Data Processing and Quality Control (QC)
Overview

• Main concepts
  • Controls
  • Probe & sample QC
  • Filtering
  • Normalisation & batch effect

• Examples
Assessing controls

• Most platforms for methylation quantification use controls for quality assurances

• Locus-specific assays (e.g. Pyrosequencing)
  • Addition of bisulphite conversion control to dispensation sequence
  • Use of methylated and unmethylated controls
  • Use of negative (no DNA) wells

• Types of controls in Illumina platforms: visualise in genome studio
  
  STAINING CONTROLS                   BISULFITE CONVERSION CONTROLS
  EXTENSION CONTROLS                  SPECIFICITY CONTROLS
  HYBRIDIZATION CONTROLS             NEGATIVE CONTROLS
  TARGET REMOVAL CONTROLS            NON-POLYMORPHIC CONTROLS
  G/T MISMATCH CONTROLS (HUMANMETHYLATION27)
Probe & sample QC

- Locus-specific platforms: perform duplicate or triplicate measures
  - For all samples and all probes:
    - Replicates with large discrepancies in methylation levels are removed
    - Common thresholds being 5-10% difference
    - Mean methylation across replicates can be utilised for the remaining samples
  - Across a subset of samples e.g. 10%, 20%
    - Drop the entire assay if the replicates, on average, show large discrepancies

- Illumina assays: detection p-value for each beta-value
  - Probability that the target sequence signal was distinguishable from the background
  - Common practice: drop individual beta value if detection p-value >0.05
    - Drop probes where median p-value >0.05

- Drop probes that are unsuccessfully measured in n\(^{th}\)% of samples
  - Common thresholds are 20%, 10%, 5%

- Drop samples that failed in n\(^{th}\)% of probes
  - Common thresholds are 50%, 20%
Filtering probes

• Reduce the number of CpG sites taken forward for analysis

• Common practices related to technical issues:
  • Drop those with known SNPs residing in the probe sequence
  • Drop those for which the CpG site contains a SNP
  • Drop those in which probes anneal to multiple genomic locations

• Common practices related to analysis issues:
  • Drop those on X and Y chromosomes
  • Drop those with lowest variation
  • Drop those with extreme methylation levels (e.g. median = 0% or 100%)
  • Only consider those in regions of interest (e.g. CpG island, shore, other)
Non-biological variation

- Variation within measurements caused by technical factors rather than biological differences between samples

- Batch effect: systematic differences across groups of samples

- Causes of technical variation & batch effect include:
  - Differences in sample handling and preparation (e.g. extraction methods)
  - DNA processing (e.g. amplification, labelling, hybridisation)
  - Scanning of arrays/chips (e.g. background noise)
  - Location of sample on chips
  - Technical biases
    - Dye bias
    - Probe bias
Normalisation & batch correction

- Ideally, potential sources of non-biological variation are identified and minimised at the planning & experimental stage
  - Eg. Arrangement of samples on arrays, identical treatment of all samples
- In reality, it’s not possible to remove all experimental artefact
- Normalisation and batch correction techniques aim to identify and remove any remaining non-biological variation by modifying and standardising the measurements
- Not removing potential experimental variation reduces the ability to detect true biological variation.
Normalisation & batch correction

• Current methods (adapted from expression studies) focus on:
  • Batch effect
  • Background noise
  • Dye bias
  • Probe type bias

• Numerous R packages have been developed for this and include:

<table>
<thead>
<tr>
<th>Package</th>
<th>Processes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lumi &amp; Methylumi</td>
<td>Colour adjustment, background correction, normalisation, transformations, QC reports (simple scaling normalisation (ssn), quantile, smooth quantile normalisation, variance-stabilisation normalisation, M-value)</td>
</tr>
<tr>
<td>Minfi</td>
<td>Normalisation, transformations, QC reports &amp; analysis (Background subtraction/normalisation, control normalisation, subset-quantile within array normalisation (SWAN), logit)</td>
</tr>
<tr>
<td>IMA</td>
<td>Type I/II bias adjustment, normalisation, transformations, SNP filtering, QC reports &amp; analysis (Peak-base correction, quantile normalisation, variance-stabilisation normalisation)</td>
</tr>
<tr>
<td>HumMeth27 QCReport</td>
<td>Colour adjustment, normalisation, transformations, QC reports (see Lumi &amp; Methylumi)</td>
</tr>
</tbody>
</table>
Batch effect

- Lumi R package
- Example Dataset
  - Infinium array
  - 4 controls & 4 treatment samples
  - Technical replicates on separate plates
- Batch effect correction following:
  - Background & colour correction
  - SmoothQuantileNormalisation
  - Simple scaling normalisation

Comparison of batch effect before and after correction & normalisation

Taken from Lumi tutorial example dataset
Dye bias in beta values

Comparison of 2 colour channels before and after correction & normalisation

- Lumi R package
- Example Dataset
  - Infinium array
  - 4 controls & 4 treatment samples
  - Technical replicates on separate plates
- Colour correction method
  - smoothQuantileNormalisation

 Taken from Lumi tutorial example dataset
Type I/II bias can create a problem if probes are to be compared and ranked according to absolute differential methylation.

- IMA R package
- Infinium 450k data in a range of tissues and samples
- Peak-based correction

Comparison of beta intensity values between Type I & II probes

Dedeurwaerder et al. Epigenomics, 2011
However!

• The unique characteristics of methylation data is problematic

• Many assumptions for these methods are violated:
  • Variance of methylation is a function of the mean (heteroscedasticity)
  • CpG site density and correlation is not constant across the genome or the chips
  • DNA methylation is associated with CpG density
  • Fluorescence signals, and methylation levels, is influenced by GC content
  • Different probe types, measuring different CpGs, present on one chip
  • DNA methylation shows inter-individual variation

• But, this is a highly active field of research! So we may have some better solutions soon
Quantitative, high-resolution epigenetic profiling of CpG loci identifies associations with cord blood plasma homocysteine and birth weight in humans

- 27k Infinium chip in 12 infant cord blood samples
- Dropped any probe in which at least 1 sample demonstrated a detection p-value >0.05 or null (missing) beta value
- Maximum failure rate in samples was 22%, none dropped
- Eliminated probes on sex chromosomes and those containing known SNPs, deletions & microsats
- Removed non-variable CpG sites i.e. those where beta values for all samples were ≤20% or ≥80%
- Bisulphite sequenced 8 candidate CpG sites in 3 replicates

Fryer et al. Epigenetics, 2011
Analysed all probes in 178 samples using mixed linear regression modelling adjusting for sex and batch, followed by non-parametric tests.

**Example 2**

**Tobacco-smoking-related differential DNA methylation: 27K discovery and replication**

- **27k Infinium chip in 180 adult blood samples**
  - Performed Illumina recommended QC.
  - I assume this includes dropping beta values with detection p-value >0.05 (?)
  - Dropped two samples due to aberrant correlation clustering

- **No probes or samples were eliminated based on success rates but they were reported!**
  - ~88% of probes had complete data, ~98% had ≤ 1 missing value, ~99% had ≤ 2 missing values, only 0.1% of probes had ≥ 12 missing values.
  - Missing values per subject: median (IQR) = 18 (10-30)

- **Analysed all probes in 178 samples using mixed linear regression modelling adjusting for sex and batch, followed by non-parametric tests**

- **Performed subsequent validation & replication by Sequenom**

*Breitling et al.* Am J Hum Genet. 2011
<table>
<thead>
<tr>
<th>QC Process</th>
<th>Locus-specific (e.g. Pyrosequencing, Sequenom)</th>
<th>Microarray (e.g. Illumina 27k, 450k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessing controls (various types depending on platform)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Probe &amp; sample QC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replicates</td>
<td>Ideally N/A</td>
<td>Unlikely Yes</td>
</tr>
<tr>
<td>Detection p-value</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Drop unsuccessful probes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Drop unsuccessful samples</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filtering probes (various options)</td>
<td>N/A</td>
<td>Optional</td>
</tr>
<tr>
<td>Normalisations (For: Batch effect, background noise, dye bias, probe-type bias)</td>
<td>Unlikely</td>
<td>Something is better than nothing!</td>
</tr>
</tbody>
</table>
References

• R: [http://www.r-project.org/](http://www.r-project.org/)
• Cran: [http://cran.r-project.org/](http://cran.r-project.org/)
  • HumMeth27QCReport
• Bioconductor: [http://www.bioconductor.org/](http://www.bioconductor.org/)
  • Minfi, lumi, methylumi
• RForge: [https://www.rforge.net/](https://www.rforge.net/)
  • IMA

