BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 1 2009, pages 134–136
doi:10.1093/bioinformatics/btn581

Genetics and population analysis

Genetic association analysis with FAMHAP: a major program update

Christine Herold* and Tim Becker*
Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Sigmund-Freud-Str. 25, D-53105 Bonn, Germany

Received on August 18, 2008; revised on October 29, 2008; accepted on November 6, 2008
Advance Access publication November 9, 2008
Associate Editor: Alex Bateman

ABSTRACT

Summary: FAMHAP is an established software for haplotype association analysis of nuclear families. We have released a major update that comprises various new features for case-control data. Furthermore, we provide an additional program runFamhap that allows users to start the same method repeatedly for varying sets of genetic markers. In addition, a platform-independent graphical user interface (GUI) was developed to simplify the usage of both FAMHAP and runFamhap. The runFamhap program greatly facilitates the application of FAMHAP to genome-wide association studies (GWAS) and supports flexible genome-wide haplotype analysis. As an example, we describe application to HapMap data.

Availability: The software is available at http://famhap.meb.uni-bonn.de

Contact: herold@imbie.meb.uni-bonn.de; becker@imbie.meb.uni-bonn.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

FAMHAP (Becker and Knapp, 2004a) was initially developed for the estimation of haplotype frequencies in nuclear families. Haplotyped-based association test for nuclear families and case-control data were published subsequently (Becker and Knapp, 2004b; Becker *et al*., 2005). The goal of our new FAMHAP release is, first, to enlarge the range of the programs applicability in the context of association studies, and, second, to make the program easier to use for non-experts by providing a graphical user interface (GUI). We have implemented new methods for the gene-based analysis of case-control data, as well as single-marker analysis features for genome-wide association studies. Options to conduct standard single-marker tests, Armitage (1955) test for trend and the TDT (Spielman *et al*., 1993), are implemented in such a way that they allow easy application to genome-wide association studies (GWAS). The new program runFamhap allows users to start the same FAMHAP options repeatedly for varying marker sets.

2 METHODS AND IMPLEMENTATION

2.1 New methods for gene-based analysis

FAMHAP implements new gene-based analysis methods for data of case-control type. We have previously introduced Monte-Carlo (MC) simulation-based haplotype tests for case-control data allowing multiple comparison adjustment (Becker *et al*., 2005). In brief, the algorithm described there works as follows: first, haplotype frequencies are estimated from the joint case-control sample. Next, weighted haplotype explanation lists (WHLs), i.e. lists of possible haplotype assignments per individual with conditional likelihood weights w, are used to construct contingency tables. The weights w are the conditional probabilities of the different possible haplotype explanations of an individual, given the maximum likelihood haplotype frequency estimates. An omnibus or maximum statistic—can be computed from the contingency tables. In each replication, affection status is randomly permuted. The ratio of cases and controls is kept constant. P-values are computed as s/n, where n is the number of permutation replicates, and where s is the number of permutation replicates leading to a test statistic higher than or equal to that of the real data. By altering the contingency table or simulation scheme, the algorithm can now be applied in additional situations.

Analysis of diplotypes: a diplotype is a multi-marker genotype with phase information, or, in other words, a pair of haplotypes. The WHLs described above contain information on diplotypic status and can directly be used to build up contingency tables for diplotypes. The same MC simulation procedure as used for haplotypes then naturally yields a valid test for diplotypes.

Testing case sib-pairs against control sib-pairs: this feature can be used to compare a sample of affected sib-pairs against a sample of unaffected sib-pairs. We have implemented a modified MC simulation scheme that accounts for within sibship dependency of diplotypic status in regions of linkage. Affection status is permuted or not with equal chance simultaneously for siblinks.

Using discordant sib-pairs for association testing: for discordant sib-pairs, the MC procedure was modified in such a way, that affection status is interchanged between the discordant sibs with a probability of 0.5.

We conducted a simulation study under the null hypothesis H_0: 'linkage but no association' and did not find deviations from the nominal level for any of the new methods (data not shown).

Furthermore, we have implemented an association test that is useful in the presence of multiple, rare variants within one gene. The method complements analysis methods that are powerful under the common-disease common-variants hypothesis (Lander, 1996). The idea of our rare variants test is that in the presence of multiple rare variants there should be an excess of rare haplotypes in cases. The presence of such an excess can be tested using the following algorithm:

I. Haplotype frequencies are estimated from the compound sample of cases and controls. To each individual the list of its possible haplotype explanations with respective conditional likelihood weights w is assigned.

II. For each cutoff x, $0 < x < 1$, the class $H_{\leq x}$ of haplotypes with a frequency less than or equal to x and the class $H_{>x}$ of haplotypes with a frequency greater than x are considered. Furthermore, a 2×2 contingency table T_x whose rows correspond to the two classes and whose columns correspond

*To whom correspondence should be addressed.

to case/control status is constructed. Cell counts are computed by summing, according to case/control status, the per individual likelihood weight w of a haplotype assignment h_1, h_2 into class $H_{\leq x}$, if both h_1 and h_2 have a frequency lower than x, by summing w into $H_{>x}$ if both haplotypes have a frequency greater than x, and by summing $w/2$ into both classes if exactly one of h_1, h_2 has a frequency greater than x. The corresponding standard test statistic t_x for 2×2 tables is computed. Note that only a finite number of values for x has to be considered, since there is only a finite number of haplotypes.

III. A test statistic t is defined by $t := \max_x t_x$.

IV. The distribution of t is determined using MC simulations. For each permutation replicate i of these simulations, case/control status is randomly permuted such that the ratio of cases to controls is kept constant. For the simulated data, $t_i = \max_x t_{x,i}$ is computed from the contingency tables $T_{x,i}$, which are constructed using the weights from step I. Finally, the P-value is computed as $P = \#\{i : t_i \geq t\}/n$.

By construction, the algorithm returns an optimal cut-off x and accounts for the optimization by evaluating the maximum statistic $t := \max_x t_x$ within the MC framework.

2.2 The runFamhap program

With the new runFamhap program it is possible to run the same analysis option repeatedly, for varying marker combinations (subsets of markers that shall be analyzed simultaneously). Data file, map file, analysis method, size of the marker combination and maximal marker distance in kilobases can be specified using the graphical user interface (GUI). An obvious application to a GWAS is to conduct a haplotype analysis for all pairs of SNPs less than, for instance, 30 kb apart. Thus, haplotype analysis that is more flexible than a sliding window approach is possible. In particular, sets that leave out in between SNPs can be considered. This is useful, since SNPs that are younger than the disease mutation event will only 'split up' the primary haplotype the disease variant arose on. Leaving out the younger SNPs, however, may restore the full association signal. Since in practice the relative age of the SNPs is unknown, the implementation considers all sets of SNPs that come from the same region in terms of kilobase distance. Thus, our philosophy is to do too many tests and to adjust for multiple testing, rather than to do too few tests.

2.3 The GUI

The GUI (Supplementary Figure 1) is implemented in the object orientated language C#, which was developed for the .NET-platform. In order to start the GUI, Microsoft .NET has to be installed on Windows-machines. Detailed installation and running instructions, also for Linux users, are available at our website.

The GUI is divided into eight parts, which are separated by boxes with capital letter headings. The box on top of the GUI is important to guarantee that FAMHAP runs on different operating systems without problems. First, the operating system, Windows or Linux, has to be selected and additionally the Dos version (32-Bit or 64- Bit) can be specified.

In the first box the input file and, if necessary, a map file can be selected. The map file is only needed if we use the additional program runFamhap. General options are shown in the second box and can be selected together with the different association analysis methods. There is also the opportunity to set the allowed missing rate per person, separate the data by sex, modify the estimation mode of haplotype frequencies and improve the output, e.g. choose the kind of decimal separator or more detailed output. Another useful box is the box 'selected markers'. Here, certain markers from the input file, identified by their number of occurrence, can be selected.

The main part is the middle box which lists all association methods that are implemented in FAMHAP. To have a better overview, this box is divided into different smaller boxes which show the different kinds of association analysis: single-marker analysis, likelihood-ratio tests, MC methods and analysis of imputed SNPs. Each of these methods can be combined with the general options and additional parameters. To avoid senseless combinations of options, the methods and parameters which cannot be combined will appear in gray.

The command line that will be executed by FAMHAP is created by selecting the different methods, options and parameters on the GUI. Before starting FAMHAP, it can be useful to examine the command line via the button 'COMMAND LINE: SHOW' to check that the correct input file and desired kind of analysis have been selected.

The 'RUN' box gives the user two options to start the program: besides the conventional option which starts the program once, the new program runFamhap allows the user to start FAMHAP repeatedly by choosing the required marker combinations and distances between the markers.

The output files can be opened by selecting a certain type of output (P-values, haplotypes, WHLs, tagging markers, Mendelian errors) and an editor.

3 RESULTS AND DISCUSSION

According to Clark (2004), the genetic variation of a population is intrinsically organized into haplotypes. Therefore, we investigated whether genetic population differences between Chinese (CHB) and Japanese (JPT) can be found in the HapMap (The International HapMap Consortium, 2007), using FAMHAP to conduct single marker and haplotype-based tests based on chromosome 22 (build 35) SNP data. After removing mono-allelic SNPs, SNPs with a call-rate of <95% in at least one of the groups and SNPs with significant deviation from Hardy–Weinberg equilibrium ($P \leq 0.001$) in at least one of the groups, 28 867 SNPs remained for analysis. Table 1 contains all QC-SNPs with a P-value $< 3 \times 10^{-5}$. The best P-value ($= 7.89 \times 10^{-6}$) was obtained for rs5762375. After Bonferroni correction, this result was not significant ($P = 0.228$).

Table 2 contains the best results of a chromosome-wide haplotype analysis of all pairs of SNPs <50-kb apart. We compared haplotype P-values (P_{hap}) to the best single-marker P-value (P_{best}) of the linkage disequilibrium (LD) region in order to judge whether the haplotype P-value could be regarded as an improvement. Criteria for inclusion into Table 2 were either a haplotype $P < 1 \times 10^{-7}$ or that the quotient $\frac{p_{hap}}{p_{best}}$ was greater than 1000. We decided to consider this quotient in order to provide local comparison of single-marker analysis and haplotype analysis. In practice, a smaller threshold than 1000 could be useful, for instance, a threshold equal to the number of haplotype tests divided by the number of single-marker tests. The two-marker-haplotype distribution of rs226507 and rs11704481 (LD region 5) gave the best haplotype P-value ($P = 7.65 \times 10^{-9}$). Of note, these SNPs are not direct neighbors. The haplotype P-value was better by a factor of 2026.14 than the best single-marker P-value of the region. Moreover, the result remained significant after correction for 662 705 tests that were performed ($P = 0.0051$). The haplotype P-values found in regions 2 and 3 also withstood correction ($P = 0.01$ and $P = 0.028$). Regions 1, 4 and 6 were not significant after correction, but reflect interesting results since those regions would not have been prioritized for further investigation based on single-marker results.

Although we do not know if the regions we identified are of biological relevance, they support the idea that our haplotype approach can lead to the identification of additional disease genes. Indeed, an intrinsic organization into haplotypes in the sense of Clark (2004) will exist also in 'true' case and control populations. We note that other software packages are not as flexible as FAMHAP when it comes to the choice of marker sets. Even PLINK

Table 1. Excerpt of output file obtained with FAMHAP option singlecc

SNP ID	P_HWE_Ca	P_HWE_Co	P_Armitage	A_Ca	A_Co	OR_A	left_A	right_A
rs16985013	0.664	0.556	1.08E-05	0.522	0.189	4.694	2.400	9.178
rs5762375	**0.35**	**0.9**	**7.89E-06**	**0.878**	**0.568**	**5.458**	**2.556**	**11.657**
rs6001900	0.202	0.905	2.58E-05	0.163	0.467	0.222	0.110	0.450
rs2205661	0.719	0.137	1.55E-05	0.244	0.567	0.247	0.130	0.471
rs136612	0.93	0.0678	2.29E-05	0.221	0.523	0.259	0.134	0.501

'Ca' stands for cases, 'Co' for controls. The last three columns contain odds ratio and confidence interval. The best result is marked bold.

Table 2. Best haplotype results in physical order

LD region	SNP ID 1	SNP ID 2	haplotype *P*-value[a]	best SNP *P*-value[b]	Quotient[c]
1	rs4822896	rs5762224	2.71E-07	0.000716	2642.07
2	rs16984994	rs12628574	1.56E-08	0.000011	692.31
3	rs5762375	rs5762379	4.19E-08	0.000008	188.31
4	rs9626680	rs8142684	1.13E-06	0.003630	3212.39
5	rs226507	rs11704481	7.65E-09	0.000016	2026.14
6	rs8135489	rs7410305	2.05E-07	0.000332	1619.51

[a] *P*-value obtained with omnibus likelihood-ratio test (3 degrees of freedom).
[b] Best *P*-value for a single SNP (Armitage's trend test, 1d.f.) within the LD region.
[c] Quotient of columns haplotype *P*-value and best SNP *P*-value.

(Purcell *et al.*, 2007), which is in general much more comprehensive than FAMHAP, supports only sliding windows and the analysis of predefined haplotype lists.

In order to estimate running time, we reanalyzed a GWAS (643 individuals and 550 000 SNPs) published by Hillmer *et al*. (2008). Haplotype analysis of chromosome 1 (38 635 SNPs; 363 711 SNP pairs) took ∼17 h and analysis of chromosome 22 (7792 SNPs; 89 320 SNP pairs) took ∼2 h on a Windows 64-Bit desktop computer with 2.66 GHz. Genome-wide haplotype association analysis with FAMHAP is thus computationally feasible, even for larger samples and marker panels. Nevertheless, our application is slower than PLINK. This is presumably due to the fact, that FAMHAP does not load the complete data into the working memory, but operates with a program that makes repeated calls to FAMHAP. This is an advantage when working memory is a limitation, at the price of running time.

Finally, we note that our software does not implement methods for case/control data to adjust for population stratification. Thus, if population stratification is a concern, additional *post hoc* validation with software like EIGENSTRAT (Price *et al*., 2006) should be considered.

Funding: Deutsche Forschungsgemeinschaft (BE 3828/3-1).

Conflict of Interest: none declared.

REFERENCES

Armitage,P. (1955) Tests for linear trends in proportions and frequencies. *Biometrics*, **11**, 375–386.

Becker,T. and Knapp,M. (2004a) Maximum-likelihood estimation of haplotype frequencies in nuclear families. *Genet. Epidemiol.*, **27**, 21–32.

Becker,T. and Knapp,M. (2004b) A powerful strategy to account for multiple testing in the context of haplotype analysis. *Am. J. Hum. Genet.*, **75**, 561–570.

Becker,T. *et al*. (2005) Multiple testing in the context of haplotype analysis revisited: application to case-control data. *Ann. Hum. Genet.*, **69**, 747–756.

Clark,A.G. (2004) The role of haplotypes in candidate gene studies. *Genet. Epidemiol.*, **27**, 321–333.

The International HapMap Consortium. (2007) A second generation human haplotype map of over 3.1 milllion SNPs. *Nature* **449**, 851–862.

Hillmer,A.M. *et al*. (2008) A genome-wide association scan identifies new susceptibility variants for male pattern baldness on chromosome 20p11. *Nat. Genet.*, (in press).

Lander,E.S. (1996) The new genomics: global views of biology. *Science*, **274**, 536–539.

Price,A.L. *et al*. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.

Purcell,S. *et al*. (2007) PLINK: a tool set for Whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

Spielman,R.S. *et al*. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.*, **52**, 506–516.