

Introduction to transcriptome

Isheng Jason Tsai

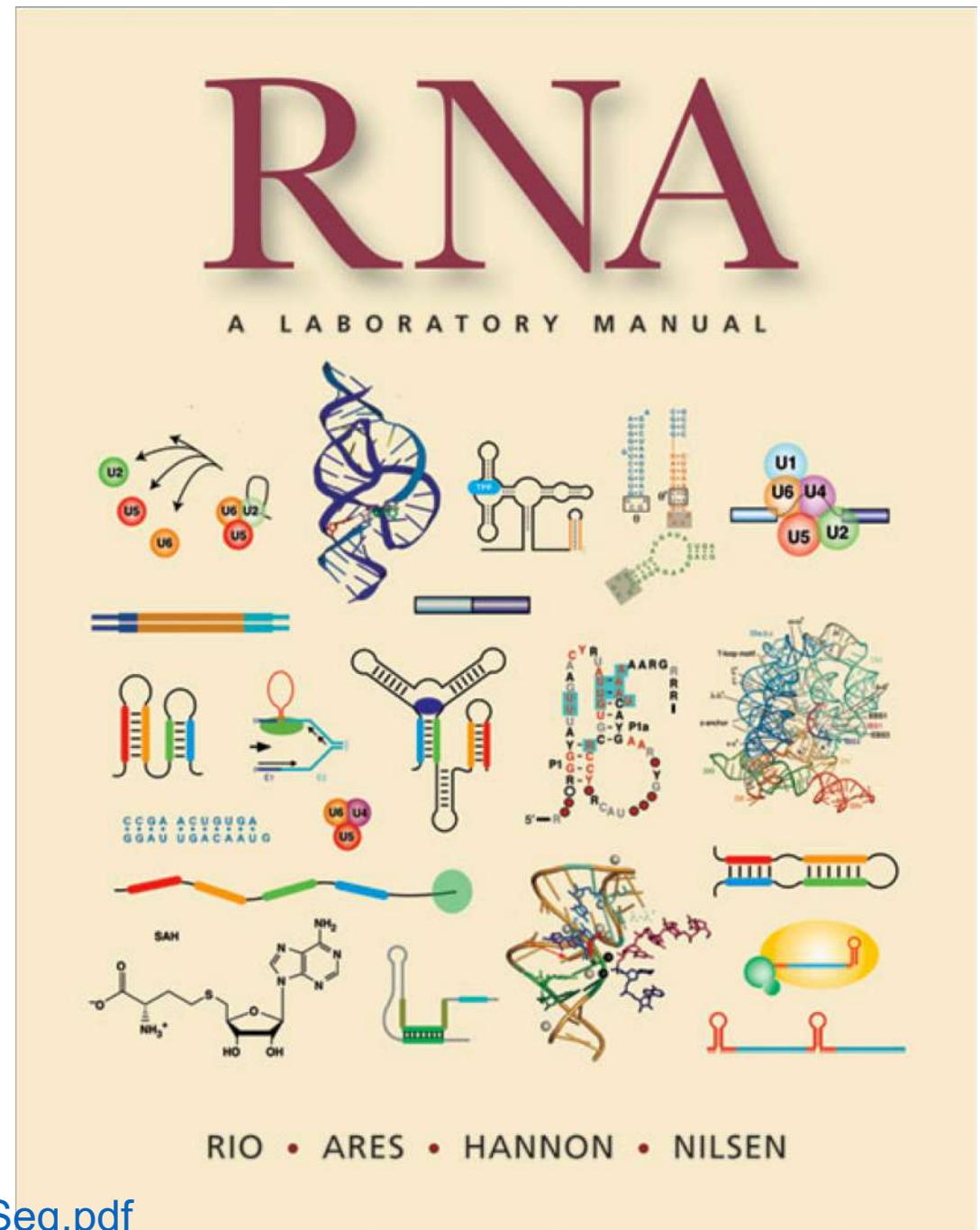
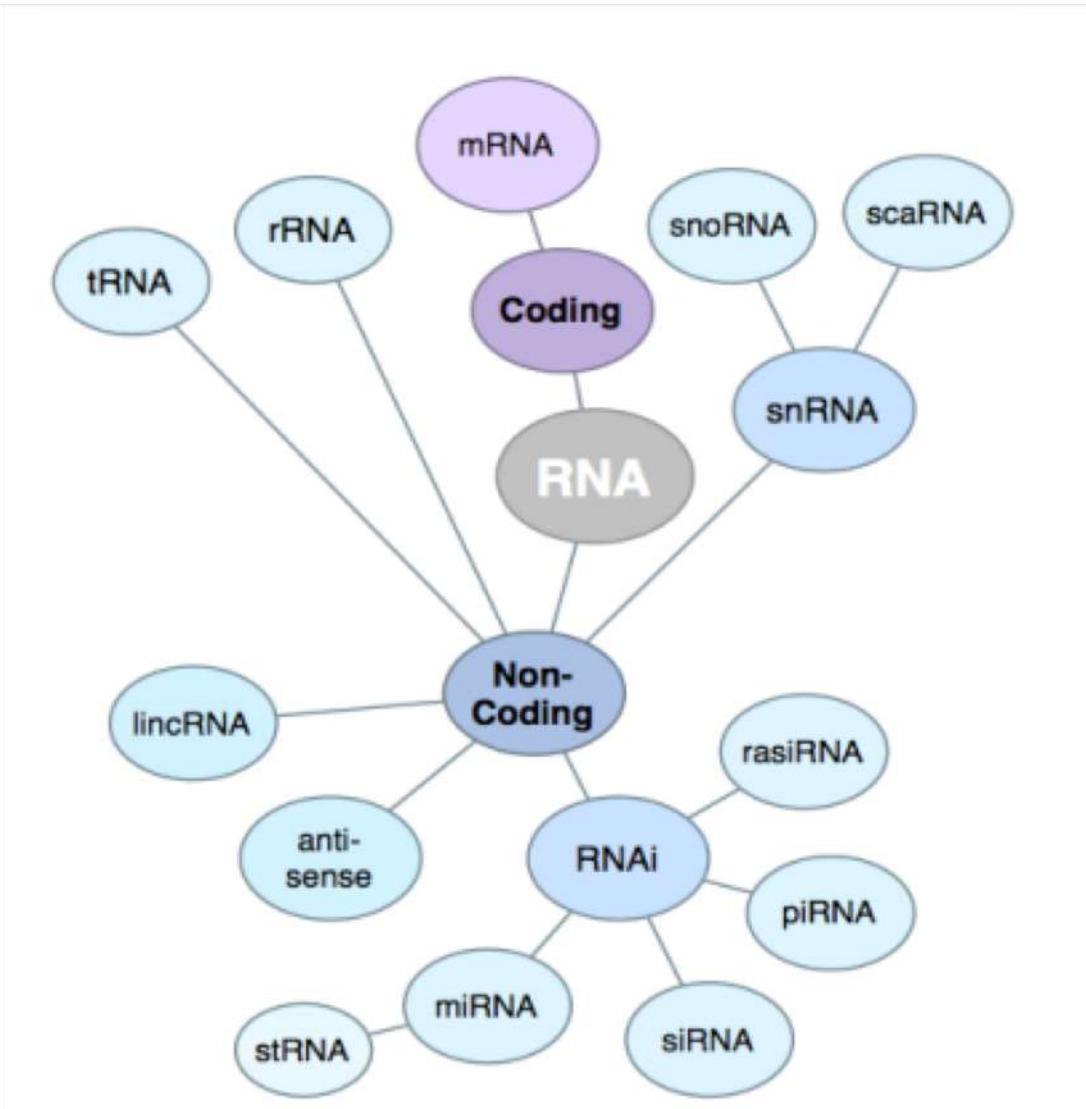
29th November 2018



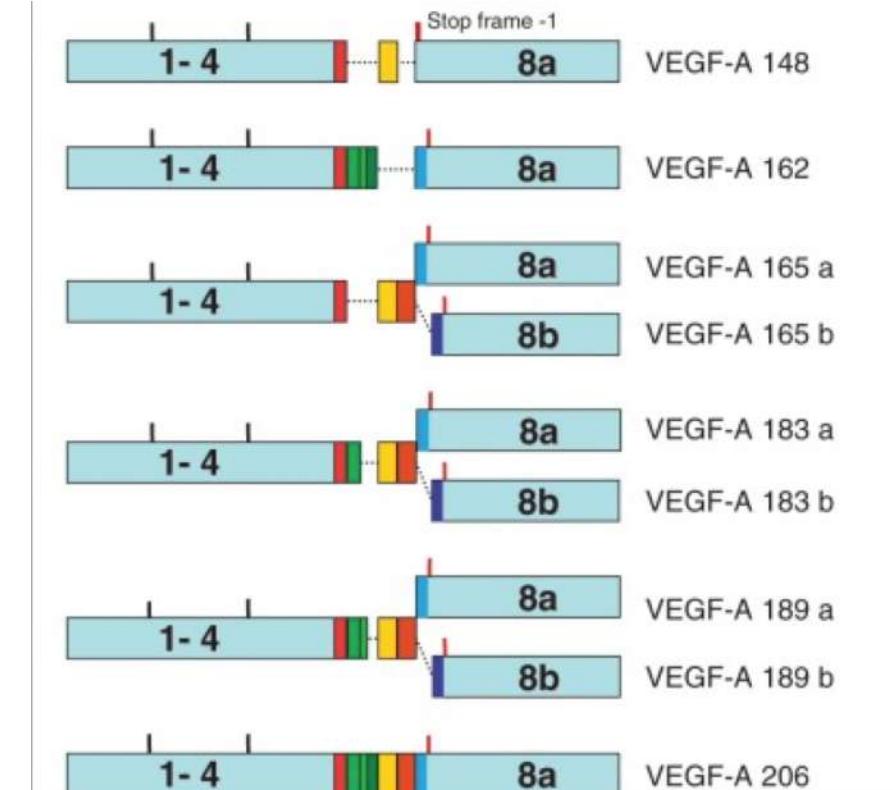
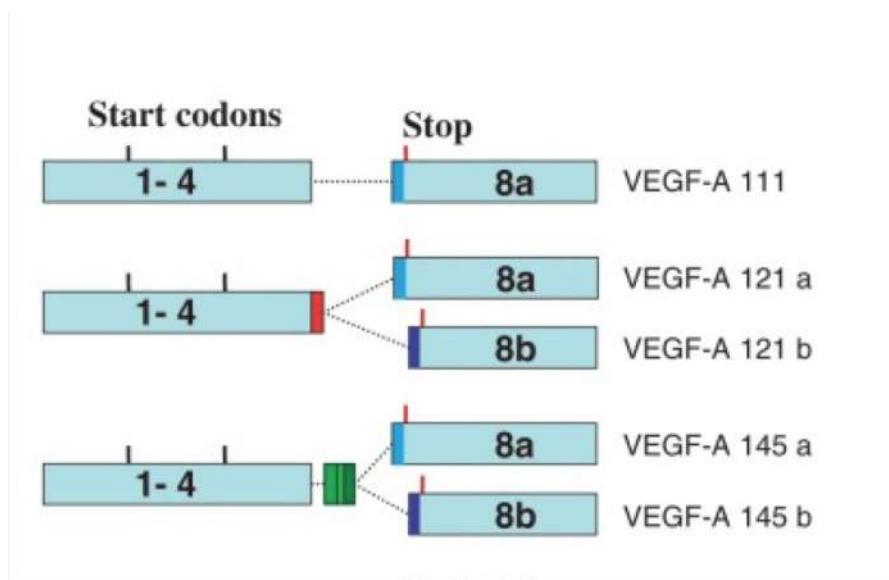
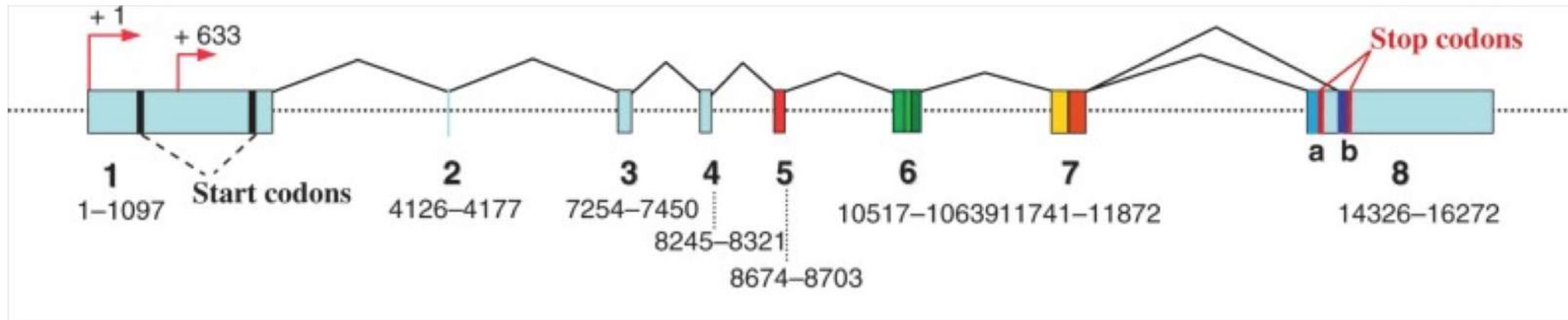
For this lecture, I gathered information from :

- **Introduction to differential gene expression analysis using RNA-seq**
<http://chagall.med.cornell.edu/RNASEQcourse/Intro2RNAseq.pdf>
(updated: October 8th 2018; highly recommended)
- **RNA sequencing data: hitchhiker's guide to expression analysis**
<https://peerj.com/preprints/27283/> (Nov 24th 2018!; also highly recommended)
- Reviews
- Slideshare (with permission)
 - Especially <http://www.slideshare.net/aubombarely/rnaseq-analysis-19910448>
- Twitter
- RNAseq blog <http://www.rna-seqblog.com/>
- <http://chagall.med.cornell.edu/RNASEQcourse>

Types of RNA



Gene and isoforms



REVIEW

Open Access



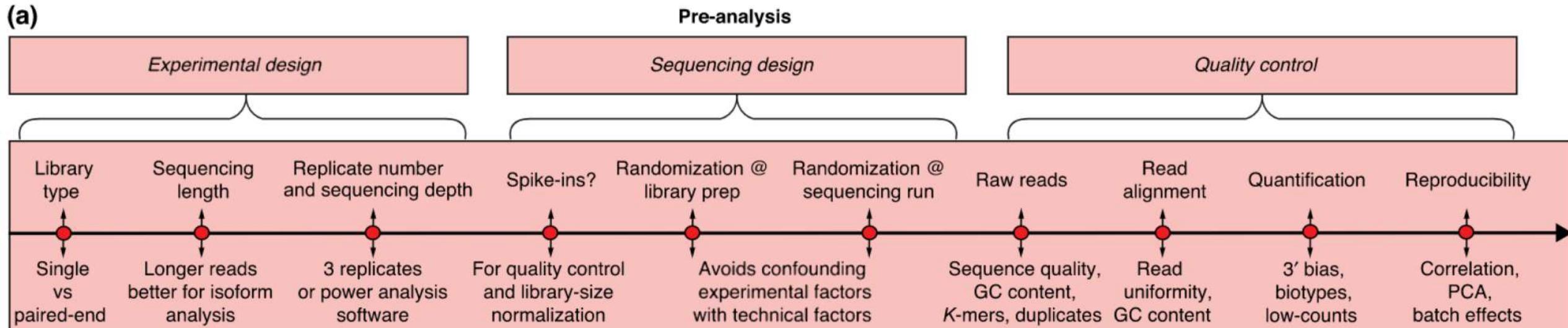
CrossMark

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 Cervera¹⁰, Andrew McPherson¹¹, Michał Wojciech Szcześniak¹², Daniel J. Gaffney³, Laura L. Elo¹³, Xuegong Zhang^{14,15} and Ali Mortazavi^{16,17*}

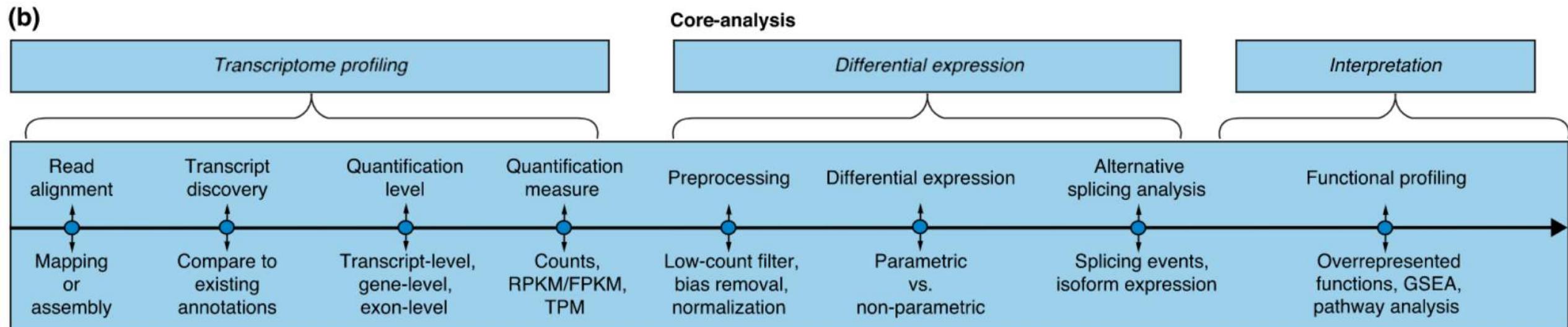
Pre-analysis

(a)



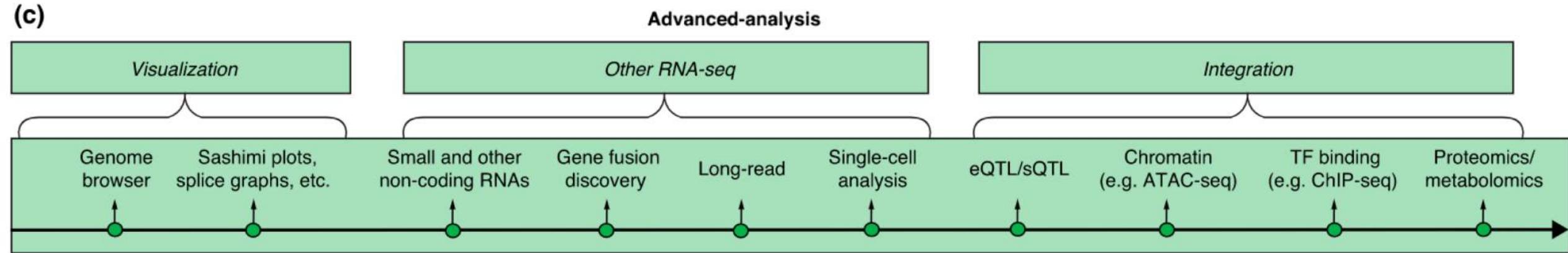
Core-analysis

(b)



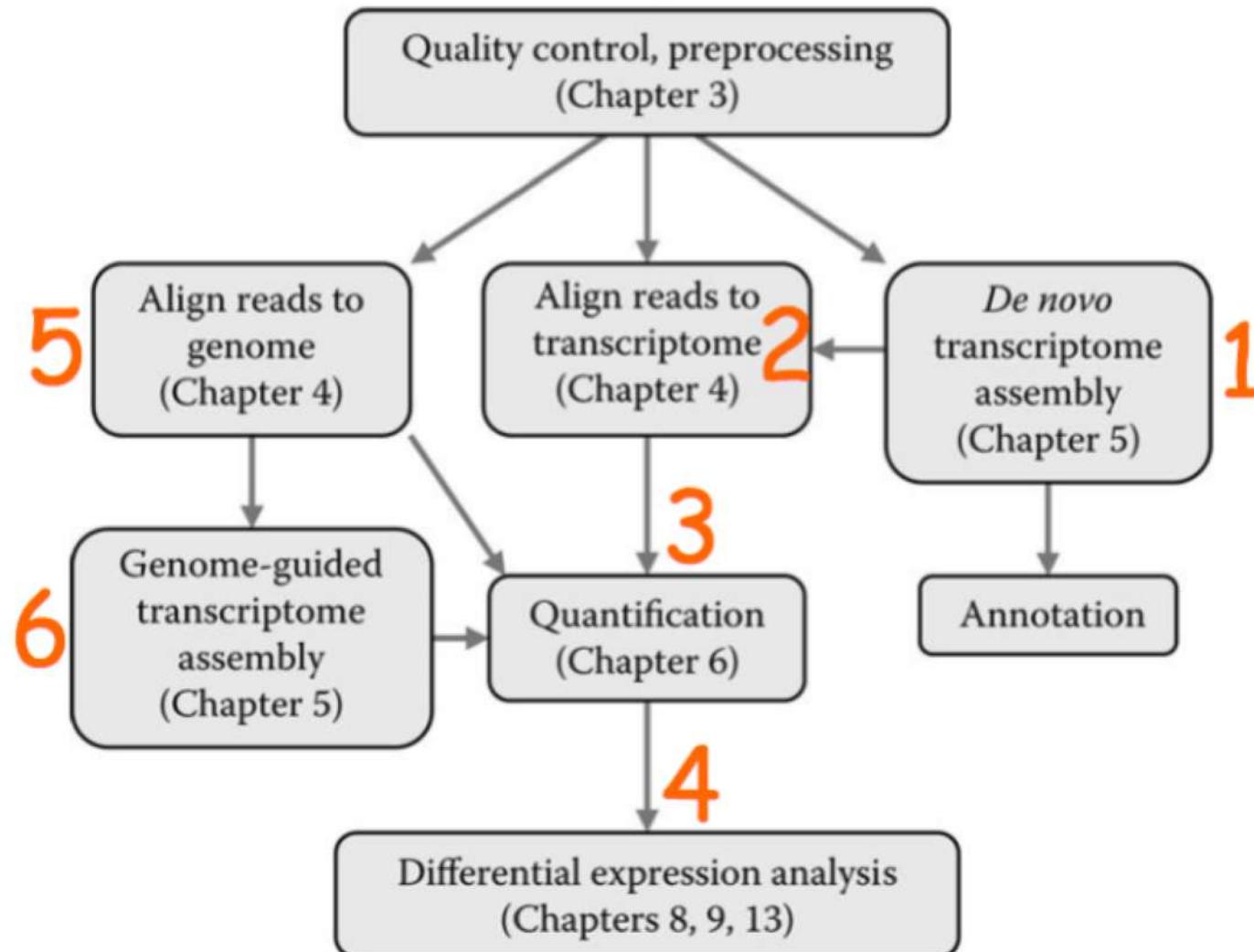
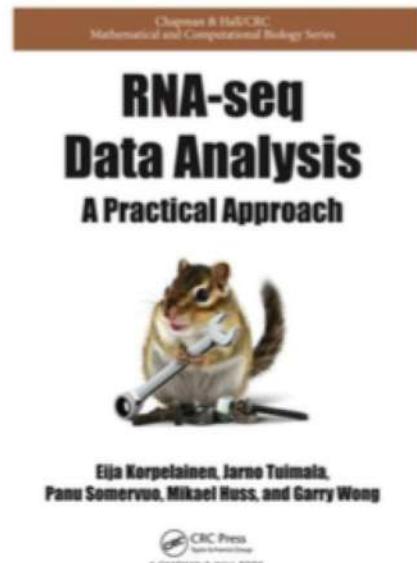
Advanced-analysis (not covered in this lecture but should be mentioned)

(c)



Expression quantification and transcript assembly

Obsolete



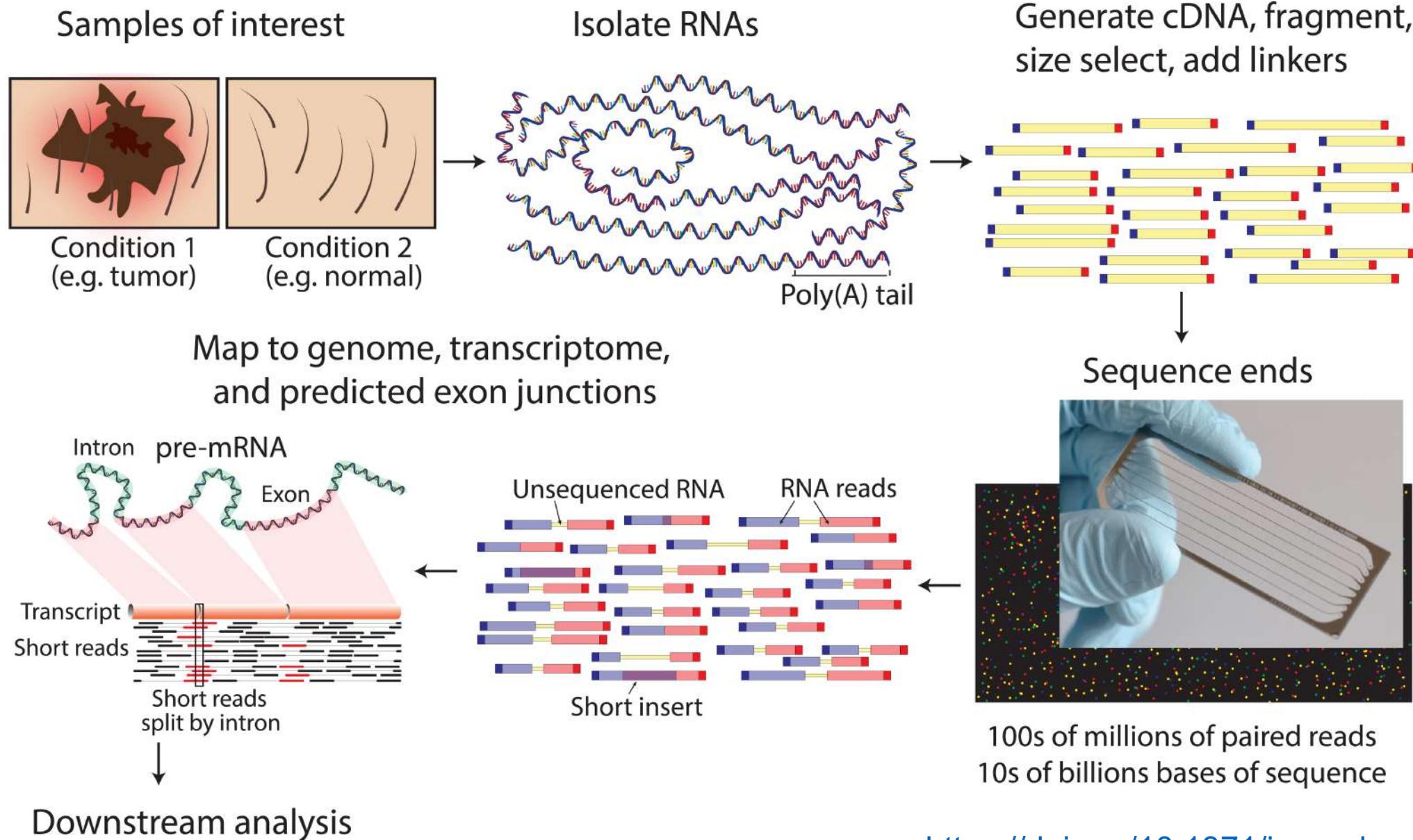
Introduction to differential gene expression analysis using RNA-seq

Written by Friederike Dündar, Luce Skrabaneck, Paul Zumbo

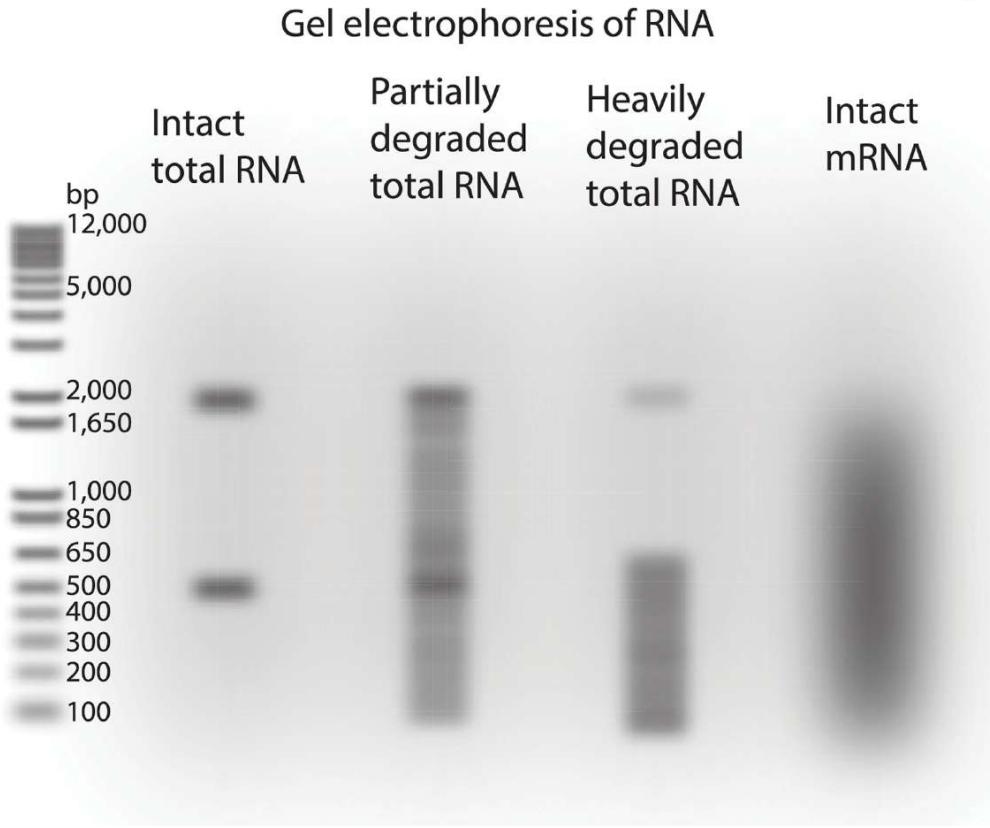
September 2015
updated October 8, 2018

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

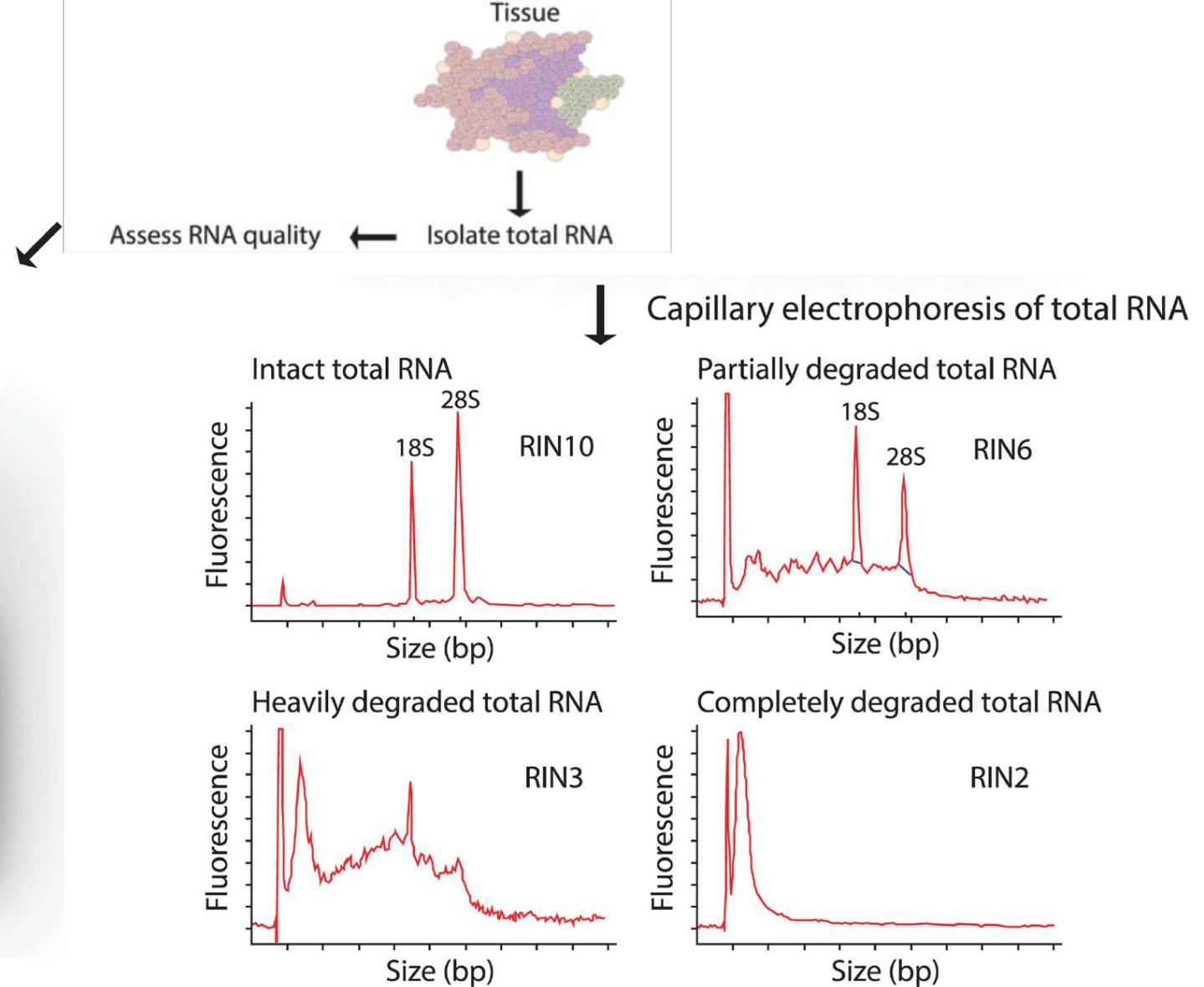
RNA-seq data generation



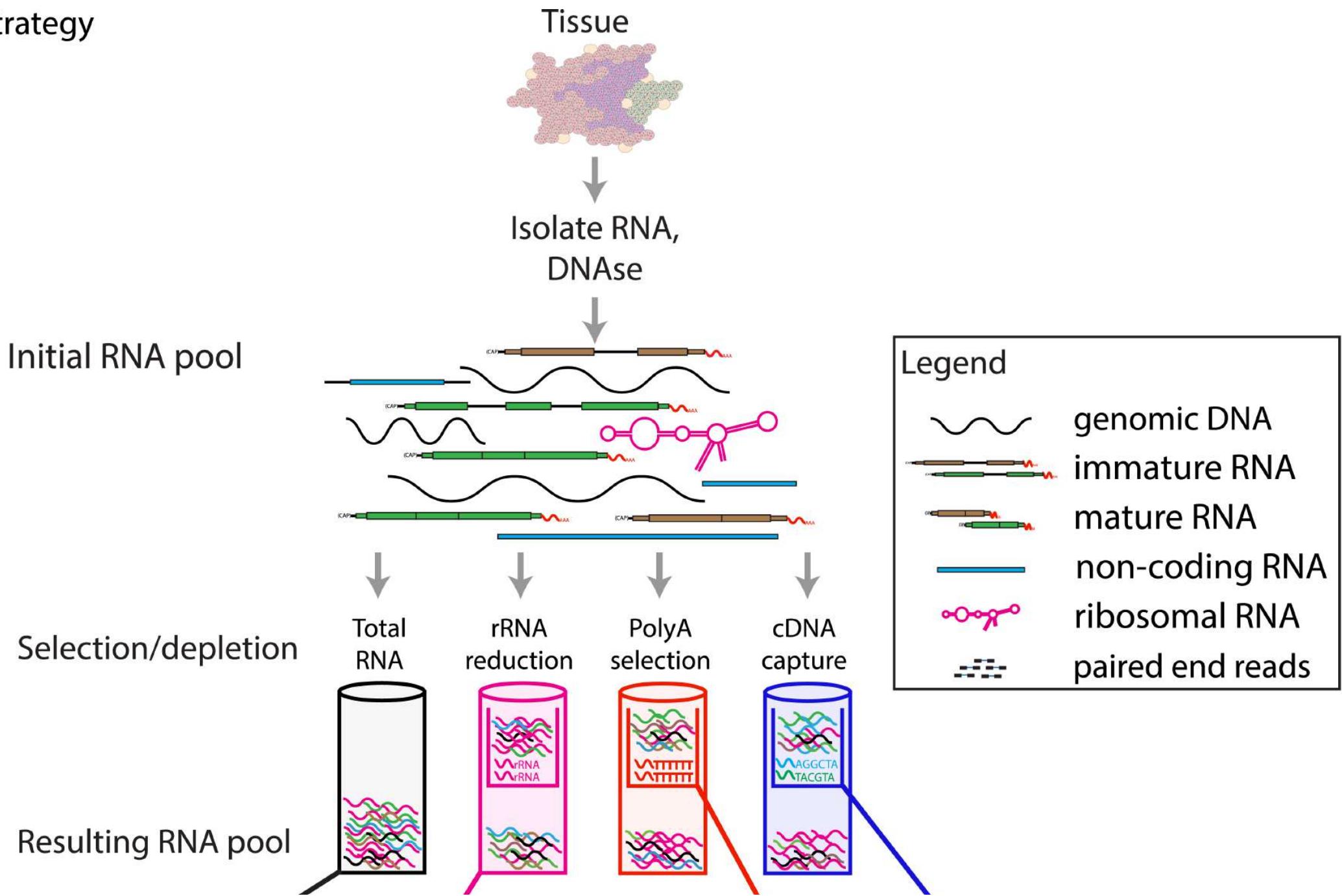
RNA-seq data generation



RIN = 28S:18S ratio



RNA-seq Strategy



A. Total RNA

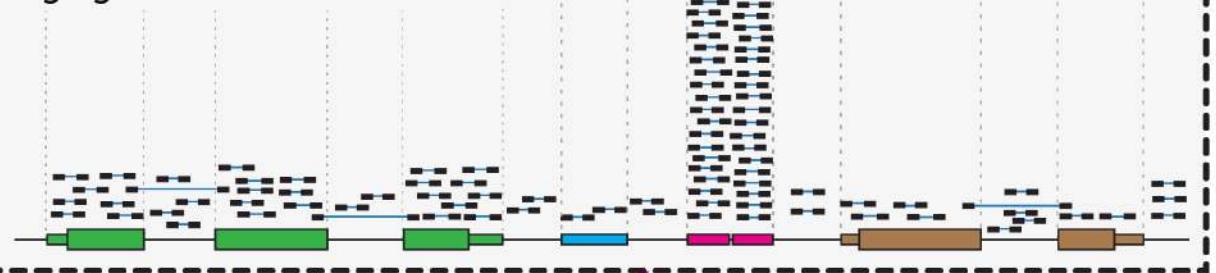
Broad transcript representation*

High rRNAs

Abundant mRNAs dominate

High unprocessed RNA

High genomic DNA



D. cDNA capture

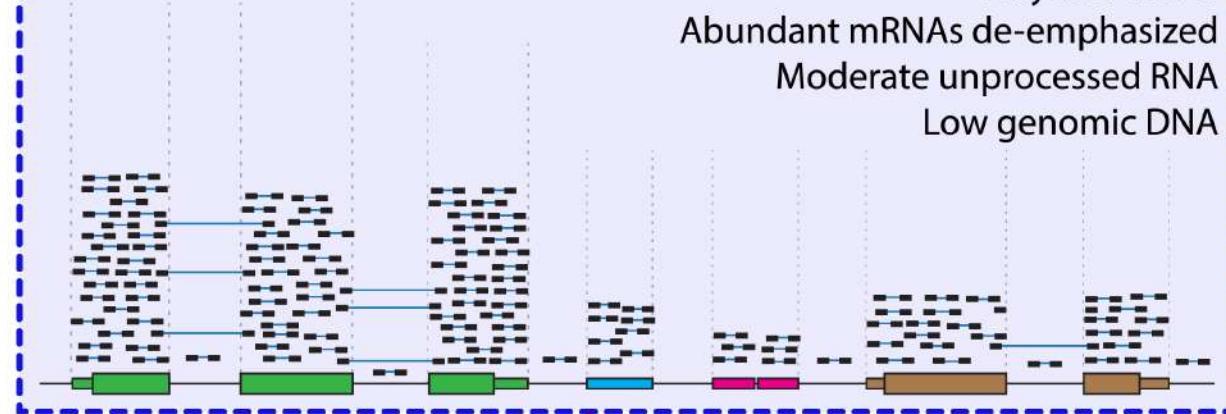
Limited transcript representation (targeted)

Very low rRNAs

Abundant mRNAs de-emphasized

Moderate unprocessed RNA

Low genomic DNA



B. rRNA reduction

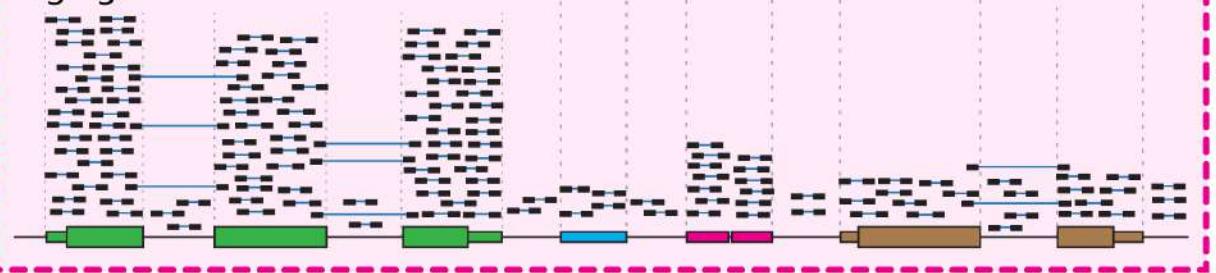
Broad transcript representation

Low rRNAs

Abundant mRNAs dominate

High unprocessed RNA

High genomic DNA



C. PolyA selection

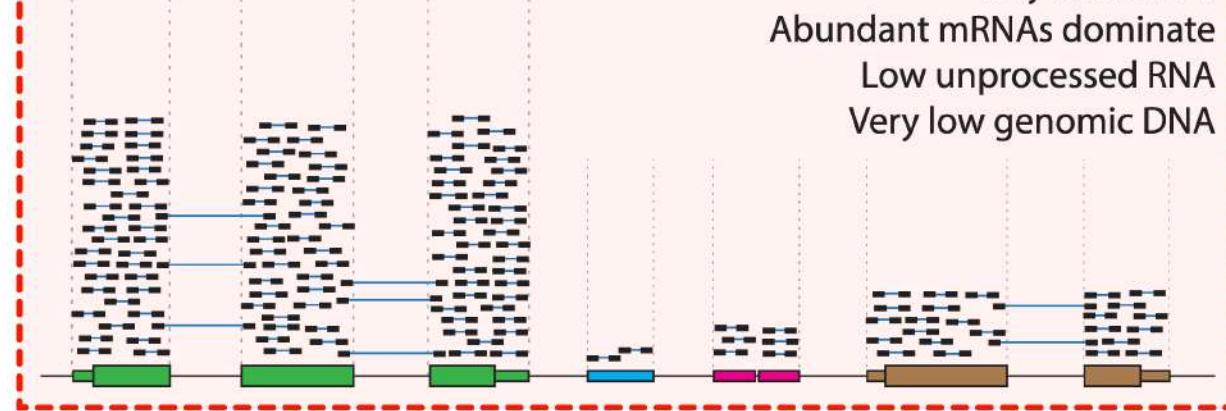
Limited transcript representation (polyA)

Very low rRNAs

Abundant mRNAs dominate

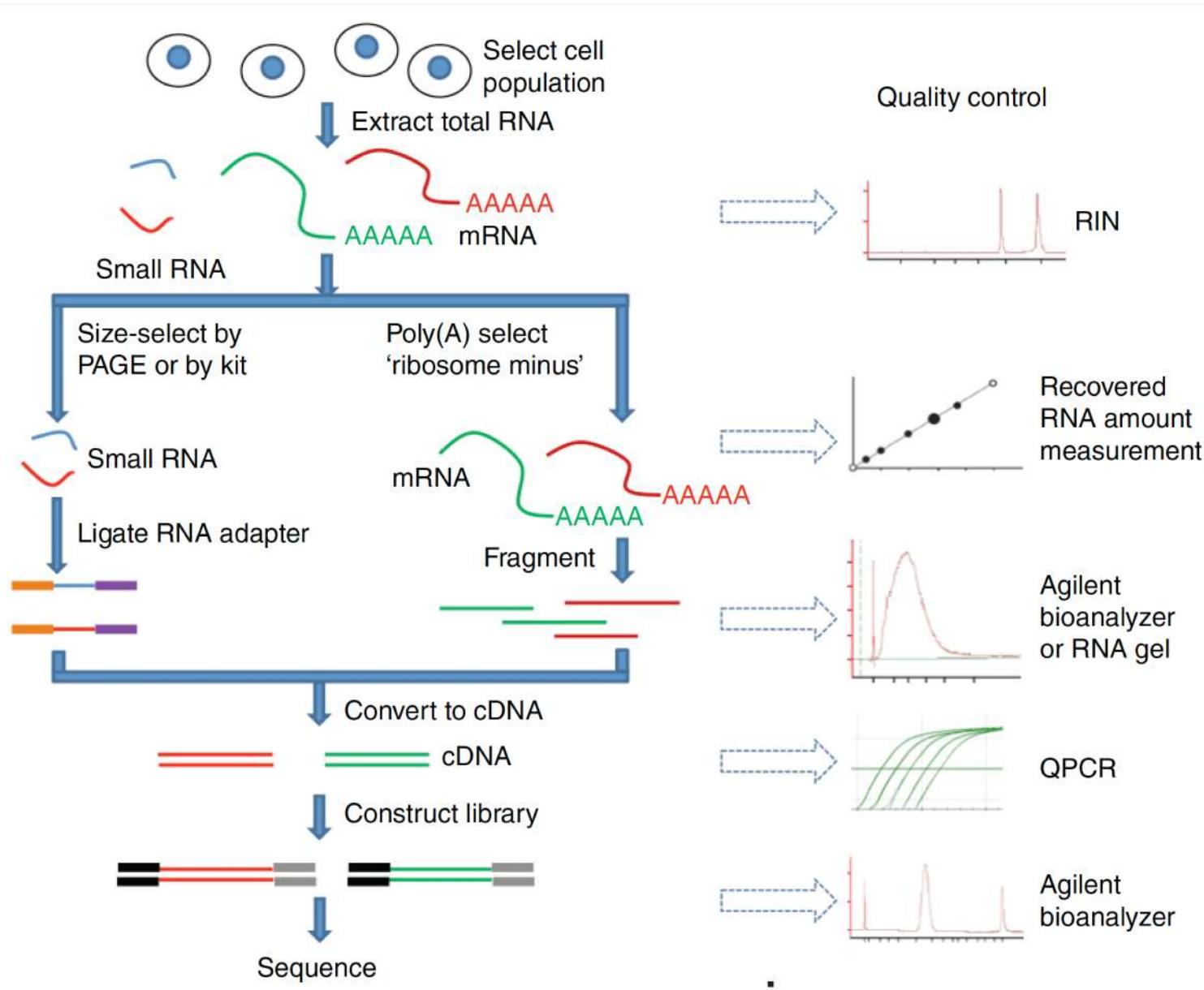
Low unprocessed RNA

Very low genomic DNA



Expected Alignments

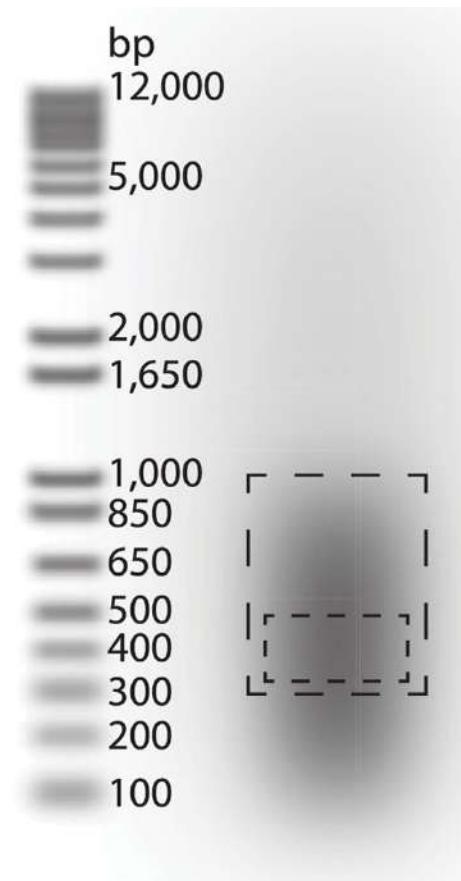
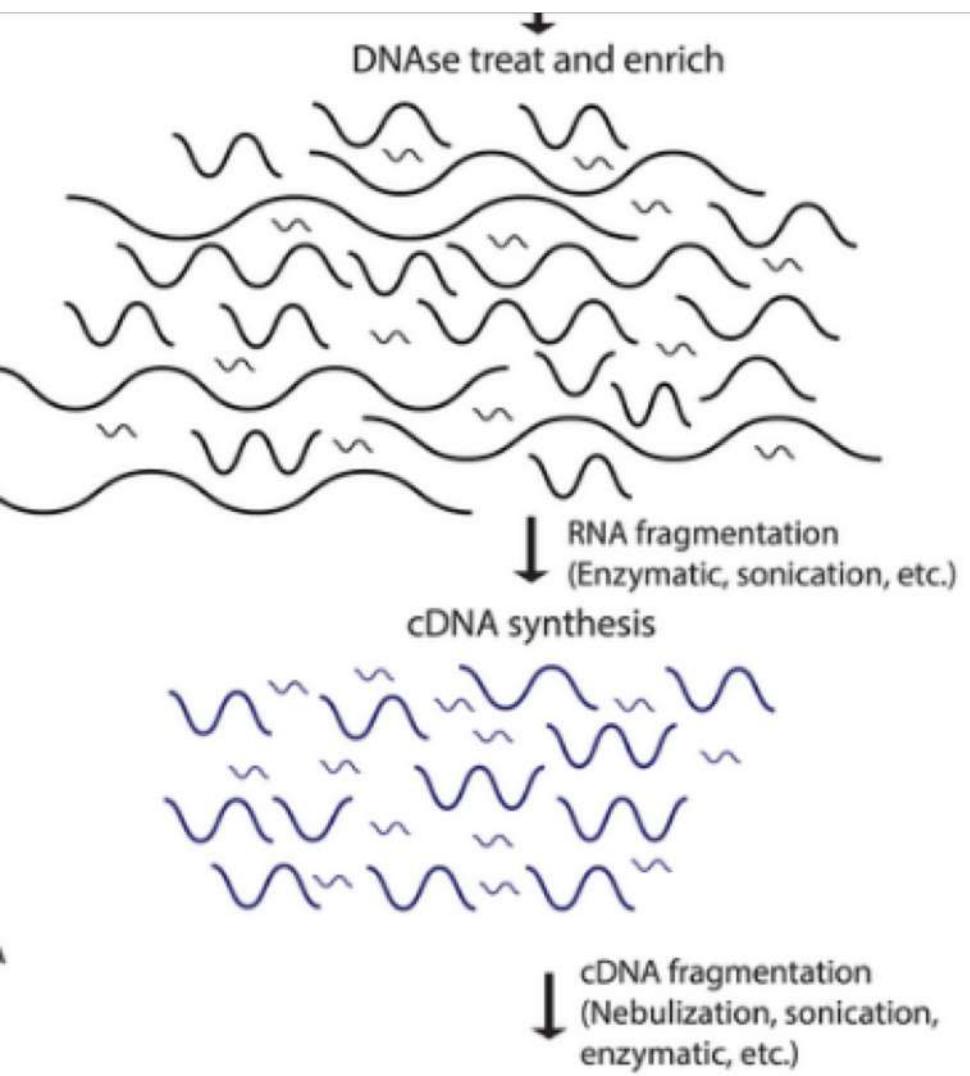
General RNA library preparation workflow



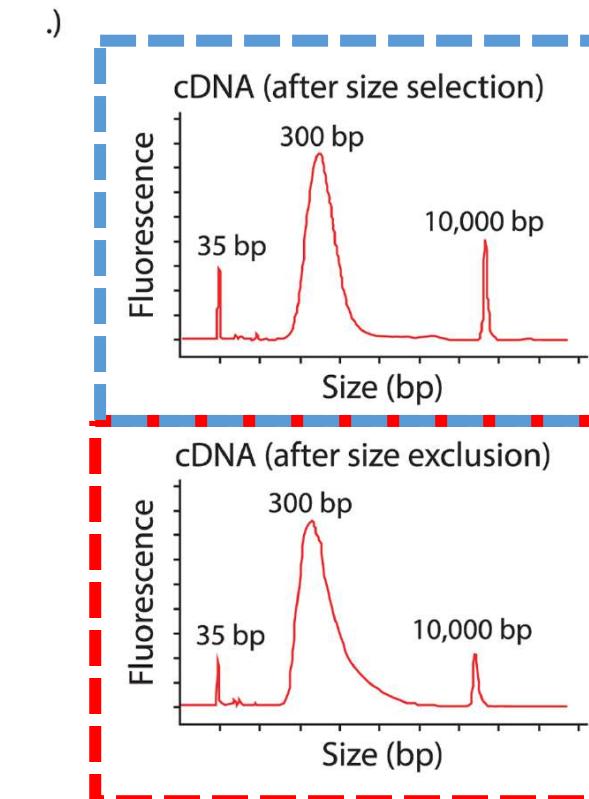
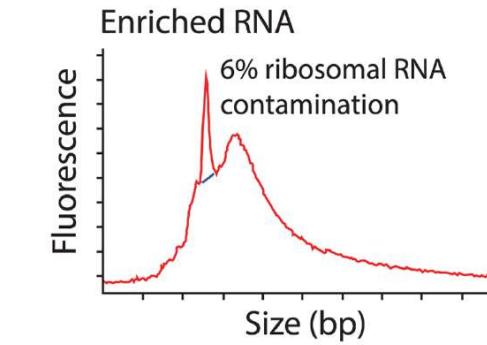
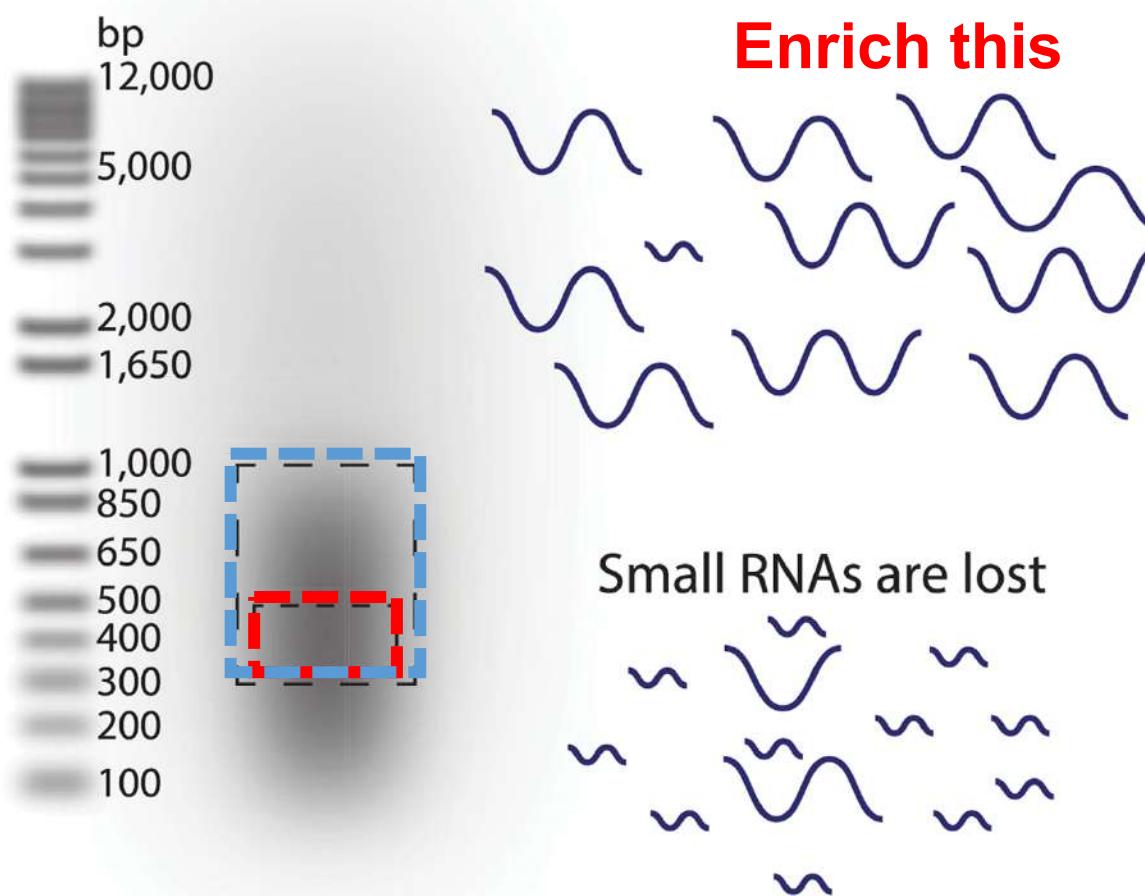
- i) RNA extraction and measuring its integrity,
- ii) rRNA is depleted (either using poly(A)-selection or rRNA depletion)
- iii) the remaining RNA molecules are fragmented, ideally achieving a uniform size distribution.
- iv) Double-stranded cDNA is synthesized and the adapters for sequencing are added to construct the final library whose fragment size distribution should be unimodal and well-defined.

RNA-seq data generation ; cDNA synthesis

Total RNA



RNA-seq data generation

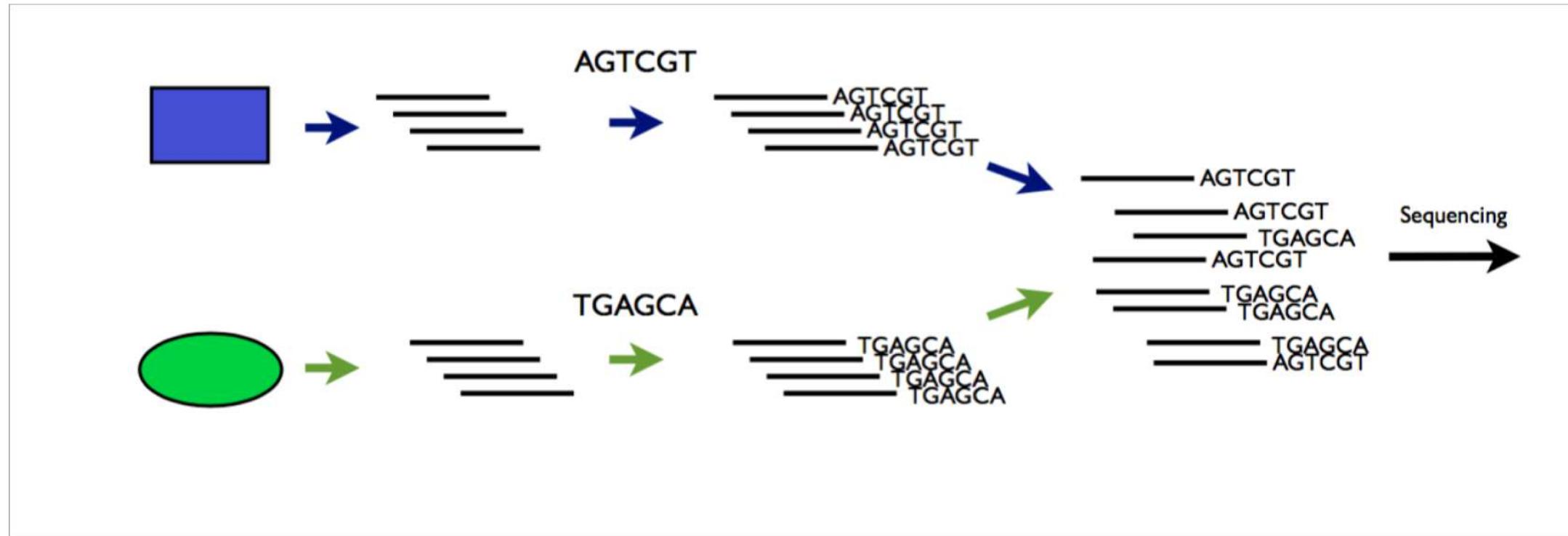


Column selection

Gel selection

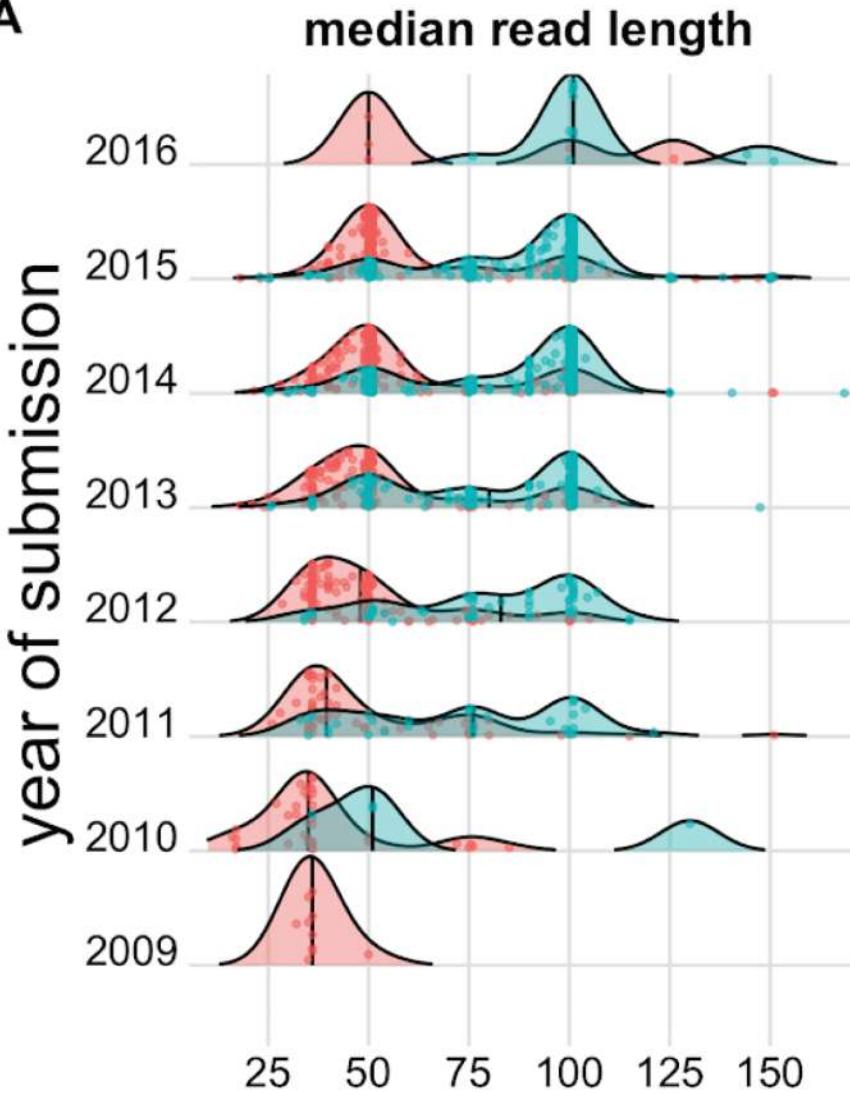
Enriched cDNA -> Added sequencing adaptors -> Sequencing

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

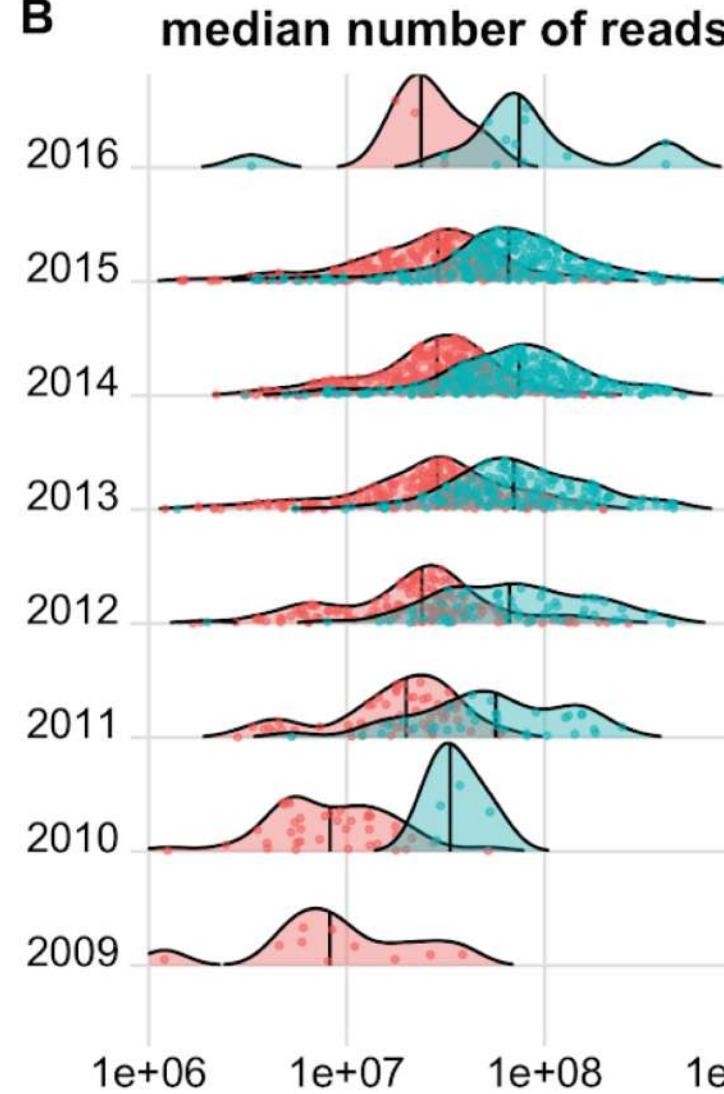


Evolution of RNAseq over time (from SRA)

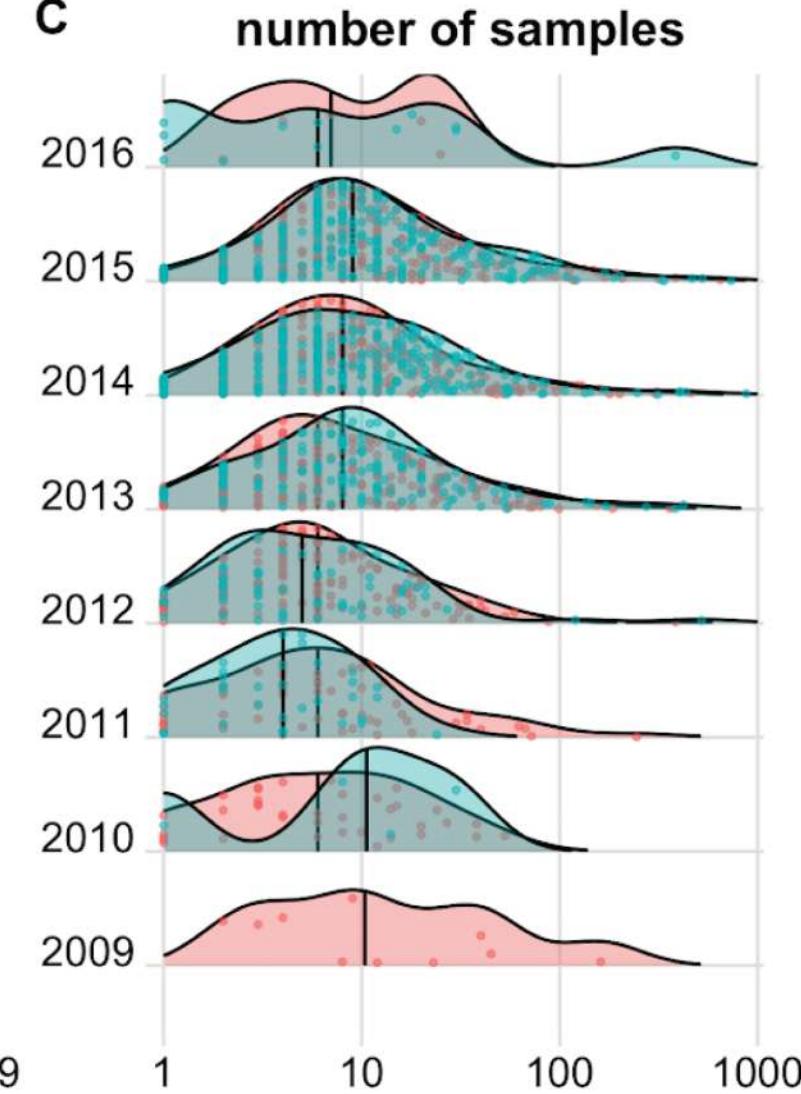
A



B



C



■ single-end (787) ■ paired-end (1008)

<https://doi.org/10.7287/peerj.preprints.27283v2>

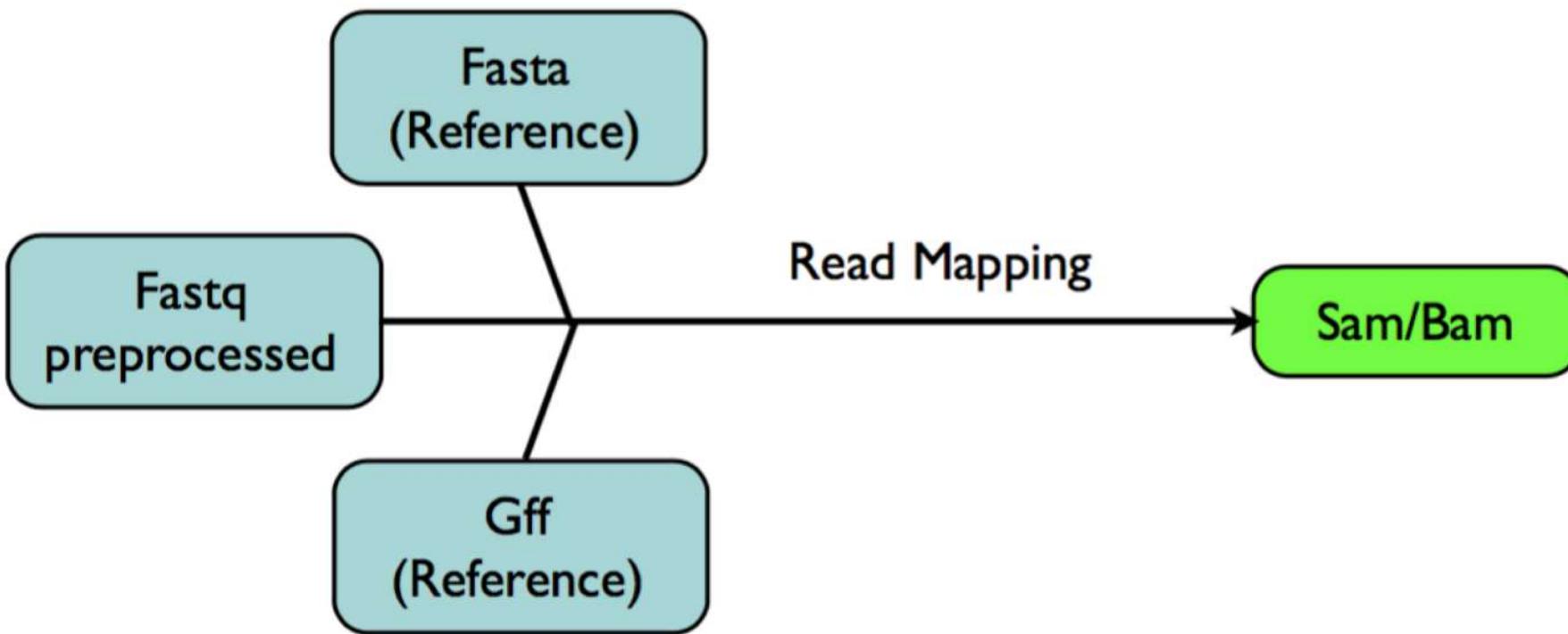
Fastq files:

FASTQ format is a **text-based format** for storing both a biological **sequence** (usually nucleotide sequence) and its corresponding **quality scores**.

-Wikipedia

```
@SEQUENCE_ID1
ATGCGCGCGCGCGCGCGGGTAGCAGATGACGACACAGAGCGAGGATGCGCTGAGAGTA
GTGTGACGACGATGACGGAAAATCAGA
+
BBBBBPPPPPXXXXX ^^^^^^————— _ ^—————^ _eeeeeee
[ [ [ [ [ ^ ^ ^ ] ] ] ] XXXXXPPPPPBBB
```

1. Single line ID with at symbol (“@”) in the first column.
 2. There should be not space between “@” symbol and the first letter of the identifier.
 3. Sequences are in multiple lines after the ID line
 4. Single line with plus symbol (“+”) in the first column to represent the quality line.
 5. Quality ID line can have or have not ID
 6. Quality values are in multiple lines after the + line



Note about the hardware and mapping software:

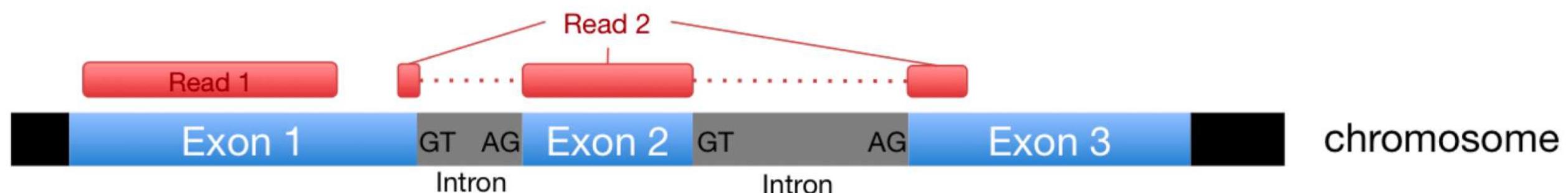
- + Bigger is the reference, more memory the programs needs
(example: Bowtie2 ~2.1 Gb for human genome with 3 Gb)
- + Longer are the reads, more time the program needs for the mapping.

Read mapping and transcript identification strategies

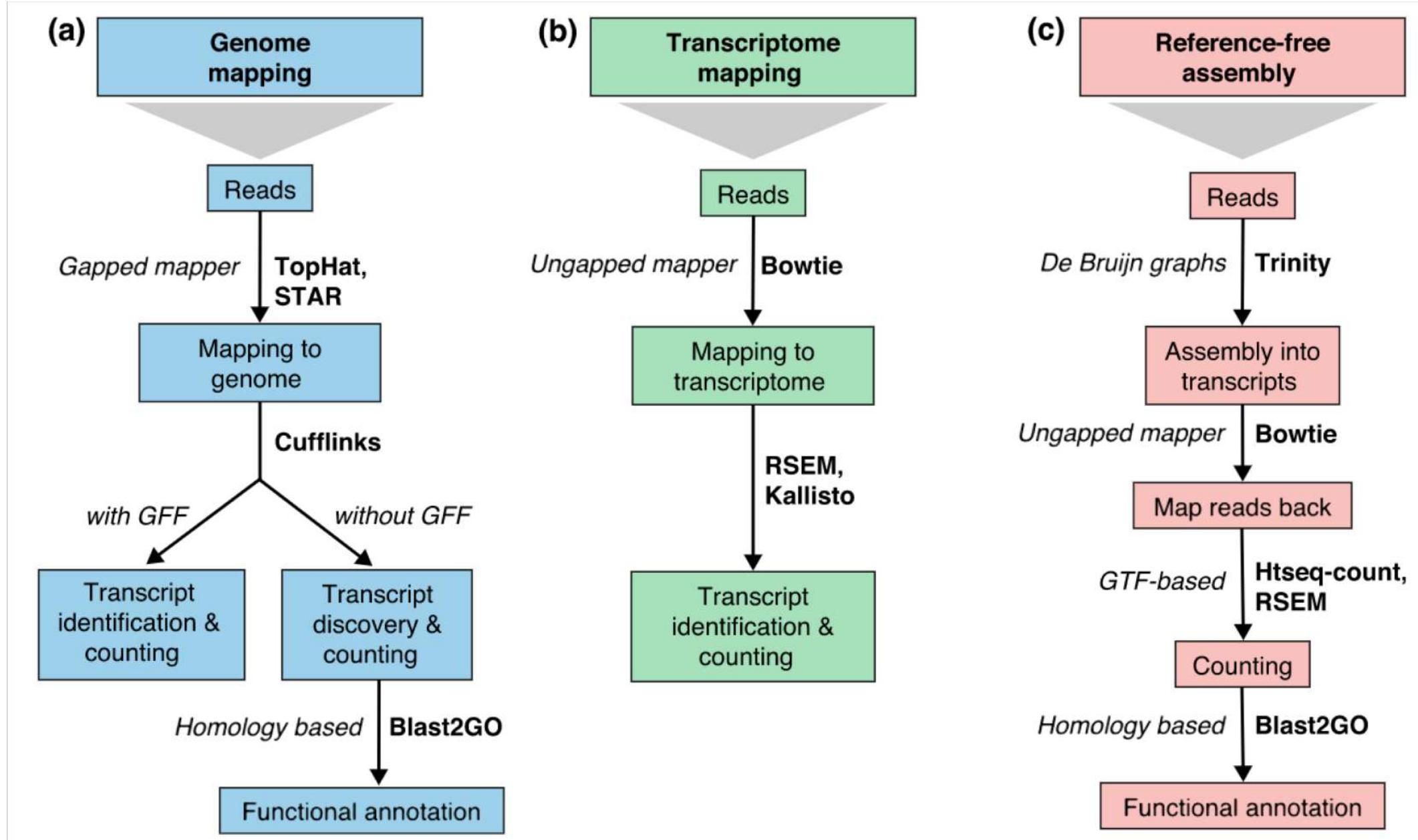
(a) Aligning to the transcriptome



(b) Aligning to the genome



Read mapping and transcript identification strategies

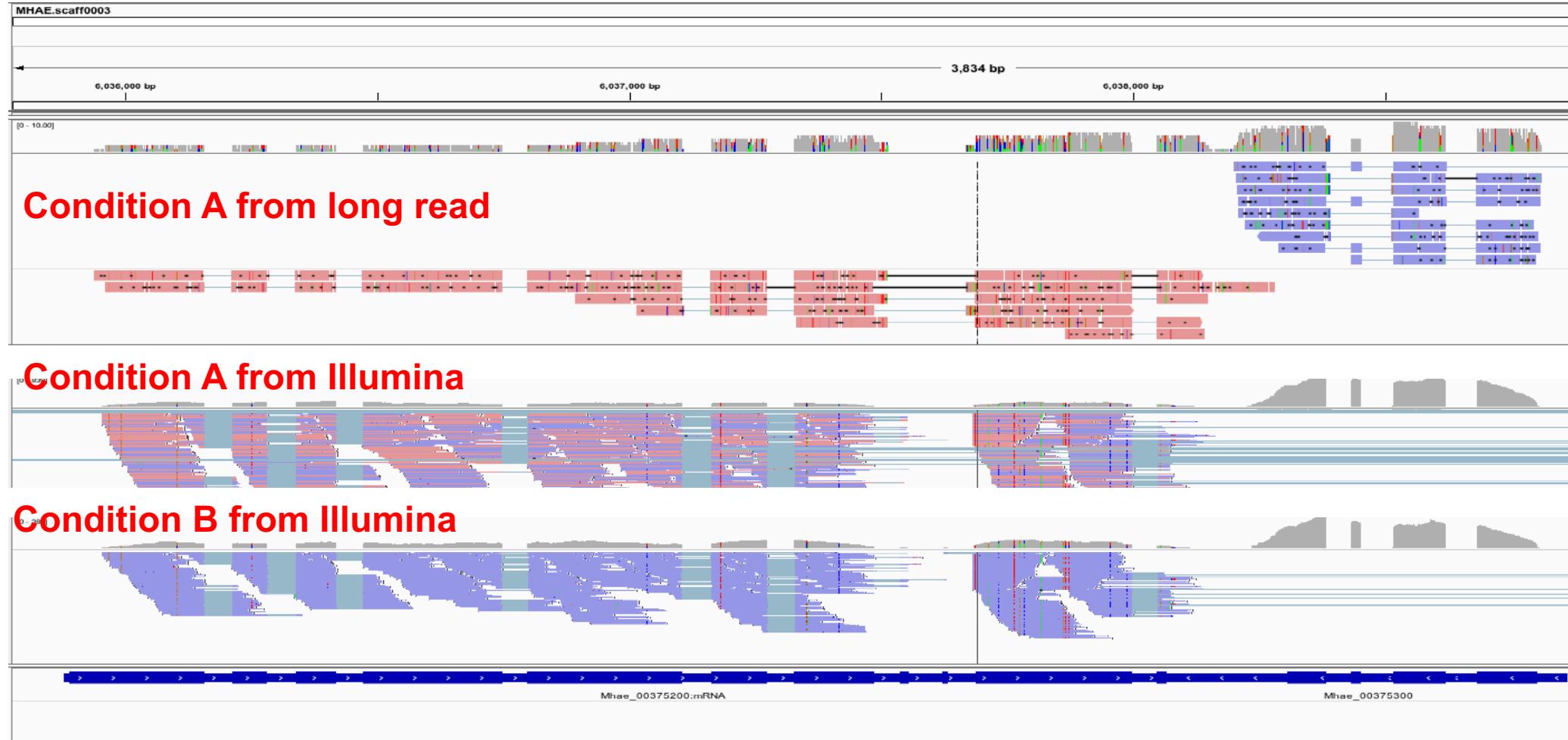


Read visualisation

Load reference, annotation and bam into a program

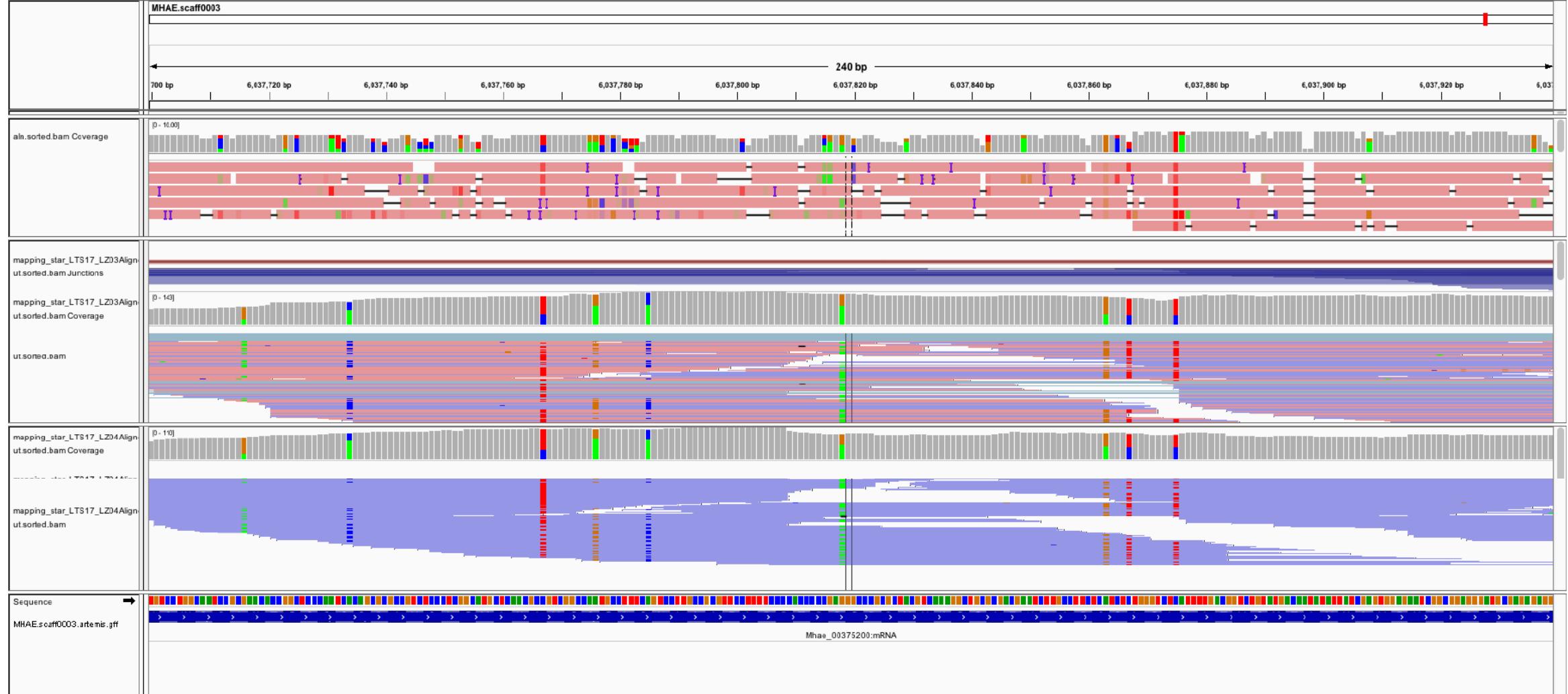
- Artemis <http://www.sanger.ac.uk/science/tools/artemis>
- IGV <http://software.broadinstitute.org/software/igv/>

Scenario 1



Annotation: two genes of two orientations

Long reads have more errors

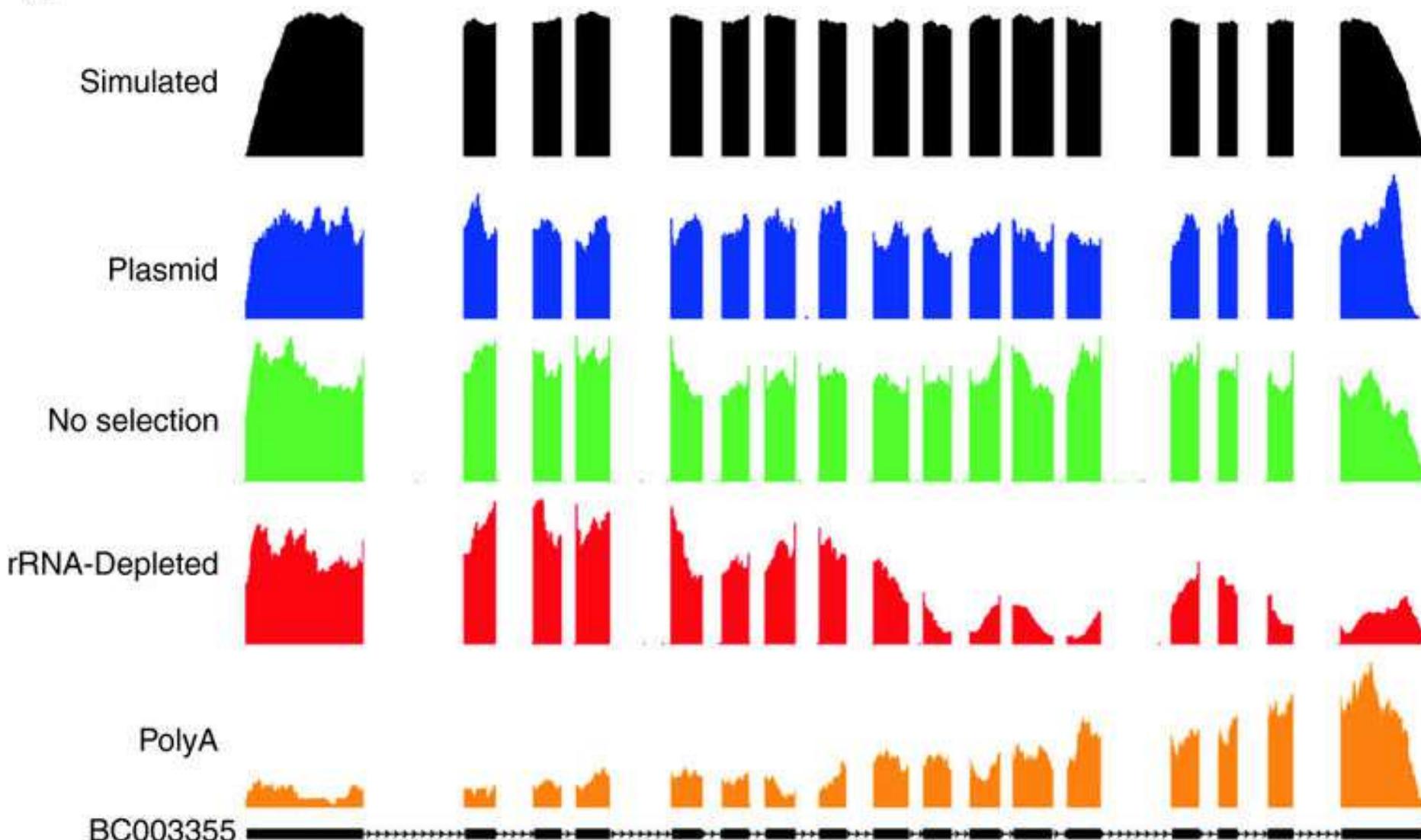


Wrong annotation (wrong gene fusion / wrong exons)



Library enrichment result in sequencing bias

A



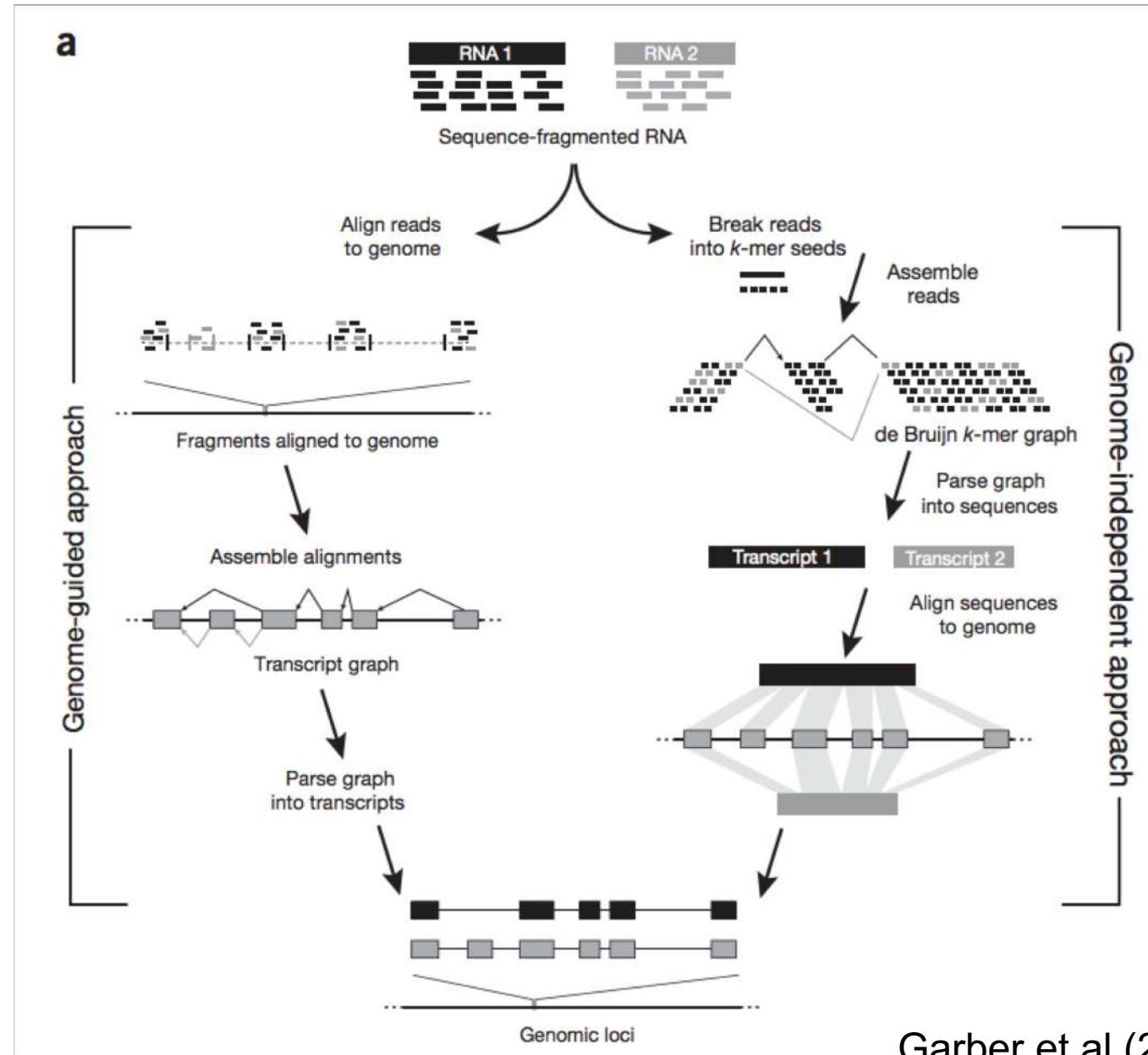
Mapping

General workflow for RNAseq to produce annotation

Options:

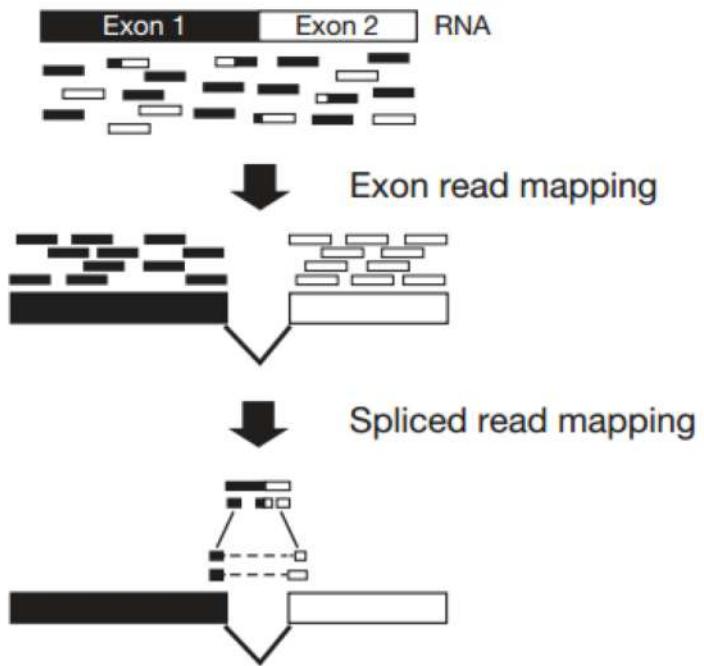
- Align and then assemble
- Assemble and then align

Align to
Genome
Transcriptome (if no genome)

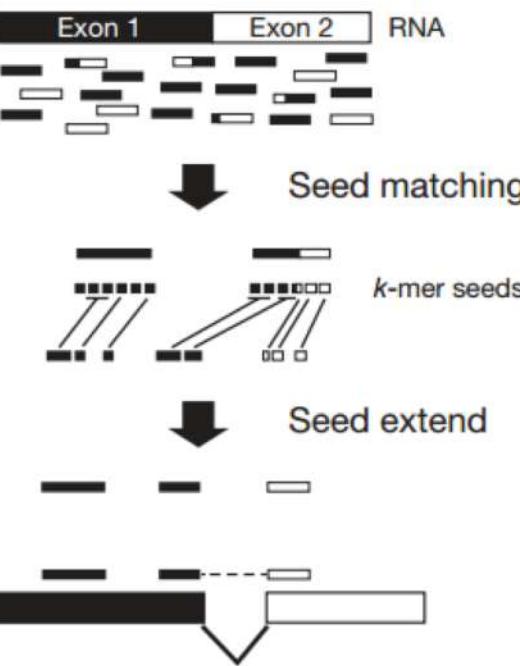


Strategies for gapped alignments of RNAseq reads

a Exon-first approach

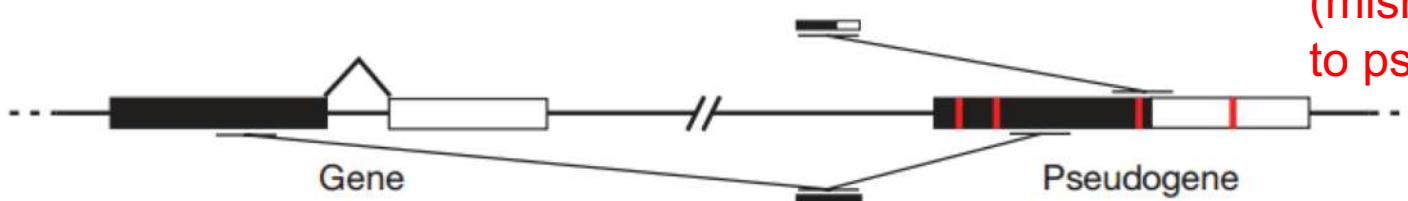


b Seed-extend approach

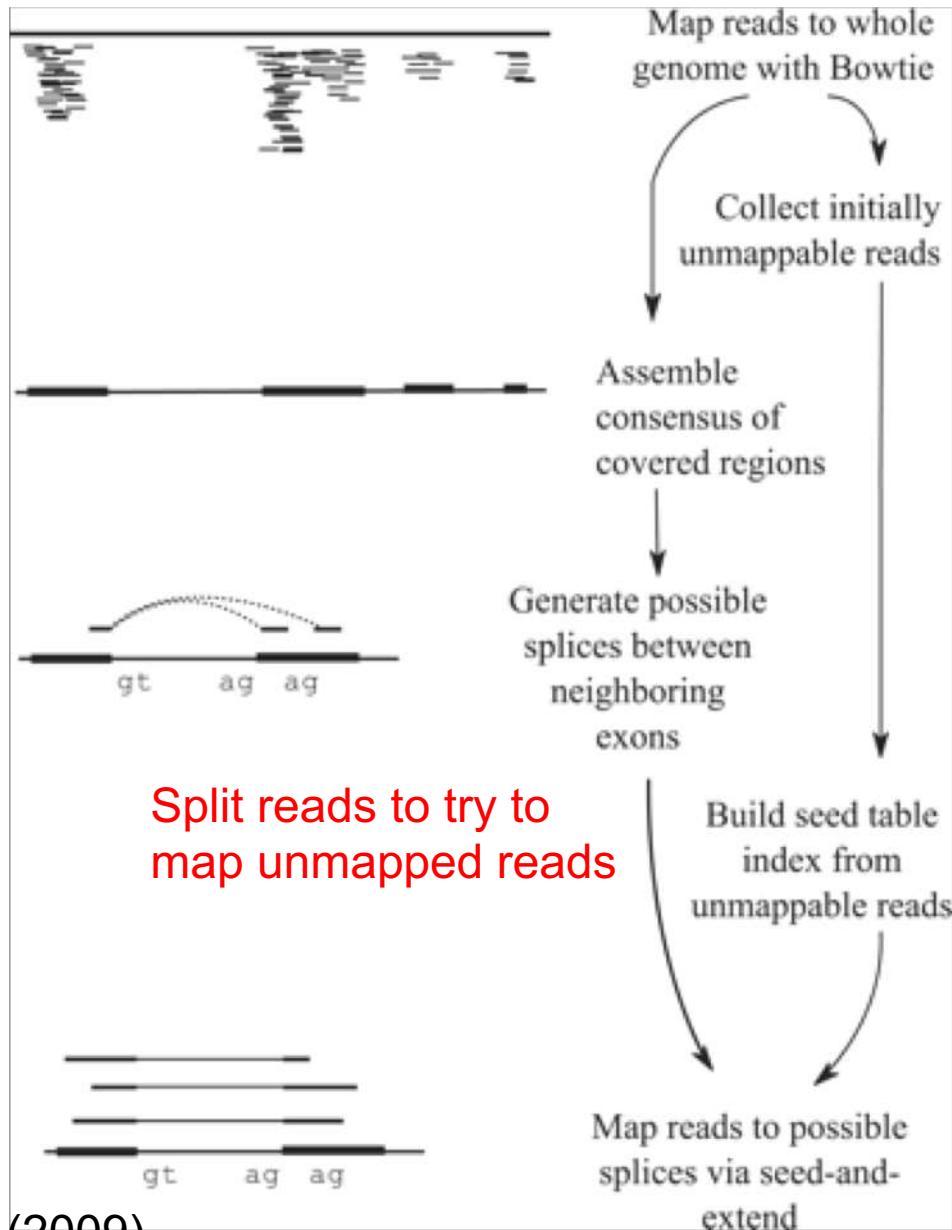


c Potential limitations of exon-first approaches

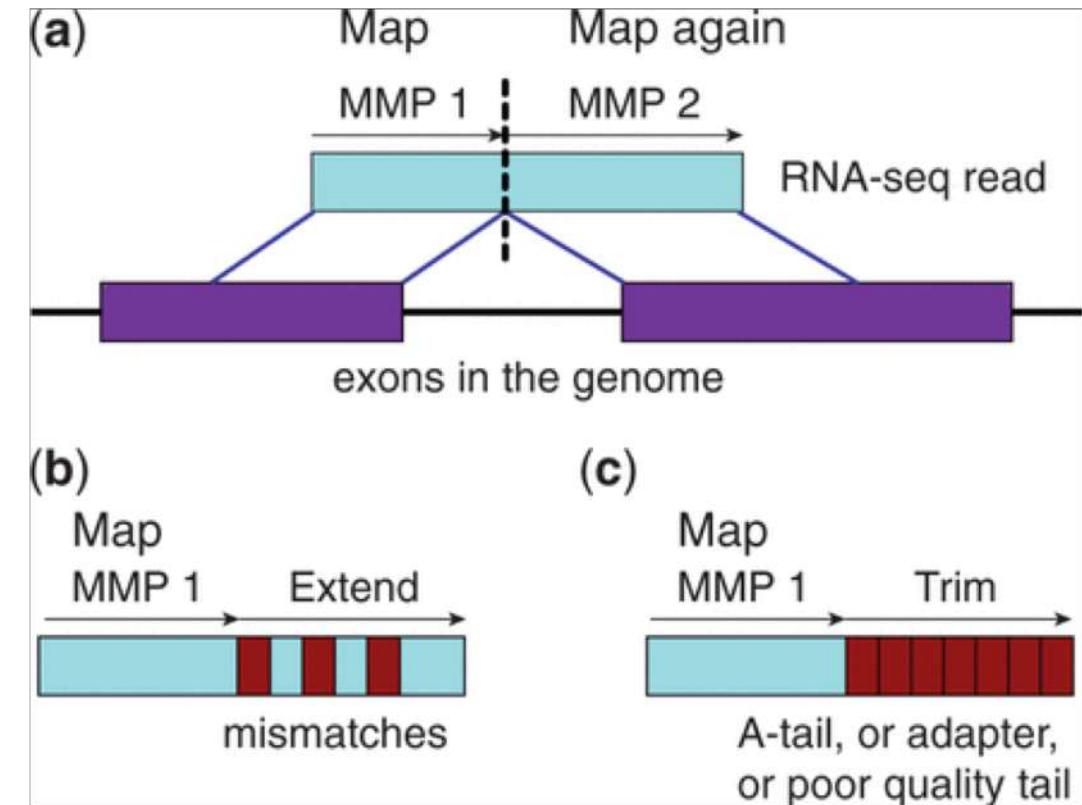
Preferential alignment
(mismatch rather than split)
to pseudogene



Tophat and STAR (different ways to handle split reads)



Trapnell et al (2009)



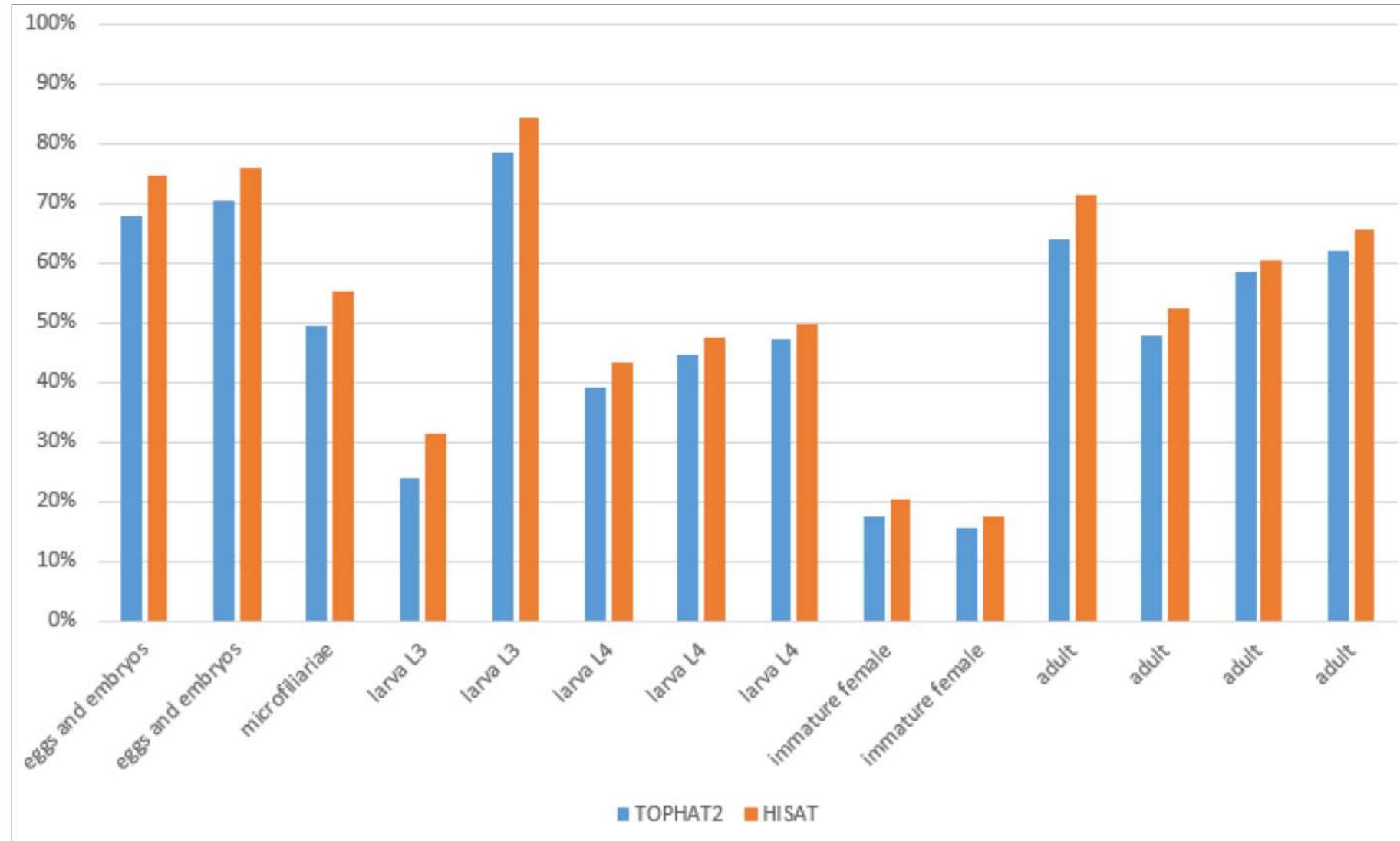
Dobin et al (2013)

Tophat2 is being obsoleted

As reads get longer, better methods are implemented

This leads to more reads get mapped and in a much faster speed

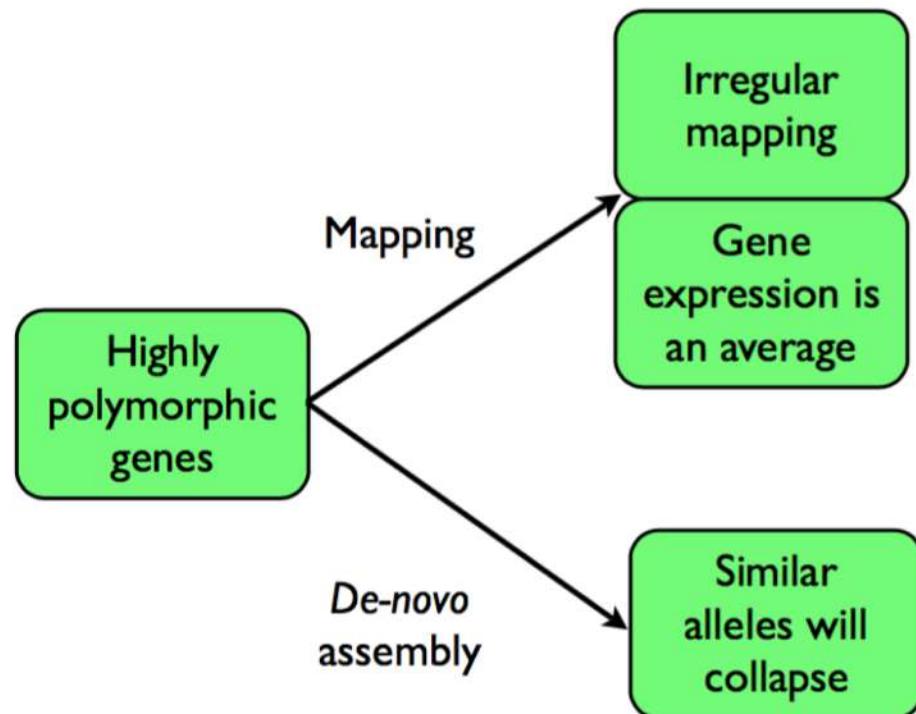
Obsoleted



A potential mapping problem

High heterozygosity/Polypliod problem:

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



Reference Gene I

ATGCGCGCTAGACGACATGACGACA

CACTT GACGACATGACG **Gene I A**

CTT GACGACATGACGAC

CCCTT GACGACATGACG

CGCCCTT GACGACATGA **Gene I B**

Expression Gene I = A + B

CACTT GACGACATGACG **Gene I A**

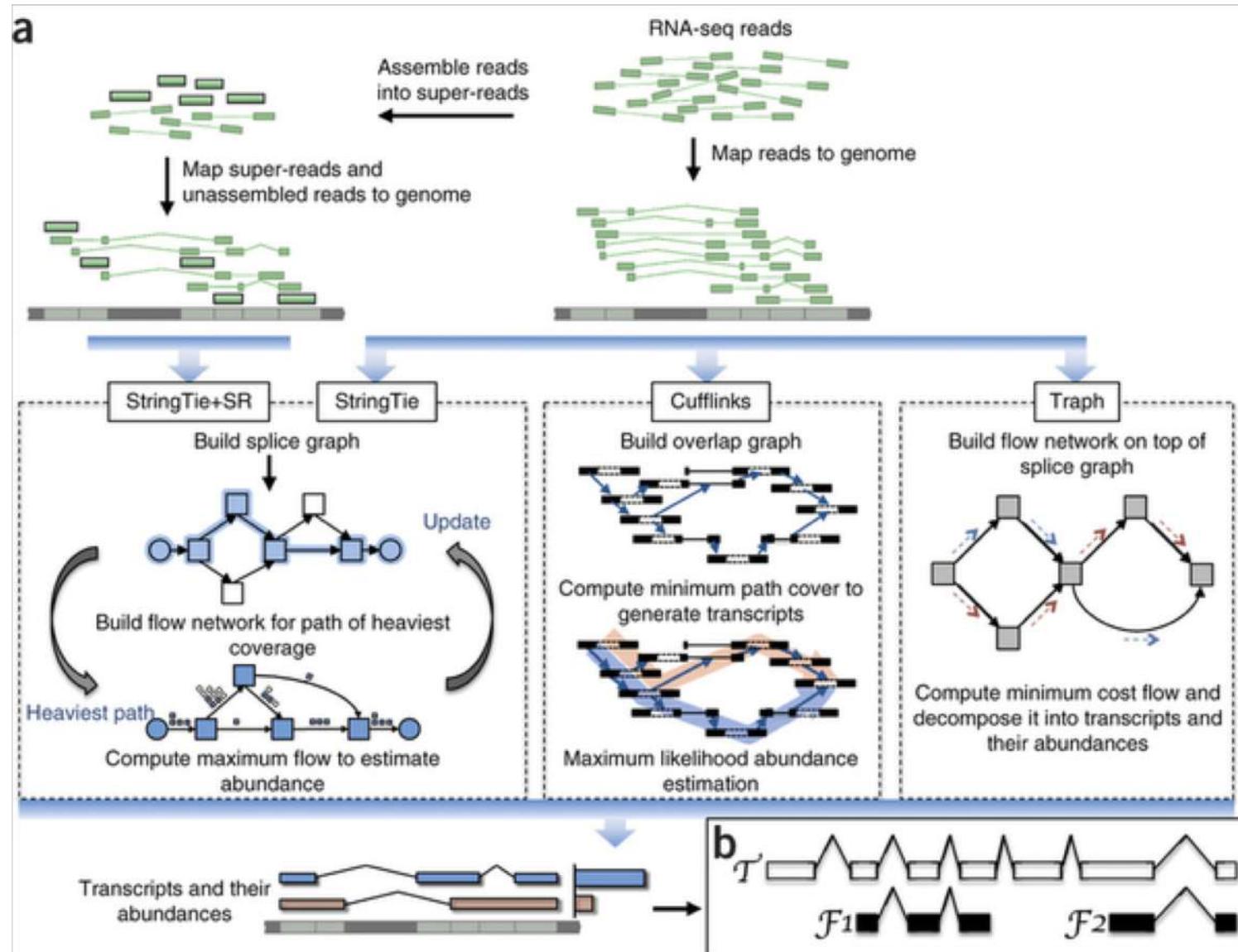
CTT GACGACATGACGAC

CCCTT GACGACATGACG **Gene I B**

CGCCCTT GACGACATGA

Collapsed consensus Gene A + Gene B

Transcript reconstruction: Cufflinks and StringTie



De novo assembly of transcriptomes

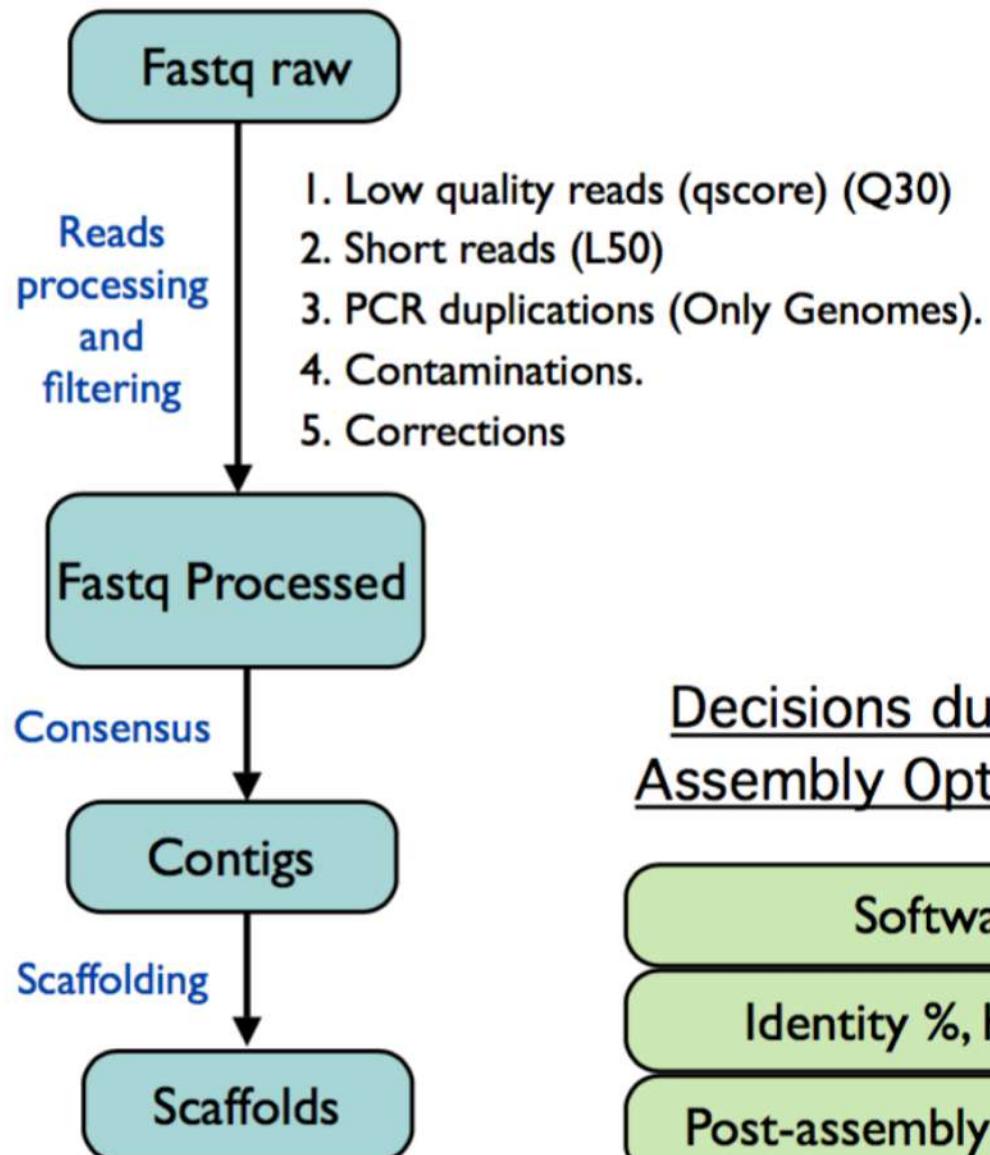
Transcriptome *de novo* assembly

Decisions during the Experimental Design

Technology

Library Preparation

Sequencing Amount



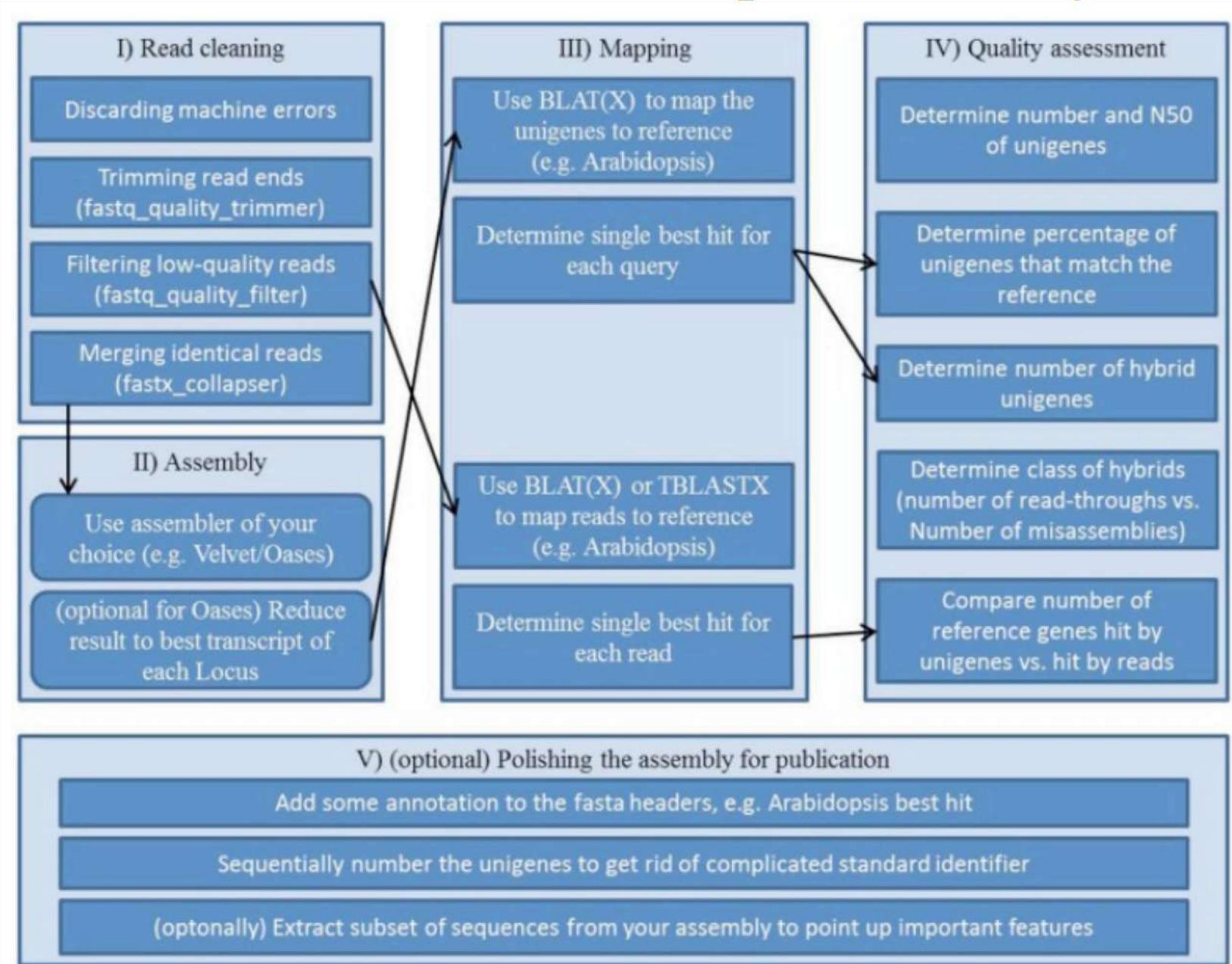
Decisions during the Assembly Optimization

Software

Identity %, Kmer ...

Post-assembly Filtering

Workflow



Obsoleted



Software	Sequencing technology	Type	Features	URL
MIRA	Sanger, 454	Overlap-layout-consensus	Highly configurable	http://sourceforge.net/apps/mediawiki/mira-assembler
gsAssembler	Sanger, 454	Overlap-layout-consensus	Splicings	http://454.com/products/analysis-software/index.asp
iAssembler	Sanger, 454	Overlap-layout-consensus	Improves MIRA	http://bioinfo.bti.cornell.edu/tool/iAssembler
Trans-ABySS*	454 or Illumina	Bruijn graph	Splicings, Gene fusions	http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss
SOAPdenovo-trans*	454 or Illumina	Bruijn graph	Fastest	http://soap.genomics.org.cn/SOAPdenovo-Trans.html
Velvet/Oases	454 or Illumina or SOLiD	Bruijn graph	SOLiD	http://www.ebi.ac.uk/~zerbino/oases/
Trinity*	454 or Illumina	Bruijn graph	Downstream expression	http://trinityrnaseq.sourceforge.net/

Transcriptome assembly benchmark 2018 version

Table 1: Benchmarking of Trans-ABySS, **IDBA**-tran, **SOAPdenovo-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

If you really have no reference and have to do assembly, then:

- A good assembly is important for transcriptome analysis for non-model organisms
- Strategy 1 (all samples in one assembly) delivers longer transcripts, but results in more mis-assembly
 - Be careful to fusion transcripts
- Redundancy does matter quantification
- Strand-specific poly-A > non-strand-specific poly-A > strand-specific ribo-minus

What is the best protocol for RNAseq analysis?

(Quick answer: no quick answer)

What analysis combinations should we do?

ARTICLE

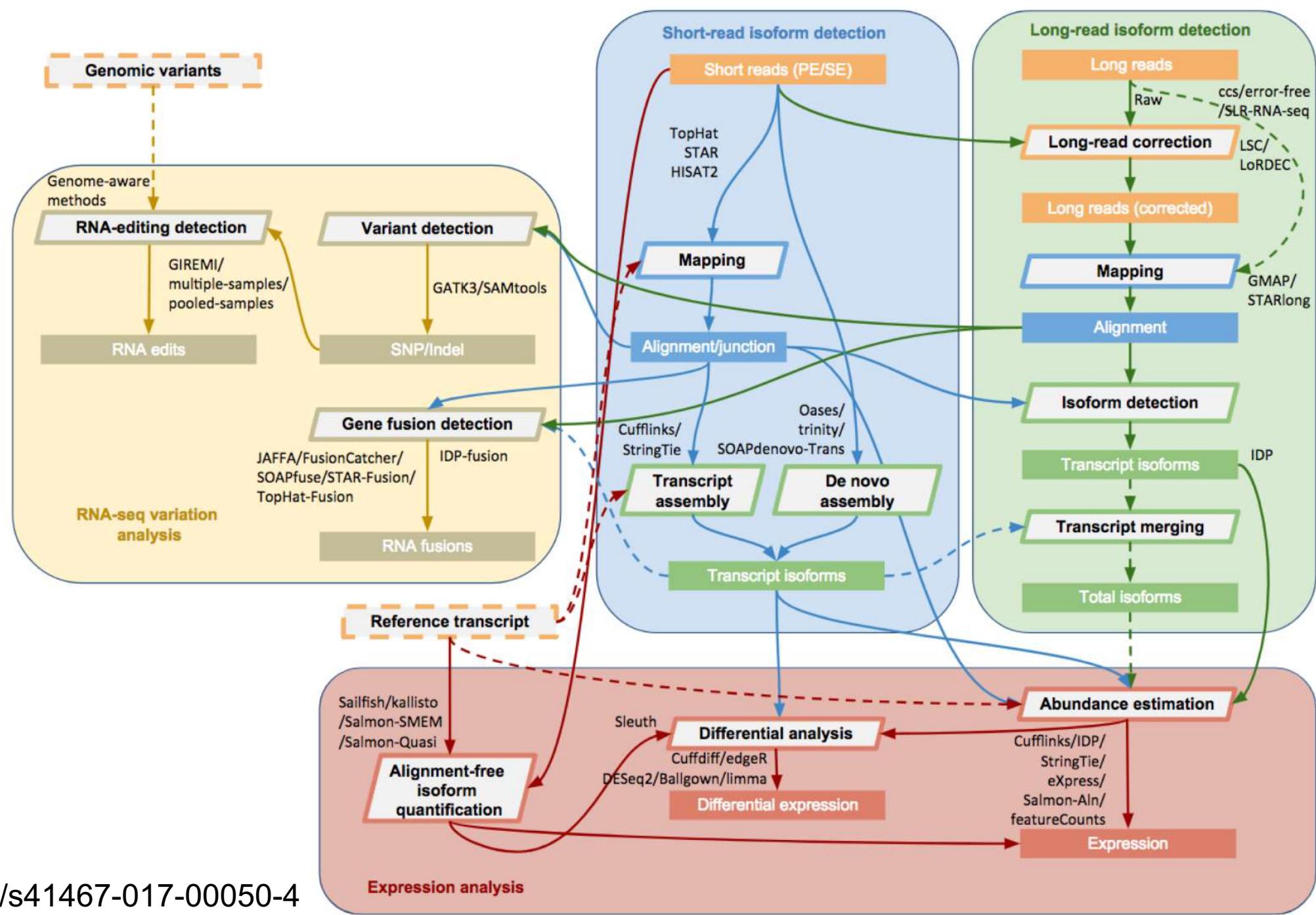
DOI: 10.1038/s41467-017-00050-4

OPEN

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

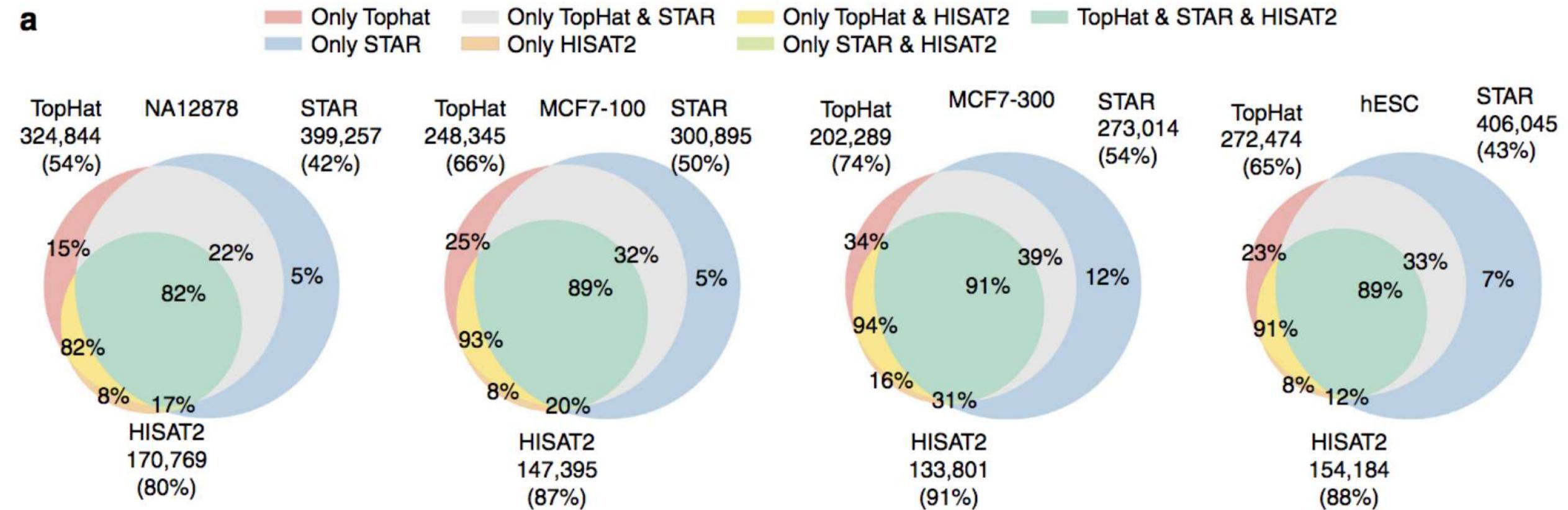
Sayed Mohammad Ebrahim Sahraeian¹, Marghoob Mohiyuddin¹, Robert Sebra², Hagen Tilgner³, Pegah T. Afshar⁴, Kin Fai Au⁵, Narges Bani Asadi¹, Mark B. Gerstein⁶, Wing Hung Wong⁷, Michael P. Snyder³, Eric Schadt² & Hugo Y.K. Lam¹

... 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.

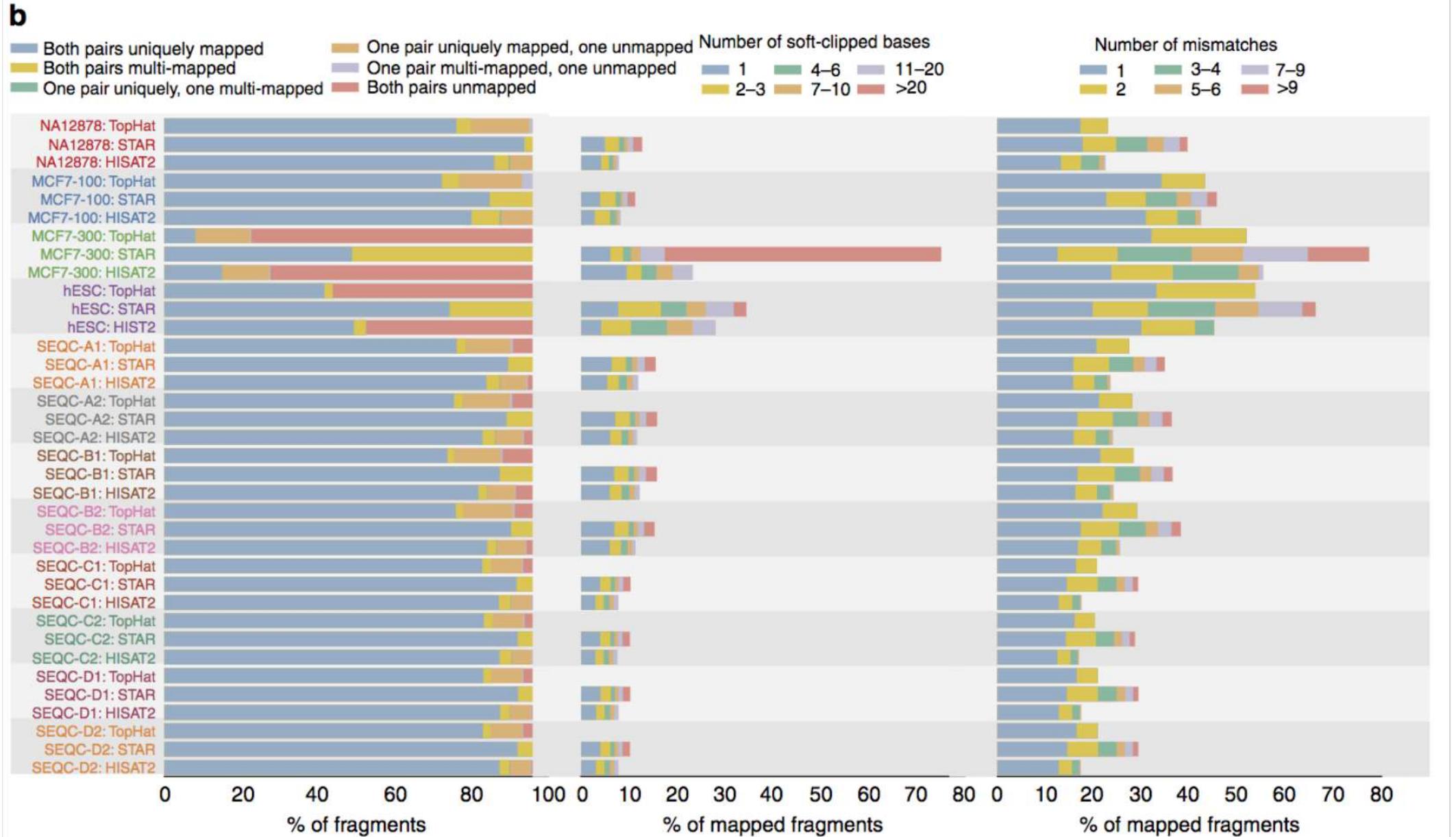


Can be ~10% difference in mapping

a

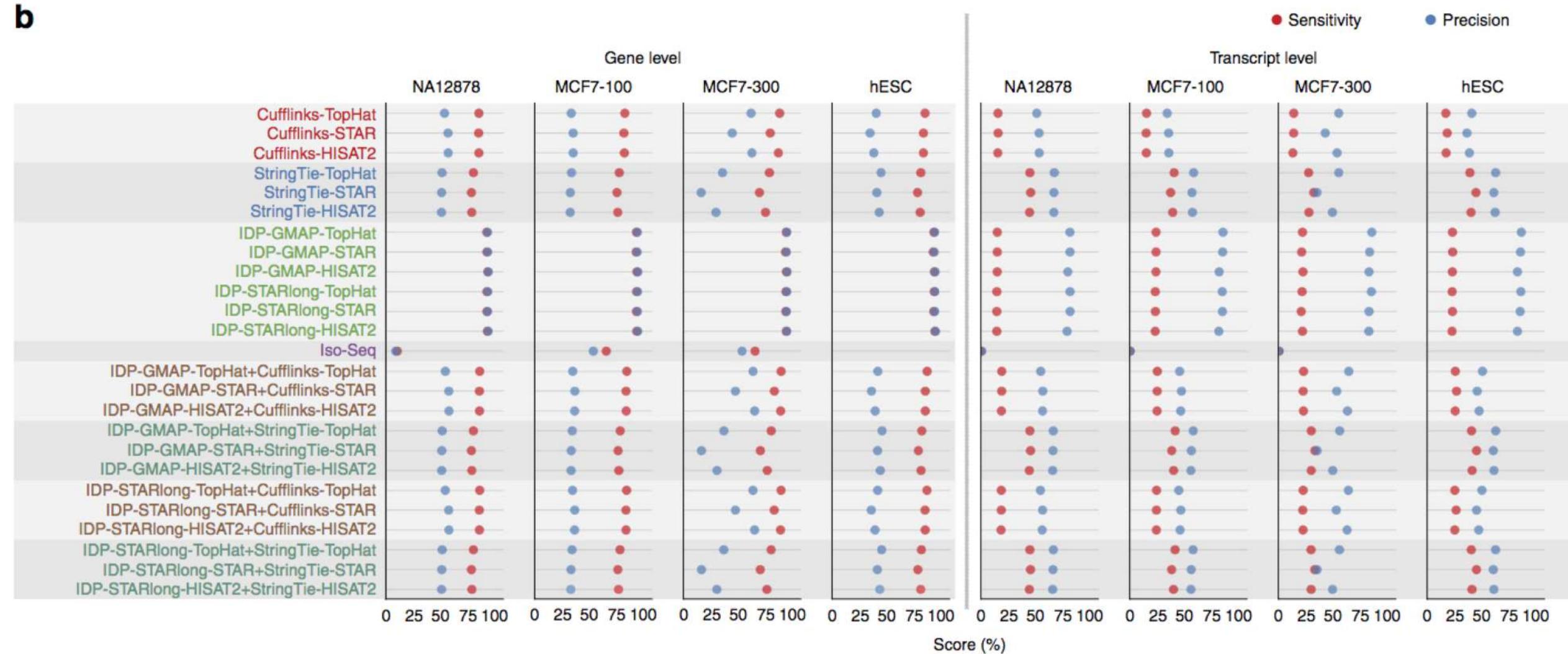


Read mapping different amongst tools



Performance of different transcriptome reconstruction schemes

b



Annotation
(focus only on gene annotation)

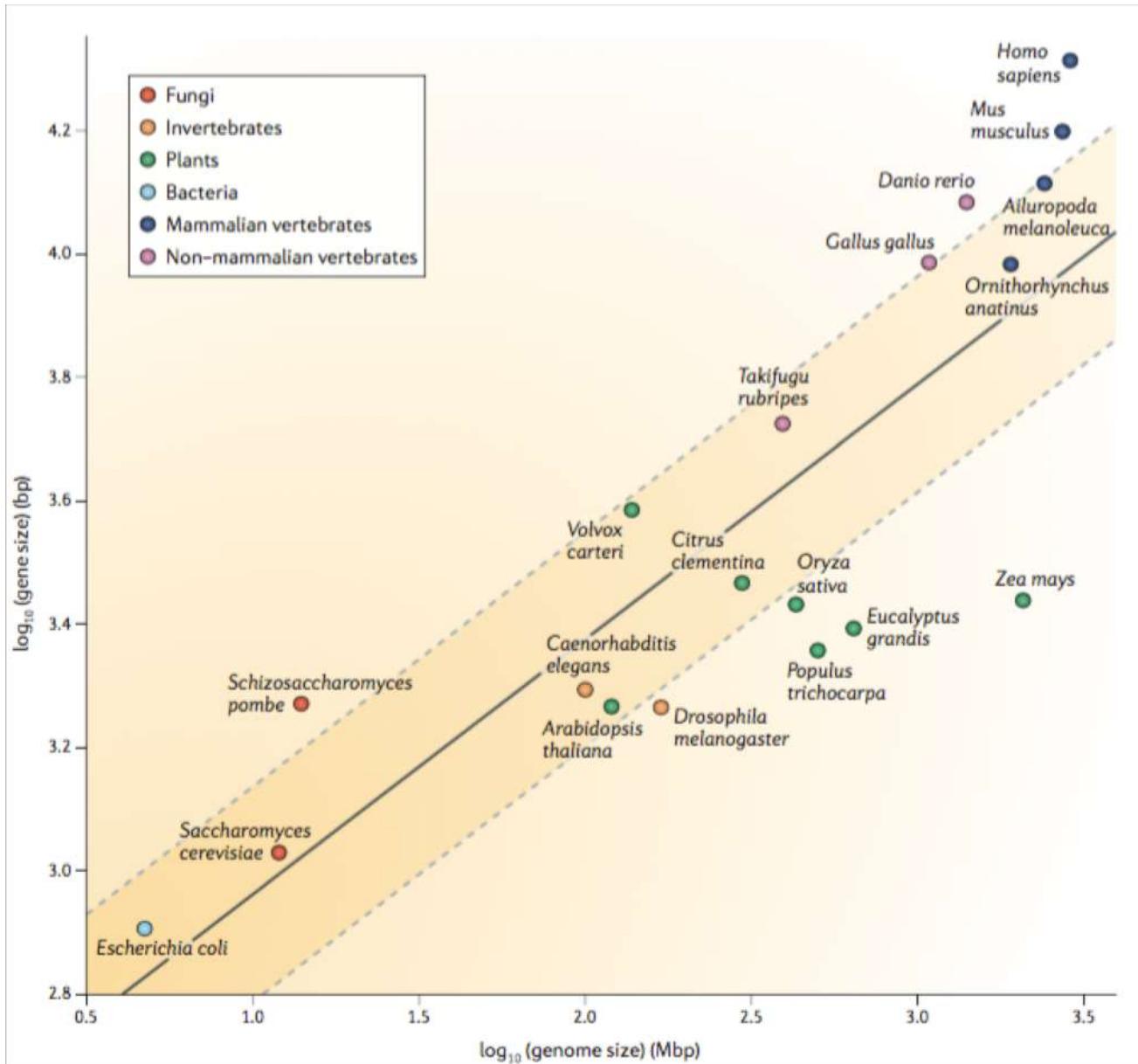


A beginner's guide to eukaryotic genome annotation

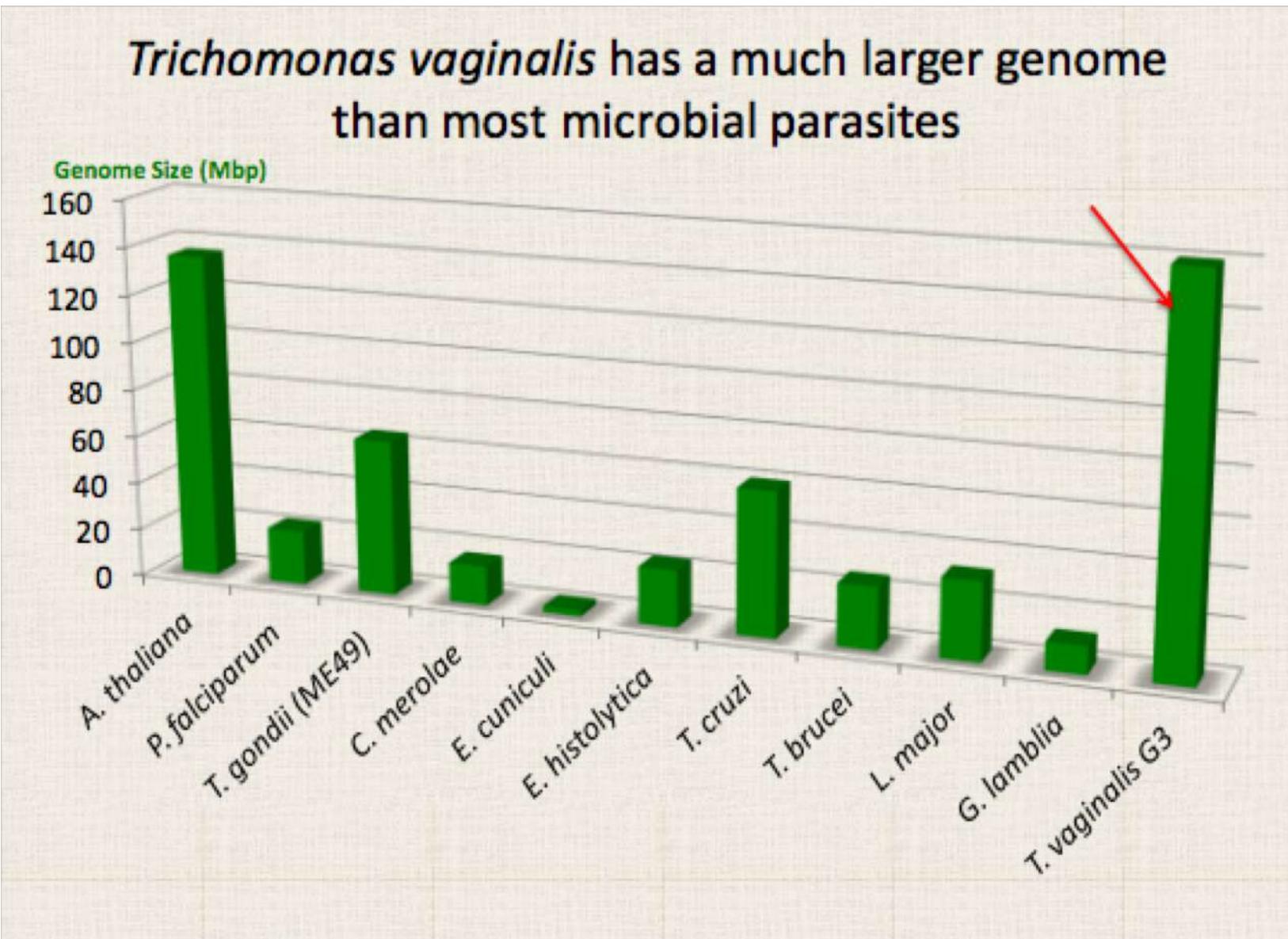
Mark Yandell and Daniel Ence

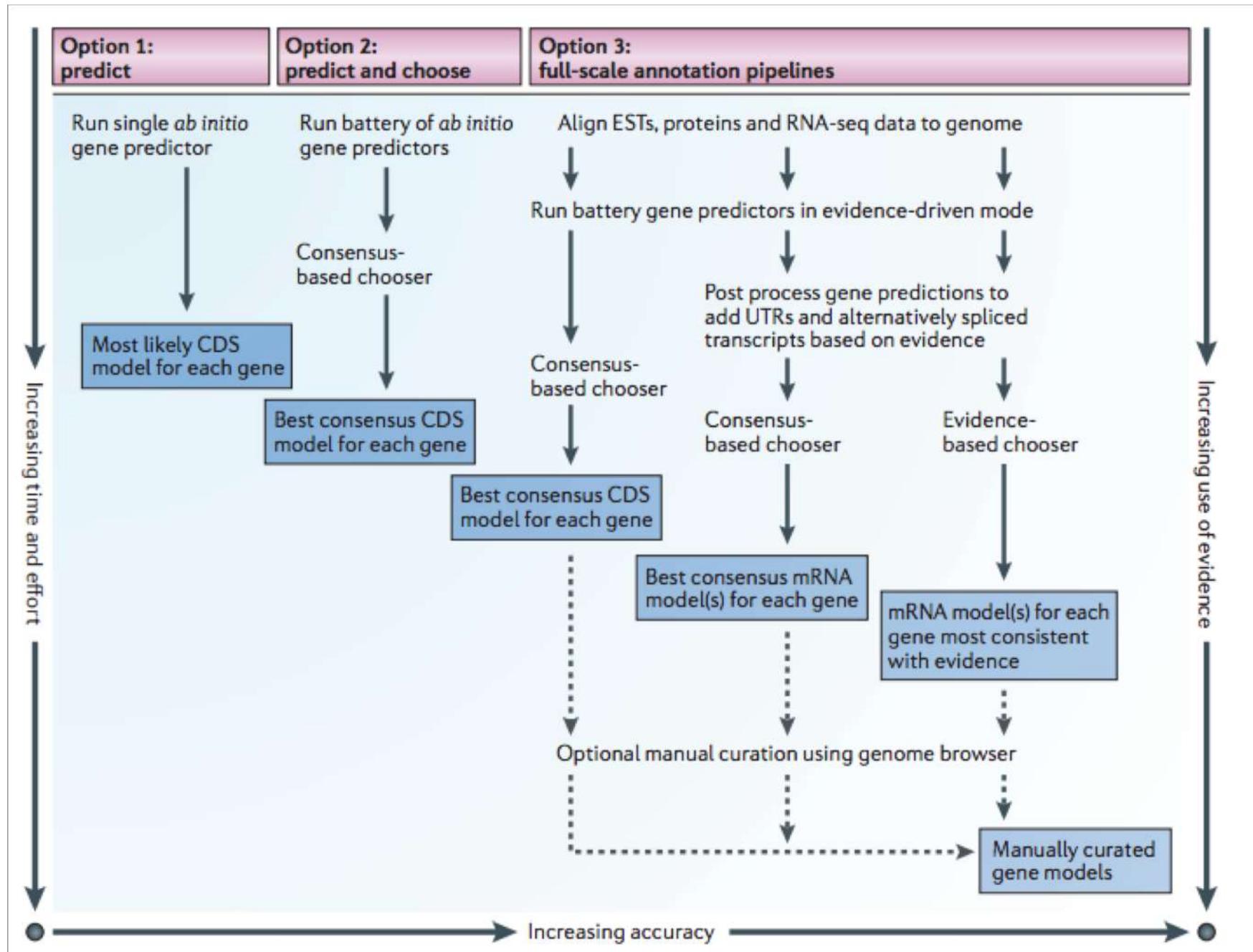
Abstract | The falling cost of genome sequencing is having a marked impact on the research community with respect to which genomes are sequenced and how and where they are annotated. Genome annotation projects have generally become small-scale affairs that are often carried out by an individual laboratory. 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.

Know your genome size (and gene numbers)

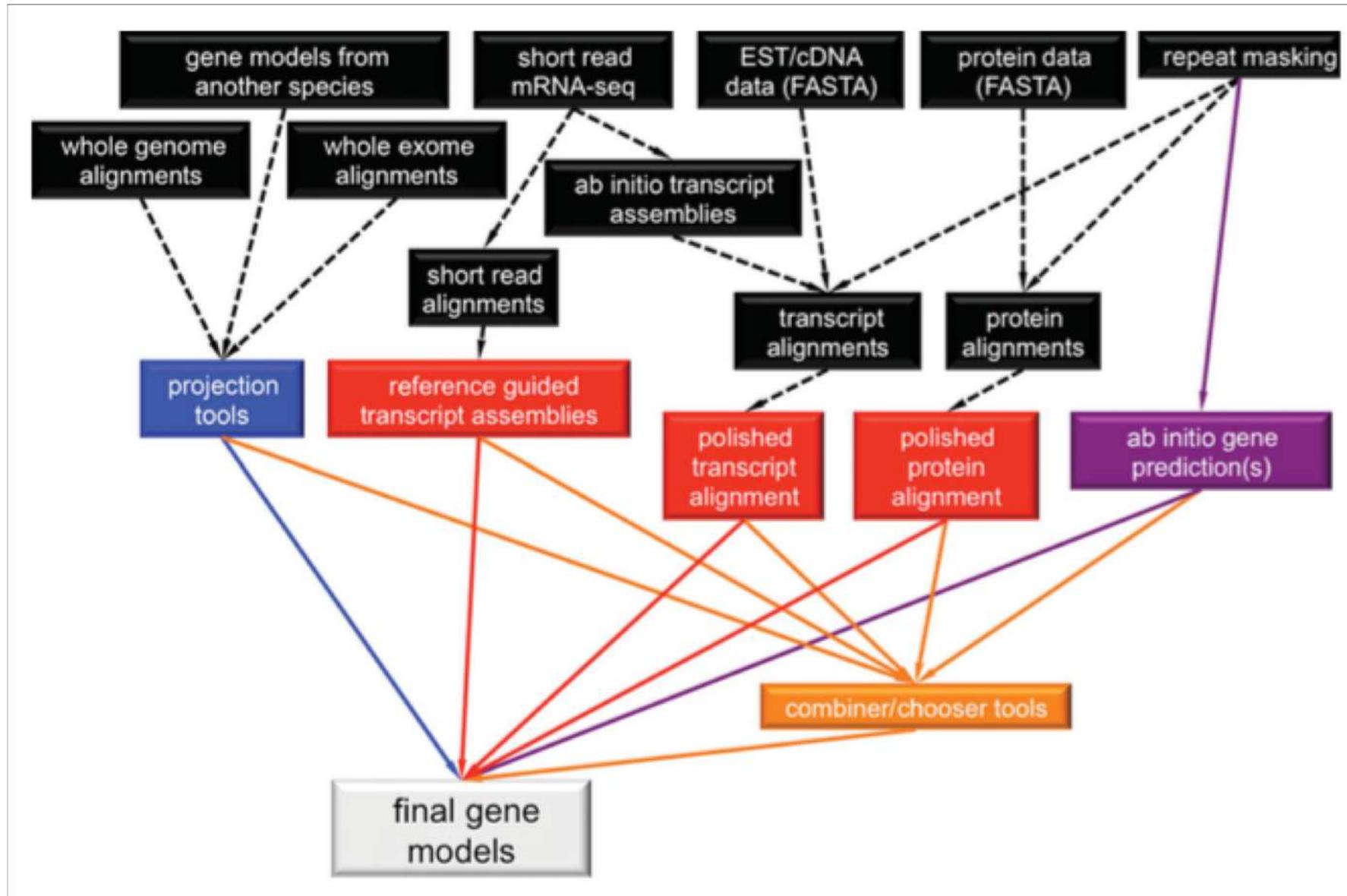


There's always exceptions (due to TE Maverick expansion)





Multiple evidences; Update



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

Obsolete

Basic rule of thumb 2018 edition

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 + long and short read 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

Prediction and Evidence aligners

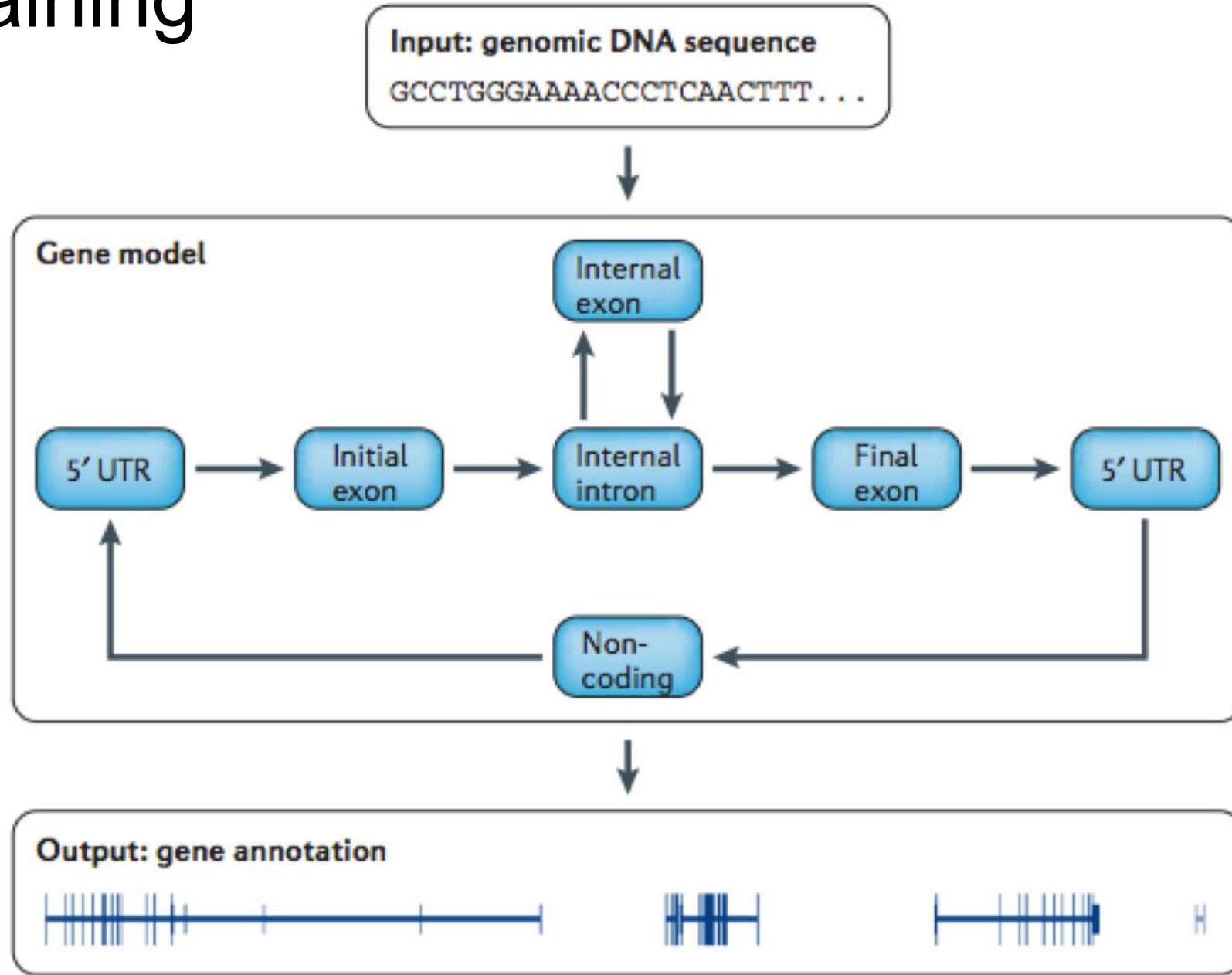
Table 4.1.2 Gene Predictors

Software Package	Features	Reference
Augustus	Can incorporate mRNA-seq data. Predicts alternatively spliced transcripts.	(Stanke and Waack, 2003; Stanke et al., 2008; Hoff and Stanke, 2013)
Genemark	Self-training. Performs well on fungal genomes. Versions available for prokaryotic and eukaryotic gene prediction.	(UNIT 4.5 and 4.6; Lomsadze et al., 2005; Ter-hovhannisyan et al., 2008; Borodovsky and Lomsadze, 2011a,b; Lomsadze et al., 2014)
Fgenesh	Runs locally or by Web service. Fee-for-use. Trained by softberry (no local training option).	(Solovyev et al., 2006)
SNAP	Easily trained. Incorporates hints from mRNA-seq and protein alignments.	(Korf, 2004)
Gnomon	Uses a combination of ab initio modeling and homology searching. Accepts mRNA-seq and protein data.	(Souverov et al., 2010)
mGene	Utilizes multiple machine learning techniques, including generalized hidden Markov models and Support Vector Machines.	(Schweikert et al., 2009)

Table 4.1.1 Evidence Aligners and Assemblers

Software package	Features	Reference
BLAST	A suite of tools that can align any combination of protein and nucleotide sequences. Uses Karlin-Altschul statistics.	(UNIT 3.4; Korf et al., 2003; Ladunga, 2009)
BLAT	Faster than BLAST but not as configurable.	(UNIT 10.8; Kent, 2002; Bhagwat et al., 2012)
Tophat2	Memory efficient splice-junction mapper for RNA-seq reads.	(Kim et al., 2013)
StringTie	Assembles transcripts from Tophat-aligned RNA-seq reads, and estimates transcript abundance. Designed to succeed Cufflinks.	(Pertea et al., 2015)
Trinity	Assembles transcripts <i>de novo</i> or with reference guidance.	(Grabherr et al., 2011)
NovoAlign	Aligns RNA and DNA short-read sequences. Can use ambiguous nucleotide codes in the reference sequence. Requires purchased license.	(see Internet Resources)
GSNAP	Single-nucleotide-variant tolerant aligner for splice site detection. Available as part of the GMAP package.	(Wu and Nacu, 2010)
Splign	Combines global and local alignment algorithms in a splice-aware manner to align transcript sequences to a reference.	(Kapustin et al., 2008)
MapSplice	Splice-junction mapper for RNA-seq reads.	(Wang et al., 2010)
STAR	Very fast and accurate RNA-seq aligner uses sequential mappable seed search in uncompressed suffix arrays.	(UNIT 11.14; Dobin et al., 2013; Dobin and Gingeras, 2015)
Exonerate	Aligns proteins and assembled transcripts to a reference in a splice-aware manner.	(Slater and Birney, 2005)

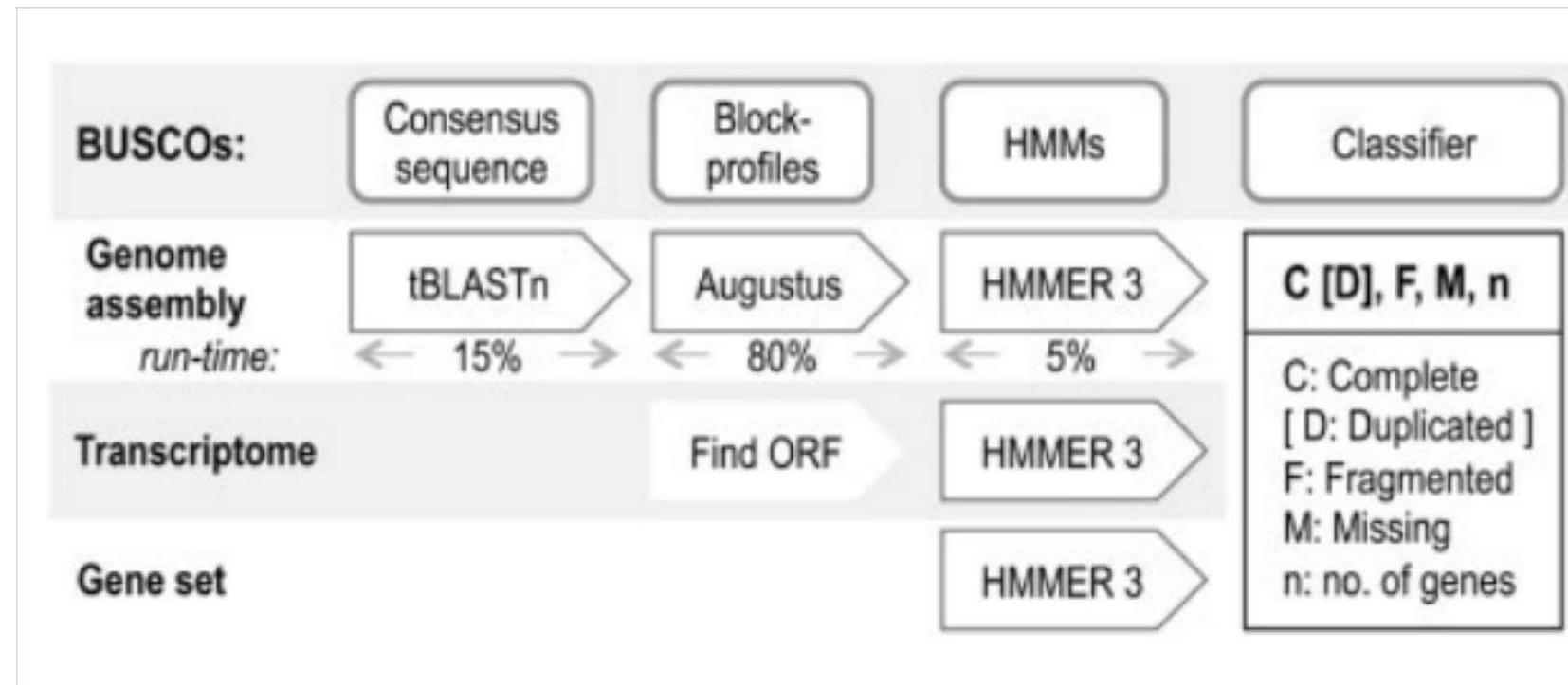
Training



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*



Watch out for latest tool improvement

Research

Open Access

AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome

Mario Stanke, Ana Tzvetkova and Burkhard Morgenstern

Genome analysis

BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS

Katharina J. Hoff^{1,*}, Simone Lange¹, Alexandre Lomsadze³,
Mark Borodovsky^{2,3,4,*} and Mario Stanke¹

Nov. 2015

Genome annotation pipelines

Table 4.1.5 Genome Annotation Pipelines

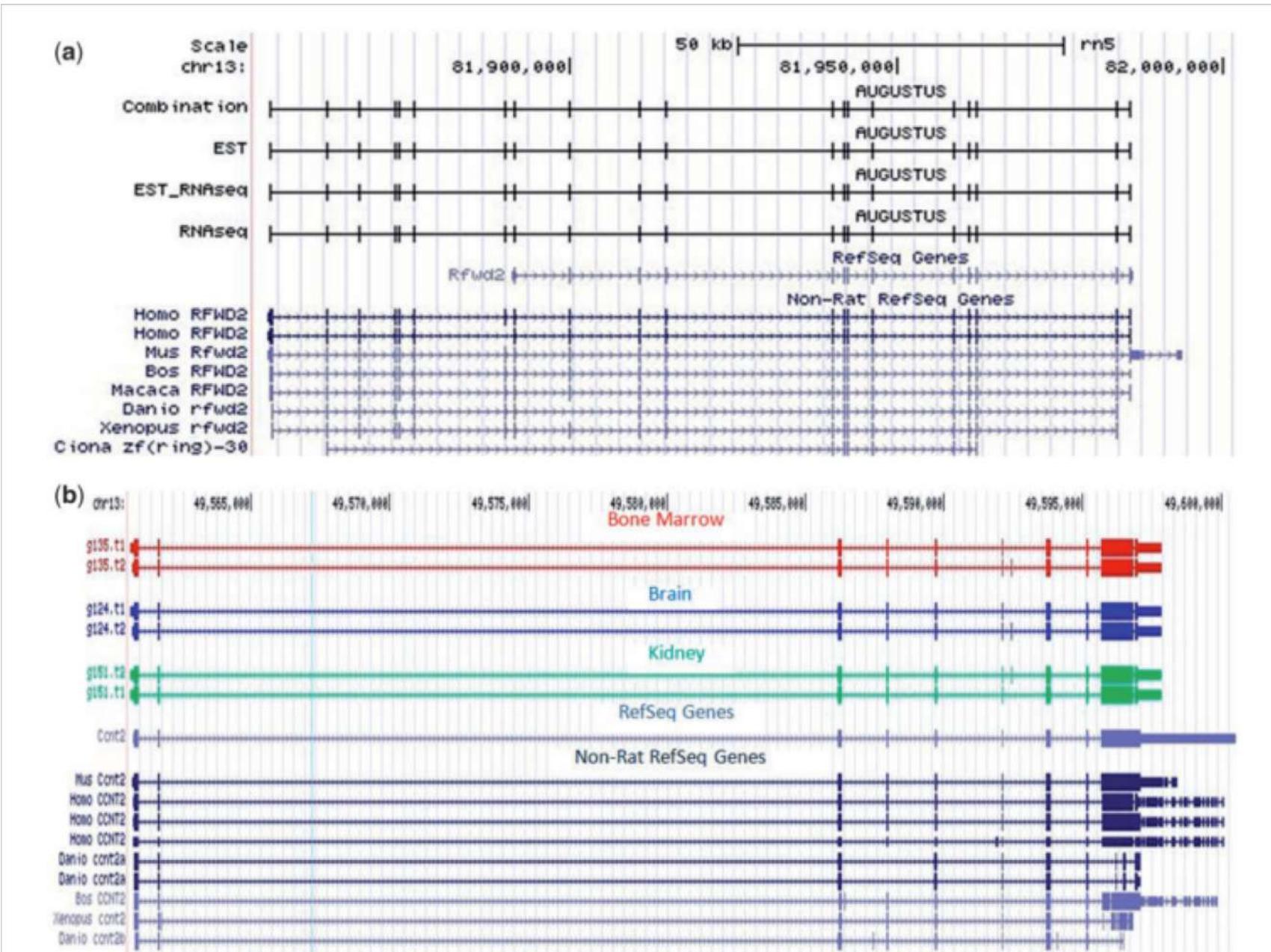
Software package	Features	Reference
EuGene	Annotation pipeline that integrates multiple evidence types using a C++ based plugin system.	(Foissac et al., 2008)
MAKER	Annotation pipeline that aligns and polishes protein and transcriptome data with BLAST and Exonerate, provides evidence-based hints to gene predictors, and provides an evidence trail and quality metrics for each annotation. Highly parallelizable (a single command can use thousands of CPUs if available).	(Cantarel et al., 2008; Holt and Yandell, 2011; <i>UNIT 4.11</i> ; Campbell et al., 2014a,b)
Ensembl	Annotation pipeline that builds gene models from aligned and polished protein and transcript data. Identical transcripts are merged and a non-redundant set of transcripts is reported for each gene.	(<i>UNIT 1.15</i> ; Curwen et al., 2004; Fernández-Suárez and Schuster, 2010)
NCBI	Annotation pipeline that aligns and polishes protein and transcript data. Generates Gnomon gene predictions. Weights gene models based on manually curated evidence higher than computationally derived models.	(<i>UNIT 1.3</i> ; Gibney and Baxevanis, 2011; Thibaud-Nissen et al., 2013)
PASA	Annotation pipeline that aligns transcripts to the genome using BLAT, GMAP, or sim4. Can generate annotations based only on transcript data or on pre-existing gene models or predictions.	(Haas et al., 2008)

Annotation done; next is manual/community curation

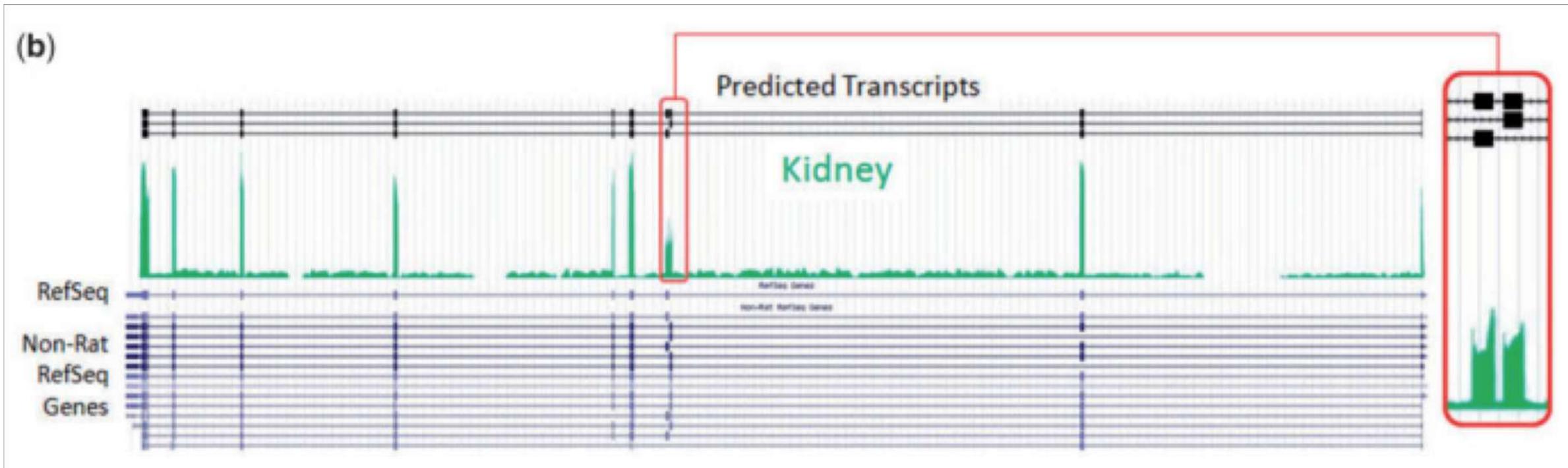
Table 4.1.6 Genome Browsers for Community Curation

Software package	Features	Reference
WebApollo	Web-based plug-in for Jbrowse with an editable user-created annotation track. Edits are visible in real time to all curators.	(Lee et al., 2013)
Argo	Stand-alone Java application for viewing and editing gene annotations.	(see Internet Resources)
IGV	Genome viewer that supports a variety of data types, including BAM and array based data. Also available for iPad.	(Robinson et al., 2011; Thorvaldsdóttir et al., 2015)
GenomeView	Stand-alone genome viewer and editor. Supports visualization of synteny and multiple-alignment data.	(Abeel et al., 2012)
Artemis	Browser and annotation tool than can read EMBL and GENBANK database entries; FASTA sequence formats (indexed or raw); and other features in EMBL, GENBANK, or GFF formats.	(Carver et al., 2012)
Jbrowse	Fast, embeddable genome browser. Supports multiple data formats, including VCF visualization.	(UNIT 9.13; Skinner et al., 2009; Skinner and Holmes, 2010)
Gbrowse	Feature-rich, highly customizable, Web-based genome browser. Predecessor of Jbrowse.	(UNIT 9.9; Stein et al., 2002; Donlin, 2009)

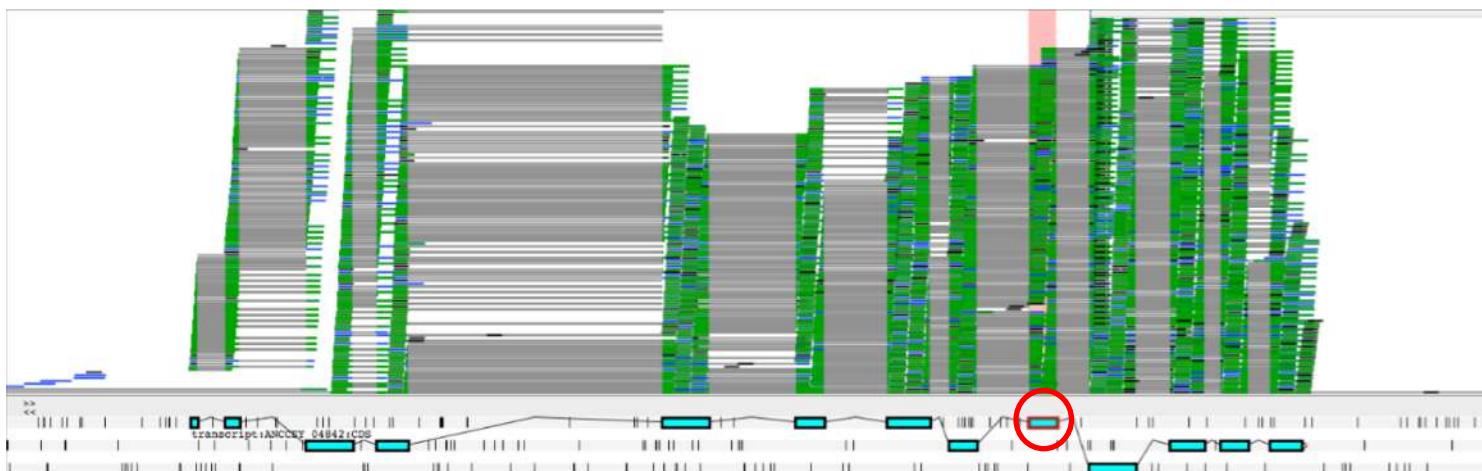
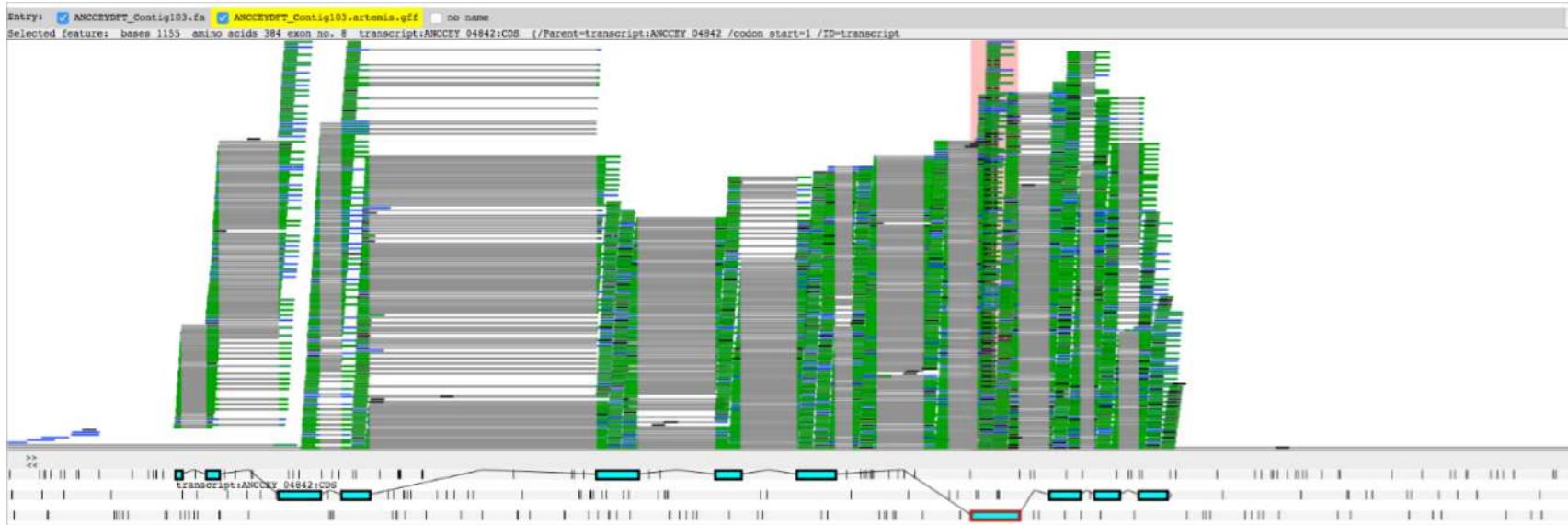
Combine multiple evidence will improve annotation



Novel isoforms

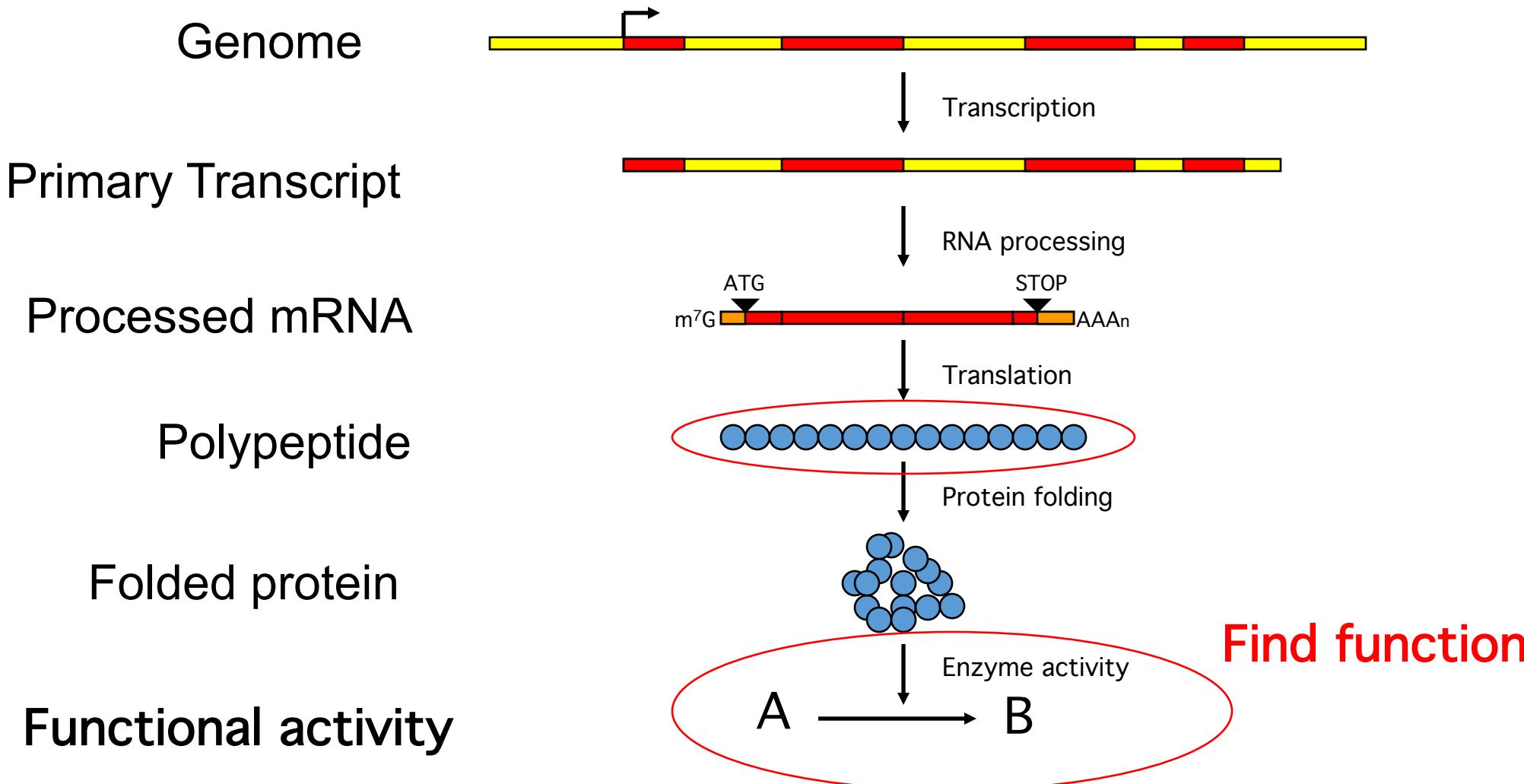


Manual curation using artemis



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

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

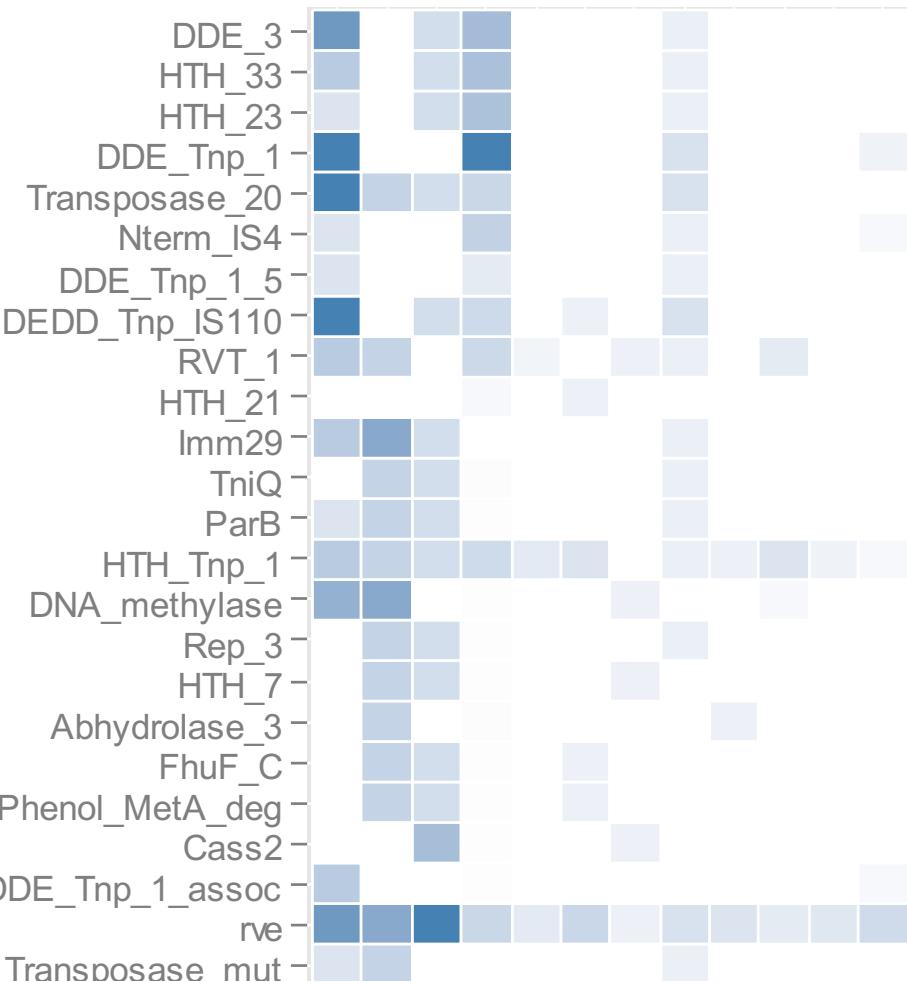
Visually assess the **top 5-10 hits** to identify whether these have been assigned a function

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

Protein families (Pfam)s



Pfam 32.0 (September 2018, 17929 entries)

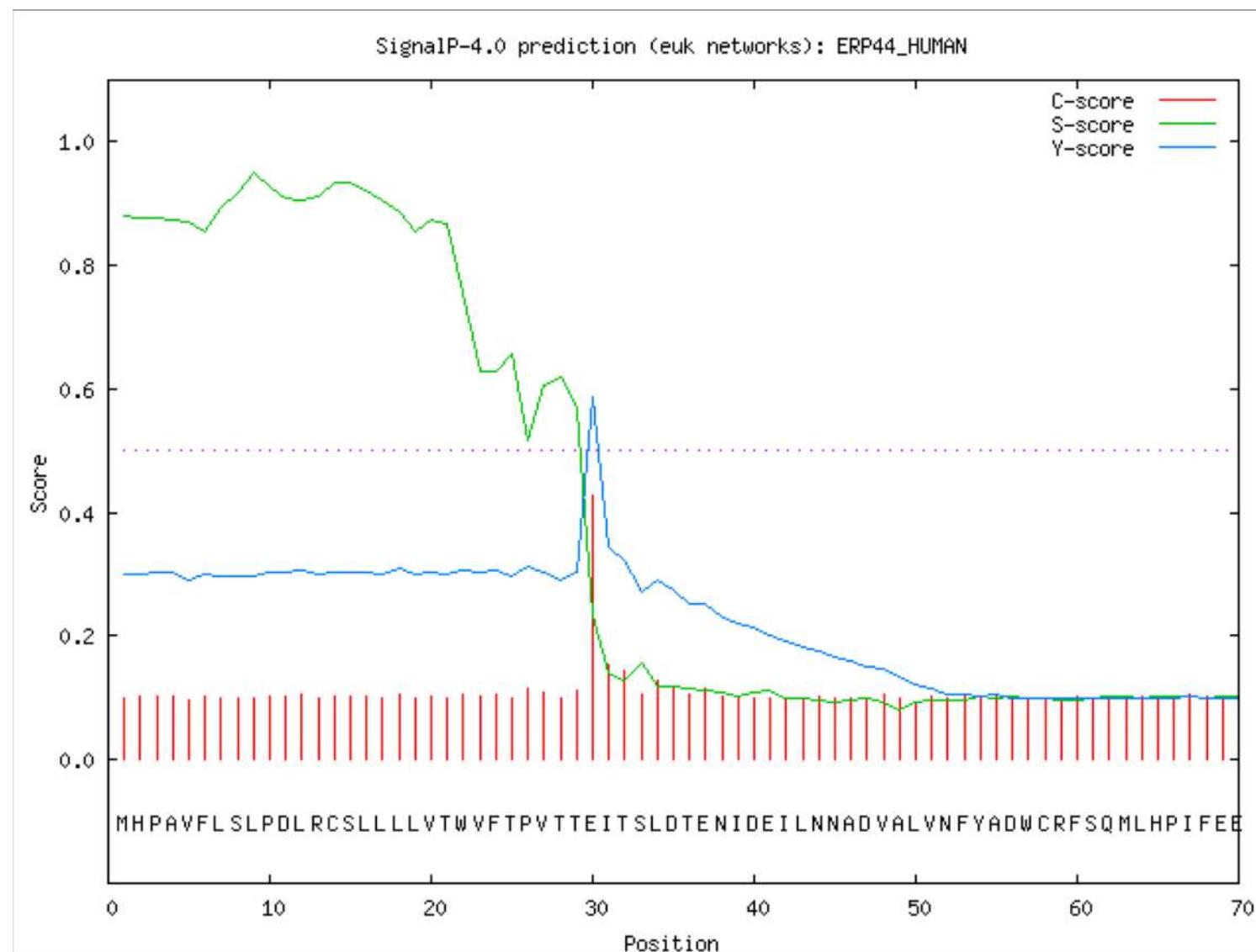
The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [Less...](#)

Proteins are generally composed of one or more functional regions, commonly termed **domains**. Different combinations of domains give rise to the diverse range of proteins found in nature. The identification of domains that occur within proteins can therefore provide insights into their function.

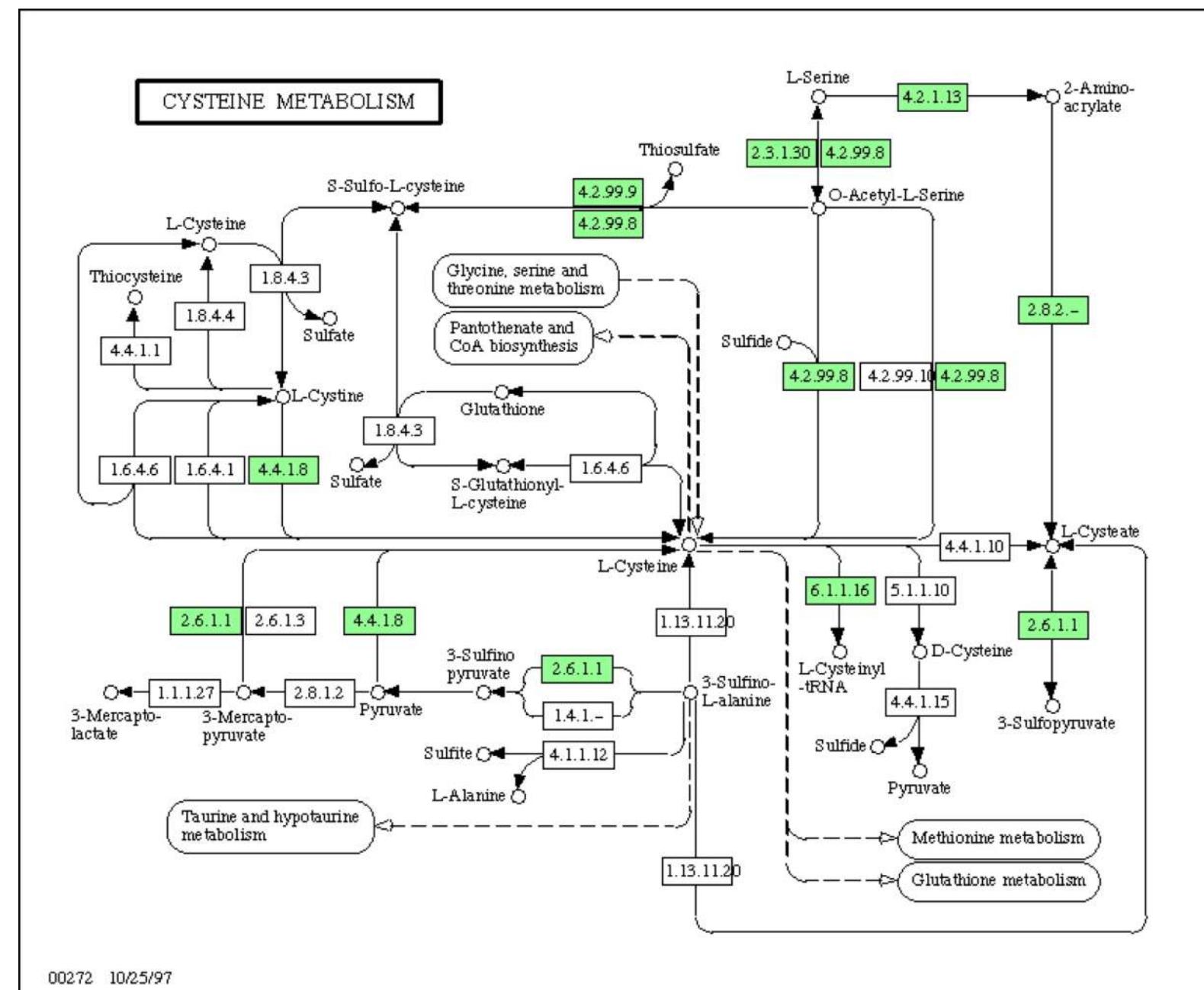
Pfam also generates higher-level groupings of related entries, known as **clans**. A clan is a collection of Pfam entries which are related by similarity of sequence, structure or profile-HMM.

The data presented for each entry is based on the [UniProt Reference Proteomes](#) but information on individual UniProtKB sequences can still be found by entering the protein accession. Pfam *full* alignments are available from searching a variety of databases, either to provide different accessions (e.g. all UniProt and NCBI GI) or different levels of redundancy.

SignalP: predicts the presence and location of signal peptide



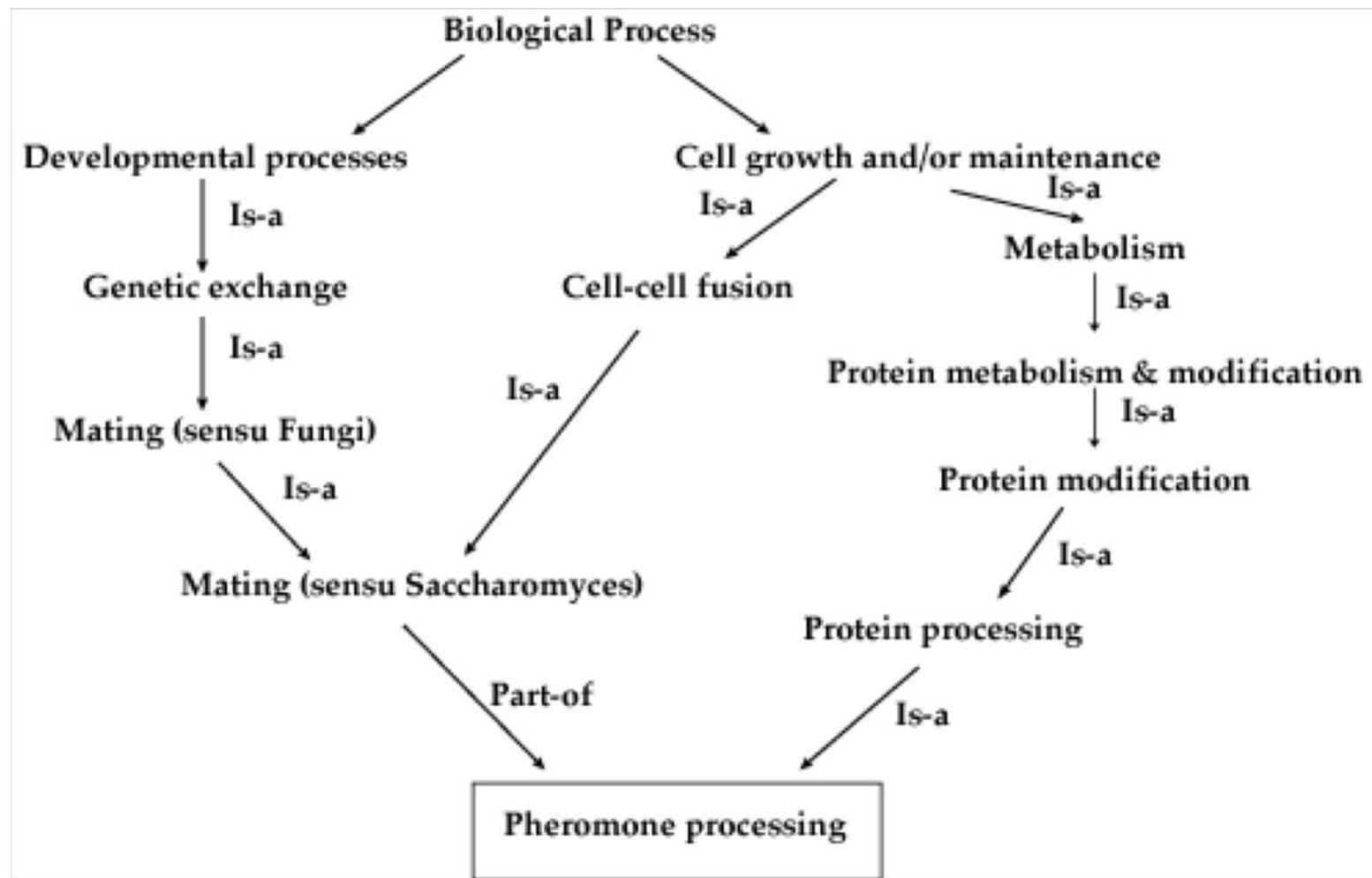
Help improve annotation by
showing missing genes in
essential pathways



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



BLAST2GO

Blast2GO PRO

b2g 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...	cyt_i_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							

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.

Welcome Message Blast Result of C04018A12

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

Sequences Producing Significant Alignments Gene Name

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

Default < Back Next > Cancel Run

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

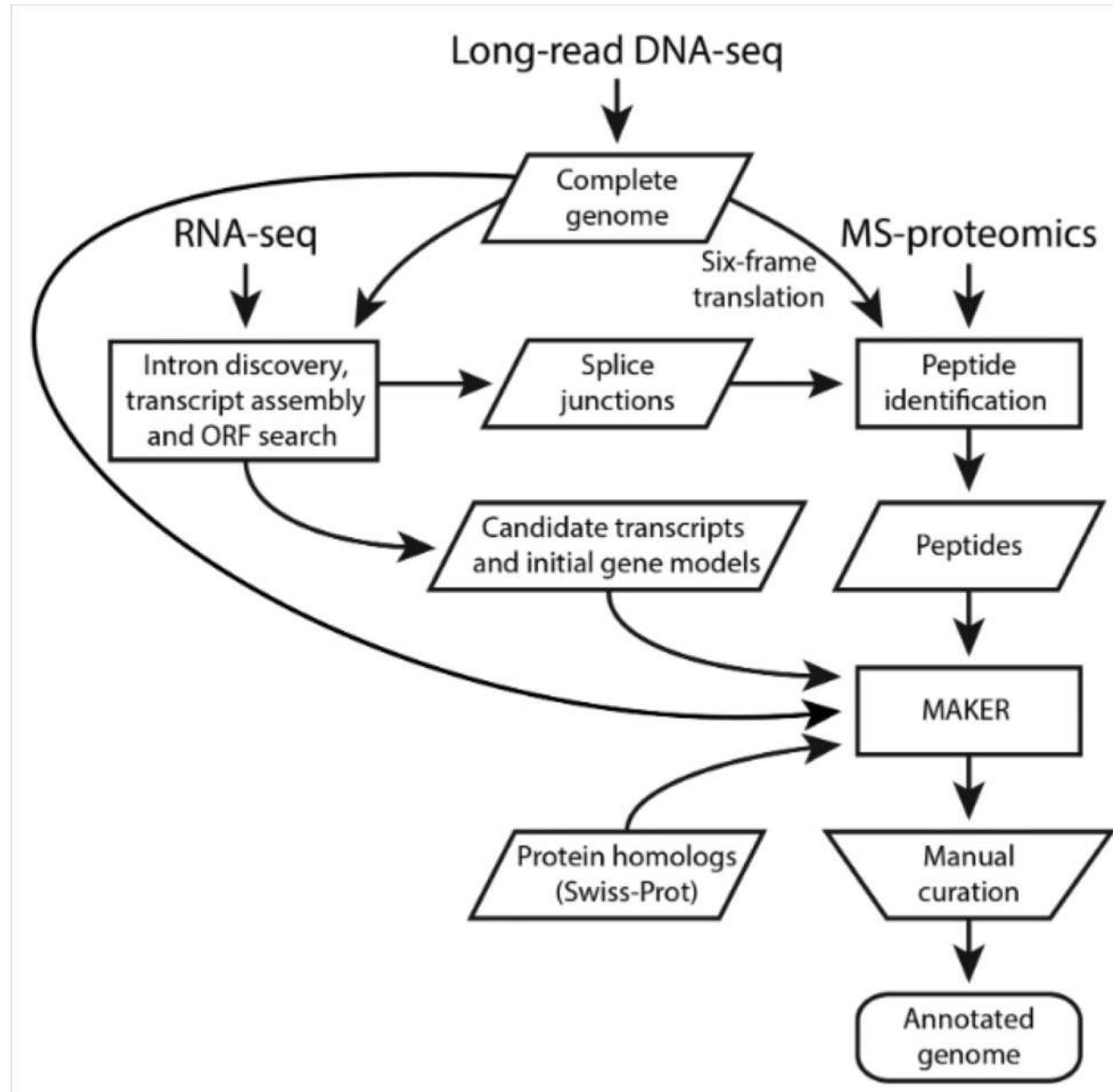
Case study: eukaryote annotation (2018)

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-Roth³, Charles D. Baudo⁴, John C. Kennell⁴, Sheng Sun⁵, R. Blake Billmyre⁵, Markus S. Schröder⁶, Anna Andersson⁷, Tina Holm⁷, Benjamin Sigurgeirsson⁸, Guangxi Wu⁹, Sundar Ram Sankaranarayanan¹⁰, Rahul Siddharthan¹¹, Kaustuv Sanyal¹⁰, Joakim Lundeberg⁸, Björn Nystedt¹², Teun Boekhout¹³, Thomas L. Dawson, Jr.¹⁴, Joseph Heitman⁵, Annika Scheynius^{15,*‡} and Janne Lehtiö^{1,*‡}



Current annotation



MAKER prediction (homology, RNA-seq and peptides)



Gioti *et al* annotation



RNA-seq coverage



Peptide evidence

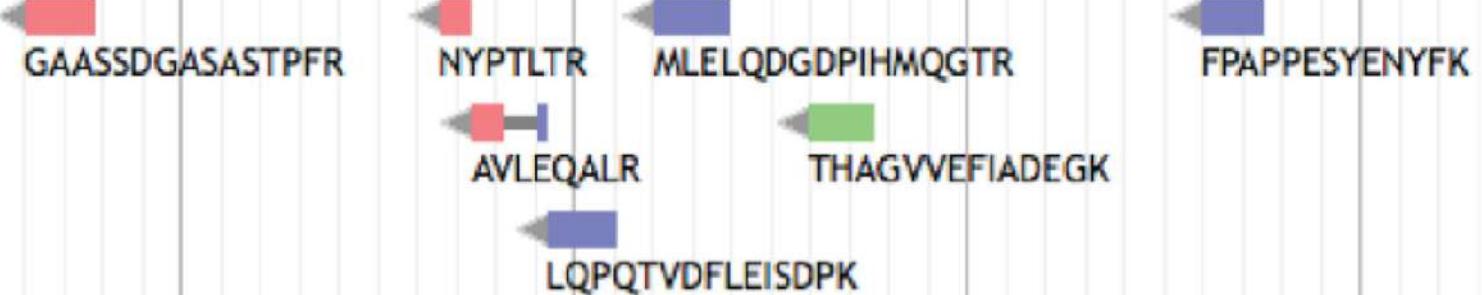


Table 2. Characteristics of *M. sympodialis* gene sets

	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%)

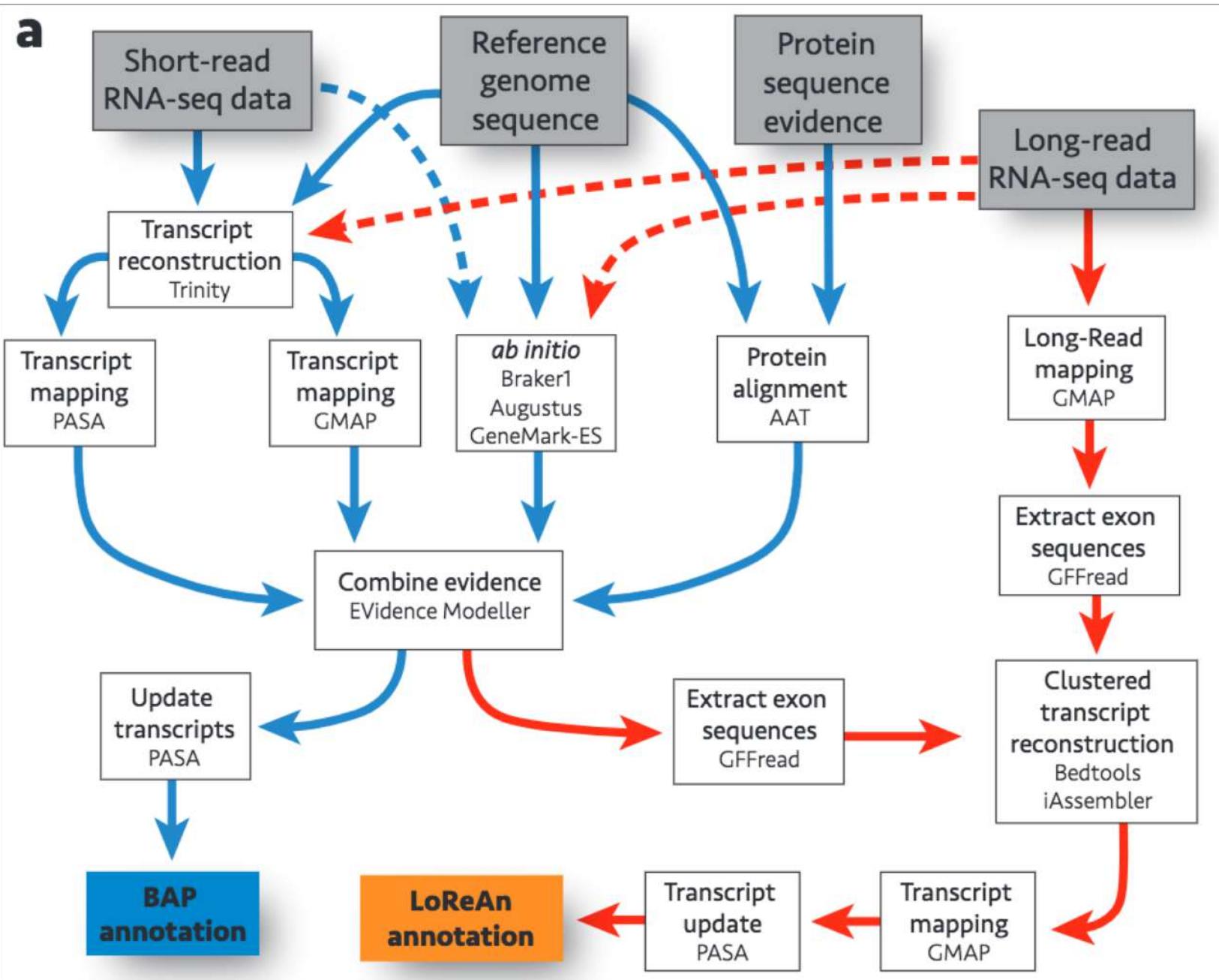
¹ Gene density was computed relative to the size of the corresponding genome assembly (7.71 Mb for the draft assembly of Gioti *et al.* (5) and 7.79 Mb for the current assembly).

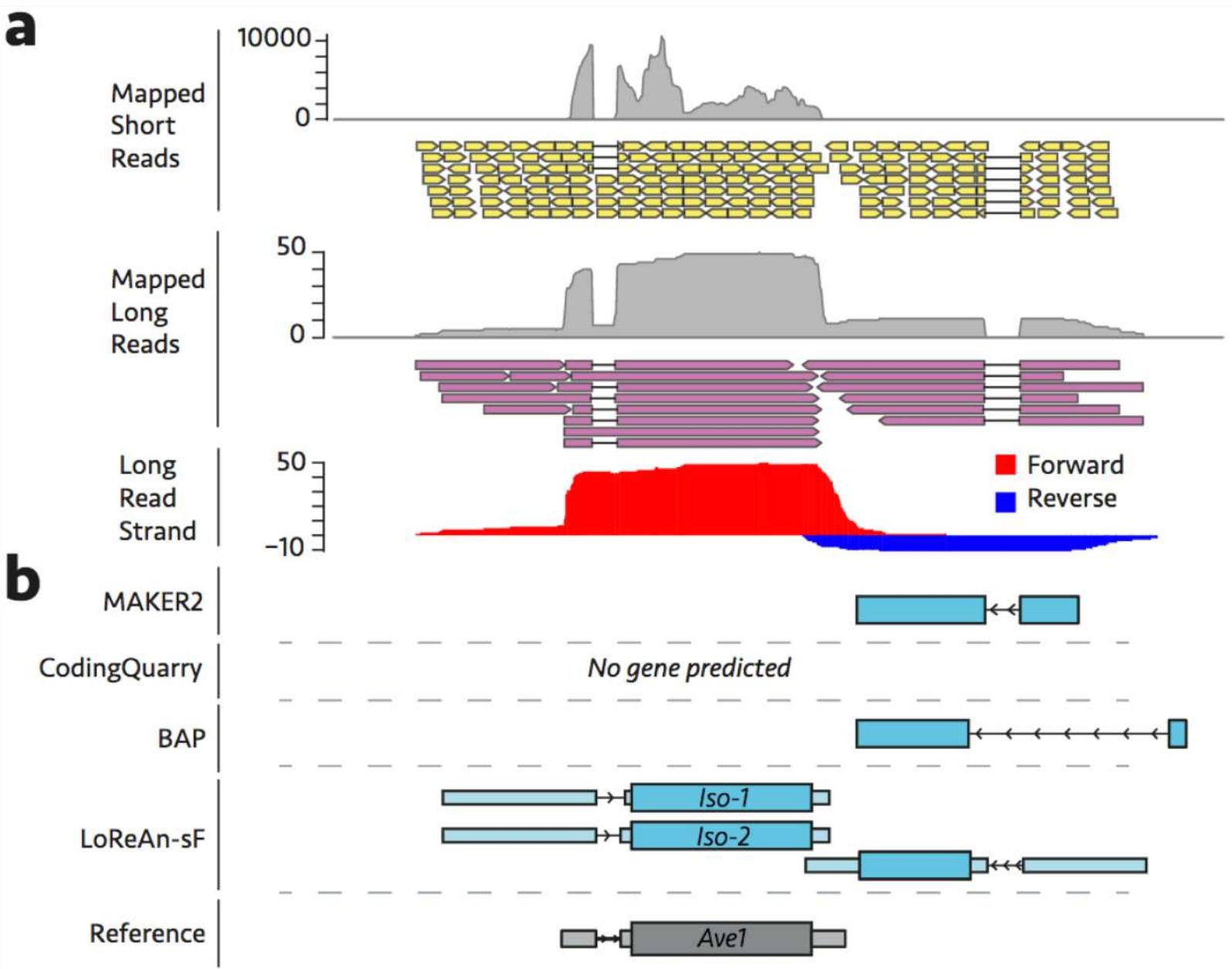
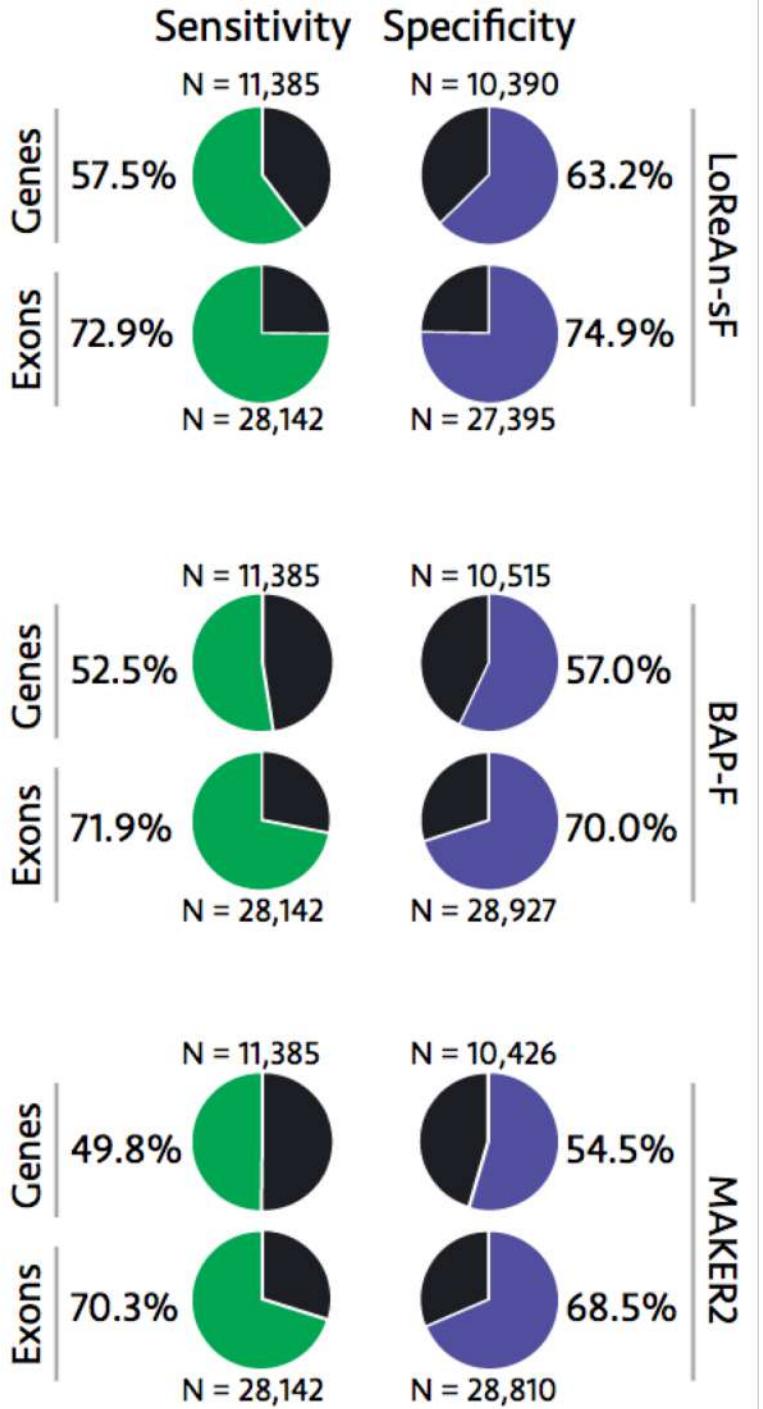
² Excluding untranslated regions.

Case study: Annotation using long reads

**Long Read Annotation (LoReAn): automated eukaryotic genome annotation
based on long-read cDNA sequencing**

David E. Cook^{1,2†}, Jose Espejo Valle-Inclan^{1,3†}, Alice Pajoro^{4,5}, Hanna Rovenich^{1,6}, Bart PHJ Thomma^{1\$*}, and Luigi Faino^{1,7\$*}





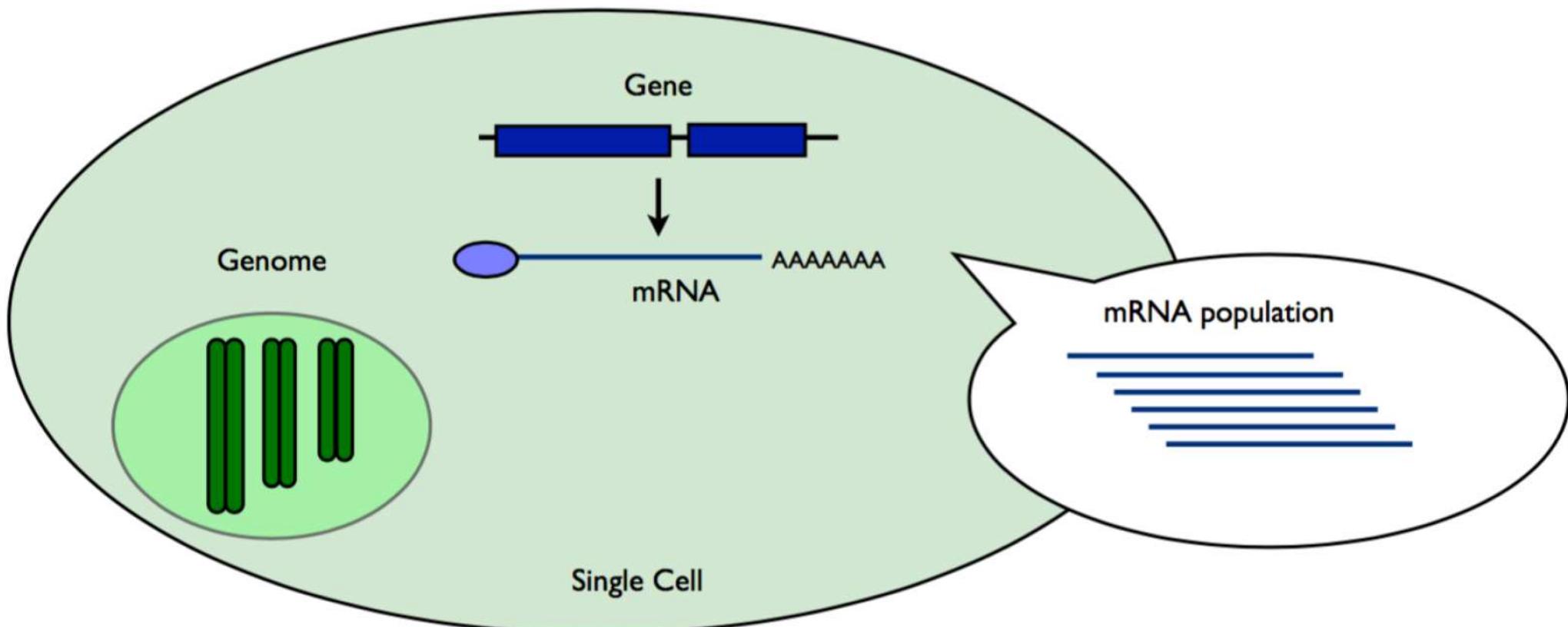
Differential expression

Types of experiments

Transcriptome Complexity:

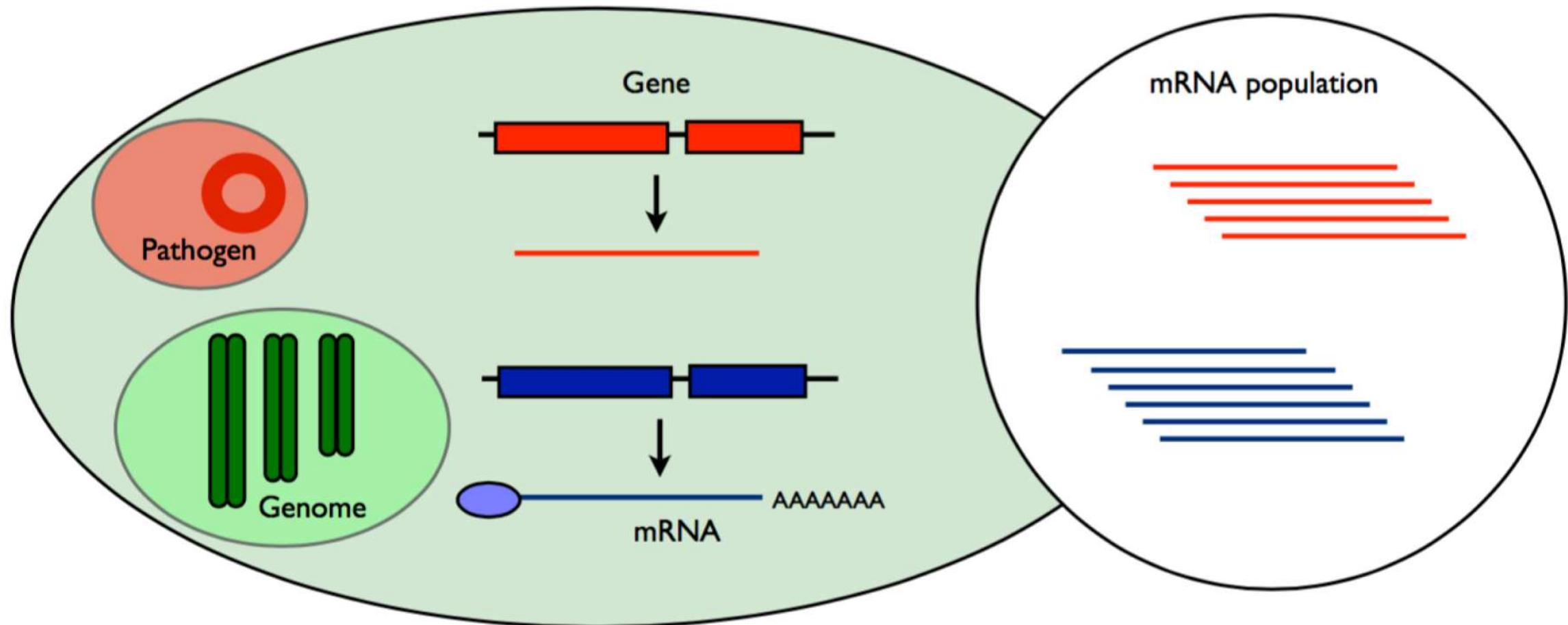
Simple System:

One Genome => Gene 1 copy => Single mRNA



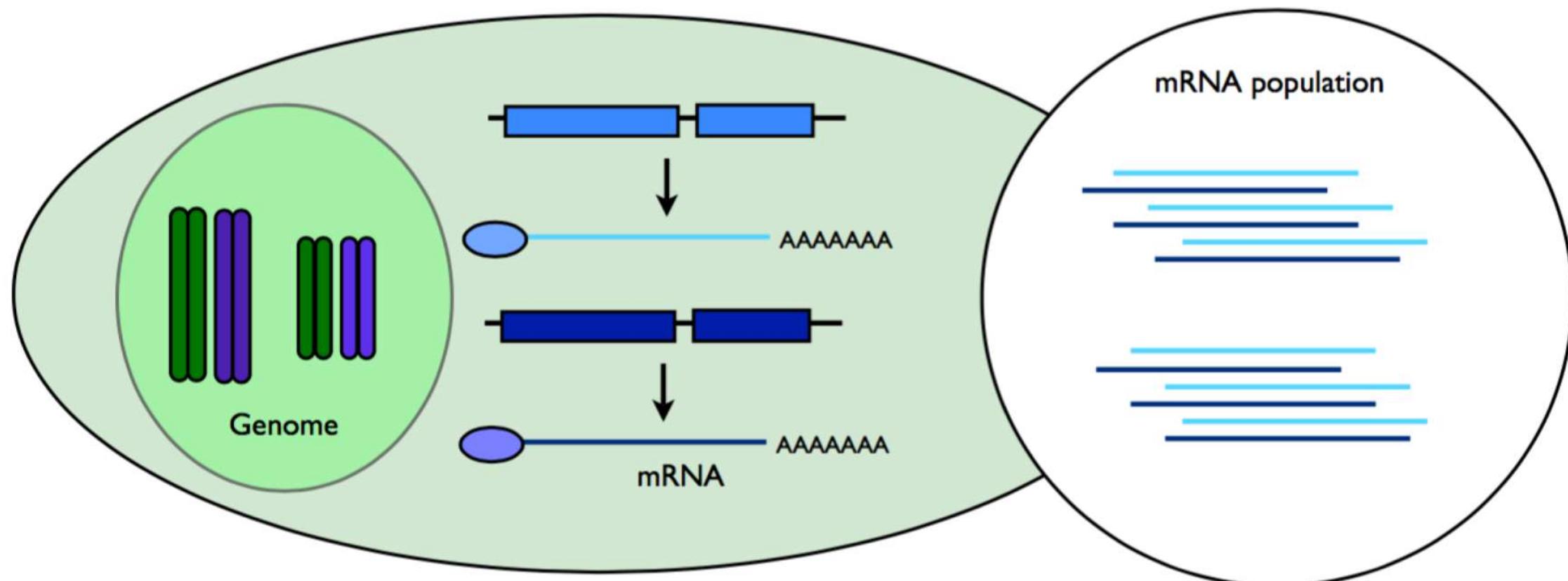
How many species we are analyzing ?

- 1) Problems to isolate a single species (rhizosphere)
- 2) Species interaction study (plant-pathogen)



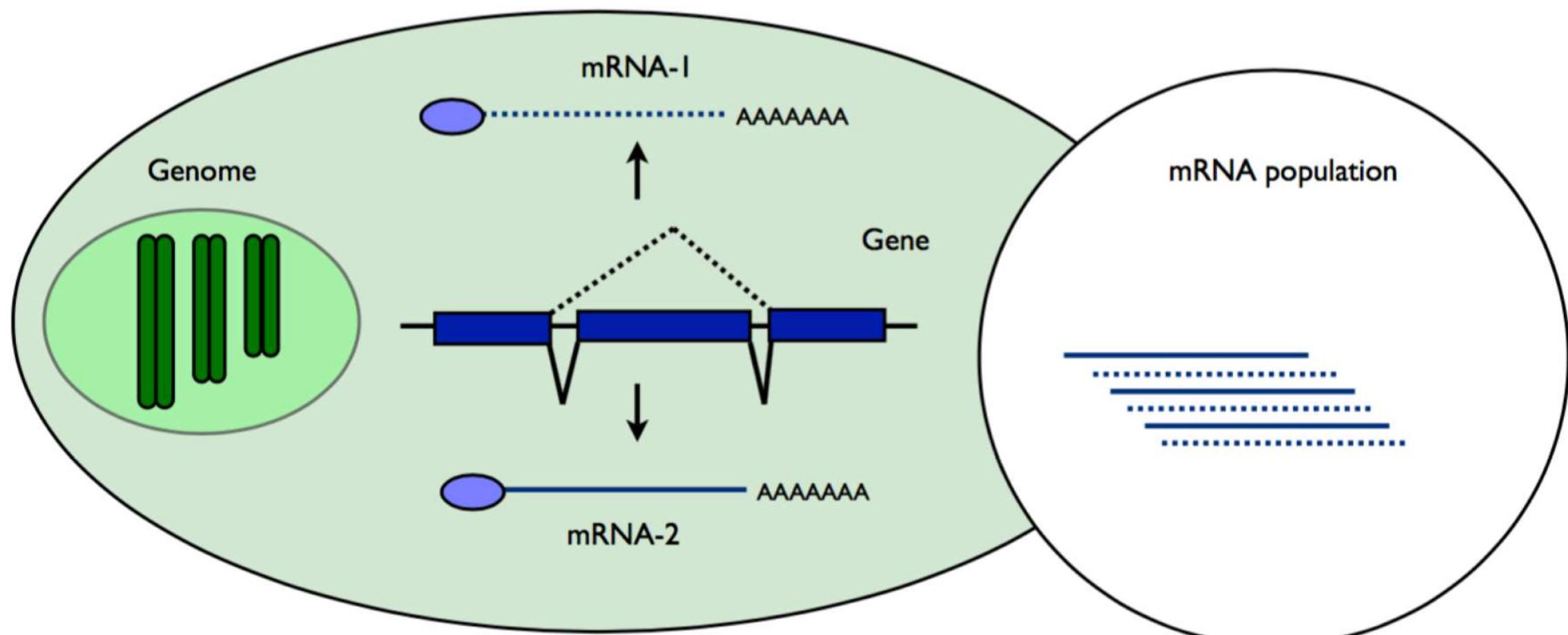
How many possible alleles we expect per gene ?

- 1) Polyploids (autopolyploids, allopolyploids).
- 2) Heterozygosity
- 3) Complex Gene Families (tandem duplications)



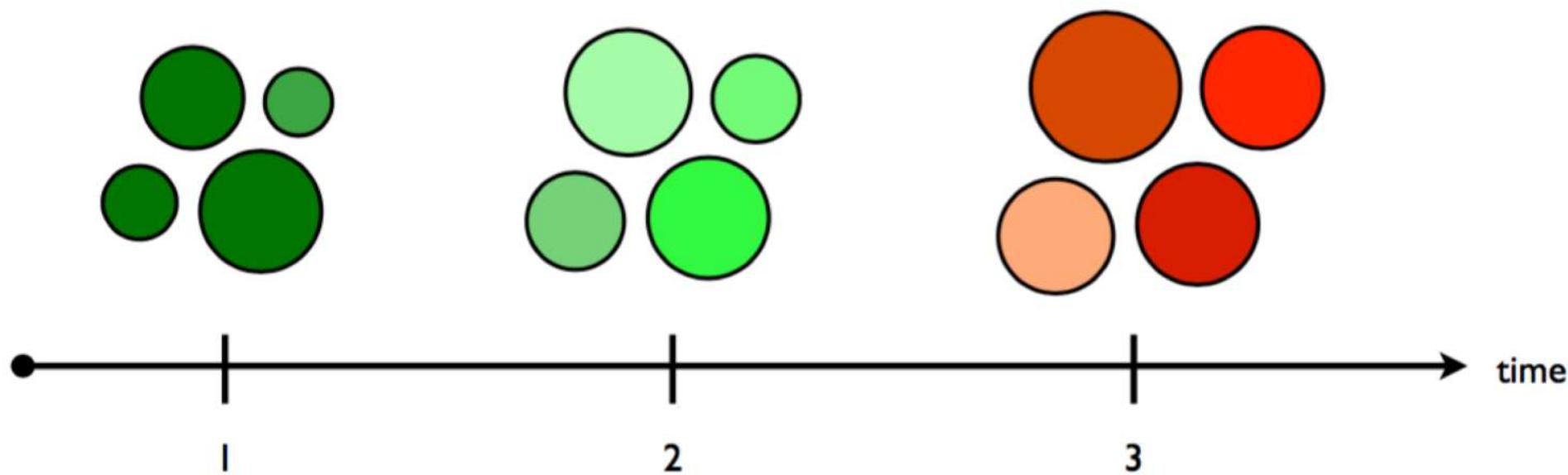
How many isoforms we expect for each allele ?

1) Alternative splicings



Is the study performed at different time points?

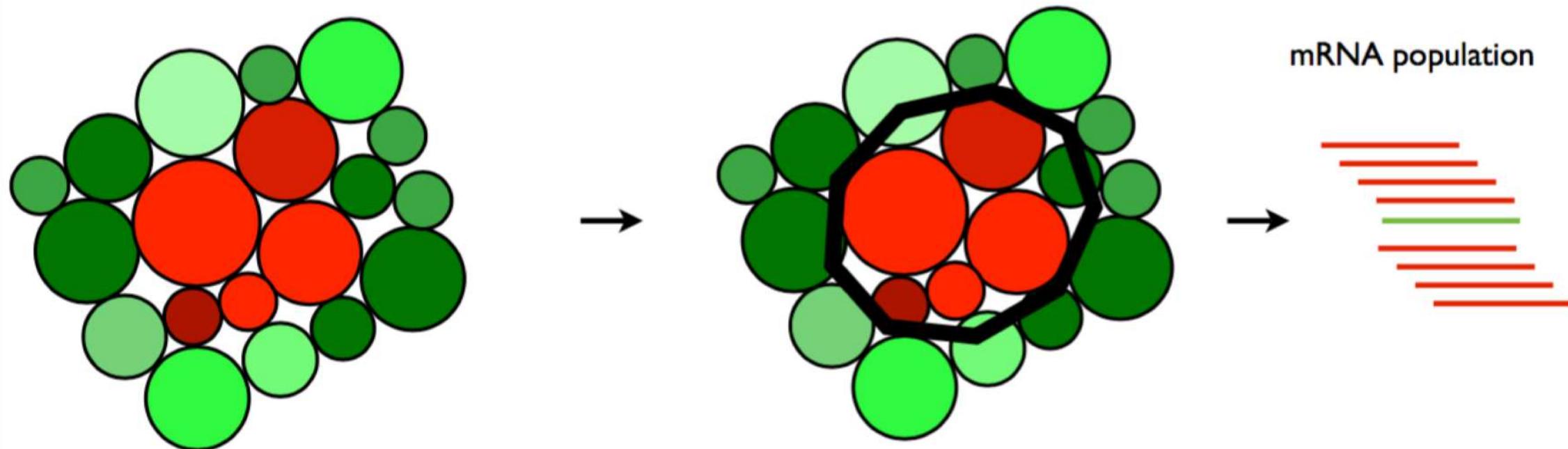
- 1) Developmental stages (difficult to select the same)**
- 2) Response to a treatment**



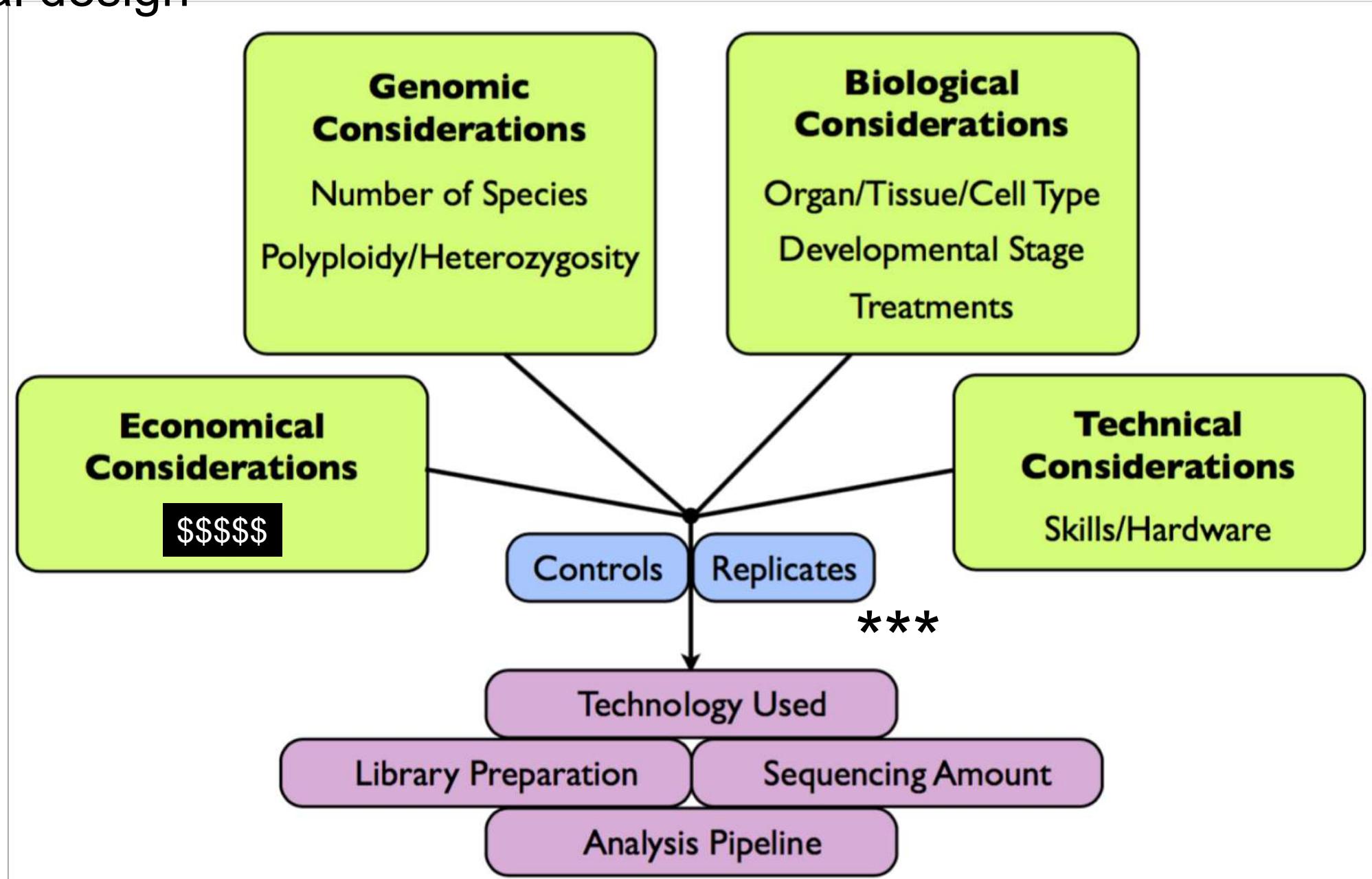
Is the study performed with different parts?

- 1) Organ specific**
- 2) Tissue/Cell type specific**

(Laser Capture Microdissection, LCM)



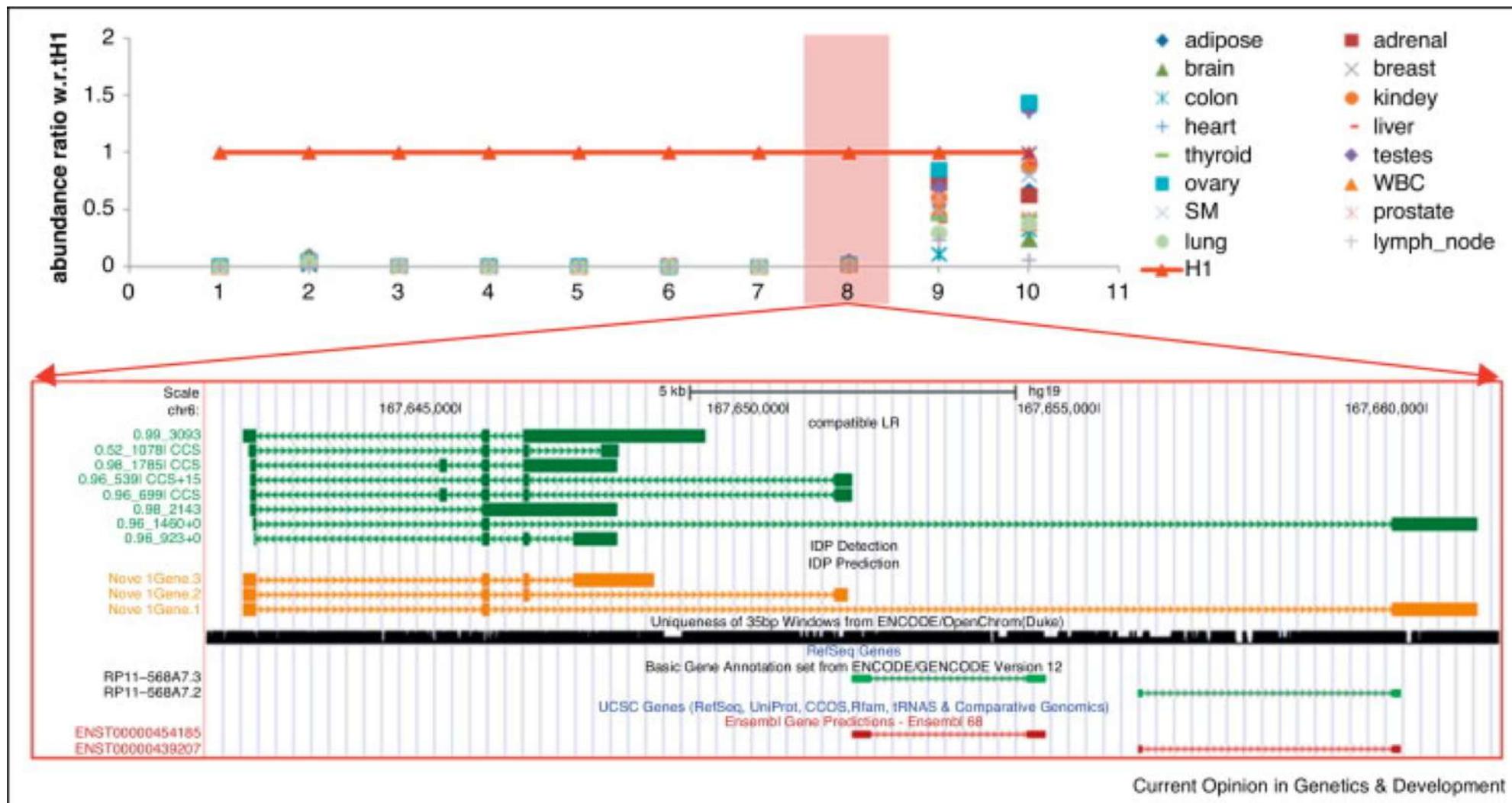
Experimental design



How many reads are enough?

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

For isoform discovery, longer sequences are better



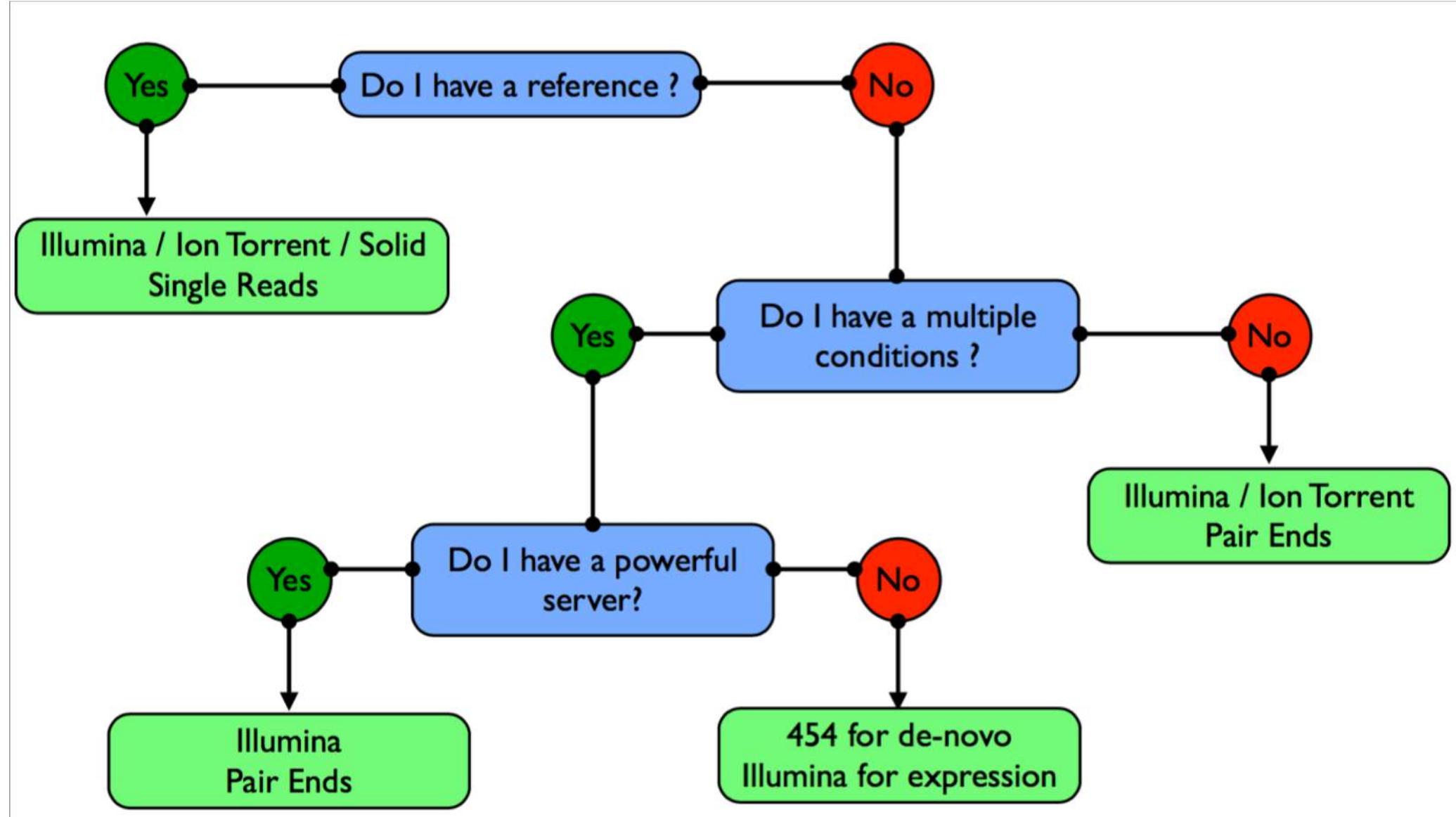
Selecting the right technology

Obsoleted

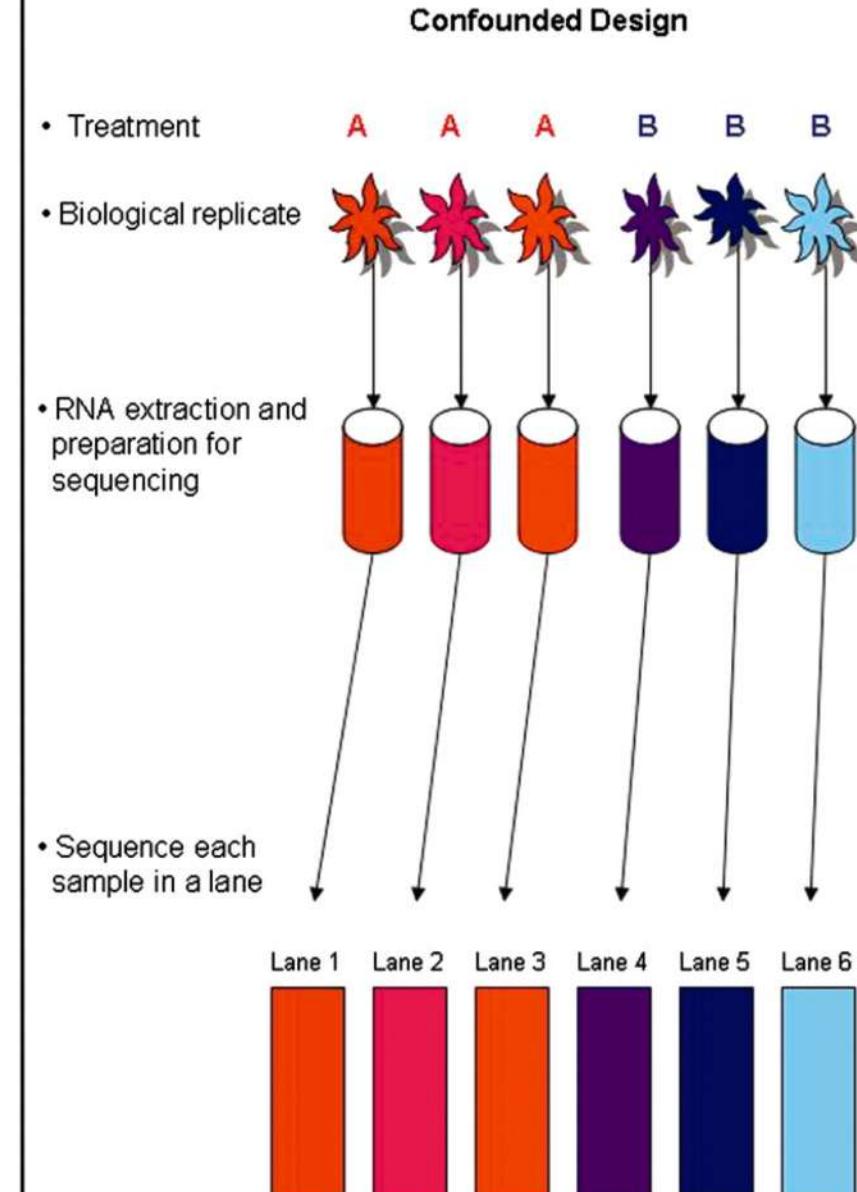
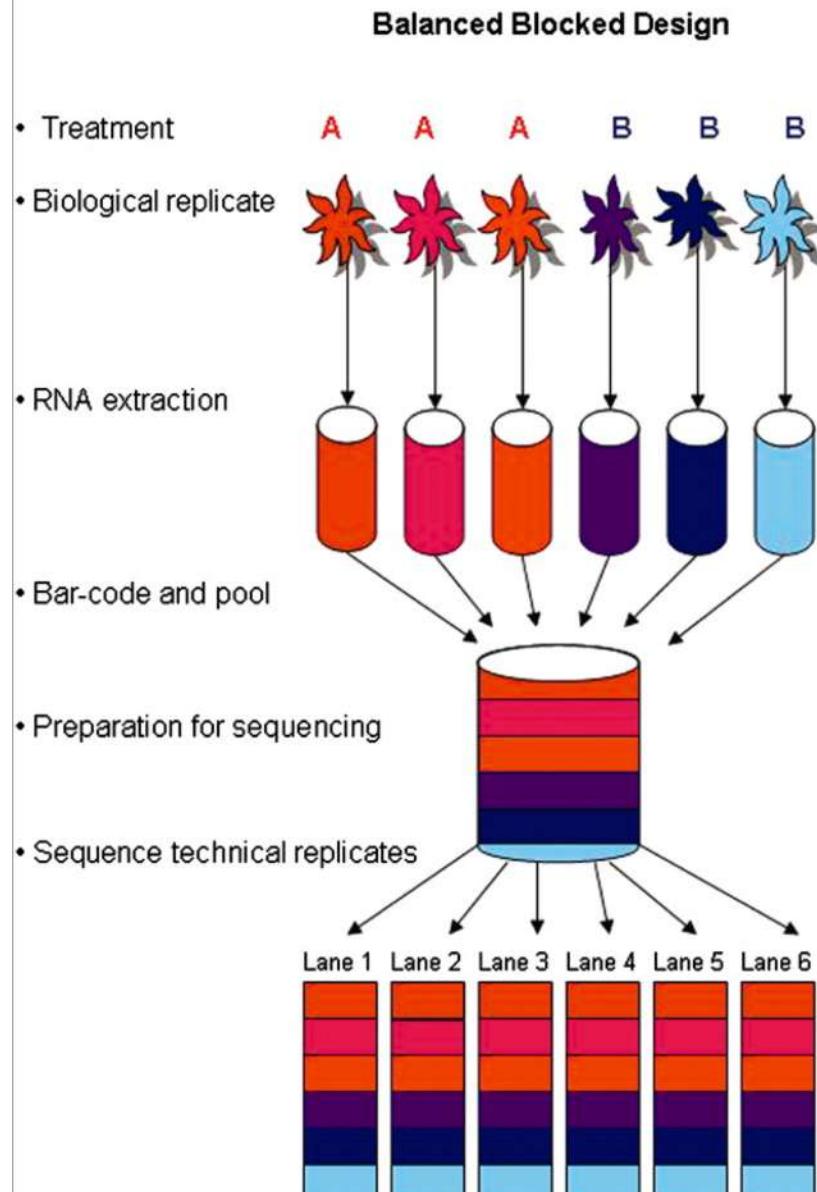
	Run Time	Sequence Length	Reads/Run	Total nucleotides sequenced per run
Capillary Sequencing (ABI37000)	~2.5 h	800 bp	386	0.308 Mb
454 Pyrosequencing (GS FLX Titanium XL+)	~23 h	700 bp	1,000,000	700 Mb (0.7 Gb)
Illumina (HiSeq 2500)	264 h / 27 h (11 days)	2 x 100 bp 2 x 150 bp	2 x 3,000,000,000 2 x 600,000,000	600,000 / 120,000 Mb (600 / 120 Gb)
Illumina (MiSeq)	39 h	2 x 250 bp	2 x 17,000,000	8,500 Mb (8.5 Gb)
SOLID (5500xl system)	48 h (2 days)	75 bp	400,000,000	30,000 Mb (30 Gb)
Ion Torrent (Ion Proton I)	2 h	100 bp	100,000,000	10,000 Mb (10 Gb)
PacBio (PacBioRS)	1.5 h	~3,000 bp	25,000	100 Mb (0.1 Gb)



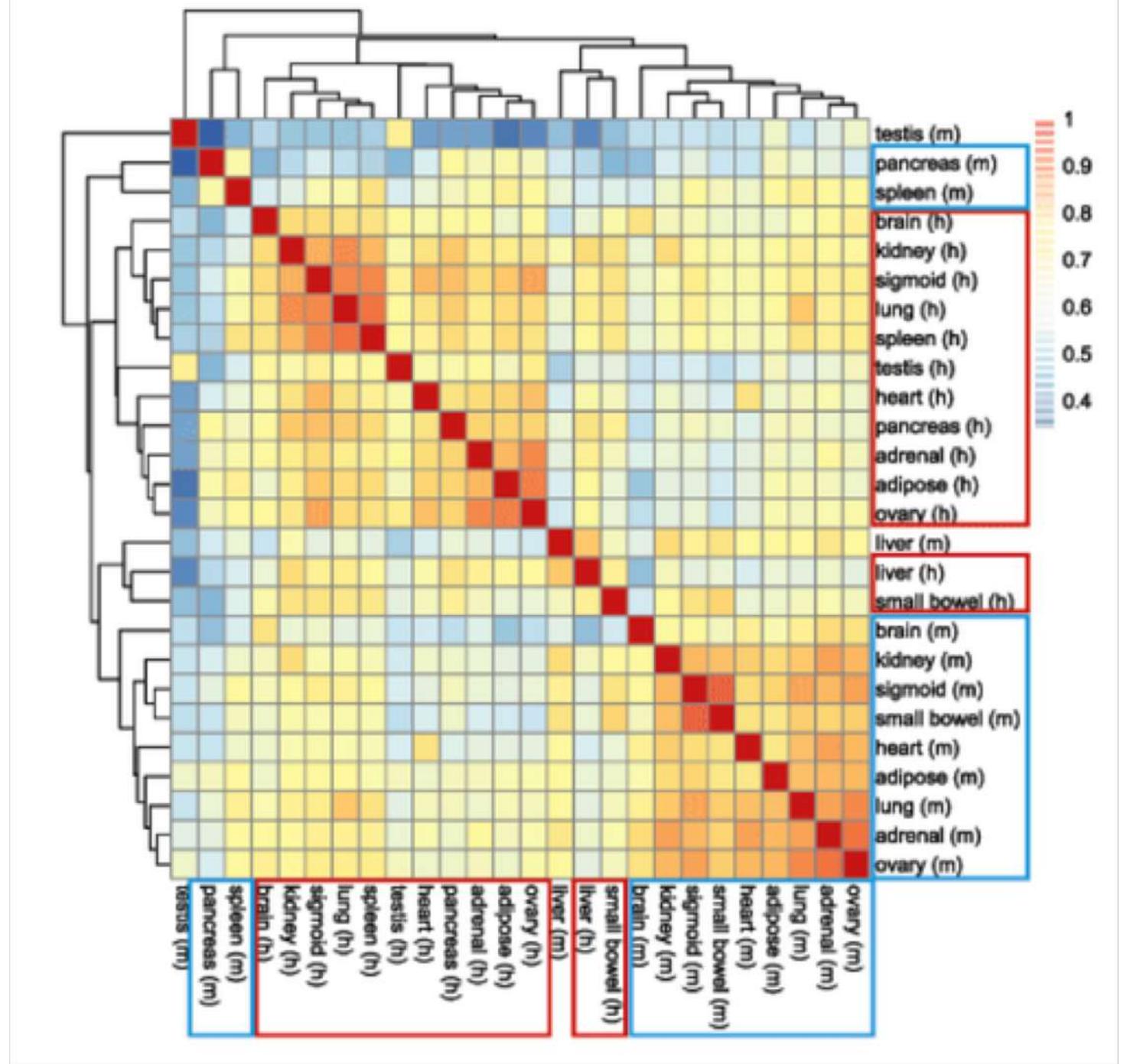
Selecting the right technology (depends on your purpose)



You need to design experiment carefully



Example of batch effect:



Example of batch effect:

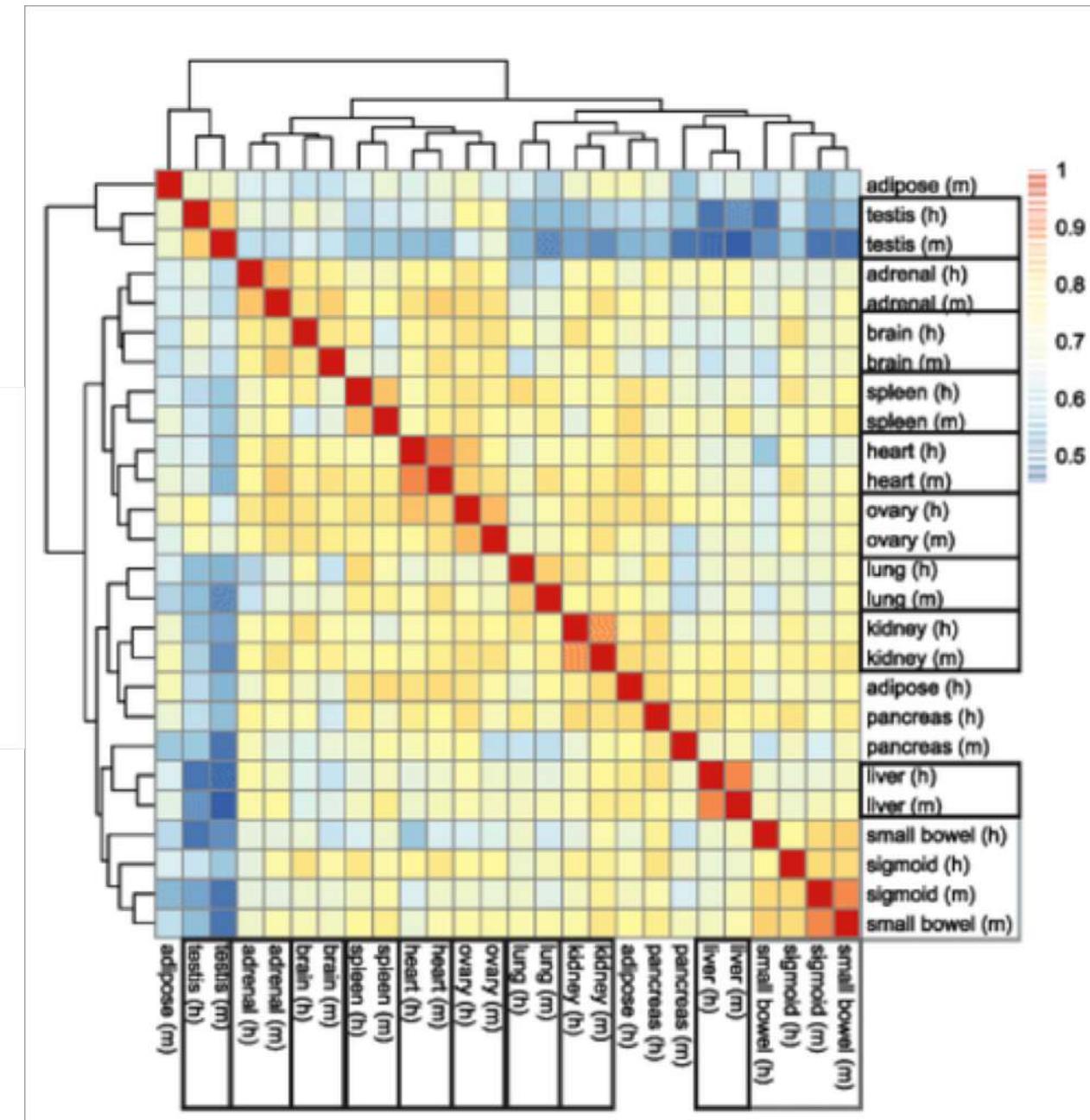


Yoav Gilad
@Y_Gilad

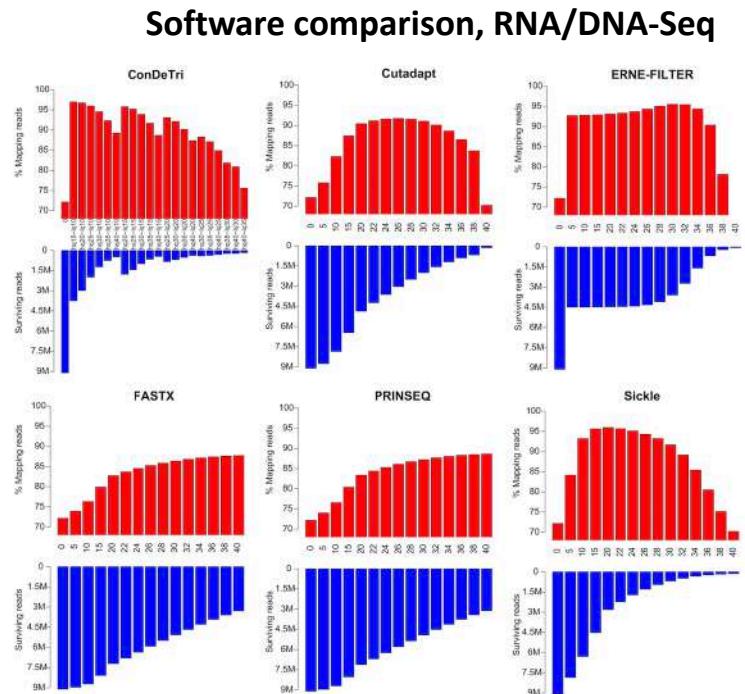


Following

We reanalyzed the data from
[pnas.org/content/111/48...](http://pnas.org/content/111/48/) and found the
following:



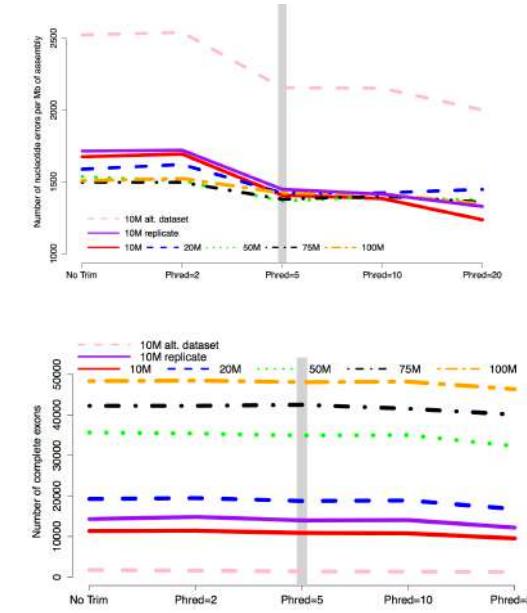
Is trimming beneficial?



"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



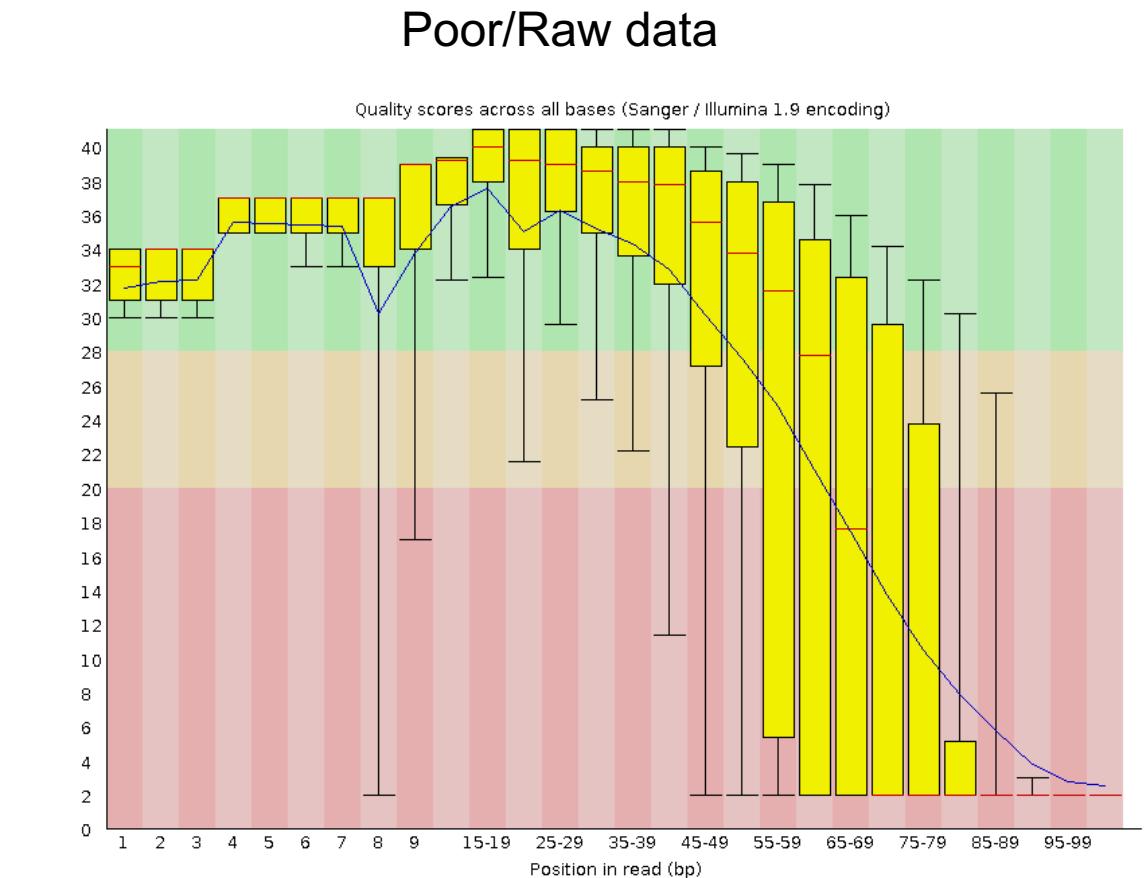
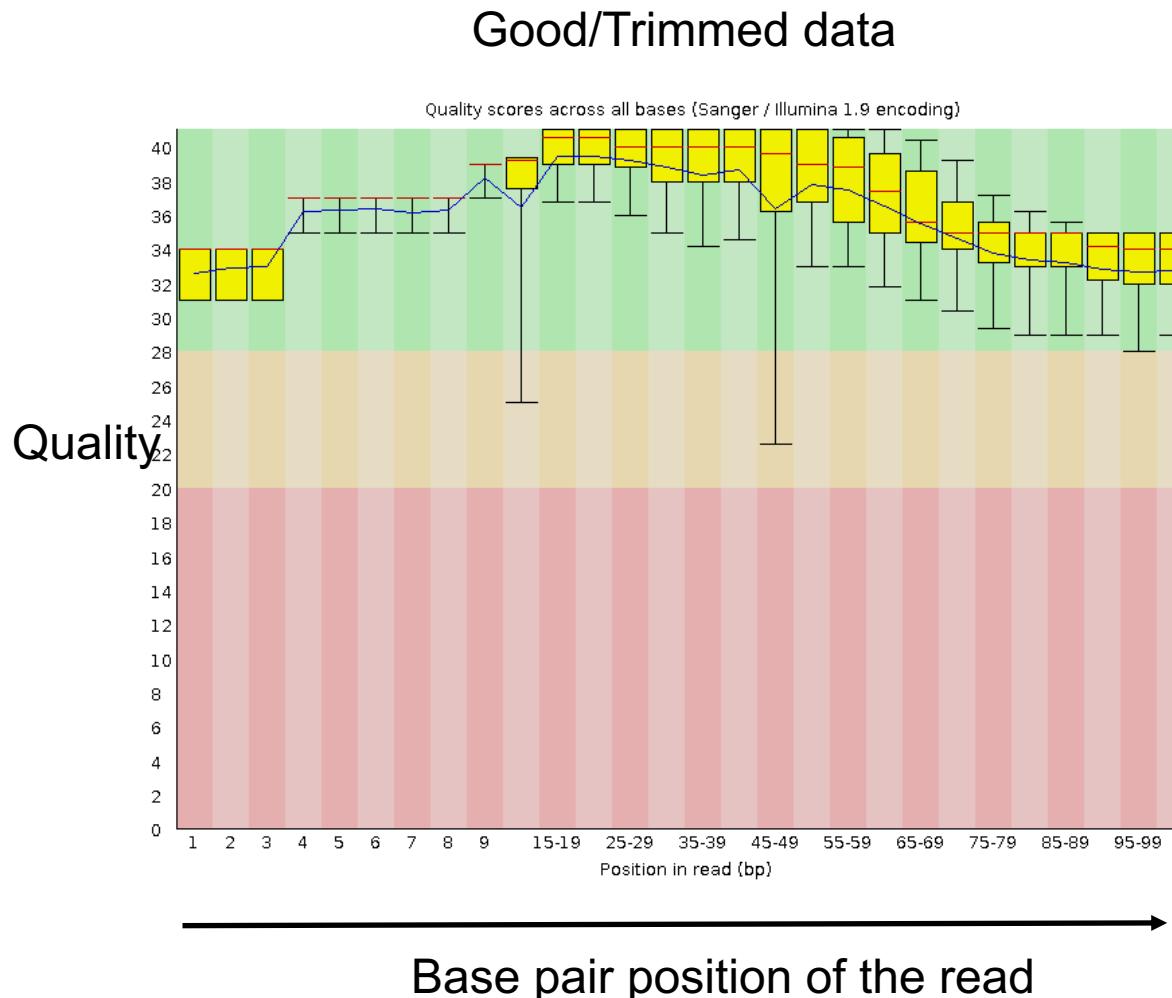
Erroneous bases
in assembly

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."

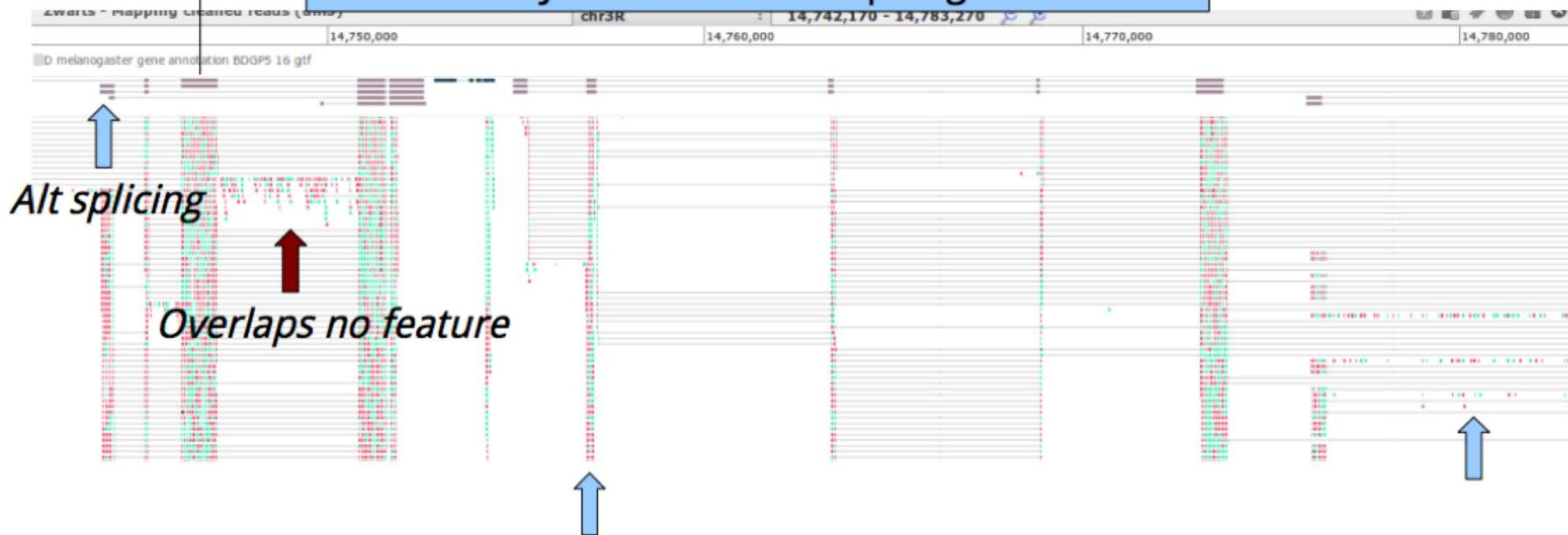
MacManes MD (2013)
On the optimal trimming of high-throughput mRNAseq data doi: 10.1101/000422

My take: only trim data when you have to



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

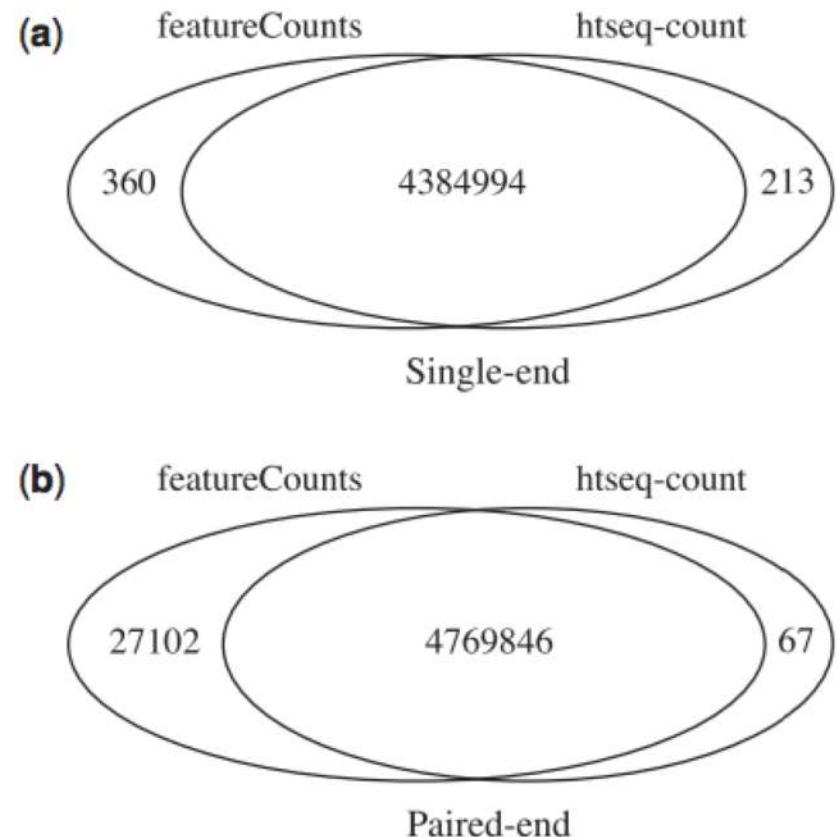


Table 3. Performance with RNA-seq reads simulated from an annotated assembly of the Budgerigar genome

Methods	Number of reads	Time (mins)	Memory (MB)
<i>featureCounts</i>	7 924 065	0.6	15
<i>summarizeOverlaps</i> (whole genome at once)	7 924 065	12.6	2400
<i>summarizeOverlaps</i> (by scaffold)	7 924 065	53.3	262
<i>htseq-count</i>	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.

Some QC is needed

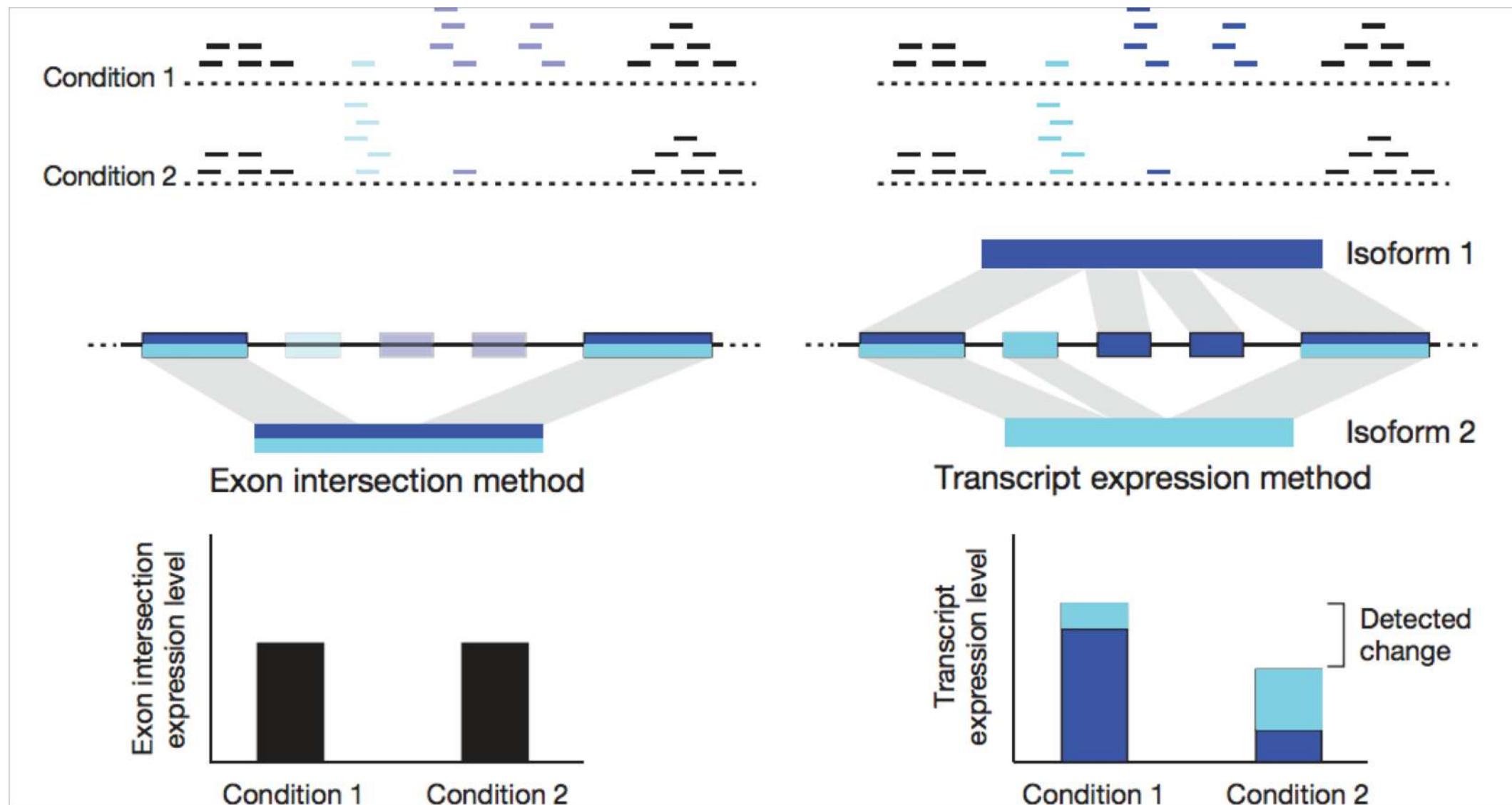
- Which genes are highly counted?
- Any samples with a lot of missing/no count genes?
- Anything that may affect sample counts (like batch effect?)

Ambiguity in counting

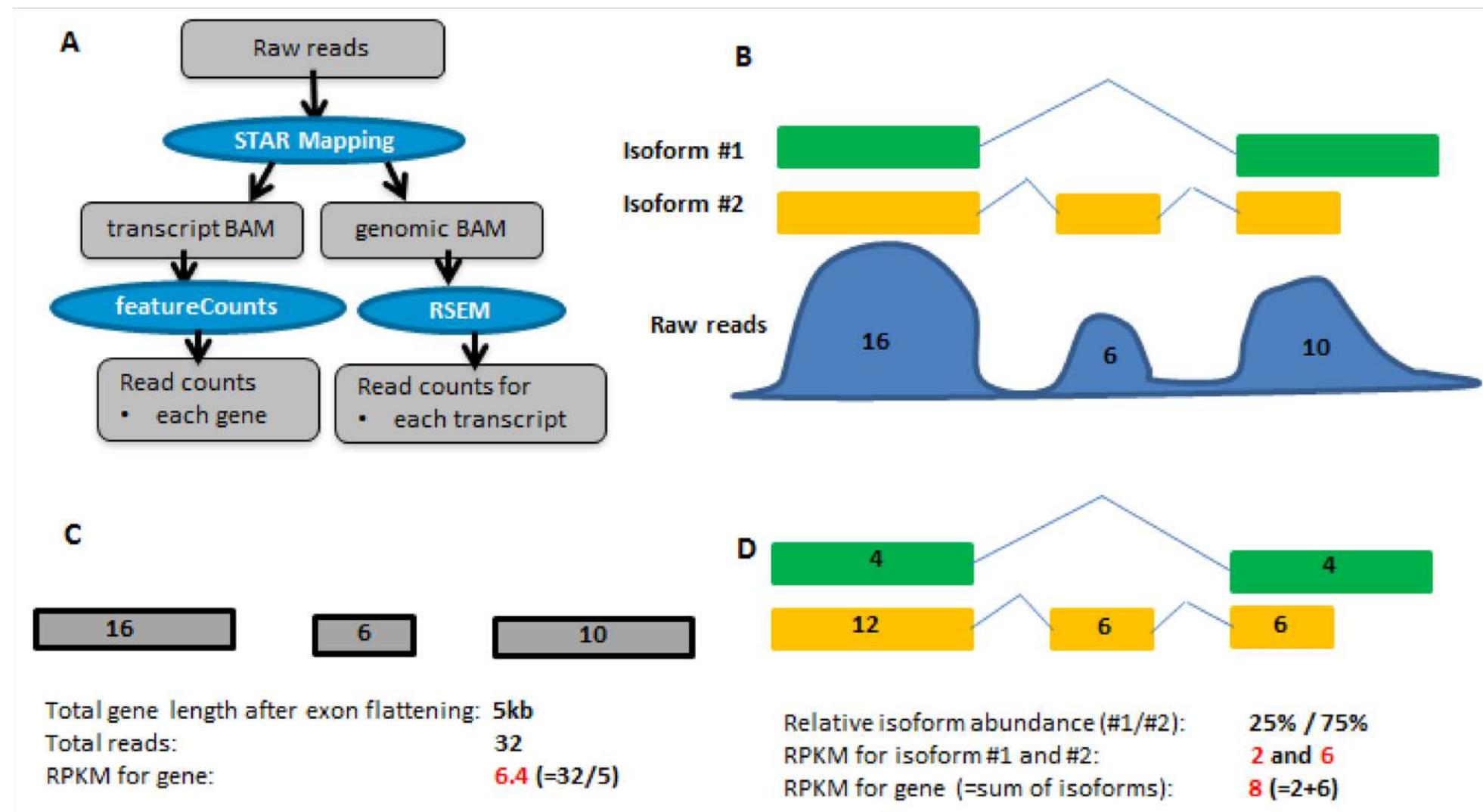
We focus on the **gene level**: merge all counts over different **isoforms** into one, taking into account:

- 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



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

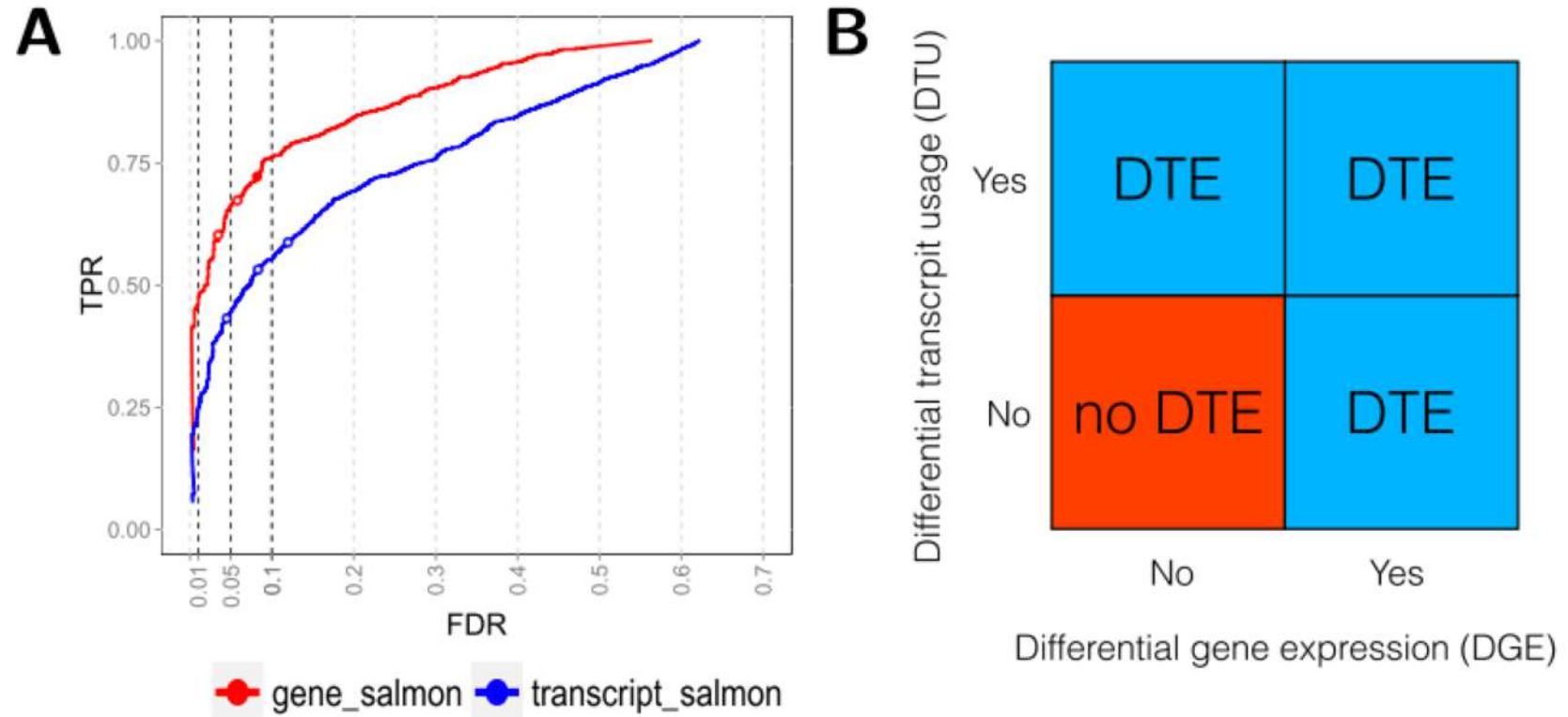


Figure 2 (sim2). A: DTE detection performance on transcript- and gene-level, using *edgeR* applied to transcript-level estimated counts from *Salmon*. The statistical analysis was performed on transcript level and aggregated for each gene using the *perGeneQValue* function from the *DEXSeq* R package; 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).

So use isoform or not?

Modern RNA-seq differential expression analyses: transcript-level or gene-level

Posted by: RNA-Seq Blog in Presentations ⌂ February 11, 2016 ⌂ 1,733 Views

Modern RNA-seq differential expression analyses: transcript-level or gene-level

SIB Virtual CB Seminar Series, 3 Feb 2016, Lausanne

University of Zurich
URPP Systems Biology / Functional Genomics

SIB
Swiss Institute of Bioinformatics

Modern differential analyses for RNA-seq: transcript-level or gene-level

Mark D. Robinson
Institute of Molecular Life Sciences,
University of Zurich

@markrobinsonca

figshare

Share

Download (6.53 MB)

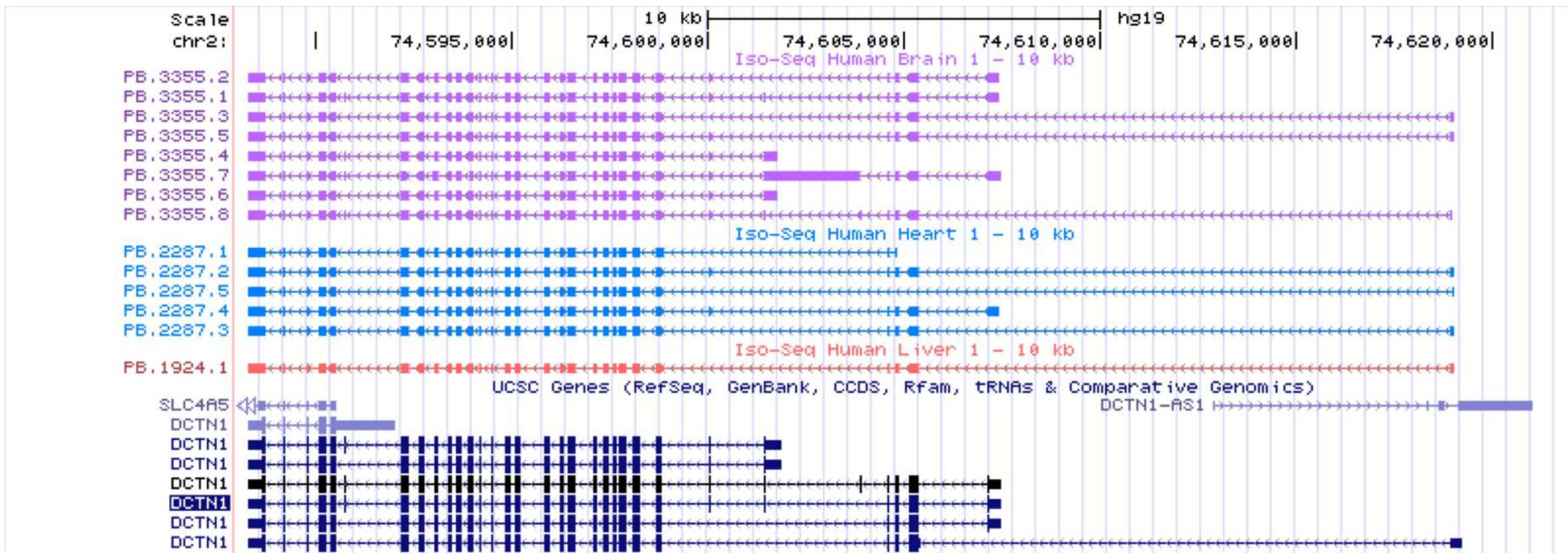
“There is no crisis; the impact of union vs. transcript counting in many datasets is rather small”

“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

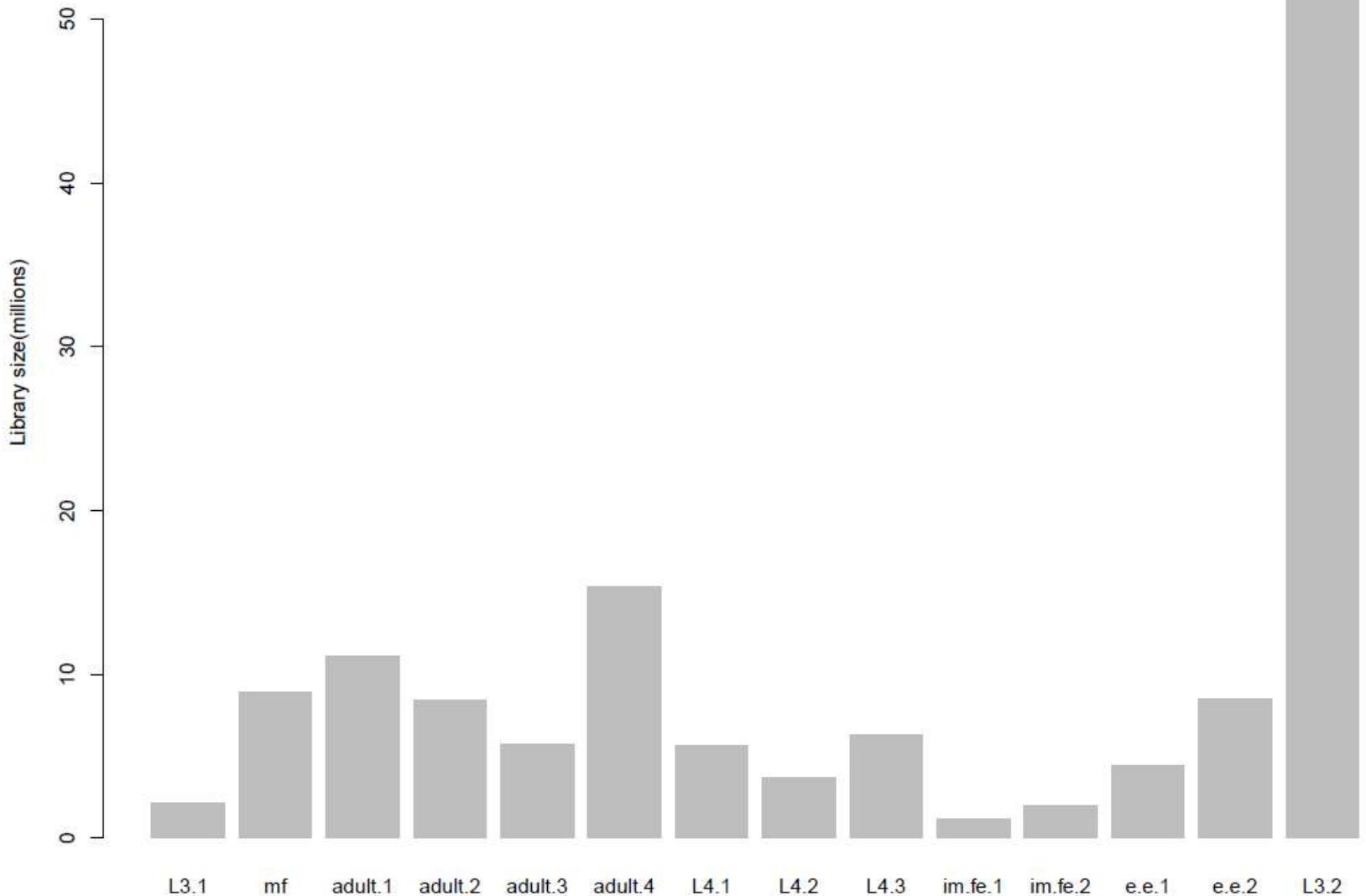


This is the bit we care about!

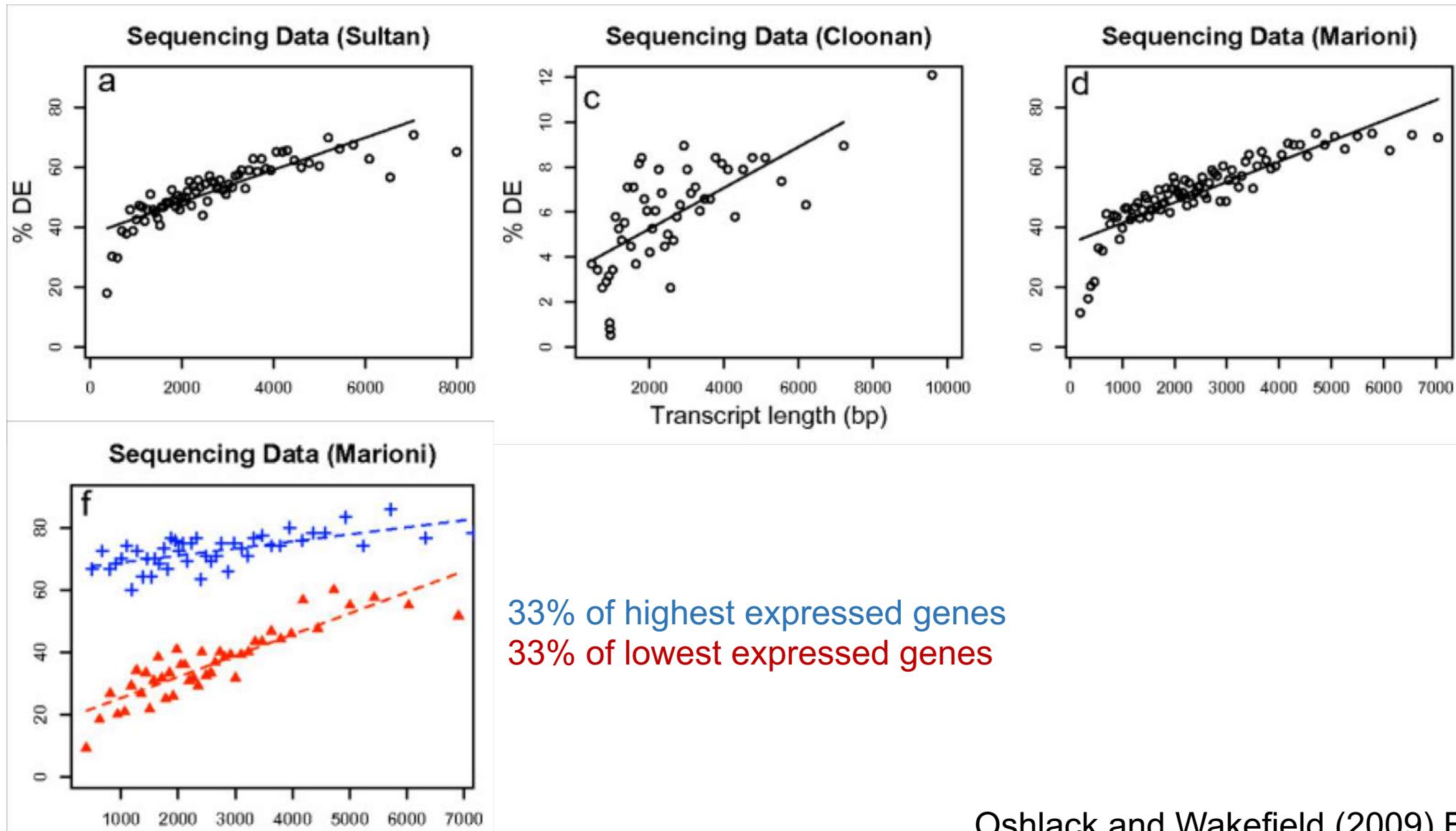


Counts of the gene depends on **expression** ,transcript length
,sequencing depth and simply chance

Higher sequencing depth equals more counts



Counts are proportional to the transcript length x mRNA expression level



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):** (Law et al 2013)
 - **Aiming to:** stabilize variance; remove dependence of variance on the mean

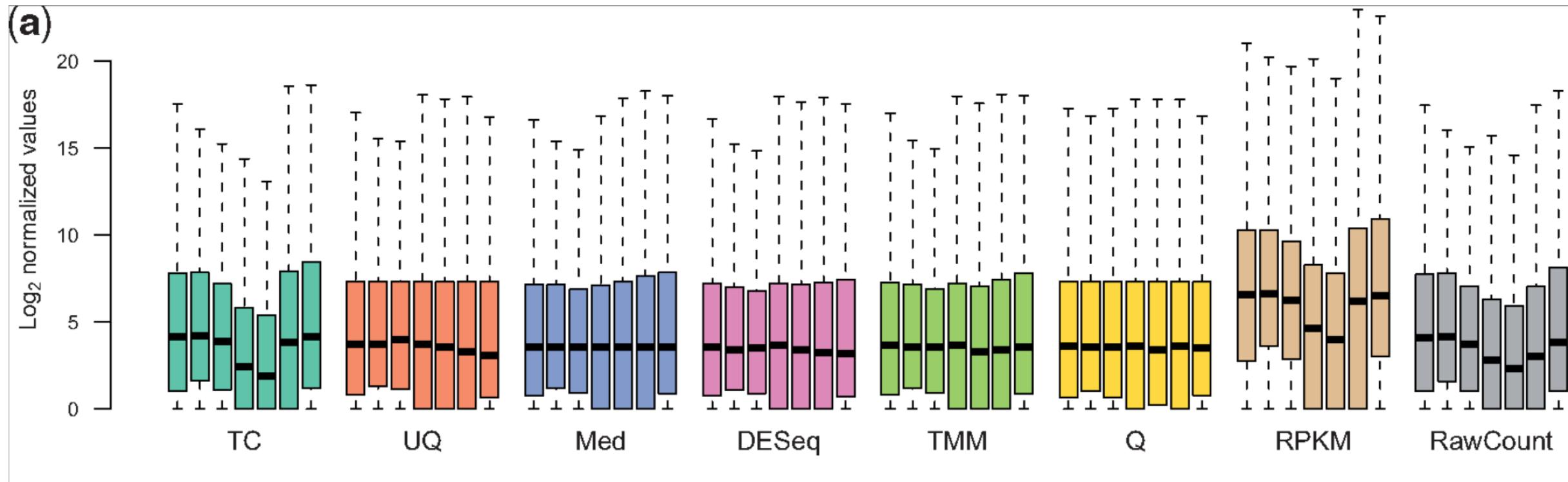
Optimal Scaling of Digital Transcriptomes

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

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

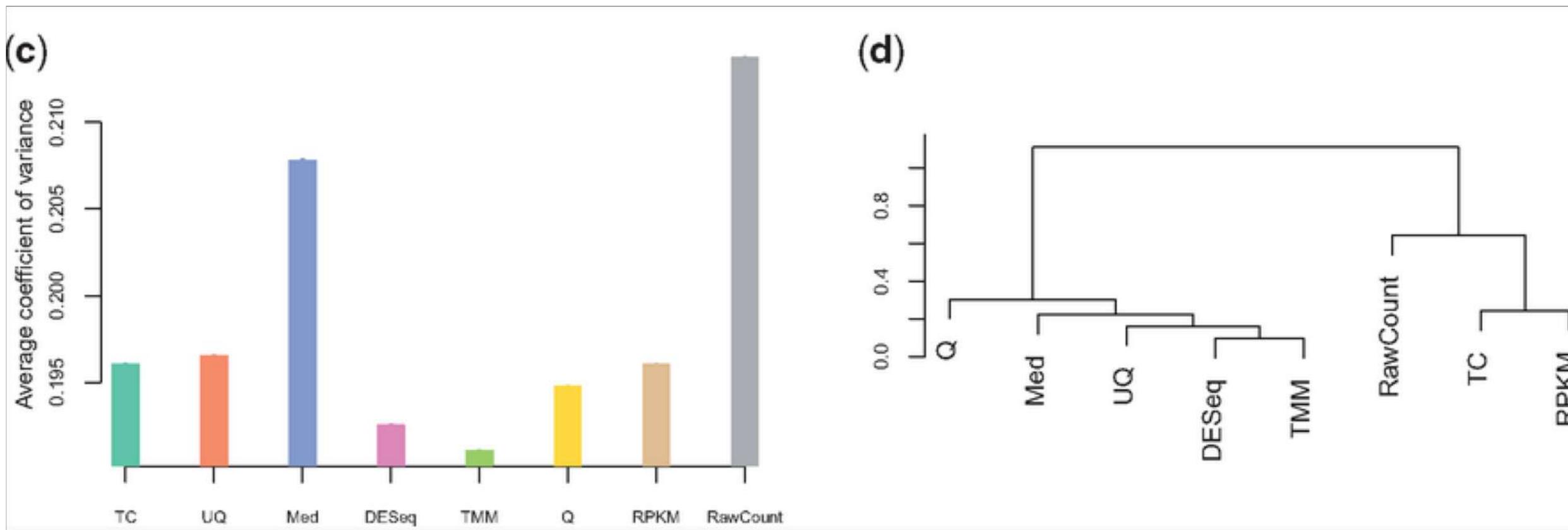
RPKM shouldn't be used for between sample comparisons

Boxplots of $\log_2(\text{counts} + 1)$ for **seven** replicates in the *M. musculus* data, by normalization method.



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



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.

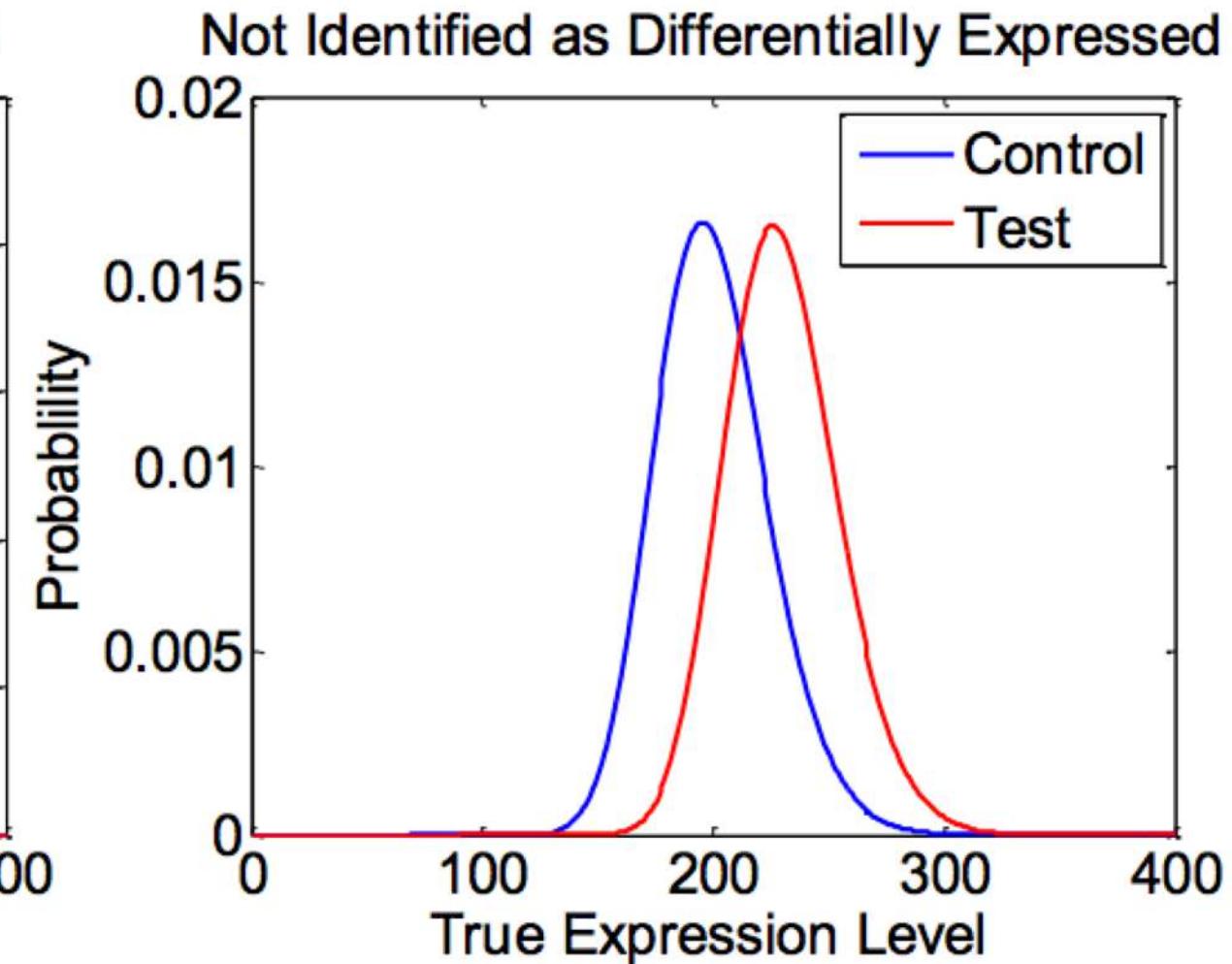
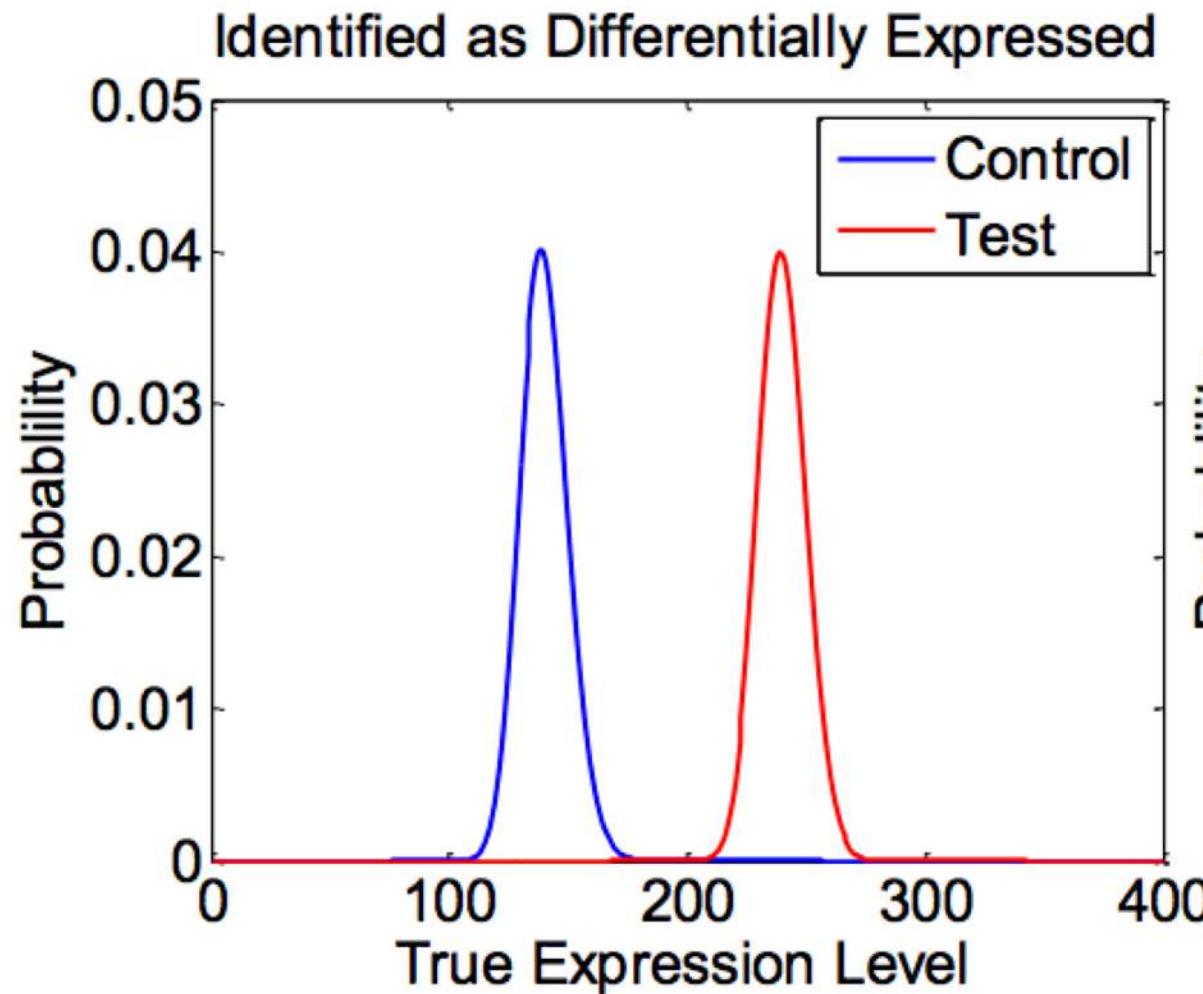
Between sample comparisons

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

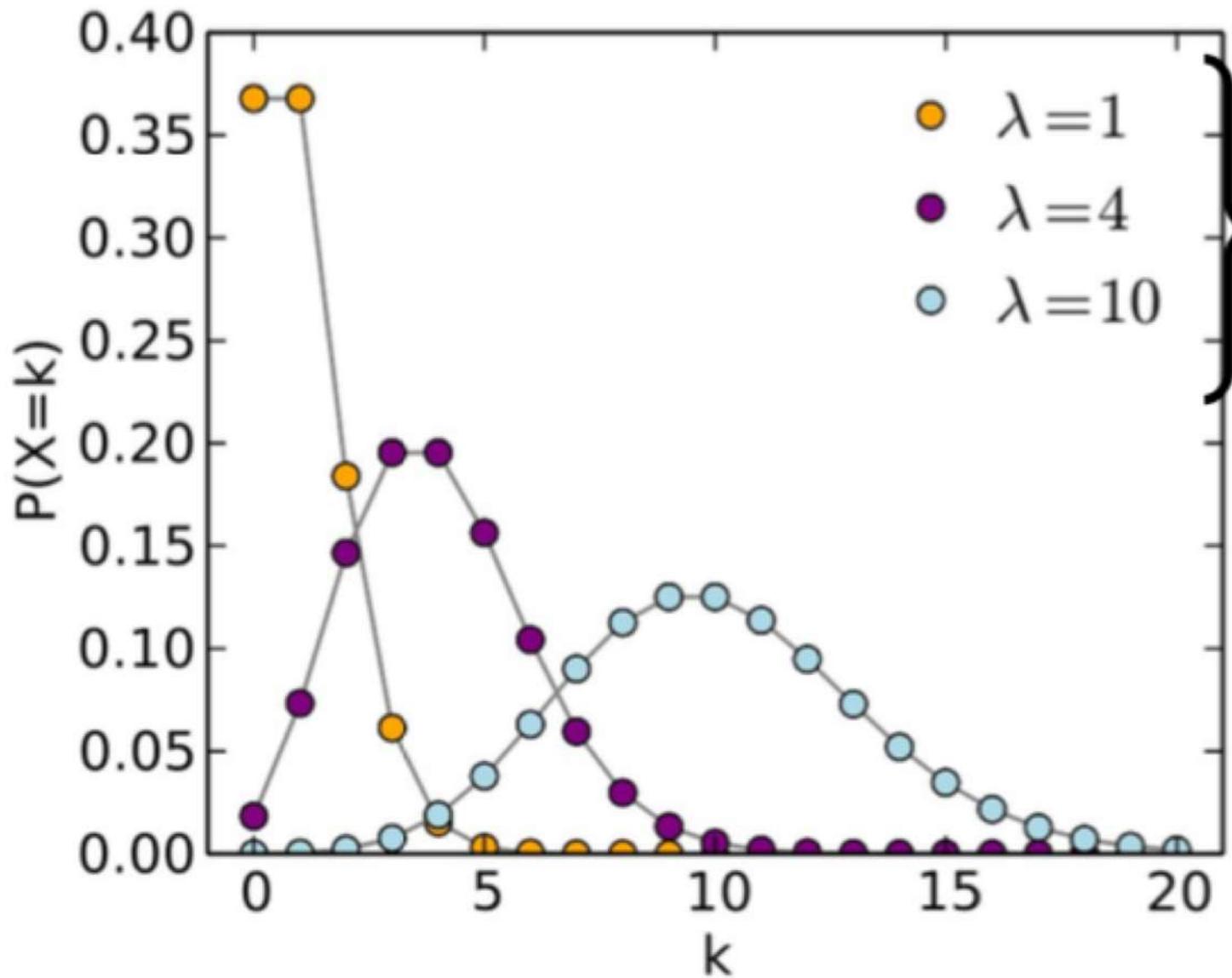
Replicate categories

	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
Sequencing	RNA-seq library construction	Technical
	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

Fitting a distribution for every gene for DE



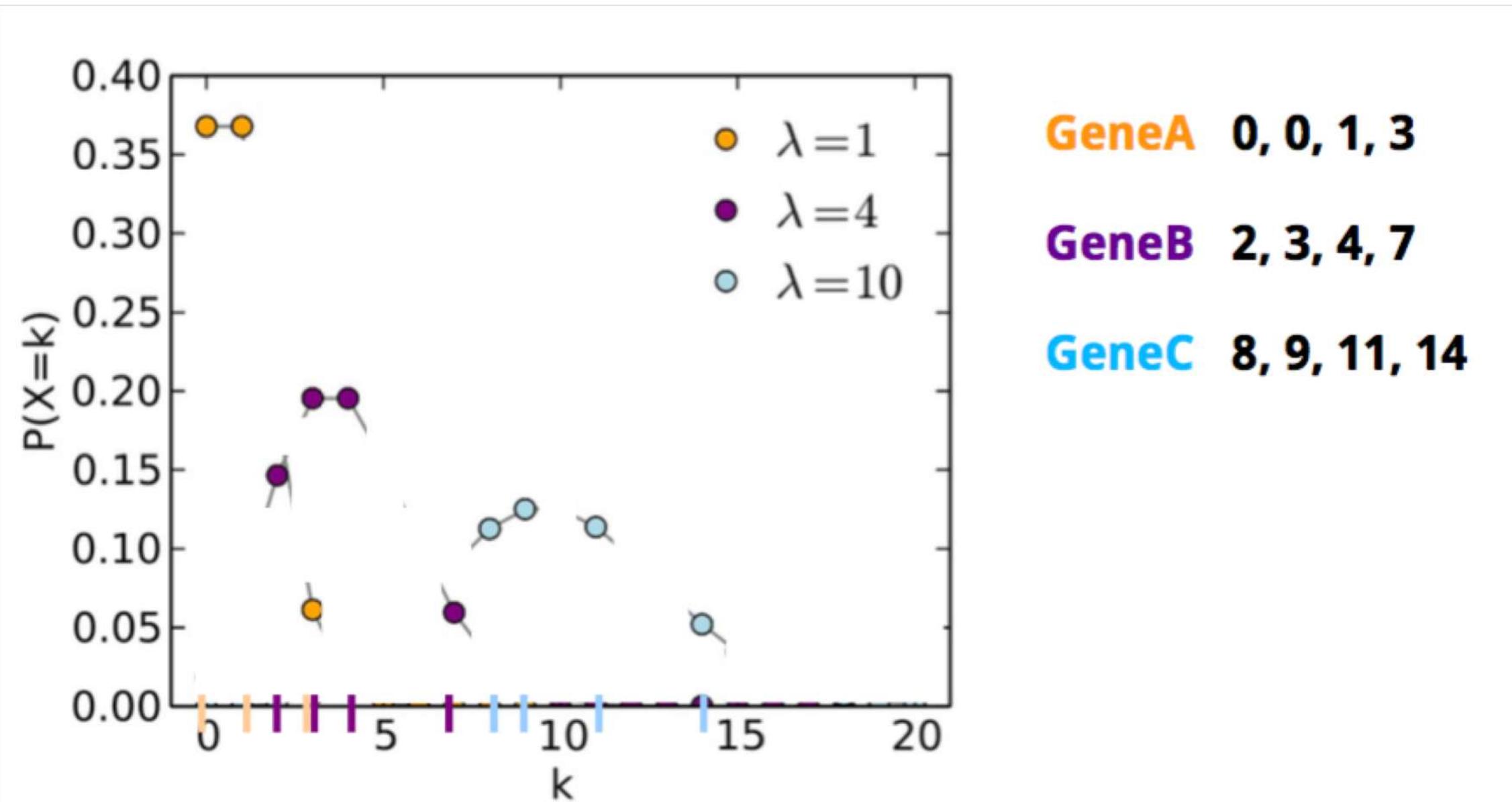
The counts of technical replicates follow a **poisson** distribution (Marioni *et al.*, 2008). So mean = count, variance = count



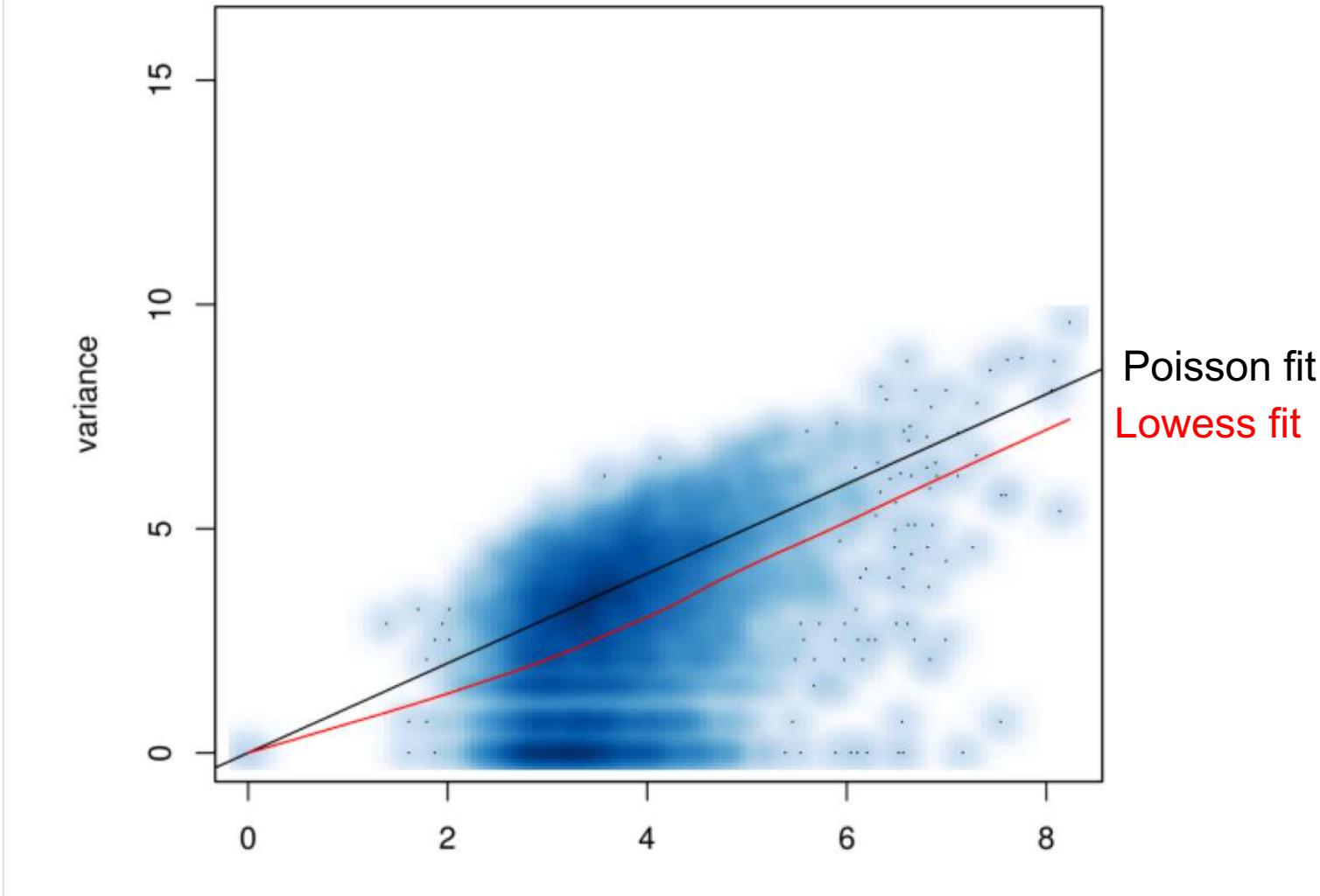
From Wikipedia. Can be 3 different genes, each with their own poisson distribution. Lambda is the mean of the gene's distribution, with a certain number of reads.

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

Four technical replicates

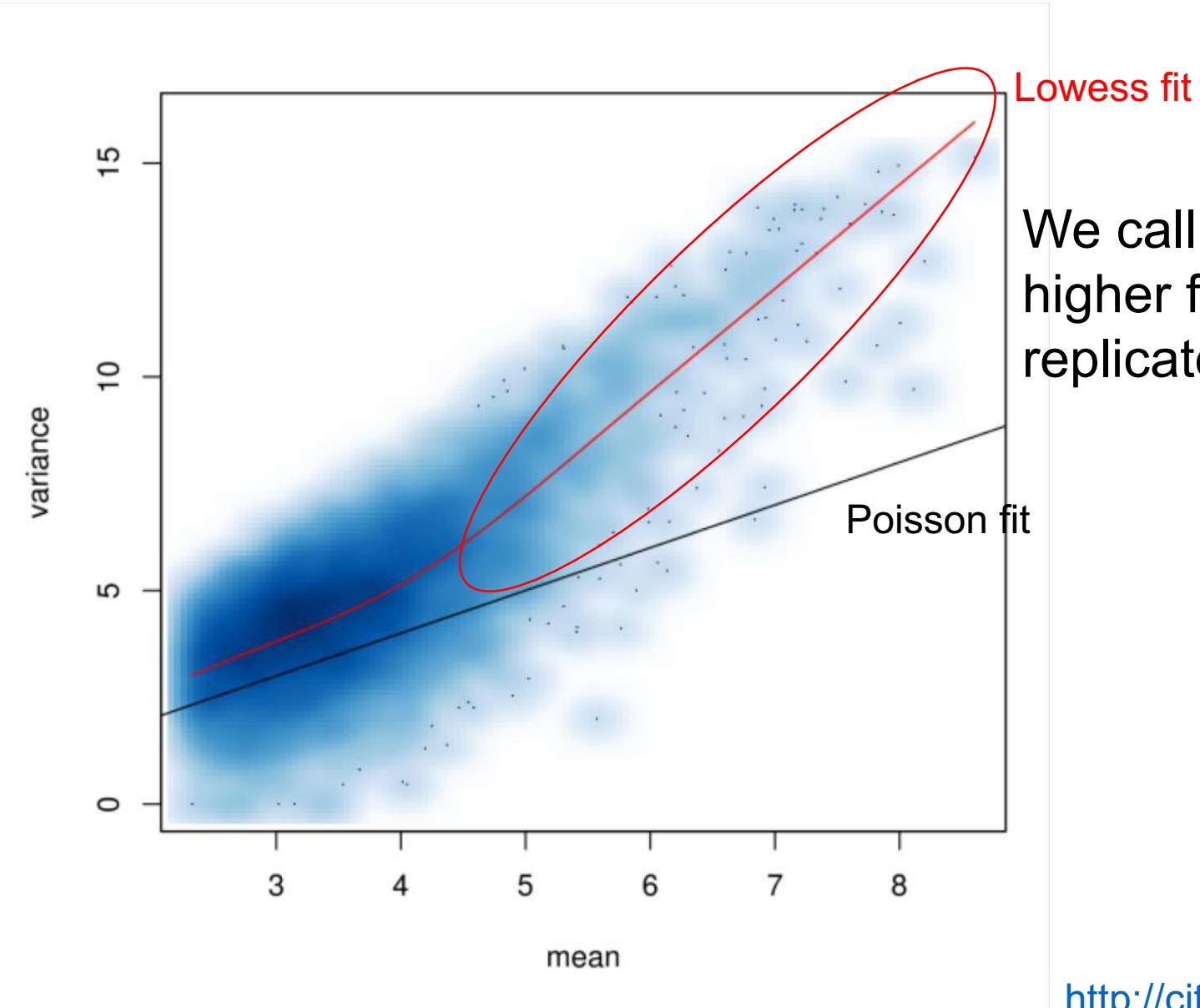


Poisson model seems good fit in technical replicates



Poisson fit
Lowess fit

Poisson model seems good fit in technical replicates



Lowess fit

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

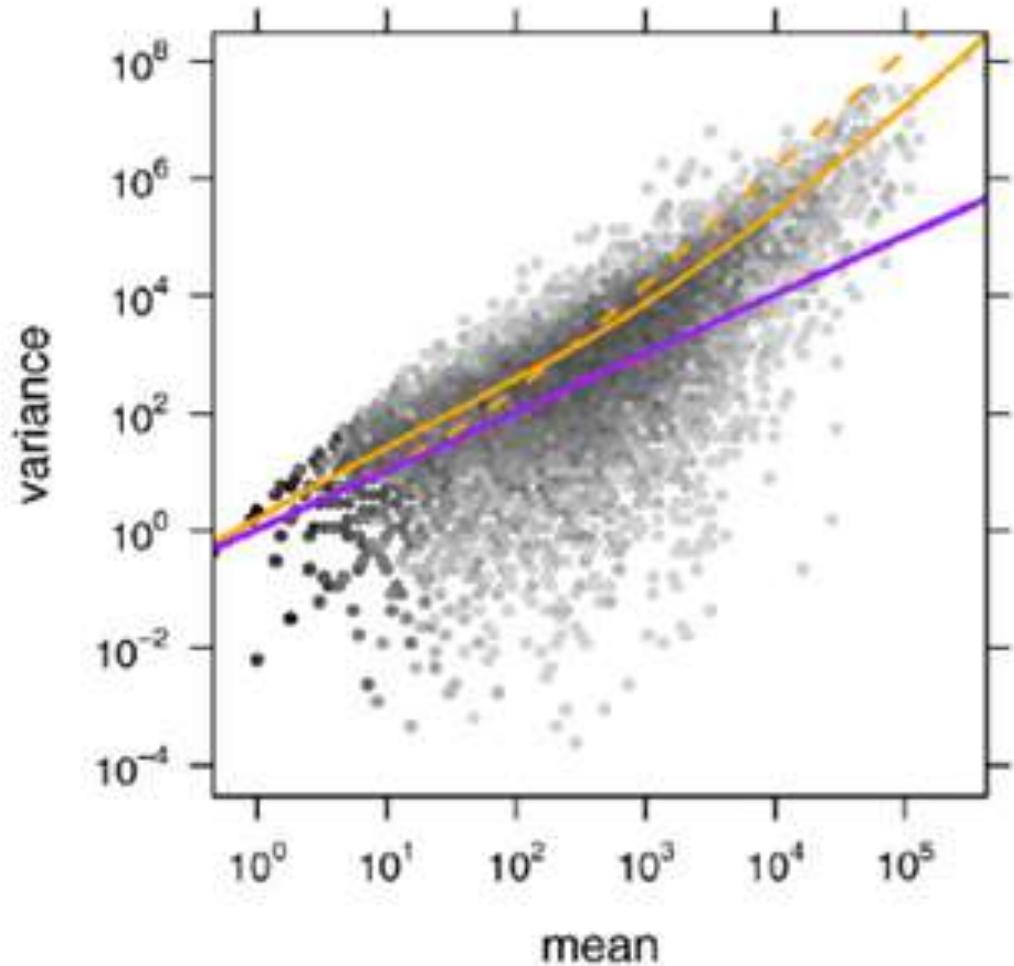
Variance depends strongly on the mean

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 $v = \mu$ Poisson distribution
- - - Poisson + constant CV $v = \mu + \alpha \mu^2$ (edgeR)
- Poisson + local regression $v = \mu + f(\mu^2)$ (DESeq) }

Negative binomial distribution

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

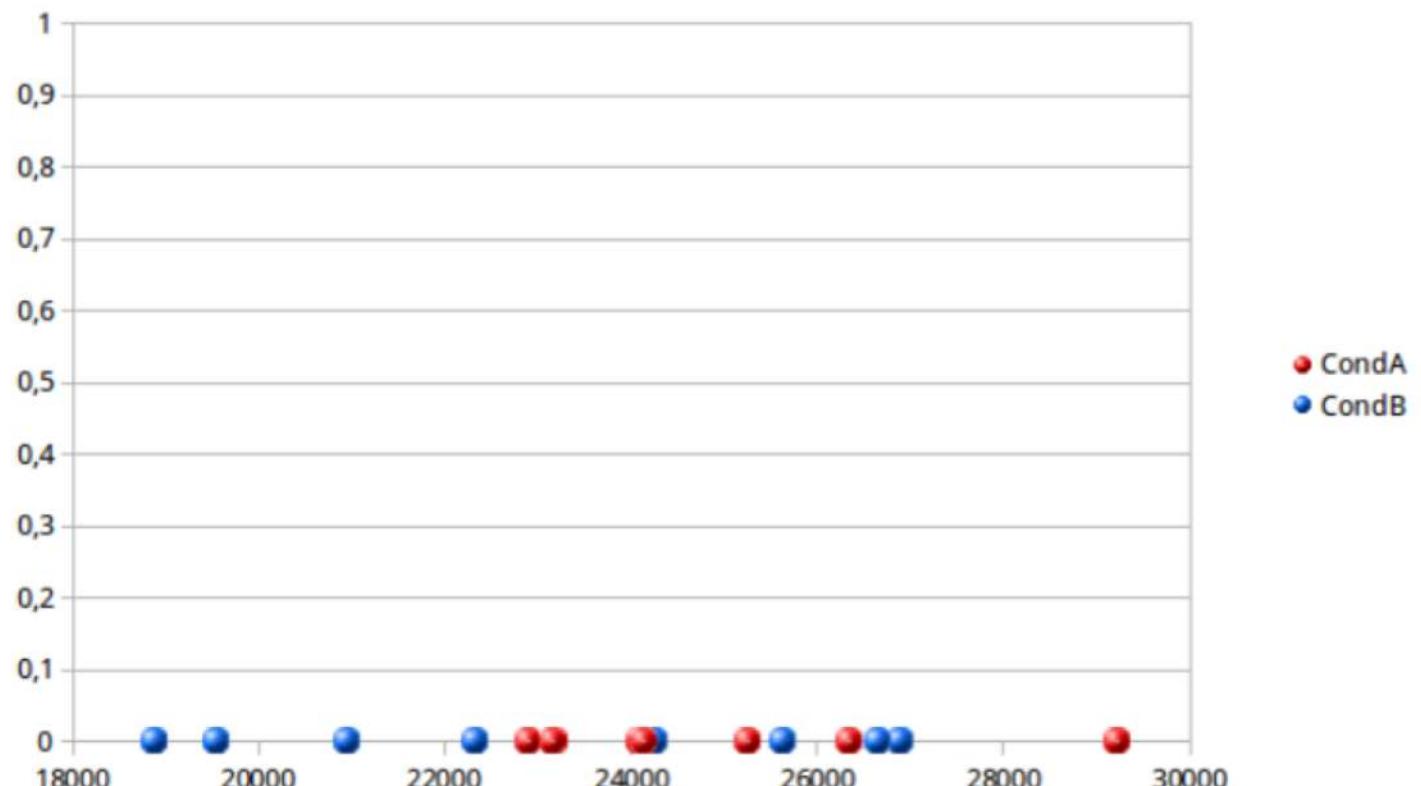
Table 7: Comparison of programs for differential gene expression identification (Rapaport et al., 2013; Seyednasrollah et al., 2015; Schurch et al., 2015).

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	<i>log</i> -normal	Neg. binomial
Test for DE	Exact test (Wald)	Exact test for over-dispersed data	Generalized linear model	<i>t</i> -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

Scenario

gene_id CAF0006876

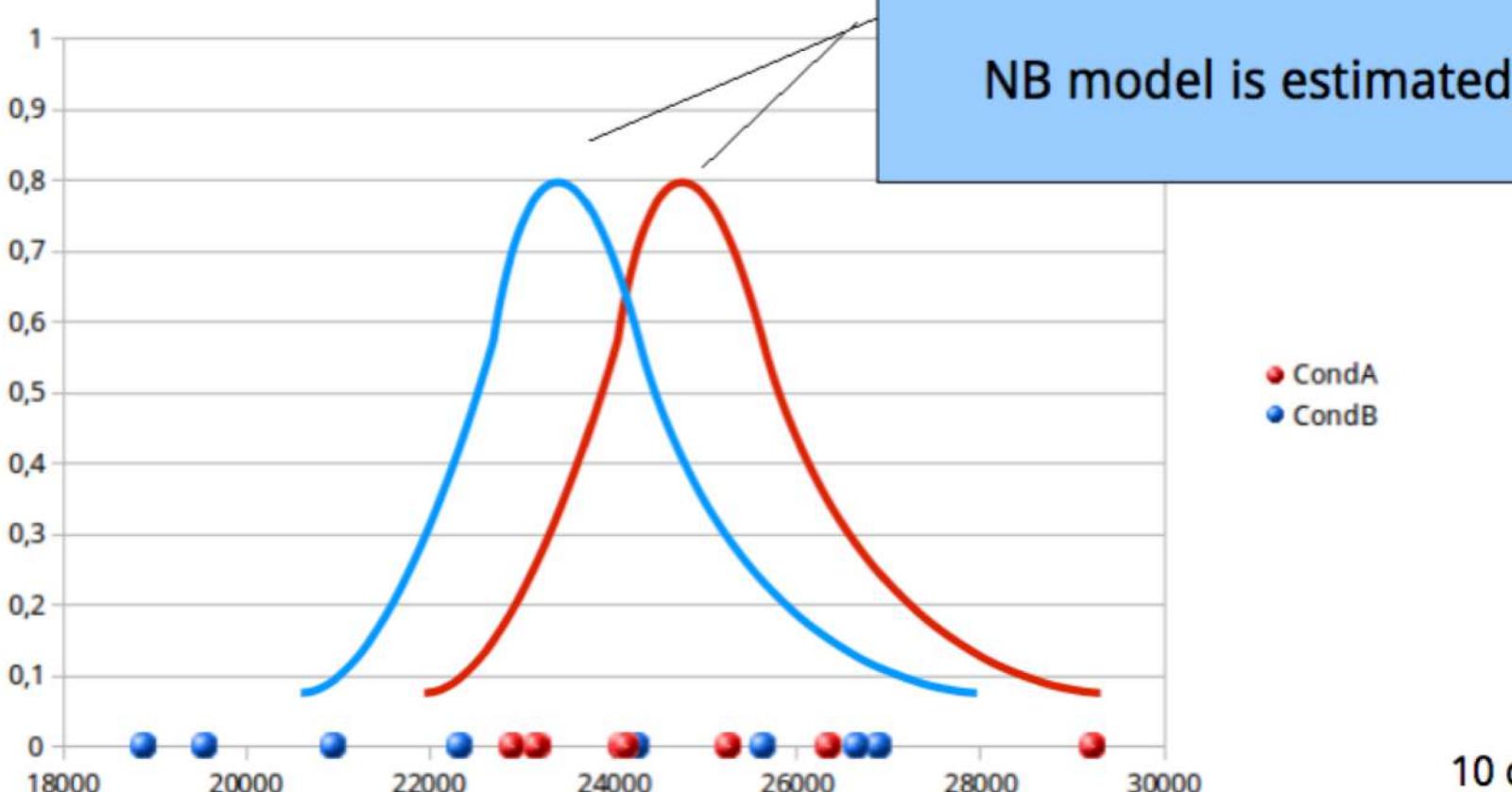
	sample1 23171	sample2 22903	sample3 29227	sample4 24072	sample5 23151	sample6 26336	sample7 25252	sample8 24122
Condition A								
Condition B	Sample9 19527	sample10 26898	sample11 18880	sample12 24237	sample13 26640	sample14 22315	sample15 20952	sample16 25629



Scenario

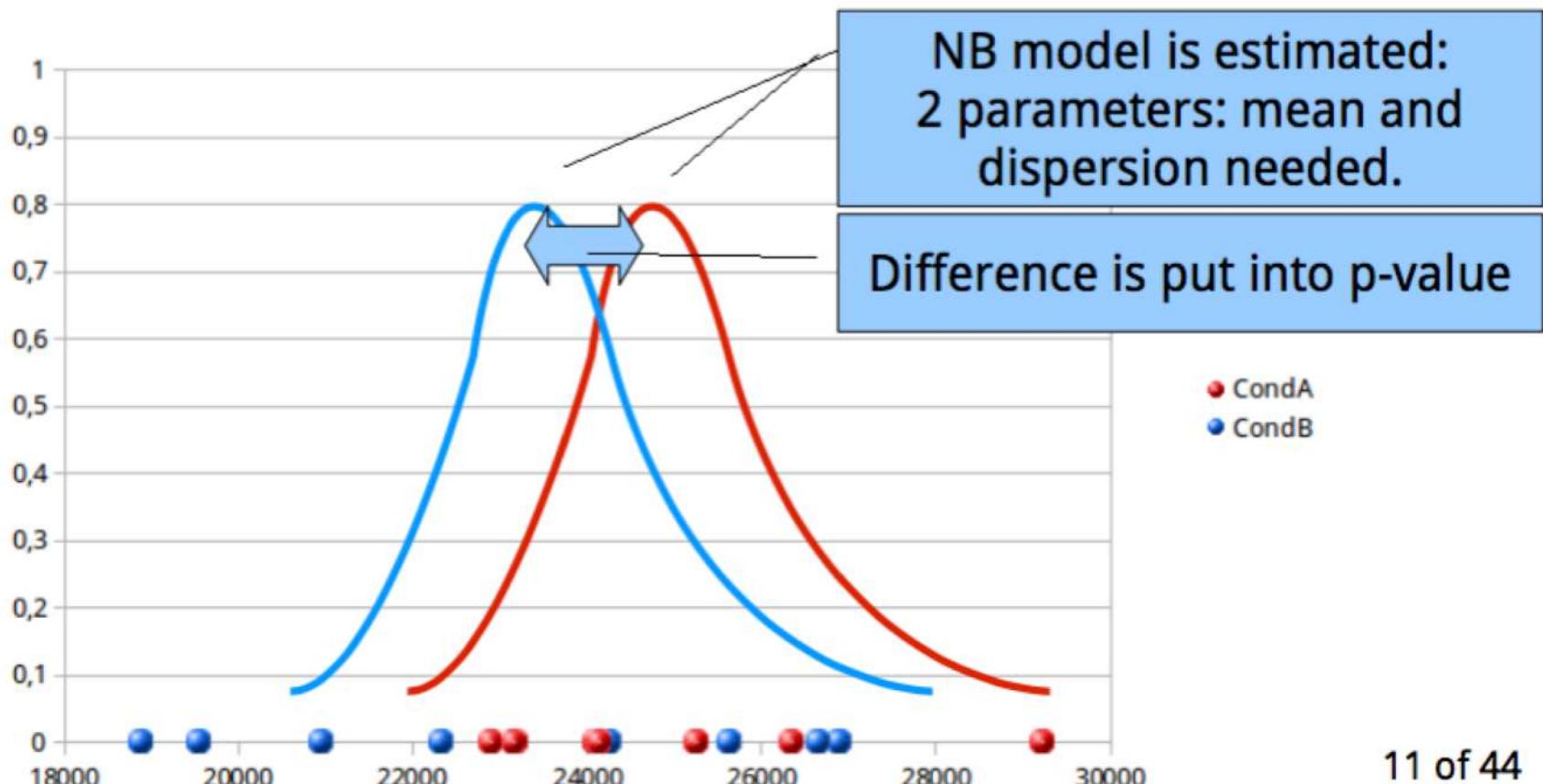
gene_id CAF0006876

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

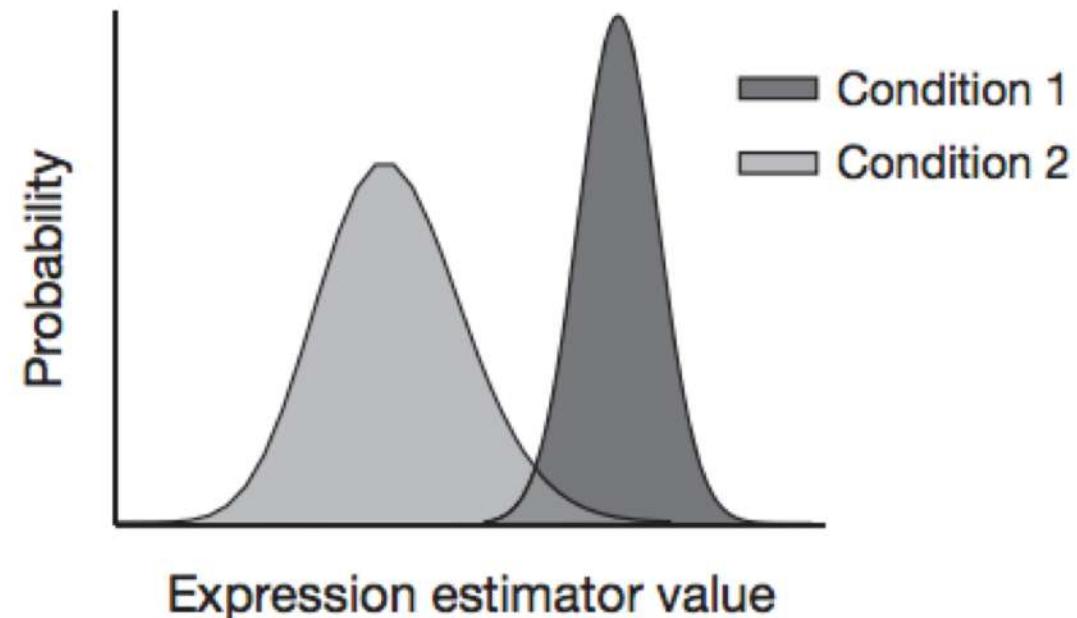
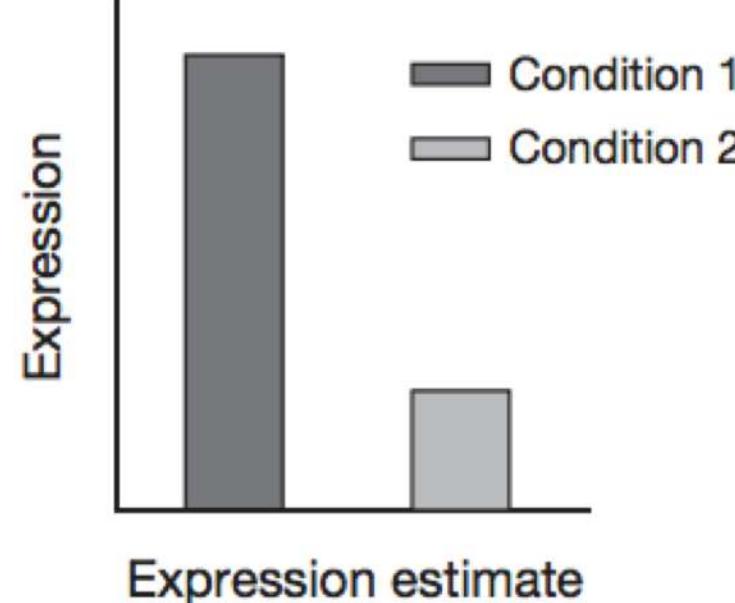
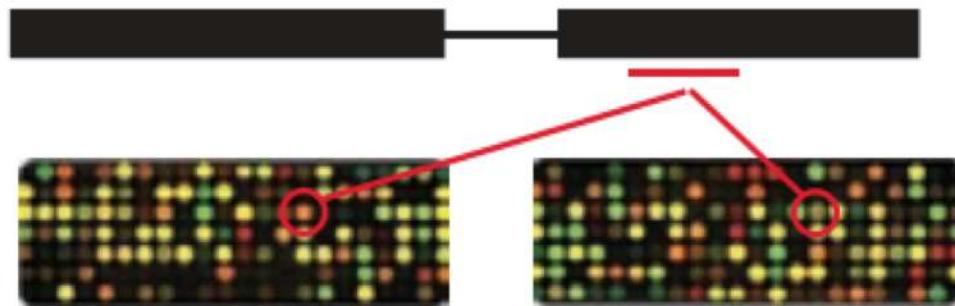


Scenario

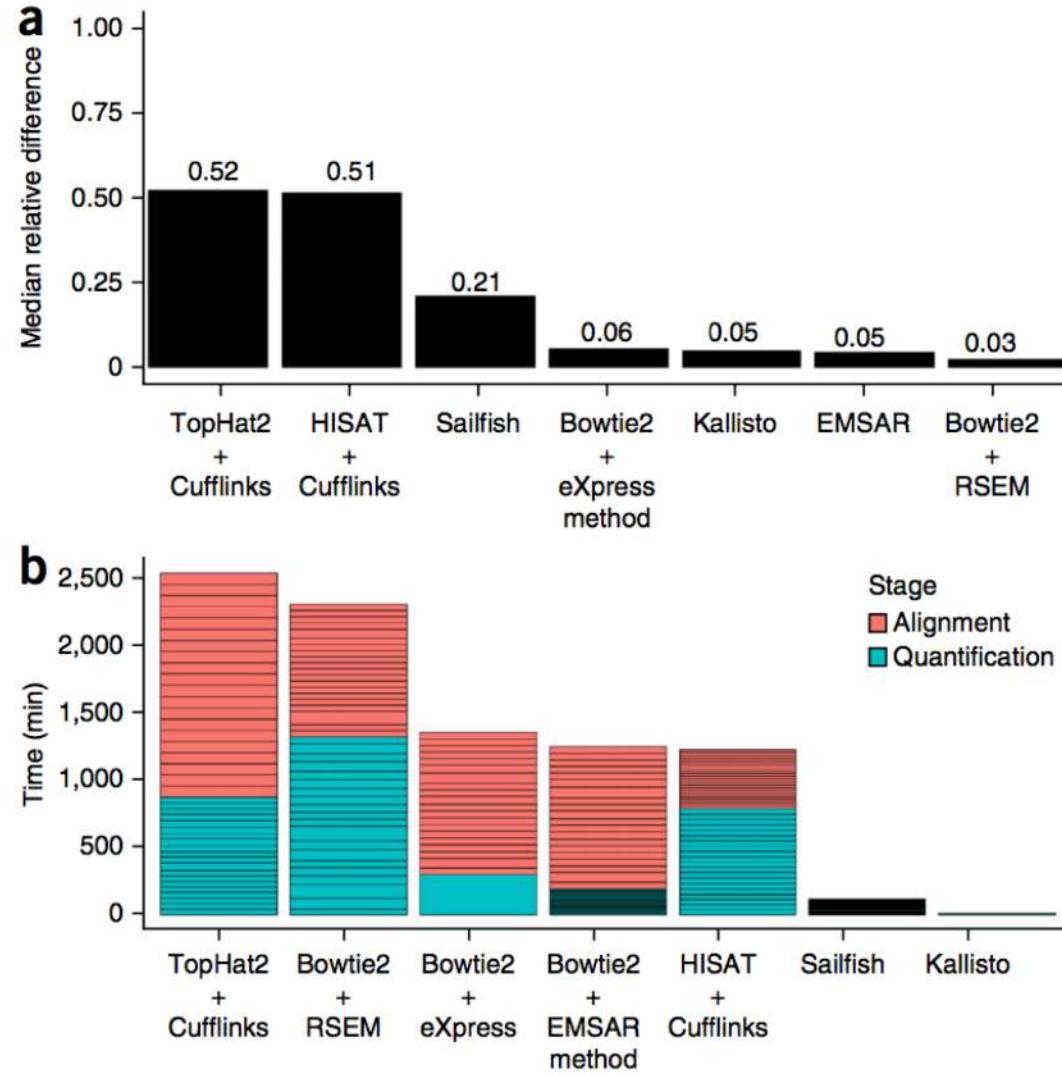
gene_id	CAF0006876	sample1	sample2	sample3	sample4	sample5	sample6	sample7	sample8
Condition A	23171	22903	29227	24072	23151	26336	25252	24122	
Condition B	19527	26898	18880	24237	26640	22315	20952	25629	



RNAseq vs Microarray

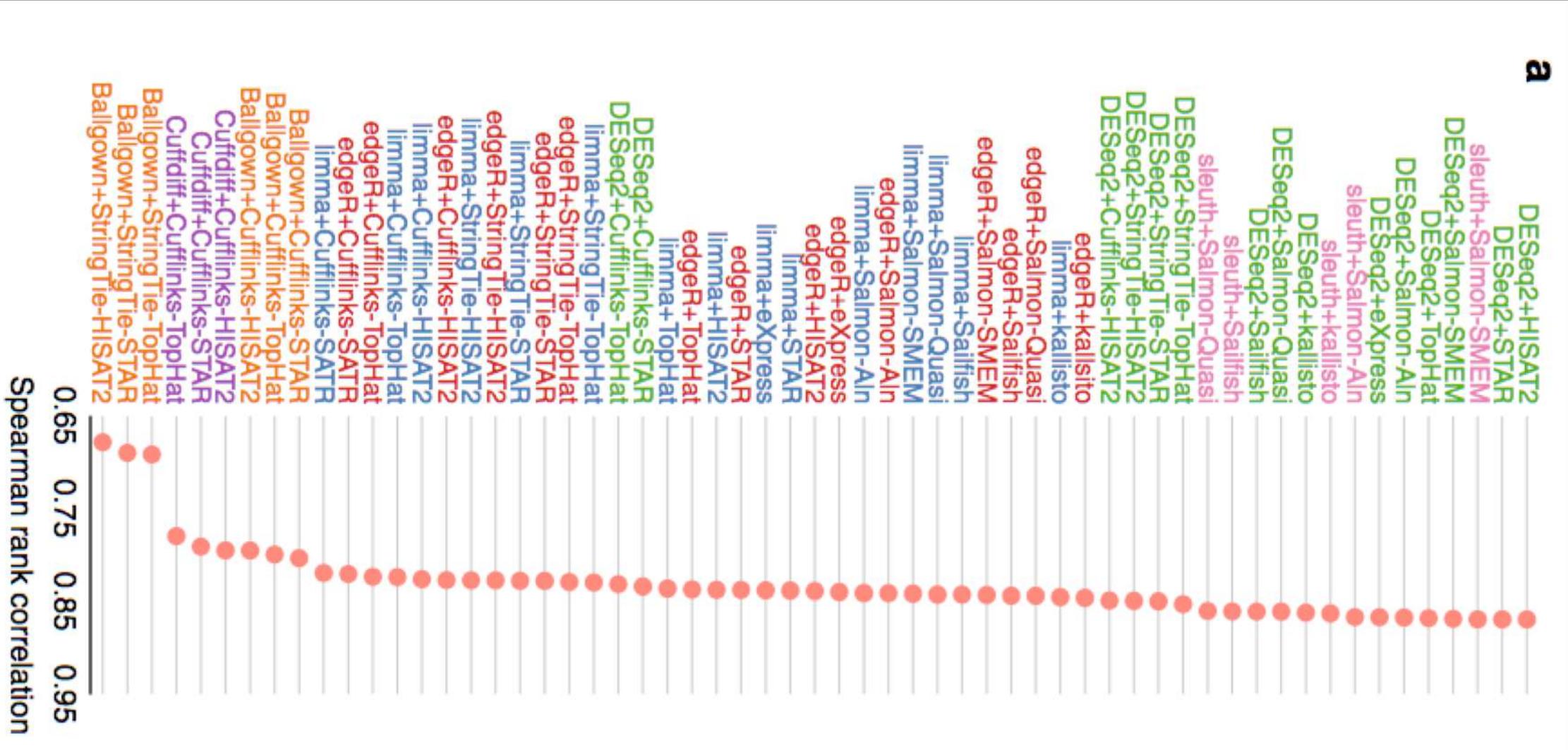


Advances in quantification



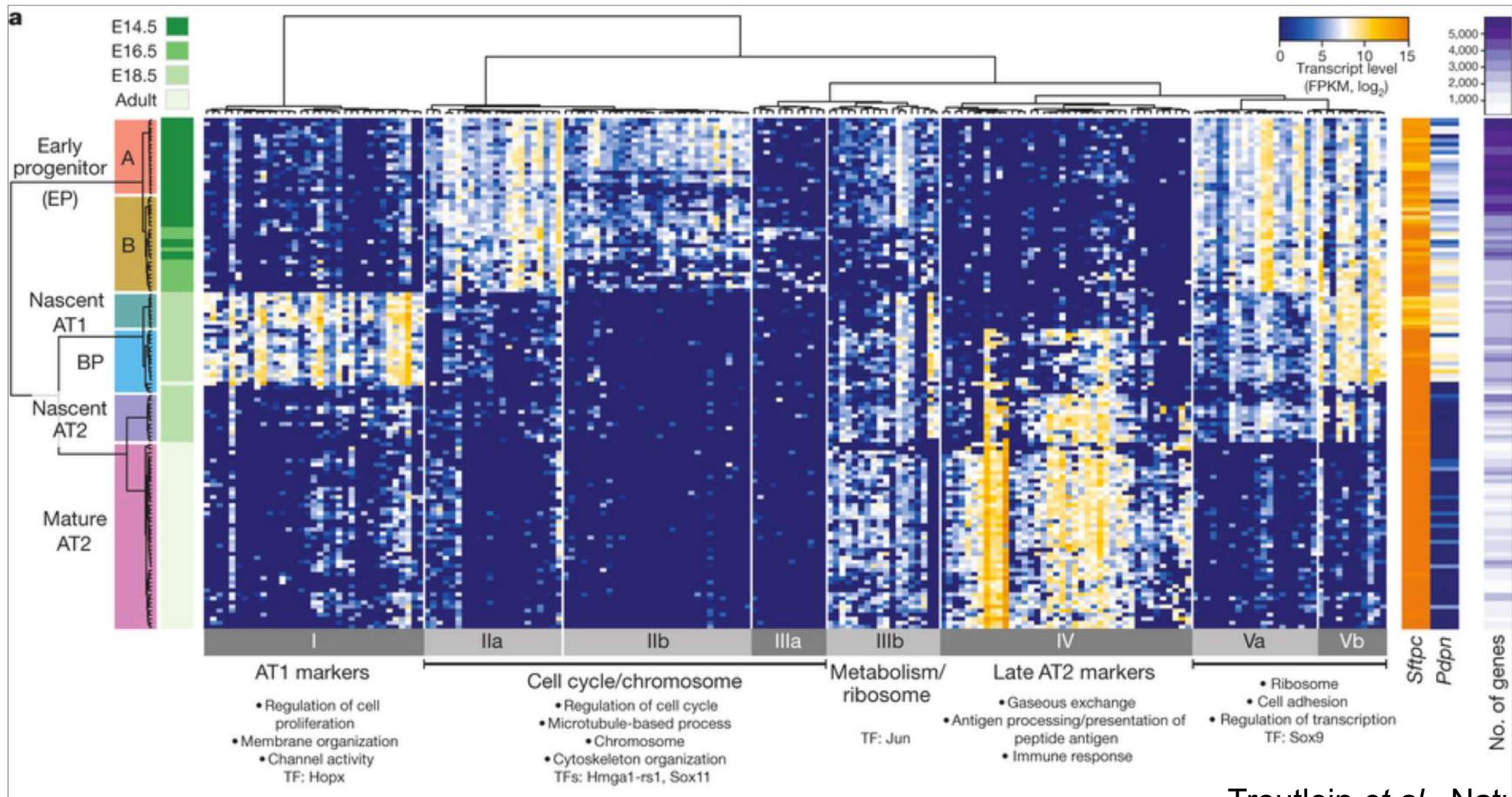
We present kallisto, an RNA-seq quantification program that is two orders of magnitude faster than previous approaches and achieves similar accuracy. Kallisto pseudoaligns reads to a reference, producing a list of transcripts that are compatible with each read while avoiding alignment of individual bases. 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.

Spearman rank correlation of DEG results to qPCR measured genes

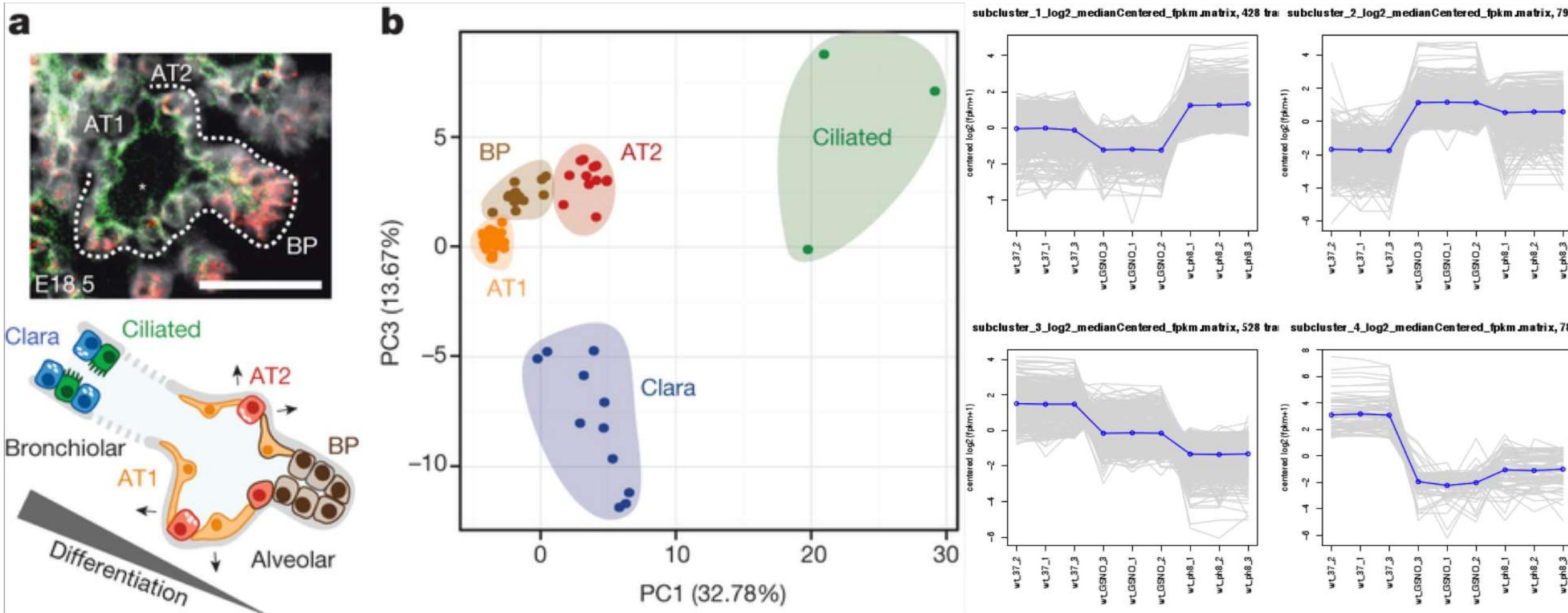


Once you have set of differentially expressed genes

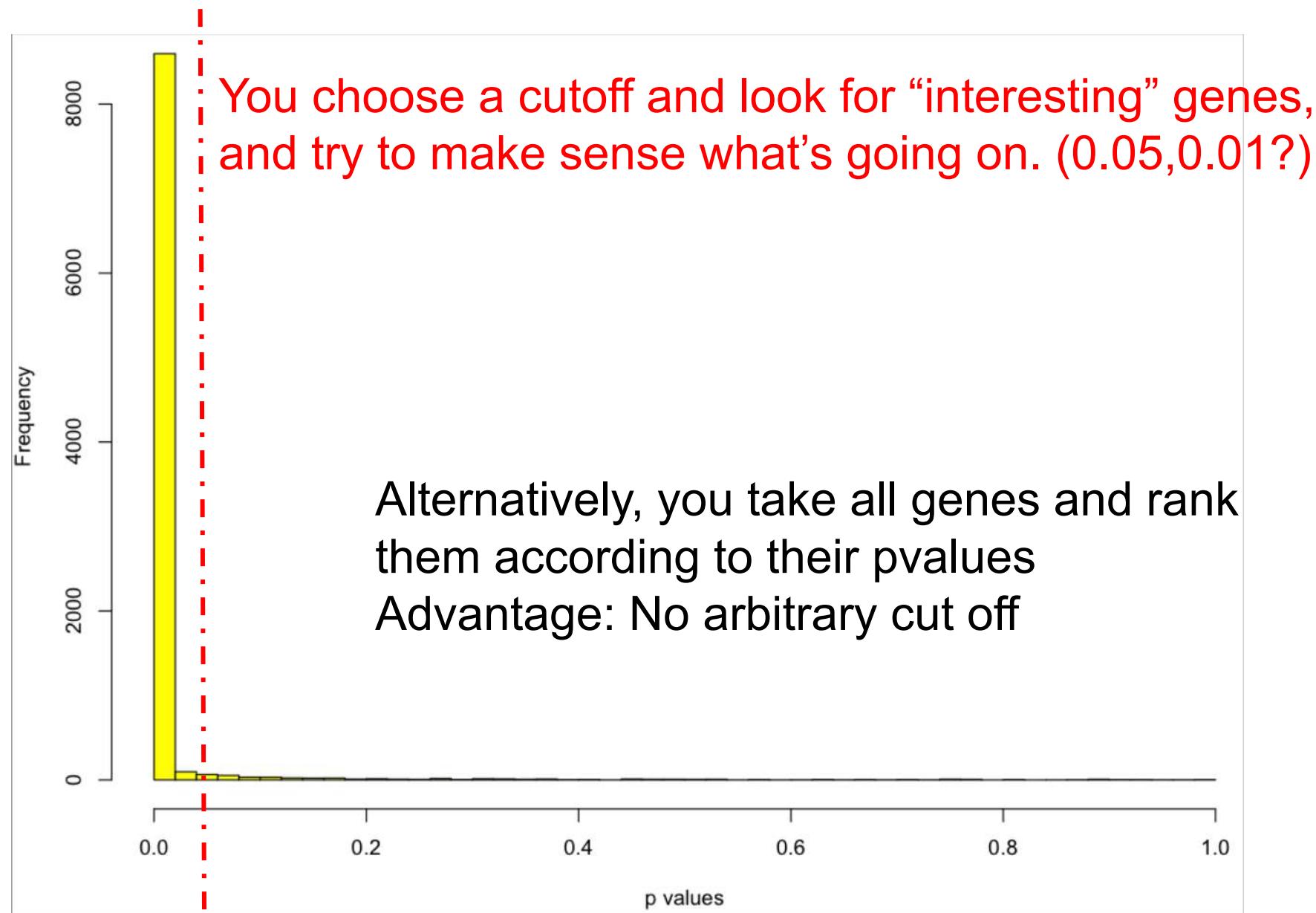
Summarization visualizing the expression data through heatmap ; Classification using Gene Ontology terms and metabolic annotations



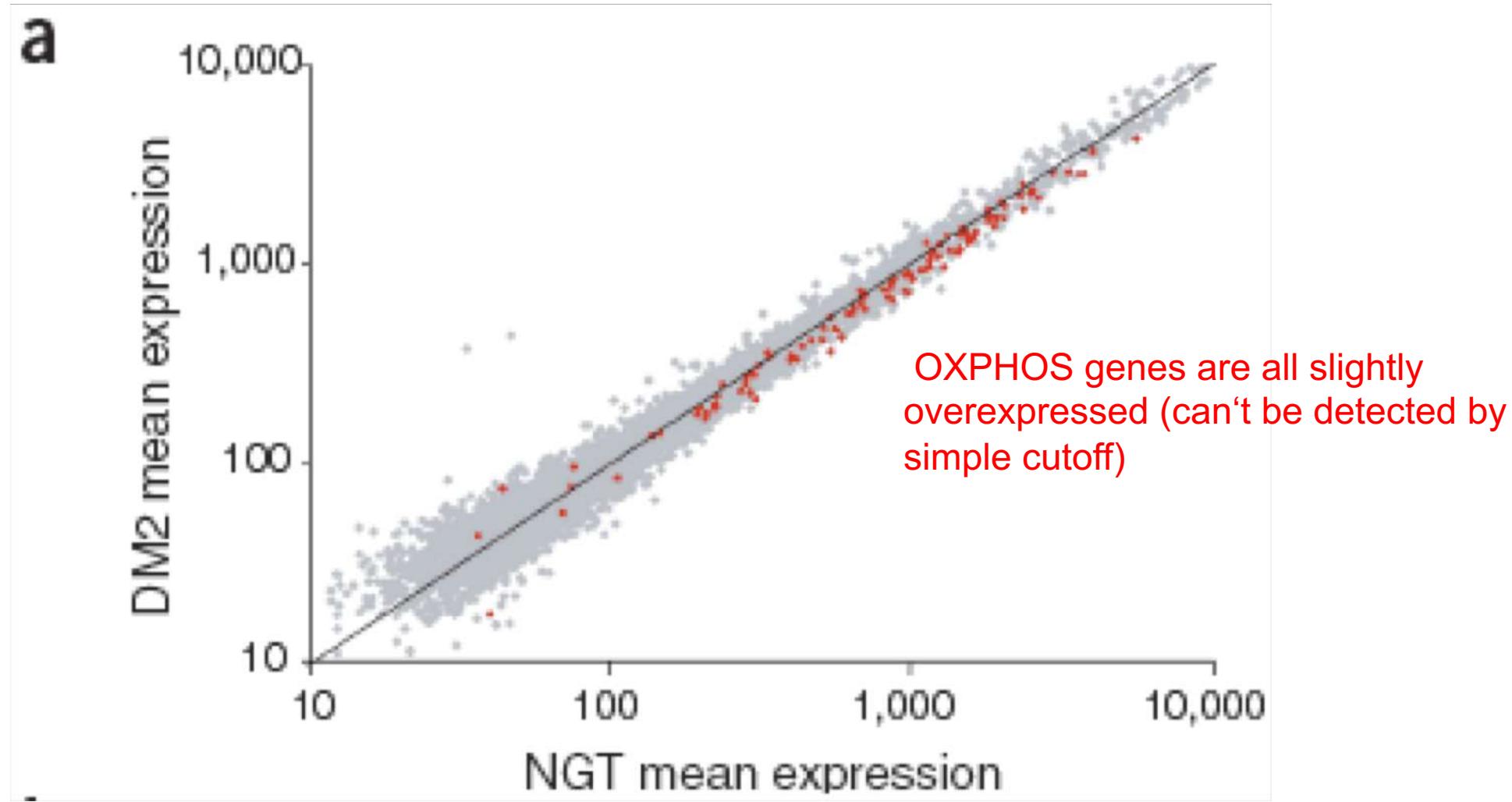
Clustering of the expression values and principal component analysis to reduce the variables.



Now, setting a cut-off

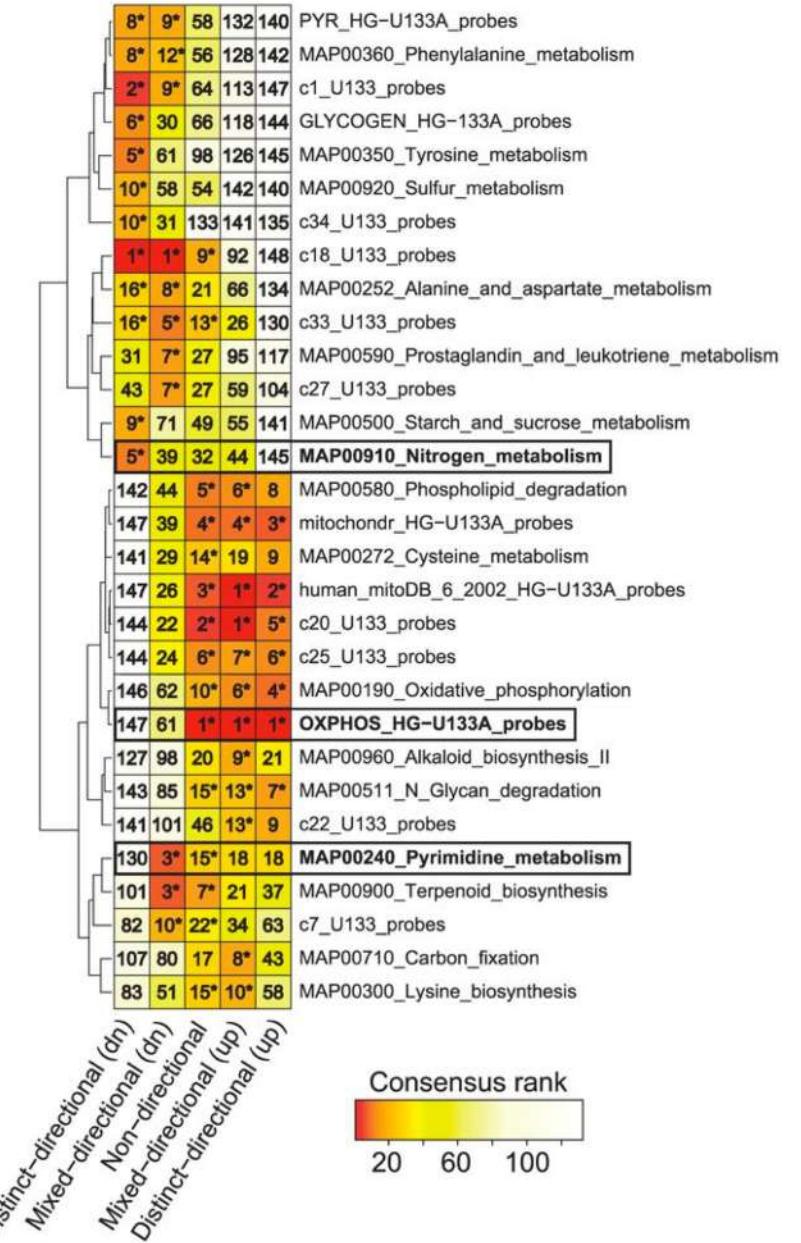


GSEA (Gene Set Enrichment Analysis) methods (cut-off free approach)



- Piano combined different methods and calculates a consensus score
- Output a heatmap of different gene set significantly enriched

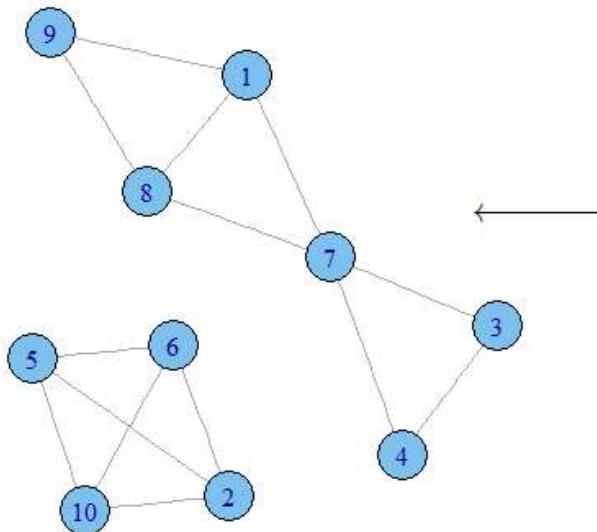
B Heatmap of consensus scores for all directionality classes (gene sets that have median rank 1-10 in at least one class)



Gene co-expression network

	S_1	S_2	S_3		G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	43.26	40.89	5.05		1.00	0.23	0.61	0.71	0.03	0.35	0.86	1.00	0.97	0.37
G_2	166.6	41.87	136.65		0.23	1.00	0.63	0.52	0.98	0.99	0.29	0.30	0.46	0.99
G_3	12.53	39.55	42.09		0.61	0.63	1.00	0.99	0.77	0.53	0.93	0.56	0.41	0.51
G_4	28.77	191.92	236.56		0.71	0.52	0.99	1.00	0.69	0.41	0.97	0.66	0.52	0.40
G_5	114.7	79.7	99.76		0.03	0.98	0.77	0.69	1.00	0.95	0.48	0.09	0.27	0.94
G_6	119.1	80.57	114.59		0.35	0.99	0.53	0.41	0.95	1.00	0.17	0.41	0.57	1.00
G_7	118.9	156.69	186.95		0.86	0.29	0.93	0.97	0.48	0.17	1.00	0.83	0.72	0.16
G_8	3.76	2.48	136.78		1.00	0.30	0.56	0.66	0.09	0.41	0.83	1.00	0.98	0.42
G_9	32.73	11.99	118.8		0.97	0.46	0.41	0.52	0.27	0.57	0.72	0.98	1.00	0.58
G_{10}	17.46	56.11	21.41		0.37	0.99	0.51	0.40	0.94	1.00	0.16	0.42	0.58	1.00

Gene expression values Similarity (Co-expression) score

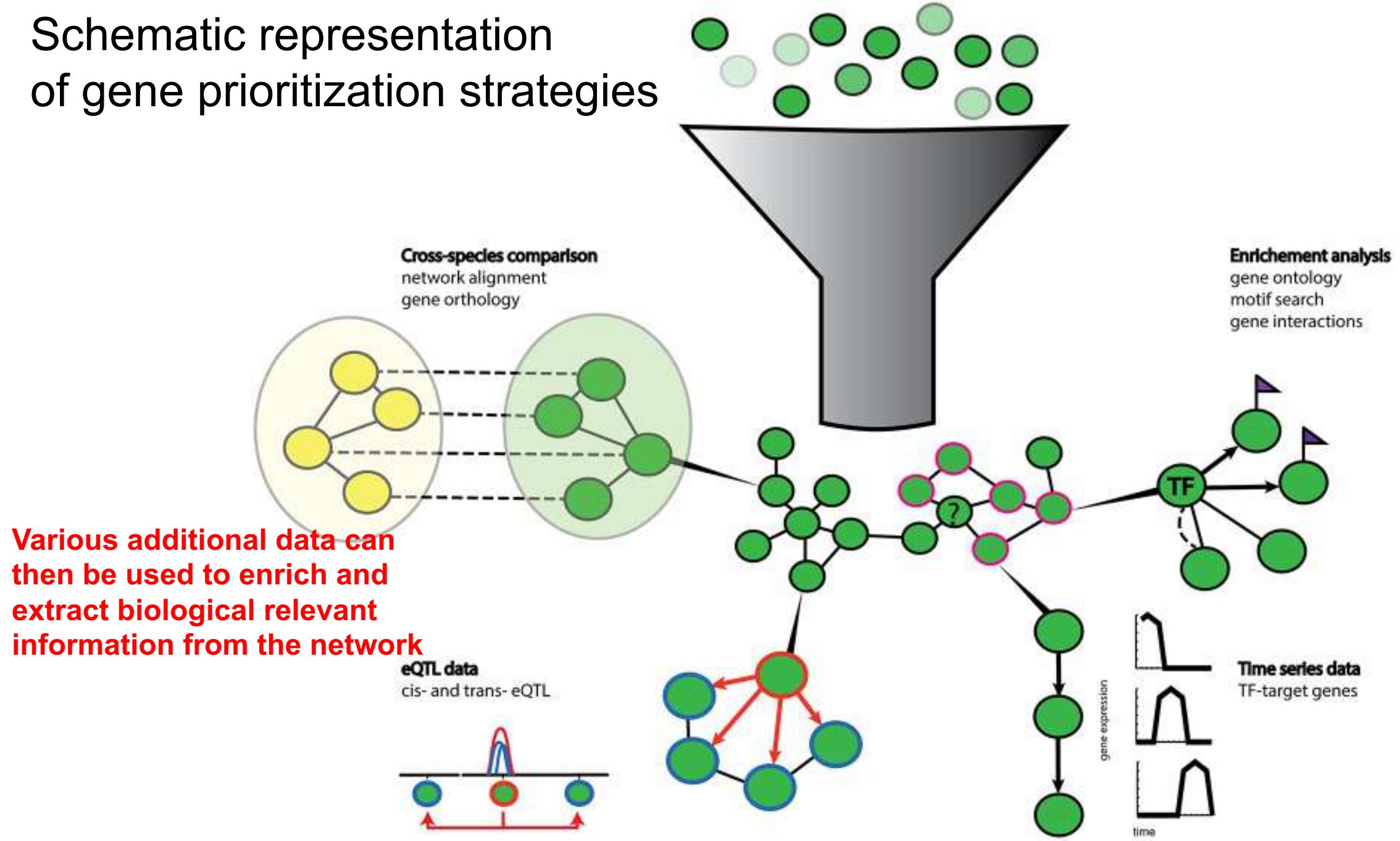


	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}
G_1	0	0	0	0	0	0	1	1	1	0
G_2	0	0	0	0	1	1	0	0	0	1
G_3	0	0	0	1	0	0	1	0	0	0
G_4	0	0	1	0	0	0	1	0	0	0
G_5	0	1	0	0	0	1	0	0	0	1
G_6	0	1	0	0	1	0	0	0	0	1
G_7	1	0	1	1	0	0	0	1	0	0
G_8	1	0	0	0	0	0	1	0	1	0
G_9	1	0	0	0	0	0	0	1	0	0
G_{10}	0	1	0	0	1	1	0	0	0	0

Network adjacency matrix

$|r(G_i, G_j)| \geq 0.8$ Significance threshold

Schematic representation of gene prioritization strategies



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 (isoform quantification, assembly, single cell sequencing)