##
######## R code for Illumina beadchips data analysis using lumi and limma ########
##

####### DIFFERENTIAL EXPRESSION BETWEEN TWO GROUPS OF SAMPLES FOLLOWING AN UNPAIRED DESIGN ######

##
################# HELP on FORMAT #################
##

METADATA

Metadata is a table containing two columns the sample names (sampleName) and array barcodes (ArrayBarcode). The metadata can be done by creating an excel spreadsheet that is saved as a tab delimited file (.txt) and it should contain the information regarding which sample belongs to which group and the chip number for each sample.
To be able to run the code, it is important to follow the format described in this section. The headers (sampleName, ArrayBarcode) should be identical.
ArrayBarcode - combining chip IDs and 12 capital letters (A-L).
Each of the sample names is a combination of a sample ID (eg patient ID) and a marker (group) separated by an underscore (_).

sampleName	ArrayBarcode	variableX
65055_T	9683758015_A	1
61634_C	9683758015_B	1
61210_T	9683758015_C	1
61634_T	9683758015_D	1
61420_T	9683758015_E	1
48610_C	9683758015_F	1
65043_T	9683758015_G	1

ANNOTATION FILE
This file is ususally provided by the sequencing facility or can be obtained from the Illumina website (http://support.illumina.com/array/downloads.html)
Here below is the top of the manifest file called HumanHT-12_V4_0_R2_15002873_B.txt

? Illumina, Inc.
[Heading]

Date	15/4/2010
ContentVersion	4
FormatVersion	1.0.0
Number of Probes	47323
Number of Controls	887

[Probes]

Species	Source	Search_Key	Transcript	ILMN_Gene	Source_Refe	Accession	Symbol	Protein_Proc	Probe_Id	Array_Addre	Probe_Type	Probe_Start	Probe_Sequ	Chromosom	Probe_Chr_C	Probe_Coorc	Cytoband	Definition	Ontology_Co	Ontology_Pr
ILMN Controls	ILMN_Contr	ERCC-00162	ILMN_33373	ERCC-00162	ERCC-00162	DQ516750	ERCC-00162		ILMN_31666	5270161	S	12	CCCATGTGTCCAATTCTGAATATCTTTCCAGCTAAGTGCTTCTGCCCACC	Methanocaldococcus jannaschii spike-ir						
ILMN Controls	ILMN_Contr	ERCC-00071	ILMN_33364	ERCC-00071	ERCC-00071	DQ883654	ERCC-00071		ILMN_31655	4230037	S	563	ACAGAGTTAAGACTTAGATCAGCGAGCAGGTGTACGCCCCGGACCTTC	Synthetic construct clone NISTag13 exte						
ILMN Controls	ILMN_Contr	ERCC-00009	ILMN_33358	ERCC-00009	ERCC-00009	DQ668364	ERCC-00009		ILMN_31648	60372	S	889	GACACGCGCTTGACACGACTGAATCCAGCTTAAGAGCCCTGCAACGCC	Synthetic construct clone TagJ microarr						
ILMN Controls	ILMN_Contr	ERCC-00053	ILMN_33362	ERCC-00053	ERCC-00053	DQ516785	ERCC-00053		ILMN_31653	5260356	S	873	CTGCAATGCCATTAACAACCTTAGCACGGTATTTCCAGTAGCTGGTGAC	Methanocaldococcus jannaschii spike-ir						
ILMN Controls	ILMN_Contr	ERCC-00144	ILMN_33371	ERCC-00144	ERCC-00144	DQ854995	ERCC-00144		ILMN_31665	6060692	S	398	GCTCGTCACCAACTCGTCACGCGATCGAAATAGCTTGGACTAATGTCCC	Synthetic construct clone AG006.1100 e						
ILMN Controls	ILMN_Contr	ERCC-00003	ILMN_33357	ERCC-00003	ERCC-00003	DQ516784	ERCC-00003		ILMN_31647	6370471	S	901	ATTGAAAGTTTGGGAGGGACTATTCACAGTATAGATGAGGTTGTTGCA	Methanocaldococcus jannaschii spike-ir						
ILMN Controls	ILMN_Contr	ERCC-00138	ILMN_33371	ERCC-00138	ERCC-00138	DQ516777	ERCC-00138		ILMN_31664	1710435	S	768	CCACAGCATCCCAGTCGTGAATTAAGTATAAAGCAACTCCACCAATGTT	Methanocaldococcus jannaschii spike-ir						

ILLUMINA BEADCHIPS RAW DATA

This file is provided by the sequencing facility.

It contains the following columns: TargetID, ProbeID, AVG_Signal, BEAD_STDERR, Avg_NBEADS, and Detection Pval. The total number of columns is 2 + 4 * Nsamples.

The code will have to be adapted if another file is used.

[Header]

GSGX Version	1.9.0
Report Date	6/10/2014 2:01:32 PM
Project	
Group Set	
Analysis	_nonorm_nobkgd
Normalization	none

[Sample Probe Profile]

TargetID	ProbeID	9683758015_A.AVG_Signal	9683758015_A.BEAD_STDERR	9683758015_A.Avg_NBEADS	9683758015_A.Detection Pval
7A5	6450255	120.8	4.596	14	0.05065
A1BG	2570615	2070.3	112.449	22	0
A1BG	6370619	102.9	3.512	16	0.67662
A1CF	2600039	108.8	3.052	26	0.3039
A1CF	2650615	111.8	3.828	15	0.17662
A1CF	5340672	104.2	3.514	21	0.58571
A26C3	2000519	100.4	3.866	16	0.82208

##
##

This code is covering the following topics:

1 install R packages

2. Pre-processing using lumi package, quality control plots and clustering.

3. Statistical testing of differential expressions using limma package. The code is designed to calculate differential expression (unpaired design) between 2 groups of samples defined by the user in the metadata as in our metadata example, treated T versus contrl C samples.

4. Create a heatmap of the top 500 most differentially expressed genes, generate an image of the heatmap as well as an excel table.

5. Prepare the rank fike and the expression file to run GSEA in a pre-ranked mode.

##
##

Tips on how to run the code:

create your metadata file in excel (save as tab .txt)
on your computer create a folder containing the 3 files that you need and where the output files from the code will be saved. Don't change the name of this analysis folder.
open R
do a ls() to make sure that you are working on a clean workspace
assign the variables as indicated at the top of the code script : indicate the path to your analysis folder in workDirectory, the name of your metadata file name in metadataFileName , name of the Illumina raw data in IlluminaDataFileName and annotation file name in annotationFileName
run each line of the code carefully , stop if you see an error message, fix the issue and rerun the line
if this is the first time you run the script, you need to install packages (in the section R packages and work directory) and load them into the workspace using the library() function , if you already installed the package, you just need to load them.
Once the script is completed, save your script (.R) as well as the R workspace (.RData) by writing save.image("myanalysis.RData").
Lines that start with an hash character # are comment lines, you don't need to run these lines.

##
##################### R packages and work directory ###########################
##

To install the packages, uncomment the source line (source("http://bioconductor.org/biocLite.R")) as well as the lines corresponding to the package(s) (librarires) you want to install and run these lines (e.g. biocLite("lumi")). then, load the libraries.

The work directory corresponds to the path to the analysis folder containing your raw data, metadata and annotation file. Setting the working directory will tell R where to find these files on your computer. The R script will create in this folder a subfolder called temp that will contain the output files generated by this script.

Use getwd() to see if you set the directory correctly and dir() to see the files in the analysis folder and if temp has been created.

```
################################################################################
########## METADATA: SAMPLE NAME AND ARRAY BARCODE ######################
################################################################################
```

This section of the code takes the metadata table and create 3 additional columns that will be called ‘patient’, ‘marker’ and ‘chip’. The patient and marker column are made by splitting the sampleName column. These columns will be used to create the design matrix in limma and will be used to define the comparisons to make (eg treated vs control or group2 vs group3).

Running the command “sample_barcode” (last line of this section) will output the new metadata table and you can check if it is correct.

```
################################################################################
#####  PREPROCESSING USING LUMI PACKAGE ##################################
################################################################################
```

This section reads the data into a lumi object (lumiR) and perform normalization (lumiEpxresso). The names of the raw data will now be changed for more meaningful names taken from the column sampleName of the metadata file. It will be for example useful to see this more informative names in a hierarchical clustering plot later.

```
################################################################################
##################  QUALITY CONTROL PLOTS  ###############################
################################################################################
```

##boxplot and density plot of both raw and normalized intensities on log2 scale
The density plot shows the distribution shape for each sample (each beadchip). Before normalization, the distributions are not identical and many lines are visible. After the normalization process, the lines should be all identical and superimposed. A boxplot divides each sample data into quartiles. The first and fourth quartiles are represented by dotted lines (whiskers) and the second and third as a rectangle box separated by the median.
Before normalization, the median of all samples may not be aligned whereas it should be aligned after the normalization process.
The lines starting with pdf are there to save the plot image in a pdf format and the png will save the image as a png format. If you like the png format, uncomment the corresponding line

of code. The files will be saved when the dev.off() line is run. They will be saved in the temp directory.

##Pairs plot of sample intensities in an ExpressionSet object
The dots that are close to the red lines indicate probes with similar intensities between the 2 pairs of samples. The dots away from the diagonal represent probes showing differential expression.

##MA plots
MA plots can be used to visualize differentially expressed genes
M versus A plot for each array can be drawn with M and A defined as :
M = intensity ratio between 2 samples
A = average intensity of the 2 samples
The MA-plots have loess lines in red and the M = 0 horizontal axis in blue.
Situations where an array has a clearly aberrant loess line on these MA-plots often are indicative of potential quality problems.The median and IQR values appearing on each plot relate to the center and vertical spread of the M values. You can see a classical result with a skewed MAplot before normalization (A) towards low intensity points given to this graphics a banana-like shape.After normalization, using the "lowess whitin-tip" option the graphics becomes more symmetric but there are still some points with very high ratio at low intensity levels that are possibly future false positives.

##
################## CLUSTERING ##################################
##

This section of the code presents different plots that visualize the relationships between the samples. It is based on all probes. If samples are close together, they will be linked by a common branch on the dendrogram; on the PCA plot (2D or 3D), the dots will be close to each other.
These quality control plots can help for detection of outliers, detection of possible mismatch between samples, to see if groups cluster as expected (e.g, clean data have a good separation between the control and treated group; noisy data do not show clear separation). These plots could also indicate the need for differential expression analysis performed as a paired design (e.g.pair of patients cluster together regardless of treatment).
Looking at genes in additional principal components (e.g PC3 , PC4) of the PCA could also in some cases help identify a noise signal coming from a mix of populations.

##
################## PRESENT COUNTS ##################################

##

Present counts use the detection pvalues that are included in the raw data (in the lumi object) for each probe in each sample. It helps remove probes which correspond to genes that are likely not expressed in the tissue type under study. Removing these probes is usually recommended as they are adding noise to the data as well as adding more statistical tests to perform, increasing therefore the magnitude of correction for multiple hypothesis testing.
The first step is to set a threshold (detection.p.th); a probe is present if its detection pvalue is less than or equal to the threshold. The probes that are absent in each sample should be removed of the differential expression analysis (presentCount).
There are several ways to define a threshold using an arbitrary value (e.g 0.1) or looking a the distribution of the samples and remove the probes that have a pvalue that falls into the upper quartile for 50% (or as defined by users) of the samples. In this code, the detection threshold will be set by estimating a FDR (false discovery rate) value.

##
############ DIFFERENTIAL EXPRESSION (DE) USING LIMMA ################
##

The aim of the differential expression analysis is to test if a probe representing a gene has a mean intensity that is statistically different between the 2 groups under comparison (e.g treated and control). As the distribution of the log2 normalized intensities values are following a normal distribution, a t-test can be applied. A t-test is applied for each probe that were called present. The result of t-test is a t value. A t value is the ratio of the difference between the mean of the 2 groups (e.g. the logFC) divided by the standard error which is a measure of the variability of the intensity of the probe within the samples of] 2 groups. A t value will be high (positive or negative) if the difference in intensities between the 2 groups is high and the variance is small. Limma used a moderated t-test meant to correct for the high effect that the variance has on the t value. Biological data are variable by nature or the variance can be attributed to manipulation of the samples. pvalues are calculated from the t values (using the degree of freedom and magnitude of the t values) and pvalues are corrected for multiple correction testing (false discovery rate FDR, adjusted or corrected pvalue).

You can run this part of the code without running the quality control plots and clustering sections but you need to have run the preprocessing with lumi part.

##############################
Annotation file
##############################

This part of the code adds the annotation details for each probe using the Illumina annotation file to the normalized data table. It takes the manufacturer's version available from the Illumina website. At the end of this section, the annotated normalized data are copied in the temp folder on your computer. These normalized data are needed for GEO (Gene Expression Omnibus) submission for example.

```
#####################################################################
###### Calculating differential expression (DE) between markers ############
#####################################################################

##################
# I. Data filtering ###
##################
```

The data are first filtered to remove probes that are called absent in the samples included in the 2 comparison groups and all tables are reduced to the included samples.

```
###################
# II. Model fittings  ###
###################
```

Differential expression testing in limma consists of 3 steps :

1) lmFit: estimate the coefficient (b), standard deviation (u), and residual variance (s^2) for each gene. Then the usual t-statistic for this gene is calculated as b/(u*s),which follows t-distribution with a degree of freedom (d) . But for small sample size, the variance estiamte s^2 will be larger, and then reduces the test power. So a called moderated t-test, i.e, the next step, is further used in limma. If this is paired design, the information about which samples are paired will be included in the model design. This is also where batch effect (using the chip information) or other confounding variables (column variableX of metadata) will be used here to correct for these effects.
2) contrast fitting: What is makeContrasts doing? (CHANGJIANG: can you please add a sentence here)
3) ebay: empirical Bayes smoothing is applied to the standard errors (or residual variances), as Shaheena mentioned. A posterior value for the residual variance is calculated, denoted as sp2. Then the moderated t-statistic is b/(u*sp), which follows t-distribution with degree of freedom d0+d. The extra degree of freedom, d0, represents the extra information which is borrowed from the ensemble of genes. The value of the moderated variance, sp2, is between s2 and a prior value s02.

The output value of this part will be for each gene a t value , a pvalue and an adjusted pvalue (FDR). The output is created by the topTable() function.

Note: the limma report names the coefficient b as the " logFC" column; we also add a column called logFC2 which follows the definition mean group1(log2) - mean group2(log2).

##
Updating annotation using the biomaRt package
##

As the manufacturer's annotation files are created once a chip becomes commercially available and not updated afterwards, this part of the code use biomart to retrieve most recent probe annotation using the latest genome annotation from Ensembl.
If a probe corresponds to multiple genes, the corresponding genes are all listed for this specific probe and each gene are separated by a slash character (geneA/ GeneB).

####################
save the result table in the temp folder on the local computer
####################

The output table that contains the limma output results, the original logFC, the original and updated annotation as well as the normalized data are saved on the local computer. The results are ordered by default using the best FDR.

##
hierarchical clustering on the 2 groups
##

This part of the code creates a hierarchical clustering that display the similarities only between the samples included in the 2 groups comparison.

#############
STARTING FROM THIS POINT, TABLES ARE REDUCED AT THE GENE LEVEL BY SELECTION OF ONE PROBE CORRESPONDING TO THE BEST T VALUE
############

################
Volcano plots
################

A volcano plot enables to quickly identify genes with most differential expression. It compares the logFC with the FDR. The logFC represents only the amplitude of the signal, ie the differences between the intensity average of the 2 groups whereas the FDR calculated from the t values takes also the heterogeneity or variability within the groups into account.

```
################################################
## Heatmaps of the top 500 and top 50 genes
################################################
```

This part of the code represents a colored image (heatmap) with a clustering on samples and genes for the top 500 and top 50 most differentially expressed genes. It also produces a .csv file (comma separated value format) that can be opened in Excel and contains the genes in the same order of genes and columns as the top 500 heatmap. The last columns of these files contain the scaled (Z score) values.

```
#########################################################
## Preparing rank file and expression file for GSEA #
#########################################################
```

This code prepared a rank file (.RNK) that can be used to rank GSEA using the pre-ranked mode. It also created the expression file needed for the creation of an enrichment map.