Expression and Methylation: QC and Pre-Processing

RANDA STRINGER

Epigenetics in NutriGen

- Examine epigenetic mediation of prenatal environment
- Subset of NutriGen cohorts
  - ~500 each from START and CHILD
- Matched samples (where possible)
  - Integrate methylation and expression changes

Expression Data

- Ongoing
- Majority of samples available
  - START = 496
  - CHILD = 467
- QC/pre-processing is underway

Expression QC/Pre-Processing

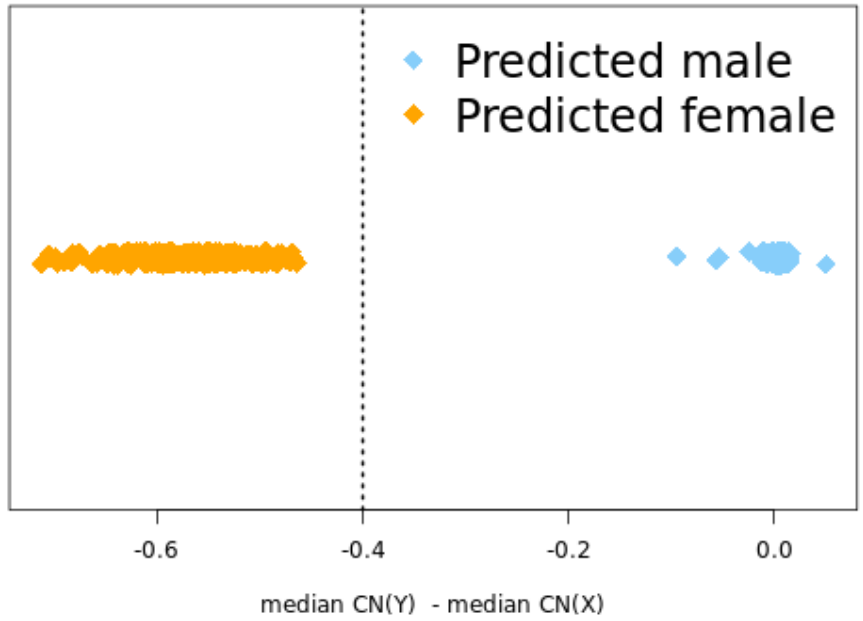
- Quality assessment
  - Probe boxplots (signal distributions, unusual samples)
- Background correction
  - Uses negative controls
- Normalization
- Batch effects
  - MDS plots/ComBat
- Transformation
- Probe filtering
  - Detection $p < 0.01$ in $> 50\%$ of samples

Methylation Data

- In analysis stage
- QC/pre-processing complete*
  - *Pending further advancements in the field
- Good final sample size for both
  - START = 506
  - CHILD = 491

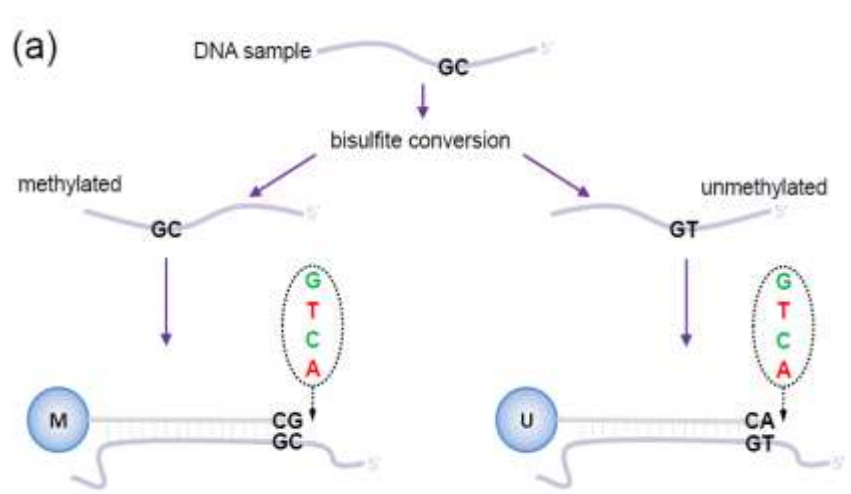
Methylation QC/Pre-Processing

- Sample Quality
  - Compare reported vs. predicted sex
  - Remove samples where proportion of failed probes is > 0.01
- Probe Quality
  - Remove probes that failed to be detected in > 5% of samples
  - Remove cross-hybridizing and polymorphic probes
    - Chen et al. 2013

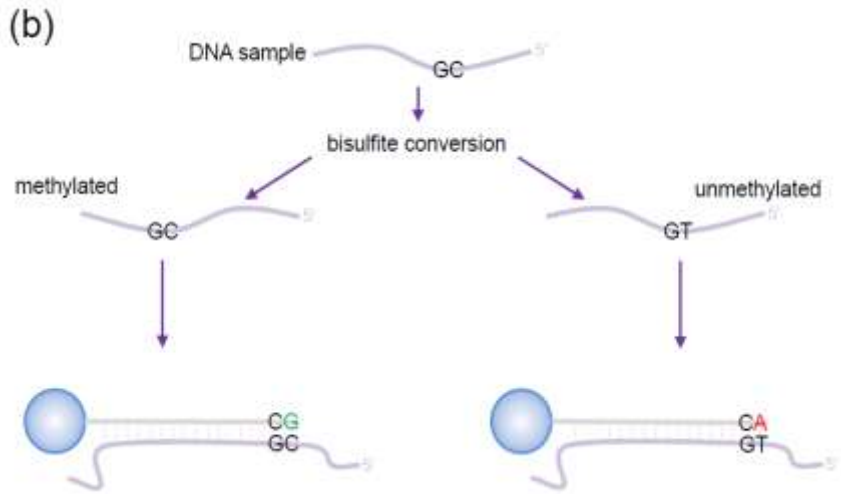


Methylation QC/Pre-Processing

- Normalization
  - 2 probe types with different distributions



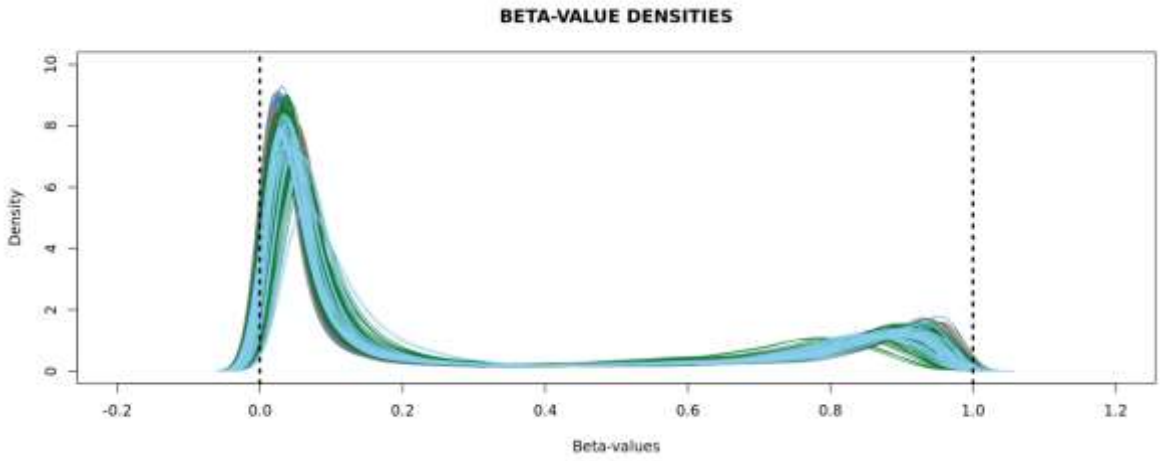
Infinium I Probe
2 different probes per CpG

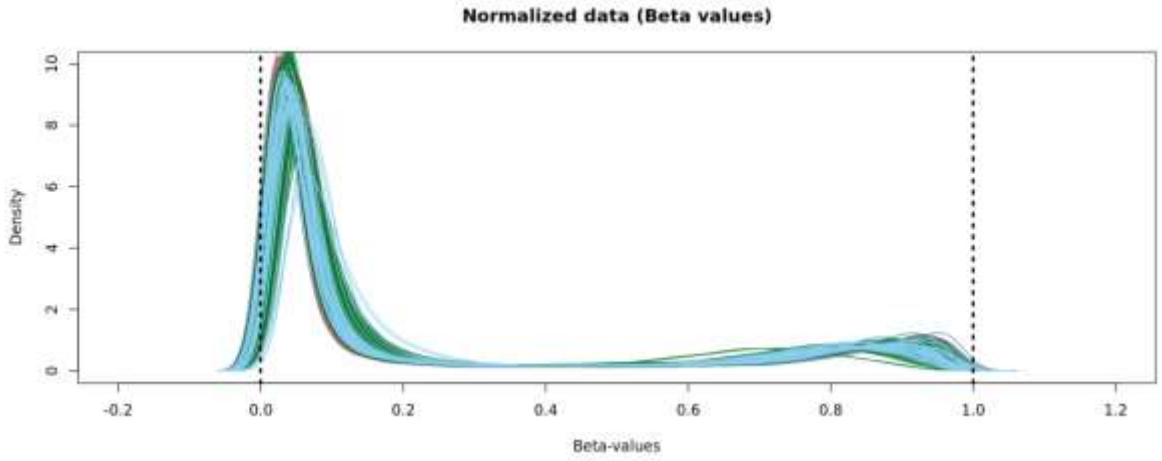


Infinium II Probe
Single base extension at CpG

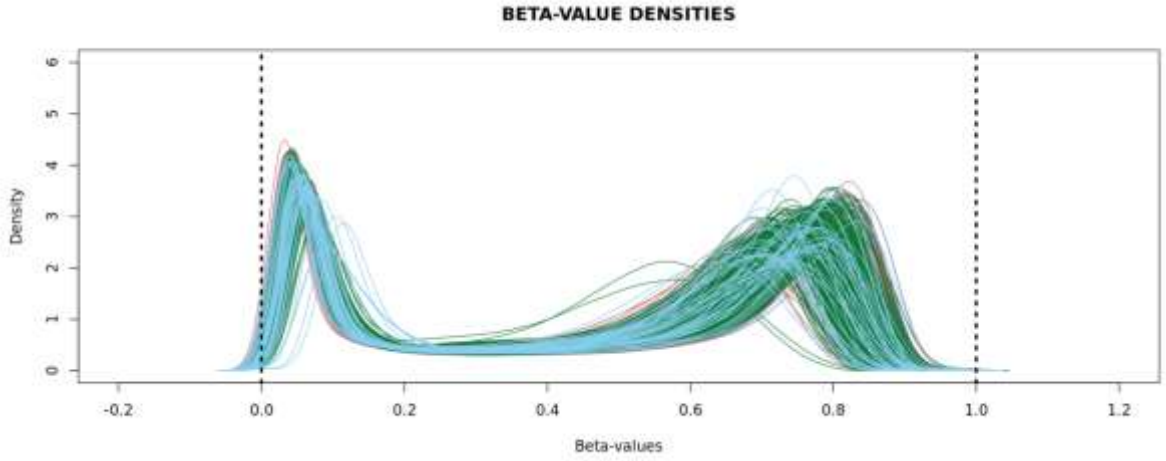
Maksimovic et al. *Genome Biology* 2012

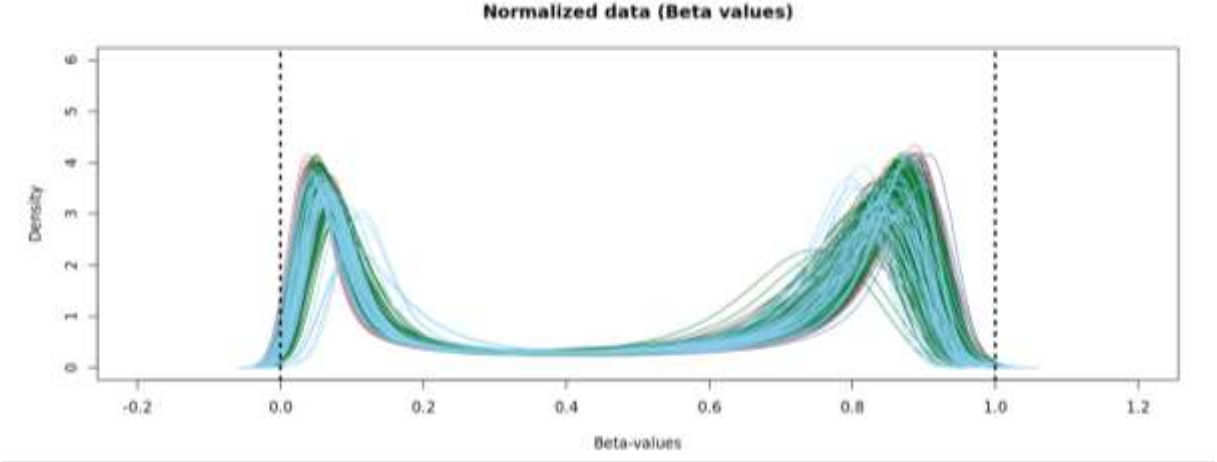
Type I Grn Probes





Type II Probes





Methylation QC/Pre-Processing

- Batch effects
  - Adjust for technical variation
  - Corrected by plate
- Cellular composition
  - Crucial issue in methylation studies
  - Cord blood not well characterized
  - ReFACTor (Rahmani et al., 2016)
    - Reference-free
    - Utilizes PCA

Other Considerations

- Background correction
- Bead count (> 3)
- SNP probe definition
  - MAF > 0.01
- Other normalization methods
  - BMIQ vs SWAN
- Cellular composition adjustment

QC Summary

Samples		
	START	CHILD
Initial	512	511
Sex Check	5	14
Missingness	2	7
Final	506	491

Probes		
	START	CHILD
Initial	> 485 000	
Failed	756	634
Polymorphic	70 889	
Cross-Reactive	29 233	
Final	393 400	393 449

Questions?

Normalization

Goal: reduce non-biological variation

Equalizes probe intensity and signal distributions across arrays and between colour channels

New challenges with DNA methylation vs. gene expression techniques

- Systematic/technical variation
- Novel probe design

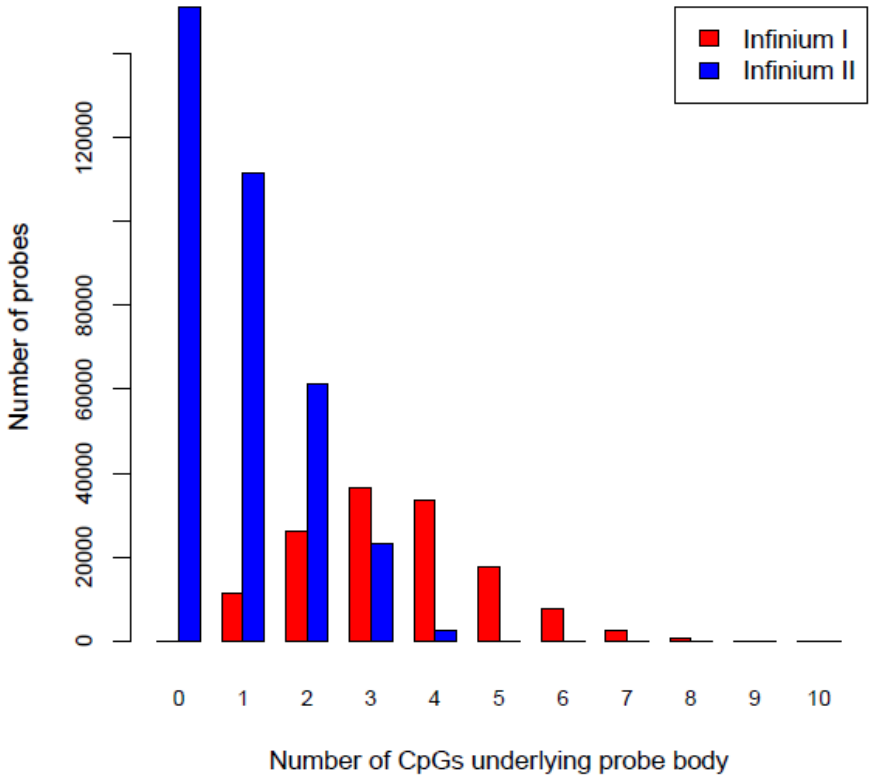
Maksimovic et al. *Genome Biology* 2012

CpG Content

Infinium II ≤ 3 Infinium I ≥ 3

Compressed β value distribution in InfII

Solution: scale Infinium II probes to InfI probes

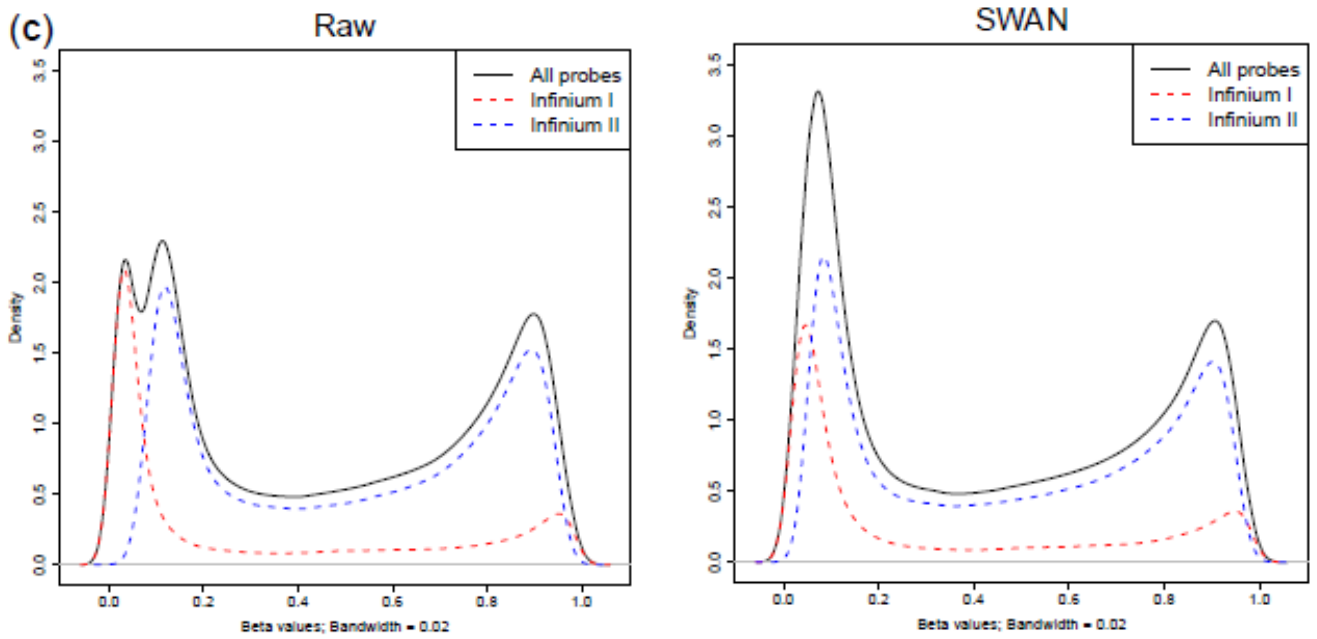


Maksimovic et al. *Genome Biology* 2012

Subset Within-Array Normalization (SWAN)

Allows InfI and InfII probes to be normalized together

- Subset of N InfI and InfII probes chosen based on underlying CpG content
- Separate methylated and unmethylated channels
- Mean intensity for each of 3N calculated
- InfI and II probes adjusted separately by linear interpolation

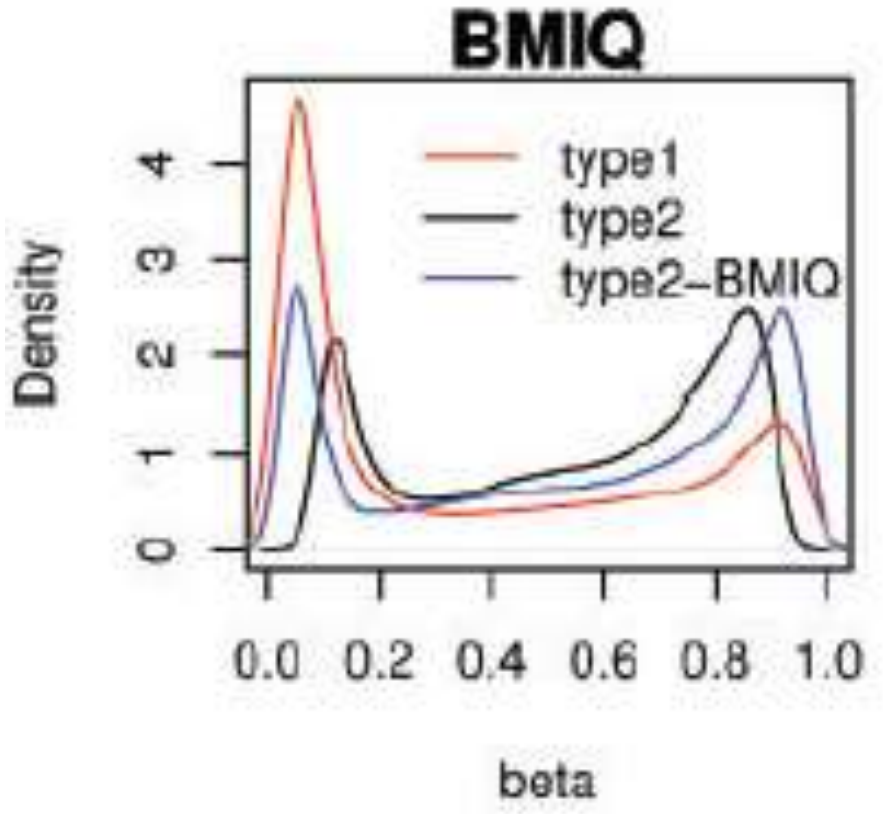


Maksimovic et al. *Genome Biology* 2012

Beta-MIxture Quantile normalization (BMIQ)

Novel normalization method

- Fit 3-state (U/H/M) to InfI and InfII probes separately
- Transform InfI U and M probes using the inverse of the cumulative beta distribution estimated from the respective InfII probes
- For H probes perform dilation transformation to fit the data into the gap



Teschendorff et al. *Bioinformatics* 2012