Dhanikonda\_Exam2

Nagendra Dhanikonda

2022-11-07

## Question 1

1. The two files are not the same because the checksums are different [nagendra@Nagdhanis-MacBook-Air](mailto:nagendra@Nagdhanis-MacBook-Air) exam\_2 % shasum Sample2Data\_1.txt b6e740582bc1a02cb92bee0b55b9f2bfa1a8763b Sample2Data\_1.txt [nagendra@Nagdhanis-MacBook-Air](mailto:nagendra@Nagdhanis-MacBook-Air) exam\_2 % shasum ‘Sample2Data (1).txt’ cfe3aa7935db51c916e479e97e1d300f7176b625 Sample2Data (1).txt
2. [nagendra@Nagdhanis-MacBook-Air](mailto:nagendra@Nagdhanis-MacBook-Air) exam\_2 % diff -u Sample2Data\_1.txt ‘Sample2Data (1).txt’ There was a line called Pikachurin that was added (indicated by a +) and a line called TAM410 was removed. (indicated by a -) [nagendra@Nagdhanis-MacBook-Air](mailto:nagendra@Nagdhanis-MacBook-Air) exam\_2 % diff -u ‘Sample2Data (1).txt’ Sample2Data\_1.txt — Sample2Data (1).txt 2022-11-02 16:49:21.000000000 -0400 +++ Sample2Data\_1.txt 2022-11-02 16:49:39.000000000 -0400 @@ -123,7 +123,7 @@ GSCACGGCYTAYMATGCMGTCSAGCGGGGTTTTCSGACCWWTCGGCKGAGGGTGASTAATGCGTTGAAGTGCCCGATWRAGGGGGATACCARTTGGAAACCACYGTCAATACCACWTAATGTCTACRGACCAAAGTGTGGGACCTTCKGGCCACATGCTATCSGATGCGCCTACSTGGGATTAGCTAGTTGGTGAGGTAATGGCTCACCAAGGCSACGATCTCTASCTGGTTTGAKAGGATGATCAGCCMCACTGGAACTGAGACACKGTCCACACTCCTACGGGAGGCMSCAGTGGGGAATATTGGACAATGGGCKCMMGCCTGATCCMGCCMTGCCSCSTGTGTGAARAAAGCCTTCGTGTTGTAAAGCACTTTCACCAARGASGAAGGGGTGTGTGTTAATAKTACAACCCTTTGACRTTACTCGTMKAASAATCMCCGGCKGAMTCTGTGCCACCMKCCRCGGTAATACAGAKGGTGCRAGCGTTAATCGGAATTACYGGGCGTAAAACGCAMRTARGCTGTTTTTTAASTCGKATGTGAAASCCCCRGGCTCATCCAGGGAATTGCWTTTGAAACTGGAAAACTAGAGTGTGTGAKAGGGGGGTAGAATTCCWASTGTMRCTRTRAAATGCSTAAARATTTGKAGAAATACCTKTGGCSATGGACGCCCCCGGGCACWACWCTKACTCTCTCGTGCSACAGCSTGAGGAGCRMATMRGAKTAGATMCCCGGGTMCTCCATCCCSTMTGGTATKTCTACTARCTGTTCSTGRTCTTGTACTGTGASTAGCGCATCTAMCSCMCTMWSTAGACCKCCWGTGGASTACTGTCRCAWGATTAWMACTCGTATGACTYGACGGGGKGCMCRCMYAAGCKRTGGAGCATGTGRTTTRMTTCYKWCSCMCCRCTAATAMCCWTACCTRCTCKCGACMTCTACAKWASMSYGCRCTAGCATGCRSWTGTGYYMTTCCGGKARCTSTACGACAGGTRMYGYMTGRCKTKYGTCMKMTCRTGWTRTRMAACTGTTCGKSTTMAGTCCMMGCWGCTAKMSCAMCCCCMWAKCCTYRTTCRCCAGCGCAKCAGGTCGKGGAYYMTGRCGTGASACTSACTGTTSATAATCCCSCAGCAGCGTGCGGACGACGKCAGKYTCGATMCGKGYCCTACTACTAGGTSTACTSAATCGATKCCTTACCYATWGAGCTASGCTWGCG >TAM210 GYGSGCGGSGTACCATGCAGTCGAGGGGTAGARGAAGCWTGCTTCCTTGAKACCGGCGCASGGGTGCGTAACGCGTATGCAATCTACCTTGTACAGGGGGATAGCCCARASAAATTTGGATTAATACCCCATASTGTAATGAATAGGCATCTATTTATTACTAAAGTTCCAACGGTACWWGATGAGCATGCGTCCCATTAGCTAGTTGGTGTGGTAACGGCRCACCRAGGCAACGATGGGTAGGGGTCCTGAGAGGGAGATCCCCCACRCTGGTACTGARACMCGGACCAGACTCCTACGGGAGGCASCARTGAGGAATATTGGTCAATGGACKCMAGTCTGAACCARCCATGCCRCGTGCMGGATGACSGTCCTATGGATTGTAAACTGCTTTTGTACAGGAAGAAACACTCCCTCRTGAGGGAGCTTGACGGTACTGTAAGAATAARGATCGGCTAACTCCRTGCCAGCAGCCRCGGTAATACSGAGGATCCAAGCGTTATCCGGAATCATTGGGTTTAAAGGGTCCGTAGGCGGTTTTATAAGTCWRTGGTGAAATCTGGTCGCTCAACGATCAAACGGCCATTGATACTGTAGAACTTGAATTACTTGGAAGTAACTAGAAYATRTARTGTARCKGTGAAATGCTTAGAGATTACATGGAATACCAATTGCGAAGGCAGGTTACTACAAGTACRTTGACKCTGATGGACGAAAGCGTGGGGAGCGAACAGGATTAGATACCCTGGTAGTCCACGCCSTAWACGATGGATACTAGCTGTTCRGAGCARTCTGAKTGGCTRAGCGAAAGTGATAAGTATCCCACCTGGGGAGTACGTTCGCAAGAATGAAACTCAAASGATTGACGGGGGCCCGCACAAGCRGAKGAGCATGTGGTTTATTCGATGATACGCGAGGAACCTTAYCAAAGCTTAATGGGAGACGACASATWTSGAAACAGATCTTTCTTCGGACGTCTTTCAAGGTGCTGCATGGTTGTCRTCARCTCRTGCCGTGAGGTGTCATGTTAAGTCCTATACYGAKCRCACCCCTGTTGATARCTTGTCAGCGAGGTCATGYCKGACTCTAGTCACGACTGCCAGTGCAGACTGTGAGGATGGTGGGGGATGGACGGTCAACTCGATCRCAGGCGCTTGACGCATGGCTWACCAACGTGCCTCAATGCAAYGGTTACCGWCAGSTAGTCMRRTC ->TAM410 +>Pikachurin CCGCGGRAATCTTAWSATGCAGGTCGAGCGASGKATCACGASCTTGCTCCTTTGAAGTTARCGGCSGACSGGTGASTAACACGTGGGTAACCTACCTATRARACTGGAATAACTTCSGGAAACCGGAGCTAATGCCGGATAACMTATAKAACCGCATGGTTCTWTAGTGAAAGATGGTTTTGCTATCACTTATWKATGGACCCGCGCCGTATTAGCTAKTTGGTAAGGTAAAGGCTTACCMAKGCKACSATACGTRKCCCACCTGAGAGGGTGATCGGCCACACTGGAACTGAGACMCGGTCCASACTCCTACGGGAGGCASCAKTAGGGAATCTTCCSCAATGGGCGAAAGCCTGACRGASCAACGGCKCGTGAKTGATGAAGGTTTTCRGATCGTAAMACTCTRTTATTASGGAAGAACAMATGCSTAARTMACTGTGCKCATCTYGACRGTGCCTAATCWSAAWGCCACGGCTAAYTACRTGCCAKCARCCKCGGTAATACKTAKGTGGCWWGCGTTATCCKGAATTATTGGGCSTAMAGCGCGCGTAGGCSGTTTCYTAWKTCTKATGTKAAARCCCWCGGCTCAACCSTGSASGGTCATTGKAAACTGGGAAWCTTSAGTGCMGAARAGGARAGTGGAATTCCATGTGTAGCGGTGAAATGCKCAGAGATATGGASGAACACCRGTGGCGAACGCGACTTTCTGGTCTGTAACTGACGCTSATGTGCGATAGCGTGSGGATCAAWCACGATTASATACCCTGGTASTCCACGCCGTARACGATGAGTGCTAAGTGTTAGGGGGTTTCCKCCCCTCAGTGCTGCWGCTRWYKCACTAWKCACTCCGMCYGGCGAGTRCGACCGCRWGGTTGTAACTCWCAGGATTTGAMAGSGACCCRCACAMGCKGTGSAGCATGTGRTTTCAAKTCGACGCWCGCGMMKACCTTACCAATCTTGACAKCCTTTGAATACTCTAKAGATAKAKCMTTCSCTTCGKGGGAMAAMGTGACATGTSGTRCACGGTTGTCGTCAGCTMGTRTCRTRAKATKRTGAGGYTAARTCCKTCATCAGCGCAACSCTTWGCTTCAGTGGSCATCATTAMATGGGCGMATSAATGGCTTGACTGCCTGCGACGWCGGAATSAAAGYGGGATGACGTCCAGTCRTSACGCACWTACTGMATCTAGGGTSCTGACACAGAGATGCG >TPS43 CGGGGCGTGCASAMTGATSCTAGGCTCAGCTCATTTARTGTGACACTTGCTWCTARMTWTKAGAGCGGGACRGGGGAGCGTAACGCGTRTGCCMCCTGCCCTTCACTGGGGTATAKCCCGGAAAAATTCGGATTAATCCCCCATARTATTATGACATCKGGTGGTTTTATAATTAAAGGTTACGGTGAAGGATGGGCATGCGTCCTATTASCTAKATGGTGAKGTAACGGCTCACCRTGGCSACRATMRGTARGGGGCCTGARASGGTGGTCCCCCACMCTGGTACTGAGACMCGGACCASACTCCTACSGGAGGCAGCAGTGAGGAATATTGGTCMATGGGCGCAAGCCTGAACCAGCCATGCCSCGTGCMGGAAGAAKGTCCTATGGATTGTAAACTGCTTTTWTCTGGGAATAAACCTCCTTACKTGTAGGGARCTGAAGGTRCCMSATGAATRARCACCGGCWAACTCCSTGCCAKCASCCSCKGTAATACGGAGGGTGCAAGCGTTATCCGGAATCATTGGGTTTAAAKGGTCCGCAKGCGGTCCTATMAGTCAGTGKTGAAAGCCTACWGCTCAACTGTAGAACTGCCATTGATACTGTAGGACTTGAATTCGATMSAAGTGKGCRGAATGTGACATGTARCKKTGAARTGCTTAGATATGTCRCAKAACACCGATRGCGAACGCAGCTCRCTAGGTCTGGATTGACGCTYAKGGACGAAAGCGTGGGGAGCAAACRKGATTAGATACCCTGGTARTCCACGCCGTAAACYATCAATACWCGTTTTCWGKGTCGTRWGACTTCGGAKACTMACCKAAAGTGATAARTATTGCASCTGGGGAGTACSAYCGCRRGGTTGAAACTCAAAGGAATTRACCGGRGGCCRCWCAMGCGGTGGASCATGTGGTTTAATTCTATGATACKCKAKGAAMCTTACCAGGGCTTAAATGCMAASKCATAACGTGGTAAACMTGTWATCWTCAKGACGGTCTGCCATGTGCTGCAWGGMTGATCGTCAGCTCGTGCCGTGACGTSATCGGTTAAGTCCAGAYWAYRGAGCWCAACYCCCYAWCMTTARCTSGCMCAGCKGAATTATGACATGGSGACCYCTAAAGAATRCKGCCCAKCGCAAGCTGASAAMGGAKGGCAAGGAGMCGACGGWCAGGTTGAKCAMKGACCCTTACGTCATGGCCCATACCGAGTGTCSTAGCRCATTSKAGCTA

## Question 2

Using wc -l on unix to find number of lines per file Not Species1Data.txt because has 6305 lines Not Species2Data.txt because has 128251 lines Could be either Species3Data.txt or Species4Data.txt because both have 1596 lines

The correct file is Species4Data.txt because upon using grep to fine Terrapene carolina, this text was found in the file: Similarly , Terrapene carolina induced SMA expression and reduced VE-cadherin levels in RBECs , and these effects could be prevented in the presence of the TGF- beta inhibitor SB-431542 ( XREF\_FIG ) .

## Question 3

An organisms genes are linked to the proteins that they produce. Proteins are then responsible for many different functions and changes in protein function reflect accordingly in the phenotype. For example, in the case of cystic fibrosis, a 3 nucleotide deletion (ΔF508) results in the deletion of a phenylalanine residue, causing improper CFTR protein, folding (Verkman 2012). The lack of functional CFTR, a chloride channel,decreases the amount of chloride ions that are exported from within epithelial cells onto their apical surfaces in passageways such as lungs and pancreatic ducts, which in turn decreases the osmotic gradient, resulting in thick mucus formation in the airways and other passages. This increase in thick mucus is one of the many phenotypes caused by a single mutation in the CFTR gene.

(Verkman 2012) : 10.1016/j.molmed.2011.10.003

## Question 4

library(IRanges)

## Loading required package: BiocGenerics

##   
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:stats':  
##   
## IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':  
##   
## anyDuplicated, append, as.data.frame, basename, cbind, colnames,  
## dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,  
## grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,  
## order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,  
## rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,  
## union, unique, unsplit, which.max, which.min

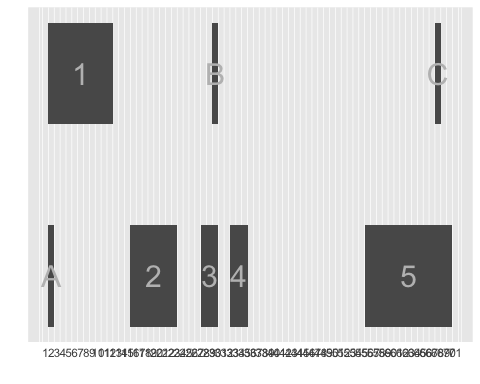
## Loading required package: S4Vectors

## Loading required package: stats4

##   
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:base':  
##   
## expand.grid, I, unname

x <- IRanges(start=c(2, 16, 28, 33, 56, 2, 30, 68), end=c(12, 23, 30, 35, 70, 2, 30, 68))  
names(x) <- c('1','2','3','4','5','A','B','C')  
  
#This code was adapted from Dr. Hansen`s lecture file named RangeAndSequence.Rmd  
  
# Housekeeping - defining boundaries so graph looks nice (don't need to touch)  
sep=0.5  
height=0.5  
# Convert ranges to dataframe for ease of reference  
out <- as.data.frame(x)  
out$y <- disjointBins(x)  
out$ymin <- out$y \* (sep + height) - height  
out$ymax <- out$ymin + height  
xmin <- min(out$start)  
xmax <- max(out$end)  
xbreaks <- seq(xmin - 1L, xmax + 1L)  
  
# Plot the ranges  
ggplot(out) + geom\_rect(aes(ymin=ymin, ymax=ymax, xmin=start-0.5, xmax=end+0.5)) + geom\_text(aes(x=start + width/2 - 0.5, y=ymin+(ymax-ymin)/2, label=names), size=8, color="grey") + scale\_x\_continuous(breaks=xbreaks) + xlab("") + ylab("") + theme(axis.text.y=element\_blank()) + theme(panel.grid.major=element\_blank(), panel.grid.minor.y=element\_blank(), axis.ticks=element\_blank())



#As seen on the graph, there is an overlap of SNP A in gene1, B in gene 3, and C in gene 5. Thus we can conclude that SNP A affects gene1, B affects gene 3, and C affects gene 5.

## Question 5

library(GenomicRanges)

## Loading required package: GenomeInfoDb

gr <- GRanges(seqname=c("chr1", "chr1", "chr2", "chr3","chr1", "chr6", "chr2", "chr3","chr1", "chr4", "chr2", "chr4","chr10", "chr1", "chr2", "chr6","chr1", "chr", "chr6", "chr3"),  
ranges=IRanges(start=c(2, 3, 5, 7, 9, 12, 13, 16, 18, 22, 23, 26, 30, 36, 38, 41, 46, 56, 57, 59), end=c(19,19,18,26,13,24,21,31,35,25,26,44,33,39,54,50,57,60,74,74)),  
strand=c("+", "-", "-", "+","+", "-", "-", "+","+", "-", "-", "+","+", "-", "-", "+","+", "-", "-", "+"),identifier=letters[1:20])  
  
width(gr)

## [1] 18 17 14 20 5 13 9 16 18 4 4 19 4 4 17 10 12 5 18 16

gr[4]

## GRanges object with 1 range and 1 metadata column:  
## seqnames ranges strand | identifier  
## <Rle> <IRanges> <Rle> | <character>  
## [1] chr3 7-26 + | d  
## -------  
## seqinfo: 7 sequences from an unspecified genome; no seqlengths

# From the width function, it can be seen that the largest width range is 20 and it is in the 4th position in the list of widths. The 4th position in the gr grange was indexed which revealed that the range was on chr3 and the identifier character is d.

## Question 6

#FASTA import code adapted from Dr. Hansen`s lecture file named RangeAndSequence.Rmd  
  
df <- read.table('/Users/nagendra/Desktop/Bioinformatics/Test\_submissions/Test\_2/Exam\_2/Forest\_Seqs.fasta',header=FALSE,sep="|",fill=TRUE)  
# Found out about the stringr package and str\_detect at: https://www.statology.org/r-check-if-column-contains-string/  
  
library(stringr)  
#Empty vector made  
Fasta<- c()  
#For each row in dataframe df, if the column contains 'MPM' add the row to the Fasta vector as well as the row below it. Else, go to the next line  
for( i in 1:length(df$V1)){  
 if(str\_detect(df$V1[i], 'MPM')){  
 Fasta<- append(Fasta, df$V1[i]) ; Fasta <- append(Fasta, df$V1[i+1])  
 }else{  
 next  
 }  
}  
#save the Fasta vector as a dataframe  
MPMData <- data.frame(Fasta)  
  
#save dataframe as a text file  
  
write.table(MPMData,'/Users/nagendra/Desktop/Bioinformatics/Test\_submissions/MPMdata.txt')

## Question 7

#code to import FASTQ files and convert quality scores from ASCII to PHRED was adapted from Dr. Hansen`s lecture file named RangeAndSequence.Rmd  
  
fastq <- read.table('/Users/nagendra/Desktop/Bioinformatics/Test\_submissions/Test\_2/Exam\_2/Gene\_compare.fastq',header=FALSE,sep="|",fill=TRUE)  
  
name <- seq(1, length(fastq$V1), by=4)  
seq <- seq(2, length(fastq$V1), by=4)  
score <- seq(4, length(fastq$V1), by=4)  
  
Name <- fastq$V1[name]  
Sequence <- fastq$V1[seq]  
Score\_char <- fastq$V1[score]  
Score\_val = c()  
  
fastq\_data <- data.frame(Name, Sequence, Score\_char)  
  
for (x in 1:length(fastq\_data$Name)){  
 score\_num = c()  
 char <- unlist(strsplit(fastq\_data$Score\_char[x],""))  
 for (a in char){  
 score\_num <- append(score\_num,utf8ToInt(a)-33)  
 }  
 Score\_val <- append(Score\_val,list(score\_num))  
}  
  
for (i in 1:(length(Score\_val))){  
 fastq\_data$Score\_val[i] <- list(Score\_val[i])  
}  
#sample 1 scores   
fastq\_data[1,4]

## [[1]]  
## [[1]][[1]]  
## [1] 40 41 41 40 41 40 37 39 39 38 38 38 40 39 39 39 37 37 35 36 36 34 34 34 35  
## [26] 34 34 34 35 34 34 35 35 35 34 35 35 35 35 35 35 35 34 34 33 33 31 30 33 33  
## [51] 33 27

# sapmle 2 scores  
fastq\_data[2,4]

## [[1]]  
## [[1]][[1]]  
## [1] 40 40 40 40 33 39 40 40 40 40 40 40 39 39 39 39 37 39 36 36 35 34 32 32 29  
## [26] 7 26 31 32 33 33 34 34 34 34 33 33 33 33 33 34 34 25 34 34 34 32 27 32 34  
## [51] 32 33

#A higher PHRED score indicates higher probability that the nucleotide was identified correctly. It can be seen that the PHRED scores for Sample 2 are lower in the majority of corresponding nucleotides when compared to sample 1. Thus, we can conclude that sample 1 is of higher quality due to the higher certainty of the sequencing.