Inference of Ancestry, Sex and Trait Liabilities From Whole Genome Single Nucleotide Polymorphisms

Lorcan Purcell

September 1, 2025

1 Introduction

As the human genome is 99.6% identical between individuals, comparing genomes in their entirety would be both a waste of computational resources and effort (1). Instead, the unique mutations possessed by each individual are used for comparison, such mutations include: repeat regions, inversions and deletions However, these classes of variation are not assayed by the approaches used in this study. Rather single nucleotide polymorphisms (SNPs) are used, which are single nucleotide changes that are found within at least 1% of the global population (distinct from rare mutations that may only be carried by a single individual) (2). SNPs most common form of genetic variance between people and most often occur in the region of DNA between genes (2). While a significant majority of SNPs have benign effects, there are others that can help explain an individual's susceptibility to a particular disease or condition (3). One major application of SNP based analysis is within genealogical companies such as Ancestry.com. As part of their analysis process, Ancestry claims to genotype 730,525 SNPs, ones they claim "account for majority of common genetic variation in European and other worldwide populations", which are then compared to population reference panels in order to infer an individuals genetic makeup (4). Through this investigation, a similar, albeit slightly reductive, process will be carried out to determine the ancestry as well as sex and predisposition to certain traits for an individual. While this investigation does not fall under the realm of hypothesis testing, it remains an important proof-of-principle project with significant didactic value.

To go into greater depth on the nature of this investigation, we will use SNP data gathered from Ancestry.com on a specific individual to discover specific attributes about that individual. Combining this data with reference samples from HapMap3 we will infer: genetic ancestry, biological sex, relatedness and their liability to certain traits.

2 Investigation

2.1 Gathering SNP Data

The first step was gather SNP data. This was done through downloading data from the Illumina OmniExpress platform with the genotyping having been done by Ancestry.com based on a saliva sample from the individual. The data provided 677,437 variants across 23 autosomes, sex chromosomes and mitochondrial data.

Chr	# variants		Chr	# variants		Chr	# variants		Chr	# variants
1	50554		8	32960		15	21344		22	10985
2	54847		9	29616		16	23350		X	25242
3	43310		10	32763		17	22891		Y	1665
4	36843		11	32527		18	18986		Y (pseudoautosomal region)	36
5	38796		12	31283		19	16969		Mitochondrial genome	263
6	43301		13	24919		20	18072			
7	34621		14	21126		21	10167			

Figure 1: Table illustrating gathered SNPs from Illumina OmniExpress platform. Chr indicates chromosome number or type, variants indicates the number of variants found within the chromosome on the same row

The data downloaded also contained RefSeq IDs and the genomic coordinates of each SNP relative to the Genome Reference Consortium Human Build 37.1.

2.2 Gathering Reference Panel Genotype Data

The reference genotype data was then obtained from the HapMap3 project, which was an early genotyping project attempting to map the patterns of individual differences in the 0.6% of the unique genome between individuals. The 11 populations sampled by the HapMap3 are listed here:

Label	N	Population
ASW	90	African ancestry in Southwest USA
CEU	180	Utah residents with Northern and Western European ancestry from the CEPH collection
CHB	90	Han Chinese in Beijing, China
CHD	100	Chinese in Metropolitan Denver, Colorado
GIH	100	Gujarati Indians in Houston, Texas
JPT	91	Japanese in Tokyo, Japan
LWK	100	Luhya in Webuye, Kenya
MEX	90	Mexican ancestry in Los Angeles, California
MKK	180	Maasai in Kinyawa, Kenya
TSI	100	Toscans in Italy
YRI	180	Yoruba in Ibadan, Nigeria

Figure 2: Population data from HapMap3. Only the data from the CEU and TSI populations was used. Links to the data can found in references

The HapMap data contained 1,490,422 genotyped variants which were primarily SNPs.

2.3 Merging Datasets

Next, PLINK (v1.9) was used to generate binary files for both the HapMap dataset SNPs and the individual SNP data downloaded from ancestry.com (5). These files were then merged using the –make-bed and –merge commands. Some SNPs were dropped due to having inconsistent alleles between the two datasets. After the merging, there were 451,565 identified SNPs present in both the Ancestry dataset and the HapMap dataset.

2.4 Ancestry Estimation

In the merged dataset, identity-by-state values (sharing 0, 1 or 2 alleles per SNP), was calculated for every pair of individuals using PLINK's –genome command. These values were stored in an N by N matrix (for N individuals), which captures genetic distance between individuals from the datasets. This matrix was reduced to 4 dimensions by applying principal components analysis (PCA) through PLINK's –mds command. PCA is a technique that transform data into a new coordinate system where the axis capture the most significant variation in the dataset. The first two components with the full HapMap dataset, as they capture the broadest variations, were visualized below.

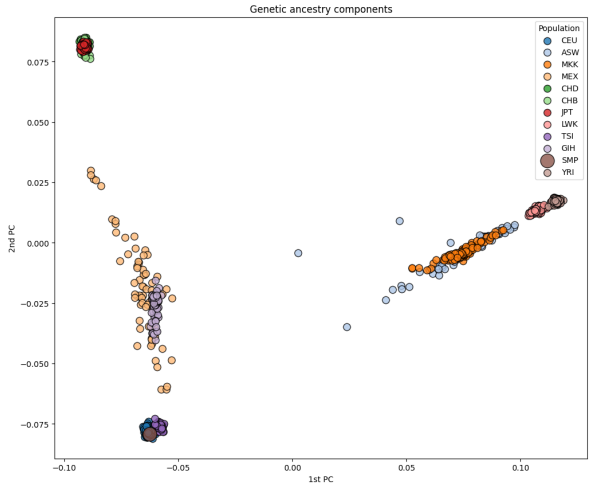

Figure 3: Principle component analysis with full HapMap dataset and individual. Individual is labeled as SMP, in brown

It is clearly visible that the individual is most closely related to the TSI and CEU populations, which are in blue and purple respectively. This makes sense, as they were the populations from Utah and Italy and the individual being studied is likewise of European descent. However, even if the ancestry of the individual had not been previously known, this result would showcase the most likely possibility. The next figure illustrates PCA of the individual with the TSI and CEU populations only.

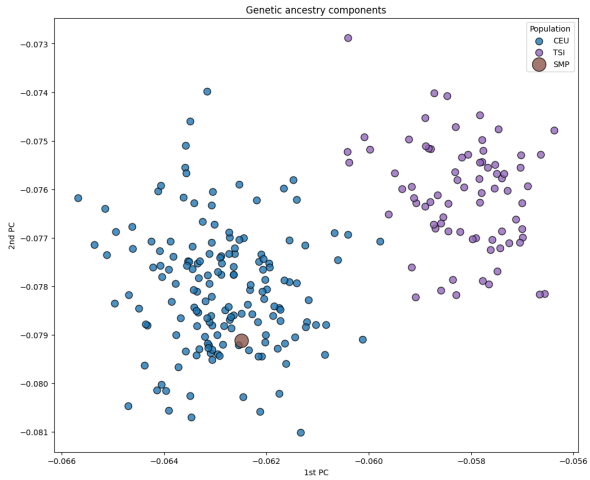

Figure 4: Principle component analysis with TSI and CEU populations from HapMap dataset and individual. Individual is labeled as SMP, in brown

This result was also expected as the individual being studied is of English descent, which places them as more closely related to the Utah residents with ancestry stemming from Western Europe than the Italians. This result also shows how precisely ancestry can be triangulated. The HapMap dataset is relatively small and old, yet in this case provides great accuracy.

2.5 Sex Inference

The next part of the investigation was to determine the individual's sex. When dealing with SNP data, this is done through analyzing patterns of homozygosity and heterozygosity on the X chromosome. In practice, software like PLINK examines the genotype at each X-linked SNP. If there is only one x chromosome, then the value would be homozygous, e.g. A/A. However, if there are two x chromosomes, then the SNP could be homozygous or heterozygous. PLINK then calculates based on the proportion

of heterozygous SNPs on the X chromosome to infer biological sex. The resulting value is an F statistic, which indicates the ratio of variability between groups. A value near 0 suggests female, while a value near 1 suggests male. Using PLINK's –check-sex command, the result for the individual was calculated to be 0.999800, which heavily suggests the individual to be male. The F scores for the European populations are visualized in Figure 6.

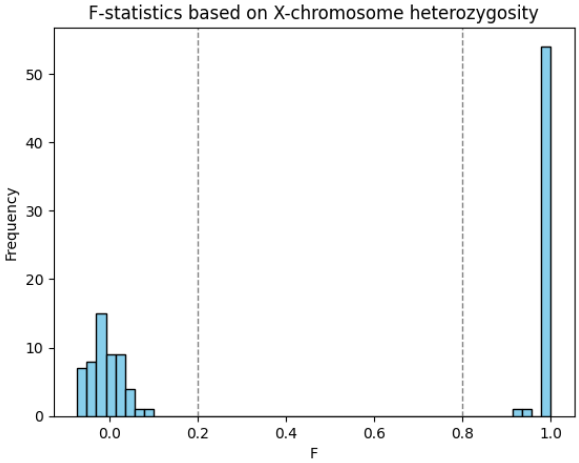

Figure 5: Distribution of calculated F statistics for 110 Europeans from HapMap dataset.

The distribution reveals an approximately even split between males and females within the dataset. The reason for the larger cluster around 0.0 is due to females still possessing the chance to have homogeneous x-linked SNPs. In contrast, Males will almost always have hemizygous (which software counts as homozygous) SNPs, the exception being in the pseudoautosomal regions, which are shared between the X and Y chromosomes and have the chance of showing heterozygosity. These regions are small, however, and thus barely impact the data.

2.6 Polygenic Risk Score Estimation

Polygenic Risk Scores (PRS) are in essence a weighted sum of risk allele counts across many locations on the genome and estimate the genetic risk for a known trait or outcome. PRS can be derived from genome-wide association studies, which for the more ambiguous traits, such as levels of psychopathy, are often self-reported and thus have potential inaccuracies. Effectively, each allele is weighted by the extent to which is predicts the outcome in the original study. People with more "risk-increasing" alleles are therefore assumed to be at higher risk.

Whereas the previous analyses could be performed on relatively sparse genomic coverage (e.g. the 500K variants used in the ancestry estimations discussed above), PRS estimation requires a fuller coverage so the SNPs from the individual can match with the published PRS databases. A process called imputation is used, where missing SNPs are statistically filled in, given a large reference panel in which all variants have been genotyped. The reason for the missing SNPs is that ancestry.com only uses a fraction of total SNPs to perform their analysis, so when data is downloaded from them, it is incomplete. This step was performed by uploading the genotype data for the European HapMap samples plus the individual to the Michigan Imputation Server (MIS). The data had to first be reformatted to Variant Call Format (VCF) using PLINK as that is the file format required by MIS. After imputing the missing SNPs, MIS also generates PRS estimates for all individuals based on over 4000 previously published studies.

To visualize the scores for the index case, the mean and standard deviation from the HapMap3 European samples was used to normalize the scores for the index. As such, absolute scores greater than say 3 reflect unusual scores, assuming a normal distribution. Figure 6 shows the distribution of PRS values for the individual.

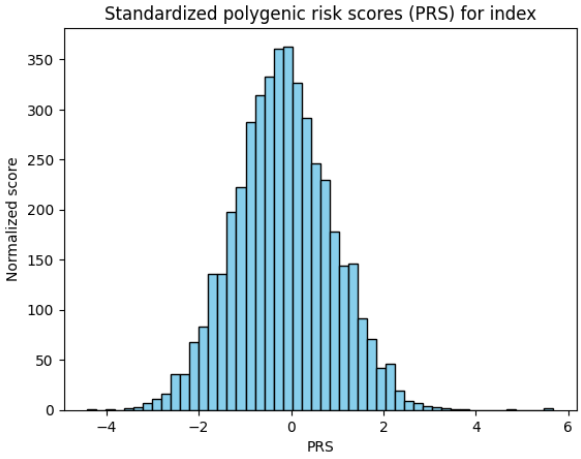

Figure 6: Distribution of PRS values for the index. Y axis represents number of PRS. X axis represents the z score the individual has for each PRS

The distribution indicates that the individual has a relatively balanced genetic risk profile. Furthermore, most PRS fall near zero, which indicates no significant propensities towards certain outcomes or traits. With that said, the individual still had some unusual scores as seen in the table below.

	score	Z	coverage	variants_used	variants_total	variants_flipped	variants_ambiguous	trait	trait_efo
711	PGS000714	3.145271	0.600000	33	55	0	0	Prostate cancer	EFO_0001663
772	PGS000777	5.491575	1.000000	3	3	0	0	Parkinson's disease dementia	EFO_0007710
854	PGS000854	-3.087658	0.740741	20	27	0	0	Type 2 diabetes (based on SNPs associated with...	MONDO_0005148
1037	PGS001072	3.746534	0.717949	28	39	0	0	Facial aging looking 'older than you are'	EFO_0008006
1217	PGS001300	3.282516	0.444444	4	9	0	0	Celiac disease or gluten sensitivity diagnosed	EFO_0001060
1346	PGS001430	3.08708	0.917038	1647	1796	0	0	Mean FA in posterior limb of internal capsule ...	EFO_0006930
1841	PGS001923	-3.01105	0.999031	50534	50583	0	0	Time spent watching television (TV) or using c...	EFO_0010724
2312	PGS002395	-3.05822	0.646439	2823	4367	0	0	Eczema	HP_0000964
2361	PGS002444	-3.829093	0.518986	8720	16802	0	0	Eczema	HP_0000964
2556	PGS002640	-3.145535	0.719111	173687	241530	0	0	Eczema	HP_0000964
2686	PGS002767	-3.0279	0.998676	1050882	1052275	0	0	Knee osteoarthritis	EFO_0004616
2793	PGS002878	3.327859	0.996768	1110194	1113794	0	0	C-reactive protein	EFO_0004458
2808	PGS002888	3.610076	0.996548	1110906	1114754	0	0	C-reactive protein	EFO_0004458
2836	PGS002921	5.672684	0.193878	19	98	0	0	Estradiol	EFO_0004697
2837	PGS002922	4.847576	0.806510	8297111	10287670	0	0	Estradiol	EFO_0004697
3042	PGS003128	-3.421052	0.996722	1112765	1116425	0	0	Thinness	EFO_0004340
3050	PGS003133	-3.075531	0.996478	1113533	1117469	0	0	Thinness	EFO_0004340
3337	PGS003418	3.112321	0.857143	6	7	0	0	Prostate cancer	MONDO_0008315
3543	PGS003627	-4.417088	0.948206	3039	3205	0	0	Right CA3 body volume	EFO_0009395
4229	PGS004378	-3.241257	0.797475	845275	1059939	0	0	Body mass index (BMI)	EFO_0004340
4246	PGS004389	-3.046116	0.797475	845275	1059939	0	0	Leg fat percentage (left)	EFO_0007800
4267	PGS004409	-3.225447	0.797475	845275	1059939	0	0	Body mass index (BMI)	EFO_0004340
4374	PGS004522	-3.498045	0.797475	845275	1059939	0	0	G47 (Sleep disorders)	EFO_0008568
4404	PGS004551	-3.384122	0.797475	845275	1059939	0	0	M20 (Acquired deformities of fingers and toes)	MONDO_0017427

Figure 7: PRS scores with absolute value greater than 3

The z score column indicates the number of standard deviations the individual is from the mean in regards to the number of SNPs they have that correlate to a specific outcome. The coverage column shows what fraction of the total varients associated with a certain outcomes overlap with the variants possessed by the individual. The flipped column is for cases where strand orientation had to be flipped in order to match the reference panel. This can happen when reference data is on the opposite strand than the uploaded genotype data. The ambiguous column has to do with palindromic SNPs, though they were removed by PLINK during the merging done earlier in the process. Looking at Figure 7, many PRS can be seen that look potentially problematic. However, there are many limitations with PRS, such as that they are not transferable between populations. This is due to the slightly different genetic makeup that can be found within different populations, and therefore the SNPs associated with particular outcomes can also look slightly different, causing the PRS to be invalid. Another critique is that they are not truly indicative of an outcome, instead only hinting at potential predisposition. In this way, they lack clinical utility and cannot be used to diagnose patients. Criticism aside, the individual did test highly for basal cell carcinoma, having a z score of 2.92 with 99.7% coverage. This aligns with reality, as the individual did in fact have basal cell carcinoma a few years ago, indicating that PCR tests do have some applicability as a preventative measure.

3 Conclusion

This investigation demonstrated how SNP data can be used to infer a wide range of relevant characteristics about a person. Through analysis with the HapMap3 dataset as a reference, it was possible to precisely triangulate an individuals ancestry despite having a limited sample size of only around $\tilde{1}000$. Sex was also deduced based on SNP heterozygosity within the X chromosome. Lastly, polygenic risk scores were calculated and revealed several elevated risk levels, one of which aligns with previous medical history of the individual. Though limitations still exist, especially regarding PRS clinical utility, this investigation illustrated the potential of genomic data in revealing identity, heritage, and predisposition.

4 References

N.B. All graphs and tables were made using Matplotlib and Python
Big shout out to

- (1) Human Genomic Variation - Genome.gov
- (2) What are SNPs? - MedlinePlus
- (3) The use of SNPs in pharmacogenomics - NCBI
- (4) AncestryDNA Ethnicity Estimate White Paper
- **HapMap3 Dataset:** Genotype Data (PED Format)
- **SNP Positions:** MAP File
- (5) PLINK: Whole genome data analysis toolset
- (6) Imputation and Reference Panels - NCBI
- (7) Michigan Imputation Server