MRC Integrative Epidemiology Unit

University of BRISTOL

Federated analyses for large epigenetic studies

Josine Min

Senior Research Fellow in Genetic and Epigenetic Epidemiology

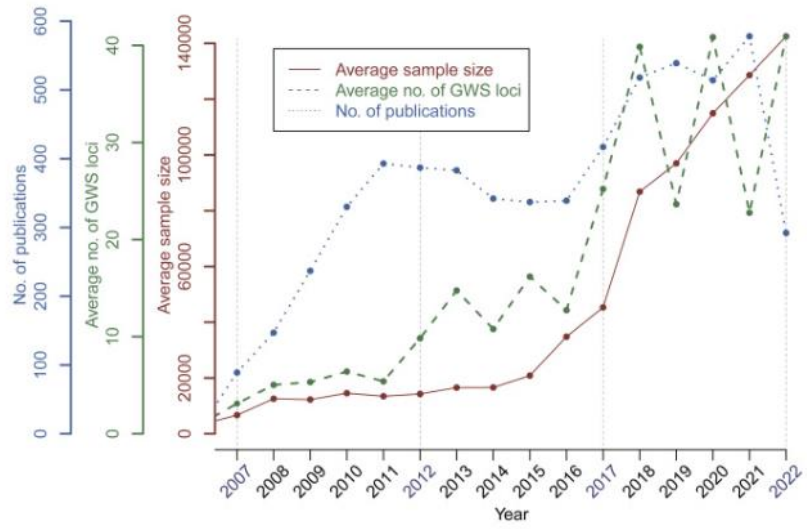
MRC IEU

Population Health Sciences

Association of common genetic variation with diseases and traits

- Genome wide association studies led to the discovery of >70,000 genetic associations to common diseases and traits

- 90% of GWAS variants are located in non-coding regions

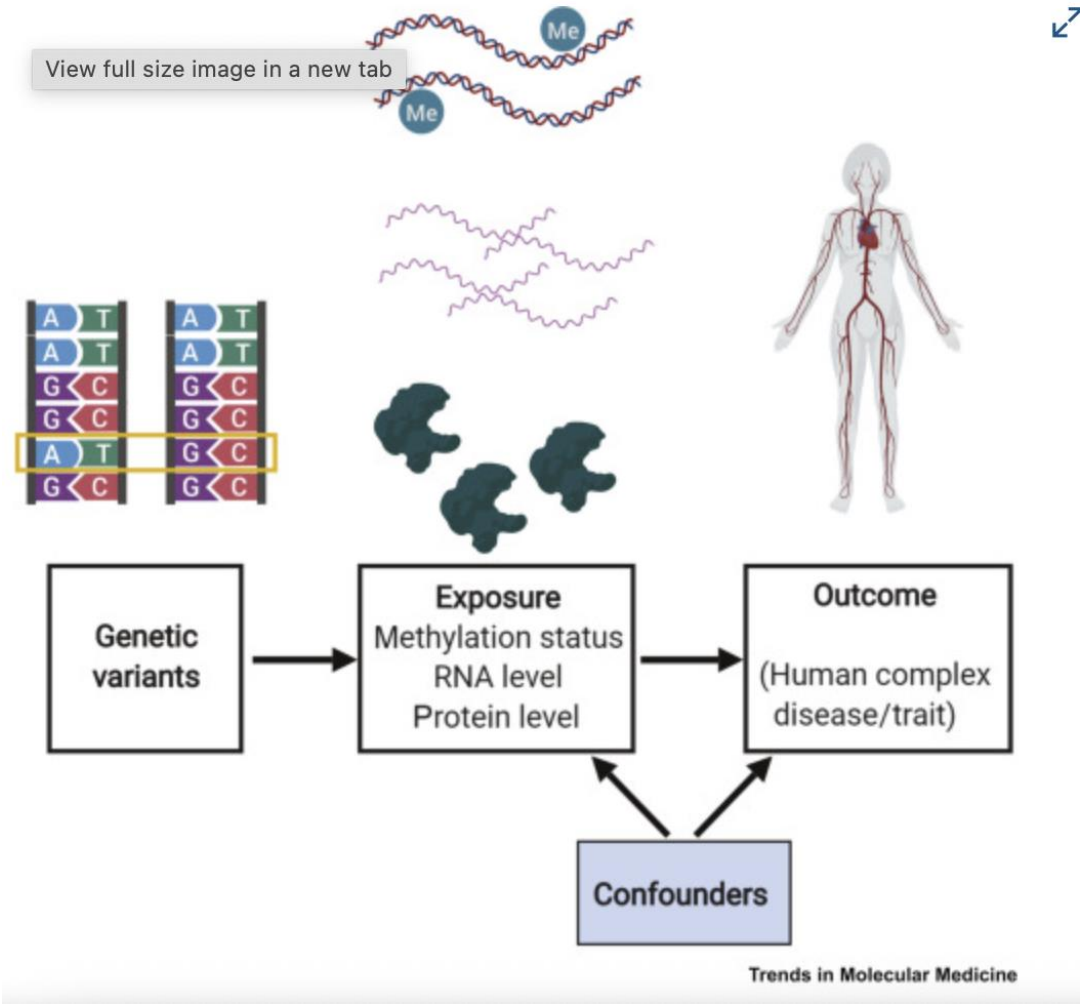


Abdellaoui et al 2023

MRC Integrative Epidemiology Unit

University of BRISTOL

Integration with other genomic layers

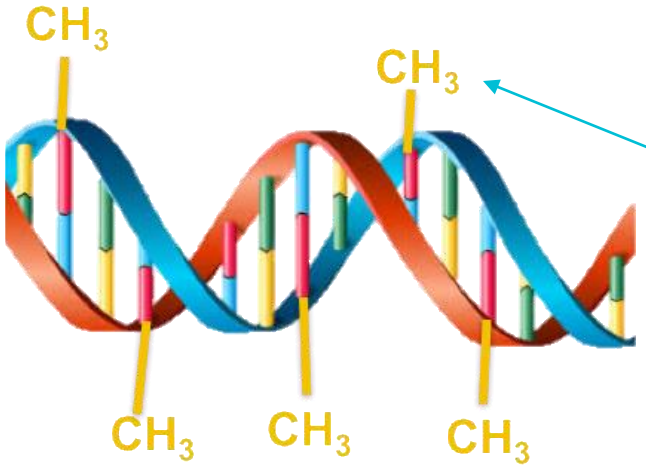


Zeggini et al. 2020

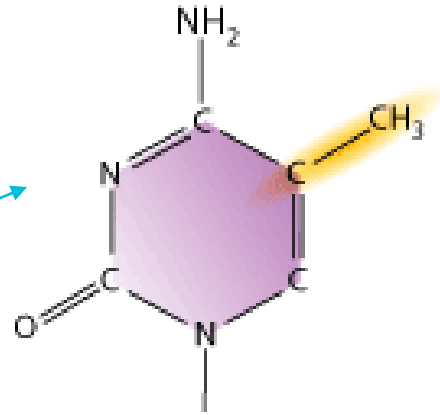
What does “epigenetics” mean?

Epigenetics is the study of mitotically heritable changes in gene expression that occur without changes in DNA sequence.

DNA methylation

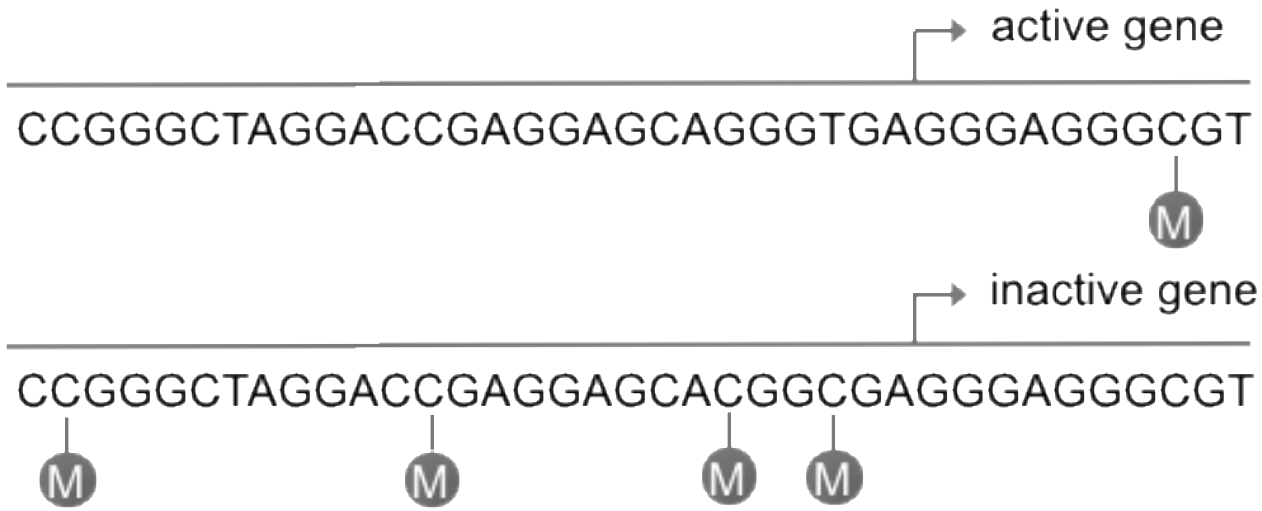


Methyl group

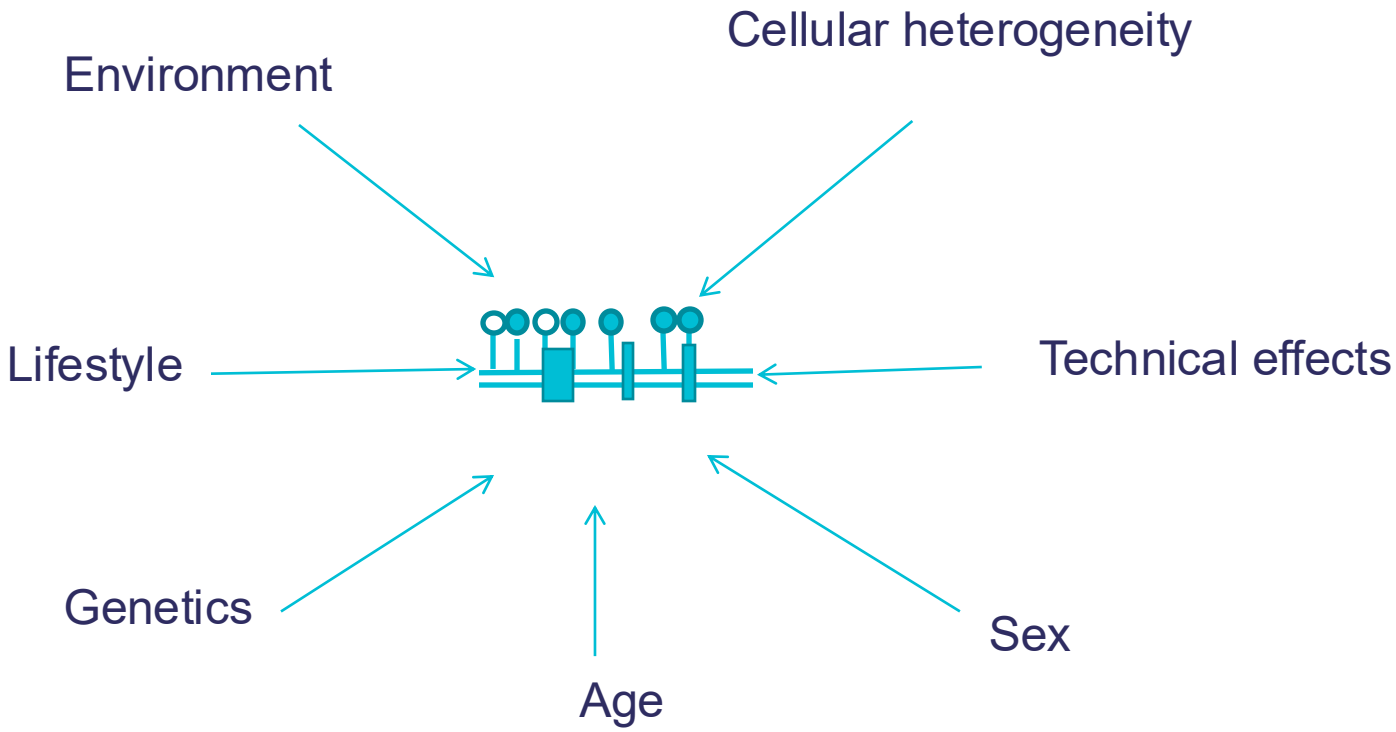


CpG Sites

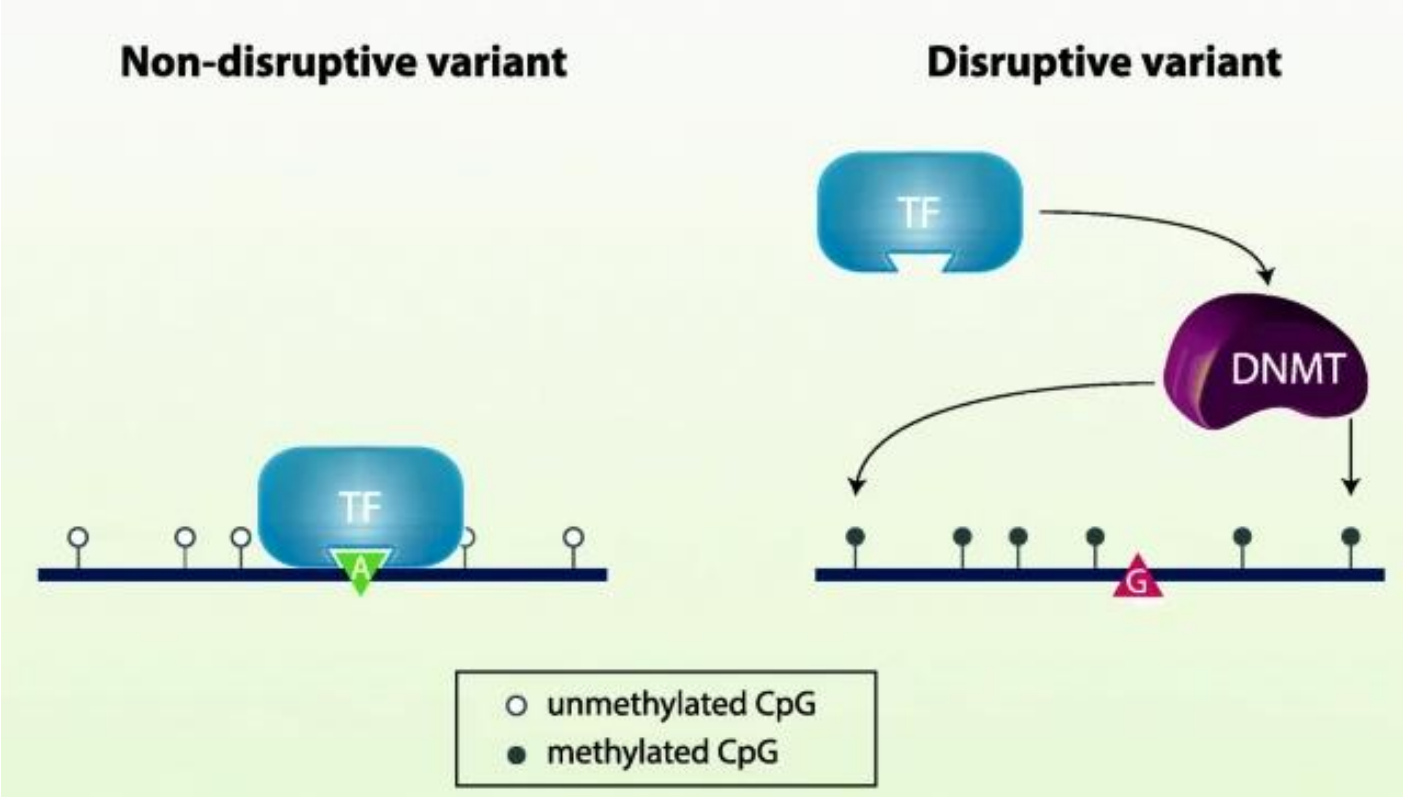
TTTATCGGTAGATC



What influences DNA methylation?



mQTL as regulatory variant

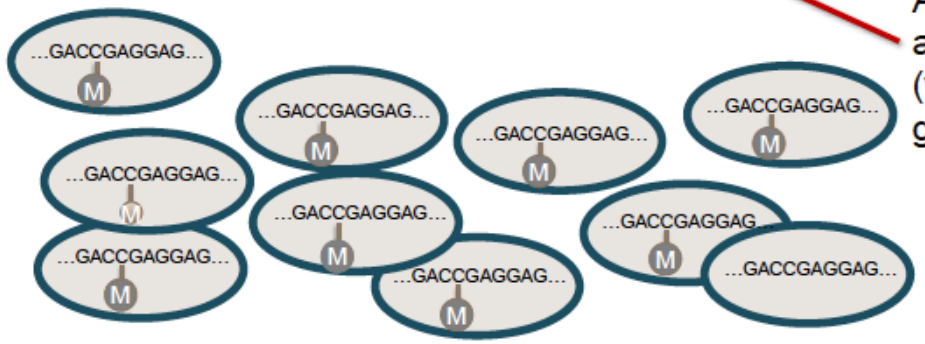


mQTL: a genetic variant that is associated with DNA methylation at a CpG site.

Villicana et al. 2021

DNA methylation dataset

	AAT0	A0SE	A0SF	A0SJ	A04R
cg00000029	0.24	0.16	0.16	0.11	0.14
cg00000165	0.12	0.14	0.45	0.45	0.25
cg00000236	0.86	0.93	0.94	0.94	0.93
cg00000289	0.53	0.71	0.80	0.74	0.60
cg00000292	0.41	0.65	0.47	0.90	0.90



90% of cells from sample A04R have methylation at cytosine cg00000292 (which is in the *ATP2A1* gene)

- Arrays measure 3-5% of the genome
- On an array we measure methylation in a population of cells
- One CpG can be 0, 0.5 or 1
- Across a population we get a methylation level between 0 and 1

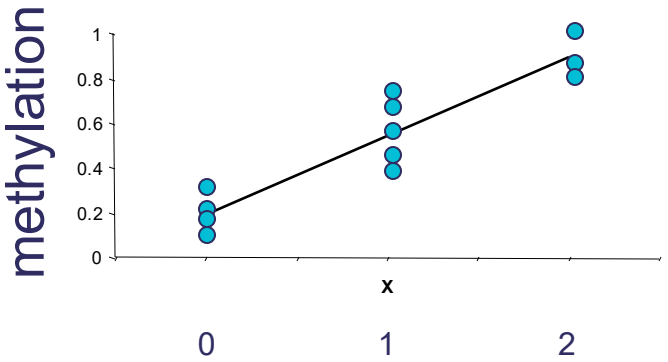
mQTL analysis - analogous to GWAS for continuous traits

Linear regression equation $Y_i = b_0 + b_1 X_i + e_i$

where

Y_i = continuous trait value for individual i

X_i = number of 'A' alleles an individual has

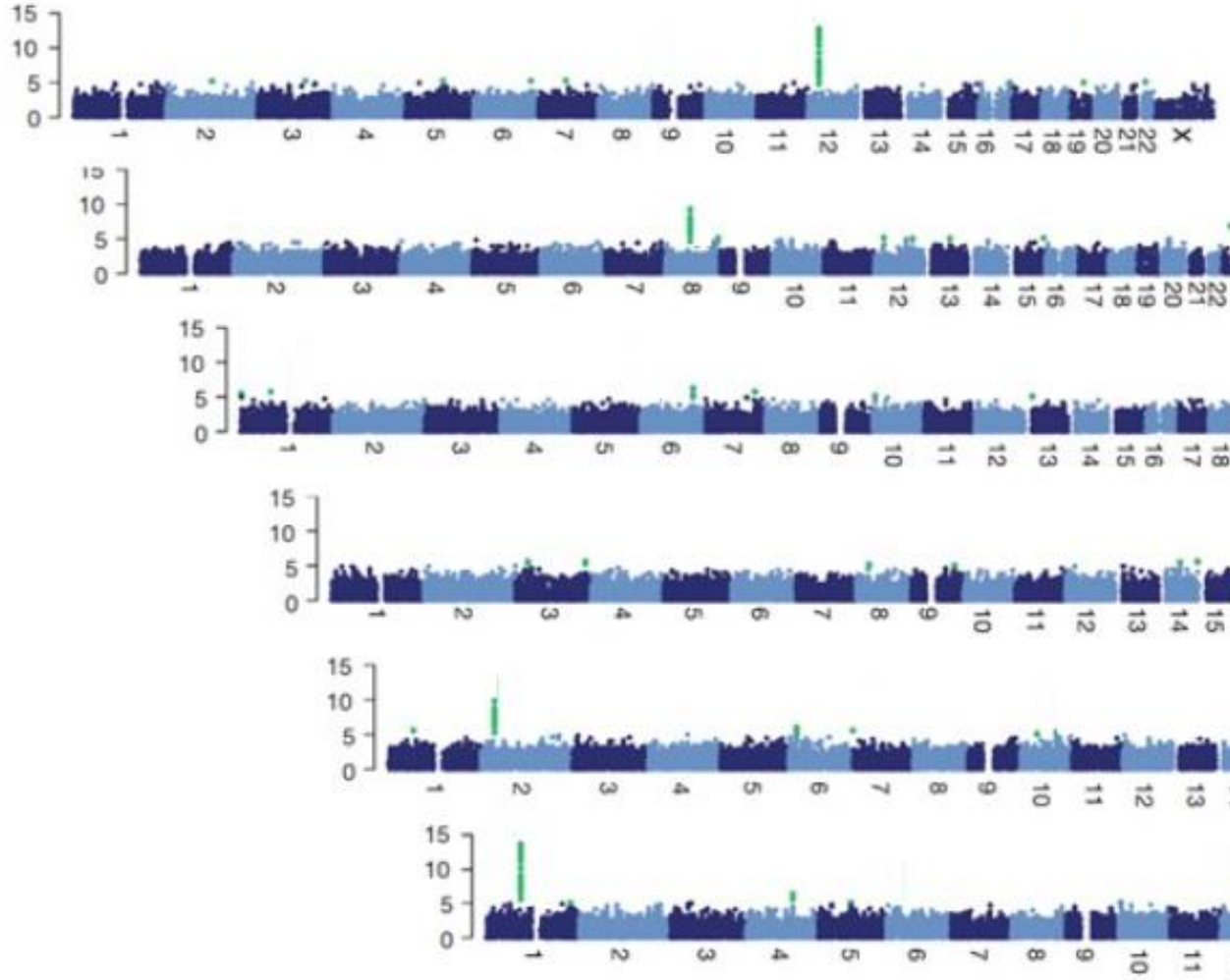


Genotype (e.g. number of "A" alleles")

	sample 1	sample 2	sample 3	sampl
rs1050745	1	2	1	
rs10507452	0	0	0	
rs10507456	2	1	1	
rs10507457	2	1	2	
rs10507458	0	1	0	
rs10507462	2	2	2	
rs10507463	2	2	2	
rs10507465	0	1	1	
rs10507466	1	1	2	
rs10507467	1	2	1	
rs10507481	1	2	1	
rs10507482	1	0	2	
rs10507483	0	0	0	
rs10507489	1	2	2	
rs10507490	2	2	2	
rs10507491	0	0	0	
rs10507495	0	0	0	
rs10507496	2	1	2	
rs10507498	0	0	0	

Association test is whether b ≠ 0

Whole genome mQTL analysis is a GWAS for each CpG



CpG 1

CpG 2

CpG 3

CpG 4

CpG 5

CpG N

Total tests:

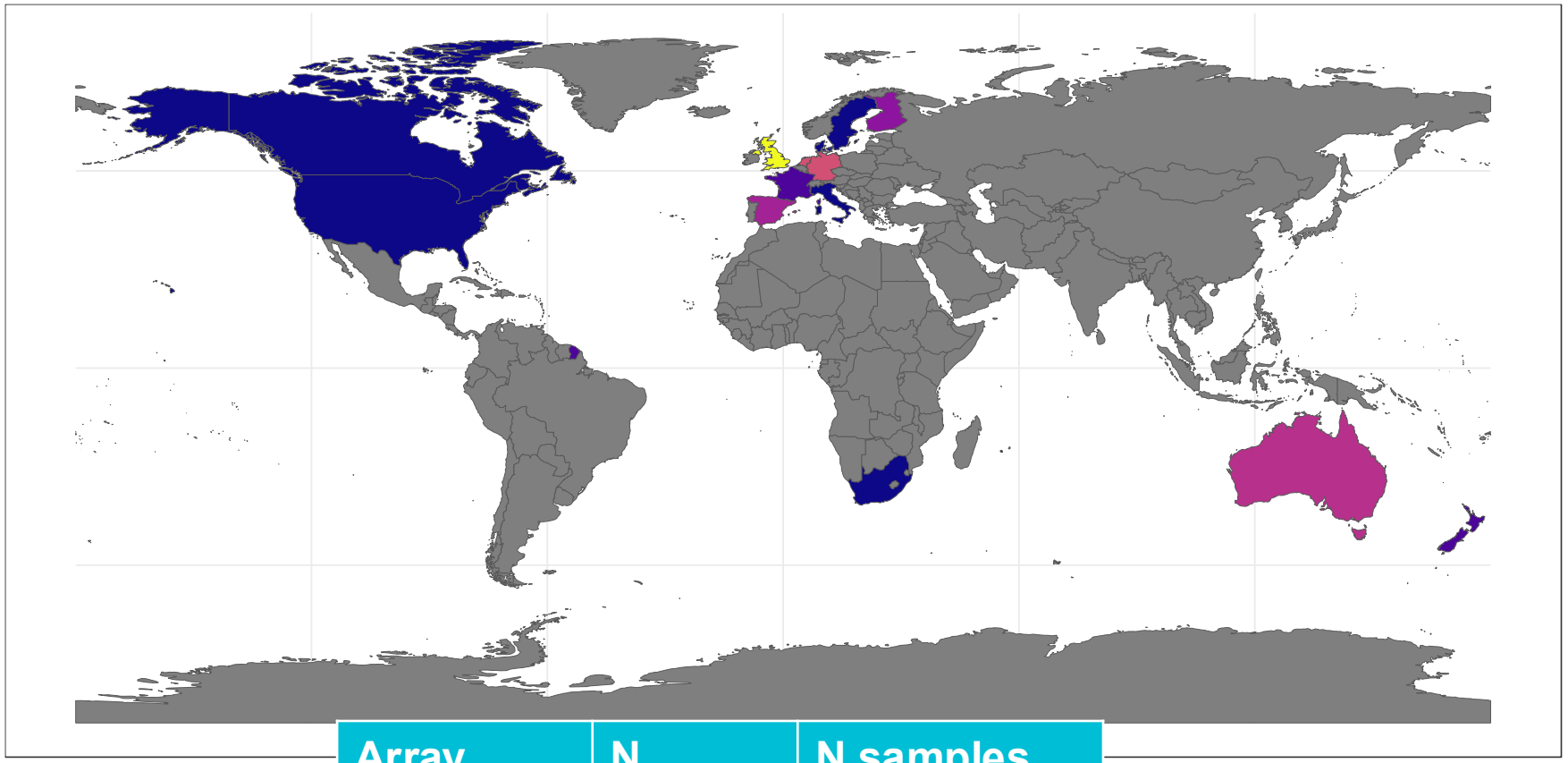
800,000 CpGs x 10M SNPs
800,000 GWAS!

Why federated analyses?

- Larger sample sizes are needed
  - Combining multiple studies
- Relevance of distant associations
  - To increase power to identify distant mQTLs have smaller effect sizes
- Reproducible and efficient
  - Sharing code improves research quality -> no analysis plans
  - Consistency across large-scale analyses of genetic and DNAm datasets
  - Automated control of confounding and batch effects
- To enable collaboration
  - No sharing of individual data

Genetics of DNA methylation Consortium

>60 cohorts with genetic and epigenetic data on blood samples



Inclusion criteria

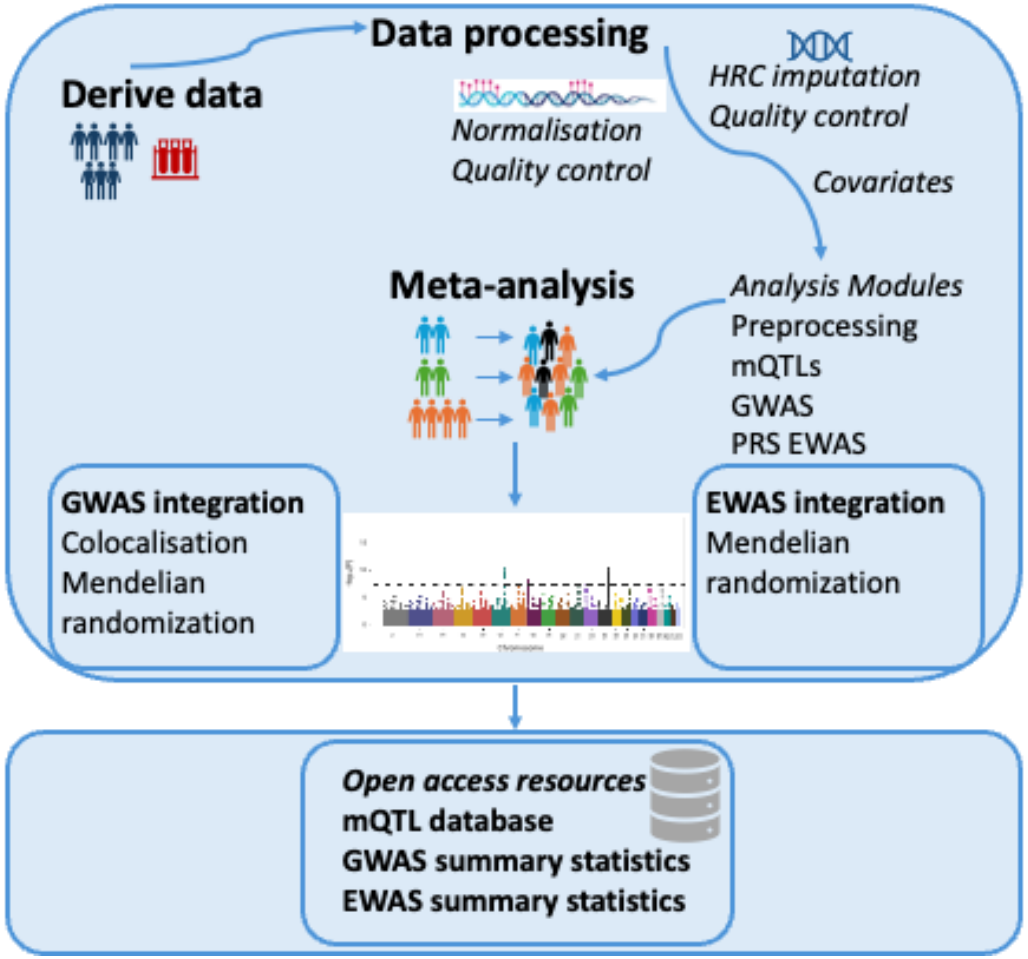
- Genetic data
- Methylation array data
- Blood samples/sorted celltypes
- N>200
- European

Array	N cohorts	N samples
EPIC	37	65186
450k	23	21022
Total	60	86208

- Multi-ancestry
  - Diverse epigenetic epidemiology partnership (DEEP)
  - Fine mapping

Github repository

https://github.com/genetics-of-dna-methylation-consortium/godmc_phase2/wiki



- Input data folder
- Config file
- Processed data
- Results
- Results upload

- 55 cohorts ran first module

GoDMC2 wiki

Home

Edit | New page

ejh243 edited this page on Jan 17 · 12 revisions

Welcome to the GoDMC Phase 2 pipeline. This wiki will guide you through the analysis. The pipeline has been designed so that minimal coding is required by analysts, which hopefully ensures that it is quick to execute, avoids duplication of efforts, increases reproducibility and reduces potential for error. The main objectives of the pipeline are:

Perform a mQTL analysis

Perform a full GWAS on every methylation probe available in the sample. Variations of this analysis will be performed, including:

- a standard mQTL analysis using bestguess imputed SNPs covering the full surface using new software
- cell-type interacting mQTL analyses for abundant cell types
- an inversion-mQTL analysis using inversions inferred from SNP data
- a var-mQTL analysis which looks for SNPs that influence the variance of a probe.
- a sex-stratified mQTL analysis for Chromosome X and Y

IMPORTANT: Please note that in GoDMC Phase 2 we are focusing on blood samples of Europeans only that are profiled with EPIC v1 or EPIC v2 arrays. Datasets should have at least 200 individuals and ideally be imputed with the HRC panel.

GWAS of DNAm derived phenotypes

We will be performing a GWAS on:

- biological age proxies derived from epigenetic clocks
- cumulative smoking exposure derived from DNA methylation data
- blood cell type proportions derived by DNA methylation
- MZ twinning

EWAS of polygenic risk scores

We will be conducting PRS EWASs on:

Pages 34

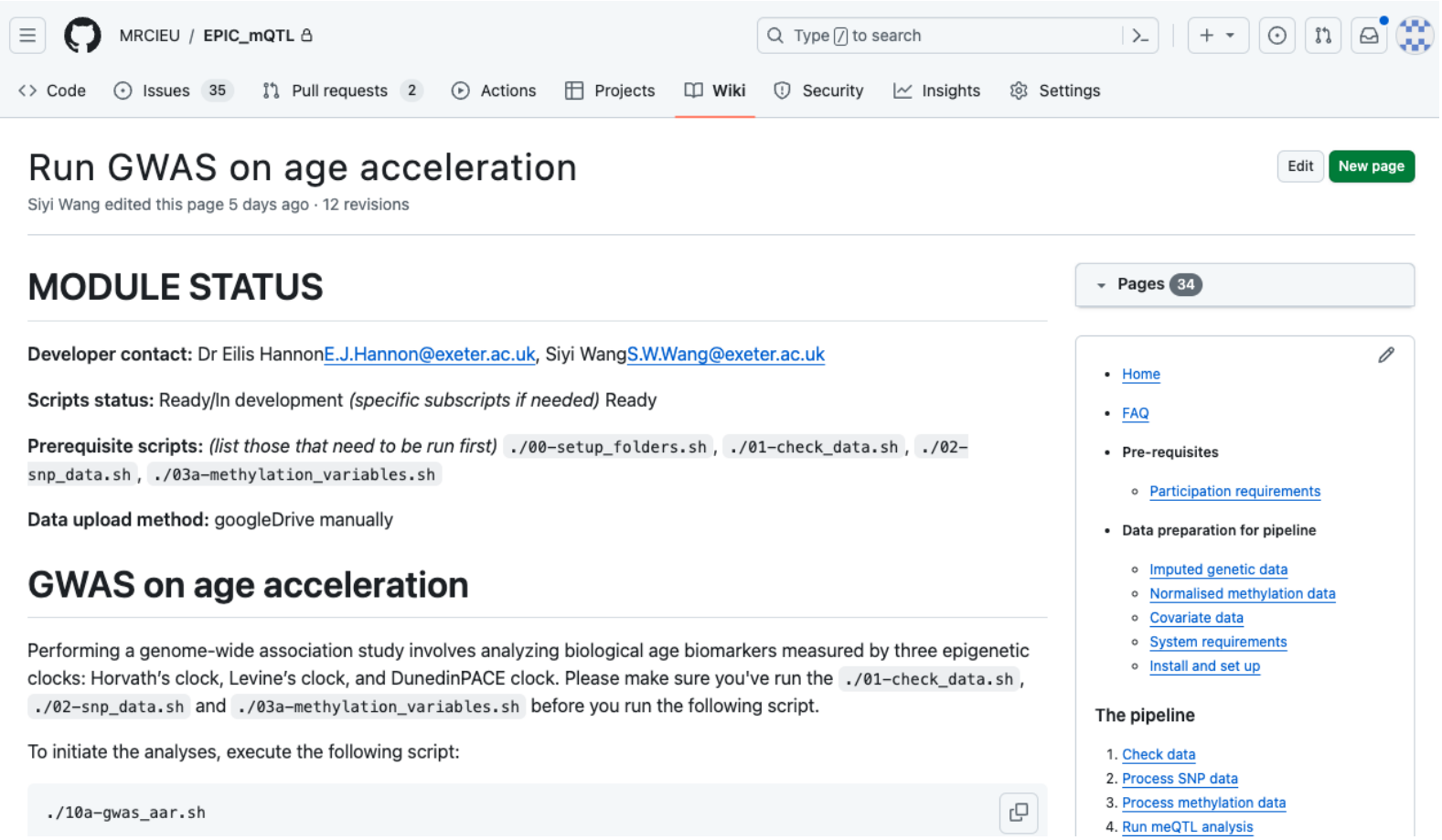
- Home
- FAQ
- **Pre-requisites**
  - Participation requirements
- **Data preparation for pipeline**
  - Imputed genetic data
  - Normalised methylation data
  - Covariate data
  - System requirements
  - Install and set up

The pipeline

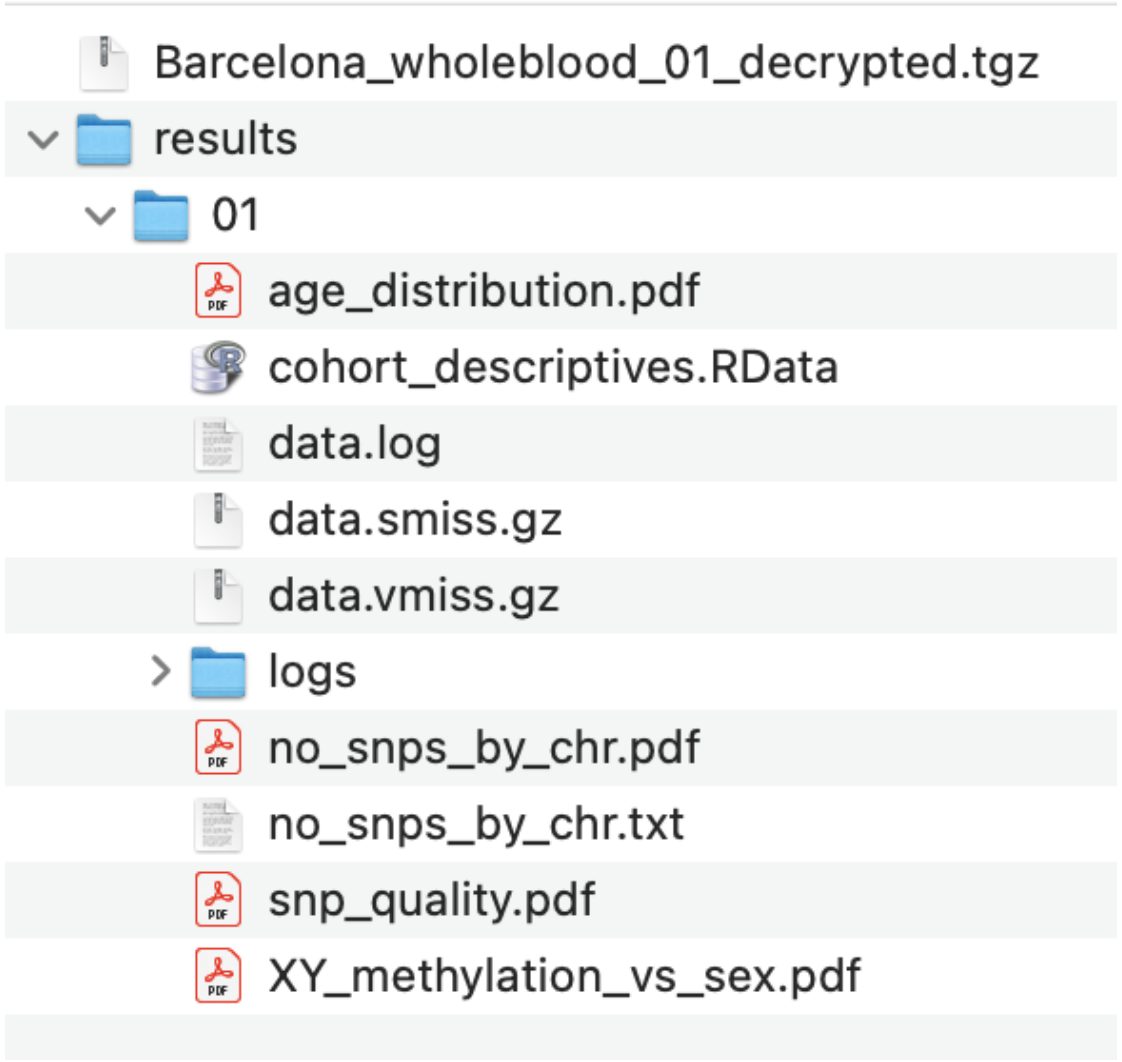
1. Check data
2. Process SNP data
3. Process methylation data
4. Run meQTL analysis
5. meQTL analysis of chromosome X and Y CpGs
6. Run cell type interacting meQTL analysis
7. Run variance meQTL analysis
8. Run inversion meQTL analysis
9. Run PRS EWAS
10. Run GWAS and heritability on age acceleration
11. Run GWAS and heritability on smoking
12. Run GWAS on cell counts
13. Run GWAS on MZ twinning
14. Run GWAS on the imprinted gene nc886

https://github.com/genetics-of-dna-methylation-consortium/godmc_phase2

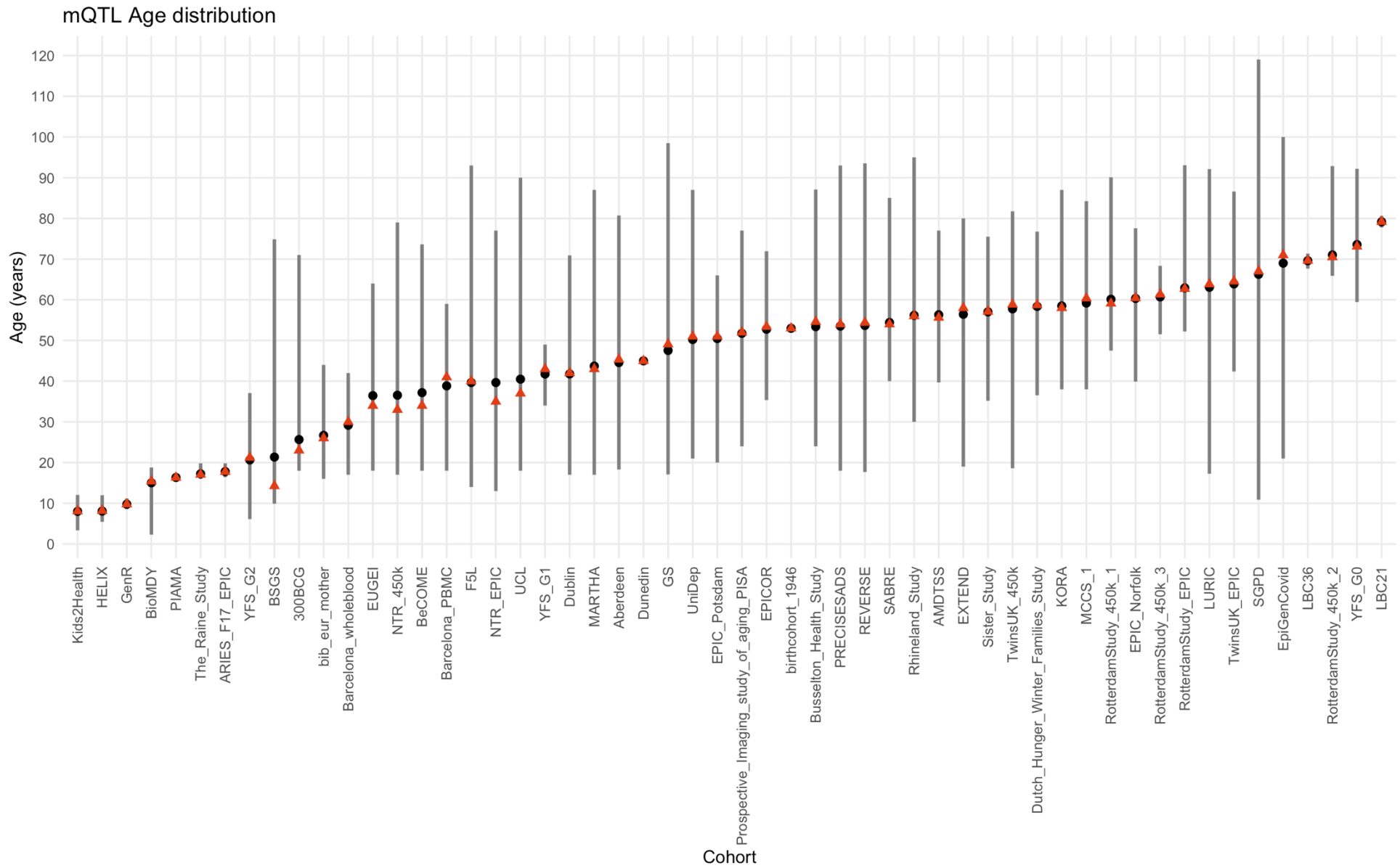
Example of module



Example Module 1



Age check

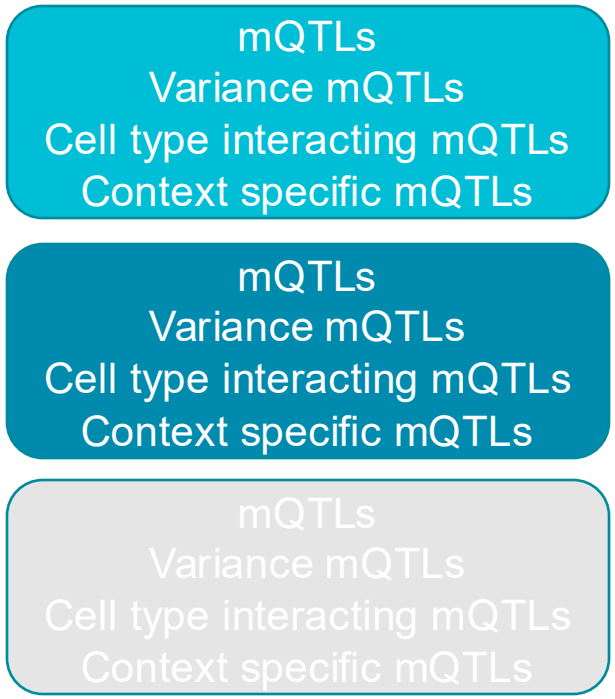

mQTL Age distribution
Age (years)
0
10
20
30
40
50
60
70
80
90
100
110
120
Kids2Health
HELIX
GenR
BioMDY
PIAMA
The_Raine_Study
ARIES_F17_EPIC
YFS_G2
BSGS
300BCG
bib_eur_mother
Barcelona_wholeblood
EUGEI
NTR_450k
BeCOME
Barcelona_PBMC
F5L
NTR_EPIC
UCL
YFS_G1
Dublin
MARTHA
Aberdeen
Dunedin
GS
UniDep
EPIC_Potsdam
Prospective_Imaging_study_of_aging_PISA
EPICOR
birthcohort_1946
Busselton_Health_Study
PRECISESADS
REVERSE
SABRE
Rhineland_Study
AMDTSS
EXTEND
Sister_Study
TwinsUK_450k
Dutch_Hunger_Winter_Families_Study
KORA
MCCS_1
RotterdamStudy_450k_1
EPIC_Norfolk
RotterdamStudy_450k_3
RotterdamStudy_EPIC
LURIC
TwinsUK_EPIC
SGPD
EpiGenCovid
LBC36
RotterdamStudy_450k_2
YFS_G0
LBC21
Cohort

GoDMC2 mQTL catalog

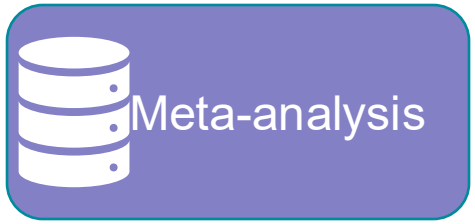
Software/Package	Language	Approx Speed	Approx Storage
MatrixEQTL	R	<1 day	>500T
TensorQTL	python3	~1-2 days	>500T
HASE	python2	~1-2 days	~100G
OSCA	C++	~1 day	30T

HASE can store complete surface -> partial derivatives

Building a mQTL resource for the community



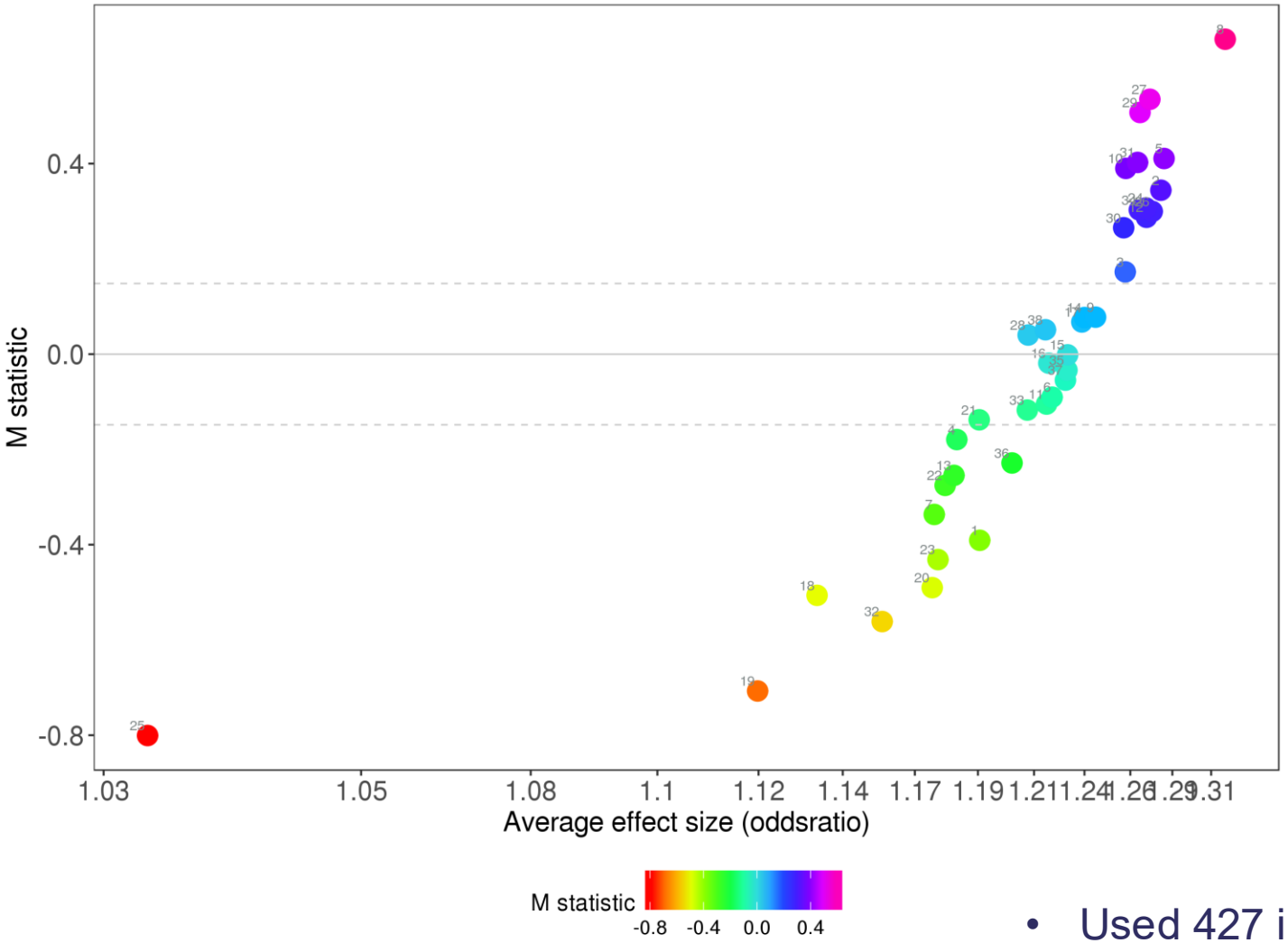
Federated analysis



Research:

- Open source mQTL catalog
- Comparison with variance mQTLs and interaction mQTLs
- Causal relationship with clinical traits
- Genetic correlations
- Natural selection
- Functional properties
- ...
- ...

QC- systematic heterogeneity



Meta-regression

Number of SNPs
Number of CpGs
Sample size
Methylation array
Average MAF
Average Info
Lambda
Average beta value across all probes
Average sd across all probes
Ancestry (UK vs Non UK, NL vs non NL etc.
Average age
Fraction of males
Relatedness yes/no
Cord vs whole blood
Height (N=24)
BMI (N=24)
Normalization method

- Used 427 independent trans SNPs with p<10-14

Diverse DNA methylation data to understand disease discordance

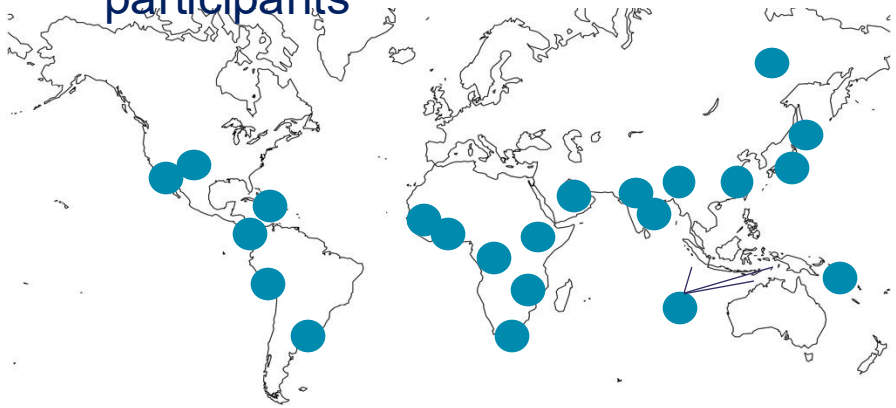
↑ Genetic variation

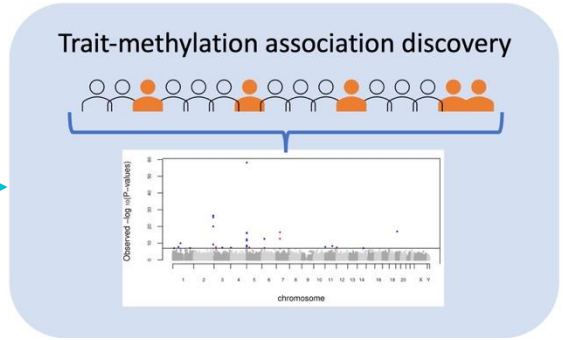
↑ Environmental variation

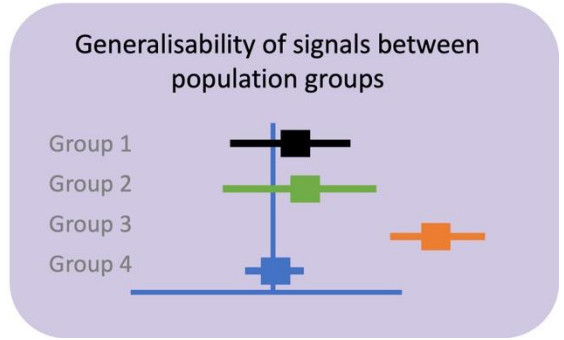
↑ DNA methylation variation

The Diverse Epigenetic Epidemiology Partnership (DEEP) study aims to improve population health by exploring the effects of genomic and environmental diversity on DNA methylation and disease risk across the global human population

N=13,278 diverse human participants







Medical Research Council

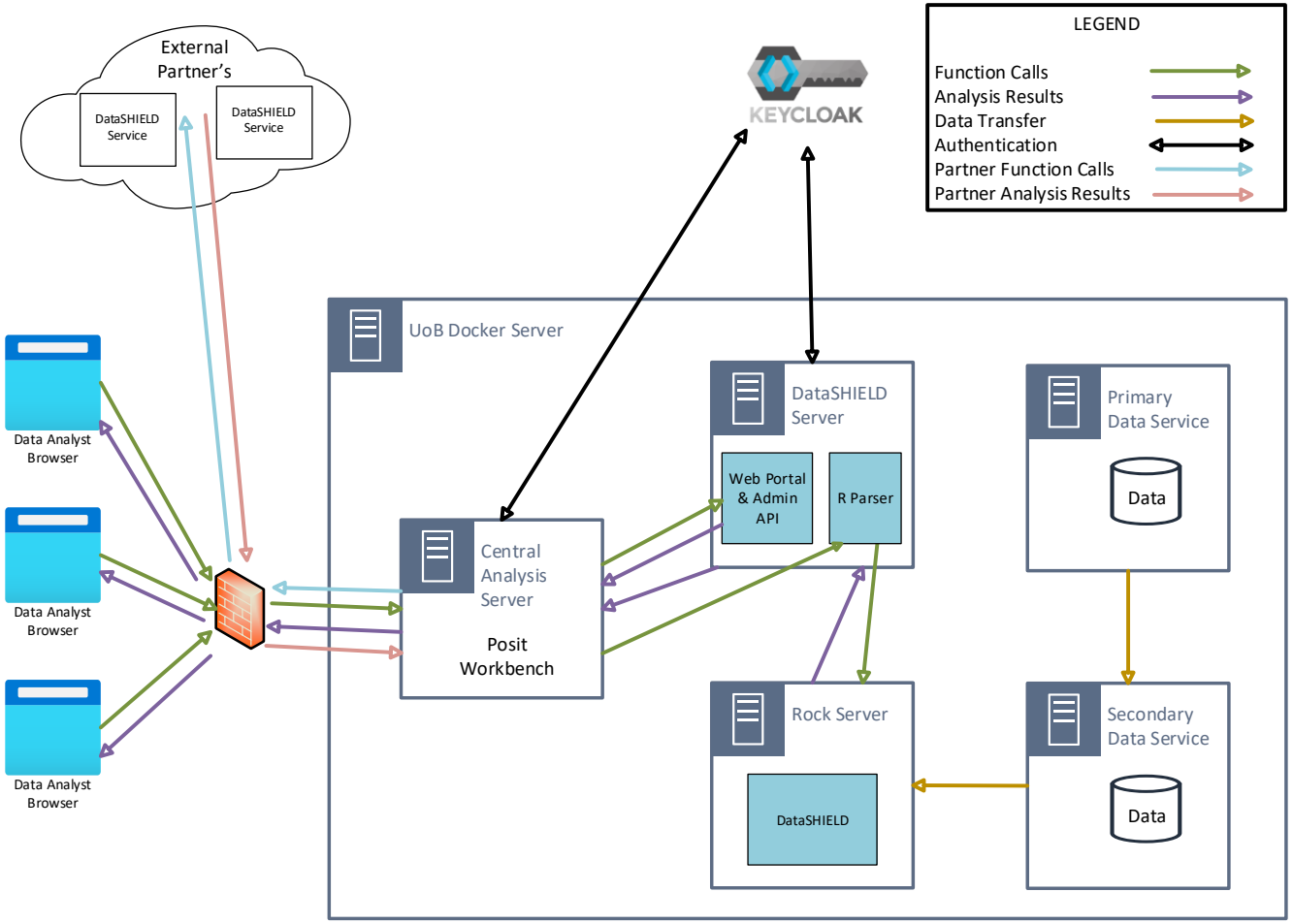
DIVERSE EPIGENETIC EPIDEMIOLOGY PARTNERSHIP

PIs Josine Min & Hannah Elliott (UoB)
Co-Is Prachand Issarapu, Andrew Prentice (LSHTM/MRCG), Giriraj Chandak (CCMB)
& 20 partners

DEEP Study

www.deep-epigenetics.org

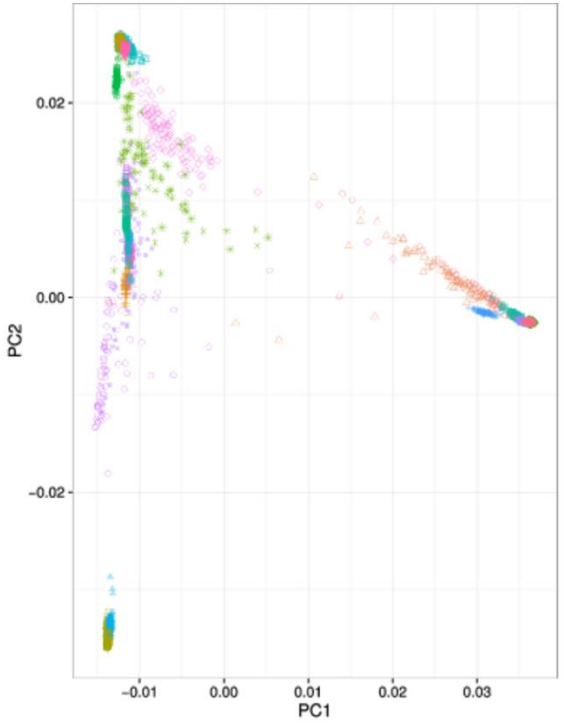
DATASHIELD
nondisclosive federated analysis



- Pooled analysis
- No individual level data sharing

Non genetic variation of DNA methylation

Nondisclosive Principal component analysis



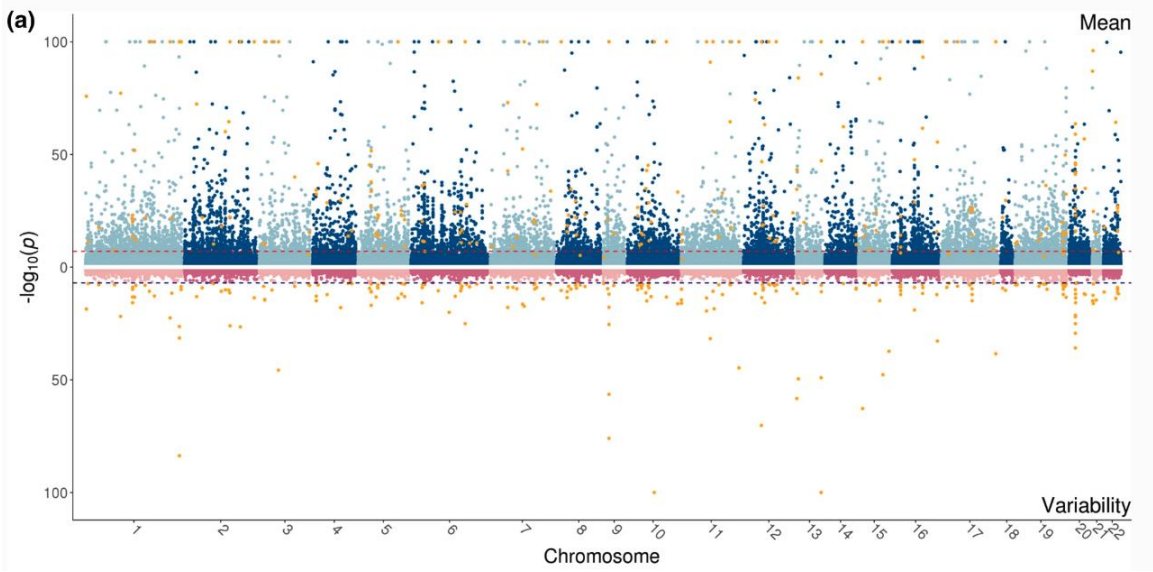
- Batch and cohort effects
- Heritable and nonheritable probes
- Remove genetic component
- Correlate PCs to
    - Cell counts
    - Inflammation measures (IL6)
    - Age at infection
    - Climate variation
    - ...

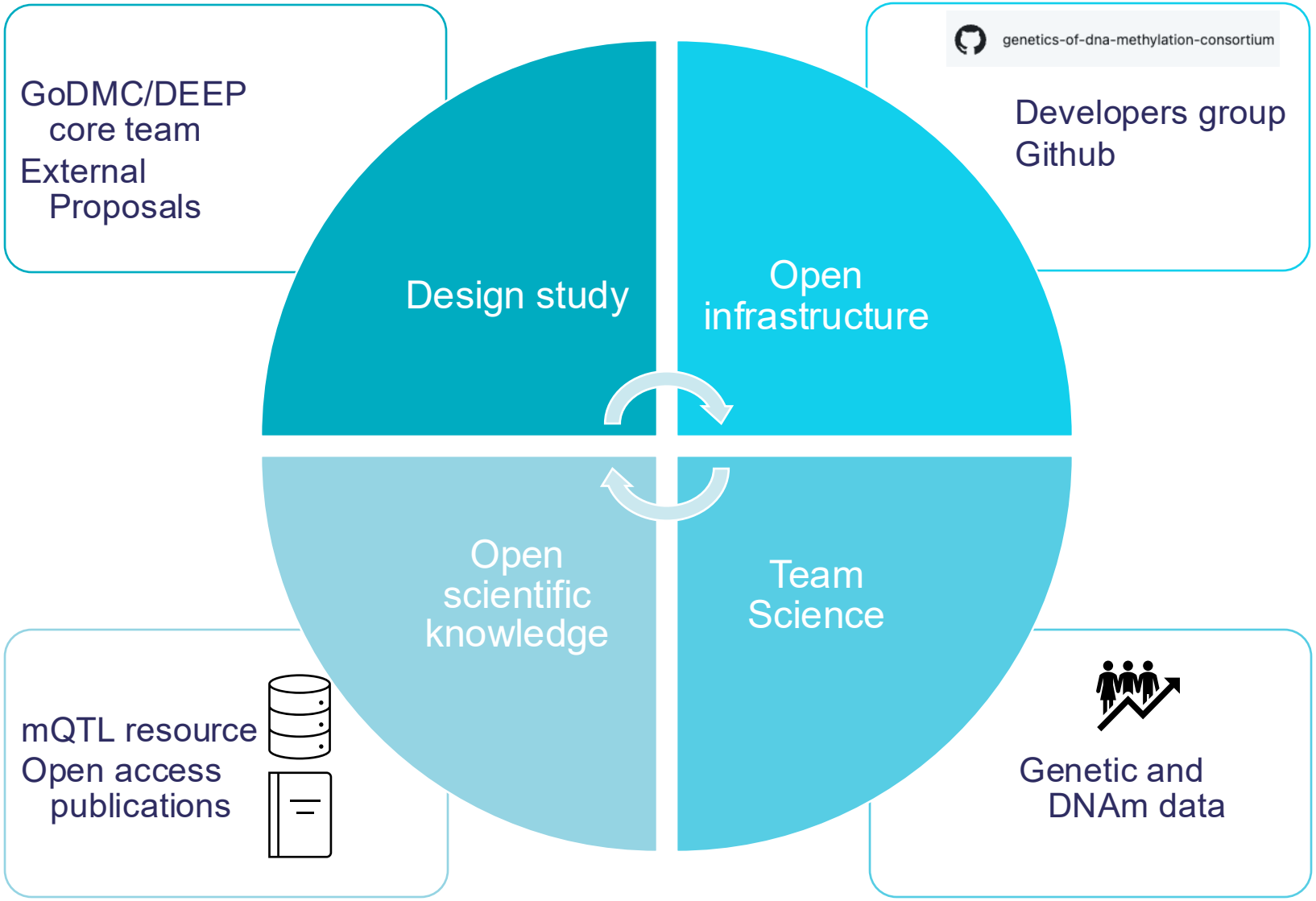
- Inform future analysis of health outcomes by identifying factors that contribute to inter-individual DNAm variation

Non genetic variation of DNA methylation

- Comparison between populations
  - EWAS of ancestry/ethnicity
  - EWAS of environmental exposures

  - little variation within population
  - large variation between populations
  - age at infection (CMV, EBV)




GoDMC/DEEP
core team
External
Proposals
Design study
Open
infrastructure
genetics-of-dna-methylation-consortium
Developers group
Github
Open
scientific
knowledge
Team
Science
mQTL resource
Open access
publications
Genetic and
DNAm data

Acknowledgements

IEU
Caroline Relton
Gibran Hemani
Hannah Elliott
Tom Gaunt
Haotian Tang

GoDMC
Developer's team
GoDMC core team

Diverse Epigenetic Epidemiology Partnership