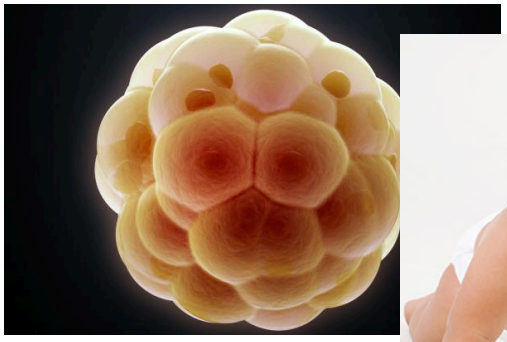
Epigenetics and EWAS

Sylvane Desrivières, PhD
Institute of Psychiatry, Psychology & Neuroscience
King's College London, United Kingdom
sylvane.desrivieres@kcl.ac.uk

Epigenetics drives phenotype

- Epigenetics = literally 'outside conventional genetics'
- The study of changes in gene expression that are '*heritable through cell division*' and that occur without a change in the sequence of the DNA

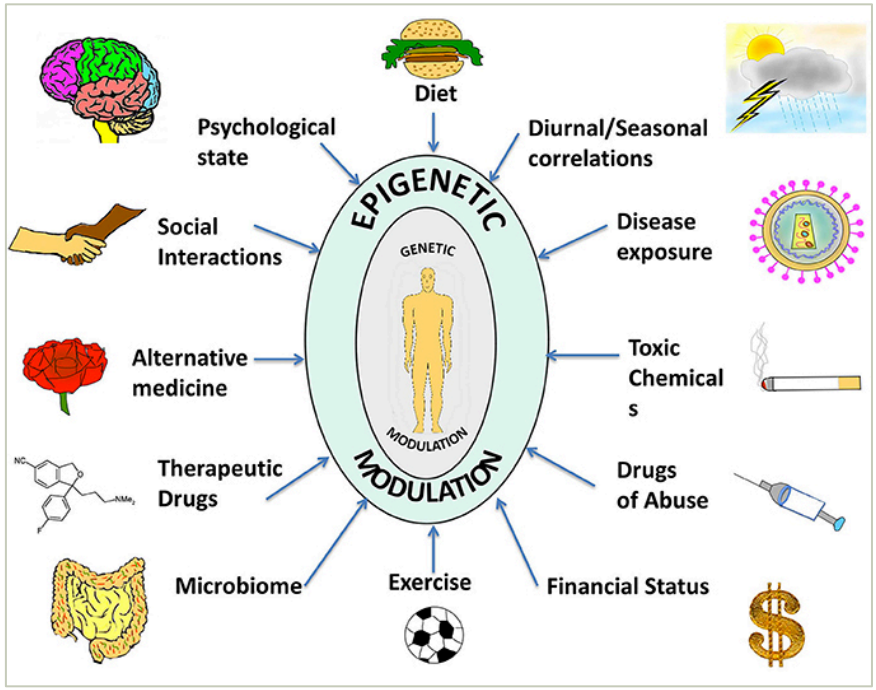
  — **a change in phenotype without a change in genotype**
- Critical for development and differentiation
- Dysregulation of the epigenome is associated with Cancer, Autoimmune diseases, Diabetes & Mental disorders



The Epigenome reacts to the environment

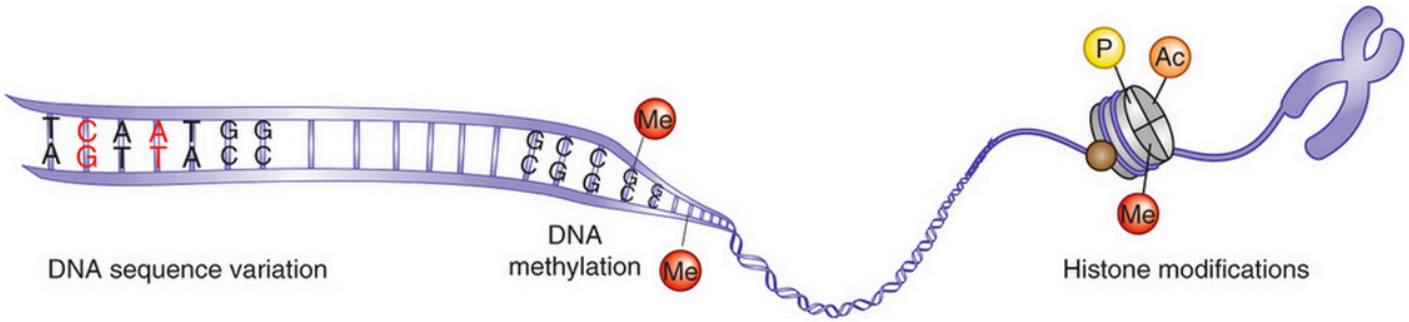
The Epigenome:

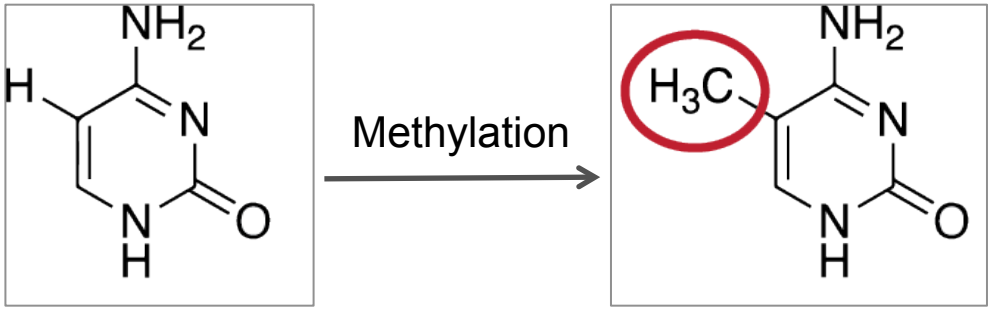
- is characterized by a dynamic response to intra- and extra-cellular stimuli & by environmental and lifestyle factors
- integrates the information encoded in the genome with all the molecular and chemical cues of cellular, extracellular, and environmental origin
- represents the ability of an organism to adapt and evolve in response to environmental stimuli



Front. Cell Dev. Biol., 2014
https://doi.org/10.3389/fcell.2014.00049

DNA Methylation





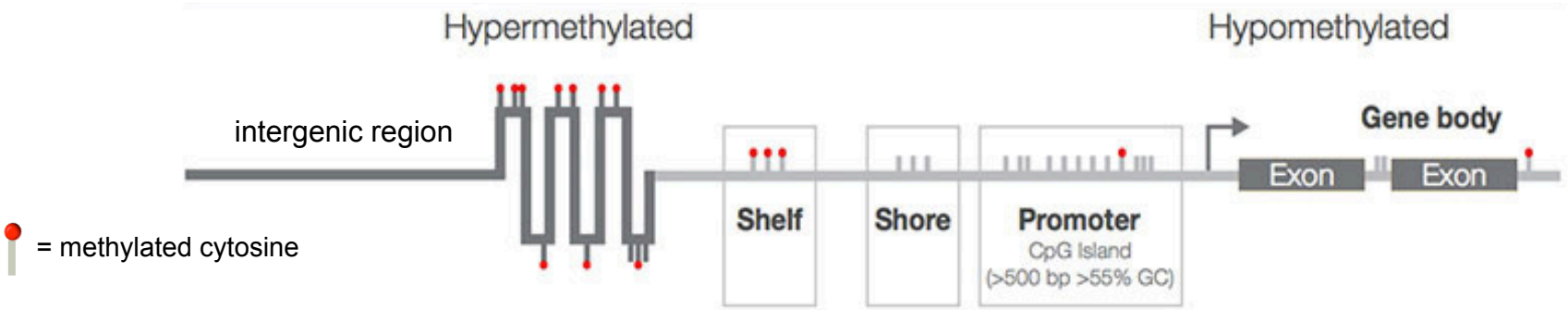
Cytosine **5-Methylcytosine**

- In adult mammals, largely restricted to CpG dinucleotides
- Mediated through the addition of a methyl group to a cytosine base at a CpG site
- Reversible chemical modification

DNAm stabilises the genome & silences gene expression

Non-random genomic distribution of DNAm

- CpGs clusters in promoter regions (CpG islands). Methylation at these sites leads to gene silencing.
- At intergenic regions and repetitive elements, methylation usually adds to genomic stability



If hypomethylated → Genomic instability

If hypermethylated → Transcriptional repression

Disease

Fragile X, Rett syndrome, autism, SCZ, dementia, suicide risk, Alzheimers, Prader-Willi, Angelman, Beckwith-Wiedermann syndromes

GWAS vs. EWAS

Screening for 100Ks to millions of loci in the genome:

- GWAS: Single nucleotide polymorphisms (SNPs)
  - test for association with disease/phenotype
- EWAS: CpG sites
  - test for association with exposure/risk factor
  - test for association with disease/phenotype
- The EWAS field is relatively new
- Several tools are methods are inferred from GWAS, but important considerations specific for EWAS!

EWAS-specific challenges

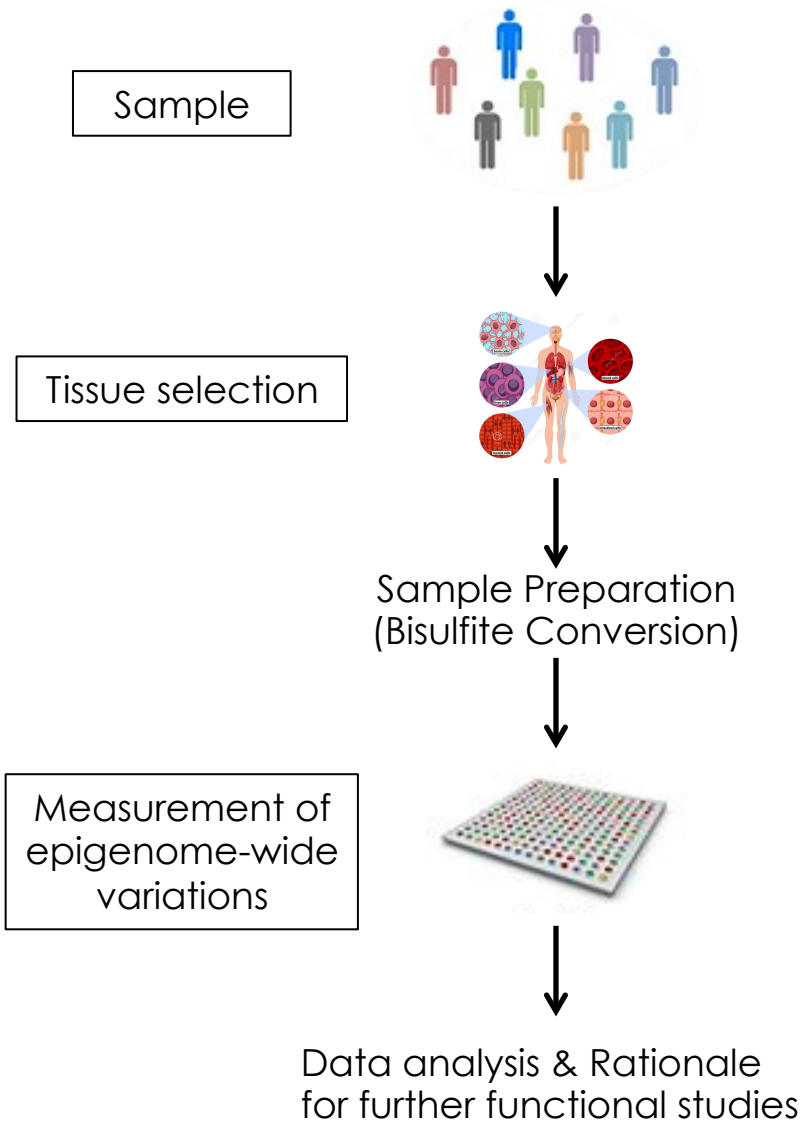
- **Tissue choice**: Disease/Phenotype-relevant vs. accessible
  - Epigenetic variation can be tissue-specific.
  - But most EWAS use blood as a surrogate tissue, due to its availability and ease of collection. Epigenetic changes in the blood may not be found in other tissues.

- **Cell type heterogeneity**
  - Sample may contain different cell types (e.g., blood) each of which have a unique epigenetic signature.
  - **Essential to control for cell type diversity** to make sure that variation in epigenetic marks are due to the differences in phenotype rather than due to the sample heterogeneity.

- **Sample size and power**

- **Causality?**
  - Variations in the epigenome could be the cause or the consequence of differences in phenotype, and distinguishing between the two is a major limitation in DNAm analyses.

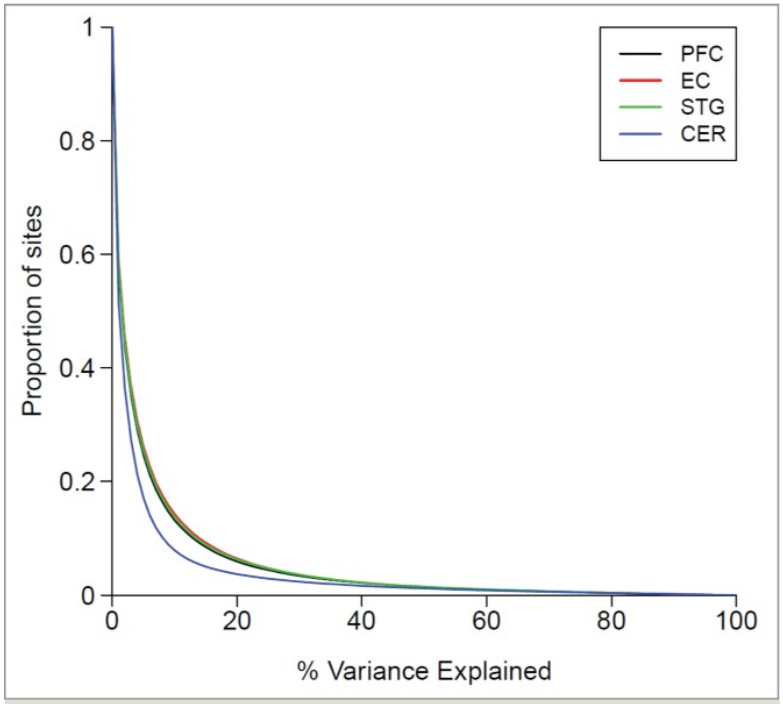
EWAS Workflow



Tissue selection
DNAm in blood as a predictor of DNAm in brain

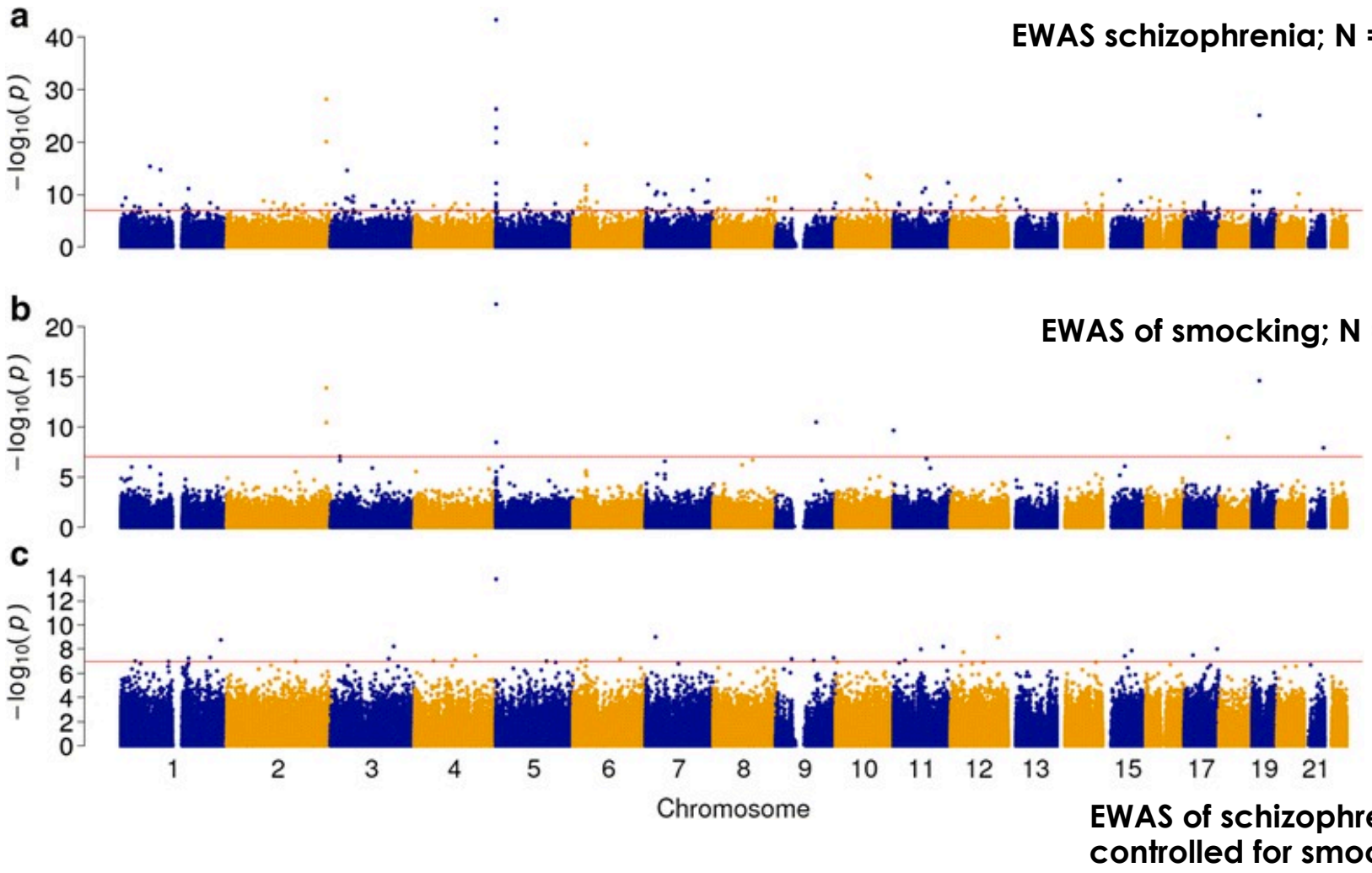
- For most CpG sites, interindividual variation in blood is not a strong predictor of interindividual variation in the brain.
- DNAm variation at a subset of probes strongly correlates across tissues.
- Blood-based EWAS for disorders where brain is presumed to be the primary tissue of interest,
  - May give limited information relating to underlying pathological processes.
  - May be used to identify biomarkers of phenotypes manifest in the brain.

Proportion of CpG sites for which variation in blood explains a certain % of DNAm variance in brain tissues from the same individuals



Epigenetics. 2015; 10: 1024–1032.

Tissue selection
Blood DNAm & Biomarkers for Mental Health



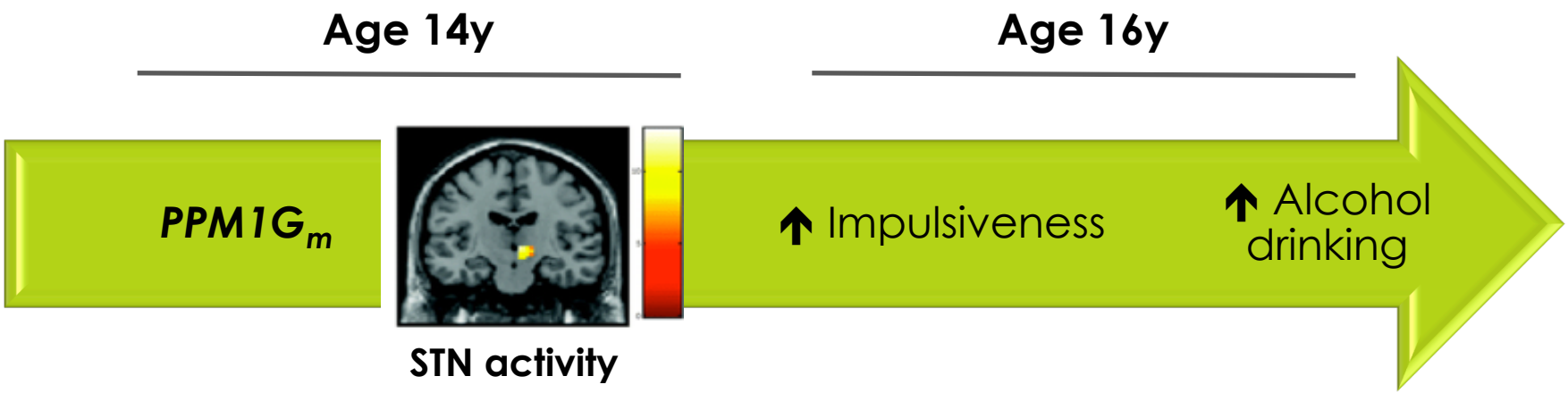
EWAS schizophrenia; N = 1714

EWAS of smocking; N = 192

EWAS of schizophrenia controlled for smocking

Genome Biol. 2016; 17: 176

Blood DNAm as predictor of brain function and behaviour



N=499

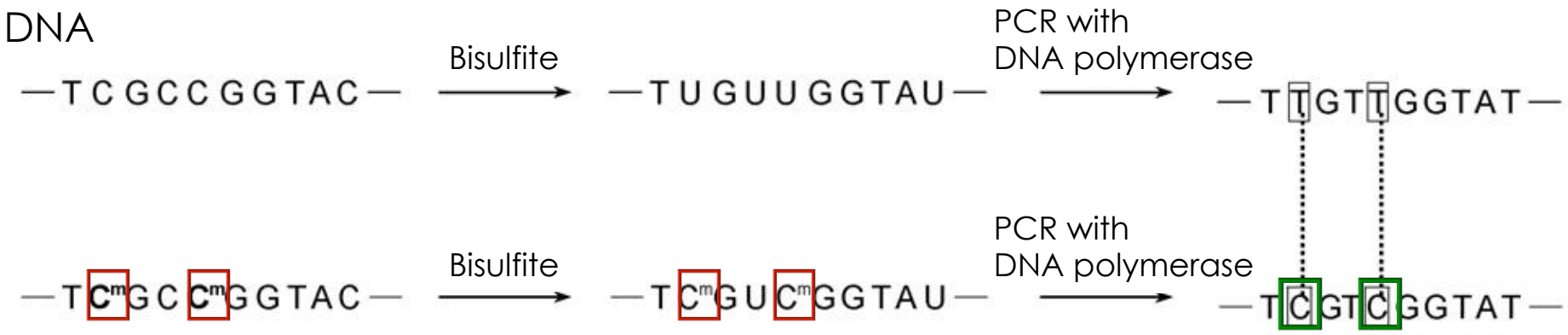
Association of Protein Phosphatase *PPM1G* With Alcohol Use Disorder and Brain Activity During Behavioral Control in a Genome-Wide Methylation Analysis

Barbara Ruggeri, Ph.D., Charlotte Nymberg, Ph.D., Eero Vuoksimaa, Ph.D., Anbarasu Lourdusamy, Ph.D., Cybele P. Wong, Ph.D., Fabiana M. Carvalho, Ph.D., Tianye Jia, Ph.D., Anna Cattrell, Ph.D., Christine Macare, M.Sc., Tobias Banaschewski, M.D., Ph.D., Gareth J. Barker, Ph.D., Arun L.W. Bokde, Ph.D., Uli Bromberg, M.Sc., Christian Büchel, M.D., Ph.D., Patricia J. Conrod, Ph.D., Mira Fauth-Bühler, Ph.D., Herta Flor, Ph.D., Vincent Frouin, Ph.D., Jürgen Gallinat, M.D., Ph.D., Hugh Garavan, Ph.D., Penny Gowland, Ph.D., Andreas Heinz, M.D., Ph.D., Bernd Ittermann, Ph.D., Jean-Luc Martinot, M.D., Ph.D., Frauke Nees, Ph.D., Zdenka Pausova, M.D., Ph.D., Tomáš Paus, M.D., Ph.D., Marcella Rietschel, Ph.D., Trevor Robbins, Ph.D., Michael N. Smolka, M.D., Ph.D., Rainer Spanagel, Ph.D., Georgy Bakalkin, Ph.D., Jonathan Mill, Ph.D., Wolfgang H. Sommer, Ph.D., Richard J. Rose, Ph.D., Jia Yan, Ph.D., Fazil Aliev, Ph.D., Danielle Dick, Ph.D., Jaakko Kaprio, M.D., Ph.D., Sylvane Desrivières, Ph.D., Gunter Schumann, M.D., the IMAGEN Consortium

Am J Psychiatry. 2015;172:543-52

Sample Preparation for DNAm analyses
Bisulfite Conversion

Converting non-methylated cytosines [**C**] **to** uracil [**U**]



- Harsh Reaction Conditions (low pH/high temperature), which can degrade DNA
- Experimental conditions (pH, temperature and incubation time) are important considerations for preserving DNA quality which impact downstream analyses
- Commercial kits have improved protocols that often yield less fragmented DNA compared to earlier methods

Epigenome-wide DNAm profiling

Wide range of techniques used to study DNAm post-bisulfite conversion

- Methylation Specific Restriction Enzymes
- PCR Techniques (e.g., Bisulfite Specific PCR)
- Sequencing (High-throughput, polymerase or ligase-based, very complicated data analysis)
- **Microarrays** (High-throughput, Hybridization/probe-based)

Infinium® HumanMethylation Arrays:
- 27k BeadChip: ~27k CpGs (14,495 genes)
- **450k BeadChip:** ~450k CpGs. Covers 99% of RefSeq genes
- MethylationEPIC Array: ~850K CpGs.

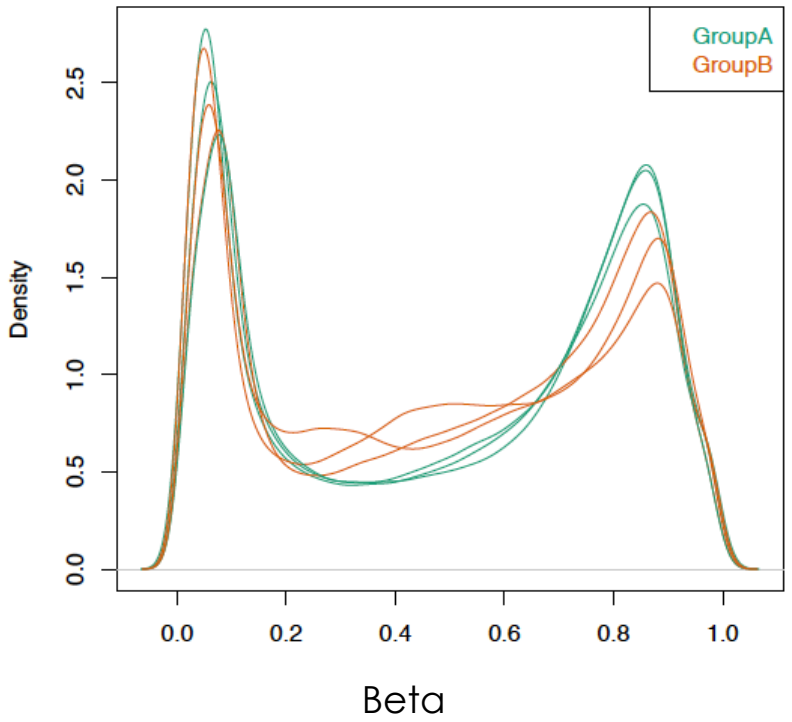
Estimating DNA methylation levels

Beta value (β) = ratio of intensities between methylated and unmethylated alleles

$$\beta = \frac{M}{M + U + 100}$$

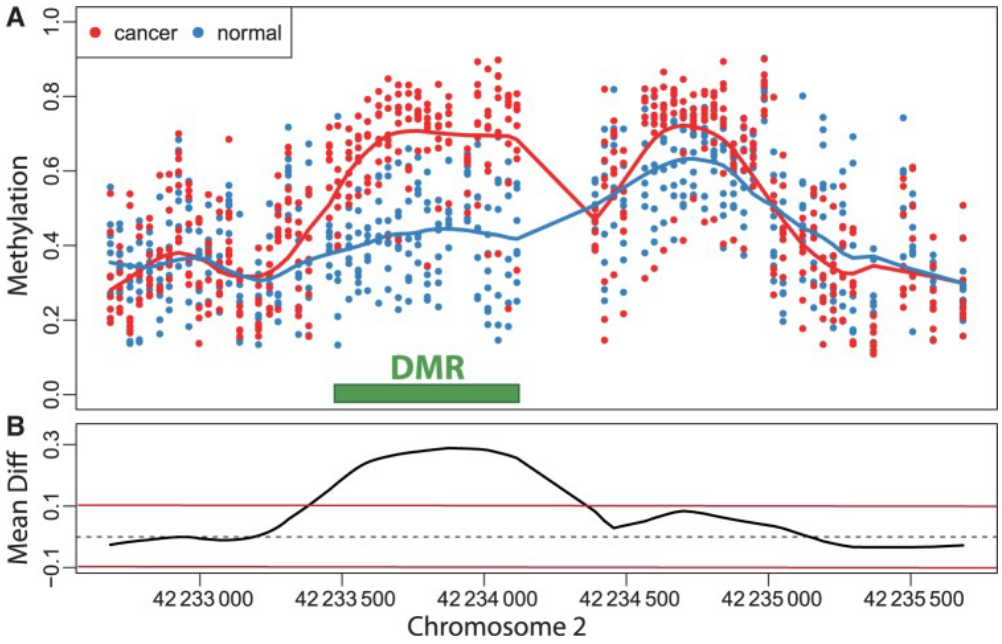
M = methylated signal
U = unmethylated signal

β lies between zero and one



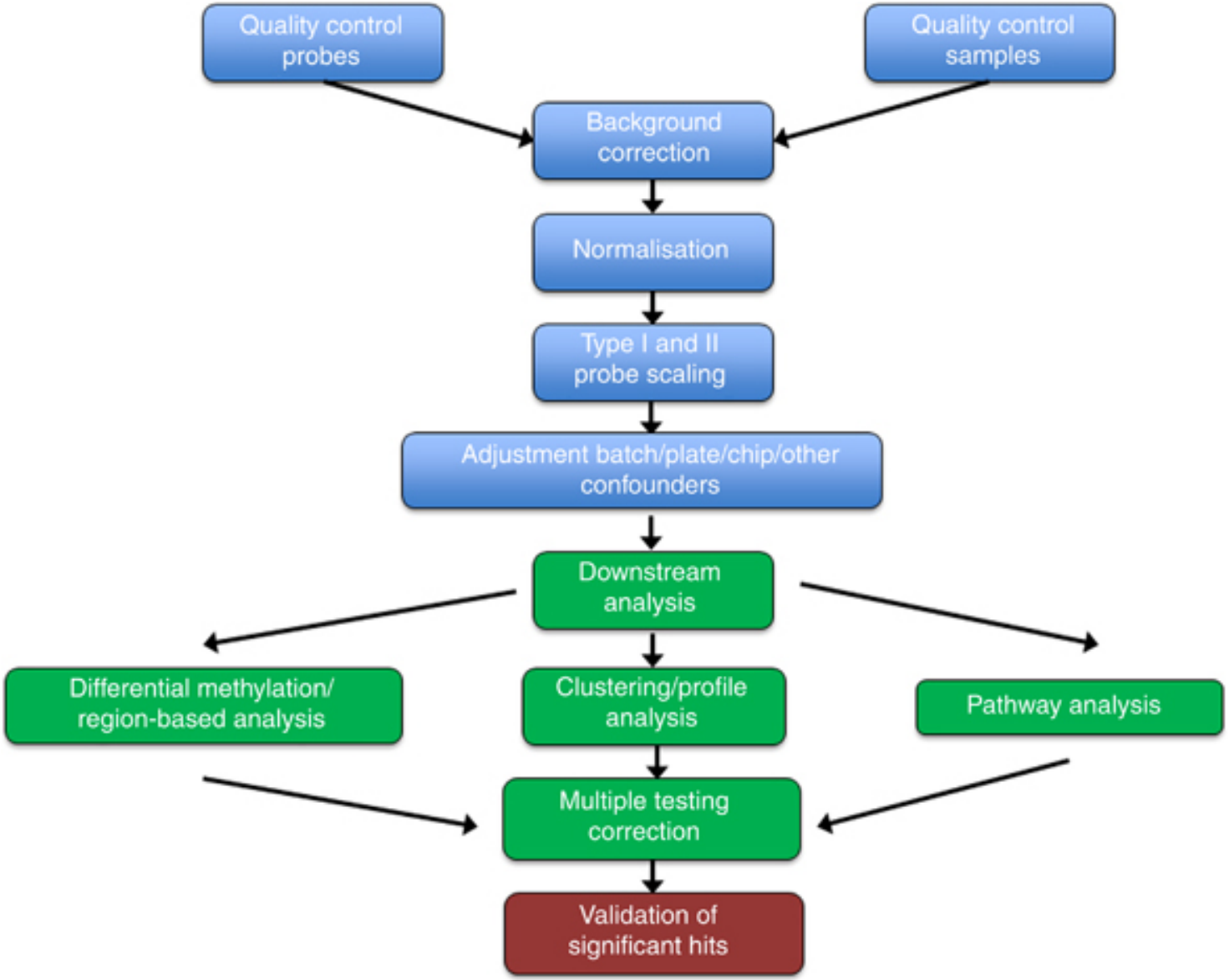
DNAm variation
Single-base vs regional definitions

- DNAm variation at a single CpG site ~ epigenetic equivalent of a SNP
- If DNAm is altered at multiple adjacent CpG sites, this is referred to as a differentially methylated region (DMR)



Int J Epidemiol. 2012 Feb; 41(1): 200–209.

Methylation array data processing & analysis pipeline



Probe & sample QC

- Detection p-value for each methylation beta-value
  - Probability that the target sequence signal is distinguishable from the background
  - Common practice: drop individual beta value if detection p-value >0.05
  - drop probes where median p-value >0.05
- Drop probes that are unsuccessfully measured in n^{th}% of samples
  - Common thresholds are 20%, 10%, 5%
- Drop samples that failed in n^{th}% of probes
  - Common thresholds are 50%, 20%

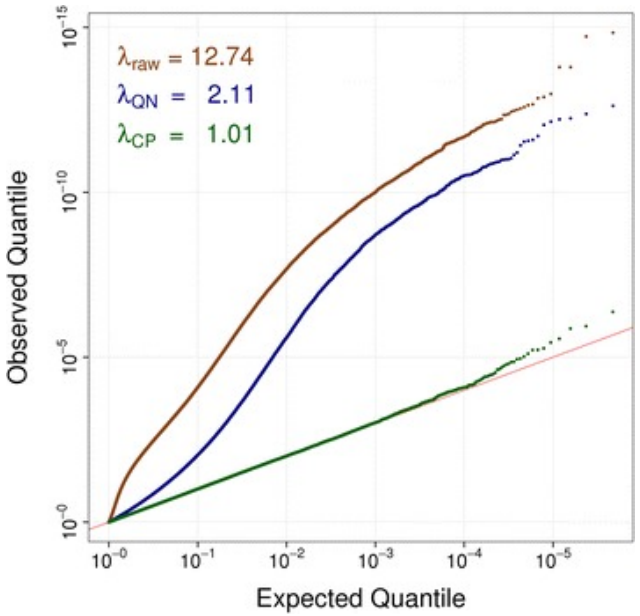
Filtering out probes

- Reduces the number of CpG sites taken forward for analysis
- Common practices related to technical issues:
  - Drop CpGs with known SNPs residing in the probe sequence
  - Drop CpG probes for which the CpG site contains a SNP
  - Drop CpGs in which probes anneal to multiple genomic locations
- Common practices related to analysis:
  - Drop CpGs on X and Y chromosomes
  - Drop CpGs with lowest variation
  - Drop CpGs with extreme methylation levels
  - Only consider those in regions of interest (e.g. CpG island, shore, other)

Normalisation & batch correction
Removing non-biological variations

- Variation within measurements caused by technical factors & batch effects –systematic differences across groups of samples
- Causes:
  - Differences in sample handling
  - DNA processing
  - Scanning of arrays (e.g. background noise)
  - Location of sample on chips
  - Technical biases

Correcting for statistical inflation due to technical biases

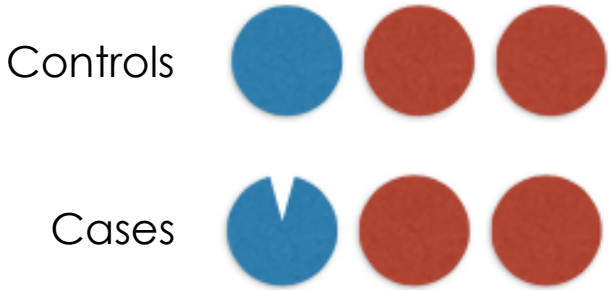


λ raw: uncorrected
λ QN: normalisation
λ CP: batch-correction

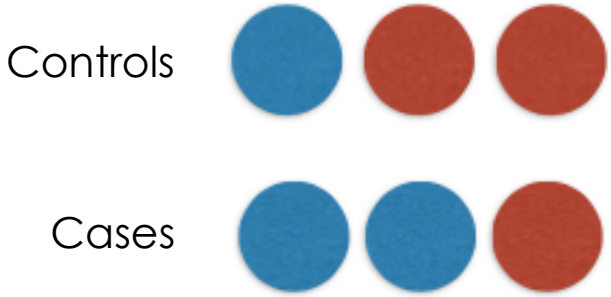
Genome Biol. 2015

Consequences of cellular heterogeneity

- **Differential expression within a cell type**
  - If the disease-associated DNAm is restricted to a certain cell type that represents only a small proportion of the tissue sampled, then the variation may not be detected.



- **Differential cell type composition between groups**
  - The disease state itself can also alter the composition of cell types in a tissue, and hence measured DNAm differences may only reflect differences in cell type composition and not true epigenetic differences.



Correcting for cellular heterogeneity

- Use direct cell counts for the major cell types in the sample
- Use reference information on cell-specific methylation signatures to estimate cell proportions from genome-scale methylation data
- Several 'reference-free' approaches identify clusters of covariation in the data, removing this covariation by adjustment

Houseman *et al. BMC Bioinformatics* 2012, **13**:86
http://www.biomedcentral.com/1471-2105/13/86

BMC Bioinformatics

RESEARCH ARTICLE Open Access

DNA methylation arrays as surrogate measures of cell mixture distribution

Eugene Andres Houseman[1*], William P Accomando[2], Devin C Koestler[3], Brock C Christensen[3], Carmen J Marsit[3], Heather H Nelson[4], John K Wiencke[5] and Karl T Kelsey[2,6]

R/Bioconductor packages for DNA methylation

Processing/analysis step	Packages
QC (samples)	IMA, HumMethQCReport, methylkit, MethyLumi, preprocessing and analysis pipeline, minfi
QC (probes)	IMA, HumMethQCReport, lumi, LumiWCluster, preprocessing and analysis pipeline, wateRmelon
Background correction	Limma, lumi, MethyLumi, minfi, preprocessing and analysis pipeline
Normalisation	Combat, HumMethQCReport, lumi, minfi, TurboNorm, MethyLumi, wateRmelon
Type 1 and 2 probe scaling	IMA, minfi, wateRmelon
Batch/plate/chip/confounder adjustment	Combat, CpGassoc, ISVA, MethLAB
Data dimension reduction	MethyLumi
Differential methylation analysis/ region-based analysis	CpGAssoc, IMA, limma, methylkit, MethLAB, MethVisual, minfi, EVORA
Clustering/profile analysis	Lumi, ISVA, HumMeth27QCReport, methylkit, RPMM, SS-RPMMb
Multiple testing correction	CpGAssoc, methylkit, MethLAB, NHMMfdr

Needs for analytical improvements!

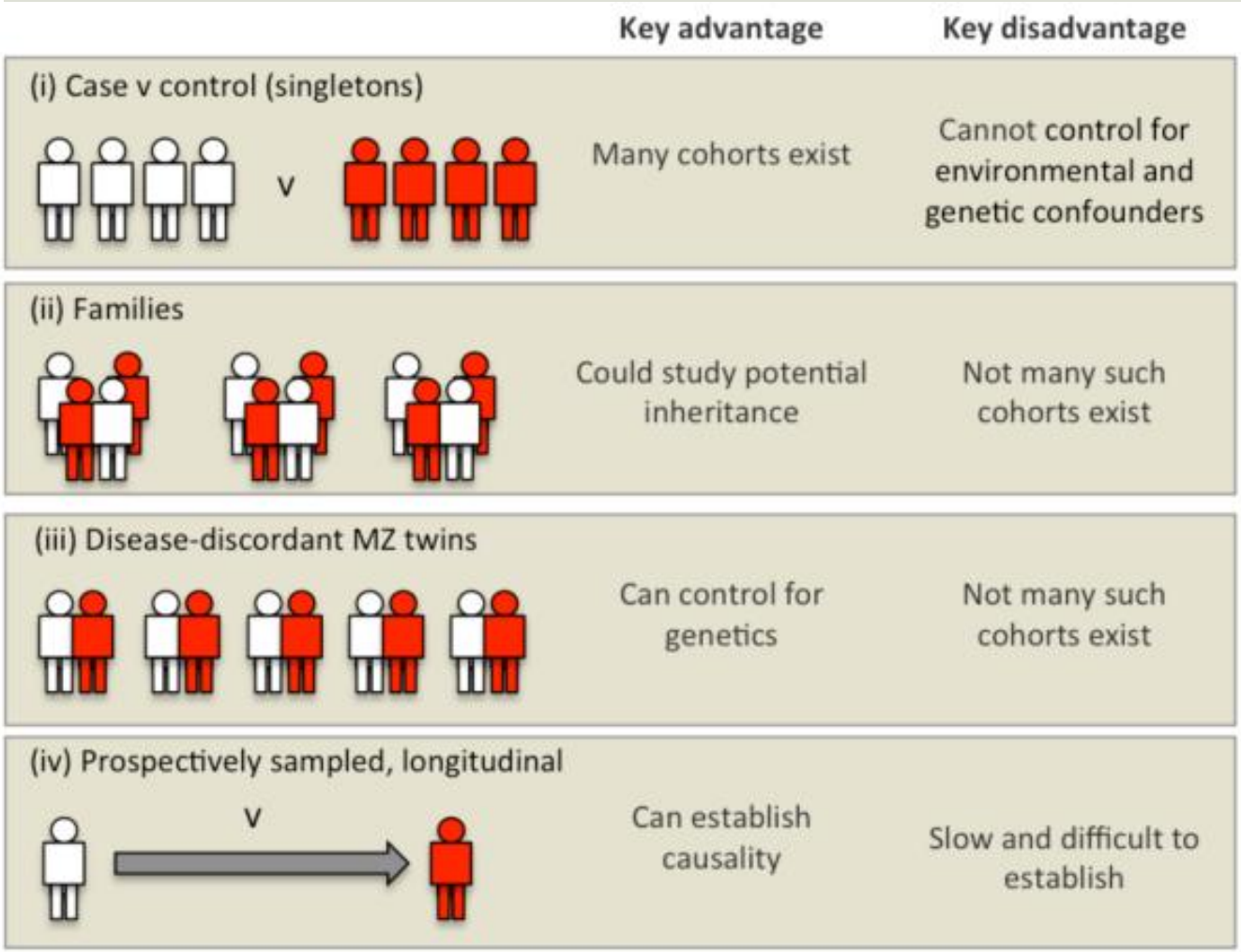
- EWAS = Novel, evolving field of study
- Many assumptions for the methods used are violated
  - Variance of DNAm is a function of the mean (heteroscedasticity)
  - CpG site density and correlation is not constant across the genome
  - DNAm is associated with CpG density
  - Fluorescence signals, and methylation levels influenced by GC content
  - Different probe types, measuring different CpGs, present on one chip

    => *Improved solutions needed for the statistical analysis of DNAm*

Sample size and power?

Difficult to calculate:

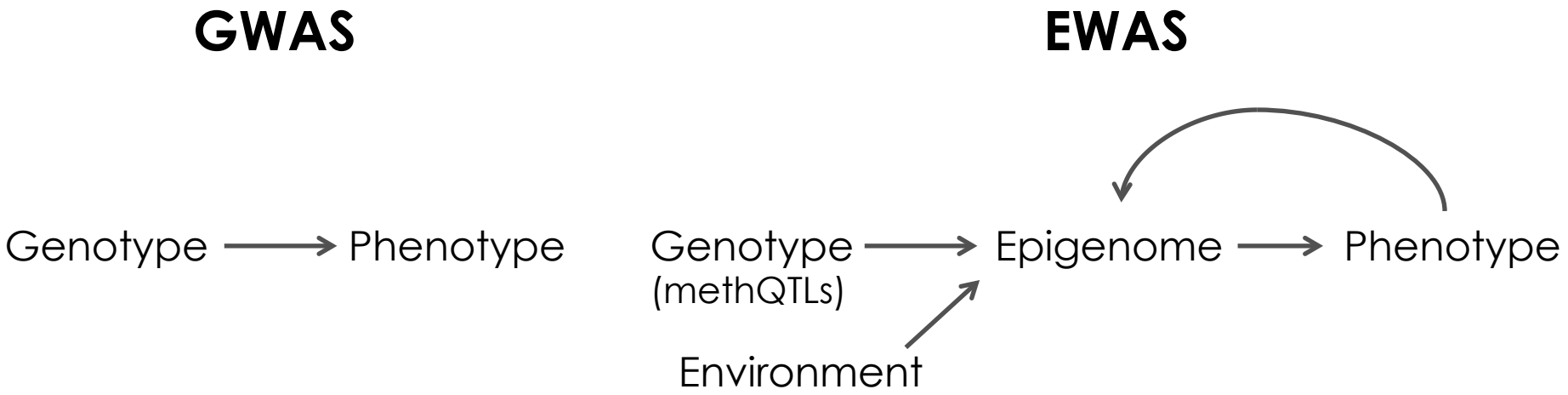
- Little information available about frequency spectra of DNAm variants and their effect sizes for common diseases
- Great variation of DNAm across genomic contexts & cell types
- Likely that effect sizes and hence power will vary substantially according to genomic context



Nat Rev Genet. 2011 12: 529–541

DNAm as Cause or Consequence?

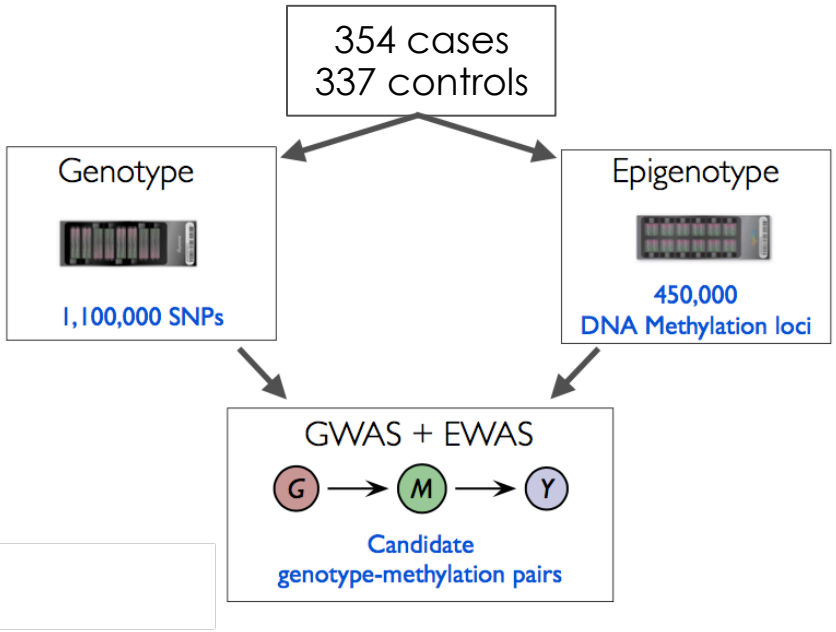
Study Design can help



DNAm: Cause or Consequence?

Integrating GWAS data can help

Mediation analysis to filter out associations likely consequential to disease

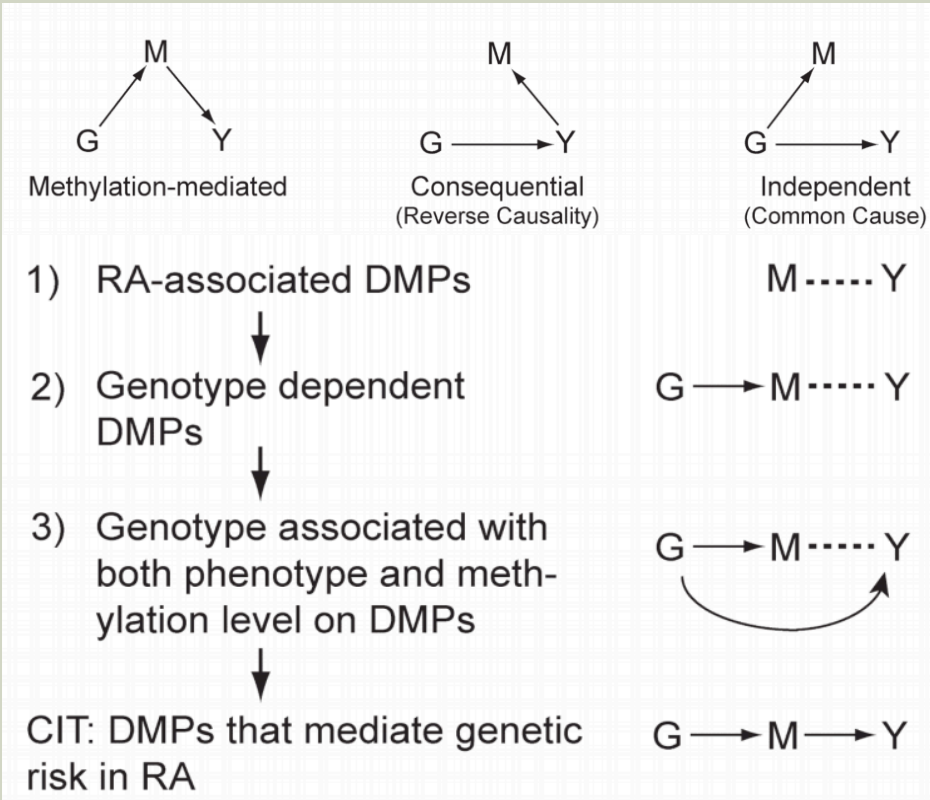


Nat Biotechnol. 2013: 142–147.

Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in Rheumatoid Arthritis

Yun Liu[1,2,*], Martin J. Aryee[1,3,*], Leonid Padyukov[4,5,*], M. Daniele Fallin[1,8,9,*], Espen Hesselberg[4,5], Arni Runarsson[1,2], Lovisa Reinius[6], Nathalie Acevedo[7], Margaret Taub[1,8], Marcus Ronninger[4,5], Klementy Shchetynsky[4,5], Annika Scheynius[7], Juha Kere[6], Lars Alfredsson[10], Lars Klareskog[4,5,†], Tomas J. Ekström[5,11,†], and Andrew P. Feinberg[1,2,8,†]

Causal Inference Test



Useful links

- R: http://www.r-project.org/
- Bioconductor: http://www.bioconductor.org/
  (Minfi, lumi, methylumi)
- Dedeurwaerder et al. Evaluation of the Infinium Methylation 450K technology. Epigenomics. 2011;3(6):771-84.

Acknowledgments

IMAGEN Consortium

Gunter Schumann
Tianye Jia
Barbara Ruggeri
Bing Xu
Anna Cattrell
Alex Ing
Erin Quinlan

imagen

ENIGMA Consortium

Paul Thompson
Derrek Hibar
Jason Stein
Alejandro Arias Vasquez
Neda Jahanshad
Nick Martin
Margie Wright
Barbara Franke
Sarah Medland

ENIGMA

ENIGMA Epigenetics Working Group

Yun Liu
Tianye Jia
Paul Thompson
+ PIs from participating cohorts