Clock like mutational signatures and age in HL

Introduction: clock like mutational processes and aging

Alexandrov et al. (Nat Gen 2015) demonstrated how SBS1 and SBS5 mutational signatures are acquired at constant rate over time across multiple tumor types. These two clokc like mutational processes start to accumulate mutations since the fertilized egg. Based on this assumptions, different groups have investigated the correlation between age and SBS1 and SBS5 using regression models with the intercept restricted to zero (e.g., Gerstung et al Nature 2020; Mitchell et al. Cell 2018).

Here we will investigate the correlation between age and SBS1 and SBS5 in cHL. Only WGS data will be used for this purpuse (n=25). Additional WGS from normal Naive B-cells (n=85) and Memory B-cells (n=53) were included as well from Machado et al. Nature 2022.

```
knitr::opts_chunk$set(echo = TRUE)
library(stringr)
library(MASS)
library(lme4)
```

```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.0.5
```

```
library(RColorBrewer)
```

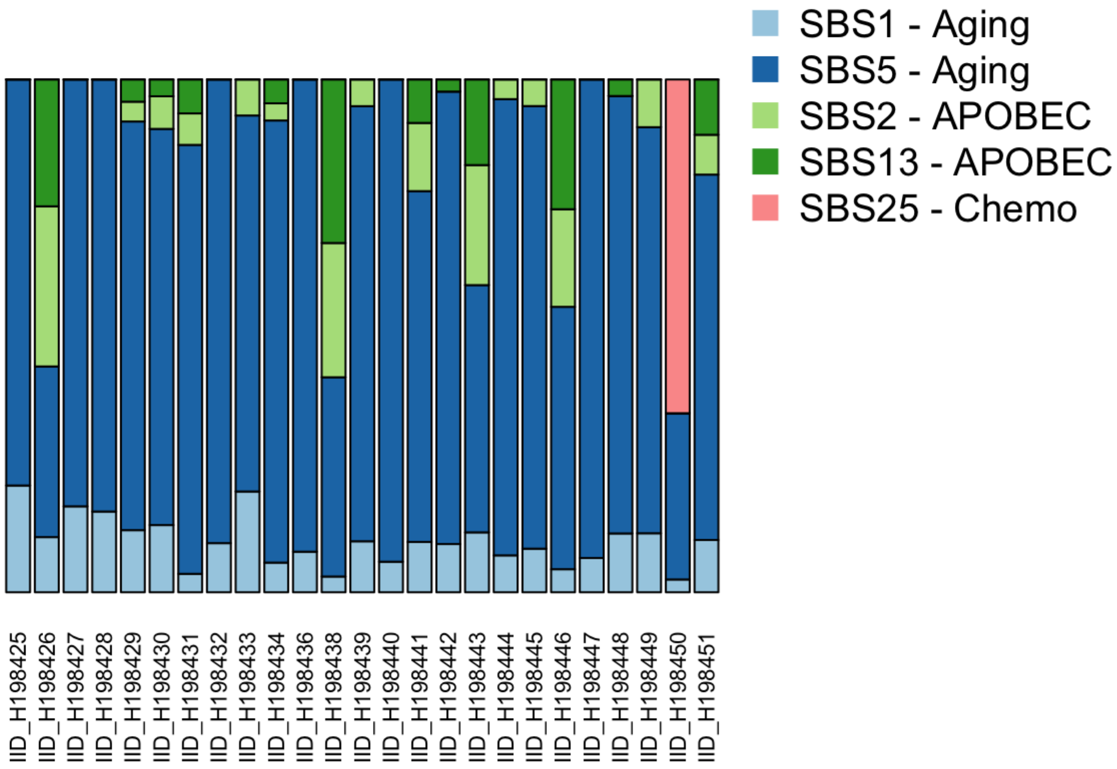
```
## Warning: package 'RColorBrewer' was built under R version 4.0.5
```

Upload files from HL WGS

```
### file with mutational signature contribution for all cHL included in the study

comb_all<- read.delim("signatures.counts.txt")
comb_all_wgs<- comb_all[comb_all$seq.x=="wgs",]
head(comb_all_wgs)
```

```
##        sample   SBS1   SBS5    SBS2   SBS13 SBS25  tot Age_Category Dx.Status
## 1 IID_H198425 0.2082 0.7918 0.00000 0.00000     0 2634  Older Adult diagnosis
## 2 IID_H198426 0.1076 0.3329 0.31220 0.24737     0 1880  Older Adult diagnosis
## 3 IID_H198427 0.1676 0.8324 0.00000 0.00000     0 3547  Older Adult diagnosis
## 4 IID_H198428 0.1575 0.8425 0.00000 0.00000     0 2574  Older Adult diagnosis
## 5 IID_H198429 0.1211 0.7969 0.03859 0.04334     0 4324  Older Adult diagnosis
## 6 IID_H198430 0.1311 0.7726 0.06359 0.03279     0 5367     AYA_Peds diagnosis
##   EBV.Status age seq.x  col_dg seq.y Purity Ploidy WGD
## 1        neg  58   wgs #82ed82   wgs   0.91   2.20   0
## 2        neg  55   wgs #82ed82   wgs   0.50   3.45   1
## 3        neg  76   wgs #82ed82   wgs   0.74   2.25   0
## 4        neg  69   wgs #82ed82   wgs   0.61   2.40   0
## 5        neg  66   wgs #82ed82   wgs   0.85   3.35   1
## 6        neg  26   wgs #82ed82   wgs   0.86   2.25   0
```

Plot SBS signatures contribution across 25 cHL WGS

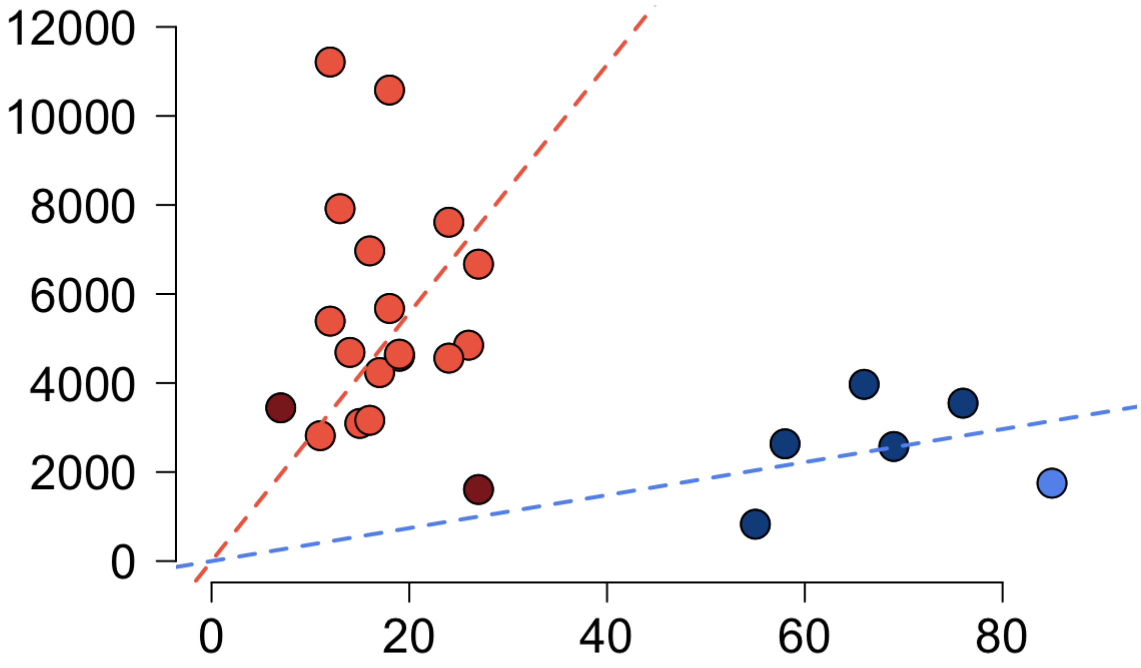


WGS are divided according to EBV status and age

```
comb_all_wgs$SBS1_5_abs<- (comb_all_wgs$SBS1+comb_all_wgs$SBS5)*comb_all_wgs$tot
comb_all_wgs$color<- "cornflowerblue"
comb_all_wgs$color[comb_all_wgs$age<40 & comb_all_wgs$EBV.Status=="neg"]<-"coral2"
comb_all_wgs$color[comb_all_wgs$age>40 & comb_all_wgs$EBV.Status=="neg"]<-"dodgerblue4"
comb_all_wgs$color[comb_all_wgs$age<40 & comb_all_wgs$EBV.Status=="pos"]<-"brown4"
comb_all_wgs$color[comb_all_wgs$age<40 & is.na(comb_all_wgs$EBV.Status)]<-"coral2"
```

Plot correlation between age and SBS/SBS5 mutational burden

```
par(xpd=F, mar=c(5,5,5,5), mfrow=c(1,1))
plot(comb_all_wgs$age, comb_all_wgs$SBS1_5_abs, pch=21, ylim=c(0,12000),bty="n",
     xlim=c(0,90), bg=comb_all_wgs$color, las=2, yaxt="n", xaxt="n", yla="", xlab="", ce
x=2)
par(new=T)
plot(comb_all_wgs$age, comb_all_wgs$SBS1_5_abs, pch=21, ylim=c(0,12000),bty="n",
     xlim=c(0,90), bg=comb_all_wgs$color, las=2, yaxt="n", xaxt="n", yla="", xlab="", ce
x=2)
abline((lm(SBS1_5_abs~0+age , data= comb_all_wgs[comb_all_wgs$Age_Category =="AYA_Ped
s",])), col="coral2", lty=2, lwd=2)
abline((lm(SBS1_5_abs~0+age , data= comb_all_wgs[comb_all_wgs$Age_Category !="AYA_Ped
s",])), col="cornflowerblue", lty=2, lwd=2)
axis(side = 1, at=seq(0,90, 20), labels = seq(0,90, 20), las=1, cex.axis=1.5)
axis(side = 2, at=seq(0,12000, 2000), labels = seq(0,12000, 2000), las=1, cex.axis=1.5)
```



Correlation between age and SBS1/SBS5 mutational burden in Pediatric and young adolescent cHL (Ped/AYA). The intercept was constrained to zero.

```
summary(lm(SBS1_5_abs~0+age , data= comb_all_wgs[comb_all_wgs$Age_Category =="AYA_Ped
s",]))
```

```
##
## Call:
## lm(formula = SBS1_5_abs ~ 0 + age, data = comb_all_wgs[comb_all_wgs$Age_Category ==
##     "AYA_Peds", ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5919   -972   -249   1768   7867
##
## Coefficients:
##     Estimate Std. Error t value  Pr(>|t|)
## age    278.8       38.8    7.19 0.0000011 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3130 on 18 degrees of freedom
## Multiple R-squared:  0.742,  Adjusted R-squared:  0.727
## F-statistic: 51.7 on 1 and 18 DF,  p-value: 0.00000108
```

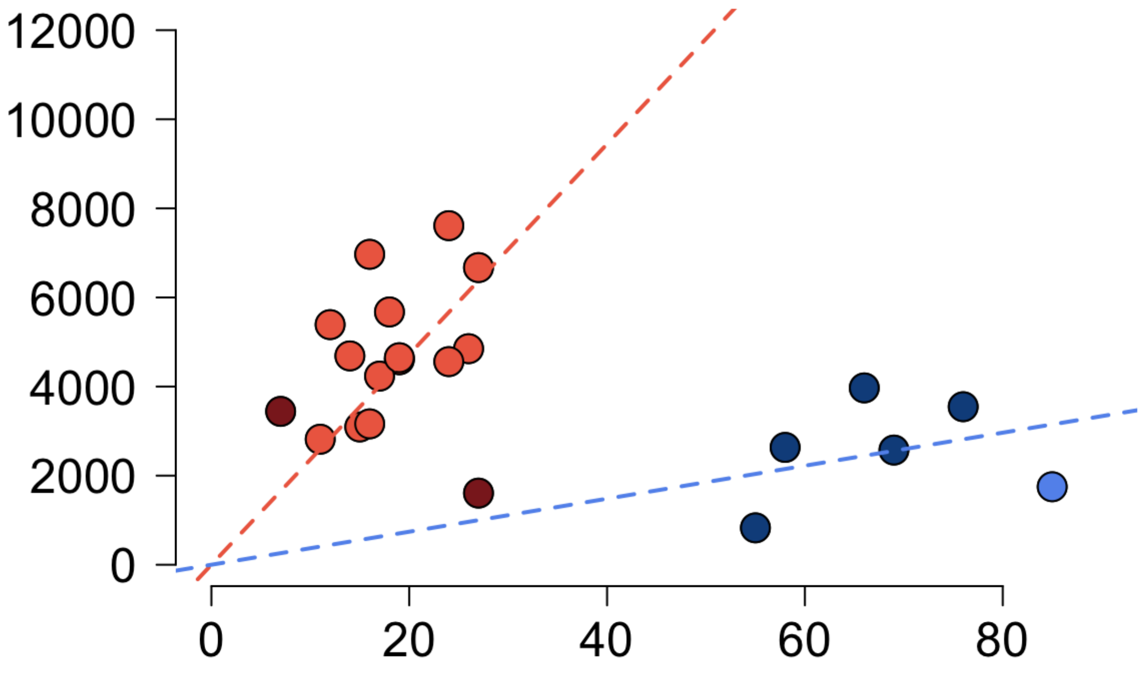
Correlation between age and SBS1/SBS5 mutational burden in Older Adults cHL (Ped/AYA). The intercept was constrain to zero

```
summary(lm(SBS1_5_abs~0+age , data= comb_all_wgs[comb_all_wgs$Age_Category !="AYA_Ped
s",]))
```

```
##
## Call:
## lm(formula = SBS1_5_abs ~ 0 + age, data = comb_all_wgs[comb_all_wgs$Age_Category !=
##     "AYA_Peds", ])
##
## Residuals:
##       1       2       3       4       5      24
##   484.7 -1209.9   731.1    17.5  1524.2 -1399.3
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## age    37.05       6.76    5.48   0.0028 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1140 on 5 degrees of freedom
## Multiple R-squared:  0.857,  Adjusted R-squared:  0.829
## F-statistic:   30 on 1 and 5 DF,  p-value: 0.00276
```

Test the same correlation by removing hypermutated samples, similarly to what has been done in other clock-like analyses (e.g. Maura et al. Nat Comm 2020; Gerstung et al. Nature 2020).

```
comb_all_wgs_no_hyper<- comb_all_wgs[comb_all_wgs$tot<10000,]
par(xpd=F, mar=c(5,5,5,5), mfrow=c(1,1))
plot(comb_all_wgs_no_hyper$age, comb_all_wgs_no_hyper$SBS1_5_abs, pch=21, ylim=c(0,12000),bty="n",
     xlim=c(0,90), bg=comb_all_wgs_no_hyper$color, las=2, yaxt="n", xaxt="n", yla="", xlab="", cex=2)
par(new=T)
plot(comb_all_wgs_no_hyper$age, comb_all_wgs_no_hyper$SBS1_5_abs, pch=21, ylim=c(0,12000),bty="n",
     xlim=c(0,90), bg=comb_all_wgs_no_hyper$color, las=2, yaxt="n", xaxt="n", yla="", xlab="", cex=2)
abline((lm(SBS1_5_abs~0+age , data= comb_all_wgs_no_hyper[comb_all_wgs_no_hyper$Age_Category =="AYA_Peds",])), col="coral2", lty=2, lwd=2)
abline((lm(SBS1_5_abs~0+age , data= comb_all_wgs_no_hyper[comb_all_wgs_no_hyper$Age_Category !="AYA_Peds",])), col="cornflowerblue", lty=2, lwd=2)
axis(side = 1, at=seq(0,90, 20), labels = seq(0,90, 20), las=1, cex.axis=1.5)
axis(side = 2, at=seq(0,12000, 2000), labels = seq(0,12000, 2000), las=1, cex.axis=1.5)
```



```
summary(lm(SBS1_5_abs~0+age , data= comb_all_wgs_no_hyper[comb_all_wgs_no_hyper$Age_Category !="AYA_Peds",]))
```

```
##
## Call:
## lm(formula = SBS1_5_abs ~ 0 + age, data = comb_all_wgs_no_hyper[comb_all_wgs_no_hyper
$Age_Category !=
##     "AYA_Peds", ])
##
## Residuals:
##       1       2       3       4       5      24
##   484.7 -1209.9   731.1    17.5  1524.2 -1399.3
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## age    37.05       6.76    5.48   0.0028 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1140 on 5 degrees of freedom
## Multiple R-squared:  0.857,  Adjusted R-squared:  0.829
## F-statistic:   30 on 1 and 5 DF,  p-value: 0.00276
```

```
summary(lm(SBS1_5_abs~0+age , data= comb_all_wgs_no_hyper[comb_all_wgs_no_hyper$Age_Cate
gory =="AYA_Peds",]))
```

```
##
## Call:
## lm(formula = SBS1_5_abs ~ 0 + age, data = comb_all_wgs_no_hyper[comb_all_wgs_no_hyper
$Age_Category ==
##     "AYA_Peds", ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4767   -490    220   1516   3191
##
## Coefficients:
##     Estimate Std. Error t value    Pr(>|t|)
## age    236.1       24.8    9.53 0.000000094 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1900 on 15 degrees of freedom
## Multiple R-squared:  0.858,  Adjusted R-squared:  0.849
## F-statistic: 90.8 on 1 and 15 DF,  p-value: 0.0000000942
```

Upload SBS data from Machado et al. Nature 2022 (SBS data were re-analyzed using mmsig

mmsig code can be found here: https://github.com/UM-Myeloma-Genomics/mmsig (https://github.com/UM-Myeloma-Genomics/mmsig) Machado et al. Nature 2022 data can be found here: https://github.com/machadoheather/lymphocyte_somatic_mutation (https://github.com/machadoheather/lymphocyte_somatic_mutation)

```
machado<- read.delim("machado_mmsig.txt")
head(machado)
```

```
##    sample CellType Cell.type2 Tissue Age Nmut    SBS1   SBS5 SBS8   SBS9 SBS18
## 12 B1_C10  B Naive    Naive B  blood  63  998 0.11534 0.8847    0 0.0000     0
## 13  B1_D8  B Naive    Naive B  blood  63  838 0.10360 0.8964    0 0.0000     0
## 14  B1_G7 B Memory   Memory B  blood  63 1970 0.05884 0.4786    0 0.4626     0
## 15 B10_G7 B Memory   Memory B  blood  63 1289 0.08855 0.5926    0 0.3189     0
## 16 B11_A7  B Naive    Naive B  blood  63  732 0.14148 0.8585    0 0.0000     0
## 17 B12_B4  B Naive    Naive B  blood  63  446 0.13450 0.8655    0 0.0000     0
##    SBS7a SBS17b mutations  id
## 12     0      0       997  B1
## 13     0      0       836  B1
## 14     0      0      1960  B1
## 15     0      0      1287 B10
## 16     0      0       730 B11
## 17     0      0       445 B12
```

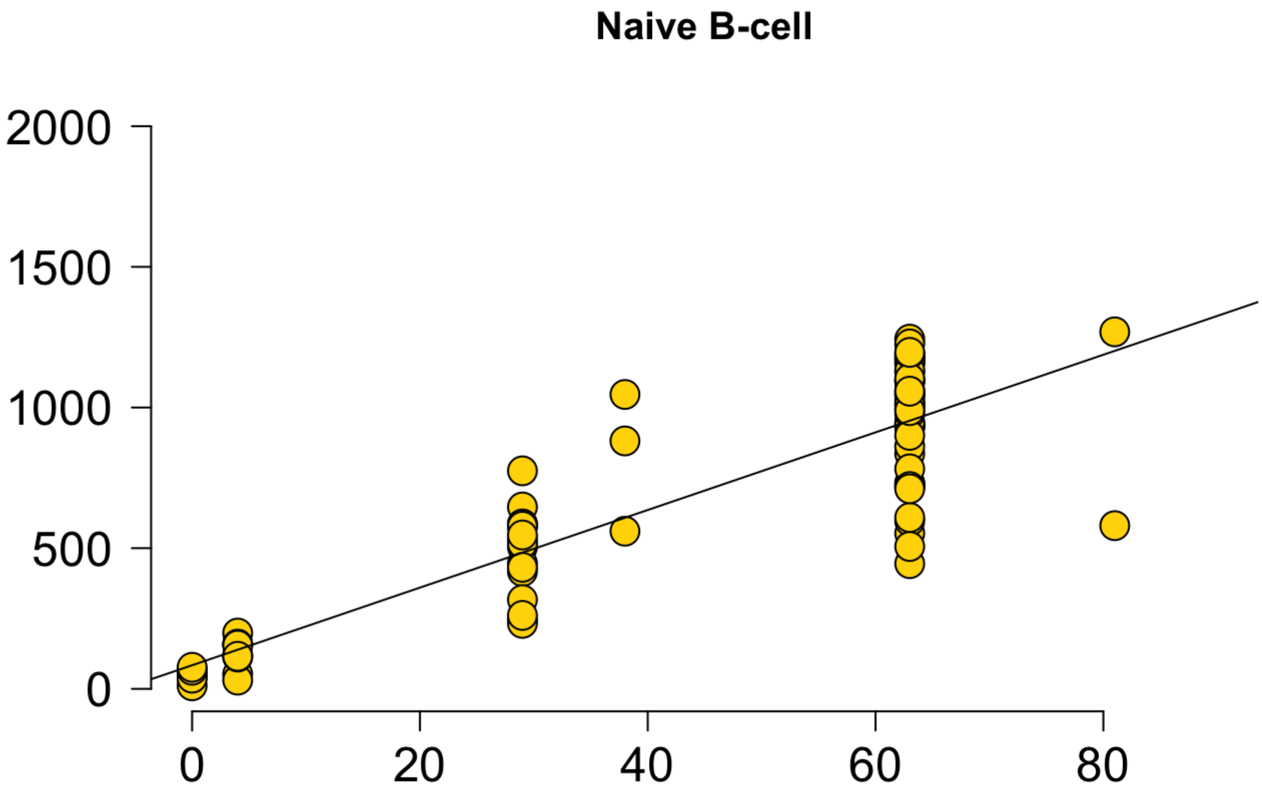
```
# remove hypermutated cases as done in the original paper
machado<- machado[machado$mutations<2000,]

machado$ageing<- (machado$SBS1+machado$SBS5)*machado$mutations
mycelltypes = c("Naive B", "Memory B")
out<- str_split_fixed(machado$id, "", 8)
machado$sample_ID<- paste0(out[,1],out[,2],out[,3],out[,4],
                           out[,5],out[,6],out[,7])

machado$Num.mutations<- (machado$SBS1+ machado$SBS5)*machado$mutations
machado_naive<- machado[machado$Cell.type2 == "Naive B",]
machado_mem<- machado[machado$Cell.type2 != "Naive B",]
```

Plot correlation between SBS1 and SBS5 mutational burden among naive B-cell using linear mixed effect model.

```
plot( machado_naive$Age, ((machado_naive$SBS1+ machado_naive$SBS5)*machado_naive$mutatio
ns),pch=21, ylim=c(0,2000),bty="n",
      xlim=c(0,90), bg="gold", las=2, yaxt="n", xaxt="n", yla="", xlab="", cex=2, main
="Naive B-cell")
muts.naive.per.year.lmer <- lmer(Num.mutations ~ Age + (1 + Age | sample_ID ), data=mach
ado_naive, REML=FALSE)
```

```
## boundary (singular) fit: see help('isSingular')
```

```
model_naive<- coef(summary(muts.naive.per.year.lmer))[ , "Estimate"]
abline(a = model_naive[1], b = model_naive[2])
axis(side = 1, at=seq(0,90, 20), labels = seq(0,90, 20), las=1, cex.axis=1.5)
axis(side = 2, at=seq(0,2000, 500), labels = seq(0,2000, 500), las=1, cex.axis=1.5)
```



```
summary(muts.naive.per.year.lmer)
```

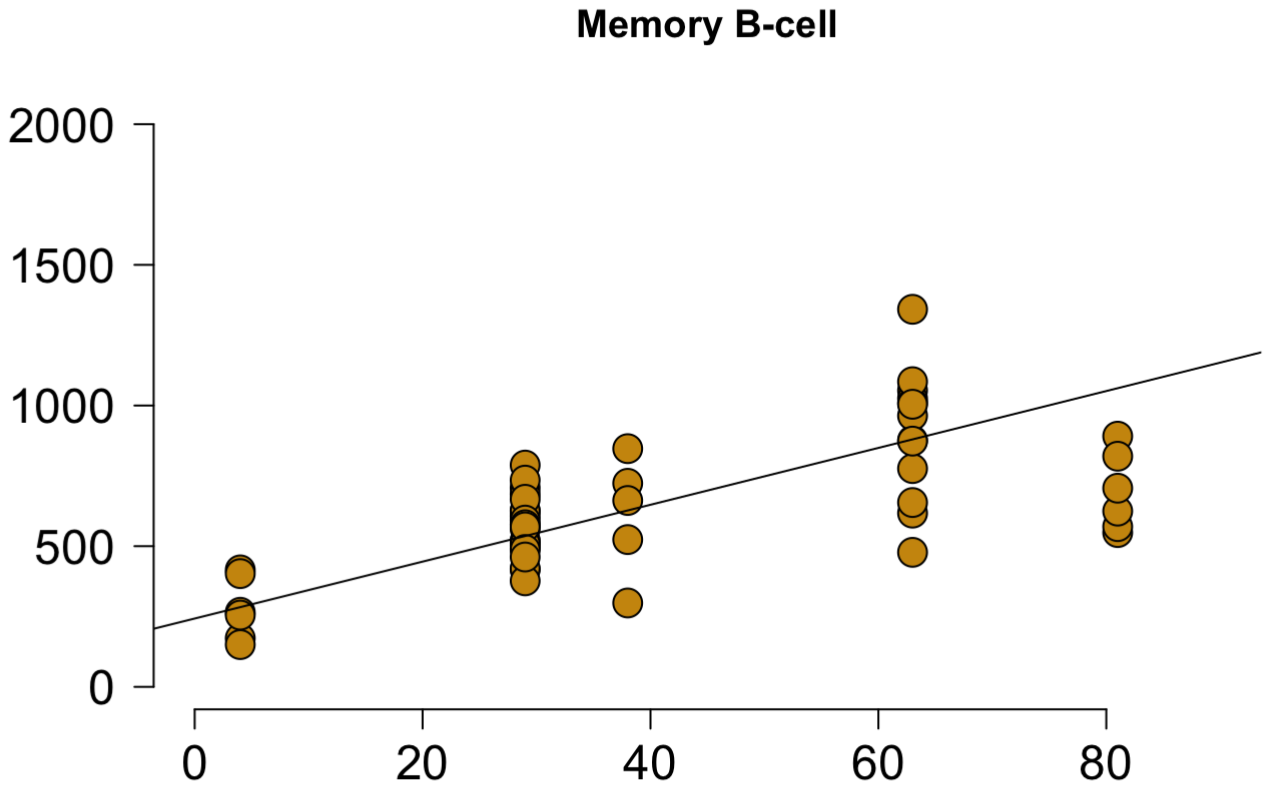
```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Num.mutations ~ Age + (1 + Age | sample_ID)
##    Data: machado_naive
##
##      AIC      BIC   logLik deviance df.resid
##   1143.7   1158.4   -565.9   1131.7       79
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -3.300 -0.356  0.101  0.556  2.327
##
## Random effects:
##  Groups    Name        Variance           Std.Dev.    Corr
##  sample_ID (Intercept)     0.000000000000   0.0000000
##            Age             0.000000000253   0.0000159  NaN
##  Residual              35468.033532631984 188.3295875
## Number of obs: 85, groups:  sample_ID, 21
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   83.862     41.034    2.04
## Age           13.799      0.817   16.89
##
## Correlation of Fixed Effects:
##     (Intr)
## Age -0.867
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```

Plot correlation between SBS1 and SBS5 mutational burden among memory B-cell using linear mixed effect model.

```
plot( machado_mem$Age, ((machado_mem$SBS1+ machado_mem$SBS5)*machado_mem$mutations),pch=21, ylim=c(0,2000),bty="n",
      xlim=c(0,90), bg="darkgoldenrod3", las=2, yaxt="n", xaxt="n", yla="", xlab="", cex=2, main="Memory B-cell")
muts.mem.per.year.lmer <- lmer(Num.mutations ~ Age + (1 + Age | sample_ID ), data=machado_mem, REML=FALSE)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge with max|grad| = 0.0072715 (tol = 0.002, component 1)
```

```
model_mem<- coef(summary(muts.mem.per.year.lmer))[ , "Estimate"]
abline(a = model_mem[1], b = model_mem[2])
axis(side = 1, at=seq(0,90, 20), labels = seq(0,90, 20), las=1, cex.axis=1.5)
axis(side = 2, at=seq(0,2000, 500), labels = seq(0,2000, 500), las=1, cex.axis=1.5)
```



```
summary(muts.mem.per.year.lmer)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: Num.mutations ~ Age + (1 + Age | sample_ID)
##    Data: machado_mem
## 
##      AIC      BIC   logLik deviance df.resid 
##    693.6    705.4   -340.8    681.6       47 
## 
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -2.4161 -0.6298 -0.0355  0.5411  2.0157 
## 
## Random effects:
##  Groups    Name        Variance Std.Dev. Corr 
##  sample_ID (Intercept)  1891.0   43.49        
##            Age            13.6    3.69   -1.00
##  Residual              15320.8  123.78        
## Number of obs: 53, groups:  sample_ID, 15
## 
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   242.91      53.32    4.56
## Age            10.11       1.43    7.06
## 
## Correlation of Fixed Effects:
##     (Intr)
## Age -0.777
## optimizer (nloptwrap) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.0072715 (tol = 0.002, component 1)
```

Naive and Memory B-cell showed a similar mutation rate per year, but different intercepts. This might be explained by the fact that memory B-cell experience germinal center and poly-eta exposure. This process increases the mutational burden, in particular through a distinct SBS signature: SBS9. SBS9 shared similar trinucleotides with SBS5 and this might create an inter-bleeding of signatures.

Depite these considerations, we re-analyzed the WGS HL data constraining the intercept to the memory and naive B-cell values.

```
### naive B-cell in AYA/Ped
summary(lm(I(SBS1_5_abs - 83) ~ 0 +age, data= comb_all_wgs[comb_all_wgs$Age_Category =
="AYA_Peds",]))
```

```
##
## Call:
## lm(formula = I(SBS1_5_abs - 83) ~ 0 + age, data = comb_all_wgs[comb_all_wgs$Age_Categ
ory ==
##     "AYA_Peds", ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5886   -966   -285   1725   7835
##
## Coefficients:
##     Estimate Std. Error t value  Pr(>|t|)
## age    274.5       38.6    7.11 0.0000012 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3110 on 18 degrees of freedom
## Multiple R-squared:  0.738,  Adjusted R-squared:  0.723
## F-statistic: 50.6 on 1 and 18 DF,  p-value: 0.00000125
```

```
### naive B-cell in Older Adults
summary(lm(I(SBS1_5_abs - 83) ~ 0 +age, data= comb_all_wgs[comb_all_wgs$Age_Category !
="AYA_Peds",]))
```

```
##
## Call:
## lm(formula = I(SBS1_5_abs - 83) ~ 0 + age, data = comb_all_wgs[comb_all_wgs$Age_Categ
ory !=
##     "AYA_Peds", ])
##
## Residuals:
##       1       2       3       4       5      24
##   470.8 -1227.4   738.6    16.7  1519.8 -1381.1
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## age    35.86       6.75    5.31   0.0032 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1140 on 5 degrees of freedom
## Multiple R-squared:  0.85,   Adjusted R-squared:  0.819
## F-statistic: 28.2 on 1 and 5 DF,  p-value: 0.00316
```

```
### Memory B-cell in Ped/AYA

summary(lm(I(SBS1_5_abs - 243) ~ 0 +age, data= comb_all_wgs[comb_all_wgs$Age_Category =
="AYA_Peds",]))
```

```
##
## Call:
## lm(formula = I(SBS1_5_abs - 243) ~ 0 + age, data = comb_all_wgs[comb_all_wgs$Age_Cate
gory ==
##     "AYA_Peds", ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5824   -953   -354   1644   7774
##
## Coefficients:
##     Estimate Std. Error t value  Pr(>|t|)
## age    266.3       38.2    6.97 0.0000017 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3080 on 18 degrees of freedom
## Multiple R-squared:  0.729,  Adjusted R-squared:  0.714
## F-statistic: 48.5 on 1 and 18 DF,  p-value: 0.00000166
```

```
### Memory B-cell in Older Adults
summary(lm(I(SBS1_5_abs - 243) ~ 0 +age, data= comb_all_wgs[comb_all_wgs$Age_Category !
="AYA_Peds",]))
```

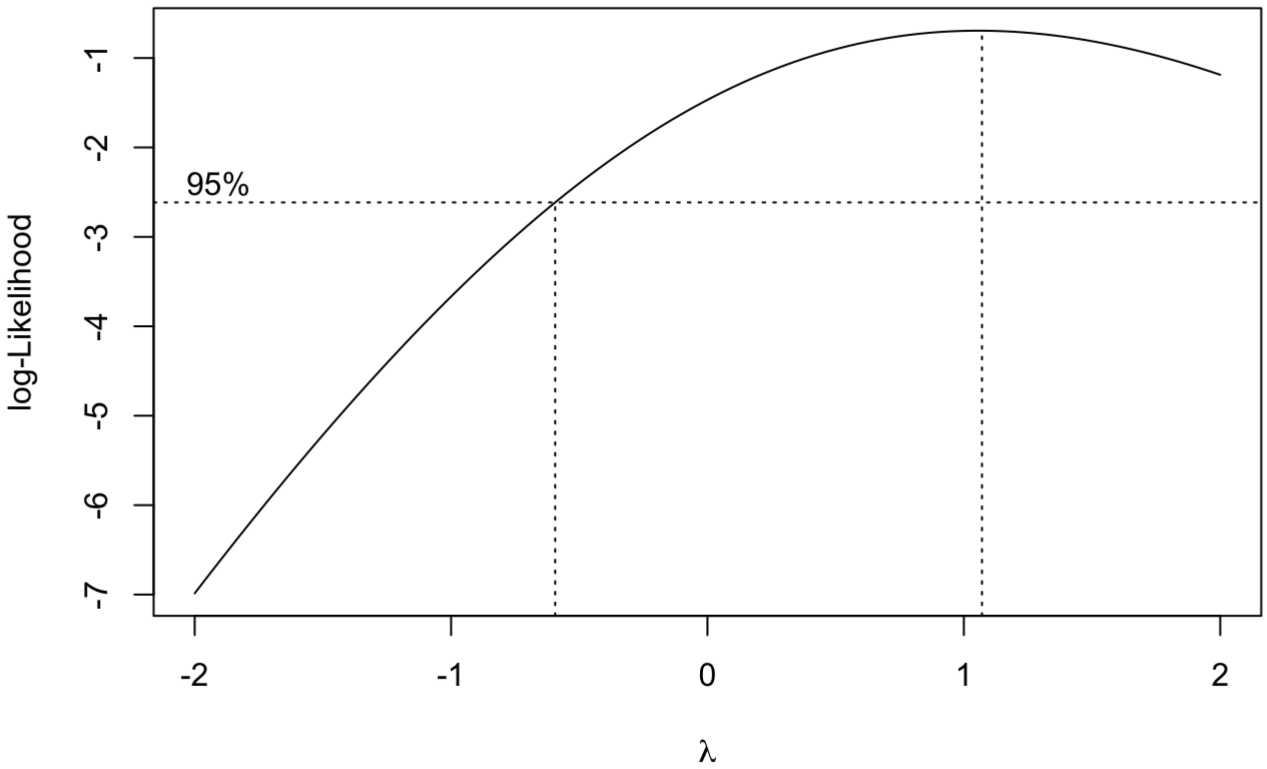
```
##
## Call:
## lm(formula = I(SBS1_5_abs - 243) ~ 0 + age, data = comb_all_wgs[comb_all_wgs$Age_Cate
gory !=
##     "AYA_Peds", ])
##
## Residuals:
##       1       2       3       4       5      24
##   444.0 -1261.2   753.1    15.1  1511.3 -1346.0
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## age    33.57       6.73    4.99   0.0041 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1140 on 5 degrees of freedom
## Multiple R-squared:  0.833,  Adjusted R-squared:  0.799
## F-statistic: 24.9 on 1 and 5 DF,  p-value: 0.00414
```

Linear model vs non-linear model

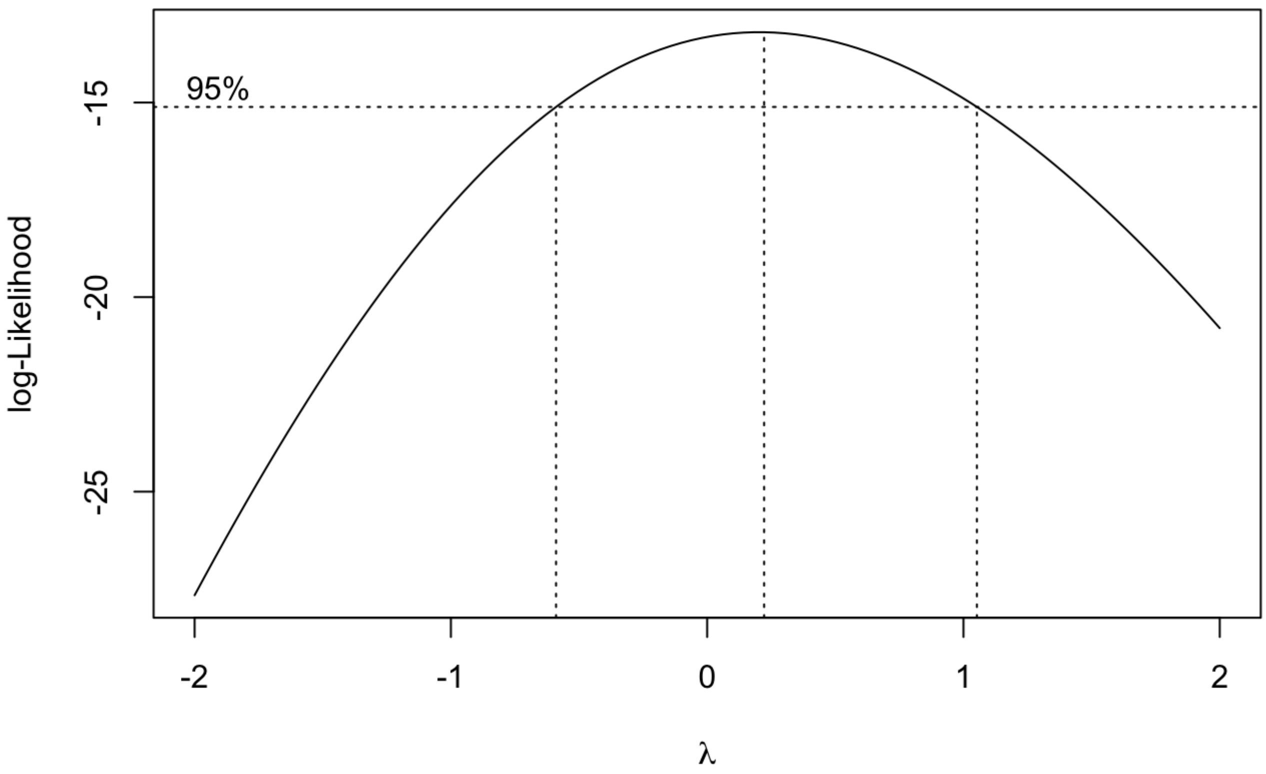
To better explore the clock-like mutational rate across our cohort we have tested multiple models using the R function boxcox to see which one better explains the clock-like mutation distribution.

We observed that the linearity of SBS1-SBS5 in pediatric/AYA can explain the distribution better than other models (see below).

```
### Older Adults
comb_all_wgs_old<- comb_all_wgs[comb_all_wgs$Age_Category !="AYA_Peds",]
boxcox(lm(comb_all_wgs_old$SBS1_5_abs~0+comb_all_wgs_old$age))
```



```
### Ped/AYA
comb_all_wgs_40<- comb_all_wgs[comb_all_wgs$Age_Category =="AYA_Peds",]
boxcox(lm(comb_all_wgs_40$SBS1_5_abs~0+comb_all_wgs_40$age))
```


95%
log-Likelihood
-15
-20
-25
-2
-1
0
1
2
λ