THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Methylome-wide Association Studies

Part 1: Data preparation

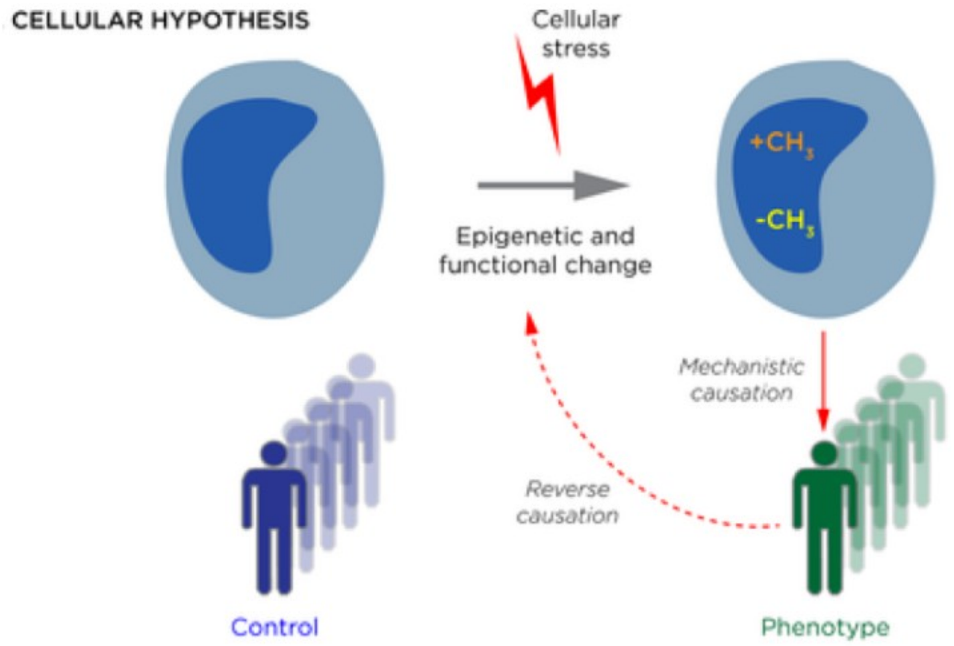
Methylome-wide Association Studies

- Also known (incorrectly) as Epigenome-wide association studies
- Identifies changes in methylation levels at single CpG sites that are associated with human phenotype/disease

- Similar to GWAS
  - Association analysis between each CpG and phenotype of interest (~450,000 association analyses)
  - Unlike SNPs, DNA methylation measurements considered as quantitative measure.
  - Linear or logistic regression (for binary dependent variables)
  - Interpretation of effect depends on whether methylation is your dependent or independent variable

Methylation Can be Cause or Consequence

- Methylation changes can be driven by disease
- e.g. alterations in white blood cell proportions in autoimmune disorders or altered metabolic regulation in type 2 diabetes
- This is different to SNPs which are fixed from conception

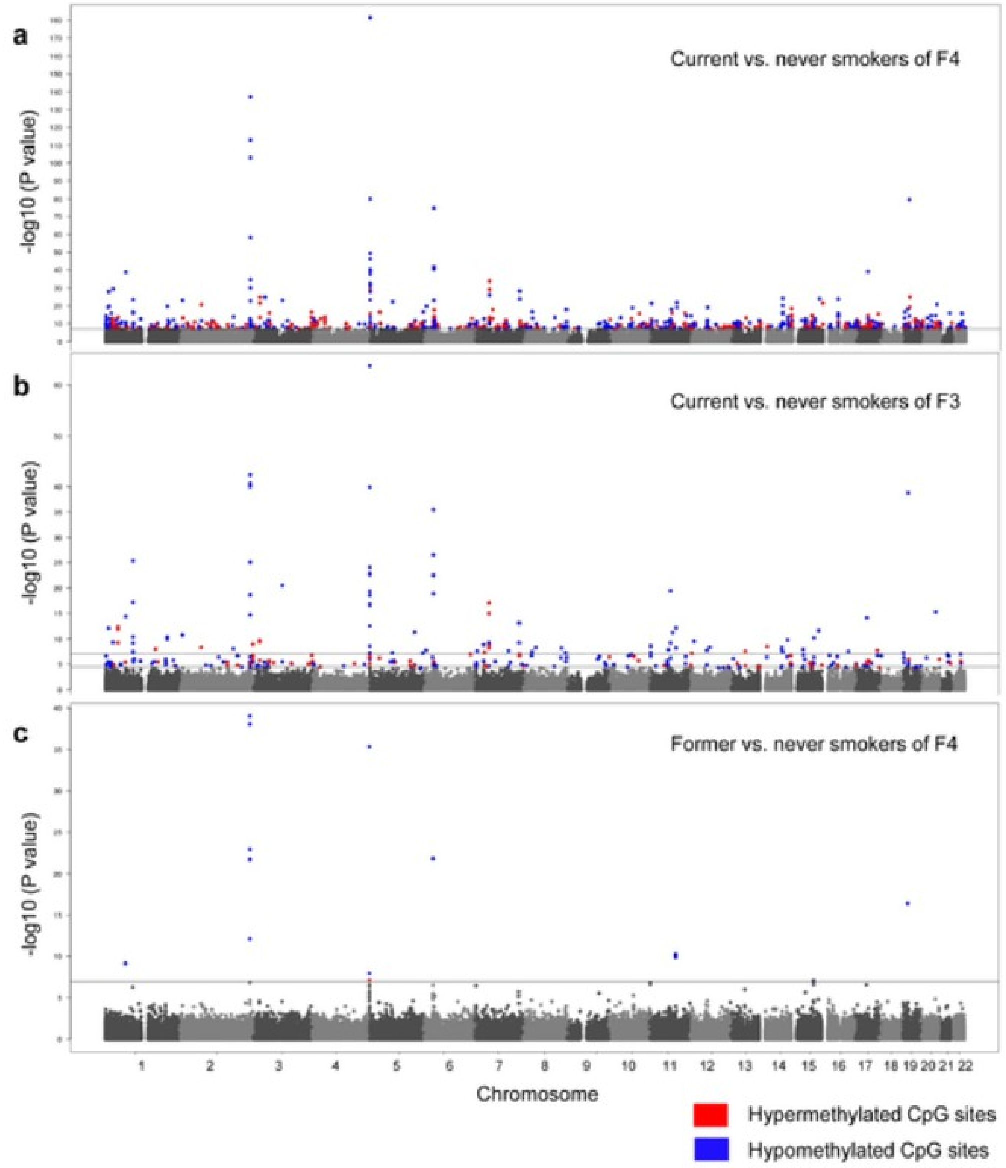


MWAS

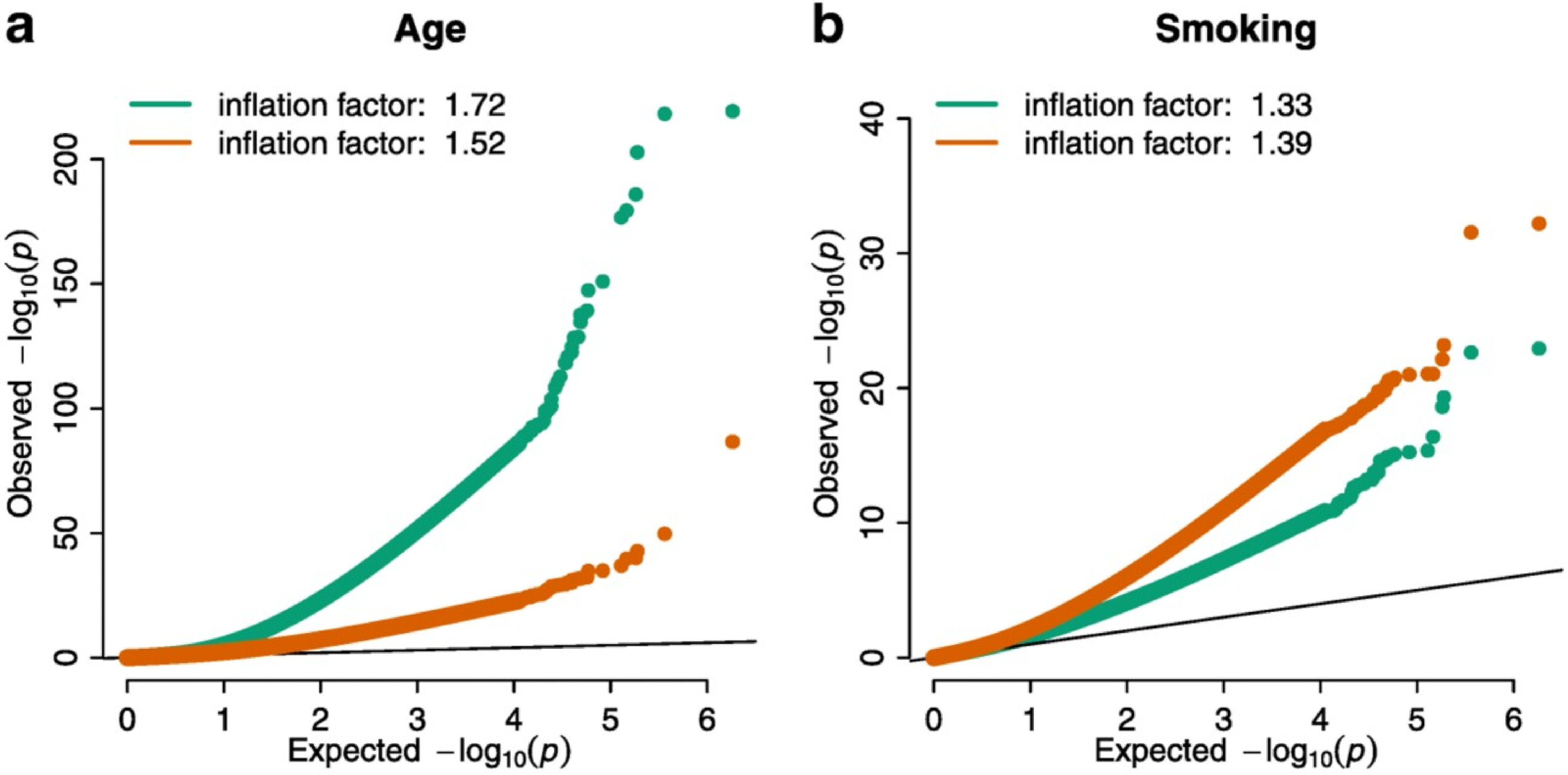
- Very similar to GWAS...

- Test each DNA methylation site across the genome for association with your trait of interest

- “Manhattan” plot and QQ plots to assess confounding



Inflation in MWAS



Why is Inflation More of an Issue in MWAS

- Methylation probes have more extensive correlations than SNPs
- Associations at one probe cause inflated test statistics beyond local genomic region

- Confounders are much more of an issue
    - e.g. case-control study on blood DNA methylation could be confounded by inflamation
    - Differences in age, sex ratios, smoking, ….

Controlling inflation in MWAS

- 1) Correct for known covariates
- 2) Predict unknown covariates
- 3) Explicitly model unknown covariates

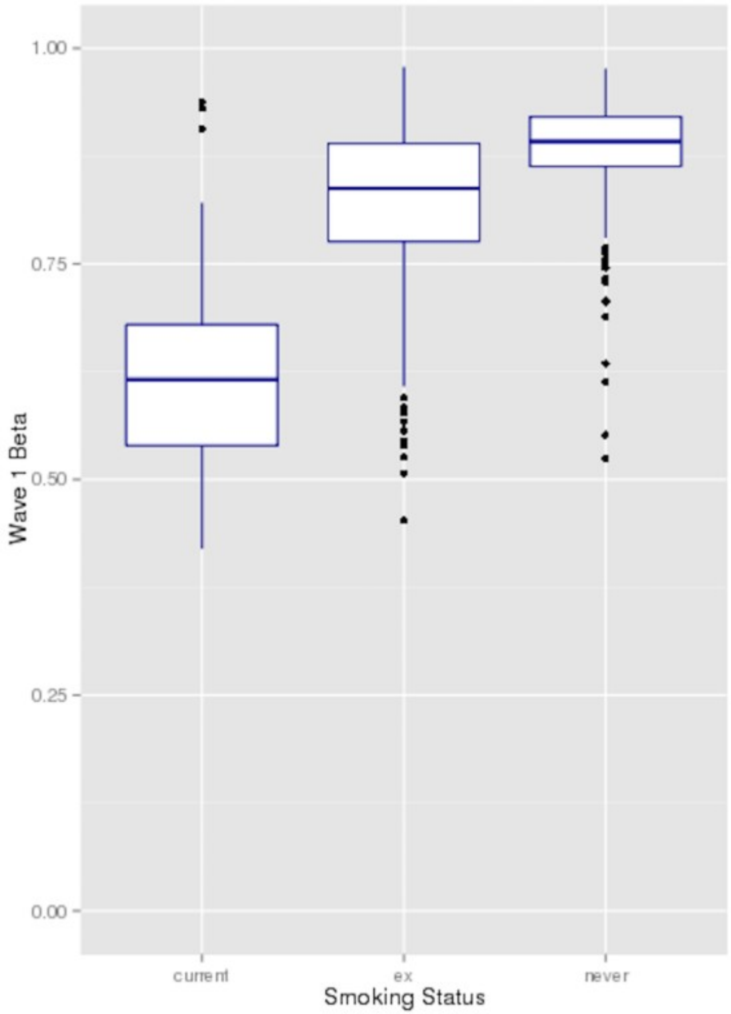
Correcting for Known Covariates

- Experimental design is particularly important in ‘omics studies
- Randomisation is important when generating DNA methylation data
- Record potential batch effects to correct for in analysis
  - Array ID
  - Position on array
  - Extraction date
  - Lab technician
  - ….

Prediction of Unknown Covariates

- The “confounding” in ‘omics data can be used to estimate covariates to include in your analysis

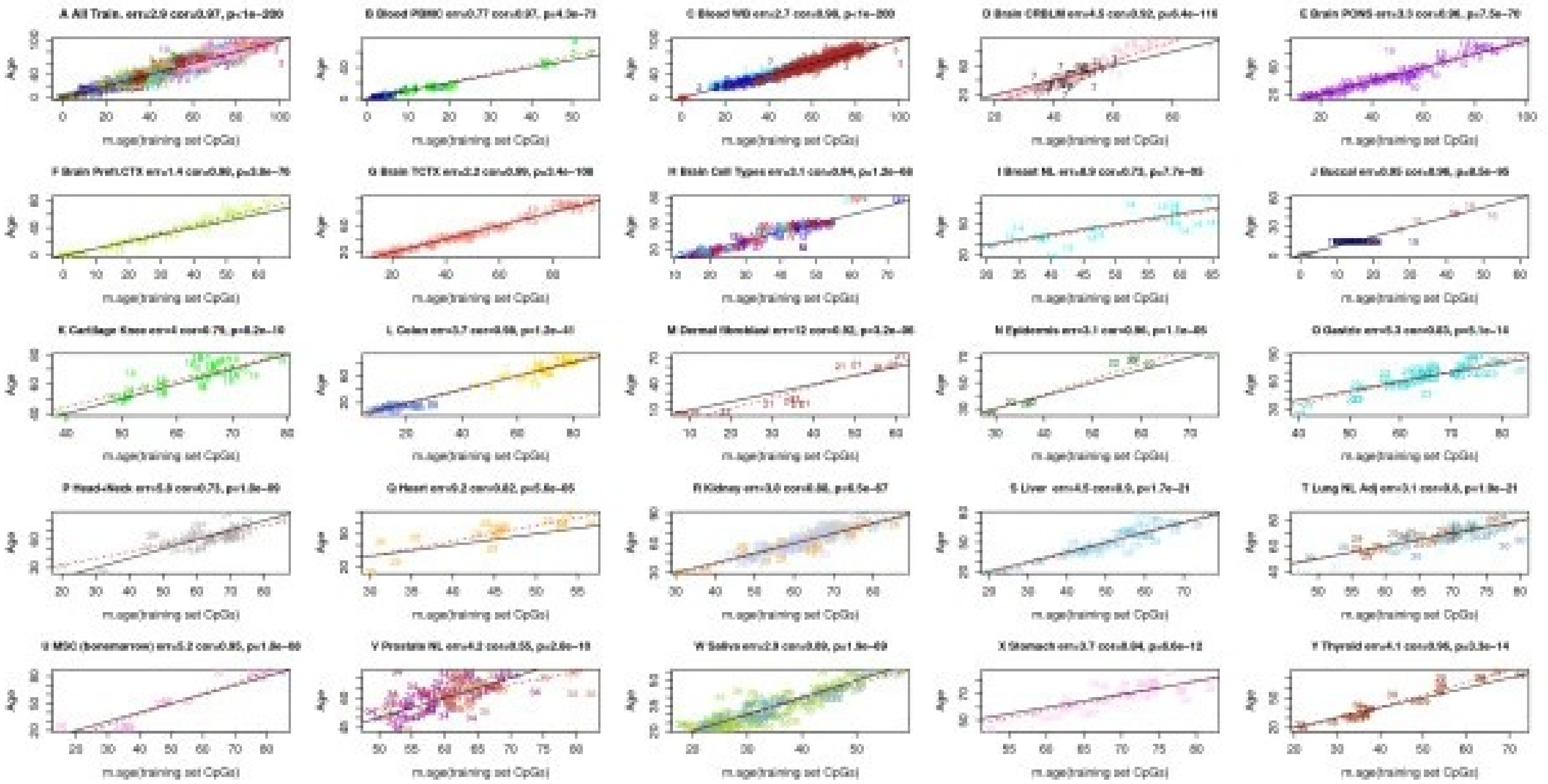
- e.g. Smoking has strong associations with DNA methylation
- The most associated CpG in the AHRR gene can imputed smoking status with an accuracy of >90% on its own.
- The imputed value may be more epidemiologically relevant than the reported smoking measure



Prediction of Unknown Covariates

- Age can be accurately imputed from DNA methylation

- Several “clocks” available for human age imputation



Prediction of Unknown Covariates

- DNA methylation varies by cell type
- Cell composition of blood/tissues can vary with disease state

- Cell composition can be estimated provided a good reference panel is available
- We estimated blood cell proportions in the last practical

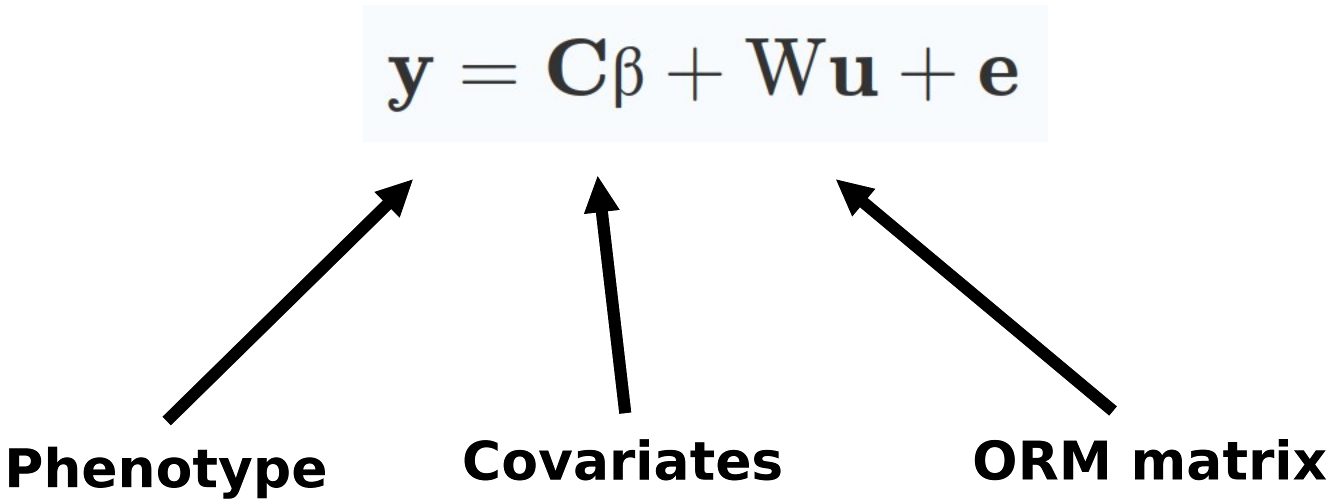
Modelling Unknown Confounders

- Many, many methods are available to model unknown confounders

- PCA – removes axis explaining most variation in the data – could include your trait
- SVA – think of it like a PCA that does not remove variation associated with the trait
- ReFACTor – specifically for cell type composition (?)
- RUV – run EWAS, pick unassociated probes to do PCA on, rerun EWAS with covariates
- ...
- …
- ...

OSCA – OREML

- Model the covariance of all ‘omics measures at the same time in a mixed linear model as a random effect
- Create an ‘Omics Relationship Matrix (ORM), which measures the similarity between individuals



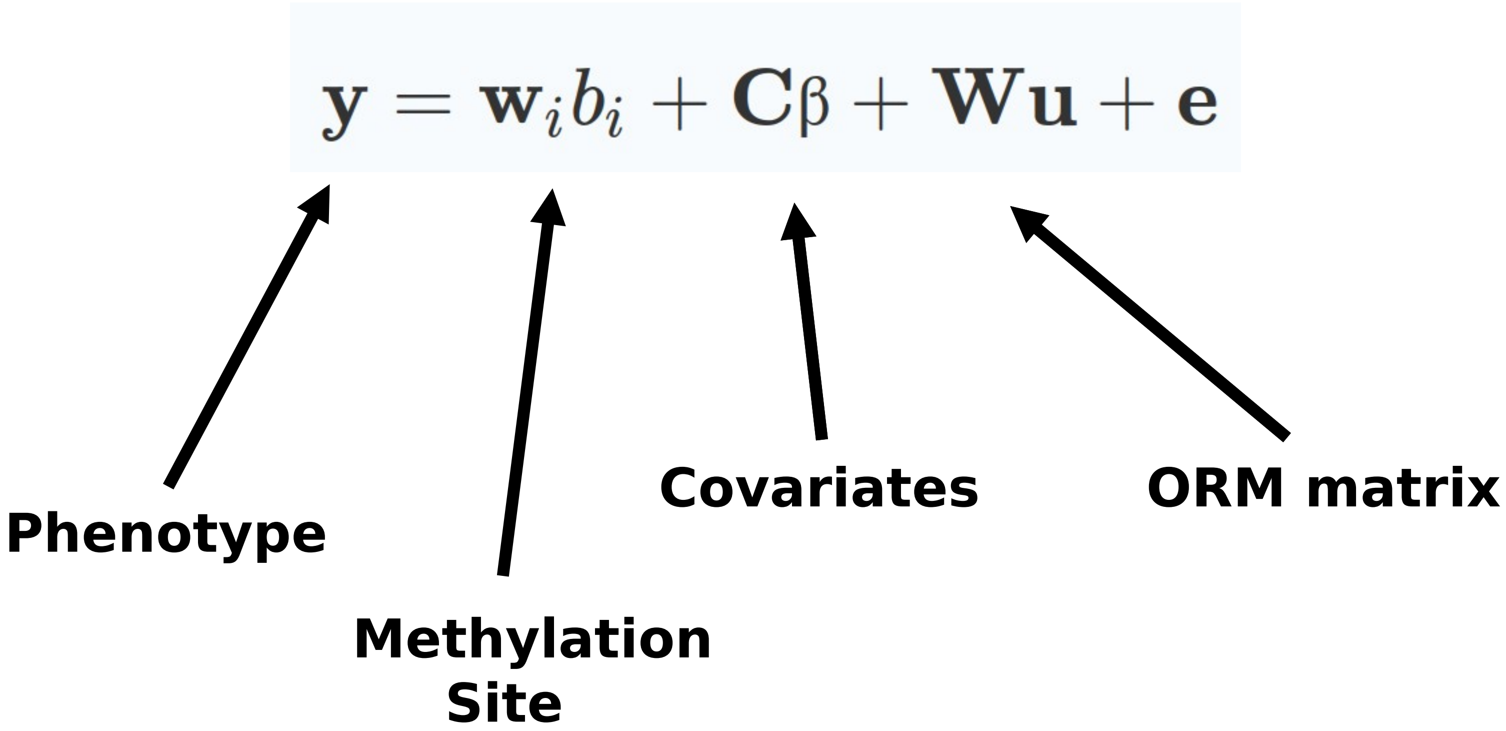
$$\mathbf{y} = \mathbf{C}\beta + \mathbf{W}\mathbf{u} + \mathbf{e}$$

Phenotype **Covariates** **ORM matrix**

- Can estimate the proportion of variation in a trait captured by ‘omics measures

OSCA – MOA Method

- Test for association at a probe while modelling the covariance across all probes
-



$$\mathbf{y} = \mathbf{w}_i b_i + \mathbf{C}\boldsymbol{\beta} + \mathbf{W}\mathbf{u} + \mathbf{e}$$

Phenotype

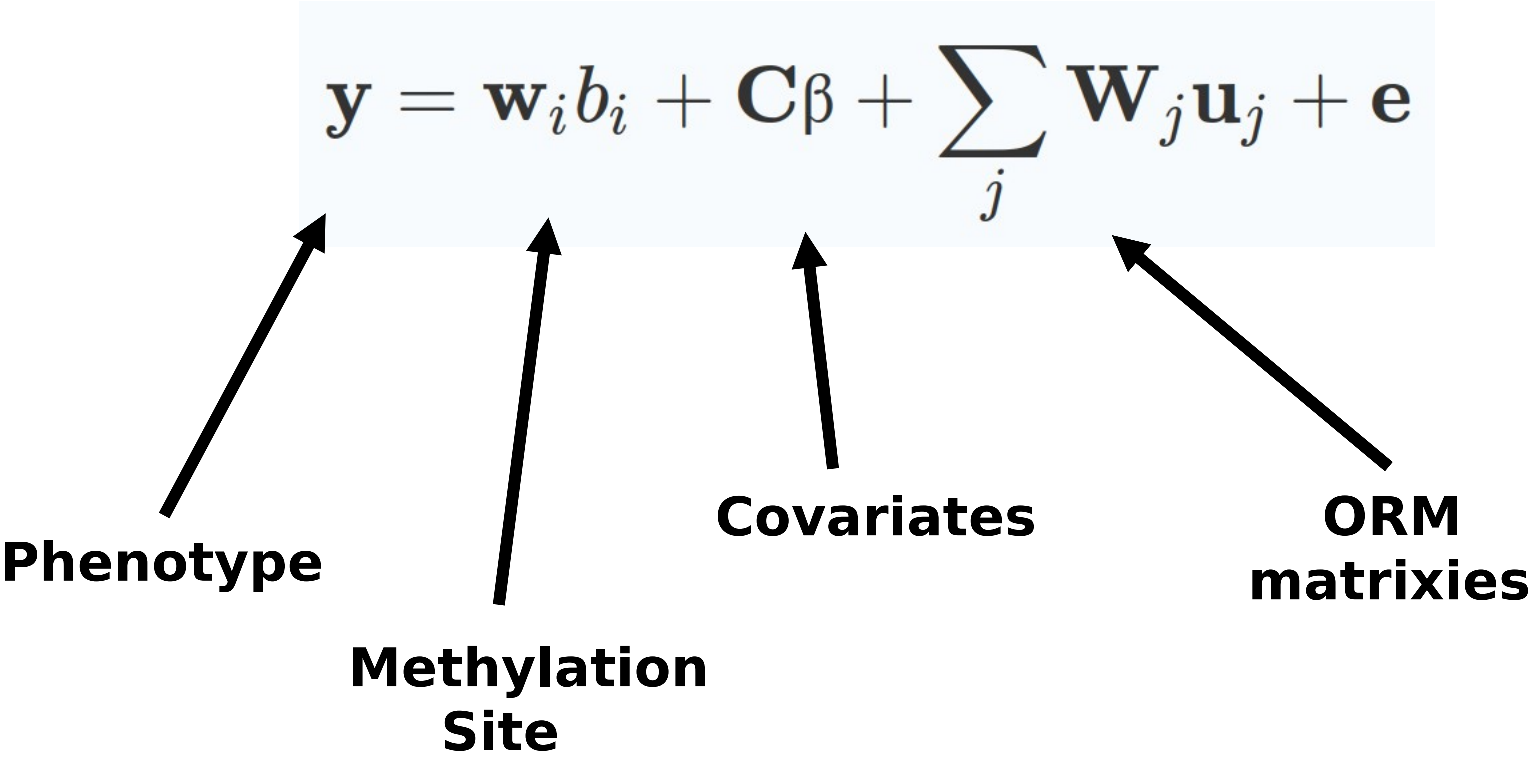
Methylation Site

Covariates

ORM matrix

OSCA – MOMENT Method

- Model multiple ORMs
- One ORM made from probes associated in a linear regression analysis, and one ORM with the rest of the probes



$$\mathbf{y} = \mathbf{w}_i b_i + \mathbf{C}\beta + \sum_j \mathbf{W}_j \mathbf{u}_j + \mathbf{e}$$

Phenotype

Methylation Site

Covariates

ORM matrixies