R Notebook

Luiza Moore

26062019

Modelling the effect of menstrual phase on driver mutation burden

Markdown file to document methods used in the analysis of the driver mutation burden in normal endometrium.

Load Libraries

```
library(tidyverse)
library(magrittr)
library(lme4)
library(lmerTest)
library(rlang)
library(knitr)
library(kableExtra)
library(pbkrtest)
```

Load in data

Load in sample level data for all 28 donors, but exclude post-menopausal women and women with undetermined menstrual phase.

```
  endom_burden <- read.csv("Endometrium_for_model_26062019.csv", stringsAsFactors = F, na.s
trings = c("", "NA", "Unknown", "Uncertain"))
  dim(endom_burden)
```

```
## [1] 257  25
```

```
# Make BMI and Parity numeric
  endom_burden %<>%  mutate(BMI.QC = as.numeric(BMI))
  endom_burden %<>%  mutate(Parity.QC = as.numeric(Parity))

# Exclude post-menopausal women
  endom_burden.qc <- endom_burden %>% filter(Menopause_status_num == 0)
  dim(endom_burden.qc)
```

```
## [1] 218  27
```

```
# Exclude cases with undetermined menstrual phase
  endom_burden.qc <- endom_burden.qc %>% filter(Menstrual_phase_num >0)
  dim(endom_burden.qc)
```

```
## [1] 208  27
```

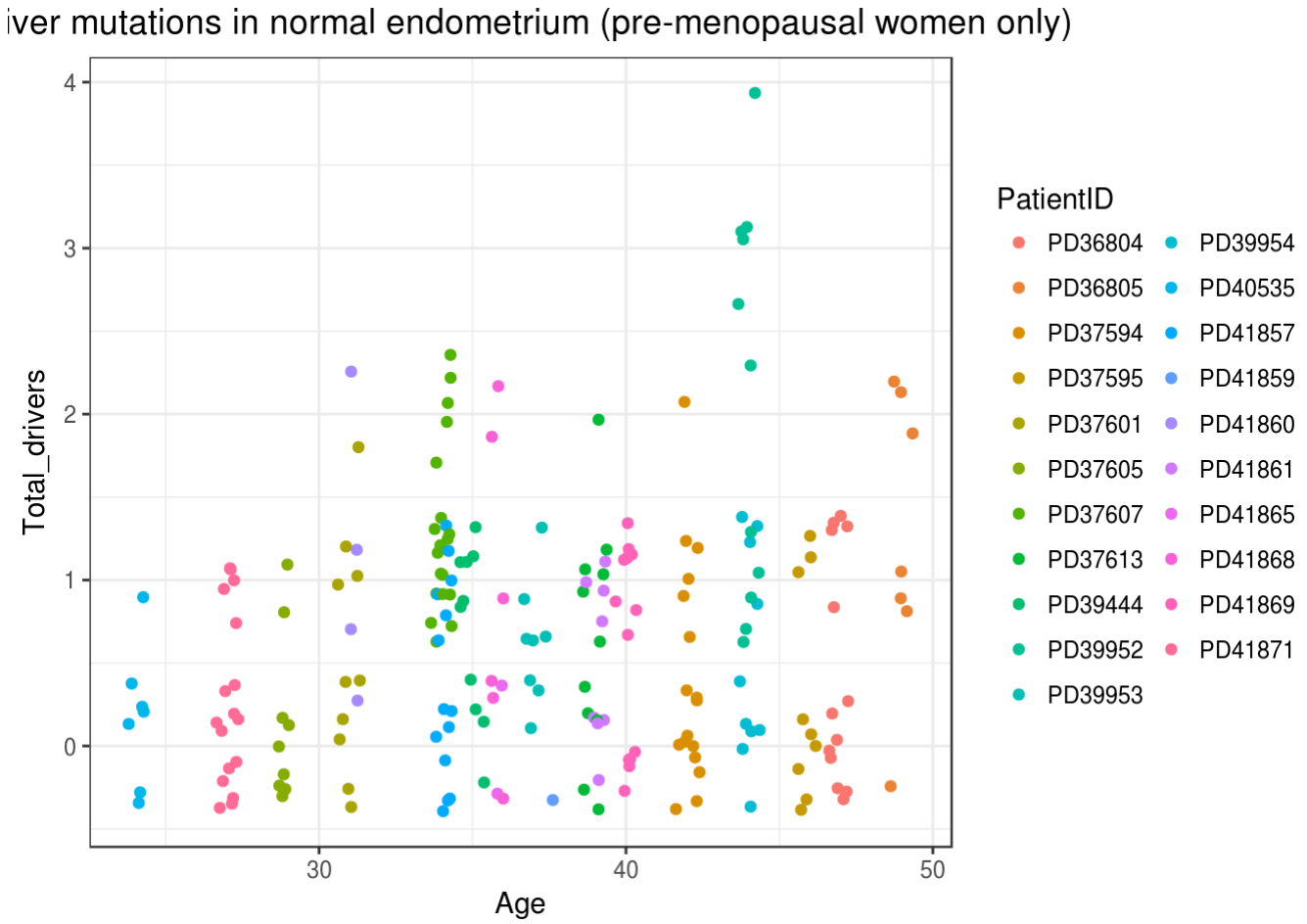
```
# Remove samples with no Parity information
  endom_burden.qc %<>% filter(!is.na(BMI.QC), !is.na(Parity.QC))
  dim(endom_burden.qc)
```

```
## [1] 206  27
```

```
 # Samples per patient
endom_burden.qc %>% group_by(PatientID) %>%  count(PatientID) %>%  rename(`Sample count` =
 n) %>% arrange(desc(`Sample count`)) %>%  kable() %>%  kable_styling(bootstrap_options = c
("striped", "condensed"), full_width = F, position = "left")
```

PatientID	Sample count
PD37607	19
PD37594	17
PD41871	17
PD41857	14
PD36804	13
PD41869	13
PD37613	11
PD39952	11
PD37601	10
PD39444	10
PD39954	10
PD37595	9
PD37605	9
PD39953	8
PD41861	8
PD36805	7
PD40535	7
PD41868	6
PD41860	4
PD41865	2
PD41859	1

```
# Look at the raw data
  endom_burden.qc %>% ggplot(aes(Age, Total_drivers, colour = PatientID)) +
  geom_jitter() +
  theme(plot.title = element_text(size = 8)) +
  ggtitle("Driver mutations in normal endometrium (pre-menopausal women only)") +
  theme(plot.title = element_text(size = 14)) + theme_bw() +theme(plot.title = element_text
(hjust = 0.5))
```



Does menstrual phase have an effect on the driver mutation burden?

To test the effect of menstrual phase on the driver mutation burden we add Menstrual phase to the final generalized linear mixed-effects model with Poisson distribution with features that have been shown to be significant in the full cohort of patients.

The significant features are:

- Age
- Read depth & VAF ('Vafdepth')
- Parity

We use backwards elimination to define the final model

Define full model and drop each fixed effect in turn

```
# Combine read depth and median sample depth (Seq_X) as 'Vafdepth'
  endom_burden.qc %<>%  mutate(Vafdepth = Seq_X*SampleMedianVAF)

# Make BMI and Parity numeric
  endom_burden.qc %<>%  mutate(BMI.QC = as.numeric(BMI))
  endom_burden.qc %<>%  mutate(Parity.QC = as.numeric(Parity))

# Define the full model containing all features
  full_glmer_model = glmer(Total_drivers ~ Age + Parity.QC + Menstrual_phase_num +(Age - 1|
PatientID), data=endom_burden.qc, family = poisson(link = "log"), control =   glmerControl
(optimizer="bobyqa", optCtrl = list(maxfun = 100000)))

  print(summary(full_glmer_model))
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: Total_drivers ~ Age + Parity.QC + Menstrual_phase_num + (Age -
##     1 | PatientID)
##    Data: endom_burden.qc
## Control:
## glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e+05))
##
##      AIC      BIC   logLik deviance df.resid
##    403.8    420.4   -196.9    393.8      201
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.0933 -0.6763 -0.5314  0.6787  2.0963
##
## Random effects:
##  Groups    Name Variance  Std.Dev.
##  PatientID Age  7.543e-05 0.008685
## Number of obs: 206, groups:  PatientID, 21
##
## Fixed effects:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.97608    0.95243  -1.025   0.3054
## Age                  0.04002    0.01914   2.091   0.0366 *
## Parity.QC           -0.24689    0.10749  -2.297   0.0216 *
## Menstrual_phase_num -0.46049    0.32828  -1.403   0.1607
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Age    Prt.QC
## Age         -0.751
## Parity.QC   -0.031 -0.101
## Mnstrl_phs_ -0.649  0.011  0.024
## convergence code: 0
## Model failed to converge with max|grad| = 0.0018568 (tol = 0.001, component 1)
```

```
# "user" parametric boot function as defined in drop1.merMod help example
  PBSumFun <- function(object, objectDrop, ...) {
    pbnames <- c("stat", "p.value")
    r <- if (missing(objectDrop)) {
      setNames(rep(NA, length(pbnames)), pbnames)
    } else {
      pbtest <- PBmodcomp(object, objectDrop, nsim = nsim, ref = NULL, seed=12345, details
 = 0)
      unlist(pbtest$test[2, pbnames])
    }
    attr(r, "method") <- c("Parametric bootstrap via pbkrtest package")
    r
  }
# Drop each fixed effect from model and test significance
# Use 1000 samples to form the reference distribution
nsim <- 1000
drop1(full_glmer_model, test = "user", sumFun = PBSumFun)
```

```
## Single term deletions
##
## Model:
## Total_drivers ~ Age + Parity.QC + Menstrual_phase_num + (Age -
##     1 | PatientID)
## Method:
## Parametric bootstrap via pbkrtest package
##
##
##                       stat  p.value
## <none>
## Age                 4.2999 0.056701
## Parity.QC           5.1460 0.048857
## Menstrual_phase_num 1.7141 0.260549
```

Remove feature with the largest P > 0.05 to make reduced model

```
# Remove Menstrual phase from the full model
  reduced_glmer_model <- update(full_glmer_model, ~ . -Menstrual_phase_num, control=glmerCo
ntrol(optimizer="bobyqa", optCtrl = list(maxfun = 100000)))
# Drop each fixed effect from the model and test significance
  drop1(reduced_glmer_model, test = "user", sumFun = PBSumFun)
```

```
## Single term deletions
##
## Model:
## Total_drivers ~ Age + Parity.QC + (Age - 1 | PatientID)
## Method:
## Parametric bootstrap via pbkrtest package
##
##
##             stat  p.value
## <none>
## Age       3.8150 0.067708
## Parity.QC 4.3927 0.063017
```

Define the final model

```
# Define the final model keeping only the significant features  (P < 0.05)
  final_glmer_model <- reduced_glmer_model

# Print summary for the final model
  print(summary(final_glmer_model))
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: Total_drivers ~ Age + Parity.QC + (Age - 1 | PatientID)
##    Data: endom_burden.qc
## Control:
## glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 1e+05))
##
##      AIC      BIC   logLik deviance df.resid
##    403.5    416.8   -197.8    395.5      202
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.1113 -0.6955 -0.4384  0.6488  2.1040
##
## Random effects:
##  Groups    Name Variance  Std.Dev.
##  PatientID Age  9.975e-05 0.009987
## Number of obs: 206, groups:  PatientID, 21
##
## Fixed effects:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.87925    0.77163  -2.435   0.0149 *
## Age          0.04092    0.02062   1.985   0.0471 *
## Parity.QC   -0.24412    0.11431  -2.136   0.0327 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##           (Intr) Age
## Age       -0.978
## Parity.QC  0.024 -0.145
```

```
# Estimate confidence intervals using "likelihood profile" method
  confint.merMod(final_glmer_model, method = "profile")
```

```
## Computing profile confidence intervals ...
```

```
##                    2.5 %      97.5 %
## .sig01       0.0049871530  0.01716773
## (Intercept) -3.5142119925 -0.37232360
## Age         -0.0001535045  0.08423708
## Parity.QC   -0.4821811109 -0.01665213
```