Epigenetics & High Dimensional Epigenome-Wide Association Studies (EWAS) for Perinatal Epidemiology

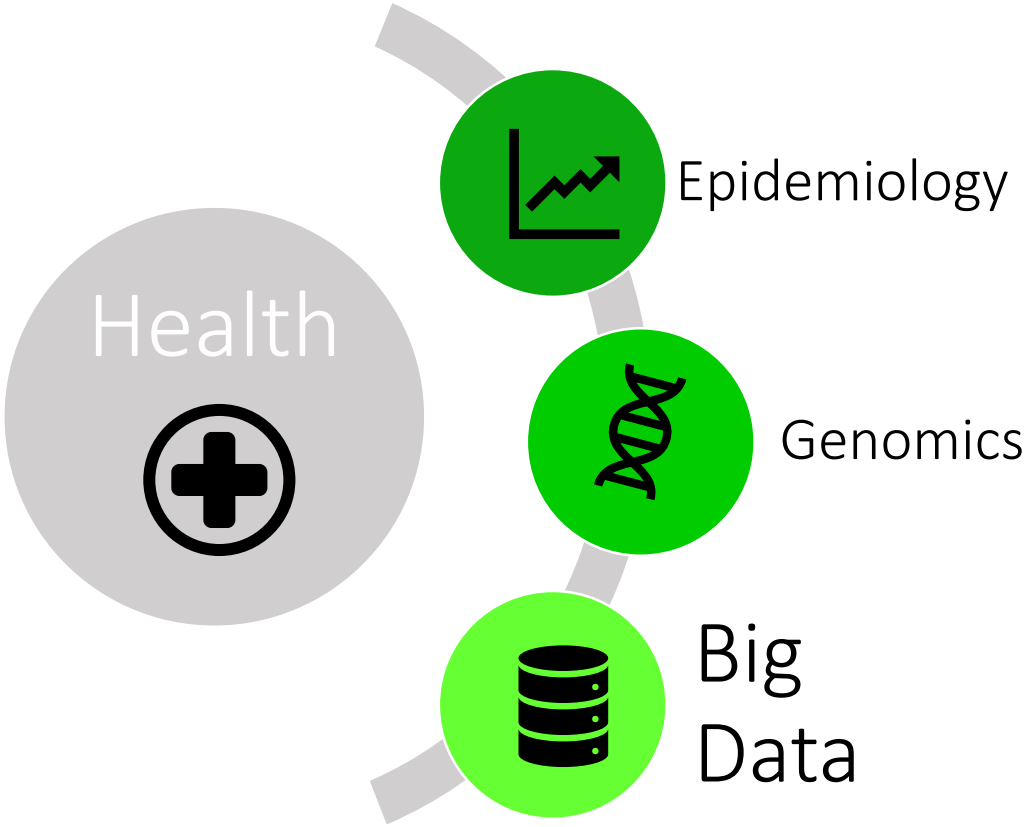
SPER Advanced Methods Workshop 2018

Baltimore, Maryland

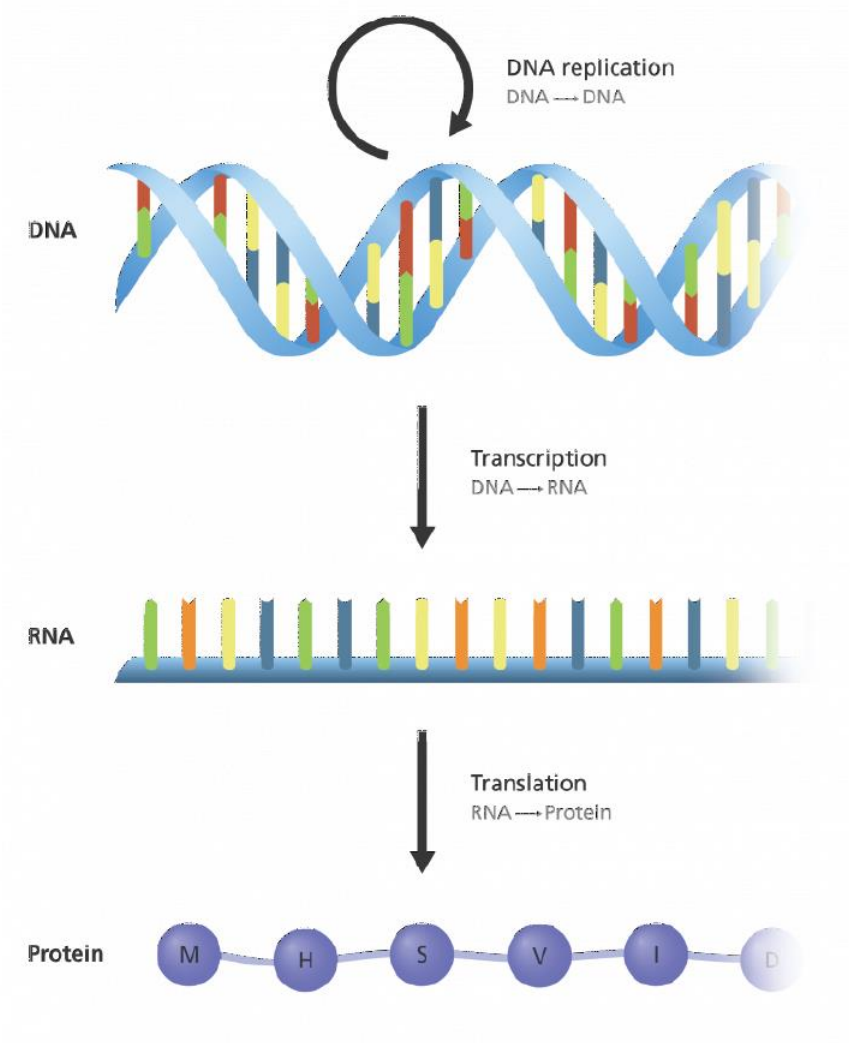
Andres Cardenas, PhD, MPH
Postdoctoral Research Fellow
Department of Population Medicine
Harvard Medical School

Outline

- **Epigenetics**
  - Working understanding of Epigenetics
- **DOHaD**
  - Fetal epigenetic programming
- **Applications to Birth Cohorts**
  - Methods/tools in Epigenetic Epidemiology
- **Methods & Study Design Considerations**
  - Strengths and Limitations
- **Examples - Implementation**
  - Interpreting studies

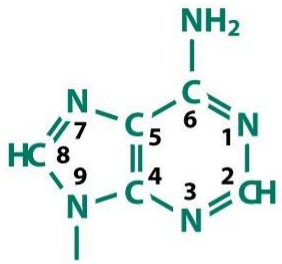


Central Dogma of Molecular Biology

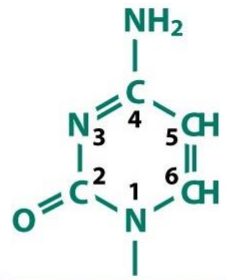


- Replication (DNA -> DNA)
- Transcription (DNA -> RNA)
- Translation (RNA-> Protein)

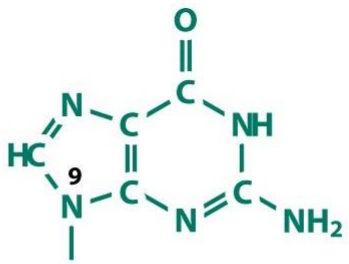
Universal Code of Life



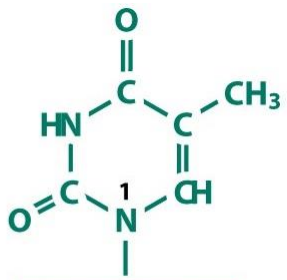
Adenine (A)



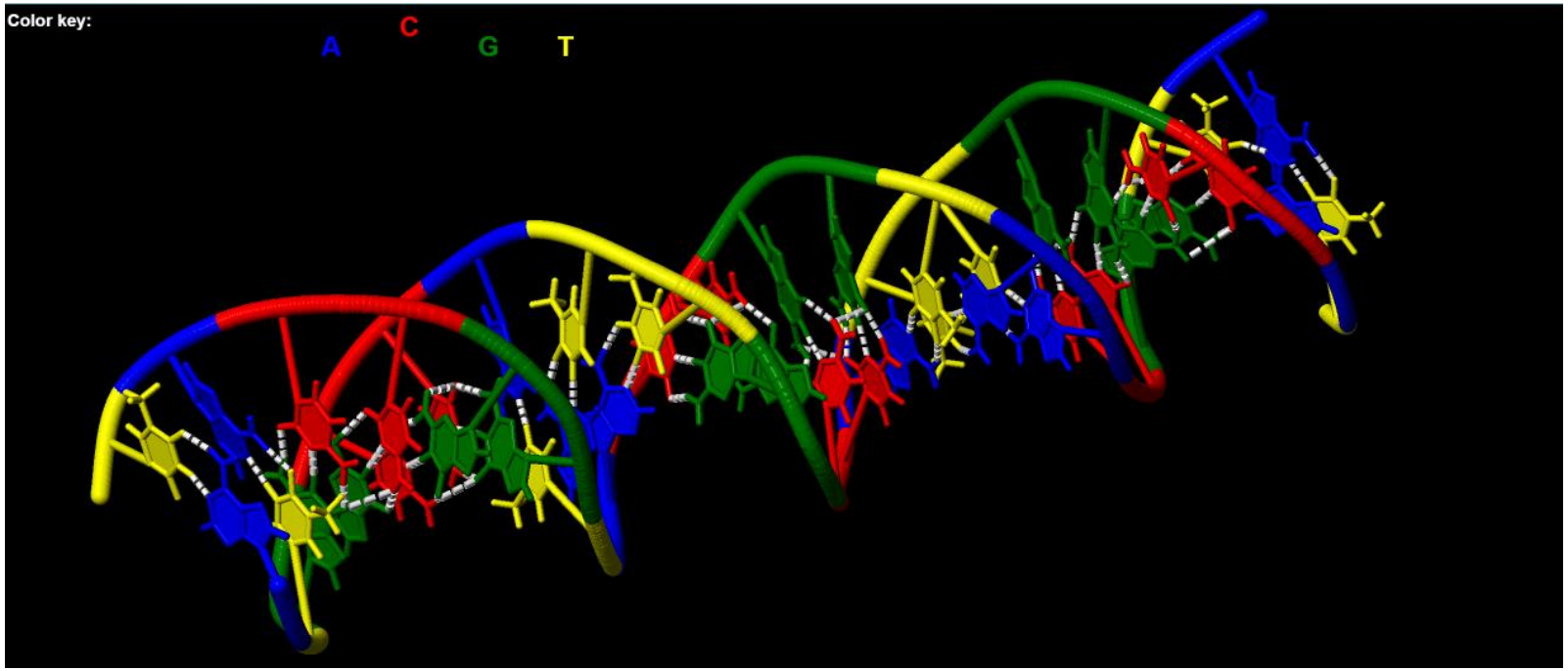
Cytosine (C)



Guanine (G)



Thymine (T)



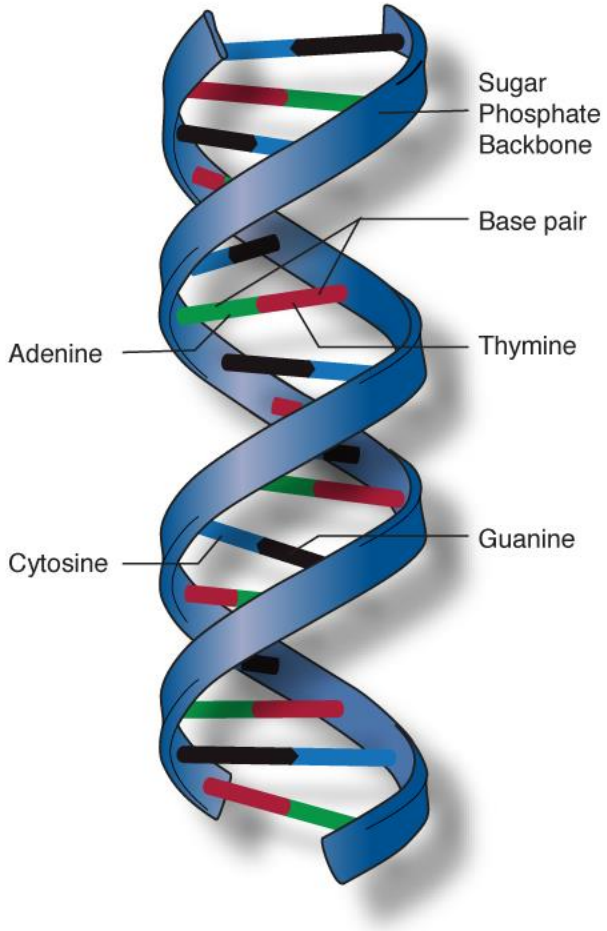
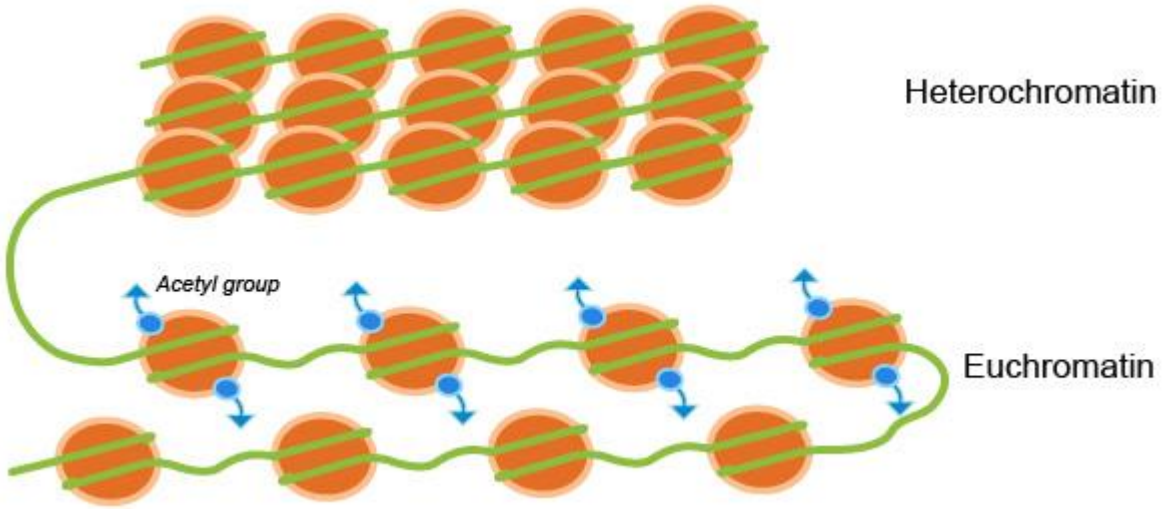
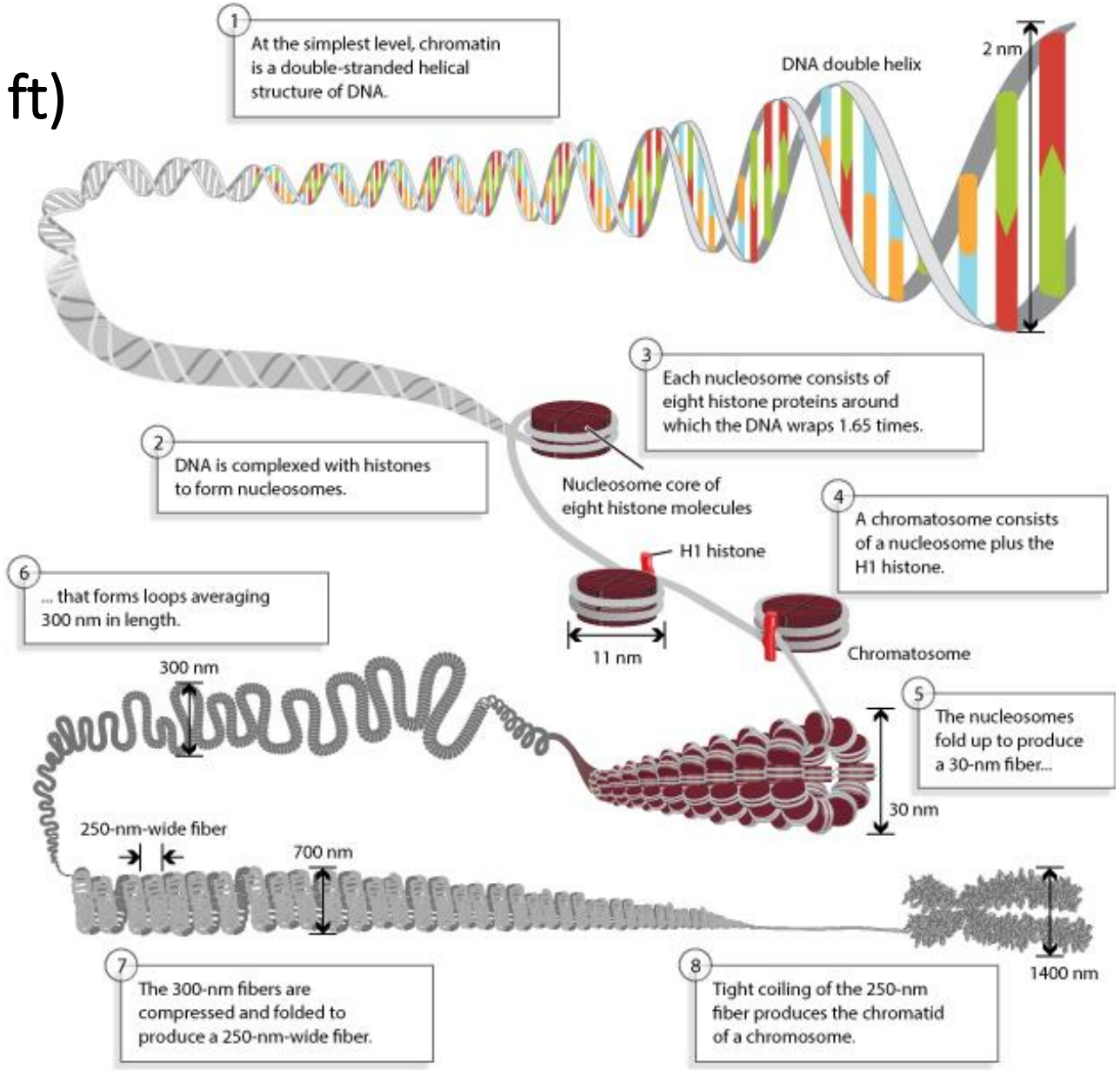


Image credit: https://geneed.nlm.nih.gov

DNA Inside the Cell

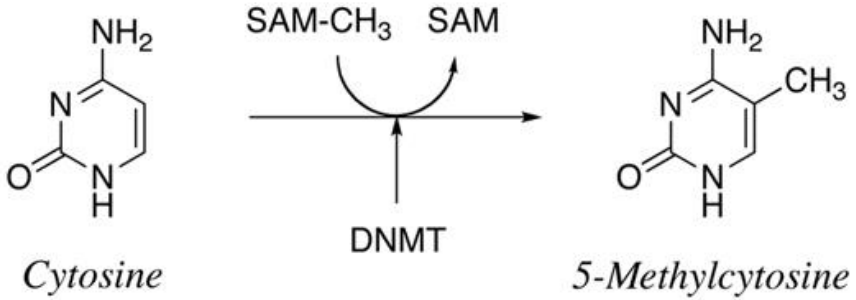
- One human cell ~2 meters DNA (6.6 ft)



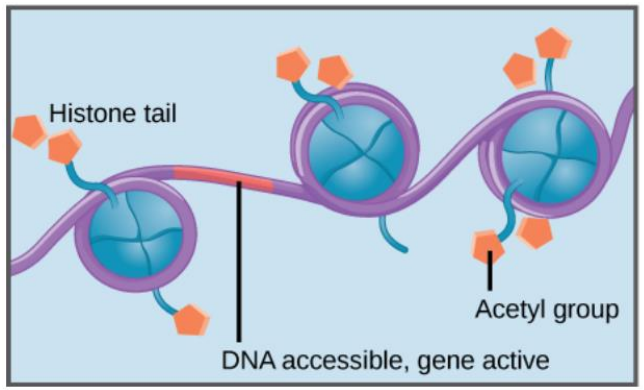


Common Epigenetic Modifications

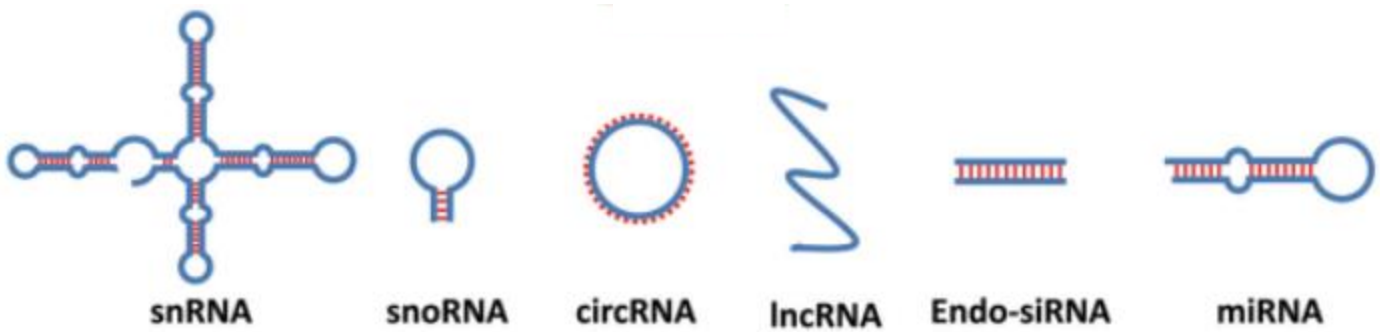
- DNA Methylation



- Histone Modifications

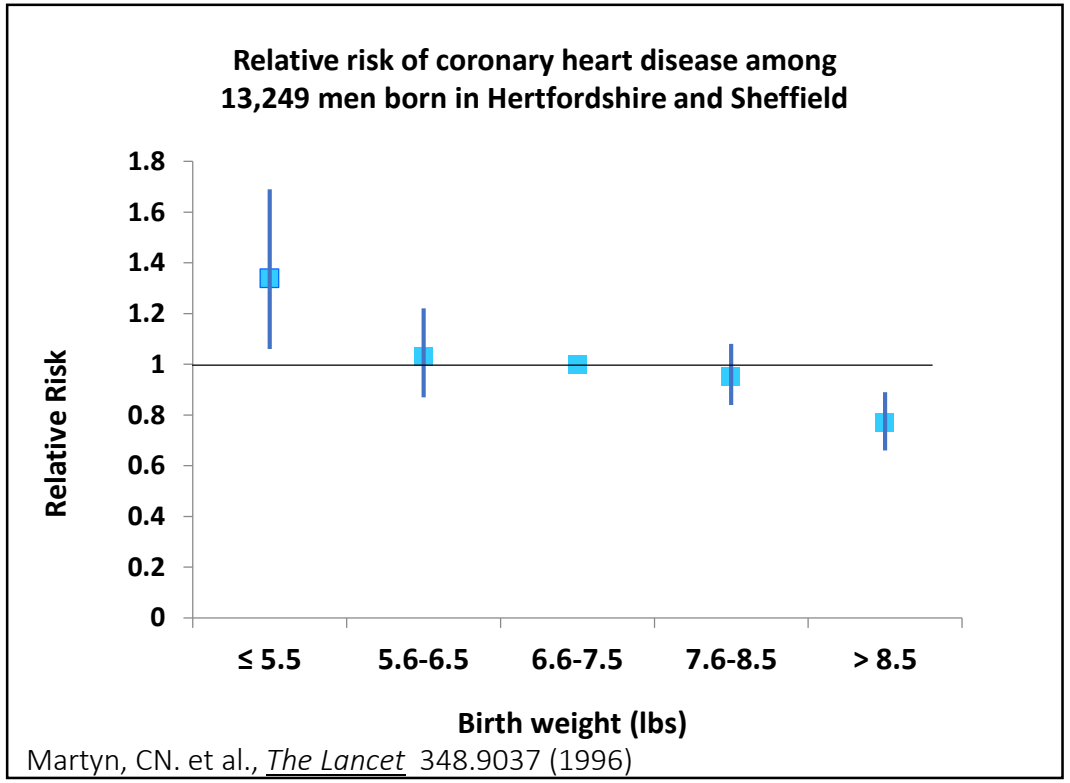


- Non-coding RNAs



Developmental Origins of Health and Disease

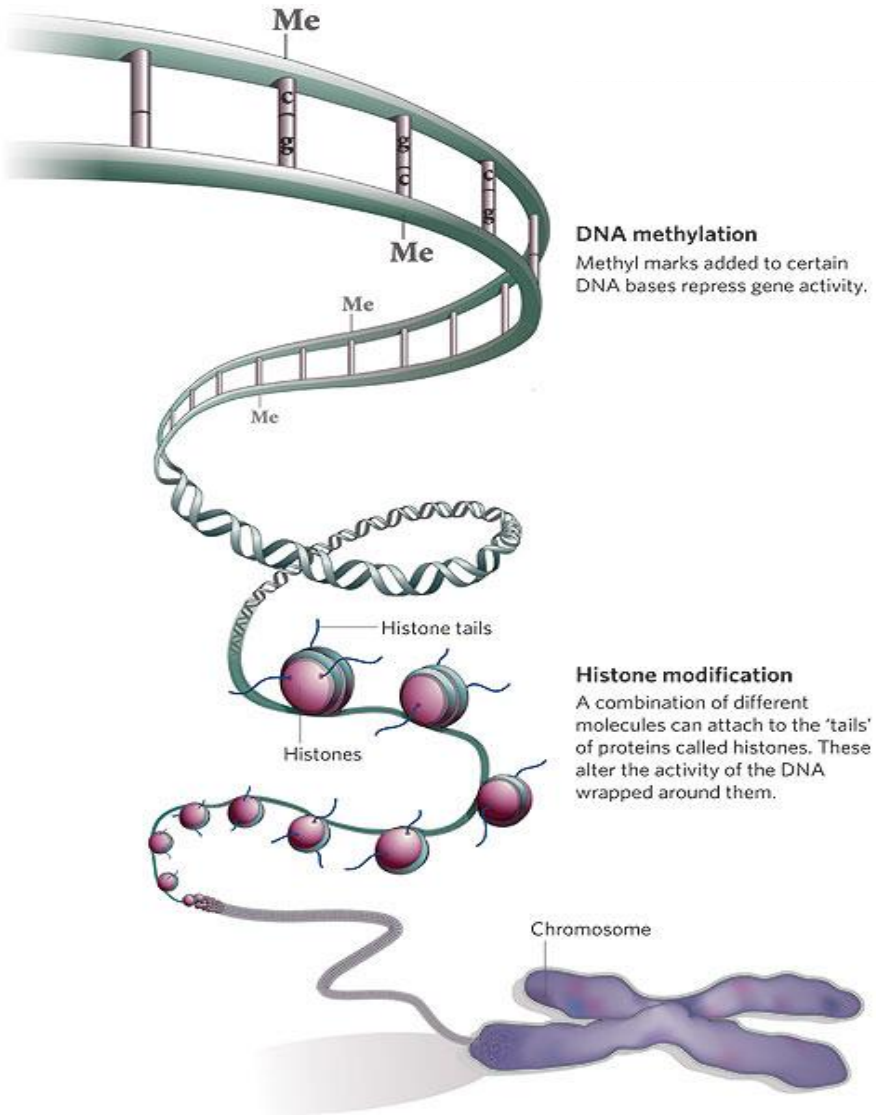
- Barker's hypothesis
  - Fetal programming of adult disease
  - Low birth weight/size -> disease (CHD)



Martyn, CN. et al., *The Lancet* 348.9037 (1996)

- Epigenetics as a potential mechanism/biomarker
  - Interface between genome and the environment

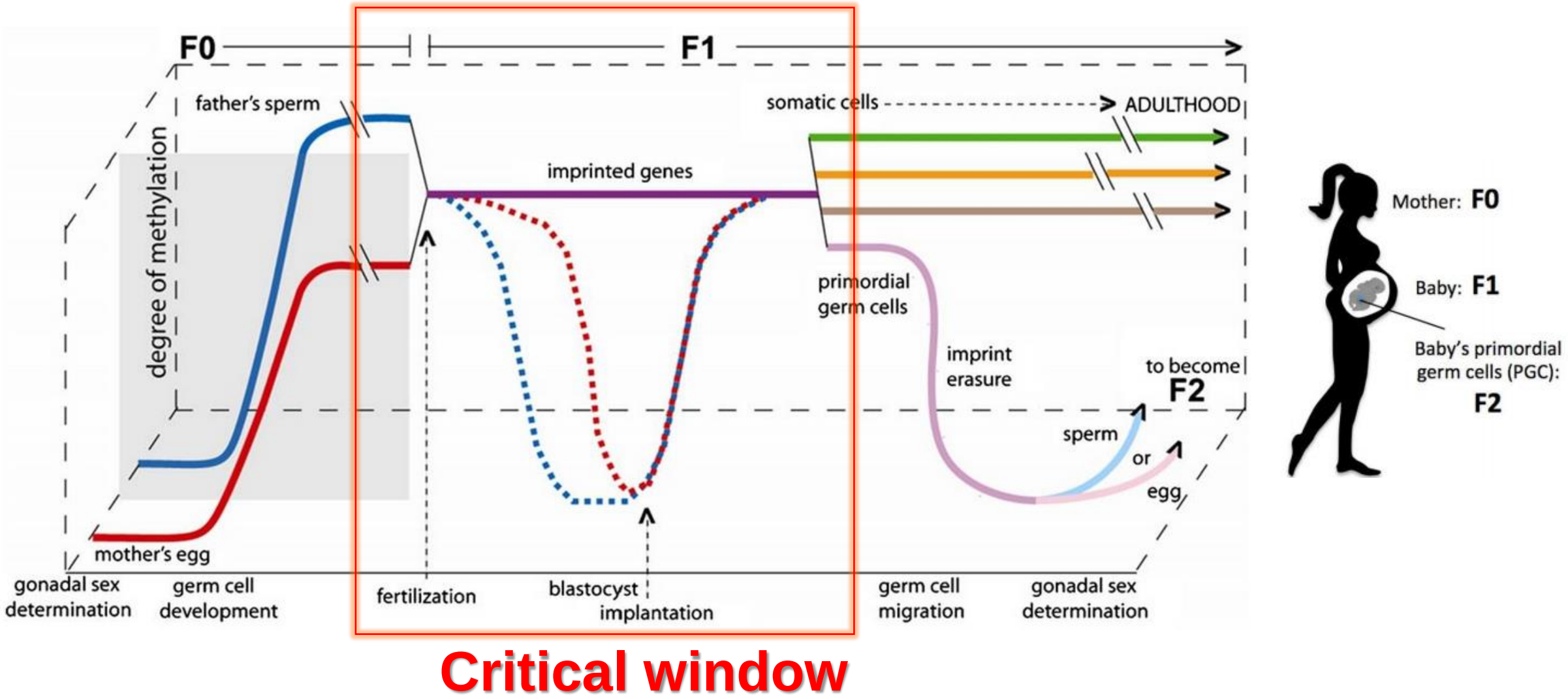
Epigenetics



Epigenetics

- Changes in gene expression that:
  - Do not depend on the DNA sequence
  - Can be stable
    - Through cell division (mitotically stable)
    - Transgenerational inheritance (limited evidence in humans)
- May persist even in the absence of the conditions that established them (Biological memory-> Biosensor)

Epigenetics - Fetal Programming *in Utero*

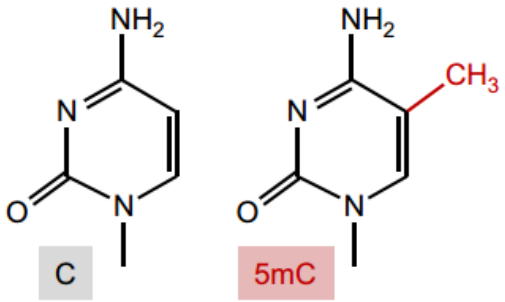


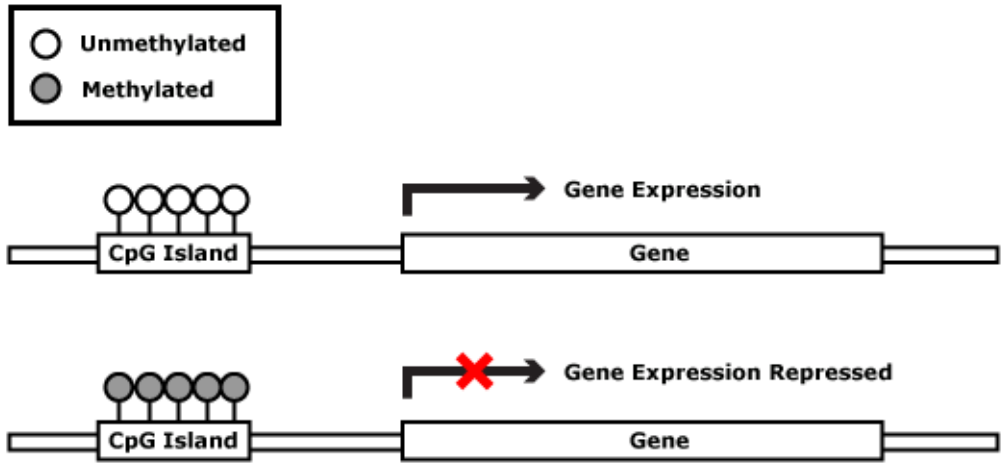
Critical window

Adapted from F. Perera & J. Herbstman *Reproductive Toxicology* 31.3 (2011): 363-373

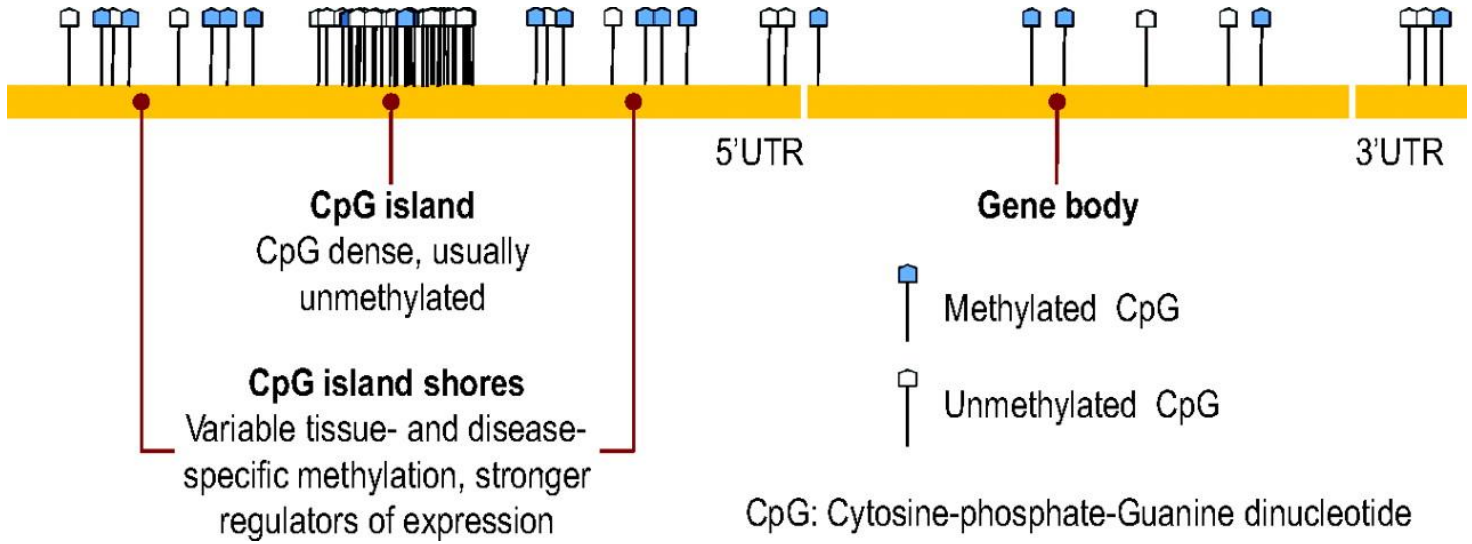
DNA Methylation - Fetal Programming *in Utero*

- DNAm as a molecular switch



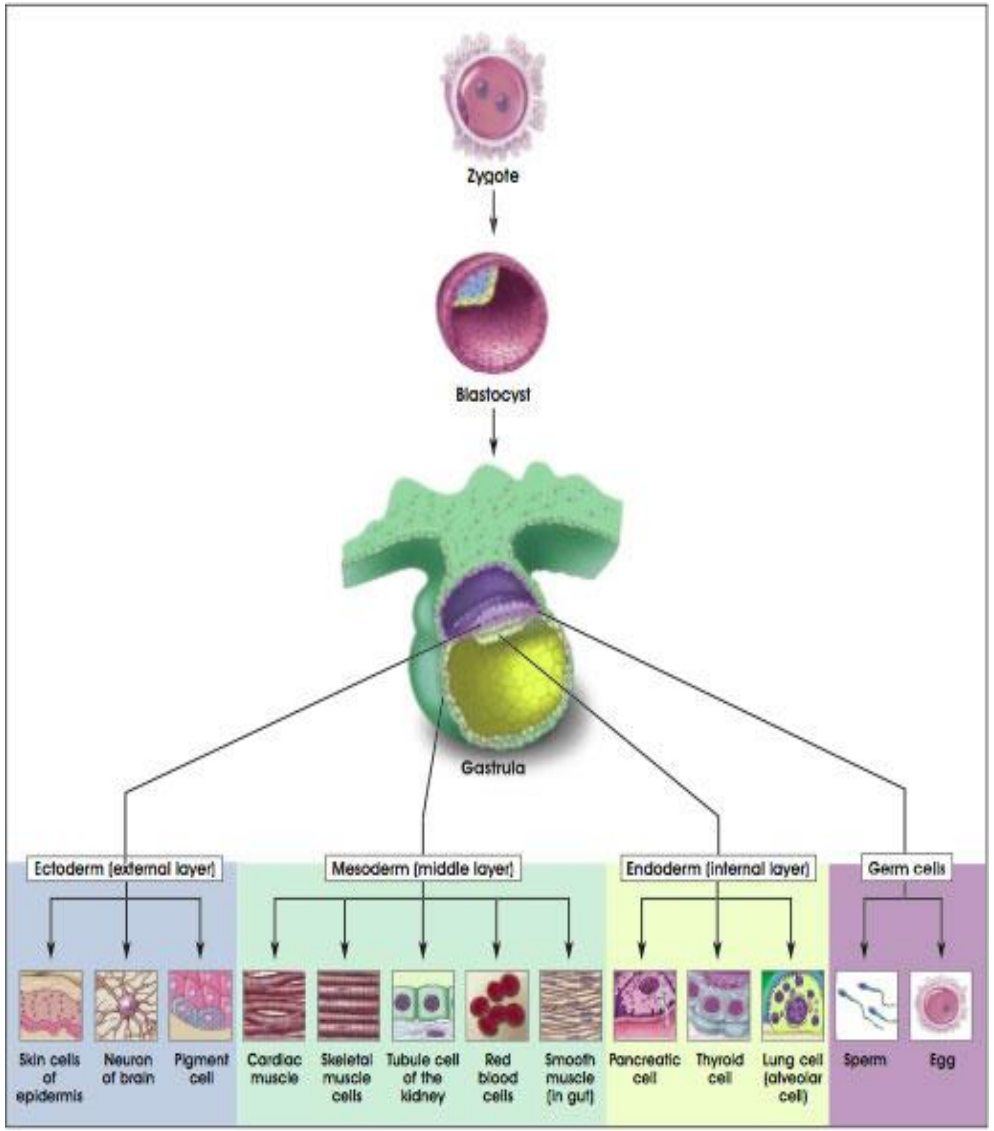


- Genomic regions associated with tissue differentiation and oncogenesis



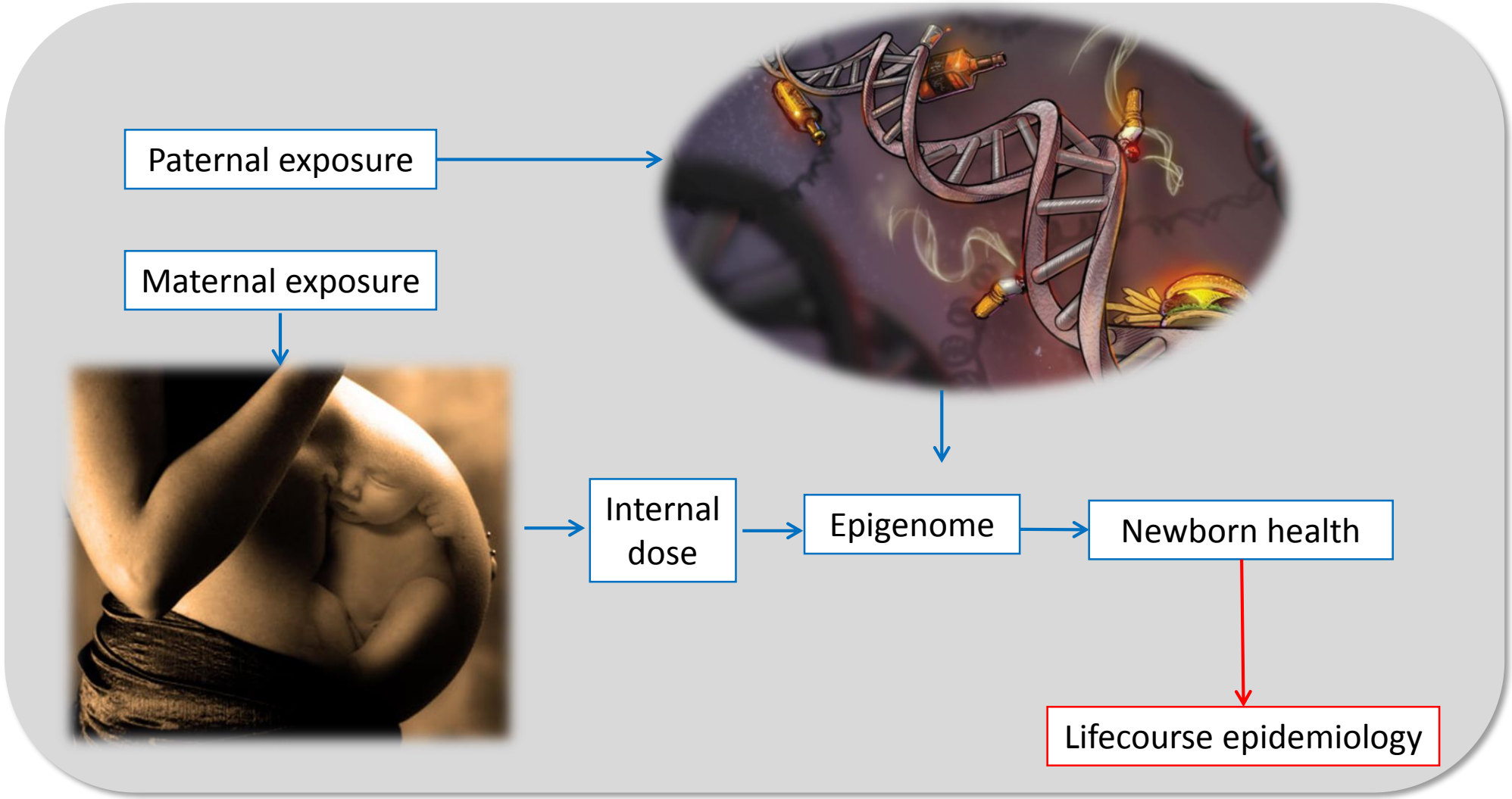
DNAm: tissue specificity

- Epigenetics contributes to tissue differentiation
- Epigenetic marks are cell-type and tissue specific
- Each cell type has a unique epigenetic signature
- A challenge and opportunity for epidemiological studies



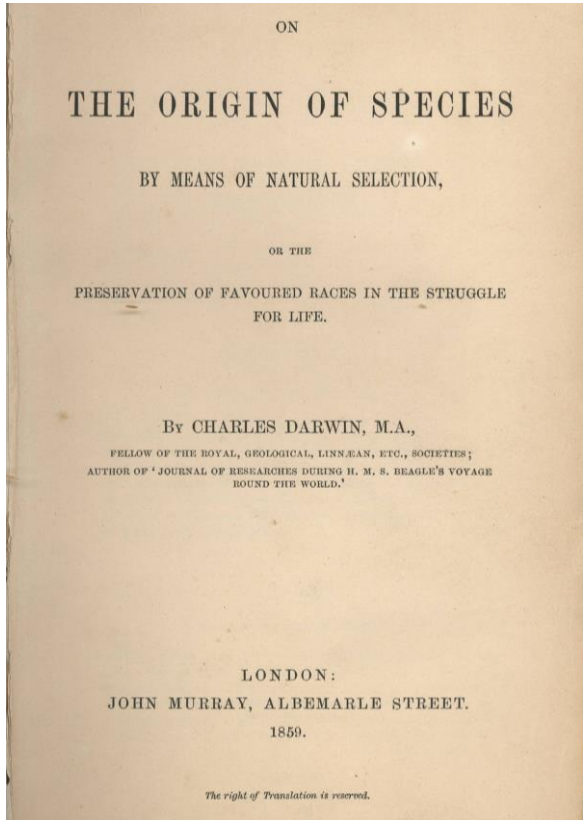
Early Life Exposure & Disease Risk

- Epigenome: biomarkers of exposure and response to the early life environment



Epigenetics

(Window into the past)


ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

By CHARLES DARWIN, M.A.,

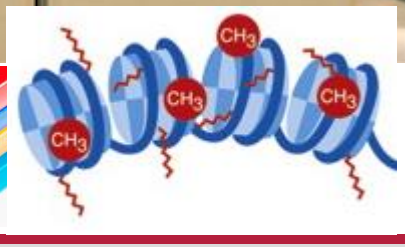
FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;
AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE ROUND THE WORLD.'

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

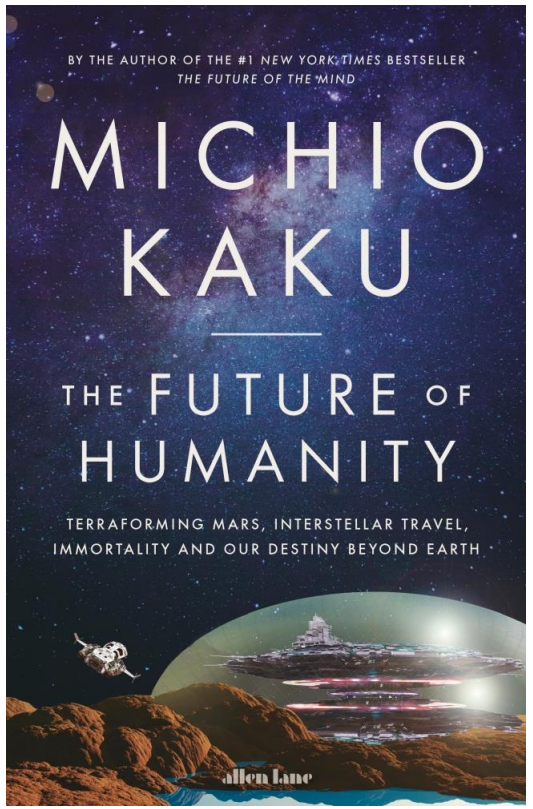
The right of Translation is reserved.

(Biological Memory)

Library



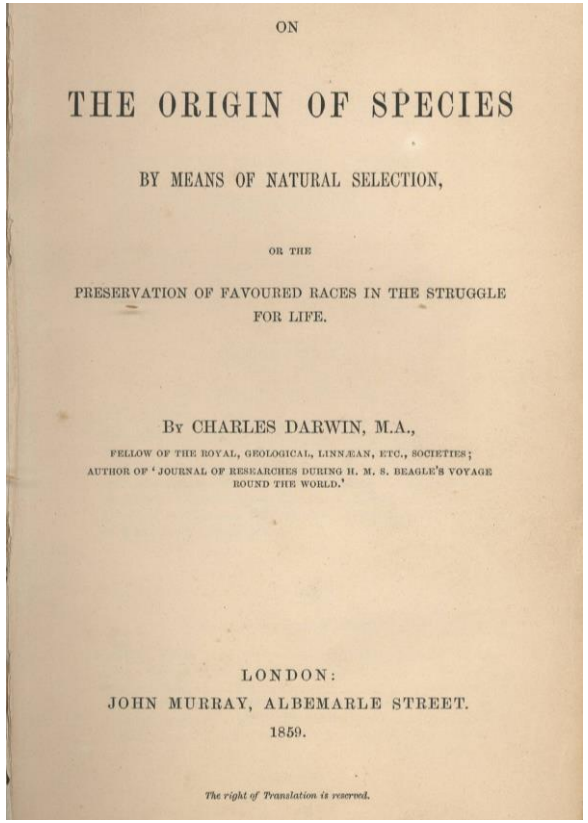
(Snapshot of the future)



(Epigenetic Programming)

Epigenetics

(Window into the past)



ON

THE ORIGIN OF SPECIES

BY MEANS OF NATURAL SELECTION,

OR THE

PRESERVATION OF FAVOURED RACES IN THE STRUGGLE FOR LIFE.

By CHARLES DARWIN, M.A.,

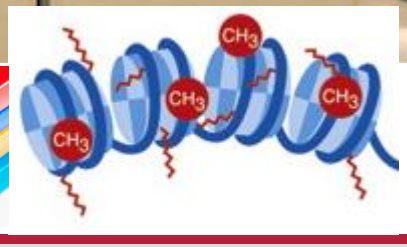
FELLOW OF THE ROYAL, GEOLOGICAL, LINNÆAN, ETC., SOCIETIES;
AUTHOR OF 'JOURNAL OF RESEARCHES DURING H. M. S. BEAGLE'S VOYAGE ROUND THE WORLD.'

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
1859.

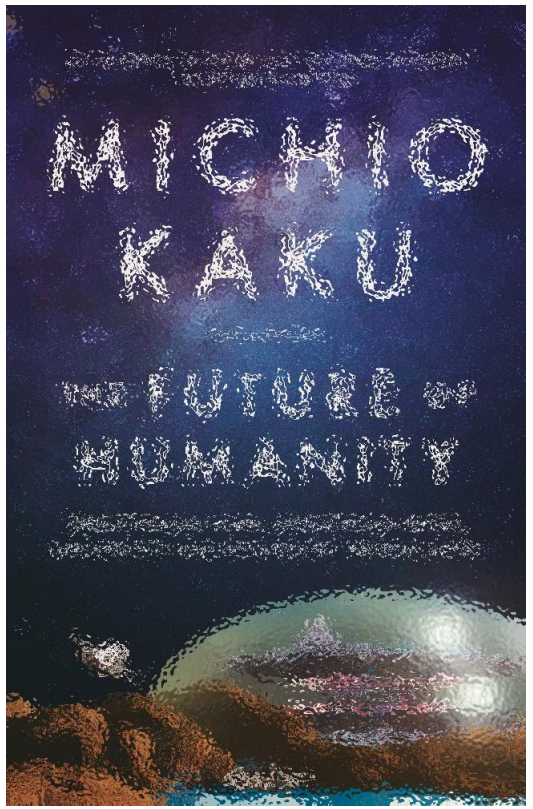
The right of Translation is reserved.

(Biological Memory)

Library



(Snapshot of the future)



(Epigenetic Programming)

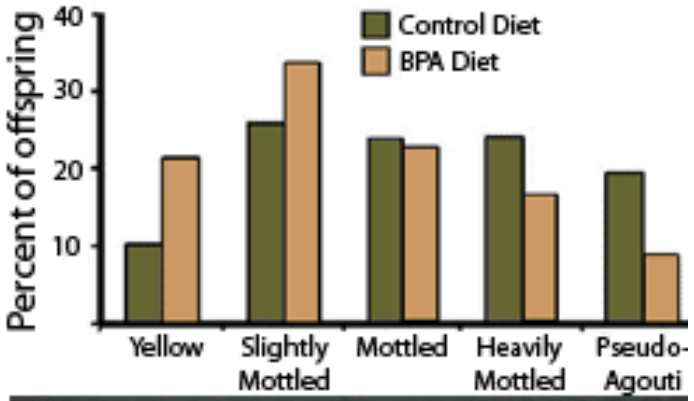
Vocabulary

- **DNA Methylation (DNAm):** A covalent epigenetic modification of the nucleotide cytosine, which is heritable (cell-division)
- **CpG site**: Cytosine followed by a Guanine nucleotide
- **Epigenome:** The epigenetic information in a cell, comprising DNA methylation, post-translational modifications of histones and higher-order chromatin structure
- **EWAS:** Epigenome-Wide Association Study. Usually referring to genome-wide DNAm
- **Genomic imprinting:** Parent-of-origin–specific epigenetic marks generally associated with comparative silencing of the allele transmitted to the offspring

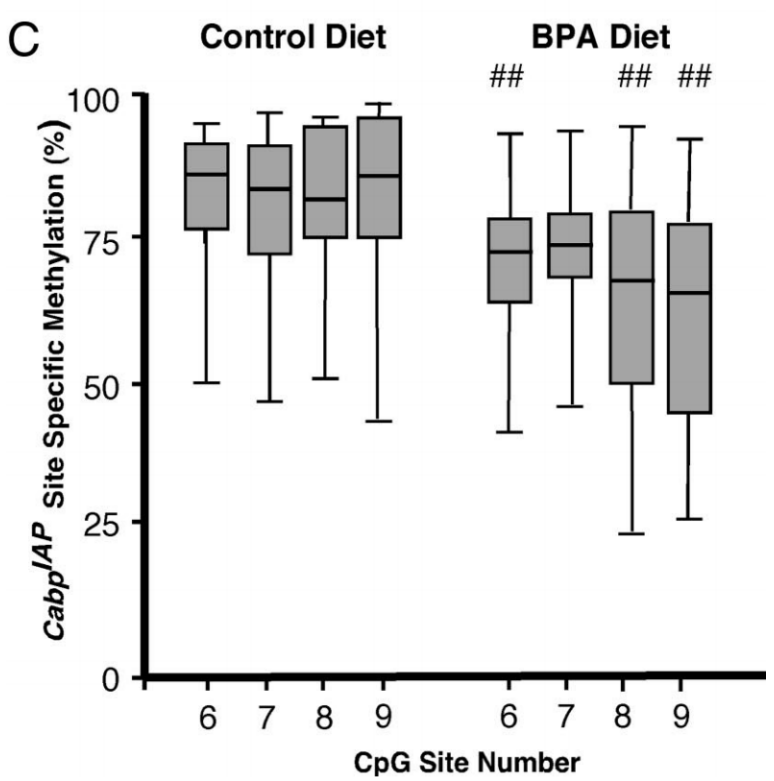
Feinberg, Andrew P. *NEJM* 378.14 (2018): 1323-1334

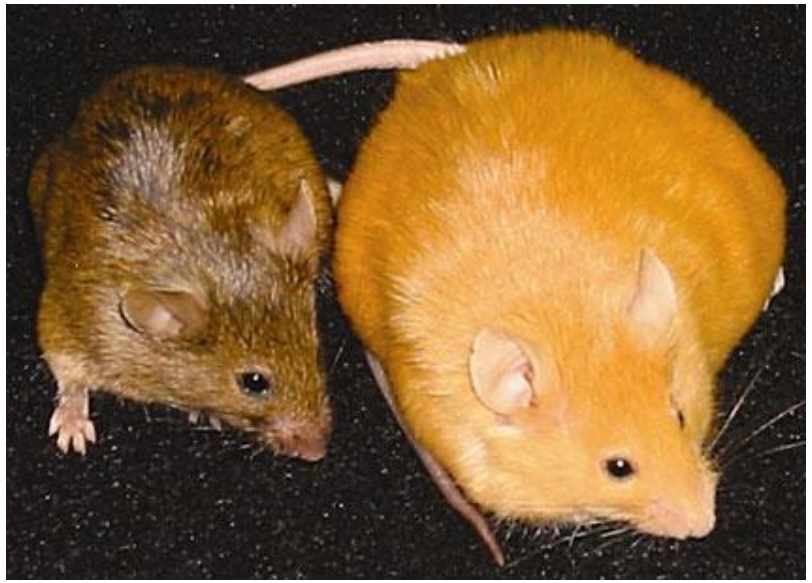
The Agouti Mouse-Model

- Exposure to BPA *in utero* and DNAm of the A^{vy} locus



BPA Exposure





Dolinoy, DC., et al. *PNAS*. 104.32 (2007)

Candidate Gene Approach

Persistent epigenetic differences associated with prenatal exposure to famine in humans

Bastiaan T. Heijmans[a,1,2], Elmar W. Tobi[a,2], Aryeh D. Stein[b], Hein Putter[c], Gerard J. Blauw[d], Ezra S. Susser[e,f], P. Eline Slagboom[a], and L. H. Lumey[e,1]

Departments of [a]Molecular Epidemiology, [c]Medical Statistics, and [d]Gerontology and Geriatrics, Leiden University Medical Center, Leiden, The Netherlands; [b]Hubert Department of Global Health, Rollins School of Public Health, Emory University Atlanta, GA 30322; [e]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY 10032; and [f]New York State Psychiatric Institute, New York, NY 10032

Edited by Charles R. Cantor, Sequenom Inc., San Diego, CA, and approved September 17, 2008 (received for review July 7, 2008)

- *IGF2* (insulin-like growth factor II):
  - Maternally imprinted
  - Similar structure to insulin & promotes growth during gestation
  - Highly active during fetal development
- Dutch Hunger Winter (German-occupied Netherlands):
  - Severe caloric restriction during gestation
  - Exposed individuals compared to same-sex siblings (unexposed)

Heijmans, BT., et al. *PNAS* 105.44 (2008): 17046-17049

Candidate Gene Approach

- Prenatal famine exposure associated with lower DNAm of *IGF2*

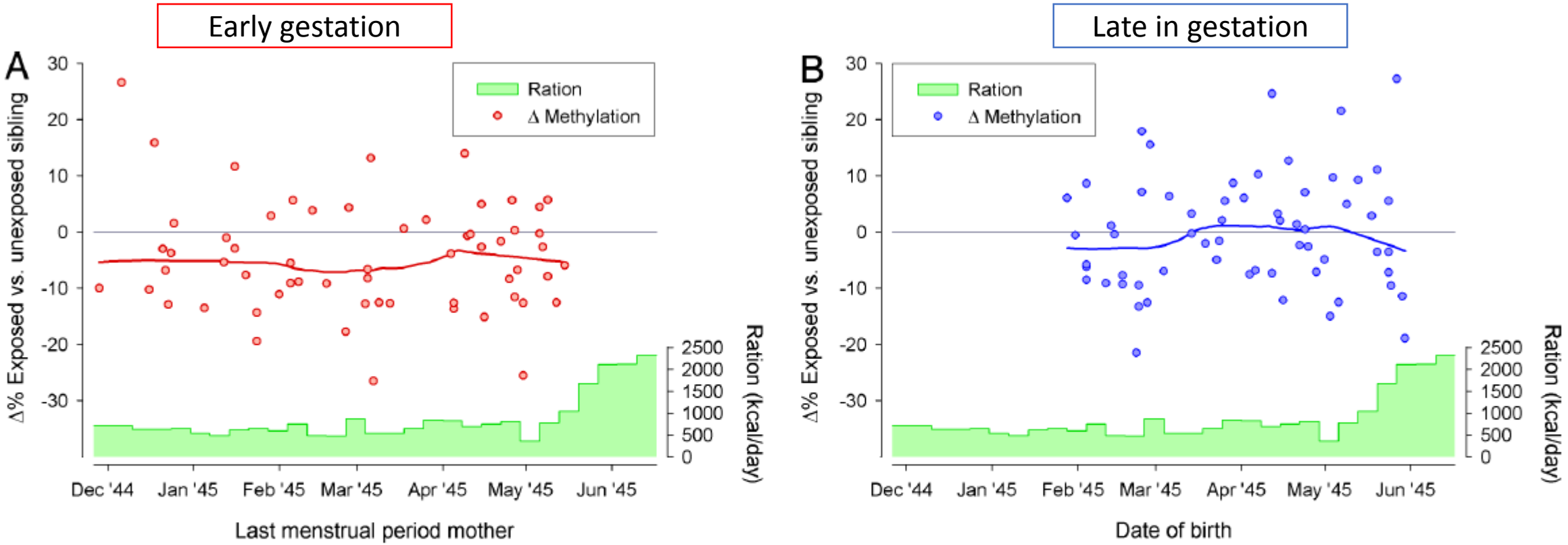
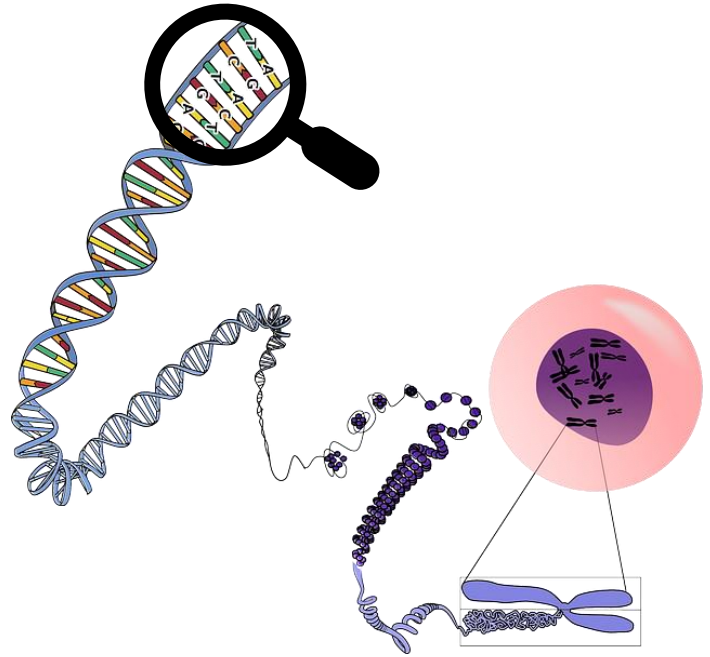


Fig. 1. Difference in *IGF2* DMR methylation between individuals prenatally exposed to famine and their same-sex sibling. (*A*) Periconceptional exposure: Difference in methylation according to the mother's last menstrual period (a common estimate of conception) before conception of the famine-exposed individual. (*B*) Exposure late in gestation: Difference in methylation according to the date of birth of the famine-exposed individual. To describe the difference in methylation according to estimated conception and birth dates, a lowess curve (red or blue) is drawn. The average distributed rations (in kcal/day) between December 1944 and June 1945 are depicted in green.

Heijmans, BT., et al. *PNAS* 105.44 (2008): 17046-17049

DNAm: Human Studies

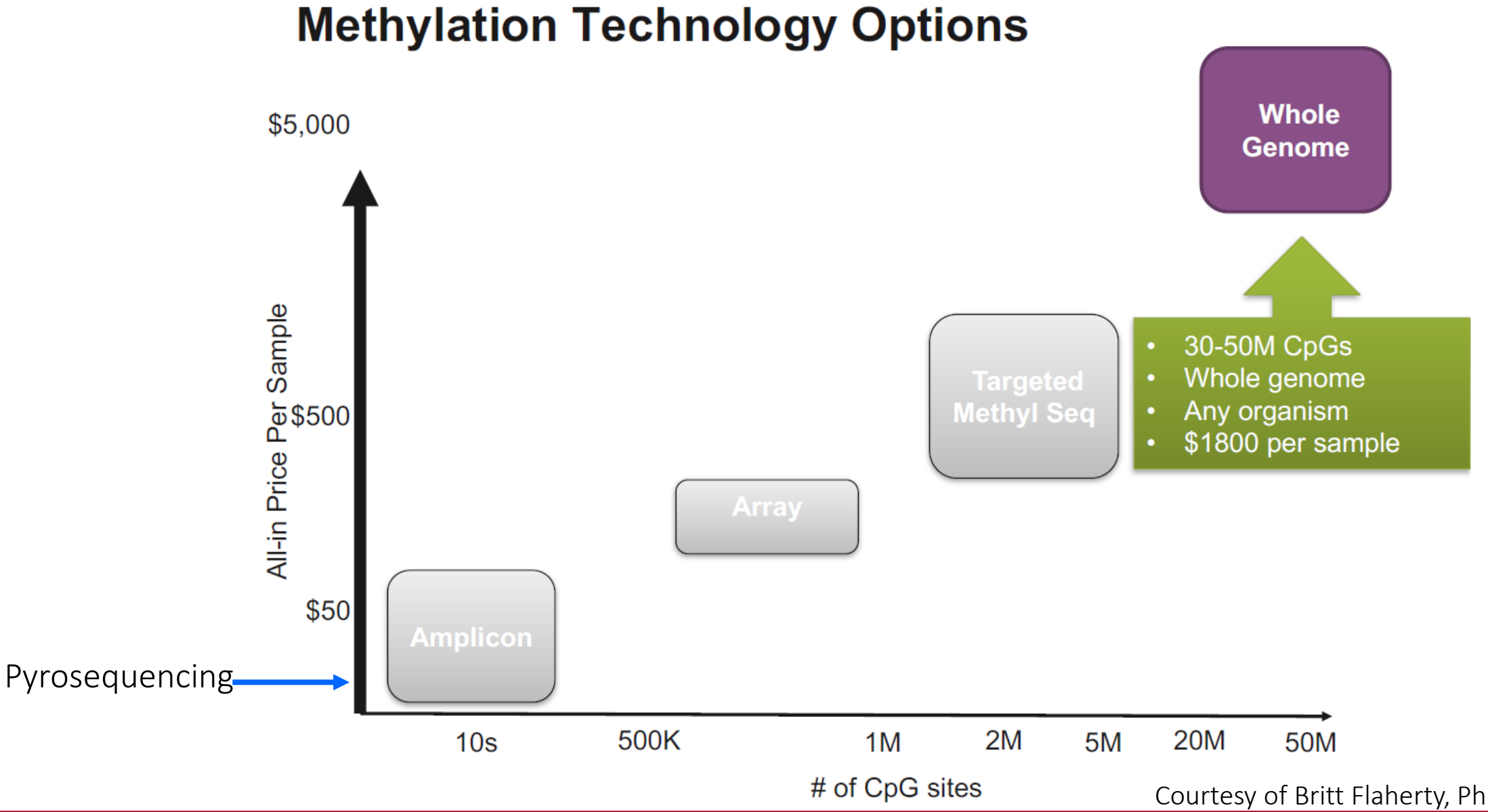
- Look at candidate genes
  - ~20,000 genes x 100 - 1,000s of CpGs



- High density microarrays (agnostic approach)
  - Covers 99% of the RefSeq. genes
  - Screens >850,000 CpG sites (EPIC) / 450K CpGs (450K)
  - Single nucleotide resolution

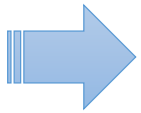
- Hypothesis-free Epigenome-Wide Association studies (EWAS)
  - High dimensional genomic data
  - Parameters >> N

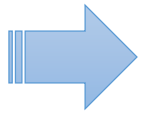
Costs/Sample size

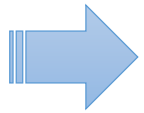


Courtesy of Britt Flaherty, PhD (Illumina, 2017)

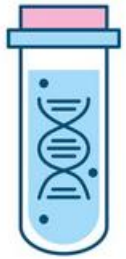
Study Design Considerations

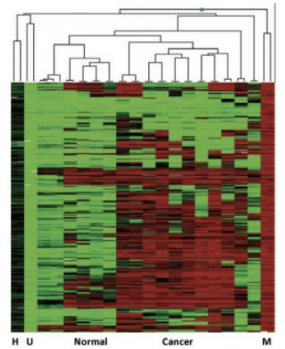






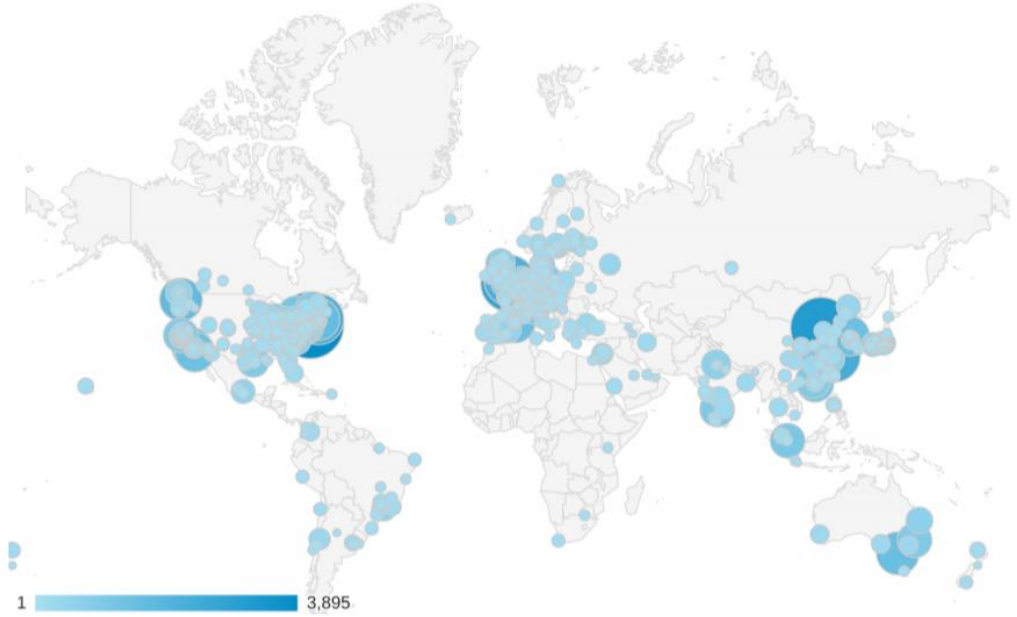
- **Periconception period**
  - Questioners
  - Bio-specimens
- **Birth**
  - Questioners
  - DNA isolation (placenta/CB)
- **Processing**
  - Storage
  - Analyses
- **Analyses**
  - Hypothesis testing
  - EWAS





Processing: Bioconductor

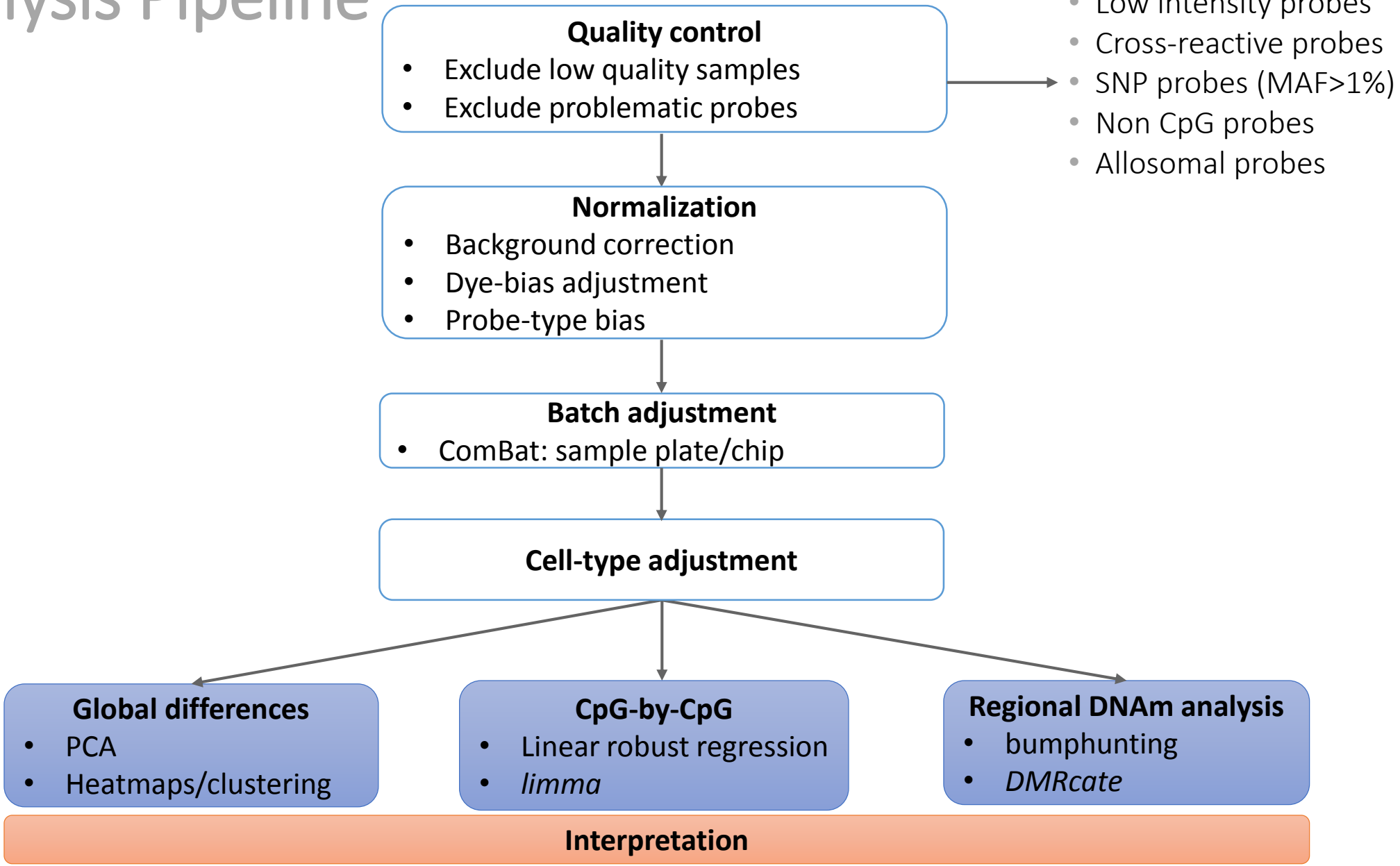
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS



- Open source, open development software project
- Based primarily on the *R* programming language
- Widespread access to a broad range of omics tools
- Complete workflows for epigenomic data
- Enables high-quality documentation and reproducible research

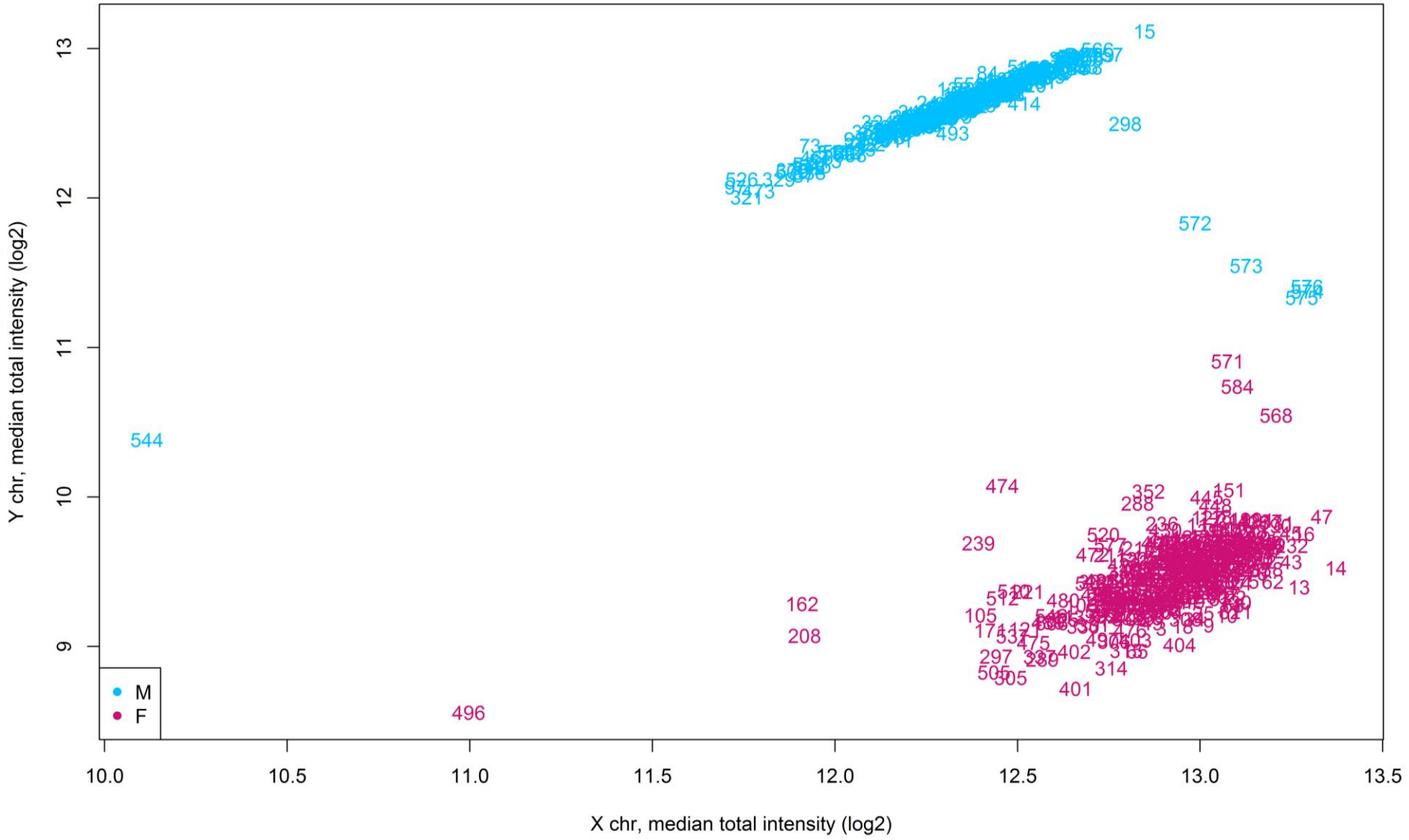
https://www.bioconductor.org/

Analysis Pipeline



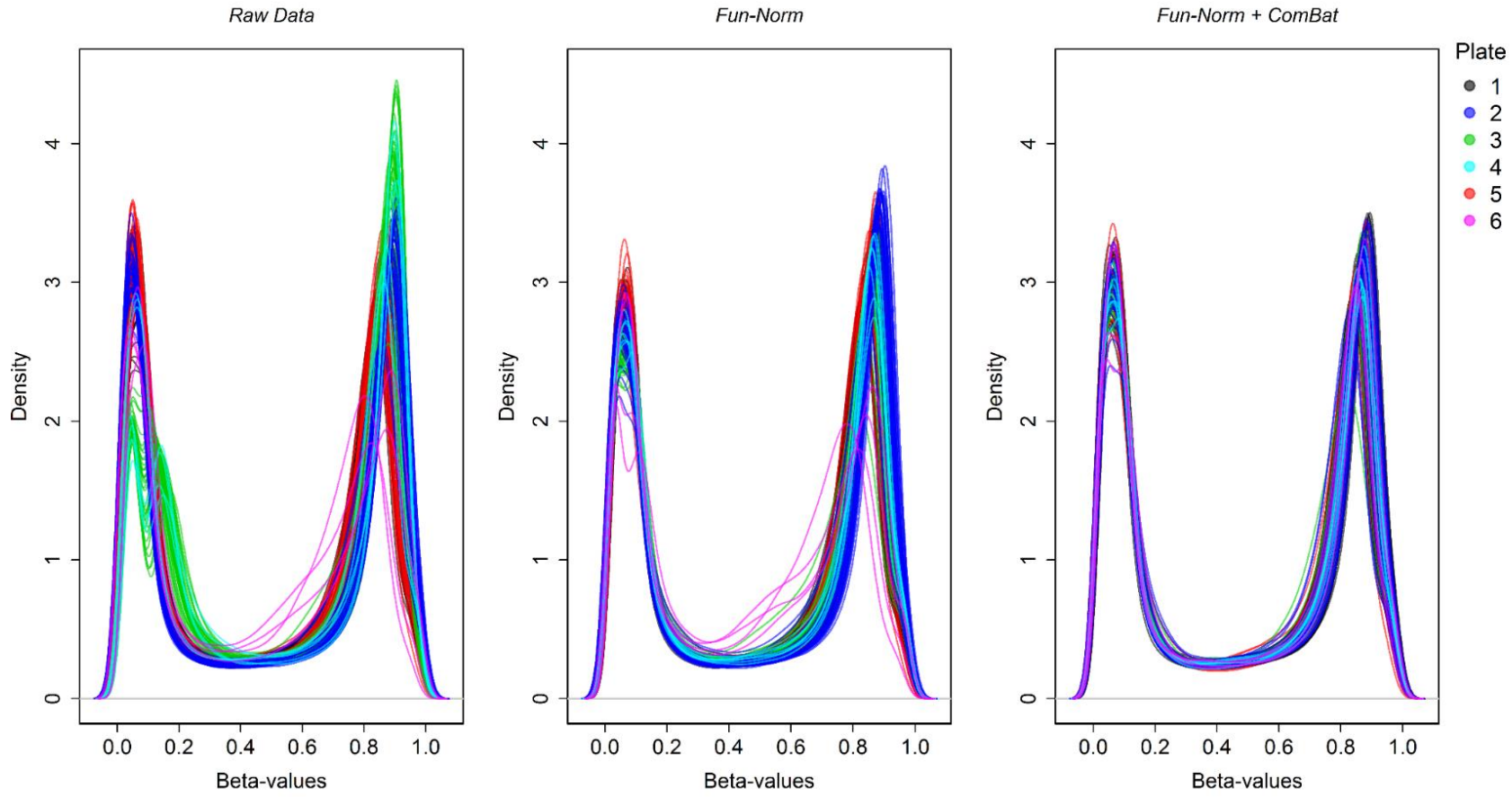
Quality Control of Samples

- Sex-prediction



Normalization

- Removing unwanted variation in microarray data (technical variation)
  - Background correction (between-array variation)
  - Dye-bias adjustment
  - Batch effects



Plate

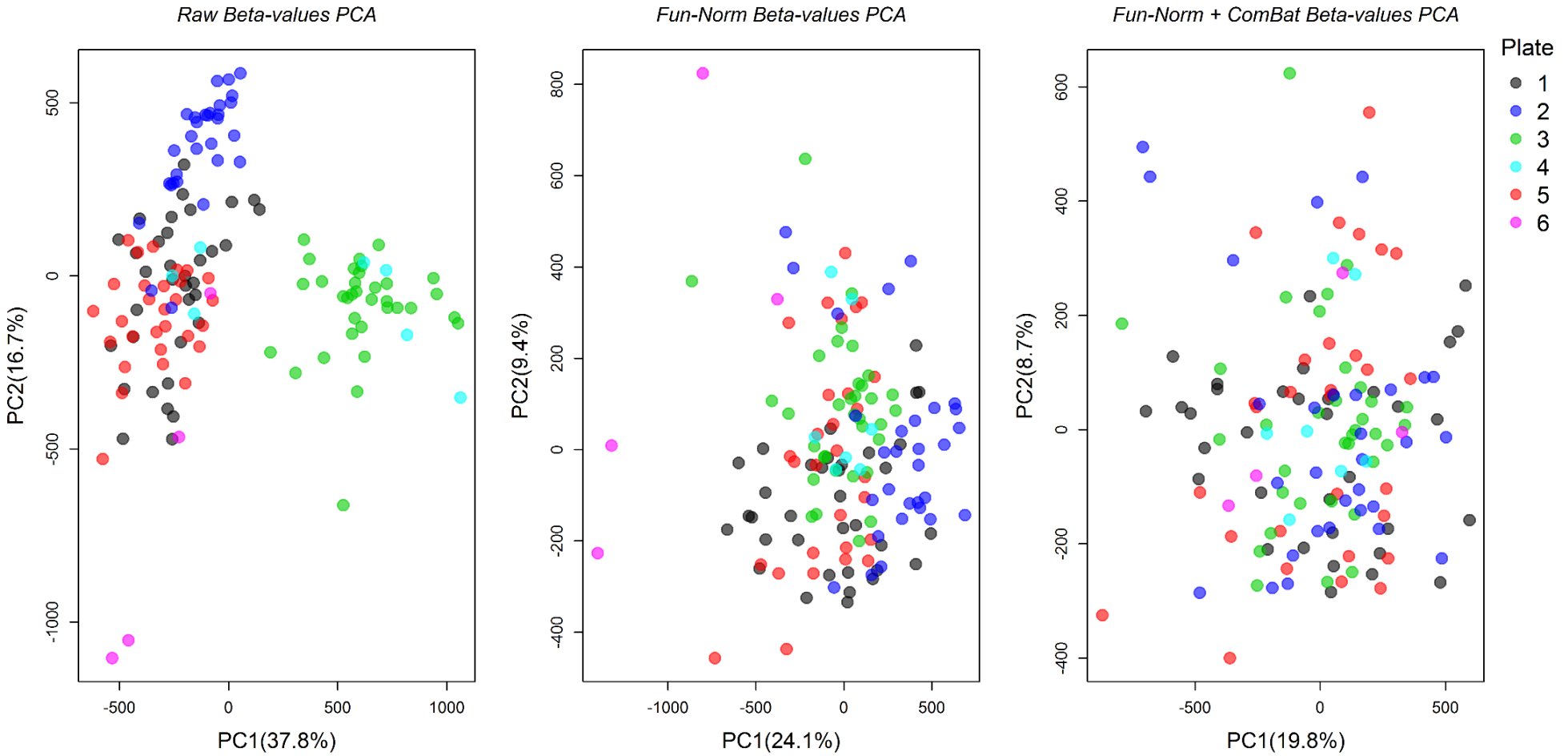
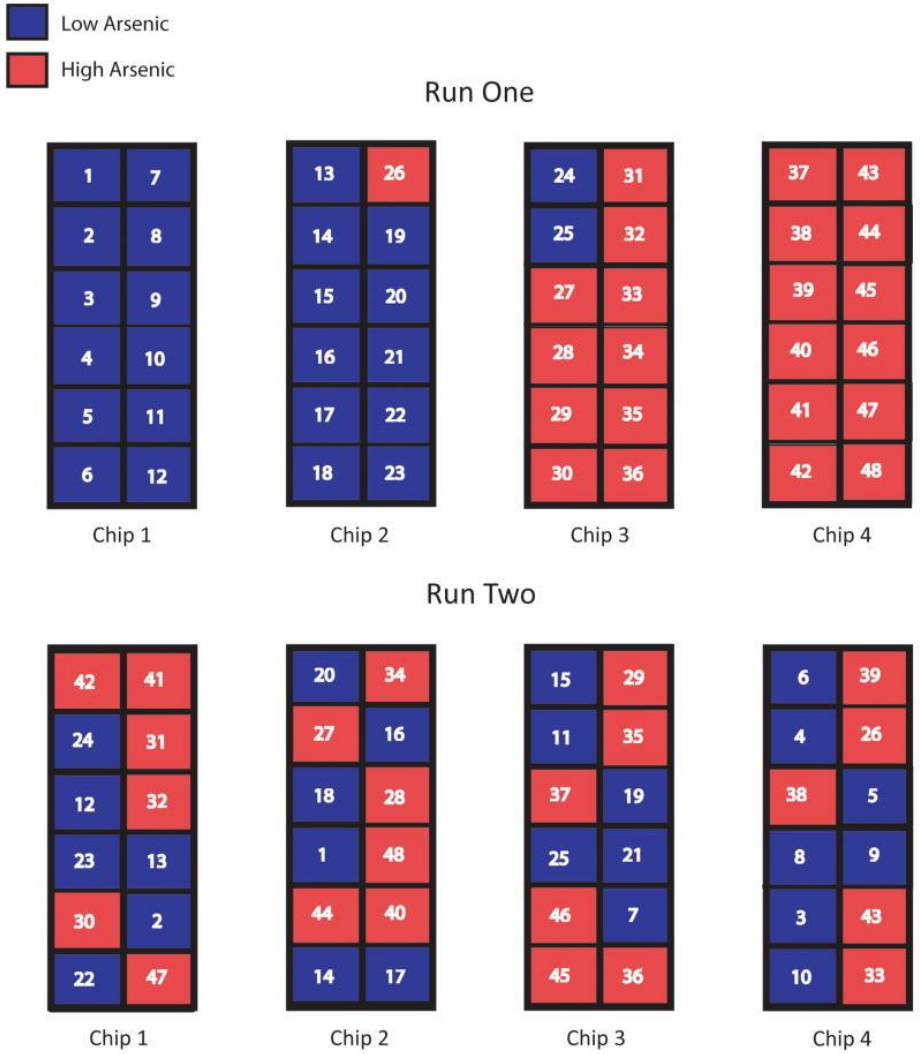


Plate Associations with Top three PCs						
	Raw PCA		Fun-Norm PCA		Fun-Norm + ComBat Adjusted	
Principal Component	Variance explained	P-value	Variance explained	P-value	Variance explained	P-value
PC_1	37.8%	<0.001	24.1%	<0.001	19.8%	0.99
PC_2	16.7%	<0.001	9.4%	0.001	8.7%	0.95
PC_3	6.2%	<0.001	6%	<0.001	4.5%	0.99

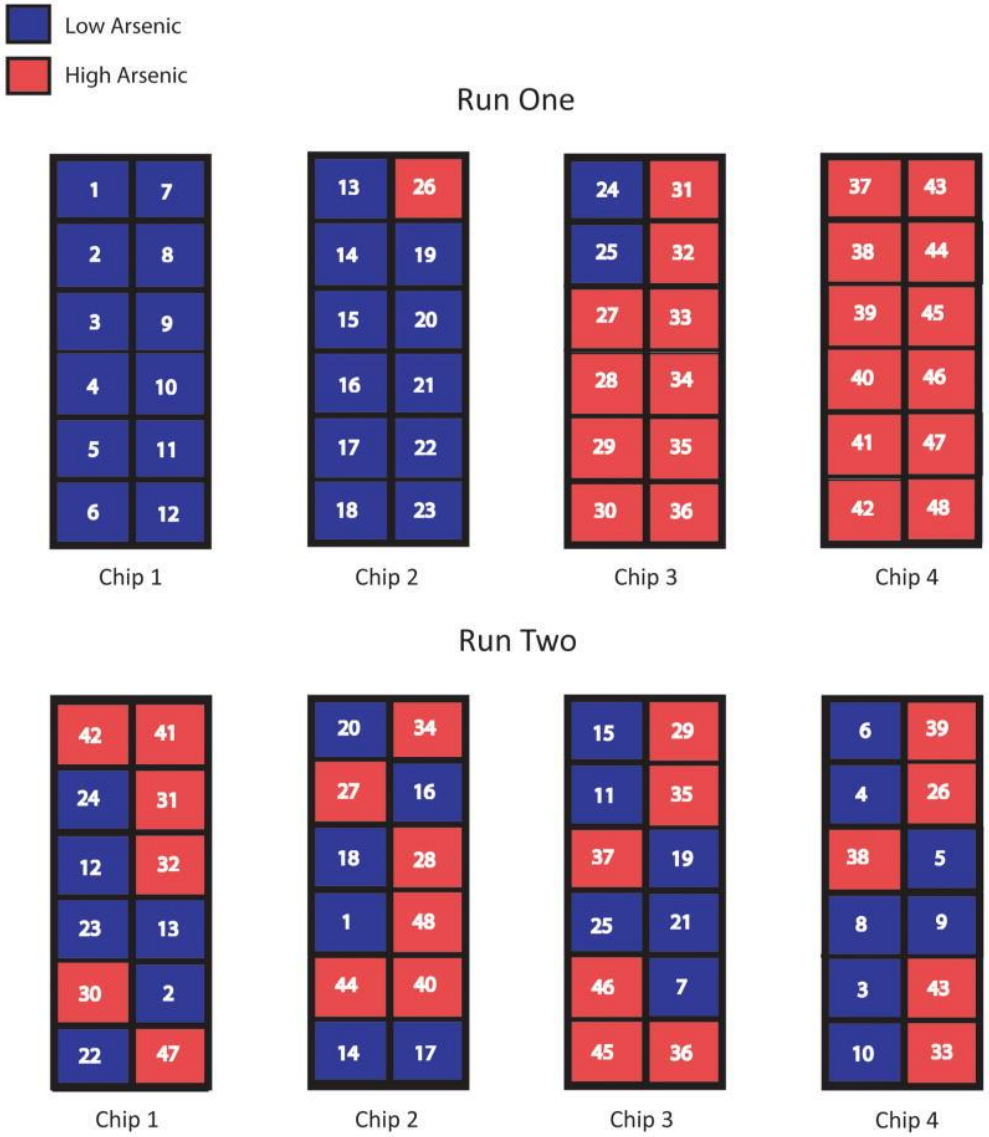
Batch effects

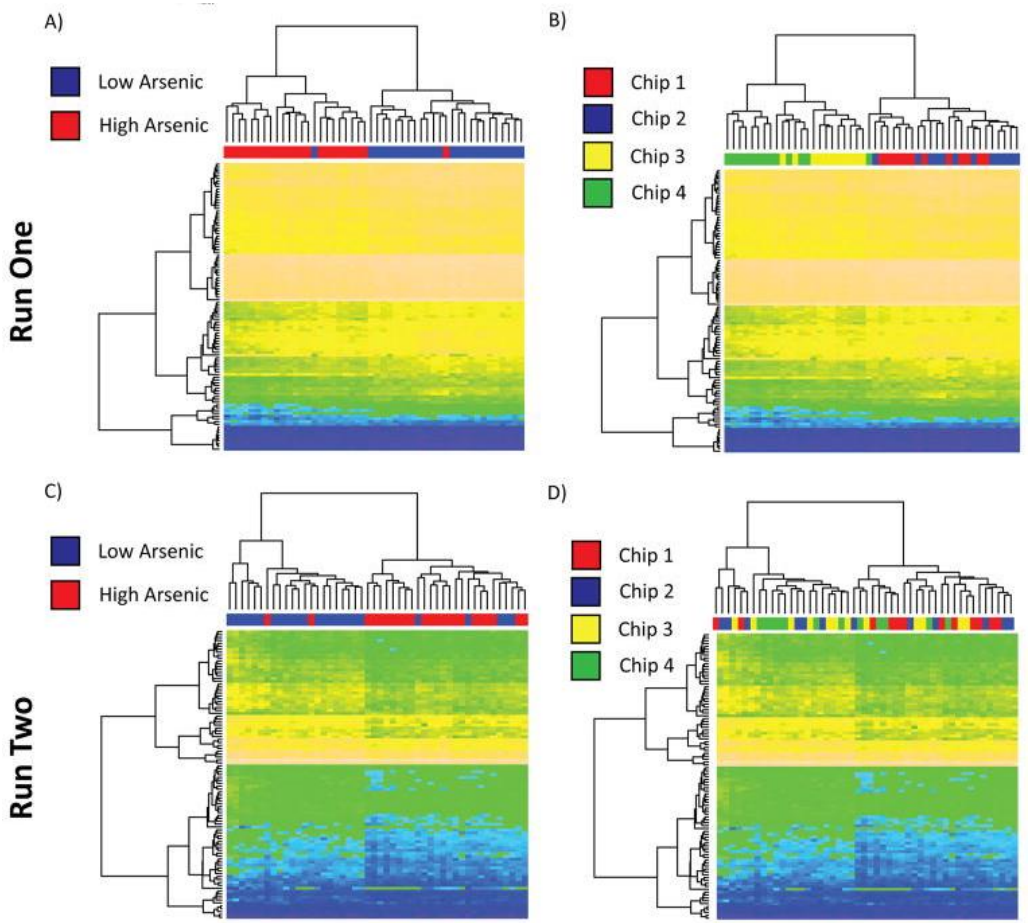
- Batch effects can completely ruin experiments!
- How are batch effects generated?



Harper, KN et al. *Cancer Epidemiol Biomarkers Prev. (2013) 22.6 : 1052-1060.*

Batch effects: a cautionary tale

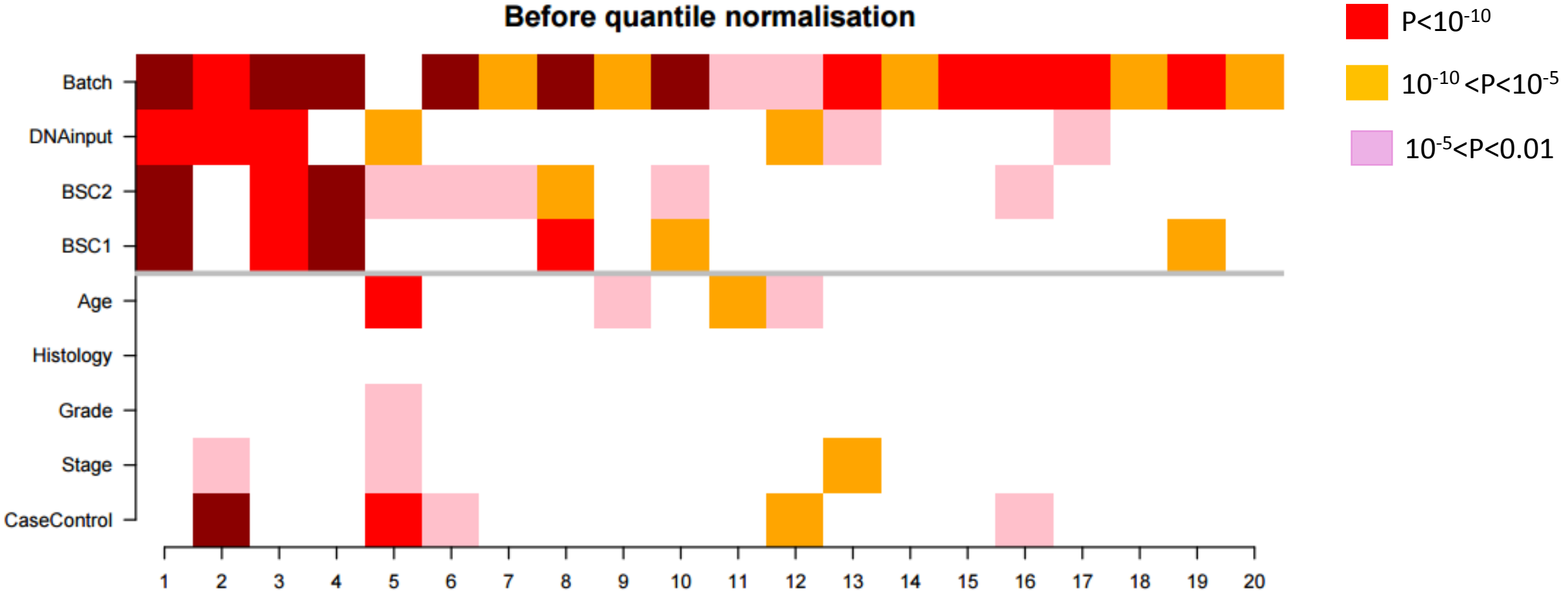




Significant CpGs	Run One (Non-Randomized)	Run Two (Randomized)
No chip adjustment	1,203	0
Chip adjustment	24,184	187
Overlap	25	

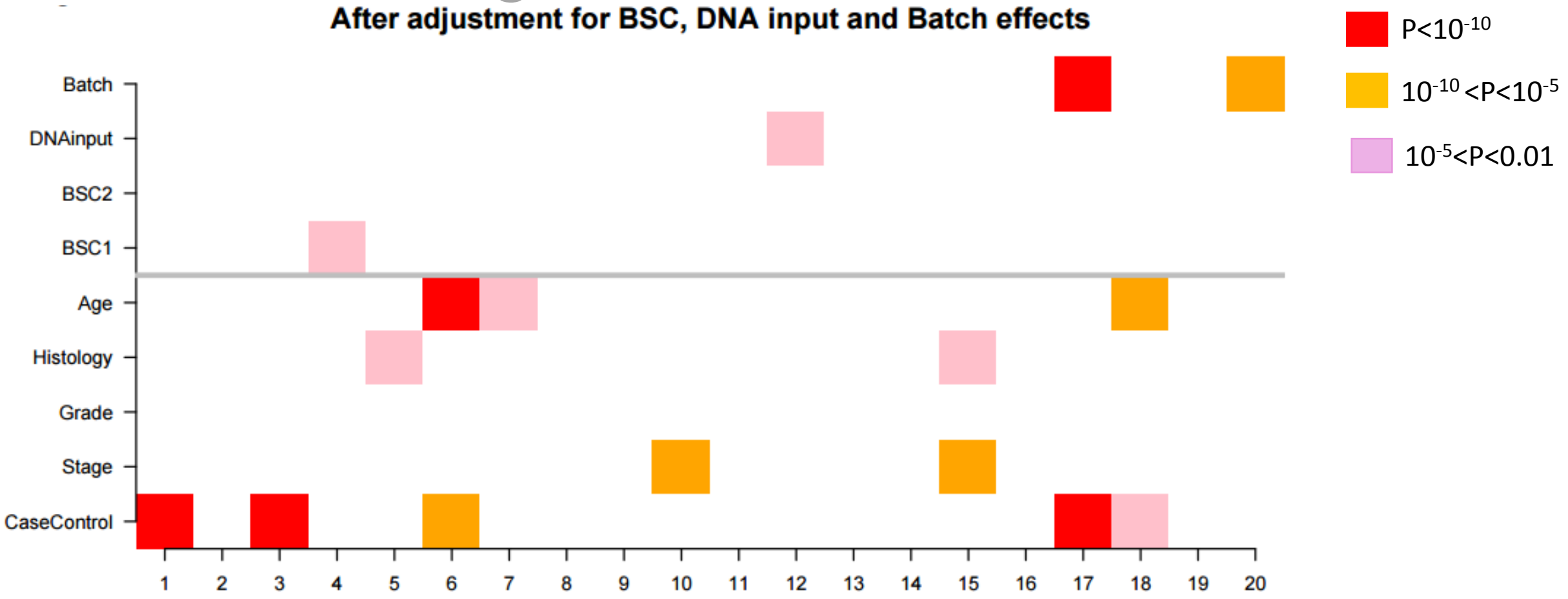
Harper, KN et al. *Cancer Epidemiol Biomarkers Prev. (2013) 22.6 : 1052-1060.*

PCA: a tool for high dimensional data



Teschendorff et al. *PLoS One* (2009); e8274.

PCA: a tool for high dimensional data



- If known batch effects can be adjusted
- Sample plate is a good surrogate batch variable
- Randomization of samples is key!

Teschendorff et al. *PLoS One* (2009); e8274.

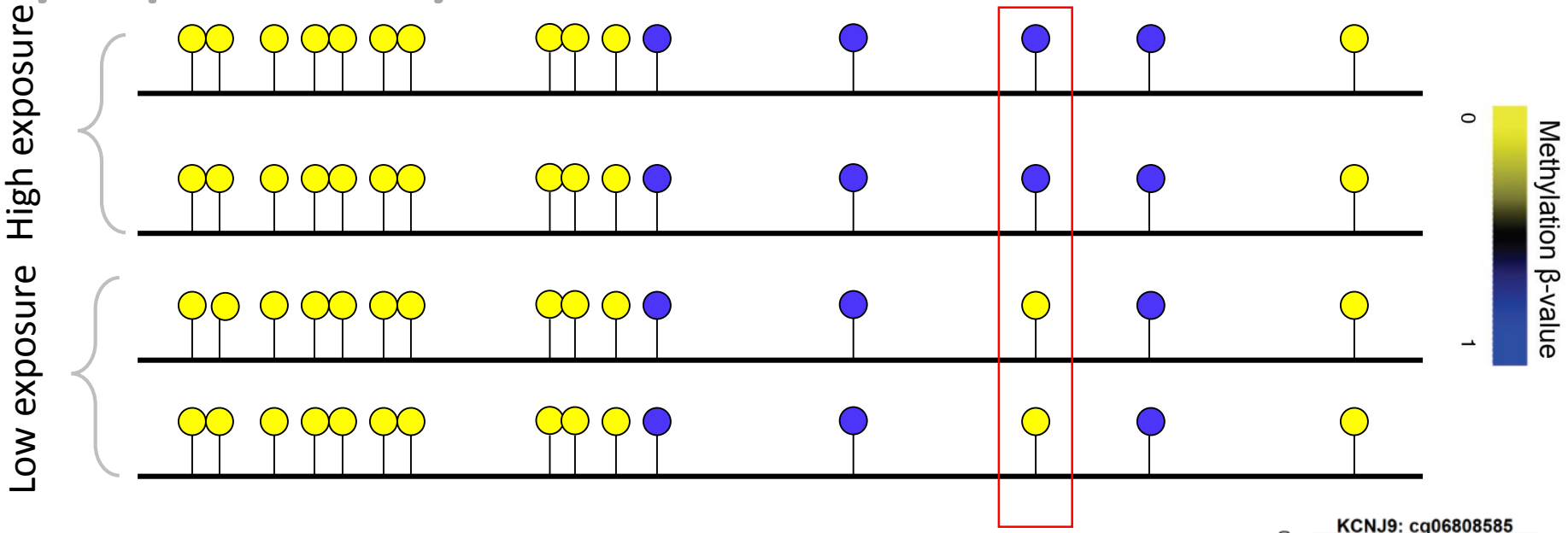
EWAS: CpG-by-CpG Analysis

- "Backbone" of all DNAm analyses
- Fit adjusted linear model for each CpG (450-850K models)
- Estimate coefficients and *P*-values (decide on significance *a priori*)
- Several methods/models
  - OLS
  - Linear robust regression
  - *Limma*
  - t-test
- Model methylation on native β-scale or M-value
  - β-values (0-1)
  - Logit transformation of β-values -> M-values

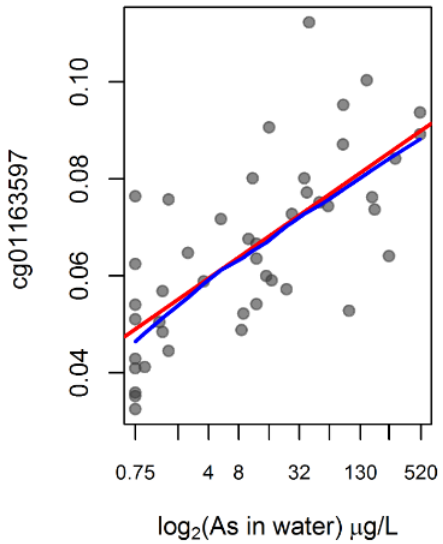
$$M_i = \log_2\left(\frac{\max(y_{i,methy},0)+\alpha}{\max(y_{i,unmethy},0)+\alpha}\right)$$

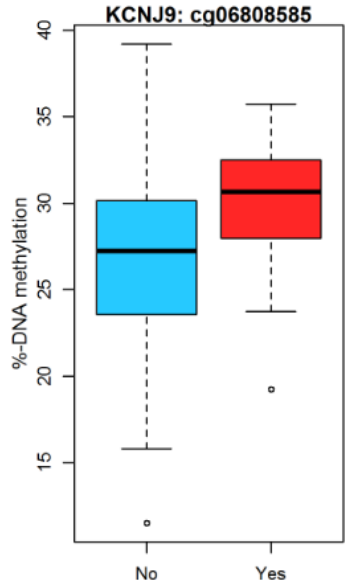
$$Beta_i = \frac{2^{M_i}}{2^{M_i}+1}; M_i = \log_2\left(\frac{Beta_i}{1-Beta_i}\right)$$

CpG-by-CpG Analysis



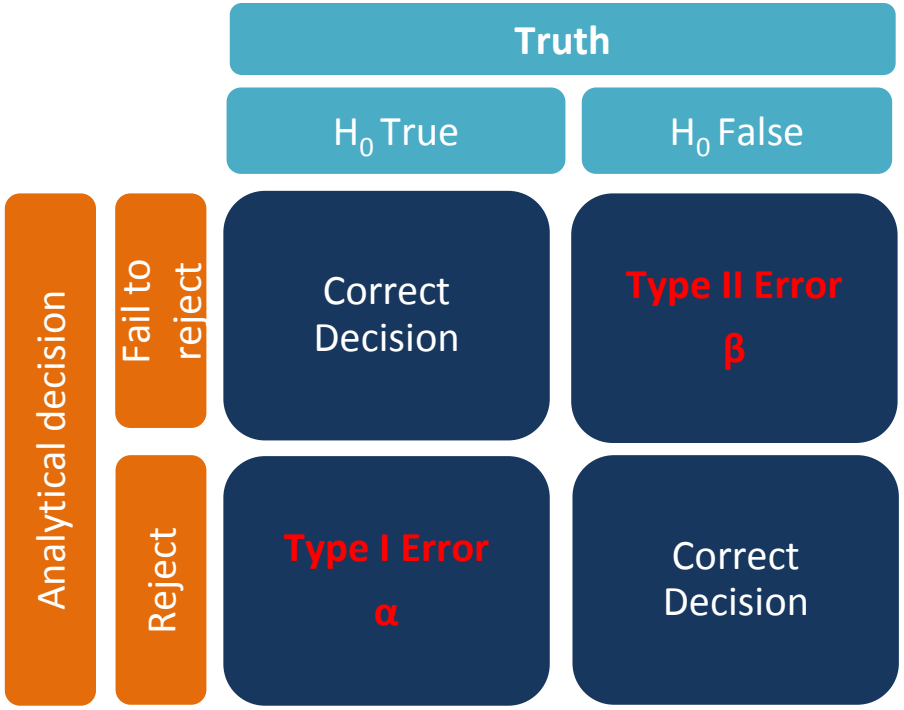
- Fitting a statistical model
  - Assumptions still apply
  - Confounding
  - Mediators/colliders etc.
  - Think about the model carefully





Multiple testing burden (omics)

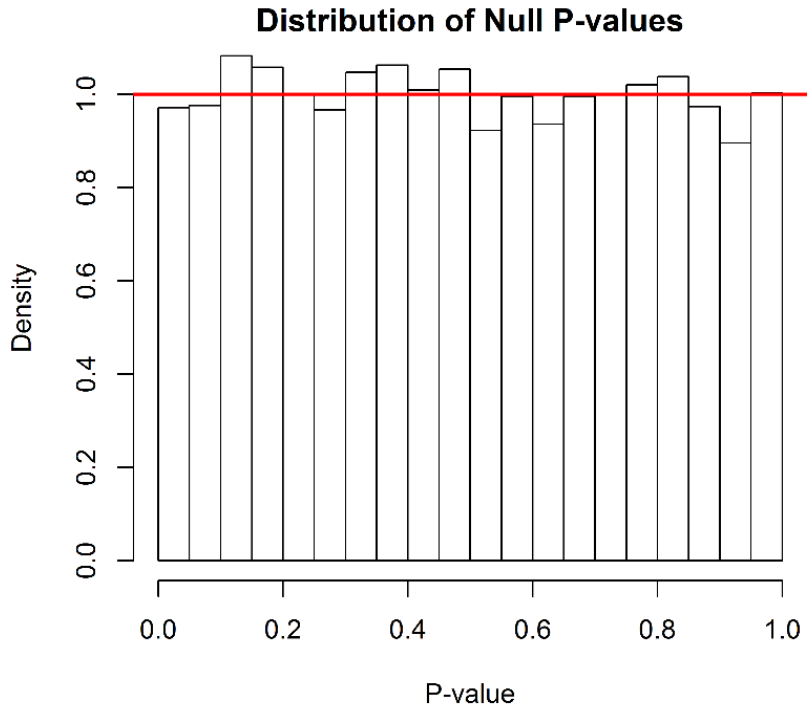
- Experimental design testing many features (CpGs, SNPs, miRNA, etc.)
- Example: Illumina's EPIC chip (~850K tests)

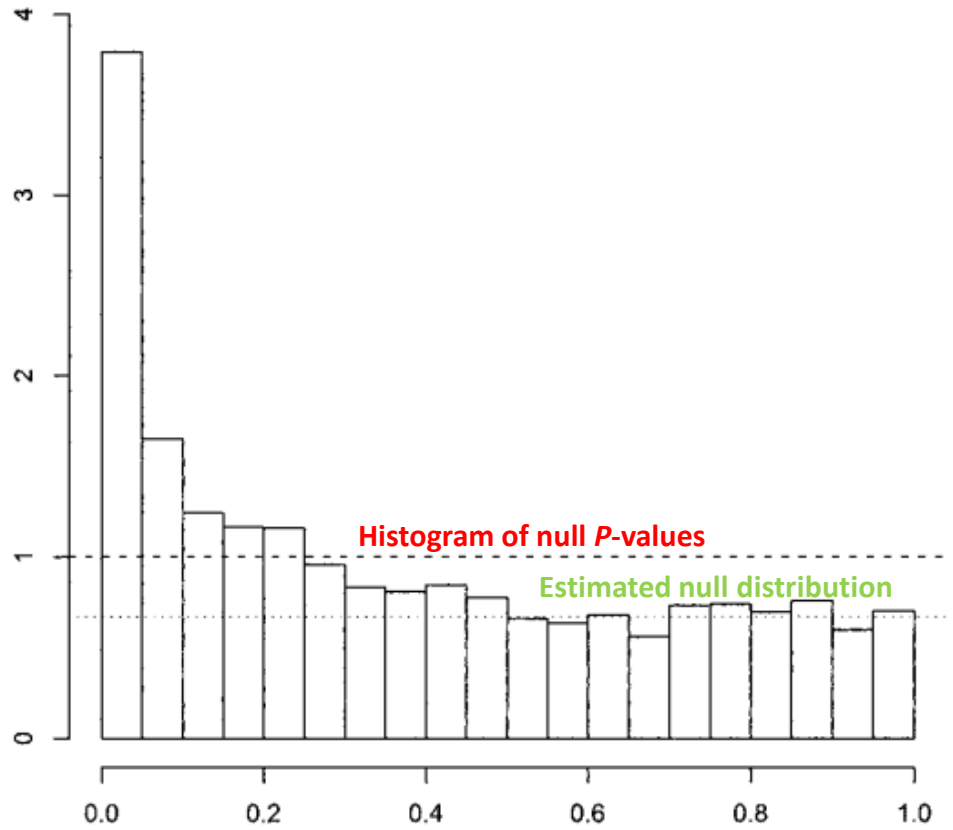


#-tests	Probability Type I error	False Discoveries (Expected)
1	5%	0.05
2	10%	0.10
3	14%	0.15
4	19%	0.20
5	23%	0.25
….	….	….
450K	~100%	22,500
850K	~100%	42,500

EWAS: multiple testing burden (omics)

- Bonferroni correction (α/850,000): *P*<5.8x10-8
- Bonferroni correction: increases type II errors
- We can control the FDR at 0.05
- Controlling the FDR: q-value

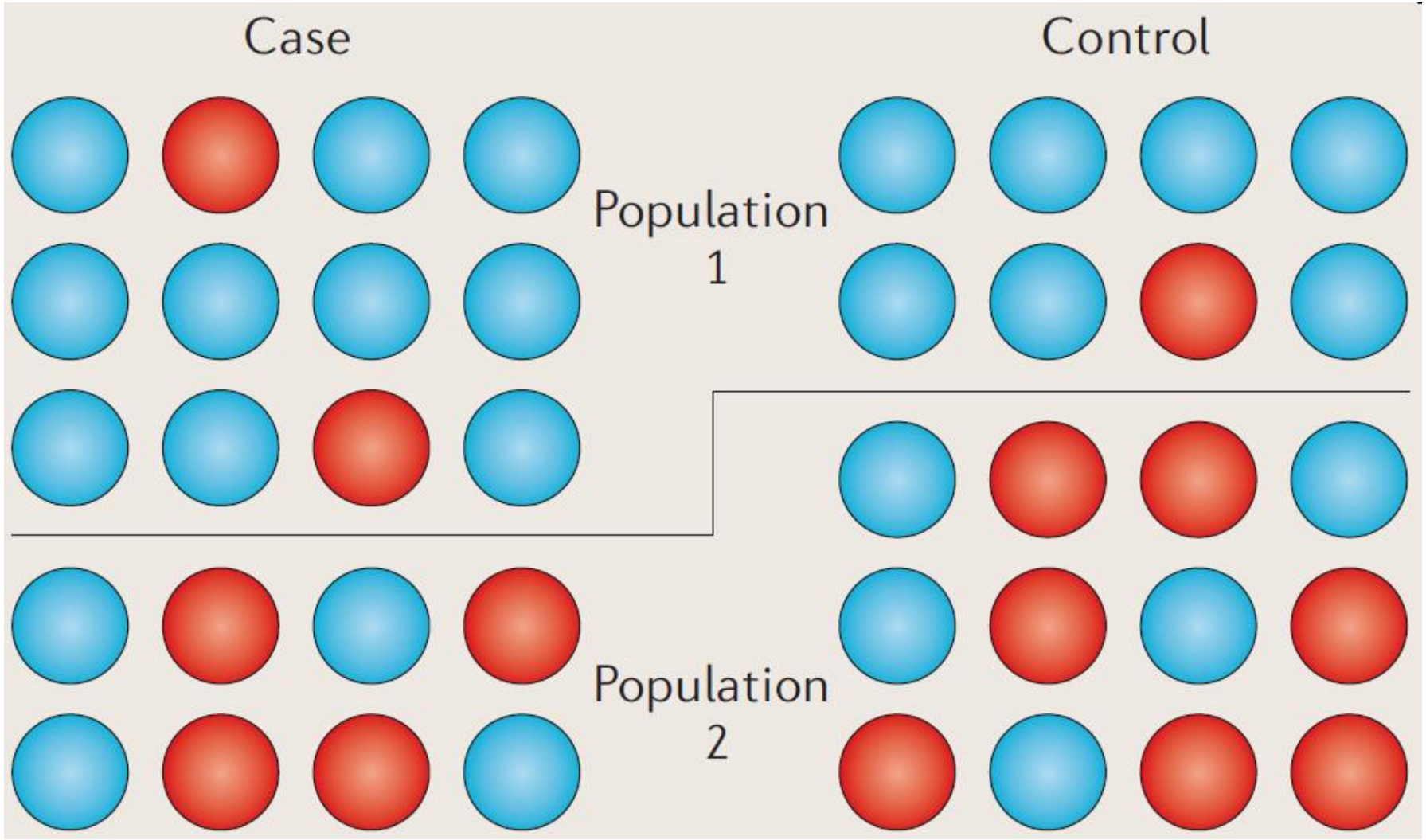




Storey & Tibshirani *PNAS* 100.16; 440-445 (2003)

Population Stratification

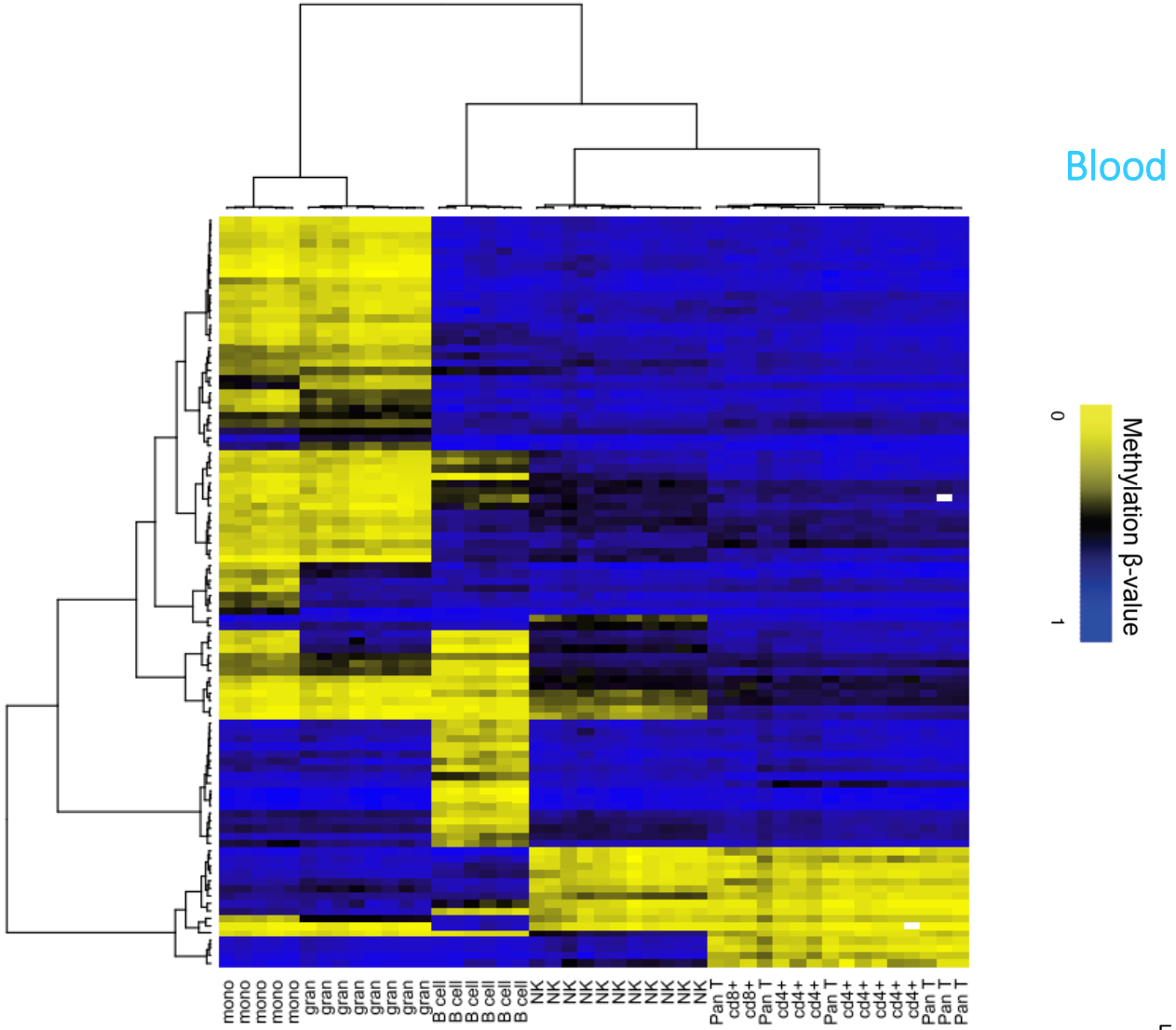
- Spurious association of cases with the blue SNP feature (similar with CpGs)



Balding DJ, *Nature Reviews Genetics* (2006): 7.10 781.

DNA methylation – cellular heterogeneity

- Cell-types have a “unique” DNA methylation fingerprint



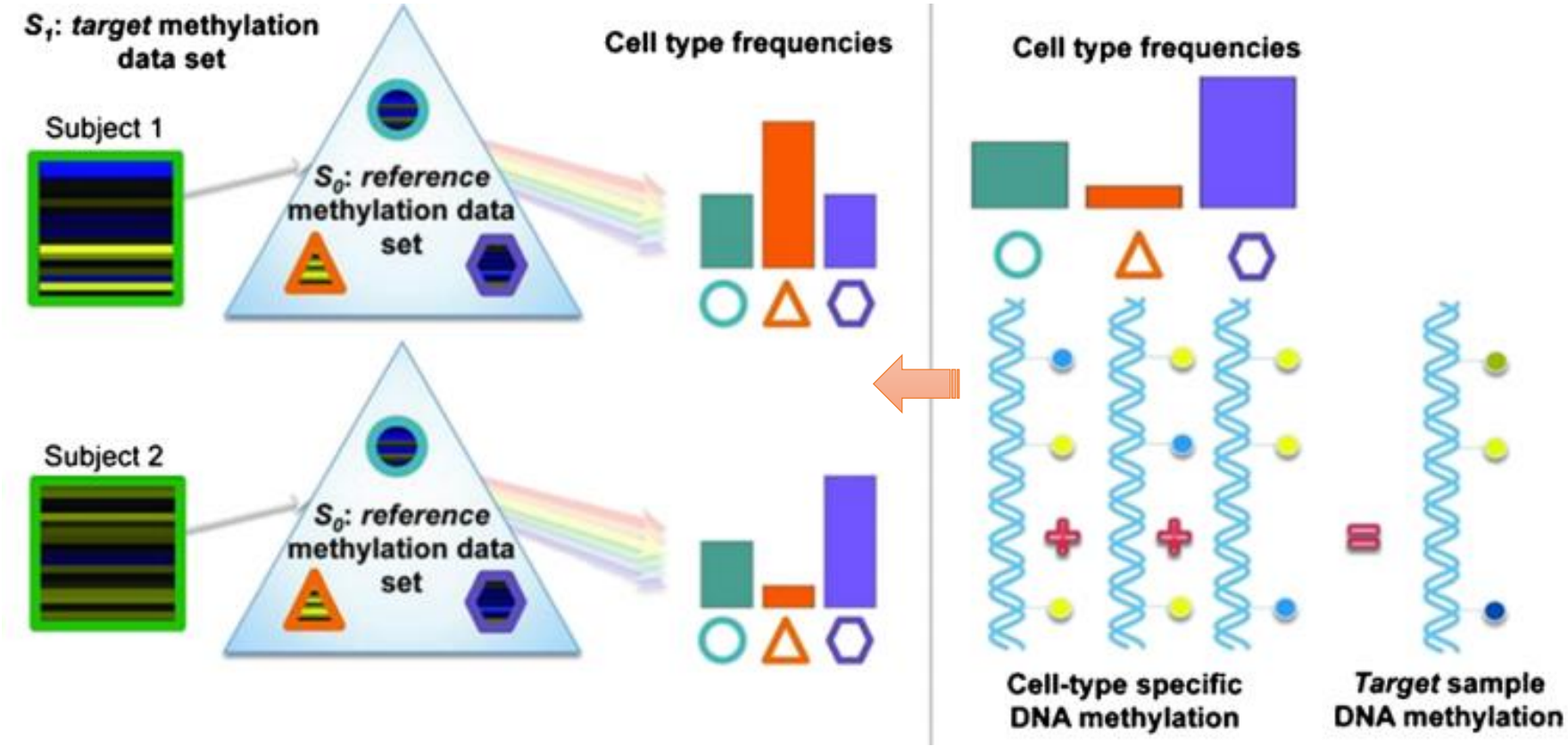
Blood commonly used in Epidemiological studies!

E.A. Houseman et al. BMC *Bioinformatics* (2012) 86; 13.1

Population stratification

- Adjustment for blood cell-type composition is possible

- Adjustment reduces confounding and improves genomic inflation

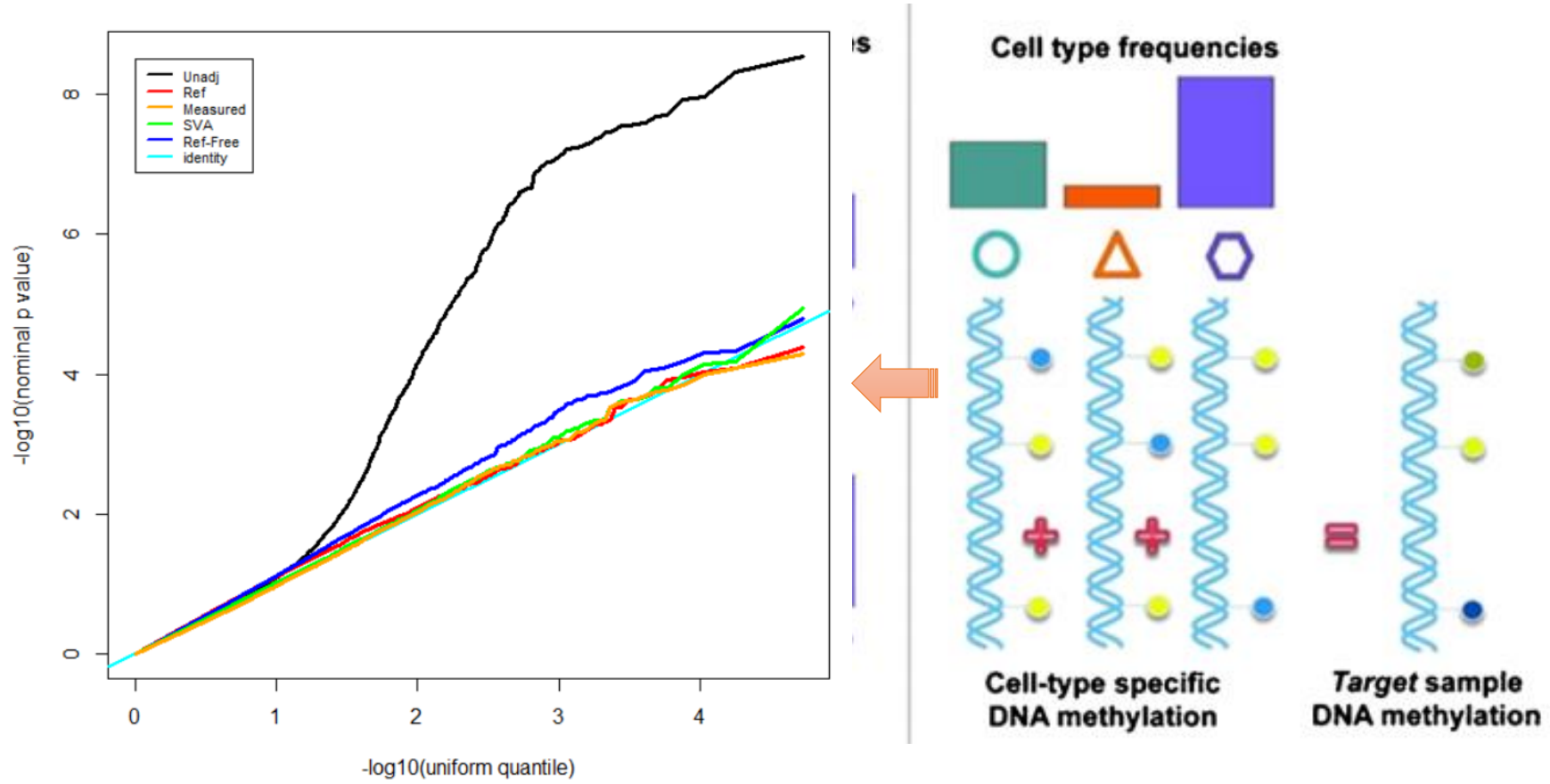


E.A. Houseman et al. *Curr Env Health Rept* (2015); 2.2 145-154

Population stratification

- Adjustment for blood cell-type composition is possible

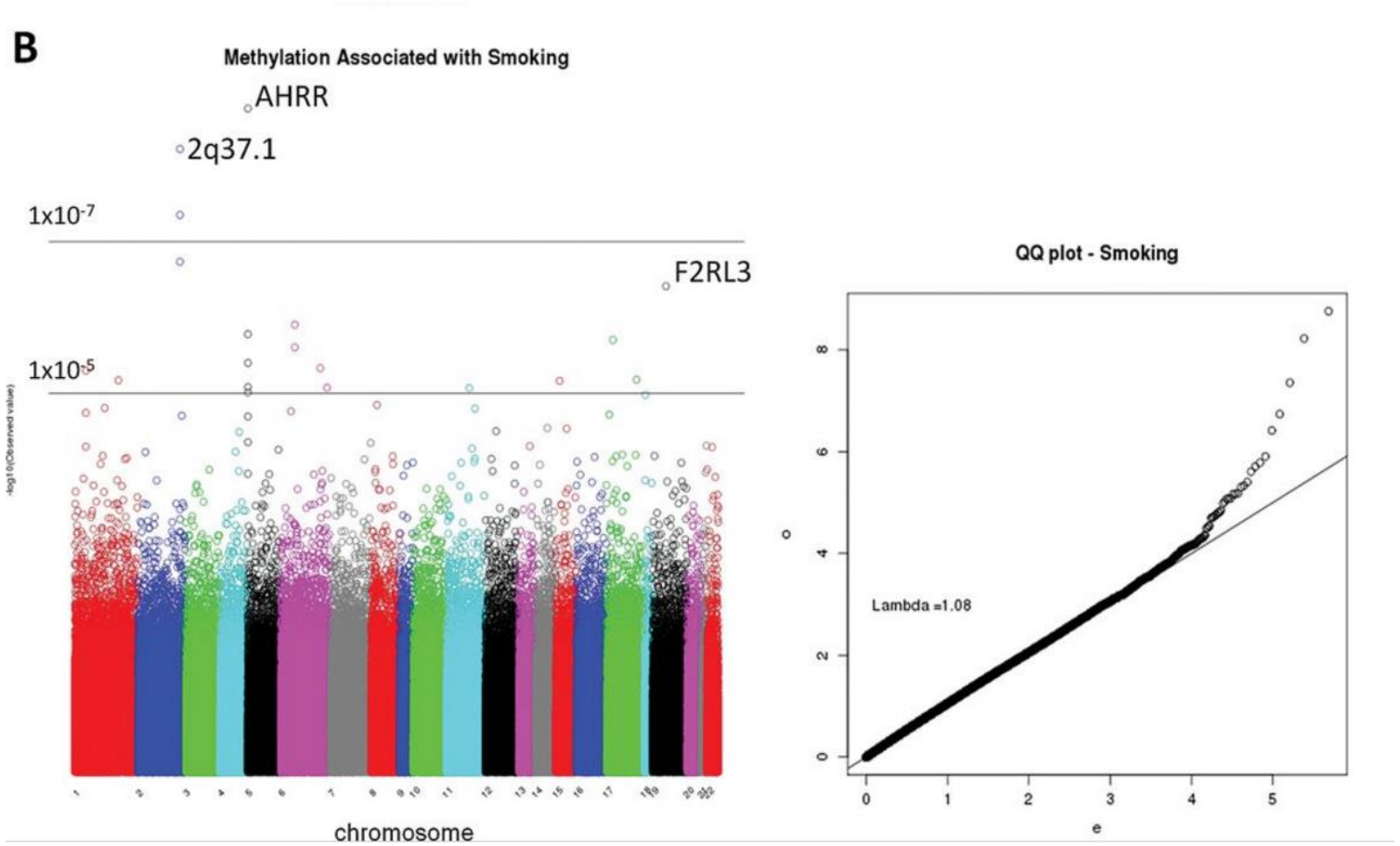
- Adjustment reduces confounding and improves genomic inflation



E.A. Houseman et al. *Curr Env Health Rept* (2015); 2.2 145-154

Population stratification

- Q-Q plots and lambda (λ) values are useful to evaluate population stratification
- λ can help detect systematic bias and population stratification
- Healthy Q-Q plot

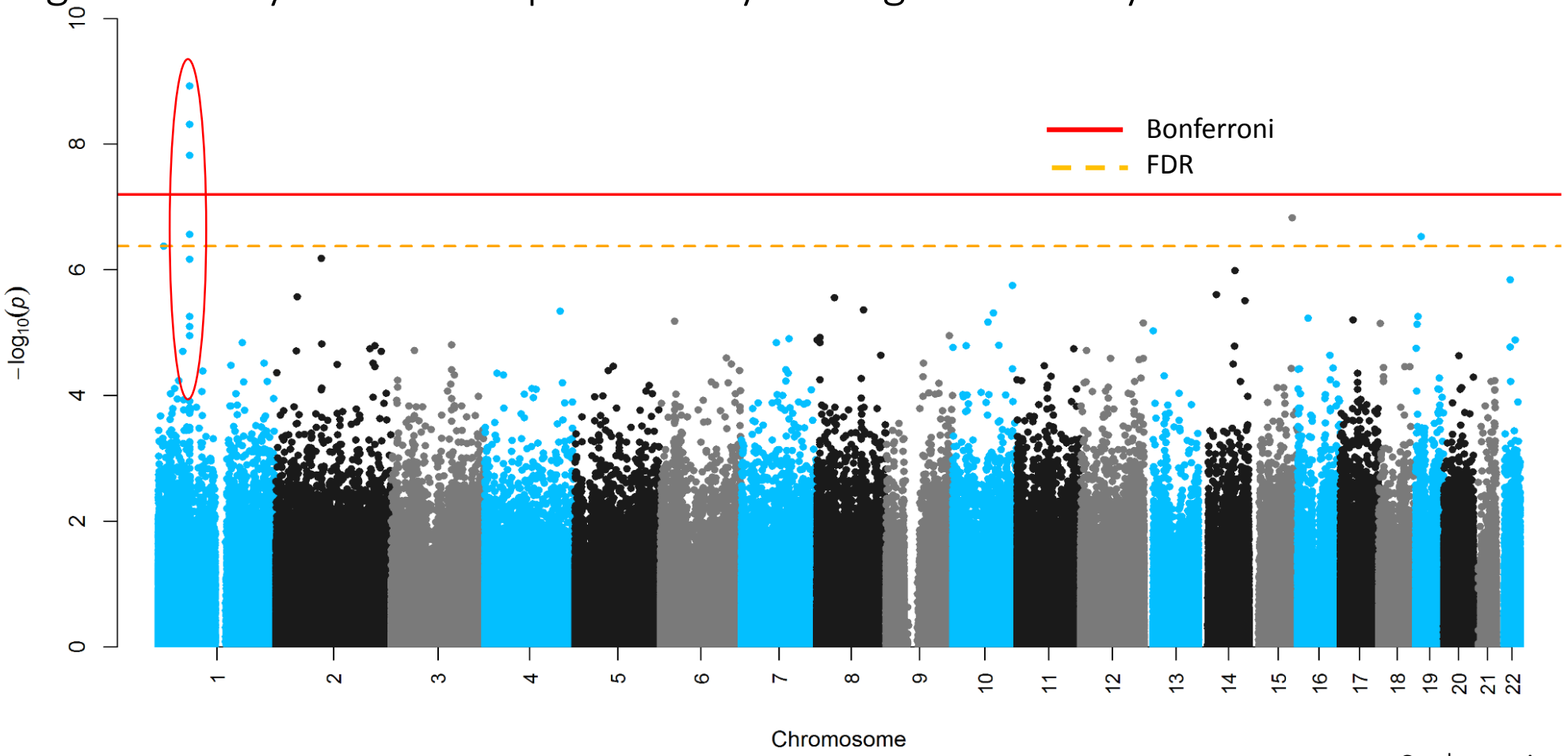


Shenker et al. *Hum. Mol. Genet.* (22.5 (2012): 843-851.

Manhattan plots: CpG-by-CpG Analyses

- Help identify more relevant biological signal

- Regional analyses are complementary to single site analyses

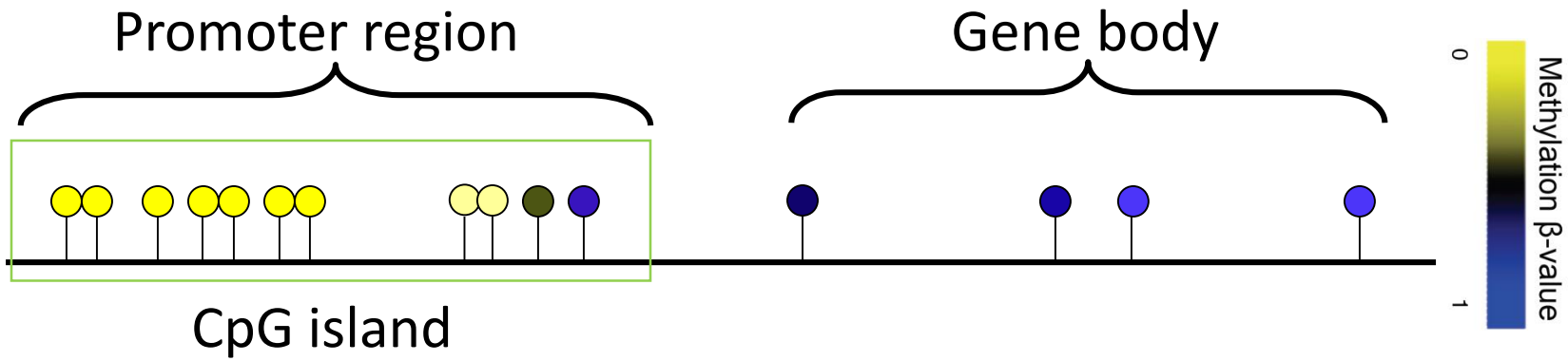


Cardenas, A., et al. *Diabetes* (*2018 Epub.*)

Regional Analyses

- CpG-by-CpG analyses: are we using the right tools?

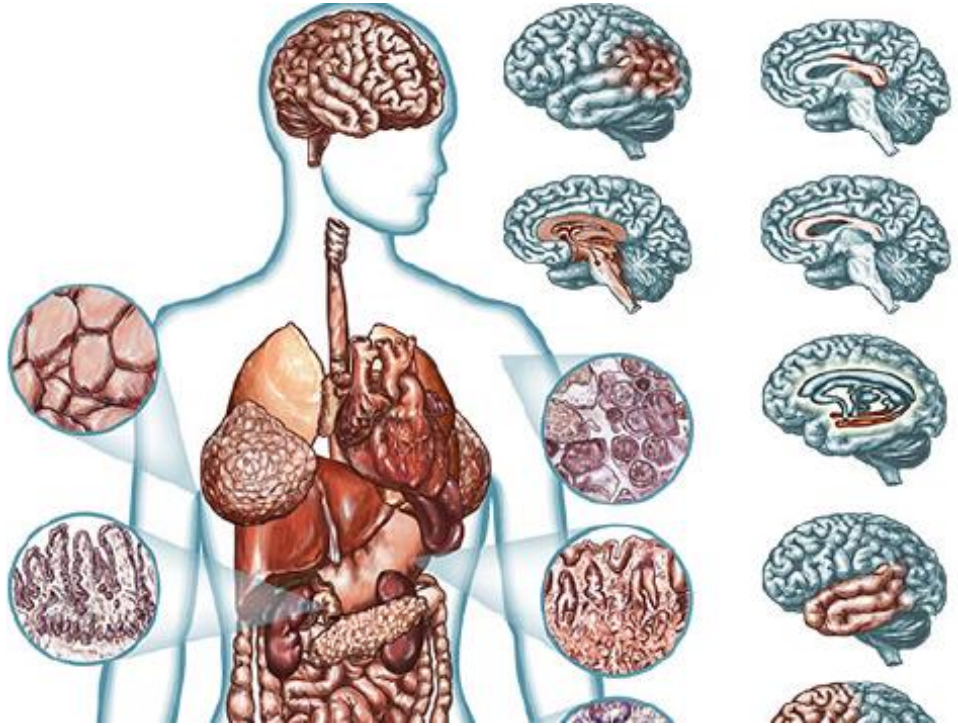
- What about CpGs that are close to each other (correlated)



- Methods are now available to model clusters of CpGs
  - Bump-hunting (Jaffe, et al. *Int J Epidemiol*, 2012)
  - Comb-P (Pedersen, et al. *Bioinformatics*, 2012)
  - DMRcate (Peters, et al. *Epigenetics Chromatin*, 2015)
  - Aclust (Tamar, et al. *Bioinformatics*, 2013)
  - Probe Lasso (Butcher, et al. *Methods*, 2015)

Target tissue

- Think about the cell-type composition of tissue
  - Blood: DNA from leukocytes
  - Cord blood: DNA from leukocytes and nucleated RBCs
  - Placenta: trophoblast, endothelial cells, vascular tissue, etc.



- Relevance for the disease/exposure of interest
  - Cardiovascular disease → blood might be OK
  - Cognition → Is blood relevant?

- Aim: discovery of biomarkers or mechanism?
  - Accessibility *vs.* biological relevance

- Top CpG sites might be passengers or drivers of associations
  - Causal claims must be critically examined
  - Replication ensures generalizability
  - Mendelian randomization methods

Cell Mixture in Epidemiological Studies

- Cell mixture remains an issue for other commonly used tissues
  - Placenta
  - Buccal cells
  - Nasal cells

- Reference free methods for adjustment are available
  - *EWASher* (Zou, et al. Nature Methods, 2014)
  - *RefFreeEWAS* (Houseman, et al. Bioinformatics, 2014)
  - *ReFACTor* (Elior, et al. Nature Methods, 2016)
  - *MeDeCom* (Pavlo, et al. Genome Biology, 2017)

- Assumes that some factors reflect cell-type distribution (SVA, PCA, etc..)

- Improvement of signal to noise ratio

- Overcorrection is a risk with reference-free methods

EWAS: Example

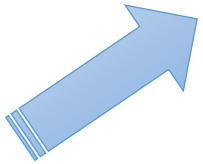
SCIENCE ADVANCES | RESEARCH ARTICLE

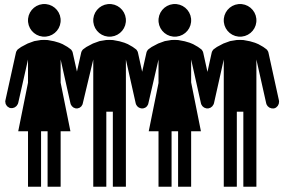
HUMAN GENETICS

DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood

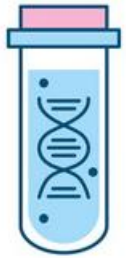
Elmar W. Tobi,[1,2] Roderick C. Slieker,[1] René Luijk,[1,3] Koen F. Dekkers,[1] Aryeh D. Stein,[4] Kate M. Xu,[3,5] Biobank-based Integrative Omics Studies Consortium,* P. Eline Slagboom,[1] Erik W. van Zwet,[3] L. H. Lumey,[1,6†] Bastiaan T. Heijmans[1†‡]

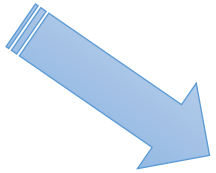
- Dutch Hunger Winter
  - Prenatal famine exposure
    - <900 kcal/day





- DNAm in whole blood
  - Illumina 450K ~ 450,000 CpGs





- Serum triglycerides
- BMI

Tobi, EW., et al. *Science Advances* 4.1 (2018): eaao4364.

EWAS: Example

- Mediation on triglycerides levels and BMI

Table 3. Mediation analysis: DNAm and the association between famine exposure and triglycerides.

CpG	Location (hg19)	Nearest gene*	Methylation (SD)†	Rank	EWAS P_{FDR}	P_{famine}	P_{TG}	Previous studies	$\beta_{mediation}$‡	*P* mediation	Proportion mediated (%) [95% CI]§	$P_{proportion}$
cg19693031	chr1:145441552	*TXNIP*	77.5 (4.3)	6	2.6×10^{-5}	4.8×10^{-3}	2.3×10^{-11}	(*41–45*)	2.6 [0.7–4.8]	0.005	28.0 [5.7–100]	0.026
cg18120259	chr6:43894639	*LOC100132354*	60.4 (4.7)	10	1.8×10^{-3}	6.6×10^{-4}	6.4×10^{-8}	(*44*)	2.3 [0.8–4.1]	0.001	24.9 [7.5–100]	0.021
cg15020801	chr17:46022809	*PNPO*	36.1 (3.4)	12	3.5×10^{-3}	7.1×10^{-4}	6.0×10^{-8}	(*30*)	2.3 [0.9–4.2]	0.001	25 [7.0–100]	0.022
cg06983052	chr1:90288099	*LRRC8D*	64.8 (3.8)	13	4.2×10^{-3}	1.0×10^{-5}	5.3×10^{-6}		2.6 [1.1–4.5]	<0.001	28.0 [8.8–100]	0.024
cg07397296	chr21:43655316	*ABCG1*	26.9 (3.8)	14	0.021	5.1×10^{-3}	1.9×10^{-7}	(*49*)	1.9 [0.6–3.6]	0.005	20.5 [4.6–97.4]	0.027
cg20496314	chr22:39759864	*SYNGR1*	40.2 (4.3)	15	0.032	3.9×10^{-3}	1.5×10^{-7}	(*45, 46*)	1.8 [0.5–3.5]	0.007	19.6 [3.6–88.3]	0.026

*Nearest gene within 100 kb. †The Illumina 450k array β value (ranging from 0 to 1) multiplied by 100 for easy interpretation. ‡This is the estimate and CI based on 10K Monte Carlo simulations of the indirect effect or mediation effect, which is often referred to as the "$a \times b$" effect. §The percentage of the total exposure-phenotype relationship explained by the indirect (mediated) effect as based on 10K Monte Carlo simulations.

Tobi, EW., et al. *Science Advances* 4.1 (2018): eaao4364.

Future: Biomarker Development

- DNAm in **cord-blood** predicts (prenatal) maternal smoking (3-24 CpGs)

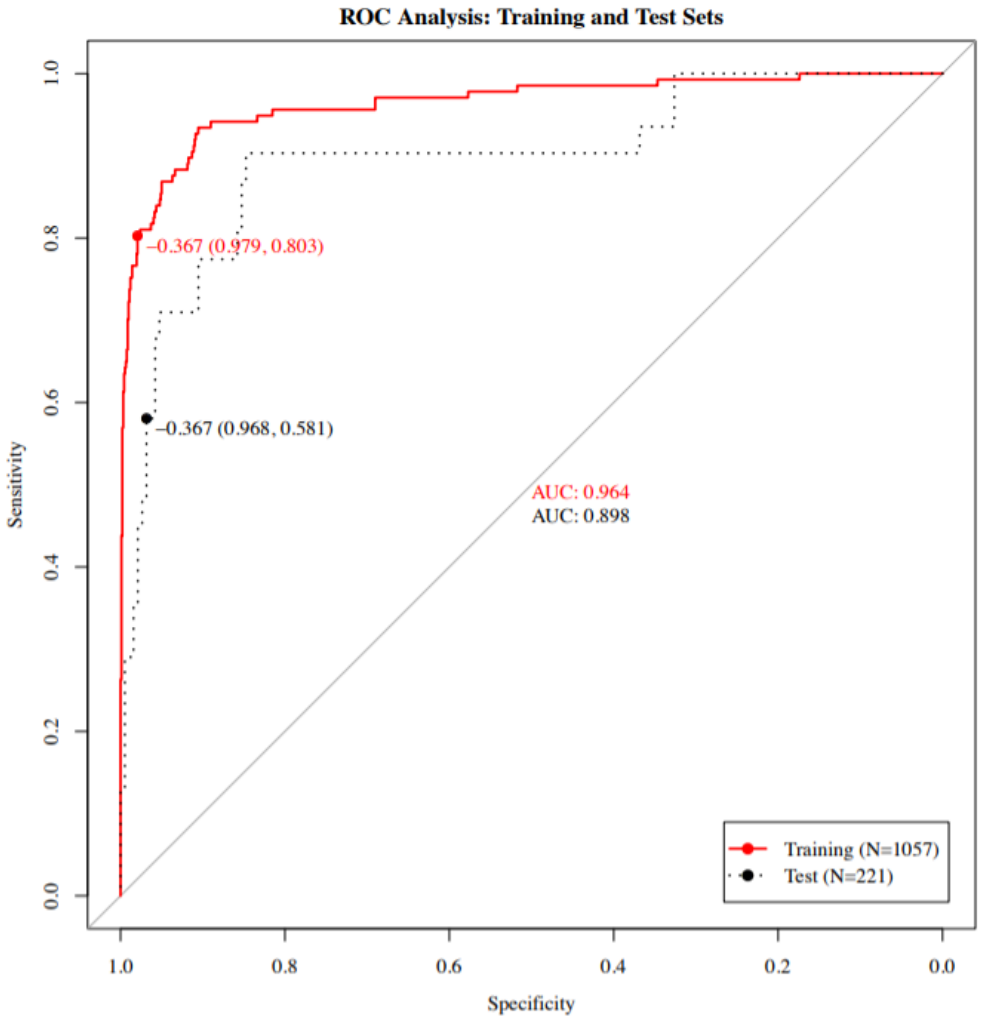


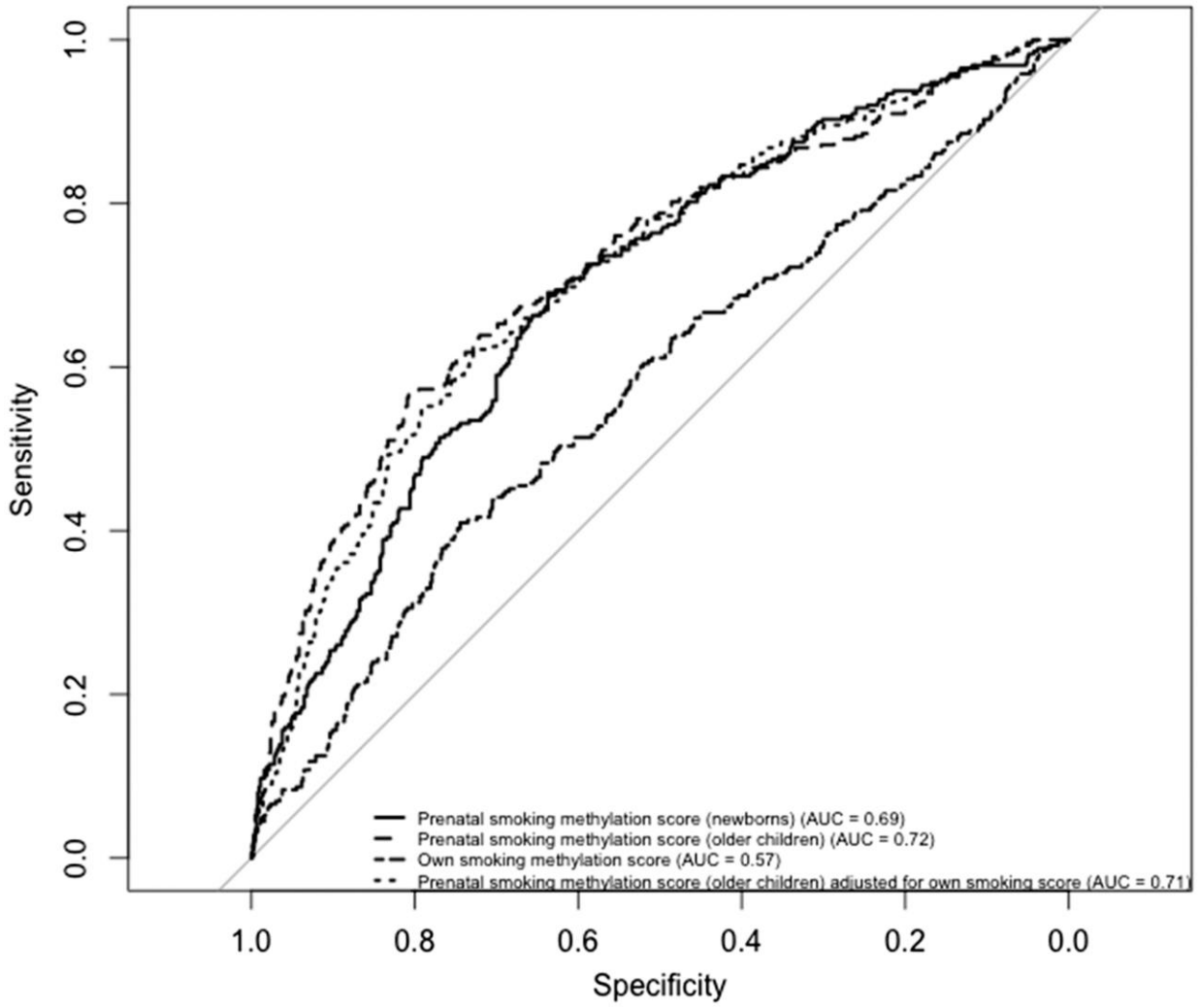
Table 2. Logistic LASSO results for main and additional analyses.

Model	Data set	*q*	AUC (CI)
a Cotinine-based sustained smoking	Training[a]	24	0.97 (0.95, 0.99)
	Test[a]		0.88 (0.80, 0.96)
b Self-reported sustained smoking	Training[a]	12	0.93 (0.90, 0.96)
	Test[a]		0.82 (0.74, 0.91)
c Combined sustained smoking[b]	Training[a]	28	0.96 (0.95, 0.98)
	Test[a]		0.90 (0.83, 0.97)
d Naïve CpG selection[c]	Training[a]	3	0.89 (0.86, 0.92)
	Test[a]		0.82 (0.73, 0.91)

Reese, SE., et al. *Environmental Health Perspectives* 125.4 (2017): 760

Future: Biomarker Development

- DNAm of blood in **adults** predicts (prenatal) maternal smoking

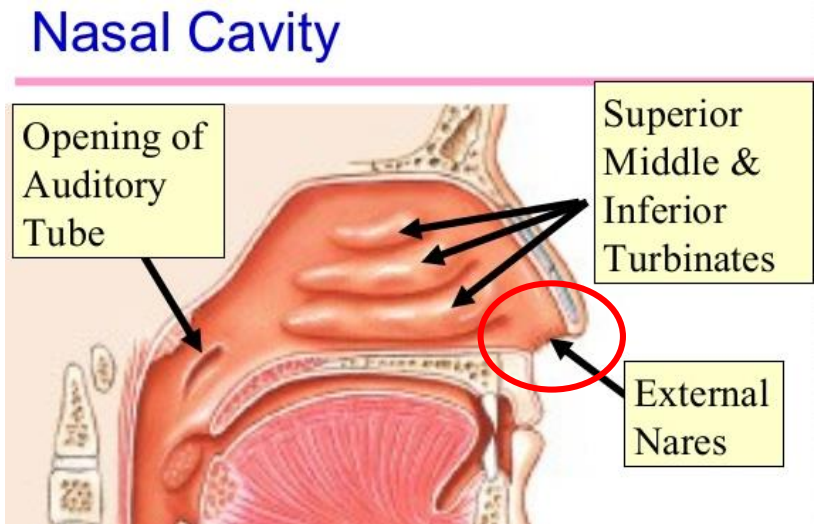


- 15 CpGs predicting prenatal smoking
- AUC= **0.72** (95% CI: 0.69, 0.76)
- Persistence of signature

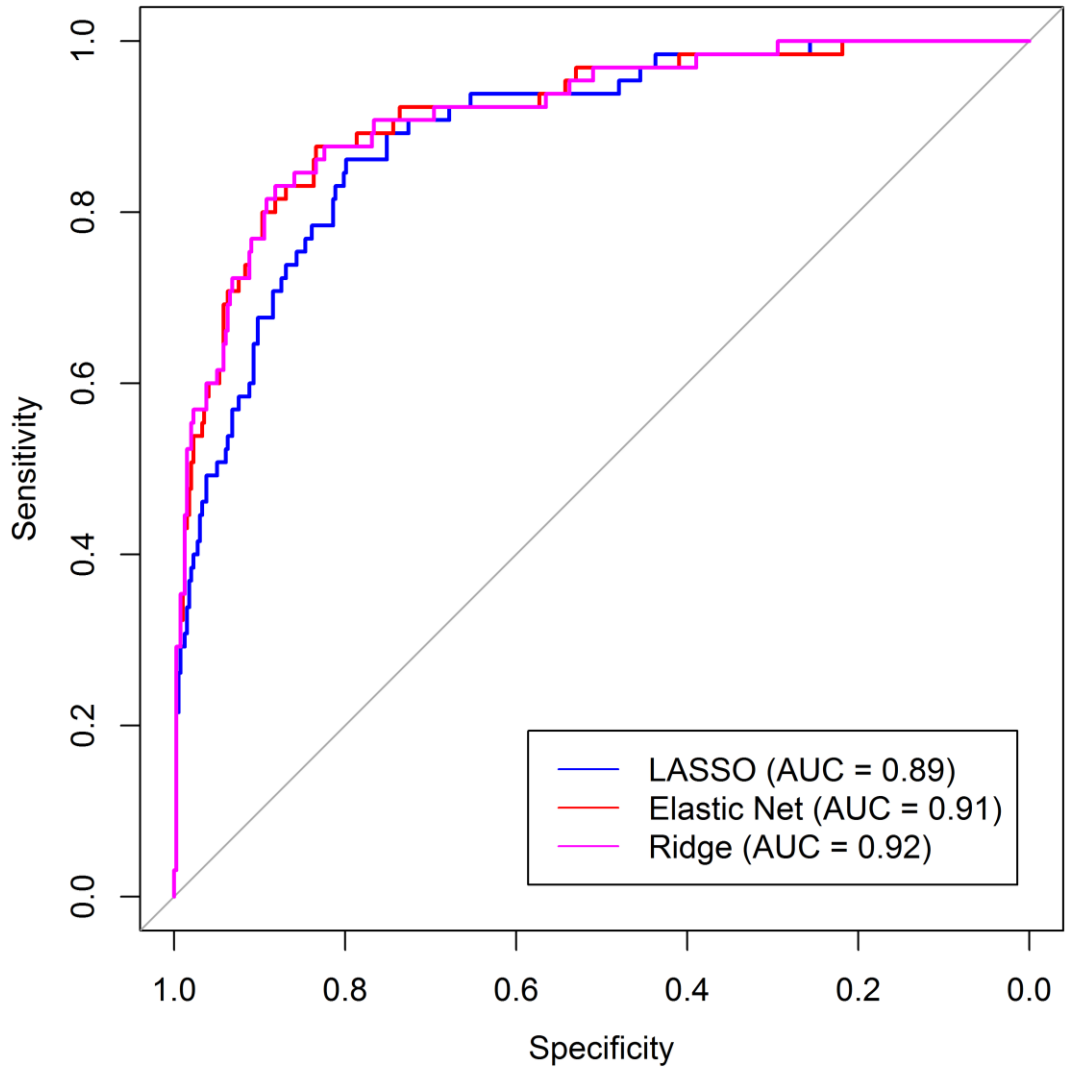
Richmond, R. et al. *International Journal of Epidemiology* (2018) Epub.

Future: Biomarker Development

- Nasal Methylome
  - In direct contact with the environment
  - Different tissues give you different information



Asthma prediction



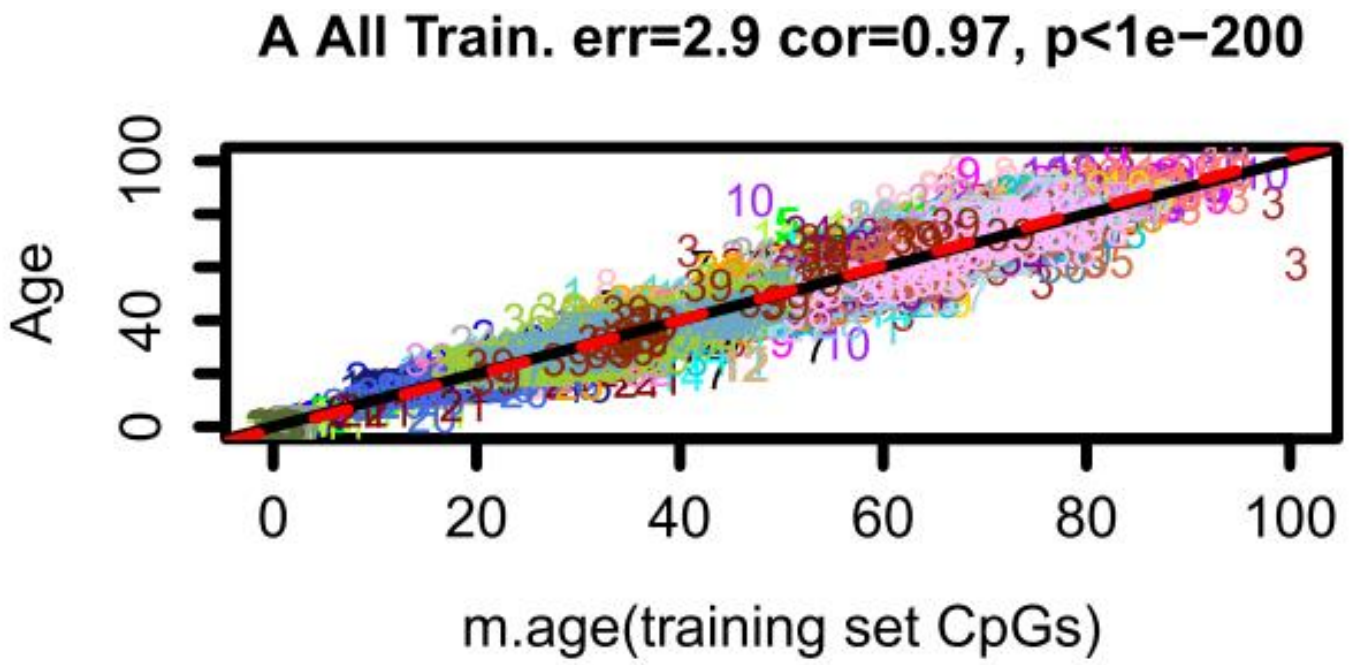
Cardenas, A., et al. (In Preparation)

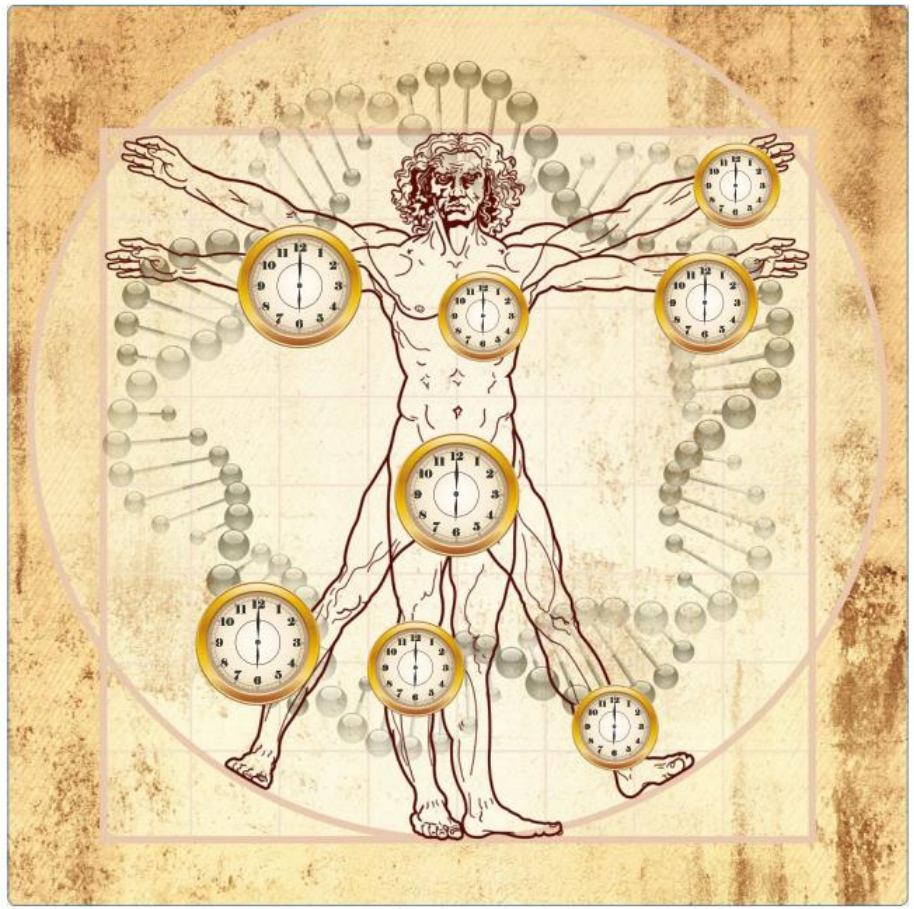
A note on effect sizes: smoking – DNAm

Joubert, Bonnie, et al. *Environmental Health Perspectives*. (2012)

Epigenetic Clock (DNAm Age)

- **DNAm age (Horvath clock)**
  - Highly correlated with chronological age
  - Accelerated by environmental exposures
  - Predicts mortality
  - Multi-tissue predictor



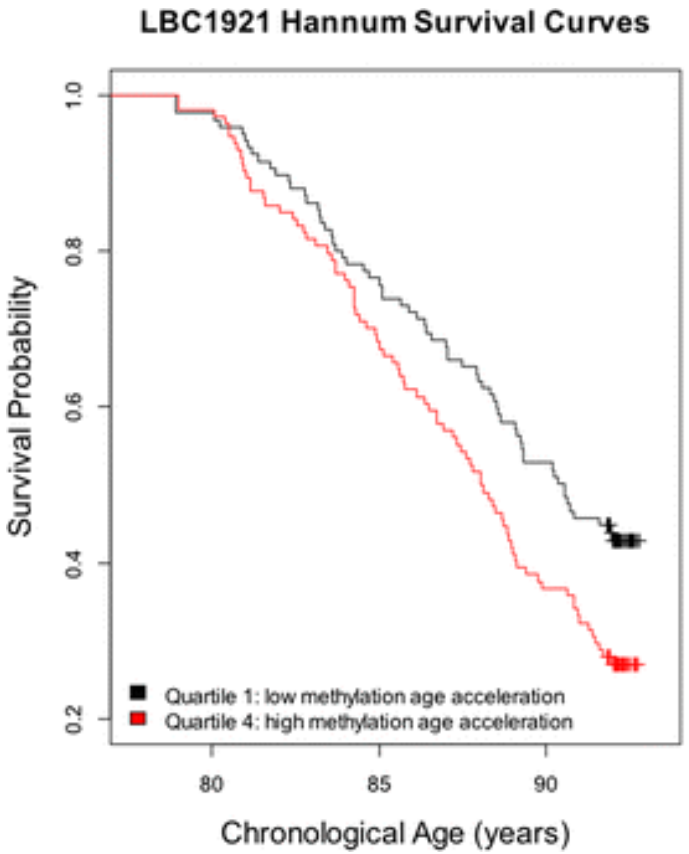


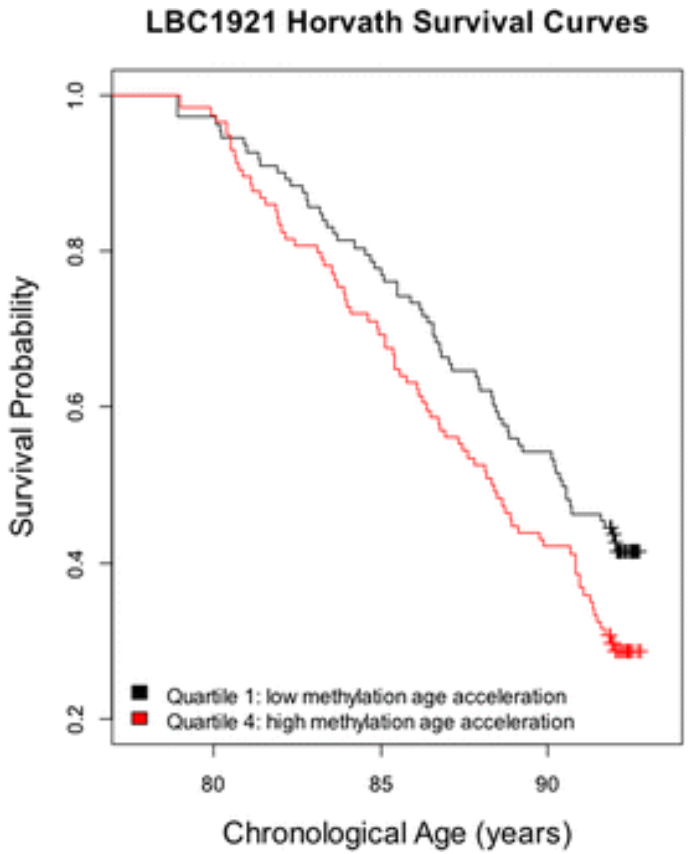
DNA methylation age of human tissues and cell types

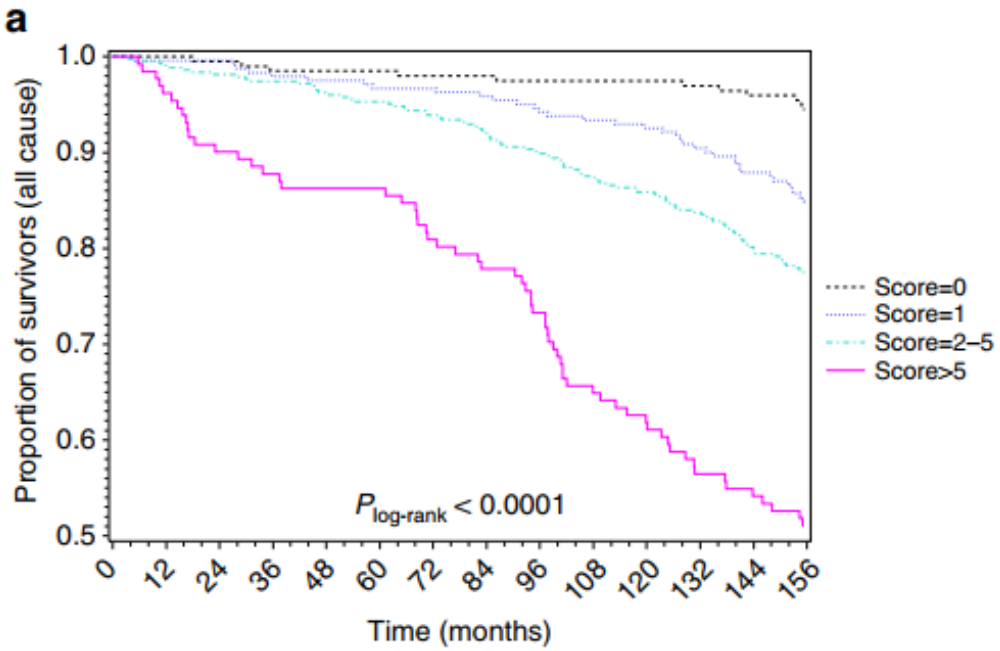
Horvath

Horvath, Steve. *Genome Biology* 14.10 (2013)

Age Acceleration: DNAm Age > Chronological Age

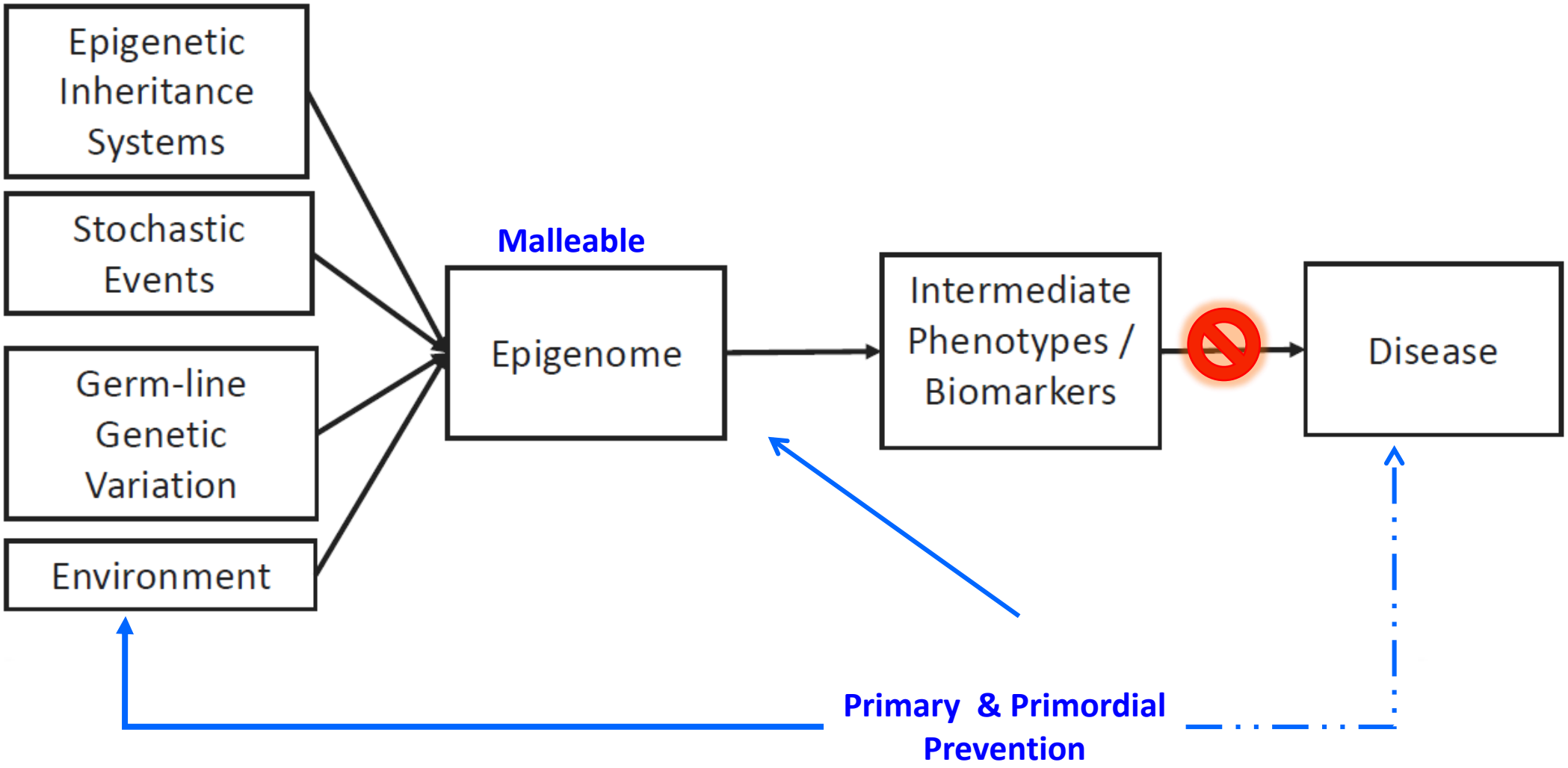






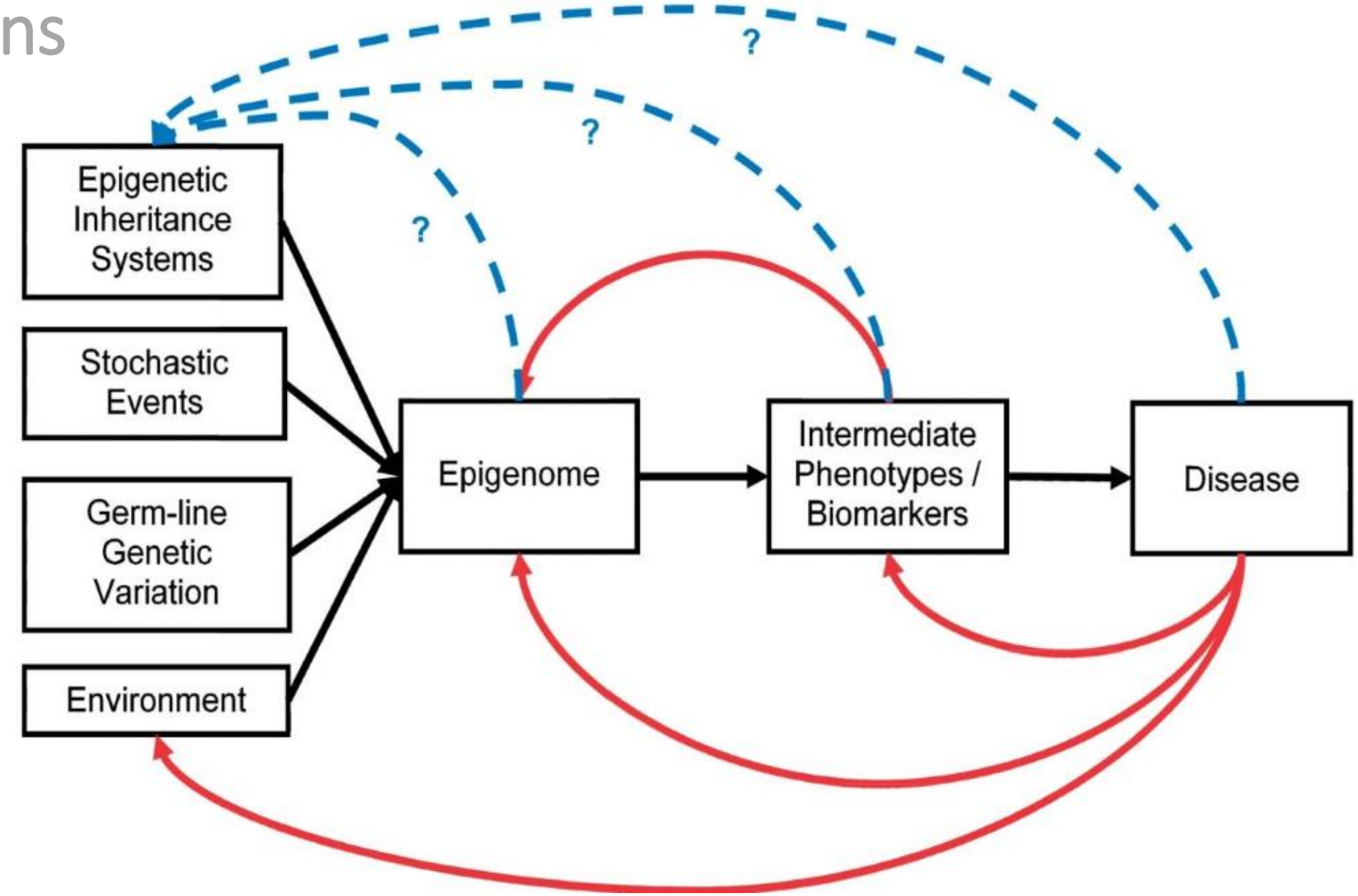
Marioni, Riccardo E., et al. *Genome Biology* 16.1 (2015): 25 & Zhang, Yan, et al. *Nature Communications* 8 (2017): 14617

Public Health Significance



Relton C. & Davey-Smith G. *International Journal of Epidemiology* 41.1 (2012): 5-9

Limitations



Relton C. & Davey-Smith G. *International Journal of Epidemiology* 41.1 (2012): 5-9

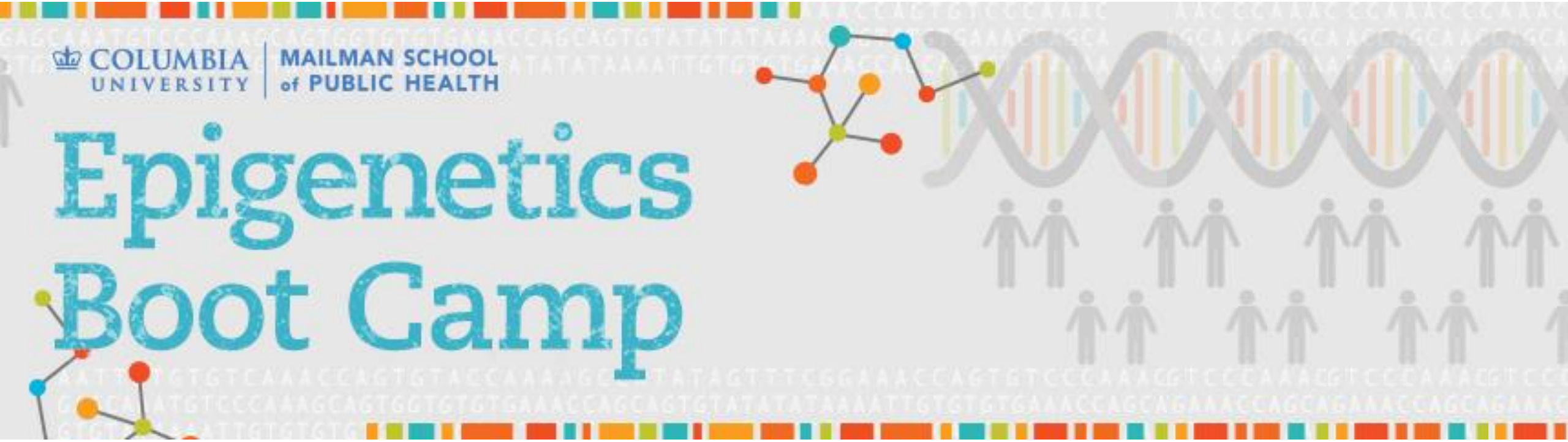
Summary

- Epigenetics as the interface between the genome and the environment
  - Address hypothesis on fetal origins of adult disease
  - Develop biomarkers of exposures and outcomes
  - Powerful tool for birth cohorts & early life events

- Study considerations
  - Big data does not circumvent epidemiological issues
  - Design matters (case-control/prospective/retrospective)
  - Collect relevant tissues/samples at appropriate time-points

- Analytical issues
  - Sample size (Parameters>>N)
  - Randomize samples and ensure balance by trait of interest
  - Collaborative science

- Welcome to the Epigenetic Revolution!



Planning and Analyzing DNA Methylation Studies

Columbia University – New York City

- 2-day Boot Camp – Instructors with 40+ years of combined epigenetics experience from Columbia, Harvard, and Icahn School of Medicine at Mount Sinai
- Learn more & sign up to hear about next course: mailman.columbia.edu/bootcamp

Resources

- Coursera free online class: Epigenetic Control of Gene Expression
  - https://www.coursera.org/learn/epigenetics
- NIH Roadmap Epigenomics Project
  - *http://www.roadmapepigenomics.org/*

- Bioconductor workflow for data analysis
  - https://f1000research.com/articles/5-1281/v3
- DNA methylation: roles in mammalian development
  - https://www.nature.com/articles/nrg3354
- Consortium: Pregnancy And Childhood Epigenetics (PACE)
  - https://www.niehs.nih.gov/research/atniehs/labs/epi/pi/genetics/pace/index.cfm

PACE
Pregnancy And Childhood Epigenetics

Discussion & Questions

Contact: cardenas@hsph.harvard.edu
@cardenaasca