R Notebook

Luiza Moore

26062019

Modelling the effect of menstrual phase on total mutation burden and clonality

Markdown file to document methods used in the analysis of the menstrual phase and its effect on the total mutation burden and clonality

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
```

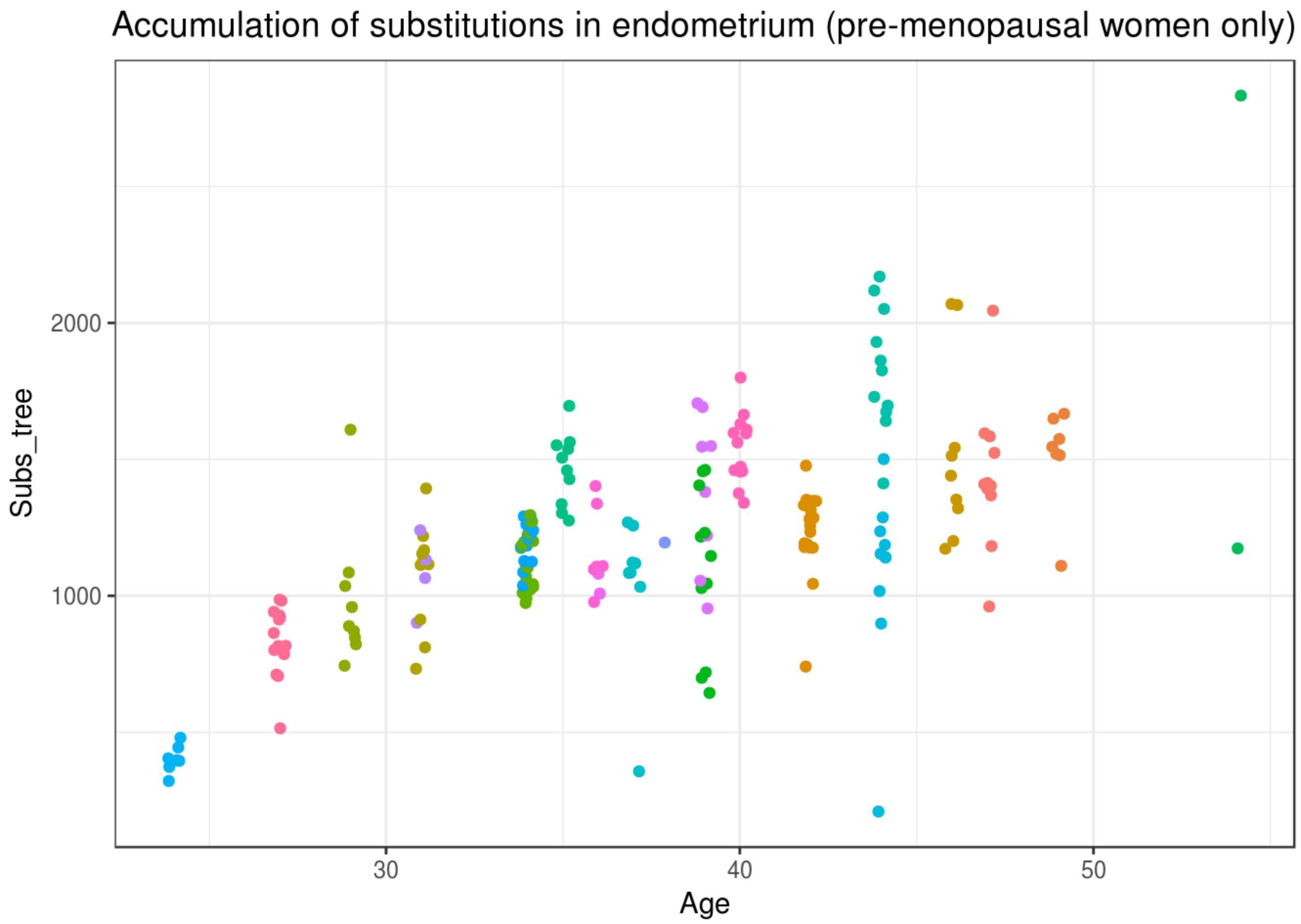
```
# Exclude cases with undetermined menstrual phase
  endom_burden.qc <- endom_burden.qc %>% filter(Menstrual_phase_num >0)
  dim(endom_burden.qc)
```

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

```
 # 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
PD38812	2
PD41865	2
PD41859	1

```
# Plot data
endom_burden.qc %>% ggplot(aes(Age, Subs_tree, colour = PatientID)) +
  geom_jitter(width = 0.2) +
  theme(plot.title = element_text(size = 3)) +
  ggtitle("Accumulation of substitutions in endometrium (pre-menopausal women only)") +
  theme(plot.title = element_text(size = 3)) + theme_bw() +theme(plot.title = element_text
(hjust = 0.5)) +
  theme(legend.position="none")
```



Accumulation of substitutions in endometrium (pre-menopausal women only)

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

To test the effect of menstrual phase on the total mutation burden we apply the final mixed-effect model with features that have been shown to be significant in the full cohort of patients.

These significant features are:

- Age
- Read depth & VAF (‘Vafdepth’)
- Driver mutations

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

# Total mutation burden
  full_lmer_model1 = lmer(Subs_tree ~ Age + Vafdepth + Driver_status + Menstrual_phase_num
 + (Age - 1|PatientID),  data=endom_burden.qc, REML=F)
  summary(full_lmer_model1)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use
##   Satterthwaite's method [lmerModLmerTest]
## Formula:
## Subs_tree ~ Age + Vafdepth + Driver_status + Menstrual_phase_num +
##     (Age - 1 | PatientID)
##    Data: endom_burden.qc
##
##      AIC      BIC   logLik deviance df.resid
##   2854.9   2878.2  -1420.4   2840.9      201
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.6221 -0.4167  0.0248  0.4660  3.8938
##
## Random effects:
##  Groups    Name Variance Std.Dev.
##  PatientID Age     13.19   3.632
##  Residual       42550.40 206.277
## Number of obs: 208, groups:  PatientID, 22
##
## Fixed effects:
##                      Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)          -358.353    221.519   28.399  -1.618    0.117
## Age                    37.250      4.669   21.003   7.977 8.61e-08 ***
## Vafdepth               22.215      5.423  207.973   4.097 6.01e-05 ***
## Driver_status         131.593     33.017  200.942   3.986 9.41e-05 ***
## Menstrual_phase_num  -74.194     84.986   22.755  -0.873    0.392
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Age    Vfdpth Drvr_s
## Age         -0.667
## Vafdepth    -0.241  0.027
## Driver_stts  0.041 -0.121 -0.192
## Mnstrl_phs_ -0.591 -0.120 -0.067  0.046
```

```
  reduced_lmer_model1 = lmer(Subs_tree ~ Age + Vafdepth + Driver_status  + (Age - 1|Patient
ID),  data=endom_burden.qc, REML=F)
  summary(reduced_lmer_model1)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use
##   Satterthwaite's method [lmerModLmerTest]
## Formula:
## Subs_tree ~ Age + Vafdepth + Driver_status + (Age - 1 | PatientID)
##    Data: endom_burden.qc
##
##      AIC      BIC   logLik deviance df.resid
##   2853.6   2873.6  -1420.8   2841.6      202
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5372 -0.4404  0.0263  0.4820  4.0069
##
## Random effects:
##  Groups    Name Variance Std.Dev.
##  PatientID Age     14.5    3.807
##  Residual       42357.8  205.810
## Number of obs: 208, groups:  PatientID, 22
##
## Fixed effects:
##                Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)   -474.721    184.103   30.774  -2.579   0.0149 *
## Age             36.876      4.798   23.455   7.685 7.43e-08 ***
## Vafdepth        21.747      5.419  207.876   4.013 8.36e-05 ***
## Driver_status  132.336     32.969  201.308   4.014 8.42e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Age    Vfdpth
## Age         -0.925
## Vafdepth    -0.338  0.018
## Driver_stts  0.083 -0.113 -0.190
```

```
anova(full_lmer_model1,reduced_lmer_model1)
```

```
## Data: endom_burden.qc
## Models:
## reduced_lmer_model1: Subs_tree ~ Age + Vafdepth + Driver_status + (Age - 1 | PatientID)
## full_lmer_model1: Subs_tree ~ Age + Vafdepth + Driver_status + Menstrual_phase_num +
## full_lmer_model1:     (Age - 1 | PatientID)
##                     Df    AIC    BIC  logLik deviance  Chisq Chi Df
## reduced_lmer_model1  6 2853.6 2873.6 -1420.8   2841.6
## full_lmer_model1     7 2854.9 2878.2 -1420.4   2840.9 0.7026      1
##                     Pr(>Chisq)
## reduced_lmer_model1
## full_lmer_model1        0.4019
```

Does menstrual phase have an effect on clonality?

To test the effect of menstrual phase on clonality, we used a linear mixed-effect model with SampleMedianVAF as a proxy for clonality

```
  full_lmer_model2 = lmer(SampleMedianVAF ~ Age + Vafdepth + Driver_status + Menstrual_phas
e_num + (Age - 1|PatientID),  data=endom_burden.qc, REML=F)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 0.0184371 (tol =
## 0.002, component 1)
```

```
  summary(full_lmer_model2)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use
##   Satterthwaite's method [lmerModLmerTest]
## Formula:
## SampleMedianVAF ~ Age + Vafdepth + Driver_status + Menstrual_phase_num +
##     (Age - 1 | PatientID)
##    Data: endom_burden.qc
##
##      AIC      BIC   logLik deviance df.resid
##   -584.8   -561.5    299.4   -598.8      201
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.97712 -0.48971  0.05725  0.56190  2.74962
##
## Random effects:
##  Groups    Name Variance  Std.Dev.
##  PatientID Age  2.486e-07 0.0004986
##  Residual       3.055e-03 0.0552702
## Number of obs: 208, groups:  PatientID, 22
##
## Fixed effects:
##                       Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)          2.236e-01  4.256e-02  3.449e+01   5.253 7.75e-06 ***
## Age                  5.753e-04  8.292e-04  1.954e+01   0.694    0.496
## Vafdepth             1.390e-02  1.365e-03  1.827e+02  10.185  < 2e-16 ***
## Driver_status       -4.209e-03  8.558e-03  2.072e+02  -0.492    0.623
## Menstrual_phase_num  2.068e-04  1.540e-02  2.217e+01   0.013    0.989
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Age    Vfdpth Drvr_s
## Age         -0.659
## Vafdepth    -0.348  0.058
## Driver_stts  0.037 -0.162 -0.198
## Mnstrl_phs_ -0.586 -0.083 -0.071  0.077
## convergence code: 0
## Model failed to converge with max|grad| = 0.0184371 (tol = 0.002, component 1)
```

```
  reduced_lmer_model2 = lmer(SampleMedianVAF ~ Age + Vafdepth + Driver_status + (Age - 1|Pa
tientID),  data=endom_burden.qc, REML=F)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control
## $checkConv, : Model failed to converge with max|grad| = 0.0180755 (tol =
## 0.002, component 1)
```

```
summary(reduced_lmer_model2)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use
##   Satterthwaite's method [lmerModLmerTest]
## Formula: SampleMedianVAF ~ Age + Vafdepth + Driver_status + (Age - 1 |
##     PatientID)
##    Data: endom_burden.qc
## 
##      AIC      BIC   logLik deviance df.resid 
##   -586.8   -566.8    299.4   -598.8      202 
## 
## Scaled residuals: 
##      Min       1Q   Median       3Q      Max 
## -2.97672 -0.49076  0.05746  0.56106  2.74980 
## 
## Random effects:
##  Groups    Name Variance  Std.Dev. 
##  PatientID Age  2.486e-07 0.0004986
##  Residual       3.055e-03 0.0552703
## Number of obs: 208, groups:  PatientID, 22
## 
## Fixed effects:
##                 Estimate Std. Error         df t value Pr(>|t|)    
## (Intercept)    2.239e-01  3.448e-02  3.567e+01   6.495 1.59e-07 ***
## Age            5.762e-04  8.264e-04  1.987e+01   0.697    0.494    
## Vafdepth       1.390e-02  1.361e-03  1.836e+02  10.212  < 2e-16 ***
## Driver_status -4.218e-03  8.532e-03  2.063e+02  -0.494    0.622    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Correlation of Fixed Effects:
##             (Intr) Age    Vfdpth
## Age         -0.876              
## Vafdepth    -0.483  0.053       
## Driver_stts  0.102 -0.157 -0.194
## convergence code: 0
## Model failed to converge with max|grad| = 0.0180755 (tol = 0.002, component 1)
```

```
anova(full_lmer_model2,reduced_lmer_model2)
```

```
## Data: endom_burden.qc
## Models:
## reduced_lmer_model2: SampleMedianVAF ~ Age + Vafdepth + Driver_status + (Age - 1 |
## reduced_lmer_model2:     PatientID)
## full_lmer_model2: SampleMedianVAF ~ Age + Vafdepth + Driver_status + Menstrual_phase_num
+
## full_lmer_model2:     (Age - 1 | PatientID)
##                     Df     AIC     BIC logLik deviance Chisq Chi Df
## reduced_lmer_model2  6 -586.84 -566.82 299.42  -598.84
## full_lmer_model2     7 -584.84 -561.48 299.42  -598.84 2e-04      1
##                     Pr(>Chisq)
## reduced_lmer_model2
## full_lmer_model2        0.9893
```