Lab 3: Handling genomic data using Bioconductor

The main purpose of this lab is to introduce basic functionalities of several powerful Bioconductor package through analyzing some features of human genome. The packages to be used are: `Biostrings, GenomicRanges, GenomicFeatures, RMariaDB, BSgenome.Hsapiens.UCSC.hg19` (900Mb). Students should have Bioconductor and above packages installed before the lab.

It was known that in many mammalian genomes, including human, C and G bases are underrepresented and the occurrences of CG dinucleotide (where C and G are at consecutive bases) are depleted. One explanation for the phenomena is DNA methylation. When C and G appear at consecutive bases, C tend to be methylated and the methylated C resembles T in its chemical structure so it's very easily to be mutated to T. Over generations, the methylation-mutation process accumulates, and CG's in many unimportant positions are turned into TG's. The rest of the CGs tends to cluster in small regions called "CpG islands" (CGIs). The canonical definition of CGI looks at following two characteristics of a genomic window:

- GC content: pC+pG, where pC and pG denote the percentages of bases being G or C.
- Observed-to-expected CG ratio (OE ratio): computed at *pCG/(pC*pG)*. Here pCG denotes the percentage of dinucleotides being CG. If the occurrences of C and G are independent, this ratio should be close to 1, e.g., if pC=1/4, pG=1/4, then pCG should be roughly 1/16. This ratio is however only about 0.2 in human genome.

A genomic window is defined as a CGI if its GC content>50% and observed-to-expected CG ratio>0.6. CGIs mark important regions in the genome, for example, over 60% of the gene promoter regions overlap with CGI. In this lab, we will explore these characteristics of human genome. To be specific, we will:

1. Examine the sequence composition of human genome.

2. Explore GC content and OE ratio at gene TSS (transcriptional starting site).

3. Obtain list of CpG islands from UCSC table browser and check their overlap with TSS.

Some homework questions need to be answered from the results generated at this lab.