# Practical – Microarray analysis – Data loading and exploration

(STAT2)

*Jacques van Helden*

*2020-03-06*

## Contents



### Introduction

In this practical, we will load a dataset that will be used as study case to apply different approaches of multivariate analysis:

- data exploration
- multidimensional scaling
- differential analysis
- clustering (unsupervised classification)
- supervised classification

### Study case

**Reference:** Den Boer ML *et al.* (2009). A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. Lancet Oncol. 2009 10:125-34. [doi: 10.1016/S1470-2045(08)70339-5], [PMID 19138562]. Data available at Gene Expression Omnibus, series [GSE13425]

### Data pre-processing

The raw microarray data has been pre-processed in order to dispose of a ready-to-use dataset. pre-processing included

- filtering of barely detected or poorly expressed genes,
- log2 transformation to normalise the raw measurements
- between-sample standardisation to enable comparison between the different samples.

### Availability of the pre-processed data

The preprocessed data is available here: https://github.com/jvanheld/stat1/data/DenBoer_2009.

It contains the following files.

| File | Contents | Structure |
|---|---|---|
| GSE13425_group_descriptions.tsv.gz | Description of the patient groups | Tab-delimited file with one row per group and one column per type of description (group name, label) |

| File | Contents | Structure |
|---|---|---|
| phenoData_GSE13425.tsv.gz | Metadata (sample descriptions) | Tab-delimited file with one row per biological sample (one per patient) and one column per type of information about the biological sample |
| GSE13425_AMP_Whole.tsv.gz | Normalised microarray data | Tab-delimited file with one row per gene and one column per patient |
| GSE13425_AMP_Whole.tsv.gz | Detection status of each gene in each sample (Absent, Marginal, Present) | Tab-delimited file with one row per gene and one column per patient. |