Analysis of DNA methylation data

EWAS

JON BOHLIN BIOINFORMATICS NIPH (FHI)

Methods applied in EWAS analyses

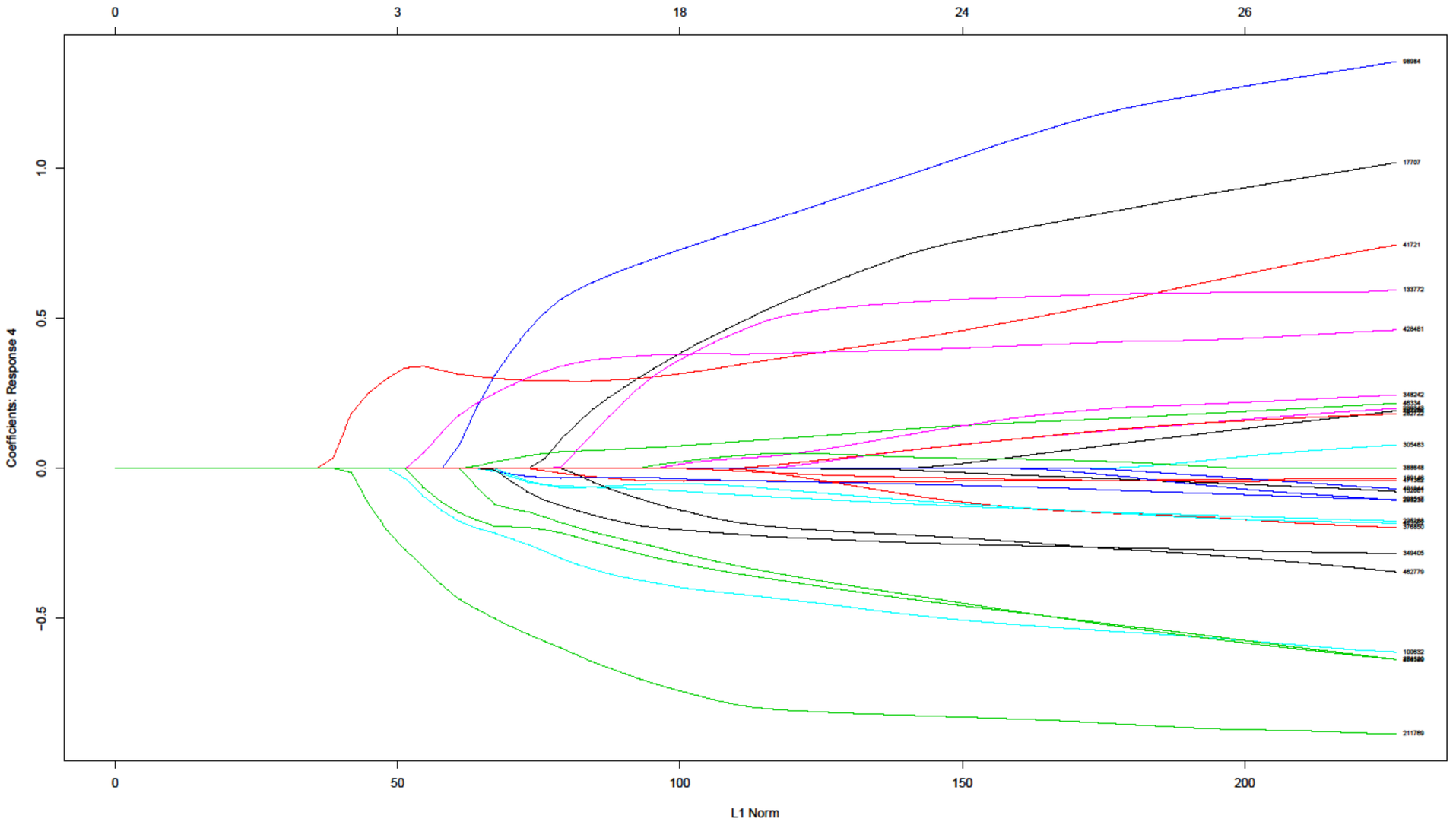
- The data is really just a matrix of samples vs. Transformed intensity values from hybridised methylation sites
- That is, Illumina DNAm data is just a NxK matrix of values between 0 and 1
- We are interested in whether specific probes present in all or most samples exhibit association with a phenotype (i.e. trait/disease)

Other quantitative methods that are used in EWAS analyses

- T-test
- 1-1 regression (Limma, robust, GLM, etc.)
- Shrinkage methods (LASSO/RIDGE+variants)
- LASSO+RIDGE=ELNET
- (CP)PLS/PCA regression
- ++ active research field

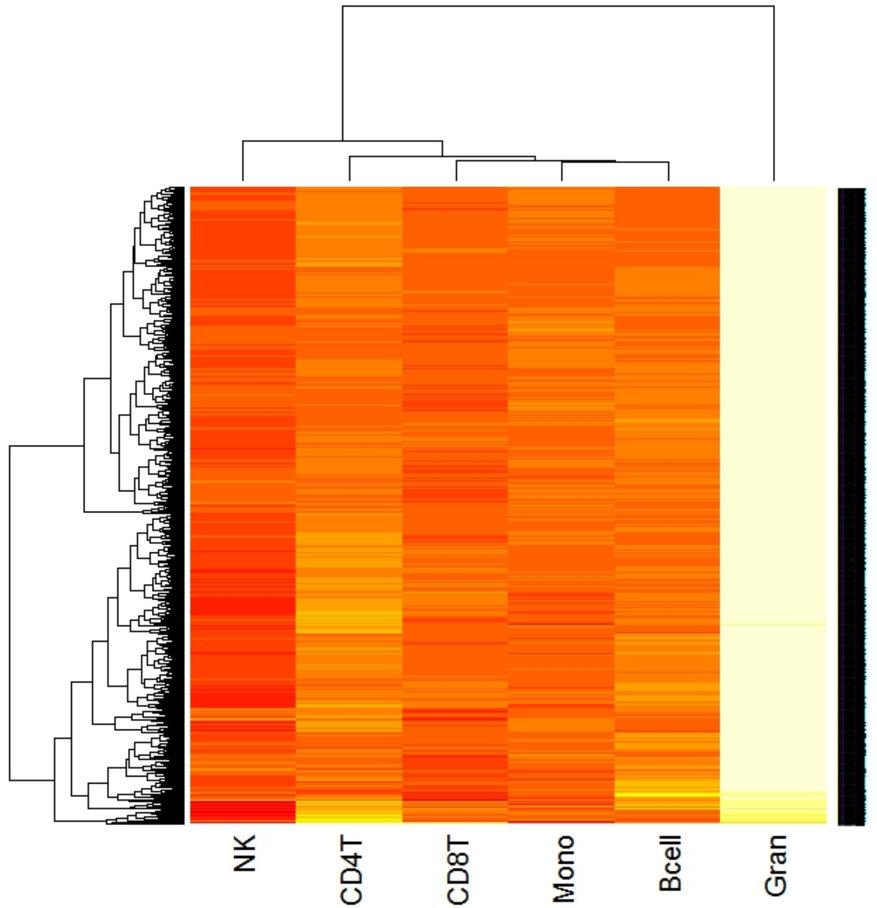
Methods for screening

- The elastic net (glmnet) is EXTREMELY fast, and works well on a standard PC (not to great with categorical variables)
- Can run GLMs, *i.e.* «Poisson», «multinomial», «binomial» also survival
- Takes whole methylation matrix as explanatory variables in one go!
- RIDGE and LASSO special cases in a continum of methods set with a parameter
- Variance not computed


0
3
18
24
26
1.0
0.5
0.0
−0.5
Coefficients: Response 4
0
50
100
150
200
L1 Norm
98984
17707
41721
133772
428481
348242
46334
305483
388648
349405
462779
100632
211769

EWAS analyses:

- Further analyses can be subsequently carried out by using standard linear regression to get the «variance explained» R^2 statistic
- Adjust for: Sex, Smoking, age, cell type
- Ethnicity (often not performed)
- Family background (random effects models)
- First PCA components often related to cell type and sex
- Age may also exhibit a a strong «global» influence
- PCA may adjusting away effects of phenotype.



Regression results

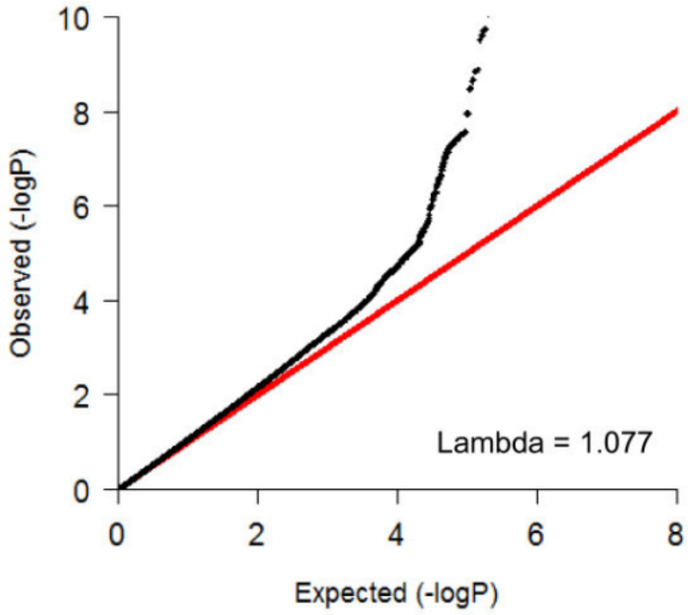
Table 1

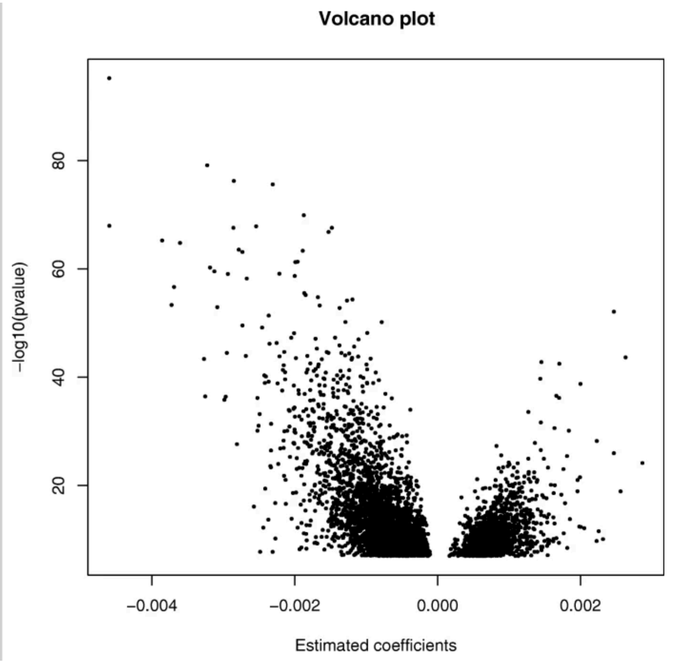
	est	se	pval
cg00005974	0.004296261	0.0009989015	4.345633e-05
cg00004962	0.003799851	0.0009631020	1.586828e-04
cg00001070	-0.003374497	0.0008764390	2.219065e-04
cg00007232	0.003697749	0.0009846563	3.084131e-04
cg00000845	0.003388276	0.0009250157	4.229506e-04
cg00009598	0.003565897	0.0009857787	4.933585e-04
cg00000941	0.003430470	0.0009732136	6.716065e-04
cg00002848	-0.003285014	0.0009320666	6.726013e-04
cg00004718	0.003507581	0.0009975113	6.908778e-04
cg00002858	0.003445350	0.0009992500	8.648064e-04
cg00002893	-0.003197148	0.0009326856	9.231261e-04
cg00004361	0.003326078	0.0009714479	9.353788e-04
cg00007169	0.003255150	0.0009585613	1.024061e-03
cg00001373	0.003154435	0.0009526732	1.343493e-03
cg00003000	0.003189906	0.0009679909	1.412346e-03
cg00009140	0.003138950	0.0009540474	1.436013e-03
cg00006890	0.003201494	0.0009814987	1.570154e-03

Multiple testing

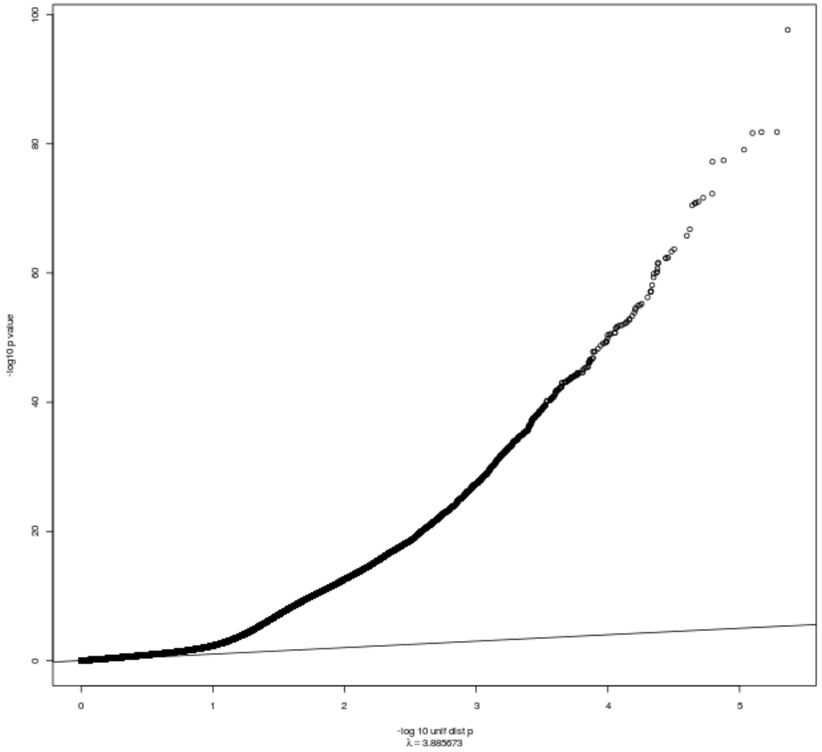
- Expect 5 significant p<0.05 for every 100 CpGs from pure randomness (if independent)
- Bonferroni correct (multiply p value with number of tests)
- Methylome highly correlated so Bonferroni is too strict
- FDR a good alternative (Q values), however FDR 0.05 difficult to reproduce
- Many more genomic CpG loci than array probes…
- …probes may correlate with missing CpG loci => increase sample size
- CpGs vary check genes and regions (CpG islands, shelfs, shores, promoter regions, TSSs, enhancers, gene body, etc)
- P value “torturing” common in methylation analyses

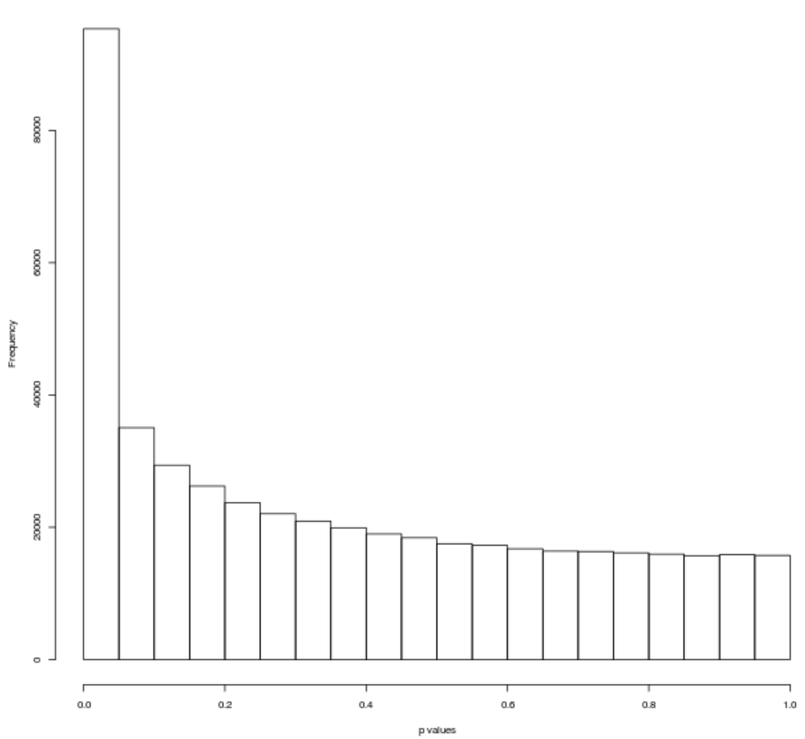
QQ plot and Volcano plot

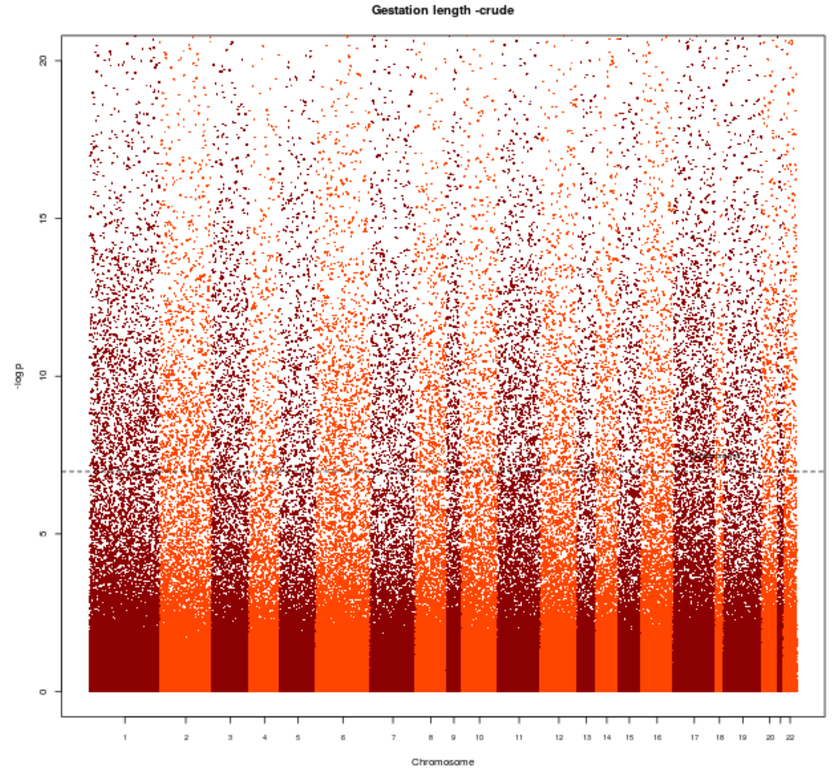




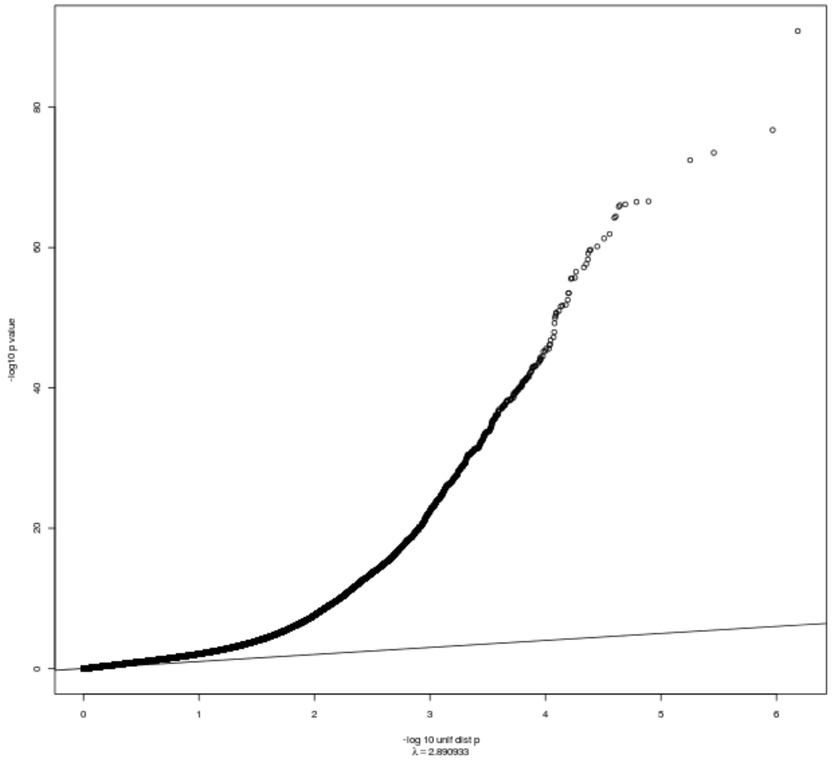
Associations between methylation and gestational age?

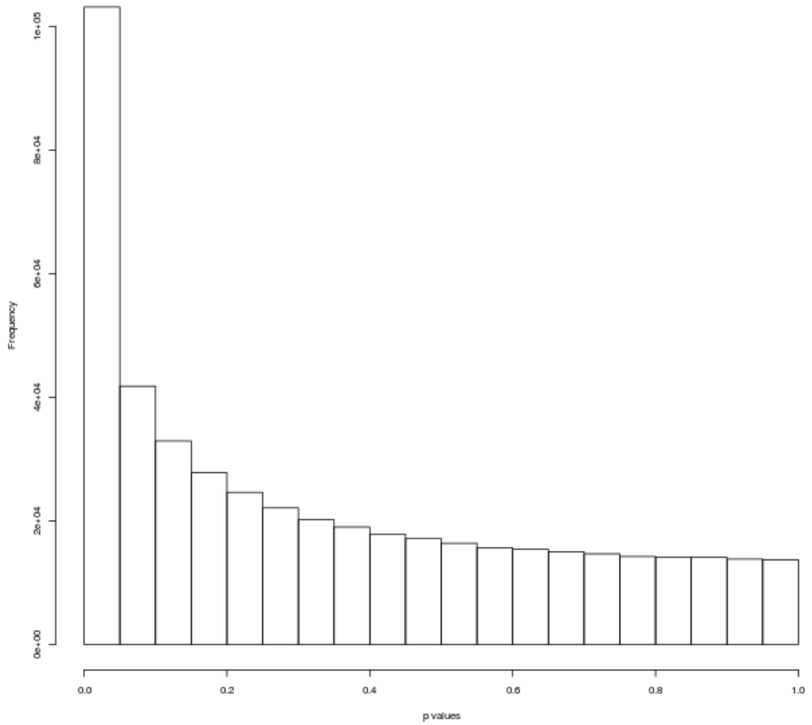


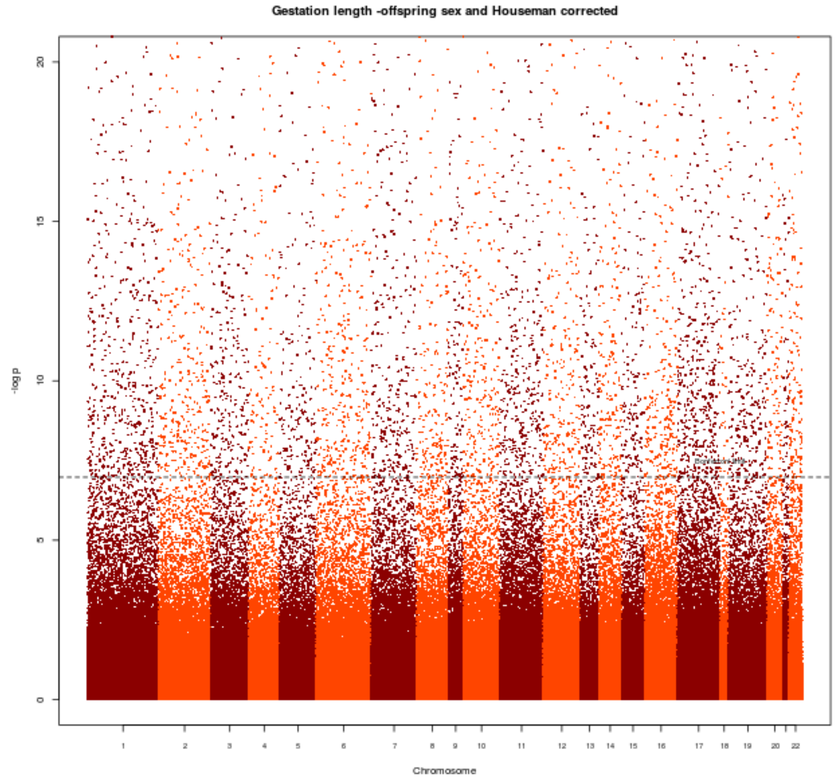




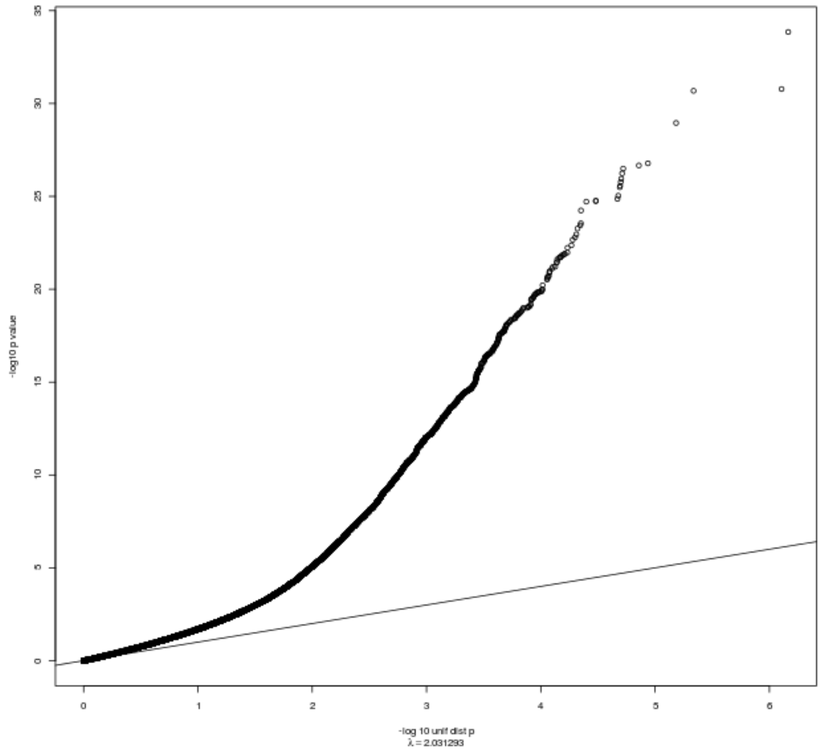
Adjustment for cell type (Houseman) and sex

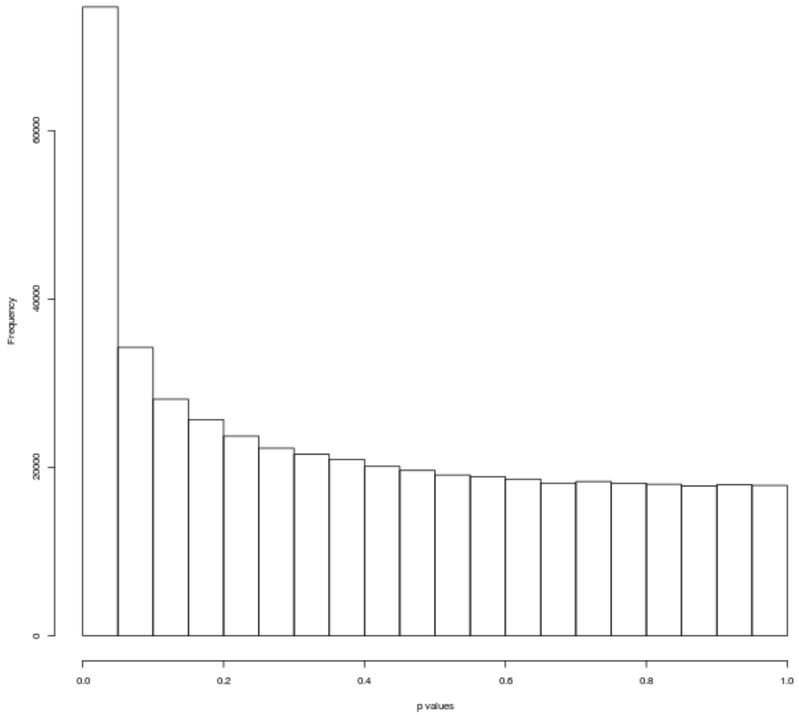


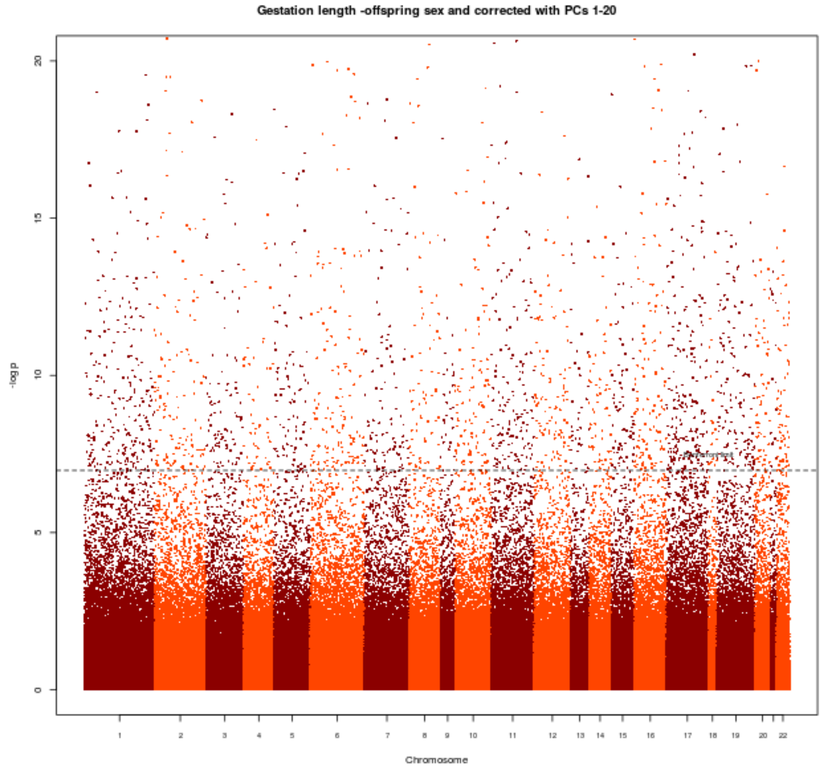




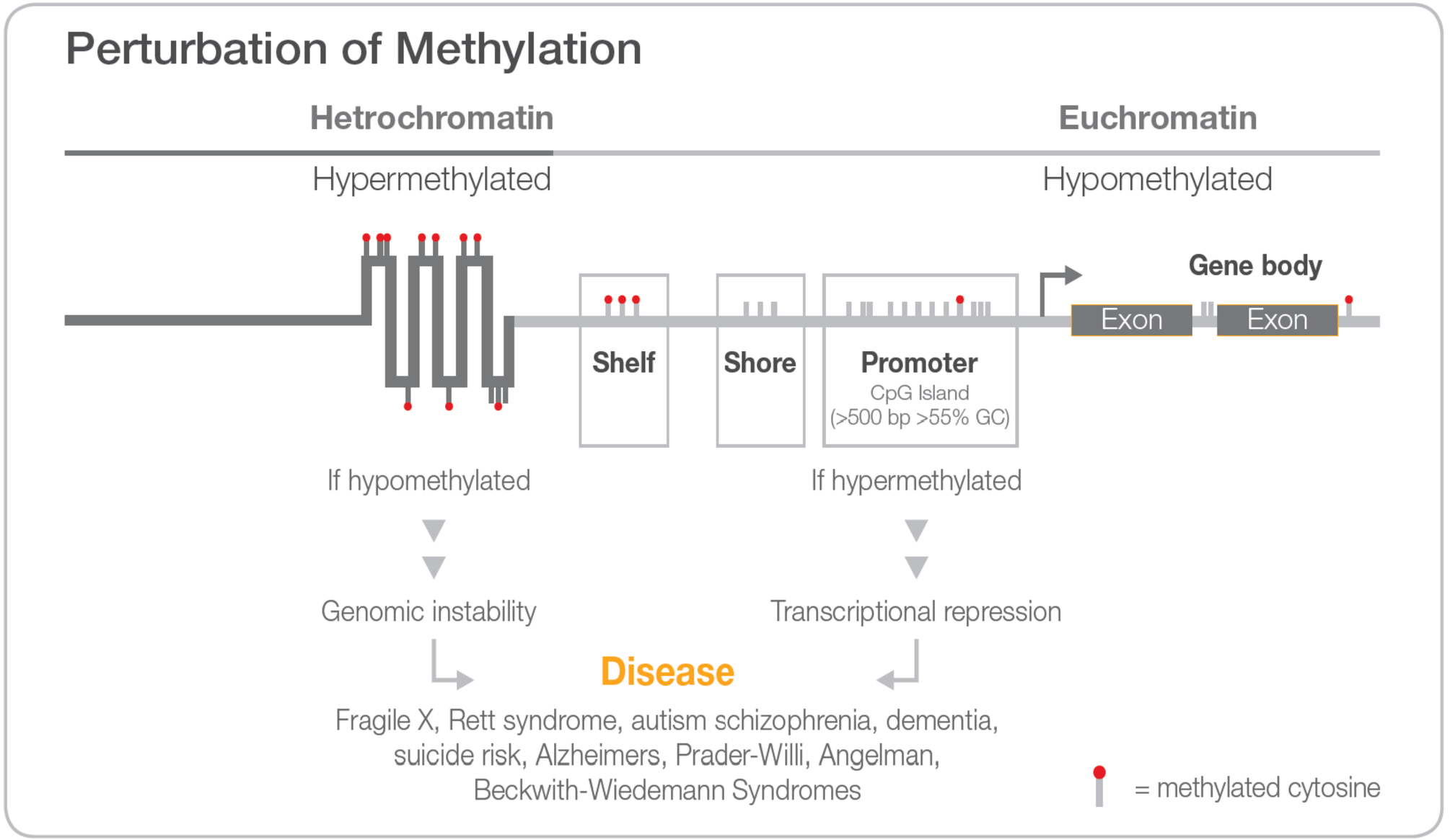
…added 20 PC’s


-log10 p value
-log 10 unif dist p
λ = 2.031293


Frequency
p values


Gestation length -offspring sex and corrected with PCs 1-20
-log p
Chromosome

Even more refined analyses…



Results/impressions from experience…

- Physical measurements seems to be significantly correlated with methylation data
- Data from questionnaires difficult due to a number of reasons (missing, «subjective answers»)
- «Kostdata» (nutrition) difficult, effects maybe too weak to discover
- Behavioural/cognitive data difficult to detect
- DNA methylation mark does not seem to be easily altered

All is not lost…

- Bureaucracy in human data horrible
- Often problems with availability and delays
- GDPR can suddenly break the flow
- NCBI - GENBANK - ton’s of data available (Gene Expression Omnibus)
- https://www.ncbi.nlm.nih.gov/gds