KTH Royal Institute of Technology
School of Biotechnology

Analysis of data from high-throughput molecular biology experiments BB2490

Transcriptomics: RNA-seq

Olof Emanuelsson
olofem@kth.se

Lecture 6 2016-02-01 10:15-12:00 FB55

Enabler for Life Sciences

Karolinska Institutet · KTH · Stockholms universitet · Uppsala Universitet · SciLifeLab

Todo-list for next lecture (Wednesday):

read the Furey paper (ChIP-seq) (as much as possible)

one question for the student paper Dobin et al (STAR)
=> email this to me before next lecture

Today's lecture:

1. **Discussion based on reading assignments**
2. **Background: some definitions; gene expression (and its history); the structure of a gene**
3. **RNA-seq, experimental procedure**
4. **RNA-seq, bioinformatics**
5. **Some comments on current status of transcriptome research**
6. **Summary**

[1] Discussion based on reading assignments

Reading for today's lecture:

[1] Garber, Grabherr, Guttman, and Trapnell:

"Computational methods for transcriptome annotation and quantification using RNA-seq". *Nature Methods* vol. **8**, p 469-477 (2011).

Optional:

[2] BB2440 lecture about gene expression.

[3] Wang, Gerstein, and Snyder:

"RNA-seq: a revolutionary tool for transcriptomics". *Nature Rev Genet* vol. **10,** p. 57-63 (2009)

Discuss in pairs your reflections on Garber *et al*.

... and answer these questions:

1. What is, according to Garber *et al.*, the 3 major areas of RNA-seq bioinformatics?

2. Name one topic that you had trouble understanding while reading the paper.

3. What do you think is the main message of this paper?

Time for this: 5 minutes, including taking notes.
Format: write a summary in pairs. Hand it in to me.

A few words about…

Wang, Gerstein, and Snyder:

Too optimistic claims about the lack of need for any normalization in RNA-seq.

Garber, Grabherr, Guttman, and Trapnell:

Sums up the computational tasks nicely but omits (at least) one crucial task: quality control of reads (experiments and alignments).

Table 1, needs completion (e.g., Trinity, STAR).

Table 1 | Selected list of RNA-seq analysis programs

Class	Category	Package	Notes	Uses	Input
Read mapping					
Unspliced aligners[a]	Seed methods	Short-read mapping package (SHRiMP)[41]	Smith-Waterman extension	Aligning reads to a reference transcriptome	Reads and reference transcriptome
		Stampy[39]	Probabilistic model		
	Burrows-Wheeler transform methods	Bowtie[43]			
		BWA[44]	Incorporates quality scores		
Spliced aligners	Exon-first methods	MapSplice[52]	Works with multiple unspliced aligners	Aligning reads to a reference genome. Allows for the identification of novel splice junctions	Reads and reference genome
		SpliceMap[50]			
		TopHat[51]	Uses Bowtie alignments		
	Seed-extend methods	GSNAP[53]	Can use SNP databases		
		QPALMA[54]	Smith-Waterman for large gaps		
		STAR			
Transcriptome reconstruction					
Genome-guided reconstruction	Exon identification	G.Mor.Se	Assembles exons	Identifying novel transcripts using a known reference genome	Alignments to reference genome
	Genome-guided assembly	Scripture[28]	Reports all isoforms		
		Cufflinks[29]	Reports a minimal set of isoforms		
Genome-independent reconstruction	Genome-independent assembly	Velvet[61]	Reports all isoforms	Identifying novel genes and transcript isoforms without a known reference genome	Reads
		TransABySS[56]			
		Trinity			
		Oases			
Expression quantification					
Expression quantification	Gene quantification	Alexa-seq[47]	Quantifies using differentially included exons	Quantifying gene expression	Reads and transcript models
		Enhanced read analysis of gene expression (ERANGE)[20]	Quantifies using union of exons		
		Normalization by expected uniquely mappable area (NEUMA)[82]	Quantifies using unique reads		
	Isoform quantification	Cufflinks[29]	Maximum likelihood estimation of relative isoform expression	Quantifying transcript isoform expression levels	Read alignments to isoforms
		MISO[33]			
		RNA-seq by expectaion maximization (RSEM)[69]			
Differential expression		Cuffdiff[29]	Uses isoform levels in analysis	Identifying differentially expressed genes or transcript isoforms	Read alignments and transcript models
		DegSeq[79]	Uses a normal distribution		
		EdgeR[77]			
		Differential Expression analysis of count data (DESeq)[78]			
		Myrna[75]	Cloud-based permutation method		

[2] Background: some definitions; gene expression and its history; the structure of a gene.

Some definitions:

Gene – a genomic sequence encoding a functional product (or several functional products). Approximately 20,000 genes in the human genome.

Transcript – an RNA species transcribed from a *gene.*

- One gene may produce many different transcripts ("isoforms").
- Each transcript is typically represented by many identical RNA molecules (dynamic range of transcription).

Transcriptome – the set of *transcripts* present in a cell/tissue/organism (at a particular time point or integrated over many time points)

Transcriptomics – finding out everything about the *transcriptome*

Why study transcription and transcriptomics?

transcription ≠ gene expression

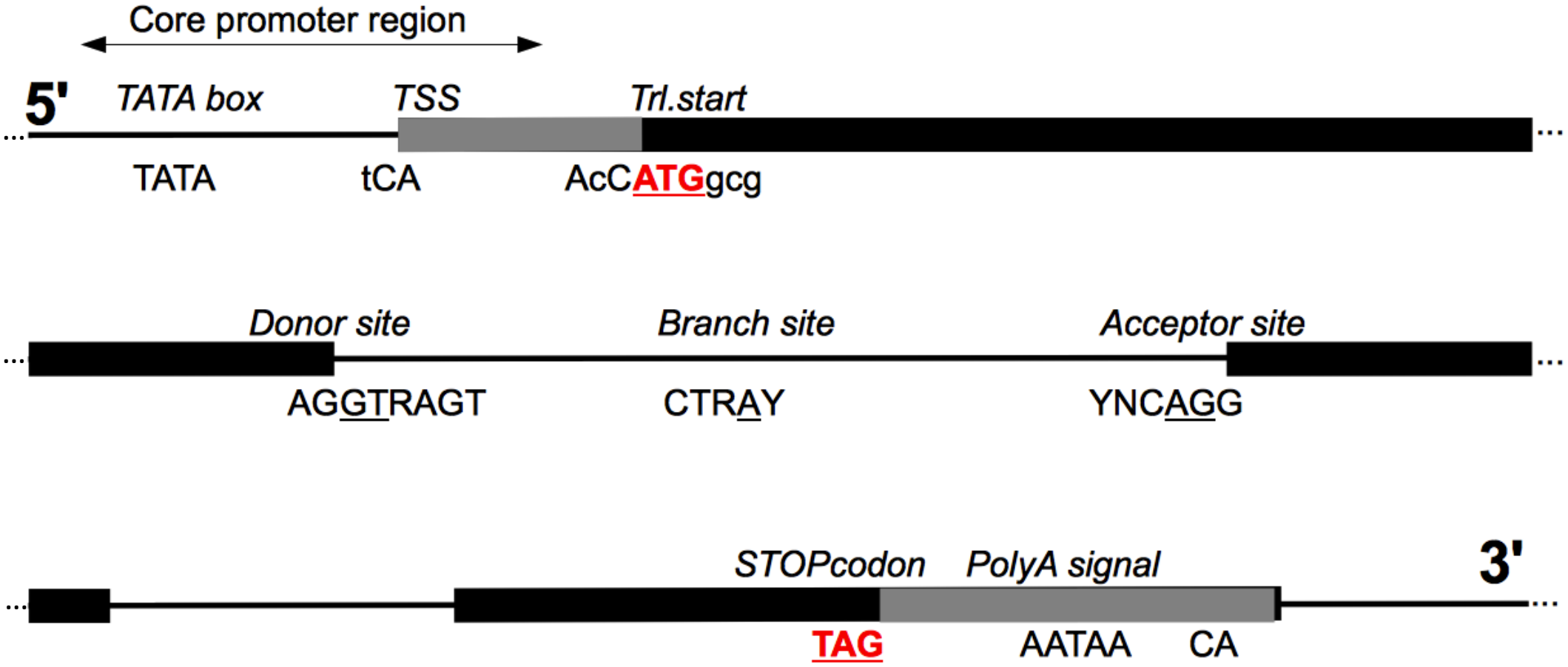
Historically:

- Most interest in finding genes that are **differentially expressed**, e.g., to figure out what genes are involved in a disease (expressed differently in healthy vs disease tissue)

Nowadays, in addition to the above:

- Finer **details** of gene expression: isoforms, haplotypes etc
- The importance of various RNA molecules in **regulation** of cellular behaviour
- Curiosity – how does the cellular machinery work?

A eukaryotic protein coding gene

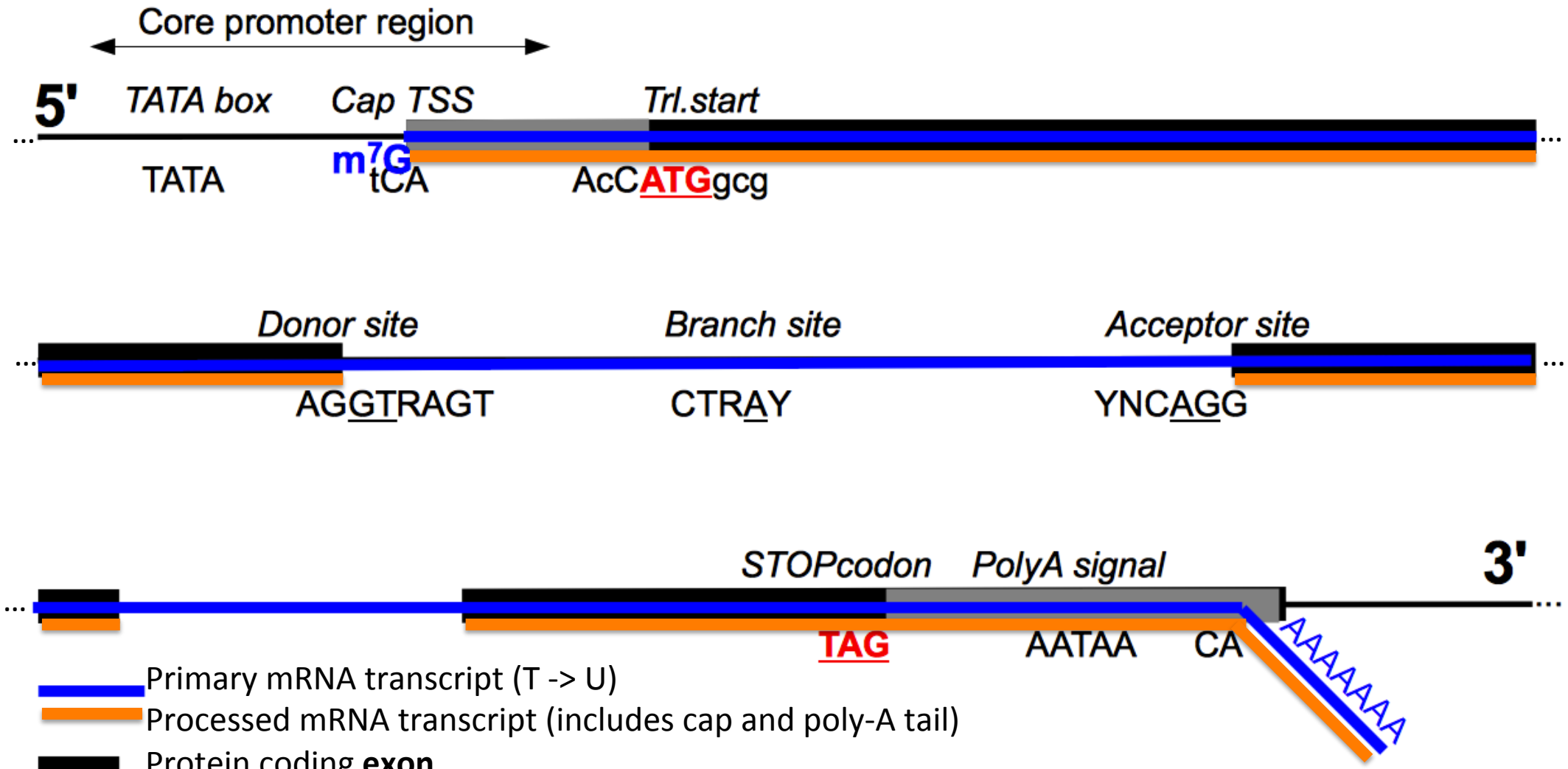


Protein coding **exon**

Non-protein coding part of exon (*untranslated region*, **UTR**)

Intron or intergenic

A eukaryotic protein coding gene

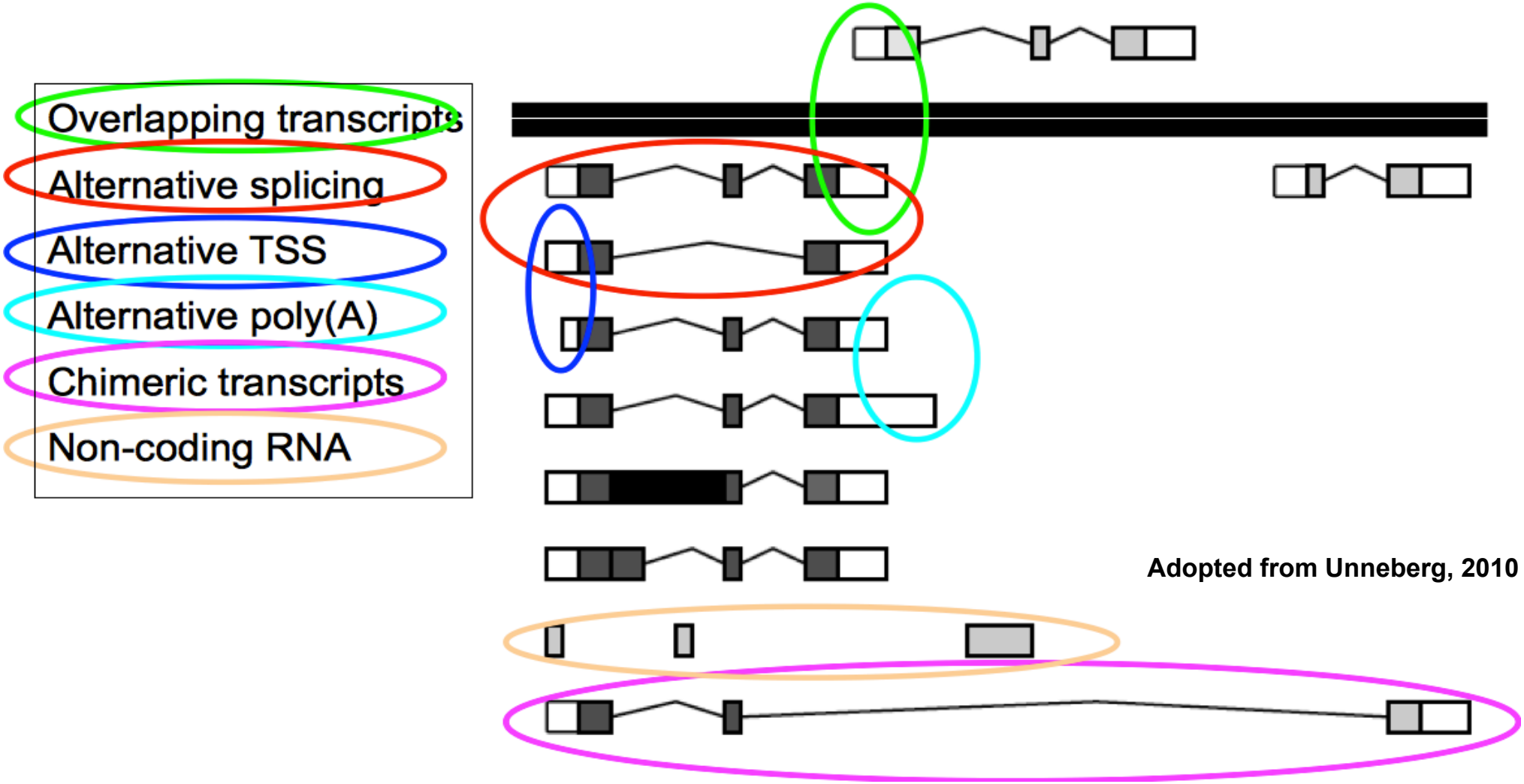


RNA types

rRNA – ribosomal RNA
mRNA – protein coding RNA (messenger RNA)
primary mRNA *vs.* processed mRNA
ncRNA – collective name for all non-protein coding RNA

miRNA – micro RNA, regulates mRNA levels
piRNA – piwi RNA, silencing transposons
lncRNA – long non-coding RNA, e.g., *Xist* – X-chr inactivation
tRNA – transfer RNA, for building protein chains
snRNA – small nuclear RNA, involved in splicing
snoRNA – small nucleolar RNA, modifying mRNA/rRNA/snRNA
eRNA – enhancer RNA (RNA transcribed from enhancers)
PROMPT – promoter upstream transcripts (RNA transcribed from promoters)

Transcriptome complexity



One **gene** can produce many possible **transcripts**.

[3] RNA-seq: experimental procedure

RNA-seq

The state-of-the-art approach for high-throughput analysis of gene expression

The most widely used sequencing based molecular biology assay

PubMed papers with "RNA-seq" in title or abstract

Year	Papers
2015	2289
2014	1562
2013	1054
2012	837
2011	472
2010	217
2009	58
2008	16

RNA-seq

What application(s) are you interested in:

- mRNA abundance
- differential expression
- novel transcription
- antisense transcription
- transcriptome reconstruction without reference genome
- allele-specific expression
- non-coding RNAs

An RNA-seq experiment

1. Sample collection
2. RNA extraction
3. Library preparation
4. Sequencing
5. Data analysis

Sample collection and RNA extraction

Sample collection

Important considerations:

- number of biological and technical replicates
- is it possible to get enough material (i.e., RNA)?

Why are replicates important?

Many other crucial aspects. e.g. ethical permits.

RNA extraction

Follows standardized protocols

Library preparation

From RNA to sequencing-ready DNA molecules

1. Quality checks on samples. E.g., RIN value, 0-10.
2. **Fragmentation** of RNA molecules
3. **Depletion/enrichment:** get the RNA types you need
4. Random priming [or poly(A) priming then fragmentation]
5. cDNA generation (reverse transcription)
6. Adapter ligation, cluster generation

=> RNAs converted to cDNA and ready for sequencing

(RIN = RNA integrity number; 10 is best quality)

Fragmentation

Processed mRNAs are typically on the order of 1-2 kbases (in humans).

They are fragmented during library generation.

Typical fragment size: ~300 bases.

Depletion/enrichment

rRNA is ~ 80-90% of total RNA

rRNA typically uninteresting

rRNA should (typically) be removed

Strategies:

- Remove rRNA (e.g. RiboZero or RiboMinus kit)
- Enrich for mRNA through extracting poly-adenylated RNA (most mRNA and some lncRNA)
- size selection (if interested in e.g. miRNA)

Sequencing

Several different technologies available for sequencing: sequencing chemistry, amplification strategy, read length, number of reads, base calling accuracy, sequencing errors (rates and types), … all this differ.

The output is called “a **read**” – a stretch of DNA sequence.

Sequencing output

Fastq: *de facto* standard for output files

(1) DNA sequence for each read

(2) Quality for each base in each read

```
@SEQ_ID_1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Illumina HiSeq: Typically >10^9 reads of 2x150 bases, corresponding to >300 Gbases per run.

Illumina MiSeq: 2x300 bases.

Base quality

Base quality: $Q = -10 \log_{10} P$

Probability of wrong base: $P = 10^{\frac{-Q}{10}}$

Quality scores:

Quality score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

Sequencing depth (= coverage)

Sequencing depth is the total number of nucleotides sequenced divided by genome (or transcriptome) length:

Genomic DNA sequencing

depth = $L_{read}N/G$

L_{read} = read length

N = number of reads

G = genome length

RNA-sequencing

depth$_{RNA\text{-}seq}$ = $L_{read}N/T$

L_{read} = read length

N = number of reads

T = transcriptome length

Sequencing depth (= coverage)

Calculate the sequencing depth for an RNA-seq experiment. 12 M paired-end reads (2x125) from a transcriptome with estimated total exon length of 60 M base pairs.

$depth_{RNA\text{-}seq} = L_{read}N/T$

L_{read} = read length

N = number of reads

T = transcriptome length

Average $depth_{RNA\text{-}seq}$ *= ?*

Is the coverage even over the transcriptome?

"The power to identify and accurately quantify RNA molecules is dependent on their lengths and abundance, and on the number of sequenced reads." (Sims et al Nat Rev Genet 2014)

[4] RNA-seq: bioinformatics

RNA-seq bioinformatics

Reads have been generated. Then three main tasks:

0. Remove low quality reads, adapter sequences etc.
1. Read mapping

   Map (align) reads to reference genome or transcriptome

2. Transcriptome reconstruction

   Reconstruct the transcriptome from the reads without a reference

3. Expression quantification

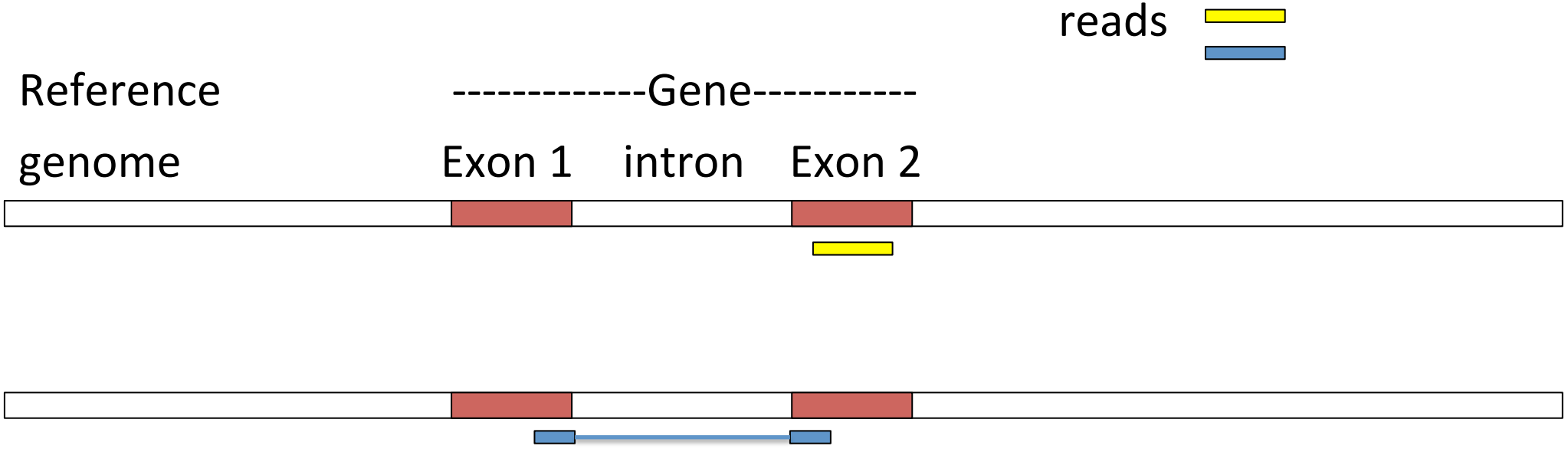
   Transcript expression levels

   Differential expression analysis

Map reads to a reference genome

Find the place (=map or align) on the genome (=reference) from which the read originated (=was transcribed). *Important: allow spliced reads*.

The reference genome is the genomic DNA sequence of the organism.

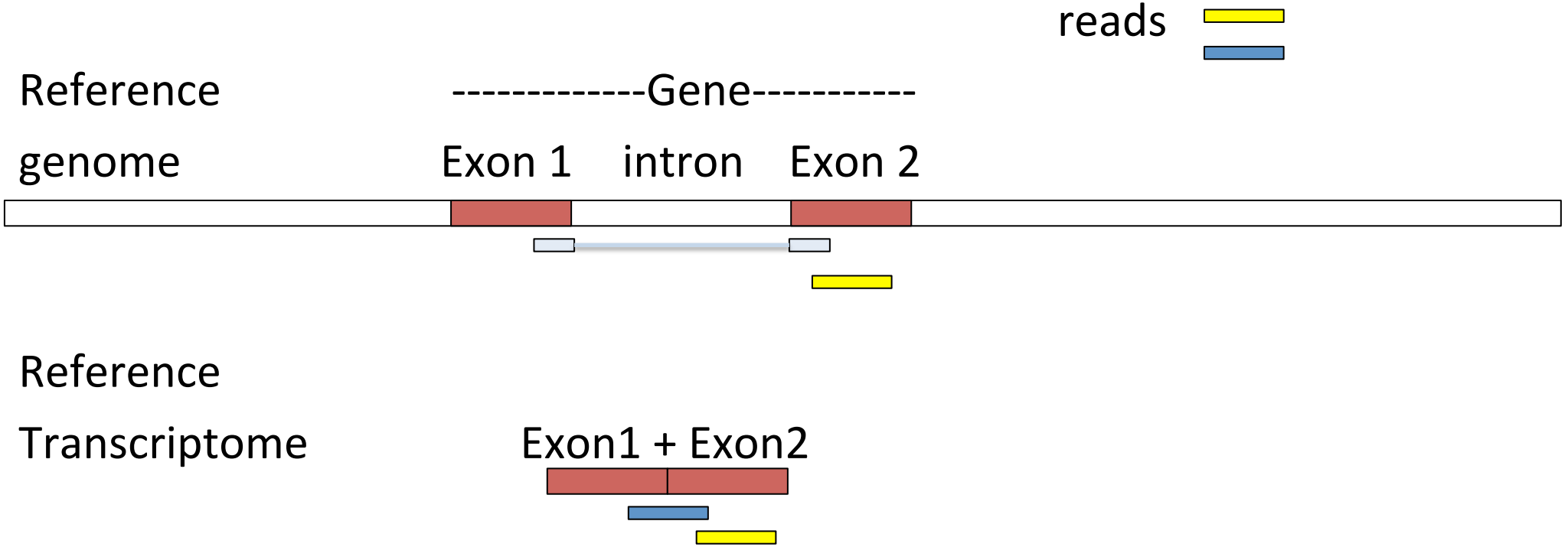


Tools for this: TopHat, GSNAP, STAR, ...

Map reads to a reference transcriptome

Find the place (=map or align) on the *transcriptome* (=reference) from which the read originated.

The reference transcriptome is the total set of RNAs of the organism.



Tools for this: BWA, Bowtie, Maq, SOAP, Gnumap, ...

Gene body coverage

Are the mapped reads equally distributed over the length of genes?

Depends on

- RNA sample quality (RIN)
- Depletion/enrichment strategy
- Sequencing depth

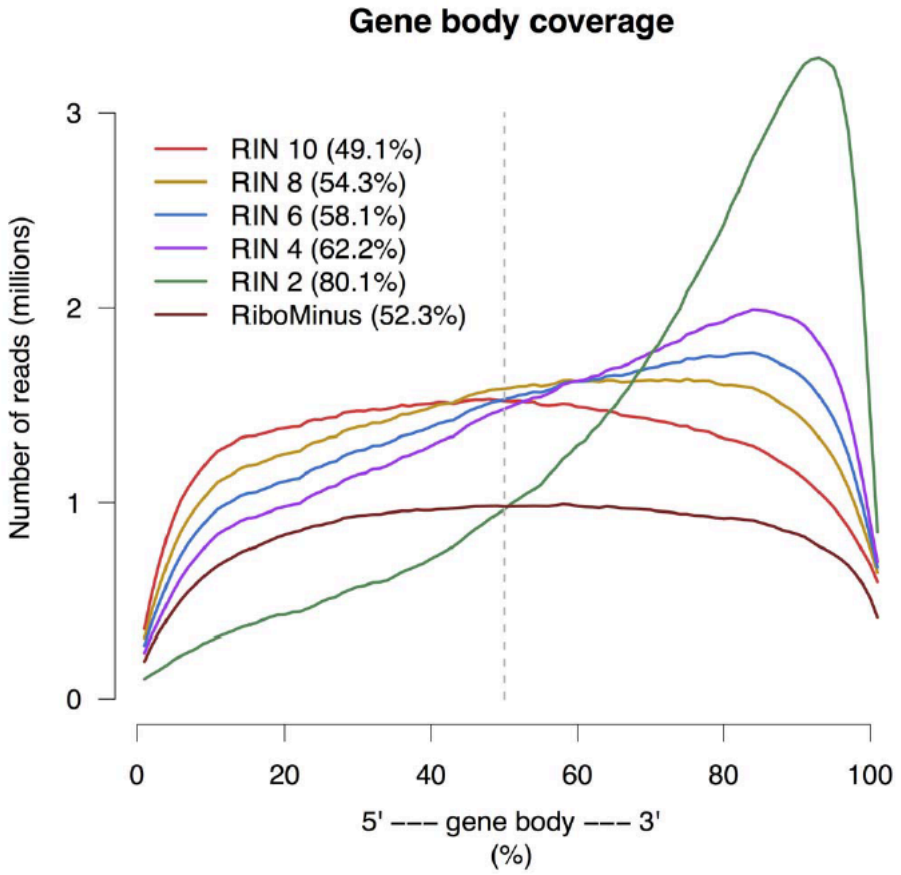
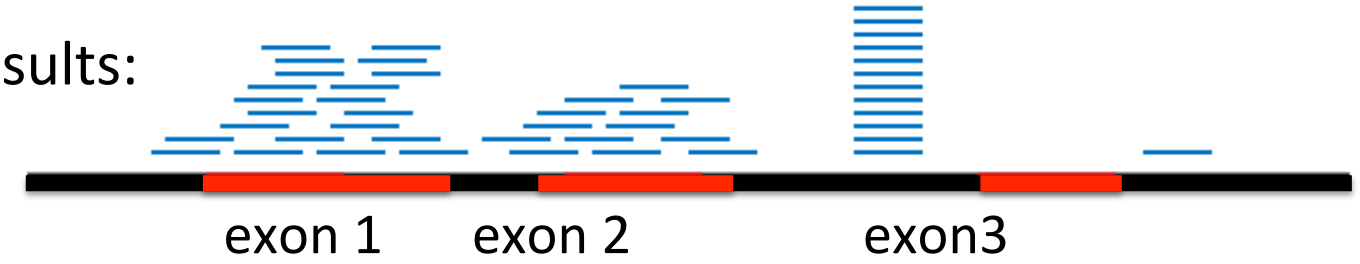


Figure 5. Gene body coverage on average for each group. Both RIN 10 and RiboMinus show even coverage. The percentages in the paranthesis show the relative amount of reads that map closer to the 3′ end than to the 5′ end, i.e. the amount of reads that map to the right of the dashed vertical line. Each step of decreasing RIN shows an increase in 3′ bias.
doi:10.1371/journal.pone.0091851.g005

Ambiguous reads

Mapping results:



Ambiguous reads:

one read matching to *many* places in the reference

=> these reads are usually discarded as they are ambiguously mapped and hence their origin is unclear

=> or they are added to one, or all, of its possible origins according to certain algorithms. (*Can you think of one principle?*)

many reads matching to **one** exact place in the reference

=> referred to as “PCR duplicates” and usually discarded [e.g. Picard, GATK]

=> but if sequencing depth is reasonably large, we would expect duplicate reads and they should not be blindly removed

Transcriptome reconstruction

The goal is to find out: what transcripts are present in a sample.

For humans, approximately 40-80% of all possible genes are actually expressed in a given sample.

If you know the reference genome: use this knowledge to guide your reconstruction. If you also know the gene annotation (and hence the exon-intron structure of the genes), use this knowledge as well.

If you do not know the reference genome: then you have to perform a *de novo* assembly, in many ways similarly to how you would assemble the genomic DNA sequence. But with the obvious difference that your final goal is not to assemble *n* chromosomes, but *N* different transcripts, where $N>>n$, and where a subset of the *N* transcripts may be overlapping (isoforms).

Transcriptome reconstruction

i.e., reconstruct the transcripts present in the sample

<u>Genome-guided approach:</u>

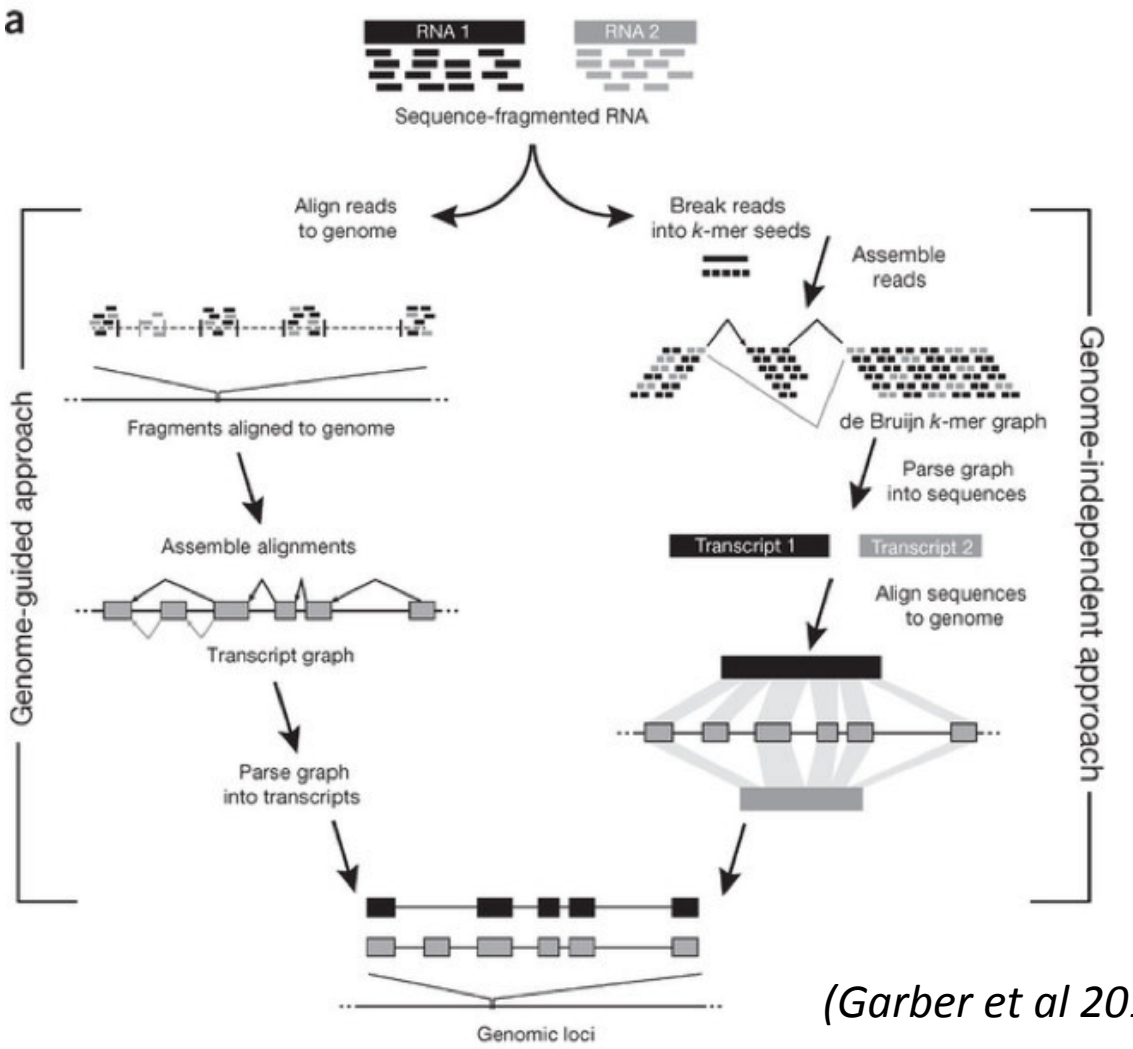
If there is a reference genome

(*e.g.*, Cufflinks)

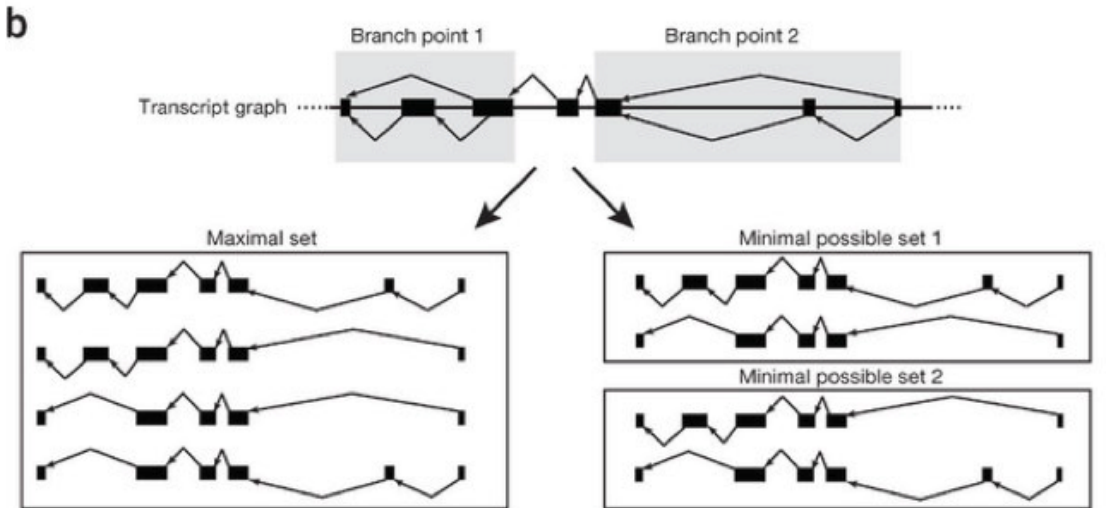
<u>Genome-independent approach:</u>

The only option if there is no reference genome

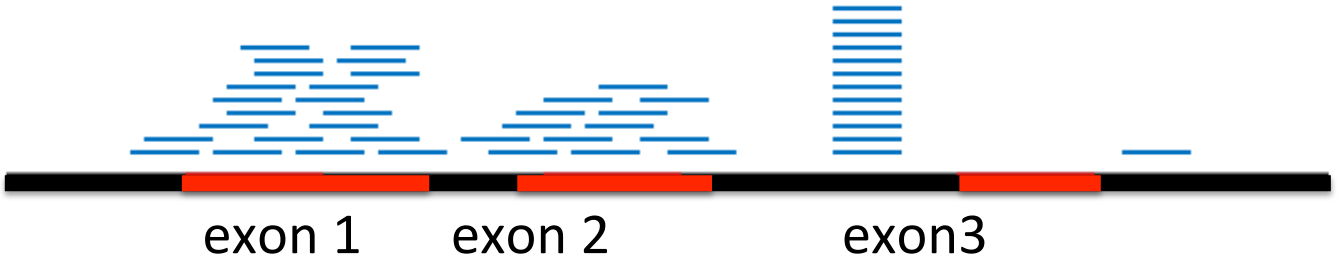
(*e.g.*, Trinity, Oases, transABySS)



(Garber et al 2011)



Transcript quantification: count the reads



Exon 1: 21

Exon 2: 13

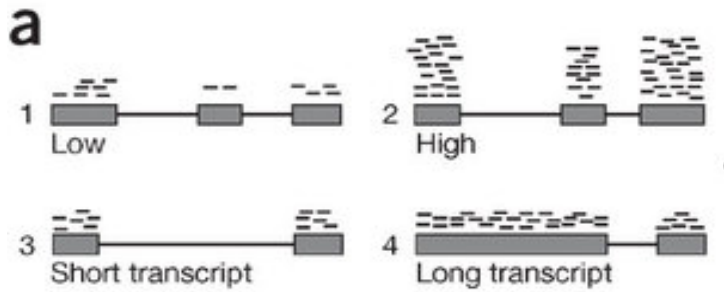
Exon 3: 0

Total for the gene: 34

I.e., if a transcript reconstruction method was applied to this particular data set, it would produce a transcript consisting of exon1 and exon2.

Estimate abundance

Read counts can be misleading:



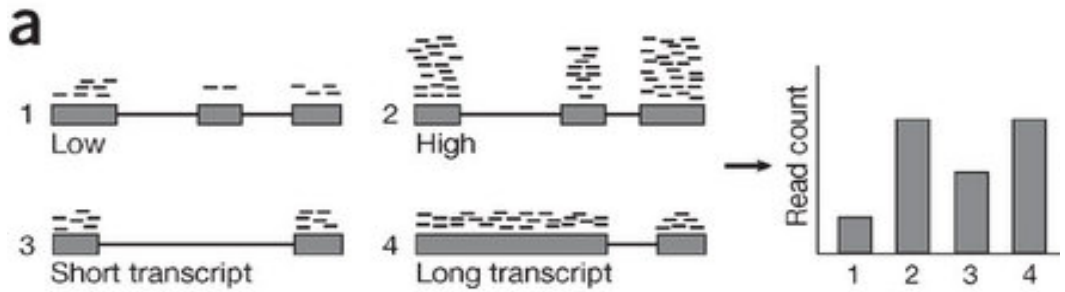
Transcript 3: 12 reads

Transcript 4: 31 reads

But… are 3 and 4 unequally expressed?

Estimate abundance

Read counts can be misleading:



Transcript 3: 12 reads

Transcript 4: 31 reads

But… are 3 and 4 unequally expressed?

Normalize RNA-seq read counts: RPKM and FPKM

RPKM – reads per kilobase of transcript per million mapped reads

$$RPKM_g = \frac{C_g \times 10^9}{L_g \times N}$$

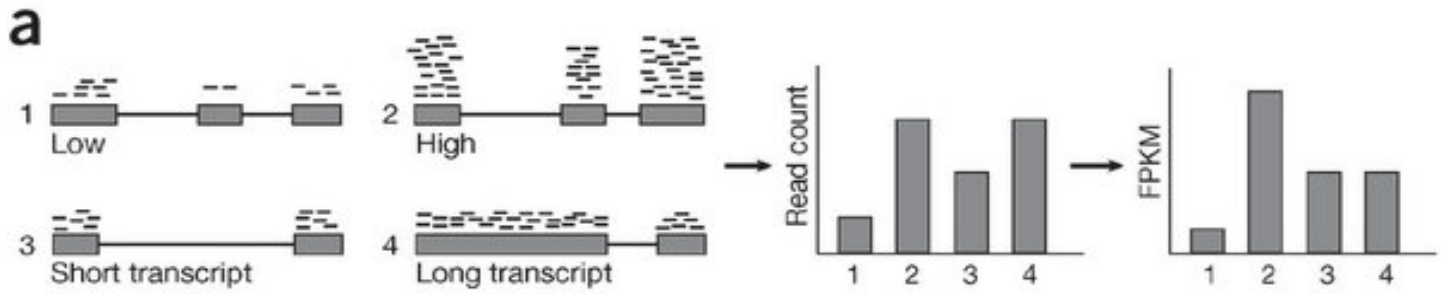
- $RPKM_g$ value for gene (transcript) *g* (or $FPKM_g$ for paired end reads)
- C_g= number of reads mapped to the gene
- L_g= length of gene *g*
- *N*= number of mappable reads; $N = \sum_{i \in G} C_i$, (where *G* is the total set of genes

Normalizes for (i) transcript length and (ii) number of reads.

FPKM is similar to RPKM but takes into account the fact that in paired-end sequencing, each fragment may be represented by either one or two reads.

Estimate abundance

Use *RPKM* instead of read counts:



Transcript 3: 12 reads

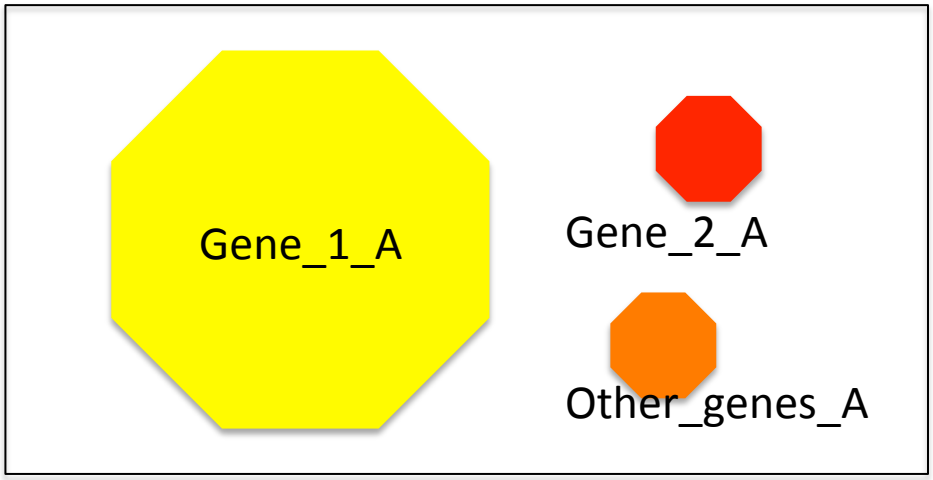
Transcript 4: 31 reads

But their RPKM is the same.

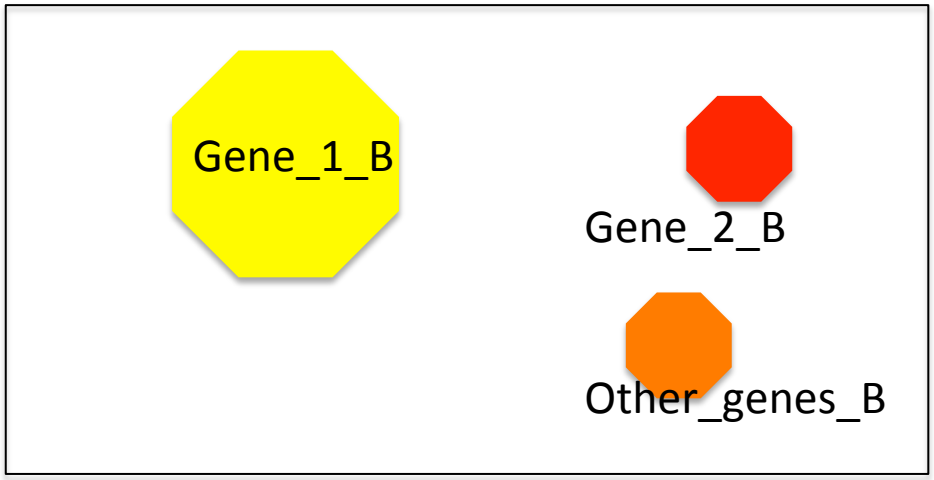
Tools for this: ERANGE, Myrna, eXpress, ...

Quiz!

RNA population sample A



RNA population sample B



(Coloured areas represent number of present mRNA molecules)

The two samples A and B are from two different tissues A, B, of the same individual. Genes 1 and 2 are, by far, the most expressed genes in both samples.

Gene 1 (yellow) is expressed much more in sample A than in sample B

Gene 2 (red) is expressed equally much in sample A and sample B.

Both samples are sequenced to $10 * 10^6$ mappable reads.

*Which statement is **true**?*

a. The RPKM for every gene is the same in sample A as in sample B

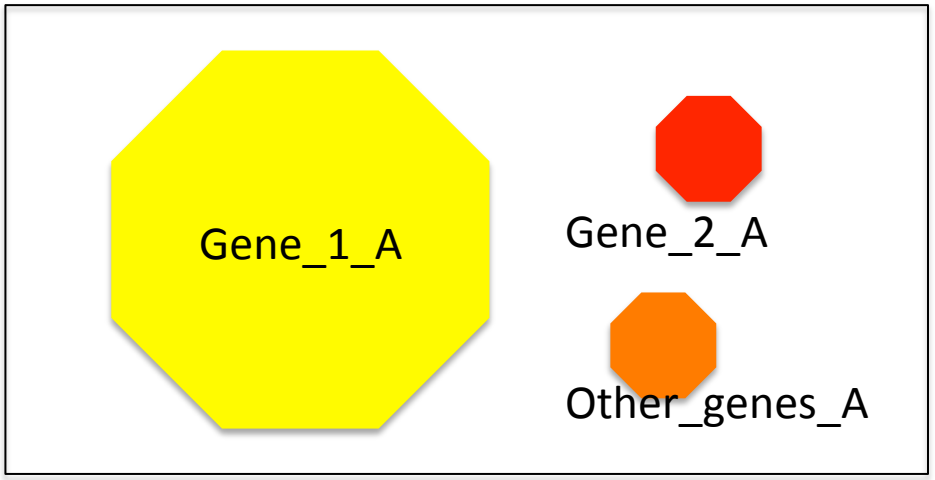
b. Gene 2 will have the same RPKM in both samples

c. Gene 2 will have lower RPKM in sample A than in sample B

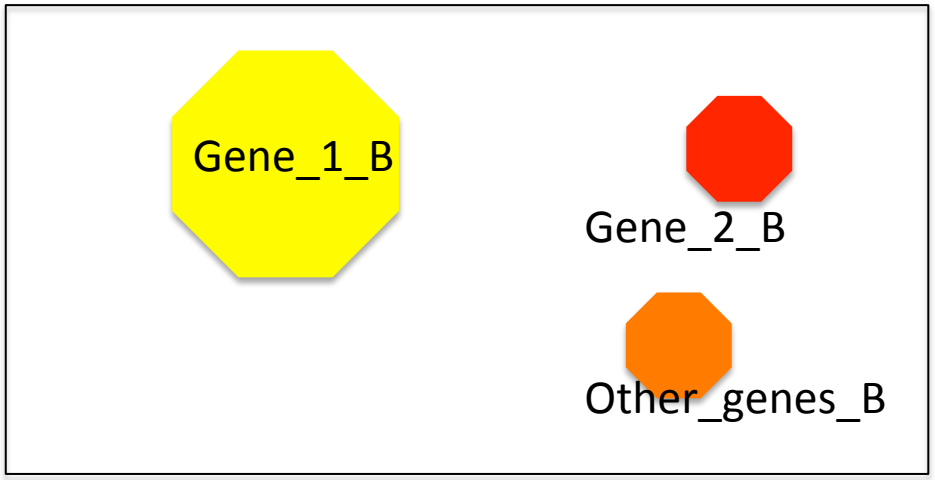
d. Gene 2 will have higher RPKM in sample A than in sample B

Quiz!

RNA population sample A



RNA population sample B



(Coloured areas represent number of present mRNA molecules)

The two samples A and B are from two different tissues A, B, of the same individual. Genes 1 and 2 are, by far, the most expressed genes in both samples.

Gene 1 (yellow) is expressed much more in sample A than in sample B

Gene 2 (red) is expressed equally much in sample A and sample B.

Both samples are sequenced to 10 * 10^6 mappable reads.

*Which statement is **true**?*

a. The RPKM for every gene is the same in sample A as in sample B

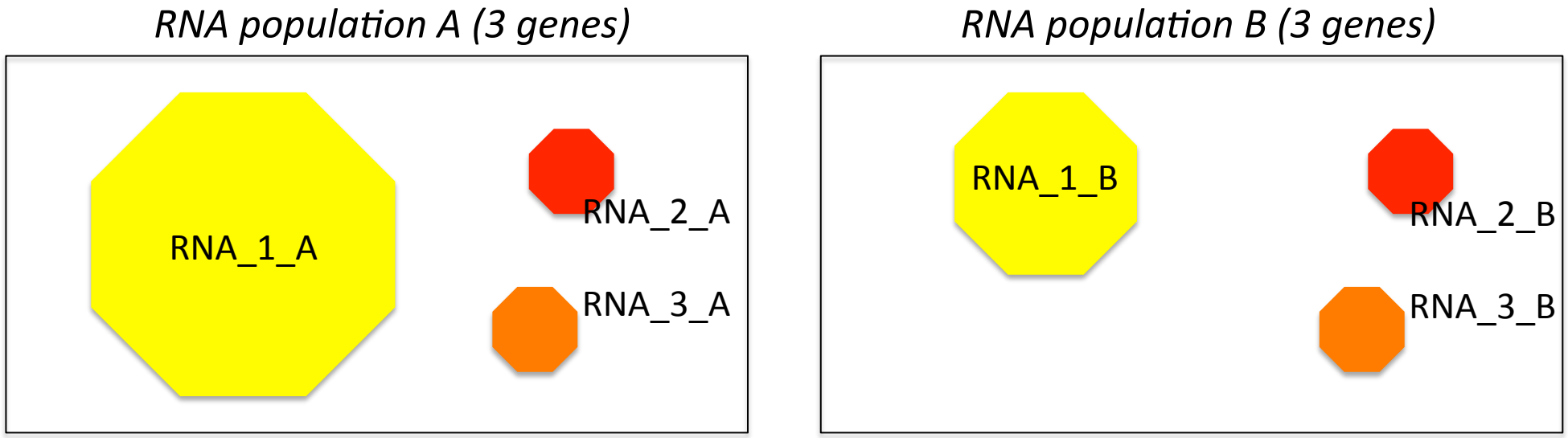
b. Gene 2 will have the same RPKM in both samples

c. Gene 2 will have lower RPKM in sample A than in sample B

d. Gene 2 will have higher RPKM in sample A than in sample B

Estimate abundance

RPKM has widespread use. But consider this:



Given the same sequencing depth (i.e., the same number of reads), RNA 2 and 3 will get lower RPKM in population A than in population B, although the expression levels (the actual number of mRNA molecules) are the same in the two populations. Thus, need to normalize for the composition of the RNA pool.

Tools for this: EdgeR, DEseq

Normalize RNA-seq read counts: TPM

TPM: “transcript per million”

$$TPM_g = \frac{\frac{C_g}{L_g} \times 10^6}{\sum_{i \in G} \frac{C_i}{L_i}}$$

I.e., very similar to RPKM, but takes the length of each transcript into account when summing the total number of reads, in order to better represent the **true total abundance of the transcripts**.

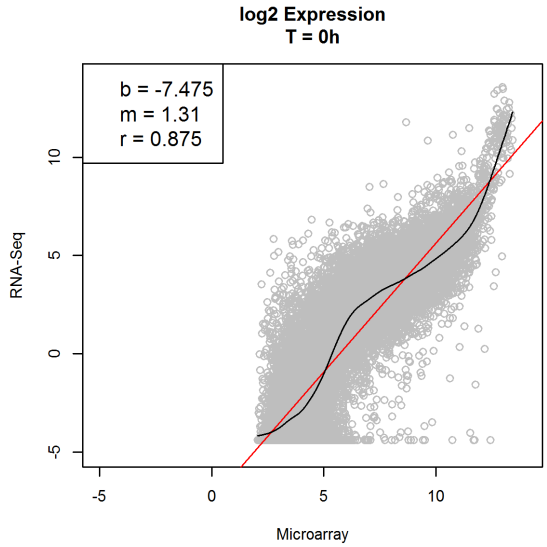
- TPM_g = “transcript per million” for gene *g*
- C_i= number of reads mapped to gene *i*
- L_i= length of gene *i*

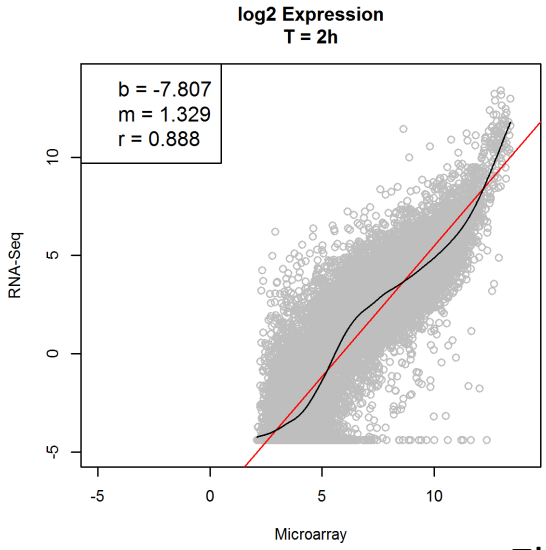
Dynamic range

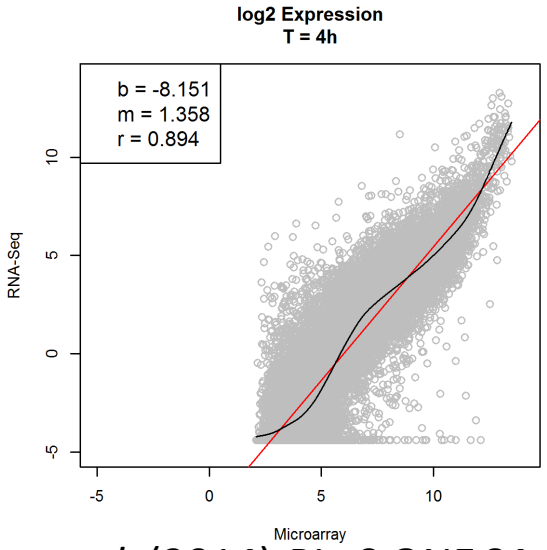
The expression level of a transcript influences how well we can measure it:

- Low expressed transcripts:
  - Microarrays have high background signal -> poor measurement
  - RNA-seq can measure well if you sequence very deeply
- Medium expressed transcripts:
  - Microarrays measure well
  - RNA-seq measures well if sequenced relatively deeply
- High expressed transcripts:
  - Microarrays measure poorly because of saturation
  - RNA-seq measures well

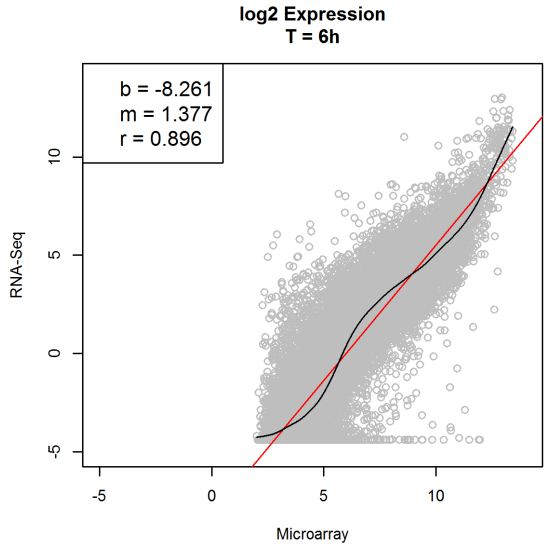
Dynamic range – one example (T-cells)

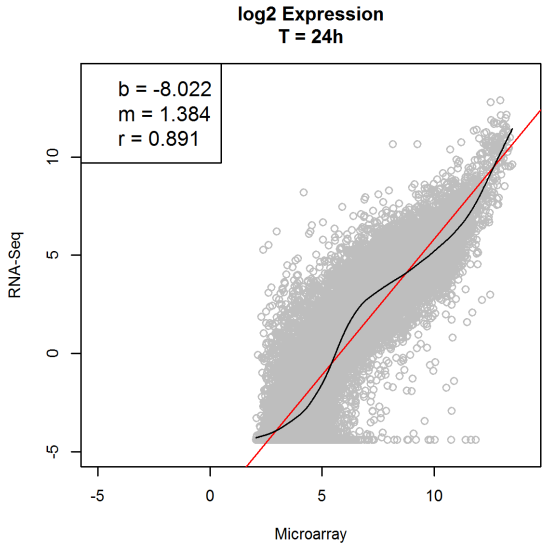


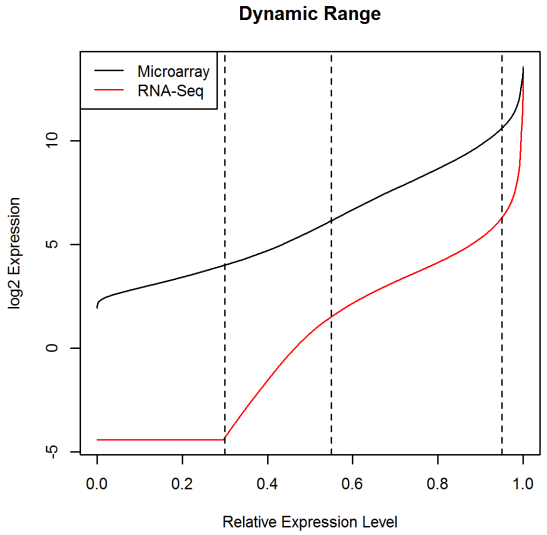




Zhao *et al*. (2014) *PLoS ONE* **91**:e78644







Microarray: ~4×10^3

RNA-seq: ~3×10^5 (general RNA-seq estimation ~10^6)

Differential gene expression (DE)

When is a difference in read count for a gene also statistically significant?

=> Model the variability in read count for each gene across replicates.

The read count variability has been modelled using

Poisson distribution (e.g., DEGseq, Myrna)

models the variance between technical replicates

Negative binomial (e.g. edgeR, DEseq, CuffDiff)

also models the overdispersion of read counts between biological replicates (biological replicates are less similar than technical replicates)

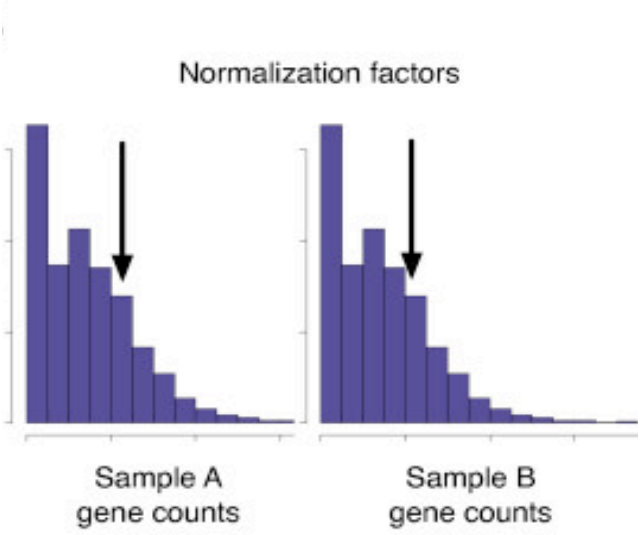
=> near consensus that this is the preferred modelling

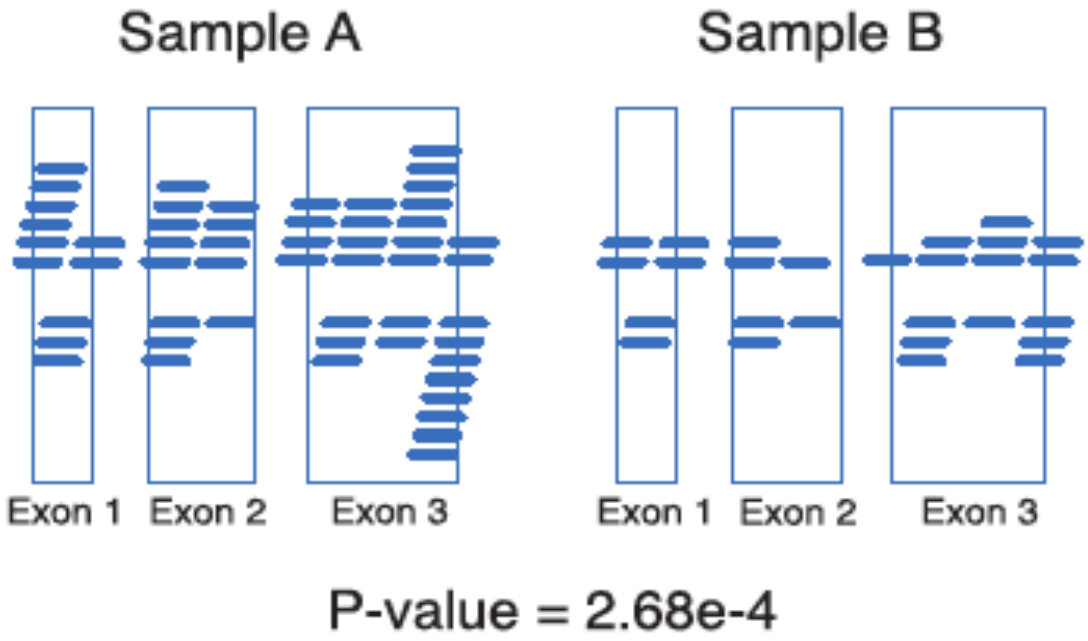
<u>Output is, for each gene</u>

⇒ a ***P*-value** describing the probability that the difference in counts is due to chance. (Should be corrected for multiple hypothesis testing).

⇒ **Effect size** (fold change)

Differential gene expression (DE)





Counts for the genes in the different samples are used as input:

[1] Variability of read counts for each gene across replicates is modeled

[2] *P*-value is calculated – what is the probability that this difference (or larger) in counts/RPKMs would occur by random chance

[3] Effect size is calculated (fold change)

[4] Multiple testing correction is (or should be) performed

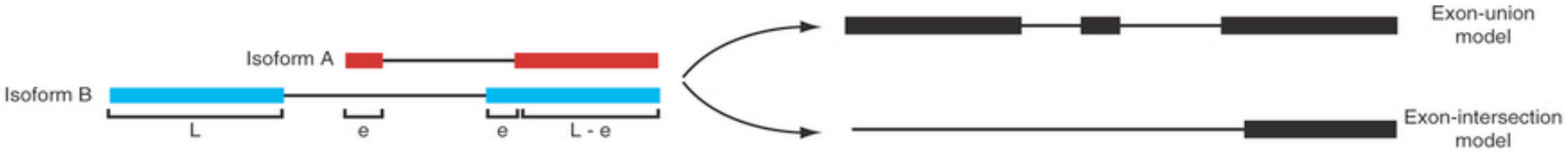
NOTE: instead of explicitly using normalized counts or RPKM/FPKM, adjust for sequencing depth and composition within the DE software. E.g. median-based normalization.

Alternative: use non-parametric methods (i.e. rank), then no need to assume a distribution (skip [1])

Differential gene expression

Transcriptome complexity: isoforms (=splice variants = transcripts).

We do not know from which isoform (transcript) a particluar read originated.



(Isoform B is 2x the length of Isoform A)

<u>Two approaches:</u>

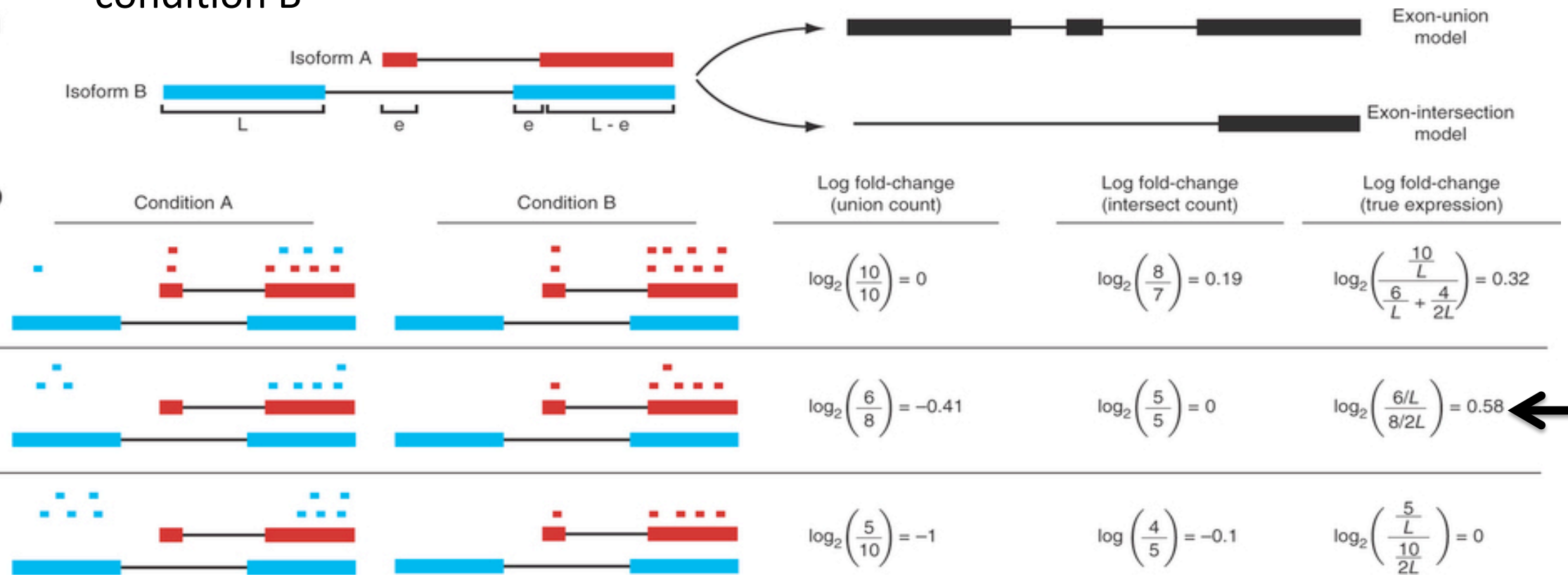
Exon union – use the read counts from all exons in all isoforms

Exon intersection – use the read counts only from exons present in all isoforms

Differential gene expression

These approaches could both provide the wrong answer.

E.g.: if one isoform is upregulated and the other downregulated in condition B



Trapnell et al., Nature Biotech. 2013

Differential transcript expression

Isoform-aware (a.k.a., transcript-based) differential expression analysis:

CuffDiff

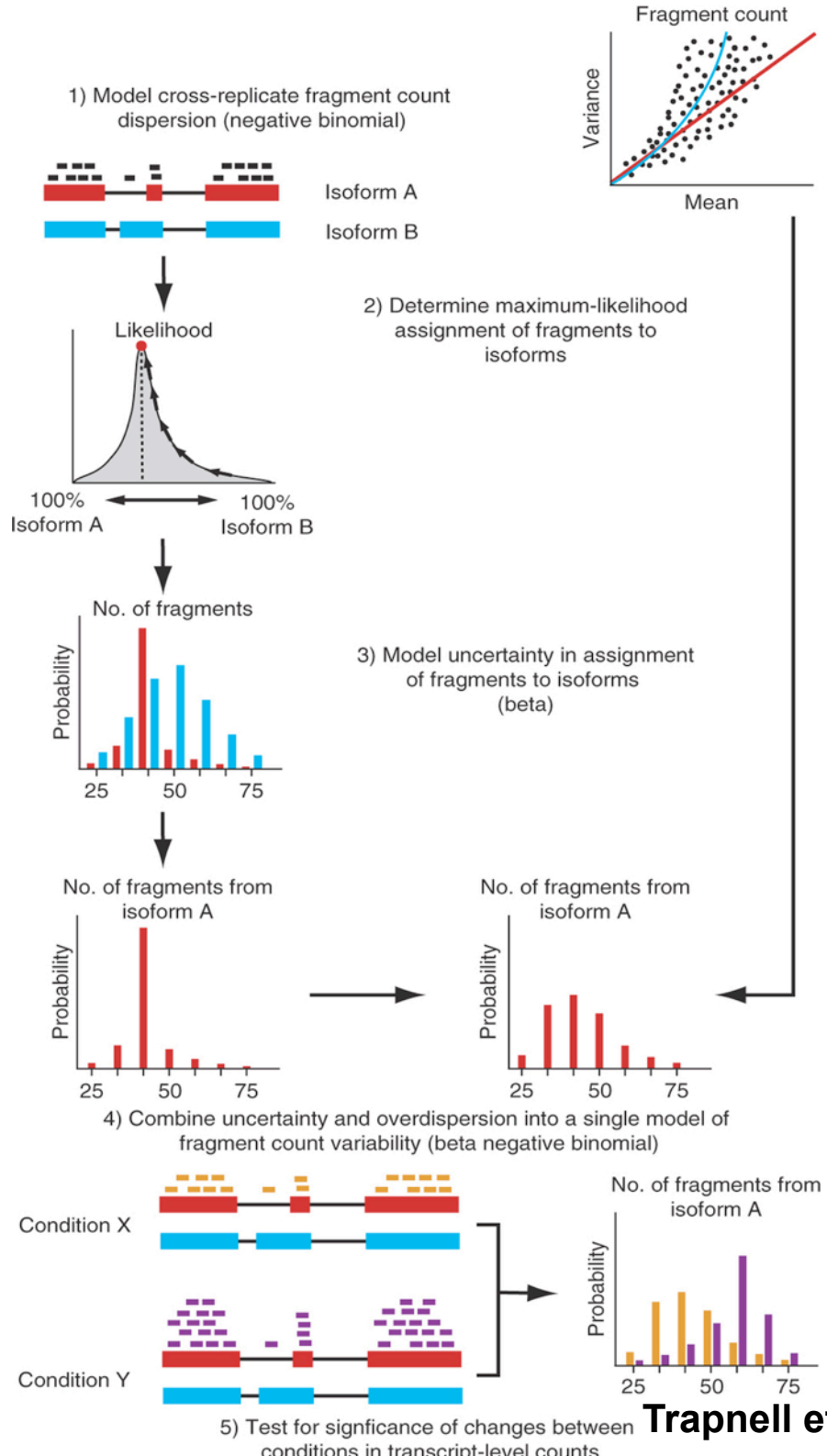
Model overdispersion of data (like previously mentioned tools)

Maximum-likelihood assignment of fragments to isoforms

Model uncertainty in assignment of fragments to isoforms

Combine these into a final model of fragment count variability, which is used to calculate a *P*-value describing the probability that the difference occurred by random chance.

=> Output is a *P*-value and effect size



1.Model count variability across replicates

2.Estimate fragment count for each transcript

3.Model uncertainty in fragment count for transcripts

4.Combine (1) and (3) to generate final model of fragment count variability

5.These variance estimates are used in the statistical testing of differential expression.

Trapnell et al., Nature Biotech. 2013

Differential expression

Which method is best – isoform aware *or* union/intersection methods?

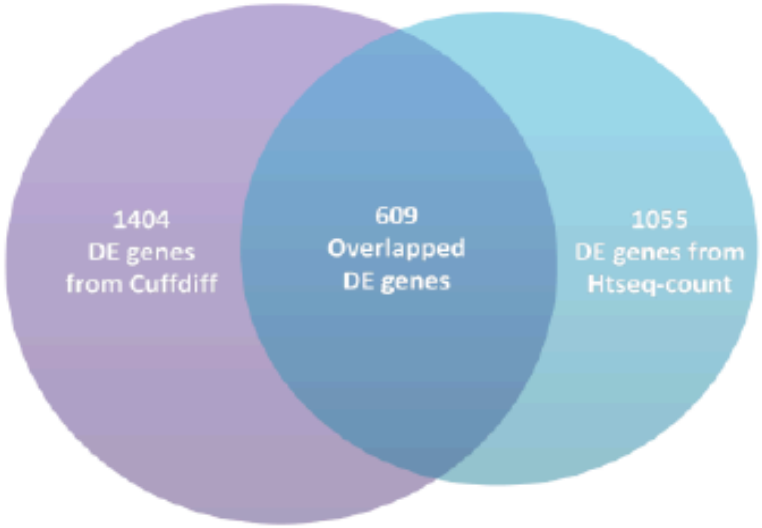
Benchmarking studies show that exon union/inter-section methods hold up surprisingly well.

[E.g., Soneson & Delorenzi 2013]

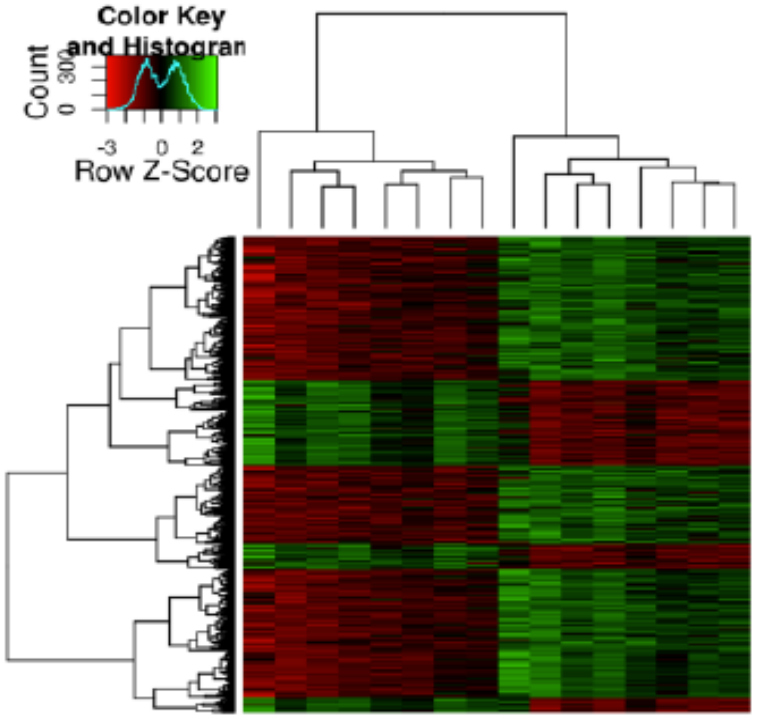
A

software	Expressed Genes	DE genes	Up-regulated genes	Down-regulated genes
Cuffdiff	40435	1404	901	503
Htseq-DESeq	45250	1055	789	266
Overlap	39425	609	448	161

B



C



[5] Some comments on the current status of transcriptome research

Quiz: How much of the human genome is transcribed? (how large fraction of the 3 billion bases).

Approximately…

a. 0.5%
b. 2%
c. 55%
d. 75%

How much of the human genome is transcribed?

ENCODE project: Nature, 2007, 2012 (www.nature.com/encode/#/threads)
Aim is to find the function of each base in the human genome

- The human genome is pervasively transcribed, such that the majority of its bases are associated with at least one primary transcript and many transcripts link distal regions to established protein-coding loci.
- Many novel non-protein-coding transcripts have been identified, with many of these overlapping protein-coding loci and others located in regions of the genome previously thought to be transcriptionally silent.
- Numerous previously unrecognized transcription start sites have been identified, many of which show chromatin structure and sequence-specific protein-binding properties similar to well-understood promoters.

⇒ 74.7% of bases are represented in primary transcripts
⇒ 62.1% of bases represented in processed transcripts

From "Landscape of transcription in human cells" (ENCODE companion paper *Nature* **489**:101-108 (**2012**))

Our genome-wide compilation of subcellular localized and product-precursor-related RNAs serves as a public resource and reveals new and detailed facets of the RNA landscape.

- Cumulatively, we observed a total of 62.1% and 74.7% of the human genome to be covered by either processed or primary transcripts, respectively, with no cell line showing more than 56.7% of the union of the expressed transcriptomes across all cell lines. The consequent reduction in the length of 'intergenic regions' leads to a significant overlapping of neighbouring gene regions and prompts a redefinition of a gene.

- Isoform expression by a gene does not follow a minimalistic expression strategy, resulting in a tendency for genes to express many isoforms simultaneously, with a plateau at about 10–12 expressed isoforms per gene per cell line.

- Cell-type-specific enhancers are promoters that are differentiable from other regulatory regions by the presence of novel RNA transcripts, chromatin marks and DNase I hypersensitive sites.

- Coding and non-coding transcripts are predominantly localized in the cytosol and nucleus, respectively, with a range of expression spanning six orders of magnitude for polyadenylated RNAs, and five orders of magnitude for non-polyadenylated RNAs.

- Approximately 6% of all annotated coding and non-coding transcripts overlap with small RNAs and are probably precursors to these small RNAs. The subcellular localization of both annotated and unannotated short RNAs is highly specific.

Some current topics in RNA-seq bioinformatics research

Allele-specific expression

Variant calling from RNA-seq data

Phasing of RNA-seq data

Evaluation of assembled transcriptomes (many false transcripts)

Normalization procedures

How to assign multireads to genes

Validation of results

What is a gene – an updated definition

1. A gene is a genomic sequence (DNA or RNA) directly encoding functional product molecules, either RNA or protein.

2. In the case that there are several functional products sharing overlapping regions, one takes the union of all overlapping genomic sequences coding for them.

3. This union must be coherent—i.e., done separately for final protein and RNA products—but does not require that all products necessarily share a common subsequence.

=> The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products.

What is a gene – an updated definition

Gerstein et al Genome Res 2007

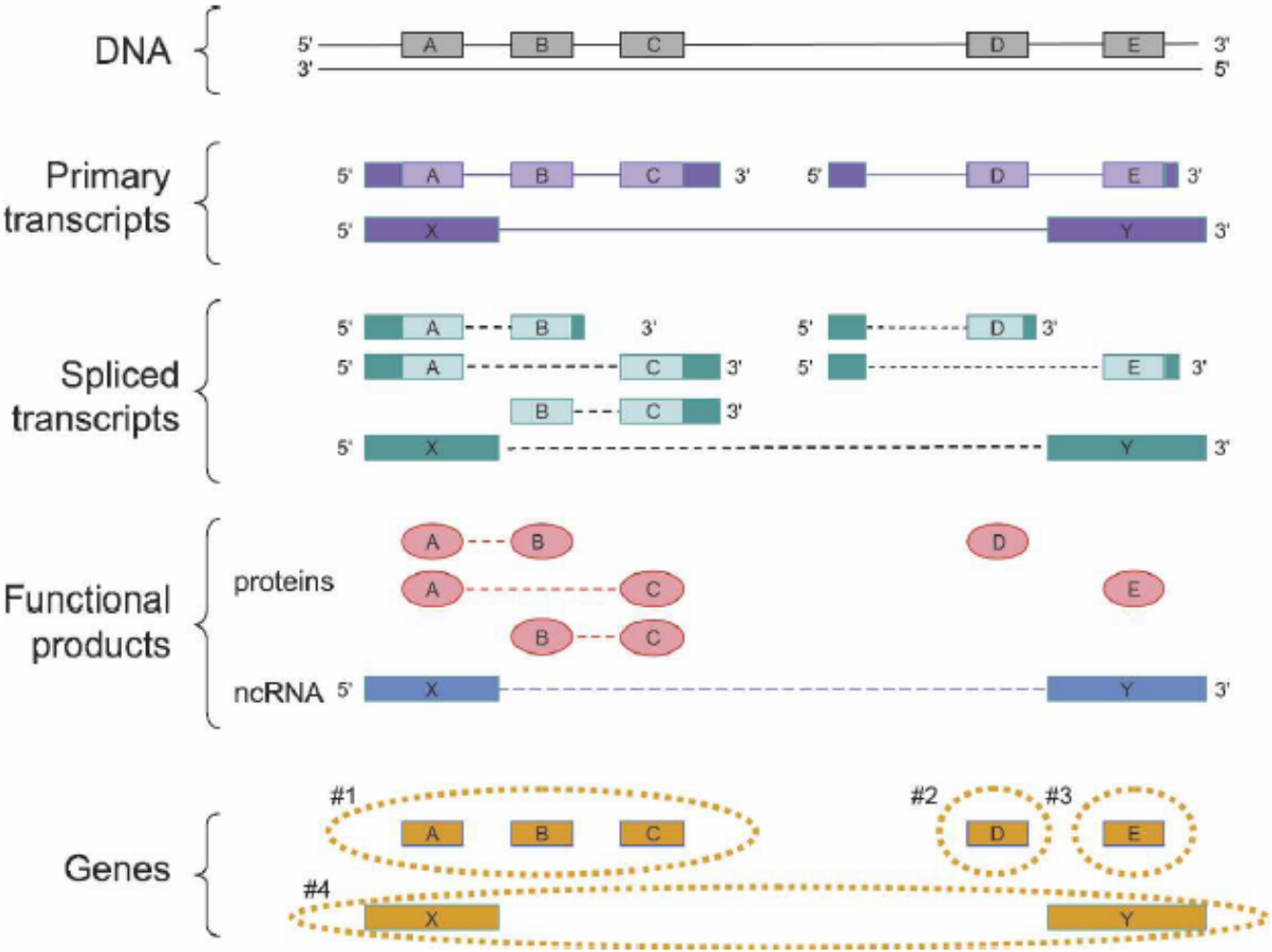


Figure 5. How the proposed definition of the gene can be applied to a sample case. A genomic region produces three primary transcripts. After alternative splicing, products of two of these encode five protein products, while the third encodes for a noncoding RNA (ncRNA) product. The protein products are encoded by three clusters of DNA sequence segments (A, B, and C; D; and E). In the case of the three-segment cluster (A, B, C), each DNA sequence segment is shared by at least two of the products. Two primary transcripts share a 5′ untranslated region, but their translated regions D and E do not overlap. There is also one noncoding RNA product, and because its sequence is of RNA, not protein, the fact that it shares its genomic sequences (X and Y) with the protein-coding genomic segments A and E does not make it a co-product of these protein-coding genes. In summary, there are four genes in this region, and they are the sets of sequences shown inside the orange dashed lines: Gene 1 consists of the sequence segments A, B, and C; gene 2 consists of D; gene 3 of E; and gene 4 of X and Y. In the diagram, for clarity, the exonic and protein sequences A–E have been lined up vertically, so the dashed lines for the spliced transcripts and functional products indicate connectivity between the proteins sequences (ovals) and RNA sequences (boxes). (Solid boxes on transcripts) Untranslated sequences, (open boxes) translated sequences.

Concluding task

Write down your reflections from the RNA-seq lecture on:

1. Something that you found interesting and/or fun.
2. Something that you found hard to grasp.
3. Something that you think this lecture should cover better (either something that wasn't covered at all, or something that you'd like to be covered in more detail).

Format: one sentence per question.

Time: 3 minutes.

Hand in your paper to me when you leave the room today.

Please write your name on it!