Project Report: A Reproduction Analysis of Allele Frequency Differences Between Populations*

Bianca Kirsh Josh De Freitas

April 21, 2021

Abstract

This document presents the re-analysis of a published genomics study, including the written report, integrated code, and graphs.

Introduction

Overview of the original experiment

Large research efforts have been made to identify genomic regions, variants in candidate genes, and environmental factors that contribute to disease-associated health disparities across populations. In particular, SNPs have often been studied as a factor in a person's susceptibility of developing diseases.

In a study by Mao et al., worldwide differences in the effect allele frequency of 225 obesity-associated SNPs were investigated. The original analysis was conducted to identify significantly enriched or depleted effect alleles across 26 populations, which were later clustered using hierarchical clustering to gain insight into significance across continental groups. Additionally, researchers calculated the composite risk scores for each population to investigate the effect significantly enriched or depleted alleles have on individuals' genetic risk for obesity. Their results suggest that over 85% of effect alleles exhibit significant allele frequency differences and that population-level differences in genetic risk scores are correlated with a 2011 report from the Global Health Observatory, World Health Organization on obesity prevalence.

Goals and scope of the re-analysis

The original publication concluded that populations belonging to the same continental group showcased similar patterns in allele enrichment and depletion, suggesting that continent-specific environmental factors may shape the allele abundance of obesity-associated SNPs. In this re-analysis, however, we focus on the role of genetic factors, particularly ancestry-enrichment, by investigating admixed populations from geographically distant regions.

*The original study, **Population differentiation in allele frequencies of obesity-associated SNPs**, is available in BMC Genomics (https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-4262-9#Equ1).

For these purposes, we reduced our scope to a subset of samples, namely the African American (ASW), African (YRI), and European American (CEU) populations. The main goal of this project is to examine the relationship between obesity-related variants and ancestry to elucidate the genetic underpinnings of obesity susceptibility across populations. We will also explore the correlation between obesity prevalence and the effect allele frequency of obesity-associated SNPs in different populations.

Ultimately, this re-analysis will seek to illustrate the spectrum of genetic diversity between geographically distant genomes, as well as reveal ancestral biological signatures and systematic differences underlying the obesity epidemic. In doing so, we hope to encourage diversity and inclusion in genomics research, in addition to the investigation of genetic variants with extreme allele frequency differences (EAFD) between populations to help address health disparities across the world.

Methods

This project implements a pipeline that utilizes the R programming language to automate key components of data processing, analysis, and visualization. The pipeline employs a wide set of user-defined helper functions, which are documented and showcased in the Appendix, to break up large tasks into smaller steps, in addition to making the analysis easily customizable and expandable.

The current setup consists of a re-analysis of the original study, where only 3 out of the 26 populations will be investigated to explore the impact of ancestral genetics rather than environmental factors on obesity prevalence with greater depth.

Overview of the pipeline stages

Data download

For the first stage of the pipeline, we obtained a list of obesity-associated SNPs and their effect allele frequencies from the supplementary material available in the original publication. The authors compiled 225 SNPs from 29 GWA studies in the NHGRI-EBI GWAS Catalog, 19 of which were performed in European populations, 3 in East Asians, 2 in South Asians, 3 in Africans, and 2 in mixed ethnic populations. The corresponding effect allele frequencies were based on genotype information for 26 populations surveyed by the 1000 Genomes Project.

By utilizing pre-compiled, but not pre-processed data, we overcame computationally-intensive and resource-demanding challenges associated with retrieving data from the 1000 Genomes Project. At the same time, we ensured that this re-analysis was conducted on the raw data set used by the original publication, prior to any major filtering or manipulation steps made by the authors.

We retrieved obesity prevalence data from the same WHO report as the original study, but utilized the most recent statistics instead (2016).

As a limitation, data for the calculation of composite risk scores was not published in the original study. The formula used by the authors required information about the 225 SNPs for each person present in the 1000 Genomes Project, including the number of copies of effect alleles at obesity-associated variants and their effect size. Since these were not retrievable given the computational resources at hand, two additional analyses on the enrichment and depletion data were performed. An investigation of Venn diagrams and allele frequency distribution plots were also added to supplement the re-analysis.

Data pre-processing

The next stage involved selecting a subset of the data from the original publication. To compare the association signals of each variant as they relate to admixture, we chose to investigate the African American (ASW), African (YRI), and European American (CEU) populations.

The helper function created to assist in generating and populating data frames with the raw downloaded data is described in Appendix A: populateDfs.

Data analysis

The bulk of this re-analysis consisted of two main interrogations, namely the investigation of significant enrichment or depletion patterns and of extreme allele frequency differences.

For the first portion, we conducted hypergeometric tests to assess if the effect allele of each SNP was enriched or depleted in the 3 populations as compared to the overall population average. The results of these tests were the foundation for many subsequent steps, including the production of Venn diagrams, heat maps, and tabulated data. For the second portion, we utilized the allele frequency data from the original publication to plot the distribution of effect allele frequency across the three populations and investigate its correlation with obesity prevalence as reported by the WHO survey.

Quality control and filtering

As mentioned previously, the inputs of this re-analysis are essentially raw. Thus, we applied a series of processing steps for both statistical and visualization purposes.

The first filtering step consisted of selecting a subset of the 3 populations out of the original 26, as alluded to in the section above. As a caveat, the original publication included a 27th "population"," which pulls all 26 populations together and is referred to as the **global population** for hypothesis testing purposes. Thus, a "4th" population was similarly created for this smaller data set consisting of the ASW, YRI, and CEU groups in order to be used as the background set for the hypergeometric tests.

Then, we conducted two separate hypergeometric tests for each of the 225 SNPs, one for enrichment and one for depletion. This yielded 6 (2 x 3) tests per SNP, for a total of 1350 (6 x 225) tests across all 225 SNPs and 3 populations. With an overall significance level of 0.01, we applied Bonferroni correction to control the family-wise error rate (FWER) at a raw p-value cutoff of 7.4E-06 (0.01/1350). This first data set was used to produce the Venn diagrams and heat maps pertaining to the enrichment analysis.

For further inspection of a smaller subset of SNPs from the original 225 obesity-related variants, we replicated the orinal methods and selected those with enrichment or depletion log10 transformed p-values of at least 10E-100 and genome-wide significance ($5 \times$ 10E-08) in GWA studies. This second data set was used to produce the risk (effect) allele frequency distribution plots.

Once again, a set of helper functions were created to aid in data wrangling and hypothesis testing. These are available in Appendix B: hyperTest, renameColsRows, and combineHyper.

Data visualization

In this re-analysis, many hypothesis can be made based on the graphical depiction of data. Thus, we created different plots to assist in drawing conclusions about ancestry and admixture as they relate to obesity prevalence. The helper functions are available in Appendix C: makeVenn, makeHeatmap, plotDist, and plotCorrPops.

For clustering, we used the centroid option as the linkage method and a Pearson correlation-based distance to compute the dissimilarity matrix. These choices are common in gene expression analysis, where clusters of observations are formed based on overall profiles rather than magnitudes (e.g., similar genes are "up" and "down" together, regardless of their absolute expression value). Furthermore, they were made to reproduce the clustering algorithm employed by the tool used in the original study, dChip.

The plot that illustrates the correlation between effect allele frequencies and obesity prevalence was based on the average obesity rate of both sexes for two WHO regions: Africa and Europe. Because a linear relationship between the two variables was expected, we utilized Pearson's correlation.

Results

Finally, we ran the pipeline for the three pre-selected populations, CEU, ASW, and YRI, using another helper function described in Appendix D: runPipeline.

```
data <- read_excel("12864_2017_4262_MOESM1_ESM.xlsx")  # read the data from the original publi
colnames(data) <- data[2, ]  # the second row is now the header
data <- data[-c(1, 2), ]  # and the first two rows can be removed

myPop <- c("CEU", "ASW", "YRI")  # make a list of the populations under investigation

sigRes <- runPipeline(data, myPop, 0.01)
```

Patternicity of enrichment and depletion analysis

The first data interpretation step was based on the results of hypergeometric testing. At the FWER-adjusted p-value of 0.01, the enrichment analysis revealed significant results for the effect alleles of 80 SNPs out of the original 225 SNPs (35.55%). More specifically, 77 SNPs were significantly enriched in at least one population and 3 SNPs were significantly enriched in at least two populations (Figure 1). Similarly, 75 SNPs were significantly depleted in at least one population and 5 SNPs were significantly enriched in at least two populations (Figure 2). In comparison with the original analysis, where 195 out of 225 SNPs (86.7%) were significantly enriched or depleted in at least one of the 26 populations, we obtained far less significant results.

We produced a set of Venn diagrams to illustrate the differences and commonalities in enriched or depleted SNPs across the three populations. In general, the CEU population presented the greatest number of significantly enriched or depleted alleles, 76 and 71, respectively. On the other hand, the ASW yielded the least amount of significant results, with only 3 alleles in each category. The code blocks below indicate how the diagrams were generated.

```
enrVenn <- makeVenn(enrDf, myPop, 0.01)
enrVenn
```

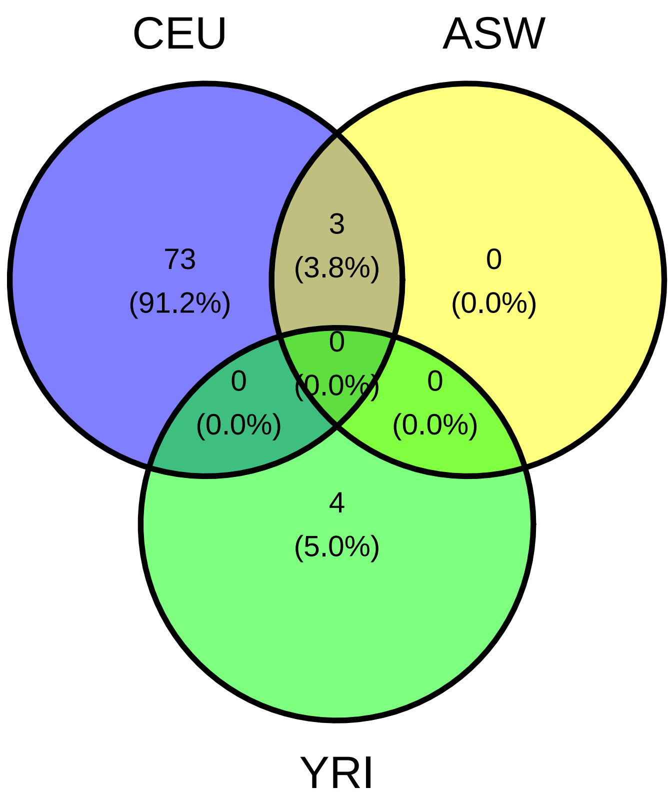

Figure 1: A Venn diagram of significantly enriched effect alleles in the ASW, YRI, and CEU populations

```
depVenn <- makeVenn(depDf, myPop, 0.01)
depVenn
```

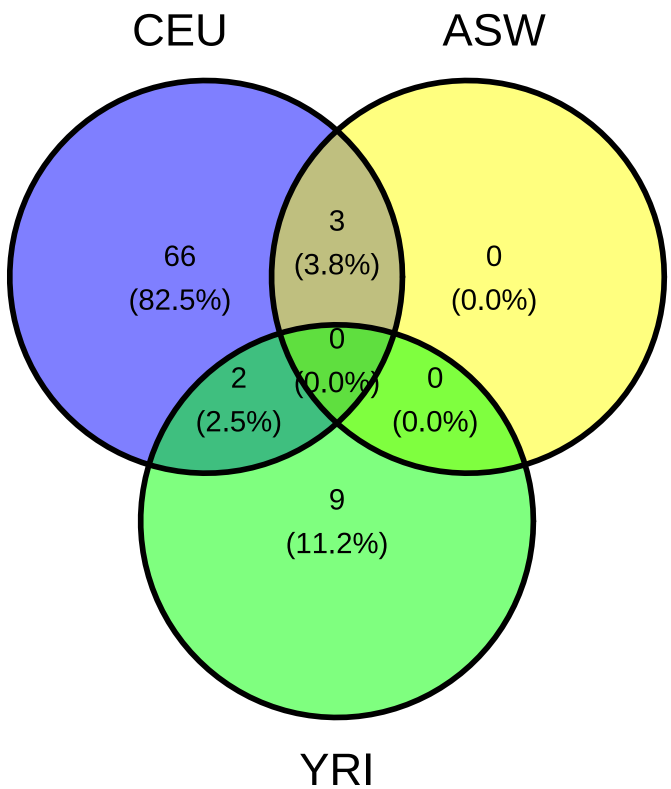

Figure 2: A Venn diagram of significantly depleted effect alleles in the ASW, YRI, and CEU populations

The results from hypergeometric testing were also utilized to generate a heat map and visualize allele enrichment and depletion patterns among the three populations (Figure 3). If an effect allele was enriched, the cell was colored red based on the negative value of log10P, whereas it was colored green based on the actual value of log10P if it was depleted. A hierarchical clustering of the three populations shows that the YRI and ASW populations generally share similar allele enrichment and depletion patterns. Nonetheless, there are also commonalities between the YRI and CEU populations. Lastly, cases of simultaneous significant enrichment in one population and depletion in another are observed between the CEU and ASW populations. In other words, they exhibit multiple SNPs with opposite directions in allele frequency changes, where effect alleles that are enriched in one population appear depleted in the other.

Overall, the YRI population appears to share intermediate features between ASW and CEU, which does not align with our initial hypothesis that the admixed ASW population would display features of both the acestral populations, CEU and YRI.

```
makeHeatmap(sigRes, "heatmap1.jpg")
```

```
## pdf
##   2
```

```r
knitr::include_graphics("heatmap1.jpg")
```

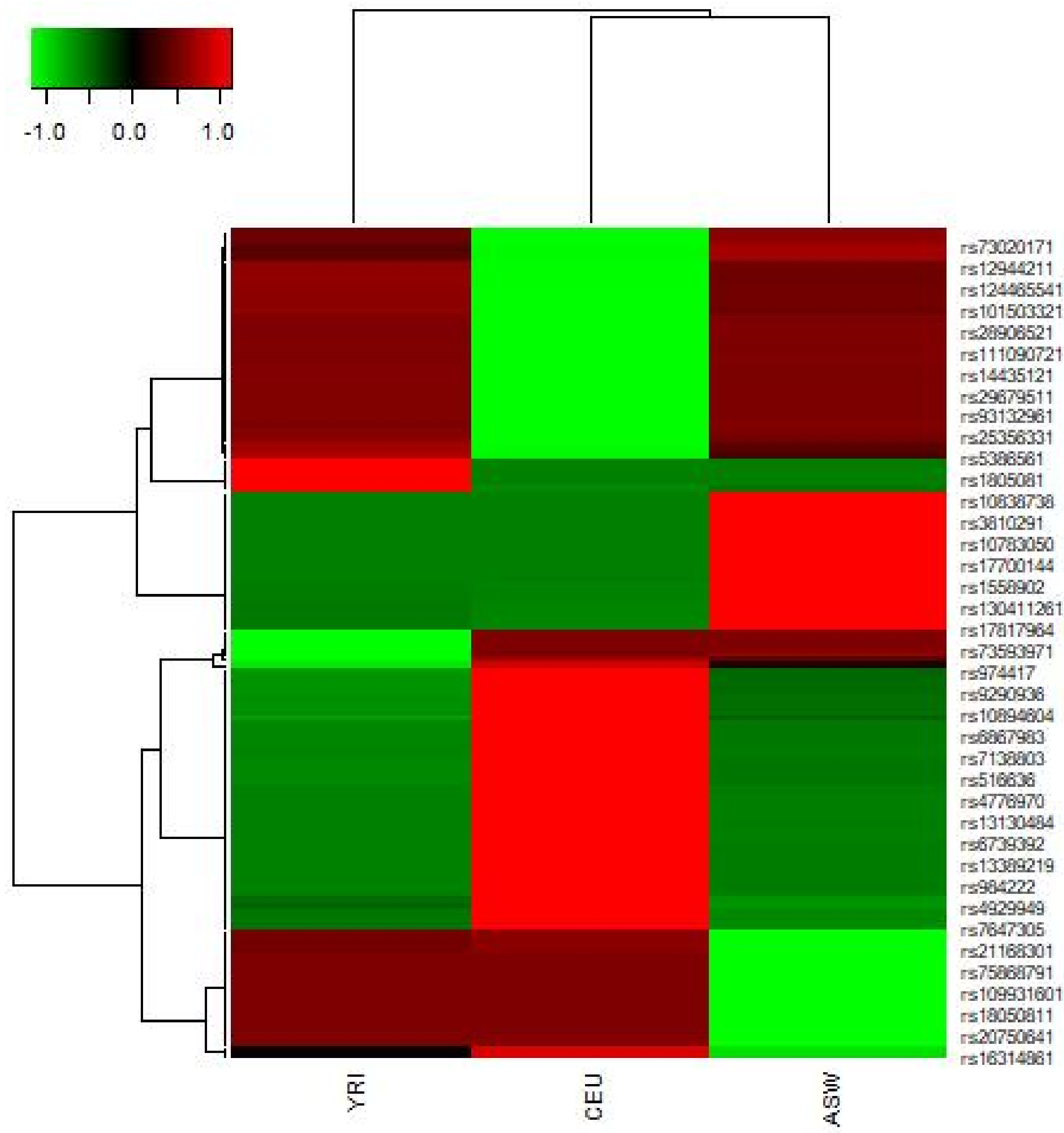

Figure 3: A heat map of significantly enriched and depleted effect alleles across the ASW, YRI, and CEU populations

Distribution of risk allele frequencies

The second investigation consisted of analyzing the differences in effect allele frequency by plotting their distributions across the three populations (Figure 4). Only a small subset of SNPs was considered at this stage, as determined by their significance in both the enrichment analysis and GWAS studies and described in Quality control and filtering.

In a second study by Mersha et al., the ASW population showcased allele frequencies intermediate to the ancestral CEU and YRI populations for asthma-related GWAS SNPs. The same statement applies to our subset of 8 obesity-related SNPs. Furthermore, we observed a bias towards YRI allele frequencies in the admixed ASW population, suggesting that it shares more similarities than with the CEU population for this criteria. It is important to notice the difference between our re-analysis and this study, both in terms of the studied disease (i.e., asthma vs. obesity) and number of plotted variants (i.e., 78 vs. 8), which makes the concordance of results especially convincing.

```
distPlot <- plotDist(sigRes, freqInfo)
distPlot
```

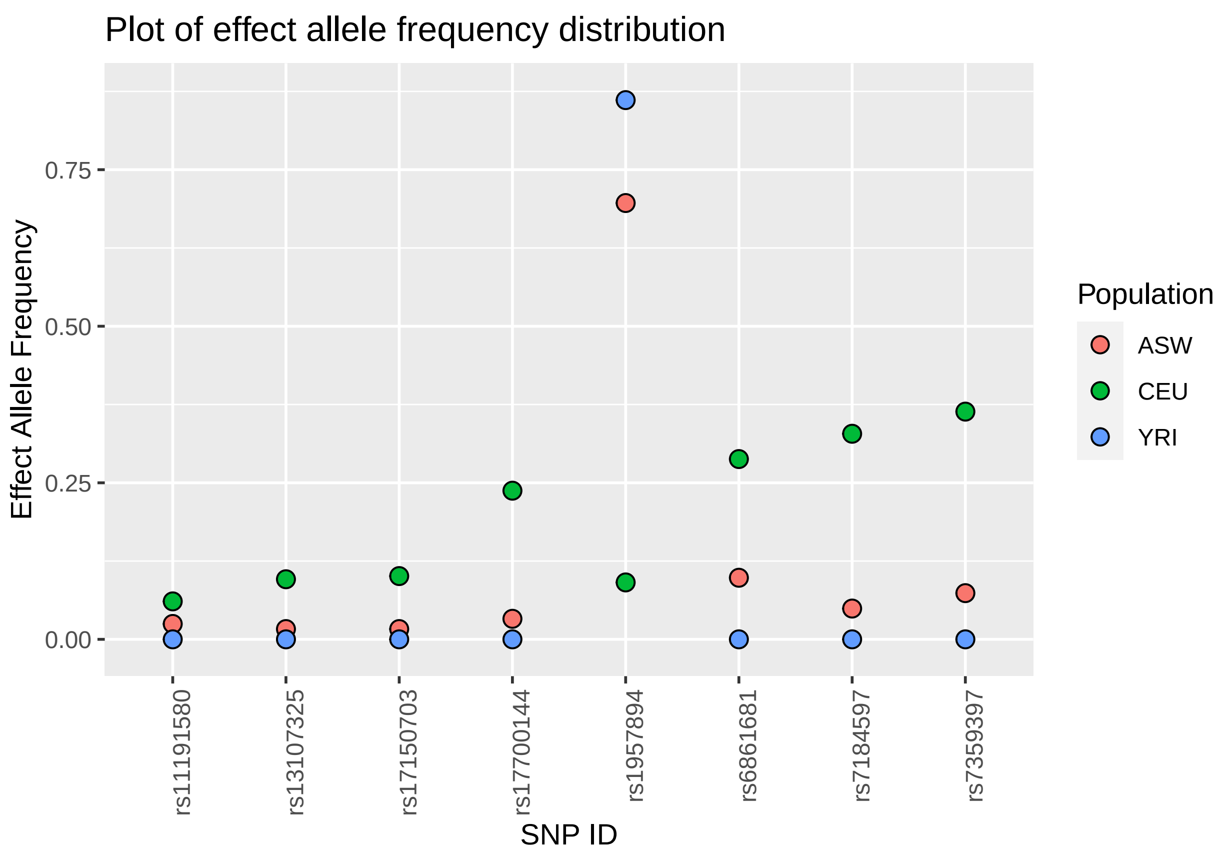

Figure 4: A frequency distribution plot of effect alleles for the most significant SNPs across the ASW, YRI, and CEU populations

The minor allele frequency (MAF) distribution of asthma-related GWAS SNP's across ASW, CEU, and YRI populations is shown below for comparison with the aforementioned publication (Figure 5).

```
## [1] TRUE
```

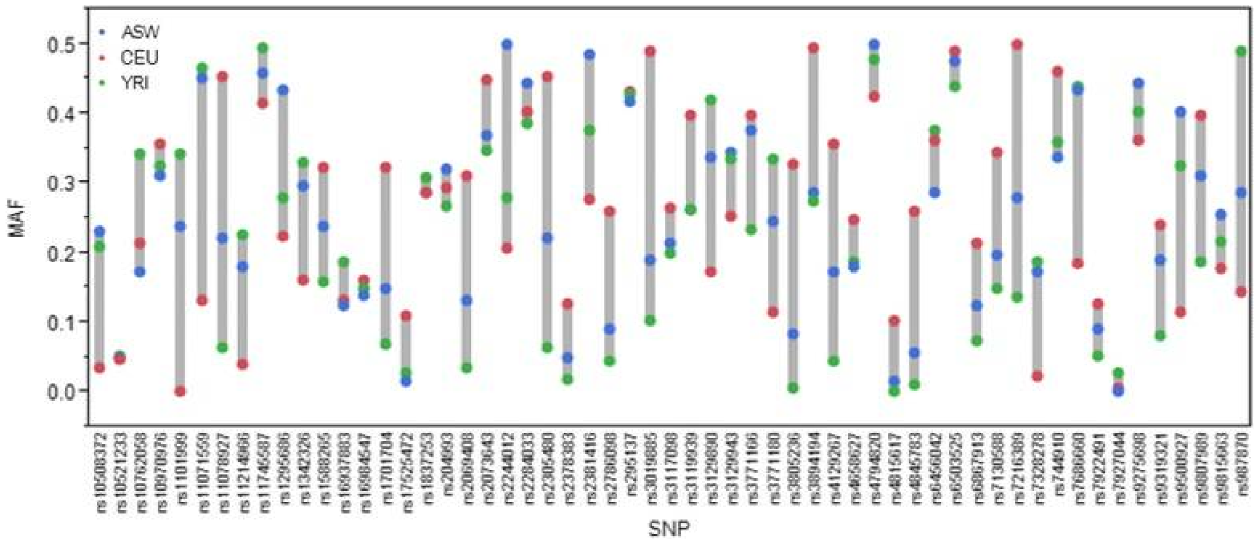

Figure 5: The minor allele frequency (MAF) distribution of asthma-related GWAS SNP's across the ASW, YRI, and CEU populations

Correlation between effect allele frequency and obesity prevelence

According to data published by the World Health Organization in 2016, the average obesity rates are much larger in Europe than in African countries. Similarly, the average effect allele frequency of the 225 obesity-associated SNPs in African sub populations, including ASW and YRI, is noticeably lower than that of European ones such as CEU (Figure 6). This suggests that the influence of ancestry and genetics on effect allele frequency may play a role in obesity prevalence among populations.

```
sex <- c("BTSX")
year <- c("2016")
superPop <- c("EUR", "AFR")
corrPlotPops <- plotCorrPops(WHO_BMI, freqInfo, sex, year, superPop,
    myPop)
corrPlotPops
```

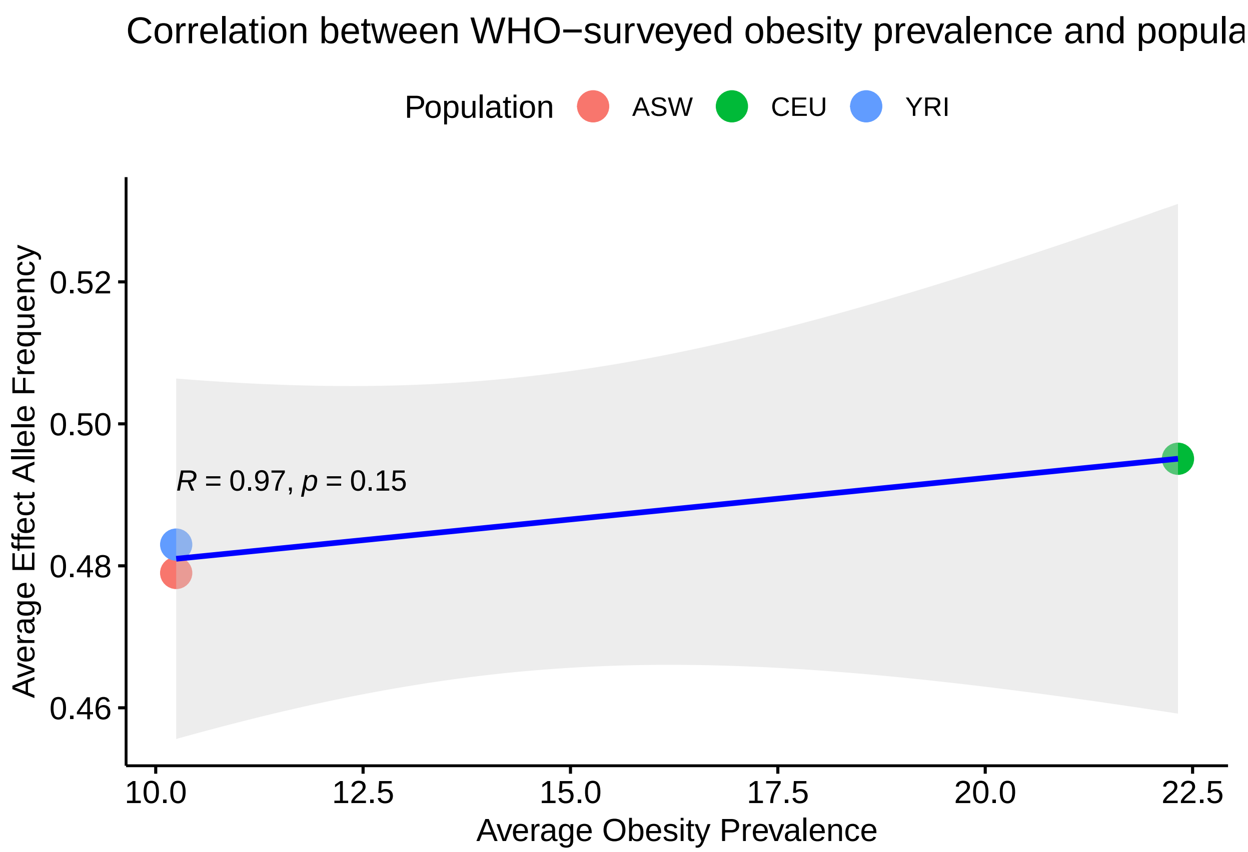

Figure 6: Correlation plot of WHO obesity prevelence and effect allele frequency for the ASW, YRI, and CEU populations

Lastly, to confirm that the hypergeometric tests and clustering methods applied to the 3 populations were in line with those from the original publication and to test the capabilities of extending the pipeline to other user-defined lists of populations, we also generated a heat map of the enrichment/depletion p-values (log10 based) of effect alleles across all 26 populations (Figure 7). The use of a pipeline for most aspects of this project greatly facilitated this step, as the only parameter to be modified was the populations under investigation. The last helper function is described in Appendix E: allPops.

This final stage of the re-analysis was crucial to validate the conclusions made using the replicated methods for a smaller subset of sampled populations, as it would reveal whether any inconsistencies were due to a faulty approach or simply an inability to reproduce the analysis with a different, reduced data set.

```
allPop <- allPops(data)  # make a list of all 26 populations included in the study
# now, re-run the pipeline up to the heat map generation
# step:
valRes <- runPipeline(data, allPop, 0.01)
```

```
## pdf
```

```
## 2
```

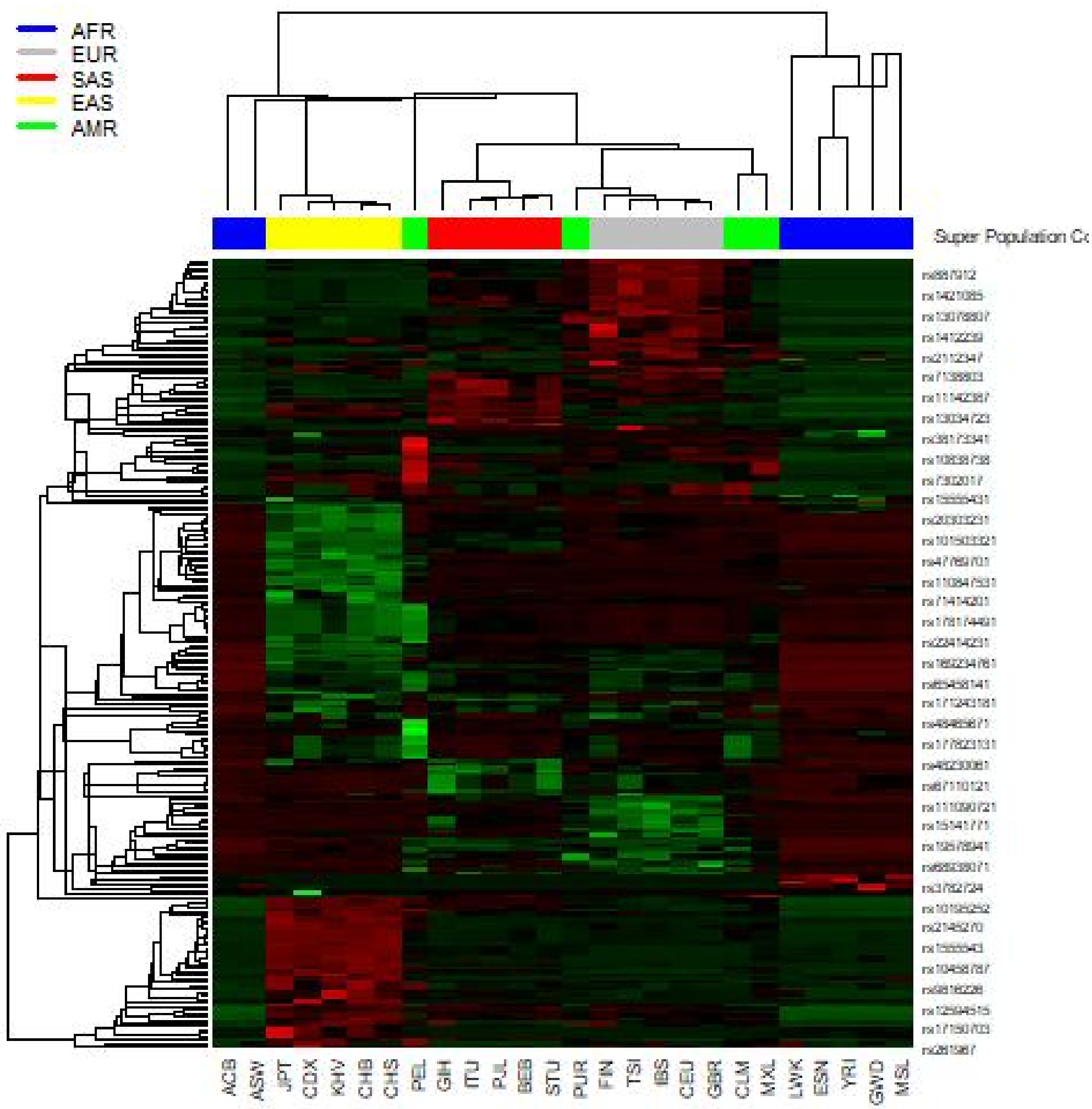

Figure 7: A heat map of significantly enriched and depleted effect alleles across the 26 populations

The original heat map showing significant enrichment and depletion of the effect alleles of 225 obesity risk SNPs across the 26 populations is shown below for comparison (Figure 8).

```
## [1] TRUE
```

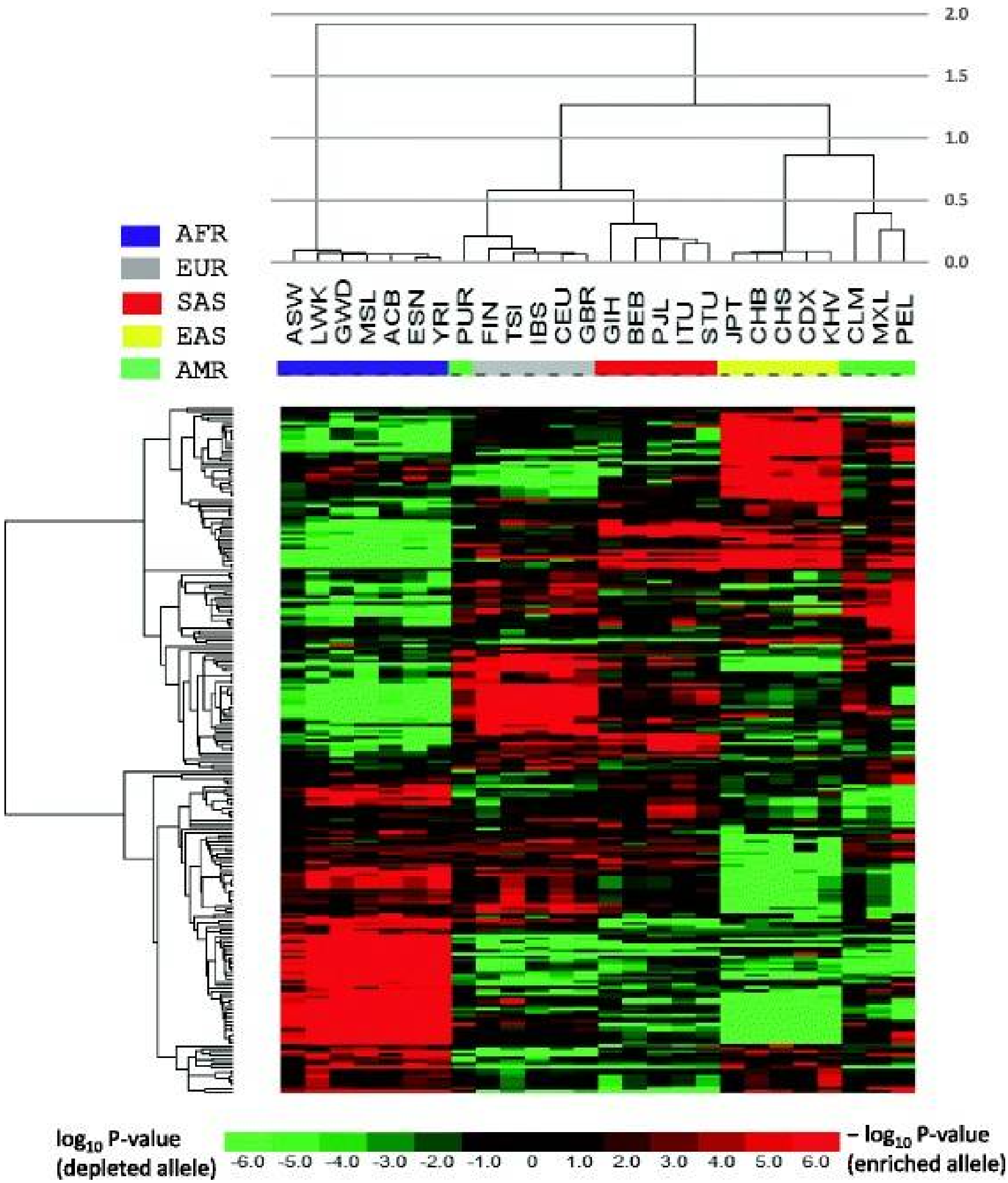

Figure 8: The original heat map showing significant enrichment and depletion of the effect alleles of 225 obesity risk SNPs across the 26 populations

Conclusions

In general, the results of this re-analysis are mostly consistent with the literature. The additional study that investigated asthma-related variants in precisely the same three populations, ASW, CEU, and YRI, came to the strikingly similar conclusion that the ASW population displayed intermediate allele frequencies of obesity-related SNPs to the ancestral YRI and CEU populations. This conclusion cannot be extended to enrichment and depletion patterns, as the YRI population appeared to showcase intermediate characteristics between the ASW and CEU populations. Nonetheless, this is still somewhat in line with another conclusion made by the original study, which stated that "it is the allele frequencies, not the number of obesity-associated SNPs, that determine the outcome of composite scores." Thus, in our case, it seems as though it is in fact the allele frequencies, not the under- or over-representation of SNPs, that determine obesity prevalence.

There are many possible reasons for the discrepancies between this project and the original publication. Firstly, the large difference in the proportion of SNPs that possess effect alleles significantly enriched or depleted in at least one of the populations (i.e., 86.7% vs. 35.55%) is most likely due to the gross reduction in sample size from 26 down to 3. This impacted the predictive power of hypergeometric tests and, thus, all the analysis downstream of it.

A second source of errors could have also been the tests themselves, or the choice of background set to be more specific. When we leveled-off the comparison of this re-analysis to the original study by considering all 26 populations, much more similar results were obtained. However, disparities still remained in the clustering of certain populations. We believe that this is attributed to the fact that the original authors utilized dCHIp, while we employed heatmap.3's library to generate results.

All together, we conclude that ancestral genetics might in fact play a role in the disease susceptibility. However, obesity and obesity-related traits are the result of the interaction between and within variants and the surrounding environment, so socio-economic factors such as diet, climate, local pathogens and lifestyle must also be taken into account.

Appendix

Appendix A: populateDfs

```
# function: 'populateDfs': a function to populate data frames
# with user-specified population data from the original study
# argument(s): 'input': the table from the original
# publication 'list': a vector of population codes to be
# investigated in the re-analysis, as per 1000 Genomes
# convention 'rows': a vector of row names (i.e., SNP IDs)
# 'cols': a vector of column names (i.e., SNP information)
# return value(s): 'popInfo': a data frame of
# population-specific effect allele frequency data
# 'backgroundInfo': a data frame of global population SNP
# data, which consists of all user-specified populations
# pulled together 'XINfo': n data frames of
```

```r
# population-specific SNP data, whcih consists of the
# breakdown of 'backgroundInfo' into the distinct populations
populateDfs <- function(input, list, rows, cols) {
    popInfo <- data.frame(row.names = rows)
    backgroundInfo <- data.frame(matrix(nrow = length(rows),
        ncol = length(cols)))

    colnames(backgroundInfo) <- cols
    backgroundInfo$population <- "ALL"
    backgroundInfo$`GWAS P-value` <- input$`GWAS P-value`
    backgroundInfo[is.na(backgroundInfo)] <- 0
    for (i in seq(from = 4, to = ncol(input), by = 5)) {
        population <- toString(input[1, i])
        if (population %in% list) {
            popInfo[population] <- input[, i + 4]
            assign(paste(population, "Info", sep = ""), input[,
                i:(i + 4)], envir = .GlobalEnv)
            backgroundInfo["effect allele number"] <- as.numeric(backgroundInfo[["effect allele
                as.numeric(input[["effect allele number"]])
            backgroundInfo["other allele number"] <- as.numeric(backgroundInfo[["other allele n
                as.numeric(input[["other allele number"]])
            backgroundInfo["total allele number"] <- as.numeric(backgroundInfo[["total allele n
                as.numeric(input[["total allele number"]])
            backgroundInfo["effect allele frequency"] <- as.numeric(backgroundInfo[["effect al
                as.numeric(input[["effect allele frequency"]])
        }
    }

    backgroundInfo["effect allele frequency"] <- backgroundInfo["effect allele frequency"]/leng
    return(list(popInfo, backgroundInfo))

}
```

Appendix B: hyperTest, renameColsRows, and combineHyper

```r
# function: 'hyperTest': a function to perform the enrichment
# analysis using Fisher's two-tail exact test (two
# hypergeometric tests, one for enrichment and another for
# depletion, are conducted per SNP) argument(s): 'list': a
# vector of population codes to be investigated in the
# re-analysis, as per 1000 Genomes convention 'background': a
# data frame of global population SNP data, which consists of
# all user-specified populations pulled together return
# value(s): 'enrichmentDf': a data frame containing the
# results of the enrichment analysis 'depletionDf': a data
# frame containing the results of the depletion analysis
```

```
hyperTest <- function(list, background) {
    enrichmentDf <- data.frame()
    depletionDf <- data.frame()

    popDfs <- paste(list, "Info", sep = "")

    for (i in 1:nrow(background)) {
        a <- c()
        b <- c()
        for (df in popDfs) {
            info <- get(df)
            m <- as.numeric(background$`effect allele number`[i])  # total number of effect al
            n <- as.numeric(background$`other allele number`[i])  # total number of other alle
            k <- as.numeric(info$`total allele number`[i])  # total number of alleles (effect
            q <- as.numeric(info$`effect allele number`[i])  # number of effect alleles select

            # phyper(q,m,n,k) = phyper(success-in-sample,
            # success-in-bkgd, failure-in-bkgd, sample-size)
            a <- cbind(a, phyper(q = q - 1, m = m, n = n, k = k,
                lower.tail = FALSE))  # over-representation (enrichment)
            b <- cbind(b, phyper(q = q, m = m, n = n, k = k,
                lower.tail = TRUE))  # under-representation (depletion)
        }
        enrichmentDf <- rbind(enrichmentDf, a)
        depletionDf <- rbind(depletionDf, b)
    }
    return(list(enrichmentDf, depletionDf))
}
# function: 'renameColsRows': a function to re-name data
# frames, modifying rows to SNP IDs columns to population IDs
# argument(s): 'dfs': a list of data frames (as strings) to
# be modified 'rows': a vector of row names (i.e., SNP IDs)
# 'list': a vector of population codes to be investigated in
# the re-analysis, as per 1000 Genomes convention return
# value(s): NA (modifies the data frames in the global
# environment)
renameColsRows <- function(dfs, rows, list) {
    for (df in dfs) {
        df.tmp <- data.frame(get(df))
        colnames(df.tmp) <- list
        rownames(df.tmp) <- rows
        assign(df, df.tmp, envir = .GlobalEnv)
    }
}
# function: 'combineHyper': a function to combine two data
# frames, namely the ones produced by hyperTest (enrichment
# and depletion) argument(s): 'dfs': a list of data frames
```

```r
# (as strings) to be modified 'rows': a vector of row names
# (i.e., SNP IDs) return value(s): hyperRes': a data frame
# holding the combined results of the enrichment/depletion
# analysis
combineHyper <- function(dfs, rows) {
    hyperRes <- data.frame()
    for (df in dfs) {
        df.tmp <- get(df)
        df.tmp$type <- sub("Df", "", df)
        df.tmp$`GWAS P-value` <- as.numeric(data$`GWAS P-value`[match(rownames(df.tmp),
            rows)])
        assign(df, df.tmp)
        hyperRes <- rbind(hyperRes, df.tmp)
    }
    return(hyperRes)

}
```

Appendix C: makeVenn, makeHeatmap, plotDist, and plotCorrPops

```r
# function: 'makeVenn': a function to create a Venn diagram
# that illustrates the differences and similarities between
# the populations under investigation argument(s): 'df': a
# data frame of p-values (either enrichment or depletion)
# 'list': a vector of population codes to be investigated in
# the re-analysis, as per 1000 Genomes convention 'sig': the
# significance level (alpha) return value(s): 'plot': a Venn
# diagram
makeVenn <- function(df, list, sig) {
    cutoff <- sig/(2 * length(list) * nrow(df))  # control the family-wise error rate (FWER) f

    filteredDf <- data.frame(df < cutoff)
    table <- table(rowSums(filteredDf))
    plot <- ggvenn(filteredDf)
    return(plot)
}
# function: 'getSig': a function to filter the output of the
# enrichment/depletion analysis for significant results
# argument(s): 'df': a data frame of p-values (either
# enrichment or depletion) 'list': a vector of population
# codes to be investigated in the re-analysis, as per 1000
# Genomes convention 'sig': the significance level (alpha)
# return value(s): 'logDf': a data frame with the log10
# transformed p-values of significantly enriched and depleted
# effect alleles
getSig <- function(df, list, sig) {
```

```
    cutoff <- sig/(2 * length(list) * nrow(df))
    sigDf <- df[!rowSums(df[, 1:(ncol(df) - 2)] < cutoff) ==
        0, , drop = FALSE]  # remove rows whose raw p-value is not significant at alpha
    logDf <- sigDf %>% mutate_at(1:(ncol(sigDf) - 2), log10)  # log10 transform p-values
    return(logDf)
}
# function: 'makeHeatmap': a function to create a heat map
# that illustrates the enrichment/depletion of significant
# SNPs across the populations under investigation
# argument(s): 'df': a data frame with the log10 transformed
# p-values of significantly enriched and depleted effect
# alleles 'groupColours': a vector that annotatess the column
# values with colours 'fname': output file name for heat map
# return value(s): NA (saves the heat map as a .jpeg file to
# the current directory)
makeHeatmap <- function(df, fname, groupColours) {
    heatmapDf <- df
    heatmapDf[heatmapDf$type == "enr", ] <- heatmapDf[heatmapDf$type ==
        "enr", ] %>% mutate_at(1:(ncol(heatmapDf) - 2), ~. *
        -1)
    # the negative of log10 of the p-value (a positive number) is
    # used to represent enriched effect allele of a SNP for a
    # population in the heatmap the actual value of log10 of the
    # p-value (a negative number) is used to represent depleted
    # effect allele of a SNP for a population in the heatmap

    heatmapMtx <- as.matrix(heatmapDf[, 1:(ncol(heatmapDf) -
        2)])  # create a matrix from the FWER filtered, combined, log-transformed, and additiv
    heatmapMtx[!is.finite(heatmapMtx)] <- 0
    rm <- apply(heatmapMtx, 1, var) == 0
    y <- heatmapMtx[!rm, ]
    hr <- hclust(as.dist(1 - cor(t(y), method = "pearson")),
        method = "centroid")
    hc <- hclust(as.dist(1 - cor(y, method = "pearson")), method = "centroid")
    col_fun = colorRampPalette(c("green", "black", "red"))(1024)
    jpeg(file = fname)
    if (missing(groupColours)) {
        heatmap3(y, Rowv = as.dendrogram(hr), Colv = as.dendrogram(hc),
            col = col_fun, scale = "row", margins = c(3, 7),
            cexRow = 0.8, cexCol = 1, file = "heatmap.pdf")
    } else {
        heatmap3(y, Rowv = as.dendrogram(hr), Colv = as.dendrogram(hc),
            ColSideColors = groupColours, legendfun = function() showLegend(legend = c("AFR",
                "EUR", "SAS", "EAS", "AMR"), col = c("blue",
                "grey", "red", "yellow", "green")), col = col_fun,
            scale = "row", margins = c(3, 7), cexRow = 0.8, cexCol = 1,
            file = fname)
```

```r
    }
    dev.off()
}

# function: 'plotDist': a function to plot the distribution
# of effect allele frequencies for select SNPs argument(s):
# 'df': a data frame with the log10 transformed p-values of
# significantly enriched and depleted effect alleles
# 'frequency': a data frame containing the effect allele
# frequencies for the select populations return value(s):
# 'plot': a distribution frequency plot
plotDist <- function(df, frequency) {
    freqDf <- df[!rowSums(abs(df[, 1:(ncol(df) - 2)]) < 1e-99) ==
        0, , drop = FALSE]  # remove rows whose log10 transformed p-value is not significant a
    freqDf <- freqDf[freqDf$`GWAS P-value` < 5e-08, ]  # remove rows whose SNPs has not reache
    filteredSNP <- rownames(freqDf)

    frequency["id"] <- rownames(frequency)
    meltedPopDf <- melt(frequency, id.vars = "id", value.name = "MAF",
        variable.name = "Population")
    filteredPopDf <- meltedPopDf[meltedPopDf$id %in% filteredSNP,
        ]
    plot <- ggplot(filteredPopDf, aes(x = id, y = as.numeric(MAF),
        fill = Population)) + geom_dotplot(binaxis = "y", stackdir = "center")
    plot <- plot + ggtitle("Plot of effect allele frequency distribution") +
        xlab("SNP ID") + ylab("Effect Allele Frequency")
    plot <- plot + theme(axis.text.x = element_text(angle = 90,
        hjust = 1))
    return(plot)
}

# function: 'plotCorrPops' argument(s): 'df': a data frame
# with the WHO data (BMI = 30, age-standardized) 'freq': a
# data frame containing the effect allele frequencies for the
# select populations 'sex': a vector containing the sexes to
# filter by 'superPop': a vector containing the super
# populations to be investigated 'list': a vector of
# population codes to be investigated in the re-analysis, as
# per 1000 Genomes convention return value(s): 'plot': a
# correlation plot of average obesity prevalence and average
# effect allele frequency
plotCorrPops <- function(df, freq, sex, year, superPop, list) {
    filtdf <- df[df$REGION %in% superPop, ]
    filtdf <- filtdf[filtdf$SEX %in% sex, ]
    filtdf <- filtdf[filtdf$YEAR %in% year, ]
    filtdf <- select(filtdf, REGION, Numeric)
    freq <- freq[, colnames(freq) %in% list]
```

```
    meanAF <- as.data.frame(colMeans(mutate_all(freq, function(x) as.numeric(as.character(x))))
    colnames(meanAF) <- "MeanEffectAlleleFrequency"
    pop1 <- subset(filtdf, REGION == superPop[1])
    pop2 <- subset(filtdf, REGION == superPop[2])
    pop2 <- na.omit(pop2)
    pop1 <- na.omit(pop1)
    pop2_mean <- mean(pop2$Numeric)
    pop1_mean <- mean(pop1$Numeric)

    avg_prev <- c(pop2_mean, pop1_mean, pop2_mean)
    meanAF$AveragePrevalence <- avg_prev

    corr_data <- meanAF
    corr_data$Population <- rownames(corr_data)

    plot <- ggscatter(corr_data, x = "AveragePrevalence", y = "MeanEffectAlleleFrequency",
        add = "reg.line", conf.int = TRUE, color = "Population",
        shape = 19, size = 5, add.params = list(color = "blue",
            fill = "lightgray"), cor.coef = TRUE, xlab = "Average Obesity Prevalence",
        ylab = "Average Effect Allele Frequency") + stat_cor(aes(color = Population),
        method = "pearson", label.x = 0.3)
    plot <- plot + ggtitle("Correlation between WHO-surveyed obesity prevalence and population-
    return(plot)
}
```

Appendix D: runPipeline

```
# function: 'runPipeline': a function to run the entire
# pipeline for the oriinal study data and user-defined
# populations argument(s): 'input': a data frame containing
# data from the original publication 'list': a vector of
# population codes to be investigated in the re-analysis, as
# per 1000 Genomes convention 'sig': the significance level
# (alpha) return value(s): 'results': a data frame holding
# the combined results of the enrichment/depletion analysis
runPipeline <- function(input, list, sig) {
    varInfo <- input[, 1:3]  # gather general SNP data (SNP ID, effect allele, other allele)
    rowData <<- input$`SNP ID`
    colData <<- c("population", "effect allele number", "other allele number",
        "total allele number", "effect allele frequency")
    otherInfo <- populateDfs(input, list, rowData, colData)
    freqInfo <<- otherInfo[[1]]  # gather effect allele frequency data for the select populati
    bckgdInfo <<- otherInfo[[2]]  # gather general SNP data for the global average of select p
    hyperRes <<- hyperTest(list, bckgdInfo)
    enrDf <<- hyperRes[[1]]  # gather enrichment analysis results for the 225 SNPs and select p
    depDf <<- hyperRes[[2]]  # gather depletion analysis results for the 225 SNPs and select p
```

```
    renameColsRows(c("enrDf", "depDf"), rowData, list)
    hyperComb <<- combineHyper(c("enrDf", "depDf"), rowData)
    results <- getSig(hyperComb, list, sig)
    return(results)
}
```

Appendix E: allPops

```
# function: 'allPops': a function to retrieve the names of
# all surveyed populations (26) argument(s): 'input': the
# table from the original publication return value(s):
# 'list': a vector of population codes reported in the study,
# as per 1000 Genomes convention
allPops <- function(input) {
    list <- c()
    for (i in seq(from = 4, to = ncol(input) - 5, by = 5)) {
        population <- toString(input[1, i])
        list <- append(list, population)
    }
    return(list)
}
```

Bibliography

1. Mao, L., Fang, Y., Campbell, M., & Southerland, W. M. (2017). Population differentiation in allele frequencies of obesity-associated SNPs. BMC genomics, 18(1), 861. https://doi.org/10.1186/s12864-017-4262-9

2. Mersha, T. B., & Abebe, T. (2015). Self-reported race/ethnicity in the age of genomic research: its potential impact on understanding health disparities. Human genomics, 9(1), 1. https://doi.org/10.1186/s40246-014-0023-x

3. Global Health Observatory, World Health Organization (The 2017 update). Prevalence of obesity among adults, BMI 30, age-standardized [Data table]. Retrieved from https://apps.who.int/gho/data/view.main.CTRY2450A?lang=en

4. Li, C., & Wong, W. H. (2001). Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection (dChip). Proceedings of the National Academy of Sciences of the United States of America, 98(1), 31–36. https://doi.org/10.1073/pnas.011404098