dna\_methylation

Charles

Last compiled on 30 Juli, 2023

Table of Contents

## Introduction

Epigenetic mechanisms regulate gene expression. Certain technologies enable us measure some epigentic endpoints. One example is DNA methylation. DNA methylation is a chemical process that occurs around genes and which is capable of silencing gene expression. In this project, I will analyse DNA methylation data for the whole genome and relate it to phenotypic variation.

It has been studies that when part of a genome is methylated, the gene close to it is not expressed. It is inherited at mitosis.

Methylation often occurs at CpG islands. A **CpG** is a C, followed by a G but from the end to the end. When one DNA strand is methylated, so is the other. if we also have a CpG on one strand, we have it as well on the other strand.When DNA replicates, its methylation characteristics are preserved.

In this project I will be performing exploratory data analysis on public DNA methylation data.

## Load libraries

library(BSgenome.Hsapiens.UCSC.hg19) # human genome package

## Exploratory data analysis

### Calculate GC content on chromosome 22.

The GC-content are the proportion of bases that are either “G” or “C”.

#### Subset gene sequence of chr22

# subset chr22  
chr22 <- Hsapiens[["chr22"]]  
  
# select start region  
s <- subseq(chr22,   
 start = 23456789,   
 width = 1000)  
print(as.character(s))

## [1] "AGTCACTTGTGCCTGGGTGTGGGGACTAAGCTGTCCATGTGTATCCCACATCAGCCCTCAGCCCCCTCACGAGGGAAGTATGCTTACATATTCTGTGTGTGTGTGGTTATTTTTTTGGAGGGGGGACAGGTTCAGAACAGCTGGCAGAGCAGGAGGTTGGGCCTGCATTGGCCACTCCGCAGCTGCCACACCGGCTCCAGCATTCCCTCCTGCCTGTCCTCACCCACCTTGTTGAGAGCAAGGCTGTCTGCTTGCCTCCATCCCTCACACAAGAGCACCTGGAGAGCCTCTGCCCAGGGAGGTGAGGAAAGGGAGAGGGCAGTGGGGACTGCCATGAAGGCATCTCAGAAGCTGGCAGGGCTGGGGGTTCCAGGTCATCTGTGTCCCAGGGATGATGCTGGTTCCAGGAAGAGCTGAAACCTTAATGTCACGTGCATTTTATGTGAGGTTTGAGGCCCCTAGTTGGGCCAGCTGGCCTTGCTCTGTGCTGTGGCCACAGCAAACTGCATGGGGTCCTGTGGAGTAAGGGACCTAGTGGAGAGTGAGTGGACAGGGAGACCAGATGGGTCAGTACAGAGCCTTCCCCAGCCATGTGGCCATATCACACTGGCATCACCCACCTATTCCCCAATACAACACAGTCCGGGGGCCCCTCACCTGGCTGAGGTATGTGACAAGGCTCTGAGCTGATGGTGTACGTTCAGCCTGAGTAGGTGTTTGCATGTGGTTGCCCCCAAACCAGTGTGATTTGGGGCTGCATGGATCCAGAGTAGGGAGGTGACGGTCCCACCATCTCTGGGGTGTGTGCAAGGGTCAGGGTTTGGGAAGGGGTTGGACAAGACCAGTGTCCCTGAGGACTGCCTACCTGCTGGGAGCTGACATGGGAATCTGGGAGTGCGTCTGGCCTCCTGGAGGGGGTCTCGGGTTGCCCAGAGGGGAGCGAGGAGGAGGCCCAGAACATACCCCTTCCTAGAACAGGAAGGTGGGGTGACCCTGCAGG"

The string above is the DNA sequence 1000 basepairs centered around the point 23456789.

#### Calculate the GC content

cg\_prop <- letterFrequency(s,   
 letters = "CG", # specify string of interest  
 as.prob = TRUE) # specify results output  
cg\_prop

## C|G   
## 0.583

The results show that GC content of the slected resion of chr22 is **58.3%**.

### Calculate the number of CpGs

n\_cg <- countPattern(s,   
 pattern = "CG")  
n\_cg

## [1] 10

There are 10 CpGs in the DNA string of interest on chr22.

### Calculate the number of GpCs

n\_gc <- countPattern(s,   
 pattern = "GC")  
n\_gc

## [1] 65

There are 65 GpCs in the DNA string of interest on chr22.