Assessing Disease Risk Through the Use of Individual-Level 23andMe Data

Jay Mantuhac

EPIDEM 275: Bioinformatics

Professor Trina Norden-Krichmar

June 9, 2021

Abstract

Direct-to-consumer testing is a powerful tool that allows individuals to learn about their own genetic risk for disease. When given a list of all single nucleotide polymorphisms that are present in an individual's genome and merging it with a category of genetic wide association studies, we can learn more about the kinds of diseases an individual may be at risk for developing over the life course. Using 3 semi-randomly selected individual-level 23andMe datasets extracted from the Personal Genome Project (based in Harvard University), I aimed to replicate the process that direct to consumer testing companies use in order to provide these insights, with the goal of coming up with a list of 10 diseases and disease traits that each selected individual is at most risk for. After extracting the data from the Personal Genome Project website in R, each dataset was then merged with the EMBL-EBI GWAS Catalog, a catalog that links together studied SNPs with respective genome wide association studies, using the gwas package in R. Finally, risk calculations were conducted in order to assess the true risk of disease traits for each subject, with the end result being a "Top 10" list of disease traits subjects demonstrated. All 3 selected subjects were found to have disease traits that related to cardiovascular disease, including the presence of SNPs that code for pulse pressure and HDL cholesterol. Upon generating the results of this analysis, I found that the primary limitation of this analysis was the fact that merging data with the catalog of genetic wide association studies allowed me to generate a list of disease-linked traits, rather than linking to diseases themselves. As such, I found that the entries seen in the top 10 list are not necessarily correlated with the presence of actual disease. Other limitations of the study include the overrepresentation of White Americans in the final sample and the fact that this analysis did not take into account the role that environmental (and other non-genetic factors) can play in the overall risk of disease. Future

research should expand on the methods used in this project by quantifying the true role that genetics plays in the development of certain diseases given the influence of environmental factors.

Introduction

In-vitro diagnostics (IVDs), also known as Direct to Consumer (DTC) tests, allow for individuals to gain insights on their own genetic data without the presence of a healthcare provider.[1] Through this process, interested consumers send a sample containing genetic data, usually saliva or urine, to a company, which the company then uses to sequence the consumer's entire genome. The resulting output of this process, which the consumer receives, can include a wide variety of documents, including a dataset that details the single nucleotide polymorphisms (SNPs) present in the consumer's genome, as well as reports that a company can produce that give insights from the resulting genetic data, including the consumer's evolutionary origins and potential risk for diseases. 23andMe is one company that provides DTC testing for individual genetic data, and, depending on the type of product purchased from 23andMe, can provide consumers with reports that relate to an individual's biological ancestry and/or reports that give consumers insights about the consumer's Genetic Health Risk (GHR).[2] For companies such as 23andMe, the primary appeal for consumers to purchase their service is the fact that consumers can be empowered to make decisions revolving their own health and behaviors in response to the insights that are provided by the company. For instance, a consumer who discovers an increased genetic risk for Type II Diabetes may be more empowered to control their diet in an effort to counteract the inherent genetic risk that 23andMe may discover.

After consumers get the corresponding outputs from using direct to consumer testing services, which typically includes a dataset that details all SNPs present in their genome, consumers may voluntarily submit their genome dataset to the Personal Genome Project, an international, volunteer built database of individual-level genomes, traits, and health data. For the United States, this project was established in 2005 by Harvard University and serves to build a

database for consumers who are based in the United States.[3] The data that is submitted to Harvard's Personal Genome Project is publicly available, and as such, individuals and organizations are able to access the genetic data of other people for use in studies and personal projects.

The following project is an exploratory project meant to replicate the process that 23andMe (among other companies) engages in after sequencing an interested client's genotype in order to provide its consumers with insights related to individual genetic health risk. The primary question of interest that this project answers is the following: Given a person's entire dataset of single nucleotide polymorphisms present in their genome, what kinds of diseases (as well as other traits) would that individual be most at risk for?

Materials and Methods

For this project, 3 individual-level 23andMe datasets were semi-randomly selected and extracted directly from the Personal Genome Project website. Initially, 2 subjects were selected specifically due to reporting lactose intolerance, with 1 subject without lactose intolerance acting as a control. Although this criteria of lactose intolerance was eventually thrown out in the selection of subjects, the 3 subjects that were selected under this pretense were still allowed to be included in the final sample.

The datasets were individually loaded into R, either by downloading the .txt file onto my local machine or by using R code to read the dataset from a URL to the dataset on the Personal Genome Project website. All datasets that were selected followed the same format, in which all datasets contained 4 columns: the rsid number of a given SNP, the chromosome the SNP is located, the location of the SNP on the chromosome, and the actual genotype of the given SNP.

Upon reading all 3 datasets into R, the first set of analysis conducted was to visualize the SNP distribution across all chromosomes in R, in which a histogram is created that displays the number of SNPs present across each chromosome.

Following this extraction of SNP distribution, each of the 3 datasets was merged (via an inner join) with the EMBL-EBI GWAS Catalog using the gwascat() package in R. The package allows us to link together a given genotype in the subject's dataset to a studied and documented disease trait of interest in the larger catalog of genome wide association studies.

Once all 3 datasets have been merged to the EMBL-EBI GWAS Catalog, the overall disease risk for a given disease was calculated given the number of risk alleles that were present in an individual's genome. Overall risk, for this project, is described as a function of alleles that have 2 copies of the risk SNP of interest (i.e. "double risk alleles") and alleles that have only 1 copy of the risk SNP of interest (i.e. "single risk alleles"). Overall risk is defined as the proportion of double risk alleles (risk $count_2$) over a denominator that represents the number of potential alleles that could have been present in the genome added to the proportion of single risk alleles (risk $count_1$) given that same denominator.[4] The resulting overall risk value is then multiplied by 100 in order to display the value as a percentage. The formula for calculating overall risk is summarized below.

$$overall\ risk = \left(\frac{risk\ count_2}{(risk\ count_1 * 2) + risk\ count_2} + \frac{risk\ count_1}{(risk\ count_1 * 2) + risk\ count_2}\right) * 100$$

Figure 1. *The formula used for calculating overall risk*

The resulting output from this calculation for overall risk (after cleaning the data for any unnecessary columns that are a part of the calculation) is a data frame that consists of linked disease traits and the calculated overall risk. This data table is then organized by decreasing overall risk values, such that disease traits that have the highest risks are displayed first. The

kable() package in R was then used in order to format and display the “Top 10” disease traits that had the highest risk values and export the final tables as .png files for display.

As secondary analysis, I also referred back to the Personal Genome Project pages for each of the selected subjects, which contained responses to surveys (distributed by the Personal Genome Project) that allowed participants to report any health conditions that they had at the time of submitting their genome. There, I checked to see if the top hits for disease traits corresponded to the manifestation of corresponding diseases.

All data wrangling and analysis, as well as table formatting was conducted using R Version 4.0.5 (“Shake and Throw”).

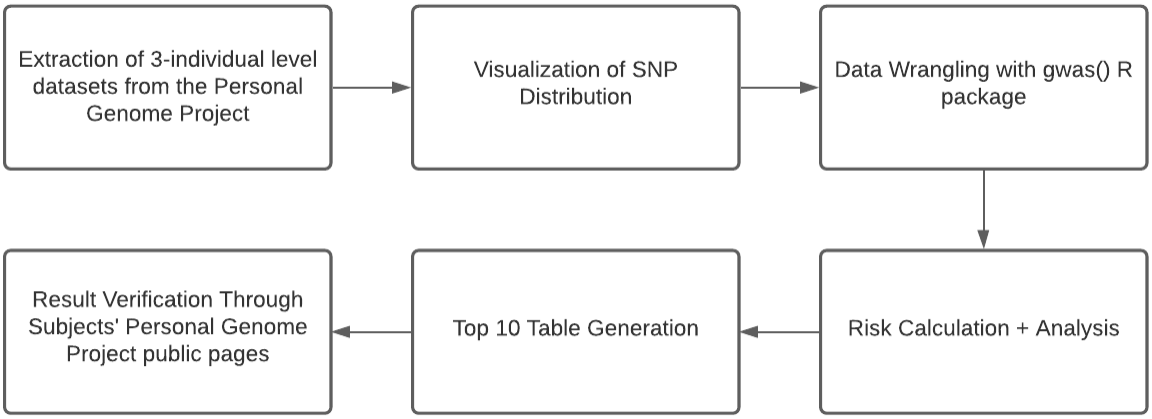

Figure 2. *Summary flow chart of project methods*

Results

One of the first things that can be pointed out are similarities found in the overall distribution of SNPs among all 3 subjects. Subject 1 (Figure 3) and Subject 3 (Figure 5) have incredibly similar SNP distributions, in which the largest number of SNPs can be found on either Chromosome 1 or 2, which contain approximately 75,000 SNPs on both chromosomes each. On the other hand, Subject 2 (Figure 4) seems to show a SNP distribution similar to what is seen in

the other subjects, however, the maximum number of SNPs found out of all chromosomes is approximately 45,000 SNPs found on Chromosome 1, which is approximately half of the maximum number of chromosomes present for Subject 1 and Subject 3.

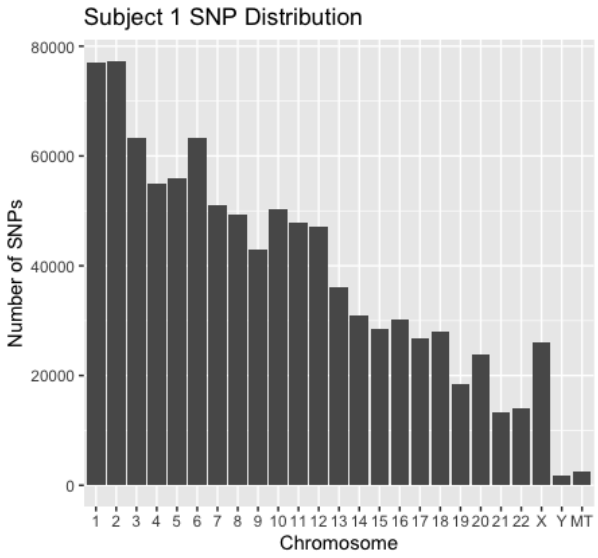

Figure 3. *Distribution of SNPs for Subject 1*

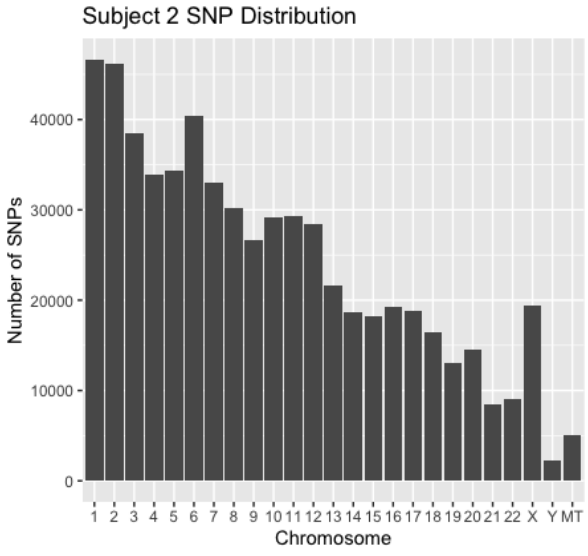

Figure 4. *Distribution of SNPs for Subject 2*

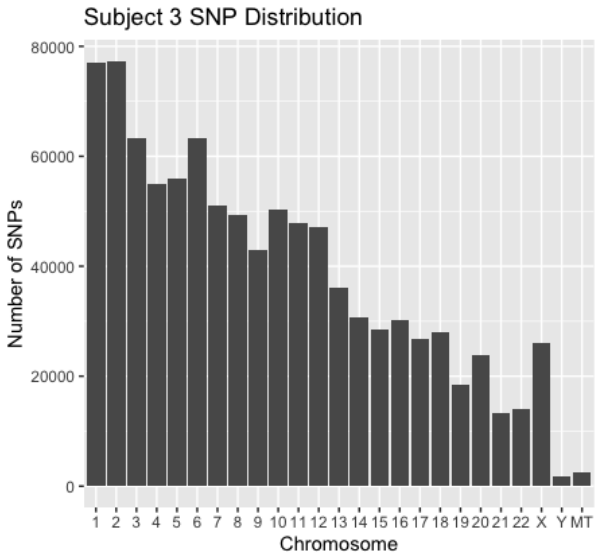

Figure 5. *Distribution of SNPs for Subject 3*

When looking at the Top 10 disease traits for each subject, several things can be noticed. Two main disease traits that are present among all subjects (Figure 6, Figure 7, and Figure 8) include traits that code for LDL cholesterol levels, HDL cholesterol levels, IgG glycosylation, and Heel bone mineral density. In addition, 2 of the 3 subjects demonstrated disease traits that correspond to Pulse Pressure (Figure 6 and Figure 7), C-reactive protein levels (Figure 6 and Figure 8), High density lipoprotein cholesterol levels (Figure 6 and Figure 8), and Total cholesterol levels (Figure 6 and Figure 8).

Disease Trait	Overall Risk
Total cholesterol levels	63.73626
Post bronchodilator FEV1/FVC ratio	62.50000
Pulse pressure	62.35294
C-reactive protein levels	61.68224
High density lipoprotein cholesterol levels	61.25000
Low density lipoprotein cholesterol levels	61.00324
Post bronchodilator FEV1	60.86957
LDL cholesterol levels	60.81081
Heel bone mineral density	60.58932
IgG glycosylation	60.35156

Figure 6. *Subject 1 Top Disease Traits*

Disease Trait	Overall Risk
Hip circumference adjusted for BMI	65.21739
High density lipoprotein cholesterol levels	62.97071
HDL cholesterol levels	62.86645
Pulse pressure	61.65644
Waist circumference adjusted for BMI in active individuals	61.44578
Triglycerides	61.17424
Heel bone mineral density	60.92896
LDL cholesterol levels	60.77586
IgG glycosylation	60.57971
Waist circumference adjusted for body mass index	60.13072

Figure 7. *Subject 2 Top Disease Traits*

Disease Trait	Overall Risk
C-reactive protein levels	63.46154
HDL cholesterol levels	63.19613
Heel bone mineral density	62.66667
Blood metabolite levels	62.17228
Blood protein levels	61.94030
Total cholesterol levels	60.70959
IgG glycosylation	60.35156
LDL cholesterol levels	60.20067
Estimated glomerular filtration rate	60.05587
High density lipoprotein cholesterol levels	60.03063
Waist circumference adjusted for BMI	59.77864

Figure 8*. Subject 3 Top Disease Traits*

One common theme that was seen among all subjects is the presence of disease traits that relate to cardiovascular disease. This trend brought up the question of whether or not the disease traits that were listed in these "Top 10" tables may actually correlate to the manifestation of disease, which allowed me to explore this question.

When individuals submit to the Personal Genome Project, they are also given a variety of surveys that allow these individuals to voluntarily disclose traits such as their zip code, their race/ethnicity, and any medical conditions that they may have at the time of submission. These survey results are displayed as part of an individual's public page on the Personal Genome Project website, which allows all interested individuals and parties to access this information directly from the website. By going through the individual public pages for each of my selected subjects, I was able to learn about the health conditions that they may have at the time of submitting their genome.

On the note of cardiovascular diseases, Subject 1 reported no cardiovascular diseases upon genome submission,[5] Subject 2 reported having chronic chest pain that is not diagnosed to any specific health condition,[6] and Subject 3 reported having Hypertension.[7]

Discussion and Limitations

Addressing the Inherent Assumptions of this Analysis

The results of this analysis revealed that among all 3 subjects, the "Top 10" hits for risk were actual traits that could be related to disease, rather than linking to actual diseases. For instance, although the charts for all 3 subjects contained a hit coding for low-density lipoprotein (LDL) cholesterol, this disease trait by itself, while associated with cardiovascular conditions such as coronary heart disease,[8] does not actually detail the overall risk of such diseases.

It is important to note that the formula used to quantify overall risk (Figure 1) assumes that the presence of certain SNPs is inherently linked to the development of disease. As such, the formula implies that increased risk for a certain disease trait is inherently linked to the number of SNPs that are present in an individual's genome. However, given that these SNPs are a reflection of an individual's disease traits, rather than actual disease risk, this formula is not an accurate metric for determining actual disease risk.

The Role of Age in Interpreting Project Results

One interesting aspect of the results is the secondary analysis conducted to assess if the top disease traits are linked to the actual manifestation of diseases. When looking at the public profiles of each of the selected subjects on the Personal Genome Project database, one major factor that can influence the kinds of diseases that participants reported is the age of the subject at the time that they voluntarily submit their 23andMe data to the Personal Genome Project. Subject 1, who reported no cardiovascular conditions,[5] also reported being between 20 and 29

years of age at the time of submitting their genome data to the project, whereas Subjects 2 and 3, both of whom reported some sort of cardiovascular condition, were at least 40 years of age at the time of their genome submission to the project.[6,7] Given how certain diseases, such as coronary heart disease, are a result of multiple years' worth of environmental exposure, the fact that the first subject reported having no cardiovascular conditions could be as a result of that subject's young age, while the other 2 subjects, who were much older, may have had more time for these conditions to develop and eventually manifest into a noted medical condition.

Other Limitations

One limitation to note is in the way the 3 subjects were selected to be a part of this individual-level analysis. All subjects that were selected identified as White and were based in the United States,[5–7] which fails to account any other individuals of other races and ethnicities into this analysis. Although the semi-random selection of subjects has contributed to analysis of only White individuals, there is also the question of how represented other races and ethnicities are in the pool of individuals in the Personal Genome Project. Essentially, an overrepresentation of White Americans is more likely to result in a 3-person sample consisting of mostly (if not all) White Americans. This limitation does not allow us to look into the effects of race and ethnicity of individual genomes, which in turn, prevents us from assessing how race and ethnicity play a role in genetic health risk and genetic risk for disease. Ways to remedy this include increasing the final sample size used in this analysis as well as advocating for the submission of genome data from individuals of other races and ethnicities.

A second limitation to note is that the analysis fails to account for the effects of the environment on disease risk, which can potentially have a much larger effect on the development of disease compared to the presence of SNPs in an individual's genome. In the context of

cardiovascular diseases and metabolic syndrome, non-genetic influences such as an individual's diet and exercise are directly associated with the development of disease. Although it is still a unclear as to how much of a role the environment plays on the influence of cardiovascular diseases, we still cannot ignore the potentially greater influence that non-genetic factors play in the development of these diseases.[9]

Next Steps

Given the limitation that this analysis conducted on 23andMe data fails to take into account the effects of an individual's environment on disease risk, future research should work to not only assessing the true effect that genetics may play on disease risk given environmental factors, but also quantifying that effect in the form of a numerical covariate that can be applied into statistical models. Specifically, incorporating both data on true genetic risk as well as the influence of environmental factors can allow for the development of generalized linear mixed models to predict the odds for developing a certain disease. Incorporating these covariates into such models would not only allow us to make predictions on a population level, but also on the individual level (which is accomplished through the incorporation of random effects into these models).[10] In order to create the most accurate models, however, it is important to further our understanding of the true influence that the presence of certain SNPs can play and to come up with metrics that quantify that risk in the presence of environmental conditions.

References

1. Research C for DE and. Direct-to-Consumer Tests. *Food Drug Adm*. Published online December 20, 2019. Accessed June 2, 2021. https://www.fda.gov/medical-devices/in-vitro-diagnostics/direct-consumer-tests
2. DNA Genetic Testing & Analysis - 23andMe. Accessed June 2, 2021. https://www.23andme.com/?hphec=tog
3. The Harvard Personal Genome Project (PGP) – enabling participant-driven science. Accessed June 2, 2021. https://pgp.med.harvard.edu/
4. Lecker HK. 23 and Me Raw Data Risk Analysis in R Tutorial. RPubs by RStudio. Published September 4, 2020. https://rpubs.com/cursingwords/Raw-Data-23-and-Me-Analysis-Part-1
5. Personal Genome Project. Personal Genome Project: Public Profile -- huF6CDC1. my.pgp-hms.org. Accessed June 2, 2021. https://my.pgp-hms.org/profile_public?hex=huF6CDC1
6. Personal Genome Project. Personal Genome Project: Public Profile -- huB173FD. my.pgp-hms.org. Accessed June 2, 2021. https://my.pgp-hms.org/profile/huB173FD
7. Personal Genome Project. Personal Genome Project: Public Profile -- hu47A9D1. my.pgp-hms.org. Accessed June 2, 2021. https://my.pgp-hms.org/profile/hu47A9D1
8. LDL Cholesterol as a Strong Predictor of Coronary Heart Disease in Diabetic Individuals With Insulin Resistance and Low LDL | Arteriosclerosis, Thrombosis, and Vascular Biology. Accessed June 6, 2021. https://www.ahajournals.org/doi/full/10.1161/01.ATV.20.3.830
9. Elder SJ, Lichtenstein AH, Pittas AG, et al. Genetic and environmental influences on factors associated with cardiovascular disease and the metabolic syndrome. *J Lipid Res*. 2009;50(9):1917-1926. doi:10.1194/jlr.P900033-JLR200

10. University of California, Los Angeles. Introduction to Generalized Linear Mixed Models. UCLA Institute for Digital Research & Education Statistical Consulting. Accessed June 6, 2021. https://stats.idre.ucla.edu/other/mult-pkg/introduction-to-generalized-linear-mixed-models/

Appendix

Data acquired from the Personal Genome Project (Harvard University):

https://pgp.med.harvard.edu/

Written code for data analysis in R was primarily adapted from the following source on rpubs.com: https://rpubs.com/cursingwords/Raw-Data-23-and-Me-Analysis-Part-1

The following other sources were also consulted for writing R Code:

- https://dabblingwithdata.wordpress.com/2018/07/16/analysing-your-23andme-genetic-data-in-r-part-1-importing-your-genome-into-r/
- https://dabblingwithdata.wordpress.com/2018/09/07/analysing-your-23andme-genetic-data-in-r-part-2-exploring-the-traits-associated-with-your-genome/
- https://vincebuffalo.com/blog/2012/03/12/using-bioconductor-to-analyze-your-23andme-data.html

Example Code Used in Data Analysis

```
#Subject 1 EDA
subject_1_EDA <- subject_1 %>%
  ggplot(aes(chrom)) +
  geom_bar() +
  labs(x = "Chromosome", y = "Number of SNPs") +
  ggtitle("Subject 1 SNP Distribution")

subject_1_joined <- inner_join(subject_1, updated_gwas_data,
                               by = c('rsid' = "SNPS"))

subject_1_joined$risk_allele_clean <-
str_sub(subject_1_joined$STRONGEST.SNP.RISK.ALLELE, -1)
subject_1_joined$my_allele_1 <- str_sub(subject_1_joined$genotype, 1,
1)
subject_1_joined$my_allele_2 <- str_sub(subject_1_joined$genotype, 2,
2)
subject_1_joined$have_risk_allele_count <-
if_else(subject_1_joined$my_allele_1 ==

subject_1_joined$risk_allele_clean, 1, 0) +
  if_else(subject_1_joined$my_allele_2 ==
subject_1_joined$risk_allele_clean, 1, 0)

subject_1_risk_data <- dplyr::select(subject_1_joined,
                                     rsid,
                                     have_risk_allele_count,
                                     DISEASE.TRAIT,
                                     risk_allele = risk_allele_clean,
                                     your_genotype = genotype,
                                     RISK.ALLELE.FREQUENCY,
                                     MAPPED_TRAIT,
                                     REPORTED.GENE.S.)
subject_1_risk_data

#Get Count of Disease Traits by DISEASE.TRAIT
subject_1_trait_count_disease <- subject_1_risk_data %>%
  group_by(DISEASE.TRAIT) %>%
```

```
  summarise(risk_count = n())

#Count of Disease Traits for Risk Allele = 2
subject_1_trait_count_disease2 <- subject_1_risk_data %>%
  group_by(DISEASE.TRAIT, have_risk_allele_count) %>%
  filter(have_risk_allele_count == 2) %>%
  summarise(risk_count_2 = n()) %>%
  subset(select = -c(have_risk_allele_count))

#Mutating Risk Counts for Allele = 1 and Allele = 2
subject_1_trait_count_disease3 <-
  merge(subject_1_trait_count_disease,
subject_1_trait_count_disease2,
        by = c("DISEASE.TRAIT", "DISEASE.TRAIT"), all = TRUE)
subject_1_trait_count_disease3[is.na(subject_1_trait_count_disease3)]
<- 0

#Risk Count for Allele 1 equation
risk_d_1 <- (subject_1_trait_count_disease3$risk_count -
subject_1_trait_count_disease3$risk_count_2)

#Adding risk 1 column in
subject_1_trait_count_disease4 <-
dplyr::select(subject_1_trait_count_disease3,
                                        DISEASE.TRAIT,
                                        risk_count,
                                        risk_count_2) %>%
  mutate(risk_d_1)

#Adding Overall Risk Column and Equation
overall_risk_d <- (subject_1_trait_count_disease4$risk_count_2 /
((subject_1_trait_count_disease4$risk_d_1 * 2) +
subject_1_trait_count_disease4$risk_count_2)) +
  (subject_1_trait_count_disease4$risk_d_1 /
((subject_1_trait_count_disease4$risk_d_1 * 2) +
subject_1_trait_count_disease4$risk_count_2))

subject_1_overall_risk_count_d <-
dplyr::select(subject_1_trait_count_disease4,
```

```
                                DISEASE.TRAIT,
                                risk_count,
                                risk_1 = "risk_d_1",
                                risk_count_2) %>%
  mutate(overall_risk_d*100)

overall_risk_count_d[is.na(overall_risk_count_d)] <- 0

#Table of Disease Traits with over 100 risk alleles for DISEASE.TRAIT
subject_1_table1 <- subject_1_overall_risk_count_d %>%
  filter(risk_count > 99) %>%
  arrange(desc(`overall_risk_d * 100`)) %>%
  subset(select = -c(risk_count, risk_1, risk_count_2))

names(subject_1_table1)[1] <- 'Disease Trait'
names(subject_1_table1)[2] <- 'Overall Risk'

subject_1_table1 <- subject_1_table2 %>%
  kable() %>%
  kable_paper("hover", full_width = FALSE)

subject_1_table1
```