Investigation on the GWAS of Variants Controlling Eye Colour Phenotype

Helen King

Abstract

This lab report demonstrates that SNPs in regions of genes that encode for melanin metabolism related proteins contribute towards eye colour phenotype variation in two European populations. It also shows the statistical significance of removing covariates from population sets to ensure data sets suitable for downstream processing within the GWAS pipeline.

Introduction

In this lab, our main aim is to determine the genomic loci that govern eye colour phenotype and then testing the accuracy of these phenotype-associated SNPs on a sample set. We performed a GWAS using plink tools for samples with blue, brown and other eyes. This was visualised using Manhattan plots and QQ plots. The list of putative SNPs is compared with previously found eye colour related SNPs and then used for trait prediction and comparison between CEU and TSI population groups.

GWAS, genome wide assisted studies, are a tool into gaining more accurate understanding of complex, continuous traits. There is still a lot of improvement needed to ascertain all variants that contribute to disease risk. Only then can it be used as information for individual disease risk profiles. [1] However, there is a big problem with the individuals within most GWAS that they are majority of European descent. With this phenotype, blue eye colour is mainly seen in European background so, is not a problem; in others, it means a massive amount of genomic variance is being ignored.

Methods

Dataset description

- Describe our GWAS cohort. How many samples were collected? How many have blue vs. brown eyes? How many SNPs were genotyped?

The original file is a SNP genotype array that has been cleaned to remove both duplicates and other erroneous SNPs. SNP arrays are the best format to analyse the polymorphisms between different individual genomes within a population with high throughput. There is three different file types that displayed the array, .ped, .map and .fam. The .ped file contains all the information on the genotype arranged into a separate sample ID for each line. The .map file contains the SNP information with each line being a different rsID for a different SNP.

There was 261 samples in the cohort found by doing wc on the .ped file. To find the frequency of brown and blue eyes in our samples, we cat the .phen file then pipe it into cut (with the flag –f 3) then sort, then finally the uniq command (-c flag). This outputs the number of blue as 104 and brown as 157. Overall, the number of SNPs, if you wc the .map file is 911774.

- Describe the cohort used for eye color prediction. How many samples were included? How many SNPs did we look at?

Here we used a .vf.gz file which contains a smaller number of samples independent from the orginal GWAS. This has to be manipulated using a series of commands to extract the sample data and removing the header and other redundant information. This is done by using the functions, bcftools query, datamash transpose and awk to get a header for the file. Then the next set of commands bcftools query with the flag –f and sed to assign the ID and genotypes for each SNP. The final set of commands are cat then datamash transpose to out put the final desired table. There are 206 samples in total in the file, with 6 SNPs being looked at.

GWAS

- Describe how you used plink (and which version) to perform the GWAS. How did you control for population structure?

We used plink function with the flags –-file, --pca 10 and then the –-out flag used to specify the output file name. This forms 10 PCAs or principle component analysis of the data. PCAs cluster together samples based on similarity.

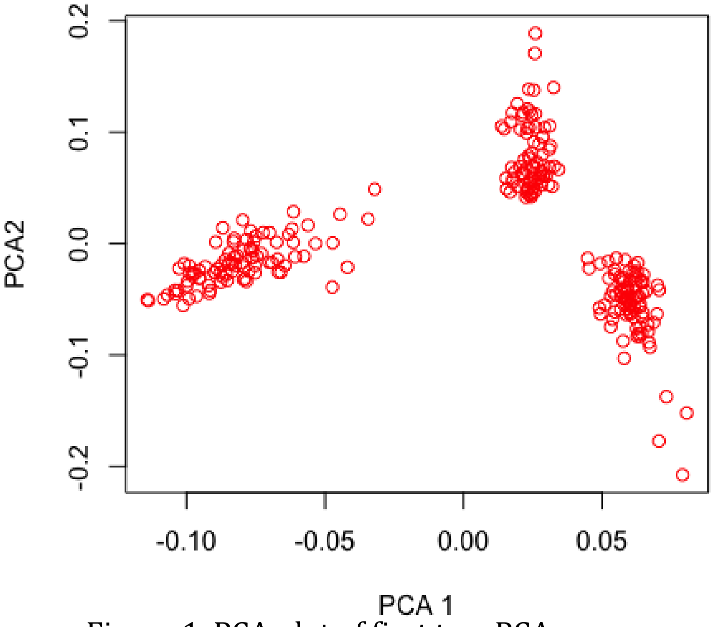


Figure 1. PCA plot of first two PCAs

The sample used to make the SNP genotype array was collected from different ethnic backgrounds, to validate the GWAS with these heterogeneous results, further data manipulation is needed. To look at the population structure of the GWAS, we looked at the PCA plot. The three clusters suggest three different ancestry backgrounds. We ran plink twice. The first time to perform the GWAS using 'logistic regression'. Using the plink command with the flags –-file (the prefix used for the three plink files and the path to it), --pheno (the path to and the .phen file) , --out (specify an outprefix), --logistic and –allow-no-sex. However, this doesn't take into account the different covariates. So we ran it a second time with the flag –covar with the .eigenvec file, then use cat and awk commands in order to extract one test for each SNP for further downstream analysis. The population structure was controlled by removing these covariates which would create bias in the dataset due to ethnic background.

- How did you "clump" variants into independent signals? Describe the parameters you used for plink.

Due to the QQ plot of the GWAS without the covariates being more linear, we use that data set for all the downstream processing. We also use plink again but this time use the –-clump flag on the .assooc.logistic.no_covars file, --clump-field P. The –clump-p1 flag is used for the threshold p-value for the variants. The –clump-p2 flag is the number for the clumped variant p-value threshold. The –clump-r2 flag is the r^2 threshold. The clump-kb flag is used is the kb radius. Don't forget to specify the prefix for the GWAS files and the output prefix. The output will be a table with $OUTPREFIX.clumped file name.

Eye color prediction

- What method did you use to compute eye color (e.g. script? excel spreadsheet?)

For eye colour prediction, I used a Google sheets spread sheet with the one given. It uses the probability equations below found in the lab report, as a model for the probability of getting blue, brown and other eyes for each sample. α_i is the intercept term for the model , i. $\beta_{i,k}$ is the effect size for the model (at the kth SNP).

$$p_{blue} = \frac{e^{\alpha_1 + \sum_k \beta_{1,k} X_k}}{1 + e^{\alpha_1 + \sum_k \beta_{1,k} X_k} + e^{\alpha_2 + \sum_k \beta_{2,k} X_k}}$$

$$p_{other} = \frac{e^{\alpha_2 + \sum_k \beta_{2,k} X_k}}{1 + e^{\alpha_1 + \sum_k \beta_{1,k} X_k} + e^{\alpha_2 + \sum_k \beta_{2,k} X_k}}$$

$$p_{brown} = 1 - p_{blue} - p_{other}$$

For each sample, taking into account each SNPs weighting produces a probability of eye colour for each sample. The highest probability could accredit the most likely eye colour out of blue, brown and other.

- Describe any additional methods or datasets you used to interpret what the variants might be doing.

We were then given a list of samples CEU and TSI populations. CEU is Utah Residents (CEPH) with Northern and Western European Ancestry and TSI means Toscani in Italia. Some of these don't list the known samples given in the previous list so couldn't match them to the original list. However, with the samples that overlapped between both the original list and the given list, we averaged the probabilities of all the eye colours and then created figure 8.

We used the IGV viewer to look at the regions the SNPs reside in within the hg 19 human genome.

Results

GWAS

- How many variants pass genome-wide significance in the GWAS with and without controlling for population structure? After clumping variants, how many independent signals remain?

32 SNPs passed the genome wide significance test with covariates. 15 passed the genome wide significance test in the no covariates. The removal of the covariates due to the genomic variance due to ethnic background means the 15 SNPs are more accurately indicative of causing change in eye colour. After clumping only around four independent signals remain. (see below in Figure 6).

- Include Manhattan plots and QQ plots for each GWAS carried out and compare the results with and without controlling for population structure.

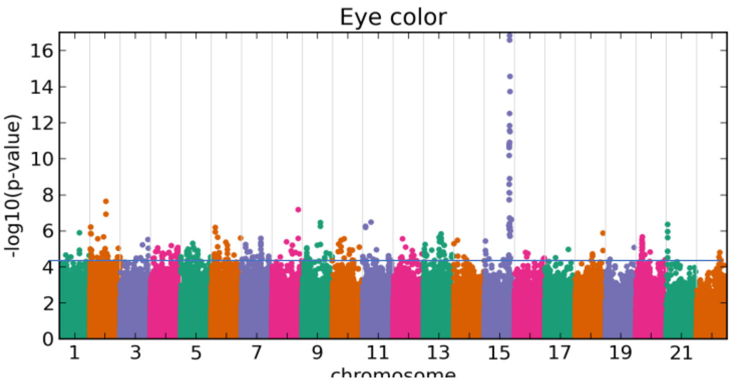


Figure 2. Manhattan Plot based on GWAS with covariates not included.

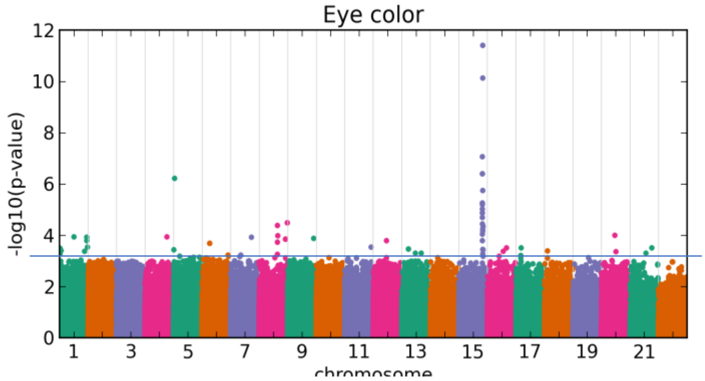


Figure 3. Manhattan Plot based on GWAS with covariates included.

Looking at the genome wide significance line of the GWAS without covariates included in figure 2 shows a reading of around of –log10 (4) compared to a –log10 (3) with covariates. This decrease indicates a larger p-value and the SNPs have lower statistical significance and weaker associations when covariates are included. In both set of data, the largest peak is on chromosome 15 as that is the area with largest number of SNPs that have strongest association with eye colour.

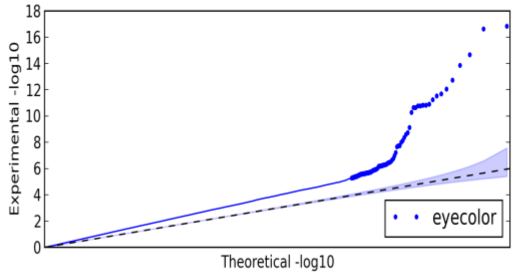


Figure 4. QQ plot based on the GWAS with covariates not included.

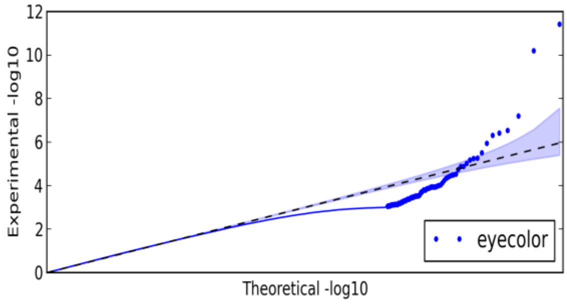
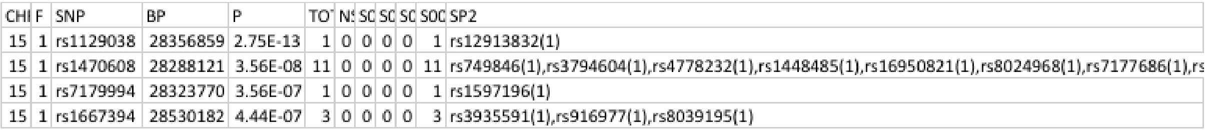


Figure 5. QQ plot based on the GWAS with covariates included.

With the comparative Q-Q plots, it is a comparison between the observed values on the y axis compared to the expected values on the x axis. When covariates are included in the dataset, there is a dip below the line which suggests a not true association between the observed and the expected. Whereas when covariates are excluded the points have "inflation", indicating a true association. Both four figures confirm that covariates should be removed for further downstream analysis.

- Include a table of genome-wide significant results after clumping.

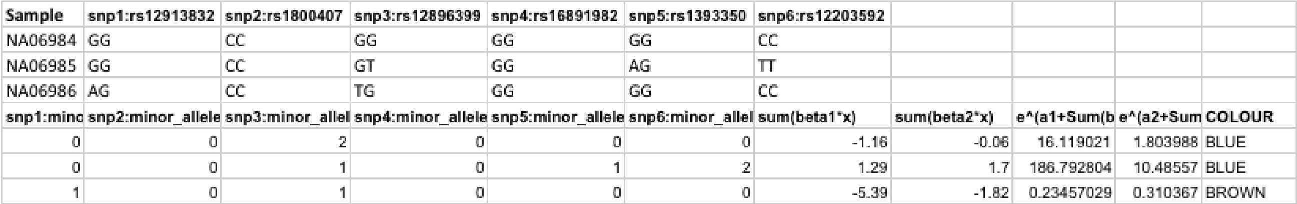


CHI	F	SNP	BP	P	TO	N:	SC	SC	SC	SOC	SP2
15	1	rs1129038	28356859	2.75E-13	1	0	0	0	0	1	rs12913832(1)
15	1	rs1470608	28288121	3.56E-08	11	0	0	0	0	11	rs749846(1),rs3794604(1),rs4778232(1),rs1448485(1),rs16950821(1),rs8024968(1),rs7177686(1),rs
15	1	rs7179994	28323770	3.56E-07	1	0	0	0	0	1	rs1597196(1)
15	1	rs1667394	28530182	4.44E-07	3	0	0	0	0	3	rs3935591(1),rs916977(1),rs8039195(1)

Figure 6. Table on the genome wide significant results after clumping

Eye colour prediction

- Include and reference a supplementary table of results with eye color predictions for each sample



Sample	snp1:rs12913832	snp2:rs1800407	snp3:rs12896399	snp4:rs16891982	snp5:rs1393350	snp6:rs12203592			
NA06984	GG	CC	GG	GG	GG	CC			
NA06985	GG	CC	GT	GG	AG	TT			
NA06986	AG	CC	TG	GG	GG	CC			
snp1:minc	snp2:minor_allele	snp3:minor_allel	snp4:minor_allele	snp5:minor_allele	snp6:minor_allel	sum(beta1*x)	sum(beta2*x)	e^(a1+Sum(b	e^(a2+Sum	COLOUR
0	0	2	0	0	0	-1.16	-0.06	16.119021	1.803988	BLUE
0	0	1	0	1	2	1.29	1.7	186.792804	10.48557	BLUE
1	0	1	0	0	0	-5.39	-1.82	0.23457029	0.310367	BROWN

Figure 7. Example of the table (formatted to fit on page) that displays the eye colour predictions within the samples given in the second sample set.

https://docs.google.com/spreadsheets/d/1rySg2f16fC1RMNXZ6u-GH9sg1b66oqIH-2XXhzpJat0/edit?usp=sharing

	Probability of this Eye Colour occurring within Population set			
	Blue	Other	Brown	Total
CEU	0.6313029304	0.1300114549	0.2386856147	1
TSI	0.2501047608	0.1666767793	0.5832184599	1

Figure 8. Table showing the probability of the eye colour occurring with northern Europeans (CEU) and southern Europeans (TSI)

- Include at least one genome browser screenshot of one of the six eye colour SNPs and your interpretation of how that SNP might affect eye colour.

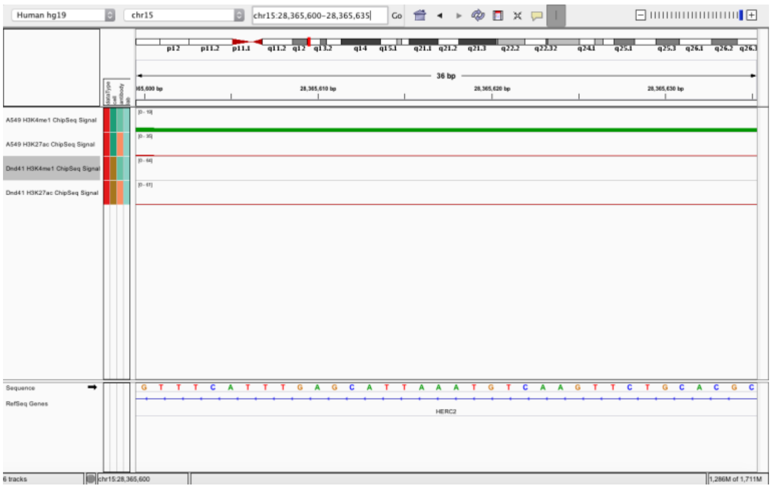
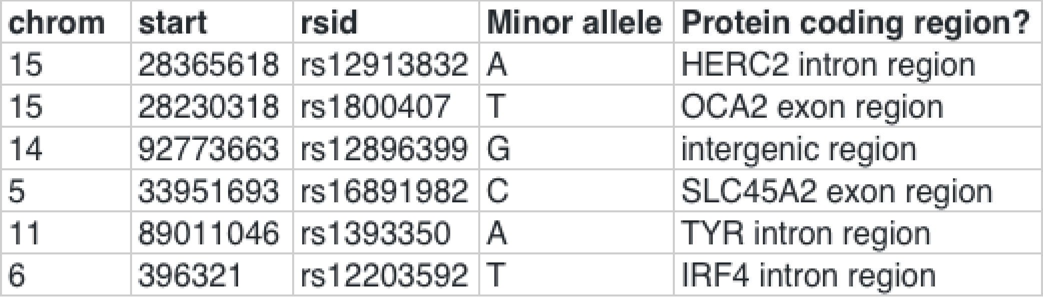


Figure 9. Screen shot of IGV with loaded A549 and Dnd H3K4me1 and H23K27ac tracks loaded, along with the h19 human genome at the rsid – rs12913832.



chrom	start	rsid	Minor allele	Protein coding region?
15	28365618	rs12913832	A	HERC2 intron region
15	28230318	rs1800407	T	OCA2 exon region
14	92773663	rs12896399	G	intergenic region
5	33951693	rs16891982	C	SLC45A2 exon region
11	89011046	rs1393350	A	TYR intron region
6	396321	rs12203592	T	IRF4 intron region

Figure 10. table of the SNPs and the region they reside in within the h19 human genome.

Considering the regions (seen in Figure 10) of what the SNPs do by looking at their function NCBI Gene database. I followed the lab report suggestion of looking at the histone methylation and acetylation marks (H3K4me1 and H3K27ac). However, none of the available tracks used a cell line that would be in the iris (where the colour of the eye is) so, as the histone marks aren't tissue specific they hold little or no weight.

Even though the SNP rs12913832 is seen in the intron, it probably causes a missense mutation producing a different protein product. HERC2 is a gene that codes for an unusually large protein product that has a HECT and RLD domain containing ubiquitin protein ligase. OCA2 gene codes for a melanosomal transmembrane protein which is believed to be an 'integral membrane protein involved in small molecule transport, specifically tyrosine, which is a precursor to melanin synthesis.' [2] SLC45A2 encodes a transporter protein that mediates melanin synthesis. TYR encodes for the enzyme that catalyses one of the steps that converts tyrosine to melanin. IRF4 codes for a transcription factor important in the body's response to a viral infection

Discussion

- What affect did controlling for population structure have on your results?

Using plink to remove the covariates reduced the number of genome wide significant variants found to reduce from 32 to 15. This was because the ethnic background of an individual created a skew seen in both the QQ plots and the Manhattan plots (Figures 2-5). The removal of the covariate ensured that only SNPs with the strongest association with eye colour were found. This means that the downstream results were more accurate.

- Did you identify all the known SNPs contributing to eye colour? If not, why do you think you might have missed them?

Clumping them into to group correlated SNPs, only one of these four variant groups (Figure 6), overlap with the known 6 SNPs associated with eye colour (Figure 10). This SNP is rs12913832. We might have missed them because the genome-wide significance P value is common place. However, it would remove lower allele frequency variants. [3] Looking at the eye colour phenotype it is difficult to qualify if the genetic architechture underpinning this phenotype is caused by small number of rare variants of large effect or numerous small, common variants

Both ethnic backgrounds, CEU and TSI are of European background. We might have missed SNPs of consequence because we haven't looked at the full spectrum of different ancestries. This is seen in most GWAS as, even though there has been an increase from 4% to 20% of GWAS participants from non-European descent, it is not representative of the world population. [4] In eye colour however, most non-European descendants have got brown eyes.

- Hypothesize how the 6 SNPs used in Part II might be affecting eye colour.

Looking at all the genes found in Figure 10 and finding their gene and their corresponding protein product that might be effected was insightful. Rs12896399 is found in an intergenic region, so be an enhancer or repressor region for genes involved in melanin synthesis in the eye but it is not even on a chromosome where other top SNPs that indicate eye phenotype. The variants rs1800407 (OCA2), rs16891982 (SLC45A2), rs1393350 (TYR) and rs12203592 (IRF4) all reside in genes whose protein products directly affect melanin production which, attributes brown eyes. The SNP rs12913832 seems to lie in a gene which encodes for nothing to do with melanin synthesis, a ubiquitin protein ligase. After research, I found an article [5] by Eiberg that this SNP downregulates the OCA2 promoter and therefore reduces expression of the OCA2 gene and causing blue eye colour.

Citations

[1] Genome-wide association studies for complex traits: consensus, uncertainty and challenges.- McCarthy MI1, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Nat Rev Genet. 2008 May;9(5):356-69. doi: 10.1038/nrg2344

[2] The NCBI website – Gene tool https://www.ncbi.nlm.nih.gov/gene/4948

[3]The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. Fadista J1,2, Manning AK3,4, Florez JC3,4,5,6, Groop L2,7. Eur J Hum Genet. 2016 Aug;24(8):1202-5. doi: 10.1038/ejhg.2015.269. Epub 2016 Jan 6.

[4] Genomics is failing on diversity. Alice B. Popejoy& Stephanie M. Fullerton. Nature Comment

[5] Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. Human Genetics, 2008, Volume 123, Number 2, Page 177. Hans Eiberg, Jesper Troelsen, Mette Nielsen

Extra Credit

See the extra credit on the sheet in the attached google document with the sheet labelled extra credit.

Looking at the four possible combination of mother and father gametes, I then narrowed down the number of combinations due to being homozygous or heterozygous for the G and C alleles of each of the SNPs. Therefore, there is 3x2x2x1x1x1=12 possible ways to arrange the different genotypes of the child's SNPs. In the end, the child is most likely to have blue eyes due to the following probability statistics. blue- 0.540056947 other-0.1506096358 brown-0.3093334172

https://docs.google.com/spreadsheets/d/1rySg2f16fC1RMNXZ6u-GH9sg1b66oqIH-2XXhzpJat0/edit?usp=sharing