Illumina Infinium 450K Array

Alan Harris

Baylor College of Medicine

Epigenomics Workshop

BCM
Baylor College of Medicine

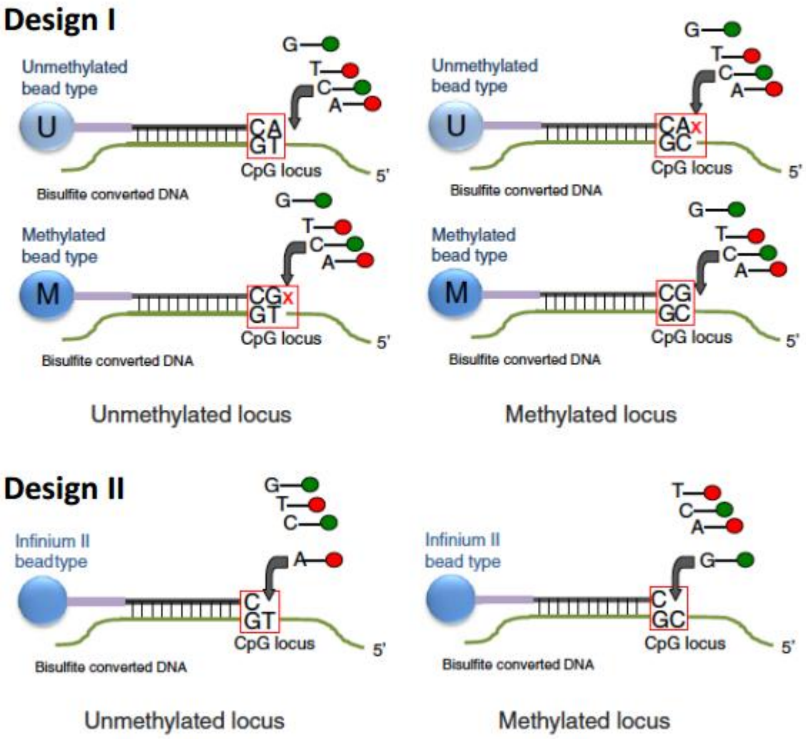
Bioinformatics
Research
Laboratory

Array Design

- 487,557 probes assaying 12 samples
  - CpG 482,421
  - CpH 3,091 – methylated in embryonic stem cells
  - rs SNPs 65

Color of bead types:

350,076 (70%) Both (M,U) Design II
46,298 (10%) Green (M,U) Design I
89,203 (20%) Red (M,U) Design I



Beta Values

- Beta value (β) - estimate of methylation level using ratio of intensities between methylated and unmethylated alleles

β = Methylated allele intensity (M) /
(Unmethylated allele intensity (U) + Methylated allele intensity (M) + 100)

- Genome Studio Methylation Module Normalization
  - **Normalization to internal controls** targeting same region in housekeeping genes with no CpG sites. Intensity multiplied by a constant normalization factor (for all samples) and divided by the average of normalization controls in the probe's channel in the given sample
  - **Background subtraction** derived by averaging the signals of built-in negative control probes

- High correlation with other bisulfite-based data[1]
  - technical replicates - $R^2 > 0.992$
  - 27K BeadChip data - $R^2 > 0.95$ (94% of 27K probes in 450K)
  - whole-genome bisulfite sequencing data - $R^2 > 0.95–0.96$

[1] Bibikova *et al.* (2011) *Genomics* 98:288

Probe Annotations

In Illumina Manifest

- Genomic Coordinates
- UCSC RefGene Name
- UCSC RefGene Accession
- UCSC RefGene Group
- UCSC CpG Islands Name
- Relation to UCSC CpG Island (Island, Shore, Shelf)
- Phantom
- DMR
- Enhancer
- HMM Island
- Regulatory Feature Name
- Regulatory Feature Group

Preprocessing – Convert Beta to M values?

- Comparison of Beta and M values[1]
  - Relationship between Beta-value and M-value is a logit transformation
  - Beta-value method has severe heteroscedasticity for highly methylated or unmethylated CpG sites
  - M-value method provides much better performance in terms of detection rate and true positive rate for both highly methylated and unmethylated CpG sites
  - Beta-value has a more intuitive biological interpretation, but the M-value is more statistically valid

- Software for Beta to M value conversion
  - Lumi[2] (R)
  - Methylumi[3] (R)

[1] Du *et al.* (2010) *BMC Bioinformatics.* 11:587

[2] Du *et al.* (2008) *Bioinformatics.* 24:1547

[3] Davis *et al*. *Bioconductor R package*

Preprocessing - Normalization

- Illumina claims probe design differences do not significantly affect differential methylation detection; can detect delta beta of 0.2 with 99% confidence[1]

- Design I signals more stable and have an extended dynamic range of methylation values compared with design II signals[2]

- Software to normalize between probe designs
  - Illumina Methylation Analyzer (IMA)[3] (R) – peak correction
  - Complete Pipeline[4] (R) – subset quantile normalization
  - BMIQ[5] (R) - beta-mixture quantile normalization

[1] Bibikova *et al.* (2011) *Genomics.* 98:288

[2] Dedeurwaerder *et al.* (2011) *Epigenomics.* 3:771

[3] Wang *et al.* (2012) *Bioinformatics.* 28:729

[4] Touleimat *et al.* (2012) *Epigenomics*. 4:325

[5] Teschendorff *et al*. (2013) *Bioinformatics*. 29:189

Preprocessing – Remove SNPs

•SNPs in probes can lead to incorrect methylation measurements

•File of SNP containing probes can be downloaded from here:
https://www.rforge.net/IMA/snpsites.txt

•91988 cg probes contain SNPs

•Software to remove probes containing SNPs

- Illumina Methylation Analyzer (IMA)[1] (R)
- Genboree Workbench Array Data Importer has option to exclude SNP containing probes

[1] Wang *et al.* (2012) *Bioinformatics.* 28:729

Differentially Methylated Regions

- Detection of statistically significant differentially methylated regions (DMRs) is primary analysis

- Multiple testing correction should be applied to statistical results

- A number of software packages have been developed to identify DMRs

Illumina Methylation Analyzer (IMA)

- Calculates methylation indices for 5’ UTR, first exon, gene body, 3’ UTR, CpG island, CpG shore, CpG shelf
  - Mean
  - Median
  - Tukey’s Biweight robust average

- Identifies DMRs in regions
  - Wilcoxon rank-sum test
  - Student’s t-test
  - Empirical Bayes
  - Generalized linear models

- Multiple Testing Correction
  - Bonferroni
  - False Discovery Rate

Limma

- Originally designed for detecting differential expression from arrays[1]

- Also widely used for Infinium methylation arrays

- Fits a linear model to the data for each gene

- Empirical Bayes method to moderate standard deviations between genes constraining the within-block correlations to be equal between genes

- Accessible through the Genboree Workbench

[1] Smyth. (2004) *Statistical Applications in Genetics and Molecular Biology*. 3, No. 1, Article 3