Analysis of heuristically identified germline de novo mutations

Arslan Zaidi

4/12/2019

Introduction

Here, we heuristically identified putative de novo mutations that may have arisen in the germline and tested whether they increase with age of the mother.

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

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

#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)
```

Heteroplasmies shared by both tissues are more likely to be germline

To find putative de novo germline mutations, we we first identify samples who are heteroplasmic for both tissues but no one in their family shares that heteroplasmy.

To do this, we first tabulate the number of samples that are heteroplasmic in the family. Then, select families where this number is 2.

```
hq.2t.fam<-allfam%>%
  group_by(FID,position)%>%
  summarize(nhets=length(which(maf>0.01)))%>%
  filter(.,nhets==2)%>%
  mutate(fam_het_id=paste(FID,position,sep="_"))
```

Subset cases in which both of these heteroplasmies are present in the same individual. This would only retain cases where no one else in the family has a heteroplasmy at MAF of 0.01.

```
hq.2t.ind<-allfam%>%
  filter(fam_het_id%in%hq.2t.fam$fam_het_id)%>%
  group_by(FID,fam_het_id,fam_cat,individual_id,level,position)%>%
  summarize(nhets=length(which(maf>0.01)))%>%
  filter(nhets==2)
```

Select cases only where these are present in 'kids'. This is to remove cases where a mother was heteroplasmic in her germline but the heteroplasmy was lost in the kids. Also, to exclude cases where we don't have data from the previous generation.

```
hq.denovo<-hq.2t.ind%>%
  filter(
    fam_cat%in%c("g1m1c2","tg1m1c1c2g2","g1m1c2m2c0","g1m1c3") & level=="m1"|
    fam_cat%in%c("m1c2","m1c3","m1c4","m1c5") & level%in%c("c1","c2","c3","c4","c5"))%>%
```

```
  mutate(individual_het_id=paste(individual_id,position,sep="_"))

#103 cases

#extract data for sites which are denovo from all family members
allfam.denovo<-allfam%>%
  filter(fam_het_id%in%hq.denovo$fam_het_id)%>%
  mutate(individual_het_id=paste(individual_id,position,sep="_"))
```

Heteroplasmies shared by other family members less likely to be *de novo*

Count number of individuals in each family who are not the "proband" and have heteroplasmy at that site less than 0.2%. This is to reduce the probability of false-positives.

```
hq.denovo.red<-allfam.denovo%>%
  group_by(FID,position)%>%
  summarize(
    nhets.2p = length(
      setdiff(which(maf>0.002),
              which(individual_het_id%in%hq.denovo$individual_het_id)
              )))%>%
  mutate(fam_het_id=paste(FID,position,sep="_"))%>%
  filter(nhets.2p==0)



#78 cases

#isolate these cases - frequency info for all family members
allfam.denovo.red<-allfam%>%
  filter(fam_het_id%in%hq.denovo.red$fam_het_id)%>%
  mutate(individual_het_id=paste(individual_id,position,sep="_"))

#output these heteroplasmies for supplement
hq.ind.denovo<-allfam.denovo.red%>%
  filter(maf>0.01)%>%
  select(FID,mother_id,individual_id,level,tissue,position,maf)

# fwrite(hq.ind.denovo,"files/Analysis/Denovo_mutations/
#hq_denovo_mutations_03202019.txt",
#sep="\t",col.names=T,row.names=F,quote=F)
```

Count the number of de novo heteroplasmies per individual

```
hq.ndenovo.per.ind<-allfam.denovo.red%>%
  filter(maf>0.01)%>%
  group_by(FID,level,individual_id,age_collection,age_birth)%>%
  summarise(ndenovo.hets=length(unique(position)))%>%
  group_by(FID,level,individual_id,age_collection,age_birth)%>%
  summarise(nhets=sum(ndenovo.hets))%>%
  mutate(age.c=age_collection/365,age.b=age_birth/365)
```

Now add 0s - the kids who don't have any denovo mutations - for the regression.

```
#for this, read in the whole famfile
famfile<-fread("~/Documents/mtproj_files/M2_new/files/Analysis/
               Fixing_heteroplasmy_table/
               famfile_cleared_conservative_09272018.txt",header=T,sep="\t")
```

```
#"not in" function
"%ni%"<-Negate("%in%")

allfam.nodenovo<-famfile[
  which(
    (famfile$individual_id%ni%hq.ndenovo.per.ind$individual_id &
       famfile$fam_cat%in%c("g1m2","g1m1c2","g1m1c2m2c0","g1m1c3") &
       famfile$level%in%c("m1","m2","c1","c2","c3")) |
      (famfile$individual_id%ni%hq.ndenovo.per.ind$individual_id &
         famfile$fam_cat%in%c("tg1m1c3","tg1m1c1c2g2") &
         famfile$level%in%c("g1","g2","m1","c1","c2","c3")) |
      (famfile$individual_id%ni%hq.ndenovo.per.ind$individual_id &
         famfile$fam_cat%in%c("m1c2","m1c3","m1c4","m1c5") &
         famfile$level%in%c("c1","c2","c3","c4","c5"))),]


nodenovo.per.ind<-allfam.nodenovo%>%
  group_by(FID,level,individual_id,age_collection,age_birth)%>%
  summarize(nhets=0)%>%
  mutate(age.c=age_collection/365,age.b=age_birth/365)


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

hq.ndenovo.per.ind<-hq.ndenovo.per.ind[order(hq.ndenovo.per.ind$individual_id),]
```

Number of putative *de novo* heteroplasmies ~ age of mother

Carry out Poisson regression of the number of heuristically identified de novo heteroplasmies on the age of the mother at child birth. This will tell us if older mothers pass on more germline mutations than younger mothers.

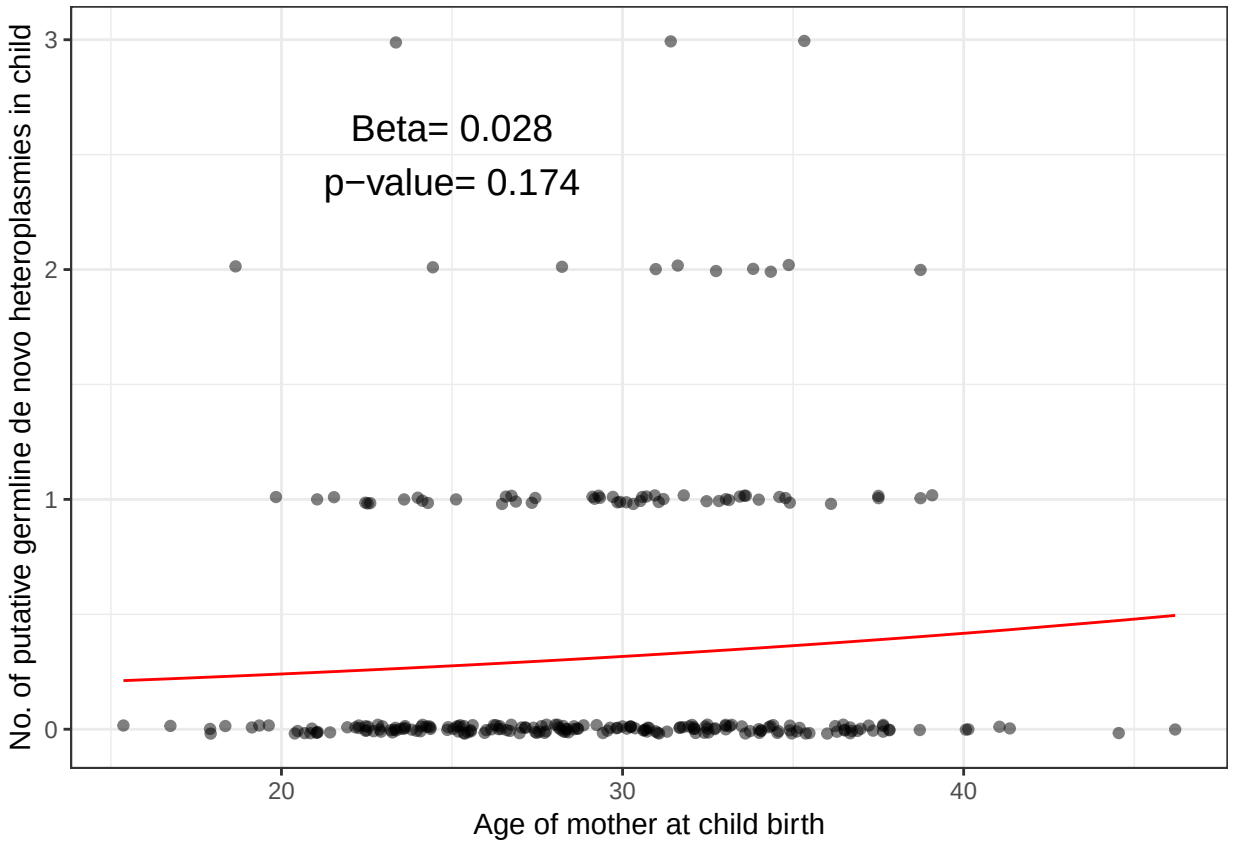
```
#poisson regression of nhets ~ age at birth
pred.age.b<-hq.ndenovo.per.ind%>%
  filter(is.na(age.b)=="FALSE")
hq.age.b<-glm(data=pred.age.b,nhets~age.b,family="poisson")
pred.age.b$pred.nhets=predict(hq.age.b,type="response")

beta=round(summary(hq.age.b)$coeff[2,1],3)
pval=round(summary(hq.age.b)$coeff[2,4],3)

#plot this
#plot relationship between nhets and age at birth
plt.nhets.ageb<-ggplot(hq.ndenovo.per.ind,aes(age.b,nhets))+
  geom_point(alpha=0.5,position=position_jitter(height=0.02,width=NULL))+
  geom_line(data=pred.age.b,aes(x=age.b,pred.nhets),color="red")+
  theme_bw()+
  labs(x="Age of mother at child birth",
```

```
      y="No. of putative germline de novo heteroplasmies in child")+
  annotate(geom="text",x=25,y=2.5,
           label=paste("Beta= ",beta,"\np-value= ",pval,sep=""),size=5)

plt.nhets.ageb
```



So, even though the number of putative de novo germline mutations increases with the age of the individual at collection, this relationship is not statistically significant at the 0.05 level.

$$N_{0-40} = exp(\beta_{intercept}).exp(\beta_{age}.80) = exp(-1.978).exp(0.0276 \times 40)$$

This amounts to ~0.42 new mutations arising in the germline between the ages of 0 and 40, which is not a large number.