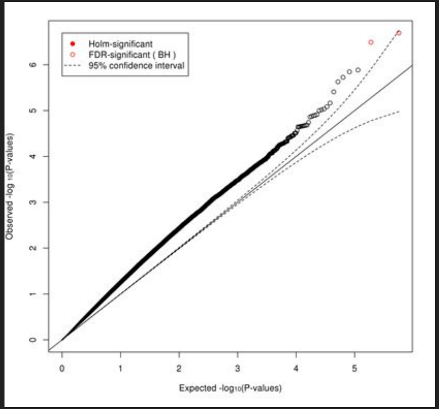
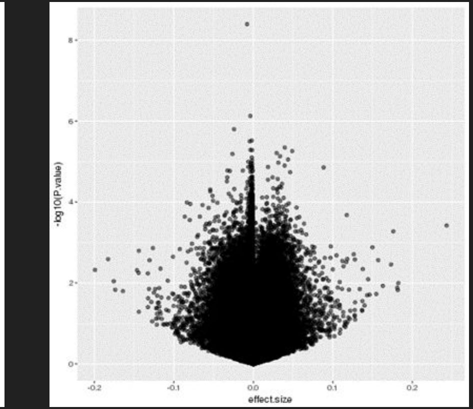
Introduction to Bioinformatics

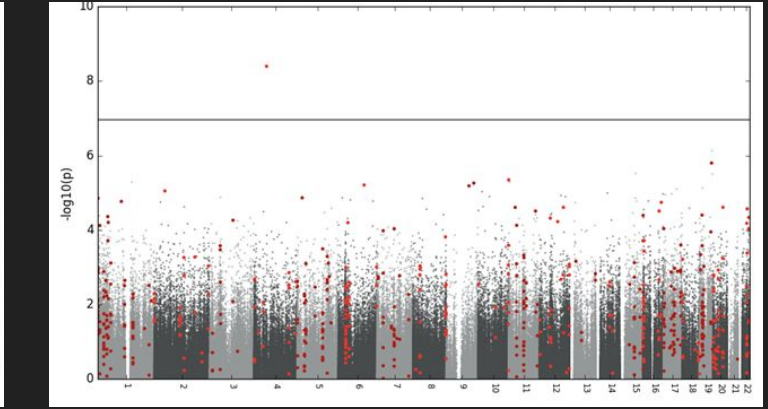
More on Methylation Analysis (Interpreting and Replicating plots from publications in R)

Introduction

- When reading paper on methylation studies (DMR, EWAS) you may encounter the following plots



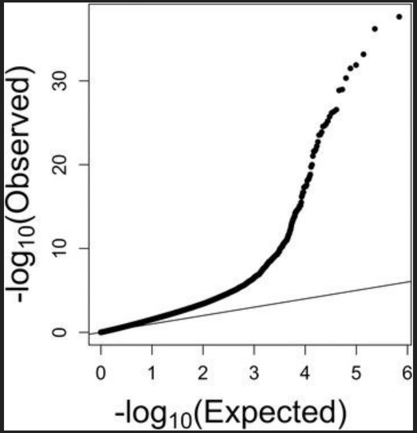


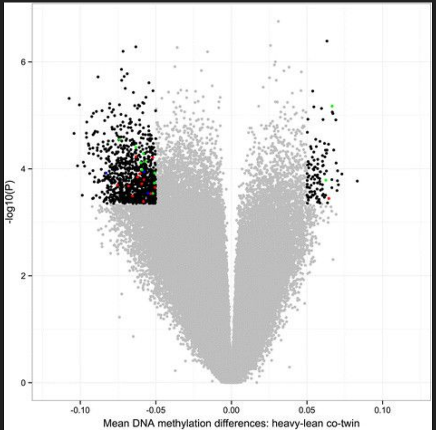


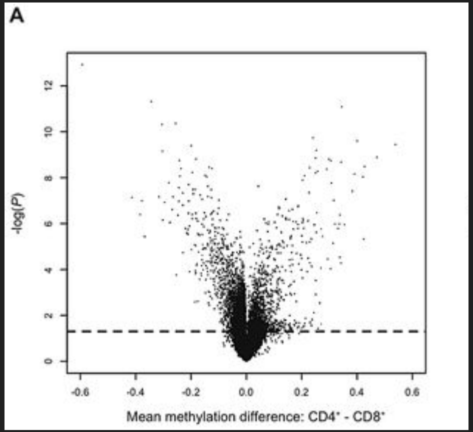
Richmond et al. 2018. The long-term impact of folic acid in pregnancy on offspring DNA methylation: follow-up of the Aberdeen Folic Acid Supplementation Trial (AFAST)

Introduction

- Today we will learn on how to interpret those and how to generate those using the data that we have and using R



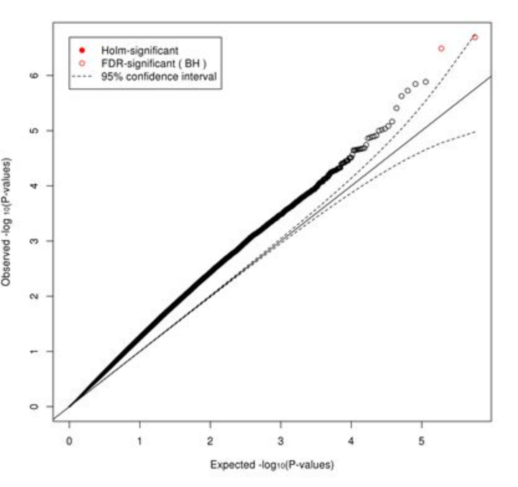




Gervin et al. 2012. DNA Methylation and Gene Expression Changes in Monozygotic Twins Discordant for Psoriasis: Identification of Epigenetically Dysregulated Genes

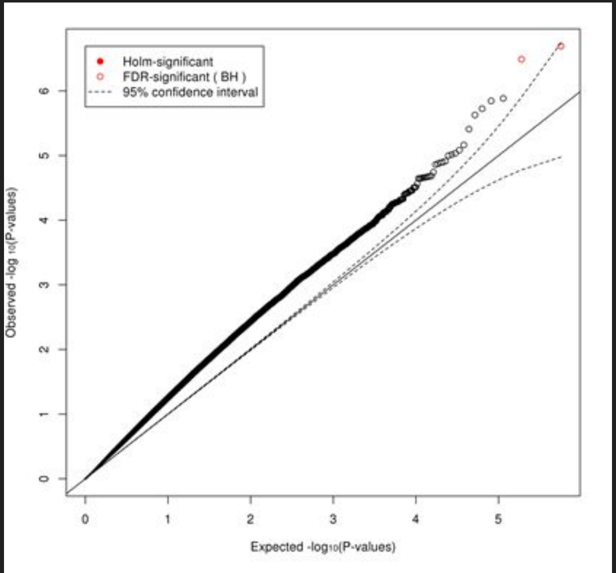
qqplot

- quantile - quantile plot,
- In general it’s used to assess whether our data is coming from some theoretical distribution
- Visual check, not an air-tight proof -> somewhat subjective
- For methylation study we’re using it see the evidence of DNA methylation differences between our sample



Richmond et al. 2018. The long-term impact of folic acid in pregnancy on offspring DNA methylation: follow-up of the Aberdeen Folic Acid Supplementation Trial (AFAST)

First glance on qqplot



- basically a scatter plot
- y axis represents the observed -log10(p-values)
- x axis represents the expected -log10(p-values)
- y axis is from our results
- X axis is generated theoretically
- Small circles are our instances??

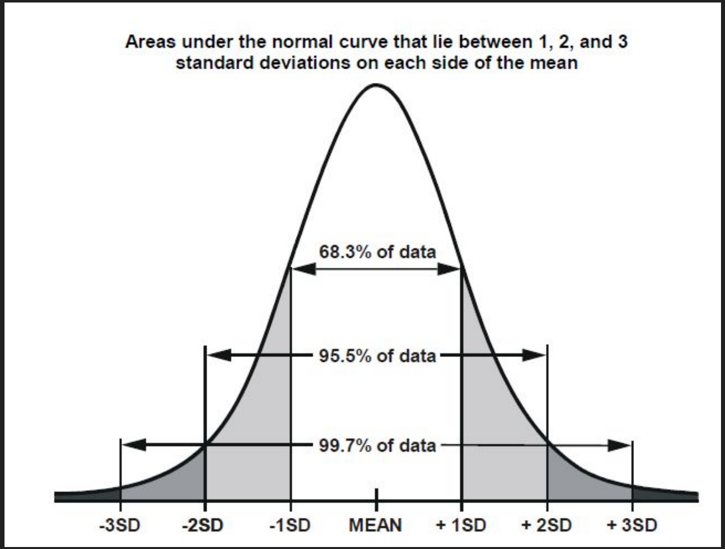
Richmond et al. 2018. The long-term impact of folic acid in pregnancy on offspring DNA methylation: follow-up of the Aberdeen Folic Acid Supplementation Trial (AFAST)

How is it generated?

- To generate Q-Q plots take your sample data, sort it in ascending order, and then plot them versus **quantiles** calculated from a theoretical distribution.

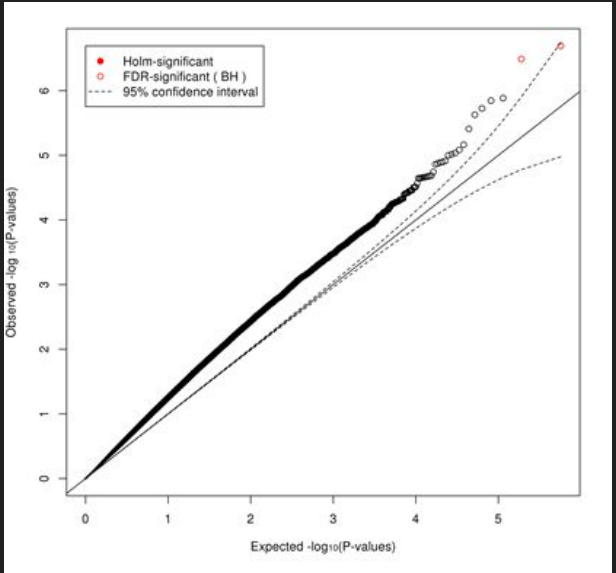
↓

- referred to as “percentiles”
- points in your data below which a certain proportion of your data fall.
- e.g, in the classic bell-curve standard Normal distribution with a mean of 0. The 0.5 quantile, or 50th percentile, is 0.


Areas under the normal curve that lie between 1, 2, and 3 standard deviations on each side of the mean
68.3% of data
95.5% of data
99.7% of data
-3SD
-2SD
-1SD
MEAN
+ 1SD
+ 2SD
+ 3SD

Sharma et al. 2019. Why is Normal Distribution Bell Shaped?

Quick Way to Interpret it

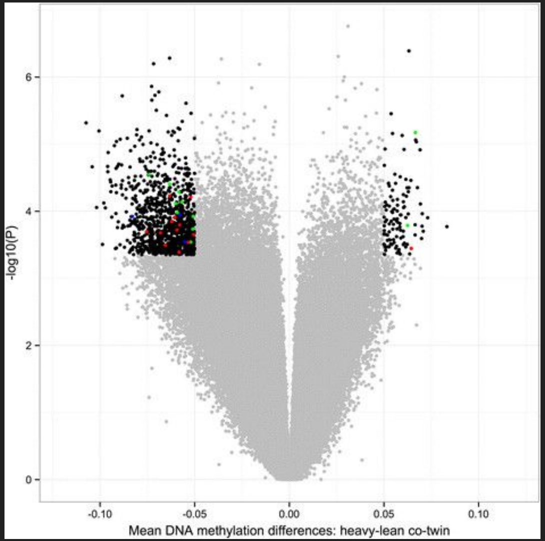


- If the line between x and y axes aligned well making a 45 degree line then it means that there's no evidence of DNA methylation differences between our sample groups at the CpG sites.
- - If you see deviation (they don't align well), it means that there's evidence of methylation differences between our sample groups.

Richmond et al. 2018. The long-term impact of folic acid in pregnancy on offspring DNA methylation: follow-up of the Aberdeen Folic Acid Supplementation Trial (AFAST)

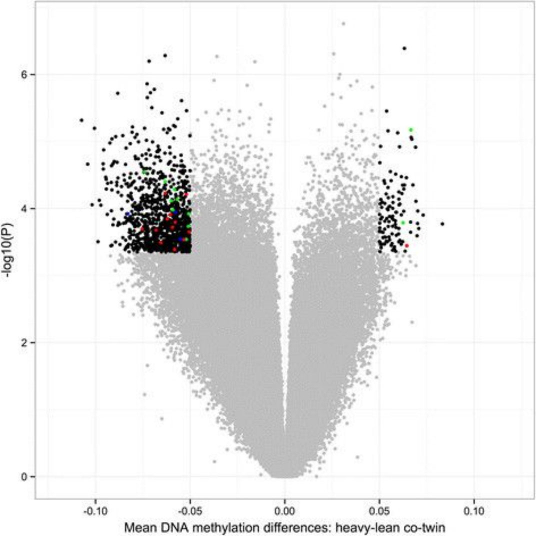
Volcano plot

- named because it looks like a volcano
- Basically just a scatter plot
- visualize the p-values to the magnitude of change (fold change, mean B-value difference)
- helps us see or identify which cpg sites/genes that are significantly differentially methylated and have large fold changes.



Richmond et al. 2018. The long-term impact of folic acid in pregnancy on offspring DNA methylation: follow-up of the Aberdeen Folic Acid Supplementation Trial (AFAST)

First glance on volcano plot

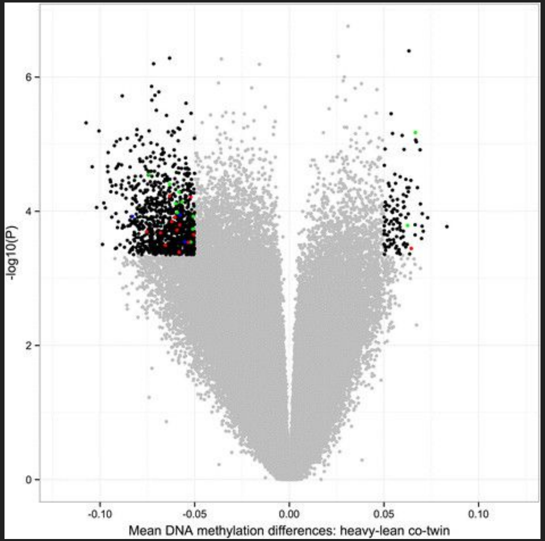


- basically a scatter plot
- y axis represents the observed -log10(p-values)
- x axis represents the Fold Change
- y axis is from our results
- x axis is also from our results
- Small circles are our instances??

Richmond et al. 2018. The long-term impact of folic acid in pregnancy on offspring DNA methylation: follow-up of the Aberdeen Folic Acid Supplementation Trial (AFAST)

Fold Change

- Fold change = treated expression level / control expression level
- Mostly they are usually written and converted using the log function



Richmond et al. 2018. The long-term impact of folic acid in pregnancy on offspring DNA methylation: follow-up of the Aberdeen Folic Acid Supplementation Trial (AFAST)

Pathway Enrichment Analysis

In general the procedure can be summarized into this 3 big steps:

1. Definition of a gene list of interest using omics data
2. Pathway enrichment analysis
3. Visualization and interpretation of pathway enrichment analysis results

Reimand J. 2019. Pathway enrichment analysis and visualization of omics data

Continuation from the previous training

In general the procedure can be summarized into this 3 big steps:

1. Definition of a gene list of interest using omics data
2. Pathway enrichment analysis
3. Visualization and interpretation of pathway enrichment analysis results

Reimand J. 2019. Pathway enrichment analysis and visualization of omics data

Stage 1: definition of a gene list of interest using omics data

We mostly have done this stage

During this stage we perform the differential methylation analysis using R.

- Doing this step we then can discover our gene of interest based on their differentially methylated significance of adjusted p-value.

Reimand J. 2019. Pathway enrichment analysis and visualization of omics data

Stage 2: Pathway Enrichment Analysis

We also have done this stage, we’re using one of the way, in this case GSEA

- GSEA is a threshold-free method that analyzes all genes on the basis of their differential expression rank, or other score
- GSEA is particularly suitable and is recommended when ranks are available for all or most of the genes in the genome
- GSEA searches for pathways whose genes are enriched at the top or bottom of the ranked gene list. For instance, if the topmost differentially expressed genes are involved in the cell cycle, this suggests that the cell cycle pathway is regulated in the experiment.

Reimand J. 2019. Pathway enrichment analysis and visualization of omics data

Stage 3: visualization and interpretation of pathway enrichment analysis results

This one we haven’t done it yet

- **Pathway information is inherently redundant**, as genes often participate in multiple pathways, and databases may organize pathways hierarchically by including general and specific pathways with many shared genes
- Consequently, pathway enrichment analysis often **highlights several versions of the same pathway**
- To address the redundancy problem we usually use the following tools EnrichmentMap and ClueGO

Reimand J. 2019. Pathway enrichment analysis and visualization of omics data

Stage 3: visualization and interpretation of pathway enrichment analysis results

An enrichment map helps identify interesting pathways and themes.

1. Expected themes **should be identified to help validate the pathway enrichment analysis results** . For instance, growth-related pathways and other hallmarks of cancer are expected to be identified in analyses of cancer genomics datasets
2. Pathways **not previously associated with the experimental context are evaluated more carefully** as potential discoveries. Pathways and themes with the strongest ESs should be studied first, followed by progressively weaker signals
3. **Interesting pathways are examined in more detail**, examining genes within the pathways (e.g., expression heat maps and the GSEA leading edge genes)

Reimand J. 2019. Pathway enrichment analysis and visualization of omics data