NAME: K UGENDER
SR.NO: 21008
ugenderk@iisc.ac.in

Color Blindness
Data Analytics Assignment 6

Prog: M-tech
Dept: CSA
2023-11-05

Task

Align reads to the reference sequence with up to two mismatches, count reads mapping to exons of the red and green genes, considering 1 for unambiguous mappings and 1/2 for ambiguous mappings. Interpret 'N' in the reads file as 'A'. Calculate the probabilities of generating these counts for each red-green gene configuration in the presentation and determine the most likely configuration leading to Color Blindness.

1 Implementation

I have set Delta = 1 to improve efficiency and have stored the rank of every element to facilitate this process. To allow for a maximum of two mismatches in read alignment, I then divided each read into three equal parts. If any of these parts have an exact match with the reference sequence, we proceed to the next step.
After finding a partial match, we then check for mismatches between the complete read and the portion of the reference sequence calculated using the exact match of the subpart. If the number of these mismatches is less than or equal to 2, we count the read as a match.
Subsequently, we align the reads to the reference sequence with up to two mismatches and count reads mapping to exons of the red and green genes. For unambiguous mappings to one of the two genes, we count 1 for each read, and for reads that map ambiguously, we assign 1/2 to each gene involved.
To calculate the conditional probabilities of observing the counts based on different configurations, we assume that each read mapped to the two genes is a Bernoulli random variable with 'p' determined by the given configuration.
By comparing the probabilities associated with each configuration, we can select the one with the highest probability, which is the most likely to lead to color blindness. In our analysis, Configuration 3 has the greatest probability of the observed counts, it indicates the maximum likelihood for color blindness.

2 Results

We do not need to align all the reads with the reference sequence. The reads are arranged in increasing order, so the initial reads in the file match with the beginning of the reference sequence, and the later reads correspond to the later part. The reads associated with the red and green genes are located between positions 2936000 and 2948000 in the reads file.

Red Exons	Counts
Exon 1	157
Exon 2	175
Exon 3	109
Exon 4	166
Exon 5	309
Exon 6	425

Table- 1: Counts for Red Exons

Green Exons	Counts
Exon 1	157
Exon 2	256
Exon 3	132
Exon 4	142
Exon 5	363
Exon 6	425

Table- 2: Counts for Green Exons

Configuration	Probability of Count
Configuration 1	1.649×10^{-33}
Configuration 2	0.0
Configuration 3	1.352×10^{-28}
Configuration 4	1.462×10^{-53}

Table- 3: Probability of Count for Different Configurations

Hence, as the probability of observing the given read counts is highest for configuration 3, it is the most likely gene configuration.