RNAseq and Annotation

Isheng Jason Tsai

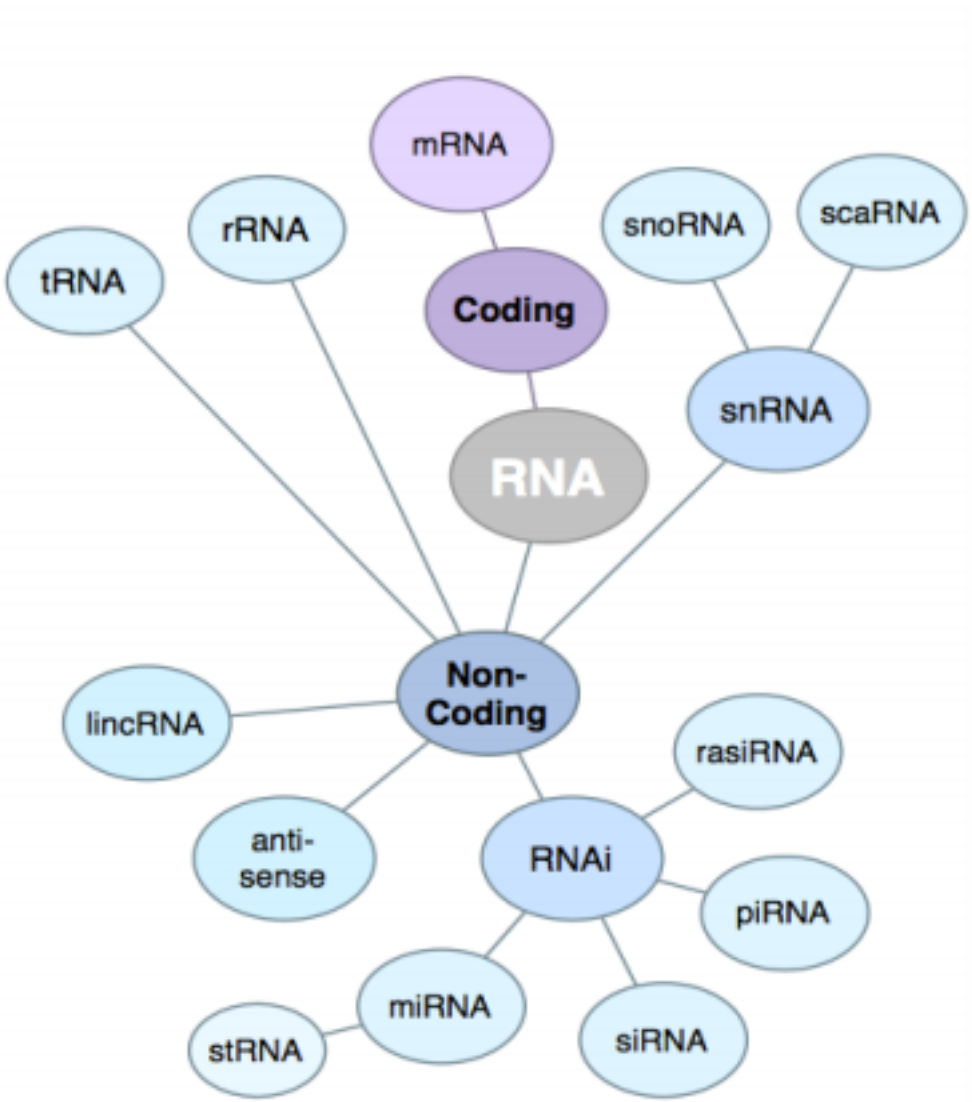
Introduction to NGS Data and Analysis
Lecture 8

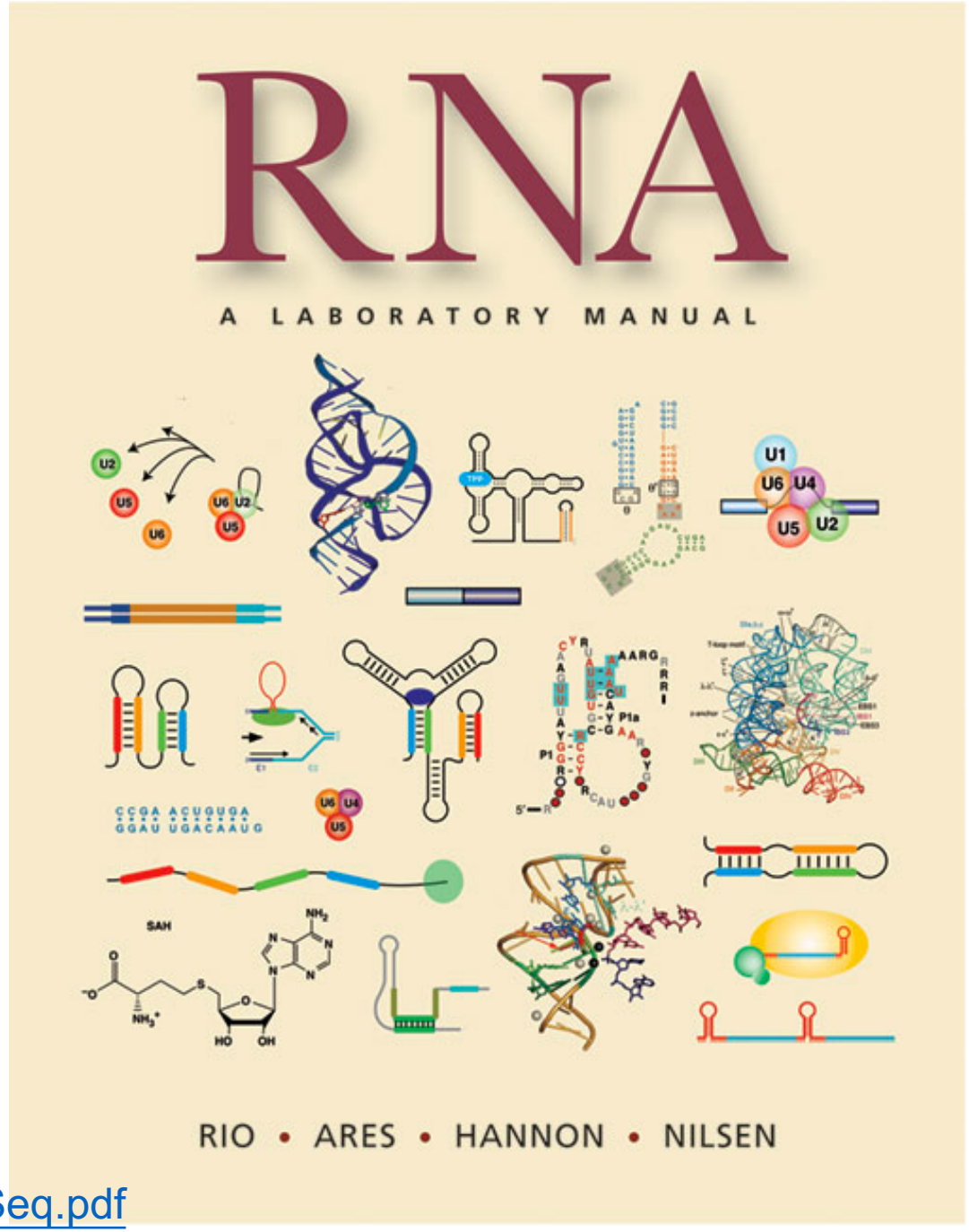
For this lecture, I gathered information from :

- Reviews
- Slideshare (with permission)
  - Especially http://www.slideshare.net/aubombarely/rnaseq-analysis-19910448

- Twitter
- RNAseq blog http://www.rna-seqblog.com/
- Prof. Chien-Yu Chen's slides from ISEGB, 2015

Types of RNA





http://jura.wi.mit.edu/bio/education/hot_topics/RNAseq/RNA_Seq.pdf

Applications of RNAseq

Discovery / Annotation

- Find new genes
- Find new transcripts
- Find new ncRNAs, xxx, xxx ….
- Gene fusion

Comparison / Quantification : given X conditions, find the effect of Y on

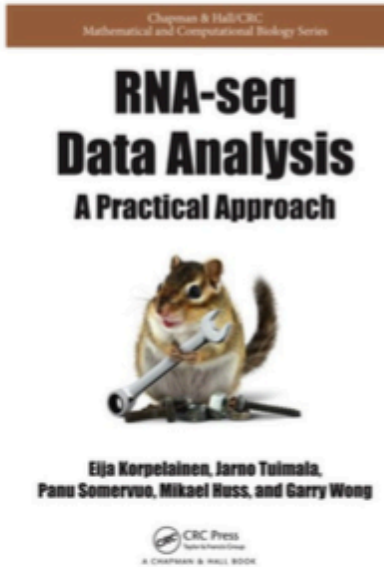
- expression
- Isoform abundance, splice patterns, transcript boundaries

New field:

RNA:RNA (CLASH) (Travis *et al.,* 2014)

RNA:protein (RIP-seq) (Cook *et al.,* 2015)

Expression quantification and transcript assembly



September 9, 2014
E-Book - ISBN 9781482262353

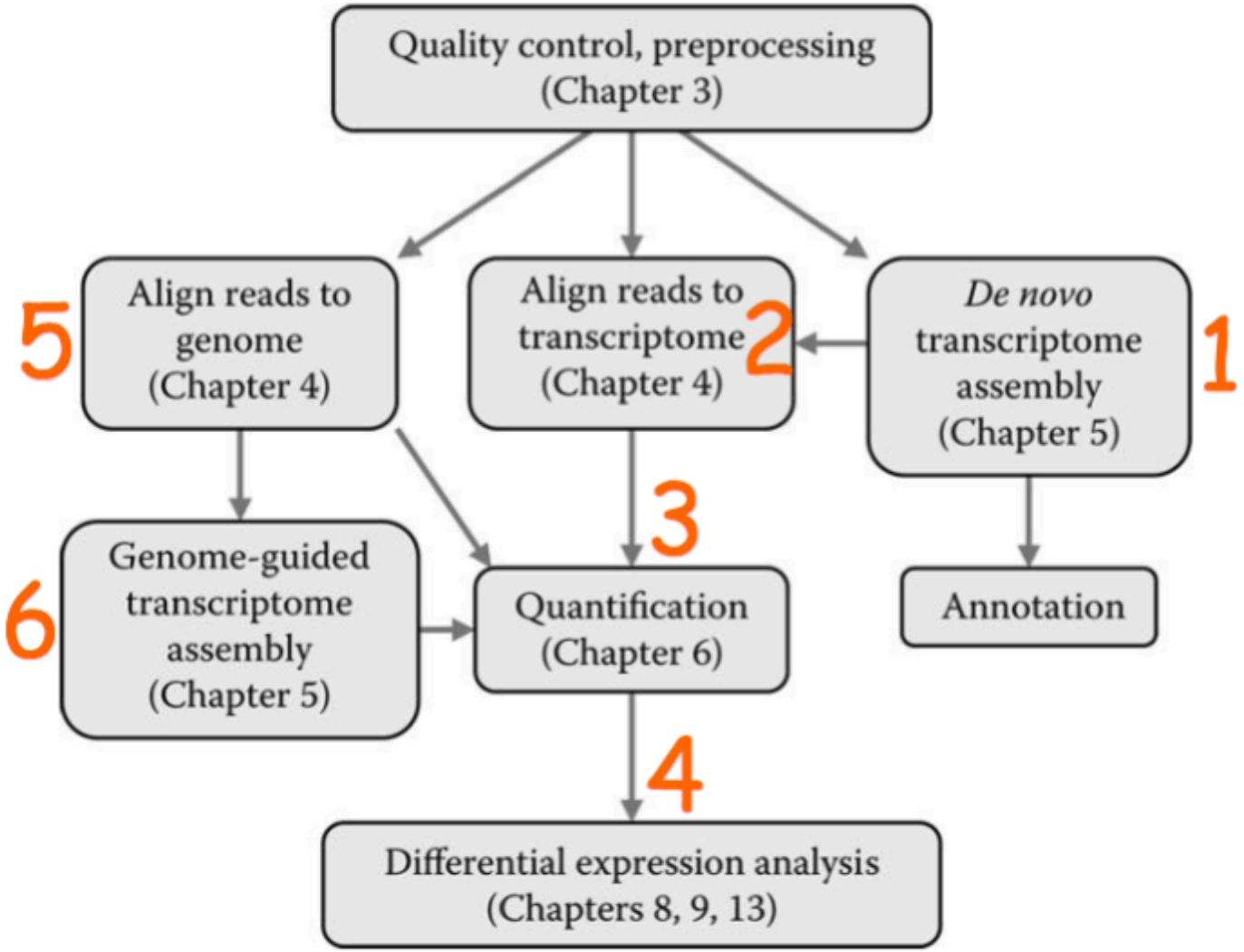
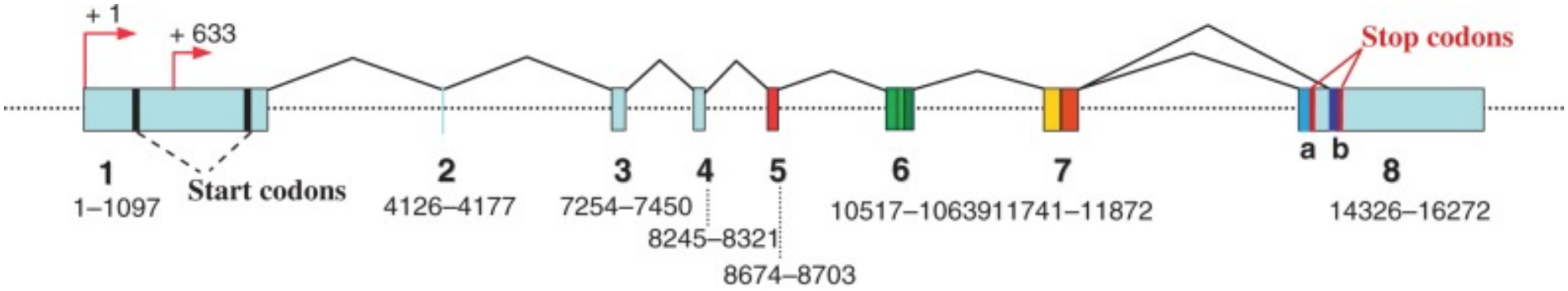


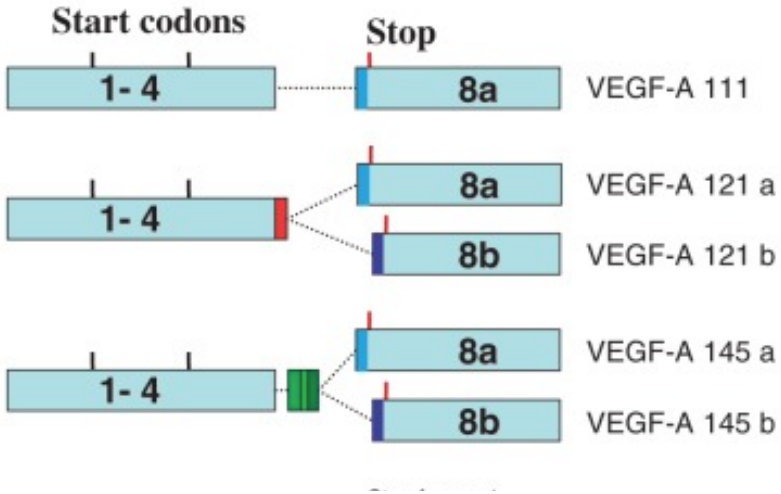
FIGURE 2.1 Possible paths in RNA-seq data analysis.

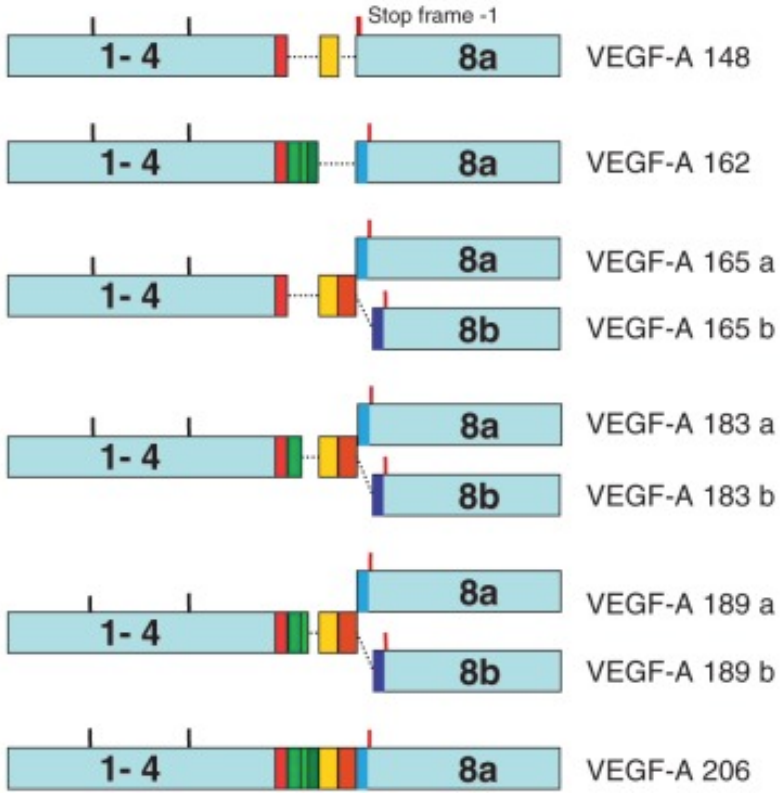
Annotation

(focus only on gene annotation)

Gene and isoforms







Arcondeguy et al (2015)

STUDY DESIGNS

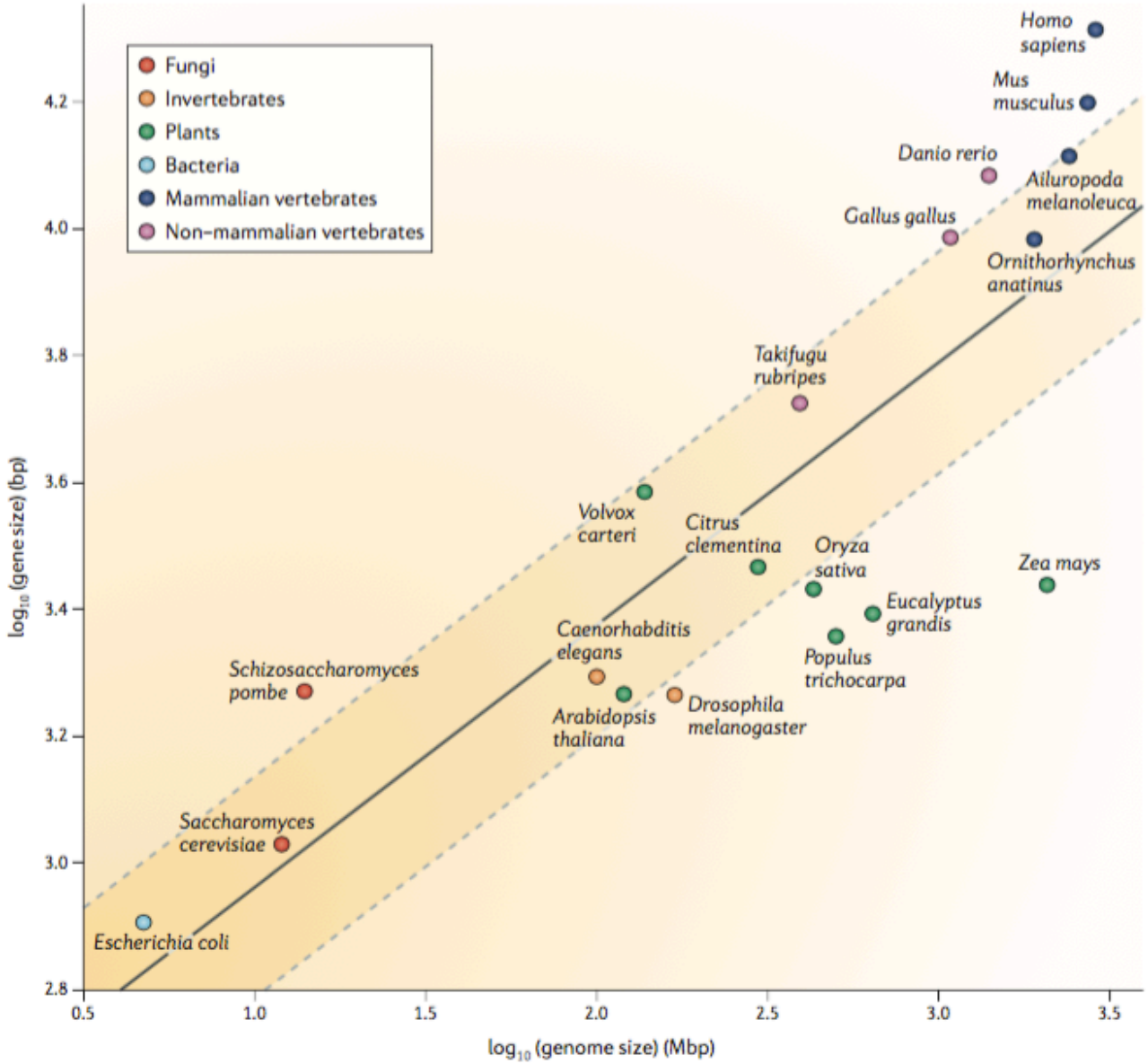
A beginner's guide to eukaryotic genome annotation

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.

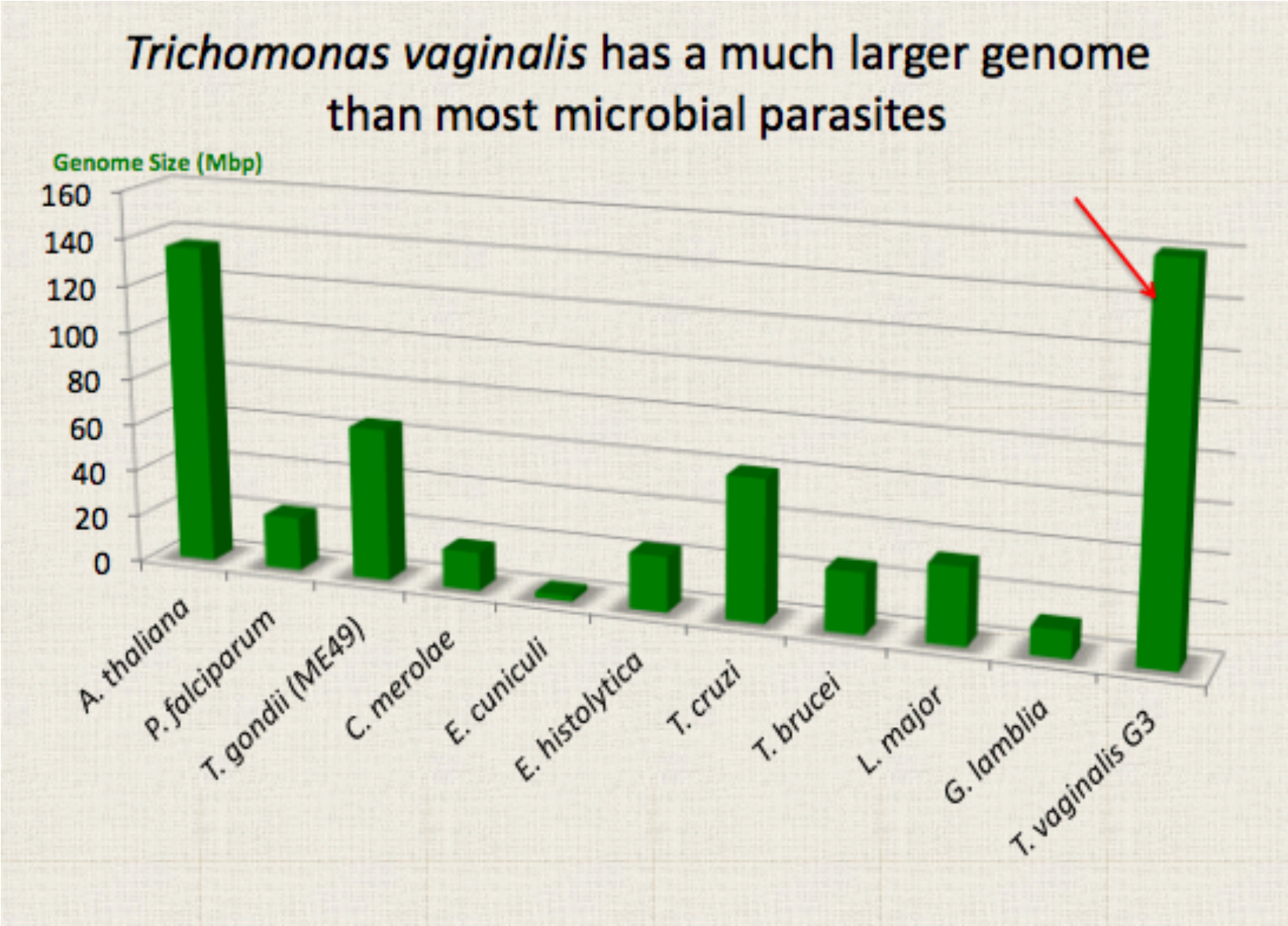
Yandell and Ence (2012)

Know your genome size (and gene numbers)

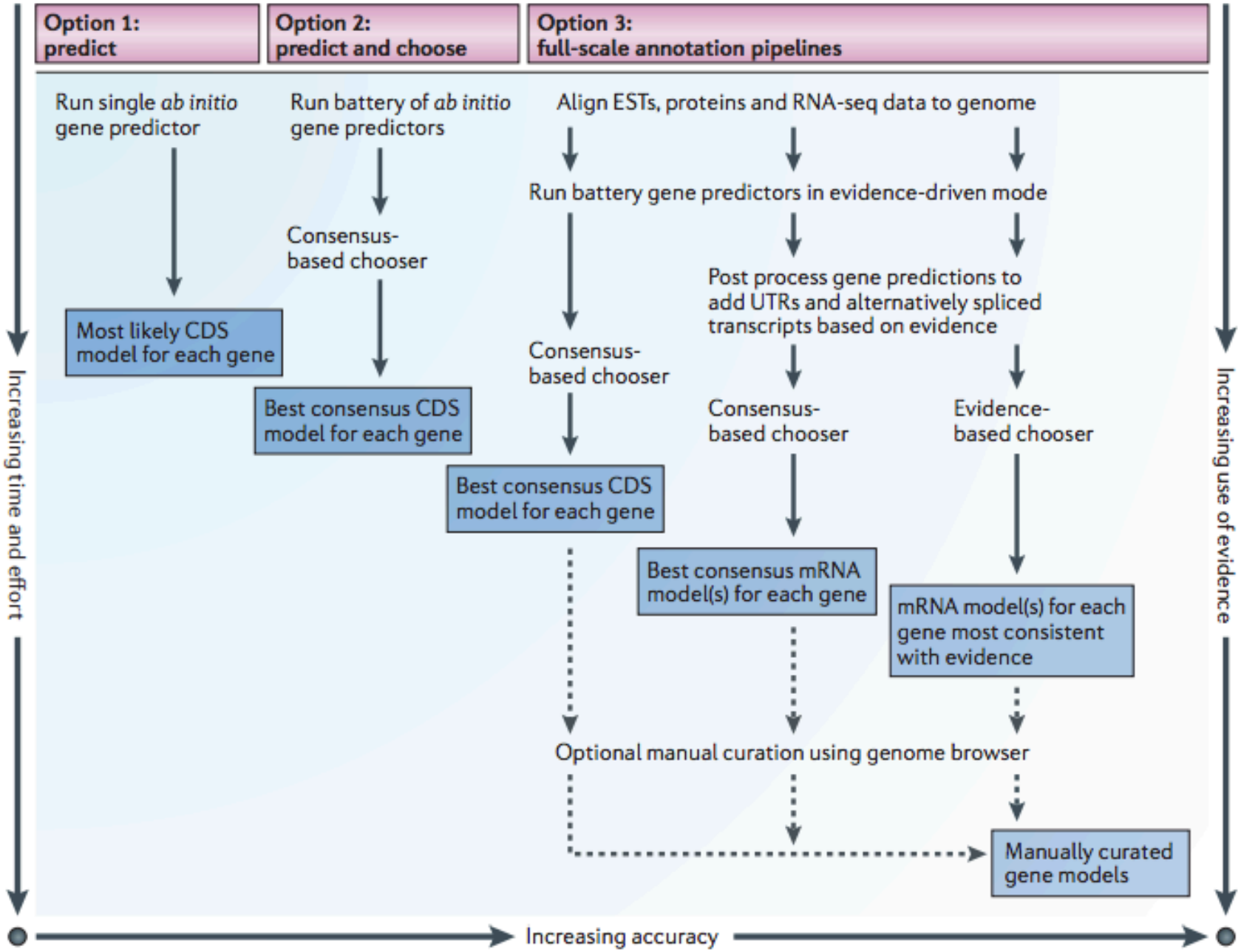


Yandell and Ence (2012)

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

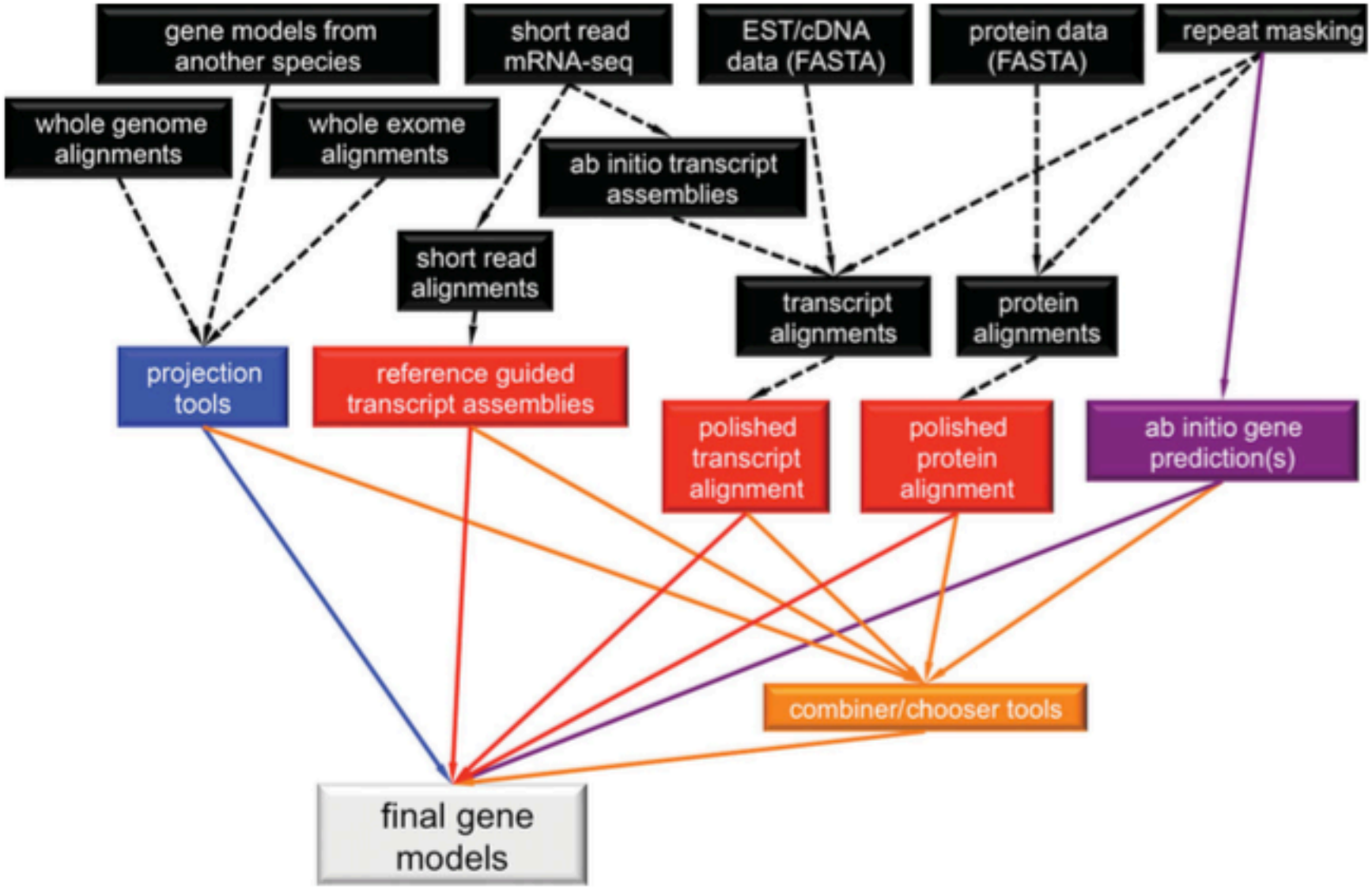


Ence (2014)



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

Prediction and Evidence aligners

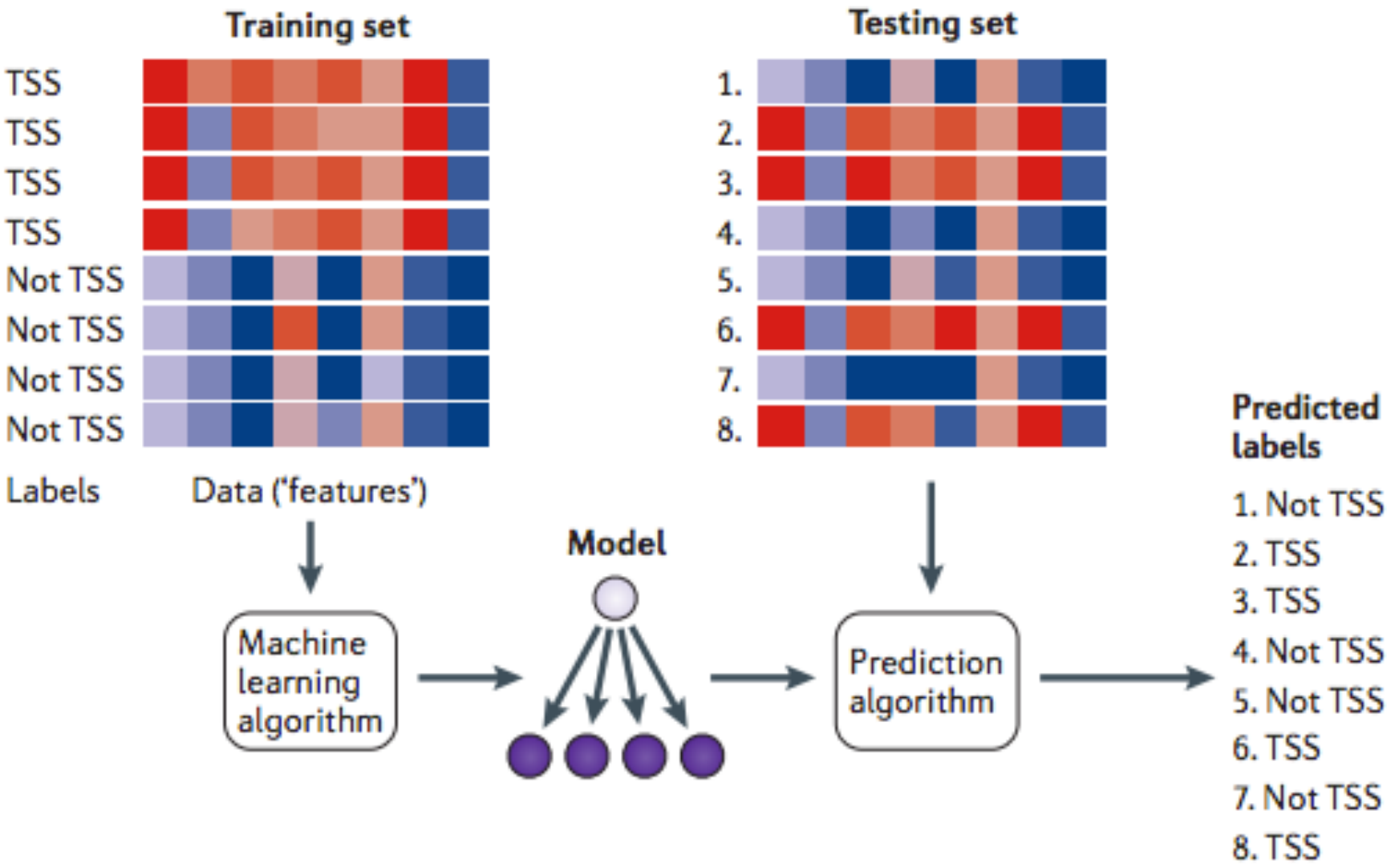
Table 4.1.2 Gene Predictors

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.	(*UNITS 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.	(Souvorov 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

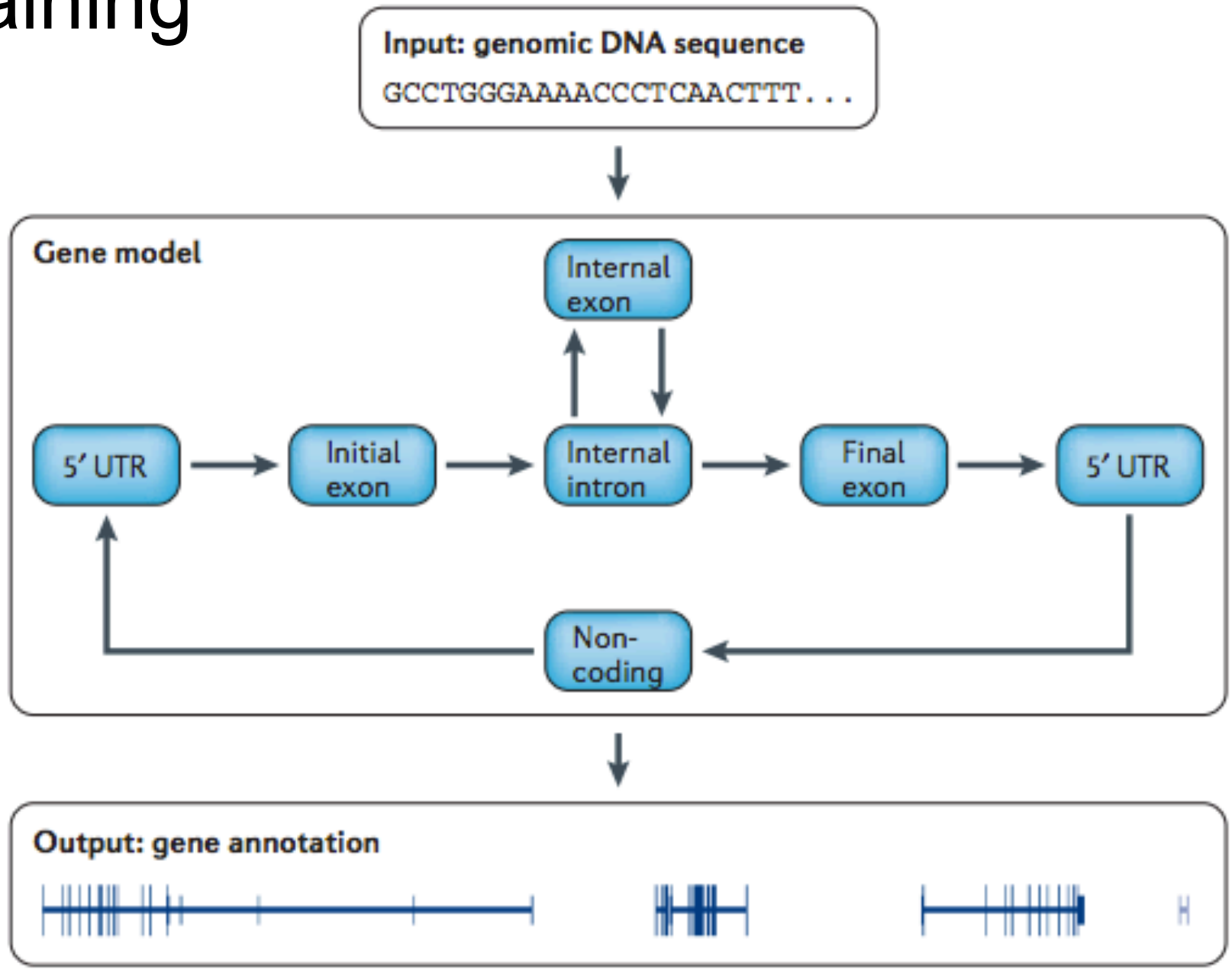
Software package	Features	Reference
BLAST	A suite of tools that can align any combination of protein and nucleotide sequences. Uses Karlin-Altschul statistics.	(*UNIT 3.4*; Korf et al., 2003; Ladunga, 2009)
BLAT	Faster than BLAST but not as configurable.	(*UNIT 10.8*; Kent, 2002; Bhagwat et al., 2012)
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 *de novo* 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)

Machine learning; An example on TSS (Transcription start site)



Libbrecht and Noble (2015)

Training



Libbrecht and Noble (2015)

Where to find initial “correct” genes

BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 9 2007, pages 1061–1067
doi:10.1093/bioinformatics/btm071

Genome analysis

CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes

Genis Parra[1], Keith Bradnam[1] and Ian Korf[1,2,*]

[1]UC Davis Genome Center, 451 E. Health Sciences Drive and [2]Department of Molecular and Cellular Biology, University of California Davis, Davis, CA 95616, USA

Received on December 7, 2006; revised on January 26, 2007; accepted on February 22, 2007

Advance Access publication March 1, 2007

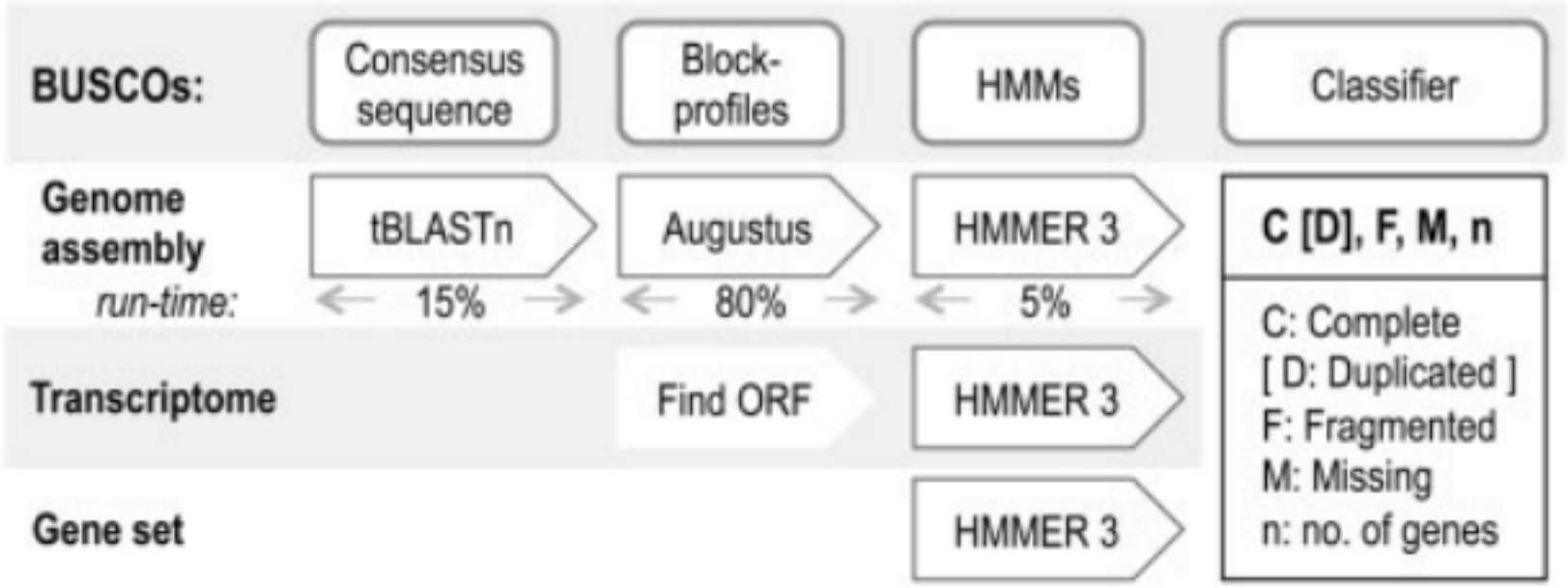
Associate Editor: Alex Bateman

CEGMA: a pipeline to accurately annotate core genes in ...
www.ncbi.nlm.nih.gov/pubmed/17332020 ▾ 翻譯這個網頁
由 G Parra 著作 - 2007 - 被引用 318 次 - 相關文章
Bioinformatics. 2007 May 1;23(9):1061-7. Epub 2007 Mar 1. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Parra G(1), Bradnam ...

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*



RNAseq Raw data type

Platform	GS jnr	FLX plus	MiSeq	Next Seq 500	HiSeq 2500 RR	Hiseq 2500 V3	HiSeq 2500 V4	HiSeq 4000	HiSeq X	SOLiD 4	5000 XL	318 HiQ 520	Ion 530	Ion Proton P1	PGM HiQ 540	RS P6-C4	Sequel	Mini ION	Prome thION	QiaGen Gene Readr	BGI SEQ 500	#
Reads: (M)	0.1	1.25	25	400	600	3000	4000	5000	6000	1400	--	5	15-20	165	60-80	5.5	38.5	0.05	--	400	--	--
Read length: (paired-end*)	400	700	300*	150*	100*	100*	125*	150*	150*	50	75	200 400	200 400	200	220	15K	12K	10K	10K	--	--	--
Run time: (d)	0.4	0.9	2	1.2	1.125	11	6	3.5	3	12	7	0.37	--	--	--	4.3	4.3	2	--	--	1	--
Yield: (Gb)	0.035	0.7	15	120	120	600	1000	1500	1800	100	180	1.2-2	6-8	10	10-15	12	84	0.5	600	80	200	--
Rate: (Gb/d)	0.2	0.75	7.5	100	106.6	55	166	400	600	8.3	30	--	--	--	--	2.8	19.5	0.25	--	--	--	--
Reagents: ($K)	1.1	6.2	1	4	6.145	23.47	29.9	29.9	12.75	9	10.5	0.6	--	1	1.2	2.4	11.2	1	--	0.5	--	--
per-Gb: ($)	31K	8K	93	33.3	51.2	39.1	29.9	20	7	90	58.33	--	--	100	--	200	80	2000	20	--	--	--
Machine: ($)	110K	500K	99K	250K	740K	690K	690K	900K	1M	500K	595K	50K	65K	243K	242K	695K	350K	1000	30K	--	--	--

#Page maintained by http://www.vilellagenomics.com #Editable version: tinyurl.com/ngsspecsshared

#curl "https://docs.google.com/spreadsheets/d/1GMMfhyLK0-q8Xklo3YxlWaZA5vVMuhU1kg41g4xLkXc/export?gid=4&format=csv" | grep -v '^#' | column -t -s\, | less -S

Albert Vilella @AlbertVilella · 18h

Updated NGS specs @BGI_Genomics @PacBio @illumina @thermofisher @nanopore @QIAGENscience tinyurl.com/ngsspecs

1 5

Next Generation Sequencing technologies

	Strenghs	Weaknesses
454 Pyrosequencing (GS FLX Titanium XL+)	– Long reads (450/700 bp). – Long insert for mate pair libraries (20Kb). – Low observed raw error rate (0.1%) – Low percentage of PCR duplications for mate pair libraries	– Homopolymer error. – Low sequence yield per run (0.7 Gb). – Preferred assembler (gsAssembler) uses overlapping methodology.
Illumina (HiSeq 2500)	– High sequence yield per run (600 Gb) – Low observed raw error rate (0.26%)	– High percentage of PCR duplications for mate pair libraries. – Long run time (11 days) – High instrument cost (~ $650K)
Illumina (MiSeq)	– Medium read size (250 bp) – Faster run than Illumina HiSeq	– Medium sequence yield per run (8.5 Gb)
SOLID (5500xl system)	– 2-base encoding reduce the observed raw error rate (0.06%)	– 2-base color coding makes difficult the sequence manipulation and assembly. – Short reads (75 bp)
Ion Torrent (Ion Proton I)	– Fast run (2 hours) – Low instrument cost ($80K). – Medium read size (200 bp)	– Medium sequence yield per run (10 Gb) – Medium observed raw error rate (1.7%)
PacBio (PacBioRS)	– Long reads (3000 bp) – Fast run (2 hours)	– Really high observed raw error rate (12.7%) – High instrument cost (~ $700K) – No pair end/mate pair reads



Next Generation Sequencing technologies

	Inputs	Outputs
454 Pyrosequencing (GS FLX Titanium XL+)	– Single Reads Library. – Pair End Library (3 to 20 Kb insert size). – Multiplexed sample.	– sff files – (fasta and fastq files)
Illumina (HiSeq 2500)	– Single Reads Library. – Pair End Library (170-800 bp insert size). – Mate Pair Library (2 to 10 Kb insert Size) – Multiplexed sample.	– fastq files (Phred+64) – fastq files (Phred+33, Illumina 1.8+)
Illumina (MiSeq)		
SOLID (5500xl system)	– Single Reads Library. – Mate Pairs Library (0.6 to 6 Kb insert size). – Multiplexed sample.	– fastq files (Phred+33)
Ion Torrent (Ion Proton I)	– Single Reads Library. – Pair End Library (0.6 to 6 Kb insert size). – Multiplexed sample.	
PacBio (PacBioRS)	– Single Reads Library.	

Types of reads

★ Library types (orientations):

- Single reads

- Pair ends (PE) (150-800 bp insert size)

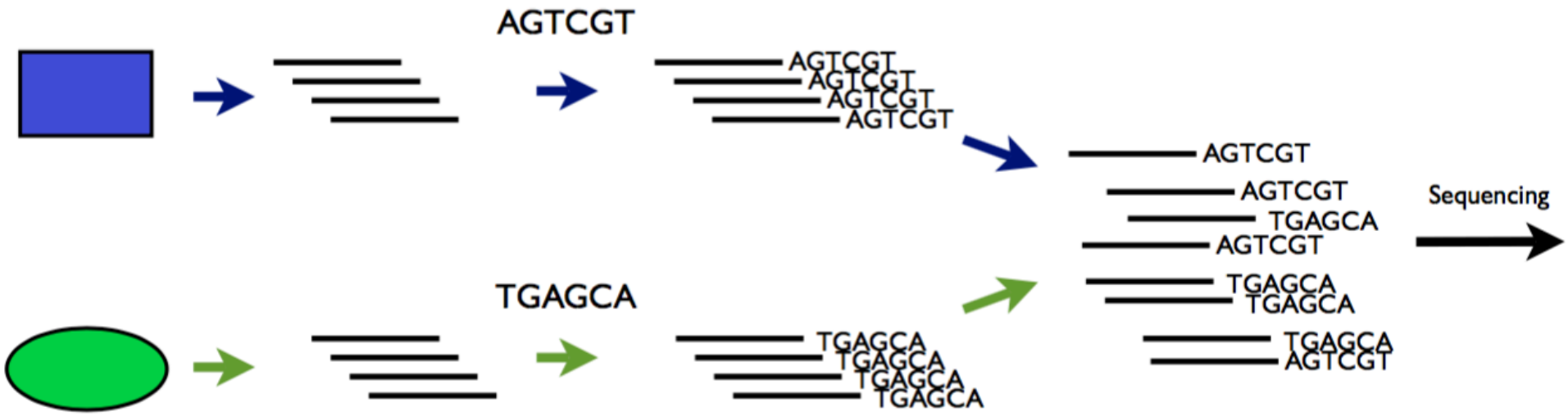
Illumina

- Mate pairs (MP) (2-20 Kb insert size)

Illumina

454/Roche

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



Fastq files:

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

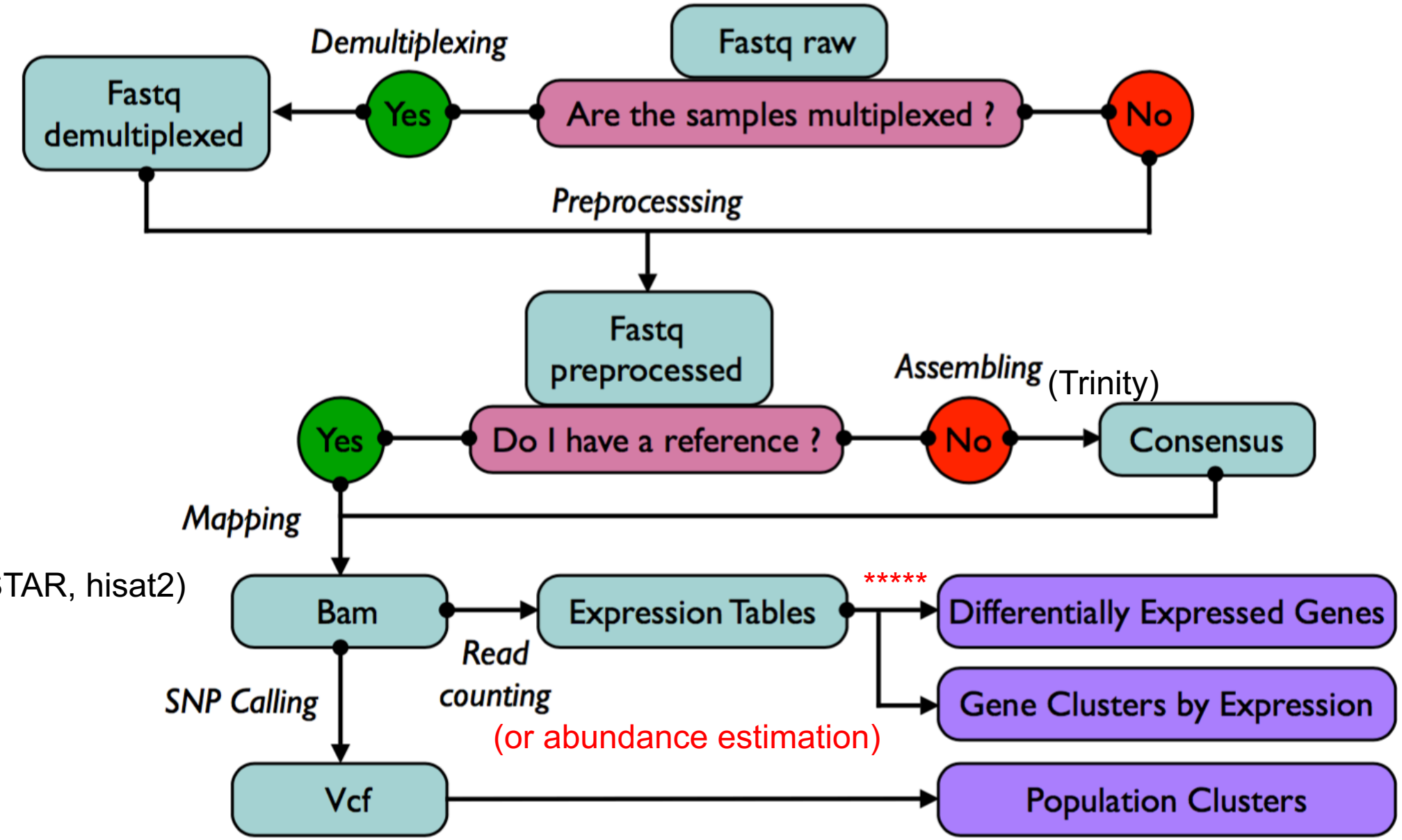
-Wikipedia

```
@SEQUENCE_ID1
ATGCGCGCGCGCGCGCGCGGGTAGCAGATGACGACACAGAGCGAGGATGCGCTGAGAGTA
GTGTGACGACGATGACGGAAAATCAGA
+
BBBBBPPPPPXXXXX^^^^^^^^^^^^^^^^^^^^^^^^__^^^^^^^^^^^^_eeeeeee
[[[[[^^^^]]]]]XXXXXPPPPPBBBBB
```

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

Once you have the raw data

General workflow



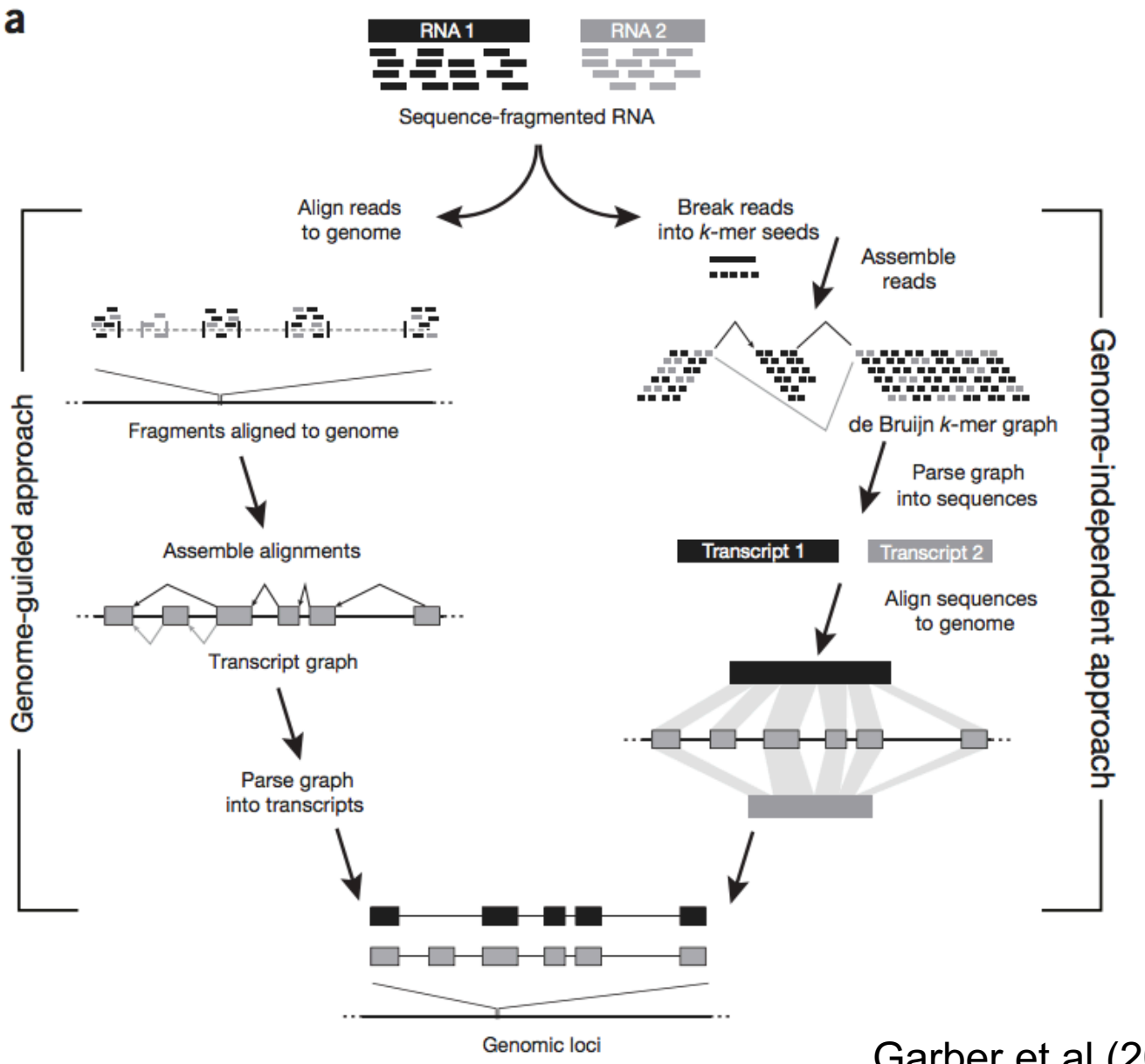
General workflow for RNAseq to produce annotation

Options:

- Align and then assemble
- Assemble and then align

Align to

- Genome
- Transcriptome (if no genome)



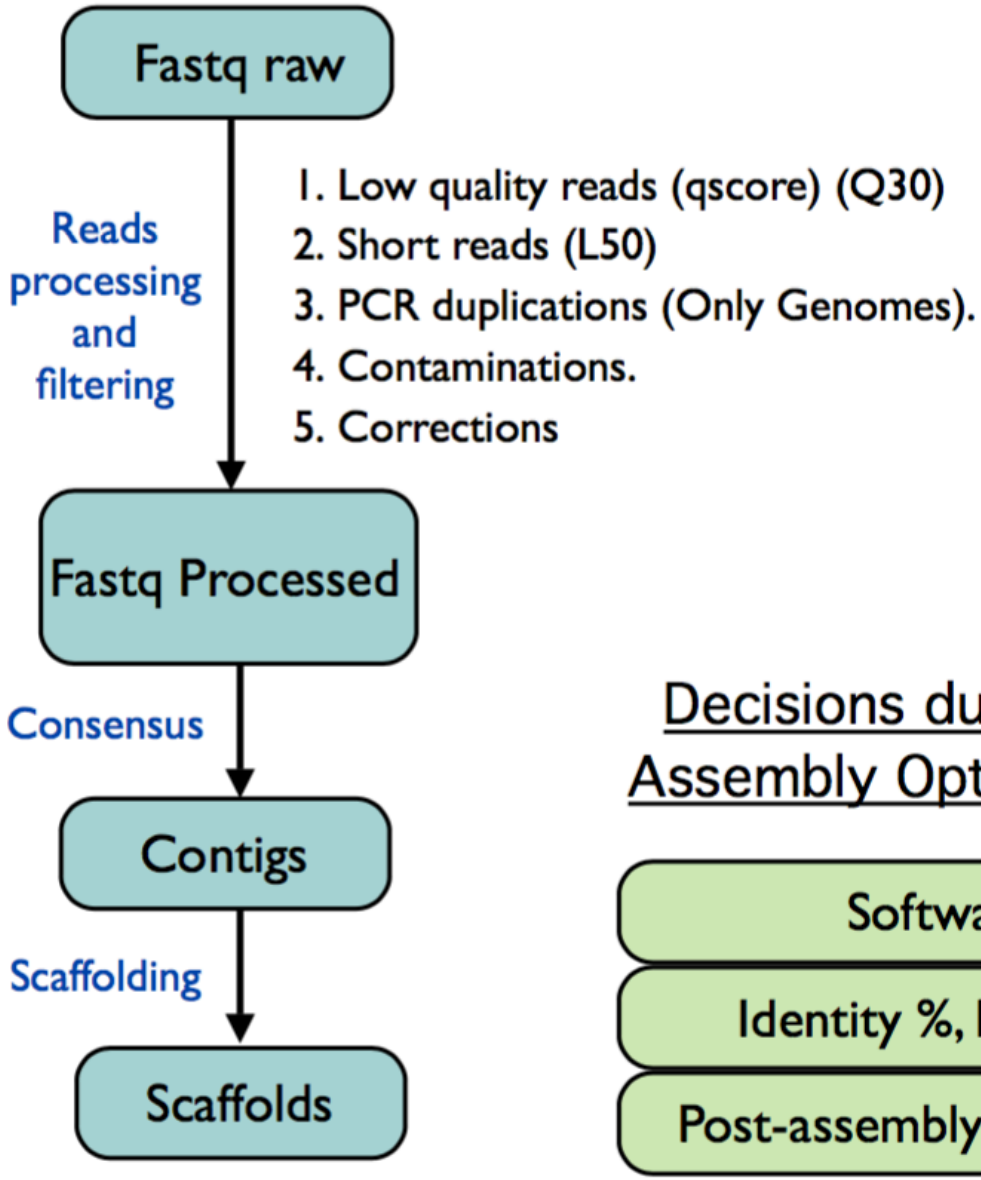
Garber et al (2010)

De novo assembly of transcriptomes

Transcriptome *de novo* assembly

Decisions during the Experimental Design

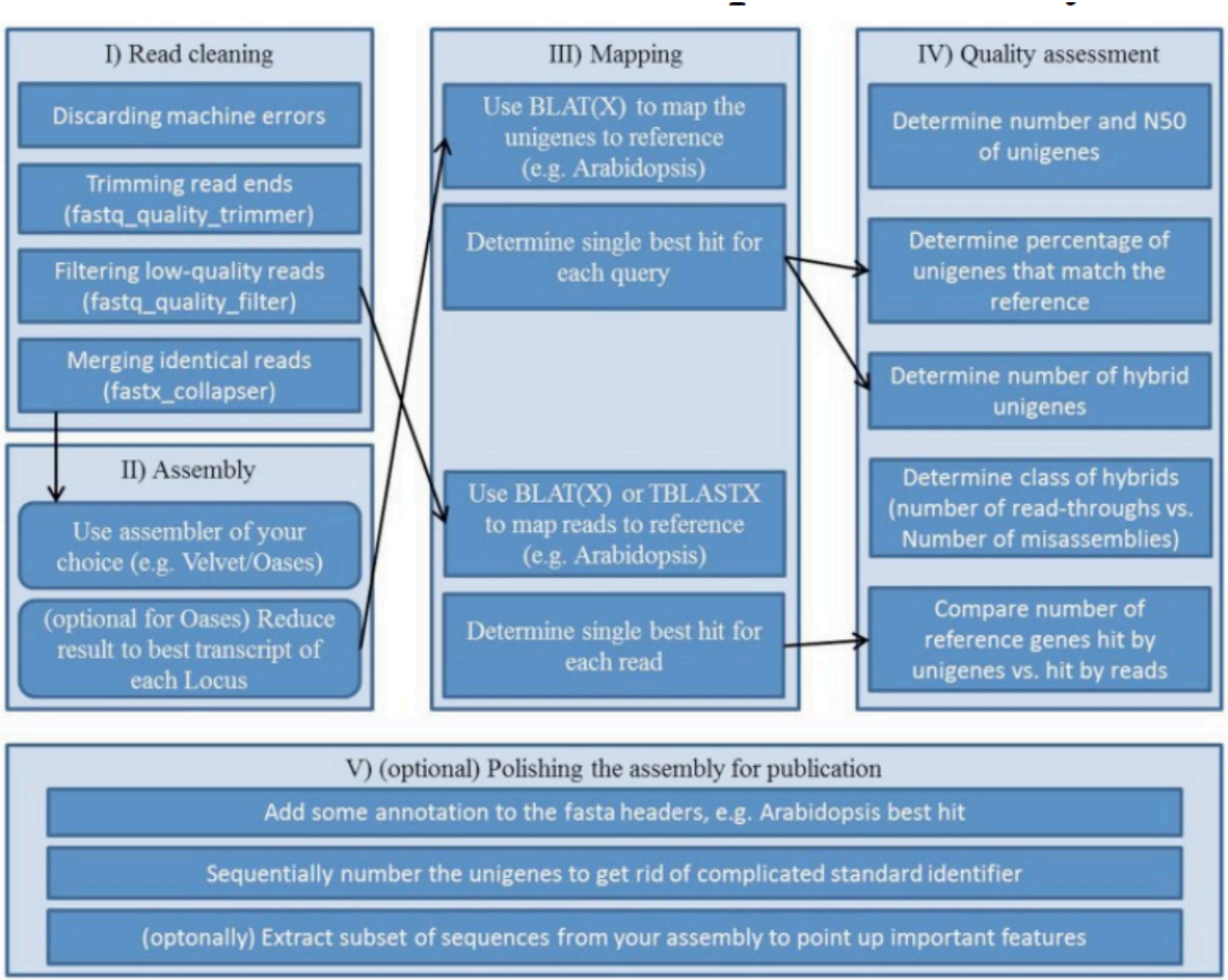
- Technology
- Library Preparation
- Sequencing Amount



Decisions during the Assembly Optimization

- Software
- Identity %, Kmer ...
- Post-assembly Filtering

Workflow



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*

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/

* Comparisons in the Article: Vijay N. *et al* (2012) *Molecular Ecology* DOI: 10.1111/mec.12014

But transcriptome assembly always produce more transcripts than expected

	Species	Estimated genome size	Total length of reads	Assembler	Number of transcripts	Number of transcripts with protein homologues
	Bactrocera dorsalis	~414 Mbp	2.4 Gbp	Velvet + Oases	71,722 (>100bp)	29,067
	Bactrocera cucurbitae	~375 Mbp	8.1 Gbp	Trinity	55,141 (>100 bp)	25,370
	Plutella xylostella	~394 Mbp	14 Gbp	Trinity	67,002 (>200 bp)	27,144
	Ambystoma mexicanum	~30 Gbp	4.9 Gbp	SOAPdenovo	116,787	39,200
	Syrmaticus mikado	~1 Gbp	7.6 Gbp	Trinity	93,617 (>300 bp)	26,703
	Lophura swinhoii	~1 Gbp	9.6 Gbp	Trinity	142,371 (>300 bp)	38,690

Prof. Chien-Yu Chen

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

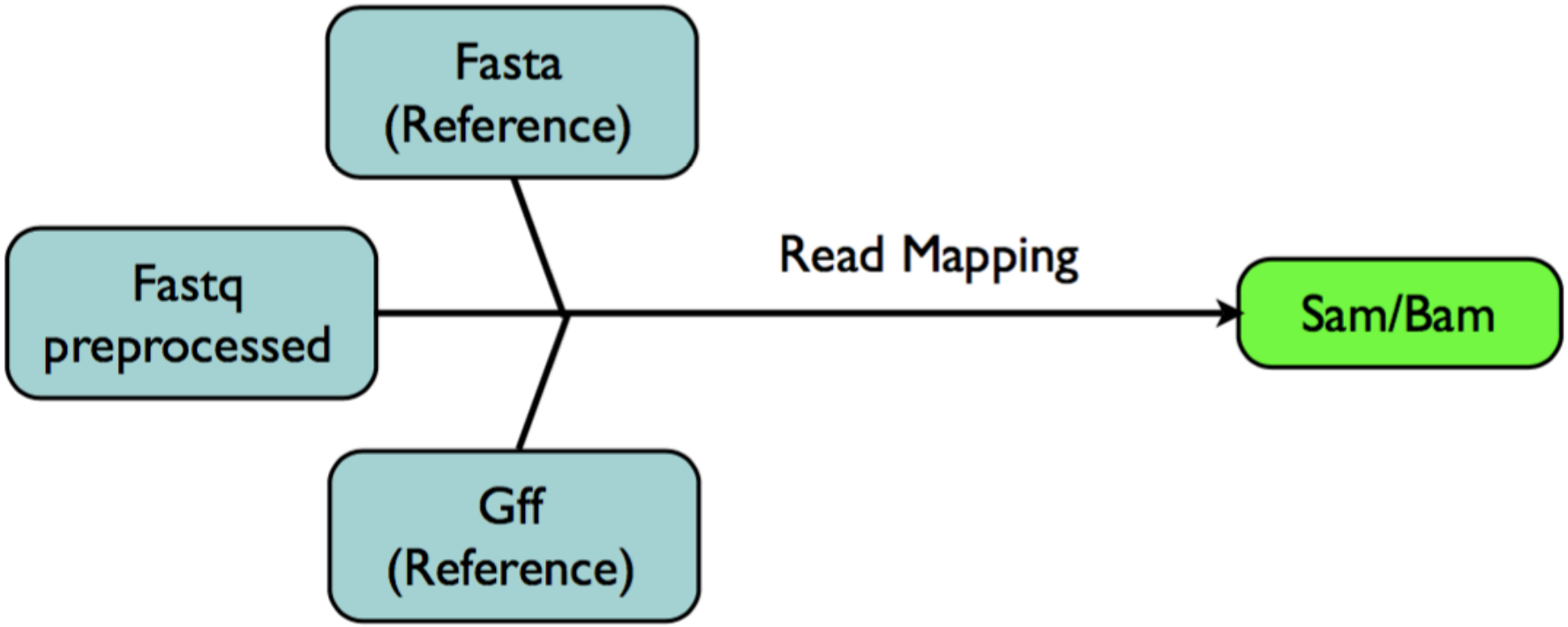
RNAseq read mapping

Read mapping

1. **Fasta** file with genome sequence
2. **Gff** file with gene model annotations

```
##gff-version   3
##sequence-region   ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
```

http://www.sequenceontology.org/resources/gff3.html

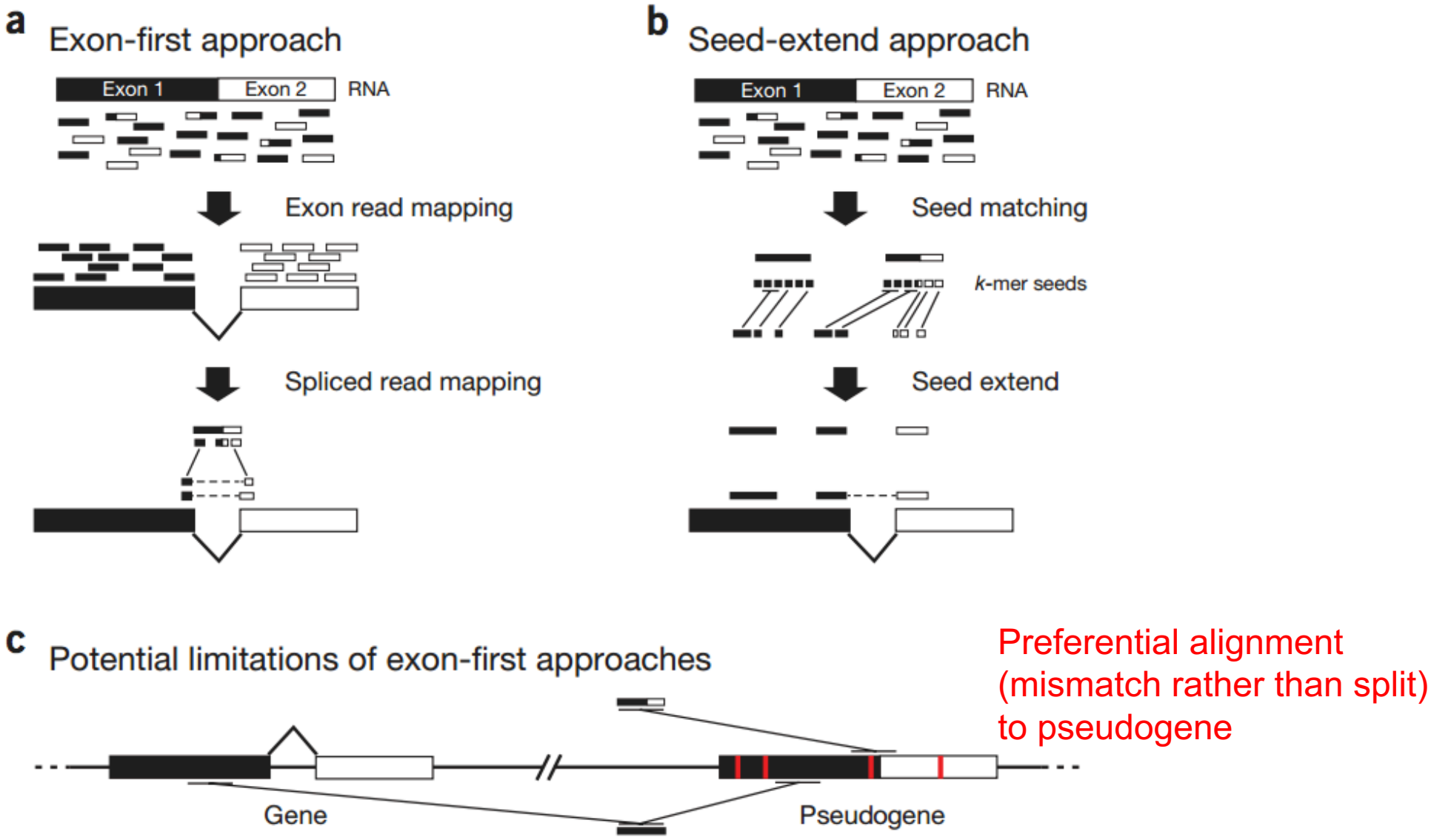


Note about the hardware and mapping software:

+ 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.

RNAseq aligners

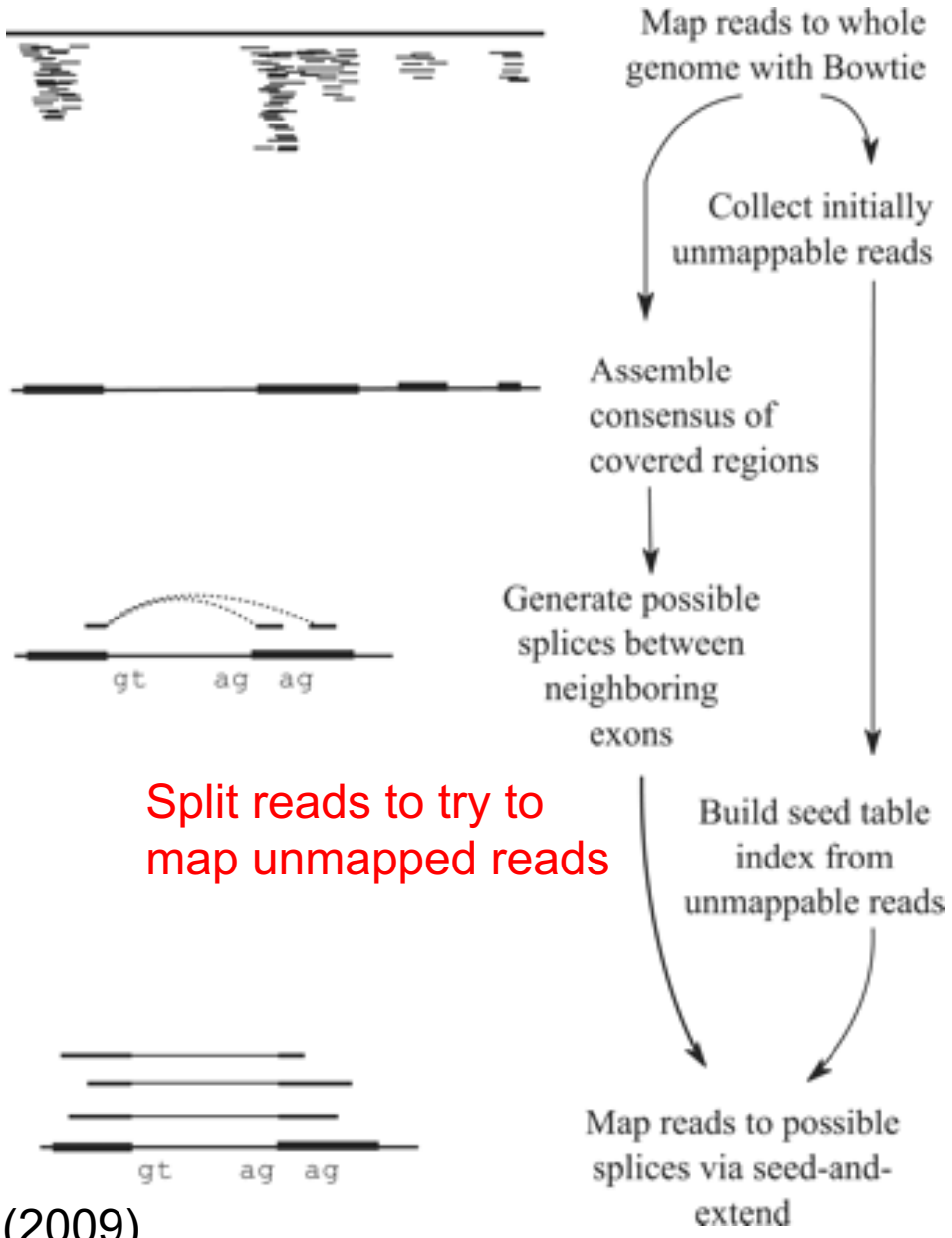
Strategies for gapped alignments of RNAseq reads



Preferential alignment (mismatch rather than split) to pseudogene

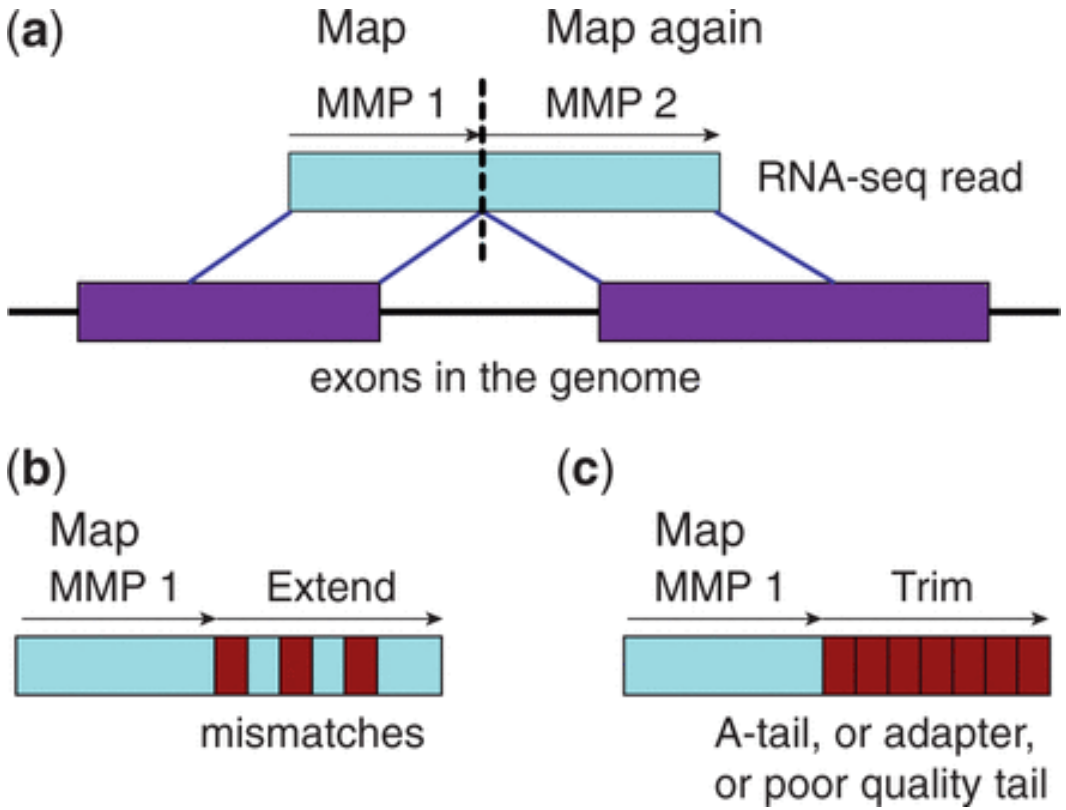
Garber et al (2010)

Tophat and STAR (different ways to handle split reads)



Split reads to try to map unmapped reads

Trapnell et al (2009)

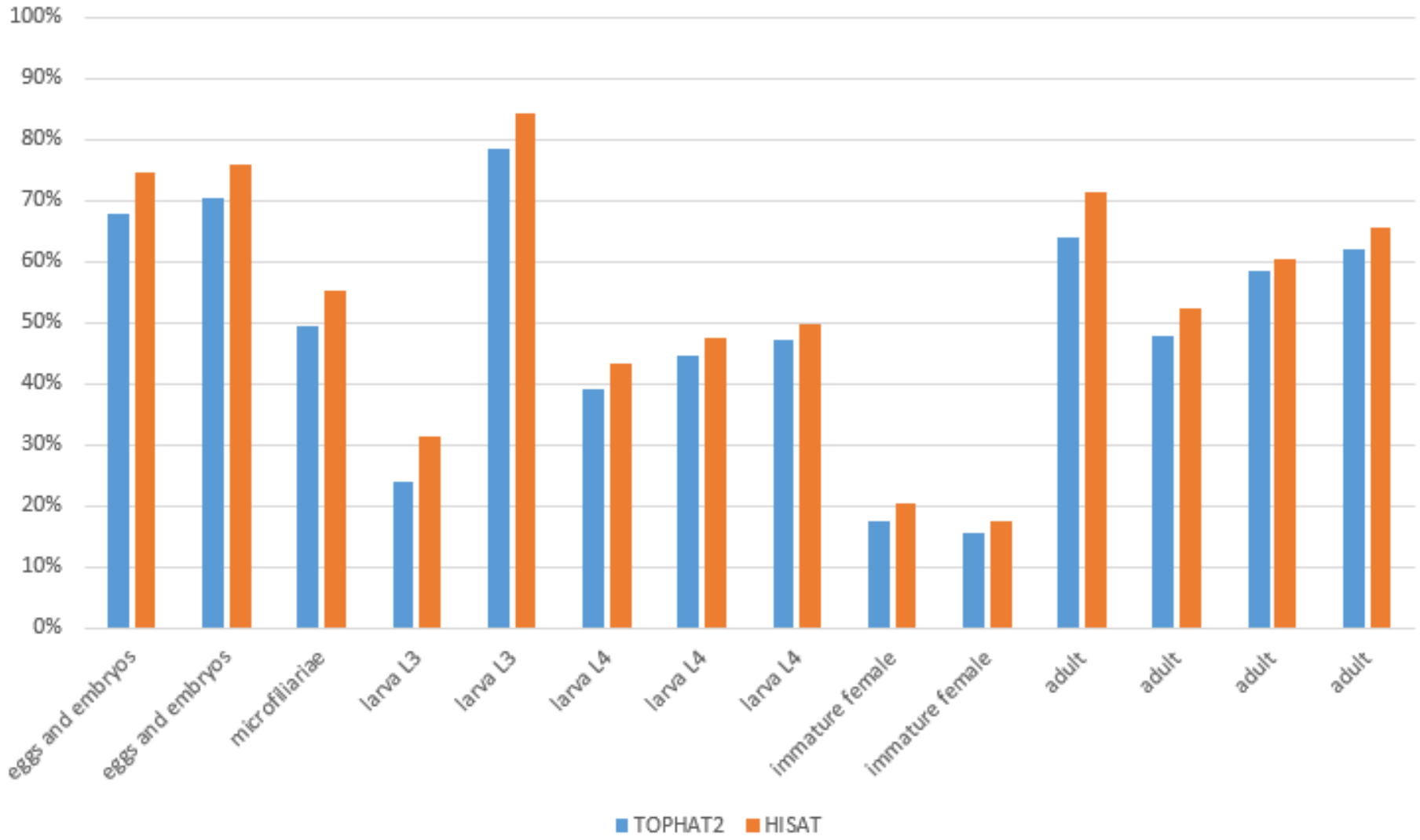


Dobin et al (2013)

Tophat2 is being obsoleted

As reads get longer, better methods are implemented

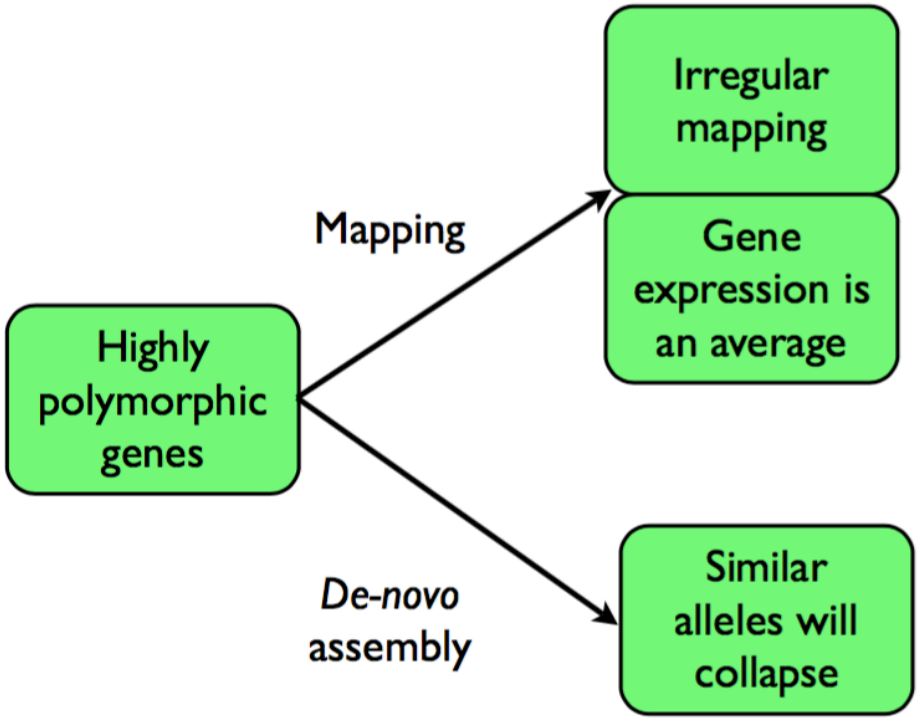
This leads to more reads get mapped and in a much faster speed



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**

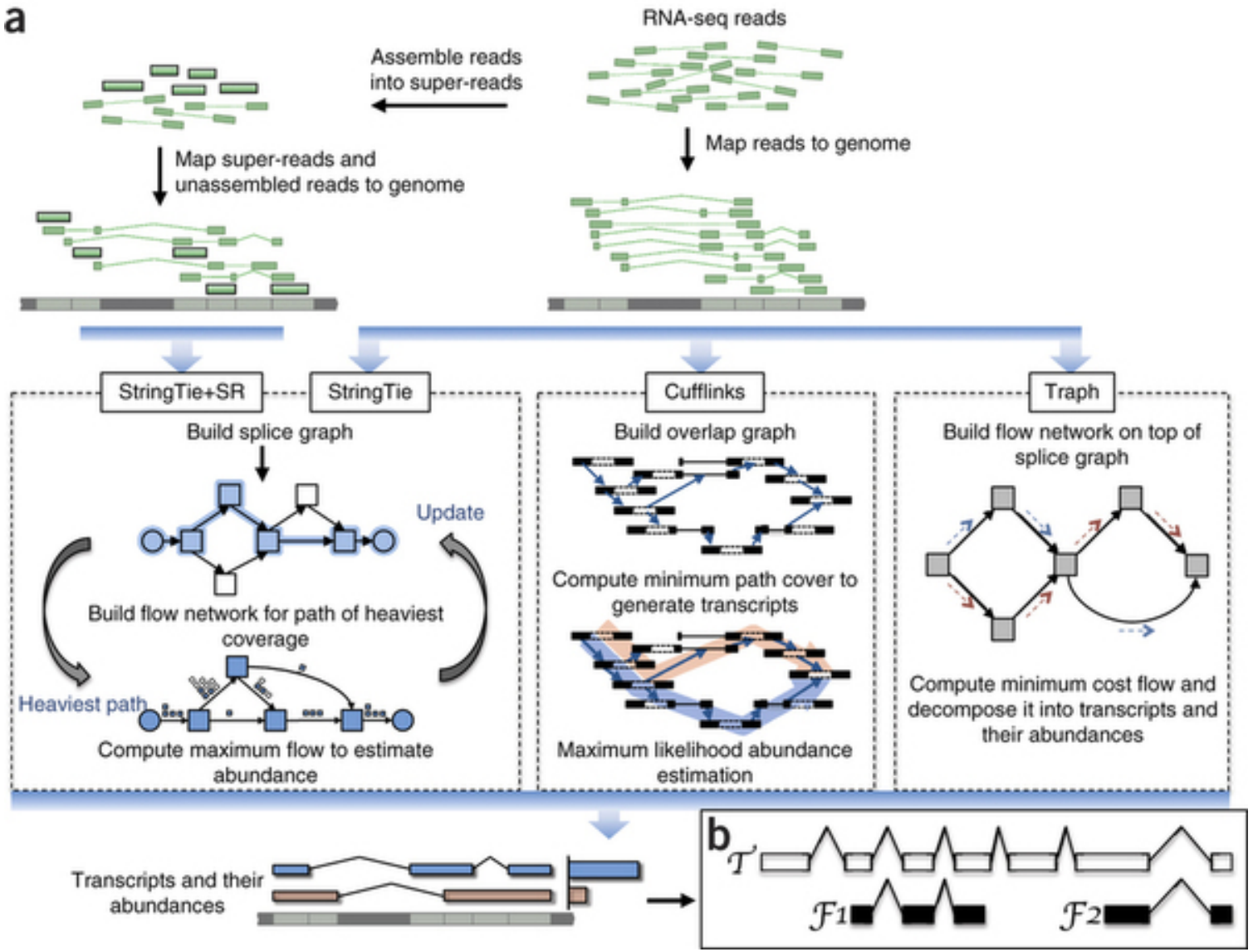
CTTGACGACATGACGAC
CCCTTGACGACATGACG **Gene I B**
CGCCCTTGACGACATGA

CGCCCTTGACGACATGACGACA

Collapsed consensus Gene A + Gene B

Transcript reconstruction

Cufflinks and StringTie



Pertea et al (2015)

Genome annotation pipelines

Table 4.1.5 Genome Annotation Pipelines

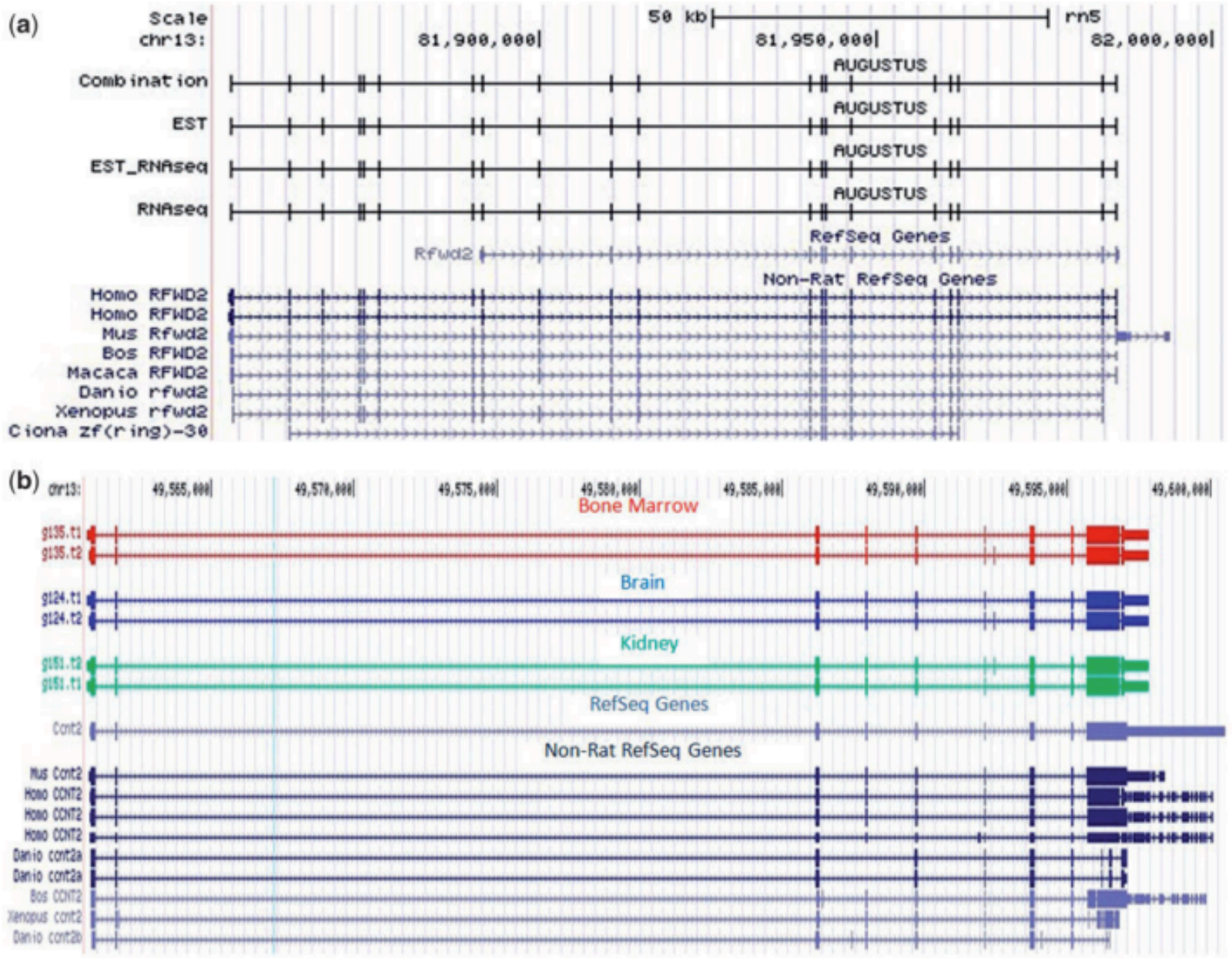
Software package	Features	Reference
EuGene	Annotation pipeline that integrates multiple evidence types using a C++ based plugin system.	(Foissac et al., 2008)
MAKER	Annotation pipeline that aligns and polishes protein and transcriptome data with BLAST and Exonerate, provides evidence-based hints to gene 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; *UNIT 4.11*; 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.	(*UNIT 1.15*; 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.	(*UNIT 1.3*; 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

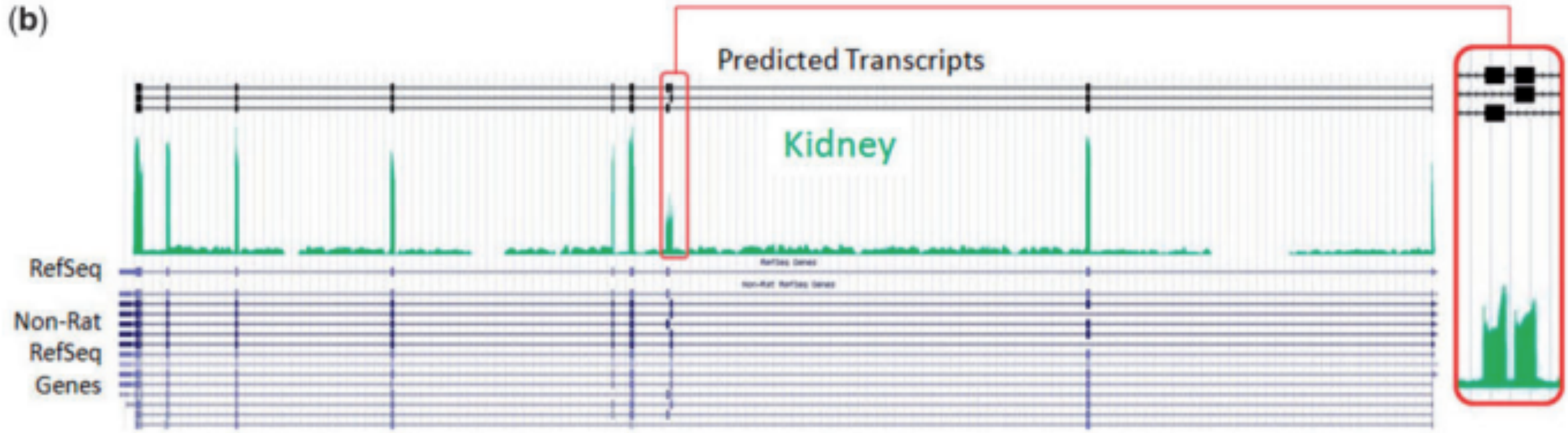
Software package	Features	Reference
WebApollo	Web-based plug-in for Jbrowse with an editable user-created annotation track. Edits are visible in real time to all currators.	(Lee et al., 2013)
Argo	Stand-alone Java application for viewing and editing gene annotations.	(see Internet Resources)
IGV	Genome 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



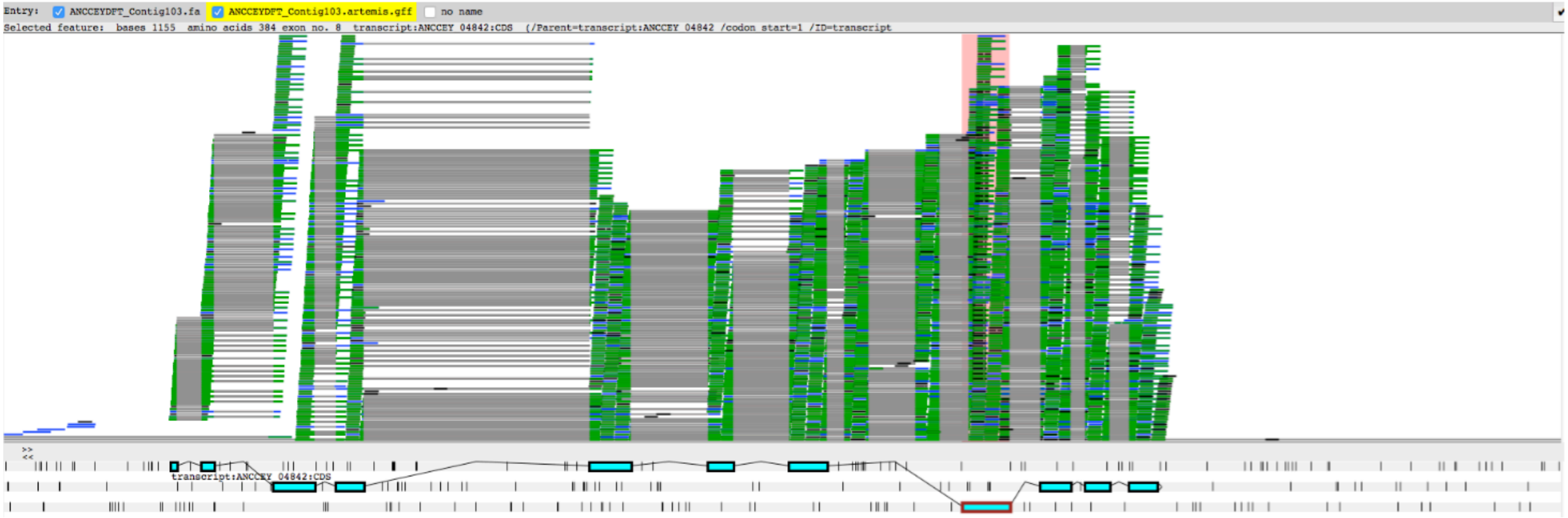
Li et al (2015)

Novel isoforms

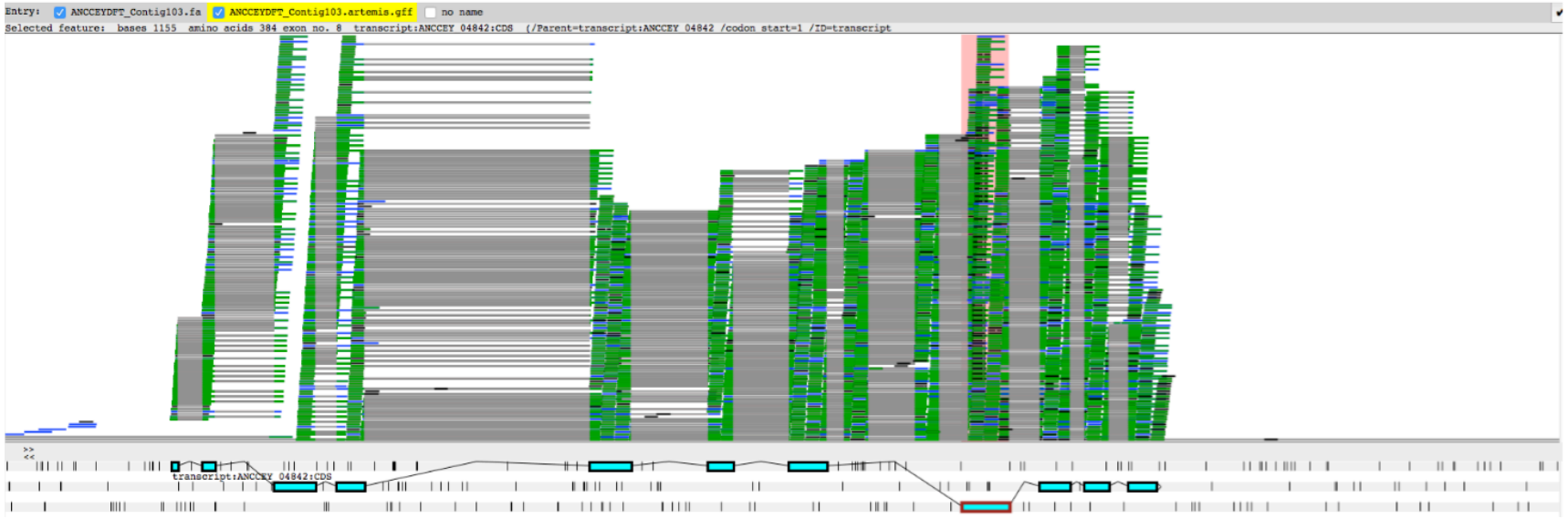


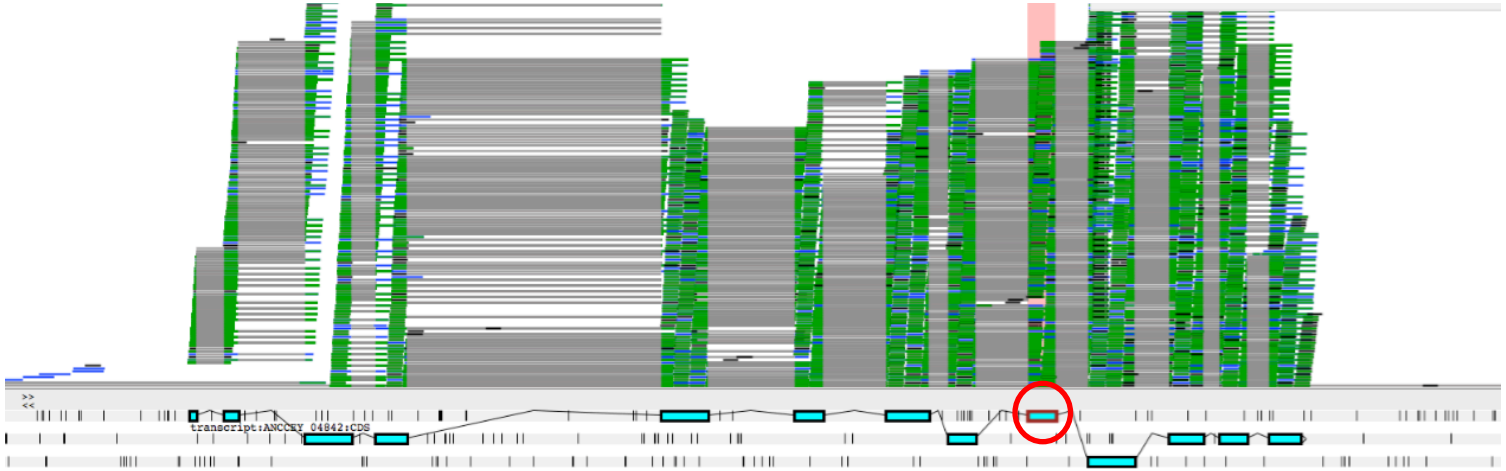
Li et al (2015)

Artemis (will be taught later)



Manual curation using artemis





Hueimien Ke

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 Lange[1], Alexandre Lomsadze[3], Mark Borodovsky[2,3,4,*] and Mario Stanke[1]

Nov. 2015

Functional annotation

Functional annotation

Genome

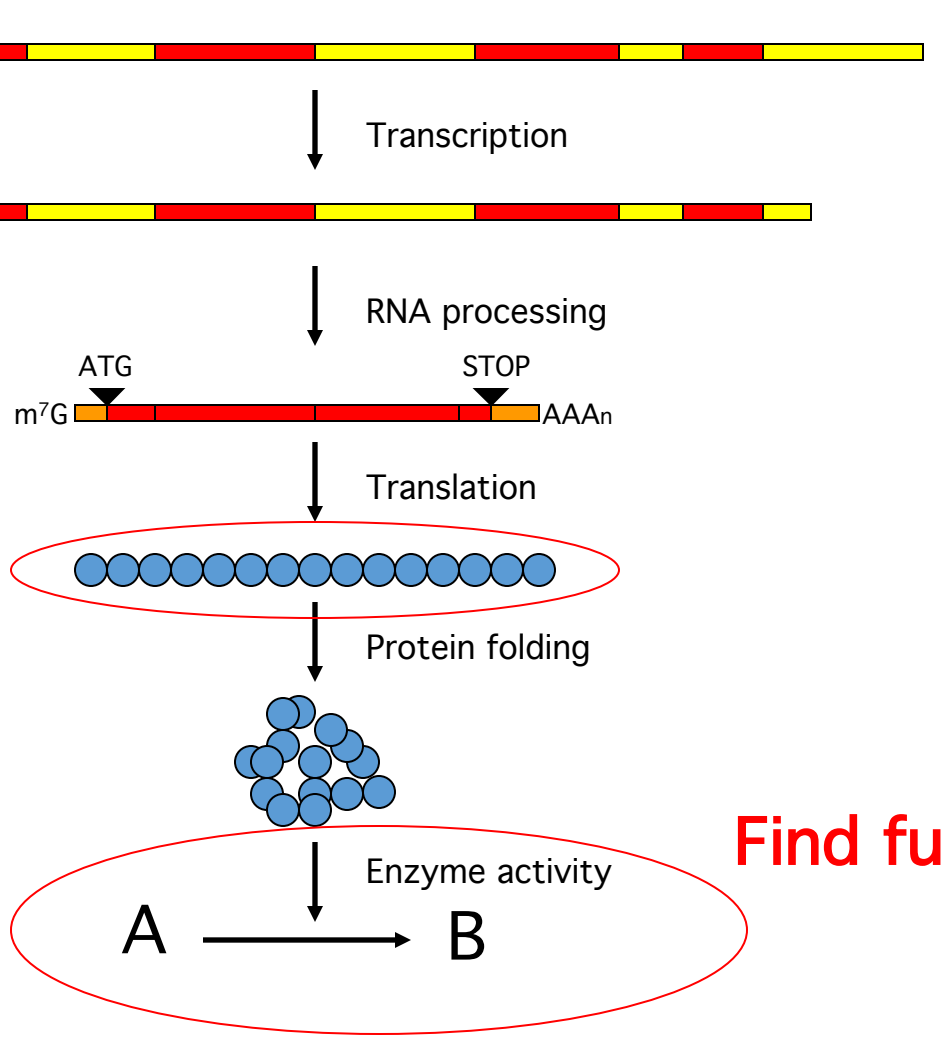
Primary Transcript

Processed mRNA

Polypeptide

Folded protein

Functional activity



Find function

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

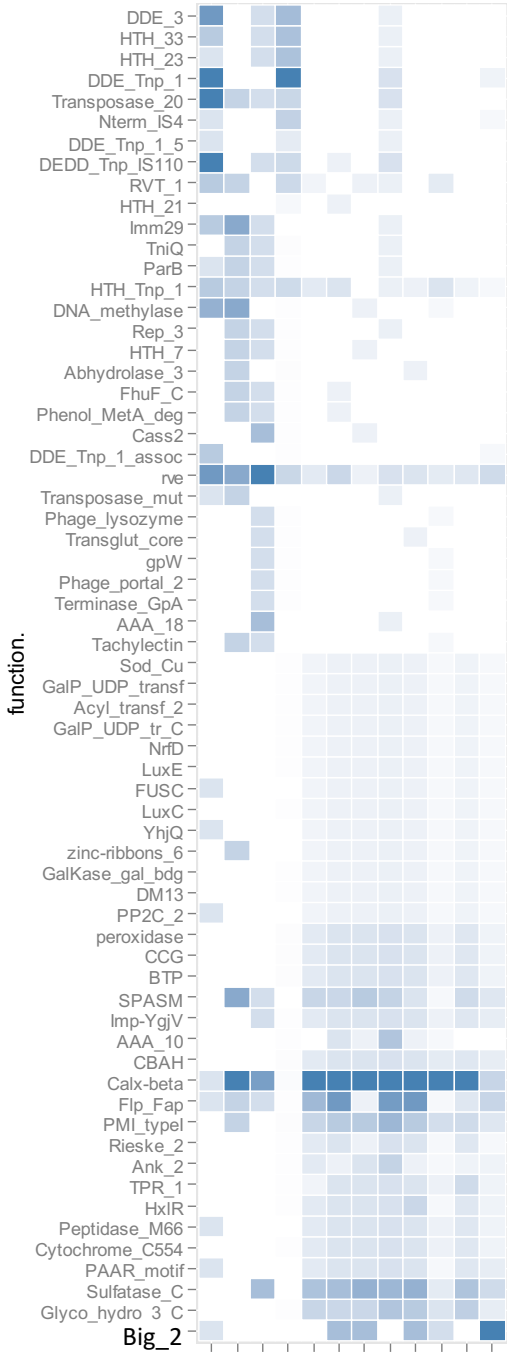
Visually assess the **top 5-10 hits** to identify whether these have been assigned a function

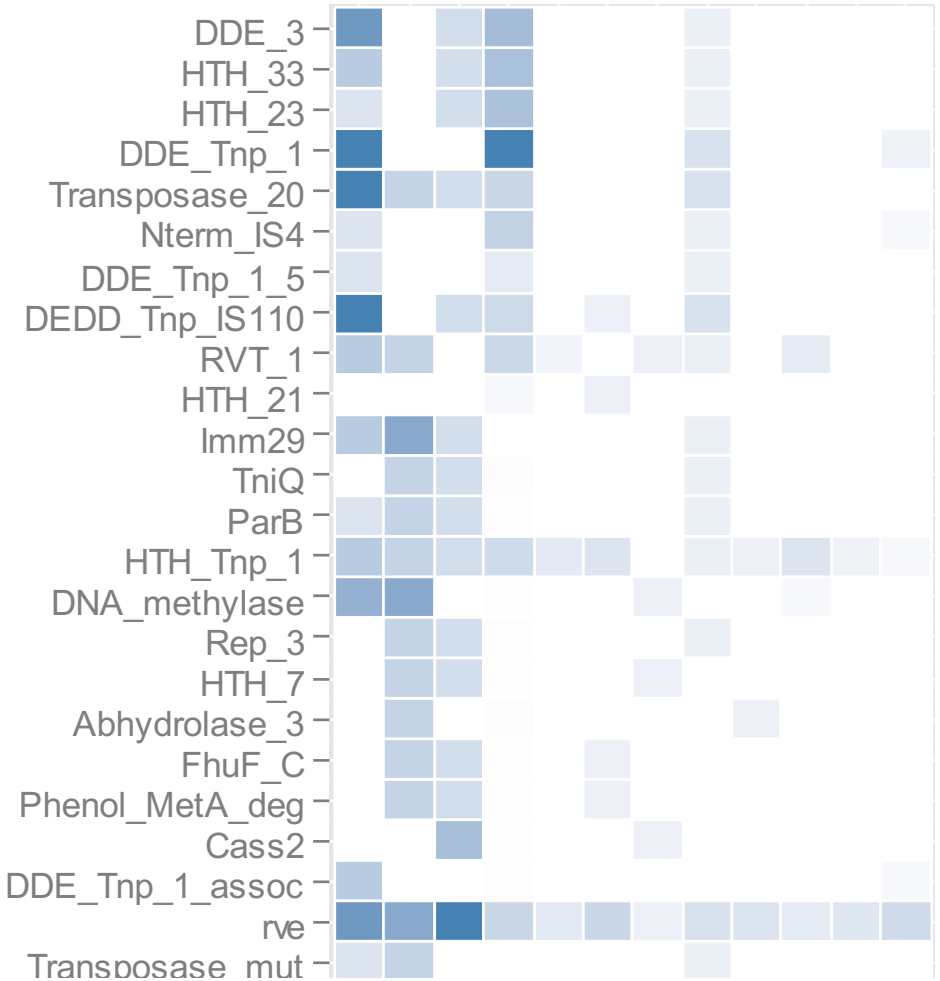
Functions (**and names**) are assigned

Other features which can be determined

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

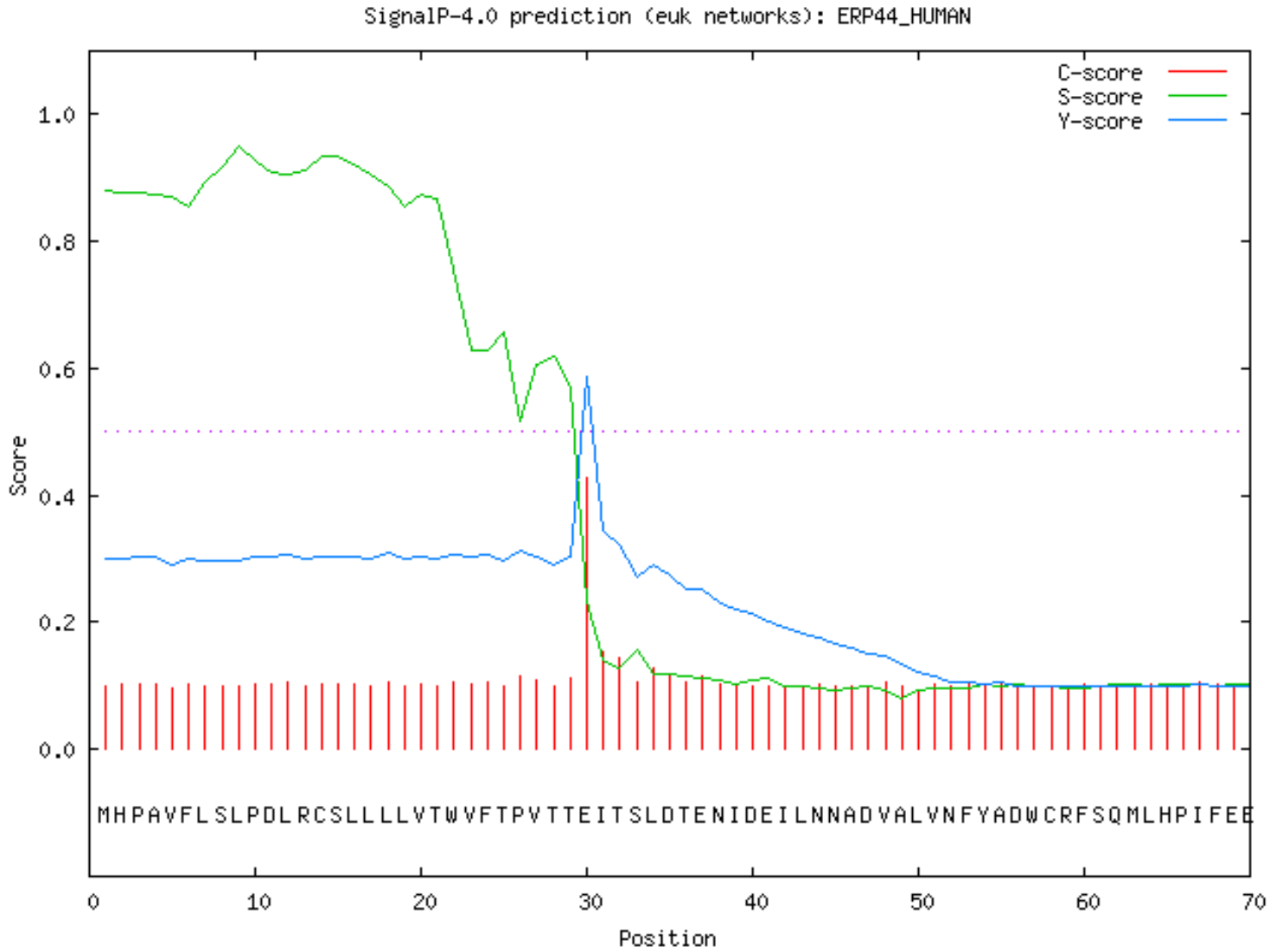
PFAM





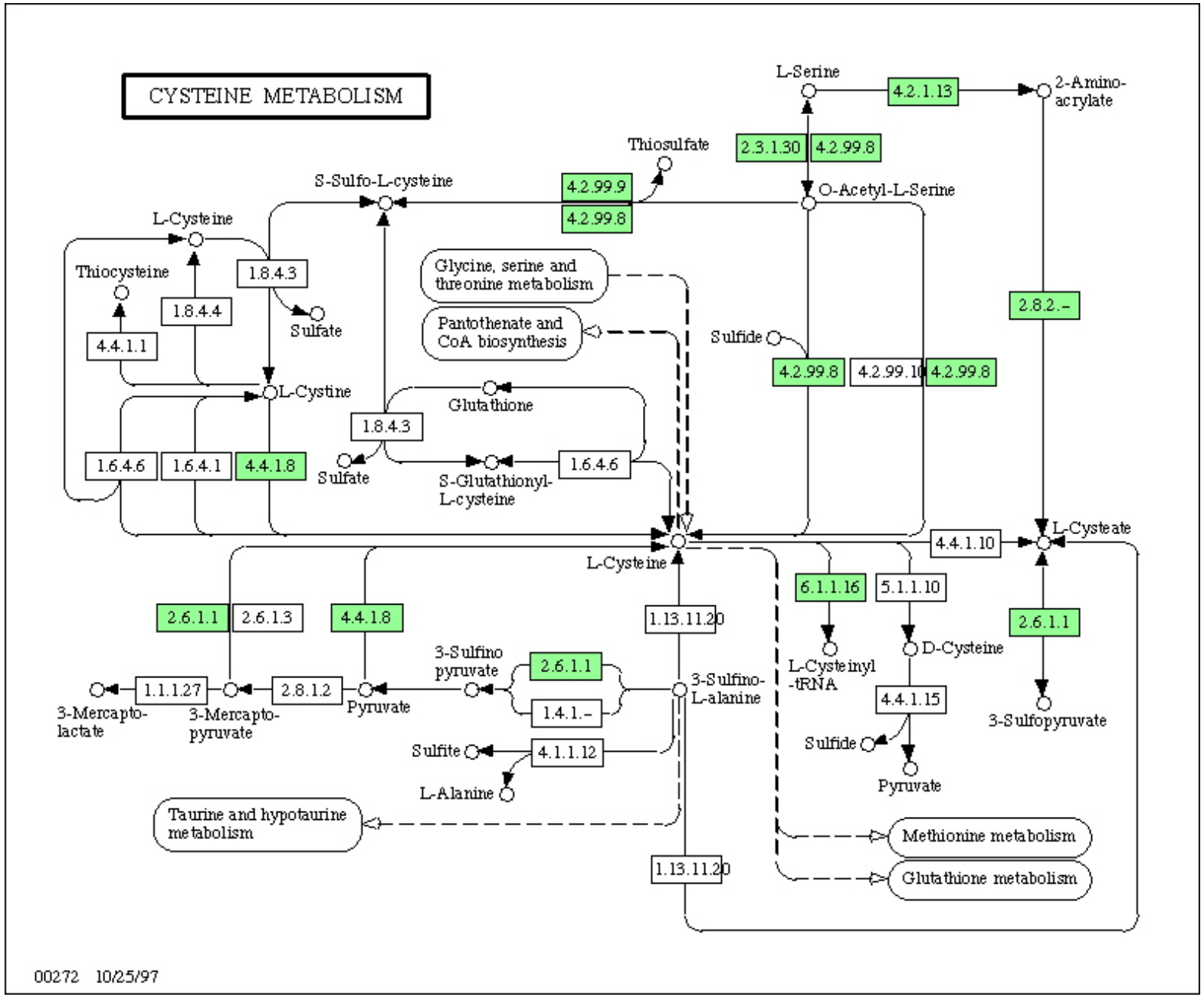
Hueimien Ke

SignalP: predicts the presence and location of signal peptide



KEGG

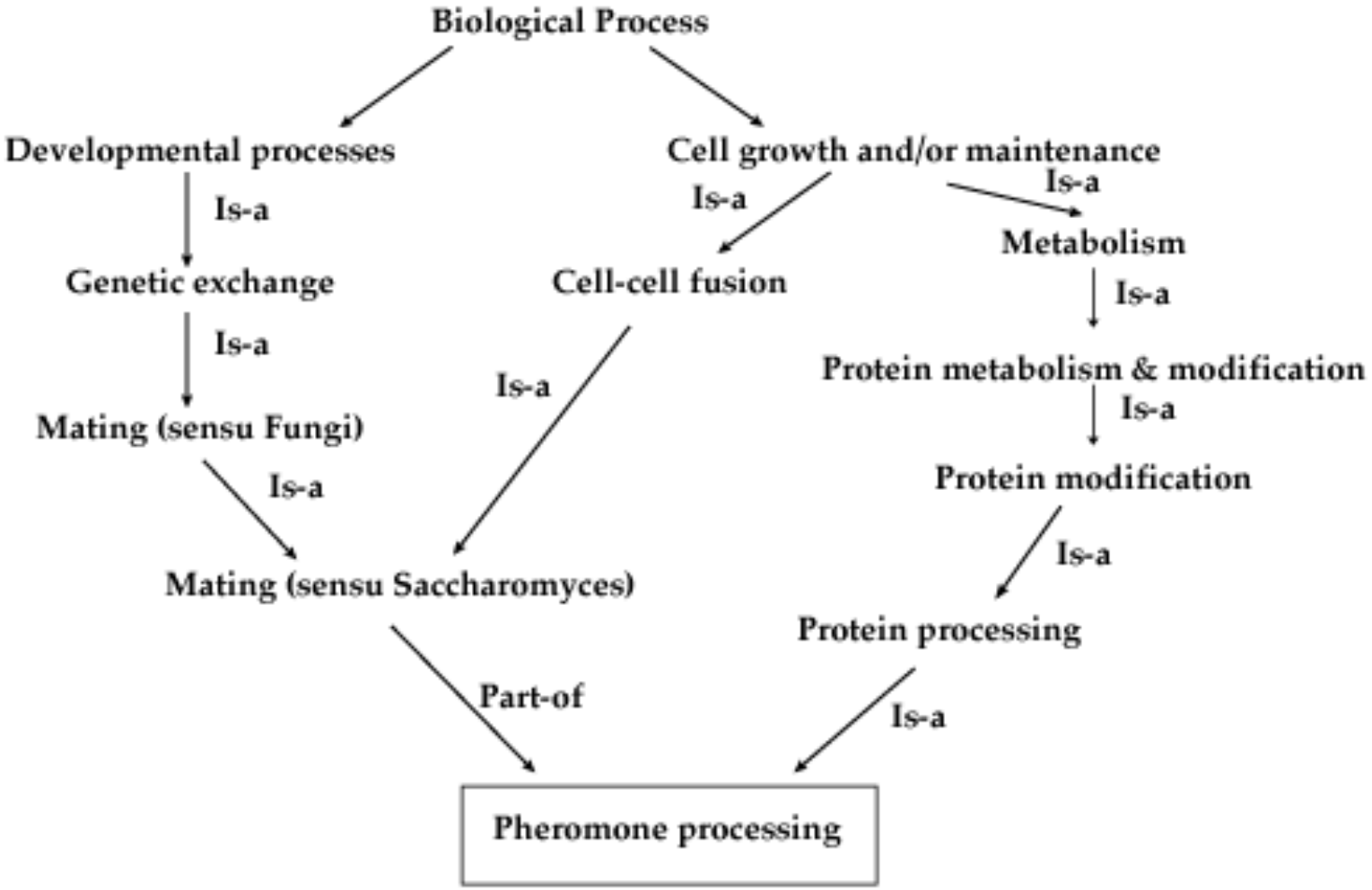
Help improve annotation by showing missing genes in essentail pathways



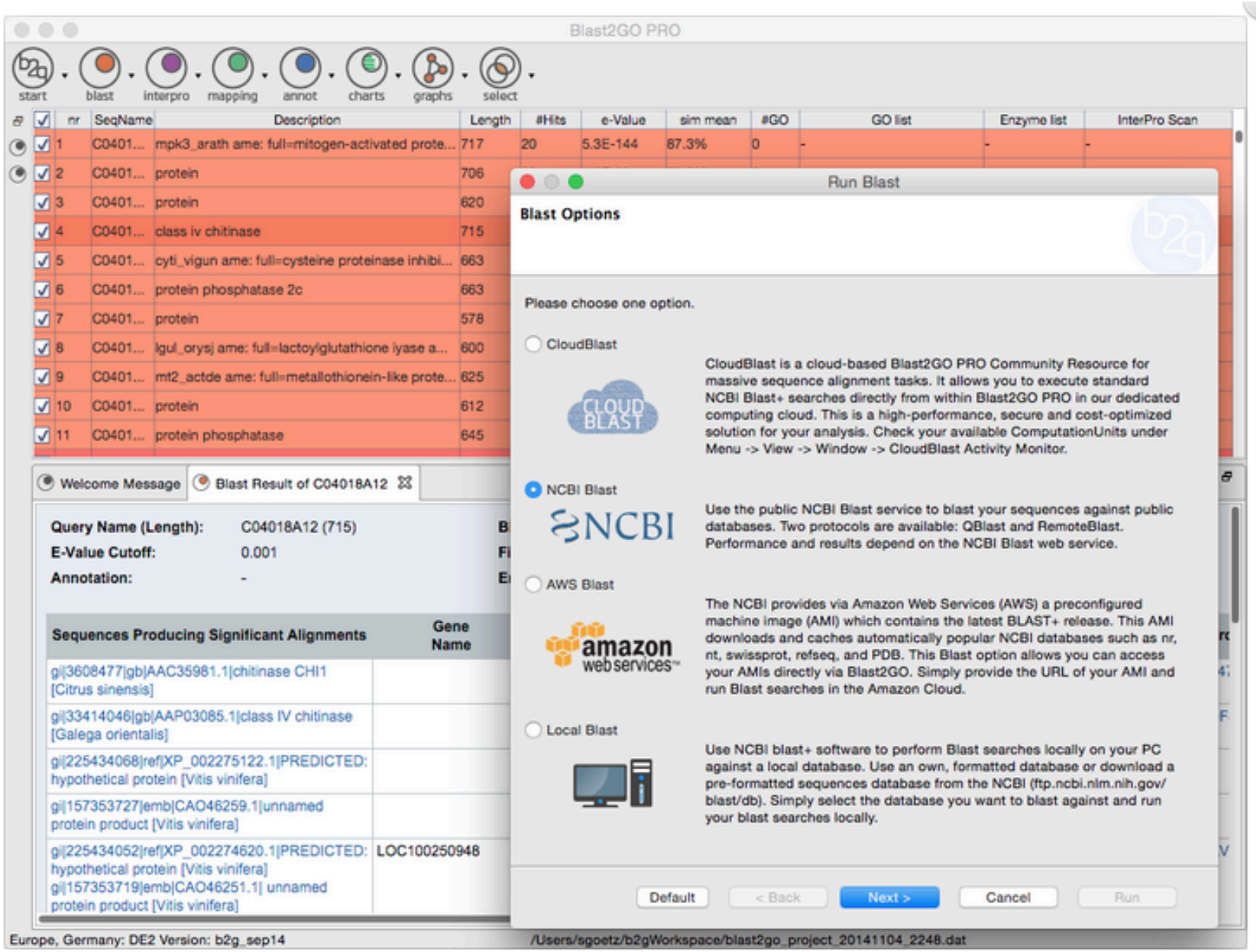
Gene Ontology

- 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



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

When everything’s done well and tested: the case of bacteria annotation

Key bacterial features

- tRNA
  - easy to find and annotate: anti-codon
- rRNA
  - easy to find and annotate: 5s 16s 23s
- CDS
  - straightforward to find candidates
    - false positives are often small ORFs
    - wrong start codon
  - partial genes, remnants
  - pseudogenes
  - assigning function is the bulk of the workload

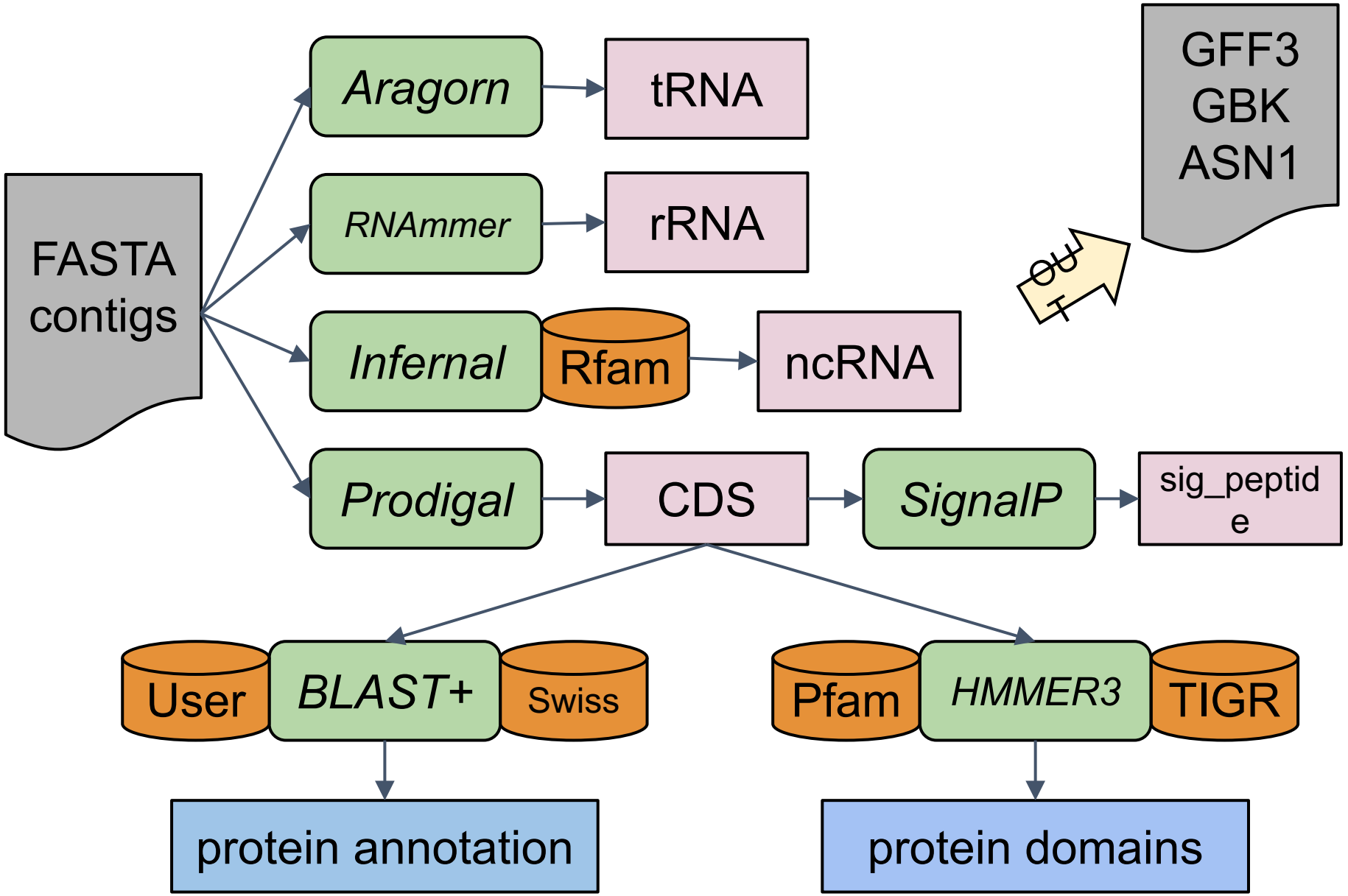
Automatic annotation in bacteria is possible

Two strategies for identifying coding genes:

- **sequence alignment**
  - find known protein sequences in the contigs
    - transfer the annotation across
  - will miss proteins not in your database
  - may miss partial proteins

- ***ab initio* gene finding**
  - find candidate open reading frames
    - build model of ribosome binding sites
    - predict coding regions
  - may choose the incorrect start codon
  - may miss atypical genes, overpredict small genes

Prokka pipeline (simplified)



Core bacterial proteome

- Many bacterial proteins are conserved
  - experimentally validated
  - small number of them
  - good annotations
- Prokka provides this database
  - derived from UniProt-Swissprot
  - only bacterial proteins
  - only accept evidence level 1 (aa) or 2 (RNA)
  - reject "Fragment" entries
  - extract /gene /EC_number /product /db_xref
- First step gets ~50% of the genes
  - BLAST+ blastp, multi-threading to use all CPUs

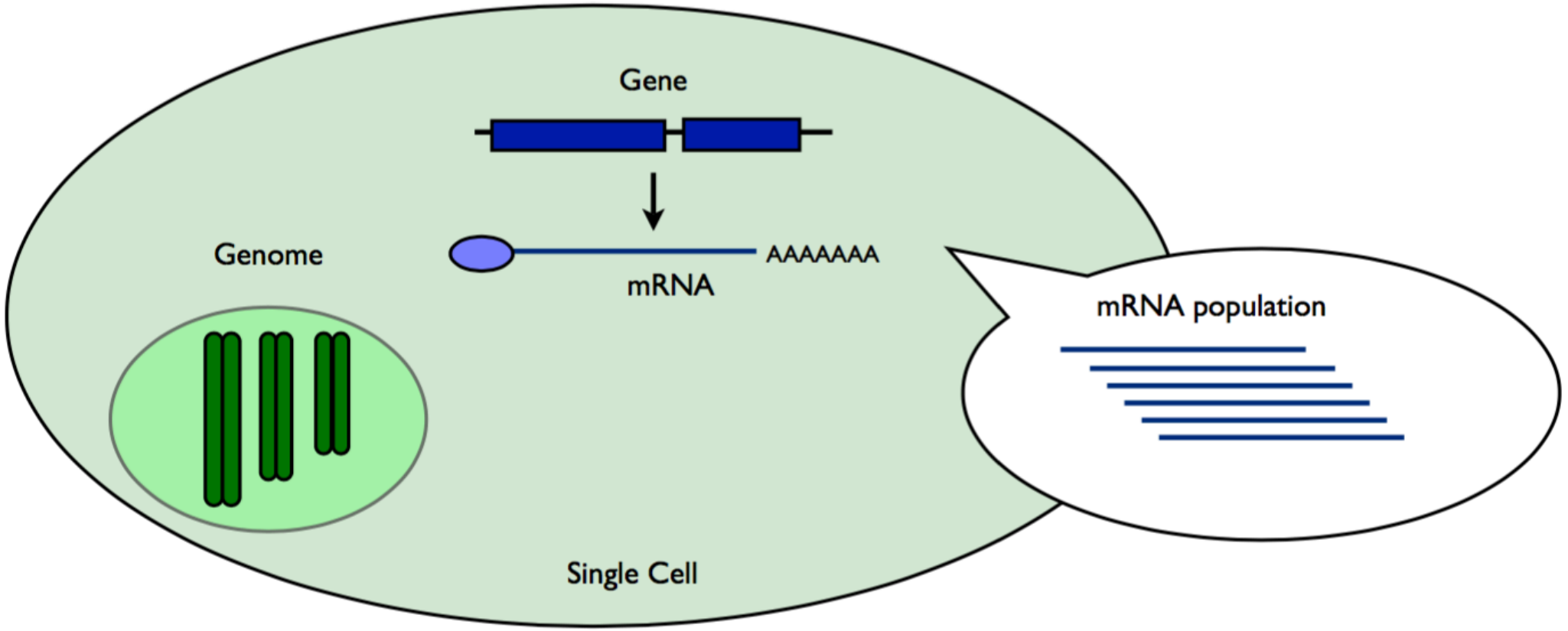
Differential expression

Types of experiments

Transcriptome Complexity:

Simple System:

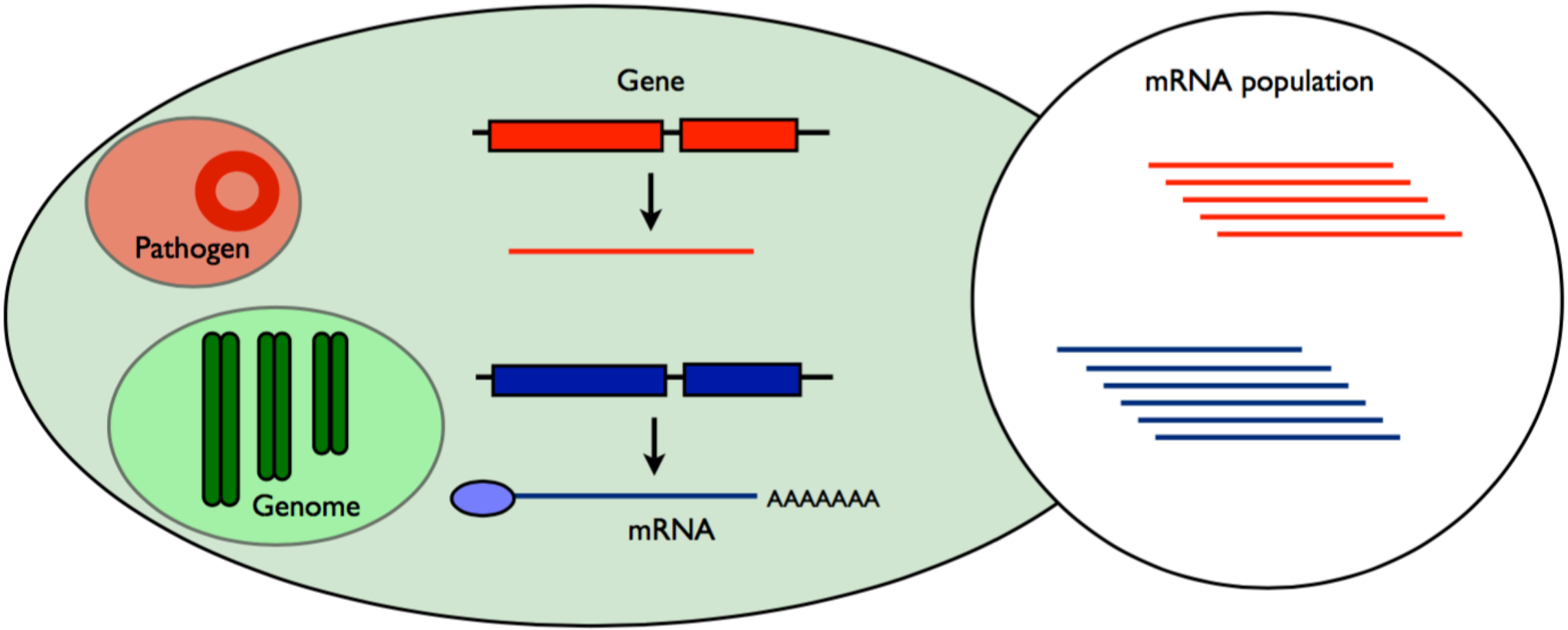
One Genome => Gene 1 copy => Single mRNA



How many species we are analyzing ?

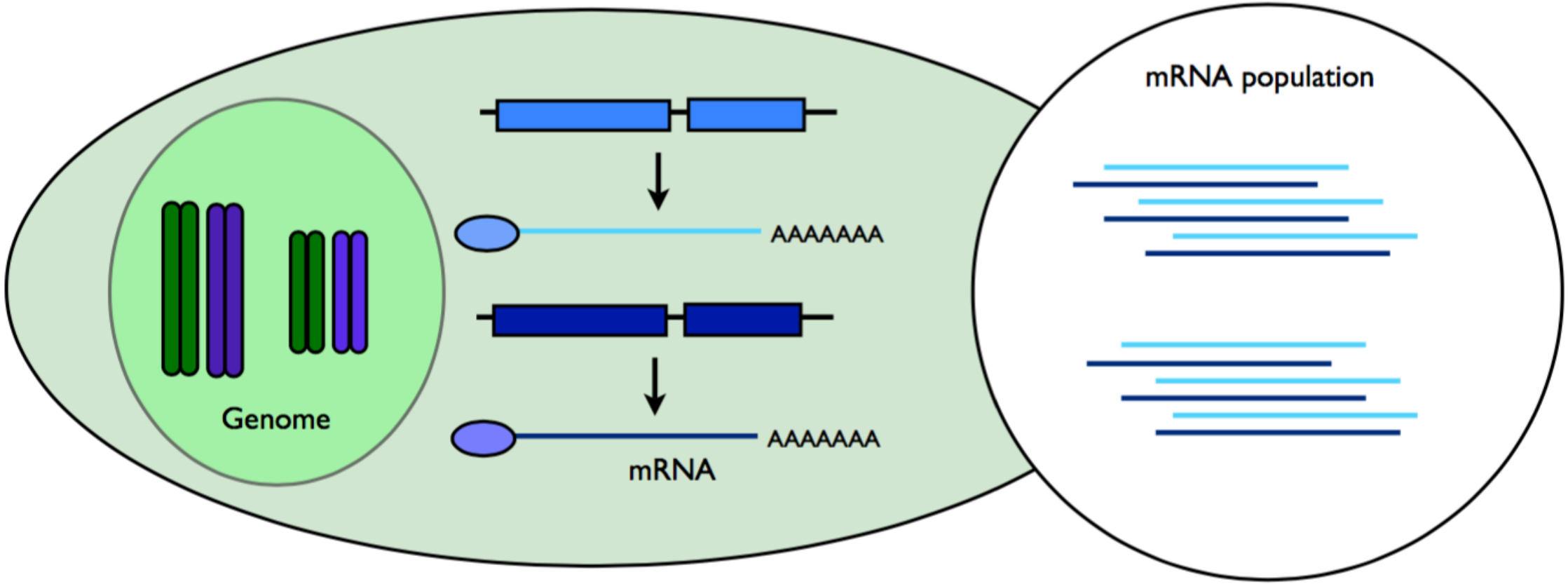
1) Problems to isolate a single species (rhizosphere)

2) Species interaction study (plant-pathogen)



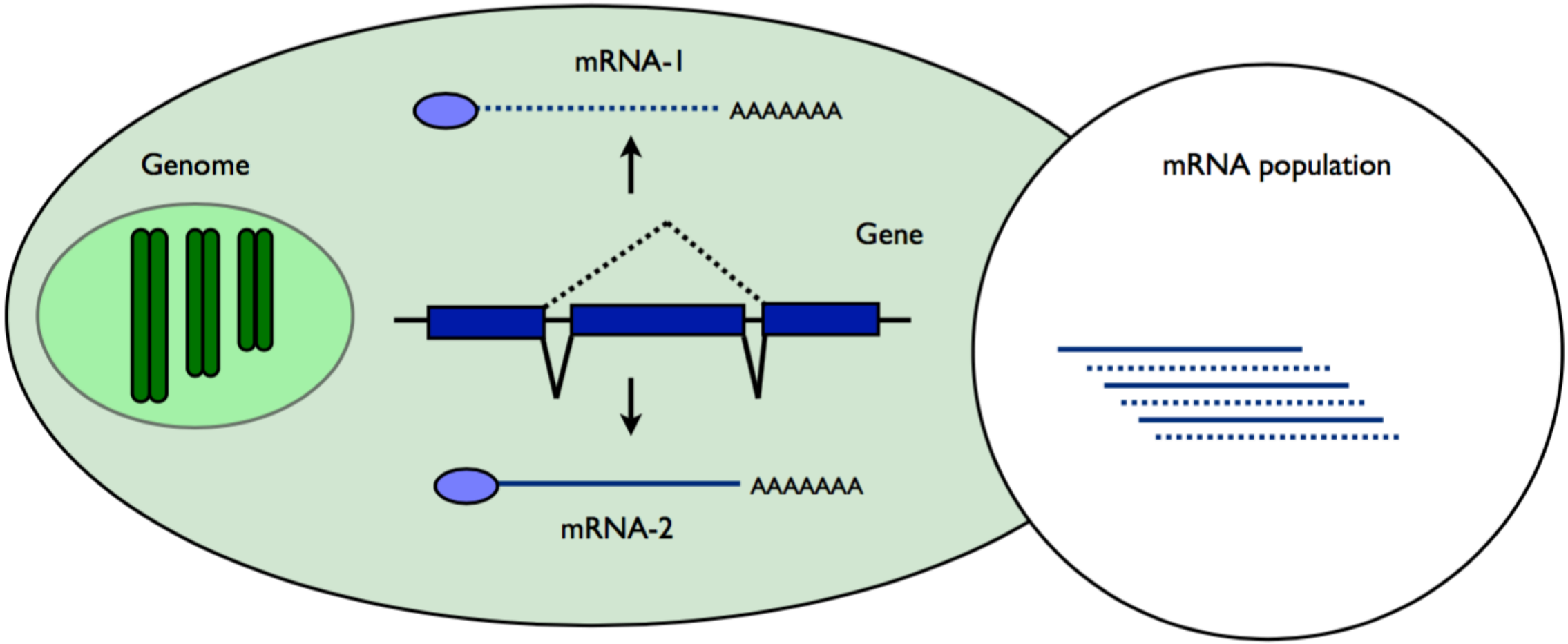
How many possible alleles we expect per gene ?

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



How many isoforms we expect for each allele ?

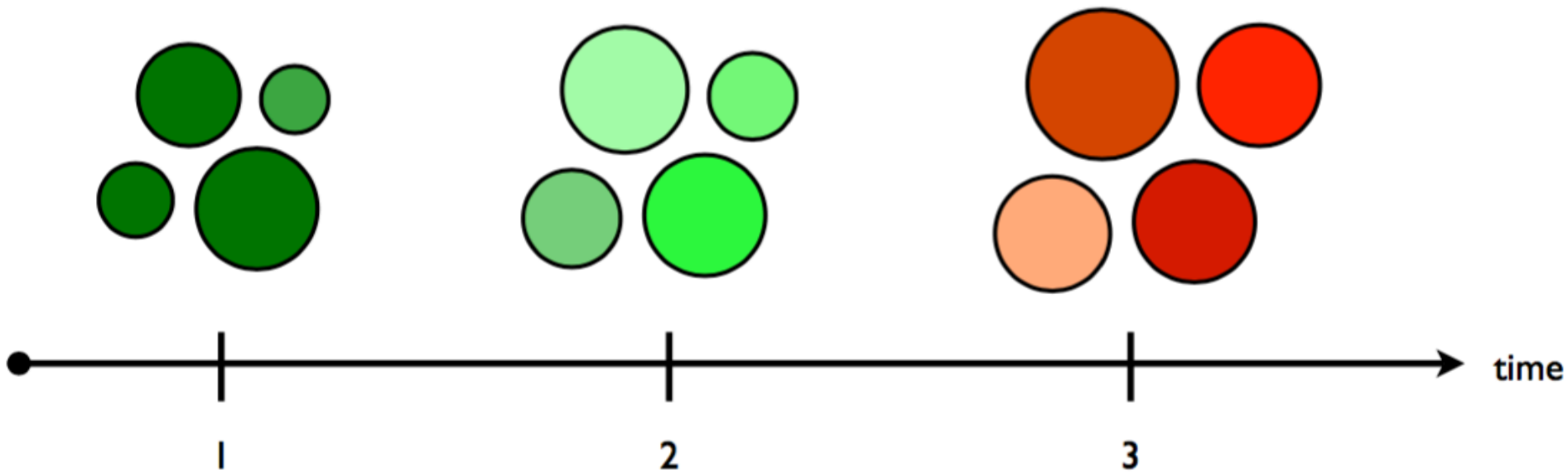
1) Alternative splicings



Is the study performed at different time points?

1) Developmental stages (difficult to select the same)

2) Response to a treatment

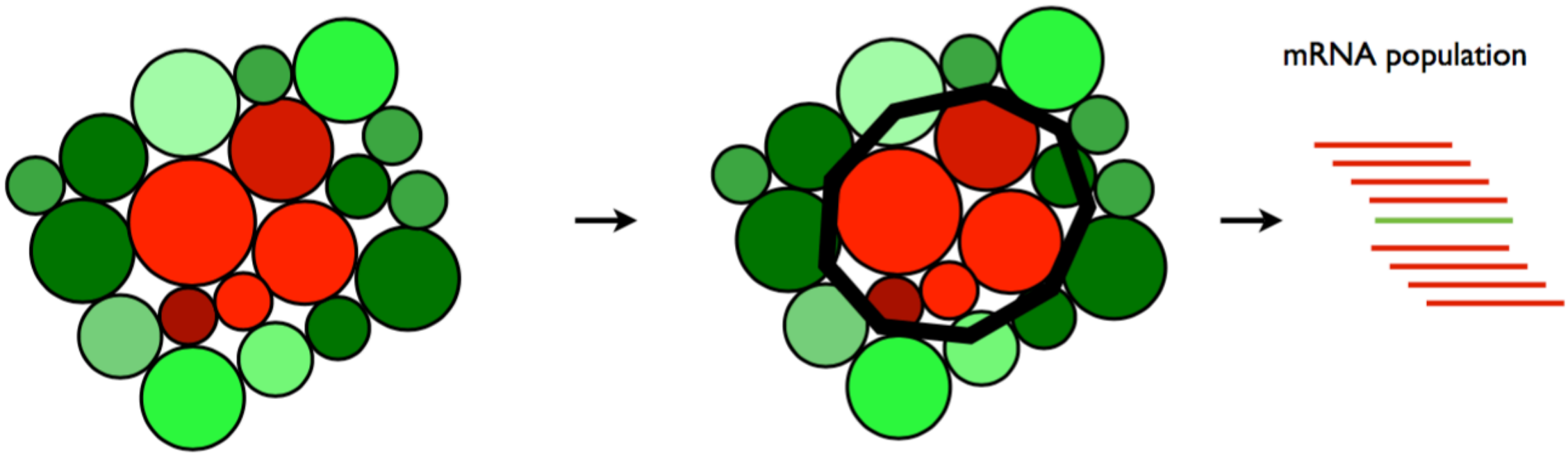


Is the study performed with different parts?

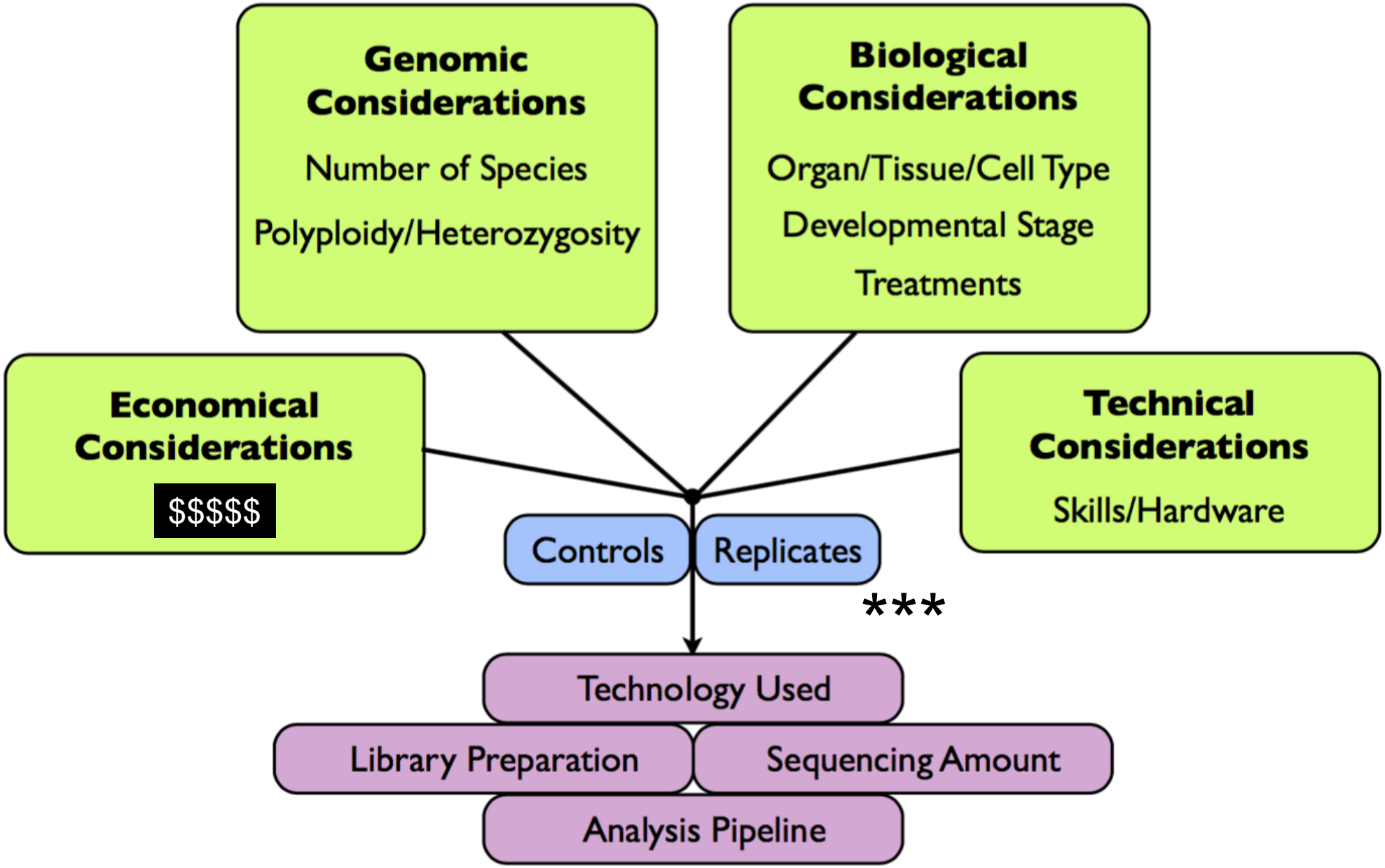
1) Organ specific

2) Tissue/Cell type specific

(Laser Capture Microdissection, LCM)



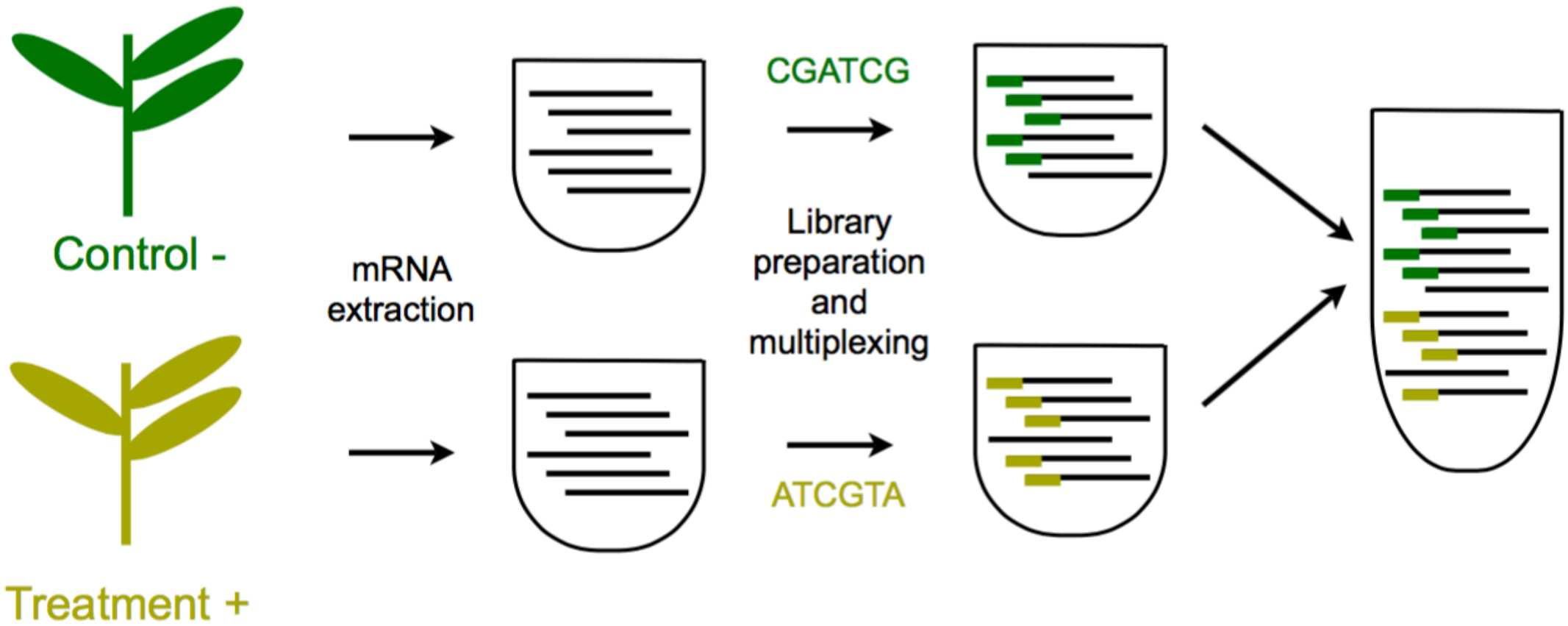
Experimental design



Prep and treatment

Sequencing of multiple samples can be performed using **multiplexing.**

The multiplexing add a tag/**barcode** of 4-6 nucleotides during the library preparation to identify the sample. Common kits can add up to 96 different tags.



Different organs, tissues or cell types can produce different mRNA extraction yields.

For samples where a low yield is expected a common practice is 1 to 3 rounds of **cDNA amplification**, specially using techniques such as LCM.

↓

Amplifications produce severe **bias** for between low/high represented transcripts

↓

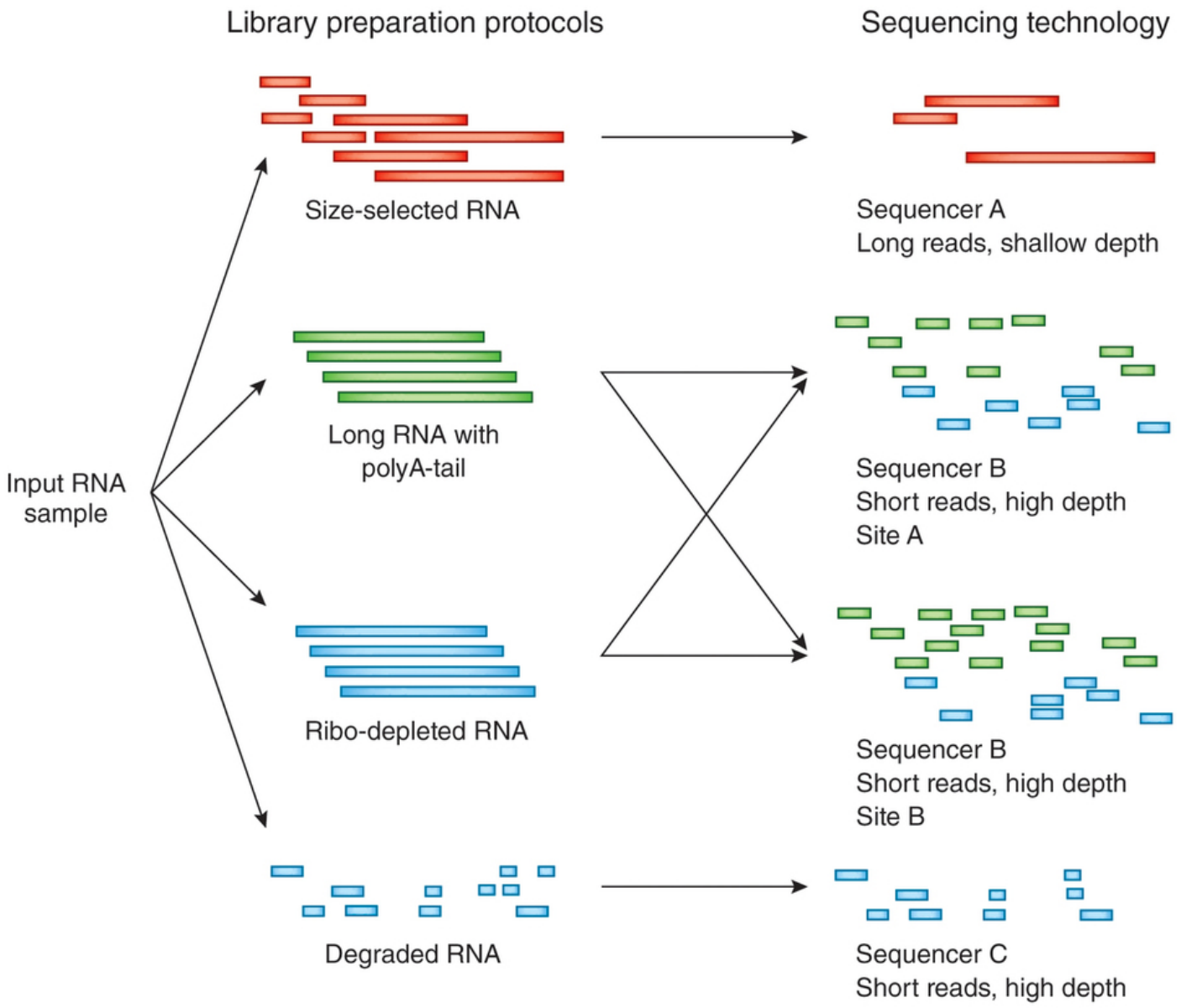
Best Practices:

1) Compare samples with same number of amplification rounds

2) Use software to measure and correct the bias

(example: *seqbias* from R/Bioconductor, Jones DC *et al.* 2012)

Combinations of input RNA, library preparation protocols and sequencing technology generate different read output



Kratz *et al.,* (2014)

Figure 1. Considerations in choosing RNAseq workflow.

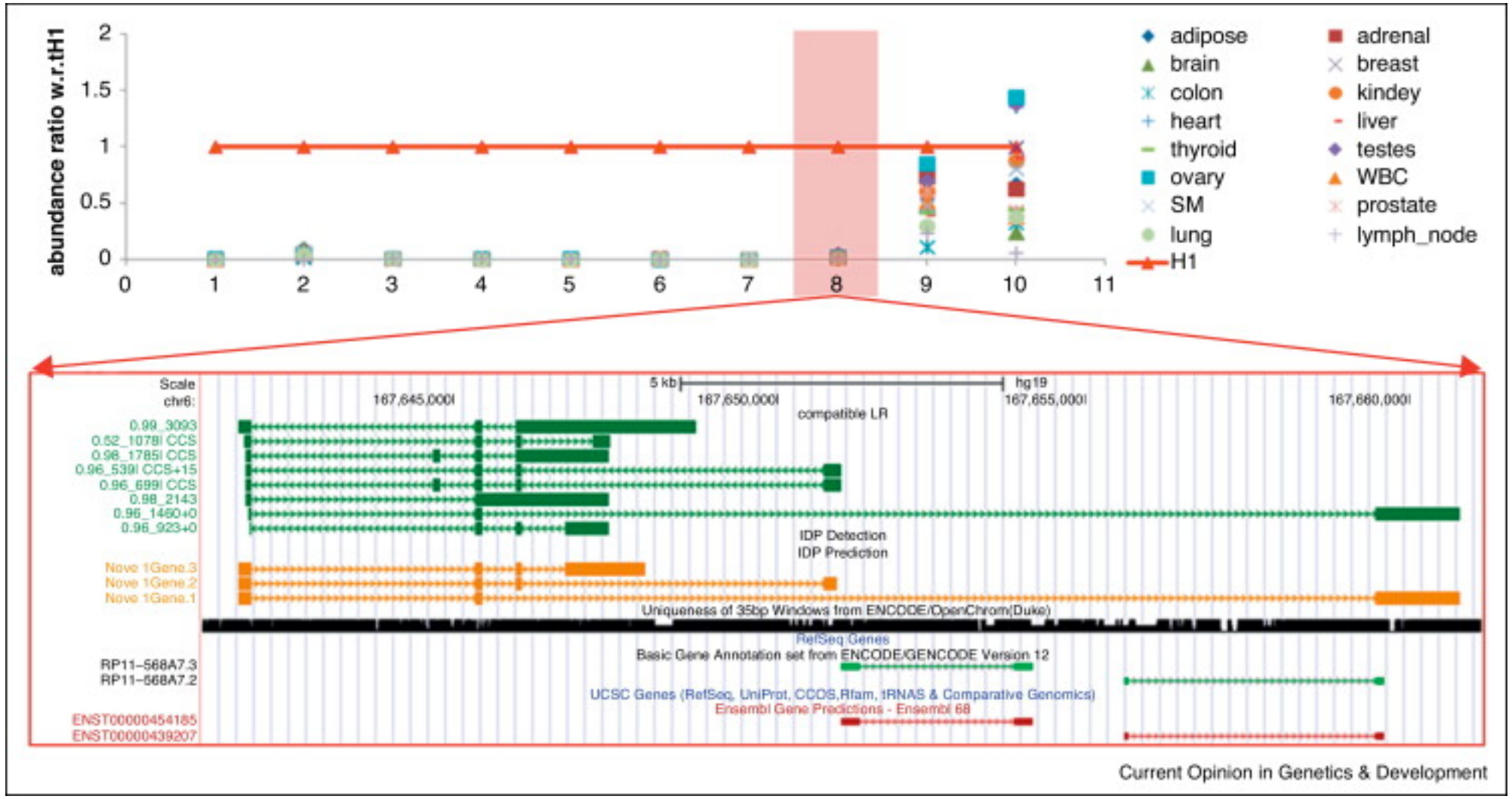
How many reads are enough? (for differential gene expression)

"Experiments whose purpose is to evaluate the similarity between the transcriptional profiles of two polyA+ samples may require only modest depths of sequencing (e.g. 30M pair-end reads of length > 30NT, of which 20-25M are mappable to the genome or known transcriptome."

The analysis from the current study demonstrated that 30 M (75 bp) reads is sufficient to detect all annotated genes in chicken lungs. Ten million (75 bp) reads could detect about 80% of annotated chicken genes. Furthermore, the depth of sequencing had a significant impact on measuring gene expression of low abundant genes.

http://www.rna-seqblog.com/information/how-many-reads-are-enough/

For isoform discovery, longer sequences are better



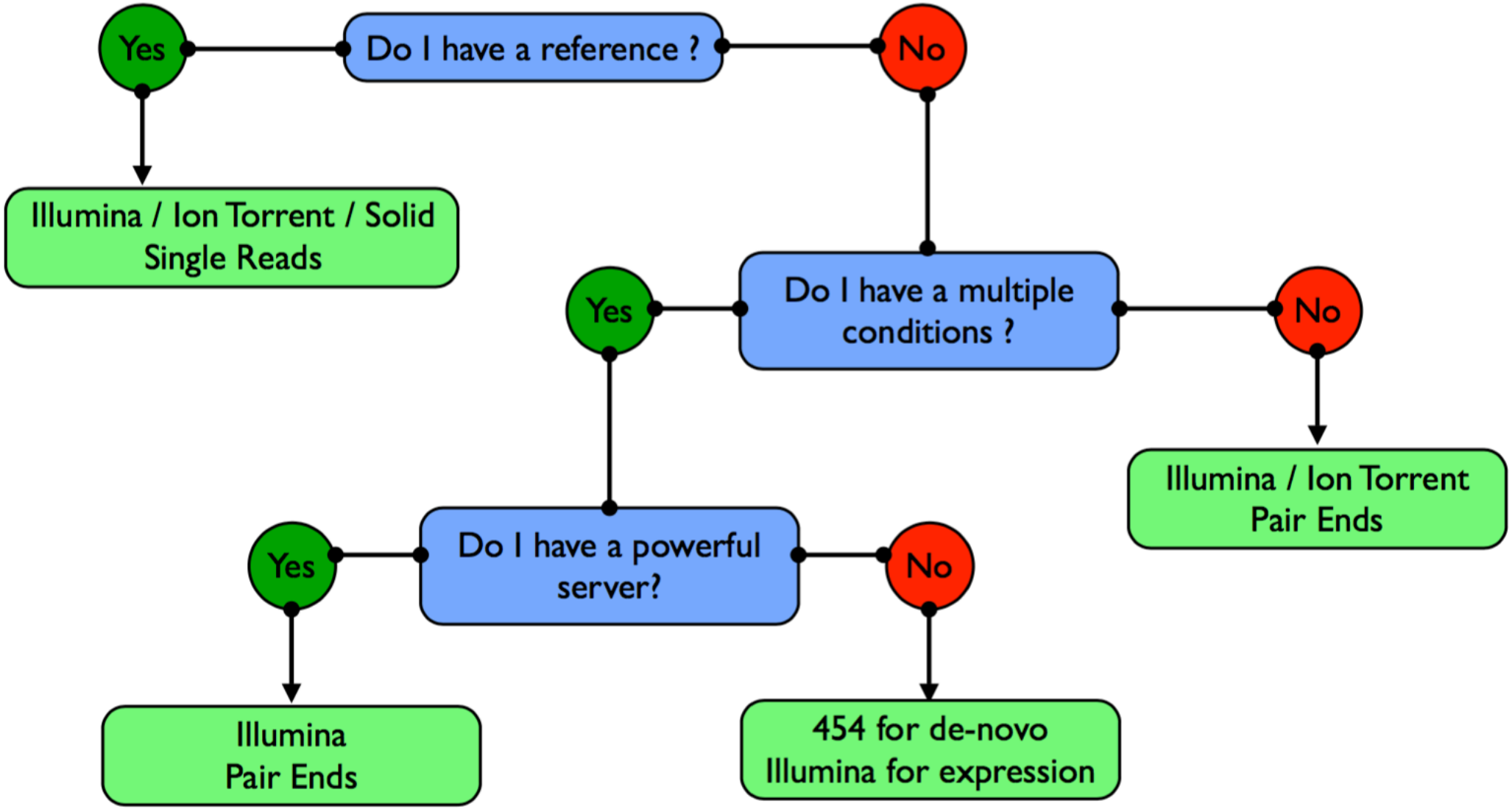
Au *et al.,* (2014)

Selecting the right technology

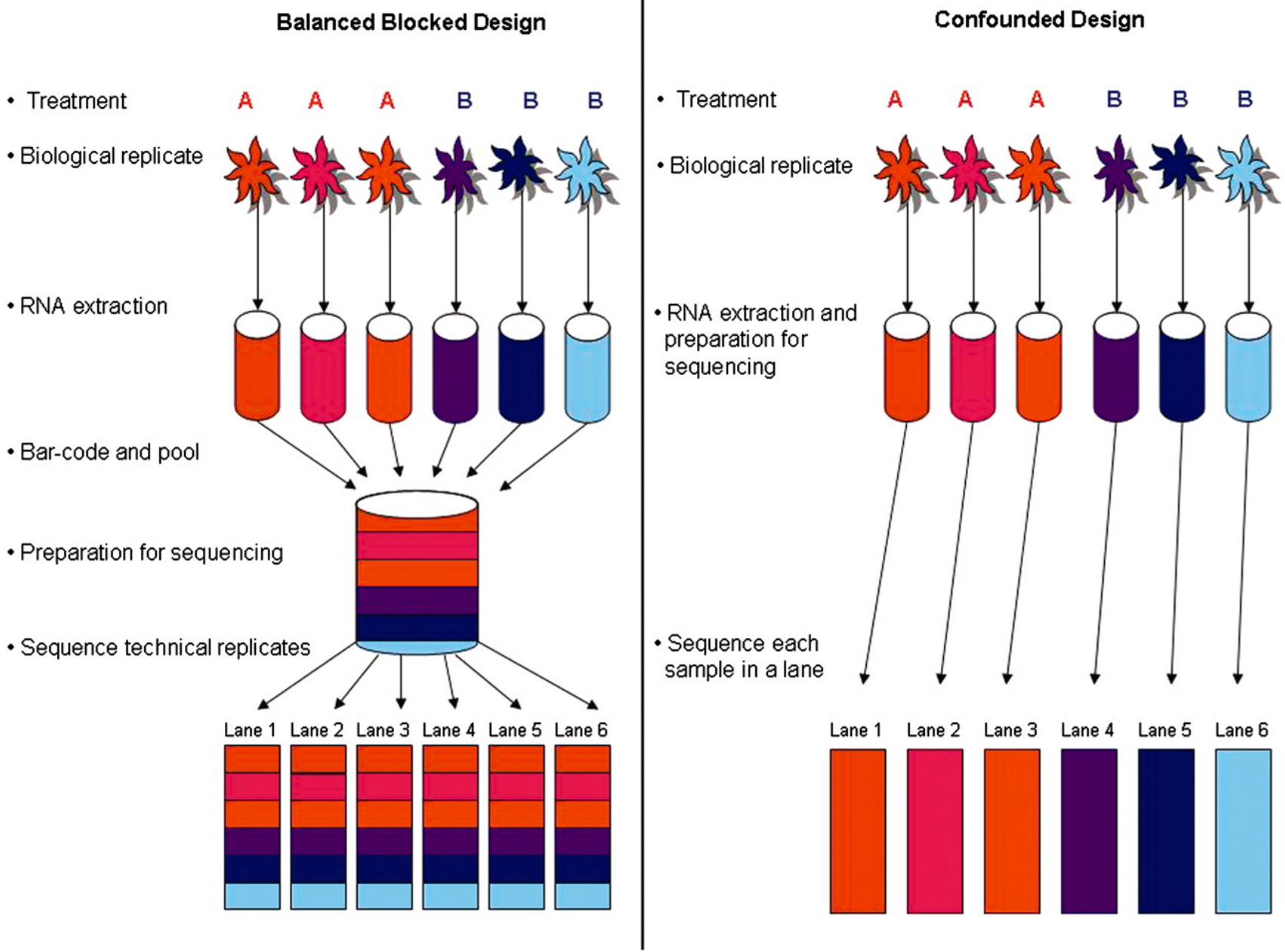
	Run Time	Sequence Length	Reads/Run	Total nucleotides sequenced per run
Capillary Sequencing (ABI37000)	~2.5 h	800 bp	386	0.308 Mb
454 Pyrosequencing (GS FLX Titanium XL+)	~23 h	700 bp	1,000,000	700 Mb (0.7 Gb)
Illumina (HiSeq 2500)	264 h / 27 h (11 days)	2 x 100 bp 2 x 150 bp	2 x 3,000,000,000 2 x 600,000,000	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)

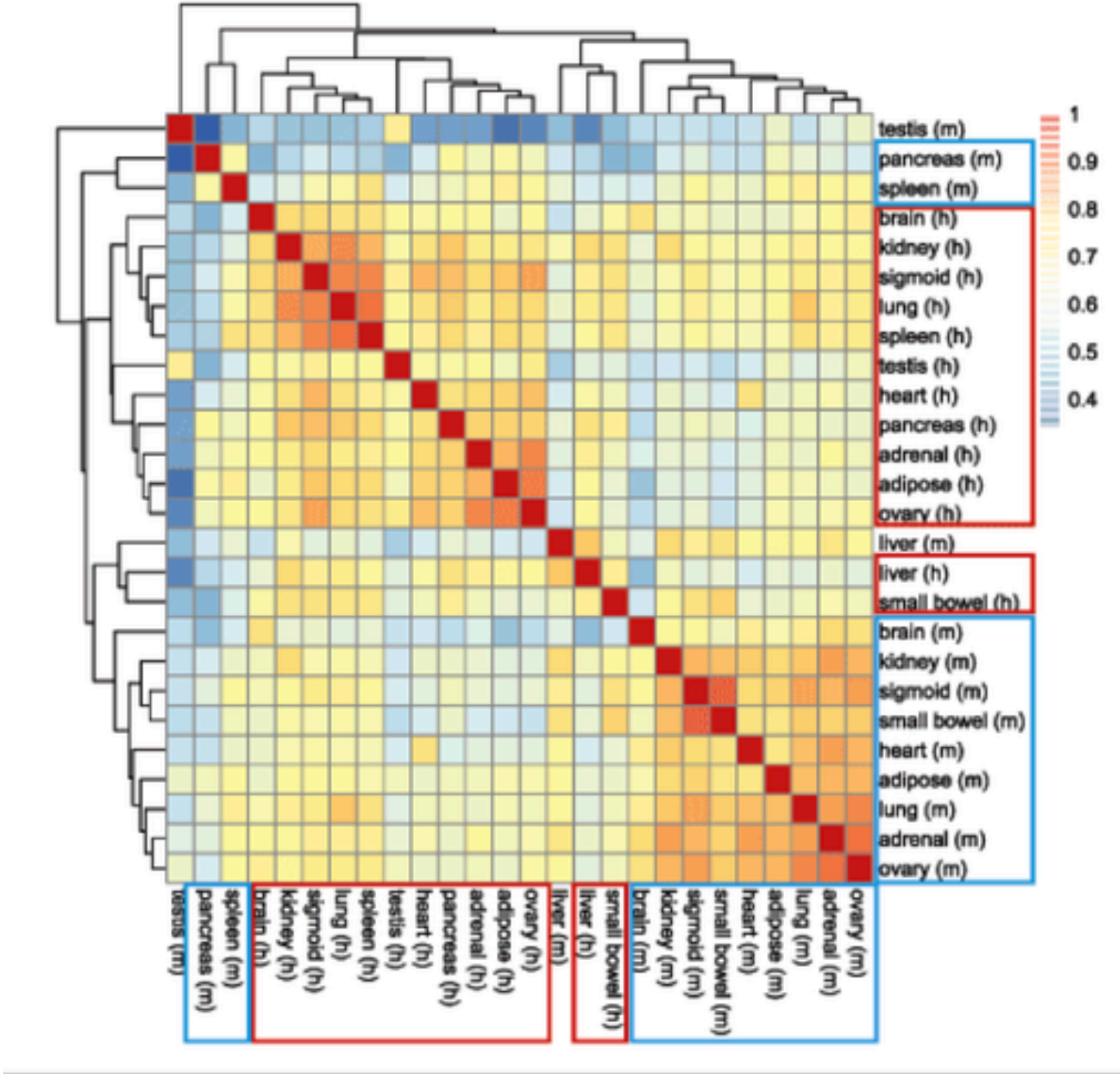


You need to design experiment carefully



Auer *et al.,* Genetics (2010)

Example of batch effect:



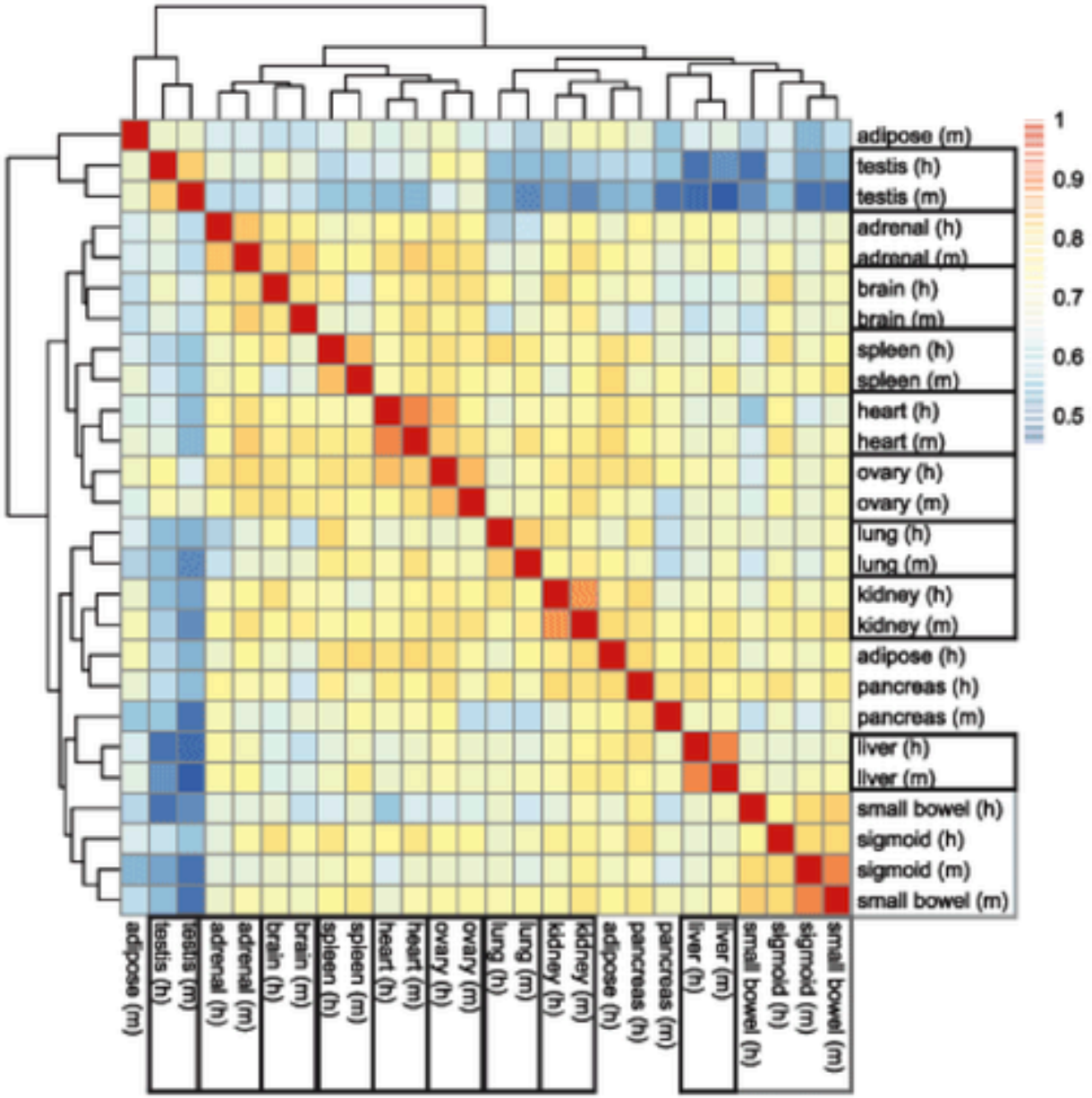
Lin *et al.,* (PNAS) 2014

Example of batch effect:

Yoav Gilad
@Y_Gilad

Following

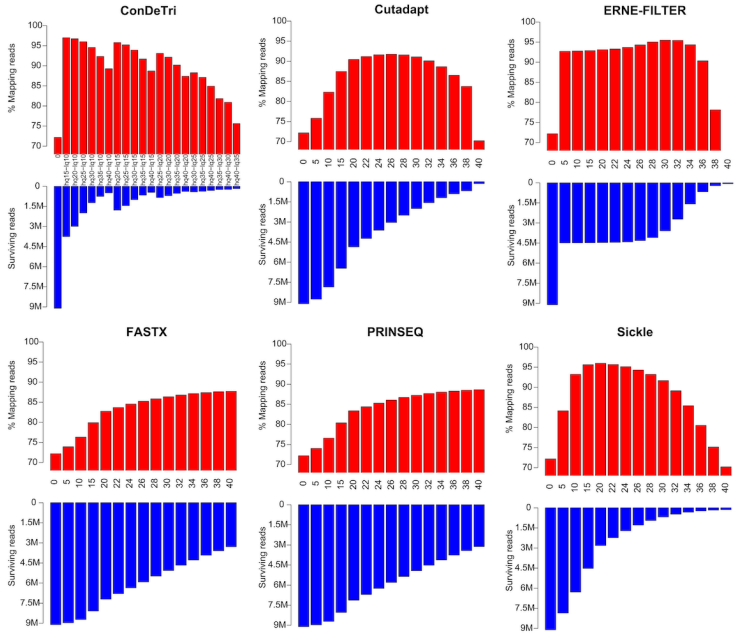
We reanalyzed the data from pnas.org/content/111/48… and found the following:



http://f1000research.com/articles/4-121/v1

Is trimming beneficial?

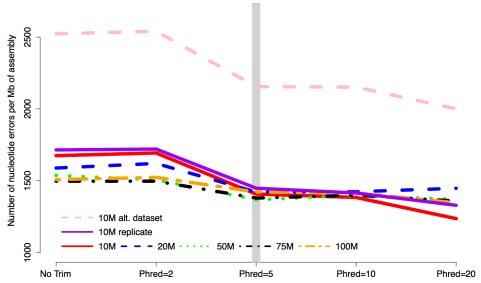
Software comparison, RNA/DNA-Seq



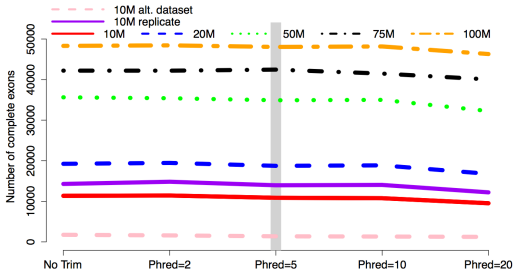
"trimming is beneficial in RNA-Seq, SNP identification and genome assembly procedures, with the best effects evident for intermediate quality thresholds (Q between 20 and 30)"

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

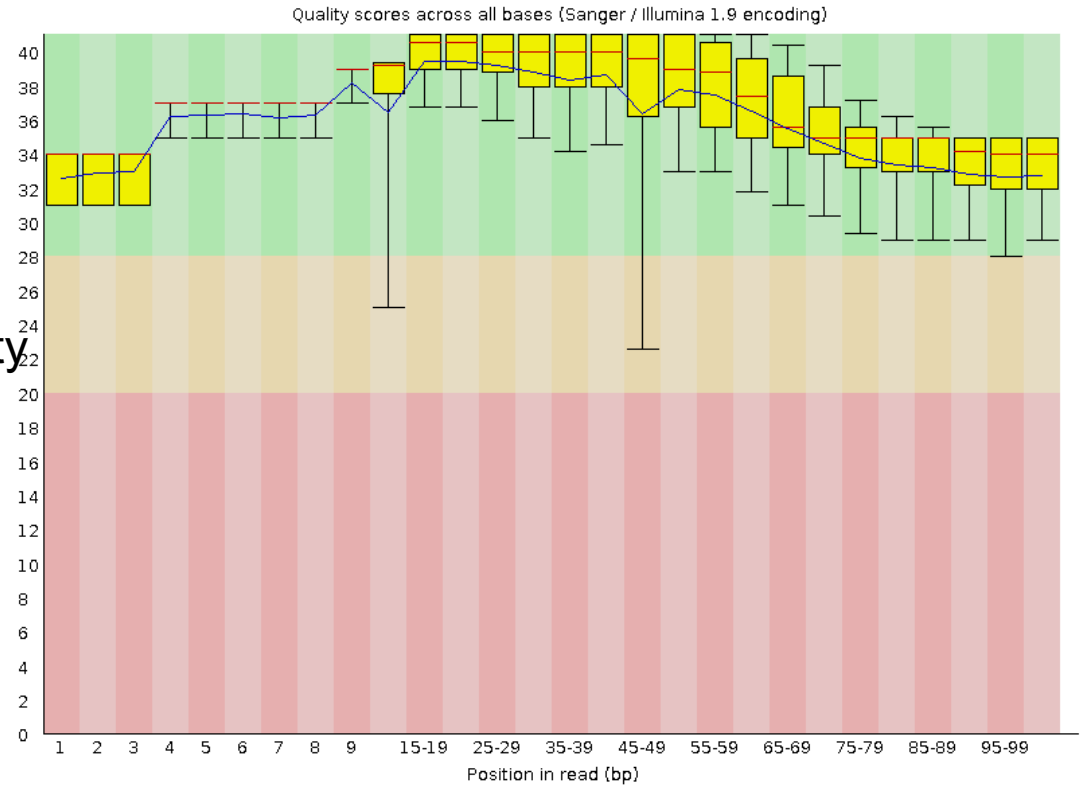
"Although very aggressive quality trimming is common, this study suggests that a more gentle trimming, specifically of those nucleotides whose Phred score < 2 or < 5, is optimal for most studies across a wide variety of metrics."

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

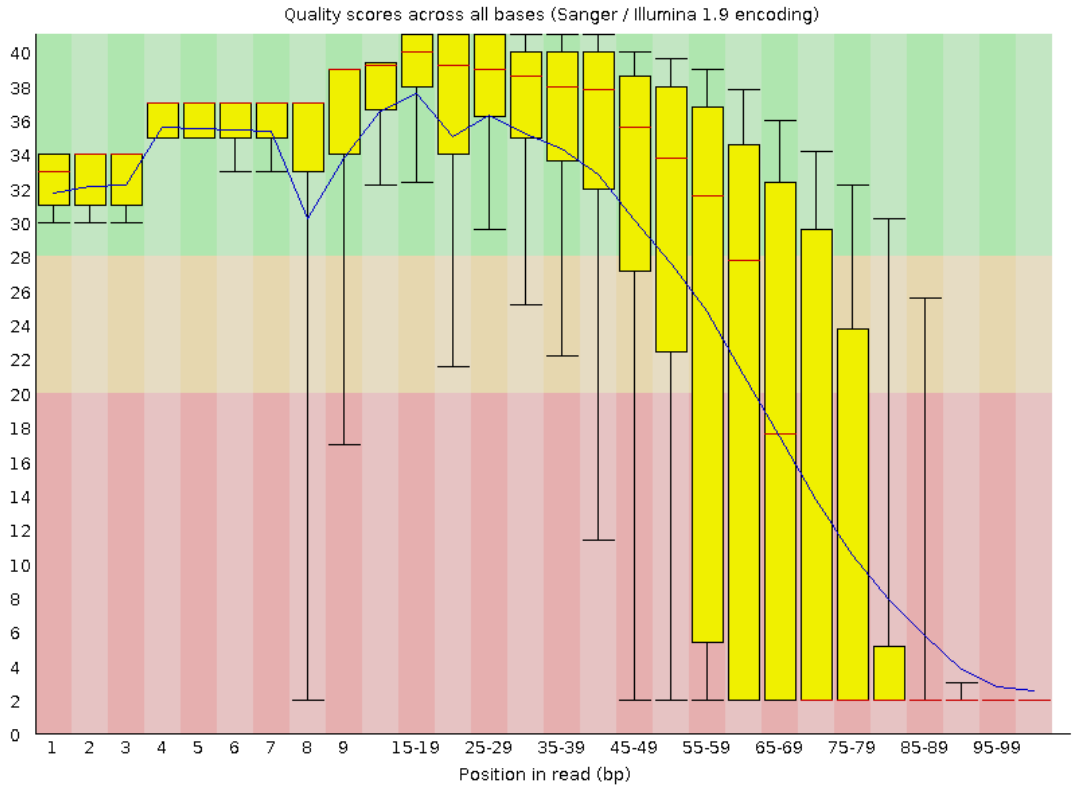
Good/Trimmed data



Quality

Base pair position of the read

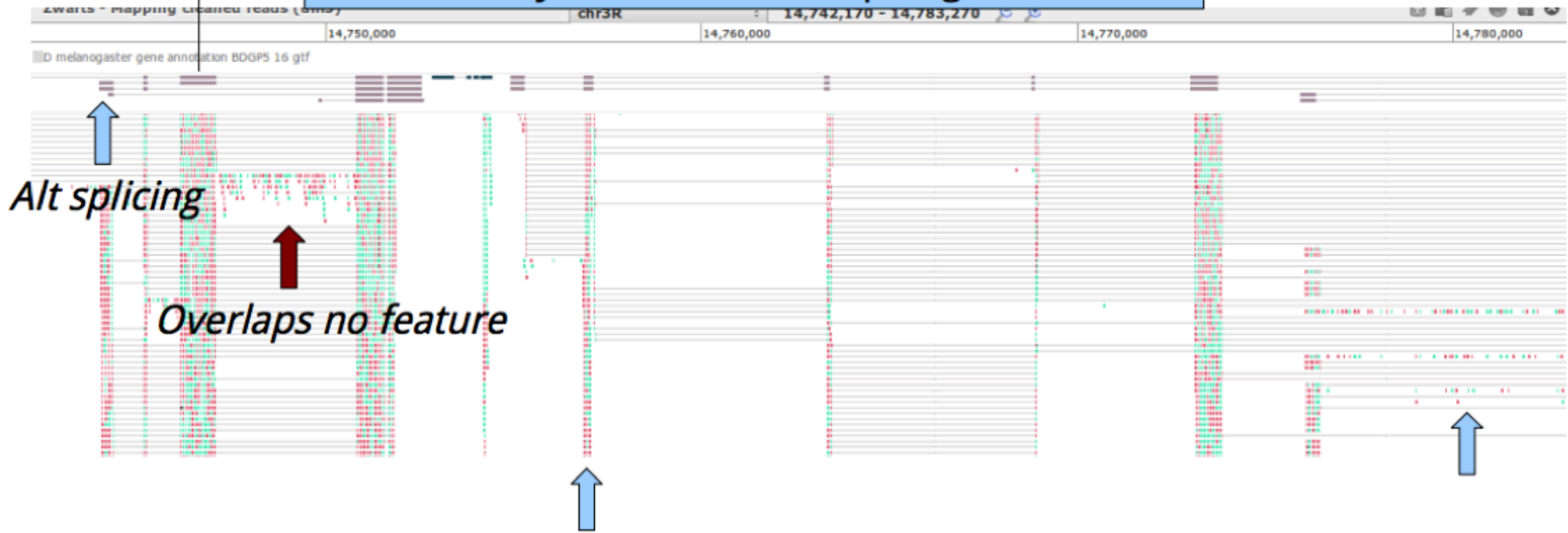
Poor/Raw data



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

Once you have mappings, you can start counting

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

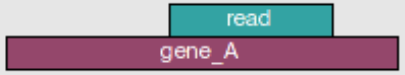
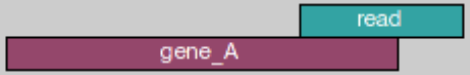
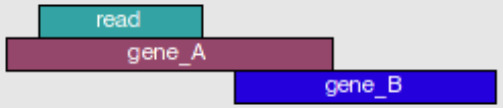
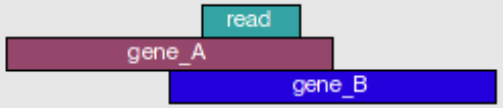


Concept:

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

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

HTseq (the most popular program) -> Replacing featurecount

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Union mode is recommended in most cases

http://www-huber.embl.de/HTSeq/doc/count.html#count

Featurecount (much faster!)

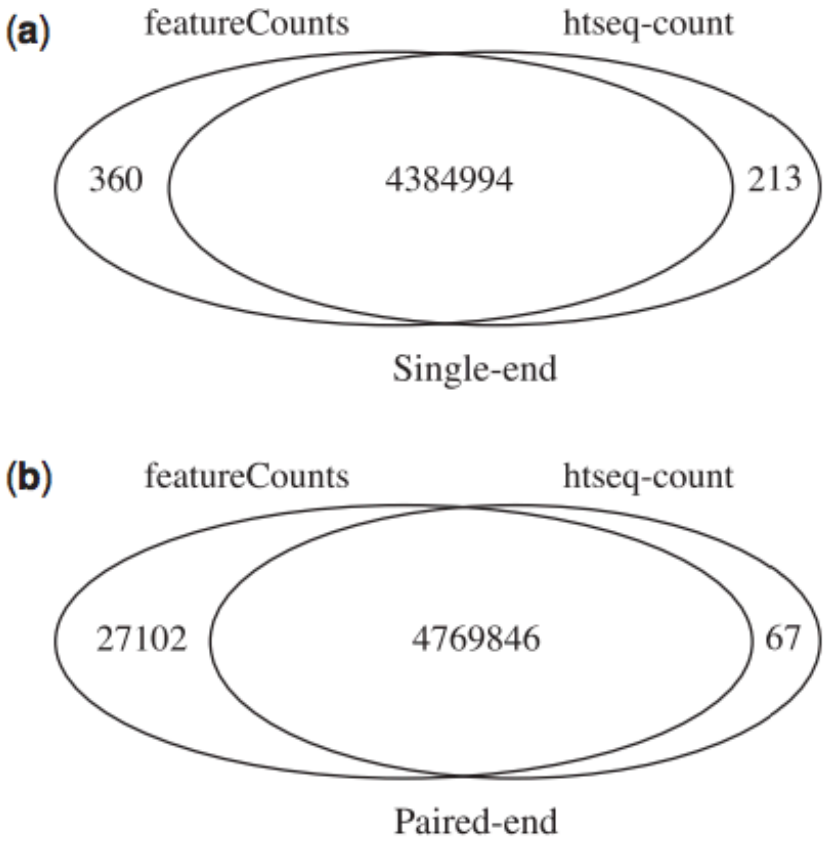


Table 3. Performance with RNA-seq reads simulated from an annotated assembly of the Budgerigar genome

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

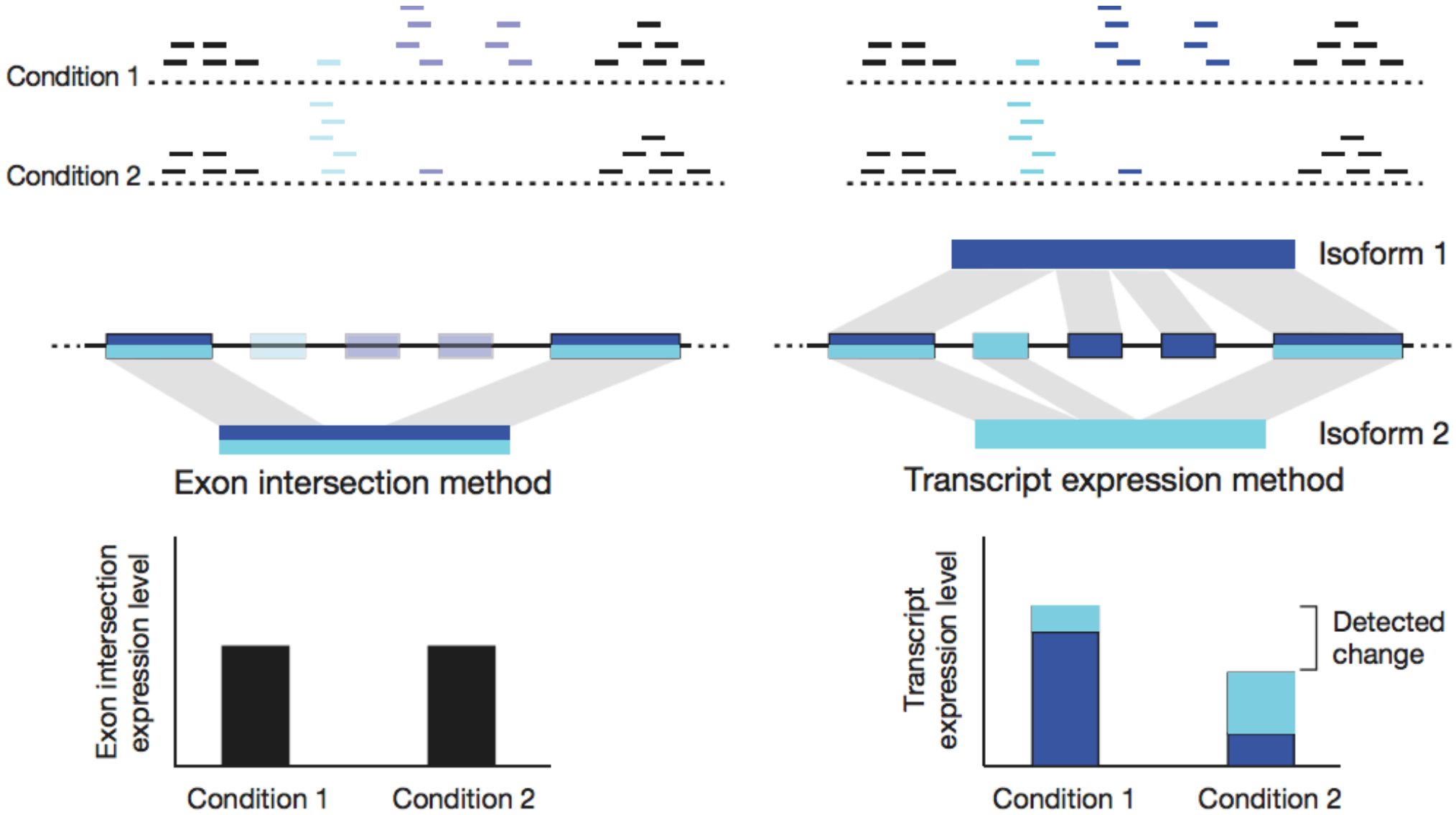
- Which genes are highly counted?
- Any samples with a lot of missing/no count genes?
- Anything that may affect sample counts (like batch effect?)

Ambiguity in counting

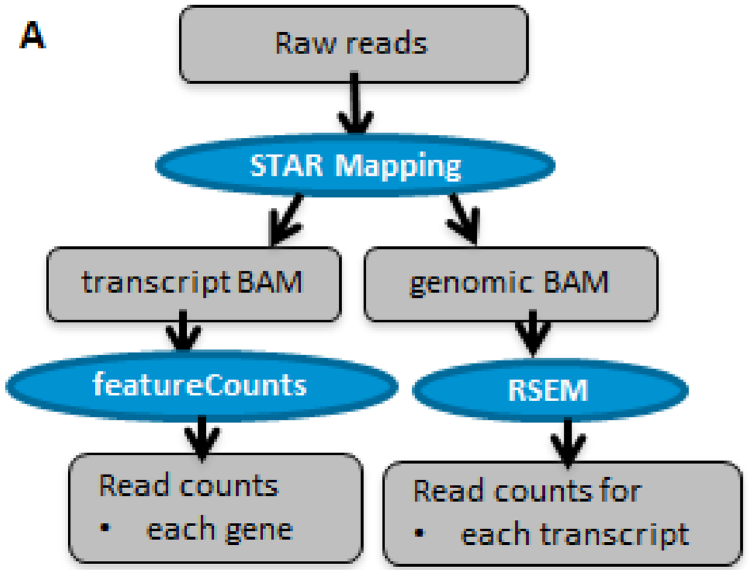
We focus on the **gene level**: merge all counts over different **isoforms** into one, taking into account:

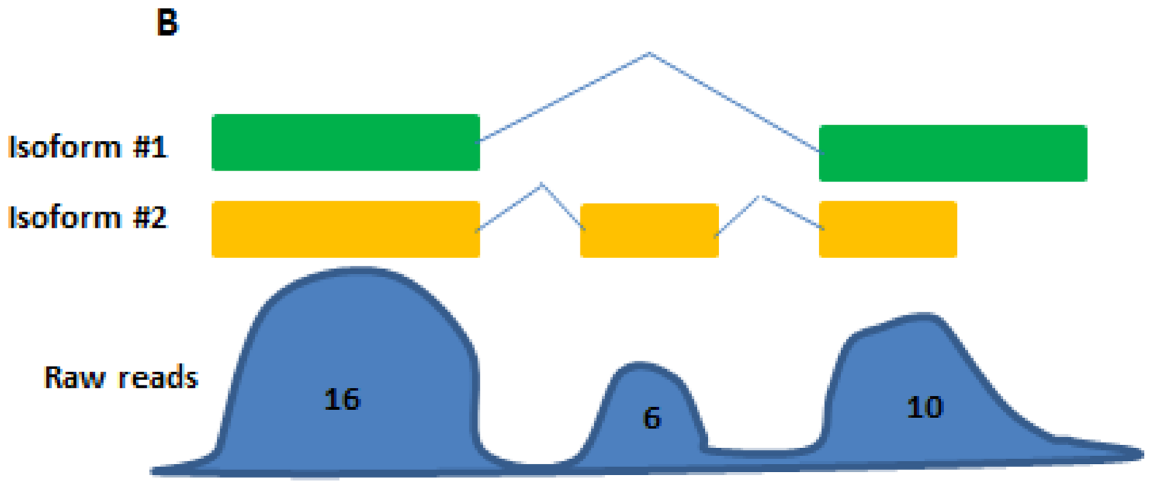
- Reads that do **not overlap** a feature, but appear in introns. Take into account?
- Reads that align to **more than one feature** (exon or transcript)**.** Transcripts can be overlapping - perhaps on different strands. (PE, and strandedness can resolve this partially).
- Reads that **partially** overlap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

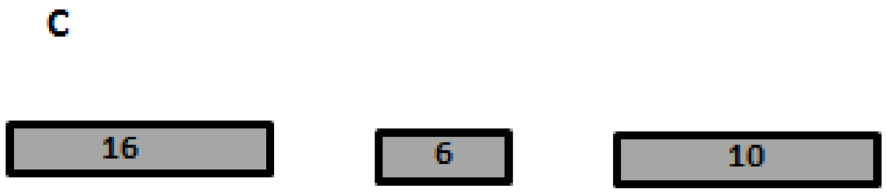
Transcript counting could be more robust in detecting changes



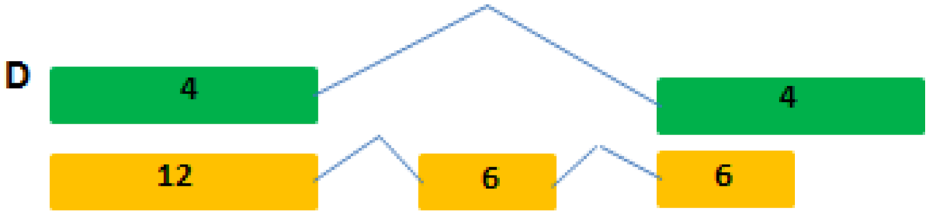
Outstanding problems in counting with exon merging model







Total gene length after exon flattening: **5kb**
Total reads: **32**
RPKM for gene: **6.4 (=32/5)**



Relative isoform abundance (#1/#2): **25% / 75%**
RPKM for isoform #1 and #2: **2 and 6**
RPKM for gene (=sum of isoforms): **8 (=2+6)**

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

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

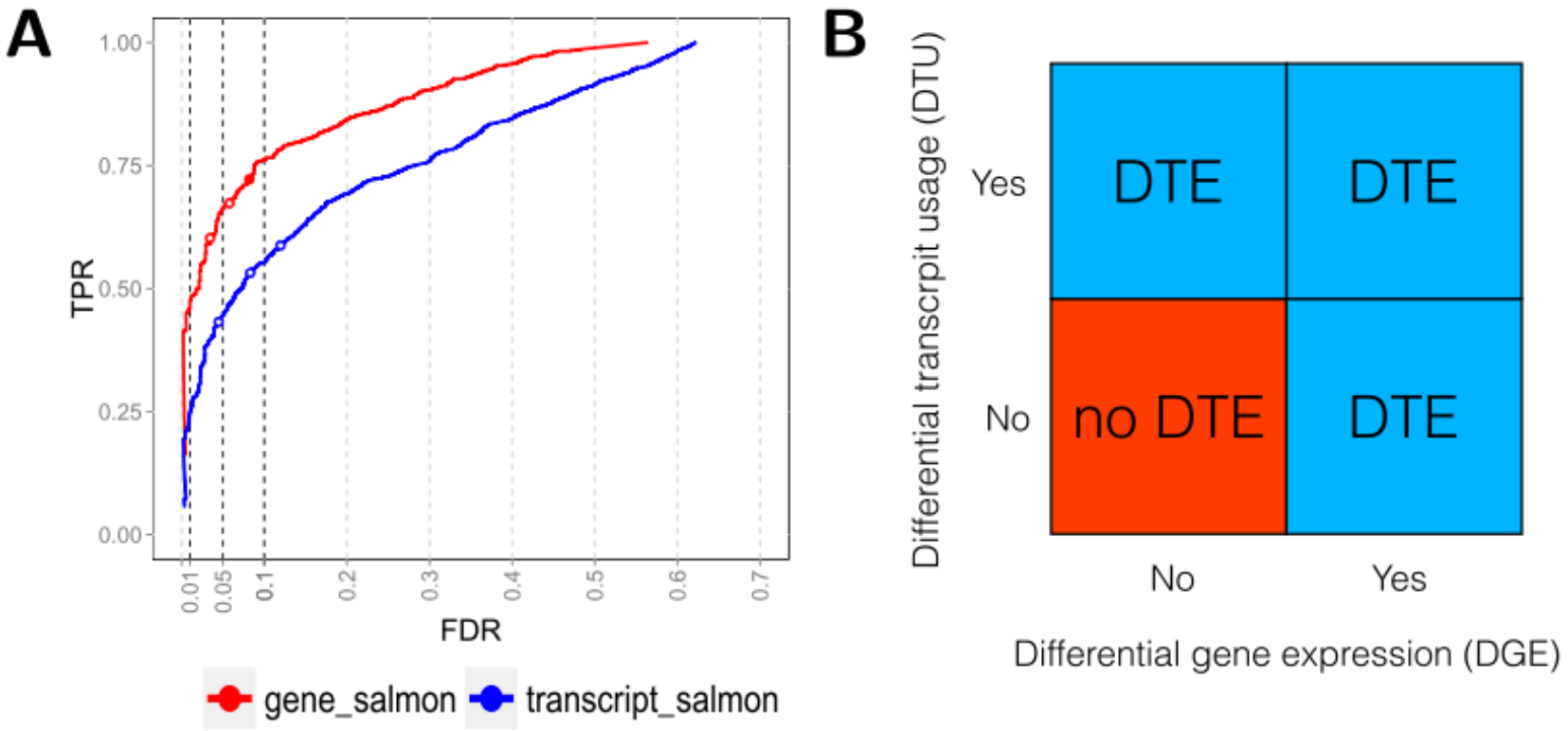
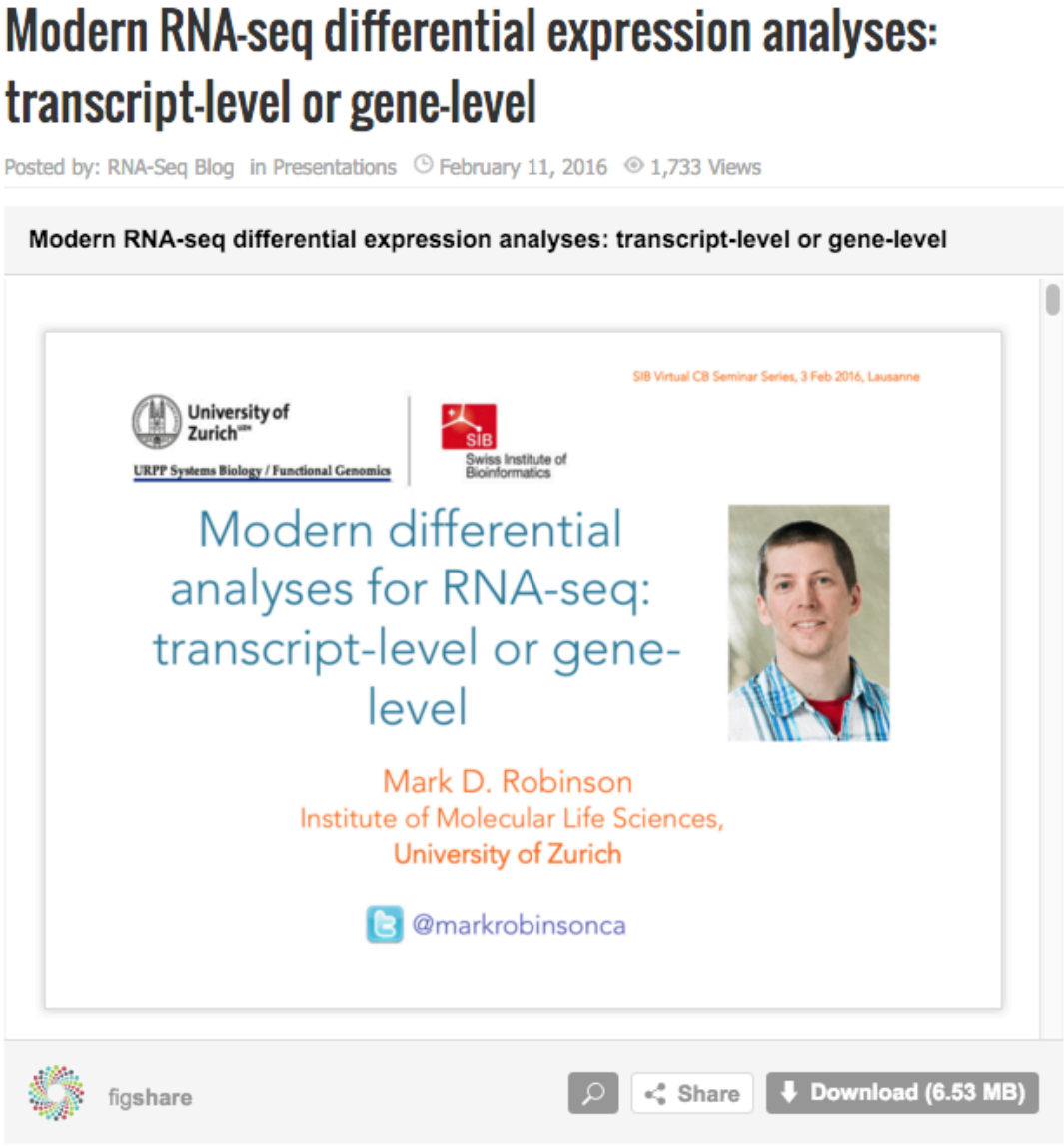


Figure 2 (sim2). A: DTE detection performance on transcript- and gene-level, using *edgeR* applied to transcript-level estimated counts from *Salmon*. The statistical analysis was performed on transcript level and aggregated for each gene using the *perGeneQValue* function from the *DEXSeq* R package;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



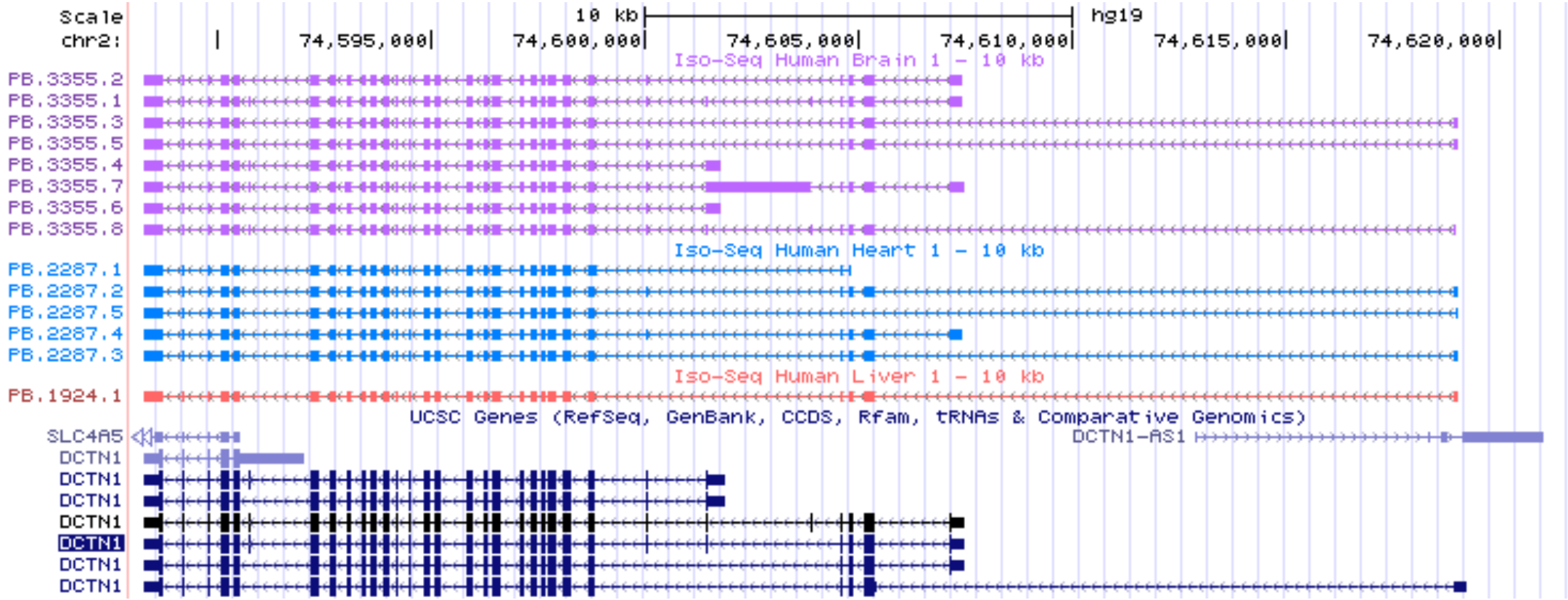
"There is no crisis; the impact of union vs. transcript counting in many datasets is rather small"

"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

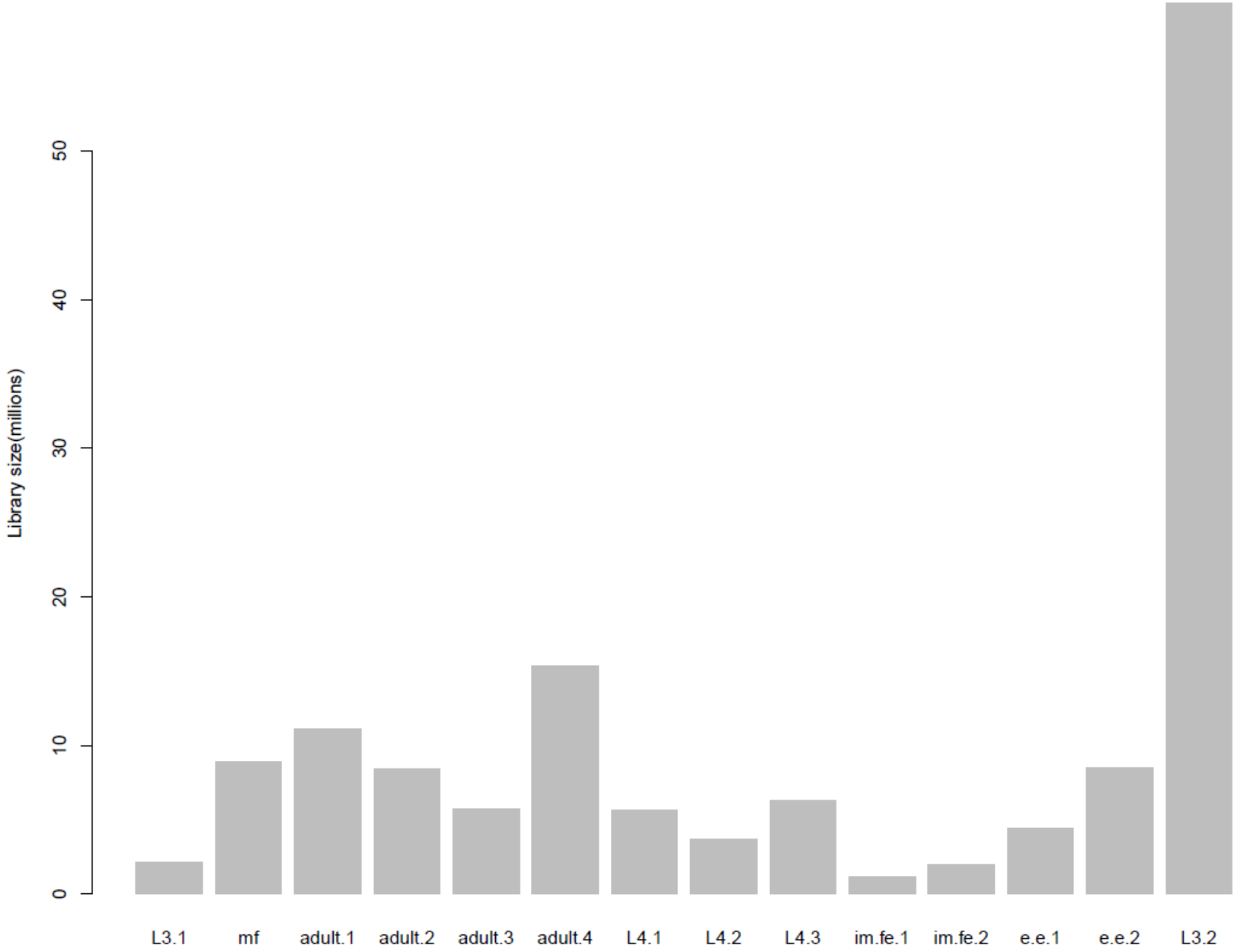
Pacbio IsoSeq



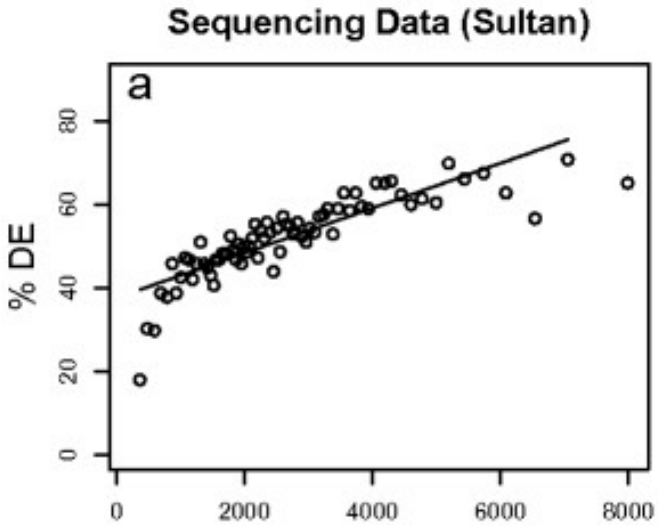
This is the bit we care about!

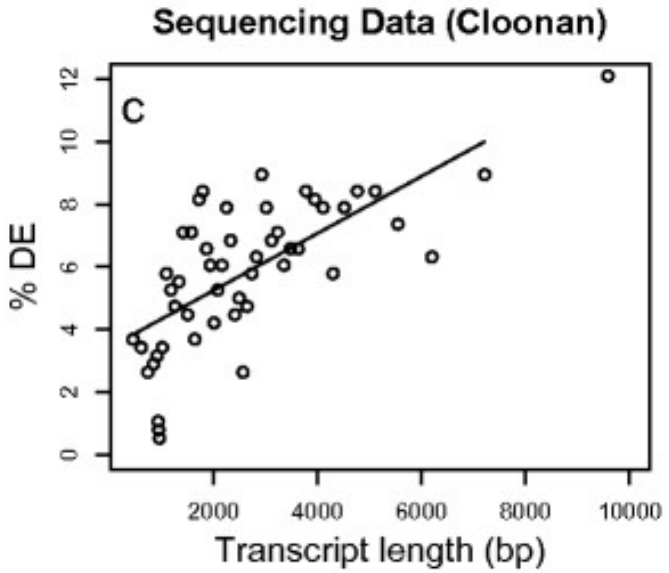
Counts of the gene depends on **expression** ,transcript length ,sequencing depth and simply chance

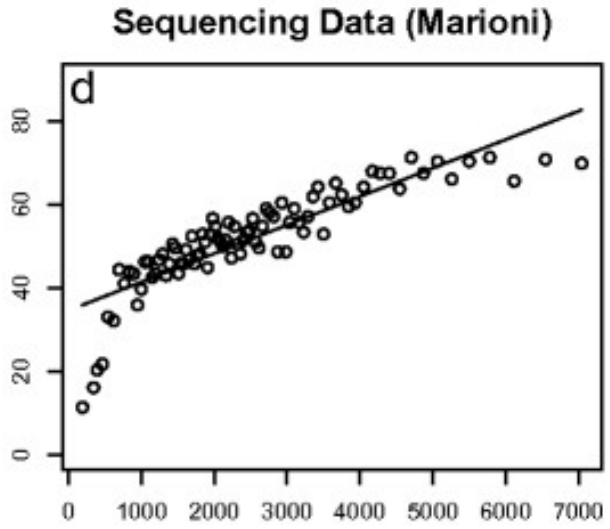
Higher sequencing depth equals more counts

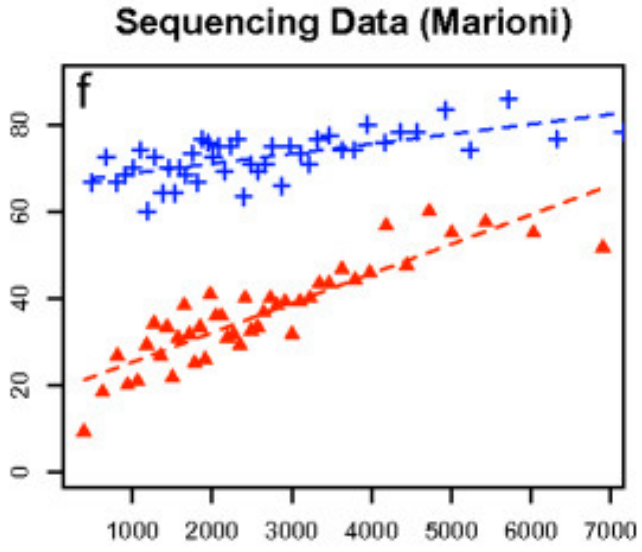


Counts are proportional to the transcript length x mRNA expression level









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

Oshlack and Wakefield (2009) Biology Direct.

Normalization: different goals

- **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

Optimal Scaling of Digital Transcriptomes

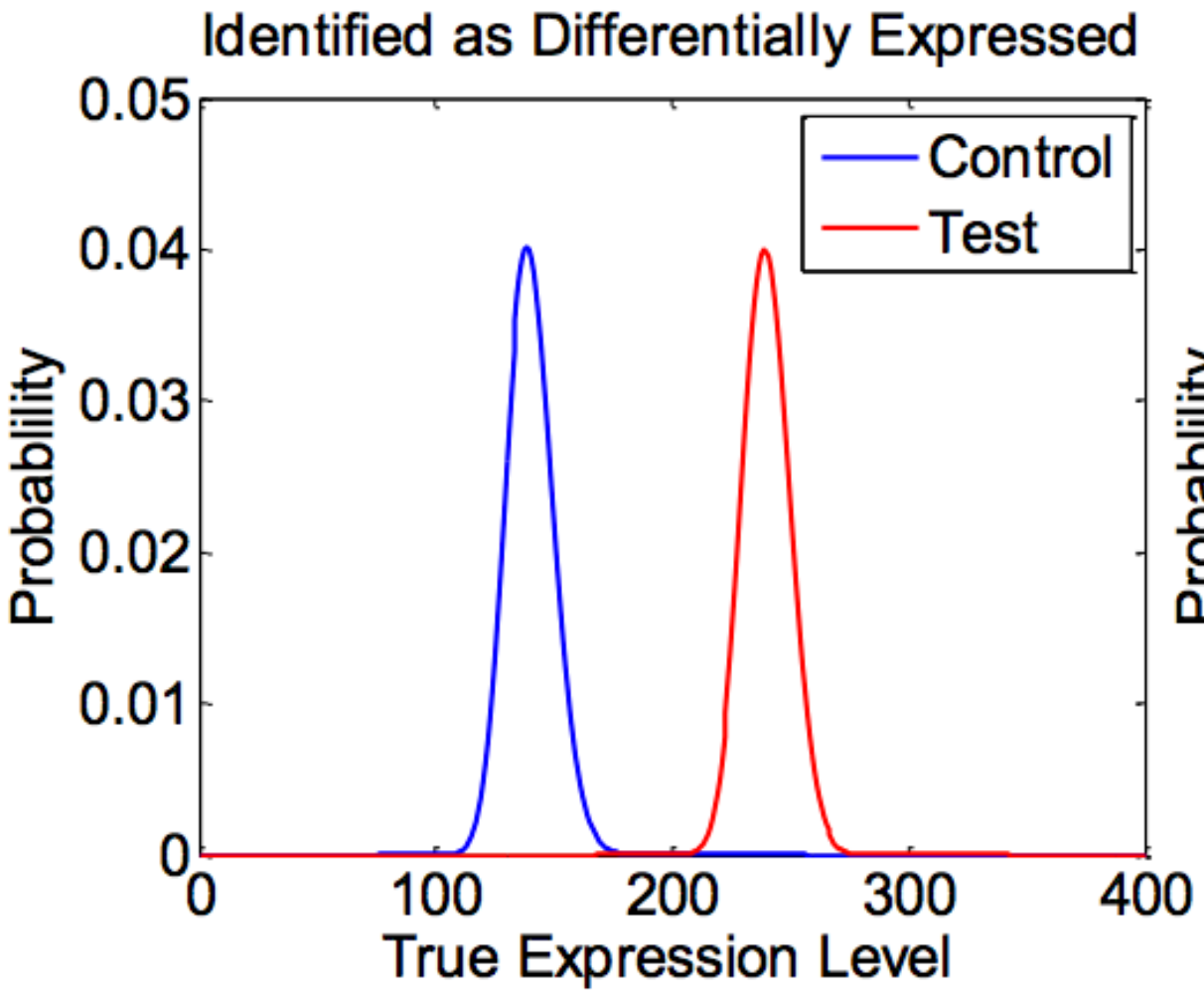
Gustavo Glusman, Juan Caballero, Max Robinson, Burak Kutlu, Leroy Hood

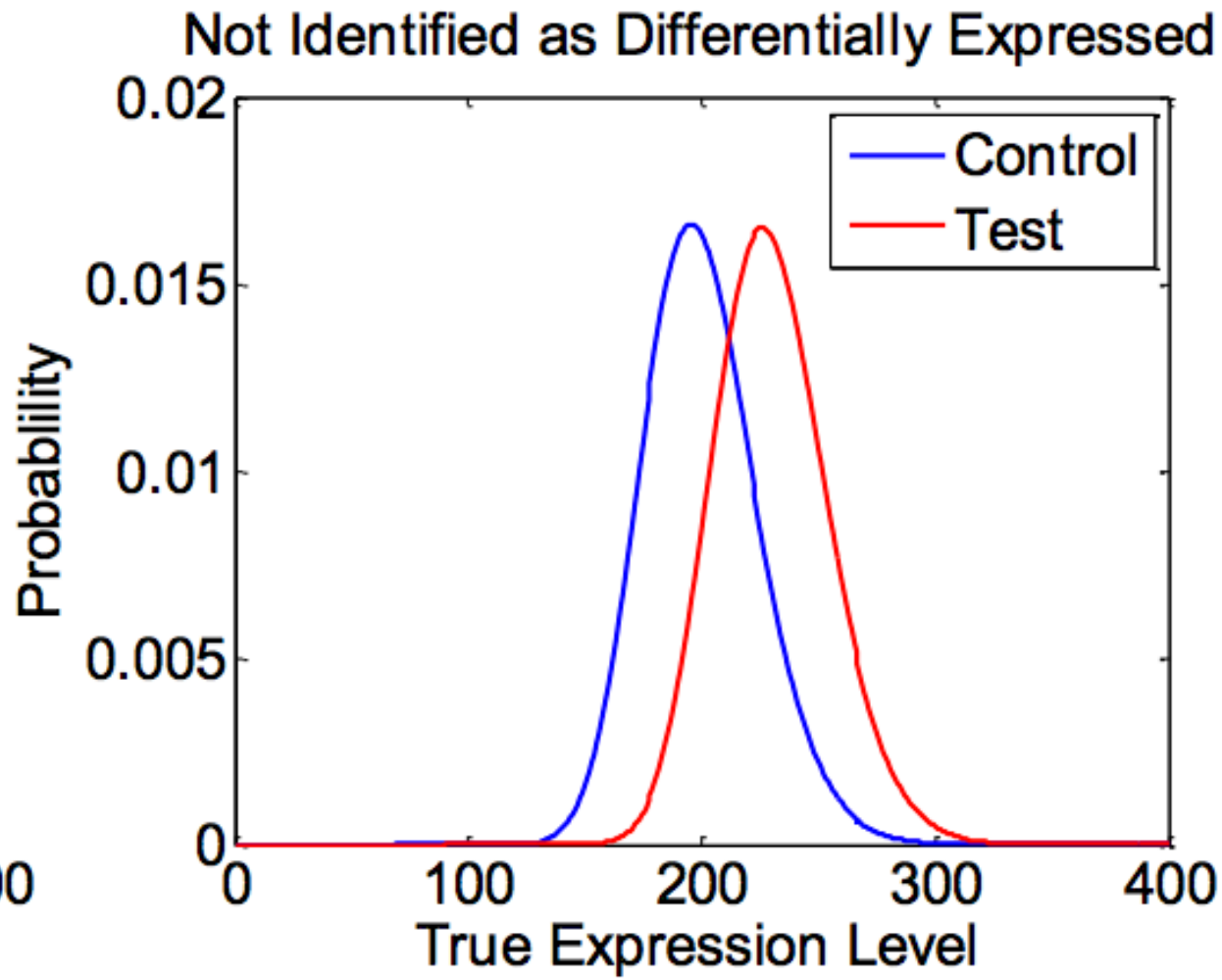
Published: Nov 06, 2013 • DOI: 10.1371/journal.pone.0077885

But how do you know your count = 2 is really 2?

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

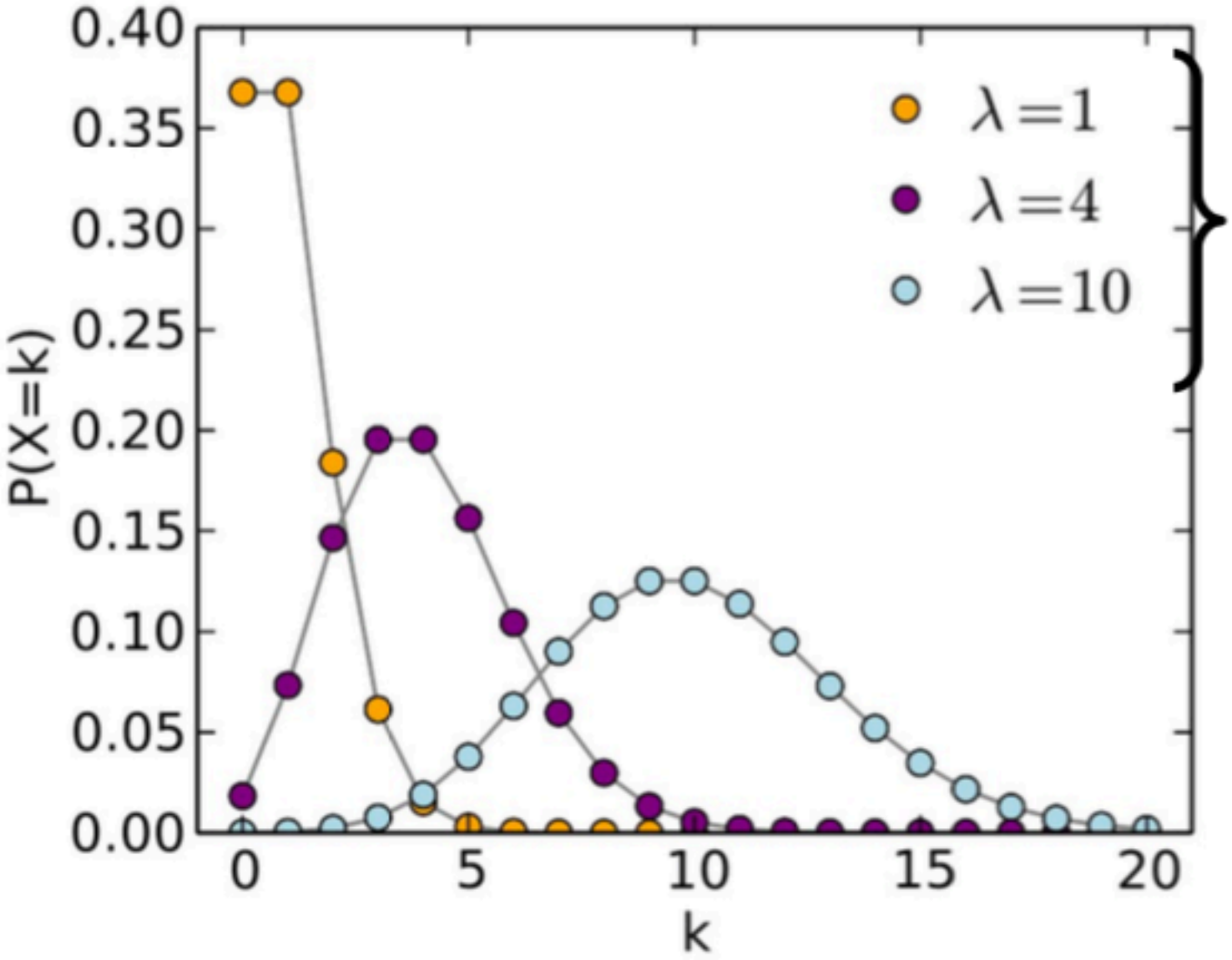
Fitting a distribution for every gene for **DE**





. Busby *et al.,* Bioinformatics (2013)

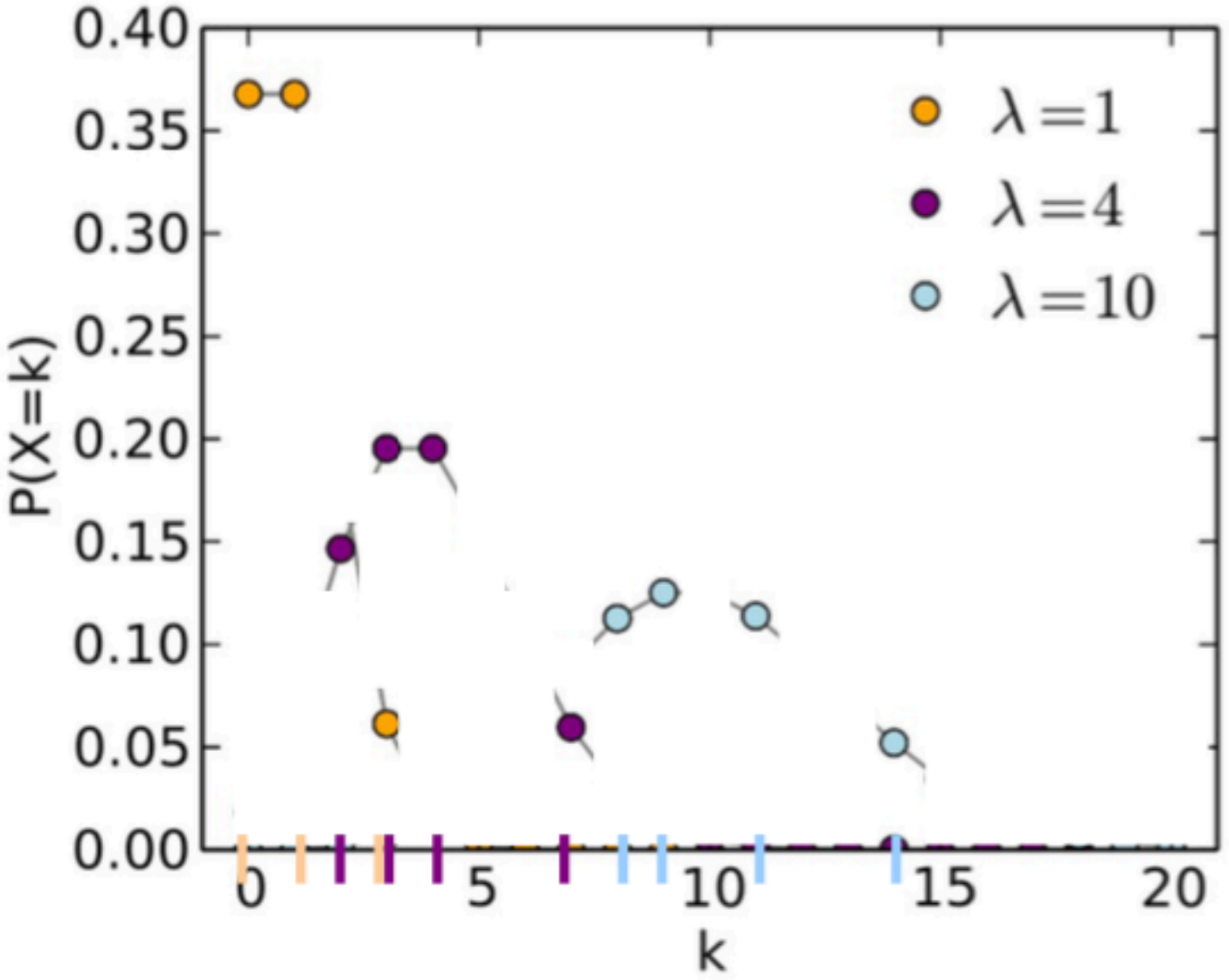
The counts of technical replicates follow a **poisson** distribution (Marioni *et al.,* 2008). So mean = count, variance = count



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

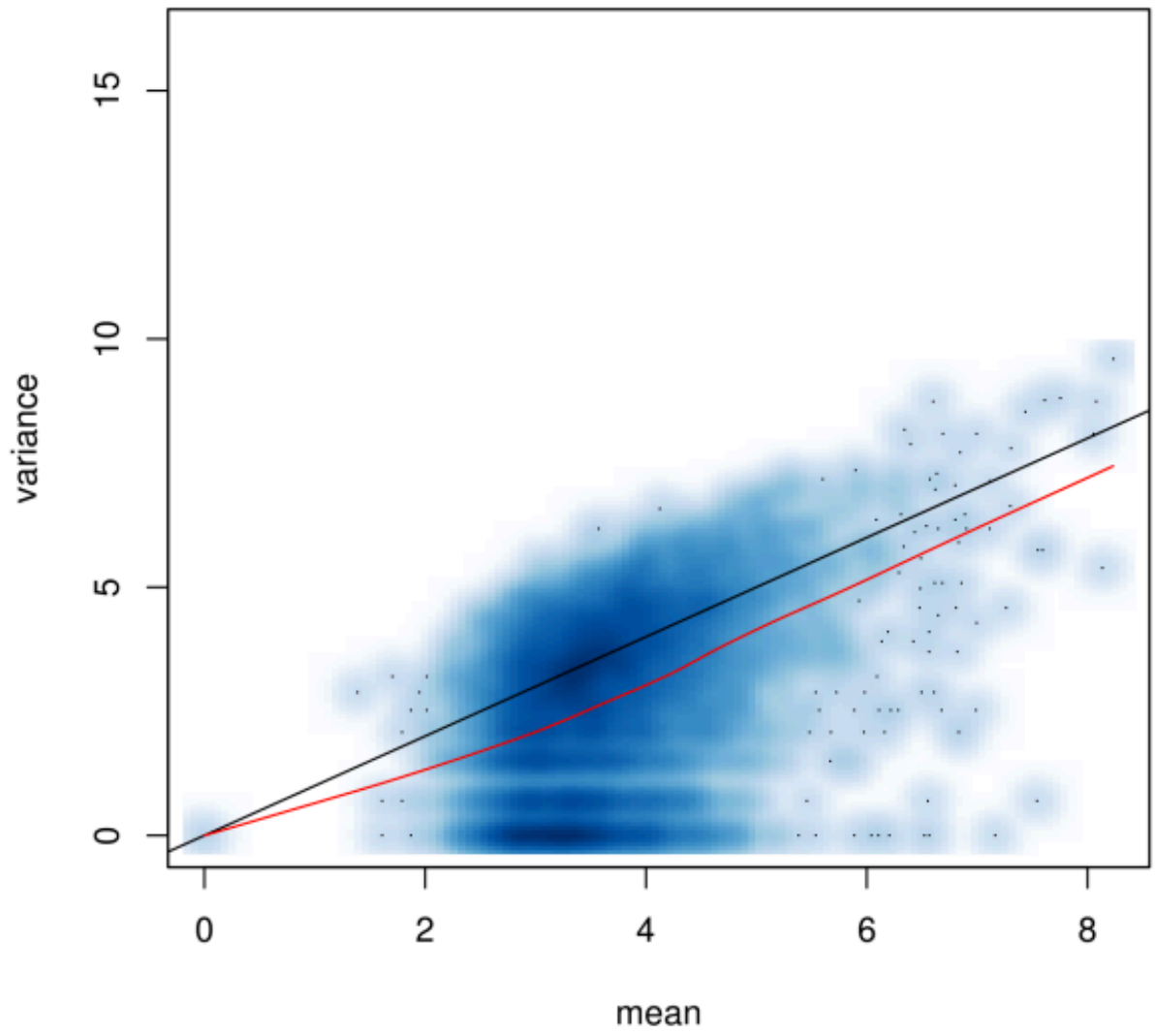


GeneA **0, 0, 1, 3**

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

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

Poisson model seems good fit in technical replicates

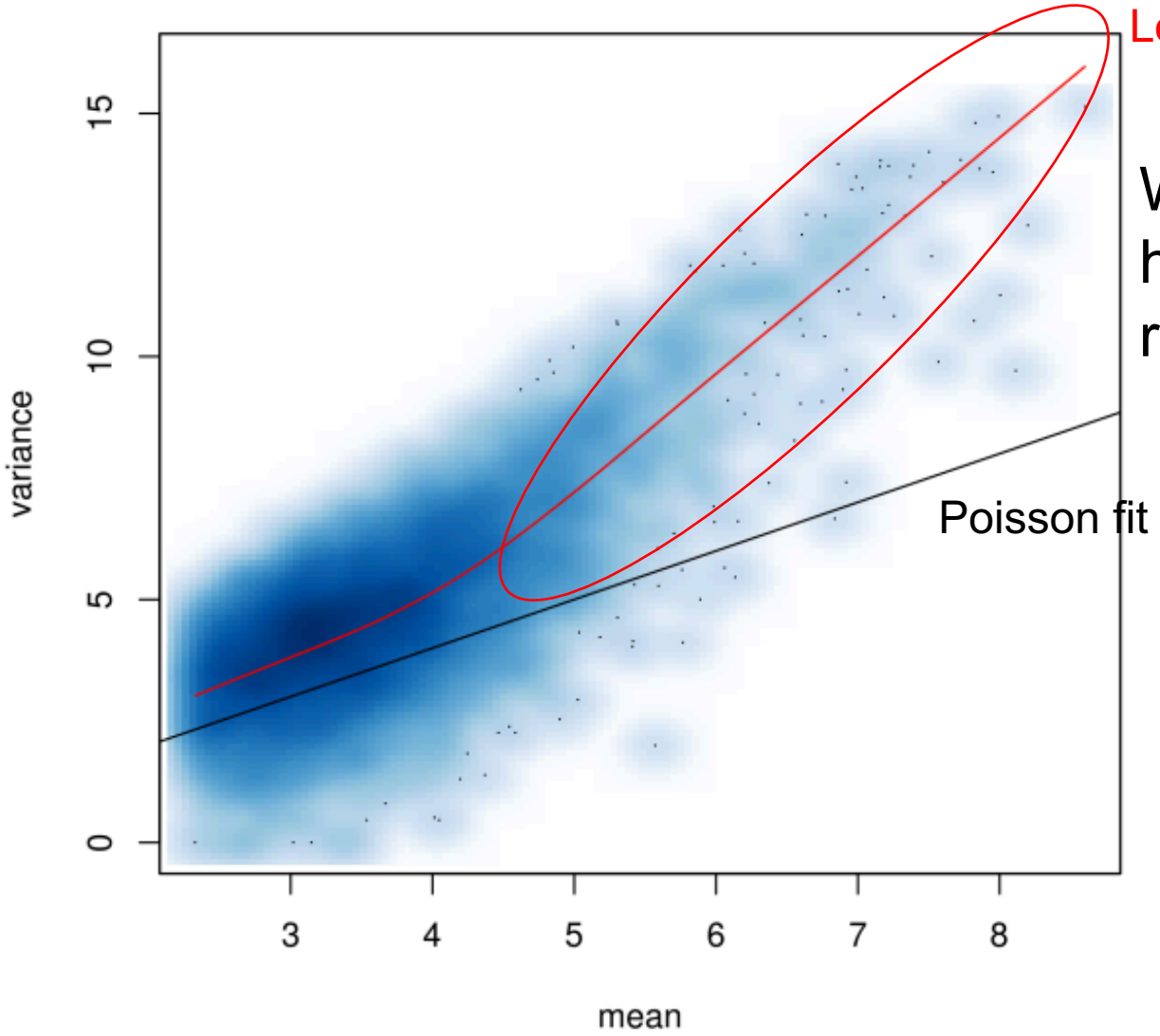


Poisson fit

Lowess fit

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

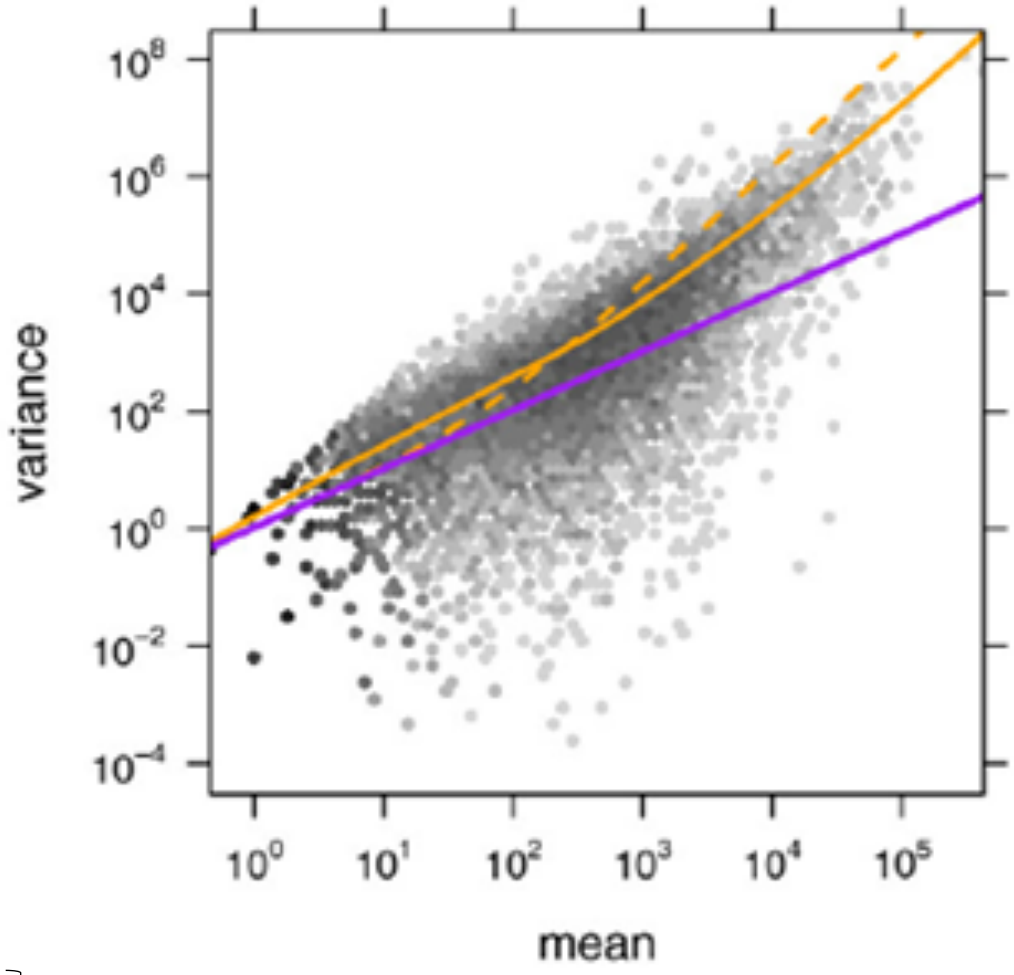
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.367.1606&rep=rep1&type=pdf

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

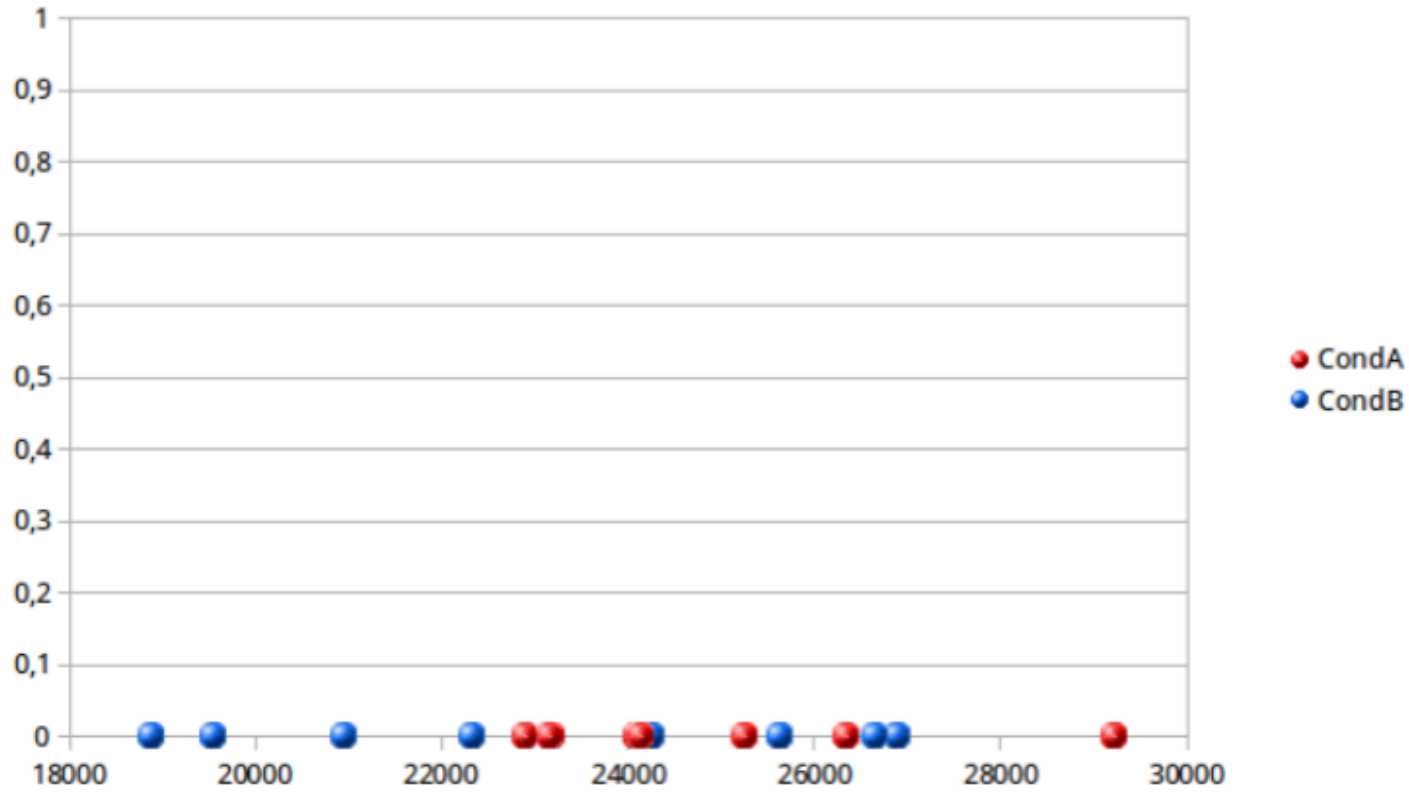
Table 1 Methods for calling differentially expressed genes in RNA-seq data analysis. Total citations were based on Google Scholar search result as of 22 September 2015, and normalized by number of years since formal publication. The methods were ranked according to their citations per year.

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)

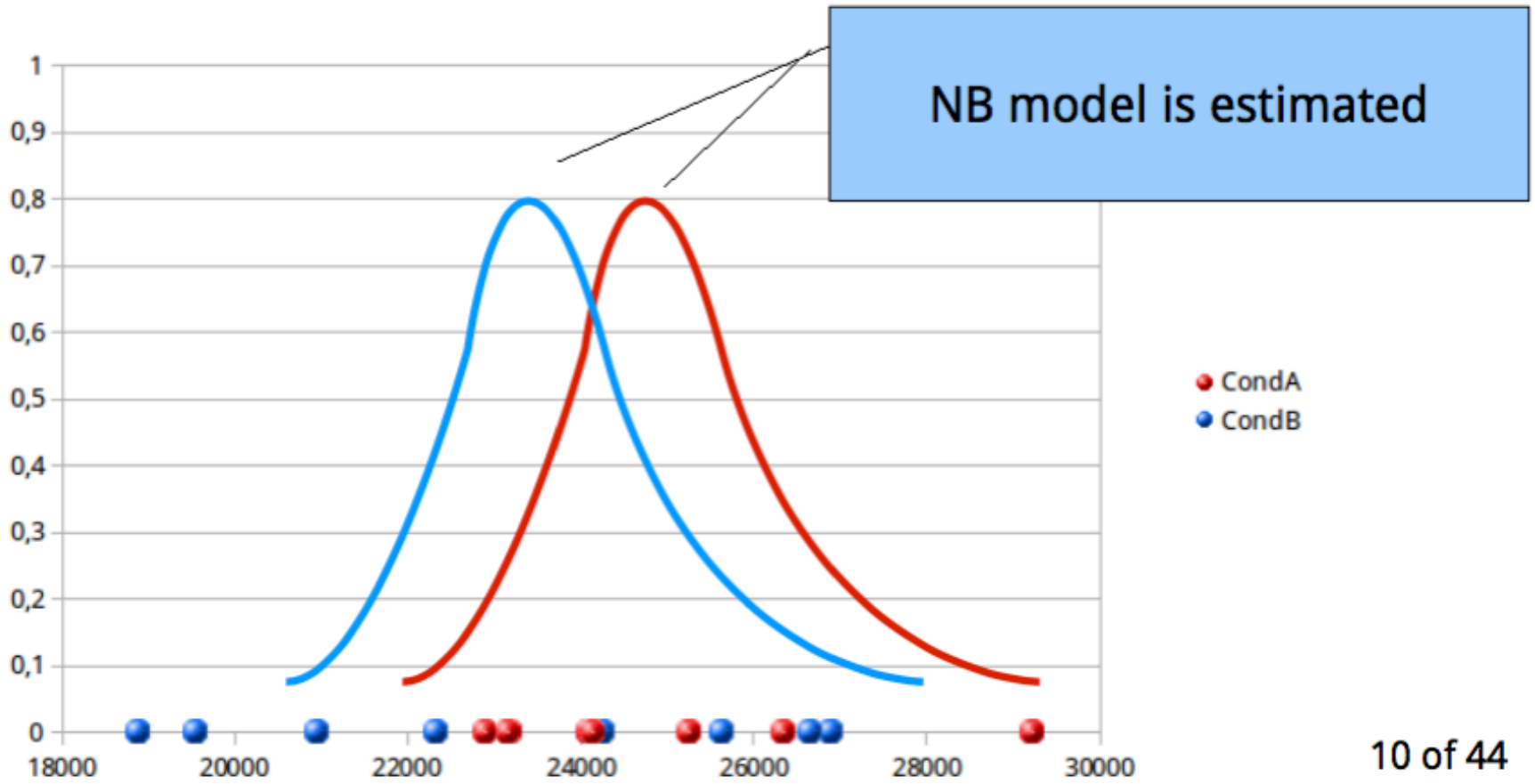
Scenario

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



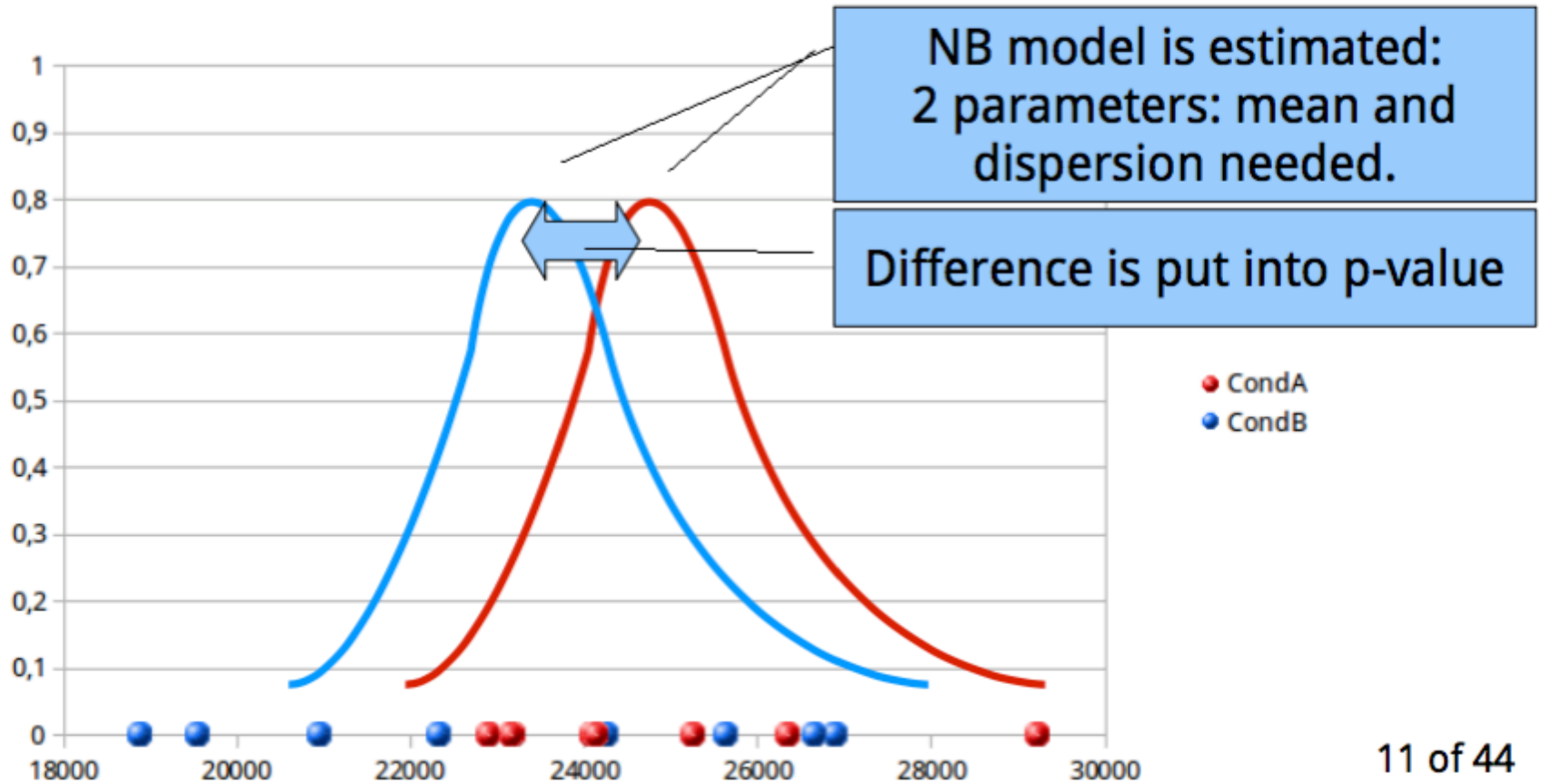
Scenario

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



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

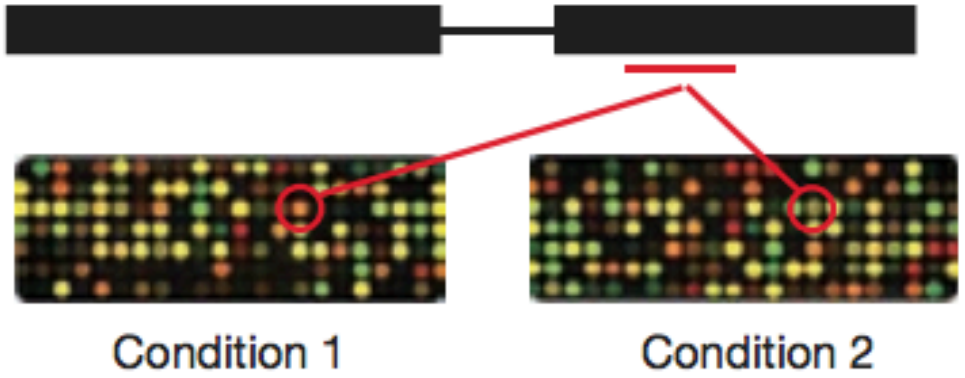


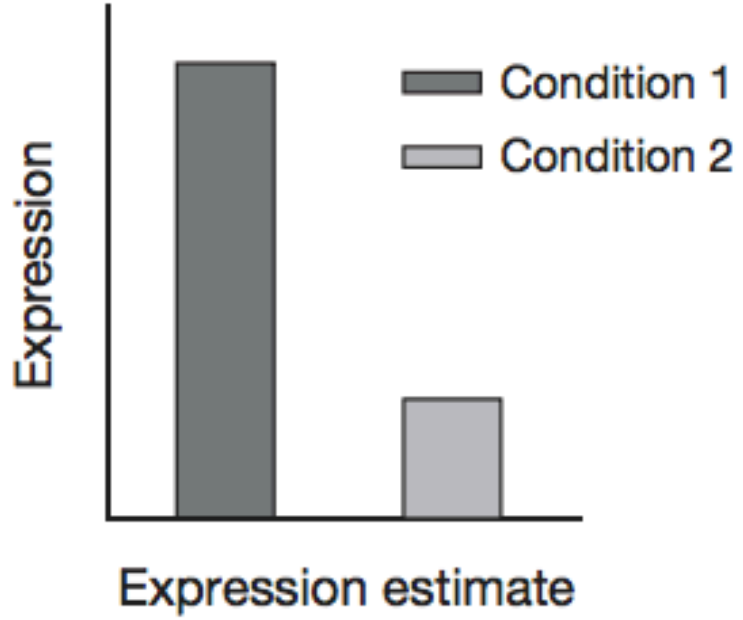
NB model is estimated:
2 parameters: mean and dispersion needed.

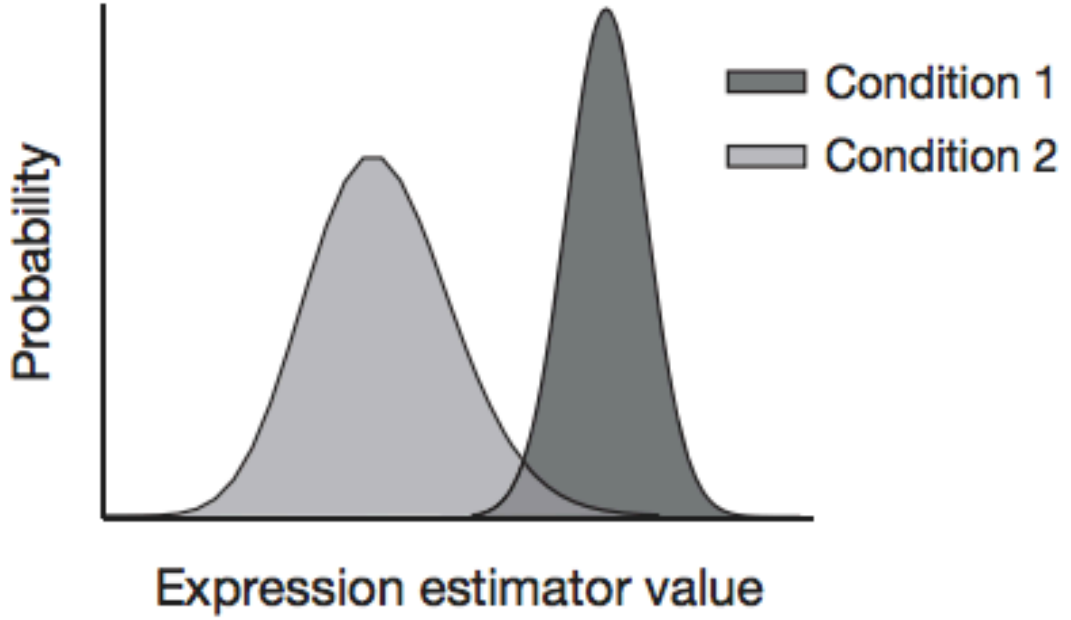
Difference is put into p-value

VIB JAKONIX

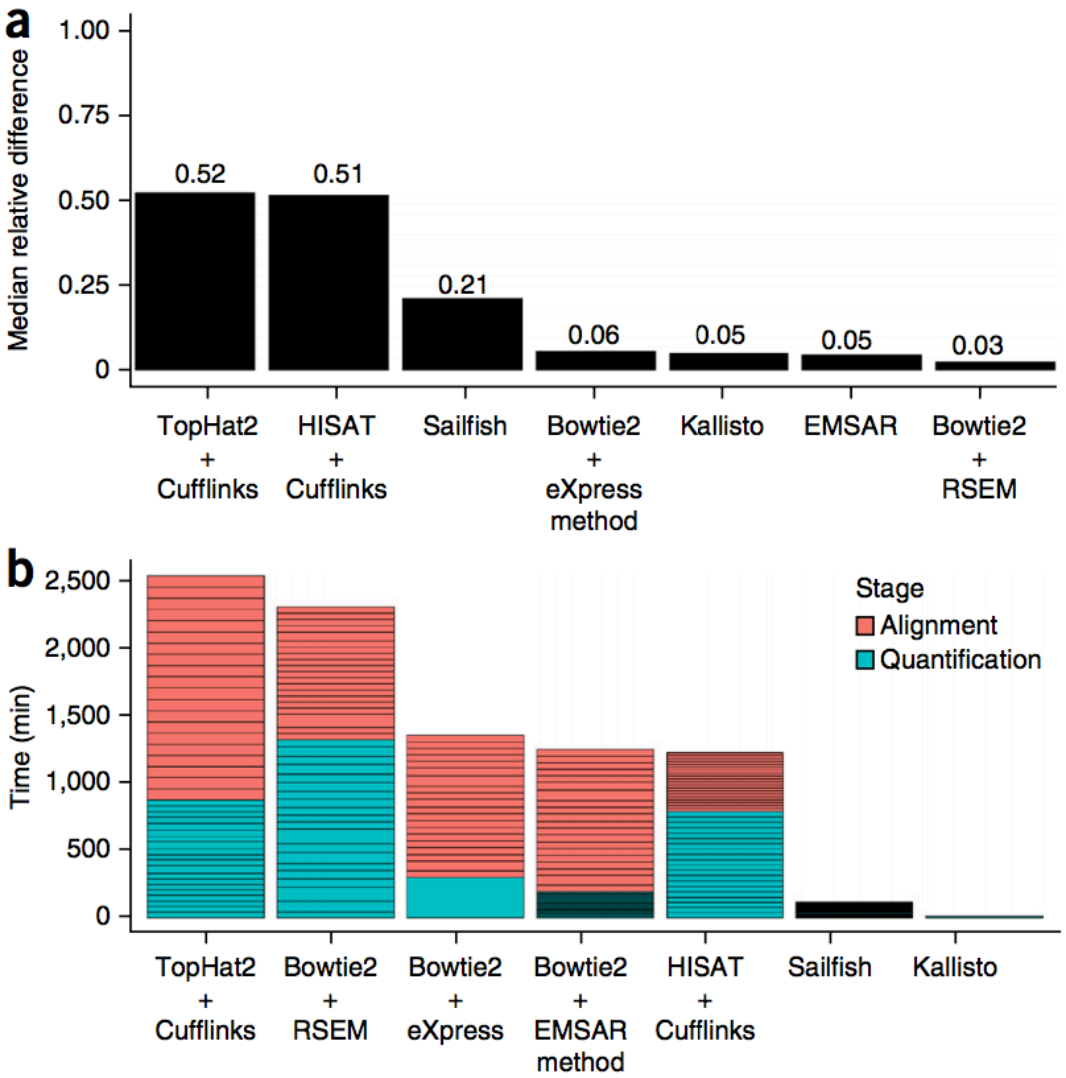
RNAseq vs Microarray







Advances in quantification

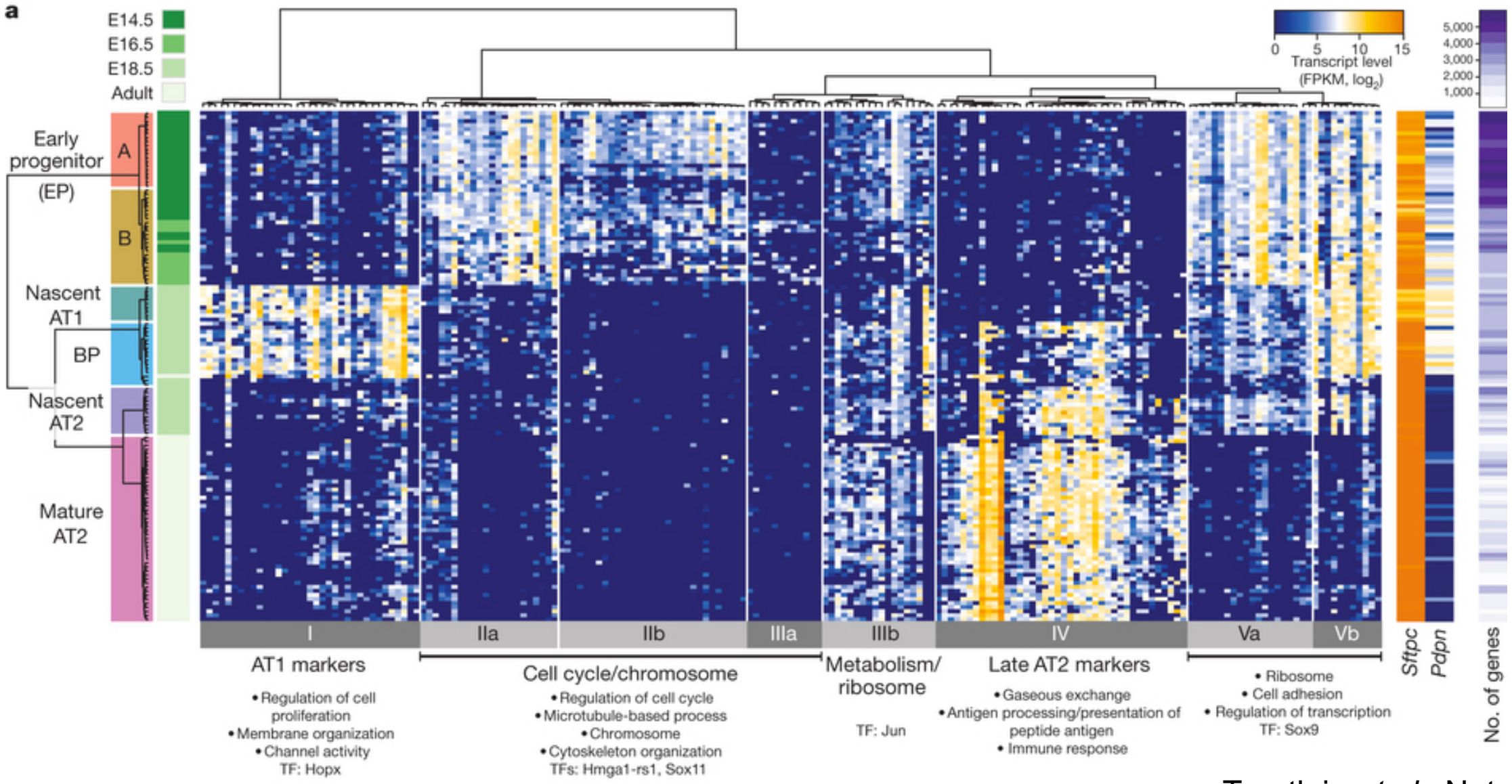


We present kallisto, an RNA-seq quantification program that is two orders of magnitude faster than previous approaches and achieves similar accuracy. Kallisto pseudoaligns reads to a reference, producing a list of transcripts that are compatible with each read while avoiding alignment of individual bases.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

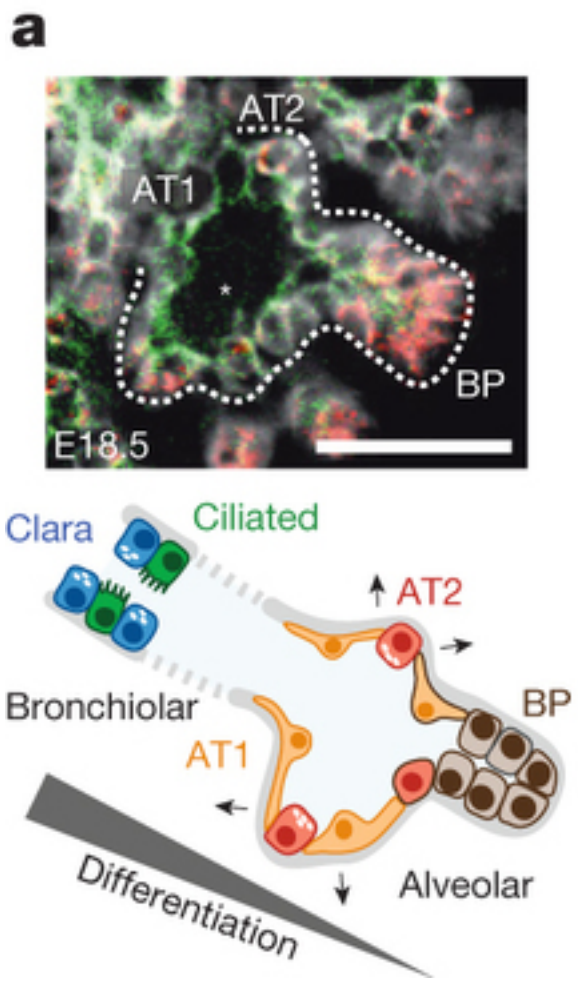
Once you have set of differentially expressed genes

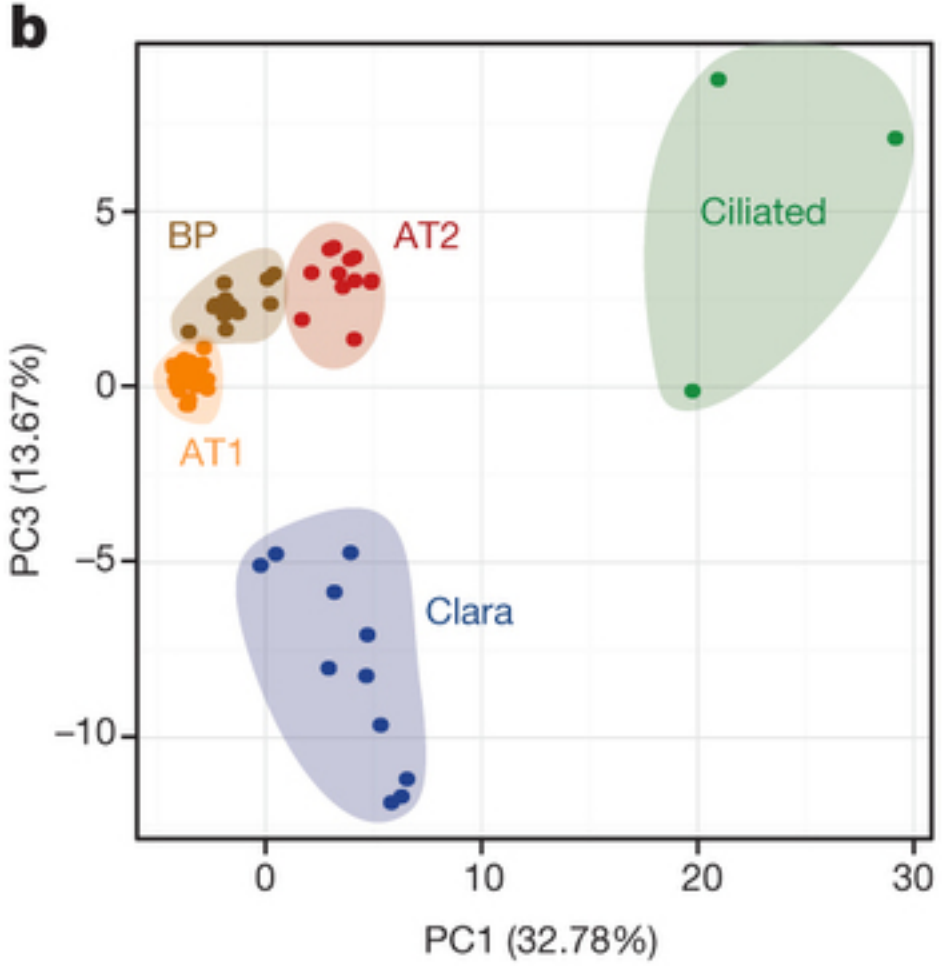
Summarization visualizing the expression data through heatmap ; Classification using Gene Ontology terms and metabolic annotations

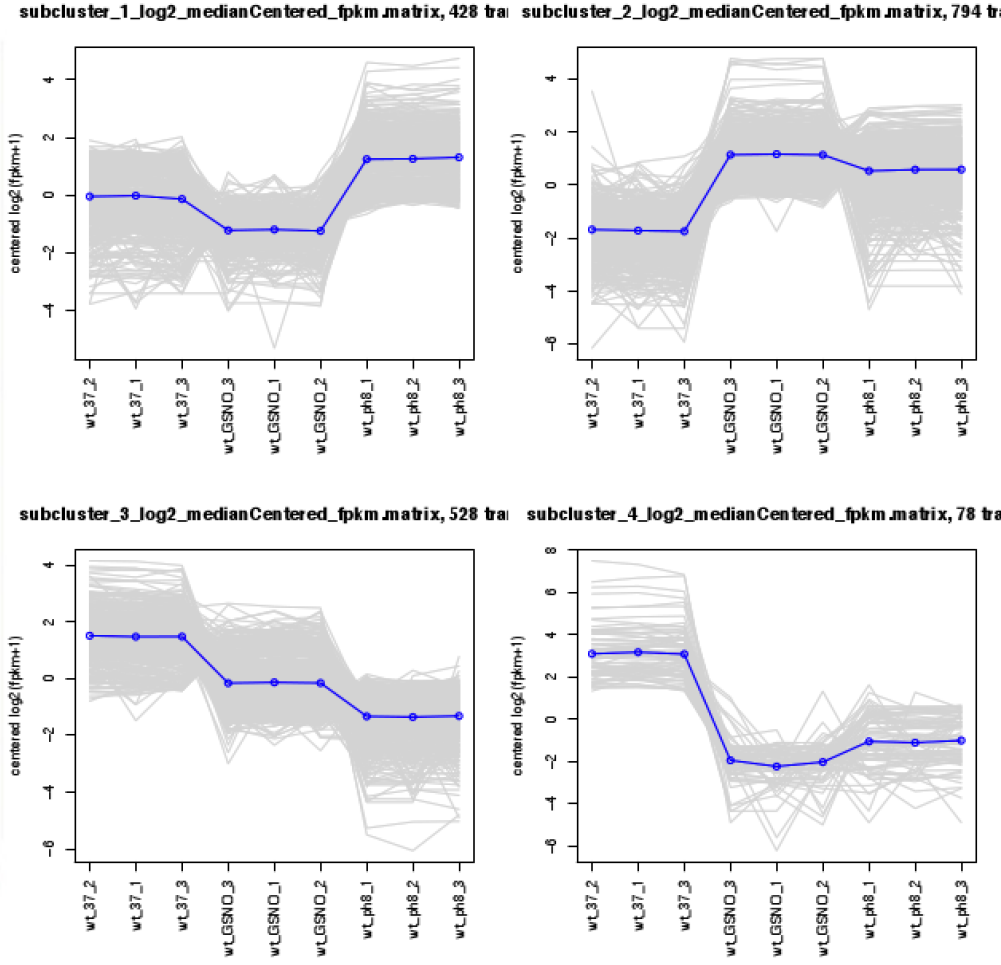


Treutlein *et al.,* Nature (2014)

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

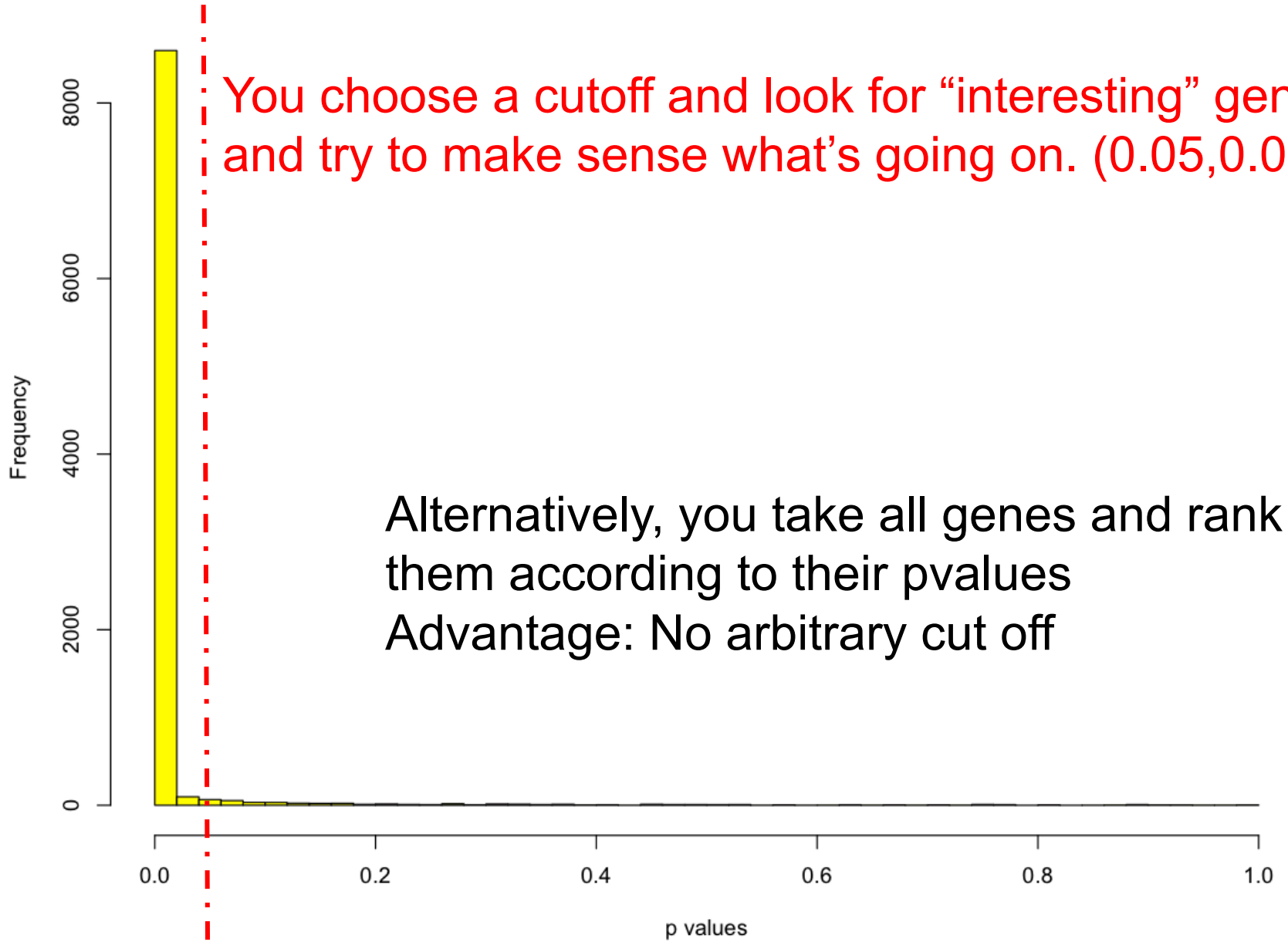






Treutlein *et al.,* Nature (2014)

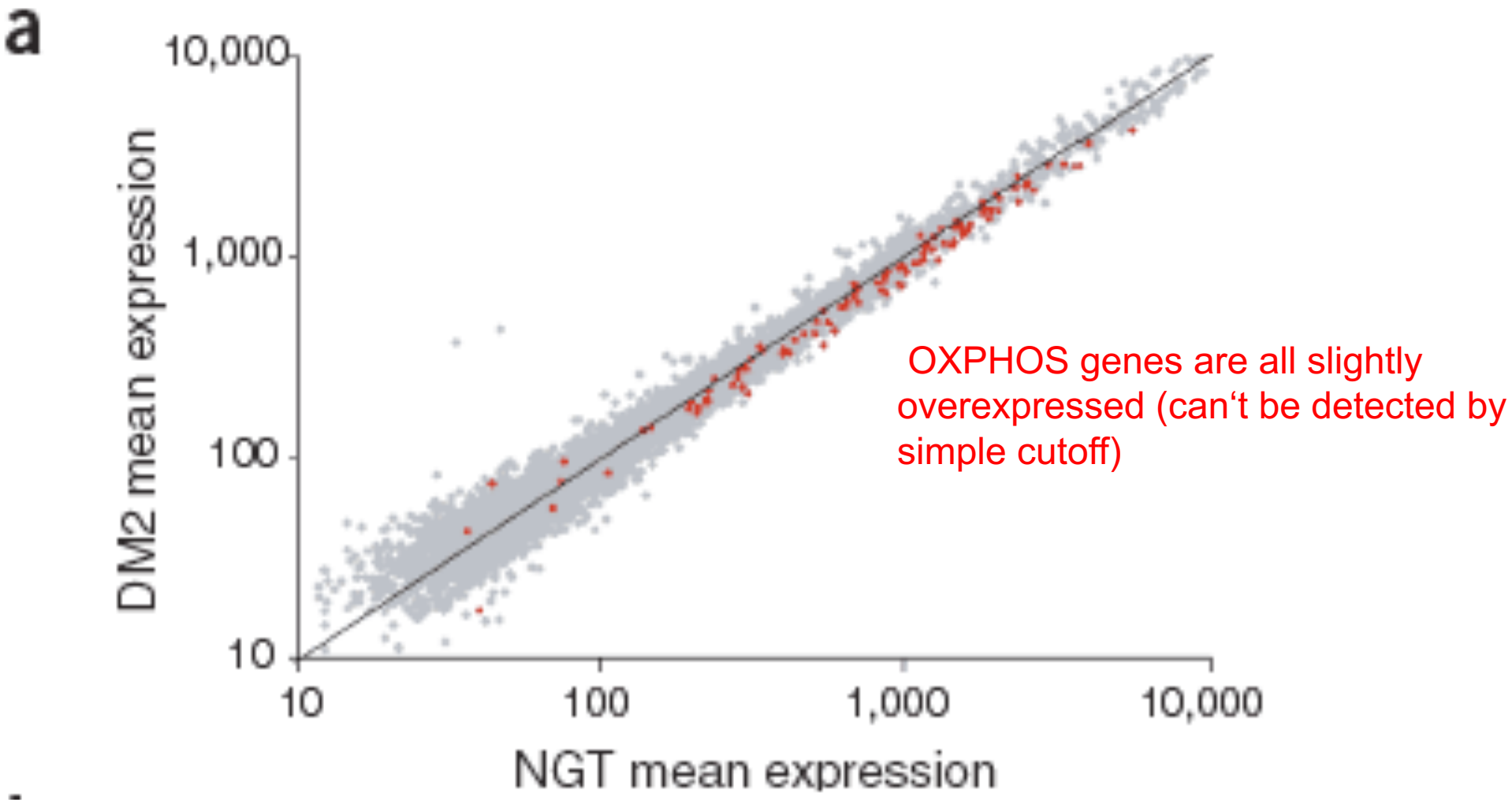
Now, setting a cut-off



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

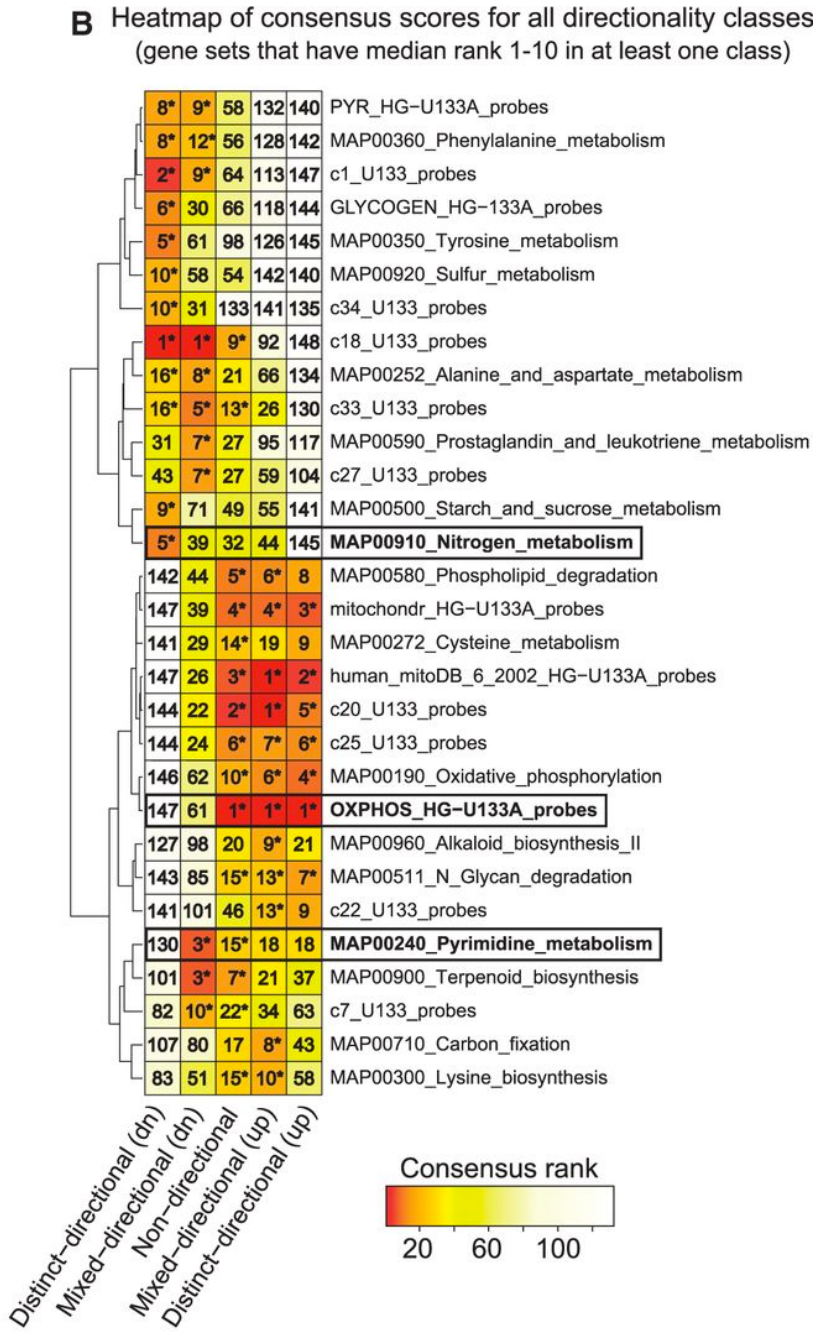
GSEA (Gene Set Enrichment Analysis) methods (cut-off free approach)



OXPHOS genes are all slightly overexpressed (can‘t be detected by simple cutoff)

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

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



Varemo *et al.,* Nucleic Acid Res (2013)

Summary / Our experiences

- Experimental design is key to correctly address your biological question
- Always use replicates (at least 4 if got $$$$)
- Avoid *de novo* transcriptome assembly if you can

- EdgeR and DEseq 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)