UMD\_Natural\_Infection\_Analysis.R

jbueno

Thu Feb 14 11:20:51 2019

# EMIT\_UMD\_Natural\_Infection\_Analysis.R  
# Program Objective: Run analysis on EMIT Natural Infection datasets  
# Author: Jacob Bueno de Mesquita using material from Jing Yan and Don Milton  
# Date: December 14, 2018 -- February 2019  
# Summary: While most of the analysis for the EMIT\_UMD study was done in SAS (especially that as part of the PNAS manuscript), here are some others analyses that we will include as well. Here, we estimate the mean and variance for the ratio of ffu/copy number in fine aerosols from EMIT using a mixed model with random effect of person and fixed effect of day since onset of symptoms.   
  
#### Load required packages and set working directory ####  
  
library(tidyverse)  
library(RcppRoll)  
library(readxl)  
library(knitr)  
library(data.table)  
library(lubridate)  
library(arsenal)  
library(lme4)  
  
# setwd("/Users/jbueno/Box Sync/EMIT/EMIT\_Data\_Analysis\_Jake/EMIT\_UMD\_Natural\_Infection")  
# Commented out the setwd() command in order to properly run the markdown report compilation.  
  
sessionInfo() # for reproducibility

## R version 3.5.1 (2018-07-02)  
## Platform: x86\_64-apple-darwin15.6.0 (64-bit)  
## Running under: macOS 10.14.3  
##   
## Matrix products: default  
## BLAS: /System/Library/Frameworks/Accelerate.framework/Versions/A/Frameworks/vecLib.framework/Versions/A/libBLAS.dylib  
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/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] stats graphics grDevices utils datasets methods base   
##   
## other attached packages:  
## [1] bindrcpp\_0.2.2 lme4\_1.1-19 Matrix\_1.2-14   
## [4] arsenal\_1.5.0 lubridate\_1.7.4 data.table\_1.11.8  
## [7] knitr\_1.20 readxl\_1.1.0 RcppRoll\_0.3.0   
## [10] forcats\_0.3.0 stringr\_1.3.1 dplyr\_0.7.7   
## [13] purrr\_0.2.5 readr\_1.1.1 tidyr\_0.8.2   
## [16] tibble\_1.4.2 ggplot2\_3.1.0 tidyverse\_1.2.1   
## [19] htmlTable\_1.12 rmarkdown\_1.10 markdown\_0.8   
##   
## loaded via a namespace (and not attached):  
## [1] Rcpp\_0.12.19 lattice\_0.20-38 utf8\_1.1.4 assertthat\_0.2.0  
## [5] rprojroot\_1.3-2 digest\_0.6.18 R6\_2.3.0 cellranger\_1.1.0  
## [9] plyr\_1.8.4 backports\_1.1.2 evaluate\_0.12 highr\_0.7   
## [13] httr\_1.3.1 pillar\_1.3.0 rlang\_0.3.0.1 lazyeval\_0.2.1   
## [17] rstudioapi\_0.8 minqa\_1.2.4 nloptr\_1.2.1 checkmate\_1.8.5   
## [21] splines\_3.5.1 htmlwidgets\_1.3 munsell\_0.5.0 broom\_0.5.0   
## [25] compiler\_3.5.1 modelr\_0.1.2 pkgconfig\_2.0.2 htmltools\_0.3.6   
## [29] tidyselect\_0.2.5 fansi\_0.4.0 crayon\_1.3.4 withr\_2.1.2   
## [33] MASS\_7.3-51 grid\_3.5.1 nlme\_3.1-137 jsonlite\_1.5   
## [37] gtable\_0.2.0 magrittr\_1.5 scales\_1.0.0 cli\_1.0.1   
## [41] stringi\_1.2.4 testthat\_2.0.1 xml2\_1.2.0 tools\_3.5.1   
## [45] glue\_1.3.0 hms\_0.4.2 yaml\_2.2.0 colorspace\_1.3-2  
## [49] rvest\_0.3.2 bindr\_0.1.1 haven\_1.1.2

#### Estimate the mean and variance for the ratio of ffu/copy number in fine aerosols from EMIT ####  
# Use a mixed model with random effect of person and fixed effect of day since onset of symptoms.   
# First have to cut the right df  
  
PNAS\_data\_full <- read.csv("/Users/jbueno/Box Sync/EMIT/EMIT\_Data\_Analysis\_Jake/EMIT\_UMD\_Natural\_Infection/Curated Data/Analytical Datasets/PNAS\_data\_full.csv")  
PNAS\_data\_full$date.visit <- as.Date(PNAS\_data\_full$date.visit)  
  
# Find each sample where we have PCR and focus assay  
  
focus\_subjects <- PNAS\_data\_full %>%  
 filter(!is.na(focus.ct)) %>%  
 distinct(subject.id, date.visit, .keep\_all = TRUE) %>%  
 distinct(subject.id)  
  
focus <- PNAS\_data\_full %>%  
 filter(!is.na(focus.ct))  
  
focus\_pos\_subjects <- focus %>%  
 filter(focus.ct > 0) %>%  
 distinct(subject.id, date.visit, .keep\_all = TRUE) %>%  
 distinct(subject.id)  
  
focus\_pos <- focus %>%  
 filter(focus.ct > 0)  
  
focus\_pos\_fine\_subjects <- focus\_pos %>%  
 filter(sample.type == "GII condensate NO mask") %>%  
 distinct(subject.id, date.visit) %>%  
 distinct(subject.id)  
print(nrow(focus\_pos\_fine\_subjects))

## [1] 53

focus\_pos\_person\_visits <- focus\_pos %>%  
 distinct(subject.id, date.visit)  
print(nrow(focus\_pos\_person\_visits))

## [1] 111

focus\_pos\_fine <- focus\_pos %>%  
 filter(sample.type == "GII condensate NO mask")  
  
# (Could also find the number of fine positive where focus was negative (make sure this uses the set where focus was taken))  
  
# Find the number focus positive where fine was negative (both replicates)  
focus\_pos\_fine\_ever\_pos\_person\_visits <- focus\_pos\_fine %>%  
 filter(!is.na(final.copies)) # %>%  
# distinct(subject.id, date.visit)  
# print(nrow(focus\_pos\_fine\_ever\_pos\_person\_visits))  
  
focus\_pos\_fine\_ever\_pos\_subjects <- focus\_pos\_fine %>%  
 filter(!is.na(final.copies)) %>%  
 distinct(subject.id)  
print(nrow(focus\_pos\_fine\_ever\_pos\_subjects))

## [1] 43

focus\_pos\_fine\_neg <- focus\_pos\_fine %>%  
 anti\_join(focus\_pos\_fine\_ever\_pos\_subjects) %>%  
 distinct(subject.id, date.visit)

## Joining, by = "subject.id"

print(nrow(focus\_pos\_fine\_neg))

## [1] 10

focus\_pos\_fine\_pos <- focus\_pos\_fine %>%  
 anti\_join(focus\_pos\_fine\_neg)

## Joining, by = c("subject.id", "date.visit")

focus\_pos\_fine\_pos\_person\_visits <- focus\_pos\_fine %>%  
 anti\_join(focus\_pos\_fine\_neg) %>%  
 distinct(subject.id, date.visit)

## Joining, by = c("subject.id", "date.visit")

print(nrow(focus\_pos\_fine\_pos\_person\_visits))

## [1] 47

# Find the set of sample instances where focus was positive and fine was 1 positive and 1 negative  
# Use the LOD/sqrt(2) to impute a value for these nondetects.   
# Looks like there are 6 instances where there are NA for final.copies (5 fluA and 1 fluB)  
# focus\_pos\_fine\_pos$final.copies[is.na(focus\_pos\_fine\_pos$final.copies)] <- (500/(2)^(.5))  
  
# Now we are changing the plan and will exclude pcr results if one of the replicates was a nondetect  
one\_or\_both\_replicates\_negative <- focus\_pos\_fine %>%  
 filter(is.na(final.copies)) %>%  
 distinct(subject.id, date.visit)  
  
focus\_pos\_fine\_pos <- focus\_pos\_fine %>%  
 anti\_join(one\_or\_both\_replicates\_negative)

## Joining, by = c("subject.id", "date.visit")

focus\_pos\_fine\_pos\_averaged\_replicates <- focus\_pos\_fine\_pos %>%  
 group\_by(subject.id, date.visit) %>%  
 mutate(avg.final.copies = mean(final.copies)) %>%  
 mutate(avg.focus.ct = mean(focus.ct)) %>%  
 mutate(copy\_number\_ffu\_ratio = avg.final.copies / avg.focus.ct) %>%  
 mutate(log10\_avg.final.copies = log10(avg.final.copies)) %>%  
 mutate(log10\_avg.focus.ct = log10(avg.focus.ct)) %>%  
 mutate(log10\_copy\_number\_log10\_ffu\_ratio = log10\_avg.final.copies / log10\_avg.focus.ct) %>%  
 mutate(log10\_ratio = log10(copy\_number\_ffu\_ratio)) %>%  
 distinct(subject.id, date.visit, .keep\_all = TRUE)  
  
## Estimate the mean and variance for the ratio of ffu/copy number in fine aerosols.  
## Use a mixed model with random effect of person and fixed effect of day since onset of symptoms.   
  
# Random intercept (to take by-subject variability into account) with fixed mean model   
# (assumes the effect of dpo to be the same across all subjects)  
model\_1 <- lmer(copy\_number\_ffu\_ratio ~ dpo + (1|subject.id), data = focus\_pos\_fine\_pos\_averaged\_replicates)  
summary(model\_1)

## Linear mixed model fit by REML ['lmerMod']  
## Formula: copy\_number\_ffu\_ratio ~ dpo + (1 | subject.id)  
## Data: focus\_pos\_fine\_pos\_averaged\_replicates  
##   
## REML criterion at convergence: 1118.6  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -0.7734 -0.2149 -0.1154 -0.0241 3.2578   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## subject.id (Intercept) 1.091e+11 330311   
## Residual 4.232e+10 205717   
## Number of obs: 41, groups: subject.id, 37  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) -56241 164426 -0.342  
## dpo 75082 75852 0.990  
##   
## Correlation of Fixed Effects:  
## (Intr)  
## dpo -0.923

coef(model\_1)

## $subject.id  
## (Intercept) dpo  
## 32 269880.511 75081.92  
## 62 -71689.528 75081.92  
## 63 -85837.429 75081.92  
## 79 -8541.462 75081.92  
## 84 1537666.487 75081.92  
## 88 -64553.343 75081.92  
## 92 -123790.836 75081.92  
## 94 -123686.696 75081.92  
## 97 -65647.008 75081.92  
## 105 -66452.816 75081.92  
## 110 -174704.793 75081.92  
## 112 -177932.525 75081.92  
## 113 -69644.023 75081.92  
## 117 -109811.879 75081.92  
## 120 -114828.874 75081.92  
## 123 -117225.134 75081.92  
## 138 -22845.603 75081.92  
## 140 -123138.498 75081.92  
## 142 -177932.016 75081.92  
## 145 -69761.485 75081.92  
## 150 -121979.780 75081.92  
## 194 -123654.562 75081.92  
## 228 -104562.624 75081.92  
## 235 -101200.148 75081.92  
## 238 -175696.265 75081.92  
## 244 -177890.426 75081.92  
## 250 -122908.278 75081.92  
## 252 -170223.428 75081.92  
## 253 -123753.742 75081.92  
## 255 -69000.961 75081.92  
## 265 -130764.963 75081.92  
## 276 -165474.813 75081.92  
## 288 -110480.584 75081.92  
## 298 -117465.800 75081.92  
## 313 -63625.155 75081.92  
## 314 -173121.071 75081.92  
## 326 -68629.279 75081.92  
##   
## attr(,"class")  
## [1] "coef.mer"

summary(focus\_pos\_fine\_pos\_averaged\_replicates$copy\_number\_ffu\_ratio)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 76.1 696.8 7304.3 133011.0 26858.3 2433098.0

# This is quite skewed  
  
model\_1\_log\_scale <- lmer(log10\_copy\_number\_log10\_ffu\_ratio ~ dpo + (1|subject.id), data = focus\_pos\_fine\_pos\_averaged\_replicates)  
summary(model\_1\_log\_scale)

## Linear mixed model fit by REML ['lmerMod']  
## Formula: log10\_copy\_number\_log10\_ffu\_ratio ~ dpo + (1 | subject.id)  
## Data: focus\_pos\_fine\_pos\_averaged\_replicates  
##   
## REML criterion at convergence: 171.2  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -1.3772 -0.4881 -0.2218 0.5736 2.0550   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## subject.id (Intercept) 1.889 1.374   
## Residual 2.176 1.475   
## Number of obs: 41, groups: subject.id, 37  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 4.09544 0.89974 4.552  
## dpo -0.07711 0.41948 -0.184  
##   
## Correlation of Fixed Effects:  
## (Intr)  
## dpo -0.932

summary(focus\_pos\_fine\_pos\_averaged\_replicates$log10\_copy\_number\_log10\_ffu\_ratio)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.923 2.596 3.315 4.001 5.445 9.604

# This is less skewed but the team decided to go with a different plan - see below.  
  
# But if we want the fixed effect of dpo to not be continuous, then we will convert this dpo variable to character  
# focus\_pos\_fine\_pos\_averaged\_replicates$dpo <- ordered(focus\_pos\_fine\_pos\_averaged\_replicates$dpo)  
# Could use dpo as ordinal - as done above, but decided to go with using dpo as a character variable for now.   
  
focus\_pos\_fine\_pos\_averaged\_replicates$dpo <- as.character(focus\_pos\_fine\_pos\_averaged\_replicates$dpo)  
  
# Also, in fitting this second model, we want to use the log10(ratio) because the model was somewhat skewed before  
model\_2 <- lmer(log10\_ratio ~ dpo + (1|subject.id), data = focus\_pos\_fine\_pos\_averaged\_replicates)  
summary(model\_2)

## Linear mixed model fit by REML ['lmerMod']  
## Formula: log10\_ratio ~ dpo + (1 | subject.id)  
## Data: focus\_pos\_fine\_pos\_averaged\_replicates  
##   
## REML criterion at convergence: 126.4  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -1.28110 -0.83862 0.09108 0.46243 1.59578   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## subject.id (Intercept) 0.5102 0.7143   
## Residual 0.8365 0.9146   
## Number of obs: 41, groups: subject.id, 37  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 3.76758 0.34530 10.911  
## dpo2 0.05015 0.42121 0.119  
## dpo3 -0.30167 0.48746 -0.619  
##   
## Correlation of Fixed Effects:  
## (Intr) dpo2   
## dpo2 -0.783   
## dpo3 -0.706 0.579

summary(focus\_pos\_fine\_pos\_averaged\_replicates$log10\_ratio)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.881 2.843 3.864 3.764 4.429 6.386

table(focus\_pos\_fine\_pos\_averaged\_replicates$dpo)

##   
## 1 2 3   
## 11 19 11

# writing this out so that Charles Ma can work on it related to the power calculation for new RO1 grants due Februrary 5, 2019  
# write.csv(focus\_pos\_fine\_pos\_averaged\_replicates, "/Users/jbueno/Desktop/focus\_pos\_fine\_pos.csv")