Fiber_Comparison

Jacob T. Nearing Ph.D.

2023-09-14

Load in ParathaakSGB_table

```
parathaa_ksgb_assignments <- read.table("~/Repos/Parathaa2_OP3/Datasets/fiber_amplicons_true/taxonomic_a

parathaa_abundance_table <- read.table("~/Repos/Parathaa2_OP3/Datasets/fiber_amplicons_true/taxonomic_as

colnames(parathaa_abundance_table)[1] <- "query.name"

parathaa_abundance_table[,-1] <- sweep(parathaa_abundance_table[,-1], 2, colSums(parathaa_abundance_tabl

parathaa_abundance_table_kSGB <- full_join(parathaa_ksgb_assignments, parathaa_abundance_table, by="quer

#remove control samples
parathaa_abundance_table_kSGB <- parathaa_abundance_table_kSGB[,-c(256:263, 10)]
```

Load in MetaPhlAn4 table

```
m4_assignments <- read.table("~/Repos/Parathaa2_OP3/Datasets/fiber_amplicons_true/taxonomic_assignments_

colnames(m4_assignments) <- gsub("_.*", "", colnames(m4_assignments))

indexs <- match(colnames(parathaa_abundance_table), colnames(m4_assignments))
##ignore NAs for the moment..

indexs <- indexs[!is.na(indexs)]


m4_filt <- m4_assignments[,c(1,indexs)]
m4_filt <- m4_filt[-which(rowSums(m4_filt[,-1])==0),]

spec_index <- grep("s__", m4_filt$X..taxonomy)

SGB_index <- grep("t__", m4_filt$X..taxonomy)
only_spec <- setdiff(spec_index, SGB_index)


m4_filt$Species <- gsub(".*s__", "", m4_filt$X..taxonomy)
m4_filt$Species <- gsub("\\|.*", "", m4_filt$Species)

m4_filt$Genus <- gsub(".*g__", "", m4_filt$X..taxonomy)
m4_filt$Genus <- gsub("\\|.*", "", m4_filt$Genus)

m4_filt$Family <- gsub(".*f__", "", m4_filt$X..taxonomy)
m4_filt$Family <- gsub("\\|.*", "", m4_filt$Family)
```

```
m4_filt$Order <- gsub(".*o__", "", m4_filt$X..taxonomy)
m4_filt$Order <- gsub("\\|.*", "", m4_filt$Order)

m4_filt$Class <- gsub(".*c__", "", m4_filt$X..taxonomy)
m4_filt$Class <- gsub("\\|.*", "", m4_filt$Class)

m4_filt$Phylum <- gsub(".*p__", "", m4_filt$X..taxonomy)
m4_filt$Phylum <- gsub("\\|.*", "", m4_filt$Phylum)

m4_filt_spec <- m4_filt[c(3,only_spec),]
```

Load in Parathaa_SILVA.SEED

```
parathaa_seed_assignments <- read.table("~/Repos/Parathaa2_OP3/SILVA_run_V3V4/Fiber_assignments/taxonom

parathaa_abundance_table_seed <- full_join(parathaa_seed_assignments, parathaa_abundance_table, by="que

parathaa_abundance_table_seed <- parathaa_abundance_table_seed[,-c(1, 256:263)]
```

Load in Dada2 assignments

```
DADA2_assignments <- read.table("~/Repos/Parathaa2_OP3/Datasets/fiber_amplicons_true/taxonomic_assignme
colnames(DADA2_assignments)[8] <- "query.name"

dada2_abundance_table <- full_join(DADA2_assignments, parathaa_abundance_table, by="query.name")
dada2_abundance_table <- dada2_abundance_table[,-c(1, 256:263)]
```

Load in Dada2 seed assignments

```
DADA2_assignments_seed <- read.table("~/Repos/Parathaa2_OP3/Datasets/fiber_amplicons_true/taxonomic_ass

colnames(DADA2_assignments_seed)[8] <- "query.name"
dada2_seed_abundance_table <- full_join(DADA2_assignments_seed, parathaa_abundance_table, by="query.nam
dada2_seed_abundance_table <- dada2_seed_abundance_table[,-c(1,256:263)]
```

Number of assignments

Functions

```
tax_levels <- c("Species", "Genus", "Family", "Order", "Class", "Phylum")
calculate_assignments <- function(tax_assignments){

  ret_frame <- data.frame(matrix(nrow=12, ncol=1))
  multi_names <- paste("multi", tax_levels, sep=" ")
  rownames(ret_frame) <- c(tax_levels, multi_names)
  colnames(ret_frame) <- "Number of Assignments"


  for(i in 1:length(tax_levels)){
    ret_frame[tax_levels[i],1] <- length(which(!is.na(tax_assignments[,tax_levels[i]])))
    ret_frame[multi_names[i],1] <- length(which(grepl(";", tax_assignments[,tax_levels[i]])))
  }
  return(ret_frame)
}
```

	Number of Assignments
Species	285
Genus	1437
Family	2630
Order	3021
Class	3486
Phylum	3628
multi Species	89
multi Genus	116
multi Family	107
multi Order	74
multi Class	72
multi Phylum	35

	Number of Assignments
Species	193
Genus	1704
Family	2929
Order	3498
Class	3727
Phylum	3885
multi Species	58
multi Genus	64
multi Family	49
multi Order	17
multi Class	3
multi Phylum	2

Parathaa kSGB

```
Assignment_levels_kSGB <- calculate_assignments(parathaa_ksgb_assignments)
kable_styling(kable((Assignment_levels_kSGB)))
```

Parathaa SEED

```
Assignment_levels_SEED <- calculate_assignments(parathaa_seed_assignments)
kable_styling(kable(Assignment_levels_SEED))
```

Dada2

```
Assignment_levels_DADA2 <- calculate_assignments(DADA2_assignments)
kable_styling(kable(Assignment_levels_DADA2))
```

Dada2_seed

```
Assignment_levels_DADA2_seed <- calculate_assignments(DADA2_assignments_seed)
kable_styling(kable(Assignment_levels_DADA2_seed))
```

Yikes these are a bit concerning...

	Number of Assignments
Species	176
Genus	2835
Family	3664
Order	3824
Class	3930
Phylum	3961
multi Species	0
multi Genus	0
multi Family	0
multi Order	0
multi Class	0
multi Phylum	0

	Number of Assignments
Species	48
Genus	2467
Family	3243
Order	3494
Class	3728
Phylum	3872
multi Species	0
multi Genus	0
multi Family	0
multi Order	0
multi Class	0
multi Phylum	0

	# of Assign. kSGB	# of Assign. SEED	# of Assign. DADA2	# of Assign. DADA2.Seed
Species	285	193	176	48
Genus	1437	1704	2835	2467
Family	2630	2929	3664	3243
Order	3021	3498	3824	3494
Class	3486	3727	3930	3728
Phylum	3628	3885	3961	3872
multi Species	89	58	0	0
multi Genus	116	64	0	0
multi Family	107	49	0	0
multi Order	74	17	0	0
multi Class	72	3	0	0
multi Phylum	35	2	0	0

Combined table

```
colnames(Assignment_levels_kSGB) <- "# of Assign. kSGB"
colnames(Assignment_levels_SEED) <- "# of Assign. SEED"
colnames(Assignment_levels_DADA2) <- "# of Assign. DADA2"
colnames(Assignment_levels_DADA2_seed) <- "# of Assign. DADA2.Seed"

merged_assignments <- cbind(Assignment_levels_kSGB, Assignment_levels_SEED, Assignment_levels_DADA2,
                            Assignment_levels_DADA2_seed)
kable_styling(kable(merged_assignments))
```

Composition Analysis

```
generate_tax_level_table <- function(table, level, prop=1){
  res_list <- list()
  Other <- c()
  for(i in unique(table[,level])){
    tmp_df <- table[which(table[,level]==i),]
    tmp_df <- tmp_df %>% select_if(is.numeric)

    tmp <- colSums(tmp_df)
    if(mean(tmp) < prop){
      if(length(Other)==0){
        Other <- tmp
      }else{
      Other <- Other + tmp
      }
    }else
      res_list[[i]] <- tmp
  }
  res_list[["Other"]] <- Other
  res_df <- data.frame(do.call(rbind, res_list))
  if("maxDist" %in% colnames(res_df)){
      res_df <- res_df[,-which(colnames(res_df) == "maxDist")]
  }
  res_df$Taxon <- rownames(res_df)
  return(res_df)
```

```
}
```

Phylum

```
parathaa_kSGB_Phylum <- generate_tax_level_table(parathaa_abundance_table_kSGB, "Phylum")
parathaa_seed_Phylum <- generate_tax_level_table(parathaa_abundance_table_seed, "Phylum")

dada2_full_Phylum <- generate_tax_level_table(dada2_abundance_table, "Phylum")
dada2_seed_Phylum <- generate_tax_level_table(dada2_seed_abundance_table, "Phylum")

m4_Phylum <- generate_tax_level_table(m4_filt_spec, "Phylum")

unique(c(parathaa_kSGB_Phylum$Taxon, parathaa_seed_Phylum$Taxon, dada2_full_Phylum$Taxon,
       dada2_seed_Phylum$Taxon, m4_Phylum$Taxon))
```

```
## [1] "Firmicutes"       "Proteobacteria"   "Bacteroidetes"    "Actinobacteria"
## [5] "Other"            "Actinobacteriota" "Fusobacteriota"   "UNCLASSIFIED"
```

```
parathaa_kSGB_Phylum_melt <- melt(parathaa_kSGB_Phylum)
```

```
## Using Taxon as id variables
```

```
parathaa_seed_Phylum_melt <- melt(parathaa_seed_Phylum)
```

```
## Using Taxon as id variables
```

```
dada2_full_Phylum_melt <- melt(dada2_full_Phylum)
```

```
## Using Taxon as id variables
```

```
dada2_seed_Phylum_melt <- melt(dada2_seed_Phylum)
```

```
## Using Taxon as id variables
```

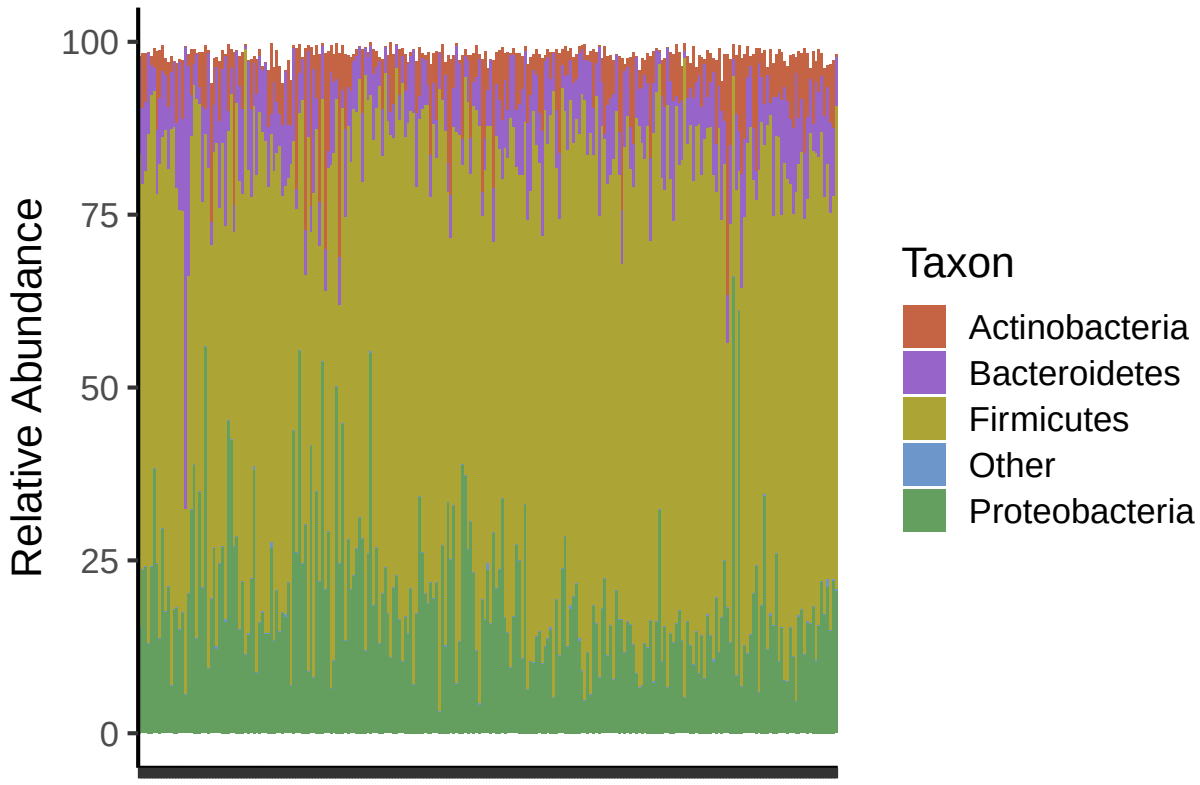
```
m4_Phylum_melt <- melt(m4_Phylum)
```

```
## Using Taxon as id variables
```

```
Phyla_colors <- c('Actinobacteria'="#c56445", "Bacteroidetes"="#9764c9", 'Firmicutes'="#aca535",
                  'Other'="#6d96ca", "Proteobacteria"="#659f60", "UNCLASSIFIED"="#c65c8a",
                  'Actinobacteriota'="#c56445", 'Bacteroidota'="#9764c9", "Fusobacteriota"="cyan")
```

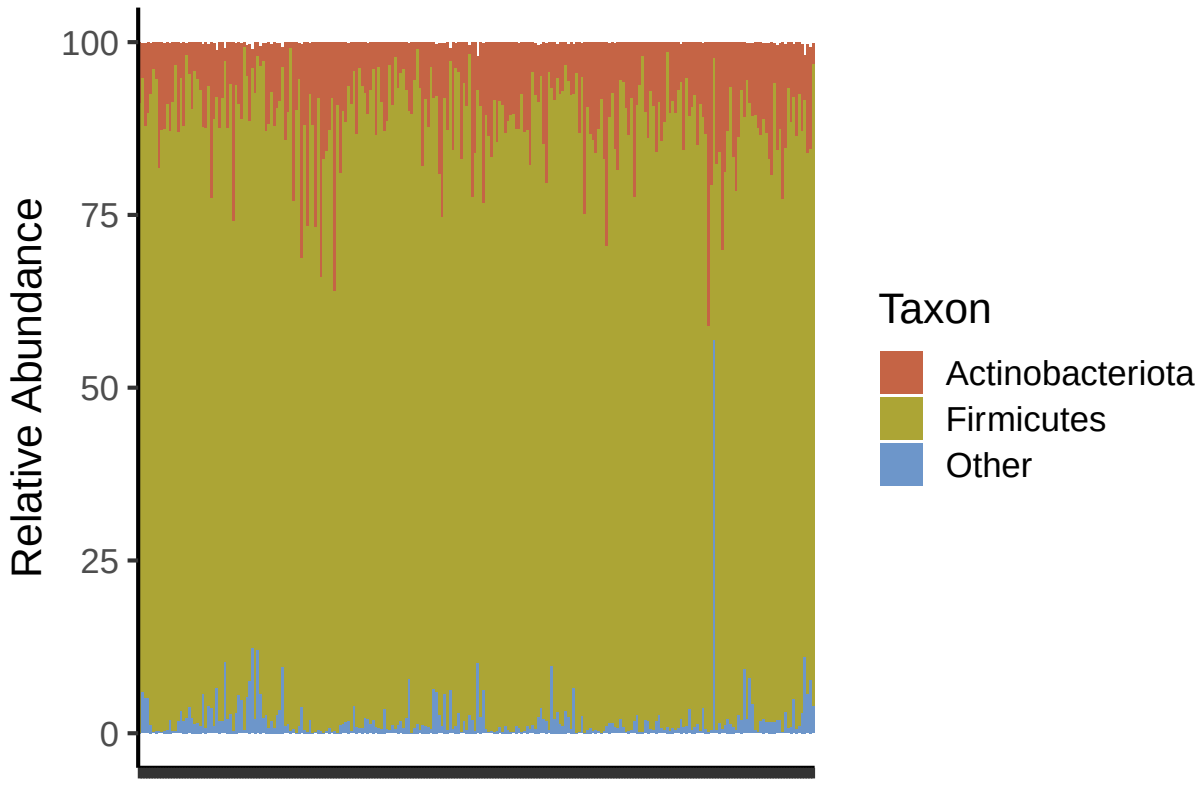
Parathaa kSGB

```
parathaa_kSGB_phyla_plot <- parathaa_kSGB_Phylum_melt %>%
  ggplot(aes(x=variable, y=value, fill=Taxon)) + geom_col()+
  theme_classic(base_size = 16) + theme(axis.text.x = element_blank()) +
  xlab("") + ylab("Relative Abundance") +
  scale_fill_manual(values=Phyla_colors)

parathaa_kSGB_phyla_plot
```



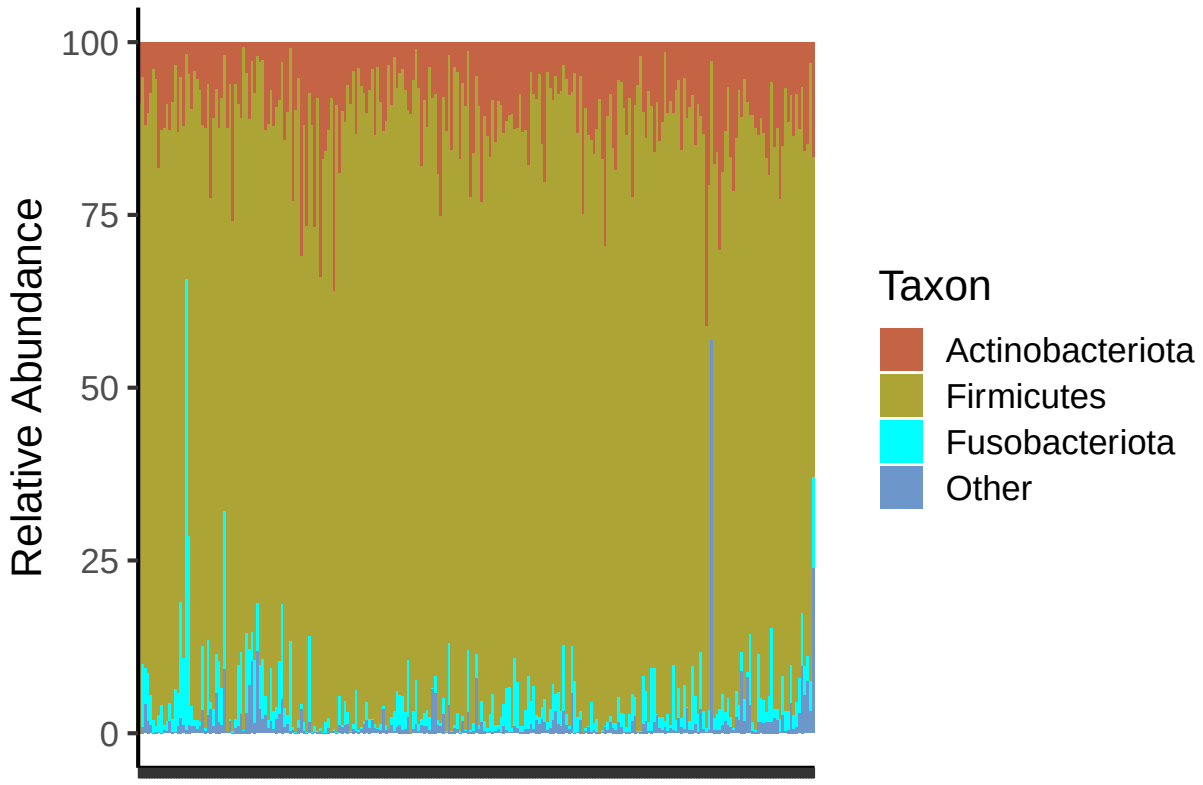
Parathaa SEED

```
parathaa_seed_phyla_plot <- parathaa_seed_Phylum_melt %>%
  ggplot(aes(x=variable, y=value, fill=Taxon)) + geom_col()+
  theme_classic(base_size = 16) + theme(axis.text.x = element_blank()) +
  xlab("") + ylab("Relative Abundance") +
  scale_fill_manual(values=Phyla_colors)

parathaa_seed_phyla_plot
```



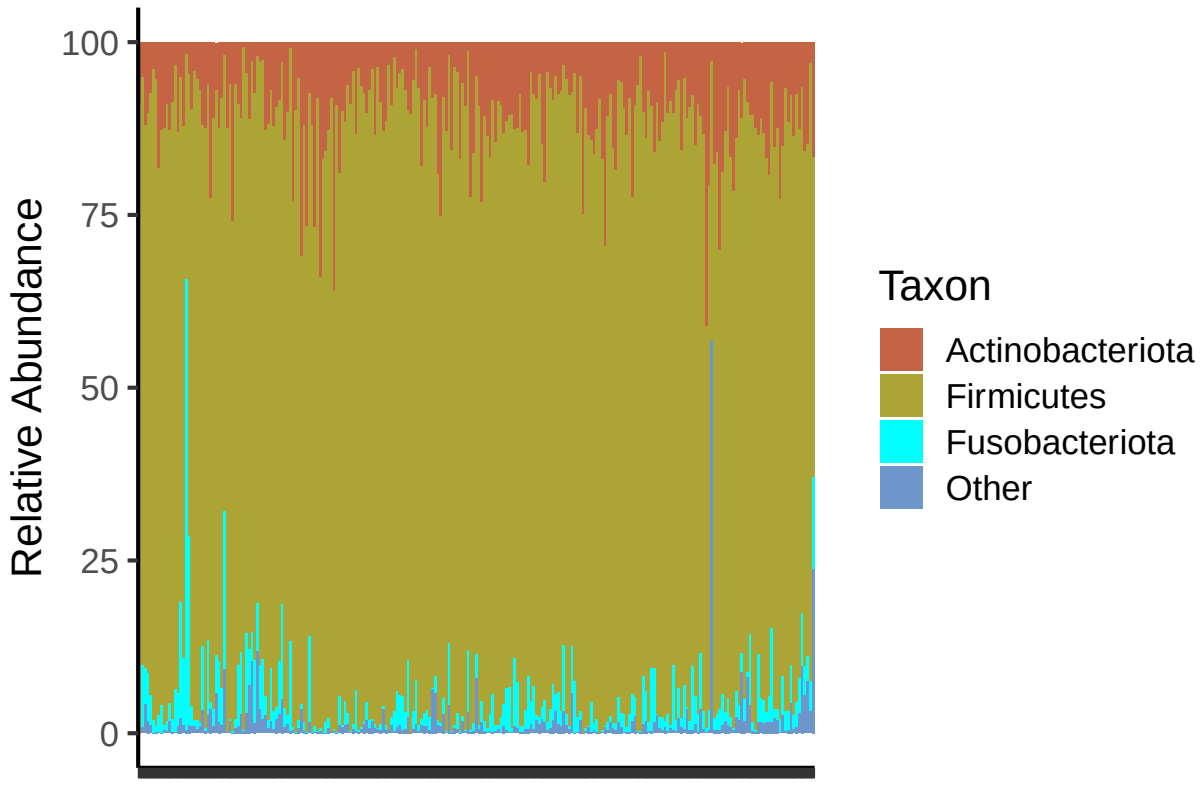
Dada2 Full

```
dada2_full_phyla_plot <- dada2_full_Phylum_melt %>%
  ggplot(aes(x=variable, y=value, fill=Taxon)) + geom_col()+
  theme_classic(base_size = 16) + theme(axis.text.x = element_blank()) +
  xlab("") + ylab("Relative Abundance") +
  scale_fill_manual(values=Phyla_colors)

dada2_full_phyla_plot
```



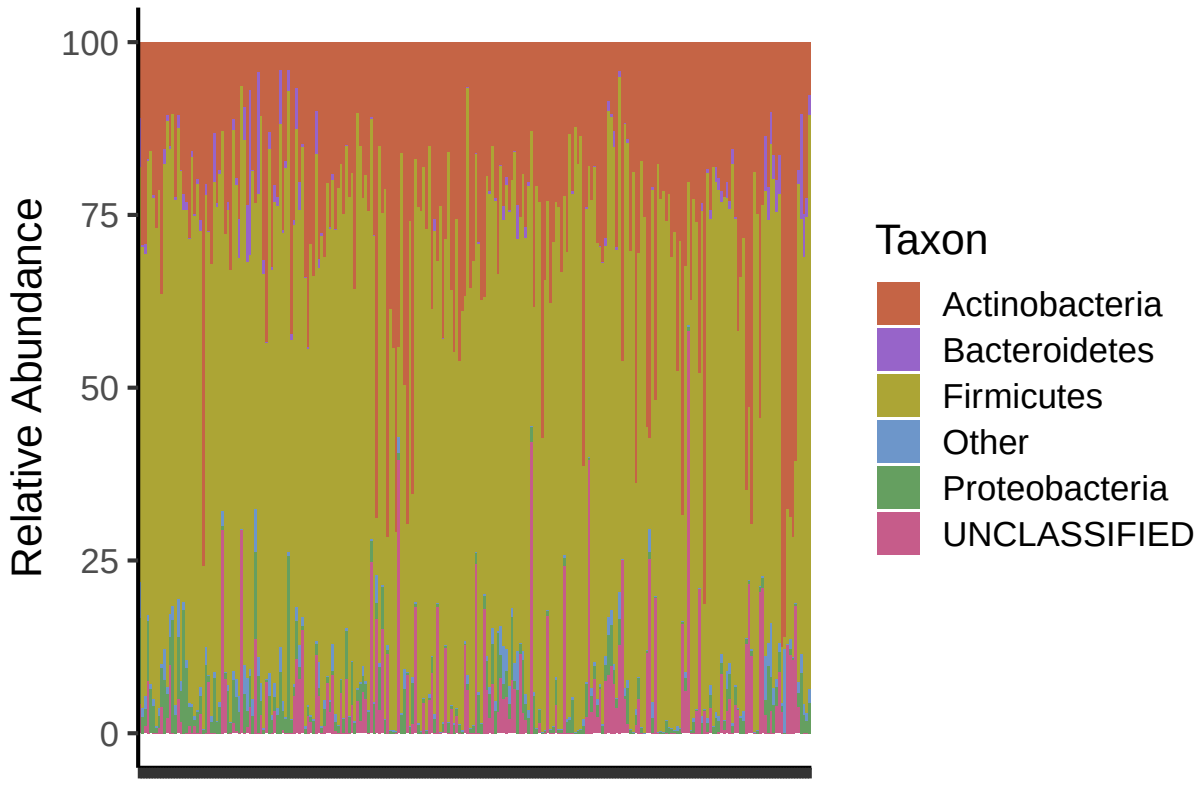
Dada2 SEED

```
dada2_seed_phyla_plot <- dada2_seed_Phylum_melt %>%
  ggplot(aes(x=variable, y=value, fill=Taxon)) + geom_col()+
  theme_classic(base_size = 16) + theme(axis.text.x = element_blank()) +
  xlab("") + ylab("Relative Abundance") +
  scale_fill_manual(values=Phyla_colors)

dada2_seed_phyla_plot
```



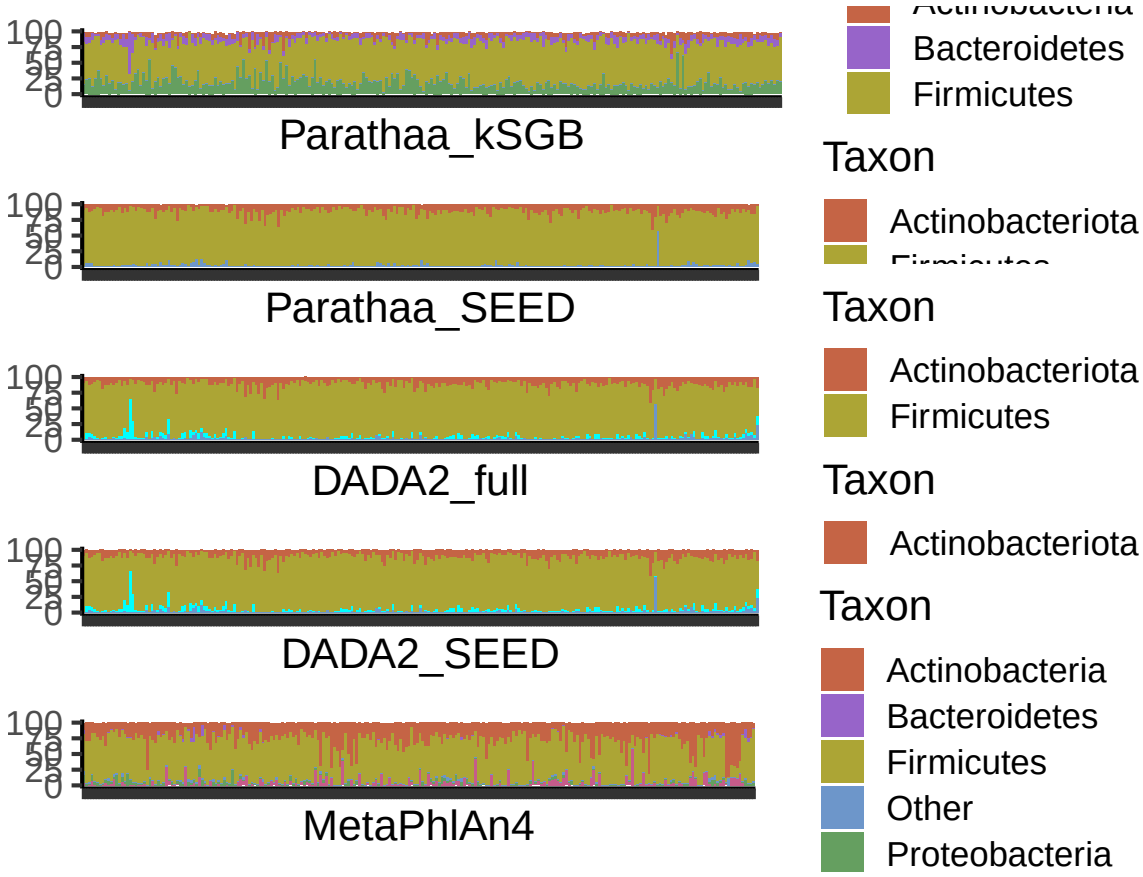
MetaPhlAn4

```
m4_Phyla_plot <- m4_Phylum_melt %>%
  ggplot(aes(x=variable, y=value, fill=Taxon)) + geom_col()+
  theme_classic(base_size = 16) + theme(axis.text.x = element_blank()) +
  xlab("") + ylab("Relative Abundance") +
  scale_fill_manual(values=Phyla_colors)

m4_Phyla_plot
```



Combined

```
plot_grid(parathaa_kSGB_phyla_plot +xlab("Parathaa_kSGB") + ylab(""),
          parathaa_seed_phyla_plot + xlab("Parathaa_SEED") + ylab(""),
          dada2_full_phyla_plot + xlab("DADA2_full") + ylab(""),
          dada2_seed_phyla_plot + xlab("DADA2_SEED") +ylab(""),
          m4_Phyla_plot + xlab("MetaPhlAn4") + ylab(""),
          ncol=1)
```



Genus Comparison

```
parathaa_kSGB_Genus <- generate_tax_level_table(parathaa_abundance_table_kSGB, "Genus", prop=5)
parathaa_seed_Genus <- generate_tax_level_table(parathaa_abundance_table_seed, "Genus", prop=5)

dada2_full_Genus <- generate_tax_level_table(dada2_abundance_table, "Genus", prop=5)
dada2_seed_Genus <- generate_tax_level_table(dada2_seed_abundance_table, "Genus", prop=5)

m4_genus <- generate_tax_level_table(m4_filt_spec, "Genus", prop=5)

unique(c(parathaa_kSGB_Genus$Taxon, parathaa_seed_Genus$Taxon, dada2_full_Genus$Taxon, dada2_seed_Genus$
```

```
##  [1] "Desulfofarcimen"         "Firmicutes_unclassified"
##  [3] "Other"                   "Bacillus"
##  [5] "Clavibacter"             "Peptoclostridium"
##  [7] "Turicibacter"            "Romboutsia"
##  [9] "Blautia"                 "Bifidobacterium"
## [11] "Terrisporobacter"        "Paraclostridium"
## [13] "GGB51725"                "GGB47957"
## [15] "GGB77090"                "Collinsella"
```

```
parathaa_kSGB_Genus_melt <- melt(parathaa_kSGB_Genus)
```

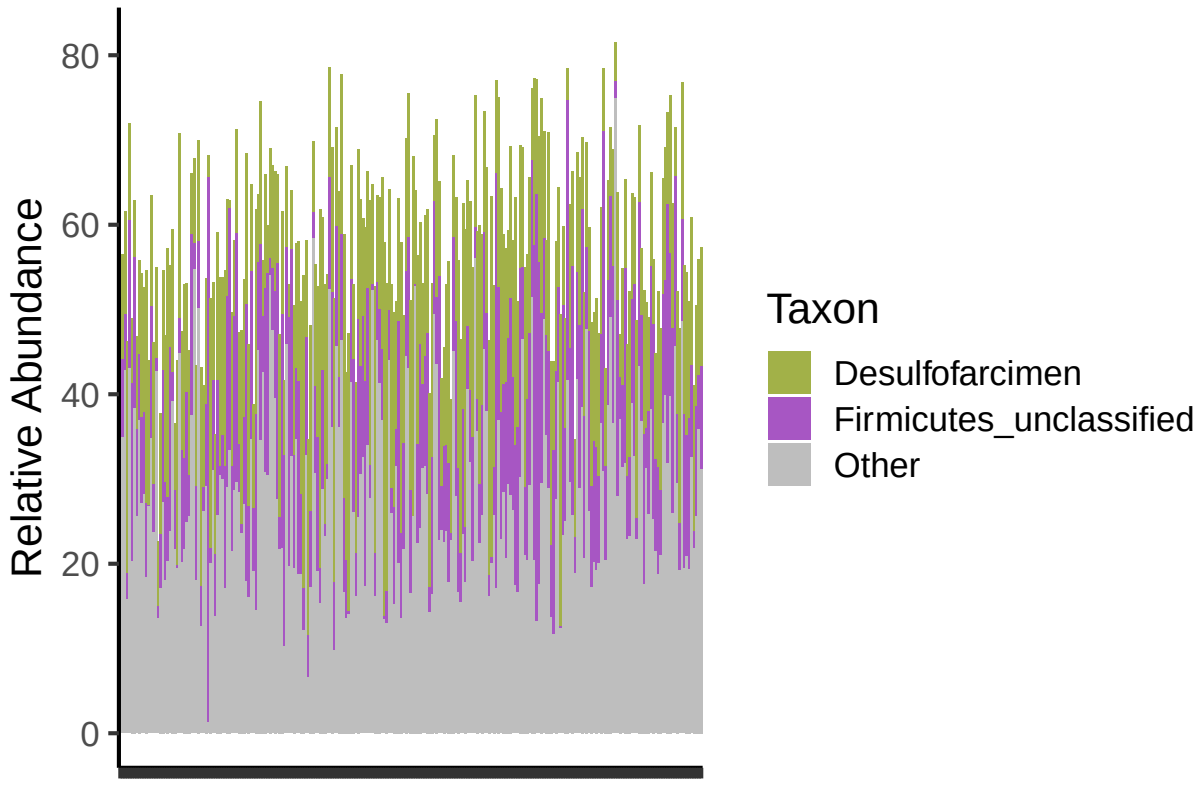
```
## Using Taxon as id variables
```

```
parathaa_seed_Genus_melt <- melt(parathaa_seed_Genus)
```

```
## Using Taxon as id variables
dada2_full_Genus_melt <- melt(dada2_full_Genus)

## Using Taxon as id variables
dada2_seed_Genus_melt <- melt(dada2_seed_Genus)

## Using Taxon as id variables
m4_genus_melt <- melt(m4_genus)

## Using Taxon as id variables
Genera_colors_5 <- c("Desulfofarcimen"="#a2b148",
                     "Firmicutes_unclassified"="#a756c3",
                     "Other"="grey",
                     "Bacillus"="#5bb94e",
                     "Clavibacter"="#6b6dc6",
                     "Peptoclostridium"="#d39a3b",
                     "Turicibacter"="#6999d4",
                     "Romboutsia"="#ca542a",
                     "Blautia"="#54c0a9",
                     "Bifidobacterium"="#d3425f",
                     "GGB51725"="#49874d",
                     "GGB47957"="#da73b6",
                     "GGB77090"="#83732f",
                     "GGB77090"="#a34d72",
                     "Collinsella"="#cc7d62")
```

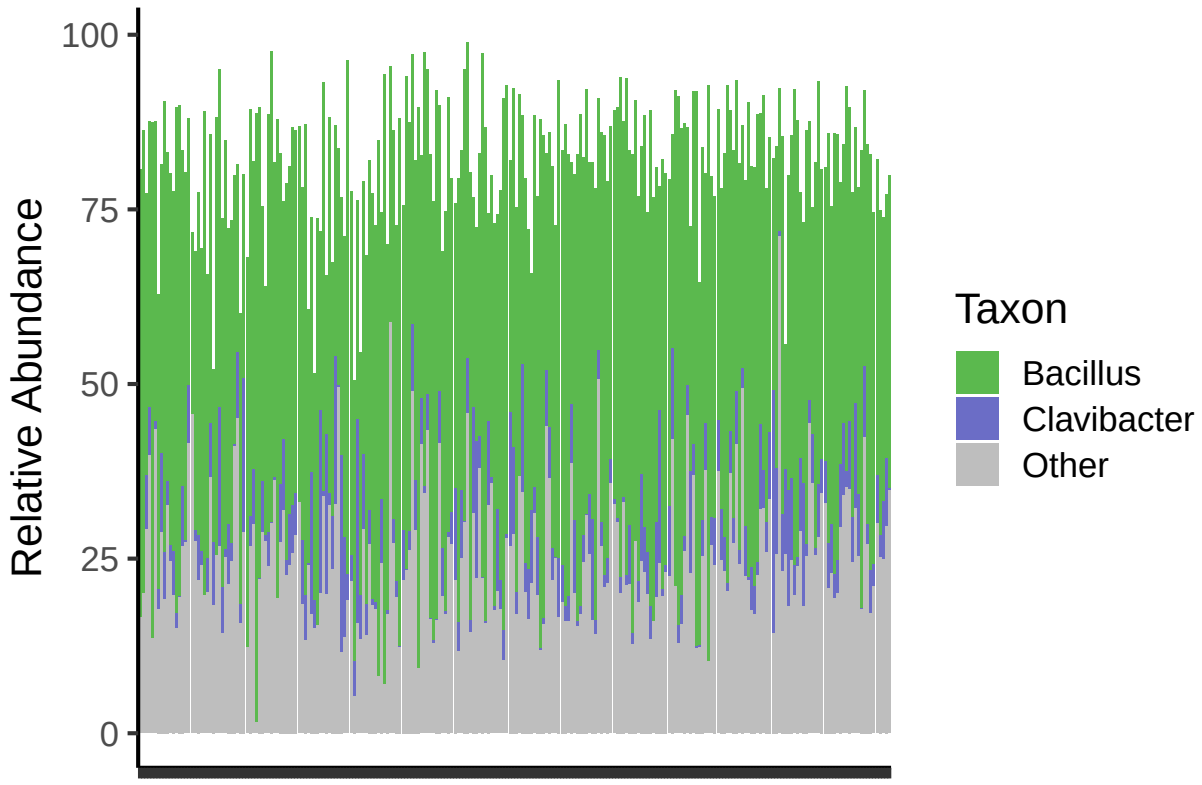
Parathaa kSGB

```
parathaa_kSGB_genus_plot <- parathaa_kSGB_Genus_melt %>%
  ggplot(aes(x=variable, y=value, fill=Taxon)) + geom_col()+
  theme_classic(base_size = 16) + theme(axis.text.x = element_blank()) +
  xlab("") + ylab("Relative Abundance") +
  scale_fill_manual(values=Genera_colors_5)

parathaa_kSGB_genus_plot
```



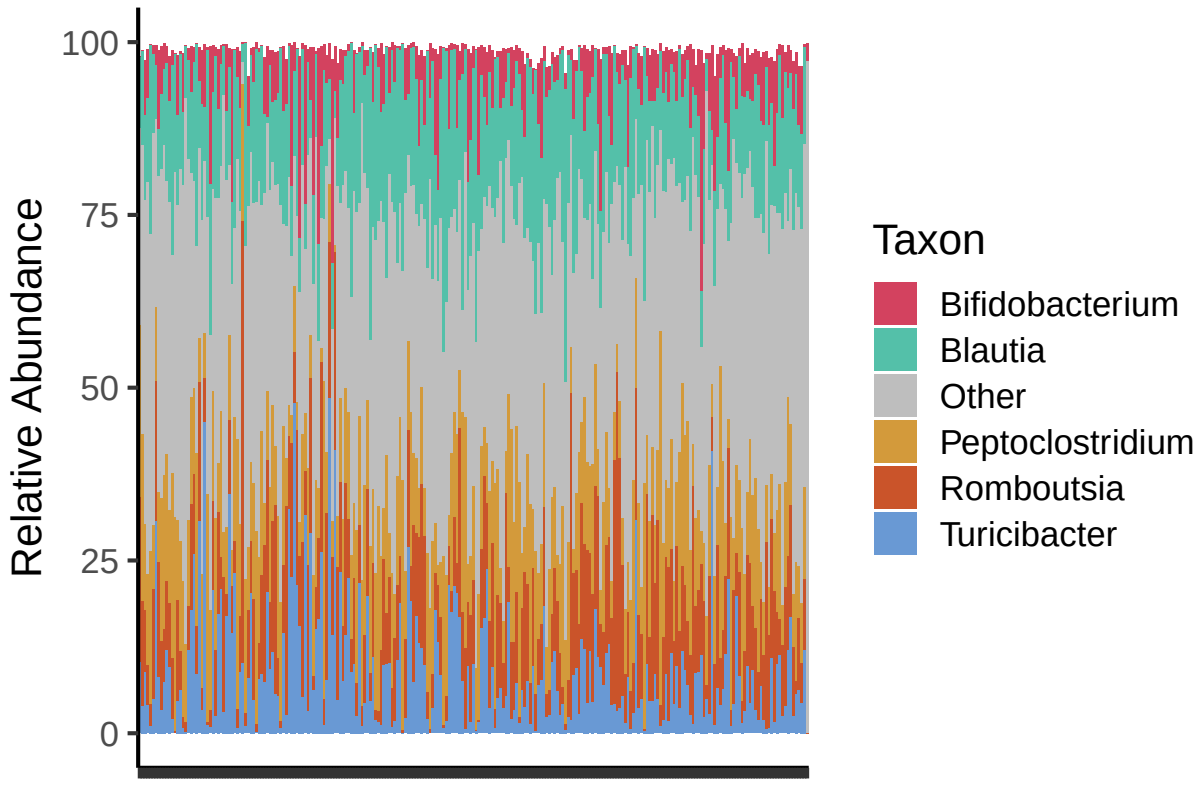
Parathaa SEED

```
parathaa_seed_genus_plot <- parathaa_seed_Genus_melt %>%
  ggplot(aes(x=variable, y=value, fill=Taxon)) + geom_col()+
  theme_classic(base_size = 16) + theme(axis.text.x = element_blank()) +
  xlab("") + ylab("Relative Abundance") +
  scale_fill_manual(values=Genera_colors_5)

parathaa_seed_genus_plot
```



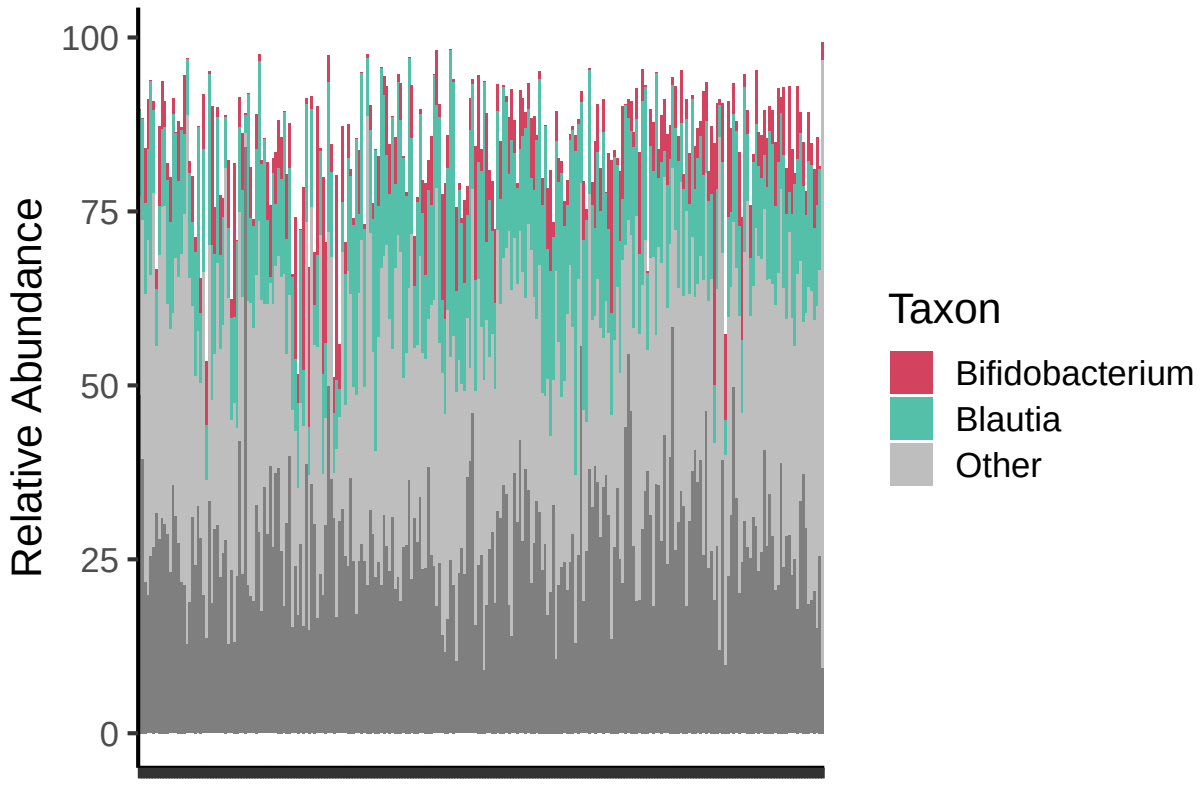
dada2 full

```
dada2_full_Genus_plot <- dada2_full_Genus_melt %>%
  ggplot(aes(x=variable, y=value, fill=Taxon)) + geom_col()+
  theme_classic(base_size = 16) + theme(axis.text.x = element_blank()) +
  xlab("") + ylab("Relative Abundance") +
  scale_fill_manual(values=Genera_colors_5)

dada2_full_Genus_plot
```



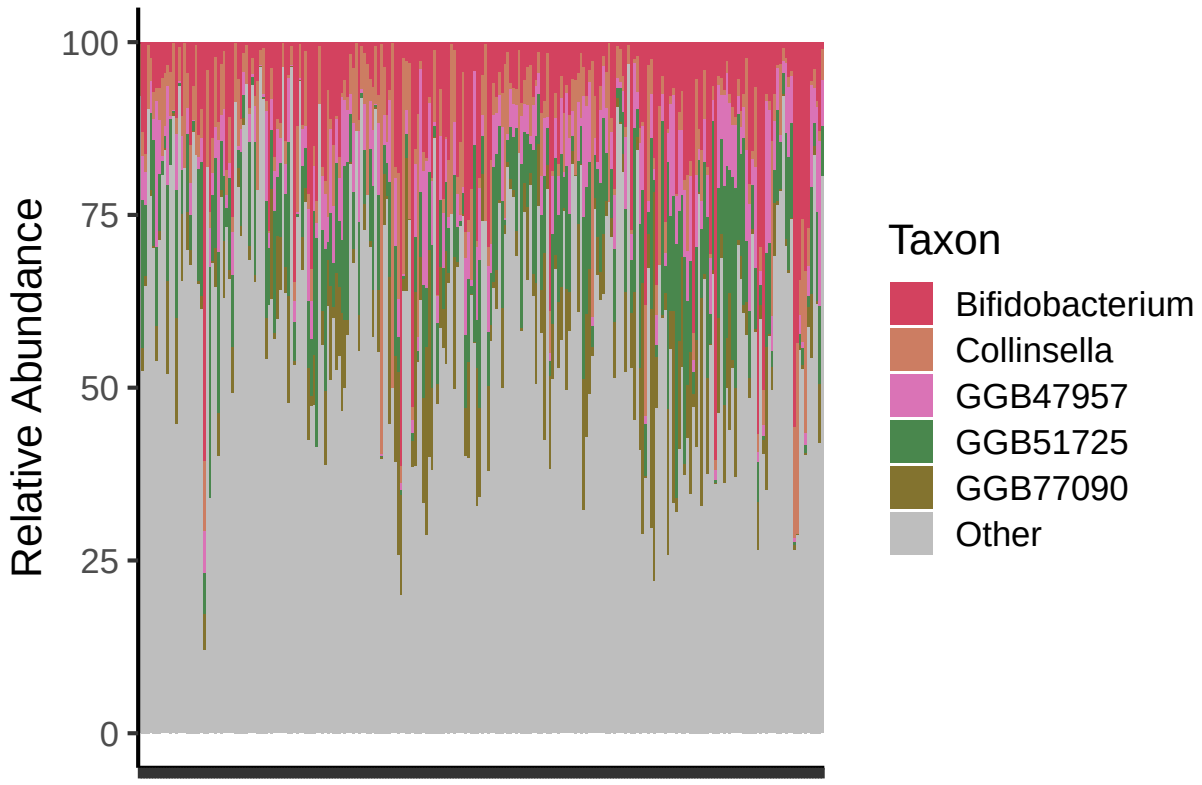
dada2 SEED

```
dada2_seed_genus_plot <- dada2_seed_Genus_melt %>%
  ggplot(aes(x=variable, y=value, fill=Taxon)) + geom_col()+
  theme_classic(base_size = 16) + theme(axis.text.x = element_blank()) +
  xlab("") + ylab("Relative Abundance") +
  scale_fill_manual(values=Genera_colors_5)

dada2_seed_genus_plot
```



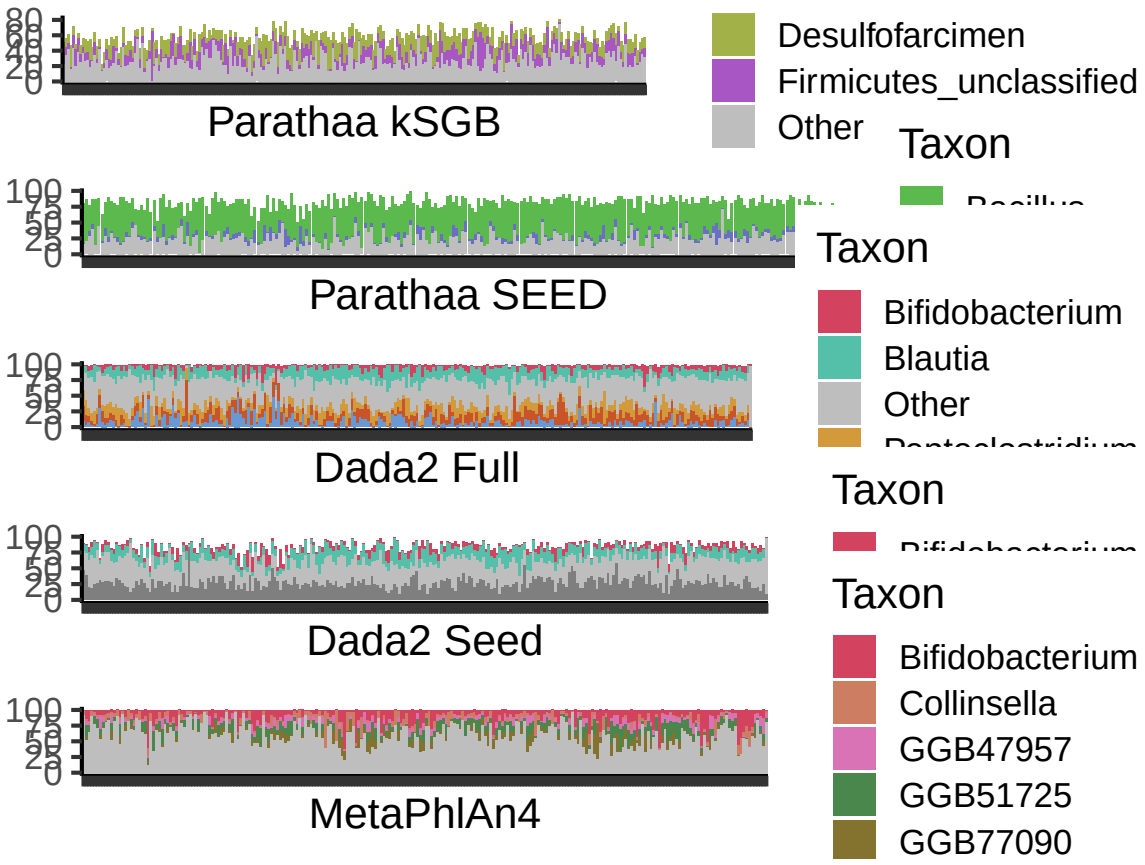
MetaPhlAn4

```
m4_genus_plot <- m4_genus_melt %>%
  ggplot(aes(x=variable, y=value, fill=Taxon)) + geom_col()+
  theme_classic(base_size = 16) + theme(axis.text.x = element_blank()) +
  xlab("") + ylab("Relative Abundance") +
  scale_fill_manual(values=Genera_colors_5)

m4_genus_plot
```



Combined

```
plot_grid(parathaa_kSGB_genus_plot + ylab("") + xlab("Parathaa kSGB"),
          parathaa_seed_genus_plot + ylab("") + xlab("Parathaa SEED"),
          dada2_full_Genus_plot + ylab("") + xlab("Dada2 Full"),
          dada2_seed_genus_plot + ylab("") + xlab("Dada2 Seed"),
          m4_genus_plot + ylab("") + xlab("MetaPhlAn4"),
          ncol=1)
```



PCA Analysis

```
#takes in a list of feature tables
#takes in a distance type
#generates PCA
#its not the perfect analysis since B/C doesn't do well with multi-assignments..

generate_PCA <- function(feat_tables, distance="bray"){
  merged_data <- Reduce(function(x, y) merge(x, y, by="Taxon", all=T), feat_tables)
  rownames(merged_data) <- merged_data$Taxon
  merged_data <- merged_data[,-1]
  merged_data[is.na(merged_data)] <- 0

  merged_data <- data.frame(t(merged_data))

  merged_dist <- vegdist(merged_data, method=distance)
  merged_PCA <- cmdscale(merged_dist, k=2, eig=T)

  plot_df <- data.frame(sample=rownames(merged_PCA$points),
                        PC1=merged_PCA$points[,1],
                        PC2=merged_PCA$points[,2])
  plot_df <- plot_df %>% mutate(Method = case_when(
    grepl("P_kSGB_", sample) ~ 'Parathaa_kSGB',
    grepl("P_SEED_", sample) ~ 'Parathaa_SEED',
    grepl("D_Full_", sample) ~ 'Dada2_Full',
    grepl("D_SEED_", sample) ~ 'Dada2_SEED',
```

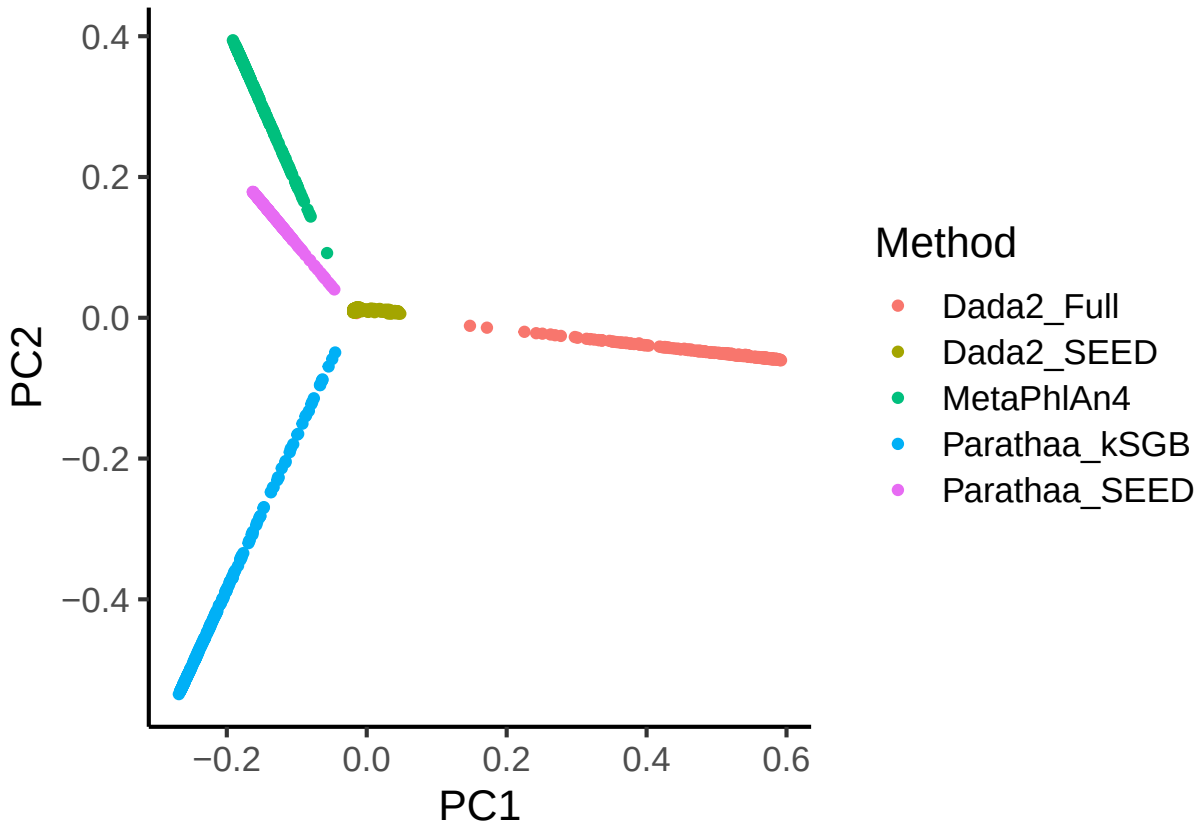
```
    grepl("M4_", sample) ~ 'MetaPhlAn4'
  ))

  ret_plot <- plot_df %>% ggplot(aes(x=PC1, y=PC2, colour=Method)) + geom_point() +
    theme_classic(base_size=16)
  return(ret_plot)
}
```

Generate data

```
parathaa_kSGB_Species <- generate_tax_level_table(parathaa_abundance_table_kSGB, level="Species", prop=0
colnames(parathaa_kSGB_Species)[-ncol(parathaa_kSGB_Species)] <-
  paste("P_kSGB_", colnames(parathaa_kSGB_Species)[-ncol(parathaa_kSGB_Species)], sep="")

parathaa_seed_Species <- generate_tax_level_table(parathaa_abundance_table_seed, level="Species", prop=0
colnames(parathaa_seed_Species)[-ncol(parathaa_seed_Species)] <-
  paste("P_SEED_", colnames(parathaa_seed_Species)[-ncol(parathaa_seed_Species)], sep="")


dada2_full_Species <- generate_tax_level_table(dada2_abundance_table, level="Species", prop=0)
colnames(dada2_full_Species)[-ncol(dada2_full_Species)] <-
  paste("D_Full_", colnames(dada2_full_Species)[-ncol(dada2_full_Species)], sep="")

dada2_seed_Species <- generate_tax_level_table(dada2_seed_abundance_table, level="Species", prop=0)
colnames(dada2_seed_Species)[-ncol(dada2_seed_Species)] <-
  paste("D_SEED_", colnames(dada2_seed_Species)[-ncol(dada2_seed_Species)], sep="")


m4_Species <- generate_tax_level_table(m4_filt_spec, level="Species", prop=0)
colnames(m4_Species)[-ncol(m4_Species)] <-
  paste("M4_", colnames(m4_Species)[-ncol(m4_Species)], sep="")

spec_feat_tabs <- list(parathaa_kSGB_Species, parathaa_seed_Species, dada2_full_Species, dada2_seed_Spe
                       m4_Species)
```

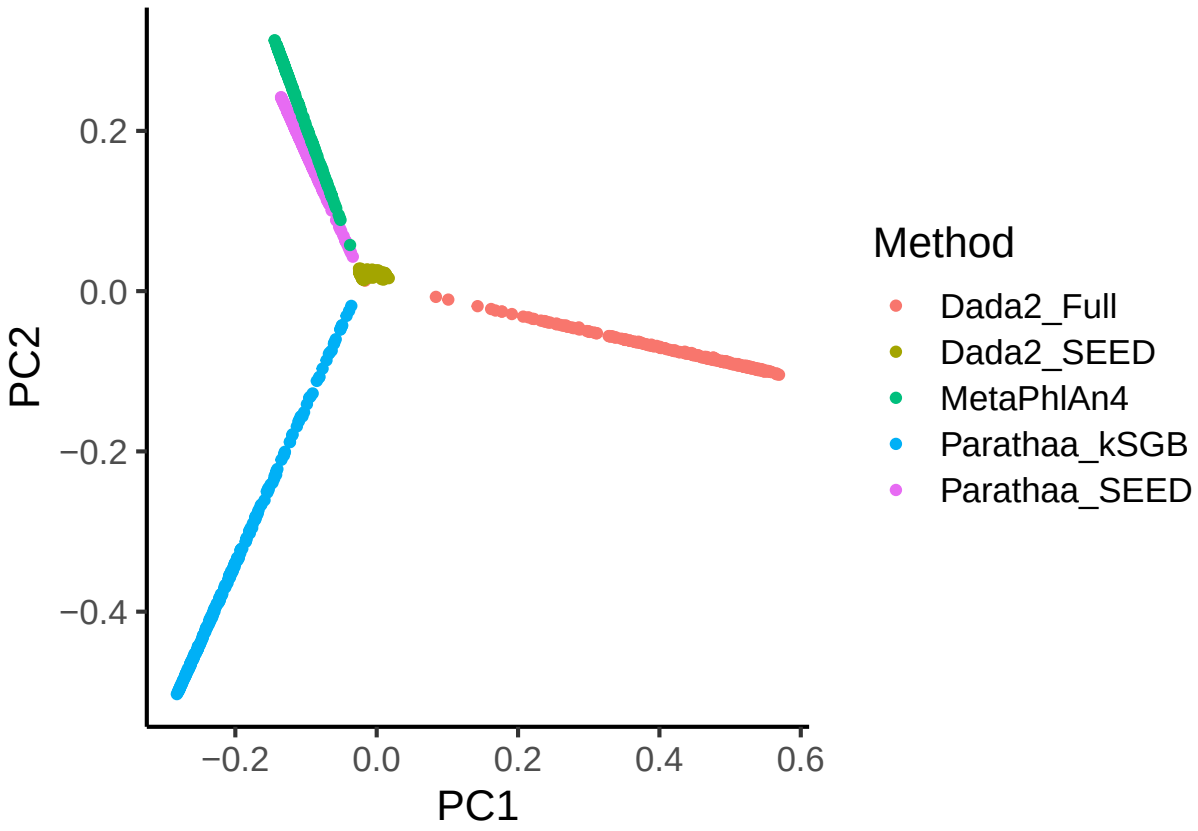
Bray

```
Spec_PCA_bray <- generate_PCA(spec_feat_tabs)
Spec_PCA_bray
```



Jaccard

```
Spec_PCA_Jac <- generate_PCA(spec_feat_tabs, distance="jaccard")
Spec_PCA_Jac
```



Deeper investigation

```
region_tree_kSGB <- "~/Repos/Parathaa2_OP3/M4_Oct22/M4_kSGB_V3V4/region_specific.tree"
taxafile_kSGB <- "~/Repos/Parathaa2_OP3/M4_Oct22/taxmapper.tsv"
load("~/Repos/Parathaa2_OP3/M4_Oct22/M4_kSGB_V3V4/resultTree_bestThresholds.RData")
Parathaa_kSGB_tree <- resultData$tax_bestcuts
Parathaa_kSGB_placements <- read.jplace("~/Repos/Parathaa2_OP3/M4_Oct22/M4_kSGB_V3V4/Fiber_assignments/
```

```
region_tree_SEED <- "~/Repos/Parathaa2_OP3/SILVA_run_V3V4/region_specific.tree"
taxafile_SEED <- "~/Repos/Parathaa/parathaa/input/silva_v138/taxmap_slv_ssu_ref_138.1.txt"

load("~/Repos/Parathaa2_OP3/SILVA_run_V3V4/resultTree_bestThresholds.RData")
Parathaa_SEED_tree <- resultData$tax_bestcuts
Parathaa_SEED_placements <- read.jplace("~/Repos/Parathaa2_OP3/SILVA_run_V3V4/Fiber_assignments/merged_
```

Phylum investigation...

```
merged_parathaa_assignments <- full_join(parathaa_ksgb_assignments, parathaa_seed_assignments, by="query

phylum_non_agreement <- merged_parathaa_assignments[which(merged_parathaa_assignments$Phylum.x !=
                                                          merged_parathaa_assignments$Phylum.y),]

setdiff(phylum_non_agreement$Phylum.x, phylum_non_agreement$Phylum.y)
```

```
##  [1] "Bacteroidetes"
```

```
##  [2] "Actinobacteria"
##  [3] "Bacteroidetes;Cyanobacteria"
##  [4] "Acidobacteria"
##  [5] "Cyanobacteria;Tenericutes"
##  [6] "Tenericutes"
##  [7] "Bacteria_unclassified"
##  [8] "Chlorobi"
##  [9] "Ignavibacteriae"
## [10] "Candidatus_Kryptonia;Ignavibacteriae;Synergistetes"
## [11] "Candidatus_Margulisbacteria"
## [12] "Synergistetes"
## [13] "Nitrospirae"
## [14] "Fusobacteria"
## [15] "Candidatus_Eremiobacteraeota"
```

```
setdiff(phylum_non_agreement$Phylum.y, phylum_non_agreement$Phylum.x)
```

```
##  [1] "Actinobacteriota"                "Bacteroidota"
##  [3] "Myxococcota"                     "Desulfobacterota"
##  [5] "Acidobacteriota"                 "Bdellovibrionota"
##  [7] "Campylobacterota"                "Nitrospirota"
##  [9] "Synergistota"                    "Nitrospinota"
## [11] "Acidobacteriota;Actinobacteriota" "Chloroflexi"
## [13] "Fusobacteriota"                  "Fibrobacterota"
## [15] "Patescibacteria"                 "Gemmatimonadota"
## [17] "Dependentiae"                    "Marinimicrobia (SAR406 clade)"
## [19] "Caldatribacteriota"
```

```
phylum_non_agreement$Phylum.y[which(phylum_non_agreement$Phylum.y=="Actinobacteriota")] <- "Actinobacte
phylum_non_agreement$Phylum.y[which(phylum_non_agreement$Phylum.y=="Bacteroidota")] <- "Bacteroidetes"
phylum_non_agreement$Phylum.y[which(phylum_non_agreement$Phylum.y=="Acidobacteriota")] <- "Acidobacteri
phylum_non_agreement$Phylum.y[which(phylum_non_agreement$Phylum.y=="Fusobacteriota")] <- "Fusobacteria"

phylum_non_agreement <- phylum_non_agreement[which(phylum_non_agreement$Phylum.x !=
                                                   phylum_non_agreement$Phylum.y),]
```

Bacterodietes dive...

```
non_agree_bacts <- phylum_non_agreement[which(phylum_non_agreement$Phylum.x=="Bacteroidetes"),]
non_agree_bacts$query.name[1]
```

```
## [1] "008424a175a5cc75c61aa1a547a93f2b"
```

```
non_agree_bacts$query.name[2]
```

```
## [1] "00fd4ac0447c6b6539db1bcc40240e81"
```

```
non_agree_bacts[which(non_agree_bacts$query.name=="008424a175a5cc75c61aa1a547a93f2b"),]
```

008424a175a5cc75c61aa1a547a93f2b

```
##                         query.name Kingdom.x      Phylum.x        Class.x
## 5 008424a175a5cc75c61aa1a547a93f2b  Bacteria Bacteroidetes Flavobacteriia
##           Order.x          Family.x         Genus.x Species.x  maxDist.x
## 5 Flavobacteriales Flavobacteriaceae Salinimicrobium      <NA> 0.04222775
```

```
##   Kingdom.y   Phylum.y    Class.y         Order.y        Family.y      Genus.y
## 5  Bacteria Firmicutes Clostridia Oscillospirales Ruminococcaceae Ruminococcus
##   Species.y  maxDist.y
## 5      <NA> 0.03843471
```

```
DADA2_assignments[which(DADA2_assignments$query.name=="008424a175a5cc75c61aa1a547a93f2b"),]
```

```
##
## ATTGGACAATGGACCAAAAGTCTGATCCAGCAATTCTGTGTGCACGATGAAGTTTTTCGGAATGTAAAGTGCTTTCAGTTGGGAAGAAGAAAGTGACGGTA
##
## ATTGGACAATGGACCAAAAGTCTGATCCAGCAATTCTGTGTGCACGATGAAGTTTTTCGGAATGTAAAGTGCTTTCAGTTGGGAAGAAGAAAGTGACGGTA
##
## ATTGGACAATGGACCAAAAGTCTGATCCAGCAATTCTGTGTGCACGATGAAGTTTTTCGGAATGTAAAGTGCTTTCAGTTGGGAAGAAGAAAGTGACGGTA
##
## ATTGGACAATGGACCAAAAGTCTGATCCAGCAATTCTGTGTGCACGATGAAGTTTTTCGGAATGTAAAGTGCTTTCAGTTGGGAAGAAGAAAGTGACGGTA
##
## ATTGGACAATGGACCAAAAGTCTGATCCAGCAATTCTGTGTGCACGATGAAGTTTTTCGGAATGTAAAGTGCTTTCAGTTGGGAAGAAGAAAGTGACGGTA
##
## ATTGGACAATGGACCAAAAGTCTGATCCAGCAATTCTGTGTGCACGATGAAGTTTTTCGGAATGTAAAGTGCTTTCAGTTGGGAAGAAGAAAGTGACGGTA
##
## ATTGGACAATGGACCAAAAGTCTGATCCAGCAATTCTGTGTGCACGATGAAGTTTTTCGGAATGTAAAGTGCTTTCAGTTGGGAAGAAGAAAGTGACGGTA
##
## ATTGGACAATGGACCAAAAGTCTGATCCAGCAATTCTGTGTGCACGATGAAGTTTTTCGGAATGTAAAGTGCTTTCAGTTGGGAAGAAGAAAGTGACGGTA
```

HMMM so dada2 assigns it Fusobacterium but the other two do not… that is very concerning…

All blast searches give: Fusobacterium..

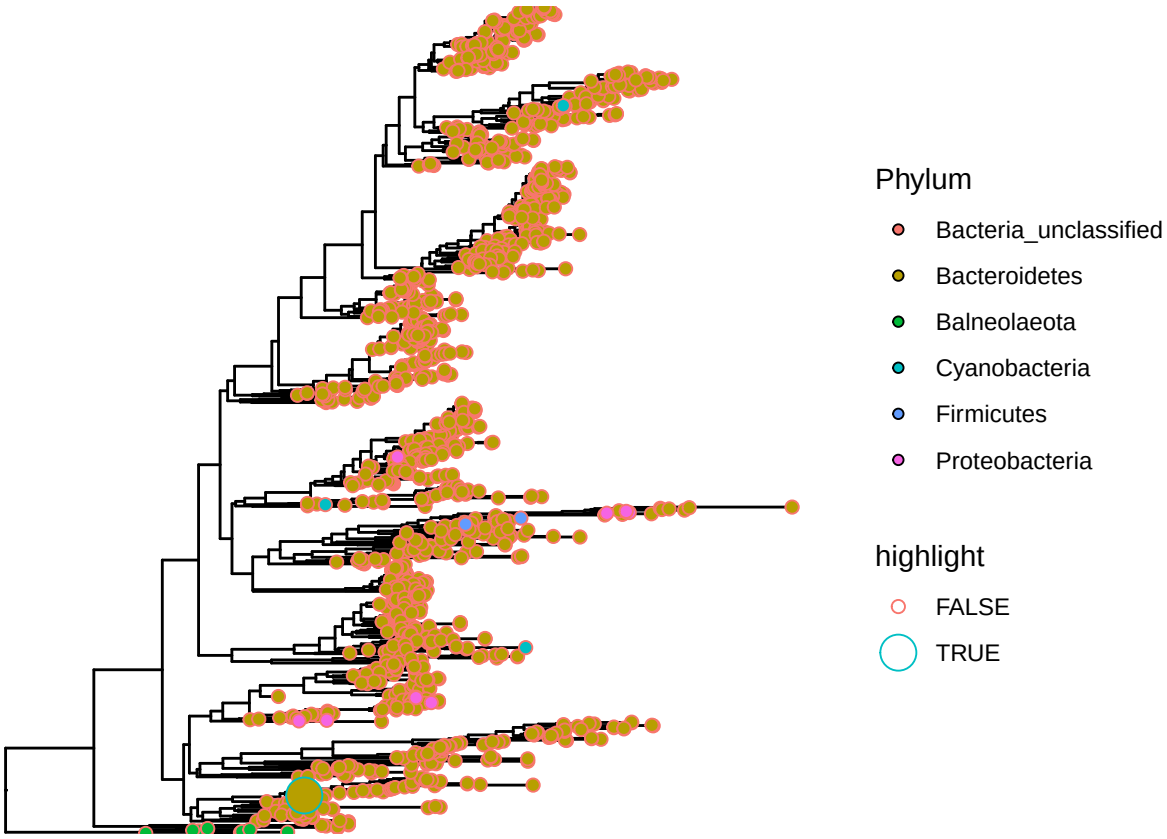
Lets look at the tree placements…

```
parthaa_kSGB_placements <- get.placements(Parathaa_kSGB_placements)

### hmmm I think something is going wrong here...
parthaa_kSGB_placement_nodes <- plot_placement(Parathaa_kSGB_tree, "Phylum", Parathaa_kSGB_placements,
                                               label="008424a175a5cc75c61aa1a547a93f2b")

get_n_parents(Parathaa_kSGB_tree, 2320, 10)
```

```
## [1] 13294
```

```
subtree_parent <- get_subtree_plot_data(Parathaa_kSGB_tree, 13294, isTip=FALSE,
                                        highlight_lab = "PEMPGKGO_00004_16S_ribosomal_RNA|M1878371282")

ggtree(as.phylo(subtree_parent)) %<+% subtree_parent + geom_tippoint(aes(fill=Phylum, size=highlight, c
```

```
## Warning: Using size for a discrete variable is not advised.
```

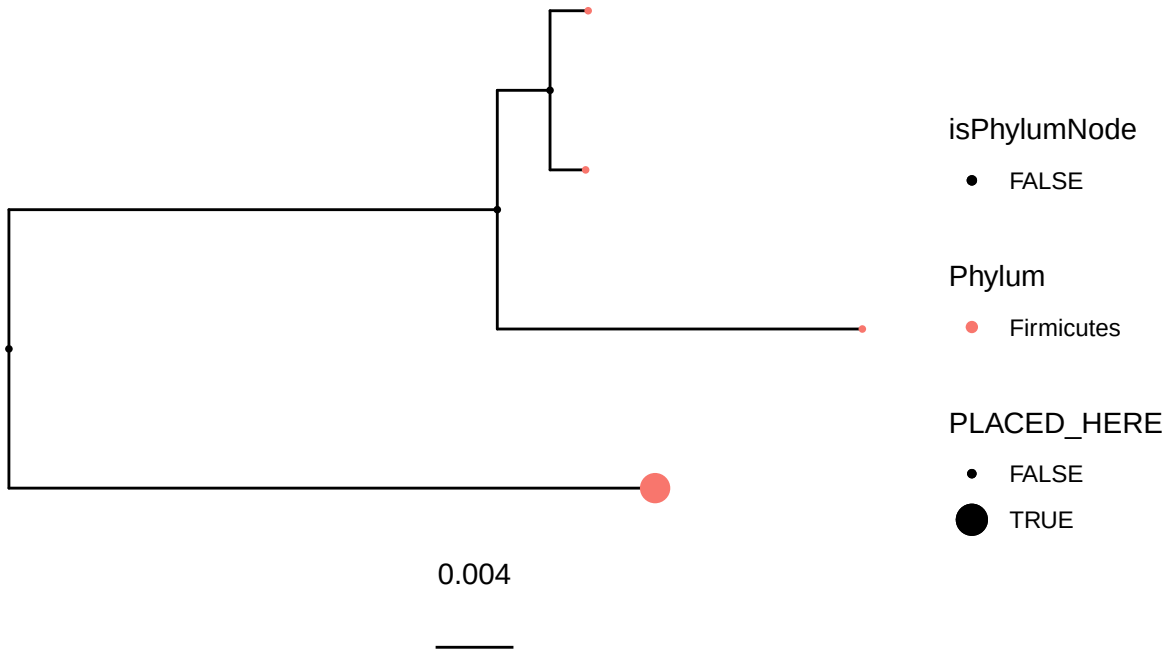


It seems the placements around it are all "bacteroidetes" lets blast the sequence at its tip..

It does blast to Bacteroidetes... so doesn't seem like a huge issue...

Why is it being placed there though????? Something is clearly going wrong here during placement not sure what it is though... Is it the bifurcation enforcement?

```
parthaa_SEED_placement <- plot_placement(Parathaa_SEED_tree, "Phylum", Parathaa_SEED_placements,
                                         label="008424a175a5cc75c61aa1a547a93f2b")
parthaa_SEED_placement
```

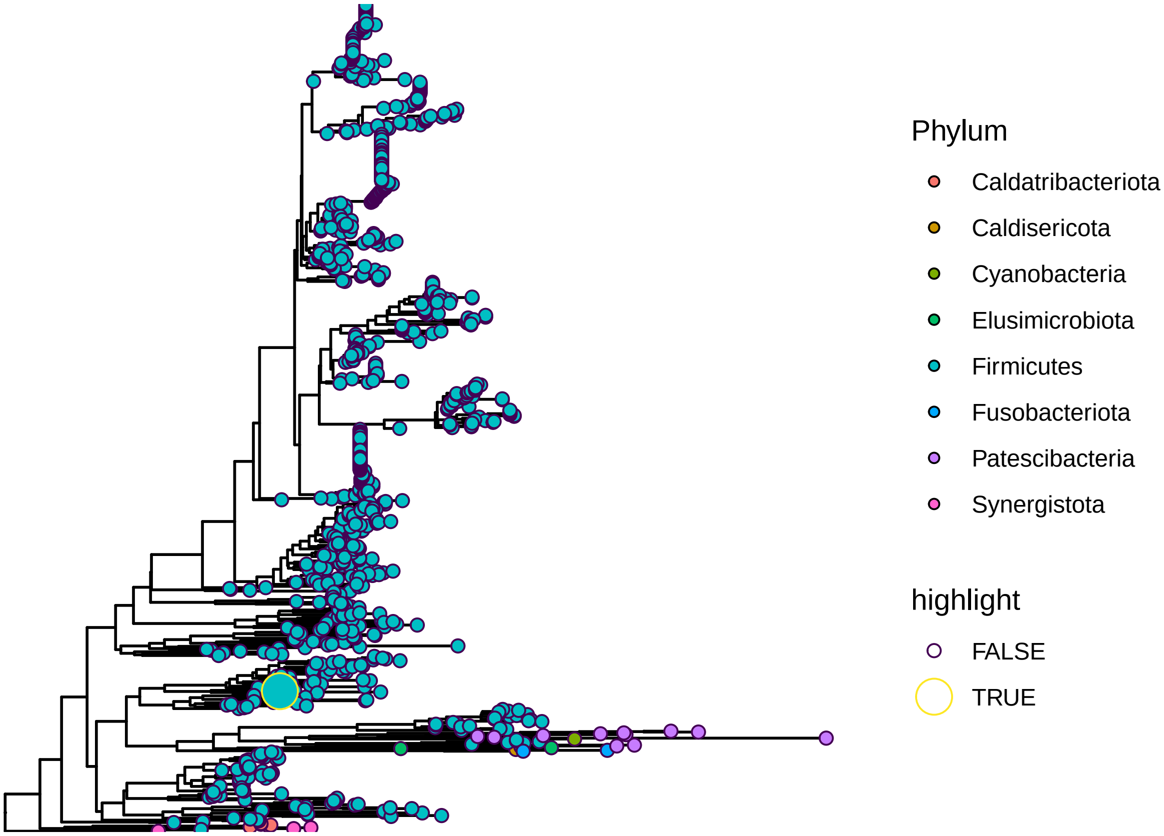



```
get_n_parents(Parathaa_SEED_tree, 2030, 10)
```

```
## [1] 5918
```

```
subtree_parent_SEED <- get_subtree_plot_data(Parathaa_SEED_tree, 5918, isTip = FALSE,
                                             highlight_lab="AF030451.Ru2Albu9")

ggtree(as.phylo(subtree_parent_SEED)) %<+% subtree_parent_SEED + geom_tippoint(aes(fill=Phylum, size=hi
```

```
## Warning: Using size for a discrete variable is not advised.
```



HMMM it places it in the Firmicutes... There are Fuso's in the subtree but there are a bit far away...

Again looks like it is a placement issue?

```
non_agree_bacts[which(non_agree_bacts$query.name=="00fd4ac0447c6b6539db1bcc40240e81"),]
```

00fd4ac0447c6b6539db1bcc40240e81

```
##                         query.name Kingdom.x      Phylum.x Class.x Order.x
## 9 00fd4ac0447c6b6539db1bcc40240e81  Bacteria Bacteroidetes    <NA>    <NA>
##   Family.x Genus.x Species.x maxDist.x Kingdom.y       Phylum.y        Class.y
## 9     <NA>    <NA>      <NA> 0.1187457  Bacteria Actinobacteria Actinobacteria
##                 Order.y Family.y Genus.y Species.y maxDist.y
## 9 Streptosporangiales     <NA>    <NA>      <NA> 0.1434061
```

```
DADA2_assignments[which(DADA2_assignments$query.name=="00fd4ac0447c6b6539db1bcc40240e81"),]
```

```
##
## CTTCGGCAATGGGCGAAAGCCTGACCGAGCAACGCCGCGTGAATGAAGAAGGCCTTCGGGTTGTAAACTCCTGTTGTTGGGGAAGATAATGACGGTACCC
##
## CTTCGGCAATGGGCGAAAGCCTGACCGAGCAACGCCGCGTGAATGAAGAAGGCCTTCGGGTTGTAAACTCCTGTTGTTGGGGAAGATAATGACGGTACCC
##
## CTTCGGCAATGGGCGAAAGCCTGACCGAGCAACGCCGCGTGAATGAAGAAGGCCTTCGGGTTGTAAACTCCTGTTGTTGGGGAAGATAATGACGGTACCC
##
## CTTCGGCAATGGGCGAAAGCCTGACCGAGCAACGCCGCGTGAATGAAGAAGGCCTTCGGGTTGTAAACTCCTGTTGTTGGGGAAGATAATGACGGTACCC
##
## CTTCGGCAATGGGCGAAAGCCTGACCGAGCAACGCCGCGTGAATGAAGAAGGCCTTCGGGTTGTAAACTCCTGTTGTTGGGGAAGATAATGACGGTACCC
##
```

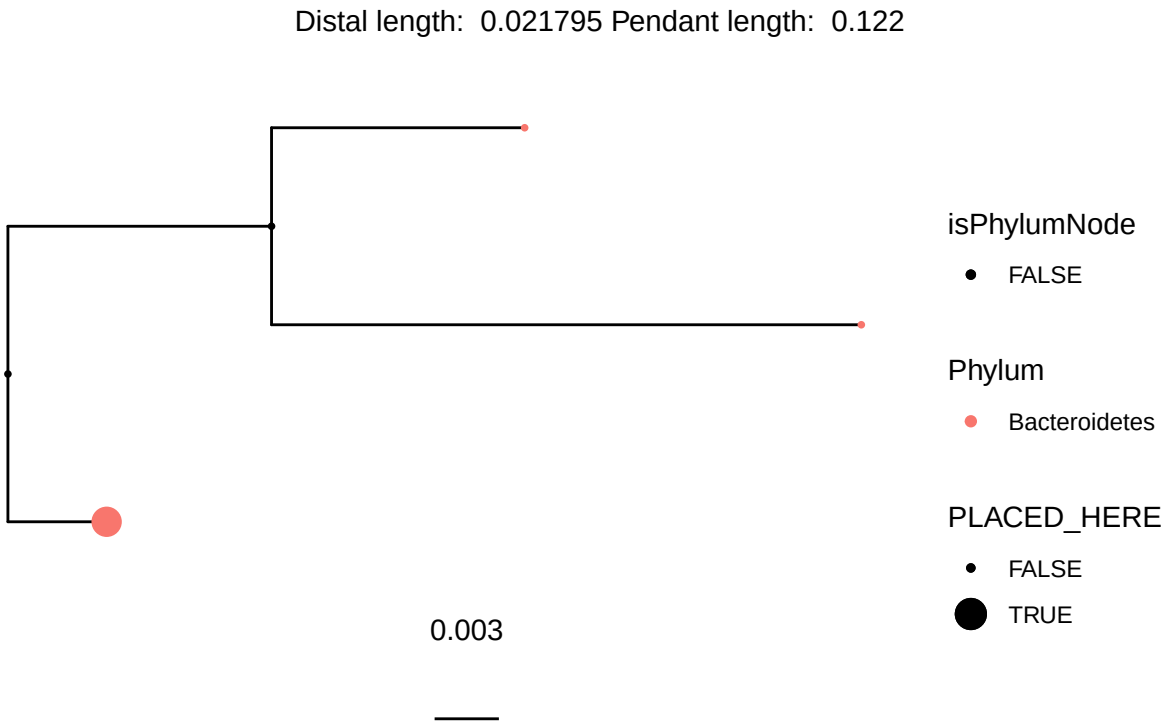
```
## CTTCGGCAATGGGCGAAAGCCTGACCGAGCAACGCCGCGTGAATGAAGAAGGCCTTCGGGTTGTAAACTCCTGTTGTTGGGGAAGATAATGACGGTACCCA
##
## CTTCGGCAATGGGCGAAAGCCTGACCGAGCAACGCCGCGTGAATGAAGAAGGCCTTCGGGTTGTAAACTCCTGTTGTTGGGGAAGATAATGACGGTACCCA
##
## CTTCGGCAATGGGCGAAAGCCTGACCGAGCAACGCCGCGTGAATGAAGAAGGCCTTCGGGTTGTAAACTCCTGTTGTTGGGGAAGATAATGACGGTACCCA
```

So parathaa_kSGB says it should be Bacteroidetes... (no classification after Phylum...) Parathaa_SEED says it should be Actinobacteria...(Order is Streptosporangiales...) DADA2 says it should be Actinobacteria... (Family is Eggerthellaceae, Genus is Asaccharobacter...)

So DADA2 and Parathaa do not agree on Order assignment... lets dig in...

BLAST says: it should be in Eggerthellaceae....

```
### hmmm I think something is going wrong here...
parthaa_kSGB_placement_nodes <- plot_placement(Parathaa_kSGB_tree, "Phylum", Parathaa_kSGB_placements,
                                               label="00fd4ac0447c6b6539db1bcc40240e81")
parthaa_kSGB_placement_nodes
```

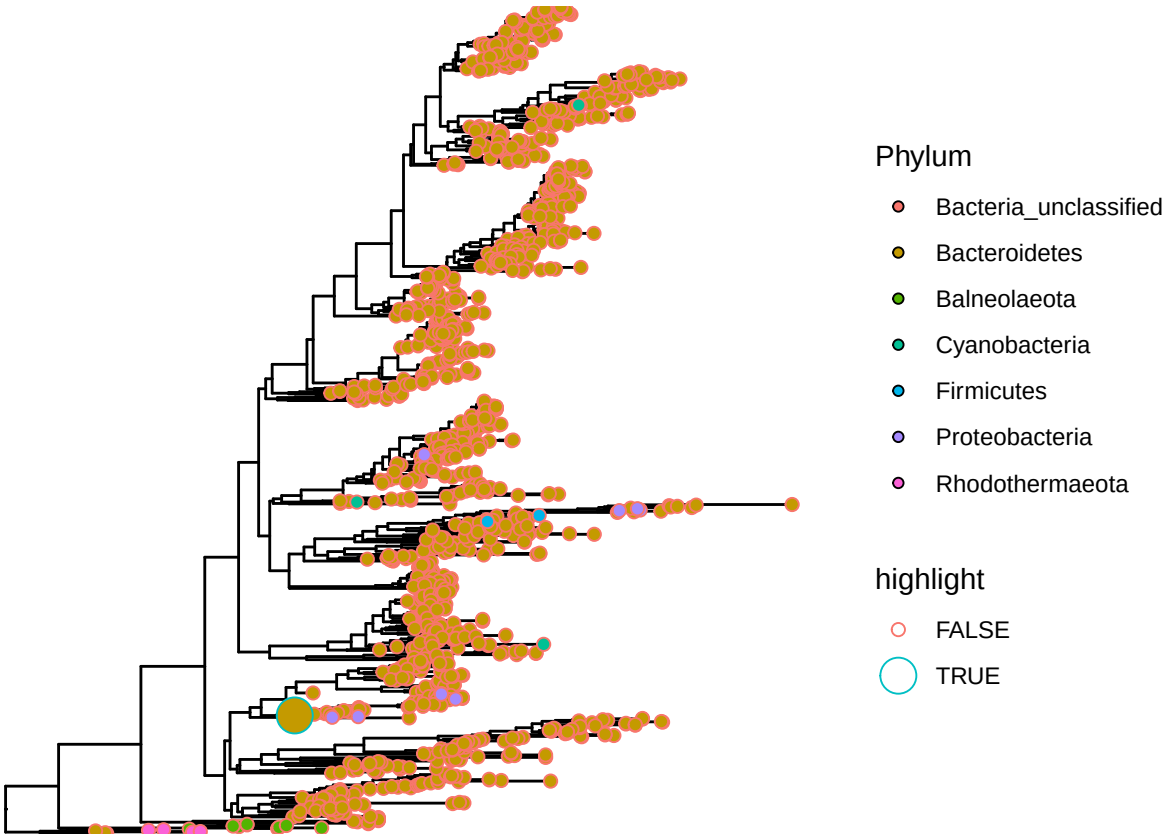



```
get_n_parents(Parathaa_kSGB_tree, 2456, 10)
```

```
## [1] 13287
```

```
subtree_parent <- get_subtree_plot_data(Parathaa_kSGB_tree, 13287, isTip=FALSE,
                                        highlight_lab = "MDNOLOHA_04242_16S_ribosomal_RNA|M1873143409")

ggtree(as.phylo(subtree_parent)) %<+% subtree_parent + geom_tippoint(aes(fill=Phylum, size=highlight, c
```

```
## Warning: Using size for a discrete variable is not advised.
```



Bad placement??

```
parthaa_SEED_placement_nodes <- plot_placement(Parathaa_SEED_tree, "Phylum", Parathaa_SEED_placements,
                                               label="00fd4ac0447c6b6539db1bcc40240e81")

## placed at a bifurcated branch so it won't plot...
#parthaa_SEED_placement_nodes

get_n_parents(Parathaa_SEED_tree, 2692, 10)
```

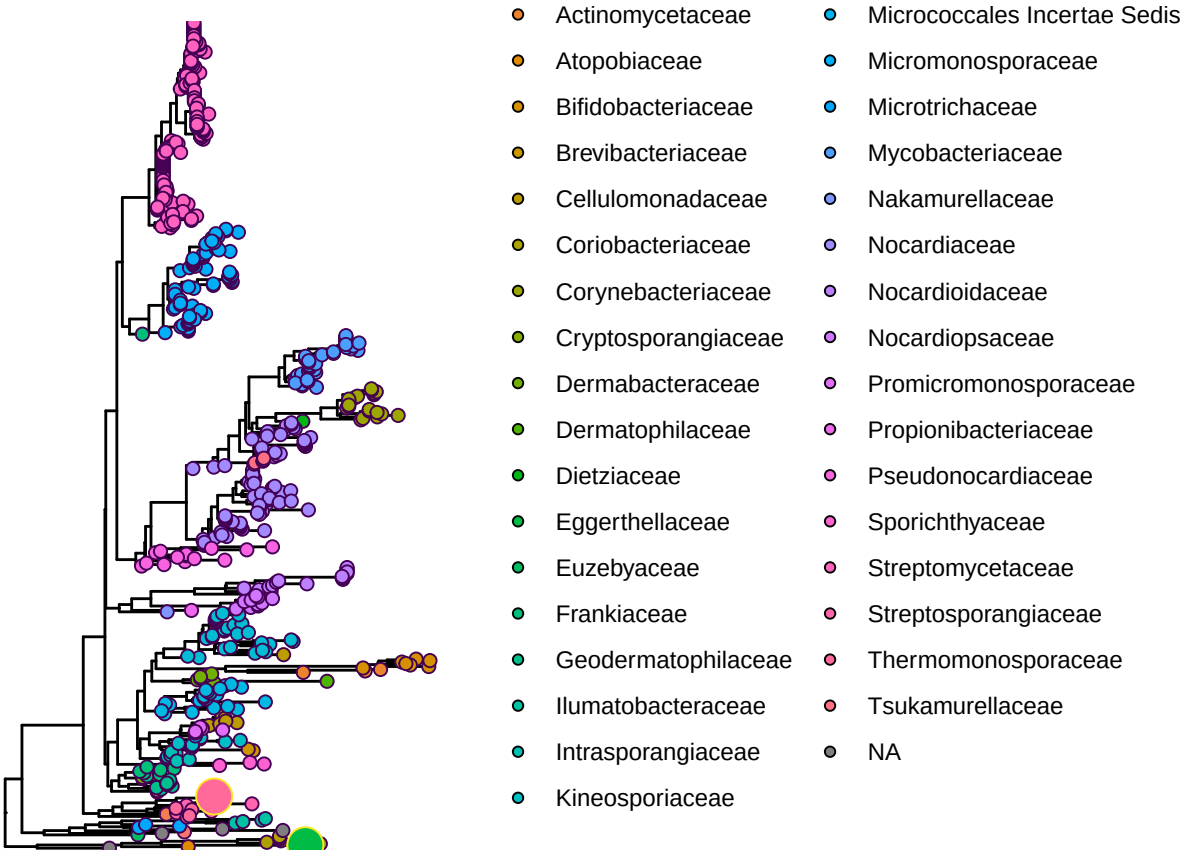
```
## [1] 6921
```

```
subtree_parent <- get_subtree_plot_data(Parathaa_SEED_tree, 6920, isTip=FALSE,
                                        highlight_lab = c("AF116563.GWJChro2", "AB011817.I97Lent5"),
                                        level="Family")

get_mrca_of_set(as.phylo(Parathaa_SEED_tree), c(2717, 2692))
```

```
## [1] 6920
```

```
ggtree(as.phylo(subtree_parent)) %<+% subtree_parent + geom_tippoint(aes(fill=Phylum, size=highlight, c
```

```
## Warning: Using size for a discrete variable is not advised.
```



This placement looks fine at genus level but is pretty far away from Eggerthelleace at the family level (which is its most likely origin given the dada2 and blast search)