**Practical 3: biological databases, data repositories and basics of data processing**

In this practical you will learn how to use repositories of public omics data and how to process and interpret such data and its quality. You will compute a list of differentially expressed genes for a transcriptomics experiment related to thyroid hyperplasia and cancer. Also, you will consult databases with biological knowledge to find out about the properties and functions of genes and metabolites. Next week, in practical 4, you will apply pathway and network analysis to further biologically interpret the outcome of the experiment.

**Part I: Consulting omics repositories: ArrayExpress**

One of the databases that contains data generated from experiments is ArrayExpress (<https://www.ebi.ac.uk/arrayexpress/>). Besides the data itself, this database contains information on the setup of the experiment itself and on the different samples that were generated (called *metadata*). This additional information is required to understand and interpret the generated data as well as to be able to reuse it.

Open ArrayExpress and find the experiment with identifier E-GEOD-54958 using the search box.

Now consult the information on this experiment and answer the following question:

1. What is the title of the experiment?

Expression profiling in proliferative thyroid lesions of different histologies and normal thyroid tissue

1. In which year was this data set released?

2015

1. Which platform was used to generate this data (microarray, sequencing) and which brand and type?

Platform: microarray; brand: Affymetrix; type: GeneChip Human Gene 1.0 ST Array

1. Which experimental groups were compared in the experiment?

Papillary thyroid carcinoma (PTC) with a BRAF mutation; PTC without a BRAF mutation; follicular adenomas (FA); follicular variant of PTC (FVPTC); normal thyroid tissue (TN)

1. Open the detailed information at ‘Protocols’. Which amount of RNA (in μg) was used for each sample?

6 μg

1. Open the detailed information at ‘Samples’. How many samples were in each of the experimental groups?

PTC with mutation: 14; PTC without mutation: 11; FN: 7; FVPTC: 6; TN: 7 – total: 45

1. For each sample ‘raw’ and ‘processed’ data are available. What is the difference (without opening them, you cannot open the raw data)?

The raw data has not been normalised, the processed data has been normalised (in this case with the RMA method, as you can read in the ‘Protocols’ section)

1. Open the normalised data for one sample. What information is in the two columns it contains?

It contains an ID\_REF column, representing the identifier for a certain probeset (Affymetrix identifier representing the probes for a gene); furthermore, it contains a column VALUE, containing the normalised expression value for that gene – note: the normalised expression value is given on a 2log scale (this is commonly done)

As you can see from the main page of the experiment, the raw and normalised data and the experimental descriptions can also be downloaded as tables and zip files. This allows you to further study the experiment and to reprocess the data. We have done this for you, and created some results that you will explore in the next part of the practical.

**Part II: Processing publicly available data: quality control using ArrayAnalysis.org**

ArrayAnalysis.org is a web portal for microarray data analysis. It allows you to upload data for online processing, or you can download scripts to process the data on your own computer. We have applied the ArrayAnalysis.org workflow to generate quality control (QC) images, to estimate the quality of the data set. You can download the images from student portal (‘QC images.zip’).

1. Open the boxplot of raw signal intensities per sample (‘RawDataBoxplot.png’). What does this tell you about the data quality?

The signal intensity seems to be a bit variable between samples: across all groups some samples have lower signal strength; also, the group of normal samples seems to have a lower signal overall; signals are never totally comparable, it requires experience and consultation of all QC plots together to say what difference is to be expected / acceptable.

1. Open the boxplot of normalised signal intensities. What does this tell you about the data quality?

The boxplot shows that the signal distributions between the samples have been fairly well normalised, except for the group of normal controls, which is still very much deviating; it is hard to tell whether this is biological information (the normal tissue may be quite different from all neoplastic tissue) or an experimental artefact (maybe the normal tissue was obtained from a different hospital or processed in a different lab).

1. Open the density plot of normalised signal intensities. Explain what this tells you.

The density plot allows to check the signal distributions of the samples in more detail and shows that they all are still somewhat deviating (also within the normal and neoplastic groups); based on this plot alone, it is hard to say whether this poses a problem or not.

1. Open the cluster plot. What does this show about the samples?

The cluster plot first of all shows that the normal tissues samples cluster quite apart from all other samples (as we would have expected); in the other cluster it shows two subclusters, one containing almost only the FA and FVPTC samples, the other one containing almost only PTC samples (with or without BRAF mutation); there are no individual outliers, but some samples are in the ‘wrong’ group: PTC\_BRAF\_mut\_9, PTC\_BRAF\_wt\_10, FVPTC\_1, and FA\_5.

When we look at the subclusters in more detail, we see that the FA / FVPTC one is again subdivided in two clusters: one with almost only FA samples (except for FVPTC\_6) and one with almost only FVPTC samples (except for FA\_1); the PTC one is subdivided in three (be careful!) subgroups: one with mut\_1, 2, and 3; one with almost only mut samples (expect for wt\_4); one with almost only wt samples (except for mut\_13).

1. Open the array-array correlation plot. What does this indicate?

The array-array correlation plot shows similar results, mainly that the normal tissue samples are quite different from the other samples; this plot shows – besides a clustering – also the correlation coefficients between the arrays (based on the expression of all genes in the samples): these show that correlation coefficients range between 0.625 and 0.975 (1 in fact, but these are the autocorrelations for each sample with itself, which is always 1); correlations also show you that normal tissue samples TN\_5 and TN\_6 are less deviating from the neoplastic samples as the other normal tissue samples.

1. Open the principal component analysis (PCA) plot. What does this show about the samples?

A PCA plot shows a representation of the most important variation in the data set (in this case between the samples) in a reduced dimensionality, which allows us to visualise the data (details on how this works are beyond scope now); with each axis, the percentage indicates which fraction of the total variance is explained by that axis.

This plot shows again that the normal tissue samples are quite different from the others: they cluster separately, in the first component which represents over 40% of total variation in the data set. Then, for the neoplastic samples, the second component (almost 7%) seems to mostly separate the PTC samples from the PA and FVPTC samples, though the clusters overlap in the middle (and some samples are in the ‘wrong’ group). The third component (a bit over 5%) does not do much more, one sample in the PTC group with BRAF mutation seems to be separated out from the others, but that probably does not mean much. Note: without labels on the plot, it is hard to tell which samples is exactly which sample, ArrayAnalysis.org code also allows to add labels, but that is not needed now. Also note that the symbols in the legend are accidentally partially incorrect, the colours are correct.

**Part III: Processing publicly available data: statistics using ArrayAnalysis.org**

We have also applied the ArrayAnalysis.org workflow to generate statistical results based on the data from this experiment. In order to do so, one has to first determine which groups one wants to compare.

If we name (as on the QC plots):

* PTC\_mut = PTC with BRAF mutation
* PTC\_wt = PTC without BRAF mutation
* FVPTC = follicular variant of PTC
* FA = follicular adenoma
* TN = normal thyroid tissue

Then we compared each group to each other group: PTC\_mut versus PTC\_WT; PTC\_mut versus FVPTC ; PTC\_mut versus FA, PTC\_mut versus TN; PTC\_wt versus FVPTC; PTC\_wt versus FA; PTC\_wt versus TN; FVPTC versus FA; FVPTC versus TN; FA versus TN. So ten comparisons in total. (I WILL PICK SOME)

HERE THE QUESTIONS: something on P value histograms, FC histograms, expected number of significant genes, and some on the table of statistics (most significant genes, most changed, …)

Now we will look up some more information about the most clearly changed genes. Of course, this does not take into account the totality of changes observed in the data set. Next time, in practical 4, you will apply pathway and network analysis methods that allow you to take all results into consideration and interpret them as a whole.

**Part IV: Evaluating findings by consulting genomics databases: Ensembl**

For one or two genes find something about what they do and some functional annotation

**Part V: Metabolomics**

look very briefly into (another) metabolomics data set: only open results tables from the paper and confirm how limited the number of identified metabolites are

**Part VI: Finding more information about metabolites: HMDB**

Look up some information about T3/T4 in HMDB