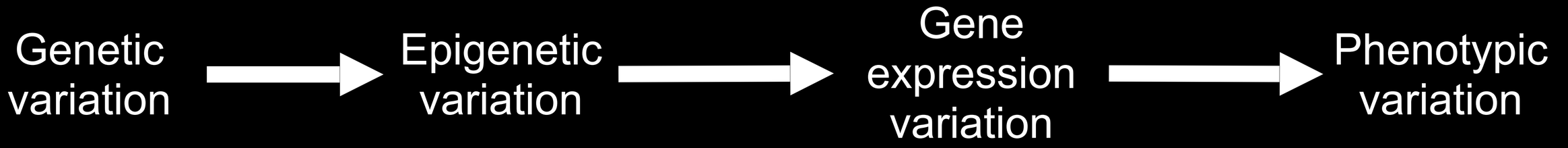
Introduction to genome- and epigenome-wide association studies

Nuno Sepúlveda, 28.11.2025

Course content

3. Introduction to epigenome-wide association studies (EWAS)
    A. Basics of epigenetic data - proportion of methylation per probe
    B. Quality controls
    C. Data analysis methods - simple statistical tests, linear regression, beta regression, multiple testing corrections
    D. Main outputs - Manhattan plots, qq-plots

Genotype-phenotype mapping



Single nucleotide polymorphisms
Copy number variation
Inversion/deletion

Post-translation modification
DNA methylation
microRNA
Chromosomal modifications

Different symptoms for the same disease

Supporting material (check Materials & Methods of this paper)

Heliyon 7 (2021) e07665

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

The SARS-CoV-2 receptor angiotensin-converting enzyme 2 (ACE2) in myalgic encephalomyelitis/chronic fatigue syndrome: A meta-analysis of public DNA methylation and gene expression data

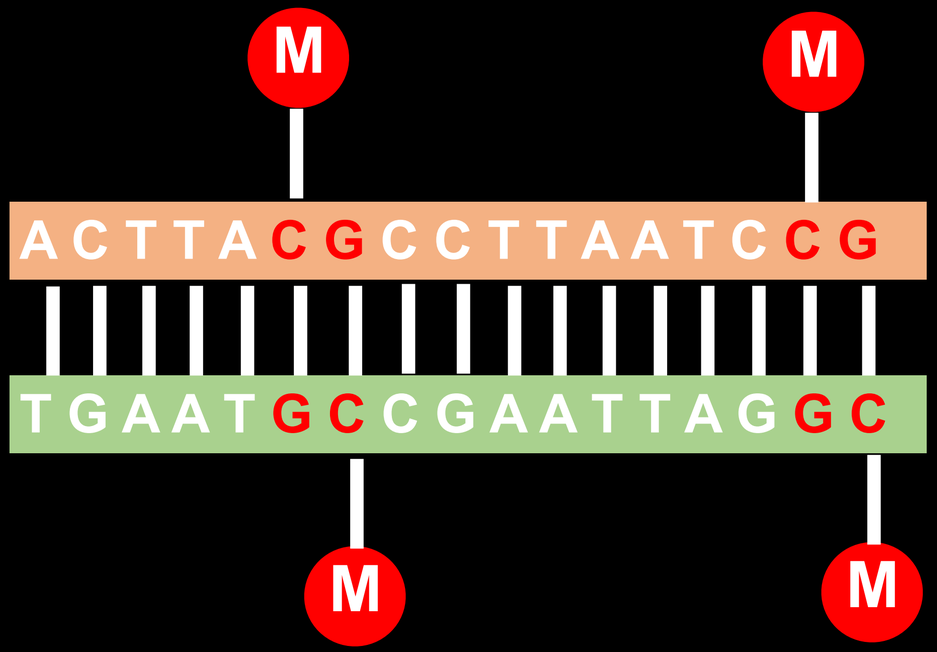
João Malato [a,b], Franziska Sotzny [c,**], Sandra Bauer [c], Helma Freitag [c], André Fonseca [d], Anna D. Grabowska [e], Luís Graça [a], Clara Cordeiro [b,d], Luís Nacul [f,g], Eliana M. Lacerda [f], Jesus Castro-Marrero [h], Carmen Scheibenbogen [c], Francisco Westermeier [i,j], Nuno Sepúlveda [b,c,k,*]

[a] Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal
[b] CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal
[c] Charité - Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt Universität zu Berlin and Berlin Institute of Health, Institute of Medical Immunology, Berlin, Germany
[d] Faculdade de Ciências e Tecnologia, Universidade do Algarve, Faro, Portugal
[e] Department of Biophysics, Physiology, and Pathophysiology, Medical University of Warsaw, Warsaw, Poland
[f] Department of Clinical Research, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom
[g] Complex Chronic Diseases Program, British Columbia Women's Hospital and Health Centre, Vancouver, British Columbia, Canada
[h] Vall d'Hebron Hospital Research Institute, Division of Rheumatology, ME/CFS Unit, Universitat Autònoma de Barcelona, Barcelona, Spain
[i] Institute of Biomedical Science, Department of Health Studies, FH Joanneum University of Applied Sciences, Graz, Austria
[j] Centro Integrativo de Biología y Química Aplicada (CIBQA), Universidad Bernardo OHiggins, Santiago, Chile
[k] Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom

DNA methylation (CG sites)

Gene might be expressed

Production of the protein



Gene might not be expressed

No production of the protein

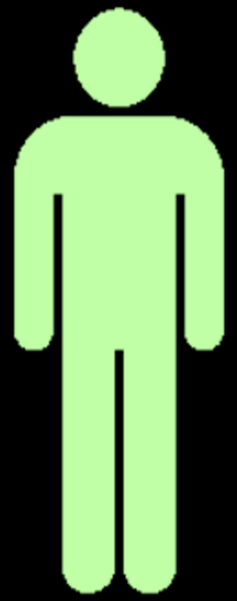
Epigenome-wide association studies

Objective:

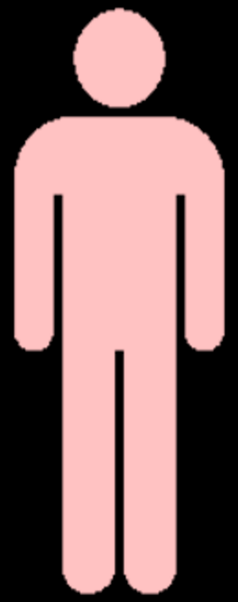
Search the whole genome for methylation modifications associated with the phenotype (i.e., presence of disease).

Find sites where the level of methylations is different between cases and controls.

Example: case-control study



n=48
75% Women
Mean age of 37 years old
Mean BMI of 27 kg/m^2



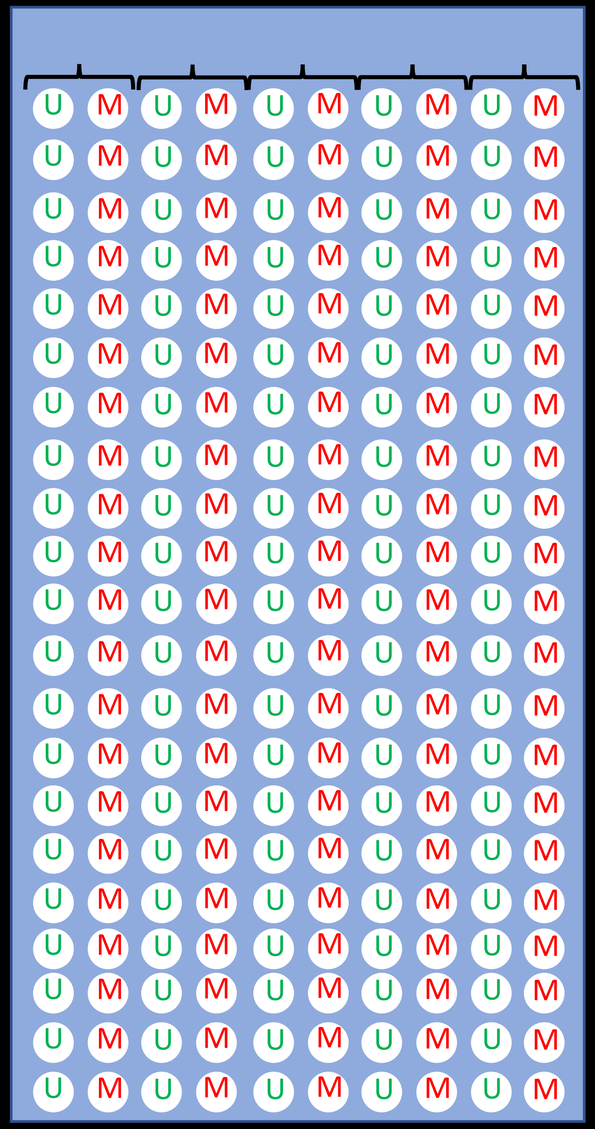
n=61
79% Women
Mean age of 41 years old
Mean BMI of 27 kg/m^2

Human Methylation 450K Array

Herrera et al (2018). Epigenetics 13:12, 1174-1190

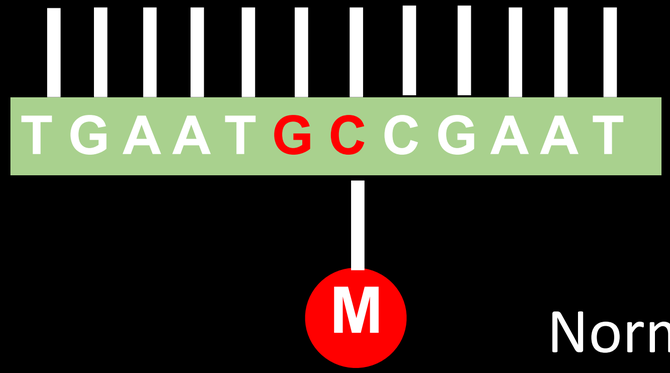
Basics of epigenetic data: methylation arrays

Array



Unmethylated probe

Green light



Methylated probe

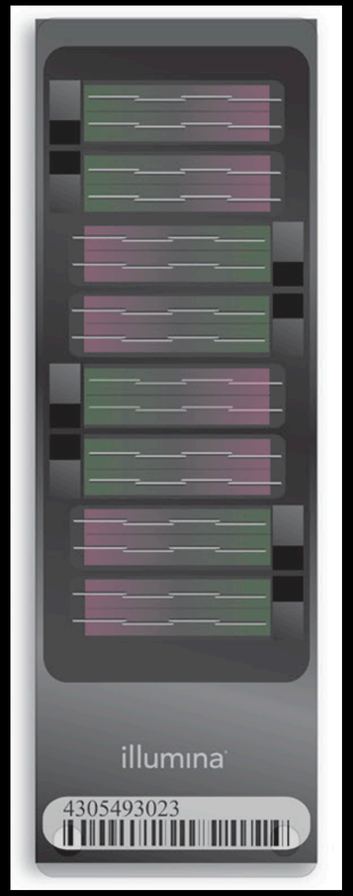
Red light

Intensity of red light / (Intensity of red light + Intensity of green light)

Normalize the raw data within each array and across arrays (R package mini)

Illumina Arrays

Array	Number of methylation sites including not CpG sites	Number of samples per array
HumanMethylation450K	> 450,000	12
MethylationEPIC V2.0	> 935,000	8

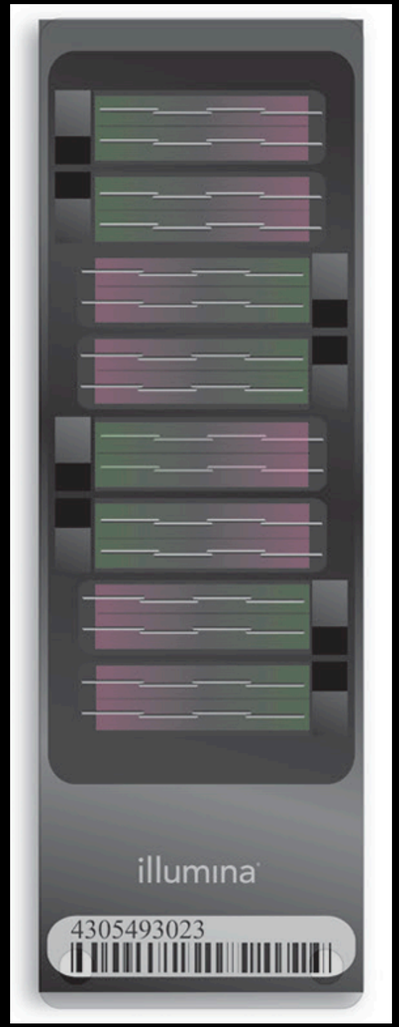


Reduced representation Bisulfite sequencing (outside the scope of this course)

Location of methylation sites

Table 3: Infinium MethylationEPIC v2.0 coverage of genomic regions

Feature type	No. of features mapped	% features covered	Avg no. of loci/ feature
RefSeq			
NM_TSS200[a]	51,688	82%	2.8
NM_TSS1500	59,981	96%	5.6
NM_5′ UTR	42,051	67%	1.7
NM_1stExon	44,471	71%	1.8
NM_3′ UTR	39,407	63%	1.3
NM_Exonic	207,398	28%	0.5
NR_TSS200	12,706	68%	2.0
NR_TSS1500	15,961	86%	3.9
NR_1stExon	9810	53%	1.4
NR_Exonic	30,211	25%	0.5
GenCode Basic v41			
TSS200	160,572	79%	1.7
TSS1500	197,603	80%	3.9
5′ UTR	61,823	59%	1.4
First Exon	118,516	47%	1.1
3′ UTR	41,659	53%	1.2
Exonic	417,055	26%	0.5
Enhancers			
DNase hypersensitivity sites[b]	432,393	16%	0.2
FANTOM5 Enhancers[c]	23,852	84%	1.0
CisReg Site Evid 40-50[d]	19,159	70%	1.3
CisReg Site Evid 50-60	21,609	67%	1.2
CisReg Site Evid 60-70	30,152	61%	1.1
CisReg Site Evid 70-80	66,446	47%	0.8
CisReg Site Evid > 80	153,712	19%	0.3
Cancer driver mutations			
Cancer driver mutations[e]	473	81%	0.8



Software

GenomeStudio™ Methylation Module Software

SeSAMe

minfi (R software)

Basics of epigenetic data: data normalization

Similar to microarrays technology

Normalize the raw data within each array and across arrays (R package mini)

GEO database - NCBI website

Loads of free datasets to play around.

Data quality controls

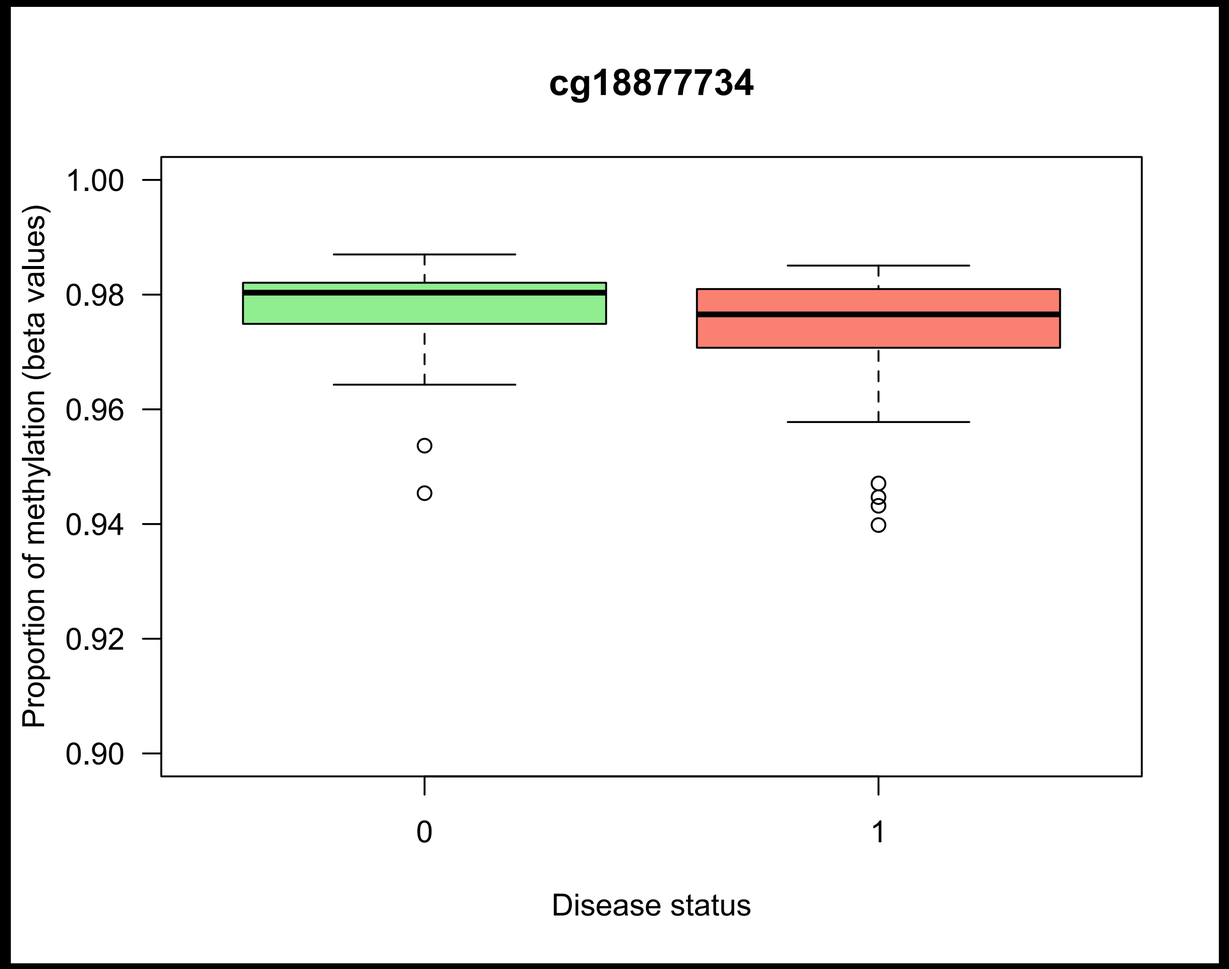
Check on probes:

- showing a high probability of detection (provided by the machine);
- not cross-reactive with other genomic regions;
- not affected by the presence of single nucleotide polymorphisms (SNPs) with high minor allele frequencies;
- having sufficient variability (0.01<proportion of methylation<0.99)

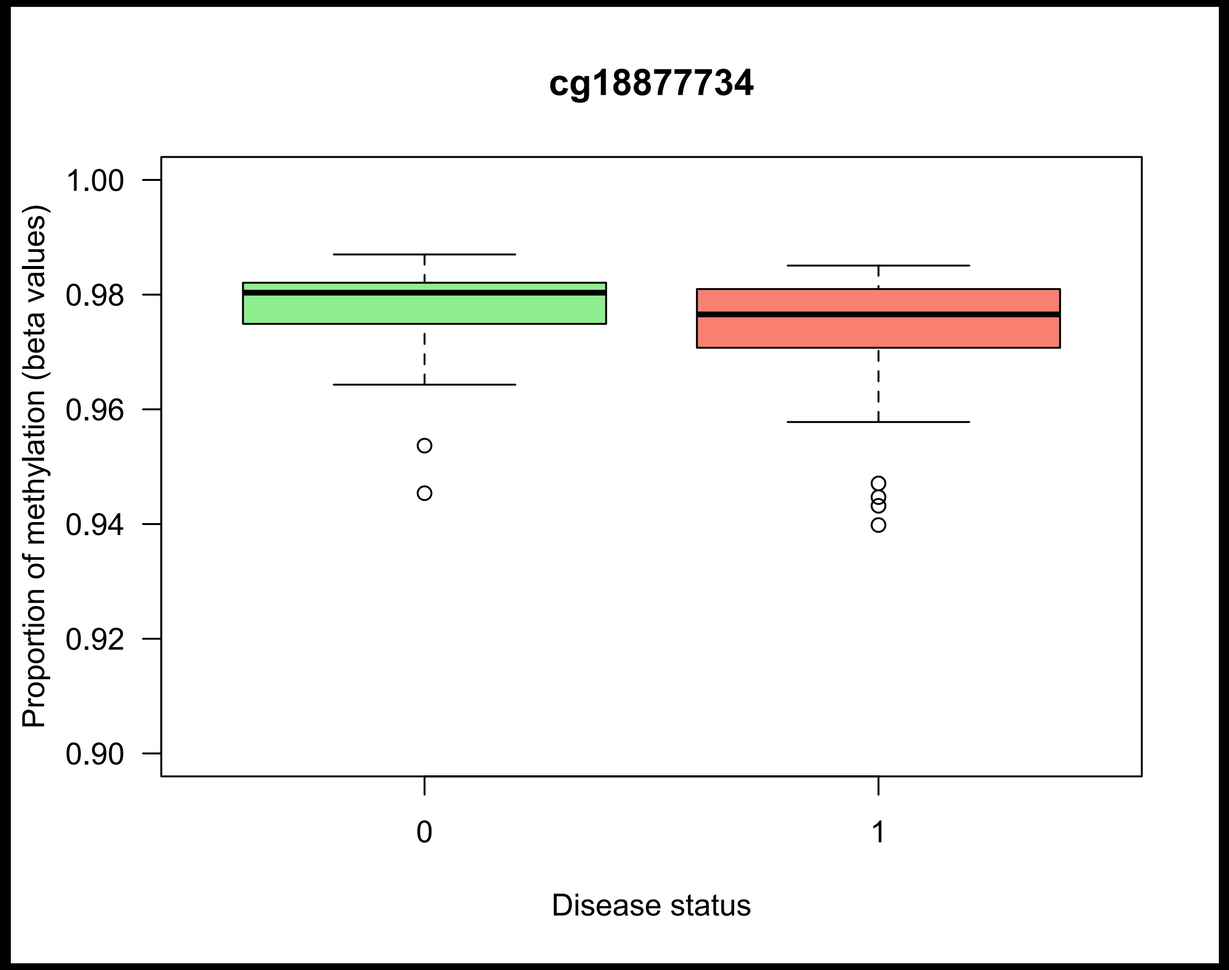
Check on individuals:

- Biological sex (X-linked CG probes)

Data from a single probe

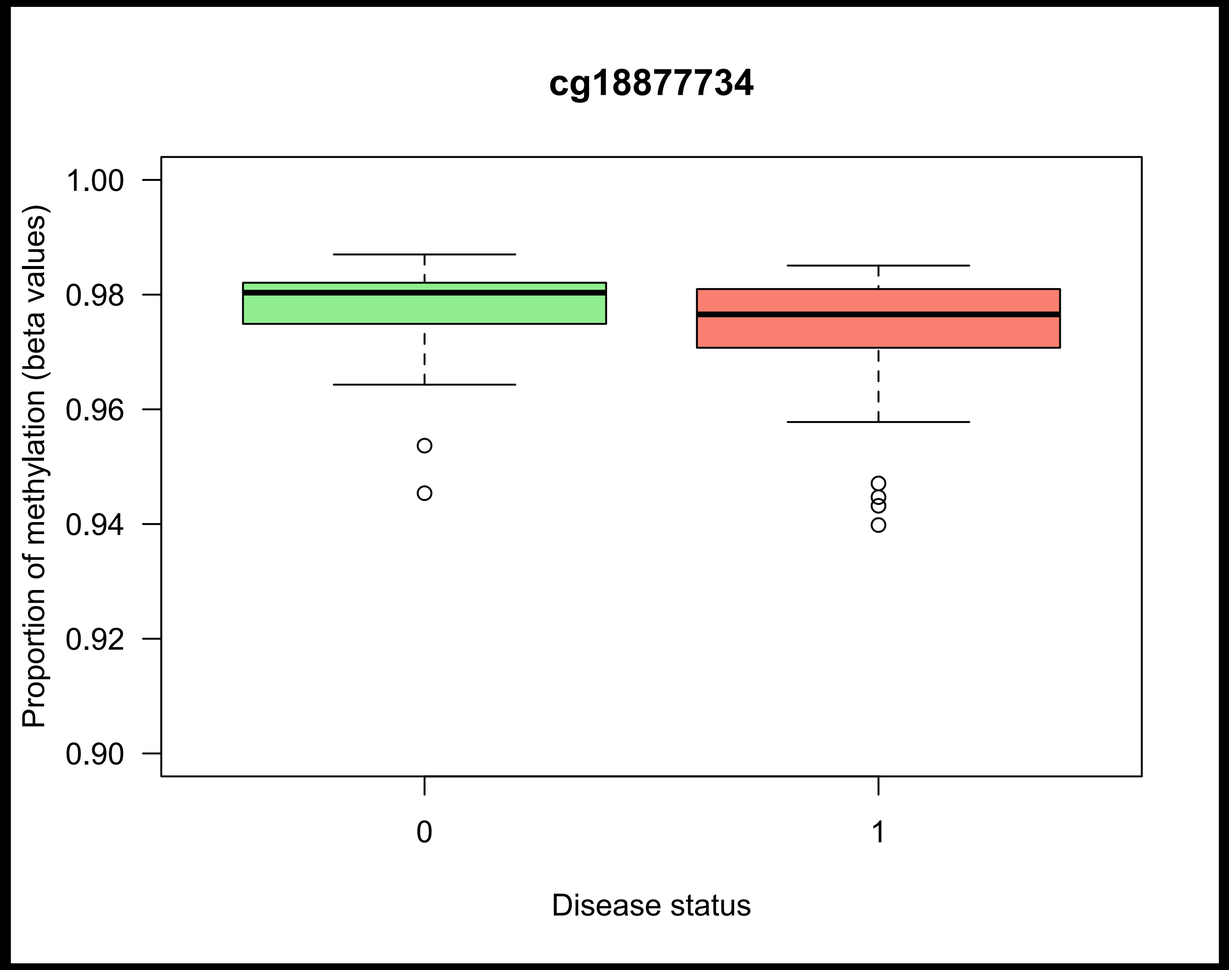


Data from a single probe



Which statistical tests can we used to check whether patients differ from health controls in terms of the methylation level for this probe?

Simple statistical analysis of a single probe

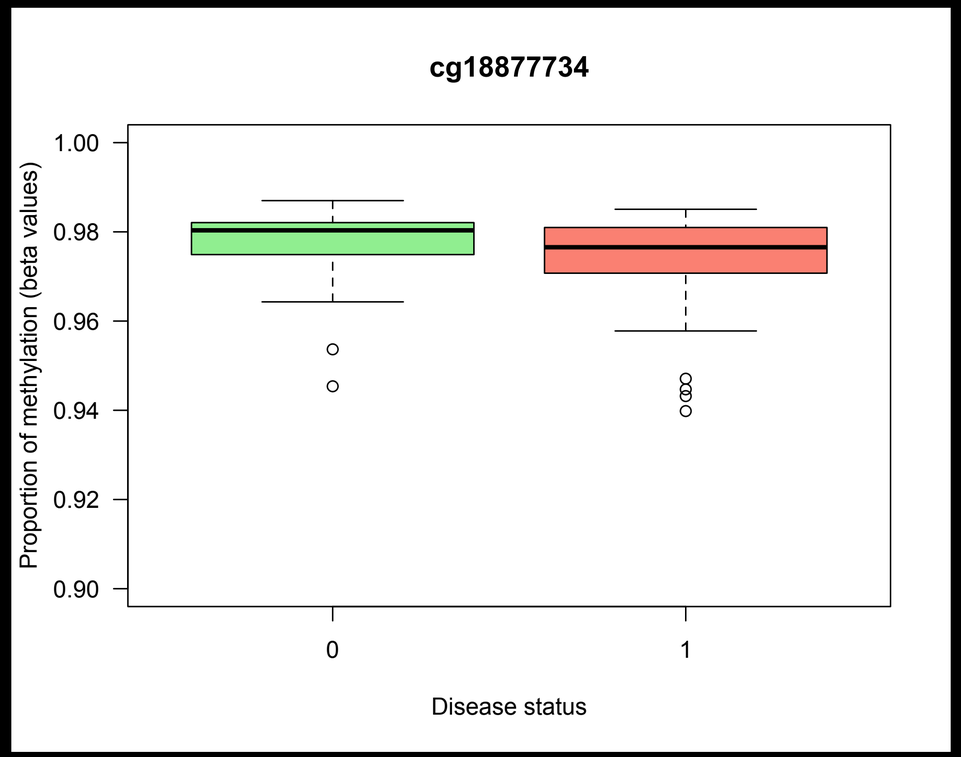


T test

Mann-Whitney test

Statistical analysis of a single probe adjusting for covariates

Linear regression (as in GWAS for quantitative traits)



Y_{ij} = methylation levels of probe j in individual i

$x_{group,i}$ = group of individual i

$x_{k,i}$ = value of covariate k for the individual i

$$Y_{ij} = \beta_{0j} + \beta_{1j} x_{group,i} + \sum_{k=2}^{p} \beta_{kj} x_{k,i} + \underbrace{\epsilon_{ij}}_{\text{residuals}}$$

$\epsilon_{ij} \rightsquigarrow \text{Normal}\left(0; \sigma^2\right)$

$H_0 : \beta_{1j} = 0$ versus $H_1 : \beta_{1j} \neq 0$

Epigenome-wide association studies

Perform this test on data of each probe

$$H_0 : \beta_{1j} = 0 \text{ versus } H_1 : \beta_{1j} \neq 0$$

$$j = 1, \ldots, M \text{ (number of probes)}$$

Wald's score test

Wilks' likelihood ratio test

Correct the p-values of each individual test by a procedure controlling the false discovery rate (e.g., Benjamini-Hochberg procedure)

Construct a manhattan plot as learned for GWAS

Report the significant probes and their location

Exercise: data_ace_ace2.csv

Information about the CG probes: ace_ace2_annotation.csv

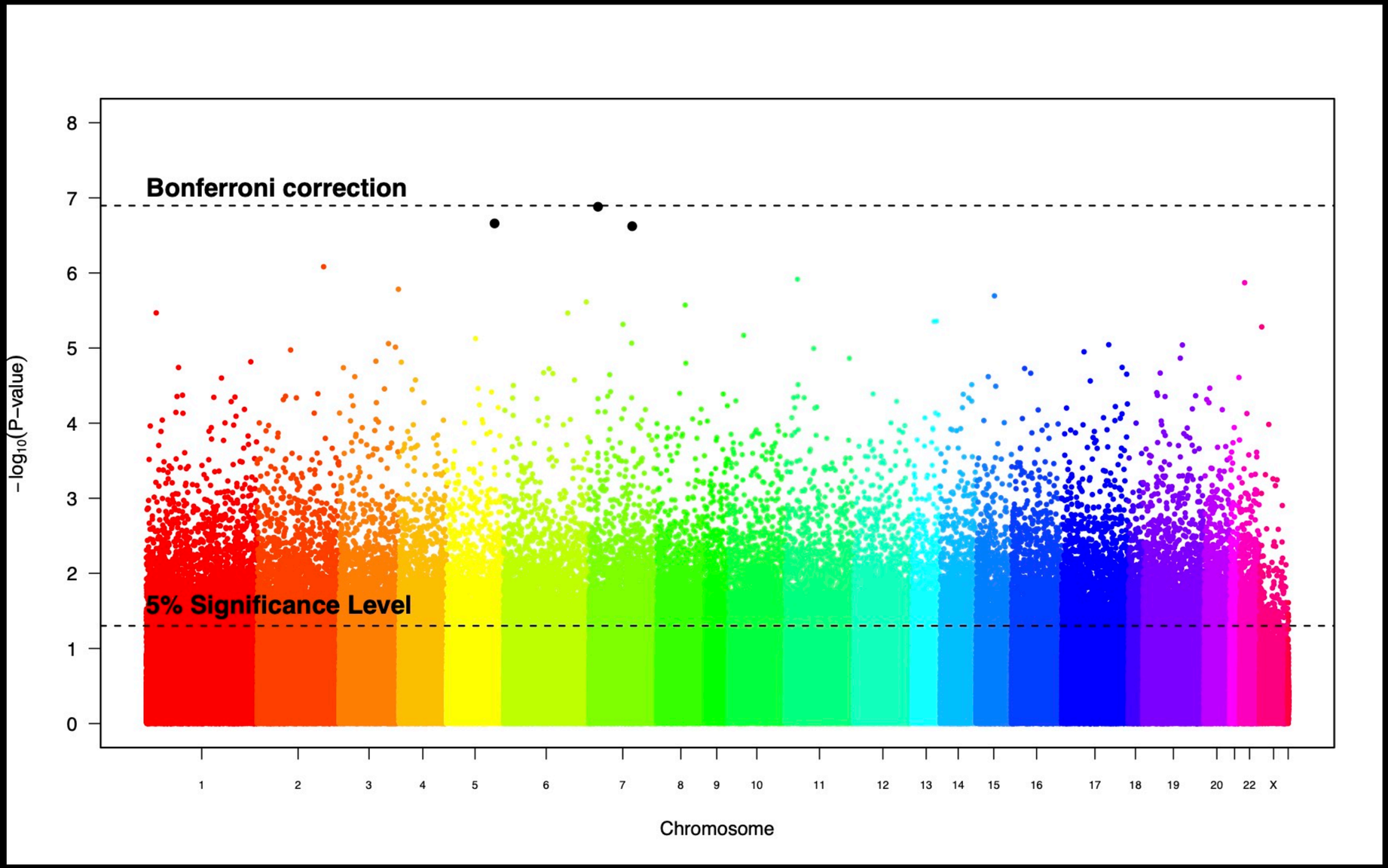
Compare the methylation levels of cg18877734 between patients with chronic fatigue syndrome and healthy controls using T test and Mann-Whitney test on beta values.

Which test is preferable to analyse data of this CG probe?

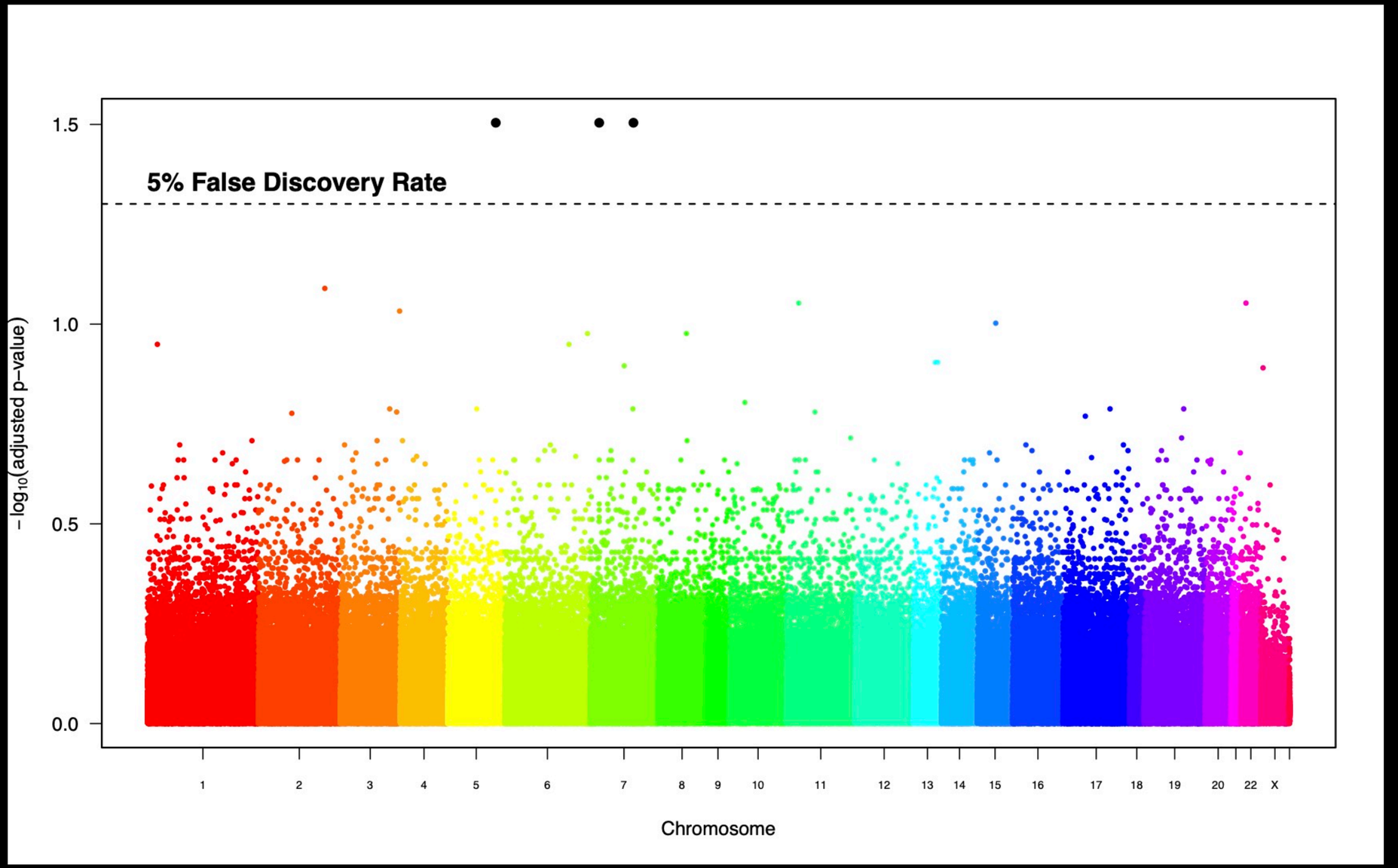
Repeat previous analysis using linear regression in which “cg18877734” is the outcome and Disease and Female as covariate?

Is there any evidence that this probe is differentially methylated between patients and controls?

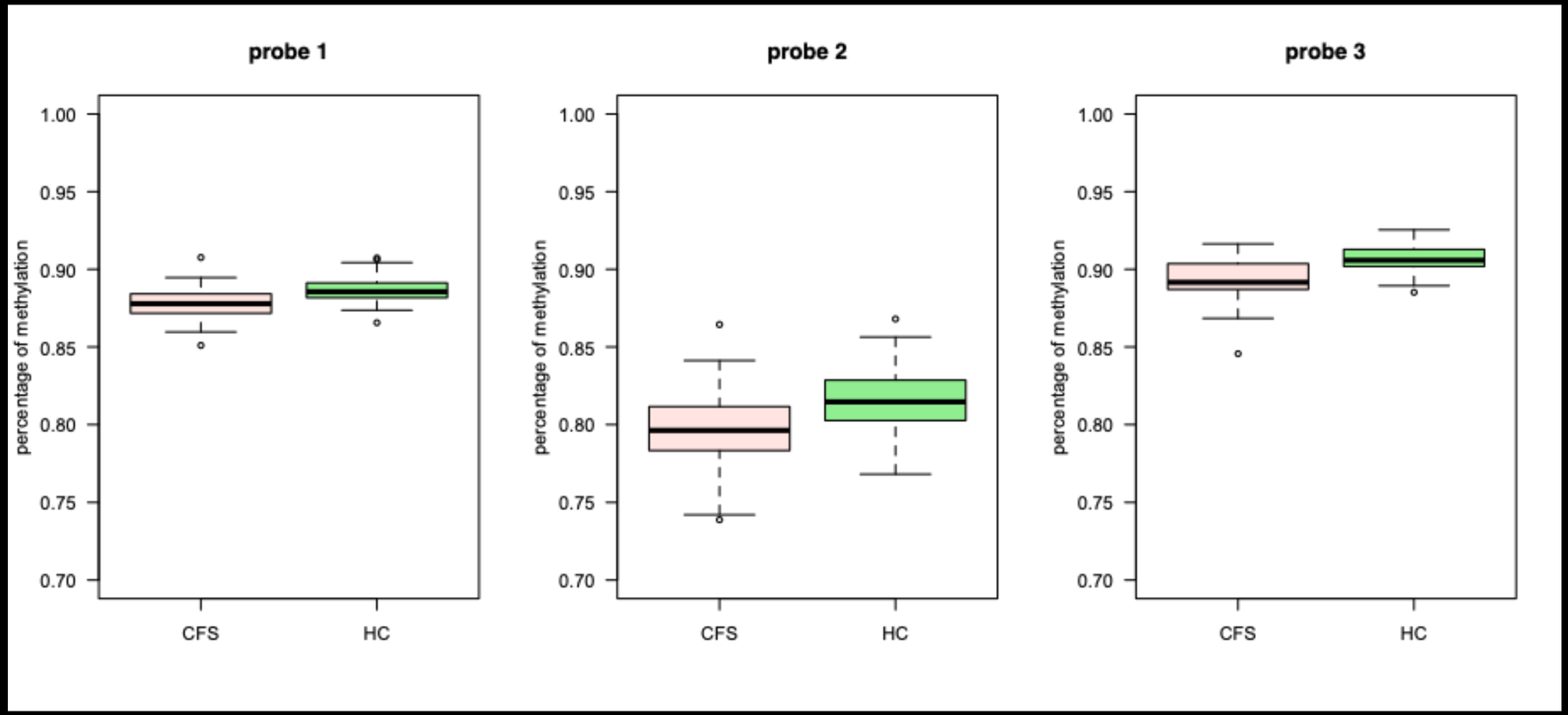
Manhattan plot based on Bonferroni correction for multiple testing



Manhattan plot based on bonferroni correction for multiple testing



Decreased methylation in patients with ME/CFS



More expression for genes associated with these probes

Analysis based on M values

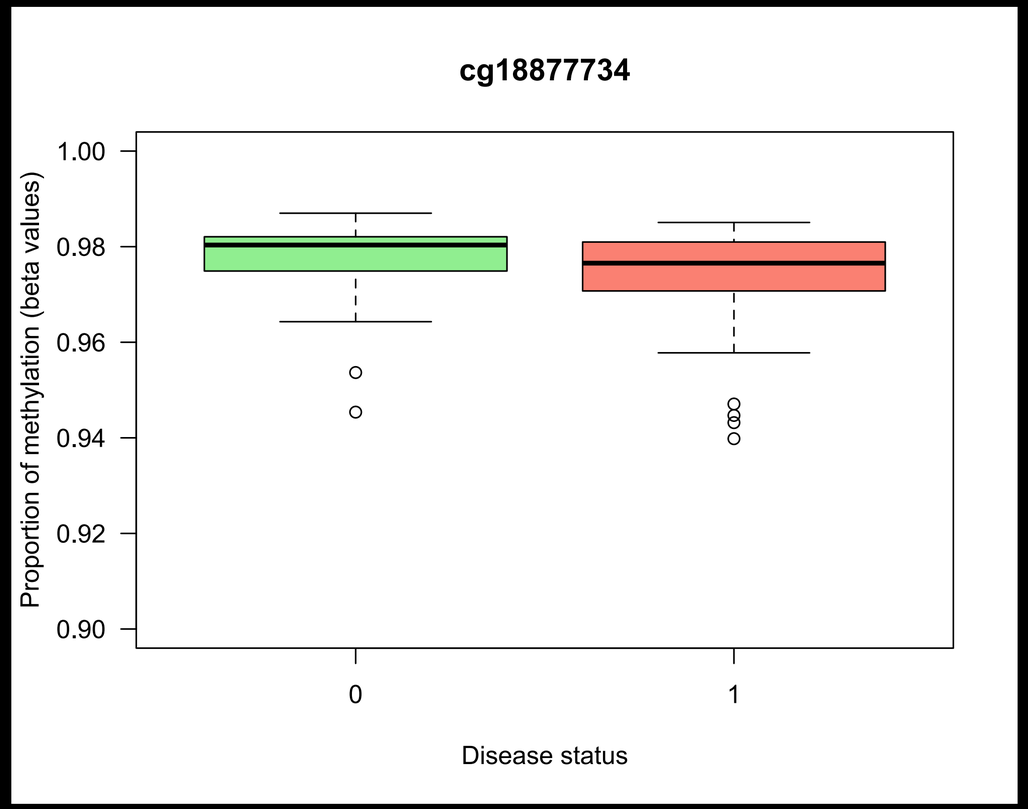
Use of linear regression again under an appropriate transformation of the outcome

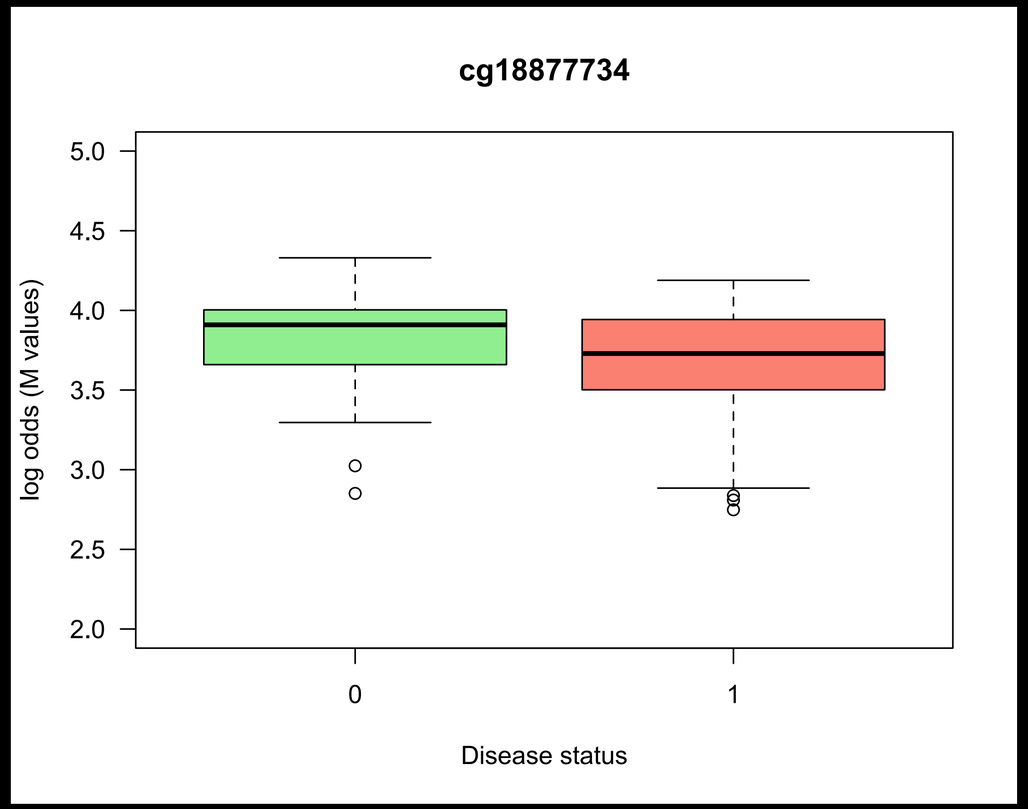
Y_{ij} = methylation levels of probe j in individual i

$$\underbrace{Y_{ij}}_{\text{Beta values}} \rightarrow \underbrace{Y^*_{ij}}_{\text{M values}} = \log \frac{Y_{ij}}{1 - Y_{ij}} \text{ or } \log_2 \frac{Y_{ij}}{1 - Y_{ij}}$$

What are the theoretical advantages of this approach?

Analysis based on M values





What are the theoretical advantages of this approach?

Analysis based on M values

Similar tests to the analysis of beta values

T test

Mann-Whitney test

Linear regression

Analysis based on M values

Du et al. BMC Bioinformatics 2010, 11:587
http://www.biomedcentral.com/1471-2105/11/587

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis

Pan Du[1,3*], Xiao Zhang[2], Chiang-Ching Huang[2], Nadereh Jafari[4], Warren A Kibbe[1,3], Lifang Hou[2,3], Simon M Lin[1,3*]

Abstract

Background: High-throughput profiling of DNA methylation status of CpG islands is crucial to understand the epigenetic regulation of genes. The microarray-based Infinium methylation assay by Illumina is one platform for low-cost high-throughput methylation profiling. Both Beta-value and M-value statistics have been used as metrics to measure methylation levels. However, there are no detailed studies of their relations and their strengths and limitations.

Results: We demonstrate that the relationship between the Beta-value and M-value methods is a Logit transformation, and show that the Beta-value method has severe heteroscedasticity for highly methylated or unmethylated CpG sites. In order to evaluate the performance of the Beta-value and M-value methods for identifying differentially methylated CpG sites, we designed a methylation titration experiment. The evaluation results show that the M-value method provides much better performance in terms of Detection Rate (DR) and True Positive Rate (TPR) for both highly methylated and unmethylated CpG sites. Imposing a minimum threshold of difference can improve the performance of the M-value method but not the Beta-value method. We also provide guidance for how to select the threshold of methylation differences.

Conclusions: The Beta-value has a more intuitive biological interpretation, but the M-value is more statistically valid for the differential analysis of methylation levels. Therefore, we recommend using the M-value method for conducting differential methylation analysis and including the Beta-value statistics when reporting the results to investigators.

M values can be positive and negative

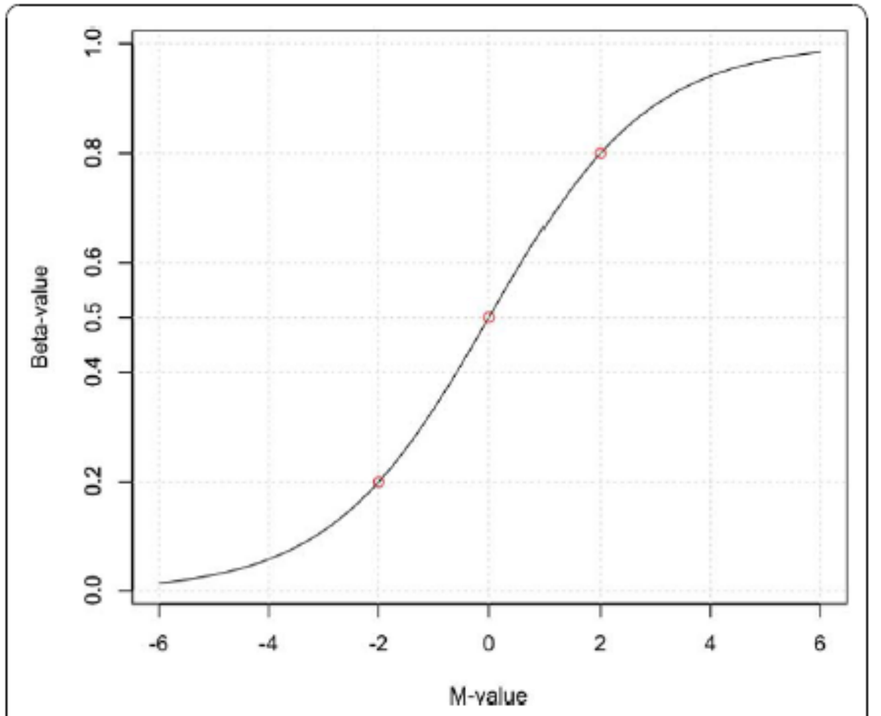


Figure 1 The relationship curve between M-value and Beta-value.

The range of M values is wider especially at the extremes of the Beta value scale

Exercise: data_ace_ace2.csv

Repeat previous analysis using M values adjusting the effect of study and gender.

Perform a residual analysis to validate the model for each probe.

Exercise: data_ace_ace2.csv

Repeat previous analysis of the methylation levels of cg18877734 using M values.

Is there any evidence that this probe is differentially methylated between patients and controls?

Analysis based on beta values: beta regression (more advanced)

Y_{ij} = methylation levels of probe j in individual i

$Y_{ij} \in (0,1)$ $\qquad$ $Y_{ij} \rightsquigarrow Beta(\alpha_{ij}, \beta_{ij})$

$$\mu_{ij} = \frac{\alpha_{ij}}{\alpha_{ij} + \beta_{ij}} \qquad \phi_{ij} = \alpha_{ij} + \beta_{ij}$$

Useful reparametrization

$$Y_{ij} \rightsquigarrow Beta(\mu_{ij}, \phi_{ij})$$

Beta distribution is very flexible

$$f_{X|\alpha,\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{Be(\alpha,\beta)} I_{(0,1)}(x)$$

Analysis based on beta values: beta regression (more advanced)

Beta regression

$$Y_{ij} \rightsquigarrow Beta(\mu_{ij}, \phi_{ij}) \rightarrow Y_{ij} \rightsquigarrow Beta(\mu_{ij}, \phi_{j})$$

$$g(\mu_{ij}) = \beta_{0j} + \beta_{1j} x_{group,i} + \sum_{k=2}^{p} \beta_{kj} x_{k,i}$$

where $g(\cdot)$ is a link function (e.g., logit function, $g(x) = \log \frac{x}{1-x}$)

Homework: data_ace_ace2.csv

Repeat previous analysis of the methylation levels of cg18877734 using beta regression where your outcome variable is beta values and Disease and Female are the covariates of the model.

Is there any evidence that this probe is differentially methylated between patients and controls?