Biostatistics

Applications in Genetic and Epigenetic data

Nuno Sepúlveda, 22.12.2025

Syllabus

1. General review
    a. Population/Sample/Sample size
    b. Type of Data – quantitative and qualitative variables
    c. Common probability distributions/popular tests
2. Applications in Medicine
    a. Construction and analysis of diagnostic tools – Binomial distribution, ROC curve, sensitivity, specificity, Rogal-Gladen estimator
    b. Estimation of treatment effects - generalized linear models
    c. Survival analysis - Kaplan-Meier curve, log-rank test, Cox's proportional hazards model
3. Applications in Genetic and Epigenetic Data
    a. Genetic association studies – Hardy-Weinberg test, homozygosity, minor allele frequencies, additive model, multiple testing correction
    b. Methylation association studies – M versus beta values
4. Applications in Serological Data Analysis
    a. Determination of seropositivity using Gaussian mixture models
    b. Reversible catalytic models for estimating seroconversion rate
    c. Sample size calculation for estimating seroconversion rate

Genotypic-phenotype mapping

Single nucleotide polymorphisms
Copy number variation
Inversion/deletion

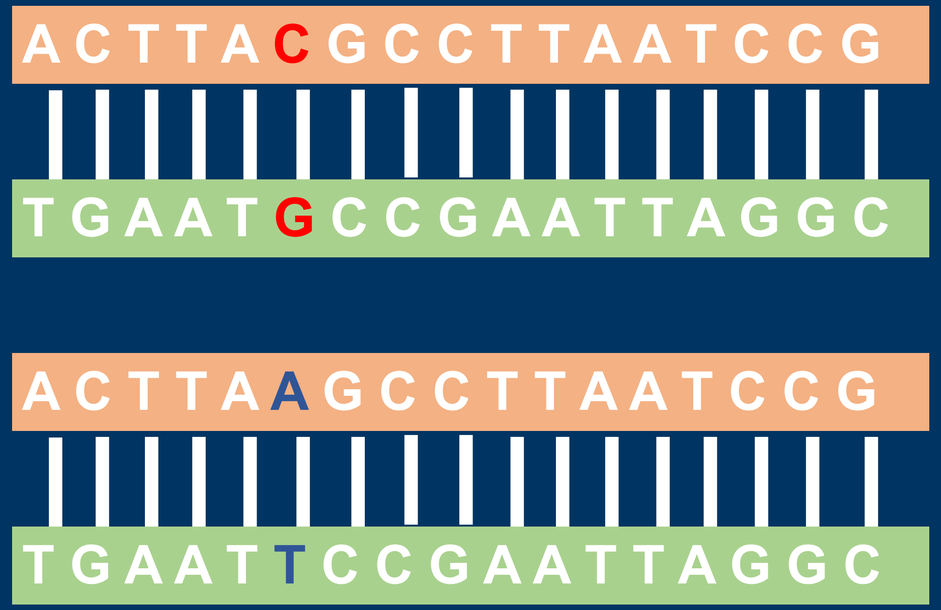
Post-translation modification
DNA methylation
microRNA

Different symptoms for the same disease

Genetic association studies

Genetic variation → Disease

Single nucleotide polymorphisms
Copy number variation
Inversion/deletion



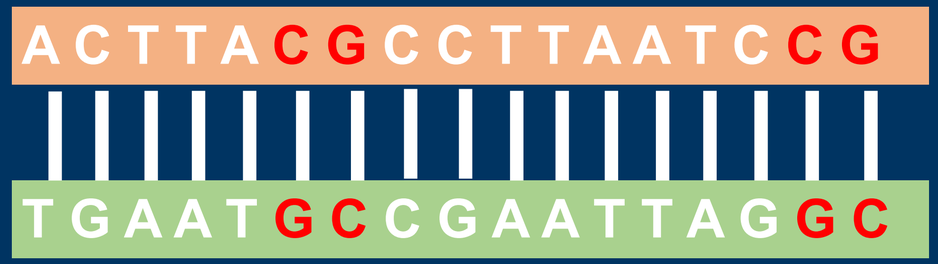
Genotypic-phenotype mapping

Single nucleotide polymorphisms
Copy number variation
Inversion/deletion

Post-translation modification
DNA methylation
microRNA

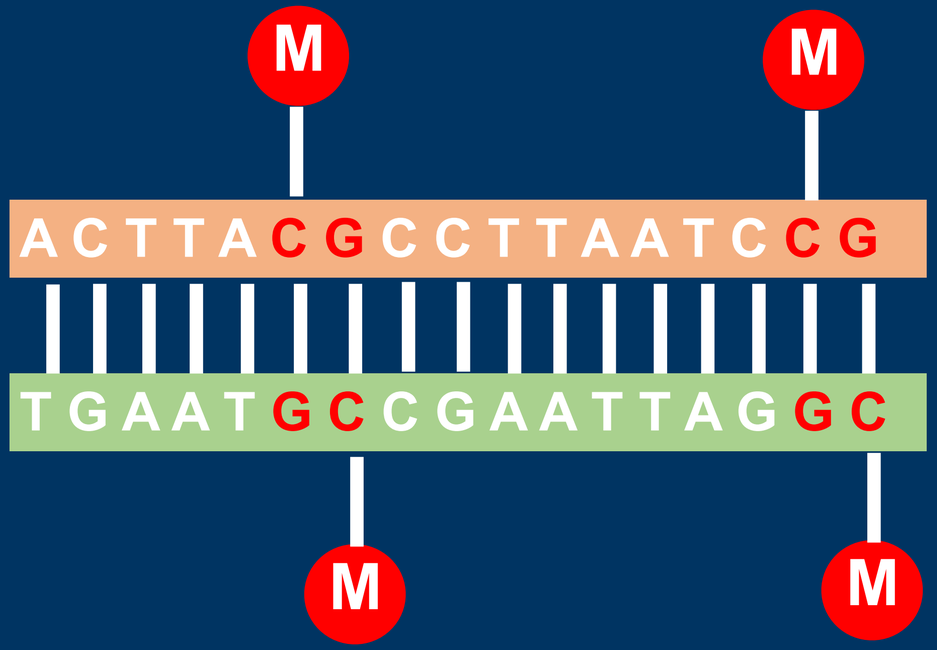
Different symptoms for the same disease

DNA methylation



Gene might be expressed

Production of the protein



Gene might not be expressed

No production of the protein

Epigenome-wide association studies

Aim:

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

Example: Case-control study

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

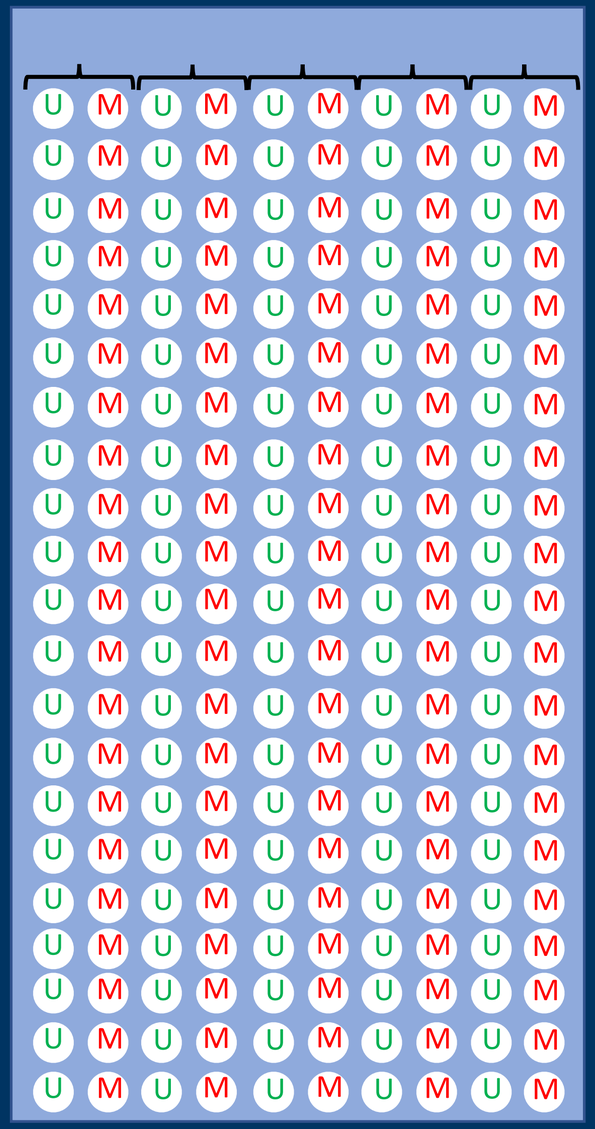
n=61
79% Women
Mean age of 41 years old
Mean BMI of 27 kg/m^2

Human Methylation 450K Array

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

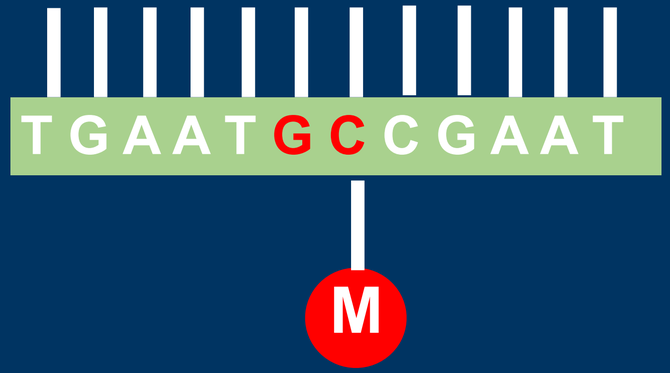
DNA methylation array

Array



Unmethylated probe

Green light



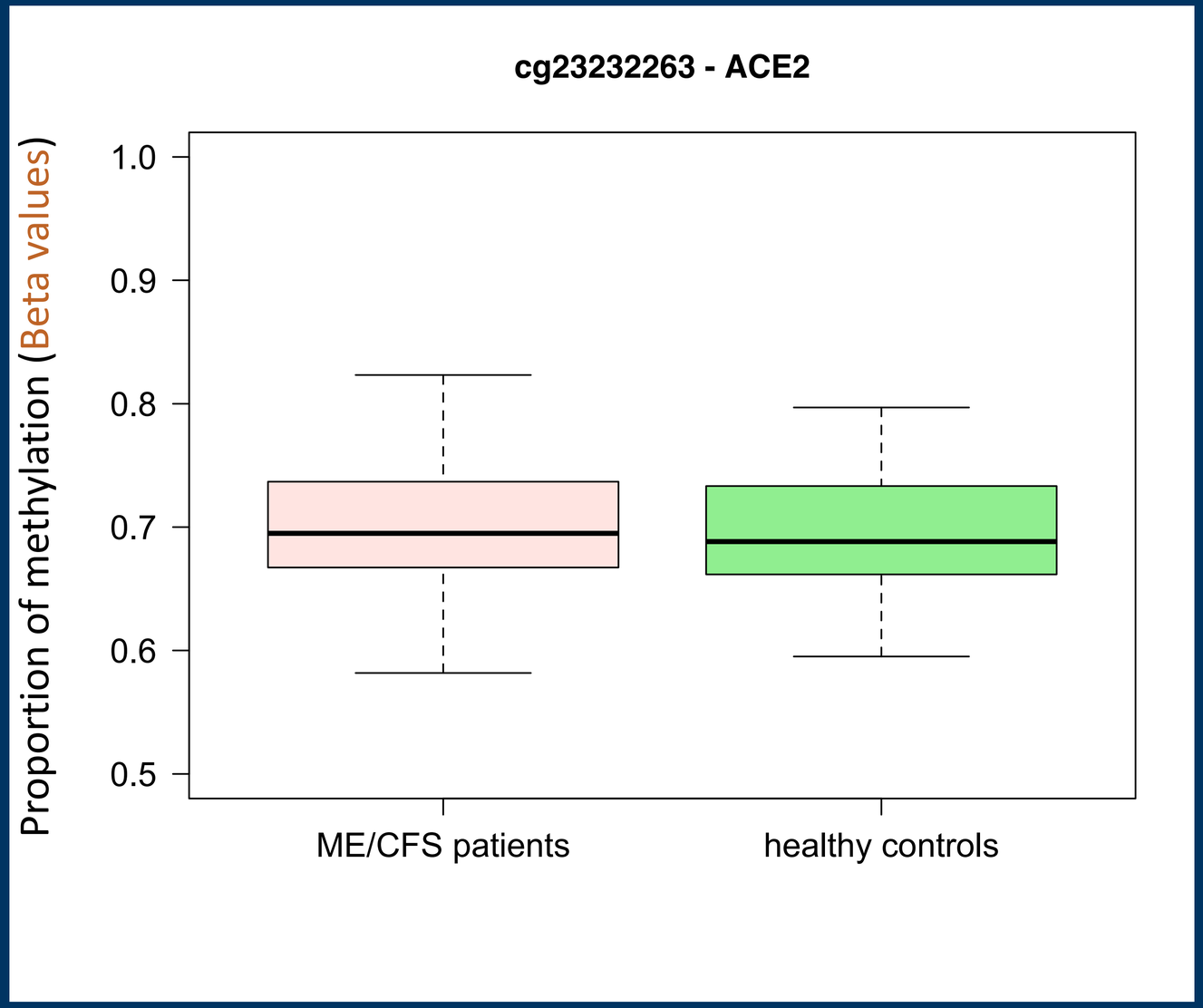
Methylated probe

Red light

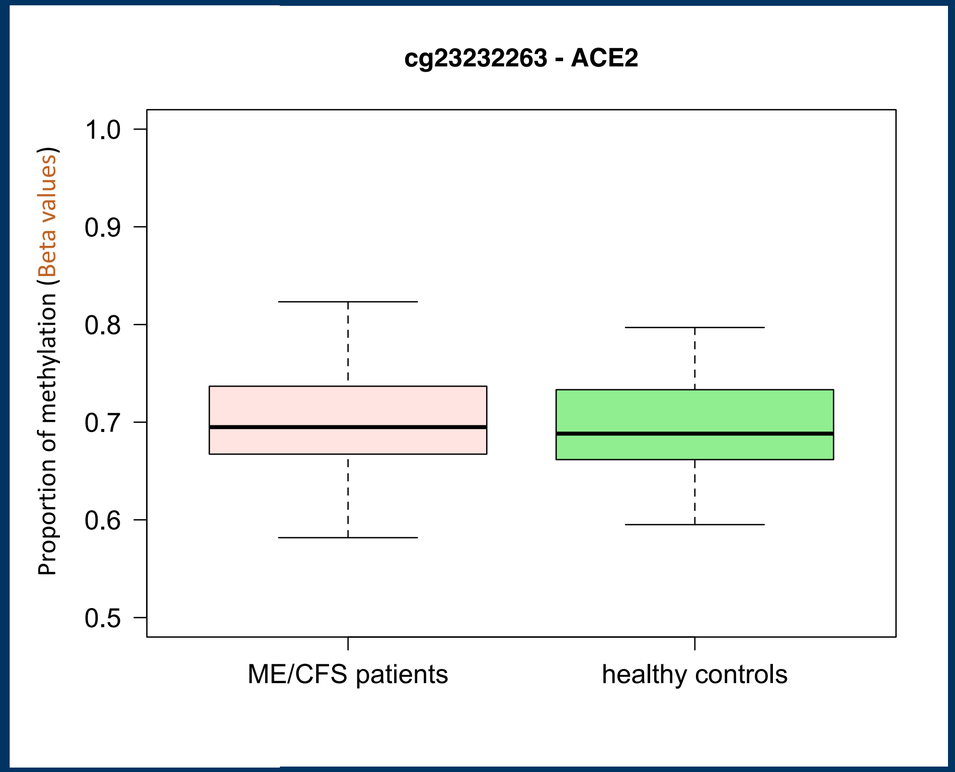
Beta values

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

Example of data for a single probe

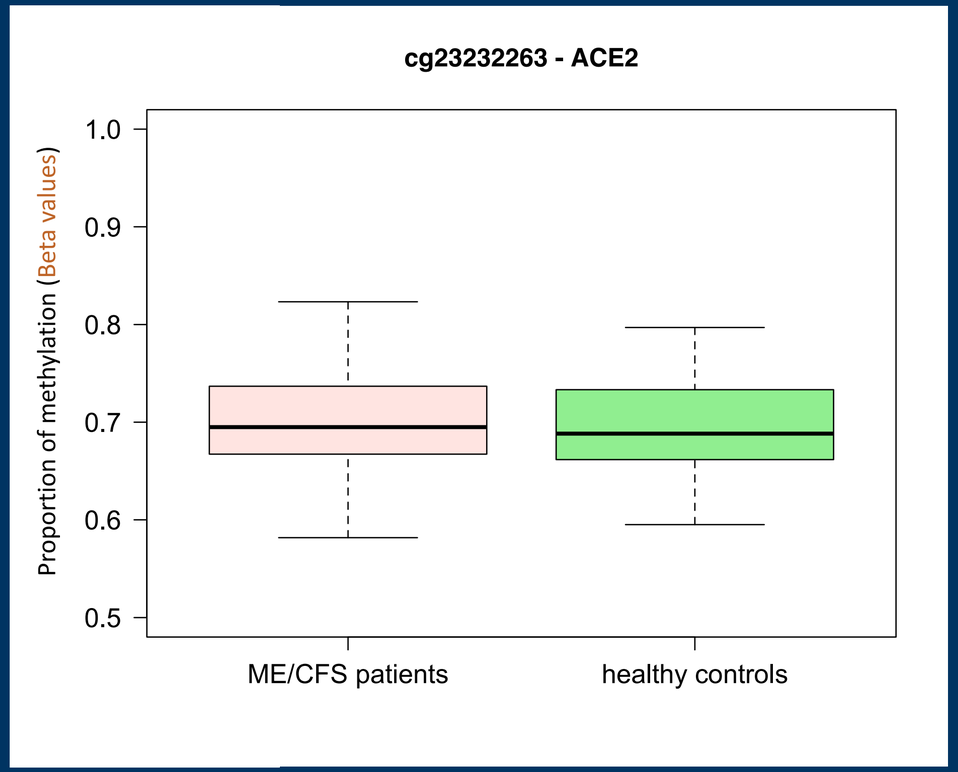


Simple statistical analysis of a single probe



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

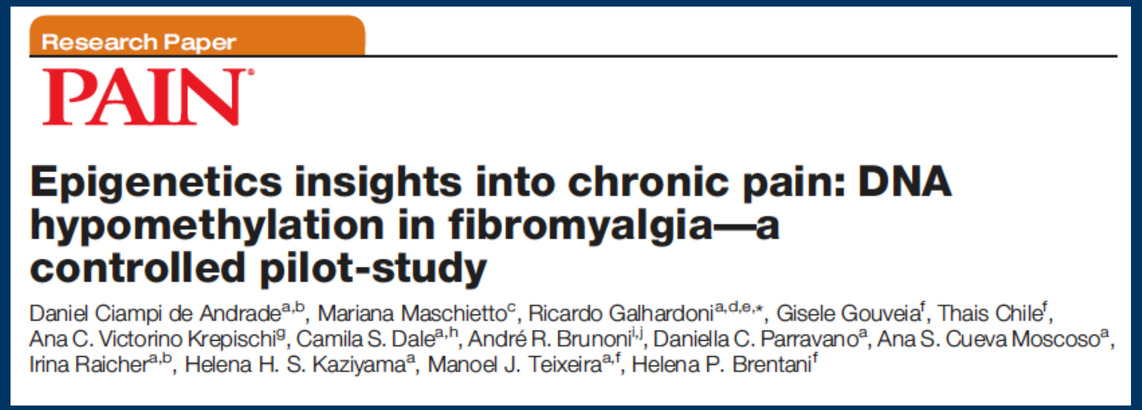
Simple statistical analysis of a single probe



T test using original or transformed data

Mann-Whitney test

Exercise: data_epigenetic_fm.csv



Research Paper

PAIN

Epigenetics insights into chronic pain: DNA hypomethylation in fibromyalgia—a controlled pilot-study

Daniel Ciampi de Andrade[a,b], Mariana Maschietto[c], Ricardo Galhardoni[a,d,e,*], Gisele Gouveia[f], Thais Chile[f], Ana C. Victorino Krepischi[g], Camila S. Dale[a,h], André R. Brunoni[i,j], Daniella C. Parravano[a], Ana S. Cueva Moscoso[a], Irina Raicher[a,b], Helena H. S. Kaziyama[a], Manoel J. Teixeira[a,f], Helena P. Brentani[f]

24 patients with fibromyalgia
24 healthy controls

Data repository
https://www.ncbi.nlm.nih.gov/geo/

Dataset
GSE85506 (IDAT files)

Ciampi de Andrade D, Maschietto M, Galhardoni R, et al. Epigenetics insights into chronic pain: DNA hypomethylation in fibromyalgia-a controlled pilot-study. Pain. 2017;158(8):1473-1480.

Exercise: data_epigenetic_fm.csv

Compare the methylation levels of 100 CpG probes between patients with fibromyalgia (cases) and healthy controls using T test and Mann-Whitney test.

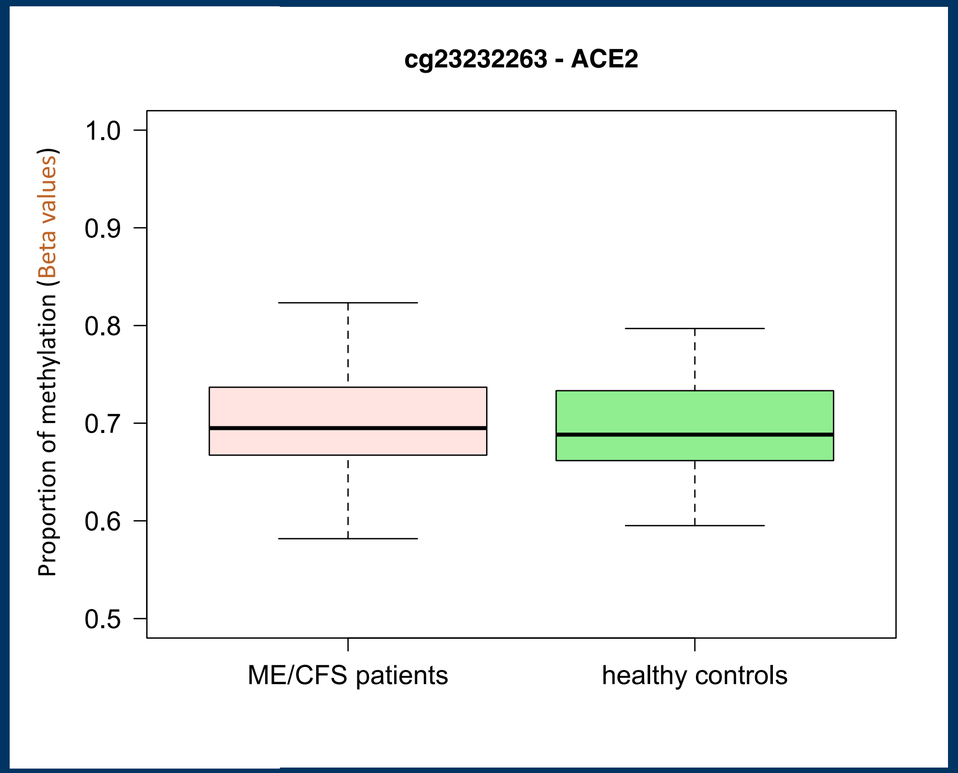
Which test is preferable to analyse data of each probe?

Which probes are differentially methylated when comparing patients and healthy controls after adjusting for multiple testing?

Are these probes hypo or hypermethylated in patients?

Statistical analysis of a single probe adjusting for covariates

Linear regression



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

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

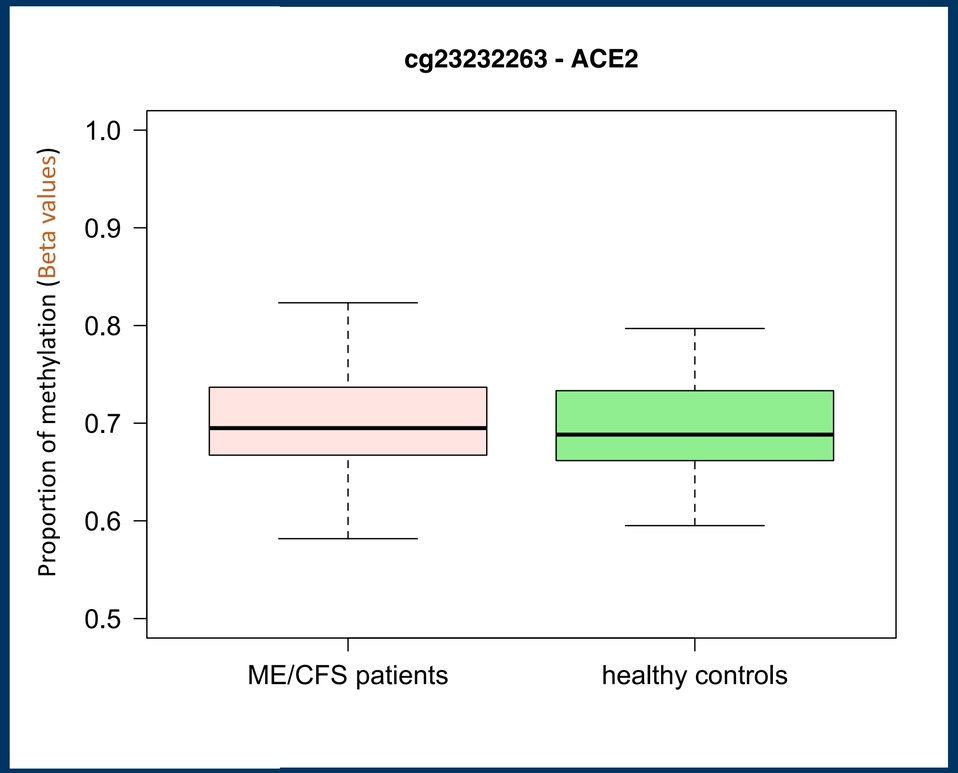
$x_{k,i}$ = value of covariate k for the individual i

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

$$\epsilon_{ij} \rightsquigarrow N\left(\mu = 0; \sigma = \sigma_0\right)$$

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

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

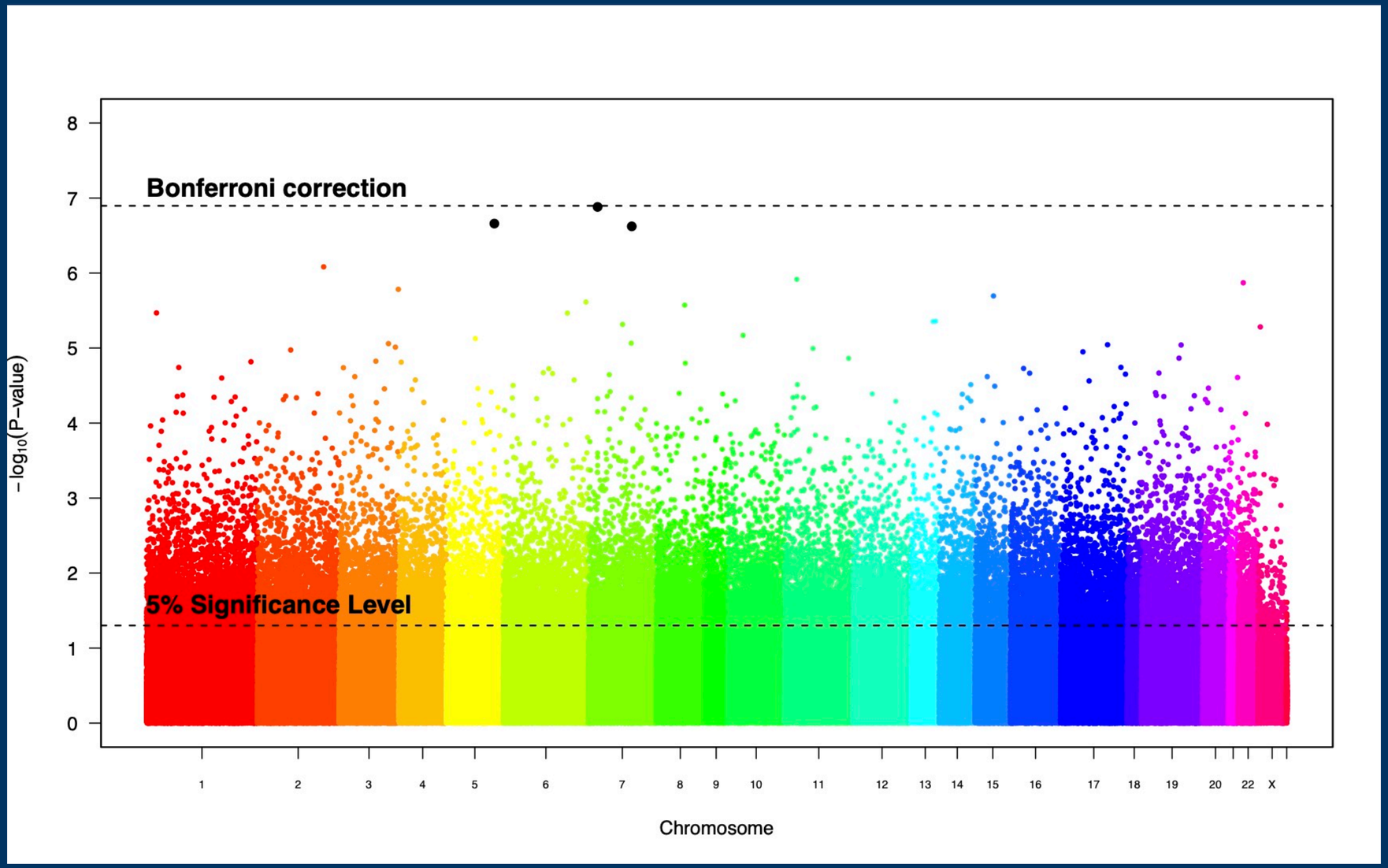
Wilks' likelihood ratio test

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

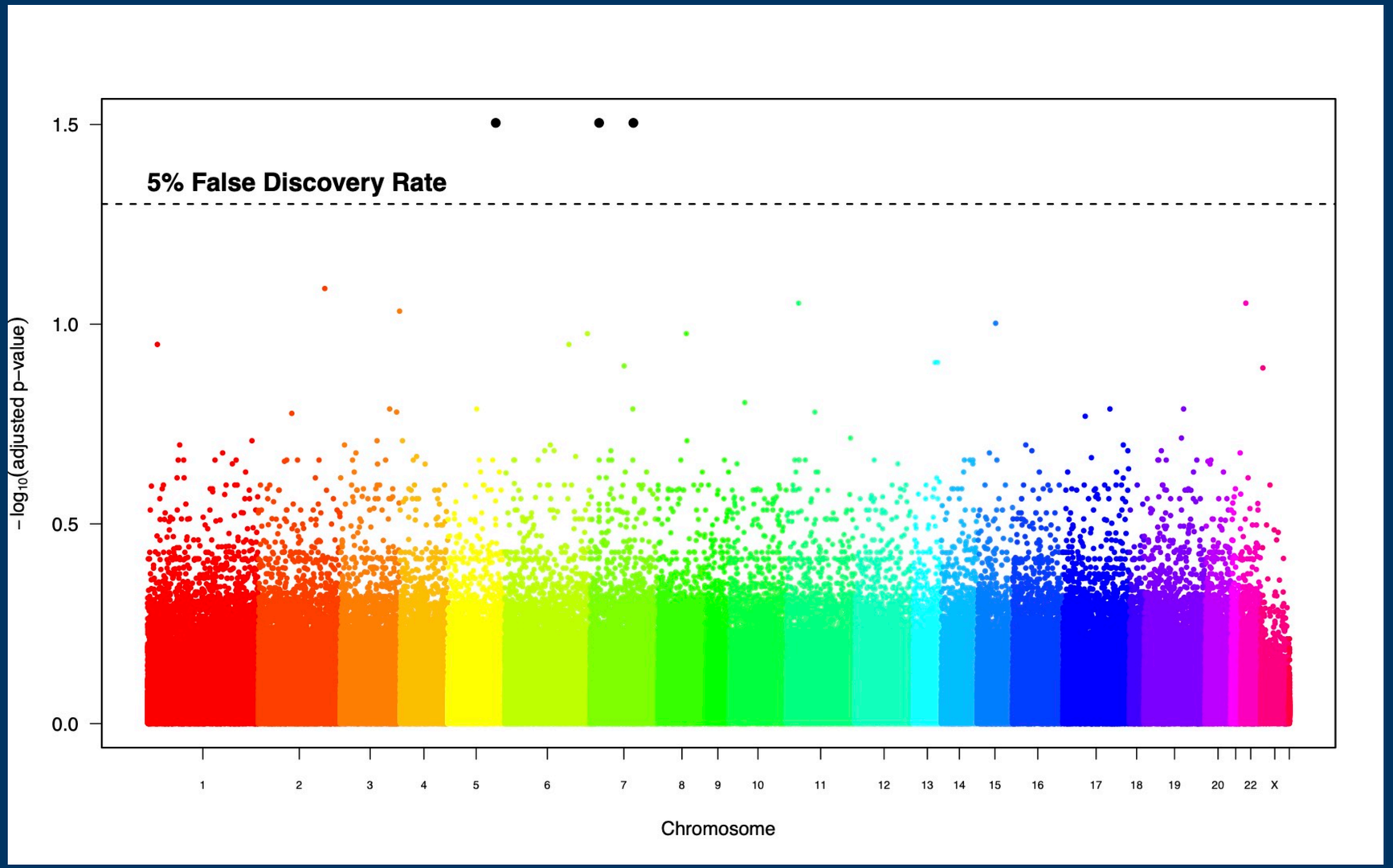
Construct a manhattan plot as learned for the genome-wide association studies

Report the significant probes and their location

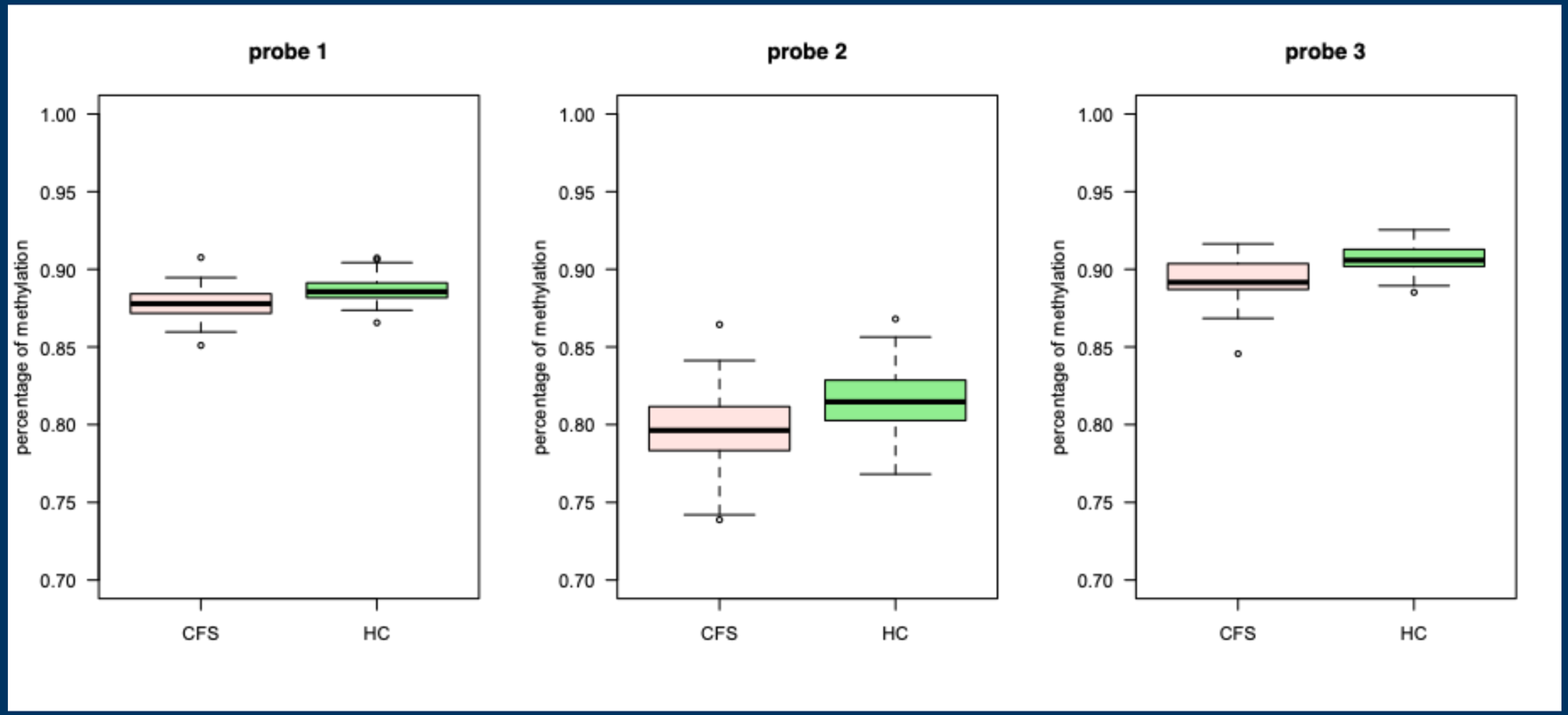
Manhattan plots adjusting the significance level via Bonferroni correction for multiple testing



Manhattan plots adjusting p-values via Benjamini-Hochberg procedure

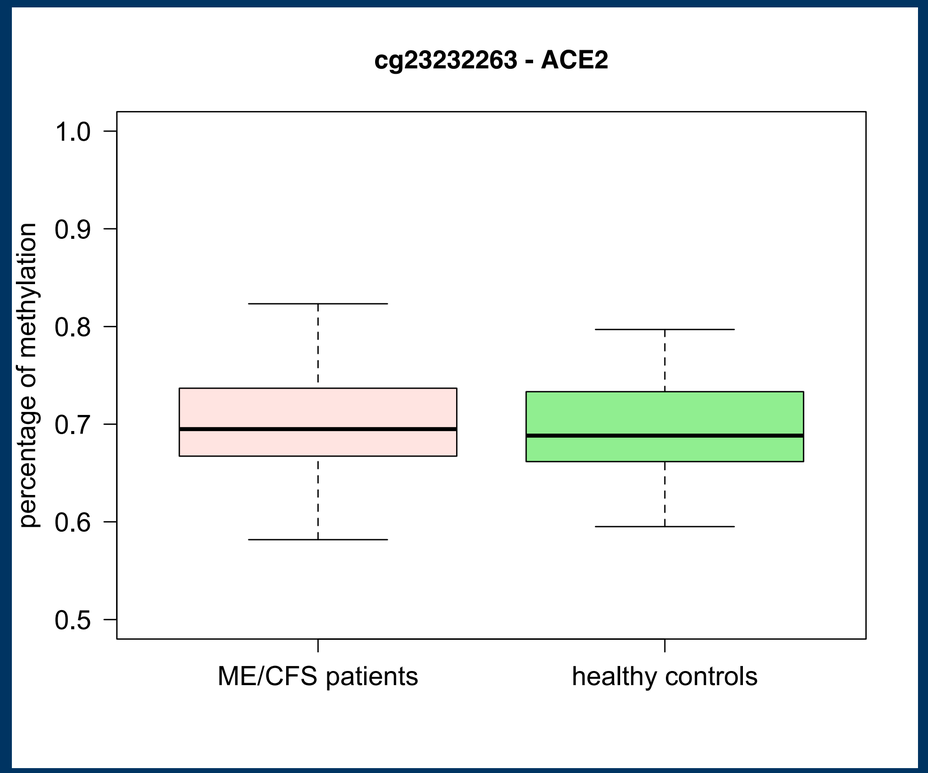


Decreased methylation in patients with ME/CFS



More expression for genes associated with these probes

Discussion



Is it reasonable to use linear regression in this kind of data?

What are the eventual problems?

Exercise: data_epigenetic_data.csv

Repeat previous analysis using linear regression. Include the effects of group and age. (Perform a residual analysis to validate the model for each probe.)

Analysing M values instead

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

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

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

$$Y^*_{ij} = \beta_{0j} + \beta_{1j} x_{group,i} + \sum_{k=2}^{p} \beta_{kj} x_{k,i} + \epsilon_{ij}$$

$$\epsilon_{ij} \rightsquigarrow N\left(\mu = 0; \sigma = \sigma_0\right)$$

What are the theoretical advantages of this approach?

Alternative statistical analysis of a single probe adjusting for covariates

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.

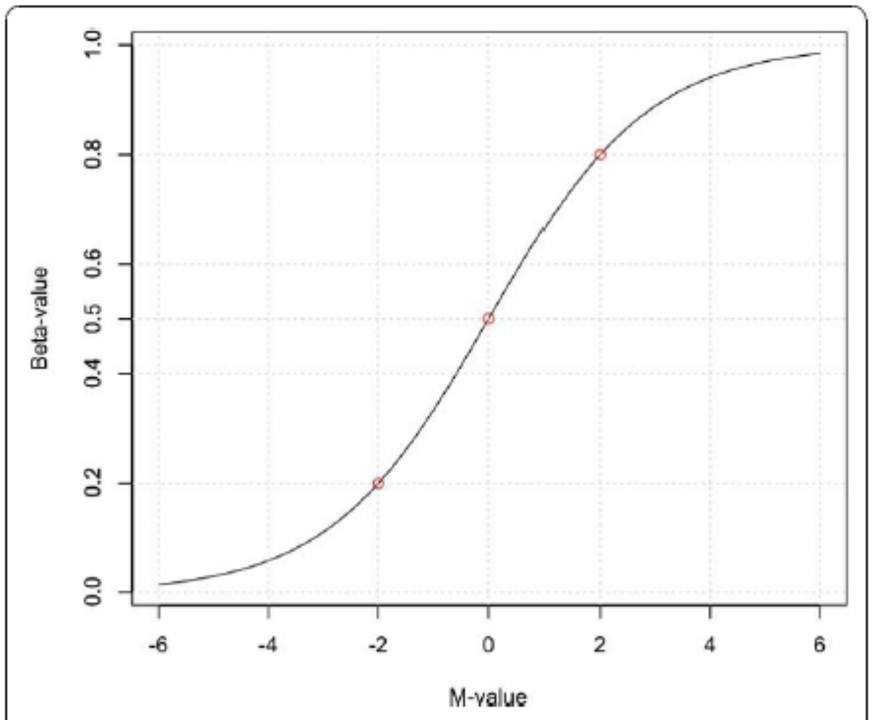


Figure 1 The relationship curve between M-value and Beta-value.

M values can be positive and negative

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

Alternative statistical analysis of a single probe adjusting for covariates

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.

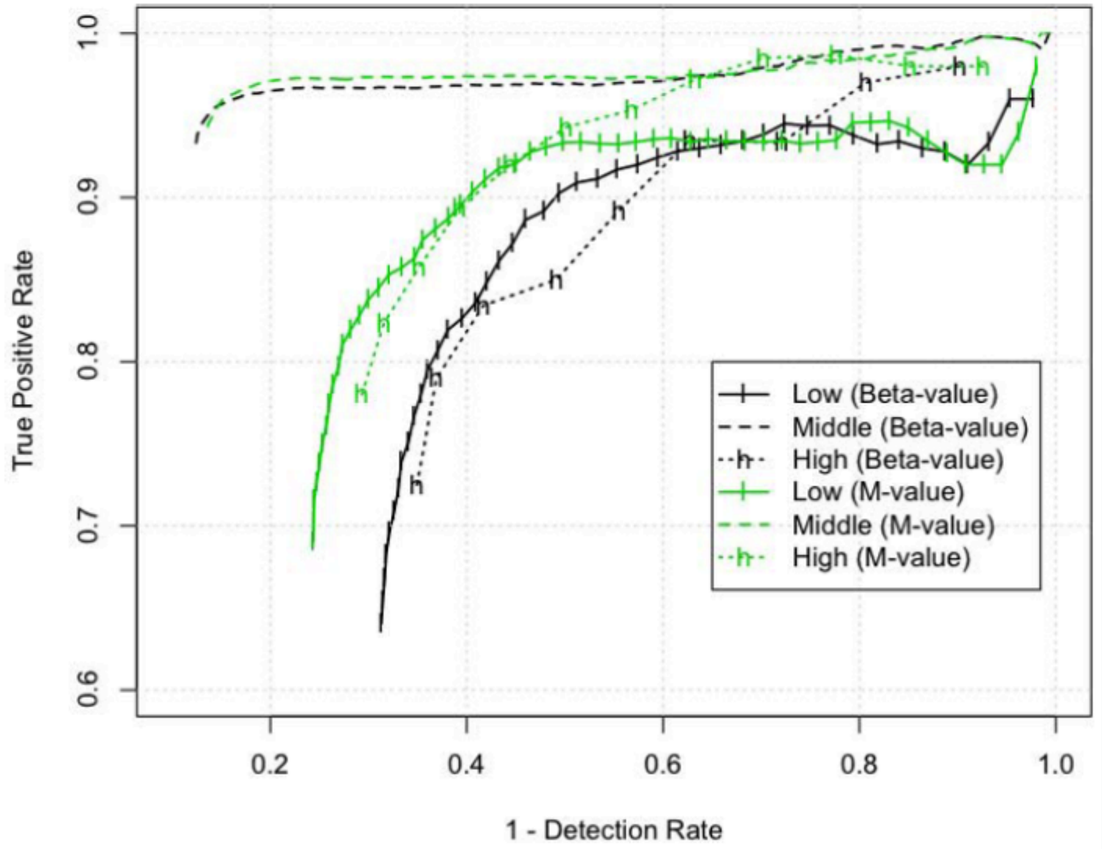


Figure 4 Performance comparisons of Beta- and M-value in the range of low, middle and high methylation levels based on the relationship of 1 - Detection Rate versus True Positive Rate.

Exercise: data_epigenetic_data.csv

Repeat previous analysis using linear regression on the M values of each probe. Include the effects of group and age in the model. (Perform a residual analysis to validate the model for each probe.)

Analysing Beta values with beta regression

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

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

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

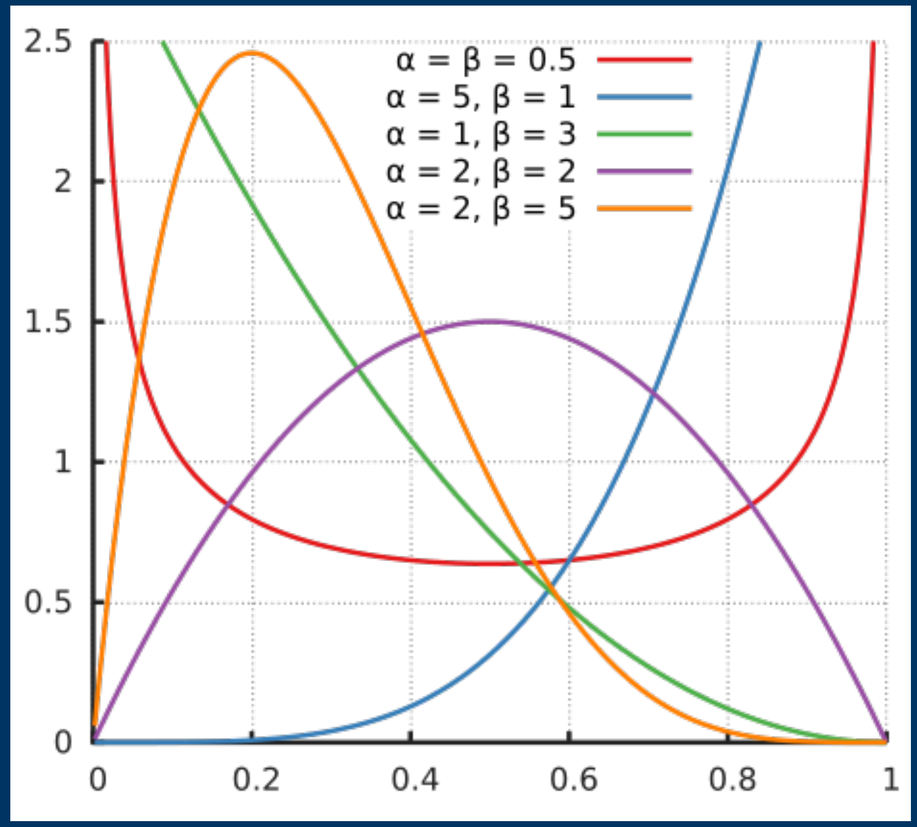
Useful reparametrization

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

Beta regression

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

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



Beta distribution is very flexible

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

Exercise: data_epigenetic_data.csv

Install and use the package “betareg”.

Repeat previous analysis on the Beta values using Beta regression. Include the effect of group and age.

Compare the results from all the analyses done so far.

What are your conclusions?