Genetic Data Analysis of Rheumatoid arthritis and Clouston's disease

Module: M1 Méthodologie de traitement de données NGS

Authors

Ngoc VU, Denys BURYI

Introduction

We have carried out linkage analyses (LOD score or affected sib-pairs) and association analyses (case-control or familial) for a monogenic disease, then for a multifactorial disease:

1. Study of a monogenic disease: Clouston's disease
   (a) Genetic linkage analysis (LOD score)
   (b) Familial association analysis
2. Study of a multifactorial disease: rheumatoid arthritis
   (a) Genetic linkage analysis (affected sib-pairs)
   (b) Genome-Wide case-control association analysis

The data and programs were provided by our professor VALÉRIE CHAUDRU.

1 Monogenic disease: Clouston's disease

Clouston syndrome is mainly characterized by abnormal nails (thickened, slow-growing, brittle) and alopecia (hair loss). *Palmoplantar hyperkeratosis* (commonly known as horniness) is also sometimes observed. This is a rare disease, with a prevalence ranging from 1 to 9 in 100,000. It is autosomal dominant. Penetrance of the disease is complete, but its expression is variable, even in patients from the same family. Our aim in this part of the analysis was to:

- Perform linkage analysis using the LOD score method to locate the Clouston disease gene using genetic markers located on chromosome 13.
- Interpret data from a `.vcf` file containing genetic variants identified in a case-control sample and perform association analysis for a candidate variant.
- Search for the causal genetic variant using familial association analysis.

1.1 Genetic linkage analysis - LOD score method

We have analyzed the data contained in the `fam.txt` file using `paramlink` package in R. First we performed some initial data exploration using `linkdat()` and `summary()` functions:

```
fam = read.table('fam.txt')
x = linkdat(fam)
summary(x)
```

We found that the file contains data of a large family of 47 individuals, comprising 22 affected and 25 unaffected members. These individuals are organized into 10 nuclear subfamilies and have been genotyped for 13 genetic markers on chromosome 13. For 11 individuals in the dataset no parent data was available (further referred to as "founders"). Figure 1 shows the pedigree plot of the family.

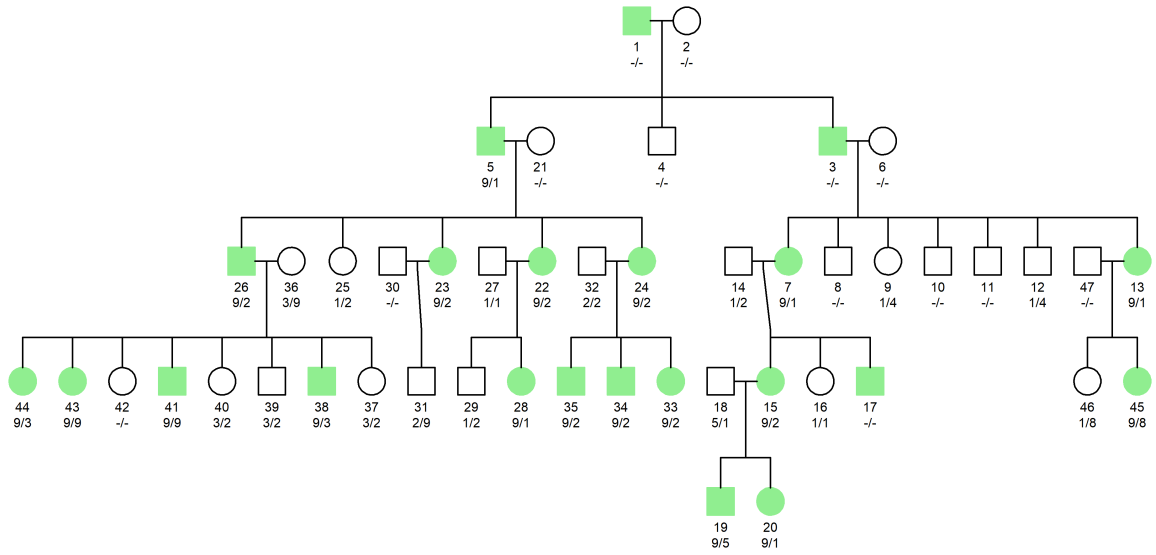

Figure 1: Pedigree plot of the family along with their affection status and genotypes with regards to marker 1. Affected - light green, Unaffected - white. Sex: square/circle - M/F. Genotype: alleles of the marker 1.

In addition to helping summarize the data, the `linkdat()` function also transforms the original data and outputs a `linkdat` object that can be used to perform linkage analysis. `linkdat` object is a list containing among others, the following important components:

- `pedigree` - data.frame with 5 columns (ID, FID, MID, SEX, AFF) describing the pedigree in linkage format.
- `orig.ids` - vector of original IDs of the individuals in the pedigree.
- `subnucs` - list of vectors of IDs of individuals in each nuclear family.
- `markerdata` - a list of `marker` objects describing information about the genetic markers.

- `model` - a list of parameters for the linkage analysis.

1.2 Mode of inheritance: Autosomal Dominant

First, we have analyzed our data assuming an autosomal dominant mode of inheritance. Using the `setmodel()` function, we have set the parameters to:

- phenocopies = 10^{-5}
- complete penetrance
- disease allele frequency = 10^{-5}

We then performed the linkage analysis using the `lod()` function for a range of θ from 0 to 0.5 with a step of 0.05. We have saved the results in the `result_dom` data frame for further analysis (see code below).

```
xdom = setModel(x, model=1, penetrances = c(0.00001, 1, 1), dfreq = 0.00001)
result_dom = lod(xdom, theta=c(0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5))
result__dom_df = as.data.frame(result_dom)
```

MARKER	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
θ\LOD	7.6738	7.247411	4.762693	8.171219	6.005044	4.969872	5.714132	5.255074	4.253735	3.564322	3.570878	4.285454	2.385958
0	7.6738	7.247411	4.762693	8.171219	6.005044	4.969872	5.714132	5.255074	4.253735	3.564322	0.287406	1.061945	-32.17327
0.05	7.035453	6.653626	4.309984	7.511664	5.460603	4.555357	5.174228	4.749973	3.855585	3.252373	3.570878	4.285454	2.159062
0.1	6.365581	6.030791	3.848484	6.819437	4.889819	4.121439	4.608351	4.227196	3.44244	2.931911	3.436845	4.088351	2.385958
0.15	5.661047	5.376151	3.371107	6.091242	4.290372	3.666446	4.01416	3.683289	3.013695	2.60297	3.15687	3.741883	2.34275
0.2	4.918254	4.68658	2.871736	5.323255	3.659971	3.18876	3.389412	3.115148	2.568844	2.265492	2.803243	3.318095	2.17572
0.25	4.133297	3.958642	2.346428	4.511135	2.996853	2.687328	2.732778	2.521333	2.107762	1.919199	2.398286	2.839221	1.930144
0.3	3.302875	3.189035	1.79561	3.650405	2.301237	2.162844	2.046245	1.905112	1.631701	1.563422	1.953307	2.316724	1.626211
0.4	1.528959	1.530367	0.682811	1.782355	0.864361	1.069246	0.669406	0.692381	0.673462	0.817007	0.986735	1.184283	0.884744
0.5	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 2: LOD score analysis results for autosomal dominant mode of inheritance. Each column shows LOD score for a given marker, for a given θ value. Top row shows maximum LOD score for each marker. Conditional coloring: green to red - higest to lowest LOD score.

When performing LOD score analysis, under the null hypothesis H_0 we assume that there is no linkage between the markers and the disease gene and $\theta = 0.5$. In other words if the two loci are unlinked, the probability of observing a recombinant event $c = 50\%$. Under an alternative hypothesis H_1, if the two loci are indeed linked, we expect the

MARKER	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
θ\LOD	7.6738	7.247411	4.762693	8.171219	6.005044	4.969872	5.714132	5.255074	4.253735	3.564322	3.570878	4.285454	2.385958
0	7.6738	7.247411	4.762693	8.171219	6.005044	4.969872	5.714132	5.255074	4.253735	3.564322	0.287406	1.061945	-32.17327
0.01	7.548517	7.130858	4.672514	8.041779	5.898157	4.888445	5.608121	5.155308	4.175274	3.502616	3.206201	3.96898	1.047643
0.02	7.422062	7.013221	4.582208	7.911126	5.790286	4.806293	5.501142	5.05494	4.096234	3.440568	3.42465	4.175555	1.579161
0.03	7.294414	6.894484	4.491721	7.779238	5.681415	4.723402	5.393179	4.953946	4.016609	3.378178	3.517565	4.25648	1.860084
0.04	7.165552	6.774625	4.400999	7.646092	5.571527	4.639761	5.284213	4.852299	3.936394	3.315446	3.558611	4.285417	2.037972
0.05	7.035453	6.653626	4.309984	7.511664	5.460603	4.555357	5.174228	4.749973	3.855585	3.252373	3.570878	4.285454	2.159062

Figure 3: LOD score analysis results for autosomal dominant mode of inheritance for more granular θ values. Each column shows LOD score for a given marker, for a given θ value. Top row shows maximum LOD score for each marker. Conditional coloring: green to red - highest to lowest LOD score.

probability of observing a recombinant event c to fall within $0 < c < 50\%$. LOD score is a logarithmic likelihood ratio:

$$Z(x) = \log_{10}\left[\frac{L(c = x)}{L(c = 0.5)}\right]$$

A LOD score of 3 indicates a 1,000:1 likelihood that two genes are linked and inherited together with a given recombination rate(e.g. $c{=}10\%$), compared to the likelihood under the null hypothesis. For all markers except M13, the maximum LOD score exceeds 3 (see Figure 2). These results suggest that the disease gene is likely in extremely close proximity to markers 1–10, and approximately 5 cM away from markers 11 and 12.

As for marker 13, it is highly probable that adding more families to the dataset would increase $Z(max)$ above 3. For now, the data suggest that the disease gene might be located approximately 15 cM from marker 13.

To validate these findings, we performed the same analysis using θ values ranging from 0 to 0.05 in increments of 0.01 (see Figure 3).

Next, we calculated the confidence intervals for each marker. Visualizing the LOD score curve can be helpful for this purpose, so we plotted it for marker 1 as a representative example for markers 210 (Figure 4). To determine the precise confidence intervals for the results as well as more precise Z_{max} and θ_{Zmax}, we ran the script located in the `ci_bounds.R` file. Results are shown in the Table 1.

1.3 Allele frequencies and LOD score

So far in our analysis we have assumed that marker allele frequencies(AF) in the general population are equal, which isn't always true. If one of the alleles has a higher AF, it will have implication for the LOD score. The reason changing AF influences the LOD score is because of the way genotypes are estimated for founders in the sample. If every allele is equally likely, the probability for each genotype (assuming 4 alleles) is:

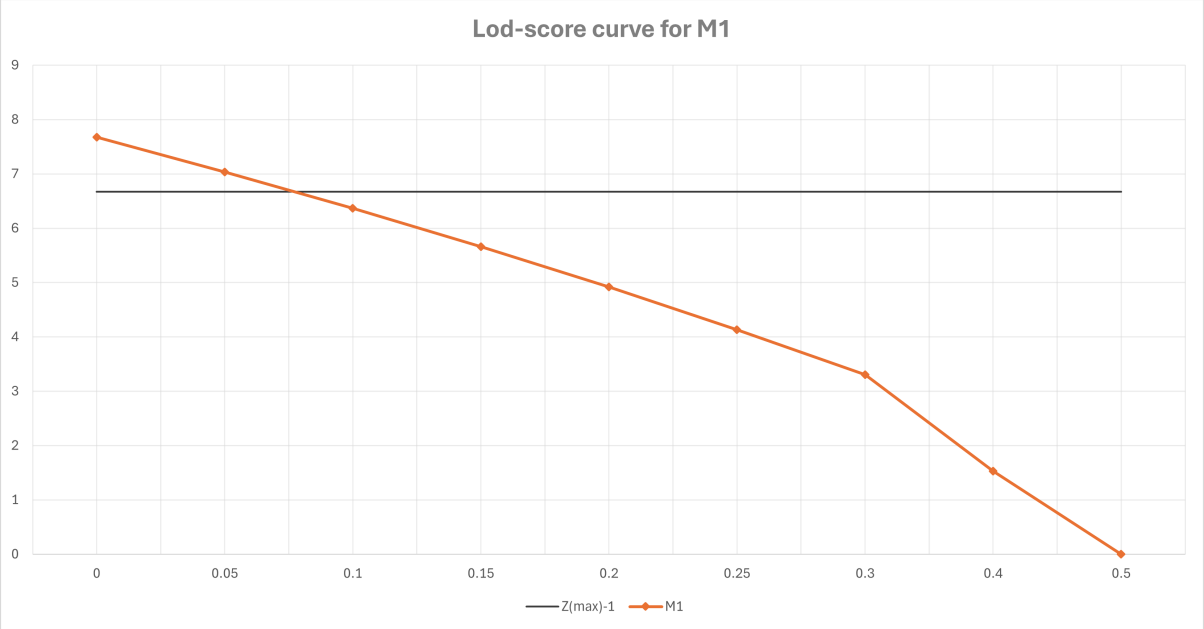

Figure 4: LOD score curve for marker 1. LOD score values on the y-axis, θ values on the x-axis. Grey line represents the $Z_{max} - 1$ value.

Marker	θ_{Zmax}	Z_{max}	**Lower**	**Upper**
1	0.000	7.6738	0.000	0.077
2	0.000	7.2474	0.000	0.082
3	0.000	4.7627	0.000	0.109
4	0.000	8.1712	0.000	0.074
5	0.000	6.0050	0.000	0.090
6	0.000	4.9699	0.000	0.116
7	0.000	5.7141	0.000	0.090
8	0.000	5.2551	0.000	0.097
9	0.000	4.2537	0.000	0.122
10	0.000	3.5643	0.000	0.089
11	0.051	3.5710	0.019	0.201
12	0.045	4.2882	0.019	0.201
13	0.111	2.3914	NA	NA

Table 1: Confidence intervals for each marker, including the maximum LOD score (Z_{max}) and the corresponding recombination fraction (θ_{Zmax}). For marker 13, results are inconclusive.

$0.25 * 0.25 = 0.0625$ *or* 6.25% for each.

But let's assume that for alleles 1-4 if known allele frequencies are 0.1, 0.1, 0,1 and 0.7 respecivelly. In this case the probabilities of occurence for genotypes 4/4 and 1/4 are:

$$\text{Genotype } 4/4 \;=\; 0.7 \times 0.7 \;=\; 0.49 \;=\; 49\%$$

$$\text{Genotype } 1/4 \;=\; 0.1 \times 0.7 \;=\; 0.07 \;=\; 7\%.$$

This has implications on LOD score when considering founders in the sample. When you change the allele frequencies for a marker, you alter the prior probabilities of founder genotypes, since founders rely on these frequencies to estimate their most likely alleles. This shifts the overall likelihood of observing the pedigree data under each hypothesis (linked vs. unlinked). Formally, the LOD score is defined as the base-10 logarithm of the ratio of the pedigree likelihood under the linked hypothesis to that under the unlinked hypothesis:

$$\text{LOD} \;=\; \log_{10}\Big(\frac{P(\text{data} \mid \theta < 0.5)}{P(\text{data} \mid \theta = 0.5)}\Big).$$

Since updated allele frequencies can either make the observed genotypes more plausible under linkage or under no linkage, the resulting LOD score can go up or down depending on whether these new priors improve the fit of the data more for the linked model than for the unlinked model (or vice versa).

Next we have performed the same analysis as above for just marker 5, with the updated allele frequencies for it's 4 alleles(0.1, 0.1, 0,1 and 0.7). Considering previous results have shown that linkage between marker 5 and the disease gene is likely, we expect the LOD score to increase, which was confirmed by our results (see Table 2).

θ	**LOD-mod AF**	**LOD-equal AF**
0.0	6.1601	6.0050
0.05	5.6090	5.4606
0.1	5.0311	4.8898
0.15	4.4243	4.2904
0.2	3.7862	3.6600
0.25	3.1152	2.9969
0.3	2.4110	2.3012
0.4	0.9454	0.8644
0.5	0.0000	0.0000

Table 2: LOD score comparison between data with modified and default allele frequencies for different θ values. The left column remains uncolored, while the other two columns are colored to highlight differences. The first row with headers is explicitly white.

1.4 Misspecifying the genetic model

To demonstrate the impact of misspecifying the genetic model on the LOD score, we performed a linkage analysis on the same dataset, assuming an autosomal recessive mode of transmission while keeping the other parameters identical:

- phenocopies = 10^{-5}
- complete penetrance
- disease allele frequency = 10^{-5}

MARKER	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
θ\LOD	0.308428	0.011001	0.839489	0	0.314571	0.008245	0.905735	0.519404	1.408298	0.488132	0.011489	0.09737	1.309171
0	-11.5274	-20.3452	0.839489	-20.8017	-11.4692	-16.2429	-6.4832	-3.08997	1.408298	-14.5118	-16.1315	-8.40399	-3.2749
0.05	-0.75935	-2.48558	0.736062	-2.88036	-0.62992	-2.03986	0.607967	0.414956	1.250479	-0.61849	-1.95385	-0.92119	1.223133
0.1	-0.06925	-1.21595	0.632743	-1.54557	-0.00204	-1.01197	0.887882	0.519404	1.076408	0.153685	-0.94676	-0.34357	1.309171
0.15	0.20839	-0.61015	0.528933	-0.87518	0.237629	-0.51618	0.905735	0.499014	0.893526	0.422983	-0.46823	-0.07638	1.233101
0.2	0.307679	-0.2826	0.424775	-0.48595	0.314571	-0.24459	0.813197	0.430097	0.708601	0.488132	-0.21078	0.050381	1.079326
0.25	0.308428	-0.10678	0.321871	-0.25337	0.303688	-0.09633	0.662825	0.339125	0.528559	0.447142	-0.07379	0.09737	0.880117
0.3	0.252541	-0.02188	0.223599	-0.11864	0.243624	-0.02296	0.484876	0.241074	0.361077	0.349382	-0.0091	0.096886	0.655873
0.4	0.085466	0.011001	0.063681	-0.01421	0.080598	0.008245	0.147265	0.070054	0.099886	0.113427	0.011489	0.037314	0.214991
0.5	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 5: LOD score analysis results for autosomal recessive mode of inheritance. Top row shows maximum LOD score for each marker. Conditional coloring: green to red - higest to lowest LOD score.

Based on the results shown in Figure 5, we were unable to reject the null hypothesis (no linkage) for any marker at any θ, as none of the LOD scores exceeded 3. For marker 9 maximum LOD score is 1.4 at $\theta = 0$, which could suggest possible linkage if more families were added to the dataset. But for the other markers that are very close to the disease gene based on previous results, the LOD scores are actually the lowest at $\theta = 0$. This outcome aligns with expectations, given the prior knowledge of autosomal dominant mode of inheritance of the disease. This analysis shows the danger of misspecifying the mode of inheritance of the disease.

2 Familial association analysis

The gene GJB6 is situated within the linkage region identified in the previous section and appears to be a strong candidate for a driver gene. To investigate further, we conducted a familial association analysis using data from 652 nuclear families, with an average of 3.1 individuals per family, genotyped for six SNP markers within the gene. We used `fbat.exe` program to analyze the data, with the following parameters:

- **Trait: affection**
  Specifies that the test is focused on a affection status (e.g., affected vs. unaffected).
- **Offset: 0.000**
  The offset defines an adjustment for the null hypothesis. An offset of `0.000` assumes no deviation from the null expectation in the absence of association.
- **Model: additive**
  The additive model assumes that the genetic effect of the allele increases additively

with the number of copies of the allele (e.g., heterozygotes have an intermediate effect between homozygotes for the major and minor alleles).

- **Test: bi-allelic**
  Specifies that the test is performed on bi-allelic markers (e.g., SNPs), considering only two alleles per locus.

- **Min family size: 10**
  Results are given for SNPs with at least 10 informative families.

- **Minimum allele frequency: 0.000**
  No exclusion based on allele frequency.

- **p-value threshold: 1.000**
  All p-values are reported, regardless of significance.

- **Maximum CMH statistic: 1000**
  Defines the upper limit for the Cochran-Mantel-Haenszel (CMH) statistic.

An informative family is one where at least one parent is heterozygous for the marker in question. Such families contribute to the analysis because heterozygosity allows the transmission of alleles to offspring to depend on the recombination rate. In contrast, families where parents are homozygous for the marker cannot provide information about linkage, as the likelihood of transmission does not vary with recombination. The results are shown in the Table 3 below.

SNP1 is not present in the above results since it didn't fulfill the criteria of having at least 10 informative families. For the rest of the SNPs no significant association was found, with once exception being SNP6. This SNP showed very high **S-E(S)** score with very high level of significance. Based on this score and its sign for the two alleles, we can say that allele 2 is transmitted from heterozygous parents to affected children much more frequently than allele 1. This suggests that allele 2 might be associated with the disease. We then explored linkage disequilibrium (LD) between this SNP and other SNPs using Ensembl. We used GRCh37 version of human genome and the population of `1000GENOMES:phase_3:CEU` since they are appropriate for the data we have. On the Figure 6 we can see teh result of our search. There are no SNPs in high LD with SNP6, which suggests that the association is not due to another nearby variant, but rather to SNP6 itself.

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-8
SNP6	2	0.788	283	52.000	91.000	5.451	5.01e-8

Table 3: Results of the FBAT analysis for SNPs with at least 10 informative families. Each SNP has two rows of results, one for each allele. The columns are defined as follows: **Afreq** represents the allelic frequency, **Fam#** is the number of informative families, **S-E(S)** is the score used to test association minus its expected value under the null hypothesis (with the sign indicating whether an allele is more or less frequently transmitted from heterozygous parents to affected children), **Var(S)** is the variance of the score, **Z** is the test statistic calculated as $(S - E(S))/\sqrt{Var(S)}$, and **P** is the p-value associated with the test ($p \leq 0.05$ indicates rejection of H_0).

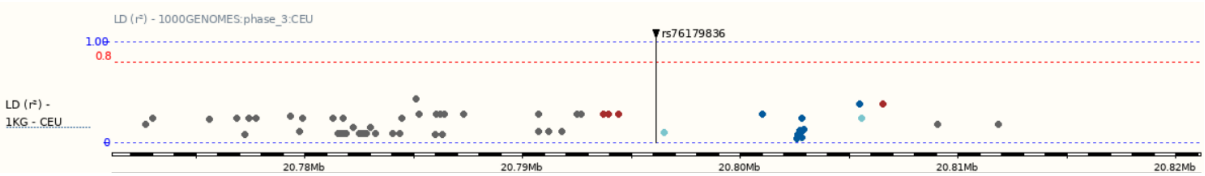

Figure 6: Linkage disequilibrium plot for SNP6 - rs76179836 and other SNPs in the GJB6 gene. The plot shows the r2 values for each pair of SNPs. On the x-axis shown are $\mathtt{r}^2$ values for each SNP. The blue and red dashed lines represent $\mathtt{r}^2$ of 1 and 0.8, respectively. On the y-axis are the coordinates on the chromosome.

3 Multifactorial disease: Rheumatoid Arthritis