##code for paternal age project

#load relevant packages

library(ggplot2)

library(grid)

library(RColorBrewer)

library(colorRamps)

library(colorspace)

library(cowplot)

library(ggthemes)

#keep aesthetic parameters the same across plots

title.size <- 14

axistitle.size <- 10

axistext.size <- 8

setwd([redacted])

##read in files we need for analyses and perform initial explorations of SSC data:

setwd([redacted])

#read in data from SSC (ASD and unaffected siblings)

cases.whichdata <- read.csv("SSC\_cases\_imputed\_sequenced.csv")

cases.parentage <- read.csv("cases\_parent\_age.csv")

cases.SNVs <- read.csv("case\_counts\_SNVs.csv")

controls.whichdata <- read.csv("SSC\_controls\_imputed\_sequenced.csv")

controls.parentage <- read.csv("controls\_parent\_age.csv")

controls.SNVs <- read.csv("control\_counts\_SNVs.csv")

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

##2) Relationship between paternal age and DNVs

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

##produce an estimate for the rate of increase in dnSNVs with respect to paternal age:

#Determine number of control subjects with both SNV data and parental age data

SNV.IDs <- controls.SNVs[,1] #all individuals with SNV data

parentage.IDs <- controls.parentage[,1][which(controls.parentage[,2] != ".")]

s.IDs <- SNV.IDs[SNV.IDs%in%parentage.IDs]

parentage.index <- which(controls.parentage[,1]%in%s.IDs)

dad.ages <- as.numeric((as.vector(controls.parentage

[parentage.index,]$dadage\_years)))

###12.4 hack to account for sperm not starting to divide until puberty

#dad.ages <- dad.ages - 13

SNVs <- controls.SNVs[parentage.index, c(3, 5, 7, 9, 11, 13, 15, 17)]

SNVs$nonsyn <- SNVs$num\_missense + SNVs$num\_PTV

length(dad.ages) #1827

#nonsynonymous SNVs vs. dad's age

x <- glm(SNVs[,9] ~ dad.ages, family = "poisson")

intercept.nonsyn.dad <- summary(x)$coefficients[1,1]

intercept.SE.nonsyn.dad <- summary(x)$coefficients[1,2]

beta.nonsyn.dad <- summary(x)$coefficients[2,1]

beta.SE.nonsyn.dad <- summary(x)$coefficients[2,2]

intercept.nonsyn.dad #-1.45

exp(intercept.nonsyn.dad) #0.23

beta.nonsyn.dad #0.031

summary(x) #p = 2e-10

#Any SNV (including synonymous) vs. dad's age

x <- glm(SNVs[,10] ~ dad.ages, family = "poisson")

intercept.any\_SNV.dad <- summary(x)$coefficients[1,1]

intercept.SE.any\_SNV.dad <- summary(x)$coefficients[1,2]

beta.any\_SNV.dad <- summary(x)$coefficients[2,1]

beta.SE.any\_SNV.dad <- summary(x)$coefficients[2,2]

intercept.any\_SNV.dad #-1.00

exp(intercept.any\_SNV.dad) #0.37

beta.any\_SNV.dad #0.027

summary(x)

##3) Identify the synonymous rate for each phenoytpe and determine whether it is statistically

#different from the rate among controls

#rate among controls

##should count even those controls for whom paternal age data is unavailable:

dim(controls.SNVs) #n = 1902

sum(controls.SNVs$num\_synonymous) #468

contr.syn.rate <- sum(controls.SNVs$num\_synonymous)/dim(controls.SNVs)[1]; contr.syn.rate #0.25

#rate in ASD

ASD.syn.rate <- sum(cases.SNVs$num\_synonymous)/dim(cases.SNVs)[1]; ASD.syn.rate #0.25

#clearly not significantly different than rate in controls

poisson.test(c(468, 626), c(1902, 2508)) #0.83

sum(cases.SNVs$num\_synonymous) #626

ASD.syn.rate/contr.syn.rate #1.01

#rate in SCZ

#number of synonymous variants: 242 in 1077 trios. though one study (comprising 14 trios) didn't report any

#synonymous variants

SCZ.syn.rate <- 242/1077 #0.22

#Test whether rate differs from SSC controls:

poisson.test(c(468, 242), c(1902, 1077)) #p = 0.26

SCZ.syn.rate/contr.syn.rate #0.91

#rate in CHD

#n = 2645

#syn variants = 701

CHD.syn.rate <- 701/2645 #0.27

#Test whether rate differs from SSC controls:

poisson.test(c(468, 701), c(1902, 2645)) #p = 0.22

CHD.syn.rate/contr.syn.rate #1.08

#rate in ID:

#n = 5264

#syn variants = 1529

ID.syn.rate <- 1529/5264 #0.29 (SE = sqrt(1529/5264^2))

#Test whether rate differs from SSC controls:

poisson.test(c(468, 1529), c(1902, 5264)) #p = 0.002

ID.syn.rate/contr.syn.rate #1.18

#For epilepsy a different method was used to assign annotations to variants which lowers synonymous rate

#n = 1942

#348 = nonsynonymous

#there are 1911 controls and 317 synonymous variants in controls

EPI.syn.rate <- 348/1942 #0.18 variants per patient

#using her method there are only 317 synonymous variants in controls.

EPIcontrol.syn.rate <- 317/1911 #0.17 variants per control

#Test whether rate differs from SSC controls:

poisson.test(c(317, 348), c(1911, 1942)) #p = 0.33

EPI.syn.rate/EPIcontrol.syn.rate #1.08

##rate of nonsynonymous dnSNVs in each disorder: comparing observed rate in each disorder to observed

#control rate; determine rate assuming true synonymous rates are equalized

#canonical control rate

control.dnSNVs <- sum(controls.SNVs$num\_missense, controls.SNVs$num\_PTV); control.dnSNVs; #1238

dim(controls.SNVs) #1902

1238/1902 #0.65

#ASD:

ASD.dnSNVs <- sum(cases.SNVs$num\_missense, cases.SNVs$num\_PTV); ASD.dnSNVs; #1992

dim(cases.SNVs) #2508

1992/2508 #0.79

#estimated OR:

(1992\*1902)/(2508\*1238) #1.22

#test whether rate differs from SSC controls

poisson.test(c(1238, 1992), c(1902, 2508)) #3e-8

#number of dnSNVs assuming synonymous rates equalized

ASD.adj.dnSNVs <- ASD.dnSNVs\*contr.syn.rate/ASD.syn.rate #1964

ASD.adj.dnSNVs/2508 #0.78

#Estimated OR

ASD.adj.dnSNVs\*1902/(2508\*1238)

#test whether rate still differs

poisson.test(c(1238, 1964), c(1902, 2508)) #3e-7

##for schizophrenia

#number of trios = 1077

#number of PTVs = 114

#number of missense = 664

114 + 664 -> SCZ.dnSNVs

#778 total nonsynonymous:

778/1077 #0.72

#Estimated OR:

#estimated OR:

(778\*1902)/(1077\*1238) #1.11

#test whether rate differs from SSC controls

poisson.test(c(1238, 778), c(1902, 1077)) #0.02

#number of dnSNVs assuming synonymous rates equalized

SCZ.adj.dnSNVs <- SCZ.dnSNVs\*contr.syn.rate/SCZ.syn.rate #852

SCZ.adj.dnSNVs/1077 #0.79

#Estimated OR

SCZ.adj.dnSNVs\*1902/(1077\*1238)

#test whether rate still differs

poisson.test(c(1238, 852), c(1902, 1077)) #1e-5

##for intellectual disability used data published in Kaitlin's paper on missense constraint (bioarchive):

#n = 5264

#missense = 4835

#PTV = 1249

4835 + 1249 -> ID.dnSNVs #6084

6084/5264 #1.16

#estimated OR:

(6084\*1902)/(5264\*1238) #1.78

#test whether rate differs from SSC controls

poisson.test(c(1238, 6084), c(1902, 5264)) #<2e-16

#number of dnSNVs assuming synonyms rates equalized

ID.adj.dnSNVs <- ID.dnSNVs\*contr.syn.rate/ID.syn.rate #5154

ID.adj.dnSNVs/5264 #0.98

#Estimated OR

ID.adj.dnSNVs\*1902/(5264\*1238)

#test whether rate still differs

poisson.test(c(1238, 5154), c(1902, 5264)) #<2e-16

#For congenital heart defects

# n = 2645 trios

2431 -> CHD.dnSNVs

2431/2645 #0.92

#estimated OR:

(2431\*1902)/(2645\*1238) #1.41

#test whether rate differs from SSC controls

poisson.test(c(1238, 2431), c(1902, 2645)) #<2e-16

#number of dnSNVs assuming synonyms rates equalized

CHD.adj.dnSNVs <- CHD.dnSNVs\*contr.syn.rate/CHD.syn.rate #2257

CHD.adj.dnSNVs/2645 #0.85

#Estimated OR

CHD.adj.dnSNVs\*1902/(2645\*1238)

#test whether rate still differs

poisson.test(c(1238, 2257), c(1902, 2645)) #e-14

#For epilepsy

#n = 1942 as above

#nonsynoymous = 2118 (see excel file referenced above)

EPI.dnSNVs <- 2118

#control rate for epilepsy; n = 1911. nonsynon = 1307

2118/1942 #1.09 = rate in epilepsy

1307/1911 #0.68 = relevant control rate

#estimated OR:

(2118\*1911)/(1942\*1307) #1.59

#test whether rate differs from SSC controls

poisson.test(c(1307, 2118), c(1911, 1942)) #<2e-16

#number of dnSNVs assuming synonyms rates equalized

EPI.adj.dnSNVs <- EPI.dnSNVs\*EPIcontrol.syn.rate/EPI.syn.rate #1960

EPI.adj.dnSNVs/1942

#Estimated OR

EPI.adj.dnSNVs\*1911/(1942\*1307)

#test whether rate still differs

poisson.test(c(1307, 1960), c(1911, 1942)) #<2e-16

##4) Generate parameters (beta2s) using observed nonsynonymous and nonsynonymous adj for syn

##Will create a function generate\_Beta2s which will take as its inputs:

#number of DNVs in disease population (dnvA)

#size of disease population (nA)

#number of DNvs in control population (dnvB)

#size of control population (nB)

##It will return as its outputs:

#(1) OR of having disease given one additional DNV

#(1) natural log of odds ratio of having disease given one additional DNV

#(2) SE for this estimate

generate\_Beta2s <- function(dnvA, nA, dnvB =1238, nB = 1902) {

A <- dnvA/nA

B <- dnvB/nB

SE.A <- sqrt(A/nA)

SE.B <- sqrt(B/nB)

c(A/B, log(A/B),

sqrt((SE.A/A)^2 + (SE.B/B)^2))

}

##for adjusted rate, uncertainty in synonymous rate must also be taken into account:

generate\_Beta2s\_adj <- function(n.case, nonsyn.case, syn.case,

n.contr = 1902, nonsyn.contr = 1238, syn.contr = 468) {

A <- syn.contr/n.contr #control synonymous rate

B <- syn.case/n.case #case synonymous rate

C <- nonsyn.contr/n.contr #control nonsynonymous rate

D <- nonsyn.case/n.case #case nonsynonymous rate

SE.A <- sqrt(A/n.contr)

SE.B <- sqrt(B/n.case)

SE.C <- sqrt(C/n.contr)

SE.D <- sqrt(D/n.case)

c(A\*D/(B\*C), log(A\*D/(B\*C)),

sqrt((SE.A/A)^2 + (SE.B/B)^2 + (SE.C/C)^2 + (SE.D/D)^2))

}

#ASD:

x <- generate\_Beta2s(ASD.dnSNVs, 2508)

Beta2.nonsyn.ASD <- x[2]

Beta2.SE.nonsyn.ASD <- x[3]

x[1] #OR = 1.22

2\*min(pnorm(Beta2.nonsyn.ASD/Beta2.SE.nonsyn.ASD, lower.tail = TRUE),

pnorm(Beta2.nonsyn.ASD/Beta2.SE.nonsyn.ASD, lower.tail = FALSE))

#p = 4e-8

x <- generate\_Beta2s\_adj(n.case = 2508, nonsyn.case = ASD.dnSNVs, syn.case = 626)

Beta2.nonsyn.ASD.adj <- x[2]

Beta2.SE.nonsyn.ASD.adj <- x[3]

x[1] #adjusted OR = 1.20

2\*min(pnorm(Beta2.nonsyn.ASD.adj/Beta2.SE.nonsyn.ASD.adj, lower.tail = TRUE),

pnorm(Beta2.nonsyn.ASD.adj/Beta2.SE.nonsyn.ASD.adj, lower.tail = FALSE))

#p = 0.009

#SCZ

x <- generate\_Beta2s(SCZ.dnSNVs, 1077)

Beta2.nonsyn.SCZ <- x[2]

Beta2.SE.nonsyn.SCZ <- x[3]

x[1] #OR = 1.11

2\*min(pnorm(Beta2.nonsyn.SCZ/Beta2.SE.nonsyn.SCZ, lower.tail = TRUE),

pnorm(Beta2.nonsyn.SCZ/Beta2.SE.nonsyn.SCZ, lower.tail = FALSE))

#p = 0.02

x <- generate\_Beta2s\_adj(n.case = 1077, nonsyn.case = SCZ.dnSNVs, syn.case = 242)

Beta2.nonsyn.SCZ.adj <- x[2]

Beta2.SE.nonsyn.SCZ.adj <- x[3]

x[1] #adjusted OR = 1.22

2\*min(pnorm(Beta2.nonsyn.SCZ.adj/Beta2.SE.nonsyn.SCZ.adj, lower.tail = TRUE),

pnorm(Beta2.nonsyn.SCZ.adj/Beta2.SE.nonsyn.SCZ.adj, lower.tail = FALSE))

#p = 0.03

#ID:

x <- generate\_Beta2s(ID.dnSNVs, 5264)

Beta2.nonsyn.ID <- x[2]

Beta2.SE.nonsyn.ID <- x[3]

x[1] #OR = 1.78

2\*min(pnorm(Beta2.nonsyn.ID/Beta2.SE.nonsyn.ID, lower.tail = TRUE),

pnorm(Beta2.nonsyn.ID/Beta2.SE.nonsyn.ID, lower.tail = FALSE))

#p = e-75

x <- generate\_Beta2s\_adj(n.case = 5264, nonsyn.case = ID.dnSNVs, syn.case = 1529)

Beta2.nonsyn.ID.adj <- x[2]

Beta2.SE.nonsyn.ID.adj <- x[3]

x[1] #adjusted OR = 1.50

2\*min(pnorm(Beta2.nonsyn.ID.adj/Beta2.SE.nonsyn.ID.adj, lower.tail = TRUE),

pnorm(Beta2.nonsyn.ID.adj/Beta2.SE.nonsyn.ID.adj, lower.tail = FALSE))

#p = 3e-11

#CHD:

x <- generate\_Beta2s(CHD.dnSNVs, 2645)

Beta2.nonsyn.CHD <- x[2]

Beta2.SE.nonsyn.CHD <- x[3]

x[1] #OR = 1.41

2\*min(pnorm(Beta2.nonsyn.CHD/Beta2.SE.nonsyn.CHD, lower.tail = TRUE),

pnorm(Beta2.nonsyn.CHD/Beta2.SE.nonsyn.CHD, lower.tail = FALSE))

#p = 5e-23

x <- generate\_Beta2s\_adj(n.case = 2645, nonsyn.case = CHD.dnSNVs, syn.case = 701)

Beta2.nonsyn.CHD.adj <- x[2]

Beta2.SE.nonsyn.CHD.adj <- x[3]

x[1] #adjusted OR = 1.31

2\*min(pnorm(Beta2.nonsyn.CHD.adj/Beta2.SE.nonsyn.CHD.adj, lower.tail = TRUE),

pnorm(Beta2.nonsyn.CHD.adj/Beta2.SE.nonsyn.CHD.adj, lower.tail = FALSE))

#p = 9e-5

#EPI

x <- generate\_Beta2s(EPI.dnSNVs, 1942, 1307, 1911) #using different number of dnSNVs in epilepsy

Beta2.nonsyn.EPI <- x[2]

Beta2.SE.nonsyn.EPI <- x[3]

x[1] #OR = 1.59

2\*min(pnorm(Beta2.nonsyn.EPI/Beta2.SE.nonsyn.EPI, lower.tail = TRUE),

pnorm(Beta2.nonsyn.EPI/Beta2.SE.nonsyn.EPI, lower.tail = FALSE))

#p = 4e-40

x <- generate\_Beta2s\_adj(n.case = 1942, nonsyn.case = EPI.dnSNVs, syn.case = 348,

n.contr = 1911, nonsyn.contr = 1307, syn.contr = 317)

Beta2.nonsyn.EPI.adj <- x[2]

Beta2.SE.nonsyn.EPI.adj <- x[3]

x[1] #adjusted OR = 1.48

2\*min(pnorm(Beta2.nonsyn.EPI.adj/Beta2.SE.nonsyn.EPI.adj, lower.tail = TRUE),

pnorm(Beta2.nonsyn.EPI.adj/Beta2.SE.nonsyn.EPI.adj, lower.tail = FALSE))

#p = 5e-6

##5) basic functions for running the model:

##create a function that takes a baseline age, another age, beta1 and beta2 and returns

#expected disease risk associated with increased paternal age mediated by

#increased de novo mutations. Beta1 reflects the association between paternal age

#and number of new mutations.

#Beta2 reflects the association between number of new mutations and disease risk.

##12.4 edit to account for puberty (age 13)

denovo.risk <- function(age1, age2, beta1, intercept, beta2) {

exp(beta2)^(exp(intercept)\*(exp(beta1\*(age2)) - exp(beta1\*(age1))))

}

##create a function that takes all of the above parameters, in addition to standard errors,

#i, which determines number of permutations to perform (set at 10000) and CI (set at 0.95),

#which determines how wide a confidence interval:

#which determines how wide a confidence interval:

denovo.risk.CI <- function(age1, age2, beta1, beta1.SE,

intercept, intercept.SE, beta2, beta2.SE, i = 10000, CI = 0.95) {

beta1.sim <- rnorm(i, beta1, beta1.SE)

intercept.sim <- rnorm(i, intercept, intercept.SE)

beta2.sim <- rnorm(i, beta2, beta2.SE)

sims <- NULL

for (x in 1:i) {

sims[x] <- denovo.risk(age1, age2, beta1.sim[x], intercept.sim[x], beta2.sim[x])

}

CI.range <- quantile(sims, c(0.5\*(1 - CI), 1 - 0.5\*(1 - CI)))

c(denovo.risk(age1, age2, beta1, intercept, beta2), CI.range)

#returns estimate, lower limit, upper limit

}

###6) prep supplementary tables with results from model vs age 25 (ages 30-75):

#Create a table showing risk of each disease (due to all nonsynonymous dnSNVs)

#in older fathers compared to 25 year old fathers

#repeat using adjusted estimates for impact of dnSNVs

intercept <- intercept.nonsyn.dad

intercept.SE <- intercept.SE.nonsyn.dad

beta1 <- beta.nonsyn.dad

beta1.SE <- beta.SE.nonsyn.dad

beta2s <- c(Beta2.nonsyn.ASD, Beta2.nonsyn.SCZ, Beta2.nonsyn.ID, Beta2.nonsyn.CHD, Beta2.nonsyn.EPI)

beta2s.SE <- c(Beta2.SE.nonsyn.ASD, Beta2.SE.nonsyn.SCZ, Beta2.SE.nonsyn.ID, Beta2.SE.nonsyn.CHD,

Beta2.SE.nonsyn.EPI)

disorder.names <- c("ASD", "SCZ", "ID", "CHD", "EPI")

ages <- seq(30, 75, 5)

ests <- matrix(nrow = length(ages), ncol = length(beta2s) + 1)

colnames(ests) <- c("older age", disorder.names)

ci.ll <- matrix(nrow = length(ages), ncol = length(beta2s) + 1)

colnames(ci.ll) <- c("older age", disorder.names)

ci.ul <- matrix(nrow = length(ages), ncol = length(beta2s) + 1)

colnames(ci.ul) <- c("older age", disorder.names)

set.seed(12345)

for (p in 2:(1 + length(beta2s))) {

for(a in 1:length(ages)) {

x <- denovo.risk.CI(25, ages[a], beta1, beta1.SE, intercept, intercept.SE,

beta2s[p - 1], beta2s.SE[p - 1])

ests[a, p] <- x[1]

ci.ll[a, p] <- x[2]

ci.ul[a, p] <- x[3]

}}

ests[,1] <- ages

ci.ll[,1] <- ages

ci.ul[,1] <- ages

results.table <- ests

for (a in 1:nrow(ests)) {

for (b in 2:ncol(ests)) {

results.table[a,b] <-

paste(round(ests[a,b], digits = 2), " (",

round(ci.ll[a,b], digits = 2), "-",

round(ci.ul[a,b], digits = 2), ")", sep = "")}}

setwd("../figures and tables2")

write.csv(results.table, "all pheno results table2 NOT ADJUSTED for SYN.csv", row.names = FALSE)

##7) main figures NOT ADJUSTED

##MAIN FIGURE 1: paternal age effect

#create a figure showing dn-only paternal age effect for all phenotypes mediated by

#nonsynonymous SNVs:

dat <- data.frame(comparison = rep(c("35 v. 25", "45 v. 25", "55 v. 25"), 5),

condition = c(rep("SCZ", 3), rep("ASD", 3), rep("CHD", 3),

rep("EPI", 3), rep("ID", 3)),

ests = c(as.numeric(ests[c(2,4,6), 3]), #SCZ

as.numeric(ests[c(2,4,6), 2]), #ASD

as.numeric(ests[c(2,4,6), 5]), #CHD

as.numeric(ests[c(2,4,6), 6]), #EPI

as.numeric(ests[c(2,4,6), 4])), #ID

ul = c(as.numeric(ci.ul[c(2,4,6), 3]), #SCZ

as.numeric(ci.ul[c(2,4,6), 2]), #ASD

as.numeric(ci.ul[c(2,4,6), 5]), #CHD

as.numeric(ci.ul[c(2,4,6), 6]), #EP

as.numeric(ci.ul[c(2,4,6), 4])), #ID

ll = c(as.numeric(ci.ll[c(2,4,6), 3]), #SCZ

as.numeric(ci.ll[c(2,4,6), 2]), #ASD

as.numeric(ci.ll[c(2,4,6), 5]), #CHD

as.numeric(ci.ll[c(2,4,6), 6]), #EP

as.numeric(ci.ll[c(2,4,6), 4]))) #ID

dat$condition <- factor(dat$condition, levels = c("SCZ", "ASD", "CHD", "EPI", "ID"))

setwd("../figures and tables2")

#bar

pdf("bar graph 5 conditions 3 ages2 UNADJUSTED.pdf")

Palette <- c("#FECC5C", "#FD8D3C", "#F03B20","#BD0026")

ggplot(data = dat, aes(comparison, fill = comparison))+

geom\_bar(stat = "identity", position = "dodge", aes(y = ests))+

geom\_errorbar(aes(ymax = ul, ymin = ll),

position = position\_dodge(width = 0.9),width = 0.25)+

facet\_wrap(~condition, switch = "x", nrow = 1, ncol = 5)+

coord\_cartesian(ylim=c(1,2.77))+

scale\_y\_continuous(breaks = seq(1, 2.75, 0.25))+

labs(y = "IRR", x = element\_blank())+

ggtitle("Predicted impact of paternal age-related\ndnSNVs (unadjusted by synonymous rate) on five disorders")+

theme(plot.title = element\_text(hjust = 0.5, size = rel(1)))+

geom\_hline(yintercept = 1, linetype = 2, lwd = 0.2)+

scale\_fill\_manual(values = Palette)+

theme(plot.title = element\_text(face="bold", size = title.size),

axis.title.y = element\_text(face = "bold", size = axistitle.size),

axis.text.x = element\_blank(),

axis.text.y = element\_text(size = axistext.size),

legend.position = "bottom",

legend.title = element\_text(size = axistitle.size),

legend.text = element\_text(size = axistext.size),

strip.text = element\_text(size = axistext.size),

panel.background = element\_blank(),

panel.grid.major.y = element\_blank(),

panel.grid.major.x = element\_blank(),

axis.line.y = element\_line(),

axis.ticks.y = element\_blank(),

axis.ticks.x = element\_blank())

dev.off()

##figure 2: comparing model to epidemiology in Denmark

dat.model <- data.frame(condition = c("SCZ", "ASD", "CHD", "EPI", "ID"), ests = NA, ul = NA, ll = NA, comparison = "dnSNV only")

dat.model$condition <- factor(dat.model$condition, levels = dat.model$condition[1:5])

ests <- NULL

ll <- NULL

ul <- NULL

##Run our model comparing offspring of 44.1 year olds to offspring of 26.2 year olds

set.seed(2468)

for(a in c(2, 1)){ #SCZ, ASD

x <- denovo.risk.CI(26.2, 44.1, beta1, beta1.SE, intercept, intercept.SE, beta2s[a], beta2s.SE[a])

ests <- c(ests, x[1])

ll <- c(ll, x[2])

ul <- c(ul, x[3])

}

#CHD: note different young and old ages

x <- denovo.risk.CI(27.1, 43.7, beta1, beta1.SE, intercept, intercept.SE, beta2s[4], beta2s.SE[4])

ests <- c(ests, x[1])

ll <- c(ll, x[2])

ul <- c(ul, x[3])

for(a in c(5, 3)){ #EP, ID

x <- denovo.risk.CI(26.2, 44.1, beta1, beta1.SE, intercept, intercept.SE, beta2s[a], beta2s.SE[a])

ests <- c(ests, x[1])

ll <- c(ll, x[2])

ul <- c(ul, x[3])

}

dat.model$ests <- ests; dat.model$ul <- ul; dat.model$ll <- ll

#Danish data:

#include crude Danish data (see JC email from 6/17, 7/3) on comparison for >39 vs 20s

#use age 18 for epilepsy

Danish.ests <- c(1.31, 1.68, 1.06, 1.39, 1.44)

Danish.ll <- c(1.23, 1.51, 0.89, 1.08, 1.33)

Danish.ul <- c(1.4, 1.86, 1.18, 1.79, 1.56)

#now create dat that takes this all into account:

Danish.dat <- data.frame(condition = dat.model$condition,

ests = Danish.ests, ul = Danish.ul, ll = Danish.ll, comparison = "Epidemiologic")

dat2 <- rbind(dat.model, Danish.dat)

row.names(dat2) <- 1:10

#data for table comparing dnSNV model to Danish epidemiologic data:

dat2

#ASD order of magnitude bigger effect size

(1.68 - 1)/(1.073551 - 1) #9.25

#make figure

#make table comparing:

setwd("../figures and tables2")

pdf("epi vs model figure2 UNADJUSTED.pdf")

Palette <- c("#FECC5C", "#FD8D3C", "#F03B20","#BD0026")

ggplot(data = dat2, aes(comparison, fill = comparison))+

geom\_bar(stat = "identity", position = "dodge", aes(y = ests))+

geom\_errorbar(aes(ymax = ul, ymin = ll),

position = position\_dodge(width = 0.9),width = 0.25)+

facet\_wrap(~condition, switch = "x", nrow = 1, ncol = 5)+

scale\_fill\_manual(values = Palette)+

coord\_cartesian(ylim=c(0.9,2.05))+

scale\_y\_continuous(breaks = seq(1, 2.05, 0.25))+

labs(y = "Fold increase in risk", x = element\_blank())+

ggtitle("Disease risk in offspring of fathers older than 39 vs. offspring of fathers in their 20s\nin dnSNV model (unadjusted by synonymous rate) and Danish population")+

theme(plot.title = element\_text(hjust = 0.5, size = rel(1)))+

theme(plot.title = element\_text(hjust = 0.5, face="bold", size = axistitle.size),

axis.title.y = element\_text(face = "bold", size = axistitle.size),

axis.text.x = element\_blank(),

axis.text.y = element\_text(size = axistext.size),

legend.position = "bottom",

legend.title = element\_text(size = axistitle.size),

legend.text = element\_text(size = axistext.size),

strip.text = element\_text(size = axistext.size),

panel.background = element\_blank(),

panel.grid.major.y = element\_blank(),

panel.grid.major.x = element\_blank(),

axis.line.y = element\_line(),

axis.ticks.y = element\_blank(),

axis.ticks.x = element\_blank())+

geom\_text(x = 2, y = c(rep(dat2$ul[6] + 0.02, 2), rep(dat2$ul[7] + 0.02, 2), rep(0, 6)),

label = c("", "\*", "", "\*", "", "", "", "", "",""))

dev.off()

##8) statistical tests comparing model to epidemiology

##Statistical tests comparing Danish data to our model and looking for heterogeneity in ratios:

older.age <- 44.2

base.age <- 26.1

#create a function that takes data from JC (6/15) and create probability cloud around data for hypothesis testing:

Make.Danish.Distr <- function(est, se.log, n = 100000) {

exp(rnorm(n, log(est), se.log))}

#create a function that returns 10,000 values for de novo risk for a given comparison

denovo.risk.dist <- function(age1, age2, beta1, beta1.SE,

intercept, intercept.SE, beta2, beta2.SE, i = 100000) {

beta1.sim <- rnorm(i, beta1, beta1.SE)

intercept.sim <- rnorm(i, intercept, intercept.SE)

beta2.sim <- rnorm(i, beta2, beta2.SE)

sims <- NULL

for (x in 1:i) {

sims[x] <- denovo.risk(age1, age2, beta1.sim[x], intercept.sim[x], beta2.sim[x])

}

sims

}

##For each phenotype find empiric 2-sided p value for null hypothesis Danish estimate = dn only estimate:

##also find estimate and 95%

set.seed(12345)

##Schizophrenia:

#From JC: HR 1.31; SE(log(HR)) = 0.0334

danish.distr.SCZ <- Make.Danish.Distr(1.31, 0.0334)

model.distr.SCZ <- denovo.risk.dist(base.age, older.age, beta1, beta1.SE,intercept, intercept.SE,

beta2s[2], beta2s.SE[2])

p <- sum(danish.distr.SCZ < model.distr.SCZ)/100000

2\*min(p, 1 - p) #2e-5

##ASD:

#From JC: HR 1.68; SE(log(HR)) = 0.0538

danish.distr.ASD <- Make.Danish.Distr(1.68, 0.0538)

model.distr.ASD <- denovo.risk.dist(base.age, older.age, beta1, beta1.SE,intercept, intercept.SE,

beta2s[1], beta2s.SE[1])

p <- sum(danish.distr.ASD < model.distr.ASD)/100000

2\*min(p, 1 - p) #0 (< 2x10^-5)

##CHD:

#From JC: OR 1.06; 95% CI 0.94 - 1.20

#to Find the SE(log(OR):

mean(log(1.20) - log(1.06), log(1.06) - log(0.94))/1.96 #0.0633

danish.distr.CHD <- Make.Danish.Distr(1.06, 0.0633)

model.distr.CHD <- denovo.risk.dist(base.age, older.age, beta1, beta1.SE,intercept, intercept.SE,

beta2s[4], beta2s.SE[4])

p <- sum(danish.distr.CHD < model.distr.CHD)/100000

2\*min(p, 1 - p) #0.27

##EP:

#From JC: HR 1.39; SE(log(HR)) = 0.128

danish.distr.EP <- Make.Danish.Distr(1.39, 0.128)

model.distr.EP <- denovo.risk.dist(base.age, older.age, beta1, beta1.SE,intercept, intercept.SE,

beta2s[5], beta2s.SE[5])

p <- sum(danish.distr.EP < model.distr.EP)/100000

2\*min(p, 1 - p) #0.33

##ID:

#From JC: HR 1.44; SE(log(HR)) = 0.0406

danish.distr.ID <- Make.Danish.Distr(1.44, 0.0406)

model.distr.ID <- denovo.risk.dist(base.age, older.age, beta1, beta1.SE,intercept, intercept.SE,

beta2s[3], beta2s.SE[3])

p <- sum(danish.distr.ID < model.distr.ID)/100000

2\*min(p, 1 - p) #0.13

##9) Repeat sections 6-8 above with adjusted nonsynonymous rates:

beta2s <- c(Beta2.nonsyn.ASD.adj, Beta2.nonsyn.SCZ.adj, Beta2.nonsyn.ID.adj, Beta2.nonsyn.CHD.adj,

Beta2.nonsyn.EPI.adj)

beta2s.SE <- c(Beta2.SE.nonsyn.ASD.adj, Beta2.SE.nonsyn.SCZ.adj, Beta2.SE.nonsyn.ID.adj,

Beta2.SE.nonsyn.CHD.adj, Beta2.SE.nonsyn.EPI.adj)

ages <- seq(30, 75, 5)

ests <- matrix(nrow = length(ages), ncol = length(beta2s) + 1)

colnames(ests) <- c("older age", disorder.names)

ci.ll <- matrix(nrow = length(ages), ncol = length(beta2s) + 1)

colnames(ci.ll) <- c("older age", disorder.names)

ci.ul <- matrix(nrow = length(ages), ncol = length(beta2s) + 1)

colnames(ci.ul) <- c("older age", disorder.names)

set.seed(12345)

for (p in 2:(1 + length(beta2s))) {

for(a in 1:length(ages)) {

x <- denovo.risk.CI(25, ages[a], beta1, beta1.SE, intercept, intercept.SE,

beta2s[p - 1], beta2s.SE[p - 1])

ests[a, p] <- x[1]

ci.ll[a, p] <- x[2]

ci.ul[a, p] <- x[3]

}}

ests[,1] <- ages

ci.ll[,1] <- ages

ci.ul[,1] <- ages

results.table <- ests

for (a in 1:nrow(ests)) {

for (b in 2:ncol(ests)) {

results.table[a,b] <-

paste(round(ests[a,b], digits = 2), " (",

round(ci.ll[a,b], digits = 2), "-",

round(ci.ul[a,b], digits = 2), ")", sep = "")}}

setwd("../figures and tables2")

write.csv(results.table, "adjusted pheno results table2.csv", row.names = FALSE)

##ADJUSTED FIGURE 1: paternal age effect

#create a figure showing dn-only paternal age effect for all phenotypes mediated by

#nonsynonymous SNVs:

dat <- data.frame(comparison = rep(c("35 v. 25", "45 v. 25", "55 v. 25"), 5),

condition = c(rep("SCZ", 3), rep("ASD", 3), rep("CHD", 3),

rep("EPI", 3), rep("ID", 3)),

ests = c(as.numeric(ests[c(2,4,6), 3]), #SCZ

as.numeric(ests[c(2,4,6), 2]), #ASD

as.numeric(ests[c(2,4,6), 5]), #CHD

as.numeric(ests[c(2,4,6), 6]), #EPI

as.numeric(ests[c(2,4,6), 4])), #ID

ul = c(as.numeric(ci.ul[c(2,4,6), 3]), #SCZ

as.numeric(ci.ul[c(2,4,6), 2]), #ASD

as.numeric(ci.ul[c(2,4,6), 5]), #CHD

as.numeric(ci.ul[c(2,4,6), 6]), #EP

as.numeric(ci.ul[c(2,4,6), 4])), #ID

ll = c(as.numeric(ci.ll[c(2,4,6), 3]), #SCZ

as.numeric(ci.ll[c(2,4,6), 2]), #ASD

as.numeric(ci.ll[c(2,4,6), 5]), #CHD

as.numeric(ci.ll[c(2,4,6), 6]), #EP

as.numeric(ci.ll[c(2,4,6), 4]))) #ID

dat$condition <- factor(dat$condition, levels = c("SCZ", "ASD", "CHD", "EPI", "ID"))

setwd("../figures and tables2")

#bar

pdf("adjusted bar graph 5 conditions 3 ages2.pdf")

Palette <- c("#FECC5C", "#FD8D3C", "#F03B20","#BD0026")

ggplot(data = dat, aes(comparison, fill = comparison))+

geom\_bar(stat = "identity", position = "dodge", aes(y = ests))+

geom\_point(stat = "identity", aes(y = ests))+

geom\_errorbar(aes(ymax = ul, ymin = ll),

position = position\_dodge(width = 0.9),width = 0.25)+

facet\_wrap(~condition, switch = "x", nrow = 1, ncol = 5)+

coord\_cartesian(ylim=c(1,2.77))+

scale\_y\_continuous(breaks = seq(1, 2.75, 0.25))+

labs(y = "IRR", x = element\_blank())+

ggtitle("Impact of paternal age-related dnSNVs on five disorders")+

theme(plot.title = element\_text(hjust = 0.5, size = rel(1)))+

geom\_hline(yintercept = 1, linetype = 2, lwd = 0.2)+

scale\_fill\_manual(values = Palette)+

theme(plot.title = element\_text(face="bold", size = title.size),

axis.title.y = element\_text(face = "bold", size = axistitle.size),

axis.text.x = element\_blank(),

axis.text.y = element\_text(size = axistext.size),

legend.position = "bottom",

legend.title = element\_text(size = axistitle.size),

legend.text = element\_text(size = axistext.size),

strip.text = element\_text(size = axistext.size),

panel.background = element\_blank(),

panel.grid.major.y = element\_blank(),

panel.grid.major.x = element\_blank(),

axis.line.y = element\_line(),

axis.ticks.y = element\_blank(),

axis.ticks.x = element\_blank())

dev.off()

##ADJUSTED figure 2: comparing model to epidemiology in Denmark

dat.model <- data.frame(condition = c("SCZ", "ASD", "CHD", "EPI", "ID"), ests = NA, ul = NA, ll = NA, comparison = "dnSNV only")

dat.model$condition <- factor(dat.model$condition, levels = dat.model$condition[1:5])

ests <- NULL

ll <- NULL

ul <- NULL

##Run our model comparing offspring of 44.1 year olds to offspring of 26.2 year olds

set.seed(12345)

for(a in c(2, 1)){ #SCZ, ASD

x <- denovo.risk.CI(26.2, 44.1, beta1, beta1.SE, intercept, intercept.SE, beta2s[a], beta2s.SE[a])

ests <- c(ests, x[1])

ll <- c(ll, x[2])

ul <- c(ul, x[3])

}

#CHD: note different young and old ages

x <- denovo.risk.CI(27.1, 43.7, beta1, beta1.SE, intercept, intercept.SE, beta2s[4], beta2s.SE[4])

ests <- c(ests, x[1])

ll <- c(ll, x[2])

ul <- c(ul, x[3])

for(a in c(5, 3)){ #EP, ID

x <- denovo.risk.CI(26.2, 44.1, beta1, beta1.SE, intercept, intercept.SE, beta2s[a], beta2s.SE[a])

ests <- c(ests, x[1])

ll <- c(ll, x[2])

ul <- c(ul, x[3])

}

dat.model$ests <- ests; dat.model$ul <- ul; dat.model$ll <- ll

#Danish data:

#include crude Danish data (see JC email from 6/17, 7/3) on comparison for >39 vs 20s

#use age 18 for epilepsy

Danish.ests <- c(1.31, 1.68, 1.06, 1.39, 1.44)

Danish.ll <- c(1.23, 1.51, 0.89, 1.08, 1.33)

Danish.ul <- c(1.4, 1.86, 1.18, 1.79, 1.56)

#now create dat that takes this all into account:

Danish.dat <- data.frame(condition = dat.model$condition,

ests = Danish.ests, ul = Danish.ul, ll = Danish.ll, comparison = "Epidemiologic")

dat2 <- rbind(dat.model, Danish.dat)

row.names(dat2) <- 1:10

#data for table comparing dnSNV model to Danish epidemiologic data:

dat2

#ASD order of magnitude bigger effect size

(1.68 - 1)/(1.073551 - 1) #9.25

#make figure

#make table comparing:

setwd("../figures and tables2")

pdf("adjusted epi vs model figure2.pdf")

Palette <- c("#FECC5C", "#FD8D3C", "#F03B20","#BD0026")

ggplot(data = dat2, aes(comparison, fill = comparison))+

geom\_bar(stat = "identity", position = "dodge", aes(y = ests))+

geom\_point(aes(y = ests))+

geom\_errorbar(aes(ymax = ul, ymin = ll),

position = position\_dodge(width = 0.9),width = 0.25)+

facet\_wrap(~condition, switch = "x", nrow = 1, ncol = 5)+

scale\_fill\_manual(values = Palette)+

coord\_cartesian(ylim=c(0.9,2.05))+

scale\_y\_continuous(breaks = seq(1, 2.05, 0.25))+

labs(y = "Fold increase in risk", x = element\_blank())+

ggtitle("Disease risk in offspring of older fathers vs. offspring of younger fathers in\ndnSNV model and Danish population")+

theme(plot.title = element\_text(hjust = 0.5, size = rel(1)))+

theme(plot.title = element\_text(hjust = 0.5, face="bold", size = axistitle.size),

axis.title.y = element\_text(face = "bold", size = axistitle.size),

axis.text.x = element\_blank(),

axis.text.y = element\_text(size = axistext.size),

legend.position = "bottom",

legend.title = element\_text(size = axistitle.size),

legend.text = element\_text(size = axistext.size),

strip.text = element\_text(size = axistext.size),

panel.background = element\_blank(),

panel.grid.major.y = element\_blank(),

panel.grid.major.x = element\_blank(),

axis.line.y = element\_line(),

axis.ticks.y = element\_blank(),

axis.ticks.x = element\_blank())+

geom\_text(x = 2, y = c(rep(dat2$ul[6] + 0.02, 2), rep(dat2$ul[7] + 0.02, 2), rep(dat2$ul[10] + 0.02, 6)),

label = c("", "\*", "", "\*", "", "", "", "", "",""))

dev.off()

## statistical tests comparing model to epidemiology

##Statistical tests comparing Danish data to our model and looking for heterogeneity in ratios:

older.age <- 44.2

base.age <- 26.1

#create a function that takes data from JC (6/15) and create probability cloud around data for hypothesis testing:

Make.Danish.Distr <- function(est, se.log, n = 100000) {

exp(rnorm(n, log(est), se.log))}

#create a function that returns 10,000 values for de novo risk for a given comparison

denovo.risk.dist <- function(age1, age2, beta1, beta1.SE,

intercept, intercept.SE, beta2, beta2.SE, i = 100000) {

beta1.sim <- rnorm(i, beta1, beta1.SE)

intercept.sim <- rnorm(i, intercept, intercept.SE)

beta2.sim <- rnorm(i, beta2, beta2.SE)

sims <- NULL

for (x in 1:i) {

sims[x] <- denovo.risk(age1, age2, beta1.sim[x], intercept.sim[x], beta2.sim[x])

}

sims

}

##For each phenotype find empiric 2-sided p value for null hypothesis Danish estimate = dn only estimate:

##also find estimate and 95%

set.seed(12345)

##Schizophrenia:

#From JC: HR 1.31; SE(log(HR)) = 0.0334

danish.distr.SCZ <- Make.Danish.Distr(1.31, 0.0334)

model.distr.SCZ <- denovo.risk.dist(base.age, older.age, beta1, beta1.SE,intercept, intercept.SE,

beta2s[2], beta2s.SE[2])

p <- sum(danish.distr.SCZ < model.distr.SCZ)/100000

2\*min(p, 1 - p) #0.006

##ASD:

#From JC: HR 1.68; SE(log(HR)) = 0.0538

danish.distr.ASD <- Make.Danish.Distr(1.68, 0.0538)

model.distr.ASD <- denovo.risk.dist(base.age, older.age, beta1, beta1.SE,intercept, intercept.SE,

beta2s[1], beta2s.SE[1])

p <- sum(danish.distr.ASD < model.distr.ASD)/100000

2\*min(p, 1 - p) #0

##CHD:

#From JC: OR 1.06; 95% CI 0.94 - 1.20

#to Find the SE(log(OR):

mean(log(1.20) - log(1.06), log(1.06) - log(0.94))/1.96 #0.0633

danish.distr.CHD <- Make.Danish.Distr(1.06, 0.0633)

model.distr.CHD <- denovo.risk.dist(base.age, older.age, beta1, beta1.SE,intercept, intercept.SE,

beta2s[4], beta2s.SE[4])

p <- sum(danish.distr.CHD < model.distr.CHD)/100000

2\*min(p, 1 - p) #0.49

##EP:

#From JC: HR 1.39; SE(log(HR)) = 0.128

danish.distr.EP <- Make.Danish.Distr(1.39, 0.128)

model.distr.EP <- denovo.risk.dist(base.age, older.age, beta1, beta1.SE,intercept, intercept.SE,

beta2s[5], beta2s.SE[5])

p <- sum(danish.distr.EP < model.distr.EP)/100000

2\*min(p, 1 - p) #0.23

##ID:

#From JC: HR 1.44; SE(log(HR)) = 0.0406

danish.distr.ID <- Make.Danish.Distr(1.44, 0.0406)

model.distr.ID <- denovo.risk.dist(base.age, older.age, beta1, beta1.SE,intercept, intercept.SE,

beta2s[3], beta2s.SE[3])

p <- sum(danish.distr.ID < model.distr.ID)/100000

2\*min(p, 1 - p) #0.015

##

p <- sum(danish.distr.ASD < danish.distr.ID)/100000

2\*min(p, 1 - p)

##ASD:

#From JC: HR 1.68; SE(log(HR)) = 0.0538

##ID:

#From JC: HR 1.44; SE(log(HR)) = 0.0406

##Calculation of p value comparing epidemiologic estimate of ASD and ID:

(log(1.68) - log(1.44))/(sqrt(0.0538^2 + 0.0406^2))

pnorm(2.29, lower.tail = F)

#0.011