ISFG-GHEP Advanced Theoretical Challenge 2025: Kinship

Prepared by: Magnus Dehli Vigeland

General instructions

This is a multiple-choice test consisting of 20 questions. For each question exactly one alternative is correct. You are free to use whatever software you like, but keep in mind that some programs have built-in conventions (e.g., rounding) that may affect the output. If your answer does not precisely match any of the options, choose the closest one.

Assumptions: Throughout, we assume no deviations from Hardy-Weinberg equilibrium, no linkage disequilibrium, no drop-outs, drop-ins or silent alleles, and no mutations unless explicitly stated.

Introduction

The topic of this year's challenge is *linkage* and the effect of linked markers in kinship testing. For convenience we recall some basics here; for more details, some references are included at the end.

The *recombination rate* θ between two markers is the long term fraction of gametes (meiotic products) whose alleles at the markers come from different parental strands. For instance, Figure 1 shows a mother with genotypes A/a and B/b at two markers. For simplicity we assume known *phase* as indicated by the vertical line. Four different haplotypes may result from this situation, as shown in the bottom part of Figure 1 along with their transmission probabilities. The two outer haplotypes are non-recombinant; the inner ones are recombinant.

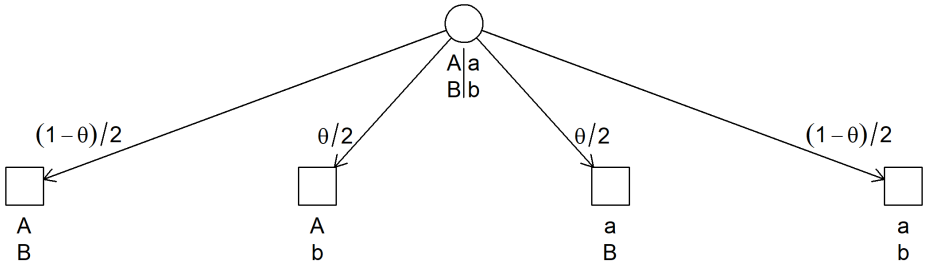


Figure 1. Illustration of recombination and haplotype transmission probabilities in terms of θ.

For two independent loci, e.g., situated on different chromosomes, the four haplotypes in Figure 1 are equally likely, and $\theta = 0.5$. Such loci are called *unlinked*. Conversely, loci with $\theta < 0.5$ are *linked*.

The *genetic map distance* between two loci on the same chromosome is defined as the expected number of crossovers between them per meiosis. The basic unit is 1 Morgan, corresponding to 1 expected crossover per meiosis. A more common unit is 1 centiMorgan (cM) = 0.01 Morgan.

Several approximations exist for converting recombination rates to genetic distance, and *vice versa*. The simplest is Haldane's map function, relating the recombination rate θ between two loci, and their genetic distance d measured in M:

$$\theta = \frac{1}{2}(1 - e^{-2d}).$$

Software: The following is an (non-exhaustive) list of programs that support linked markers, and which may be useful for this challenge. (Click on the links for more information.) Some instructions for KLINK are included at the end.

- *KLINK*: Recently developed app based on the pedsuite, compatible with the Familias software.
- *FamLink2*: Well-established stand-alone software for Windows. Accepts Familias files as input.
- *pedsuite*: A collection of R packages for pedigree analysis, including a wide range of functions for kinship testing.
- *MERLIN*: A popular software for linkage analysis, though not specifically designed for kinship testing.

Part I: A delicate inheritance case

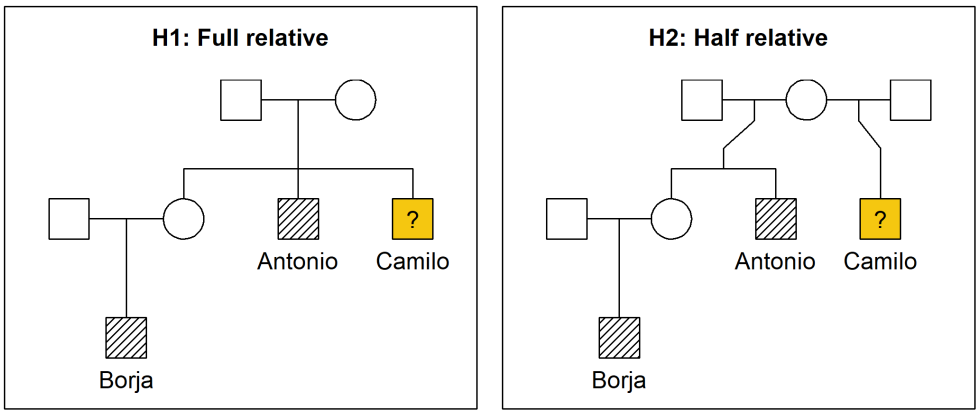
Detective Vargas had seen his share of family disputes, but this one had trouble written all over it. Antonio and his nephew Borja, both with piercing blue eyes, leaned over his desk, aggressively arguing that Camilo, seated quietly next to them, had no claim to the inheritance. Antonio's father had been a rich man, and now millions were up for grabs.

"He is not my full brother," said Antonio. "I always suspected my mother had something going on with the gardener. Or that low-life juggler she used to invite to parties."

"Just look at his brown eyes!" added Borja. "And he can't even twirl his tongue like we can!"

They both started rolling their tongues frenetically. Vargas closed his eyes and leaned back in his chair. Mendelian party tricks wouldn't solve this one—distinguishing full siblings from half siblings required a more scientific approach. Would the standard kinship tests even suffice? With a sigh, he reached for the phone and called the best forensic geneticist in town …

After the phone call from Detective Vargas, you sketch the pedigrees forming the competing hypotheses of the case:



The following files are available for this exercise:

- *vargasA.ped* and *vargasA.fam*: Dataset A in standard ped format and Familias format
- *vargasCombined.ped* and *vargasCombined.fam*: Datasets A and B combined
- *db35.txt*: Frequency database for all markers used in this case
- *map35.txt*: Marker map with chromosome and cM position of all markers

For a given dataset involving linked markers, we introduce the following notation:

- LR^{ℓ} denotes the linkage-aware LR, i.e., where linkage is taken into account.
- LR^{u} denotes the LR when linkage is ignored, i.e., obtained by simply multiplying all the marker-wise LRs.

All LRs refer to the hypotheses H1 and H2 shown above, with H1 as the numerator and H2 as the denominator.

Initially you type the three men with a standard kit including 19 autosomal STRs. This is dataset A (Table 1).

1. The total LR^{u} for dataset A is approximately

a) 0.71
b) 2.01
c) 35.78
d) 109.23
e) 213.07

Table 1: Dataset A

Marker	Antonio	Borja	Camilo
D1S1656	16.3/20.3	16.3/16.3	12/20.3
TPOX	11/11	8/11	8/11
D2S1338	17/23	17/23	17/19
D3S1358	16/16	14/15	16/16
FGA	19/23	20/21	19/20
CSF1PO	10/12	9/12	8/12
D6S474	17/18	14/16	15/16
D7S820	8/12	8/8	10/12
D8S1179	12/13	12/14	12/13
D10S1248	13/16	13/16	14/16
TH01	9/9.3	7/9	9/9.3
vWA	17/18	17/17	17/18
D13S317	12/12	8/12	12/12
PentaE	5/12	5/12	12/20
D16S539	9/11	9/11	9/11
D18S51	19/20	13/16	14/19
D19S433	13/14	14/15	13/14
PentaD	11/12	9/11	11/12
D22S1045	15/16	16/16	11/15

Table 2: Dataset B

Marker	Antonio	Borja	Camilo
D3S3045	11/15	11/14	9/15
D5S2800	17/17	14/17	14/17
D7S3048	21/24	21/23	21/24
D8S1132	19/22	20/22	19/22
D9S1122	12/13	12/12	10/13
D10S1435	11/12	12/13	11/12
D11S2368	19/20	18/19	19/20
D12S391	20/24	17/24	17/20
D13S325	18/19	18/19	18/19
D14S1434	10/11	10/10	10/11
D15S659	17/18	14/18	17/18
D17S1301	11/11	11/12	11/12
D18S1364	13/18	14/14	13/16
D19S253	7/12	12/13	7/12
D20S482	11/14	14/15	12/16
D21S11	28/31.2	28/28	28/30

2. Two of the markers in dataset A are located on the same chromosome. According to Haldane's map function, the recombination rate between them, rounded to two decimals, is

a) 0.41
b) 0.42
c) 0.45
d) 0.48
e) 0.49

3. Let FC denote the *fold change* ratio $\mathrm{LR}^{\ell}/\mathrm{LR}^{u}$ for the two markers in the previous question. Then

a) $FC < 1$
b) $FC = 1$
c) $1 < FC \leq 1.0001$
d) $1.0001 < FC \leq 1.001$
e) $FC > 1.001$

You decide to type the three individuals with another kit, adding 16 new STR markers. This is dataset B (Table 2). In the following, two markers are said to form a *linked pair* if they are located on the same chromosome.

4. The number of linked pairs in the total dataset (including the pair in Exercise 2) is

a) 11
b) 12
c) 13
d) 14
e) 15

5. The total LR^{ℓ} with all 35 markers, after rounding, is

a) 213
b) 1 031
c) 7 624
d) 15 401
e) 23 125

6. The total fold change $\mathrm{LR}^{\ell}/\mathrm{LR}^{u}$ with all 35 markers, is approximately

a) 0.55
b) 1.01
c) 2.02
d) 3.03
e) 4.04

7. The number of linked pairs for which LR^{ℓ} is larger than LR^{u}, is

a) 4
b) 6
c) 8
d) 10
e) 12

8. The number of linked pairs for which LR^{ℓ} and LR^{u} support *different* hypotheses, is

a) 0
b) 1
c) 2
d) 3
e) 4

A natural measure of how much linkage affects LR is the absolute value of the base-2 log of the fold change, i.e.,

$$W = \left|\log_2\left(\mathrm{LR}^{\ell}/\mathrm{LR}^{u}\right)\right|.$$

9. Find the value of W for each of the linkage pairs. Then,

a) the closest pair has the largest W, and the furthest pair has the smallest W.
b) the closest pair has the largest W, but the furthest pair does not have the smallest W.
c) the closest pair does not have the largest W, but the furthest pair has the smallest W.
d) the furthest pair have the largest W.
e) none of the above.

In your report to Detective Vargas, you derive a conclusion based on the complete dataset. Your lab's protocol considers $\mathrm{LR} > 10\,000$ as strong support for H1 and $\mathrm{LR} < 0.0001$ as strong support for H2; thresholds of $1\,000$ and 0.001 provide moderate support. Values between these thresholds are considered inconclusive.
If LR^{ℓ} and LR^{u} give different conclusions, you must decide how to deal with this.

10. You conclude that the data

a) strongly supports a full relationship (as in H1).
b) moderately supports a full relationship.
c) strongly supports a half relationship (as in H2).
d) moderately supports a half relationship.
e) provides inconclusive evidence.

Part II: Linkage Lab

The next exercises are based on the Shiny app **Linkage Lab**, available online at *https://magnusdv.shinyapps.io/linkagelab/*. This app is designed to explore the effect of linkage in common kinship scenarios.

Each scenario compares two possible relationships between individuals A and B, who are genotyped at two markers, M1 and M2. Both markers have four alleles, labelled 1, 2, 3, and 4, whose allele frequencies and mutation rates can be set by the user (these parameters are assumed to be the same for both markers). The distance between the markers can be given in centiMorgans or as a recombination rate. The app calculates the likelihood ratio (LR) for the two relationships and presents the results in a plot, where one of the parameters is free to vary.

All questions below pertain to the `Sibs : half-sibs` comparison. Initially, set all allele frequencies to 0.25 and the mutation rate to 0.

11. Let A have genotype 1/2 for both markers, while B is 3/4 for both markers. Then it is *not* true that

a) the LR increases as the linkage gets stronger.
b) the LR under complete linkage is precisely twice that under no linkage.
c) for both hypotheses, the likelihood is largest when the recombination rate is 0.
d) the data supports the full-sib hypothesis if the recombination rate is 0.
e) the data supports the half-sib hypothesis if the recombination rate is 0.5.

12. Now change B's genotype at M1 to 1/2. Then it is *not* true that

a) the LR increases as the linkage gets weaker.
b) the LR under no linkage is precisely twice that under complete linkage.
c) for both hypotheses, the likelihood is largest when the recombination rate is 0.5.
d) the data supports the half-sib hypothesis if the recombination rate is 0.
e) the data supports the full-sib hypothesis if the recombination rate is 0.5.

13. Now change B's genotype at M2 to 1/3 (keeping M1 as in the previous exercise). Then it is *not* true that

a) the LR is below 1 if the markers are completely linked.
b) the LR is largest if the recombination rate is 0.25.
c) the likelihood of Ped 1 depends on linkage, but the likelihood of Ped 2 does not.
d) the data supports the full-sibs hypothesis if the markers are unlinked.
e) the LR equals 1 if the markers are x cM apart, for some value of x between 0 and 10.

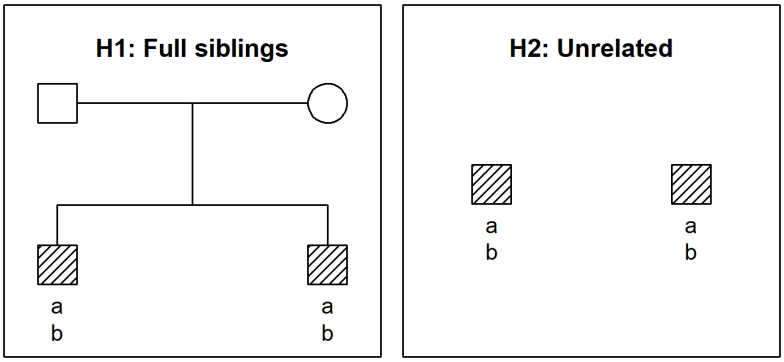
Finally, let A and B both be 1/1 at M1, and respectively 2/2 and 3/3 at M2. Set the `Plot variable` to "Frequency of '1' allele", the marker distance to 50 cM, and the mutation rate to 0.005. Keep the frequencies of alleles '2' and '3' at 0.25 as before.

14. The frequency of '1' for which $LR = 1$, is approximately

a) 0.1
b) 0.2
c) 0.3
d) 0.4
e) 0.5

Part III: Linkage on X

The remaining exercises explore the effect of linkage between X-chromosomal SNPs when testing for brotherhood. Suppose that two boys are both hemizygous for allele a at marker X1, and allele b at marker X2, as shown in the figure below.



Let $p = P(a)$ and $q = P(b)$ be the allele frequencies of a and b. In all exercises below we assume that p and q are both close to 0. Let θ denote the recombination rate between the markers.

15. The probability that a random female person carries the a allele at locus X1, is approximately

a) $p^2 - p$
b) $p/2$
c) p
d) $2p$
e) p^2

16. The likelihood of hypothesis H_1, ignoring a constant factor[1], is best approximated by

a) $pq\,\theta(1-\theta)$
b) $pq\,(1-\theta)^2$
c) $p^2q^2\,(1-\theta)^2$
d) $pq\,(\theta^2+(1-\theta)^2)$
e) $p^2q^2\,(\theta^2+(1-\theta)^2)$

17. The LR comparing H_1 to H_2, ignoring a constant factor, is best approximated by

a) $(\theta^2+(1-\theta)^2)/(pq)$
b) $1/(pq)$
c) $pq/(\theta^2+(1-\theta)^2)$
d) $pq/\big(\theta(1-\theta)\big)$
e) pq

18. In this situation,

a) complete linkage is impossible unless a mutation has happened.
b) complete linkage roughly doubles the LR compared to unlinked markers.
c) the LR is largest if $\theta = 0.25$.
d) the LR is largest if the markers are unlinked.
e) the LR is approximately independent of linkage.

Now consider the same situation, but where the leftmost boy has the *common* allele A at marker X1.

19. In this situation,

a) complete linkage is impossible unless a mutation has happened.
b) complete linkage roughly doubles the LR compared to unlinked markers.
c) the LR is largest if $\theta = 0.25$.
d) the LR is largest if the markers are unlinked.
e) the LR is approximately independent of linkage.

Finally, consider the same situation, but where *both* boys have the common allele A at marker X1.

20. In this situation,

a) complete linkage is impossible unless a mutation has happened.
b) complete linkage roughly doubles the LR compared to unlinked markers.
c) the LR is largest if $\theta = 0.25$.
d) the LR is largest if the markers are unlinked.
e) the LR is approximately independent of linkage.

GOOD LUCK!

Magnus

[1] For example, if the truth is $3pq\theta$, then $pq\theta$ is considered correct. (The point of this phrasing is to avoid trivial solutions by checking the alternatives directly.)

Using KLINK

Here are some pointers to the KLINK app for kinship testing with pairwise linked markers:

- The online app is available at *https://magnusdv.shinyapps.io/klink/*.
- Load the .fam file with the dataset you want to analyse.
- Set *Marker map* to "Custom" and load the map file.
- Set *Map function* to "Haldane".
- You might want to adjust the *Ignore linkage above (cM)* setting.
- Click *Calculate LR*. Note that the results include both linked LRs and unlinked LRs.
- Increase *Decimals* if you need higher precision.

Further reading

The following books provide more details on linkage in general, and on linked markers in kinship testing.

- Thompson, *Statistical Inference from Genetic Data on Pedigrees*. Institute of Mathematical Statistics, 2000.
- Egeland, Kling & Mostad, *Relationship Inference with Familias and R*. Academic Press, 2016.