Welcome to The Carpentries Etherpad!

This pad is synchronized as you type, so that everyone viewing this page sees the same text. This allows you to collaborate seamlessly on documents.

Use of this service is restricted to members of The Carpentries community; this is not for general purpose use (for that, try https://etherpad.wikimedia.org).

Users are expected to follow our code of conduct: https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html

---

https://nadinebestard.github.io/2023-10-23-CRM/

data4Carp

ec2-54-165-88-227.compute-1.amazonaws.com - Nadine
ec2-100-26-155-211.compute-1.amazonaws.com - becky
ec2-34-229-56-237.compute-1.amazonaws.com - Alessia
ec2-35-175-63-153.compute-1.amazonaws.com -Sonia
ec2-54-87-158-131.compute-1.amazonaws.com - Katie
ec2-54-175-146-140.compute-1.amazonaws.com - Richardmd
ec2-35-153-16-155.compute-1.amazonaws.com

In SRR098026.fastq,
What are the next three nucleotides (characters) after the first instance of TTTTT?
What is the last line of the file?

1. Search for the sequence GNATNACCACTTCC in the SRR098026.fastq file. Have your search return all matching lines and the name (or identifier) for each sequence that contains a match.

2. Search for the sequence AAGTT in both FASTQ files. Have your search return all matching lines and the name (or identifier) for each sequence that contains a match.

```
cd ~/dc_workshop/data/untrimmed_fastq
curl -O ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/004/SRR2589044/SRR2589044_1.fastq.gz
curl -O ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/004/SRR2589044/SRR2589044_2.fastq.gz
curl -O ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/003/SRR2584863/SRR2584863_1.fastq.gz
curl -O ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/003/SRR2584863/SRR2584863_2.fastq.gz
curl -O ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/006/SRR2584866/SRR2584866_1.fastq.gz
curl -O ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR258/006/SRR2584866_2.fastq.gz
```

```
ls  ~/Desktop/
```

```
mkdir ~/Desktop/fastqc_html
 scp "dcuser@ec2-54-165-88-227.compute-1.amazonaws.com:~/dc_workshop/results/
fastqc_untrimmed_reads/*html" ~/Desktop/fastqc_html

 ~/.miniconda3/pkgs/trimmomatic-0.38-0/share/trimmomatic-0.38-0/adapters/NexteraPE-PE.fa

 trimmomatic PE SRR2589044_1.fastq.gz SRR2589044_2.fastq.gz \
          SRR2589044_1_trim.fastq.gz SRR2589044_1_un_trim.fastq.gz \

 trimmomatic PE SRR2589044_1.fastq.gz SRR2589044_2.fastq.gz \
          SRR2589044_1_trim.fastq.gz SRR2589044_1_un_trim.fastq.gz \
           SRR2589044_2_trim.fastq.gz SRR2589044_2_un_trim.fastq.gz \
           SLIDINGWINDOW:4:20 MINLEN:25 ILLUMINACLIP:NexteraPE-PE.fa:2:40:15



 for filename1 in *_1.fastq.gz
do

   base=$(basename ${filename1} _1.fastq.gz)

  trimmomatic PE ${base}_1.fastq.gz ${base}_2.fastq.gz \
           ${base}_1_trim.fastq.gz ${base}_1_un_trim.fastq.gz \
           ${base}_2_trim.fastq.gz ${base}_2_un_trim.fastq.gz \
           SLIDINGWINDOW:4:20 MINLEN:25 ILLUMINACLIP:NexteraPE-PE.fa:2:40:15
           done

 ${base}_1

 SRR2584866
```

R day

```
 snp_genes <- c("OXTR", "ACTN3", "AR", "OPRM1", "CYP1A1", NA, "APOA5")
 snp_positions <- c(8762685, 66560624, 67545785, 154039662)
 snps <- c("rs53576", "rs1815739", "rs6152", "rs1799971")
snp_chromosomes <- c("3", "11", "X", "6")
```

REVIEW EXERCISE 1

What data modes are the following vectors?
a. snps
b. snp_chromosomes
c. snp_positions

REVIEW EXERCISE 2
To the snps vector add: “rs662799”

EXERCISE3
Your vector should look like this in ‘Environment’: chr [1:7] "OXTR" "ACTN3" "AR" "OPRM1" "CYP1A1" NA "APOA5".
If not recreate the vector by running this expression: snp_genes <- c("OXTR", "ACTN3", "AR", "OPRM1", "CYP1A1", NA, "APOA5")
Create a new version of snp_genes that does not contain CYP1A1

REVIEW EXERCISE 4
Using indexing, create a new vector named combined that contains:

- The the 1st value in snp_genes
- The 1st value in snps
- The 1st value in snp_chromosomes
- The 1st value in snp_positions

EVIEW EXERCISE 5
What type of data is combined?

Work Localy

https://figshare.com/articles/dataset/Data_Carpentry_Genomics_beta_2_0/7726454

```
git config --global user.email "you@example.com"
git config --global user.name "Your Name"
```

```
#load the data.
variants <- read_csv(file.path("data/combined_tidy_vcf.csv"))
```

```
# Starting with the variants data frame, use pipes to subset the data to include
# only the columns REF, ALT, and POS andonly observations from SRR2584863 sample,
# where the filtered depth (DP) is at least 10.
```

PLOTS

```
# FORMULA.
ggplot(data = <DATA>, mapping = aes(<MAPPINGS>)) +  <GEOM_FUNCTION>()
```

```
# Dplyr  + Ggplot
variants %>%
```

```
  filter(sample_id %in% c("SRR2584863", "SRR2589044" )) %>%
ggplot(aes(x = POS , y=MQ, color = sample_id)) +
  geom_point() +
  scale_color_discrete(type = c("blue", "red")) +
  labs(x= "position in the genome",
       y = "mapping quality (MQ)")
```

Install RNASEQ workflow packages:

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install("rnaseqGene")
```

Other option:

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install(c( "BiocStyle", "airway", "tximeta", "magrittr", "DESeq2", "apeglm", "vsn",
"dplyr", "ggplot2", "hexbin", "pheatmap", "RColorBrewer", "PoiClaClu", "glmpca", "ggbeeswarm",
"genefilter", "AnnotationDbi", "org.Hs.eg.db", "ReportingTools", "Gviz", "sva", "RUVSeq", "fission"))
```

www.bioconductor.org/packages/3.17/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html

```
 colorRampPalette( rev(brewer.pal(9, "Blues")) )(255)
```

```
 sample_dist <- dist(t(assay(rld)))
sample_dist_matrix <- as.matrix(sample_dist)
rownames(sample_dist_matrix) <- paste(rld$cell, rld$dex)
colnames(sample_dist_matrix) <- paste(rld$cell, rld$dex)

pheatmap(sample_dist_matrix,
         clustering_distance_rows = sample_dist,
         clustering_distance_cols = sample_dist,
         color = colorRampPalette( rev(brewer.pal(9, "Blues")) )(255)
         )
```

```
ggplot(data = pca_data, mapping = aes(x = PC1, y = PC2, colour = dex, shape = cell)) +
  geom_point(size = 4) +
  ggtitle("PCA plot for Dexhamatosone treatment") +
  labs( x = paste("PC1:", variance[1],  "% of variance"),
        y = paste("PC2:", variance[2],  "% of variance"),
        colour = "Treatment",
        shape = "Cells donor")
```

```
    var_genes <- head(order(rowVars(assay(rld)), decreasing = TRUE),20)

matrix <- assay(rld)[var_genes, ]
annotation <- as.data.frame(colData(rld)[, c("cell","dex")])
pheatmap(matrix, annotation_col = annotation)
```