Hospital Universitari Vall d'Hebron
Institut de Recerca - VHIR

*Institut **d'Investigació Sanitària de l'Instituto de Salud Carlos III (ISCIII)***

Introduction to RNA-seq and RNA-seq Data Analysis

Bioinformatics Course

Ferran Briansó and Alex Sánchez
ferran.brianso@vhir.org

Vall d'Hebron
Institut de Recerca

Disclaimer

- This lecture is based on many presentations freely available in the web.
- We wish to acknowledge the authors for their efforts and for making their work available

Vall d'Hebron
Institut de Recerca

Transcriptomics by NGS

Evolution of transcriptomics technologies

- Northern Blot
- RT-PCR
- Microarrays
- (NGS) RNA-seq

- Single Genes
- Multiple genes
- Whole Genomes
- Populations of genomes

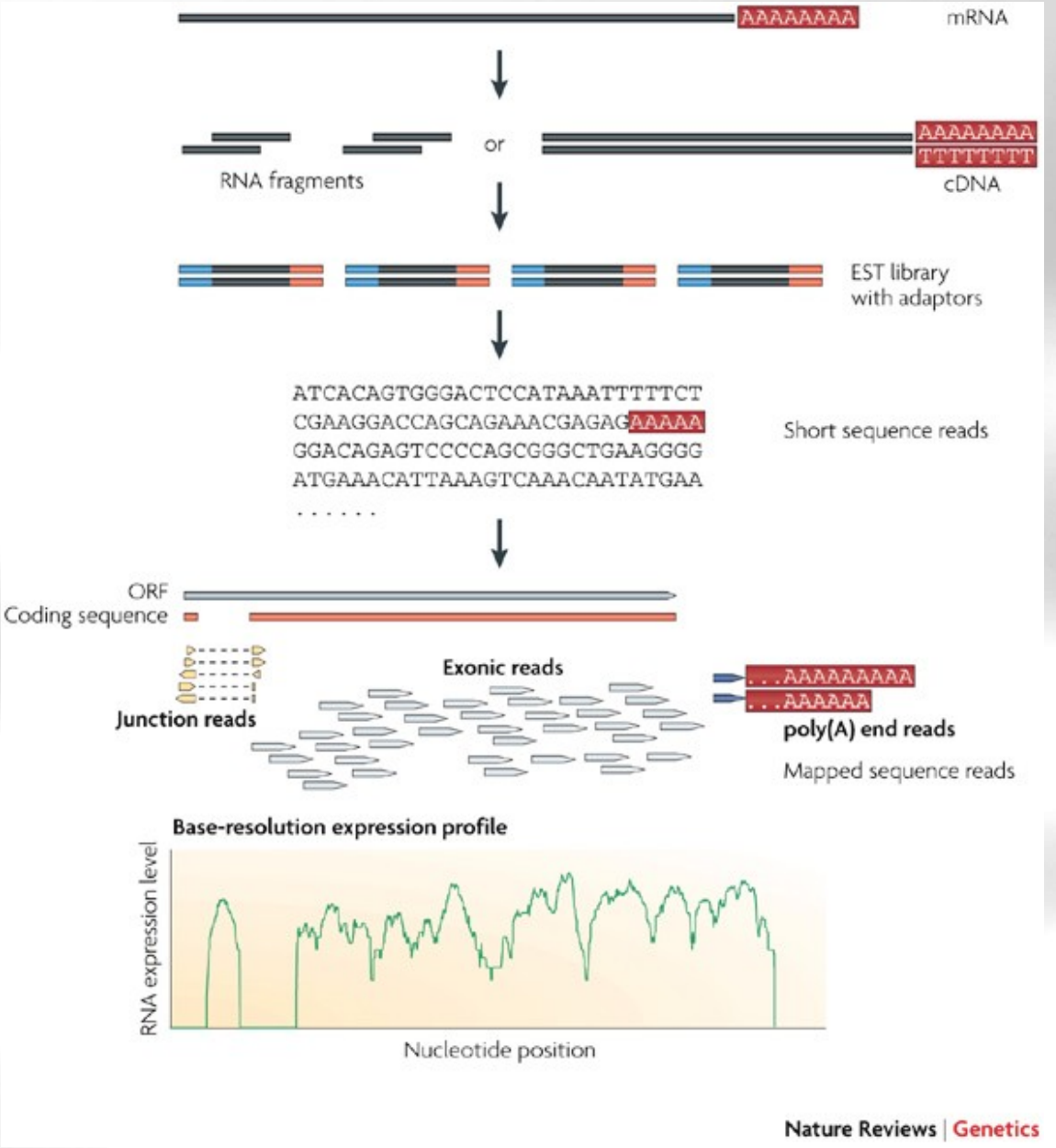
What is RNA-seq?

- RNA-seq is the high throughput sequencing of cDNA using NGS technologies
- RNA-seq works by sequencing every RNA molecule and profiling the expression of a particular gene by counting the number of time its transcripts have been sequenced.
- The summarized RNA-Seq data is widely known as *count data*

	Condition A			Condition B		
Gene1	4	0	2	12	14	13
Gene2	0	23	50	47	22	0
Gene3	0	2	6	13	11	15
...	...	...	...	...	...	...
GeneG	156	238	37	129	51	118

A typical RNA-seq experiment

Briefly, long RNAs are first converted into a library of cDNA fragments through either RNA fragmentation or DNA fragmentation (see main text). Sequencing adaptors (blue) are subsequently added to each cDNA fragment and a short sequence is obtained from each cDNA using high-throughput sequencing technology. 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. These three types are used to generate a base-resolution expression profile for each gene, as illustrated at the bottom; a yeast ORF with one intron is shown.



Nature Reviews Genetics **10**, 57-63 (January 2009)

Vall d'Hebron
Institut de Recerca

Why use RNA-seq

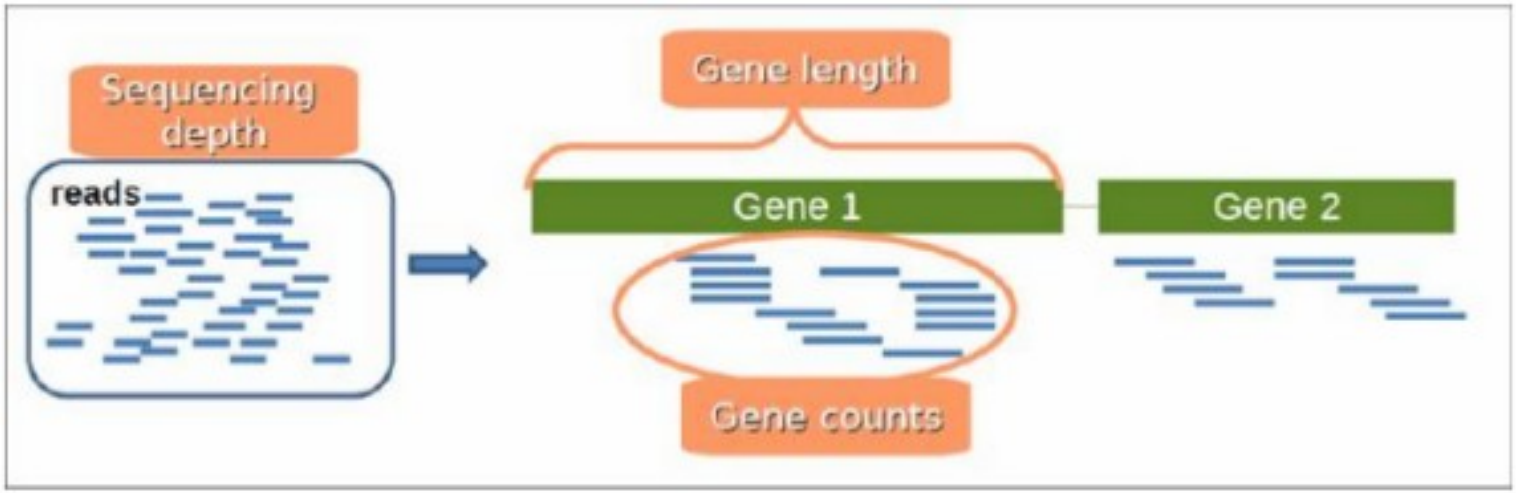
- Unique new possibilities
  - Evaluate absolute transcript level of sequenced and unsequenced organisms.
  - Detect novel transcripts and isoforms
  - Map exon/intron boundaries, splice junctions
  - Analyze alternative splicing
  - Reveal sequence variations (e.g. SNPs) and splice variants

It has limitations too ...

- Non-uniformity coverage of the genome due to experimental factors
- Transcript-length bias
- Read mapping uncertainty caused by sequencing error rates, repetitive elements, incomplete genome sequence, etc
- Downstream bioinformatics algorithm/software need to be improved
- Cost more than microarray

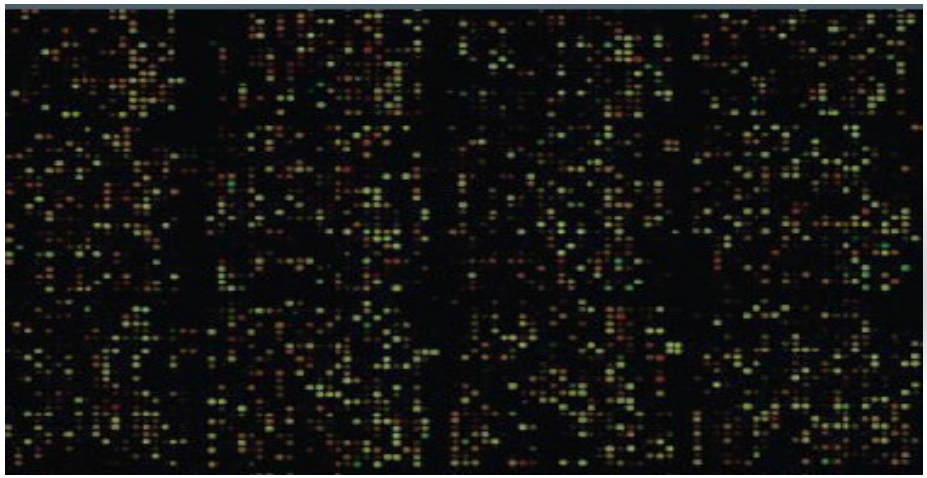
Important concepts

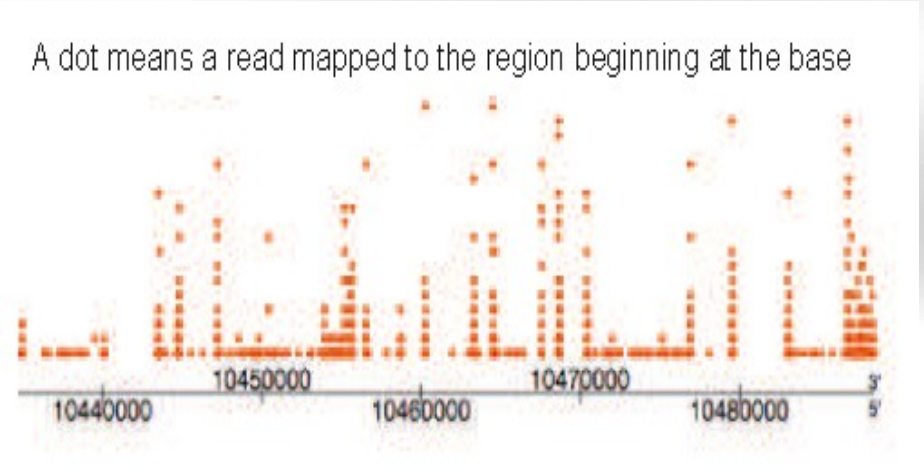
- Sequencing depth (Library size): Total number of reads mapped to the genome.
- Gene length: Number of bases that a gene has.
- Gene counts: Number of reads mapping to that gene (expression measurement).



Microarrays vs NGS

RNA-seq can be seen as the NGS-counterpart of microarrays





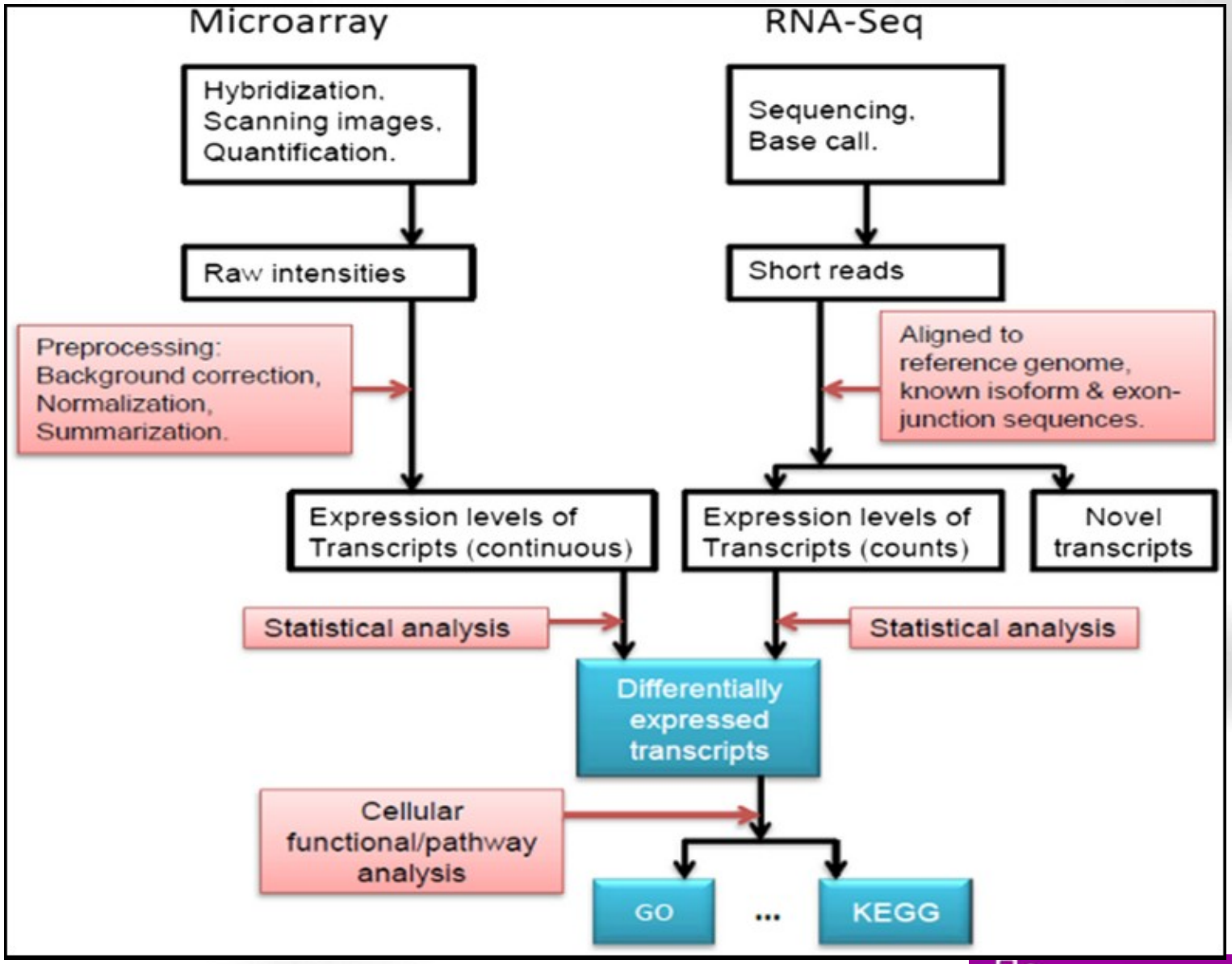
- Analog Signal

  - Easy to convey the signal's information
  - Continuous strength
  - Signal loss and distortion

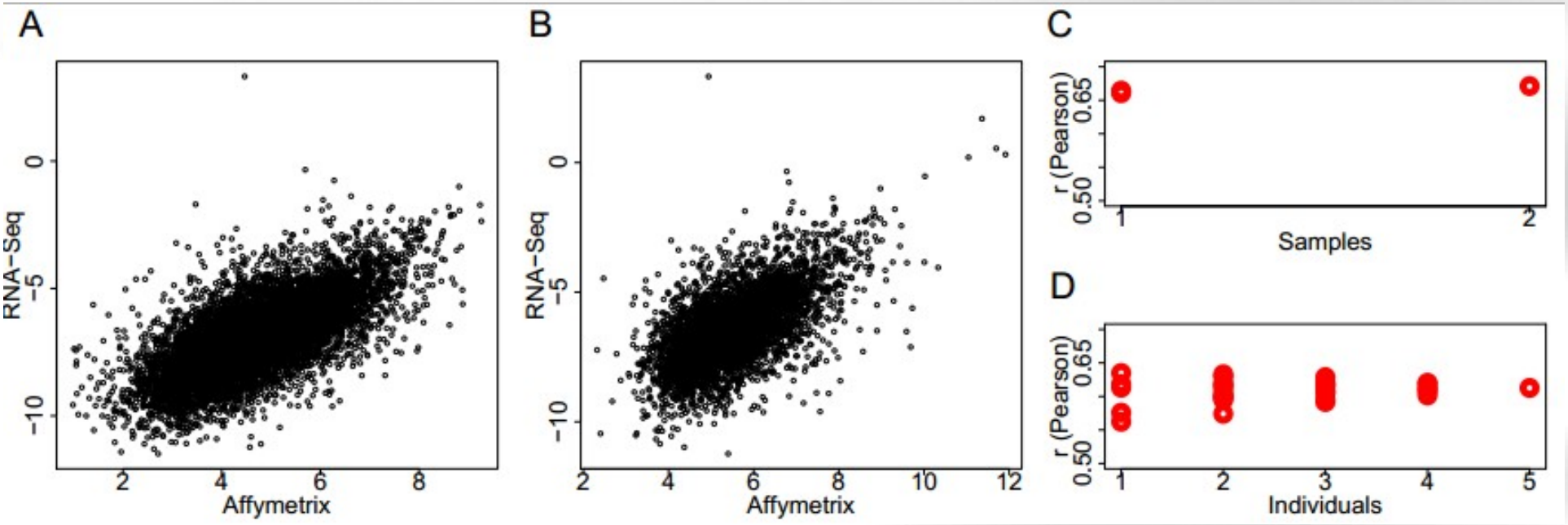
- Digital Signal

  - Harder to achieve & interpret
  - Reads counts: discrete values
  - Weak background or no noise

Microarrays and NGS pipelines



RNA seq and microarrays yield correlated results



Vall d'Hebron
Institut de Recerca

Pros and cons of RNA-seq and microarrays

Microarrays

- Costs,
- well established methods, small data

- Hybridization bias,
- sequence must be known

RNA-seq

- High reproducibility,
- not limited to expression

- Cost
- Complexity of analysis

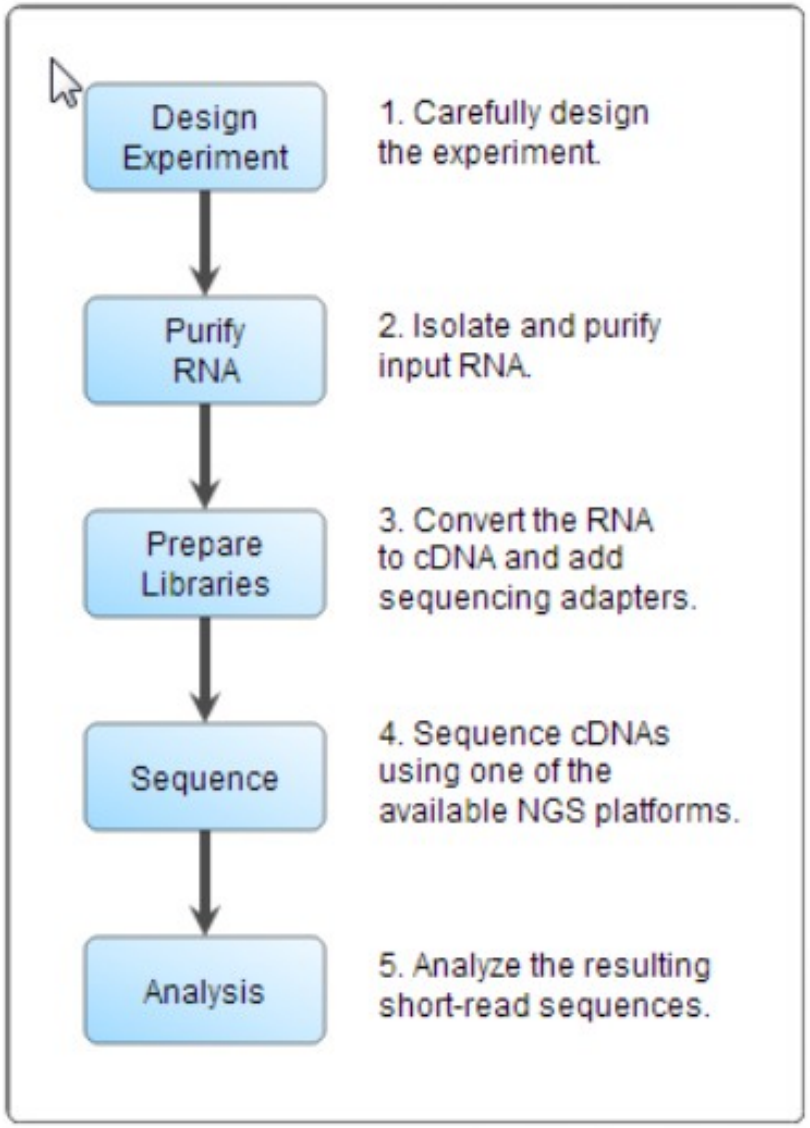
So what?

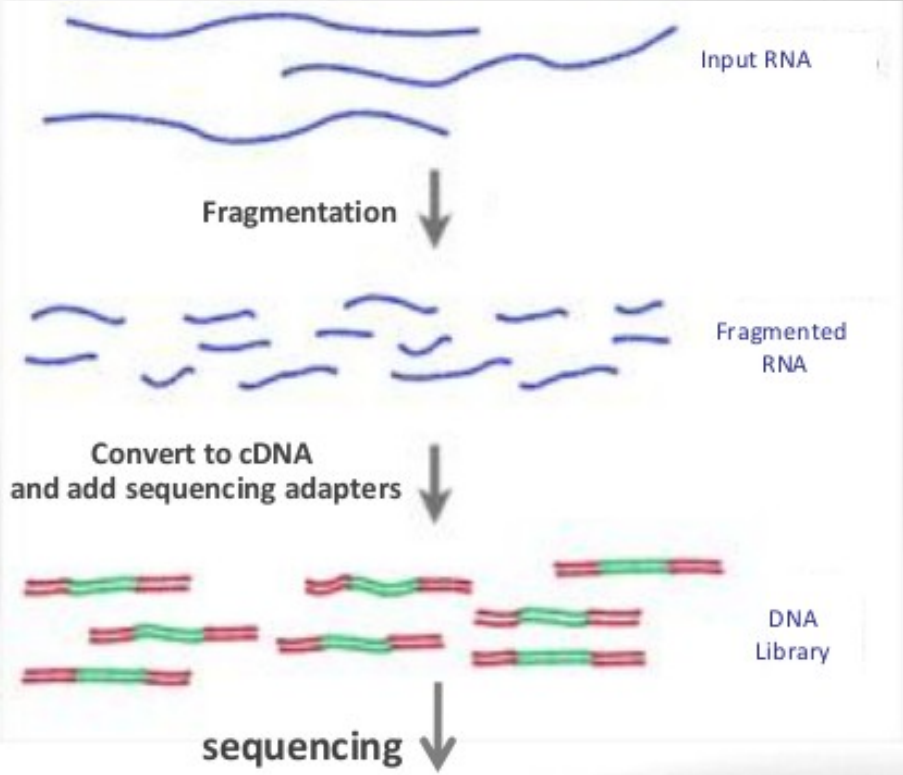
- It is generally agreed/believed/expected that RNA-seq will *soon* replace microarrays for many uses. But not for all uses
- There are still situations where the “simplicity” of microarrays yields the necessary information at an optimal cost.
  - Microarrays are now part of the standard molecular biology toolbox whereas RNA-seq is still in development.

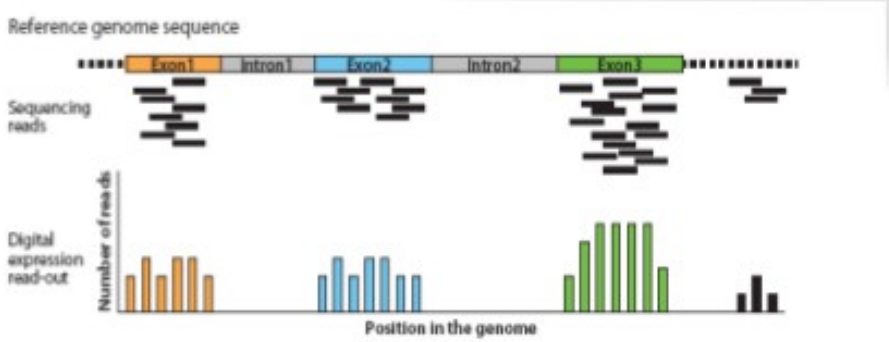
RNA-seq analysis

(workflows, pipelines, protocols…)

RNA-seq analysis workflow

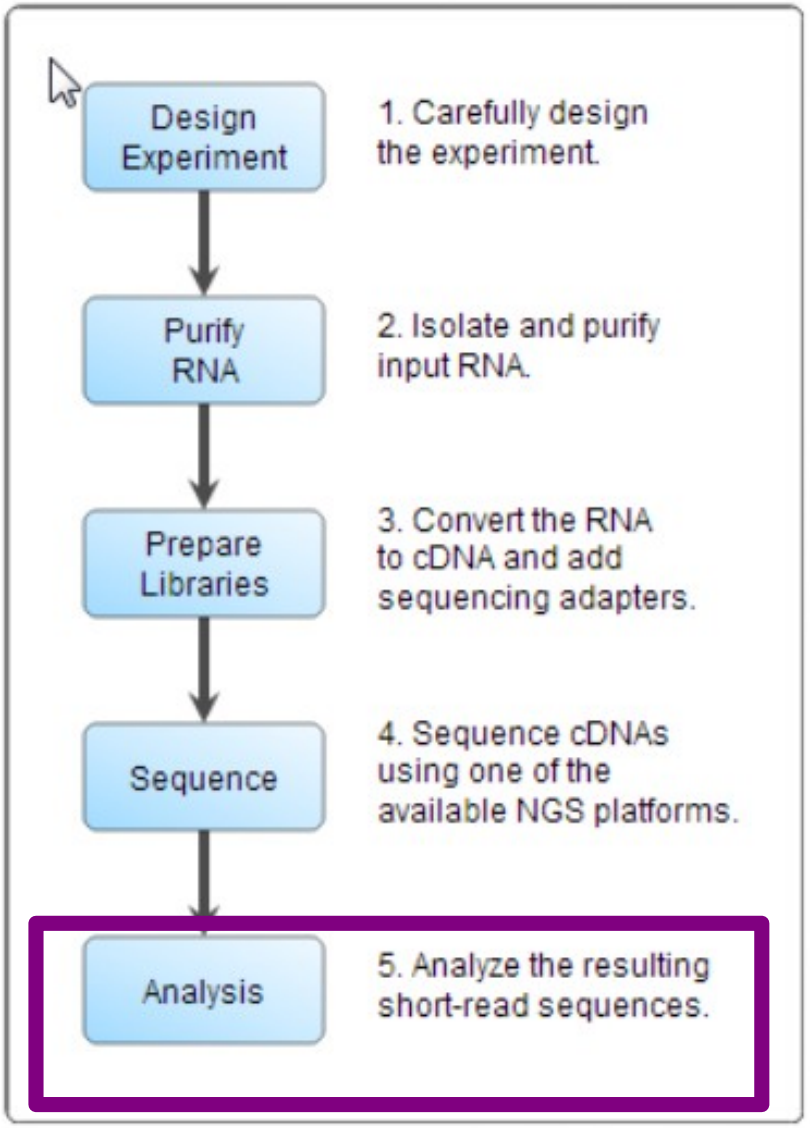


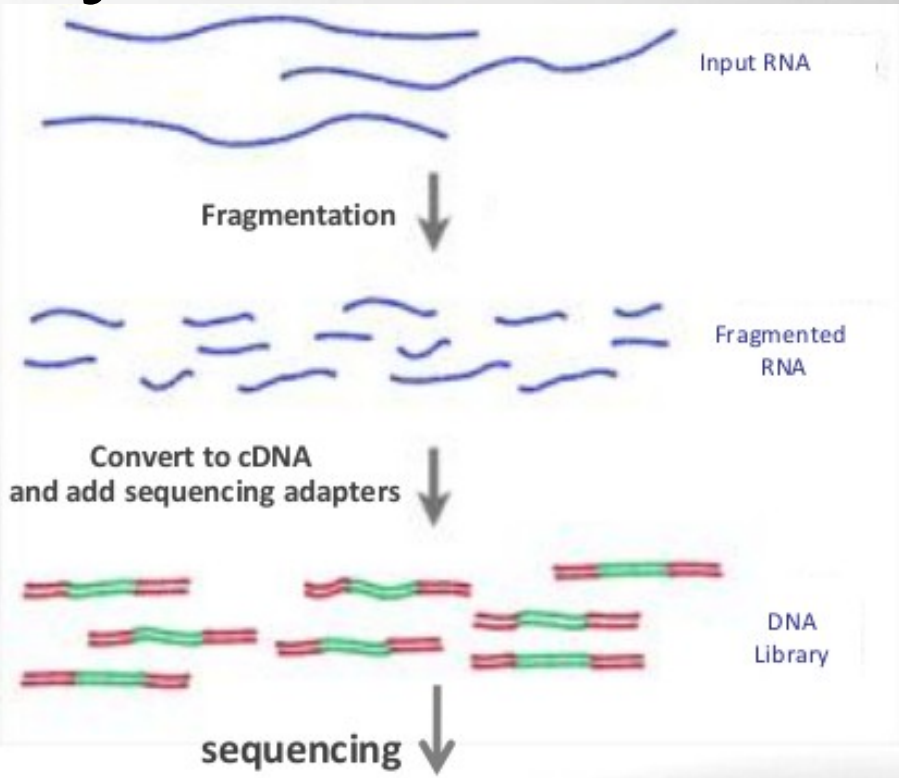


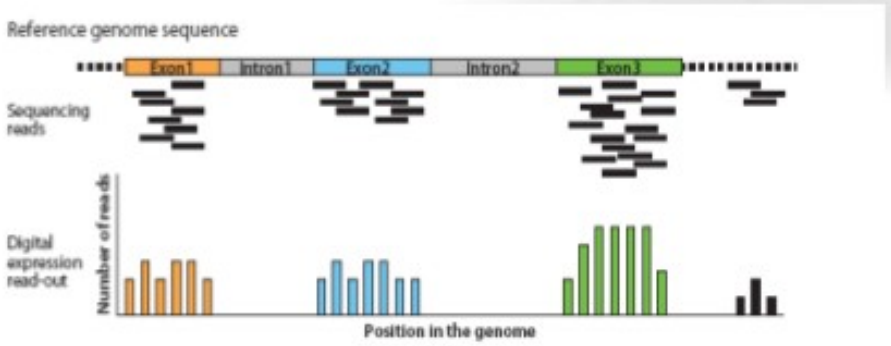


Vall d'Hebron
Institut de Recerca

RNA-seq analysis workflow



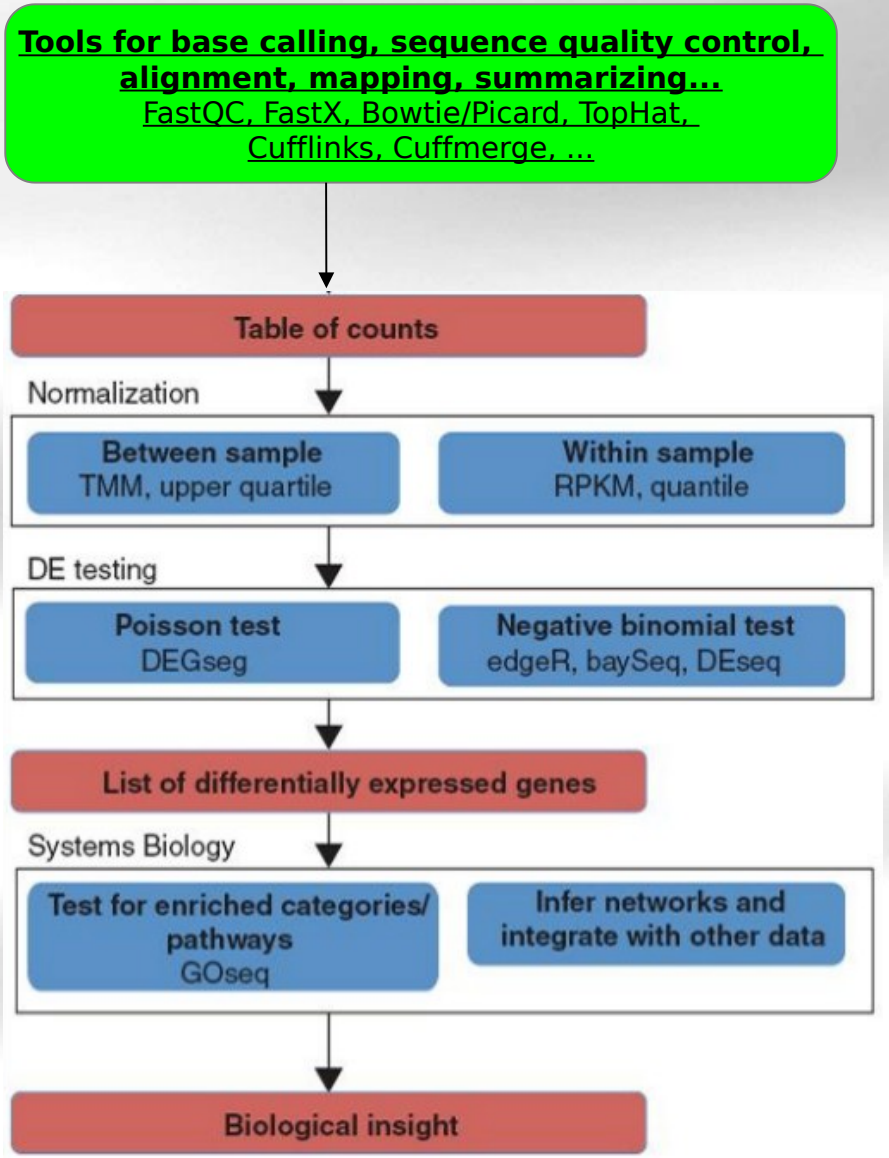




Vall d'Hebron
Institut de Recerca

RNA-seq analysis workflow

- **Reads are mapped to the reference genome or transcriptome**
- **Mapped reads are assembled into expression summaries (tables of counts, showing how may reads are in coding region, exon, gene or junction)**
- **Data is normalized**
- **Statistical testing of differential expression (DE) is performed, producing a list of genes with p-values and fold changes.**
- **Similar downstream analysis than microarray results (Functional Annotations, Gene Enrichment Analysis; Integration with other**



RNA Seq data analysis (1)-Mapping

Main Issues:

- Number of allowed mismatche
- Number of multihits
- Mates expected distance
- Considering exon junctions

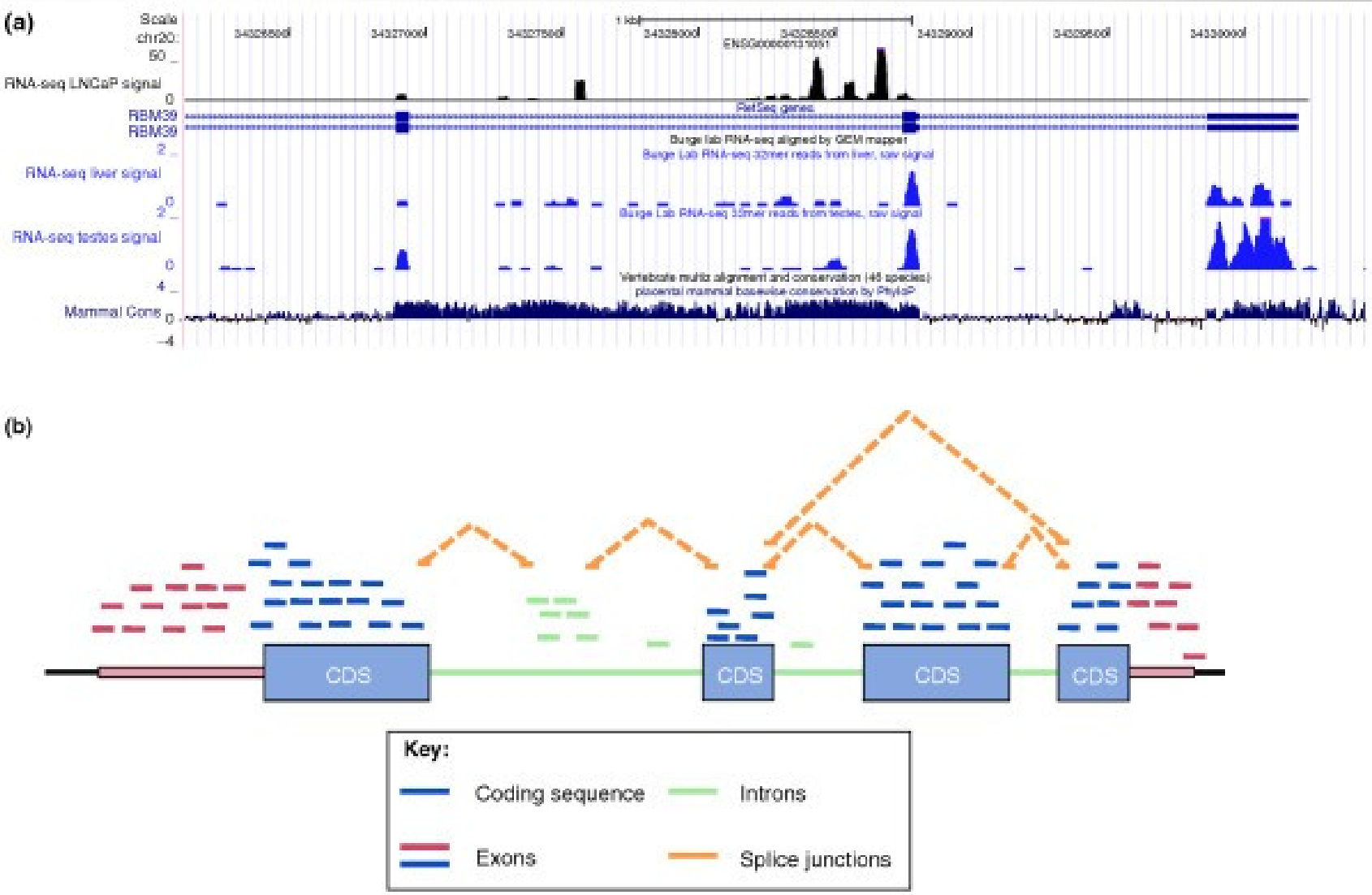
End up with a list of
# of reads per transcript

These will be our (discrete)
response variable

RNA Seq data analysis (2)-Summarization

- Sequencing ☾ genomic locations of **many** reads
- Next task : Summarize & aggregate reads over *some biologically meaningful unit*, such as exons, transcripts or genes.
- Many methods available
  - Counts # of reads overlapping the exons in a gene,
  - Include reads along the whole length of the gene and thereby incorporate reads from 'introns'.
  - Include only reads that map to coding sequence or…

RNA Seq data analysis (2)-Summarization



RNA Seq data analysis (3)-Normalization

- Two main sources of bias
  - Influence of **length**: Counts are proportional to the transcript length times the mRNA expression level.
  - Influence of **sequencing depth**: The higher sequencing depth, the higher counts.

- How to deal with this
  - **Normalize** (correct) gene counts to minimize biases.
  - Use **statistical models** that take into account l**ength** and **sequencing depth**

RNA-seq normalization methods

- RPKM (Mortazavi et al., 2008): Counts are divided by the transcript length (kb) times the total number of millions of mapped reads.

$$\text{RPKM} = \frac{\dfrac{\text{number of reads in region}}{\text{region length}\times 10^3}}{\text{total reads} \times 10^6}$$

- TMM (Robinson and Oshlack, 2010): Trimmed Mean of M values.

- EDAseq (Risso et al., 2011): Within-lane gene-level GC-content normalization (corrects for library size, gene length, GC-content)

- cqn (Hansen et al., 2011): Conditional quantile normalization (CQN) algorithm combining robust generalized regression (corrects for library size, gene length, GC-content)

- Others: Upper-quartile (Bullard et al., 2010); FPKM (Trapnell et al., 2010): Instead of counts, Cufflinks software generates FPKM values (Fragments Per Kilobase of exon per Million fragments mapped) to estimate gene expression, which are analogous to RPKM.

RNA Seq (4)- Differential expression analysis

- The goal of a DE analysis is to *highlight genes that have changed significantly in abundance across experimental conditions*.
- In general, this means
  - taking a table of summarized count data for each library and
  - performing statistical testing between samples of interest

RNA Seq (4)- Methods for Differential Expression Analysis

- Transform count data to use existing approaches for microarray data.
- Use Fisher's exact test or similar approaches.
- Use statistical models appropriate for count data such as *Generalized Linear Models* using
  - Poisson distribution.
  - Negative binomial distribution.

RNA-Seq example: Marioni *et al.* (2008)

Human LIVER and KIDNEY samples.

Illumina (5 lanes per sample). 32-bp reads. Mapping allowing at most two missmatches.

Sequencing depth:

Kidney = 9.293.530 reads

Liver = 8.361.601 reads

Which genes are differentially expressed in kidney and liver?

EnsemblGeneID	kidney	liver
ENSG00000146556	0	0
ENSG00000197194	0	0
ENSG00000197490	0	0
ENSG00000205292	0	0
ENSG00000177693	0	0
ENSG00000209338	0	0
ENSG00000196573	0	0
ENSG00000177799	0	0
ENSG00000209341	0	0
ENSG00000209342	9	1
ENSG00000209343	0	0
ENSG00000209344	0	0
209346	0	0
349	0	0
50	27	161
351	0	0
00209352	2	2
ENSG00000212679	620	746
ENSG00000212678	64591	44870
ENSG00000185097	0	0
ENSG00000209353	0	0
ENSG00000197049	1	1
ENSG00000215918	0	0
ENSG00000177757	8	3

RNA seq example (Normalized values)

EnsemblGeneID	Length	kidney	liver	RPKM kidney	RPKM liver	UQUA kidney	UQUA liver
ENSG00000187642	3035	39	7	1.38	0.28	382.37	74.66
ENSG00000188290	877	54	9	6.63	1.23	981.59	182.39
ENSG00000187608	634	59	63	10.01	11.88	1252.58	1484.31
ENSG00000188157	7353.5	2108	259	30.85	4.21	13193.25	1796.14
ENSG00000131591	2039.83	54	34	2.85	1.99	637.77	445.46
ENSG00000215916	2008	57	34	3.05	2.03	678.67	448.97
ENSG00000207730	95	0	0	0	0	0	0
ENSG00000207607	90	0	0	0	0	0	0
ENSG00000198976	83	2	0	2.59	0	113.62	0
ENSG00000205231	3532	4	0	0.12	0	35.71	0
ENSG00000162571	2060.25	4	0	0.21	0	47.82	0
ENSG00000186891	964	0	3	0	0.37	0	57.78
ENSG00000186827	987	5	1	0.55	0.12	86.6	19.42
ENSG00000078808	1870.67	1136	883	65.34	56.45	14095	12172.11
ENSG00000176022	2793	143	165	5.51	7.07	1453.42	1853.37
ENSG00000184163	1036	16	14	1.66	1.62	268.45	258.3
ENSG00000160087	1614.14	315	290	21	21.49	4220.61	4320
ENSG00000162572	2635.86	47	28	1.92	1.27	489.42	323.8
ENSG00000131584	3861.75	379	216	10.56	6.69	3281.75	2076.27
ENSG00000169972	1239	105	143	9.12	13.8	1611.32	2423.83
ENSG00000127054	1813.5	496	330	29.43	21.76	6250.88	4618.39
ENSG00000187488	2193	17	6	0.83	0.33	194.62	76.17
ENSG00000215792	2193	17	6	0.83	0.33	194.62	76.17
ENSG00000169962	3402	2	0	0.06	0	18.2	0
ENSG00000107404	2554	51	16	2.15	0.75	542.14	188.61
ENSG00000162576	2169.67	56	11	2.78	0.61	645.79	138.29
ENSG00000175756	874.25	0	0	0	0	0	0
ENSG00000131586	676	9	0	1.43	0	187.59	0
ENSG00000205116	492	7	0	1.53	0	175.4	0
ENSG00000179403	2442.5	53	46	2.33	2.25	578.29	553.48
ENSG00000215915	2800.5	0	4	0	0.17	0	46.38
ENSG00000160072	1718.2	16	17	1	1.18	206.19	243.58
ENSG00000197785	2530	96	117	4.08	5.53	1023.05	1382.85

RNA seq example – Analysis (Fisher test)

Better using normalized values, e.g. RPKM.

Conservative when expression values are close to 0.

	KIDNEY	LIVER	Total
ENSG00000188157	30.85	4.21	35.06
Remaining genes	809347.05	799467.99	1608815.04
Total	809377.90	799472.20	

```
> cont.table <- matrix(data = c(31, 4, 809347, 799468), nrow = 2, ncol = 2)
> cont.table
     [,1]   [,2]
[1,]   31 809347
[2,]    4 799468
> fisher.test(cont.table)
```

```
        Fisher's Exact Test for Count Data
data:  cont.table
p-value = 3.511e-06
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 2.706480 29.858019
sample estimates:
odds ratio
  7.65533
```

This gene is differentially expressed in both tissues

REMEMBER: When computing a test per gene, gene p-values must be adjusted for multiple testing correction.

Statistical models for count data

- The number of reads that are mapped into a gene was first modelled using a Poisson distribution
  - Poisson distribution appears when things are counted
  - It assumes that mean and variance are the same

- However biological variability of RNA-seq count data cannot be captured using the Poisson distribution because data present overdispersion (i.e., variance of counts larger than mean)
  - Negative Binomial (NB) distribution takes into account overdispersion; hence, it has been used to model RNA-seq data
  - Poisson distribution has only one parameter λ, while NB is a two-parameter distribution λ and φ.

Analysis methods based on assuming statistical models

- Basic analysis methods use the exact test approach:
  - for each gene (t = 1, ...), and groups A and B, $H_0 : \lambda_{tA} = \lambda_{tB}$
- There are better options if data are assumed to follow a Negative Binomial Distribution or some generlization of this.
  - edgeR allows the option of estimating a different φ parameter for each gene
  - baySeq uses Poisson-Gamma and BN models estimating parameters by bootstrapping from the data.
  - DESeq assumes that the mean is a good predictor of the variance.

RNA Seq (5)-Going beyond gene lists (1)

- DE analysis yields lists of *differentially expressed* genes [transcripts, …]
- Traditionally these lists arae explored by some type of *gene set analysis*
- RNA-seq has biases (e.g. due to gene length) that require adapting methods developed with microarray
  - GO-Seq is such a method

RNA Seq (5)-Going beyond gene lists (2)

- Results of RNA-seq data can be ***integrated with other sources of biological data*** e.g. to establish a more complete picture of gene regulation
  - RNA-seq has in conjunction with genotyping data identify genetic loci responsible for variation in gene expression between individuals
  - Integration of expression data & epigenomic information (transcription factor binding, histone modification, methylation) has the potential for greater understanding of regulatory mechanisms.

Additional topics

Additional topics

- Transcriptome assembly
- Alignment methods and tools
- Alternative splicing and isoforms

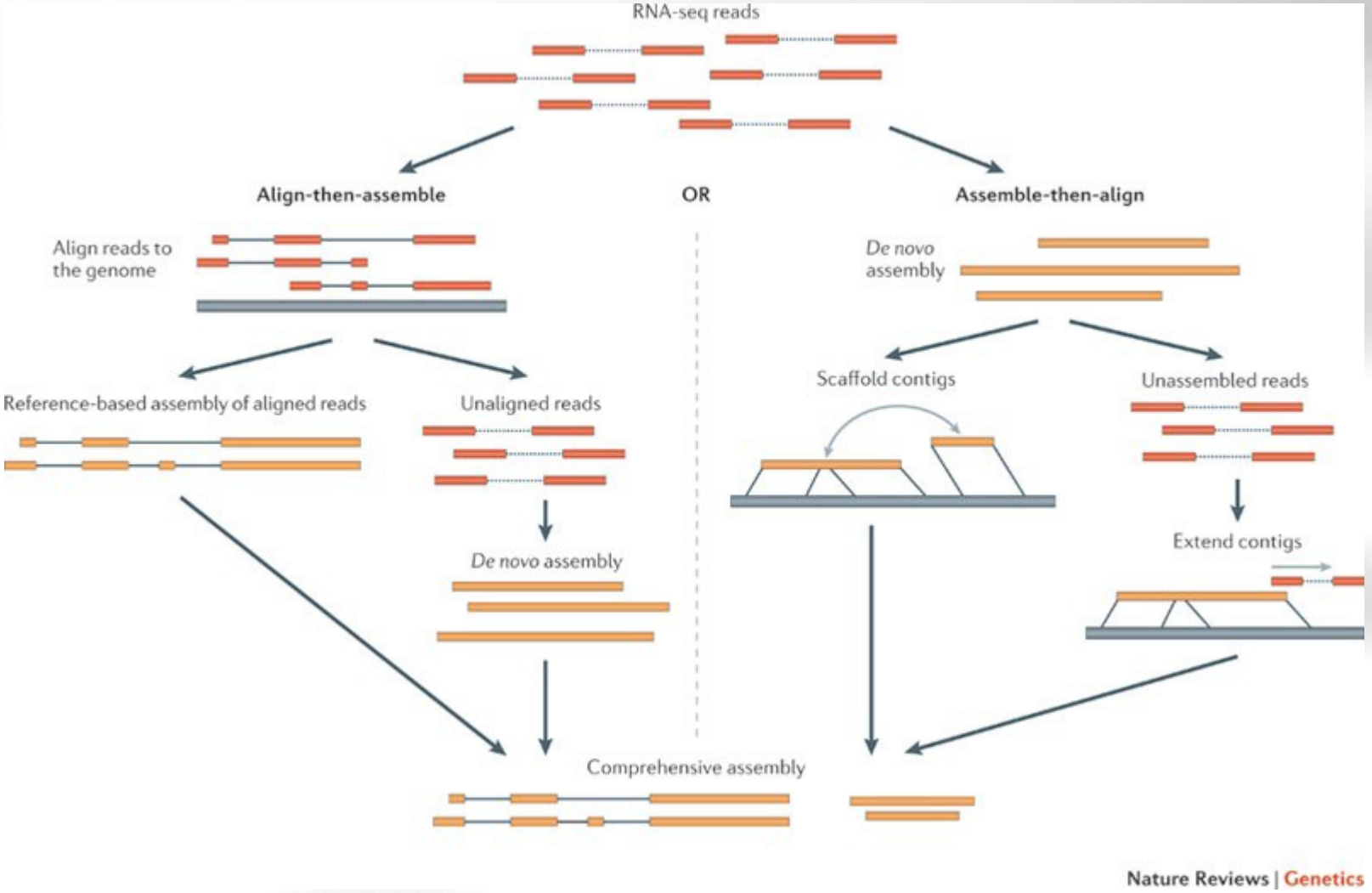
- List of software methods and tools for differential expression analysis of RNA-seq

http://genomebiology.com/2010/11/12/220/table/T1

De novo assembly

- Underlying assumptions relative to RNA expression
  - sequence coverage is similar in reads of the same transcript
  - strand specific (sense and antisense transcripts)
- Assemblers:
  - Velvet (Genomic and transcriptomic)
  - Trinity (Transcriptomic)
  - Cufflinks (Transcriptominc, reassemble pre-aligned transcripts to find alternative splicing based on differential expression)

Transcriptome assembly



Alignment methods

- Two different approach are possible:
  - Align vs the transcriptome
  - faster, easier
- Align vs the whole genome
  - the complete information

Alignment tools

- NGS common alignment program:
  - BWA
  - Bowtie (Bowtie2)
  - Novoalign
- Take into account splice-junction
  - Tophat/Cufflinks