A tutorial on LHiSA

or how to use the Local Score statistic in Genetic Epidemiology and the analysis of (genome-wide) association studies.

15/09/07

Authors : Mickaël Guedj, Jérôme Wojcik, Grégory Nuel.

Laboratoire Statistique et Génome, UMR CNRS 8071, INRA 152, University of Evry, France.

Statistics for System Biology group, Paris, France.

Merck-Serono, France.

Email: mickael.guedj@gmail.com

LHiSA website: http://stat.genopole.cnrs.fr/software/lhisa

Corrections of this document, questions about LHiSA or the use of the Local Score to Genetic Epidemiology should be addressed to the author. Any suggestion can help to improve this document and the implementation of LHiSA.

If you refer to the application of the Local Score to Genetic Epidemiology and the analysis of (genome-wide) association studies, please cite the initial publication by Guedj, Robelin et al in *Stat. Appl. Genet. Mol. Bio.* (2006).

I - Introduction

In genetic association studies, we can expect to observe an accumulation of high values of test statistics around a disease suscptibility locus (DSL). This accumulation can be due to (i) linkage disequilibrium with surrounding markers or (ii) an aggregation of several DSLs in the same genomic region. Consequently by detecting such accumulations and by considering at the same timein the analysis associated regions instead of markers taken independently, we should improve the discovery of DSLs.

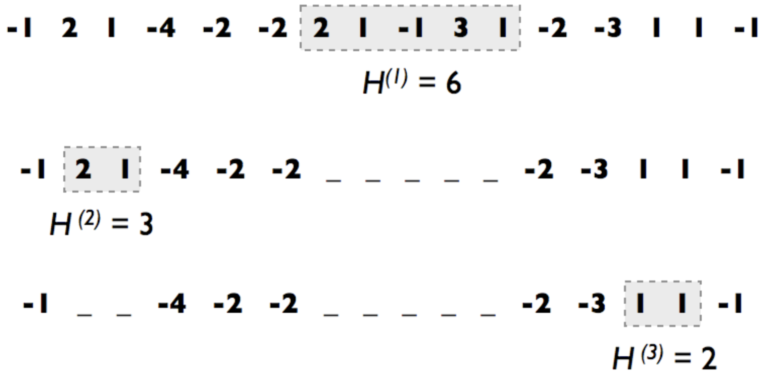
A classical way to consider the detection of accumulations in a sequence is to move a sliding-frame (or window) along the sequence; but it requires to specify the size of the frame which is in general not easy to make objectively. A natural and efficient alternative to sliding-frames is to use the Local Score statistic.

II - Local Score

Definition

A complete and formal definition of the Local Score is given in Guedj, Robelin et al (2006). Briefly, given a sequence of random variable ***X***, the positive Local Scores of this sequence correspond to the sub-sequences (or regions, segments...) with positive sums of individual test statistics. The first Local Score ($H^{(1)}$) corresponds to the value of the region (called the best region) with the maximal sum of test statistics. Since we do not want the different regions to overlap, the second Local Score ($H^{(2)}$) is defined as Local Score of the sequence, disjoint from the first Local Score and more generally the *k*-th Local Score ($H^{(k)}$) is defined as the Local Score of the sequence disjoint from the *k*-1 best Local Scores ($H^{(1)}$,..., $H^{(k-1)}$).

Example



You can notice the regions can not be increased or decreased without reducing at the same time the value of the corresponding Local Scores. Moreover the sequence must be negative on average, otherwise, the best region would easily span the entire sequence.

Statistical Significance

We know from the extrem-values theory that the Local Scores follow a Gumbel distribution under several restrictive assumptions. Most of the probalistic results about the Local Score are recalled in Guedj, Robelin et al (2006). See Nuel (2006) for an exact computation.

In LHiSA, calculation of p-values is driven through Monte-Carlo simulations to take into account the potential amount of LD between the markers.

Algorithms

Several algorithms have been proposed to find the best Local Score and the set of positive Local Scores in the most effective way. To date, the ‘best’ and fastest algorithm (which is used in LHiSA) has been proposed by Ruzzo and Tompa (1999).

Selection of the regions:

to do

III - Application to large-scale association studies

In large-scale association studies such as genome-wide association studies (GWAS) or studies on large chromosomal regions (linkage regions for instance) the Local Score has been proposed as a fast, simple tool to detect associated region at the first-stage of studies. In Guedj, Robelin et al (2006), the authors outline the assets of the method over single-marker analyses in a power study and an application to a linkage region for schizophrenia. In Aschard, Guedj et al (2007), the authors propose to use the Local Score at the first step of a two-stage multi-marker analysis and show on the GAW15 data that this approach performs better than traditionnal marker-based multi-stage studies.

In association studies, the sequence ***X*** is based on the single-marker test statistic considered, for each marker, along with the genome.

The delta parameter

Since association statistic are generally positive, and the Local Score approach require to work on a sequence that is on average negatice, we need to decrease the whole sequence by a constant *delta*. This controled parameter of the method corresponds actually to the level upon which we consider that a test statistic should positively contribute to the Local Score. *Delta* is generally set to the 1, 5 or 10% level. Within this range of values, it does not sensibly change the nature of the results (Guedj, Robelin et al 2006).

Multiple-testing

By considering regions instead of single markers, the Local Score reduces the number of tests from *n* markers to *k* regions with positive Local Scores and hence contributes to reduces at the same time the multiple-testing problem. In addition all the methods proposed to select the final set of regions controled the genome-wide type-I error-rate (or family-wise error-rate) to the specified level.

Flexibility of this approach

The flexibility of LHiSA relies on the choice ot the test statistic at the basis of ***X*** which allows to consider any phenotype (disease, severity, GxE interactions, LD between to contigous markers ...) and any data structure (families, case-control, case-only...).

IV - Extension to replicated association studies

Local Replications: definition

In both linkage and association studies, replication of initial findings in independent populations has been put forward as the gold standard for results validation in order to filter false positives from true signals. Historically, association has referred primarily to marker association, implicating the marker as the basis unit of the analysis. But with the increasing marker density and the use of an indirect approach to association through Linkage Disequilibrium, association is now often considered at the haplotypic level. Consequently, the current tendency is to perform replications on the basis of the marker or the haplotype.

However in practice, such replications are generally difficult to obtain. Among the different possible causes, lack of power, multiple-testing, genotyping-error, missing-values and population stratifications are often invoked.

But beside these study-design and data-analysis-related factors, inconsistent findings may also result from real biological differences between populations. The complex nature of etiologies under investigation including differences in allele frequencies, allele and locus heterogeneity as well as a high degree of variation in the strength of LD among populations of different origins, is a major challenge for the discovery of disease susceptibility loci. As a result, a given locus may have very different patterns of association accross different populations.

In this context, the marker or haplotype-based analysis can appear limited and we believe that this problem can be reduced by considering Local Replications instead of strict replications of markers or haplotypes. Indeed, in genetic association studies, we can expect to observe an accumulation of high statistics of association around a disease susceptibility locus. Such an accumulation may be due to: (1) Linkage Disequilibrium with surrounding markers, or (2) to an agregation of several susceptibility loci in a same genomic location. Consequently, these accumulations may be locally replicated across populations without restraint about the specific allele or pattern of alleles to be replicated. So we define a Local Replication as the presence of a local accumulation of high statistics of association in a given genomic region, which is replicated among the different populations.

Local Score applied to Local Replications

In this situation again, the Local Score appears as a natural framework to deal with Local replication. The signal ***X*** to which the Local Score is applied result from a combination of the different population, the simplest combination being the sum of test statistics through the populations for each marker (implemented in lhisa).

V - Other applications

Deleted chromosomal regions

Deleted chromosomal regions result in a high proportions of homozygotes for the corresponding markers. The Local Score can easily applied to detect such regions.

Localization of blocs of LD

When ***X*** is based on a LD mesure between two contigous markers for all the markers of the dataset, the Local Score can be used to delimit blocks of LD.

VI - LHiSA in R

LHiSA is implemented in a R function called **lhisa()**. It is less fast than the C++ implementation (section VII) but it is much more easy to use and flexible.

>> Input parameters

'**x**' is a sequence of random variables (a numerical vector in R) in which we want to search for Local Scores.

```
> x
 [1] 0.37645872 0.25000349 0.41333509 0.45063096 0.59252997 0.87846796
 [7] 0.01767701 0.81648020 0.17573147 0.86087791
```

'**geno**' and '**pheno**' are used instead of '**x**' if you want to use genetic data from which '**x**' is derived by **lhisa()**. The advantage to use 'geno' and 'pheno' over '**x**' is to take the pattern of LD between the markers into account. The advantage to use '**x**' is to run the algorithm faster and to apply **lhisa()** on other dataset than genetic association studies.

Genotypes (for instance 0 for *aa*, 1 for *aA*, 2 for *AA* and NA for missing values) are stored into '**geno**' with one column for each marker. Phenotypic information such as the case-control status (*e.g.* 0 for controls and 1 for case) are stored in '**pheno**'. It can contain more information if required but the first column must correspond to the trait to study (case-control for instance). Each row in '**pheno**' and '**geno**' corresponds to one subject and they should be R data.frames.

```
> geno
  SNP1 SNP2 SNP3
1   NA    0    2
2    2    2    2
3    2    1    2
4    0    2    0
```

```
5    1   1   0

6    2   2   2

7    2   1   0

8    1   0   1

9    0   1   2

10    1    1    2

> pheno

   status age  ...

1       1  21

2       0  39

3       0  39

4       0  45

5       0  37

6       0  23

7       1  29

8       0  23

9       1  24

10      1  35
```

‘**info**’ is a data.frame containing additional information on the markers (null by default) such as chromosome, position on the chromosome in bp ... The number of rows in ‘**info**’ should correspond to the number of columns in ‘**geno**’.

‘**pop**’ is a vector of factors. It is NULL by default. If specified, **lhisa()** perform Local Replications by stratifying the data according to the value of each subject in ‘**pop**’. By default the combination is made *via* a sum. In the futur, the user will be allow to specify its own way to combine signals from different populations by using the ‘**my.pop**’ and ‘**my.sat**’ options in ‘**association.test**’ >> not implemented in the current version.

>> Optional parameters

‘**coding**’ corresponds to the possible genotypes to find in geno (NA not included). By default it is (0,1,2) for *aa*, *aA* and *AA*.

‘**association.test**’ stands for the test statistic applied to each marker in order to contruct the sequence ‘**x**’ if ‘**geno**’ and ‘**pheno**’ are specified instead of ‘**x**’ directly. It can be a string to use the function available in **lhisa()** or a function specified by the user which apply to ‘**geno**’ and ‘**pheno**’. Test statistics implemented in lhisa() are the fast exact and unbiased allelic test (“fuea.test”, requires to load the “allelic” R package) proposed by Guedj, Wojcik et al (2006), the Fisher exact test (“fisher.test”) or the chi-square test (“chisq.test”) applied on genotypes.

Using an association test specified by the user make lhisa very flexible. However it makes the implementation of lhisa more complex and consequently, all the association should have the same R form and input parameters even if all are not used, should apply to one marker of ‘**geno**’ and should return a p-value for this marker :

```
my.association.test = function(my.geno, my.pheno, my.coding, other.geno, my.pop,
my.sat) {...}
```

‘**my.geno**’ stands for the column of ‘**geno**’ to treat, ‘**my.pheno**’ = ‘**pheno**’ in **lhisa()**, ‘**my.coding**’ = ‘**coding**’ in **lhisa()** and ‘**other.geno**’ = ‘**geno**’ in **lhisa()** if information on other markers are required. ‘my.pop’ and ‘my.sat’ are not to use at the moment.

‘**p.value**’ is TRUE (T) or FALSE (F, by default) if you want to compute p-values of Local Scores by Monte-Carlo. Applied on ‘**x**’ it consider each marker as independent. Applied on ‘**geno**’ and ‘**pheno**’ it takes the pattern of LD into account.

‘**B**’ is the number of Monte-Carlo simulation used to calculate p-values (2,000 by default).

‘**delta**’ is the parameter defined in section III (5% by default).

‘**selection**’ is the selection method used to select a subset of significant regions at a given significance level specified in ‘**level**’.

‘**level**’ is the global significance level to reject or not the null hypothesis that there is no region associated to the trait under investigation.

>> Ouput

lhisa() returns a R list (let say ‘**toto**’) containing the following items:

‘**toto$losc**’ is the set of Local Scores. ‘**start**’ corresponds to the first marker of the region and ‘**end**’ to the last one. ‘**H**’ is the Local Score and ‘**T**’ the cumulative sum of the first Local Scores. ‘**p.values**’ gives the p-value for each region.

‘**toto$x**’ is the signal ‘**x**’ on which the method has been applied (taking the parameter ‘**delta**’ into account).

‘**toto$raw.x**’ is the raw signal not decreased by ‘**delta**’, computed from ‘**geno**’ and ‘**pheno**’ or directly given by the user in the ‘**x**’ input parameter.

‘**toto$info**’ recall the ‘**info**’ input parameter is specified by the user. ‘**start**’ and ‘**end**’ in ‘**toto$losc**’ correspond in rows in ‘**toto$info**’. Specifying ‘**info**’ allows to easily determine the chromosome and genomic position (in bp for instance) for each high-scoring region.

VIII - References

Website

http://stat.genopole.cnrs.fr/software/lhisa

First application of the Local Score to Genetic Epidemiology

- Guedj, Robelin et al. Detecting local high-scoring segments: a first stage approach to genome-wide association studies. 2006. *Stat. Appl. Genet. Mol. Bio*. **5:** 22

Other applications in Genetics and Genomics

- Guedj, Wojcik et al. Catching local replications: a local score-based approach to replicated association studies. 2007. In proceedings of EMGM (Heidelberg) and IGES (York).

- Aschard, Guedj et al. A multiple-marker two-step approach to genome-wide association studies. 2007. *BMC Genetics*. [in press]

- Karlin. Statistical signals in Bioinformatics. 2005. *PNAS*. **102:** 13355-13362.