université
PARIS-SACLAY

Genetic Data Analysis of a Monogenic & a Multifactorial Disease

Module: M1 Méthodologie de traitement de données NGS

Written by

Christos Mitsakopoulos 20245433
Raul Duran De Alba 20245534

Introduction

Working on a precompiled dataset of two different diseases, namely Clouston's syndrome (Hidrotic ectodermal dysplasia, HED) and Rheumatoid arthritis, the following work attempts to perform a set of genetic linkage as well as association analyses. Both have a strong hereditary background, though both conditions vary greatly in their genetic architecture. Clouston's syndrome is monogenic in nature, caused by mutations in a single gene with a classic Mendelian pattern, whereas Rheumatoid arthritis is multifactorial, resulting from an interaction of multiple genes and environmental factors.

Because of these differences, our analytic strategies were adapted to each disease's specifications. We calculated the lod-score for the monogenic form of Clouston's in order to narrow down the most likely region harboring the causative mutation. This classic linkage approach is well-suited to Mendelian disorders, where tracking a single mutation through a pedigree can clearly reveal inheritance patterns. In the case of Rheumatoid arthritis, we used sib-pair analysis using family-based data to identify the genotype landscape of the families in our dataset. This was followed up by a genome-wide association test to look for causal variants directly (single nucleotide polymorphisms, SNPs).

Clouston's Syndrome: A Monogenic Disease

Genetic Linkage Analysis: LOD-score

To identify the disease-causing gene, we performed parametric linkage analysis on markers from chromosome 13 using a LOD-score approach. Specifically, the LOD score is calculated as the log of the ratio between the likelihood of linkage (with a recombination fraction $\theta < 0.5$) and the likelihood of no linkage (where $\theta = 0.5$ under the null hypothesis) (Nyholt, 2000). The raw data for this analysis are stored in a text file, fam.txt, where each column corresponds to one of the following data points:

- V1: Family number (identifier)
- V2: Individual number
- V3: Father's number (0 = no father in the sample analysed)
- V4: Mother's number (0 = no mother in the sample analysed)
- V5: Sex (1=male; 2 = female)
- V6: Disease status (1=not affected; 2 = affected; 0 = unknown)
- V7 onwards: Genotypes for the 13 markers (2 columns per marker; 1 allele per column) where 0 0 is an unknown genotype

To analyse the linkage data the R package paramlink (Vigeland et al., 2022) is used on the R Studio platform. After uploading fam.txt to a readable data frame, the first five rows of the table are observed to get a visual idea of the data's shape. The table below provides an example of pedigree data in linkage format. In the fourth row, for instance, the individual with ID 5 (from column V2) has an unknown genotype for marker 1 (as indicated by columns V7 and V8). To facilitate analysis with the paramlink package, this file must be converted into a linkdat format by using its homonym function. The file consists of the pedigree of 1 family tree of 45 individuals with 10 different sub-nuclear families and a total of 22 affected and 23 unaffected individuals. Of these individuals, 11

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	1	1	0	1	2	0	0	0	0	0
2	1	2	0	0	2	1	0	0	0	0
3	1	21	0	0	2	1	0	0	0	0
4	1	5	1	2	1	2	9	1	5	10
5	1	4	1	2	1	1	0	0	0	0

Table 1: Data from Columns V1 to V10 on Linkage format

have no parents, also known as founders, and 14 have an unknown genotype. This can also be observed in the pedigree chart on marker 1, Figure 1. The 13 markers analysed are composed of a different number of alleles, and due to software limitations on the R package, only those of 4 alleles or less can the maximum LOD-score be easily obtained.

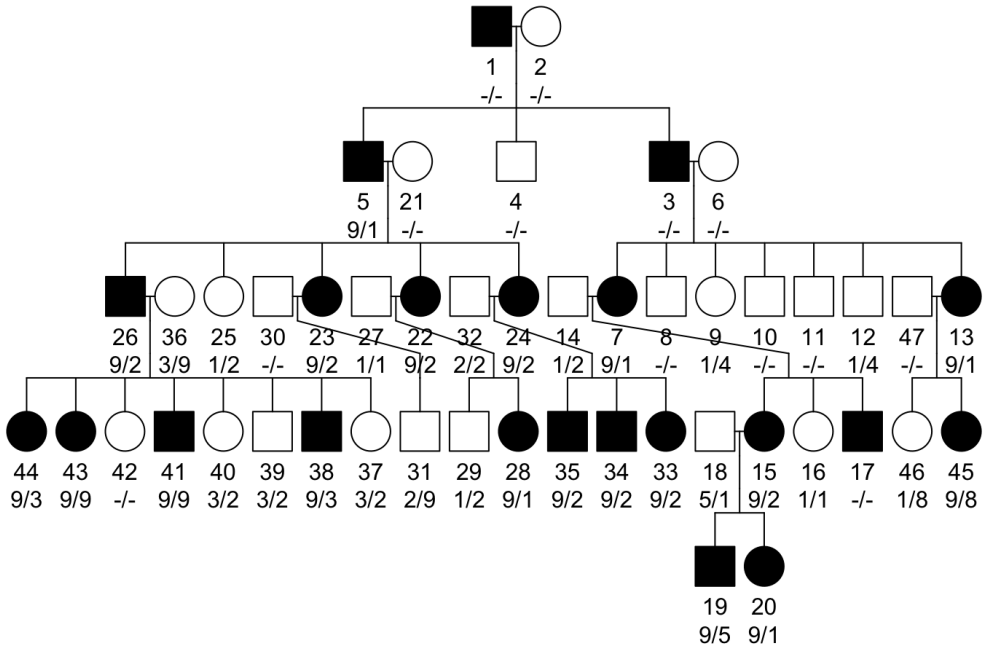

Figure 1: Pedigree Chart for Genetic Marker 1. Females are marked by a circle and males by a square, the colour white marks unaffected individuals while black marks those with a phenotypic expression of the disease. Each individual has their identifying number and their genotype coding for marker 1.

Afterward, the disease model must be input to properly analyse the data using the setModel function. Clouston's disease is known to be an autosomal dominant disease with full penetrance, as per a traditional Mendelian unit. the model also needs to include the probability of the different penetrances and the deleterious allele for the disease gene. A new variable stores the model and the previously formatted data. The model parameter defines the mode of inheritance and an autosomal dominant disease is number 1 out of 4 options: 1. Autosomal Dominant, 2. Autosomal Recessive, 3. X-linked Dominant and 4. X-linked Recessive, all with a frequency of deleterious alleles, dfraq, of 1×10^{-5} and full penetrance by default. In our disease model, we modified the penetrance of the affected DD allele to be 0.00001, where the deleterious allele is d and the unaffected allele is D, while maintaining that of the rest of the alleles, affected Dd and affected dd, at 1, thus

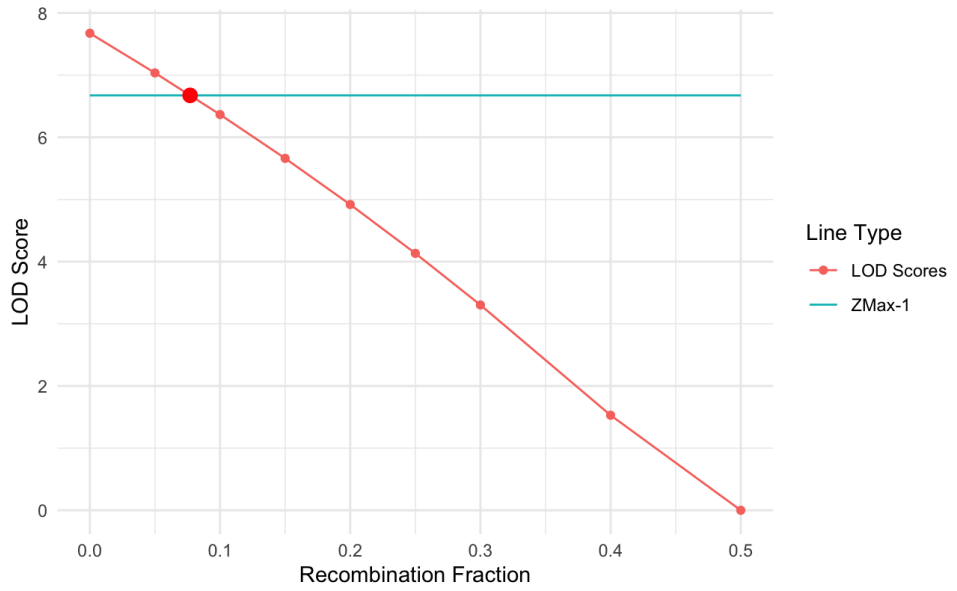

Figure 2: LOD Curve of Marker 1 intersecting with the $Z_{max} - 1$ line. Made with the GGplot package for R

considering the existence of phenocopies but maintaining full penetrance. By default, to be fully penetrant means no phenocopy and complete penetrance. These values can be modified with the use of the penetrance parameter on the function itself.

The lod function is used to obtain the LOD-score at specified θ values. Therefore, we calculate the LOD-score from 0 to 0.5, in a stepwise manner. We used the following function:

```
> lod(xdom, theta=c(0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4,
   0.5))
```

The result of this function is observed in the next table:

θ	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
0.00	7.67	7.25	4.76	8.17	6.01	4.97	5.71	5.26	4.25	3.56	0.29	1.06	−32.17
0.05	7.04	6.65	4.31	7.51	5.46	4.56	5.17	4.75	3.86	3.25	3.57	4.29	−2.15
0.10	6.37	6.03	3.85	6.82	4.89	4.12	4.61	4.23	3.44	2.93	3.44	4.09	2.39
0.15	5.68	5.36	3.37	6.10	4.28	3.68	4.03	3.72	3.01	2.61	3.31	3.74	2.34
0.20	4.92	4.69	2.87	5.32	3.64	3.19	3.39	3.12	2.57	2.27	3.17	3.46	2.18
0.25	4.13	3.96	2.35	4.51	2.90	2.69	2.73	2.52	2.11	1.92	3.00	3.14	1.83
0.30	3.30	3.19	1.80	3.65	2.13	2.19	2.04	1.89	1.64	1.57	2.83	2.78	1.36
0.40	1.53	1.53	0.68	1.78	0.86	1.07	0.67	0.90	0.67	0.82	0.99	1.18	0.80
0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 2: Table of LOD-score at different θ values

We can use this data to estimate the maximum LOD-score of each marker by graphing them. The maximum LOD-score would tell us if the marker is significantly linked to the disease by having a $Z(\theta_{max}) \geq 3$ or possibly correlated if it is between 2.5 and 3. Also, it is possible to obtain the most likely regions around the recombination fraction where the disease gene will be located, and the confidence interval surrounding the most likely location obtained by the maximum LOD-score. For example, marker 1 has a maximum

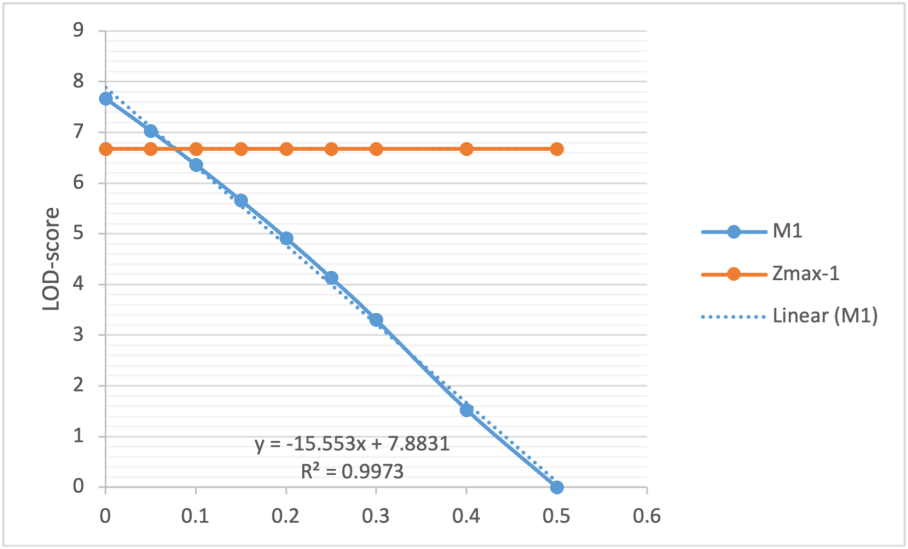

Figure 3: LOD Curve of Marker 1 intersecting with the $Z_{max} - 1$ line, and linear regression of the LOD Curve. Made with Excel software by Microsoft corporation.

LOD-score of 7.67 at $\theta = 0$ (see 2) which would mean that the most likely location of the disease gene is on the marker, with a confidence interval of 0.077 defined by the intersection of the LOD curve with $Z_{max} - 1$, obtained by linear interpolation on R. Due to the perceivable linearity of our LOD-score curve and given that there are no LOD-score below the -2 threshold, we can say that there is no "impossible" area in which the alleles and markers cannot coexist, this area would be obtained from the -2 threshold on the curve.

Talking advantage of the perceived linearity of the LOD curve, it is possible to create a linear regression out of the obtained LOD-scores values. This was done using the Excel software. The result is seen in the next Figure 3, where the linear model is defined as $y = -15.553x + 7.8831$ and a coefficient of determination of 0.9973, confirming a decent adjustment to the data points. In this instance the intersection between the model and the $Z_{max} - 1$ value of 6.67 is calculated to be 0.0777, which is equal to the value obtained by the parametric estimation obtained using the R software package.

As previously stated, the old function of the paramlink package (Vigeland et al., 2022) also allows us to obtain the maximum LOD-score and their corresponding recombination fraction for markers up to 4 alleles which we only have 5 in our dataset, 1 composed of 3 alleles and 4 composed of 4 alleles, shown on the Table 3.

	M5	**M7**	**M8**	**M9**	**M12**
LOD	6.005044	5.741132	5.255074	4.253735	4.288297
t_max	0.000000	0.000000	0.000000	0.000000	0.045049

Table 3: Max θ and LOD-score values of markers 5, 7, 8, 9, and 12

In the above results, we observe that markers 5, 6, 8, and 9 obtain their Z-max at $\theta = 0$, whilst marker 12 at a $\theta \approx 0.05$. Particularly, we could see these markers as having a descending order of proximity to the allele, with marker 12 being the farthest away compared to the rest.

If we use the "MAX" function in Excel on data from Table 2 the max LOD-scores for

the other markers are, from one to thirteen respectively: 7.673800, 7.247411, 4.762693, 8.171219, 6.0050443, 4.969872, 5.7141323, 5.2550744, 4.2537353, 3.5643218, 3.5708778 ($\theta = 0.1$), 4.285454 ($\theta = 0.1$), 2.3859584 ($\theta = 0.15$).

Markers 1 to 10 are in remarkable proximity to the disease gene, given they all obtain their Z_{max} at $\theta = 0$.

The ModifyMarker function allows for adjustments to marker allele frequencies. By default, the frequencies of a marker's alleles are taken to be equi-frequent. If we for example were to modify the allele frequencies of marker 5, where alleles 1 to 3 have a frequency of 10% and allele 4 has a frequency of 70%. The code used is the following to create this modified model and obtain the corresponding LOD-scores of marker 5.

```
> xdom5=modifyMarker(xdom,marker = 5, afreq = c(0.1, 0.1, 0.1,
   0.7))
> lod(xdom5, marker=5, theta=c(0, 0.05, 0.1, 0.15, 0.2, 0.25,
   0.3, 0.4, 0.5))
```

θ	**Same frequency**	**Modified frequency**
0	6.005044318	6.160138
0.05	5.460603228	5.608978
0.1	4.889818834	5.031102
0.15	4.290372115	4.42426
0.2	3.659970869	3.786227
0.25	2.996853174	3.115204
0.3	2.30123708	2.411001
0.4	0.864360545	0.9454197
0.5	0	0

Table 4: The θ and LOD-score for marker 5 with equi-frequent alleles compared with LOD-score of the modified allele frequencies model.

Following modification of the allele frequencies, we observe higher LOD-scores which are still contained within a descending order; indicative of a nearer localization of marker 5 to the disease gene but it's understood that an error in the marker allele frequencies leads to a risk of false positives.

We restarted the genetic linkage analyses using the first dataset obtained directly from the fam.txt file, assuming an autosomal recessive mode of transmission, a phenocopy rate of 10^{-5}, complete penetrance, and a deleterious allele frequency of 10^{-5} for the disease gene; to reaffirm the use of the autosomal dominant genetic model.

```
> xrec=setModel(x, model=2, penetrances=c(0.00001,0.00001, 1),
   dfreq=0.00001)
> lod(xrec)
```

As LOD-score linkage analysis is completely model dependent, changing it to an autosomal recessive model of disease will completely alter the results obtained, for example, the highest LOD-score of the original model was localized at marker 4 at a value of 8.171219 at $\theta = 0$ defining a consistent genetic location for the disease gene in chromosome 13 near or at marker 4, while in the new model is located at the marker 9 with a value of 1.408298 at $\theta = 0$ which no longer significantly supports the linkage alternative hypothesis, see Table 5.

Marker	LOD-score
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 $\theta = 0$ of all the markers of an autosomal receive model of the disease.

By looking at the $\theta = 0$ of all the different markers we can gather that most of them have an absence of genetic linkage with the Clouston's disease gene as they have a LOD-score $Z(\theta = 0) \leq -2$ there is an absence of genetic linkage. This would be a false exclusion of the markers of the disease gene created by using the wrong genetic model.

Familial Association Analysis: TDT

After analyzing the previous linkage zone on chromosome 13, the gene GJB6 (JHU, 2022) appears to be a good candidate for the disease gene. SNPs located in the GJB6 gene were genotyped in a sample of trios in order to perform a Transmission Disequilibrium Test carried out using the fbat.exe program. The data is located in the fbat.ped file with the same formatting as the files used for linkage analysis. The FBAT program runs on the Windows Command Line. The file contains the pedigree data of 652 nuclear families and 2011 individuals, an average of 3 individuals per nuclear family, with 6 markers genotyped. The results for the SNPs with at least 10 informative families; where an informative family is such that has a present copy of the SNPs in one of the individuals conforming it, thus one of the parents is heterozygous and an affected child; are available in the following Table 6, where only SNP1 has an insufficient number of informative families.

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	4.500	140.500	0.379	0.704363
SNP3	2	0.630	402	-4.500	140.500	-0.379	0.704363
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	5.010e-08
SNP6	2	0.788	283	52.000	91.000	5.451	5.010e-08

Table 6: Genetic Analysis with FBAT for Clouston's Disease Markers

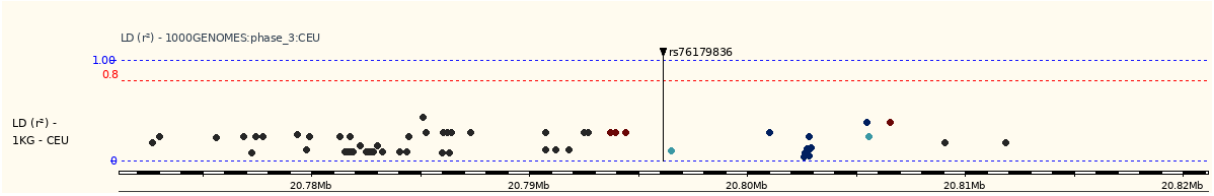

Figure 4: Linkage disequilibrium data (r^2 score) for the variant rs76179836 and surrounding variants in the 1000GENOMES phase_3 CEU population.

In Table 6, results for each SNP are displayed across two rows, with each row representing a different allele. The columns in the table are explained as follows:

- Afreq stands for allelic frequency
- Fam# denotes the number of informative families
- S represents the score utilized to assess the association
- E(S) refers to the expected score under the null hypothesis of no association, where the sign of $S - E(S)$ reveals which allele is typically more frequently transmitted from heterozygous parents to affected offspring
- V(S) is the score's variance
- Z is the test statistic calculated as $(S - E(S))/\sqrt{V(S)}$
- P is the p-value corresponding to the test ($p \leq 0.05$ indicates a rejection of H_0)

Upon examination of the table, it is evident that allele 2 of SNP 6 exhibits the most significant association with the disease. Acknowledging that SNP 6 corresponds to SNP rs76179836 (GRCh37, 2024), we conducted an analysis using the ENSEMBL Linkage Disequilibrium website, obtaining a chart for the European population as depicted in Figure 4. The genes in proximity predominantly represent regulatory region variants (shown in red), a 3' untranslated region (UTR) variant in light blue, which may impact mRNA stability, and intron variants in dark blue, potentially influencing splicing or regulatory motifs. However, none of these demonstrate a high degree of linkage disequilibrium (LD) to the reference variant, with the highest of the near variants having an $r^2 \approx 0.4$.

The SNP identified as rs76179836 of the GJB6 gene, appears to be the causal or threshold variant, as evidenced by LD values. It is preserved through recombination due to its proximity to the gene, thereby being inherited across generations. This allele has a global minor allele frequency (MAF) of 0.005391, which makes it a rare variant globally as it is less than 5% of the population and in the European population it has a MAF of 0.03 or 3%. The GJB6 gene is known to encode for connexin-30, on chromosome 13q12.11,

Rheumatoid Arthritis: A Multifactorial Disease

Genetic Linkage Analysis: Affected Sib-Pairs

To identify genomic regions that contained Rheumatoid Arthritis (RA) susceptibility genes, we used the MERLIN program for non-parametric genetic linkage analysis on a database of family pedigrees (sib-pair data). Note that, due to the disease being multifactorial, the use of a database with large families and adequate generation history, is vital to perform accurate sib-pair analysis. As such, before running the genetic linkage analysis, we also performed a preliminary analysis with 'pedstats', to verify the integrity of our database.

This is the command we used for pedstats, note the inclusion of '–ignoreMendelianErrors', aimed at preventing the inclusion of polluting-misinformative family data in the statistics:

```
pedstats -d fam.dat -p fam.ped --ignoreMendelianErrors >
   stats_fam_MK
```

- Number of families analyzed: 88
- Number of generations per family: 2 (100.0%); all families have two generations accounted for in the data.
- Average number of individuals per family: 6.50; Distr. 5 (20.5%), 6 (19.3%) and 4 (18.2%)
- Total number of individuals: 572
- Total number of affected individuals: 453
- Number of genetic markers studied: 1089

As you can see above the database not only contains large families, but also adequate generations per family, which are required to correctly infer genotypes from parents to siblings when possible.

Having run this check, we also looked further into the genetic markers themselves observing their heterozygosity levels which is another key element in ensuring genetic heterogeneity (quality assurance). Below, you will observe the highest and lowest heterozygosity genetic markers (respectively):

```
MARKER    RANK     PROP    N_GENO |          MARKER    RANK     PROP
   N_GENO
a055zg1/(AC)n       1    74.1%       424 | a102wf9/(AC)n    1089
   1.7%        10
```

With our preliminary quality assurance concluded, we commenced with the MERLIN based genetic linkage analysis. An assortment of relevant result files was produced, key amongst them being the "merlin.pdf" containing a set of lod score curves to represent genetic region association to RA. Out of all of the returned graphs, only one proved to be informational, where lodscores surpassed a threshold of 3 (See Figure 5).

As can be seen in Figure 5, we observe two distinct but closely positioned peaks that indicate a strong correlation of genomic region to RA; those regions (and peaks) are located at $\tilde{1}$48cM and $\tilde{1}$75cM. The identification of RA-associated genetic regions exclusively on the X-chromosome is particularly intriguing and goes against expectations. Recent studies

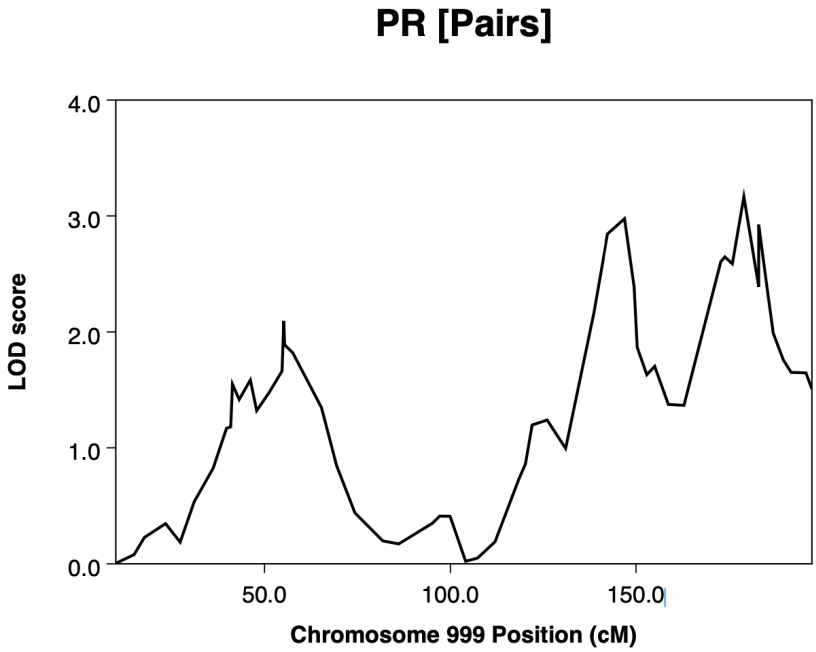

Figure 5: LOD score curve against genetic positioning of centiMorgan (cM) scale, on chromosome 999 (X-chromosome).

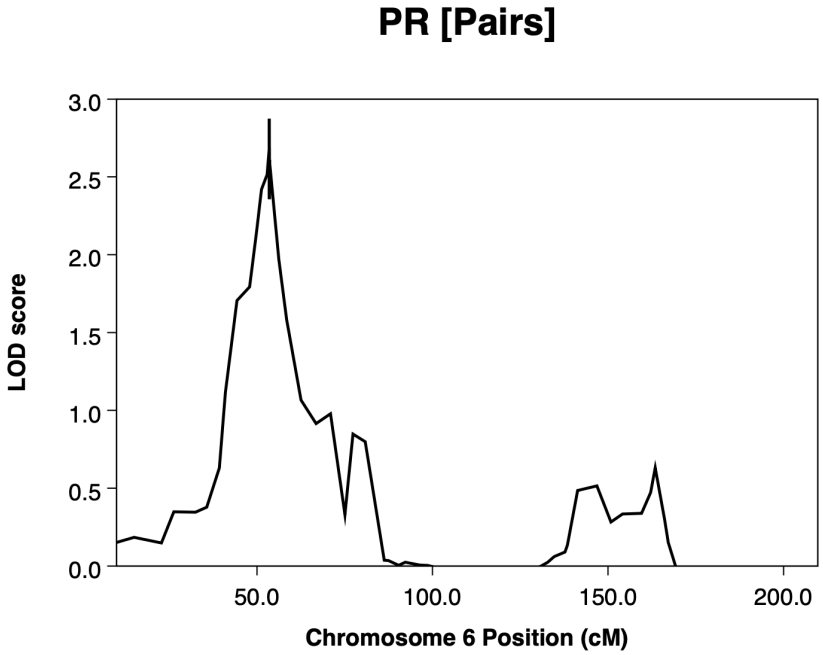

Figure 6: LOD score curve against genetic positioning of centiMorgan (cM) scale, on chromosome 6; the most confidently associated chromosome to autoimmune diseases such as Rheumatoid Arthritis (Kanaan et al., 2016).

on autoimmune diseases have linked their association to the Human Leucocyte Antigen (HLA) locus, located on chromosome 6 (Kanaan et al., 2016). By lowering the LOD-score threshold, we can see in our results that chromosome 6 also presents possibly associated regions, with lod scores (association) peaking at $\tilde{2}.8$ and 55cM (See Figure 6).

Genomic Wide Association Analysis

After ascertaining that there is strong evidence for the presence of susceptibility genes on chromosome X and evidence for chromosome 6, we proceeded with genome wide association studies. By doing so, we could distinguish the causal variants within the database, to investigate them further. The program we used for both quality assurance and the association study itself, was PLINK.

Just like previously, before proceeding with our main aim we performed quality assurance on our data. First step was to eliminate individuals of a call rate below 95%, for which none were removed by PLINK. Next, all SNPs that have not been confidently genotyped in more than 95% of calls. In this case, terminal return from PLINK contained (keep in mind for future reference):

```
50971 variants loaded from .bim file.
Total genotyping rate is 0.994134.
1509 variants removed due to missing genotype data (--geno).
49462 variants and 89 people pass filters and QC.
Among remaining phenotypes, 44 are cases and 45 are controls.
```

As seen above, 1509 variants failed to pass our filter and therefore were eliminated, no phenotypes were lost.

The next step was to filter for variants with a minor allele frequency (MAF) less than 5%, to eliminate variants of low disease relevance. Following filtering, we observed a considerable reduction in variants passing the filter:

```
35293 variants and 89 people pass filters and QC.
Among remaining phenotypes, 44 are cases and 45 are controls.
```

We now observe 35293 variants from an initial number of 50971. A final test to run was to test for Hardy-Weinberg (HW) equilibrium across the SNPs and their associated genotypes. While there is some literature on the option of using SNPs lost to the HW-test and maintaining statistical validity in the results (Fardo et al., 2009), we opted to use the test in order to follow protocol. As you will observe below, only SNPs observed in individuals with no parents were removed;

```
--hardy: Writing Hardy-Weinberg report (founders only) to
   IndMkMAFHWE.hwe ...
done.
--hwe: 945 variants removed due to Hardy-Weinberg exact test.
34348 variants and 89 people pass filters and QC.
```

With all quality control steps completed, we proceeded to identify SNPs associated to RA through allelic and genotype association analysis. As such, we could examine the impact of both allele and genotype frequencies on RA development in the families of our database, where the confidence interval was set to 95%. We also performed Bonferonni

correction following each analysis, which due to its strictness, observed no SNPs passing its filter. All of the discussed results from PLINK are as such:

```
Allelic association before bonferonni correction: 2290
Genotypic association before bonferonni correction: 2033
```

To disprove the null hypothesis under the Bonferonni principle, we would need to apply the following cut-off threshold formula: α/m. Where α is the standard significance p-value of 0.05 and m the number of hypotheses (Gao, 2011). Given we observe 2290 Allelic associations before correction, the Bonferonni cut-off would become 0.05/2290, or 2.18×10^{-5}. Equally, for genotypic association, we would observe a cut-off of 2.46×10^{-5}. As one would expect, a p-value returned which matches the strictness of the Bonferonni correction is very rare and in our case, no SNPs were able to pass the filter:

```
Following Bonferonni correction: NONE
```

This means that while false positive allelic and genotypic associations have been eliminated from the results, the true positives have been lost in the process. In retrospect, one could apply the simpleM method which is believed to be powerful without discarding as many positives (Gao, 2011). Since that option has to be implemented in future experimentation, we decided to sort the allelic and genotypic association file contents, sorting at smallest p-value and extracting the smallest 5. An example command for this process is attached below (Unix command):

```
Allelic association results sorting:
(head -n 1 res_allelic && tail -n +2 res_allelic | sort -k9,9g |
   head -n 5)
Genotypic association results sorting:
(head -n 1 res_geno && tail -n +2 res_geno | awk -F '\t' '$12 !=
   "NA"' | sort -k12,12g | head -n 5)
```

In doing so, we would could get the SNPs of highest statistical significance and thereby possible causal association to RA, which would otherwise be ignored. The top 5 allelic and genotypic SNPs that we found are below:

Table 7: Allelic association: Table depicts the top 5 most statistically significant allelic based associations of SNPs to RA; all SNPs are sorted based on ascending order of p-value (smallest to highest).

CHR	SNP	BP	A1	F_A	F_U	A2	CHISQ	P	OR	SE	L9	U9
2	rs2222162	10602	1	0.2841	0.6222	2	20.51	5.918e-06	0.2409	0.3212	0.2319	0.2503
9	rs10810856	46335	1	0.2955	0.04444	2	20.01	7.723e-06	9.016	0.5623	8.431	9.642
2	rs4675607	13220	1	0.1628	0.4778	2	19.93	8.05e-06	0.2125	0.3603	0.2036	0.2219
2	rs1375352	13219	1	0.1818	0.5	2	19.83	8.485e-06	0.2222	0.3491	0.2132	0.2317
2	rs4673349	13218	1	0.1818	0.5	2	19.83	8.485e-06	0.2222	0.3491	0.2132	0.2317

Following some research on the web, we identified that rs4078404 is a contained within a database of non-coding mutants with a considerable "0.555" allele frequency ("Turkish Genome Project - Transcript ENST00000271971", 2025). On the other hand, rs1375352 appears within a database of SNPs related to major depressive disorder on NCBI. rs4675607 appears on a repository of variants related to ADHD, hosted by the UCLA (Ding et al., 2021). While we were able to locate multiple and interesting results from the variant identifiers, we did not find any pertaining directly to empirical annotation data on RA. This is not to say that we should disqualify them as irrelevant/non causal SNPs.

Table 8: Genotypic association: Table depicts the top 5 most statistically significant genotype based associations of SNPs to RA; all SNPs are sorted based on ascending order of p-value (smallest to highest).

CHR	SNP	BP	A1	TEST	NMISS	OR	SE	L9	U9	STAT	P
1	rs4078404	6200	2	ADD	89	4.663	0.3983	4.447	4.89	3.865	1.11×10^{-4}
2	rs4673349	13218	1	ADD	88	0.251	0.365	0.2403	0.2622	-3.787	1.53×10^{-4}
2	rs1375352	13219	1	ADD	88	0.251	0.365	0.2403	0.2622	-3.787	1.53×10^{-4}
2	rs4675607	13220	1	ADD	88	0.2417	0.3757	0.2311	0.2528	-3.779	1.57×10^{-4}
9	rs10810856	46335	1	ADD	89	8.937	0.5893	8.33	9.588	3.716	2.02×10^{-4}

Considering that we did not find any satisfactory results above and to make better use of our data from the affected sib-pairs analysis, we also tried to identify if there are any SNPs of "great" statistical correlation which are located on either chromosome 6 or X. Below are the SNPs located on chromosome 6 from the genotypic and allelic association; chromosome X was absent from these results.

Table 9: Table depicts the top 5 most statistically significant genotype based associations of SNPs to RA; all SNPs are sorted based on ascending order of p-value (smallest to highest).

CHR	SNP	BP	A1	TEST	NMISS	OR	SE	L9	U9	STAT	P
6	rs9464779	30394	1	ADD	89	0.3137	0.3576	0.3006	0.3274	-3.242	0.001188
6	rs4839919	33088	2	ADD	89	0.3372	0.3364	0.3239	0.351	-3.231	0.001232
6	rs2206517	34789	1	ADD	89	5.909	0.557	5.529	6.315	3.189	0.001427
6	rs2234079	31064	1	ADD	85	0.2359	0.4648	0.2232	0.2494	-3.107	0.001889
6	rs3823017	30256	2	ADD	89	2.928	0.3599	2.805	3.056	2.985	0.002839

Table 10: Table depicts the top 5 most statistically significant allele based associations of SNPs to RA; all SNPs are sorted based on ascending order of p-value (smallest to highest).

CHR	SNP	BP	A1	F_A	F_U	A2	CHISQ	P	OR	SE	L9	U9
6	rs4839919	33088	2	0,2614	0,5222	1	12,69	0,0003678	0,3237	0,3216	0,3115	0,3364
6	rs9464779	30394	1	0,25	0,5	2	11,85	0,0005774	0,3333	0,3241	0,3207	0,3465
6	rs13217899	31955	1	0	0,1222	2	11,46	0,0007096	0	inf	0	nan
6	rs2234079	31064	1	0,09524	0,3023	2	11,39	0,0007385	0,2429	0,4396	0,2305	0,256
6	rs12191259	34136	1	0,1705	0,02222	2	11,32	0,0007677	9.041	0,7692	8.248	9,91

Particularly, all SNPs depicted in tables 9 and 10 do not belong to the most confidently predicted to be causal to RA; such as the SNPs of tables 7 and 8.

Looking to improve on our analytical methodology, we also decided to perform a principal component analysis (PCA). To perform the PCA, we used PLINK to obtain a number of tagged eigenvectors and eigenvalues. Importantly, the eigenvectors were the ones used to create the graphical representation needed for the PCA. We ran the analysis on the filtered '.bed' and '.fam' files which we produced before the allelic and genotypic association analysis. Our aim was to identify if there is a factual reason for which Chinese and Japanese patients should be treated together in the same batch, or separately. The code we used to generate the graph is below:

```
library(ggplot2)
setwd("~/Desktop/TP_DATA_PROG/II.a.Plink")
df <- read.table("plink.eigenvec", header = FALSE,
   stringsAsFactors = FALSE)
num_pcs <- ncol(df) - 2
```

```
colnames(df) <- c("FID", "IID", paste0("PC", 1:num_pcs))
df$Population <- ifelse(grepl("^HCB", df$FID), "HCB",
                        ifelse(grepl("^JPT", df$FID), "JPT", NA))
pac_scatter <- ggplot(df, aes(x = PC1, y = PC2, color =
   Population)) +
  geom_point(size = 2, alpha = 0.7) +
  theme_minimal() +
  labs(
    title = "PCA: HCB vs. JPT",
    x = "PC1",
    y = "PC2"
  ) +
  theme(plot.title = element_text(hjust = 0.5))
ggsave("pca_scatter_plot.png", pca_scatter, width = 8, height =
   6, dpi = 300)
```

From this code we were able to produce Figure 6, which demonstrates clear sepratation between ethnic backgrounds and therefore serves as compelling evidence to deem the use of Chinese and Japanese samples together as misguided.

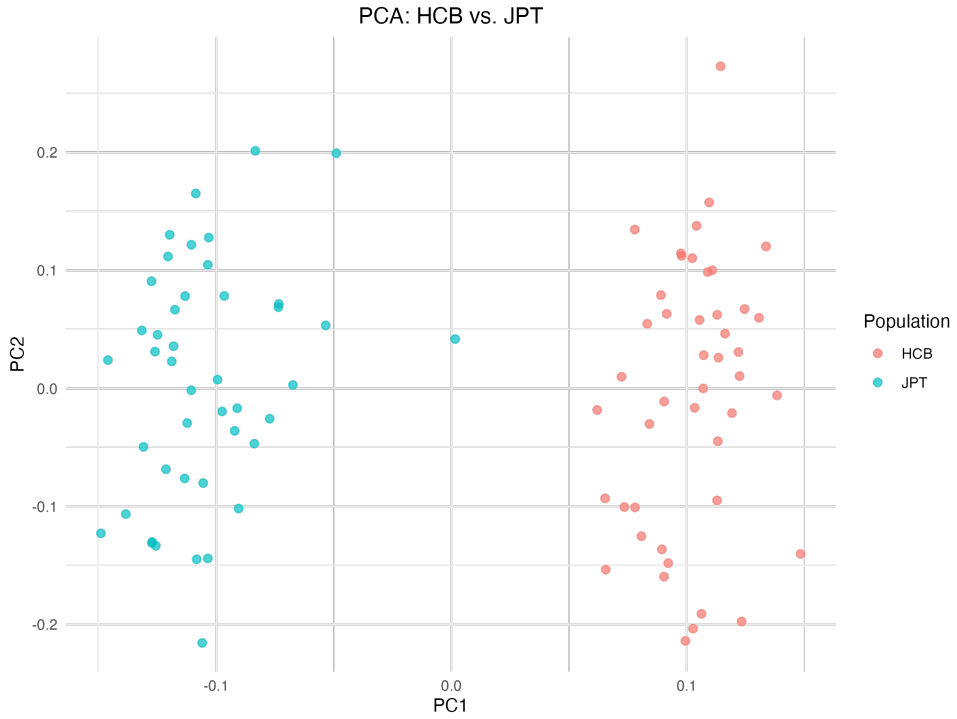

Figure 7: Scatter plot depicting the spread of SNPs for Chinese (HCB) and Japanese (JPT) individuals, following PCA analysis on filtered SNP data.

Conclusion

In this study, we employed targeted genetic analysis strategies for two diseases with significant genetic differences: the monogenic Clouston's syndrome and the multifactorial Rheumatoid arthritis. For Clouston's syndrome, the LOD-score method based on an autosomal dominant genetic model successfully identified a high-confidence linkage peak on chromosome 13, which aligns well with the monogenic inheritance pattern. Further fine-mapping of this region and the Transmission Disequilibrium Test (TDT) of the GJB6 gene demonstrated that rs76179836 is the variant most strongly associated with the disease and is a strong candidate for the causative mutation. This series of analyses highlights the relative simplicity and certainty of genetic analysis for monogenic diseases, as the causal gene can be accurately located through traditional linkage analysis methods.

In contrast, the genetic analysis of Rheumatoid arthritis presented a more complex scenario. The affected sib-pair linkage analysis using MERLIN and the Genome-Wide Association Study (GWAS) using PLINK revealed the intricate genetic pattern of this disease. Although linkage peaks were detected on the X chromosome in the sib-pair analysis and suggestive signals were identified in the HLA region of chromosome 6, no significant association loci were found across the genome after strict Bonferroni correction. This suggests that the genetic architecture of Rheumatoid arthritis involves multiple low-effect risk variants scattered throughout the genome, requiring larger sample sizes and refined statistical methods for accurate identification. Additionally, the results of the Principal Component Analysis (PCA) revealed clear genetic differences between samples from different ethnic groups (such as the Chinese and Japanese populations), underscoring the importance of considering population stratification in research.

Overall, this study not only successfully identified the causative gene locus of Clouston's syndrome but also offered valuable guidance and methodological references for the genetic study of Rheumatoid arthritis. Meanwhile, we acknowledge that for the genetic analysis of complex diseases, increasing sample size, optimizing statistical methods, and thoroughly considering population stratification are critical directions for future research. Further studies can better integrate multi-omics data and combine functional experiments to explore the pathogenesis of these complex diseases in depth, providing a stronger theoretical foundation for the early diagnosis, precise treatment, and prevention of diseases.

References

Ding, J., Blencowe, M., Nghiem, T., Ha, S.-m., Chen, Y.-W., Li, G., & Yang, X. (2021). Mergeomics 2.0: A web server for multi-omics data integration to elucidate disease networks and predict therapeutics. *Nucleic Acids Research*, *49*(W1), W375–W387. https://doi.org/10.1093/nar/gkab405

Fardo, D. W., Becker, K. D., Bertram, L., Tanzi, R. E., & Lange, C. (2009). Recovering unused information in genome-wide association studies: The benefit of analyzing snps out of hardy–weinberg equilibrium. *European Journal of Human Genetics*, *17*(12), 1676–1682.

Gao, X. (2011). Multiple testing corrections for imputed snps. *Genetic Epidemiology*, *35*(3), 154–158. https://doi.org/10.1002/gepi.20563

GRCh37, E. (2024, October). Rs76179836 (SNP).

JHU. (2022, May). 604418 GAP JUNCTION PROTEIN, BETA-6; GJB6.

Kanaan, S. B., Onat, O. E., Balandraud, N., Martin, G. V., Nelson, J. L., Azzouz, D. F., Auger, I., Arnoux, F., Martin, M., Roudier, J., Ozcelik, T., & Lambert, N. C. (2016). Evaluation of x chromosome inactivation with respect to hla genetic susceptibility in rheumatoid arthritis and systemic sclerosis (C. Feghali-Bostwick, Ed.). *PLOS ONE*, *11*(6), e0158550. https://doi.org/https://doi.org/10.1371/journal.pone.0158550

Nyholt, D. R. (2000). All LODs Are Not Created Equal. *American Journal of Human Genetics*, *67*(2), 282–288.

Turkish Genome Project - Transcript ENST00000271971. (2025).

Vigeland, M. D., Egeland, T., & Doerum, G. (2022, April). Paramlink: Parametric Linkage and Other Pedigree Analysis in R.