Outline for Session 2 (1.30 – 3.00pm)

- Experimental Design
- Normalisation
  - Background correction
  - Colour Bias
  - Across Array
  - Probe Bias
- Batch Effect Correction

Normalisation

Goal: reduce non-biological variation

A. Experimental design is critical for reducing technical variation:

- Randomising cases and controls on plates, arrays, run times etc.
- Repeated samples run on across plates, arrays, run times etc.

B. Statistical methods to reduce technical variation:

1. Within array normalisation - correcting for intensity-related dye biases
2. Between array normalisation - removing technical artifacts between samples on different arrays

No consensus on best normalisation approach.

Experimental Design

- This is the most critical part of any study
- Poor experimental design can result in not being able to draw any conclusions from a study

- Record as much information as possible about the experiement
  - DNA extraction dates/batches
  - Bisulphite conversion dates/batches
  - Array processing dates/batches
  - …

Experimental Design – Example #1

- Case-control study looking at methylation and disease

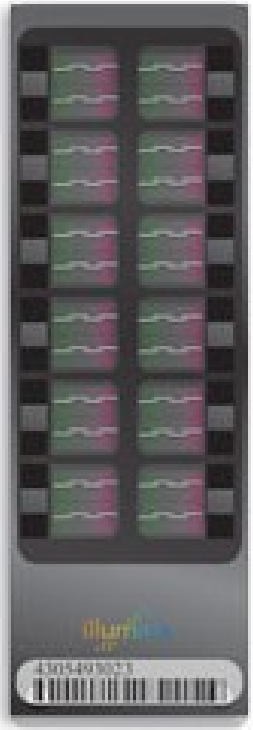
- How were the cases and controls collected?
- Was the DNA extracted and stored in the same fashion?
- How was bisulphite conversion done?
- How are cases and controls placed on methylation arrays?

Experimental Design – Example #1

- Cases and controls should be collected using common methods
  - Over a similar time frame
  - Have similar demography (age, sex, ancestry, smoking, …)
  - DNA should be extracted and stored by a common method
    - Ideally not in batches of cases or controls, but if necessary there should not be a single batch of each…
    - Record all information on DNA extraction batches (date, operator, …)
  - Cases and controls should be randomly placed in batches for bisulphite conversion
  - Cases and controls should be randomly placed on arrays
  - Consider using control samples and duplicates to track quality over time

Experimental Design – Example #2

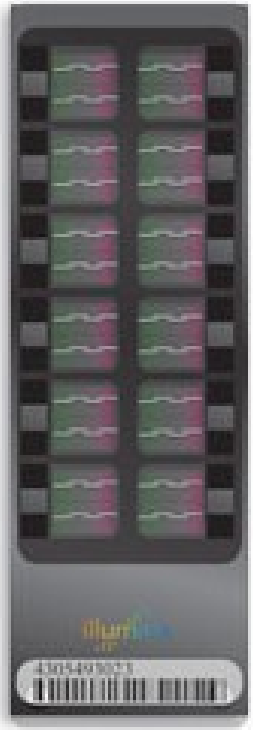
- Investigating transmission of DNA methylation across generations (e.g. mother -> daughter)



Experimental Design – Example #2

- Investigating transmission of DNA methylation across generations (e.g. mother -> daughter)

- Do not put mother and daughter beside each other
- Do not put them on the same array
- Do not have all mothers on one array and daughters on the other



Normalisation

- Although good experimental design is key to a successful experiment, power can be gained by removing batch or processing effects in that data

- This is a landmine…
- There is consensus on the best method to use
- New methods claiming to be the best are released weekly

- No normalisation method should be used blindly

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)
- Other methods include
  - Smoothing data
  - Fitting complex mixture distributions to model signal + noise and subtracting noise

Normalisation – Background Correction

- Usual approach is to subtract the estimated noise from the signal
- Can result in negative intensity values
  - Truncate to zero

- Implemented in a wide variety of R packages
- Often occurs during initial data reading

- The Illumina GenomeStudio default is widely used

Normalisation – Colour Bias

- The two colour channels are know to perform differently
- Usually higher overall intensities on the red channel that the green channel
  (extreme differences in colour intensities should be caught when cleaning bad samples from the data)
- Large number of methods to handle this….

Normalisation – Colour Bias

- 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 (is this a good idea?)
- R methylumi
  - Same as above but scales to sample on array with least difference in average dye intensities
- ASMN (All Sample Mean Normalisation)
  - Modifies above to scale to the average across all samples

Normalisation – Colour Bias

- R wateRmelon - nanes and nanet
  - Quantile normalisation for methylation and unmethylation intensity values either for both Type I & II probe types (nanes) or separately (nanet)

- R lumi
  - Implements a variant of quantile normalisation

- The Illumina GenomeStudio version is still widely used

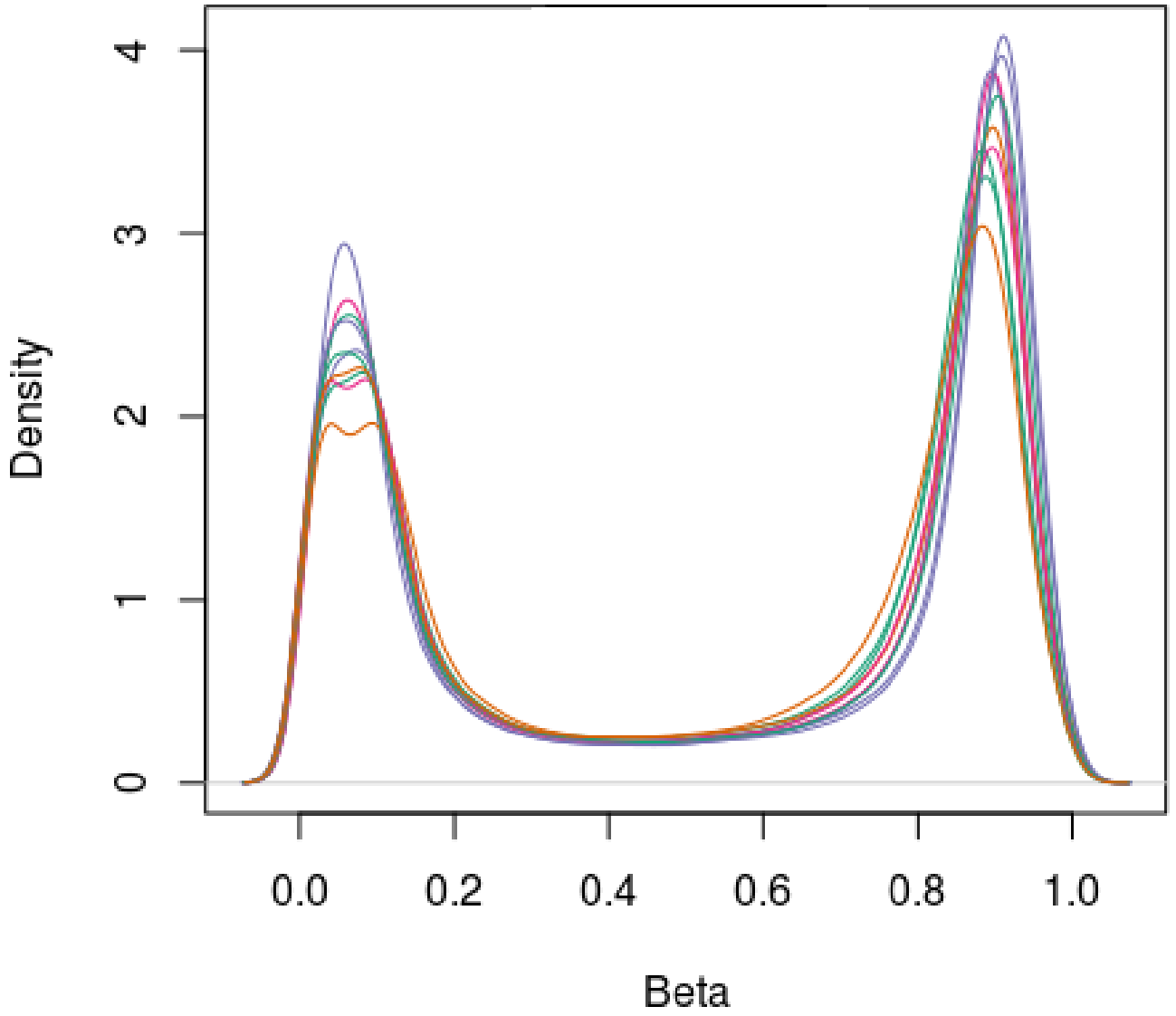
Normalisation – Across Array

- Beta values are calculated once background correction and colour bias removal is performed

$$\beta = \frac{M}{M+U+\varepsilon}$$

- Next stage is normalisation is to normalise the beta values

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 people
  - Case/control studies can have vastly different methylation profiles (e.g. cancer)

Normalisation – Across Array

- Functional normalisation

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

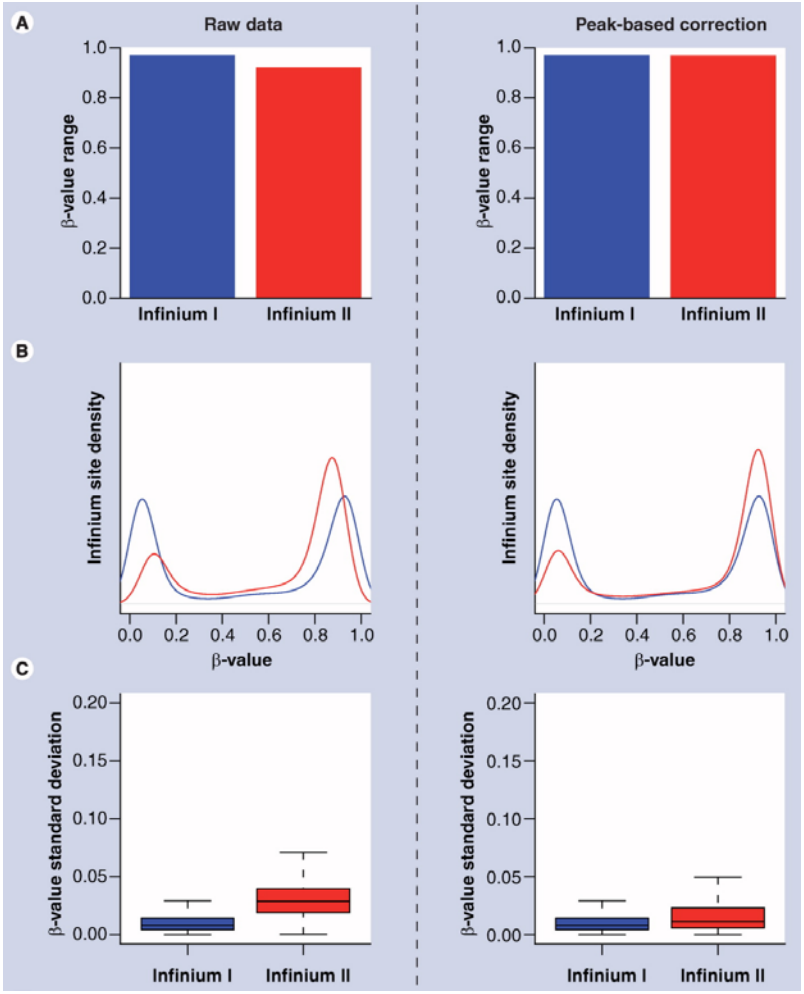
- We will use this method in the practicals

Normalisation – Probe Bias

- Some measurement bias is shown between Type I and II probes
- 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

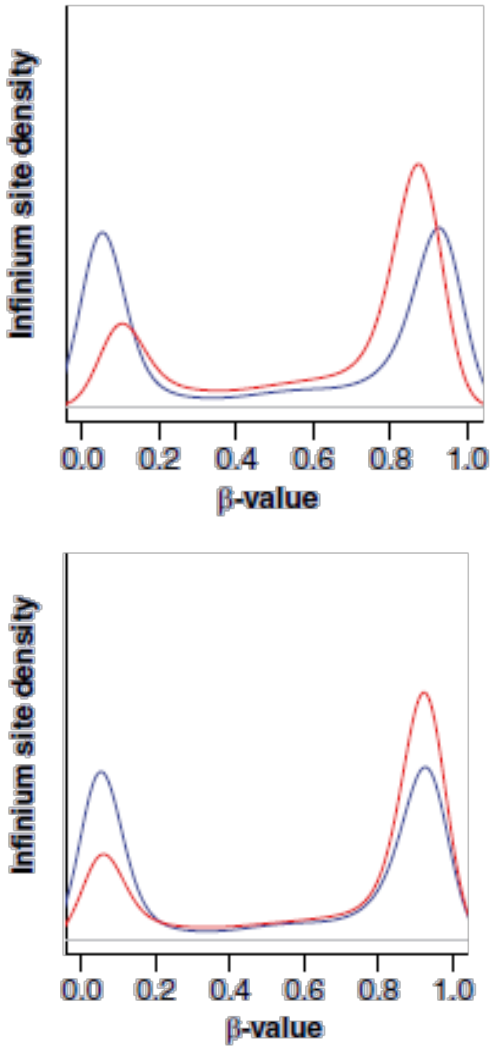
- Type II probes have a smaller range of beta values than Type I probes
- Type II probes are more variable than Type I probes
- This may be expected given biology…



Normalisation – Probe Bias

- Peak Based Correction

- Uses peak summits to correct β values
  - Convert β to M values
  - Determine peaks for I and II probes with kernel density estimation
  - Rescale M values by peak summits
  - Convert these corrected M values back to β values



Dedeurwaerder et al. *Epigenomics* 2011

Normalisation – Probe Bias

- Beta MIxture Quantile Dilation (BMIQ)

- “The strategy involved application of a three-state beta-mixture model to assign probes to methylation states, subsequent transformation of probabilities into quantiles and finally a methylation-dependent dilation transformation to preserve the monotonicity and continuity of the data”

- Currently a widely used approach…

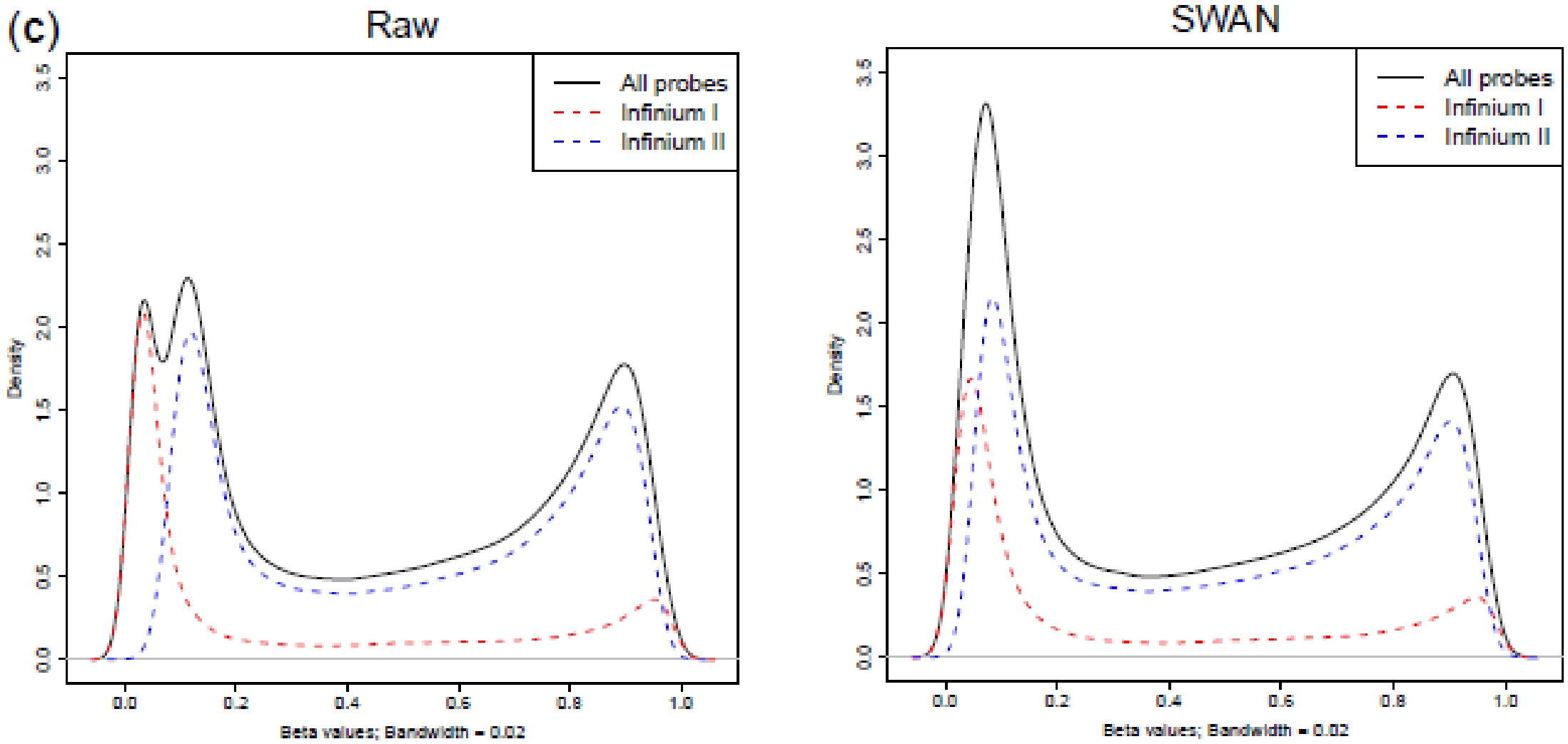
Normalisation – Probe Bias

- Subset Within-Array Normalization (SWAN)

- Normalises TypeI and TypeII probes together
  - Subsets all probes that cover the same number of CpG sites

  - Takes the methylated and unmethylated channels, calculates mean intensity
  - Scales TypeI and TypeII probes to this mean separately by linear interpolation

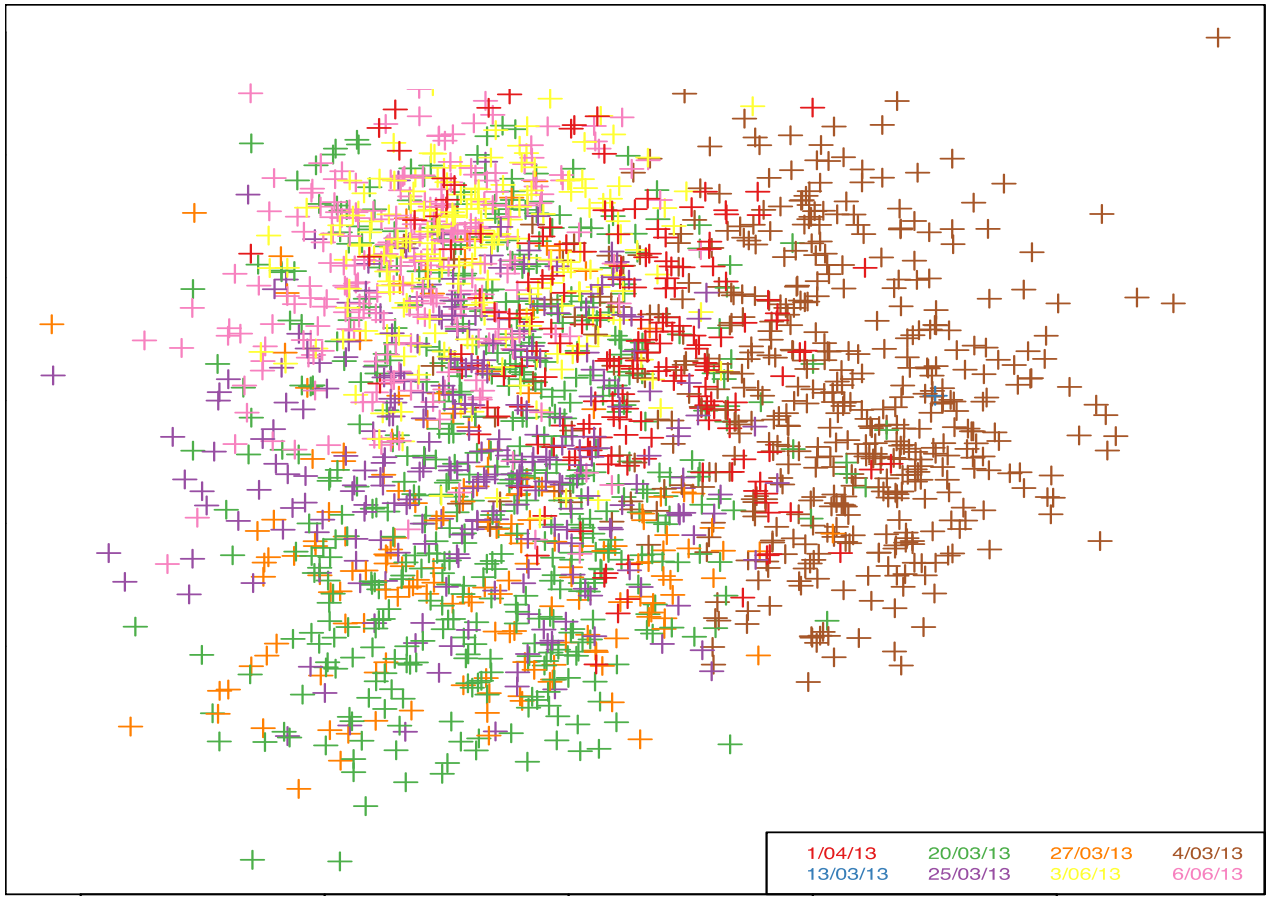
Normalisation – Probe Bias



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 (SVA, ISVA - attempt to infer the unwanted variation from the data itself)

Batch effects

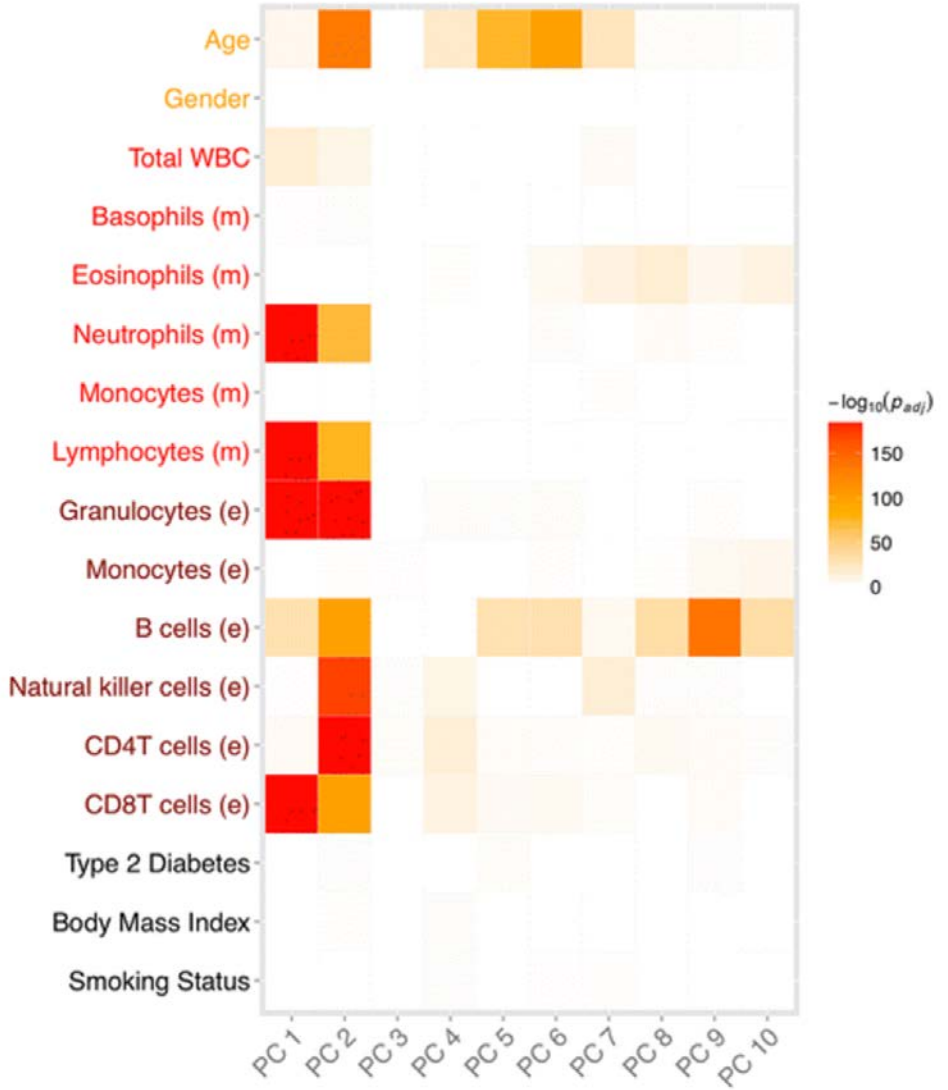


Batch Effects

- We have carefully recorded all information from our experimental design…
- We can correlate each of these with the Principle Components of the DNA methylation data to test if they explain variation in the data
- Once we know which effects to correct for we can either include them in our analysis model (if possible) or pre-correct the data.

Principal Component Analysis

- Understand the major sources of variation in methylation across array.
  - Can highlight possible confounders that should be adjusted for
- Identify outlying samples that may need to excluded from analysis



Batch Effects – Linear Regression

- The most basic correction for batch effects is to perform a linear regression with known batch effects as covariates

- Convert to M values and then back to Beta values
- Take the residuals of the model through to further analysis

- Different regression for each probe

Batch Effects – COMBAT

- Method designed for gene-expression data
- Can be used for DNA methylation after transforming to M values

- Uses information across probes to scale the residual variance to provide more accurate estimates of corrected values

Further normalisation

- The normalisation methods covered so far are at the limit of the corrections that can be done given the recorded information
- Further corrections may remove genuine biological differences between the groups
- We can attempt to recover unobserved batch effects from a variety of methods

Unobserved Batch Effects

- Principle Component Analysis…
- When used on all probes at once, there is a great risk of removing the biological effects you are trying to detect

- Compromise: Use PCA on the control probes
  - Is known to capture effects of array and array position
  - Is unlikely to capture all unobserved effects due to the small number of control probes (and the fact that control probes have very specific design)

- How many PCs to include?

Unobserved Batch Effects

- Remove Unwanted Variation (RUV)
- A suit of methods to try capture unobserved batch effects from the data

- General approach with DNA methylation data
  - Perform analysis
  - Take bottom 50% least associated probes
  - Do a PCA on those probes…

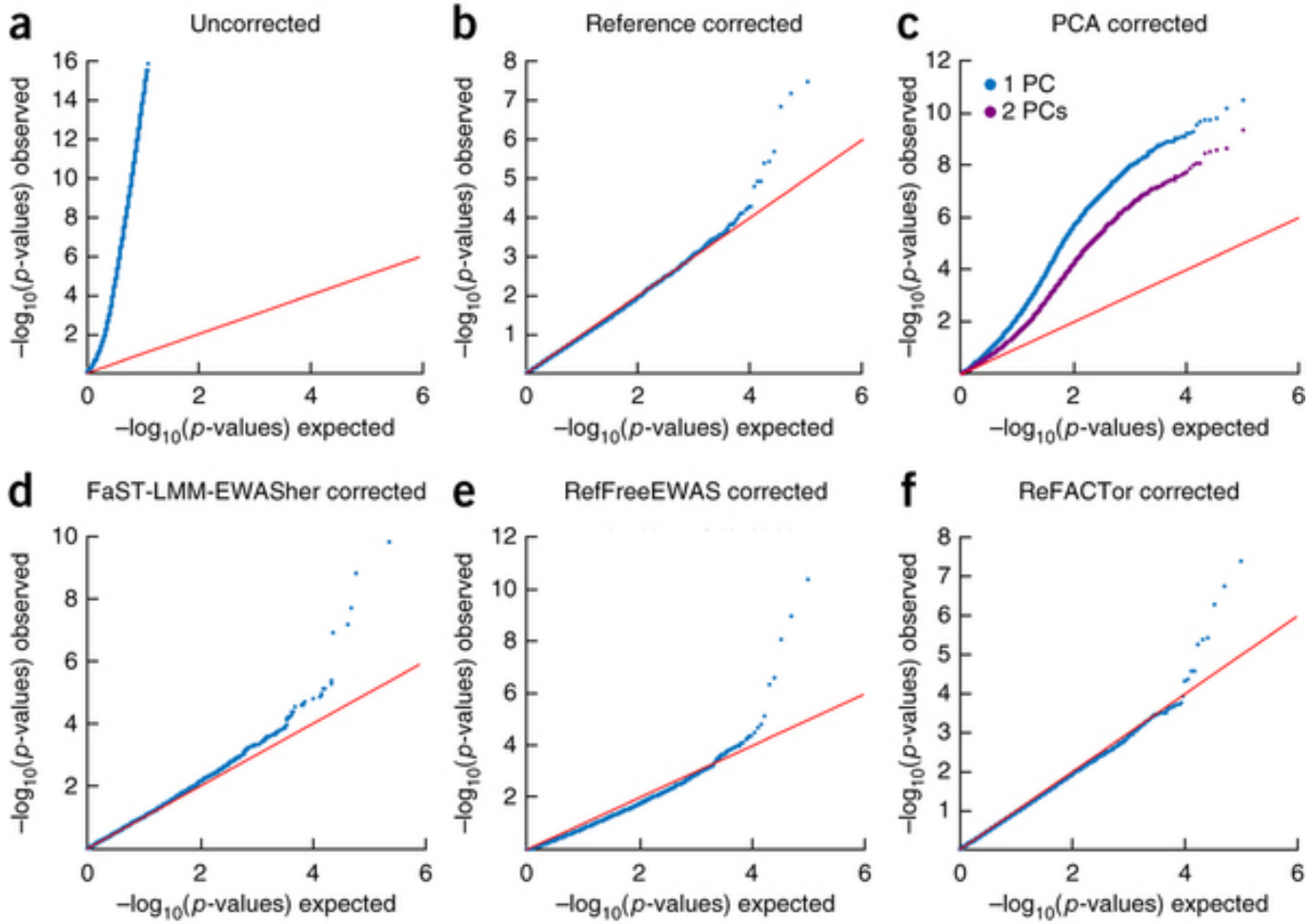
Unobserved Batch Effects

- Surrogate Variable Analysis (SVA)
- Space PCA (sPCA)

- Both try to capture unobserved technical variation without removing signal being tested

- SVA uses correlation with phenotype to select probes
- sPCA does not (same correction can be used for many phenotypes)

Unobserved Batch Effects



Rahmani et al., Nat Methods 2016

Estimating Unobserved Batch Effects

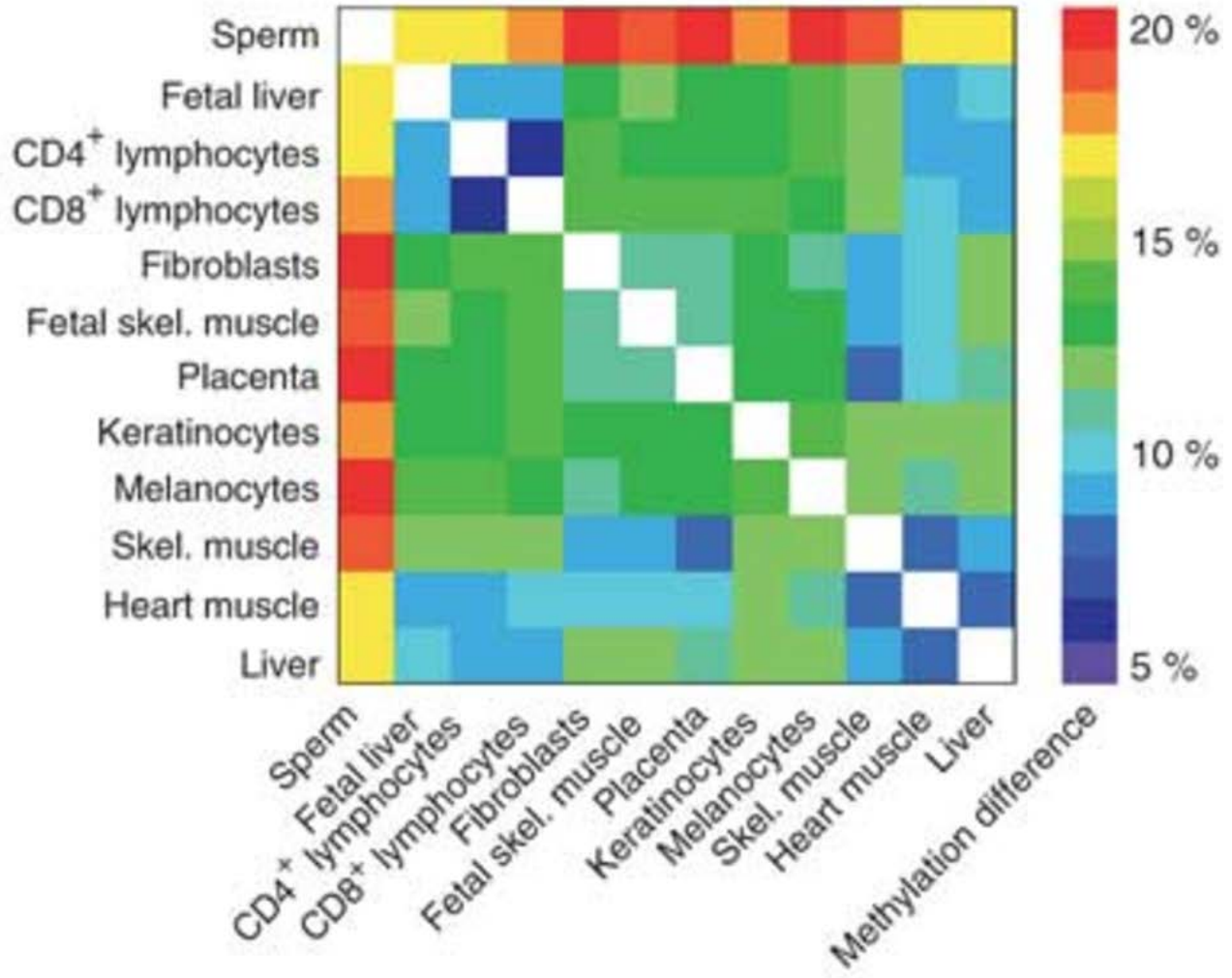
- If we know about a batch/technical effect but do not have data to correct for it, we may be able to estimate it from the data

- E.g. Blood cell counts, age, …

Cell composition

- Methylation plays a large role in cellular differentiation
- Substantial variation across tissue types as well as individual cell types (well demonstrated in WBCs).
- Measured methylation levels represent weighted averages of cell-type-specific methylation levels with weights corresponding to the proportion of the different cell types in a sample.
- Cell-type proportions can vary across individuals, and can be associated with diseases or phenotypes
- Cell composition a potential confounder in MWAS

Cell composition



Eckhardt et al *Nat Gen* 2011

Estimating Blood Cell Counts

- We can “easily” sort blood into its component cell types and measure the DNA methylation differences in each.
- Using the differences of DNA methylation across cell types, we can model the proportion of each cell type in whole blood
- These values can be used as covariates in analyses
- Particularly important in analysis of disease that affect immune function

Estimating Age

- DNA generally becomes more methylated with age
- We can use these changes in DNA methylation with age to make a predictor to estimate a persons age
- Accurate within +/-10 years – so real age preferred!