R Notebook

Luiza Moore

26062019

Modelling total mutation burden in normal endometrium

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

Load Libraries

```
library(tidyverse)
library(magrittr)
library(lme4)
library(lmerTest)
library(rlang)
library(knitr)
library(kableExtra)
library(sjPlot)
library(sjmisc)
```

Load in data

Load in sample level data for 28 donors with associated meta-data on age, body mass index (BMI) and parity.

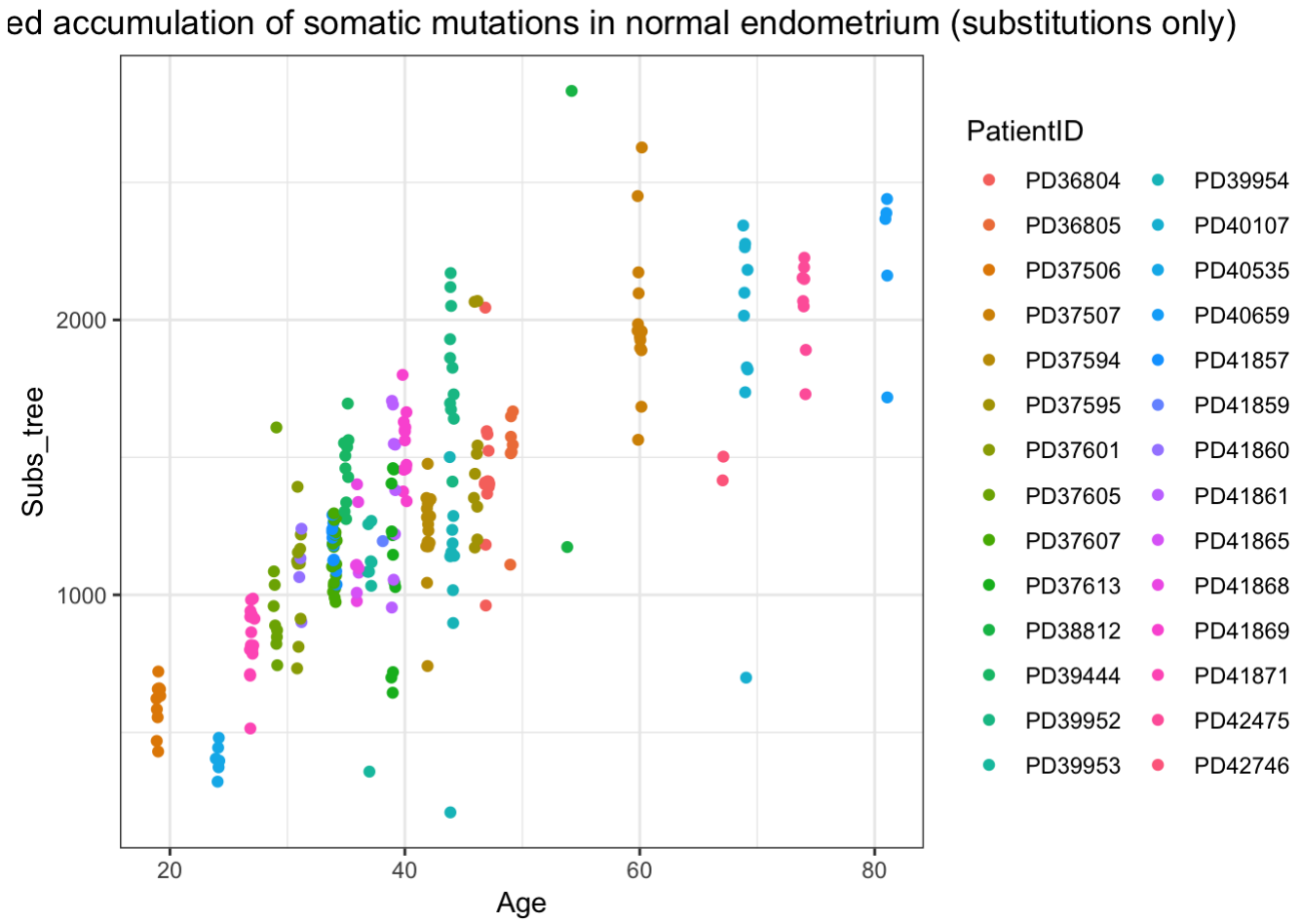
```
endom_burden <- read.csv("~/Desktop/Endometrium_for_model_26062019.csv")

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

PatientID	Sample count
PD37607	19
PD37594	17
PD41871	17
PD37507	14
PD41857	14
PD36804	13
PD41869	13
PD37613	11
PD39952	11

PatientID	Sample count
PD37506	10
PD37601	10
PD39444	10
PD39954	10
PD40107	10
PD37595	9
PD37605	9
PD39953	8
PD41861	8
PD42475	8
PD36805	7
PD40535	7
PD41868	6
PD40659	5
PD41860	4
PD38812	2
PD41865	2
PD42746	2
PD41859	1

```
# Look at raw data
endom_burden %>% ggplot(aes(Age, Subs_tree, colour = PatientID)) +
  geom_jitter(width = 0.2) +
  theme(plot.title = element_text(size = 8)) +
  ggtitle("Age-associated accumulation of somatic mutations in normal endometrium (substitu
tions only)") +
  theme(plot.title = element_text(size = 14)) + theme_bw() +theme(plot.title = element_text
(hjust = 0.5))
```



Fit linear mixed effects models and estimate mutation rate per year

To account for the non-independent sampling per patient we use a linear mixed-effects model as the observed frequencies of all substitutions approximates a normal distribution. We also use a random slope with fixed intercept as most women will start menarche at a similar age (~13 years), but to account for the potential differences in the rates at which mutations were acquired in different individuals due to variation in parity, contraception and other factors.

We test features with a known affect on mutation burden or endometrial cancer risks:

- Age
- Read depth & VAF ('Vafdepth')
- Driver mutations
- BMI
- Parity
- Cohort

We use backwards elimination to define the final model

Make the full model and drop each fixed effect in turn

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

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

# Exclude cases without Parity data
  endom_burden.qc <- endom_burden %>% filter(!is.na(Parity.QC))

# Build the full model

  full_lmer_model = lmer(Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + Parity.QC +
 Cohort + (Age - 1|PatientID),  data=endom_burden, REML=F)

  print(full_lmer_model)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerModLmerTest']
## Formula:
## Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + Parity.QC +
##     Cohort + (Age - 1 | PatientID)
##    Data: endom_burden
##       AIC       BIC    logLik  deviance  df.resid
##  3566.797  3605.836 -1772.398  3544.797       246
## Random effects:
##  Groups    Name Std.Dev.
##  PatientID Age    3.651
##  Residual       219.661
## Number of obs: 257, groups:  PatientID, 28
## Fixed Effects:
##            (Intercept)                     Age                Vafdepth
##               -280.880                  29.666                  27.855
##          Driver_status                  BMI.QC               Parity.QC
##                110.348                   7.572                 -16.138
##      CohortPost-mortem              CohortTAH  CohortTransplant donor
##                 30.250                 -56.199                 -97.972
```

```
# Drop each fixed effect
  lme4:::drop1.merMod(full_lmer_model, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + Parity.QC +
##     Cohort + (Age - 1 | PatientID)
##               Df    AIC    LRT   Pr(Chi)
## <none>           3566.8
## Age            1 3611.0 46.170 1.084e-11 ***
## Vafdepth       1 3590.9 26.116 3.215e-07 ***
## Driver_status  1 3575.2 10.362  0.001286 **
## BMI.QC         1 3565.2  0.436  0.509086
## Parity.QC      1 3565.1  0.299  0.584717
## Cohort         3 3562.8  1.979  0.576675
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Remove feature with largest P > 0.05 to make reduced model 1

```
# Remove Parity from full model
  reduced1_glmer_model <- update(full_lmer_model, ~ . -Parity.QC )
  anova(full_lmer_model,reduced1_glmer_model)
```

```
## Data: endom_burden
## Models:
## reduced1_glmer_model: Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + Cohort +
## reduced1_glmer_model:     (Age - 1 | PatientID)
## full_lmer_model: Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + Parity.QC +
## full_lmer_model:     Cohort + (Age - 1 | PatientID)
##                       Df    AIC    BIC  logLik deviance  Chisq Chi Df
## reduced1_glmer_model 10 3565.1 3600.6 -1772.5   3545.1
## full_lmer_model      11 3566.8 3605.8 -1772.4   3544.8 0.2987      1
##                       Pr(>Chisq)
## reduced1_glmer_model
## full_lmer_model          0.5847
```

```
  print(reduced1_glmer_model)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerModLmerTest']
## Formula: Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + Cohort +
##     (Age - 1 | PatientID)
##    Data: endom_burden
##       AIC       BIC    logLik  deviance  df.resid
##  3565.095  3600.586 -1772.548  3545.095       247
## Random effects:
##  Groups    Name Std.Dev.
##  PatientID Age    3.654
##  Residual       219.783
## Number of obs: 257, groups:  PatientID, 28
## Fixed Effects:
##            (Intercept)                     Age                Vafdepth
##               -327.209                  29.847                  28.011
##          Driver_status                  BMI.QC       CohortPost-mortem
##                111.647                   9.277                 -64.864
##              CohortTAH  CohortTransplant donor
##                -77.080                -115.590
```

```
  lme4:::drop1.merMod(reduced1_glmer_model, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + Cohort +
##     (Age - 1 | PatientID)
##               Df    AIC    LRT   Pr(Chi)
## <none>           3565.1
## Age            1 3610.2 47.111 6.707e-12 ***
## Vafdepth       1 3589.5 26.442 2.716e-07 ***
## Driver_status  1 3573.7 10.629  0.001113 **
## BMI.QC         1 3563.8  0.705  0.401140
## Cohort         3 3561.5  2.387  0.496036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Remove next feature with largest P > 0.05 to make reduced model 2

```
# Remove Cohort from reduced model 1
reduced2_glmer_model <- update(reduced1_glmer_model, ~ . -Cohort)
anova(reduced1_glmer_model,reduced2_glmer_model)
```

```
## Data: endom_burden
## Models:
## reduced2_glmer_model: Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + (Age -
## reduced2_glmer_model:     1 | PatientID)
## reduced1_glmer_model: Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + Cohort +
## reduced1_glmer_model:     (Age - 1 | PatientID)
##                       Df    AIC    BIC  logLik deviance  Chisq Chi Df
## reduced2_glmer_model  7 3561.5 3586.3 -1773.7   3547.5
## reduced1_glmer_model 10 3565.1 3600.6 -1772.5   3545.1 2.3871      3
##                       Pr(>Chisq)
## reduced2_glmer_model
## reduced1_glmer_model      0.496
```

```
print(reduced2_glmer_model)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerModLmerTest']
## Formula: Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + (Age -
##     1 | PatientID)
##    Data: endom_burden
##       AIC       BIC    logLik  deviance  df.resid
##  3561.482  3586.326 -1773.741  3547.482       250
## Random effects:
##  Groups    Name Std.Dev.
##  PatientID Age    3.771
##  Residual       220.280
## Number of obs: 257, groups:  PatientID, 28
## Fixed Effects:
##   (Intercept)            Age       Vafdepth  Driver_status         BMI.QC
##      -323.464         28.952         28.681        110.772          6.553
```

```
lme4:::drop1.merMod(reduced2_glmer_model, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + (Age -
##     1 | PatientID)
##               Df    AIC    LRT   Pr(Chi)
## <none>           3561.5
## Age            1 3605.6 46.093 1.128e-11 ***
## Vafdepth       1 3587.3 27.855 1.308e-07 ***
## Driver_status  1 3569.9 10.413  0.001251 **
## BMI.QC         1 3560.1  0.593  0.441211
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Remove next feature with largest P > 0.05 to make reduced model 3

```
# Remove BMI information from reduced model 2
reduced3_glmer_model <- update(reduced2_glmer_model, ~ . -BMI.QC)
anova(reduced2_glmer_model,reduced3_glmer_model)
```

```
## Data: endom_burden
## Models:
## reduced3_glmer_model: Subs_tree ~ Age + Vafdepth + Driver_status + (Age - 1 | PatientID)
## reduced2_glmer_model: Subs_tree ~ Age + Vafdepth + Driver_status + BMI.QC + (Age -
## reduced2_glmer_model:     1 | PatientID)
##                       Df    AIC    BIC  logLik deviance  Chisq Chi Df
## reduced3_glmer_model  6 3560.1 3581.4 -1774.0   3548.1
## reduced2_glmer_model  7 3561.5 3586.3 -1773.7   3547.5 0.5931      1
##                       Pr(>Chisq)
## reduced3_glmer_model
## reduced2_glmer_model     0.4412
```

Define the final model

```
# Define final model keeping all features that are significant with P < 0.05
  final_glmer_model <- reduced3_glmer_model

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

```
## 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
##
##      AIC      BIC   logLik deviance df.resid
##   3560.1   3581.4  -1774.0   3548.1      251
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.0371 -0.4099  0.0067  0.4361  3.9936
##
## Random effects:
##  Groups    Name Variance Std.Dev.
##  PatientID Age     14.78   3.845
##  Residual       48474.42 220.169
## Number of obs: 257, groups:  PatientID, 28
##
## Fixed effects:
##               Estimate Std. Error       df t value Pr(>|t|)
## (Intercept)   -267.398    120.757   57.039  -2.214  0.03082 *
## Age             28.620      2.732   28.290  10.477 3.02e-11 ***
## Vafdepth        29.028      5.266  255.958   5.513 8.61e-08 ***
## Driver_status  109.881     33.881  249.039   3.243  0.00134 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) Age    Vfdpth
## Age         -0.829
## Vafdepth    -0.543  0.081
## Driver_stts  0.131 -0.220 -0.161
```

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

```
##                    2.5 %    97.5 %
## .sig01                NA        NA
## .sigma                NA        NA
## (Intercept)   -504.07833 -30.71845
## Age             23.26647  33.97419
## Vafdepth        18.70793  39.34852
## Driver_status   43.47519 176.28725
```

```
# Calculate mutation rates for each donor from this model
# # randomEffects.df <- as.data.frame(ranef(final_glmer_model))
# write_csv(randomEffects.df, "model_rates.csv")
```