Genetic Data Analysis

NGS Methodologies Report

Aya BEN TAGHALINE, Joelle ASSY, Rayane ADAM

Introduction

Next-Generation Sequencing (NGS) has become an essential tool in genetic research, allowing for high-throughput analysis of genetic variations linked to various diseases. This report presents the results and interpretation of genetic data analyses for Clouston disease, a monogenic disorder, and rheumatoid arthritis, a complex multifactorial disease.

Clouston disease follows an autosomal dominant inheritance pattern. It's a monogenic rare disease characterized by the presence of a mendelian entity, which is a rare mutation variant with high disease risk. It is known to have complete penetrance, meaning that the presence of the mutation will lead to the disease. To locate the gene, there is linkage analysis based on the Lod score method that estimates recombinant distance between the disease variant and the marker. In this report, we performed genetic linkage analysis using the LOD score method to locate the disease-associated gene on chromosome 13. This analysis was carried out using the paramlink package in R, enabling the processing of familial data to calculate LOD scores for different recombination fractions (θ values). Additionally, to identify the causal variant, association studies can be performed, like the Transmission Disequilibrium Test (TDT), which is a trios family-based method. This test was conducted to explore potential associations between candidate variants and the disease within affected families.

In the case of Rheumatoid Arthritis, a multifactorial autoimmune disease with a complex genetic background. It can result from genetic and ennviromental factors. As a multifactorial disease, it is characterized by the presence of susceptibility genes, which are known to be frequent variants with low disease risk. It is also subject to gene-gene and gene-environment interactions, making the identification of causal factors difficult. Nevertheless, some common approaches to study these complex diseases are known. To identify causal variants, case-control population studies are conducted. To locate a susceptibility gene, sib-pairs method based on affected siblings is usually performed. In this report, we applied non-parametric linkage analysis using the sib-pairs method with the MERLIN tool to identify genomic regions potentially linked to the disease. A genome-wide association study (GWAS) was then performed using PLINK, involving a series of quality control steps, including checks for missing data, minor allele frequency (MAF), and Hardy-Weinberg equilibrium. Both allelic and genotypic association tests were conducted to identify significant SNPs associated with the disease.

The methodologies applied in this study provide insights into the genetic basis of these diseases, demonstrating the importance of accurate statistical and computational approaches in genetic data analysis using a variety of bioinformatics tools. The findings contribute to the understanding of genetic susceptibility and highlight the potential for further research with larger datasets and refined analytical methods.

Results and Interpretation

Clouston Disease

Starting with the analysis of the Clouston monogenic disease, genetic linkage and familial association analyses were performed.

Genetic Linkage Analysis: lodscore method

To locate the Clouston disease gene, linkage analysis was performed using the LOD score method by analyzing genetic markers located on chromosome 13. This analysis was conducted using the `paramlink` package in R.

The familial data set contains information on the family identifier, individual number, father's number, mother's number, sex, disease status, and genotypes of 13 markers (two columns per marker, with one allele per column). Data transformation was first performed to merge the two columns of alleles for each marker into a single column representing genotypes as `A1/A2`.

A total of one family, comprising 10 nuclear subfamilies, was studied. This family included 47 individuals, 22 of whom were affected and 25 non-affected. Among them, there were 11 founder individuals (without parents) and 13 individuals with unknown genotypes. The allele number distribution for the markers is summarized in Table 1. Markers with a high number of alleles are important because they increase the chance of having a double heterozygous parent, making the family informative for the analysis. An example family tree for marker 1 is shown in Figure 1.

Number of Alleles	Number of Markers
3	1
4	4
5	2
6	2
7	4

Table 1: Distribution of allele numbers across markers.

To perform the lod score method, which is parametric, the genetic model must be specified to attribute disease genotypes to individuals. An autosomal dominant mode of inheritance was considered, with a phenocopy value of 0.00001, complete penetrance, and a frequency of the deleterious allele for the disease gene of 10^{-5}. The analysis provides a lod score between the disease gene and each marker taken separately. The results are shown in Table 2, where the lod scores for each marker are provided for different θ values ranging from 0 to 0.5, with an increment of 0.05.

To better visualize the results, the lod score curve was plotted for marker 1, as shown in Figure 2. As observed, the maximum lod score occurs at $\theta = 0$, and the values of the lod score decrease as θ increases, reaching zero when $\theta = 0.5$.

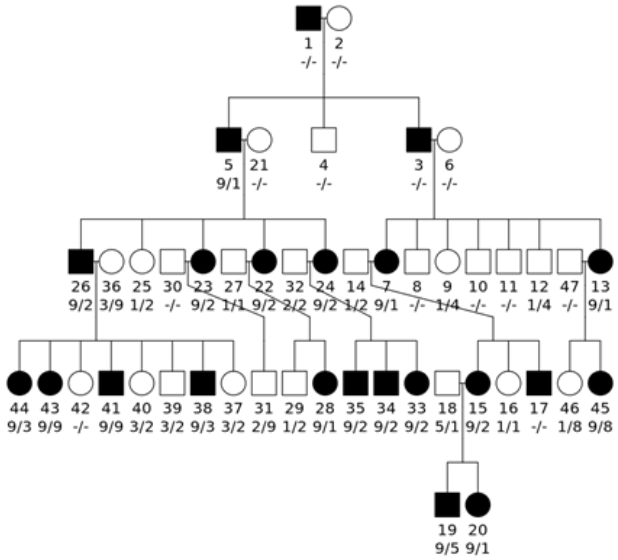

Figure 1: Example family tree for marker 1. Affected individuals are represented in black, while unaffected individuals are shown in white. Below each individual, their identifier number is provided, followed by the genotype for marker 1.

θ	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
0	7.673800	7.247411	4.762693	8.171219	6.0050443	4.969872	5.7141323	5.2550744	4.2537353	3.5643218	0.2874061	1.061945	-32.1732679
0.05	7.035453	6.653626	4.309984	7.511664	5.4606032	4.555357	5.1742276	4.7499732	3.8555848	3.2523734	3.5708778	4.285454	2.1590621
0.1	6.365581	6.030791	3.848484	6.819437	4.8898188	4.121439	4.6083512	4.2271963	3.4424404	2.9319108	3.4368453	4.088351	2.3859584
0.15	5.661047	5.376151	3.371107	6.091242	4.2903721	3.666446	4.0141596	3.6832892	3.0136955	2.6029704	3.1568696	3.741883	2.3427504
0.2	4.918254	4.686580	2.871736	5.323255	3.6599709	3.188760	3.3894121	3.1151478	2.5688442	2.2654917	2.8032430	3.318095	2.1757203
0.25	4.133297	3.958642	2.346428	4.511135	2.9968532	2.687328	2.7327781	2.5213329	2.1077616	1.9191995	2.3982856	2.839221	1.9301442
0.3	3.302875	3.189035	1.795610	3.650405	2.3012371	2.162844	2.0462455	1.9051123	1.6317013	1.5634218	1.9533074	2.316724	1.6262113
0.4	1.528959	1.530367	0.682811	1.782355	0.8643605	1.069246	0.6694062	0.6923811	0.6734622	0.8170073	0.9867355	1.184283	0.8847441
0.5	0.000000	0.000000	0.000000	0.000000	0.0000000	0.000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.000000	0.0000000

Table 2: Lod scores for different θ values and markers.

Given that the maximum lod score is greater than 3, we can reject the null hypothesis (H_0: no linkage, i.e, $\theta = 0.5$), indicating a strong genetic linkage between marker 1 and the disease locus. To estimate the confidence interval, the Zmax-1 line was also drawn in Figure 2. Since the recombination rate is between 0 and 0.5, the confidence interval should be within this range. The maximum lod score corresponds to $\theta = 0$, which implies that the lower bound of the confidence interval is 0, and the upper bound is determined by the point of intersection. Based on this, we approximate the confidence interval for marker 1 to be between $[0, 0.07]$.

Looking at the lod-score values for the other markers provided in Table 2, we observe a similar pattern for markers M_1 to M_{10}: the maximum lod score is obtained at $\theta = 0$, and the lod score value for each of these markers is greater

Figure 2: Lod score curve for marker 1, represented in blue, with the Zmax - 1 line shown in orange.

than 3. This allows us to reject the null hypothesis, indicating that the disease gene is very close to each of these markers, implying that markers M_1 to M_{10} form a cluster. For markers M_{11} and M_{12}, the maximum lod score is obtained at $\theta = 0.05$, and the value is higher than 3. This also leads to rejecting the null hypothesis and concluding that the probable location of the disease locus is approximately 0.05 cM from these markers. In contrast, for marker M_{13}, there is no value θ where the lod score exceeds 3. The maximum lod score is 2.38, obtained at $\theta = 0.1$, but since this value does not exceed 3, we cannot draw any definitive conclusion. Perhaps adding more families might provide clearer results. However, at $\theta = 0$, the lod score is -32, which is less than -2, meaning the null hypothesis is not rejected. Therefore, we can exclude the region very close to marker 13, both on the left and right, as it is an impossible location for the disease locus.

The analysis was conducted again with the determination of the maximum lod score, which automatically determines the value of θ that maximizes the lod score. This approach is useful when identifying the most likely genetic recombination rate (or the most likely position of the disease gene relative to the markers) based on the data, eliminating the need to test each value of θ manually. However, this can only be done for markers with up to 4 alleles. The method was applied to markers M_4, M_7, M_8, and M_{12}, and the results are provided in Table 3. As seen, for markers M_5, M_7, and M_8, the results are the same as before, with the maximum lod score obtained at $\theta = 0$. For marker M_{12}, the maximum lod score is now 4.288, slightly higher than the previous lod score of 4.285, which was obtained for $\theta = 0.05$. Now, the value of θ that maximizes the lod score is $\theta = 0.045$, providing more specificity. Based on this, the most probable distance of the disease locus from the marker M_{12} is around 0.045 cM.

The analysis was repeated by modifying the allele frequencies of marker M_5. By default, allele frequencies are assumed to be equifrequent. In this case, the frequencies were adjusted such that one allele had a significantly higher

Marker	M5	M7	M8	M12
LOD	6.005044	5.714132	5.255074	4.28823977
$\theta_{\max}$	0.000000	0.000000	0.000000	0.04508497

Table 3: Maximum LOD scores and corresponding θ values for markers M_5, M_7, M_8, and M_{12}.

frequency than the others. The results are shown in Table 4, comparing the new findings with the previous ones. Overall, the same pattern is observed, with LOD scores decreasing as θ values increase but with slightly higher LOD scores after modifying the allele frequencies and reaching 0 when $\theta = 0.5$. This slight difference can be attributed to the incorporation of allele frequencies into the calculation due to the presence of founders (individuals without parents) and individuals with missing genotypes in the dataset. For these cases, all possible genotypes must be considered, making the likelihood computation more complex. The probabilities of these genotypes are calculated using the allele frequencies, with Hardy-Weinberg equilibrium applied to estimate the expected genotype frequencies. If the results had been identical, it would suggest the absence of founders with missing genotypes, eliminating the need to account for different genotypes.

Θ	**Before**	**After**
0	6.0050443	6.1601384
0.05	5.4606032	5.6089781
0.1	4.8898188	5.0311016
0.15	4.2903721	4.4242603
0.2	3.6599709	3.7862270
0.25	2.9968532	3.1152044
0.3	2.3012371	2.4110015
0.4	0.8643605	0.9454197
0.5	0.0000000	0.0000000

Table 4: Comparison of LOD scores for marker M5 at various θ values before and after modifying allele frequencies.

The genetic model was adjusted to assess its impact on the analysis. With an autosomal recessive mode of inheritance, complete penetrance, a phenocopy rate of 10^{-5}, and a deleterious allele frequency of 10^{-5}, the results were recalculated. As shown in Table 5, markers M_1, M_2, M_4, M_5, M_6, M_7, M_8, M_{10}, M_{11}, M_{12}, and M_{13} displayed LOD scores less than -2 for $\theta = 0$, suggesting that the disease locus is not in close proximity to these markers. Markers M_3 and M_9 yielded positive LOD scores, but these fell within the range $-2 <$ LOD score < 3, preventing any definitive conclusions. This difference underscores the substantial influence of the disease model on linkage analysis. Shifting from an autosomal dominant to an autosomal recessive inheritance model changes the at-risk geno-

types and alters the disease genotype assignment for each individual, which in turn impacts the LOD score calculations. This emphasizes the importance of selecting an accurate disease model to ensure reliable results.

Marker	$\theta = 0$
M1	-11.52737
M2	-20.34523
M3	0.8394887
M4	-20.80169
M5	-11.46918
M6	-16.24291
M7	-6.483199
M8	-3.089972
M9	1.408298
M10	-14.51178
M11	-16.13149
M12	-8.403986
M13	-3.274896

Table 5: LOD scores at $\theta = 0$ for different markers after changing the genetic model to an autosomal recessive inheritance pattern.

Familial Association Analysis: TDT

Based on the previously identified linkage region, the `GJB6` gene appears to be a promising candidate for further investigation. Six SNPs located within the `GJB6` gene were genotyped in a sample of trios for familial association analysis using the Transmission Disequilibrium Test (TDT). The analysis was performed using the `fbat.exe` program. The dataset consisted of 652 nuclear families, with an average of 3 individuals per family.

The results are presented in Table 6. Results are reported only for SNPs with at least 10 informative families. As shown in the table, no results are available for SNP 1, indicating that it did not meet the threshold of having at least 10 informative families. An informative family is defined as a family in which at least one parent is heterozygous. The remaining SNPs are included in the results, as they each have at least 10 informative families, with the smallest number being 283.

The results from the table 6 indicate that SNP6 is significantly associated with the disease, as it is the only SNP with a p-value less than 0.05, thus confirming a significant association. Allele 2 is identified as the at-risk allele, as it is more frequently transmitted from a heterozygous parent to a diseased child. This is supported by the fact that the $S - E(S) > 0$, where S is the score used to test the association, $E(S)$ is the expected score under the null hypothesis of no association, and the sign of $S - E(S)$ indicates which allele is more or less often transmitted from heterozygous parents to affected children.

Marker	Allele	afreq	fam#	S-E(S)	Var(S)	Z	P
SNP2	1	0.636	409	3.500	138.750	0.297	0.766365
SNP2	2	0.364	409	-3.500	138.750	-0.297	0.766365
SNP3	1	0.370	402	2.000	140.500	0.169	0.866009
SNP3	2	0.630	402	-2.000	140.500	-0.169	0.866009
SNP4	1	0.403	425	5.000	148.500	0.410	0.681582
SNP4	2	0.597	425	-5.000	148.500	-0.410	0.681582
SNP5	1	0.626	393	-4.500	136.750	-0.385	0.700377
SNP5	2	0.374	393	4.500	136.750	0.385	0.700377
SNP6	1	0.212	283	-52.000	91.000	-5.451	5.01e-008
SNP6	2	0.788	283	52.000	91.000	5.451	5.01e-008

Table 6: Familial Association Analysis Results for SNPs in the GJB6 Gene

Based on these results, it is possible that SNP6 is either the causal variant or is in linkage disequilibrium with the causal variant.

To verify which case, knowing that SNP n°6 corresponds to SNP rs76179836, linkage disequilibrium between SNP rs76179836 and nearby SNPs was searched on the Ensembl website. The plot in Figure 3 showed r^2 values between the SNP of interest and surrounding variants. As observed, r^2 values are below the threshold for all nearby markers, indicating that none of the markers is in linkage disequilibrium with SNP rs76179836. Therefore, it can be concluded that SNP6 (rs76179836) is the causal variant.

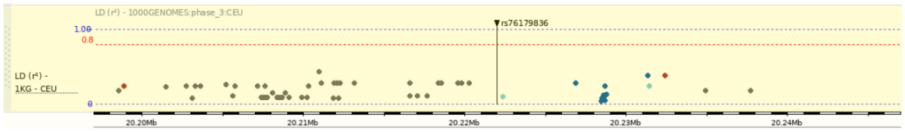

Figure 3: Linkage disequilibrium plot between SNP rs76179836 and surrounding SNPs.

Rheumatoid arthritis

Moving to the multifactorial Rheumatoid Arthritis disease, a non-parametric linkage analysis and a genome wide association study have been conducted.

Genetic Linkage Analysis: sib-pairs method

Carrying a Non-Parametric Linkage (NPL) study, we might be able to locate susceptibility genes for Rheumatoid Arthritis (RA). It's an equivalent to the lod-score method and is conducted using the command-line tool `MERLIN` starting from familial data containing family structure, marker information and chromosomal location.

There is 89 nuclear families consisting of 572 individuals, 453 of them are affected. The markers used for this analysis are microsatellites located on 23 chromosomes, there is 1089 with a heterozygosity range between 14.7% and 99.6%.

Performing the analysis we get curve maps for each chromosome, where we have lod-scores on the y-axis and position in cM from the end of the short arm of the chromosome on the x-axis. In here the null hypothesis is no linkage:

$$H_0 : locus\ is\ not\ linked\ to\ a\ disease - susceptibility\ gene$$

We reject H_0 for the same thresholds as the lod-score method, thus whenever $lod-sore \geq 3$. From the resulting map, only one chromosome show markers that have surpasses this threshold, denoted as "chromosome 999" which signifies the chromosome X. In this chromosome only one marker rejects the null hypothesis with a lod-score value of '3.17' at position 179 cM from the end of the short arm (value retrieved from the generated results in text file). This can be seen in Figure 4.

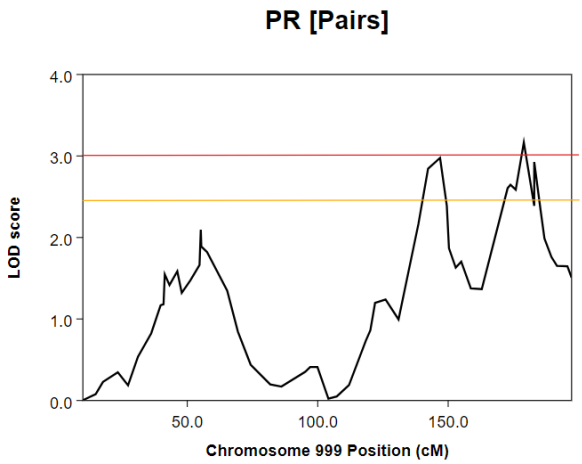

Figure 4: Lod score plot of different markers on chromosome X. Red line is the threshold 3 to reject the null hypothesis, and the yellow one is threshold 2.5 to indicate a suggestive linkage

While there seem to be a value very close to 3, the value is in fact '2.98' at around 147 cM, as can be concluded from the output file. Conducting another study with more data might result in a lod-score value rejecting H_0

This marker rejecting H_0 means that there seem to be linkage with the disease gene with the polymorphic marker at 179 cM. This means that the marker locus is at proximal distance of a linked Rhematoid Arthritis susceptibility gene.

Furthermore, lod score values that are between 2.5 and 3 belong to regions suggestive of RA. In the same chromosome X, there are markers in this zone (between red and yellow lines in Figure 4) in 143-150 cM region and within the 172-175 cM region. Similalry, chromosome 6 also present markers with lod-scores $\in [2.5, 3]$ as seen in Figure 5. Thus 50-53 cM region of chromsome 6 is also sugegstive of RA.

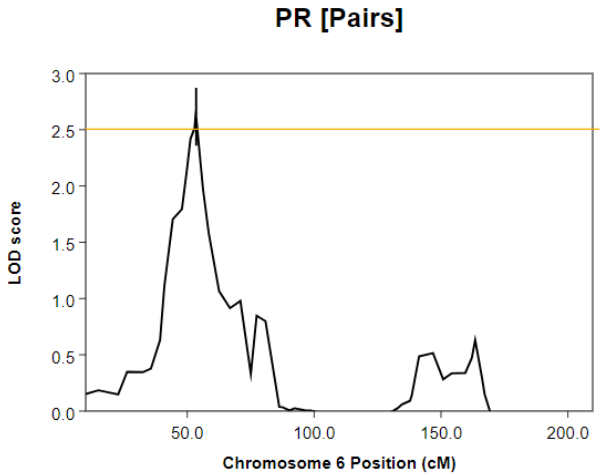

Figure 5: Lod score plot of different markers on chromosome 6.

Overall, RA is a multifactorial disease having many susceptibility genes involved in the disease. Therefore, different loci can suggest linkage with the disease-susceptibility genes since they are numerous, unlike mendelian entities. The results present:

- 1 region suggestive of RA on chromosome 6, hence there might be one or several disease genes around it at 50 cM from the end of the chromosome 6 short arm
- 1 region on chromosome X with a locus linked to RA at around 170-176 cM and many suggestive loci around it
- 1 region suggestive of RA on chromosome X near to the one mentioned before with many suggestive markers at around 146-150 cM

The 2 regions on the sex chromosome can suggest linked RA genes at the intersection of both loci (between 150-175 cM) or/and possibly RA genes at the regions inferior to 150 cM or superior to 175 cM.

Genome Wide Association Analysis

Since RA is a complex disease, a genome-wide association study can be conducted to find associated variants from the whole genome. This will be done on

command-line using the `PLINK` software.
For this part of the analysis, the starting data consist of 89 individuals and 50,971 variants. Unlike linkage analysis, the preferable variants here are SNPs. It's important to start with quality assessment before the statistical analysis. These steps narrow down the data to 89 individuals (no change) and 34,384 variants.
The following QC steps have been performed:

- Removing individuals with more than 5% missing data (0 removed)
- Removing SNPs with more than 5% missing data (1509 removed)
- Removing SNPs with Minor Allele Frequency less than 5% (14169 removed)
- Removing variants that are not in Hardy-Weinberg equilibrium in the control sample (945 removed)

However, one of the most important quality checks is to test for homogeneity for neutral variants which is done using PCA has not been performed here. This can largely affect the results: if data is not homogeneous then there might be different sub-populations leading to population stratification and false positive associations. The data in question is coming from Chinese and Japanese subpopulation, and while they are of the same East-Asian ancestry, they are 2 different ethnic groups and might present some genetic differences.

Additionally, the population consists of only 89 individuals, which is a relatively small number of participants for such a large scale analysis at the genome-wide level. For instance, with 34,384 variants, there will be 34,384 association tests conducted for a low number of participants. This adds up to the lack of homogeneity quality test as a limitation that might result in biased or wrong association results.

GWAS will be perfomed using both allelic and genotypic tests with χ^2 and Cochram-Armitage trend tests respectively.
In here the null hypothesis $H_0 : no\ association$. It's worth noting with 34,384 statistical tests, a Multiple Testing Correction approach must be used in order not to fall for multiple testing errors, for that Bonferroni correction (FWER method) will be used.
With a type I error rate $\alpha = 0.05$, the Bonferroni adjusted p-value threshold where $n = no.\ of\ tests = 34,384$ will be:

$$\alpha_{Bonf} = \frac{\alpha}{n}$$

$$\alpha_{Bonf} = \frac{0.05}{34,384} = 1.45 \times 10^{-6}$$

Starting with an allelic association test, the flag `--assoc` in 'PLINK' indicates a χ^2 statistical test to be performed according to `PLINK`'s documentation.

Results contain p-values for each variant, variants with $p-values \leq 1.45 \times 10^{-6}$ are considered associated with the disease.
Using unix commands to extract association results, 0 variants had $p-value \leq 1.45 \times 10^{-6}$ thus none rejected the null hypothesis (no association after performing the Bonferroni correction).

On the other hand `2287` SNPs showed association with RA where no correction has been performed (comparing with unadjusted 0.05 threshold).

Now performing a genotypic test with `--logistic` flag produces results of the Cochram-Armitage trend test. The same corrected Bonferroni p-value threshold will be taken since the same number of tests will be held (34,348 as validated from the output file results count).
Similarly to the allelic association test, no variants are associated with RA after Bonferroni correction.

Concerning results prior to Bonferroni correction, with unadjusted 0.05 level of significance, 2032 RA associate variants have been found.

Table 7: Results of Association Tests

Association Test	**Rejected**	**Rejected after correction**
Allelic	2287	0
Genotypic	2032	0

This is a manhattan plot visualization of the results done using the `qqman` R package from `PLINK` resulting files `res_allelic.assoc` and `res_geno.assoc.logistic`. As a validation to the results, no variants have been marked above the adjusted p-value threshold (figure 6)

This can be due to several reasons:

- Bonferroni is too conservative and the number of tests is large, it is known to produce a low number of discoveries to avoid false ones. To increase the number might consider changing the method to FDR Benjamini-Hochberg for instance to produce more potential associated variants.
- Missing data, low sample number (89) is not enough to make powerful statistical tests and conclude associations. Moreover, variants are only belonging to the first 10 chromosomes, some important loci that are indeed associated might belong to the other 13 chromosomes. Data is also biased due potential population stratification that hasn't been checked with PCA during quality control.

As the reasons for corrected tests being null have been mentioned, it's worth mentioning that between the allelic and genotypic tests there is small difference in the results. There is slightly more variants rejected in the allelic test (257 more) then the genotypic one. To understand why, it's important to clarify the differences between the 2 tests:

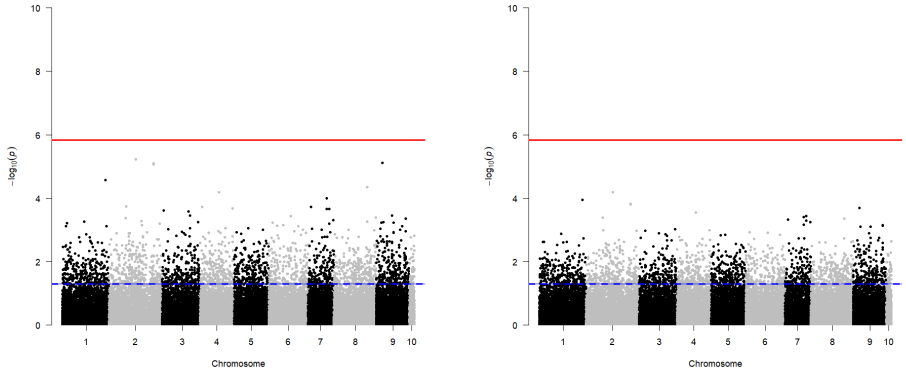

Figure 6: Manhattan plots of allelic and genotypic association tests respectively. The red line shows the log_{10} adjusted p-value threshold, and the blue line is the log_{10} unadjusted p-value threshold

The allelic association test compares the frequency of the alleles between cases and controls. We have 2 allelic frequencies (for alleles 1 and 2) and the statistical test would be checking the risk of allele 2 in comparison to allele 1.

The genotypic association test compares the frequency of the genotypes between cases and controls. For this test, we have 3 genotypes in the additive model that has been used here: 0, 1, 2 for 11, 12 and 22 respectively. So we have 3 genotypic frequencies between cases and controls and 2 statistical tests will be done to compare genotype 12 to genotype 11 and genotype 22 to genotype 11.

This being said, we would need twice as many tests as in the allelic test to compare all genotypes. However, the number of tests is the same as can be seen in the results. In the genotypic test where association test was done through a logistic model, the test statistic is solely comparing the heterozygous genotype 12 to the homozygous genotype 11. And genotype 12 is intermediate between the 2 homozygous 11 and 22 genotypes. For instance if 12 is 2 times more likely to develop the disease than 11, then 22 can be 4 times more likely to develop the disease than 11. As a result, there can be cases where 12 is not significantly associated with RA but 22 can be at risk of developing it which we can only know if we conduct a statistical test comparing 22 to 11. It might particularly impact the cases where 2 is an at risk recessive allele because 12 would not be associated with the disease but 22 would be an at risk genotype. To conclude, this explains why less variants have been rejected in the genotypic test than in the allelic test, to avoid missing some associations, making the double number of tests necessary to compare all genotypes together.

Conclusion

In this study, genetic linkage and association analyses were conducted to investigate the underlying genetic factors of two distinct diseases, Clouston and Rheumatoid Arthritis.

In the context of the monogenic disease Clouston, the analysis highlighted the utility of the LOD score method in identifying candidate loci. The parametric nature of the LOD score, influenced by factors such as the genetic model of the disease and allelic frequencies, underscores the importance of accurately specifying these parameters to obtain reliable results, as seen in this study. Through genetic linkage analysis, the disease locus was determined to be in close proximity to Markers 1 to 10, which likely form a cluster, approximately 0.05 from Markers 11 and 12, while a region near Marker 13 was excluded. This approach identified the *GJB6* gene as a strong candidate, prompting the genotyping of 6 SNPs for association analysis. The association study identified SNP6, corresponding to *rs76179836*, as a variant associated with the disease, with allele 2 being the at-risk allele. Verification using the linkage disequilibrium plot from Ensembl confirmed that SNP6 can be considered the causal variant, as it is not in linkage disequilibrium with nearby SNPs.

For the complex disease Rheumatoid Arthritis, many susceptibility genes can be associated with the disease. The non-parametric linkage analysis suggests loci in chromosome 6 (50-53 cM) and 23 (143-150 and 172-176 cM) linked to disease susceptibility genes. From this information, positional cloning can be performed on markers from around the 50-53 cM region in chromosome 6 and around 143-176 cM in the sex chromosome to gain more insights about associated genes in these regions. However, with the available data, GWAS was conducted on a small Chinese-Japanese sample, testing for RA association in variants belonging to the first 10 chromosomes. Due to insufficient data that did not get checked to the important quality step of homogeneity, and after performing Bonferroni correction, no variants showed association. Throughout the analysis, a difference has been perceived between genotypic and allelic tests, as the genotypic would require twice as many tests as the allelic to conclude a comprehensive list of associated markers.

In summary, this analysis has demonstrated the utility of genetic approaches in understanding disease etiology, underscoring the importance of accurate methodology, comprehensive data, and robust quality control to ensure reliable results in genetic research.

Code Availability

Kindly find the associated code with this report in a jupyter notebook comprising the R code and bash commands used along with a documentation of the steps taken and answers for each question: code file link