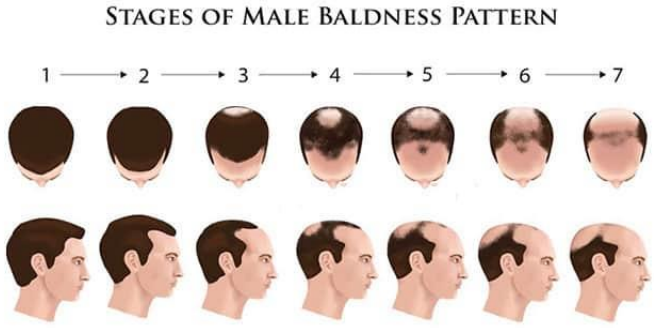
Kyungjin Cho
BMI 826 – Statistics in Human Genetics

Written Report of Final Project

Genetic Analysis of Male Pattern Baldness

I. Project Background

Male Pattern Baldness (MPB) is an age-related condition that characterizes hair thinning and loss. It is a common trait that 80% of European male are affected by this condition. The balding process is highly patterned as it can be shown in the image below:



<Image 1.1 – Stages of Male Baldness Pattern>

Before performing the analysis, I was curious about the environmental features of what makes a person bald. I came up with the ideas such as medical conditions, stress level, age, hairstyle, expose to radiation, and sleep pattern.

II. Project Introduction

As mentioned above several times, this project is about a male pattern baldness which is a study from UK Biobank. There are sample of 205,327 European males who are aged from 40 to 73 years old. Subjects were given the survey and were asked to respond to it conveying their hair pattern with MPB score of 1 to 4. Please refer to the image below:

<Image 2.1 – Survey taken from subjects of MPB study>

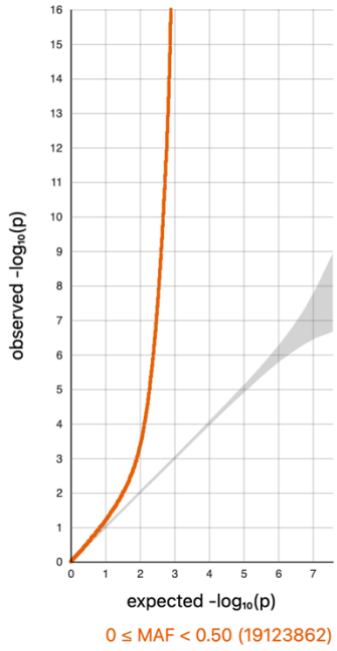
III. Methods used for this Project

Before getting deeper into this project, this section explains the methods used. There are several methods implemented for several sections. For the Manhattan, QQ plots, and LocusZoom plot, I have used the LocusZoom website (http://locuszoom.sph.umich.edu/). For getting the intercept, heritability, global correlation and h2 enrichment plot, I have implemented the LDSC Score Regression tool, and finally, for the local genetic correlation, I have used the SUPERGNOVA.

IV. Results

a) Q-Q Plot

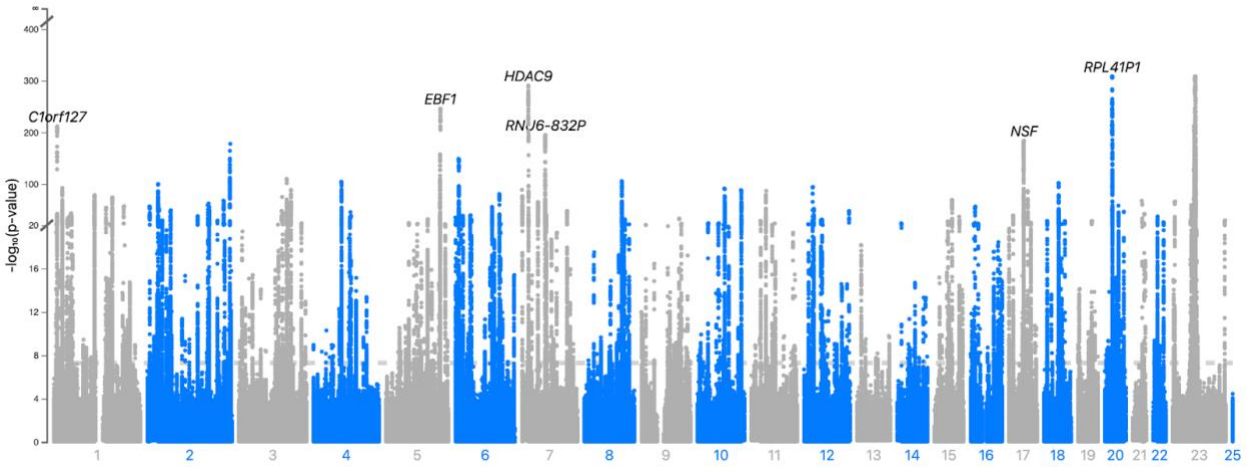
The QQ Plot of male pattern baldness summary statistics is shown below:



<Figure 5.1 – QQ plot of the MPB>

As shown on the above, the QQ plot of MPB is not diagonal with the existence of so-called "tail". Furthermore, the line is skewed, rather, exponential. This QQ plot does not represent the standard QQ plot that have to be observed and thus the p-value is not uniformly distributed. Therefore, further analysis strategy will be needed.

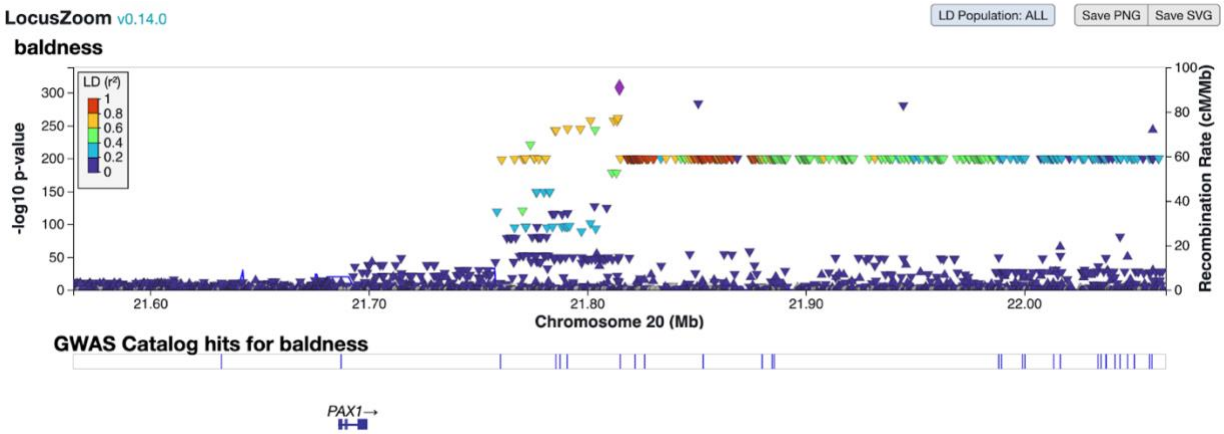
b) Manhattan Plot



<Figure 5.2 – Manhattan Plot of MPB>

The Manhattan plot is the plot that displays large number of data points, especially represents significant SNPs. In this plot, there are lot of significant points on the graph. However, three most significant points among them would be from chromosome 20, chromosome 23, and chromosome 7. From the most significant chromosome, which is chromosome 20, the nearest gene would be RPL41P1.

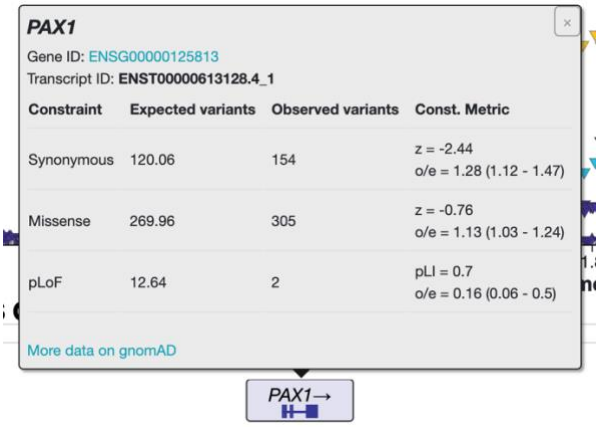
c) LocusZoom Plot



<Figure 5.3 – LocusZoom Plot for Chromosome 20, Marker Name 20: 21,814,739>

The LocusZoom plot for SNP 20:21,814,739 in Chromosome 20 is shown above. The GWAS catalog hits for baldness is shown below the image. The points with color represent how significant those genes are. Looking at the GWAS catalog, the most relevant gene was PAX1 and the description is shown below:

Kyungjin Cho
BMI 826 – Statistics in Human Genetics



<Figure 5.4 – The most significant gene: PAX

<Image 5.1 - Otofaciocervial Syndrome>

I have used the website called National Library of Medicine to search for the SNP that contains this gene, rs6047658. Under the related disease, phenotype, and trait section, I was able to find the related disease that is called Otofaciocervical syndrome. This syndrome is
A disease trait about the developmental defect characterized by a distinct facial feature and I thought that the reason why this gene is related to the baldness is because both relates to the facial part of the human body.

d) LDSC Intercept and Heritability

The LDSC intercept and heritability results are shown below:

- LDSC results on Intercept:

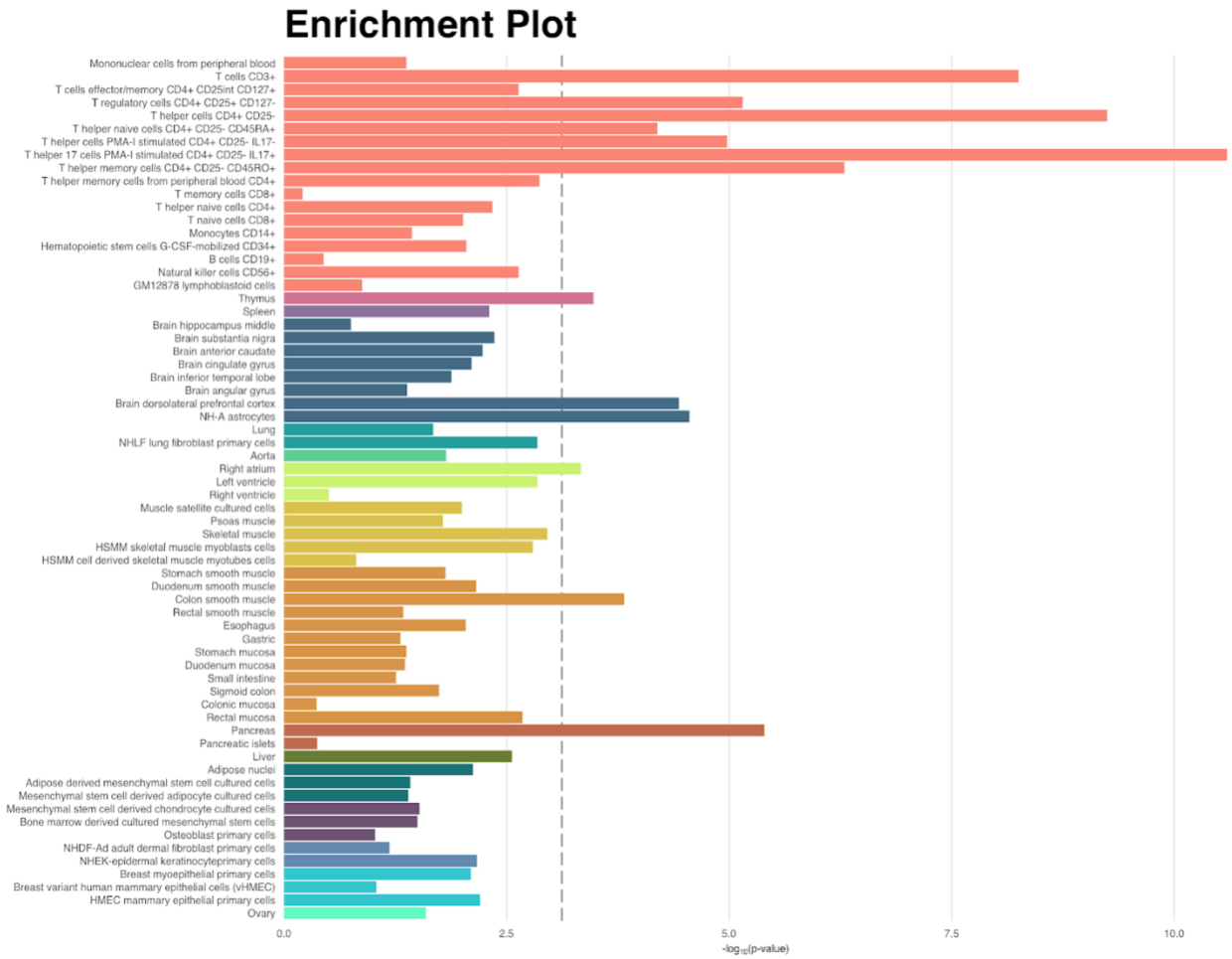
	Point Estimate	Standard Error	P-value
Value	1.227	0.020	< 0.05

- LDSC results on Heritability:

	Point Estimate	Standard Error	P-value
Value	0.324	0.027	< 0.001

As shown above, they are highly significant with p-value less than the Bonferroni Correction of 0.05. The results on heritability shows that about 32% of the male baldness is attributable to the genetic individual difference.

e) Heritability Enrichment Plot



<Figure 5.5 – Heritability Enrichment Plot of MPB>

For the heritability enrichment plot of MPB, there were lot of significant tissue cells that are related to it. Among them, the most significant one was about the T-helper 17 Cell. According to the research I did, this cell plays role in getting the autoimmune disease with all the other T-helper cells. Surprisingly, autoimmune diseases are associated with the hair loss as their side effects, and I was able to figure out why this tissue cells are so closely related to the pattern of baldness.

The second most important tissue cell is Pancreatic cell, the brown-colored bar on the plot. This pancreas cells play a role in producing and transporting chemicals that exists the body through digestive system. As similar to the autoimmune disease, pancreatic related diseases, cancers, or disorders causes hair loss as a side effect and this could be the reason why this cell is so closely related to the baldness.

f) Introduction to 5 other GWAS

GWAS	Reasons of Interest
• Cannabis Use Disorder	Misuse of drugs can cause a lot of side effects to our body
• Smoking Initiation	Smoking is always considered unhealthy
• Daytime Napping	Napping might relieve the stress and prevent the hair loss
• Loneliness	Loneliness might lead to mental health problem and may be correlated with hair loss
• Household Income	Higher income people are likely to consume better quality of foods such as non-GMO or organic foods

The five other GWAS I picked are listed above in the table.

g) Global Genetic Correlation

The result of global genetic correlation implemented by LDSC is listed below:

	SMOKE	CANNABIS	LONELINESS	NAP	INCOME
BALDNESS	+0.030	-0.006	-0.017	+0.041	-0.004

According to the LDSC correlation results, all five GWAS were not significantly correlated with baldness. However, it was reasonable because of several reasons. First, all five GWAS did not take gender into the account. Male Pattern Baldness is specifically focused on one gender, but all the other GWAS did not specifically gave information about the gender. Also, as it was mentioned on the research paper of MPB, the authors mention that the baldness is closely related to gender, bone mineral density and pancreatic cell function, which I also got the results from enrichment heritability plot that those are significant. Yet, none of the GWAS I picked are not that closely related to those three categories.

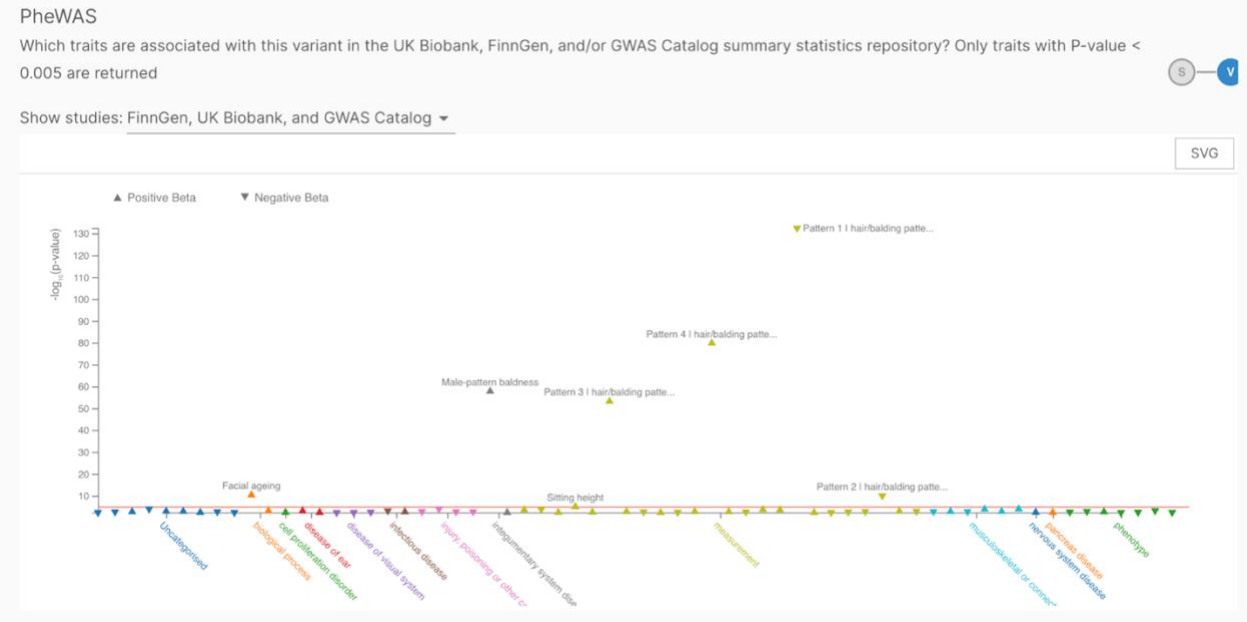
h) Local Genetic Correlation

For the local genetic correlation, I have implemented the SUPERGNOVA. The result file example is shown below:

chr	start	end	rho	corr	h2_1	h2_2	var	p	m
1	1103150	2320702	1.7893588090311074e-05	4.155906995077646	8.177887801636582e-06	2.266846347923122e-06	1.473029244518748e-09	0.6410573810542222	288
1	2321099	3065568	-1.2071137680794549e-05	-0.46423537975402535	5.209673023466997e-05	1.2978050801389823e-05	7.118312515806925e-10	0.6509530235763833	184
1	3066761	3679461	-1.1867509067197825e-05	NA	9.881776891085647e-06	-7.932987120714685e-06	1.0309550750400134e-09	0.7116756428633237	211
1	3679775	4749076	2.0550574818088745e-05	NA	2.7143021914871337e-05	-3.6475508492623204e-05	2.170389996422865e-09	0.6591273489332017	547
1	4750013	5000086	7.342302354336776e-06	NA	2.0532725619778442e-05	-1.5452640979039876e-05	6.195880088738611e-10	0.7680152849841652	156

<Figure 5.6 – Results.txt File of the Correlation between Baldness and Smoke Initiation>

With all the five results file, each five GWAS summary statistics compared with baldness summary statistics, I have figured out that the chromosome 20 is highly correlated with baldness and five other GWAS. To further analyze, I have compared this chromosome to the Manhattan plot I have created above; Manhattan Plot also shows that the chromosome 20 is highly related to the male pattern baldness.



<Figure 5.7 – PreWAS results plot>

The tool PreWAS pre-processes the GWAS studies, and here is the result of PreWAS for SNP rs6047658 in chromosome 20. The Figure 5.7 shows that this SNP rs6047658 is highly correlated with hair/balding pattern, as five points out of seven on the plot shows the correlation of this SNP with hair/balding pattern, highest correlated GWAS being the same. Furthermore, the p-value is less than 0.00001 which means that it is clearly significant study.

V. Conclusion

The male pattern baldness is closely related to genetics. The analysis of GWAS correlation shows that it is highly gender specific, and age related. From the heritability enrichment plot and Manhattan plot one can realize how baldness is specifically related to the genetics as they significantly convey lot of cells and genes related to the male pattern of baldness. One thing that

can be improved from the study is if the five GWAS specifically focused on male as gender, the correlation between 5 other GWAS would be more credible. The main chromosome to focus on this research is chromosome 20 with the most significant SNP being rs6047658. As there are still no medical treatment to treat baldness, the biostatistical analysis on baldness would be one of the most important step for the improvement of treatment for hair loss.

VI. References:

MPB Research Paper: https://www.nature.com/articles/s41467-018-07862-y
LDSC score regression: https://github.com/bulik/ldsc/
SUPERGNOVA: https://github.com/qlu-lab/SUPERGNOVA
SNP search: https://www.ncbi.nlm.nih.gov/snp/
PreWAS: https://cran.r-project.org/web/packages/prewas/index.html