GWAS - genome-wide association search - a study aimed at finding relationships between the phenotype (for example, disease) of an organism and a set of SNPs, deletions, insertions and other individual characteristics of the genome (down to individual genes). The analysis is usually performed on two samples. The test sample has the phenotype of interest and the control sample is collected from people of the same sex, age, etc. Today this approach is used to search for associated morbidity of COVID-19.

Our work pursued two goals:

1. Familiarization the PLINK tool for GWAS analysis
2. Comparative analysis of articles on GWAS studies of patients with COVID-19

**Step 1. How to work with PLINK**

**Software**

- PLINK version 1.09 [1], downloaded from http://zzz.bwh.harvard.edu/plink/: QC procedures and statistical analyses
- R version 3.6.3, https://www.r-project.org/ (for correct operation it is recommended to use versions> 3.0.0): generated graphs by the GitHub example scripts (https://github.com/MareesAT/GWA_tutorial/)

**Data**

To be able to illustrate all analysis steps using realistic genetic data, we simulated a dataset (N = 207) with a binary outcome measure using the publicly available data from the International HapMap Project [2]. Here in order to create an ethnically homogenous dataset, Utah residents with ancestry from Northern and Western Europe (CEU) were only included. Because of the relatively small sample size of the HapMap data, genetic effect sizes in these simulations were set at values larger than usually observed in genetic studies of complex traits. It is important to note that larger sample sizes (e.g., at least in the order of thousands but likely even tens or hundreds of thousands) will be required to detect genetic risk factors of complex traits.

**Quality Control of genetic data**

The seven QC steps consist of filtering out of SNPs and individuals based on the following: (1) individual and SNP missingness, (2) inconsistencies in assigned and genetic sex of subjects (see sex discrepancy), (3) minor allele frequency (MAF), (4) deviations from Hardy–Weinberg equilibrium (HWE), (5) heterozygosity rate, (6) relatedness, and (7) ethnic outliers.

1. **Individual and SNP missingness**

Individual-level missingness -- this is the number of SNPs that are missing for a specific individual. High levels of missingness can be an indication of poor DNA quality or technical problems.

SNP-level missingness -- this is the number of individuals in the sample for whom information on a specific SNP is missing. SNPs with a high level of missingness can potentially lead to bias.

Investigate missingness per individual and per SNP and make histograms:

```
plink --bfile HapMap_3_r3_1 --missing
```

output: plink.imiss and plink.lmiss, these files show respectively the proportion of missing SNPs per individual and the proportion of missing individuals per SNP

Generate plots to visualize the missingness results (Fig.1):

```
Rscript hist_miss.R
```

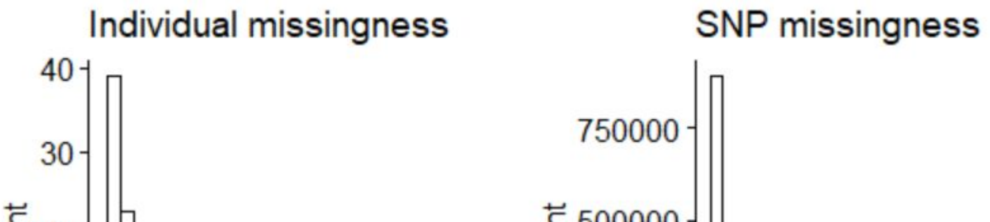

Fig. 1. Histograms of individual and SNP missingness

Next, we need to delete SNPs and individuals with high levels of missingness. For these two commands are used:

`--geno`: excludes SNPs that are missing in a large proportion of the subjects. In this step, SNPs with low genotype calls are removed. It is recommended to first filter SNPs and individuals based on a relaxed threshold (0.2; >20%), as this will filter out SNPs and individuals with very high levels of missingness. Then a filter with a more stringent threshold can be applied (0.02)

`--mind`: excludes individuals who have high rates of genotype missingness. In this step, individuals with low genotype calls are removed. SNP filtering should be performed before individual filtering.

Delete SNPs with missingness >0.2:

```
plink --bfile HapMap_3_r3_1 --geno 0.2 --make-bed --out HapMap_3_r3_2
```

Delete individuals with missingness >0.2:

```
plink --bfile HapMap_3_r3_2 --mind 0.2 --make-bed --out HapMap_3_r3_3
```

Delete SNPs with missingness >0.02:

```
plink --bfile HapMap_3_r3_3 --geno 0.02 --make-bed --out HapMap_3_r3_4
```

Delete individuals with missingness >0.02:

```
plink --bfile HapMap_3_r3_4 --mind 0.02 --make-bed --out HapMap_3_r3_5
```

## 2. Check for sex discrepancy

Sex discrepancy -- this is the difference between the assigned sex and the sex determined based on the genotype. A discrepancy likely points to sample mix-ups in the lab. Note, this test can only be conducted when SNPs on the sex chromosomes (X and Y) have been assessed.

We will use the command `--check-sex`, which checks for discrepancies between sex of the individuals recorded in the dataset and their sex based on X chromosome heterozygosity/homozygosity rates. Subjects who were a priori determined as females must have a F-value of <0.2, and subjects who were a priori determined as males must have a F-value >0.8. This F-value is based on the X chromosome inbreeding (homozygosity) estimate. Subjects who do not fulfil these requirements are flagged "PROBLEM" by PLINK.

```
plink --bfile HapMap_3_r3_5 --check-sex
```

Generate plots to visualize the sex-check results (Fig. 2):

```
Rscript gender_check.R
```

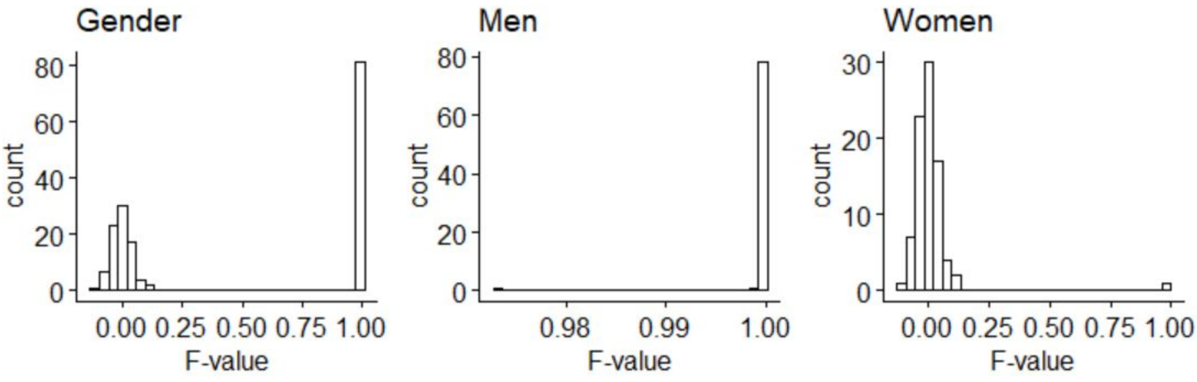

Fig. 2. Sex-check results for all data and for subsets of men an women. Subjects who were a priori determined as females must have a F-value of <0.2, and subjects who were a priori determined as males must have a F-value >0.8. This F-value is based on the X chromosome inbreeding (homozygosity) estimate.

These checks indicate that there is one woman with a sex discrepancy, F value of 0.99. (When using other datasets often a few discrepancies will be found).

Next, we will delete individuals with sex discrepancy.

```
grep "PROBLEM" plink.sexcheck| awk '{print$1,$2}'> sex_discrepancy.txt
```

This command generates a list of individuals with the status PROBLEM.

```
plink --bfile HapMap_3_r3_5 --remove sex_discrepancy.txt --make-bed --out HapMap_3_r3_6
```

This command removes the list of individuals with the status PROBLEM.

### 3. Generate a bfile with autosomal SNPs only and delete SNPs with a low minor allele frequency (MAF)

Minor allele frequency (MAF) -- this is the frequency of the least often occurring allele at a specific location. Most studies are underpowered to detect associations with SNPs with a low MAF and therefore exclude these SNPs.

SNPs with a low MAF are rare, therefore power is lacking for detecting SNP-phenotype associations. These SNPs are also more prone to genotyping errors. The MAF threshold should depend on the sample size, larger samples can use lower MAF thresholds. Respectively, for large (N = 100.000) vs. moderate samples (N = 10000), 0.01 and 0.05 are commonly used as MAF threshold.

The command `--maf` is used for including in the following analysis only SNPs above the set MAF threshold.

Select autosomal SNPs only (i.e., from chromosomes 1 to 22):

```
awk '{ if ($1 >= 1 && $1 <= 22) print $2 }' HapMap_3_r3_6.bim > snp_1_22.txt
plink --bfile HapMap_3_r3_6 --extract snp_1_22.txt --make-bed --out HapMap_3_r3_7
```

Generate a plot of the MAF distribution (Fig. 3):

```
plink --bfile HapMap_3_r3_7 --freq --out MAF_check
Rscript MAF_check.R
```

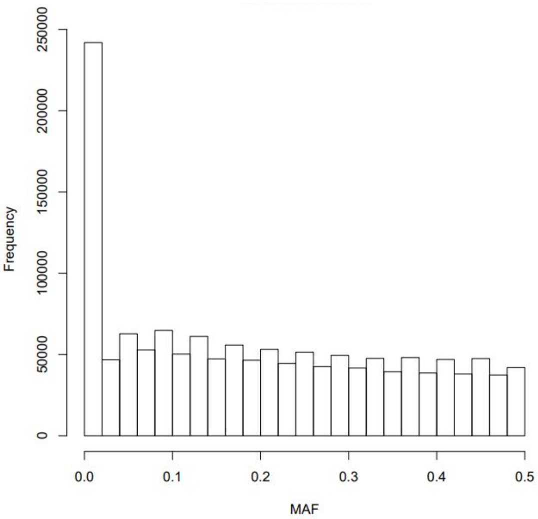

Fig. 3. MAF distribution. SNPs with a low MAF are rare, therefore power is lacking for detecting SNP-phenotype associations. These SNPs are also more prone to genotyping errors.

Remove SNPs with a low MAF frequency:

```
plink --bfile HapMap_3_r3_7 --maf 0.05 --make-bed --out HapMap_3_r3_8
```

1073226 SNPs are left

### 4. Hardy–Weinberg equilibrium (HWE)

The Hardy–Weinberg (dis)equilibrium (HWE) law concerns the relation between the allele and genotype frequencies. It assumes an indefinitely large population, with no selection, mutation, or migration. The law states that the genotype and the allele frequencies are constant over generations. Violation of the HWE law indicates that genotype frequencies are significantly different from expectations (e.g., if the frequency of allele

A = 0.20 and the frequency of allele T = 0.80; the expected frequency of genotype AT is 2*0.2*0.8 = 0.32) and the observed frequency should not be significantly different. In GWAS, it is generally assumed that deviations from HWE are the result of genotyping errors. The HWE thresholds in cases are often less stringent than those in controls, as the violation of the HWE law in cases can be indicative of true genetic association with disease risk.

Command `--hwe` excludes markers which deviate from Hardy–Weinberg equilibrium. For binary traits, it is proposed to exclude: the value of HWE p-value <1e-10 in cases and <1e-6 in the control. A less strict case threshold avoids exclusion of disease-related SNPs in selection. For quantitative traits, a p HWE value <1e-6 is recommended.

Check the distribution of HWE p-values of all SNPs:

```
plink --bfile HapMap_3_r3_8 --hardy
```

Selecting SNPs with HWE p-value below 0.00001, required for one of the two plots generated by the next Rscript, allows zooming in on strongly deviating SNPs (Fig. 4).

```
awk '{ if ($9 <0.00001) print $0 }' plink.hwe > plinkzoomhwe.hwe
Rscript hwe.R
```

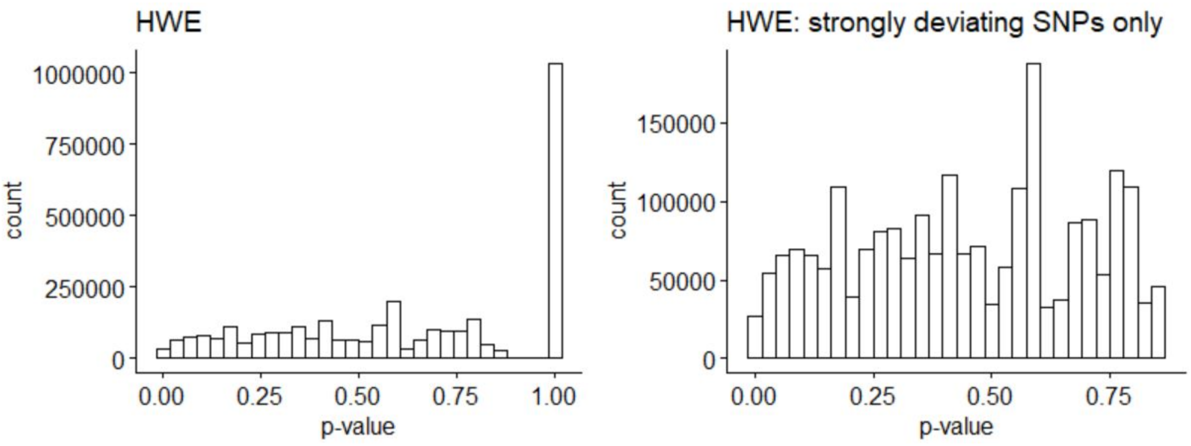

Fig. 4. The distribution of HWE p-values of all SNPs and strongly deviating SNPs.

By default the `--hwe` option in plink only filters for controls.

Therefore, we use two steps, first we use a stringent HWE threshold for controls, followed by a less stringent threshold for the case data:

```
plink --bfile HapMap_3_r3_8 --hwe 1e-6 --make-bed --out HapMap_hwe_filter_step1
```

The HWE threshold for the cases filters out only SNPs which deviate extremely from HWE. This second HWE step only focuses on cases because in the controls all SNPs with a HWE p-value < hwe 1e-6 were already removed:

```
plink --bfile HapMap_hwe_filter_step1 --hwe 1e-10 --hwe-all --make-bed --out HapMap_3_r3_9
```

## 5. Heterozygosity

Heterozygosity -- this is the carrying of two different alleles of a specific SNP. The heterozygosity rate of an individual is the proportion of heterozygous genotypes. High levels of

heterozygosity within an individual might be an indication of low sample quality whereas low levels of heterozygosity may be due to inbreeding.
We will exclude individuals with high or low heterozygosity rates. It is recommended to remove individuals who deviate ±3 SD from the samples' heterozygosity rate mean.
Generate a plot of the distribution of the heterozygosity rate of your subjects.
Checks for heterozygosity are performed on a set of SNPs which are not highly correlated.
Therefore, to generate a list of non-(highly)correlated SNPs, we exclude high inversion regions (inversion.txt [High LD regions]) and prune the SNPs using the command `--indep-pairwise`.
The parameters 50 5 0.2 stand respectively for: the window size, the number of SNPs to shift the window at each step, and the multiple correlation coefficient for a SNP being regressed on all other SNPs simultaneously:

```
plink --bfile HapMap_3_r3_9 --exclude inversion.txt --range --indep-pairwise 50 5
0.2 --out indepSNP
plink --bfile HapMap_3_r3_9 --extract indepSNP.prune.in --het --out R_check
```

Plot of the heterozygosity rate distribution (Fig. 5):

```
Rscript check_heterozygosity_rate.R
```

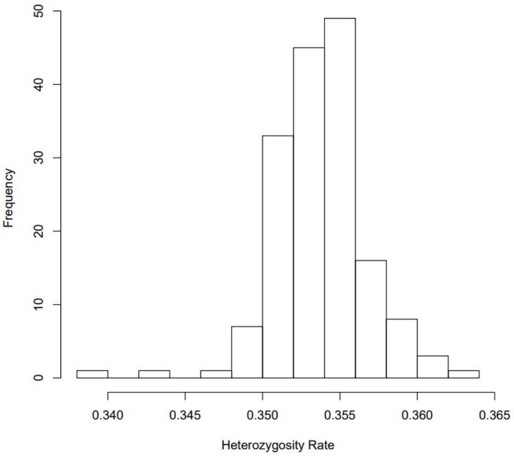

Fig. 5. Heterozygosity rate

The following code generates a list of individuals who deviate more than 3 standard deviations from the heterozygosity rate mean:

```
Rscript heterozygosity_outliers_list.R
```

The output of the command above: fail-het-qc.txt.
When using our example data/the HapMap data this list contains 2 individuals (i.e., two individuals have a heterozygosity rate deviating more than 3 SD's from the mean).
Adapt this file to make it compatible for PLINK, by removing all quotation marks from the file and selecting only the first two columns:

```
sed 's/"// g' fail-het-qc.txt | awk '{print$1, $2}'> het_fail_ind.txt
```

Remove heterozygosity rate outliers:

```
plink --bfile HapMap_3_r3_9 --remove het_fail_ind.txt --make-bed --out
HapMap_3_r3_10
```

### 6. Relatedness

Relatedness indicates how strongly a pair of individuals is genetically related. A conventional GWAS assumes that all subjects are unrelated (i.e., no pair of individuals is more closely related than second-degree relatives). Without appropriate correction, the inclusion of relatives could lead to biased estimations of standard errors of SNP effect sizes. Note that specific tools for analysing family data have been developed.
Here we will use two commands:
`--genome`: calculates identity by descent (IBD) of all sample pairs. We need to use independent SNPs (pruning) for this analysis and limit it to autosomal chromosomes only.

`--min`: sets the threshold and creates a list of individuals with relatedness above the chosen threshold. Meaning that subjects who are related at, for example, pi-hat >0.2 (i.e., second degree relatives) can be detected. Cryptic relatedness can interfere with the association analysis. If we have a family-based sample (e.g., parent-offspring), we do not need to remove related pairs but the statistical analysis should take family relatedness into account. However, for a population-based sample, it is suggested to use a pi-hat threshold of 0.2, which is in line with the literature [4].

Check for relationships between individuals with a pihat > 0.2:

```
plink --bfile HapMap_3_r3_10 --extract indepSNP.prune.in --genome --min 0.2 --out
pihat_min0.2
```

The HapMap dataset is known to contain parent-offspring relations.
The following commands will visualize specifically these parent-offspring relations, using the z values:

```
awk '{ if ($8 >0.9) print $0 }' pihat_min0.2.genome>zoom_pihat.genome
```

Generate a plot to assess the type of relationship (Fig. 6):

```
Rscript Relatedness.R
```

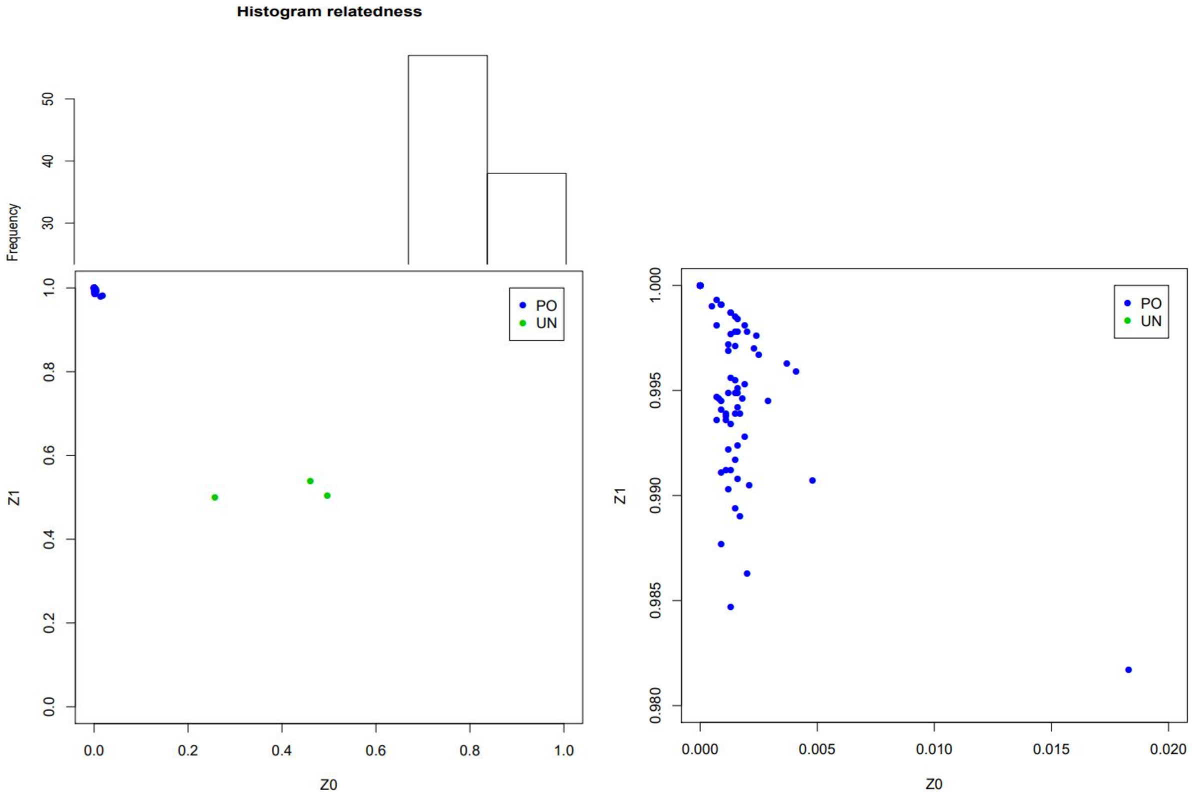

Fig. 6. Histogram of relatedness and parent-offspring relations (used the z values).

The generated plots show a considerable amount of related individuals (explentation plot; PO = parent-offspring, UN = unrelated individuals) in the Hapmap data, this is expected since the dataset was constructed as such. Normally, family based data should be analyzed using specific family based methods. To demonstrate that the majority of the relatedness was due to parent-offspring we only include founders (individuals without parents in the dataset):

```
plink --bfile HapMap_3_r3_10 --filter-founders --make-bed --out HapMap_3_r3_11
```

Now we will look again for individuals with a pihat >0.2:

```
plink --bfile HapMap_3_r3_11 --extract indepSNP.prune.in --genome --min 0.2 --out
pihat_min0.2_in_founders
```

The file 'pihat_min0.2_in_founders.genome' shows that, after exclusion of all non-founders, only 1 individual pair with a pihat greater than 0.2 remains in the HapMap data.
This is likely to be a full sib or DZ twin pair based on the Z values. Noteworthy, they were not given the same family identity (FID) in the HapMap data.

For each pair of 'related' individuals with a pihat > 0.2, it is recommended to remove the individual with the lowest call rate:

```
plink --bfile HapMap_3_r3_11 --missing
```

Use an UNIX text editor vim to check which individual has the highest call rate in the 'related pair'.

Generate a list of FID and IID of the individual(s) with a Pihat above 0.2, to check who had the lower call rate of the pair.
In our dataset the individual 13291  NA07045 had the lower call rate.

```
vi 0.2_low_call_rate_pihat.txt
i
13291  NA07045
ESC
:wq
```

In case of multiple 'related' pairs, the list generated above can be extended using the same method as for our lone 'related' pair.

Delete the individuals with the lowest call rate in 'related' pairs with a pihat > 0.2:

```
plink --bfile HapMap_3_r3_11 --remove 0.2_low_call_rate_pihat.txt --make-bed --out
HapMap_3_r3_12
```

### 7. Population stratification

Population stratification -- this is the presence of multiple subpopulations (e.g., individuals with different ethnic backgrounds) in a study. Because allele frequencies can differ between subpopulations, population stratification can lead to false positive associations and/or mask true associations. An excellent example of this is the chopstick gene, where a SNP, due to population stratification, accounted for nearly half of the variance in the capacity to eat with chopsticks [5]. There are several methods to correct for population stratification [6]. In this tutorial, we illustrate a method that is incorporated in PLINK: the multidimensional scaling (MDS) approach. This method calculates the genome-wide average proportion of alleles shared between any pair of individuals within the sample to generate quantitative indices (components) of the genetic variation for each individual. The individual component scores can be plotted to explore whether there are groups of individuals that are genetically more similar to each other than expected. For example, in a genetic study including subjects from Asia and Europe, MDS analysis would reveal that Asians are genetically more similar to each other than to Europeans. To investigate for which individuals the generated component scores deviate from the samples target population, plotting of the scores of the sample under investigation and a population of known ethnic structure (e.g., HapMap/1KG data) is helpful: this step is called anchoring. This enables the researcher to obtain ethnic information on their sample and to determine possible ethnic outliers. In this tutorial proposed to perform MDS on your own data anchored by data of the 1KG project (http://www.1000genomes.org/).

Download 1000 Genomes data, containing genetic data of 629 individuals from different ethnic backgrounds.
Due to the large file size and the lack of processing power of our computer, we are forced to skip this step. However, we can look at the results of this work taken from the article [7].

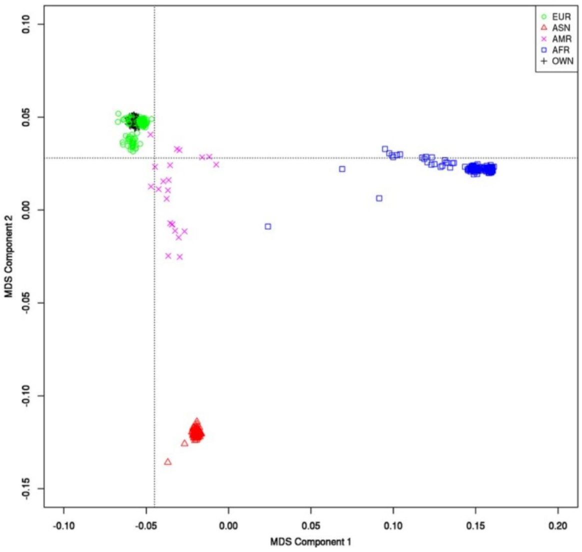

Fig. 7. Multidimensional scaling (MDS) plot of 1KG against the CEU of the HapMap data. From article [7]

This plot shows a multidimensional scaling (MDS) plot of 1KG against the CEU of the HapMap data. The black crosses (+ = “OWN”) in the upper left part represent the first two MDS components of the individuals in the HapMap sample (the coloured symbols represent the 1KG data (○ = European; □ = African; X = Ad Mixed American; △ = Asian). The MDS components representing the European samples (○) are located in the upper left, the African samples (□) are located in the upper right, the Ad Mixed American samples (X) are located near the intersection point of the dashed lines, the Asian components (△) are located in the lower left part.

This figure illustrates an example of such an analysis. Individuals who are outliers based on the MDS analysis should be removed from further analyses. After the exclusion of these individuals, a new MDS analysis must be conducted, and its main components need to be used as covariates in the association tests in order to correct for any remaining population stratification within the population. How many components need to be included depends on the population structure and the sample size, but the inclusion of up to 10 components is generally accepted within the psychiatric genetics community.

We should perform an MDS only on data without ethnic outliers. But as we said above, we cannot analyze the stratification of the population, besides, we do not see significant outliers in our graph. Therefore, below we will use the data from the previous step HapMap_3_r3_12.

```
plink --bfile HapMap_3_r3_13 --extract indepSNP.prune.in --genome --out HapMap_3_r3_13
plink --bfile HapMap_3_r3_13 --read-genome HapMap_3_r3_13.genome --cluster --mds-plot 10 --out HapMap_3_r3_13_mds
```

Change the format of the .mds file into a plink covariate file:

```
awk '{print$1, $2, $4, $5, $6, $7, $8, $9, $10, $11, $12, $13}' HapMap_3_r3_13_mds.mds > covar_mds.txt
```

The values in covar_mds.txt will be used as covariates, to adjust for remaining population stratification, onward, where we will perform genome-wide association analysis.

**Statistical tests of associations**

Now our data is ready for associative tests. Within PLINK, the association between SNPs and a binary outcome can be tested with the options `--assoc` or `--logistic`. The `--assoc` option in PLINK performs a $X^2$ test of association that does not allow the inclusion of covariates. With the `--logistic` option, a logistic regression analysis will be performed which allows the inclusion of covariates.

1. **--assoc and --logistic tests**

```
plink --bfile HapMap_3_r3_13 --assoc --out assoc_results

plink --bfile HapMap_3_r3_13 --covar covar_mds.txt --logistic --hide-covar --out logistic_results
```

Remove NA values, those might give problems generating plots in later steps:

```
awk '!/'NA'/' logistic_results.assoc.logistic > logistic_results.assoc_2.logistic
```

2. **Multiple testing**

Modern genotyping arrays can genotype up to 4 million markers concurrently, which generates a large number of tests, and thus, a considerable multiple testing burden. Three widely applied alternatives for determining genome-wide significance are the use of Bonferroni correction, Benjamini–Hochberg false discovery rate (FDR), and permutation testing. The Bonferroni correction is often too conservative and leads to an increase in the proportion of false negative findings because many SNPs are correlated, due to Linkage Disequilibrium (LD) and are thus by definition not independent.

FDR controls the expected proportion of false positives among all signals with an FDR value below a fixed threshold, assuming that SNPs are independent [8]. This method is less conservative than Bonferroni correction. To easily apply Bonferroni and FDR correction, PLINK offers the option `--adjust` that generates output in which the unadjusted p-value is displayed, along with p values corrected with various multiple testing correction methods.

Finally, permutation methods can be used to deal with the multiple testing burden. To calculate permutation-based p values, the outcome measure labels are randomly permuted multiple (e.g., 1,000–1,000,000) times which effectively removes any true association between the outcome measure and the genotype. For all permuted data sets, statistical tests are then performed. This provides the empirical distribution of the test-statistic and the p values under the null hypothesis of no association. The original test statistic or p-value obtained from the observed data is subsequently compared to the empirical distribution of p values to determine an empirically adjusted p-value. To use this method, the two PLINK options `--assoc` and `--mperm` can be combined to generate two p values: EMP1, the empirical p-value (uncorrected), and EMP2, the empirical p-value corrected for multiple testing. This procedure is computationally intensive, especially if many permutations are required, which is necessary to calculate very small p values accurately [9].

***Adjust***

```
plink --bfile HapMap_3_r3_13 -assoc --adjust --out adjusted_assoc_results
```

This file gives a Bonferroni corrected p-value, along with FDR and others.

***Permutation***

To reduce the computational time we only perform this test on a subset of the SNPs from chromosome 22. The EMP2 column provides for multiple testing corrected p-value.

Generate subset of SNPs:

```
awk '{ if ($4 >= 21595000 && $4 <= 21605000) print $2 }' HapMap_3_r3_13.bim > subset_snp_chr_22.txt
```

Filter your bfile based on the subset of SNPs generated in the step above:

```
plink --bfile HapMap_3_r3_13 --extract subset_snp_chr_22.txt --make-bed --out HapMap_subset_for_perm
```

Perform 1000000 permutations:

```
plink --bfile HapMap_subset_for_perm --assoc --mperm 1000000 --out subset_1M_perm_result
```

Order your data, from lowest to highest p-value:

```
sort -gk 4 subset_1M_perm_result.assoc.mperm > sorted_subset.txt
```

Check ordered permutation results:

```
head sorted_subset.txt
```

### 3. Generate Manhattan and QQ plots

```
Rscript --no-save Manhattan_plot.R
```

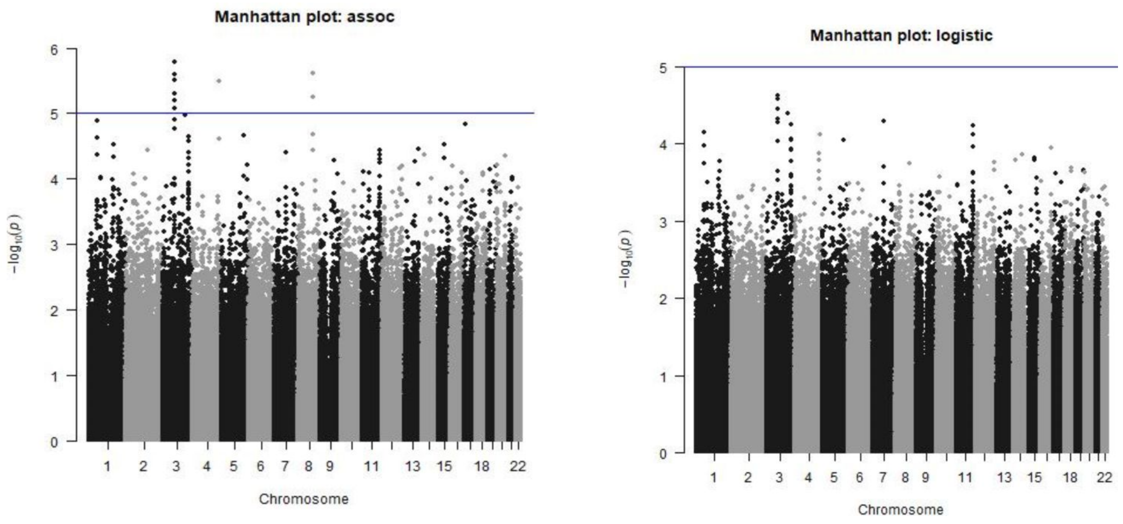

Fig. 8. Manhatten plots for two ways of analysis: assos ( $X^2$ test of association) and logistic (including covariates). Genomic coordinates are displayed along the X-axis, with the negative logarithm of the association p-value for each SNP displayed on the Y-axis, meaning that each dot on the plot signifies a SNP. Because the strongest associations have the smallest p-values , their negative logarithms are the greatest.

**Step 2. Comparative analysis of articles on GWAS studies of patients with COVID-19**

We compared the search results for gene variants associated with severe COVID-19, obtained in articles [10] and [11]. The results of these works can be seen on Fig. 9 and Fig. 10.

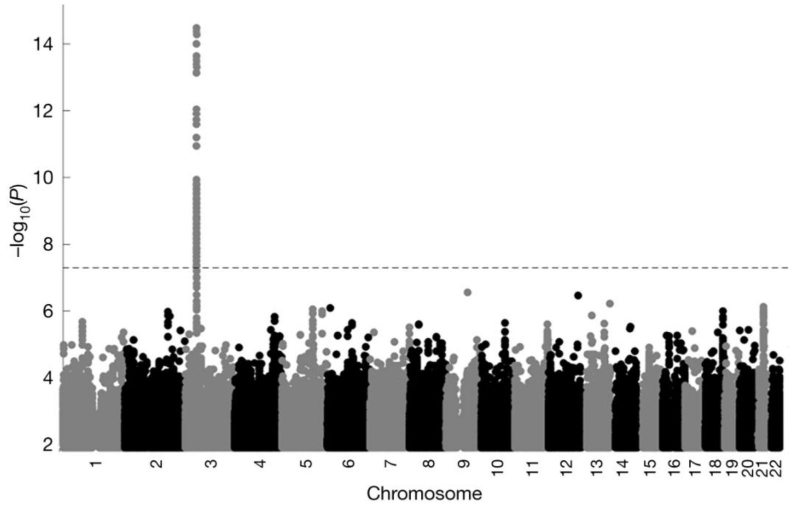

Fig. 9. A genome-wide study of associations of 3,199 hospitalized patients with COVID-19 and 897,488 controls. Genotyped data from COVID-19 Host Genetics Initiative (https://www.covid19hg.org/). From article [10].

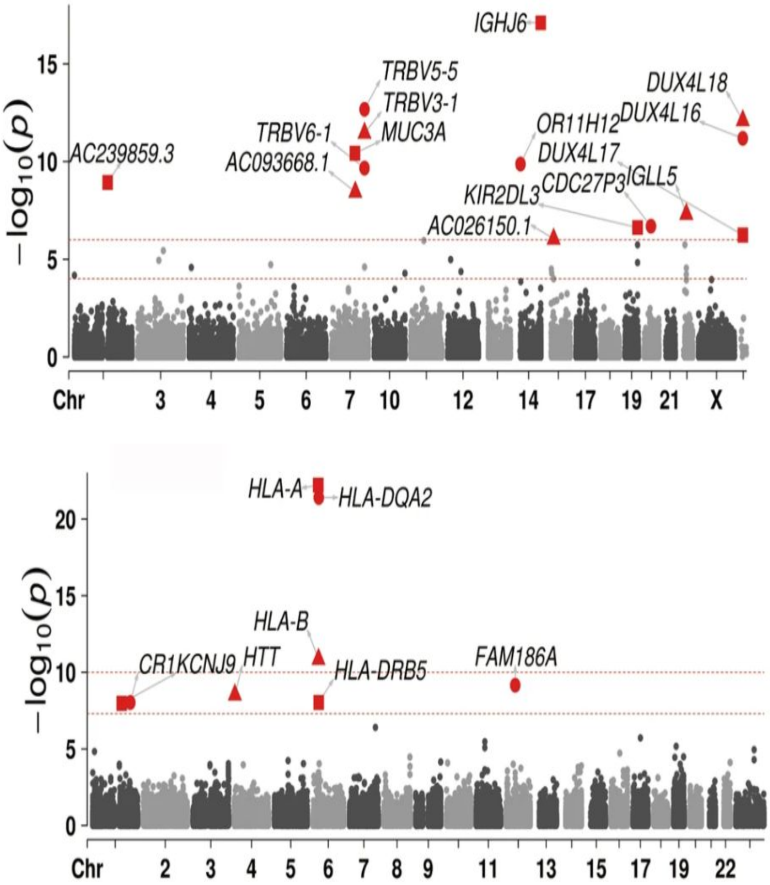

Fig. 10. Genome-wide study of associations of 284 hospitalized patients and 301 control individuals. Up: one-way association tests; down: associative tests based on gene interactions. Deep sequencing data from CNSA (China National Genebank Sequence Archive, https://db.cngb.org/cnsa/). From article [11].

The first work revealed an association of an extended region of the third chromosome, which is most likely inherited from the Neanderthals. Also, an association was found with a small region of chromosome 9 (including 1 gene), but later analysis did not confirm their relationship.

In the second work, GWAS was carried out both at the SNP level and at the level of alleles of individual genes. Two different groups were selected as control groups, depending on which the detected genes and SNPs significantly changed.

We believe that in the studies under consideration, the results differ for the following reasons:

1. Different sample sizes (3199/897488 and 284/301)

2. Different methods of obtaining SNP (genotyping using microarrays and deep sequencing)
3. Different composition of samples (in the first case, different ethnic groups, respectively, were applied stratification corrections, in the second - people of the same race).

**Conclusion**

This is where our introduction to plink and GWAS ends. We looked at the basic theory underlying GWAS analysis, as well as completed the main steps of this analysis. In addition, we analyzed the results obtained in various GWAS reseaches. We hope this knowledge will be useful to us in the future.

**References**

1. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., ... & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, *81*(3), 559-575.
2. Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F. L., Yang, H. M., ... & Tam, P. K. H. (2003). The international HapMap project.
3. Li, Y., Sheu, C. C., Ye, Y., De Andrade, M., Wang, L., Chang, S. C., ... & Cunningham, J. M. (2010). Genetic variants and risk of lung cancer in never smokers: a genome-wide association study. *The lancet oncology*, *11*(4), 321-330.
4. Anderson, C. A. , Pettersson, F. H. , Clarke, G. M. , Cardon, L. R. , Morris, A. P. , & Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. Nature Protocols, 5(9), 1564–1573. 10.1038/nprot.2010.116
5. Hamer D, Sirota L. Beware the chopsticks gene. Mol Psychiatry. 2000 Jan;5(1):11-3. doi: 10.1038/sj.mp.4000662. PMID: 10673763.
6. Price, A. L., Zaitlen, N. A., Reich, D., & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 459-463
7. Andries T. Marees, Hilde de Kluiver, et al. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research, 2018 Jun*; *27(2): e1608.*
8. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289-300.
9. North, B. V., Curtis, D., & Sham, P. C. (2003). A note on the calculation of empirical P values from Monte Carlo procedures. *The American Journal of Human Genetics*, 72(2), 498-499.
10. Zeberg, H., Pääbo, S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. *Nature* 587, 610–612 (2020). https://doi.org/10.1038/s41586-020-2818-3
11. Wang, F., Huang, S., Gao, R. et al. Initial whole-genome sequencing and analysis of the host genetic contribution to COVID-19 severity and susceptibility. Cell Discov 6, 83 (2020). https://doi.org/10.1038/s41421-020-00231-4