GWAS project - Asbjørn Kjær

Population genomics SS2021

Abstract

In this project we conduct a GWAS of human eye color using 1260 individuals of unknown ancestry. The data was obtained from the openSNP database. First, we test for association using light and dark eye color phenotypes and find significant variants in the OCA2/HERC2 region, which is known to be associated with eye color. Second, we replicate testing methods used in previously published papers and find association in the same region using these techniques as well. Due to limited sample size no method can find significant variants in other regions than OCA2/HERC2.

Introduction

Eye color is a highly genetic trait and has had the interest of geneticist for a long time, often in combination with other pigment related traits. Eye color variation is largest in European populations, but it does also vary in other populations. It was originally thought to be a monogenic trait, where blue was recessive and brown was dominant[1]. However, later studies have shown that the trait is polygenic with as much as 61 gene regions being involved[2]. In this project we use SNP data to test for association between eye color and genotype based on chip data and self-reported eye color. GWAS studies are heavily influenced by population stratification, which can inflate p-values and create false positives. It is possible to eliminate inflation of p values by doing correction for stratification. The most commonly used methods are EIGENSTRAT[3], where principal components are included in the model and mixed model methods, where a genetic relationship matrix is included in the model[4]. The main limitation of GWAS studies is power to detect significant variants. Power increases with sample size, and most modern GWAS studies uses over 100.000 individuals in order to have enough power to obtain significant p-values for rare and weak variants. In this project we have 1260 individuals, which limits power significantly.

Results/discussion

Data

The data provided is from openSNP[5], which allows users of direct-to-consumer genetic tests to share their data. This has some impact on the data that needs to be considered during analysis. Firstly, the eye color phenotype is self-reported, which makes distinction between specific eye colors subjective and less precise (if not simply wrong sometimes). Secondly, ancestry of the samples is unknown and diverse. It will later be clear that the individuals are relatively genetically diverse, which creates population stratification when testing. Normally GWAS studies use relatively homogenous populations to prevent stratification from influencing results. We could have filtered out specific ancestry based on reference genomes but have not done so due to already limited sample size (and time constraints). This approach is used in the Simcoe paper, since they also use data from direct-to-consumer genetic tests[2]. Thirdly, the data is collected from different SNP chips, which do not genotype the same set of SNPs. Therefore, some individuals have high proportion of missing data. (See supplementary fig 1). This means that individuals with high rates of missing data cannot be directly interpreted as bad quality runs. These individuals are genotyped with a chip that genotypes fewer SNPs, and these individuals should therefore not necessarily be excluded. However, this prevents us from catching and excluding individuals with bad genotyping, which should be kept in mind. This probably does not matter, since the companies most likely do quality control of their data before returning it to costumers, which means the data is prefiltered. Normally abnormal heterozygosity rate would also be used as a proxy for contaminated runs, but since the four groups in fig S1 have different mean heterozygosity, we do not filter based on heterozygosity rate. Again, bad genotyping would presumably have been filtered by the companies already. Ideally one would get the data from each chip individually and do quality control for each chip type.

Quality Control

As mentioned in the previous section the data is not filtered based on outlying heterozygosity or high proportion of missing SNP data.

We filter out related or duplicated individuals by calculating the proportion of the genome that is identical by descent (IBD). We use a IBD threshold of 0.09375, which filters out anything more related than first cousins (expected IDB: 0.125). This filters out 27 persons. SNPs are then flittered. Because of the different chips we only filter out SNPs with a very high missing rate (missing for 75% of samples). We filter out SNPs that are very far from Hardy Weinberg equilibrium (p<1e-5), since large deviations from HWE is expected to be due to miscalls. SNPs with minor allele frequency under 1% were also filtered out.
After quality control we are left with 1260 individuals and 837,498 variants.

Brown vs light phenotype

The phenotype of the samples is self-reported eye color. It is possible to choose between 12 different eye colors:

> Brown, hazel/brown-green, blue, blue-green, green, blue-grey, dark-brown, amber-brown, dark-blue, green-gray, blue-green-gold & blue-green-grey.

Based on the presumption that brown eyes represent an ancestral state, we divide the eye colors into a binary phenotype of dark/brown colored eyes and light-colored eyes. The brown category contains “brown”, “hazel/brown-green”, “dark brown” and “amber brown” and has 646 individuals, while the light contains the remaining colors and has 614 individuals.

Simple association

To test for association, we performed fishers exact test on each SNP. The results can be seen in figure 1. The p values are clearly inflated and has an inflation factor (λ) of 1.94.

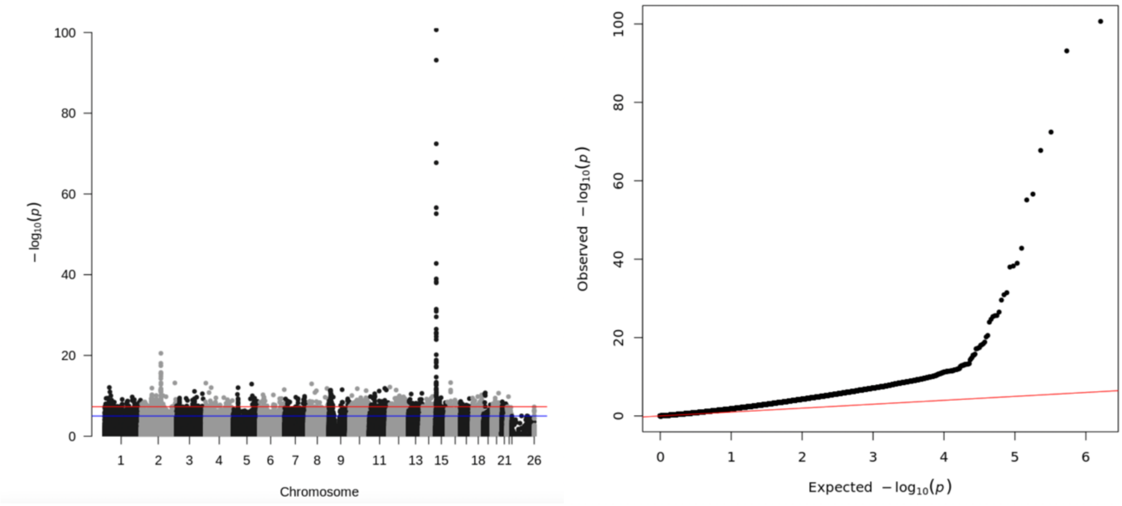


Figure 1: Manhattan- and qq-plot for fisher’s exact test

Since eye color is strongly corelated with race we expect the p values to be inflated if our sample data contains individuals of heterogenous ancestry. It is clear we need to correct for population stratification/structure. We could use genomic control, but this method is not advised for inflation factors over 1.05[4]. Instead we correct for population stratification by incorporating the principal components (PC) of the dataset into the model[3]. This works since the first couple of PCs captures the genetic stratification.

Correction for Population stratification

To calculate we first prune out SNPs in LD with each other and then perform the principal component analysis (PCA). The results are plotted in figure 2.

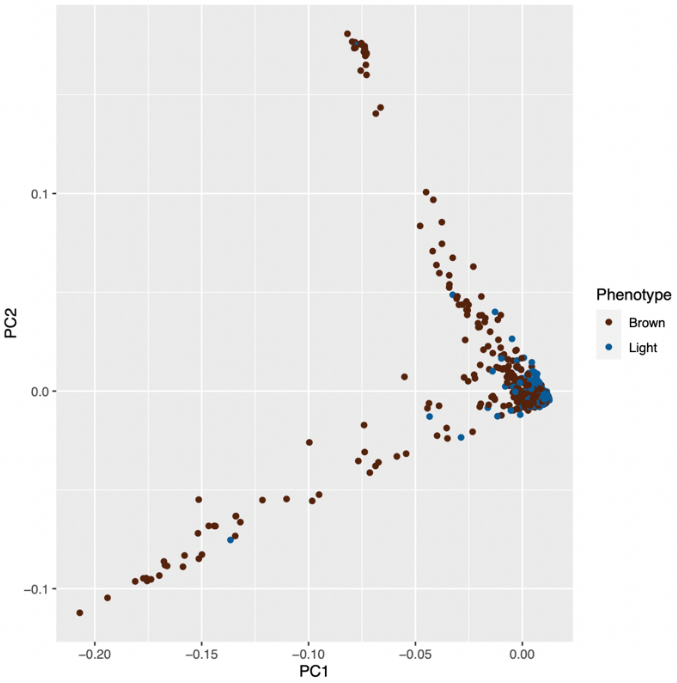


Figure 2: Plot of the 2 first PCs

As expected, the PCA plot reveals a large amount of population structure. Brown eyes have high frequency in the two "arms" of the plot. Presumably, the two arms represent individuals with either Asian or African descent, while the middle region represent European ancestry. Individuals in the middle of the arms probably represent admixed individuals. This speculation could be tested by comparing to reference genomes, but that is beyond the scope of this project. To obtain a more homogeneous sample we could exclude individuals that are not part of the middle cluster, but this would limit sample size further.

Some studies use imputation to increase the number of tested SNPs.[2] This method uses reference genomes to infer the genotype of non-genotyped SNPs. However, this only works properly if the sample data and reference genomes are from the same homogenous population. (Similar allele frequencies and LD). Imputation would probably infer wrong genotypes for this dataset, due to the relative heterogenous samples.

We now test for association using a logistic regression model that includes the first 2 PCs in the model. Results are plotted in figure 3.

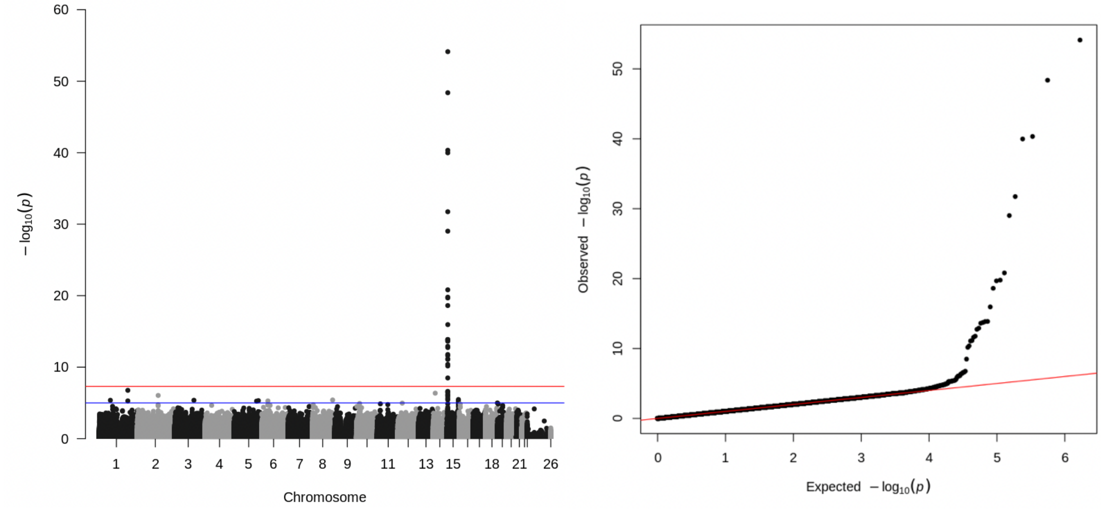


Figure 3: Manhattan- and qq-plot for the PC adjusted model. *The inflation factor has been reduced to 1.02.*

The qq-plot now looks as expected for a GWAS. However, only 1 region is still significant with a p-value threshold of 5e-8. There are 24 significant SNPs within a 261kb window that spans the start of the OCA2 gene and the end of HERC2 (downstream of OCA2). 14 variants are inside OCA2 and 10 are inside HERC2. (See source code for table of significant SNPs). The most significant SNP is rs1667394 with a p-value of 7.4e-55. Some of the significant SNPs are explored further in the following sections.

Testing using mixed model

Another way of correcting for population stratification is by using a mixed model. The model uses a matrix of pairwise relatedness (GRM) to correct for population stratification. If the tested SNP is included in the GRM you overcorrect and lose power, therefore we use the leave one chromosome out approach, where the GRM is calculated using all chromosomes, except the one that contains the SNP being tested.

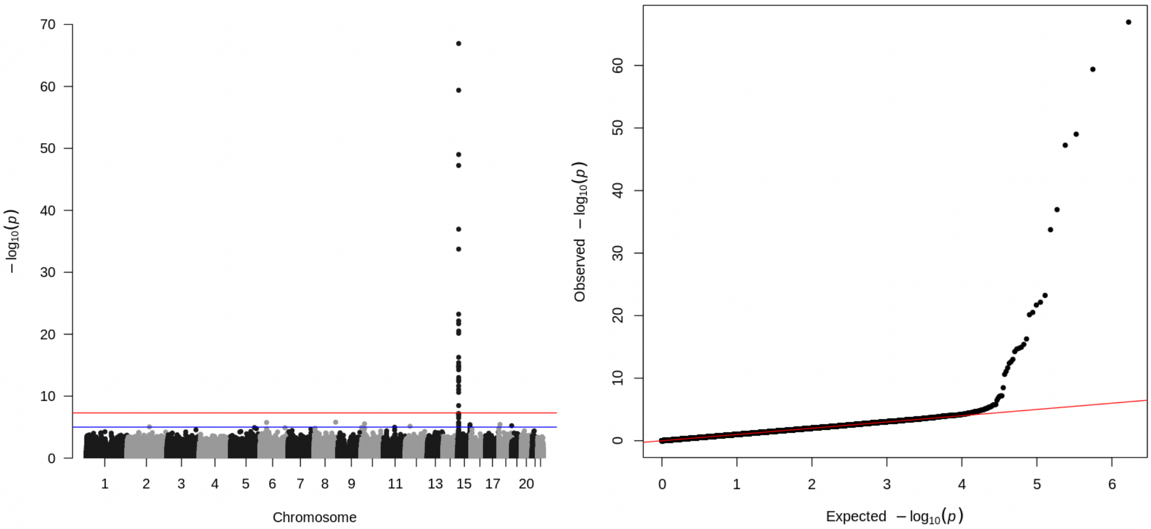


Figure 4: ***Manhattan- and qq-plot for the mixed model method.*** *The inflation factor is 1.004.*

Using this approach, we obtain very similar results. The same 24 SNPs are significant with a 5e-8 threshold. The most significant SNP is still rs1667394, now with a p value of 1.1e-67. In general, the p-values obtained by the mixed model are lower than the PCA method. This could signal that the mixed model approach has greater power, than PCA approach. The inflation factor is also lower, which could signal that mixed model method is better at correction for population stratification.

However, for this specific dataset there is no significant difference between the 2 methods.

Heritability

The GRM matrix used in the mixed model method can also be used to estimate the narrow sense heritability of a trait (additive effects). The analysis estimates that 78% of the phenotypic variation can be explained by genotypic variation. Since all significant variants are on chromosome 15, we estimate how much of the variation can be explained by chromosome 15 alone. Here we estimate that 42% of the variation can be explained by genotype on chromosome 15. This mean that even though all significant variants are situated on chromosome 15, only about half of the genetic heritability is associated with chromosome 15. (0.42/0.78=0.53) (see source code for calculations)

Note that the 78% and 42% estimates almost certainly are underestimates of the true heritability, since most individual genotypic variation aren’t genotyped by this setup.

Additionally, this method only estimates additive effects, which means that if phenotype is influenced by genetic interaction, then this will also be missed.

Phenotype distribution for the most significant SNP

To investigate the effect of the most significant SNP on phenotype, we plot phenotype/genotype distribution of rs1667394 (figure 5). It is clear, that the light-colored eye phenotype (anything over hazel) is strongly associated with the A/A genotype, while the G/G genotype is almost exclusively brown eyes. Interestingly the distribution of subcategories of the light colors seems to also be affected by the genotype. The light colors in the A/G genotype are almost exclusively green-ish colors. Another interesting point is that about 30% of people with the A/A genotype still has brown eyes. Clearly, eye color is not a simple medelian monogenic trait. If one should choose a simple genetic model, it would probably be that the light color trait is additive. As in A/G increases the chance of having light eyes, while A/A increases it further (although it does not exactly double the chance as with true additive traits) You could also argue that the behavior somewhat resembles a recessive trait (especially “pure” blue).

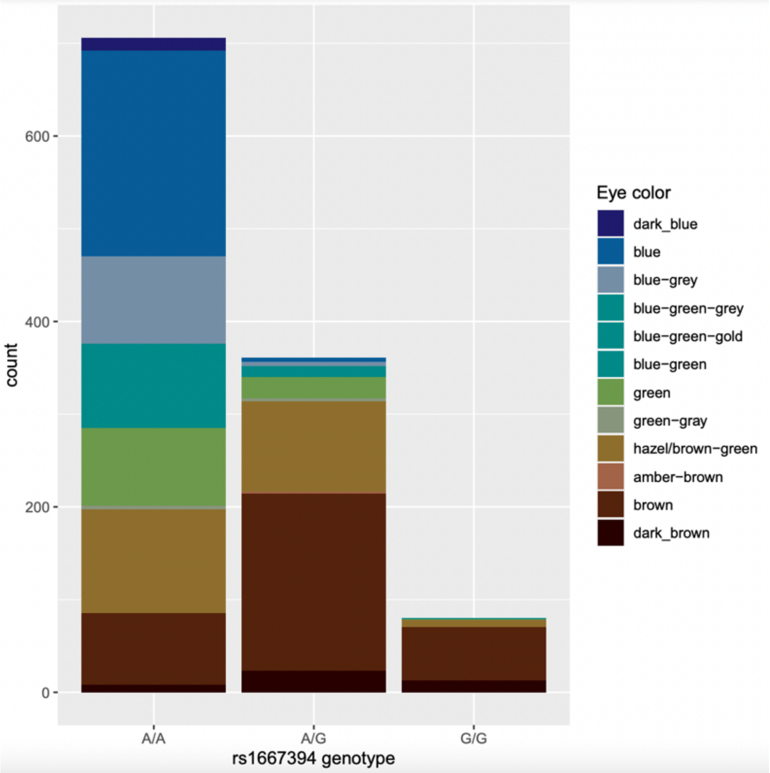


Figure 5: plot of the distribution of phenotypes associated with the genotypes of rs1667394: *(Note our data describes the phenotype as C/T, while publications refer to it as G/A phenotype. This just depends on which strand you reference. To avoid confusion, we will use G/A phenotype).*

Recreating the tests in Simcoe (2021) & Sulem (2007)

Sulem (2007): Blue vs brown & blue vs green

In Sulem et al.[6] the authors tests two hypothesis. Instead of testing light vs brown they test blue vs brown and blue vs green.

To replicate this, we create these groups from our data as described in table 1. We then perform PC on the subset of the data containing only either blue & brown or blue & green. We test for association while correction for the first two PCs as before. The results are plotted in figure 6. Qq-plots looks as expected an can be found in source code

Category	Eye colors included	N individuals
Blue	blue", "blue-grey" &"dark-blue"	374
Green	'green' & 'green-gray'	123
Brown	"brown", "dark-brown" & 'amber-brown'	413

Table 1: division of data into blue, brown & green categories. *Note some intermediate eye colors are not included in any of these categories*

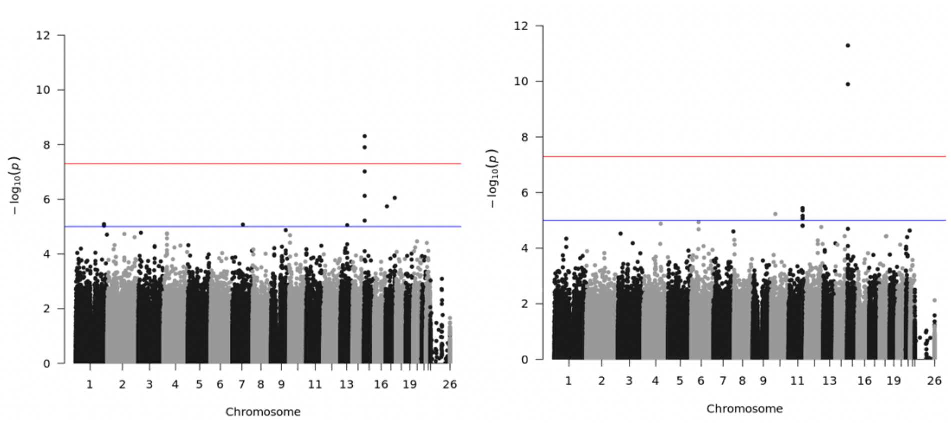


Figure 6: Manhattan-plot for "blue vs green"(left) & "Blue vs brown"(right)

In both cases there are significant SNPs in the OCA2/HERC2 region that were also significant before, and no significant SNPs elsewhere. The p values of the SNPs are less significant than in the brown vs light test. This is probably a result of the lowered sample size. The only new information obtained is that blue vs green eye color is also significantly associated with the OCA2/HERC2 SNPs. As mentioned in the previous section the blue phenotype is almost only seen in the A/A phenotype, while green is seen in both A/A and A/G phenotype. (See fig S2) The paper finds association in the OCA2/HERC2 region for both tests, in addition to association in other regions, which we cannot reproduce, presumably due to lower sample size.

Simcoe (2021): Quantitative eye color phenotype

In Simcoe et al[2] the authors use a quantitative phenotype instead of a binary phenotype. They reference other papers that claim eye color phenotype can be accurately described as a linear scale from blue to brown with green in the middle. They use 6 categories, but since we only have 1 hazel phenotype, we will use 5 phenotypes as described in table 2

Value	Eye colors included	N individuals
0	blue, blue-grey, dark-blue	380
1	blue-green-grey, blue-green-gold, blue-green	113
2	Green, green-gray	130
3	hazel/brown-green, amber-brown	244
4	brown	364
5	Dark-brown	50

Table 2: division of data into quantitative phenotype.

In their paper they use a linear regression model and correct for age, sex, platform and the five first PCs. Since we don't have sex, age and platform information we only use the 5 first PCs. The results are shown in figure 7.

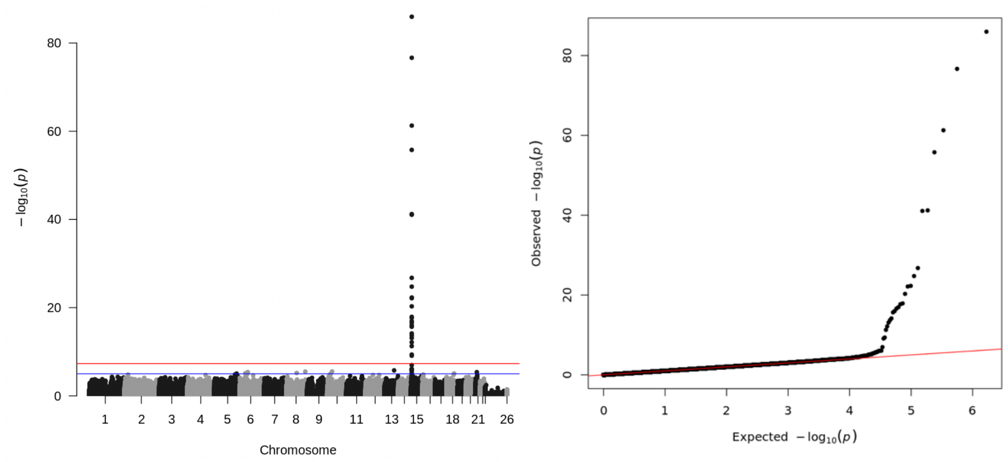


Figure 7: Manhattan and qq-plot for the quantitative phenotype model.

This approach also finds the same 24 significant variants in the OCA2/HERC2 region, that were significant in the light vs brown approach. However, the p-values are generally lower which could indicate that this model has higher power. It shows that using a quantitative phenotype is a valid approach.

The advantage of using a quantitative model is that you essentially test blue vs brown and blue vs green etc. in one model. A variant that differentiates any of the categories should be tested significant given enough statistical power, which sadly we lack in this dataset. The authors find 61 regions using this approach, while we can only find 1 (which they count as 2), but their sample size is also almost 200 times larger and from a homogenous population. We have repeated the test using mixed model instead of PCA, but the results are the same. (See source code)

Conclusion

In this project we have tested for variants associated with either light or brown eye color using genotype data from openSNP. The individuals were heterogenous, so we needed to correct for population stratification. Due to limited sample size, we could only find significant association in one region. This region contains the genes OCA2 & HERC2, which are known to play a relatively large role in eye color phenotype (and other pigment related phenotypes). The most significant SNP found is rs1667394. Almost all light color eyed individuals are heterozygotes or homozygotes for the A variant at rs1667394.

Even though OCA2/HERC2 is the only significant region we estimate that only about 42% of the phenotypic variation is explained by genotype on chromosome 15, where they are situated, while about an equal amount is explained by the rest of the genome. However, these variants are not tested significant in this dataset, presumably because of lack of power. Lastly, we replicate approaches used in other GWAS studies of eye color but are still only able to find significant variants in the OCA2/HERC2 region.

References

1 Davenport, G. C. & Davenport, C. B. HEREDITY OF EYE-COLOR IN MAN. *Science* **26**, 589-592, doi:10.1126/science.26.670.589-b (1907).

2 Simcoe, M. *et al.* Genome-wide association study in almost 195,000 individuals identifies 50 previously unidentified genetic loci for eye color. *Sci Adv* **7**, doi:10.1126/sciadv.abd1239 (2021).

3 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-909, doi:10.1038/ng1847 (2006).

4 Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11**, 459-463, doi:10.1038/nrg2813 (2010).

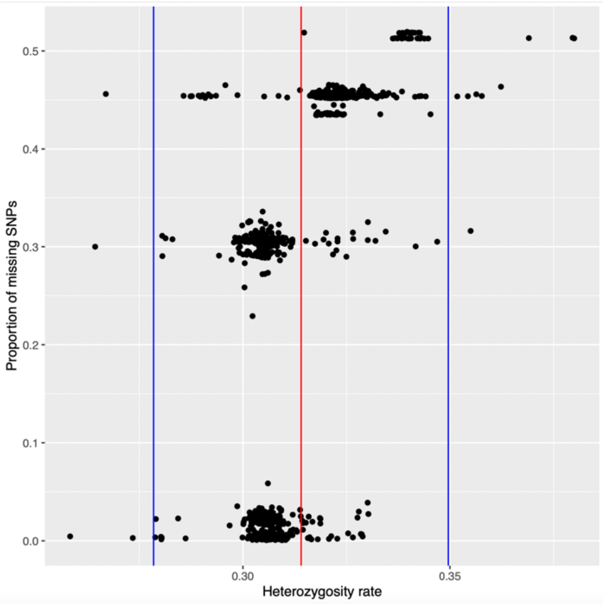
5 *OpenSNP* <https://opensnp.org> (

6 Sulem, P. *et al.* Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat Genet* **39**, 1443-1452, doi:10.1038/ng.2007.13 (2007).

Supplementary

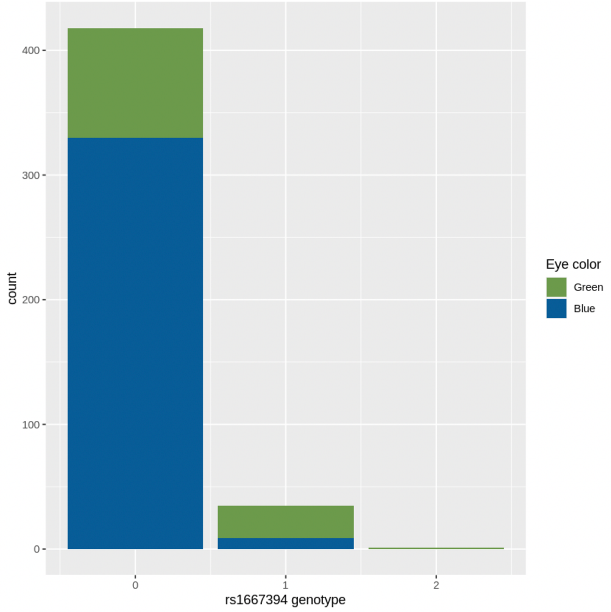
Source code can be found here:
https://github.com/asbjoernkjar/GWAS-Project.git

Fig S1



S1: Plot of proportion of missing data and heterozygosity rate. *Red line is mean heterozygosity and blue lines are +- 3sd from mean heterozygosity.*
It looks like the data comes from at least 4 different SNP chips

Fig S2



S2: distribution of blue and green phenotype for the rs1667394 SNP