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 (lodscore)
   (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 lodscore 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 - lodscore 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.

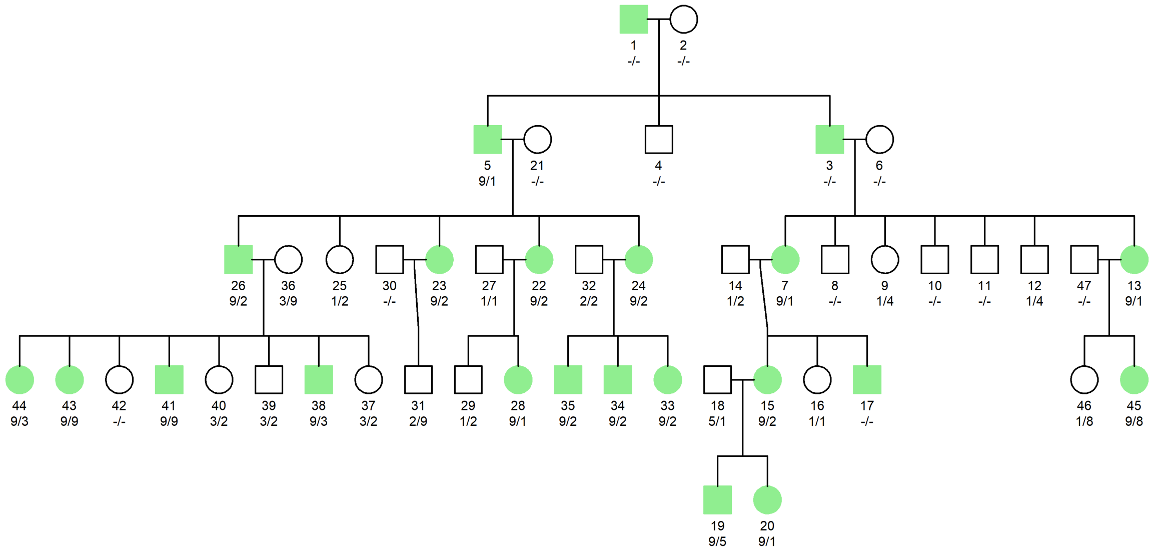


Figure 1: Pedigree plot of the family along with their affection status and genotypes with regards to marker 1. Affected - light green, Unaffected - white. Sex: square/circle - M/F. Genotype: alleles of the marker 1.

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.

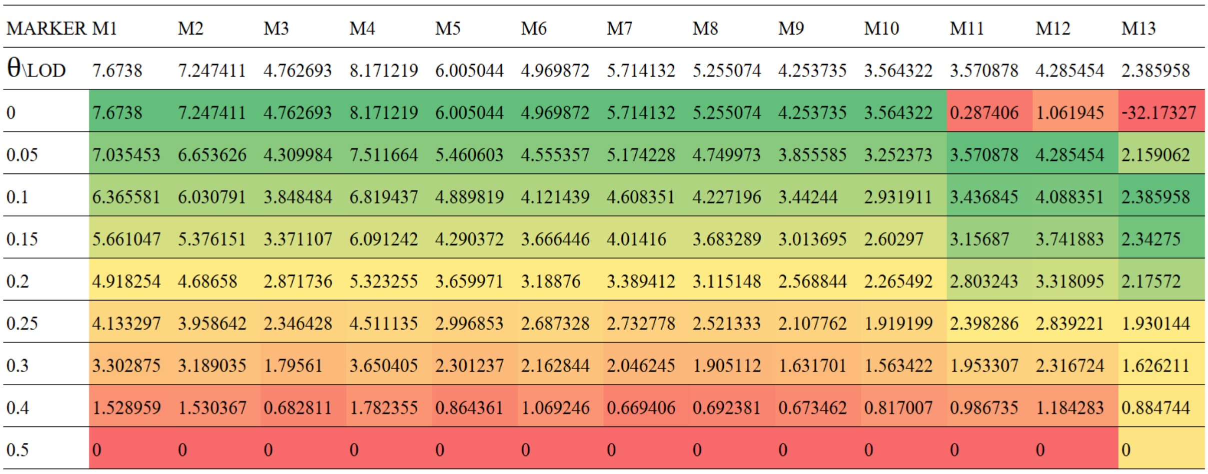
1.2 Mode of inheritance: Autosomal Dominant

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)
```

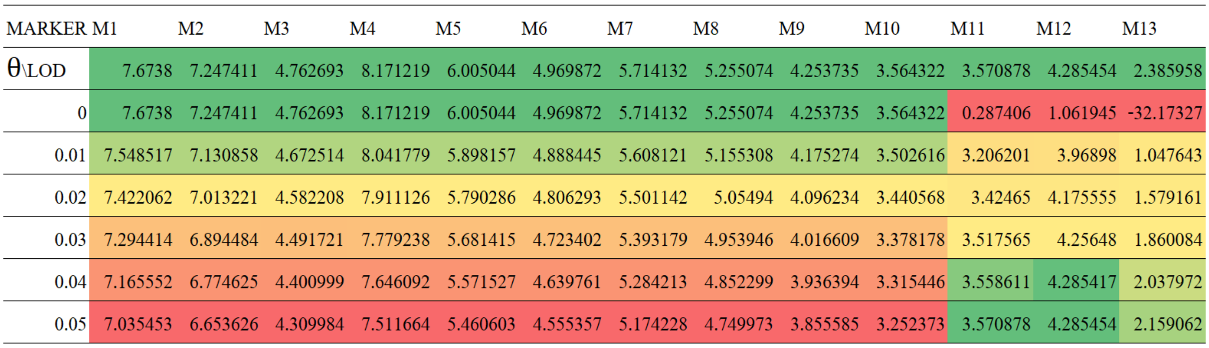


MARKER	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13
θ\LOD	7.6738	7.247411	4.762693	8.171219	6.005044	4.969872	5.714132	5.255074	4.253735	3.564322	3.570878	4.285454	2.385958
0	7.6738	7.247411	4.762693	8.171219	6.005044	4.969872	5.714132	5.255074	4.253735	3.564322	0.287406	1.061945	-32.17327
0.05	7.035453	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 probability of observing a recombinant event c to fall within $0 < c < 50\%$. LOD-score is a logarithmic likelihood ratio:

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



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.

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 2–10 (Figure 4). To determine the precise confidence intervals for the results, we ran the script located in the `paramlink.R` file. Results are shown in the table below.

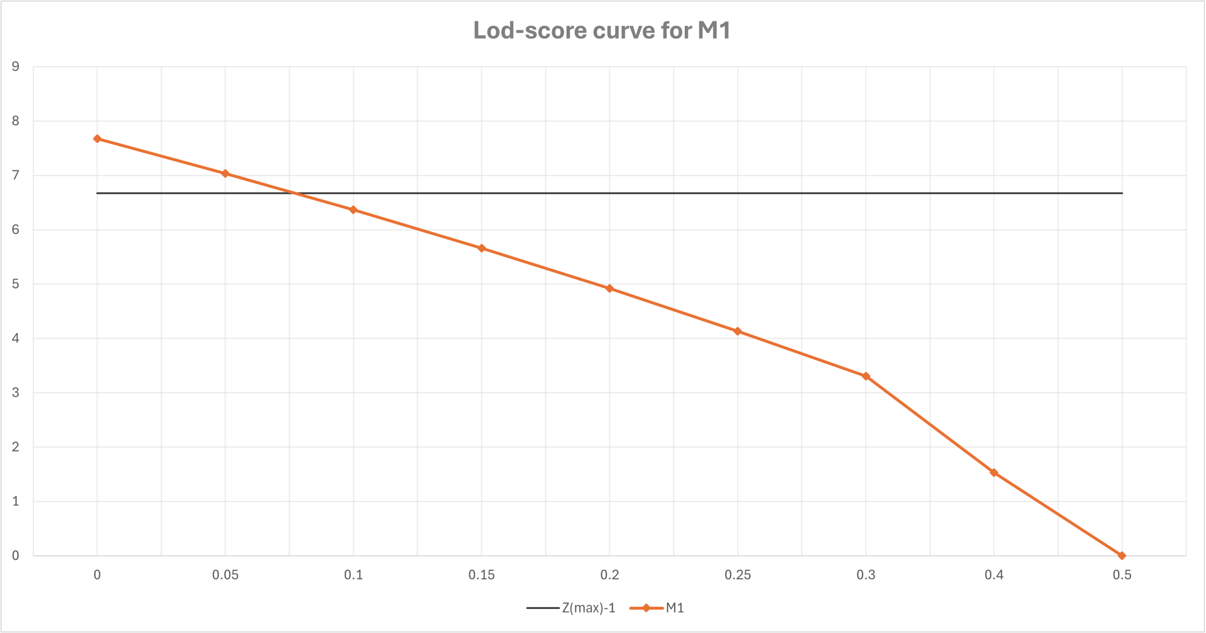


Figure 4: LOD-score curve for marker 1. LOD-score values on the y-axis, θ values on the x-axis. Grey line represents the $Z(max) - 1$ value.

2 Multifactorial disease: Rheumatoid Arthritis

Marker	**Lower**	**Upper**
1	0	0.077
2	0	0.082
3	0	0.109
4	0	0.074
5	0	0.09
6	0	0.116
7	0	0.09
8	0	0.097
9	0	0.122
10	0	0.155
11	0.059	0.201
12	0.059	0.201
13	NA	NA

Table 1: Confidence intervals for each marker, with lower and upper bounds. For marker 13, results are inconclusive.