Introduction into DNA methylation data analysis

Data Analysis project MoBi

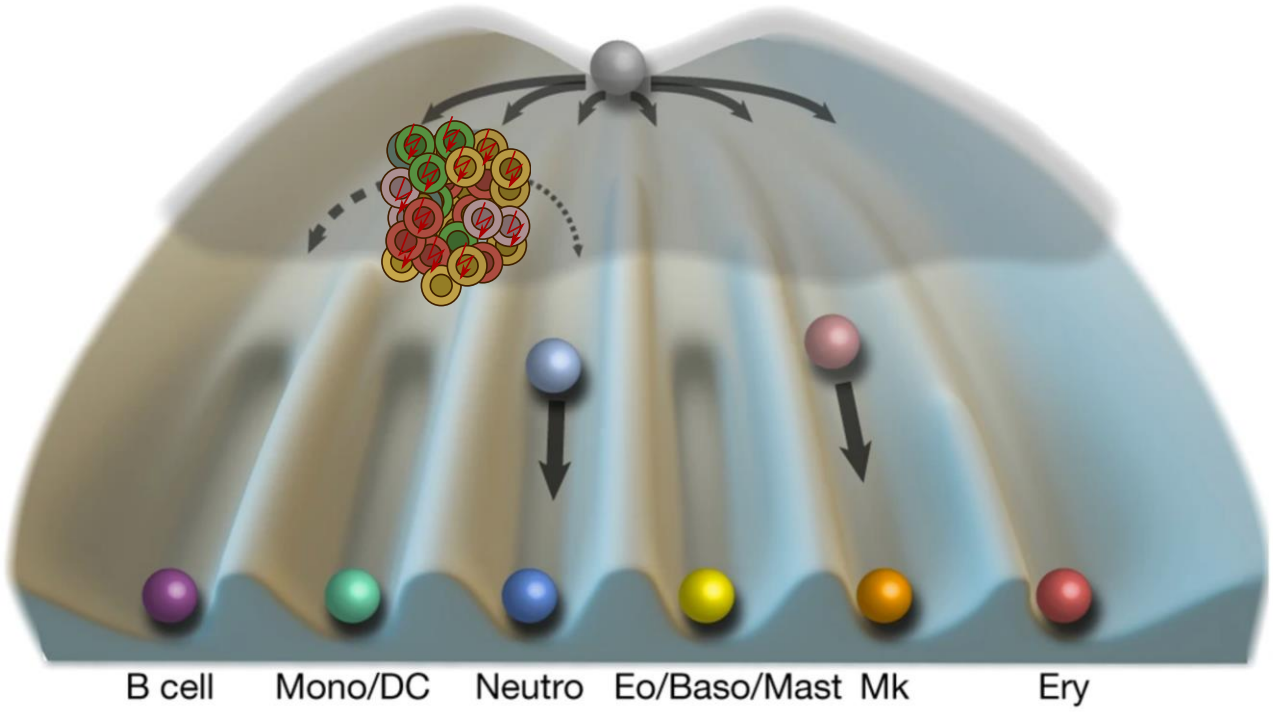
April 2024

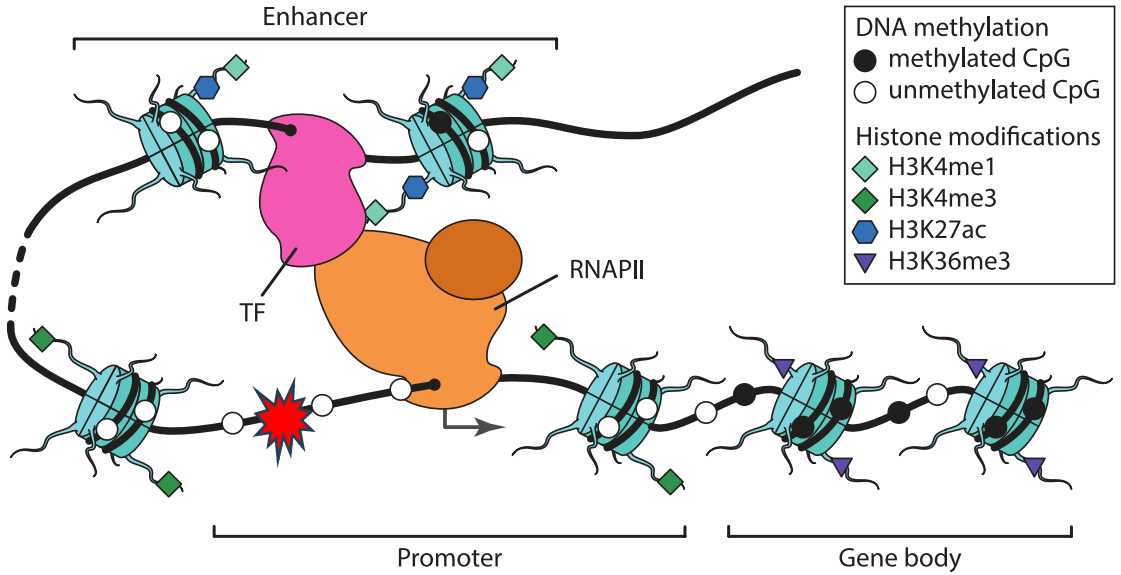
Michael Scherer (michael.scherer@dkfz.de)

Tutor: Stefanie Mantz (s.mantz@stud.uni-heidelberg.de)

dkfz.
GERMAN CANCER RESEARCH CENTER
IN THE HELMHOLTZ ASSOCIATION
Research for a Life without Cancer

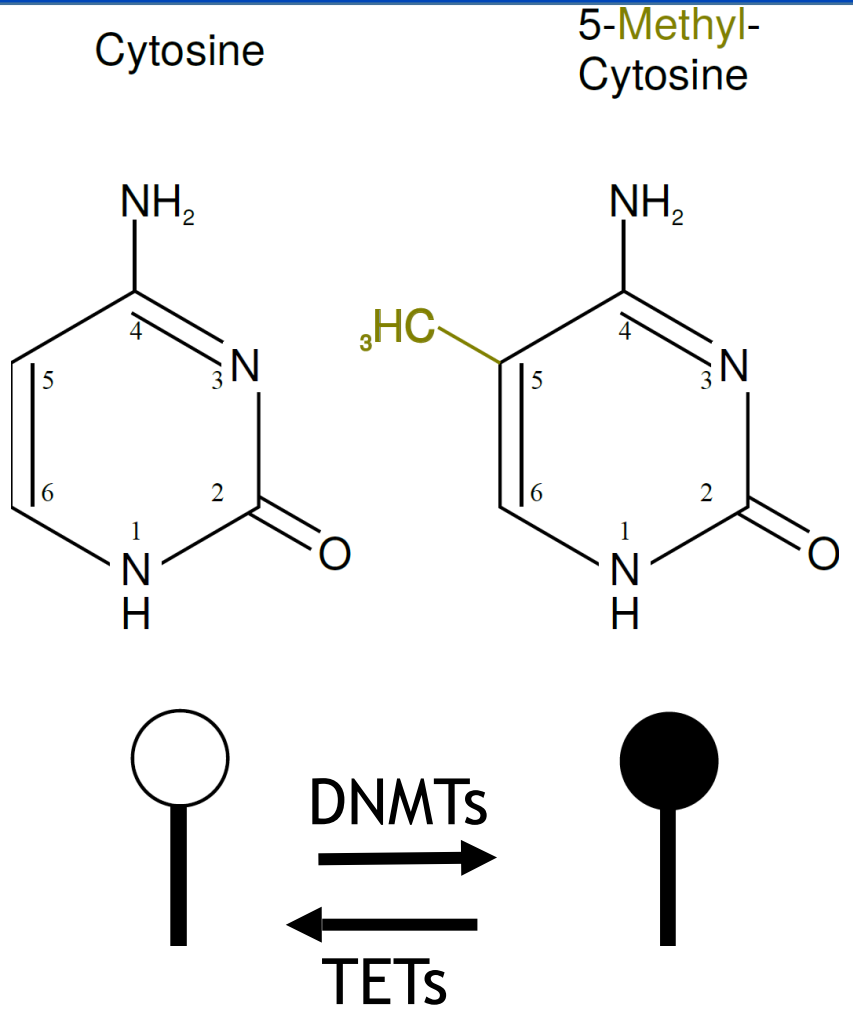
Epigenetic regulation shapes cellular identity



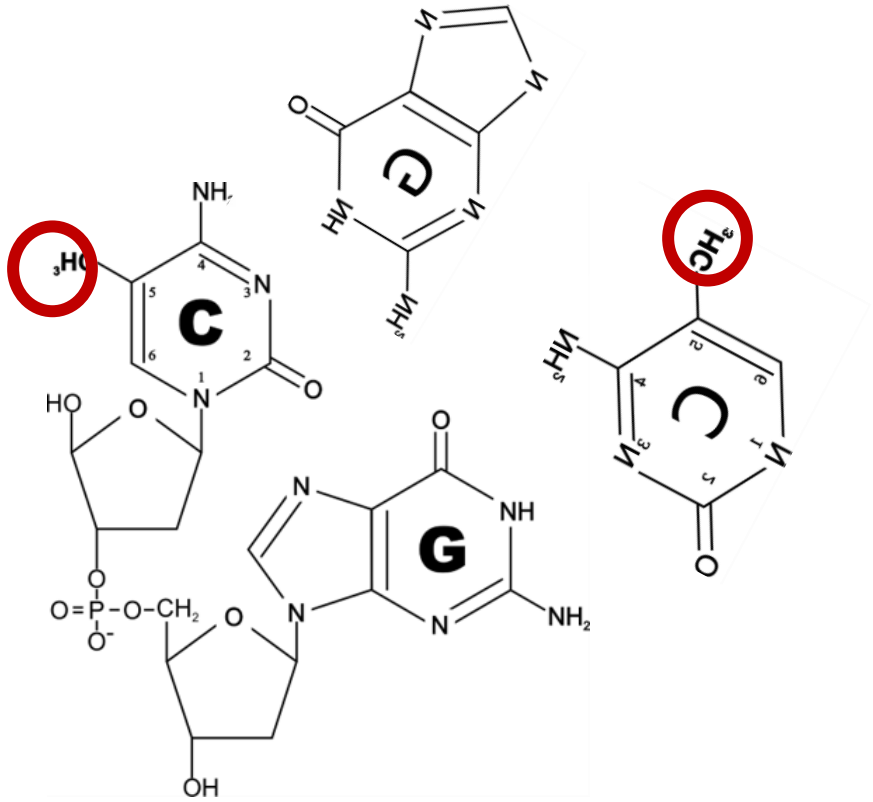


Adapted from *Velten et al. (2017) Nature Cell Biology* Created by *Fabian Müller, https://doi.org/10.6084/m9.figshare.5285473.v1*

DNA methylation is a reversible epigenetic modification



- Almost exclusively in CpG context



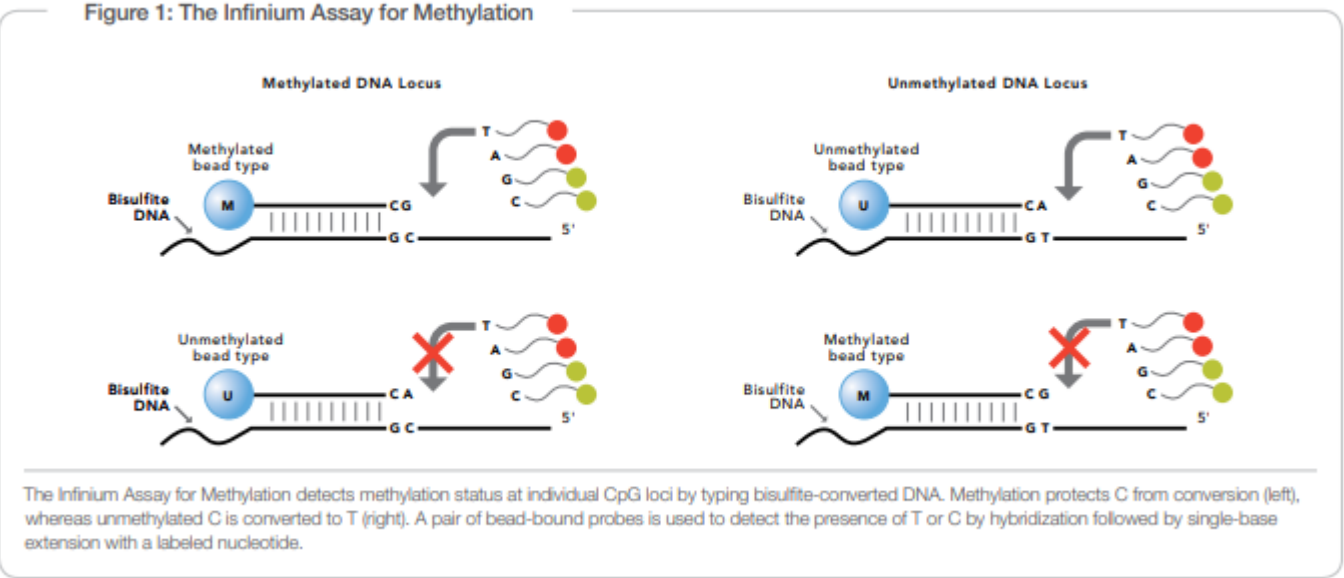
Genomics methods for measuring DNA methylation

- ~28 Mio CpGs in the human genome
- Common methods
  - WGBS = whole-genome bisulfite sequencing
  - RRBS = reduced-representation bisulfite sequencing (~4-5 Mio)
  - Microarrays
    - **450k**
    - EPIC: 850,000 CpGs
    - EPICv2: almost 1,000,000 CpGs

What kind of data are we analyzing?

Illumina 450k bead array data

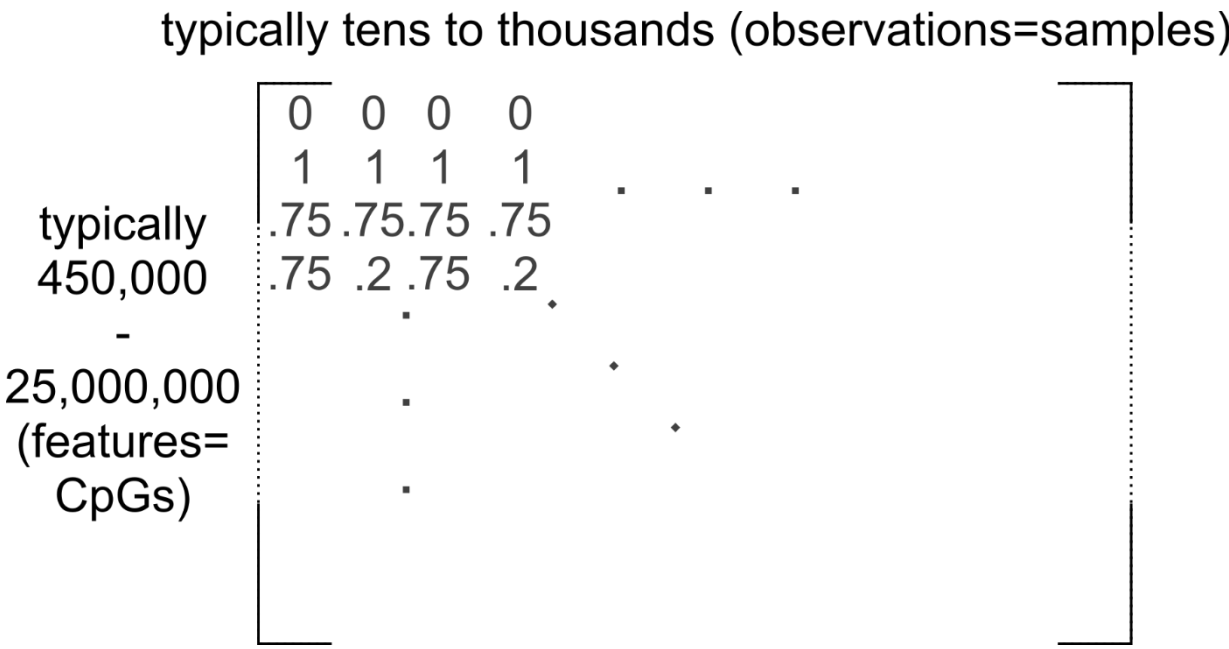
- Microarray technology
- High-throughput, relatively cheap
- High confidence DNA methylation calls
- ~450,000 CpGs



From: https://www.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/appnote_dna_methylation_analysis_infinium.pdf

DNA methylation data matrix

- Rows are typically individual CpGs, but also predefined genomic regions, e.g. promoters
- Columns are samples
- Obtained either using bisulfite sequencing or the Illumina Infinium BeadArrays



Where does the exact dataset we use come from?

Generated in the context of The Cancer Genome Atlas (TCGA) project:

- Group 1: BRCA = Breast invasive carcinoma
- Group 2: LUSC = Lung squamous cell carcinoma
- Group 3: COAD = Colon adenocarcinoma
- Group 4: THCA = Thyroid carcinoma
- Group 5: KIRC = Kidney renal clear cell carcinoma

Which types of analyses are expected?

- Removal/imputation of missing values
- Visualization
- Low-dimensional representation (PCA/MDS)
  - Of individual CpGs
  - Of aggregated values over pre-defined genomic regions
- Statistical modelling
  - Determine cancer-specific DNA methylation changes
  - Determine relationships between individual CpGs