THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Methylome-wide Association Studies

Part 1: Data preparation

Acknowledgement of Country

The University of Queensland (UQ) acknowledges the Traditional Owners and their custodianship of the lands on which we meet.

We pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.

We recognise their valuable contributions to Australian and global society.

Image: Digital reproduction of *A guidance through time* by Casey Coolwell and Kyra Mancktelow

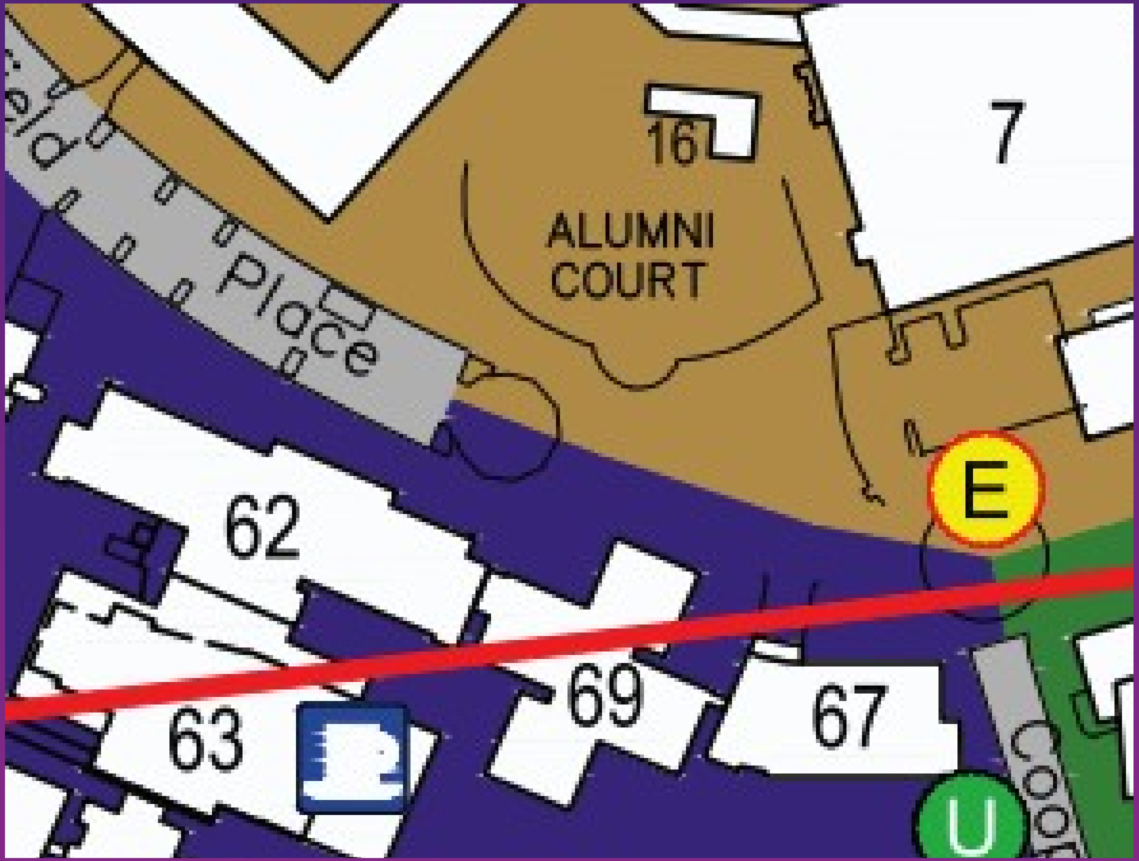
General Information:

- We are currently located in Building 69

E Emergency evacuation point

- Food court and bathrooms are located in Building 63

- If you are experiencing cold/flu symptoms or have had COVID in the last 7 days please ensure you are wearing a mask for the duration of the module



Data Agreement

To maximize your learning experience, we will be working with genuine human genetic data, during this module.

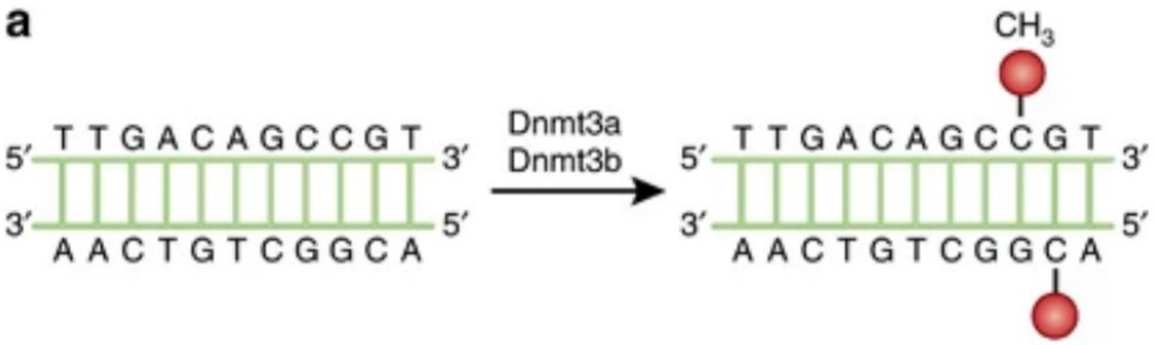
Access to this data requires agreement to the following in to comply with human genetic data ethics regulations

Please email pctgadmin@imb.uq.edu.au with your name and the below statement to confirm that you agree with the following:

“I agree that access to data is provided for educational purposes only and that I will not make any copy of the data outside the provided computing accounts.”

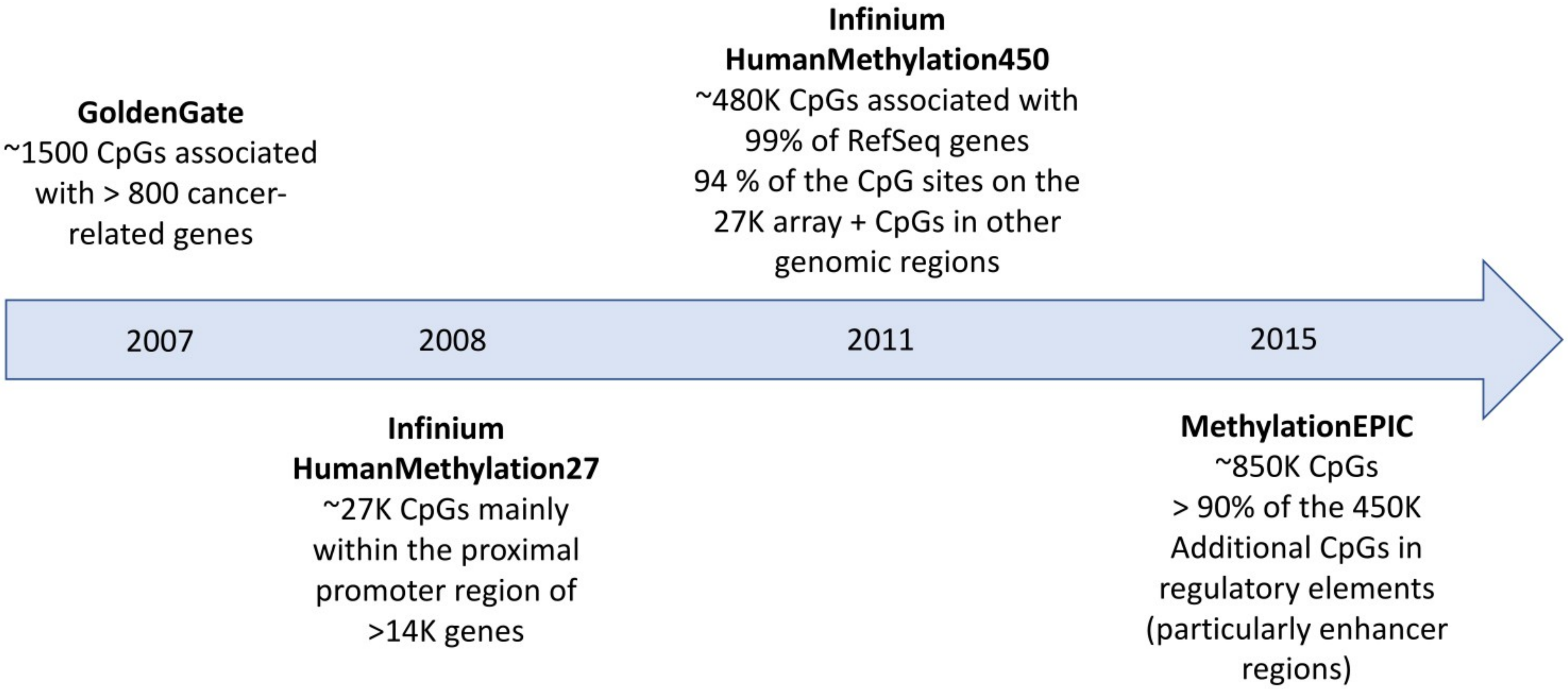
DNA Methylation

- Addition of a methyl (CH_3) group at a cytosine base
- In mammals, occurs primarily at CpG di-nucleotides
- Mediates the diversified gene expression profiles in a variety of cells and tissues in multicellular organisms


a
5′ T T G A C A G C C G T 3′
3′ A A C T G T C G G C A 5′
Dnmt3a
Dnmt3b
CH$_3$
5′ T T G A C A G C C G T 3′
3′ A A C T G T C G G C A 5′

Methylation Arrays

- Most common technology used for large cohorts



2007

GoldenGate
~1500 CpGs associated with > 800 cancer-related genes

2008

Infinium HumanMethylation27
~27K CpGs mainly within the proximal promoter region of >14K genes

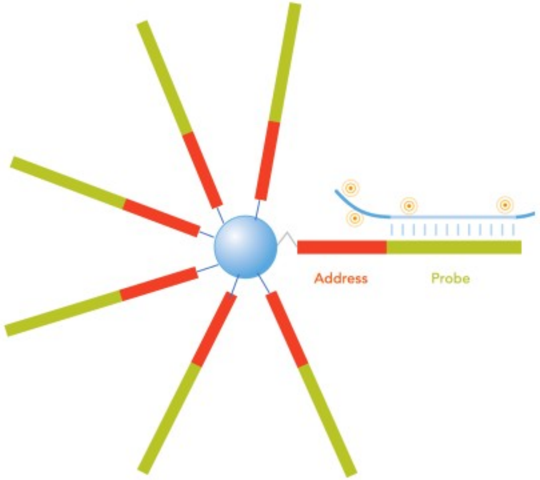
2011

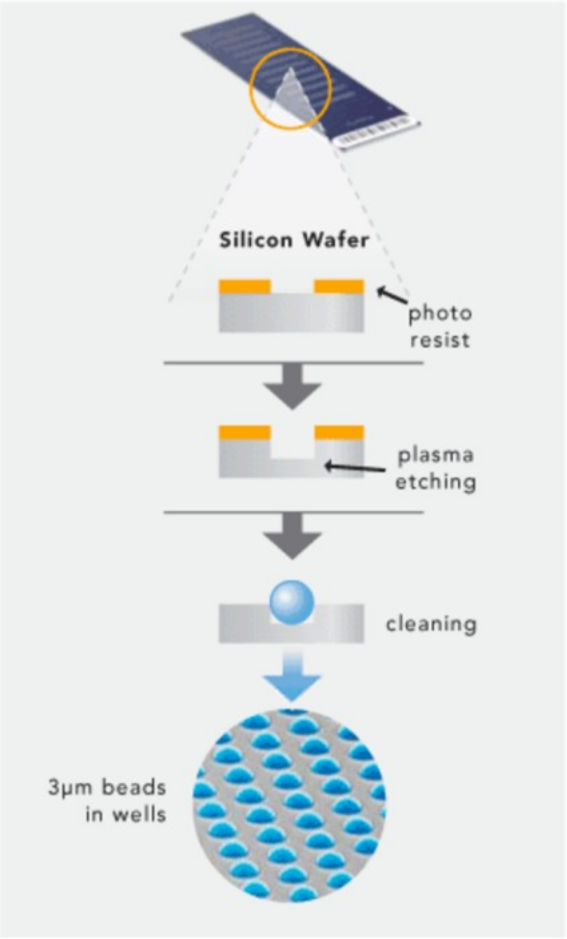
Infinium HumanMethylation450
~480K CpGs associated with 99% of RefSeq genes
94 % of the CpG sites on the 27K array + CpGs in other genomic regions

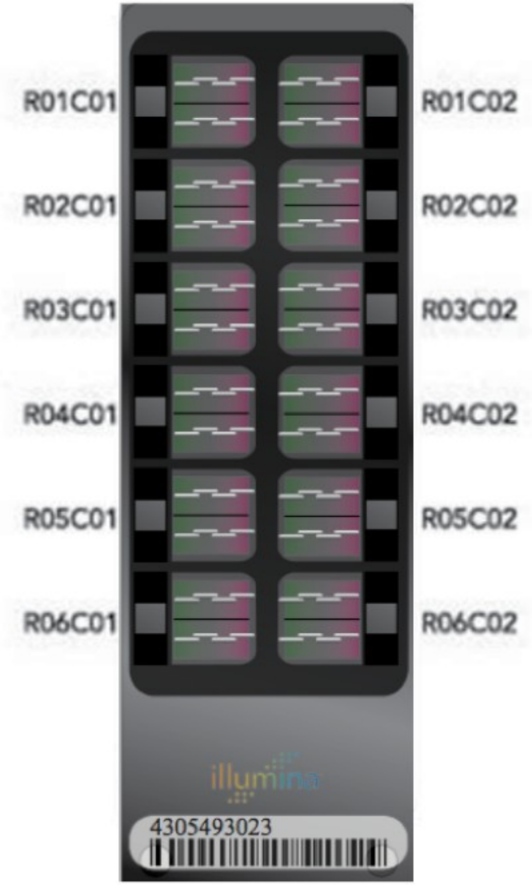
2015

MethylationEPIC
~850K CpGs
> 90% of the 450K
Additional CpGs in regulatory elements (particularly enhancer regions)

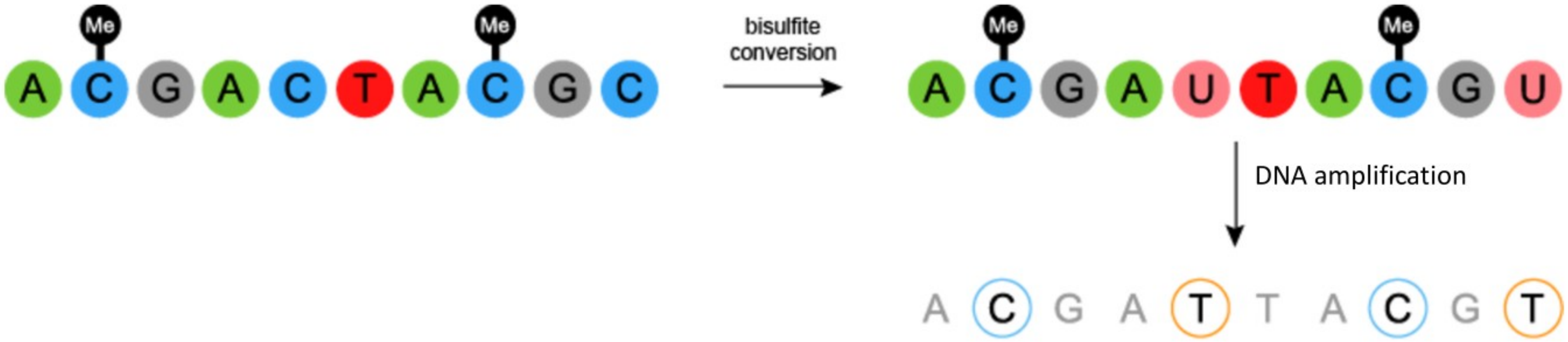
Beadchip Technology



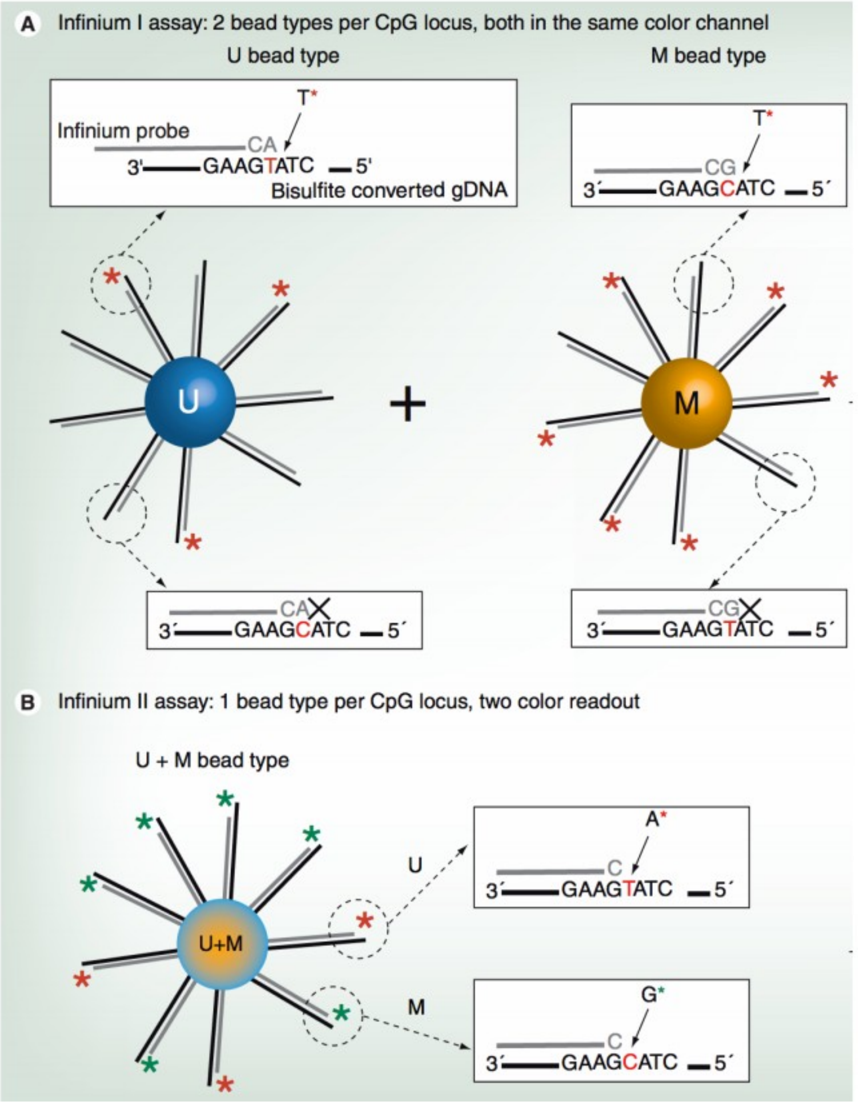


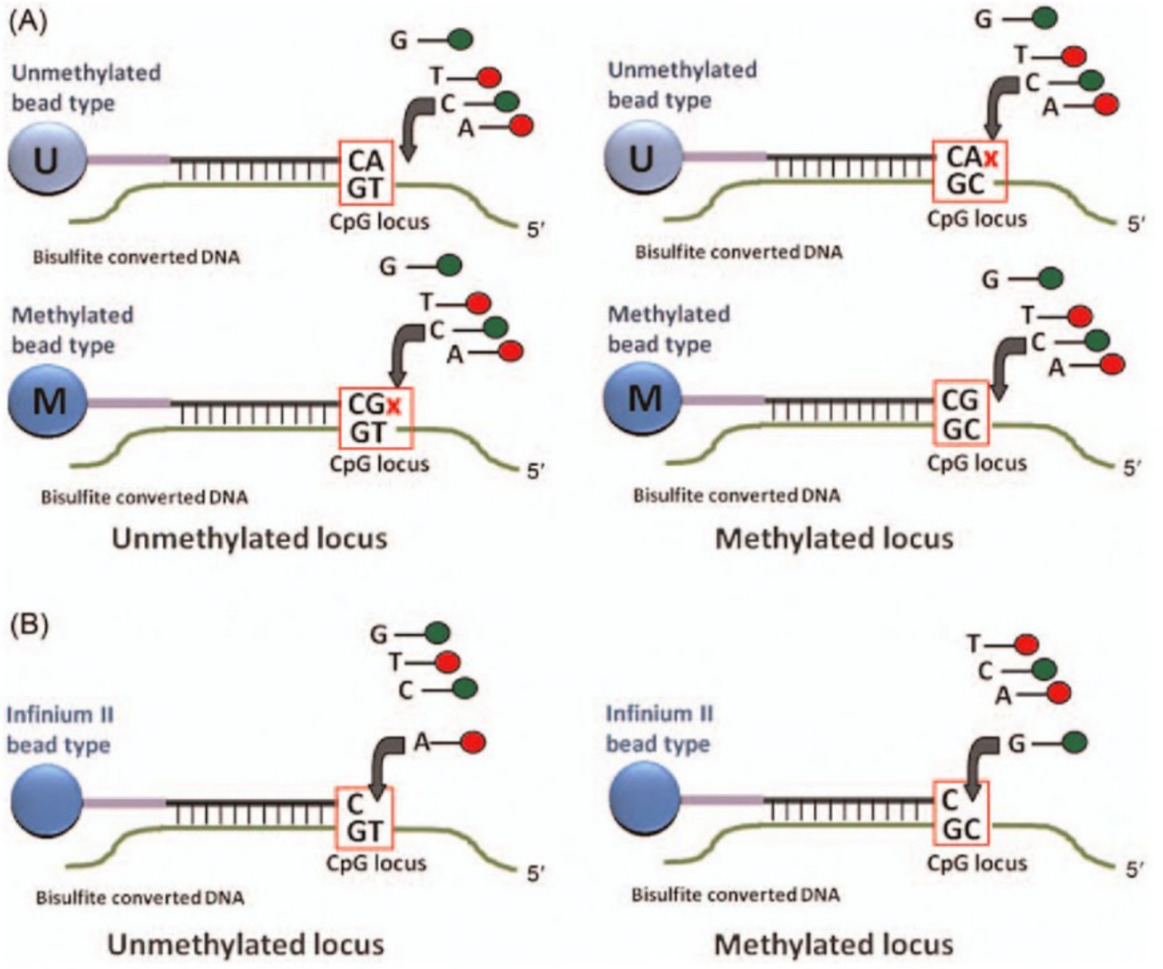


Bisulphite Conversion



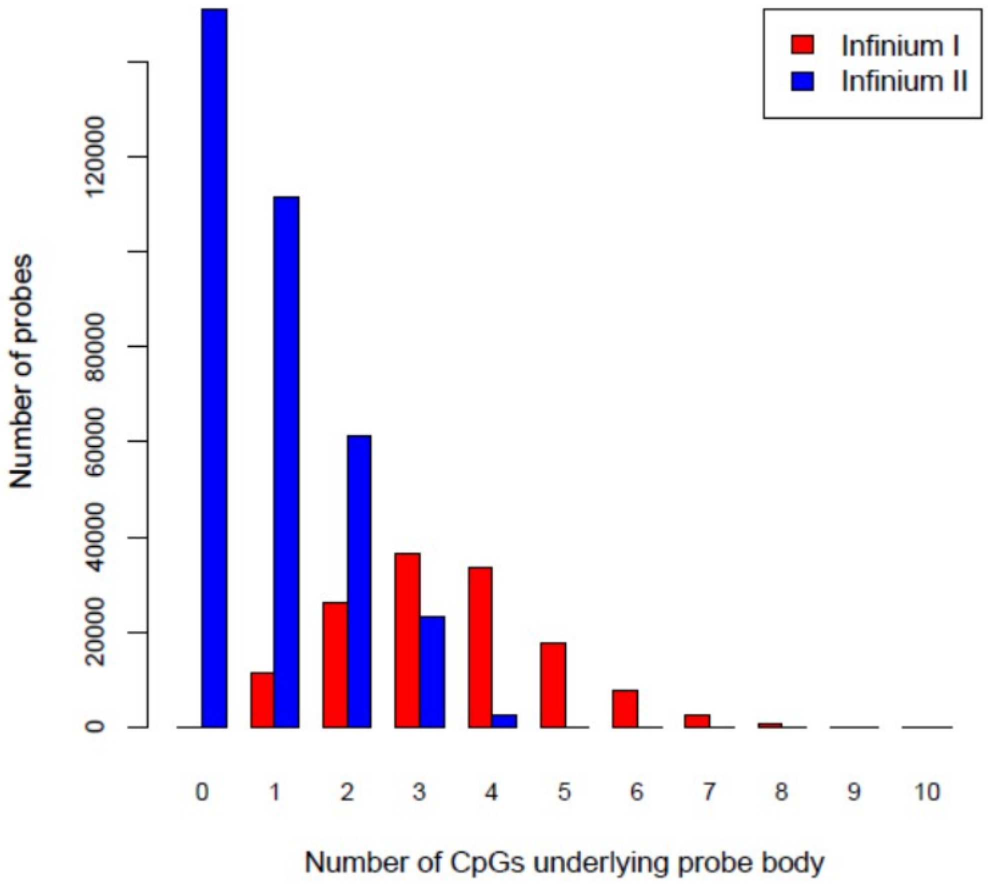
Type I & II Probes





Dedeurwaerder et al. *Epigenomics* 2011
https://www.illumina.com/documents/products/technotes/technote_hm450_data_analysis_optimization.pdf

Type I & II Probes



Raw DNA Methylation Array Data

- Raw IDAT files contained in folders whose name is the chip ID
- Red/Green signal intensities for each sample
- e.g.
  - 4305493023_R01C01_Grn.idat
  - 4305493023_R01C01_Red.idat

Quantifying Methylation – Beta Values

$$\beta = \frac{M}{M + U + \alpha}, \quad 0 \leq \beta \leq 1$$

- M and U are methylated and unmethylated signal intensities, α is an offset (usually 100) to stabilise the estimates

- Represents the proportion of chromosomes that are methylated at given site
- 0 = no methylation
- 1 = all methylation

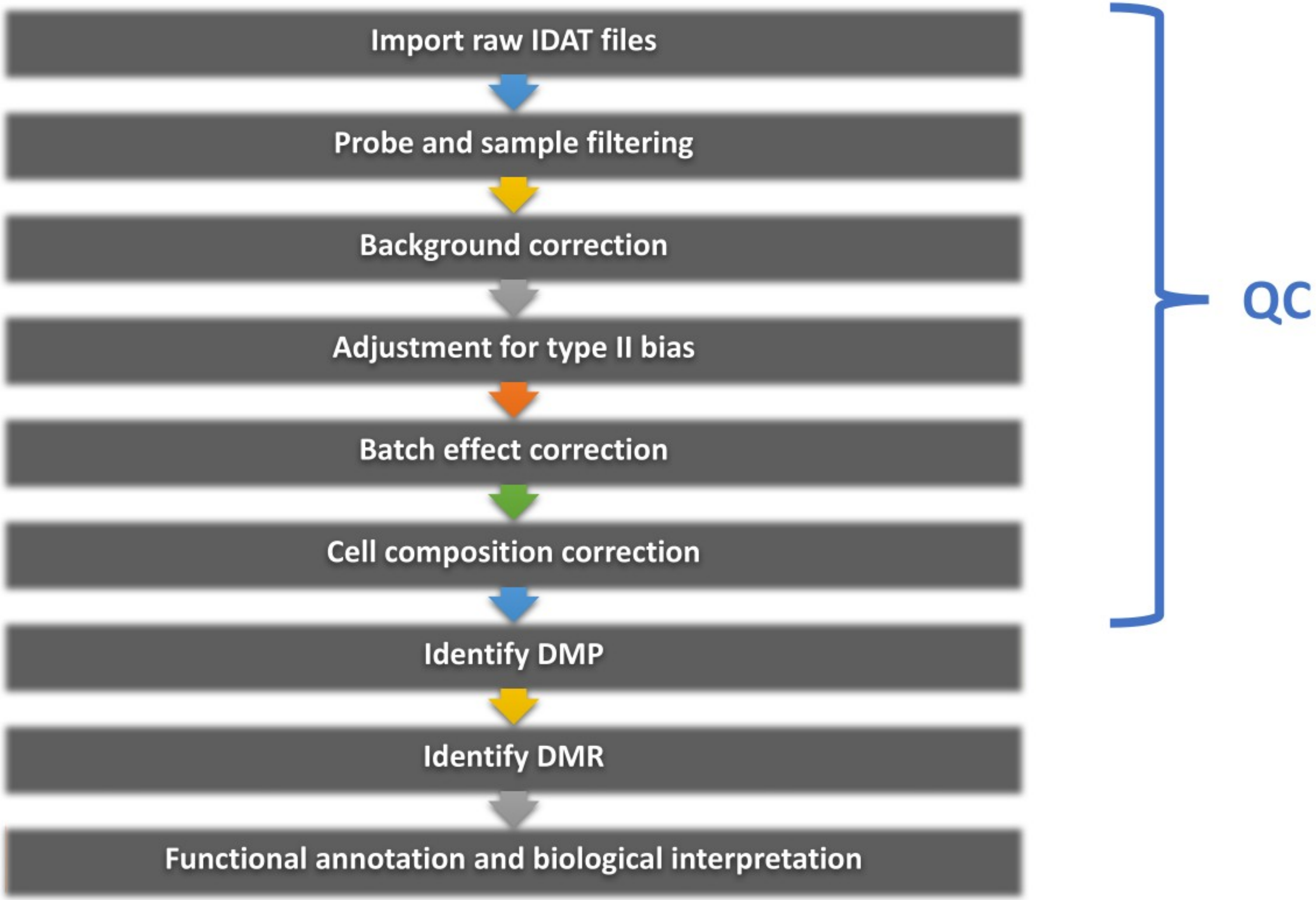
Quantifying Methylation – M Values

$$M = \log_2\left(\frac{\beta}{1-\beta}\right)$$

- Different to the methylated intensity M….
- M values are generally more robust in statistical models
- Beta values have a more intuitive biological interpretation

- Currently a move away from M values to beta values, but either acceptable

Analysis Workflow



Import raw IDAT files

Probe and sample filtering

Background correction

Adjustment for type II bias

Batch effect correction

Cell composition correction

Identify DMP

Identify DMR

Functional annotation and biological interpretation

QC

Quality Control

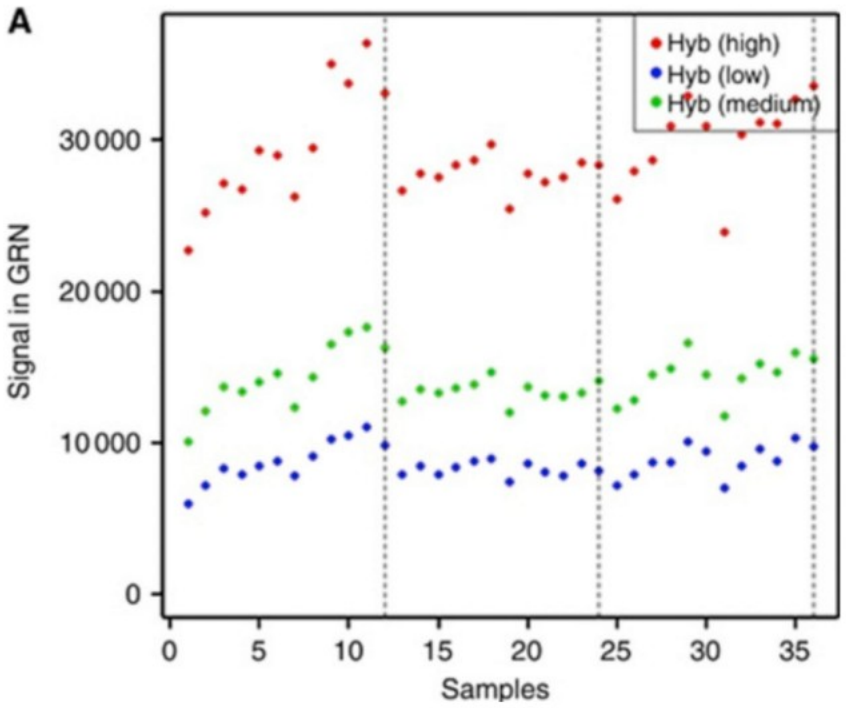
- Reduce variability introduced during the experimental process
  - E.g. Arrangement of samples on arrays, identical treatment of all samples
  - Potential experimental variation reduces the ability to detect true biological variation
  - In reality, it’s not possible to remove all experimental artifacts
- Maintain the biological variation between conditions(i.e., cases/controls)

Sample QC – Filtering on Control Probes

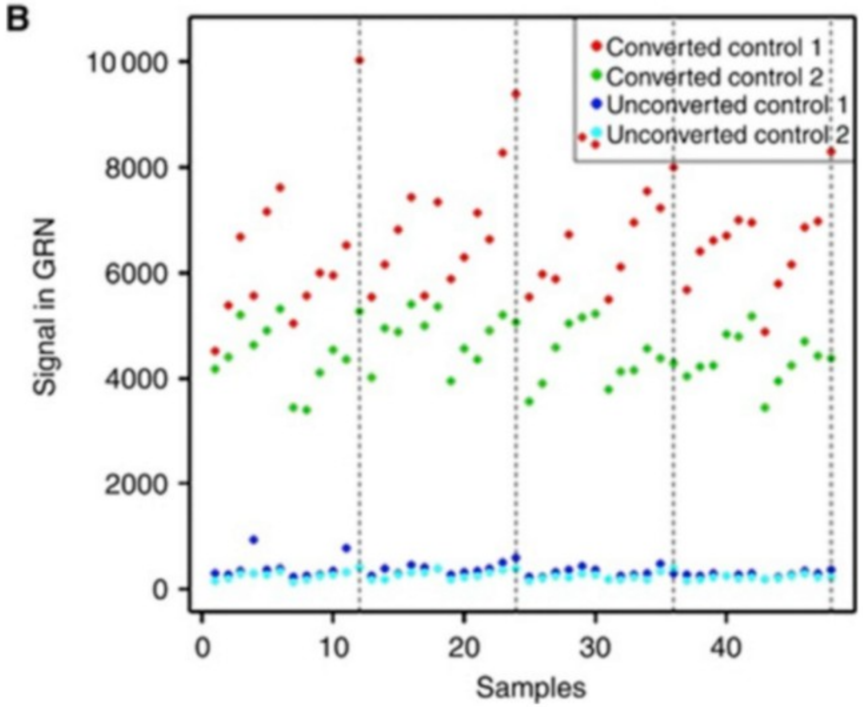
- The Illumina array contains a range of control probes to ensure quality of data:
  - STAINING CONTROLS
  - BISULFITE CONVERSION CONTROLS
  - EXTENSION CONTROLS
  - SPECIFICITY CONTROLS
  - HYBRIDIZATION CONTROLS
  - TARGET REMOVAL CONTROLS
  - NON-POLYMORPHIC CONTROLS
  - NEGATIVE CONTROLS

Sample QC – Filtering on Control Probes

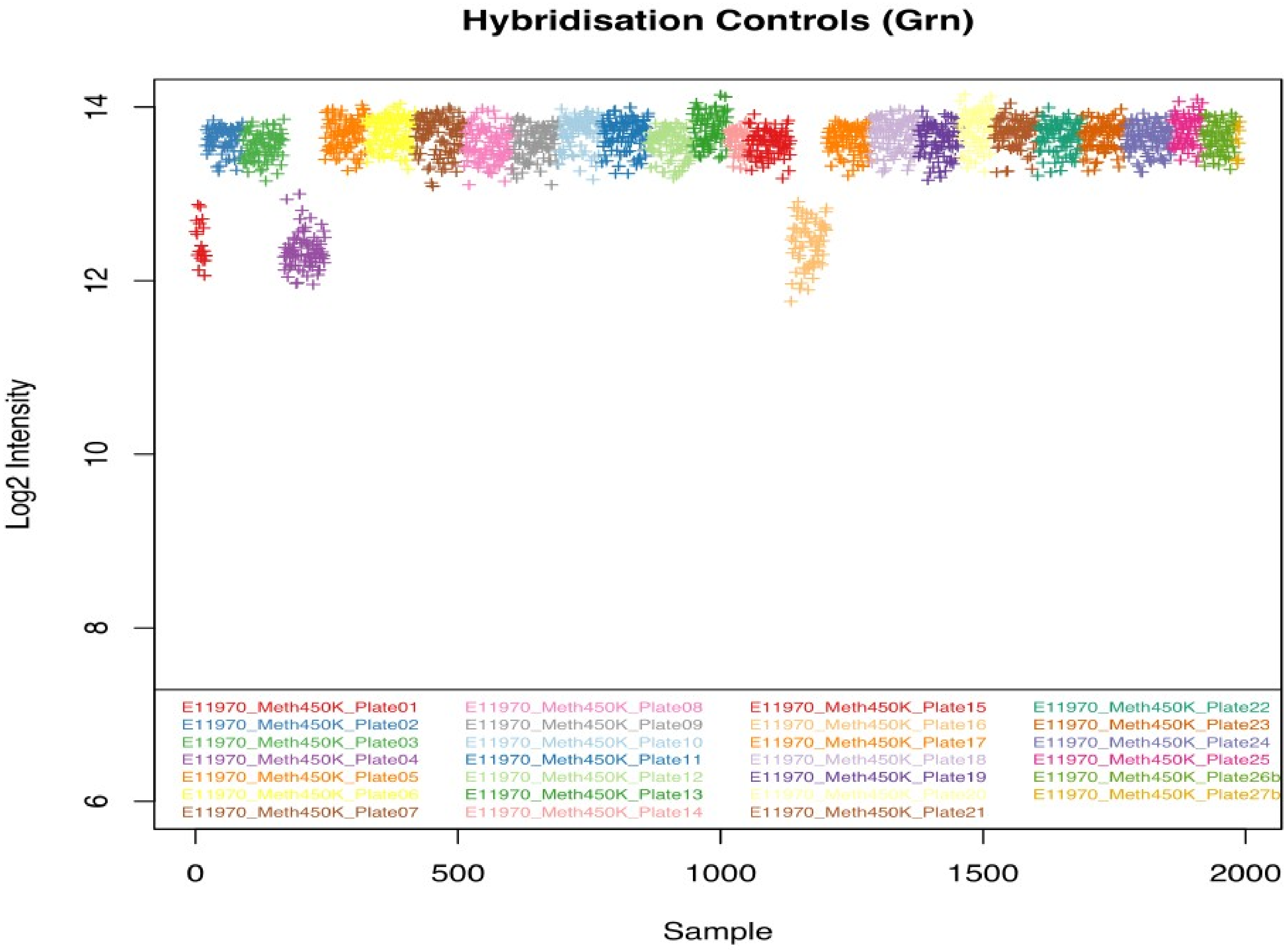
Hybridisation controls



Bisulfite conversion controls

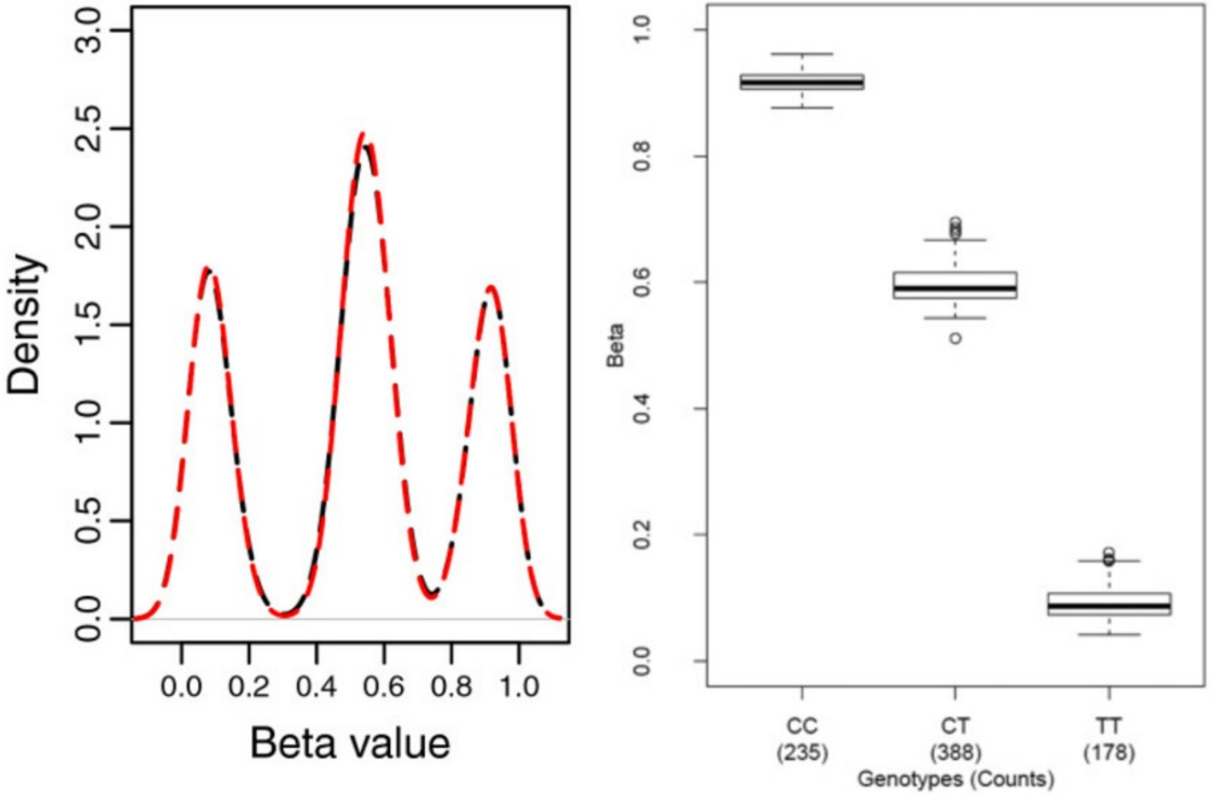


Sample QC – Filtering on Control Probes



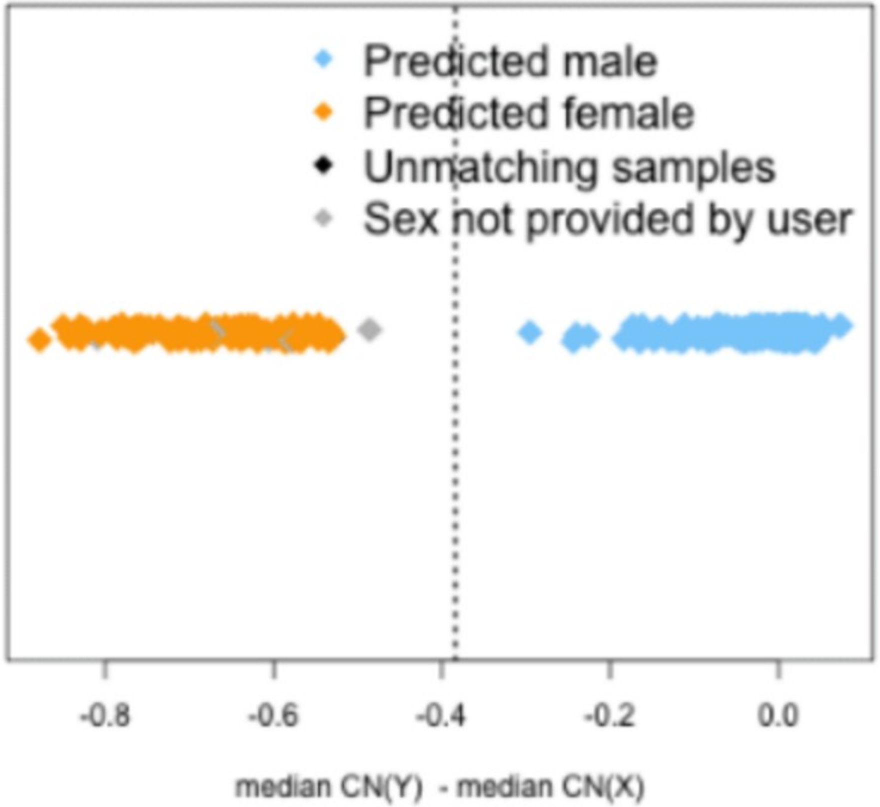
Sample QC – Filtering on Genotype

- 65 control probes on arrays whose target CpG contains a known high MAF SNP
- Can be used as a fingerprint to match to genetic data



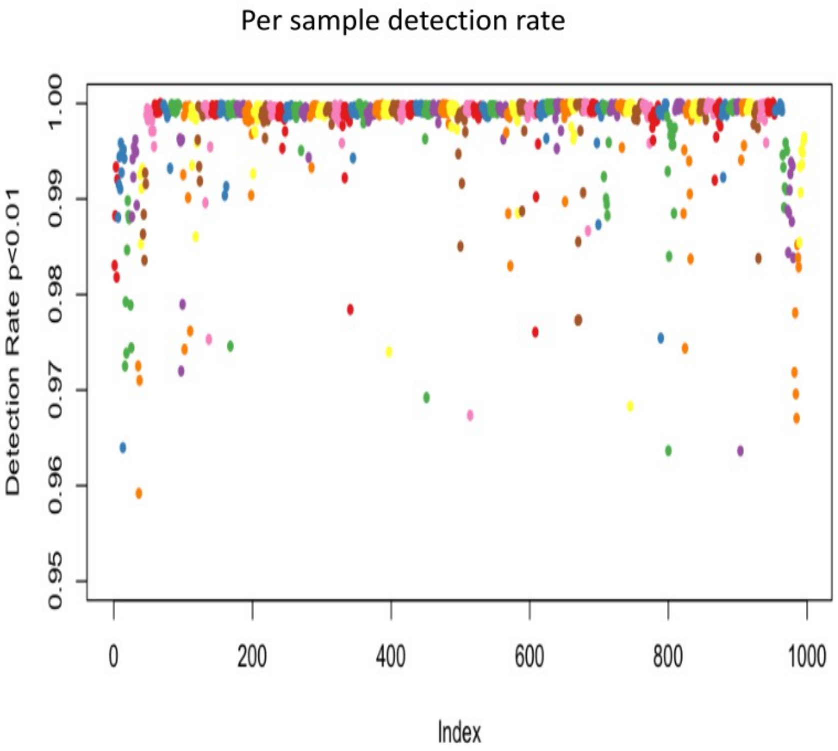
Sample QC – Filtering on Predicted Sex

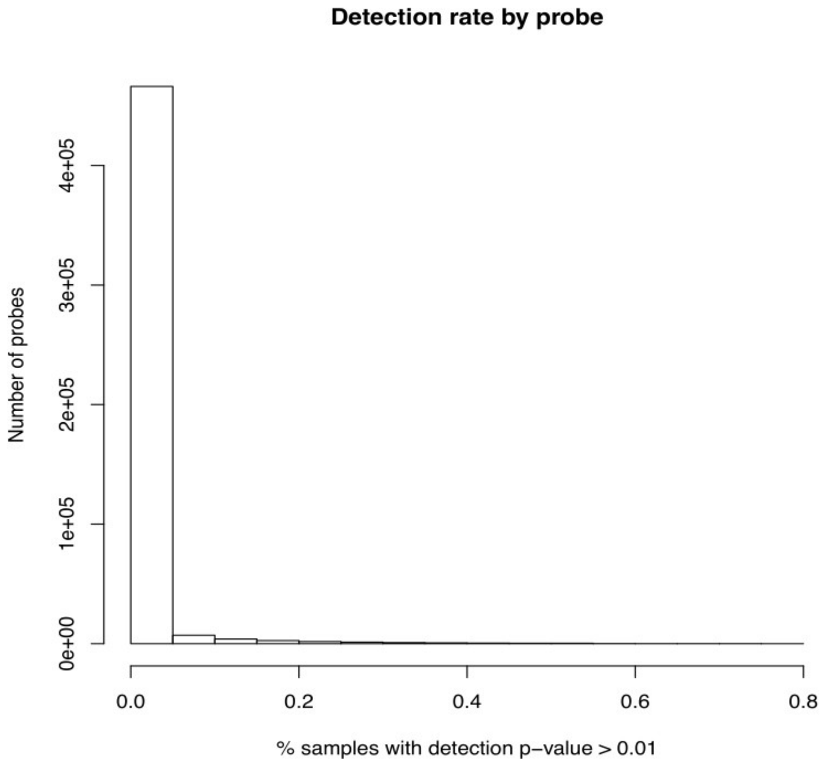
- Females have inactive-X chromosome that is heavily methylated, and have no Y chromosome



Probe and Sample QC – Detection P-value

- Compares the total DNA signal (Methylated + Unmethylated) for each probe to the background signal estimated using negative control probes.
- The detection P-value





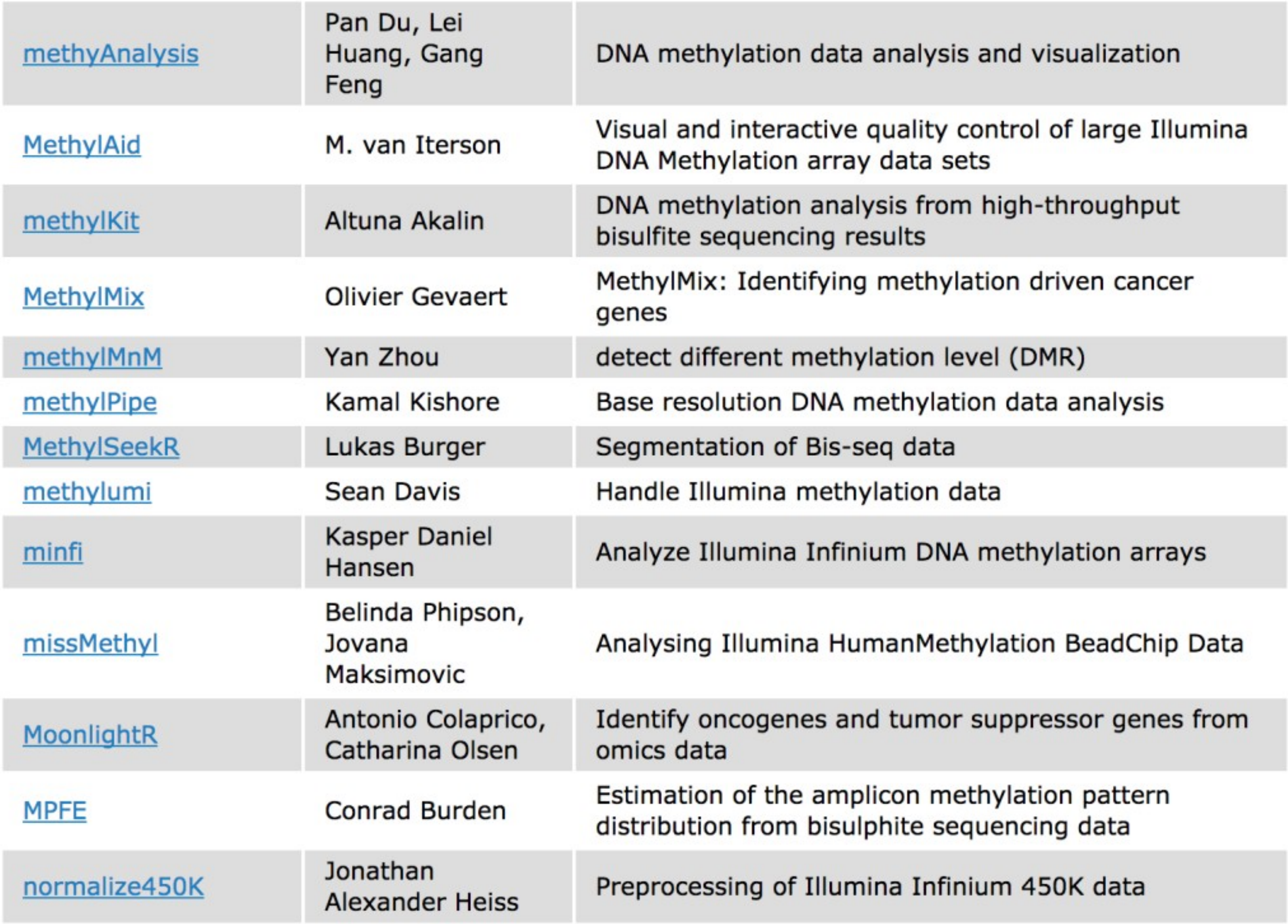
QC - Other Considerations

- Bead count – each methylation site is measured using multiple beads; remove if too few (< 3) beads for site
-
- Cross-reactive probes – some probes bind at multiple sites in the genome
    - Partial sequence overlaps
- Probes with SNPs at or near target site – reflect SNP differences and not (only) DNA methylation
    - Most are rare SNPs
    - Filter afterward?

Normalisation

- **Goal: Reduce non-biological variation**
- If the statistical design is bad, your data will be bad…
- Various statistical methods to reduce technical variation
  - Within array normalisation – correction for intensity-related dye bias
  - Between array normalisation – removing technical artifacts between samples on different arrays
- No consensus on the best approach!

R Packages for Methylation QC/normalisation



methyAnalysis	Pan Du, Lei Huang, Gang Feng	DNA methylation data analysis and visualization
MethylAid	M. van Iterson	Visual and interactive quality control of large Illumina DNA Methylation array data sets
methylKit	Altuna Akalin	DNA methylation analysis from high-throughput bisulfite sequencing results
MethylMix	Olivier Gevaert	MethylMix: Identifying methylation driven cancer genes
methylMnM	Yan Zhou	detect different methylation level (DMR)
methylPipe	Kamal Kishore	Base resolution DNA methylation data analysis
MethylSeekR	Lukas Burger	Segmentation of Bis-seq data
methylumi	Sean Davis	Handle Illumina methylation data
minfi	Kasper Daniel Hansen	Analyze Illumina Infinium DNA methylation arrays
missMethyl	Belinda Phipson, Jovana Maksimovic	Analysing Illumina HumanMethylation BeadChip Data
MoonlightR	Antonio Colaprico, Catharina Olsen	Identify oncogenes and tumor suppressor genes from omics data
MPFE	Conrad Burden	Estimation of the amplicon methylation pattern distribution from bisulphite sequencing data
normalize450K	Jonathan Alexander Heiss	Preprocessing of Illumina Infinium 450K data

Normalisation – Background Correction

- All measurements on the array are made with some noise
- It is impossible to get a “zero” measurement from the array
- Background correction attempts to remove this noise

- Often use negative control probes to remove this noise
  - Subtract 5% percentile of the negative controls from each colour channel
  - (GenomeStudio Methylation Module)
  - Subtract median intensity value of control probes (R package lumi)
- Many other methods…

- Likely to happen when reading in idat files by default

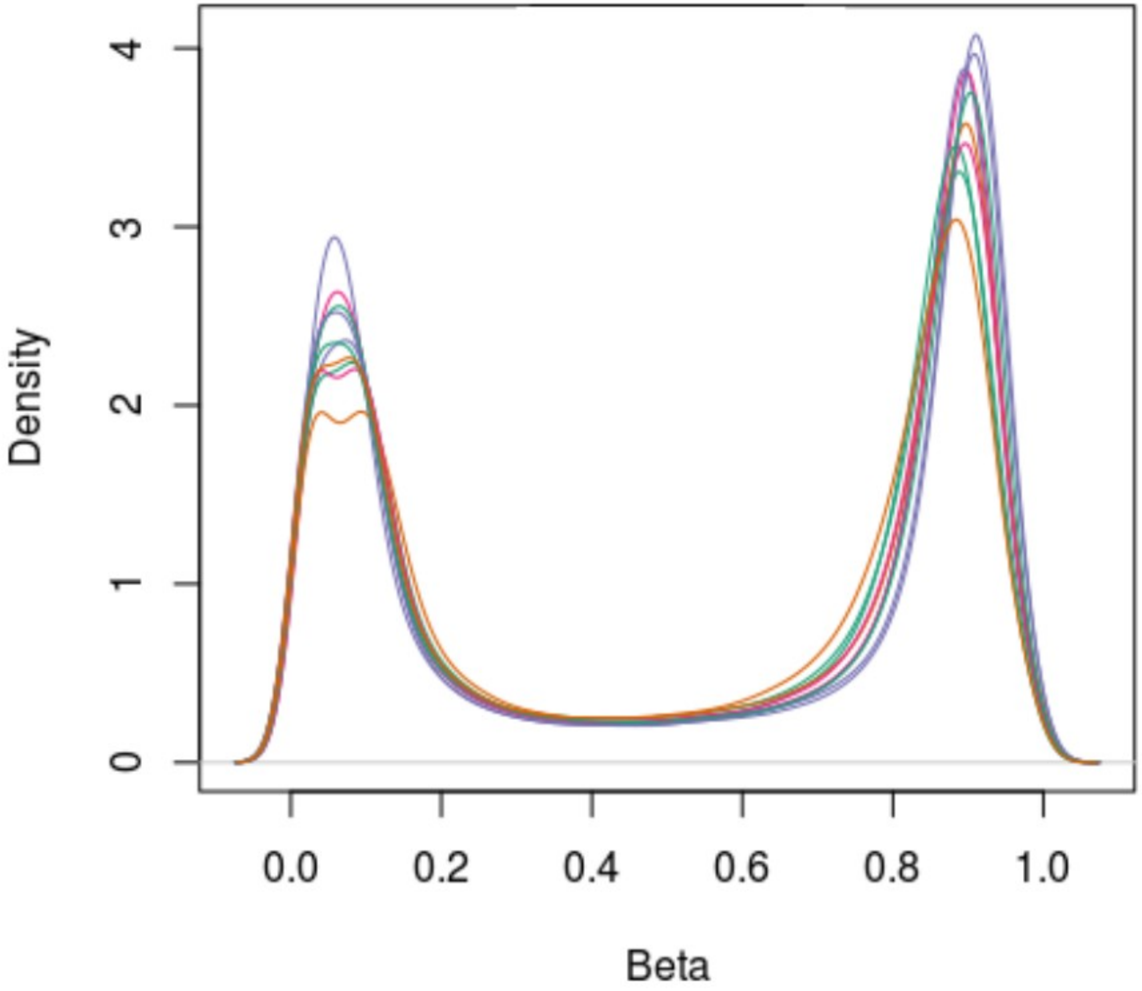
Normalisation – Colour Bias

- The two colour channels are know to perform differently

- Usually higher overall intensities on the red channel that the green channel

- Large number of methods…
- Illumina GenomeStudio:
  - Takes the average intensity of the internal normalisation control for that colour
  - Divides all intensity values by that average
  - Rescales data to the first sample on the array
- R methylumi, ASMN: scale to array with least difference in average dye intensity or average across all samples
- ….

Normalisation – Across Array



Normalisation – Across Array

- **Quantile Normalisation**
- Widely used in gene-expression studies
- Normalises data to average/median of all observations
- Makes all distributions identical

- Is this suitable for DNA methylation data?
  - Evidence for different genome-wide average methylation across peop
  - Case/control studies can have vastly different methylation profiles

Normalisation – Across Array

- Functional normalisation

- Uses quantile normalisation of control probes only
- Other array probes are scaled relative to control probes with surrounding intensities
- Fortin et al., Genome Biology 2014, 15:503

- We will use this method in the practical

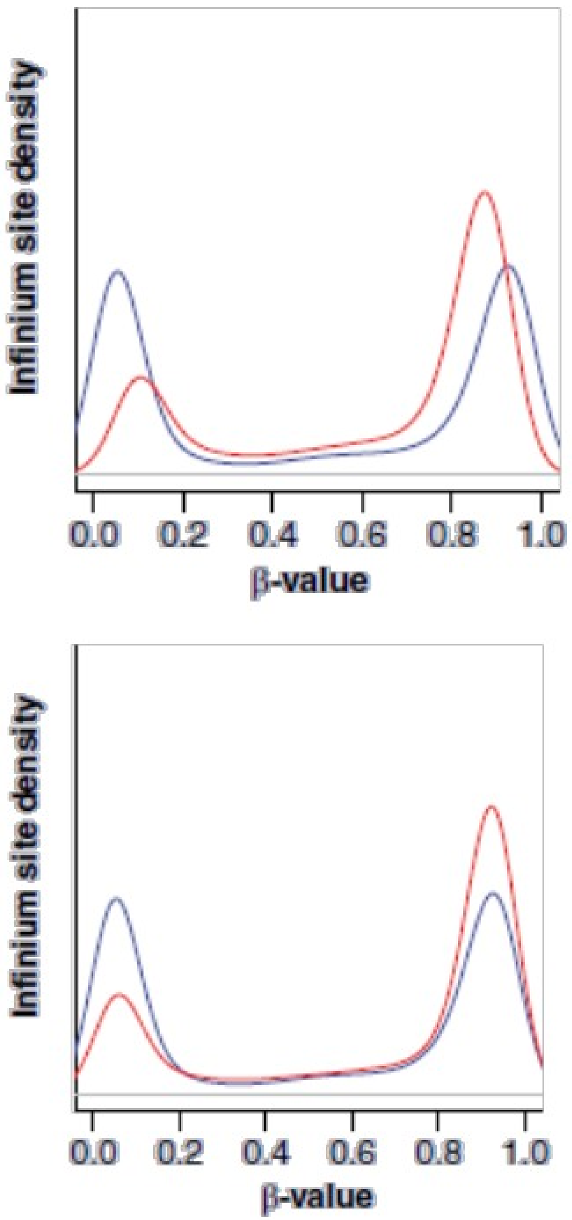
Normalisation – Probe Bias

- Some measurement bias is shown between Type I and II probes
  - May be expected giving different biology of probes…
  - Type II tend to be more variable than Type I
- This causes a problem if probes are to be ranked/combined in an analysis
  - Clustering
  - Regional approaches (“bumphunting”)
  - …
- This is “not” an issue for single probe analyses

Normalisation – Probe Bias

- Peak Based Correction

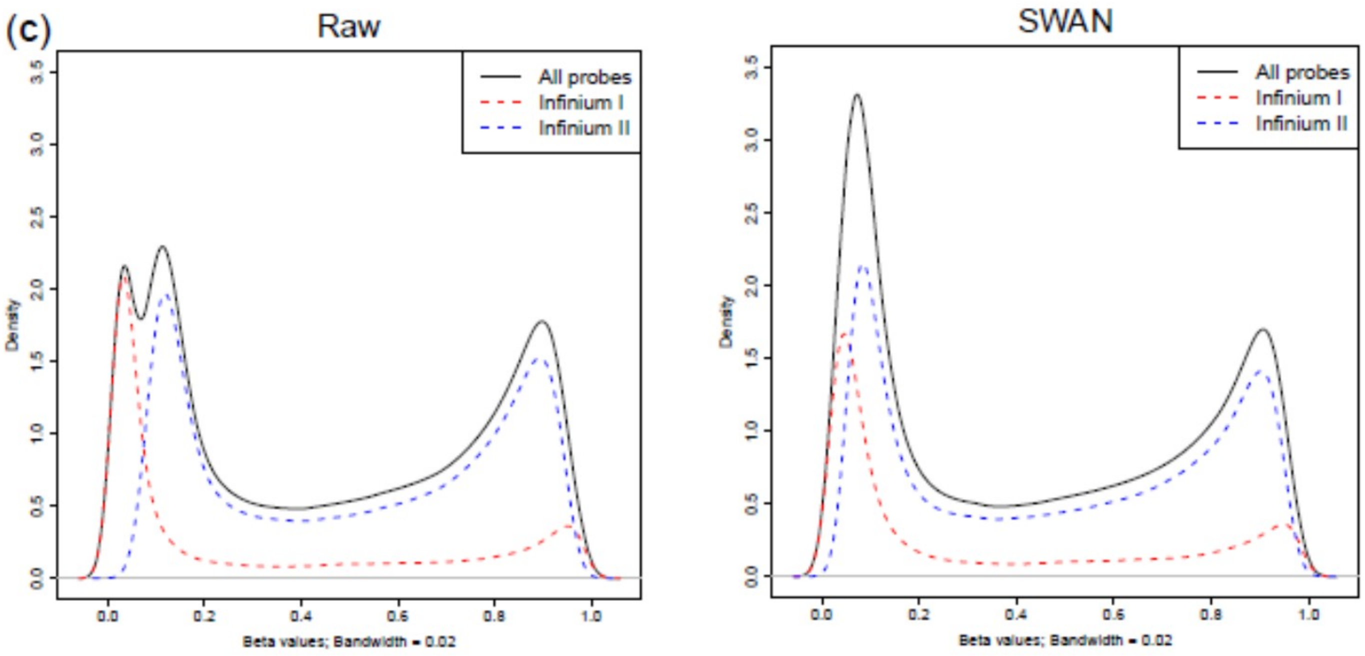
- Uses peak summits to correct β values



Normalisation – Probe Bias

- Beta MIxture Quantile Dilation (BMIQ)
  - Fits a mixture distribution to data

- Subset Within-Array Normalization (SWAN)
  - Normalise based on the number of CpG sites covered by each probe



Batch effects

- Technical artifacts (e.g. laboratory conditions, experiment time, reagent, array batch, sample plate, position on array) that are not associated with the underlying biology
- Batch effects can affect different probes in different ways
- Minimise batch effect through careful study design (e.g. randomising samples across run times, running technical replicates etc)
- Two types of methods
    - when the sources of batch effect are known
    - when batch effects are unknown