## Supplemental\_file\_S4.

## Article: "Serendipitous discovery of large potential effect of ZNF117:rs67047829 in protection against obesity."

## Authors: Jeremy Clark, Konrad Podsiadło, Marta Sobalska-Kwapis, Błażej Marciniak, Kamila Rydzewska, Andrzej Ciechanowicz, Thierry van de Wetering, Dominik Strapagiel

**## INSTRUCTIONS:** 1. Create .csv files from .xlsx Supplemental files by removing all header rows bar one from Supplemental\_Table\_S2\_C and Supplemental\_Table\_S3\_J. 2. Run code between lines below: make sure all libraries are TRUE and "mypath" to files correctly gives mydf2. Copy and paste all coding into R. (This coding has been run on Windows 11 Pro and Mac 10.15.7).

##\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ipak <- function(pkg){

new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]

if (length(new.pkg))

install.packages(new.pkg, dependencies = TRUE, repos="http://cran.r-project.org")

sapply(pkg, require, character.only = TRUE)

} ## end function

packages <- c("data.table", "car", "dplyr", "SNPassoc", "rlist", "epitools")

options(timeout=1000)

ipak(packages)

mypath <- "/Users/jeremyclark/Documents/aaJ Clark MACBOOK 2022 PUM/3 AA SUBMITTED PTCobesity Scientific Reports 140 pt/FIGS AND SUPP/"

mydf2 <- fread(paste0(mypath, "Supplemental\_Table\_S2\_C.csv"))

head(mydf2)

nrow(mydf2)

mydf2 <- as.data.frame(mydf2)

colnames(mydf2) <- make.names(colnames(mydf2), unique = TRUE)

colnames(mydf2)

##\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## Change n.d. to NA throughout dataframe:

mydf2[mydf2 == "n.d."] <- NA

levels(as.factor(mydf2[ , "Age"]))

mydf2[ , "Age"] <- as.numeric(mydf2[ , "Age"])

sum(is.na(mydf2[ , "Age"]))

levels(as.factor(mydf2[ , "BMI"]))

mydf2[ , "BMI"] <- as.numeric(mydf2[ , "BMI"])

sum(is.na(mydf2[ , "BMI"]))

levels(as.factor(mydf2[ , "SEX"]))

sum(is.na(mydf2[ , "SEX"]))

nrow(mydf2)

**## To obtain all numbers found in Supplemental\_Table\_S3:**

## only those subjects analyzed with data for Age, BMI and SEX, n = 5050:

mydfA <- mydf2[complete.cases(mydf2[ , c("Age", "BMI", "SEX")]), ]

head(mydfA)

colnames(mydfA)

nrow(mydfA)

## remove all duplicate columns with \_B:

dfNoDup <- mydfA %>% select(-contains("\_B"))

colnames(dfNoDup)

dfGENnoPHEN <- dfNoDup[ , 11:ncol(dfNoDup)]

names(dfGENnoPHEN) <- gsub("\_A", "", names(dfGENnoPHEN), fixed = TRUE)

colnames(dfGENnoPHEN)

## create genotype column dataframe:

dfgen\_1 <- dfGENnoPHEN[ , c(TRUE, FALSE)]

colnames(dfgen\_1)

ncol(dfgen\_1)

head(dfgen\_1)

dfgen\_2 <- dfGENnoPHEN[ , c(FALSE, TRUE)]

colnames(dfgen\_2)

ncol(dfgen\_2)

head(dfgen\_2)

mycol <- list();

for (i in 1:ncol(dfgen\_1)) {

mycol[[i]] <- paste0(dfgen\_1[ , i], "/", dfgen\_2[ , i])

names(mycol)[i] <- colnames(dfgen\_1)[i]

}

head(mycol[[1]])

GENdf3 <- list.cbind(mycol)

head(GENdf3)

GENdf2 <- GENdf3

for (i in 1:ncol(GENdf3)) {

GENdf2[ , i] <- gsub("NA/NA", "NA", GENdf3[ , i])

}

head(GENdf2)

## add phenotype back:

GENdf <- cbind(mydfA[ , 1:10], GENdf2)

head(GENdf)

colnames(GENdf)

for (i in 11:ncol(GENdf)) {

GENdf[ , i] <- as.factor(GENdf[ , i])

}

ANALYSIS <- setupSNP(GENdf, colSNPs = c(11:ncol(GENdf)), sep = "/")

TABLE\_BMI3 <- WGassociation(BMI ~ Age + SEX, model = c("dominant", "recessive", "overdominant", "log-additive"), data = ANALYSIS, genotypingRate = 0.1)

TABLE\_BMI3 <- as.data.frame(TABLE\_BMI3)

TABLE\_BMI3$rs\_number <- rownames(TABLE\_BMI3)

TABLE\_BMI2 <- TABLE\_BMI3 %>% dplyr::select("rs\_number", everything())

**## rearrange order of SNPs to match those in Supplemental\_Table\_3\_J:**

Table\_S3 <- fread(paste0(mypath, "Supplemental\_Table\_S3\_J.csv"))

nrow(Table\_S3)

Table\_S3 <- as.data.frame(Table\_S3)

colnames(Table\_S3)

rs\_S3 <- Table\_S3[ , "rs\_number"]

TABLE\_BMI <- TABLE\_BMI2[match(rs\_S3, TABLE\_BMI2$rs\_number), ]

TABLE\_BMI <- TABLE\_BMI[1:139, ]

TABLE\_BMI

for (i in 3:ncol(TABLE\_BMI)) {

for (j in 1:nrow(TABLE\_BMI)) {

TABLE\_BMI[j, i] <- signif(TABLE\_BMI[j, i], 3)

}

}

**## write .csv file:**

write.csv(TABLE\_BMI, paste0(mypath, "TABLE\_BMI.csv"))

**## False discovery rate correction:**

dominant <- TABLE\_BMI[ , "dominant"]

recessive <- TABLE\_BMI[ , "recessive"]

overdominant <- TABLE\_BMI[ , "overdominant"]

logadditive <- TABLE\_BMI[ , "log-additive"]

pvaluevec <- c(dominant, recessive, overdominant, logadditive)

padjustvec <- p.adjust(pvaluevec, method = c("BY"), n = length(pvaluevec))

padjustvec

## rs67047829 association with body mass Index. subjects with >5% missing SNPs removed. n = 4425

names(mydf2) <- gsub("rs67047829", "rs67047829\_1", names(mydf2), fixed = TRUE)

names(mydf2) <- gsub("rs67047829\_1.1", "rs67047829\_2", names(mydf2), fixed = TRUE)

colnames(mydf2)

rs67047829 <- paste0(mydf2[ , "rs67047829\_1"], mydf2[ , "rs67047829\_2"])

mydf2$rs67047829 <- rs67047829

levels(as.factor(mydf2[ , "rs67047829"]))

sum(is.na(mydf2[ , "rs67047829"]))

## Counts of rs67047829 genotypes - mydf2:

levels(as.factor(mydf2[ , "rs67047829"]))

Genomydf2 <- mydf2[ , "rs67047829"]

Genomydf2\_AA <- Genomydf2[Genomydf2 == "AA"]

head(Genomydf2\_AA)

length(Genomydf2\_AA)

Genomydf2\_GG <- Genomydf2[Genomydf2 == "GG"]

head(Genomydf2\_GG)

length(Genomydf2\_GG)

Genomydf2\_GA <- Genomydf2[Genomydf2 == "GA"]

head(Genomydf2\_GA)

length(Genomydf2\_GA)

Genomydf2\_AG <- Genomydf2[Genomydf2 == "AG"]

head(Genomydf2\_AG)

length(Genomydf2\_AG)

**## Remove subject if data missing from Age, BMI, SEX or rs67047829:**

mydf1B <- mydf2[complete.cases(mydf2[ , c("Age", "BMI")]), ]

head(mydf1B)

nrow(mydf1B)

colnames(mydf1B)

mydf1C <- mydf2[complete.cases(mydf2[ , c("SEX", "BMI")]), ]

head(mydf1C)

colnames(mydf1C)

nrow(mydf1C)

mydf1D <- mydf2[complete.cases(mydf2[ , c("BMI")]), ]

head(mydf1D)

colnames(mydf1D)

nrow(mydf1D)

mydf1 <- mydf2[complete.cases(mydf2[ , c("Age", "BMI", "SEX", "rs67047829")]), ]

head(mydf1)

nrow(mydf1)

colnames(mydf1)

## Counts of genotypes - mydf1

levels(as.factor(mydf1[ , "rs67047829"]))

Genomydf1 <- mydf1[ , "rs67047829"]

Genomydf1\_AA <- Genomydf1[Genomydf1 == "AA"]

head(Genomydf1\_AA)

length(Genomydf1\_AA)

Genomydf1\_GG <- Genomydf1[Genomydf1 == "GG"]

head(Genomydf1\_GG)

length(Genomydf1\_GG)

Genomydf1\_GA <- Genomydf1[Genomydf1 == "GA"]

head(Genomydf1\_GA)

length(Genomydf1\_GA)

Genomydf1\_AG <- Genomydf1[Genomydf1 == "AG"]

head(Genomydf1\_AG)

length(Genomydf1\_AG)

sum(is.na(mydf1[ , "Age"]))

sum(is.na(mydf1[ , "BMI"]))

sum(is.na(mydf1[ , "SEX"]))

sum(is.na(mydf1[ , "rs67047829"]))

## Find subjects with more than a certain percentage of genotype fails:

NoPhendf1 <- mydf1[ , 11:ncol(mydf1)]

head(NoPhendf1)

ncol(NoPhendf1)

nrow(NoPhendf1)

NAs <- rowSums(is.na(NoPhendf1))

max(NAs)

**## <8 missing gives <5% missing SNP values n = 4425**

**## <22 missing gives <15% missing SNP values n = 5026**

NAtrue <- list();

for (i in 1:length(NAs)) {

if (NAs[[i]] < 8) { **## <8 missing gives <5% missing SNP values n = 4425**

NAtrue[[i]] <- TRUE

} else {

NAtrue[[i]] <- FALSE

}

}

NAtrue <- as.vector(unlist(NAtrue))

mydf <- mydf1[NAtrue, ]

nrow(mydf)

ncol(mydf)

mydf[ , "BMI"] <- as.numeric(mydf[ , "BMI"])

mydf[ , "Age"] <- as.numeric(mydf[ , "Age"])

## Genotype <- paste0(mydf[ , "rs67047829\_1"], mydf[ , "rs67047829\_2"])

levels(as.factor(mydf[ , "rs67047829"]))

Genomydf <- mydf[ , "rs67047829"]

Genomydf\_AA <- Genomydf[Genomydf == "AA"]

head(Genomydf\_AA)

length(Genomydf\_AA)

Genomydf\_GG <- Genomydf[Genomydf == "GG"]

head(Genomydf\_GG)

length(Genomydf\_GG)

Genomydf\_GA <- Genomydf[Genomydf == "GA"]

head(Genomydf\_GA)

length(Genomydf\_GA)

Genomydf\_AG <- Genomydf[Genomydf == "AG"]

head(Genomydf\_AG)

length(Genomydf\_AG)

Genotype <- mydf[ , "rs67047829"] ## Genotype given as e.g. "GG"

levels(as.factor(Genotype))

mydf$Genotype <- Genotype

head(mydf)

**## Dominant model coding:**

Dominant <- Genotype

Dominant[(Genotype == "GA") | (Genotype == "AG") | (Genotype == "AA")] <- 1

Dominant[Genotype == "GG"] <- 0

mydf$Dominant <- as.factor(Dominant)

**## Recessive model coding:**

Recessive <- Genotype

Recessive[Genotype == "AA"] <- 0

Recessive[(Genotype == "GA") | (Genotype == "AG") | (Genotype == "GG")] <- 1

mydf$Recessive <- as.factor(Recessive)

**## Heterozygote model coding:**

Heterozygote <- Genotype

Heterozygote[(Genotype == "GA") | (Genotype == "AG")] <- 1

Heterozygote[(Genotype == "GG") | (Genotype == "AA")] <- 0

mydf$Heterozygote <- as.factor(Heterozygote)

**## Additive model coding:**

Additive <- Genotype

Additive[Genotype == "GG"] <- 0

Additive[(Genotype == "GA") | (Genotype == "AG")] <- 1

Additive[Genotype == "AA"] <- 2

mydf$Additive <- as.factor(Additive)

levels(as.factor(Genotype))

levels(as.factor(mydf[ , "BMI"]))

levels(as.factor(mydf[ , "SEX"]))

levels(as.factor(mydf[ , "Age"]))

levels(as.factor(Dominant))

levels(as.factor(Recessive))

levels(as.factor(Heterozygote))

levels(as.factor(Additive))

AA <- Genotype[Genotype == "AA"]

AA

AG <- Genotype[Genotype == "AG"]

AG

GA <- Genotype[Genotype == "GA"]

GA

GG <- Genotype[Genotype == "GG"]

GG

## BMI vs. recessive logistic model, adjusted for year of birth and SEX.

summary(glm(mydf$Recessive ~ mydf$BMI + mydf$Age + mydf$SEX, family = binomial, data = mydf))

## Call:

## glm(formula = mydf$Recessive ~ mydf$BMI + mydf$Age + mydf$SEX,

## family = binomial, data = mydf)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -3.2251 0.1127 0.1377 0.1601 0.2675

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) 2.060997 0.998425 2.064 0.03899 \*

## mydf$BMI 0.130099 0.044079 2.951 0.00316 \*\*

## mydf$Age -0.008356 0.010768 -0.776 0.43775

## mydf$SEX -0.183445 0.312723 -0.587 0.55747

## ---

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 493.31 on 4424 degrees of freedom

## Residual deviance: 483.38 on 4421 degrees of freedom

## AIC: 491.38

##

## Number of Fisher Scoring iterations: 8

**## Result: low p value for BMI, and main effects of age and sex are not significant - and these adjustors are now removed:**

## BMI vs. recessive logistic model, adjustments removed.

summary(glm(mydf$Recessive ~ mydf$BMI, family = binomial, data = mydf))

## Call:

## glm(formula = mydf$Recessive ~ mydf$BMI, family = binomial, data = mydf)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -3.2559 0.1147 0.1372 0.1598 0.2526

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) 1.83481 0.94785 1.936 0.05290 .

## mydf$BMI 0.11377 0.03995 2.847 0.00441 \*\*

## ---

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 493.31 on 4424 degrees of freedom

## Residual deviance: 484.27 on 4423 degrees of freedom

## AIC: 488.27

##

## Number of Fisher Scoring iterations: 8

**## Result: this gives the final low p value for BMI with the recessive model.**

## BMI vs. dominant model, adjusted for year of birth, SEX.

summary(glm(mydf$Dominant ~ mydf$BMI + mydf$Age + mydf$SEX, family = binomial, data = mydf))

## Call:

## glm(formula = mydf$Dominant ~ mydf$BMI + mydf$Age + mydf$SEX,

## family = binomial, data = mydf)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -0.7134 -0.6385 -0.6136 -0.5739 2.0571

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -0.900294 0.247965 -3.631 0.000283 \*\*\*

## mydf$BMI -0.027219 0.009760 -2.789 0.005291 \*\*

## mydf$Age 0.001324 0.002848 0.465 0.641995

## mydf$SEX -0.013993 0.080625 -0.174 0.862210

## ---

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 4105.9 on 4424 degrees of freedom

## Residual deviance: 4097.2 on 4421 degrees of freedom

## AIC: 4105.2

##

## Number of Fisher Scoring iterations: 4

**## Result: BMI p-value for Dominant model highly signficant, almost as low as recessive model with adjustors. Main effects of Age and SEX are not significant and these adjustors are removed below:**

## BMI vs. dominant model, adjustors removed.

summary(glm(mydf$Dominant ~ mydf$BMI, family = binomial, data = mydf))

## Call:

## glm(formula = mydf$Dominant ~ mydf$BMI, family = binomial, data = mydf)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -0.7067 -0.6382 -0.6141 -0.5744 2.0625

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -0.895394 0.229694 -3.898 9.69e-05 \*\*\*

## mydf$BMI -0.026014 0.009058 -2.872 0.00408 \*\*

## ---

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 4105.9 on 4424 degrees of freedom

## Residual deviance: 4097.5 on 4423 degrees of freedom

## AIC: 4101.5

##

## Number of Fisher Scoring iterations: 4

**## Result: Final low p value for dominant model is lower than for recessive model.**

## BMI vs. heterozygote model, adjusted for year of birth, SEX.

summary(glm(mydf$Heterozygote ~ mydf$BMI + mydf$Age + mydf$SEX, family = binomial, data = mydf))

## Call:

## glm(formula = mydf$Heterozygote ~ mydf$BMI + mydf$Age + mydf$SEX,

## family = binomial, data = mydf)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -0.6700 -0.6147 -0.5958 -0.5678 2.0462

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -1.1010704 0.2529082 -4.354 1.34e-05 \*\*\*

## mydf$BMI -0.0205800 0.0098913 -2.081 0.0375 \*

## mydf$Age 0.0008446 0.0029148 0.290 0.7720

## mydf$SEX -0.0245934 0.0824448 -0.298 0.7655

## ---

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 3966.5 on 4424 degrees of freedom

## Residual deviance: 3961.4 on 4421 degrees of freedom

## AIC: 3969.4

##

## Number of Fisher Scoring iterations: 4

**## Result: Heterozygote model significant with adjustments for Age and SEX - adjustor main effects are not significant - and can be removed:**

## BMI vs. heterozygote model, adjustors removed.

summary(glm(mydf$Heterozygote ~ mydf$BMI, family = binomial, data = mydf))

## Call:

## glm(formula = mydf$Heterozygote ~ mydf$BMI, family = binomial,

## data = mydf)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -0.6653 -0.6149 -0.5965 -0.5682 2.0485

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -1.113745 0.233854 -4.763 1.91e-06 \*\*\*

## mydf$BMI -0.020104 0.009195 -2.186 0.0288 \*

## ---

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 3966.5 on 4424 degrees of freedom

## Residual deviance: 3961.6 on 4423 degrees of freedom

## AIC: 3965.6

##

## Number of Fisher Scoring iterations: 4

**## Result: Heterozygote model is significant.**

## BMI vs. additive model, adjusted for year of birth, SEX - linear regression.

Additive <- as.factor(Additive)

## I The coefficient of determination (r 2 ) of an additive linear regression model gives an estimate of the proportion of phenotypic variation that is explained by the SNP (or SNPs) in the model, e.g., the ”SNP heritability”

## http://faculty.washington.edu/tathornt/SISG2017/lectures/SISG2017session02.pdf

Anova(lm(mydf$BMI ~ mydf$Additive + mydf$Age + mydf$SEX), type = "II", white.adjust = TRUE)

## Coefficient covariances computed by hccm()

## Analysis of Deviance Table (Type II tests)

##

## Response: mydf$BMI

## Df F Pr(>F)

## mydf$Additive 2 10.131 4.076e-05 \*\*\*

## mydf$Age 1 543.209 < 2.2e-16 \*\*\*

## mydf$SEX 1 126.717 < 2.2e-16 \*\*\*

## Residuals 4420

## ---

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

**## Result: additive model is significant - but Age and Sex main effects are also very significant - therefore assess interactions:**

Anova(lm(mydf$BMI ~ mydf$Additive \* mydf$Age \* mydf$SEX), type = "II", white.adjust = TRUE)

**##** Coefficient covariances computed by hccm()

## Analysis of Deviance Table (Type II tests)

##

## Response: mydf$BMI

## Df F Pr(>F)

## mydf$Additive 2 10.1872 3.854e-05 \*\*\*

## mydf$Age 1 518.6202 < 2.2e-16 \*\*\*

## mydf$SEX 1 161.0734 < 2.2e-16 \*\*\*

## mydf$Additive:mydf$Age 2 0.2551 0.7749

## mydf$Additive:mydf$SEX 2 0.5169 0.5964

## mydf$Age:mydf$SEX 1 58.9407 1.988e-14 \*\*\*

## mydf$Additive:mydf$Age:mydf$SEX 2 2.0019 0.1352

## Residuals 4413

## ---

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

**## Result: Age and Sex interactions with BMI are not significant - therefore remove interactions, but not adjustors**

Anova(lm(mydf$BMI ~ mydf$Additive + mydf$Age + mydf$SEX), type = "II", white.adjust = TRUE)

## Coefficient covariances computed by hccm()

## Analysis of Deviance Table (Type II tests)

##

## Response: mydf$BMI

## Df F Pr(>F)

## mydf$Additive 2 10.131 4.076e-05 \*\*\*

## mydf$Age 1 543.209 < 2.2e-16 \*\*\*

## mydf$SEX 1 126.717 < 2.2e-16 \*\*\*

## Residuals 4420

## ---

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

summary(lm(mydf$BMI ~ mydf$Additive + mydf$Age + mydf$SEX))

## Call:

## lm(formula = mydf$BMI ~ mydf$Additive + mydf$Age + mydf$SEX)

##

## Residuals:

## Min 1Q Median 3Q Max

## -11.508 -2.870 -0.556 2.363 36.397

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 19.013083 0.274758 69.199 < 2e-16 \*\*\*

## mydf$Additive1 -0.379329 0.171163 -2.216 0.02673 \*

## mydf$Additive2 -1.920195 0.640589 -2.998 0.00274 \*\*

## mydf$Age 0.101075 0.004305 23.480 < 2e-16 \*\*\*

## mydf$SEX 1.424091 0.127107 11.204 < 2e-16 \*\*\*

## ---

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

##

## Residual standard error: 4.224 on 4420 degrees of freedom

## Multiple R-squared: 0.1333, Adjusted R-squared: 0.1325

## F-statistic: 169.9 on 4 and 4420 DF, p-value: < 2.2e-16

**## Result: BMI addiitive model is significant, after adjusting for SEX and Age. r-squared is 0.1325 giving a high heritability.**

**## BMI and Age DATA FOR TABLE 1:**

Additivedf <- data.frame(Additive, mydf[ , "BMI"], mydf[ , "Age"])

colnames(Additivedf) <- c("Additive", "BMI", "Age")

head(Additivedf)

Additivedf\_0 <- Additivedf[Additivedf$Additive == 0, ]

head(Additivedf\_0)

Additivedf\_1 <- Additivedf[Additivedf$Additive == 1, ]

head(Additivedf\_1)

Additivedf\_2 <- Additivedf[Additivedf$Additive == 2, ]

head(Additivedf\_2)

length(Additivedf\_0$BMI)

length(Additivedf\_1$BMI)

length(Additivedf\_2$BMI)

length(Additivedf$BMI)

mean(Additivedf\_0$BMI)

mean(Additivedf\_1$BMI)

mean(Additivedf\_2$BMI)

mean(Additivedf$BMI)

sd(Additivedf\_0$BMI)

sd(Additivedf\_1$BMI)

sd(Additivedf\_2$BMI)

sd(Additivedf$BMI)

median(Additivedf\_0$BMI)

median(Additivedf\_1$BMI)

median(Additivedf\_2$BMI)

median(Additivedf$BMI)

mad(Additivedf\_0$BMI, constant = 1)

mad(Additivedf\_1$BMI, constant = 1)

mad(Additivedf\_2$BMI, constant = 1)

mad(Additivedf$BMI, constant = 1)

length(Additivedf\_0$Age)

length(Additivedf\_1$Age)

length(Additivedf\_2$Age)

length(Additivedf$Age)

mean(Additivedf\_0$Age)

mean(Additivedf\_1$Age)

mean(Additivedf\_2$Age)

mean(Additivedf$Age)

sd(Additivedf\_0$Age)

sd(Additivedf\_1$Age)

sd(Additivedf\_2$Age)

sd(Additivedf$Age)

median(Additivedf\_0$Age)

median(Additivedf\_1$Age)

median(Additivedf\_2$Age)

median(Additivedf$Age)

mad(Additivedf\_0$Age, constant = 1)

mad(Additivedf\_1$Age, constant = 1)

mad(Additivedf\_2$Age, constant = 1)

mad(Additivedf$Age, constant = 1)

## MALES - rs67047829 association with body mass Index.

**## MALES ONLY, n = 2933.**

MALEdf2 <- mydf2[mydf2$SEX == 1, ]

nrow(MALEdf2)

levels(as.factor(MALEdf2[ , "Age"]))

MALEdf2[ , "Age"] <- as.numeric(MALEdf2[ , "Age"])

sum(is.na(MALEdf2[ , "Age"]))

levels(as.factor(MALEdf2[ , "BMI"]))

MALEdf2[ , "BMI"] <- as.numeric(MALEdf2[ , "BMI"])

sum(is.na(MALEdf2[ , "BMI"]))

levels(as.factor(MALEdf2[ , "SEX"]))

sum(is.na(MALEdf2[ , "SEX"]))

rs67047829 <- paste0(MALEdf2[ , "rs67047829\_1"], MALEdf2[ , "rs67047829\_2"])

MALEdf2$rs67047829 <- rs67047829

levels(as.factor(MALEdf2[ , "rs67047829"]))

sum(is.na(MALEdf2[ , "rs67047829"]))

## Counts of genotypes - MALEdf2:

levels(as.factor(MALEdf2[ , "rs67047829"]))

GenoMALEdf2 <- MALEdf2[ , "rs67047829"]

GenoMALEdf2\_AA <- GenoMALEdf2[GenoMALEdf2 == "AA"]

head(GenoMALEdf2\_AA)

length(GenoMALEdf2\_AA)

GenoMALEdf2\_GG <- GenoMALEdf2[GenoMALEdf2 == "GG"]

head(GenoMALEdf2\_GG)

length(GenoMALEdf2\_GG)

GenoMALEdf2\_GA <- GenoMALEdf2[GenoMALEdf2 == "GA"]

head(GenoMALEdf2\_GA)

length(GenoMALEdf2\_GA)

GenoMALEdf2\_AG <- GenoMALEdf2[GenoMALEdf2 == "AG"]

head(GenoMALEdf2\_AG)

length(GenoMALEdf2\_AG)

## Remove subject if data missing from Age, BMI, SEX or rs67047829:

MALEdf1 <- MALEdf2[complete.cases(MALEdf2[ , c("Age", "BMI", "SEX", "rs67047829")]), ]

head(MALEdf1)

nrow(MALEdf1)

colnames(MALEdf1)

## Counts of genotypes - MALEdf1

levels(as.factor(MALEdf1[ , "rs67047829"]))

GenoMALEdf1 <- MALEdf1[ , "rs67047829"]

GenoMALEdf1\_AA <- GenoMALEdf1[GenoMALEdf1 == "AA"]

head(GenoMALEdf1\_AA)

length(GenoMALEdf1\_AA)

GenoMALEdf1\_GG <- GenoMALEdf1[GenoMALEdf1 == "GG"]

head(GenoMALEdf1\_GG)

length(GenoMALEdf1\_GG)

GenoMALEdf1\_GA <- GenoMALEdf1[GenoMALEdf1 == "GA"]

head(GenoMALEdf1\_GA)

length(GenoMALEdf1\_GA)

GenoMALEdf1\_AG <- GenoMALEdf1[GenoMALEdf1 == "AG"]

head(GenoMALEdf1\_AG)

length(GenoMALEdf1\_AG)

sum(is.na(MALEdf1[ , "Age"]))

sum(is.na(MALEdf1[ , "BMI"]))

sum(is.na(MALEdf1[ , "SEX"]))

sum(is.na(MALEdf1[ , "rs67047829"]))

## Find subjects with more than a certain percentage of genotype fails:

NoPhendf1 <- MALEdf1[ , 13:ncol(MALEdf1)]

head(NoPhendf1)

ncol(NoPhendf1)

nrow(NoPhendf1)

NAs <- rowSums(is.na(NoPhendf1))

max(NAs)

NAtrue <- list();

for (i in 1:length(NAs)) {

if (NAs[[i]] < 8) { **## <8 missing gives <5% missing SNP values**

NAtrue[[i]] <- TRUE

} else {

NAtrue[[i]] <- FALSE

}

}

NAtrue <- as.vector(unlist(NAtrue))

MALEdf <- MALEdf1[NAtrue, ]

nrow(MALEdf)

ncol(MALEdf)

MALEdf[ , "BMI"] <- as.numeric(MALEdf[ , "BMI"])

MALEdf[ , "Age"] <- as.numeric(MALEdf[ , "Age"])

## Genotype <- paste0(MALEdf[ , "rs67047829\_1"], MALEdf[ , "rs67047829\_2"])

levels(as.factor(MALEdf[ , "rs67047829"]))

GenoMALEdf <- MALEdf[ , "rs67047829"]

GenoMALEdf\_AA <- GenoMALEdf[GenoMALEdf == "AA"]

head(GenoMALEdf\_AA)

length(GenoMALEdf\_AA)

GenoMALEdf\_GG <- GenoMALEdf[GenoMALEdf == "GG"]

head(GenoMALEdf\_GG)

length(GenoMALEdf\_GG)

GenoMALEdf\_GA <- GenoMALEdf[GenoMALEdf == "GA"]

head(GenoMALEdf\_GA)

length(GenoMALEdf\_GA)

GenoMALEdf\_AG <- GenoMALEdf[GenoMALEdf == "AG"]

head(GenoMALEdf\_AG)

length(GenoMALEdf\_AG)

Genotype <- MALEdf[ , "rs67047829"] ## Genotype given as e.g. "GG"

levels(as.factor(Genotype))

MALEdf$Genotype <- Genotype

head(MALEdf)

**## Dominant model coding:**

Dominant <- Genotype

Dominant[(Genotype == "GA") | (Genotype == "AG") | (Genotype == "AA")] <- 1

Dominant[Genotype == "GG"] <- 0

MALEdf$Dominant <- as.factor(Dominant)

**## Recessive model coding:**

Recessive <- Genotype

Recessive[Genotype == "AA"] <- 0

Recessive[(Genotype == "GA") | (Genotype == "AG") | (Genotype == "GG")] <- 1

MALEdf$Recessive <- as.factor(Recessive)

**## Heterozygote model coding:**

Heterozygote <- Genotype

Heterozygote[(Genotype == "GA") | (Genotype == "AG")] <- 1

Heterozygote[(Genotype == "GG") | (Genotype == "AA")] <- 0

MALEdf$Heterozygote <- as.factor(Heterozygote)

**## Additive model coding:**

Additive <- Genotype

Additive[Genotype == "GG"] <- 0

Additive[(Genotype == "GA") | (Genotype == "AG")] <- 1

Additive[Genotype == "AA"] <- 2

MALEdf$Additive <- as.factor(Additive)

levels(as.factor(Genotype))

levels(as.factor(MALEdf[ , "BMI"]))

levels(as.factor(MALEdf[ , "SEX"]))

levels(as.factor(MALEdf[ , "Age"]))

levels(as.factor(Dominant))

levels(as.factor(Recessive))

levels(as.factor(Heterozygote))

levels(as.factor(Additive))

AA <- Genotype[Genotype == "AA"]

AA

AG <- Genotype[Genotype == "AG"]

AG

GA <- Genotype[Genotype == "GA"]

GA

GG <- Genotype[Genotype == "GG"]

GG

## BMI vs. recessive logistic model, adjusted for year of birth.

summary(glm(MALEdf$Recessive ~ MALEdf$BMI + MALEdf$Age, family = binomial, data = MALEdf))

## Call:

## glm(formula = MALEdf$Recessive ~ MALEdf$BMI + MALEdf$Age, family = binomial,

## data = MALEdf)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -3.2374 0.1057 0.1361 0.1626 0.3211

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) 1.71665 1.30792 1.313 0.1894

## MALEdf$BMI 0.15732 0.06208 2.534 0.0113 \*

## MALEdf$Age -0.01909 0.01510 -1.265 0.2060

## **---**

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 257.35 on 2286 degrees of freedom

## Residual deviance: 249.80 on 2284 degrees of freedom

## AIC: 255.8

##

## Number of Fisher Scoring iterations: 8

**## Male result: BMI p value significant, and main effect of age is not significant - and this adjustor is now removed:**

## BMI vs. recessive logistic model, adjustment removed.

summary(glm(MALEdf$Recessive ~ MALEdf$BMI, family = binomial, data = MALEdf))

## Call:

## glm(formula = MALEdf$Recessive ~ MALEdf$BMI, family = binomial,

## data = MALEdf)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -3.2852 0.1096 0.1391 0.1634 0.2583

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) 1.59456 1.28985 1.236 0.2164

## MALEdf$BMI 0.12768 0.05671 2.252 0.0243 \*

## **---**

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 257.35 on 2286 degrees of freedom

## Residual deviance: 251.35 on 2285 degrees of freedom

## AIC: 255.35

##

## Number of Fisher Scoring iterations: 8

##

## **Male result: this gives the final BMI p value for the recessive model - significant.**

## BMI vs. dominant model, adjusted for year of birth.

summary(glm(MALEdf$Dominant ~ MALEdf$BMI + MALEdf$Age, family = binomial, data = MALEdf))

## Call:

## glm(formula = MALEdf$Dominant ~ MALEdf$BMI + MALEdf$Age, family = binomial,

## data = MALEdf)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -0.7394 -0.6507 -0.6203 -0.5623 2.1022

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -0.776201 0.298567 -2.600 0.00933 \*\*

## MALEdf$BMI -0.034424 0.013171 -2.614 0.00896 \*\*

## MALEdf$Age 0.002189 0.004040 0.542 0.58806

## **---**

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 2148.1 on 2286 degrees of freedom

## Residual deviance: 2140.6 on 2284 degrees of freedom

## AIC: 2146.6

##

## Number of Fisher Scoring iterations: 4

##

## **Male result. BMI p-value for Dominant model much lower than recessive model with adjustor. ## Main effect of Age not significant and this adjustor is removed below:**

## BMI vs. dominant model, adjustors removed.

summary(glm(MALEdf$Dominant ~ MALEdf$BMI, family = binomial, data = MALEdf))

## Call:

## glm(formula = MALEdf$Dominant ~ MALEdf$BMI, family = binomial,

## data = MALEdf)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -0.7270 -0.6509 -0.6202 -0.5629 2.1017

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -0.75437 0.29512 -2.556 0.01058 \*

## MALEdf$BMI -0.03151 0.01198 -2.630 0.00854 \*\*

## **---**

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 2148.1 on 2286 degrees of freedom

## Residual deviance: 2140.9 on 2285 degrees of freedom

## AIC: 2144.9

##

## Number of Fisher Scoring iterations: 4

##

## **Male result: Final low BMI p value for dominant model is lower than for recessive model.**

## BMI vs. heterozygote model, adjusted for year of birth.

summary(glm(MALEdf$Heterozygote ~ MALEdf$BMI + MALEdf$Age, family = binomial, data = MALEdf))

## Call:

## glm(formula = MALEdf$Heterozygote ~ MALEdf$BMI + MALEdf$Age,

## family = binomial, data = MALEdf)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -0.6887 -0.6270 -0.6034 -0.5598 2.0846

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -0.982118 0.303002 -3.241 0.00119 \*\*

## MALEdf$BMI -0.026753 0.013306 -2.011 0.04436 \*

## MALEdf$Age 0.001001 0.004133 0.242 0.80864

## **---**

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 2076.4 on 2286 degrees of freedom

## Residual deviance: 2071.7 on 2284 degrees of freedom

## AIC: 2077.7

##

## Number of Fisher Scoring iterations: 4

##

## **Male result. Heterozygote model significant with adjustment for Age - adjustor main effect not significant - and can be removed:**

## BMI vs. heterozygote model, adjustor removed.

summary(glm(MALEdf$Heterozygote ~ MALEdf$BMI, family = binomial, data = MALEdf))

## Call:

## glm(formula = MALEdf$Heterozygote ~ MALEdf$BMI, family = binomial,

## data = MALEdf)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -0.6856 -0.6268 -0.6036 -0.5594 2.0846

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -0.97182 0.29966 -3.243 0.00118 \*\*

## MALEdf$BMI -0.02544 0.01212 -2.098 0.03591 \*

## **---**

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 2076.4 on 2286 degrees of freedom

## Residual deviance: 2071.8 on 2285 degrees of freedom

## AIC: 2075.8

##

## Number of Fisher Scoring iterations: 4

##

## **Male result. Heterozygote model is significant.**

## BMI vs. additive model, adjusted for year of birth - linear regression.

Additive <- as.factor(Additive)

## I The coefficient of determination (r 2 ) of an additive linear regression model gives an estimate of the proportion of phenotypic variation that is explained by the SNP (or SNPs) in the model, e.g., the ”SNP heritability”

## http://faculty.washington.edu/tathornt/SISG2017/lectures/SISG2017session02.pdf

## Anova(lm(MALEdf$BMI ~ MALEdf$Additive + MALEdf$Age), type = "II", white.adjust = TRUE)

##

## Coefficient covariances computed by hccm()

## Analysis of Deviance Table (Type II tests)

##

## Response: MALEdf$BMI

## Df F Pr(>F)

## MALEdf$Additive 2 7.8007 0.0004204 \*\*\*

## MALEdf$Age 1 447.0157 < 2.2e-16 \*\*\*

## Residuals 2283

## **---**

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

##

## **Male result: additive model is highly significant - but Age main effects is also very significant - therefore assess interaction:**

Anova(lm(MALEdf$BMI ~ MALEdf$Additive \* MALEdf$Age), type = "II", white.adjust = TRUE)

## Coefficient covariances computed by hccm()

## Analysis of Deviance Table (Type II tests)

##

## Response: MALEdf$BMI

## Df F Pr(>F)

## MALEdf$Additive 2 8.1399 0.0003002 \*\*\*

## MALEdf$Age 1 444.4754 < 2.2e-16 \*\*\*

## MALEdf$Additive:MALEdf$Age 2 1.5387 0.2148906

## Residuals 2281

## **---**

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

##

**## Male result: Age interaction with BMI is not significant - therefore remove interaction, but not adjustor:**

Anova(lm(MALEdf$BMI ~ MALEdf$Additive + MALEdf$Age), type = "II", white.adjust = TRUE)

## Coefficient covariances computed by hccm()

## Analysis of Deviance Table (Type II tests)

##

## Response: MALEdf$BMI

## Df F Pr(>F)

## MALEdf$Additive 2 7.8007 0.0004204 \*\*\*

## MALEdf$Age 1 447.0157 < 2.2e-16 \*\*\*

## Residuals 2283

## **---**

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

##

##

## **Male result: final BMI p value for additive model.**

summary(lm(MALEdf$BMI ~ MALEdf$Additive + MALEdf$Age))

## Call:

## lm(formula = MALEdf$BMI ~ MALEdf$Additive + MALEdf$Age)

##

## Residuals:

## Min 1Q Median 3Q Max

## -11.986 -3.021 -0.755 2.405 36.076

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 19.132737 0.284458 67.260 < 2e-16 \*\*\*

## MALEdf$Additive1 -0.524337 0.246741 -2.125 0.03369 \*

## MALEdf$Additive2 -2.397842 0.926156 -2.589 0.00969 \*\*

## MALEdf$Age 0.132333 0.006212 21.304 < 2e-16 \*\*\*

## **---**

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

##

## Residual standard error: 4.415 on 2283 degrees of freedom

## Multiple R-squared: 0.1694, Adjusted R-squared: 0.1683

## F-statistic: 155.2 on 3 and 2283 DF, p-value: < 2.2e-16

##

## **Male result: BMI addiitive model is significant, after adjusting for Age. r-squared is 0.1683 giving a high heritability.**

**## BMI and Age DATA FOR TABLE 1: MALES**

Additivedf <- data.frame(Additive, MALEdf[ , "BMI"], MALEdf[ , "Age"])

colnames(Additivedf) <- c("Additive", "BMI", "Age")

head(Additivedf)

Additivedf\_0 <- Additivedf[Additivedf$Additive == 0, ]

head(Additivedf\_0)

Additivedf\_1 <- Additivedf[Additivedf$Additive == 1, ]

head(Additivedf\_1)

Additivedf\_2 <- Additivedf[Additivedf$Additive == 2, ]

head(Additivedf\_2)

length(Additivedf\_0$BMI)

length(Additivedf\_1$BMI)

length(Additivedf\_2$BMI)

length(Additivedf$BMI)

mean(Additivedf\_0$BMI)

mean(Additivedf\_1$BMI)

mean(Additivedf\_2$BMI)

mean(Additivedf$BMI)

sd(Additivedf\_0$BMI)

sd(Additivedf\_1$BMI)

sd(Additivedf\_2$BMI)

sd(Additivedf$BMI)

median(Additivedf\_0$BMI)

median(Additivedf\_1$BMI)

median(Additivedf\_2$BMI)

median(Additivedf$BMI)

mad(Additivedf\_0$BMI, constant = 1)

mad(Additivedf\_1$BMI, constant = 1)

mad(Additivedf\_2$BMI, constant = 1)

mad(Additivedf$BMI, constant = 1)

length(Additivedf\_0$Age)

length(Additivedf\_1$Age)

length(Additivedf\_2$Age)

length(Additivedf$Age)

mean(Additivedf\_0$Age)

mean(Additivedf\_1$Age)

mean(Additivedf\_2$Age)

mean(Additivedf$Age)

sd(Additivedf\_0$Age)

sd(Additivedf\_1$Age)

sd(Additivedf\_2$Age)

sd(Additivedf$Age)

median(Additivedf\_0$Age)

median(Additivedf\_1$Age)

median(Additivedf\_2$Age)

median(Additivedf$Age)

mad(Additivedf\_0$Age, constant = 1)

mad(Additivedf\_1$Age, constant = 1)

mad(Additivedf\_2$Age, constant = 1)

mad(Additivedf$Age, constant = 1)

## FEMALES - rs67047829 association with body mass Index.

**## FEMALES ONLY, N = 2824.**

femdf2 <- mydf2[mydf2$SEX == 2, ]

nrow(femdf2)

levels(as.factor(femdf2[ , "Age"]))

femdf2[ , "Age"] <- as.numeric(femdf2[ , "Age"])

sum(is.na(femdf2[ , "Age"]))

levels(as.factor(femdf2[ , "BMI"]))

femdf2[ , "BMI"] <- as.numeric(femdf2[ , "BMI"])

sum(is.na(femdf2[ , "BMI"]))

levels(as.factor(femdf2[ , "SEX"]))

sum(is.na(femdf2[ , "SEX"]))

rs67047829 <- paste0(femdf2[ , "rs67047829\_1"], femdf2[ , "rs67047829\_2"])

femdf2$rs67047829 <- rs67047829

levels(as.factor(femdf2[ , "rs67047829"]))

sum(is.na(femdf2[ , "rs67047829"]))

## Counts of genotypes - femdf2:

levels(as.factor(femdf2[ , "rs67047829"]))

Genofemdf2 <- femdf2[ , "rs67047829"]

Genofemdf2\_AA <- Genofemdf2[Genofemdf2 == "AA"]

head(Genofemdf2\_AA)

length(Genofemdf2\_AA)

Genofemdf2\_GG <- Genofemdf2[Genofemdf2 == "GG"]

head(Genofemdf2\_GG)

length(Genofemdf2\_GG)

Genofemdf2\_GA <- Genofemdf2[Genofemdf2 == "GA"]

head(Genofemdf2\_GA)

length(Genofemdf2\_GA)

Genofemdf2\_AG <- Genofemdf2[Genofemdf2 == "AG"]

head(Genofemdf2\_AG)

length(Genofemdf2\_AG)

## Remove subject if data missing from Age, BMI, SEX or rs67047829:

femdf1 <- femdf2[complete.cases(femdf2[ , c("Age", "BMI", "SEX", "rs67047829")]), ]

head(femdf1)

nrow(femdf1)

colnames(femdf1)

## Counts of genotypes - femdf1

levels(as.factor(femdf1[ , "rs67047829"]))

Genofemdf1 <- femdf1[ , "rs67047829"]

Genofemdf1\_AA <- Genofemdf1[Genofemdf1 == "AA"]

head(Genofemdf1\_AA)

length(Genofemdf1\_AA)

Genofemdf1\_GG <- Genofemdf1[Genofemdf1 == "GG"]

head(Genofemdf1\_GG)

length(Genofemdf1\_GG)

Genofemdf1\_GA <- Genofemdf1[Genofemdf1 == "GA"]

head(Genofemdf1\_GA)

length(Genofemdf1\_GA)

Genofemdf1\_AG <- Genofemdf1[Genofemdf1 == "AG"]

head(Genofemdf1\_AG)

length(Genofemdf1\_AG)

sum(is.na(femdf1[ , "Age"]))

sum(is.na(femdf1[ , "BMI"]))

sum(is.na(femdf1[ , "SEX"]))

sum(is.na(femdf1[ , "rs67047829"]))

## Find subjects with more than a certain percentage of genotype fails:

NoPhendf1 <- femdf1[ , 13:ncol(femdf1)]

head(NoPhendf1)

ncol(NoPhendf1)

nrow(NoPhendf1)

NAs <- rowSums(is.na(NoPhendf1))

max(NAs)

NAtrue <- list();

for (i in 1:length(NAs)) {

if (NAs[[i]] < 8) { **## <8 missing gives <5% missing SNP values**

NAtrue[[i]] <- TRUE

} else {

NAtrue[[i]] <- FALSE

}

}

NAtrue <- as.vector(unlist(NAtrue))

femdf <- femdf1[NAtrue, ]

nrow(femdf)

ncol(femdf)

femdf[ , "BMI"] <- as.numeric(femdf[ , "BMI"])

femdf[ , "Age"] <- as.numeric(femdf[ , "Age"])

## Genotype <- paste0(femdf[ , "rs67047829\_1"], femdf[ , "rs67047829\_2"])

levels(as.factor(femdf[ , "rs67047829"]))

Genofemdf <- femdf[ , "rs67047829"]

Genofemdf\_AA <- Genofemdf[Genofemdf == "AA"]

head(Genofemdf\_AA)

length(Genofemdf\_AA)

Genofemdf\_GG <- Genofemdf[Genofemdf == "GG"]

head(Genofemdf\_GG)

length(Genofemdf\_GG)

Genofemdf\_GA <- Genofemdf[Genofemdf == "GA"]

head(Genofemdf\_GA)

length(Genofemdf\_GA)

Genofemdf\_AG <- Genofemdf[Genofemdf == "AG"]

head(Genofemdf\_AG)

length(Genofemdf\_AG)

Genotype <- femdf[ , "rs67047829"] ## Genotype given as e.g. "GG"

levels(as.factor(Genotype))

femdf$Genotype <- Genotype

head(femdf)

**## Dominant model coding:**

Dominant <- Genotype

Dominant[(Genotype == "GA") | (Genotype == "AG") | (Genotype == "AA")] <- 1

Dominant[Genotype == "GG"] <- 0

femdf$Dominant <- as.factor(Dominant)

**## Recessive model coding:**

Recessive <- Genotype

Recessive[Genotype == "AA"] <- 0

Recessive[(Genotype == "GA") | (Genotype == "AG") | (Genotype == "GG")] <- 1

femdf$Recessive <- as.factor(Recessive)

**## Heterozygote model coding:**

Heterozygote <- Genotype

Heterozygote[(Genotype == "GA") | (Genotype == "AG")] <- 1

Heterozygote[(Genotype == "GG") | (Genotype == "AA")] <- 0

femdf$Heterozygote <- as.factor(Heterozygote)

**## Additive model coding:**

Additive <- Genotype

Additive[Genotype == "GG"] <- 0

Additive[(Genotype == "GA") | (Genotype == "AG")] <- 1

Additive[Genotype == "AA"] <- 2

femdf$Additive <- as.factor(Additive)

levels(as.factor(Genotype))

levels(as.factor(femdf[ , "BMI"]))

levels(as.factor(femdf[ , "SEX"]))

levels(as.factor(femdf[ , "Age"]))

levels(as.factor(Dominant))

levels(as.factor(Recessive))

levels(as.factor(Heterozygote))

levels(as.factor(Additive))

AA <- Genotype[Genotype == "AA"]

AA

AG <- Genotype[Genotype == "AG"]

AG

GA <- Genotype[Genotype == "GA"]

GA

GG <- Genotype[Genotype == "GG"]

GG

## BMI vs. recessive logistic model, adjusted for year of birth.

summary(glm(femdf$Recessive ~ femdf$BMI + femdf$Age, family = binomial, data = femdf))

## Call:

## glm(formula = femdf$Recessive ~ femdf$BMI + femdf$Age, family = binomial,

## data = femdf)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -3.2250 0.1174 0.1371 0.1565 0.2401

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) 1.845898 1.531034 1.206 0.2280

## femdf$BMI 0.106530 0.063463 1.679 0.0932 .

## femdf$Age 0.001967 0.015525 0.127 0.8992

## **---**

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 235.96 on 2137 degrees of freedom

## Residual deviance: 232.56 on 2135 degrees of freedom

## AIC: 238.56

##

## Number of Fisher Scoring iterations: 7

## **Female result: BMI p value not significant, and main effect of age is not significant - and this adjustor is now removed:**

## BMI vs. recessive logistic model, adjustment removed.

summary(glm(femdf$Recessive ~ femdf$BMI, family = binomial, data = femdf))

## Call:

## glm(formula = femdf$Recessive ~ femdf$BMI, family = binomial,

## data = femdf)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -3.2199 0.1171 0.1372 0.1564 0.2376

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) 1.87361 1.51780 1.234 0.2170

## femdf$BMI 0.10863 0.06134 1.771 0.0766 .

## **---**

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 235.96 on 2137 degrees of freedom

## Residual deviance: 232.57 on 2136 degrees of freedom

## AIC: 236.57

##

## Number of Fisher Scoring iterations: 7

## **Female result: this gives the final BMI p value for the recessive model - not significant.**

## BMI vs. dominant model, adjusted for year of birth.

summary(glm(femdf$Dominant ~ femdf$BMI + femdf$Age, family = binomial, data = femdf))

## Call:

## glm(formula = femdf$Dominant ~ femdf$BMI + femdf$Age, family = binomial,

## data = femdf)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -0.6648 -0.6233 -0.6092 -0.5869 1.9971

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -1.1471118 0.3837484 -2.989 0.0028 \*\*

## femdf$BMI -0.0178138 0.0148319 -1.201 0.2297

## femdf$Age 0.0007577 0.0040371 0.188 0.8511

## **---**

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 1957.4 on 2137 degrees of freedom

## Residual deviance: 1955.9 on 2135 degrees of freedom

## AIC: 1961.9

##

## Number of Fisher Scoring iterations: 4

## **Female result. BMI p-value for Dominant model not significant with adjustor. Main effect of Age not significant and this adjustor is removed below:**

## BMI vs. dominant model, adjustors removed.

summary(glm(femdf$Dominant ~ femdf$BMI, family = binomial, data = femdf))

## Call:

## glm(formula = femdf$Dominant ~ femdf$BMI, family = binomial,

## data = femdf)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -0.6646 -0.6228 -0.6085 -0.5871 2.0012

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -1.13287 0.37590 -3.014 0.00258 \*\*

## femdf$BMI -0.01714 0.01438 -1.192 0.23330

## **---**

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 1957.4 on 2137 degrees of freedom

## Residual deviance: 1956.0 on 2136 degrees of freedom

## AIC: 1960

##

## Number of Fisher Scoring iterations: 4

## **Female result: Final p value for dominant model not significant.**

## BMI vs. heterozygote model, adjusted for year of birth.

summary(glm(femdf$Heterozygote ~ femdf$BMI + femdf$Age, family = binomial, data = femdf))

## Call:

## glm(formula = femdf$Heterozygote ~ femdf$BMI + femdf$Age, family = binomial,

## data = femdf)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -0.6285 -0.6001 -0.5914 -0.5774 1.9826

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -1.3822278 0.3917117 -3.529 0.000418 \*\*\*

## femdf$BMI -0.0118302 0.0150849 -0.784 0.432899

## femdf$Age 0.0009827 0.0041312 0.238 0.811987

## **---**

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 1889.7 on 2137 degrees of freedom

## Residual deviance: 1889.1 on 2135 degrees of freedom

## AIC: 1895.1

##

## Number of Fisher Scoring iterations: 3

## **Female result. Heterozygote model not significant with adjustment for Age - adjustor main effect not significant - and can be removed:**

## BMI vs. heterozygote model, adjustor removed.

summary(glm(femdf$Heterozygote ~ femdf$BMI, family = binomial, data = femdf))

## Call:

## glm(formula = femdf$Heterozygote ~ femdf$BMI, family = binomial,

## data = femdf)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -0.6253 -0.5998 -0.5913 -0.5780 1.9880

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -1.36349 0.38333 -3.557 0.000375 \*\*\*

## femdf$BMI -0.01096 0.01463 -0.750 0.453484

## **---**

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

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 1889.7 on 2137 degrees of freedom

## Residual deviance: 1889.1 on 2136 degrees of freedom

## AIC: 1893.1

##

## Number of Fisher Scoring iterations: 3

## **Female result. Heterozygote model is not significant.**

## BMI vs. additive model, adjusted for year of birth - linear regression.

Additive <- as.factor(Additive)

## I The coefficient of determination (r 2 ) of an additive linear regression model gives an estimate of the proportion of phenotypic variation that is explained by the SNP (or SNPs) in the model, e.g., the ”SNP heritability”

## http://faculty.washington.edu/tathornt/SISG2017/lectures/SISG2017session02.pdf

Anova(lm(femdf$BMI ~ femdf$Additive + femdf$Age), type = "II", white.adjust = TRUE)

## Coefficient covariances computed by hccm()

## Analysis of Deviance Table (Type II tests)

##

## Response: femdf$BMI

## Df F Pr(>F)

## femdf$Additive 2 2.7343 0.06517 .

## femdf$Age 1 132.3387 < 2e-16 \*\*\*

## Residuals 2134

## **---**

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

## **Female result: additive model not significant - and Age main effect very significant.**

**## BMI and Age DATA FOR TABLE 1: FEMALES**

Additivedf <- data.frame(Additive, femdf[ , "BMI"], femdf[ , "Age"])

colnames(Additivedf) <- c("Additive", "BMI", "Age")

head(Additivedf)

Additivedf\_0 <- Additivedf[Additivedf$Additive == 0, ]

head(Additivedf\_0)

Additivedf\_1 <- Additivedf[Additivedf$Additive == 1, ]

head(Additivedf\_1)

Additivedf\_2 <- Additivedf[Additivedf$Additive == 2, ]

head(Additivedf\_2)

length(Additivedf\_0$BMI)

length(Additivedf\_1$BMI)

length(Additivedf\_2$BMI)

length(Additivedf$BMI)

mean(Additivedf\_0$BMI)

mean(Additivedf\_1$BMI)

mean(Additivedf\_2$BMI)

mean(Additivedf$BMI)

sd(Additivedf\_0$BMI)

sd(Additivedf\_1$BMI)

sd(Additivedf\_2$BMI)

sd(Additivedf$BMI)

median(Additivedf\_0$BMI)

median(Additivedf\_1$BMI)

median(Additivedf\_2$BMI)

median(Additivedf$BMI)

mad(Additivedf\_0$BMI, constant = 1)

mad(Additivedf\_1$BMI, constant = 1)

mad(Additivedf\_2$BMI, constant = 1)

mad(Additivedf$BMI, constant = 1)

length(Additivedf\_0$Age)

length(Additivedf\_1$Age)

length(Additivedf\_2$Age)

length(Additivedf$Age)

mean(Additivedf\_0$Age)

mean(Additivedf\_1$Age)

mean(Additivedf\_2$Age)

mean(Additivedf$Age)

sd(Additivedf\_0$Age)

sd(Additivedf\_1$Age)

sd(Additivedf\_2$Age)

sd(Additivedf$Age)

median(Additivedf\_0$Age)

median(Additivedf\_1$Age)

median(Additivedf\_2$Age)

median(Additivedf$Age)

mad(Additivedf\_0$Age, constant = 1)

mad(Additivedf\_1$Age, constant = 1)

mad(Additivedf\_2$Age, constant = 1)

mad(Additivedf$Age, constant = 1)

## Calculation of odds ratios:

**## MALES AND FEMALES TOGETHER:**

OverweightObesevec <- mydf1[(mydf1$BMI\_GROUP == 3) | (mydf1$BMI\_GROUP == 4), "rs67047829"]

OverweightObesevec

NormalUnderweightvec <- mydf1[(mydf1$BMI\_GROUP == 1) | (mydf1$BMI\_GROUP == 2), "rs67047829"]

NormalUnderweightvec

length(OverweightObesevec)

length(NormalUnderweightvec)

mydf1BMI\_GROUP <- as.numeric(mydf1$BMI\_GROUP)

mydf1BMI\_GROUP

mydf1BMI\_GROUPnoNA <- mydf1BMI\_GROUP[!is.na(mydf1BMI\_GROUP)]

length(mydf1BMI\_GROUPnoNA)

length(OverweightObesevec) + length(NormalUnderweightvec)

OverweightObeseAAvec <- OverweightObesevec[OverweightObesevec == "AA"]

OverweightObeseAAvec

OverweightObeseGAGGvec <- OverweightObesevec[(OverweightObesevec == "GA") | (OverweightObesevec == "AG") | (OverweightObesevec == "GG")]

OverweightObeseGAGGvec

NormalUnderweightAAvec <- NormalUnderweightvec[NormalUnderweightvec == "AA"]

NormalUnderweightAAvec

NormalUnderweightGAGGvec <- NormalUnderweightvec[(NormalUnderweightvec == "GA") | (NormalUnderweightvec == "AG") | (NormalUnderweightvec == "GG")]

NormalUnderweightGAGGvec

length(OverweightObeseAAvec) + length(OverweightObeseGAGGvec) + length(NormalUnderweightAAvec) + length(NormalUnderweightGAGGvec)

OddsRatio <- (length(OverweightObeseAAvec) / length(OverweightObeseGAGGvec)) / (length(NormalUnderweightAAvec) / length(NormalUnderweightGAGGvec))

OddsRatio

## 0.4299128 - significant, see below:

TAB\_MC <- matrix(c(length(OverweightObeseAAvec), length(OverweightObeseGAGGvec), length(NormalUnderweightAAvec), length(NormalUnderweightGAGGvec)),

nrow = 2,

dimnames = list(c("AA", "GA+GG"),

c("OverweightObese", "NormalUnderweight")))

TAB\_MC

or\_fit <- oddsratio(TAB\_MC)

or\_fit

## $data

## OverweightObese NormalUnderweight Total

## AA 16 40 56

## GA+GG 2407 2587 4994

## Total 2423 2627 5050

##

## $measure

## NA

## odds ratio **with** 95% C.I. estimate lower upper

## AA 1.0000000 NA NA

## GA+GG 0.4326035 0.2341035 0.7611917

##

## $p.value

## NA

## two-sided midp.exact fisher.exact chi.square

## AA NA NA NA

## GA+GG 0.003242213 0.004324409 0.003461579

##

## $correction

## [1] FALSE

##

## attr(,"method")

## [1] "median-unbiased estimate & mid-p exact CI"

**## MALES ONLY**

Malesdf <- mydf1[mydf1$SEX == 1, ]

head(Malesdf)

OverweightMALEObesevec <- Malesdf[(Malesdf$BMI\_GROUP == 3) | (Malesdf$BMI\_GROUP == 4), "rs67047829"]

OverweightMALEObesevec

NormalUnderweightMALEvec <- Malesdf[(Malesdf$BMI\_GROUP == 1) | (Malesdf$BMI\_GROUP == 2), "rs67047829"]

NormalUnderweightMALEvec

length(OverweightMALEObesevec)

length(NormalUnderweightMALEvec)

MalesdfBMI\_GROUP <- as.numeric(Malesdf$BMI\_GROUP)

MalesdfBMI\_GROUP

MalesdfBMI\_GROUPnoNA <- MalesdfBMI\_GROUP[!is.na(MalesdfBMI\_GROUP)]

length(MalesdfBMI\_GROUPnoNA)

length(OverweightMALEObesevec) + length(NormalUnderweightMALEvec)

OverweightMALEObeseAAvec <- OverweightMALEObesevec[OverweightMALEObesevec == "AA"]

OverweightMALEObeseAAvec

OverweightMALEObeseGAGGvec <- OverweightMALEObesevec[(OverweightMALEObesevec == "GA") | (OverweightMALEObesevec == "AG") | (OverweightMALEObesevec == "GG")]

OverweightMALEObeseGAGGvec

NormalUnderweightMALEAAvec <- NormalUnderweightMALEvec[NormalUnderweightMALEvec == "AA"]

NormalUnderweightMALEAAvec

NormalUnderweightMALEGAGGvec <- NormalUnderweightMALEvec[(NormalUnderweightMALEvec == "GA") | (NormalUnderweightMALEvec == "AG") | (NormalUnderweightMALEvec == "GG")]

NormalUnderweightMALEGAGGvec

length(OverweightMALEObeseAAvec) + length(OverweightMALEObeseGAGGvec) + length(NormalUnderweightMALEAAvec) + length(NormalUnderweightMALEGAGGvec)

OddsRatio <- (length(OverweightMALEObeseAAvec) / length(OverweightMALEObeseGAGGvec)) / (length(NormalUnderweightMALEAAvec) / length(NormalUnderweightMALEGAGGvec))

OddsRatio

## 0.4923726 - not significant - see below:

MALE\_TAB\_MC <- matrix(c(length(OverweightMALEObeseAAvec), length(OverweightMALEObeseGAGGvec), length(NormalUnderweightMALEAAvec), length(NormalUnderweightMALEGAGGvec)),

nrow = 2,

dimnames = list(c("AA", "GA+GG"),

c("OverweightMALEObese", "NormalUnderweightMALE")))

MALE\_TAB\_MC

MALE\_or\_fit <- oddsratio(MALE\_TAB\_MC)

MALE\_or\_fit

## $data

## OverweightMALEObese NormalUnderweightMALE Total

## AA 8 24 32

## GA+GG 1027 1517 2544

## Total 1035 1541 2576

##

## $measure

## NA

## odds ratio **with** 95% C.I. estimate lower upper

## AA 1.000000 NA NA

## GA+GG 0.499083 0.2072639 1.07489

##

## $p.value

## NA

## two-sided midp.exact fisher.exact chi.square

## AA NA NA NA

## GA+GG 0.07711167 0.101396 0.07800849

##

## $correction

## [1] FALSE

##

## attr(,"method")

## [1] "median-unbiased estimate & mid-p exact CI"