**FOS Computational Biology of Complex disease and Ageing**

**Practical on Epigenomics, June 22nd 2017**

**Bas Heijmans and Roderick Slieker**

**Important:**

* **Before performing this practical, make sure that the regional settings on are right: decimals separator a point, thousands separator a comma.**
* **At the end of the practical send your work to** [**r.c.slieker@lumc.nl**](mailto:r.c.slieker@lumc.nl) **as Word document with your name(s) on top of the first page. The file name should contain your surname(s).**

­A primary function of DNA methylation, a key component of the epigenome, is to control cell differentiation. In this practical, you will identify and annotate differential methylation between tissues using genome-wide data generated with Illumina 450k chips. The coming few hours you will replicate many of our group’s findings which were published in 2013. Good luck and have fun!

Please find the syntax in the R script (*Script FOS Course 2017.R*) provided.

**Question 1**

The data that you will use in this practical consists of genome-wide DNA methylation data (Illumina 450k methylation array) of tissues from autopsy samples. Load the file ‘*Tissue data subset.csv’* in R. The file contains the DNA methylation data of a single CpG site with identifier number *cg21507095*.

**A** Which tissues were studied? And on how many cases?

**B** Make a boxplot of DNA methylation in the different tissues*.* Also,calculate the mean methylation and standard deviation for the tissues. Do you think that these differences are biologically relevant?

**C** Test whether the difference is statistically significant using ANOVA. Is the difference between tissues statistically significant?

**D** Now, use the UCSC Genome Browser to annotate the CpG site. Go to <http://genome.ucsc.edu/>, click Genomes (Assembly: Feb 2009 (GRCh37/hg19)) and look up the position of the CpG site (type the CpG identifier in the *search term* box). Then scroll down to *Regulation*, find *CpG islands*, set this feature to ‘*show*’ and click *refresh* on the right of the screen. Describe in your own words what CpG islands do and how they are defined (🡪 right click on the CpG island track, *Show details*).

**E** Go back to the genome browser. Scroll up and use ‘zoom out’ to find out: (1) in which gene the CpG is located, (2) what part of the gene it is located (intron, exon, etc.) and (3) whether it is located in a CpG island.

**Question 2**

You will investigate DNA methylation between tissues across a complete chromosome instead of a single CpG site. Load the file *Tissue data chromosome 22.csv* with DNA methylation data for chromosome 22 across three tissues.

**A** How many CpGs were measured on chromosome 22?

**B** Make a histogram for DNA methylation data for blood. Inspect the histogram and describe the characteristics of DNA methylation in the human genome (note that the histogram will look virtually identical for other chromosomes and tissues).

**C** The genome can be categorized in different genomic features. Gene-centric annotations divide the genome relative to gene structure/function. One can also annotate CpGs based on the CpG density, that is CpG islands and their surroundings. Use the annotation figure below to re-answer **1E**.

**I:\medewerkers\Roderick\Genomic features.tif**

**D** The file *Tissue data chromosome 22.csv* contains columns with annotation information. Make boxplots for the gene centric annotation for blood. Describe the differences in DNA methylation across annotations and explain whether this is according to your expectation given your knowledge of the function of DNA methylation.

**E** Do the same for theCpG island annotation. Where is DNA methylation lowest?

**F** In question Q1C we calculated the P-value for one CpG site. Now we want to run this for each of the CpGs sites across chromosome 22. Instead of saving data as csv, one can also save it as *RData* files. Load the RData file *TissueData chr22.RData.* This file contains two files, namely *samplesheet* and *TissueDataChr22.* Look at both files. What are the dimensions of both files?

**G** In R we can use functions to easily run a model across many loci. Run the function for the first and the second line.

**H** Now run the model across all CpGs on chromosome 22. Note that this takes a bit of time – after all R is running 8513 tests – and note that this couldn’t be achieved using SPSS (only when the tests were done one by one). Look at the output.

**Question 3**

**A** When measuring many features (here CpGs), one must account for the fact that many statistical tests are being performed. Small *P*-values will be observed purely by chance. In this case, the number of tests equals the number of CpGs in the file. Which *P*-value threshold is commonly used to declare an association statistically significant in case of a single statistical test? How many *P*-values lower than this threshold would you expect by chance in this chromosome 22 data set?

**B** An intuitive and popular (but perhaps somewhat conservative) way to correct the p-value is to multiply the calculated *P*-value by the number of tests performed or correct the significance level, i.e. divide the *P-*value threshold for a single test by the total number of tests. This is called *Bonferroni* correction. What is the Bonferroni-corrected *P*-value for chromosome 22 (i.e. the threshold for statistical significance after accounting for multiple testing)?

**C** For question **1C**, you calculated a *P-*value for the example CpG site. Compare this *P-*value to the *P*-value calculated in **3B** (after all it's only one of many CpG sites measured on chromosome 22). Is the difference significant after multiple testing correction?

**D** The column Pval\_correctedcontains 0 and 1’s that mark non-significant and significant CpGs based on the *P*-value calculated in **3B**. How manyCpGs (absolute and as percentage) are significant after multiple testing?

**E** Now, we want to know whether a particular genomic feature is overrepresented in the identified significant CpG sites. You will focus on the two best covered genic annotations (proximal promoters and gene bodies). The significant CpGs in these two features can be found in the file *Tissue data chromosome 22 significant.csv*. Load the file. Make a frequency table of the number of CpGs per group and split by CpG islands and non-CpG islands*.*

**F** Now you have the frequency per combined feature (so for example *proximal promoter CpG islands*). Calculate the percentages of the number of significant CpGs per feature compared to the total number of significant CpGs from **3D**. What is your interpretation?

**Question 4**

Now, you will focus on ‘top*-*hits’(=CpG sites with lowest *P-*values). Use the file *Tissue data chromosome 22.csv.*

**A** Sort the file by *P-*value (ascending). Select the two CpG sites with the lowest and second lowest *P-*value. What is the methylation level of these CpGs in the tissues investigated? What is the largest difference in DNA methylation observed?

**B** Goto the UCSC genome browser at <http://genome.ucsc.edu/>. Click on *My Data* at the top of the page and then *Custom Tracks. Upload* the *significant probe.bed* as a custom track (click *Choose File* and then *Submit*; like before, use assembly Feb 2009 (GRCh37/hg19)). Click *go to genome browser*. For clarity, display the track as ‘dense’. A red dot denotes a differentially methylated CpG (accounting for multiple testing), a blue dot a CpG for which there is no statistical evidence for differential methylation.

**C** Look up to the two top-hits (See **1E** how to if you don’t know how). In which gene centric / CGI centric feature are the top-hits? Also see/use question **2C.**

**D** Zoom out in UCSC for both top-hits. Is the differential methylation limited to a single CpG site or does it extend across a region(Use the custom track from **4B**)? Does this influence your interest in the two CpGs and why?

**E** The second lowest top-hit maps to the *PARVG* gene. How many different transcripts can arise from this gene (track UCSC genes)? Do you observe differential methylation near alternative transcript start sites for the gene?

**F** Look up the nearest gene of the tophits in GeneCards ([www.genecards.org](http://www.genecards.org)). Look at the protein expression and the mRNA expression for the three tissues studied if available. Is the gene expressed in a tissue-specific fashion?

**G** To gain insight into gene function related to differentially methylation you can use GREAT (Genomic Regions Enrichment of Annotations Tool). Go to the GREAT website (<http://bejerano.stanford.edu/great/public/html/index.php>). Select human hg19/ GRCh37. Select all significant CpG sites from the SPSS file and copy the chromosome, start and end column and paste that into test regions (after clicking BED data). Also, copy **all** CpG sites and paste that into background regions (click BED data). Click Show Settings and **select single nearest gene**. Click submit and wait (may take some time) until the site has calculated the enrichment in gene categories (Biological processes). Which biological processes are enriched for differential methylation? How does that relate to the tissues studied?

**Facultative question: Question 5**

There is increasing evidence that DNA methylation is associated with alternative transcription events. A paper showed that there is a role for intragenic (differential) DNA methylation in the regulation of alternative transcription events (Maunakea AK et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 2010; 466: 253–7).

**A** Read the abstract and inspect Figure 3a in the paper. In what process is DNA methylation implicated to play a role in? For which gene and which annotation within that gene do the authors provide experimental evidence for this role? Discuss the tracks given in Figure 3a and their implication.

**B** Look up this gene in UCSC for you own data. Display the custom track you used before and CpG island track. Argue why the data on the three tissues is either in line or at odds with the observations of Maunakea *et al.*