Pipeline to identify SNPs-density, proximity of CpG sites in a sliding window of 50bps **upstream** and **downstream** (100bp) going upto 5000bps.

The Cancer genome Atlas (https://portal.gdc.cancer.gov) dataset for DNA methylation and SNP-array were downloaded for the cancer patients and processed for downstream for analysis

R-script associated with the pipeline is as below

1. main_run_slidingwindow.R : main script to import different functions combine_data.R, bedfile_generator.R, filter_var_after_overlapSelect.R, extract_data.R, annova_surv.R, kmplot.R
2. combine_data.R: Function to combine DNA methylation and SNP array data across the patients. CpG sites having low variance across patients were filtered.
3. bedfile_generator.R: bedfile were generated for data matrix for DNA methylation, SNPs across the patients.
4. filter_var_after_overlapSelect.R: The function was used to identify all the SNPs that overlaps with CpG site.
5. extract_data.R: The function was used to identify all possible CpG-SNP pairs in proximity of 50bps to a maximum boundary of 5000bps upstream and downstream.
6. annova_surv.R: The function was used to identify SNPs that has significant effect on methylation (meQTLs) and overall survival of the cancer patients.
7. Kmplot.R: Function to generate survival plot.

Output

Mapping of significant CpG-SNP pairs in the identification of meQTLs

Screening of epigenomic modification at high resolution has disclosed a direct correlation between the underlying genetic variation and differential methylation pattern, subsequently defining the presence of meQTLs. In an attempt to identify SNPs genetically influencing methylation pattern, we integrated 905,422 SNPs and 372,626 CpG sites using ucsc tool-overlapSelect. Distribution of CpG-SNP pairs around the CpG site were identified within a base interval of 100-bases and the sliding window of 50- bases extending to the maximum boundary of 5000-bases in the upstream and downstream region. Beta value and the genotype associated with each CpG-SNP pairs were mapped across 731 samples sharing a common interface for SNP array and methylation data. An integrated two-dimensional matrix was generated for each CpG- SNP across the samples, and statistically significant CpG-SNP pairs were mapped based upon non-parametric one-way analysis of variance “ANOVA”. There were a few instances in which multiple SNPs were mapped to a single CpG site. Figure 1 shows a bell-shaped distribution of CpG-SNP pairs by applying a sliding window. From the figure, it is evident that CpG-SNP density is high across 50-bps upstream and

downstream of CpG-site. The overlapselect file constituting 7970 CpG-SNP pairs at 50- bps interval was evaluated for further analysis. The rationale for selecting the loci starting with 50-bases is to minimize the probe effect. Illumina 450K methylation chip is identified to have a "probe effect" i.e SNP within 10bp of the CpG probe may be enriched in methylation quantitative loci (meQTLs). Moreover, DNA methylation locus are primarily associated with promoter regions (besides, inter/intra-genic regions), thus localization of SNP/SNPs may interfere the interaction of DNA methyltransferases enzyme (DNMTs) with CpG loci leading to anomalous DNA methylation.

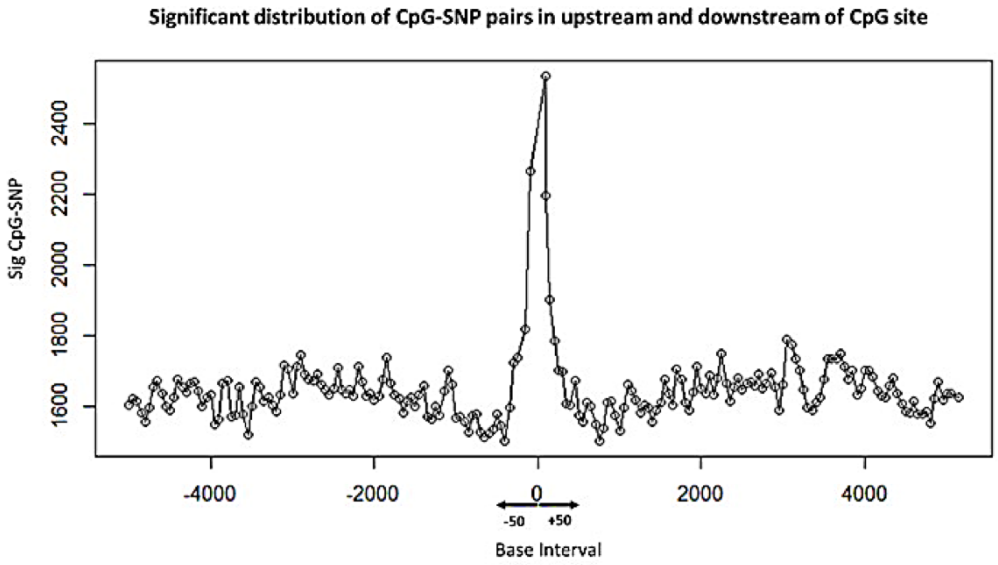

Figure 1: Significant distribution of CpG-SNP pairs around a given CpG site. The CpG-SNP density is identified to be high at 50 bases upstream and the downstream region.