Microarray data analysis lab 2 (in Feb 20)

1. Find experiment E-UMCU-12 in ArrayExpress (www.ebi.ac.uk/arrayexpress), open the full experiment view (click on the title line)
2. Explore the microarray design used in this experiment: click on Array link to A-UMCU-4, then open the spread sheet view as tab delimited file or the xls file
3. Click on link » View detailed data retrieval page... , explore some of the links, find the Processed data group 1. Do the following selection:
    a. Keep all experimental time points selected
    b. In Quantitation types select both Cy3 normalised and Cy5
    c. In Design Element Properties select Database:SGD and Reporter Name
    d. Click on Export Data, then Upload data (save data in your directory on your PC).
4. Go to Expression profiler (www.ebi.ac.uk/expressionprofiler, log in unless you've already done so), select Data import, Expression Data, Custom format, tab delimited, and enter x=3 (we exported two annotation columns), species = Sacharomices Cerevisiae
5. Observe that in the selected data there are two columns per time-point: Cy5 and Cy3 signal (the first is the reference time point 6.5h, the second the current time-point).
6. Go to Data selection, Select columns, and sub-select only the odd (Cy3) columns (note that the even columns all represent the reference time point 6.5 hours)
7. Try Hierarchical clustering – because we are not working on the logarithmic scale, the clustering is not very informative
8. We want to transform the signal into log ratios. Go to Data transformations, select the original imported data (containing both Cy3 and Cy5 columns), select Intensity -> (Log N) Ratio tab, in 'Number of data columns per experiment' type in 2, in 'Channel 1 data column (within experiment)' - 1, in 'Channel 2 data column (within experiment)' - 2, and 'What log to take' - log 2
9. Explore the data line plots - observe that first data column is an outlier – you may want to remove the first time point (by selecting columns 2-34)
10. Explore the data, observe that many rows are annotated as 'Control …'. These can be removed – go to Data selection, Select Rows, type in 'control', select 'Use approximate matching' and 'invert selection', click Execute
11. Go to Data Selection, Select Rows (Gene selection), type in a gene ID (look this up in the data matrix, e.g., type in YOL004W) and explore the result – observe replicate gene expression patterns – are they similar? Do the same (i.e., go back to Data selection, then via Select by similarity, and select 10 closest genes.
12. Go to Hierarchical clustering, select Euclidean distance and Cluster only genes (as this is a time-course, you do not want to change the order of the columns). Select one of the clusters (click on any point in the dendrogram) and save the cluster. Go back to data selection, click on the results of the selected cluster. Explore if replicate spots (e.g., the ones representing the same genes cluster together).
13. Go to Data selection, Select some number (e.g., 400) of most changing genes,
14. Go to clustering
    a. Try hierarchical clustering
    b. K-means and K-medoids clustering
    c. Explore different distance measures and clustering options
15. Go to annotation, and explore GO term overrepresentation in one of the clusters

Microarray data analysis lab 3 (February 22, 23)

1. Find experiment E-MEXP-886 in the ArrayExpress Repository. Explore it's properties – Mus muscles, 10 Affy hybridisations, 5 wilde type, 5 gene (Ataxin) knockouts (first 6 columns from the experimental design xls). Click on some of the links, e.g., experimental design xls (this will look better in Explorer than in Mozilla browser) – the first 6 columns of the xls table:

Sample name	Age	BioSourceType	DevelopmentalStage	GeneticModification	Genotype
ataxin1WT-1753	15	fresh_sample	adult		Wild type
Ataxin1WT-1756	15	fresh_sample	adult		Wild Type
Ataxin1WT-1863	15	fresh_sample	adult		Wild Type
Ataxin1WT-1869	15	fresh_sample	adult		Wild Type
Ataxin1WT-1364	15	fresh_sample	adult		Wild Type
Ataxin1KO-1307	15	fresh_sample	adult	gene_knock_out	Ataxin1-/-
Ataxin1KO-1749	15	fresh_sample	adult	gene_knock_out	Ataxin1-/-
Ataxin1KO-1750	15	fresh_sample	adult	gene_knock_out	Ataxin1-/-
Ataxin1KO-1751	15	fresh_sample	adult	gene_knock_out	Ataxin1-/-
Ataxin1KO-1919	15	fresh_sample	adult	gene_knock_out	Ataxin1-/-

2. Find Atxn1 in ENSEBL database (you will find gene ENSMUSG00000046876, Transcript ENSMUST00000091628). Look for Affymetrix mouse arrays in Ensembl. Unfortunately Affymetrix mouse array MOE430A is currently not available in Ensembl (there is MOE430). In MOE430 one finds design elements for Atxn1 – 438294_at, 1450499_at.
3. Explore array by clicking on **A-AFFY-23,** then **Spreadsheet (tab-delim.) >>**, look for Atxn1 in the annotation (edit, find in this page) – bad luck, look for 1450499_at – you will find it, but it's not annotated by anything.
4. Download raw data (E-MEXP-886.raw.zip) from this experiment to your PC
5. Click on the link » View detailed data retrieval page... , generate a Processed data matrix by making the following selections
    a. Leave all experimental conditions selected
    b. In the Quantitation types select 'Affymetrix CHPSignal
    c. In Design element properties select 'Composite sequence name, Ensembl, MGD, Swall' (you may also make additional selections)
6. Click Export data, then Download data matrix, and save on your PC
7. Now go to Expression profiler (log in, unless you've already done so), select data import, expression data
8. Select Affymetrix, E-MEXP-886.raw.zip, click Execute (note, in general if you upload zipped Affymetrix data files, you have to make sure that the all cel files in the archive are of the same tipe)
9. Explore the intensity distribution plots and the box plots
10. Go to 'Data normalisation', choose one of the four normalisation methods, and click on Execute – explore the expression value distribution plots after the normalisation by going back to Data selection. Try a different normalisation.
11. Go back to data selection, explore the normalised intensity distribution plot
12. Go to data transformations – chose transformation Abs->Rel, use average gene value as a reference, and select appropriate post transformation for the RMA/VSN normalisation, Execute

13. Go to data selection, explore the distribution plot
14. Go back to Data Import, and upload the normalised data in custom (tab delimited) format, browse and fine the right file (e.g., E-MEXP-886_1640492963_-1429975046_-556918763.txt.csv), set x=5 (because we selected 5 annotation types in the step 5c.), select species Mus musculus, Execute
15. Go to Data transformation, perform Abs->Relative, use average gene and log 2 transformation. Now you have several different normalised log scale relative datasets.
16. The remaining steps you can do either using any of the normalised relative datasets you have obtained, i.e., data from step 12 or step 15 (you can get back to the dataset from step 12 by going to File, select Data Management, and select the respective dataset)
17. Go to data selection, chose Select by similarity, type in 1450499_at, then click Execute – observe that the signal for this probe-set is rather low – probably this probe not working (in the last 5 conditions the respective gene has been knocked out, and what you observe apparently is a cross hybridisation).
18. Go to Statistics, t-test, select Two classes, type in 1-5 for the first class (normal mouse), 6-10 for the second (knock-out mouse), see the list of most differentially expressed genes.
19. Go back to Data selection, use Select by similarity and type in the probset names for the two most differentially expressed genes found in the previous step – one on the positive side, the other on the negative side
20. Go to ArrayExpress, get the array design and look up how the most differentially expressed probe-sets are annotated – try to find one differentially expressed gene that has annotation that you can then find in Ensembl, follow this up via Ensembl
21. Go to data selection, subselect genes that are at least 1 standard deviation variable in 50% of samples, and then select 500 most variable genes (or simply some, e.g., 500 most variable genes)
22. Do some clustering, see which samples cluster together
23. Go to Ordination, Between Group Analysis, click on Define new factors, define a new factor with two groups – ‘normal/knock-out’ and values ‘normal’ and ‘knock-out’. Save factor
24. Select the new factor in factors window and click Execute. Explore the output – how well the two groups are separated and what are the separating genes.
25. Try a few other steps (e.g., GO analysis, cluster comparison, etc)