Analysis of data from high-throughput molecular biology experiments

Lecture 6 (F6, "RNA-seq"), 2012-01-26

What is a gene

What is a transcriptome

History of gene expression assessment

RNA-seq

RNA-seq analysis steps

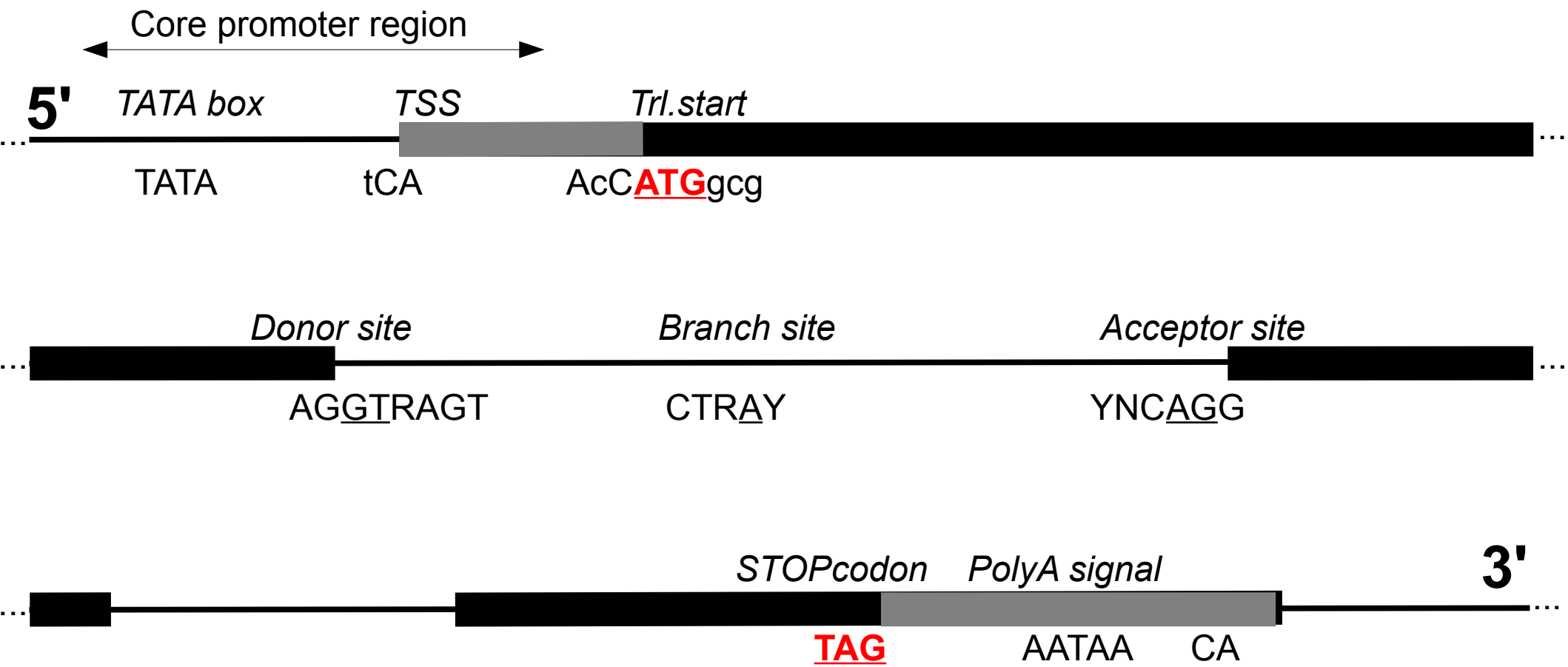
Gene expression studies

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.

A eukaryotic protein coding gene (text book style)

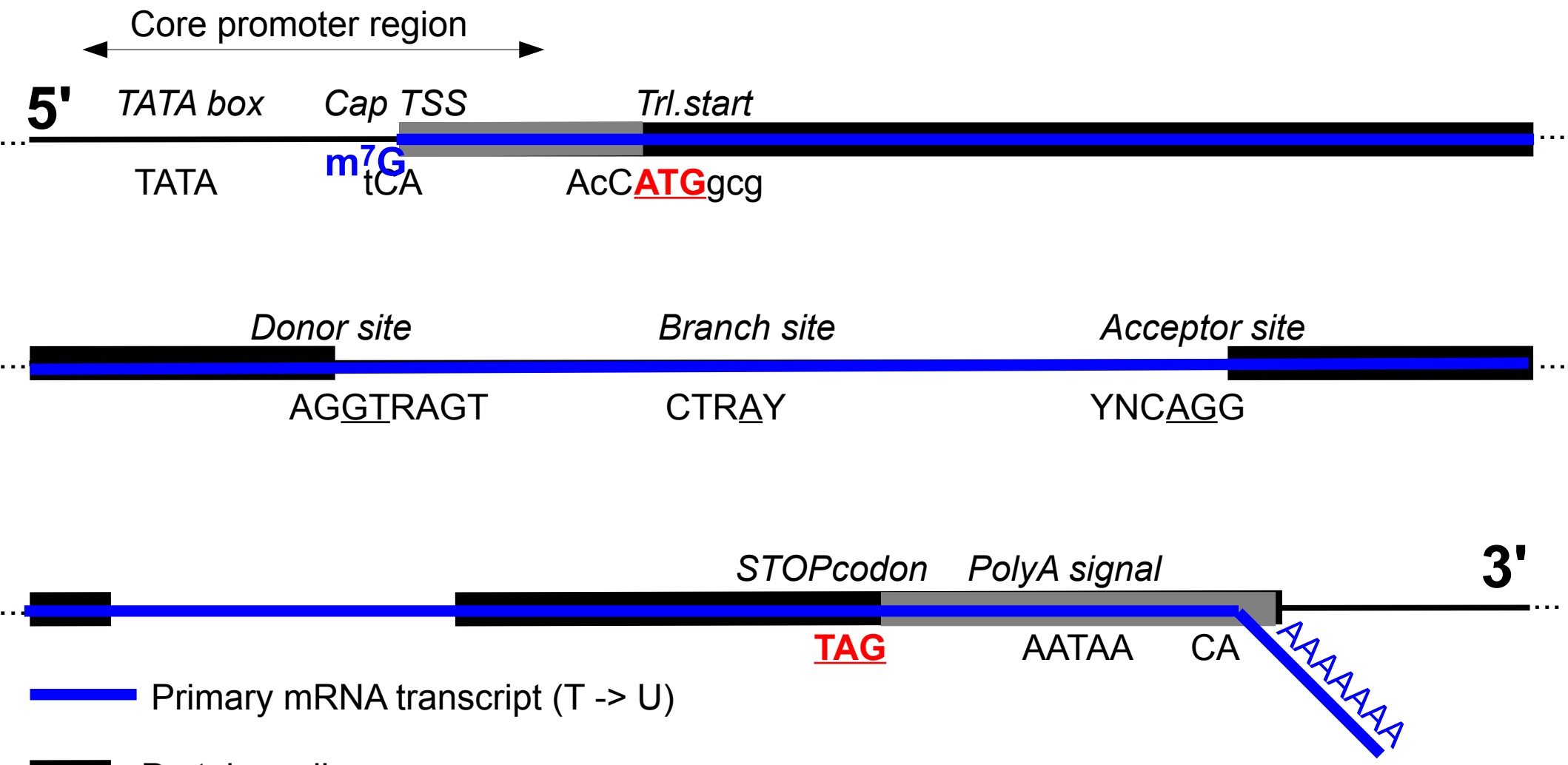


Protein coding **exon**

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

Intron or intergenic

A eukaryotic protein coding gene (text book style)



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 (extracting 16-20 nt from mRNA molecules)

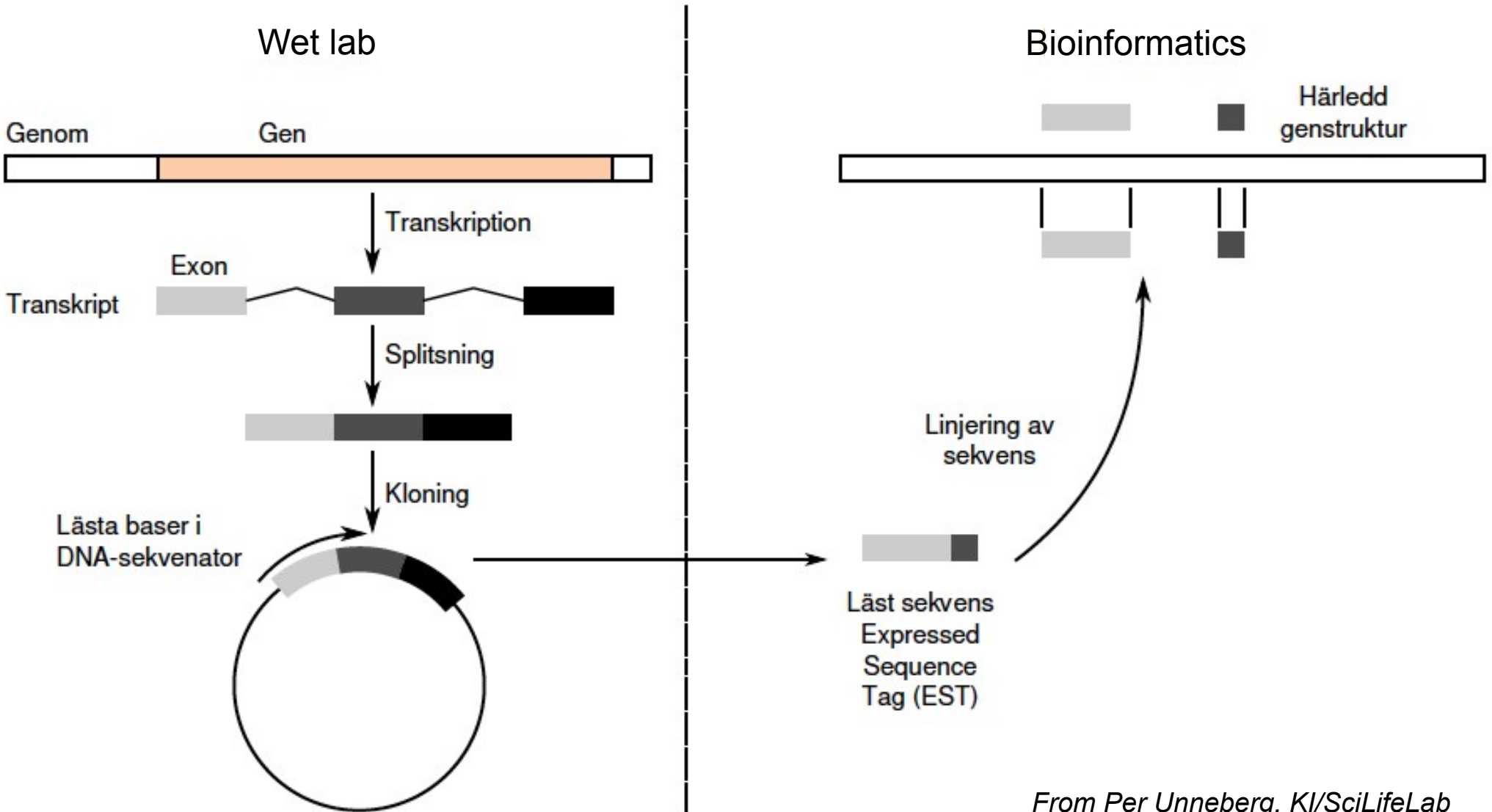
Microarray – cDNA or oligo arrays; interrogating up to 20 million features.

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

RNA-seq – the latest thing in expression analysis

The history of gene expression analysis

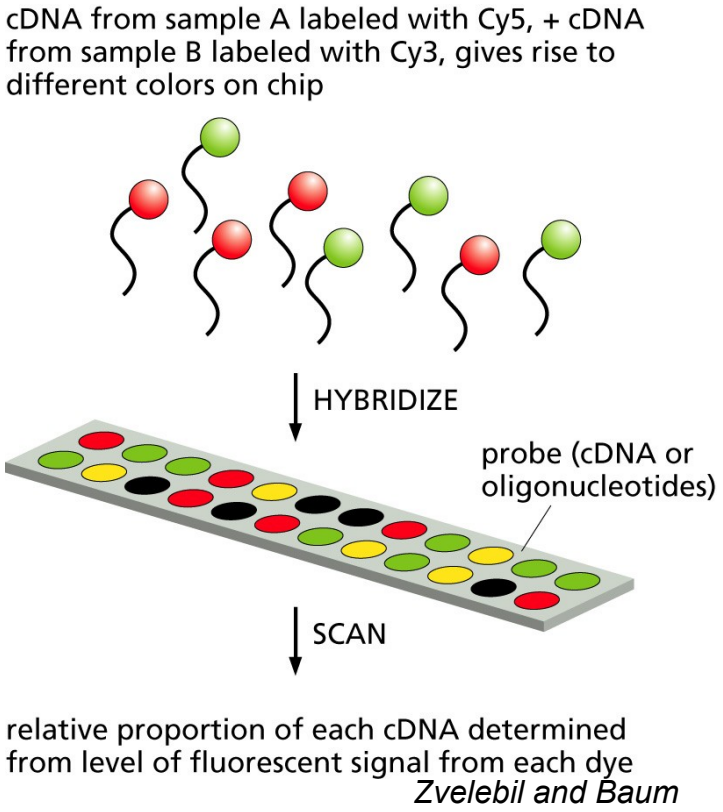
EST – expressed sequence tag (low-throughput, expensive)



From Per Unneberg, KI/SciLifeLab

The history of gene expression analysis

Microarrays (2-colour array pictured)



An image of a two-colour microarray

The probes represent genomic DNA sequences

Pros: inexepensive, “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).

How much of the human genome is transcribed?

ENCODE project:

Nature, 2007

Looking at 1% (30 Mbases)

of 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% of bases are represented in a primary transcript (supported by 2 or more experimental technologies)

=> *pre*-RNA-seq

=> will it hold in court? The conclusions have been contested.

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.

Gerstein et al Genome Res 2007

What is a gene – an updated definition

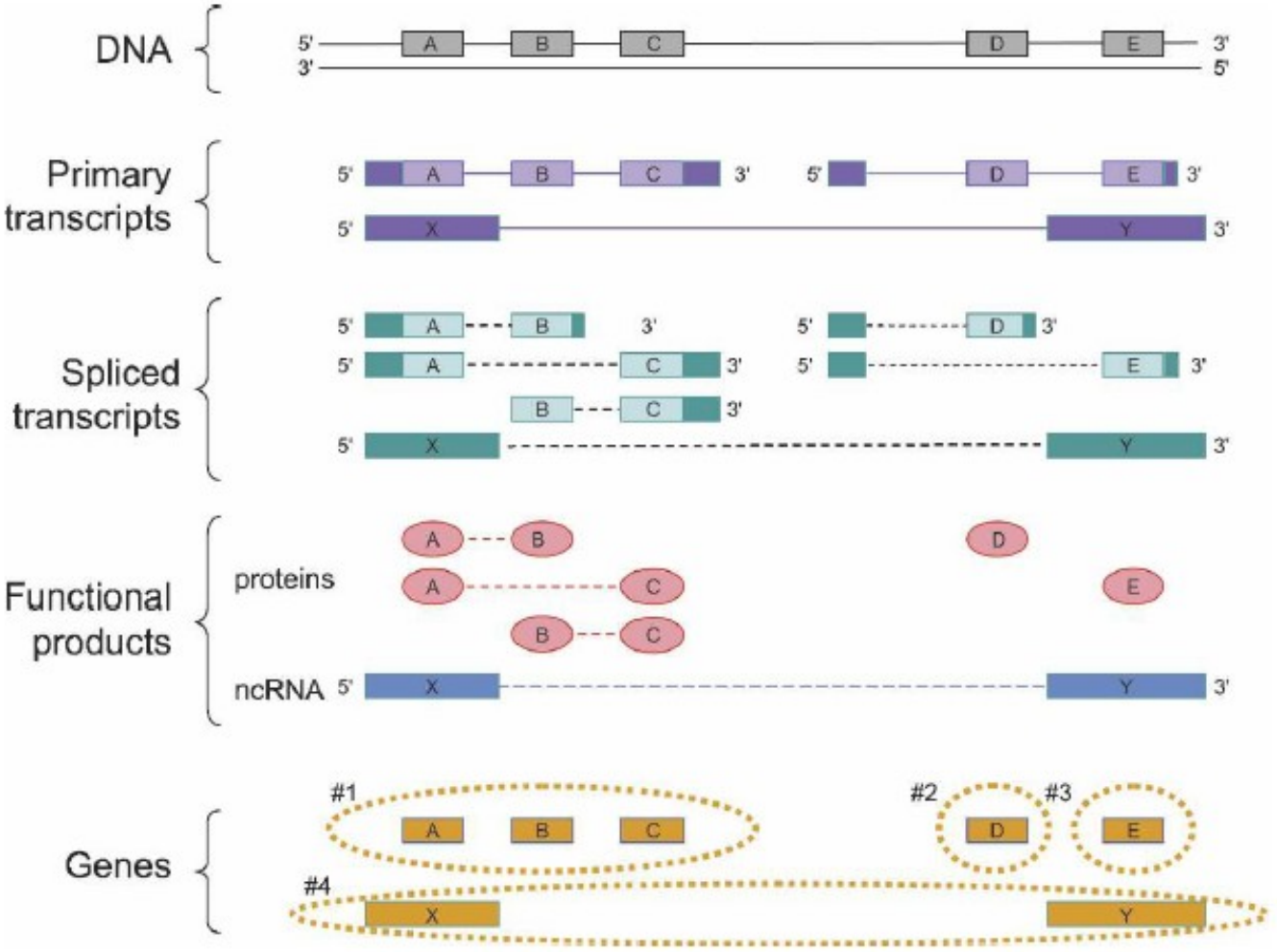


Figure 5. How the proposed definition of the gene can be applied to a sample case. A genomic region produces three primary transcripts. After alternative splicing, products of two of these encode five protein products, while the third encodes for a noncoding RNA (ncRNA) product. The protein products 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.

Gerstein et al Genome Res 2007

What is a transcriptome?

All RNA molecules transcribed in a particular cell/cell line/tissue at a certain developmental stage and under certain conditions.

<u>The transcriptome is different depending on:</u>

0. What organism (of course)
1. Which cell/cell line/tissue we're looking at

    Cell line: in culture, most cell lines derived from cancer cells

    Tissue: from a real, living organism
2. Developmental stage
3. Condition/treatment

NOTE: often, we are interested only in mRNA molecules and remove (in a wet lab procedure, or bioinformatically) *non-coding* RNA (ncRNA) (e.g. rRNA, tRNA).

Transcriptome complexity

Overlapping transcripts

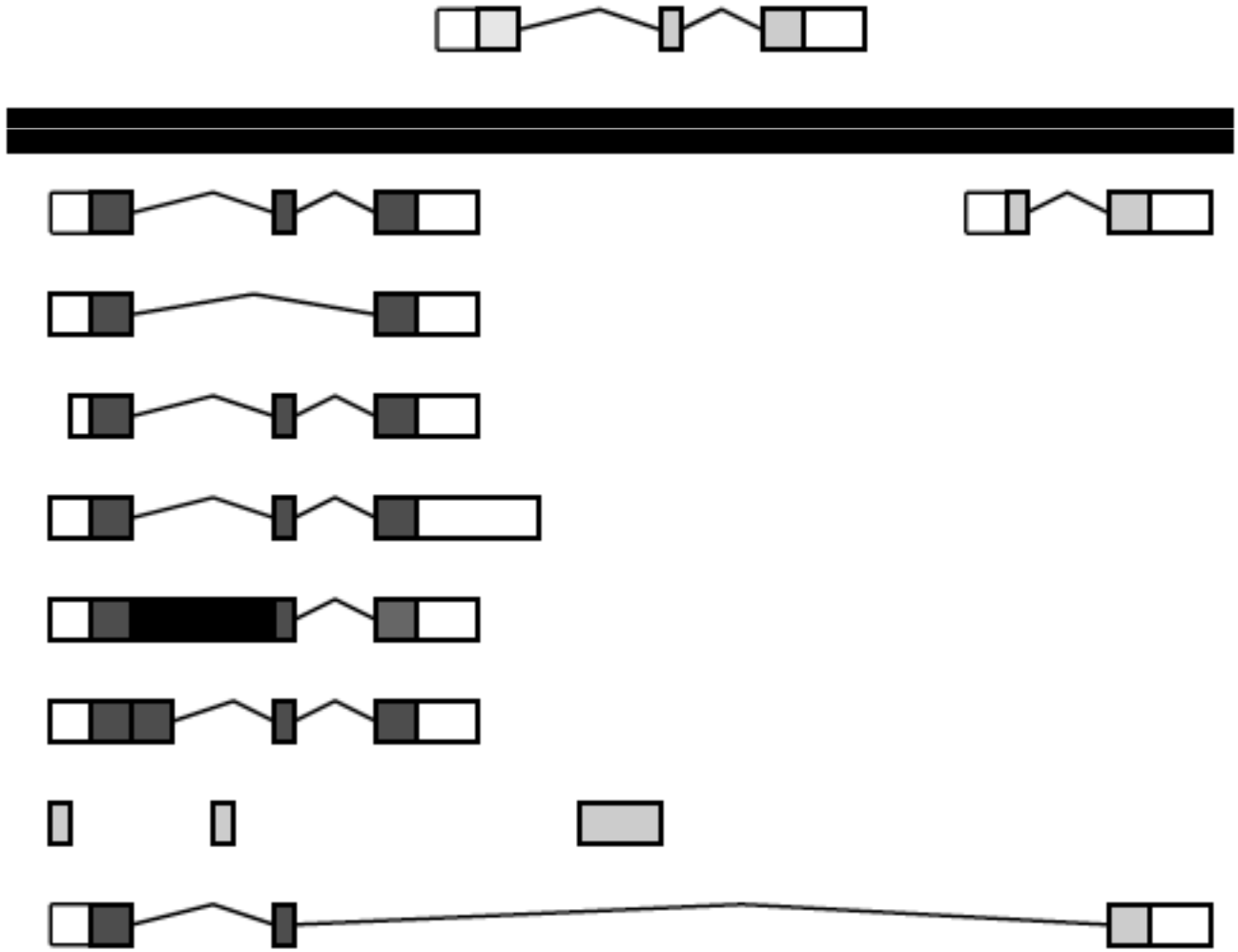
Alternative splicing

Alternative TSS

Alternative poly(A)

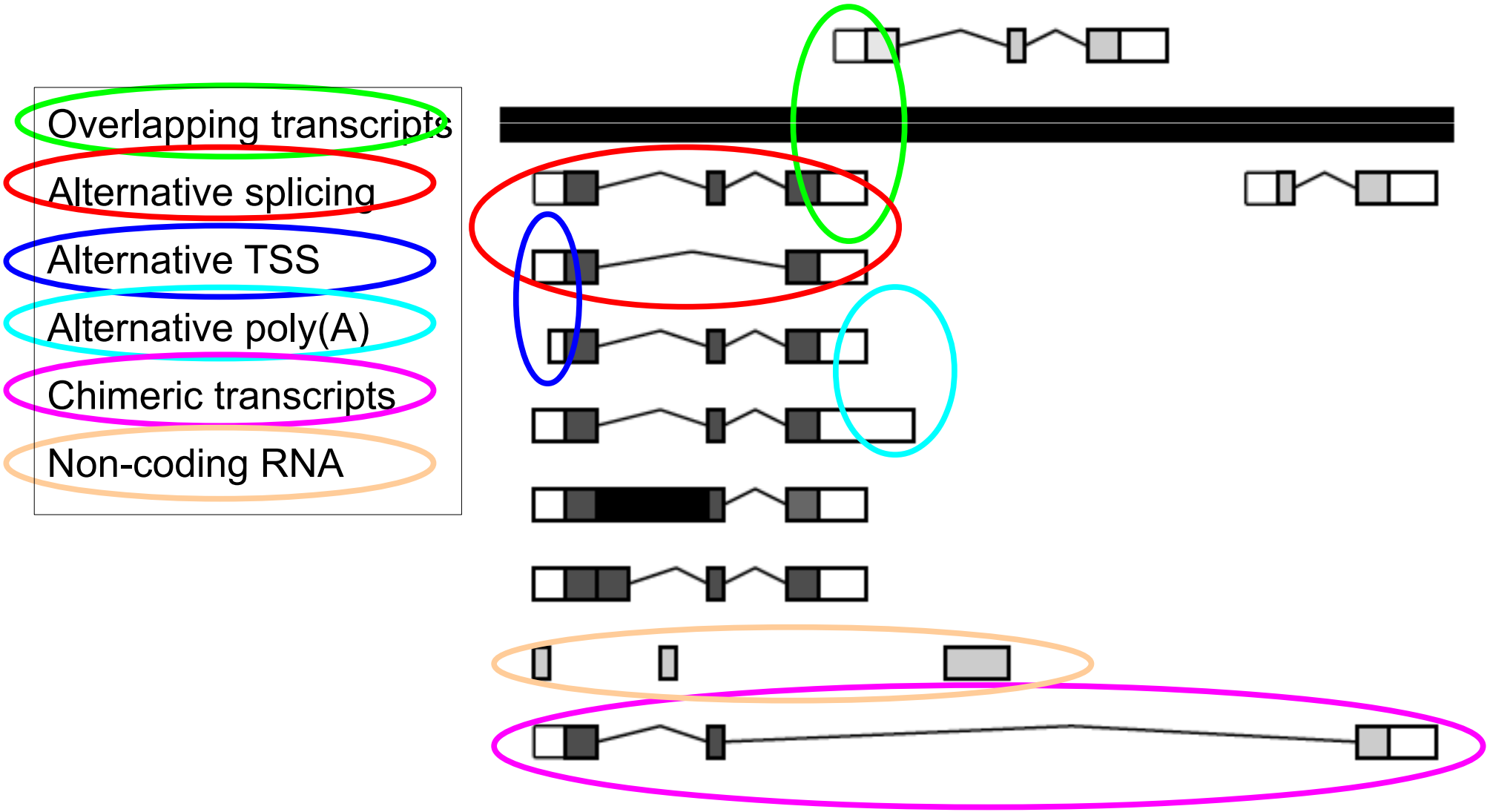
Chimeric transcripts

Non-coding RNA



From Per Unneberg, KI/SciLifeLab

Transcriptome complexity



From Per Unneberg, KI/SciLifeLab

The main tasks in transcriptome analysis

What genes (protein coding? non-coding?) are transcribed? *and*

What isoforms (splice variants; transcripts) are present?

at a certain developmental stage/condition/different cell type etc.

=> transcriptome reconstruction

At what level are these genes/transcripts expressed?

=> expression quantification

What genes (or transcripts) are up- or downregulated?

comparing 2 (or more) developmental stages/conditions/cell types etc.

=> differential expression

The main tasks in transcriptome analysis

Is there any novel transcription

=> transcription mapping

Is there a difference in transcription between gene variants in different individuals

=> analysis of coding SNPs and their impact

Is there a difference in transcription between the two gene alleles in an individual

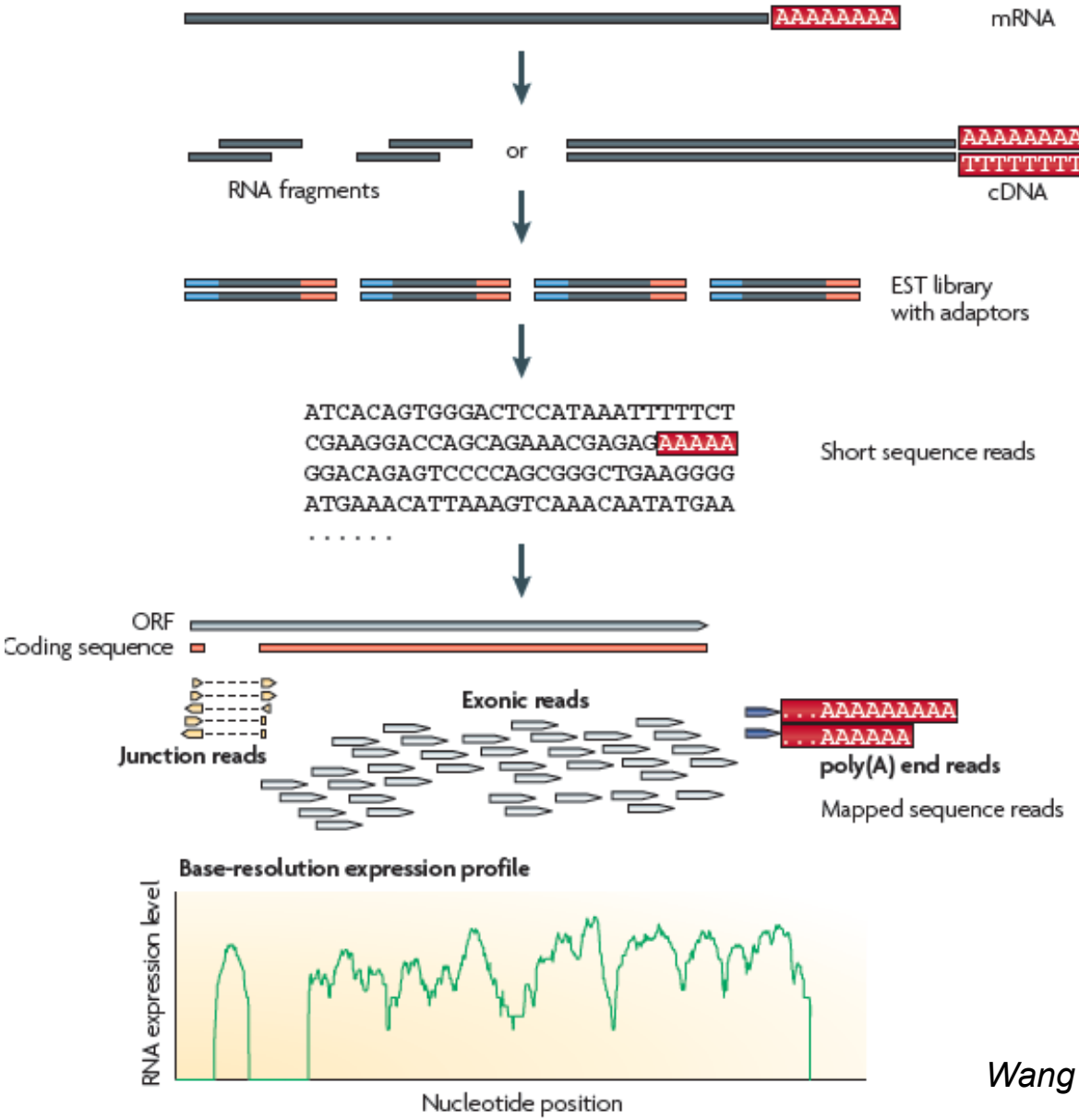
=> allele-specific transcription

=> => **RNA-seq** can help us address all of the above

RNA-seq

A typical RNA-seq experiment

RNA-seq means analyzing RNA molecules with MPS technologies.



Wang et al 2009

What is special about RNA-seq

- Compared to other MPS experiments:

  Starting material is RNA => converted to cDNA

- Compared to previous tag-based sequencing:

  Fragmentation before conversion to cDNA => uniform transcript coverage

  The throughput.

- Compared to microarrays:

  Way lower noise level

  Strand-specificity can be obtained: which DNA strand is transcribed

  You don't have to define beforehand what you're looking for

RNA-seq analysis pipeline

A “standard” RNA-seq analysis pipeline

1. Remove poor-quality sequence reads
2. Map the reads onto reference genome or reference transcriptome
   or assemble reads into transcripts or genes without a reference genome
3. Summarize the read counts for your feature of interest
4. Normalize the read counts
5. Assess differential expression – fold change and significance

---post-RNA-seq:---

6. Define what set of genes/transcripts you're interested in
7. Perform functional analysis (GO terms, Pfam domains...)
8. Validate findings using other experimental technique

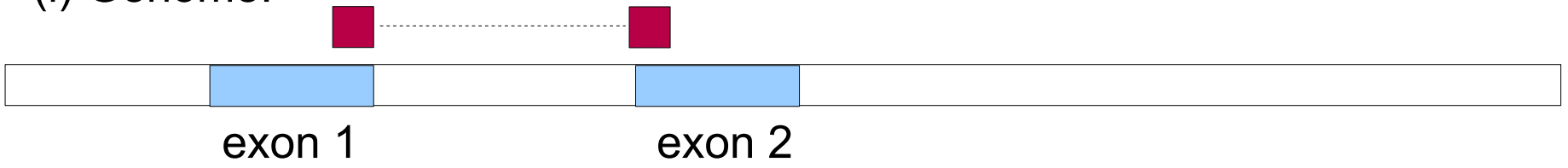
Mapping RNA-seq reads

The primary transcript (pre-mRNA) is usually processed

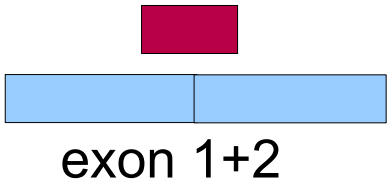
- poly(A) tail part of the read
- the RNA molecule spans exon-exon junctions (is not contiguous in gDNA)

=> mapping against (i) reference genome and (ii) exon junction database

(i) Genome:

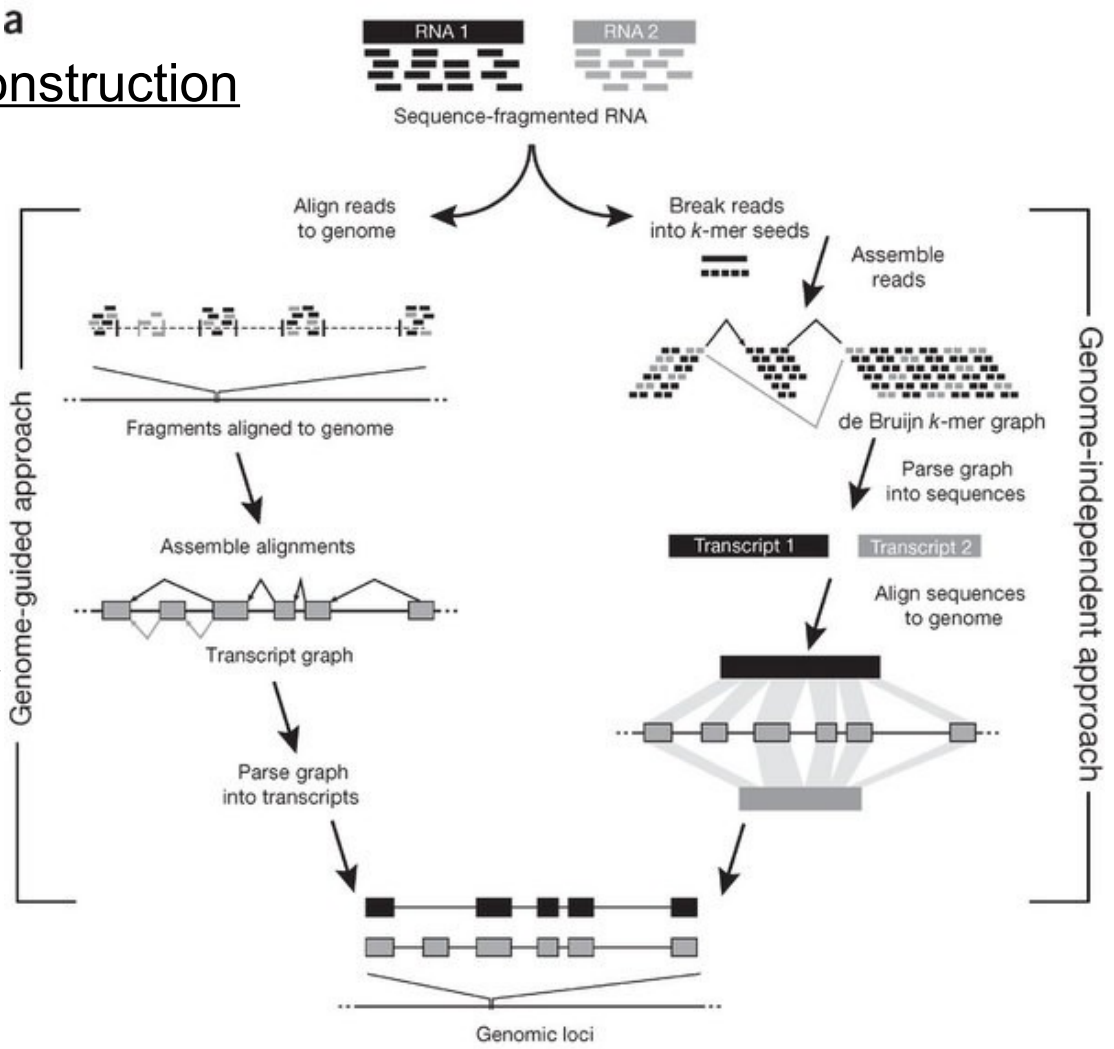


(ii) Exon junction database:



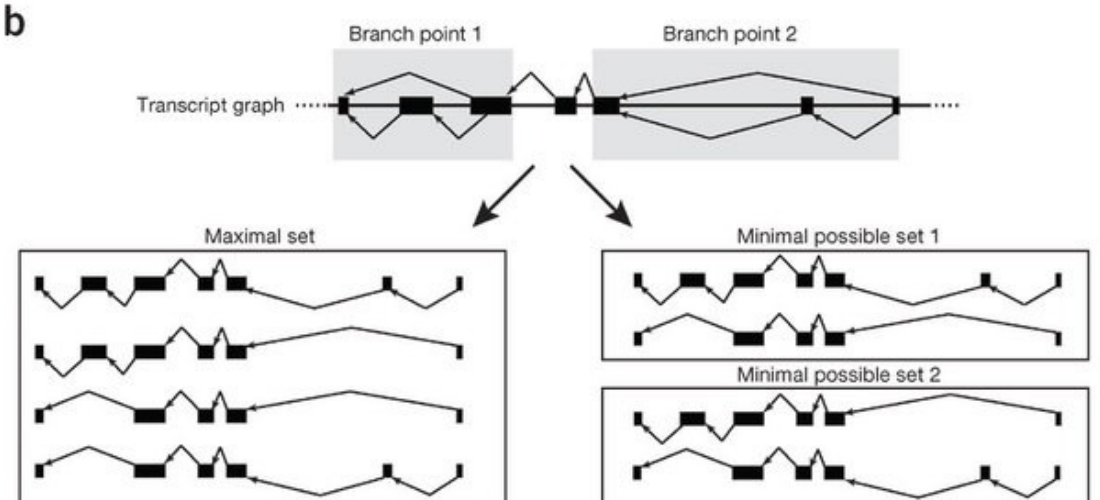
(iii) *de novo* assembly of RNA-seq reads, to discover novel splice junctions (and novel transcription in general). E.g. TopHat

Trancriptome reconstruction



"de novo"

"ab initio"



Garber et al 2011

Summarizing mapped RNA-seq reads

How many reads map to a gene? Compile a count of reads for your features of interest.

Depending on your interest and the experimental setup these features can be:

- whole gene incl. introns
- exons only
- coding sequence (CDS) only
- splicing isoforms
- antisense reads (if strand-specific protocol)

=> for each gene (or feature of interest) you obtain a **count**, which reflects the abundance of this gene (or feature of interest) in your sample

Normalization of mapped RNA-seq reads

You need to normalize the read counts to enable comparisons both within and between samples.

There is no consensus as to the best method for normalization of RNA-seq reads

Number of reads from a gene is a function of mRNA concentration and length

Normalization **within** a sample:

The most commonly used approach is the RPKM measure:

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

Mortazavi et al 2008

R=RPKM value

C= number of reads mapped to the transcript

L= gene length

N= number of million mappable reads

RPKM – reads per kilobase of transcript per million mapped reads

FPKM – fragments per kilobase of transcript per million mapped reads

Normalization of mapped RNA-seq reads

Normalization **between** samples:

In a gene-by-gene comparison, many technical biases (gene length, nucleotide composition) will cancel out

But normalization still essential to enable comparison of counts between samples

- sequence composition affect the sequencing
- a few extremely abundant genes can dominate the sequencing

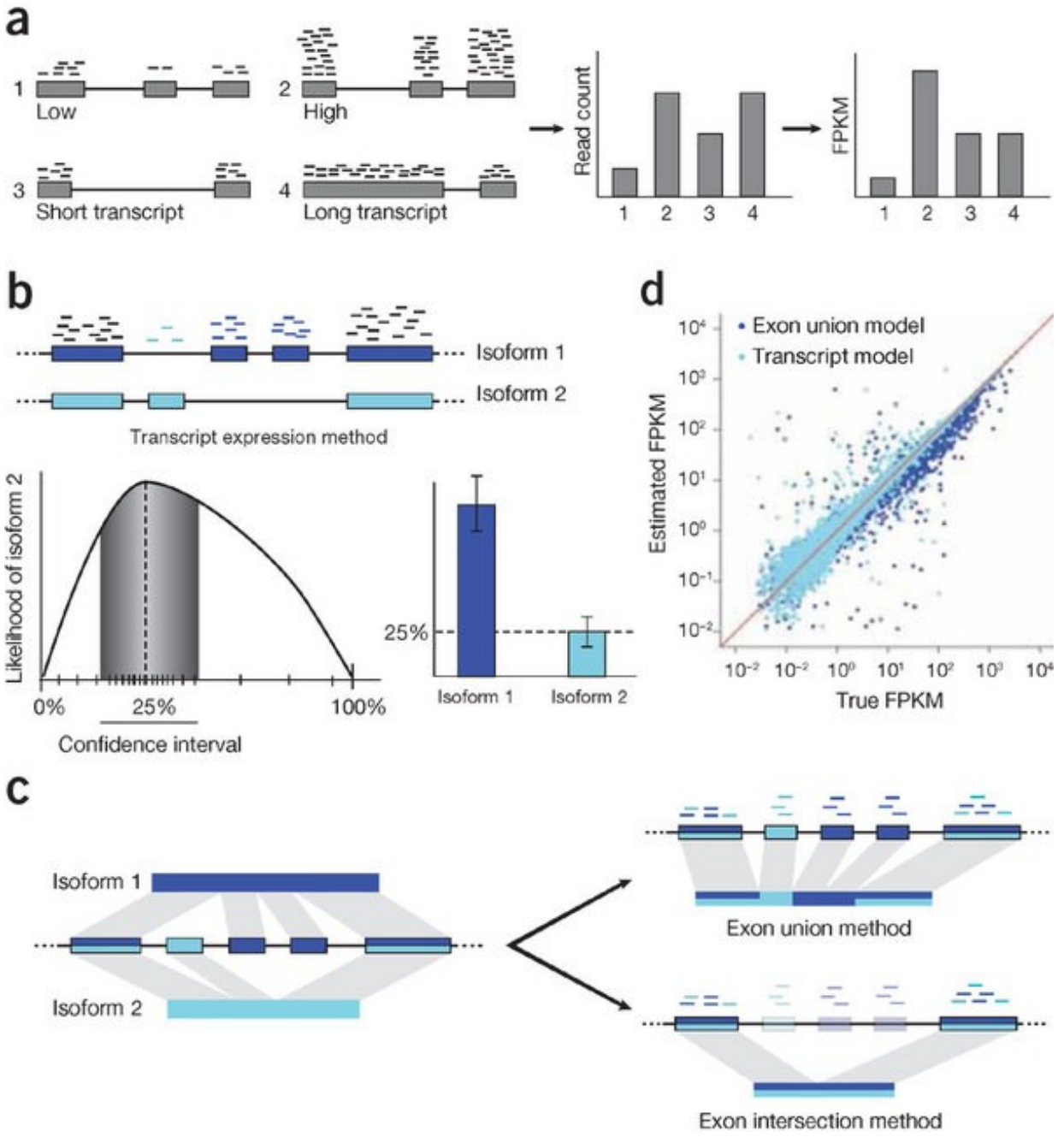
Approaches:

Normalize based on the different total number of reads in the samples

Estimate scaling factors from the data

Use quantile normalization

Expression quantification



Garber et al 2011

Assessing differential expression (DE)

When is a difference in read count also statistically significant? (I.e., the difference is greater than would be expected by random variation).

The read counts have been modelled using either:

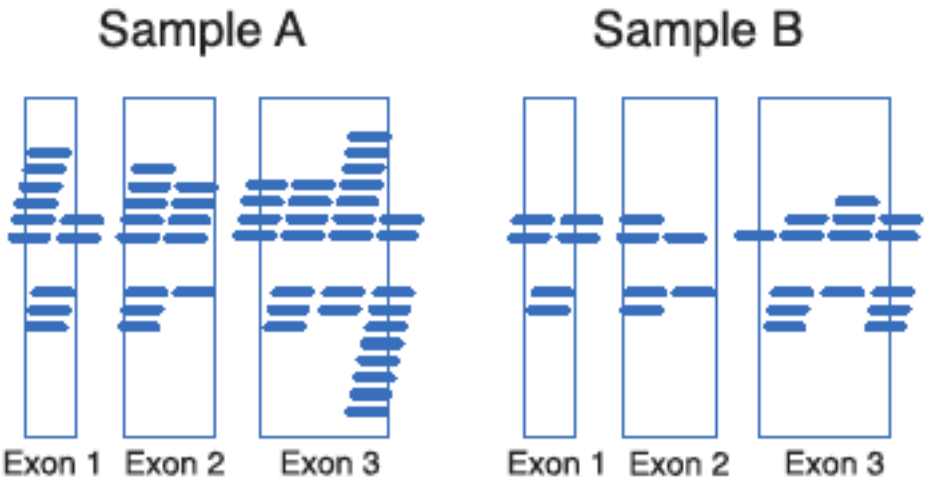
1. Poisson distribution (e.g., Myrna, DEGseq).

    - opponents claim: biological variation not well captured; sampling error underestimated => high false-positive rates (i.e., the model predicts smaller variations than is seen in real data)

2. Negative binomial (e.g. edgeR, Deseq, CuffDiff)

=> Output is, for each gene, a *P*-value describing the probability that a difference in counts is due to chance.

=> The jury is still out.



Langmead et al 2010

F8 Friday 3 Feb., 13:15, 3 papers to be presented

RPKM: **Robert Lindroos, Johannes Alneberg**

Mortazavi et al. “Mapping and quantifying mammalian transcriptomes by RNA-Seq”. *Nat Meth* (2008) vol. **5** (7) pp. 621-628

CuffLinks: **Robert Markus, Frances Vega**

Trapnell et al. “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation”. *Nat Biotechnol* (2010) vol. **28** (5) pp. 511-5

DEseq: **Benjamin Sigurgeirsson, Cecilia Lövkvist**

Anders and Huber. “Differential expression analysis for sequence count data”. *Genome Biol* (2010) vol. **11** (10) pp. R106

~12 minutes presentation, in pairs!

Email your presentation to me 1 hour before if you would like to use my computer for the presentation! [Acceptable formats: pdf, ppt, pptx, odp. **NOT** keynote].