MICB 405 Project 2

Shannah

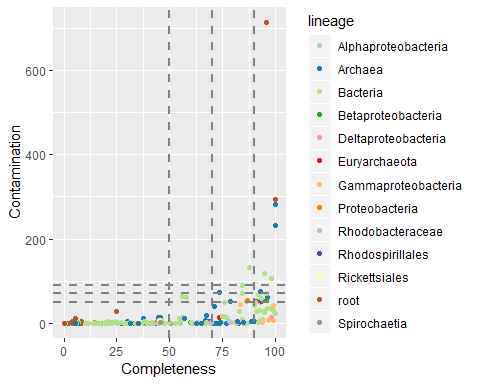
November 17, 2018

## Load the data and put it in a format that I need for plotting

#--- GETTING THE DATA ----  
# Load raw data  
raw\_dat <- readr::read\_csv("Saanich\_Data.csv")  
  
#Clean dat  
dat <-   
 raw\_dat %>%  
 dplyr::filter(!is.na(WS\_O2)) %>%  
 dplyr::rename(O2\_uM=WS\_O2, NO3\_uM=WS\_NO3, H2S\_uM=WS\_H2S) %>%  
 dplyr::mutate(Depth\_m=Depth\*1000)  
  
# -- rpkm data --  
#Read in KO Annotation tables  
ko <- read.table("SaanichInlet\_MAGs\_ORFs\_ko.cleaned.txt") %>%   
 dplyr::rename(orf = V1) %>%   
 dplyr::rename(ko = V2)  
  
#Read in all rpkm files from each cruise in August  
rpkm\_48 <- read.table("SI048\_200m.RPKM.csv", header=FALSE, sep=',') %>%   
 dplyr::rename(orf = V1) %>%   
 dplyr::rename(rpkm = V2)  
rpkm\_72 <- read.table("SI048\_200m.RPKM.csv", header=FALSE, sep=',') %>%   
 dplyr::rename(orf = V1) %>%   
 dplyr::rename(rpkm = V2)  
rpkm\_73 <- read.table("SI073\_200m.RPKM.csv", header=FALSE, sep=',') %>%   
 dplyr::rename(orf = V1) %>%   
 dplyr::rename(rpkm = V2)  
rpkm\_74 <- read.table("SI074\_200m.RPKM.csv", header=FALSE, sep=',') %>%   
 dplyr::rename(orf = V1) %>%   
 dplyr::rename(rpkm = V2)  
rpkm\_75 <- read.table("SI075\_200m.RPKM.csv", header=FALSE, sep=',') %>%   
 dplyr::rename(orf = V1) %>%   
 dplyr::rename(rpkm = V2)  
  
#Read in prokka MAG map  
prokka\_mag\_map <- read.table("Prokka\_MAG\_map\_basename.csv", header=F, sep=',') %>%   
 dplyr::rename(prokka\_id = V1) %>%   
 dplyr::rename(mag = V2)  
  
#Read in gtdbtk files  
arc\_class <- read.table("gtdbtk.ar122.classification\_pplacer.tsv", sep="\t")  
bac\_class <- read.table("gtdbtk.bac120.classification\_pplacer.tsv", sep="\t")  
  
#Combine archaea and bacteria gtdb files  
gtdb\_dat <- rbind(arc\_class, bac\_class) %>%   
 dplyr::rename(mag = V1) %>%   
 separate(V2, sep=';', into=c("Kingdom", "Phylum", "Class", "Order", "Family", "Genus", "Species"))  
  
#combine all cruises and ko values  
ko\_rpkm\_48 <- left\_join(ko, rpkm\_48, by="orf") %>%   
 separate(orf, into=c("prokka\_id", "orf\_id")) %>% # Split the Prokka ORF names into MAG identifier and ORF  
 left\_join(prokka\_mag\_map, by="prokka\_id") %>%   
 left\_join(gtdb\_dat, by="mag")  
ko\_rpkm\_72 <- left\_join(ko, rpkm\_72, by="orf") %>%   
 separate(orf, into=c("prokka\_id", "orf\_id")) %>% # Split the Prokka ORF names into MAG identifier and ORF  
 left\_join(prokka\_mag\_map, by="prokka\_id") %>%   
 left\_join(gtdb\_dat, by="mag")  
ko\_rpkm\_73 <- left\_join(ko, rpkm\_73, by="orf") %>%   
 separate(orf, into=c("prokka\_id", "orf\_id")) %>% # Split the Prokka ORF names into MAG identifier and ORF  
 left\_join(prokka\_mag\_map, by="prokka\_id") %>%   
 left\_join(gtdb\_dat, by="mag")  
ko\_rpkm\_74 <- left\_join(ko, rpkm\_74, by="orf") %>%   
 separate(orf, into=c("prokka\_id", "orf\_id")) %>% # Split the Prokka ORF names into MAG identifier and ORF  
 left\_join(prokka\_mag\_map, by="prokka\_id") %>%   
 left\_join(gtdb\_dat, by="mag")  
ko\_rpkm\_75 <- left\_join(ko, rpkm\_75, by="orf") %>%   
 separate(orf, into=c("prokka\_id", "orf\_id")) %>% # Split the Prokka ORF names into MAG identifier and ORF  
 left\_join(prokka\_mag\_map, by="prokka\_id") %>%   
 left\_join(gtdb\_dat, by="mag")  
  
#Merge all cruises except for cruise 42 (from february)  
#All cruises are from August except for Cruise 42 and b/c seasons can dramatically affect the water column AND we don't have enough data points to compare across seasons (only one from February) we are removing the February rpkm value.  
ko\_rpkm <- rbind(ko\_rpkm\_48, ko\_rpkm\_73, ko\_rpkm\_74, ko\_rpkm\_75)  
  
#Get all phylum  
lo\_phylum <- unique(ko\_rpkm$Phylum)

First I will analyse the MAGs from checkM output with rpkm abundance values

#TODO: include dotted lines to show ranges  
#TODO: include the units on each axis (should be %)  
#TODO: convert contamination to percentage ?!?!  
#TODO: include a None lineage?  
#TODO: include range labels  
  
#Probably want to include as many MAGs as possible because we are wanting to fill in a pathway  
  
#Load .tsv file from MetaBAT2 to dat  
bin\_dat <- read\_tsv(file="MetaBAT2\_SaanichInlet\_200m\_min1500\_checkM\_stdout.tsv", col\_names = TRUE)  
  
  
#Rename Marker lineage so it's one string (makes it easier later)  
bin\_dat <- bin\_dat %>%  
 dplyr::rename('Marker\_lineage' = 'Marker lineage')  
  
#Add new column with just lineage for point colors  
bin\_dat <- bin\_dat %>%  
 dplyr::mutate('lineage' = gsub(".\*\_\_(\*)", "\\1", Marker\_lineage),  
 'lineage' = gsub("\\(|\\)", "", lineage),  
 'lineage' = gsub("UID.\*", "", lineage))  
  
  
#Plotting Completeness vs Contamination of MAGs at 200m  
bin\_dat %>%  
 ggplot(aes(x=Completeness, y=Contamination, colour=lineage)) +  
 geom\_point() +  
 #added discrete scales ; palette from colorbrewer2 but added extra ("#969696") b/c 13 values  
 scale\_color\_manual(values = c('#a6cee3','#1f78b4',"#b2df8a","#33a02c","#fb9a99",  
 "#e31a1c",'#fdbf6f',"#ff7f00","#cab2d6","#6a3d9a",  
 "#ffff99","#b15928", "#969696")) +   
 geom\_vline(xintercept = 50, linetype = "dashed", size = 1, color = "#808080") +   
 geom\_vline(xintercept = 70, linetype = "dashed", size = 1, color = "#808080") + #TODO: check what this value is  
 geom\_vline(xintercept = 90, linetype = "dashed", size = 1, color = "#808080") + #TODO: check what this value is  
 geom\_hline(yintercept = 50, linetype = "dashed", size = 1, color = "#808080") + #TODO: check what this value is  
 geom\_hline(yintercept = 70, linetype = "dashed", size = 1, color = "#808080") + #TODO: check what this value is  
 geom\_hline(yintercept = 90, linetype = "dashed", size = 1, color = "#808080") #TODO: check what this value is



# Next I will create RPKM bubble-plot of each Nitrogen/Sulphur-cycling gene versus taxonomy

\*Geochemical Gradients

# Load raw data  
raw\_dat <- readr::read\_csv("Saanich\_Data.csv")  
  
#Clean dat  
dat <-   
 raw\_dat %>%  
 dplyr::select(Cruise, Date, Depth, Temperature,  
 WS\_O2, WS\_NO3, WS\_H2S) %>%  
 dplyr::filter(!is.na(WS\_O2)) %>%  
 dplyr::rename(O2\_uM=WS\_O2, NO3\_uM=WS\_NO3, H2S\_uM=WS\_H2S) %>%  
 dplyr::mutate(Depth\_m=Depth\*1000)  
  
  
#want to create a plot with all 3 variables as x axis and depth as y with shape as type  
#can manipulate data frame to do so  
dat %>%  
 dplyr::select(Depth\_m, H2S\_uM, NO3\_uM, O2\_uM) %>%  
 gather(key = "Chemical", value = "Concentration", -Depth\_m) %>%  
 ggplot() +  
 geom\_point(aes(x=Concentration, y=Depth\_m, shape=Chemical, colour=Chemical))

