CodeS2: Burden and signature analyses

Sigurgeir Ólafsson

2/20/2022

Introduction

This document describes the clonality and burden analyses carried out as part of our manuscript “Effects of psoriasis and psoralen exposure on the somatic mutation landscape of the skin” by Sigurgeir Ólafsson et al.

This analysis uses pre-calculated summary statistics provided as supplementary tables. The mutation calls can be accessed from a Mendeley-Data repository that accompanies the manuscript and the raw sequencing data has also been made publicly available, please see the manuscript for details.

```
.libPaths("/lustre/scratch126/humgen/projects/psoriasis/R_packages_farm5_R4.1.0_install/")

library(ggplot2)
library(reshape2)
library(cowplot)
library(ggsignif)
library(nlme)

## DEFINE PLOTTING VARIABLES
############################
BASESIZE=14

# Location colour vector
#Abdomen     Arm    Back   Flank     Leg
loc_colors <- c("#264653", "#2A9D8F","#E9C46A","#F4A261","#E76F51")

# Disease type (lesional vs non-lesional) colour vector
type_colours <- c("#FF7075", "#5DB4EA")

## Read in the meta-data
############################

working_dir="/nfs/users/nfs_s/so11/phd/psoriasis/bsub_jupyter_lab/psoriasis/manuscript_data_and_figures/Supplemen
tary_material/"

microd_meta <- read.table(paste(working_dir, "Supplementary_Table2_microdissection_metadata.txt", sep=""), h=T)
patient_meta <- read.table(paste(working_dir, "Supplementary_Table1_patient_metadata.txt", sep=""), h=T)

biopsy_meta <- unique(microd_meta[microd_meta$ExclusionCriteria=="PASS",c("BiopsyID", "MetaLocation", "DiseaseSta
tus")])
table(biopsy_meta$MetaLocation, biopsy_meta$DiseaseStatus, useNA="always")
```

```
##
##           Lesional Non-lesional <NA>
##   Abdomen        9            8    0
##   Arm           20           21    0
##   Back          32           27    0
##   Flank         22           25    0
##   Leg           28           25    0
##   <NA>           0            0    0
```

```
table(microd_meta$MetaLocation[microd_meta$ExclusionCriteria=="PASS"], microd_meta$DiseaseStatus[microd_meta$Excl
usionCriteria=="PASS"], useNA="always")
```

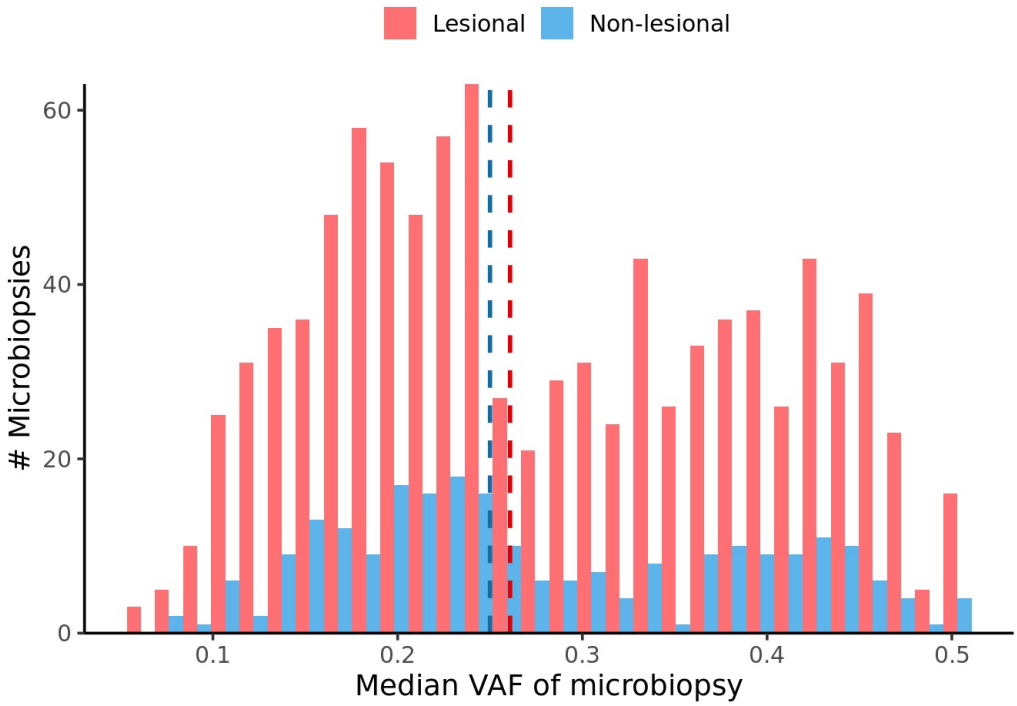
```
##
##           Lesional Non-lesional <NA>
##   Abdomen       70           20    0
##   Arm          176           45    0
##   Back         294           54    0
##   Flank        186           57    0
##   Leg          220           60    0
##   <NA>           0            0    0
```

```
table(patient_meta$Sex)
```

```
##
## Female   Male
##     23     88
```

Clonality analysis

First compare the median variant allele frequencies (VAFs) of microbiopsies derrived from lesional and non-lesional skin. We see that they are near identical. Most somatic mutations are heterozygous, so in a fully clonal sample we would expect the median VAF to be 0.5. Most microbiopsies are a mix of clones and have median VAFs lower than 0.5.



Mutation burden analyses

Since the microbiopsies tend to be a mix of cell clones, they don't give a good estimate of the per-cell mutation burden. Instead, I have computationally grouped mutations by their VAF into clusters. I have then used the pigeonhole principle to construct phylogenetic trees from the clusters. The mutation burden analyses are done on the level of the tips of the phylogenetic trees, which I refer to here as clones.

I have extracted mutational signatures for each cluster and computed the total mutation burden and the burden of each individual signature in each clone by summing accross the clusters. The scripts for doing each individual step are available on the Github page accompanying the manuscript but the below analyses will simply read in the results.

```
clone_burden <- read.table(paste(working_dir, "Supplementary_Table3_clone_mutationBurden.txt", sep=""), h=T)
clone_burden <- merge(clone_burden, microd_meta[,c("SampleID", "BiopsyID", "MetaLocation","PatientID", "DiseaseSt
atus")], by.x="HighCellFrac_sample", by.y="SampleID")
clone_burden <- merge(clone_burden, patient_meta[,c("Patient.ID", "Age_at_sampling", "Disease_duration","Sex")],
by.x="PatientID", by.y="Patient.ID")

clone_burden$Disease_duration[clone_burden$DiseaseStatus=="Non-lesional" & !is.na(clone_burden$Disease_duration)]
<- 0
```

Total mutation burden

We can first look at the total mutation burden as a function of age. Fit a linear mixed effects model with a fixed effects for age and the anatomical location of the sample and random effects for patient and for biopsy (nested within that of patient). Then add a fixed effect for disease duration (set to 0 for non-lesional samples, see above) and test if the fit of the model is improved using a likelihood ratio test.

```
model_null.lme <- lme(fixed = TotalSBS_adj ~ Age_at_sampling  + MetaLocation,
                      random = list(PatientID = pdSymm(form = ~ Age_at_sampling - 1), BiopsyID = pdSymm(form = ~
Age_at_sampling - 1)),
                      weights = varIdent(form= ~ 1 | DiseaseStatus),
                      data = clone_burden[!is.na(clone_burden$Disease_duration),], method="ML")

model_dur.lme <- lme(fixed = TotalSBS_adj ~ Age_at_sampling  + MetaLocation + Disease_duration,
                      random = list(PatientID = pdSymm(form = ~ Age_at_sampling - 1), BiopsyID = pdSymm(form = ~
Age_at_sampling - 1)),
                      weights = varIdent(form= ~ 1 | DiseaseStatus),
                      data = clone_burden[!is.na(clone_burden$Disease_duration),], method="ML")

summary(model_null.lme)
```

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: clone_burden[!is.na(clone_burden$Disease_duration), ]
##       AIC      BIC    logLik
##   17032.8 17082.84 -8506.402
##
## Random effects:
##  Formula: ~Age_at_sampling - 1 | PatientID
##         Age_at_sampling
## StdDev:        7.680529
##
##  Formula: ~Age_at_sampling - 1 | BiopsyID %in% PatientID
##         Age_at_sampling Residual
## StdDev:        6.819941 1054.584
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | DiseaseStatus
##  Parameter estimates:
## Non-lesional     Lesional
##    1.0000000    0.3868207
## Fixed effects:  TotalSBS_adj ~ Age_at_sampling + MetaLocation
##                        Value Std.Error  DF   t-value p-value
## (Intercept)       -211.14474 163.37938 902 -1.292359  0.1966
## Age_at_sampling     17.32682   2.82333 100  6.137007  0.0000
## MetaLocationArm    125.51918 165.13669  92  0.760093  0.4491
## MetaLocationBack   -41.30589 156.24096  92 -0.264373  0.7921
## MetaLocationFlank  -58.97436 157.25017  92 -0.375035  0.7085
## MetaLocationLeg      2.50996 158.13177  92  0.015873  0.9874
##  Correlation:
##                   (Intr) Ag_t_s MtLctA MtLctB MtLctF
## Age_at_sampling   -0.586
## MetaLocationArm   -0.626 -0.040
## MetaLocationBack  -0.630 -0.096  0.683
## MetaLocationFlank -0.662 -0.035  0.679  0.722
## MetaLocationLeg   -0.630 -0.082  0.674  0.718  0.708
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -4.98674062 -0.19899061 -0.04572419  0.18555987  8.88019700
##
## Number of Observations: 1100
## Number of Groups:
##               PatientID BiopsyID %in% PatientID
##                     102                     198
```

```
summary(model_dur.lme)
```

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: clone_burden[!is.na(clone_burden$Disease_duration), ]
##        AIC      BIC    logLik
##   17034.01 17089.04 -8506.003
##
## Random effects:
##  Formula: ~Age_at_sampling - 1 | PatientID
##         Age_at_sampling
## StdDev:        7.831848
##
##  Formula: ~Age_at_sampling - 1 | BiopsyID %in% PatientID
##         Age_at_sampling Residual
## StdDev:        6.686659 1052.549
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | DiseaseStatus
##  Parameter estimates:
## Non-lesional     Lesional
##    1.0000000    0.3875551
## Fixed effects:  TotalSBS_adj ~ Age_at_sampling + MetaLocation + Disease_duration
##                         Value Std.Error  DF   t-value p-value
## (Intercept)       -208.45790 164.00277 902 -1.271063  0.2040
## Age_at_sampling     18.19676   3.00343 100  6.058664  0.0000
## MetaLocationArm    131.83600 165.93155  91  0.794520  0.4290
## MetaLocationBack   -32.95523 157.08583  91 -0.209791  0.8343
## MetaLocationFlank  -58.75449 157.84629  91 -0.372226  0.7106
## MetaLocationLeg      1.99564 158.75761  91  0.012570  0.9900
## Disease_duration    -2.78454   3.06423  91 -0.908725  0.3659
##  Correlation:
##                   (Intr) Ag_t_s MtLctA MtLctB MtLctF MtLctL
## Age_at_sampling   -0.550
## MetaLocationArm   -0.624 -0.024
## MetaLocationBack  -0.628 -0.072  0.684
## MetaLocationFlank -0.661 -0.033  0.678  0.721
## MetaLocationLeg   -0.630 -0.077  0.674  0.717  0.708
## Disease_duration  -0.012 -0.329 -0.043 -0.057  0.001 -0.003
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -4.98429369 -0.20933691 -0.05395355  0.18416288  8.87382155
##
## Number of Observations: 1100
## Number of Groups:
##               PatientID BiopsyID %in% PatientID
##                     102                     198
```

```
anova(model_null.lme,model_dur.lme, test=T)$"p-value"[2]
```

```
## [1] 0.3715017
```

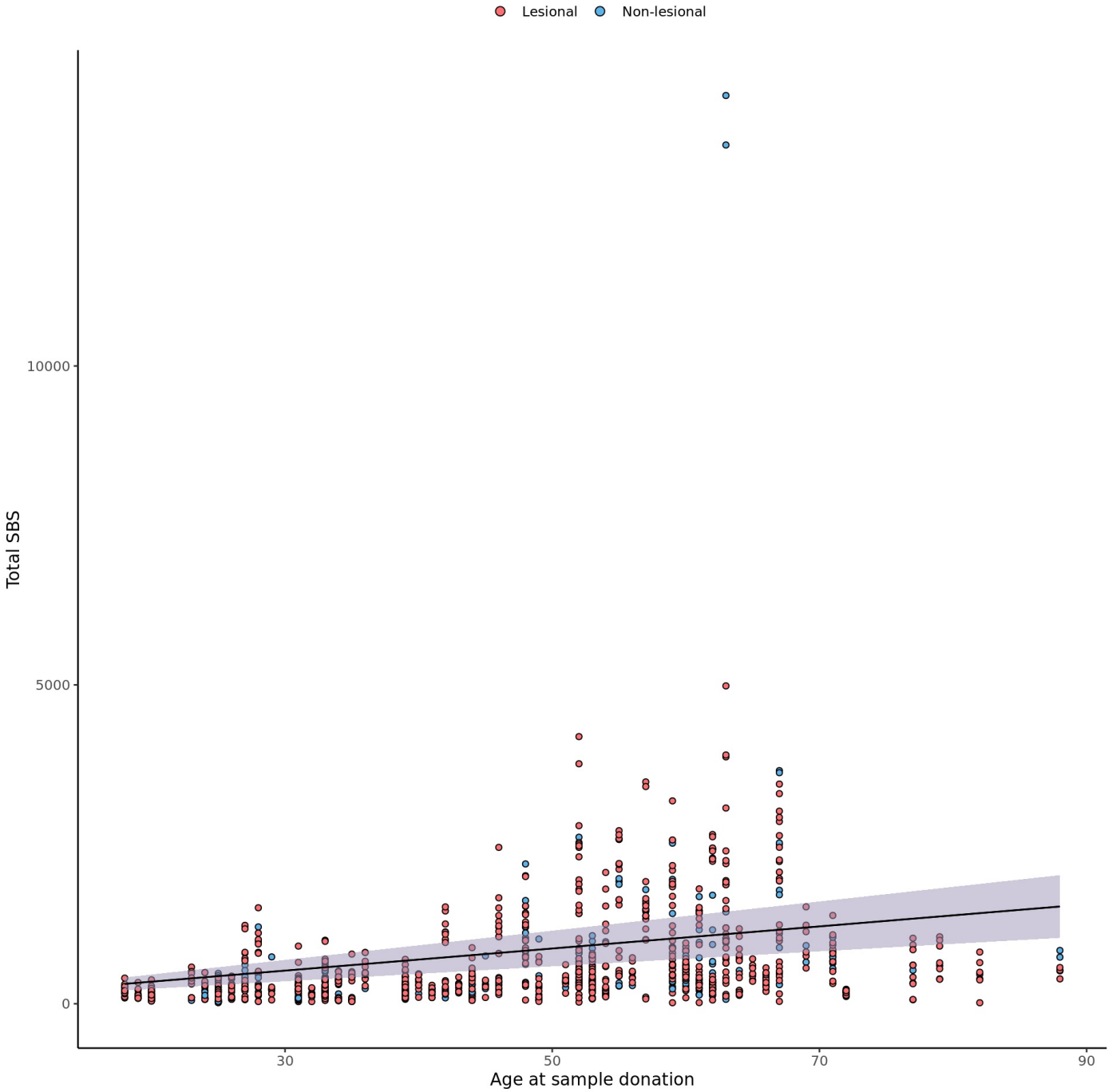
```
lme.ints <- intervals(model_null.lme, which="fixed")$fixed
lme.ints
```

```
##                       lower        est.     upper
## (Intercept)       -530.9170 -211.144742 108.62753
## Age_at_sampling     11.7407   17.326816  22.91293
## MetaLocationArm   -201.5608  125.519177 452.59920
## MetaLocationBack  -350.7665  -41.305890 268.15468
## MetaLocationFlank -370.4338  -58.974359 252.48512
## MetaLocationLeg   -310.6957    2.509962 315.71559
## attr(,"label")
## [1] "Fixed effects:"
```

We can plot the mutation burden as a function of the age of the patient. We note that there are huge outliers in the data. These are driven by a few clones having very high burden of the psoralen signature. The burden of the psoralen signature is not expected to increase linearly with age and so we wish to remove those mutations before proceeding further.

```
maxAge=max(clone_burden$Age_at_sampling)
ageEff=lme.ints["Age_at_sampling", "est."]
low <-lme.ints["Age_at_sampling", "lower"]
upp <- lme.ints["Age_at_sampling", "upper"]

ggplot(clone_burden, aes(y=TotalSBS_adj, x=Age_at_sampling, fill=DiseaseStatus)) + geom_point( colour="black", sh
ape=21) +
  scale_fill_manual(values=type_colours)  +
  labs(y="Total SBS", x="Age at sample donation") +
  theme_classic() + theme(legend.title = element_blank(), legend.position = "top") +
  geom_ribbon(aes(ymin=Age_at_sampling*low, ymax=Age_at_sampling*upp, x=Age_at_sampling), alpha = 0.3, show.legen
d=F) +
  geom_line(aes(y=Age_at_sampling*ageEff, x=Age_at_sampling)) +
  guides(fill = guide_legend(override.aes = list(size=2.5)))
```



Total burden excluding Psoralen

The model above is affected by outlier samples which have a high burden of mutations attributed to the Psoralen signature. We'll get a more representative estimate of the rate at which mutations accumulate in the skin by excluding these.

There is still no significant effect of disease duration however.

```
clone_burden$noPUVA <- clone_burden$TotalSBS_adj - clone_burden$PUVA

model_noPUVA.null <- lme(fixed = noPUVA ~ Age_at_sampling  + MetaLocation,
                      random = list(PatientID = pdSymm(form = ~ Age_at_sampling - 1), BiopsyID = pdSymm(form = ~
Age_at_sampling - 1)),
                      weights = varIdent(form= ~ 1 | DiseaseStatus),
                      data = clone_burden[!is.na(clone_burden$Disease_duration),], method="ML")

model_noPUVA.dur <- lme(fixed = noPUVA ~ Age_at_sampling  + MetaLocation + Disease_duration,
                      random = list(PatientID = pdSymm(form = ~ Age_at_sampling - 1), BiopsyID = pdSymm(form = ~
Age_at_sampling - 1)),
                      weights = varIdent(form= ~ 1 | DiseaseStatus),
                      data = clone_burden[!is.na(clone_burden$Disease_duration),], method="ML")

summary(model_noPUVA.null)
```

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: clone_burden[!is.na(clone_burden$Disease_duration), ]
##        AIC     BIC    logLik
##   16187.27 16237.3 -8083.634
##
## Random effects:
##  Formula: ~Age_at_sampling - 1 | PatientID
##         Age_at_sampling
## StdDev:        7.557317
##
##  Formula: ~Age_at_sampling - 1 | BiopsyID %in% PatientID
##         Age_at_sampling Residual
## StdDev:        3.651376 255.8907
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | DiseaseStatus
##  Parameter estimates:
## Non-lesional     Lesional
##     1.000000     1.321159
## Fixed effects:  noPUVA ~ Age_at_sampling + MetaLocation
##                         Value Std.Error  DF   t-value p-value
## (Intercept)       -165.31689 130.49624 902 -1.266833  0.2055
## Age_at_sampling     14.56302   2.26278 100  6.435883  0.0000
## MetaLocationArm    132.64561 129.52608  92  1.024084  0.3085
## MetaLocationBack    -0.87028 123.25857  92 -0.007061  0.9944
## MetaLocationFlank  -40.06695 122.94705  92 -0.325888  0.7452
## MetaLocationLeg     -8.44252 125.40753  92 -0.067321  0.9465
##  Correlation:
##                   (Intr) Ag_t_s MtLctA MtLctB MtLctF
## Age_at_sampling   -0.582
## MetaLocationArm   -0.646 -0.035
## MetaLocationBack  -0.649 -0.087  0.711
## MetaLocationFlank -0.674 -0.049  0.728  0.776
## MetaLocationLeg   -0.646 -0.073  0.696  0.746  0.740
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -6.25854478 -0.25782381 -0.05141133  0.23651517  5.67628055
##
## Number of Observations: 1100
## Number of Groups:
##               PatientID BiopsyID %in% PatientID
##                     102                     198
```

```
noPUVA.ints <- intervals(model_noPUVA.null, which="fixed")$fixed
noPUVA.ints
```

```
##                         lower         est.     upper
## (Intercept)       -420.72904 -165.3168907  90.09525
## Age_at_sampling     10.08598   14.5630168  19.04006
## MetaLocationArm   -123.90183  132.6456056 389.19304
## MetaLocationBack  -245.00388   -0.8702803 243.26332
## MetaLocationFlank -283.58353  -40.0669454 203.44964
## MetaLocationLeg   -256.83249   -8.4425205 239.94745
## attr(,"label")
## [1] "Fixed effects:"
```

```
summary(model_noPUVA.dur)
```

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: clone_burden[!is.na(clone_burden$Disease_duration), ]
##        AIC      BIC    logLik
##   16189.09 16244.12 -8083.545
##
## Random effects:
##  Formula: ~Age_at_sampling - 1 | PatientID
##         Age_at_sampling
## StdDev:        7.568998
##
##  Formula: ~Age_at_sampling - 1 | BiopsyID %in% PatientID
##         Age_at_sampling Residual
## StdDev:        3.635017 255.9926
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | DiseaseStatus
##  Parameter estimates:
## Non-lesional     Lesional
##     1.000000     1.320618
## Fixed effects:  noPUVA ~ Age_at_sampling + MetaLocation + Disease_duration
##                         Value Std.Error  DF   t-value p-value
## (Intercept)       -164.79375 130.64694 902 -1.261367  0.2075
## Age_at_sampling     14.71062   2.29255 100  6.416705  0.0000
## MetaLocationArm    132.08956 129.66993  91  1.018660  0.3111
## MetaLocationBack     0.91712 123.46018  91  0.007428  0.9941
## MetaLocationFlank  -41.15383 123.10015  91 -0.334312  0.7389
## MetaLocationLeg     -8.35641 125.54401  91 -0.066562  0.9471
## Disease_duration    -0.60387   1.43085  91 -0.422036  0.6740
##  Correlation:
##                   (Intr) Ag_t_s MtLctA MtLctB MtLctF MtLctL
## Age_at_sampling   -0.574
## MetaLocationArm   -0.646 -0.036
## MetaLocationBack  -0.649 -0.081  0.710
## MetaLocationFlank -0.674 -0.051  0.729  0.775
## MetaLocationLeg   -0.646 -0.072  0.696  0.746  0.740
## Disease_duration  -0.009 -0.153  0.007 -0.035  0.022 -0.001
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -6.24342427 -0.25400058 -0.05201199  0.23939919  5.67895083
##
## Number of Observations: 1100
## Number of Groups:
##               PatientID BiopsyID %in% PatientID
##                     102                     198
```

```
intervals(model_noPUVA.dur, which="fixed")$fixed
```

```
##                         lower         est.      upper
## (Intercept)       -420.383954 -164.7937480  90.796458
## Age_at_sampling     10.176761   14.7106207  19.244481
## MetaLocationArm   -124.662960  132.0895617 388.842084
## MetaLocationBack  -243.539804    0.9171232 245.374050
## MetaLocationFlank -284.897878  -41.1538329 202.590212
## MetaLocationLeg   -256.939420   -8.3564142 240.226592
## Disease_duration    -3.437015   -0.6038699   2.229275
## attr(,"label")
## [1] "Fixed effects:"
```

```
anova(model_noPUVA.null,model_noPUVA.dur, test=T)$"p-value"[2]
```

```
## [1] 0.6732167
```

UV-associated mutation burden

UV-light is the dominant mutagen in the skin, accounting for 80% of the mutations in this dataset (and even more if PUVA isn't considered). We may be interested in knowing the rate at which UV-associated mutations accumulate in the skin. Fit the model using only the mutation burden attributed to UV-related signatures.

```
clone_burden$UV <- clone_burden$SBS7b + clone_burden$SBS7c
model_UV.null <- lme(fixed = UV ~ Age_at_sampling  + MetaLocation,
                     random = list(PatientID = pdSymm(form = ~ Age_at_sampling - 1), BiopsyID = pdSymm(form = ~
Age_at_sampling - 1)),
                     weights = varIdent(form= ~ 1 | DiseaseStatus),
                     data = clone_burden[!is.na(clone_burden$Disease_duration),], method="ML")

model_UV.dur <- lme(fixed = UV ~ Age_at_sampling  + MetaLocation + Disease_duration,
                     random = list(PatientID = pdSymm(form = ~ Age_at_sampling - 1), BiopsyID = pdSymm(form = ~
Age_at_sampling - 1)),
                     weights = varIdent(form= ~ 1 | DiseaseStatus),
                     data = clone_burden[!is.na(clone_burden$Disease_duration),], method="ML")

summary(model_UV.null)
```

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: clone_burden[!is.na(clone_burden$Disease_duration), ]
##       AIC      BIC    logLik
##   16147.1 16197.13 -8063.551
##
## Random effects:
##  Formula: ~Age_at_sampling - 1 | PatientID
##         Age_at_sampling
## StdDev:        7.647154
##
##  Formula: ~Age_at_sampling - 1 | BiopsyID %in% PatientID
##         Age_at_sampling Residual
## StdDev:        3.516487 252.6377
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | DiseaseStatus
##  Parameter estimates:
## Non-lesional     Lesional
##     1.000000     1.310479
## Fixed effects:  UV ~ Age_at_sampling + MetaLocation
##                        Value Std.Error  DF   t-value p-value
## (Intercept)       -175.84685 130.69328 902 -1.345493  0.1788
## Age_at_sampling     13.62039   2.27119 100  5.997037  0.0000
## MetaLocationArm    132.80864 129.69847  92  1.023980  0.3085
## MetaLocationBack     6.22690 123.36307  92  0.050476  0.9599
## MetaLocationFlank  -40.40738 123.03066  92 -0.328433  0.7433
## MetaLocationLeg     -1.47761 125.60020  92 -0.011764  0.9906
##  Correlation:
##                   (Intr) Ag_t_s MtLctA MtLctB MtLctF
## Age_at_sampling   -0.582
## MetaLocationArm   -0.646 -0.036
## MetaLocationBack  -0.650 -0.088  0.712
## MetaLocationFlank -0.674 -0.050  0.731  0.780
## MetaLocationLeg   -0.646 -0.074  0.697  0.747  0.741
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -6.42862268 -0.25457082 -0.05763308  0.22646220  5.71098792
##
## Number of Observations: 1100
## Number of Groups:
##               PatientID BiopsyID %in% PatientID
##                     102                     198
```

```
summary(model_UV.dur)
```

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: clone_burden[!is.na(clone_burden$Disease_duration), ]
##        AIC      BIC    logLik
##   16148.78 16203.81 -8063.388
##
## Random effects:
##  Formula: ~Age_at_sampling - 1 | PatientID
##         Age_at_sampling
## StdDev:        7.663792
##
##  Formula: ~Age_at_sampling - 1 | BiopsyID %in% PatientID
##         Age_at_sampling Residual
## StdDev:        3.491846 252.7528
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | DiseaseStatus
##  Parameter estimates:
## Non-lesional     Lesional
##     1.000000     1.309867
## Fixed effects:  UV ~ Age_at_sampling + MetaLocation + Disease_duration
##                         Value Std.Error  DF   t-value p-value
## (Intercept)       -175.13082 130.87983 902 -1.338104  0.1812
## Age_at_sampling     13.81551   2.30035 100  6.005818  0.0000
## MetaLocationArm    131.97649 129.87536  91  1.016178  0.3122
## MetaLocationBack     8.64571 123.59234  91  0.069953  0.9444
## MetaLocationFlank  -41.87075 123.20960  91 -0.339834  0.7348
## MetaLocationLeg     -1.31038 125.76835  91 -0.010419  0.9917
## Disease_duration    -0.79604   1.39212  91 -0.571817  0.5689
##  Correlation:
##                   (Intr) Ag_t_s MtLctA MtLctB MtLctF MtLctL
## Age_at_sampling   -0.574
## MetaLocationArm   -0.646 -0.037
## MetaLocationBack  -0.649 -0.081  0.711
## MetaLocationFlank -0.674 -0.053  0.731  0.779
## MetaLocationLeg   -0.645 -0.073  0.697  0.747  0.741
## Disease_duration  -0.008 -0.149  0.008 -0.035  0.022 -0.002
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -6.4086116 -0.2510775 -0.0546847  0.2264474  5.7143715
##
## Number of Observations: 1100
## Number of Groups:
##               PatientID BiopsyID %in% PatientID
##                     102                     198
```

```
anova(model_UV.null,model_UV.dur, test=T)$"p-value"[2]
```

```
## [1] 0.5681442
```

```
UV.ints <- intervals(model_UV.null, which="fixed")$fixed
UV.ints
```

```
##                         lower        est.     upper
## (Intercept)       -431.644647 -175.846849  79.95095
## Age_at_sampling      9.126728   13.620392  18.11406
## MetaLocationArm   -124.080231  132.808635 389.69750
## MetaLocationBack  -238.113688    6.226904 250.56750
## MetaLocationFlank -284.089560  -40.407377 203.27481
## MetaLocationLeg   -250.249192   -1.477608 247.29398
## attr(,"label")
## [1] "Fixed effects:"
```

Again there is no significant effect of disease duration in this model.

SBS1/5 - associated mutation burden

The mutational signatures SBS1 and SBS5 are found in all normal cells at varying frequencies. They accumulate linearly with age but are accelerated in some inflamed tissues, including colonic mucosa affected by inflammatory bowel disease (see https://doi.org/10.1016/j.cell.2020.06.036 (https://doi.org/10.1016/j.cell.2020.06.036)). UV-exposure adds a lot of variance to the dataset and may mask any potential effects of psoriasis on the mutation burden. We should test if there is an effect of disease duration on the SBS1/5 mutation burden.

```
model_clock.null <- lme(fixed = SBS1.5 ~ Age_at_sampling  + MetaLocation,
                      random = list(PatientID = pdSymm(form = ~ Age_at_sampling - 1), BiopsyID = pdSymm(form = ~
Age_at_sampling - 1)),
                      weights = varIdent(form= ~ 1 | DiseaseStatus),
                      data = clone_burden[!is.na(clone_burden$Disease_duration),], method="ML")

model_clock.dur <- lme(fixed = SBS1.5 ~ Age_at_sampling  + MetaLocation + Disease_duration,
                      random = list(PatientID = pdSymm(form = ~ Age_at_sampling - 1), BiopsyID = pdSymm(form = ~
Age_at_sampling - 1)),
                      weights = varIdent(form= ~ 1 | DiseaseStatus),
                      data = clone_burden[!is.na(clone_burden$Disease_duration),], method="ML")

summary(model_clock.null)
```

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: clone_burden[!is.na(clone_burden$Disease_duration), ]
##        AIC      BIC    logLik
##   9229.044 9279.075 -4604.522
##
## Random effects:
##  Formula: ~Age_at_sampling - 1 | PatientID
##         Age_at_sampling
## StdDev:        0.239352
##
##  Formula: ~Age_at_sampling - 1 | BiopsyID %in% PatientID
##         Age_at_sampling Residual
## StdDev:       0.1796447  11.3497
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | DiseaseStatus
##  Parameter estimates:
## Non-lesional     Lesional
##     1.000000     1.267614
## Fixed effects:  SBS1.5 ~ Age_at_sampling + MetaLocation
##                         Value Std.Error  DF   t-value p-value
## (Intercept)         4.076509  4.704412 902  0.866529  0.3864
## Age_at_sampling     0.690176  0.079826 100  8.646062  0.0000
## MetaLocationArm   -11.719469  4.672395  92 -2.508236  0.0139
## MetaLocationBack   -6.033234  4.457973  92 -1.353358  0.1793
## MetaLocationFlank  -5.090121  4.454521  92 -1.142687  0.2561
## MetaLocationLeg    -8.356679  4.511533  92 -1.852292  0.0672
##  Correlation:
##                   (Intr) Ag_t_s MtLctA MtLctB MtLctF
## Age_at_sampling   -0.584
## MetaLocationArm   -0.647 -0.028
## MetaLocationBack  -0.646 -0.084  0.703
## MetaLocationFlank -0.675 -0.036  0.712  0.753
## MetaLocationLeg   -0.649 -0.065  0.694  0.736  0.731
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -5.8550103 -0.4753986 -0.1080387  0.4219794  4.2201898
##
## Number of Observations: 1100
## Number of Groups:
##               PatientID BiopsyID %in% PatientID
##                     102                     198
```

```
summary(model_clock.dur)
```

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: clone_burden[!is.na(clone_burden$Disease_duration), ]
##        AIC      BIC    logLik
##   9224.689 9279.723 -4601.345
##
## Random effects:
##  Formula: ~Age_at_sampling - 1 | PatientID
##         Age_at_sampling
## StdDev:       0.2448763
##
##  Formula: ~Age_at_sampling - 1 | BiopsyID %in% PatientID
##         Age_at_sampling Residual
## StdDev:       0.1695767 11.28844
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | DiseaseStatus
##  Parameter estimates:
## Non-lesional     Lesional
##      1.00000      1.27481
## Fixed effects:  SBS1.5 ~ Age_at_sampling + MetaLocation + Disease_duration
##                         Value Std.Error  DF   t-value p-value
## (Intercept)         3.886816  4.716879 902  0.824023  0.4101
## Age_at_sampling     0.651671  0.081629 100  7.983308  0.0000
## MetaLocationArm   -11.645951  4.683731  91 -2.486469  0.0147
## MetaLocationBack   -6.402939  4.468874  91 -1.432786  0.1553
## MetaLocationFlank  -4.881964  4.462627  91 -1.093967  0.2769
## MetaLocationLeg    -8.267194  4.522802  91 -1.827892  0.0708
## Disease_duration    0.163104  0.063425  91  2.571609  0.0117
##  Correlation:
##                   (Intr) Ag_t_s MtLctA MtLctB MtLctF MtLctL
## Age_at_sampling   -0.571
## MetaLocationArm   -0.647 -0.027
## MetaLocationBack  -0.646 -0.076  0.703
## MetaLocationFlank -0.675 -0.040  0.714  0.755
## MetaLocationLeg   -0.649 -0.066  0.694  0.737  0.732
## Disease_duration  -0.011 -0.190 -0.001 -0.035  0.020  0.004
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -5.8606464 -0.4886593 -0.1038414  0.4213027  4.2828992
##
## Number of Observations: 1100
## Number of Groups:
##               PatientID BiopsyID %in% PatientID
##                     102                     198
```

```
anova(model_clock.null,model_clock.dur, test=T)$"p-value"[2]
```

```
## [1] 0.01170614
```

```
clock.ints <- intervals(model_clock.dur, which="fixed")$fixed
clock.ints
```

```
##                          lower        est.      upper
## (Intercept)        -5.34101647   3.8868160 13.1146484
## Age_at_sampling     0.49023717   0.6516711  0.8131050
## MetaLocationArm   -20.91995779 -11.6459507 -2.3719435
## MetaLocationBack  -15.25151883  -6.4029389  2.4456410
## MetaLocationFlank -13.71817347  -4.8819642  3.9542451
## MetaLocationLeg   -17.22255446  -8.2671942  0.6881660
## Disease_duration    0.03751981   0.1631041  0.2886884
## attr(,"label")
## [1] "Fixed effects:"
```

```
summary(model_clock.dur)$tTable[,"p-value"]
```

```
##       (Intercept)   Age_at_sampling   MetaLocationArm  MetaLocationBack
##      4.101444e-01      2.468629e-12      1.472594e-02      1.553456e-01
## MetaLocationFlank   MetaLocationLeg  Disease_duration
##      2.768570e-01      7.084266e-02      1.174378e-02
```

When restricting to the mutation burden attributed to SBS1/5, the disease duration effect is (borderline) significant.

Pruning the phylogenetic trees

Some of the mutation clusters consisted of groups of mutations with VAFs too low for the pigeonhole principle to be incontrovertible. The calculations above assume that in such cases, the mutations all derive from a single sub-clone. However, there is a risk that the mutation burden represents not the burden of a single clone but the sum of the mutation burden for a collection of clones with similar cell fractions across all microbiopsies. This would lead to an over-estimation of the mutation rate for terminal branches of the phylogenetic trees. We performed pruning of the phylogenetic trees, retaining only branches representing nested clusters if the sum of the VAFs was greater than 1. For branches that represent single clusters (with no nesting), we pruned branches with VAF<0.3.

Unsurprisingly, this lowers the estimation of the total mutation rate. This new value should be thought of as a conservative lower bound.

```
## This file can be found in the Mendeley repository accompanying the manuscript.
clone_burden_after_pruning <- read.table("/nfs/users/nfs_s/so11/phd/psoriasis/bsub_jupyter_lab/psoriasis/manuscri
pt_data_and_figures/Supplementary_material/clone_mutation_burden_maxVAF03.txt", h=T)

clone_burden <- clone_burden_after_pruning
clone_burden <- merge(clone_burden, microd_meta[,c("SampleID", "BiopsyID", "MetaLocation","PatientID", "DiseaseSt
atus")], by.x="HighCellFrac_sample", by.y="SampleID")
clone_burden <- merge(clone_burden, patient_meta[,c("Patient.ID", "Age_at_sampling", "Disease_duration","Sex")],
by.x="PatientID", by.y="Patient.ID")

clone_burden$Disease_duration[clone_burden$DiseaseStatus=="Non-lesional" & !is.na(clone_burden$Disease_duration)]
<- 0

clone_burden$noPUVA <- clone_burden$TotalSBS_adj - clone_burden$PUVA

model_noPUVA.null <- lme(fixed = noPUVA ~ Age_at_sampling  + MetaLocation,
                      random = list(PatientID = pdSymm(form = ~ Age_at_sampling - 1), BiopsyID = pdSymm(form = ~
Age_at_sampling - 1)),
                      weights = varIdent(form= ~ 1 | DiseaseStatus),
                      data = clone_burden[!is.na(clone_burden$Disease_duration),], method="ML")

model_noPUVA.dur <- lme(fixed = noPUVA ~ Age_at_sampling  + MetaLocation + Disease_duration,
                      random = list(PatientID = pdSymm(form = ~ Age_at_sampling - 1), BiopsyID = pdSymm(form = ~
Age_at_sampling - 1)),
                      weights = varIdent(form= ~ 1 | DiseaseStatus),
                      data = clone_burden[!is.na(clone_burden$Disease_duration),], method="ML")

summary(model_noPUVA.null)
```

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: clone_burden[!is.na(clone_burden$Disease_duration), ]
##        AIC      BIC    logLik
##   11993.71 12040.99 -5986.854
##
## Random effects:
##  Formula: ~Age_at_sampling - 1 | PatientID
##         Age_at_sampling
## StdDev:        4.592135
##
##  Formula: ~Age_at_sampling - 1 | BiopsyID %in% PatientID
##         Age_at_sampling Residual
## StdDev:        3.333375  238.434
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | DiseaseStatus
##  Parameter estimates:
## Non-lesional     Lesional
##      1.00000      1.14779
## Fixed effects:  noPUVA ~ Age_at_sampling + MetaLocation
##                       Value Std.Error  DF   t-value p-value
## (Intercept)       -99.36292  92.37722 642 -1.075621  0.2825
## Age_at_sampling     9.59962   1.57251  99  6.104640  0.0000
## MetaLocationArm   137.78163  91.09434  89  1.512516  0.1339
## MetaLocationBack   19.33995  87.07279  89  0.222112  0.8247
## MetaLocationFlank  -0.35792  87.28331  89 -0.004101  0.9967
## MetaLocationLeg    18.40436  89.15607  89  0.206429  0.8369
##  Correlation:
##                   (Intr) Ag_t_s MtLctA MtLctB MtLctF
## Age_at_sampling   -0.589
## MetaLocationArm   -0.648 -0.023
## MetaLocationBack  -0.641 -0.087  0.705
## MetaLocationFlank -0.672 -0.032  0.709  0.751
## MetaLocationLeg   -0.633 -0.073  0.687  0.729  0.718
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -4.86300912 -0.32063767 -0.08139216  0.23407789  7.66261854
##
## Number of Observations: 836
## Number of Groups:
##               PatientID BiopsyID %in% PatientID
##                     101                     194
```

```
noPUVA.ints <- intervals(model_noPUVA.null, which="fixed")$fixed
noPUVA.ints
```

```
##                        lower        est.     upper
## (Intercept)       -280.108801 -99.3629158  81.38297
## Age_at_sampling      6.490635   9.5996244  12.70861
## MetaLocationArm    -42.570176 137.7816330 318.13344
## MetaLocationBack  -153.049864  19.3399461 191.72976
## MetaLocationFlank -173.164513  -0.3579178 172.44868
## MetaLocationLeg   -158.109994  18.4043631 194.91872
## attr(,"label")
## [1] "Fixed effects:"
```

We should look at what effect the pruning of the trees has on the disease duration estimate. We see that the disease duration effect is much diminished and is no longer significant.

```
model_clock.null <- lme(fixed = SBS1.5 ~ Age_at_sampling  + MetaLocation,
                     random = list(PatientID = pdSymm(form = ~ Age_at_sampling - 1), BiopsyID = pdSymm(form = ~
Age_at_sampling - 1)),
                     weights = varIdent(form= ~ 1 | DiseaseStatus),
                     data = clone_burden[!is.na(clone_burden$Disease_duration),], method="ML")

model_clock.dur <- lme(fixed = SBS1.5 ~ Age_at_sampling  + MetaLocation + Disease_duration,
                     random = list(PatientID = pdSymm(form = ~ Age_at_sampling - 1), BiopsyID = pdSymm(form = ~
Age_at_sampling - 1)),
                     weights = varIdent(form= ~ 1 | DiseaseStatus),
                     data = clone_burden[!is.na(clone_burden$Disease_duration),], method="ML")

summary(model_clock.null)
```

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: clone_burden[!is.na(clone_burden$Disease_duration), ]
##        AIC      BIC    logLik
##   6939.179 6986.465 -3459.589
##
## Random effects:
##  Formula: ~Age_at_sampling - 1 | PatientID
##         Age_at_sampling
## StdDev:       0.2075022
##
##  Formula: ~Age_at_sampling - 1 | BiopsyID %in% PatientID
##         Age_at_sampling Residual
## StdDev:       0.1145642 12.58384
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | DiseaseStatus
##  Parameter estimates:
## Non-lesional     Lesional
##     1.000000     1.079586
## Fixed effects:  SBS1.5 ~ Age_at_sampling + MetaLocation
##                       Value Std.Error  DF   t-value p-value
## (Intercept)        4.199626  4.148844 642  1.012240  0.3118
## Age_at_sampling    0.526600  0.069892  99  7.534455  0.0000
## MetaLocationArm   -8.566169  4.085368  89 -2.096792  0.0388
## MetaLocationBack  -4.841632  3.889280  89 -1.244866  0.2164
## MetaLocationFlank -2.593310  3.902734  89 -0.664485  0.5081
## MetaLocationLeg   -7.317099  3.994117  89 -1.831969  0.0703
##  Correlation:
##                   (Intr) Ag_t_s MtLctA MtLctB MtLctF
## Age_at_sampling   -0.593
## MetaLocationArm   -0.647 -0.021
## MetaLocationBack  -0.641 -0.086  0.705
## MetaLocationFlank -0.675 -0.025  0.708  0.755
## MetaLocationLeg   -0.630 -0.074  0.686  0.731  0.719
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -3.8881963 -0.5015500 -0.1786659  0.4588029  5.2659198
##
## Number of Observations: 836
## Number of Groups:
##               PatientID BiopsyID %in% PatientID
##                     101                     194
```

```
summary(model_clock.dur)
```

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: clone_burden[!is.na(clone_burden$Disease_duration), ]
##        AIC     BIC    logLik
##   6939.315 6991.33 -3458.658
##
## Random effects:
##  Formula: ~Age_at_sampling - 1 | PatientID
##         Age_at_sampling
## StdDev:       0.2078412
##
##  Formula: ~Age_at_sampling - 1 | BiopsyID %in% PatientID
##         Age_at_sampling Residual
## StdDev:       0.1116281 12.57851
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | DiseaseStatus
##  Parameter estimates:
## Non-lesional     Lesional
##     1.000000     1.080191
## Fixed effects:  SBS1.5 ~ Age_at_sampling + MetaLocation + Disease_duration
##                       Value Std.Error  DF   t-value p-value
## (Intercept)        4.109724  4.143632 642  0.991817  0.3217
## Age_at_sampling    0.507873  0.071171  99  7.135978  0.0000
## MetaLocationArm   -8.611526  4.080120  88 -2.110606  0.0376
## MetaLocationBack  -5.092668  3.887480  88 -1.310018  0.1936
## MetaLocationFlank -2.505326  3.897091  88 -0.642871  0.5220
## MetaLocationLeg   -7.239487  3.988878  88 -1.814918  0.0729
## Disease_duration   0.074812  0.054693  88  1.367868  0.1748
##  Correlation:
##                   (Intr) Ag_t_s MtLctA MtLctB MtLctF MtLctL
## Age_at_sampling   -0.579
## MetaLocationArm   -0.646 -0.018
## MetaLocationBack  -0.640 -0.075  0.705
## MetaLocationFlank -0.675 -0.028  0.708  0.754
## MetaLocationLeg   -0.630 -0.075  0.686  0.730  0.719
## Disease_duration  -0.013 -0.196 -0.011 -0.048  0.015  0.013
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -3.9160477 -0.4896735 -0.1751509  0.4563460  5.2779813
##
## Number of Observations: 836
## Number of Groups:
##               PatientID BiopsyID %in% PatientID
##                     101                     194
```

```
anova(model_clock.null,model_clock.dur, test=T)$"p-value"[2]
```

```
## [1] 0.1721982
```

```
clock.ints <- intervals(model_clock.dur, which="fixed")$fixed
clock.ints
```

```
##                         lower        est.      upper
## (Intercept)        -3.99284822  4.10972401 12.2122962
## Age_at_sampling     0.36724705  0.50787275  0.6484984
## MetaLocationArm   -16.68588828 -8.61152580 -0.5371633
## MetaLocationBack  -12.78580585 -5.09266842  2.6004690
## MetaLocationFlank -10.21748447 -2.50532616  5.2068321
## MetaLocationLeg   -15.13328578 -7.23948654  0.6543127
## Disease_duration   -0.03342182  0.07481223  0.1830463
## attr(,"label")
## [1] "Fixed effects:"
```

```
summary(model_clock.dur)$tTable[,"p-value"]
```

```
##       (Intercept)   Age_at_sampling   MetaLocationArm  MetaLocationBack
##      3.216605e-01      1.614682e-10      3.764355e-02      1.935991e-01
## MetaLocationFlank   MetaLocationLeg  Disease_duration
##      5.219795e-01      7.294338e-02      1.748365e-01
```