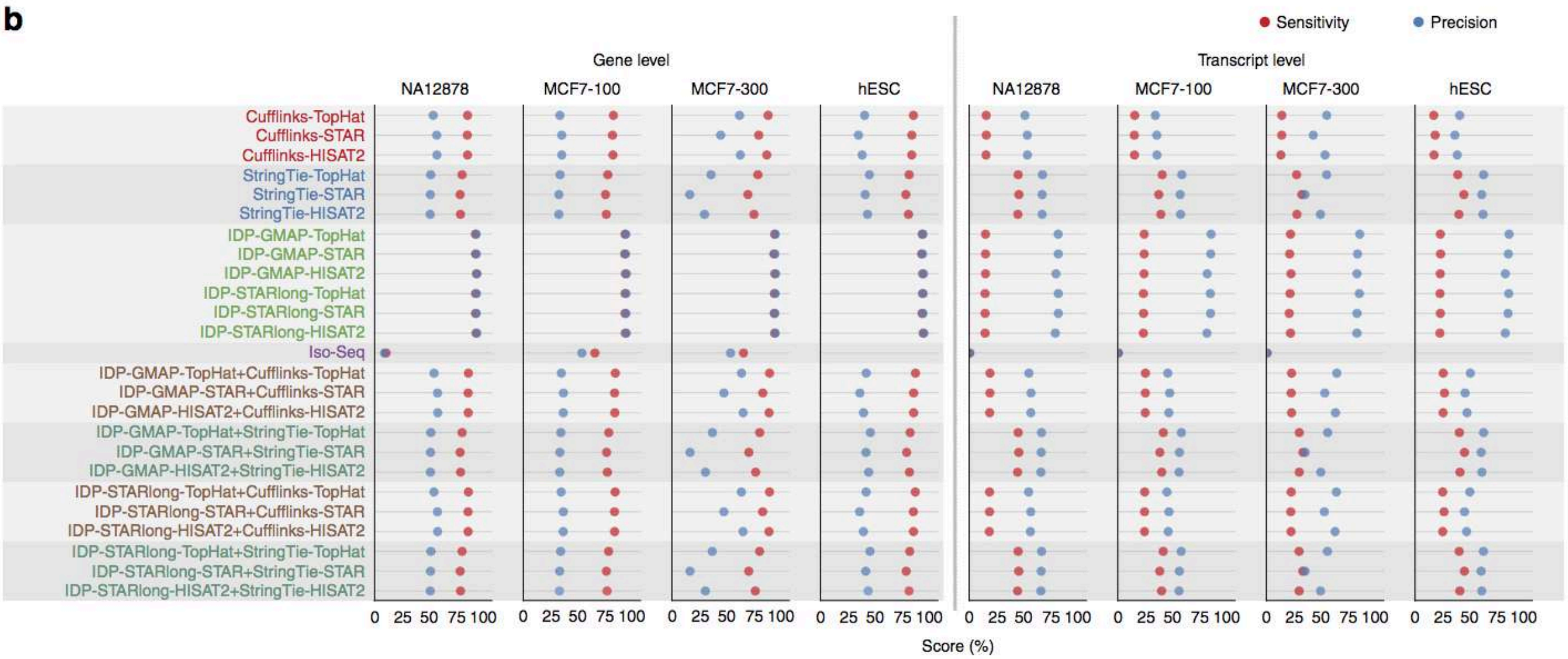
Sahraeian *et al* (2017) Nature Communications

Read mapping different amongst tools



Sahraeian *et al* (2017) Nature Communications

Performance of different transcriptome reconstruction schemes



Sahraeian *et al* (2017) Nature Communications

Typical RNAseq Workflow

2.3 Annotation
(focus only on gene annotation)

STUDY DESIGNS

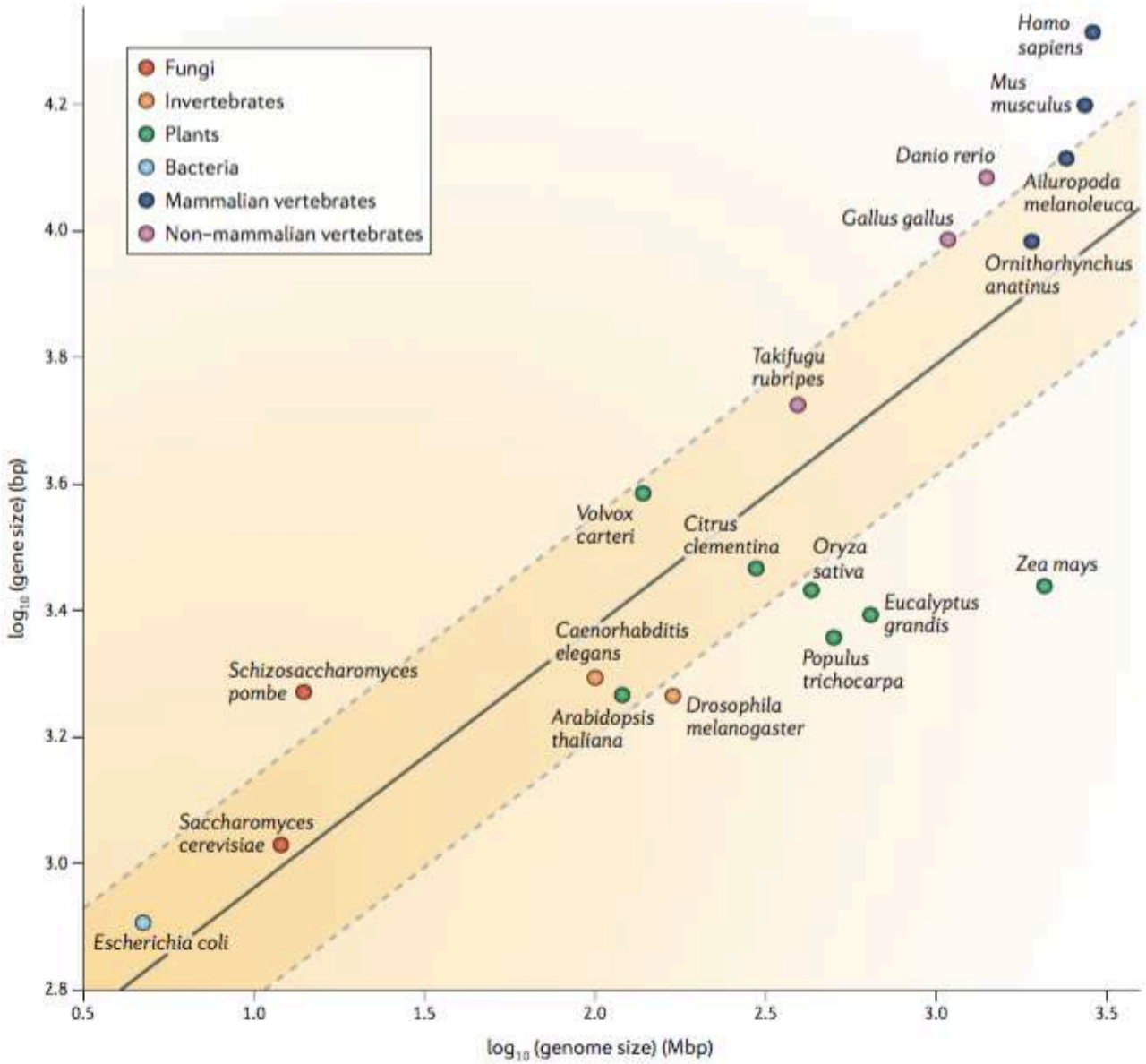
A beginner's guide to eukaryotic genome annotation

Mark Yandell and Daniel Ence

Abstract | The falling cost of genome sequencing is having a marked impact on the research community with respect to which genomes are sequenced and how and where they are annotated. Genome annotation projects have generally become small-scale affairs that are often carried out by an individual laboratory. Although annotating a eukaryotic genome assembly is now within the reach of non-experts, it remains a challenging task. Here we provide an overview of the genome annotation process and the available tools and describe some best-practice approaches.

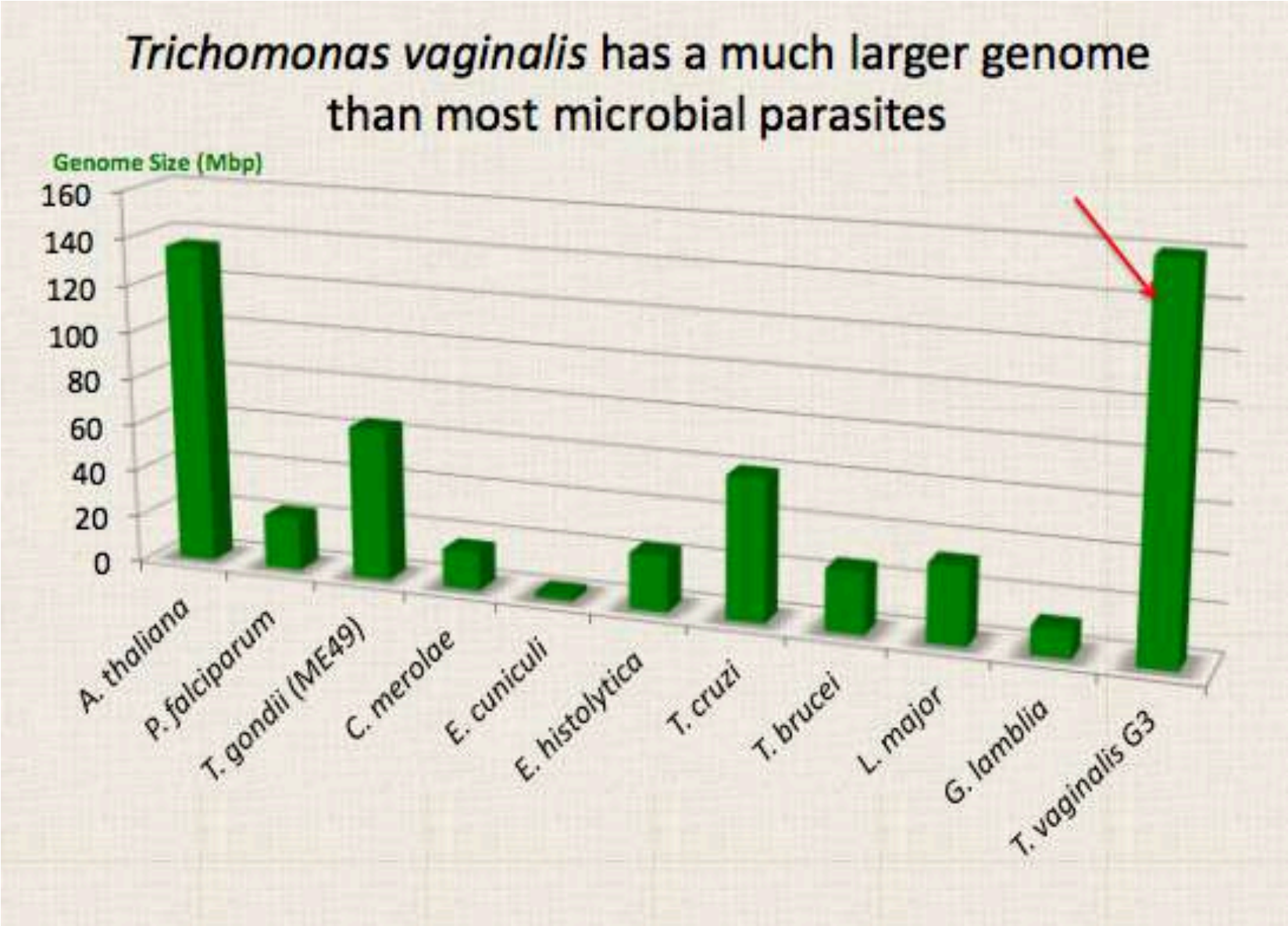
Yandell and Ence (2012)

Know your genome size (and gene numbers)

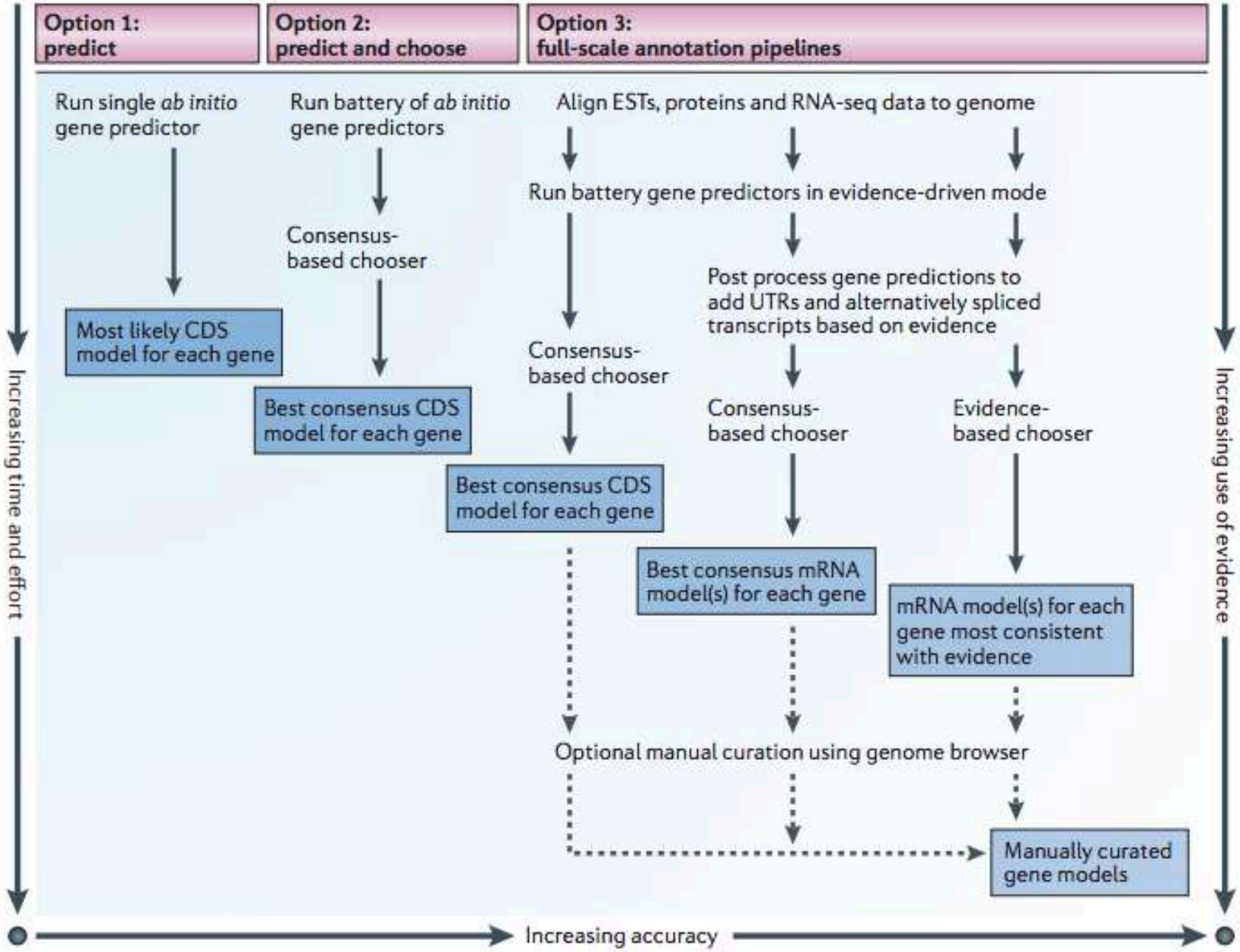


Yandell and Ence (2012)

There's always exceptions (due to TE *Maverick* expansion)

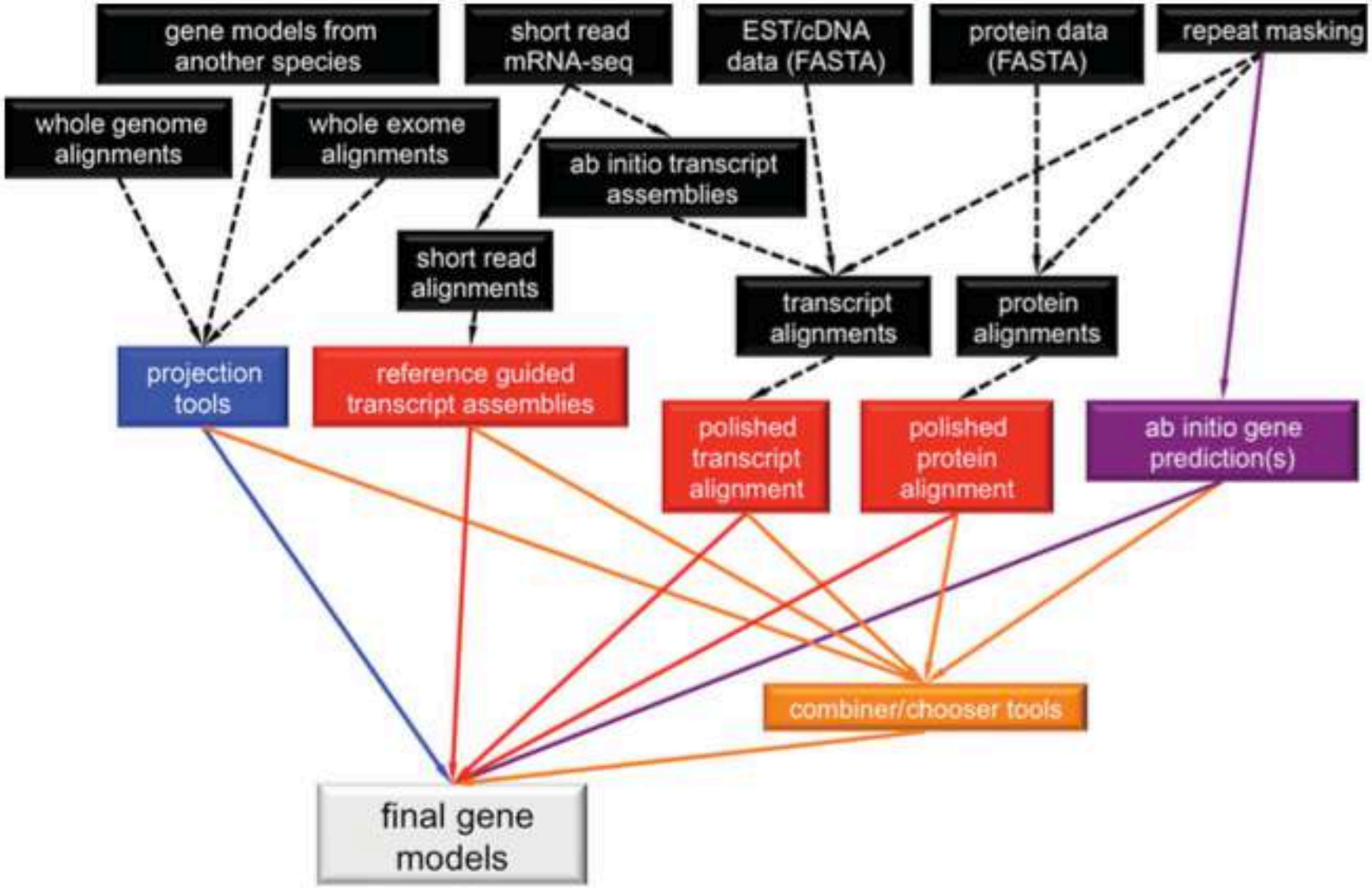


Ence (2014)



Yandell and Ence (2012)

Multiple evidences; Update



Campbell and and Yandell (2015)

Basic rule of thumb

Just genome with no closely related species

Different *de novo* predictors, and combine them with combiners

Genome + closely related species + RNAseq

de novo predictors + evidence + combiners

Genome + closely related species available + RNAseq

de novo predictors + evidence + RNAseq evidence + combiners

Genome + closely related species available + RNAseq + manual efforts

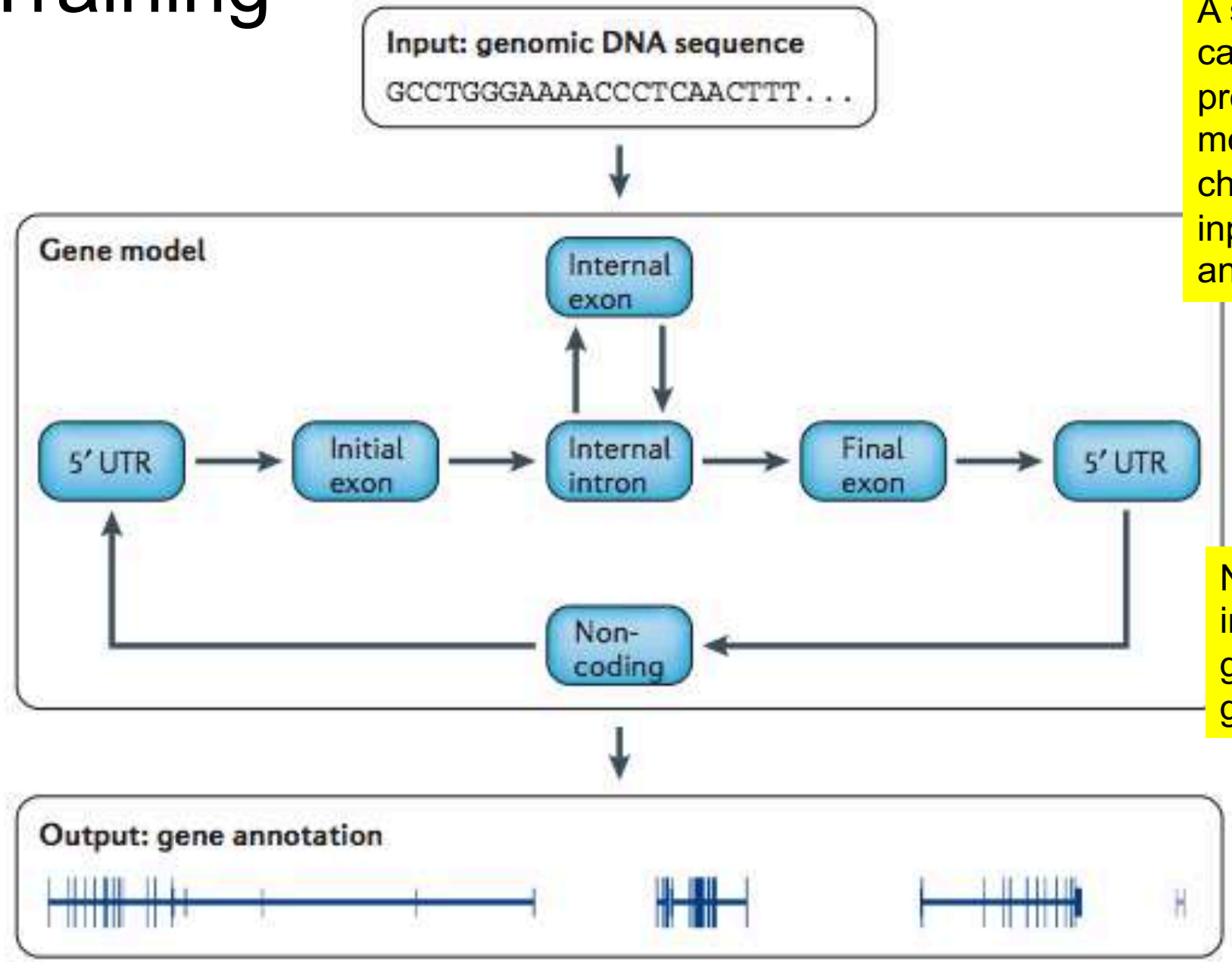
manual curation to train *de novo* predictors

Trained predictors + protein evidence + RNAseq evidence + combiners

Genome + initial annotations + RNAseq

protein evidence -> Trying to improve existing annotations

Training



A simplified gene-finding model that captures the basic properties of a protein-coding gene is shown. The model takes the DNA sequence of a chromosome, or a portion thereof, as input and produces detailed gene annotations as output.

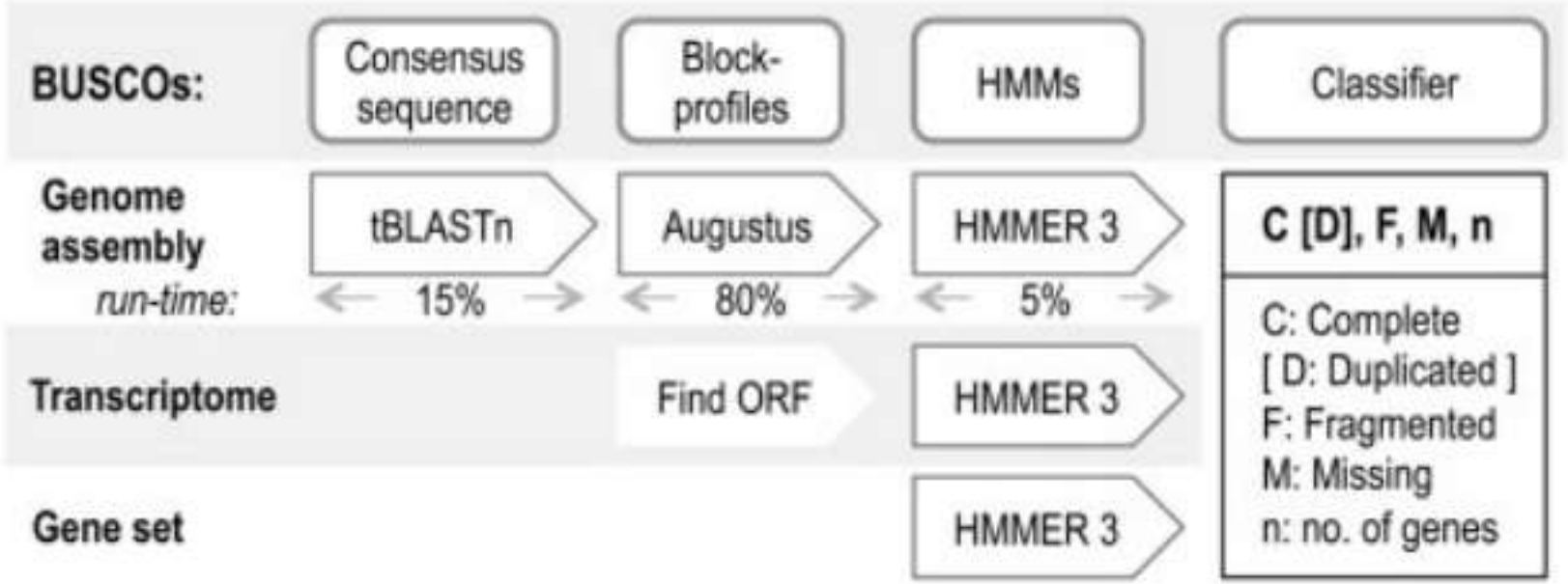
Note that this simplified model is incapable of identifying overlapping genes or multiple isoforms of the same gene. UTR, untranslated region.

Libbrecht and Noble (2015)

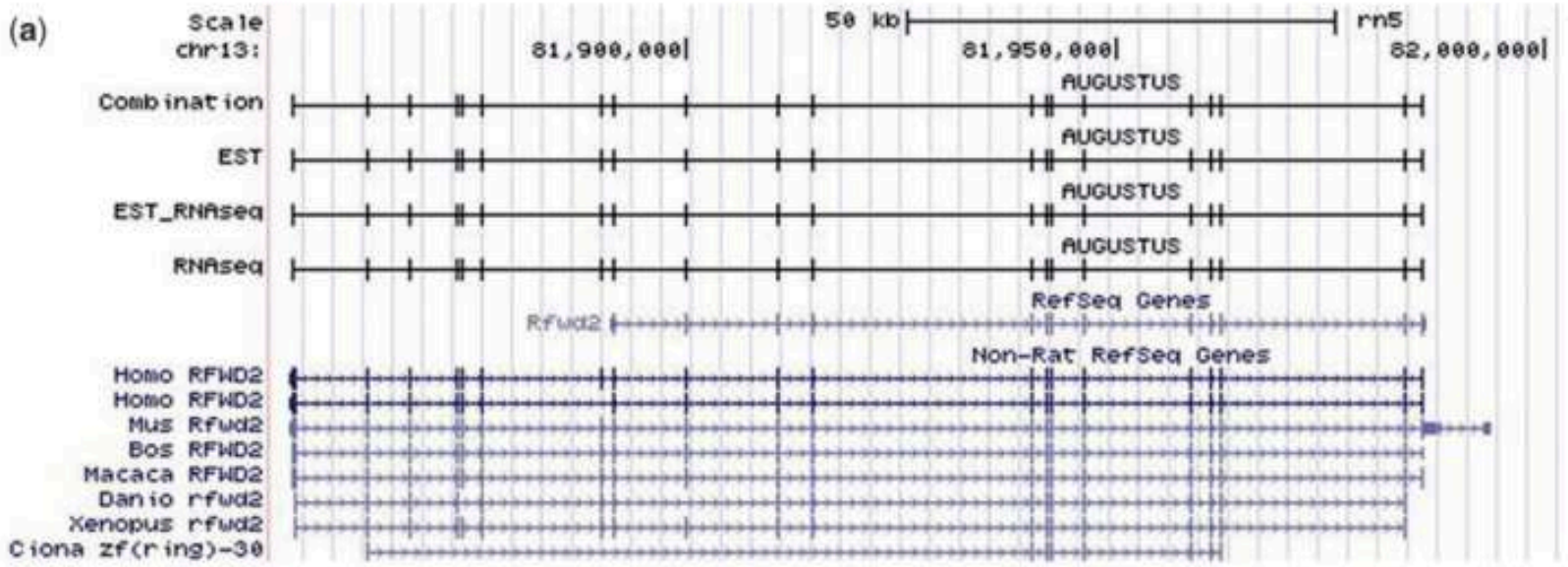
Where to find initial “correct” genes

BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs

Felipe A. Simão[†], Robert M. Waterhouse[†], Panagiotis Ioannidis, Evgenia V. Kriventseva and Evgeny M. Zdobnov*

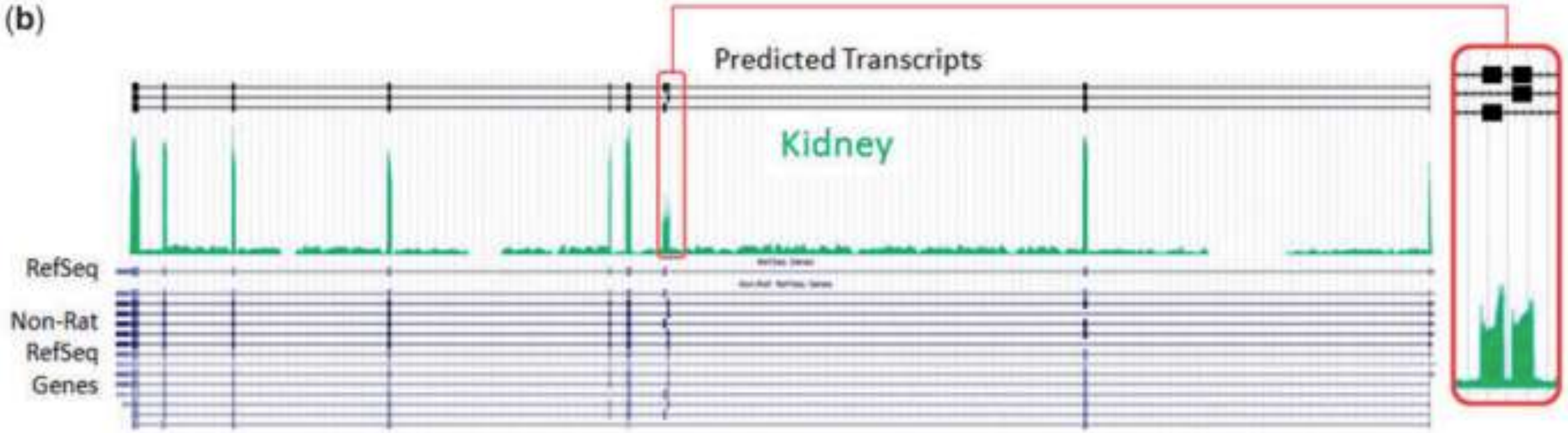


Combine multiple evidence will improve annotation



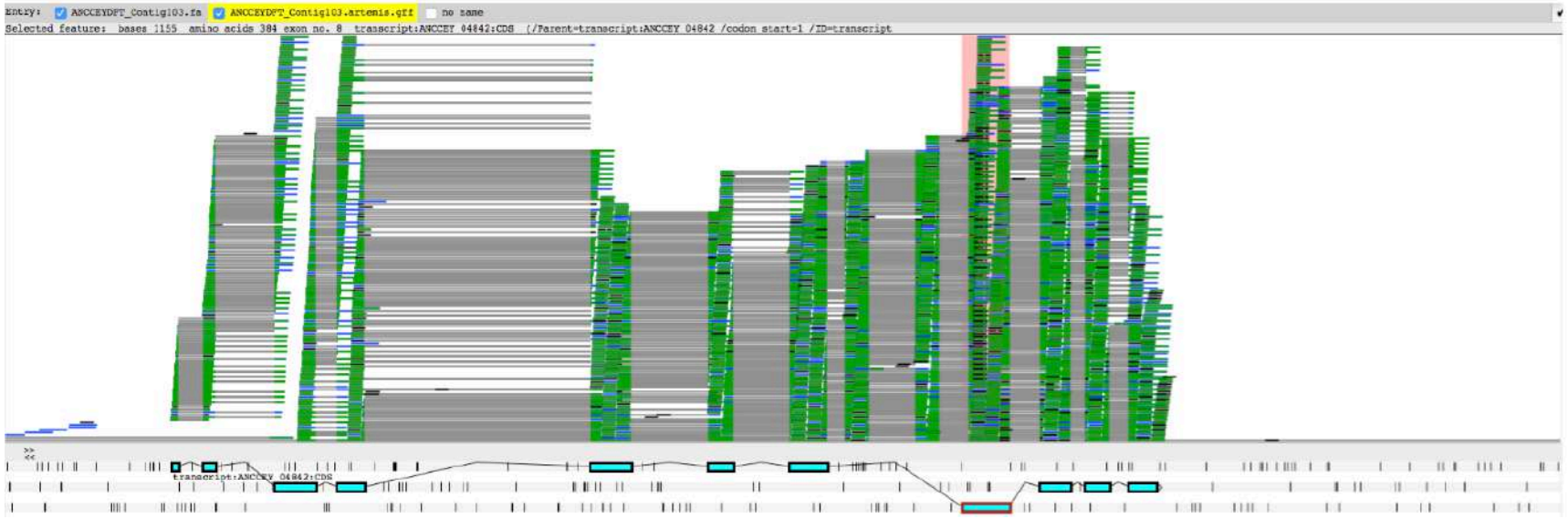
Li et al (2015)

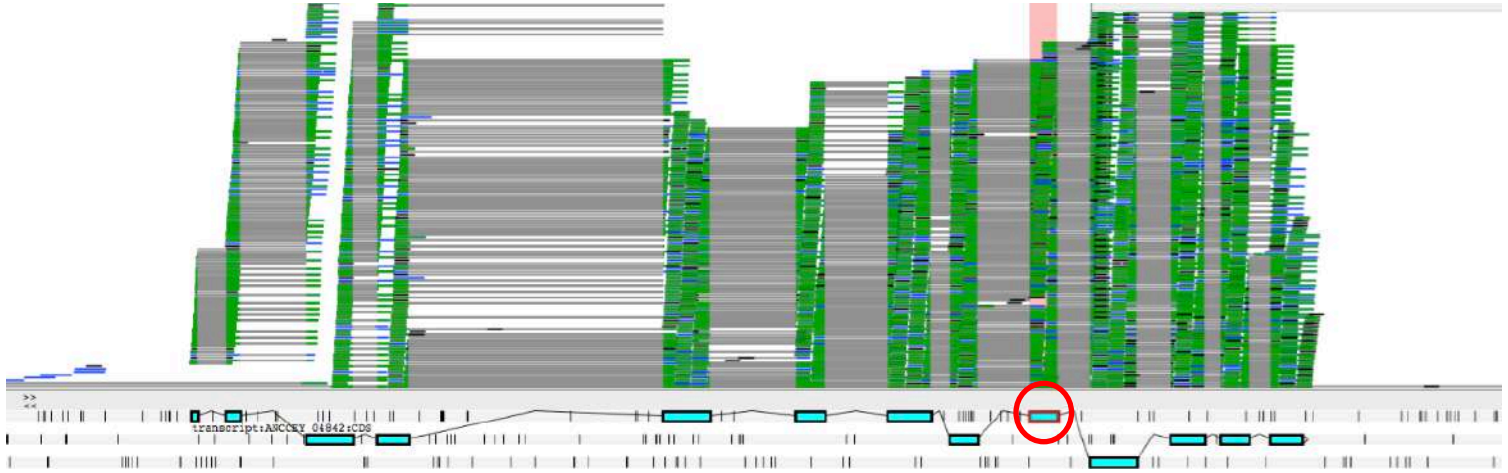
Novel isoforms



Li et al (2015)

Manual curation using artemis

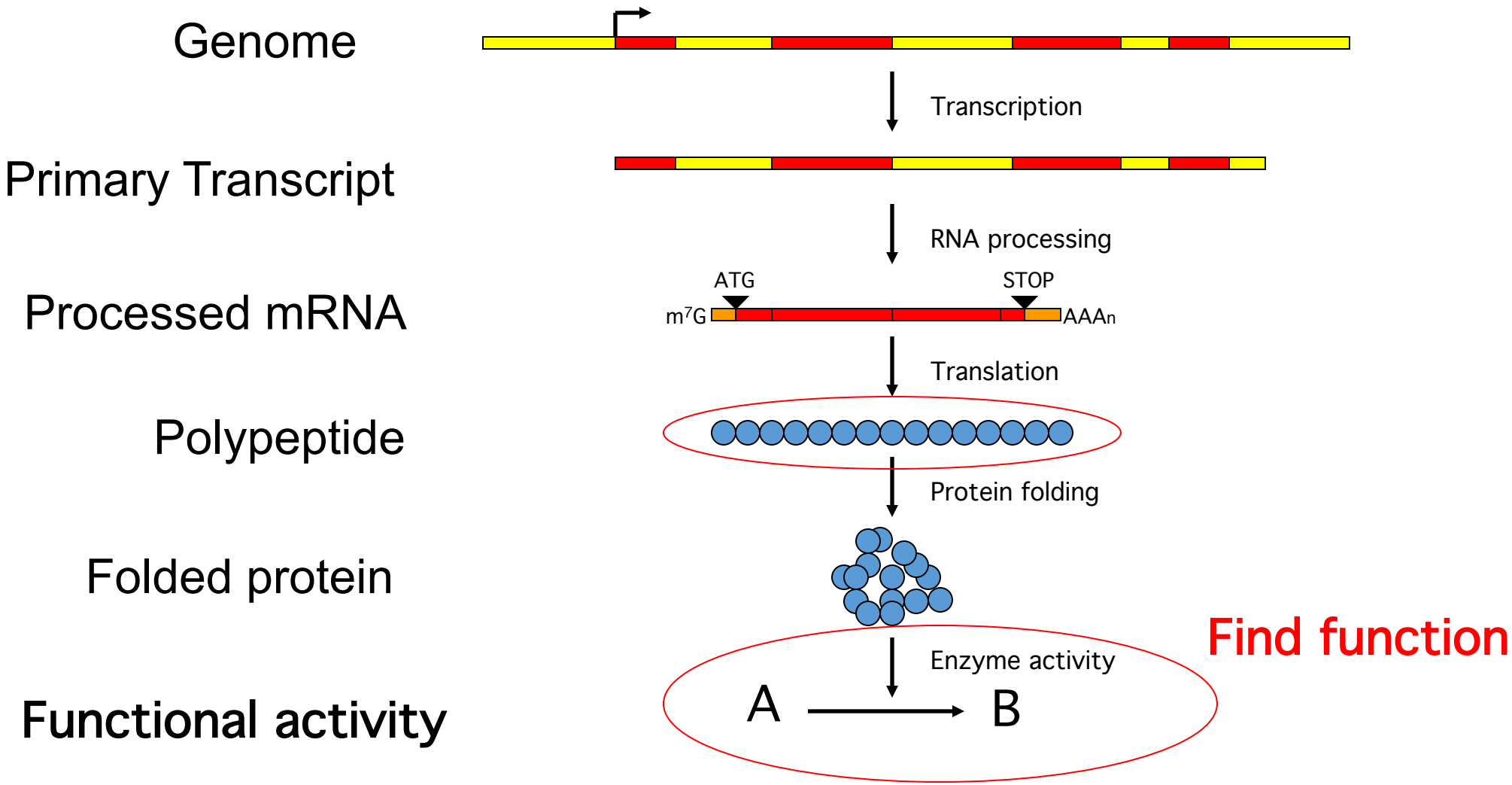




Hueimien Ke

Functional annotation

Functional annotation



Functional annotation

Name the protein correctly

Attaching biological information to genomic elements

- Biochemical function
- Biological function
- Involved regulation and interactions
- Expression

- Utilize known structural annotation to predicted protein sequence

Functional annotation – Homology Based

Most common way

Predicted Exons/CDS/ORF are searched against the non-redundant protein database (NCBI, SwissProt) to search for similarities

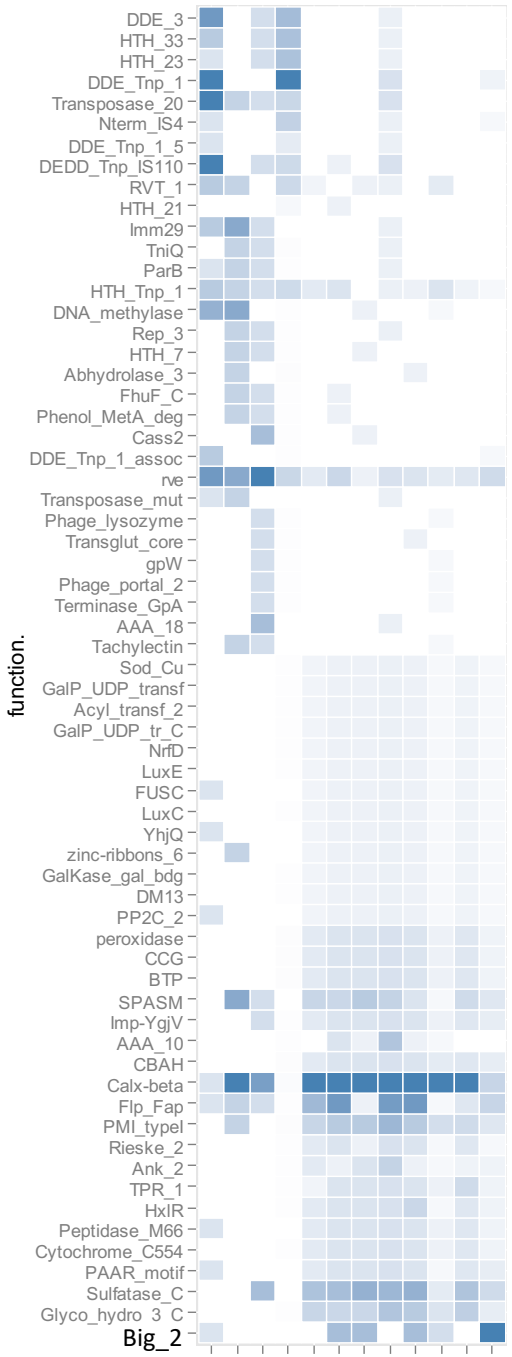
Visually assess the **top 5-10 hits** to identify whether these have been assigned a function

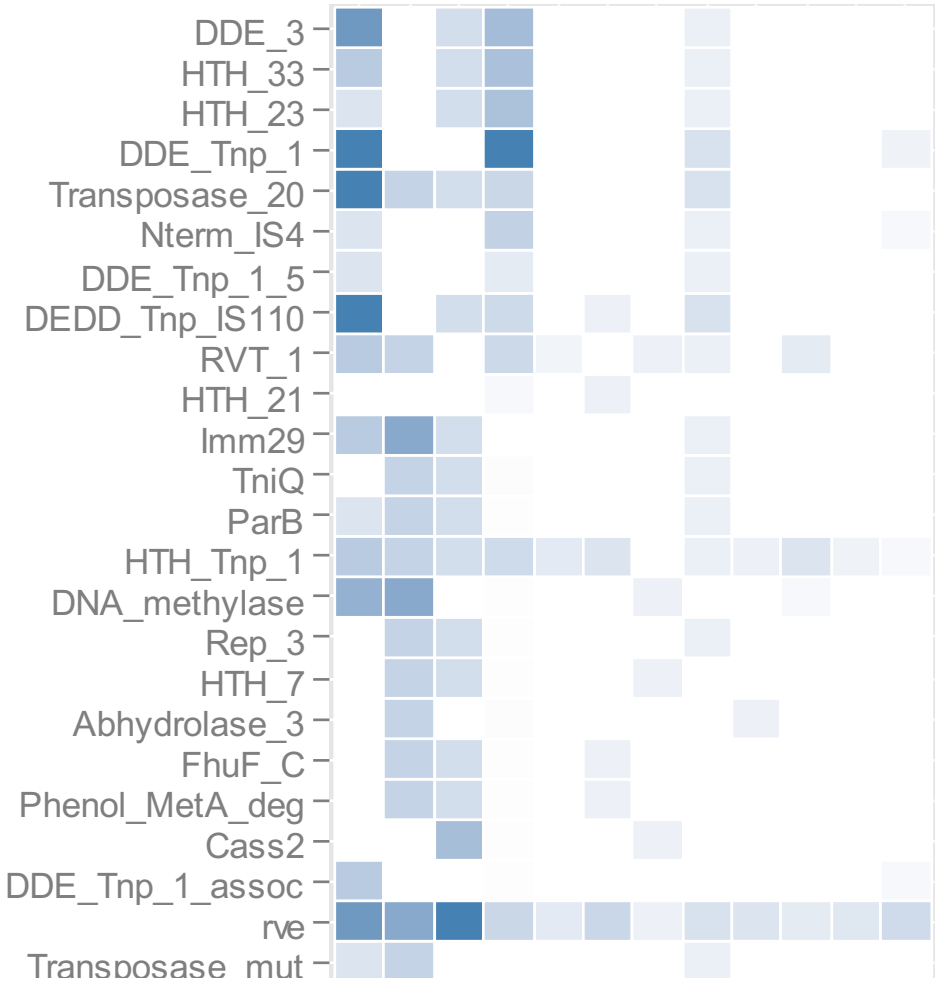
Functions (**and names**) are assigned

Other features which can be determined

- Signal peptides
- Transmembrane domains
- Low complexity regions
- Various binding sites, glycosylation sites etc.
- Protein Domain
- Secretome

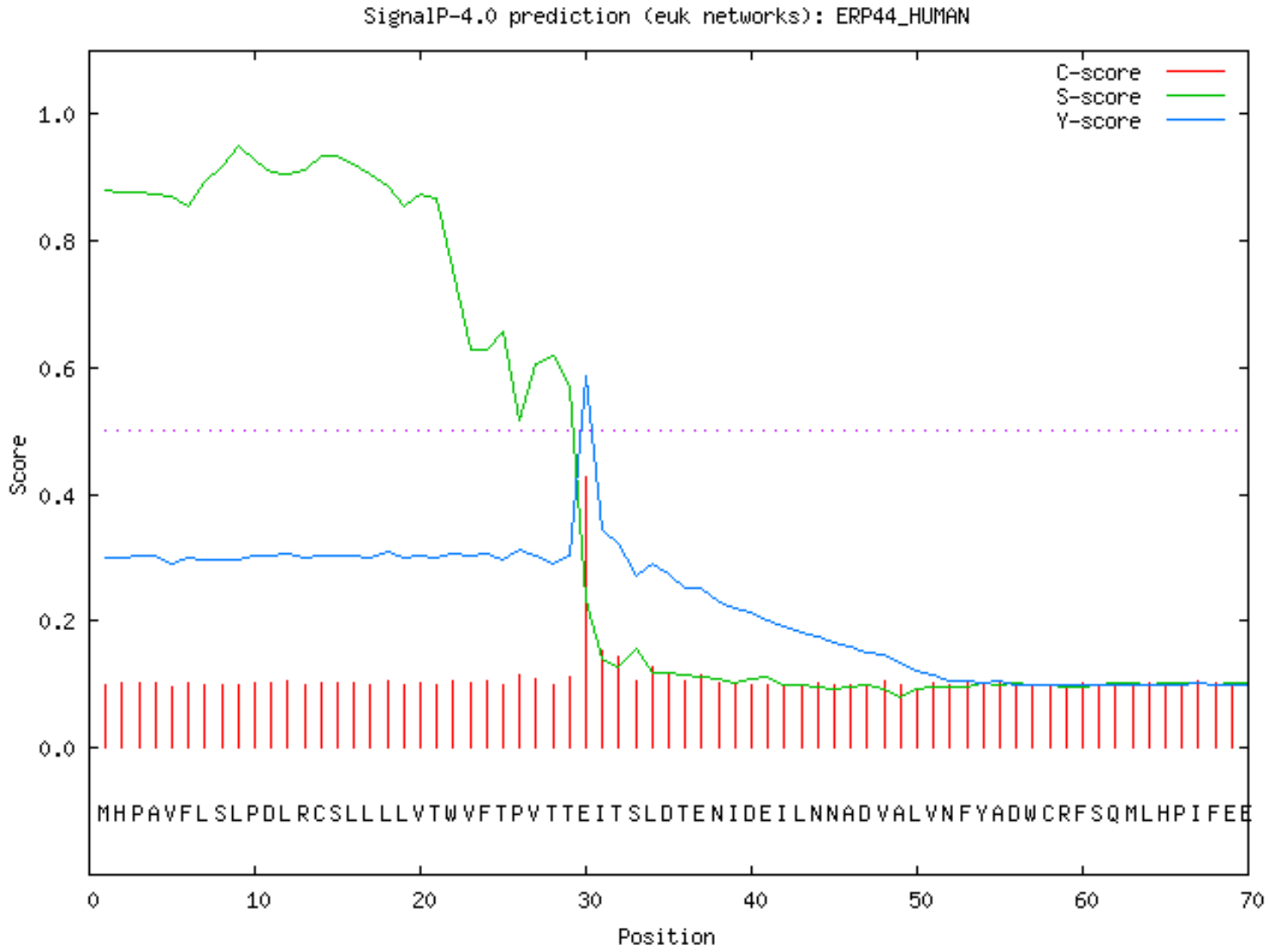
PFAM





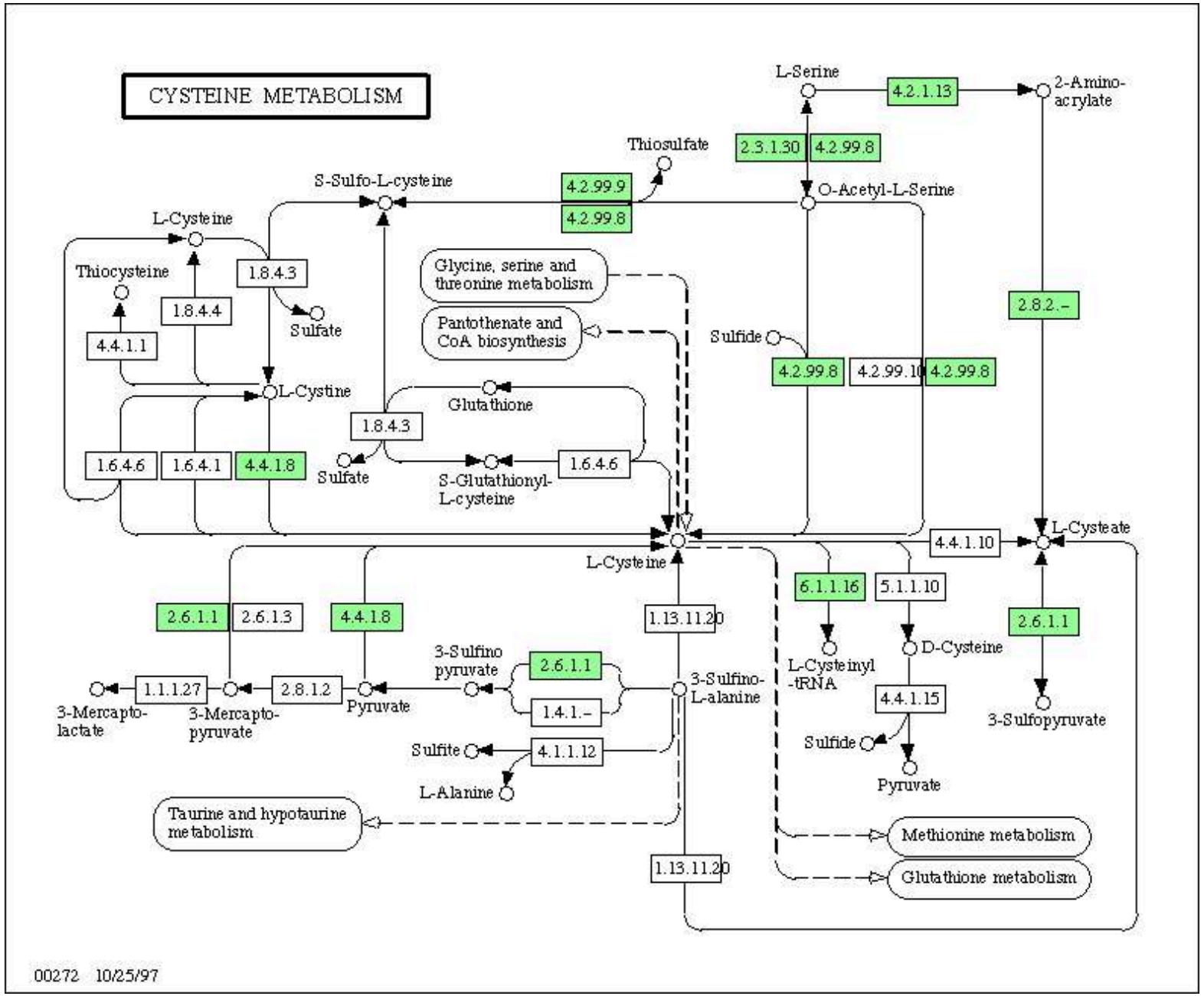
Hueimien Ke

SignalP: predicts the presence and location of signal peptide



KEGG

Help improve annotation by showing missing genes in essentail pathways



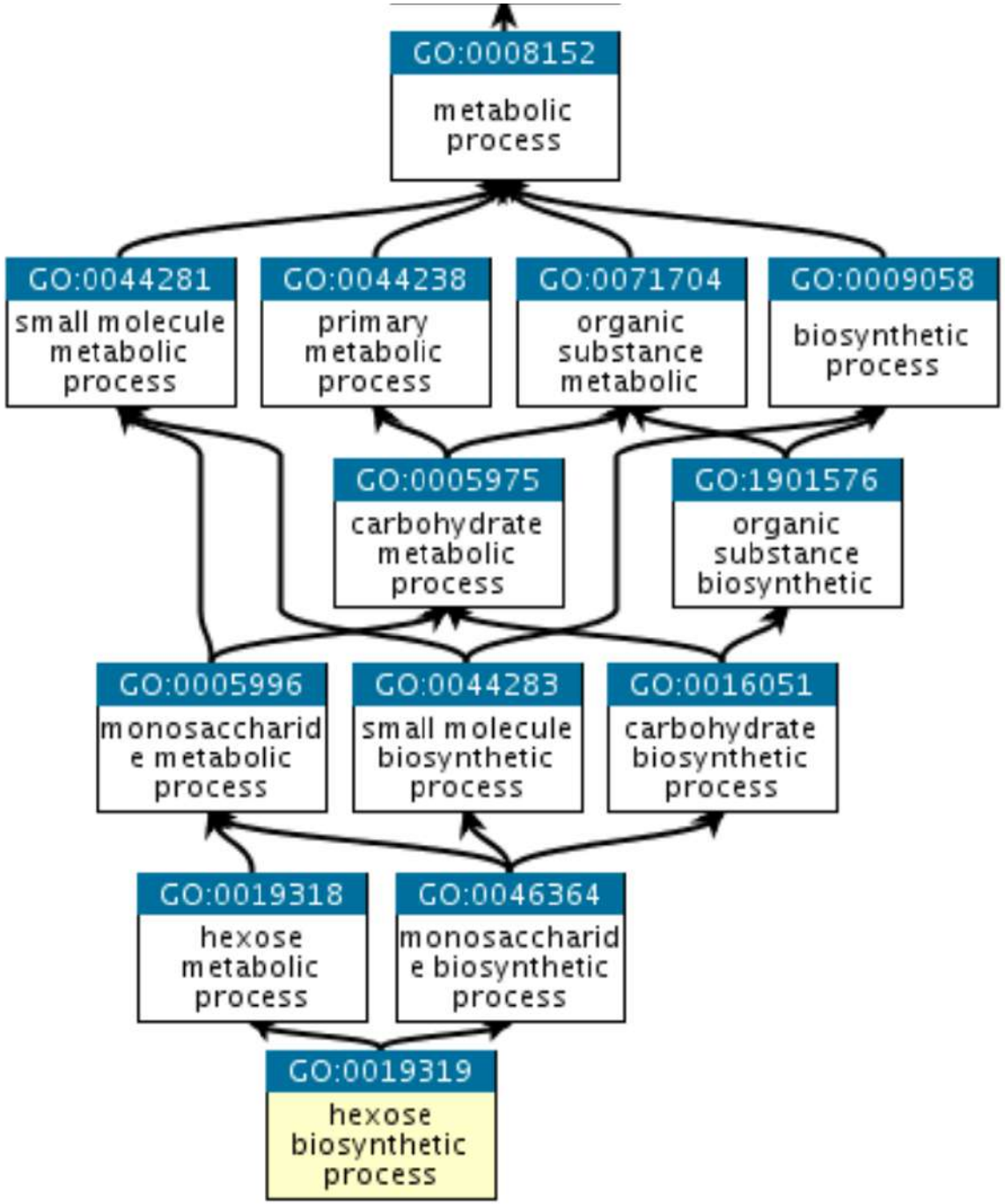
Gene Ontology

- A controlled vocabulary for annotating three aspects of a gene product's biology:
- Biological Process (BP) – the molecular, cellular, and organismal level processes in which a gene product is involved
- Molecular Function (MF) – the molecular activity of a gene product
- Cellular Component (CC) – the subcellular localization of a gene product

Gene Ontology

“An ontology is a formal representation of a body of knowledge within a given domain. Ontologies usually consist of a set of classes (or terms or concepts) with relations that operate between them.”

“GO is loosely hierarchical, with ‘child’ terms being more specialized than their ‘parent’ terms, but unlike a strict hierarchy, a term may have more than one parent term.”



http://geneontology.org/docs/ontology-documentation/

BLAST2GO

Blast2GO PRO

start · blast · interpro · mapping · annot · charts · graphs · select

nr	SeqName	Description	Length	#Hits	e-Value	sim mean	#GO	GO list	Enzyme list	InterPro Scan
1	C0401...	mpk3_arath ame: full=mitogen-activated prote...	717	20	5.3E-144	87.3%	0	-	-	-
2	C0401...	protein	706							
3	C0401...	protein	620							
4	C0401...	class iv chitinase	715							
5	C0401...	cyti_vigun ame: full=cysteine proteinase inhibi...	663							
6	C0401...	protein phosphatase 2c	663							
7	C0401...	protein	578							
8	C0401...	lgul_orysj ame: full=lactoylglutathione lyase a...	600							
9	C0401...	mt2_actde ame: full=metallothionein-like prote...	625							
10	C0401...	protein	612							
11	C0401...	protein phosphatase	645							

Welcome Message | Blast Result of C04018A12

Query Name (Length): C04018A12 (715)
E-Value Cutoff: 0.001
Annotation: -

Sequences Producing Significant Alignments	Gene Name
gi|3608477|gb|AAC35981.1|chitinase CHI1 [Citrus sinensis]	
gi|33414046|gb|AAP03085.1|class IV chitinase [Galega orientalis]	
gi|225434068|ref|XP_002275122.1|PREDICTED: hypothetical protein [Vitis vinifera]	
gi|157353727|emb|CAO46259.1|unnamed protein product [Vitis vinifera]	
gi|225434052|ref|XP_002274620.1|PREDICTED: hypothetical protein [Vitis vinifera] gi|157353719|emb|CAO46251.1| unnamed protein product [Vitis vinifera]	LOC100250948

Run Blast

Blast Options

Please choose one option.

CloudBlast

CloudBlast is a cloud-based Blast2GO PRO Community Resource for massive sequence alignment tasks. It allows you to execute standard NCBI Blast+ searches directly from within Blast2GO PRO in our dedicated computing cloud. This is a high-performance, secure and cost-optimized solution for your analysis. Check your available ComputationUnits under Menu -> View -> Window -> CloudBlast Activity Monitor.

NCBI Blast

Use the public NCBI Blast service to blast your sequences against public databases. Two protocols are available: QBlast and RemoteBlast. Performance and results depend on the NCBI Blast web service.

AWS Blast

The NCBI provides via Amazon Web Services (AWS) a preconfigured machine image (AMI) which contains the latest BLAST+ release. This AMI downloads and caches automatically popular NCBI databases such as nr, nt, swissprot, refseq, and PDB. This Blast option allows you can access your AMIs directly via Blast2GO. Simply provide the URL of your AMI and run Blast searches in the Amazon Cloud.

Local Blast

Use NCBI blast+ software to perform Blast searches locally on your PC against a local database. Use an own, formatted database or download a pre-formatted sequences database from the NCBI (ftp.ncbi.nlm.nih.gov/blast/db). Simply select the database you want to blast against and run your blast searches locally.

Default | < Back | Next > | Cancel | Run

Europe, Germany: DE2 Version: b2g_sep14 /Users/sgoetz/b2gWorkspace/blast2go_project_20141104_2248.dat

Case study: eukaryote annotation (2018)

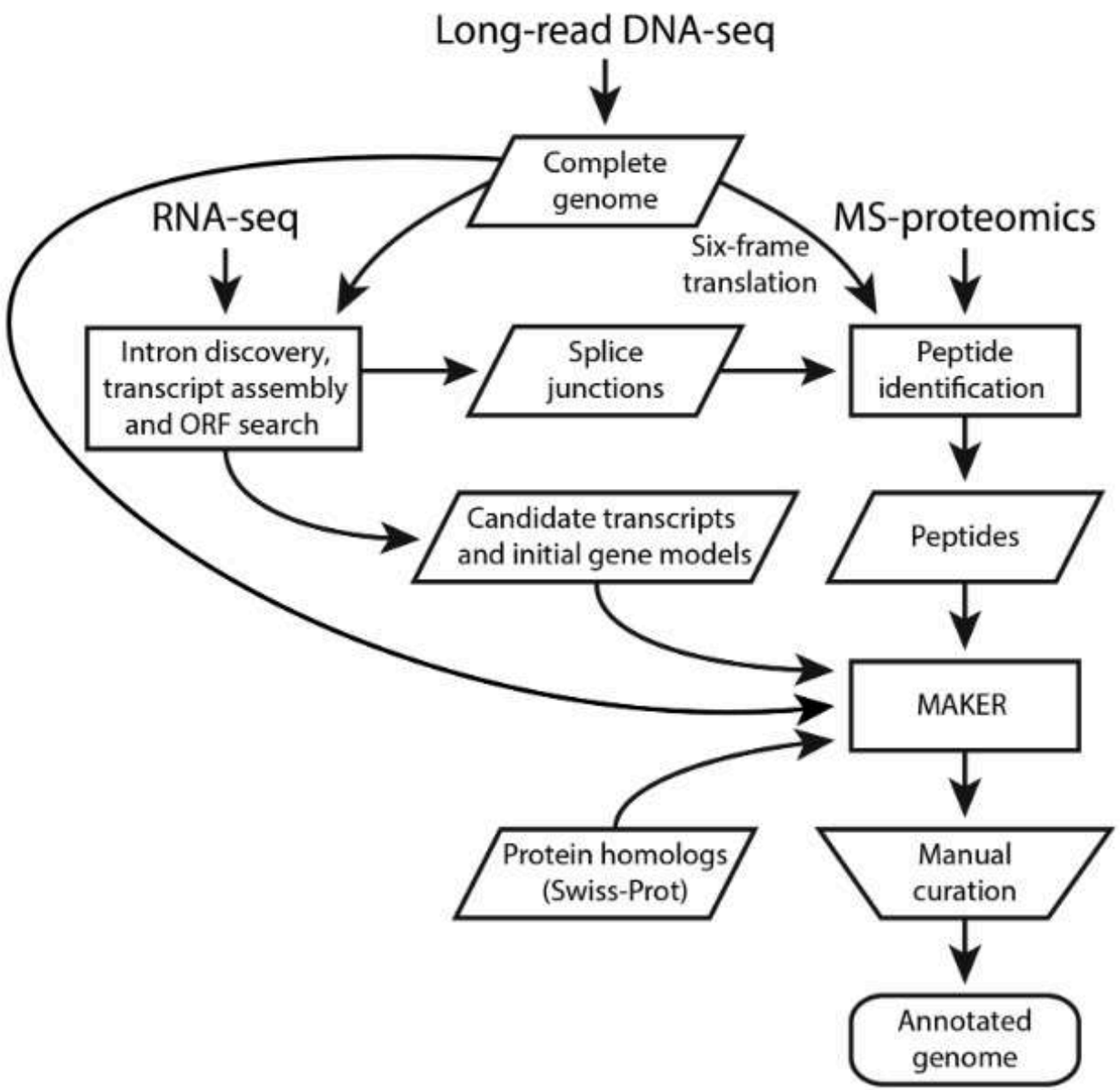
Published online 18 January 2017

Nucleic Acids Research, 2017, Vol. 45, No. 5 **2629–2643**
doi: 10.1093/nar/gkx006

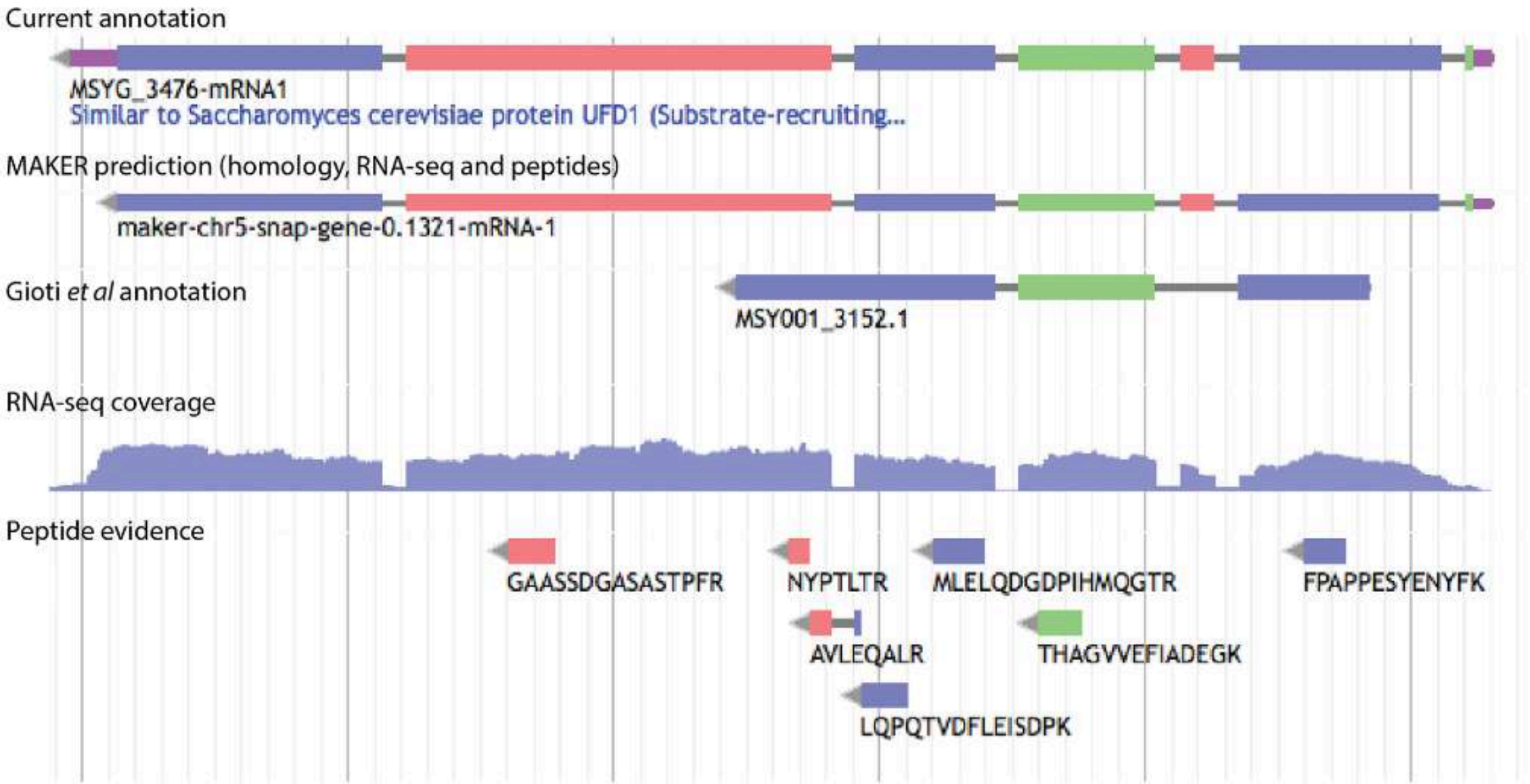
Proteogenomics produces comprehensive and highly accurate protein-coding gene annotation in a complete genome assembly of *Malassezia sympodialis*

Yafeng Zhu[1,†], Pär G. Engström[2,†], Christian Tellgren-Roth[3], Charles D. Baudo[4], John C. Kennell[4], Sheng Sun[5], R. Blake Billmyre[5], Markus S. Schröder[6], Anna Andersson[7], Tina Holm[7], Benjamin Sigurgeirsson[8], Guangxi Wu[9], Sundar Ram Sankaranarayanan[10], Rahul Siddharthan[11], Kaustuv Sanyal[10], Joakim Lundeberg[8], Björn Nystedt[12], Teun Boekhout[13], Thomas L. Dawson, Jr.[14], Joseph Heitman[5], Annika Scheynius[15,*,‡] and Janne Lehtiö[1,*,‡]

Zhu *et al* (2017) Nucleic Acids Research



Zhu *et al* (2017) Nucleic Acids Research



Zhu *et al* (2017) Nucleic Acids Research

Table 2. Characteristics of *M. sympodialis* gene sets

	Published (MAKER with homology evidence) (5)	MAKER with homology and RNA-seq evidence	MAKER with homology, RNA-seq and peptide evidence	Manually curated annotation
Protein-coding genes	3536	3612	4113	4493
Gene density (genes/kb)[1]	0.46	0.46	0.53	0.58
Coding sequence (Mb)	5.40	5.35	6.14	6.72
Coding exons	6995	8453	9212	9793
Introns	3462	5030	5267	5350
Mean exon size (bp)[2]	772	635	669	687
Mean intron size (bp)	65	52	50	30
Genes supported by peptides	3176	3176	3674	3891
Introns supported by RNA-seq	1661 (48%)	4246 (84%)	4275 (81%)	5271 (99%)
Out-frame peptides	4658 (13%)	5453 (15%)	1796 (5%)	338 (1%)

[1]Gene density was computed relative to the size of the corresponding genome assembly (7.71 Mb for the draft assembly of Gioti *et al.* (5) and 7.79 Mb for the current assembly).
[2]Excluding untranslated regions.

Zhu *et al* (2017) Nucleic Acids Research

Case study: Annotation using long reads

Check for updates

Breakthrough Technologies

Long-Read Annotation: Automated Eukaryotic Genome Annotation Based on Long-Read cDNA Sequencing[1[OPEN]]

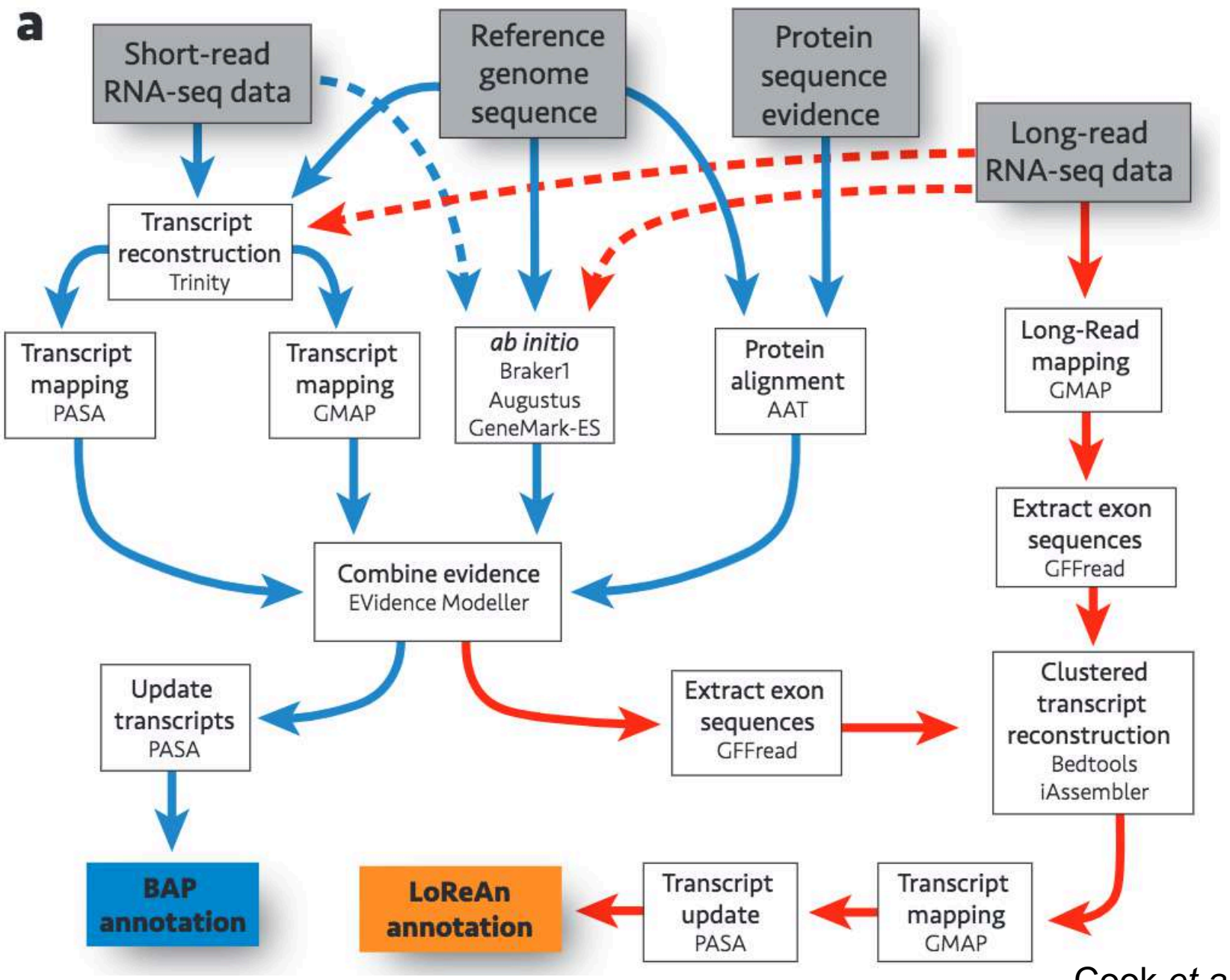
David E. Cook,[a,2,3] Jose Espejo Valle-Inclan,[a,2,4] Alice Pajoro,[b,5] Hanna Rovenich,[a,6] Bart P. H. J. Thomma,[a,7,8,9] and Luigi Faino[a,10,7]

[a]Laboratory of Phytopathology, Wageningen University and Research, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands

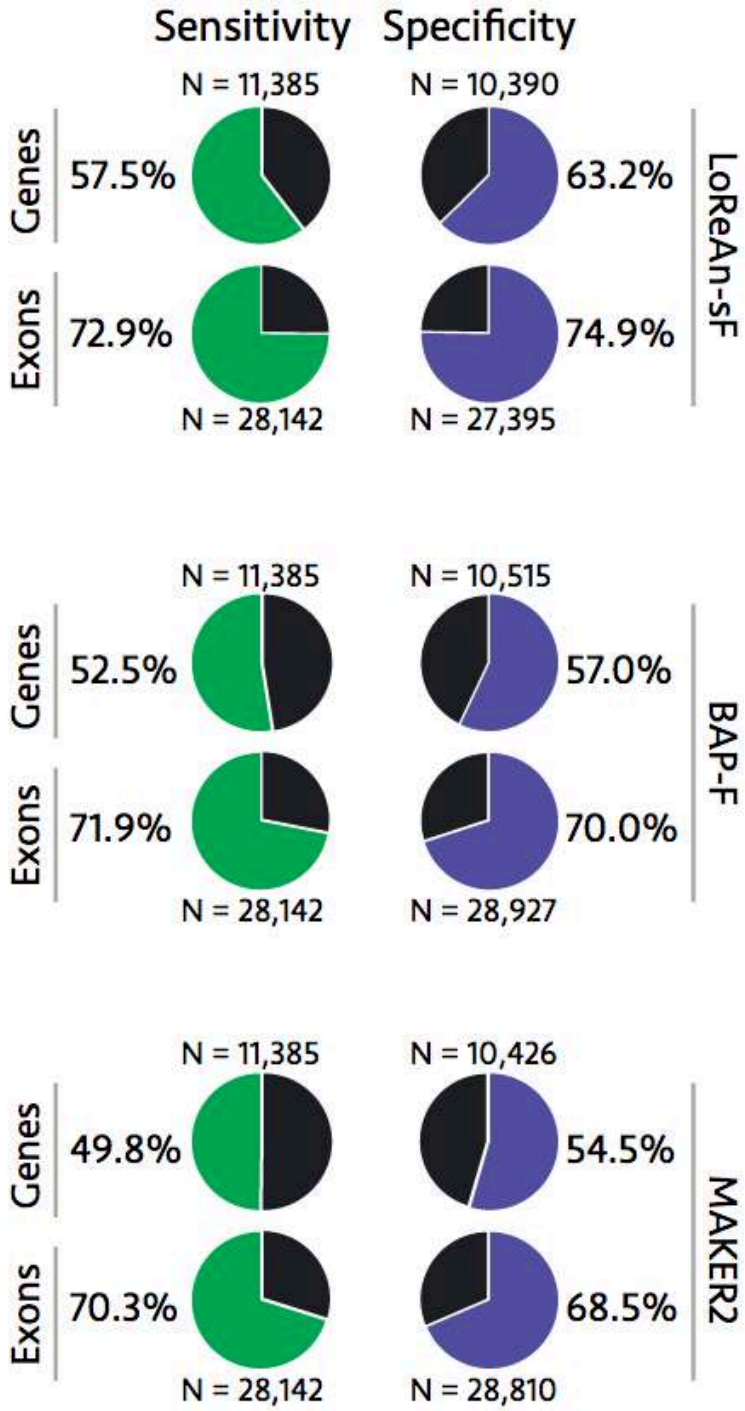
[b]Laboratory of Molecular Biology, Wageningen University and Research, Droevendaalsesteeg 1, 6708 PB Wageningen, the Netherlands

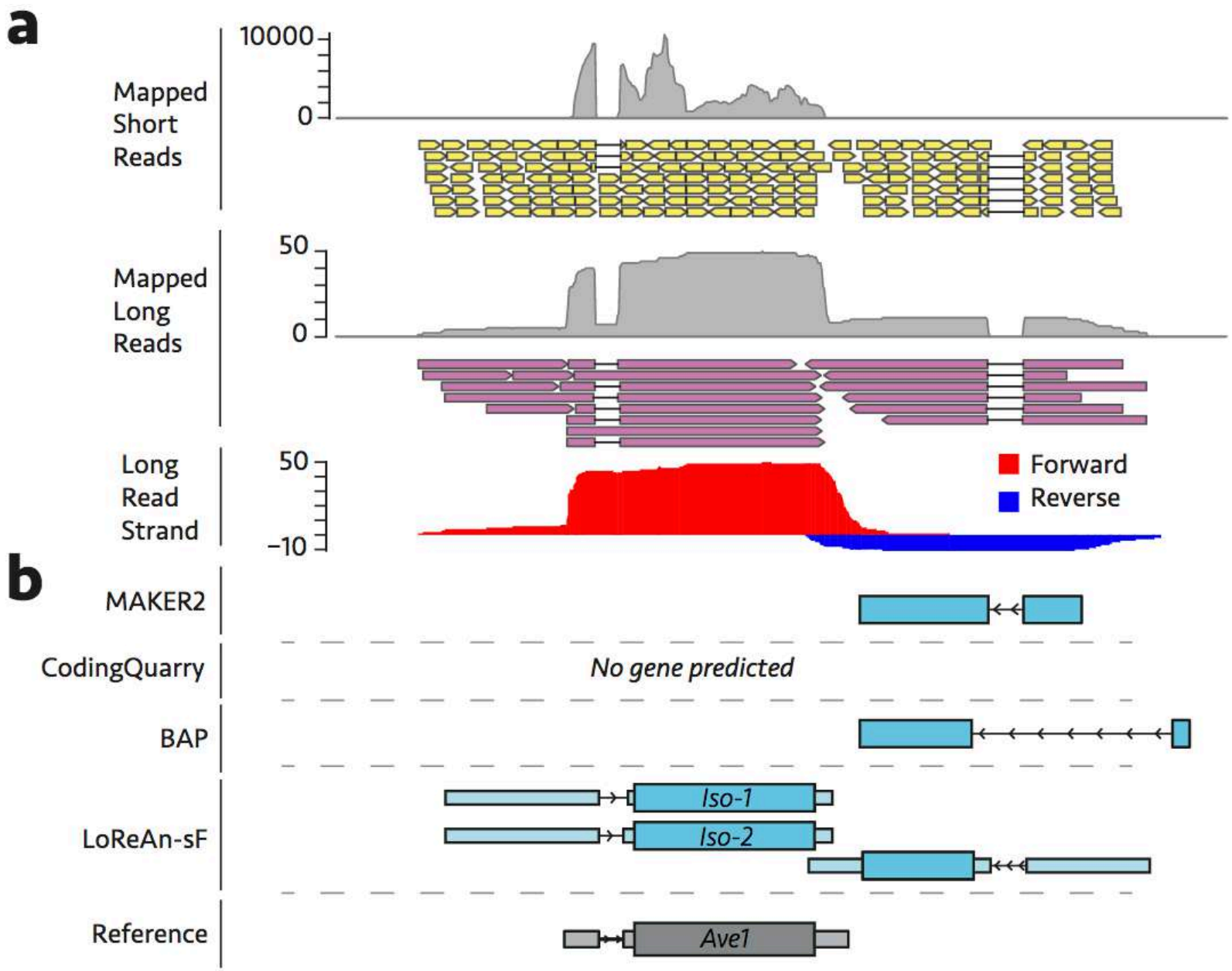
ORCID IDs: 0000-0002-2719-4701 (D.E.C.); 0000-0002-4857-5984 (J.E.V.); 0000-0003-4125-4181 (B.P.H.J.T.); 0000-0002-6807-4191 (L.F.).

Cook *et al* (2017) Plant Physiology



Cook *et al* (2017) Plant Physiology





Cook *et al* (2017) Plant Physiology

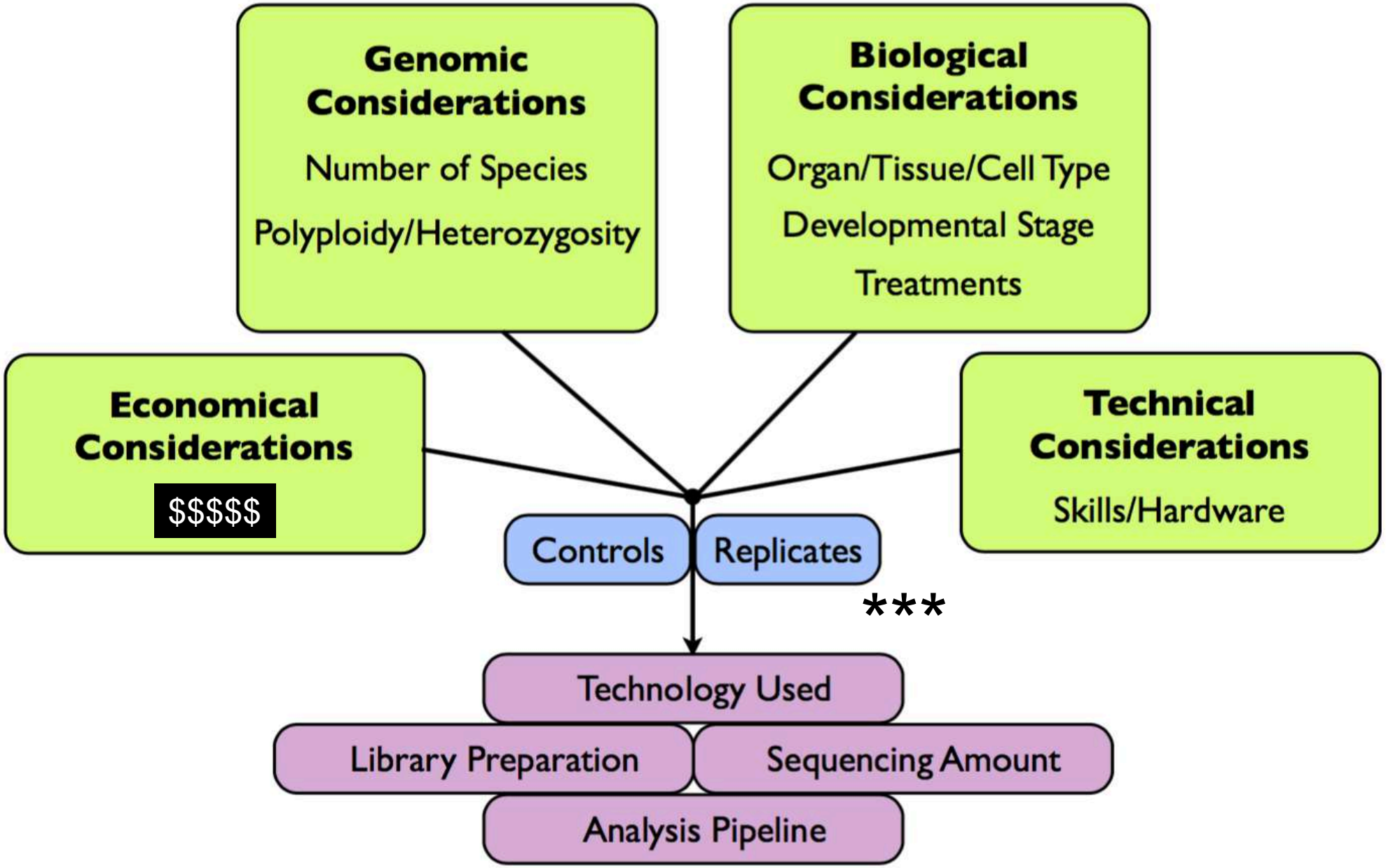
Break

3. Differential expression

Types of experiments

- One genome or multiple genomes (Host / pathogens)
- Multiple alleles
  - High heterozygosity
  - Polyploidy
  - Gene families
- Isoforms?
- Organ / Tissue / Cell type specific
  - Laser Capture Microdissection
  - single cell transcriptomics [not discussed]
- Time points
  - Development
  - Response to treatment (before, during, after)

Experimental design



How many reads are enough?

Genohub

	Small (bacteria)	Intermediate (fruit fly, worm	Large (mouse, human)
No. of reads for DGE ($x10^6$)	5 SR	10 SR	20–50 SR
No. of reads for *de novo* transcriptome assembly ($x10^6$)	30–65 PE	70–130 PE	100–200 PE
Read length (bp)	50	50–100	>100

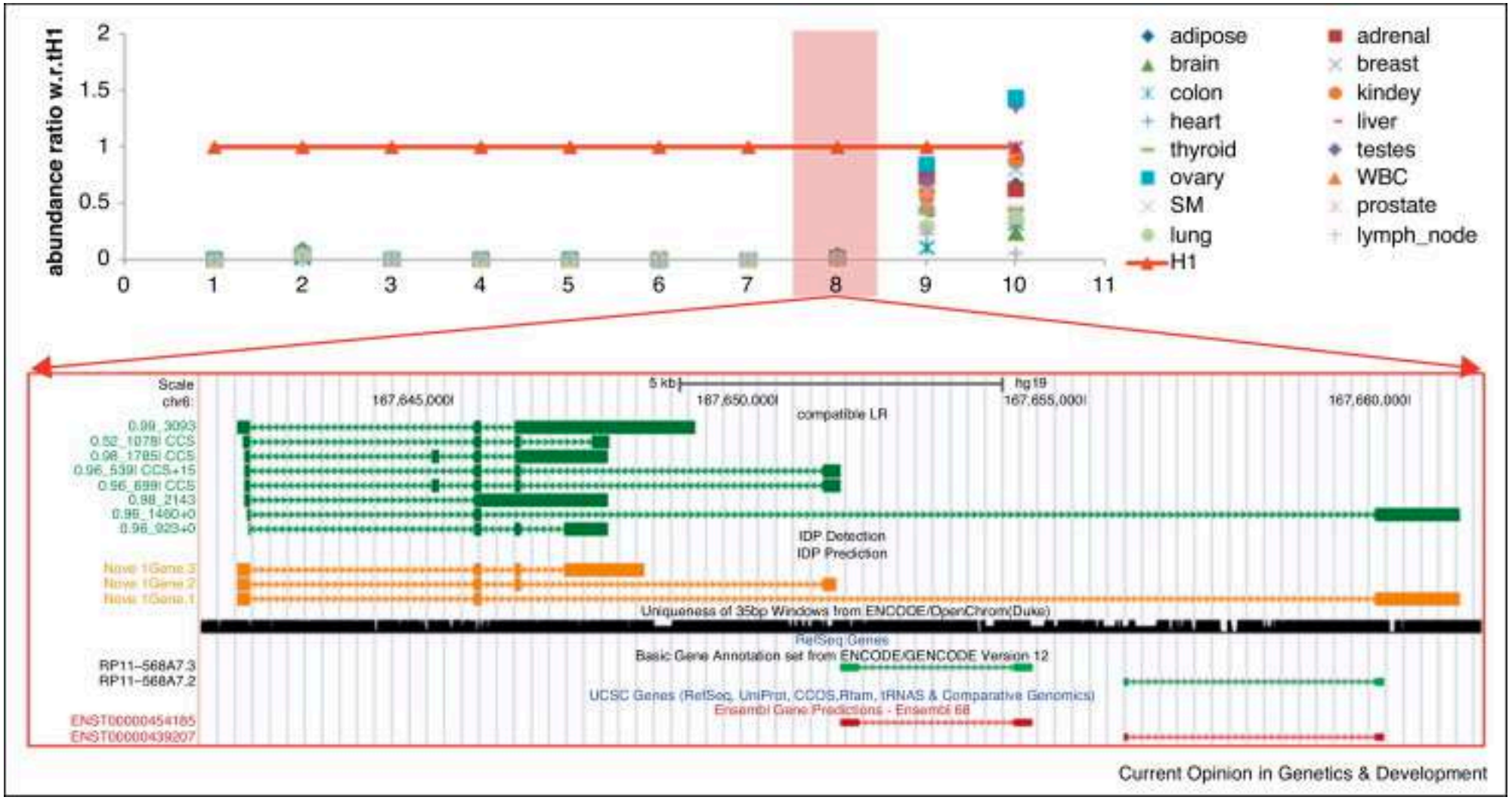
Illumina

- Gene expression profiling experiments that are looking for a quick snapshot of highly expressed genes may only need 5–25 million reads per sample. In these cases, consider pooling multiple RNA-Seq samples into one lane of a sequencing run. This allows for high multiplexing of samples.
- Experiments looking for a more global view of gene expression, and some information on alternative splicing, typically require 30–60 million reads per sample. This range encompasses most published RNA-Seq experiments for mRNA/whole transcriptome sequencing.

[1] https://genohub.com/next-generation-sequencing-guide/#depth2

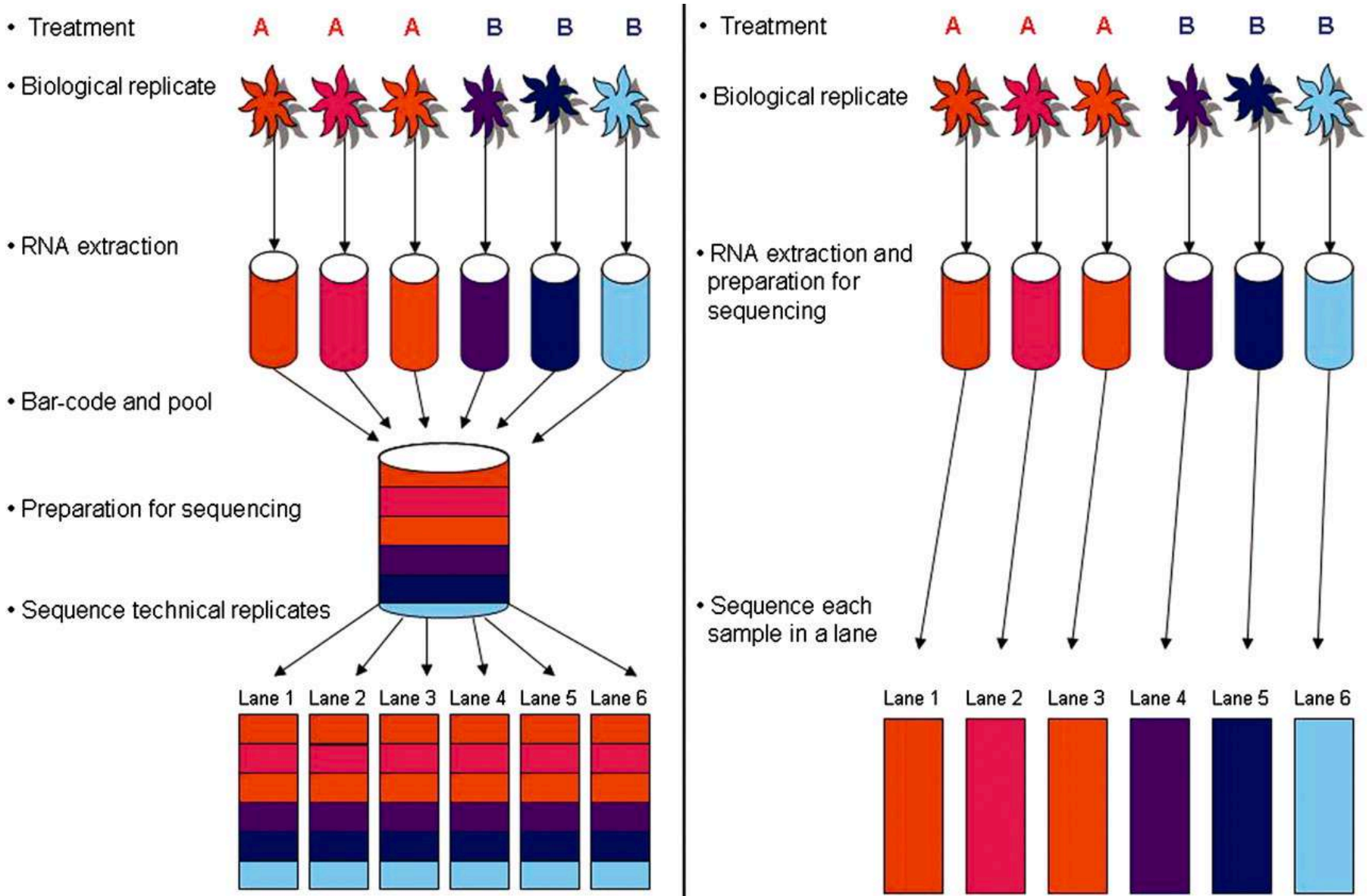
[2] https://support.illumina.com/bulletins/2017/04/considerations-for-rna-seq-read-length-and-coverage-.html

For isoform discovery, longer sequences are better



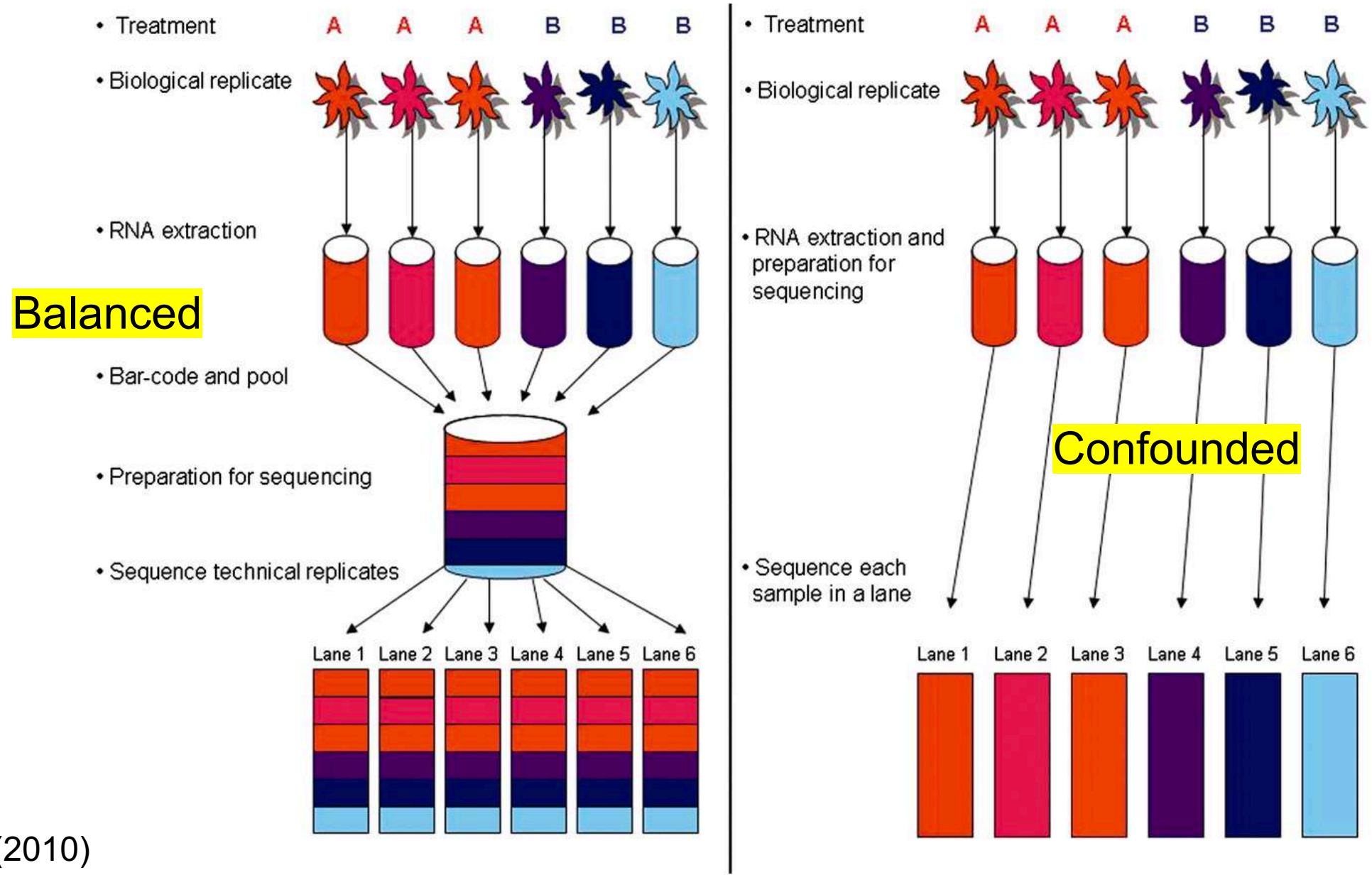
Au *et al.*, (2014)

Which of the following designs is correct?



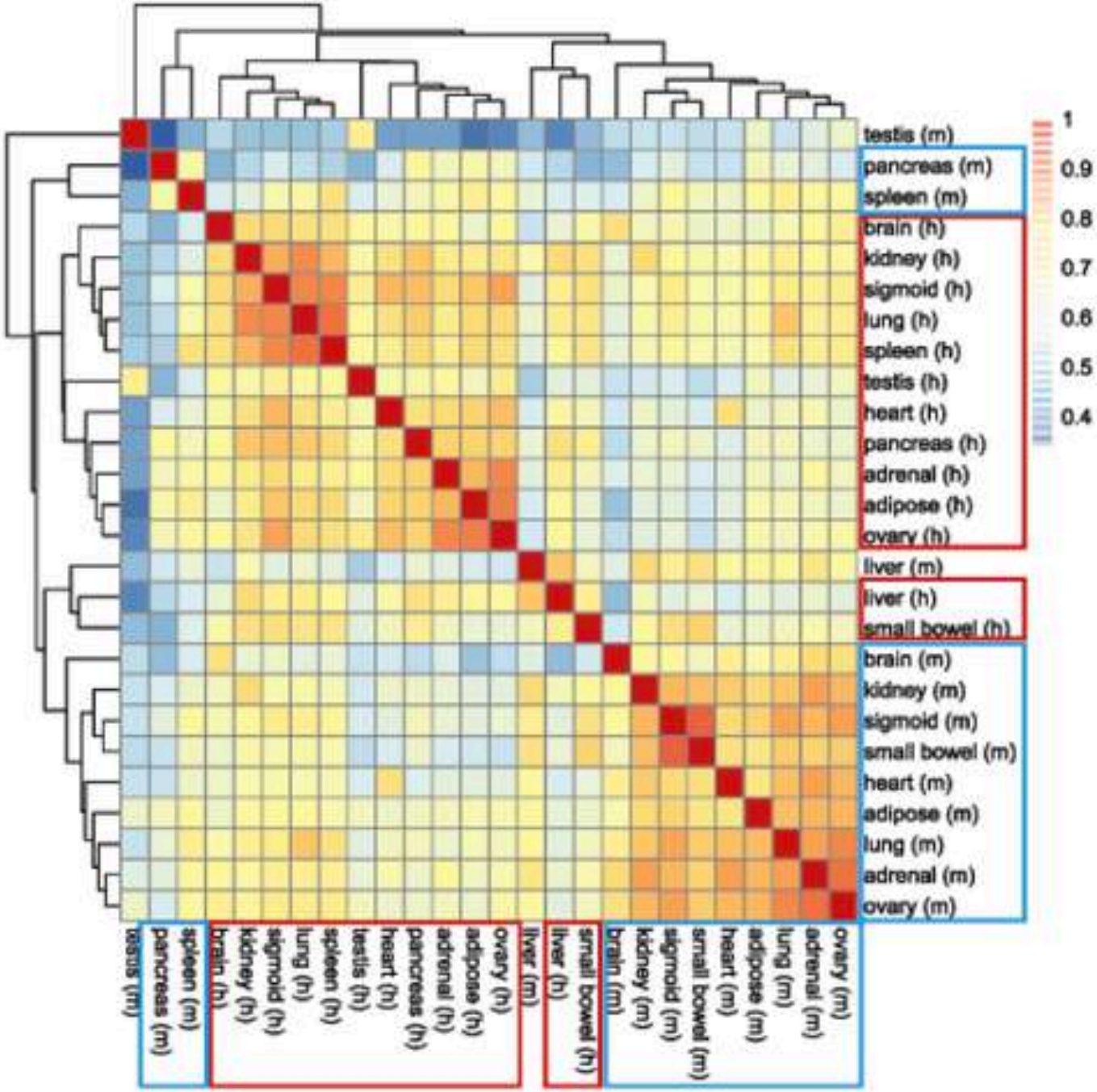
Auer *et al.,* Genetics (2010)

Which of the following designs is correct?



Auer *et al.*, Genetics (2010)

Example of batch effect:



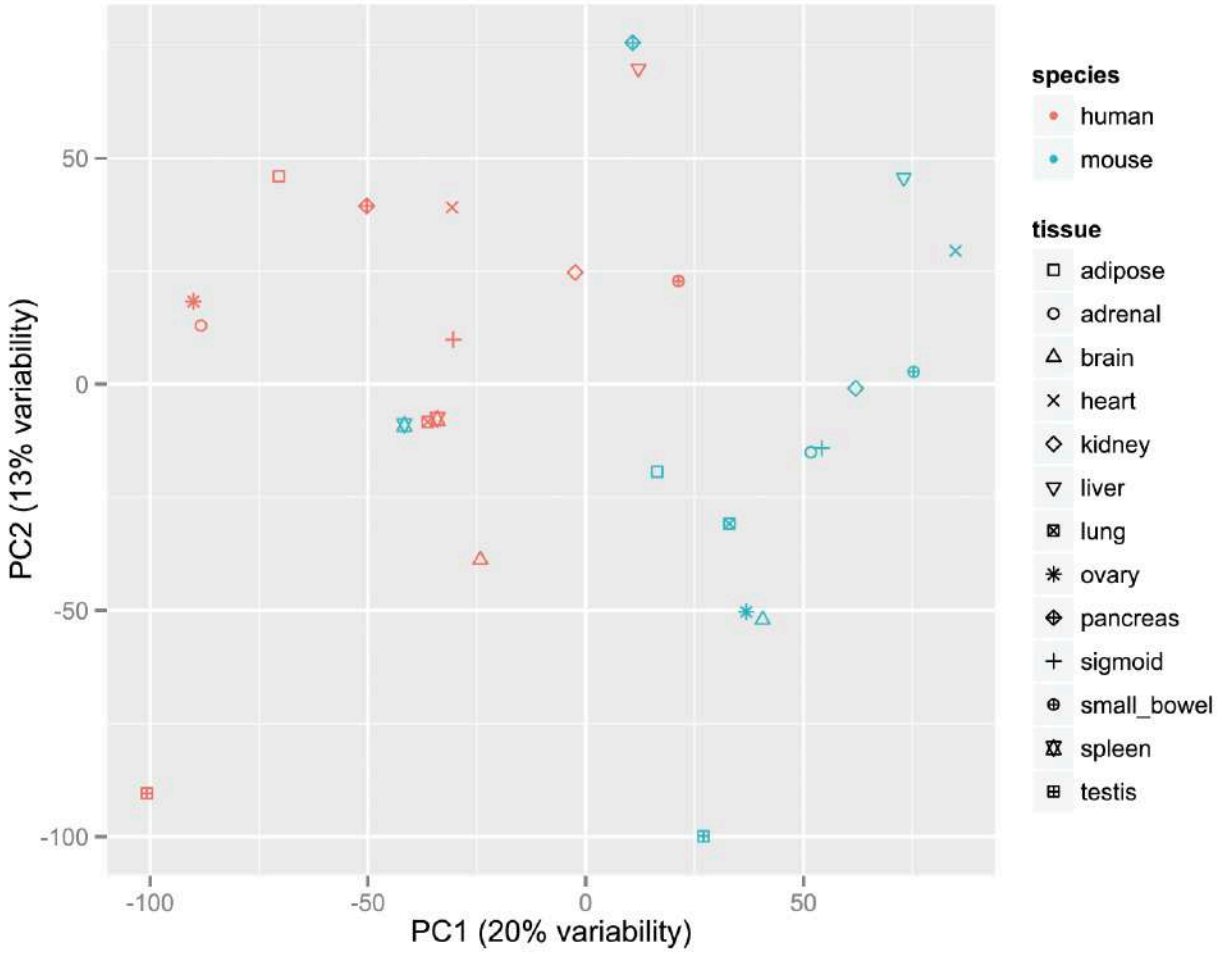
Lin *et al.,* (PNAS) 2014

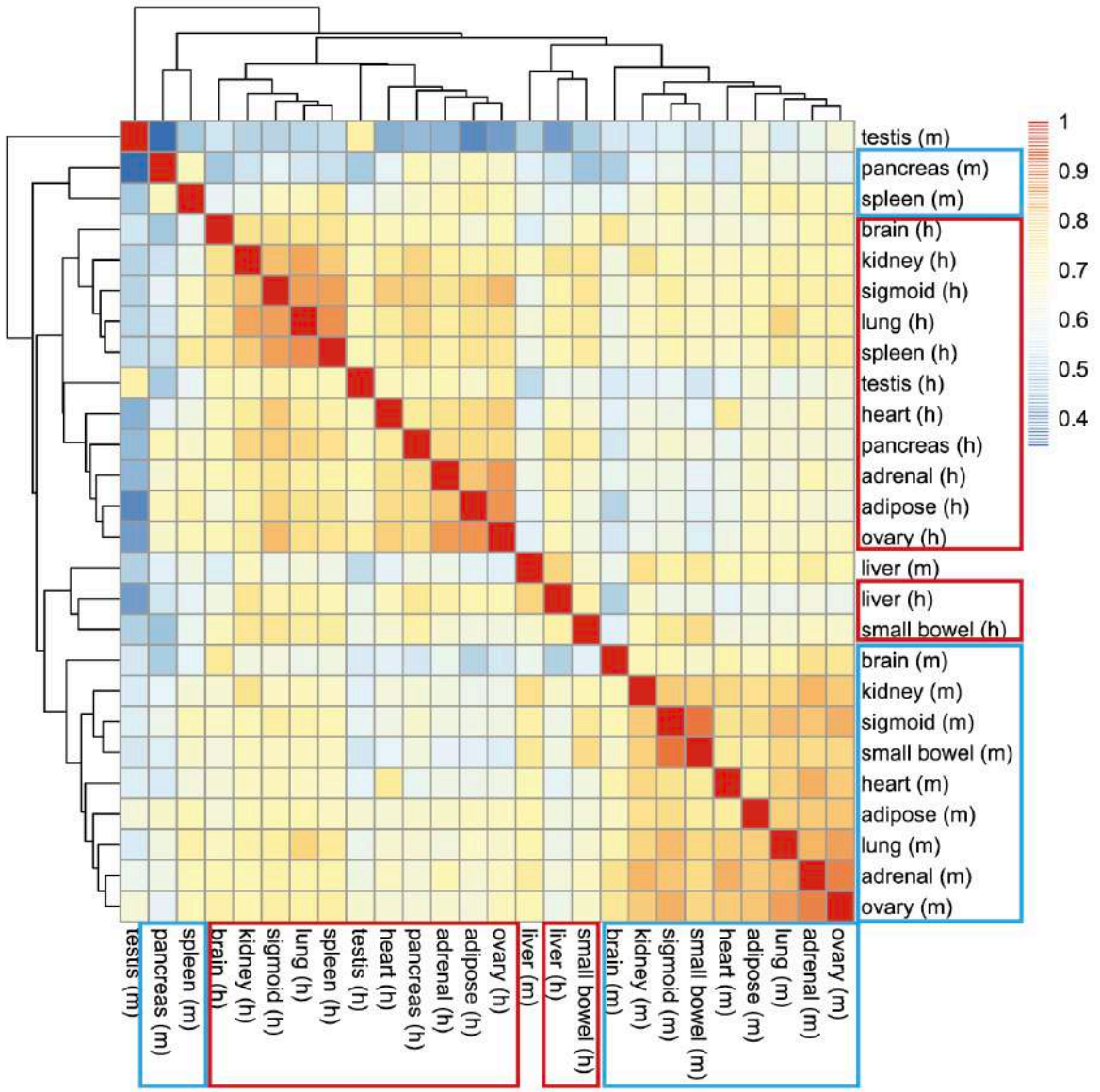
Example of batch effect:

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	● Human
testis		pancreas		● Mouse

Gilad and Mizrahi-Man (2015) F1000Research

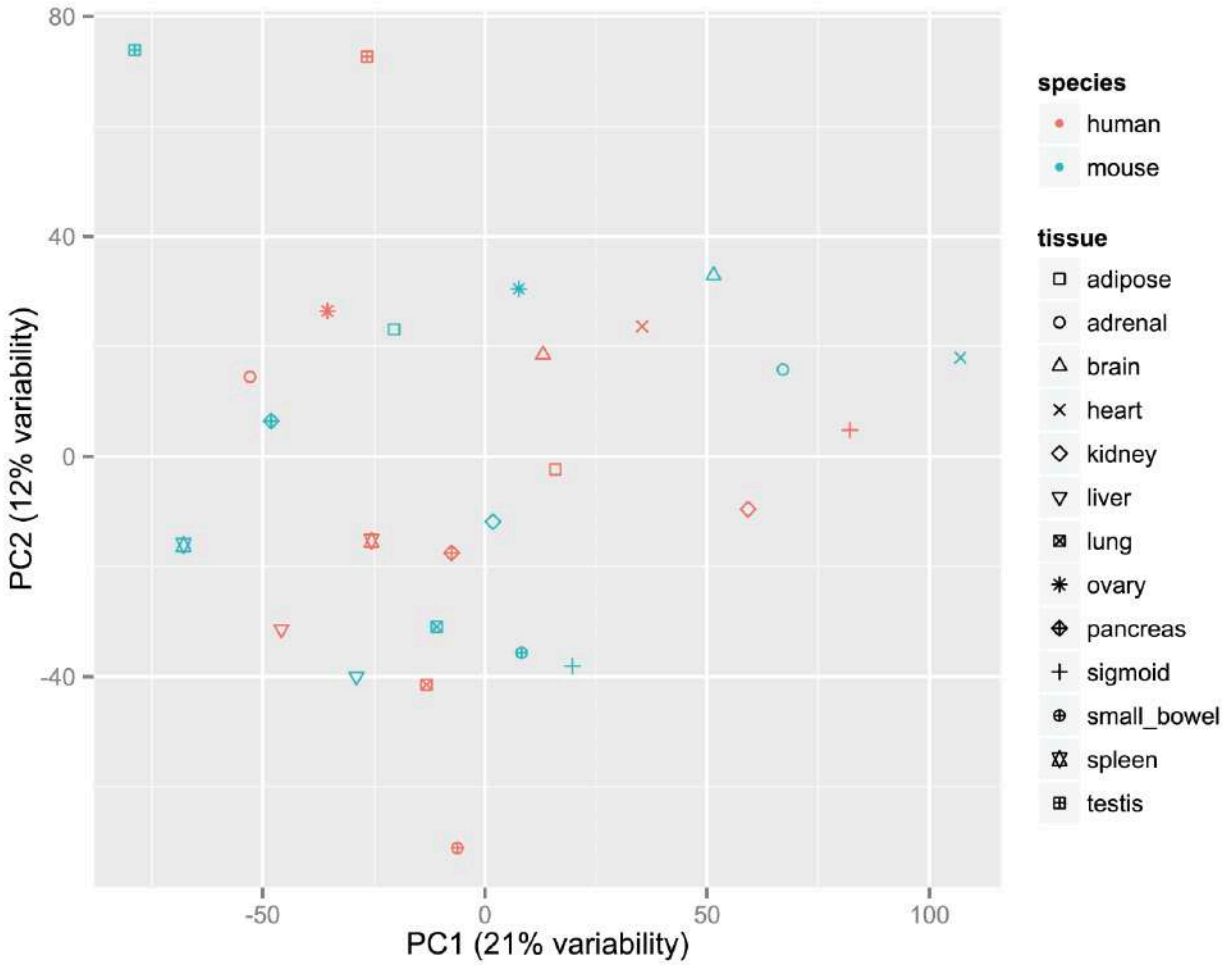
Recapitulating the patterns reported by the mouse ENCODE papers

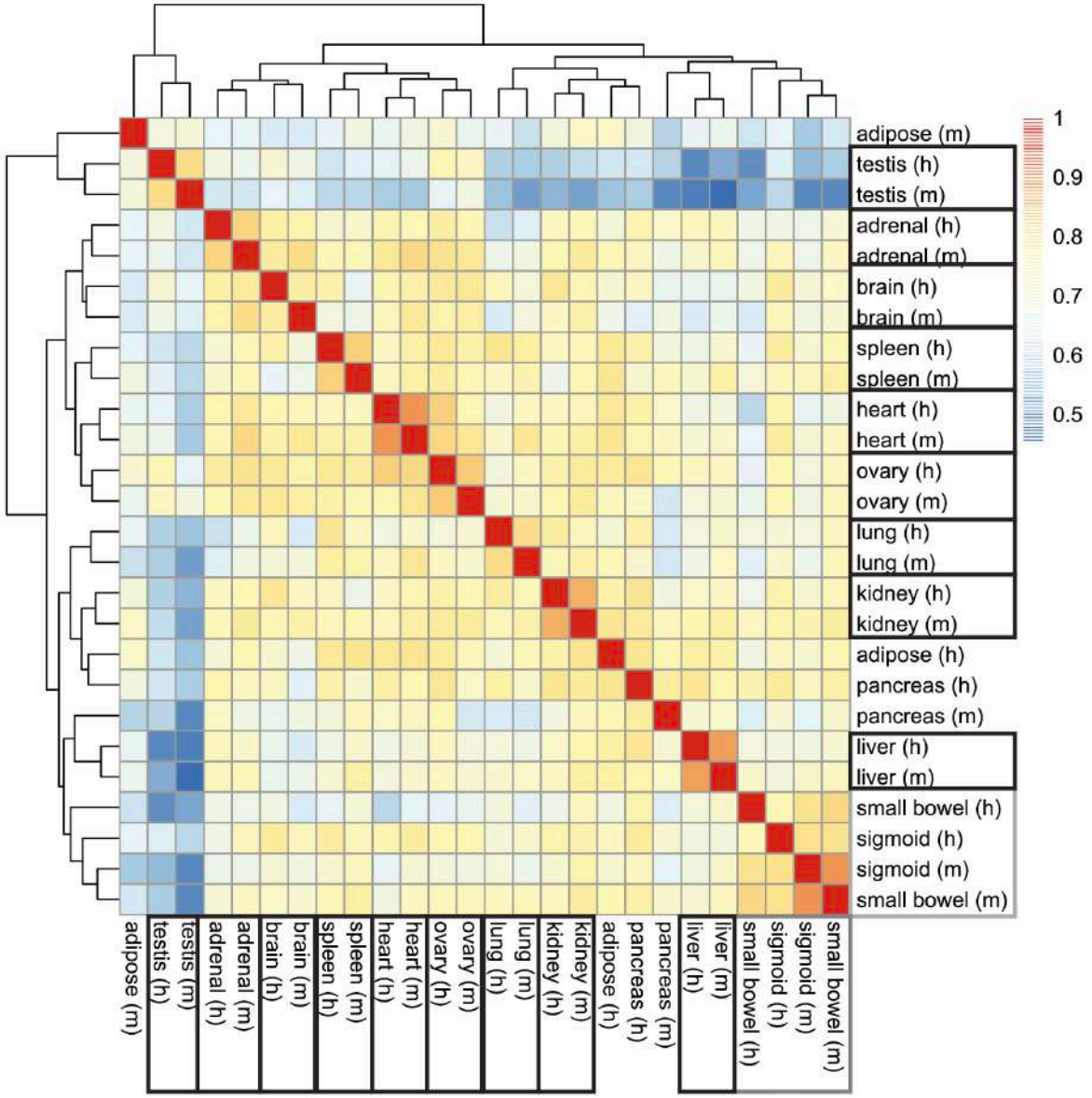




Gilad and Mizrahi-Man (2015) F1000Research

Clustering of data once batch effects are accounted for

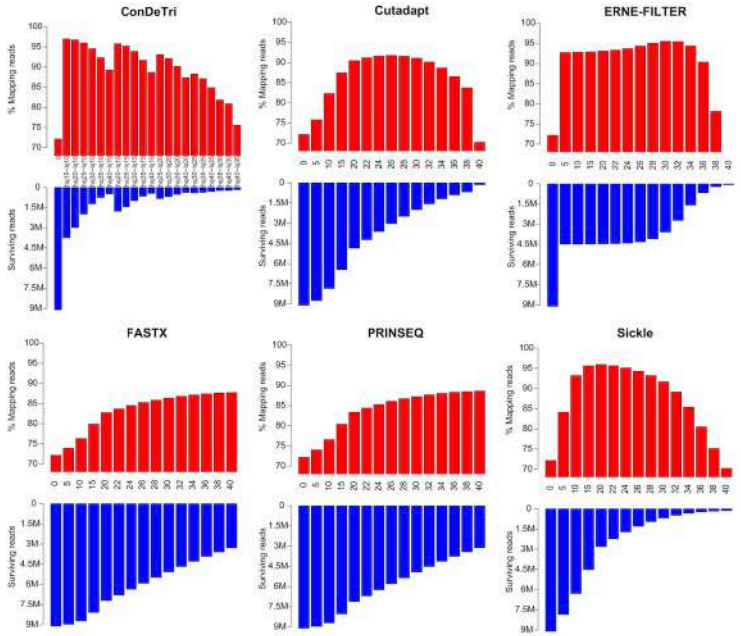




Gilad and Mizrahi-Man (2015) F1000Research

Is trimming beneficial?

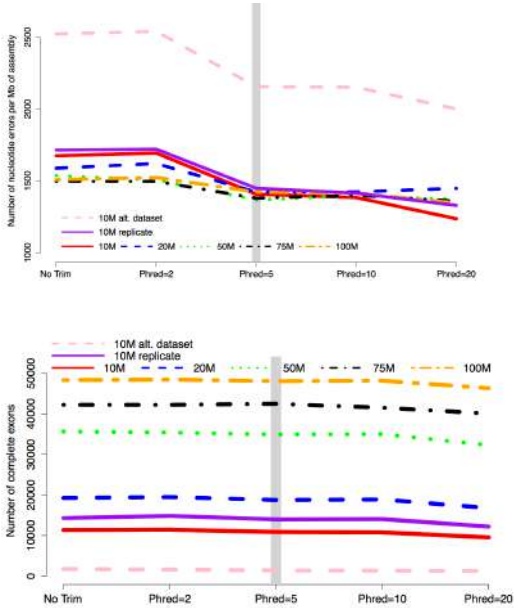
Software comparison, RNA/DNA-Seq



"trimming is beneficial in RNA-Seq, SNP identification and genome assembly procedures, with the best effects evident for intermediate quality thresholds (Q between 20 and 30)"

Del Fabbro C et al (2013) **An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis**. PLoS ONE 8(12): e85024. doi:10.1371/journal.pone.0085024

Assembly-oriented, RNA-seq only



Erroneous bases in assembly

complete exons

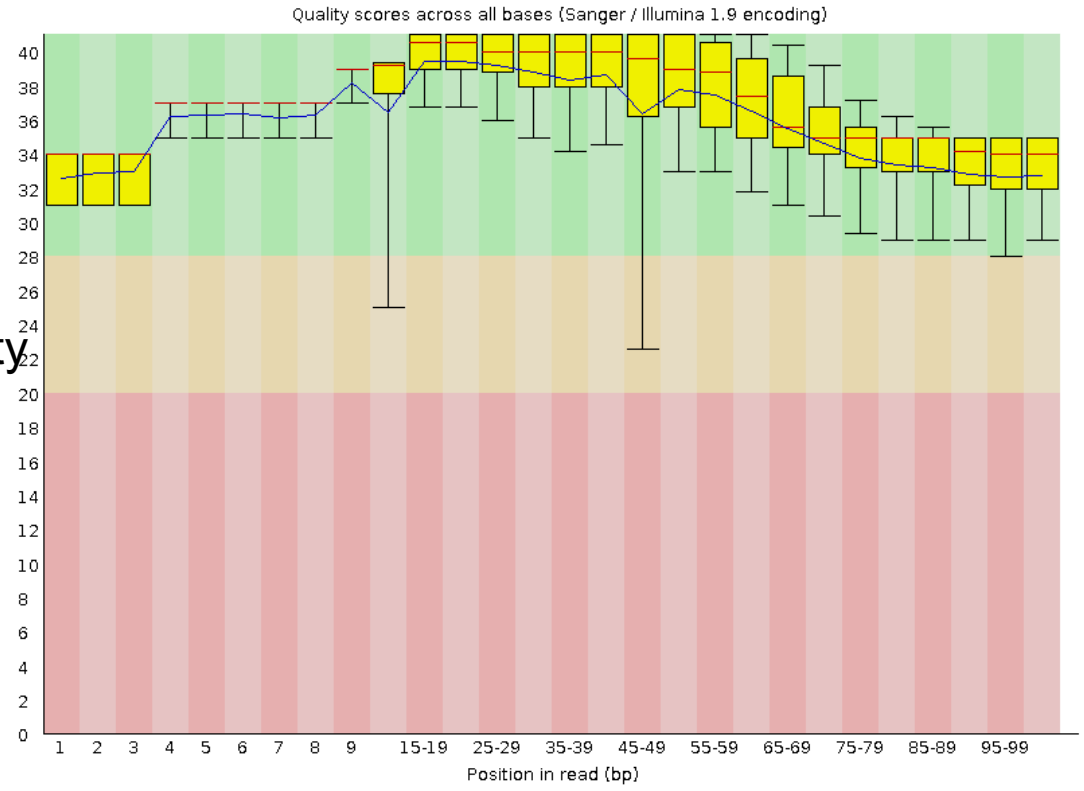
"Although very aggressive quality trimming is common, this study suggests that a more gentle trimming, specifically of those nucleotides whose Phred score < 2 or < 5, is optimal for most studies across a wide variety of metrics."

MacManes MD (2013)
On the optimal trimming of high-throughput mRNAseq data doi: 10.1101/000422

http://genomebio.org/is-trimming-is-beneficial-in-rna-seq/

My take: only trim data when you have to

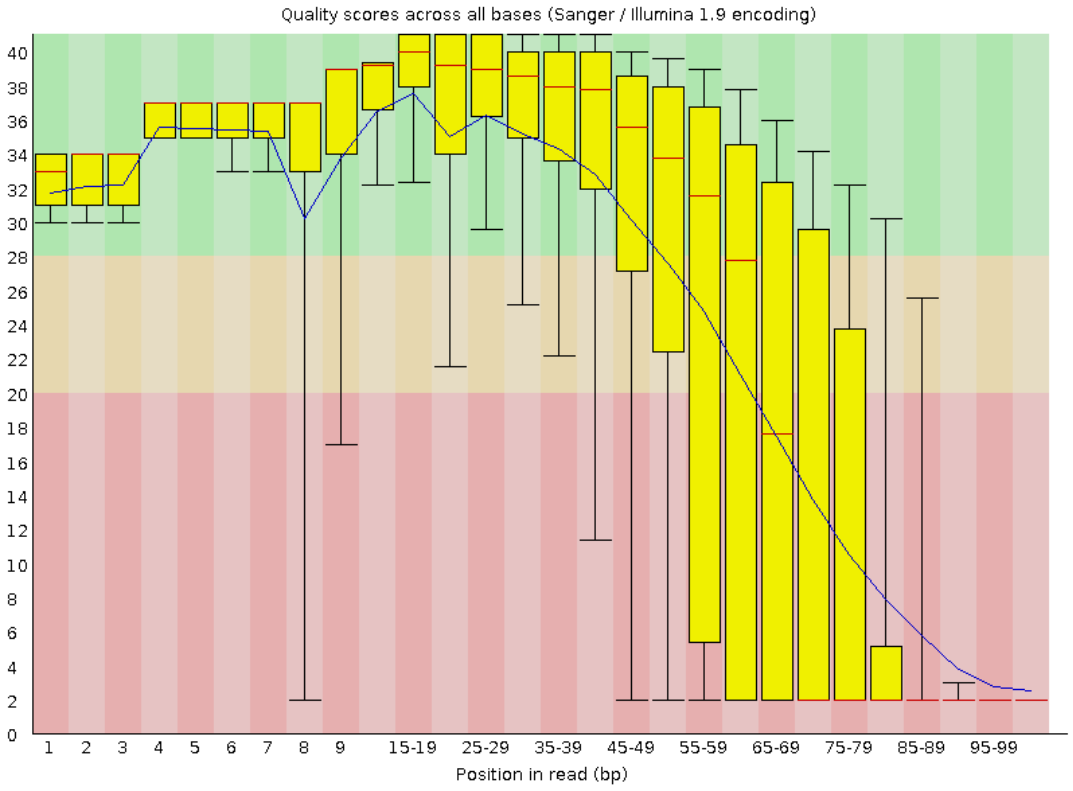
Good/Trimmed data



Quality

Base pair position of the read

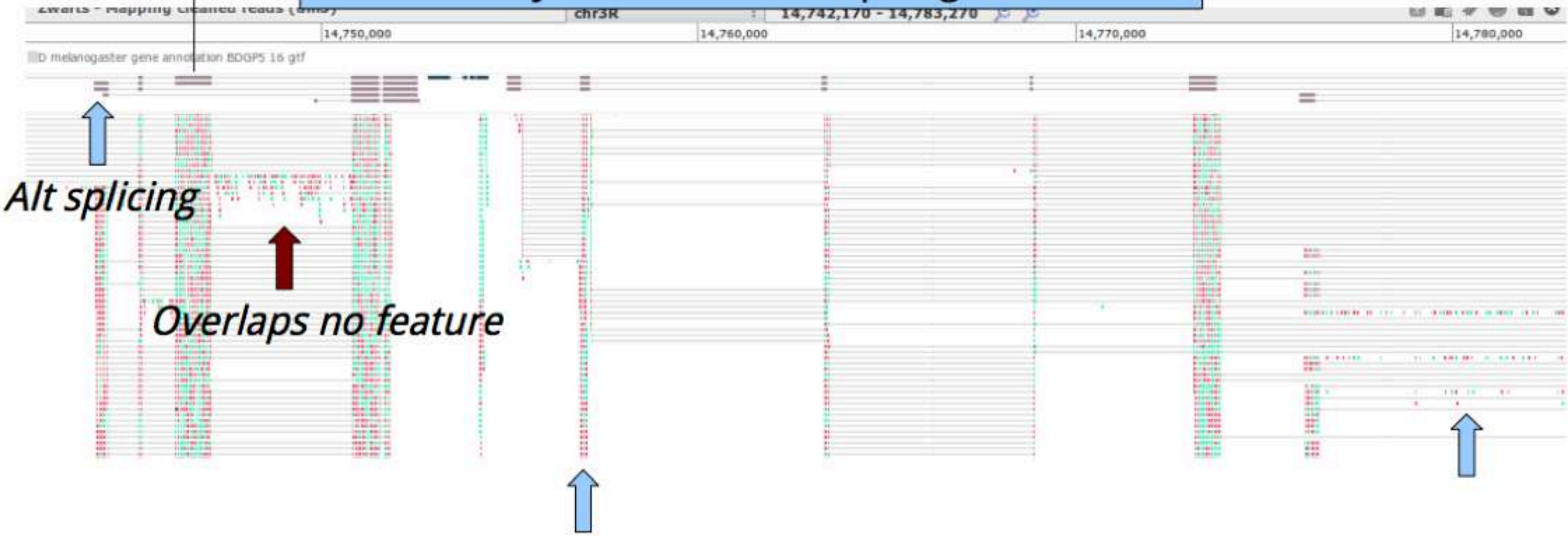
Poor/Raw data



http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

3.1 Once you have mappings, you can start counting

'Exons' are the type of *features* used here.
They are summarized per 'gene'



Concept:

GeneA = exon 1 + exon 2 + exon 3 + exon 4 = 215 reads

GeneB = exon 1 + exon 2 + exon 3 = 180 reads

Featurecount (much faster!)

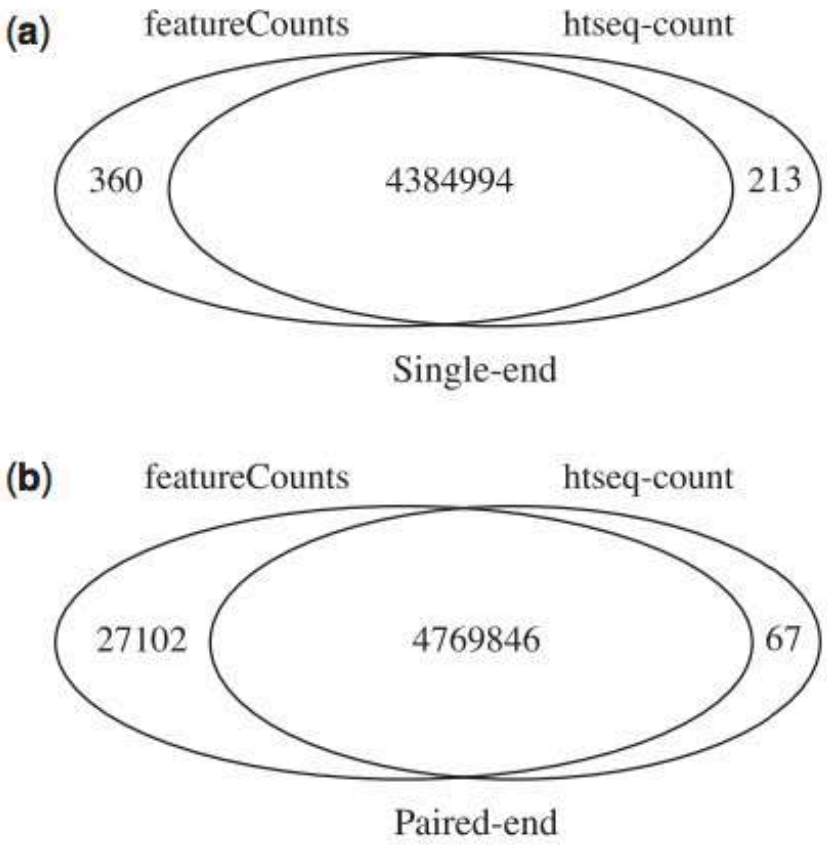


Table 3. Performance with RNA-seq reads simulated from an annotated assembly of the Budgerigar genome

Methods	Number of reads	Time (mins)	Memory (MB)
featureCounts	7 924 065	0.6	15
summarizeOverlaps (whole genome at once)	7 924 065	12.6	2400
summarizeOverlaps (by scaffold)	7 924 065	53.3	262
htseq-count	7 912 439	12.1	78

Note: The annotation includes 16 204 genes located on 2850 scaffolds. *featureCounts* is fastest and uses least memory. Table gives the total number of reads counted, time taken and peak memory used. *htseq-count* was run in 'union' mode.

Liao et al (2014)

Some QC is needed

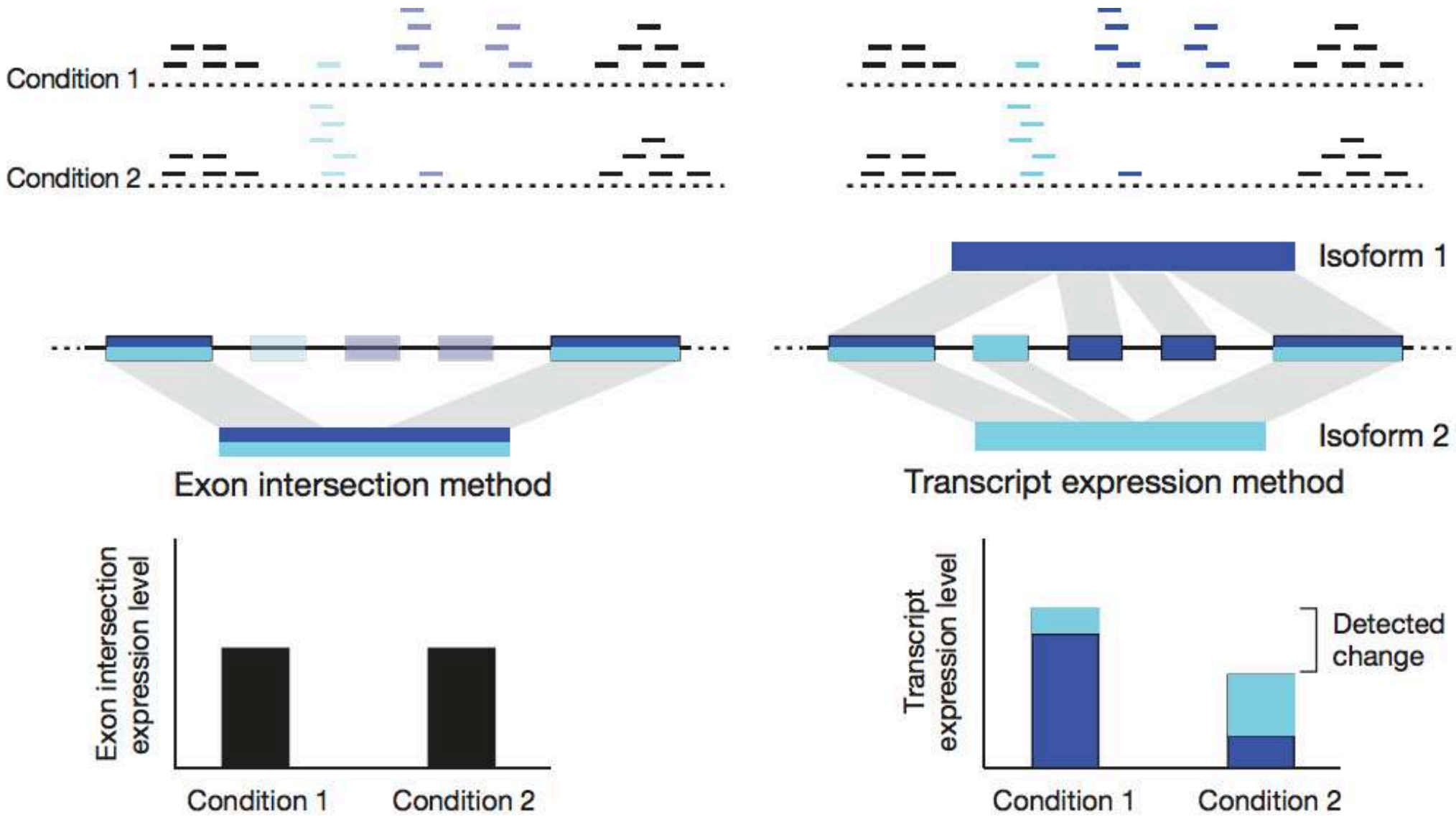
- Which genes are highly counted?
- Any samples with a lot of missing/no count genes?
- Is there any biases on sequencing?
- Anything that may affect sample counts (like batch effect?)

Ambiguity in counting

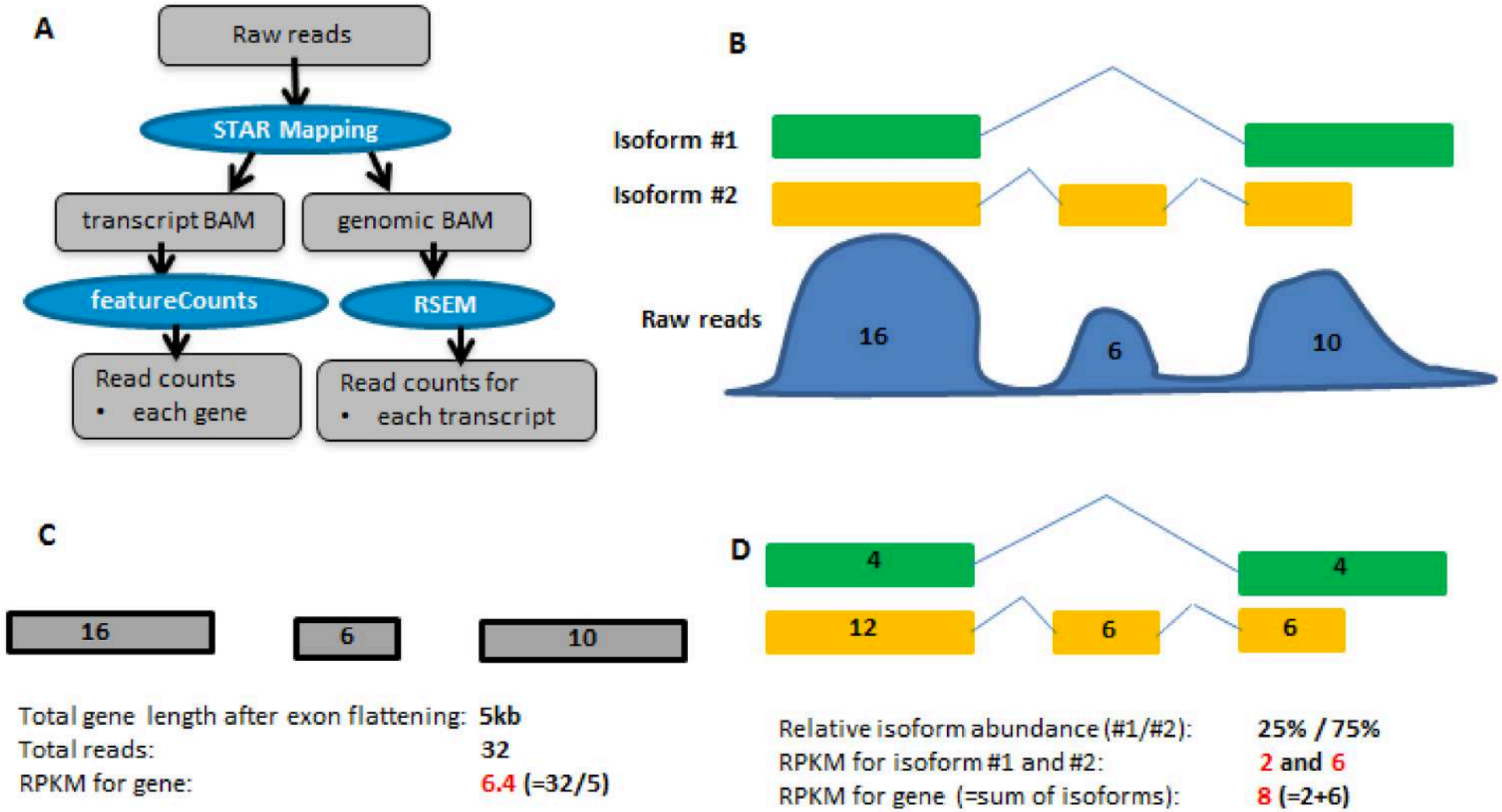
We focus on the **gene level**: merge all counts over different **isoforms** into one, taking into account:

- Reads that do **not overlap** a feature, but appear in introns. Take into account?
- Reads that align to **more than one feature** (exon or transcript)**.** Transcripts can be overlapping - perhaps on different strands. (PE, and strandedness can resolve this partially).
- Reads that **partially** overlap a feature, not following known annotations.

Transcript counting could be more robust in detecting changes



Outstanding problems in counting with exon merging model



Zhao *et al* ., PLOS one (2015)

But Differential transcript expression can lead to inflated false positive rate (and more difficult to interpret biologically)

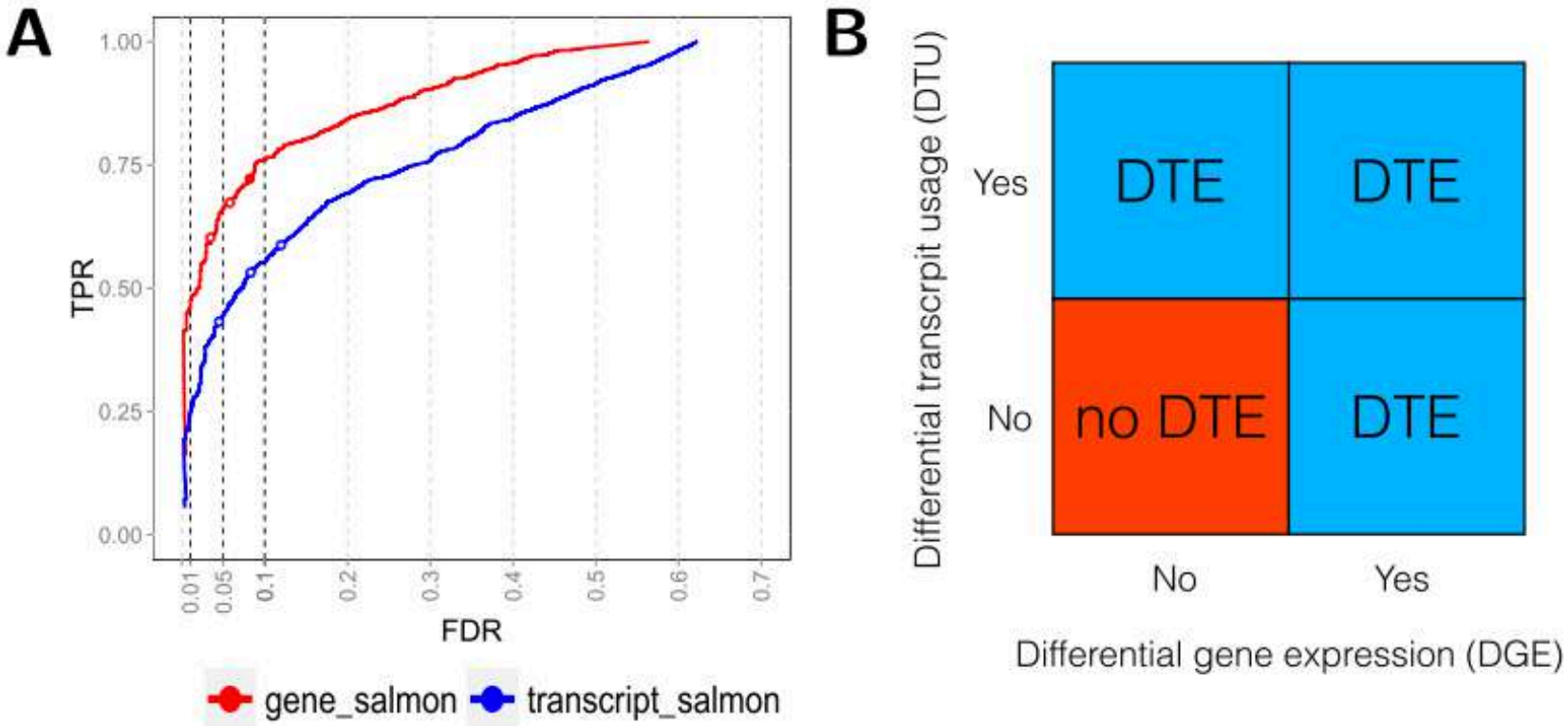
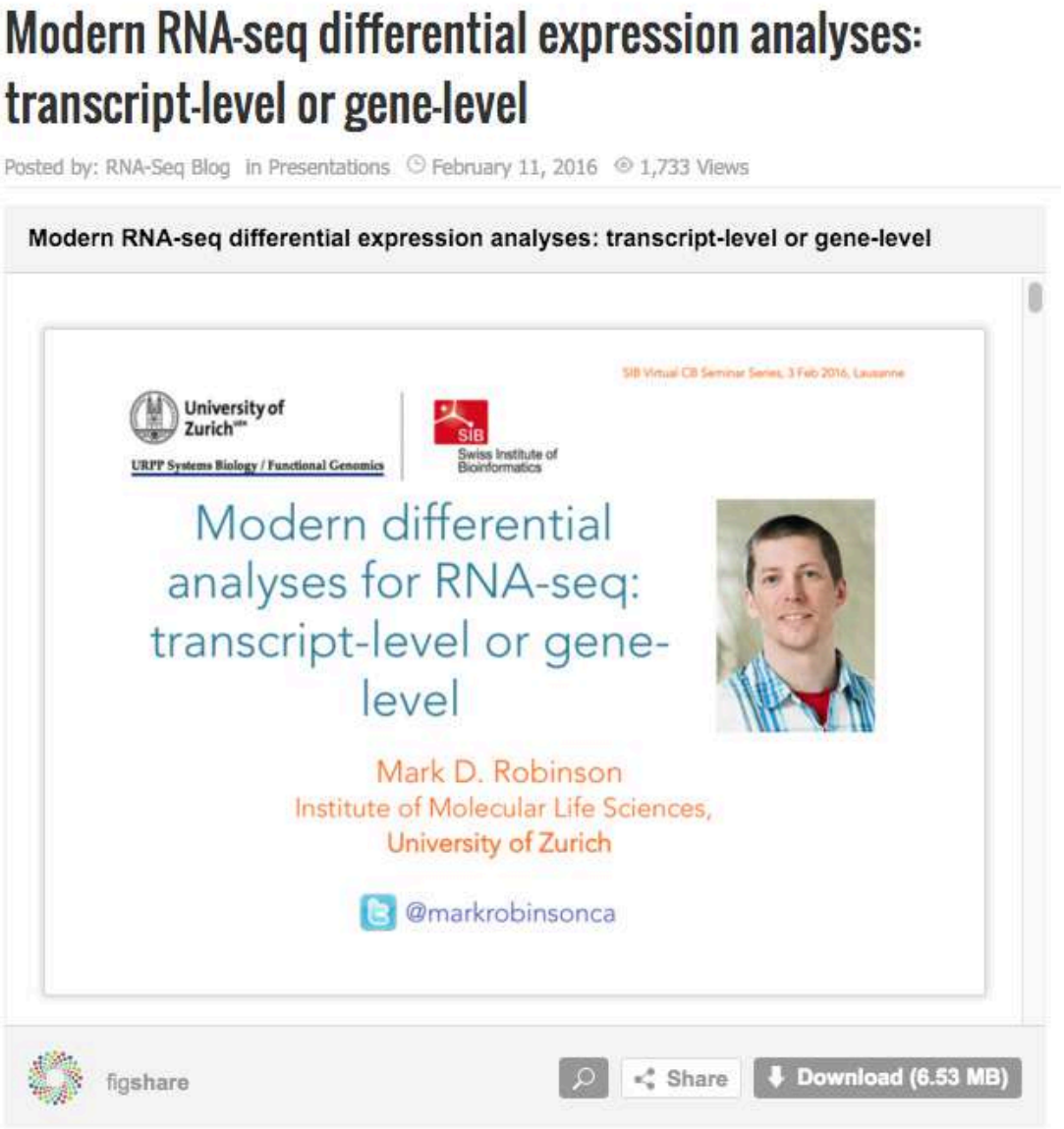


Figure 2 (sim2). A: DTE detection performance on transcript- and gene-level, using *edgeR* applied to transcript-level estimated counts from *Salmon*. The statistical analysis was performed on transcript level and aggregated for each gene using the *perGeneQValue* function from the *DEXSeq* R package; aggregated results show higher detection power. The curves trace out the observed FDR and TPR for each significance cutoff value. The three circles mark the performance at adjusted p-value cutoffs of 0.01, 0.05 and 0.1. **B**: Schematic illustration of different ways in which differential transcript expression (DTE) can arise, in terms of absence or presence of differential gene expression (DGE) and differential transcript usage (DTU).

Soneson et al (2015)

So use isoform or not?



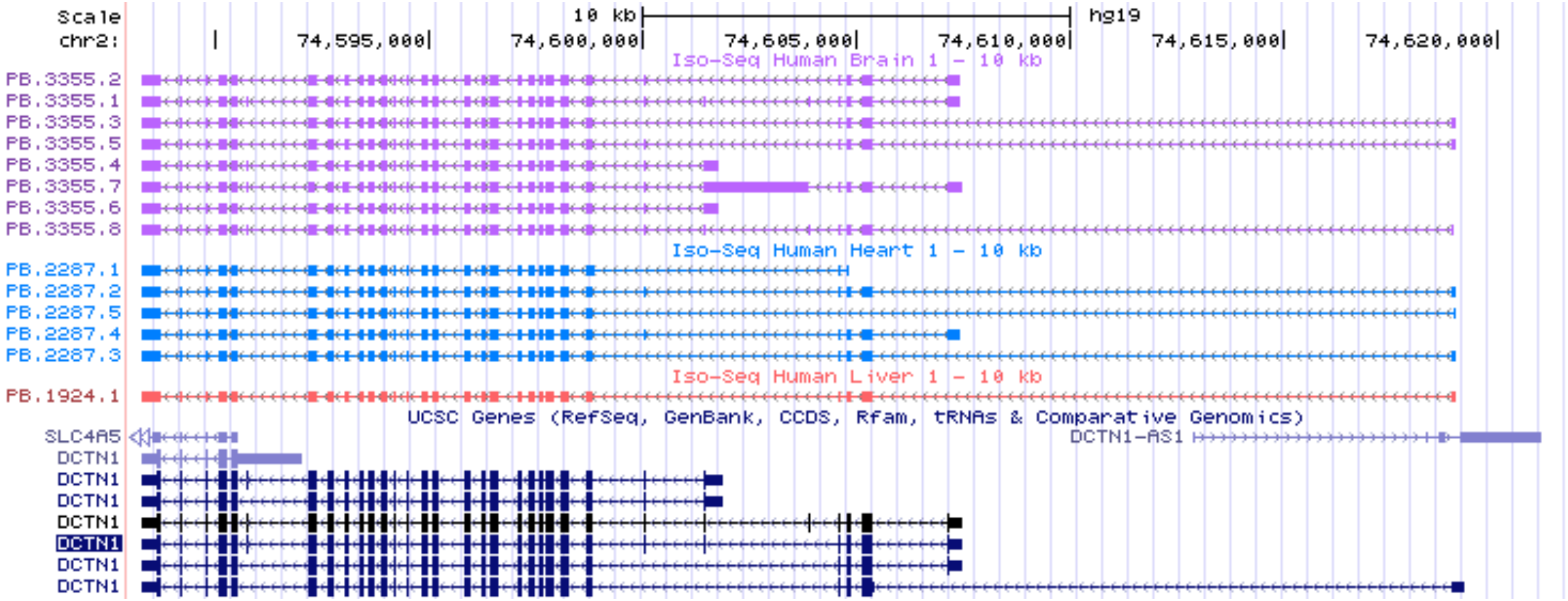
"There is no crisis; the impact of union vs. transcript counting in many datasets is rather small"

"Unless the need dictates, answer the easier questions"

http://www.rna-seqblog.com/modern-rna-seq-differential-expression-analyses-transcript-level-or-gene-level/

We may end up counting full-length transcripts anyway

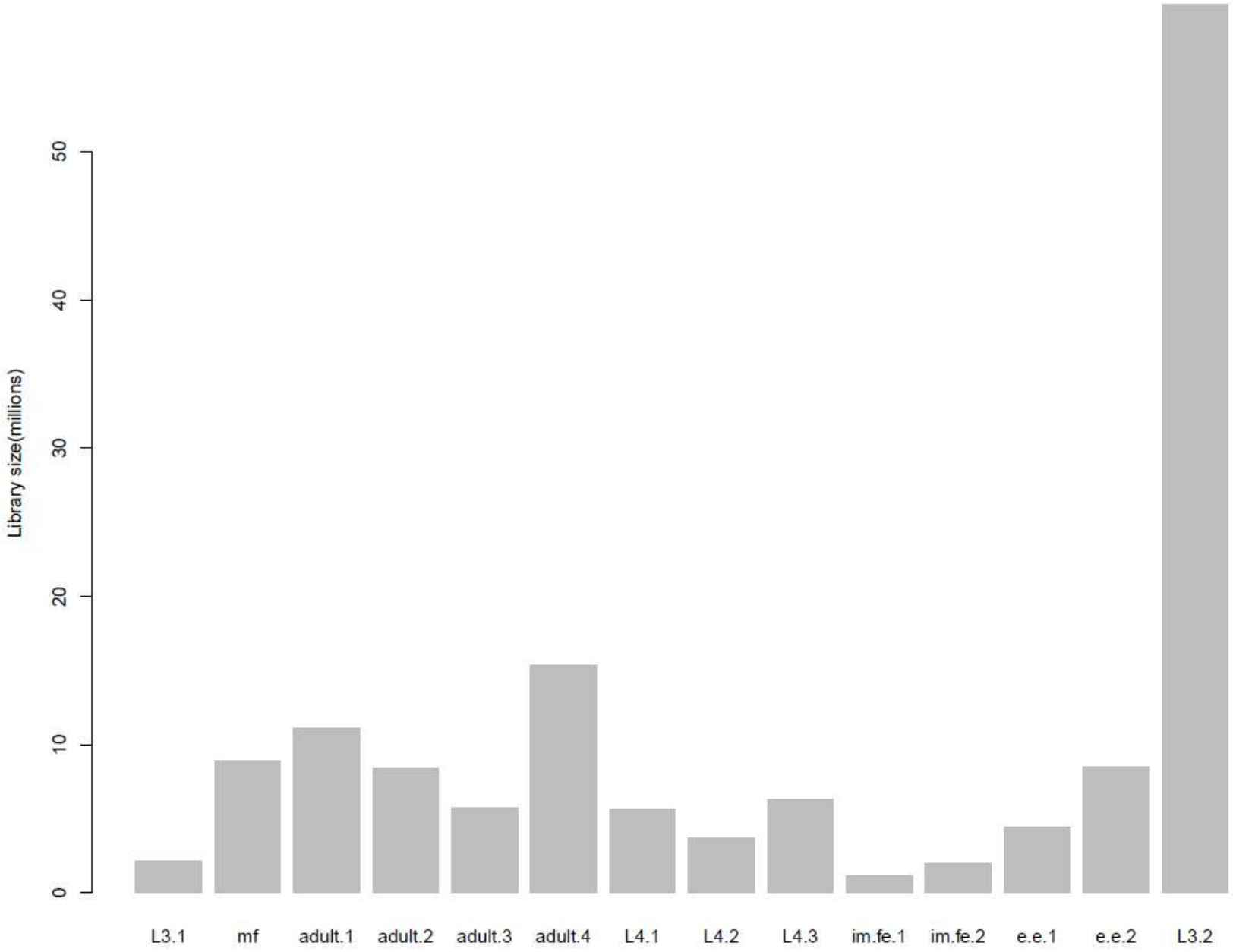
Pacbio IsoSeq



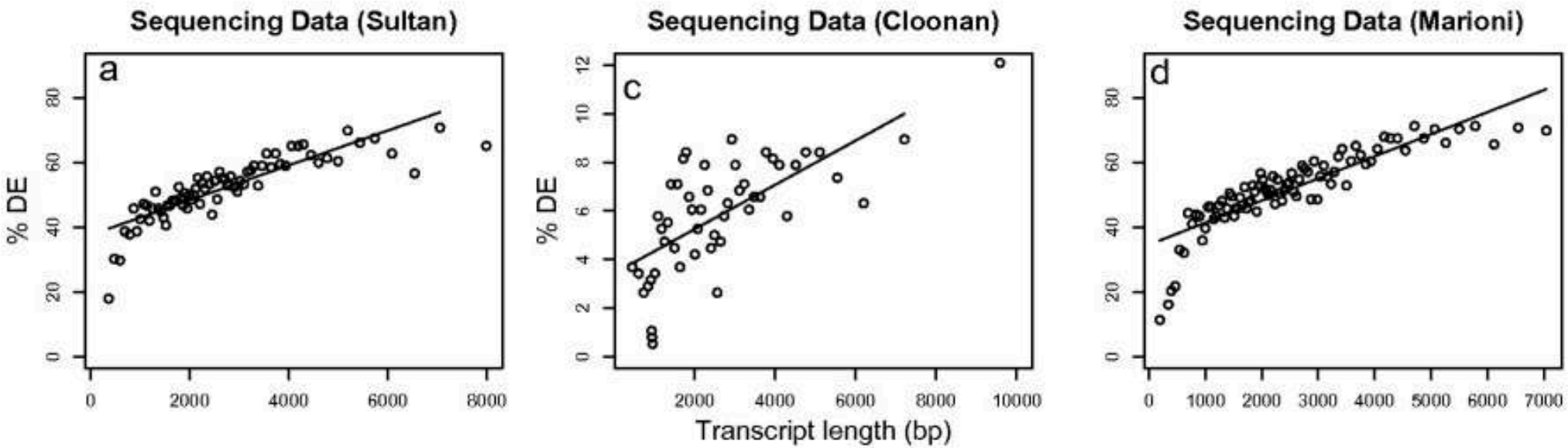
This is the bit we care about!

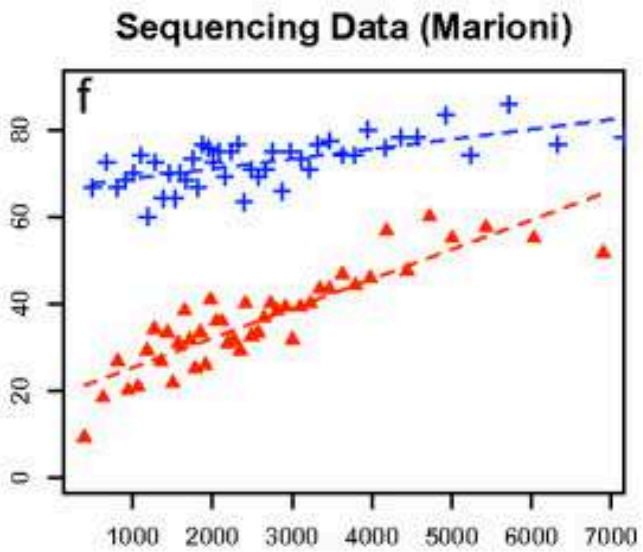
Counts of the gene depends on **expression** ,transcript length ,sequencing depth and simply chance

Higher sequencing depth equals more counts



Counts are proportional to the transcript length x mRNA expression level





33% of highest expressed genes
33% of lowest expressed genes

Oshlack and Wakefield (2009) Biology Direct.

Normalization: different goals

- **Counts per million (CPM)**

- **R/FPKM:** (Mortazavi et al. 2008)
  - Correct for: differences in sequencing depth and transcript length
  - Aiming to: compare a gene across samples and diff genes within sample

- **TMM**: (Robinson and Oshlack 2010)
  - Correct for: differences in transcript pool composition; extreme outliers
  - Aiming to: provide better across-sample comparability

- **TPM**: (Li et al 2010, **Wagner** et al 2012)
  - Correct for: transcript length distribution in RNA pool
  - Aiming to: provide better across-sample comparability

- **Limma voom (logCPM)**: (Lawet al 2013)
  - Aiming to: stabilize variance; remove dependence of variance on the mean

https://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/

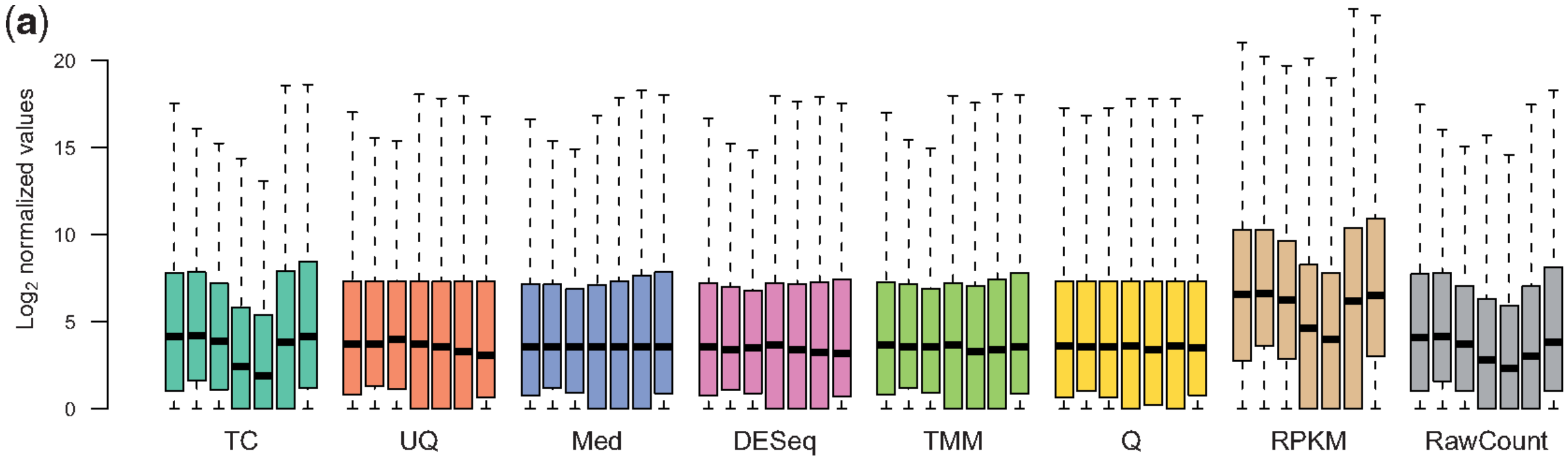
Optimal Scaling of Digital Transcriptomes

Gustavo Glusman, Juan Caballero, Max Robinson, Burak Kutlu, Leroy Hood

Published: Nov 06, 2013 • DOI: 10.1371/journal.pone.0077885

RPKM shouldn't be used for between sample comparisons

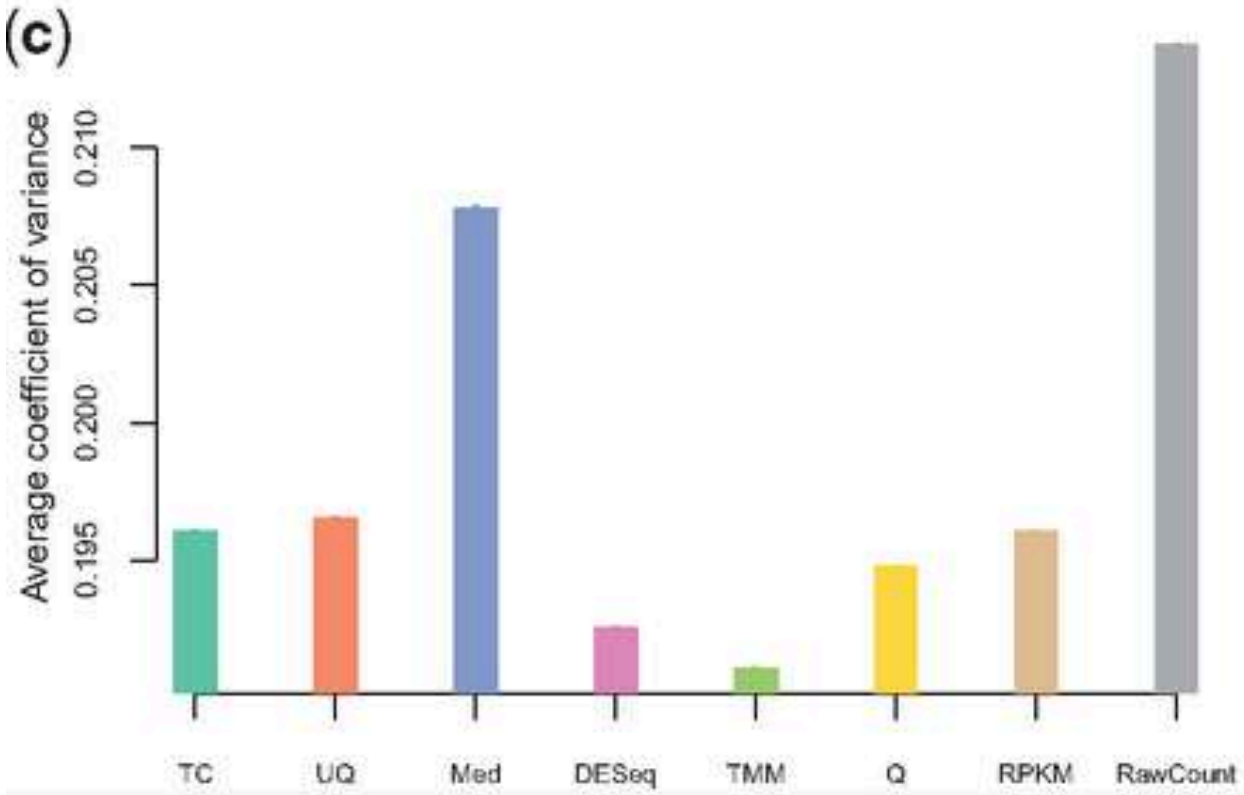
Boxplots of log2(counts + 1) for **seven** replicates in the *M. musculus* data, by normalization method.

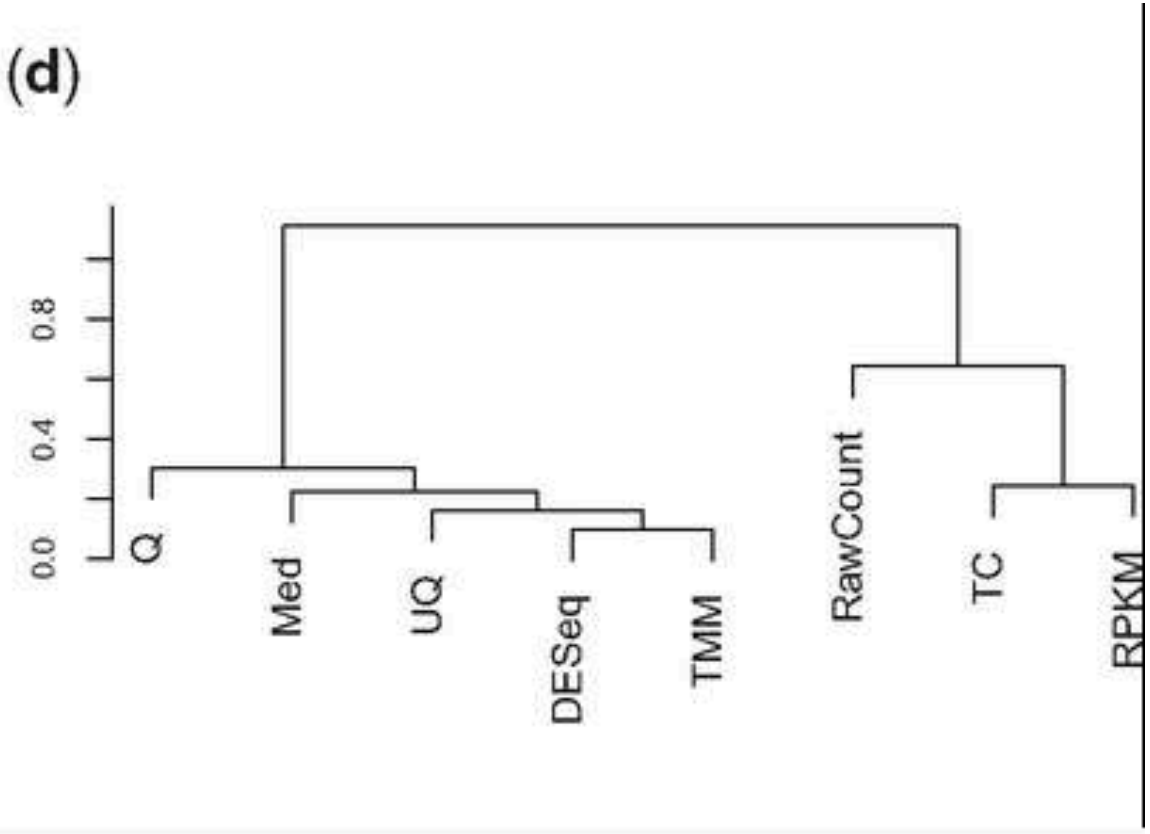


Dillies *et al* (2013) Briefings in Bioinformatics

RPKM shouldn't be used for between sample comparisons

C) Analysis of housekeeping genes for the *H. sapiens* data. (**D**) Consensus dendrogram of differential analysis results





Dillies *et al* (2013) Briefings in Bioinformatics

Summary of comparison results for the seven normalization methods under consideration

Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
TC	−	+	+	−	−
UQ	++	++	+	++	−
Med	++	++	−	++	−
DESeq	++	++	++	++	++
TMM	++	++	++	++	++
Q	++	−	+	++	−
RPKM	−	+	+	−	−

A '−' indicates that the method provided unsatisfactory results for the given criterion, while a '+' and '++' indicate satisfactory and very satisfactory results for the given criterion.

Dillies *et al* (2013) Briefings in Bioinformatics

Between sample comparisons

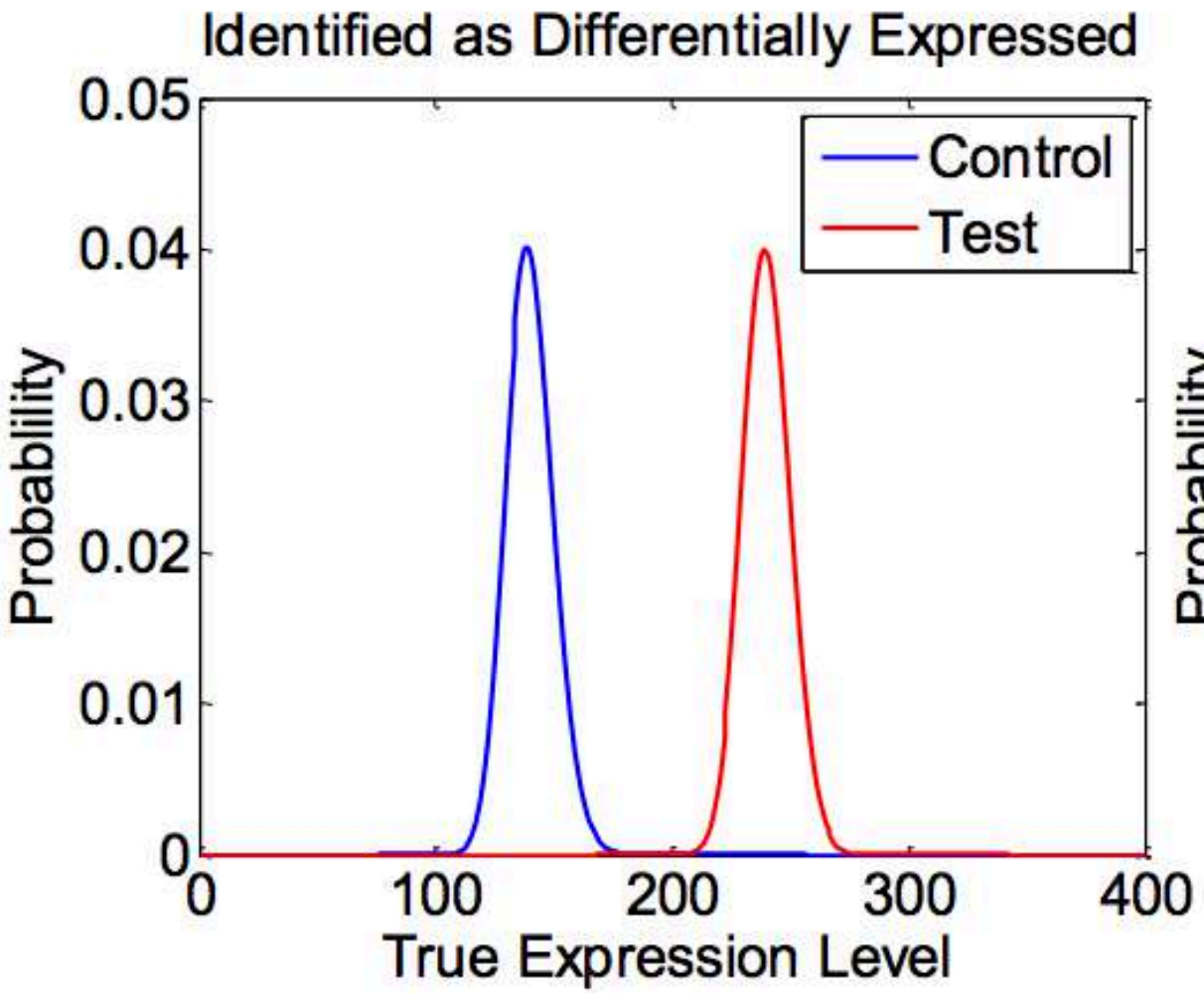
- Differentially expressed genes = counts of genes change between conditions **more systematically** than expected by chance
- Need **biological and technical replicates** to detect differential expression

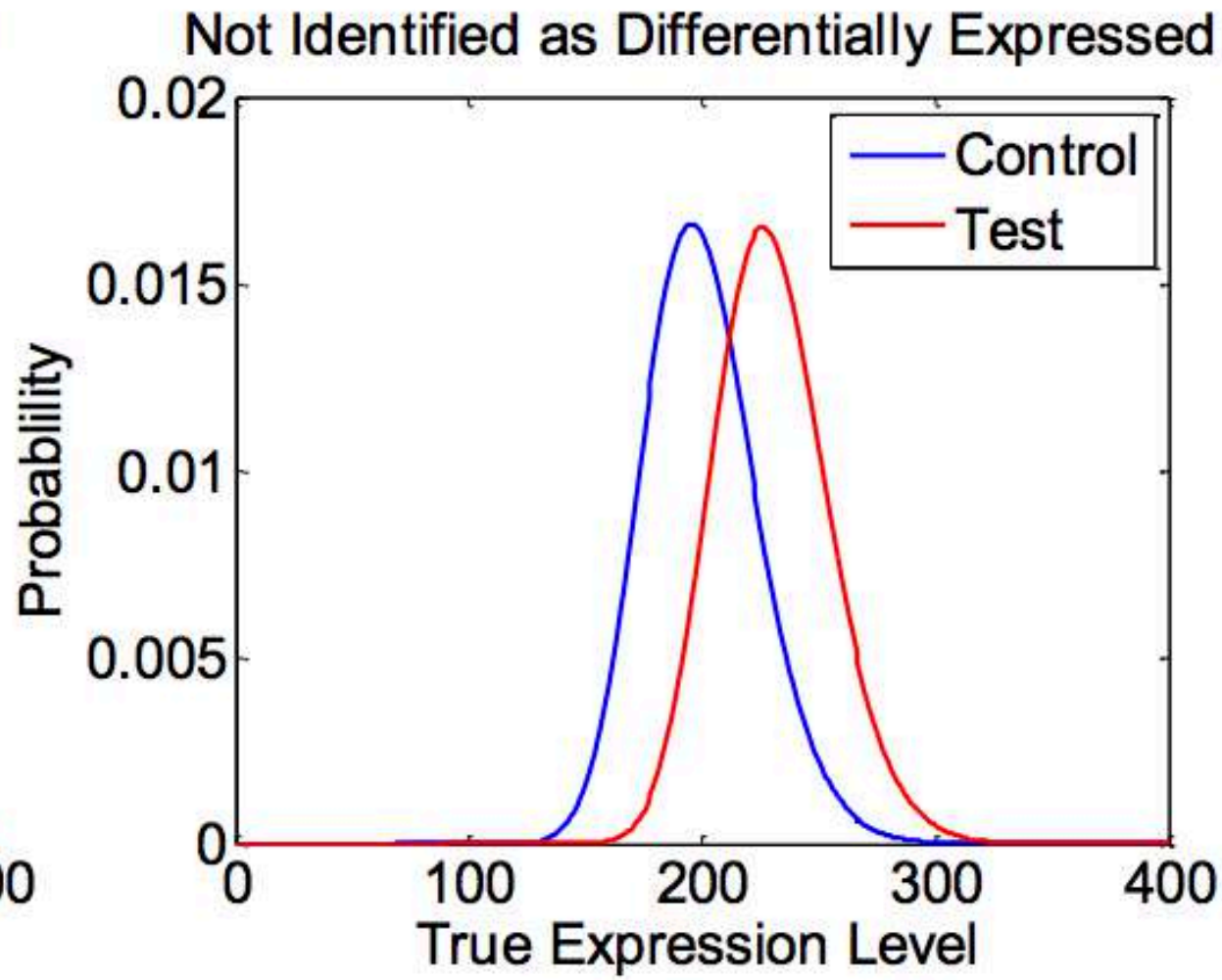
Replicate categories

	Replicate type	Category
Subjects	Colonies	Biological
	Strains	Biological
	Cohoused groups	Biological
	Gender	Biological
	Individuals	Biological
Sample preparation	Organs from sacrificed animals	Biological
	Methods for dissociating cells from tissue	Technical
	Dissociation runs from given tissue sample	Technical
	Individual cells	Biological
	RNA-seq library construction	Technical
Sequencing	Runs from the library of a given cell	Technical
	Reads from different transcript molecules	Variable
	Reads with unique molecular identifier from a given transcript molecule	Technical

http://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf

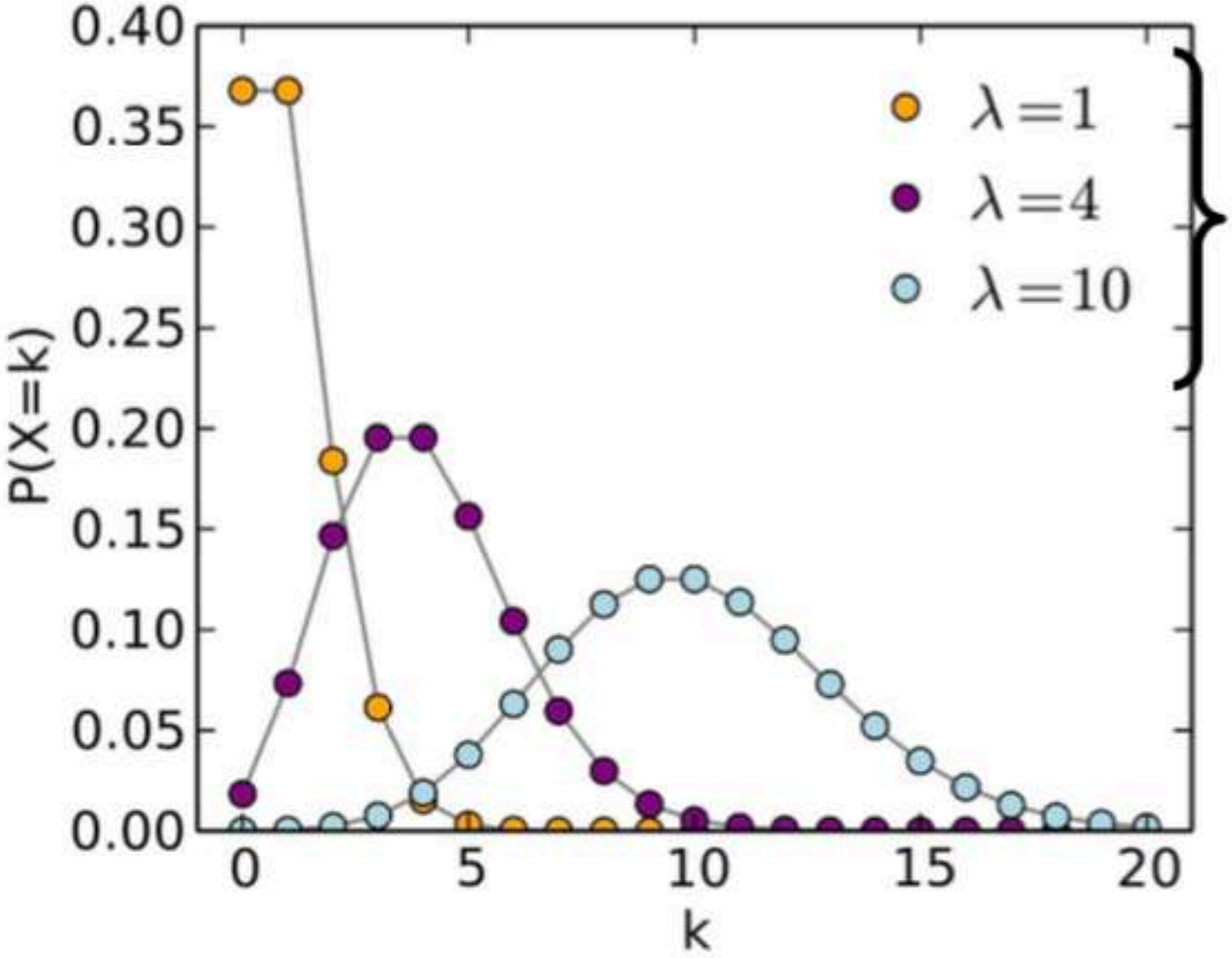
Fitting a distribution for every gene for **DE**





Busby *et al.,* Bioinformatics (2013)

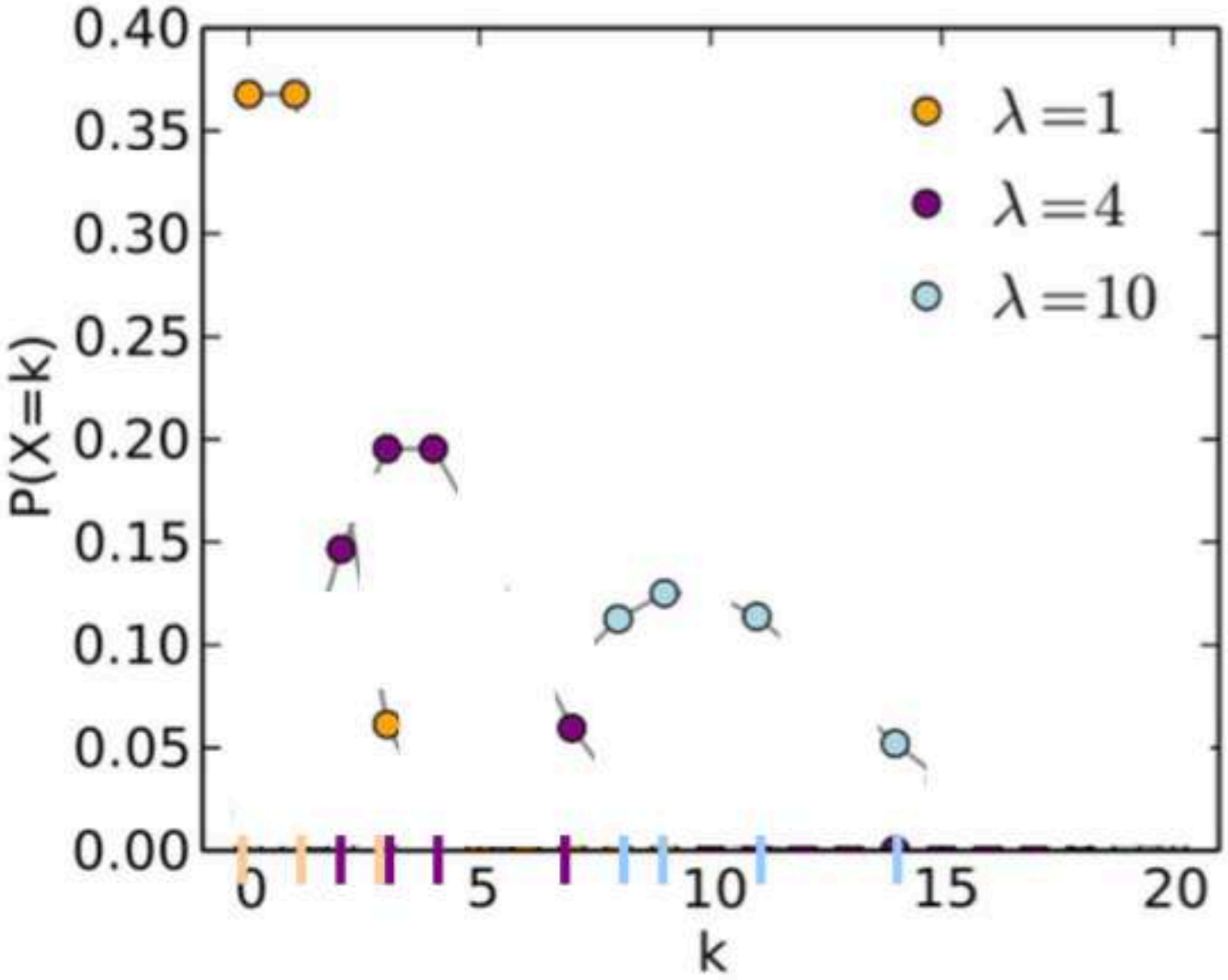
The counts of technical replicates follow a **poisson** distribution (Marioni *et al.,* 2008). So mean = count, variance = count



From Wikipedia. Can be 3 different genes, each with their own poisson distribution. Lambda is the mean of the gene's distribution, with a certain number of reads.

Y=axis: chance to pick that number of reads.

Four technical replicates

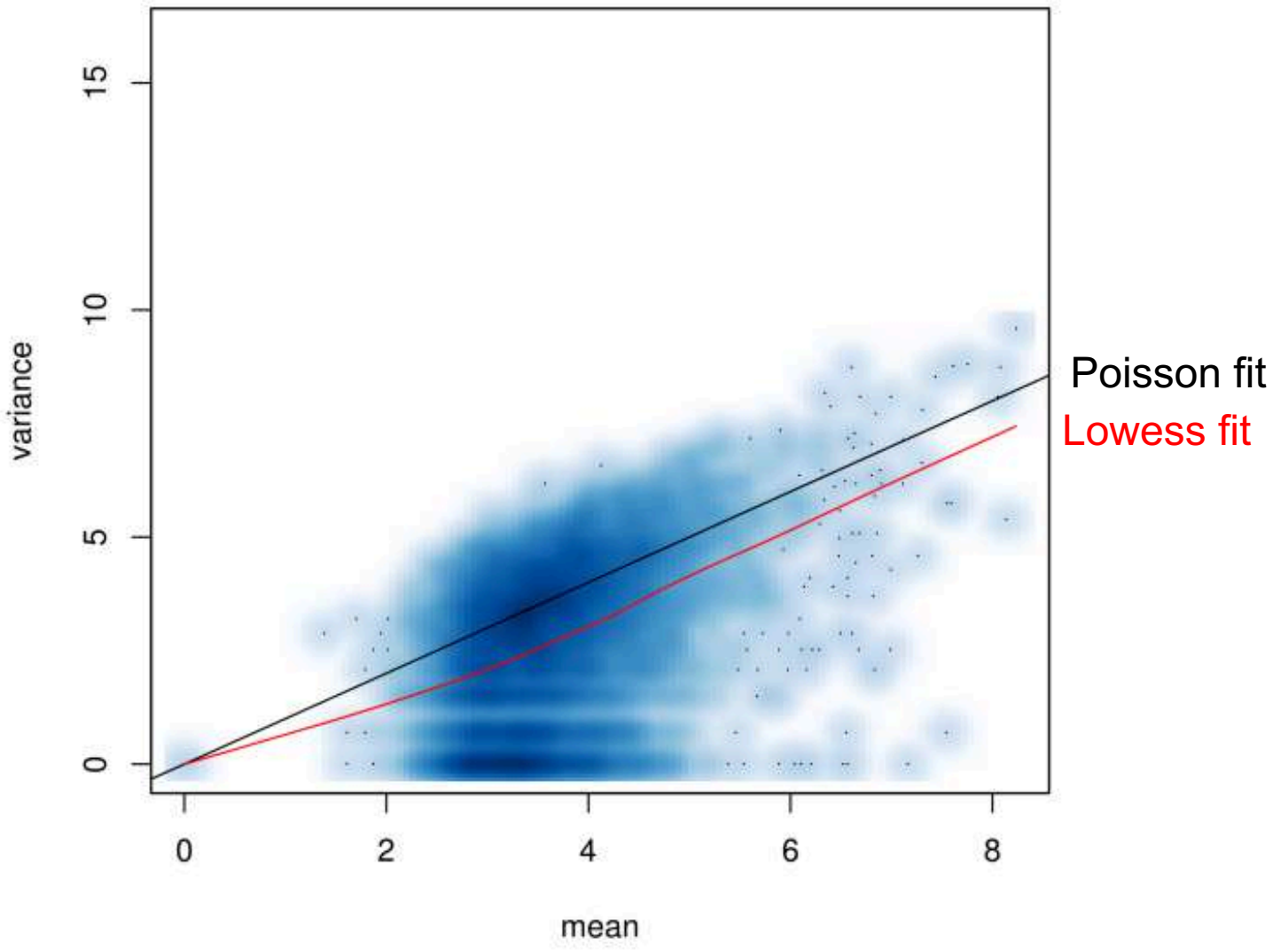


GeneA **0, 0, 1, 3**

GeneB **2, 3, 4, 7**

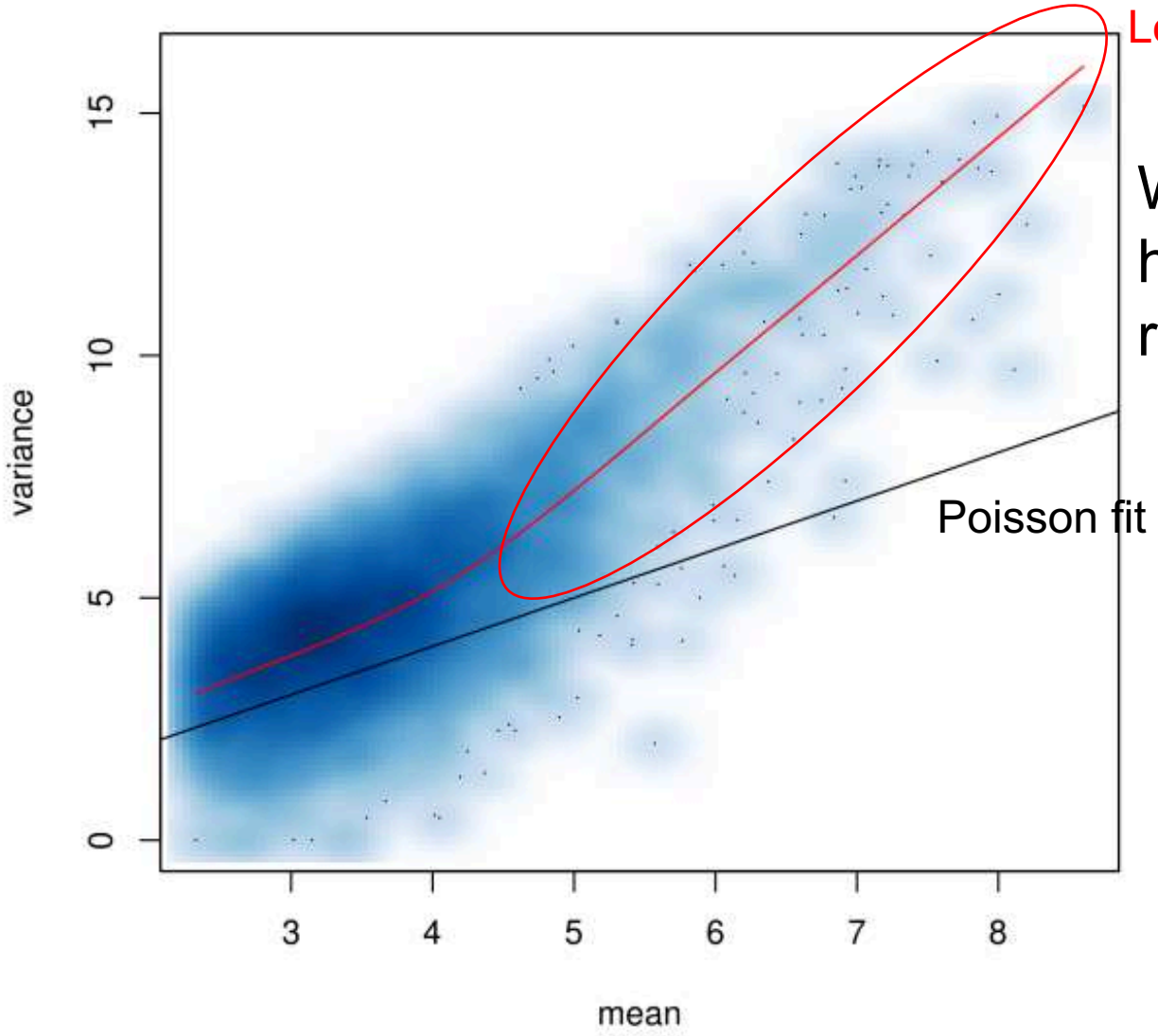
GeneC **8, 9, 11, 14**

Poisson model seems good fit in technical replicates



http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.367.1606&rep=rep1&type=pdf

Poisson model seems good fit in technical replicates



We call this **overdispersion**: the variance is higher for higher counts between biological replicates

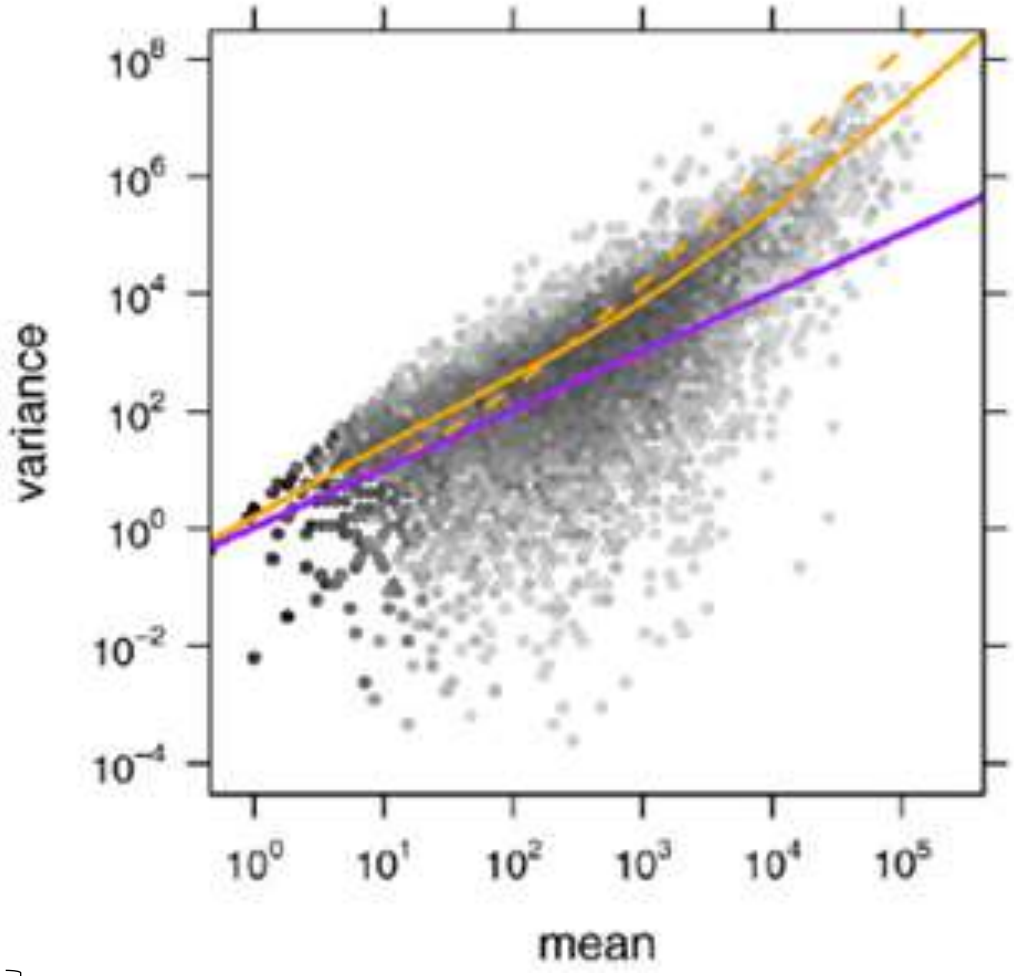
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.367.1606&rep=rep1&type=pdf

Variance depends strongly on the mean

Technical replicate: Poisson
Biological replicate: **Negative binomial**

For **low counts**, the Poisson (technical) variation or the measurement error is dominant.

For **higher counts**, the Poisson variation gets smaller, and another source of variation becomes dominant, the **dispersion** or the **biological variation**. Biological variation does not get smaller with higher counts.



Poisson v = μ Poisson distribution
Poisson + constant CV $v = \mu + \alpha\, \mu 2$ (edgeR)
Poisson + local regression $v = \mu + f(\mu 2)$ (DESeq)

Negative binomial distribution

Anders & Huber, 2010

Lots of Differential Gene Expression methods

Table 1 Methods for calling differentially expressed genes in RNA-seq data analysis. Total citations were based on Google Scholar search result as of 22 September 2015, and normalized by number of years since formal publication. The methods were ranked according to their citations per year.

Method	Total citations	Citations per year	Reference
DESeq*	2,987	597	*Anders & Huber (2010)*
edgeR*	2,260	452	*Robinson, McCarthy & Smyth (2010)*
Cuffdiff2	517	258	*Trapnell et al. (2013)*
DESeq2*	209	209	*Love, Huber & Anders (2014)*
voom*	143	143	*Law et al. (2014)*
DEGseq	592	118	*Wang et al. (2010)*
NOISeq*,a,b	324	81	*Tarazona et al. (2011)*
baySeq	310	62	*Hardcastle & Kelly (2010)*
SAMSeq[b]	114	57	*Li & Tibshirani (2013)*
EBSeq	107	53	*Leng et al. (2013)*
PoissonSeq	99	33	*Li et al. (2012)*
BitSeq	70	23	*Glaus, Honkela & Rattray (2012)*
DSS	46	23	*Wu, Wang & Wu (2013)*
TSPM	70	17	*Auer & Doerge (2011)*
GPseq	86	17	*Srivastava & Chen (2010)*
NBPSeq	65	16	*Di et al. (2011)*
QuasiSeq	47	16	*Lund et al. (2012)*
GFOLD*,a	44	15	*Feng et al. (2012)*
ShrinkSeq	30	15	*Van De Wiel et al. (2013)*
NPEBseq[b]	14	7	*Bi & Davuluri (2013)*
ASC*,a	32	6	*Wu et al. (2010)*
BADGE	2	1	*Gu et al. (2014)*

Khang *et al.*, (2015)

Table 7: Comparison of programs for differential gene expression identification (Rapaport et al., 2013; Seyednasrollah et al., 2015; Schurch et al., 2015).

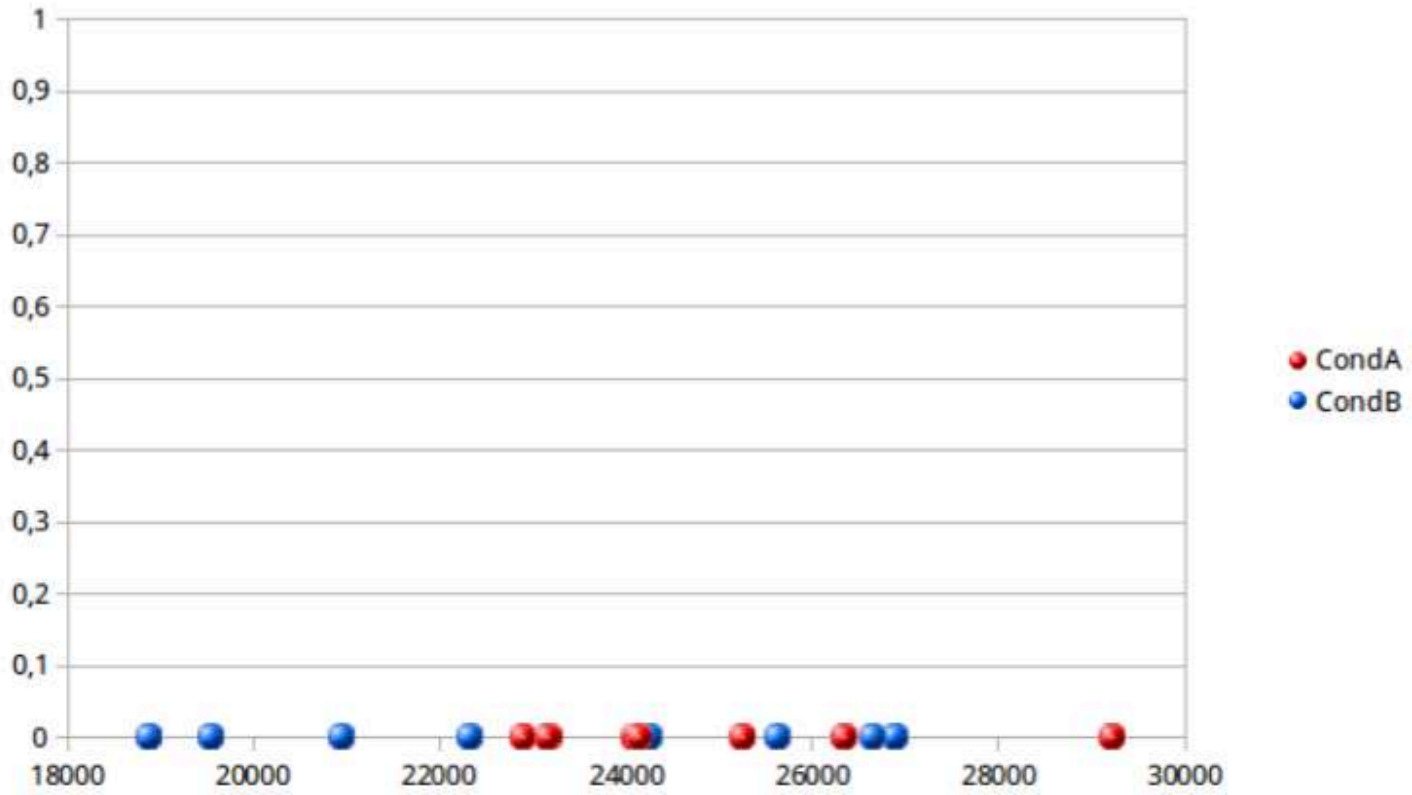
Feature	DESeq2	edgeR	limmaVoom	Cuffdiff
Seq. depth normalization	Sample-wise size factor	Gene-wise trimmed median of means (TMM)	Gene-wise trimmed median of means (TMM)	FPKM-like or DESeq-like
Assumed distribution	Neg. binomial	Neg. binomial	*log*-normal	Neg. binomial
Test for DE	Exact test (Wald)	Exact test for over-dispersed data	Generalized linear model	*t*-test
False positives	Low	Low	Low	High
Detection of differential isoforms	No	No	No	Yes
Support for multi-factored experiments	Yes	Yes	Yes	No
Runtime (3-5 replicates)	Seconds to minutes	Seconds to minutes	Seconds to minutes	Hours

http://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf

Scenario

gene_id CAF0006876

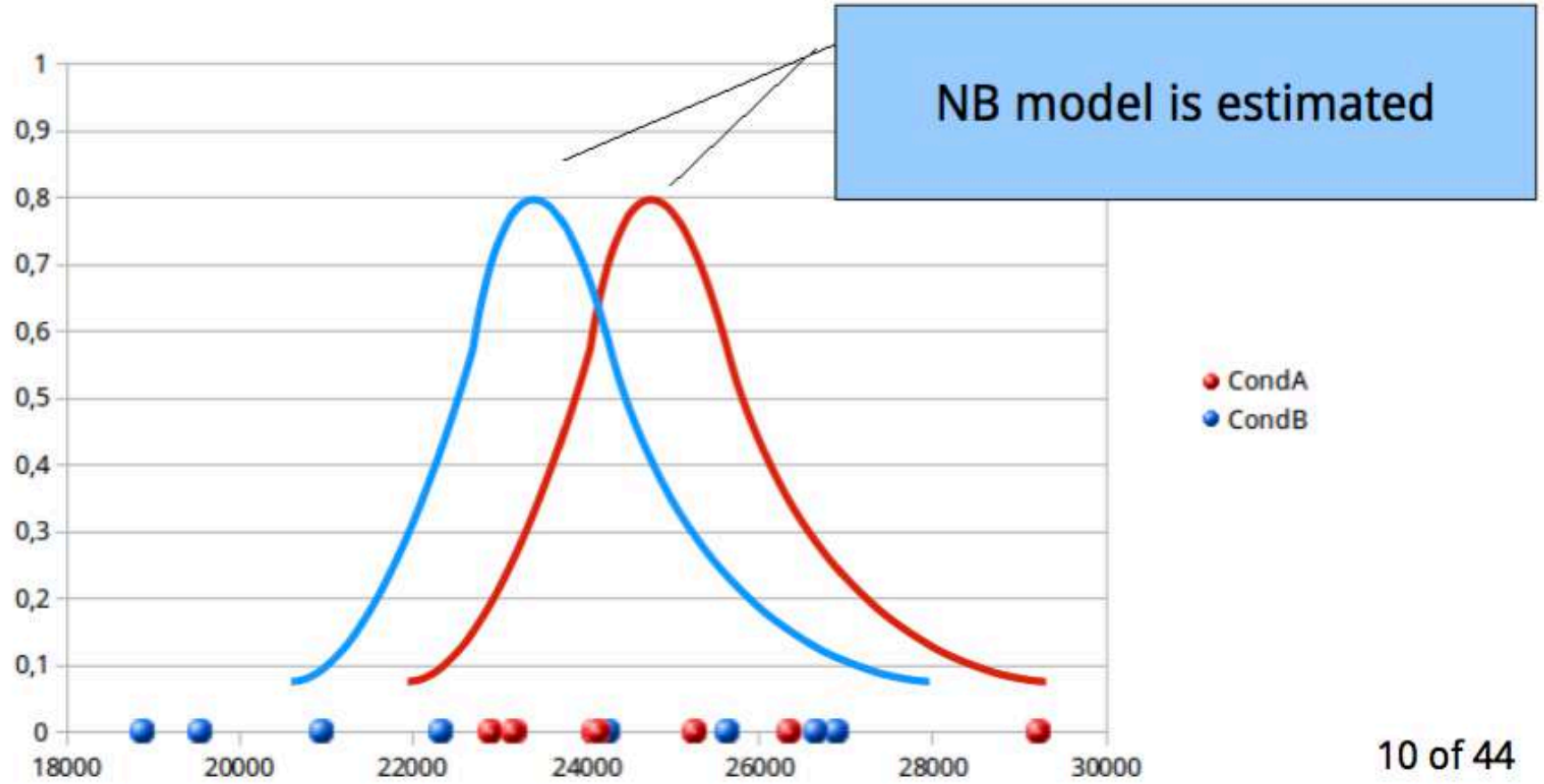
Condition A	sample1	sample2	sample3	sample4	sample5	sample6	sample7	sample8
	23171	22903	29227	24072	23151	26336	25252	24122
Condition B	Sample9	sample10	sample11	sample12	sample13	sample14	sample15	sample16
	19527	26898	18880	24237	26640	22315	20952	25629



Scenario

gene_id CAF0006876

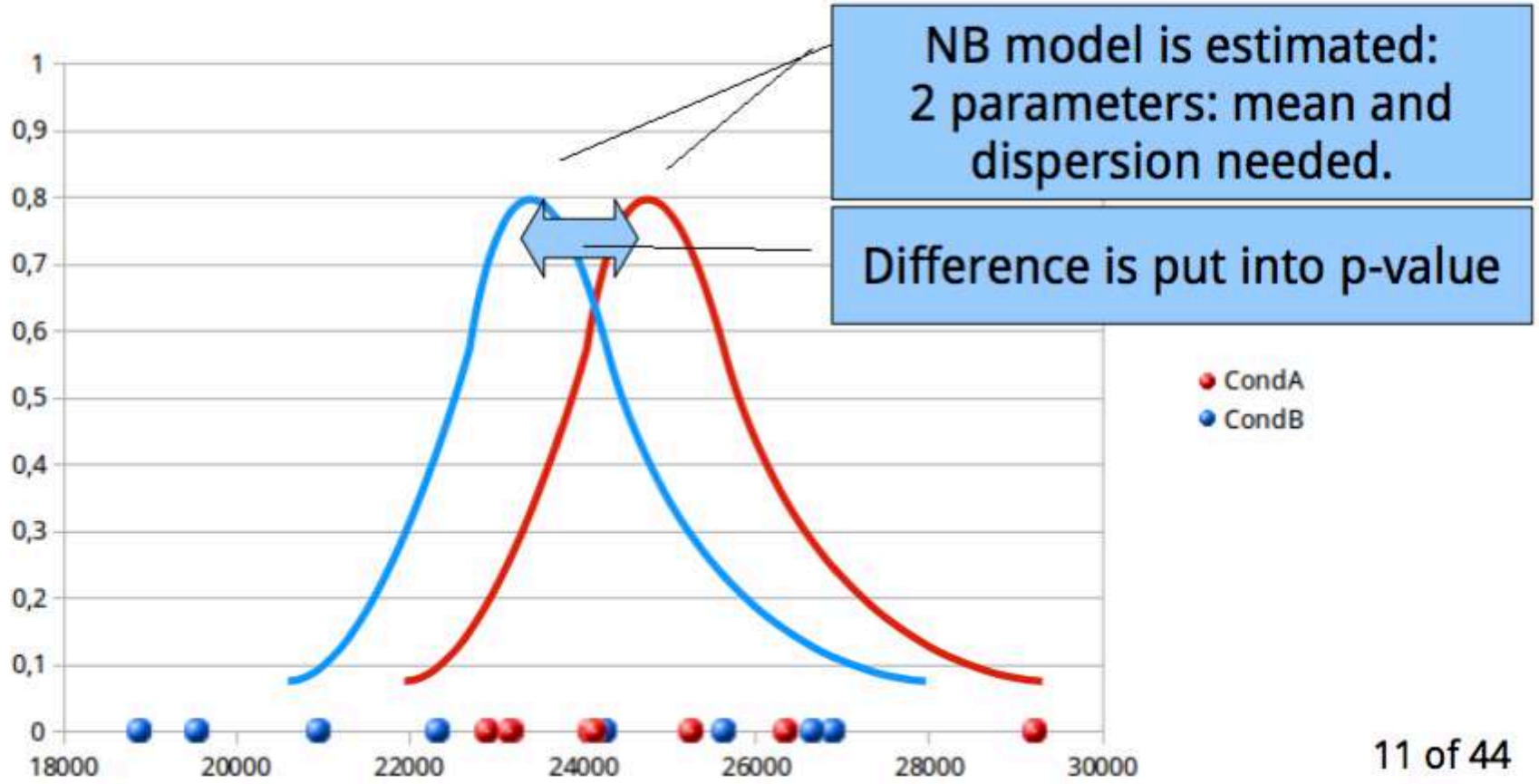
Condition A	sample1	sample2	sample3	sample4	sample5	sample6	sample7	sample8
	23171	22903	29227	24072	23151	26336	25252	24122
Condition B	Sample9	sample10	sample11	sample12	sample13	sample14	sample15	sample16
	19527	26898	18880	24237	26640	22315	20952	25629



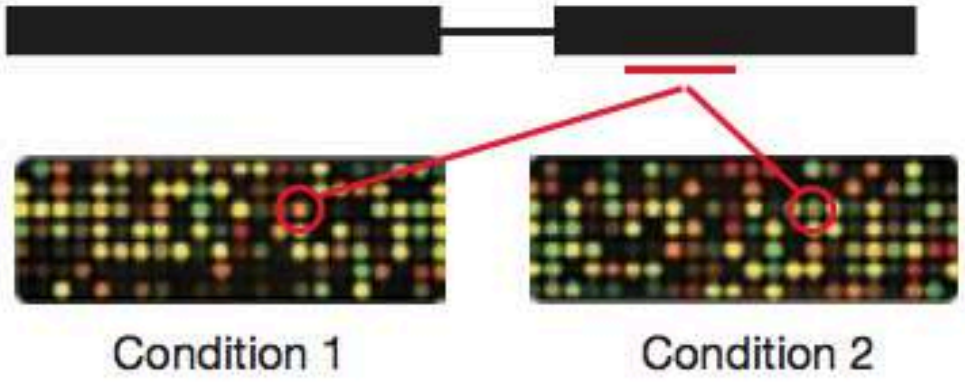
Scenario

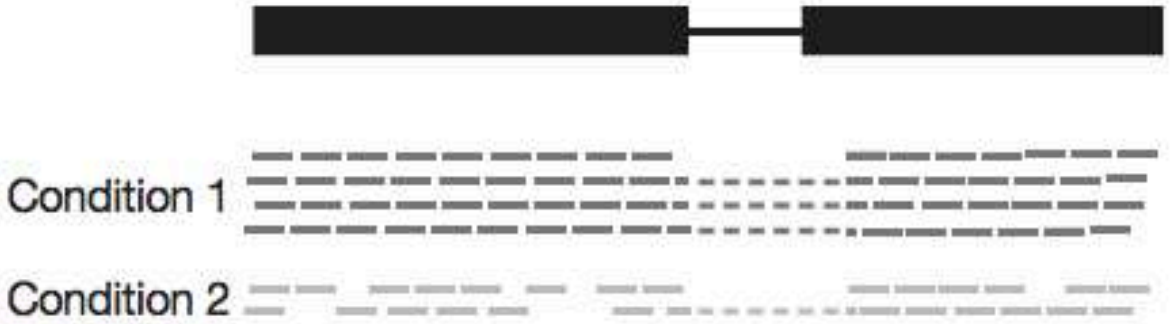
gene_id CAF0006876

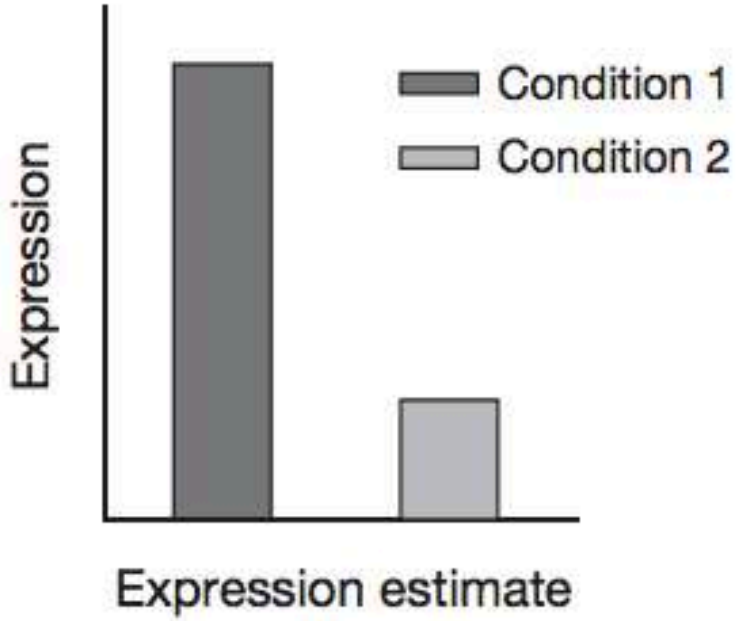
Condition A	sample1 23171	sample2 22903	sample3 29227	sample4 24072	sample5 23151	sample6 26336	sample7 25252	sample8 24122
Condition B	Sample9 19527	sample10 26898	sample11 18880	sample12 24237	sample13 26640	sample14 22315	sample15 20952	sample16 25629

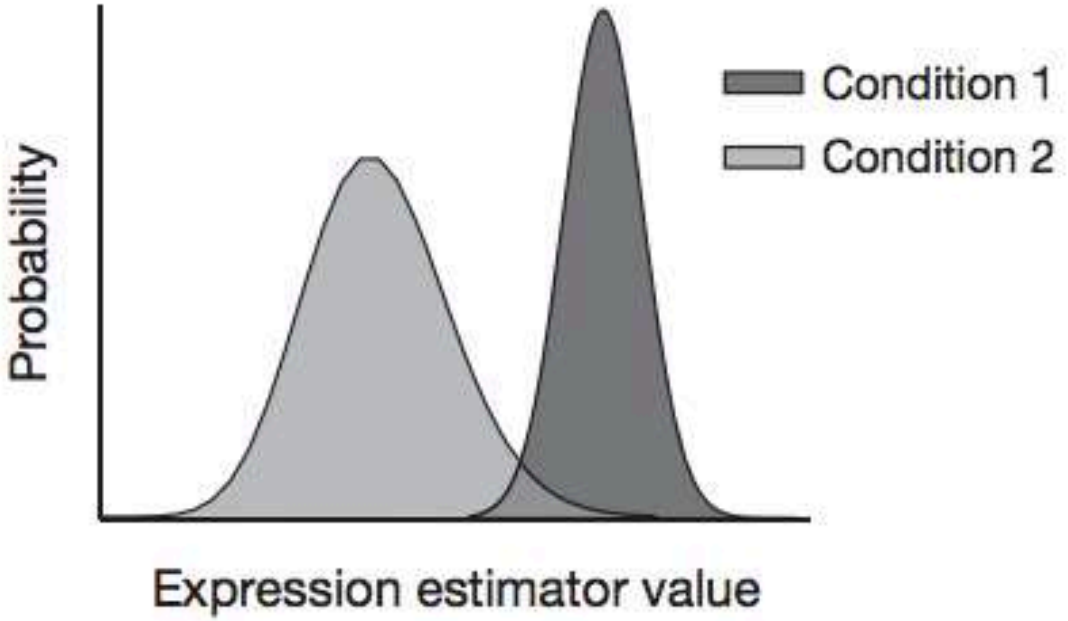


RNAseq vs Microarray

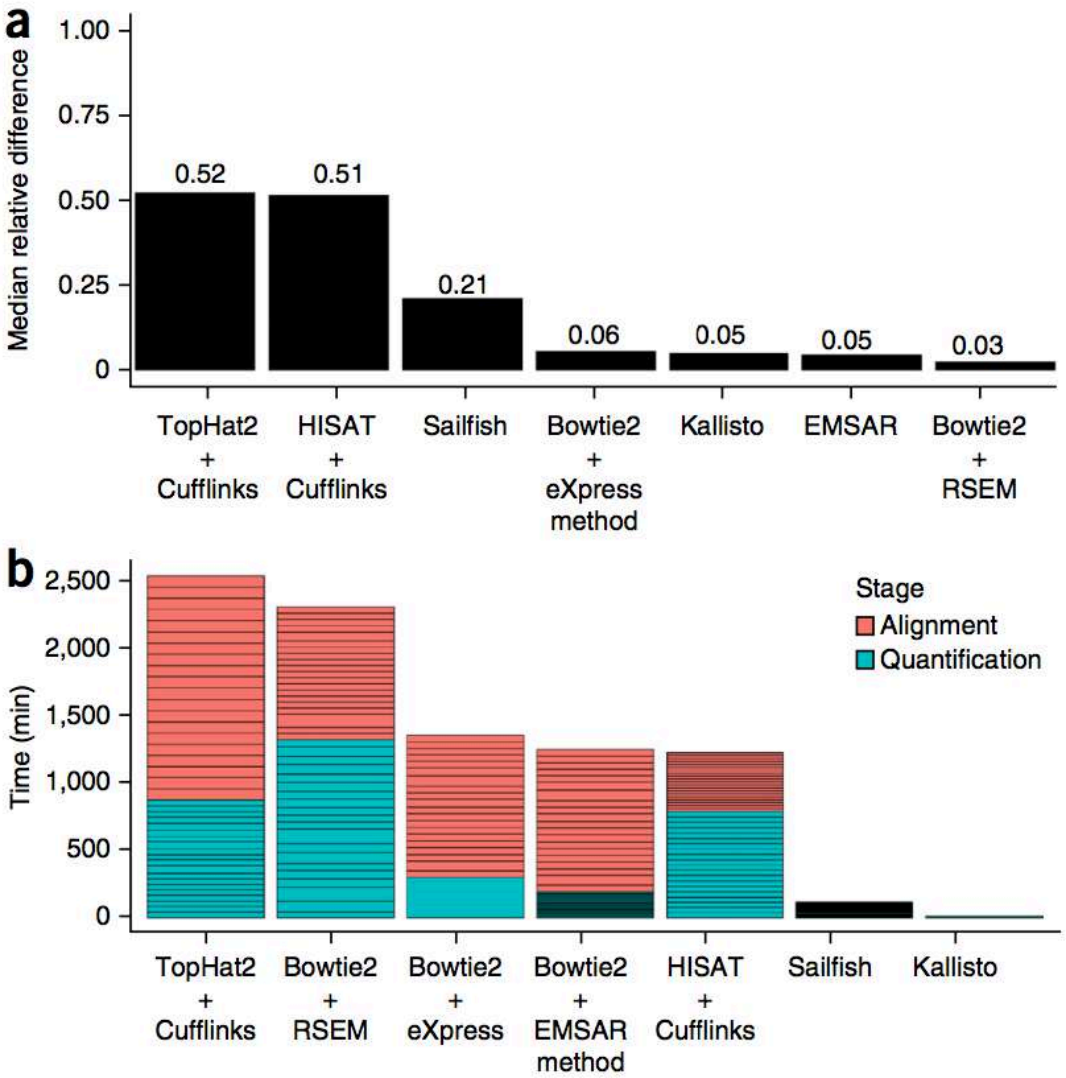








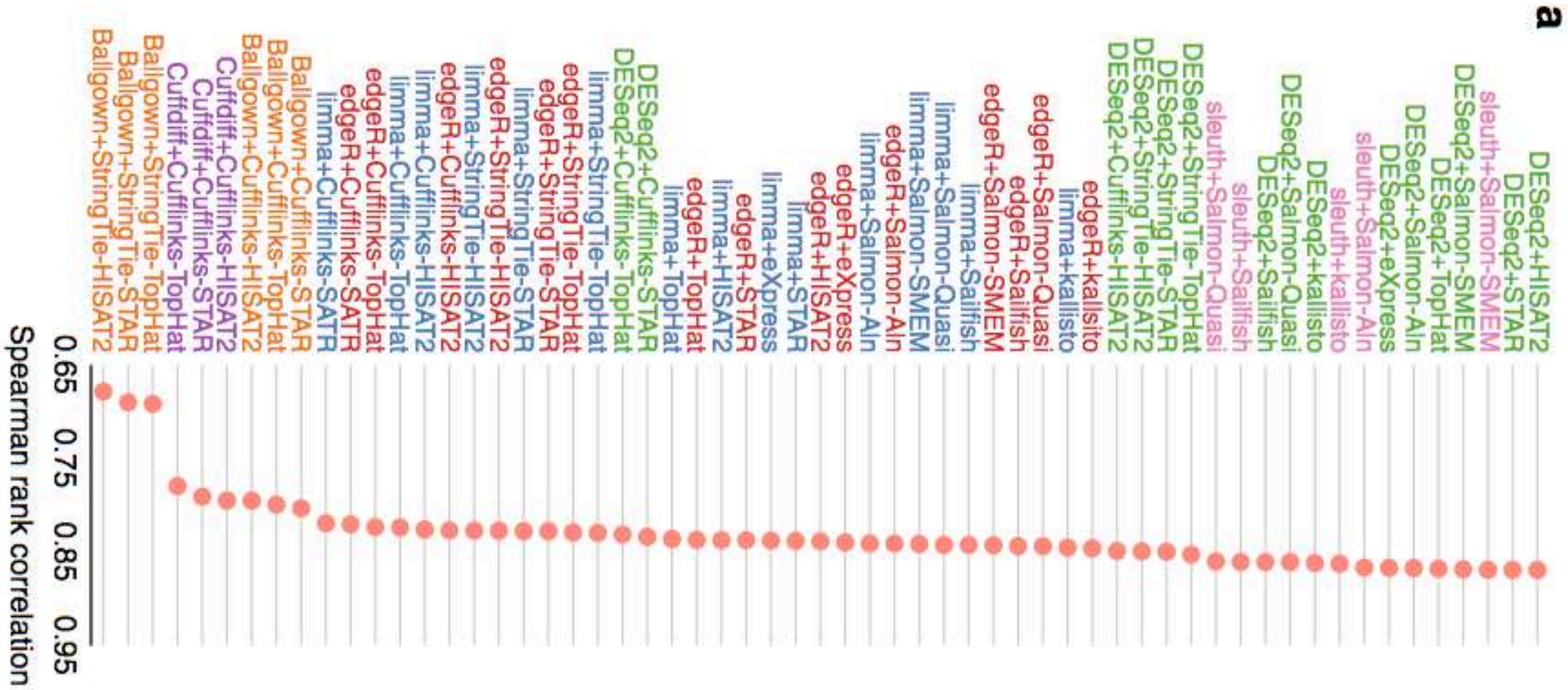
Advances in quantification



We present kallisto, an RNA-seq quantification program that is two orders of magnitude faster than previous approaches and achieves similar accuracy. Kallisto pseudoaligns reads to a reference, producing a list of transcripts that are compatible with each read while avoiding alignment of individual bases. We use kallisto to analyze 30 million unaligned paired-end RNA-seq reads in <10 min on a standard laptop computer. This removes a major computational bottleneck in RNA-seq analysis.

Bray *et al* Nature Biotechnology (2016)

Spearman rank correlation of DEG results to qPCR measured genes

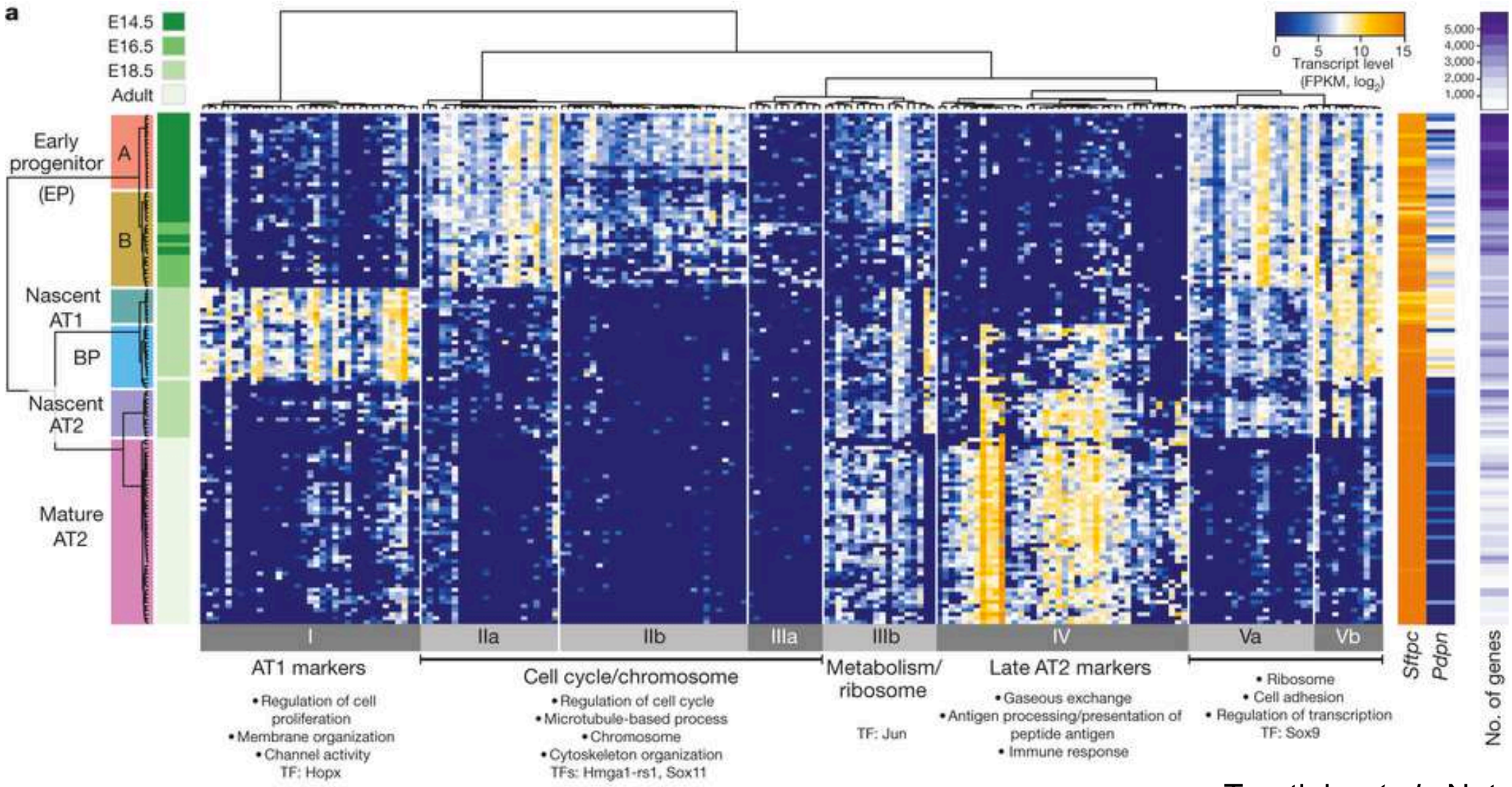


Sahraeian *et al* (2017) Nature Communications

Further analyses

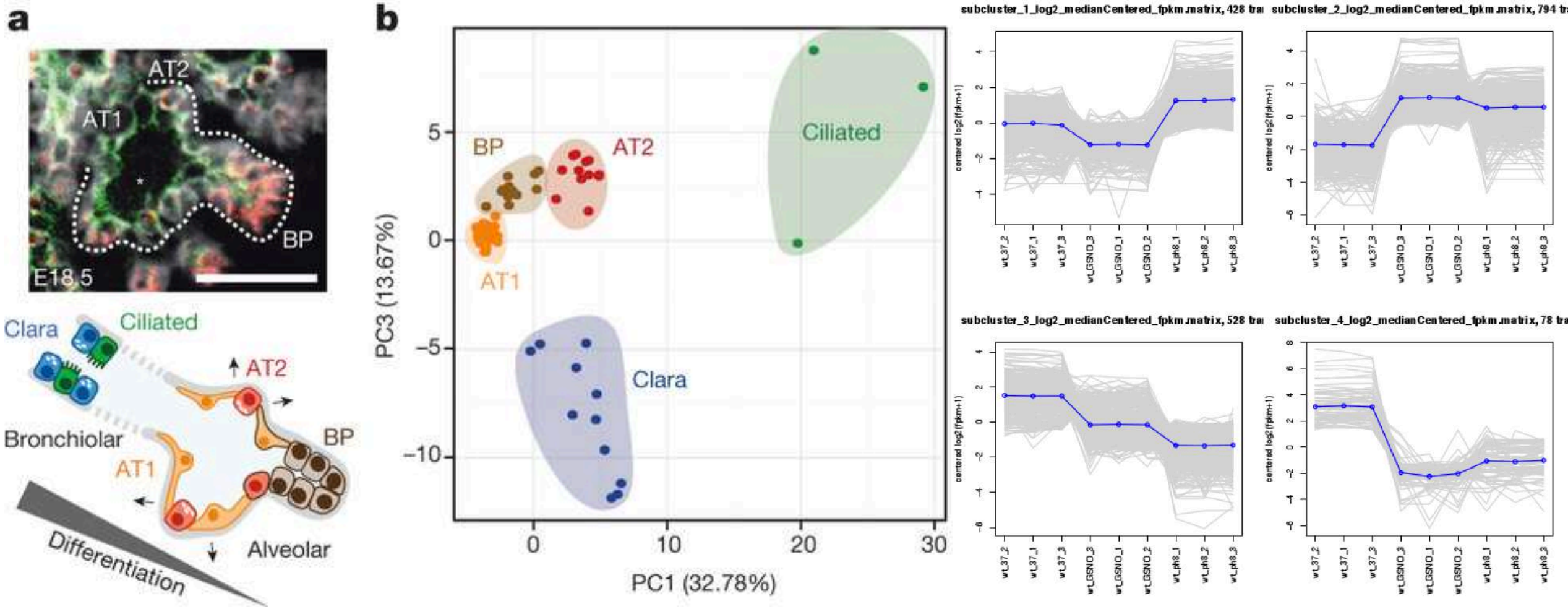
3.2 Once you have set of differentially expressed genes

Summarization visualizing the expression data through heatmap ; Classification using Gene Ontology terms and metabolic annotations



Treutlein *et al.*, Nature (2014)

Clustering of the expression values and principal component analysis to reduce the variables.

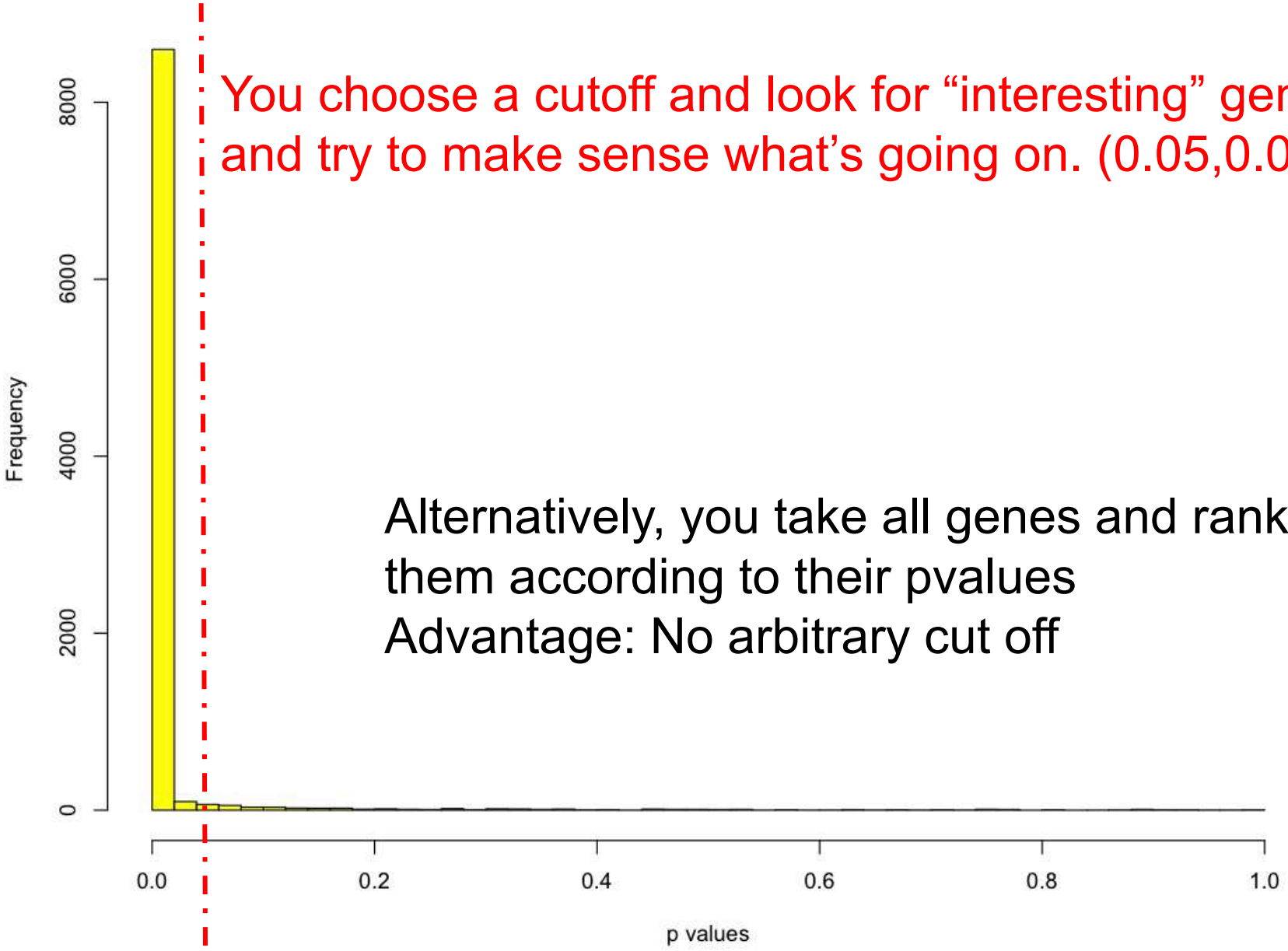


Treutlein *et al.*, Nature (2014)

Gene Ontology Enrichment analysis

GO.ID	Term	Annotated	Terms in list	Expected	p values
GO:0044281	small molecule metabolic process	481	150	54.18	< 1e-30
GO:0017144	drug metabolic process	155	72	17.46	3.20E-29
GO:0055114	oxidation-reduction process	308	103	34.7	4.30E-27
GO:0009126	purine nucleoside monophosphate metaboli.	79	47	8.9	2.10E-25
GO:0009167	purine ribonucleoside monophosphate meta	79	47	8.9	2.10E-25
GO:0072521	purine-containing compound metabolic pro..	129	61	14.53	2.30E-25
GO:0006163	purine nucleotide metabolic process	122	59	13.74	3.40E-25
GO:0009150	purine ribonucleotide metabolic process	119	58	13.41	5.30E-25
GO:0007218	neuropeptide signaling pathway	108	55	12.17	5.80E-25
GO:0019693	ribose phosphate metabolic process	138	62	15.55	3.00E-24
GO:0009161	ribonucleoside monophosphate metabolic p.	87	48	9.8	6.50E-24
GO:0009259	ribonucleotide metabolic process	129	59	14.53	1.30E-23
GO:0009117	nucleotide metabolic process	178	70	20.05	4.20E-23
GO:0006082	organic acid metabolic process	246	84	27.71	9.50E-23
GO:0019752	carboxylic acid metabolic process	232	81	26.13	1.20E-22
GO:0006753	nucleoside phosphate metabolic process	181	70	20.39	1.30E-22
GO:0009123	nucleoside monophosphate metabolic proce	97	49	10.93	4.00E-22
GO:0043436	oxoacid metabolic process	242	82	27.26	5.80E-22
GO:0055086	nucleobase-containing small molecule met..	204	74	22.98	7.10E-22
GO:0006091	generation of precursor metabolites and ...	111	52	12.5	1.70E-21

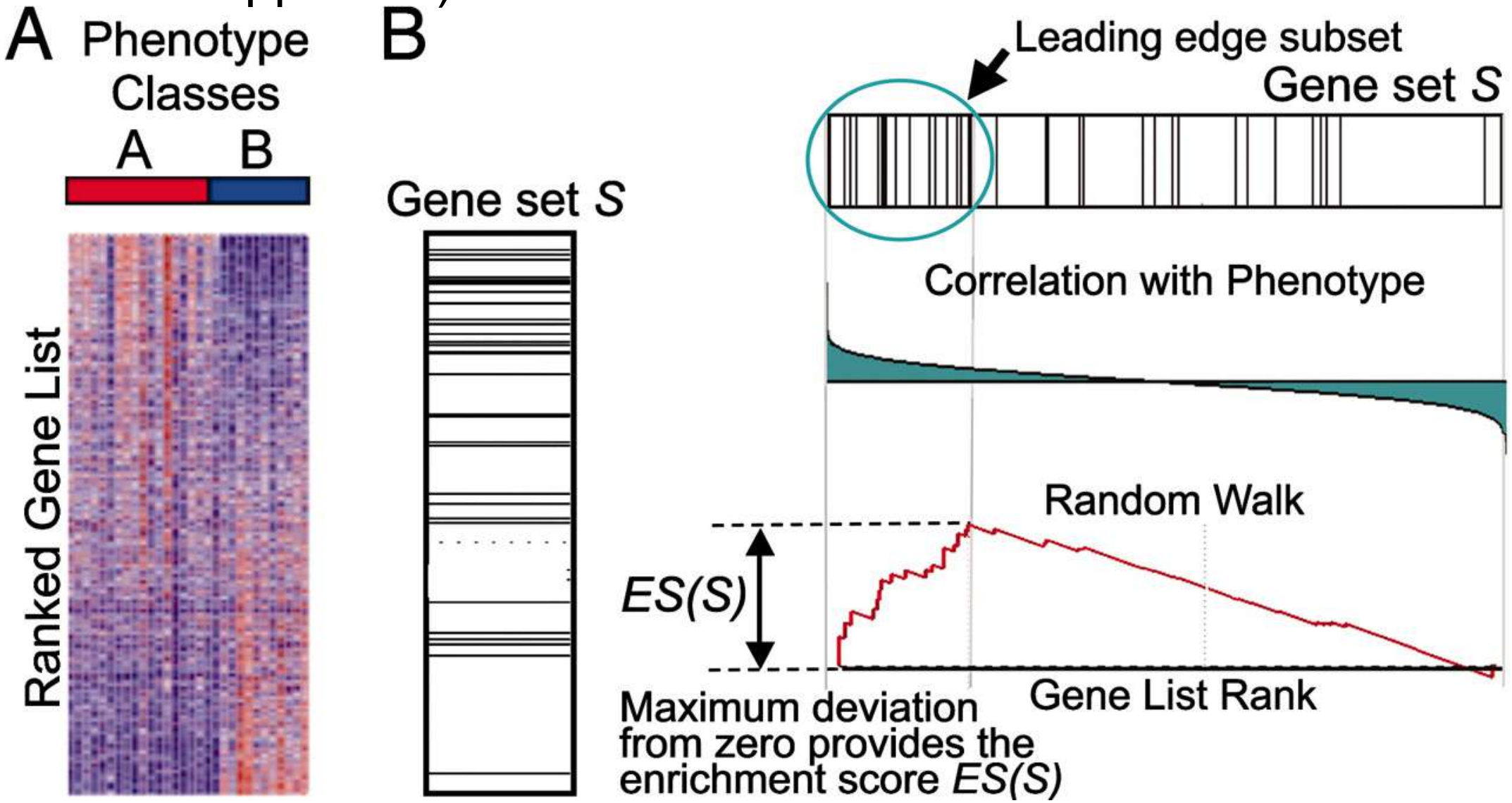
Now, setting a cut-off?



You choose a cutoff and look for “interesting” genes, and try to make sense what’s going on. (0.05,0.01?)

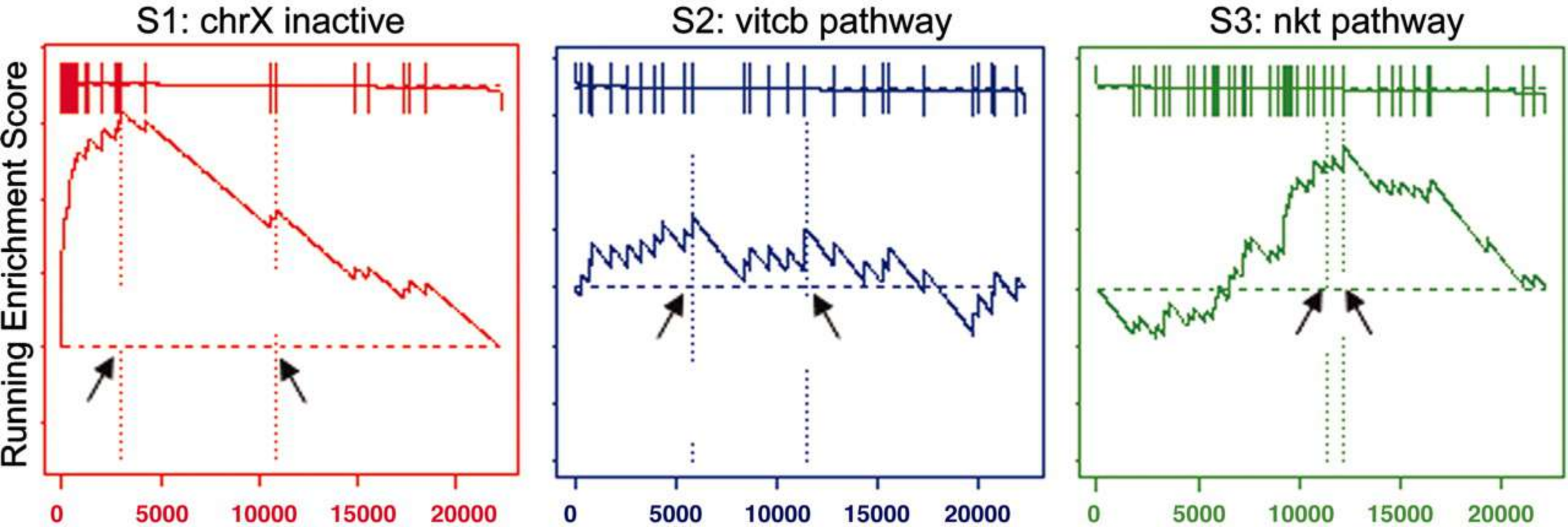
Alternatively, you take all genes and rank them according to their pvalues
Advantage: No arbitrary cut off

GSEA (Gene Set Enrichment Analysis) methods (cut-off free approach)



Subramanian *et al.*, PNAS (2005)

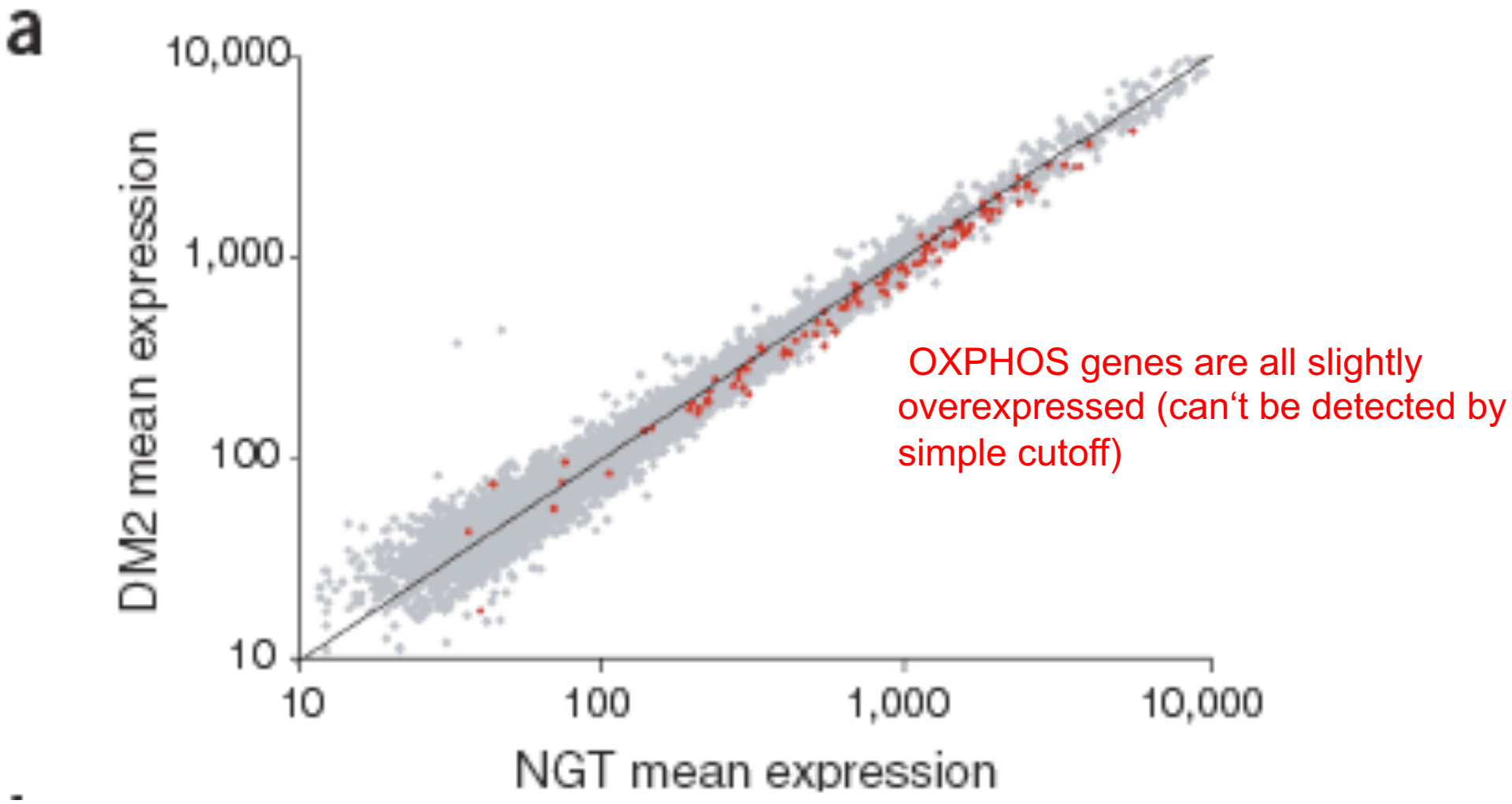
GSEA (Gene Set Enrichment Analysis) methods (cut-off free approach)



S1 is significantly enriched in females as expected, S2 is randomly distributed and scores poorly, and S3 is not enriched at the top of the list but is nonrandom, so it scores well.

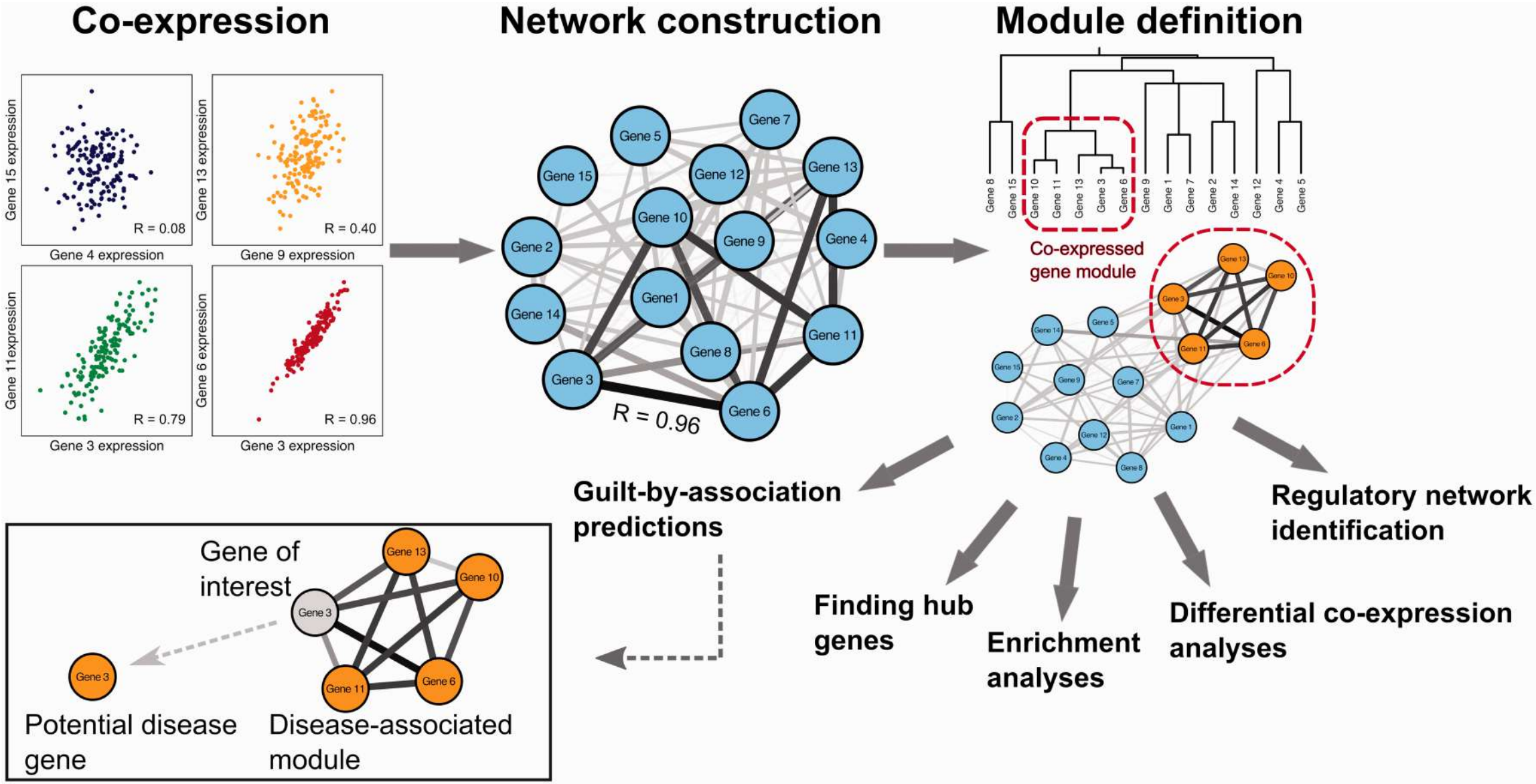
Subramanian *et al.,* PNAS (2005)

GSEA (Gene Set Enrichment Analysis) methods (cut-off free approach)



Mootha *et al.,* Nature Genetics (2003)

Gene co-expression network



Dam *et al* (2018) Briefings in Bioinformatics

Gene co-expression network

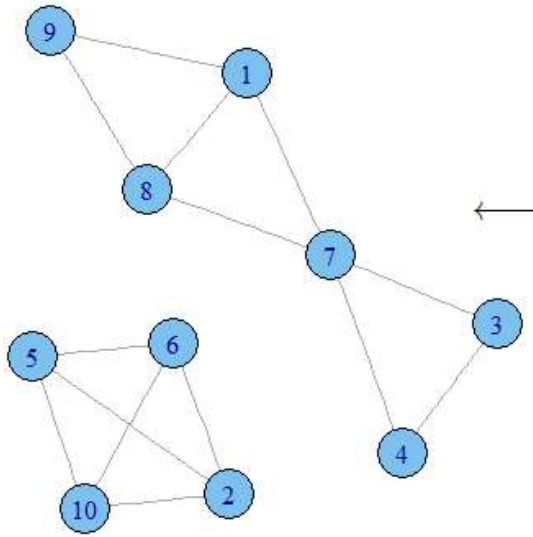
	S_1	S_2	S_3
G_1	43.26	40.89	5.05
G_2	166.6	41.87	136.65
G_3	12.53	39.55	42.09
G_4	28.77	191.92	236.56
G_5	114.7	79.7	99.76
G_6	119.1	80.57	114.59
G_7	118.9	156.69	186.95
G_8	3.76	2.48	136.78
G_9	32.73	11.99	118.8
G_{10}	17.46	56.11	21.41

Gene expression values

$\xrightarrow{|r(G_i,G_j)|}$ Pearson correlation

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	1.00	0.23	0.61	0.71	0.03	0.35	**0.86**	**1.00**	**0.97**	0.37
G_2	0.23	1.00	0.63	0.52	**0.98**	**0.99**	0.29	0.30	0.46	**0.99**
G_3	0.61	0.63	1.00	**0.99**	0.77	0.53	**0.93**	0.56	0.41	0.51
G_4	0.71	0.52	**0.99**	1.00	0.69	0.41	**0.97**	0.66	0.52	0.40
G_5	0.03	**0.98**	0.77	0.69	1.00	**0.95**	0.48	0.09	0.27	**0.94**
G_6	0.35	**0.99**	0.53	0.41	**0.95**	1.00	0.17	0.41	0.57	**1.00**
G_7	0.86	0.29	**0.93**	**0.97**	0.48	0.17	1.00	**0.83**	0.72	0.16
G_8	**1.00**	0.30	0.56	0.66	0.09	0.41	0.83	1.00	**0.98**	0.42
G_9	**0.97**	0.46	0.41	0.52	0.27	0.57	0.72	**0.98**	1.00	0.58
G_{10}	0.37	**0.99**	0.51	0.40	**0.94**	**1.00**	0.16	0.42	0.58	1.00

Similarity (Co-expression) score



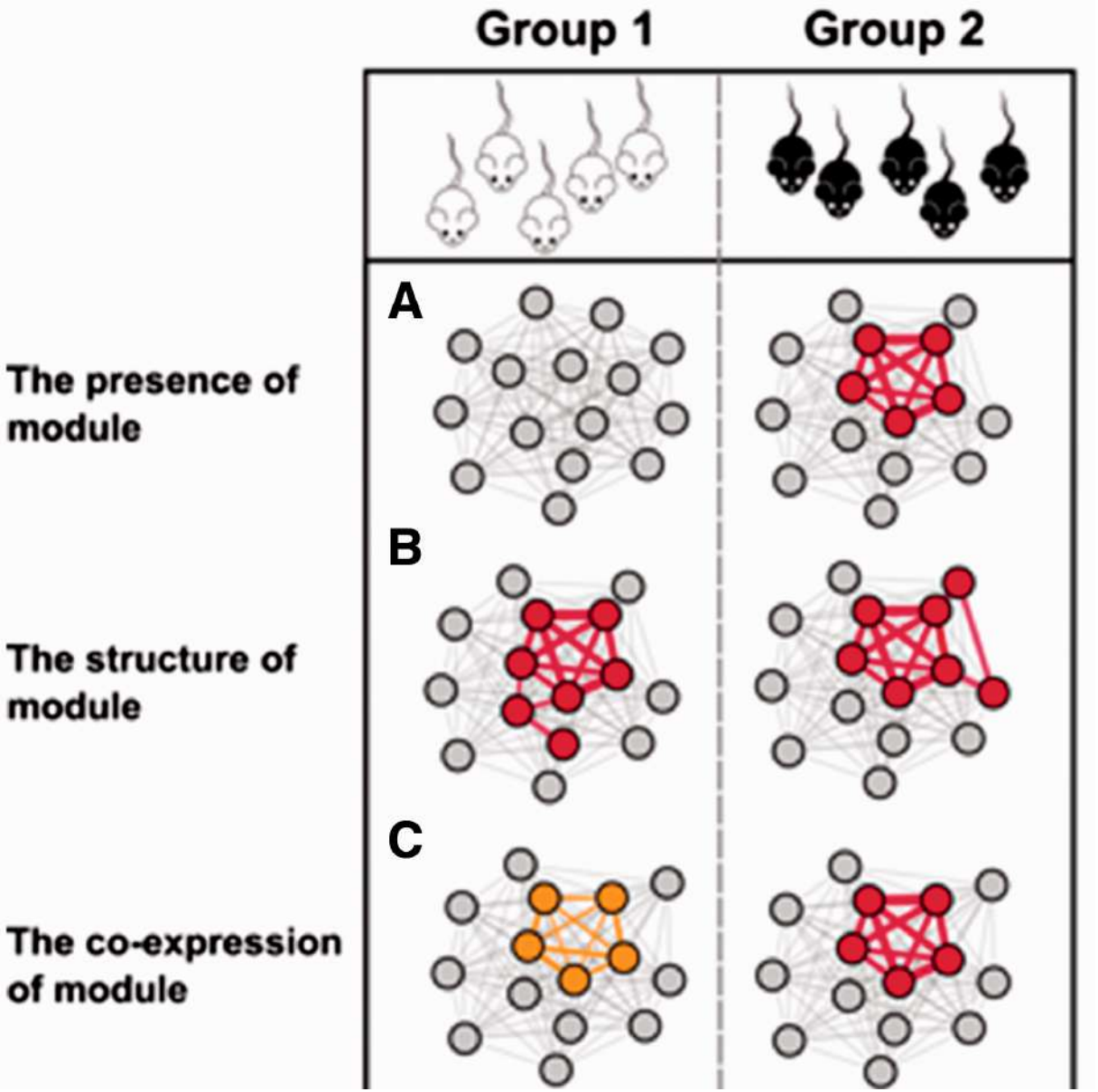
	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	0	0	0	0	0	0	1	1	1	0
G_2	0	0	0	0	1	1	0	0	0	1
G_3	0	0	0	1	0	0	1	0	0	0
G_4	0	0	1	0	0	0	1	0	0	0
G_5	0	1	0	0	0	1	0	0	0	1
G_6	0	1	0	0	1	0	0	0	0	1
G_7	1	0	1	1	0	0	0	1	0	0
G_8	1	0	0	0	0	0	1	0	1	0
G_9	1	0	0	0	0	0	0	1	0	0
G_{10}	0	1	0	0	1	1	0	0	0	0

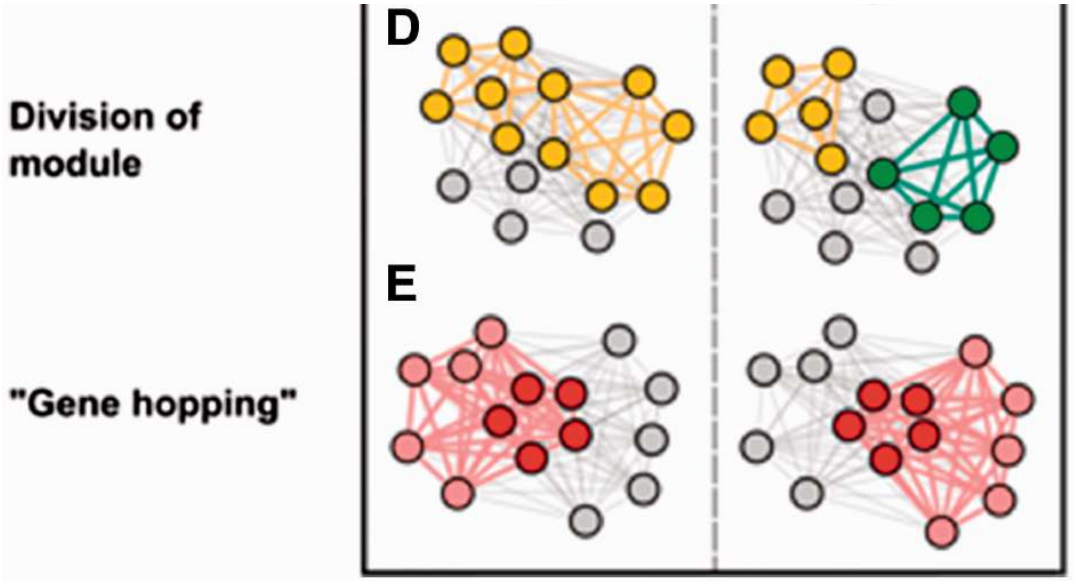
Network adjacency matrix

$\xleftarrow{|r(G_i,G_j)|>=0.8}$ Significance threshold

https://en.wikipedia.org/wiki/Gene_co-expression_network

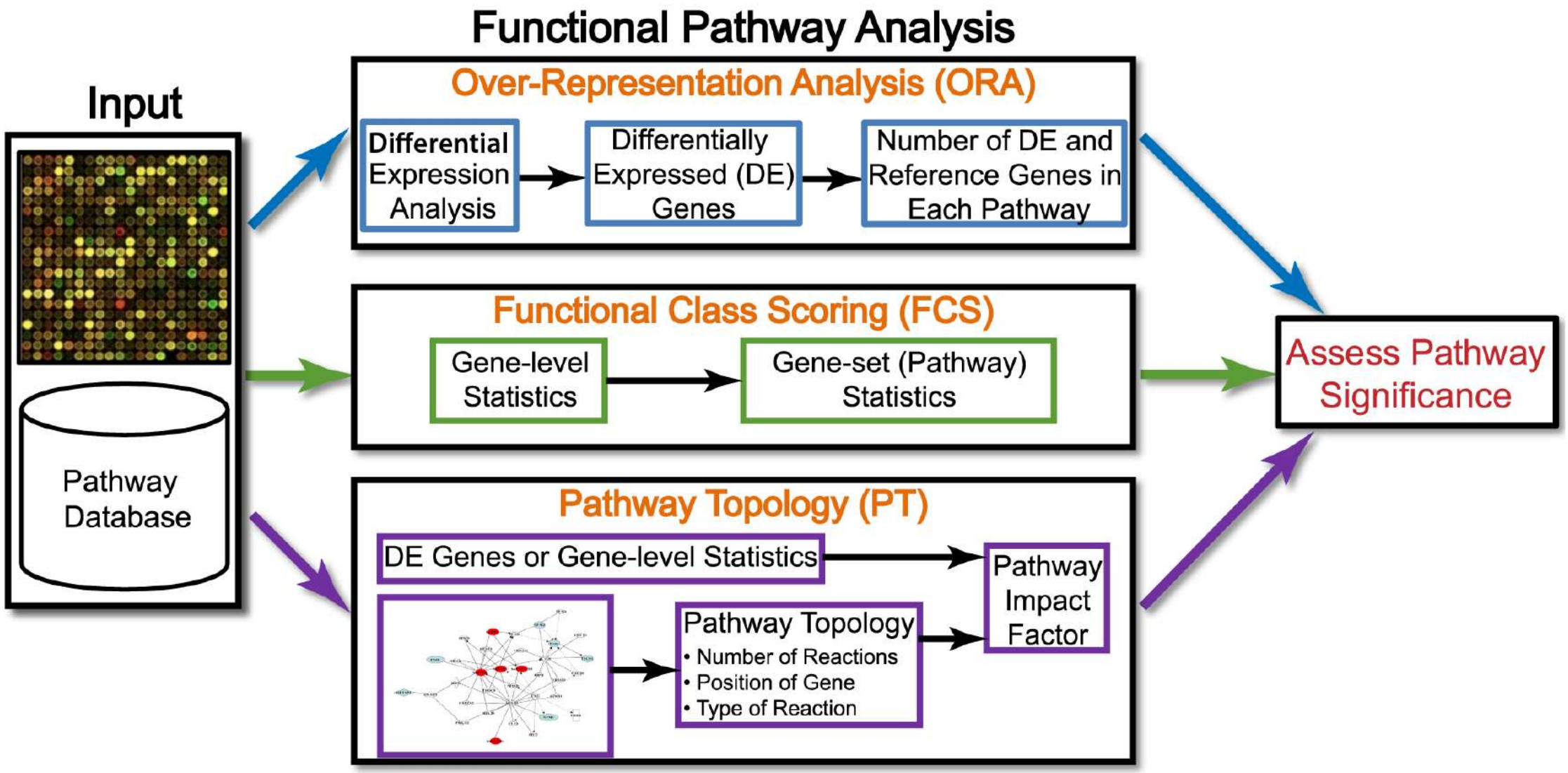
Gene co-expression network





Dam *et al* (2018) Briefings in Bioinformatics

Overview of existing pathway analysis methods using gene expression data as an example (only applicable to model species)

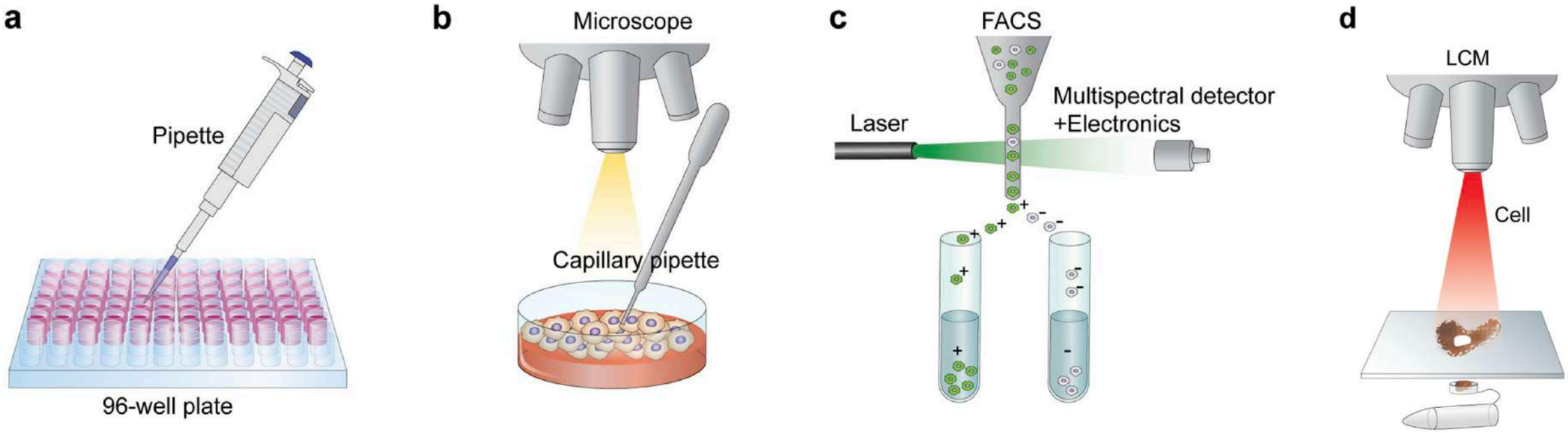


Kharati *et al.*, PLOS Computational Biology (2012)

Break

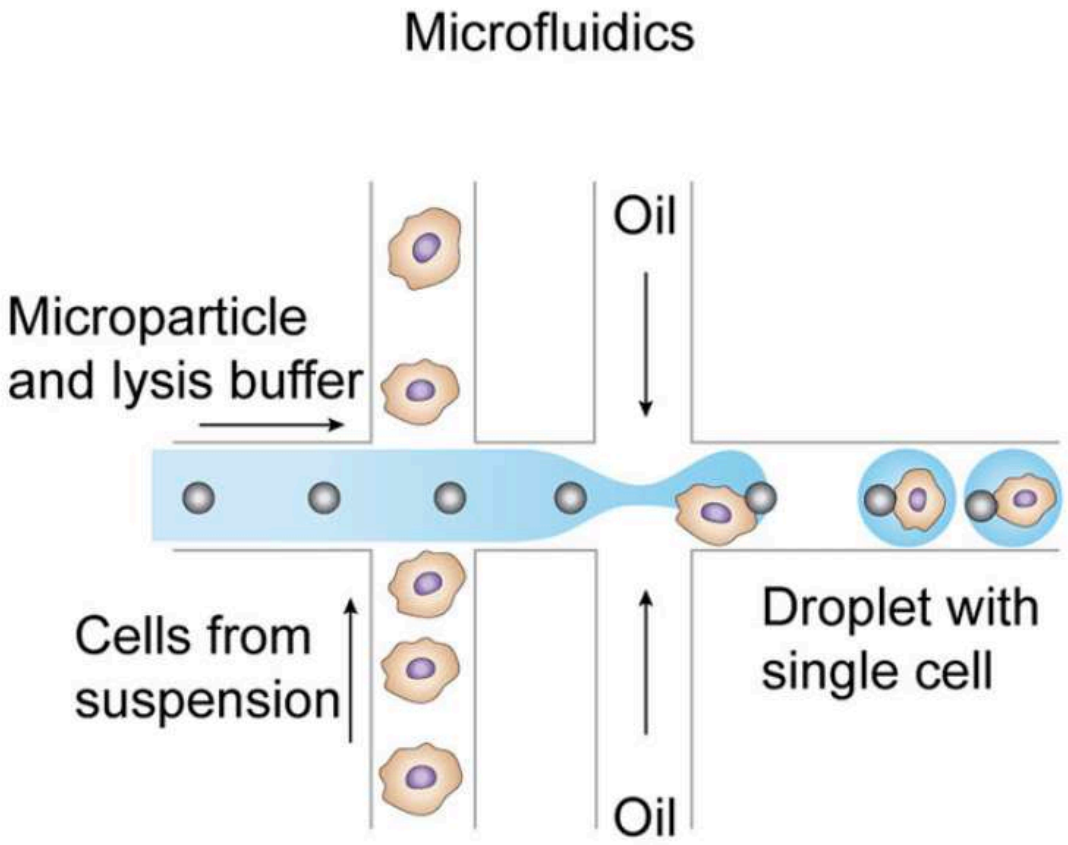
Further advances
Single cell RNAseq (ScRNAseq)

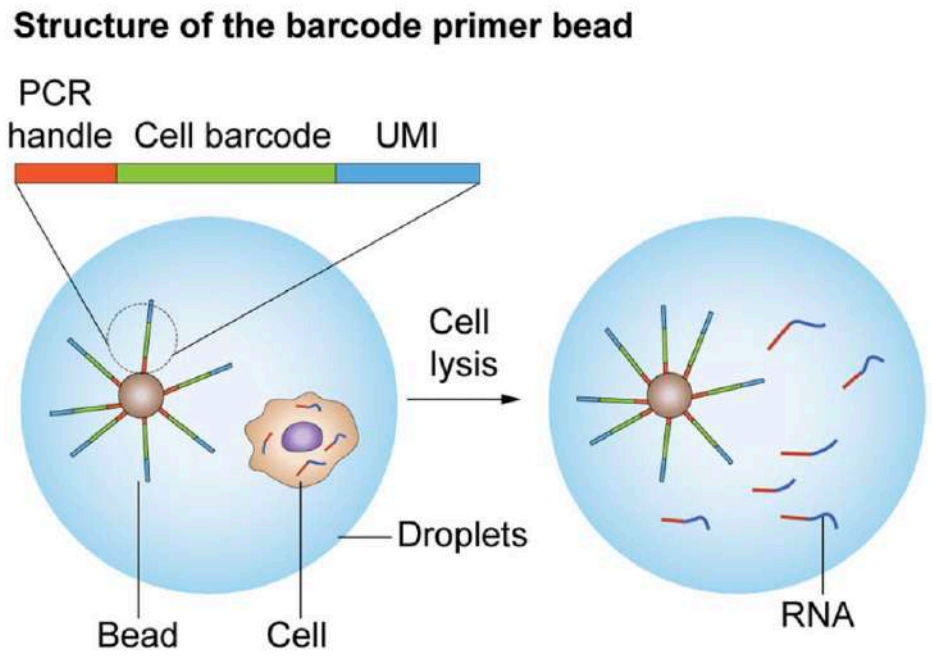
Evolution of single-cell isolation

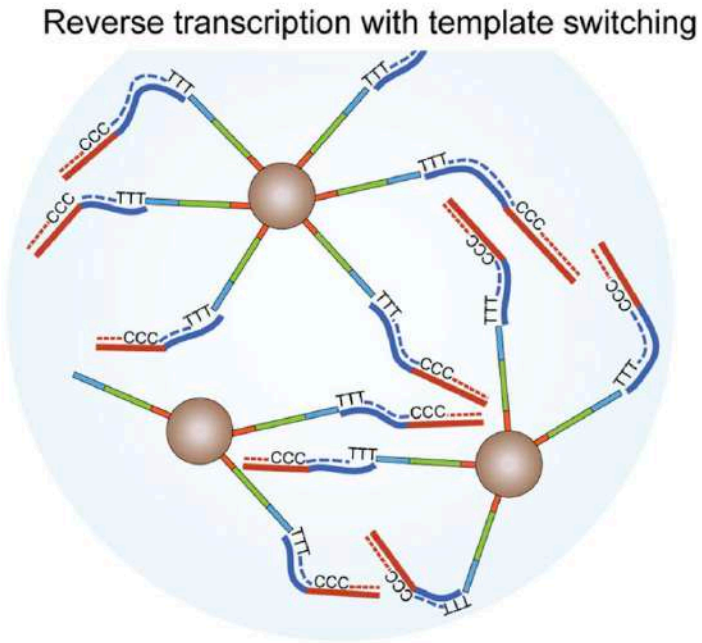


Huang *et al.,* Experimental & Molecular Medicine (2018) 50:96

Microfluidic isolation in reagent- filled droplets



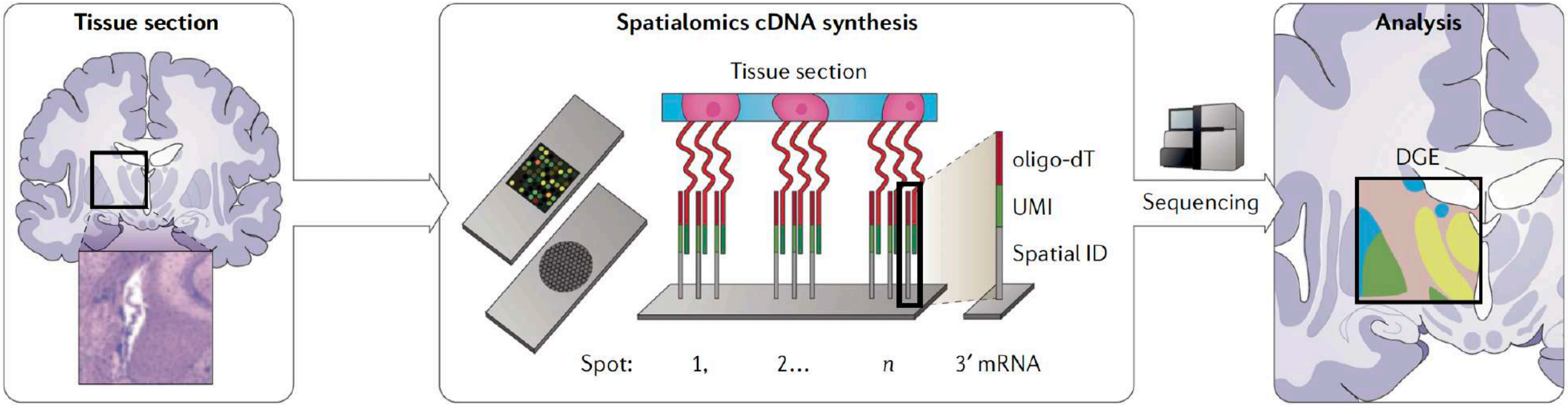




Huang *et al.,* Experimental & Molecular Medicine (2018) 50:96

Spatialomics

1. Spatial encoding requires a frozen tissue section to be applied to oligo- arrayed microarray slides or to ‘pucks’ of densely packed oligo- coated beads.

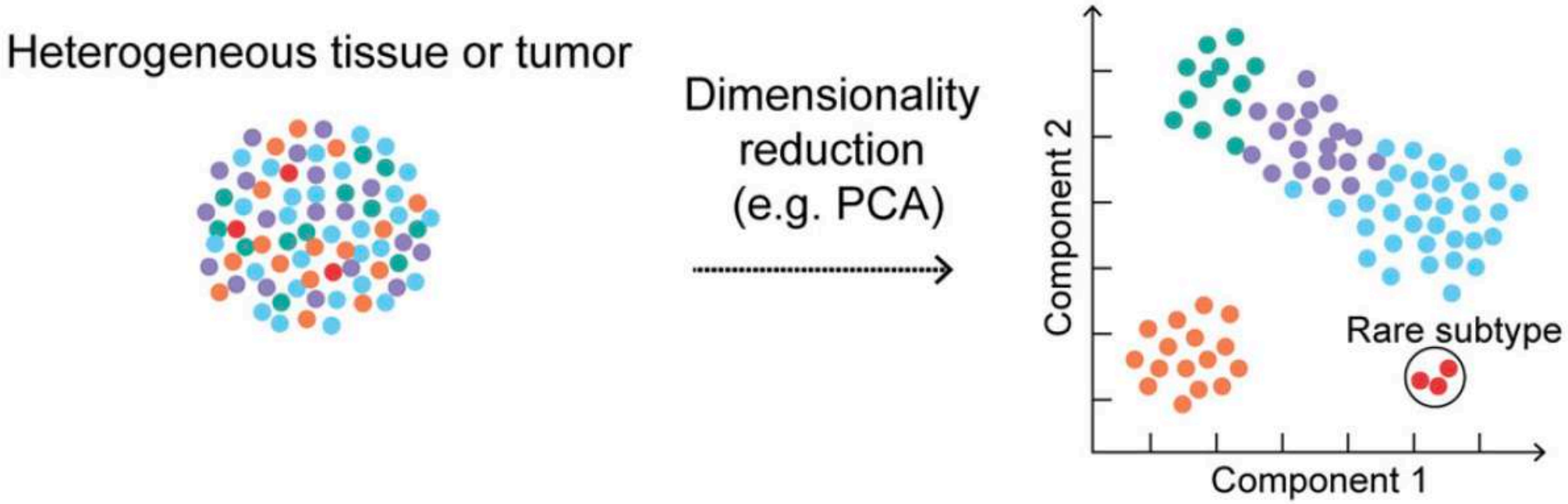


2. The mRNA diffuses to the slide surface and hybridizes to oligo- dT cDNA synthesis primers that encode UMIs and spatial barcodes. It is then reverse transcribed to produce cDNA, which is pooled for library preparation and sequencing.

3. Computational analysis of the spatialomics data maps sequence reads back to their spatial coordinates after DGE analysis and allows differential spatial expression to be visualized.

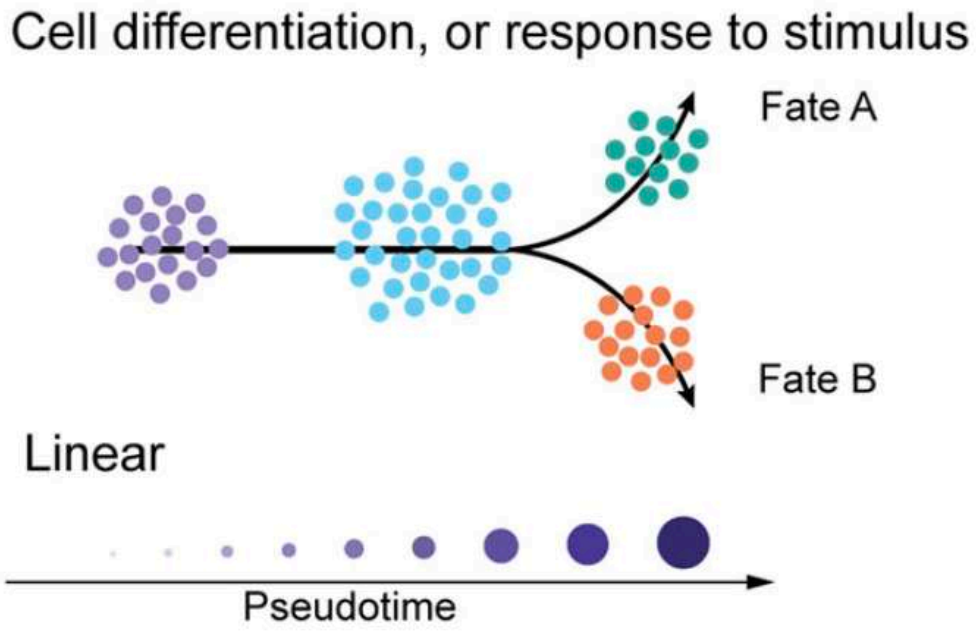
Stark *et al.,* Nature Review Genetics (2019)

Applications of scRNAseq computational approaches
1. Cell type identification



Huang *et al.,* Experimental & Molecular Medicine (2018) 50:96

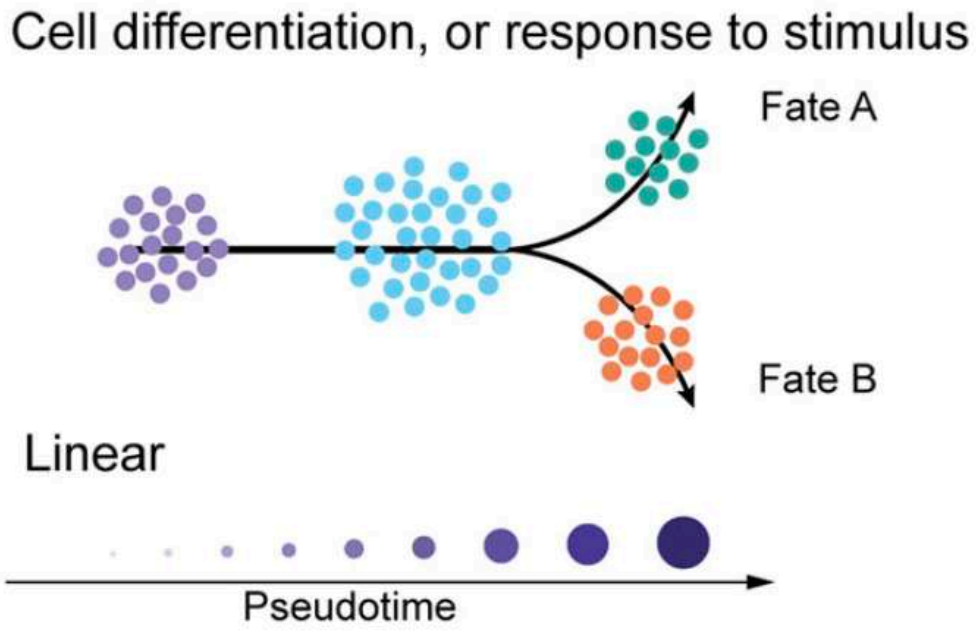
Applications of scRNAseq computational approaches
2. Cell hierarchy reconstruction

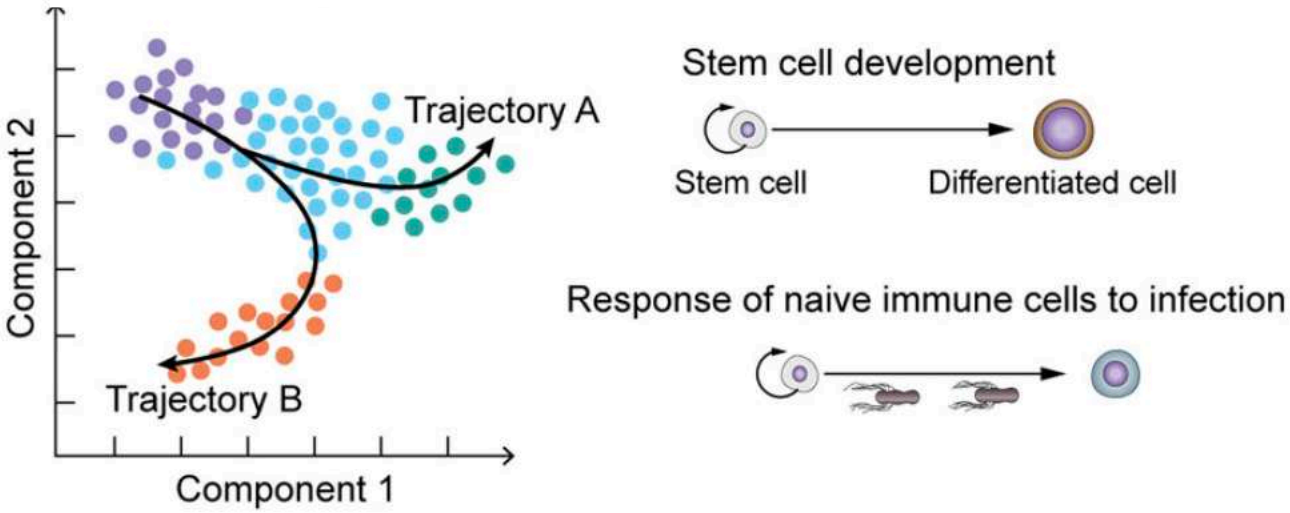


Trajectory analysis pipeline (Monocle)

Huang *et al.,* Experimental & Molecular Medicine (2018) 50:96

Applications of scRNAseq computational approaches
2. Cell hierarchy reconstruction

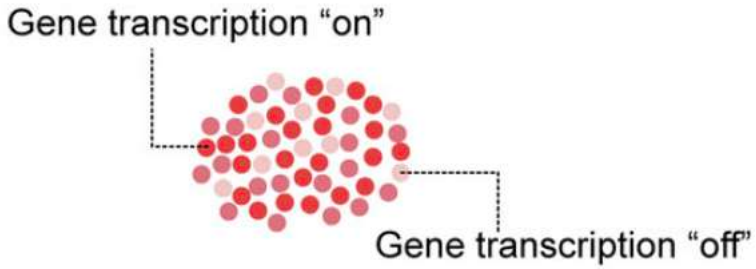




Trajectory analysis pipeline (Monocle)

Huang *et al.,* Experimental & Molecular Medicine (2018) 50:96

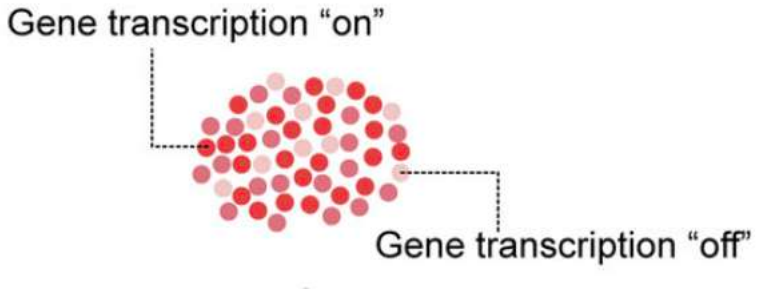
Applications of scRNAseq computational approaches
3. Inferring regulatory networks



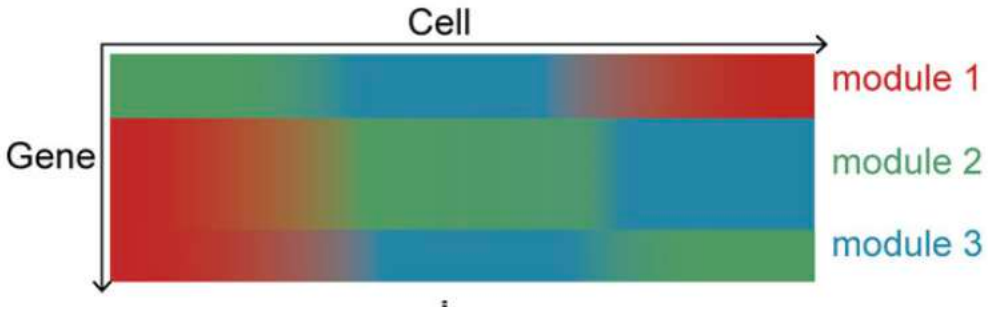
Identifying modules of
co-regulated genes

Huang *et al.,* Experimental & Molecular Medicine (2018) 50:96

Applications of scRNAseq computational approaches
3. Inferring regulatory networks

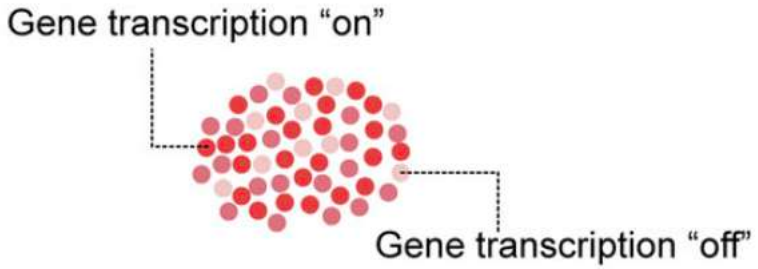


Identifying modules of co-regulated genes

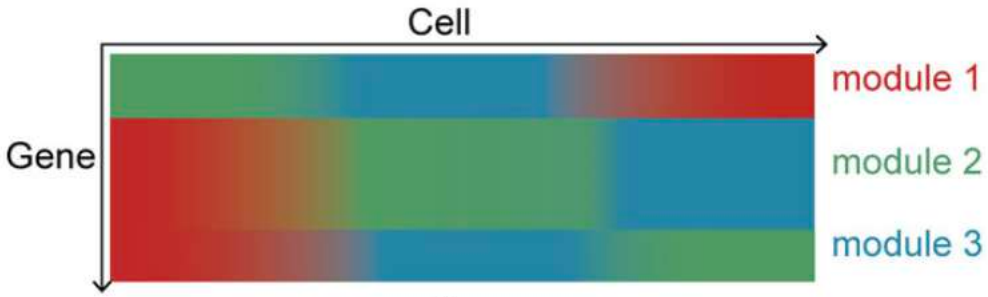


Huang *et al.,* Experimental & Molecular Medicine (2018) 50:96

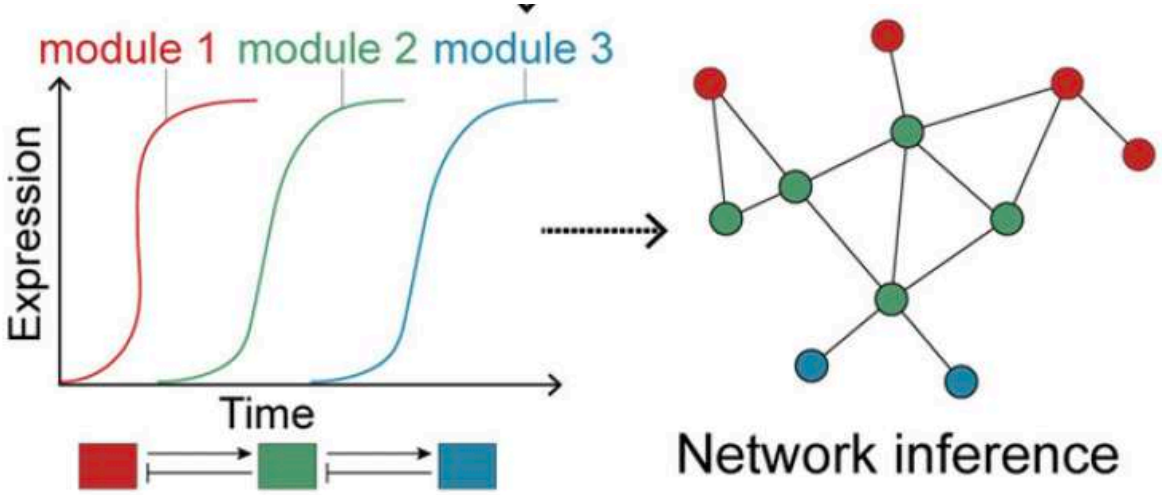
Applications of scRNAseq computational approaches
3. Inferring regulatory networks



Identifying modules of co-regulated genes



Pseudotime + Clustering

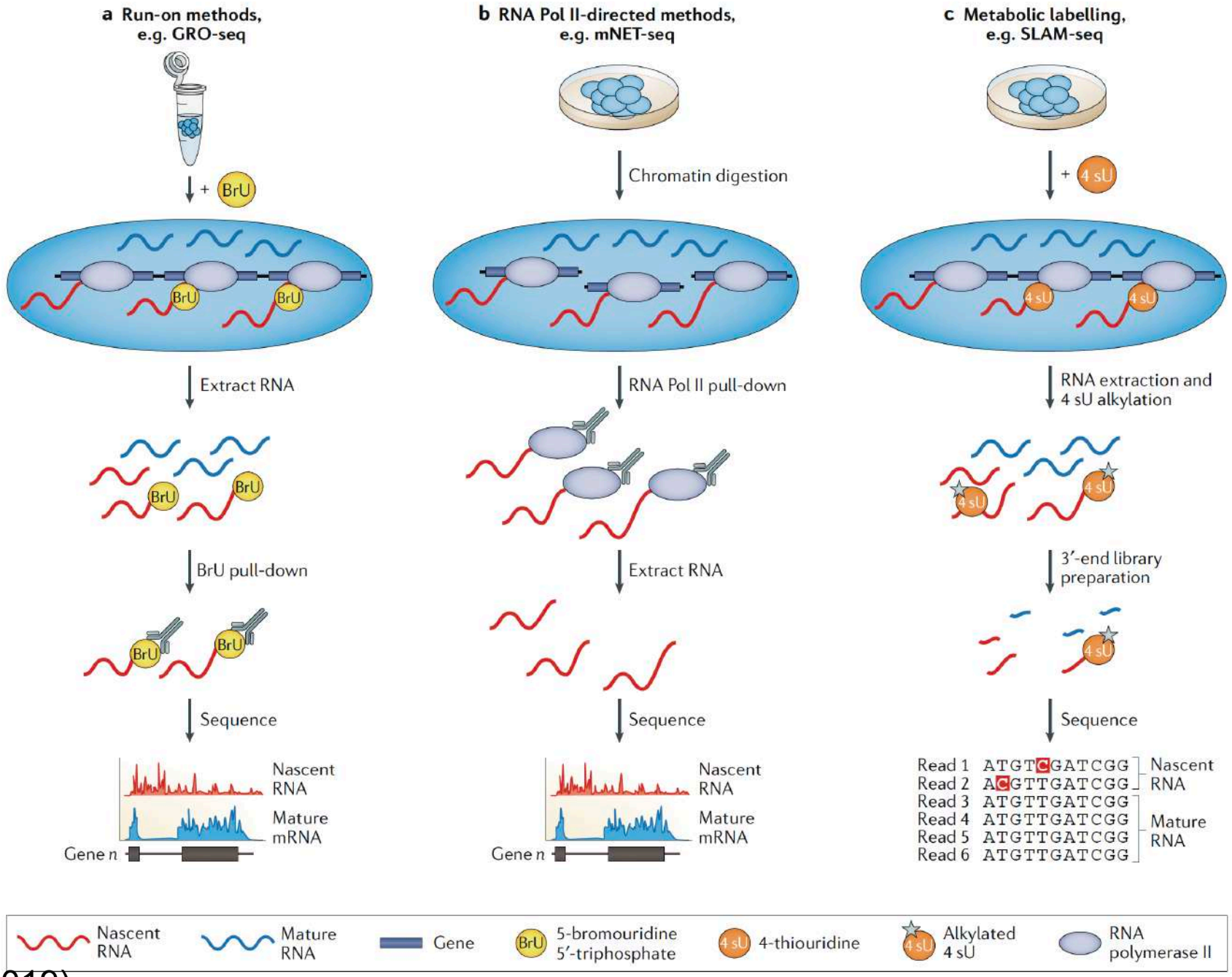


Huang *et al.,* Experimental & Molecular Medicine (2018) 50:96

Other approaches

nascent RNA

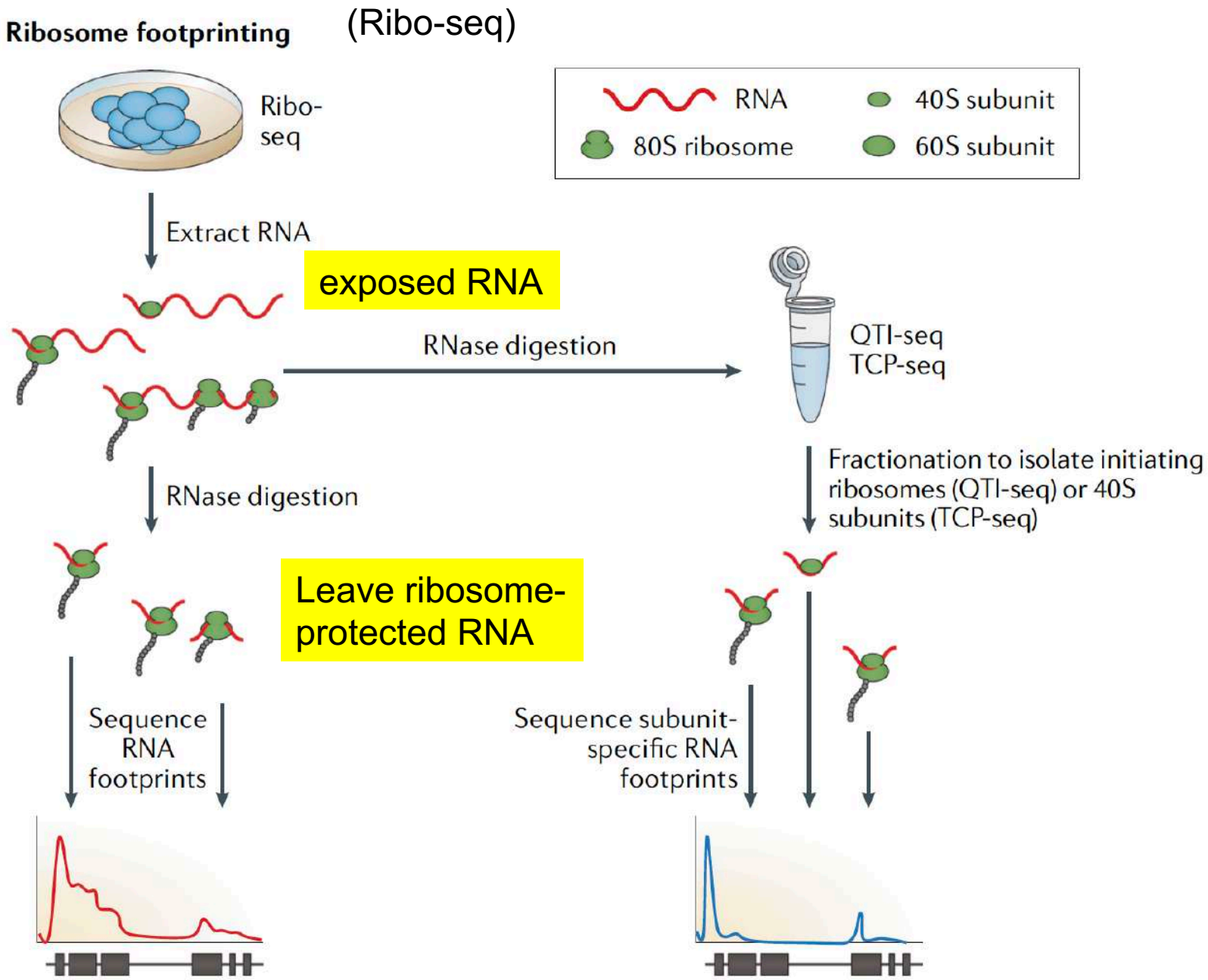
Essentially enrich newly
transcribed RNAs in a cell
and compare to control (mature RNA)



Stark *et al.,* Nature Review Genetics (2019)

translatome

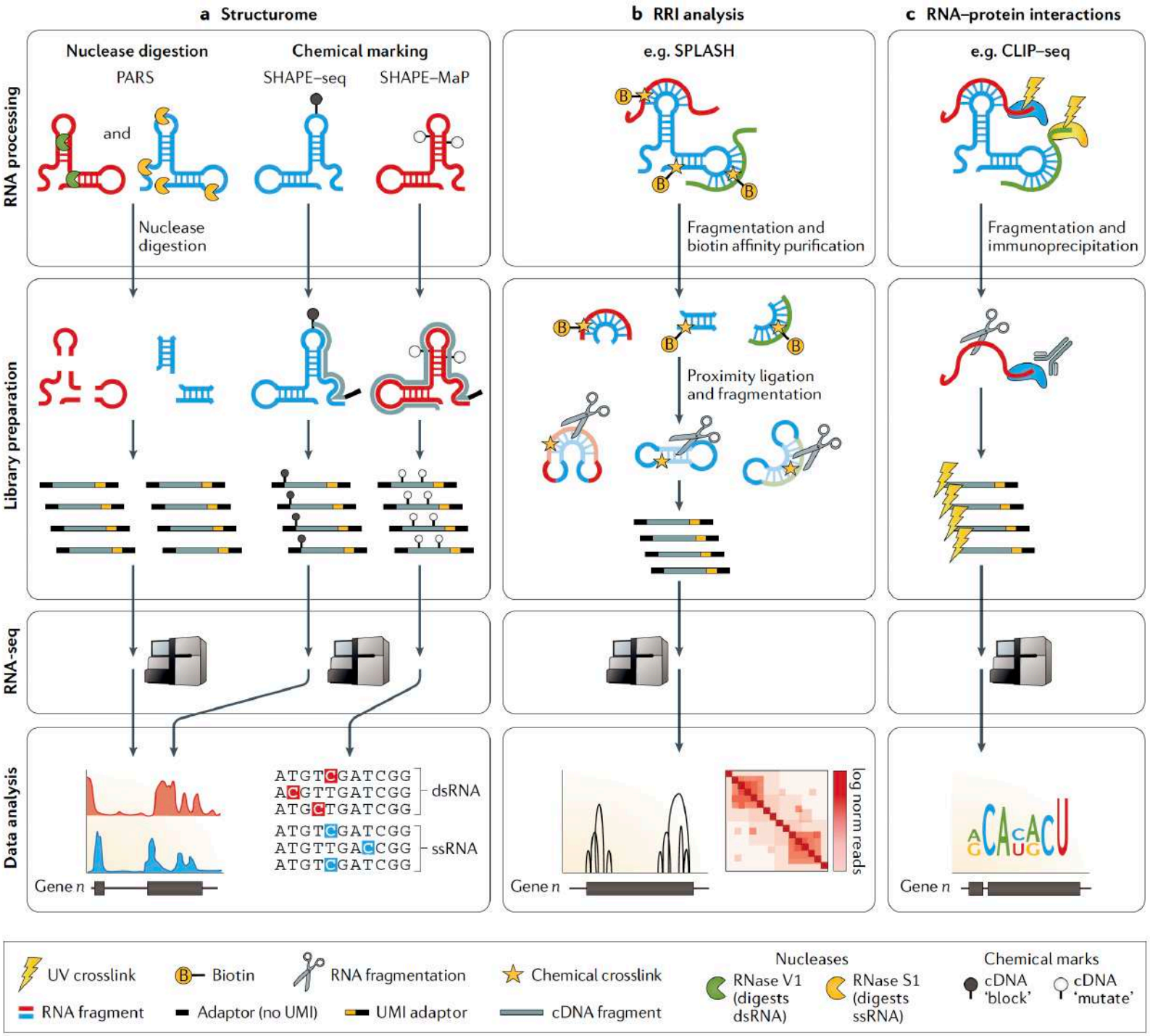
- RNA-sequencing from ribsomally bound RNA
- mRNA ribosome density correlates with the protein synthesis level



Stark *et al.,* Nature Review Genetics (2019)

RNA-RNA interaction
RNA-protein interaction

- A) Probe structured (ddRNA) or unstructured (ssRNA) RNA in transcriptome level
- B) Crosslinking interacting RNA with biotinylated psoralen
- C) Crosslinking immunoprecipitation of RNA followed by sequencing



Stark *et al.*, Nature Review Genetics (2019)

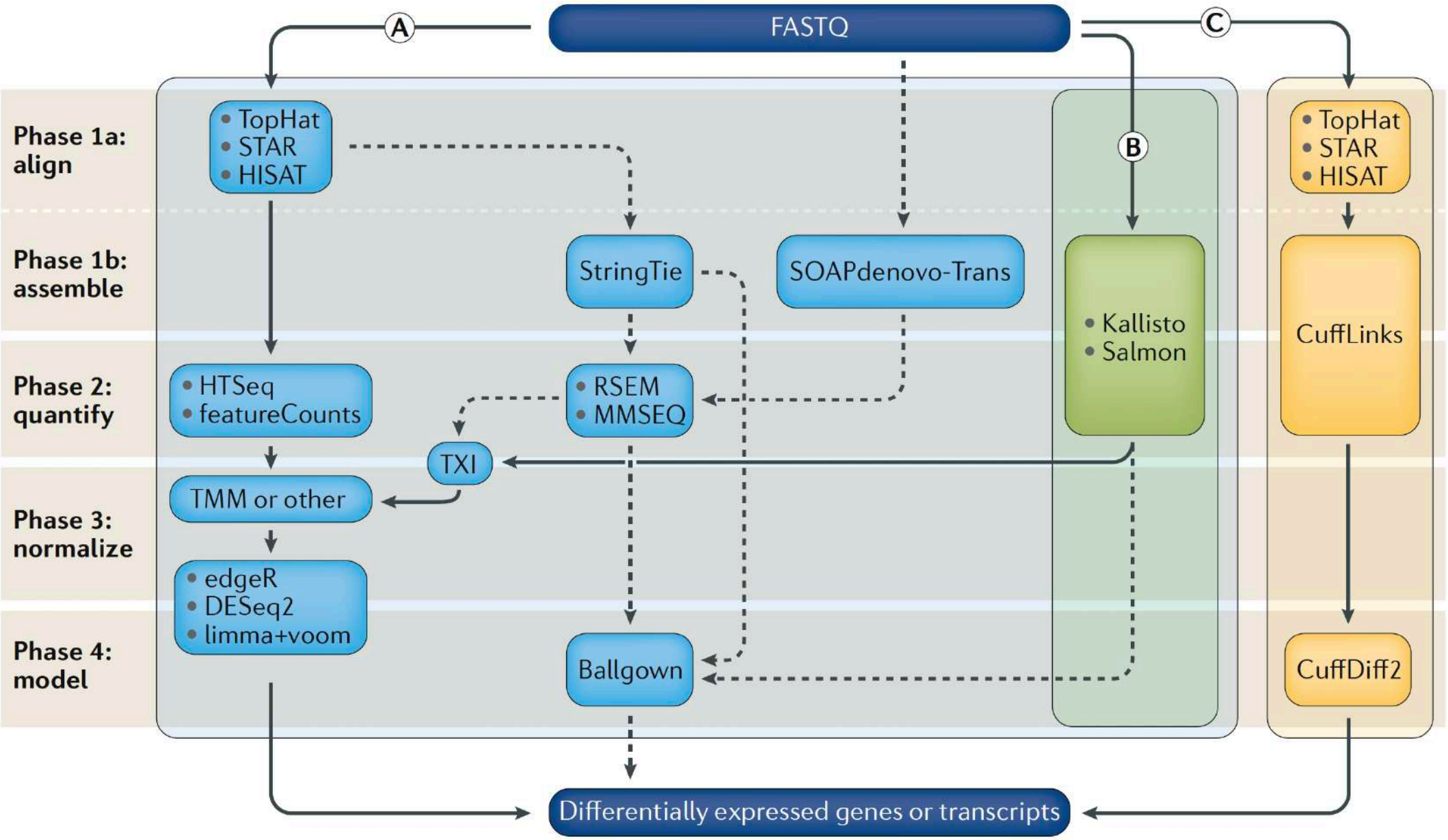
Summary (I)

Overview of technologies

Sequencing technology	Platform	Advantages	Disadvantages	Key applications
Short-read cDNA	Illumina, Ion Torrent	• Technology features very high throughput: currently 100–1,000 times more reads per run than long-read platforms • Biases and error profiles are well understood (homopolymers are still an issue for Ion Torrent) • A huge catalogue of compatible methods and computational workflows are available • Analysis works with degraded RNA	• Sample preparation includes reverse transcription, PCR and size selection adding biases to all methods • Isoform detection and quantitation can be limited • Transcript discovery methods require a de novo transcriptome alignment and/or assembly step	Nearly all RNA-seq methods have been developed for short-read cDNA sequencing: DGE, WTA, small RNA, single-cell, spatialomics, nascent RNA, translatome, structural and RNA–protein interaction analysis, and more are all possible
Long-read cDNA	PacBio, ONT	• Long reads of 1–50 kb capture many full-length transcripts • Computational methods for de novo transcriptome analysis are simplified	• Technology features low-to-medium throughput: currently only 500,000 to 10 million reads per run • Sample preparation includes reverse transcription, PCR and size selection (for some protocols), adding biases to many methods • Degraded RNA analysis is not recommended	Sequencing is particularly suited to isoform discovery, de novo transcriptome analysis, fusion transcript discovery, and MHC, HLA or other complex transcript analysis
Long-read RNA	ONT	• Long reads of 1–50 kb capture many full-length transcripts • Computational methods for de novo transcriptome analysis are simplified • Sample preparation does not require reverse transcription or PCR-reducing biases • RNA base modifications can be detected • Poly(A) tail lengths can be directly estimated from single-molecule sequencing	• Technology features low throughput: currently only 500,000 to 1 million reads per run • Sample preparation and sequencing biases are not well understood • Degraded RNA analysis is not recommended	• Sequencing is particularly suited to isoform discovery, de novo transcriptome analysis, fusion transcript discovery, and MHC, HLA or other complex transcript analysis • Ribonucelotide modifications can be detected

Stark, Grzelak and Hadfield (2019) Nature Reviews Genetics

RNAseq analysis workflow for differential expression (generalized)



Stark, Grzelak and Hadfield (2019) Nature Reviews Genetics

Summary / Our experiences

- Experimental design is key to correctly address your biological question
- Always use replicates (at least 5)
- Avoid *de novo* transcriptome assembly if you can

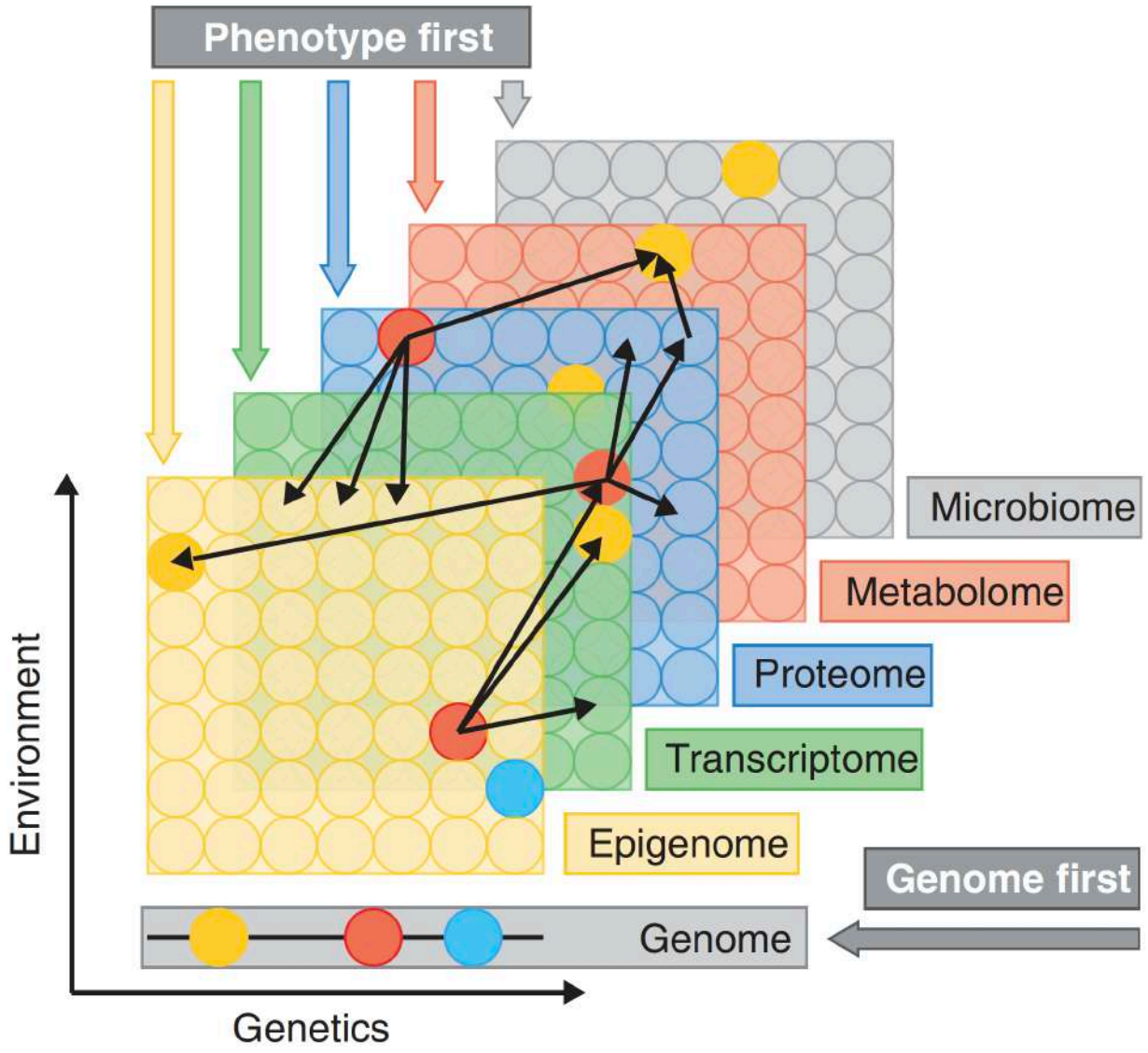
- DEseq2 are easy to use and have been standardised
- Cuffdiff2 are theoretically better but for some reasons are worse (since we used mostly 2-3 replicates)

- Still many challenges ahead

- **Question: What will be integrated/obsoleted within 5? years with the arrival of long read sequencing**

Which leads us to… a multi-omic perspective

Multiple omics data types



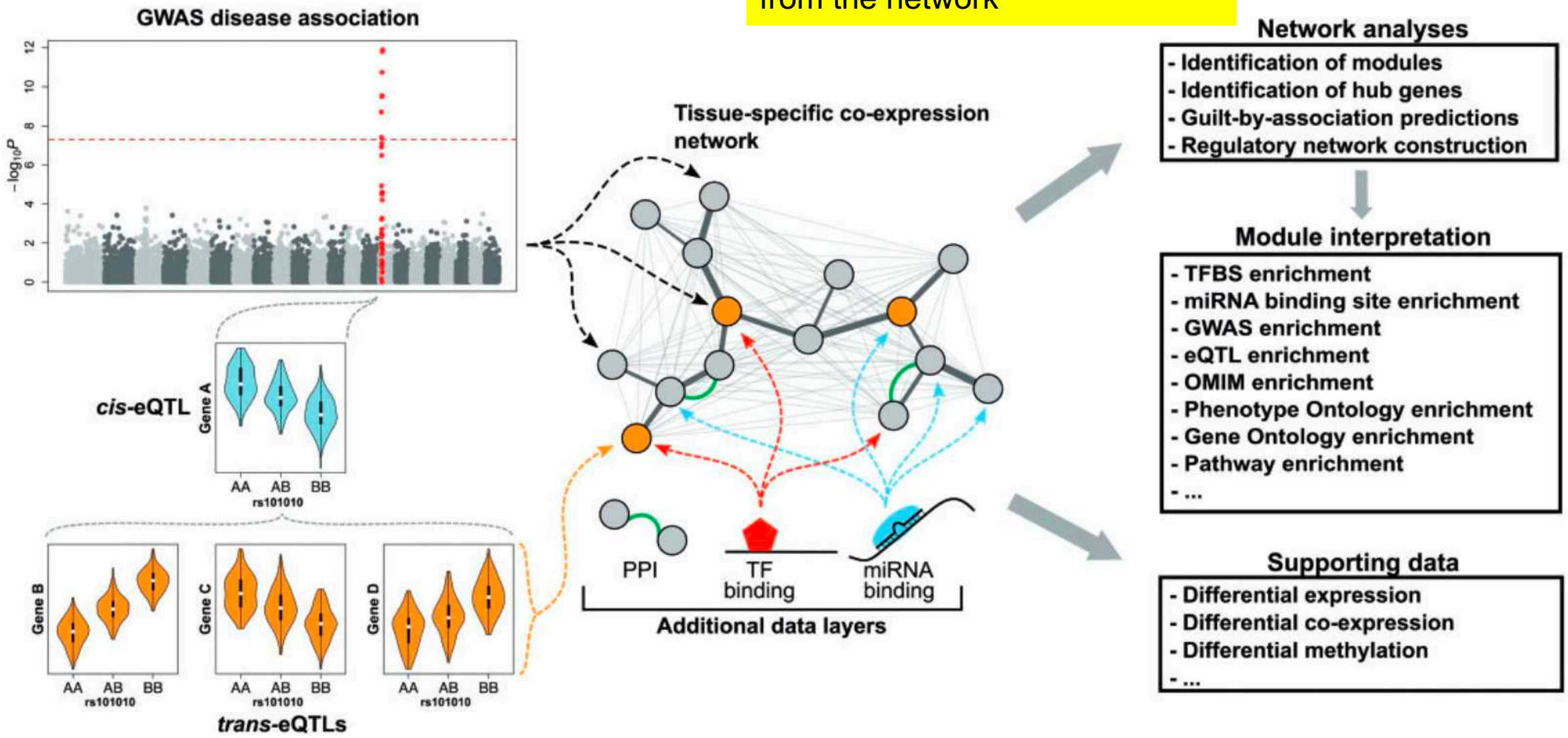
- Genome first or Phenotype first or environment first?
- Genome first -> GWAS

- "Locus-centered integration of additional omics layers can help to identify causal single nucleotide polymorphisms(SNPs) and genes at GWAS loci and then to examine how these perturb pathways leading to disease"

Hain *et al* (2017) Genome Biology

Integrating multi-omics to network

Various additional data can then be used to enrich and extract biological relevant information from the network

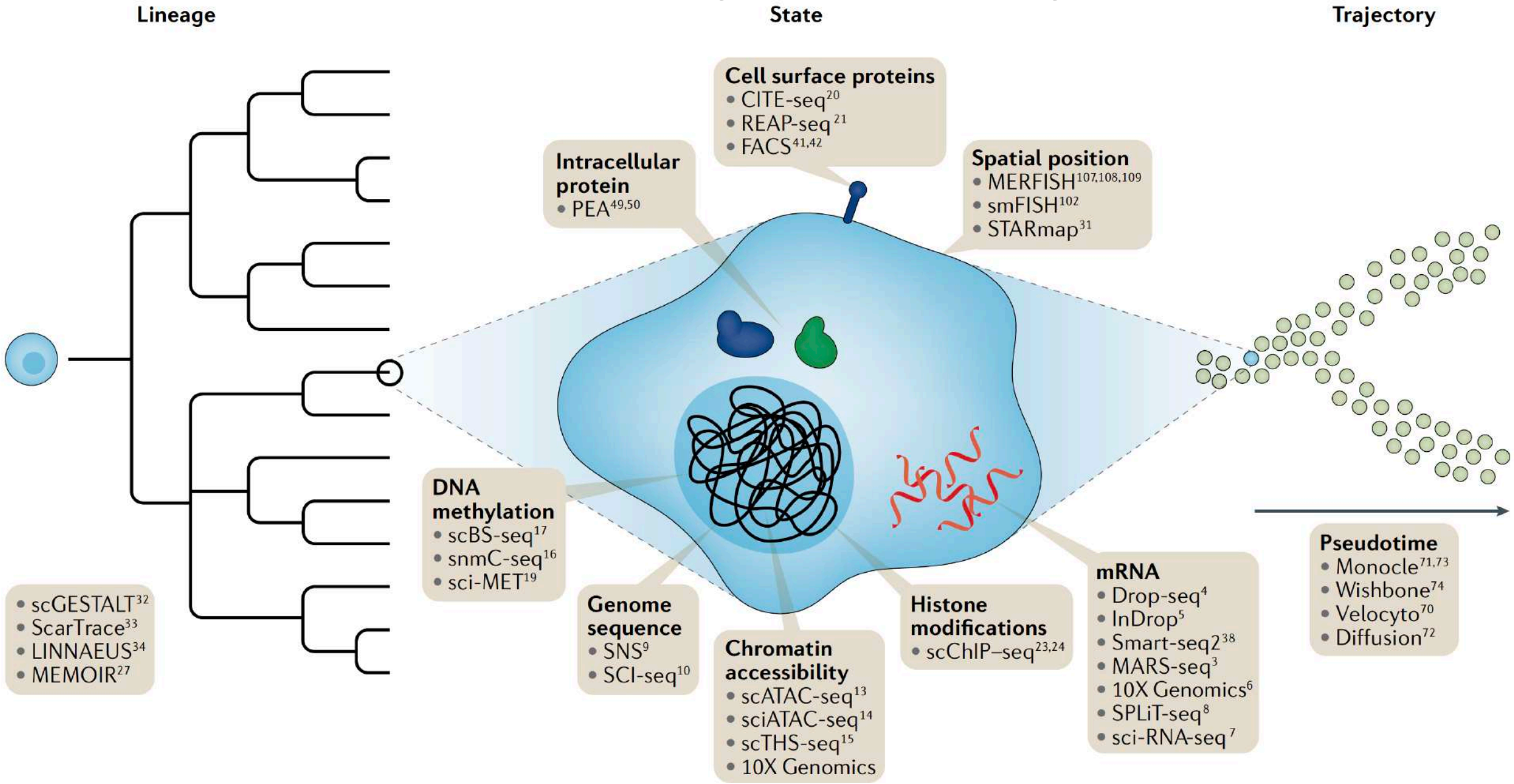


Dam *et al* (2016) Briefings in Bioinformatics

Example: FTO GWAS locus

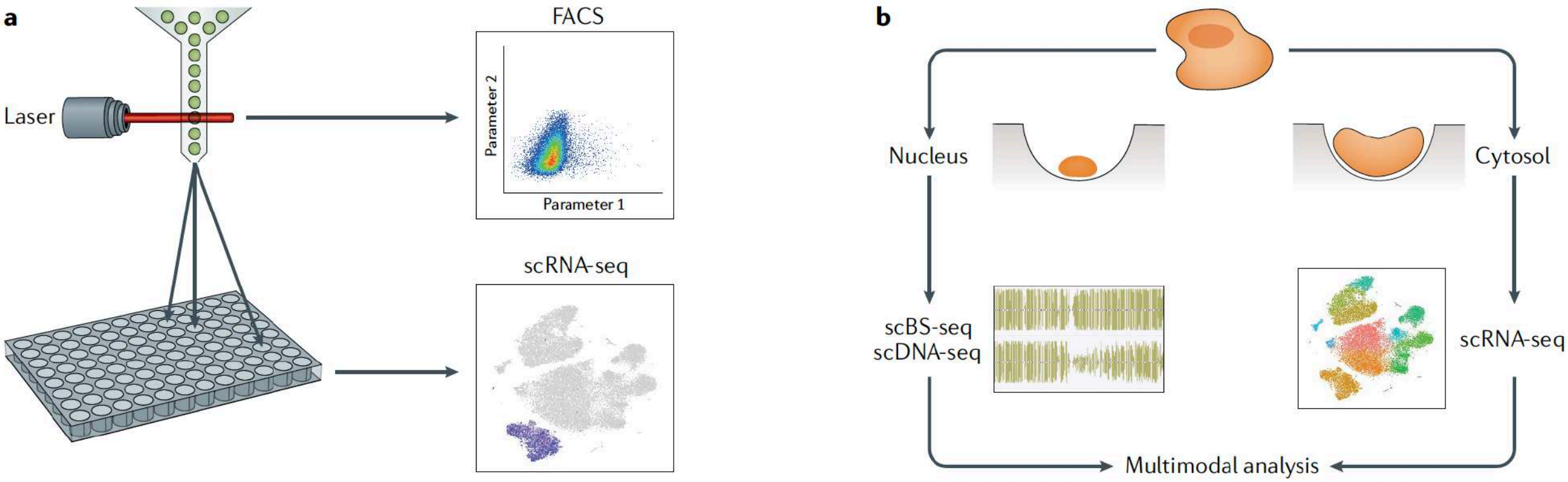
Claussnitzer *et al* (2015) N Engl J Med
Hain *et al* (2017) Genome Biology

Overview of current methods for single cell data integration



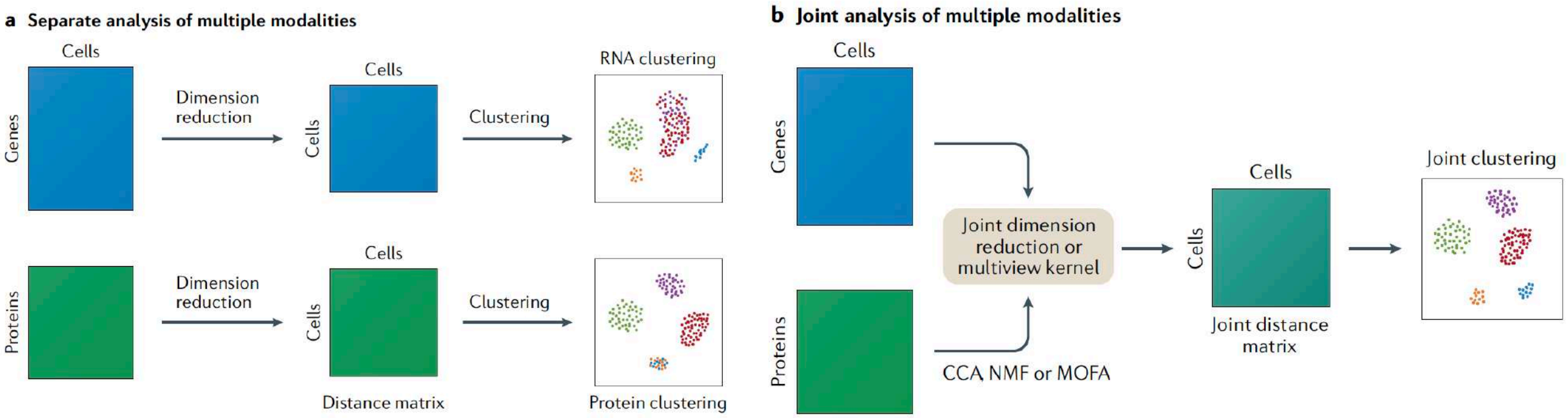
Stuart and Satija (2019) Nature Review Genetics

Example of experimental methods for performing single-cell multimodal measurements



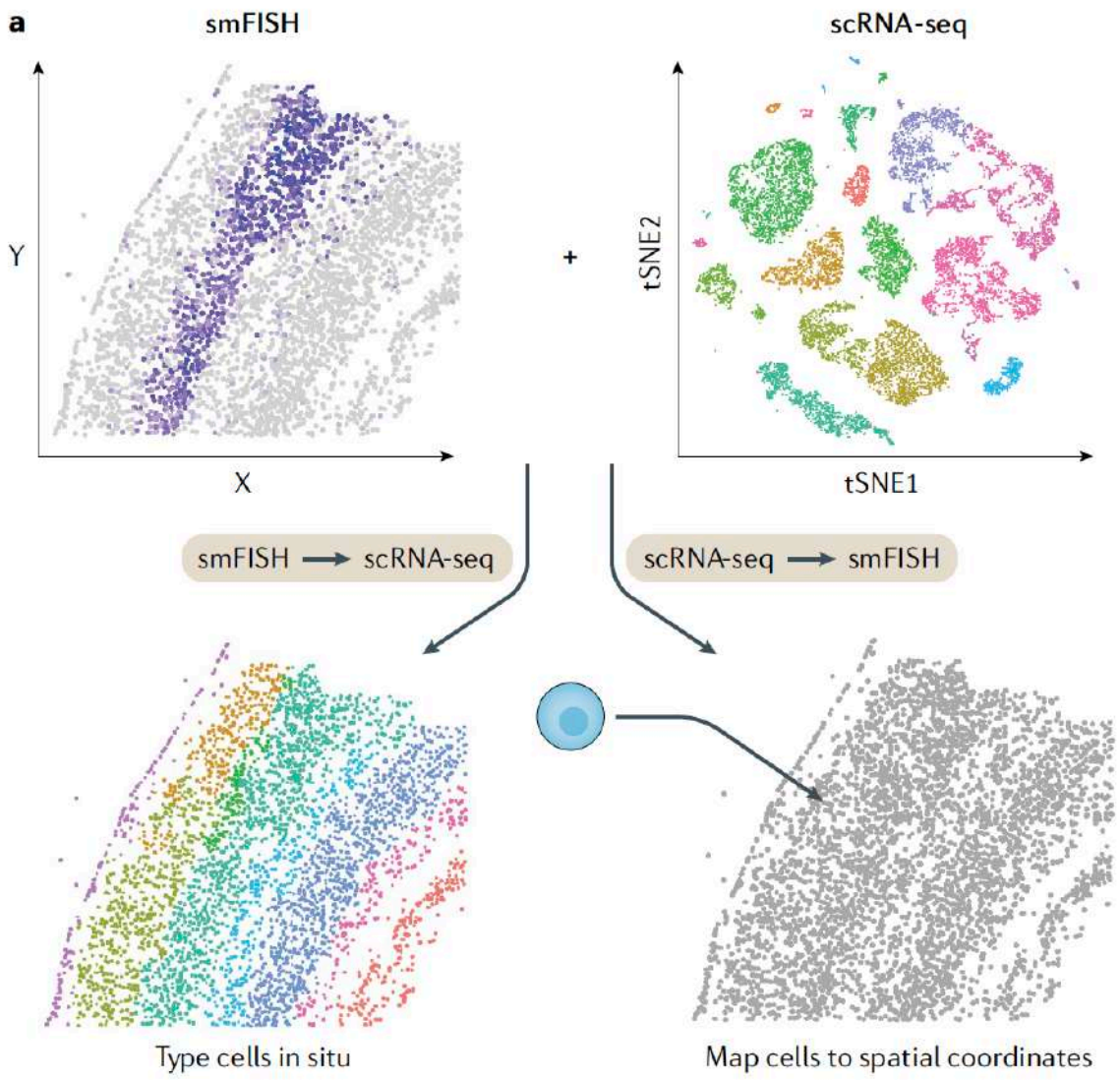
Stuart and Satija (2019) Nature Review Genetics

Multi-modal data can lead to better power at identifying cell states



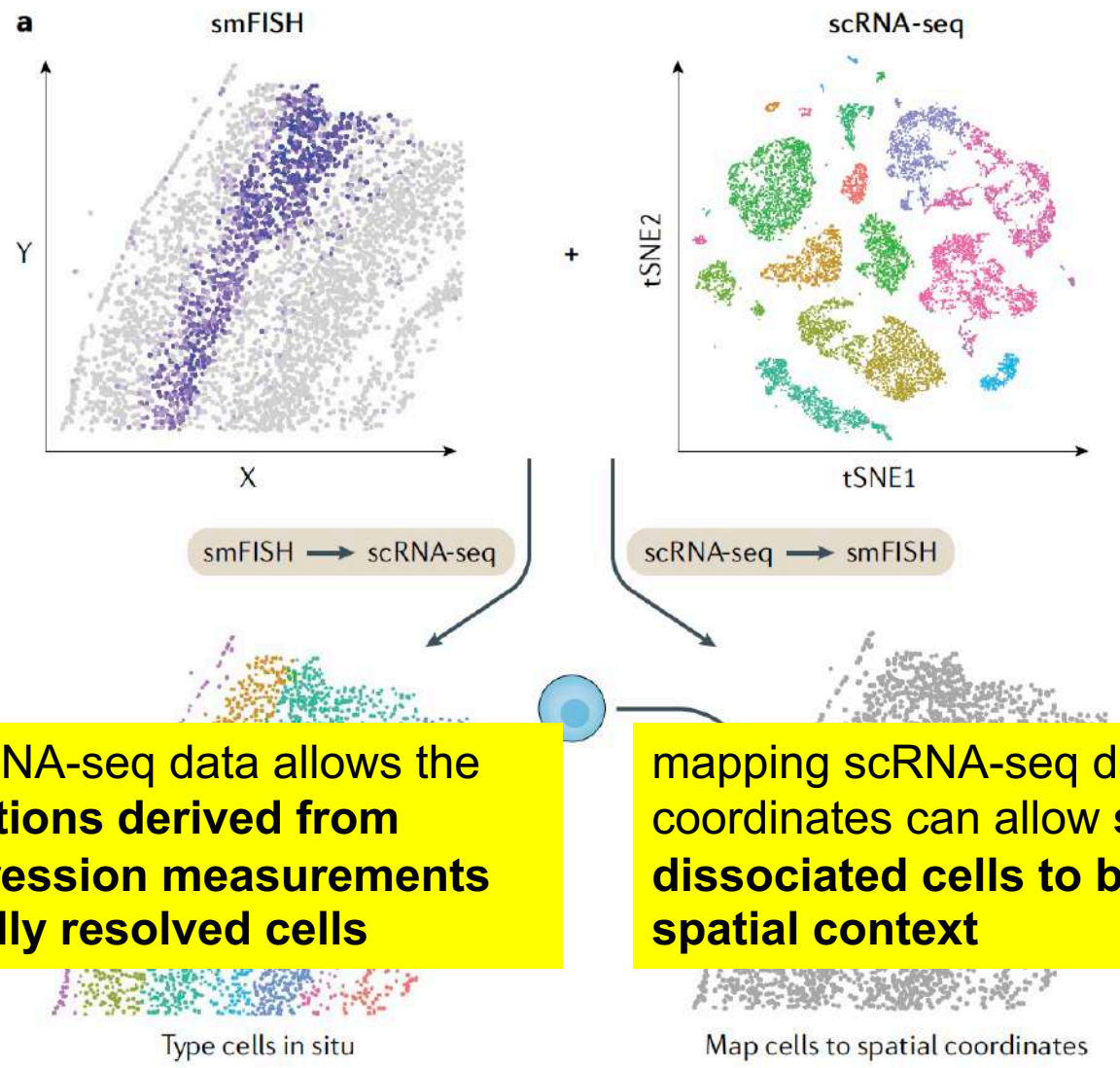
Stuart and Satija (2019) Nature Review Genetics

Spatial omics + scRNA-seq



Stuart and Satija (2019) Nature Review Genetics

Spatial omics + scRNA-seq

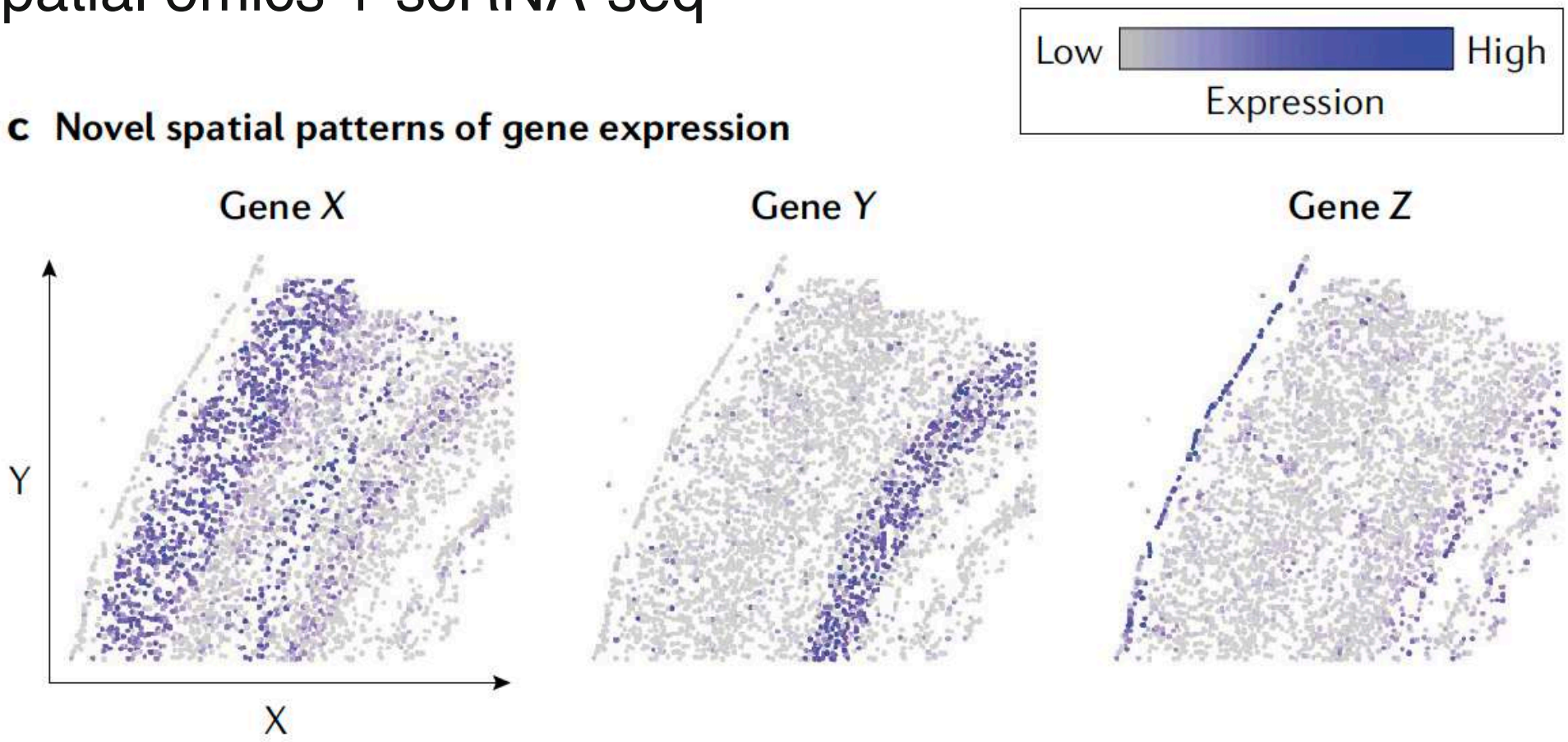


Mapping smFISH cells onto scRNA-seq data allows the **transfer of cell-type classifications derived from transcriptome-wide gene expression measurements to be transferred to the spatially resolved cells**

mapping scRNA-seq data onto smFISH-profiled spatial coordinates can allow **scRNA-seq data from dissociated cells to be placed back into their spatial context**

Stuart and Satija (2019) Nature Review Genetics

Spatial omics + scRNA-seq



By mapping scRNA-seq-profiled cells onto spatially resolved coordinates through the integration with smFISH data, **spatial patterns of gene expression can be predicted for any gene measured in the scRNA-seq data set**. Through these predictions, novel spatial patterns of gene expression may be identified through the analysis of genes that were not profiled by smFISH

Stuart and Satija (2019) Nature Review Genetics

Summary (II) and Conclusion

Potential

- Single cell/nucleus RNAseq + spatial information + long read sequencing + direct RNA sequencing?!
- It is an exciting time to be in

Challenges

- Data type gets extremely complicated
- Integrating different sources of data are powerful

References

https://www.notion.so/References-papers-links-in-start-learning-genomics-b7e57b28e9194bb29a02f483e0b894ad