

Reproducible Analysis: The Two-Wing admixed structure of environmental microbial communities

Yushi Tang

2022-08-25

```
library(ggplot2)      # for generating plots
library(latex2exp)    # for plot text latex
library(cowplot)      # for merging plots
library(gridExtra)    # for gridding plots
library(grid)         # for gridding plots
library(stringr)      # for uppercase first letter
library(ggh4x)        # for grid plot with free axis
```

Preface

This document provides reproducible research records for our manuscript about the Two-Wing admixed structure of environmental microbial communities. We provide step-by-step instructions for main results in both the original article and the supplemental materials.

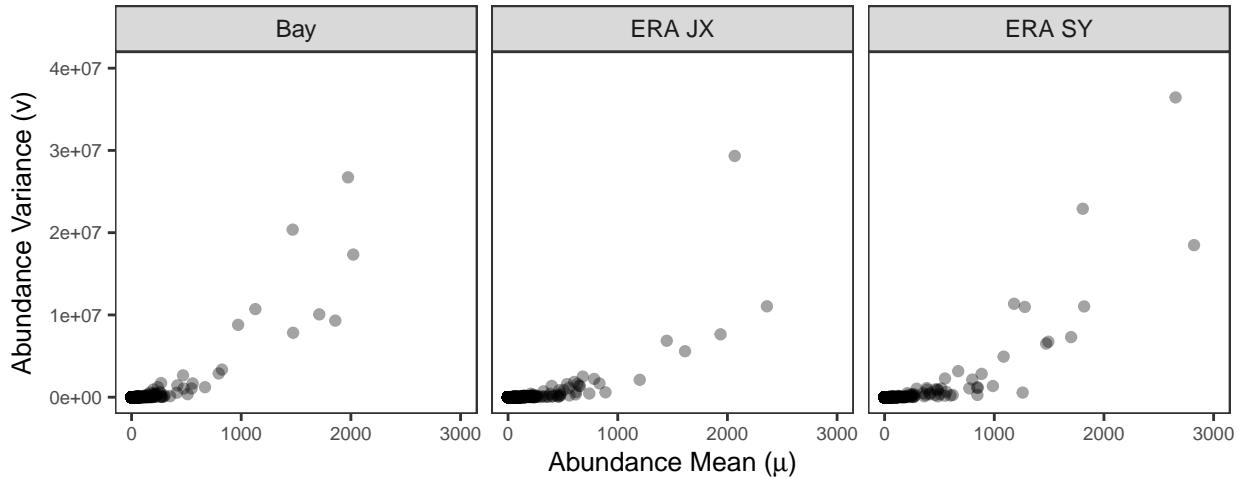
Main Figures

Figure 1. Observed abundance distribution

The over-dispersion pattern of microbial communities

```
load('../output/HZ/hz_mean_var.RData')
dat_plt$Region <- factor(dat_plt$Region, levels=c('Bay','ERA JX','ERA SY'))
ggplot(dat_plt, aes(x=miseqmean, y=miseqvar)) +
  geom_point(alpha=0.36, size=1.5) +
  xlab(TeX('Abundance Mean ($\mu$)')) +
  ylab(TeX('Abundance Variance ($v$)')) +
  xlim(0,3000) +
  ylim(0,4e7) +
  theme_bw() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size = 10),
        axis.text.x = element_text(size = 7),
        axis.text.y = element_text(size = 7)) +
```

```
theme(aspect.ratio=1) +
facet_grid(~Region)
```



```
load('../output/AO_ASV/ao_mean_var.RData')
dat_plt$Industry <- factor(dat_plt$Industry, levels=c('Dye', 'Pharmaceutical', 'Pesticide'))
dat_plt$Process <- factor(dat_plt$Process, levels=c('Influent', 'Anoxic', 'Oxic', 'Effluent'))
ggplot(dat_plt, aes(x=miseqmean, y=miseqvar)) +
  geom_point(alpha=0.36, size=1.5) +
  xlab(TeX('Abundance Mean ($\mu$)')) +
  ylab(TeX('Abundance Variance ($v$)')) +
  xlim(0,1000) +
  ylim(0,4e6) +
  theme_bw() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.title.x = element_text(size = 10),
        axis.title.y = element_text(size = 10),
        axis.text.x = element_text(size = 7),
        axis.text.y = element_text(size = 7)) +
  theme(aspect.ratio=1) +
  facet_grid(Industry~Process)
```

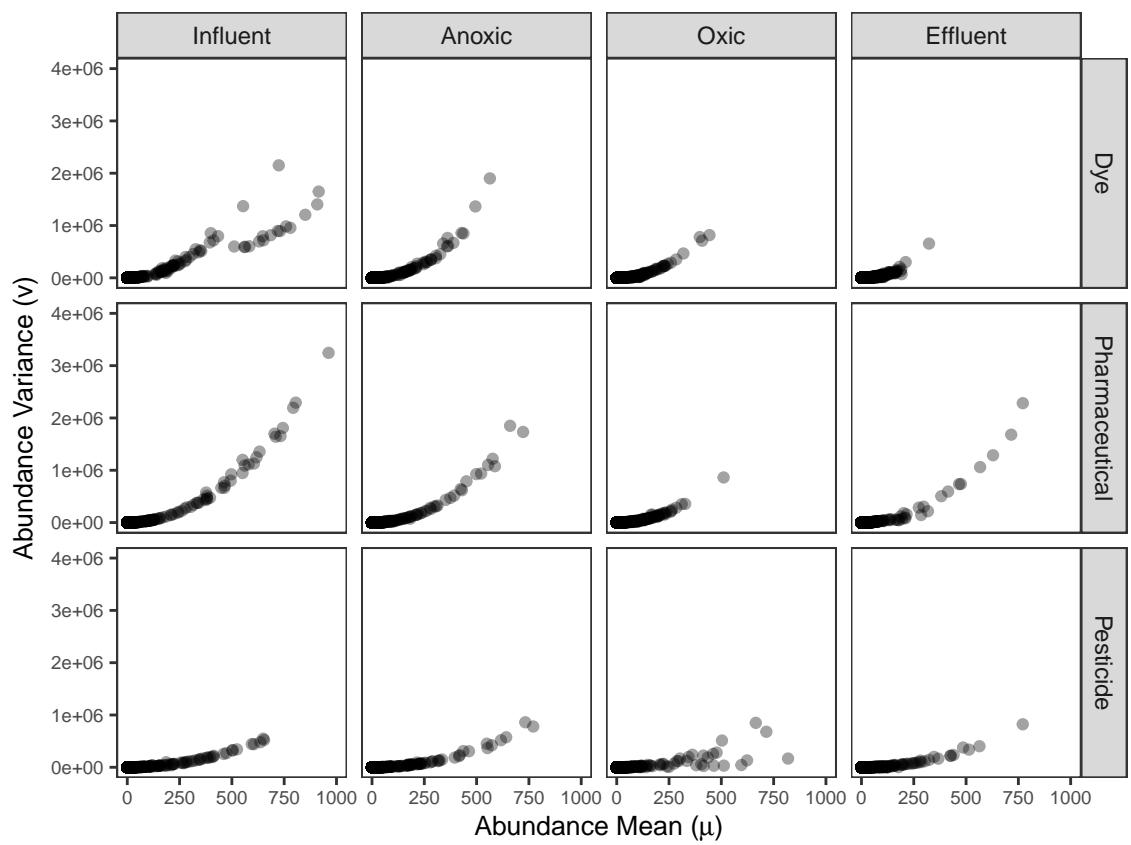
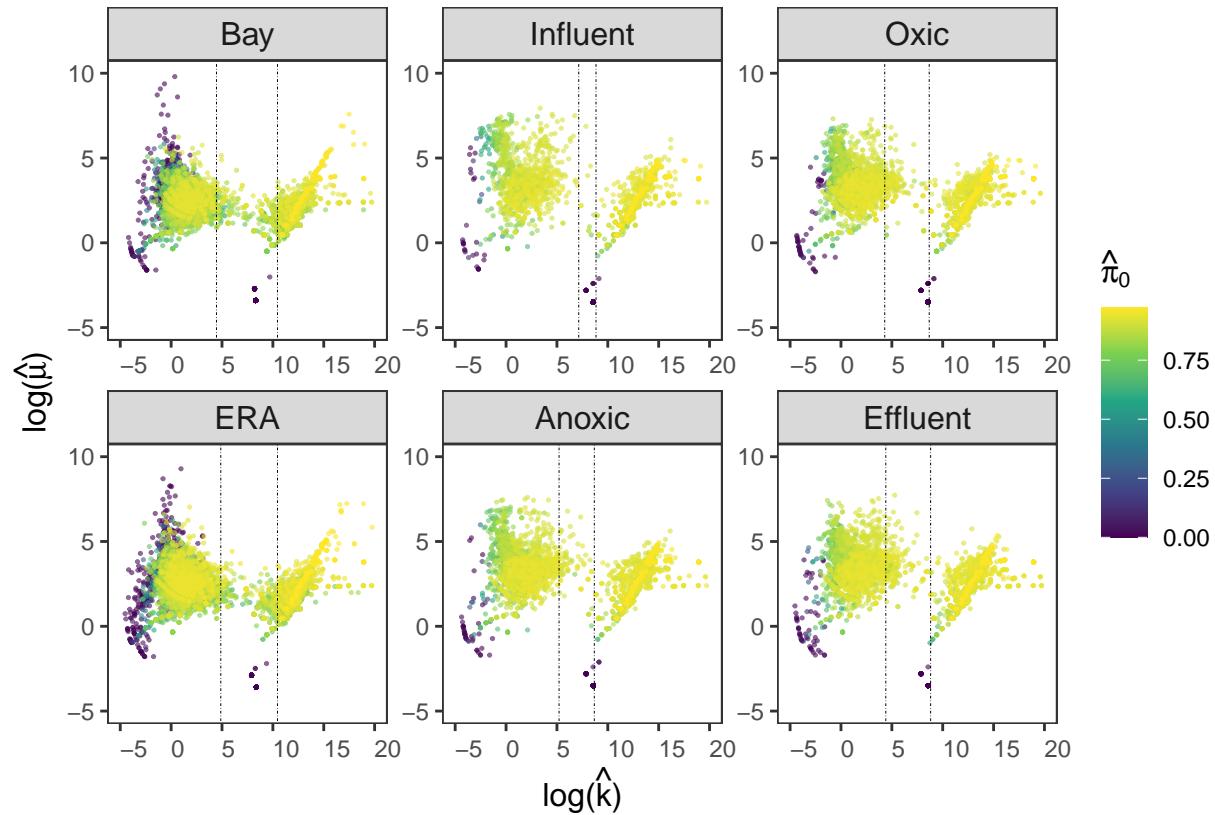


Figure 2. The Two-Wing admixed structure

Before boundary refining (Figure 2A)



After boundary refining (Figure 2B)

```
# load parameters mle
vec_process <- c('influent','anoxic','oxic','effluent')
param_plt <- data.frame()
for(item in vec_process){
  load(paste0('../output/AO_ASV/param_trio_asv_',item,'.RData'))
  load(paste0('../output/AO_ASV/mic_od_',item,'.RData'))
  param_trio$ID <- rownames(param_trio)
  # extract gamma-poisson distributed microbes
  param_gpm <- param_trio[param_trio$k!=Inf,]
  # annotate procedure
  param_gpm$Procedure <- str_to_title(item)
  # annotate taxa category
  load(paste0('../output/AO_ASV/mic_id_mu_k_',item,'.RData'))
  param_gpm$Wing <- ifelse(param_gpm$ID %in% id_mu, 'mu', 'Other')
  param_gpm$Wing <- ifelse(param_gpm$ID %in% id_k, 'k', param_gpm$Wing)
  param_plt <- rbind(param_plt, param_gpm)
}

vec_process <- c('bay','era')
```

```

for(item in vec_process){
  load(paste0('../output/HZ/param_trio_',item,'.RData'))
  load(paste0('../output/HZ/mic_od_',item,'.RData'))
  param_trio$ID <- rownames(param_trio)
  # extract gamma-poisson distributed microbes
  param_gpm <- param_trio[param_trio$k!=Inf,]
  # annotate procedure
  param_gpm$Procedure <- str_to_title(item)
  # annotate taxa category
  load(paste0('../output/HZ/mic_id_mu_k_',item,'.RData'))
  param_gpm$Wing <- ifelse(param_gpm$ID %in% id_mu, 'mu', 'Other')
  param_gpm$Wing <- ifelse(param_gpm$ID %in% id_k, 'k', param_gpm$Wing)

  param_plt <- rbind(param_plt, param_gpm)
}

rownames(param_plt) <- c(1:nrow(param_plt))
param_plt$Procedure <- ifelse(param_plt$Procedure=='Era','ERA',param_plt$Procedure)

param_plt$Procedure <- factor(param_plt$Procedure,
                               levels=c('Bay','Influent','Oxic',
                                       'ERA','Anoxic','Effluent'))
param_plt$Wing <- factor(param_plt$Wing,
                           levels=c('mu','k', 'Other'))

ggplot() +
  geom_point(data=param_plt,
             aes(x=log(k), y=log(mu), color=Wing), alpha=0.36, size=0.6) +
  ylab(TeX('\\log{(\hat{\mu})}')) +
  xlab(TeX('\\log{k}')) +
  ggtitle(TeX('')) +
  xlim(-5,20) +
  ylim(-5,10) +
  theme_bw() +
  theme(aspect.ratio=1) +
  facet_wrap(Procedure~, scale='free') +
  scale_colour_manual(values = c('#984ea3','#4daf4a'),
                      labels = unname(TeX(c('$\\mu$', '$k$')))) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12),
        axis.text.x = element_text(size = 9, vjust = 0.5, hjust = 0),
        axis.text.y = element_text(size=9),
        strip.text = element_text(size = 12),
        legend.title = element_text(color = "black", size = 10),
        legend.text = element_text(color = "black", size = 9)) +
  guides(color = guide_legend(override.aes = list(size = 1.8)))

```

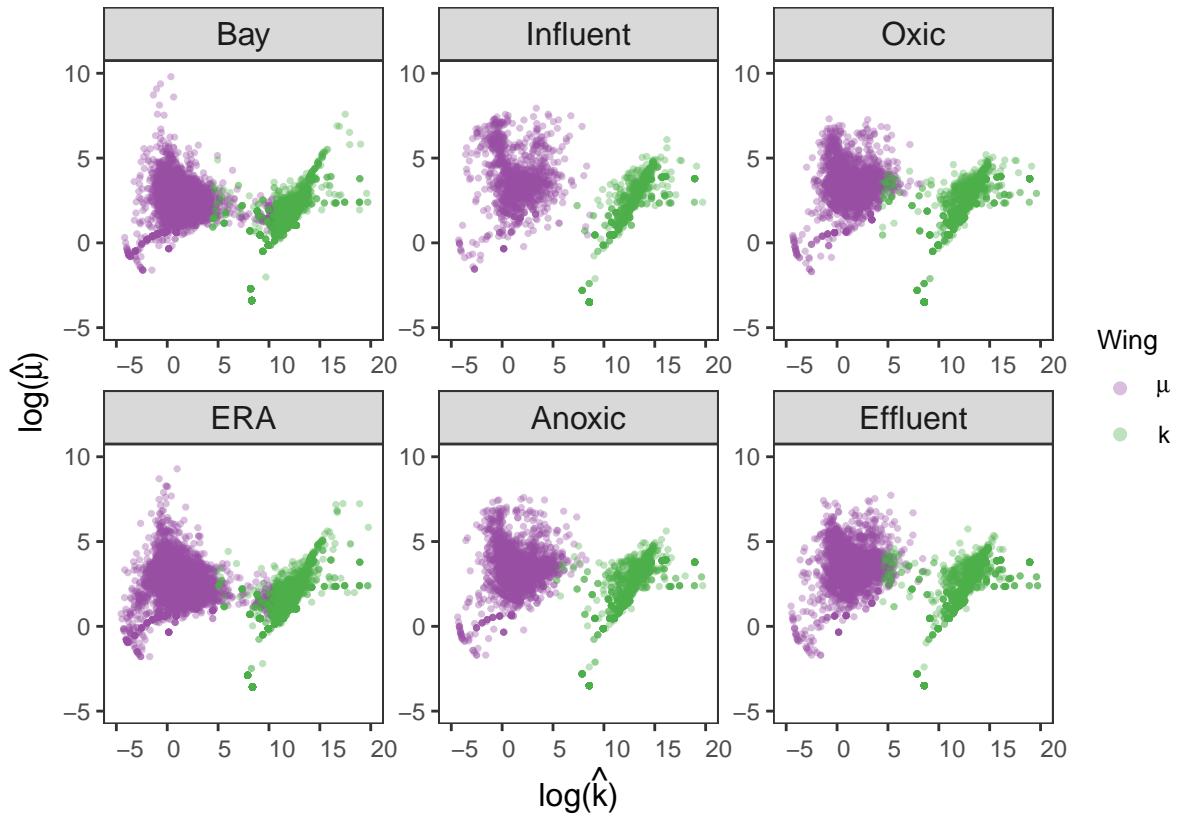


Figure 3. Dimensionality reduction analysis of μ -wing and k -wing sub-community PCA

```
# for bay
bay_plt <- data.frame()
lab_site <- c(rep(c('HB1', 'HB2', 'HB3', 'HB4', 'HB5', 'HB6', 'HB7', 'HB8', 'HB9', 'HB10'), each=3))
load('../output/HZ/pca_bay.RData')
pcs_mu <- data.frame(t(pca_mu$pc))
pcs_k <- data.frame(t(pca_k$pc))
pcs_all <- data.frame(t(pca_all$pc))
bay_plt <- rbind(bay_plt, data.frame(PC1=pcs_mu$PC1,
                                         PC2=pcs_mu$PC2,
                                         Site=lab_site,
                                         Wing='mu-wing',
                                         Region='Bay'))
bay_plt <- rbind(bay_plt, data.frame(PC1=pcs_k$PC1,
                                         PC2=pcs_k$PC2,
                                         Site=lab_site,
                                         Wing='k-wing',
                                         Region='Bay'))
bay_plt <- rbind(bay_plt, data.frame(PC1=pcs_all$PC1,
                                         PC2=pcs_all$PC2,
                                         Site=lab_site,
                                         Wing='all',
                                         Region='Bay'))
```

```

bay_plt$Site <- factor(bay_plt$Site,
                        levels=c('HB1', 'HB2', 'HB3', 'HB4', 'HB5', 'HB6', 'HB7', 'HB8', 'HB9', 'HB10'))
bay_plt$Wing <- factor(bay_plt$Wing,
                        levels=c('all', 'mu-wing', 'k-wing'))

g_bay <- ggplot(bay_plt, aes(x=PC1, y=PC2, color=Site)) +
  geom_point(size=0.8, alpha=0.8) +
  theme_bw() +
  xlab(TeX('PC1')) +
  ylab(TeX('PC2')) +
  ggtitle('') +
  theme(plot.title=element_text(hjust=0, size=rel(0.0)),
        aspect.ratio=1,
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12),
        axis.text.x = element_text(size = 8),
        axis.text.y = element_text(size = 8),
        legend.title = element_text(color = "black", size = 12),
        legend.text = element_text(color = "black", size = 8),
        strip.text = element_text(size = 12)) +
  scale_colour_manual(values=c('#a6cee3', '#1f78b4', '#b2df8a', '#33a02c',
                             '#fb9a99', '#e31a1c', '#fdbf6f', '#ff7f00',
                             '#cab2d6', '#6a3d9a')) +
  facet_grid2(Wing~Region, scales='free',
              independent = "all", labeller='label_parsed')

# for era
era_plt <- data.frame()
lab_site <- c(rep(c('SY1', 'SY2', 'SY3', 'SY4', 'SY5', 'SY6',
                    'JX1', 'JX2', 'JX3', 'JX4', 'JX5', 'JX6'), each=3))
load('../output/HZ/pca_era.RData')
pcs_mu <- data.frame(t(pca_mu$pc))
pcs_k <- data.frame(t(pca_k$pc))
pcs_all <- data.frame(t(pca_all$pc))
era_plt <- rbind(era_plt, data.frame(PC1=pcs_mu$PC1,
                                       PC2=pcs_mu$PC2,
                                       Site=lab_site,
                                       Wing='mu-wing',
                                       Region='ERA'))
era_plt <- rbind(era_plt, data.frame(PC1=pcs_k$PC1,
                                       PC2=pcs_k$PC2,
                                       Site=lab_site,
                                       Wing='k-wing',
                                       Region='ERA'))
era_plt <- rbind(era_plt, data.frame(PC1=pcs_all$PC1,
                                       PC2=pcs_all$PC2,
                                       Site=lab_site,
                                       Wing='all',
                                       Region='ERA'))
era_plt$Site <- factor(era_plt$Site,
                        levels=c('SY1', 'SY2', 'SY3', 'SY4', 'SY5', 'SY6', 'JX1', 'JX2', 'JX3', 'JX4', 'JX5', 'JX6'))

```

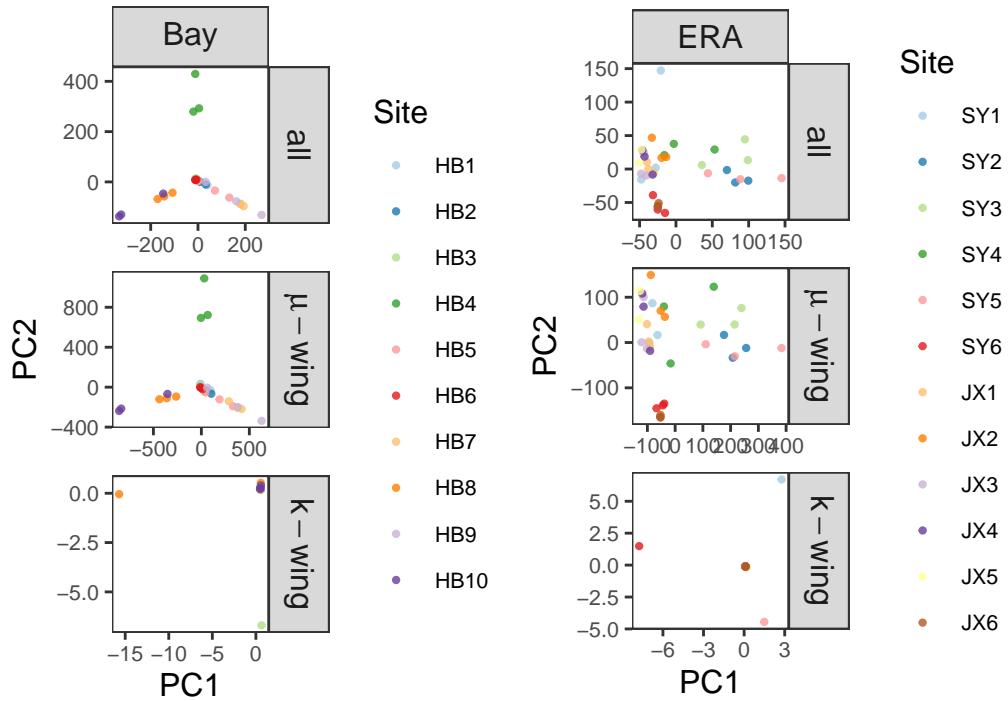
```

era_plt$Wing <- factor(era_plt$Wing,
                       levels=c('all','mu-wing','k-wing'))

g_era <- ggplot(era_plt, aes(x=PC1, y=PC2, color=Site)) +
  geom_point(size=0.8, alpha=0.8) +
  theme_bw() +
  xlab(TeX('PC1')) +
  ylab(TeX('PC2')) +
  ggtitle('') +
  theme(plot.title=element_text(hjust=0, size=rel(0.0)),
        aspect.ratio=1,
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12),
        axis.text.x = element_text(size = 8),
        axis.text.y = element_text(size = 8),
        legend.title = element_text(color = "black", size = 12),
        legend.text = element_text(color = "black", size = 8),
        strip.text = element_text(size = 12)) +
  scale_colour_manual(values=c('#a6cee3','#1f78b4','#b2df8a','#33a02c',
                             '#fb9a99','#e31a1c','#fdbf6f','#ff7f00',
                             '#cab2d6','#6a3d9a','#ffff99','#b15928')) +
  facet_grid2(Wing~Region, scales='free',
              independent = "all",labeleller='label_parsed')

plot_grid(NULL, g_bay, g_era, NULL,
          rel_widths = c(1,5,5,1), nrow=1, align='v')

```



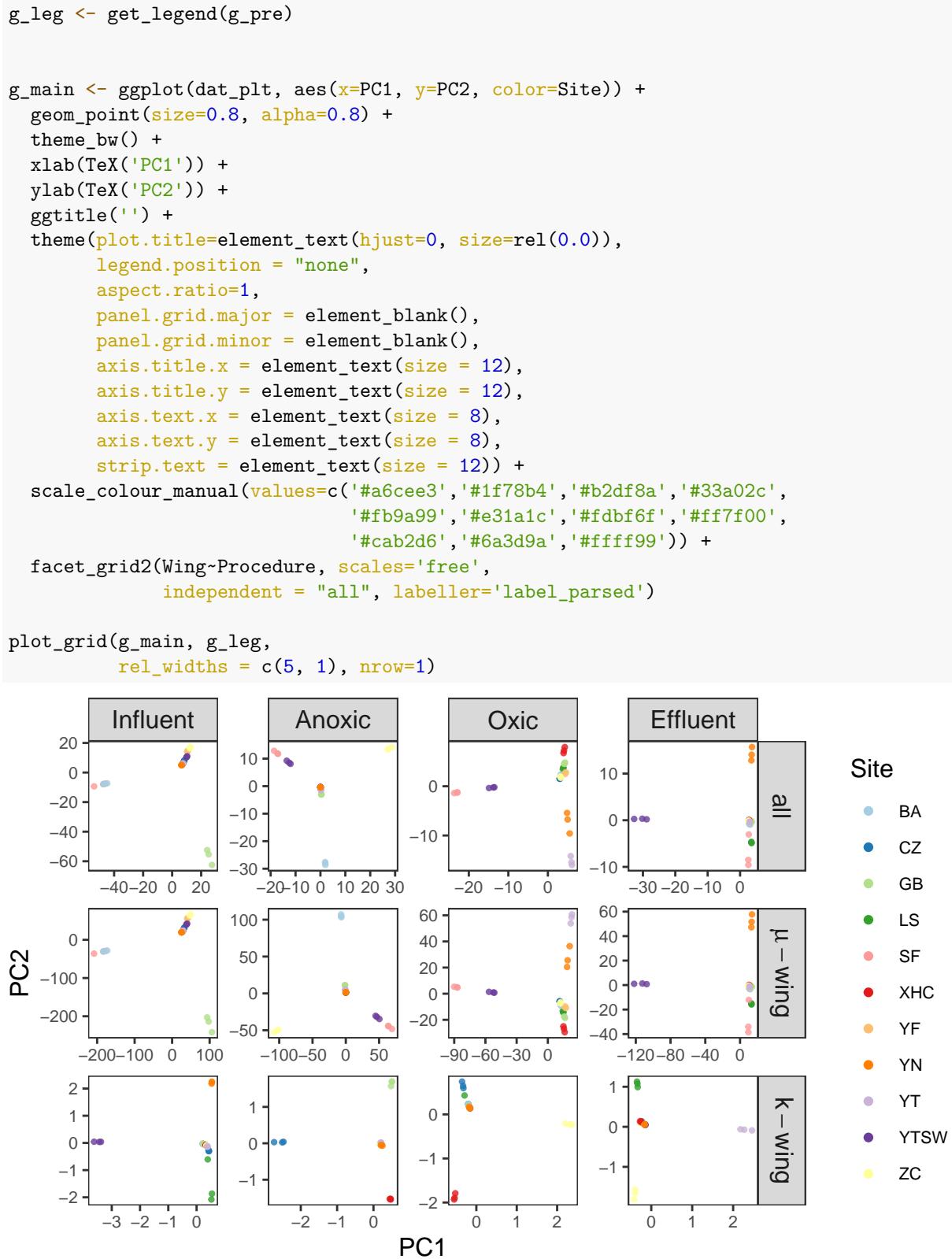
```

dat_plt <- data.frame()
vec_process <- c('influent', 'anoxic', 'oxic', 'effluent')
lab_site <- c(rep(c('LS', 'SF', 'CZ', 'BA', 'YF', 'GB', 'YTSW', 'XHC', 'ZC', 'YT', 'YN'), each=3))
for(item in vec_process){
  load(paste0('../output/AO_ASV/pca_', item, '.RData'))
  pcs_mu <- data.frame(t(pca_mu$pc))
  pcs_k <- data.frame(t(pca_k$pc))
  pcs_all <- data.frame(t(pca_all$pc))
  dat_plt <- rbind(dat_plt, data.frame(PC1=pcs_mu$PC1,
                                         PC2=pcs_mu$PC2,
                                         Site=lab_site,
                                         Wing='mu-wing',
                                         Procedure=str_to_title(item)))
  dat_plt <- rbind(dat_plt, data.frame(PC1=pcs_k$PC1,
                                         PC2=pcs_k$PC2,
                                         Site=lab_site,
                                         Wing='k-wing',
                                         Procedure=str_to_title(item)))
  dat_plt <- rbind(dat_plt, data.frame(PC1=pcs_all$PC1,
                                         PC2=pcs_all$PC2,
                                         Site=lab_site,
                                         Wing='all',
                                         Procedure=str_to_title(item)))
}

dat_plt$Wing <- factor(dat_plt$Wing, levels=c('all', 'mu-wing', 'k-wing'))
dat_plt$Procedure <- factor(dat_plt$Procedure, levels=str_to_title(vec_process))

g_pre <- ggplot(dat_plt, aes(x=PC1, y=PC2, color=Site)) +
  geom_point() +
  theme_bw() +
  xlab(TeX('PC1')) +
  ylab(TeX('PC2')) +
  ggtitle('') +
  theme(plot.title=element_text(hjust=0, size=rel(0.0)),
        #legend.position = "none",
        aspect.ratio=1,
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12),
        axis.text.x = element_text(size = 8),
        axis.text.y = element_text(size = 8),
        legend.title = element_text(color = "black", size = 12),
        legend.text = element_text(color = "black", size = 8),
        strip.text = element_text(size = 12)) +
  scale_colour_manual(values=c('#a6cee3', '#1f78b4', '#b2df8a', '#33a02c',
                               '#fb9a99', '#e31a1c', '#fdbf6f', '#ff7f00',
                               '#cab2d6', '#6a3d9a', '#ffff99')) +
  facet_grid2(Wing~Procedure, scales='free',
             labeller='label_parsed')

```



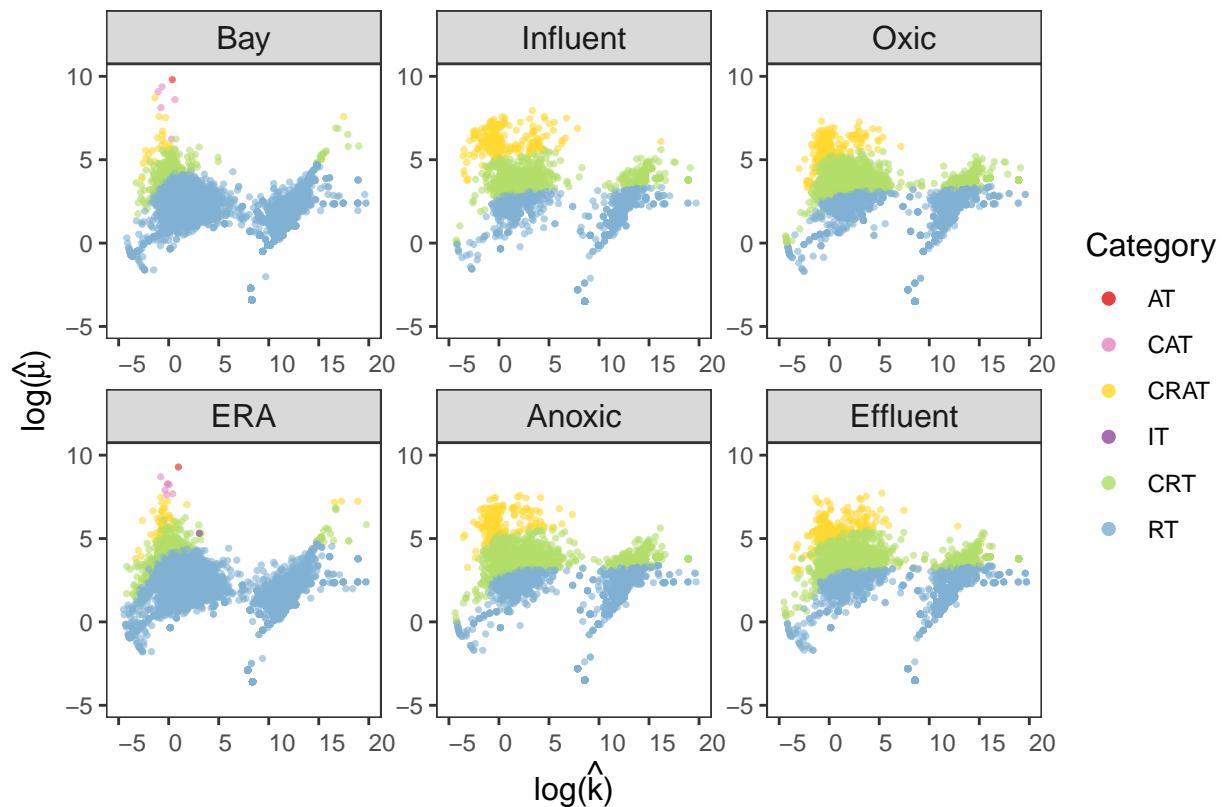
PCoA and CCA

Run `../figures/vignettes/pcoa_hz.Rmd` and `../figures/vignettes/pcoa_ao.Rmd`

CCA Run `../figures/vignettes/cca_ao_ori.Rmd`

Figure 4. The Two-Wing structure generalizes relative-abundance-based categorizing method

Using *Method 0* to identify microbial categories



Using *Method a-c* to identify microbial categories

The algorithm is similar as above. Specifically, run `../figures/vignettes/main/mu_k_category_abcd.Rmd`

```
sessionInfo()
```

```
## R version 4.2.0 (2022-04-22)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Catalina 10.15.3
##
## Matrix products: default
## BLAS:    /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
## LAPACK:  /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
##
## locale:
```

```
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] grid      stats     graphics grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
## [1] ggh4x_0.2.1    stringr_1.4.0   gridExtra_2.3   cowplot_1.1.1
## [5] latex2exp_0.9.4 ggplot2_3.3.6
##
## loaded via a namespace (and not attached):
## [1] highr_0.9       pillar_1.8.0    compiler_4.2.0  tools_4.2.0
## [5] digest_0.6.29   viridisLite_0.4.0 evaluate_0.15   lifecycle_1.0.1
## [9] tibble_3.1.8    gtable_0.3.0   pkgconfig_2.0.3 rlang_1.0.4
## [13] cli_3.3.0      rstudioapi_0.13 yaml_2.3.5     xfun_0.31
## [17] fastmap_1.1.0   withr_2.5.0    dplyr_1.0.9    knitr_1.39
## [21] generics_0.1.3  vctrs_0.4.1    tidyselect_1.1.2 glue_1.6.2
## [25] R6_2.5.1       fansi_1.0.3    rmarkdown_2.14  farver_2.1.1
## [29] purrr_0.3.4    magrittr_2.0.3   scales_1.2.0    htmltools_0.5.3
## [33] colorspace_2.0-3 labeling_0.4.2  utf8_1.2.2     stringi_1.7.8
## [37] munsell_0.5.0
```