```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
numberFamilies <- 10000
load(str_interp("../RObjects/summary_tables/summaryTable${numberFamilies}Families.Rdata"))
```

Visualize the carrier risk probabilities

when visualizing you should color by affected probands and unaffected probands

post on basecamp tonight

##this should be a covariate in my regression

##the bias is stronger for bias with very little family history but as the risk increases the bias is not as conspicuous

##population level is what we have with 0.05

##another analysis is for high risk clinics where we take a subset of the generated families and only take the families that have one affected relative or more and then refit the adjustment and compare across the two groups

see if the models are similar

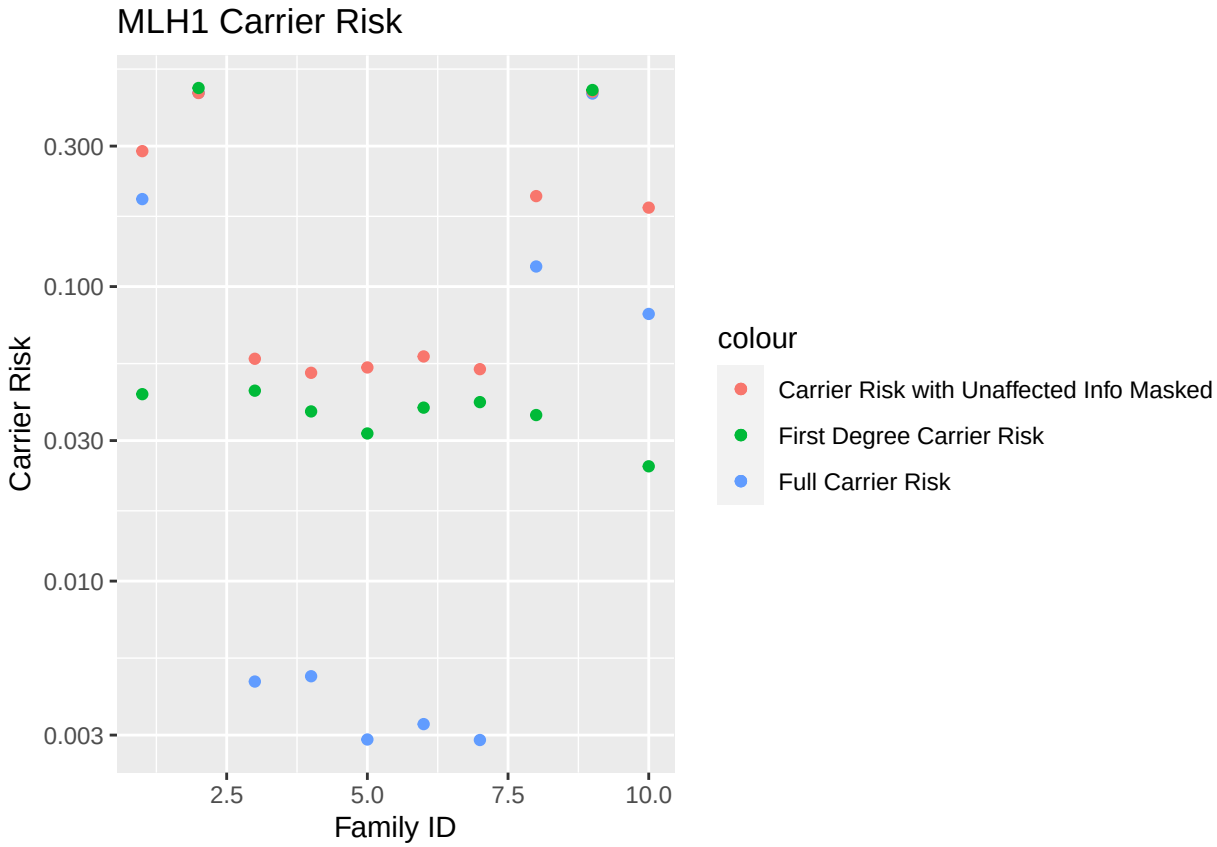
fit two separate models and if you have a very flexible model it could be applicable to both

affect of first degree relatives is probably strong, color plots binary if they have a first degree family member affected

dig into this discussion with some examples families[[2]] can be one of them

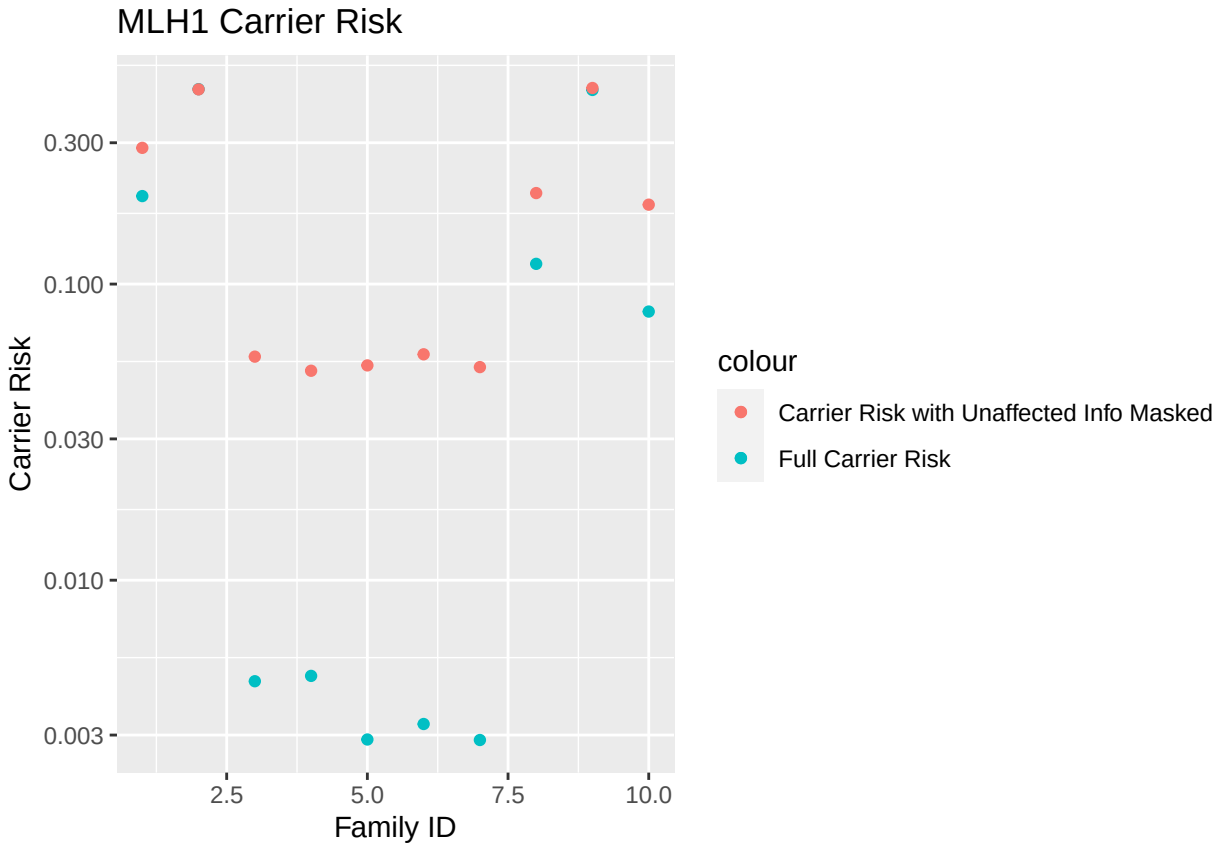
this could be factored into the model as well

```
summaryTable <- summaryTable[1:10,]
logit <- function(x){
  log(x)-log(1-x)
}
visual = ggplot(data= summaryTable) +  geom_point(aes(x=famID, y=fullCarrierRisk, color="Full Carrier Ri
  geom_point(aes(x=famID, y=carrierRiskUnaffectedInfoMasked, color = "Carrier Risk with Unaffected Info
  geom_point(aes(x=famID, y= firstDegreeCarrierRisk, color = "First Degree Carrier Risk")) +
  scale_y_log10() +
  labs(title= "MLH1 Carrier Risk", x= "Family ID", y="Carrier Risk")

print(visual)
```



Next we will generate the same plot but only focus on the full family information and the families with masked unaffected relatives

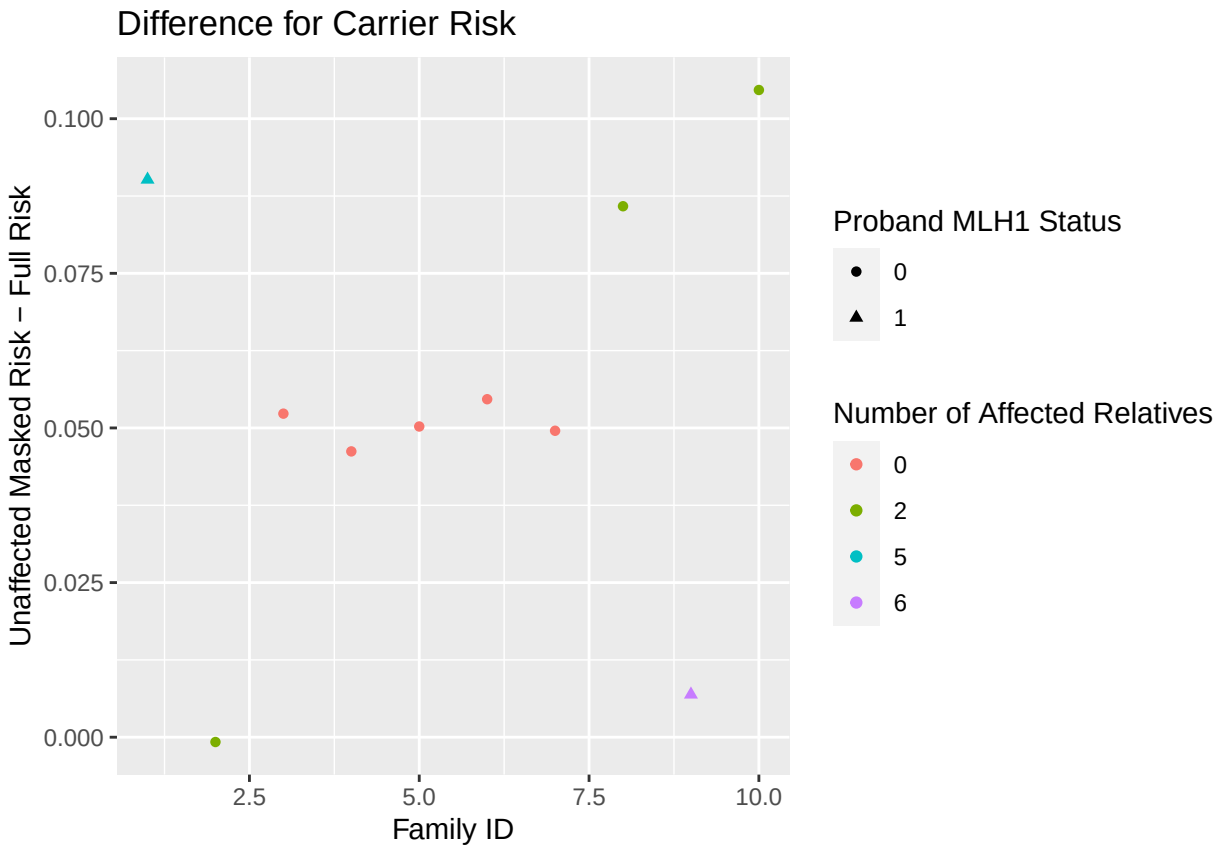
```
ggplot(data= summaryTable) +  geom_point(aes(x=famID, y=fullCarrierRisk, color="Full Carrier Risk")) +
  geom_point(aes(x=famID, y= carrierRiskUnaffectedInfoMasked, color = "Carrier Risk with Unaffected Info
  scale_y_log10() +
  labs(title= "MLH1 Carrier Risk", x= "Family ID", y="Carrier Risk")
```



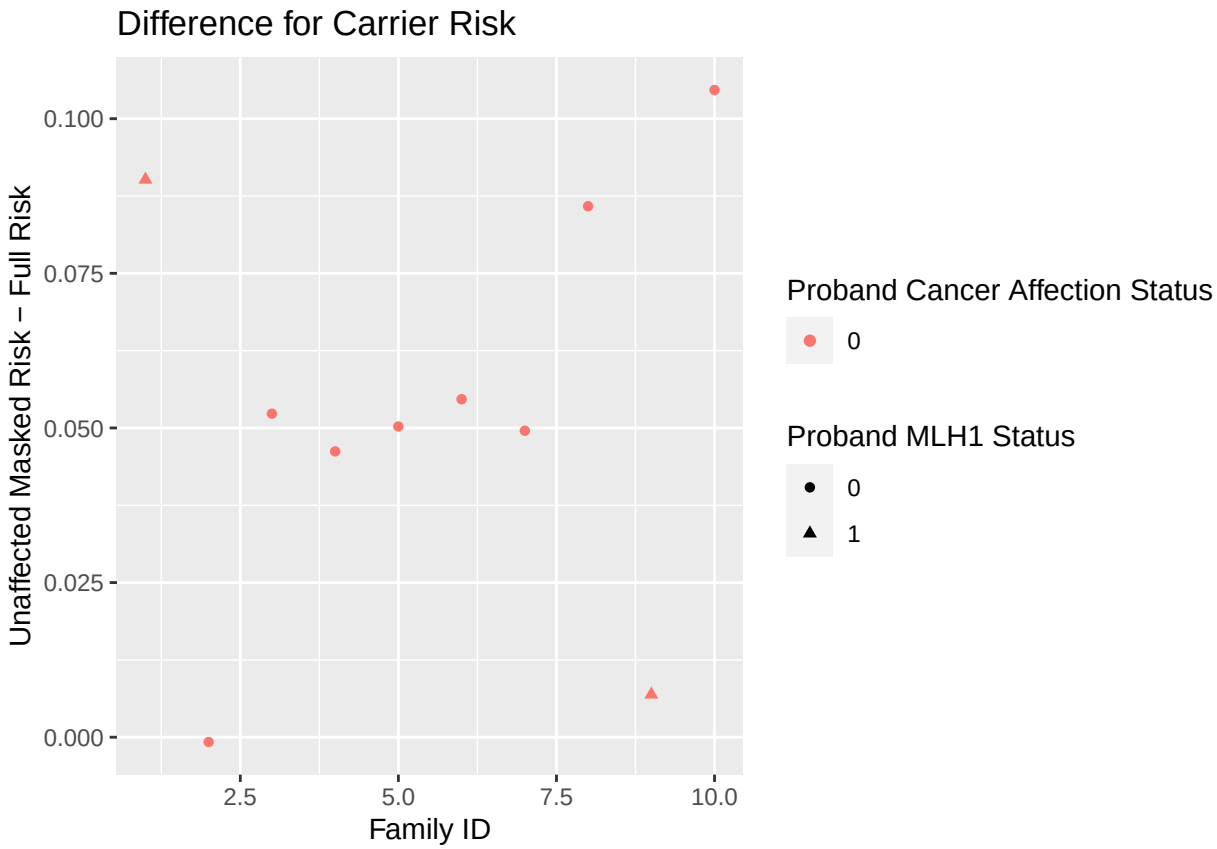
This plot shows the difference between the carrier risk with the unaffected info masked and the full carrier risk

do box plots by proband status

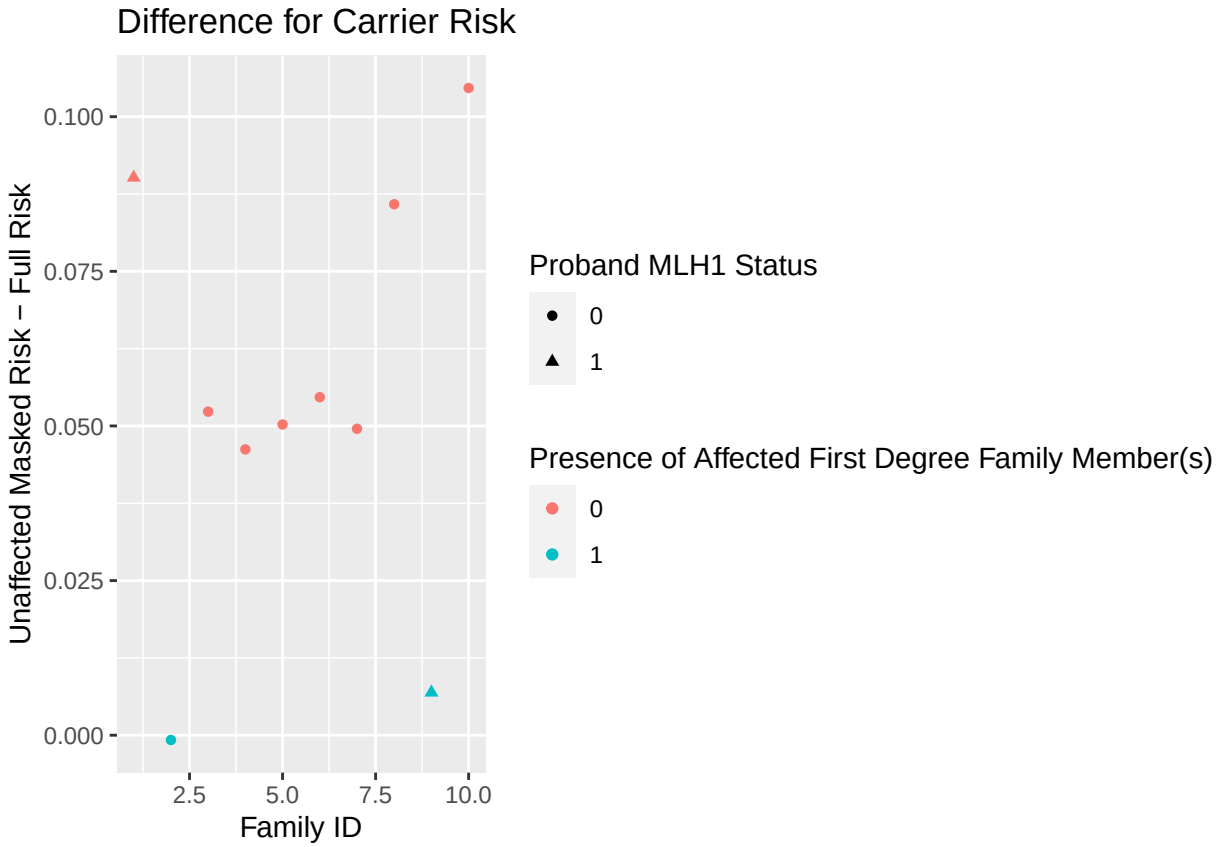
```
summaryTable <- summaryTable %>%
                    mutate(diffMaskedFull = as.double(carrierRiskUnaffectedInfoMasked) - as.double(full


ggplot(data = summaryTable, aes(x = famNumber, y = diffMaskedFull)) +
  geom_point(aes(x = famID, y = diffMaskedFull, colour= as.factor(numAffectedRelatives), shape = proban
  labs(x = "Family ID", y = "Unaffected Masked Risk - Full Risk", title="Difference for Carrier Risk",
```




```
ggplot(data = summaryTable, aes(x = famNumber, y = diffMaskedFull)) +
  geom_point(aes(x = famID, y = diffMaskedFull, colour= as.factor(probandAffectionStatus), shape = proba
  labs(x = "Family ID", y = "Unaffected Masked Risk - Full Risk", title="Difference for Carrier Risk",
```

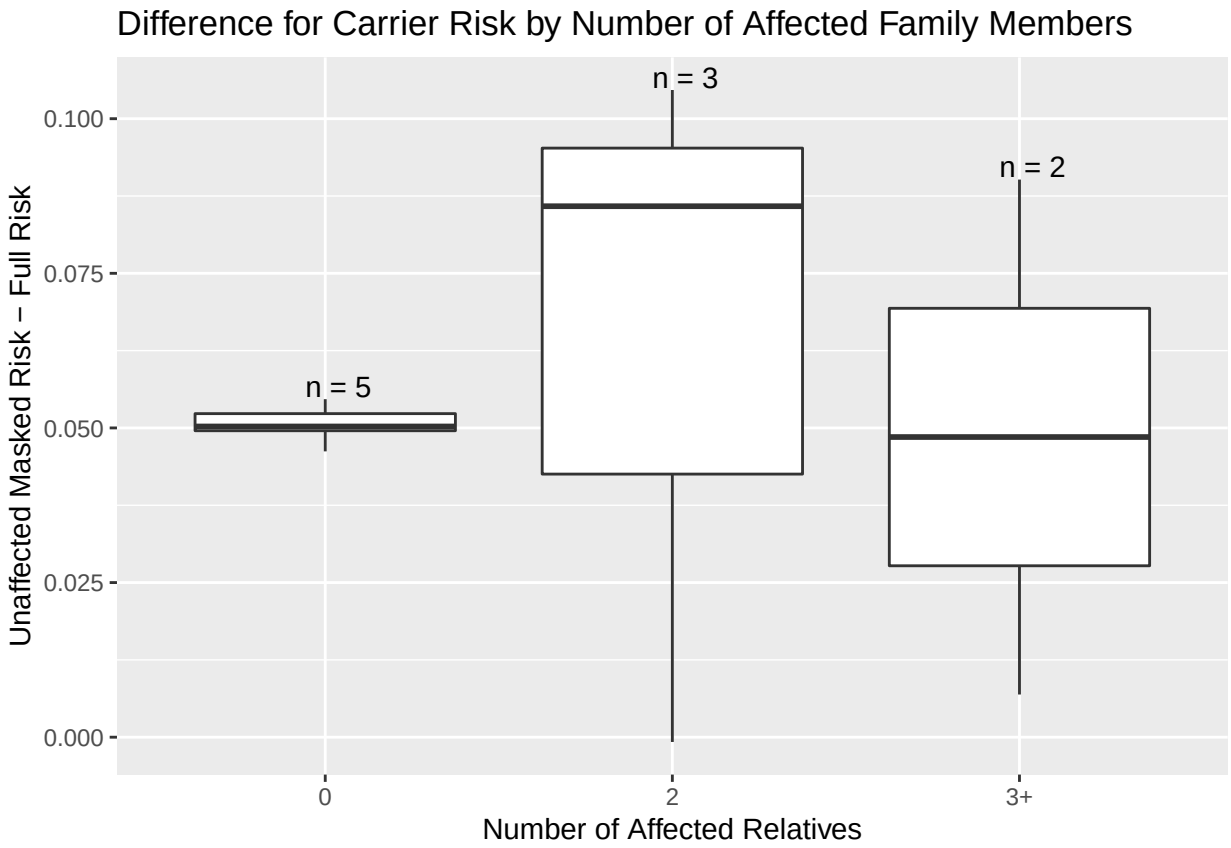



```
ggplot(data = summaryTable, aes(x = famNumber, y = diffMaskedFull)) +
  geom_point(aes(x = famID, y = diffMaskedFull, colour= as.factor(firstDegreeAffectedFamilyMembersBinary
  labs(x = "Family ID", y = "Unaffected Masked Risk - Full Risk", title="Difference for Carrier Risk",
```



Next we will make a box plot to show by number of affected relatives what the difference in risk is

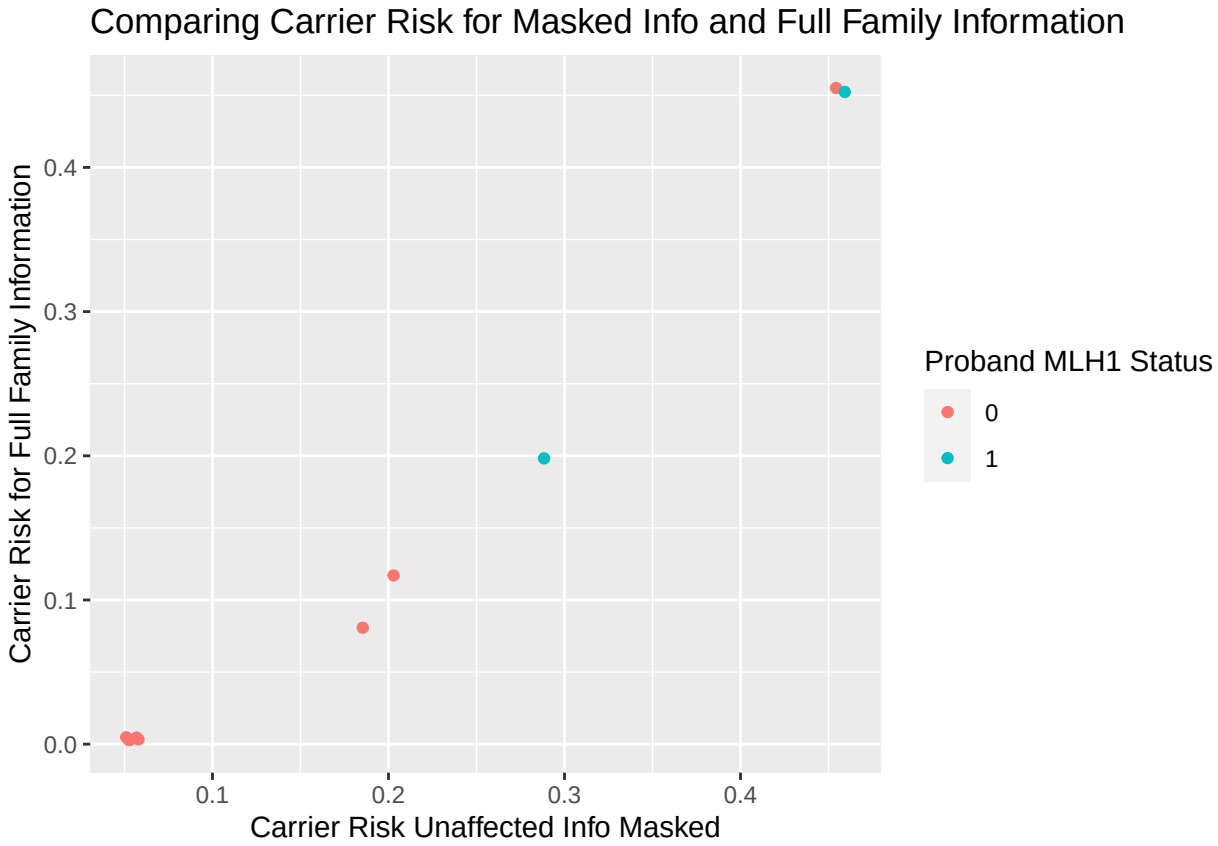
```
summaryTable$numAffectedRelativesGroup <- cut(summaryTable$numAffectedRelatives,
                   breaks=c(-Inf, 0, 1,2, Inf),
                   labels=c("0","1","2", "3+"),
                   right = TRUE) #the breaks are inclusive on the right i.e. (-inf,0], (0,1]...
ggplot(summaryTable, aes(x=as.factor(numAffectedRelativesGroup), y=diffMaskedFull)) +
  geom_boxplot() +
  labs(x = "Number of Affected Relatives", y = "Unaffected Masked Risk - Full Risk", title="Difference i
  stat_summary(
    fun.data = function(x) data.frame(y = max(x), label = paste0("n = ", length(x))),
    geom = "text", hjust = 0.3, vjust = -0.1, size = 4
  )
```



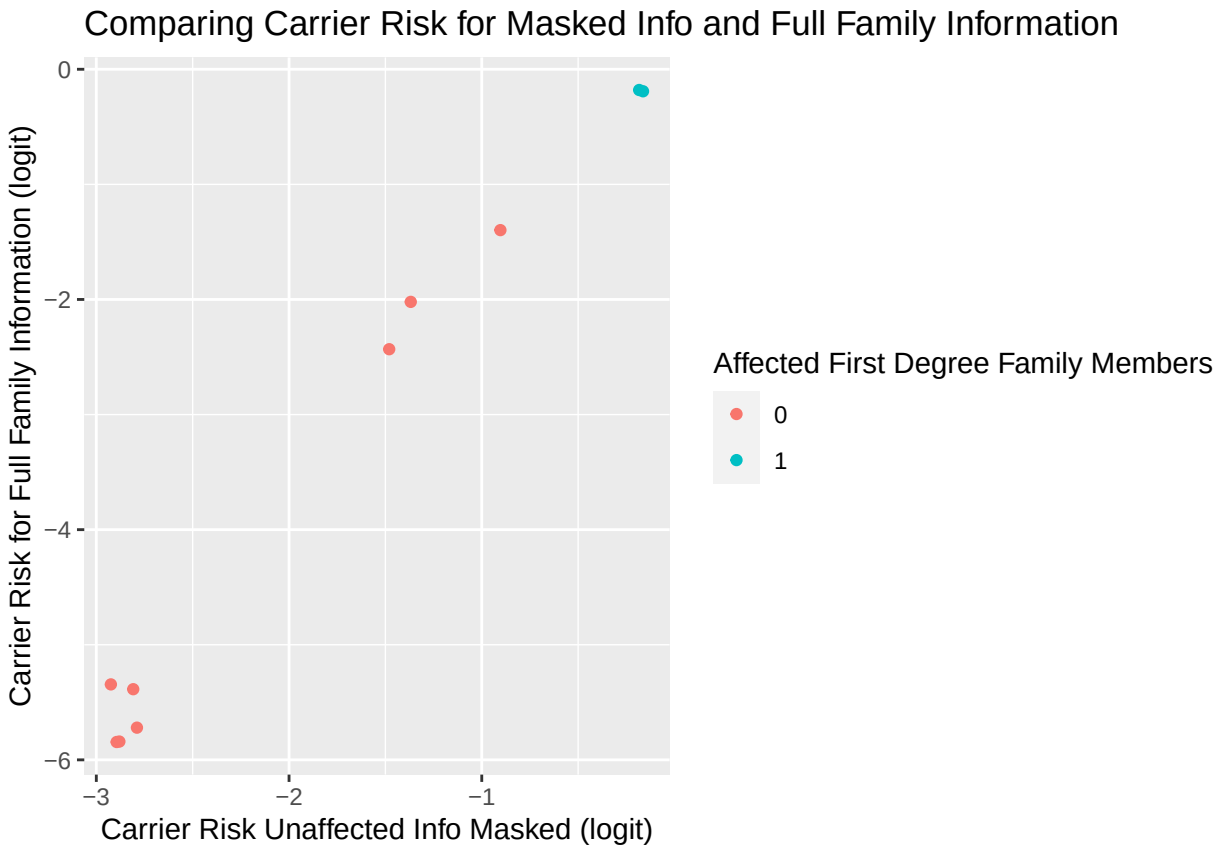
Next we will compare the carrier risk scores for the full families and the masked families

This will serve as a starting point for prediction

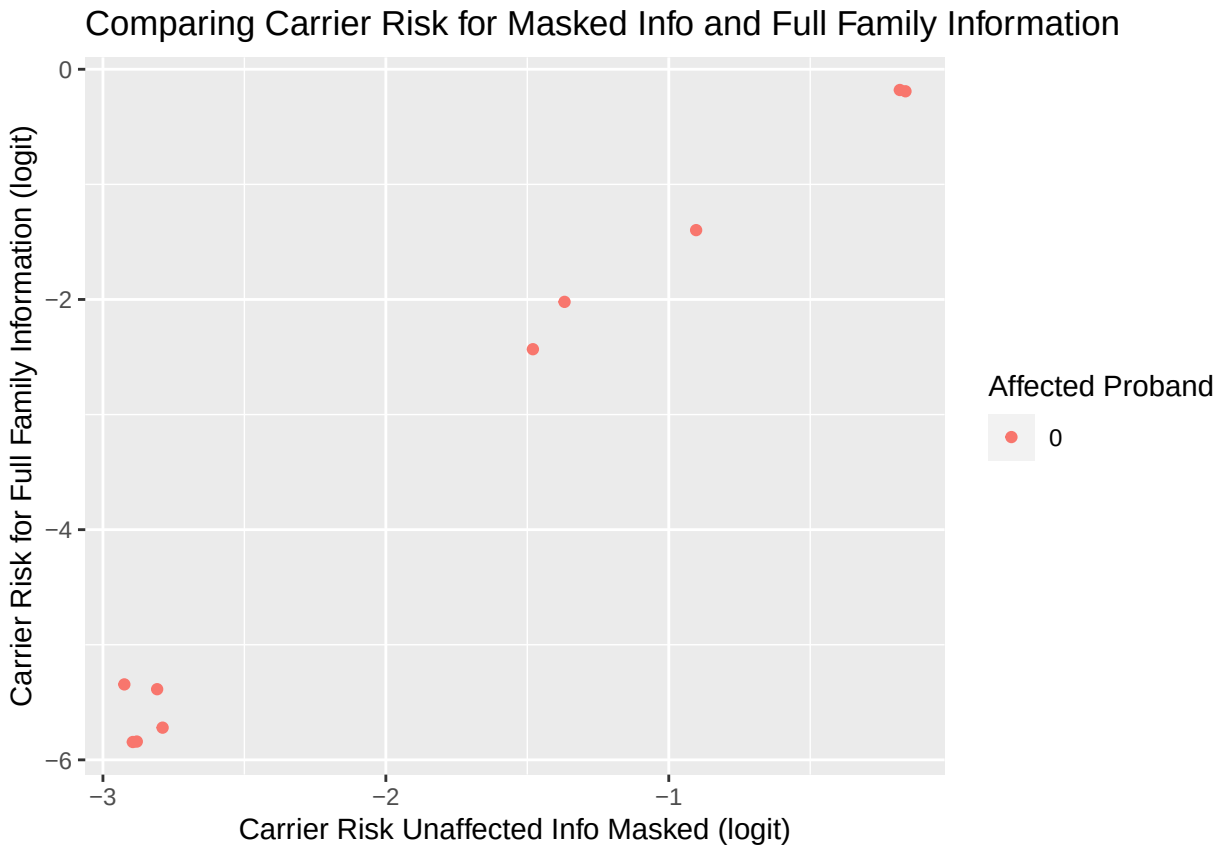
```
logit <- function(x){
  log(x)-log(1-x)
}
ggplot(data = summaryTable, aes(x= carrierRiskUnaffectedInfoMasked, y= fullCarrierRisk, color = probandl
  geom_point() +
  labs(x = "Carrier Risk Unaffected Info Masked", y= "Carrier Risk for Full Family Information", title =
```



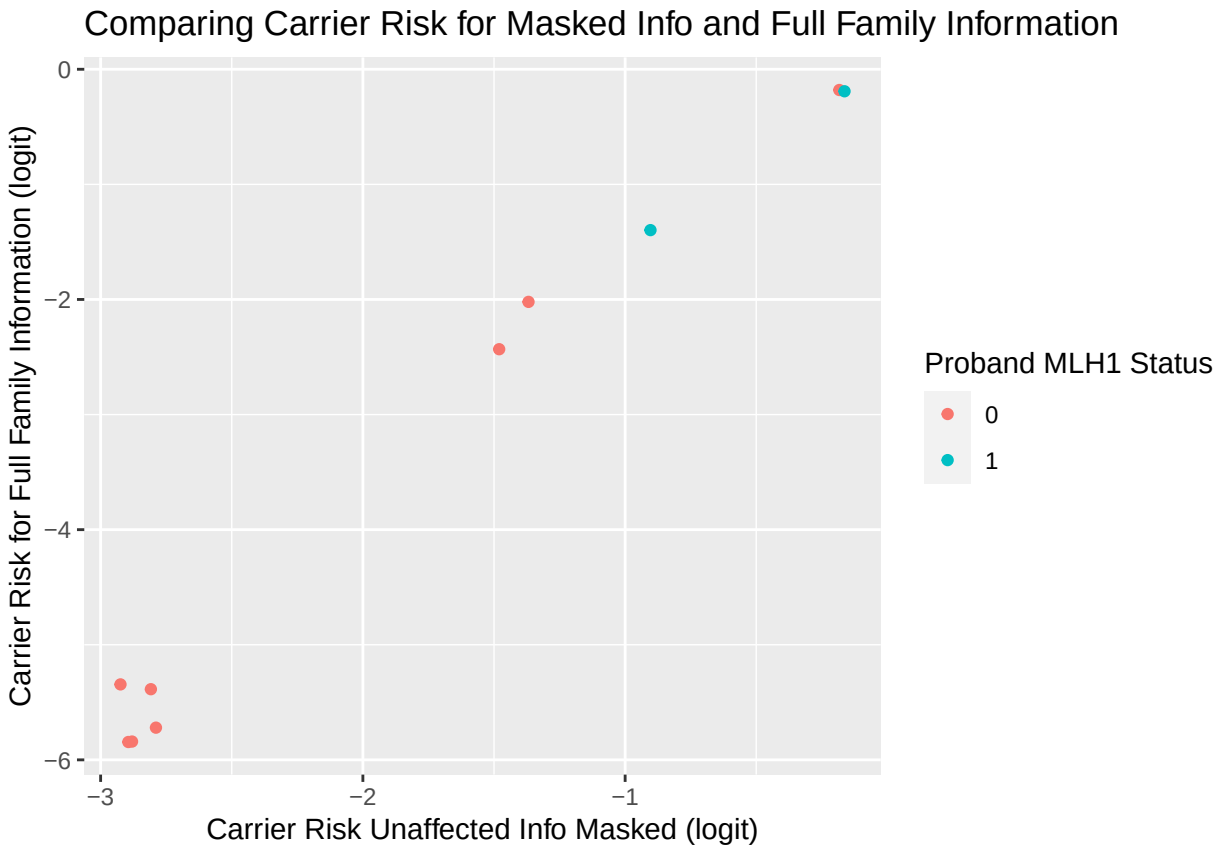

```
logit <- function(x){
  log(x)-log(1-x)
}
ggplot(data = summaryTable, aes(x= logit(carrierRiskUnaffectedInfoMasked), y= logit(fullCarrierRisk), c
  geom_point() +
  labs(x = "Carrier Risk Unaffected Info Masked (logit)", y= "Carrier Risk for Full Family Information
```



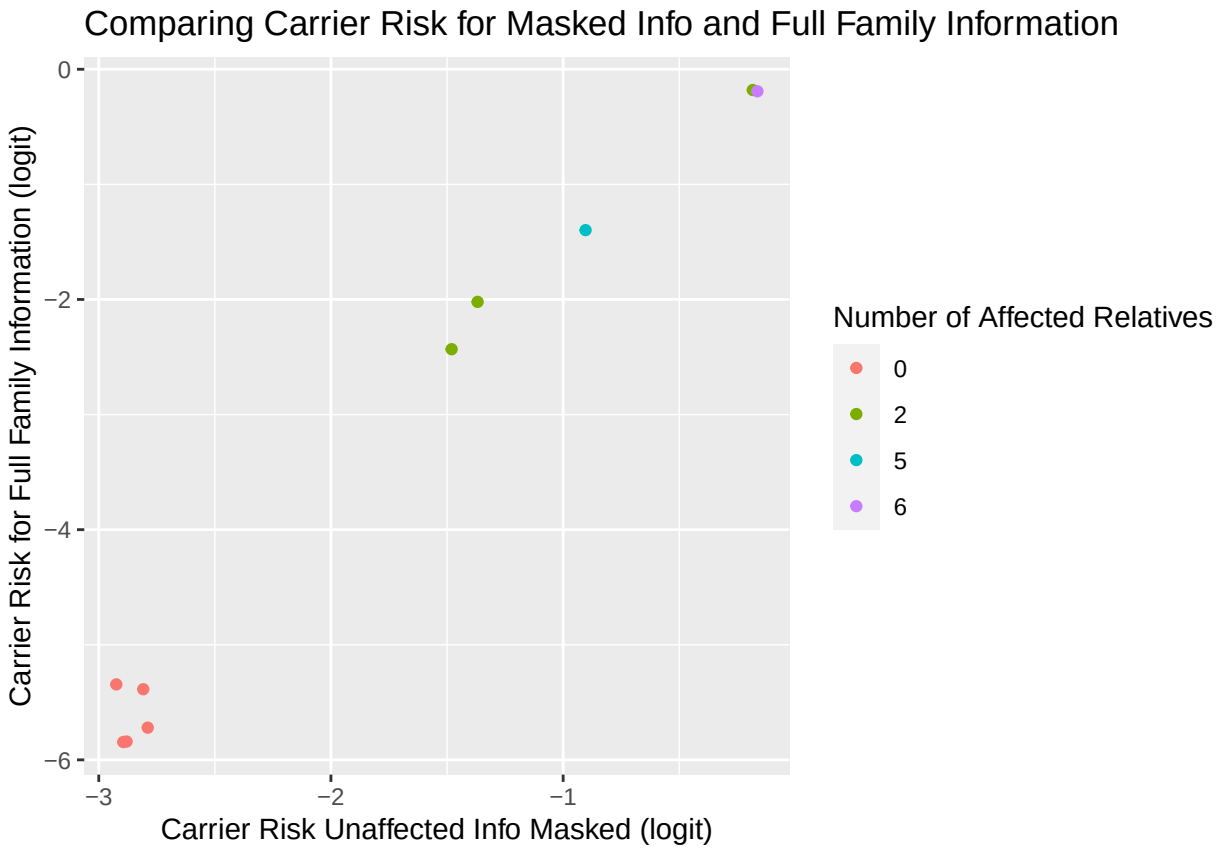

```
logit <- function(x){
  log(x)-log(1-x)
}
ggplot(data = summaryTable, aes(x= logit(carrierRiskUnaffectedInfoMasked), y= logit(fullCarrierRisk), c
  geom_point() +
  labs(x = "Carrier Risk Unaffected Info Masked (logit)", y= "Carrier Risk for Full Family Information
```




```
logit <- function(x){
  log(x)-log(1-x)
}
ggplot(data = summaryTable, aes(x= logit(carrierRiskUnaffectedInfoMasked), y= logit(fullCarrierRisk), c
  geom_point() +
  labs(x = "Carrier Risk Unaffected Info Masked (logit)", y= "Carrier Risk for Full Family Information
```

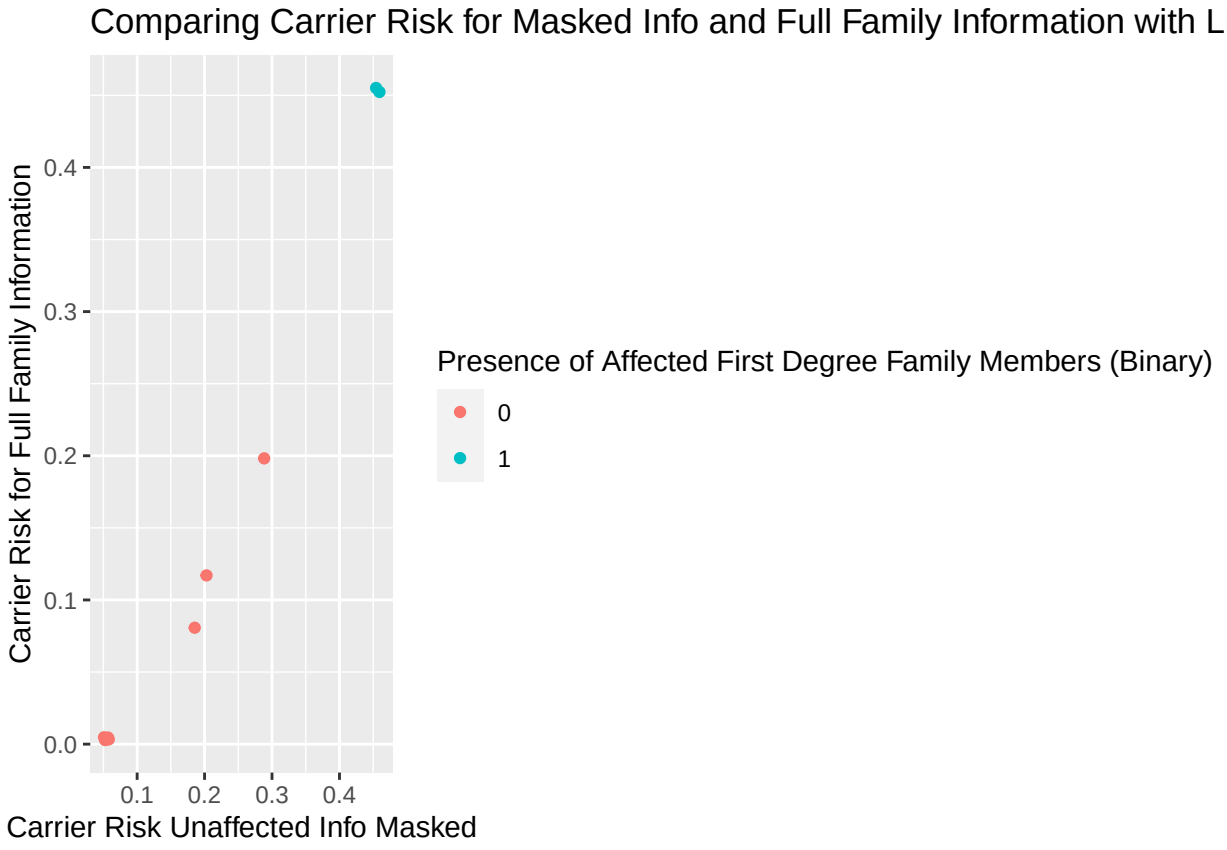



```
logit <- function(x){
  log(x)-log(1-x)
}
ggplot(data = summaryTable, aes(x= logit(carrierRiskUnaffectedInfoMasked), y= logit(fullCarrierRisk), c
  geom_point() +
  labs(x = "Carrier Risk Unaffected Info Masked (logit)", y= "Carrier Risk for Full Family Information
```

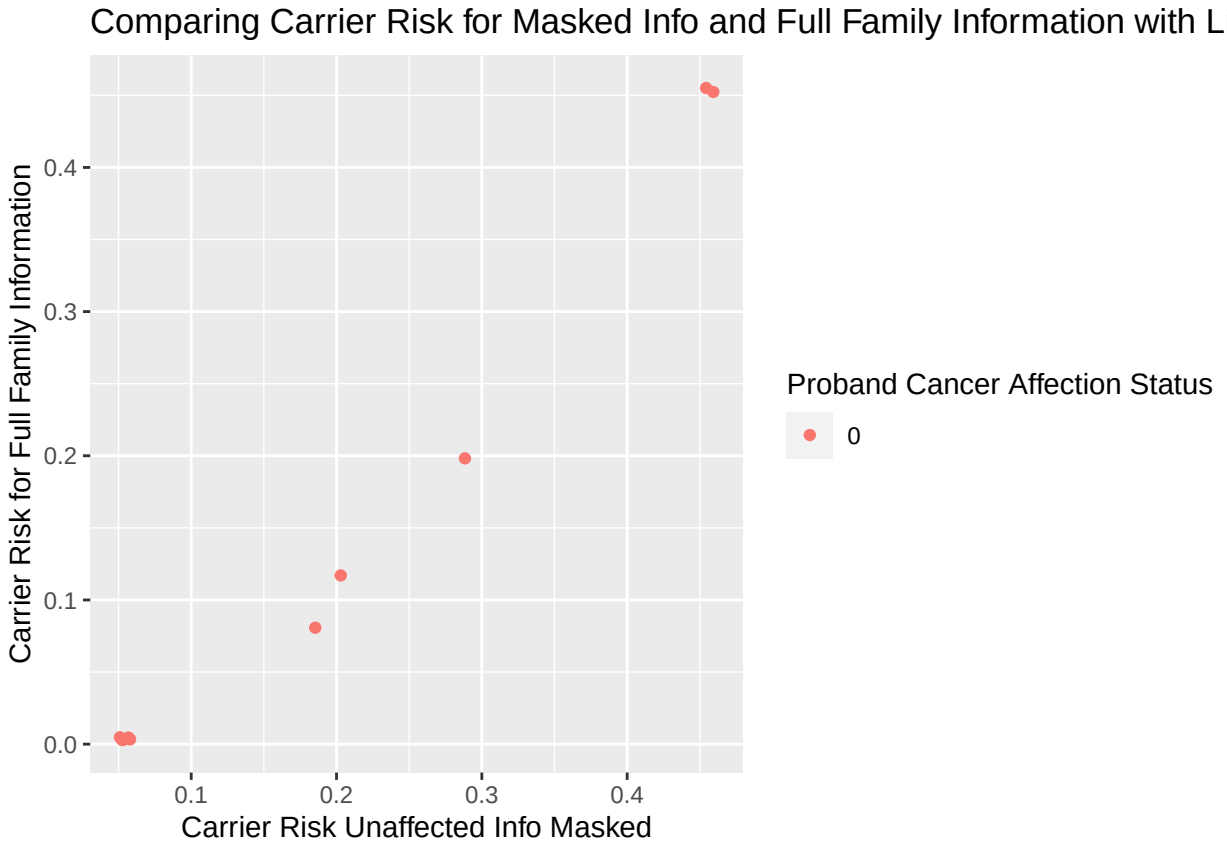


Here we will color by the number of affected relatives to see the relationship between number of affected relatives and carrier risk scores

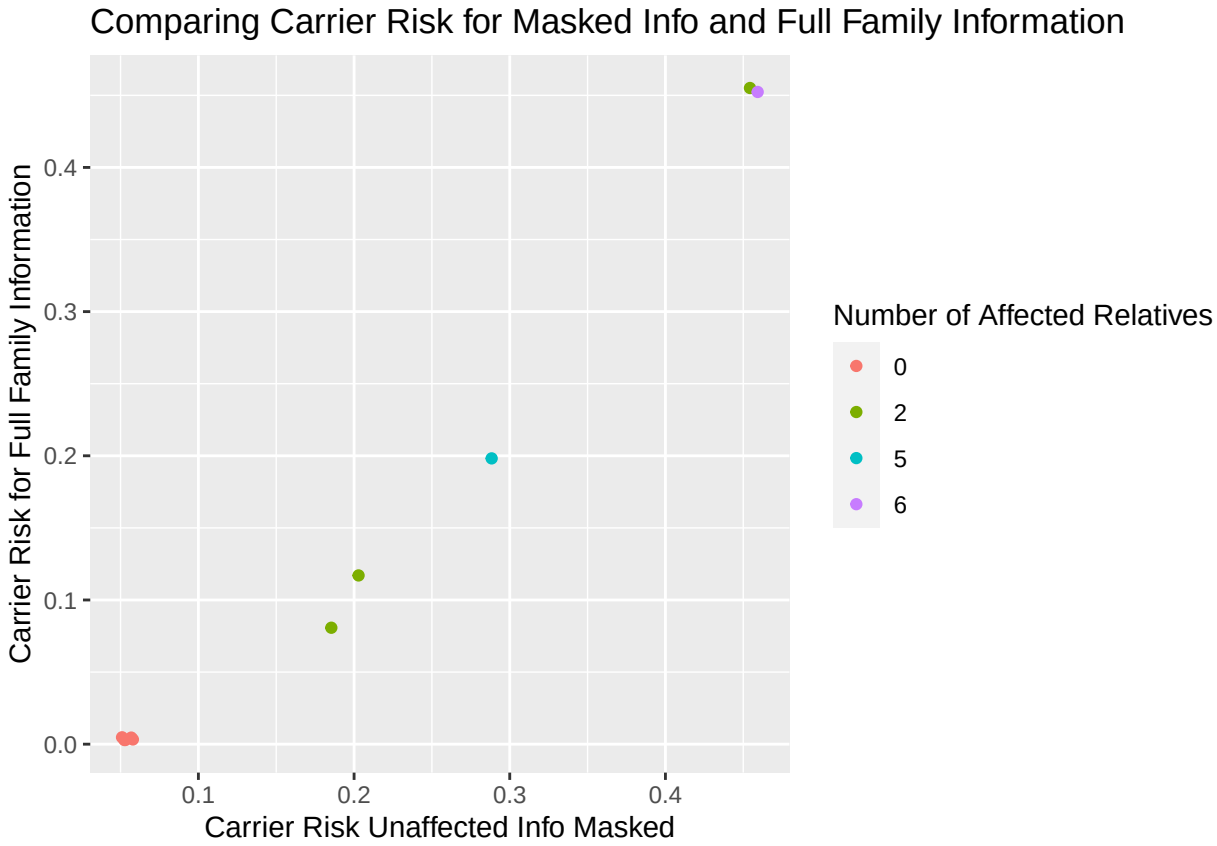
```
ggplot(data = summaryTable, aes(x= carrierRiskUnaffectedInfoMasked, y= fullCarrierRisk, color = as.facto
  geom_point() +
  labs(x = "Carrier Risk Unaffected Info Masked", y= "Carrier Risk for Full Family Information", title =
```




```
ggplot(data = summaryTable, aes(x= carrierRiskUnaffectedInfoMasked, y= fullCarrierRisk, color = as.facto
  geom_point() +
  labs(x = "Carrier Risk Unaffected Info Masked", y= "Carrier Risk for Full Family Information", title =
```



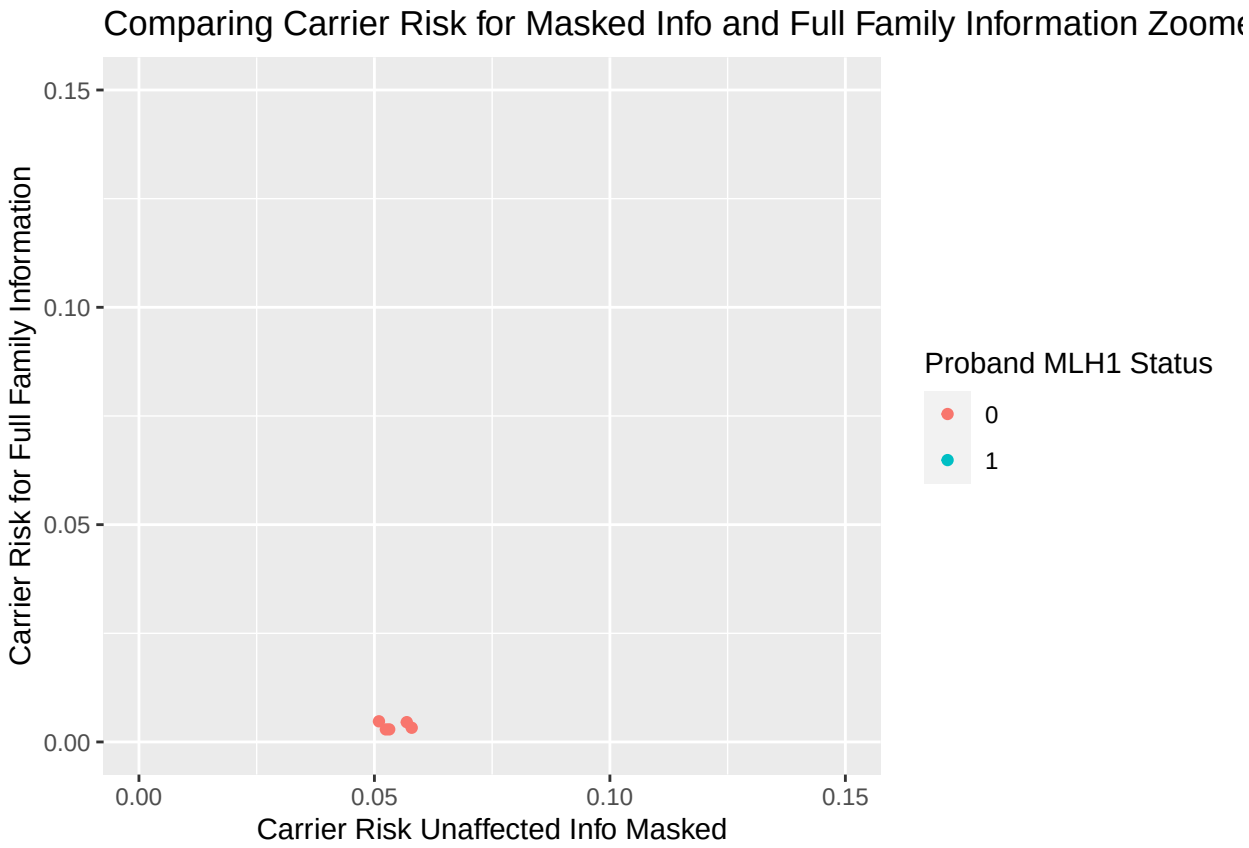

```
ggplot(data = summaryTable, aes(x= carrierRiskUnaffectedInfoMasked, y= fullCarrierRisk, color = as.facto
  geom_point() +
  labs(x = "Carrier Risk Unaffected Info Masked", y= "Carrier Risk for Full Family Information", title =
```



Zoom in to the bottom left corner to see if there are any trends there

```
ggplot(data = summaryTable, aes(x= carrierRiskUnaffectedInfoMasked, y= fullCarrierRisk, color = probandl
  geom_point() +
  labs(x = "Carrier Risk Unaffected Info Masked", y= "Carrier Risk for Full Family Information", title =
  xlim(0, 0.15) + ylim(0, 0.15)
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```

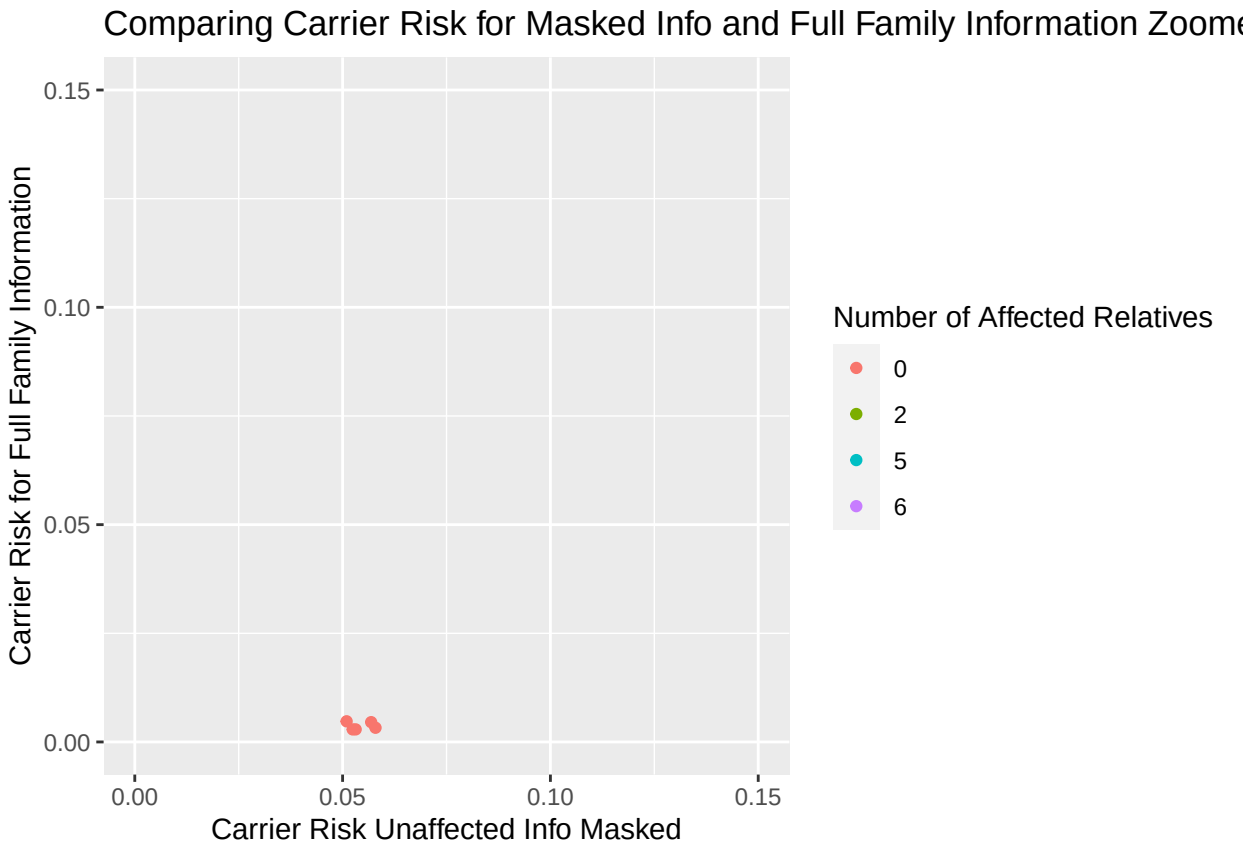


use more distinct colors for the color scale

an alternative to use 0, 1, 2, 3+ as the categories of the families numaffectedrels

```
ggplot(data = summaryTable, aes(x= carrierRiskUnaffectedInfoMasked, y= fullCarrierRisk, color = as.facto
  geom_point() +
  labs(x = "Carrier Risk Unaffected Info Masked", y= "Carrier Risk for Full Family Information", title =
  xlim(0, 0.15) + ylim(0, 0.15)
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```


Comparing Carrier Risk for Masked Info and Full Family Information Zoome
Carrier Risk for Full Family Information
0.15
0.10
0.05
0.00
0.00
0.05
0.10
0.15
Carrier Risk Unaffected Info Masked
Number of Affected Relatives
0
2
5
6