QC & Analysis of Methylation Chip Data

Allan McRae & Sonia Shah

Outline for Session 3 Lecture

- EWAS analysis
- Inflation in test-statistics
- Interpreting EWAS results
- Study design
- Examples: Smoking, age, BMI and height, ALS

Epigenome-wide association studies

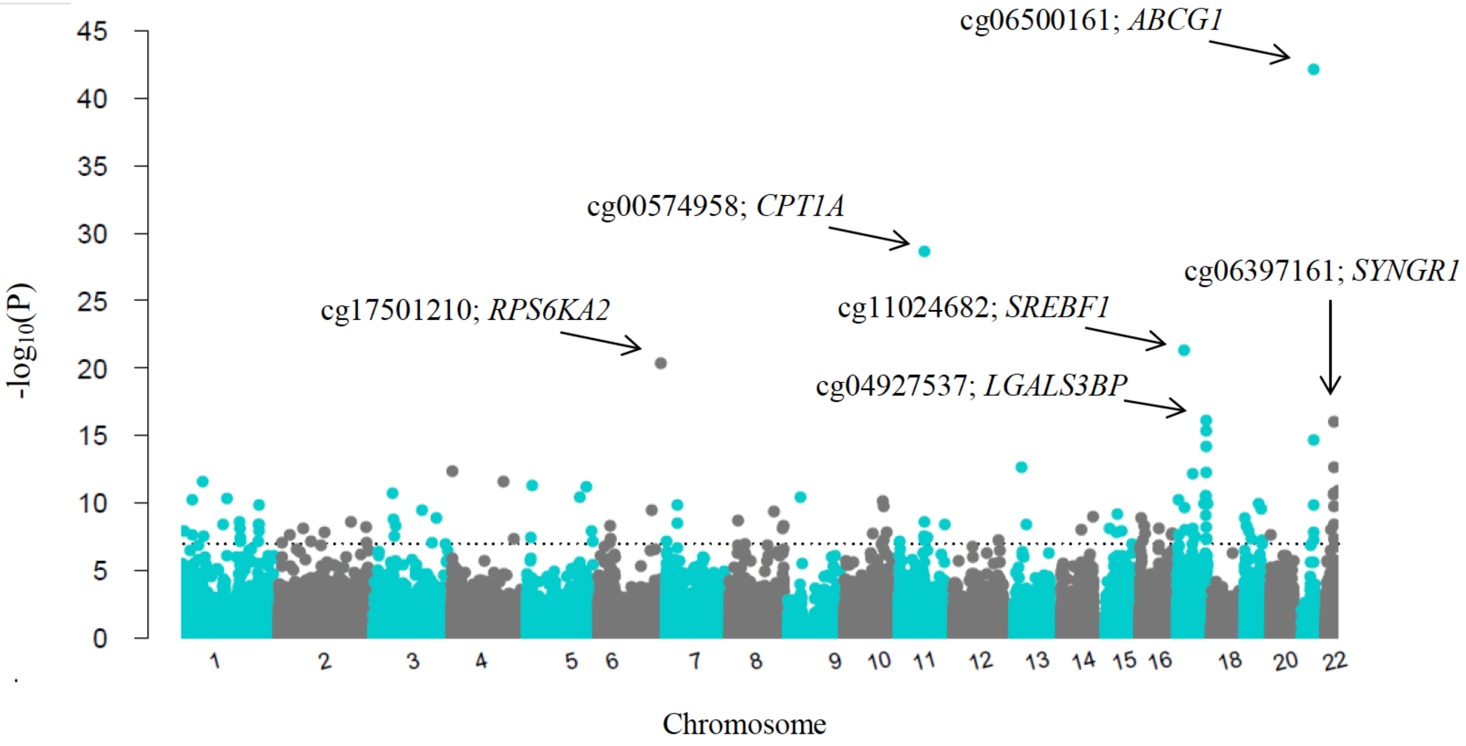
- Identifies changes in methylation levels at single CpG sites that are associated with human phenotype/disease

- Similar to analysing SNPs in 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

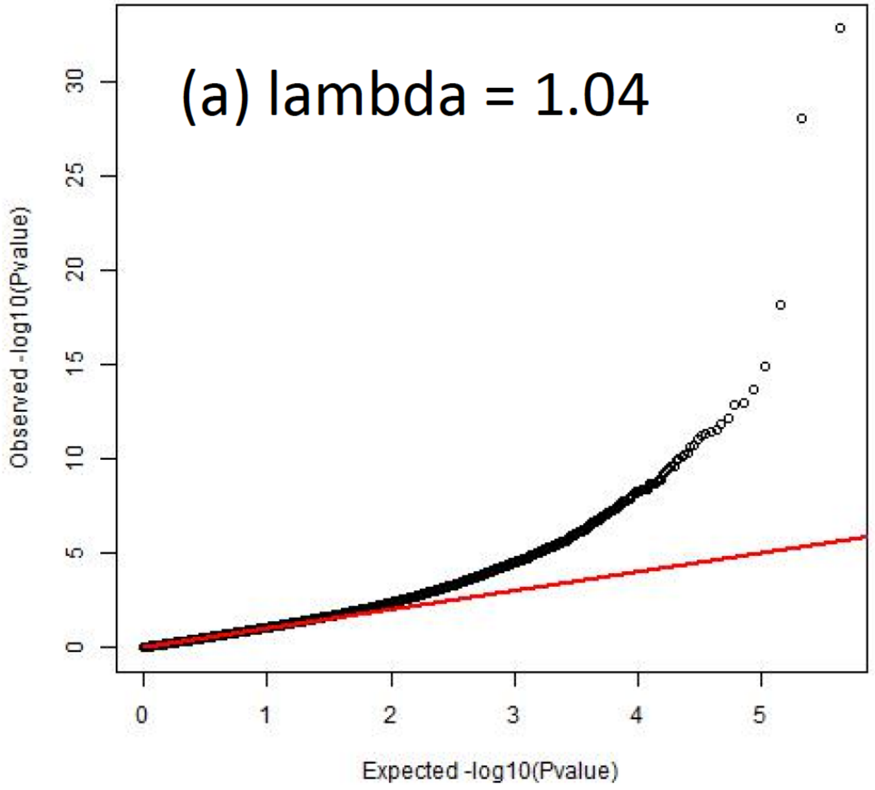
$$CpGmeth \sim smoking + covariates + PCs$$
$$disease \sim CpGmeth + covariates + PCs$$

Visualising results

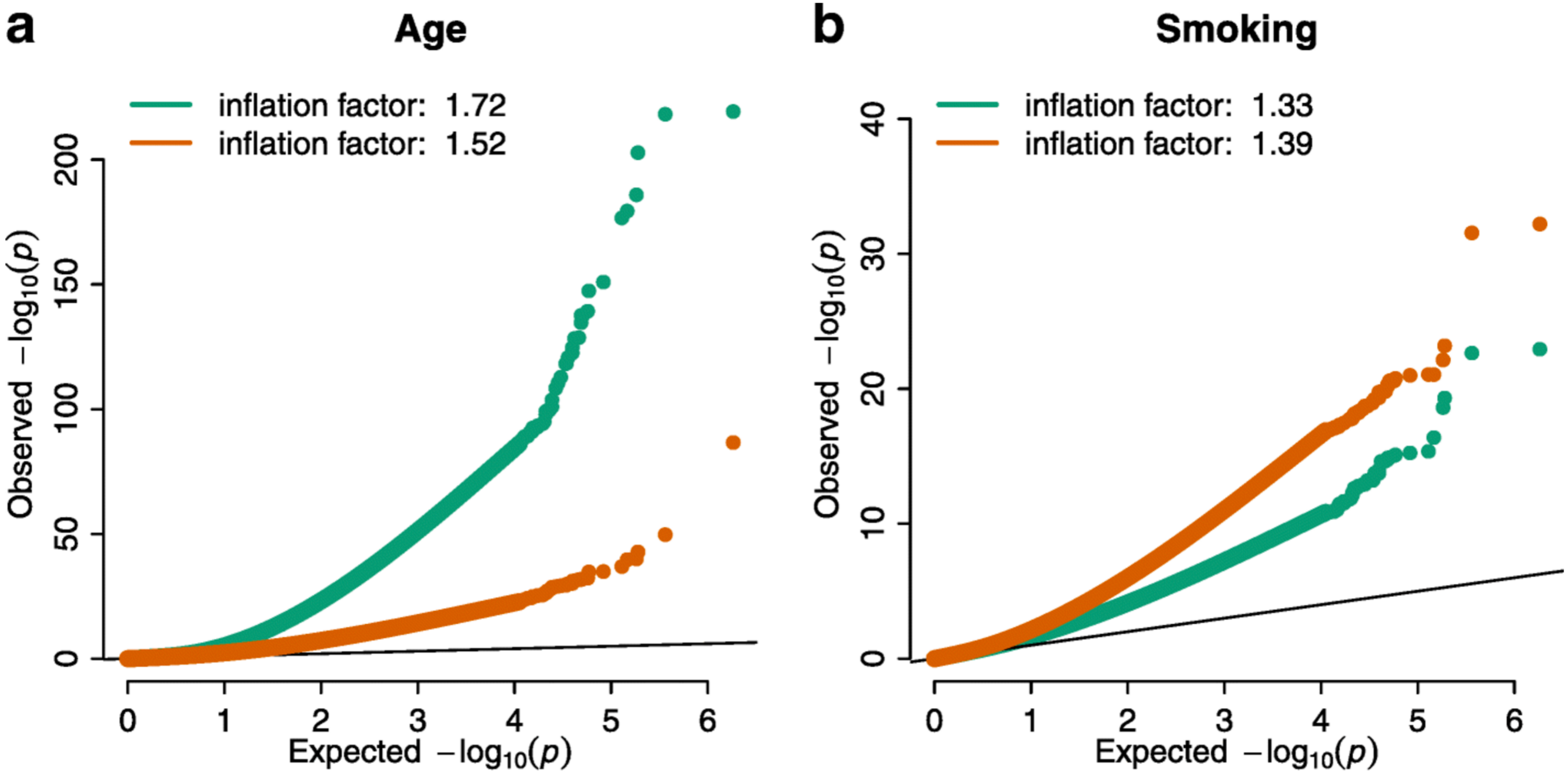
manhattan



QQ



Inflation in lambda



Iterson et al Genome Biology 2017

Controlling inflation in EWAS

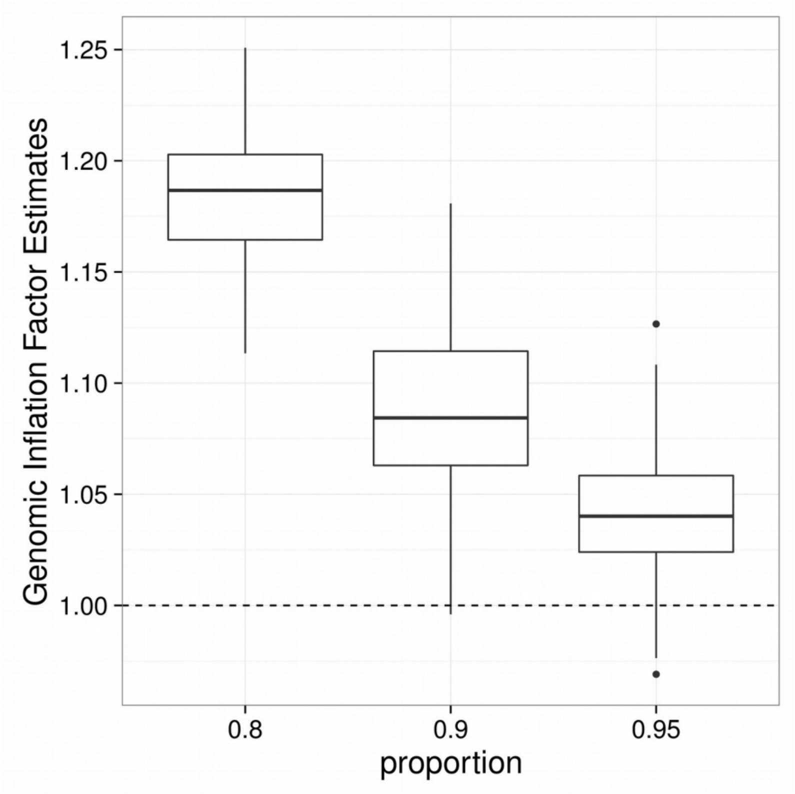


Figure 2 | The genomic inflation factor overestimates inflation if a moderated proportion of true associations is present Sets of test-statistics were generated with different amounts of true associations (20%, 10% and 5%) but without any true inflation, i.e., the inflation factor should be equal to one (**Supplemental Methods**). The genomic inflation factor was calculated as the square-root of the median of squared test-statistics divided by 0.456, the median of chi-square distribution with one degree of freedom[8].

- Simulation study showing that the genomic inflation factor depends on the number of true associations
- genomic inflation factor commonly overestimates the true level of test-statistic inflation in EWAS and TWAS

Controlling inflation in EWAS

- http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1131-9 Published Jan 2017
- EWASs and TWASs are prone not only to significant inflation but also bias of the test statistics
- Not properly addressed by GWAS-based methodology (i.e. genomic control) or approaches to control for unmeasured confounding (e.g. RUV, sva and cate).
- Method to estimate the empirical null distribution using Bayesian statistics.
- http://bioconductor.org/packages/bacon/.

Interpretation of EWAS much more complicated than GWAS

Study design very important

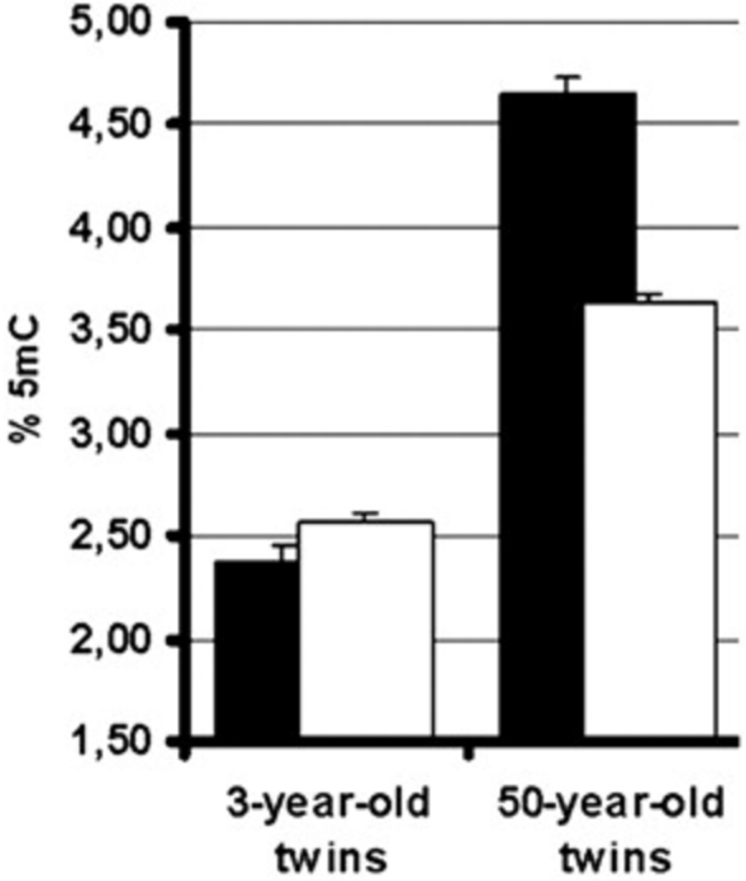
Advantage of GWAS

- Genotype is constant from birth
  - Genotype comes before phenotype
  - no issue of reverse causation i.e. phenotype does not cause changes in genotype.

- Genetic variants assumed to be randomly assigned with respect to the characteristics of individual, therefore minimised confounding bias
  - Ascertainment bias
  - Population stratification (which can be corrected for)

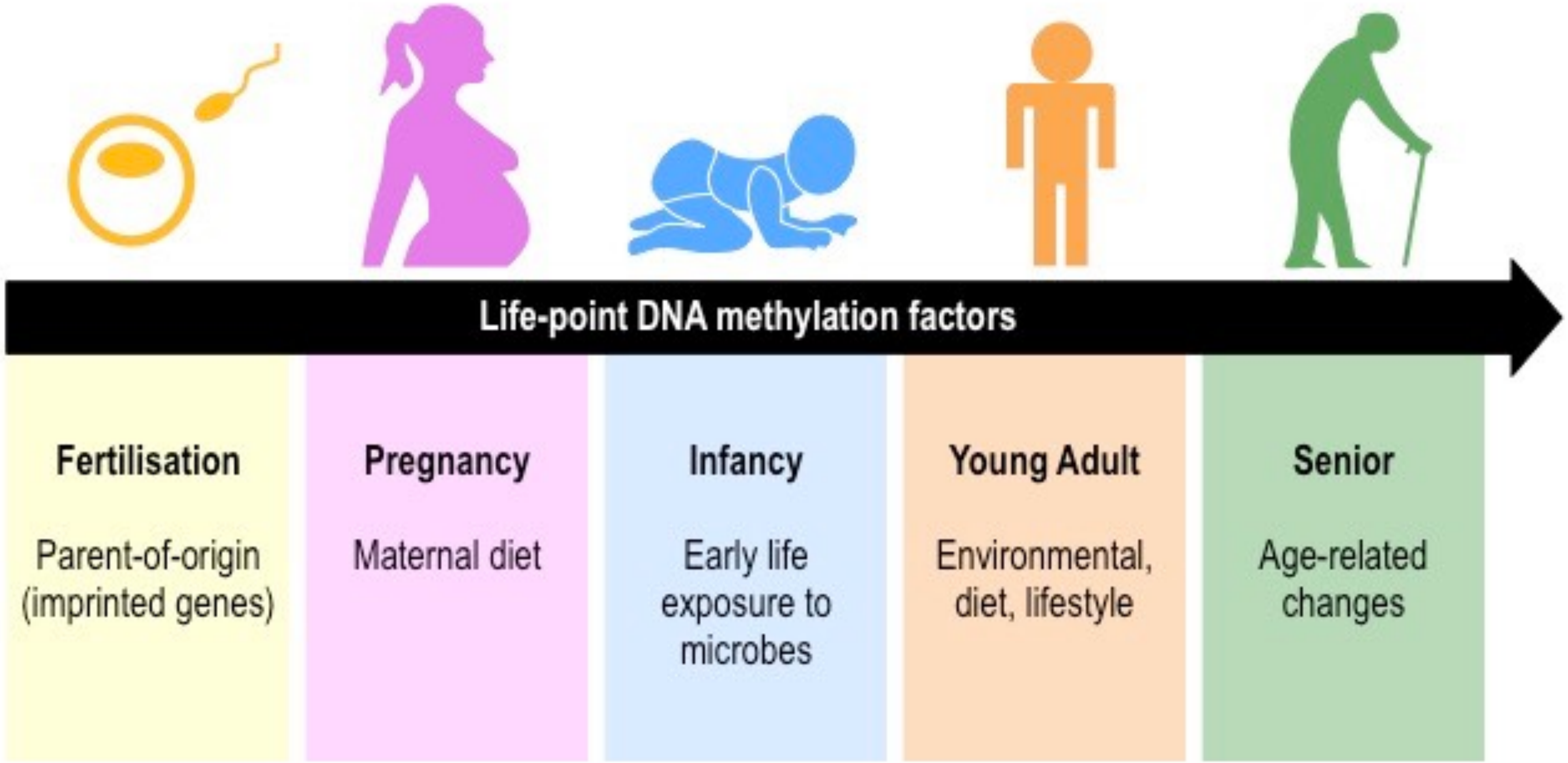
Methylation is dynamic

Differences in global 5mC DNA content in monozygotic twins



Fraga et al *PNAS* 2005

Methylation is dynamic

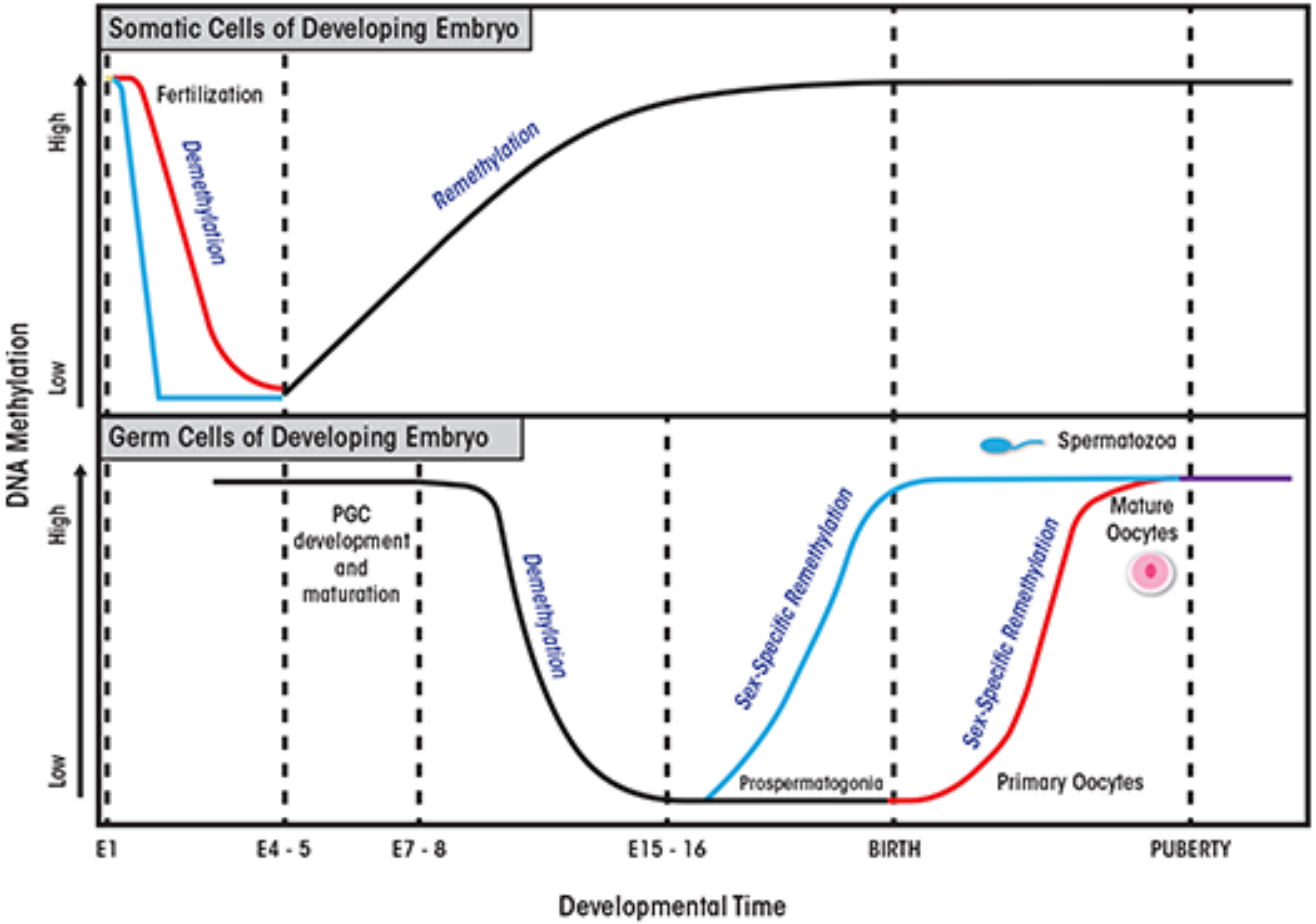


Life-point DNA methylation factors

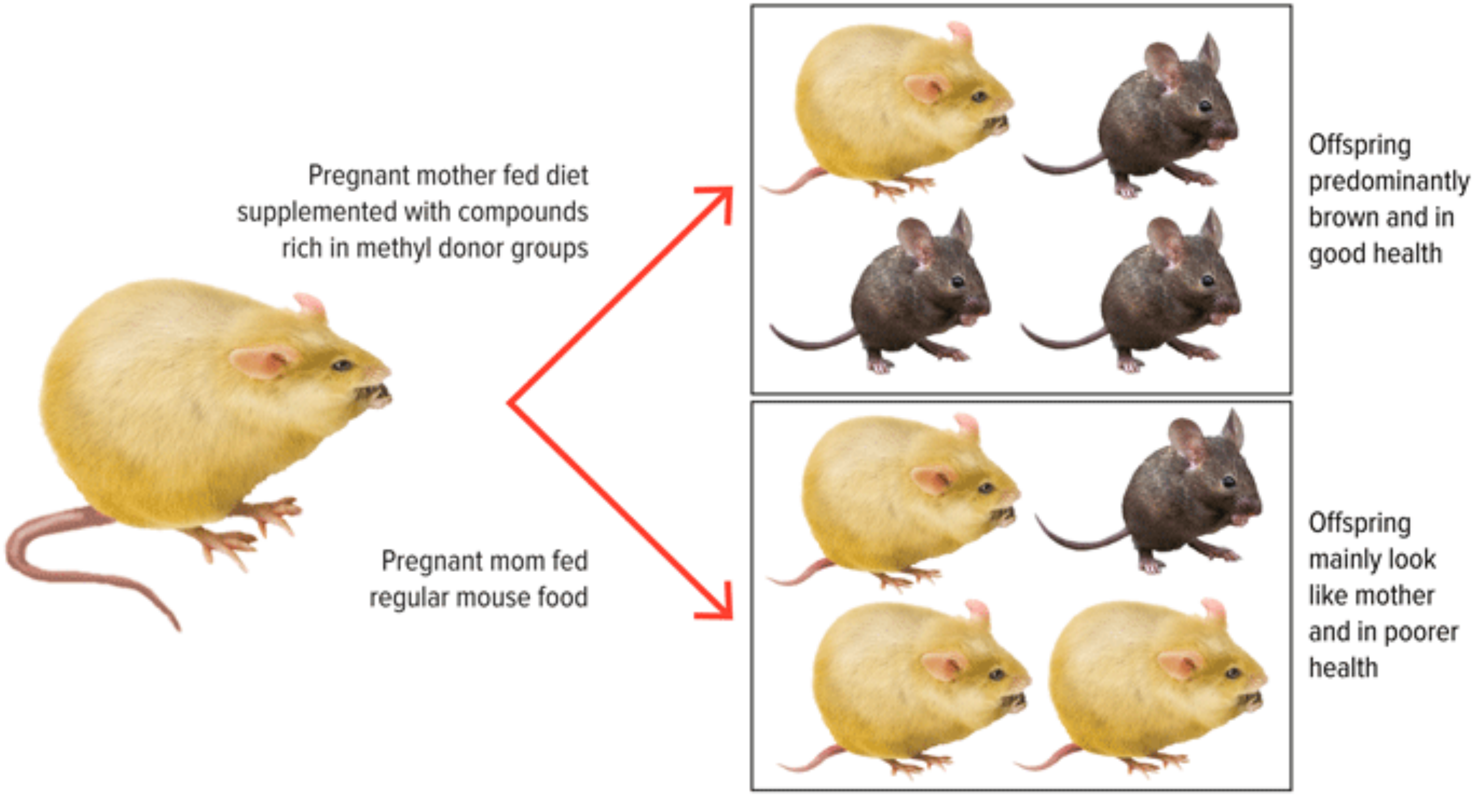
Fertilisation	Pregnancy	Infancy	Young Adult	Senior
Parent-of-origin (imprinted genes)	Maternal diet	Early life exposure to microbes	Environmental, diet, lifestyle	Age-related changes

http://ib.bioninja.com.au/_Media/methylation-factors_med.jpeg

Methylation during development



In utero environment and methylation



Pregnant mother fed diet supplemented with compounds rich in methyl donor groups

Pregnant mom fed regular mouse food

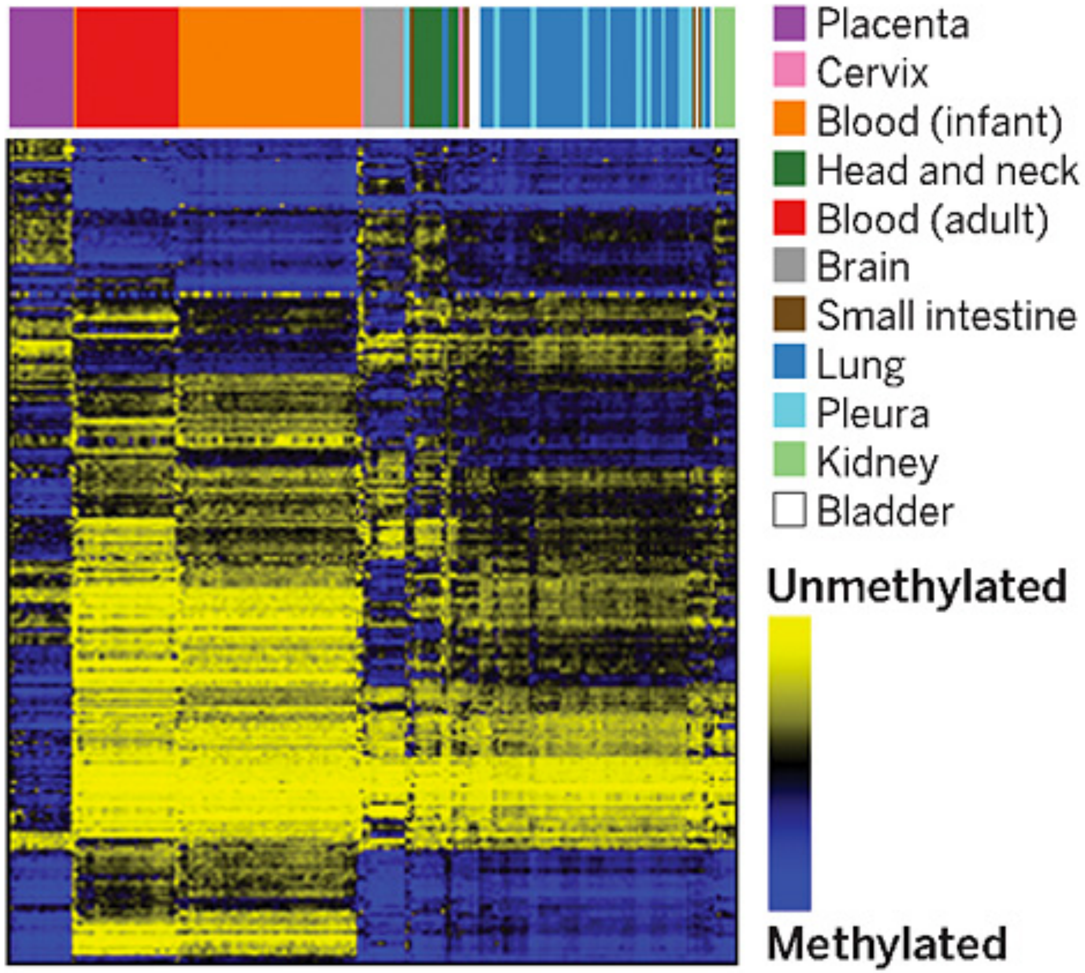
Offspring predominantly brown and in good health

Offspring mainly look like mother and in poorer health

In utero environment and methylation

- Dutch famine study
- The Dutch famine started in November 1944 - May 1945.
- Rations were as low as 400-800 calories a day; less than a quarter of the recommended adult caloric intake.
- Babies whose mothers went through the Dutch famine
    - lower birth weights
    - increased risk of cardiovascular diseases and other adverse health outcomes in adulthood

Methylation is tissue and cell-specific



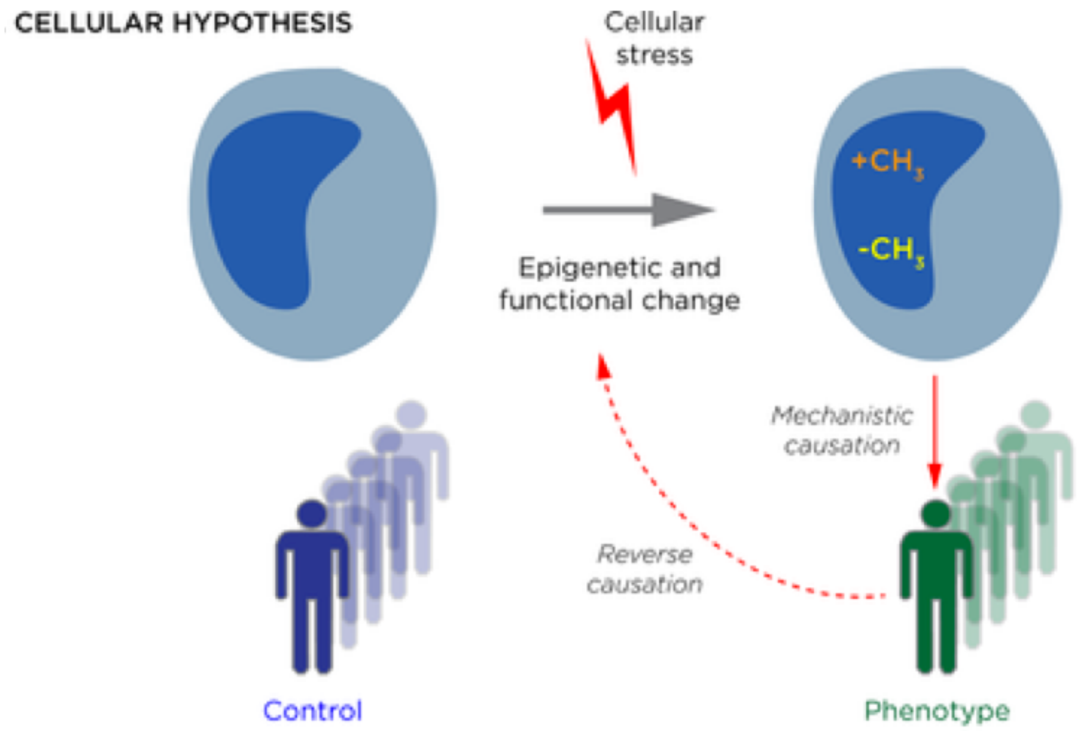
Most studies done in blood due to ease of sample collection

Methylation is tissue and cell-specific

- Any tissue suitable if the epigenetic variation is present soma-wide (e.g. if induced during developmental reprogramming in early embryogenesis).
- If changes that occur later in life, alternative tissue sources need to be explored
- Tissue heterogeneity - tissues are composed of multiple cell types (e.g. blood contains >50 distinct cell types).
- Disease state itself can also alter cell composition in a tissue (e.g. inflamed tissue vs non-inflamed tissue)

Methylation can be causal or consequential

- 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



Birney et al PLOS Genetics 2016

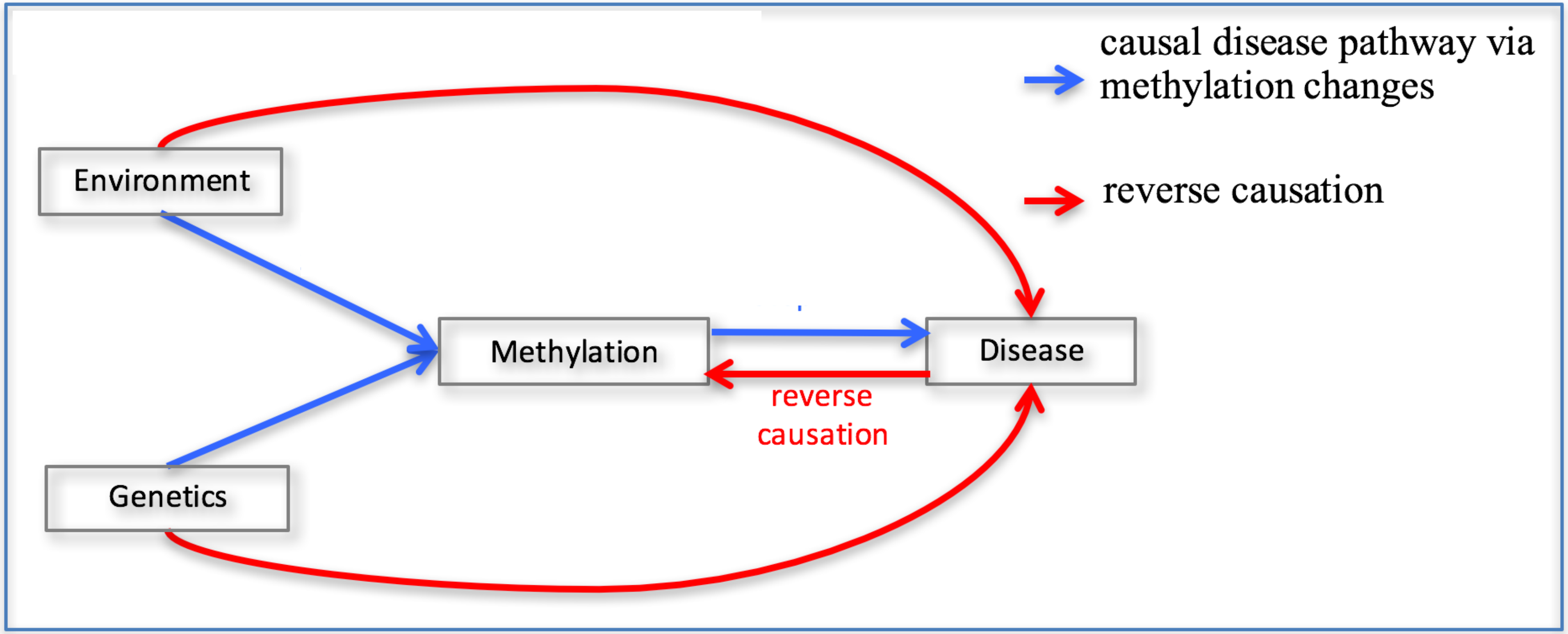
Confounding in EWAS

- Methylation may be affected by many confounding factors:
    - Environmental exposures e.g. smoking
    - Batch effects
    - Ascertainment bias
    - Population stratification
        - Could adjust for PCs generated from GWAS data if available on the same EWAS samples
- Methods such as SVA and PCA can adjust for known/unknown confounders

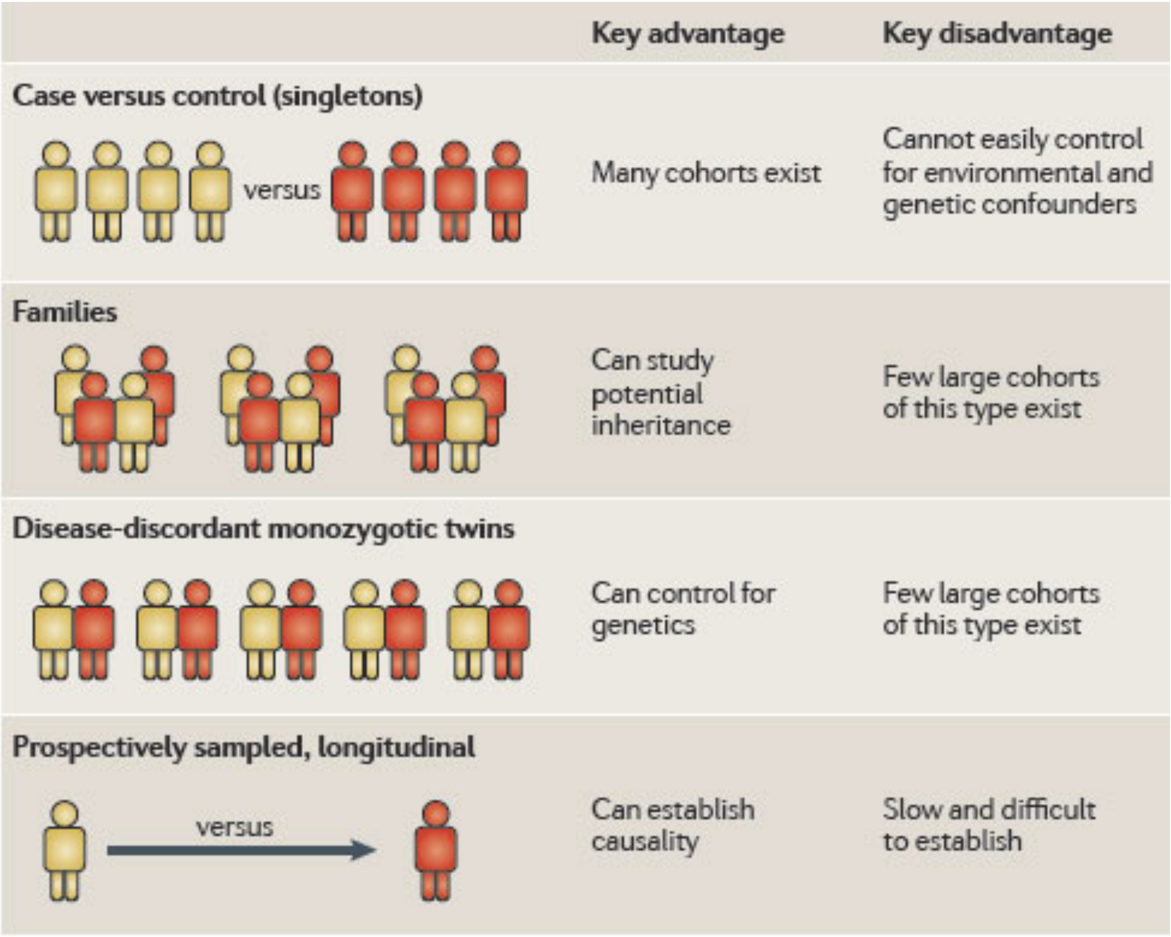
Genetic variants also affect on methylation

- McRae et al. 2013 *Genome Biology*
- Investigate the role of genetic heritability in the similarity of DNA methylation between generations
- Family based sample of 614 individuals from 117 families consisting of twin pairs, their parents and siblings
- After removing all probes overlapping SNPs (1000G EUR) average genetic heritability was 0.187
- Approximately 20% of individual differences in DNA methylation in the population are caused by DNA sequence variation that is not located within CpG sites
- SNPs associated with methylation levels of top heritable probes (mQTLs)

Methylation is dynamic



Study design



	Key advantage	Key disadvantage
Case versus control (singletons) versus	Many cohorts exist	Cannot easily control for environmental and genetic confounders
Families	Can study potential inheritance	Few large cohorts of this type exist
Disease-discordant monozygotic twins	Can control for genetics	Few large cohorts of this type exist
Prospectively sampled, longitudinal versus	Can establish causality	Slow and difficult to establish

Nature Reviews | Genetics

Rakyan et al Nat Rev Gen 2011

Study design

- Investigating causal effect of environmental exposure on disease outcome
  - 2-step design
  - EWAS of environmental exposure in healthy individuals to identify changes in methylation as a consequence of exposure
  - Look at whether the above methylation changes are associated with disease in an independent sample.
- Combine study designs e.g. a discordant monozygotic-twin stage followed by a longitudinal cohort stage.

Study design

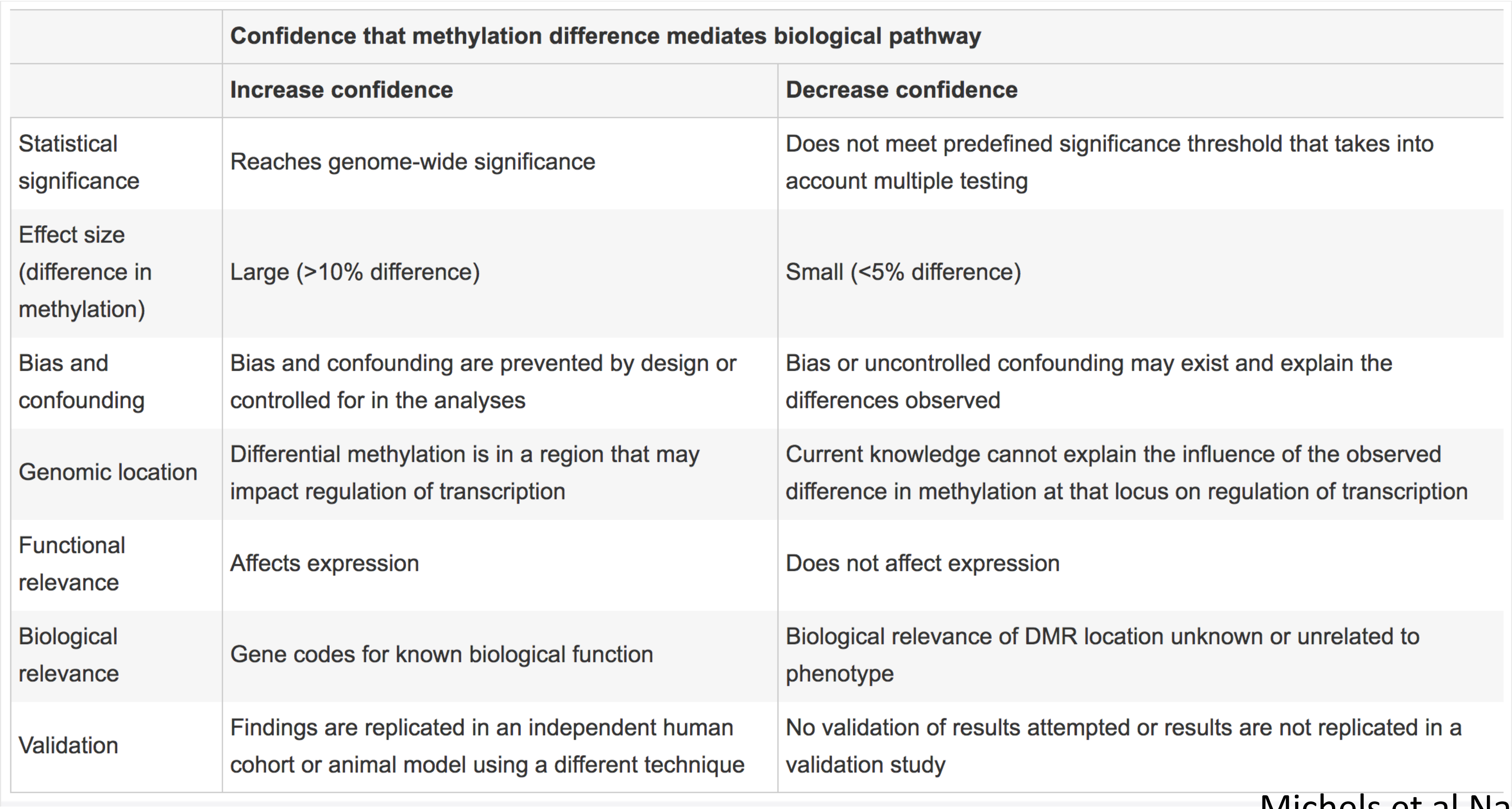
- Clearly define hypothesis
  - Understanding mechanism of disease – mediating cell type with high purity
  - Identify biomarker of exposure or predictive/prognosis – use of an accessible cell type/biological sample
- Can the study design answer this hypothesis
- Understand any cell heterogeneity in your sample
- Effect size should be evaluated in the context of functional and biological relevance. E.g. is a methylation difference of 1% large enough to have an impact on disease?
- Integrate data with genetic and transcriptomic data on same individuals to determine causality

Validation

- Technical validation using different technology - single locus–specific methylation techniques such as bisulfite (pyro)sequencing
  - ruling out technical errors such as cross-hybridising probes or unrecognised SNPs

- Biological validation of EWAS findings - replicating study results in comparable but independent sample

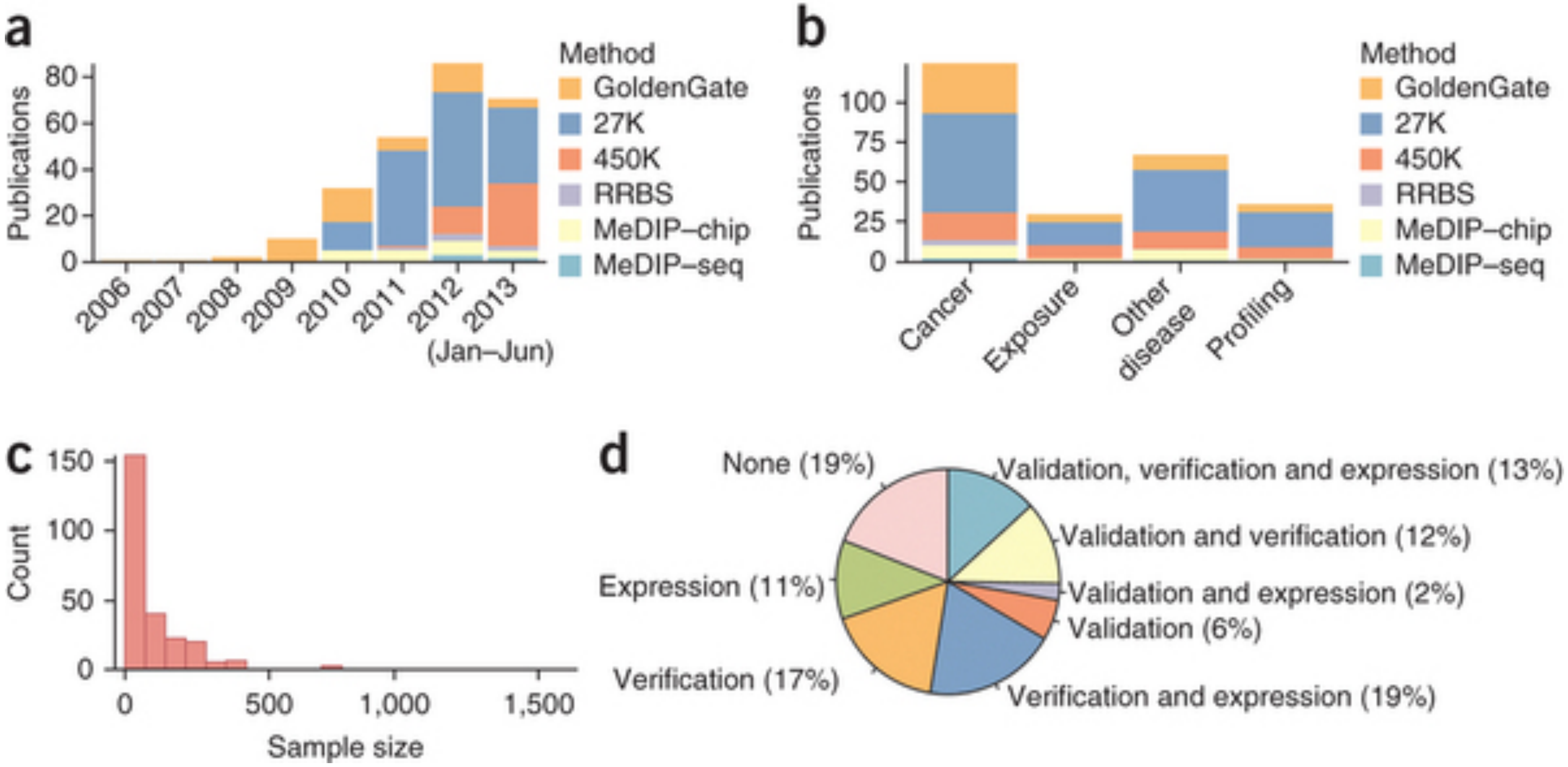
Criteria for identification of 'driver' methylation changes



	Confidence that methylation difference mediates biological pathway	
	Increase confidence	**Decrease confidence**
Statistical significance	Reaches genome-wide significance	Does not meet predefined significance threshold that takes into account multiple testing
Effect size (difference in methylation)	Large (>10% difference)	Small (<5% difference)
Bias and confounding	Bias and confounding are prevented by design or controlled for in the analyses	Bias or uncontrolled confounding may exist and explain the differences observed
Genomic location	Differential methylation is in a region that may impact regulation of transcription	Current knowledge cannot explain the influence of the observed difference in methylation at that locus on regulation of transcription
Functional relevance	Affects expression	Does not affect expression
Biological relevance	Gene codes for known biological function	Biological relevance of DMR location unknown or unrelated to phenotype
Validation	Findings are replicated in an independent human cohort or animal model using a different technique	No validation of results attempted or results are not replicated in a validation study

Michels et al Nat. Methods 2013

Summary of EWAS publications

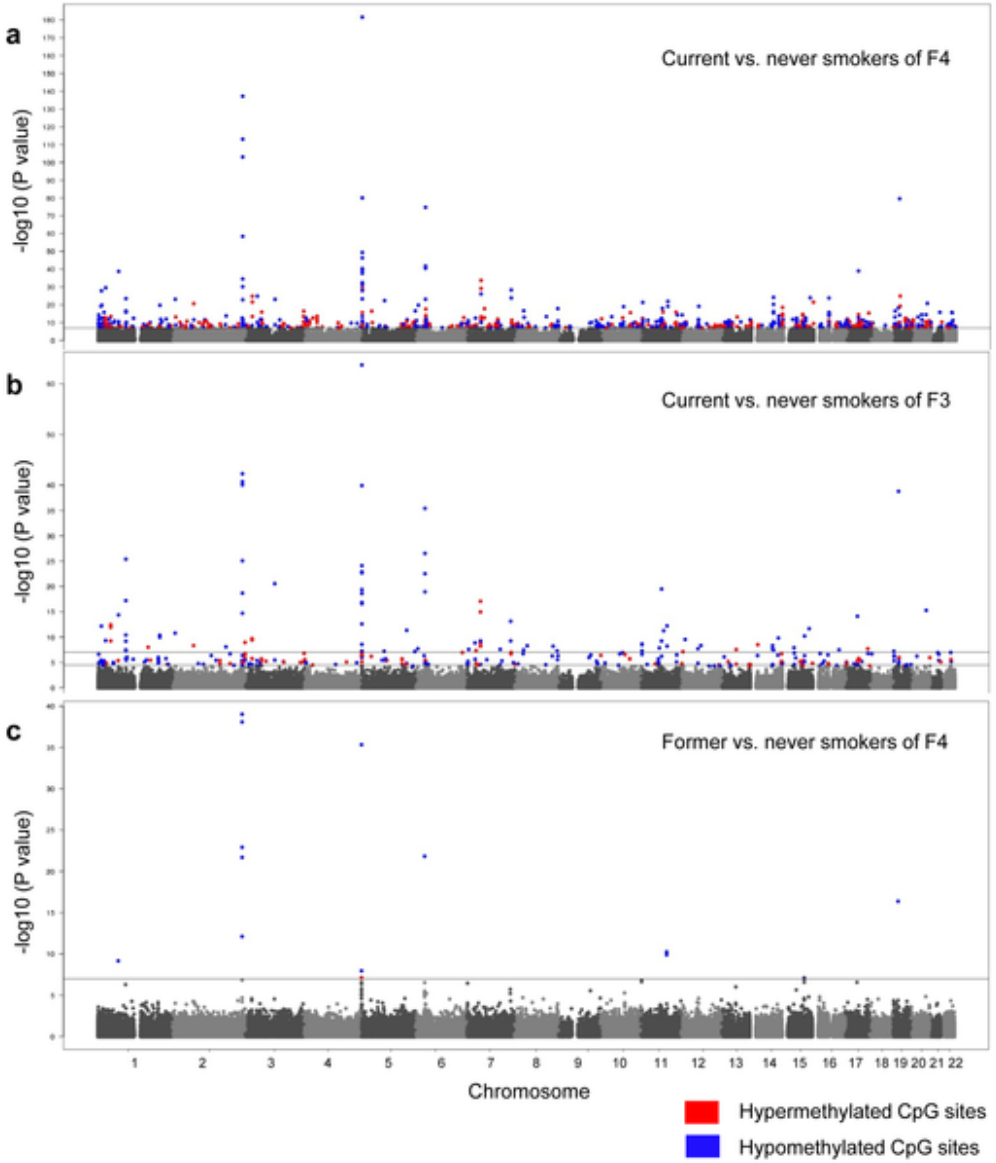


Michels et al Nat. Methods 2013

Example 1: Smoking

- Zeilinger et al
- 450K array
- Discovery sample: discovery (current N=262, never N=749)
- Replication (current N=236, never N=232)
- 972 CpG sites with differential methylation levels after Bonferroni correction (p≤1E-07)
- 187 CpG sites replicated

Smoking EWAS



Top hit
cg05575921

Effect in current smokers vs never smokers

- Discovery:

–24.40%, p=2.54E-182, explained variance=41.02%;

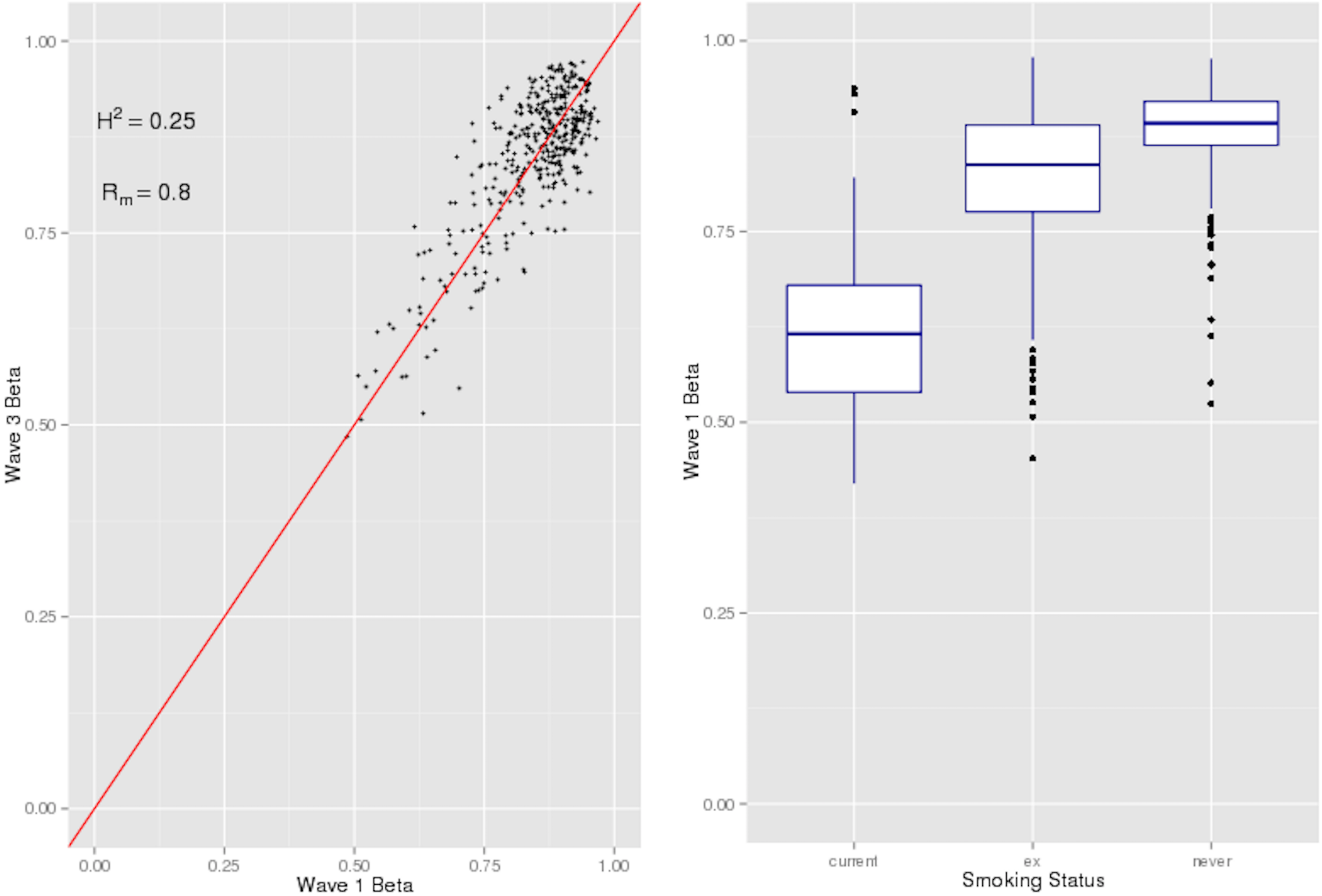
- Replication:

–23.29%, p=1.81E-64, explained variance=39.69%),

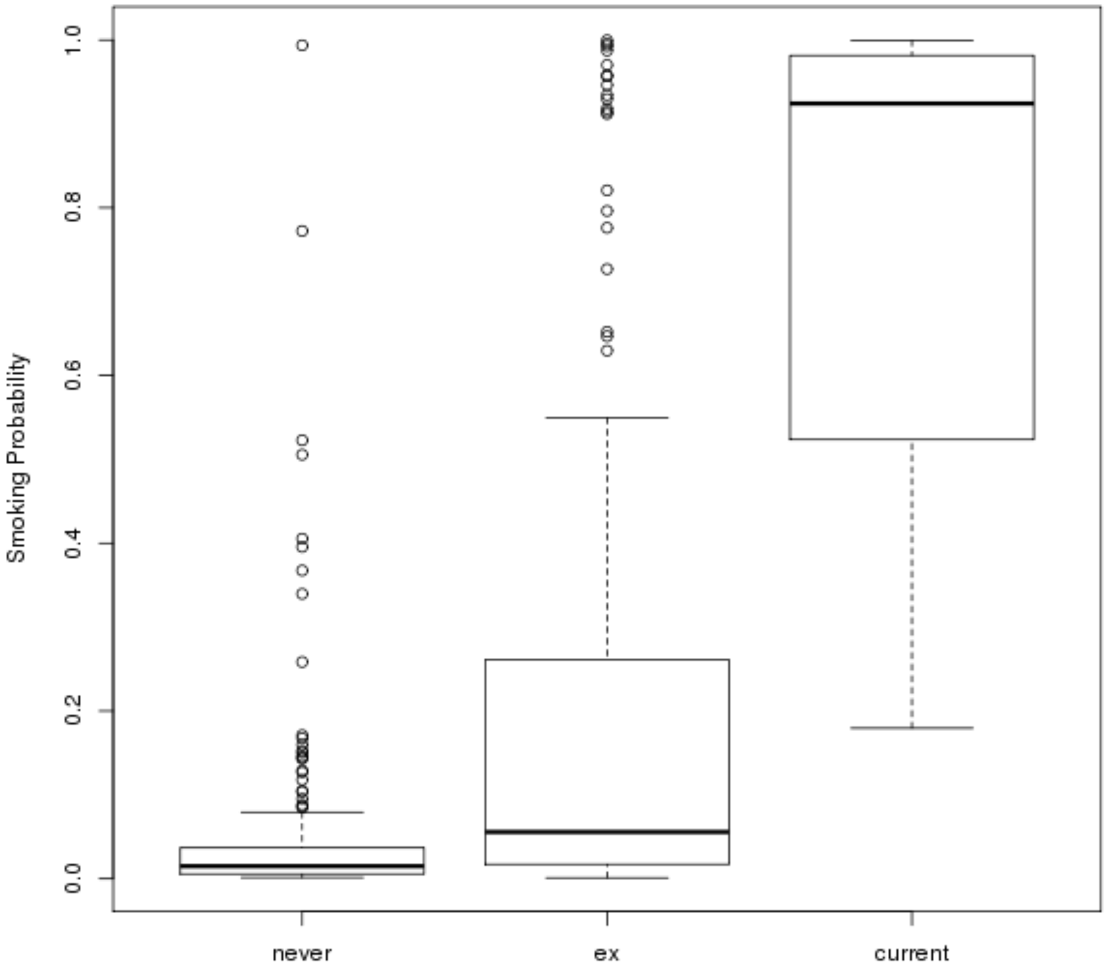
located within the *AHRR* gene (chr5)

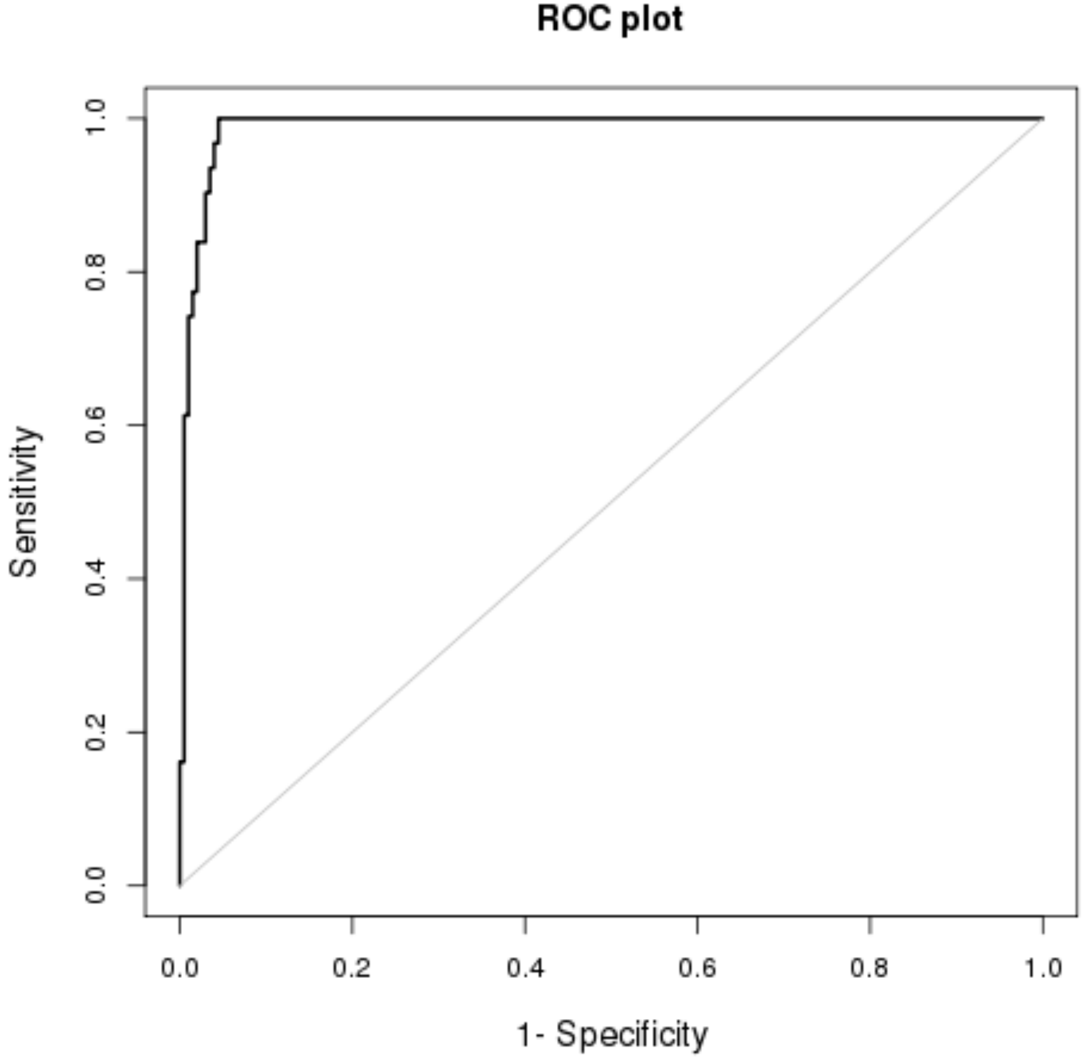
Zeilinger et al. PLOS ONE 2013

cg05575921 methylation levels



Prediction of smoking status





Longitudinal analysis of smoking

Two distinct classes of CpG sites identified:

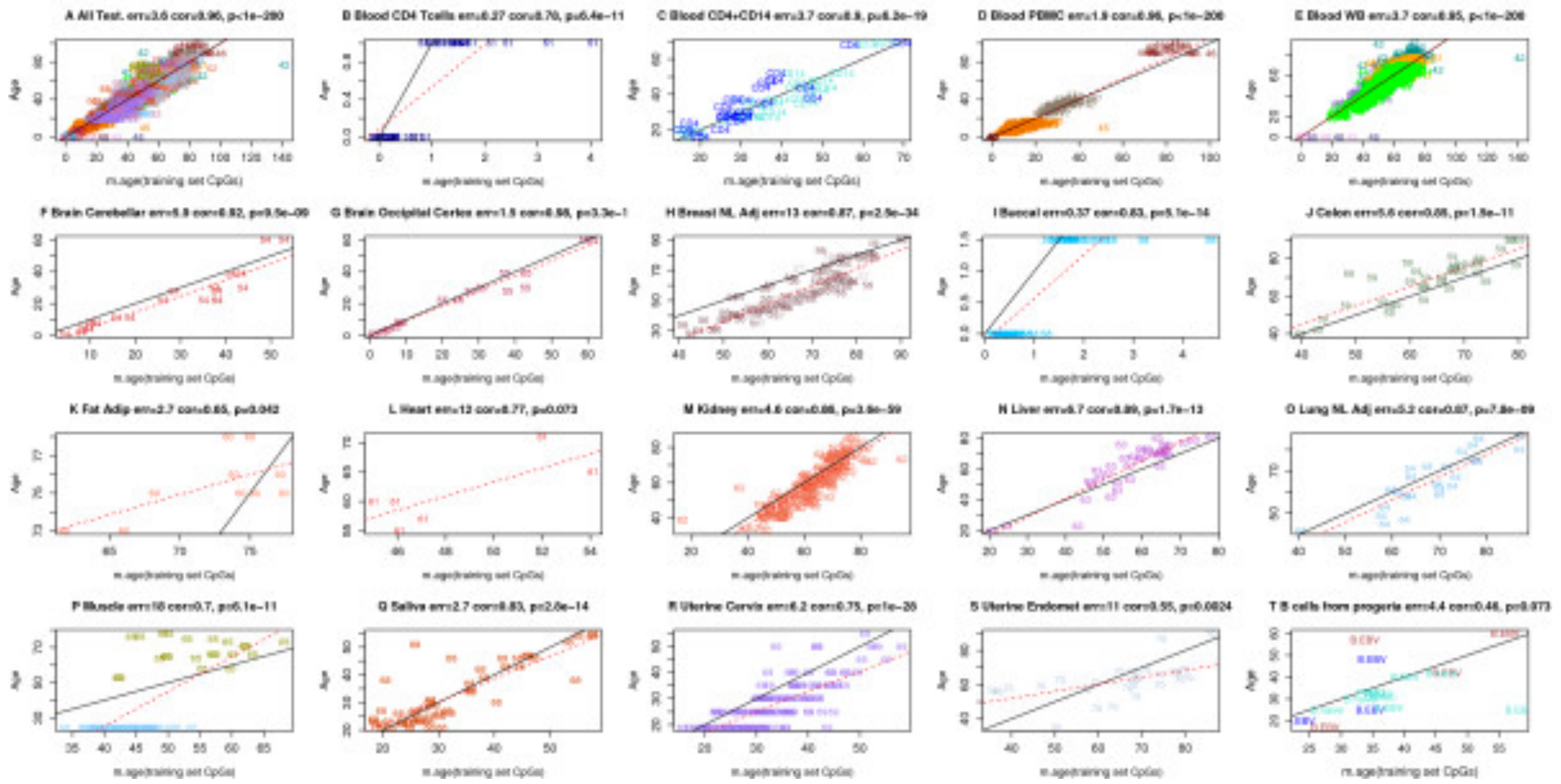
- sites whose methylation reverts to levels typical of never smokers within decades after smoking cessation

- sites remaining differentially methylated, even more than 35 years after smoking cessation.

Example 2: Age

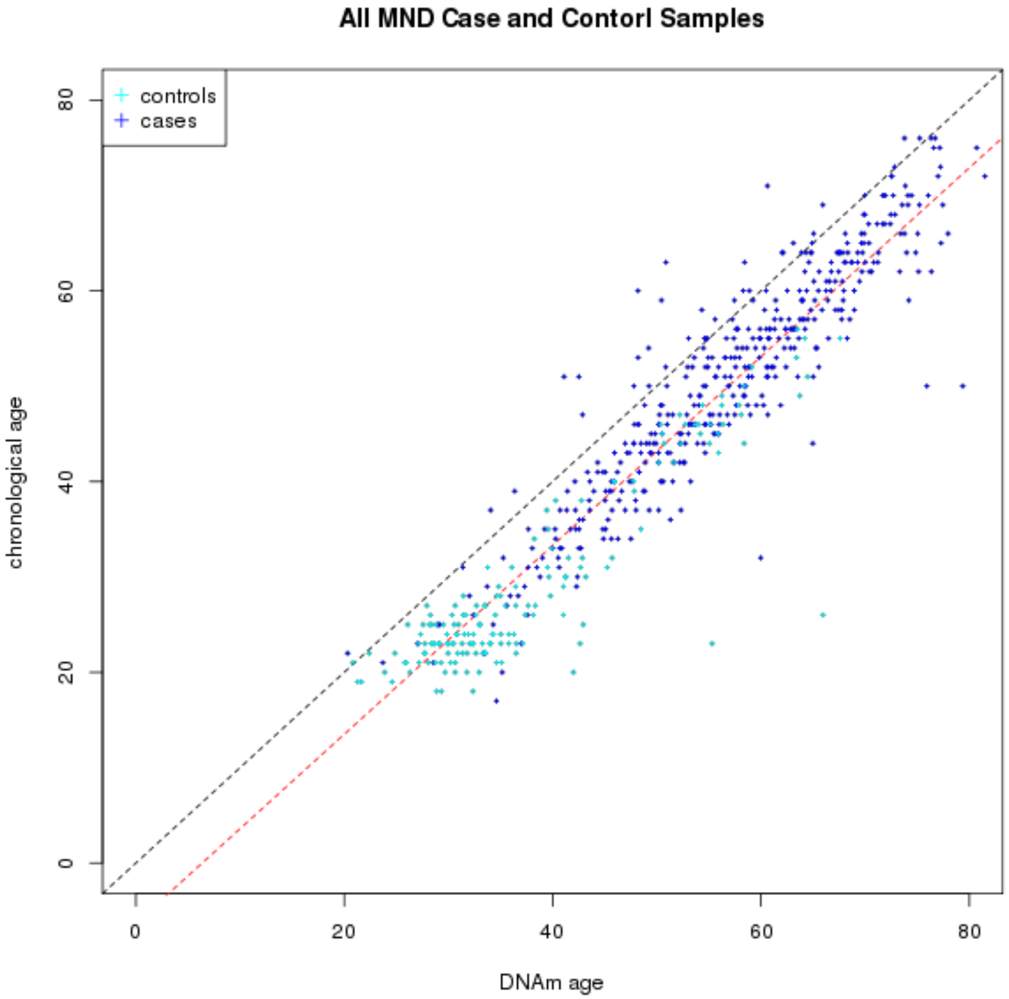
- Horvath *Genome Biology* 2013
- Identify age-associated CpGs in a training set using a penalized regression model (elastic net)
- Identified 353 CpGs
- Predicted age in independent samples and multiple tissues

Example 2: Age



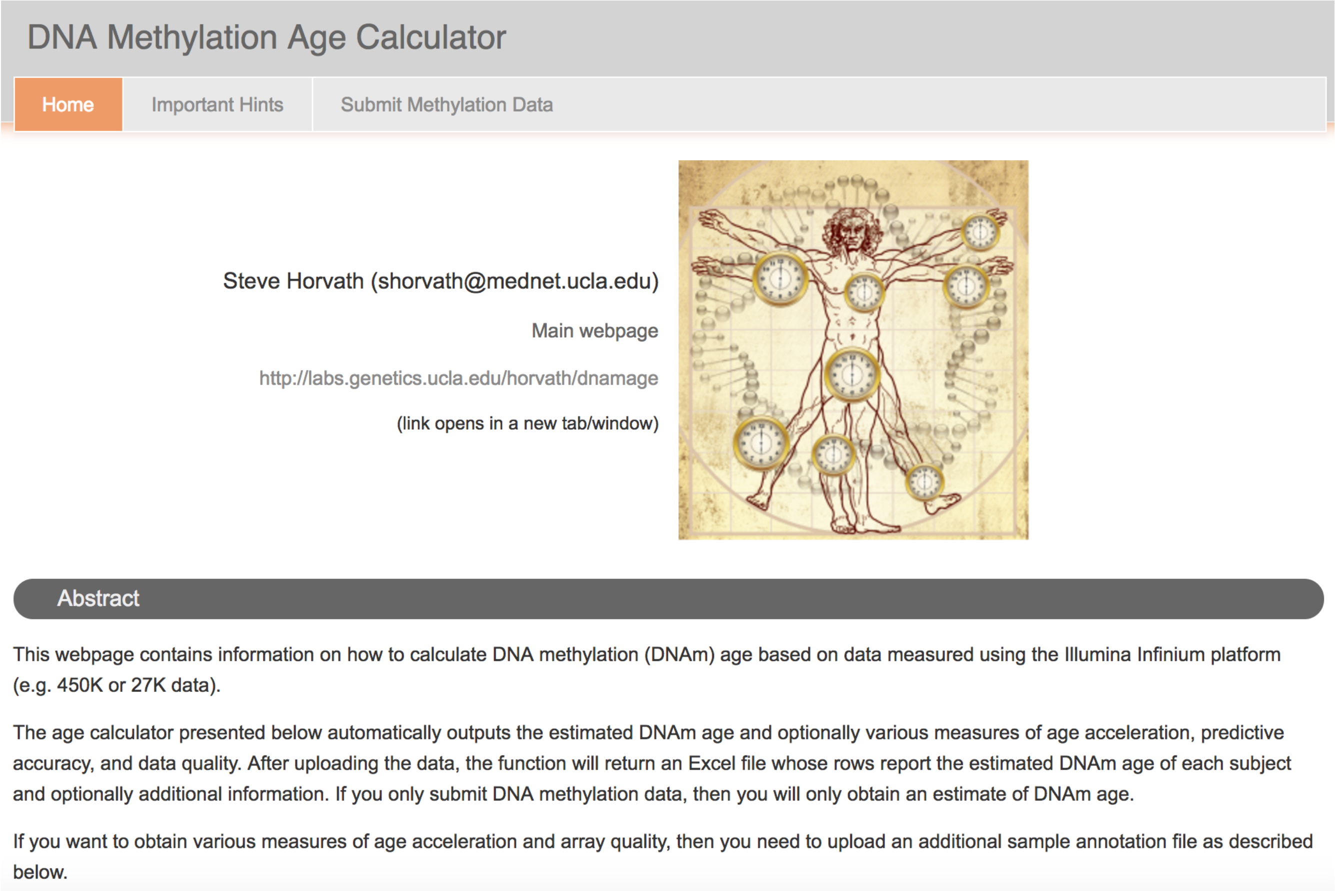
Horvath *Genome Biology* 2013

Prediction of age using Horvath CpGs in a Chinese cohort



Methylation age calculator

- https://labs.genetics.ucla.edu/horvath/dnamage/



Steve Horvath (shorvath@mednet.ucla.edu)

Main webpage

http://labs.genetics.ucla.edu/horvath/dnamage

(link opens in a new tab/window)

Abstract

This webpage contains information on how to calculate DNA methylation (DNAm) age based on data measured using the Illumina Infinium platform (e.g. 450K or 27K data).

The age calculator presented below automatically outputs the estimated DNAm age and optionally various measures of age acceleration, predictive accuracy, and data quality. After uploading the data, the function will return an Excel file whose rows report the estimated DNAm age of each subject and optionally additional information. If you only submit DNA methylation data, then you will only obtain an estimate of DNAm age.

If you want to obtain various measures of age acceleration and array quality, then you need to upload an additional sample annotation file as described below.

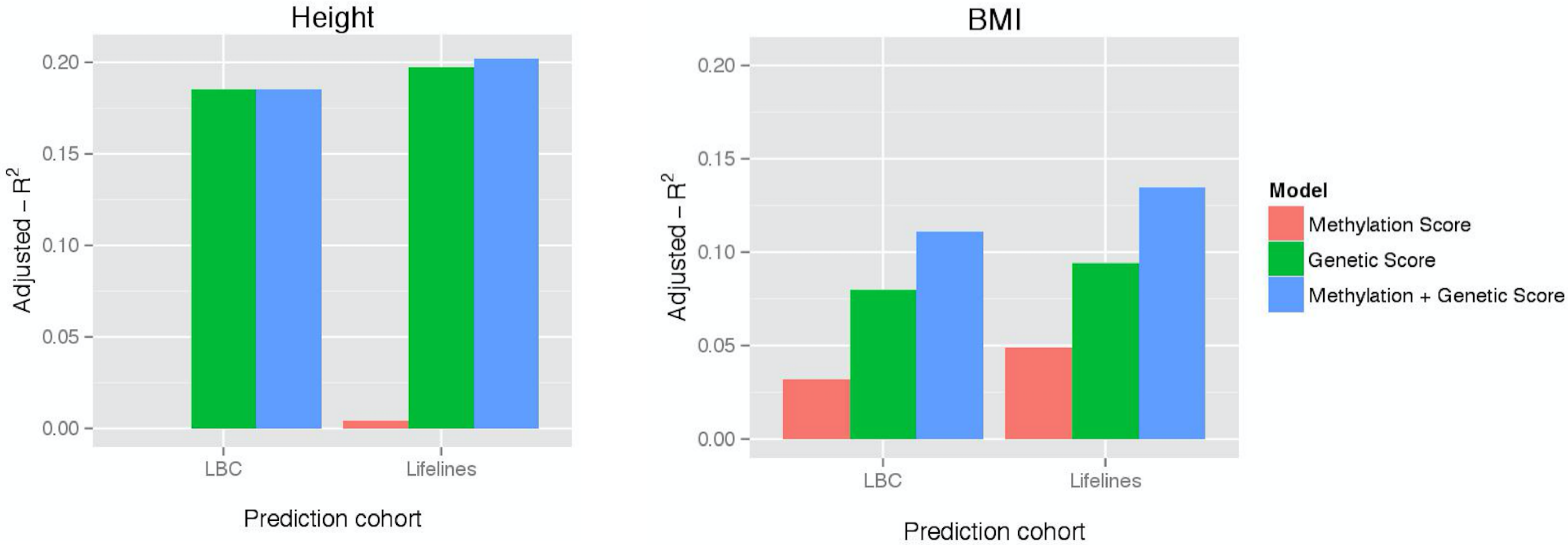
Example 3: BMI and height

- Shah et al *American Journal of Human Genetics* 2014
- Discovery of BMI-associated CpGs in 2 independent samples (LBC and Lifelines)
- Generate genetic risk scores from BMI GWAS SNPs and determine if genetic risk score and methylation risk scores are independently associated with BMI
- Repeat for height.

Methods

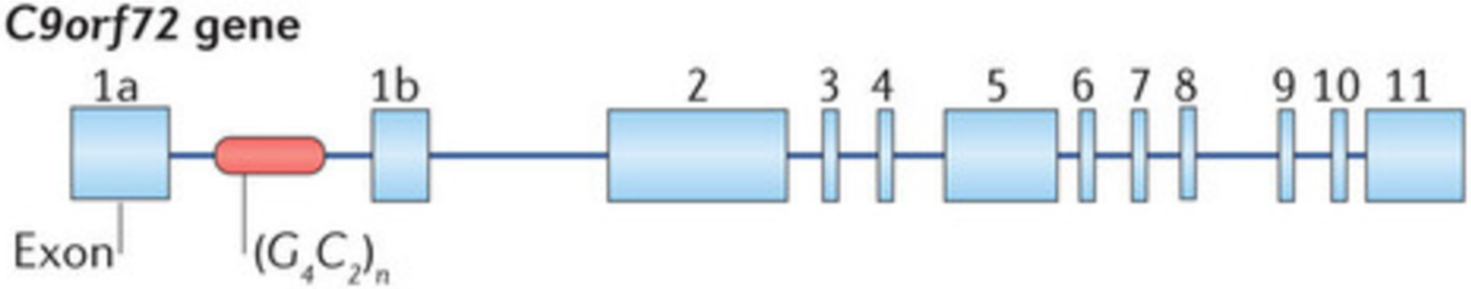
- Study A (population cohort) – EWAS on BMI -> significant probelist A
- Study B (old individuals 70+) – EWAS on BMI -> significant probelist B
- Calculate methylation BMI risk score in study A based on probelist B
- Calculate methylation BMI risk score in study B based on probelist A
- Proportion of variance in BMI explained by methylation score in each study
- Generate genetic scores for BMI in each study using SNPs identified from the largest BMI GWAS (GIANT consortium)
- Look at proportion of variance explained by genetic risk score
- Are methylation and genetic risk scores independently associated with BMI and height

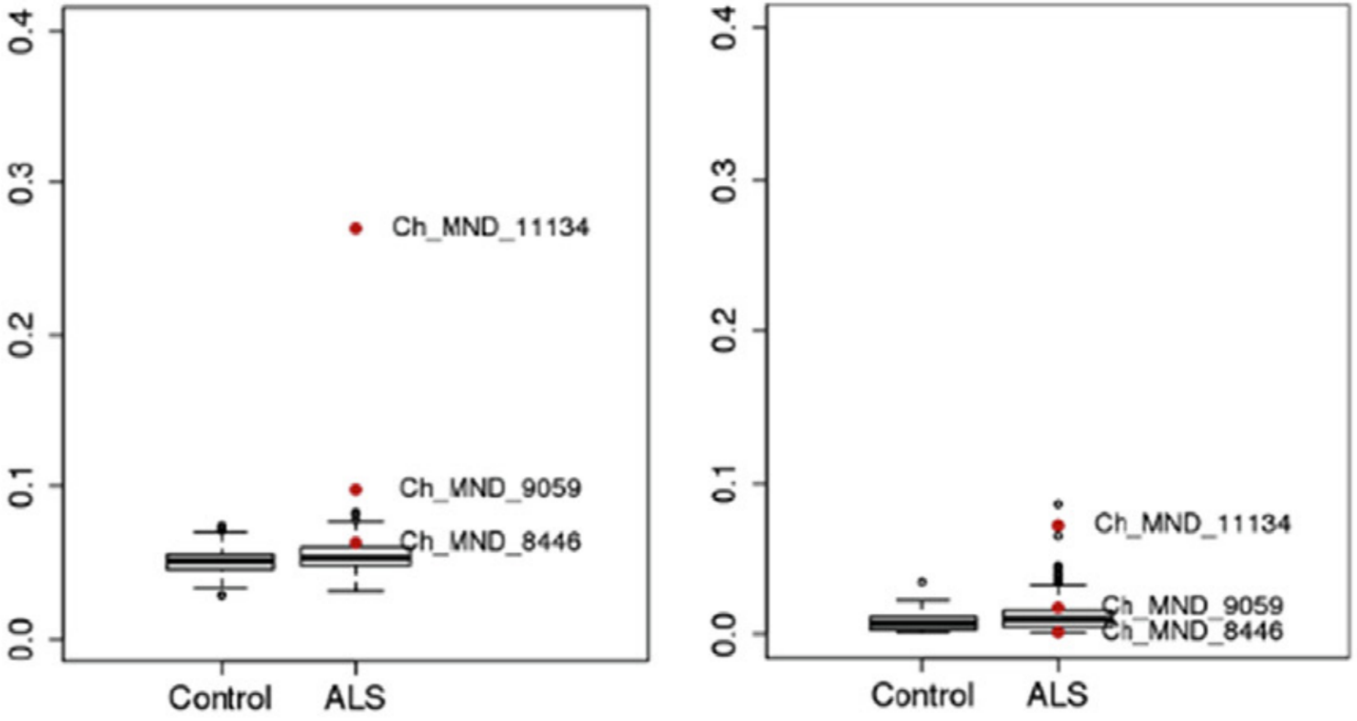
Example 3: BMI and height



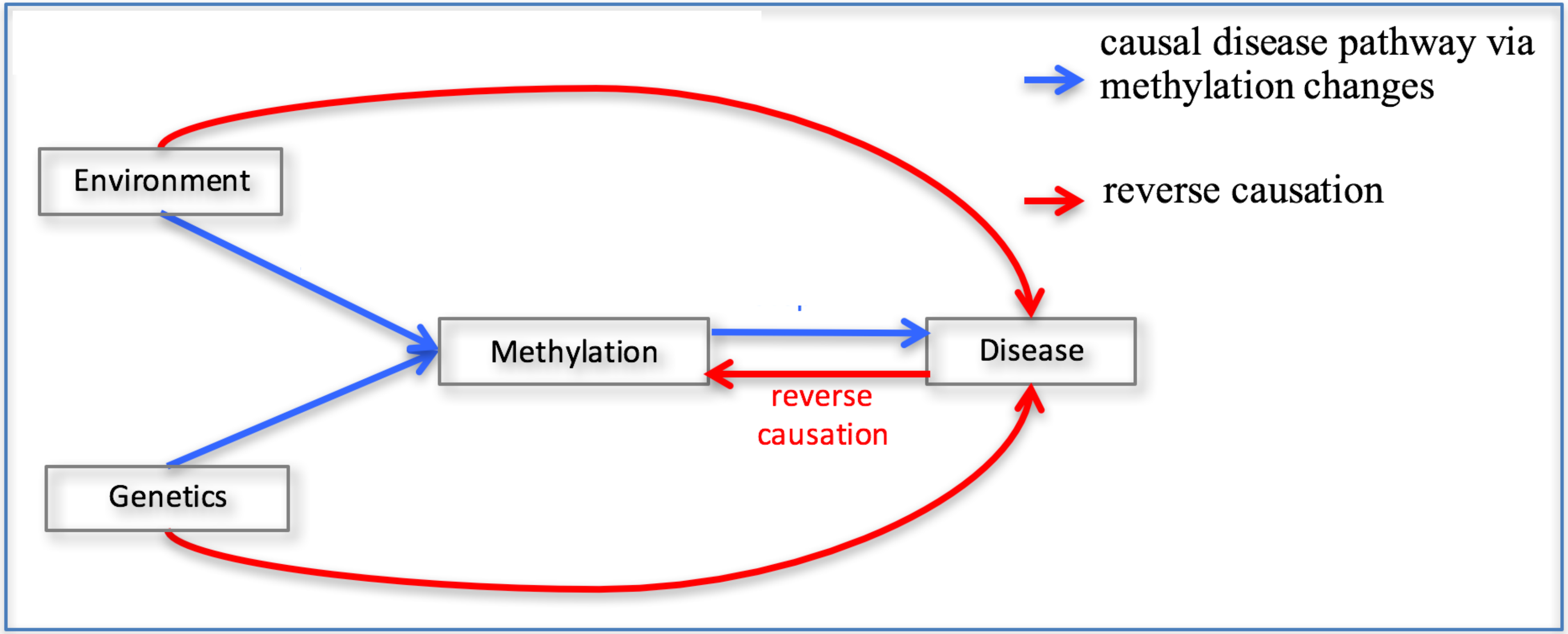
Example 4: C9orf72 repeat expansion

- hexanucleotide repeat expansion GGGGCC
- 1st Intron region of *c9orf72*
- Most common mutation identified that is associated with familial FTD and/or ALS (5–20% of patients with sporadic ALS)
- Length of repeat in cases can occur in the order of 100s and varies
- <30 repeats generally not associated with disease

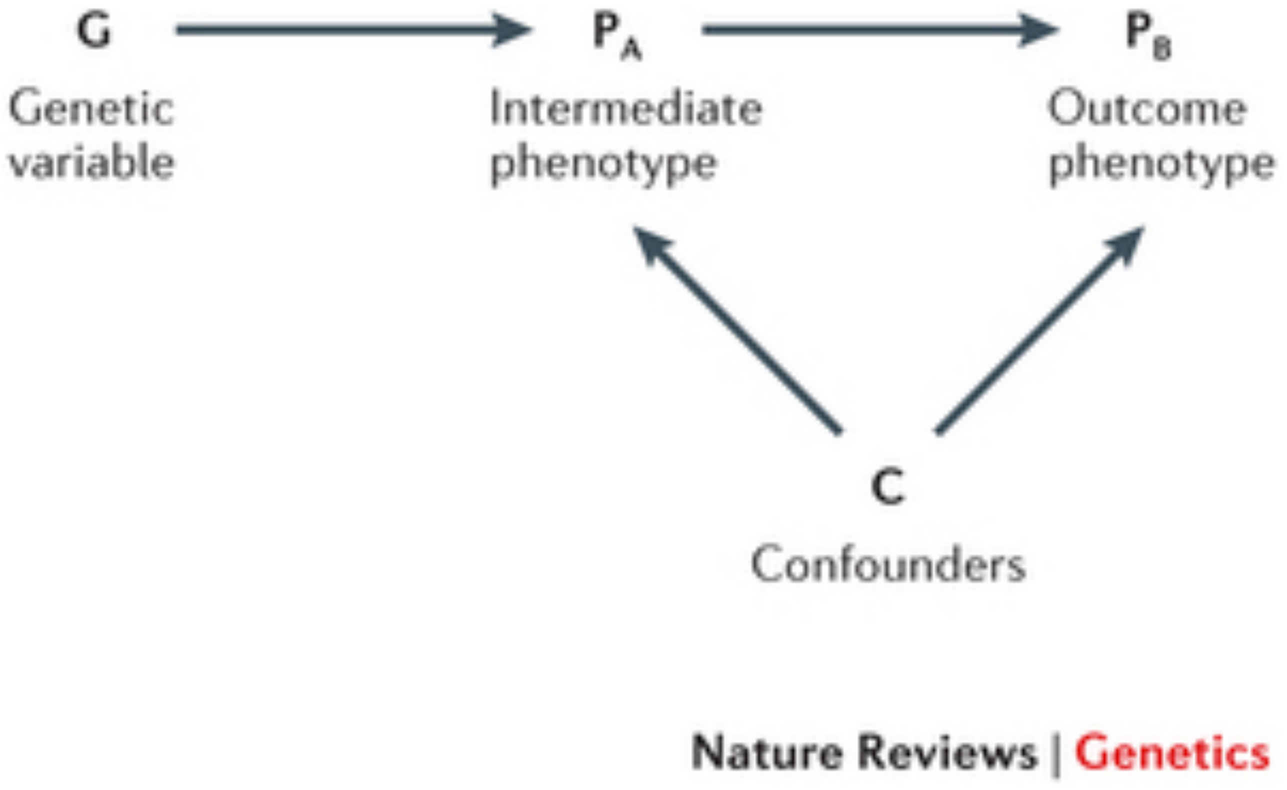




Determining causality



Mendelian randomisation



Nature Reviews | Genetics

G must be associated with intermediate phenotype P_A

G must not be associated with confounders.

G should only be related to the outcome P_B via P_A

Does genotype affect phenotype via changes in methylation?

- Instrumental variable analysis or Mendelian randomisation analysis
- Step 1: Is there a SNP (not in the probe) that is strongly associated with methylation levels (mQTL)
- Step 2: CpGmeth ~ SNP
- Step 3: BMI ~ predicted CpGmeth