INTRODUCTION TO EPIGENOME-WIDE ASSOCIATION STUDIES (EWAS)

3. EPIGENOME-WIDE ASSOCIATION STUDIES (EWAS) (THEORY)

EPIGENOME-WIDE ASSOCIATION STUDY (EWAS)

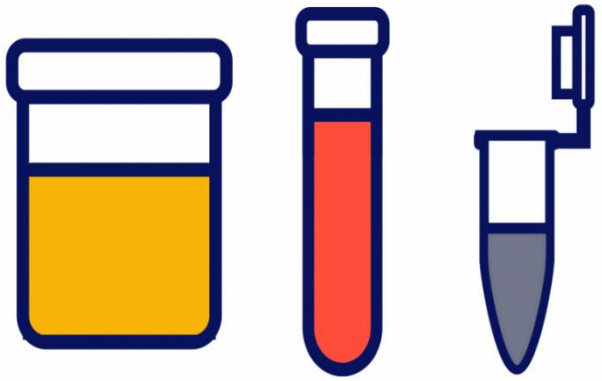
Workflow

1. Scientific question
2. Study population
3. Biological sample
4. DNA methylation data acquisition
5. Quality control of DNA methylation data
6. Epigenome-wide association study (EWAS)
7. Meta-EWAS or replication / validation
8. Biological interpretation

EPIGENOME-WIDE ASSOCIATION STUDY (EWAS)

Workflow

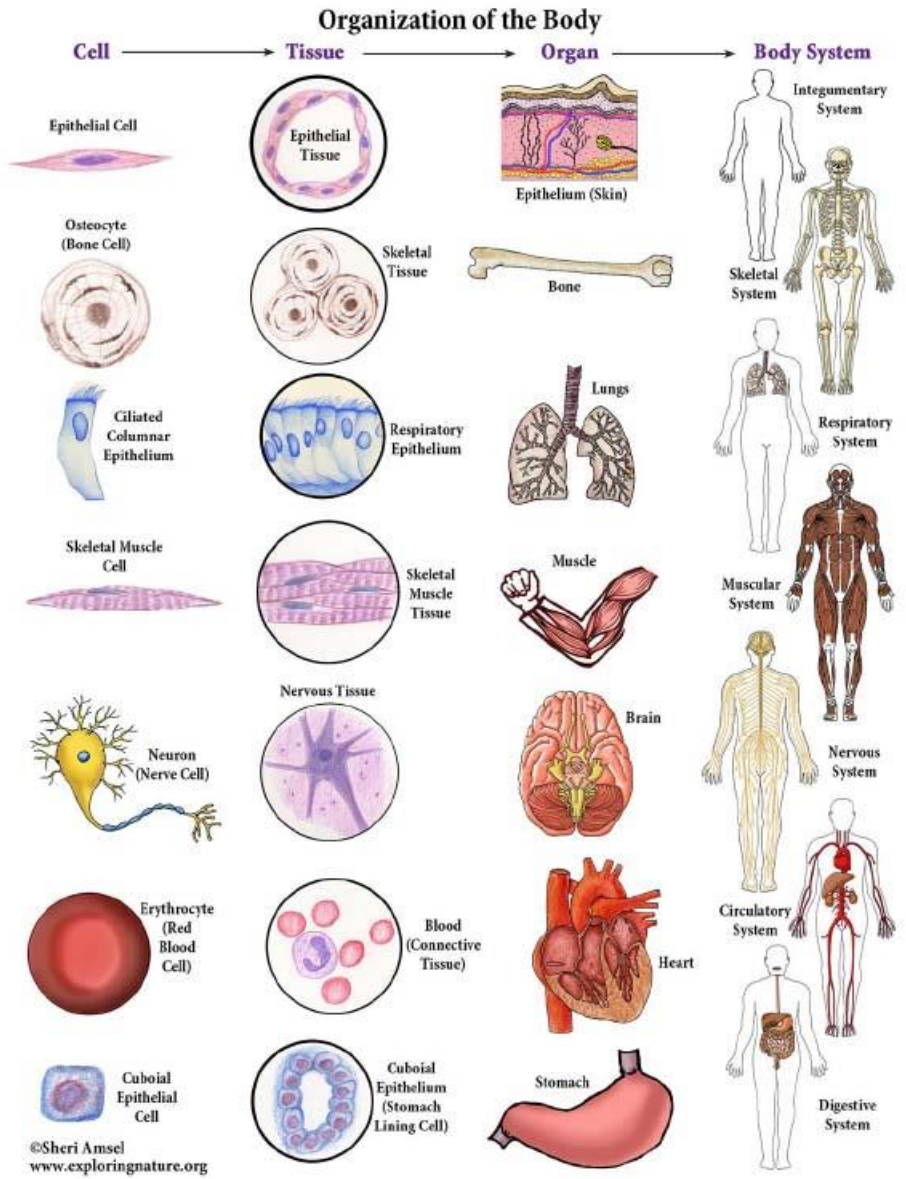
1. Scientific question
2. Study population
3. **Biological sample**
4. DNA methylation data acquisition
5. Quality control of DNA methylation data
6. Epigenome-wide association study (EWAS)
7. Meta-EWAS or replication / validation
8. Biological interpretation



3. BIOLOGICAL SAMPLE

Target tissue vs accesible tissue

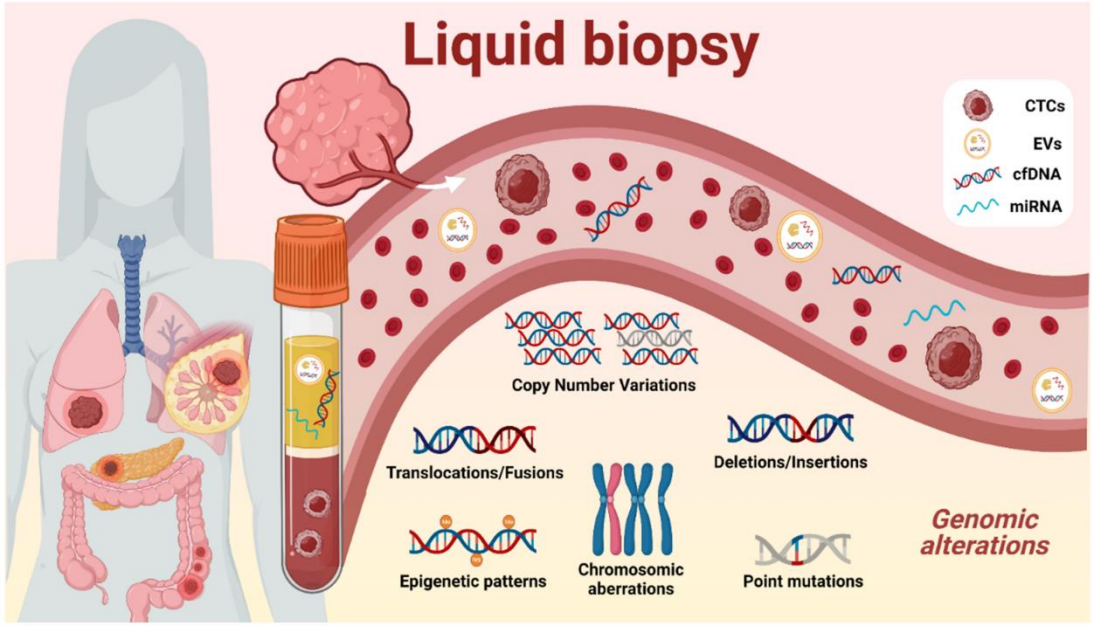
- Accessible tissues: blood, placenta, ….
- Proxy?

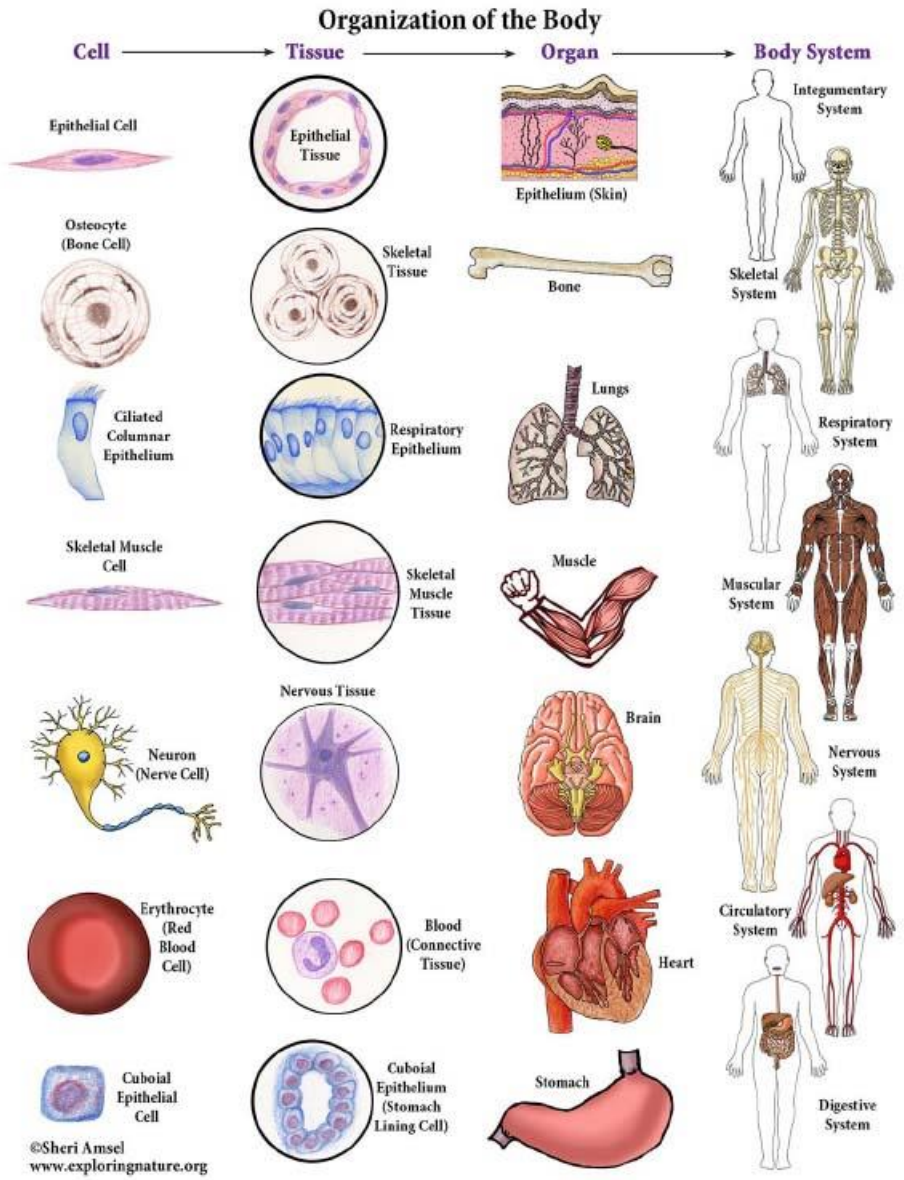


3. BIOLOGICAL SAMPLE

Target tissue vs accesible tissue

- Accessible tissues: blood, placenta, ….
- Proxy?

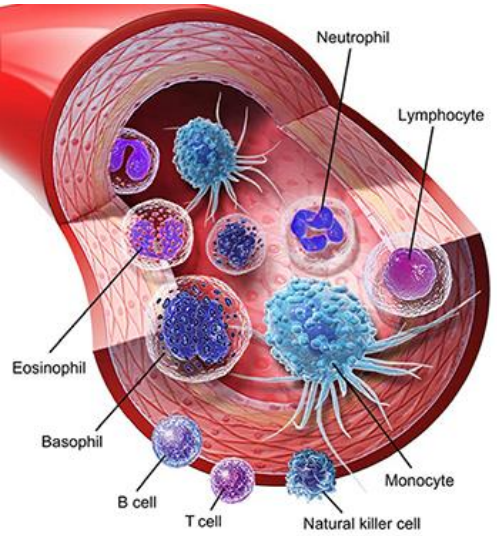


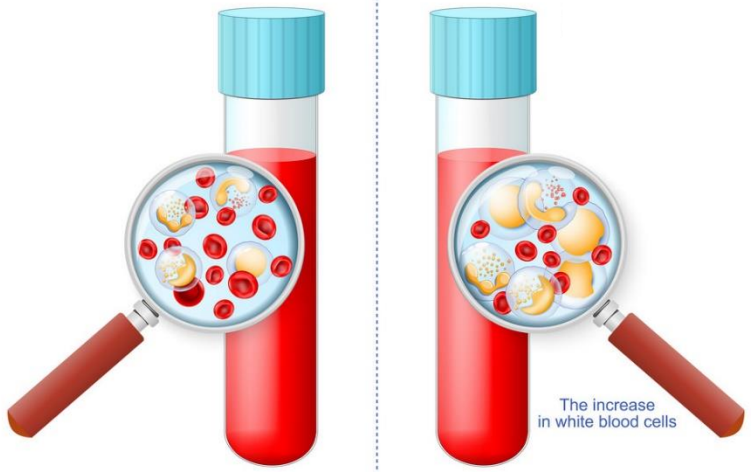


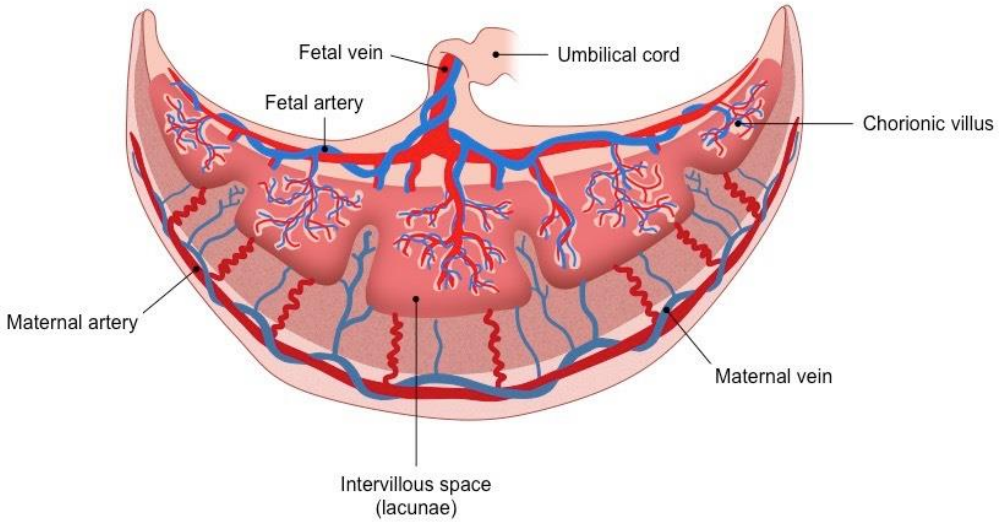
3. BIOLOGICAL SAMPLE

Tissue cellular composition

- Heterogeneity:
  - Different composition across individuals
  - Different place of biopsy (solid tissues)



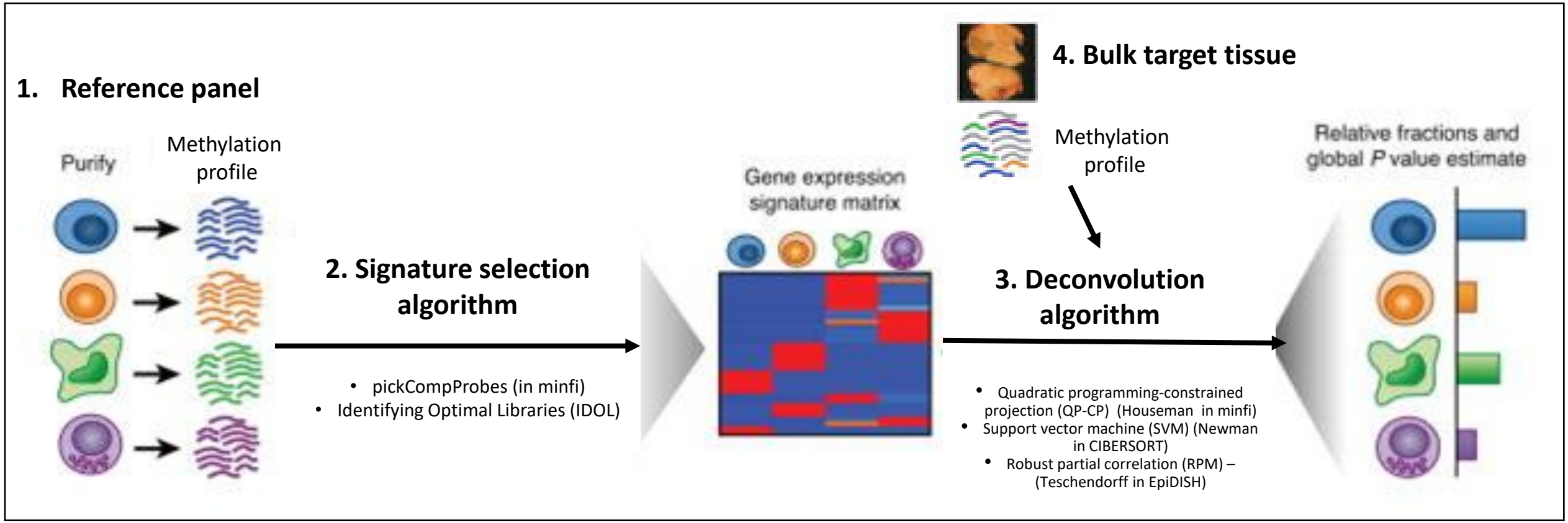




3. BIOLOGICAL SAMPLE

Cell mixture deconvolution

- Cell sorting and single cell/nuclei analysis
- Cell mixture deconvolution (statistical approach)



adapted from Newman 2015

3. BIOLOGICAL SAMPLE

Reference panels for cell deconvolution

FlowSorted.Blood.EPIC R package (but also adapted to 450K)

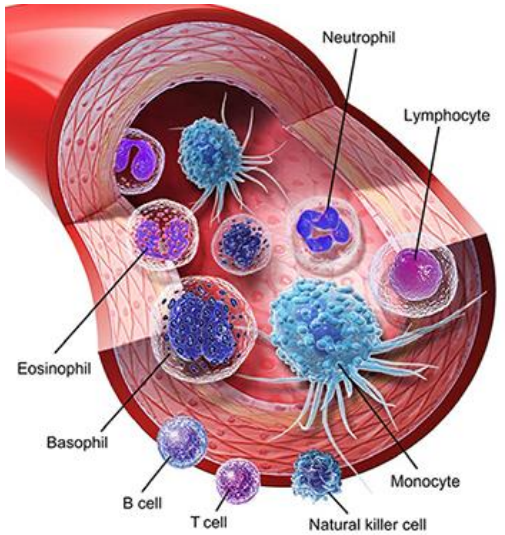
- 6 cells blood cells adults (TCD4, TCD8, Bcells, Mono, NK, Neu)

FlowSorted.BloodExtended.EPIC R package (but also adapted to 450K)

- 12 cells blood cells adults (Neu, Eos, Bas, Mono, Bnv, Bmem, CD4nv, CD4mem, Treg, CD8nv, CD8mem, and NK)

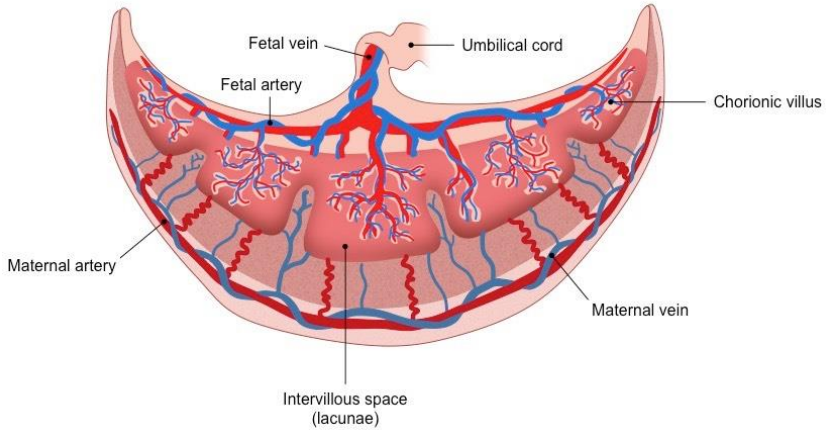
FlowSorted.CordBloodCombined.450k R package

- 7 cells cord blood cells (TCD4, TCD8, Bcells, Mono, NK, Neu, nRBC)



planet R package

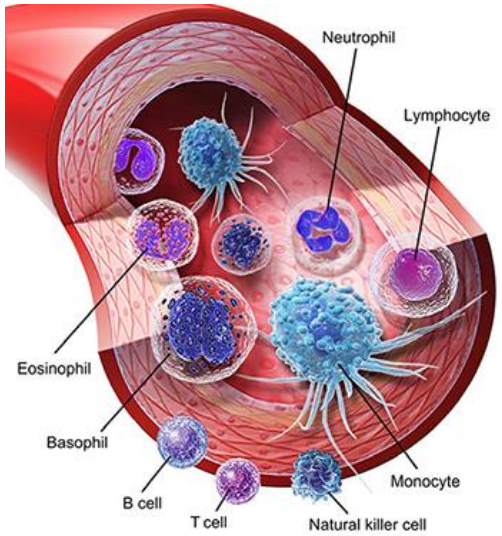
- 7 cells placenta cells (Trophoblasts, Synsitiotrophoblasts, Hofbauer,
- Stromal, Endothelial, nRBC)



3. BIOLOGICAL SAMPLE

Yesterday, already estimated with meffil R package

Reference panels for cell deconvolution



FlowSorted.Blood.EPIC R package (but also adapted to 450K)

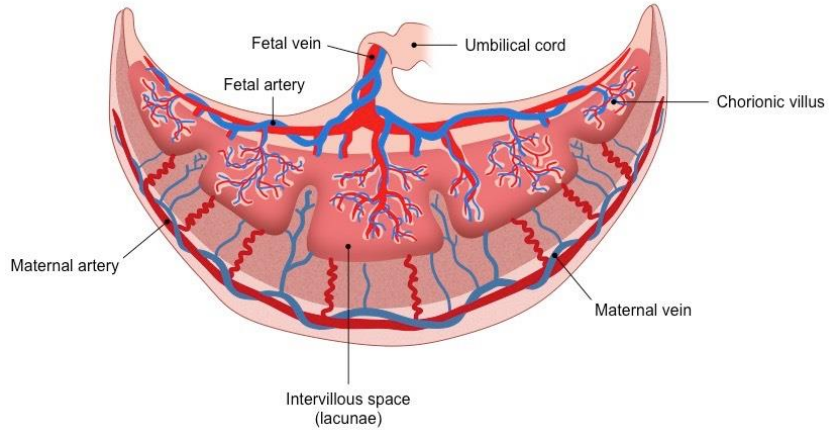
- 6 cells blood cells adults (TCD4, TCD8, Bcells, Mono, NK, Neu)

FlowSorted.BloodExtended.EPIC R package (but also adapted to 450K)

- 12 cells blood cells adults (Neu, Eos, Bas, Mono, Bnv, Bmem, CD4nv, CD4mem, Treg, CD8nv, CD8mem, and NK)

FlowSorted.CordBloodCombined.450k R package

- 7 cells cord blood cells (TCD4, TCD8, Bcells, Mono, NK, Neu, nRBC)



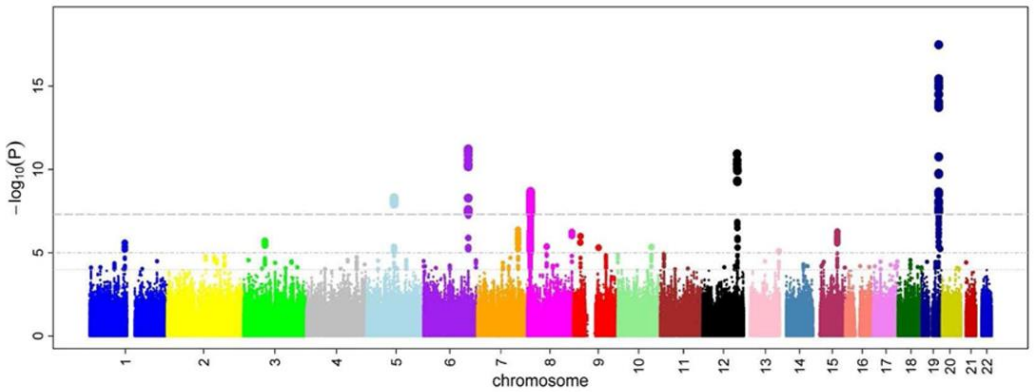
planet R package

- 7 cells placenta cells (Trophoblasts, Synsitiotrophoblasts, Hofbauer,
- Stromal, Endothelial, nRBC)

EPIGENOME-WIDE ASSOCIATION STUDY (EWAS)

Workflow

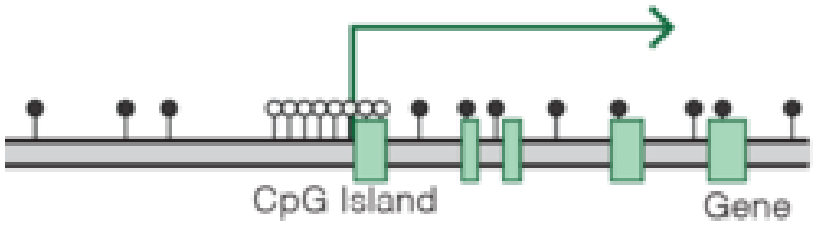
1. Scientific question
2. Study population
3. Biological sample
4. DNA methylation data acquisition
5. Quality control of DNA methylation data
6. **Epigenome-wide association study (EWAS)**
7. Meta-EWAS or replication / validation
8. Biological interpretation

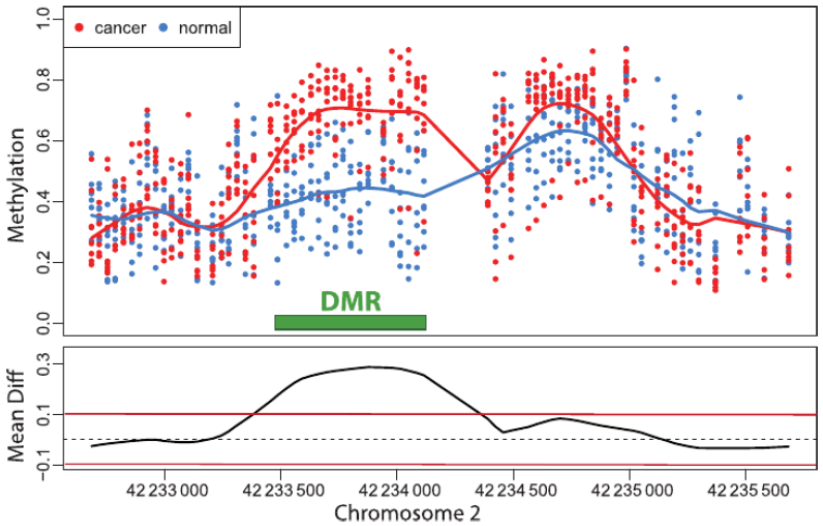


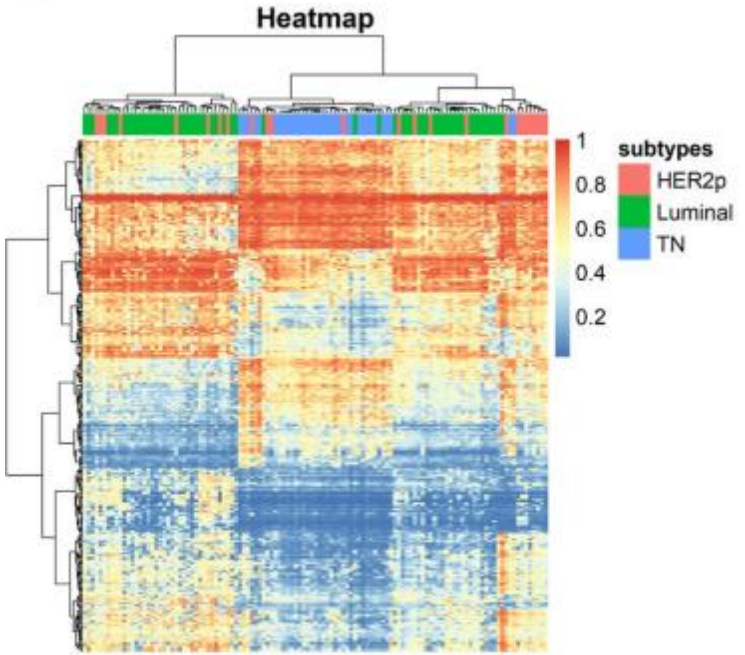
6. EWAS

Types of DNA methylation analyses

- By position: differently methylated position (DMP)
- By region: differently methylated region (DMR)
- All methylation dataset: cluster analysis, heatmap…



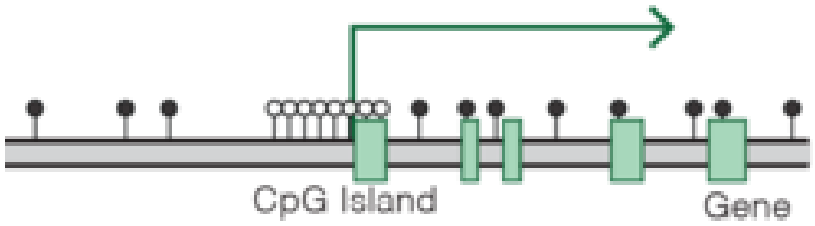


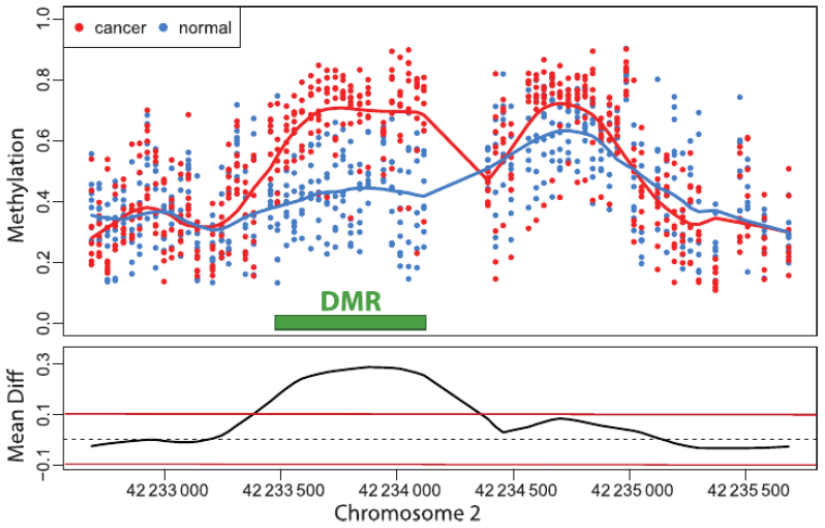


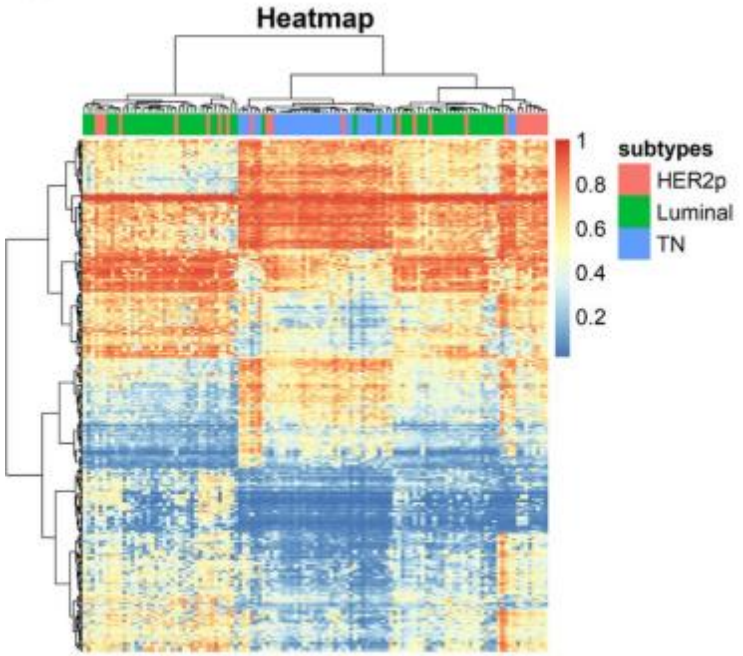
6. EWAS

Types of DNA methylation analyses

- **By position: differently methylated position (DMP)**
- By region: differently methylated region (DMR)
- All methylation dataset: cluster analysis, heatmap…







6. EWAS

Statistical test

Methylation as outcome (better to model):

- Linear or robust linear regression models (one per each CpG)
- Effect size:
  - change in methyl (from 0 to 1) by trait category (smoker/non-smoker)
  - change in methyl (from 0 to 1) by trait unit (unit of cotinine)

Methyl CpG1 (cont) = trait + cov
Methyl CpG2 (cont) = trait + cov
…

Methylation as predictor:

- Linear or logistic regression models (one per each CpG)
- Effect size:
  - Cont: change in units of disease trait (kg) by methyl unit (from 0 to 1)
  - Cat: change in disease odds by methyl unit (from 0 to 1)

Trait (cont or cat) = methyl CpG1 + cov
Trait (cont or cat) = methyl CpG2 + cov
…

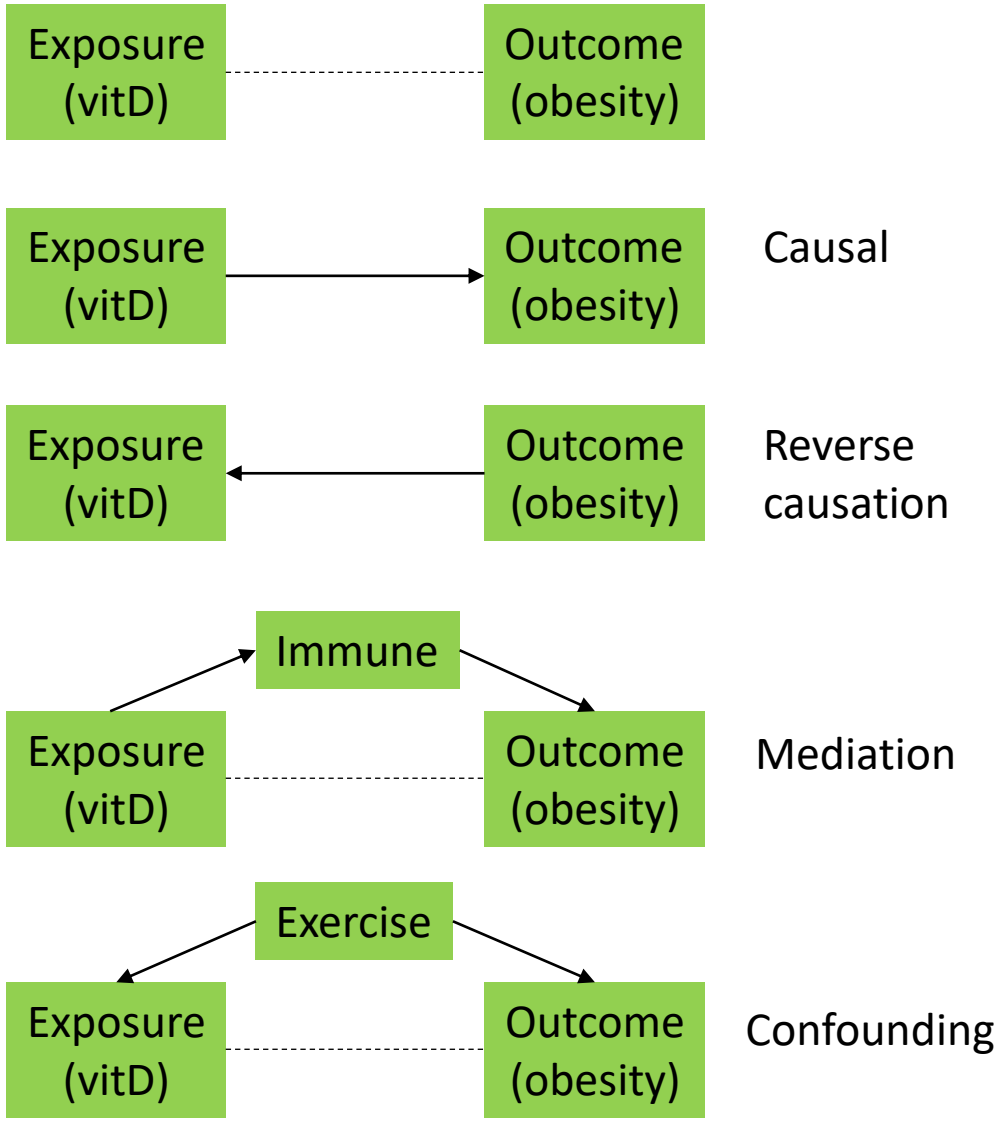
Tool: **meffil R package** (linear or robust regression)

6. EWAS

Differently from GWAS, EWAS can be confounded by several variables and can suffer from reverse causation.

Confounding

- A measured or unmeasured third variable that influences both the supposed cause and the supposed effect.
- How to select confounders:
  - A priori knowledge from literature
  - Directed Acyclic Graph (DAG) (https://www.dagitty.net/)
  - Cell types
  - Surrogate variables



6. EWAS

Surrogate variables (SVs) in EWAS

Surrogate variables are covariates constructed directly from high-dimensional data (ex. DNA methylation) that can be used in subsequent analyses to adjust for unknown, unmodeled, or latent sources of noise.

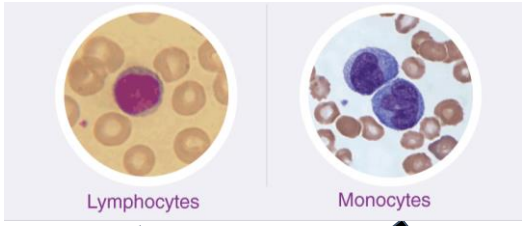
- Biological variables: sex, cell type proportions, ancestry, etc…
- Technical variables: slide, plate, etc...

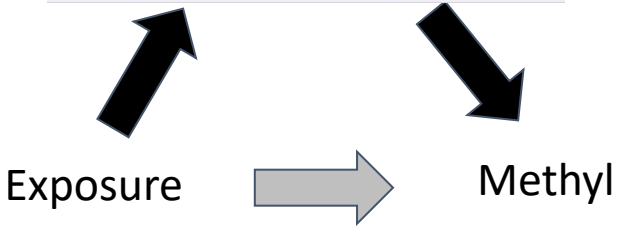
sva R package

6. EWAS

Cell type proportion in EWAS: confounding and mediation

Causation (mediation by cell%)

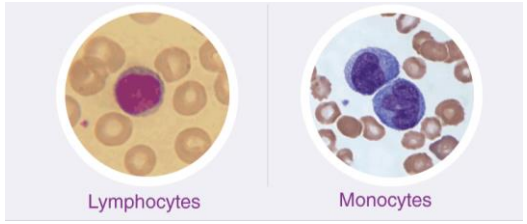


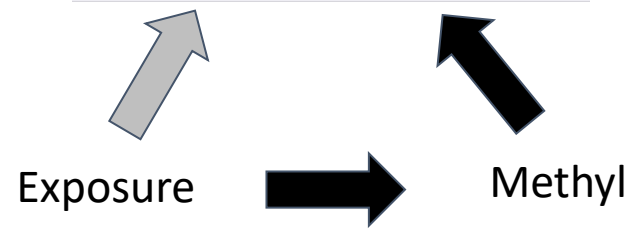


Example:
Smoking -> Inflamation -> Diff cell%
Diff cell% -> Diff methyl

Methyl = Exp + Cov

Causation (cell% final outcome)

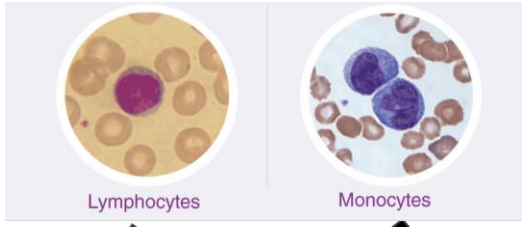


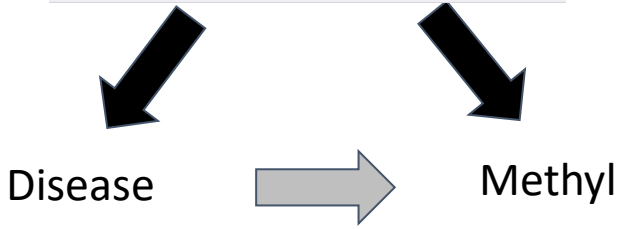


Example:
Smoking -> Diff methyl -> Diff cell%
Diff cell% -> Inflamation

Methyl = Exp + Cov

Confounding





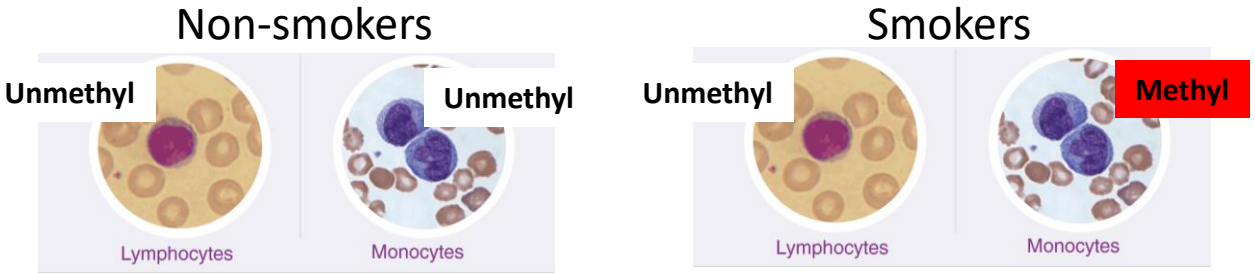
Example:
Inflamation -> Diff cell% -> Allergy
Diff cell% -> Diff methyl

Methyl = Exp + Cov + Cells

6. EWAS

Cell type proportion in EWAS: interaction or cell type specific effects

Example: smoking affects DNA methylation in CpG1, only in monocytes



Methyl = Trait * Cell type% + Cov

Tools:

- EpiDISH (CellDMC)
- RaMWAS

You et al. Nat Comm 2020

6. EWAS

Multiple-testing correction

Methods

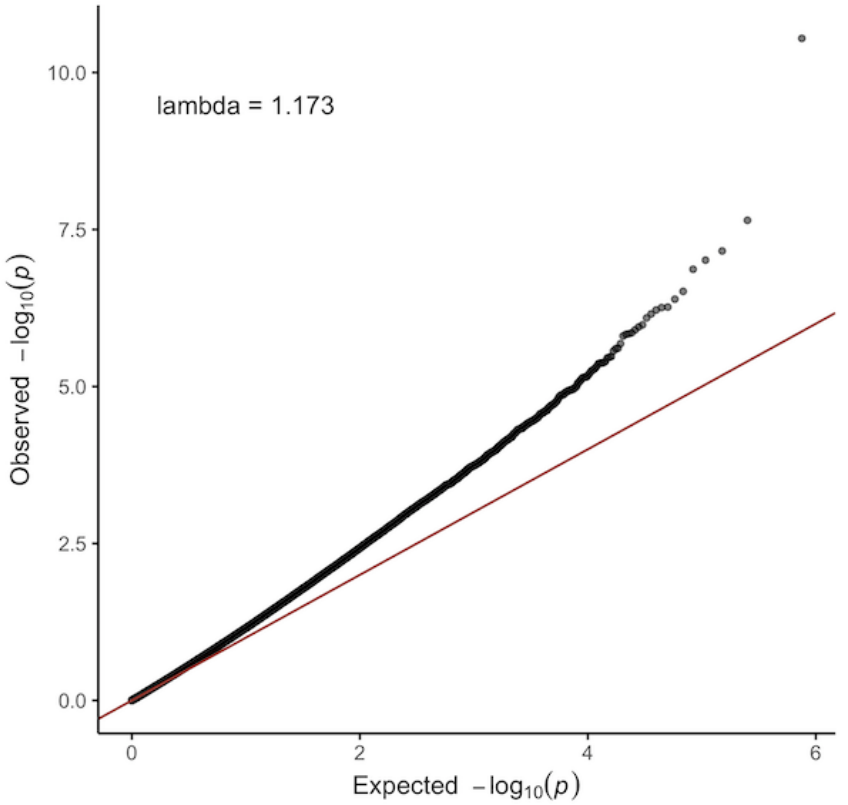
- Bonferroni (p-value 0.05 / N CpGs)
- False Discovery Rate (FDR)

QQ plots and lambda inflation factor

- Ratio of the median of the empirically observed

distribution of the test statistic to the expected median

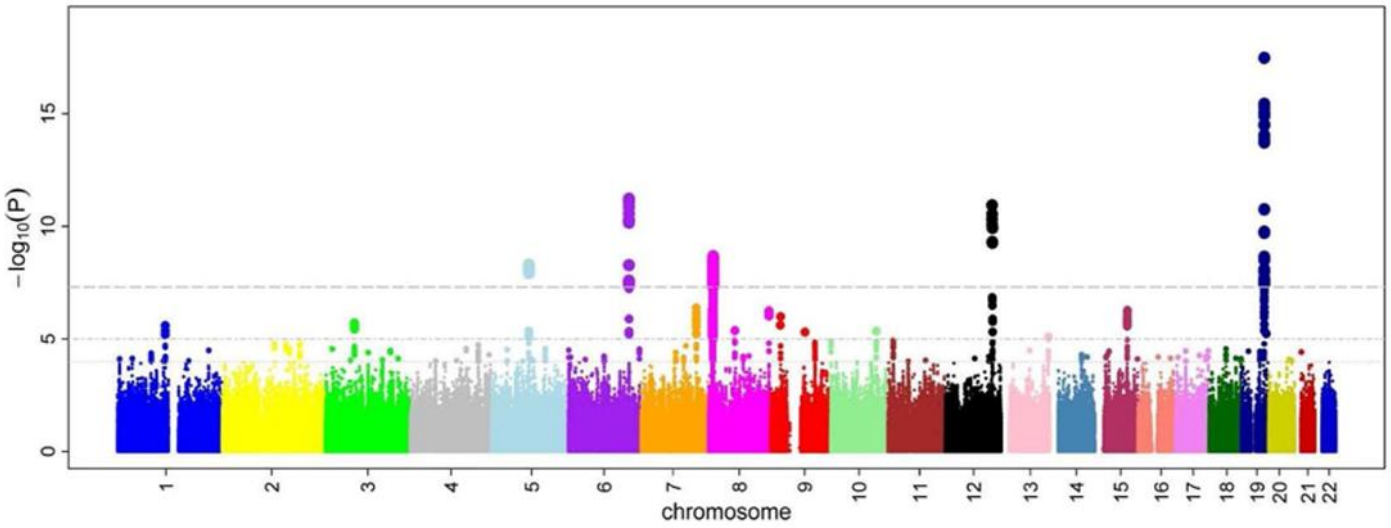
- In general close to 1
- If lower:
  - technical bias, missing adjustment…
- If higher:
  - true biological signals
  - technical bias



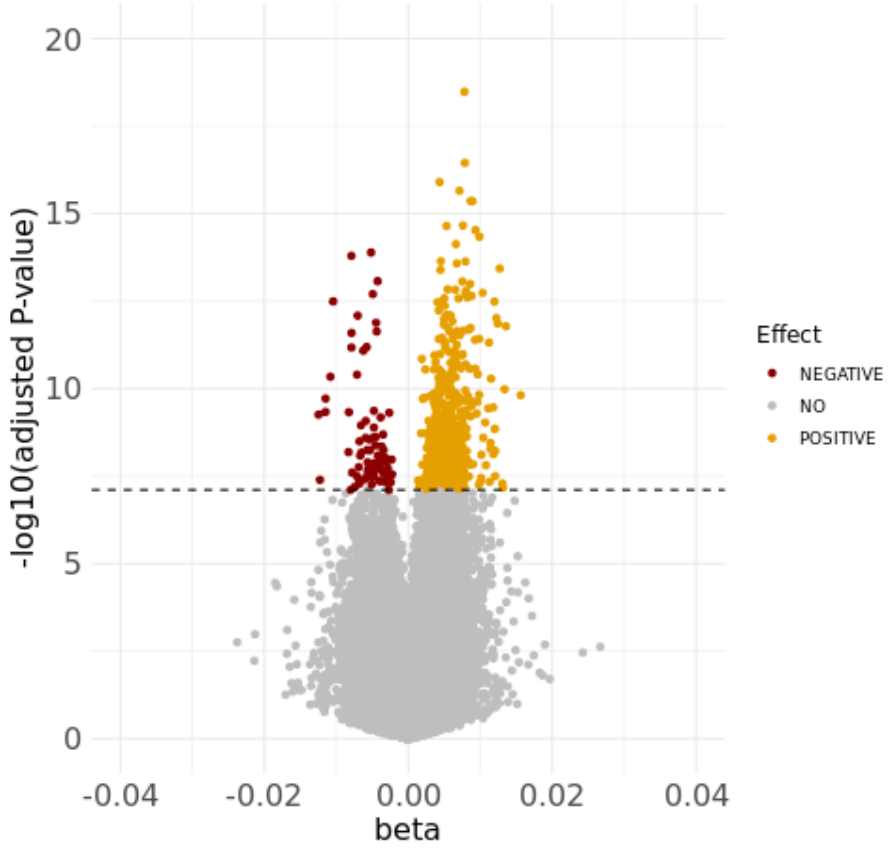
6. EWAS

Visualitzation of genome-wide results

Manhattan plot



Volcano plot



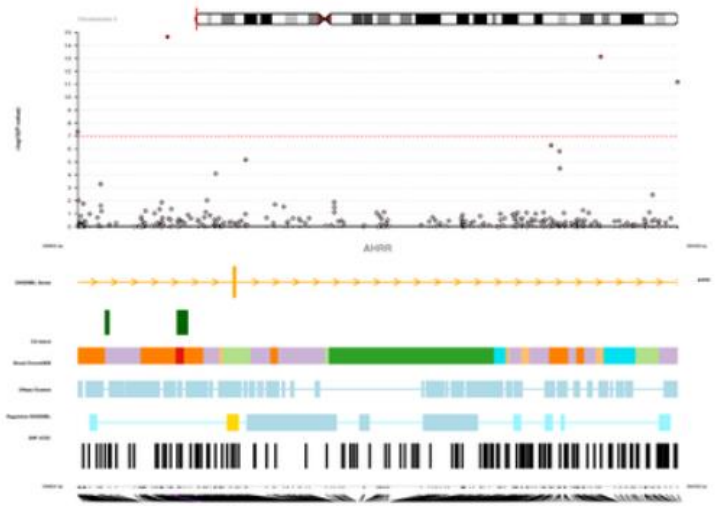
6. EWAS

Visualitzation of locus specific results

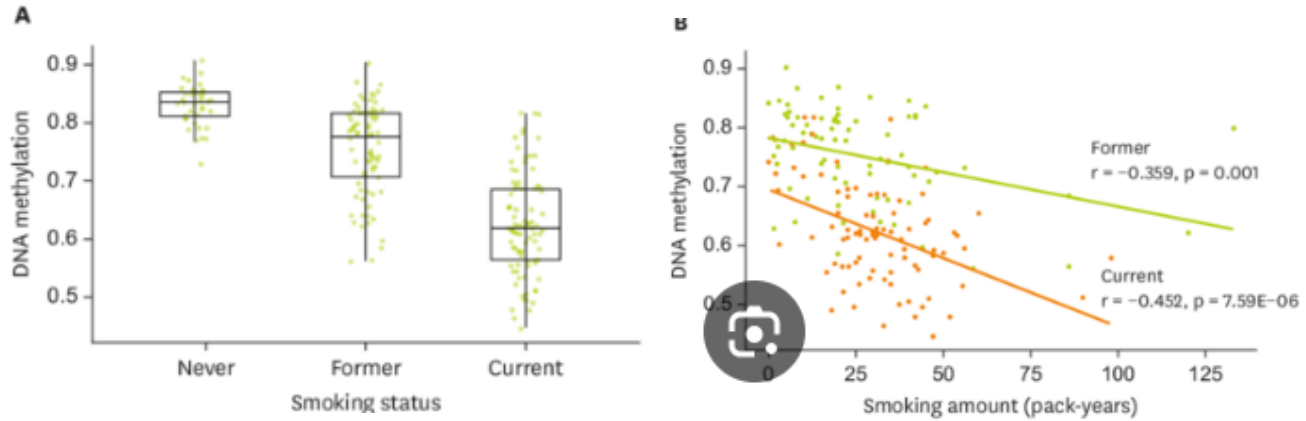
Table of results (CpG vs locus)

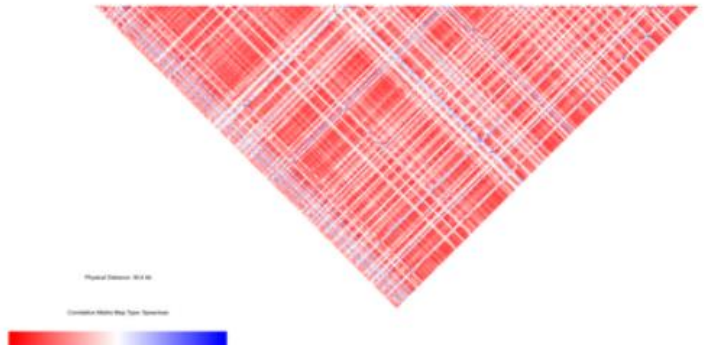
ProbeID	Beta	SE	*P* value	FDR	Bonferroni	Chromosome	Position
cg07504545	− 0.042	0.008	2.32×10^{-07}	0.109	0.109	1	203456019
cg04977770	− 0.054	0.011	7.15×10^{-07}	0.168	0.335	17	79846763
cg23625106	− 0.023	0.005	1.47×10^{-06}	0.193	0.689	8	61789727
cg08461451	− 0.036	0.008	1.73×10^{-06}	0.193	0.812	19	2295092
cg06367149	0.057	0.012	2.08×10^{-06}	0.193	0.973	15	61254575
cg13396019	− 0.026	0.005	2.57×10^{-06}	0.193	1.000	1	220564510
cg10541930	− 0.010	0.002	4.32×10^{-08}	0.020	0.020	10	131909085

Comet plot



Box plot
Scatter plot





INTRODUCTION TO EPIGENOME-WIDE ASSOCIATION STUDIES (EWAS)

3. EPIGENOME-WIDE ASSOCIATION STUDIES (EWAS) (PRACTICAL SESSION)

EWAS OF CURRENT AND FORMER SMOKING

Data: Cohort 1 (N = 294)

- Array: 450K
- Tissue: blood
- Ancestry: White European
- Sex: males and females
- Smoking: never, former, current
- Age: yes
- Array batch: yes
- Cells: yes

Input: ExpressionSet with matrix of beta values + covariates dataframe (exposure, covariates, cells)

Output (for current and former):

- results dataframes (not adj, adj, **adj and sva**)
- report (descriptive, QQ plot and lambda, Manhattan plot, Box plots)
- Volcano plot and Manhattan plot

Tool: meffil R package

Questions:

1. Which is the lambda of the unadjusted EWAS of current smoking? How does it change in adding covariates and surrogate variables?
2. How many CpGs are associated with current smoking (after False Discovery Rate – FDR - correction) in the model adjusted by the sva?
3. How many of the FDR CpGs show higher methylation and how many lower methylation?
4. Which is the top 1 CpG? In which chromosome is located?