Chapter 4 – Transcriptomes and post-transcriptional regulation

L4.1 - Transcriptomes

Post-genome

Transcriptomics

The key aims of **transcriptomics** are:

1) to catalogue all species of transcript, including mRNAs, non-coding RNAs and small RNAs;
2) to determine the transcriptional structure of genes, in terms of their start sites, 5′ and 3′ ends, splicing patterns and other post-transcriptional modifications;
3) to quantify the changing expression levels of each transcript during development and under different experimental or pathological conditions

Accessing to RNA:

RNA analysis, BASICS

1. Hybridization – based methods
2. Sequencing - based methods

Northern blotting

RNase Protection Assay (RPA)

RT-PCR

qRT-PCR

Microarray analysis (see Auxiliaries)

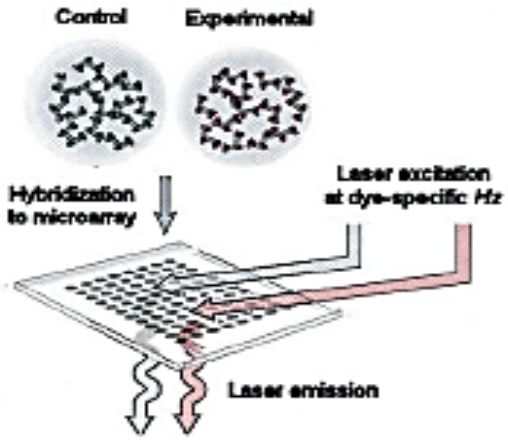
Genomic tiling microarrays.

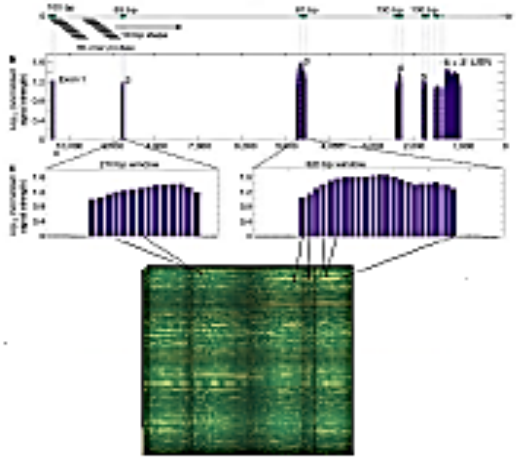
EST and SAGE, CAGE

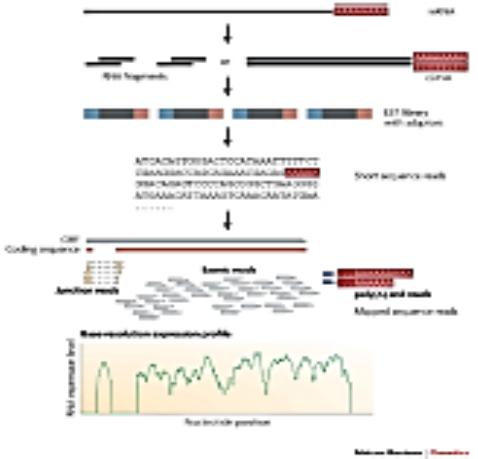
Gene-by-gene methods to measure gene expression

Serial methods to measure gene expression

The evolution of transcriptomics



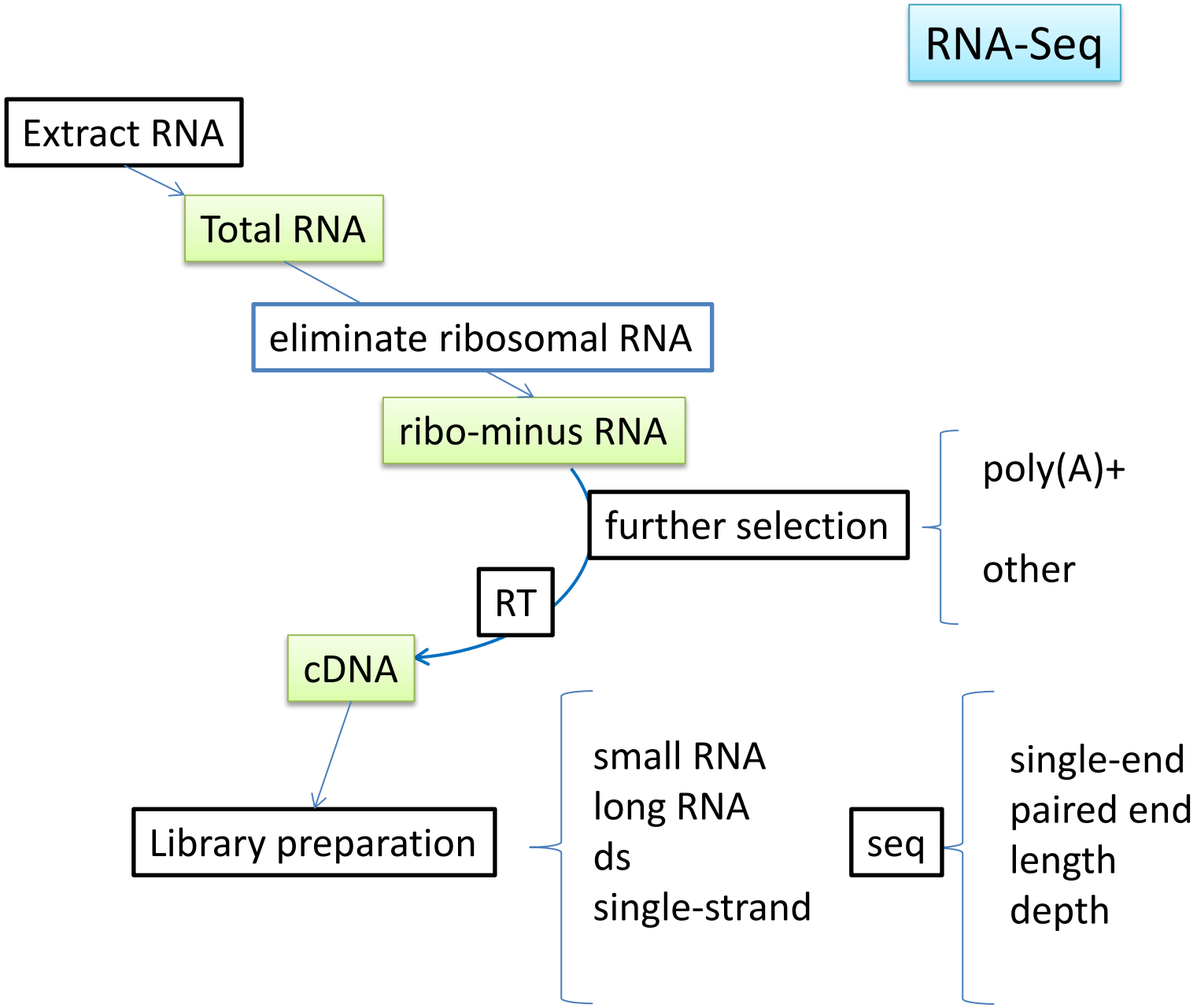




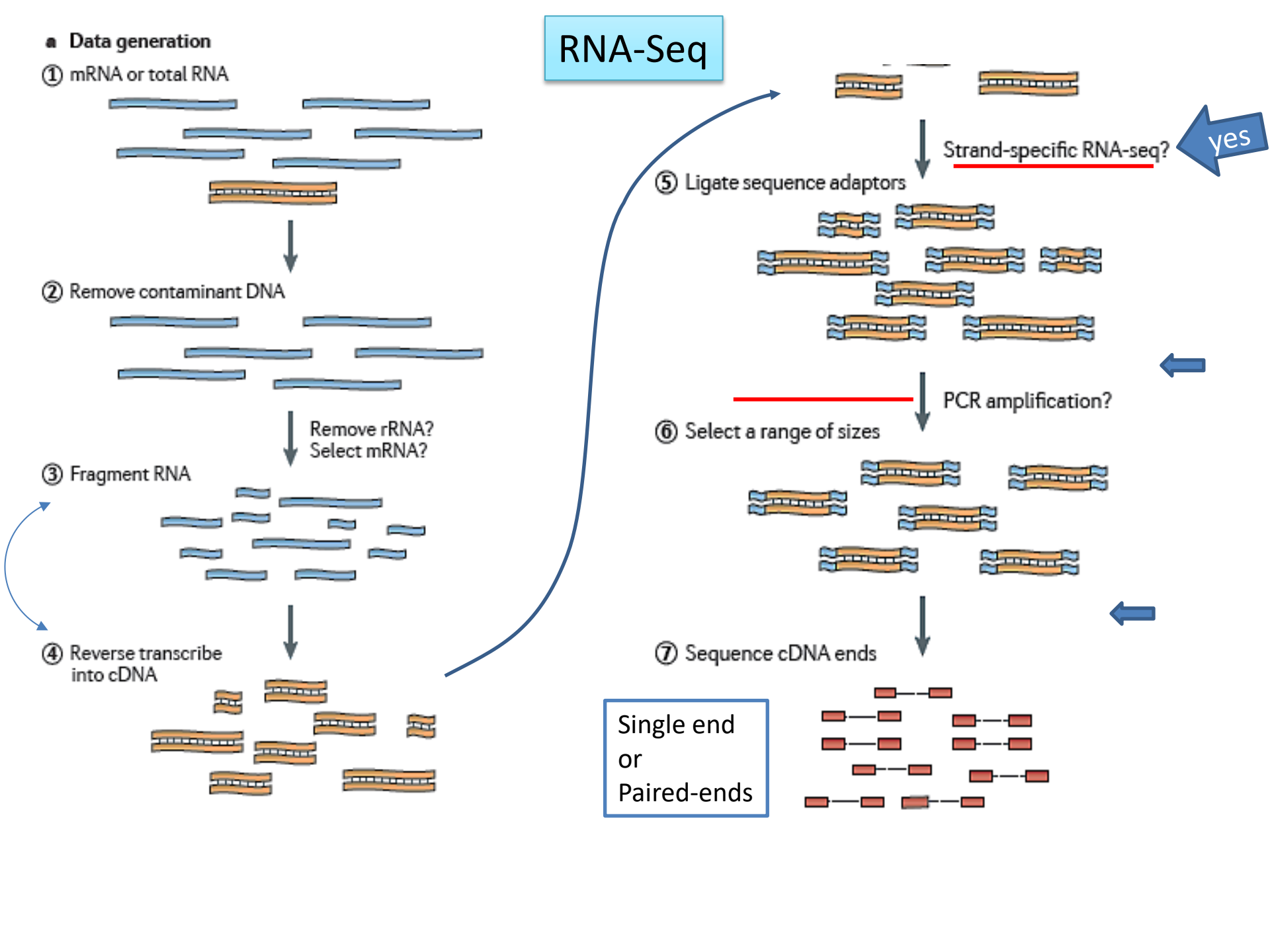
1995 P. Brown, et. al. Gene expression profiling using spotted cDNA microarray: expression levels of known genes

2002 Affymetrix, whole genome expression profiling using tiling array: identifying and profiling novel genes and splicing variants

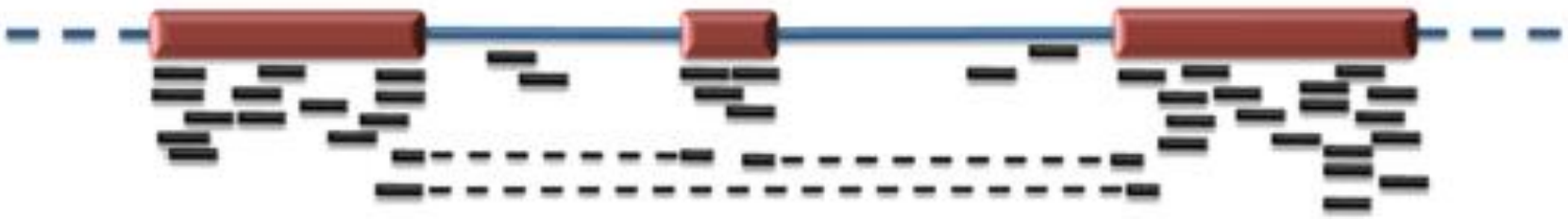
2008 many groups, mRNA-seq: direct sequencing of mRNAs using next generation sequencing techniques (NGS)


RNA-Seq
Extract RNA
Total RNA
eliminate ribosomal RNA
ribo-minus RNA
further selection
poly(A)+
other
RT
cDNA
Library preparation
small RNA
long RNA
ds
single-strand
seq
single-end
paired end
length
depth

RNA-Seq


Data generation
① mRNA or total RNA
② Remove contaminant DNA
Remove rRNA?
Select mRNA?
③ Fragment RNA
④ Reverse transcribe into cDNA
Strand-specific RNA-seq?
yes
⑤ Ligate sequence adaptors
PCR amplification?
⑥ Select a range of sizes
⑦ Sequence cDNA ends
Single end
or
Paired-ends

Mapping



Reads alignment to the genome
– Easy(ish) for genomic sequence
– Difficult for transcripts with splice junctions

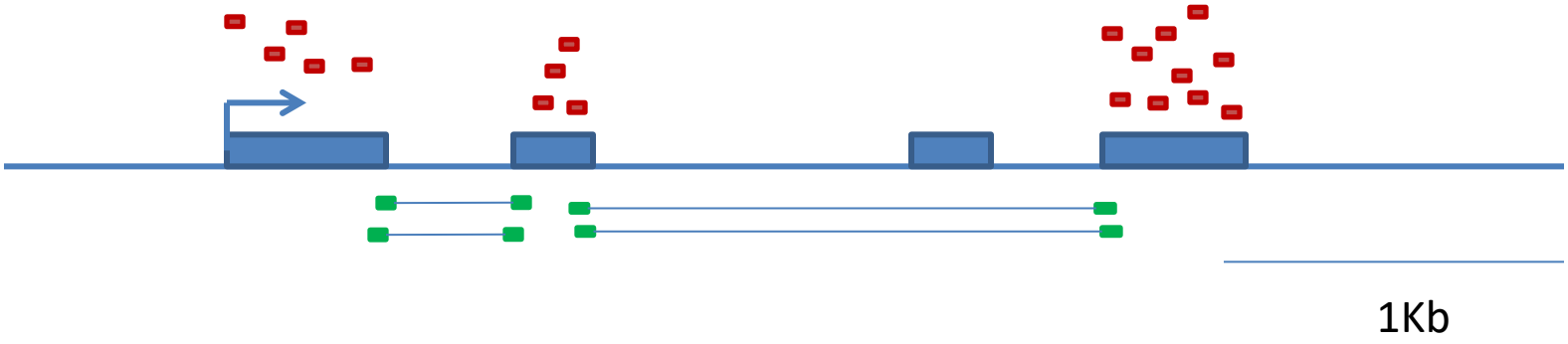
Use of specific alignment tools
(i.e. Bowtie, Tophat, MapSplice…)

Quantitative

usually one reference set of «genes» (i.e. transcripts) is chosen and reads mapped to this.

then counts are taken by integrating all the reads falling in these models.

Caution: in the example below, one exon is not expressed. Nonetheless the gene is called «expressed»: algorithms should distinguish this and map to **transcript isoforms** instead of «genes».



Quantitative (density over a region *or* transcription unit)

rpkm (reads per kilobase per million reads)

Double normalization for sequencing depth and gene length:

1- Divide the read counts by the "per million" scaling factor. This normalizes for sequencing depth, giving you reads per million (RPM)

2- Divide the RPM values by the length of the gene, in kilobases. This gives you RPKM.

fpkm = fragments per kilobase per million

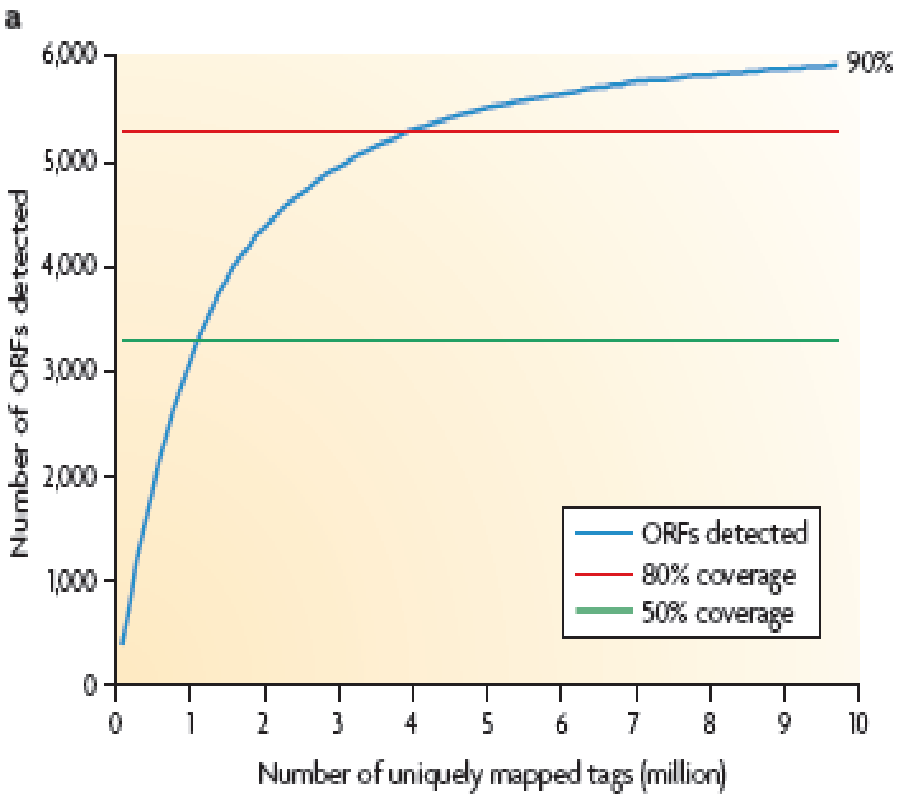
One of the major variable in RNA-seq experiments (aside the kind of RNA prep) is the sequencing depth

Sequencing depth *versus* sensitivity

Always remember that the molecules you have sequenced are a «Sample» of the total possible reads from your biological sample.

How representative this sample is will depend on the number of molecules you have sequenced (i.e. the sequencing depth).

Increasing sequencing depth (higher coverage) helps identifying new transcripts



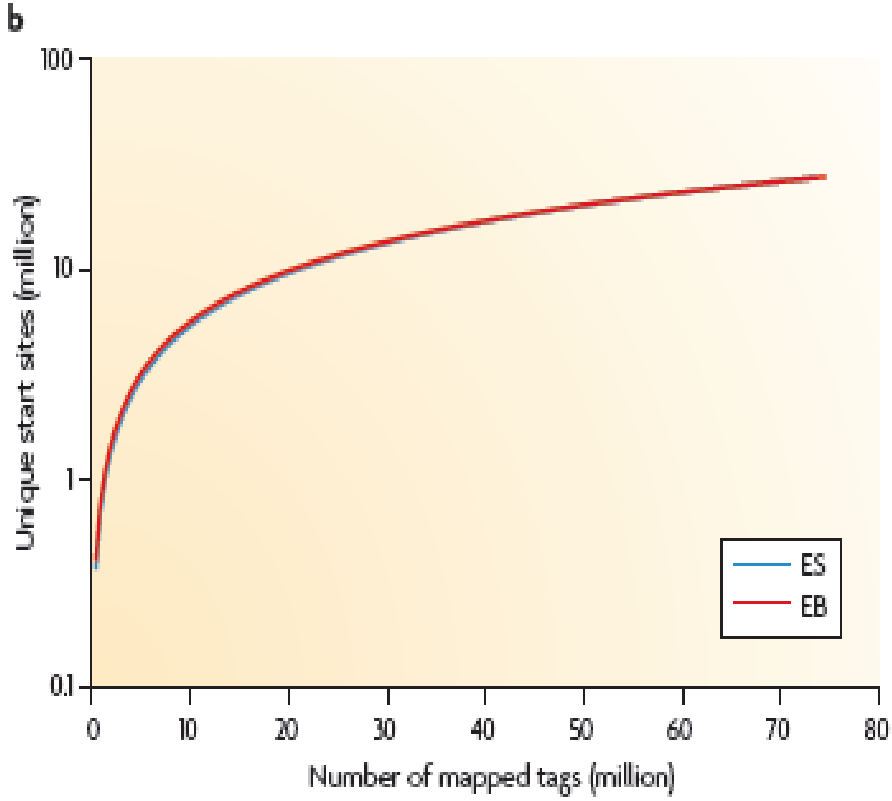


Figure 5 | **Coverage versus depth.** a | 80% of yeast genes were detected at 4 million uniquely mapped RNA-Seq reads, and coverage reaches a plateau afterwards despite the increasing sequencing depth. Expressed genes are defined as having at least four independent reads from a 50-bp window at the 3′ end. Data is taken from REF. 18.

b | The number of unique start sites detected starts to reach a plateau when the depth of sequencing reaches 80 million in two mouse transcriptomes. ES, embryonic stem cells; EB, embryonic body. Figure is modified, with permission, from REF. 22 © (2008) Macmillan Publishers Ltd. All rights reserved.

Qualitative

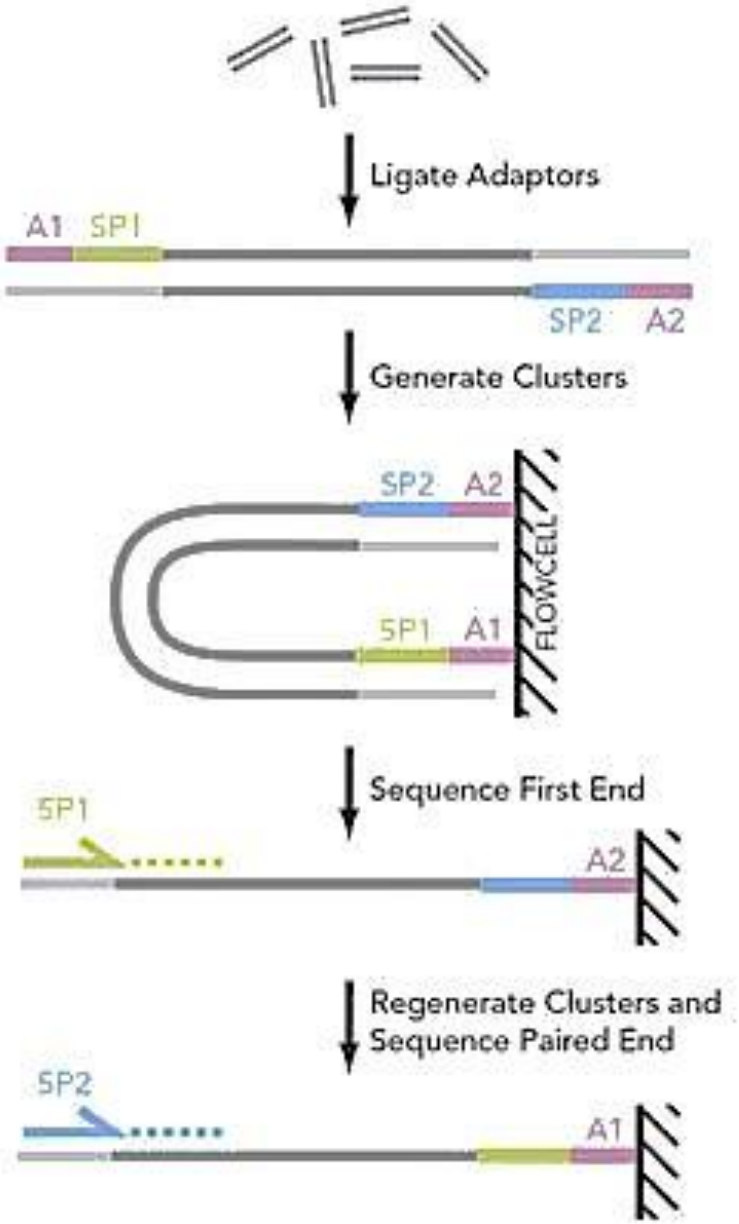
Mapping

Reads are aligned to the reference genome, or to more limited reference of your choice:

- known exons of protein-coding genes (exome)
- Spliced reads (*pay attention to this*!)
- Genes (sense and antisense)

New transcript definition : requires high sequencing depth

Fragmented cDNA

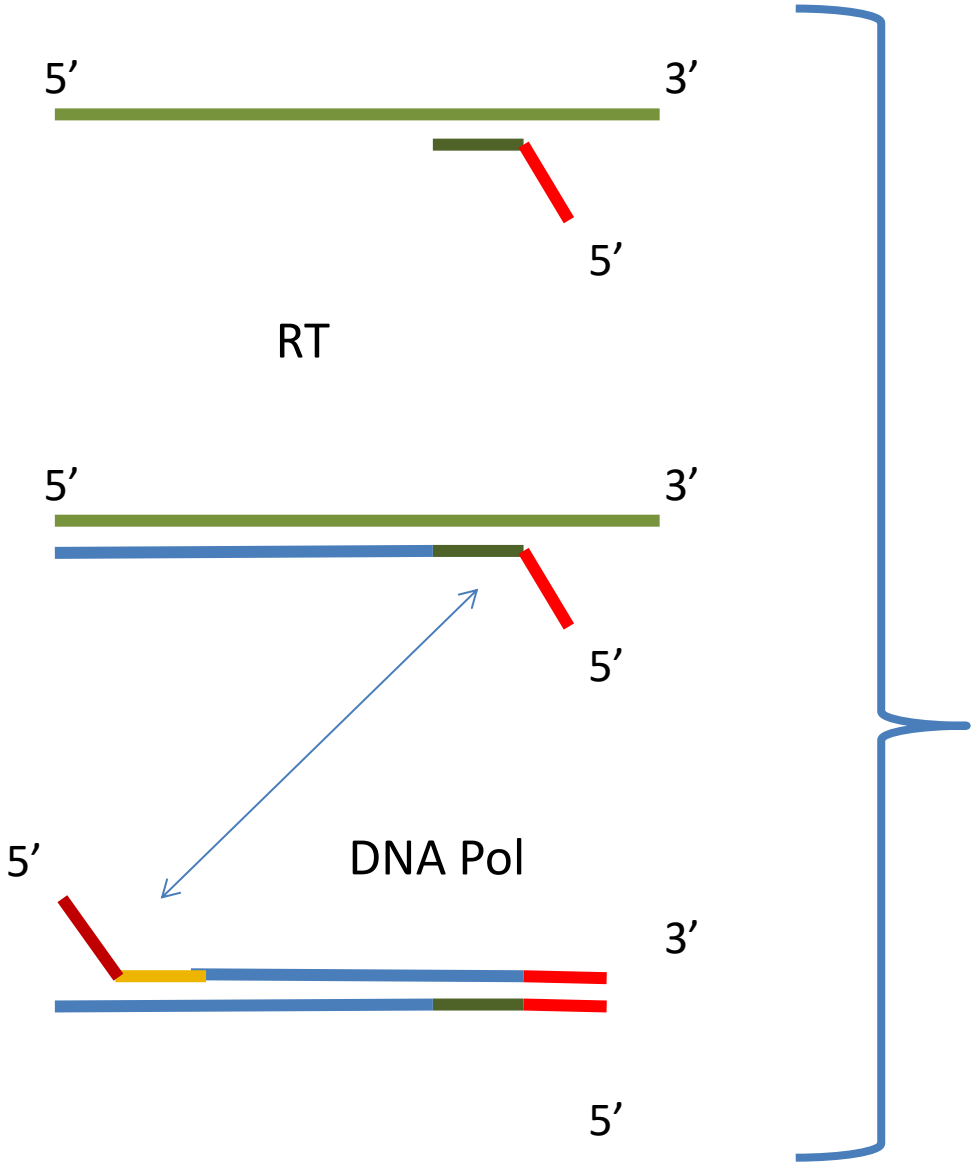


Paired-end sequencing (from Illumina)

Paired-end sequencing used in many published experiments

more common today

Strand-specific library



nature

First study by RNA-Seq

ARTICLES

Alternative isoform regulation in human tissue transcriptomes

Eric T. Wang[1,2]*, Rickard Sandberg[1,3]*, Shujun Luo[4], Irina Khrebtukova[4], Lu Zhang[4], Christine Mayr[5], Stephen F. Kingsmore[6], Gary P. Schroth[4] & Christopher B. Burge[1]

Through alternative processing of pre-messenger RNAs, individual mammalian genes often produce multiple mRNA and protein isoforms that may have related, distinct or even opposing functions. Here we report an in-depth analysis of 15 diverse human tissue and cell line transcriptomes on the basis of deep sequencing of complementary DNA fragments, yielding a digital inventory of gene and mRNA isoform expression. Analyses in which sequence reads are mapped to exon–exon junctions indicated that 92–94% of human genes undergo alternative splicing, ~86% with a minor isoform frequency of 15% or more. Differences in isoform-specific read densities indicated that most alternative splicing and alternative cleavage and polyadenylation events vary between tissues, whereas variation between individuals was approximately twofold to threefold less common. Extreme or 'switch-like' regulation of splicing between tissues was associated with increased sequence conservation in regulatory regions and with generation of full-length open reading frames. Patterns of alternative splicing and alternative cleavage and polyadenylation were strongly correlated across tissues, suggesting coordinated regulation of these processes, and sequence conservation of a subset of known regulatory motifs in both alternative introns and 3′ untranslated regions suggested common involvement of specific factors in tissue-level regulation of both splicing and polyadenylation.

Vol 456|27 November 2008|doi:10.1038/nature07509

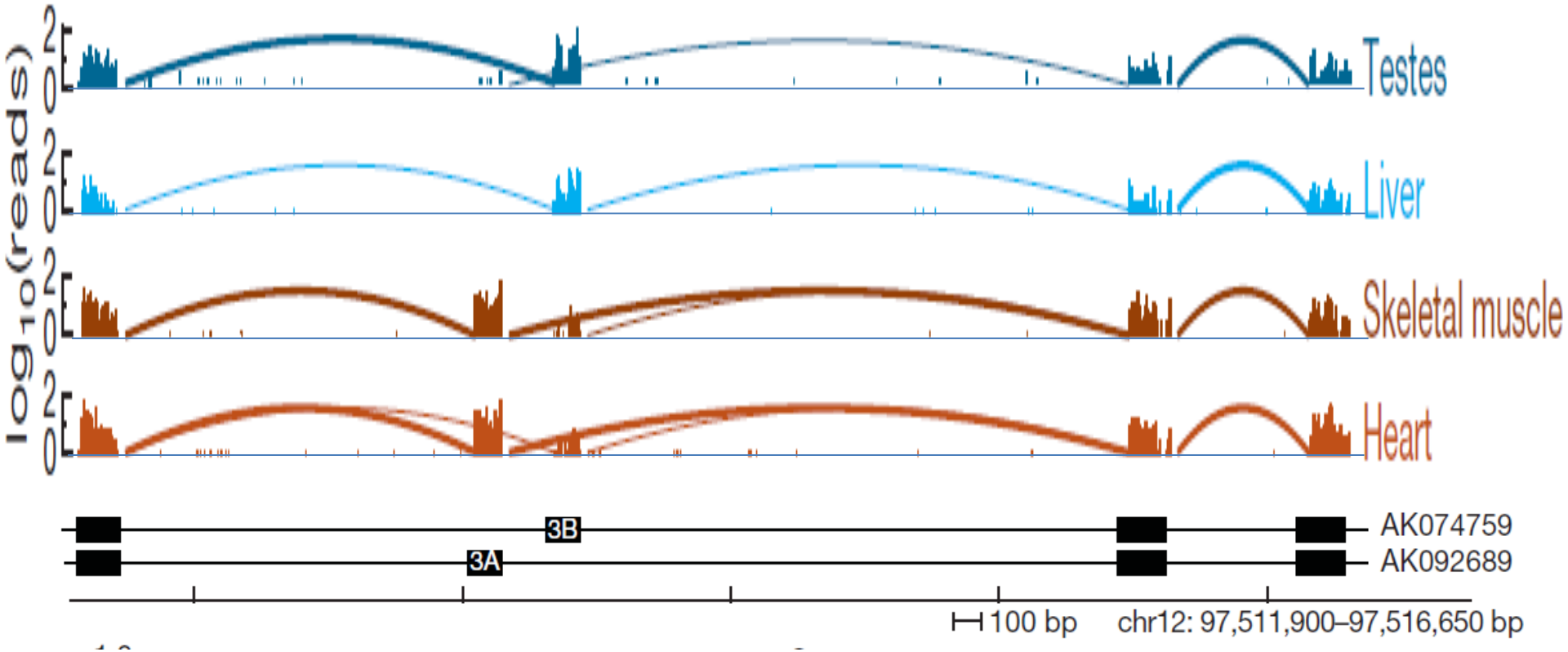
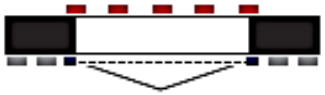
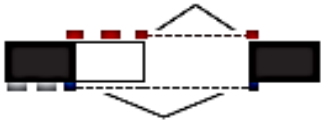
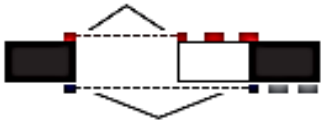
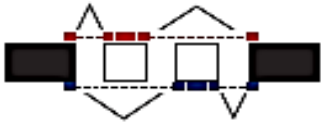
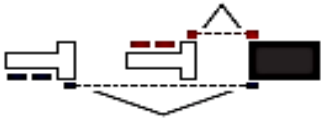
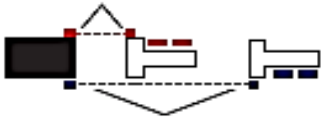


Figure 1 | Frequency and relative abundance of alternative splicing isoforms in human genes.

a, mRNA-Seq reads mapping to a portion of the SLC25A3 gene locus. The number of mapped reads starting at each nucleotide position is displayed (log10) for the tissues listed at the right. Arcs represent junctions detected by splice junction reads.

Bottom: exon/intron structures of representative transcripts containing mutually exclusive exons 3A and 3B (GenBank accession numbers shown at the right).

Alternative transcript events		Total events ($\times 10^3$)	Number detected ($\times 10^3$)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Skipped exon		37	35	10,436	6,822	65	72
Retained intron		1	1	167	96	57	71
Alternative 5′ splice site (A5SS)		15	15	2,168	1,386	64	72
Alternative 3′ splice site (A3SS)		17	16	4,181	2,655	64	74
Mutually exclusive exon (MXE)		4	4	167	95	57	66
Alternative first exon (AFE)		14	13	10,281	5,311	52	63
Alternative last exon (ALE)		9	8	5,246	2,491	47	52
Tandem 3′ UTRs		7	7	5,136	3,801	74	80
Total		105	100	37,782	22,657	60	68

Constitutive exon or region — Body read — Junction read — pA Polyadenylation site

Alternative exon or extension — Inclusive/extended isoform — Exclusive isoform — Both isoforms

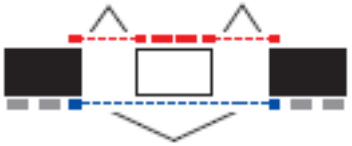
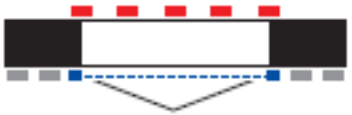
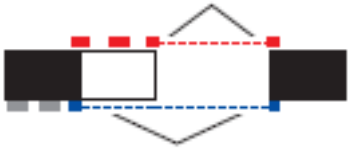
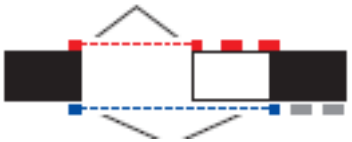
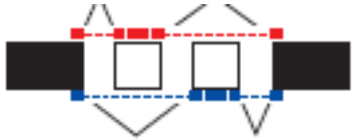
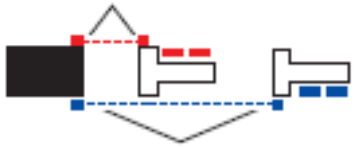
Alternative transcript events		Total events ($\times 10^3$)	Number detected ($\times 10^3$)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Skipped exon		37	35	10,436	6,822	65	72
Retained intron		1	1	167	96	57	71
Alternative 5′ splice site (A5SS)		15	15	2,168	1,386	64	72
Alternative 3′ splice site (A3SS)		17	16	4,181	2,655	64	74

Figure 2 | Pervasive tissue-specific regulation of alternative mRNA isoforms. Rows represent the eight different alternative transcript event types diagrammed. Mapped reads supporting expression of upper isoform, lower isoform or both isoforms are shown in blue, red and grey, respectively. Columns 1–4 show the numbers of events of each type: (1) supported by cDNA and/or EST data; (2) with ≥ 1 isoform supported by mRNA-Seq reads; (3) with both isoforms supported by reads; and (4) events detected as tissue regulated (Fisher's exact test) at an FDR of 5% (assuming negligible technical variation).

Alternative transcript events		Total events ($\times 10^3$)	Number detected ($\times 10^3$)	Both isoforms detected	Number tissue-regulated	% Tissue-regulated (observed)	% Tissue-regulated (estimated)
Mutually exclusive exon (MXE)		4	4	167	95	57	66
Alternative first exon (AFE)		14	13	10,281	5,311	52	63
Alternative last exon (ALE)		9	8	5,246	2,491	47	52
Tandem 3′ UTRs		7	7	5,136	3,801	74	80
Total		105	100	37,782	22,657	60	68

Constitutive exon or region — Body read — Junction read — pA Polyadenylation site

Alternative exon or extension — Inclusive/extended isoform — Exclusive isoform — Both isoforms

Columns 5 and 6 show: (5) the observed percentage of events with both isoforms detected that were observed to be tissue-regulated; and (6) the estimated true percentage of tissue-regulated isoforms after correction for power to detect tissue bias (Supplementary Fig. 6) and for the FDR. For some event types, 'common reads' (grey bars) were used in lieu of (for tandem 39UTR events) or in addition to 'exclusion' reads for detection of changes in isoform levels between tissues.
Note that Aa use the following definition for "tissue-specific": at least 10% variation in isoforms.

This paper described a number of «known» features of genes

1) the usage of Alternative Promoters

2) Alternative splicing of internal exons

3) the usage of alternative polyadenylation sites

The real news was the **frequency and extension** of these phenomena