Analysis of gene expression
From genome to transcriptome

Central dogma of molecular biology:

DNA → Transcription → RNA → Translation → Proteins

Replication

genes
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: age: conditions:
mouse liver transcriptome: embryonic transcriptome:
is different from
mouse kidney transcriptome:
is different from
adult transcriptome:
is different from
mouse eye transcriptome: 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

Cytosine
Guanine
Adenine
Uracil

Nucleobases of RNA

RNA
Ribonucleic acid

DNA
Deoxyribonucleic acid

Cytosine
Guanine
Adenine
Thymine

Nucleobases of DNA
RNA vs. DNA

DNA – double helix

RNA – single helix => secondary structure!

16S rRNA:

tRNA:
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...
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

The structure of a typical human protein coding mRNA including the untranslated regions (UTRs)

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

**Poly-A selection**

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

- Poly-A selection
  - Using a complementary probe with the **TTTTTTTTTT** motif

**rRNA depletion**

- Removes most of the rRNA

**Best choice for:** mRNA in eukaryotes

**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:

what is the response of mouse to different (temperature) treatment?
known physiology:

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

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

all mRNA sequenced!!!
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.
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:

- genes over-expressed in the warm mouse
- genes over-expressed in the cold mouse
- candidates for cold response?

*Most of the genes have similar level of expression.*
Real data:

Transcriptome diversity:

- 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...

$\Delta \Delta Ct = \text{delta-delta Ct} = \text{"difference of difference" between two samples and two genes}.$

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 earlier = eight times higher expression than control)
Ct(geneA-control) = 14
Ct(geneB-ref-treated) = 22.5
Ct(geneB-ref-control) = 23.5

$\Delta Ct1 = 22.5 - 11 = 11.5$
$\Delta Ct2 = 23.5 - 14 = 9.5$
$\Delta \Delta Ct = 11.5 - 9.5 = 2$

⇒ Normalized expression $= 2^{\Delta \Delta Ct} = 4$
⇒ 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 opsins: five "families"

- SWS1
- SWS2B
- SWS2A
- RH2B
- RH2Aβ
- RH2Aα
- LWS

Shallow-water species of Barombi Mbo cichlids:

Expression:

- Blue
- Light blue
- Green
- Red

They can see colours!
Cichlid opsins genes: five "families"

Light spectrum:
- SWS1
- SWS2B
- SWS2A
- RH2B
- RH2Aβ
- RH2Aα
- LWS

Deep-water species of Barombi Mbo cichlids:

Expression:
- They are missing the red channel...
Fluorescent in-situ hybridization = FISH

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

Retina labeled by two RNA probes (= different photoreceptors)

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

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

Applications also in SNP detection, etc...

Sample – red label, control green label

Only for model species with known genomes...
From genome to transcriptome

Central dogma of molecular biology:

DNA → Transcription → RNA → Translation → Proteins

Replication

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

CytoSine $\rightarrow$ 5’ methyl-Cytosine

$\Rightarrow$ methylation switches genes on and off!!!

Figure 1: Transcriptional silencing of gene promoters via DNA methylation.
How to sequence methylation on NG sequencers:

Bisulfite-sequencing:

1. Sulfonation
2. Deamination
3. Desulfonation

Cytosine $\rightarrow$ uracil

A C G A C T A C G C

sequencing

A C G A C T A C G C

bisulfite conversion

A C G A U T A C G U

sequencing

A C G A T T T A C G T

reconstruct sequence

A C G A C T A C G C