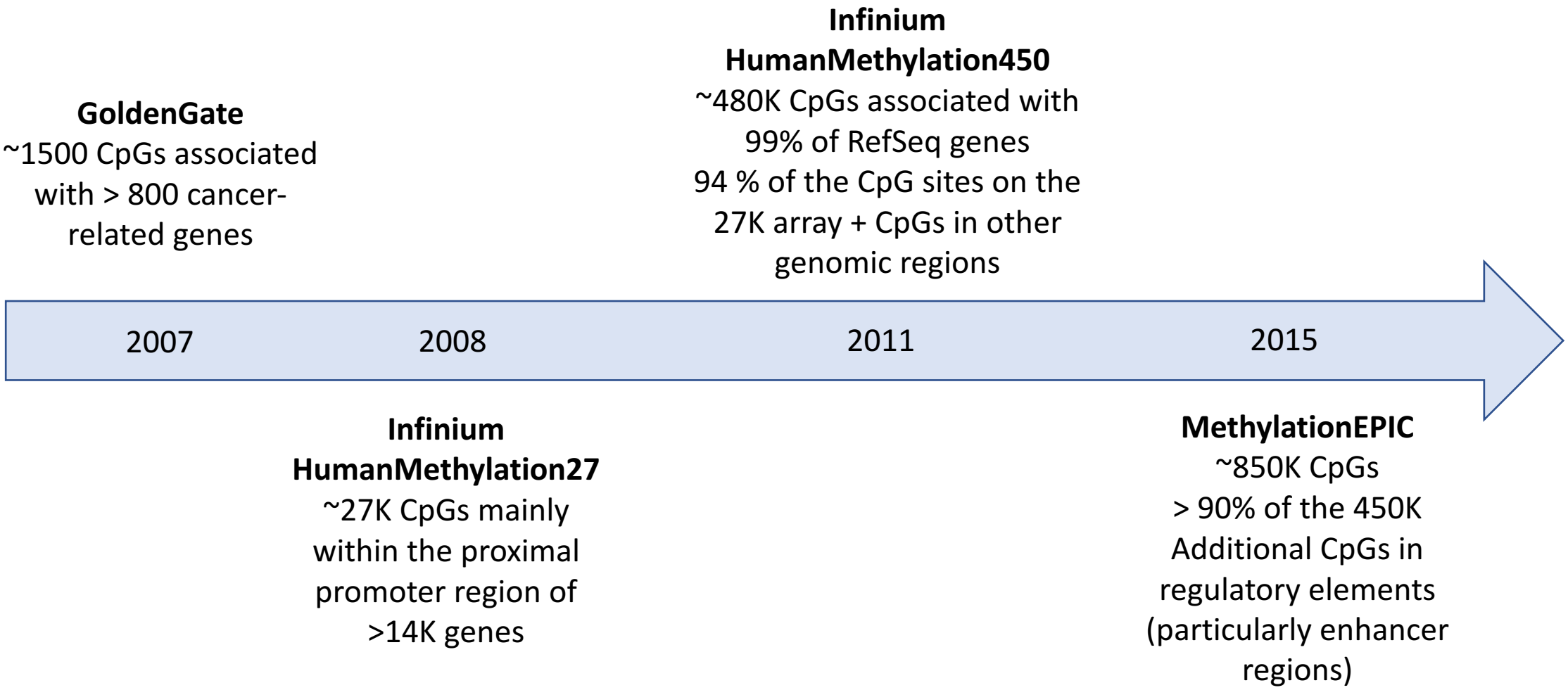
QC & Analysis of Methylation Chip Data

Allan McRae & Sonia Shah

Outline for Session 1 Lecture (9 – 10.30am)

- Overview of methylation array technology
- Quantifying methylation levels at a single CpG site
- Quality Control
  - Control probes
  - Sample and probe filtering
  - Cross-reactive probes
  - SNP probes

Illumina methylation arrays



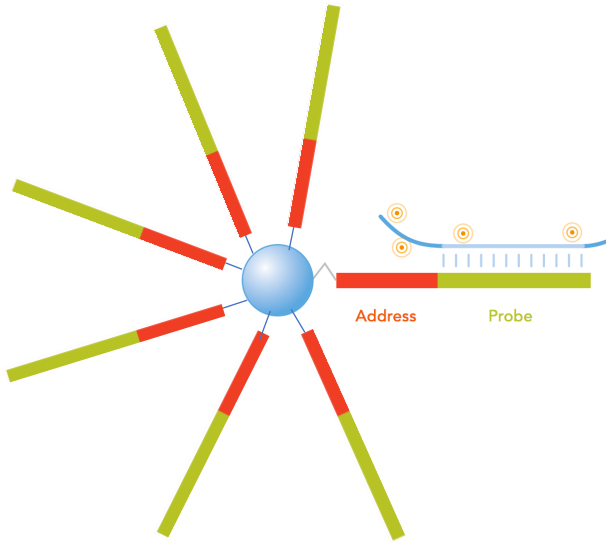
Illumina 450K array

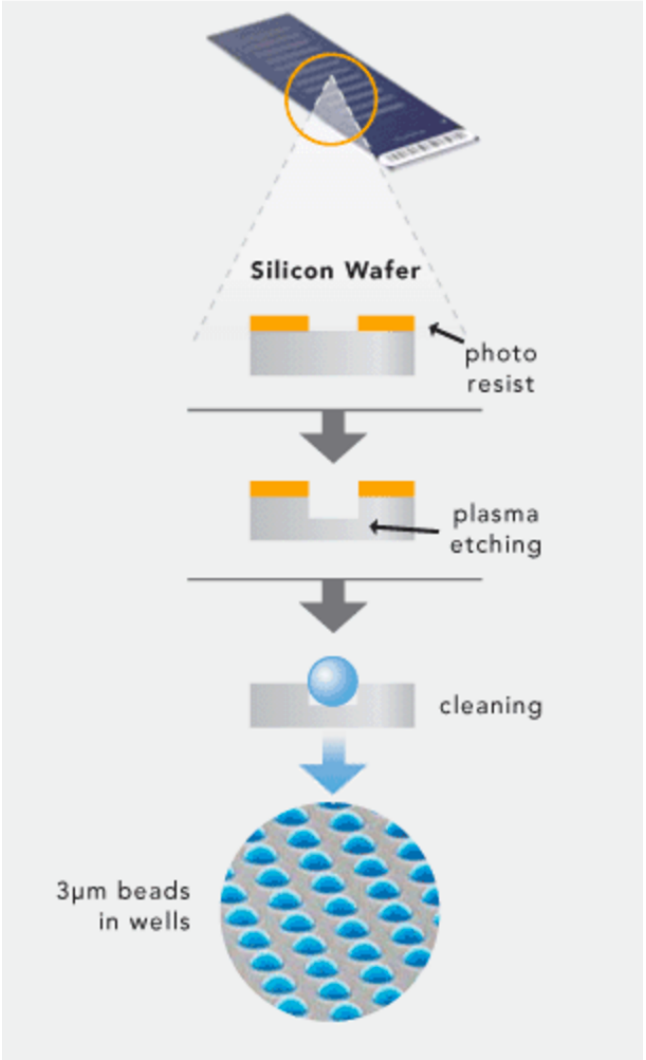
HumanMethylation450 array content.

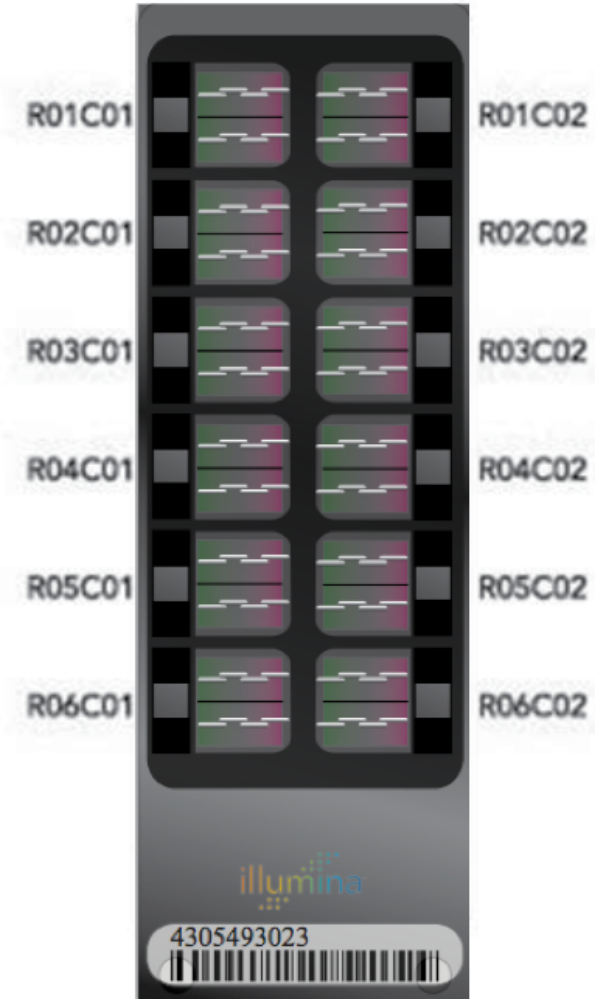
Feature type	Included on array
Total number of sites	485,577
RefSeq genes	21,231 (99%)
CpG islands	26,658 (96%)
CpG island shores (0–2 kb from CGI)	26,249 (92%)
CpG island shelves (2–4 kb from CGI)	24,018 (86%)
HMM islands[a]	62,600
FANTOM 4 promoters (High CpG content)[a]	9426
FANTOM 4 promoters (Low CpG content)[a]	2328
Differentially methylated regions (DMRs)[a]	16,232
Informatically-predicted enhancers[a]	80,538
DNAse hypersensitive sites	59,916
Ensemble regulatory features[a]	47,257
Loci in MHC region	12,334
HumanMethylation27 loci	25,978
Non-CpG loci	3091

Bibkova et al Genomics 2011

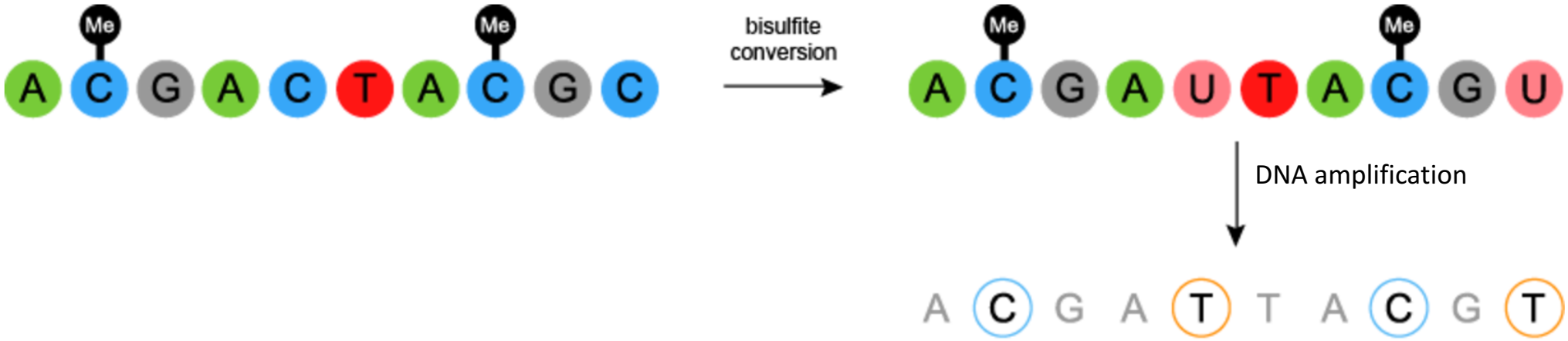
Beadchip technology



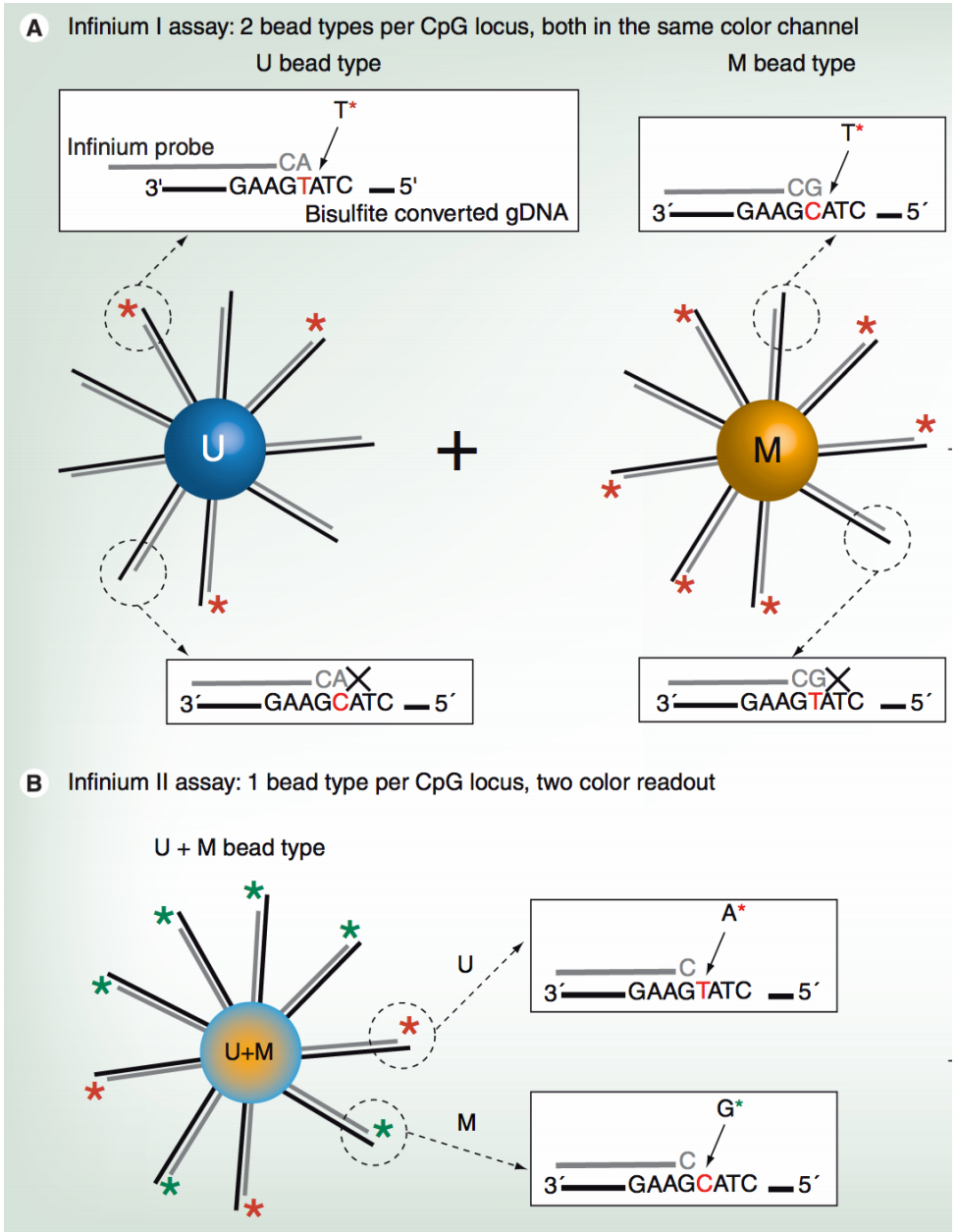


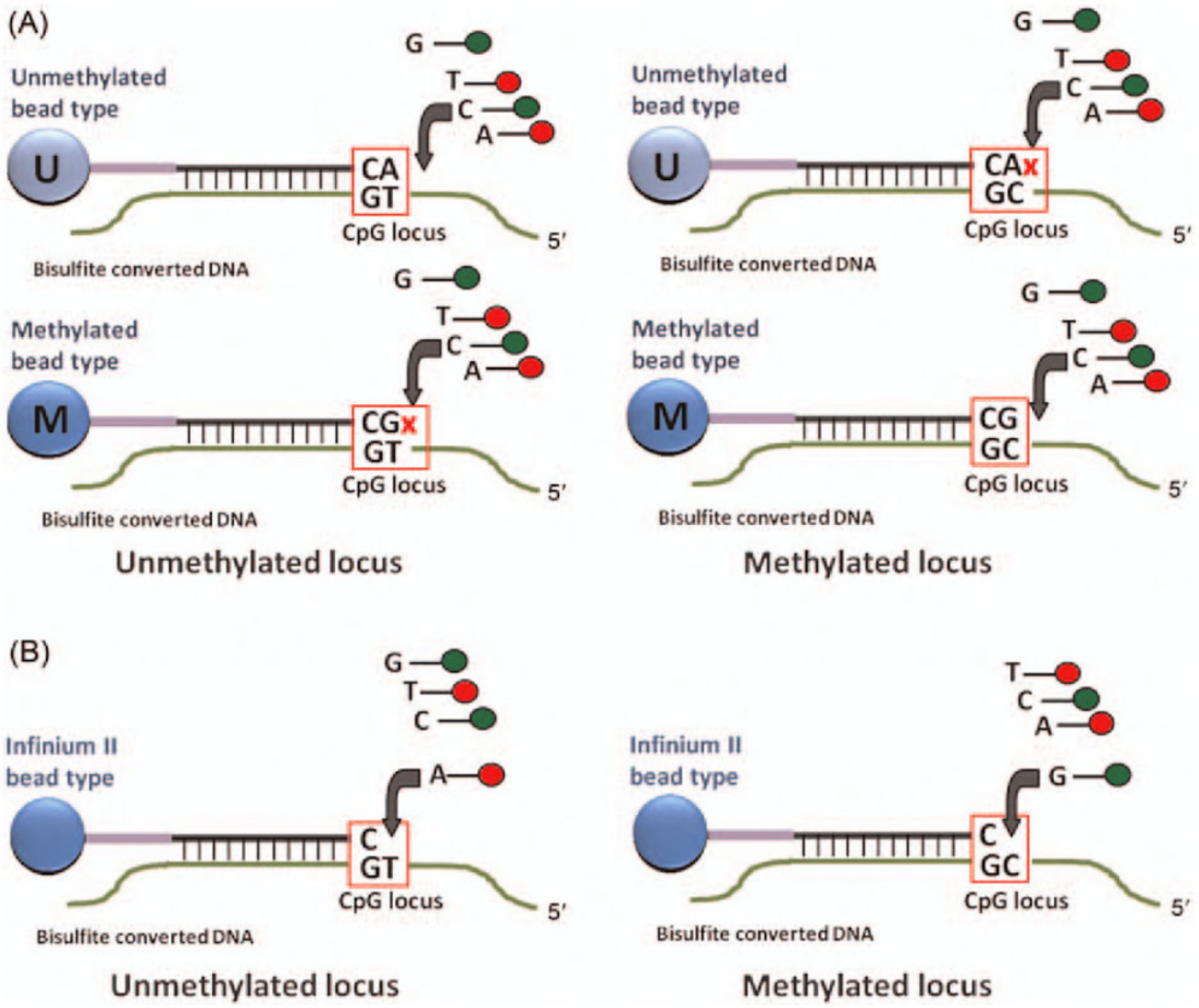


Bisulfite conversion



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

Type I probes:

- Assumes methylation is regionally correlated within a 50bp span i.e. if the target CpG is methylated, so will the nearby CpGs.

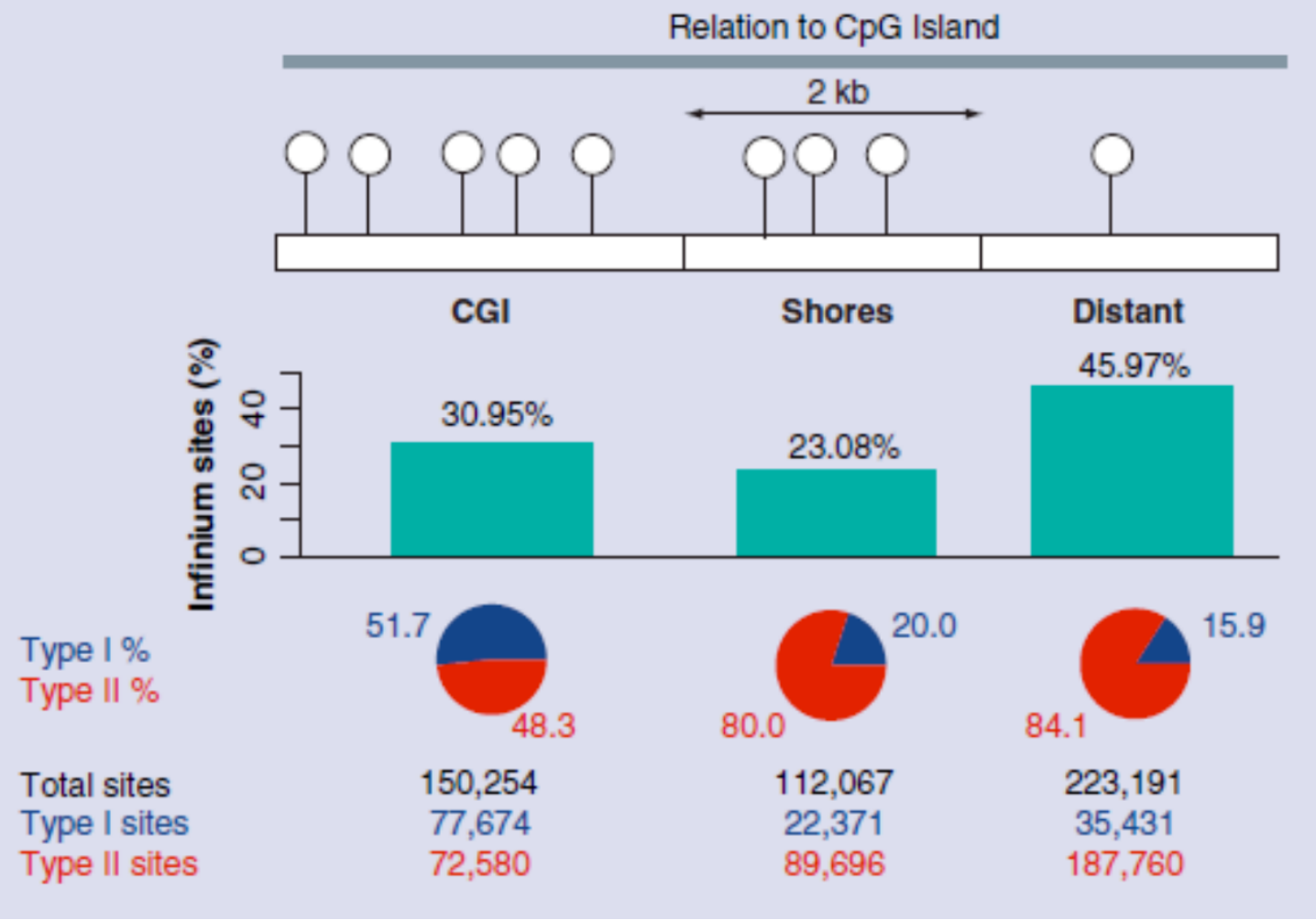
- Study using bisulfute sequencing on chr 6, 20 and 22 (Eckhardt et al):
  - >90% of CpG sites within 50 bases had the same methylation status

Type II probes:

- Targets less CpG-rich regions
- Can have up to 3 CpG sites underlying the probe without compromising data quality

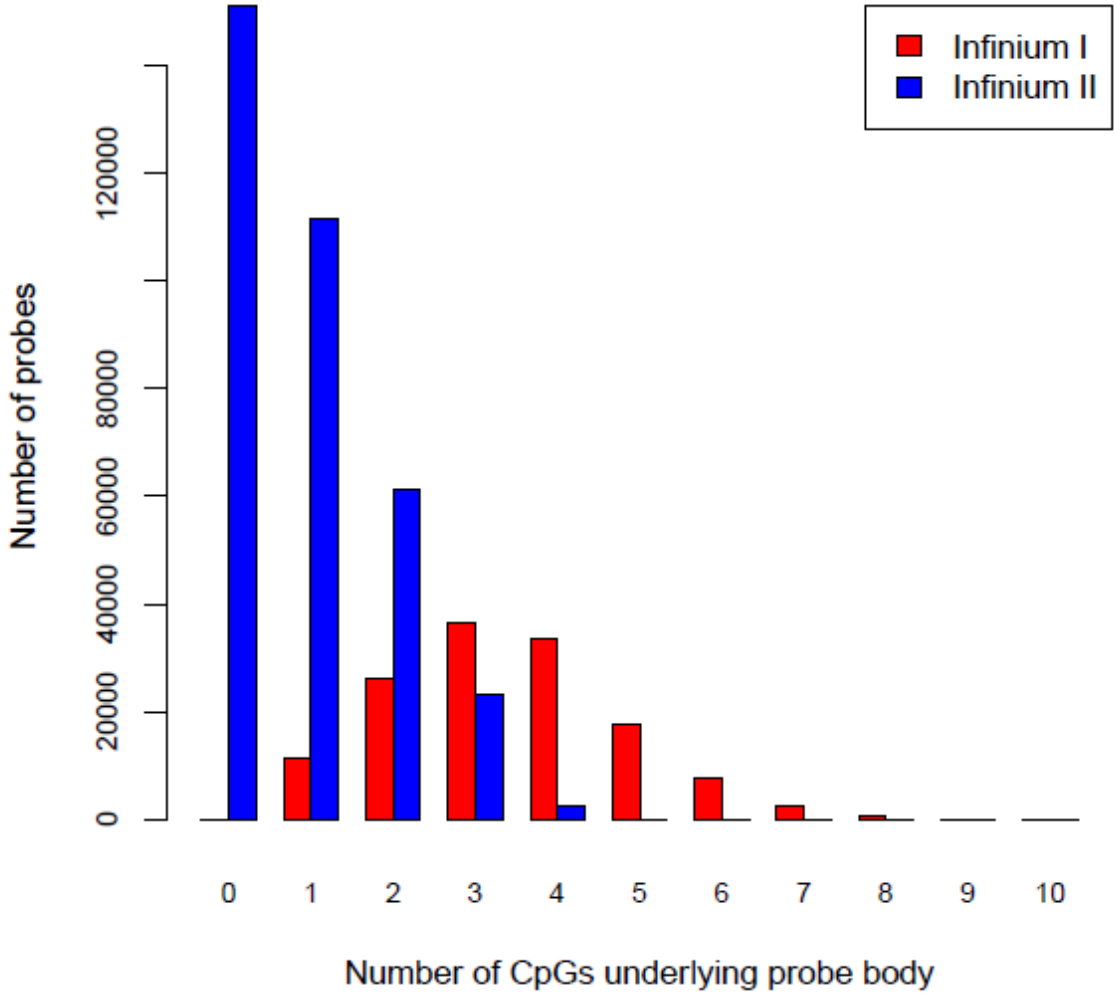
- For both probe types underlying polymorphic sites may affect hybridisation of genomic sequence to the probe

Type I/II Probes



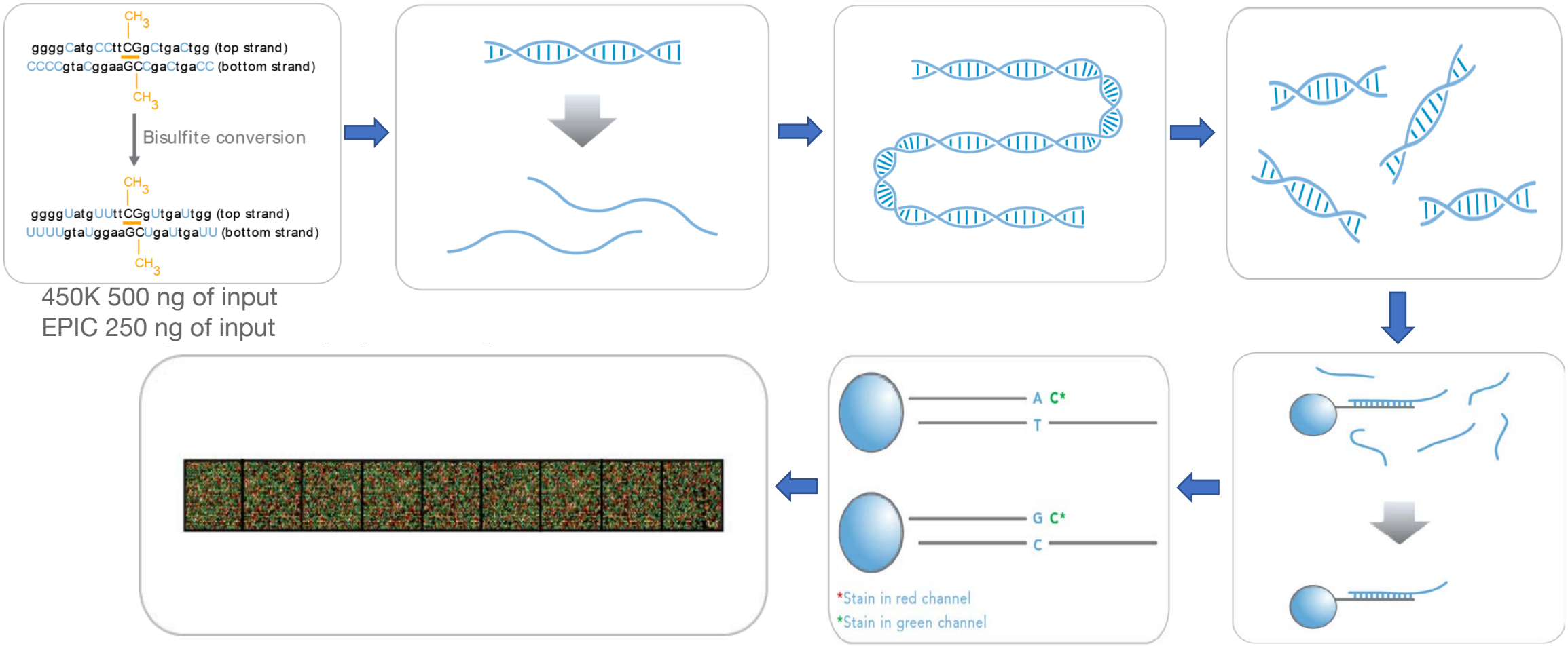
Dedeurwaerder et al. *Epigenomics* 2011

Type I/II Probes



Maksimovic et al. Genome Biology 2012

Methylation Assay



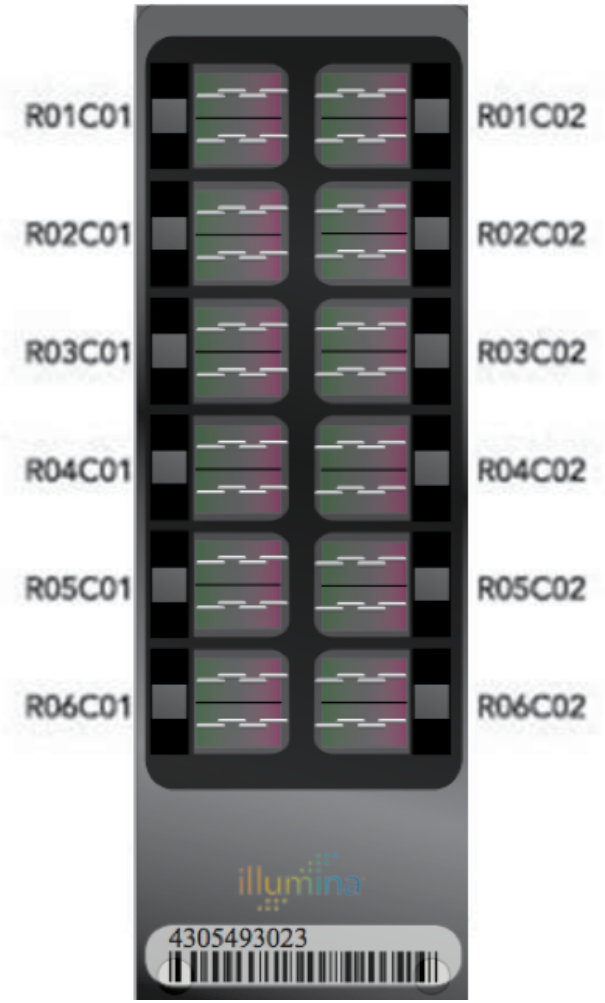
http://support.illumina.com/array/array_kits/infinium_humanmethylation450_beadchip_kit/documentation.html

IDAT Files (Raw Data)

- Raw IDAT files contained in folders whose name is the chip ID
- Red/Green signal intensities for each sample
- Default IDAT file name format:
  4305493023_R01C01_Grn.idat

  4305493023_R01C01_Red.idat

  chip.barcode_chip.position_channel
- Sample annotation also provided

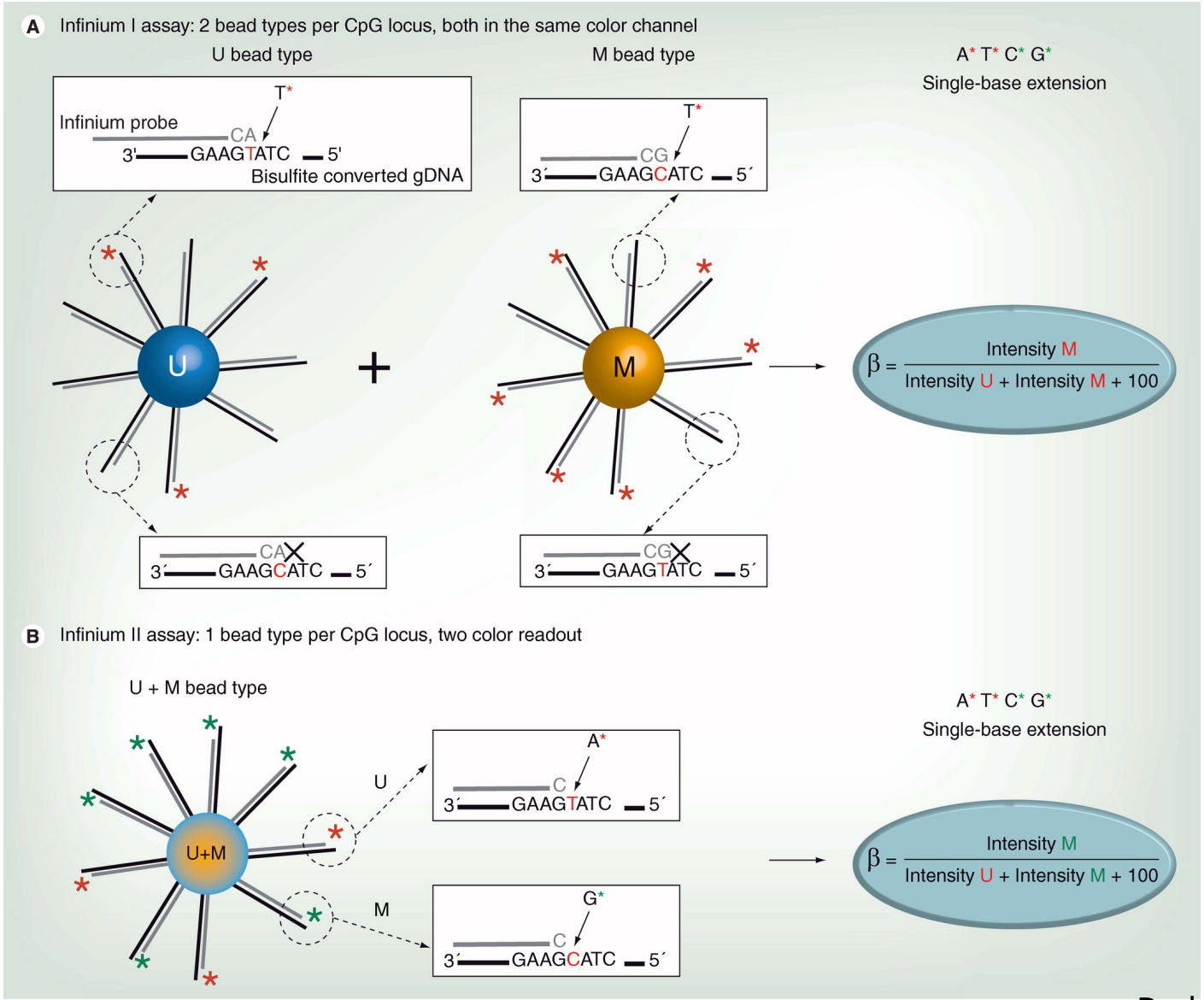


Quantifying methylation – beta values

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

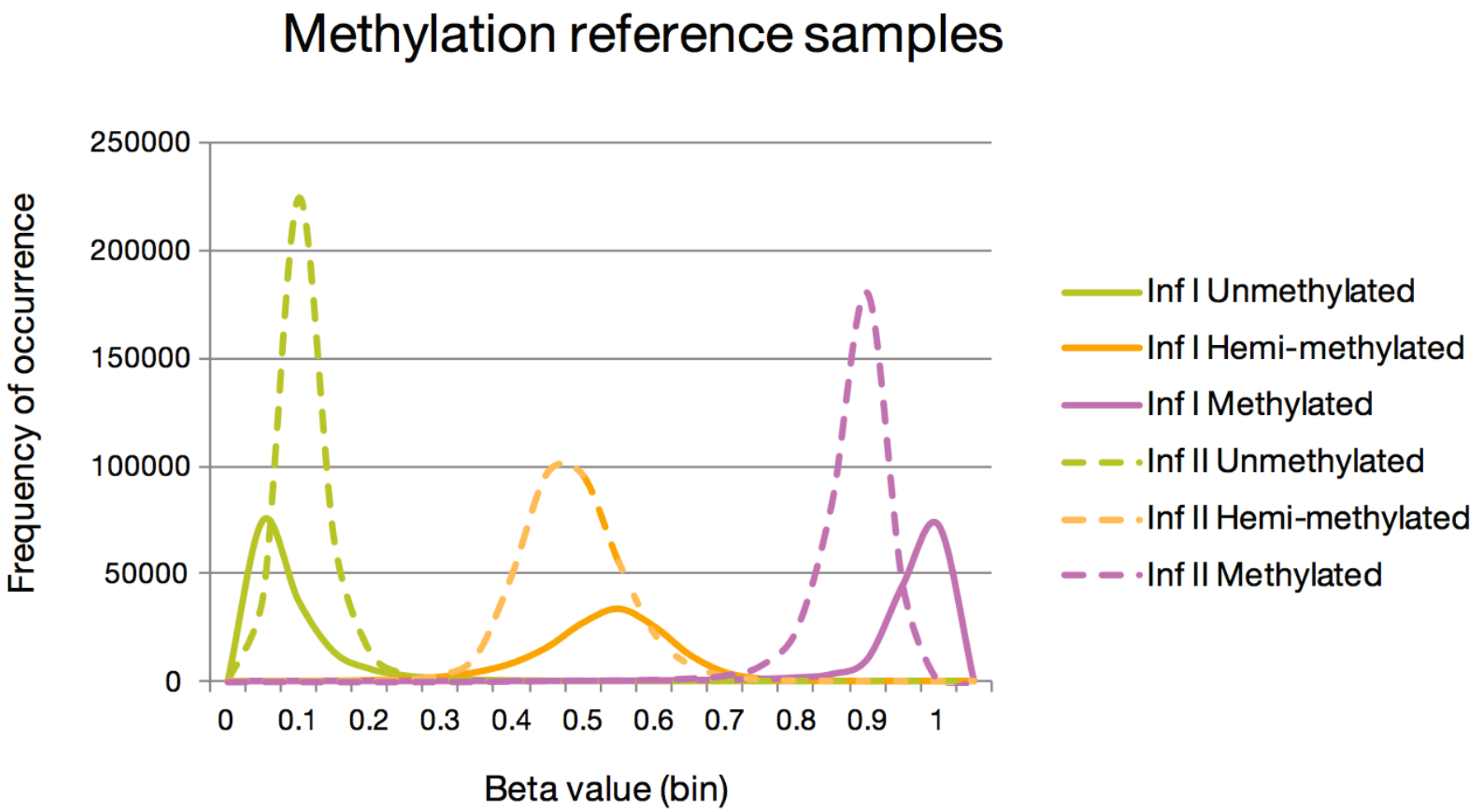
- M and U are the methylated and unmethylated signal intensities
- α is an offset (usually 100) to stabilise the beta-values
- Beta-value of 0 - all copies of the CpG site in sample were unmethylated
- Beta-value of 1 - all copies of the CpG site in the sample were methylated

Quantifying methylation – beta values



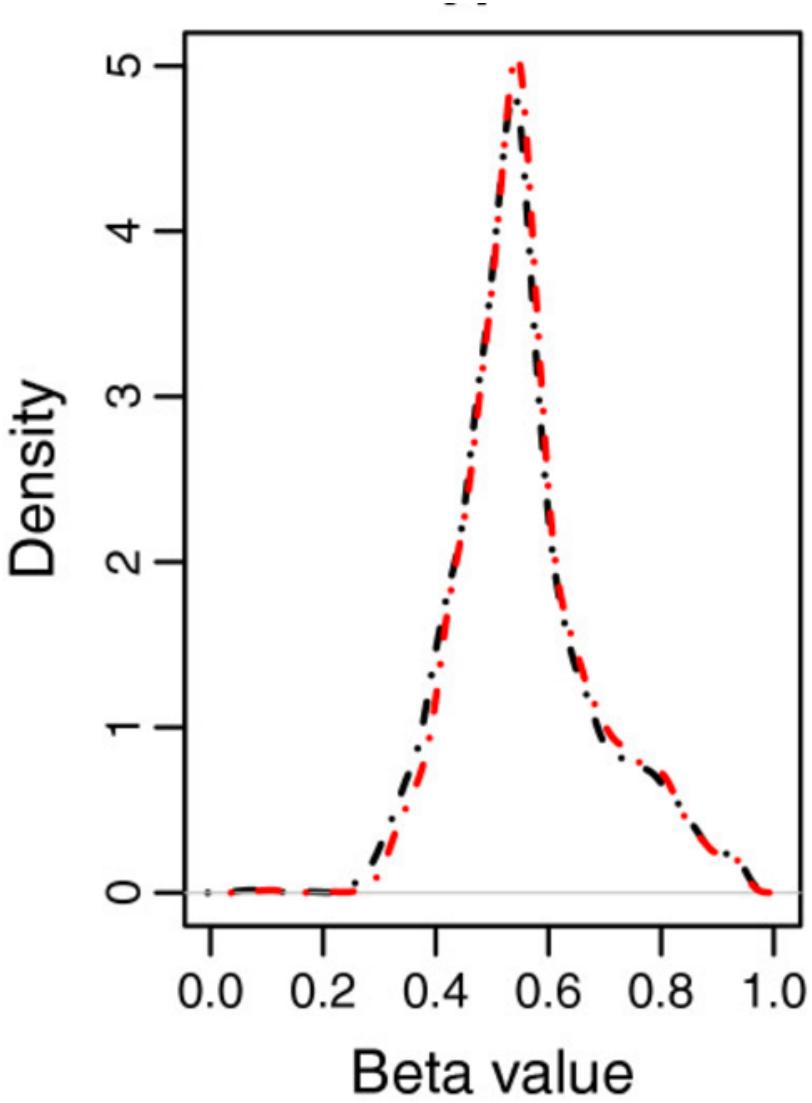
Dedeurwaerder et al. *Epigenomics* 2011

Beta distribution - Type I/II Probes



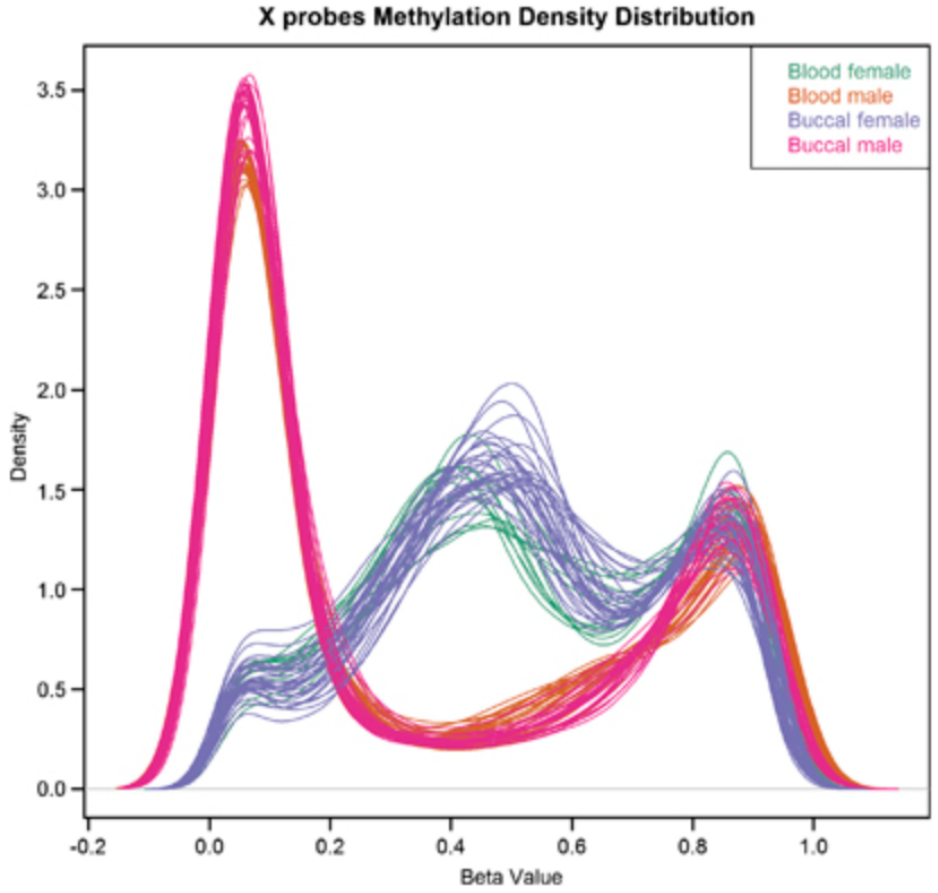
Beta distribution – imprinted alleles

- 237 probes on the array that lie within imprinted genes
- Expected *β* value of 0.5 because they are uniparentally methylated in most tissues



Beta distribution – chrX probes

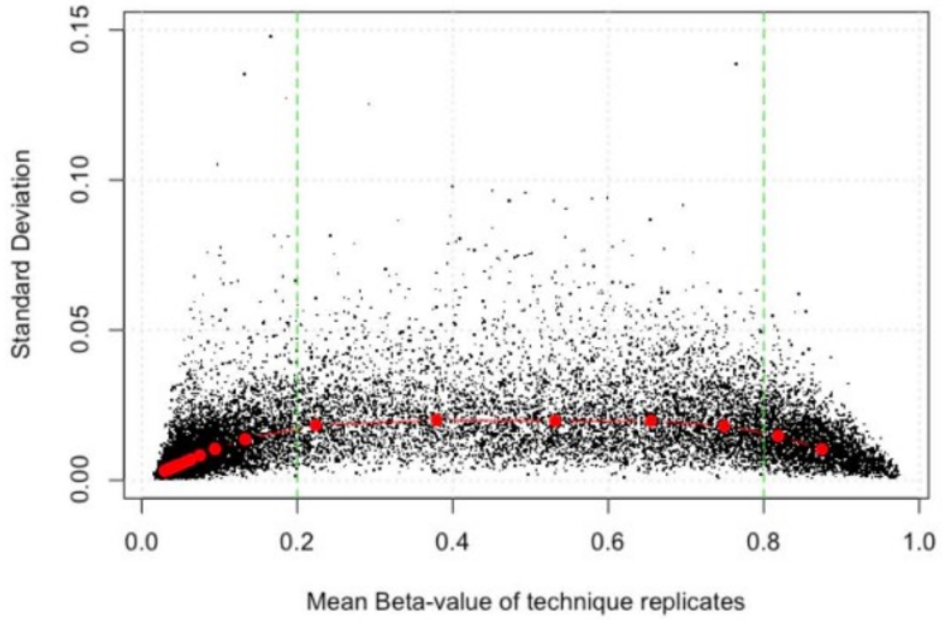
X-chromosome probes show distinctly sex-specific DNA methylation patterns irrespective of the tissue type or time point



Joo et al Nature 2014

Properties of beta distribution

- Beta-value method has severe heteroscedasticity for highly methylated or unmethylated CpG sites

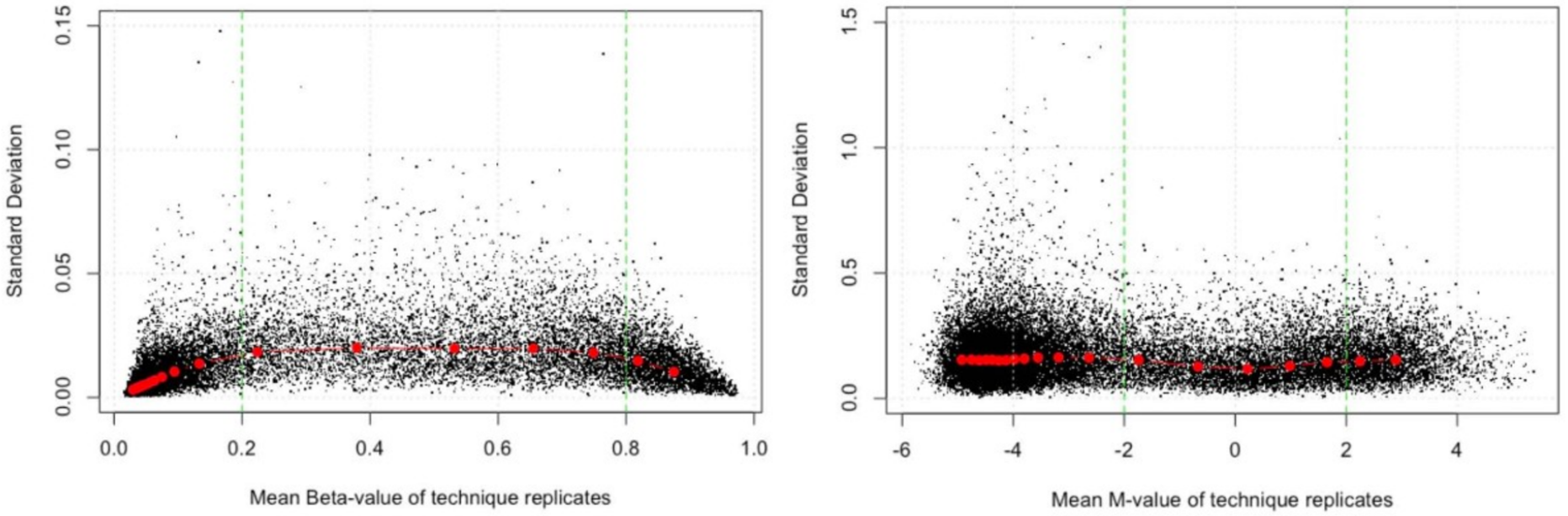


- Beta-value has a bounded range

Quantifying methylation – M-values

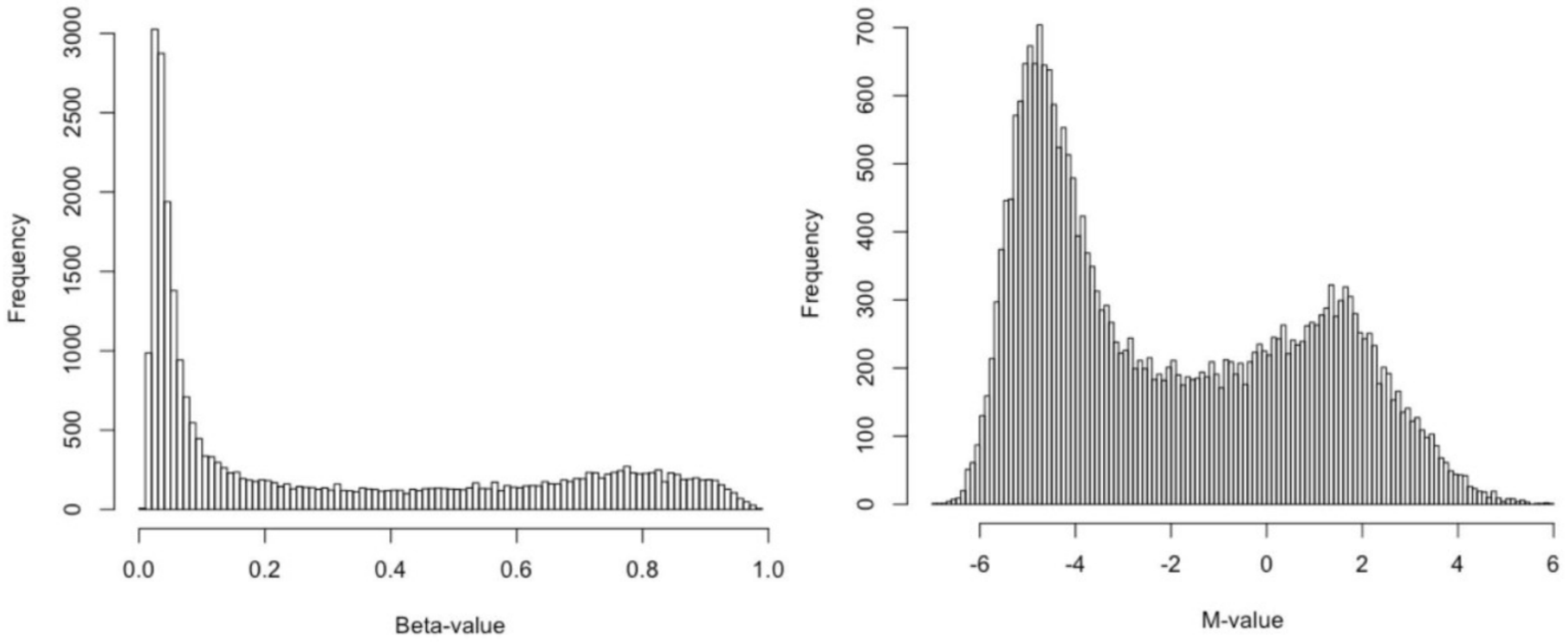
$$M = \log_2\left(\frac{\beta}{1-\beta}\right)$$

Relationship between Beta-value and M-value is a logit transformation



Du et al BMC Bioinformatics 2010

Quantifying methylation – M-values

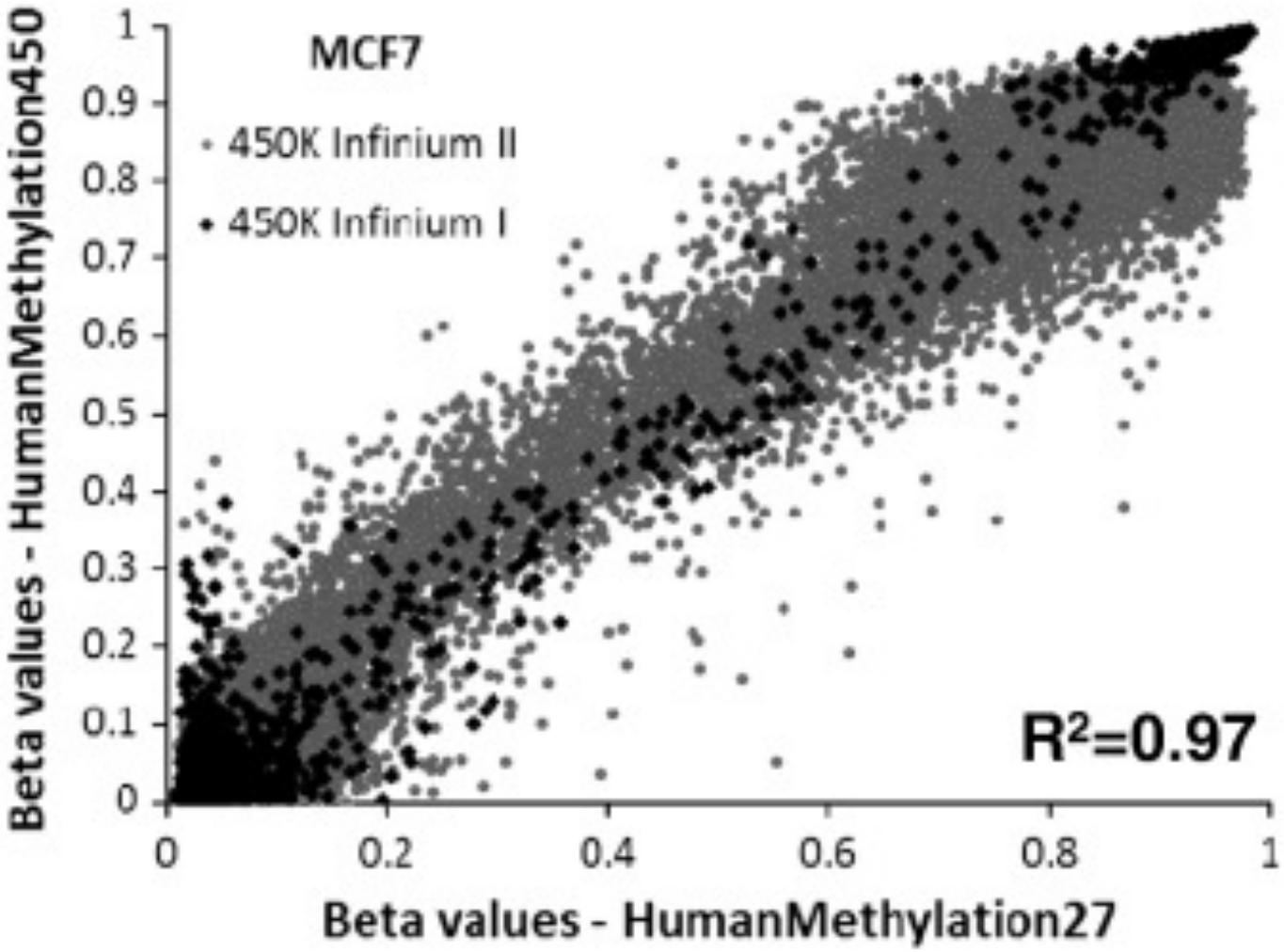


Comparison of Beta and M-values

- M-value
  - more statistically valid
  - M-value better detection rate and true positive rate
  - Difficult to directly infer the degree of methylation based on a single M-value
  - range of M-values may change across different datasets

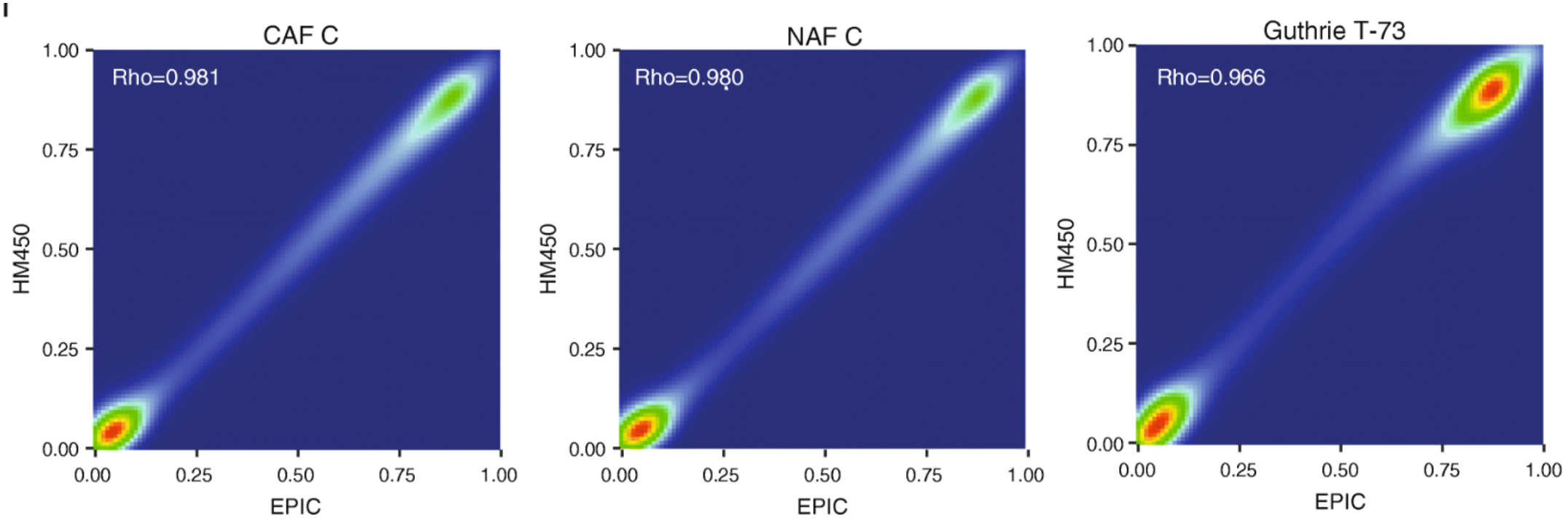
- Beta-value has a more intuitive biological interpretation,

27K vs 450K array



Bibikova et al Genomics 2011

450K vs EPIC



Pidsley et al Genome Biology 2016

450K vs whole-genome bisulfite-sequencing

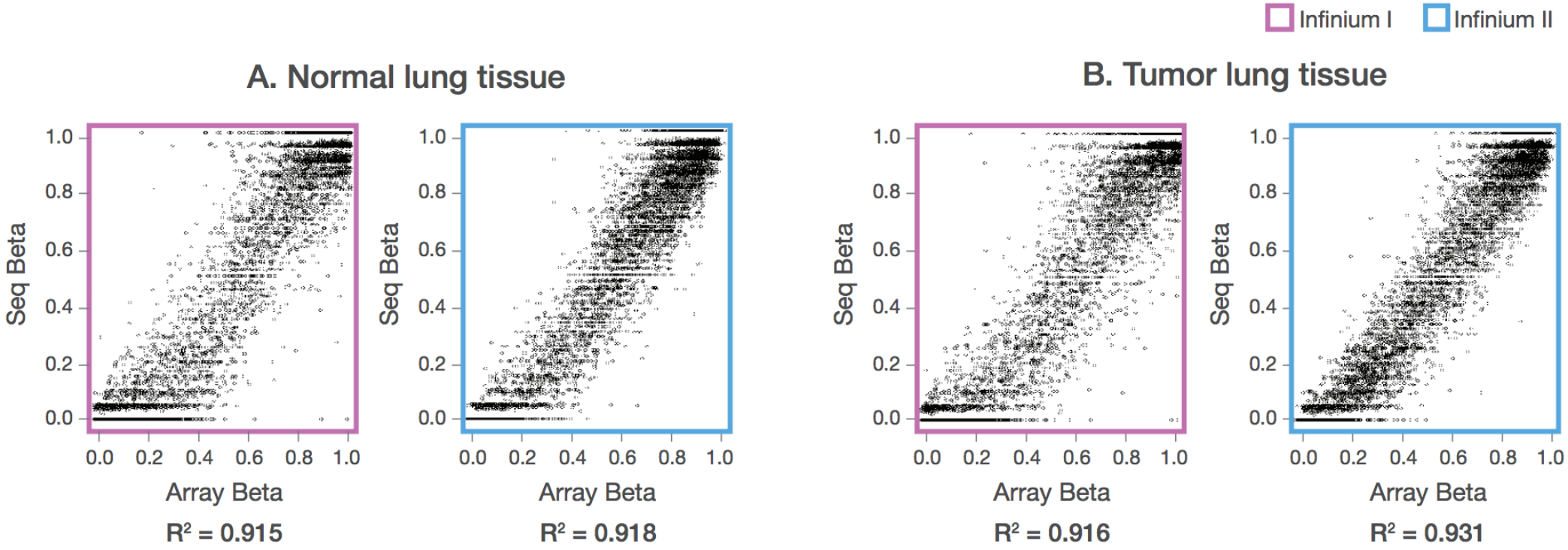
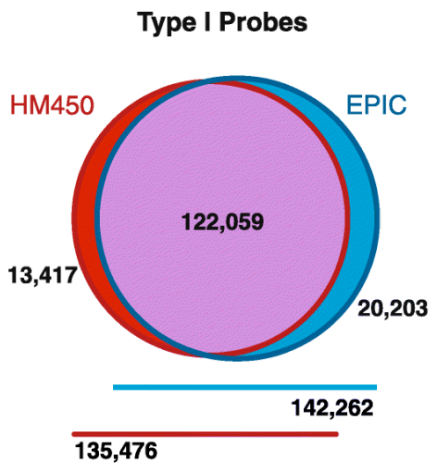
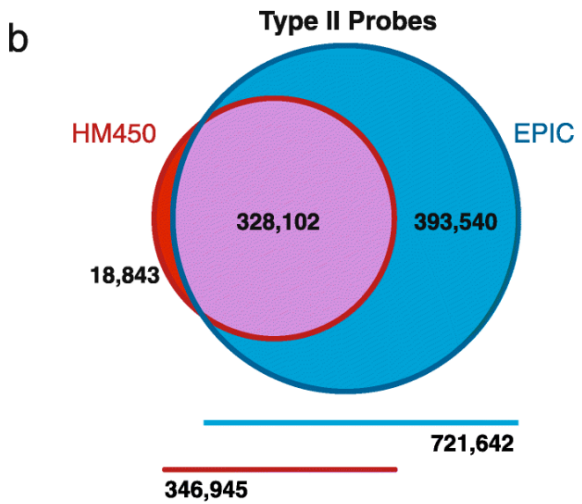


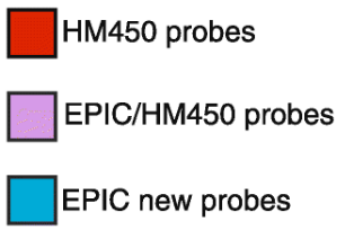
Figure 3: Relative Correlation of Infinium I and Infinium II Probes with WGBS —Normal lung (A) and tumor lung (B) tissue samples were assessed using the HumanMethylation450 BeadChip and WGBS, with high correlation seen between the Infinium I (purple) and Infinium II (blue) probes.

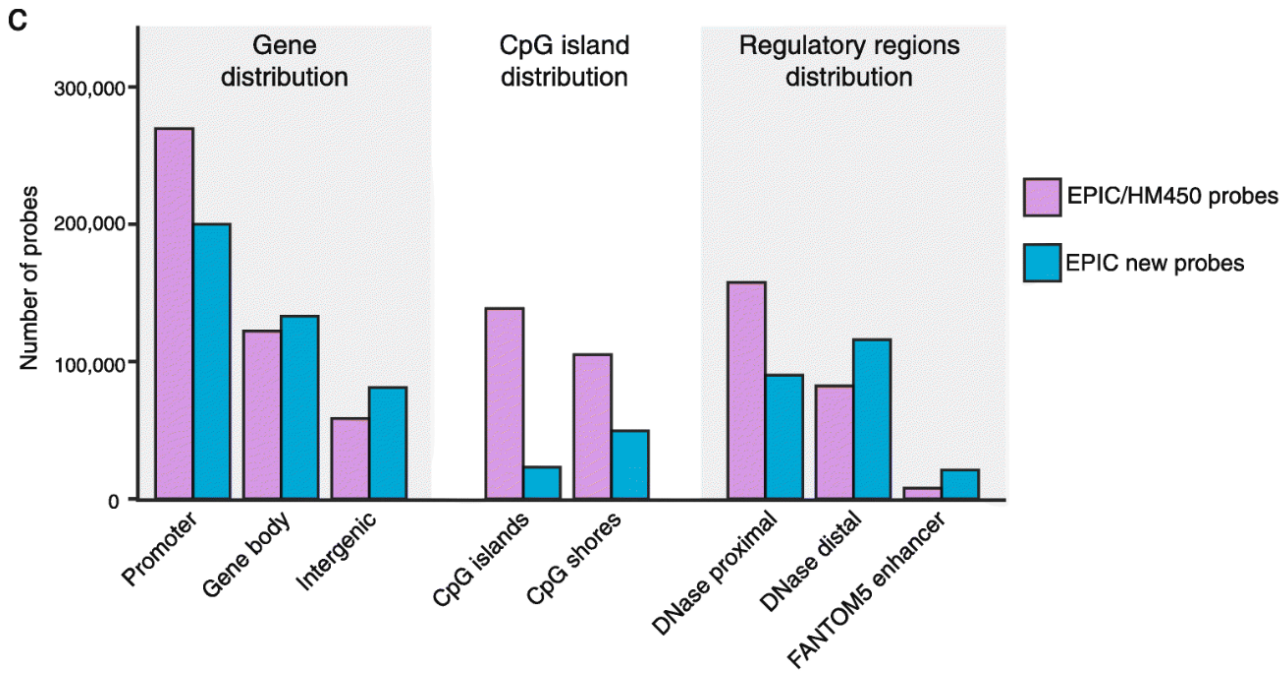
https://www.illumina.com/documents/products/technotes/technote_hm450_data_analysis_optimization.pdf

Illumina methylation arrays



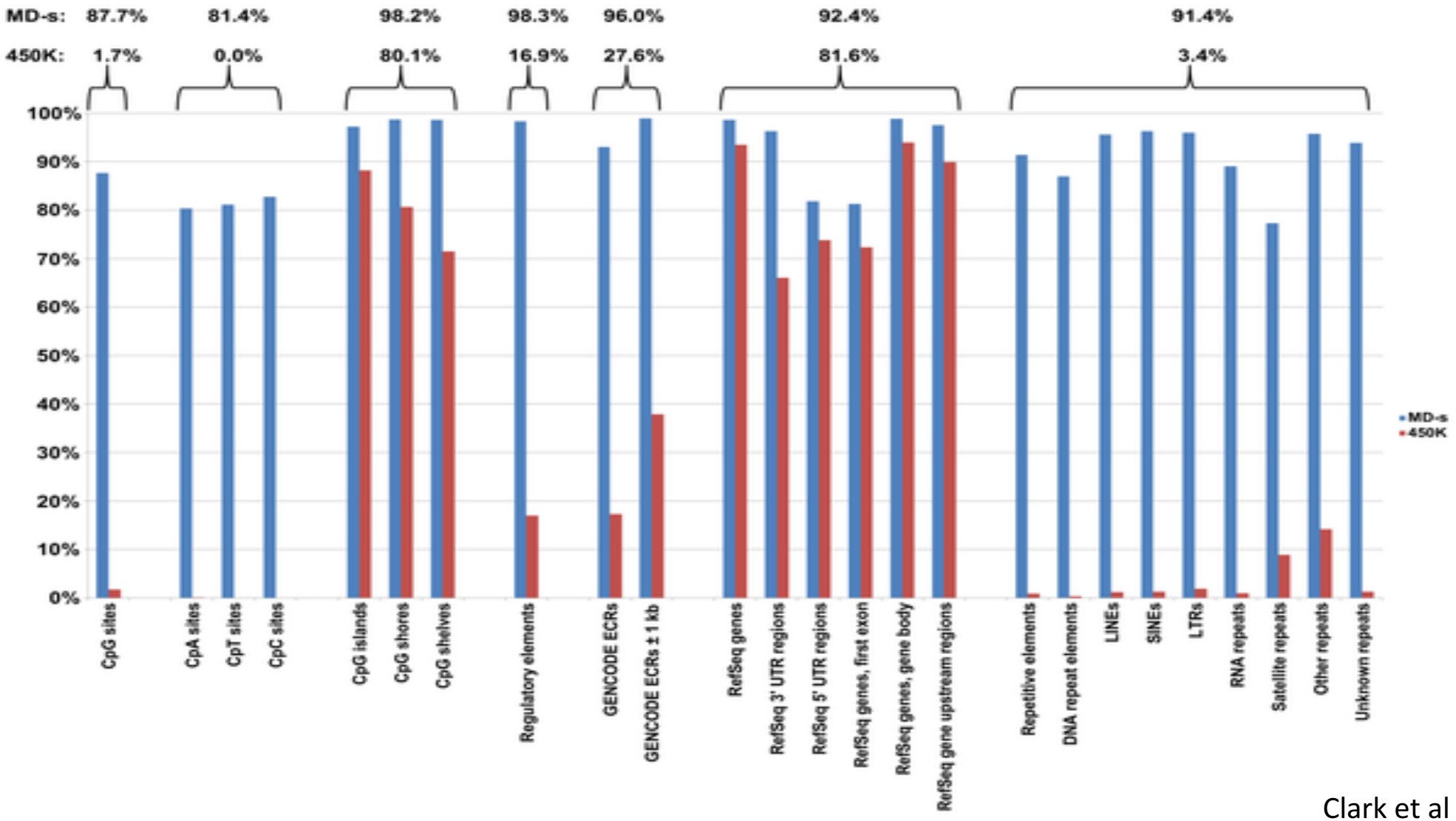






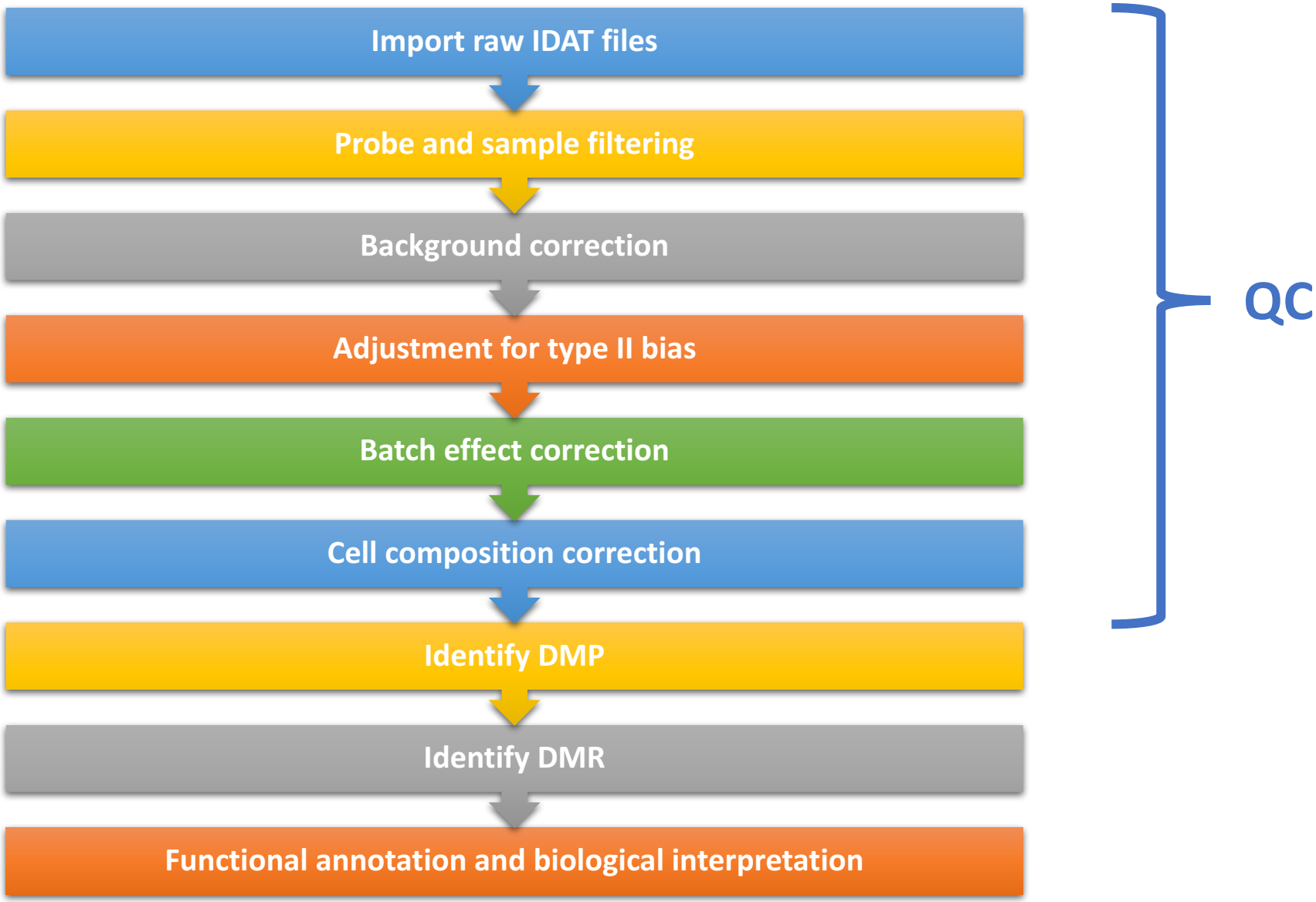
Pidsley et al Genome Biology 2016

MeDIP-seq vs Illumina 450K Array coverage



Clark et al PLOS ONE 2012

Analysis Workflow



Quality Control

Goal:

- Reduce variability introduced during the experimental process
  - Eg. 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 artefact
- Maintain the biological variation between conditions(i.e., cases/controls)

Main concepts:

- Control Probes
- Probe & sample QC
- Filtering
- Probe Type normalisation
- Batch effect

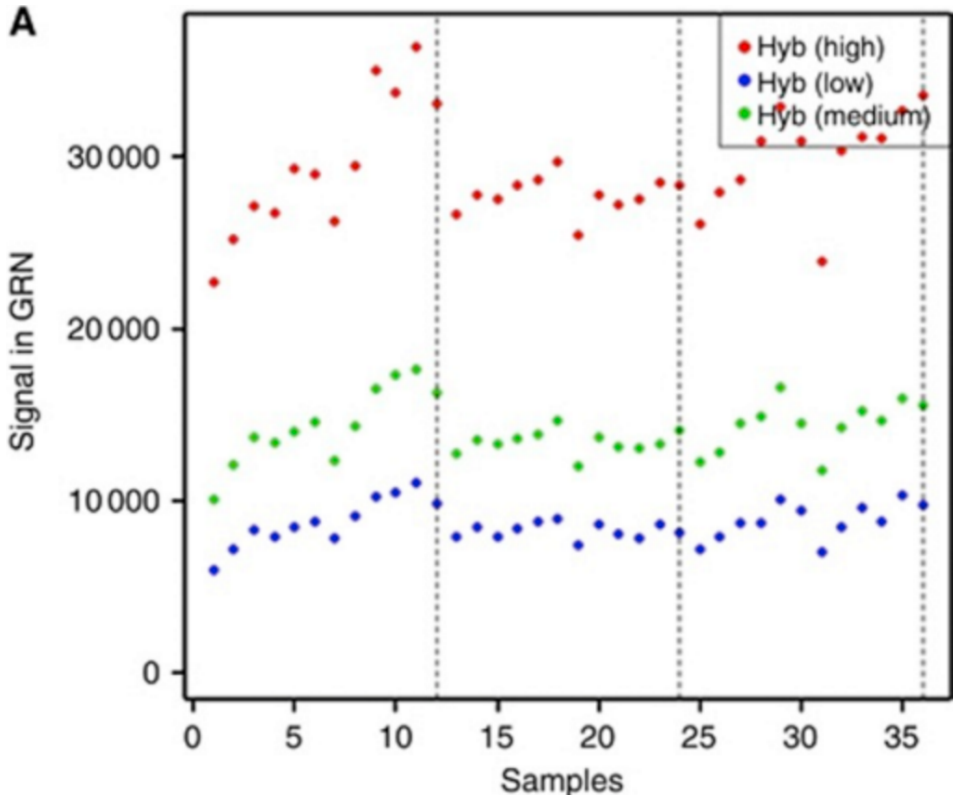
Sample QC – Filtering on control probes

- STAINING CONTROLS
- BISULFITE CONVERSION CONTROLS
- EXTENSION CONTROLS
- SPECIFICITY CONTROLS
- HYBRIDIZATION CONTROLS
- TARGET REMOVAL CONTROLS
- NON-POLYMORPHIC CONTROLS
- NEGATIVE CONTROLS

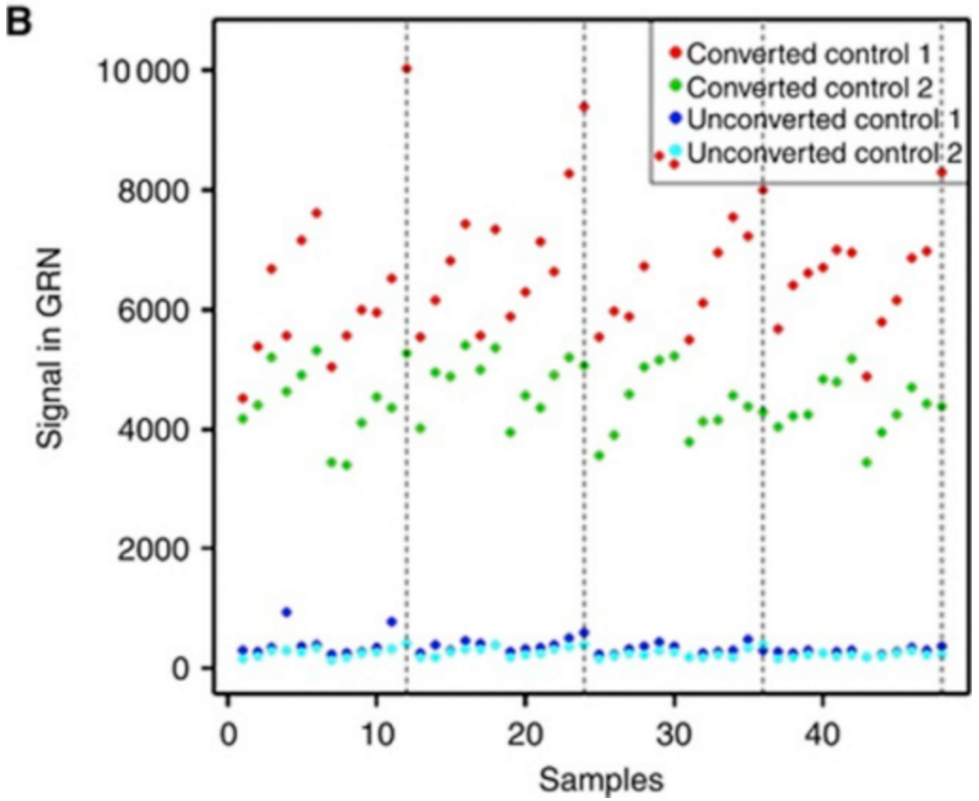
http://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/infinium_assays/infinium_hd_methylation/beadarray-controls-reporter-user-guide-1000000004009-00.pdf

Sample QC – Filtering on control probes

Hybridisation controls

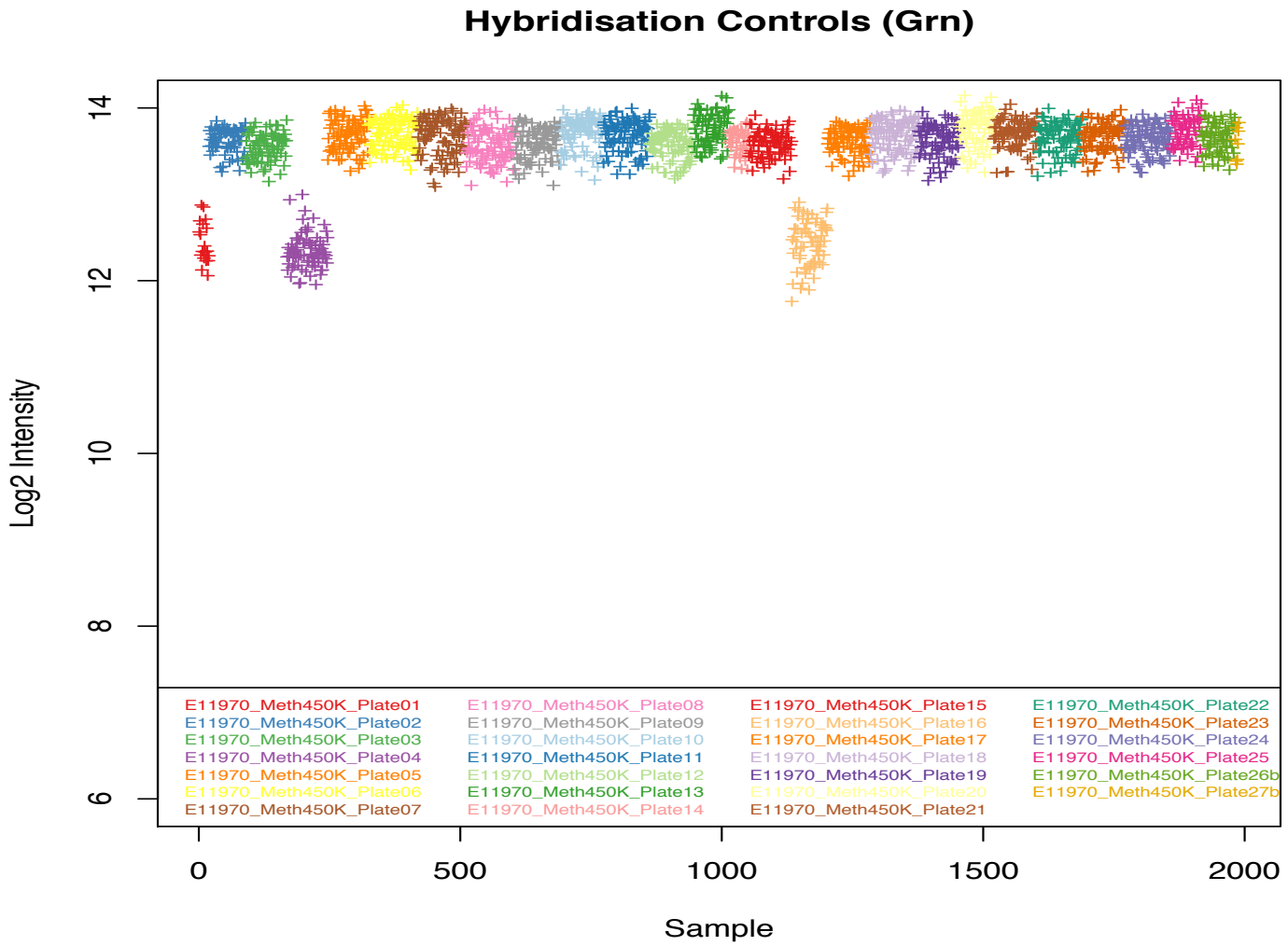


Bisulfite conversion controls



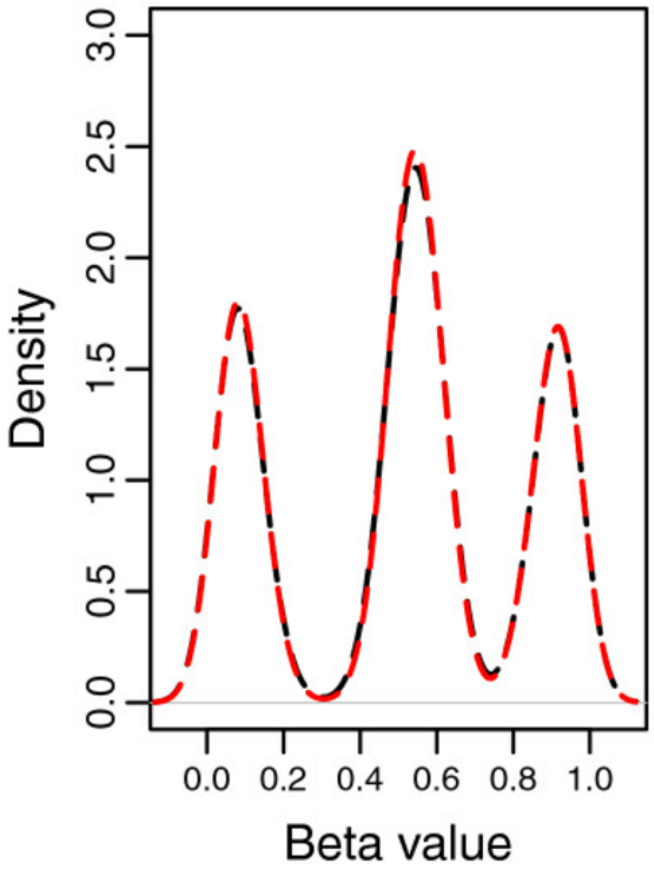
Wilhelm-Benartzi et al BJC 2013

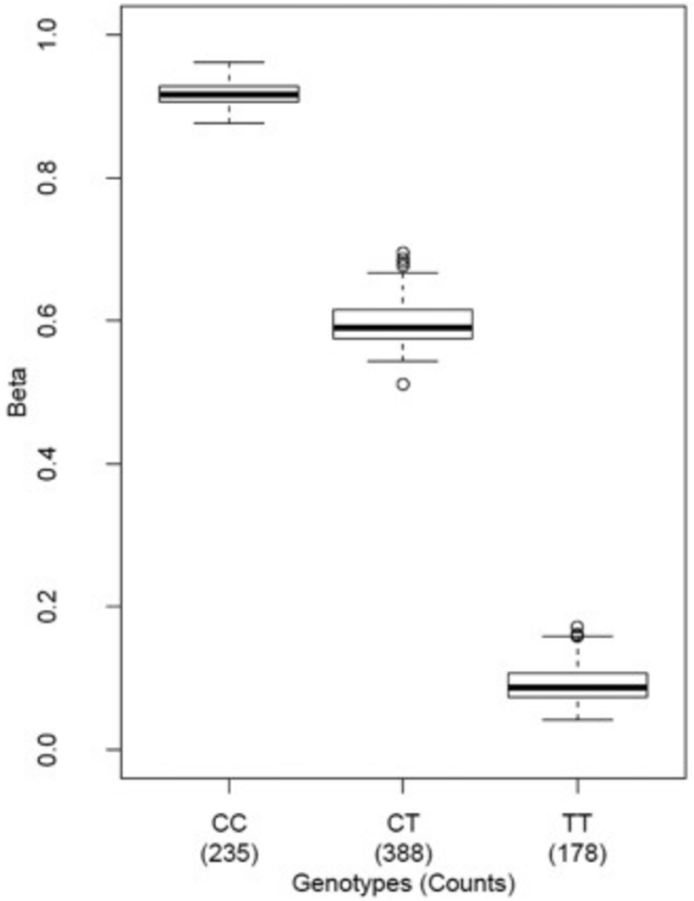
Sample QC – Filtering on control probes



Sample QC – Filtering on genotype

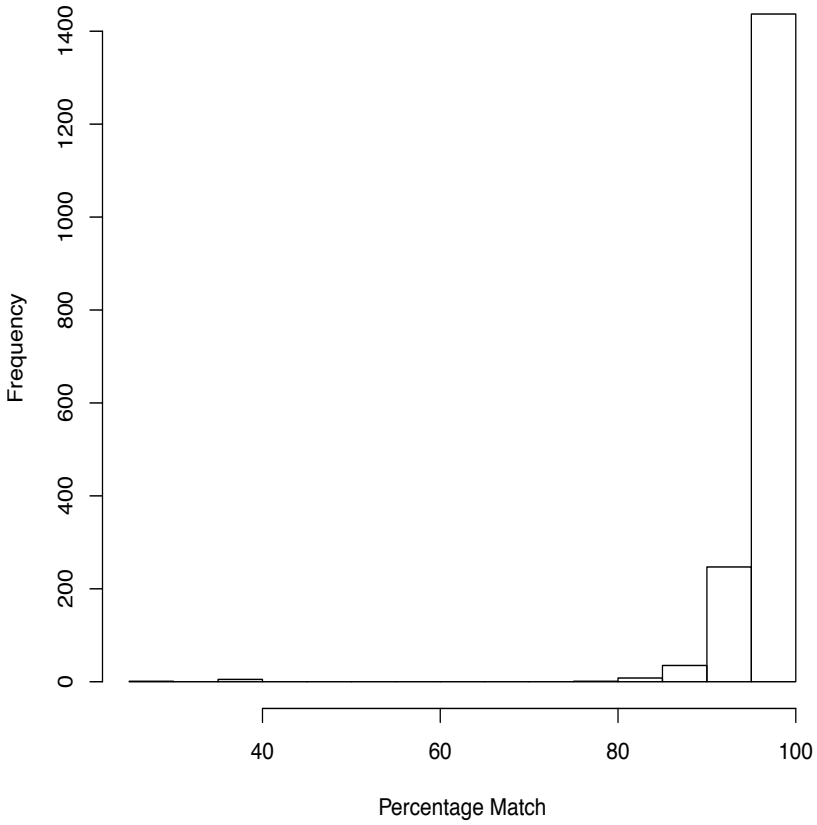
- 65 probes on 450K array whose target CpG site contains a SNP.
- Methylation signal at these probes can be used to predict sample genotype for the SNP.
- Used to generate a sample DNA "fingerprint"



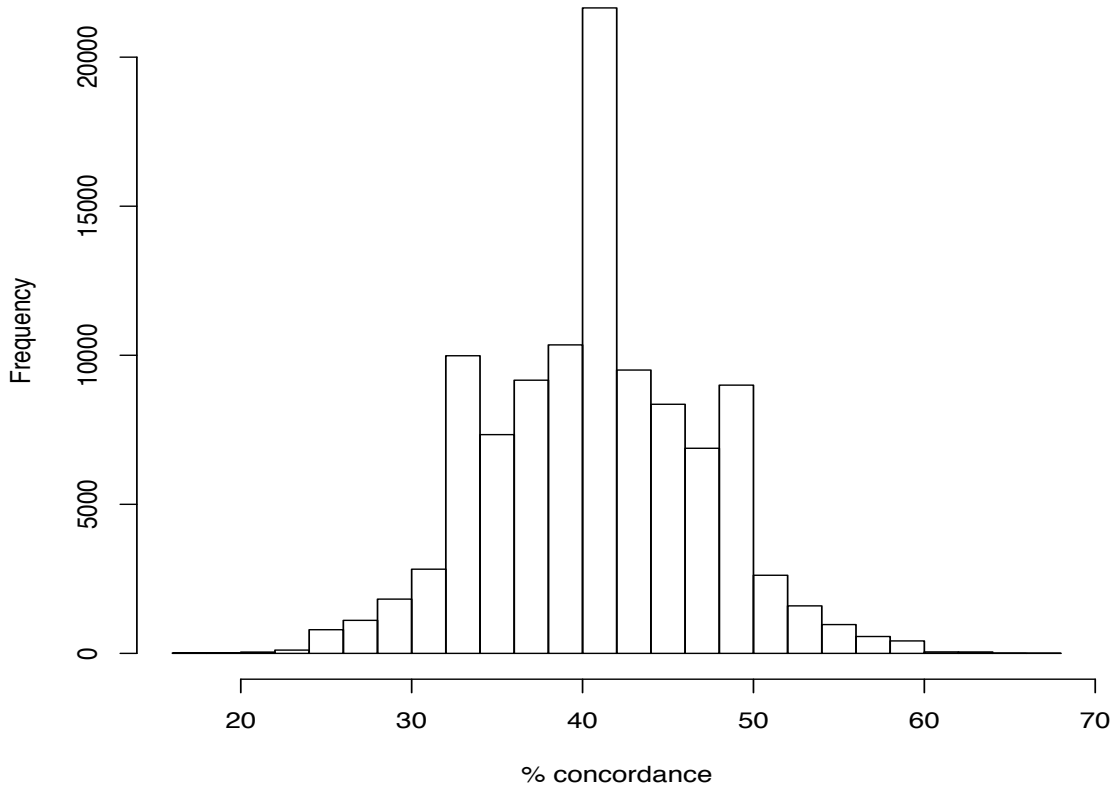


Sample QC– Genotype concordance

Duplicate pairs

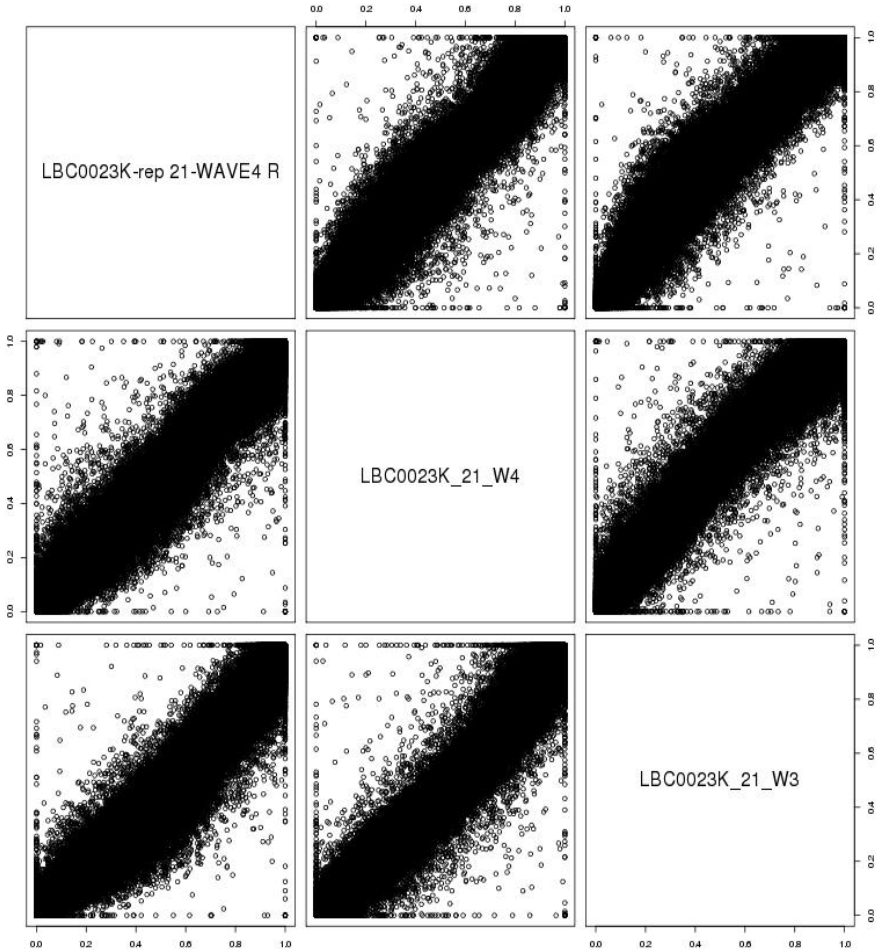


Unrelated individuals

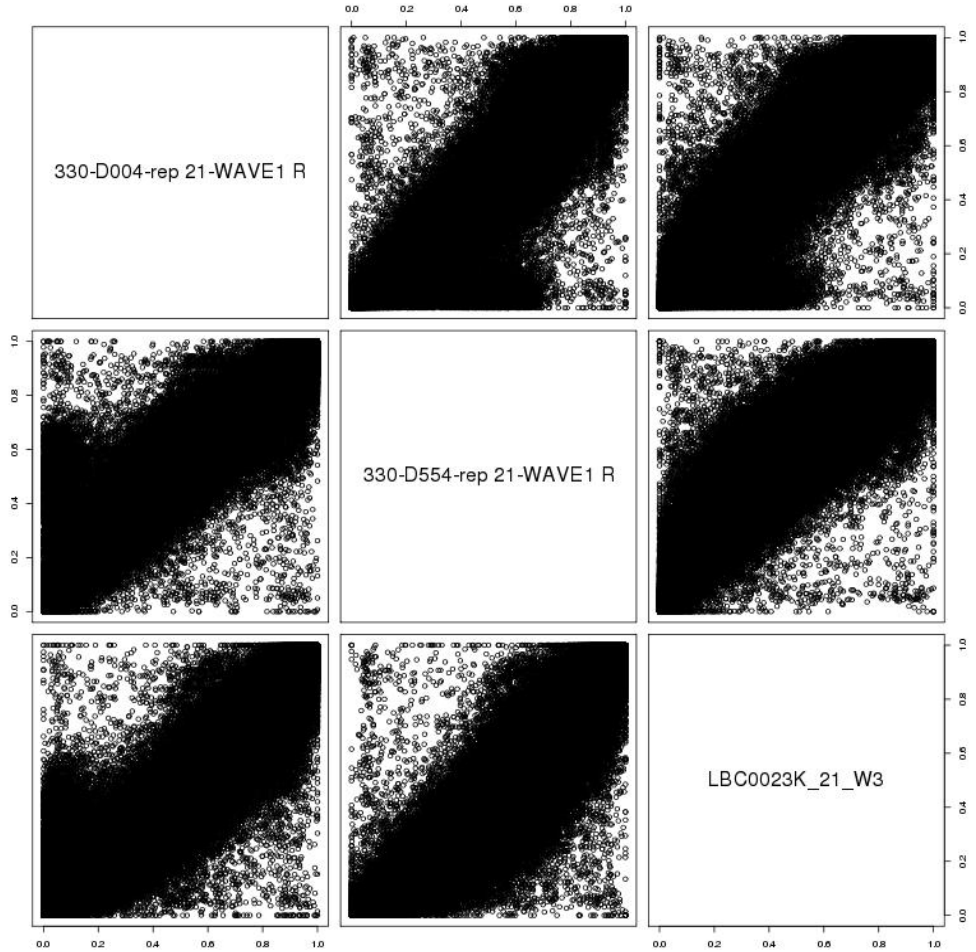


Reproducibility

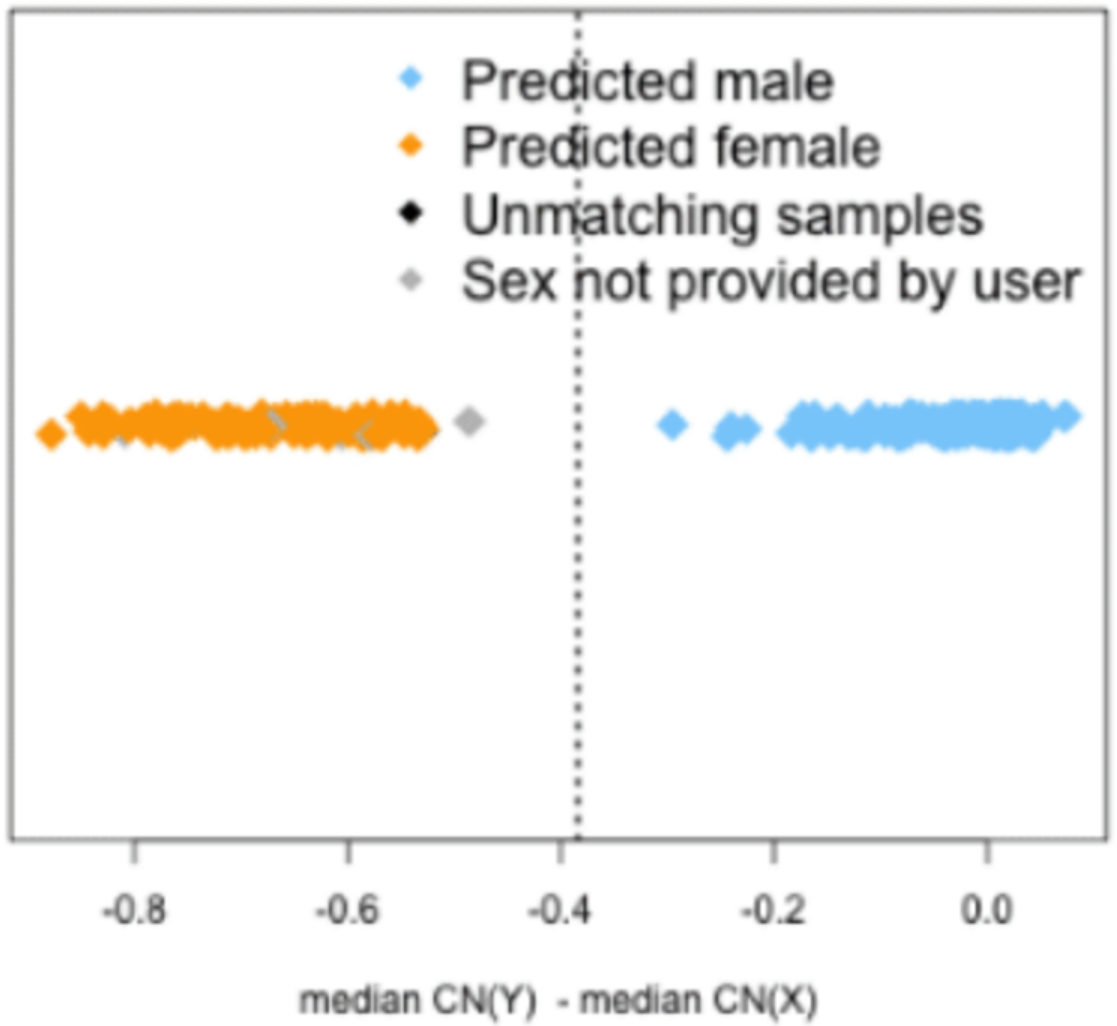
Sample repeats beta concordance



Unrelated samples beta concordance



Sample QC – Filtering on predicted sex

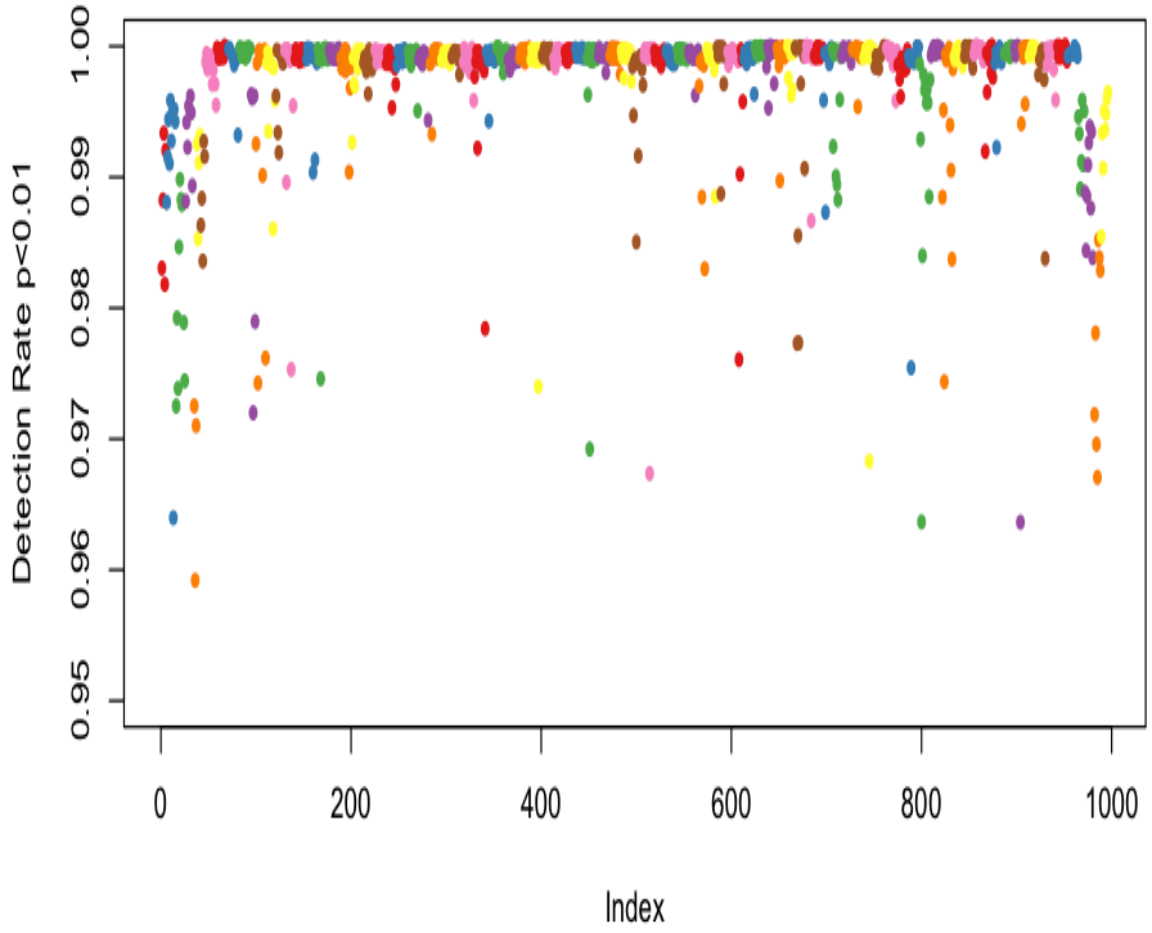


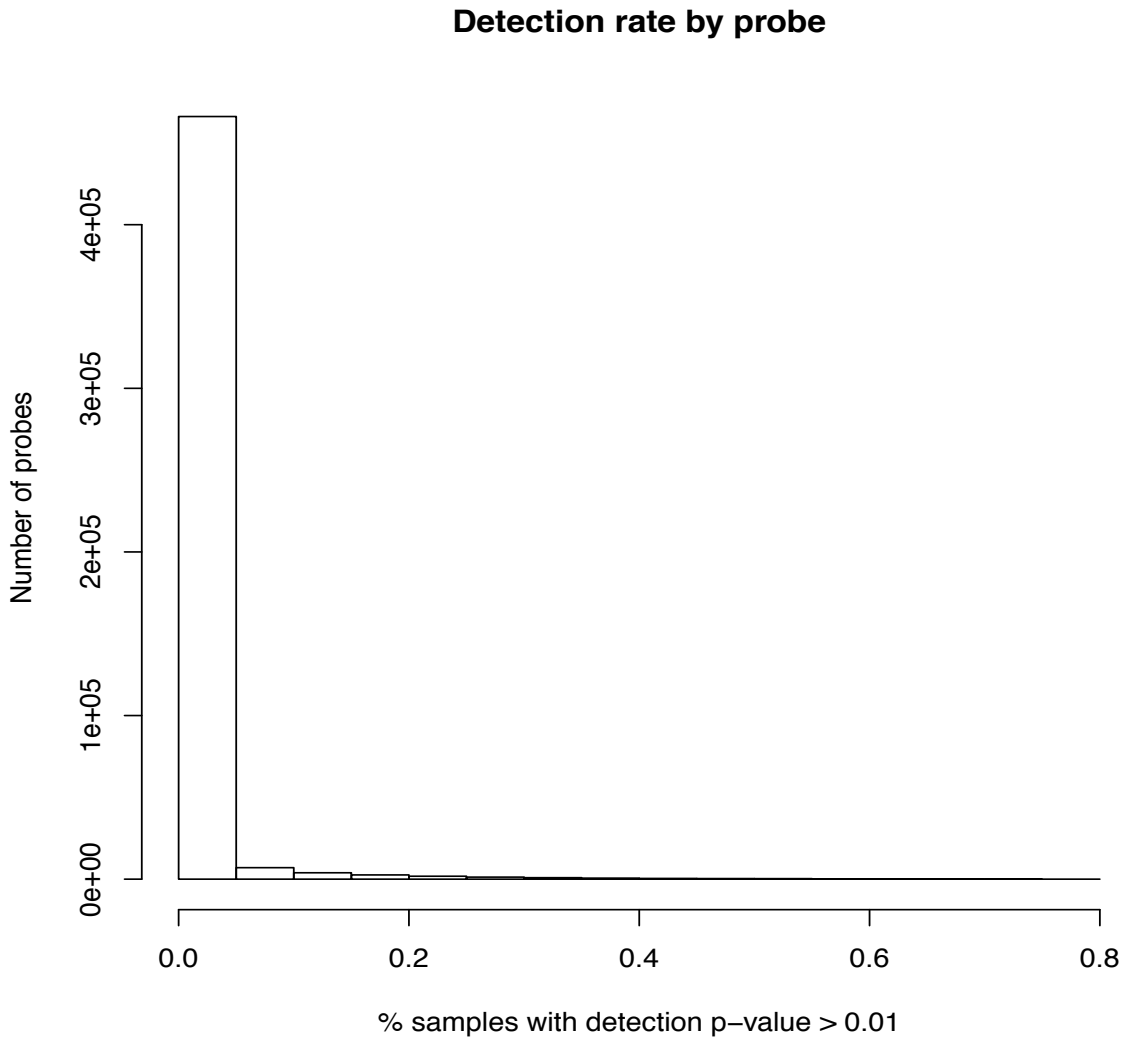
Probe and Sample QC – Detection P-value and background correction

- Compares the total DNA signal (Methylated + Unmethylated) for each probe to the background signal estimated using negative control probes.
- The detection *P*-value
- Common practice:
  - Drop probes where median p-value >0.01
  - Drop probes that are not detected in nth% of samples
  - Drop samples where nth% of probes are not detected
- Background correction commonly used – simple subtraction of background intensity from total signal
- Removes non-specific signal from total signal and corrects for between-array artefacts.

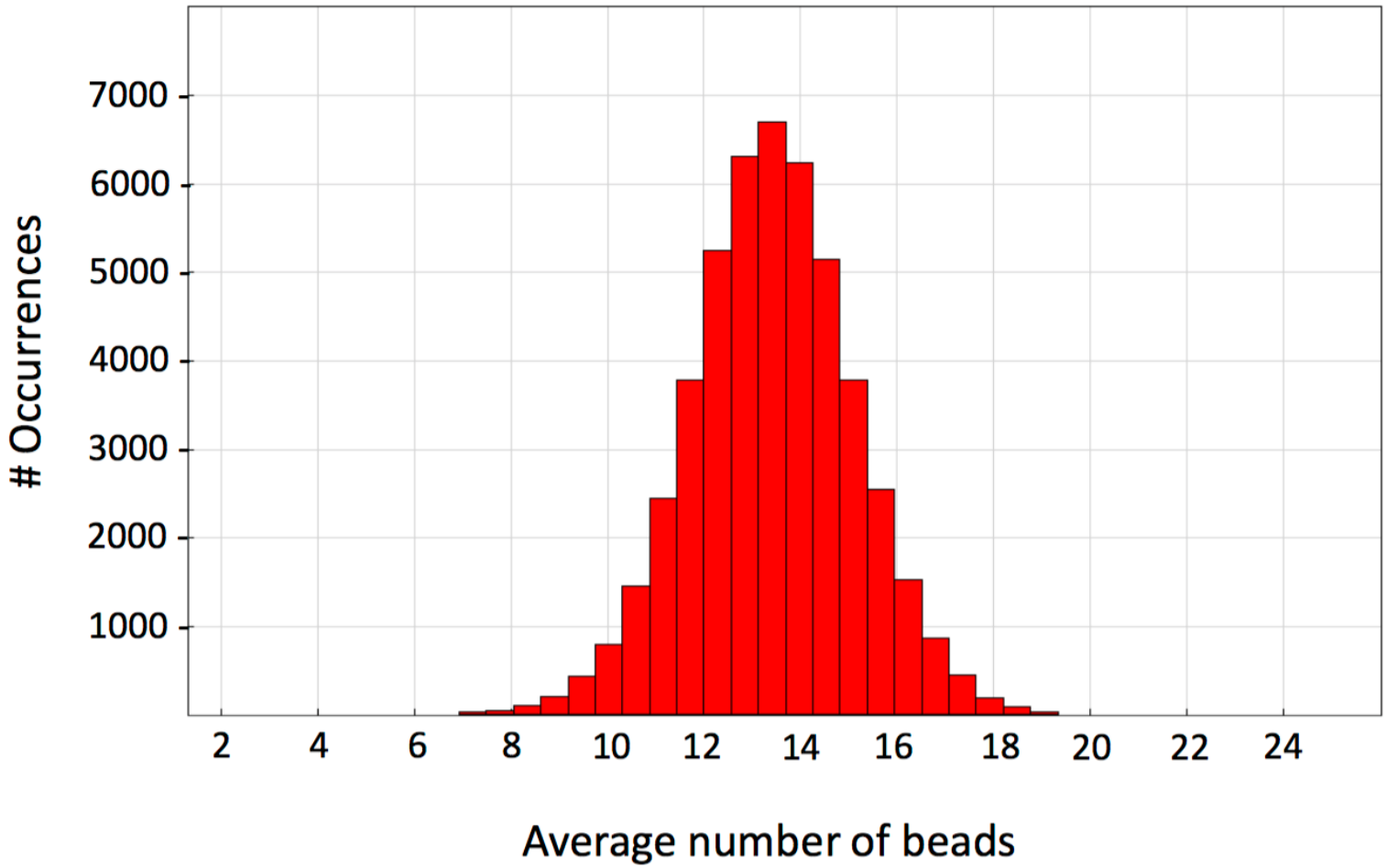
Probe and Sample QC – Detection P-value

Per sample detection rate





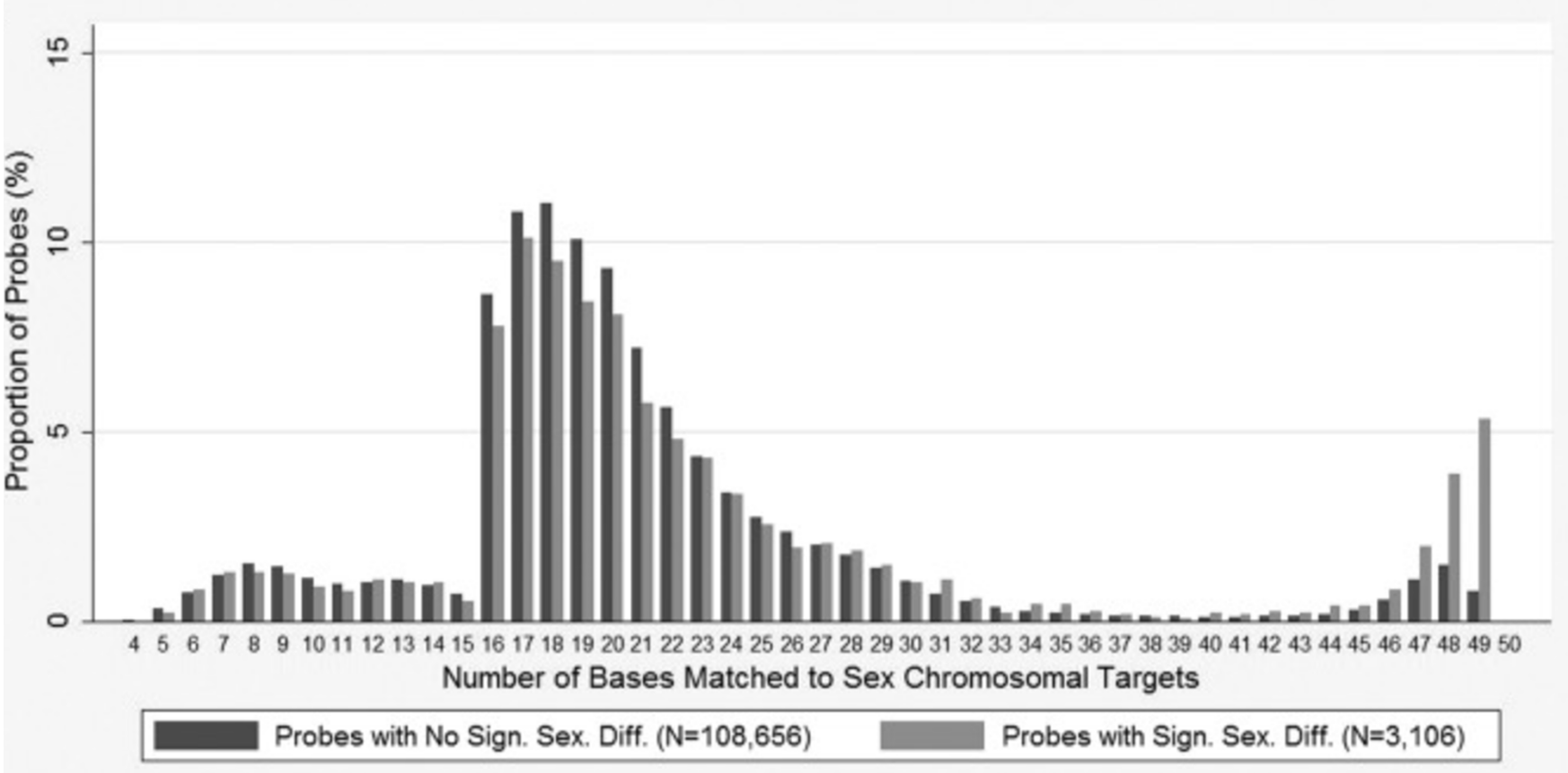
Probe QC – Filtering on bead count



Filter out probes < 3 bead counts

Probe QC - Filtering cross-reactive probes

- Large number probes cross-hybridise to non-targeted genomic regions

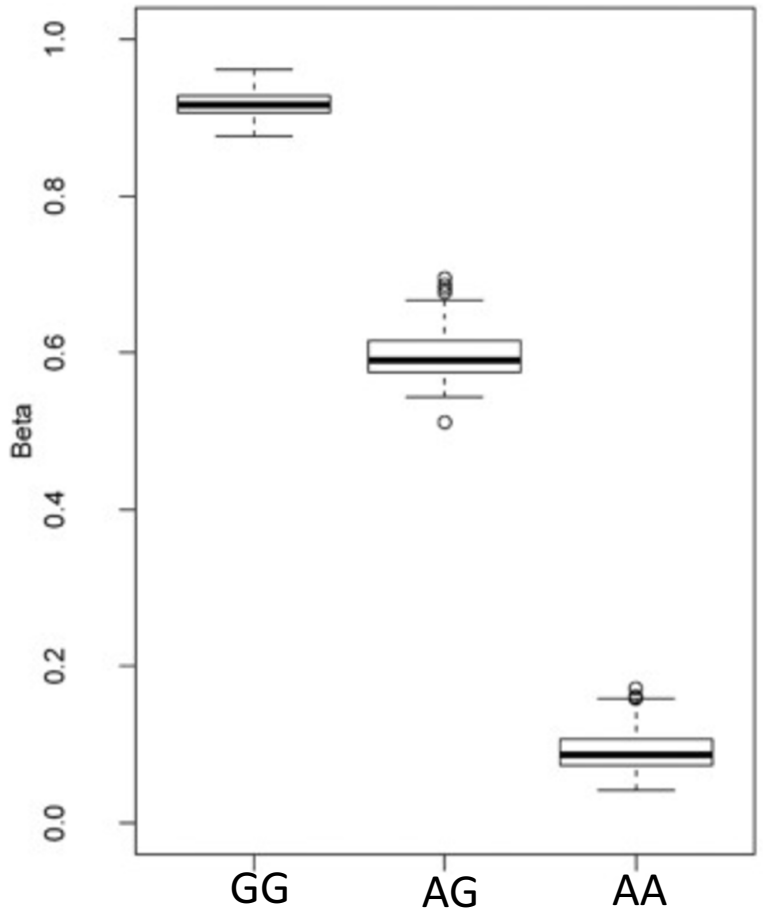


Type II probes

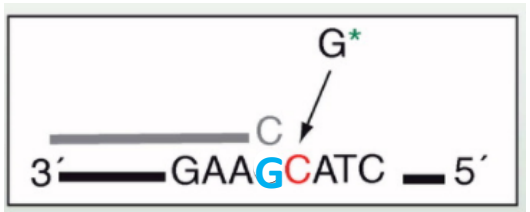
Chen et al Epigenetics 2013

Probe QC – Filtering on SNP probes

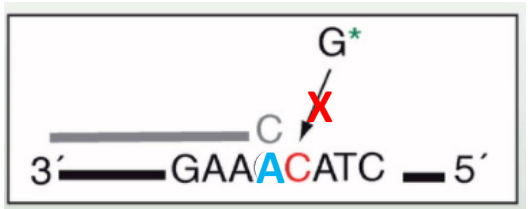
13.8% of the probes have known SNPs within the targeted CpG site



Type II probe



Signal in green channel



No signal

Polymorphic CpG sites on 450K array

Polymorphic Position	Total Probes		Infinium I		Infinium II	
	N	%	N	%	N	%
C	35524	7.3%	5956	4.4%	29568	8.4%
G	33905	7.0%	5961	4.4%	27944	8.0%
The Base Before C	1429	0.3%	1429	1.1%	-	-
Total Probes	66877	13.8%	12671	9.4%	54206	15.5%

Chen et al Epigenetics 2013