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

Primary Question of Interest

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

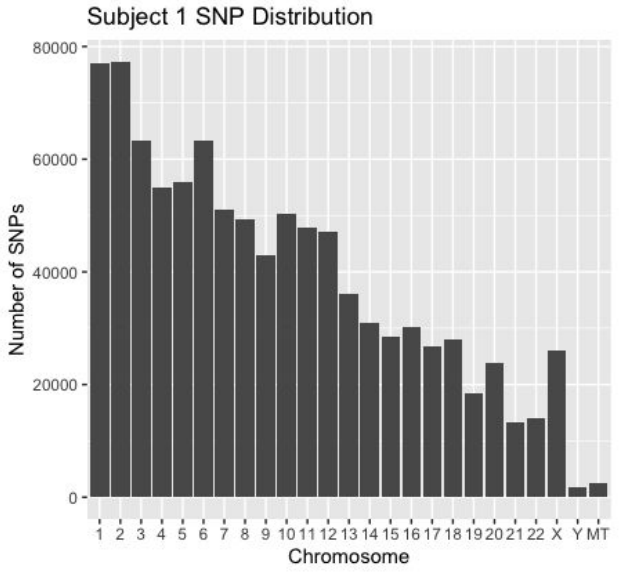
Project Methods

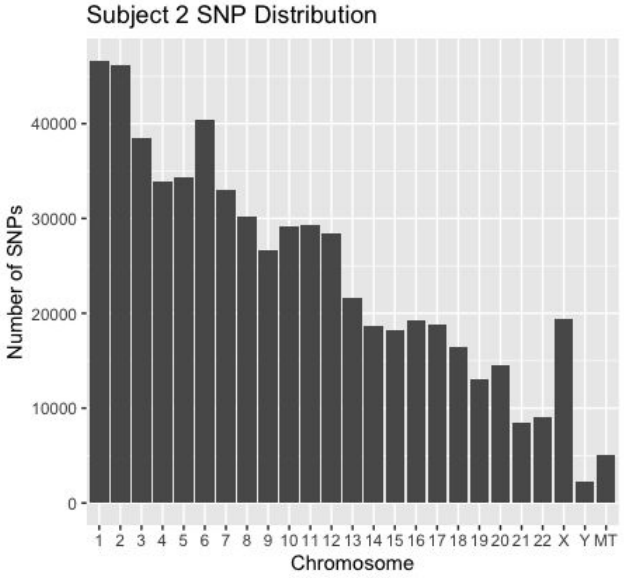
- Extraction of 3 23andMe individual-level datasets from the PGP website
  - Dataset structure: rsid, chromosome number, position, genotype
- Exploratory Data Analysis - Distribution of SNPs in Each Chromosome
- Data Wrangling (lots of it!) w/ GWAS Catalog
- Risk Calculation + Analysis

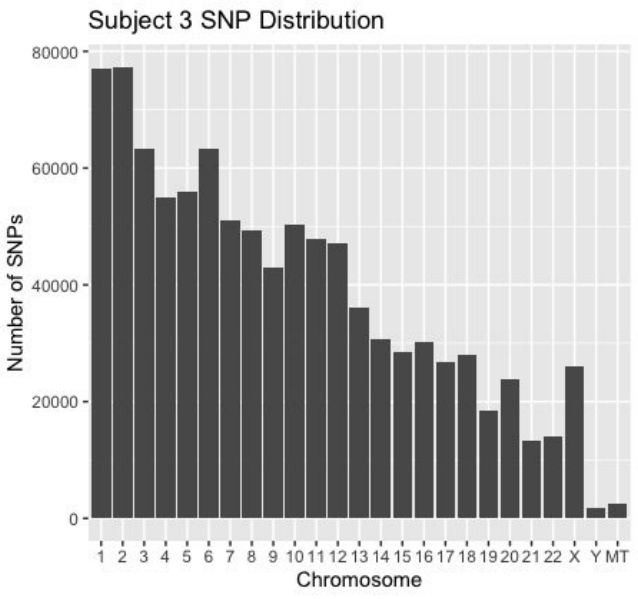
How Is Disease Risk Calculated?

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

Results: SNP Distribution







Results: Disease Risk

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

Subject 1

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

Subject 2

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	59.7786

Subject 3

Does these results correlate with reported diseases?

- Common theme of risk of cardiovascular 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 :)

Thank You!

What questions do you have?