**Instructions for SITAR script (found below)**

1. Fill in your working directory:

setwd("C:/my/directory/")

1. Fill in the name of your phenotype file. Your file should contain the following headers: height, age, sex, and ID.

# Load cohort phenotype data #

data <- read.csv(file="COHORT.csv",header=TRUE)

1. The script will remove missing cells and initialize a data frame for results.
2. Then the script will deal with each sex separately, starting with males. Double outliers will be excluded, and SITAR models using varying degrees of freedom will be generated.
3. The AIC and BIC for testing these models will be generated. The model with the lowest BIC will be chosen, along with graphs of each model. At this point, please examine the chosen model with the lowest BIC—check that the graph looks as expected (see the analysis plan for an example of an appropriate growth curve). The growth curve should not contain any “anomalies” like dips or extra bumps—it should look like a normal height growth curve. This will however depend on the ages at which height has been measured in your cohort and the spacing between measurements, the optimal parameters for which we are not yet certain.
4. The script will output the male data into “sitar\_results” and run the same for females.
5. The results will be written to a .csv file. You can rename it if you like or leave it as-is.

write.csv(file="height\_sitar\_model\_results.csv",sitar\_results)

1. The script will write out random effects for the model with the lowest BIC. This needs to be filled in manually because I want to make sure you have checked that the best model has been chosen!

# Get random effects for best models-- FILL IN BEST MODEL BASED ON BIC RESULTS ABOVE—this is a chance to check that all data looks fine#

curve\_m\_8\_random <- random.effects(curve\_m\_8)

curve\_m\_8\_random$ID <- row.names(curve\_m\_8\_random)

curve\_m\_8\_random$sex <- rep("M",length=nrow(curve\_m\_8\_random))

curve\_f\_7\_random <- random.effects(curve\_f\_7)

curve\_f\_7\_random$ID <- row.names(curve\_f\_7\_random)

curve\_f\_7\_random$sex <- rep("F",length=nrow(curve\_f\_7\_random))

1. In order to fill in some of the summary data that is required for the excel spreadsheet, you can do the following:

##You can pull out data on each model, for example the SD of the “b” tempo variable, by typing the following##

summary(model\_name)

1. Finally, the script will draw a .pdf with some graphs that will aid in visualizing the results. You may want to return these to us to make sure everything looks okay before moving forward with the GWAS analysis. If there are any questions or concerns, don’t hesitate to get in touch.

# SITAR height analysis script based on MAR SITAR documentation, Appendix B #

# Set directory #

setwd("C:/my/directory/")

# Load required packages; SITAR 1.0.2 #

install.packages("sitar")

install.packages("ggplot2")

install.packages("reshape")

require(sitar)

require(ggplot2)

require(reshape)

# Generate consistent format to be applied in plots #

format\_plots <- theme(axis.text.x=element\_text(face="bold")) + theme(axis.title.x=element\_text(face="bold")) + theme(axis.title.y=element\_text(face="bold")) + theme(axis.text.y=element\_text(face="bold")) + theme(strip.text=element\_text(face="bold")) + theme\_bw() + theme(legend.position="top") + theme(legend.text=element\_text(face="bold")) + theme(title=element\_text(face="bold"))

# Load cohort phenotype data #

data <- read.csv(file="COHORT.csv",header=TRUE)

# Remove missing cells for the phenotype; headers are “height” “age” “sex” and “ID” #

# Units for height are cm; units for age are years #

nrow(data)

summary(data$height)

data <- subset(data,(!is.na(data$height)))

data <- subset(data,(!is.na(data$age)))

data <- subset(data,(!is.na(data$sex)))

data <- subset(data,(!is.na(data$ID)))

nrow(data)

summary(data$height)

# Make data sets by sex #

data\_m <- data[which(data$sex=="M"),]

nrow(data\_m)

summary(data\_m$height)

data\_f <- data[which(data$sex=="F"),]

nrow(data\_f)

summary(data\_f$height)

# Initialize empty data frame for results #

sitar\_results <- data.frame(sex=character(2),dof=integer(2),apv=numeric(2),pv=numeric(2),stringsAsFactors=FALSE)

sitar\_results

##### 1 = Male #####

# Identify outliers prior to building the model; 5 SD outside range #

# Outliers can be excluded based on coding system; err on including for now #

# Outlier codes: 1=rare; 2=complicated, "look at curve"; 3="adjacent to simple outlier"; 4="single outlier"; 5="double outlier"; 6="edge outlier" #

data\_m\_check <- velout(age,height,ID,data\_m,limit=5)

nrow(data\_m\_check)

head(data\_m\_check)

summary(as.factor(data\_m\_check$code))

# Exclude only "double outlier", code = 5; review others #

nrow(data\_m)

data\_m <- merge(data\_m, data\_m\_check, by=c("ID","age","height"))

nrow(data\_m)

data\_m <- subset(data\_m,code!=5)

nrow(data\_m)

# Generate the SITAR models #

# A potential error may occur here, either a convergence error or singularity. This is likely due to the missingness in the data. Please move forward with any models that do work for now and contact us if any questions #

curve\_m\_4 <- sitar(x=age,y=height,id=ID,data=data\_m,4)

curve\_m\_5 <- sitar(x=age,y=height,id=ID,data=data\_m,5)

curve\_m\_6 <- sitar(x=age,y=height,id=ID,data=data\_m,6)

curve\_m\_7 <- sitar(x=age,y=height,id=ID,data=data\_m,7)

curve\_m\_8 <- sitar(x=age,y=height,id=ID,data=data\_m,8)

pdf(file="curve\_m\_4\_8\_height.pdf")

plot(curve\_m\_4)

plot(curve\_m\_5)

plot(curve\_m\_6)

plot(curve\_m\_7)

plot(curve\_m\_8)

dev.off()

# Get AIC and BIC for testing goodness of fit #

# Remove any models that failed in earlier step #

aic\_sitar <- data.frame(dof=integer(5),aic=numeric(5))

aic\_sitar$dof <- c("4","5","6","7","8")

aic\_sitar$aic <- c(summary(curve\_m\_4)$AIC,summary(curve\_m\_5)$AIC,summary(curve\_m\_6)$AIC,summary(curve\_m\_7)$AIC,summary(curve\_m\_8)$AIC)

bic\_sitar <- data.frame(dof=integer(5),bic=numeric(5))

bic\_sitar$dof <- c("4","5","6","7","8")

bic\_sitar$bic <- c(summary(curve\_m\_4)$BIC,summary(curve\_m\_5)$BIC,summary(curve\_m\_6)$BIC,summary(curve\_m\_7)$BIC,summary(curve\_m\_8)$BIC)

# Sort, identify best model by BIC #

attach(bic\_sitar)

bic\_sitar\_sorted <- bic\_sitar[order(bic),]

detach(bic\_sitar)

df=as.numeric(bic\_sitar\_sorted$dof[1])

df

### PLEASE EXAMINE THE CURVES GENERATED ABOVE TO ENSURE THE CHOSEN MODEL LOOKS REASONABLE! See analysis protocol for an example height curve. ###

### If you have questions about which model to choose, please contact us ###

# Parameters from the model; summary, fitted values, fixed effects, random effects #

apv <- ifelse(df==4,summary(curve\_m\_4)$apv[1],ifelse(df==5,summary(curve\_m\_5)$apv[1],ifelse(df==6,summary(curve\_m\_6)$apv[1],ifelse(df==7,summary(curve\_m\_7)$apv[1],ifelse(df==8,summary(curve\_m\_8)$apv[1],"missing")))))

pv <- ifelse(df==4,summary(curve\_m\_4)$apv[2],ifelse(df==5,summary(curve\_m\_5)$apv[2],ifelse(df==6,summary(curve\_m\_6)$apv[2],ifelse(df==7,summary(curve\_m\_7)$apv[2],ifelse(df==8,summary(curve\_m\_8)$apv[2],"missing")))))

# Add findings to summary results #

sitar\_results[1,] <- c("M",df,apv,pv)

sitar\_results

##### 2 = Female #####

# Identify outliers prior to building the model; 5 SD outside range #

# Outliers can be excluded based on coding system; err on including for now #

data\_f\_check <- velout(age,height,ID,data\_f,limit=5)

nrow(data\_f\_check)

head(data\_f\_check)

summary(as.factor(data\_f\_check$code))

# Exclude only "double outlier", code = 5; review others #

nrow(data\_f)

data\_f <- merge(data\_f, data\_f\_check, by=c("ID","age","height"))

nrow(data\_f)

data\_m <- subset(data\_f,code!=5)

nrow(data\_f)

# Generate the SITAR models #

# A potential error may occur here, either a convergence error or singularity. This is likely due to the missingness in the data. Please move forward with any models that do work for now and contact us if any questions #

curve\_f\_4 <- sitar(x=age,y=height,id=ID,data=data\_f,4)

curve\_f\_5 <- sitar(x=age,y=height,id=ID,data=data\_f,5)

curve\_f\_6 <- sitar(x=age,y=height,id=ID,data=data\_f,6)

curve\_f\_7 <- sitar(x=age,y=height,id=ID,data=data\_f,7)

curve\_f\_8 <- sitar(x=age,y=height,id=ID,data=data\_f,8)

pdf(file="curve\_f\_4\_8\_height.pdf")

plot(curve\_f\_4)

plot(curve\_f\_5)

plot(curve\_f\_6)

plot(curve\_f\_7)

plot(curve\_f\_8)

dev.off()

# Get AIC and BIC for testing goodness of fit #

# Remove any models that failed in earlier step #

aic\_sitar <- data.frame(dof=integer(5),aic=numeric(5))

aic\_sitar$dof <- c("4","5","6","7","8")

aic\_sitar$aic <- c(summary(curve\_f\_4)$AIC,summary(curve\_f\_5)$AIC,summary(curve\_f\_6)$AIC,summary(curve\_f\_7)$AIC,summary(curve\_f\_8)$AIC)

bic\_sitar <- data.frame(dof=integer(5),bic=numeric(5))

bic\_sitar$dof <- c("4","5","6","7","8")

bic\_sitar$bic <- c(summary(curve\_f\_4)$BIC,summary(curve\_f\_5)$BIC,summary(curve\_f\_6)$BIC,summary(curve\_f\_7)$BIC,summary(curve\_f\_8)$BIC)

# Sort, identify best model by BIC #

attach(bic\_sitar)

bic\_sitar\_sorted <- bic\_sitar[order(bic),]

detach(bic\_sitar)

df=as.numeric(bic\_sitar\_sorted$dof[1])

df

### PLEASE EXAMINE THE CURVES GENERATED ABOVE TO ENSURE THE CHOSEN MODEL LOOKS REASONABLE! See analysis protocol for an example height curve.###

### If you have questions about which model to choose, please contact us ###

# Parameters from the model; summary, fitted values, fixed effects, random effects #

apv <- ifelse(df==4,summary(curve\_f\_4)$apv[1],ifelse(df==5,summary(curve\_f\_5)$apv[1],ifelse(df==6,summary(curve\_f\_6)$apv[1],ifelse(df==7,summary(curve\_f\_7)$apv[1],ifelse(df==8,summary(curve\_f\_8)$apv[1],"missing")))))

pv <- ifelse(df==4,summary(curve\_f\_4)$apv[2],ifelse(df==5,summary(curve\_f\_5)$apv[2],ifelse(df==6,summary(curve\_f\_6)$apv[2],ifelse(df==7,summary(curve\_f\_7)$apv[2],ifelse(df==8,summary(curve\_f\_8)$apv[2],"missing")))))

# Add findings to summary results #

sitar\_results[2,] <- c("F",df,apv,pv)

sitar\_results

# Add parameter

sitar\_results$parameter <- rep("height",length=nrow(sitar\_results))

# Write SITAR modeling results to CSV file #

write.csv(file="height\_sitar\_model\_results.csv",sitar\_results)

# Get random effects for best models-- FILL IN BEST MODEL BASED ON BIC RESULTS ABOVE—this is a chance to check that all data looks fine#

curve\_m\_8\_random <- random.effects(curve\_m\_8)

curve\_m\_8\_random$ID <- row.names(curve\_m\_8\_random)

curve\_m\_8\_random$sex <- rep("M",length=nrow(curve\_m\_8\_random))

curve\_f\_7\_random <- random.effects(curve\_f\_7)

curve\_f\_7\_random$ID <- row.names(curve\_f\_7\_random)

curve\_f\_7\_random$sex <- rep("F",length=nrow(curve\_f\_7\_random))

sitar\_height\_random\_effects <- rbind(curve\_m\_8\_random, curve\_f\_7\_random)

write.csv(file="height\_sitar\_random\_effects.csv",sitar\_height\_random\_effects)

head(sitar\_height\_random\_effects)

mdata <- melt(sitar\_height\_random\_effects,id=c("ID","sex"))

head(mdata)

mdata$variable\_text <- as.factor(ifelse(mdata$variable=="a","A = Size",ifelse(mdata$variable=="b","B = Tempo",ifelse(mdata$variable=="c","C = Velocity","missing"))))

summary(mdata$variable\_text)

##You can pull out data on each model, for example the SD of the “b” tempo variable, by typing the following##

summary(model\_name)

sd(subset(sitar\_height\_random\_effects,sex=="M")$b)

sd(subset(sitar\_height\_random\_effects,sex=="F")$b)

##The following PDF is simply to help visualize the results##

pdf(file="height\_parameters\_by\_group.pdf")

ggplot(mdata,aes(x=factor(sex),y=value)) + geom\_boxplot() + facet\_wrap(~variable\_text,scales="free") + format\_plots + scale\_x\_discrete(name="\nSex") + scale\_y\_continuous(name="Parameter value, in arbitrary units,\nModeled Separately by Sex\n")

dev.off()

save.image(file="HEIGHT.RData")