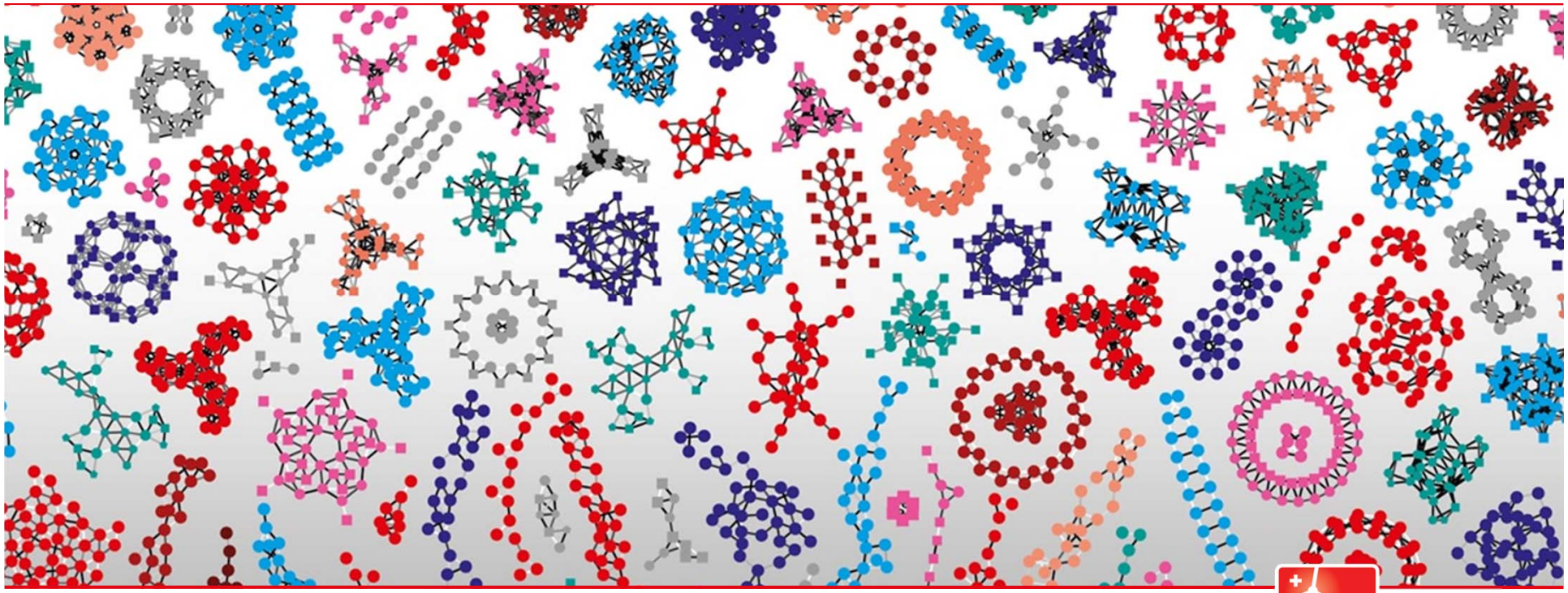


SIB
Swiss Institute of Bioinformatics

Data analysis in practice

5-6 April 2022

Isabelle Dupanloup, Rachel Marcone, Frédéric Schütz

First dataset:
qPCR dataset

- Measure of the expression of some genes in mice
- We are interested in the expression of gene AKT

- Two variables of interest:
  - **Genotype**: WT/KO mice for the MAF1 gene
  - **Treatment**: either normal food (control) or fasting (treated)

- 3 biological replicates for each group (12 in total)
- 3 technical replicates for each biological replicate

Our questions

- In MAF1 WT mice, is there a difference in AKT expression depending on the treatment ?

- Does the effect of the treatment (control/treated) depend on the MAF genotype (WT/KO) ?

[Note: if you perform the two analysis, you may get results that look contradictory. Can you explain why ?]

- The data is provided in the Excel file qPCR.xlsx, which you have to load into R for analysis

- The data will need some cleaning before being useful

- There is information in the file that can be used for normalizing the data, using the expression of GAPDH ("housekeeping gene"). You do not have to perform this step (you can do it as a second, optional, step if you want)

The data file

- sample.ID: a unique ID that identifies each biological replicate
- group: the experimental group (WT/KO and C/T)
- gene: the gene measured
- ct: the measured value (number of cycles)

Additional calculations are included in the file (but should not be loaded into R):

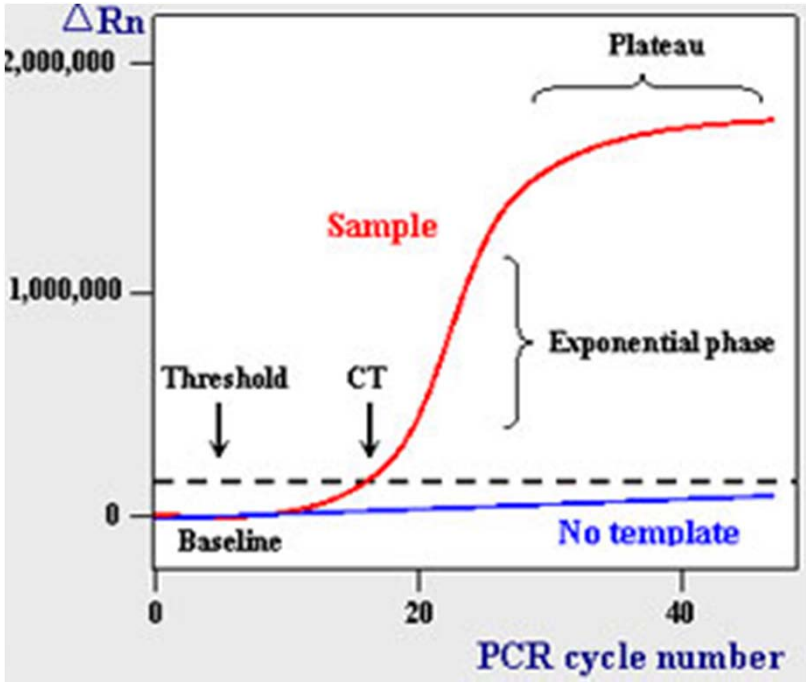
- dct: the difference between the gene of interest and the housekeeping gene
- average dct: the average dct over the technical replicates for a given biological replicate
- There is also a calculation of ddct and the log fold change (2^(-ddct))

Some background on the qPCR technology

Real Time Polymerase Chain Reaction

- Extension of the PCR technology
- Use fluorescence to measure the amount of expression at every cycle.
- Measure the number of cycles (Ct, “cycle threshold”) required until the fluorescence crosses a given threshold.
- This measure is made during the exponential phase.
- The threshold is arbitrary

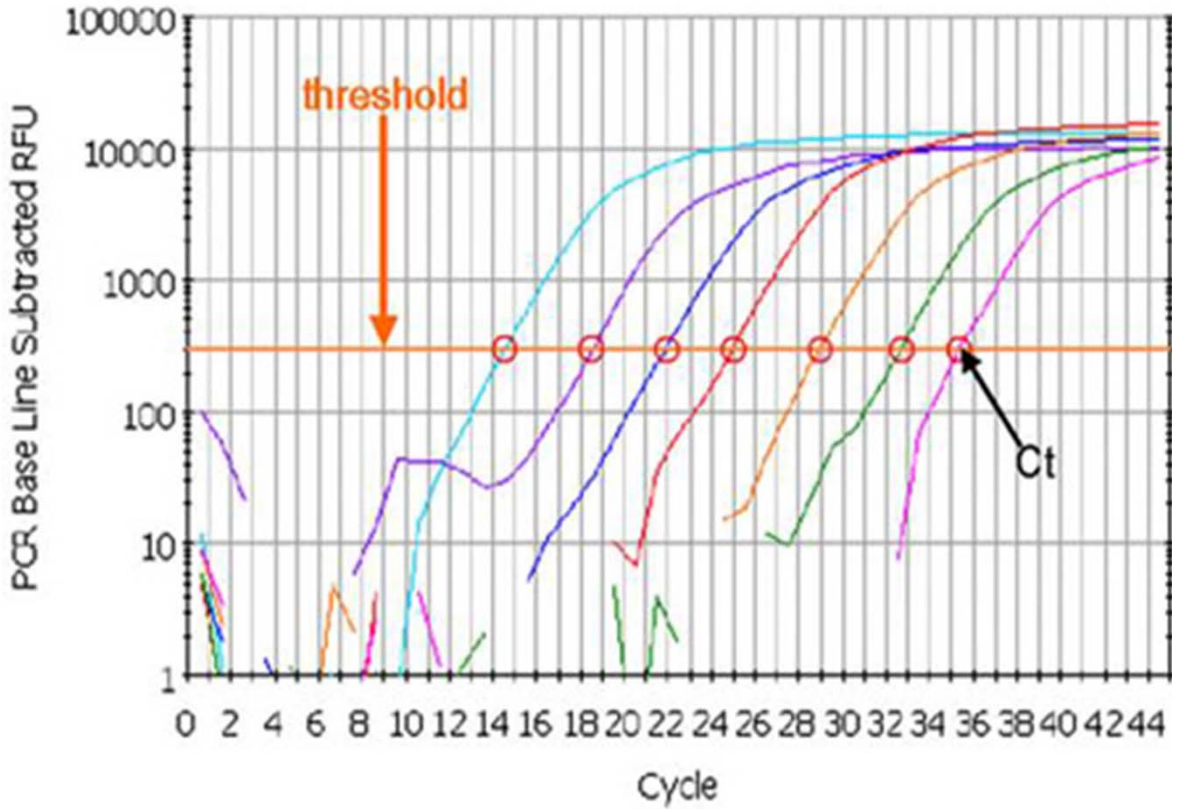
- **Gene more expressed = lower Ct value**



http://www.rt-pcr.com/

A real example

Log scale



Plateau zone

Background zone

http://bio-analyze.com/OurRealtimePrject1.htm

Normalization

- The measured expression must be normalized for difference between samples (e.g. amount of starting material).
- This is usually done using a *reference gene* (or *standard gene*, *housekeeping gene*), which should be expressed in all cells and have the same number of copies in all cells
- Examples of typical housekeeping genes:
  - Glyceraldehyde 3-phosphate dehydrogenase (GAPDH)
  - Beta actin
- The assumptions are very stringent and not always satisfied…
- Common recommendations: take several (at least 3) housekeeping genes.

http://en.wikipedia.org/wiki/Housekeeping_gene

How to measure differential expression: the ΔΔCt method

The efficiency is assumed to be perfect (100% = exact doubling at every cycle), and the efficiencies for the gene of interest and the reference should be similar.

Ct_g = Ct for gene of interest

Ct_h = Ct for housekeeping gene

$$\Delta Ct_g = Ct_g - Ct_h$$

Difference between conditions 1 and 2:

$$\Delta\Delta Ct_g = \Delta Ct_{g2} - \Delta Ct_{g1}$$

Log fold change = $-\Delta\Delta Ct_g = \Delta Ct_{g1} - \Delta Ct_{g2}$

Fold change = $2^{-\Delta\Delta Ctg}$

(The base is 2 because a difference of one cycle represents a doubling of the amount of the material. In comparison to other assays, the base of the logarithm is not arbitrary here)