

Assessing Disease Risk from 23andMe Data

Jay Mantuhac

EPIDEM 275: Bioinformatics

Project Introduction

- Direct-to-consumer (DTC) testing allows for individuals to access insights from their own genetic data without the need of a healthcare provider
 - 23andMe + Personal Genome Project (PGP) by Harvard



The Harvard
Personal Genome Project
we are open for science

Primary Question of Interest

Given the 23andMe Data from 3 individuals, what kind of diseases are individuals most at risk for?

Project Methods

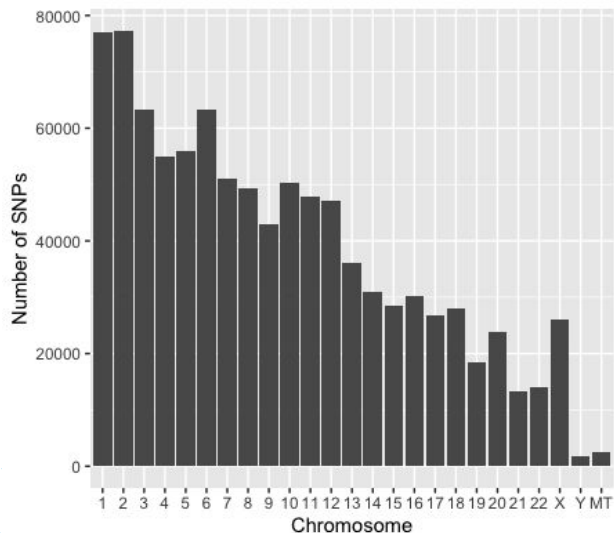
- Extraction of 3 23andMe individual-level datasets from the PGP website
 - Dataset structure: rsid, chromosome number, position, genotype
- Exploratory Data Analysis - Distribution of SNPs in Each Chromosome
- Data Wrangling (lots of it!) w/ GWAS Catalog
- Risk Calculation + Analysis

How Is Disease Risk Calculated?

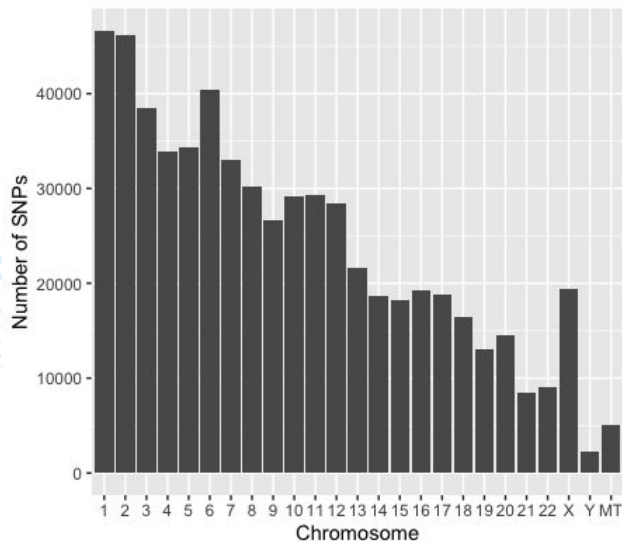
$$\text{overall risk} = \frac{\text{risk count}_2}{(\text{risk count}_1 \times 2) + \text{risk count}_2} + \frac{\text{risk count}_1}{(\text{risk count}_1 \times 2) + \text{risk count}_2}$$

Results: SNP Distribution

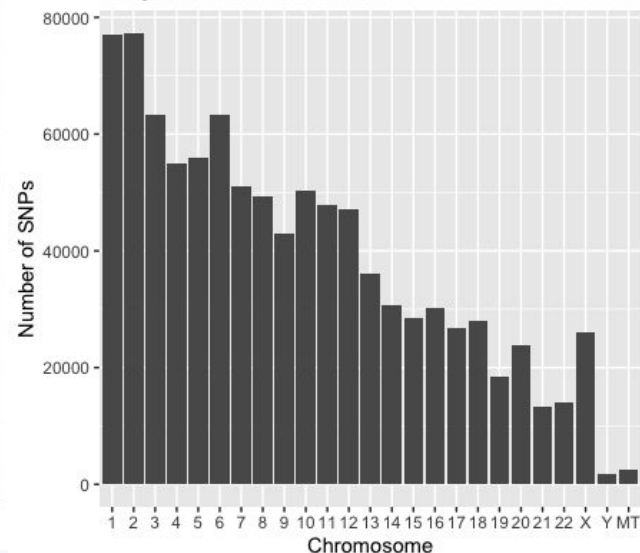
Subject 1 SNP Distribution



Subject 2 SNP Distribution



Subject 3 SNP Distribution



Results: Disease Risk

Disease Trait	Overall Risk	Disease Trait	Overall Risk	Disease Trait	Overall Risk
Total cholesterol levels	63.73626	Hip circumference adjusted for BMI	65.21739	C-reactive protein levels	63.46154
Post bronchodilator FEV1/FVC ratio	62.50000	High density lipoprotein cholesterol levels	62.97071	HDL cholesterol levels	63.19613
Pulse pressure	62.35294	HDL cholesterol levels	62.86645	Heel bone mineral density	62.66667
C-reactive protein levels	61.68224	Pulse pressure	61.65644	Blood metabolite levels	62.17228
High density lipoprotein cholesterol levels	61.25000	Waist circumference adjusted for BMI in active individuals	61.44578	Blood protein levels	61.94030
Low density lipoprotein cholesterol levels	61.00324	Triglycerides	61.17424	Total cholesterol levels	60.70959
Post bronchodilator FEV1	60.86957	Heel bone mineral density	60.92896	IgG glycosylation	60.35156
LDL cholesterol levels	60.81081	LDL cholesterol levels	60.77586	LDL cholesterol levels	60.20067
Heel bone mineral density	60.58932	IgG glycosylation	60.57971	Estimated glomerular filtration rate	60.05587
IgG glycosylation	60.35156	Waist circumference adjusted for body mass index	60.13072	High density lipoprotein cholesterol levels	60.03063
				Waist circumference adjusted for BMI	59.77866
Subject 1		Subject 2		Subject 3	

Does these results correlate with reported diseases?

- Common theme of heart related conditions among all subjects
 - Subject 1 had no reported cardiovascular conditions
 - Subject 2 reported undiagnosed chronic chest pain
 - Subject 3 reported Hypertension

Discussion + Limitations

- All 3 subjects were White and from the US
 - Underrepresentation of other races/ethnicities in the PGP database?
- Increased genetic risk does **not** guarantee a person will develop a certain disease
 - What role does environment play in disease risk?

Next Steps

- Potential other questions/avenues to explore with 23andMe data:
 - Applying statistical methods to assess the role of genetics in comparison to other factors that affect disease risk
- In all honesty, my main purpose of this project is to gain more practice/experience in using R and Bioinformatics packages in R :)

The background is a solid blue color. Overlaid on this are several wavy, horizontal lines composed of small, dark blue dots. These lines create a sense of motion and depth, flowing from the left side towards the right. The dots are arranged in a way that they form a series of overlapping, undulating bands.

Thank You!

What questions do you have?