

# Biostatistics

Applications in Genetics, Genomics, and other 'omics data

Nuno Sepúlveda, 08.01.2024

# Syllabus

## 1. General review

- a. What is Biostatistics?
- b. Population/Sample/Sample size
- c. Type of Data – quantitative and qualitative variables
- d. Common probability distributions
- e. Work example – Malaria in Tanzania

## 2. Applications in Medicine

- a. Construction and analysis of diagnostic tools – Binomial distribution, sensitivity, specificity, ROC curve, 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 Genetics, Genomics, and other 'omics data

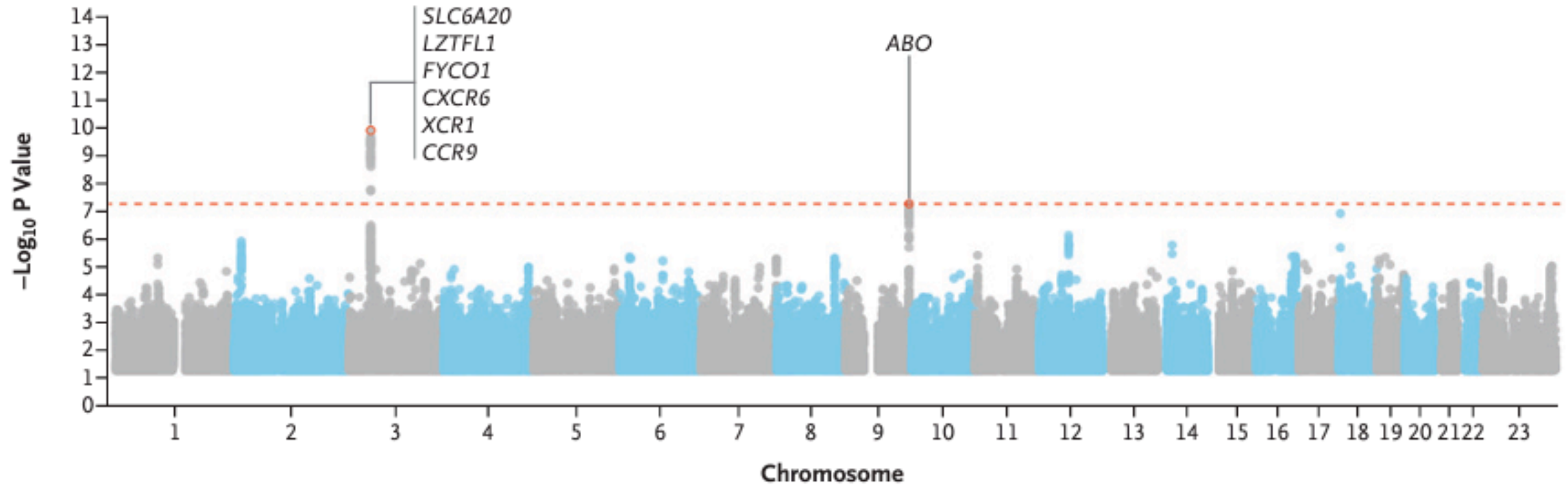
- a. Genetic association studies – Hardy-Weinberg test, homozygosity, minor allele frequencies, additive model, **multiple testing correction**
- b. **Methylation association studies – M versus beta values, estimation of biological age**
- c. Gene expression studies based on RNA-seq experiments – Tests based on Poisson and Negative-Binomial

## 4. Other Topics

- a. Estimation of Species diversity – Diversity indexes, Poisson mixture models
- b. Serological analysis – Gaussian (skew-normal) mixture models
- c. Advanced sample size and power calculations

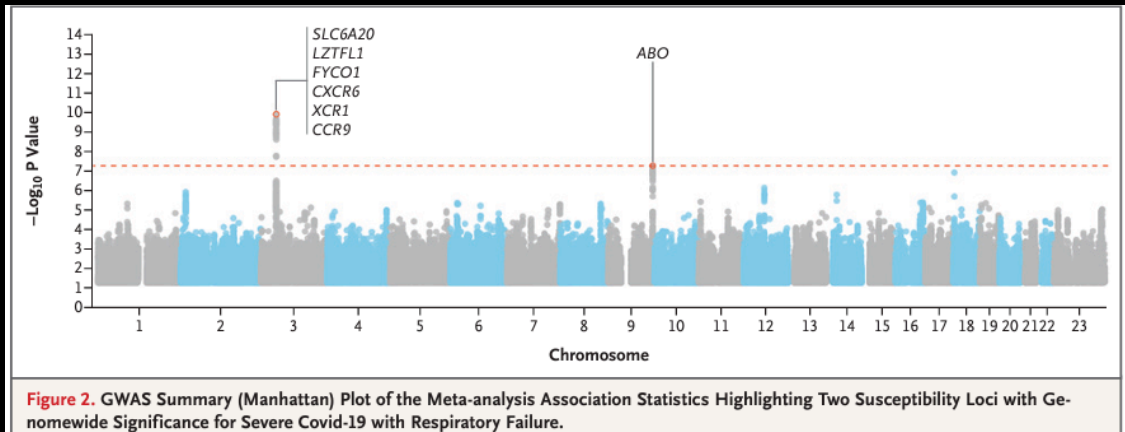
# Main outputs - GWAS

Do you know how this plot is called?

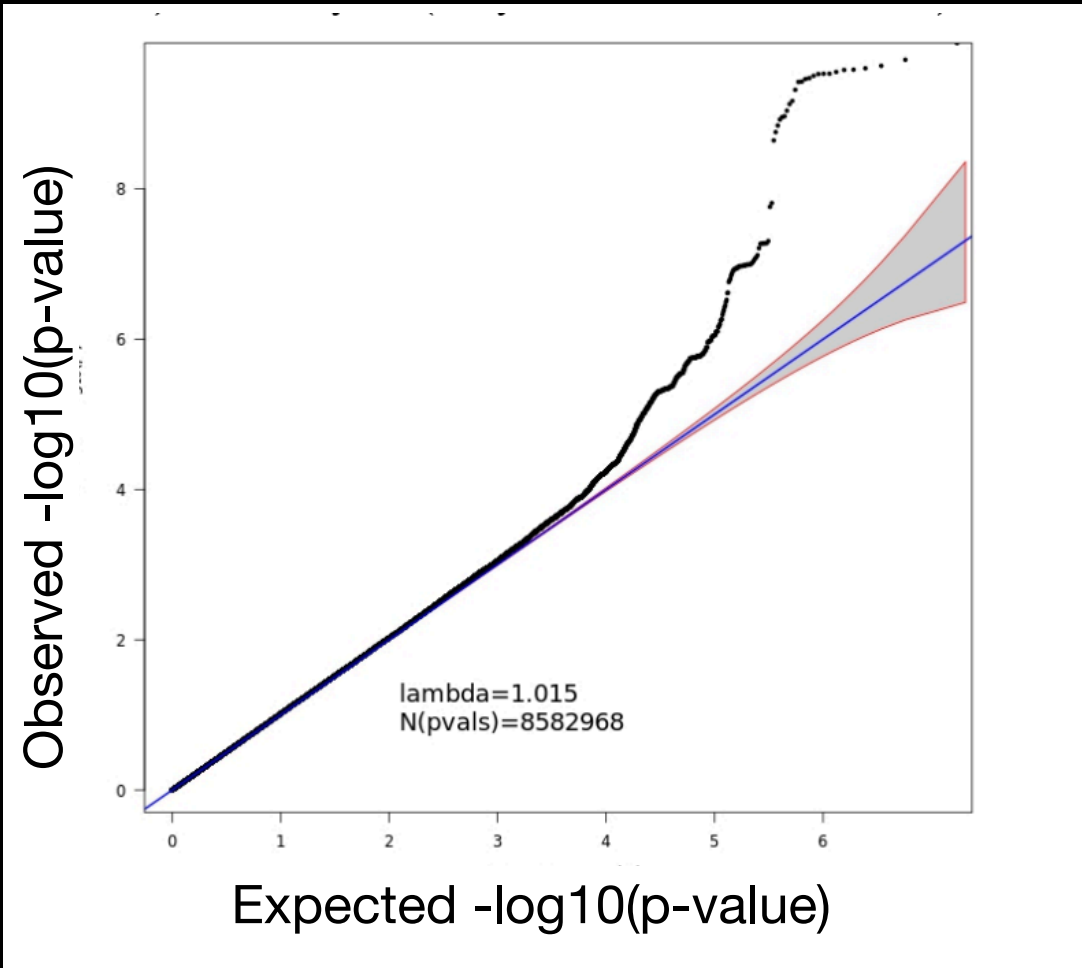


**Figure 2.** GWAS Summary (Manhattan) Plot of the Meta-analysis Association Statistics Highlighting Two Susceptibility Loci with Genomewide Significance for Severe Covid-19 with Respiratory Failure.

# Manhattan plot



# Main outputs - GWAS



## Practical - data\_lecture\_11\_gwas.csv

Let's recreate the typical visual outputs from a GWAS on COVID-19 in the Portuguese population

## Controlling the proportion of true positive among significant results

	True Positive	False Negative	Total
Significant results	m11	m12	m1
Non-significant results	m21	m22	m2
Total	m_1	m_2	m

**False discovery rate  
(Benjamini-Hochberg (BH) procedure)**

$$E [\text{True positives} \mid H_0 \text{ rejected}] = \alpha^*$$

$$E \left[ \frac{m_{11}}{m_{11} + m_{21}} \right] = \frac{q}{\alpha} = \alpha^* \qquad \alpha^* = 0.05$$



## False discovery rate (Benjamini-Hochberg procedure)

Algorithm (under the assumption of independent tests)

1. Order all the p-values by increasing order,  $p_{(1)}, \dots, p_{(n)}$
2. For  $\alpha^*$ , find  $k$  such as  $p_{(k)} \leq \frac{k}{m}\alpha^*$
3. Reject the null hypothesis (i.e., declare discoveries) for all the genetic markers associated with p-values less  $p_{(k)}$

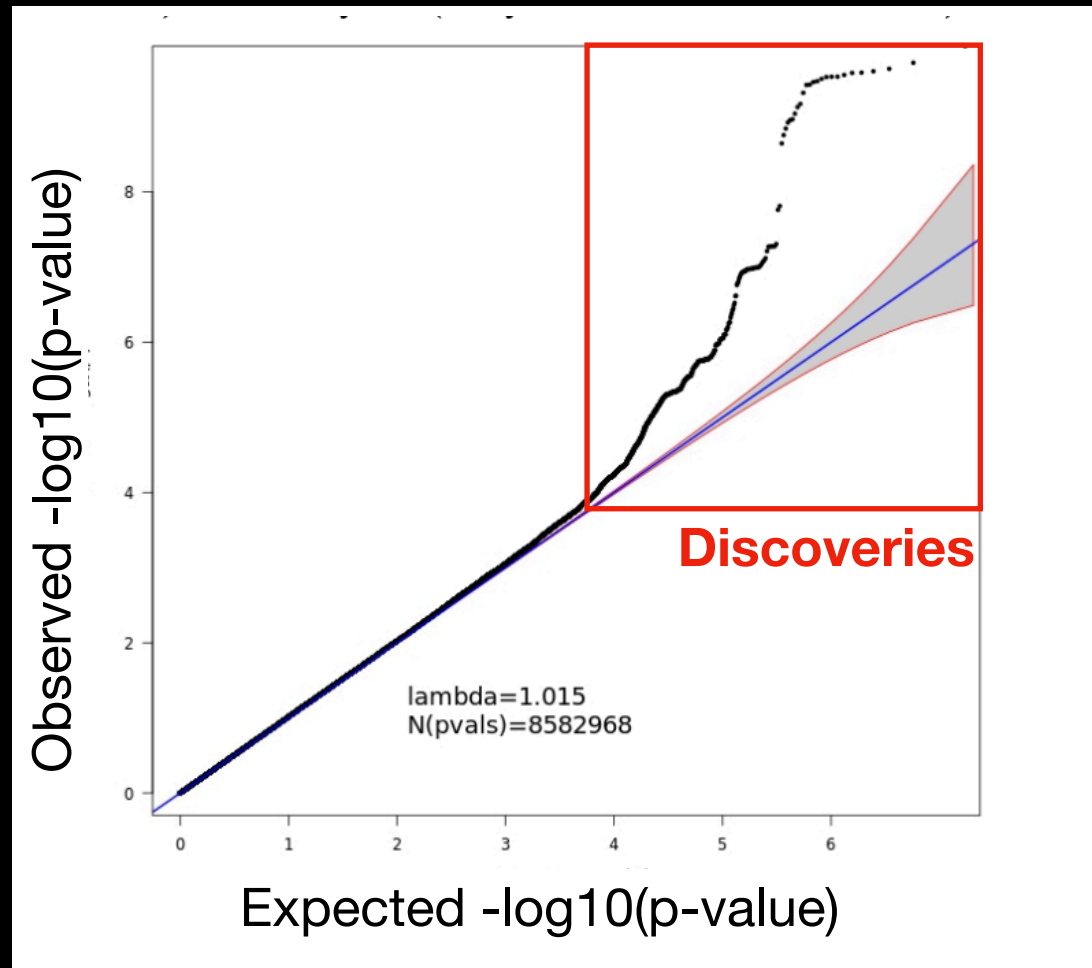
where

$p_{(1)}$  is the minimum p-value (with rank 1)

$p_{(k)}$  is the p-value with rank k

$p_{(n)}$  is the maximum p-value (with rank n)

# Visual interpretation



## Adjusting p-values

1. Order all the p-values by increasing order
2. Assign the ranking or position to each p-value
3. Calculate the adjusted by  $p_{(i)}^{adj} = \min \left\{ 1, \min_{j \geq i} \frac{mp_{(j)}}{j} \right\}$

where

$p_{(i)}^{adj}$  is the adjusted p-value with rank  $i$

$p_{(j)}$  is the p-value with rank  $j$

$m$  is the number of tests

Reject null hypothesis of tests whose the adjusted p-values are below the FDR

## **Useful variants of Benjamini-Hochberg**

### **Benjamini-Yekutieli (BY) procedure for dependent tests**

Adequate when analysing data from genetic markers in the same genetic locus

### **Benjamini-Krieger-Yekutieli (BKY) (improved) procedure for independent tests**

Package mutoss has implementations of these and other procedures

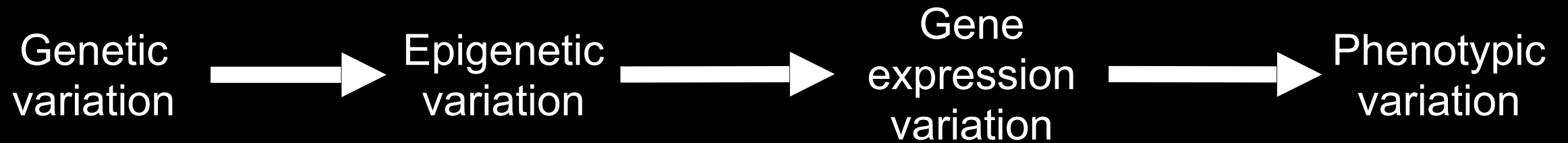
Package MASS has implementations of the BH and BY procedures

## Exercise - data\_lecture\_11\_gwas.csv

Apply Benjamini-Hochberg and Benjamini-Krieger-Yekutieli procedures to p-values from genetic markers in chromosome 10

How many genetic markers are statistically significant according to these procedures?

# 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

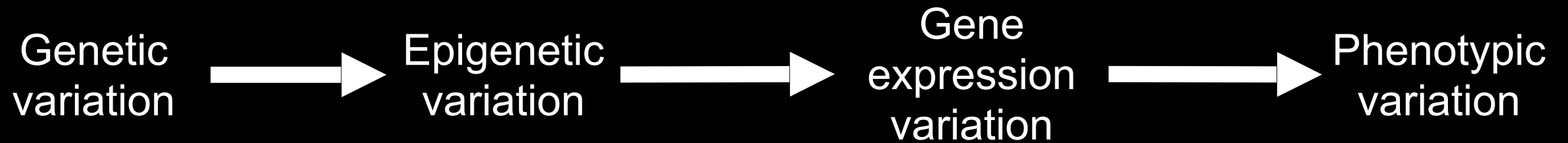
A C T T A **C** G C C T T A A T C C G

T G A A T **G** C C G A A T T A G G C

A C T T A **A** G C C T T A A T C C G

T G A A T **T** C C G A A T T A G G C

# Genotypic-phenotype mapping



Single nucleotide  
polymorphisms  
Copy number variation  
Inversion/deletion

Post-translation modification  
**DNA methylation**  
microRNA

Different  
symptoms for the  
same disease

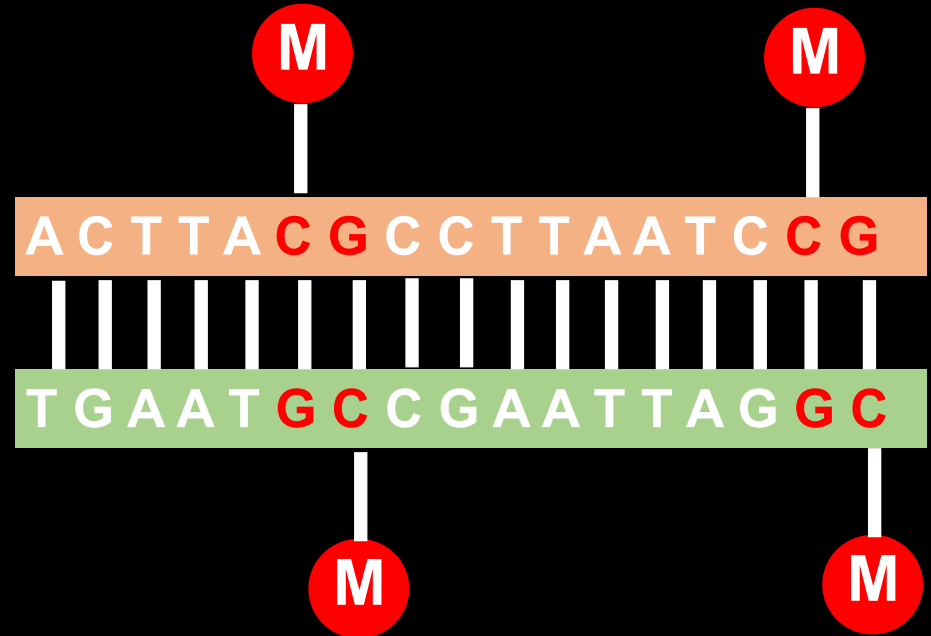


# DNA methylation



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<sup>2</sup>

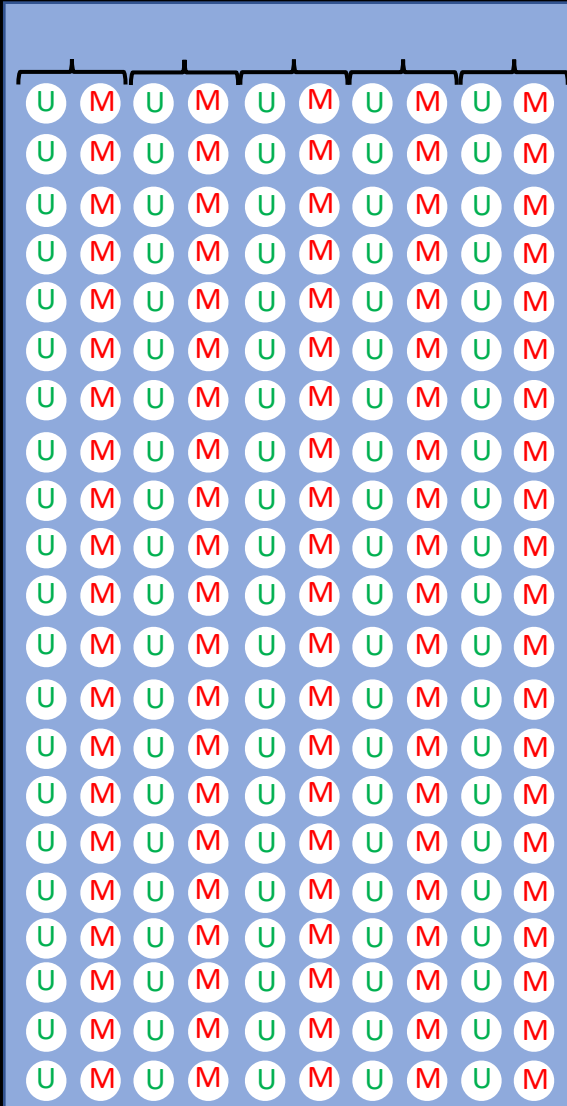


n=61  
79% Women  
Mean age of 41 years old  
Mean BMI of 27 kg/m<sup>2</sup>

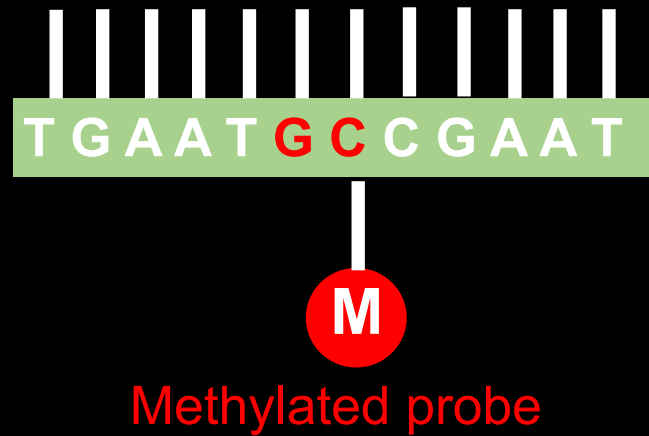
Human Methylation 450K Array

# DNA methylation array

Array



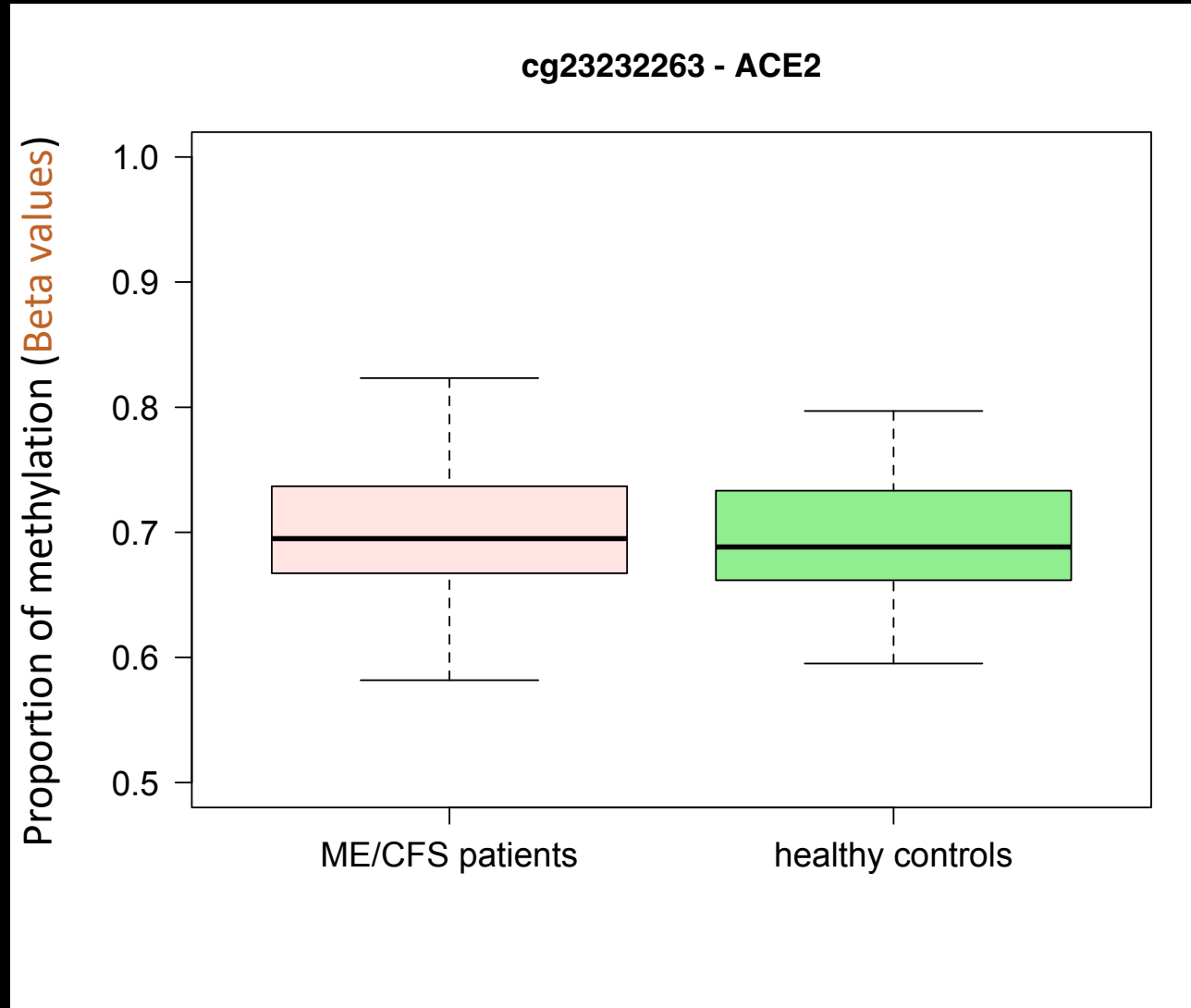
Green  
light



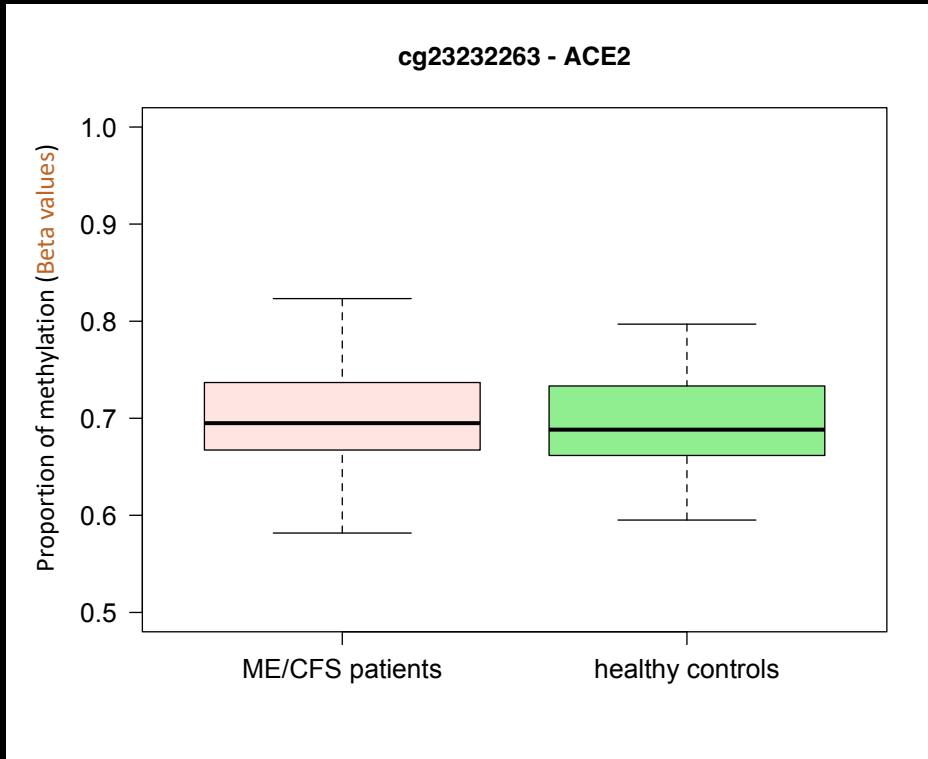
Red  
light

$$\frac{\text{Intensity of red light}}{(\text{Intensity of red light} + \text{Intensity of green light})}$$

# Example of data for a single probe

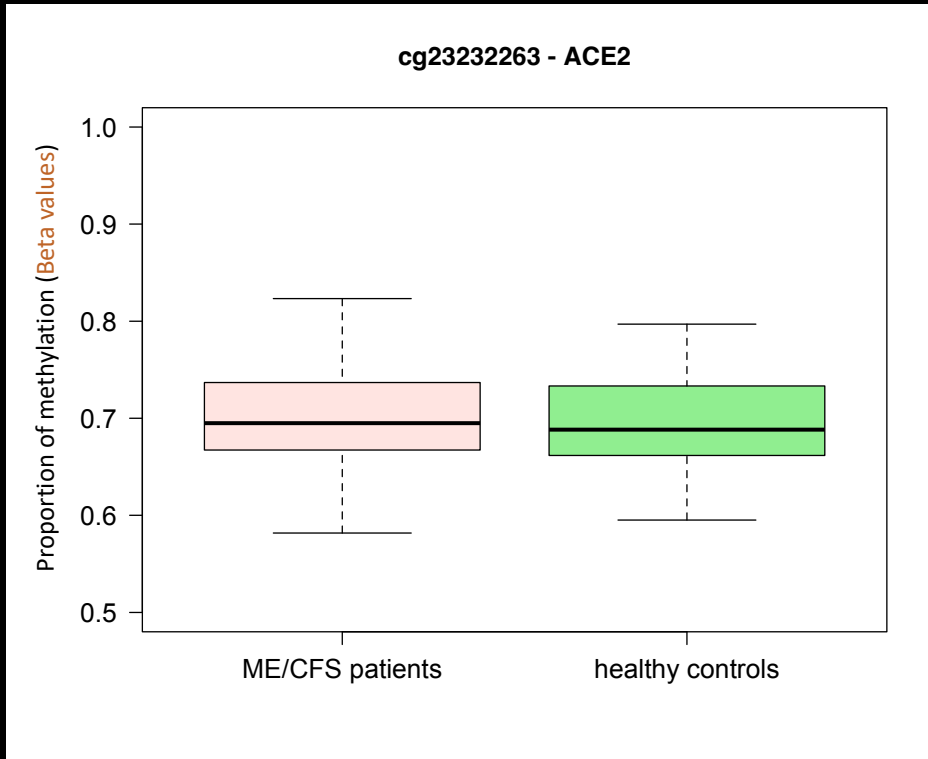


# Simple statistical analysis of a single probe



Which statistical tests can we use to check whether patients differ from healthy 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\_lecture\_12\_ACE\_ACE2.csv

Heliyon 7 (2021) e07665



Contents lists available at [ScienceDirect](#)

**Heliyon**

journal homepage: [www.cell.com/heliyon](http://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<sup>a,b</sup>, Franziska Sotzny<sup>c,\*\*</sup>, Sandra Bauer<sup>c</sup>, Helma Freitag<sup>c</sup>, André Fonseca<sup>d</sup>, Anna D. Grabowska<sup>e</sup>, Luís Graça<sup>a</sup>, Clara Cordeiro<sup>b,d</sup>, Luís Nacul<sup>f,g</sup>, Eliana M. Lacerda<sup>f</sup>, Jesus Castro-Marrero<sup>h</sup>, Carmen Scheibenbogen<sup>c</sup>, Francisco Westermeier<sup>i,j</sup>, Nuno Sepúlveda<sup>b,c,k,\*</sup>

<sup>a</sup> Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisbon, Portugal

<sup>b</sup> CEAUL – Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal

<sup>c</sup> Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin, Humboldt Universität zu Berlin and Berlin Institute of Health, Institute of Medical Immunology, Berlin, Germany

<sup>d</sup> Faculdade de Ciências e Tecnologia, Universidade do Algarve, Faro, Portugal

<sup>e</sup> Department of Biophysics, Physiology, and Pathophysiology, Medical University of Warsaw, Warsaw, Poland

<sup>f</sup> Department of Clinical Research, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom

<sup>g</sup> Complex Chronic Diseases Program, British Columbia Women's Hospital and Health Centre, Vancouver, British Columbia, Canada

<sup>h</sup> Vall d'Hebron Hospital Research Institute, Division of Rheumatology, ME/CFS Unit, Universitat Autònoma de Barcelona, Barcelona, Spain

<sup>i</sup> Institute of Biomedical Science, Department of Health Studies, FH Joanneum University of Applied Sciences, Graz, Austria

<sup>j</sup> Centro Integrativo de Biología y Química Aplicada (CIBQA), Universidad Bernardo O'Higgins, Santiago, Chile

<sup>k</sup> Department of Infection Biology, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom



# Exercise: Data\_lecture\_12\_ACE\_ACE2.csv

Information about the Probes: ACE\_ACE2\_merged\_annotation\_heliyon.csv

Compare the methylation levels of 27 CpG probes located in ACE and ACE2 genes between patients with chronic fatigue syndrome and healthy controls using T test and Mann-Whitney test.

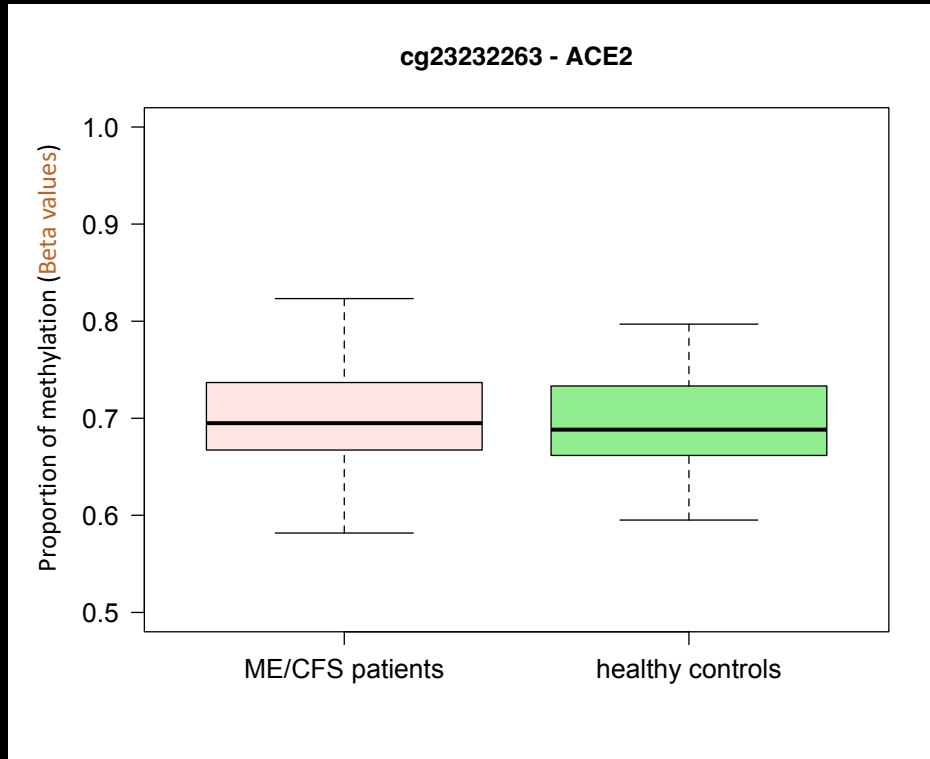
Which test is preferable to analyse data of each probe?

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 (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  $i$

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

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

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

# Epigenome-wide association studies

Perform this test on data of each probe

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

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

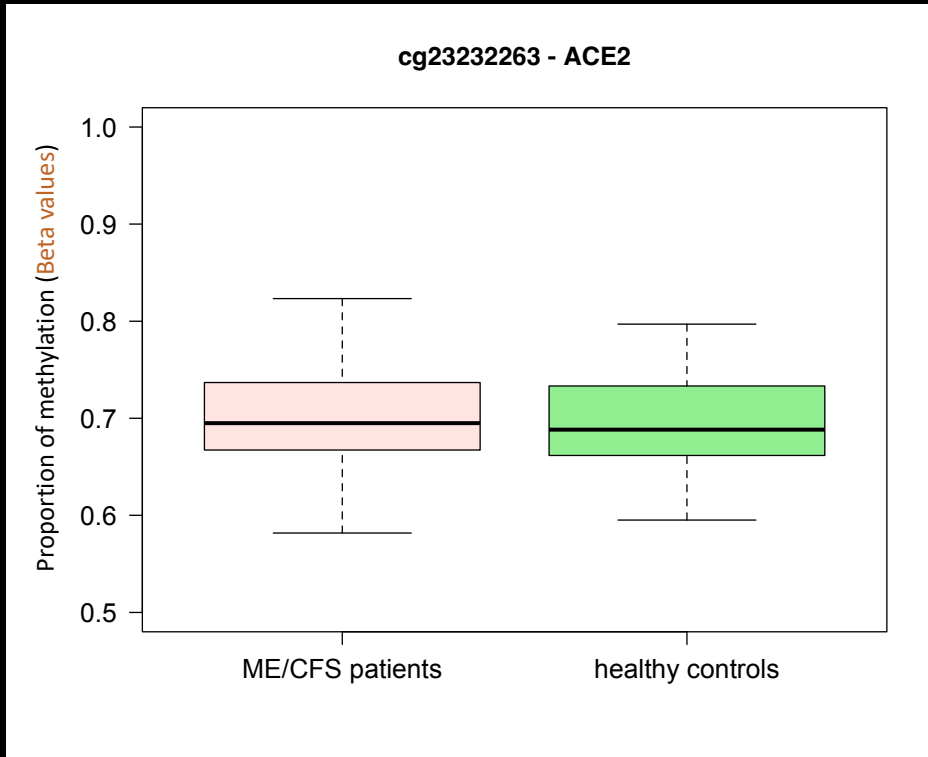
Wald's score test

Wilks' likelihood ratio test

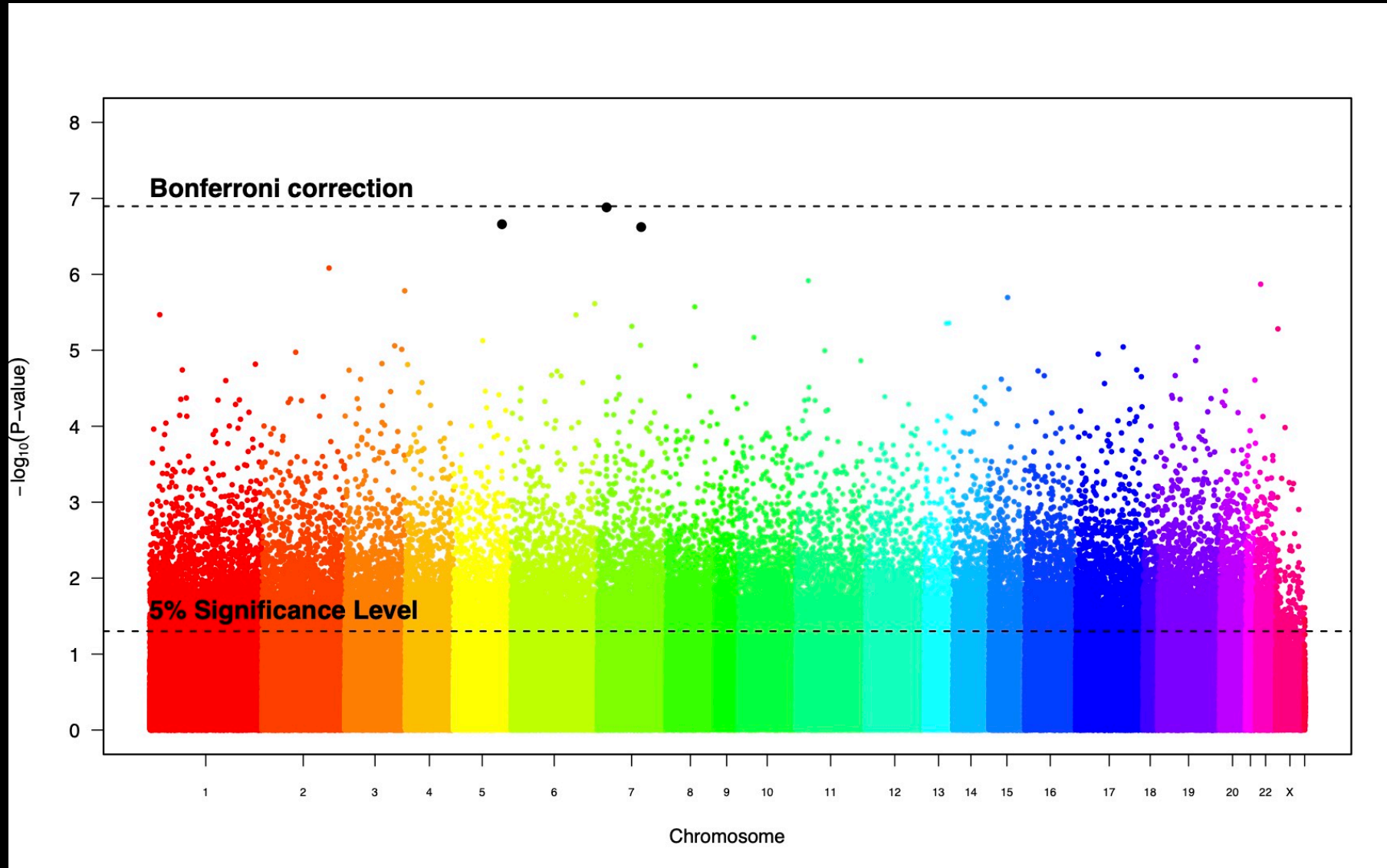
Correct the p-values of each individual test by a procedure controlling the false discovery rate (e.g., Benjamini-Hochberg procedure)

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

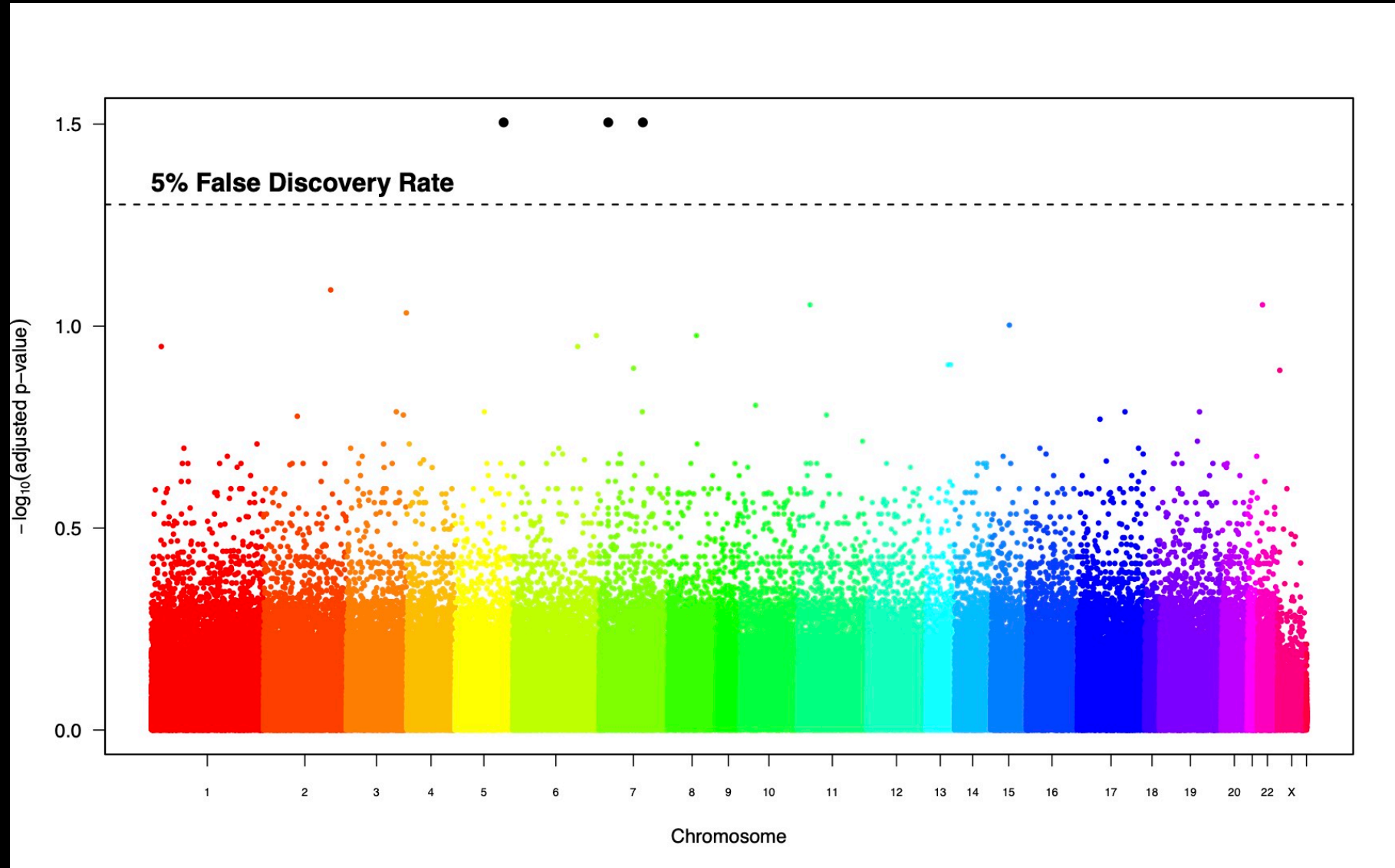
Report the significant probes and their location



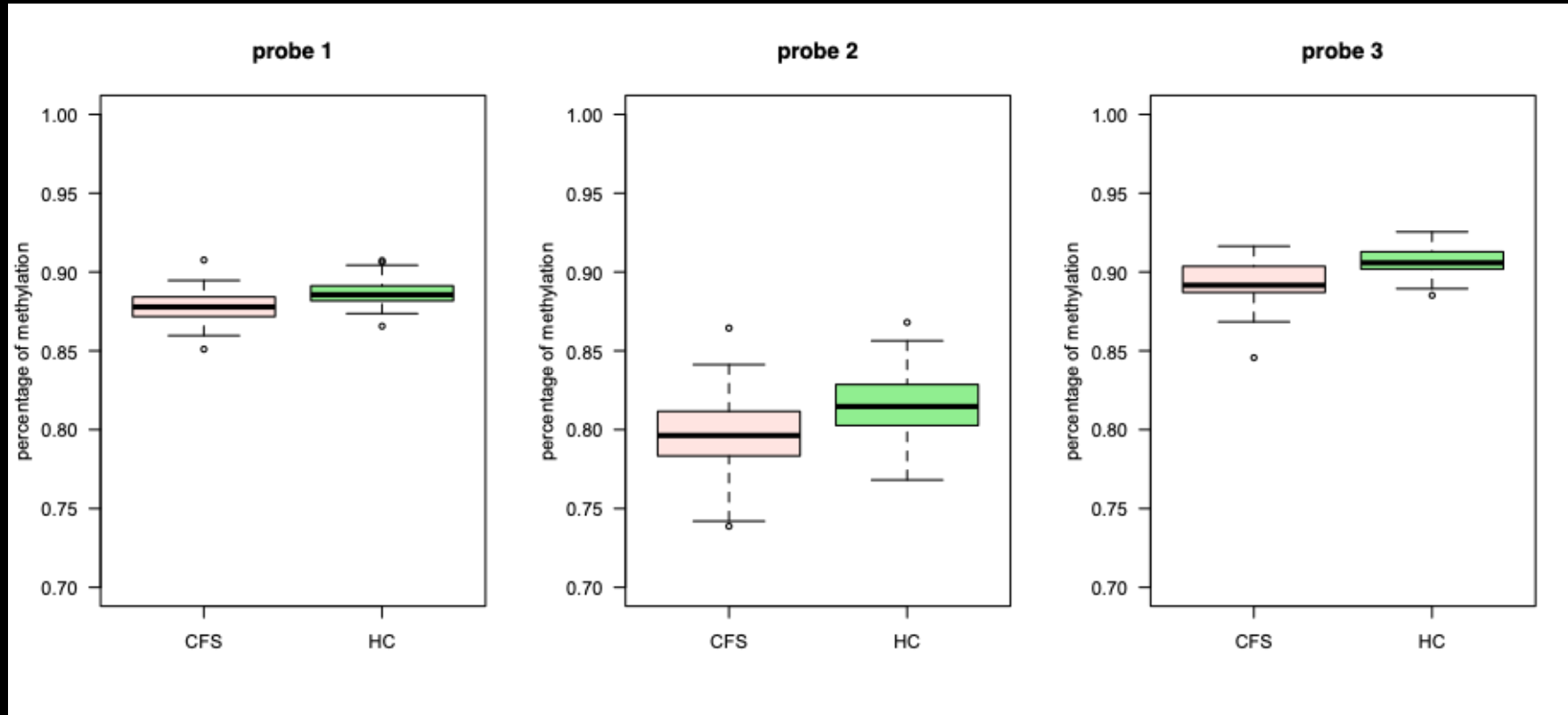
# Manhattan plots adjusting the significance level via Bonferroni correction for multiple testing



# Manhattan plots adjusting p-values via Benjamini-Hochberg procedure

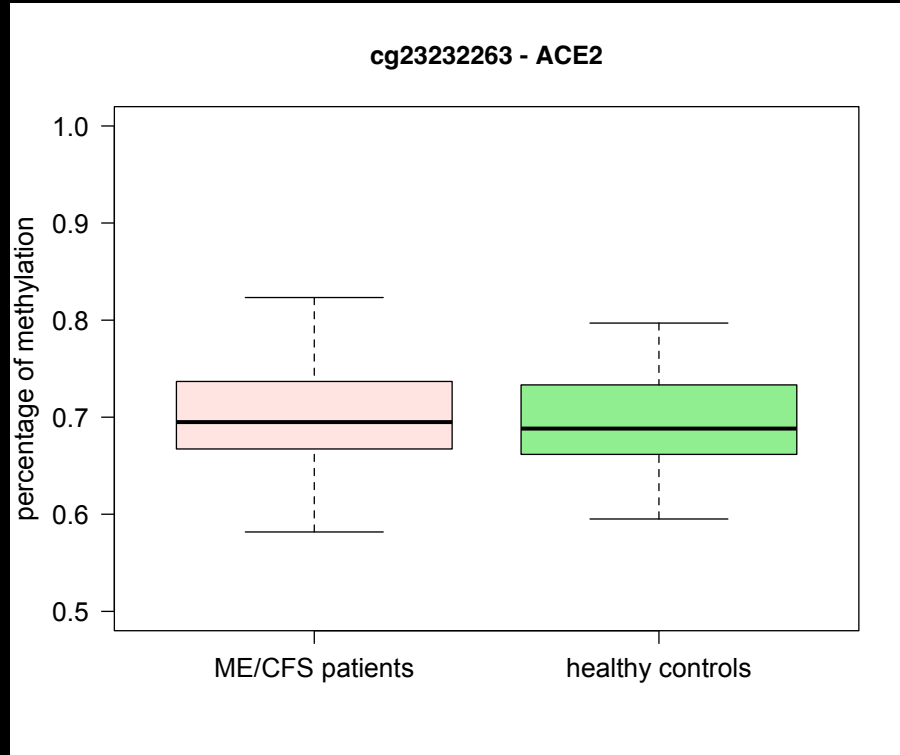


# Decreased methylation in patients with ME/CFS



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\_lecture\_12\_ACE\_ACE2.csv**

Repeat previous analysis using linear regression adjusting the effect of study and gender. 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  $i$

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

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

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

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>



## RESEARCH ARTICLE

## Open Access

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

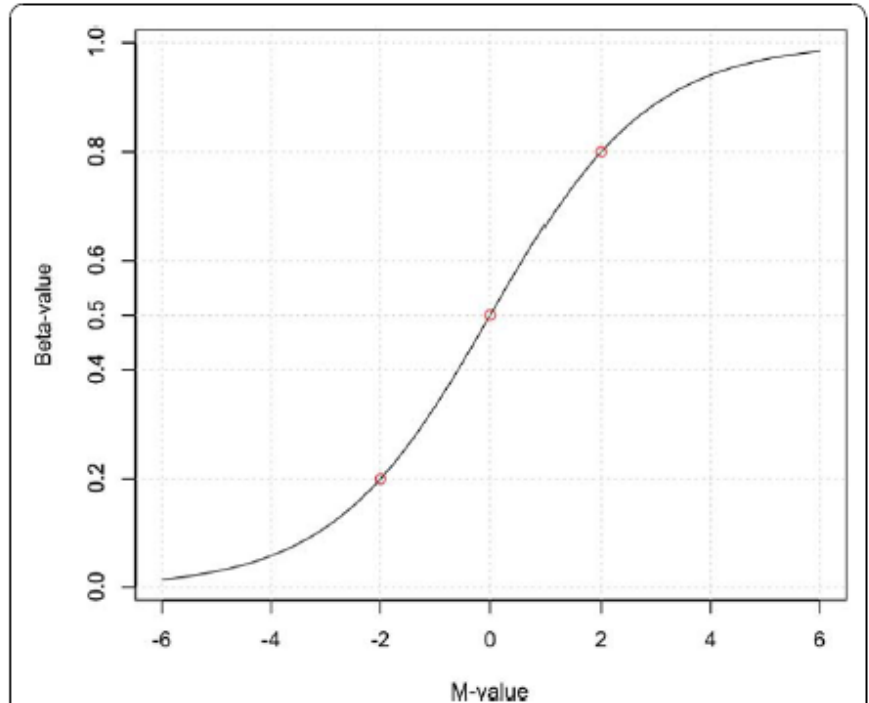
Pan Du<sup>1,3\*</sup>, Xiao Zhang<sup>2</sup>, Chiang-Ching Huang<sup>2</sup>, Nadereh Jafari<sup>4</sup>, Warren A Kibbe<sup>1,3</sup>, Lifang Hou<sup>2,3</sup>, Simon M Lin<sup>1,3\*</sup>

#### 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>



## RESEARCH ARTICLE

## Open Access

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

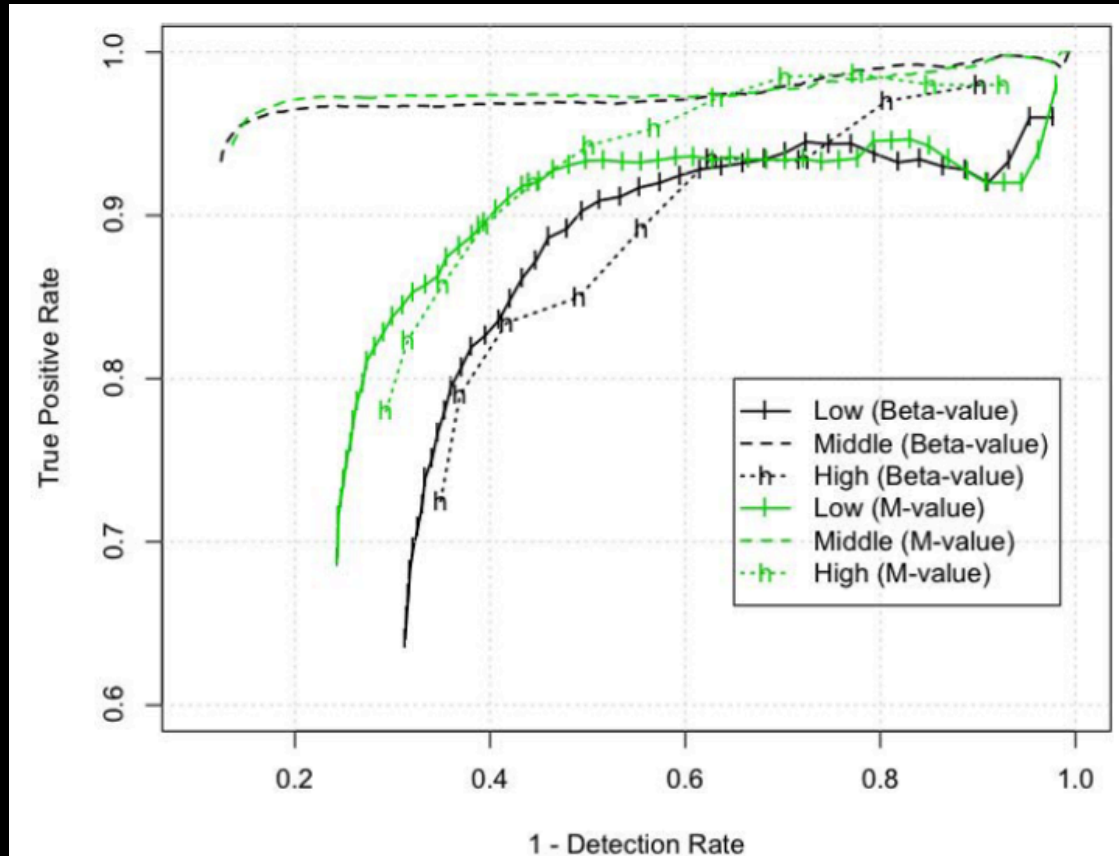
Pan Du<sup>1,3\*</sup>, Xiao Zhang<sup>2</sup>, Chiang-Ching Huang<sup>2</sup>, Nadereh Jafari<sup>4</sup>, Warren A Kibbe<sup>1,3</sup>, Lifang Hou<sup>2,3</sup>, Simon M Lin<sup>1,3\*</sup>

#### 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\_lecture\_12\_ACE\_ACE2.csv**

Repeat previous analysis using linear regression adjusting the effect of study and gender. 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  $i$

$$Y_{ij} \in (0,1)$$

$$Y_{ij} \rightsquigarrow \text{Beta}(\alpha_{ij}, \beta_{ij})$$

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

$$\phi_{ij} = \alpha_{ij} + \beta_{ij}$$

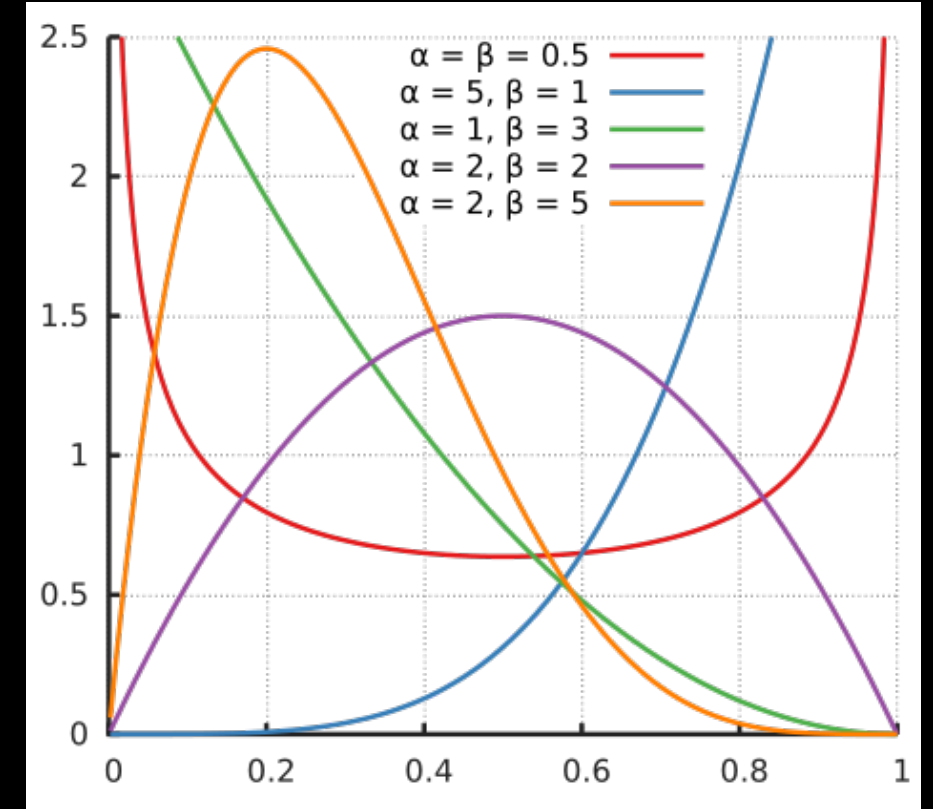
Useful  
reparametrization

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

Beta regression

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

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



Beta distribution is very flexible

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

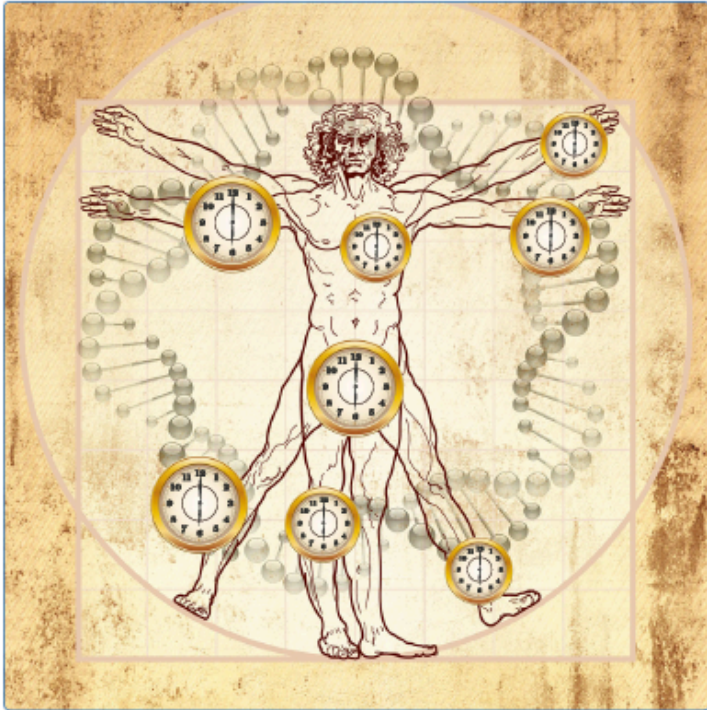
## **Exercise: Data\_lecture\_12\_ACE\_ACE2.csv**

Repeat previous analysis using Beta regression adjusting the effect of study and gender.  
Install and use the package “betareg”.

Compare the results from all the analyses done so far.

What are your conclusions?

# DNA methylation and age prediction



DNA methylation age of human tissues and cell types

Horvath

Infinium Illumina 27k - 27,578 CpG probes

Infinium Illumina 450k - 485,512 CpG probes

## Training data sets

n=7844 samples from 82 data sets

Samples from the different tissues (PBMC, lung, kidney, etc)

% methylation in more than 20k CpG

## Final linear regression model (DNAm)

Outcome = Age

Covariates = % methylation of 353 “clock” CpG probes

Penalized regression using elastic net.

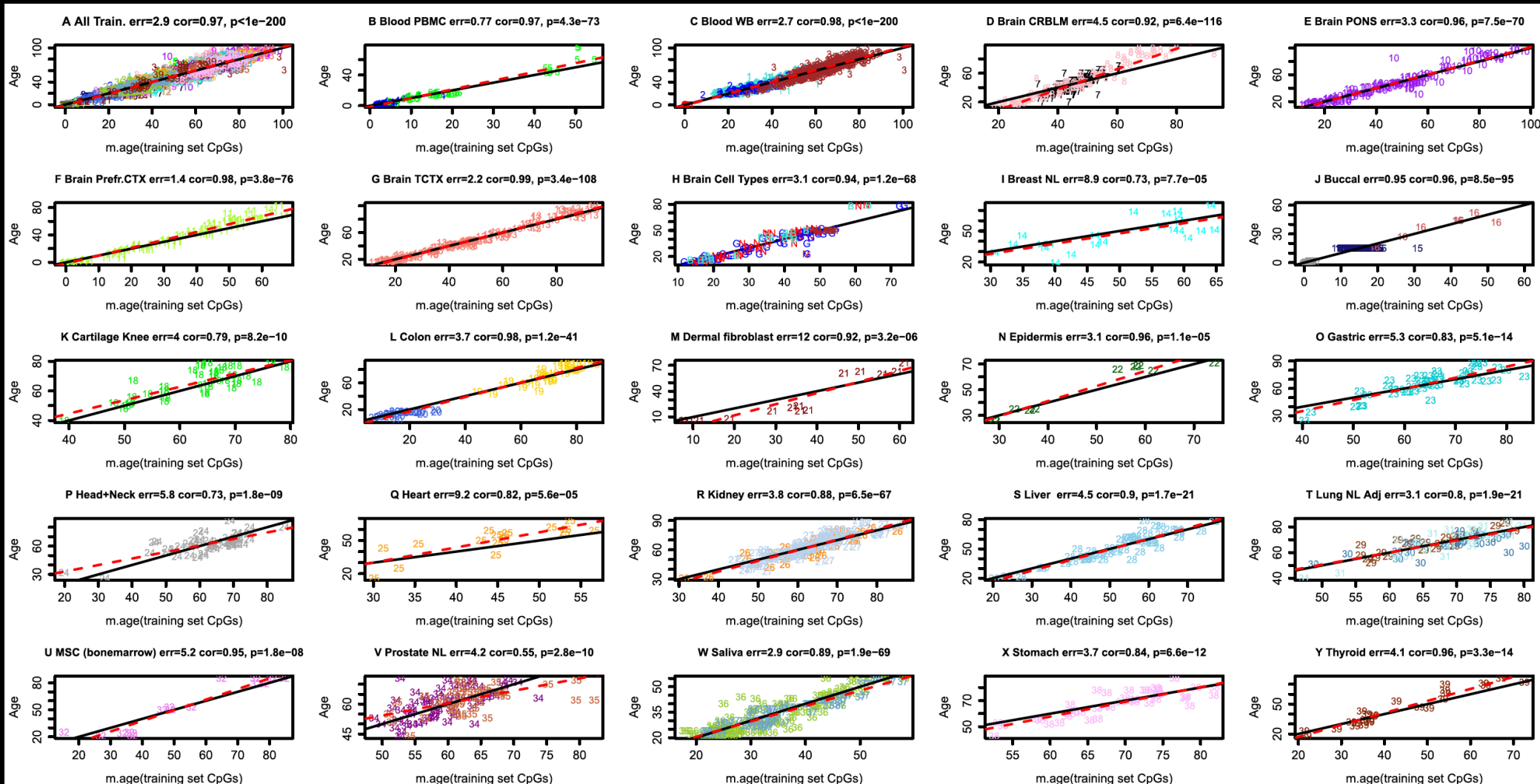
# DNA methylation and age prediction



[https://www.ted.com/talks/  
steve\\_horvath\\_epigenetic\\_clocks\\_help\\_to\\_find\\_anti\\_aging\\_treatments](https://www.ted.com/talks/steve_horvath_epigenetic_clocks_help_to_find_anti_aging_treatments)



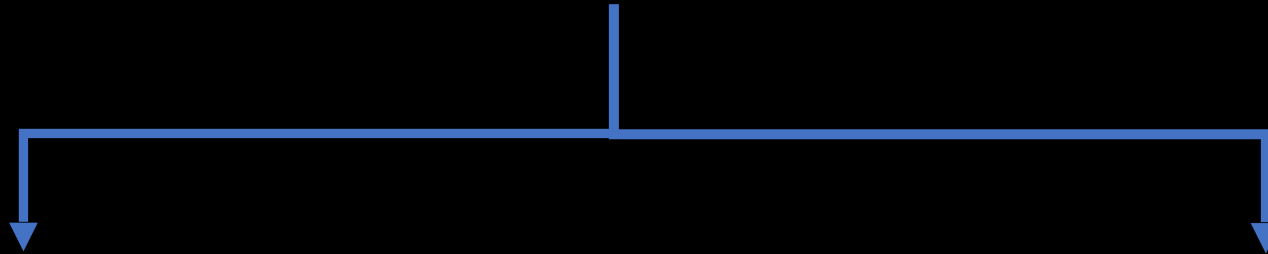
# Prediction in the training set



# Can we know whether we are aging faster than we should?

Age acceleration =

DNA methylation age - Chronological age



Age acceleration < 0

Age acceleration > 0

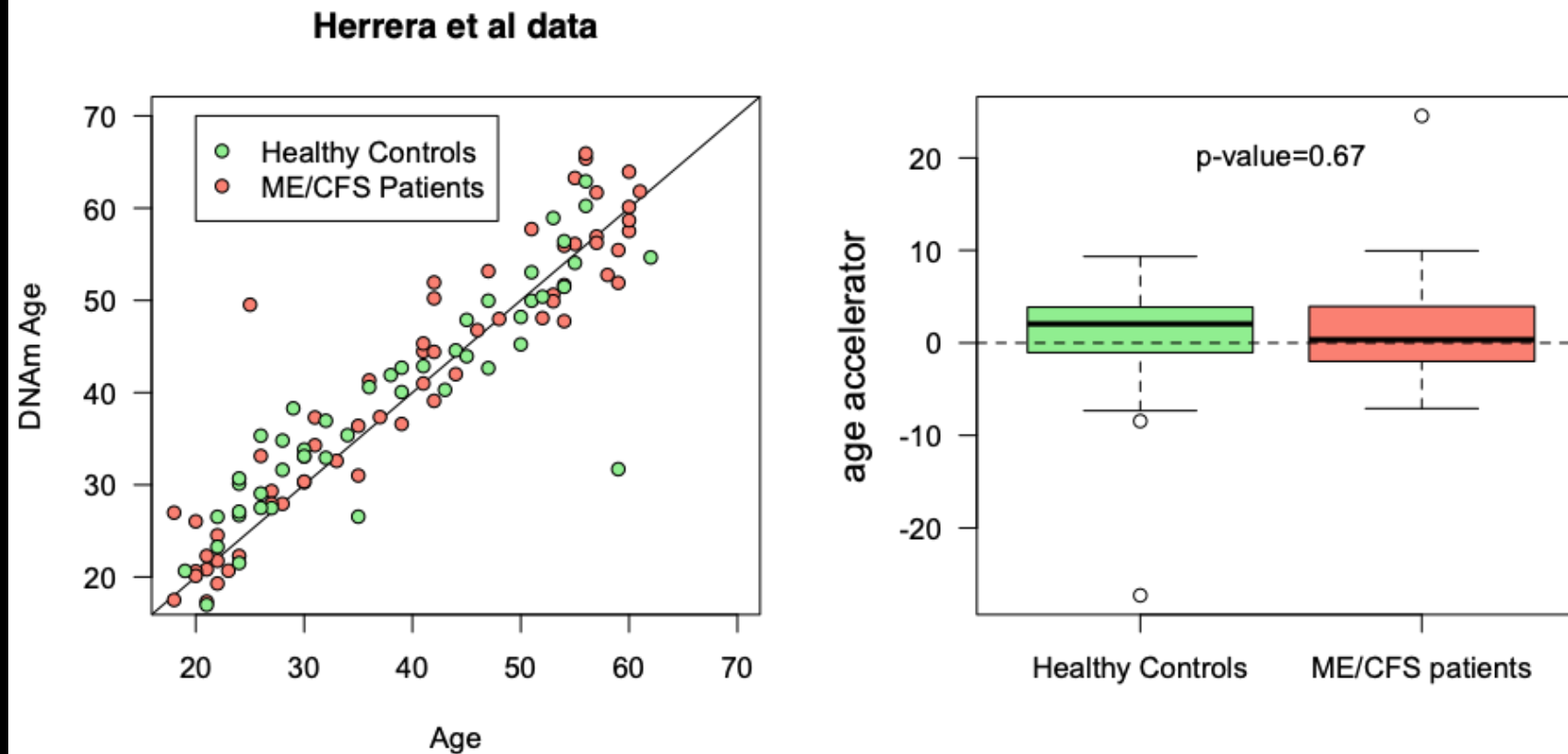
Younger than your chronological age

Older than your chronological age

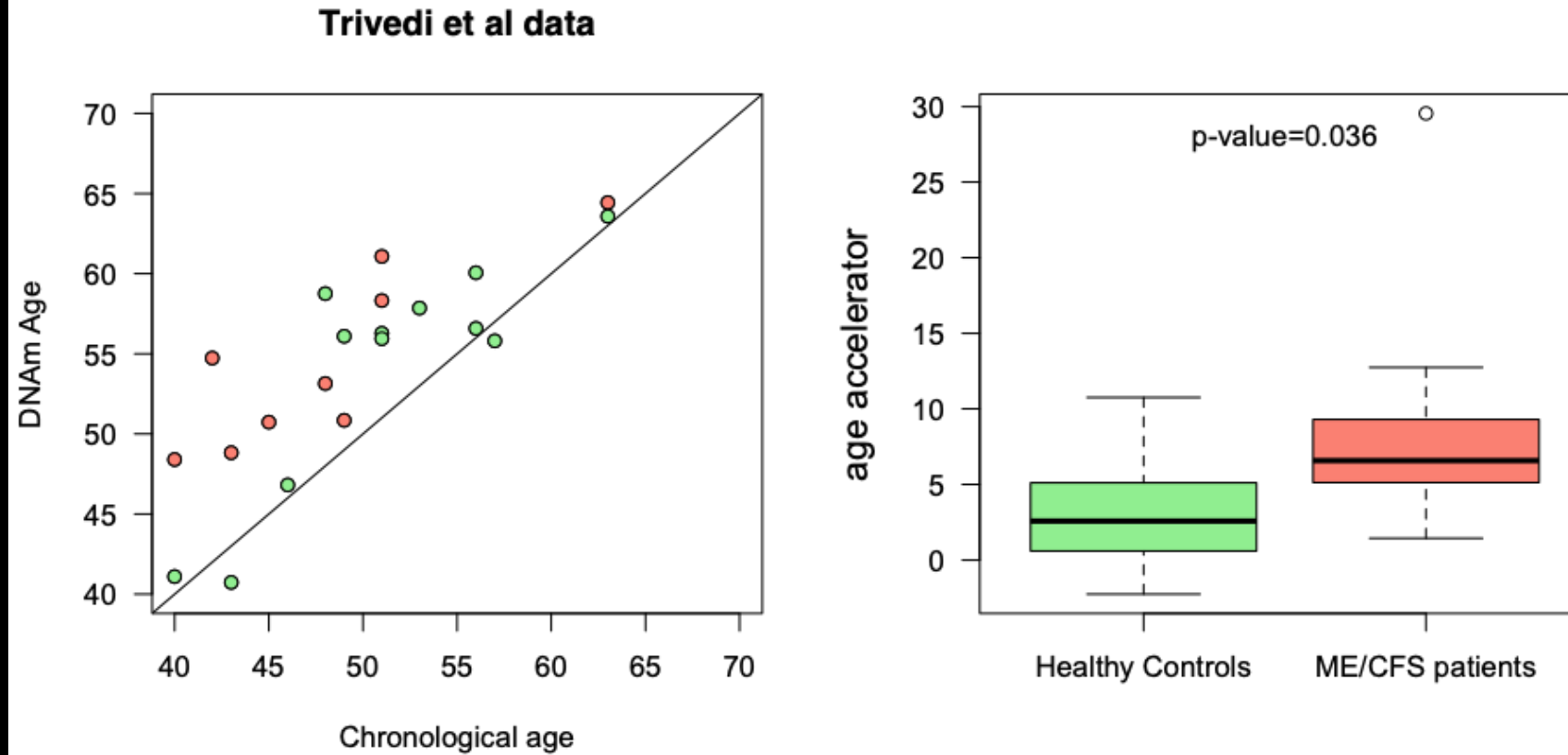
# What about the aging of patients with ME/CFS?

Reference (tissue)	ME/CFS patients			Healthy controls, n	Technology (manufacturer)	GEO ID
	n	Sample characteristics	Case definition			
<b>De Vega et al. (2014)</b> <b>(PBMC)</b>	12	Female adults Mean age: 41 yo Mean BMI: 23 kg/m <sup>2</sup>	1994 CDC/Fukuda & 2003 CCC	12	Infinium HumanMethylation450K	GSE59489
<b>De Vega et al. (2017)</b> <b>(PBMC)</b>	49	Female adults Mean age: 50 Mean BMI: 23 kg/m <sup>2</sup>	1994 CDC/Fukuda & 2003 CCC	25	Infinium HumanMethylation450 Array (Illumina)	GSE93266
<b>Herrera et al. (2018)</b> <b>(T Lymphocytes)</b>	61	<b>Female/male adults</b> <b>Mean age: 32 yo</b> <b>Mean BMI: 27 kg/m<sup>2</sup></b>	<b>1994 CDC/Fukuda</b> <b>&amp; 2003 CCC</b>	48	<b>Infinium</b> <b>HumanMethylation450</b> <b>K Array (Illumina)</b>	<b>GSE156792</b>
<b>Trivedi et al. (2018)</b> <b>(PBMC)</b>	13	<b>Female adults</b> <b>Mean age: 50 yo</b> <b>Mean BMI: 26 kg/m<sup>2</sup></b>	<b>1994 CDC/Fukuda</b> <b>&amp; 2003 CCC</b>	12	<b>Methylation EPIC Array</b> <b>(Illumina)</b>	<b>GSE111183</b>

# Herrera et al data

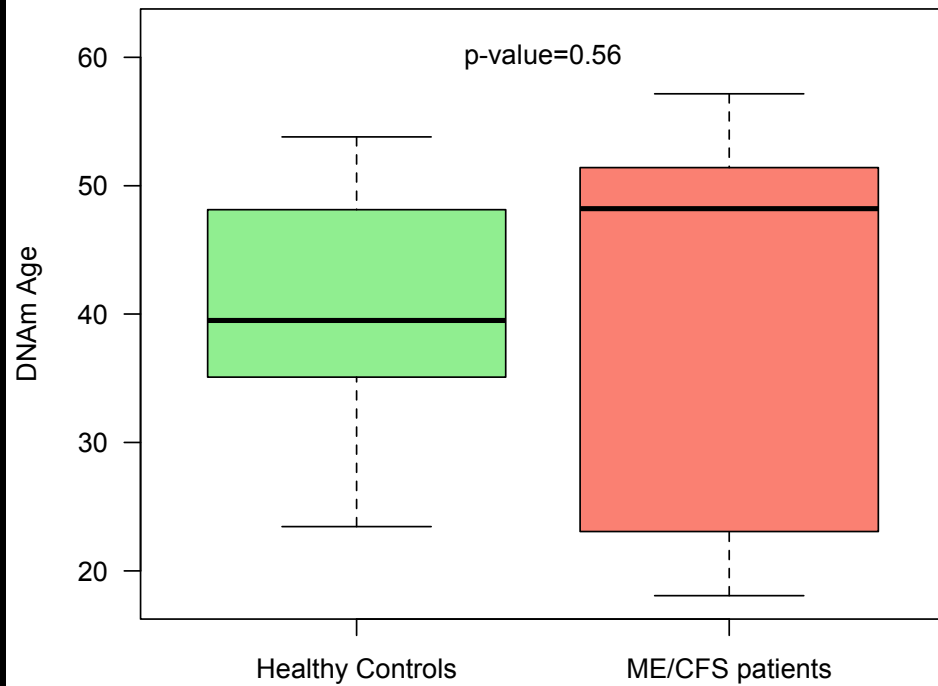


# Trivedi et al data

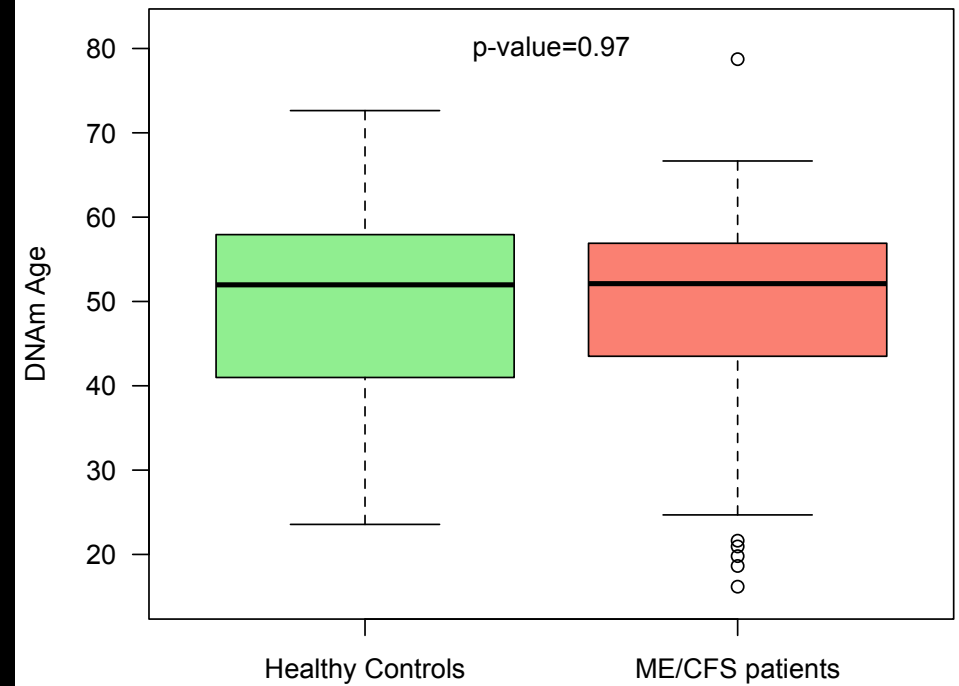


# De Vega et al studies

De Vega et al (2014) data



De Vega et al (2017) data



**How to construct your own model for predicting the biological age based on DNA methylation data?**