Analysis of heuristically identified somatic *de novo* mutations"

Arslan Zaidi

4/12/2019

```
library(data.table)
library(dplyr)
library(ggplot2)
library(cowplot)
```

Here, I heuristically identified putative *de novo* mutations that may have arisen in the **somatic** tissues of an individual and tested whether they increase in number with age of the individual.

First, load the nucleotide counts for all identified heteroplasmies from all family members.

```
#read conservative table of heteroplasmies and their counts in all family members
allfam<-fread("~/Documents/mtproj_files/M2_new/files/Analysis/
              Fixing_heteroplasmy_table/
              hqcounts_cleaned_conservative_09272018.txt",sep="\t",header=T)
```

To find potential de novo somatic mutations, we first identify sites which are heteroplasmic in only sample of the entire family.

Tabulate number of samples that are heteroplasmic in the family

```
hq.1t.fam<-allfam%>%
  group_by(FID,position)%>%
  summarize(nhets=length(which(maf>0.01)))%>%
  filter(.,nhets==1)%>%
  mutate(fam_het_id=paste(FID,position,sep="_"))

#153 cases
```

Identify these people and get the frequency data for these sites from other family members.

```
#who are these people
ind.1t<-allfam%>%
  filter(fam_het_id%in%hq.1t.fam$fam_het_id)%>%
  filter(maf>0.01)%>%
  mutate(individual_het_id=paste(individual_id,position,sep="_"))

#make table of frequencies from all family members for these sites
allfam.denovo<-allfam%>%
  filter(fam_het_id%in%hq.1t.fam$fam_het_id)%>%
  mutate(individual_het_id=paste(individual_id,position,sep="_"))
```

Check if the same site is heteroplasmic in the other samples (MAF>0.002) from the same family. We would like to exclude such sites if they are to reduce false-positives.

```
hq.denovo<-allfam.denovo%>%
  filter(fam_het_id%in%allfam.denovo$fam_het_id)%>%
  group_by(FID,fam_het_id,position)%>%
  summarize(nhets.2p=length(which(maf>0.002)))%>%
  filter(nhets.2p==1)
```

```
#57 potentially somatic denovo mutations
```

This leaves us with 57 putative somatic *de novo* mutations. Use this table to calculate the number of such mutations for each individual so we can test whether this number increases with age.

```
# #isolate the individual's frequency data
ind.denovo<-allfam.denovo%>%
  filter(fam_het_id%in%hq.denovo$fam_het_id)%>%
  filter(maf>0.01)

#calculate no. of denovo somatic mutations in each individual
hq.ndenovo.per.ind<-ind.denovo%>%
  group_by(individual_id,tissue)%>%
  summarize(nhets=length(unique(position)),
            age_collection=mean(age_collection))%>%
  mutate(age_collection=age_collection/365)
```

Split the data by tissue type (blood/cheek).

```
#split by tissue
hq.ndenovo.per.ind.bl<-hq.ndenovo.per.ind%>%
  filter(tissue=="bl")

hq.ndenovo.per.ind.ch<-hq.ndenovo.per.ind%>%
  filter(tissue=="ch")
```

Add 0s for remaining individuals - who don't carry any *de novo* heteroplasmies. These will be included in the regression.

```
#add individuals with no mutations whatsoever - which are also not present in the family
#load complete family file
famfile<-fread("~/Documents/mtproj_files/M2_new/files/Analysis/
               Fixing_heteroplasmy_table/
               famfile_cleared_conservative_09272018.txt",header=T,sep="\t")
```

```
"%ni%"<-Negate("%in%")

#assign 0 heteroplasmies to individuals with no denovo mutations
hq.nodenovo.bl<-famfile%>%
  filter(individual_id%ni%hq.ndenovo.per.ind.bl$individual_id)%>%
  select(individual_id,age_collection)%>%
  mutate(tissue="bl")%>%
  group_by(individual_id,tissue)%>%
  summarize(nhets=0,age_collection=mean(age_collection))%>%
  mutate(age_collection=age_collection/365)

#assign 0 heteroplasmies to individuals with no denovo mutations
hq.nodenovo.ch<-famfile%>%
  filter(individual_id%ni%hq.ndenovo.per.ind.ch$individual_id)%>%
  select(individual_id,age_collection)%>%
  mutate(tissue="ch")%>%
  group_by(individual_id,tissue)%>%
  summarize(nhets=0,age_collection=mean(age_collection))%>%
  mutate(age_collection=age_collection/365)

#rbind bl
```

```
hq.ndenovo.per.ind2.bl<-rbind(hq.ndenovo.per.ind.bl,hq.nodenovo.bl)

#rbind ch
hq.ndenovo.per.ind2.ch<-rbind(hq.ndenovo.per.ind.ch,hq.nodenovo.ch)
```

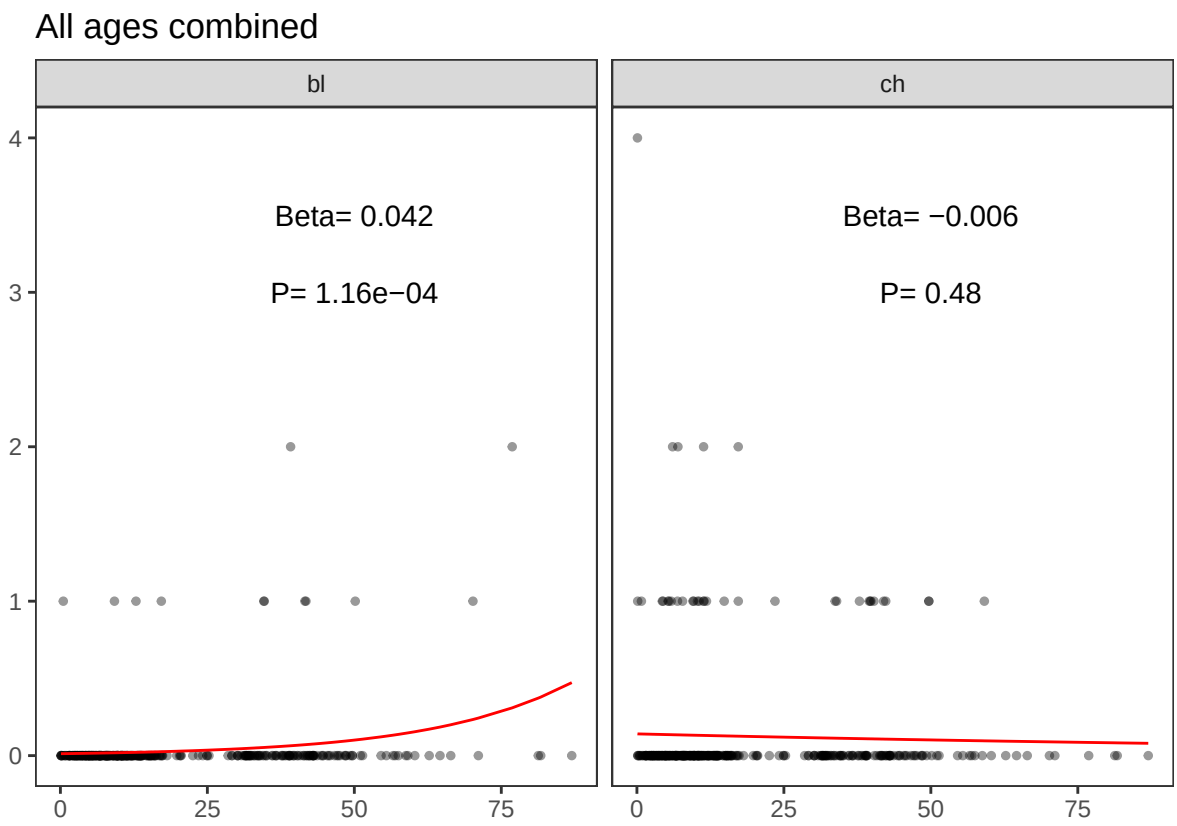
Fit a Poisson regression (number of somatic *de novo* mutations ~ age at collection) and extract regression coefficients.

```
#now separately for each tissue
#blood
hq.age.c2.bl<-glm(data=hq.ndenovo.per.ind2.bl,
                  nhets~age_collection,
                  family="poisson")
hq.ndenovo.per.ind2.bl$pred.nhets=predict(hq.age.c2.bl,type="response")


#cheek
hq.age.c2.ch<-glm(data=hq.ndenovo.per.ind2.ch,
                  nhets~age_collection,
                  family="poisson")

hq.ndenovo.per.ind2.ch$pred.nhets=predict(hq.age.c2.ch,type="response")


hq.ndenovo.per.ind2<-rbind(hq.ndenovo.per.ind2.bl,
                           hq.ndenovo.per.ind2.ch)

bl.beta2<-summary(hq.age.c2.bl)$coefficients[2]
ch.beta2<-summary(hq.age.c2.ch)$coefficients[2]

bl.p2<-summary(hq.age.c2.bl)$coefficients[8]
ch.p2<-summary(hq.age.c2.ch)$coefficients[8]


#make data.frame for glm results
glm.res2<-data.frame(tissue=c("bl","ch"),
                     x=c(50,50),
                     y=c(3.5,3.5),
                     betas=c("Beta= 0.042","Beta= -0.006"),
                     pvalue=c("P= 1.16e-04","P= 0.48"))
```

Plot the relationship between number of *de novo* somatic mutations and age at collection.

```
#plot relationship between nhets and age at collection
plt.nhets.agec2<-ggplot(hq.ndenovo.per.ind2,aes(age_collection,nhets))+
  geom_point(size=1,alpha=0.4)+
  geom_line(aes(x=age_collection,pred.nhets),color="red")+
  theme_bw()+
  theme(axis.title.y=element_blank(),
        plot.margin = margin(0.2,0.1,0.5,1,"cm"),
        panel.grid = element_blank(),
        panel.background = element_blank(),
        axis.title.x=element_blank())+
  facet_wrap(~tissue)+
  labs(x="Age at collection of individual",
```

```
      title="All ages combined")+
  geom_text(data=glm.res2,aes(x=x,y=y,label=betas))+
  geom_text(data=glm.res2,aes(x=x,y=y-0.5,label=pvalue))

plt.nhets.agec2
```



It appears that the number of somatic mutations increases with age in the blood tissue but not in cheek. We use the Poisson regression coefficients to estimate how many such mutations a person might accmuluate over their lifetime (0-80 years of age)

$$N_{0-80} = exp(\beta_{intercept}).exp(\beta_{age}.80) = exp(-4.39).exp(0.042.80)$$

This amounts to a total of 0.357 new mutations in the blood between the ages of 0 and 80, which is not a large number.