UMD\_Natural\_Infection\_Analysis.R

jbueno

Thu Feb 7 19:50:32 2019

# EMIT UMD Natural Infection Analysis  
# 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 10.14.2  
##   
## 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] htmlTable\_1.12 rmarkdown\_1.10 markdown\_0.8   
## [4] bindrcpp\_0.2.2 lme4\_1.1-19 Matrix\_1.2-14   
## [7] arsenal\_1.5.0 lubridate\_1.7.4 data.table\_1.11.8  
## [10] knitr\_1.20 readxl\_1.1.0 RcppRoll\_0.3.0   
## [13] forcats\_0.3.0 stringr\_1.3.1 dplyr\_0.7.7   
## [16] purrr\_0.2.5 readr\_1.1.1 tidyr\_0.8.2   
## [19] tibble\_1.4.2 ggplot2\_3.1.0 tidyverse\_1.2.1   
##   
## loaded via a namespace (and not attached):  
## [1] tidyselect\_0.2.5 splines\_3.5.1 haven\_1.1.2 lattice\_0.20-38   
## [5] colorspace\_1.3-2 testthat\_2.0.1 htmltools\_0.3.6 yaml\_2.2.0   
## [9] utf8\_1.1.4 rlang\_0.3.0.1 pillar\_1.3.0 nloptr\_1.2.1   
## [13] glue\_1.3.0 withr\_2.1.2 modelr\_0.1.2 bindr\_0.1.1   
## [17] plyr\_1.8.4 munsell\_0.5.0 gtable\_0.2.0 cellranger\_1.1.0  
## [21] rvest\_0.3.2 htmlwidgets\_1.3 evaluate\_0.12 fansi\_0.4.0   
## [25] highr\_0.7 broom\_0.5.0 Rcpp\_0.12.19 checkmate\_1.8.5   
## [29] scales\_1.0.0 backports\_1.1.2 jsonlite\_1.5 digest\_0.6.18   
## [33] hms\_0.4.2 stringi\_1.2.4 rprojroot\_1.3-2 grid\_3.5.1   
## [37] cli\_1.0.1 tools\_3.5.1 magrittr\_1.5 lazyeval\_0.2.1   
## [41] crayon\_1.3.4 pkgconfig\_2.0.2 MASS\_7.3-51 xml2\_1.2.0   
## [45] assertthat\_0.2.0 minqa\_1.2.4 httr\_1.3.1 rstudioapi\_0.8   
## [49] R6\_2.3.0 nlme\_3.1-137 compiler\_3.5.1

#### 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") %>%  
 select(-X)  
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: 1207.8  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -0.7592 -0.2145 -0.1173 -0.0219 3.2586   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## subject.id (Intercept) 1.074e+12 1036168   
## Residual 4.153e+11 644451   
## Number of obs: 41, groups: subject.id, 37  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) -174142 515490 -0.338  
## dpo 225655 237792 0.949  
##   
## Correlation of Fixed Effects:  
## (Intr)  
## dpo -0.923

coef(model\_1)

## $subject.id  
## (Intercept) dpo  
## 32 851068.30 225654.8  
## 62 -210672.05 225654.8  
## 63 -256739.35 225654.8  
## 79 -15083.42 225654.8  
## 84 4826459.01 225654.8  
## 88 -194827.10 225654.8  
## 92 -373611.65 225654.8  
## 94 -373285.97 225654.8  
## 97 -186095.14 225654.8  
## 105 -200767.44 225654.8  
## 110 -534520.66 225654.8  
## 112 -536459.79 225654.8  
## 113 -210747.49 225654.8  
## 117 -364658.69 225654.8  
## 120 -367981.45 225654.8  
## 123 -369568.50 225654.8  
## 138 -64391.94 225654.8  
## 140 -371571.56 225654.8  
## 142 -536658.05 225654.8  
## 145 -211114.84 225654.8  
## 150 -367947.83 225654.8  
## 194 -373185.48 225654.8  
## 228 -361182.11 225654.8  
## 235 -358955.14 225654.8  
## 238 -535177.31 225654.8  
## 244 -536630.50 225654.8  
## 250 -373332.45 225654.8  
## 252 -531552.65 225654.8  
## 253 -373495.65 225654.8  
## 255 -210746.25 225654.8  
## 265 -505419.29 225654.8  
## 276 -528407.64 225654.8  
## 288 -365101.58 225654.8  
## 298 -369727.89 225654.8  
## 313 -207185.85 225654.8  
## 314 -533471.76 225654.8  
## 326 -210500.08 225654.8  
##   
## attr(,"class")  
## [1] "coef.mer"

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

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 74 1090 6144 400707 22826 7603431

# 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: 172.5  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -1.4082 -0.4450 -0.2129 0.5450 1.8345   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## subject.id (Intercept) 2.013 1.419   
## Residual 2.201 1.484   
## Number of obs: 41, groups: subject.id, 37  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 4.4682 0.9148 4.885  
## dpo -0.2091 0.4263 -0.490  
##   
## 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.   
## 1.843 2.722 3.335 4.128 5.227 9.352

# 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: 127.4  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -1.02710 -0.62416 -0.04606 0.26238 1.70615   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## subject.id (Intercept) 0.6729 0.8203   
## Residual 0.7237 0.8507   
## Number of obs: 41, groups: subject.id, 37  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 4.06219 0.34809 11.670  
## dpo2 -0.06514 0.41960 -0.155  
## dpo3 -0.63834 0.49061 -1.301  
##   
## Correlation of Fixed Effects:  
## (Intr) dpo2   
## dpo2 -0.778   
## dpo3 -0.705 0.585

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

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.872 3.038 3.788 3.946 4.358 6.881

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")