**Methylation Filtering, Normalization, and Differential Methylation Calculation: Documentation**

1. **FILTERING**
   1. Remove all probes with a high detection p-value (p > 0.05)
   2. Remove cross-reactive probes
      1. WHICH PACKAGE
   3. Remove probes containing common SNPs (likely reflect genotype and not methylation status)
      1. WHICH PACKAGE
2. **NORMALIZATION**
   1. *Within-Array Normalization*
      1. Use Beta Mixture Quantile normalization (BMIQ) to correct for type I/II bias
         1. WHICH PACKAGE
      2. Use methylumi package to correct for color/ dye bias and background correction
   2. *Between-Array Normalization*
      1. Check for batch effects by running a PCA; if batch effects are observed, use ComBat package for batch correction removal

ALTERNATIVE: Using level 3 methylation data preprocessed using the LiftOver pipeline

* **Details of Level 3 Data**
  + GRCH38 Reference Genome, hg38
  + Use Noob for background correction, Linear Dye Bias Correction (Triche 2013) for Dye-Bias correction, Methylumi v.1.8.0 for preprocessing implementation
    - Cross-hybridization masking: Repeat masker overlap with last 15 bp of probe
    - SNP-masking: Mask SNP overlap with last 10 bp of probe
    - dbSNP SNP/INDEL annotation v135
    - P<0.05 negative control probes
  + Make genomic annotations using continuous value for distance to TSS
  + Use GENCODE v22 for gene model

1. **DIFFERENTIAL METHYLATION CALCULATION**
   1. Use the M-model-M-mean setup to calculate Δβ
      1. From the β-value, calculate the M-value
      2. Calculate ΔM as M(tumor) – M(normal)
      3. Calculate as the mean of M-values
         1. All M-values across both samples?
      4. From there, calculate Δβ as

**Notes:** the processing of raw data to Level 3 methylation (a.k.a. “beta”) values was not altered, and thus methylation beta values for individual probes were identical between the hg19 and hg38 versions. The major consequence of the probe remapping was the invalidating of a relatively small number of probes that no longer had a uniquely identifiable location in the hg38 genome (2.0% of probes for the HM27 array and 1.1% of probes for the HM450 array)

**Literature Referenced:**

* A comprehensive overview of Infinium HumanMethylation450 data processing (Dedeurwaerder et al, 2013); for filtering and normalization
* Differential methylation values in differential methylation analysis (Xie et al, 2019); for differential methylation calculation

**Other Strategies:**

* “In calculating the ratio of hyper/hypo-methylated CpGs of the tumor and normal samples, the methylation value(ranging from 0(totally unmethylated) to 1(totally methylated)) larger than 0.7 was defined as hyper-methylation, and the value less than 0.3 was defined as hypo-methylation. For comparison, the ratios of hyper/hypo-methylated CpGs of WGBS data of lung cancers were calculated. We compared the methylation levels of each CpG between tumor samples and the corresponding normal sample data, and defined a CpG site to be differentially methylated (DML) if the q-value of t-test < 0.05 and the absolute difference of methylation value > 0.1. We considered genes whose promoters contain any CpG covered by the original or the expanded methylation data. To identify the genes differentially methylated in each cancer, the methylation status of all the CpG sites covered in promoters were considered. For each promoter, the Fisher’s combined test was used to get the q-value to evaluate whether a gene is differentially methylated. Similar to call DMLs, genes with q-value < 0.05 and mean difference of DNA methylation >0.1 were selected as differentially methylated genes (DMGs).” (<https://www.nature.com/articles/s41525-019-0077-8#Sec11>)
* DORGE (2020) used Used two differential methylation features:
  + Mean methylation level (Beta value) in cancer samples minus that of normal samples, using 450K sequencing data
  + Methylation ratio (mean methylation value cancer / mean methylation value normal)
* SQN simply considers a probe as DMP if the absolute value of the difference between β-value medians of paired samples is higher than 0.2:
  + ∣∣median(βN1,…,βNn)−median(βT1,…,βTn)∣∣≥0.2
  + Generally, DMRs are detected by applying various statistical techniques such as Fisher’s exact test [26, 27], t-test [27], Wilcoxon rank sum test [28] or different regression models [29–31]. (<https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2096-3>)