Approximating Observed Microbiome Data with `microbiomeDASim`

Justin Williams

1/20/2020

An important goal of the Bioconductor package `microbiomeDASim` is the ability to closely approximate real data from longitudinal experiements where sequencing was performed. To demonstrate the ability we will approximate observed data from a longitudinal study on the human gut microbiome in gnotobiotic mice (Turnbaugh et. al., 2009 doi.10.1126/scitranslmed.3000322). This data file is available within the `metagenomeSeq` Bioconductor package.

As a first step we will load the necessary packages used during this analysis.

```r
require(metagenomeSeq)
require(tidyverse)
require(ggplot2)
devtools::install_github("williazo/microbiomeDASim")
require(microbiomeDASim)
```

We next load in the mouse data. This dataset is stored as an MRexperiment object with assay data collected at the OTU level. The raw counts represent over 10,000 OTUs that were sequences from a total of 139 samples. These samples represent repeated measurements taken on 12 gnotobiotic mice. All mice were fed the same low-fat, plant polysaccharide–rich diet for the first 21 days of the study. At this point 6 of the mice were then switched to a high-fat, high-sugar “Western” diet. The subseqeuent changes in the microbial community were then observed over a follow-up of roughly 60 days.

```r
# loading in longitudinal microbiome data from Turnbaugh et. al 2009
data("mouseData")
mouseData
```

```
## MRexperiment (storageMode: environment)
## assayData: 10172 features, 139 samples
##   element names: counts
## protocolData: none
## phenoData
##   sampleNames: PM1:20080107 PM1:20080108 ... PM9:20080303 (139 total)
##   varLabels: mouseID date ... status (5 total)
##   varMetadata: labelDescription
## featureData
##   featureNames: Prevotellaceae:1 Lachnospiraceae:1 ...
##     Parabacteroides:956 (10172 total)
##   fvarLabels: superkingdom phylum ... OTU (7 total)
##   fvarMetadata: labelDescription
## experimentData: use 'experimentData(object)'
## Annotation:
```

Many of the OTUs are low frequency observations that do not match out multivariate normal distributional assumptions, especially given our limited sample size (n=12). Therefore, we restrict our analysis to genus level counts rather than OTU level counts, which reduces the feature size from 10,172 to 61.

```
#aggregating the counts to the genus level
genus_mouseData <- metagenomeSeq::aggTax(mouseData, lvl="genus")
genus_mouseData
```

```
## MRexperiment (storageMode: environment)
## assayData: 61 features, 139 samples
##   element names: counts
## protocolData: none
## phenoData
##   sampleNames: PM1:20080107 PM1:20080108 ... PM9:20080303 (139 total)
##   varLabels: mouseID date ... status (5 total)
##   varMetadata: labelDescription
## featureData
##   featureNames: Acetanaerobacterium Acinetobacter ... Xylanibacter (61
##     total)
##   fvarLabels: superkingdom phylum ... genus (6 total)
##   fvarMetadata: labelDescription
## experimentData: use 'experimentData(object)'
## Annotation:
```

We further apply presence and depth filters, and then log normalize the counts to represent our simulated outcome of interest.

```
#additional count filters
genus_mouseData <- filterData(genus_mouseData, present = 10, depth = 1000)
genus_mouseData
```

```
## MRexperiment (storageMode: environment)
## assayData: 35 features, 137 samples
##   element names: counts
## protocolData: none
## phenoData
##   sampleNames: PM1:20080108 PM1:20080114 ... PM9:20080303 (137 total)
##   varLabels: mouseID date ... status (5 total)
##   varMetadata: labelDescription
## featureData
##   featureNames: Akkermansia Alistipes ... Turicibacter (35 total)
##   fvarLabels: superkingdom phylum ... genus (6 total)
##   fvarMetadata: labelDescription
## experimentData: use 'experimentData(object)'
## Annotation:
```

```
g_lnorm_mat <- MRcounts(genus_mouseData, norm=TRUE, log=TRUE)
```

```
## Default value being used.
```

We next randomly select a genus to use as our reference. In this case we select *Sutterella*, and this row of the log normalized count matrix will be our outcome.

```
set.seed(012020)
ex_feature <- sample(seq_len(dim(g_lnorm_mat)[1]), 1)
row.names(g_lnorm_mat)[ex_feature] #Sutterella
```

```
## [1] "Sutterella"
```

```
y <- g_lnorm_mat[ex_feature, ]
```

We then combine the observed outcome with phenotype data on the samples which includes their ID, time since study began, date sample collected, and diet. Note that the diet variable is initially saved as a time-varying covariate, but we replace all labels below 21 days as Western only for the treatment group. In our simulation step we will assume that the two groups are have identical longitudianl trends from day 0 to day 21 and that there will be subsequent changes after this date.

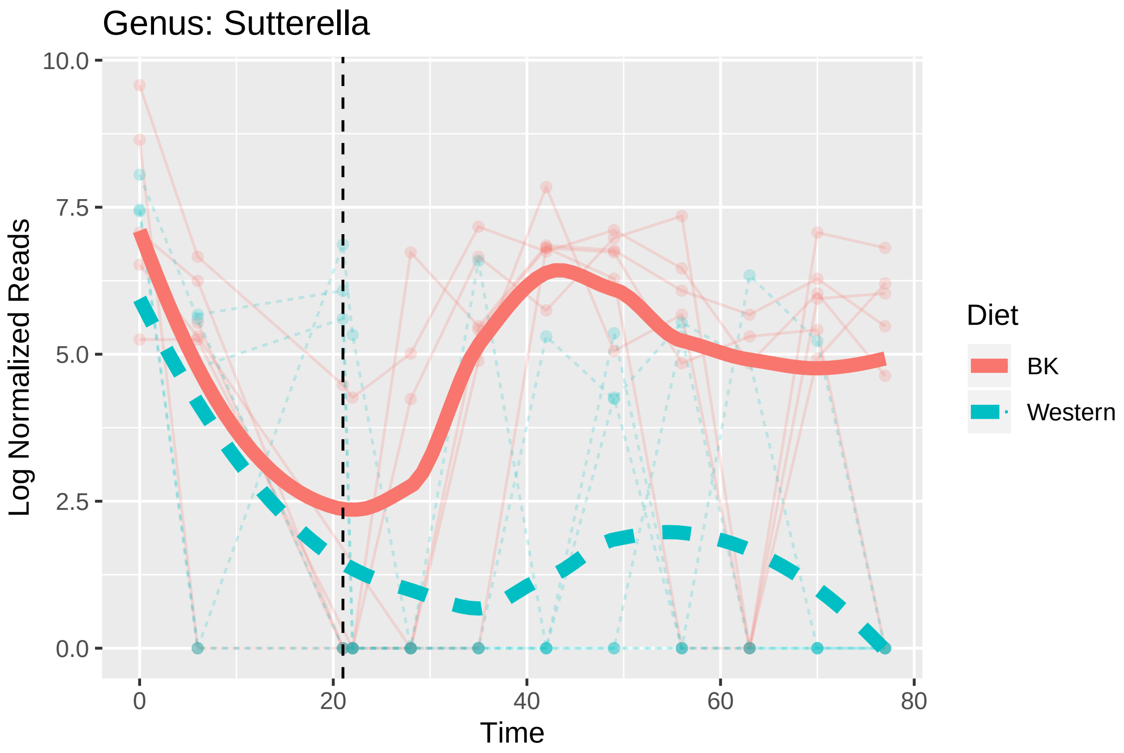
```
obs_feat <- data.frame(y, pData(genus_mouseData))
obs_feat <- obs_feat[order(obs_feat$mouseID, obs_feat$relativeTime), ]
names(obs_feat)
```

```
## [1] "y"            "mouseID"      "date"         "diet"         "relativeTime"
## [6] "status"
```

```
obs_id_split <- split(obs_feat, obs_feat$mouseID)
obs_df <- lapply(obs_id_split, function(x){
    x$diet <- ifelse(any(x$diet=="Western"), "Western", x$diet)
    return(x)
})
obs_df <- data.frame(do.call(rbind, obs_df))
obs_df$diet <- as.factor(obs_df$diet)
```

We can graph the longidutinal trends to see what sort of longitudinal differential abundant trends might be occuring.

```
obs_plot <- with(obs_df, microbiomeDASim::ggplot_spaghetti(y, mouseID, relativeTime, group=diet))+
    xlab("Time")+
    ylab("Log Normalized Reads")+
    scale_color_discrete(name="Diet")+
    scale_linetype_discrete(name="Diet")+
    ggtitle(paste0("Genus: ", row.names(genus_mouseData[ex_feature, ])))+
    geom_vline(xintercept=21, col="black", lty=2)
obs_plot
```



As both groups are changing over time we can use the `metagenomeSeq` package to obtain an estimate of the differential abundance between the groups using smoothing-spline ANOVA methodology. This allows us to directly visualize the differences between the two groups across time.

```
#smoothing spline estimate
ss_est <- fitTimeSeries(obj=genus_mouseData, formula=abundance~time*class,
                        feature=ex_feature, class="diet", id="mouseID",
                        time="relativeTime", norm=TRUE, log=TRUE, random=~1|id,
                        B=1000)
```

```
## Loading required namespace: gss
```

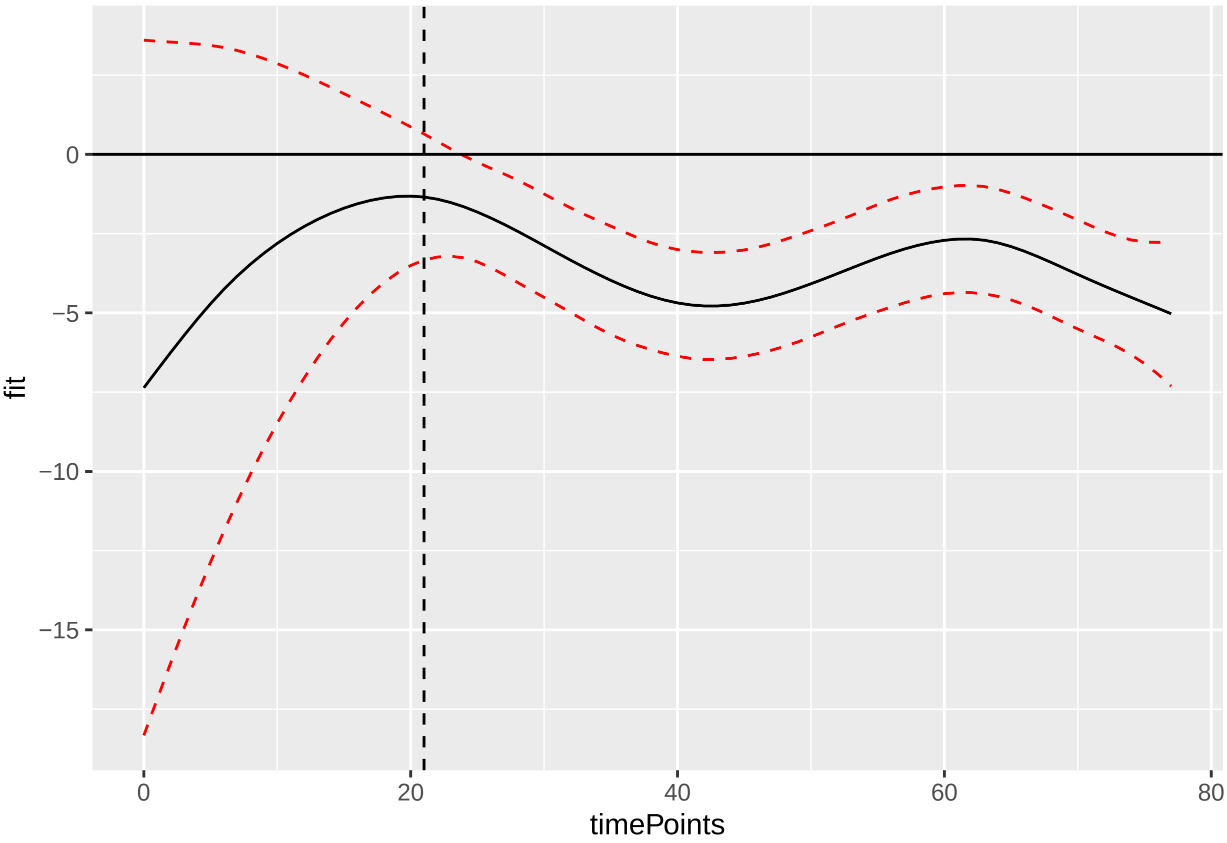
```
## Default value being used.
```

```
## [1] 100
## [1] 200
## [1] 300
## [1] 400
## [1] 500
## [1] 600
## [1] 700
## [1] 800
## [1] 900
## [1] 1000
```

```
ss_est$timeIntervals
```

```
##      Interval start Interval end      Area     p.value
## [1,]             24           77 -192.6571 0.008991009
```

```
mouse_metaSplines_est <- ggplot(data=ss_est$fit, aes(x=timePoints, y=fit))+
    geom_line()+
    geom_line(col="red", lty=2, aes(y=fit+1.96*se))+
    geom_line(col="red", lty=2, aes(y=fit-1.96*se))+
    geom_hline(col="black", yintercept=0)+
    geom_vline(xintercept=21, col="black", lty=2)
mouse_metaSplines_est
```



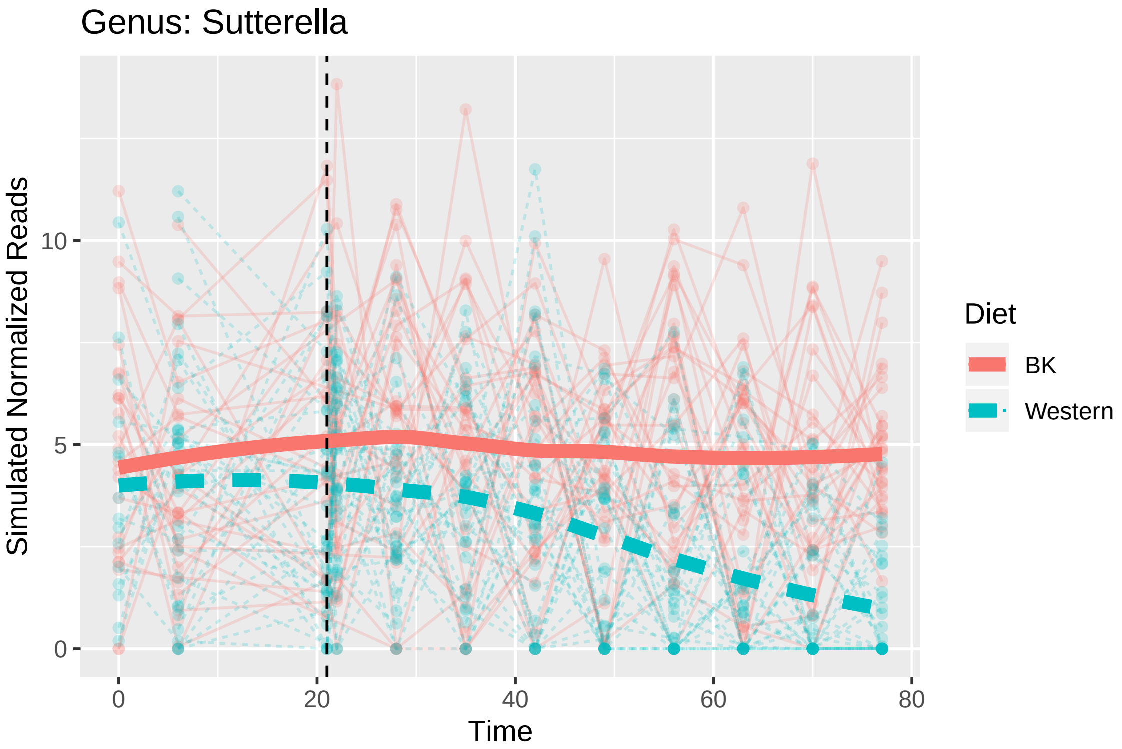
In this case we see that the confidence interval for the difference from day 0 to day 21 has no statistically significant differences. However, there are differences detected from day 24 to day 77, with the "Western" group having lower log normalized counts compared to the "BK" group. While the trend itself is somewhat non-linear, we can think that the true underlying process may be linearly decreasing from the time of diet intervention. We can simulate this process using our L_up functional form, where we assume no group differences over the first 21 days of observation and then a linearly decreasing

```
x_bar <- mean(obs_df$y[obs_df$diet=="BK"]) #control mean
s <- sd(obs_df$y) #overall standard deviation

#large scale study with 30 mice in each arm replicating each observed mice 5 times
obs_list <- lapply(unique(obs_df$mouseID), function(x){
    id_obs <- nrow(obs_df[obs_df$mouseID %in% x, ])
})
full_obs <- rep(unlist(obs_list), 5)
full_id <- unlist(mapply(x=paste0("PM", seq_len(60)), y=full_obs, function(x, y) rep(x, y)))
full_time <- rep(obs_df$relativeTime, 5)
full_diet <- rep(obs_df$diet, 5)
large_df <- data.frame(mouseID=full_id, relativeTime=full_time,
                       diet=full_diet)
large_df$diet <- factor(large_df$diet)
```

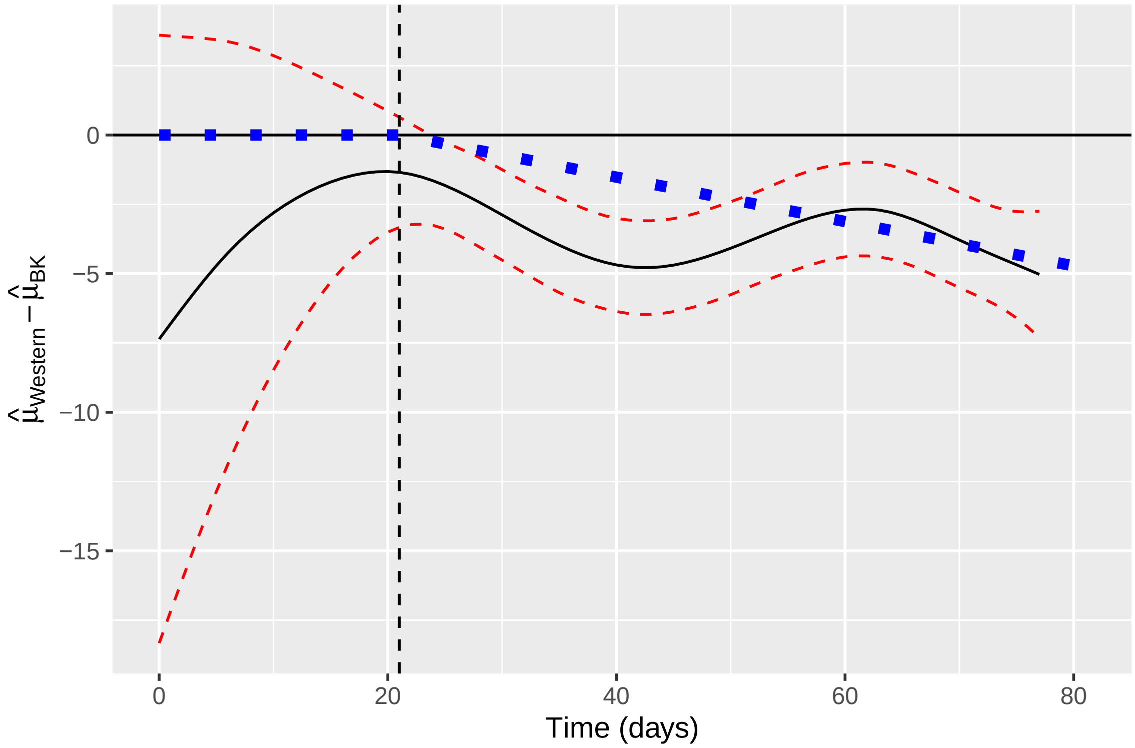
```
sim_mouse <- mvrnorm_sim_obs(id=large_df$mouseID, time=large_df$relativeTime,
                             group=large_df$diet, ref="BK", control_mean=x_bar, sigma=s,
                             rho=0.03, corr_str="ar1", func_form="L_up", IP=21, beta=-0.08)
```

```
mouse_sim_plot <- with(sim_mouse$df, microbiomeDASim::ggplot_spaghetti(Y, ID, time, group=group))+
    xlab("Time")+
    ylab("Simulated Normalized Reads")+
    scale_color_discrete(name="Diet")+
    scale_linetype_discrete(name="Diet")+
    ggtitle(paste0("Genus: ", row.names(genus_mouseData[ex_feature, ])))+
    geom_vline(xintercept=21, col="black", lty=2)
mouse_sim_plot
```



```
sim_t <- mean_trend(timepoints=seq(0, 81, 1), form="L_up", beta=-0.08, IP=21)

metaSplines_sim_trends <- ggplot(data=ss_est$fit, aes(x=timePoints, y=fit))+
    geom_line()+
    geom_line(col="red", lty=2, aes(y=fit+1.96*se))+
    geom_line(col="red", lty=2, aes(y=fit-1.96*se))+
    geom_hline(col="black", yintercept=0)+
    geom_vline(xintercept=21, col="black", lty=2)+
    geom_line(data=sim_t$trend, aes(x=timepoints, y=mu), col="blue", lty=3,
              lwd=2)+
    xlab("Time (days)")+
    ylab(bquote(hat(mu)["Western"]-hat(mu)["BK"]))
metaSplines_sim_trends
```


$\hat{\mu}_{Western} - \hat{\mu}_{BK}$
0
−5
−10
−15
0
20
40
60
80
Time (days)