

KTH Royal Institute of Technology
School of Biotechnology

Analysis of data from high-throughput molecular
biology experiments BB2490/DD2396

Transcriptomics: RNA-seq

Olof Emanuelsson
olofem@kth.se

Lecture 6 2015-02-03 13:15-15:00 FB55

Enabler for Life Sciences



SciLifeLab

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.*

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: 3 minutes, including taking notes.

Format: write an individual summary. 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 thing: 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 Smith-Waterman extension (SHRiMP) ⁴¹	Smith-Waterman extension	Aligning reads to a reference transcriptome	Reads and reference transcriptome
	Burrows-Wheeler transform methods	Stampy ³⁹ Bowtie ⁴³ BWA ⁴⁴	Probabilistic model Incorporates quality scores		
Spliced aligners	Exon-first methods	MapSplice ⁵² SpliceMap ⁵⁰ TopHat ⁵¹	Works with multiple unspliced aligners Uses Bowtie alignments	Aligning reads to a reference genome. Allows for the identification of novel splice junctions	Reads and reference genome
	Seed-extend methods	GSNAP ⁵³	Can use SNP databases		
		QPALMA ⁵⁴	Smith-Waterman for large gaps		
	Transcriptome reconstruction		STAR		
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 Cufflinks ²⁹	Reports all isoforms Reports a minimal set of isoforms		
	Genome-independent assembly	Velvet ⁶¹ TransABySS ⁵⁶	Reports all isoforms	Identifying novel genes and transcript isoforms without a known reference genome	Reads
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			
Differential expression					
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. Each transcript is typically represented by many identical mRNA 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 or all time points)

Transcriptomics – finding out everything about the *transcriptome*

Why study gene expression?

Historically the main goal of gene expression studies has been to find differential expression.

I.e. genes that are more – or less – expressed in a certain sample compared to another.

Upregulated and *downregulated* genes

- Healthy vascular tissue vs. diseased vascular tissue
- Stressed vs. not stressed flowers
- Female brain vs. male brain
- etc.

This could provide **clues to the biological function of a cell/tissue/organism**. Which could help in curing disease, optimizing crop yield, etc.

The history of gene expression analysis

EST – expressed sequence tag

Other tag-based – e.g.

CAGE (Cap analysis of gene expression; 5', 20 nt)

MPSS – massively parallel signature sequencing (16-20 nt)

Microarray – cDNA or oligo arrays; up to 20 million features.

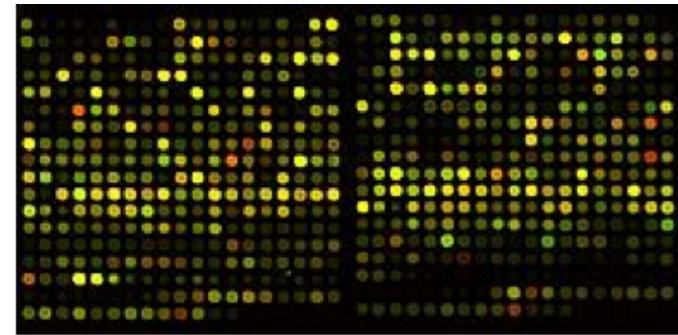
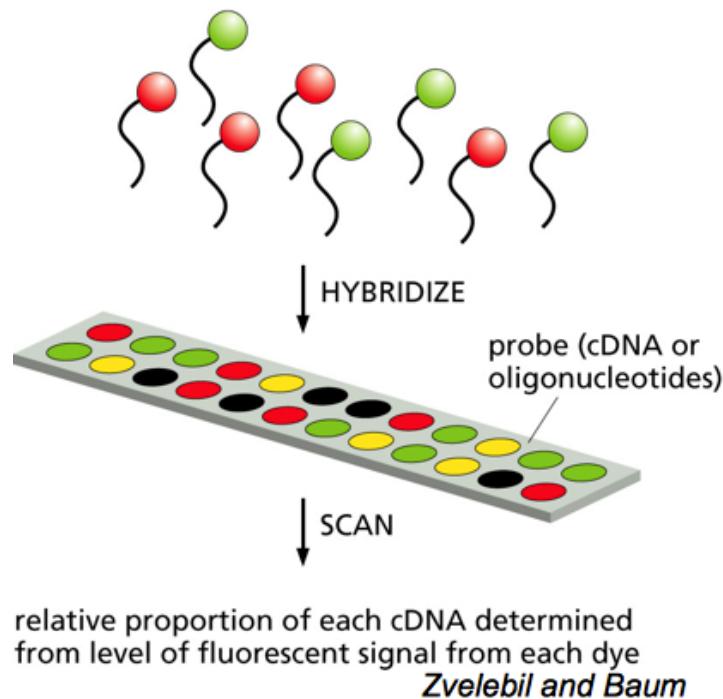
Tiling microarray – covering the entire non-repetitive part of a genome (not only genes).

RNA-seq – current state of the art

Microarrays

Microarrays (2-colour array pictured)

cDNA from sample A labeled with Cy5, + cDNA from sample B labeled with Cy3, gives rise to different colors on chip



An image of a two-colour microarray

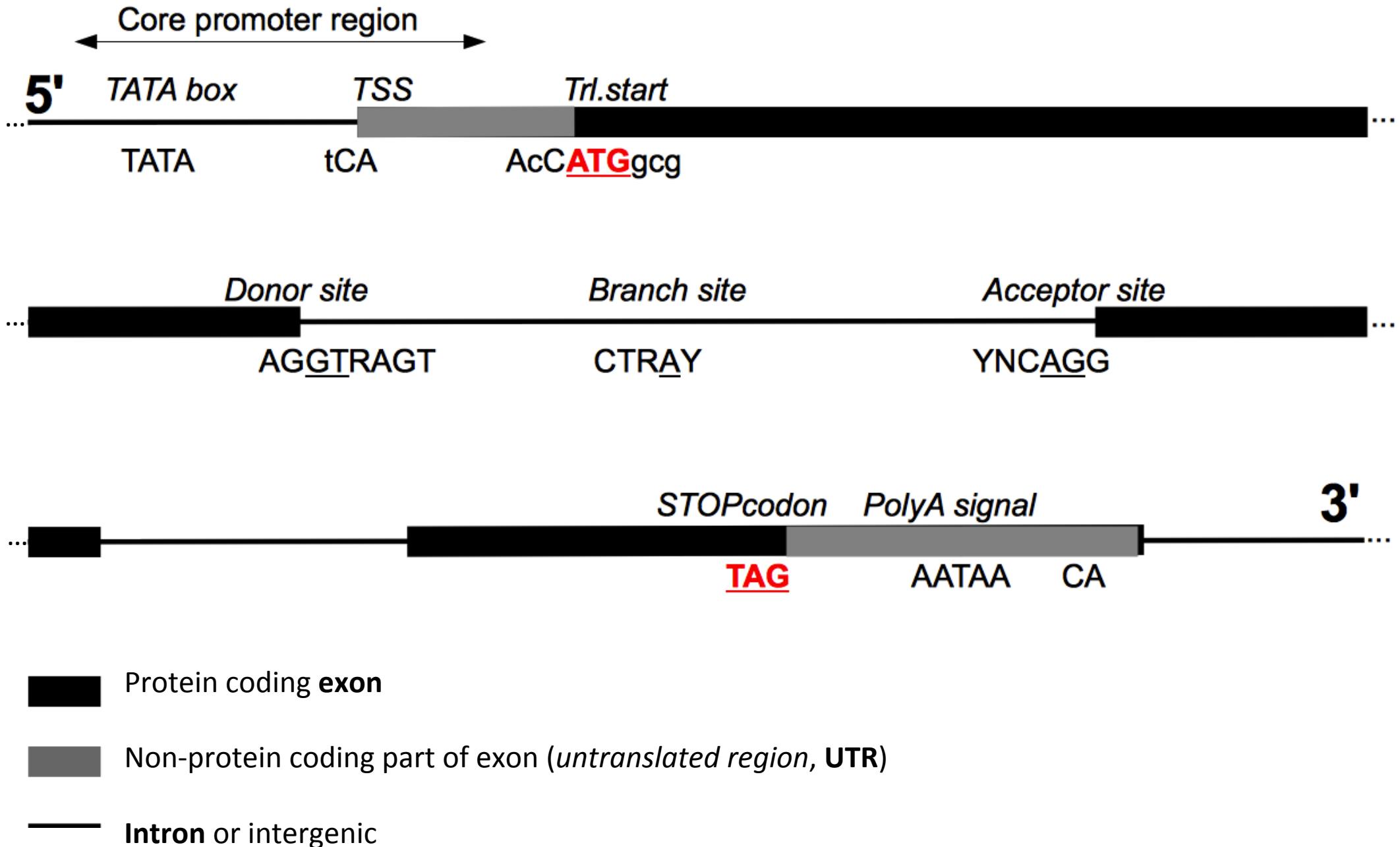
The probes represent genomic DNA sequences

Pros: inexpensive, “everybody” can do it, high throughput

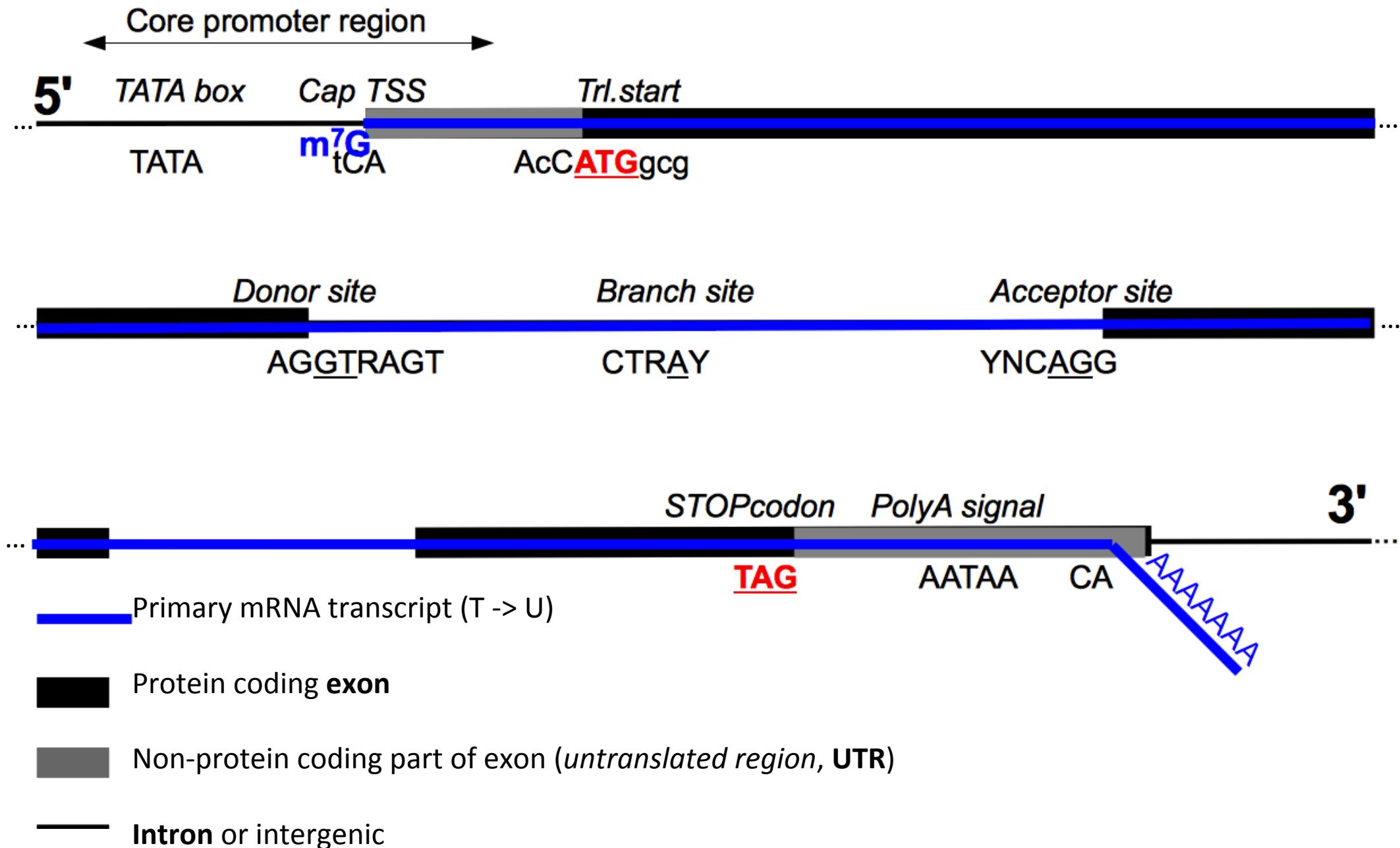
Cons: background noise, artefacts, limited dynamic range

(Tiling microarrays: oligos on the chip represent the entire non-repetitive part of the genome (“traditional” microarrays only cover the known genes)).

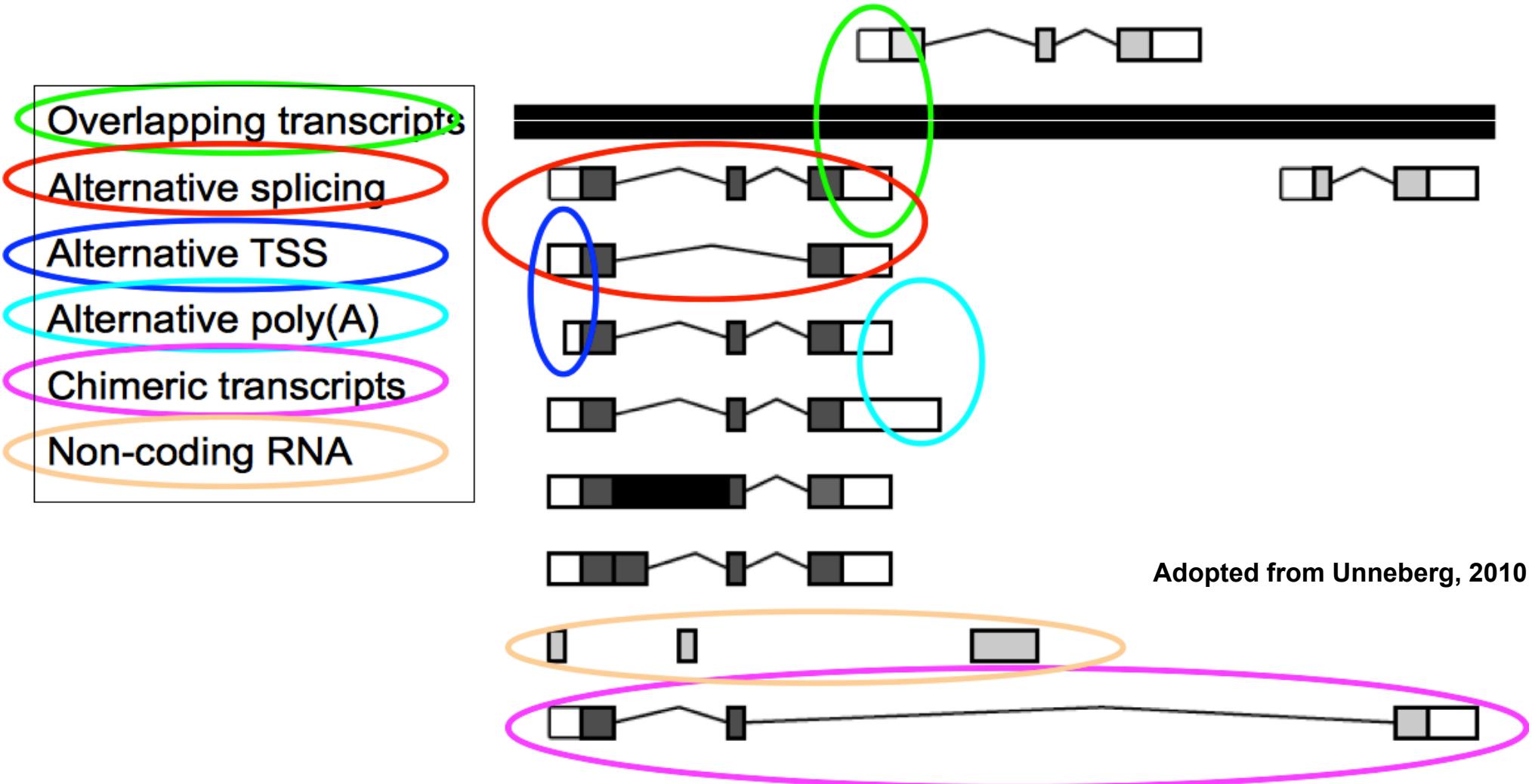
A eukaryotic protein coding gene



A eukaryotic protein coding gene



Transcriptome complexity



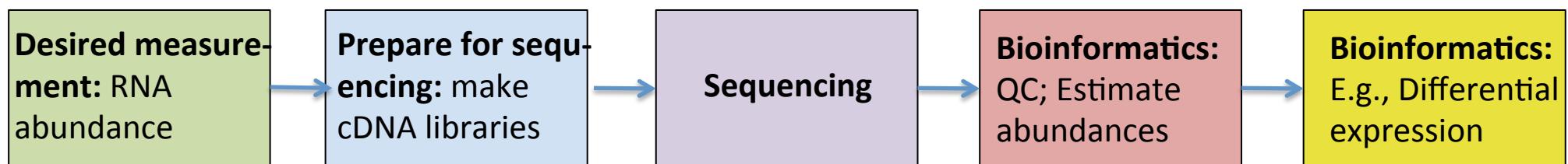
One **gene** can produce many possible **transcripts**.

[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
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

allele-specific expression

non-coding RNAs

RNA-seq

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. Selection/enrichment
4. Random priming [or poly(A) priming then fragmentation]
5. cDNA generation (reverse transcription)
6. Adapter ligation, cluster generation

=> RNAs converted to DNA and ready for sequencing

RNA-seq

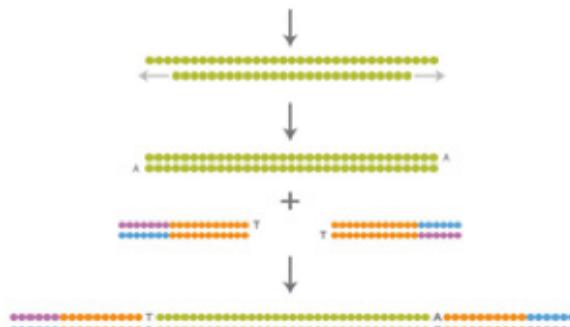
Several different technologies available for sequencing.

Sequencing chemistry, amplification strategy, read length, number of reads, base calling accuracy, sequencing errors – rates and types, ...

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

Illumina HiSeq

The most popular choice for RNA-sequencing
Read length up to 2x150 bases (MiSeq: 2x300)



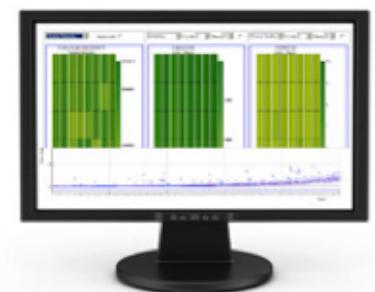
Library Preparation
<6 h (<3 h hands-on)



Cluster Generation
<4 h (<10 min hands-on)



Sequencing by Synthesis
1.5-8 days (<10 min hands-on)



RTA v1.7, CASAVA v1.7
2 days (30 min hands-on)

Typically $>10^9$ reads corresponding to >200 Gbases per run

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
GATTGGGGTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTT
+
! ' ' * ( ( ( ***+ ) ) % % % + + ) ( % % % % ) . 1 * * * - + * ' ' ) ) * * 55CCF>>>>>CCCCCCCC65
```

Sequencing output

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%

[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

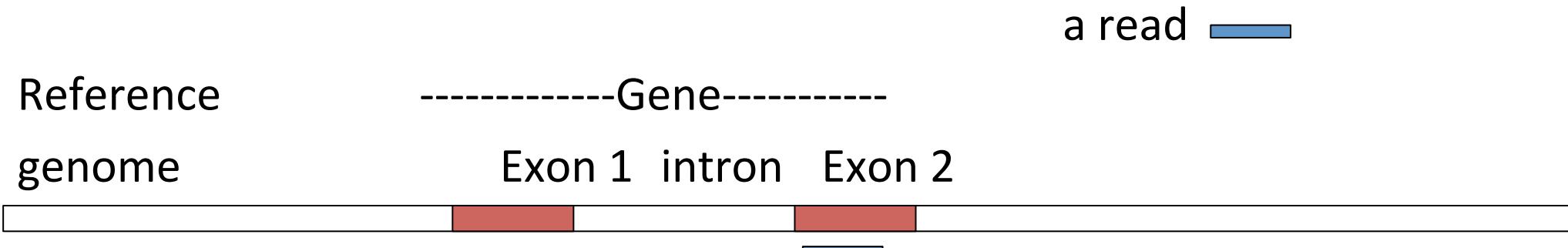
Transcript expression levels

Differential expression analysis

Map reads to a reference

Find the place (=map or align) on the genome (=reference) from which the read originated (=was transcribed).

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



The procedure of mapping the reads to the reference genome is crucial for many sequencing-based approaches.

Map reads to a reference

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.

a read —

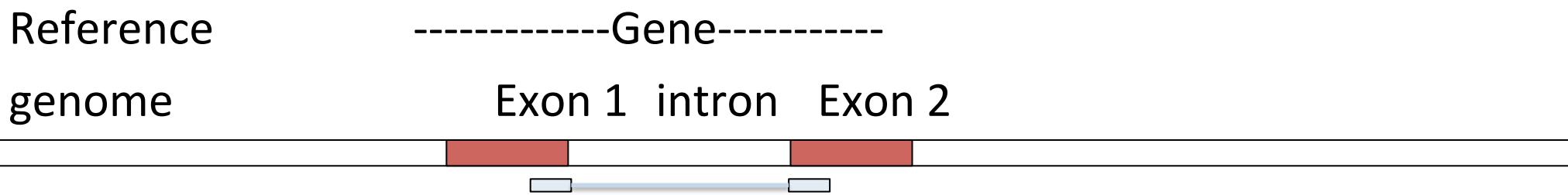


Map reads to a reference

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.

a read —



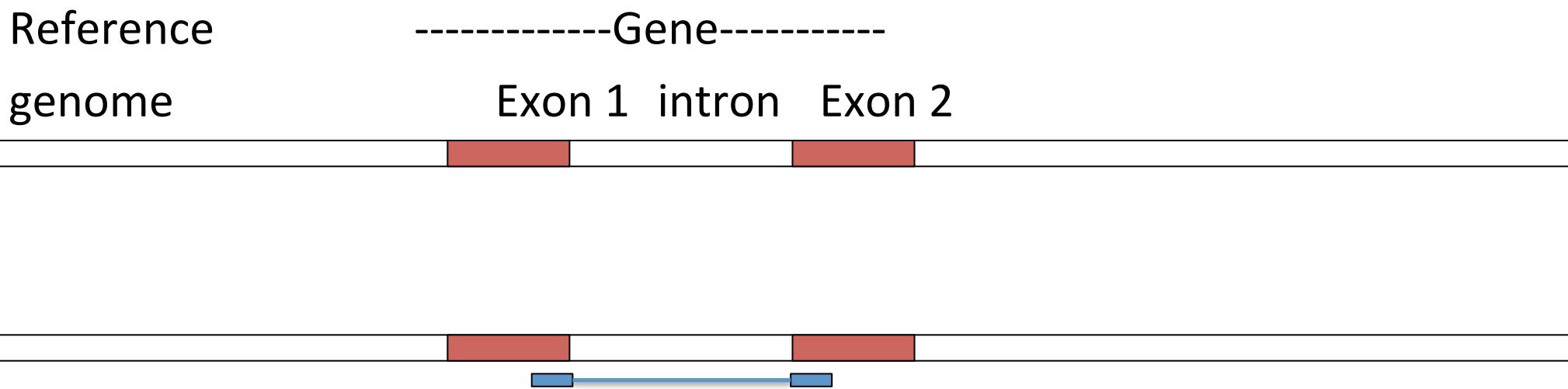
Tools for this: BWA, Bowtie, Maq, SOAP, Gnumap, ...

Map reads to a reference

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

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

a read —



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

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 $N \gg n$.

Transcriptome reconstruction

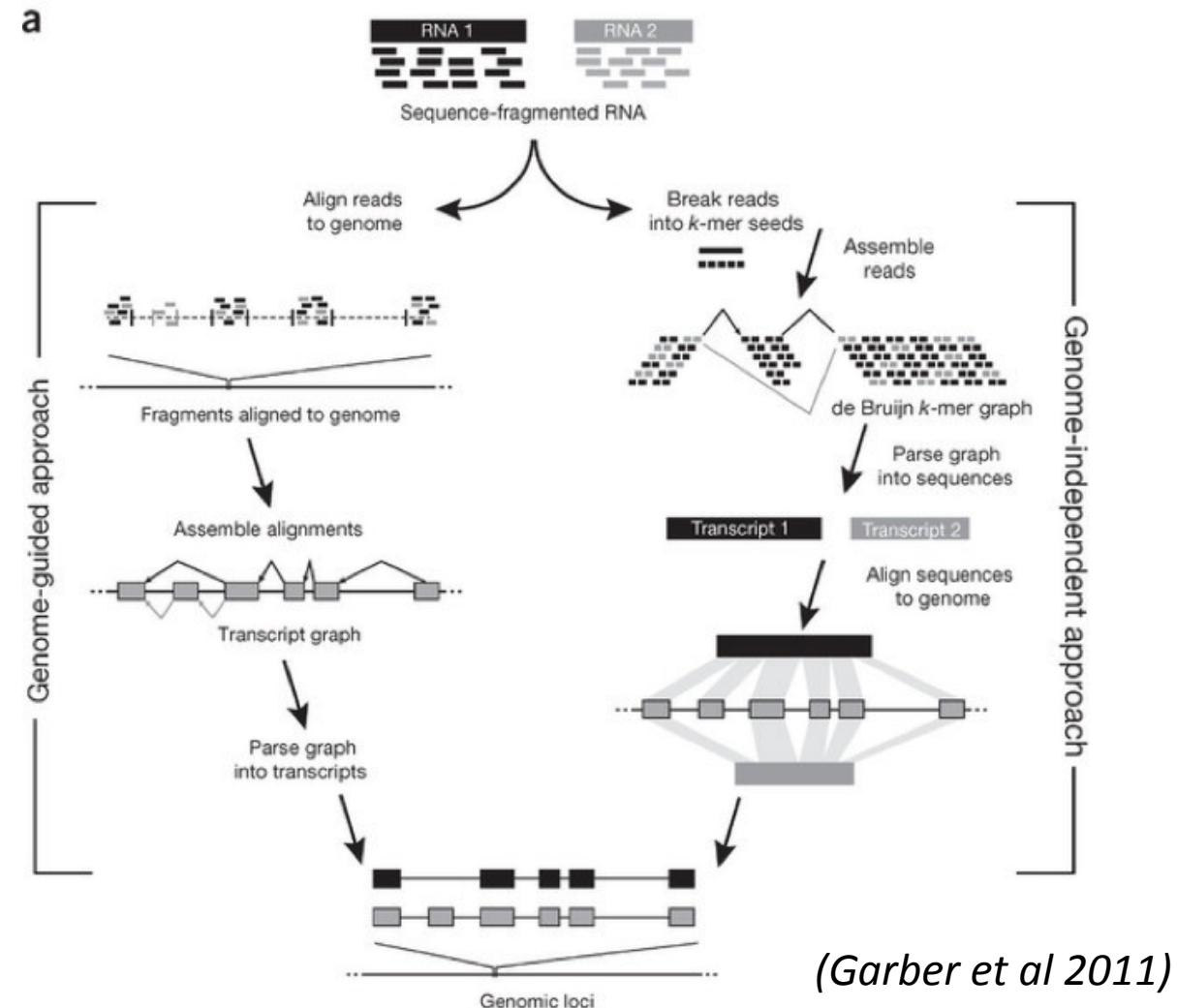
i.e., reconstruct the transcripts present in the sample

Genome-guided approach:

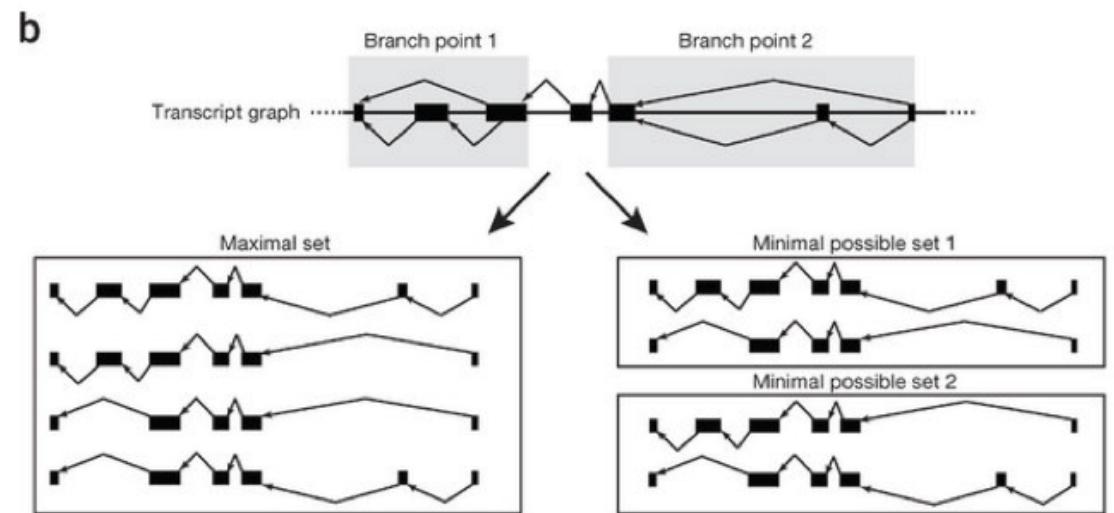
If there is a reference genome
(e.g., Cufflinks)

Genome-independent approach:

The only option if there is no reference genome
(e.g., Trinity)

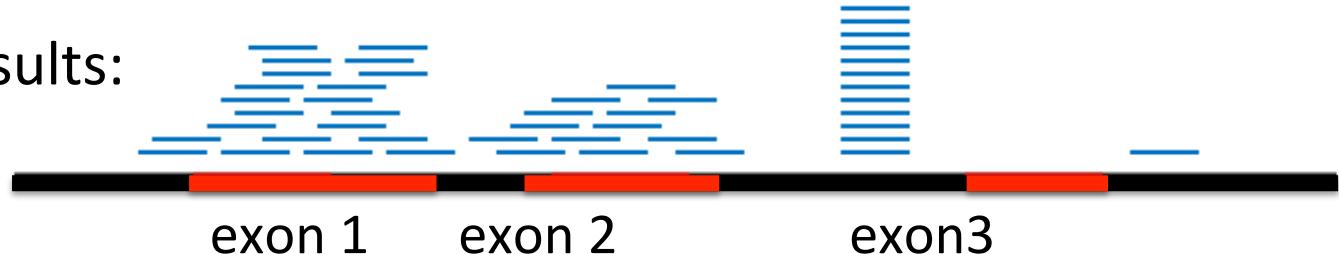


(Garber et al 2011)



Reads mapped to a gene or a transcript

Mapping results:



Ambiguous reads:

one read matching to *many* places in the reference

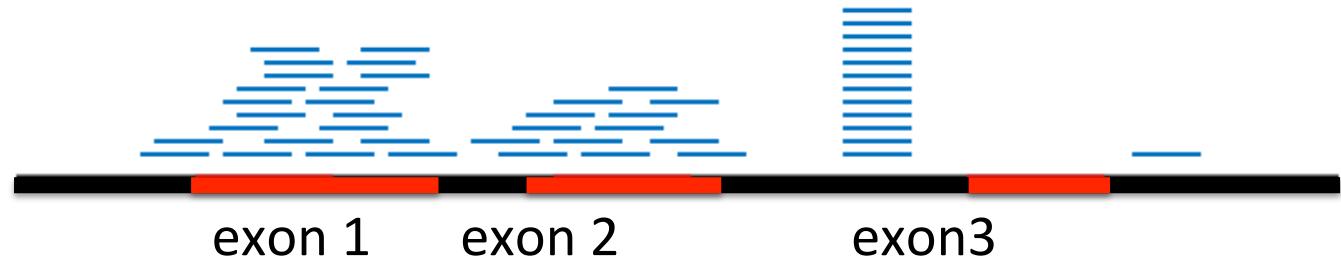
=> these reads are usually discarded as they are ambiguously mapped and hence their origin is unclear

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

=> referred to as “PCR duplicates” and usually discarded

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

Count the reads



Exon 1: 21

Exon 2: 13

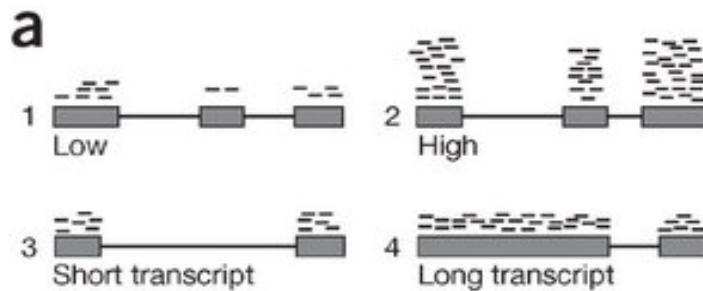
Exon 3: 0

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:



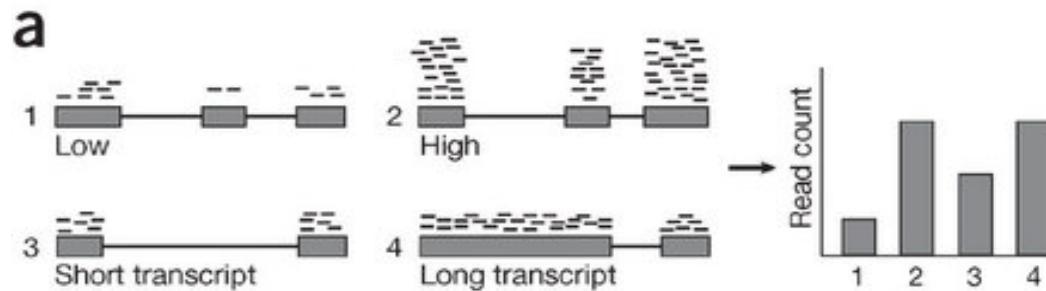
Transcript 3: 12 reads

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?

Estimate abundance

RPKM – reads per kilobase of transcript per million mapped reads

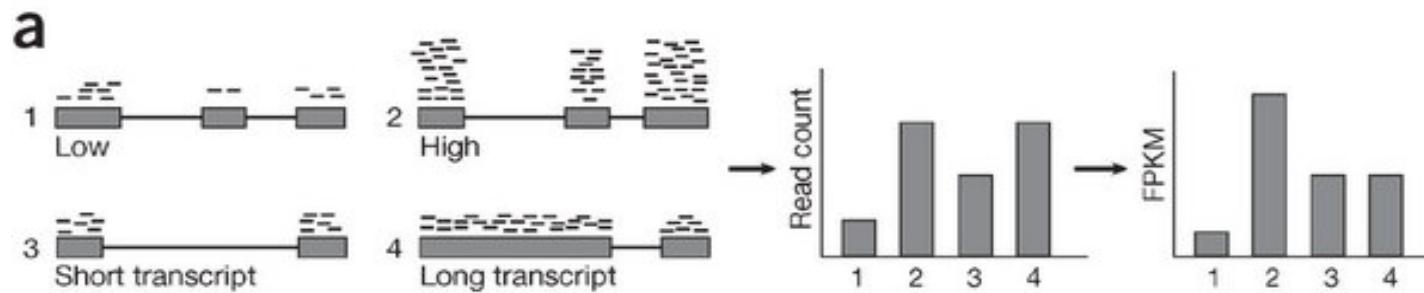
$$R = \frac{10^9 C}{LN}$$

- $R=RPKM$ value (or $FPKM$ for paired end reads)
- $C=$ number of reads mapped to the transcript
- $L=$ gene length
- $N=$ number of mappable reads

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

Estimate abundance

Use *RPKM* instead of read counts:



Transcript 3: 12 reads

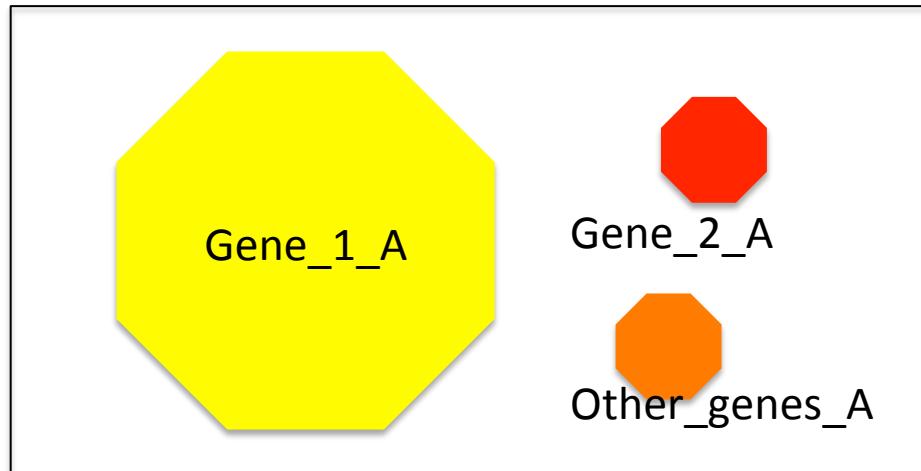
Transcript 4: 31 reads

But their RPKM is the same.

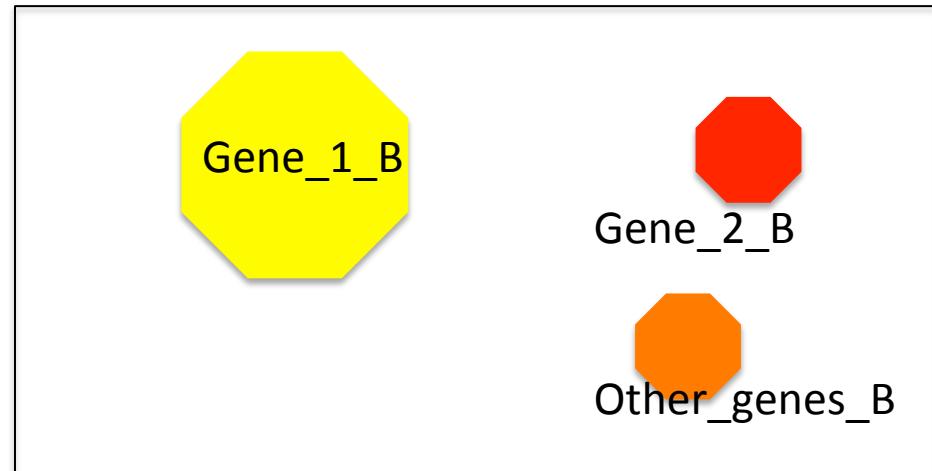
Tools for this: ERANGE, Myrna, eXpress

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.

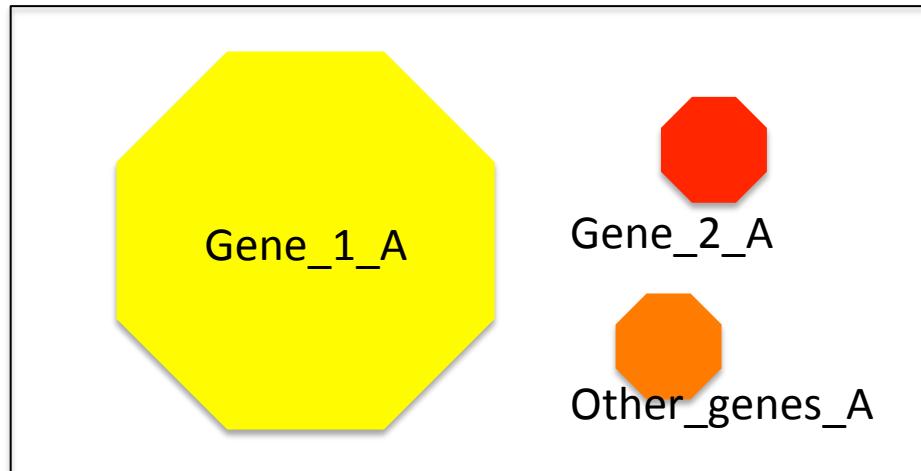
Both samples are sequenced to a depth of $10 * 10^6$ reads.

Which statement is true?

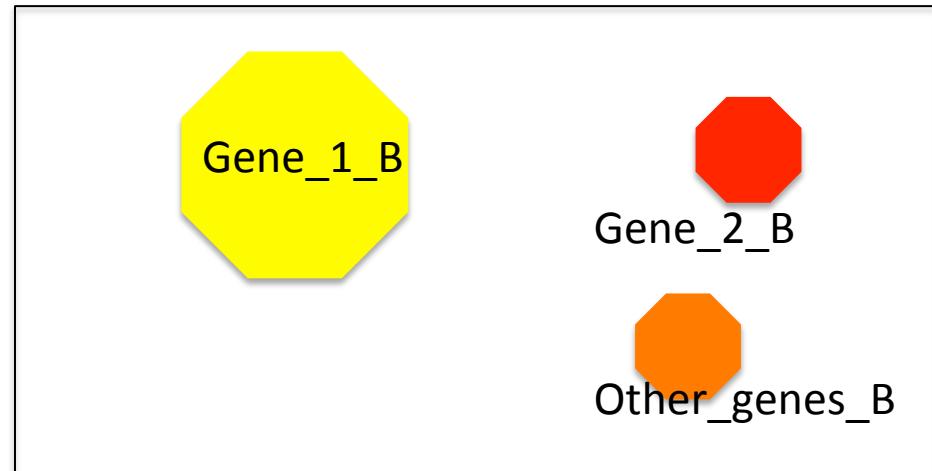
- a. The RPKM for every gene is the same in sample A as in sample B
- b. Gene 2 will have the same RPKM in both samples
- c. Gene 2 will have lower RPKM in sample A than in sample B
- d. Gene 1 will have higher RPKM in sample B than in sample A

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.

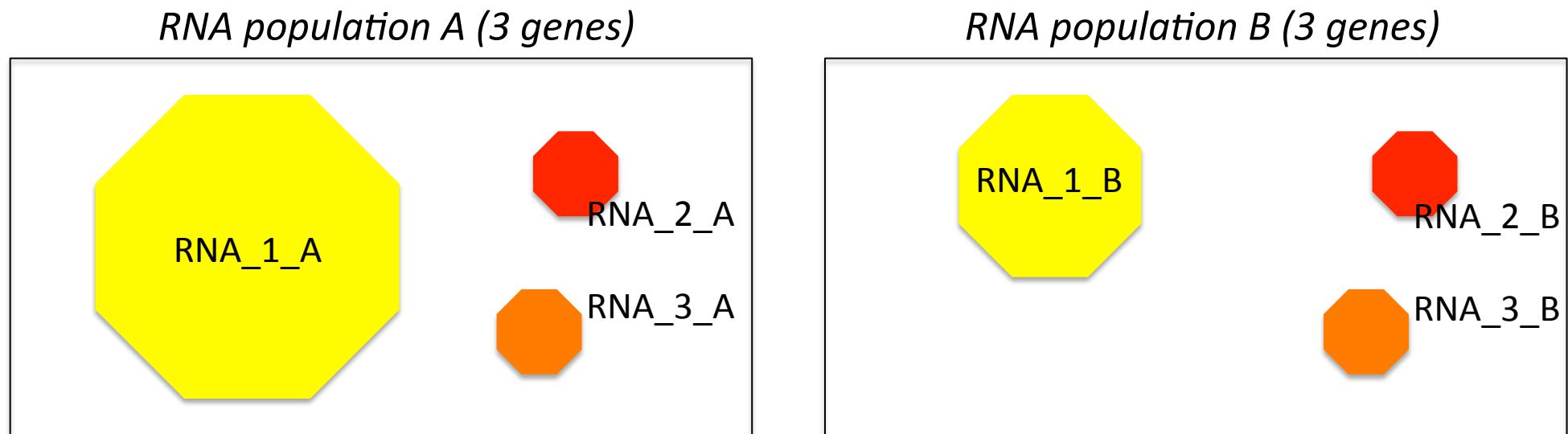
Both samples are sequenced to a depth of $10 * 10^6$ reads.

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 1 will have higher RPKM in sample B than in sample A

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.

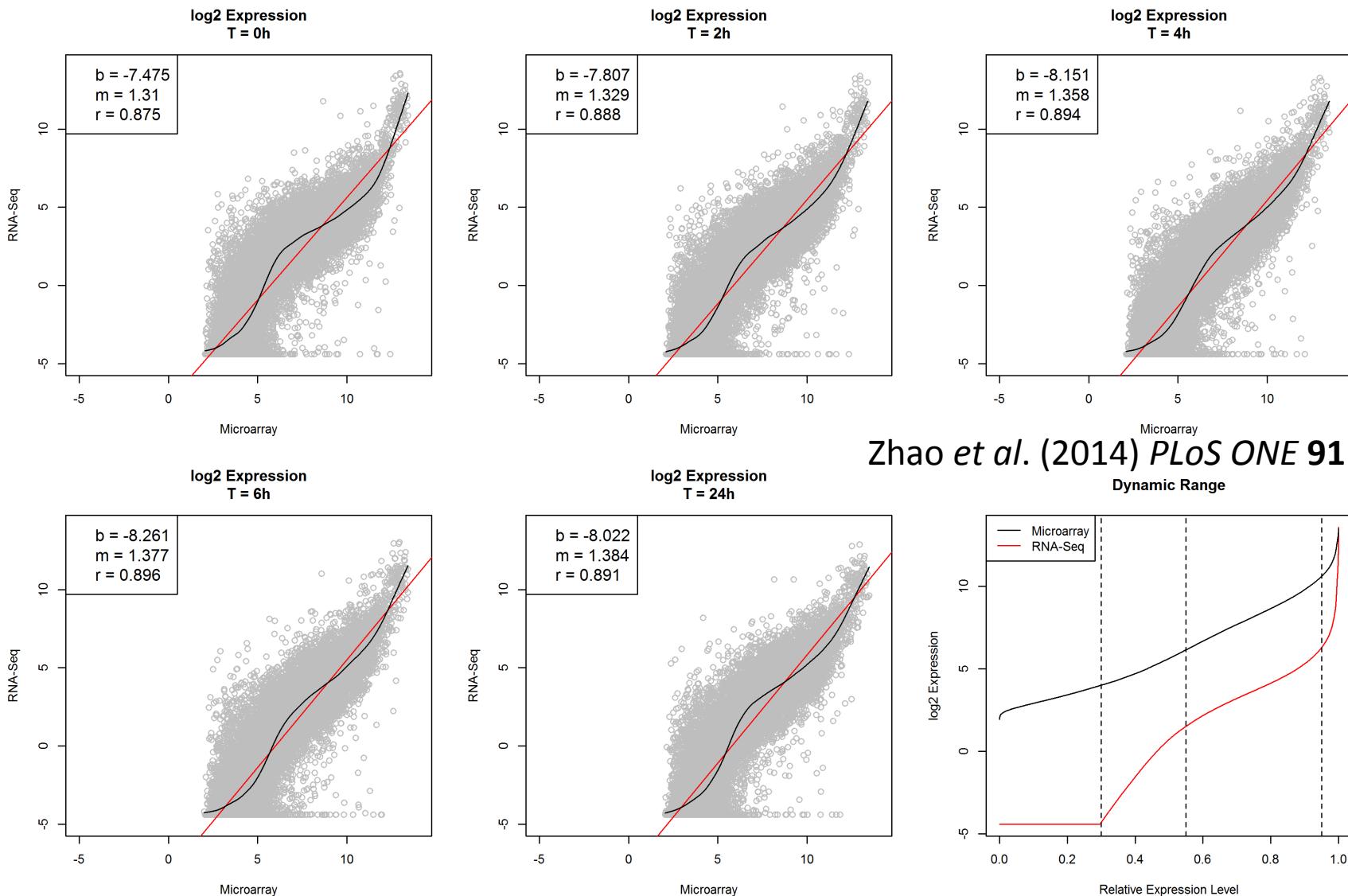
Tools for this: EdgeR, DEseq

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)



Microarray: $\sim 4 \times 10^3$

RNA-seq: $\sim 3 \times 10^5$ (general RNA-seq estimation $\sim 10^6$)

Differential expression (DE)

We now have the *normalized read count* for each gene, or for each transcript

Typically, we now want to examine the *differential expression*:

- We have samples from ≥ 2 conditions

- Find the genes or transcripts that are present in different amounts in these samples

Differential expression (DE)

When is a difference in read count 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

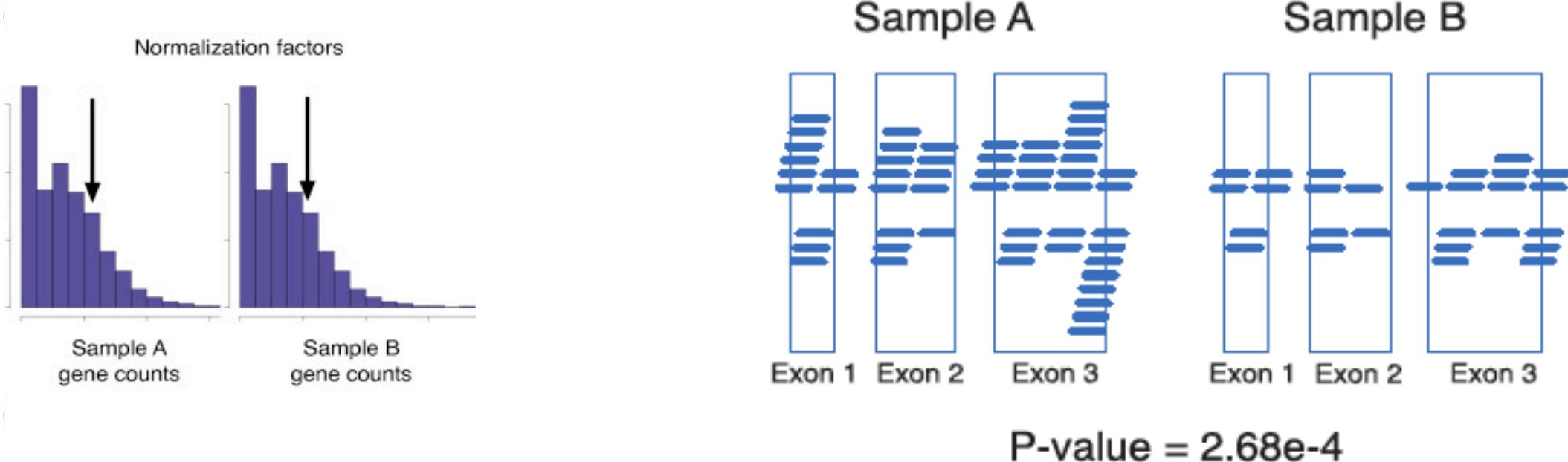
Negative binomial (e.g. edgeR, DEseq, CuffDiff)

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

Output is, for each gene

- ⇒ 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 expression (DE)



Normalized counts or RPKM obtained for each gene in the two samples:

- [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

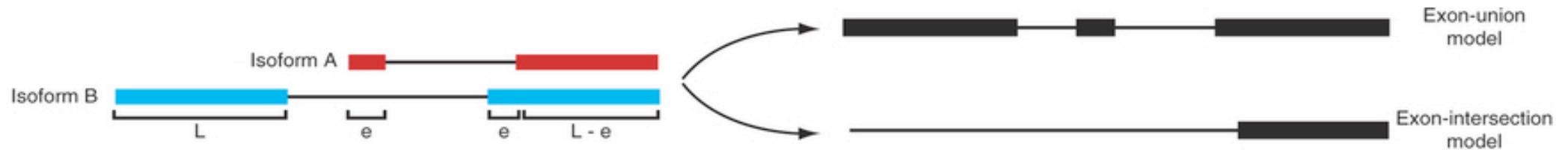
Alternatively (1): instead of explicitly using normalized counts or RPKM, adjust statistical calculation to accommodate the need for normalization

Alternatively (2): use non-parametric methods (i.e. rank), then no need to assume a distribution (skip [1])

Differential expression

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

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



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

Two approaches:

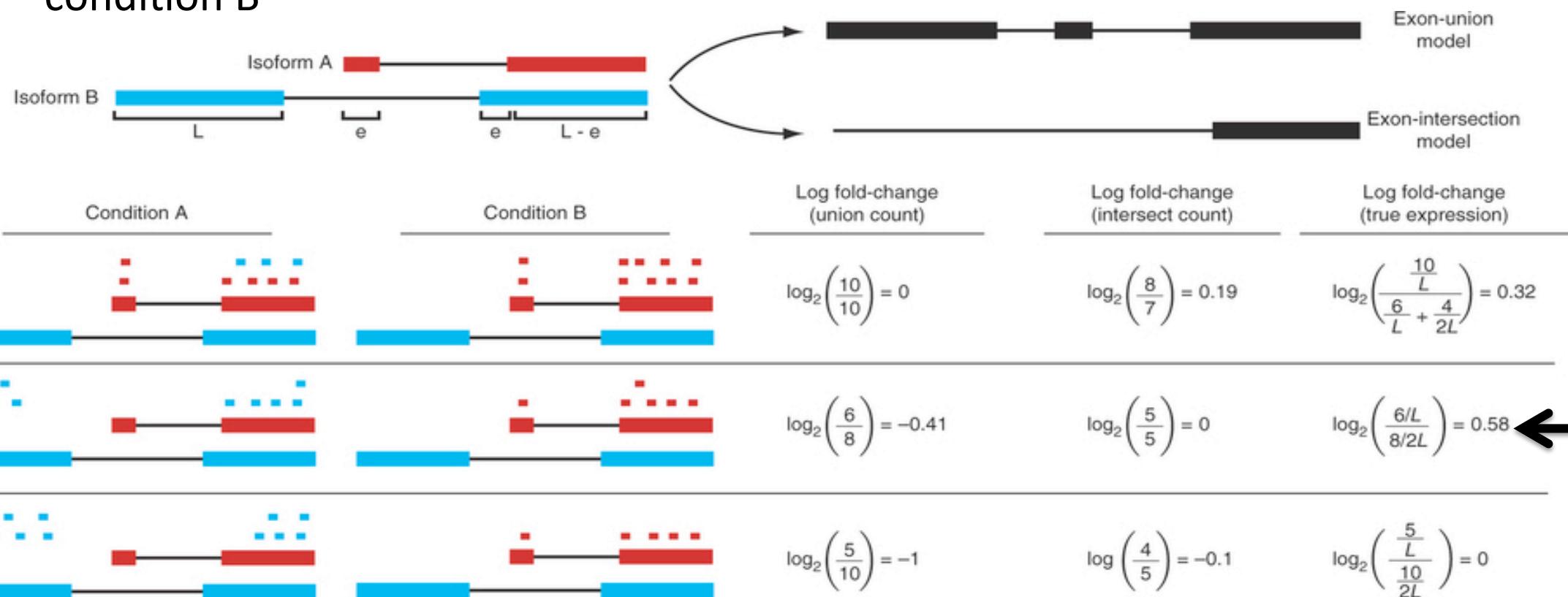
Exon union – use the read counts from all exons in all isoforms

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

Differential expression

These approaches could both provide the wrong answer.

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



Differential expression

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

CuffDiff

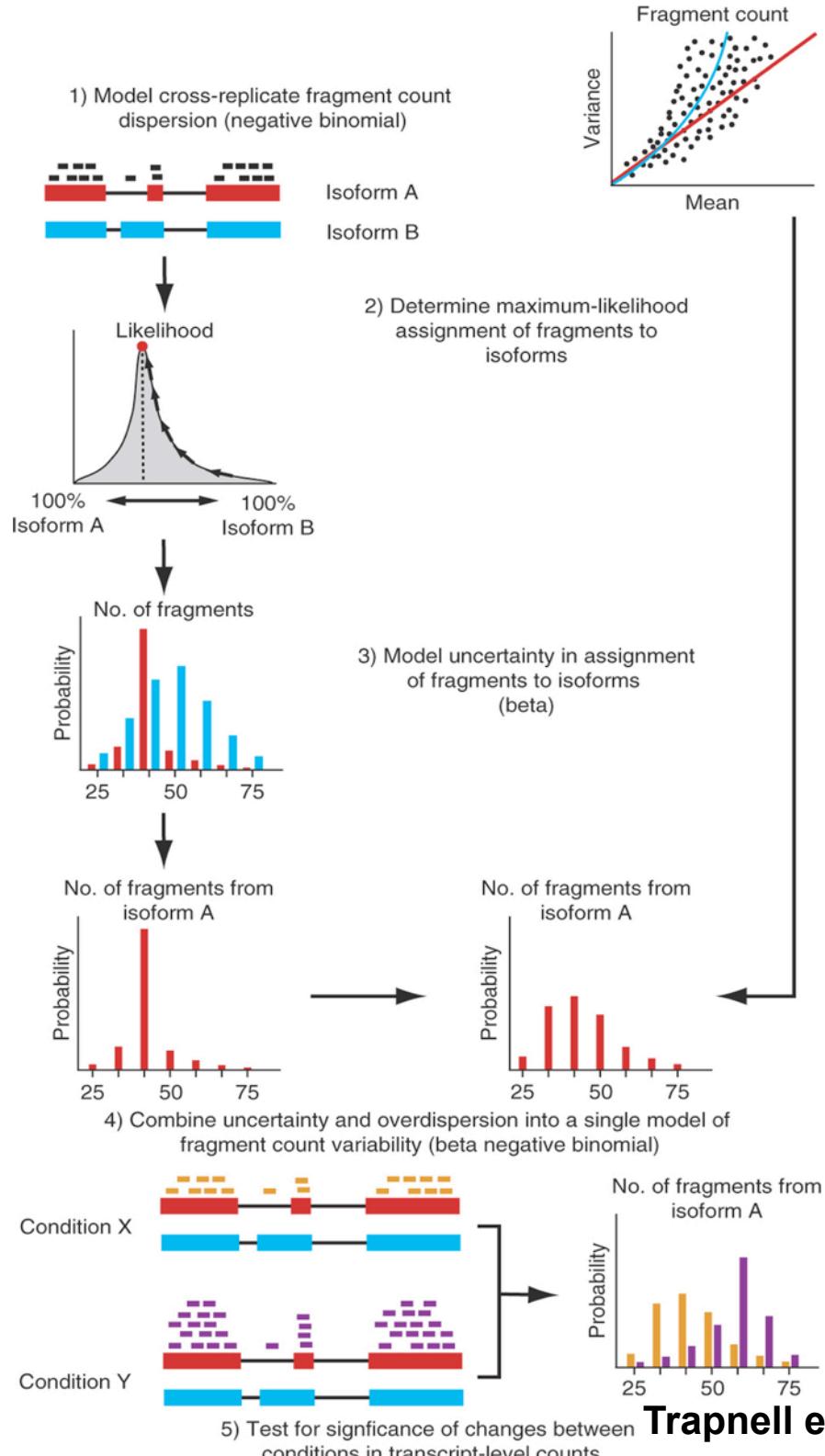
Model overdispersion of data (like previously mentioned tools)

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

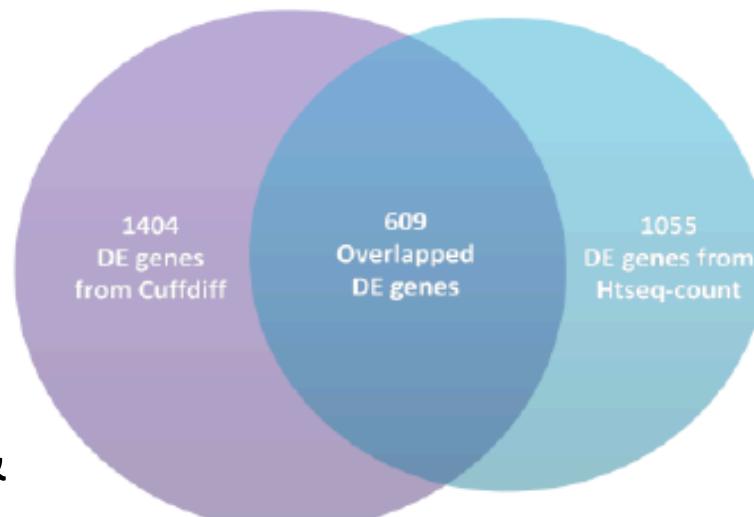
5. These variance estimates are used in the statistical testing of differential expression.

Differential expression

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

Benchmarking studies show that exon union/intersection methods hold up surprisingly well.

B

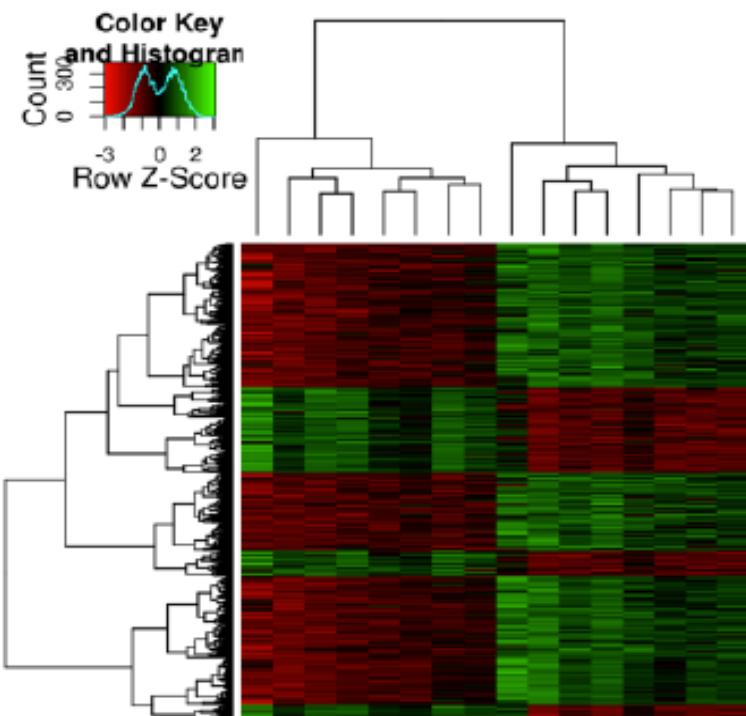


[E.g., Soneson & Delorenzi 2013]

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

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. 70%

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 important 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

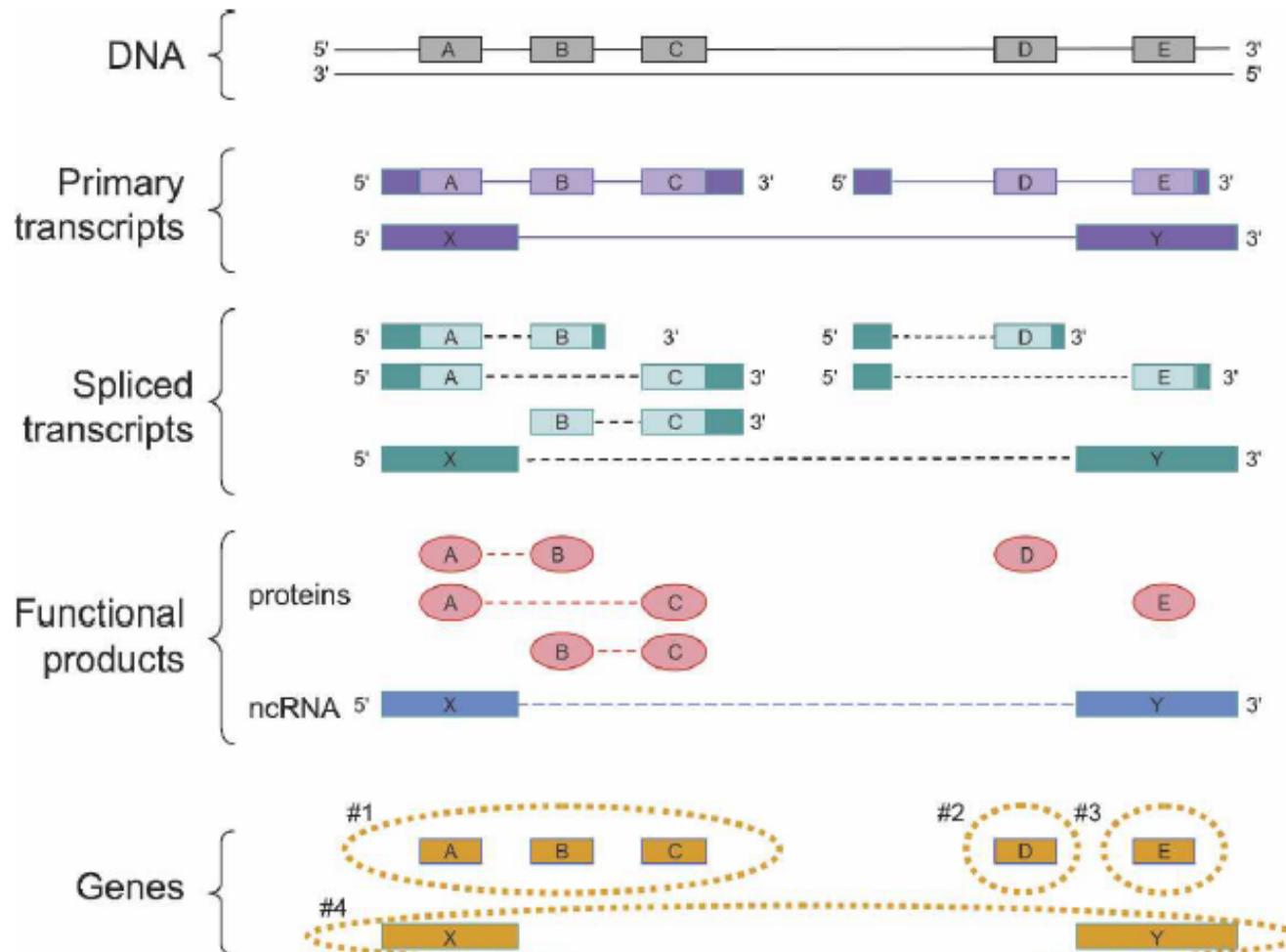
How to assign multireads to genes

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 protein 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!