BMC Proceedings 2007;1(Suppl 1):S102

On the choice of linkage statistics

Patricia Margaritte-Jeannin [1,2], Marie-Claude Babron [1,2§], Françoise Clerget-Darpoux [1,2]

[1]INSERM U535, F-94817 Villejuif, France
[2]Univ Paris-Sud, IFR69, UMR535, F-94817 Villejuif, France

[§]Corresponding author

Marie-Claude Babron
Fax : +33 1 45 59 53 31
Email : babron@vjf.inserm.fr

Email addresses:
PMJ: jeannin@vjf.inserm.fr
MCB: babron@vjf.inserm.fr
FCD: clerget@vjf.inserm.fr

Abstract

Three lod score statistics are often used for genome wide linkage analysis : the Maximum Lod Score, the lod score statistic proposed by Kong and Cox, both based on the allele-sharing between affected sib-pairs and the maximization of the lod score function of Morton on two genetic models and an heterogeneity parameter.
Using only Identity By Descent sharing between affected sibs as linkage information, we studied the behavior of these three statistics under the null hypothesis in the Rheumatoid Arthritis simulated data (problem 3 – simulating model known).
Distributions under the null hypothesis show that identical values of the statistics correspond to very different genome-wide p-values : comparison and interpretation of several linkage statistics cannot be done on the observed value.
The power of detection of the HLA region involved in Rheumatoid Arthritis shows a slight advantage of the Kong and Cox lod score statistic.
In a second step, we show that performing the analysis under a greater number of genetic models in the hope of better scanning the space of models, does not increase the power of detection.

Background

Genome wide linkage studies are often performed on affected sib pairs to detect disease susceptibility genes in multifactorial diseases. Many statistics have been proposed for such a goal.
Two methods, the lod score statistic proposed by Kong and Cox [1], hereafter denoted by KC-Lod, which is an extension of the Non Parametric Linkage statistic [2], and the Maximum Lod Score (MLS) [3], are based on the allele sharing between affected sibs and do not require the specification of a model at the disease locus which, in the case of a multifactorial disease, is unknown.
An alternative strategy, proposed by Greenberg et al [4], consists in maximizing the lod score function of Morton [5] on two genetic models at the disease locus and on an additional heterogeneity parameter. We will call this statistic, HLOD-S1. With the idea of better scanning the space of models to improve power, many authors (see for example [6-9]) considered a wider set of genetic models, without consensus on which and how many models should be employed. Here, we will focus on the maximum statistic obtained over 4 different genetic models, which we will call HLOD-S2.
These statistics are all lod scores, i.e. the decimal logarithm of the ratio of two likelihoods (linkage versus no linkage). They are computed at each marker of a given chromosome, taking into account the multipoint information provided by the entire set of markers. The maximum value observed for each chromosome is then retained to perform the linkage test. However, since these statistics differ on the parameters on which the maximization is achieved, they are likely to have different statistical properties.
In this work, we study the behavior of these statistics under the null hypothesis, and then evaluate their performance for detecting the HLA risk factor in the Rheumatoid Arthritis (RA) simulated data (problem 3). The simulating model was known prior to the analysis.

Methods

Material

The segregation of 730 microsatellite markers, spaced on 22 chromosomes with an average inter-marker distance of about 5 cM, were simulated on 100 replicates of 1500 families with at least two affected sibs.
Preliminary linkage analyses showed that, with such sample sizes, it was not possible to make any power comparison : all linkage statistics were highly significant for detecting the role of HLA, while their power was very low (less than 5%) for the other loci. Therefore, we decided to focus on the detection of the susceptibility factor in the HLA region and to split each replicate into smaller family samples, in order to have a lower, but not too low, power of detection. A sample size of 60 seemed appropriate. Each replicate was split into 25 sub-samples. The study was thus performed on 2500 replicates of 60 families each. Parental status was considered unknown in all replicates, so that linkage information consists in the Identity By Descent (IBD) sharing between affected sibs.

Linkage statistics

The data was analyzed by four lod score statistics, MLS, KC-Lod, HLOD-S1 and HLOD-S2.
The Maximum Lod Score (MLS) [3] maximizes the likelihood of the IBD sharing vector, within the Possible Triangle constraints [10]. Under the null hypothesis, the expected IBD vector is [0.25;0.50;0.25]. Calculations were performed with the Mapmaker/Sibs software [11].
The KC-Lod proposed by Kong and Cox [1], is maximized on a single parameter δ representing the degree of allele sharing among affected individuals. Under the null hypothesis, δ is equal to 0 , and the higher δ, the higher the allele sharing. KC-Lod analysis was carried out under the “score pairs” option and the exponential model proposed by Kong and Cox with Allegro v1.2 [12].
HLOD-S1 was calculated as initially proposed by Greenberg et al [4] under a dominant and recessive model, each with a disease allele frequency of 0.01, a penetrance of 0.50 and no phenocopies. The lod score function was maximized over these two models and the heterogeneity parameter α, representing the proportion of families linked to the disease locus. In HLOD-S2, two additional models were considered, with a disease allele frequency of 0.2. The lod score function was then maximized over these 4 genetic models and the parameter α. All HLOD calculations were done with the Allegro v1.2 software [12].
We first studied the distribution of these 4 statistics under H0 (no linkage), by analyzing the 16 chromosomes which did not harbor a susceptibility gene. The maximum value on each chromosome was recorded for each statistic, leading to 40000 values (2500 replicates x 16 chromosomes). This provides the distribution of the maximum of each statistics for an average chromosome. Thus, for a full genome scan, one may apply a Bonferroni correction for 22 chromosomes. This procedure can be used either to determine the threshold for a genome-wide type I error of 5% (nominal p= 0.002 per chromosome) or to determine the genome-wide p-value corresponding to a given value of the statistics.
The power for detecting linkage was calculated as the number of times a given statistic exceeded the threshold corresponding to a genome-wide type I error of 5%.
Two loci in the HLA region were known to be involved in the determinism of the

simulated disease. We considered the HLA region was detected if there was evidence for linkage in the 20cM interval around the HLA-DR locus, i.e. in the interval [STRP6_10 - STRP6_13].

Results and discussion

Distribution of the statistics under the null hypothesis

Figure 1 shows the false positive rates corresponding to the observed values of the KC-Lod, MLS and HLOD-S1 under the null hypothesis. For clarity, the graph is limited to values between 2 and 4.

Although the 3 linkage statistics are maximum lod scores, their distributions are different. This is due to the different underlying parameterization. Note, however, that MLS and HLOD-S1 have very similar distributions. The HLOD-S2 statistic, performed under 4 different genetic models, has a distribution similar to that of MLS and HLOD-S1 (data not shown).

Identical values of observed KC-Lod and MLS (or HLOD-S1) give rise to very different genome-wide p-values. For example, when a value of 2 is observed, the false positive rate is 31% for KC-Lod, while it attains almost 50% for the MLS and HLOD-S1 (49.4 and 48.4%, respectively). This shows that the comparison and interpretation of several linkage statistics cannot be done on the observed value.

The thresholds corresponding to a genome-wide type I error of 5% are 2.89, 3.23 and 3.19 for the KC-Lod, MLS and HLOD-S1, respectively.

Power for detecting linkage in the HLA region

The power of linkage detection of each statistic was determined for 2500 replicates of 60 families on chromosome 6, using the thresholds above. The power is 48.5%, 45.9% and 44.6% for KC-Lod, MLS and HLOD-S1, respectively, showing a slight advantage of KC-Lod.

In a second step, we compared the impact of 4 vs. 2 genetic models in the HLOD analysis. Both statistics have the same the 5% genome-wide threshold. The power of the 4-model HLOD-S2, is very slightly, but not significantly, increased (from 44.6 to 45.8%). This increase may be explained by the very strong correlation ($r2 > 0.97$) between the HLODs obtained for the 2 dominant models ($q=0.2$ and $q=0.01$) and for the 2 recessive models, respectively.

Conclusions

The linkage statistics studied here are all maximum lod scores. However, the KC-Lod distribution under the null hypothesis is very different from that of the MLS and HLOD-S1. The same observed value can correspond to very different p-values. We would like to stress, following Nyholt [13], that the interpretation of linkage results should not be performed in terms of observed value of the statistics, but that appropriate significance threshold should be empirically calculated on the family structures under study.

It has been claimed that HLOD-S1 had similar or even greater power than so-called "non parametric" methods, such as the MLS, or the NPL [2]. Here, under the model simulated to mimic HLA susceptibility in Rheumatoid Arthritis, the power of KC-Lod is slightly higher than HLOD-S1 and MLS. This result is not general, as it very likely depends on the underlying model and on the sampled family structures. Here, data

consisted of affected sib pairs and the information on linkage was only provided by the IBD sharing between affected individuals.
Finally, several authors apply the HLOD statistics, using a wide variety of genetic models, in the hope of better scanning the space of models, and thus increasing the power of detection [6-9]. This is not the case here: performing the analysis under four different genetic models, HLOD-S2, does not increase the power. This is due to the high correlation observed in the value of the statistics under genetic models that differ only by the disease allele frequency. When a lod score function is maximized over a set of genetic models (the so-called mod score function [14]), overparametrisation may happen for some familial structures [15]. In other words, the same maximum may be reached for an infinite set of key parameters. In particular, Clerget-Darpoux et al [14] showed that, in nuclear families with two children, the same maximum mod score was obtained for an infinite set of disease allele frequencies and recombination fractions. Similarly, many sets of disease allele frequency and heterogeneity values can explain the IBD sharing of an affected sib pair sample.

References

1. Kong A, Cox NJ: **Allele-sharing models: LOD scores and accurate linkage tests**. *Am J Hum Genet* 1997, **61**(5):1179-1188.
2. Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES: **Parametric and nonparametric linkage analysis: a unified multipoint approach**. *Am J Hum Genet* 1996, **58**(6):1347-1363.
3. Risch N: **Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs**. *Am J Hum Genet* 1990, **46**(2):242-253.
4. Greenberg DA, Abreu P, Hodge SE: **The power to detect linkage in complex disease by means of simple LOD-score analyses**. *Am J Hum Genet* 1998, **63**(3):870-879.
5. Morton NE: **Sequential tests for the detection of linkage**. *Am J Hum Genet* 1955, **7**(3):277-318.
6. Puca AA, Daly MJ, Brewster SJ, Matise TC, Barrett J, Shea-Drinkwater M, Kang S, Joyce E, Nicoli J, Benson E *et al*: **A genome-wide scan for linkage to human exceptional longevity identifies a locus on chromosome 4**. *Proc Natl Acad Sci U S A* 2001, **98**(18):10505-10508.
7. Jin Y, Teng W, Ben S, Xiong X, Zhang J, Xu S, Shugart YY, Jin L, Chen J, Huang W: **Genome-wide scan of Graves' disease: evidence for linkage on chromosome 5q31 in Chinese Han pedigrees**. *J Clin Endocrinol Metab* 2003, **88**(4):1798-1803.
8. Kenealy SJ, Schmidt S, Agarwal A, Postel EA, De La Paz MA, Pericak-Vance MA, Haines JL: **Linkage analysis for age-related macular degeneration supports a gene on chromosome 10q26**. *Mol Vis* 2004, **10**:57-61.
9. Stambolian D, Ibay G, Reider L, Dana D, Moy C, Schlifka M, Holmes T, Ciner E, Bailey-Wilson JE: **Genomewide linkage scan for myopia susceptibility loci among Ashkenazi Jewish families shows evidence of linkage on chromosome 22q12**. *Am J Hum Genet* 2004, **75**(3):448-459.
10. Holmans P: **Asymptotic properties of affected-sib-pair linkage analysis**. *Am J Hum Genet* 1993, **52**(2):362-374.
11. Kruglyak L, Lander ES: **Complete multipoint sib-pair analysis of qualitative and quantitative traits**. *Am J Hum Genet* 1995, **57**(2):439-454.

12. Gudbjartsson DF, Jonasson K, Frigge ML, Kong A: **Allegro, a new computer program for multipoint linkage analysis**. *Nat Genet* 2000, **25**(1):12-13.
13. Nyholt DR: **All LODs are not created equal**. *Am J Hum Genet* 2000, **67**(2):282-288.
14. Clerget-Darpoux F, Bonaiti-Pellie C, Hochez J: **Effects of misspecifying genetic parameters in lod score analysis**. *Biometrics* 1986, **42**(2):393-399.
15. Clerget-Darpoux F: **Extension of the lod score: the mod score**. *Adv Genet* 2001, **42**:115-124.

Figures

Figure 1 - False positive rate as a function of the observed values of 3 linkage statistics under no linkage

Solid black line: KC-Lod; Dashed line : MLS; Dotted line: HLOD-S1.