**Prospective follow‐up investigation on** **melanocytic naev development**

**(mole growth) in Colorado kids by solar elastosis**

# Abstract

This study mainly focused on the relation between the sun exposure and genetics is critical for nevus formation. Our results indicate that there is not enough evidence to confirm the causal relationship between sun exposure during vacation and the mole growth in the responding year. Overall, the specific mechanism and underline signaling pathway of navea development is still unclear. More questions should be answered by future research and studies.

# Introduction

Melanomas on sun-exposed skin increases dramatically by solar elastosis, which is a degenerative change of the elastic fibers of the dermis induced by long-term exposure to UV radiation.1 Melanocytes at diverse sites of body can develop into various phenotypes of melanoma, caused by multiple genetic, environmental, and behavioral. The most common melanoma is usually found on sun-exposed skin in Caucasians2. For example, the benign lesions, termed melanocytic naevi, intermediate in the pathway from melanocyte to melanoma, the malignant ones (Supplementary Figure 1). The researcher classifies the melanocyte by whether it is caused by chronically sun damaged (CSD). As shown in Supplementary Figure 1, CSD melanomas can be identified through physiological and anatomical signs of long-term exposure to UV. The prevalence of CSD melanomas are higher in people older than 50 year-old; those melanomas cells are more have a high mutation burden and are associated with neurofibromin 1 (NF1), NRAS, BRAFnonV600E or KIT mutations.3 In the contrast, non-CSD melanomas typically affect the less exposure body areas in younger individuals (<55 years of age) that do not show sun burnt damage.3 (See Supplementary Figure 1)

# Questions and hypothesis

Questions: to investigate which of the genetic, environmental, and behavioral factors can contribute to the development of melanocytic naev.

Hypothesis: the interaction of sun exposure and genetics is critical for nevus formation.

# Materials and Methods

Families recruited in the Denver CO area, through private and public pediatric clinics and community settings. 472 children age 6 followed from baseline to age 10; for this research only 15 variables are collected (details in Table 1 and   
Table 2). The data from the 2007 and 2008 are more appropriate for this purpose.

Table Characteristics (Categorical) of Study Participants



Table : Characteristics of Study Participants (Continued)



# Statistical Methods

All the data were analyzed with R version 3.6.1 Copyright (C) 2019 The R Foundation for Statistical Computing and multiple packages were applied, details in Appendix. The specific methods are explained and justified in the Results and Discussion. Our power analyses indicate that we have adequate statistical power to answer our research questions.

# Analysis Plan

**Examine nevus development with respect to gender, onset of puberty, and phenotypic traits (hair color, eye color, skin pigmentation).**

Most nevi develop in childhood, and aside from genetic predisposition, sun exposure is the major determinant of nevus prevalence. Based on the cohort design with collection of data, this study first tries to find the trend of genetic factors that contribute to nevus growth and development. It is essential for us to understand and recognized the sensitive group of people vulnerable to the mole growth. So, the linear model for each combination of variables (including the maximum of nevi number, gender, race, skin color, and genotype of *OCA2*). Total 12 linear models will be examined to detect the specific factor positively related to high occurrence of moles. Bootstrap will be applied to check the efficiency of two specific models.

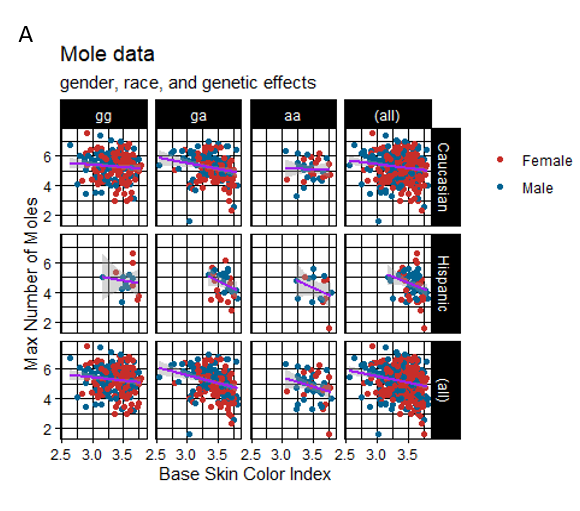
**Optimize the influence of phenotypic traits and genetic factors model on total nevus counts.**

The linear models with different variable and estimates will the compared in multiple statistical methods. And the final model will be optimized through AIC score from a stepwise test from a general linear regression model contains all the variables (including the maximum of nevi number, gender, race, skin color, eye color, hair color, and genotype of *OCA2*) and all the secondary interaction terms for major variables.

**Determine the relationship between UV exposures in vacation over a single year and the nevus counts in the corresponding year.**

The relation between the mole count and the times of vacation and sun exposures will be examined through a 2-by-2 table. The specific risk ratio and odd ratio will be reported to confirm the causal effect of vacation and mole growth. A permutation test is performed to confirm the result from the contingency table with Chi-square test.

# Results

Modeling is an essential tool for visualization. After the diagnostic tests for the assumptions of linear model and ANOVA, we found the normality, homogeneity and linearity has be violated. There are several outliners, so the log transformation was applied to meet with the assumption for linear models (the proof is shown in Supplementary Figure 3A, B, C, D). For better biological explanation, we decide the log2 transformation is more reasonable for log2(molecount) and log2(skincolor). If we back transform to the original scale, we get the ratio of molecount/skincolor, the change of 1 unite of log scale corresponds to double the original scale. The total 12 linear models with combination of gender, race, and genotype, are examined and reported. (details in Supplementary Table 1 and Supplementary Table 2). The R squared value of each model is calculated and compared in a faceted plot in Figure 1A. The studentized Jackknife residual after removing the effect from the regression model is analyzed and plotted in Figure 1A. We found the linear model just including those variables are not good enough to predict and descript the relationship between the vulnerability of mole growth with the genetic and phenetic features. As shown in Supplementary Table 2 and Supplementary Figure 2,the largest R squared values are around 0.3, which means there are only up to 30% of the residual explained by the model. 

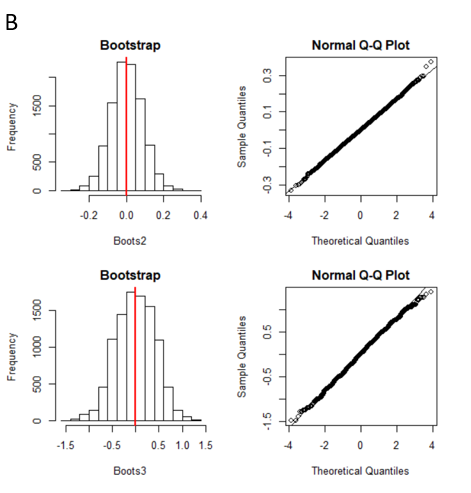


Figure : nevus development with genetic and phenotypic traits. 1A. The facets plots of maximum of the count of moles and skin color are analysis in each combination among different gender, race, and genotypes. The linear regression and the fitted range for the certain model is shown. 1B.Bootstrap simulation on the efficiency of models is testified by the comparison of two studentized residuals simulations from two models Boots2 (Caucasian, Male, OCA2/gg) and Boots3 (Hispanic, Female, OCA2/aa). Details in Supplementary Figure 2; Supplementary Table 1; Supplementary Table 2)

Hence, we specifically check the reliability of the statistics of R squared in Supplementary Table 2. The bootstrap 10000 times simulation is applied on two different linear models, Boots2 residuals from Caucasian male with *OCA2/gg* genotype (R square = 0.0063) and Boots3 residuals from Hispanic female with *OCA2/aa* (R square = 0.3036). Although the Boots3 has one of the largest R square value, the bootstrap results show that the variance is larger than Boots2, of which the model could only explain less than 1% residual of the data. We justified that the sample size of Hispanic female with *OCA2/aa* is really small (n=6) comparing to Caucasian male with *OCA2/gg* model (n=93). This might partly explain the low efficiency of the models. The result is confirmed with ANOVA (See Table 3) with female Caucasian *OCA2/gg* genotype as reference group, on the analysis on the gender, race, and genotype of *OCA2*. The results of significance groups consistent with the individual linear models with higher R square values.

Table ANOVA for log2(Molecount) ~ gender + race + OCA2 )



Signif. codes: 0.001 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

To better predict the vulnerability of mole growth and the genetic and phenetic features, we included eye color and hair color into the adjusted model. The crude model only includes the variables analyzed in last part. The details are shown in Table 4 for the crude model, and Table 5 for the adjusted model. The adjusted model does not improve very much for the model. The new variables only contribute R-square from 0.098 to 0.13, but the difference is still significant (See Table 6).

Table The Crude linear model of log2(Molecount) ~ gender + race + OCA2



Residual standard error: 0.8513 on 465 degrees of freedom; Multiple R-squared: 0.098; Adjusted R-squared: 0.09024; F-statistic: 12.63 on 4 and 465 DF, p-value: 9.169e-10  
Signif. codes: 0.001 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 5 The adjusted linear model of log2(Molecount) ~ gender + race + log2(baseskincolor) + haircolor + eye + OCA2



Residual standard error: 0.8394 on 459 degrees of freedom; Multiple R-squared: 0.13; Adjusted R-squared: 0.1154; F-statistic: 7.115 on 10 and 459 DF, p-value: 1.9e-10

Signif. codes: 0.001 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’

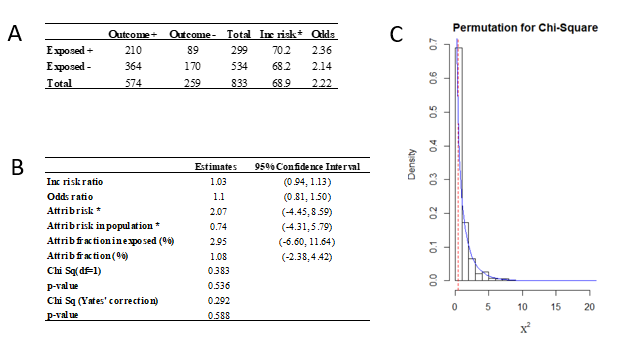
Early research shows nevus counts were strongly associated with *OCA2*; with an additive effect of g alleles on mean nevus counts. Hence the eye color and hair color can be regarded as the mediator of *OCA2*. Finally, we optimized the model by AIC stepwise with all the variables and the secondary interactions, and the hair color continue to play important role in mole numbers. Although the model is improved but it is hard to explain results included in Supplementary Table 3.

Table ANOVA for the crude and adjusted models



Signif. codes: 0.001 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’

Finally, the number of vacations numbers and the nevus counts in the corresponding year is examined for the potential risk effects. As shown in Figure 2A shown, the 2 by 2 table is summarized. The risk ratio and odd ratio are calculated in Figure 2B, but both the RR and OR are near to 1, which suggests no difference or little difference in risk and indicates that the condition or event under study is equally likely to occur in both groups. Chi square and permutation test are also used to confirm the result in Figure 2C.

Figure 2: The relation between vacation and number of moles. 2A: the 2-by-2 table is summarized based on the epidemiology studies shown in Table 2. 2B: The statistical results on the about the relation between vacation and mole in 2007 and 2008. 2C: The plot of permutation simulation on the Chi squared test, p-value > 0.05.

# Discussion

This study mainly focused on the relation between the sun exposure and genetics is critical for nevus formation. There is not enough evidence to confirm the causal relationship between sun exposure during vacation and the mole growth in the responding year. Although this study could not prove a model to predict the outcome of mole based on the genetic and physical features, due to limited samples and age group, and other genes related to melanoma, we did confirm that the positive relation of skin color and the development of moles in Caucasian and Hispanic group of people. Hair color is also one of the major factors will increase the possibility of mole growth. Overall, the specific mechanism and underline signaling pathway of navus development is still unclear. More questions should be answered by future research and studies.

# Reference

1. Crane LA, Asdigian NL, Baron AE, et al. Mailed intervention to promote sun protection of children: a randomized controlled trial. *Am J Prev Med.* 2012;43(4):399-410.

2. Wiecker TS, Luther H, Buettner P, Bauer J, Garbe C. Moderate sun exposure and nevus counts in parents are associated with development of melanocytic nevi in childhood: a risk factor study in 1,812 kindergarten children. *Cancer.* 2003;97(3):628-638.

3. Bastian BC. The molecular pathology of melanoma: an integrated taxonomy of melanocytic neoplasia. *Annu Rev Pathol.* 2014;9:239-271.

4. Crane LA, Mokrohisky ST, Dellavalle RP, et al. Melanocytic nevus development in Colorado children born in 1998: a longitudinal study. *Arch Dermatol.* 2009;145(2):148-156.

5. Shain AH, Bastian BC. From melanocytes to melanomas. *Nat Rev Cancer.* 2016;16(6):345-358.

Contents

[Abstract 1](#_Toc26527110)

[Introduction 1](#_Toc26527111)

[Questions and hypothesis 1](#_Toc26527112)

[Materials and Methods 1](#_Toc26527113)

[Statistical Methods 1](#_Toc26527114)

[Analysis Plan 1](#_Toc26527115)

[Results 2](#_Toc26527116)

[Discussion 3](#_Toc26527117)

[Reference 4](#_Toc26527118)

[Content 5](#_Toc26527119)

[Supplementary figures and tables 6](#_Toc26527120)

[Appendix 11](#_Toc26527121)

# Supplementary figures and tables

Supplementary Table : Estimates for each linear regression model on log2(Molecount) ~log2(skincolor)



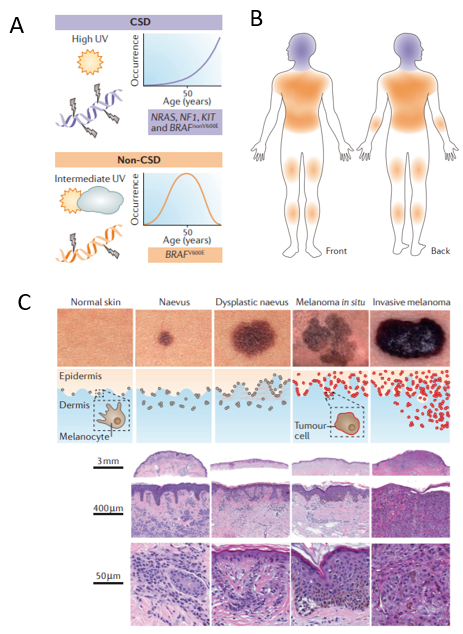
Supplementary Table : Results of the fitness and statistics of the model on log2(Molecount) ~log2(skincolor)



Supplementary Table Estimates for the optimized model for log2(Molecount) ~ log2(skincolor)   
in different combination for the race, gender, and OCR genotype

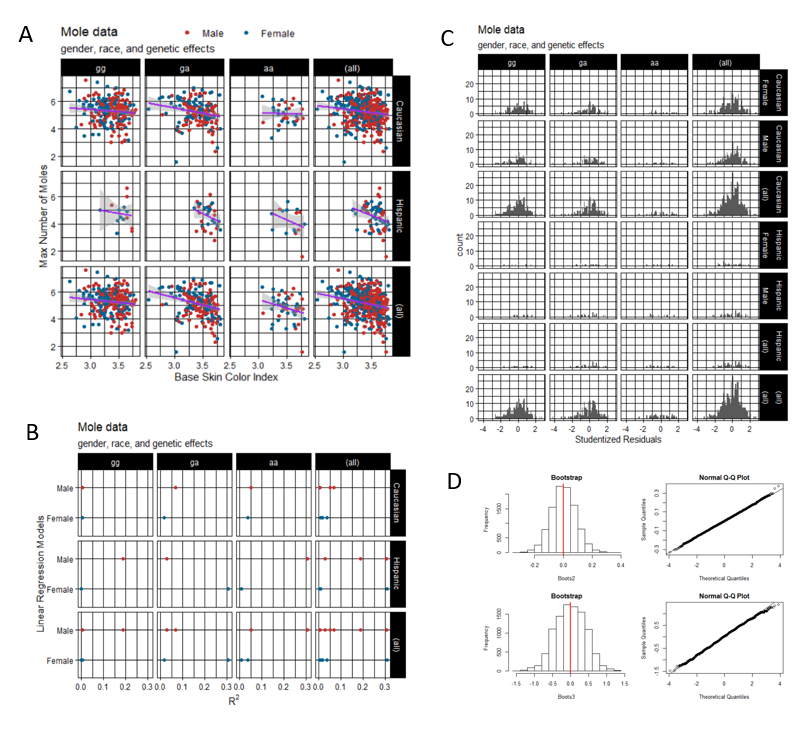


Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

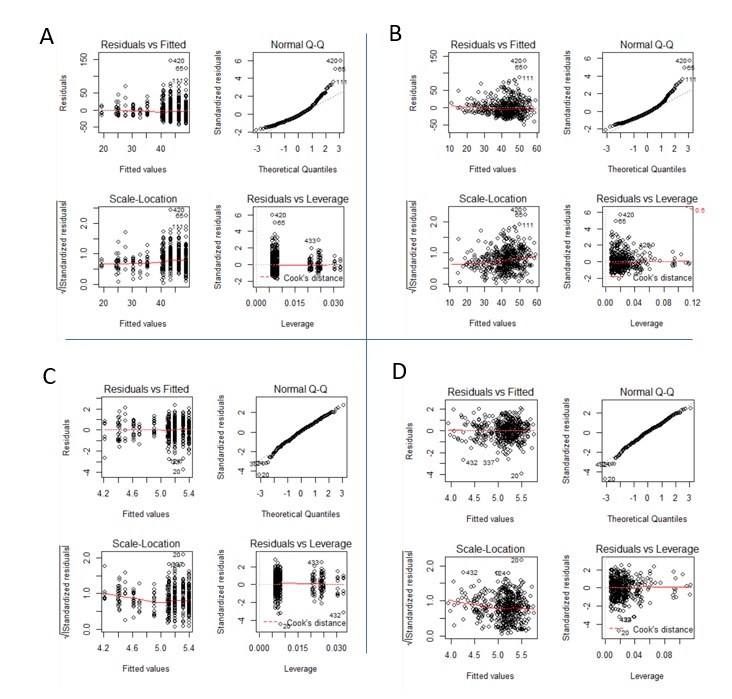


Supplementary Figure : CSD and non-CSD melanoma classifications. 1A.

CSD melanomas have higher mutation burdens and late age onset; Non-CSD melanomas present earlier in life, commonly naevi, have lower mutation and with intermediate levels of sun exposure. 1B. CSD melanomas occupy anatomical sites with the highest levels of sun exposure; Non-CSD melanomas lOCAtes at anatomical sites with intermediate levels of sun exposure. 1C. Clinical images, architectural features, and histopathological features showing a free-standing naevus, a dysplastic naevus, melanoma in situ and invasive melanoma. (Graphs are from Shain AH, Bastian BC. From melanocytes to melanomas. Nat Rev Cancer. 2016;16(6):345-358)



Supplementary Figure : Nevus development with genetic and phenotypic traits. 2A. The facets plots of maximum of the count of moles and skin color are analysis in each combination among different gender, race, and genotypes. The linear regression and the fitted range for the certain model is shown. 2B. The R squared value for each model is plotted and compared through facets. 2C. The studentized residuals are analyzed through histograms for each linear model and combinations. D. The confirmation of efficiency of models is testified by the comparison of two studentized residuals simulations from two models Boot2 (Caucasian, Male, OCA2/gg) and Boot3 (Hispanic, Female, OCA2/aa).



Supplementary Figure Diagnostic tests on the data before and after transformation. SA1: the test results and plots for the untransformed crude model: molecount~gender, race, OCA2, skincolor; SB1: the test results and plots for the untransformed adjusted model: molecount~gender, race, OCA2, skincolor, eyecolor, haircolor; SC1: the test results and plots for the transformed crude model: log2(molecount)~gender, race, OCA2, log2(skincolor); SD1: the test results and plots for the transformed adjusted model: log2(molecount)~gender, race, OCA2, log2(skincolor), eyecolor, haircolor;Topleft: the residuals versus fitted model; Topright: QQplot for the theoretical quantile and the standardized studentized residuals; Bottomleft the scale location for the residual fitted into square root residuals; Bottomright: the Cook’s distance for recidu

# Appendix

library(tidyverse)  
library(ggplot2)  
library(readr)  
library(car)  
library(MASS)  
library(broom)  
library(table1)  
library(epitools)

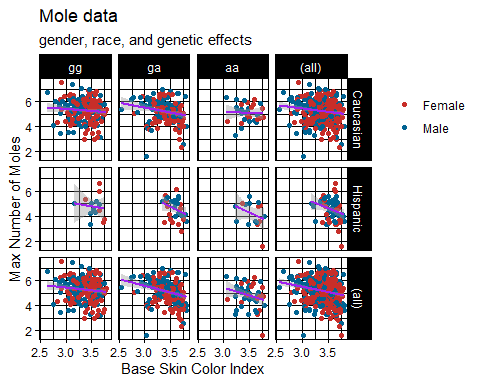
MoleData <- read\_csv("C:/Users/Goodgolden5/Desktop/BIOS6611-Alexander Kaizer/BIOS6611-Final Project/Mole Count Data 2004-2008.csv") %>%   
 as\_tibble() %>%   
 mutate(inc2005=molecount2005-molecount2004,  
 inc2006=molecount2006-molecount2005,  
 inc2007=molecount2007-molecount2006,  
 inc2008=molecount2008-molecount2007,  
 vac2006=`number vacs birth thru 2006`-`number vacs birth thru 2005`,  
 vac2007=`number vacs birth thru 2007`-`number vacs birth thru 2006`) %>%  
 gather(`molecount2004`:`molecount2008`, key="year", value ="molecount") %>%   
 gather(`number vacs birth thru 2005`:`number vacs birth thru 2007`, key="vac", value="vacnum") %>%   
 gather(`inc2005`:`inc2008`, key="incmole", value="moleinyear") %>%  
 gather(`vac2006`:`vac2007`, key="incvac", value="vacinyear") %>%  
 separate(year, into=c("M","year"), sep=9) %>%  
 separate(vac, into=c("T", "vactotalyear"), sep="thru ") %>%  
 separate(incmole, into=c("I", "moleyear"), sep=3) %>%  
 separate(incvac, into=c("V", "vacyear"), sep=3) %>%  
 rename(id=`Respondent Code Number`, oca2=`oca2 status`) %>%  
 rename(eye=eyecolor,  
 race=hispanic,  
 hair=haircolor) %>%  
 dplyr::select(-"M", -"T", -"I", -"V") %>%  
 filter(vacyear<=moleyear) %>%  
 filter(vacyear>=year) %>%  
 arrange(id, vacyear)

MoleData$oca2 <- factor(MoleData$oca2,  
 levels=c("0", "1", "2"),  
 labels=c("gg", "ga", "aa"))  
MoleData$gender <- factor(MoleData$gender,  
 levels=c("1", "2"),  
 labels=c("Female", "Male"))  
MoleData$race <- factor(MoleData$race,  
 levels=c("0", "1"),  
 labels=c("Caucasian", "Hispanic"))  
MoleData$eye <- factor(MoleData$eye,   
 levels=c("1", "2", "3"),  
 labels=c("Blue/Green", "Brown", "Hazel"))  
MoleData$hair <- factor(MoleData$hair,  
 labels=c("Blonde", "Red", "Brown", "Black"),  
 levels=c("1", "2", "3", "4"))

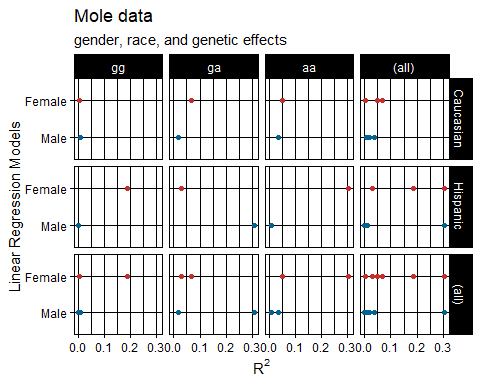
Mole0 <- MoleData %>%   
 dplyr::select(id, oca2, gender, race, eye, hair, baseskincolor, molecount) %>%  
 arrange(id, molecount) %>%  
 na.omit() %>%  
 distinct()  
table1::table1(~oca2+eye+hair+baseskincolor+molecount|gender\*race,   
 data=Mole0,   
 overall="Total",   
 topclass="Rtable1-grid",   
 main="Table 1",   
 render.continuous=c(.="Mean (SD)", .="Median [Min, Max]"))

Mole1 <- MoleData %>%   
 dplyr::select(id, oca2, gender, race, eye, hair, baseskincolor, molecount) %>%  
 na.omit() %>%  
 distinct() %>%  
 group\_by(id, oca2, gender, race, eye, hair, baseskincolor) %>%  
 summarize(Molecount = max(molecount))

p <- ggplot(Mole1, aes(x=log2(baseskincolor), y=log2(Molecount)))  
p +   
 geom\_point(aes(color=gender)) +  
 geom\_smooth(method="lm", se=T, size=1, color="purple") +   
 facet\_grid(race~oca2, margins=T) +  
 theme\_linedraw() +  
 ggthemes::scale\_colour\_wsj("colors6") +  
 ggtitle("Mole data", "gender, race, and genetic effects") +  
 labs(x="Base Skin Color Index", y="Max Number of Moles")+  
 theme(legend.title = element\_blank(), legend.justification = "top")

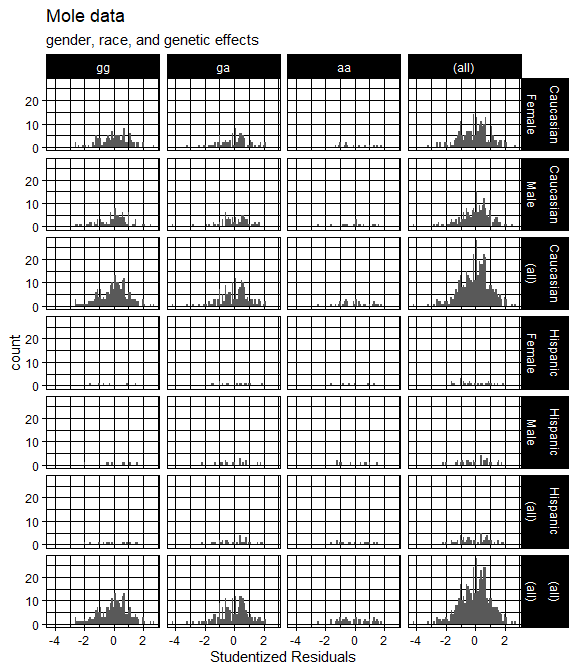


models <- Mole1 %>%  
 group\_by(race, gender, oca2) %>%  
 do(mod = lm(log2(Molecount)~log2(baseskincolor), data=., na.action=na.exclude))  
model\_sum <- models %>% broom::glance(mod)  
  
ggplot(model\_sum, aes(r.squared, reorder(`gender`, r.squared)))+  
 geom\_point(aes(color=gender)) +  
 facet\_grid(race~oca2, margins=T) +  
 theme\_linedraw() +  
 ggthemes::scale\_colour\_wsj("colors6") +  
 ggtitle("Mole data", "gender, race, and genetic effects") +  
 labs(y="Linear Regression Models", x=expression(R^2)) +  
 theme(legend.position="none")



obs\_sum <- models %>% augment(mod)  
obs\_res <- list(obs\_sum$.std.resid)

ggplot(obs\_sum, aes(.std.resid)) +  
 geom\_histogram(binwidth=0.1) +  
 ggtitle("Mole data", "gender, race, and genetic effects") +  
 labs(y="count", x="Studentized Residuals") +  
 theme\_linedraw() +  
 ggthemes::scale\_colour\_wsj("colors6") +  
 facet\_grid(race\*gender~oca2, margins=T)



tcoefs <- models %>%   
 tidy(mod)

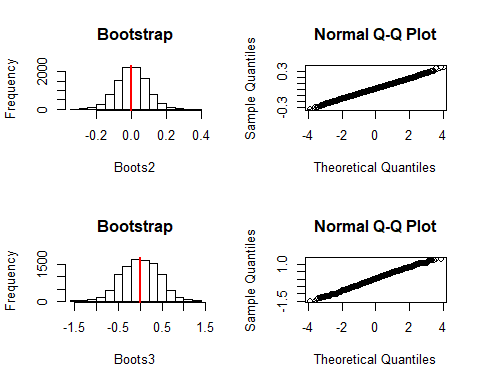
Mole2 <- Mole1 %>%   
 filter(race=="Caucasian" && gender=="Male" && oca2=="gg")   
Mole\_lm\_cmgg <- lm(log2(Molecount)~log2(baseskincolor), Mole2)  
coef(summary(Mole\_lm\_cmgg))

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 6.2787986 1.1952128 5.2532894 9.763470e-07  
## log2(baseskincolor) -0.2744894 0.3616371 -0.7590189 4.498028e-01

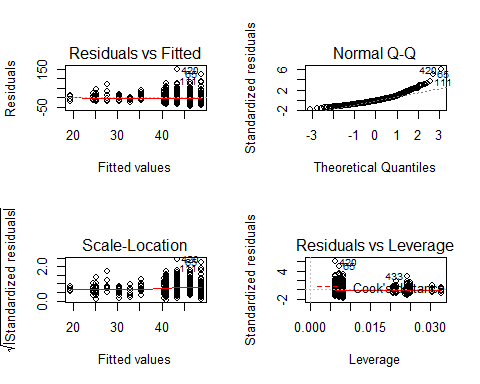
Mole2\_res <- resid(Mole\_lm\_cmgg)  
Mole3 <- Mole1 %>%   
 filter(race=="Hispanic" && gender=="Female" && oca2=="aa")   
Mole\_lm\_hfaa <- lm(log2(Molecount)~log2(baseskincolor), Mole3)  
coef(summary(Mole\_lm\_hfaa))

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 22.896001 14.417900 1.588026 0.1874755  
## log2(baseskincolor) -5.216442 3.949943 -1.320637 0.2571134

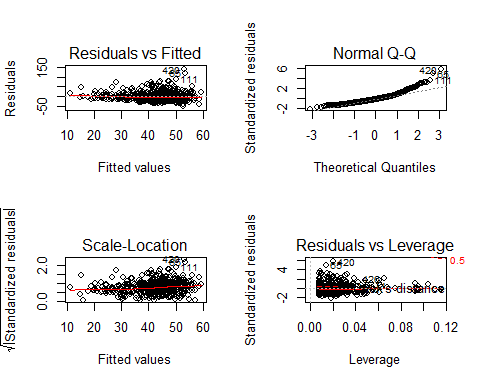
Mole3\_res <- resid(Mole\_lm\_hfaa)  
set.seed(seed=555)  
N <- 10000  
Boots2 <- vector("numeric", length(Mole2\_res))  
for(i in seq\_along(1:N)){  
 Bootstrap2 <- sample(Mole2\_res, length(Mole2\_res), replace = T)  
 Boots2[i] <- mean(Bootstrap2)  
}  
set.seed(seed=555)  
N <- 10000  
Boots3 <- vector("numeric", length(Mole3\_res))  
for(i in seq\_along(1:N)){  
 Bootstrap3 <- sample(Mole3\_res, length(Mole3\_res), replace = T)  
 Boots3[i] <- mean(Bootstrap3)  
}  
par(mfrow=c(2,2))  
hist(Boots2, main = "Bootstrap")  
abline(v = mean(Boots2), col = "red", lwd = 2)  
qqnorm(Boots2)  
qqline(Boots2)  
hist(Boots3, main = "Bootstrap")  
abline(v = mean(Boots3), col = "red", lwd = 2)  
qqnorm(Boots3)  
qqline(Boots3)



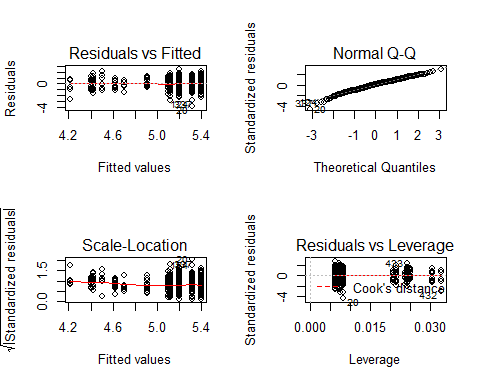
Mole\_lm\_nontrans\_adjust <- lm(Molecount~gender+race+baseskincolor+hair+eye+oca2, Mole1)  
Mole\_lm\_nontrans\_crude <- lm(Molecount~gender+race+oca2, Mole1)  
par(mfrow=c(2,2))  
plot(Mole\_lm\_nontrans\_crude)



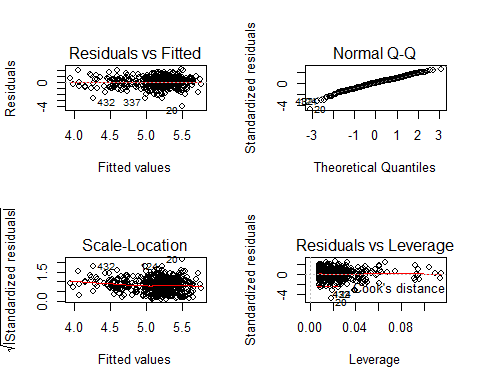
plot(Mole\_lm\_nontrans\_adjust)



Mole\_lm\_adjust <- lm(log2(Molecount)~gender+race+log2(baseskincolor)+hair+eye+oca2, Mole1)  
Mole\_lm\_crude <- lm(log2(Molecount)~gender+race+oca2, Mole1)  
par(mfrow=c(2,2))  
plot(Mole\_lm\_crude)



plot(Mole\_lm\_adjust)



summary(Mole\_lm\_crude);

##   
## Call:  
## lm(formula = log2(Molecount) ~ gender + race + oca2, data = Mole1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.7302 -0.5351 0.0625 0.5512 2.3643   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.20552 0.06692 77.791 < 2e-16 \*\*\*  
## genderMale 0.19355 0.07902 2.449 0.0147 \*   
## raceHispanic -0.69701 0.11981 -5.818 1.11e-08 \*\*\*  
## oca2ga -0.08393 0.08410 -0.998 0.3188   
## oca2aa -0.29370 0.13806 -2.127 0.0339 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8513 on 465 degrees of freedom  
## Multiple R-squared: 0.098, Adjusted R-squared: 0.09024   
## F-statistic: 12.63 on 4 and 465 DF, p-value: 9.169e-10

summary(Mole\_lm\_adjust)

##   
## Call:  
## lm(formula = log2(Molecount) ~ gender + race + log2(baseskincolor) +   
## hair + eye + oca2, data = Mole1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.9174 -0.5443 0.0494 0.5243 2.0837   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.09782 0.62852 11.293 < 2e-16 \*\*\*  
## genderMale 0.13901 0.08027 1.732 0.08398 .   
## raceHispanic -0.53688 0.13407 -4.005 7.24e-05 \*\*\*  
## log2(baseskincolor) -0.54027 0.18288 -2.954 0.00330 \*\*   
## hairRed -0.53172 0.17446 -3.048 0.00244 \*\*   
## hairBrown -0.03585 0.08780 -0.408 0.68321   
## hairBlack -0.35318 0.27542 -1.282 0.20038   
## eyeBrown -0.13787 0.15427 -0.894 0.37195   
## eyeHazel -0.12037 0.12436 -0.968 0.33360   
## oca2ga 0.05435 0.12733 0.427 0.66973   
## oca2aa -0.08698 0.18808 -0.462 0.64399   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8394 on 459 degrees of freedom  
## Multiple R-squared: 0.1342, Adjusted R-squared: 0.1154   
## F-statistic: 7.115 on 10 and 459 DF, p-value: 1.9e-10

anova(Mole\_lm\_crude, Mole\_lm\_adjust)

## Analysis of Variance Table  
##   
## Model 1: log2(Molecount) ~ gender + race + oca2  
## Model 2: log2(Molecount) ~ gender + race + log2(baseskincolor) + hair +   
## eye + oca2  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 465 336.96   
## 2 459 323.43 6 13.53 3.2001 0.004351 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

MoleAov <- aov(log2(Molecount)~gender+race+oca2+log2(baseskincolor), data=Mole1)  
summary(MoleAov)

## Df Sum Sq Mean Sq F value Pr(>F)   
## gender 1 2.5 2.499 3.494 0.06223 .   
## race 1 30.7 30.739 42.982 1.47e-10 \*\*\*  
## oca2 2 3.4 1.685 2.357 0.09585 .   
## log2(baseskincolor) 1 5.1 5.136 7.181 0.00763 \*\*   
## Residuals 464 331.8 0.715   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Mole\_glm <- glm(log2(Molecount)~gender+race+log2(baseskincolor)+hair+eye+oca2+(gender+race+log2(baseskincolor)+hair+eye+oca2)^2, data=Mole1)  
Mole\_step <- MASS::stepAIC(Mole\_glm, ~gender+race+log2(baseskincolor)+hair+eye+oca2+ (gender+race+log2(baseskincolor)+hair+eye+oca2)^2, direction="backward", trace=F)  
summary(Mole\_step)

##   
## Call:  
## glm(formula = log2(Molecount) ~ gender + race + log2(baseskincolor) +   
## hair + log2(baseskincolor):hair, data = Mole1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.1348 -0.5590 0.0480 0.5146 1.9524   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.9124 0.9032 6.546 1.58e-10 \*\*\*  
## genderMale 0.1324 0.0786 1.684 0.09279 .   
## raceHispanic -0.5179 0.1303 -3.974 8.22e-05 \*\*\*  
## log2(baseskincolor) -0.2009 0.2641 -0.761 0.44729   
## hairRed -5.3033 2.2820 -2.324 0.02056 \*   
## hairBrown 3.0878 1.2043 2.564 0.01066 \*   
## hairBlack 6.8374 6.8452 0.999 0.31838   
## log2(baseskincolor):hairRed 1.4507 0.6882 2.108 0.03557 \*   
## log2(baseskincolor):hairBrown -0.9285 0.3537 -2.625 0.00894 \*\*   
## log2(baseskincolor):hairBlack -2.0534 1.9048 -1.078 0.28158   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.6828325)  
##   
## Null deviance: 373.57 on 469 degrees of freedom  
## Residual deviance: 314.10 on 460 degrees of freedom  
## AIC: 1166.4  
##   
## Number of Fisher Scoring iterations: 2

Mole4 <- MoleData %>%   
 filter(vacyear==moleyear) %>%  
 na.omit() %>%   
 dplyr::select(id, gender, race, moleyear, moleinyear, vacinyear) %>%  
 distinct()   
  
table1::table1(~moleinyear+vacinyear|moleyear,   
 data=Mole4,   
 overall="Total",   
 topclass="Rtable1-grid",   
 main="Table 1",   
 render.continuous=c(.="Mean (SD)", .="Median [Min, Max]"))

a <- Mole4 %>% filter(moleinyear>0) %>% filter(vacinyear>0) %>% count()  
b <- Mole4 %>% filter(moleinyear<=0) %>% filter(vacinyear>0) %>% count()  
c <- Mole4 %>% filter(moleinyear>0) %>% filter(vacinyear==0) %>% count()  
d <- Mole4 %>% filter(moleinyear<=0) %>% filter(vacinyear==0) %>% count()  
MoleTable <- matrix(nrow = 2, byrow = T, c(a[[1]], b[[1]], c[[1]], d[[1]]), dimnames = list(c("Vacation", "NoVacation"), c("Mole", "NoMole")))  
epiR::epi.2by2(MoleTable)

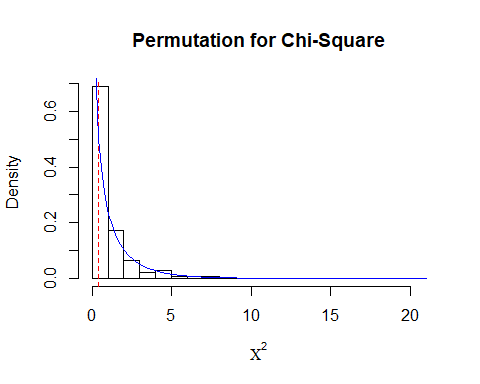
## Outcome + Outcome - Total Inc risk \*  
## Exposed + 210 89 299 70.2  
## Exposed - 364 170 534 68.2  
## Total 574 259 833 68.9  
## Odds  
## Exposed + 2.36  
## Exposed - 2.14  
## Total 2.22  
##   
## Point estimates and 95% CIs:  
## -------------------------------------------------------------------  
## Inc risk ratio 1.03 (0.94, 1.13)  
## Odds ratio 1.10 (0.81, 1.50)  
## Attrib risk \* 2.07 (-4.45, 8.59)  
## Attrib risk in population \* 0.74 (-4.31, 5.79)  
## Attrib fraction in exposed (%) 2.95 (-6.60, 11.64)  
## Attrib fraction in population (%) 1.08 (-2.38, 4.42)  
## -------------------------------------------------------------------  
## Test that odds ratio = 1: chi2(1) = 0.383 Pr>chi2 = 0.536  
## Wald confidence limits  
## CI: confidence interval  
## \* Outcomes per 100 population units

ChiYate <- chisq.test(MoleTable, correct = TRUE); ChiYate

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: MoleTable  
## X-squared = 0.29259, df = 1, p-value = 0.5886

P <- 10^5-1  
MoleExpand <- expand.table(MoleTable)  
Chisq <- function(Obs){  
 Exp <- outer(rowSums(Obs),colSums(Obs))/sum(Obs)  
 sum((Obs-Exp)^2/Exp)  
}  
Exposure <- MoleExpand[, 1]  
Symptom <- MoleExpand[, 2]  
ObsV <- Chisq(table(Exposure, Symptom))  
PerChi <- vector("numeric", P)  
for(i in seq\_along(1:P)){  
 ExPer <- sample(Exposure, replace = F)  
 PerTable <-table(ExPer, Symptom)  
 PerChi[i] <- Chisq(PerTable)  
}

hist(PerChi, freq = F, xlab = expression(Chi^2), main="Permutation for Chi-Square")  
abline(v = ObsV, col = "red", lty = 2)  
curve(dchisq(x, 1), add = T, col = "blue", lwd = 1)



Prepvalue <- (sum(PerChi >= ObsV)+1)/(P+1)  
Obspvalue <- 1 - pchisq(ObsV, df = 1)  
cat("The permutation p-value is", Prepvalue, "\n")

## The permutation p-value is 0.58433

cat("The Chi-square distribution p-value is", Obspvalue, "\n")

## The Chi-square distribution p-value is 0.5359541