**500697 Genetic Analysis**

**Bioinformatics practical**

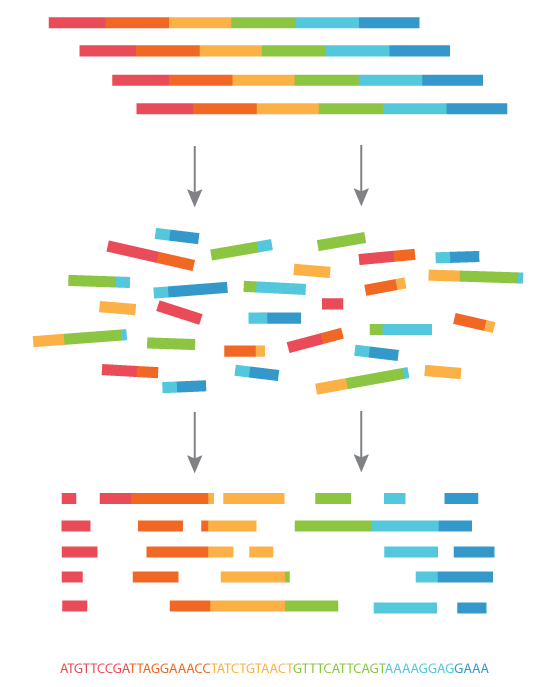
**Analysis of next-generation genetic “-omics” data**

Adam Bates & Katharina Wollenberg Valero

Date: 04/12/2018

Place: Hardy NTL

Time 9:30am-3:45pm



**Learning Outcomes**. Gain competency in navigating between different bioinformatics platforms, languages and public databases to analyze and interpret genomic data.

**Genomic resources introduced.** Python, R, Jupyterhub, Jupyter Notebook, Bioconductor, limma, Microarray data, heatmaps, differentially expressed genes, BioGPS, GENT database, GEO database, Gene Ontology annotation methods.

**Assessment.** Submit a pdf of your **completed JupyterNotebook** together with **the completed report card** at to be found at the end of this handbook by the end of this practical. You can work on your reports collaboratively, but need to individually upload the report onto Canvas by today 5pm (ideally you would finish and submit by the end of the practical today).

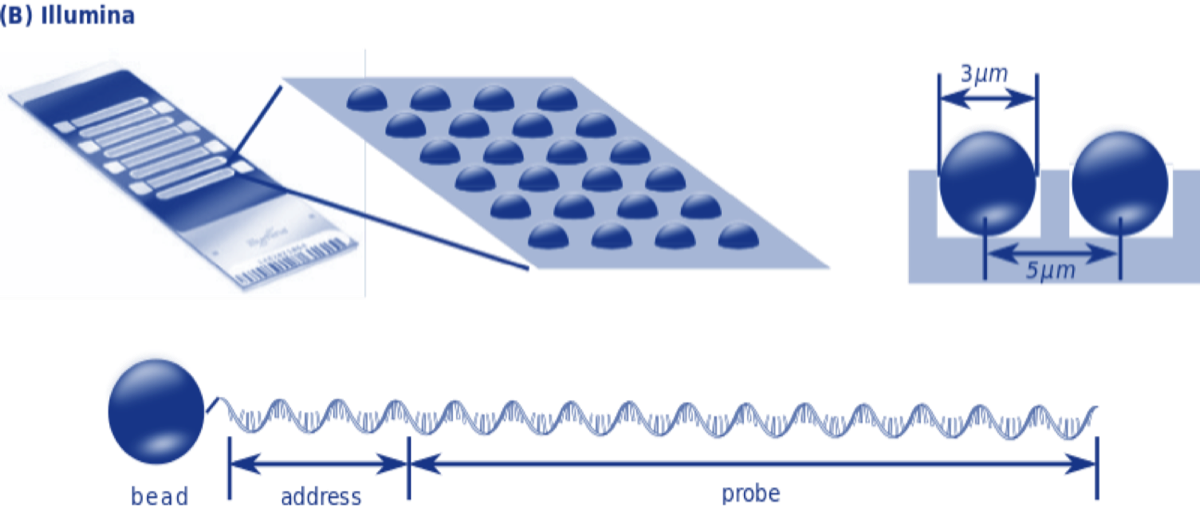
**Practical structure.**

**Morning:** 09:30-11:45 Complete JupyterNotebook and database research

**Lunch break** 11:45-13:00

**Afternoon:** 13:00-15:45 Interpret the results, write and submit your report.

**Introduction.** The previous practicals have introduced you to some of the “bread and butter” techniques that are used everyday in molecular laboratories, such as pipetting micro-volumes, DNA extraction from tissue, Polymerase chain reaction (generating Amplicons), DNA-barcoding, Agarose-Gel Electrophoresis and Sanger sequencing. You then learnt the basics of command line-based genetic data analysis and that linux/shell/bash, python and R are some programming languages that can be used to analyze data remotely on High Performance computers such as VIPER@hull. This was demonstrated on the example of generating your own blast search engine, and by building a phylogenetic tree out of Amplicon sequencing data. The data sets you practiced this on were comparably small, but today genetic data is often generated on the level of the genome which translates into many gigabytes of data per experiment. This can include the entire genome such as in whole genome sequencing WGS, or when the genome in its entirety is scanned for difference in epigenetic status such as Cytosine-Methylation (called whole genome bisulfite sequencing WGBS). Other popular methods are only sequencing parts of the genome - and a commonly used example for this is the microchip array technology. In this technique, a silica base is spotted with small spots or bead to which oligonucleotide probes are covalently linked (see figure). These probes exist for all genes of the human genome and separate chips are produced for miRNA genes or other selected genomic regions.



**Figure 1.** Illumina Bead-Chip Microarray containing oligonucleotide probes for all human genes.

In this practical, we will use Jupyter notebook to analyze such a microarray experiment using the analysis package **limma within the bioconductor platform in R**). The data was generated from four patients, and each of those patients provided a sample of tumour tissue as well as adjacent healthy tissue. The research question in this experiment was, which genes are differentially expressed between the different types of tissues so that it might direct the research into finding better treatments. Additionally, we have chosen to remove the information on which exact type of cancer was studied, to prompt you to use publicly available information to find out the answer to this question yourselves.

Most of the sequencing data generated is deposited online in the public domain. Because of that, there was in recent years an accelerated development of tools that allow users to analyze such data online as well. It is useful to know a few of such resources and tools, but it is also important to know that such resources are ever-changing and evolving.

**Procedure**

**Part 1: Jupyter Notebook**

Navigate to github.com and search for the user “cybokat”

Navigate to the github repository “gapractical” and there copy the URL via the “clone or download” button

Navigate in another browser window to jupyterhub.hull.ac.uk and log in. Please remove any data left over from the last practicals from your working directory (select the files and then click on the red bin icon)

In jupyterhub add a new terminal window (under New:Terminal)

In the terminal window,enter

>clone git URL\_to\_github\_repository

This should replicate a set of files in your directory, including these:

Gapractical\_students.ipynb

TumorAdjacent-Controls.txt

TumorAdjacent-probe-raw.txt

Asurgen\_desc.txt

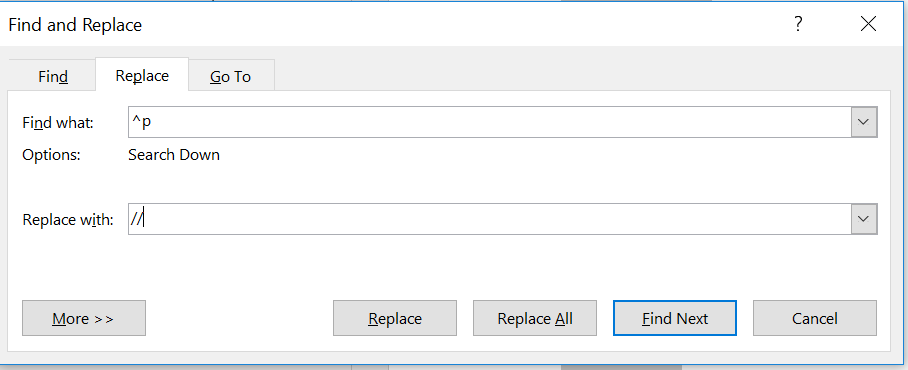
The GApractical\_students.ipynb is your JupyterNotebook for today. All code is pre-filled into the cells, so if you accidentally delete the content of a code cell, you need to type the code back in manually using the provided copy of this handbook. The first cell prompts you to set your working directory. This is likely “/home/h1/Your\_User\_ID/” so you will need to modify that line to match the correct specifications.

Read about the analysis, and **run each one cell at a time until you reach the last cell.** Because the data we are working with is much larger, it may take a while for the output to appear - don’t get impatient. The last code cell is left blank - use the instructions to modify the figure (heatmap) until it is noticeably different from the first figure generated with the provided code. Download the generated pdf (it will be in your repository) and add it to your report card as a figure.

**Part 2. Databases**

We will now analyze more in detail a subset (25) of the differentially expressed genes. Copy 25 randomly selected over/underexpressed genes from the heatmap showing the top 50 into a text-

-format. Use MSWord to make a list in which each gene symbol is separated from the next through a paragraph symbol.



It should look like this

HSP90

MAPK1

CALCR

HSBP2

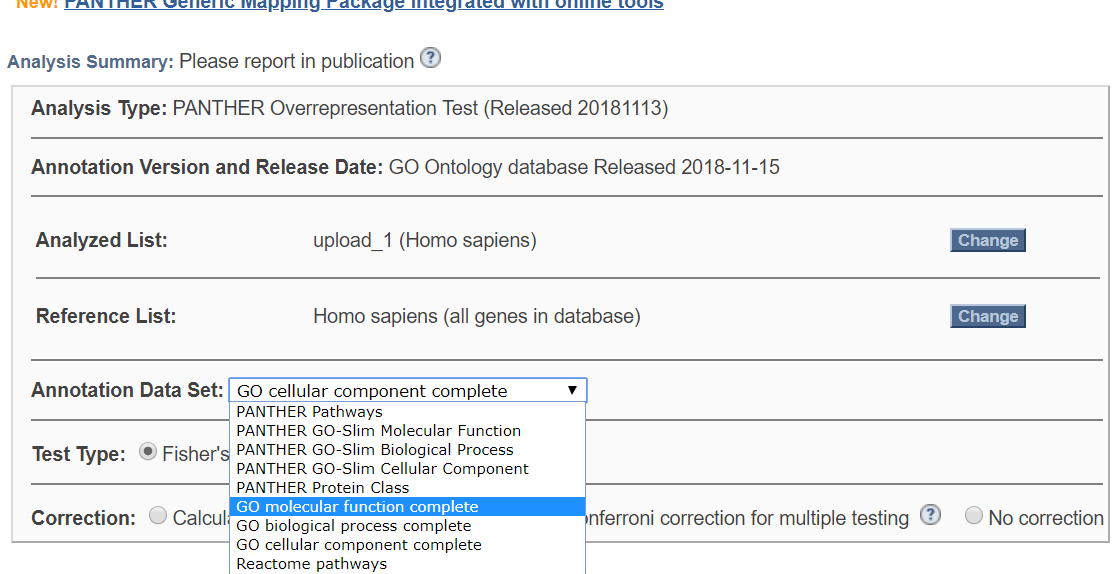
…

Such a list of gene names can be read by most online databases and tools. A few will require you to separate gene names by comma or space, but most will understand the list format. To transform the list format into a // separated list (or any other separator), you can just use Word’s find/replace command:

HSP90//MAPK1//CALCR//HSPB2

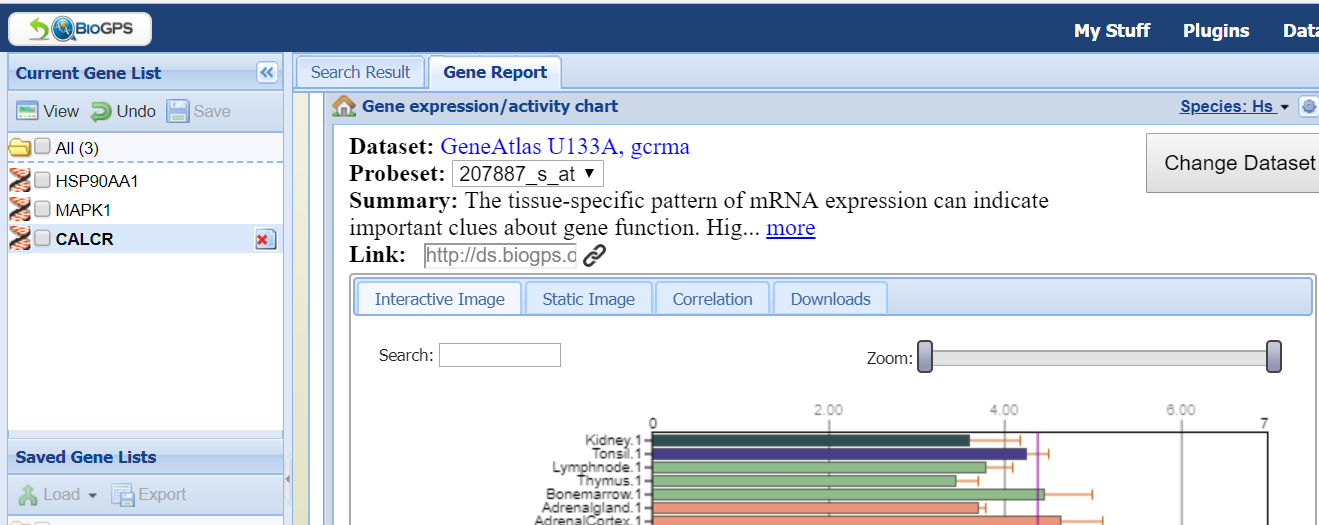
**2.1 Enrichment analysis**

To find out whether these genes have any biological functions in common, we can take advantage of systems biology knowledge and navigate in a browser to geneontology.org. Take some time to explore the website. On the landing page, then paste the list of your genes into the “enrichment analysis” window and submit. Discuss in the report which Gene Ontologies (GOs) are significantly overrepresented with their biological processes in the list of genes (“what they have in common”). Modify the search criteria (image below) to also check for overrepresented GO molecular functions and cellular components.



**2.2 Normal expression of these genes.**

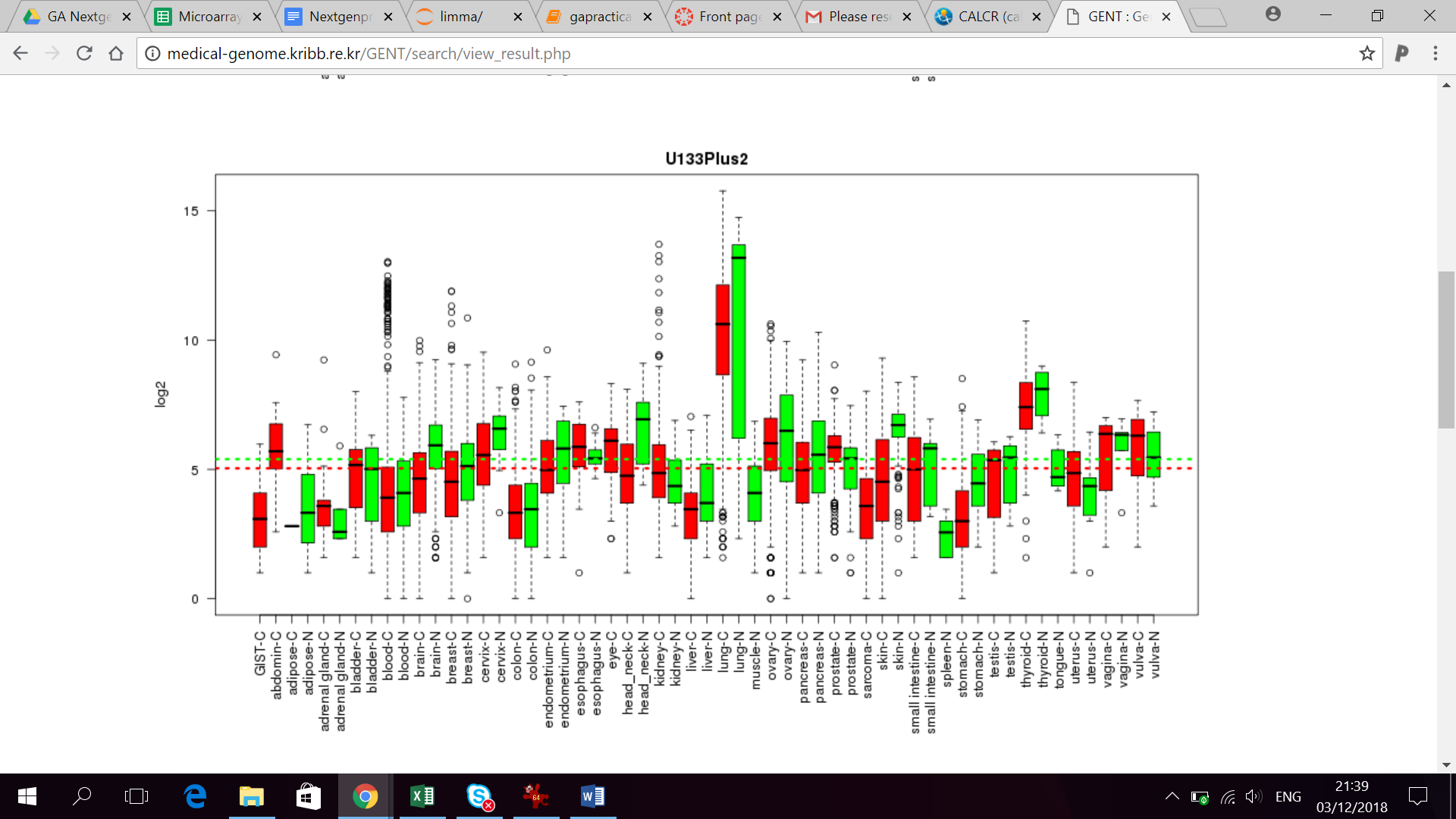
Have a more detailed look at some of these genes: which tissue are they normally expressed in? For this navigate to biogps.org where you can enter a list of genes, and then click through their names in the “current gene lists” tab to open box plots of their expression in normal tissues. Note down where these genes are highly expressed in normal circumstances.



**2.3. Abnormal expression of these genes.**

Cancer is mostly a disease of altered gene regulation -- The present dataset is only one of many that detected this. For comparison with other studies you can navigate to GENT database (Gene Expression across Normal and Tumor tissue) at <http://medical-genome.kribb.re.kr/GENT/> and there to the “search” tab. You can enter genes, one at a time, into the top box. It will open a boxplot that shows comparable expression values from other cancer microarrays as well as their log2 values. Which type of cancer © versus normal(N) ratio is noteworthy for each gene?

!!!! Mind the direction of the expression fold change: Is it up/down regulated in cancer in our data? Is it up/downregulated in the database?!!!!



**Student Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Student ID\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Report card for Bioinformatics practical on 3/12/2018**

**Hand in together (upload on canvas) with pdf of completed JupyterNotebook**

**Analysis of next-generation genetic “-omics” data (target 2 pages)**

Remove the writing prompts in blue font once you’ve answered them

**Problem:**

Discuss how the analysis of microarray data can be useful to study cancer

Discuss the challenges of large genetic data sets

**Analysis**:

How does the Illumina Beadchip Microarray work? Describe the procedure. Which type of data is generated?

What could be the problem if the four different patients had very different gene expression patterns?

How do we go about finding differentially expressed genes?

The test underlying the eBayes procedure tests for differences in the expression of each gene between the four tumour and four normal tissues (similar to a T-test). Why did we perform the FDR procedure?

**Results.**

Describe what the volcano plot shows. What is fold change?

What is the information contained within a heatmap?

Which GO functions, processes and cellular components do these genes represent? Is there a difference between up-and downregulated genes?

Where are they normally expressed, and which tissue shows a difference in cancer?

**Discussion.**

Which type of tissue or type of cancer was most likely studied in this experiment, based on your results?

For which other scientific question could you imagine using the analysis that we just performed?