RNA-Sequencing analysis

Markus Kreuz

25. 04. 2012

imise.

Institut für Medizinische Informatik, Statistik und Epidemiologie

Content:

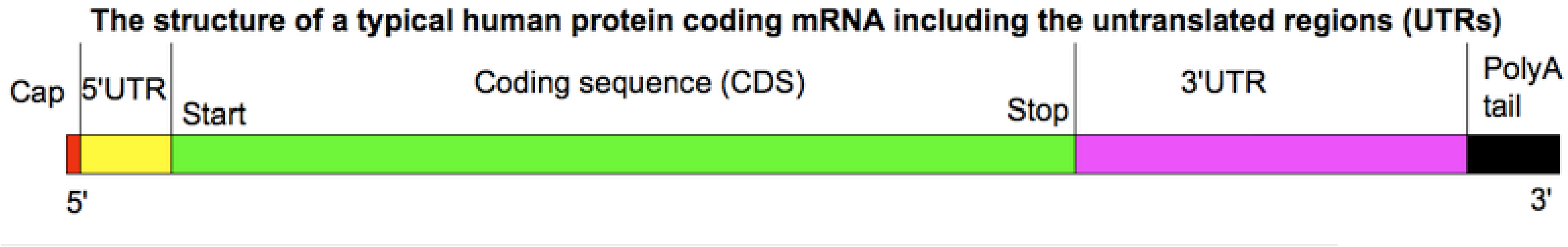
- Biological background
  - Overview transcriptomics

- RNA-Seq
  - RNA-Seq technology
  - Challenges
  - Comparable technologies

- Expression quantification
  - ReCount database

Biological background (I):

- Structure of a protein coding mRNA



The structure of a typical human protein coding mRNA including the untranslated regions (UTRs)
Cap
5'UTR
Start
Coding sequence (CDS)
Stop
3'UTR
PolyA
tail
5'
3'

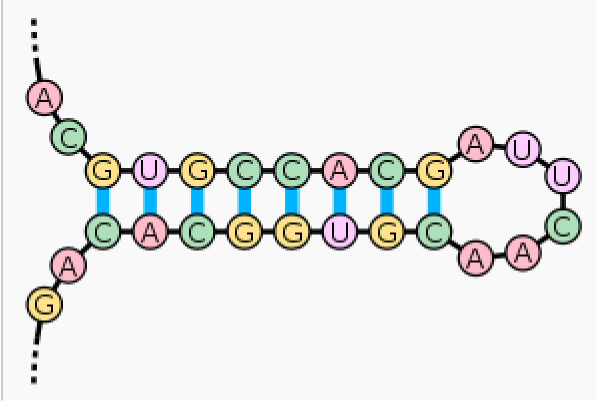
- Non coding RNAs:

Type	**Size**	**Function**
microRNA (miRNA)	21-23 nt	regulation of gene expression
small interfering RNA (siRNA)	19-23 nt	antiviral mechanisms
piwi-interacting RNA (piRNA)	26-31 nt	interaction with piwi proteins/spermatogenesis
small nuclear RNA (snRNA)	100-300 nt	RNA splicing
small nucleolar RNA (snoRNA)	-	modification of other RNAs

Biological Background (II):

- Processing
  - Splicing / Alternative Splicing / Trans-Splicing
  - RNA editing

- Secondary structures
  - Example hairpin structure:

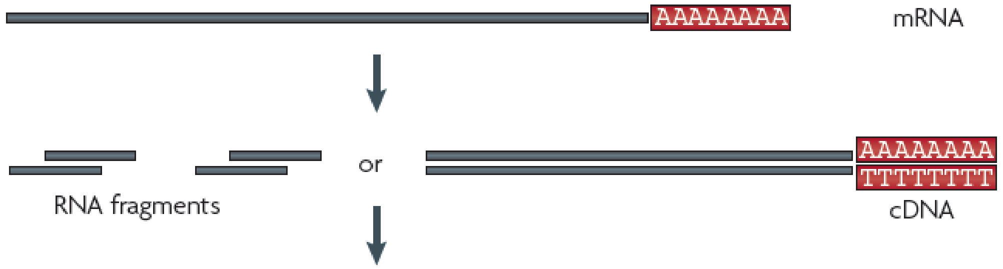


RNA-Seq technology -Aims:

- Catalogue all species of transcript including: mRNAs, non-coding RNAs and small RNAs

- Determine the transcriptional structure of genes in terms of:
  - Start sites
  - 5′ and 3′ ends
  - Splicing patterns
  - Other post-transcriptional modifications
  - Quantification of expression levels and comparison (different conditions, tissues, etc.)

RNA-Seq analysis (I):



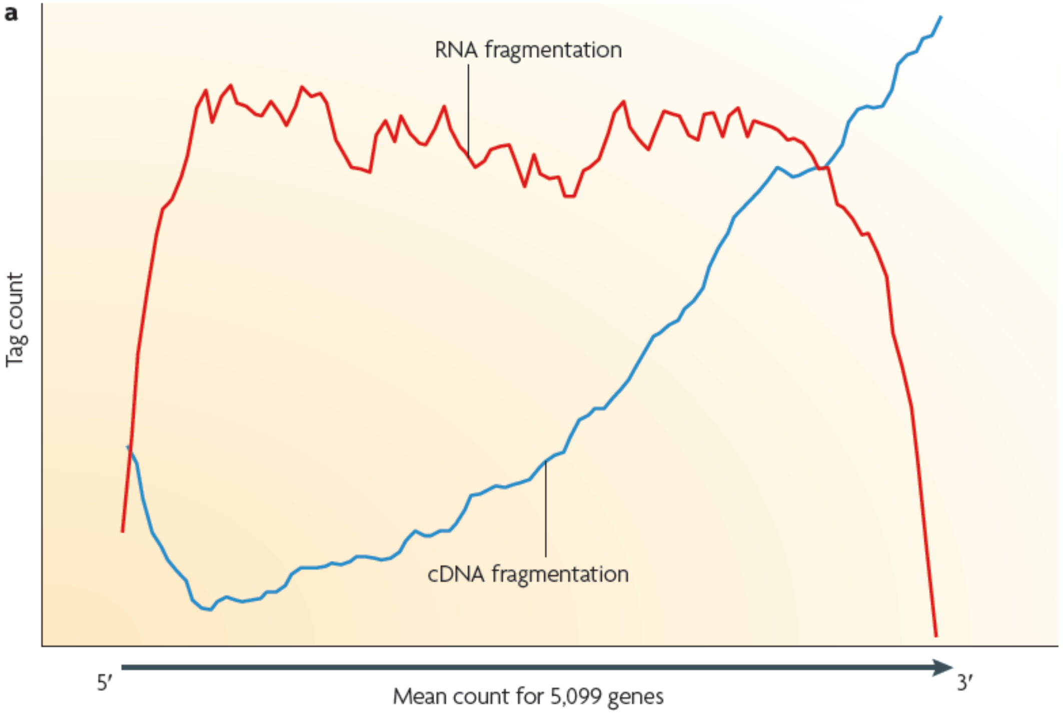
Long RNAs are first converted into a library of cDNA fragments through either:

RNA fragmentation or DNA fragmentation

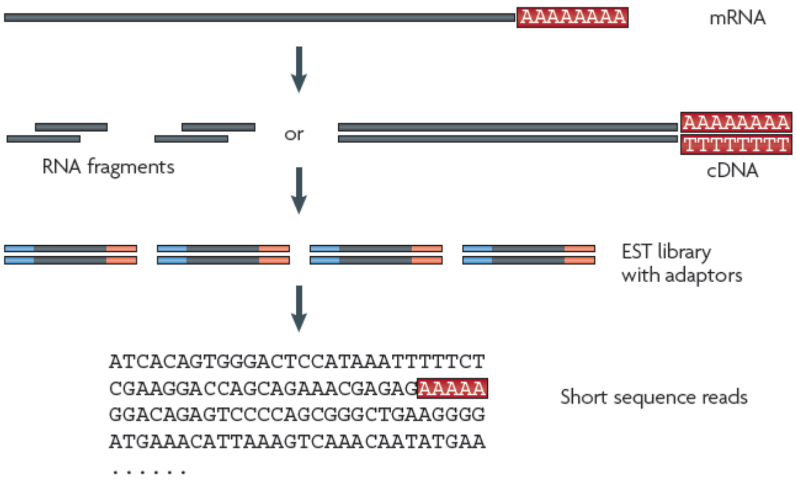
RNA-Seq analysis (II):

- In contrast to small RNAs (like piRNAs, miRNAs, siRNAs) larger RNA must be fragmented
- RNA fragmentation or cDNA fragmentation (different techniques)

- Methods create different type of bias:
  - RNA:
    depletion for ends
  - cDNA:
    biased towards 5′ end

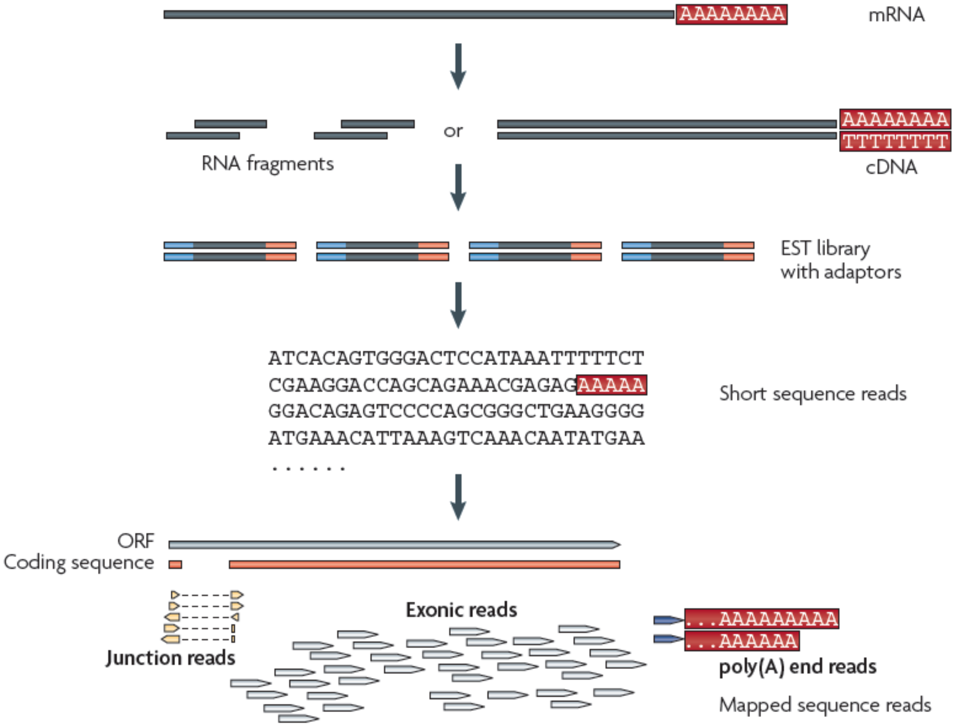


RNA-Seq analysis (III):


AAAAAAAA
mRNA
or
AAAAAAAA
TTTTTTTT
RNA fragments
cDNA
EST library
with adaptors
ATCACAGTGGGACTCCATAAATTTTTCT
CGAAGGACCAGCAGAAACGAGAGAAAAA
GGACAGAGTCCCCAGCGGGCTGAAGGGG
ATGAAACATTAAAGTCAAACAATATGAA
. . . . . .
Short sequence reads

Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing Technology
(typical read length: 30-400 bp depending on technology)

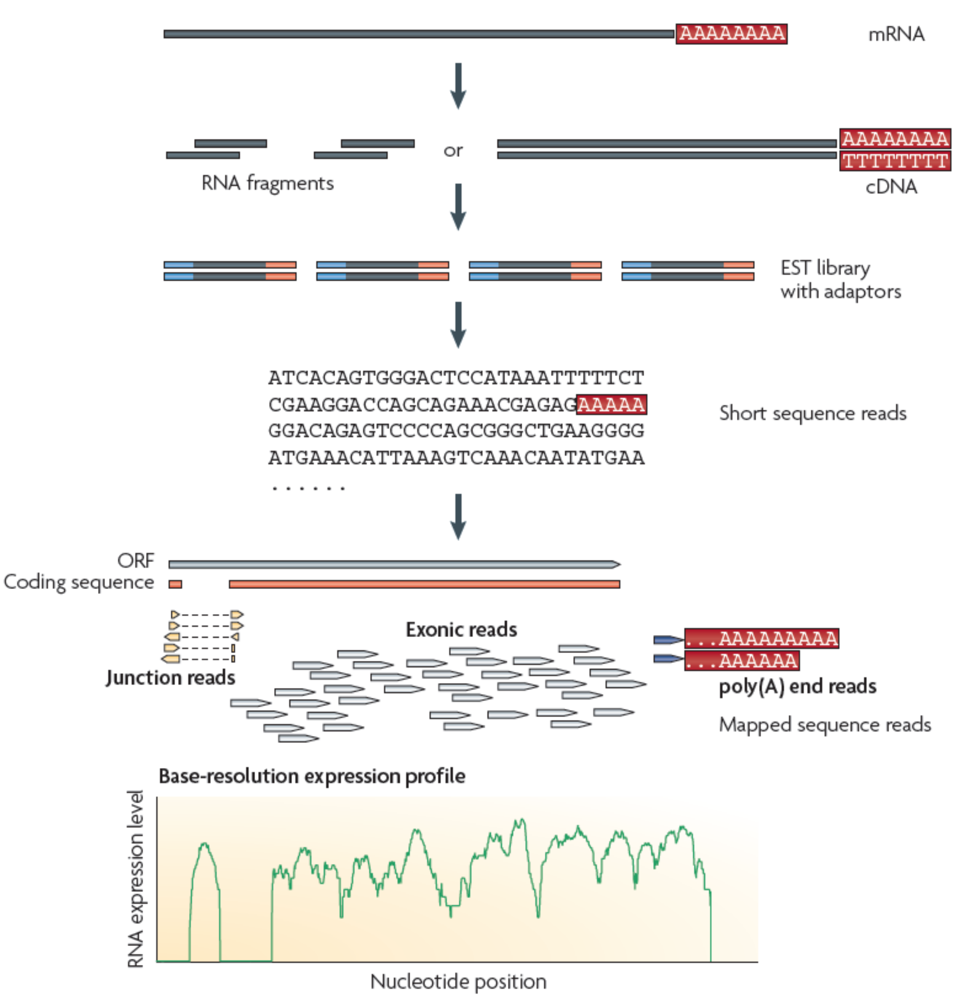
RNA-Seq analysis (IV):



The resulting sequence reads are aligned with the reference genome or transcriptome and classified as three types: exonic reads, junction reads and poly(A) end-reads.

(de novo assembly also possible => attractive for non-model organisms)

RNA-Seq analysis (V):



These three types are used to generate a base-resolution expression profile for each gene

Example:

A yeast ORF with one intron

RNA-Seq - Bioinformatic challenges (I):

- Storing, retrieving and processing of large amounts of data
- Base calling
- Quality analysis for bases and reads

  => FastQ files

- Mapping/aligning RNA-Seq reads
  (Alternative: assemble contigs and align them to genome)
  - Multiple alignment possible for some reads
  - Sequencing errors and polymorphisms

  =>SAM/BAM files

RNA-Seq - Bioinformatic challenges (II):

Specific challenges for RNA-Seq:

- Exon junctions and poly(A) ends
  - Identification of poly(A) -> long stretches of A or T at end of reads
  - Splice sites:
    - Specific sequence context: CT – AG dinucleotides
    - Low expression for intronic regions
    - Known or predicted splice sites
    - Detection of new sites (e.g. via split read mapping)
- Overlapping genes
- RNA editing
- Secondary structure of transcripts
- Quantification of expression signals

Coverage, sequencing depth and costs:

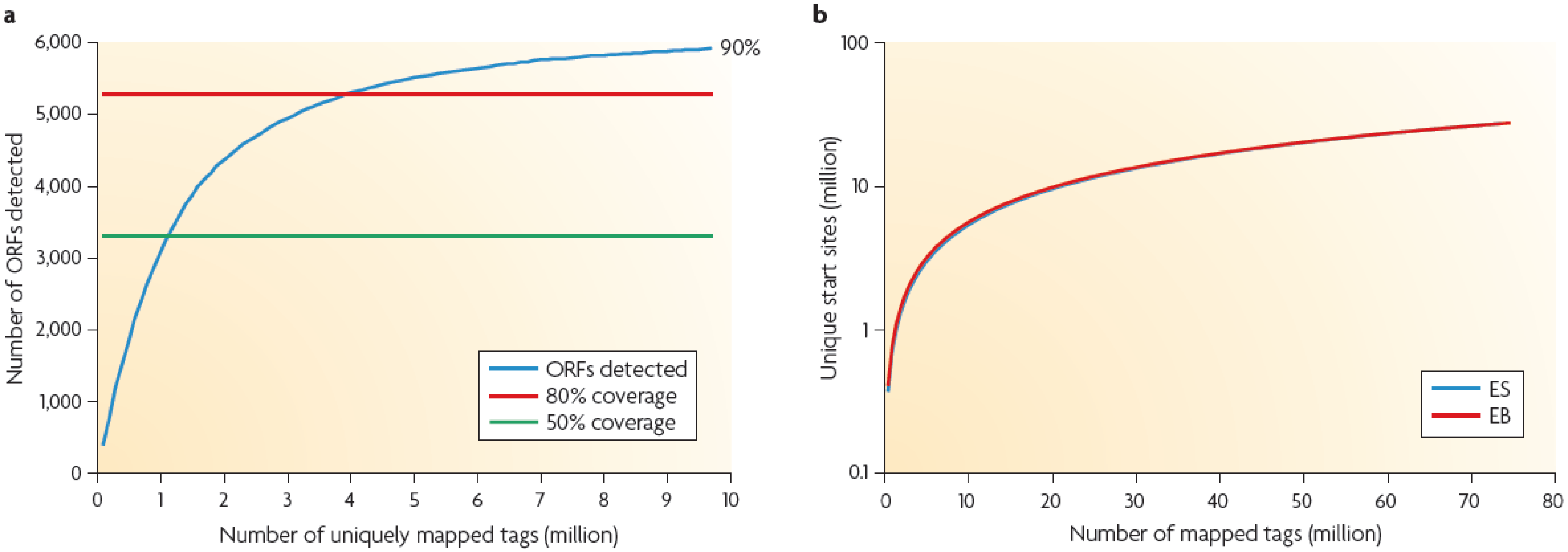


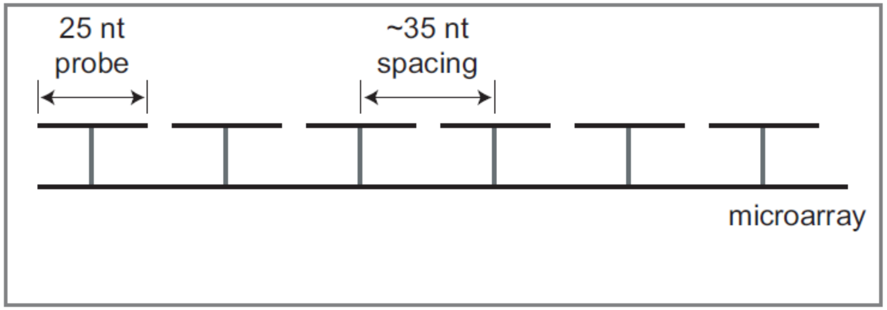
Figure 5 | **Coverage versus depth. a** | 80% of yeast genes were detected at 4 million uniquely mapped RNA-Seq reads, and coverage reaches a plateau afterwards despite the increasing sequencing depth. Expressed genes are defined as having at least four independent reads from a 50-bp window at the 3′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

- Number of detected genes (coverage) and costs increase with sequence depth (number of analyzed read)
- Calculation of coverage is less straightforward in transcriptome analysis (transcription activity varies)

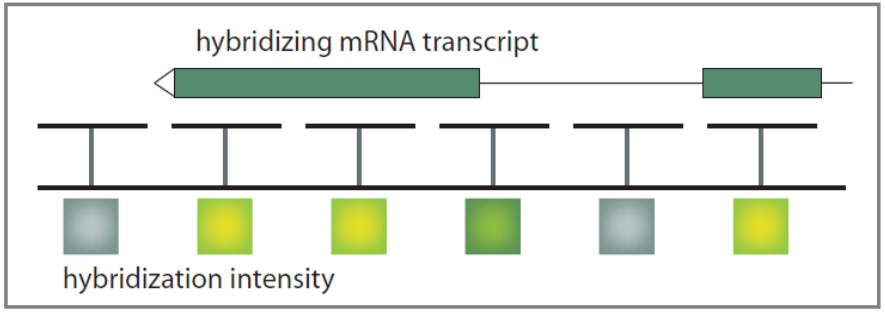
RNA-Seq - Comparable technologies:

- Tiling array analysis
- Classical sequencing of cDNA or EST
- Classical gene expression arrays

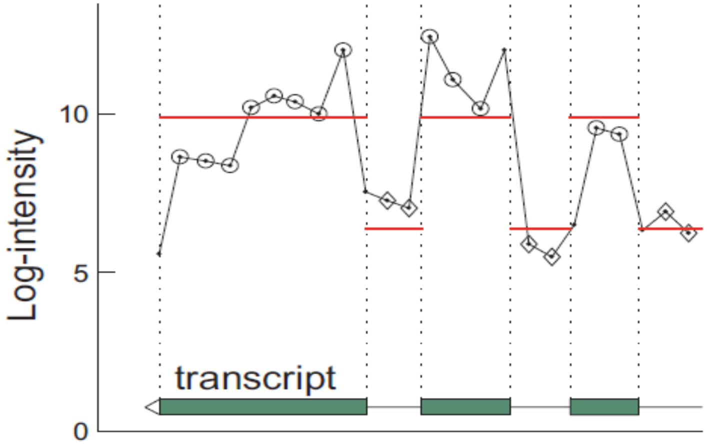
Transcriptome mapping using tiling arrays:



Chip design



Hybridization to Tiling array



Interpretation of results

Advantages of RNA-Seq:

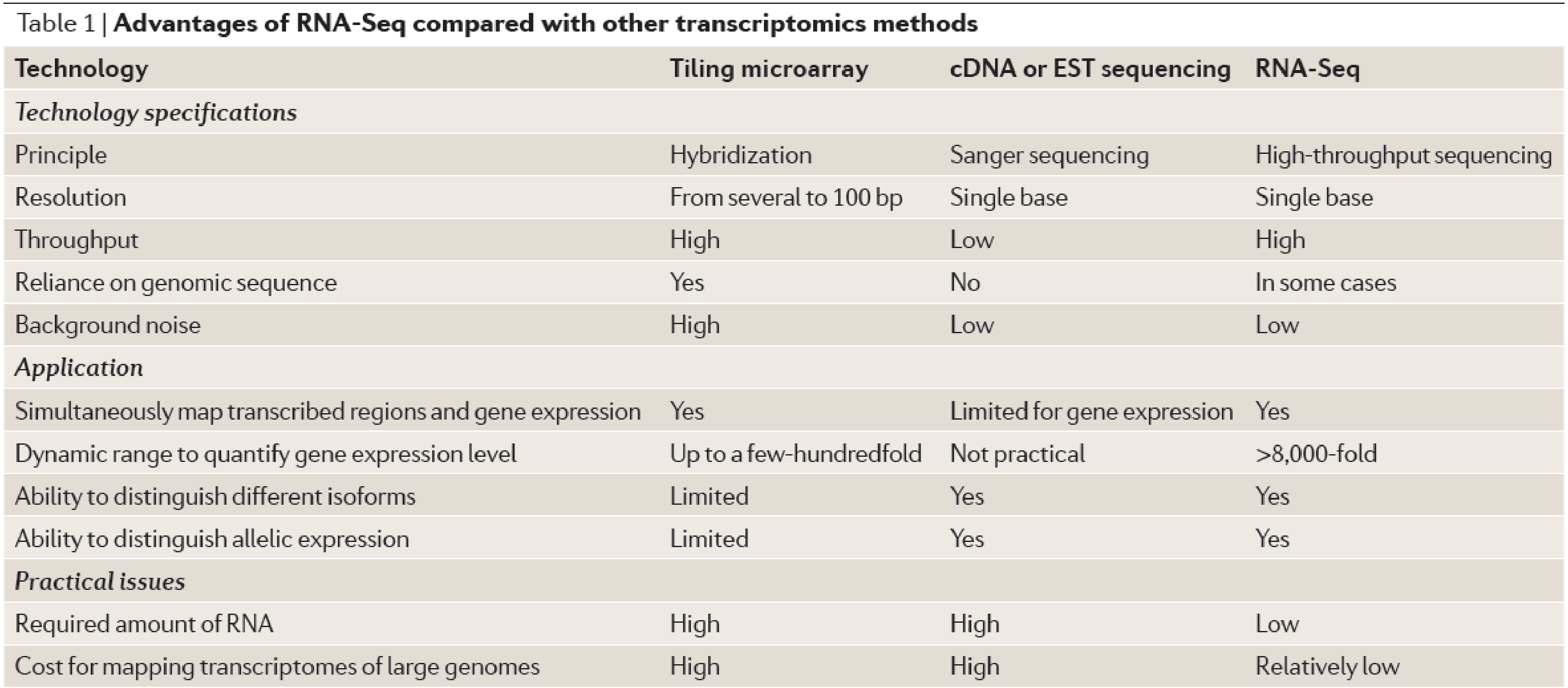


Table 1 | **Advantages of RNA-Seq compared with other transcriptomics methods**

Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

Wang Z. et al. 2009

In addition RNA-Seq can reveal sequence variation, i.e. mutations or SNPs

Advantages of RNA-Seq (II):

Background and saturation:

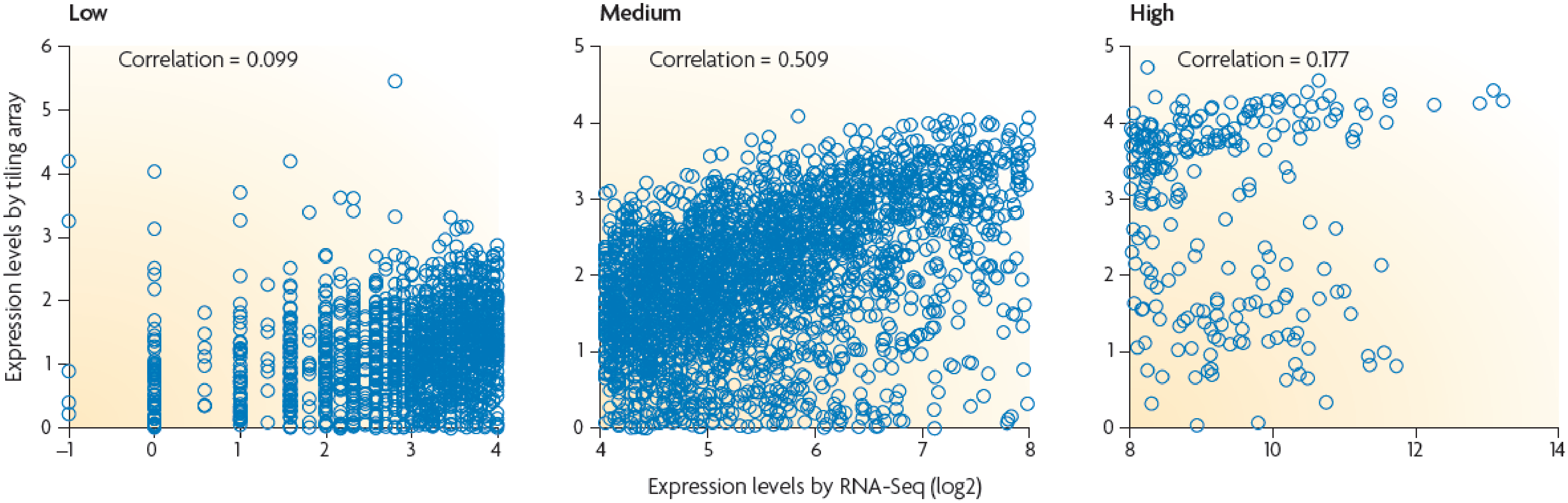


Figure 2 | **Quantifying expression levels: RNA-Seq and microarray compared.** Expression levels are shown, as measured by RNA-Seq and tiling arrays, for *Saccharomyces cerevisiae* cells grown in nutrient-rich media. The two methods agree fairly well for genes with medium levels of expression (middle), but correlation is very low for genes with either low or high expression levels. The tiling array data used in this figure is taken from REF. 2, and the RNA-Seq data is taken from REF. 18.

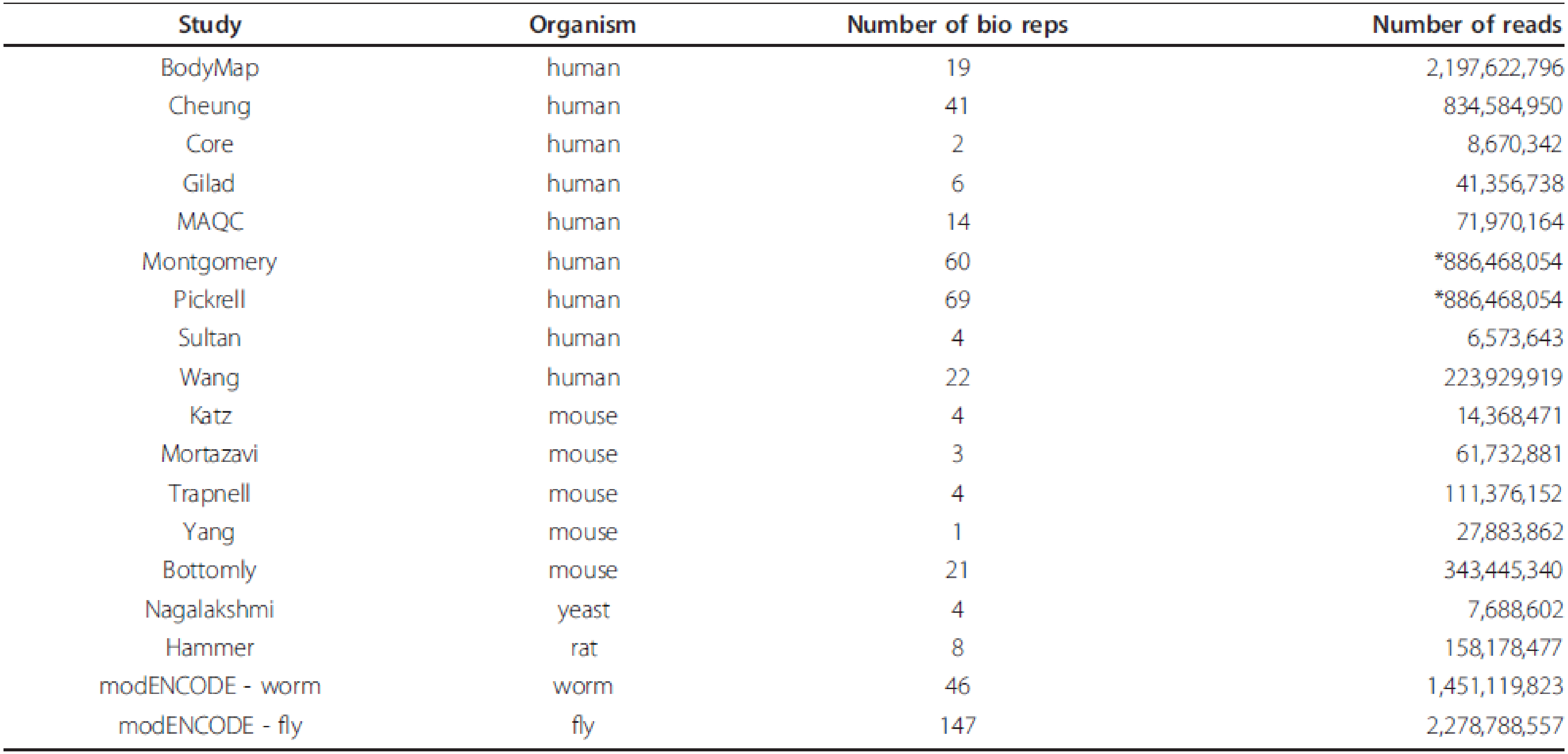
Wang Z. et al. 2009

New insights:

- More precise estimation of starts, ends and splice sites for transcripts
- Detection of novel transcribed regions
- Discovery of splicing isoforms and RNA editing
- Detection of mutations and SNPs and analysis of the influence on transcription and post-transcriptional modification

Expression quantification:

- ReCount - database:
  - Collection of preprocessed RNA-Seq data
  - **http://bowtie-bio.sf.net/recount**



Study	Organism	Number of bio reps	Number of reads
BodyMap	human	19	2,197,622,796
Cheung	human	41	834,584,950
Core	human	2	8,670,342
Gilad	human	6	41,356,738
MAQC	human	14	71,970,164
Montgomery	human	60	*886,468,054
Pickrell	human	69	*886,468,054
Sultan	human	4	6,573,643
Wang	human	22	223,929,919
Katz	mouse	4	14,368,471
Mortazavi	mouse	3	61,732,881
Trapnell	mouse	4	111,376,152
Yang	mouse	1	27,883,862
Bottomly	mouse	21	343,445,340
Nagalakshmi	yeast	4	7,688,602
Hammer	rat	8	158,178,477
modENCODE - worm	worm	46	1,451,119,823
modENCODE - fly	fly	147	2,278,788,557

Preprocessing and construction of count tables:

- For paired-end sequencing only first mate pair was considered
- Pooling of technical replicates

- Alignment using bowtie algorithm:
  - Not more than 2 mismatches per read allowed
  - Reads with multiple alignment discarded
  - Read longer than 35 bp truncated to 35 bp
  - Overlapping of alignment of reads with gene footprint from middle position of read

Example applications (I):

- Analysis of data from multiple studies
  - Comparison of the same 29 individuals from 2 studies
    - - (A) immortalized B-cells
    - - (B) lymphoblastoid cell lines
      => similar cell types

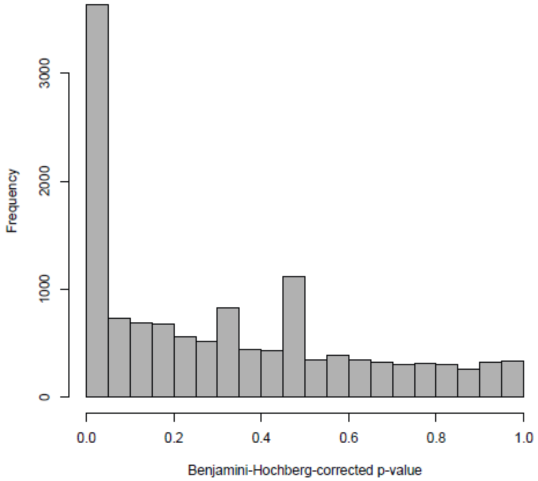


Figure 1 Histogram of adjusted p-values from differential expression analysis on the 29 samples included in both Cheung and Montgomery. The p-values in the histogram are from paired t-tests on the 25% of genes with nonzero counts in at least one of the two studies. The peak near zero is somewhat indicative of technical variability between the two studies.

- Differential gene expression
  - Paired t-test with Benjamini-Hochberg correction
  - ~28% of genes were differentially expressed

- **Evidence for dramatic batch effects!**

Example applications (II):

- Similar analysis for differential expression between different ethnicities
  - Comparison of:
    - (A) Utah resident (CEU ancestry)
    - (B) Nigeria (Yoruba ancestry)

- Differential gene expression
  - Paired t-test with Benjamini-Hochberg correction
  - ~36% of genes were differentially expressed

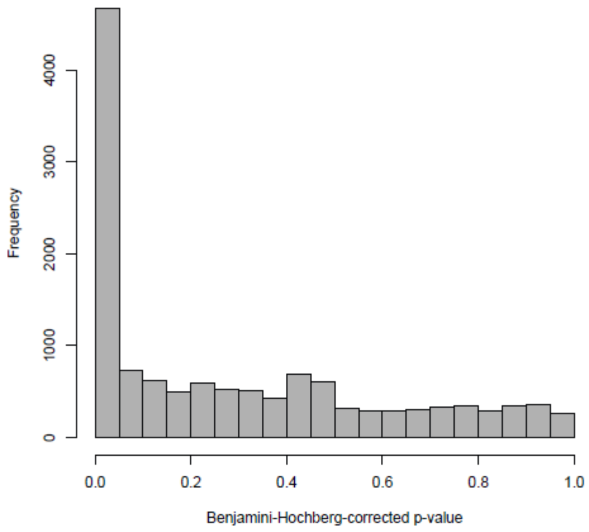


Figure 2 Histogram of adjusted p-values from analysis of differential expression between YRI and CEU populations. The p-values in the histogram are from two-sample t-tests on the 25% of genes with nonzero counts in at least one of the two studies. The peak near zero indicates differential gene expression that may result from either technical or biological variability.

- **Technical and biological variability**

Thank you for your attention!