GWAS of eye color

Abstract

Human eye colors varies in different areas, and are highly heritable. There are many potential impacts effect on human eye colors. In my analysis, I use some genome-wide association studies methods to find the correlation between phenotypes of eye colors and genotypes. I also assume it is influenced by natural selection and genetic drift, which is also proved. Overall, human eye color is a genetically difficult phenotype trait, and is influenced by many aspects.

Introduction

Eye color is one thing that varies in human population. As we know, in European populations, display the largest diversity of iris color, varying from the lightest blue to darkest brown. The outcomes of blue eyes correlates with geographic latitude, and also likely a result of human migration, sexual, and natural selection[1].

We can use some genome-wide association studies(GWAS) methods to identify the genes and alleles. The phenotype we are looking at is self-reported eye color. For eye color, there are 12 different colors, while in this project, because of the limitation of GWAS and PCA, I change the colors into brown and blue. Human eye color is significantly heritable, so we want to know what contributes to the different color of human eyes. We identify 1287 eye color samples, without caring about gender difference, trying to find the associations of it with genes. Our study reports a GWAS in these data. It is well established that the outcomes of different eye colors are effected by genes, with selection, drift, and etc.

In previous experiment, we know that iris pigmentation variation in Asians is genetically similar to Europeans, albeit with smaller effect sizes. the genetic complexity of human eye color considerably exceeds previous knowledge and expectations, highlighting eye color as a genetically highly complex human trait[2]. This can also be proved by my analysis.

Overall, our findings indicate the complex genetic effects contribute on phenotypes in human eye colors.

Results

In the data, there are 1287 eye color samples, while some of which are closely related to each other, which we should filter out. By using pruning and removing wrong ibd, 14 individuals are filtered out, and 1273 samples remained. This is shown in Fig 1A. In Fig 1B, it contains the observed number of homozygotes. In y-axis, it calculates the heterozygosity rate per individual.

A

```
1287 people (0 males, 0 females, 1287 ambiguous) loaded from .fam.
Ambiguous sex IDs written to eye_color_QC.nosex .
--remove: 1273 people remaining.
```

B

Fig. 1 (A)14 individuals have been removed. (B) heterozygosity rate per individual.

In this data, I replace all phenotypes to only brown and blue, so it's a binary-phenotype. In Table 1, it shows the phenotype information. In eye_color column, 1 represents eye color as brown, and 2 represents it as blue. Based on this, PCA test shows the distribution of PC1 versus PC2, and PC2 versus PC3. See this in Fig 2. We can see that PC1 and PC2 are orthogonal with each other, which is better than PC2 versus PC3.

id	IID	eye_color
1010	1010	1
1013	1013	1
1020	1020	2
1022	1022	2
1024	1024	2
1026	1026	1
1028	1028	2
1033	1033	1

Table 1 Phenotype data

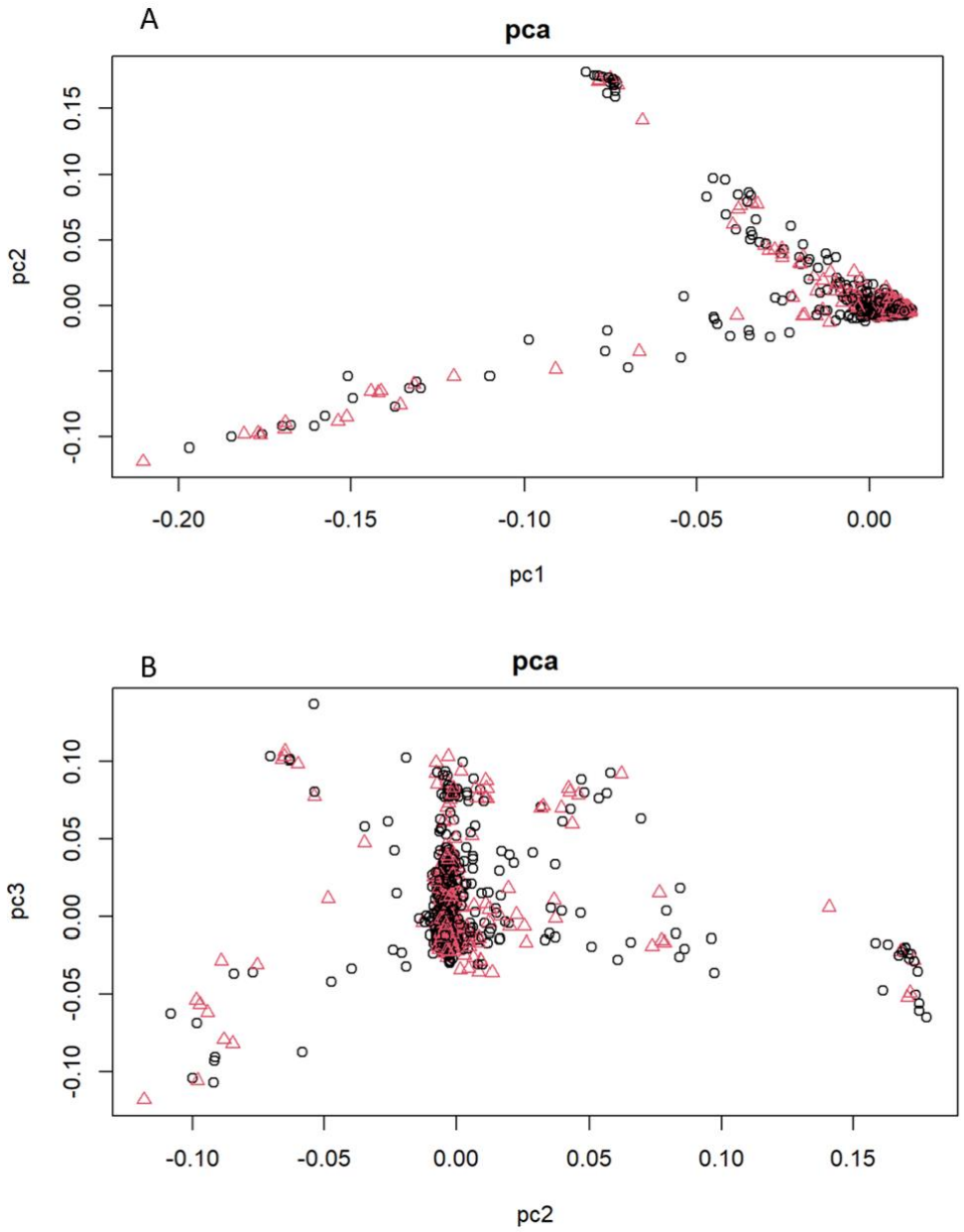

Fig 2 Overview of PCA the analysis based on 2 phenotypes. (**A**) PC1 versus PC2. (**B**) PC2 versus PC3.

Based on the bi-phenotype data and .fisher data, we can easily plot p-value of phenotype with Bonferroni corrected significance level. And we can see that, in Fig 3, there is one significant loci, rs1129038. There are also some relatively significant loci, while these are not comparable with re1129038. Using the same dataset, we can discover the comparison of two probability distribution. In Manhattan Plot, we have seen some SNP has strong correlation with phenotypes, thus the p-value of which are less than 10^(-6). The height of the locus-log10(p-value) on the Y-axis corresponds to the degree of association with the phenotype, and the stronger the association (that is, the lower the p-value), the higher. In general, due to the linkage disequilibrium (LD) relationship, those SNPs around the strongly associated loci will also show similar signal intensities and decrease in turn to both sides.

While it is still not enough considered that these loci are significantly associated with the phenotype. This is because mutations at genetic loci on the genome usually come from two sources:

The first is natural selection. It can refer to the natural selection of competition, but also refers to all the “forces” that have an impact on the fitness of species, such as high radiation environment, disease, virus, etc.

The second is genetic drift, which is a relatively random genome mutation with a large number. Although it is also an important force in species evolution, because its mutations are relatively random, it is currently believed that it is related to the environment. There is no necessary connection between the changes of the , but at some point, some random mutation brings a survival advantage, and it will show its effect in the population. But in the vast majority of cases, for the traits that have been stably existing in the population, they are not considered to have obvious effects, so GWAS research does not care about this type of mutation, and we should exclude them all.

In Fig 4, we can see the output is not concentrated in a straight line with the continuous uniform distribution(expected value). When x value is larger than 4, the y value starts rising, this is due to random drift + true site. We can conclude that phenotype is subjected to natural selection. In other words, The power of natural selection is clearly shown, causing the results to quickly move away from randomness in the population.

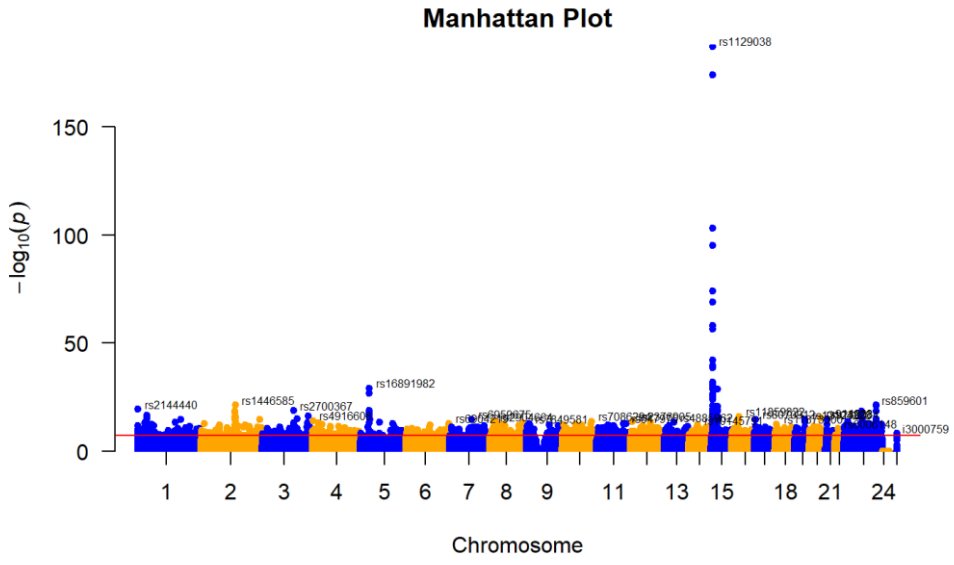

Fig 3 Manhattan Plot of p-value in each chromosome.

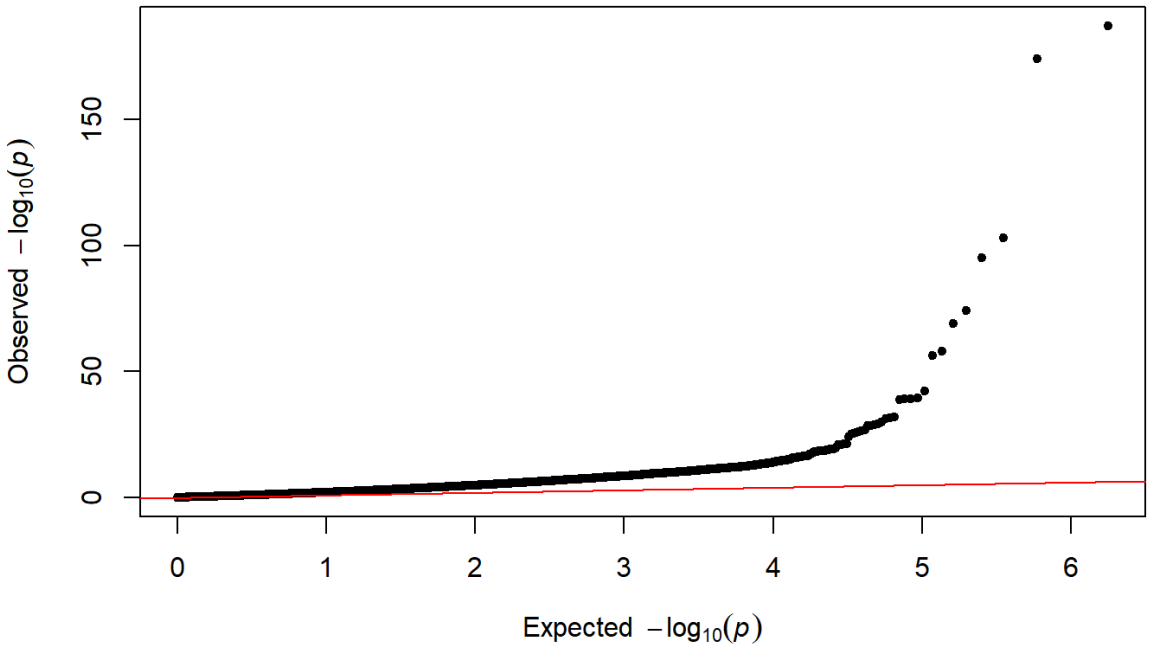

Fig 4 qqplot shows the comparison of expected and observed p-value.

Since every individual is diploid, we can analyze on phenotype with genotype. Rs1129038 is the significant SNP. In Fig 5, it shows that as the proportion of some alleles, one phenotype shows more than the other, while it has not reached to dominant, so it's an additive effect. When two dominant genes exist at the same time (A_B_), they appear as one trait. While in some other SNPs, the situation may vary. In Fig 6, the multiple plots show that some effects are additive, some are recessive. Among the two pairs of interacting genes, one pair of recessive genes plays an epistatic role to the other pair of genes, and the segregation ratio of F2 is 9:3:4. It can be understood that when aa exists, the roles of B and b are covered, that is, aaB_ and aabb both show a certain character.

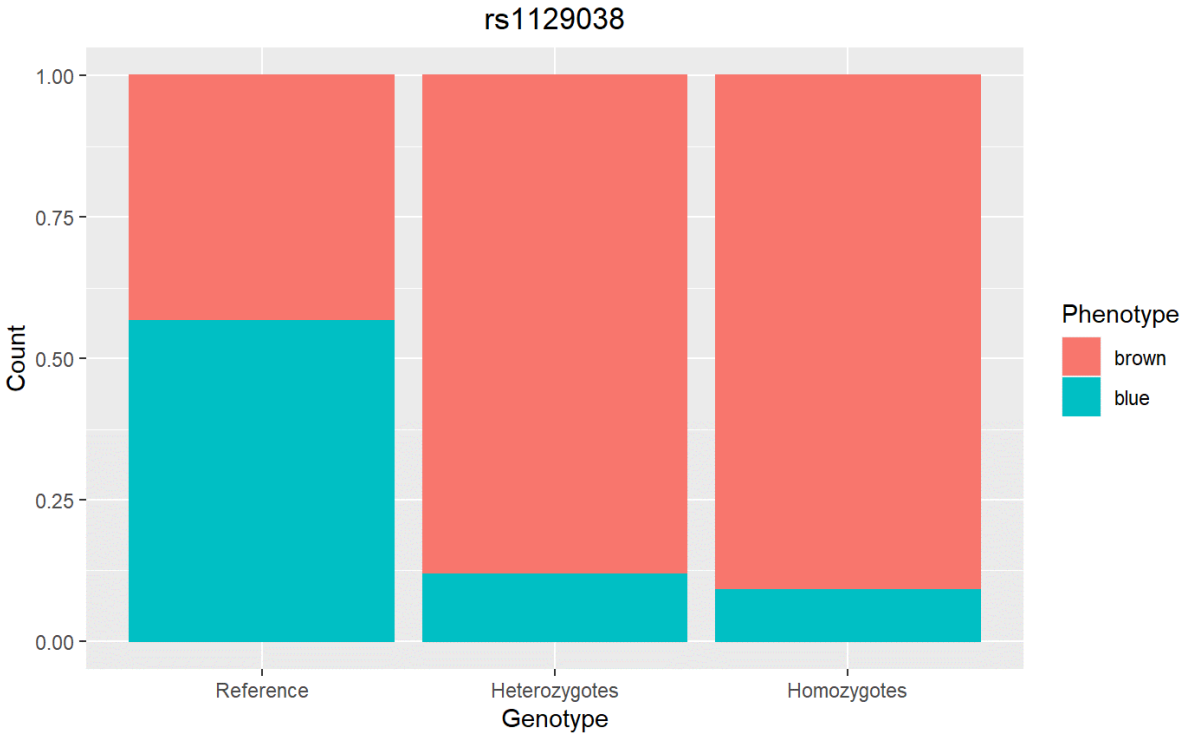

Fig 5 Distribution of eye colors for each genotype of the most significant SNP rs1129038

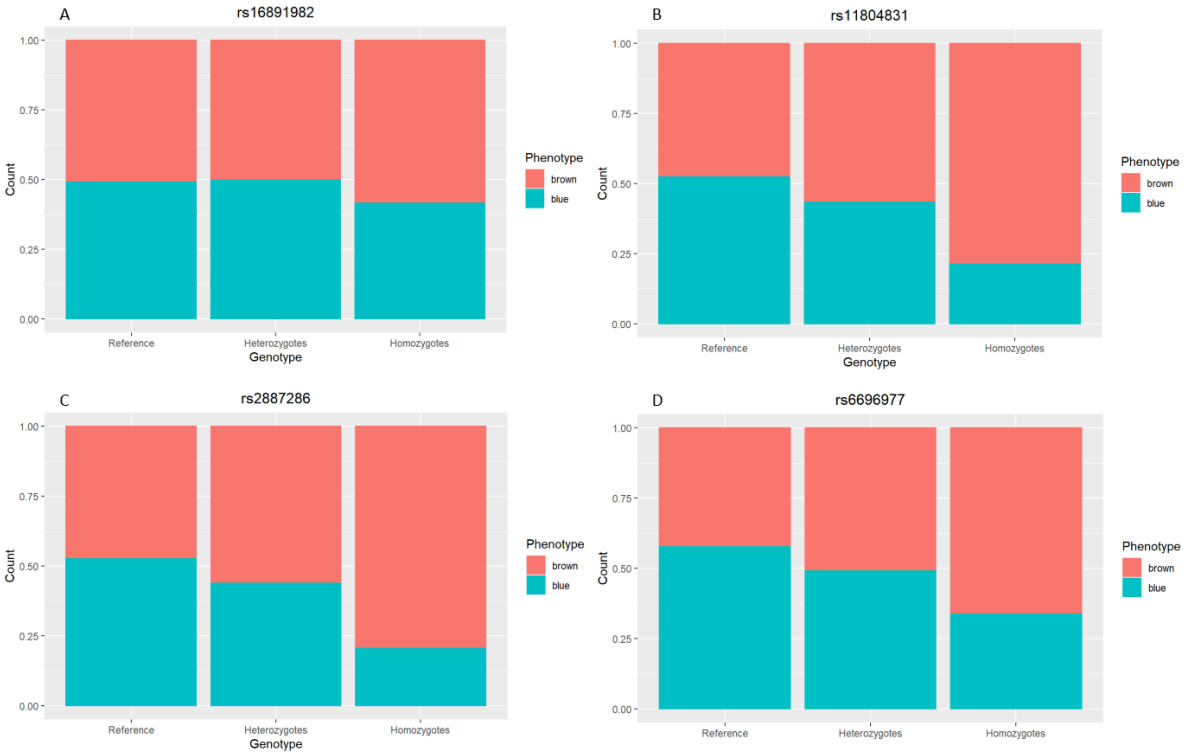

Fig 6 Distribution of eye colors for each genotype of some non-significant SNPs. (**A**) SNP of rs16891982, it has recessive effect in heterozygote, while it has additive effect in homozygote. (**B-D**) SNPs all have additive effects.

Since we have the most significant loci, I use eigenvector to do a linear association test. I still use Manhattan plot and qqplot to show it. It, Fig 7, shows a different result other than the Manhattan

and qq plot in previous test(Fig 3 and Fig 4). We can easily see that there is no pretty significant loci. There is only one solitary point with no significant surrounding points, which is likely to be a false positive. It means that it doesn't effect much on phenotype. The qqplot also indicates this. In qqplot, the distribution is a uniform distribution, which means it's an output of random drift.

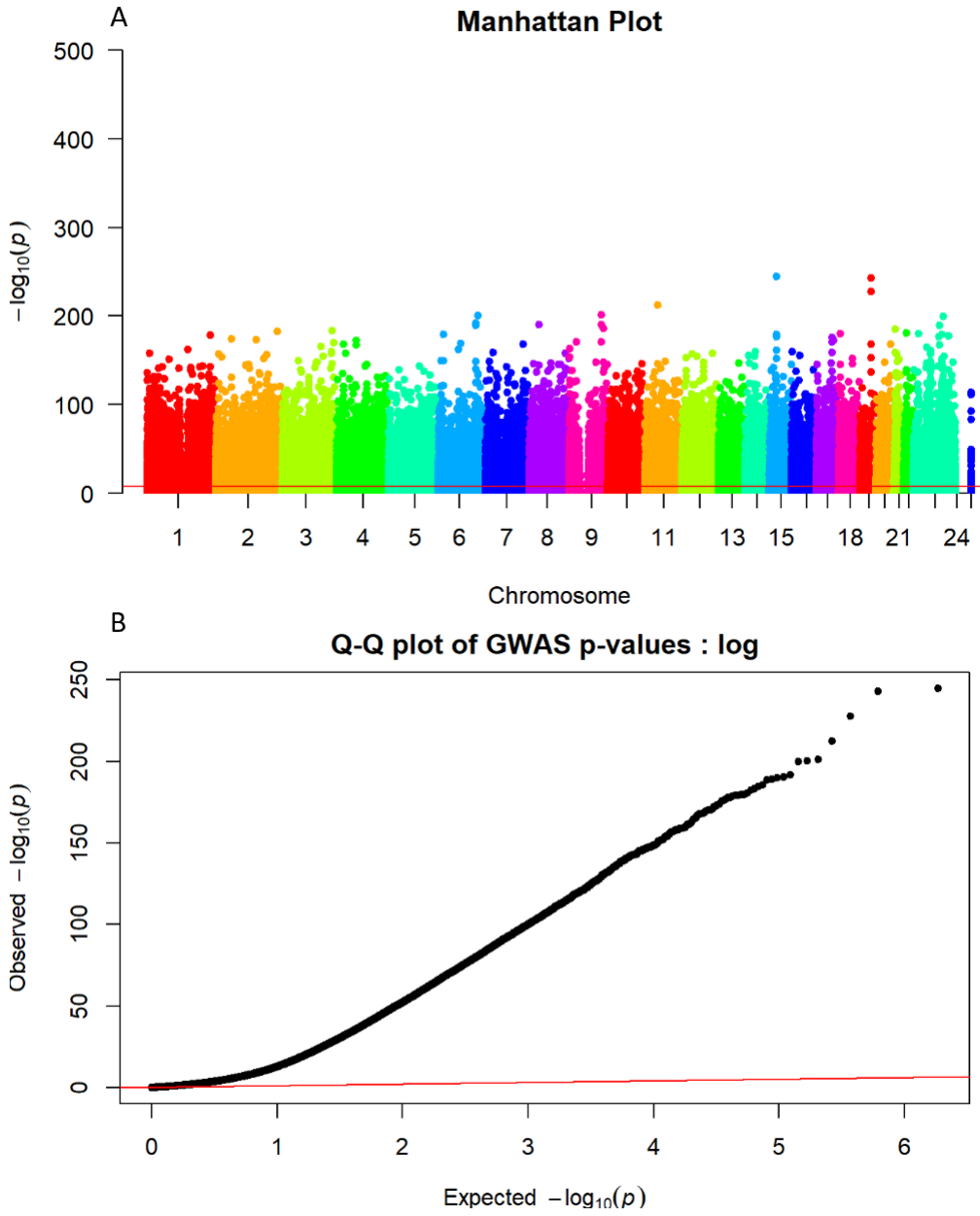

Fig 7 Linear association tests on significant loci.

Conclusion

I analyze the genotype and phenotype of human eye color, based on GWAS. In this analysis, we can see that multiple phenotypes of human eye colors have correlation with its genotype, which is effected by natural selection. In this project, I treat the phenotype as a binary case. We can see from the GWAS test, that there is only one significant SNP, and conclude that phenotype is subjected to natural selection. In other words, The power of natural selection is clearly shown, causing the results to quickly

move away from randomness in the population. There is one SNP that contributes the most, which represents the genetic loci associated with human eye color. For the most significant SNP(rs1129038), the genotype shows an additive effect on phenotype. For other SNPs, there are additive and recessive effects. When we do linear association tests on most significant SNP, I find it doesn't effect much on phenotype, which means it's an output of random drift. The findings demonstrate that eye color is effected by many aspects, such as heritance, natural selection, genetic drift, which can be used for further studies and improvements.

References

1. Simcoe M, Valdes A, Liu F, Furlotte NA, Evans DM, Hemani G, Ring SM, Smith GD, Duffy DL, Zhu G, Gordon SD, Medland SE, Vuckovic D, Girotto G, Sala C, Catamo E, Concas MP, Brumat M, Gasparini P, Toniolo D, Cocca M, Robino A, Yazar S, Hewitt A, Wu W, Kraft P, Hammond CJ, Shi Y, Chen Y, Zeng C, Klaver CCW, Uitterlinden AG, Ikram MA, Hamer MA, van Duijn CM, Nijsten T, Han J, Mackey DA, Martin NG, Cheng CY; 23andMe Research Team; International Visible Trait Genetics Consortium, Hinds DA, Spector TD, Kayser M, Hysi PG. Genome-wide association study in almost 195,000 individuals identifies 50 previously unidentified genetic loci for eye color. Sci Adv. 2021 Mar 10;7(11):eabd1239. doi: 10.1126/sciadv.abd1239. PMID: 33692100; PMCID: PMC7946369.

2. Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, Manolescu A, Karason A, Palsson A, Thorleifsson G, Jakobsdottir M, Steinberg S, Pálsson S, Jonasson F, Sigurgeirsson B, Thorisdottir K, Ragnarsson R, Benediktsdottir KR, Aben KK, Kiemeney LA, Olafsson JH, Gulcher J, Kong A, Thorsteinsdottir U, Stefansson K. Genetic determinants of hair, eye and skin pigmentation in Europeans. Nat Genet. 2007 Dec;39(12):1443-52. doi: 10.1038/ng.2007.13. Epub 2007 Oct 21. PMID: 17952075.