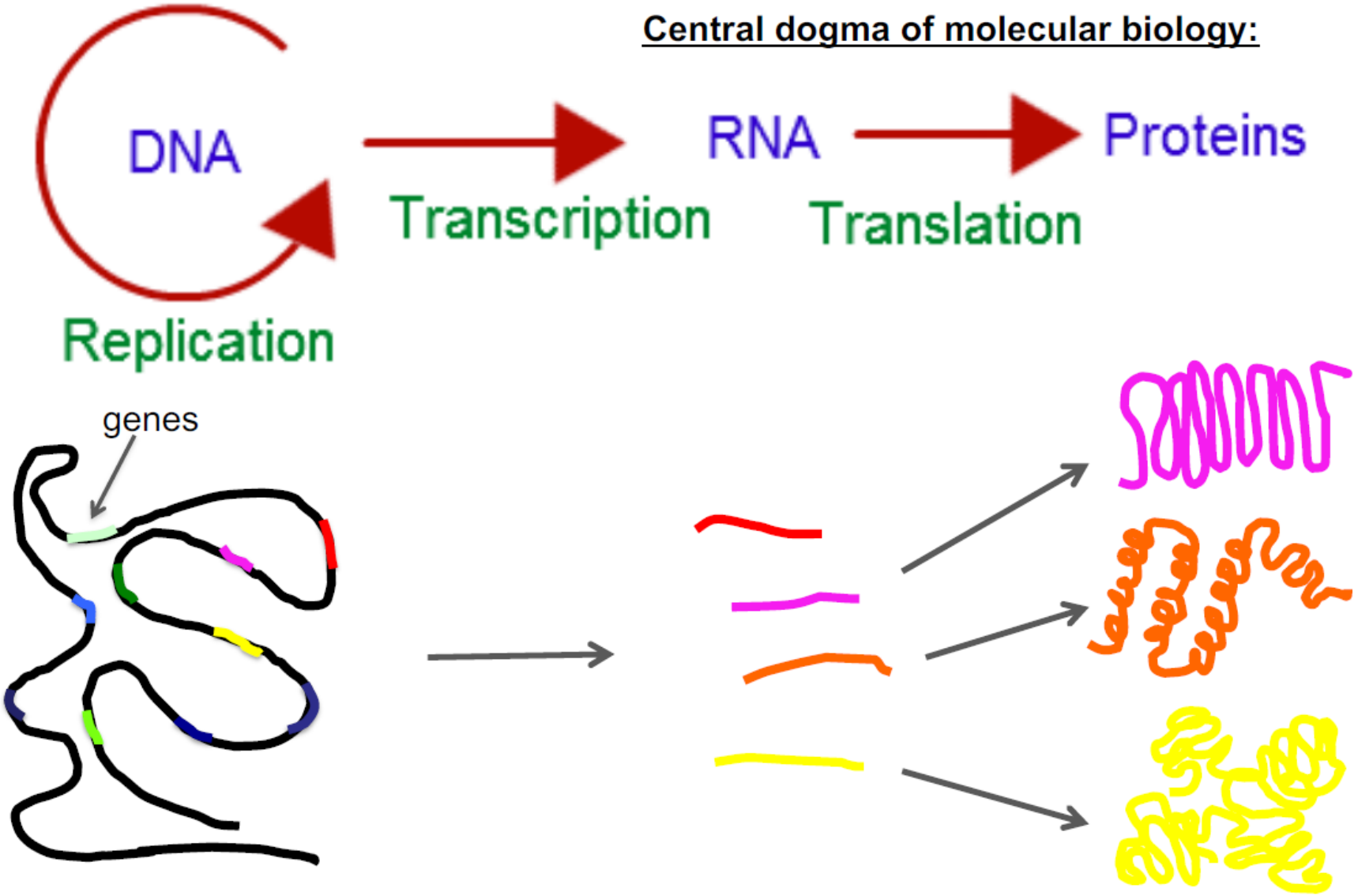
Analysis of gene expression

From genome to transcriptome

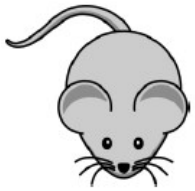


Transcriptome is dynamic and different in every tissue!

Transcriptome is a set of active (= expressed) genes at the moment of sampling.

Transcriptome is variable between tissues, during developmental stages or as a response to different conditions (stress, disease, weather…).

Mouse:



tissues:

mouse liver transcriptome:

is different from

mouse kidney transcriptome:

is different from

mouse eye transcriptome:

age:

embryonic transcriptome:

is different from

adult transcriptome:

conditions:

transcriptome of healthy mouse:

is different from

transcriptome of sick mouse:

Transcriptome = RNA – response to a need for a protein…

healthy…

infection

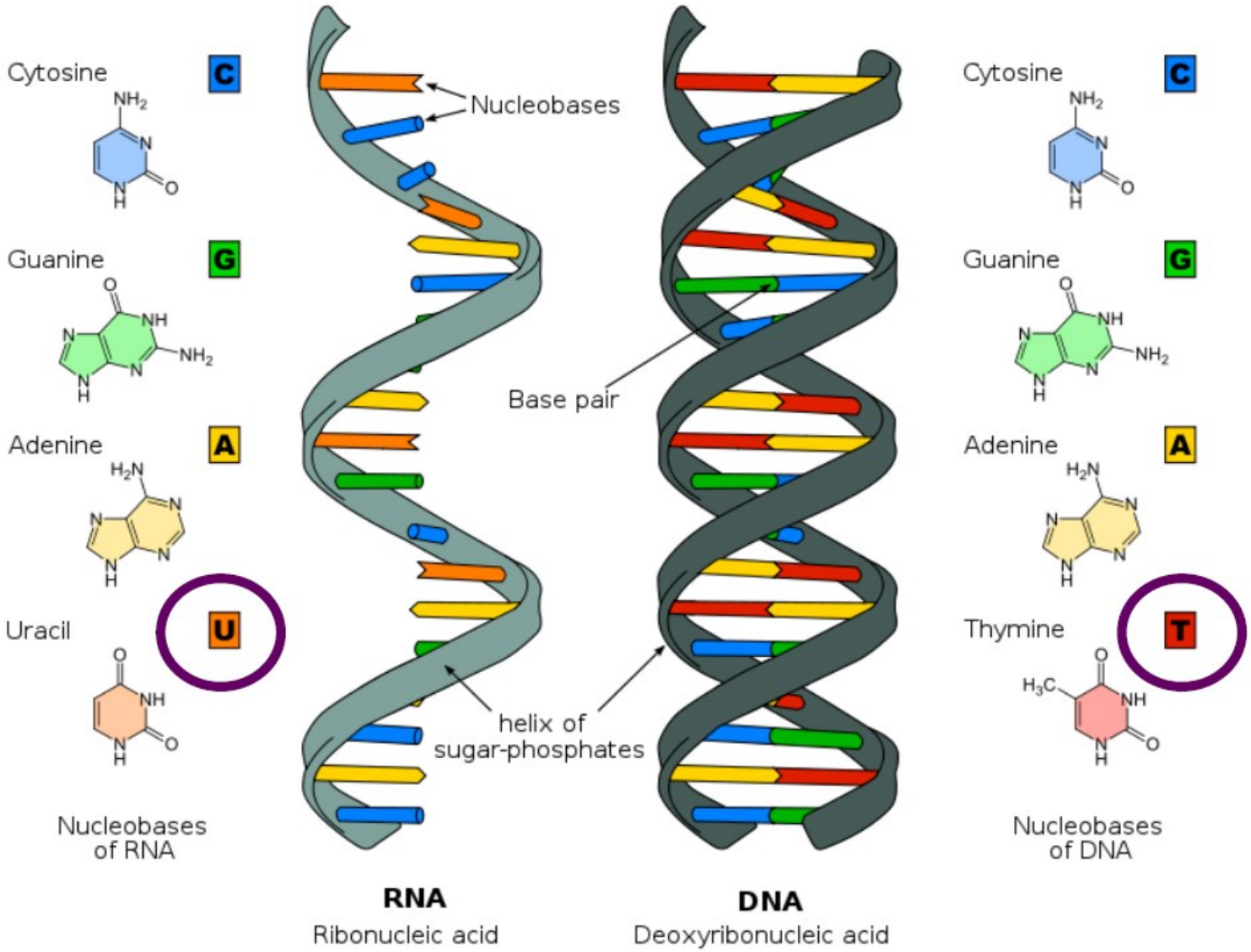
sick…

need for **immunity response**
=> need for expression of the immunity genes

Transcriptome contains less **immunity genes** → Transcriptome contains more **immunity genes**

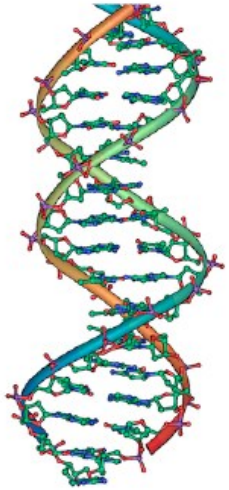
RNA vs. DNA

single strand vs. double strand

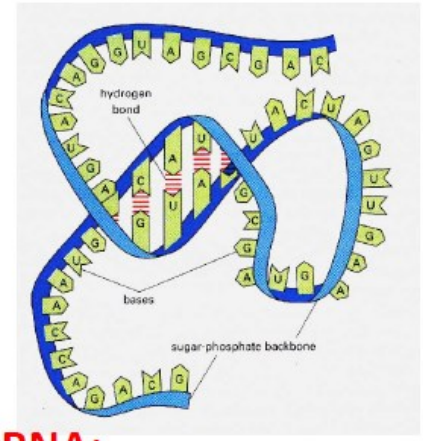


RNA vs. DNA

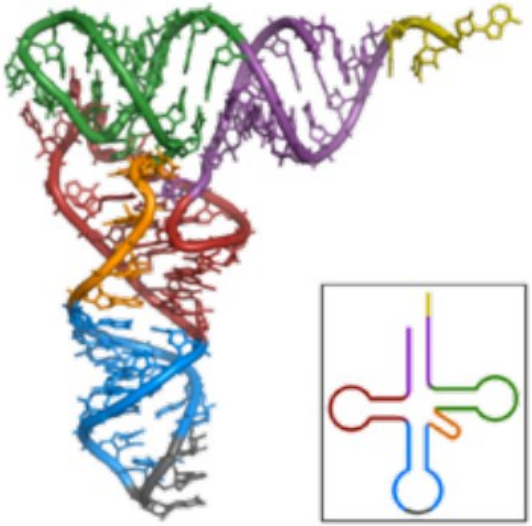
DNA – double helix



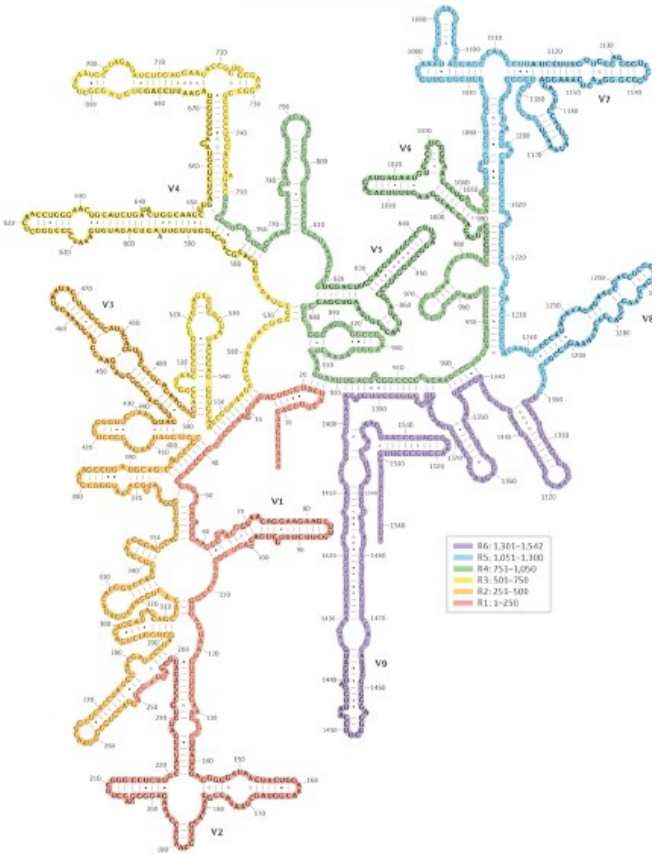
RNA – single helix => secondary structure!



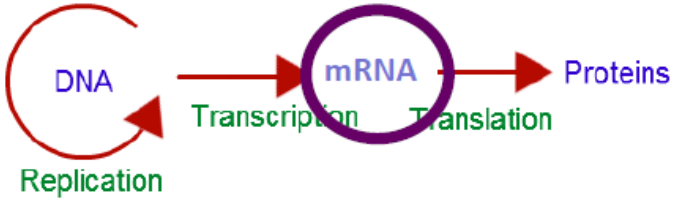
tRNA:



16S rRNA:



Many types of RNA:



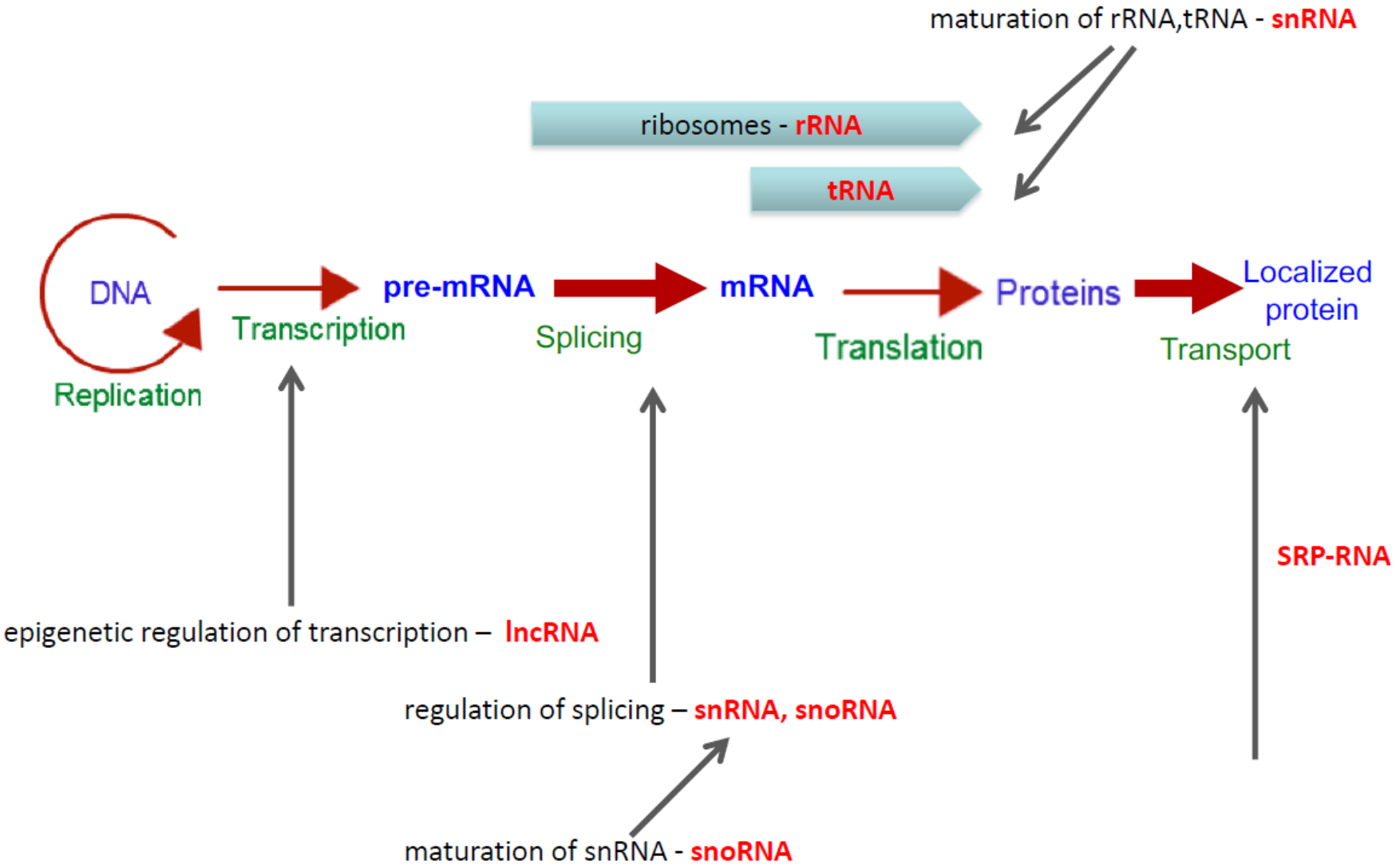
- **rRNA = ribosomal RNA – (80–90% of RNA) – ribosomes, translation**
- **tRNA = transfer RNA – (up to 15% of RNA) – translation, carries amino acids**
- **mRNA = messenger RNA (cca 1-3% of RNA) – coding proteins!!!**
- miRNA = micro RNA (21-24 nucl.) – regulation
- siRNA = small interfering RNA – gene silencing

Small non-coding RNA (26 – 31 nucl.) :

- snRNA = small nuclear RNA– maturation of rRNA, tRNA, splicing
- snoRNA = small nucleolar RNA (type of snRNA) – splicing, maturation of rRNA
- scaRNA = small Cajal body RNA (type of snRNA) – maturation of rRNA
- piRNA = piwi-interacting RNA – post-transcriptional gene silencing of retrotransposons

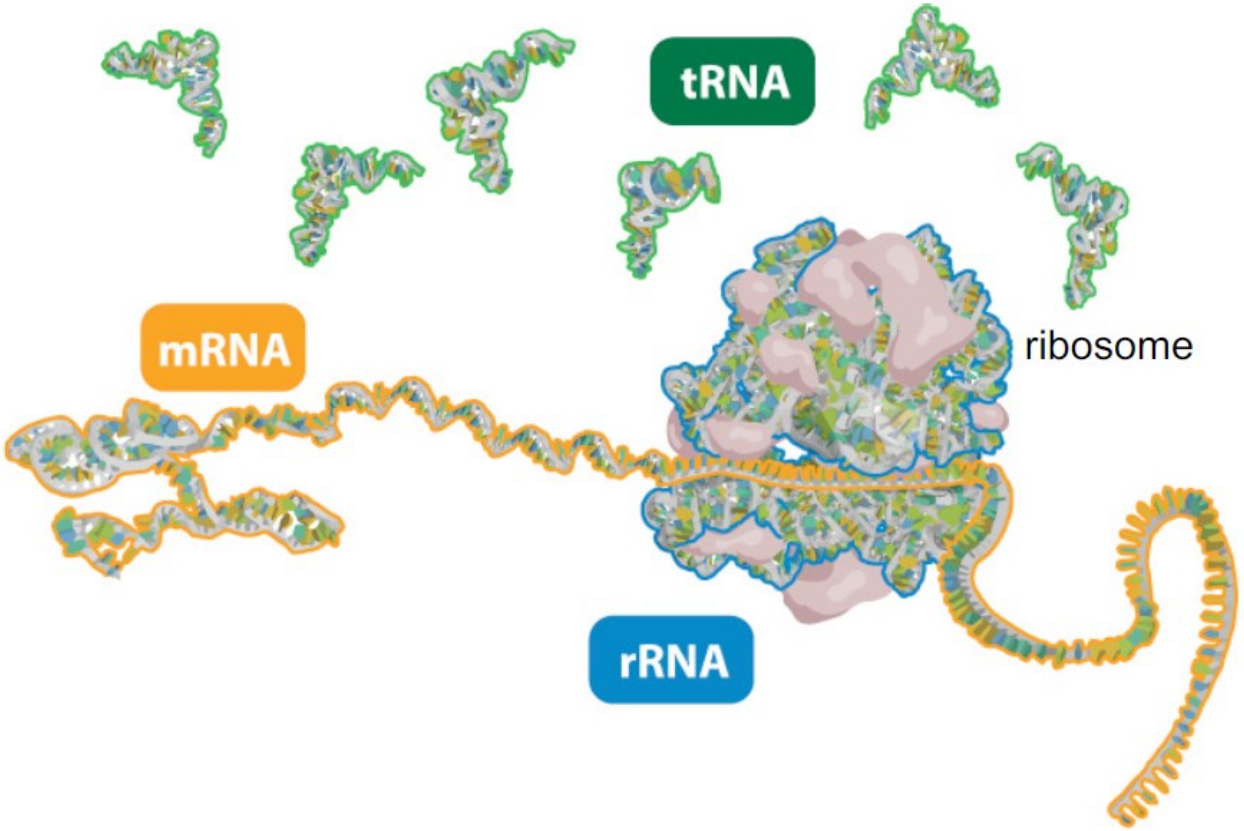
lncRNA = long non-coding RNA (>200 nt.) epigenetic regulation of transcription; Xist

- exRNA = extracellular RNA (any of the mRNA, tRNA, miRNA, siRNA, lncRN)
- SRP-RNA = signal recognition particle RNA – transport of proteins
- … + many other (unknown) types…



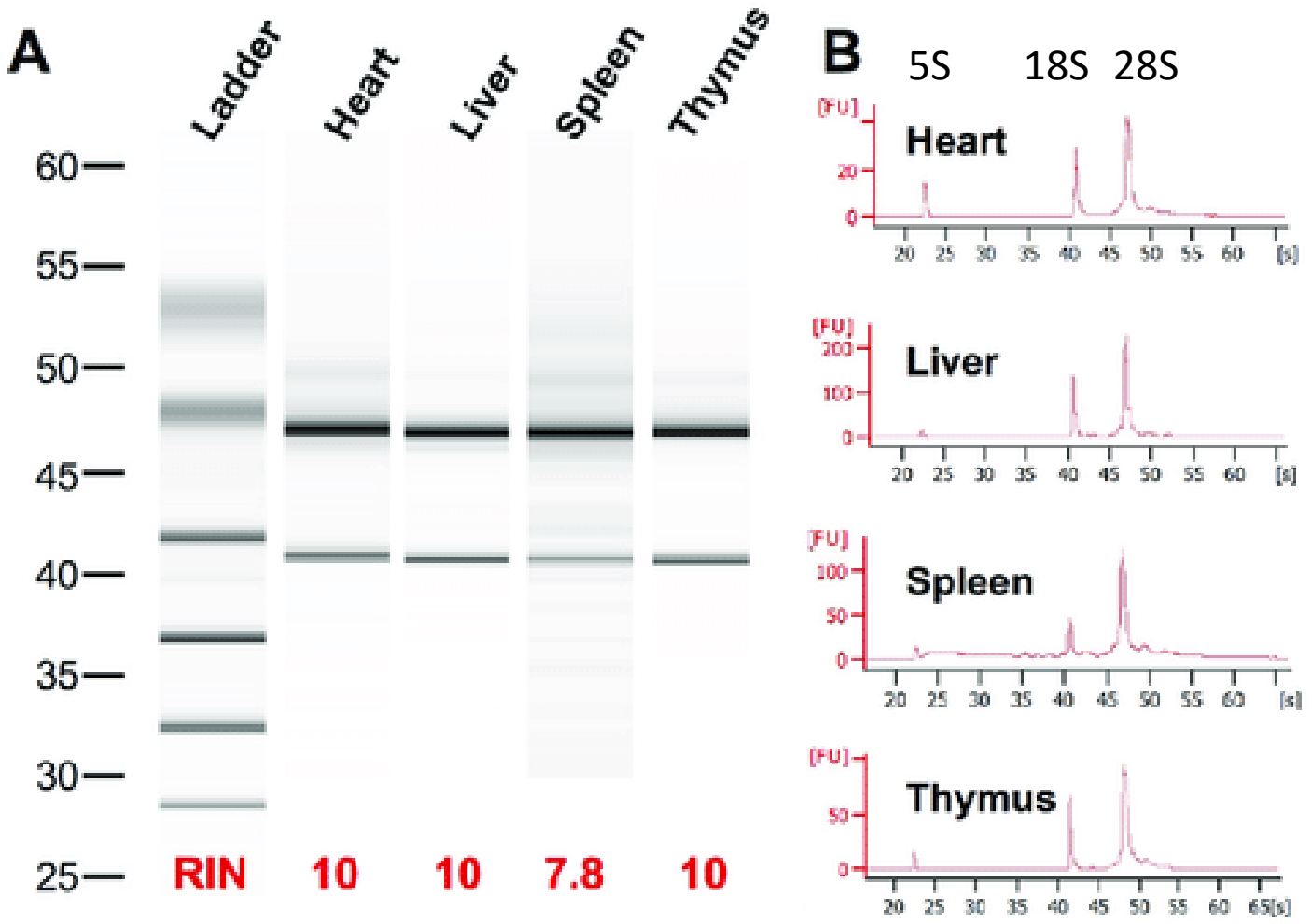
...it's way more complex and still not completely understood...

Proteosynthesis apparatus



Most of the RNA in the cell (80%) is composed of rRNA: in eukaryotes 5S rRNA, 28S rRNA (large ribosome subunit), 18S rRNA (small ribosome subunit)

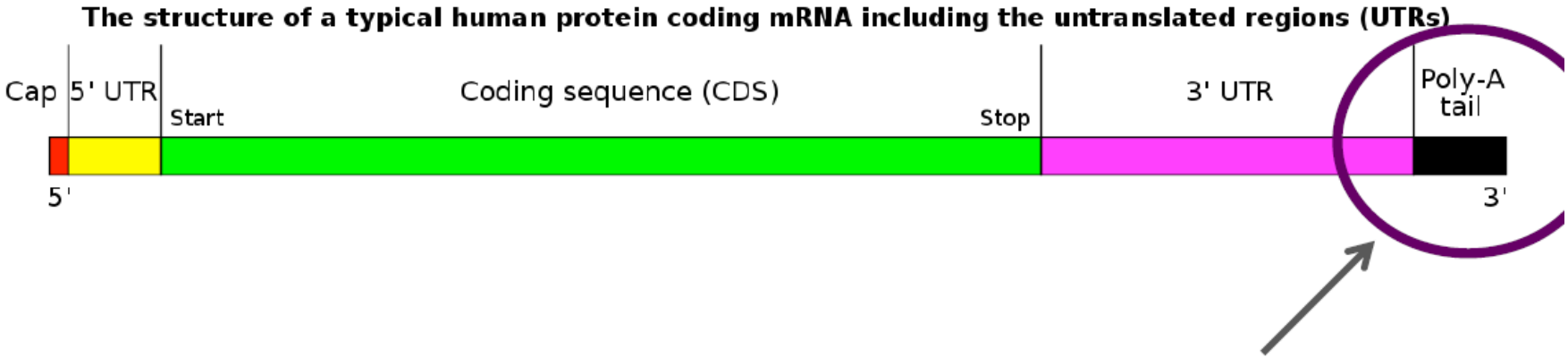
Total RNA



Agilent 2100 Bioanalyzer

Transcriptome sequencing = RNA-seq

- rRNA = ribosomal RNA – (80–90% of RNA) – ribosomes, translation
- tRNA = transfer RNA – (up to 15% of RNA) – translation, carries amino acids
- mRNA = messenger RNA (cca 1-3% of RNA) – coding proteins!!!
- miRNA = micro RNA (21-24 nucl.) – regulation



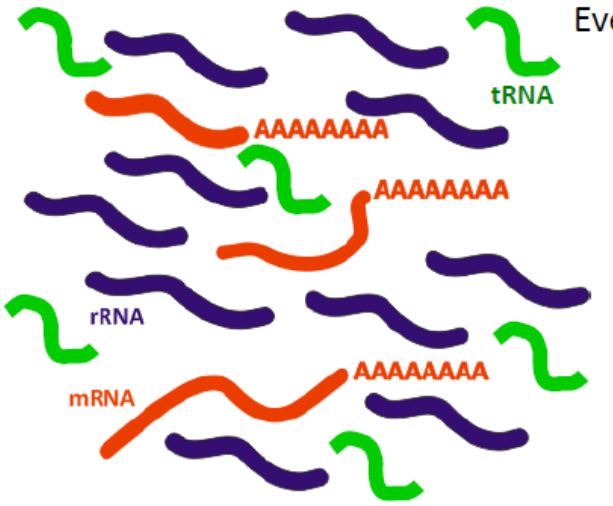
Every mRNA product has a poly-A tail (added during polyadenylation).

Challenge: how to get just mRNA?

Challenge: how to get just mRNA?

Poly-A selection vs. rRNA depletion

all RNAs

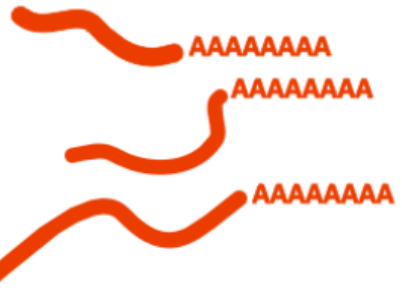


mRNA

Every mRNA product has a poly-A tail (added during polyadenylation).

poly-A selection

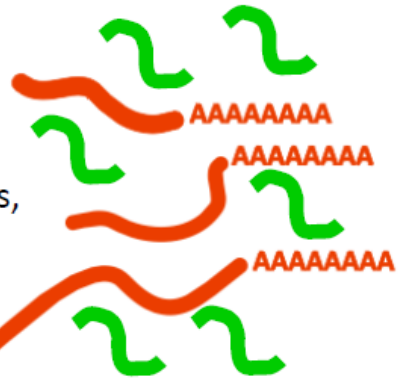
using a complementary probe with the TTTTTTTTTT motif



Best choice for: mRNA in eukaryotes

rRNA depletion

removes most of the rRNA)

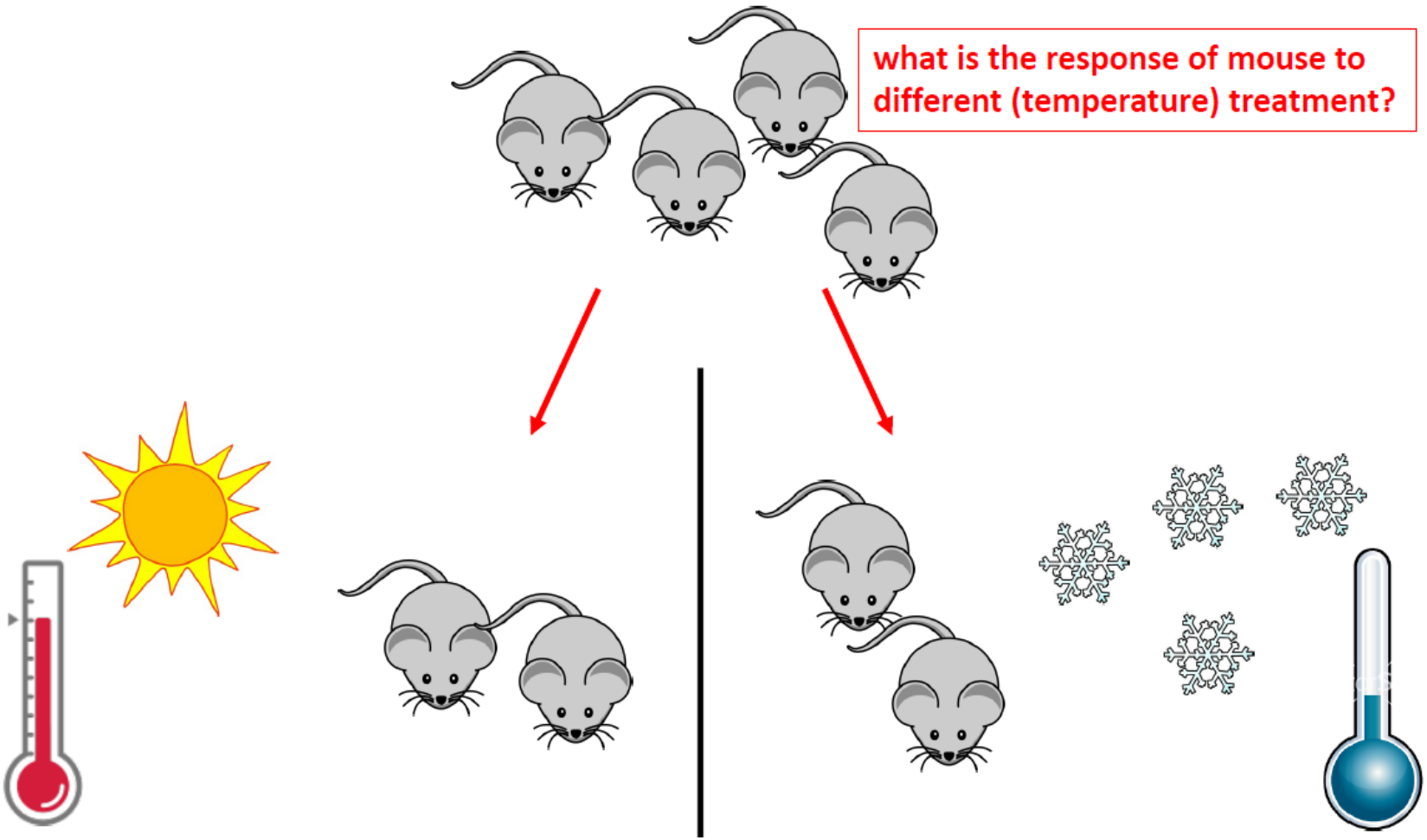


Best choice for: other types of RNA or mRNA in prokaryotes, which don't have poly-A tail

=> mitochondrial genes also don't have poly-A tail and can be removed by poly-A selection

RNA-seq and gene expression studies

example of the functional question:



known physiology:

- ***energy saved in white fat***
- ***energy consumed from brown fat***
- ***white fat turns into the brown fat in need...***

- genes for the white-to-brown-fat transformation – activated in cold

↓

- sequence all mRNA from warm and cold treatment mice...

COMPARE THEM -> find the genes with different expression (diff. amount of mRNA)

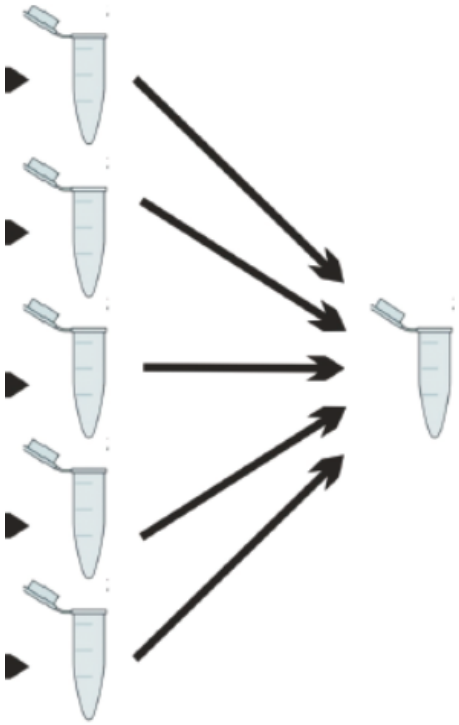
↓

candidate genes for this specific function...

e.g. neuregulin 4

Rosell et al., 2014

Nowadays mostly Illumina platform:



all mRNA sequenced!!!

-everything
-no need for primers (cool for non-model species)
-quantification
-no need for control genes, tissues etc...
-provides candidates

/s mostly Illumina platform:

NGS (many millions of short reads)

- some genes are **highly expressed**, some are rather **rare transcripts**
- Illumina HiSeq provides a **dynamic range of 5 orders of magnitude** => able to detect rare transcripts in ratio of 1:100'000! (with the linear relation)
- Minimum amount of required reads is **10'000'000 per sample** (=> i.e. many millions of reads are an advantage)
- Need for at least **3 replicates** (= samples from the same condition)
- Software for **differential expression analysis**: DESeq package in R

-everything
-no need for primers (cool for non-model species)
-quantification
-no need for control genes, tissues etc...
-provides candidates

How to calculate gene expression:

RKPM, FKPM, TPM…

1) Count all reads in the sample => divide it by 1'000'000 – that's **scaling factor**
2) Count reads of your gene / divided by scaling factor => **read per million** (RPM)
3) Normalize by length of gene in kbp => **reads per kilobase per million** (RPKM)
   - for the single-end RNA-seq (where 1 read = 1 fragment)
5) For paired-end reads - 2 reads = 1 fragment => **fragments per kilobase per million** (FPKM)

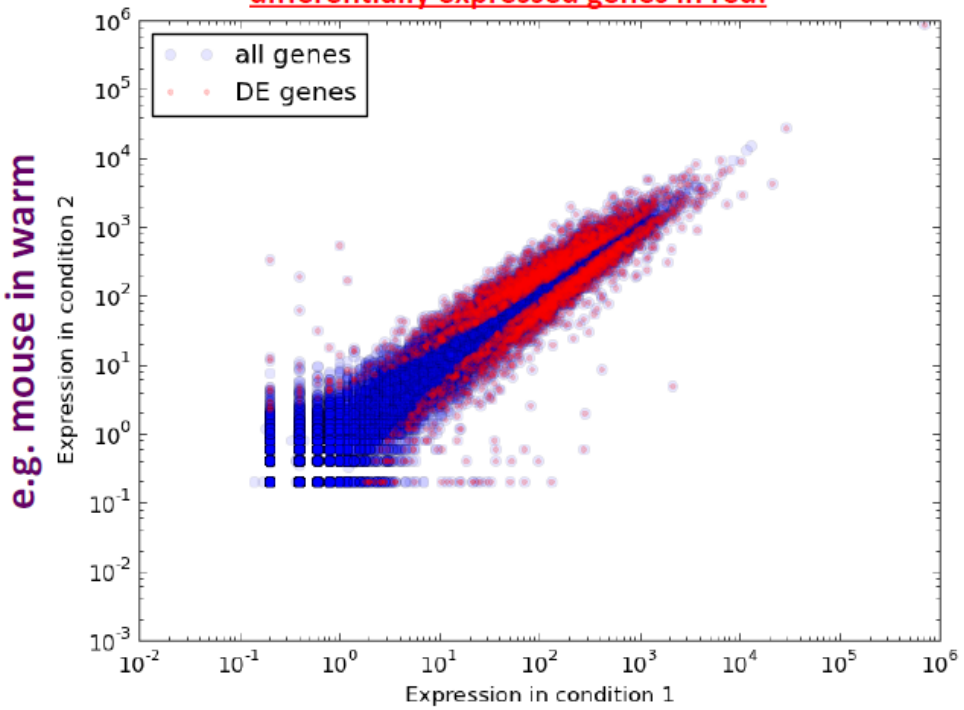
6) Alternative: TPM = **transcripts per million**
   - same method but different order: first normalize your gene reads per kilobase (RPK)
   - then sum the RPK and divide by 1'000'000 = this is a scaling factor now
   - divide you RPK by the scaling factor => TPM

RNA-seq and expression studies

Expression profiles are comparative, i.e. there is always a relative comparison

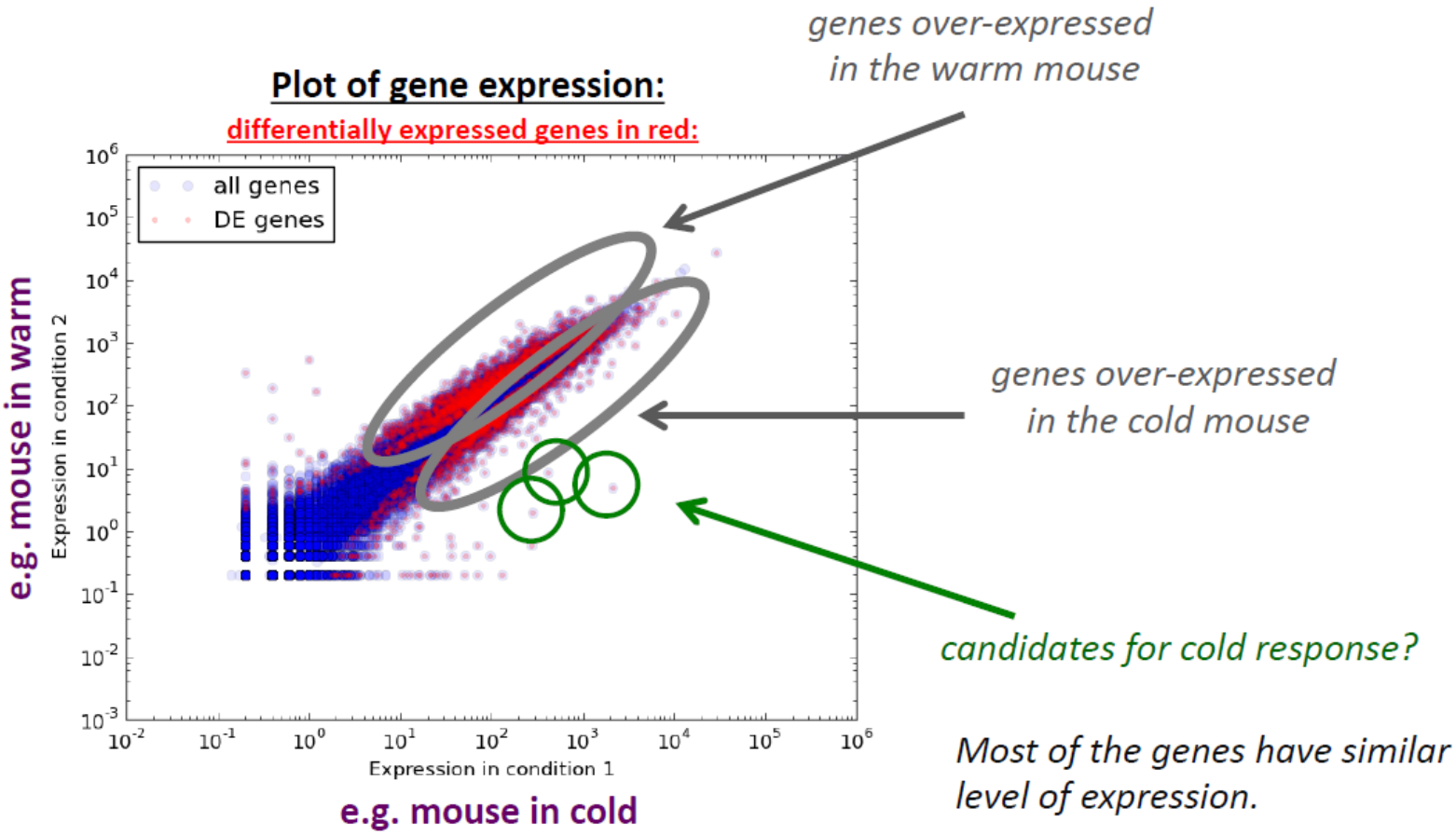
Plot of gene expression:

differentially expressed genes in red:



RNA-seq and expression studies

Expression profiles are comparative, i.e. there is always a relative comparison

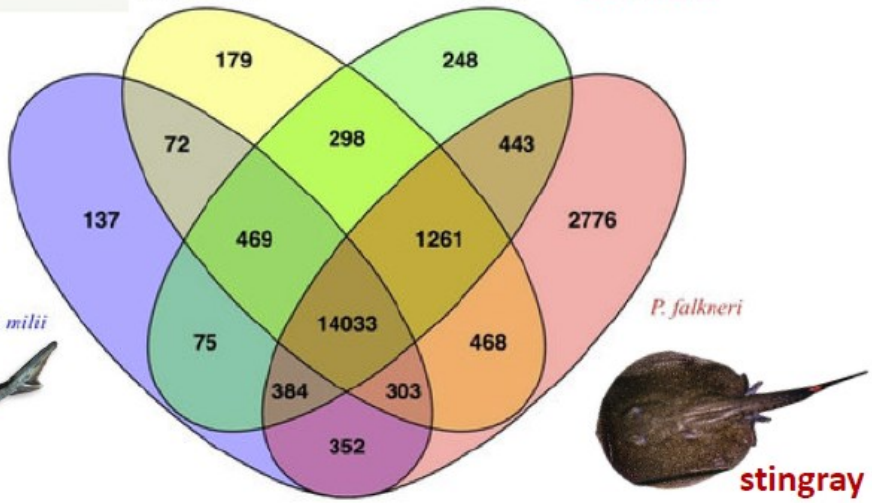


Real data:

Transcriptome diversity:

coelacanth

zebrafish



elephant shark *C. milii*

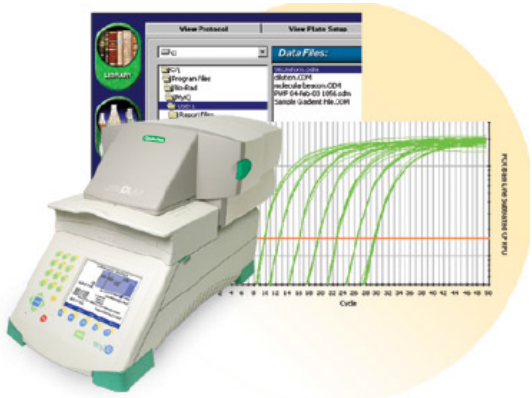
stingray

- 14‘033 genes found = expressed in all 4 species
- 2‘776 genes found only in stingray
- 179 genes found only in coelacanth
- etc...

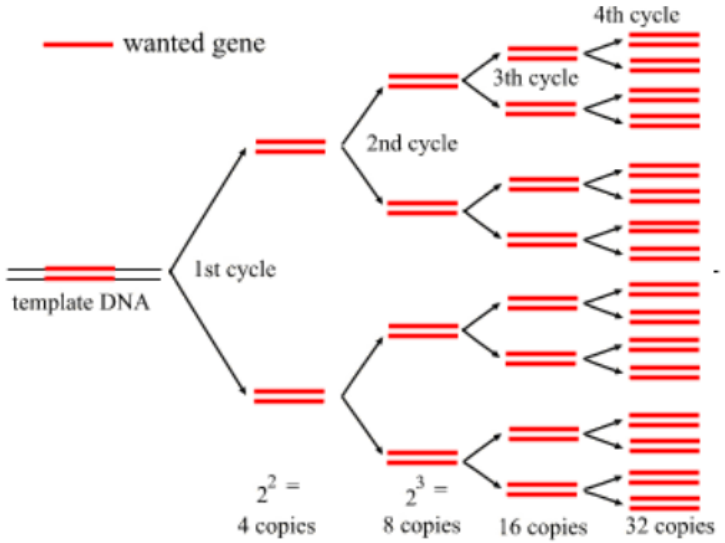
De Oliveira et al., 2016, Scientific Reports

Other gene expression methods:

traditional method – quantitative real time PCR

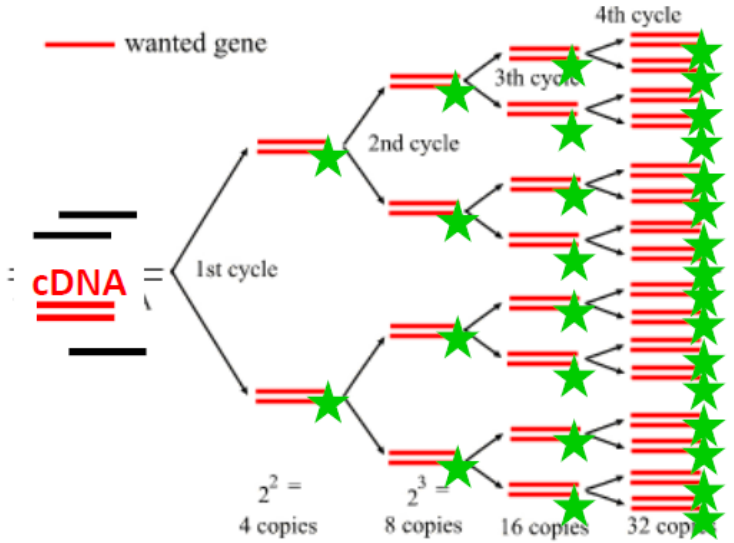


„normal“ PCR (from genome)



runs 35 cycles – in the end a lot of product –> can be used for sequencing or so...

„real time“ PCR (from RNA - cDNA)

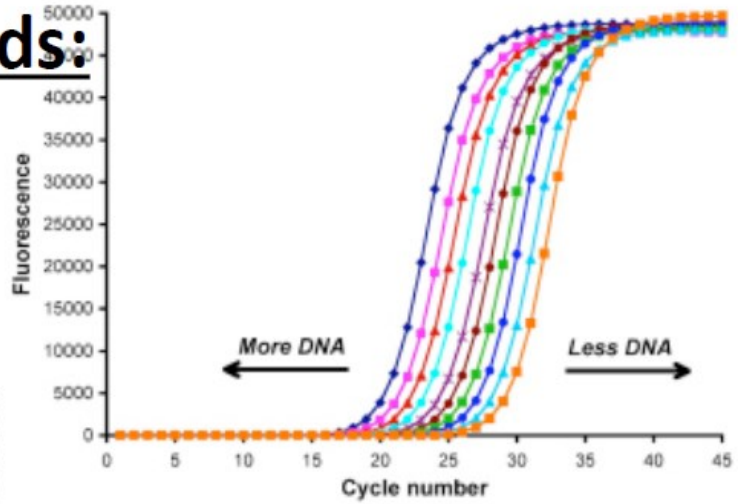


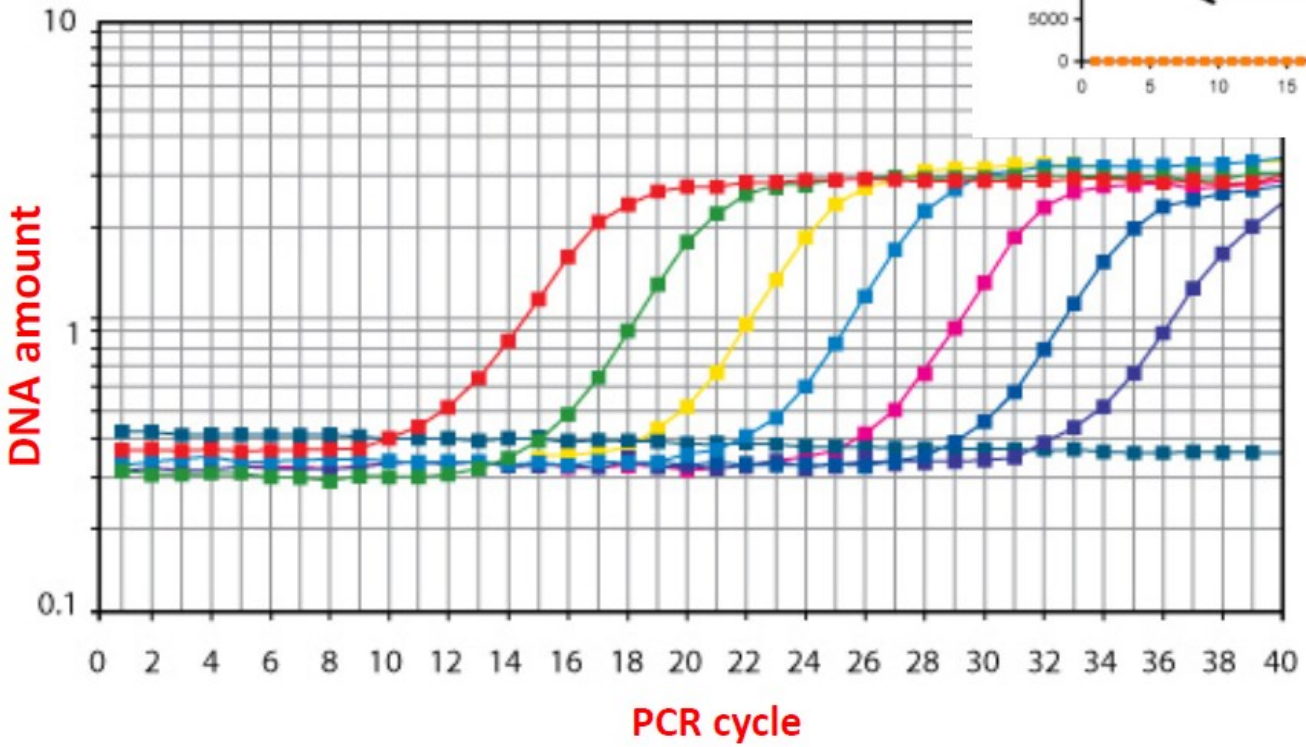
runs 35 cycles – every cycle measures content!! –> product after run is useless...

★ = fluorescent color binding to DNA

Other gene expression methods:

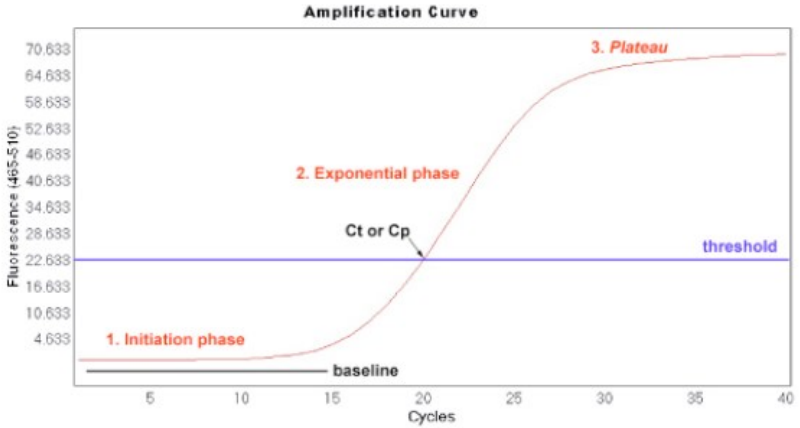
Traditional method (before NGS)
– quantitative real time PCR





difficulties: only one gene per tube; primers for each gene and/or each species...
=> can test candidates, but not search for them..

Example of real-time PCR results



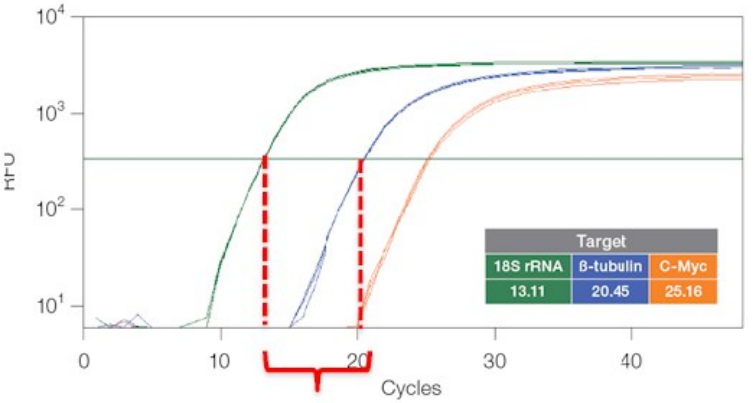
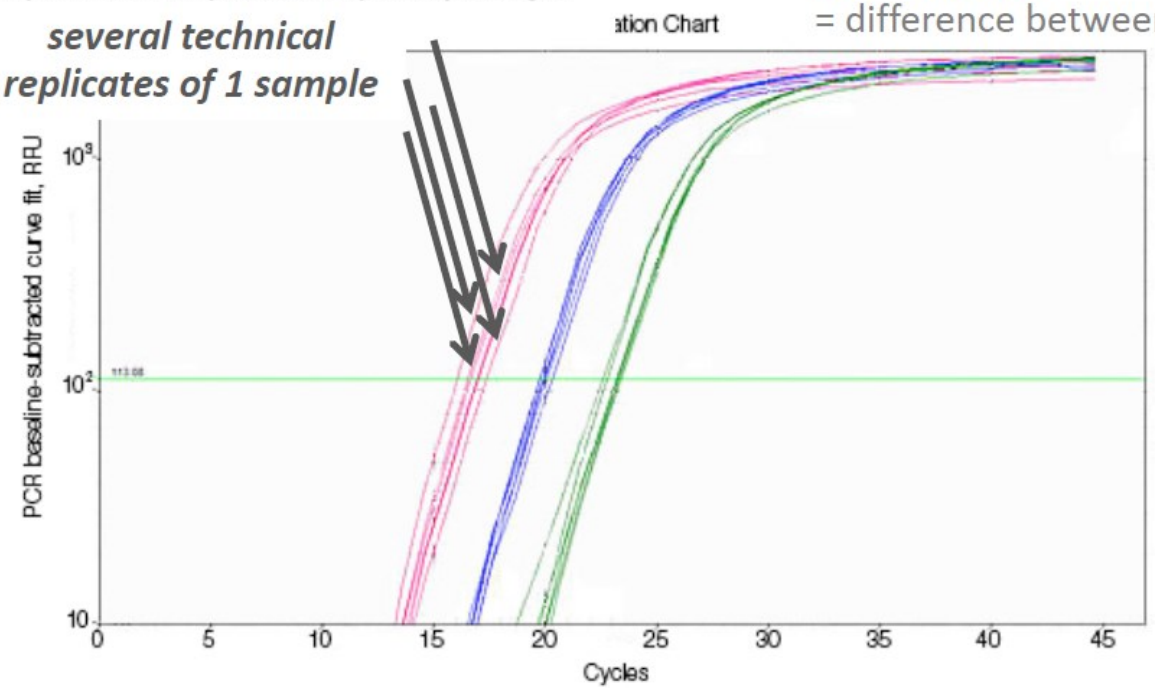


Figure 11. Phases of a PCR amplification curve. Blue: amplification curve of a positive sample. Red: negative control.

ΔCt = delta Ct

= difference between two samples

several technical replicates of 1 sample



Reference gene = usually some house-keeping gene not reacting to the treatment...

Any changes?

Target Gene A in control cells	Target Gene A in drug treated cells
Reference Gene B in control cells	Ref Gene B in drug treated cells

→ **ΔCt1 = Ct (Target A -treated) – Ct (Ref B-treated)**

→ **ΔCt2 = Ct (Target A-control) – Ct (Ref B-control)**

→ **Δ Δ Ct = ΔCt1 (treated) – ΔCt2 (control)**

ΔΔCt = delta-delta Ct
= "difference od difference"
between two samples and two genes

Normalized target gene expression level = $2^{\Delta\Delta Ct}$

Example:

gene A expression in treated cells is higher than in control and reference gene is having double expression, then:

Ct(geneA-treated) = 11 (3 cycles earlies = eight times higher expression than control)

Ct(geneA-control) = 14

Ct(geneB-ref-treated) = 22.5

Ct(geneB-ref-control) = 23.5

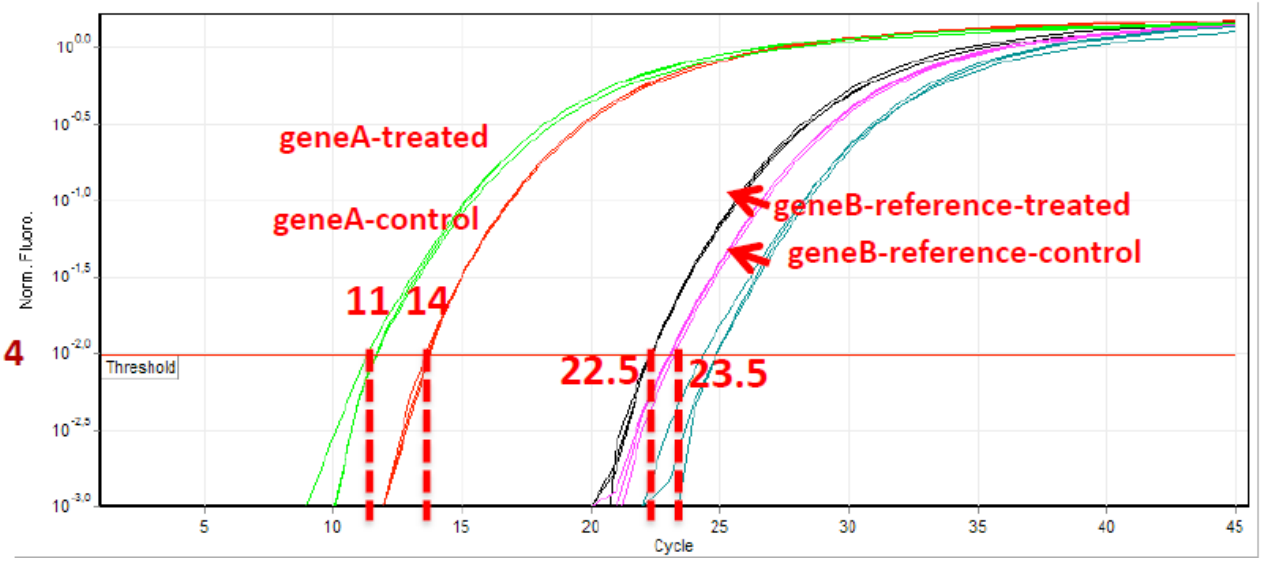
ΔCt1 = 22.5 – 11 = 11.5

ΔCt2 = 23.5 – 14 = 9.5

ΔΔCt = 11.5 – 9.5 = **2**

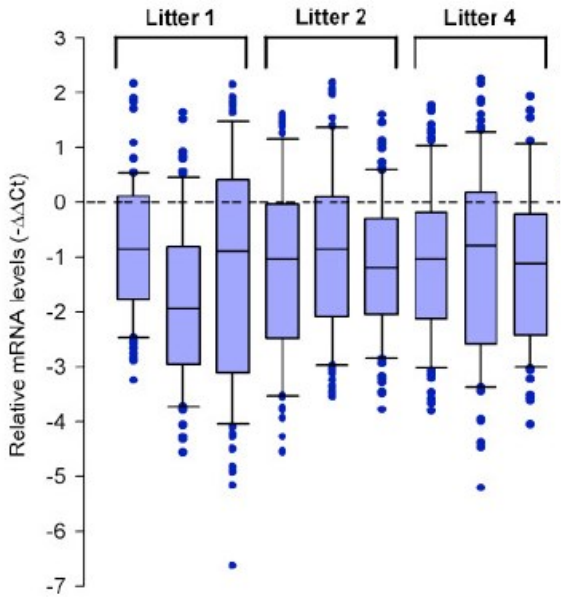
⇒ **Normalized expression = $2^2 = 4$**

⇒ **treatment causes 4 fold increase of expression**



How to present real-time PCR results

Olfactory receptors in rat (newborns vs. adults)
same data, two ways of visualization:



Rimbault et al., 2009, BMC Genomics

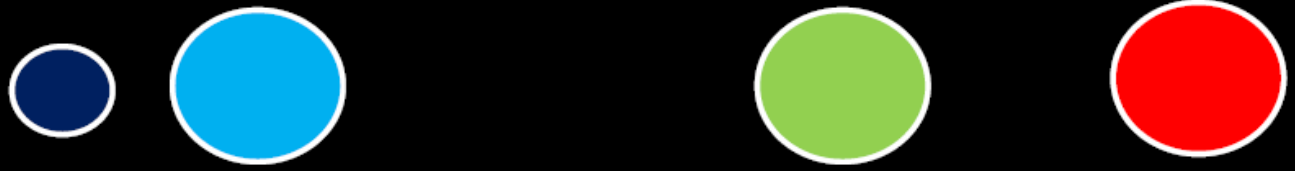
cichlid opsin genes: five „families“

cones

SWS1 SWS2B SWS2A RH2B RH2Aβ RH2Aα LWS

shallow-water species of Barombi Mbo cichlids:

expression:



they can see colours!

(c) Seafood watch; Riverpark Aquatics

cichlid opsin genes: five „families“

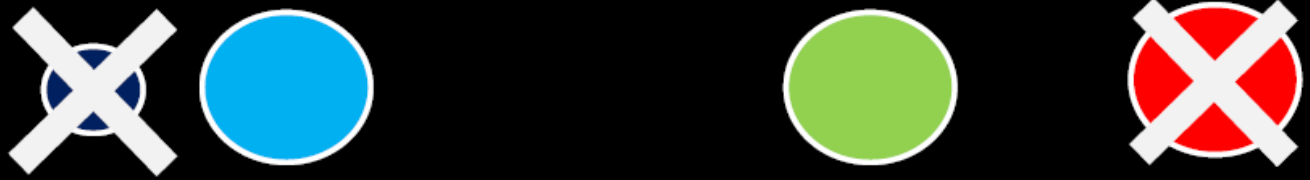
light spectrum

cones

SWS1 SWS2B SWS2A RH2B RH2Aβ RH2Aα LWS

deep-water species of Barombi Mbo cichlids:

expression:



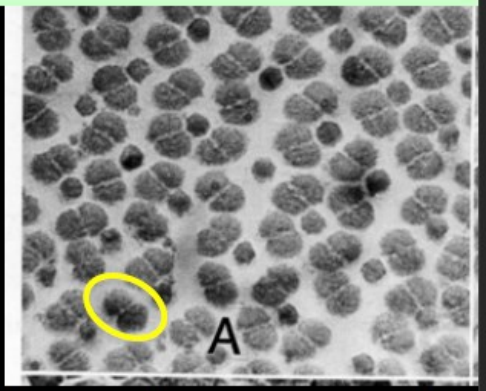
they are missing the red channel...

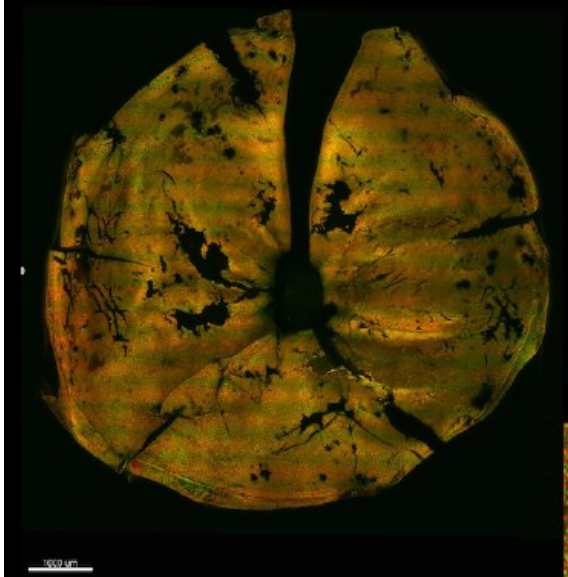
(c) Seafood watch; Riverpark Aquatics

Fluorescent in-situ hybridization

= FISH

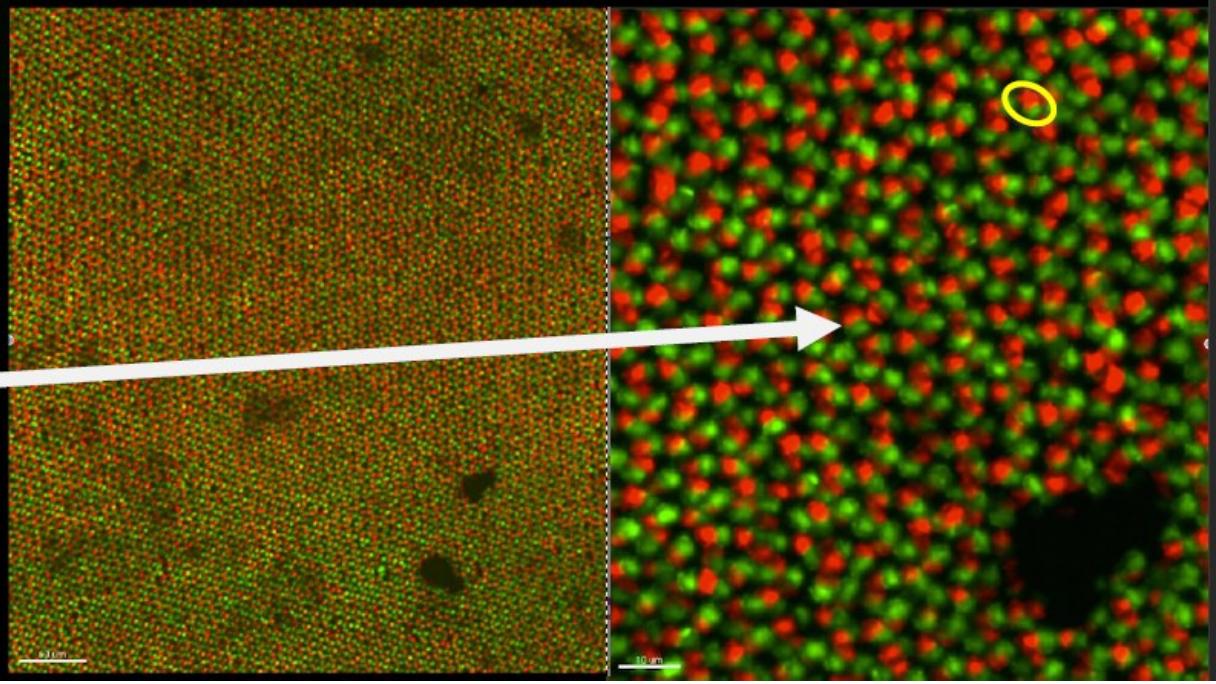
Retina of vertebrates (except for mammals) is known to be composed of double and single cones... How about photoreceptors?





Each probe – different fluorescent color (photoreceptor 1 – green, photoreceptor 2 – red

Retina labeled by two RNA probes (= different photoreceptors)



Expression of different photoreceptors is spatially separated – i.e. each cell expresses only 1 type of photoreceptors!!

Other gene expression methods:

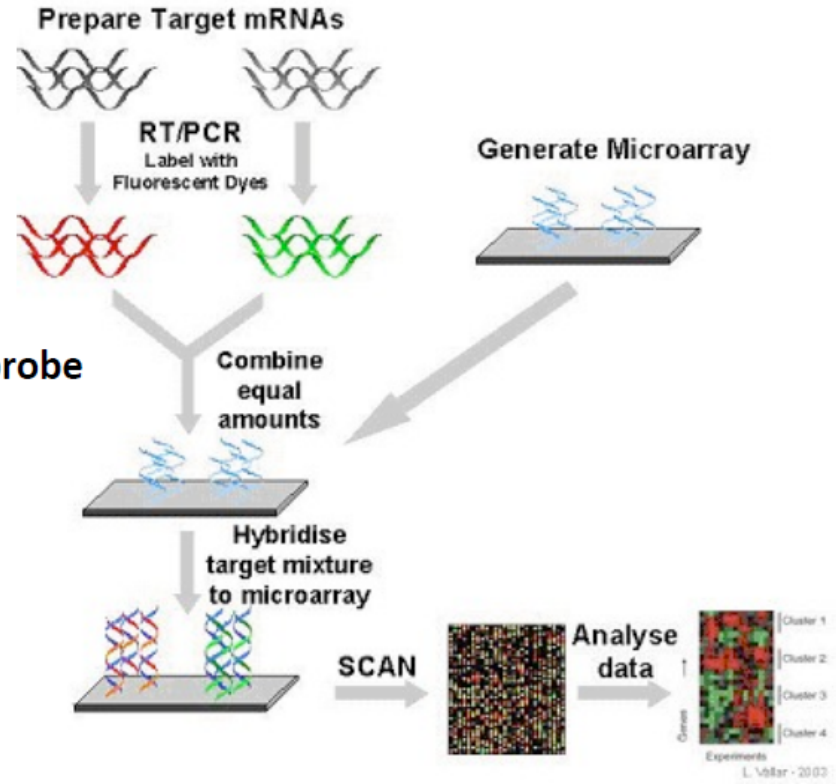
Microarrays

1) Probe = oligonucleotides covalently bound to the chip

2) Samples (cDNA) labeled with fluorescent dye

3) Sample on the chip: hybridizes with the probe

4) Fluorescent signal detected



One channel microarray
- dye Cy3 (green color) - just intensity

Two channel microarray
- two different dyes Cy3 (green), Cy5 (red)
- comparative – control / disease
- equal concentration

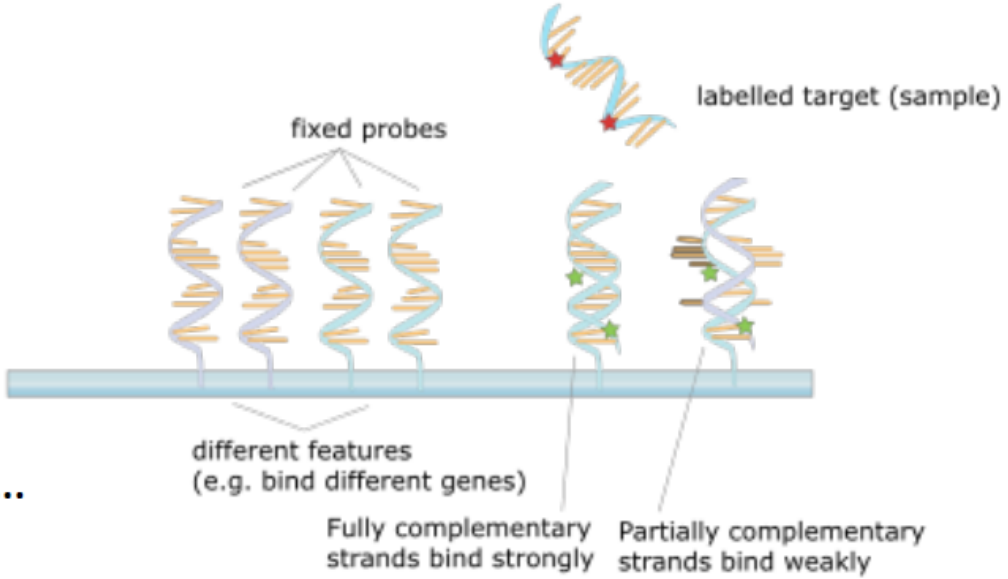
Other gene expression methods:

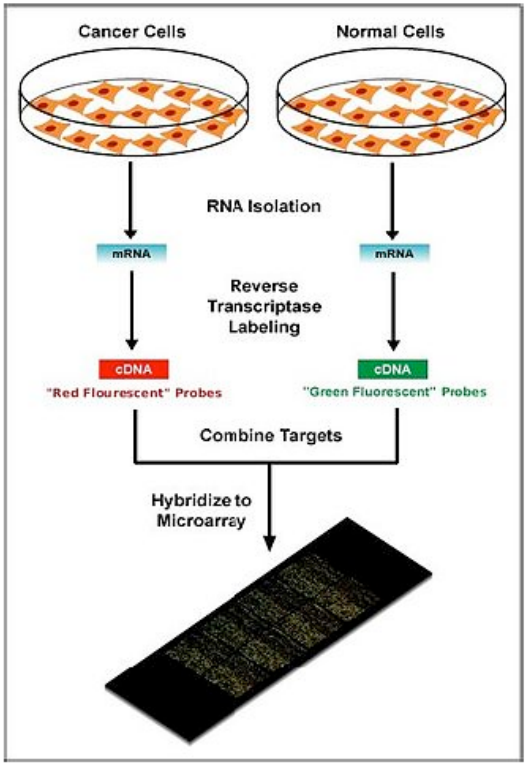
Microarrays

DNA chip, biochip

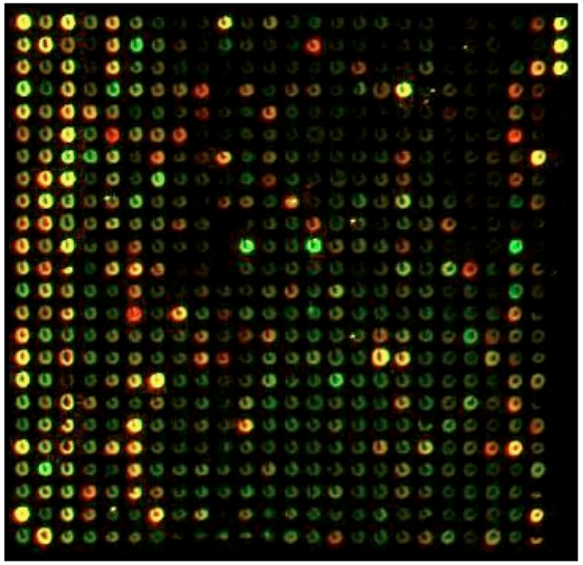
DNA hybridization

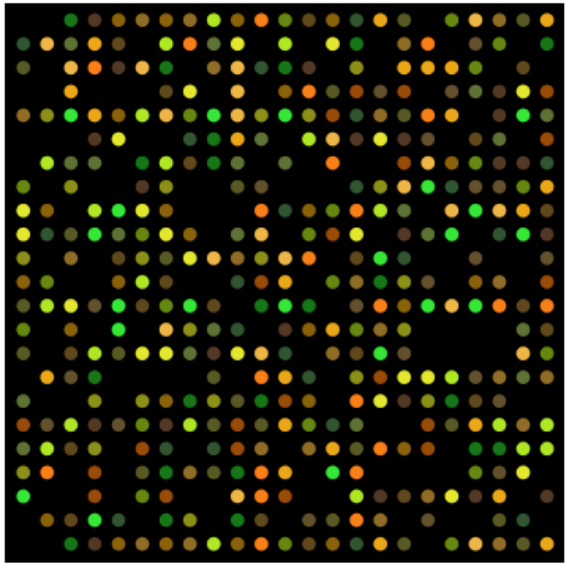
Aplications also in SNP detection, etc...





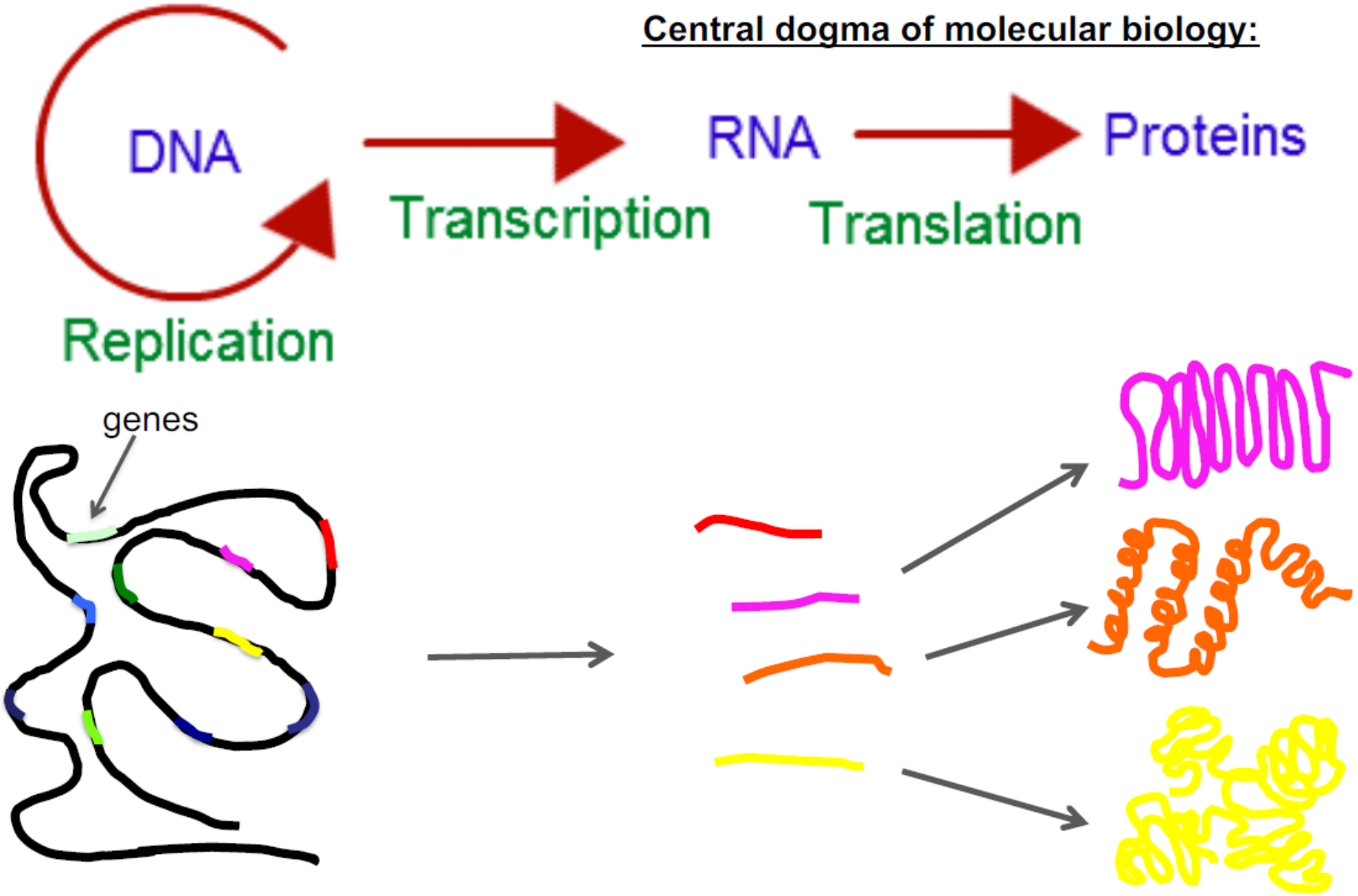
Sample – red label, control green label





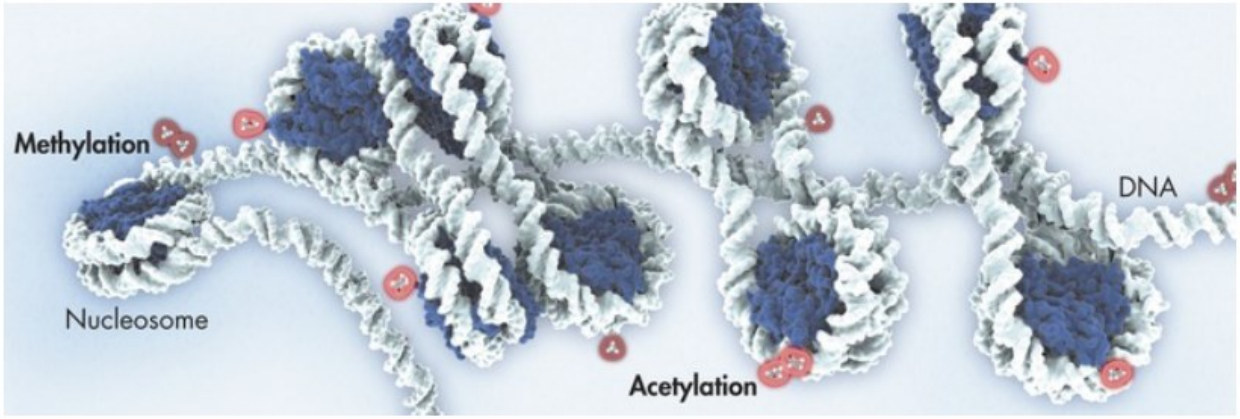
Only for model species with known genomes...

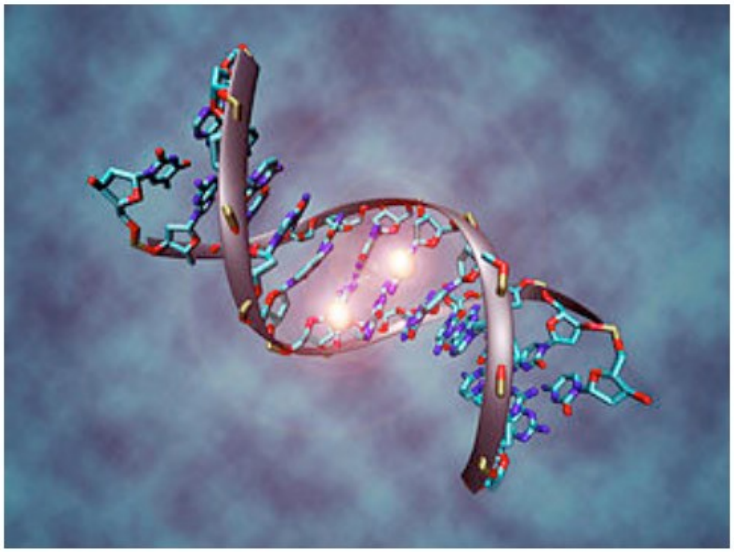
From genome to transcriptome



Epigenetic regulation

epigenetics = the study of heritable changes in gene activity that are *not* caused by changes in the DNA sequence.





Methylation of genome:

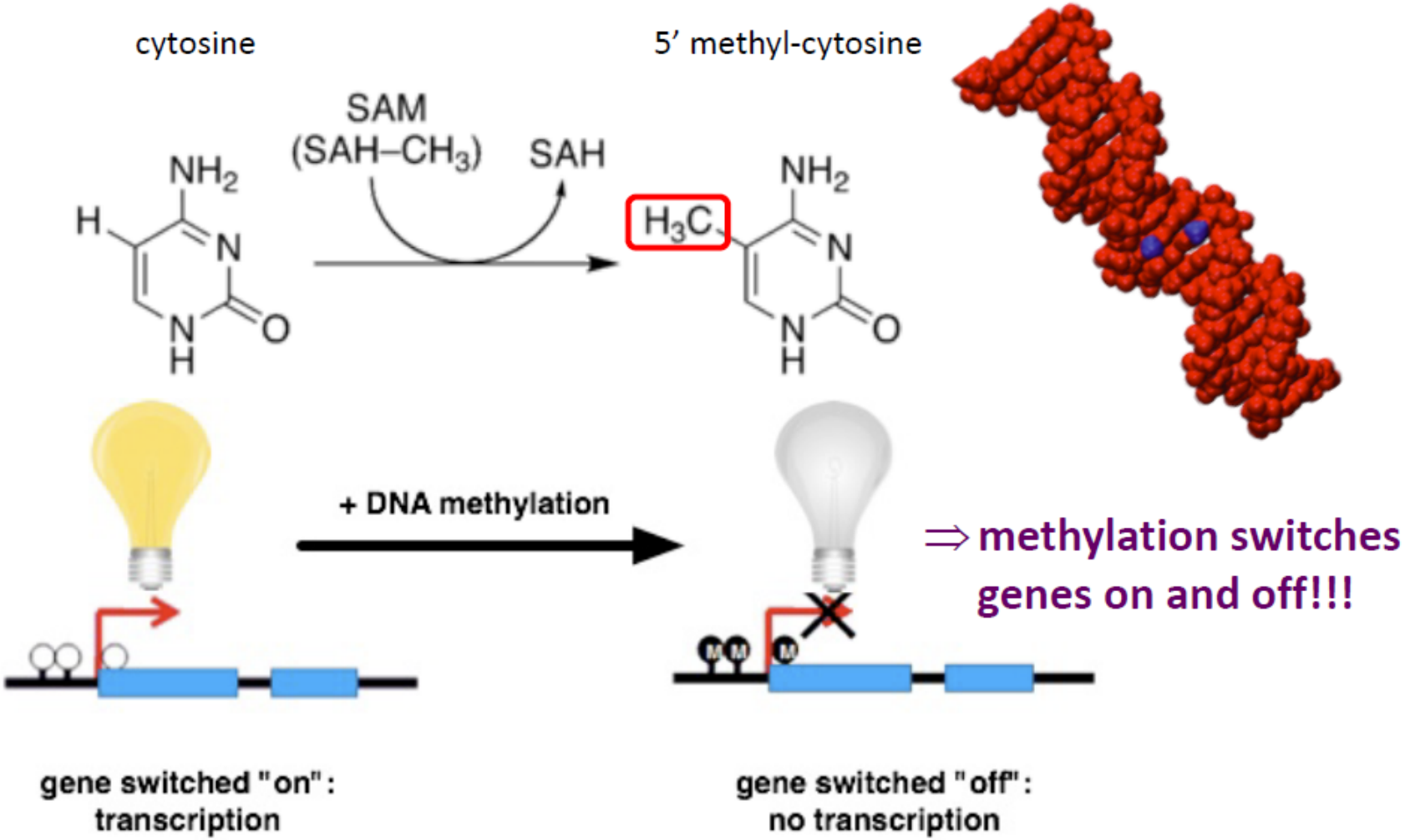


figure 1: Transcriptional silencing of gene promoters via DNA methylation

How to sequence methylation on NG sequencers:

Bisulfite-sequencing: