Genome-wide association study of eye color

Cæcilia Lind Skov-Jensen

Eye color is a highly heritable and visible trait, but there is still a lot of the genetics associated with it that is still not fully understood. Here a genome-wide association study for eye color was performed, involving 1,287 individuals with self-reported phenotypes from a direct-to-consumer genetic testing. 27 associations were identified on chromosome 15, all variants located in the genes OCA2 and HERC2, which are known to be associated with eye color. Three of the 27 found associated variants could be replicated from a previous GWAS from Simcoe et al. (2021). Overall, the study outcomes demonstrate that the biggest association is located on chromosome 15 on the two neighboring genes, OCA2 and HERC2.

INTRODUCTION

Eye color is one of the most visible variations between humans. The human eye color is highly heritable with an estimated heritability of 0.80 [1], however its genetic architecture is not yet fully understood [2]. The color of the eyes results from the amount and quality of the pigment melanin of the iris. People with brown eyes have a large amount of melanin in the iris, while people with blue eyes have much less of this pigment [3]. Previous genome-wide association studies (GWASs) have identified various genetic loci, which are significantly associated with eye color [2, 4]. The latest being Simcoe et al. (2021) [2], that identified 124 genetic loci associated with eye color. The neighboring genes OCA2 and HERC2 on chromosome 15 have been found to have the strongest genetic influence on eye color [2]. OCA2 produces a protein, known as the P protein, which is involved in the maturation of melanosomes, which are cellular structures that produce and store melanin. The P protein therefore plays a crucial role in the amount and quality of melanin that is present in the iris. A region of the nearby HERC2 gene known as intron 86 contains a segment of DNA that controls the expression of the OCA2 gene, turning it on or off as needed [3]. To better understand the genetics of human eye color, an eye color GWAS was carried out using GWAS data from 1,287 individuals obtained from openSNP [5], which is a web site where users of direct-to-customer genetic tests can upload their personal data. The phenotype data is self-reported eye colors and was processed as described below.

RESULTS/DISCUSSION

Data preprocessing

First, samples with a heterozygosity rate more than 3 standard deviations from the mean were removed, because low heterozygosity could be due to imbreeding, and high heterozygosity could be due to contamination of the sample, and both of these will bias the result. After this filtering, 13 individuals were removed from the 1,287 individuals, leaving 960,613 variants and 1,274 individuals for further filtering and quality control. Next, related individuals were removed, since

these would increase the genotype frequency regardless of what phenotype they have. This will result in a bias in the allele frequency distribution and eventually bias the case-control study. After the filtering, 14 individuals were removed from the data, leaving 960,613 variants and 1,260 individuals for further processing. Then variants were filtered using a call rate of 25%, meaning that variants with more than 75% missing data were removed. This removed 8,167 variants. Additionally, variants for which the minor allele frequency (MAF) is less than 1% were removed, as a very small MAF, so that the large majority of individuals have two copies of the major allele, results in inadequate power to infer a statistically significant relationship between the SNP and the trait [6]. This filtering removed 89,770 variants. At last, variants for which the Hardy-Weinberg equilibrium (HWE) test statistic has a corresponding p-value of less than $1 \cdot 10^{-5}$, as violations of HWE can be an indication of the presence of population substructure or the occurrence of a genotyping error. This filtering removed 24,523 variants. This results in 838,153 variants and 1,260 individuals to be considered in the association analysis.

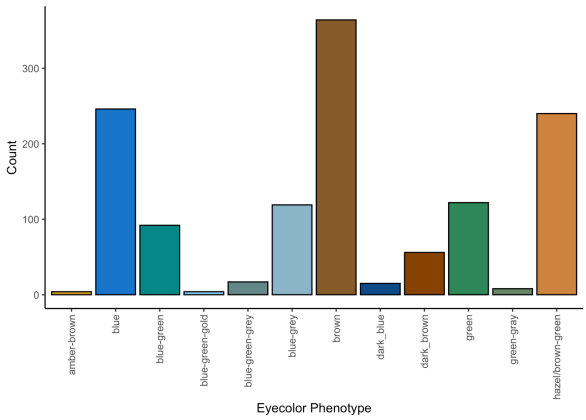

Fig. 1: Visualization of self-reported phenotype per individual.

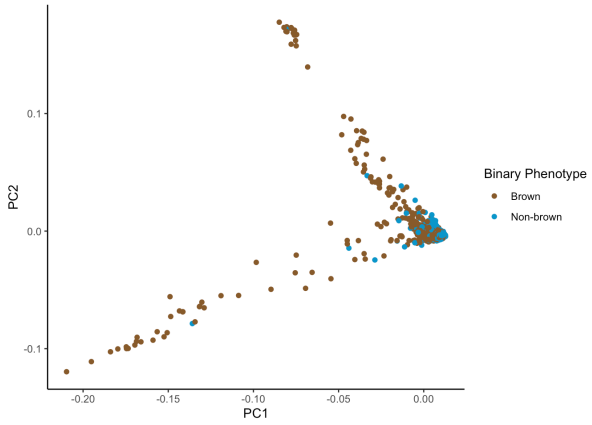

Fig. 2: PCA score plot of the first two PCs.

Association analysis

After completion of SNP and sample-level filtering, the self-reported phenotypes from **Fig. 1** were rewritten as binary phenotypes to make the analysis easier. The phenotypes that contain the word 'brown' are annotated as '1', while the other non-brown phenotypes are annotated as '2'. In total there are 646 individuals with brown phenotype and 614 individuals with non-brown phenotype.
Hereafter, the data was LD-pruned with a threshold of 0.2, and a Principal Component Analysis (PCA) was conducted where the first 20 PCs were calculated. The plot in **Fig. 2** shows that the first and the second principal component did not give a clear distinction between the two groups. This could be explained by the fact that the phenotypes were self-reported, and an individual with e.g. brown eyes could have described their eye color as green, whereby this individual would not be included in the brown phenotype. In order to make the distinction clearer, the phenotypes could have been annotated as blue, brown and others, because if an individuals reports their eye color as either blue or brown, it is quite certain that that is true.

	CHR	SNP	BP	A1	F_A	F_U	A2	P	OR
	<int>	<chr>	<int>	<chr>	<dbl>	<dbl>	<chr>	<dbl>	<dbl>
749376	15	rs1129038	28356859	C	0.06281	0.5829	T	7.533e-184	0.04795
749379	15	rs12913832	28365618	A	0.06093	0.5801	G	2.153e-172	0.04697
749399	15	rs1667394	28530182	C	0.04570	0.3961	T	4.168e-100	0.07302
749396	15	rs916977	28513364	T	0.04520	0.3836	C	1.434e-92	0.07608
749397	15	rs8039195	28516084	C	0.04461	0.3511	T	2.175e-72	0.08630
749372	15	rs4778241	28338713	A	0.07980	0.3529	C	1.582e-65	0.15900
749385	15	rs11636232	28386626	T	0.48990	0.1928	C	6.431e-56	4.02200
749383	15	rs3935591	28374012	T	0.04062	0.3769	C	4.673e-55	0.06999
749371	15	rs4778138	28335820	G	0.06426	0.2852	A	1.055e-41	0.17210
749373	15	rs7495174	28344238	G	0.02362	0.1703	A	1.052e-38	0.11790

Table 1: The ten most significant SNPs after Fisher's exact test.

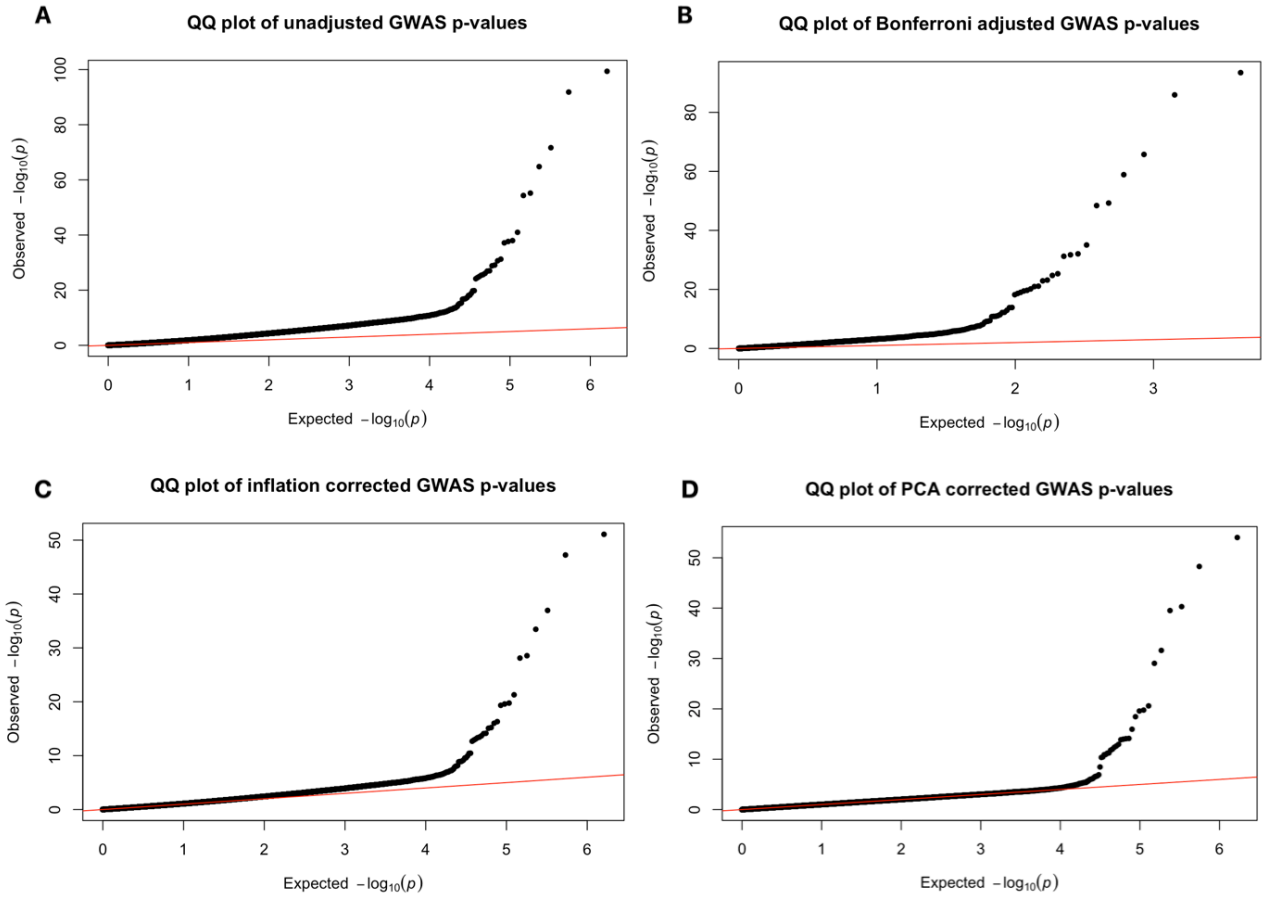

Fig. 3: QQ plot of **(A)** unadjusted, **(B)** Bonferroni adjusted, **(C)** inflation corrected, and **(D)** PCA corrected p-values.

The association between SNPs and eye color status was tested using an allelic Fisher's exact test. After the Fisher's exact test, 749 SNPs were significant ($p < 5 \cdot 10^{-8}$). The ten most significant variants were all located on chromosome 15 (**Table 1**). A quantile-quantile (QQ) plot was then made to test whether the distribution was normally distributed (**Fig. 3A**). The QQ plot indicated that the distribution of the p-values did not perfectly match with the pattern under the null hypothesis that the p-values were normally distributed. Some observed p-values were clearly more significant than expected under the null hypothesis as seen by the points moving towards the y-axis. This deviation suggested a sizable inflation. Afterwards, the p-values were adjusted using a Bonferroni correction in order to adjust for multiple comparisons. After Bonferroni correction, there are only 42 significant loci, however the ten most significant loci are still the same. The QQ plot still showed a general inflation of the test statistic (**Fig. 3B**). This was to be expected, since we had a lot of individuals coming from different populations. In order to correct for this inflation, both genomic control and adjustment for PCs were performed.

Genomic control was done by calculating the inflation factor (λ), which is calculated as the median of the chi-squared statistics computed divided by the median of the chi-squared distribution under the null. The inflation factor should be 1 as the observed chi-squared values would then be equal to the expected values under the null. However, the inflation factor calculated here was 1.9672. A correction of the observed chi-squared values was then made by dividing the chi-squared values by the inflation factor and thereafter correct the p-values. Again, a QQ plot was made of the inflation corrected p-values (**Fig. 3C**), which gave a better result, but still indicated inflation. Therefore a PCA correction was performed. In order to know how many PCs to adjust for, the amount of variance explained by each PC was calculated, and a cumulative sum plot was made (**Fig. 4**). From the plot it was clear that the two first PCs explain over 50% of the variance, and the p-values were adjusted for these two.

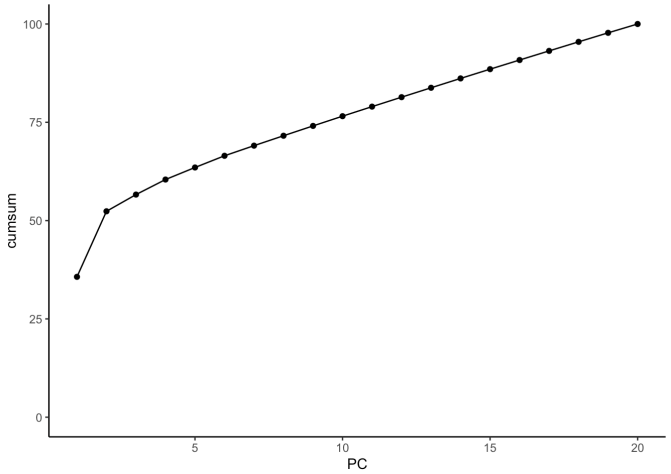

Fig. 4: Cumulative sum plot of the variance explained by each PC.

The first two PCs were adjusted for using a logistic regression test. The resulting QQ plot (**Fig. 3D**) showed that the distribution of the observed p-values performed a lot more like the expected distribution, although it still was a bit inflated. The resulting inflation factor was now 1.0238, which was a lot closer to 1. After the PCA correction a Manhattan plot was made to locate significant variants. Using a threshold of $5 \cdot 10^{-8}$, 27 SNPs were significant and all were located on chromosome 15 (**Fig. 5**). The most significant SNP was rs1667394.

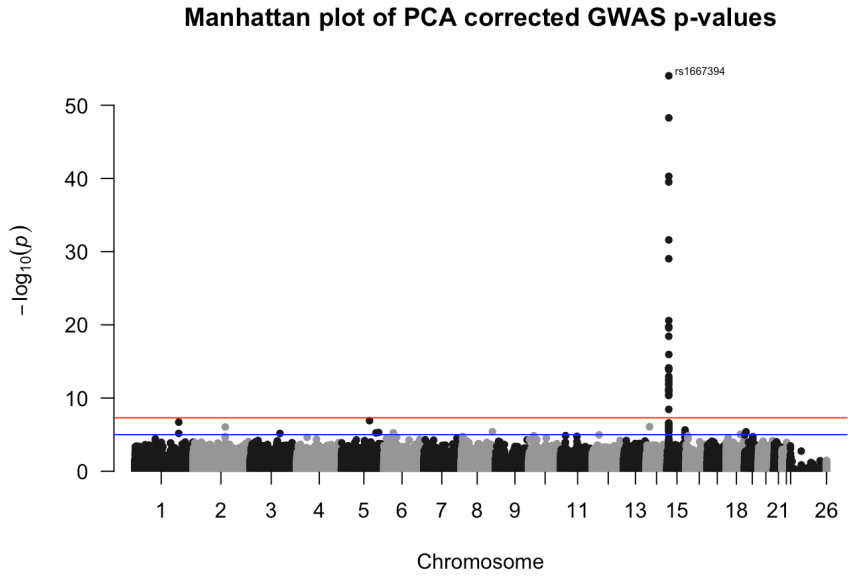

Fig. 5: Manhattan plot. Genome-wide significance line was set to $-log_{10}(5 \cdot 10^{-8})$.

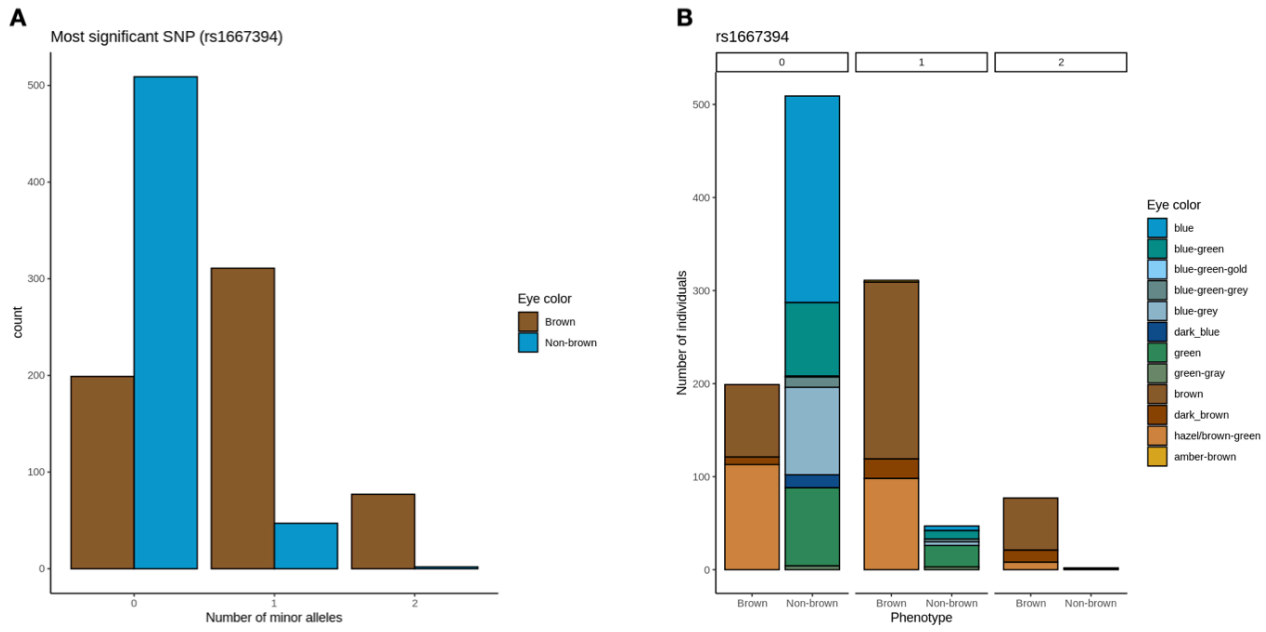

Fig. 6: Genotype distribution of rs1667394 with corresponding **(A)** binary and **(B)** self-reported phenotypes.

Genotypic effects

In order to investigate the genotypic effects of the SNPs associated with eye color, the most significant variant (rs1667394) was investigated further. This was done by 'zooming in' on this variant and the variants downstream and upstream within a window of 200 kb, and then finding the genotypes depending on the number of minor alleles. The distribution of the genotypes in rs1667394 was visualized for each individual with their corresponding binary phenotypes (**Fig. 6A**) and the self-reported phenotypes (**Fig. 6B**). The plots indicate that no individuals with non-brown eye color have a genotype with two minor alleles. Therefore, the allele that corresponds to the non-brown phenotype must have a recessive effect. On the other hand, individuals with brown eye color can have all three genotypes, which indicates that the allele for the brown phenotype has a dominant effect. However, since the heterozygous genotype is more pronounced than the homozygous genotype, it might even indicate that the allele has an overdominant effect [7].

When looking at the other variants around rs1667394 (**Fig. 7**), the same pattern can be observed for the allele associated with the non-brown phenotype, however the allele associated with the brown-phenotype does only seem to have an overdominant effect in some of the variants.

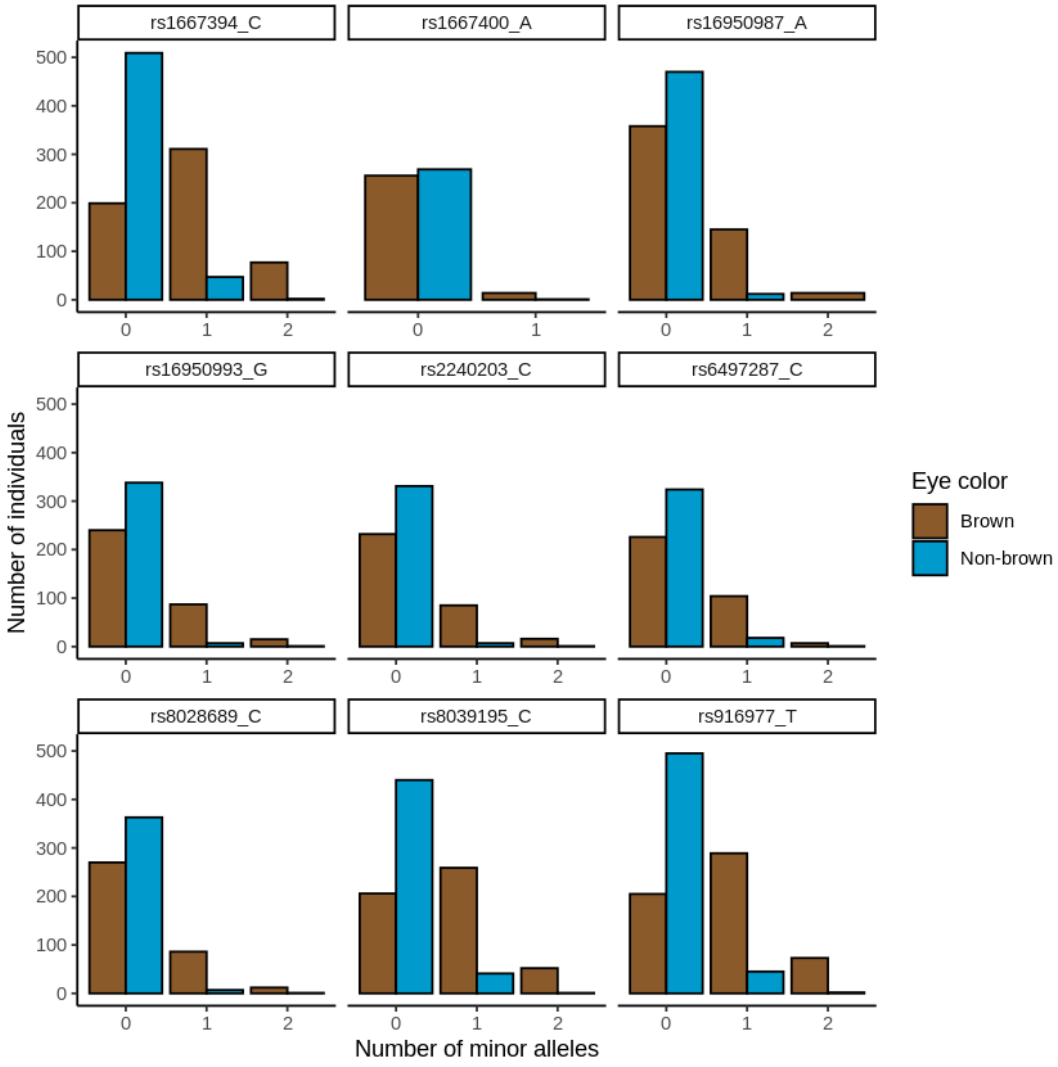

Fig. 7: Genotype distributions of nearby SNPs.

Replicating paper variants

Based on the findings in this report, it could be interesting to compare the found significant SNPs with the significant SNPs found in the Simcoe et al. paper (2021) [2]. 115 significant SNPs from the article were compared to the 27 significant SNPs from this report. In total, 3 significant SNPs were directly found in the article, namely rs1667394 (Chr15:28530182), rs12593929 (Chr 15:28359258) and rs749846 (Chr15:28268990), which are all located on the two neighboring genes, OCA2 and HERC2 (**Fig. 8**). As there were only 3 identical significant SNPs, it was then investigated if any of the significant SNPs found, was in close proximity to the ones in the article. Both the 1,000 (**Fig. 9A**) and 10,000 (**Fig. 9B**) closest SNPs were investigated. When looking at the 1,000 closest variants, 8 significant SNPs from this report were in close proximity to the ones in the article, whereas when looking at the 10,000 closest variants, 23 significant SNPs from this report were in close proximity to the ones in the article.

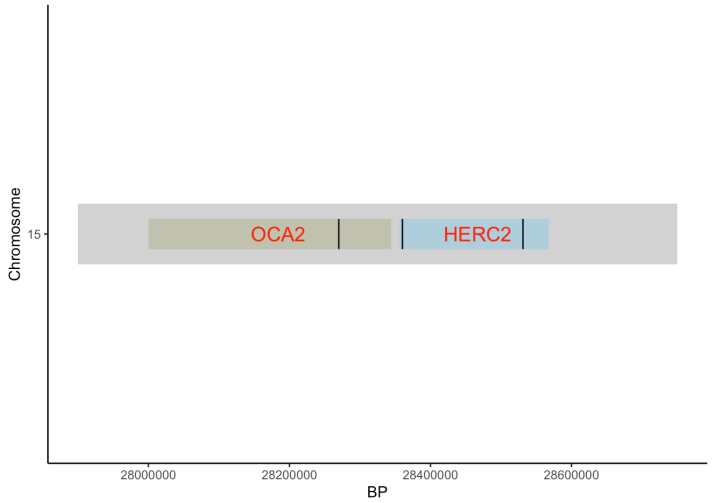

Fig. 8: Location of identical variants.

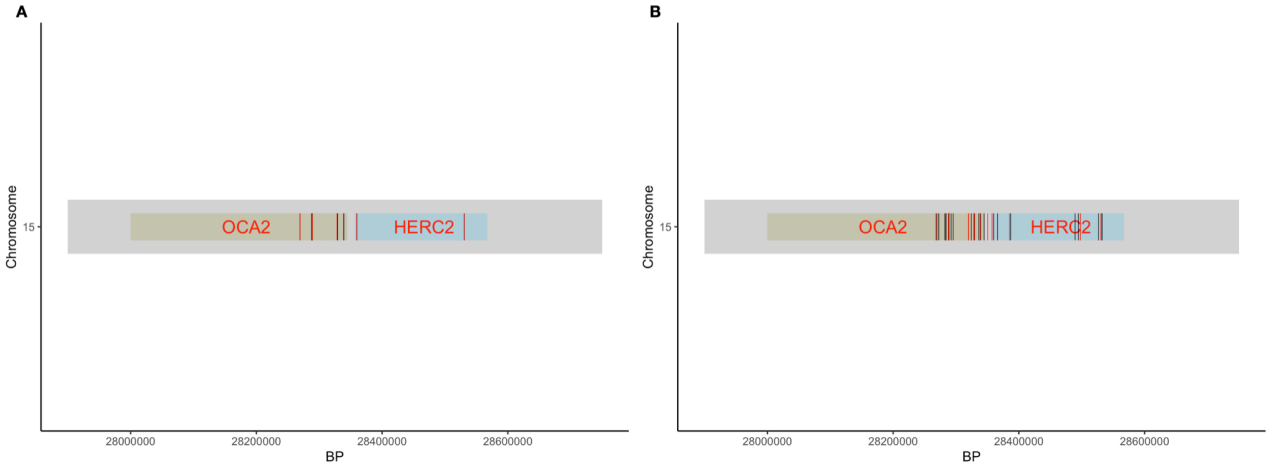

Fig. 9: Location of significant SNPs (black) in relation to **(A)** 1,000 closest (red), and **(B)** 10,000 closest (red) SNPs from the article.

CONCLUSION

According to previous genome-wide association studies of eye color, the two neighboring genes, OCA2 and HERC2, have shown the strongest genetic influence on eye color, and it would therefore be expected to find significant loci within these locations. The phenotypes in the used GWAS data were self-reported, and they were therefore reduced from 12 phenotypes to binary phenotypes. On these binary phenotypes, a Fisher's exact test was performed, which identified 749 significant loci. Hereafter a Bonferroni correction was made to correct for multiple comparison. This reduced the number of significant loci from 749 to 42, however the data was still inflated. In order to correct for this inflation, two things were done. First a genomic control was made to correct for the inflation factor, however that did not seem to help as much as would have been preferred. Therefore a PCA correction was made using logistic regression, which reduced the inflation factor from 1.9672 to 1.0238. Afterwards, 27 significant loci were found, all located on chromosome 15, using a genome-wide threshold of $p < 5 \cdot 10^{-8}$. The distribution of the genotypes for the most significant SNP and the nearby SNPs were then further investigated, and it was discovered that the allele associated with non-brown eye color had a recessive effect and the allele associated with brown eye color had a dominant effect, as would have been expected. Then 115 significant SNPs from Simcoe et al. (2021) was compared to the 27 significant SNPs found in this report, which resulted in only 3 identical variants. At last, an investigation of variants in close proximity with the ones from the article was made, which resulted in 8 SNPs in close proximity when looking at the 1,000 kb closest SNPs, and 23 SNPs in close proximity with the ones from the article, when looking at the 10,000 kb closest SNPs. All the significant SNPs found were located within either the OCA2 or HERC2 gene, confirming the importance of these two genes in the genetics of eye color.

REFERENCES

[1] G. Bräuer and V. P. Chopra, "[Estimation of the heritability of hair and eye color]," (in ger), *Anthropol Anz,* vol. 36, no. 2, pp. 109-20, Feb 1978. Schätzungen der Heritabilität von Haar- und Augenfarbe.

[2] M. Simcoe *et al.*, "Genome-wide association study in almost 195,000 individuals identifies 50 previously unidentified genetic loci for eye color," *Science Advances,* vol. 7, no. 11, p. eabd1239, 2021, doi: doi:10.1126/sciadv.abd1239.

[3] "Is eye color determined by genetics?" https://medlineplus.gov/genetics/understanding/traits/eyecolor/ (accessed 05.25, 2022).

[4] P. Sulem *et al.*, "Genetic determinants of hair, eye and skin pigmentation in Europeans," (in eng), *Nat Genet,* vol. 39, no. 12, pp. 1443-52, Dec 2007, doi: 10.1038/ng.2007.13.

[5] "openSNP." https://opensnp.org (accessed 05.25.22, 2022).

[6] E. Reed, S. Nunez, D. Kulp, J. Qian, M. P. Reilly, and A. S. Foulkes, "A guide to genome-wide association analysis and post-analytic interrogation," (in eng), *Stat Med,* vol. 34, no. 28, pp. 3769-92, Dec 10 2015, doi: 10.1002/sim.6605.

[7] "Overdominance." https://www.biologyonline.com/dictionary/overdominance (accessed 05.25, 2022).

[8] "OCA2 gene." https://medlineplus.gov/genetics/gene/oca2/ (accessed 05.25, 2022).