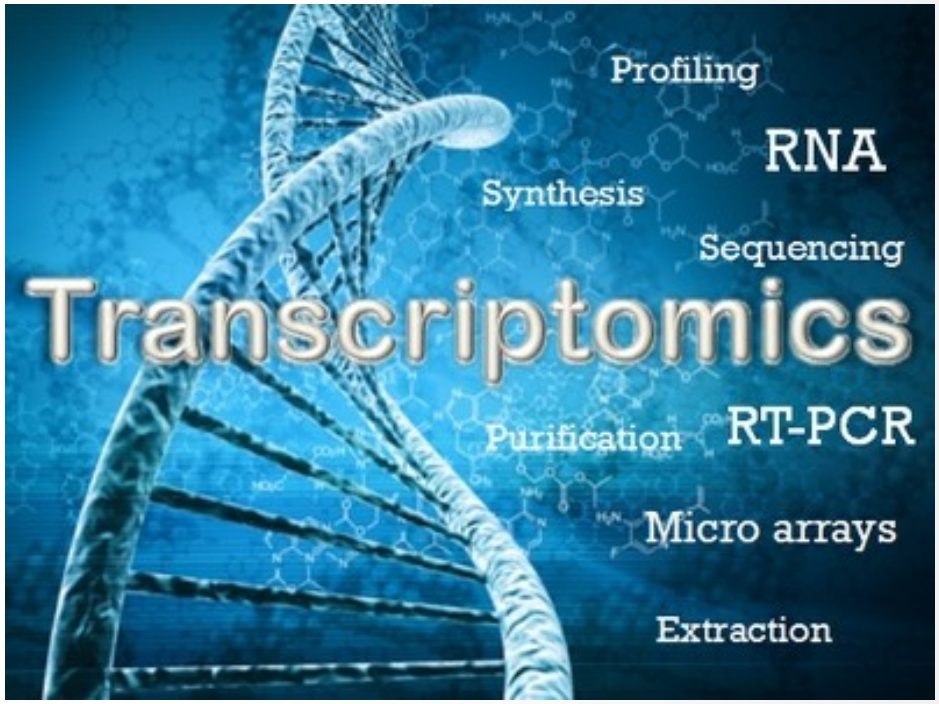
Genomics and transcriptomics II

High-throughput methods

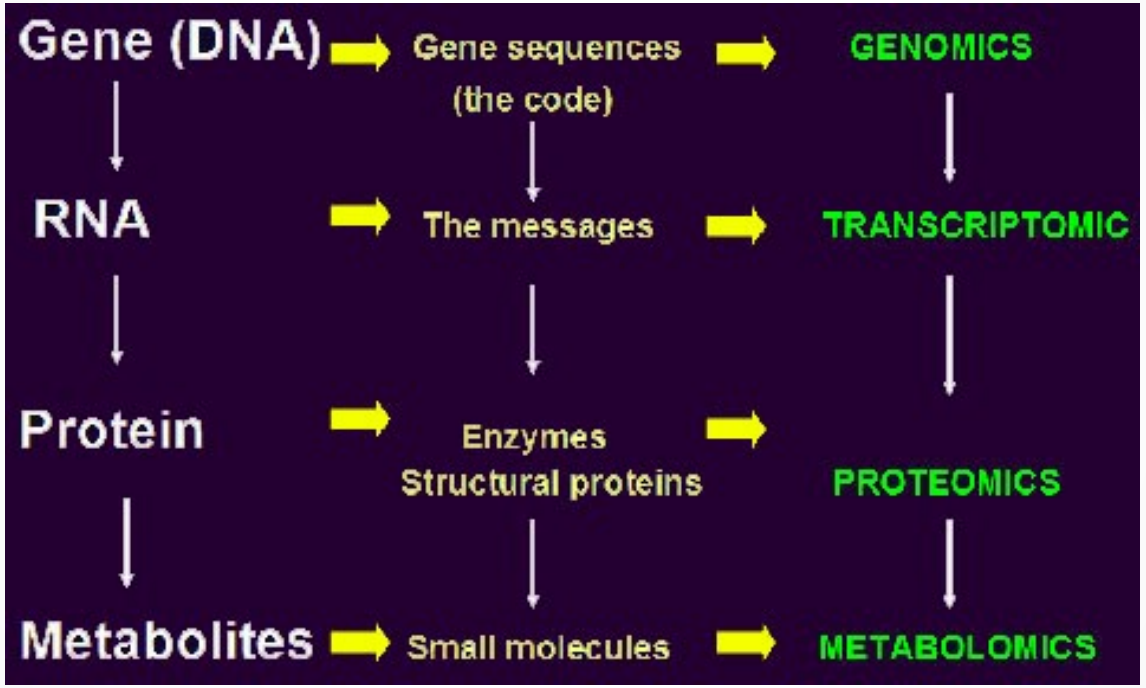
Eszter Ari
Dept. of Genetics
Eötvös Loránd Univ.
Budapest, Hungary
arieszter@gmail.com

Topics

- Transcriptomics
  - Applications
  - The microarray technology
  - RNA-Seq and its analysis
  - Differential expression analysis



Omics


Gene (DNA)
Gene sequences
(the code)
GENOMICS
RNA
The messages
TRANSCRIPTOMIC
Protein
Enzymes
Structural proteins
PROTEOMICS
Metabolites
Small molecules
METABOLOMICS

Transcriptomics

- Transcriptome:
  - the entire repertoire of transcripts in a species
  - or cells, organs, individuals, populations, etc.
  - at a specific time or under a specific set of conditions...
  - represents a key link between information encoded in DNA and phenotype
- Types of different RNAs:
  - mRNA, rRNA, tRNA
  - Post-transcriptional modificators: small nuclear snRNA, small nucleolar snoRNA, …
  - RNA regulators: micro miRNA, piwi-interacting piRNA, small interfering siRNA

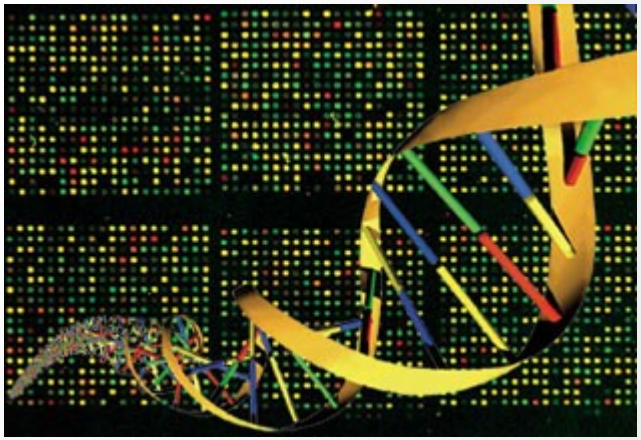
Transcriptomics

- Basis: the amount of mRNA indicates the level of gene expression and it correlates with the protein level.
- We can compare the gene expression of different cells, tissues, individuals, populations
- We can investigate the effects of different environments on gene expression
- These helps us to understand the underlying biological processes

Applications

- Genetics
  - Gene functions and regulation
- Genomics
  - Location of genes
- Systems biology
  - Co-expression networks
- Population genetics
  - Differentially expressed genes between populations
- Medical science
  - diagnostics
  - therapeutics
- Drug design

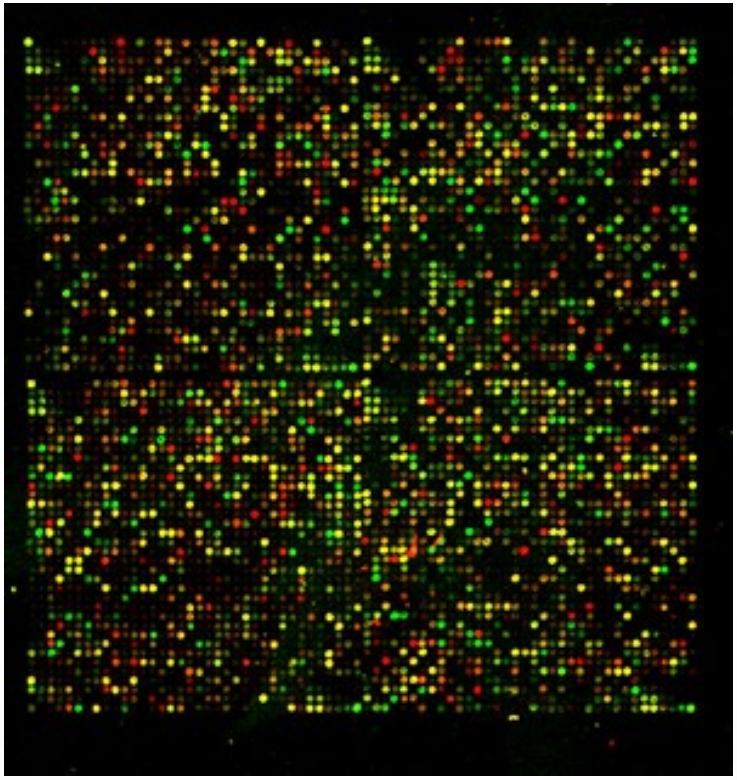
...



Methods

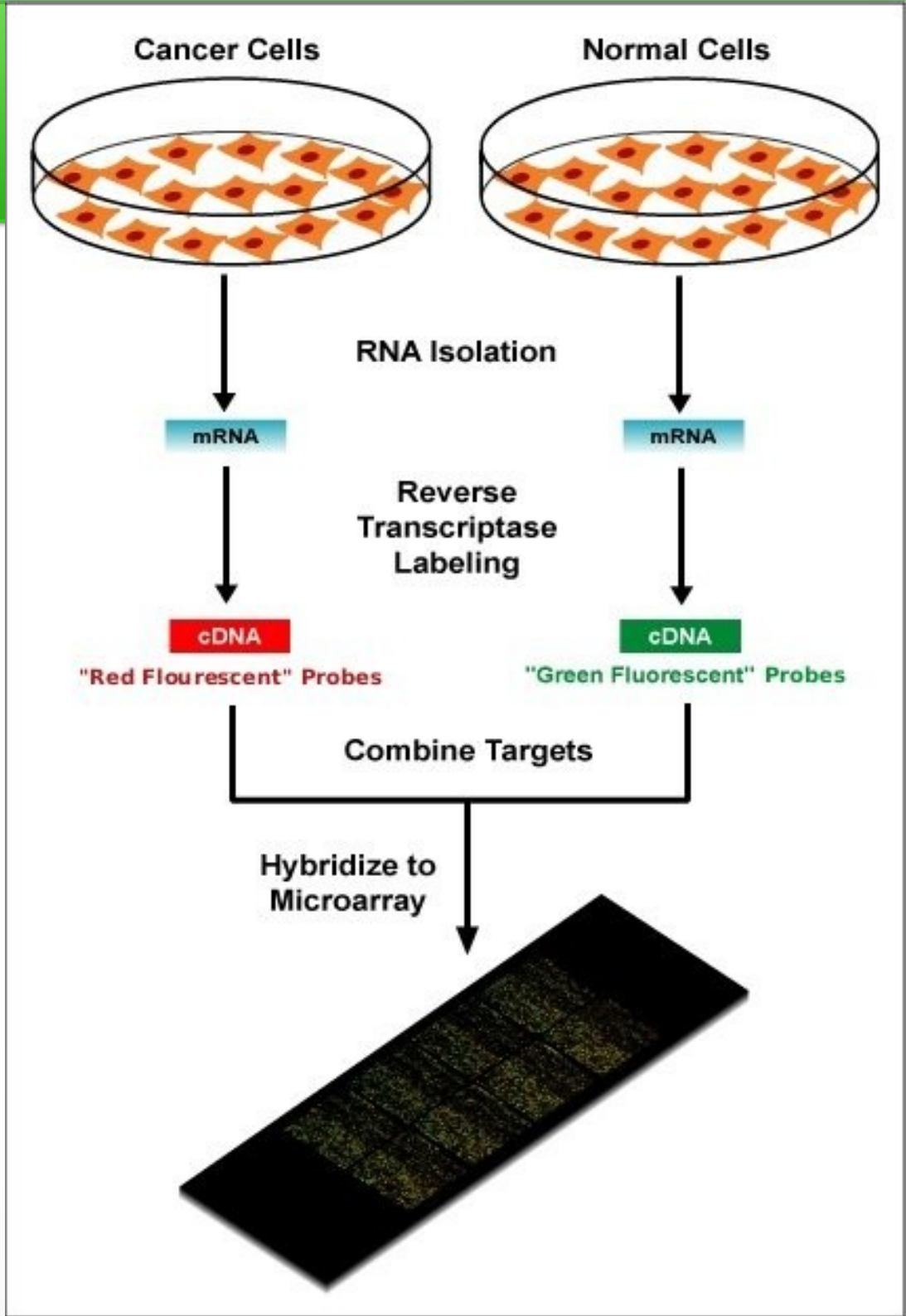
- What can we measure?
  - Levels of RNAs
  - Levels of proteins
- How?
  - Northen blot (1977)
  - reverse-transcription RT-PCR (1992)
  - Real-Time quantitative qRT-PCR
  - high-throughput methods
    - RNA Microarray or CHIP (1999)
    - High throughput sequencing - RNA-Seq (2008)
    - Protein-array

Microarray



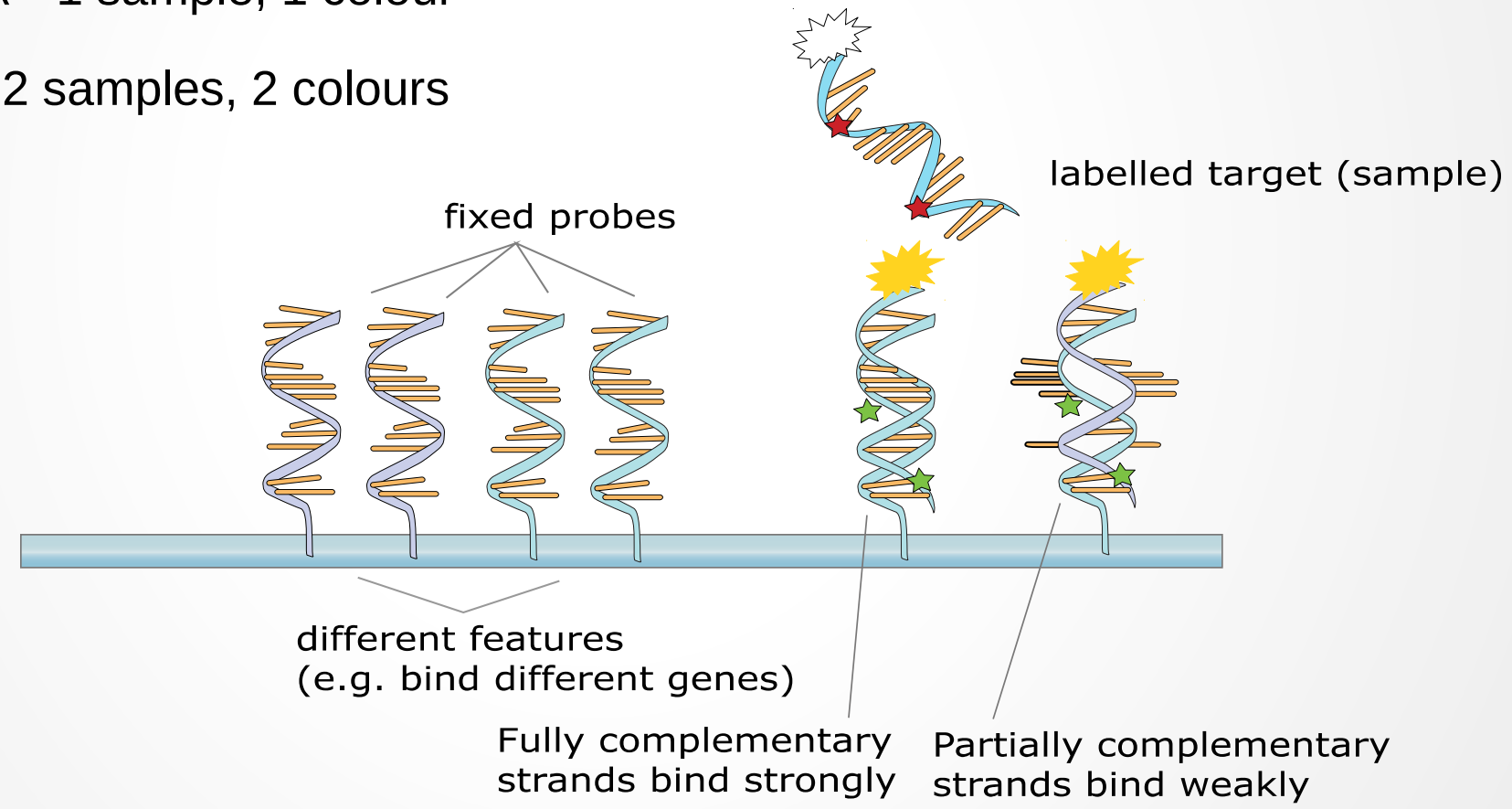
Microarray

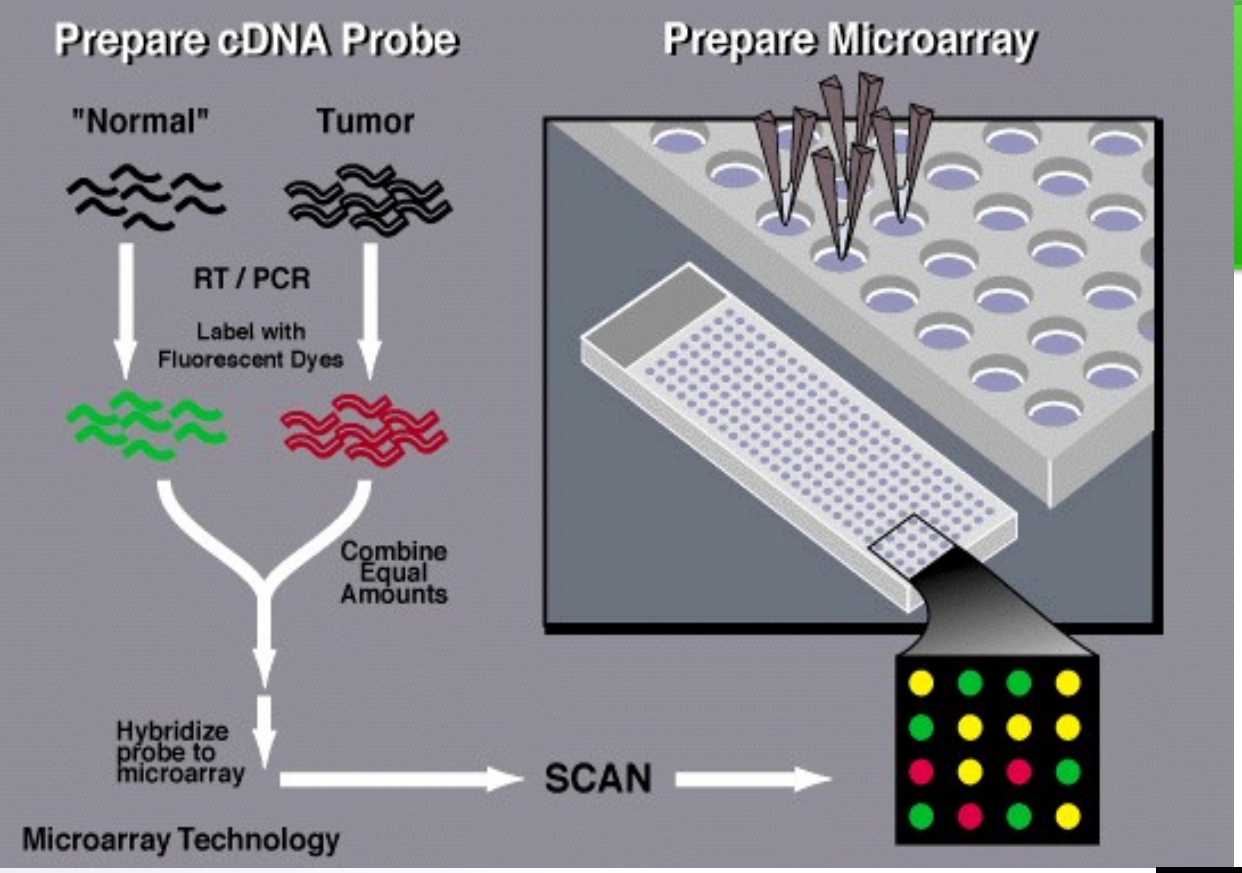
- It gives quantity information about the „whole” transcriptome using a 1×1 cm plate
  - Treatment 1 vs. treatment 2
  - Healthy vs. sick
  - Treated vs. untreated
- Which genes have significantly different expression levels?

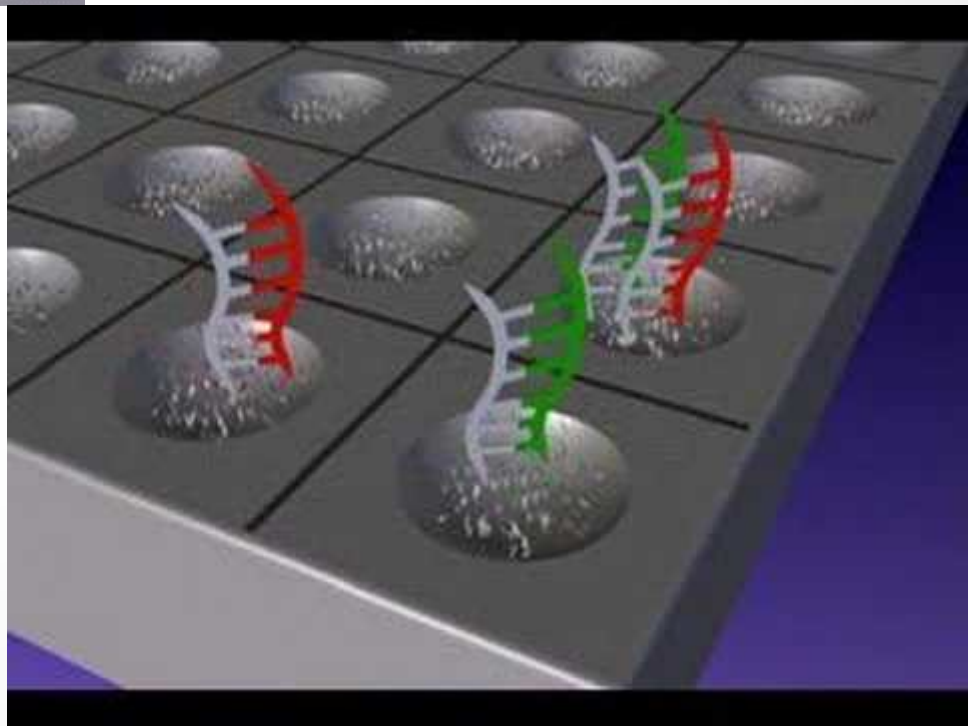


Microarray

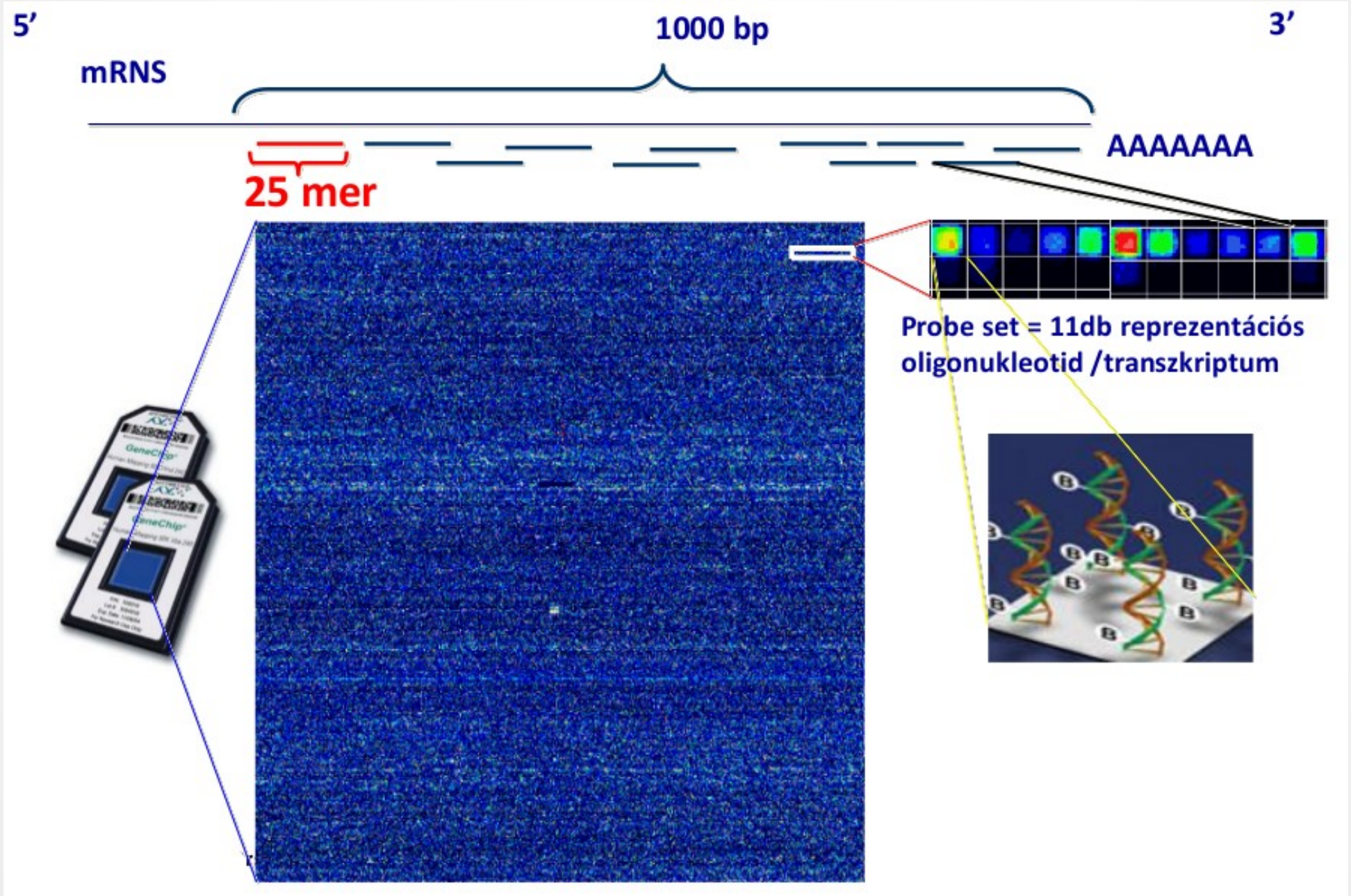
- 20-60 bp DNA oligos
- *Affimetrix* - 1 sample, 1 colour
- *Agilent* - 2 samples, 2 colours




Prepare cDNA Probe
Prepare Microarray
"Normal"
Tumor
RT / PCR
Label with
Fluorescent Dyes
Combine
Equal
Amounts
Hybridize
probe to
microarray
SCAN
Microarray Technology



Microarray


5’
1000 bp
3’
mRNS
AAAAAAA
25 mer
Probe set = 11db reprezentációs
oligonukleotid /transzkriptum

Microarray data processing

- Bioinformatics
  - Background correction
    - Romove the noise using negative controls (probes missing from the transcriptome)
  - Normalization
  - Aggregation
    - Probes in a probe set → single expression value

Differential expression analysis

- Compare two states
  - i. e. treated vs. control
- Without null hypothesis
- *Fold change*
- *t*-test: *P*-value

Fold change

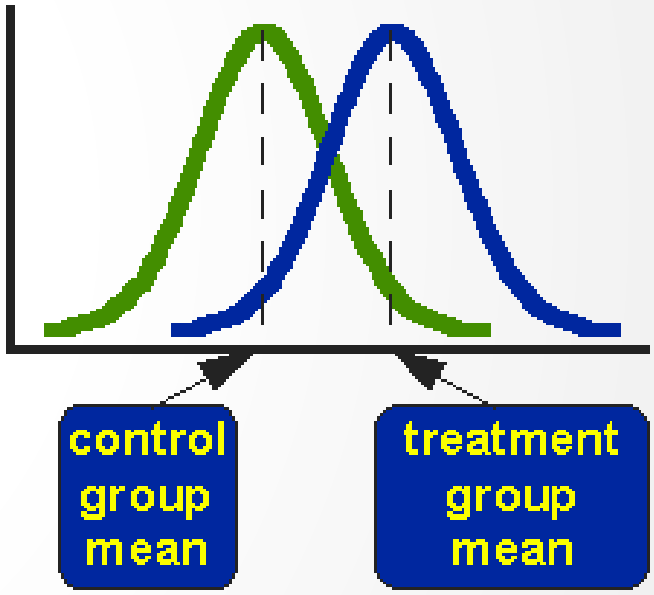
- To measure the scale and direction of expression level differences.

$$\log_2 FC = \log_2\left(\frac{\text{mean of a probeset in Treat 1}}{\text{mean of a probeset in Treat 2}}\right)$$

- 2 logFC (4x FC) is an acceptable difference
- Here we compare the means only. This is not a statistical test.

Hypothesis testing

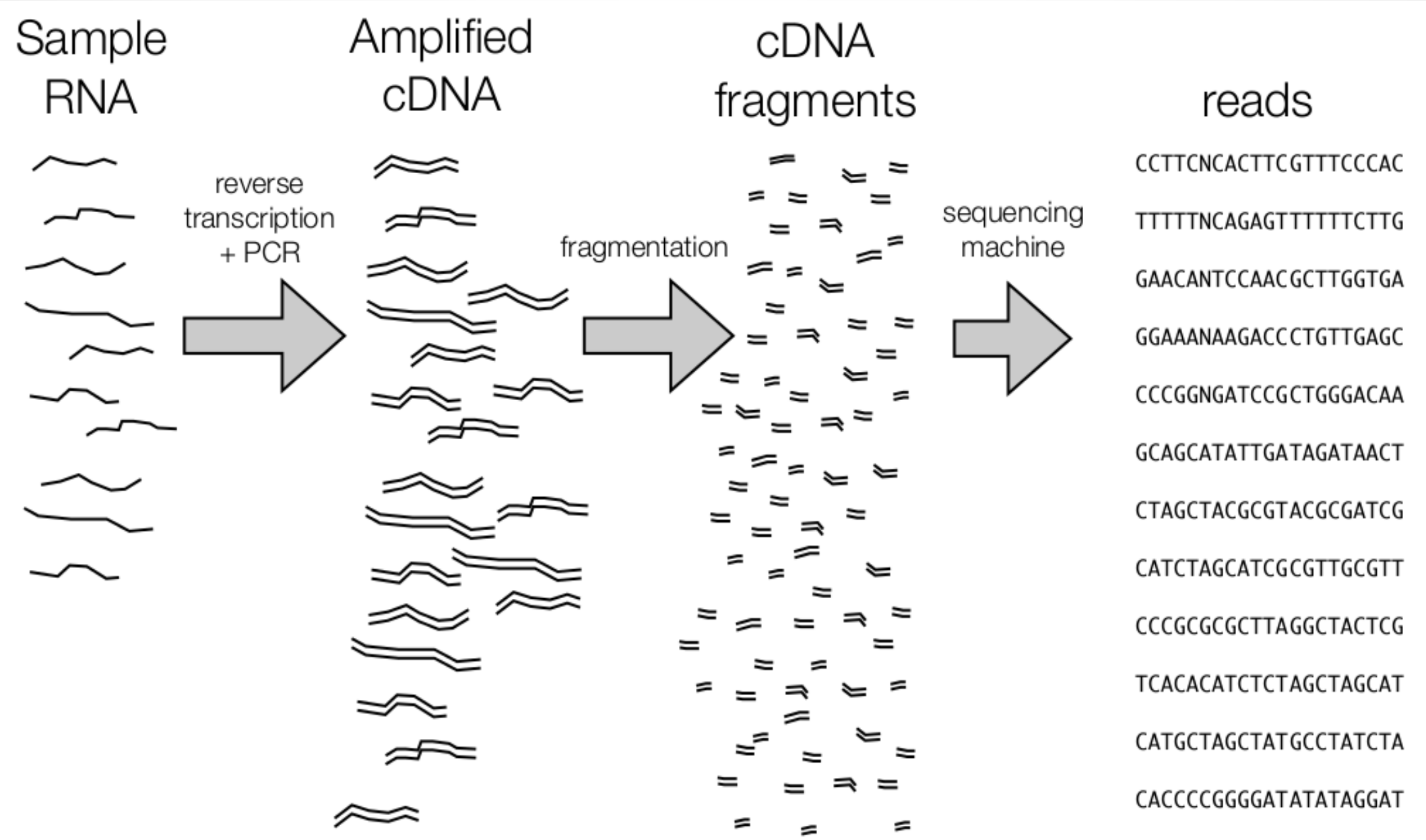
- 2 sample *t*-test
  - H0: the means of the 2 distributions are the same
    - → Where do the distributions come from?
    - → replicates (more samples of each treatment)
  - In statistical hypothesis testing, the **p-value** or **probability value** is the probability of obtaining test results at least as extreme as the results actually observed during the test, assuming that the null hypothesis is correct.



Microarray - summary

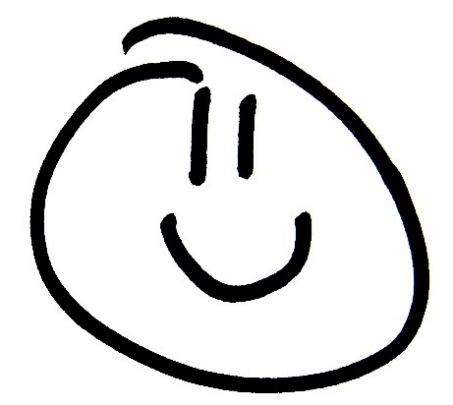
1. A lot of probes / chip
2. High-thoughput
3. Hypothesis-free research – but probe sets are pre-defined
4. Statistical testing
5. Online databases (ie. GEO, Array Express)
6. Limitations

RNA-Seq


Sample RNA
reverse transcription + PCR
Amplified cDNA
fragmentation
cDNA fragments
sequencing machine
reads
CCTTCNCACTTCGTTTCCCAC
TTTTTNCAGAGTTTTTTCTTG
GAACANTCCAACGCTTGGTGA
GGAAANAAGACCCTGTTGAGC
CCCGGNGATCCGCTGGGACAA
GCAGCATATTGATAGATAACT
CTAGCTACGCGTACGCGATCG
CATCTAGCATCGCGTTGCGTT
CCCGCGCGCTTAGGCTACTCG
TCACACATCTCTAGCTAGCAT
CATGCTAGCTATGCCTATCTA
CACCCCGGGGATATATAGGAT

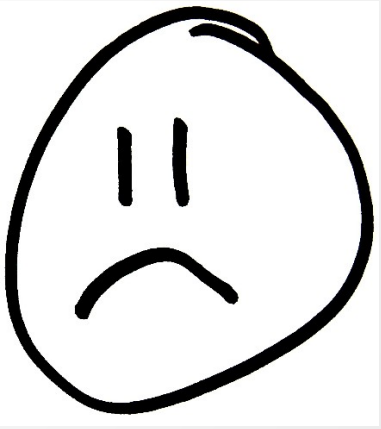
Advantages of RNA-seq

- Robustness, high reproducibility
- High sensitivity
- “Direct” measurement of gene expression at the mRNA level → absolute(?) abundance of a transcript
- The sequences of transcribed RNAs can be reconstructed
- All transcripts – even “novel” ones – present
- Detecting transcript isoformes and splicing junctions
  - → to study alternative splicing - exact start - end sites
  - → to update genome annotation
- Detecting polymorphisms (SNPs)
  - → to study allele-specific expression
- Can be used on species for which a full genome sequence is not available



Limitations of RNA-seq

- RNA-seq is (a bit) more costly than microarrays
  - RNA-seq: more extensive bioinformatic analysis and great computers are required
- Cannot detect post-transcriptional modifications
- Nor post-transcriptional regulation:
  - the amount of mRNA transcribed from geneX is not necessarily equal to the amount of proteinX
  - regulation: miRNA …
- Bias: library size, fragment length, GC content...
  - → normalization



Work-flow of RNA-seq data analysis

0. Extract expressed RNA, sequencing → fastq file
1. Pre-mapping quality checking, trimming (filtering)
2. *De novo* assembly of transcripts

   OR

   read mapping to reference genome

   Post mapping quality checking
3. Read counting
4. Differential Expression analyses: comparing expression levels
5. Functional enrichment analysis: GO, pathways...

Work-flow of RNA-seq data analysis

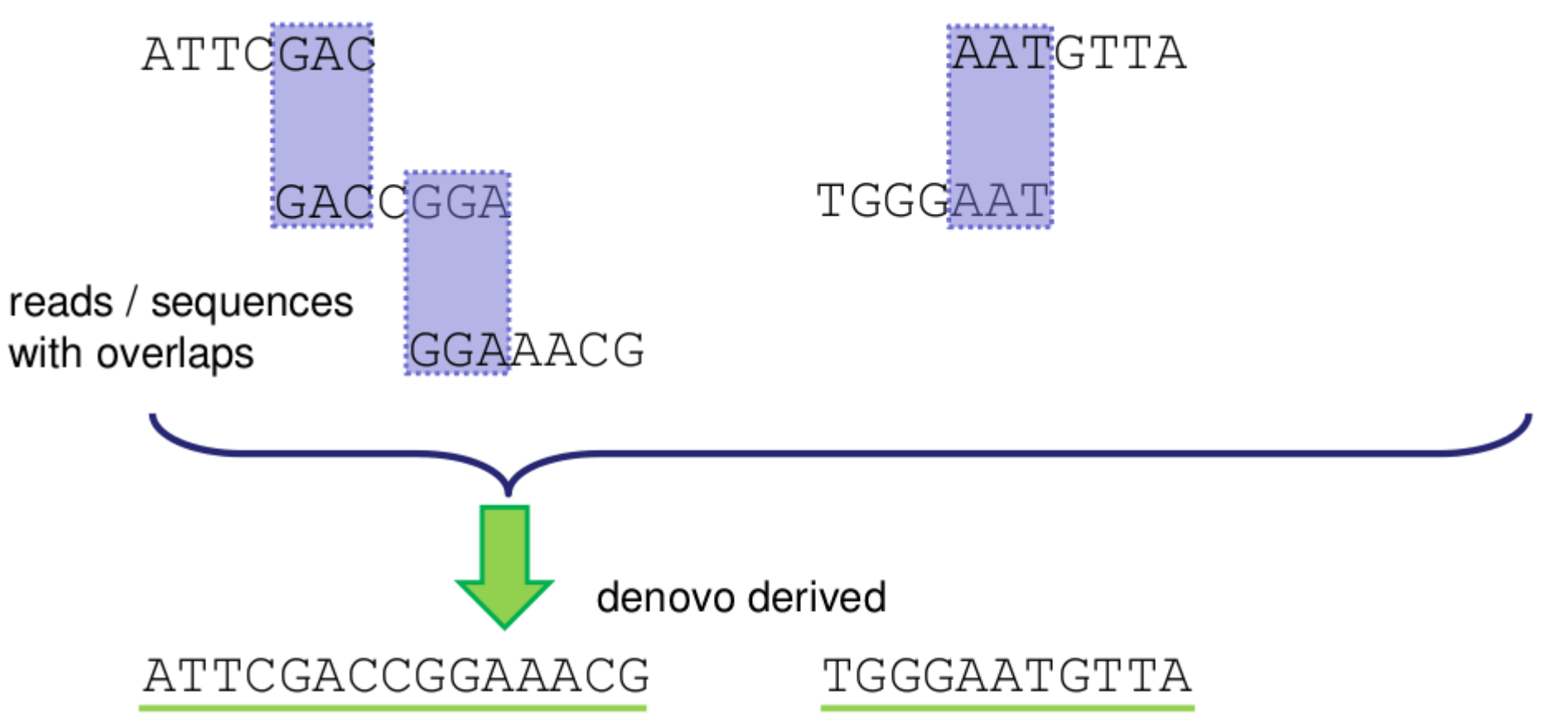
0. Extract expressed RNA, sequencing → fastq file
1. Pre-mapping quality checking, trimming (filtering)
2. *De novo* assembly of transcripts

   OR

   read mapping to reference genome

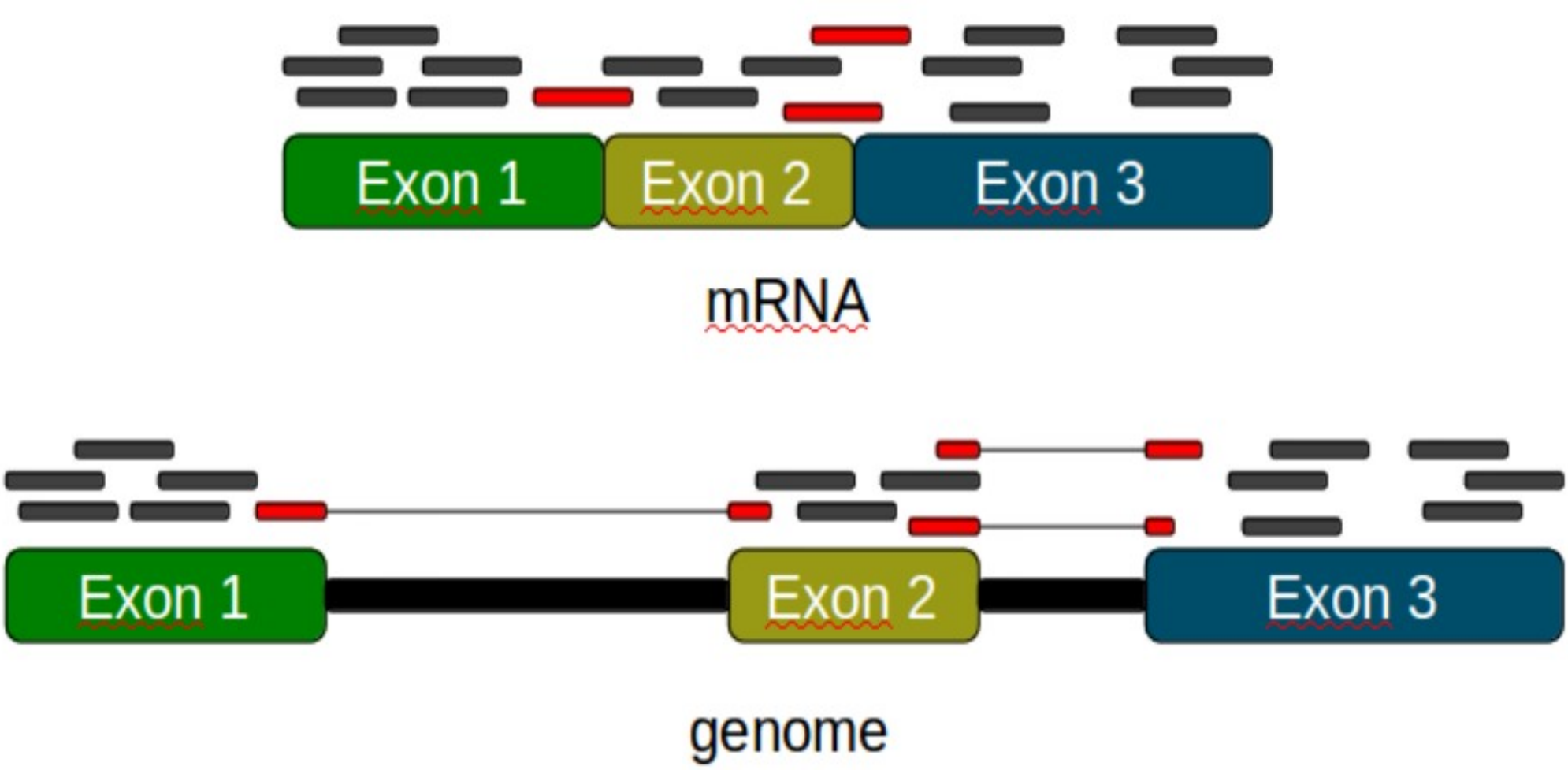
   Post mapping quality checking
3. Read counting
4. Differential Expression analyses: comparing expression levels
5. Functional enrichment analysis: GO, pathways...

De-novo transcriptome assembly




De novo assembly
ONE HUMAN BEING!
(SOME ASSEMBLY REQUIRED)
BY AUTH FOR THE PHILADELPHIA INQUIRER

Mapping reads to reference genome



TopHat, GSNAP, Star, ...

Problems

- pseudogenes (the reads were mapped to something that didn't express)
- identification and quantification of alternative transcripts
- detection of (allelespecific) SNPs
- reads mapped to intronic and intergenic regions → how should we treat them?

Work-flow of RNA-seq data analysis

0. Extract expressed RNA, sequencing → fastq file
1. Pre-mapping quality checking, trimming (filtering)
2. *De novo* assembly of transcripts OR

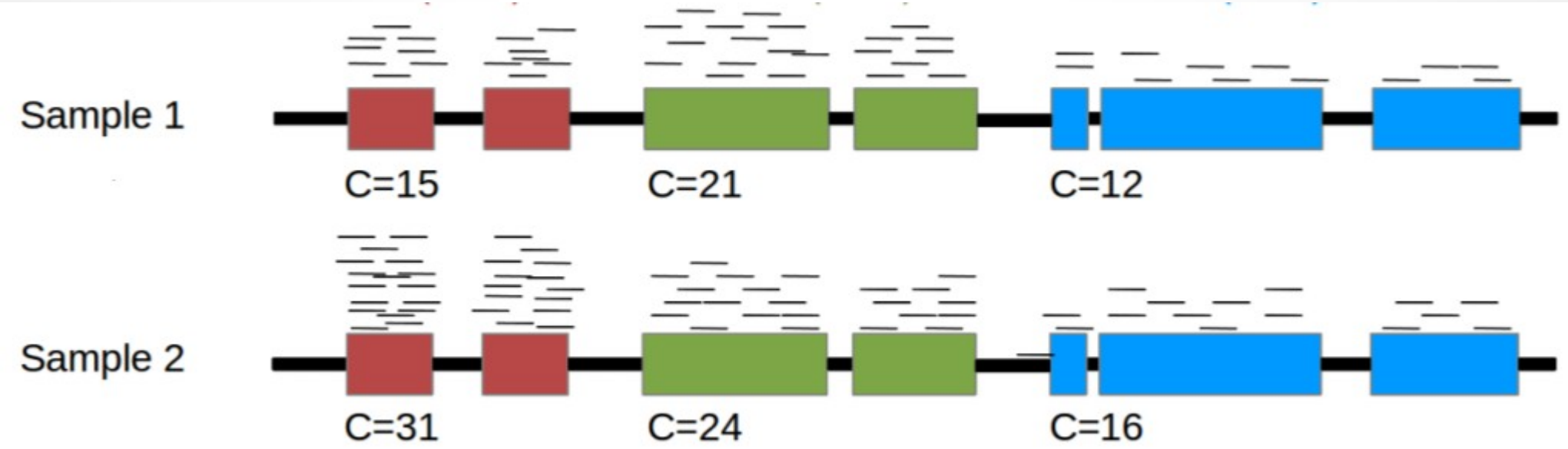
   read mapping to reference genome

   Post mapping quality checking
3. Read counting
4. Differential Expression analyses: comparing expression levels
5. Functional enrichment analysis: GO, pathways...

3. Read counting

- Find reads that map to coding sequence
  - count read(pairs) per gene, exon, transcript
    → count table
- Genome annotation: GTF (GFF, SAF, ...) file:
  - contains the location of exons, genes, other transcripts

Read counting

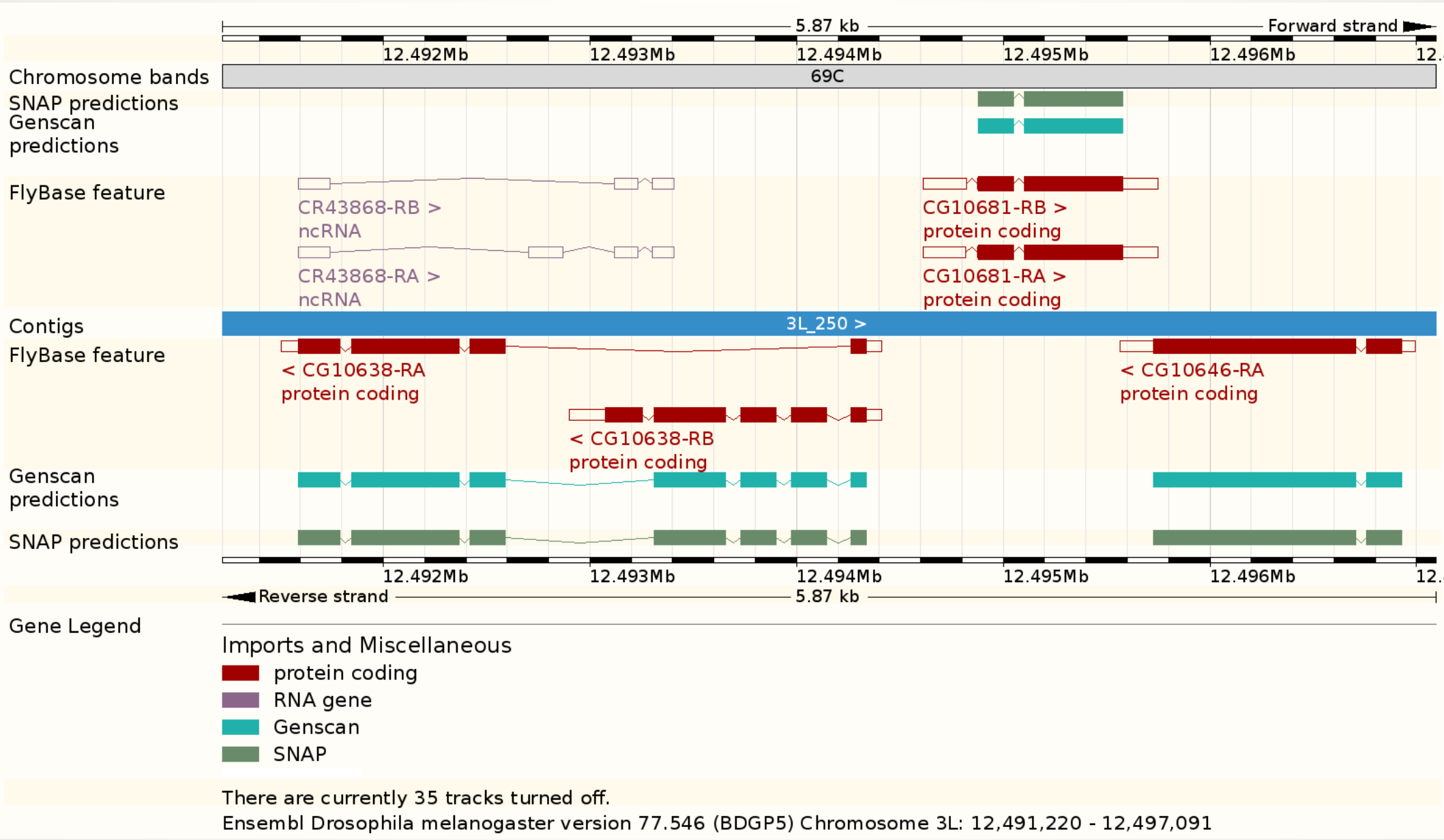


Count table

	F1	F2	F3	F4	M1	M2	M3	M4
ENSG00000127720	14	14	23	16	32	35	10	19
ENSG00000242018	24	16	11	19	21	22	13	6
ENSG00000224440	0	0	0	0	0	0	0	0
ENSG00000214453	0	0	0	0	0	0	0	0
ENSG00000237787	1	0	0	0	0	0	1	0
ENSG00000051596	220	325	450	585	475	294	224	711

...

Complexity of transcription



Work-flow of RNA-seq data analysis

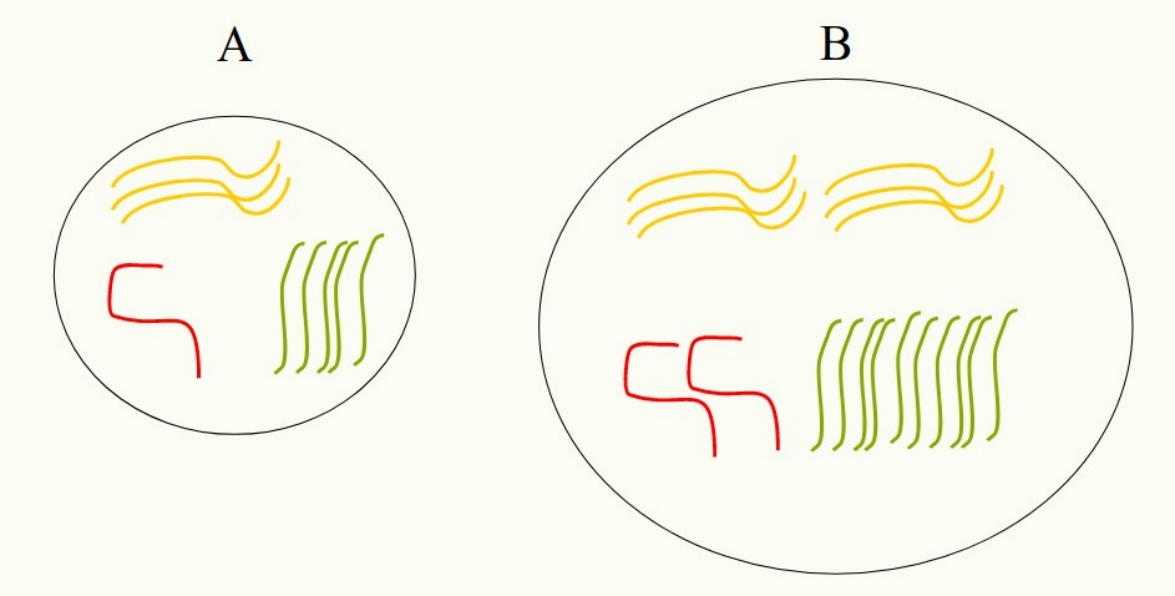
0. Extract expressed RNA, sequencing → fastq file
1. Pre-mapping quality checking, trimming (filtering)
2. *De novo* assembly of transcripts OR

   read mapping to reference genome

   Post mapping quality checking
3. Read counting
4. Differential Expression analyses: comparing expression levels
5. Functional enrichment analysis: GO, pathways...

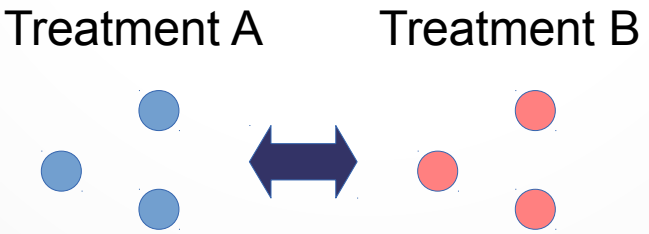
Normalization

- It is not possible to do absolute quantification using the common RNA-Seq pipeline, because it only provides RNA levels relative to all transcripts.
- The counts need to be adjusted to be comparable across samples and experiments.
  - Because the total coverage (sum of all read counts) differs accross samples
  - Relaive vs. Absolute expressions



Experimental layout

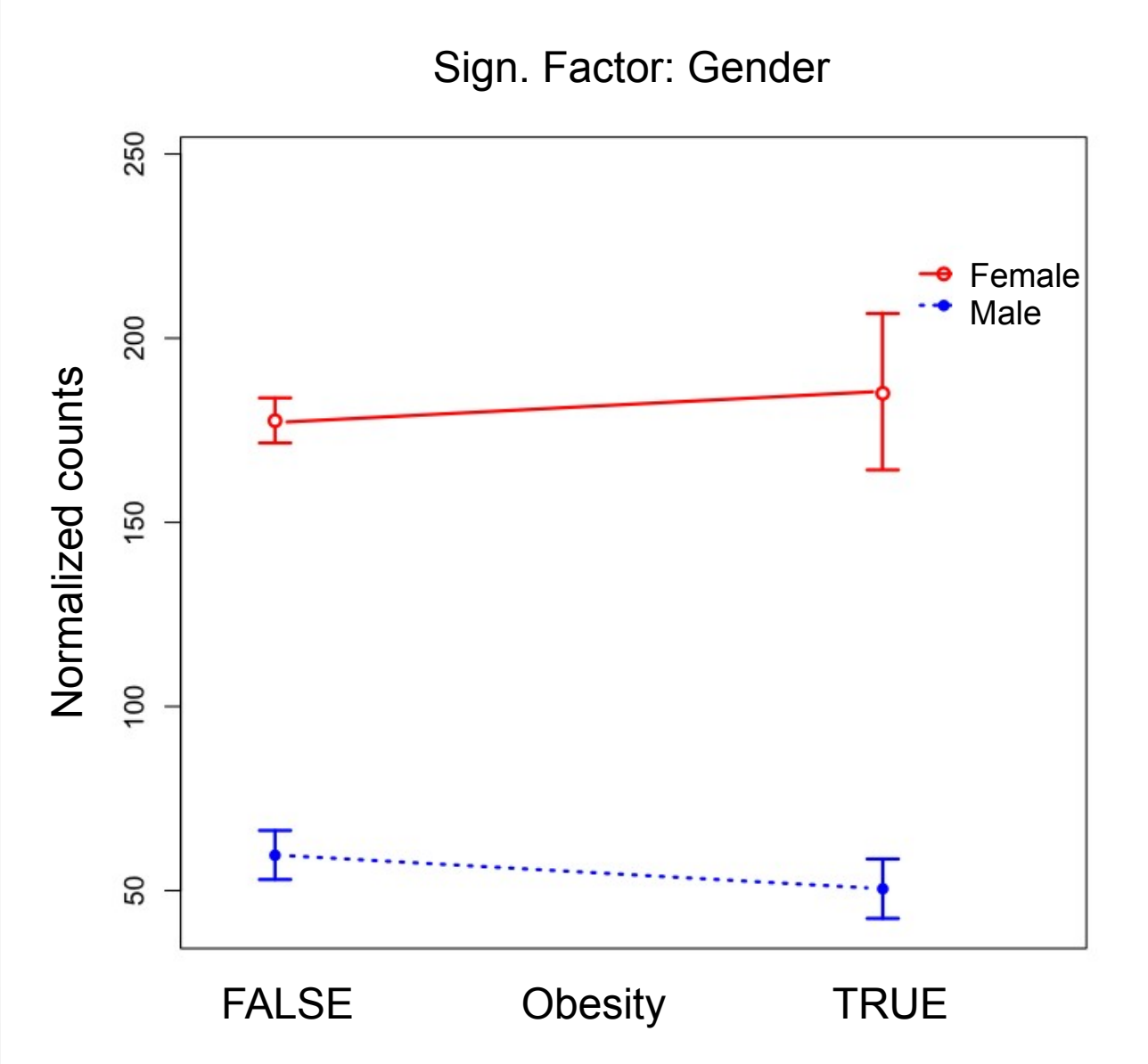
- 2 groups:
  - Question: Which genes express significantly differently between the two groups? → *p-value*
  - The direction of the difference → *Fold change*
  - *pairwise DE analysis*



Experimental layout

- More treatment types, more groups:
  - Question: Does a factor (treatment type) couse DE? In which genes?
    - Factors: i.e obese or not; male or female ...
  - We use a Generalised Linear Model (GLM) to calculate the *p*-value
  - The directin of the difference → *Fold change*

GLM


Sign. Factor: Gender
Normalized counts
250
200
150
100
50
Female
Male
FALSE
Obesity
TRUE

GLM


Sign. Factor: Obesity
Normalized counts
350
300
250
200
150
Female
Male
FALSE
Obesity
TRUE

GLM

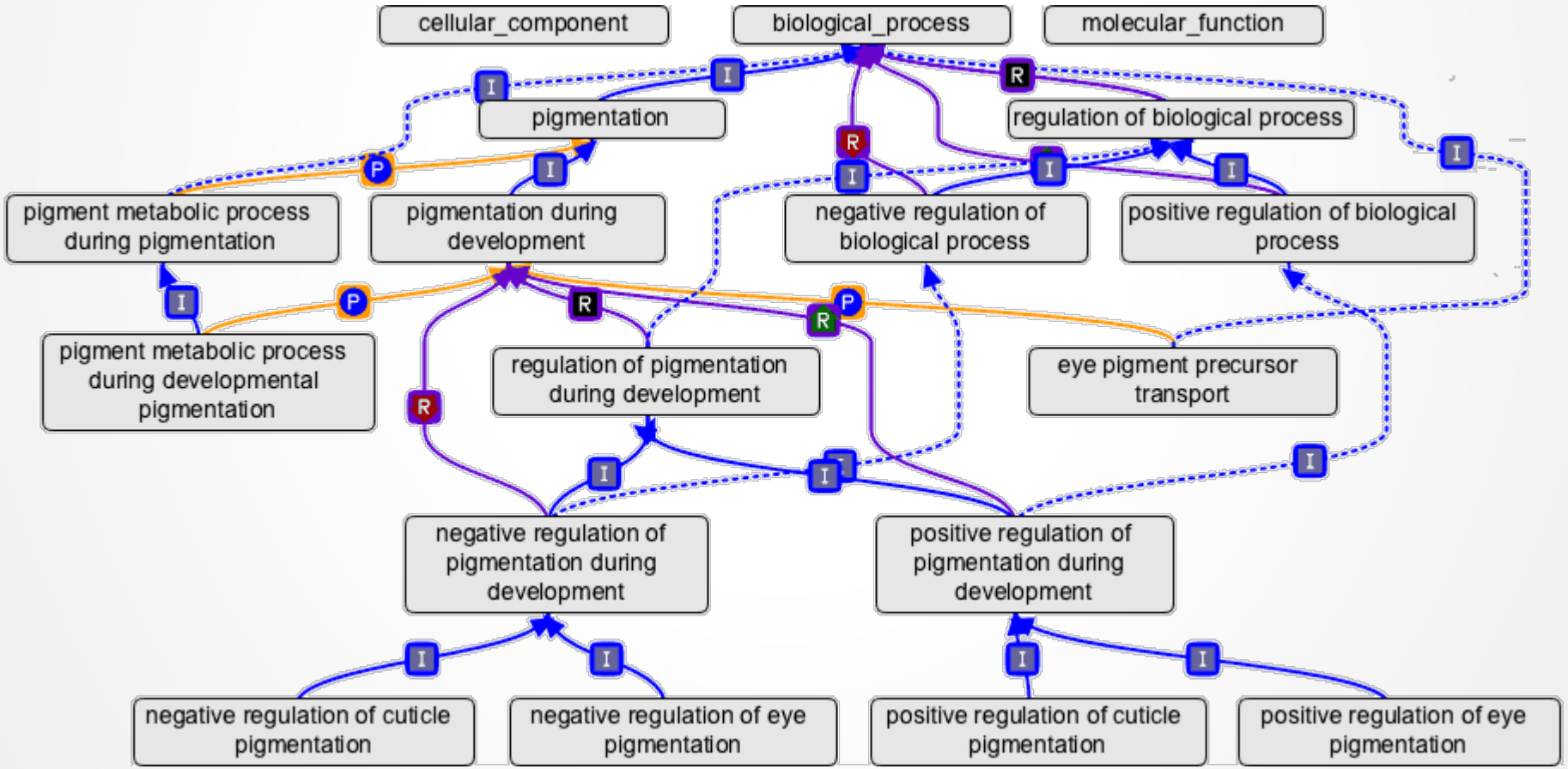

Sign. Factor: interaction of gender and obesity
Normalized counts
300
250
200
150
100
Female
Male
FALSE
Obesity
TRUE

Multiple testing correction

- We calculate the same statistical test several times → to all probe sets of the microarray / all genes or transcripts of an RNA-seq
- If we use p=0.05 as a cutoff: we have 5% chance to accept something significantly differently expressed when the expressions were not different
- → *False discovery rate* (FDR) correction based on all p-values. ie. Bonferroni or Benjamini-Hochberg correction

After the DE analysis: What is the function of DE genes

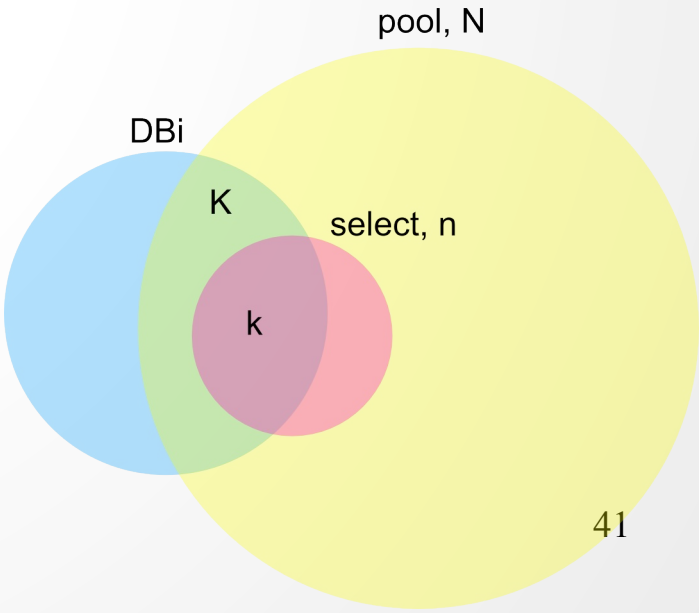
- Gene Ontology - GO: http://geneontology.org



Enrichment analysis

- Question: Is the GO category significantly overrepresented among DE genes, compared to the all genes that we investigated (background genes)?
  - → Finds the biological functions of the DE genes
- Hypergemetric test:
  - *N: Nr. of background genes (pool)*
  - *K: Size of the intersection of background genes and genes of the GO category (DBi)*
  - *n: Nr. of DE genes (select)*
  - *k: Size of the intersection of DE genes and genes of the GO category*
  - *The set of all k-combinations of a set K: K over k*

$$P = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$



Thanks for the attention