BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 24 2009, pages 3275–3281
doi:10.1093/bioinformatics/btp596

Genetics and population analysis

INTERSNP: genome-wide interaction analysis guided by a priori information

Christine Herold[1,*], Michael Steffens[1], Felix F. Brockschmidt[2,3], Max P. Baur[1,4] and Tim Becker[1,*]

[1]Institute for Medical Biometry, Informatics and Epidemiology, University of Bonn, Sigmund-Freud-Str. 25, [2]Department of Genomics, Life & Brain Center, University of Bonn, D-53105 Bonn, [3]Institute of Human Genetics, University of Bonn, D-53111 Bonn and [4]German Center for Neurodegenerative Diseases (DZNE), D-53105 Bonn, Germany

Received on August 3, 2009; revised on September 4, 2009; accepted on September 23, 2009
Advance Access publication October 16, 2009
Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: Genome-wide association studies (GWAS) have lead to the identification of hundreds of genomic regions associated with complex diseases. Nevertheless, a large fraction of their heritability remains unexplained. Interaction between genetic variants is one of several putative explanations for the 'case of missing heritability' and, therefore, a compelling next analysis step. However, genome-wide interaction analysis (GWIA) of all pairs of SNPs from a standard marker panel is computationally unfeasible without massive parallelization. Furthermore, GWIA of all SNP triples is utopian. In order to overcome these computational constraints, we present a GWIA approach that selects combinations of SNPs for interaction analysis based on a priori information. Sources of information are statistical evidence (single marker association at a moderate level), genetic relevance (genomic location) and biologic relevance (SNP function class and pathway information). We introduce the software package INTERSNP that implements a logistic regression framework as well as log-linear models for joint analysis of multiple SNPs. Automatic handling of SNP annotation and pathways from the KEGG database is provided. In addition, Monte Carlo simulations to judge genome-wide significance are implemented. We introduce various meaningful GWIA strategies that can be conducted using INTERSNP. Typical examples are, for instance, the analysis of all pairs of non-synonymous SNPs, or, the analysis of all combinations of three SNPs that lie in a common pathway and that are among the top 50 000 single-marker results. We demonstrate the feasibility of these and other GWIA strategies by application to a GWAS dataset and discuss promising results.

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

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

1 INTRODUCTION

As predicted by Risch and Merikangas (1996), genome-wide association studies (GWAS) carried out during the last years have lead to the identification of hundreds of loci associated with various complex diseases (Altshuler *et al*., 2008). However, since the effect sizes of the findings are typically small, a large portion of the genetic contribution to the phenotypes remains unexplained (Maher, 2008). Besides rare variants, so far undetected SNPs with even smaller effect size, and various other reasons, the 'missing' genetic variation could be explained by genetic interaction. Therefore, genome-wide haplotype analysis (GWHA) (Becker and Herold, 2009; Tregouet *et al*., 2009) and genome-wide interaction analysis (GWIA) are compelling next steps in the analysis of GWAS. Marchini *et al*. (2005) demonstrated that, in the presence of multi-marker disease models, GWIA can lead to increased power as compared with single-marker approaches. However, GWIA is computationally challenging. With one million SNPs, 5×10^{11} SNP pairs have to be tested for interaction. The computation of such a number of test statistics is possible when the test statistic is available in closed form (Marchini *et al*., 2005), but extraction of the respective number of contingency tables from an input file is not practicable on standard computers. We estimated a running time of 120 days for a complete GWIA with 550 000 SNPs and 1200 individuals on a 3 GHz computer. As a consequence, only massive parallelization will render a complete two-marker GWIA strategy feasible. Thinking ahead, a complete three-marker strategy would require 1.67×10^{17} tests and is undoubtedly unfeasible. An obvious way to overcome these limits is to analyze only 'interesting' combinations of SNPs (pairs or triples), selected based on an increased prior to be involved in the disease. Such priorities can be defined by statistical evidence (single-marker *P*-value in own data), genetic impact (genomic location) and potential biological relevance (SNP function class or pathway information). Here, we introduce the software INTERSNP that allows conduction of meaningful case/control GWIA strategies that make use of such a priori information.

2 METHODS

2.1 Quality control

Stringent quality control (QC) is of particular importance with joint analysis of multiple markers, since erroneous genotypes for just one of the analyzed SNP may already invalidate the analysis. We implemented the QC-criteria missing rate, deviations from Hardy–Weinberg equilibrium (HWE) and the inflation factor λ as defined by Devlin *et al*. (2004). We use an iterative QC-algorithm to address missing genotypes. The starting point is the average genotype missing rate, taken over all SNPs and individuals. In every iteration,

*To whom correspondence should be addressed.

alternately either SNPs or individuals with a missing rate worse than the average missing rate plus a user-defined missing rate difference (*mrdiff*) are discarded. Then, the new average missing rate is calculated and further SNPs or individuals are deleted, when their missing rate is higher than the new missing rate plus *mrdiff*. The algorithm terminates when there are no SNPs or individuals left that have to be discarded. Thereafter, SNPs which are not in HWE in either cases or controls are removed. The respective *P*-value thresholds for HWE are specified by the user. Furthermore, population stratification can be accounted for by treating the belonging to strata as a covariate in the logistic regression framework described below.

2.2 Tests

2.2.1 Single-marker analysis By default, a single-marker *P*-value is computed for all qc-SNPs. For the autosomes and the pseudo-autosomal region of the Y chromosome, the user can either choose Armitage's (1955) trend test or the genotype test with 2 degrees of freedom (d.f.). Y-chromosomal markers are evaluated with the χ^2-test for the 2×2 table of allele counts in male individuals. For X-chromosomal SNPs, we use the allele-based test with 1 d.f. suggested by Clayton (2008). Clayton's approach guarantees that hemizygote males and homozygote females contribute equally to the test statistic, reflecting the fact that only one of the X-chromosomes is active in females.

2.2.2 Multi-marker analysis When it comes to simultaneous analysis of SNPs, a crucial question is whether to test *for* interaction or whether to use a 'full test', i.e. a test that includes the marginal SNP effects into the analysis. In the latter case, we obtain a test that does not explicitly test *for* interaction, but a test that should be powerful in the presence of interaction. Since it is difficult to judge which strategy should be advocated, our software supports both strategies. When the goal is detection of genes involved in the etiology of a diseases, a full test can be useful, because it seems reasonable to assume that interacting genes also show some marginal effects. Even when those marginal effects are small, their inclusion into the statistical analysis improve the detection of respective SNP pairs. This has been shown by Marchini *et al*. (2005), who considered a full genotype test. A drawback of the full test strategy is that (strong) association of a particular SNP can render pairs including that SNP significant, even if the other SNP is neither marginally associated with the disease nor an interaction partner. However, this problem can be overcome by filtering the output: for each SNP, INTERSNP lists only the top 50 SNP combinations including the SNP.

On the other hand, an explicit test for interaction can be advocated when, either detection of interaction *per se* is the research goal, or when a particular region or SNP is already known to be associated. In this case, screening for interaction partners using an explicit interaction test can help to find further genes involved in the disease. INTERSNP implements various multi-marker tests that are summarized in Table 2 (tests 1–12):

Association test with genotype contingency table (test 1): for two SNPs, there are $3\times 3=9$ two-marker genotypes. We consider the respective $3\times 3\times 2$ contingency table of counts in cases and controls. Let T be the standard test statistic for contingency tables, i.e. $T=\sum_{i,j,k}\frac{(O_{i,j,k}-E_{i,j,k})^2}{E_{i,j,k}}$, where $i,j\in\{1,2,3\}$, $k\in\{1,2\}$, where $O_{i,j,k}$ is the observed number of counts in cell (i,j,k) and where $E_{i,j,k}$ is the number of counts in cell (i,j,k) expected under the null hypothesis. T is χ^2-distributed with 8 d.f. When three SNPs shall be considered simultaneously, we have 27 three-SNP-genotypes and we obtain a test statistic that is χ^2-distributed with 26 d.f. For SNP combinations containing at least one X-chromosomal marker, we consider the genotype contingency tables of males and females separately and sum up the respective test statistics and d.f. to obtain a joint *P*-value that is not biased by different ratios of cases and controls within the male and female group.

Although the contingency table test is not particularly sophisticated, we included it into the program since it is very easy to compute, and therefore can serve for screening purposes.

Test for interaction using a log-linear model (test 2): here, the observations x_{ijk} of the $3\times 3\times 2$ contingency table can be modeled by fitting a log-linear regression model of expected cell counts m_{ijk} to the cell entries (Bishop *et al.*, 2007). The model equation is (1) $\log(\hat{m}_{ijk})=u+u_{1(i)}+u_{2(j)}+u_{3(k)}+u_{12(ij)}+u_{13(ik)}+u_{23(jk)}$ without the term $u_{123(ijk)}$. Testing the null hypothesis H_0: $u_{123(ijk)}=0$ yields an explicit test for interaction. With maximum likelihood estimates $\hat{m}_{ijk}$ of the cell counts under Equation (1), we obtain the test statistic $T=-2*(\sum_{i,j,k}x_{ijk}*\log(m_{ijk})-\sum x_{ijk}\log(x_{ijk}))$ that is χ^2-distributed with 4 d.f. Computation of the number of d.f. in the presence of empty cells is straightforward. Based on the starting values $\hat{m}_{ijk}^{(0)}=1$, the maximum likelihood estimates $\hat{m}_{ijk}$ can be obtained iteratively:

$$\hat{m}_{ijk}^{(1)}=\hat{m}_{ijk}^{(0)}*\frac{X_{ij+}}{\hat{m}_{ij+}^{(1)}},\hat{m}_{ijk}^{(2)}=\hat{m}_{ijk}^{(1)}*\frac{X_{i+k}}{\hat{m}_{i+k}^{(1)}},\hat{m}_{ijk}^{(3)}=\hat{m}_{ijk}^{(2)}*\frac{X_{+jk}}{\hat{m}_{+jk}^{(2)}},$$

$$\hat{m}_{ijk}^{(4)}=\hat{m}_{ijk}^{(3)}*\frac{X_{ij+}}{\hat{m}_{ij+}^{(3)}}\text{ etc.}$$

The iteration usually converges quickly. Thus, log-linear models a recommendable method for large-scale applications. In this case of three SNPs, we obtain a test for genotypic 3-fold interaction with 8 d.f. SNP combinations containing at least one X-chromosomal marker are treated analogously to the way described for test 1 (summation of the test statistic for males and females).

Logistic regression (tests 3-12): Tests 1 and 2 can be computed quickly, but have limitations in their applicability. Therefore, we implemented logistic regression following the framework introduced by Cordell and Clayton (2002). Within this framework, it is possible to include or exclude marginal effects, distinguish allelic and genotypic tests and to adjust for covariates. We briefly described the logistic regression models we use. For further details, see the paper by Cordell and Clayton (2002). Let p_j be the probability that individual j is a case. We define $\text{logit}(p):=ln(\frac{p}{1-p})=\beta^T x$, where β is the vector of coefficients to be estimated and x is a vector that is coded depending on the genotypes as follows.

We consider three SNPs. For each SNP i, $i=1,2,3$, we model its allelic effect x_i by coding the genotypes (1,1), (1,2) and (2,2) as $x_i=-1,0,1$. Next, we model dominance effects $x_{i,D}$, $i=1,2,3$, as $x_i=-0.5,0.5,-0.5$ for the genotypes (1,1), (1,2) and (2,2), respectively. By multiplication, we obtain interaction terms, for instance, x_1x_2 represents allelic interaction between SNPs 1 and 2, while $x_{1,D}x_{2,D}$ represents interaction between the dominance terms of SNPs 1 and 2. Note that these interaction terms code interaction on an additive logit scale and, hence, on a multiplicative odds ratio scale.

Let β_0 be the intercept parameter that defines the baseline likelihood $L_0:=logit(p)=\beta_0$. Next, the likelihood $L_1^A:=\beta_0+\beta_1x_1$ models the allelic effect of SNP 1 and comparison to L_0 yields a likelihood ratio test with 1 d.f. Analogously, comparison of $L_1^G:=\beta_0+\beta_1x_1+\beta_{1,D}x_{1,D}$ to L_0 yields a genotypic test for SNP 1 with 2 d.f. In general, we let $L_{1,2}^A$ and $L_{i,j}^G$ denote likelihoods containing allelic terms, or, respectively, allelic and genotypic terms, for SNPs 1 and 2. In addition, let $L_{1,2}^{A,I}$ and $L_{1,2}^{G,I}$ be the likelihoods that also contain interaction terms, for instance, $L_{1,2}^{A,I}=\beta_0+\beta_1x_1+\beta_2x_2+\beta_{1,2}x_1x_2$ and $L_{1,2}^{G,I}=\beta_0+\beta_1x_1+\beta_{1,D}x_{1,D}+\beta_2x_2+\beta_{2,D}x_{2,D}+\beta_{1,2}x_1x_2+\beta_{1,2D}x_1x_{2,D}+\beta_{1D,2}x_{1,D}x_2+\beta_{1D,2D}x_{1,D}x_{2,D}$. The various likelihoods just introduced are summarized in Table 1.

Now, comparison of $L_{1,2}^A$ against L_0 yields a full allelic test with 3 d.f. (test 3), whereas comparison of $L_{1,2}^G$ against L_0 yields a full genotypic test with 8 d.f. (test 4). In order to test for allelic interaction, one compares $L_{1,2}^{A,I}$ against $L_{1,2}^A$ (1 d.f., test 5) and in order to test for genotypic interaction one compares $L_{1,2}^{G,I}$ against $L_{1,2}^G$ (4 d.f., test 6). Furthermore, it is possible to test for the additional effect of SNP 2, while controlling for the effect of SNP 1 by comparing $L_{1,2}^{A,I}$ against L_1^A (2 d.f., test 7), or, by comparing $L_{1,2}^{G,I}$ against L_1^G (6 d.f., test 8). These tests are summarized in Table 2 and serve as INTERSNP standard tests that can be called via their test number.

For three SNPs, likelihoods and tests can be formalized analogously. Here, we describe only allelic tests, but note that INTERSNP also allows

Table 1. Likelihoods

No.	Formula
L_0	β_0
L_1^A	$\beta_0+\beta_1 x_1$
L_1^G	$\beta_0+\beta_1 x_1+\beta_{1,D}x_{1,D}$
$L_{1,2}^A$	$\beta_0+\beta_1 x_1+\beta_2 x_2$
$L_{1,2}^G$	$\beta_0+\beta_1 x_1+\beta_{1,D}x_{1,D}+\beta_2 x_2+\beta_{2,D}x_{2,D}$
$L_{1,2}^{A,I}$	$\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_{1,2}x_1x_2$
$L_{1,2}^{G,I}$	$\beta_0+\beta_1 x_1+\beta_{1,D}x_{1,D}+\beta_2 x_2+\beta_{2,D}x_{2,D}+$ $\beta_{1,2}x_1x_2+\beta_{1,2,D}x_1x_{2,D}+\beta_{1,D,2}x_{1,D}x_2+\beta_{1,D,2,D}x_{1,D}x_{2,D}$

Table 2. Tests

Test No.	Test	Formula	d.f.	Comment
1	Chi-square-test		8	Full genotype test (contingency table)
2	Log-linear model	$l_{1,2}^{G,I}$ versus $l_{1,2}^{G}$	4	Test for genotypic interaction
3	Logistic regression	$L_{1,2}^{A,I}$ versus L_0	3	Full additive test
4	Logistic regression	$L_{1,2}^{G,I}$ versus L_0	8	Full genotype test
5	Logistic regression	$L_{1,2}^{A,I}$ versus $L_{1,2}^{A}$	1	Test for additive interaction
6	Logistic regression	$L_{1,2}^{G,I}$ versus $L_{1,2}^{G}$	4	Test for genotypic interaction
7	Logistic regression	$L_{1,2}^{A,I}$ versus L_{1}^{A}	2	Additional allelic effect of SNP 2
8	Logistic regression	$L_{1,2}^{G,I}$ versus L_{1}^{G}	6	Additional genotypic effect of SNP 2
9	Logistic regression	$L_{1,2,3}^{A,I}$ versus L_0	7	Full additive test
10	Logistic regression	$L_{1,2,3}^{A,I}$ versus$L_{1,2,3}^{A}$	4	Test for allelic interaction
11	Logistic regression	$L_{1,2,3}^{A,I}$ versus $L_{1,2,3}^{A,I_2}$	1	Test for 3-fold allelic interaction
12	Logistic regression	$L_{1,2,3}^{A,I}$ versus $L_{1,2}^{A,I}$	4	Additional allelic effect of third locus

explicit model specification, in particular genotypic 3-SNP tests. Let $L_{1,2,3}^{A}=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3$ be the three-SNP allelic likelihood, let $L_{1,2,3}^{A,I_2}=\beta_0+\beta_1x_1+\beta_2x_2+\beta_{1,2}x_1x_2+\beta_3x_3+\beta_{1,3}x_1x_3+\beta_{2,3}x_2x_3$ be the three-SNP allelic likelihood including all pairwise interactions and $L_{1,2,3}^{A,I}=\beta_0+\beta_1x_1+\beta_2x_2+\beta_{1,2}x_1x_2+\beta_3x_3+\beta_{1,3}x_1x_3+\beta_{2,3}x_2x_3+\beta_{1,2,3}x_1x_2x_3$ be the allelic likelihood including all pairwise and 3-fold interaction term. Then, testing $L_{1,2,3}^{A,I}$ against L_0 yields a full three-SNP allelic test with 7 d.f. (test 9), whereas testing $L_{1,2,3}^{A,I}$ against $L_{1,2,3}^{A}$ yields a test for 2- and 3-fold allelic interaction (4 d.f., test 10). Testing $L_{1,2,3}^{A,I}$ against $L_{1,2,3}^{A,I_2}$ yields a test for 3-fold allelic interaction (1 d.f., test 11). Finally, testing $L_{1,2,3}^{A,I}$ against $L_{1,2}^{A,I}$ yields a test for the additional allelic effect of SNP 3, while controlling for SNPs 1 and 2 (4 d.f., test 12).

For X-chromosomal markers, all dominance and dominance interaction terms are ignored. All tests can be combined with up to 10 covariates. In particular, adjustment for population strata, as derived from EIGENSOFT (Patterson *et al*., 2006; Price *et al*., 2006), for instance, is possible.

Note that at the moment only two- and three-marker analysis is implemented in INTERSNP. Since d.f. grow rapidly with even higher order interactions, it might be necessary to develop completely new statistical approaches that have enough power to detect those interactions.

2.3 Priorities

As already mentioned, complete two-marker GWIA requires computation resources that are usually not available, and complete three-marker GWIA is unfeasible. Therefore, two- or three-marker SNP combinations are selected for joint analysis based on user-specified statistic and genetic criteria.

2.3.1 Statistic criterion For each SNP a single-marker *P*-value is computed with Armitage's (1955) trend test from the own study data. Based on these *P*-values, a list of *n* top SNPs is computed. The length *n* of the list is specified by the user. The user specifies how many SNPs (0, 1, 2 or 3) of each combination shall be from the top-list.

2.3.2 Genetic criteria According to the criteria genomic location and function class, we classify SNPs into five nested groups of increasing genetic impact as follows: 0. *gene desert:* distance to nearest exon of nearest gene is >100 kb. 1. *close to gene:* distance to nearest exon of nearest gene is <100 kb or SNP lies within an intron of a gene. 2. *exon:* location within an exon of a gene. 3. *coding:* SNP lies in a coding region of a gene. 4. *non-synonymous:* SNP causes a non-synonymous amino acid exchange. The user specifies the required genetic impact and how many SNPs (0, 1, 2 or 3) of each combination shall have the selected, or a higher, impact. SNP annotation is derived from a respective annotation file that is loaded into the program. The Illumina® Human-610-chip annotation file can be directly used. For a detailed description of the annotation file format, we refer to the documentation on our web page.

2.3.3 Pathway information Pathway information is an obvious criterion to select SNPs for joint analysis. All analysis strategies (cf. next section) can be restricted to combinations of SNP that lie within genes that belong to a common pathway. Pathway information is provided by the user via a file that contains a list of pathways, together with the rs numbers of the SNPs that lie in genes from the respective pathway. In this way, expert researchers can restrict pathway-based interaction analysis to those pathways that are of potential relevance for the phenotype of interest. Since such expert knowledge is not always available, or often does not exist, we also support direct usage of all pathways from the KEGG (Kanehisa *et al*., 2006) database via a respect pathway file that can be downloaded from or web page or from the SNP ratio test homepage (O'Dushlaine *et al*. 2009; https://sourceforge.net/projects/snpratiotest/). Note, that we do not provide an explicit test for association of pathways with disease, in the sense of looking for an overrepresentation of associated SNPs in particular pathways, but that we use pathway information as a prior.

2.4 Typical GWIA strategies

Our software allows flexible combination of the selection criteria. In the following, we exemplify the usage by presenting various meaningful GWIA strategies.

(1) Single-marker analysis with Armitage's trend test.

(2) Analysis of all SNP pairs with at least one SNP among the top 10 single-marker results. (Recommended tests: 7, 8).
The idea of this strategy is to conduct a genome-wide search for possible interaction partners of the top single markers. This strategy is particularly meaningful when the top hits lie in confirmed disease loci. Therefore, we recommend to use tests 7 or 8 to test for the additional allelic or genotypic effect of further SNPs.

(3) Analysis of all pairs of SNPs which are among the top 1000 single-marker results. (Recommended tests: 1, 2, 3, 5, 6).
In this situation, tests 1–6 are useful. Test 1 tests the same hypothesis as test 4 (full genotype test) and test 2 tests the same hypothesis as test 6 (genotypic interaction). Since tests 1 and 2 are much faster to compute, we recommend to use them when no covariates shall be included into the analysis.

(4) Analysis of all pairs of SNPs which are among the top 50 000 single-marker results and also lie in a coding region of a gene. (Recommended tests: 1, 2, 3, 5, 6).

(5) Analysis of all pairs of non-synonymous SNPs. (Recommended tests: 1, 2, 3, 5, 6).

(6) Analysis of all pairs of SNPs which are among the top 5000 single-marker results and which lie in a common pathway. (Recommended tests: 1, 2, 3, 5, 6).

(7) Analysis of all SNP triples which are non-synonymous and which are among the top 10 000 single-marker results. (Recommended tests: 9, 10, 11).
In order to reduce d.f., we recommend to use allelic tests for three SNPs.

(8) Analysis of all SNP triples which are non-synonymous and which are among the top 100 000 single-marker results and which lie in a common pathway. (Recommended tests: 9, 10, 11).

2.5 Multiple testing adjustment

By default, Bonferroni correction with the number of conducted tests is provided. When combinations of SNPs from the single-marker top list are selected for joint analysis, such correction is sufficient when, either, a test for interaction is used (tests 2, 5, 6, 10 and 11), or when marginal effects are included only for those SNPs that are not required to be from the single-marker top list (strategy 2, tests 7 and 8). Thus, the marginal effects of the SNPs, the selection is based on, do not contribute to the test statistic, and there is no selection bias. However, when tests are used that include the marginal evidence into the test statistic (tests 1, 3, 4, 9 and 12), correction with the number of actually conducted tests could result in an anti-conservative procedure. Therefore, when tests including marginal effects are used and when combinations are selected based on single-marker evidence, we have to correct with the number of tests that would have been conducted without the single-marker criterion.

While Bonferroni correction avoids increased type I error, it is also known to reduce power since it ignores the correlation between tests that is caused by linkage disequilibrium (LD) and marker overlap. In order to improve adjustment for multiple testing, we have implemented Monte Carlo (MC) simulations for genome-wide application. With the MC-approach it is possible to account for the dependency of tests and to avoid conservativeness, while keeping correct type I error. We exemplify a valid GWIA MC-procedure by application to strategy IV (test 3) as follows:

(1) For each SNP compute its Armitage P-value.

(2) Determine the list of single-marker top hits (top-n-list).

(3) Compute joint P-value using test 3 for all pairs of SNPs, where both markers are in the top-n-list and both SNPs lie within the coding region of a gene.

(4) Construct a list of two-marker top hits, ordered by ascending P-value. Optionally, the best single-marker P-values can be included into the list when the goal is to adjust simultaneously for both single-maker and two-marker analysis.

(5) Do m simulations: permute case/control status such that the case/control ratio is kept constant. Conduct steps 1–4 for each simulated dataset. In particular, the SNP set of single-marker top hits differs between replicates. This is a must to mimic the selection process and to account for the fact that the marginal effects contribute to the test statistic of test 3.

(6) It is now possible to compute adjusted P-values for the best P-value of the real data, but also for k-th best P-value p_k. We compute the adjusted P-value as s/m, where s is the number of simulated datasets for which the best P-value is smaller than or equal to p_k and where m is the number of replicates. Note that, even for $k > 1$ it is necessary to compare p_k to the *best* P-value of the simulated dataset. Comparison with the k-th best P-value of the replicates could lead to the senseless situation that the corrected P-value for some p_k with $k > 1$ is better than the corrected P-value for p_1.

2.6 Implementation

INTERSNP is written in C/C++. Our data file format is identical to the transposed file set (tped/tfam format) used by PLINK (Purcell *et al.*, 2007). Analysis options (statistical tests, priorities) are specified by the user in a selection file. The user can also chose whether the complete data are read and stored in the computer's working memory, or whether for each SNP combination genotype information is reread from the respective lines of the input file. In the latter case, the genotype information is deleted from the working memory immediately after the test statistic has been computed. This technique guarantees the possibility to analyze huge datasets when computer's working memory is limited.

3 RESULTS

We analyzed a GWAS dataset that was recently published by Hillmer *et al.* (2008) and that reported a new locus for male pattern baldness on chromosome 20p11 with 296 cases and 347 controls. The strongest association signal was confirmed in an independent replication sample (combined $P = 2.7 \times 10^{-15}$). Here, we reanalyzed the initial GWAS (643 individuals), without the replication sample, and used only those 300 026 qc-SNPs that had a calling rate of at least 95% in both cases and controls. Note, that the lower calling rates for the remaining SNPs were not due to genotyping failure but arose since the sample was genotyped in batches with varying SNP content. Table 3 provides running time for strategies I–VIII for on a 3 GHz linux machine with 32 GB working memory.

Without MC-simulations, the majority of strategies is computable within a few minutes. The most demanding strategy we present (strategy V), requires only 7 min with the log-linear test (test 2). Computation time increases when logistic regression is used, since the maximization of the likelihoods requires repeated multiplication of matrices. Determinants of running time are the number of individuals and the number of parameters to be estimated. The latter is equal to the sum of parameters in the alternative and the null likelihood, and the number of individuals. As a consequence, computation time is highest with test 6, since it requires the computation of $L^G_{1,2}$ and $L^{G,I}_{1,2}$ which include 4 and 8 parameters, respectively. Still, even with strategy V, computation of test 6 is feasible in <11 h. Thus, for all strategies, it is possible to increase the number of tests by defining them based on a larger list of single-marker top hits.

Table 3 also includes running time estimates for 100 MC-simulations. We chose this number since it can give a first hint to judge whether a result is close to genome-wide significance or not. In practice, recomputation with a larger number of replicates is recommendable. Note that running time does not simply scale up by a factor of 100 since part of the running time of the real data goes into file procession and data storing and has to be done only once.

All MC-strategies are feasible, most of them need <1 day of computing time on a single computer. When test 1 or 2 are used, typically 1 h computing time is already sufficient. As before, the tests based on logistic regression take markedly longer, but remain feasible. Again, strategy V in combination with test 6 is most demanding, but about 1 month of running time still can be considered to be acceptable.

All running times in Tables 3 refer to the situation when the complete genotype data can be stored in the working storage of the computer. When we did not store the genotype data matrix, running time increased by a factor 2 on average and for none of the strategies running time increased by more than a factor of three. Thus, all strategies are also feasible when not enough computer

Table 3. Running time

Strategy	TEST	Number of tests[a]	Running Time	Running Time MC
I.	ATT[b]	300 326	1m31	7m42
II	7	3 003 205	33m34	66h18
	8		52m23	187h30
III	1	499 500	1m32	22m12
	2		1m37	33m25
	3		5m29	7h25
	5		8m25	13h20
	6		24m57	35h15
IV	1	835 728	2m14	74m6
	2		2m23	90m45
	3		8m33	11h50
	5		13m50	16h29
	6		42m22	50h24
V	1	14 180 475	4m30	6h21
	2		7m14	10h37
	3		93m21	149h7
	5		154m50	254h4
	6		645m2	798h45
VI	1	18 458	1m33	11m35
	2		1m35	12m10
	3		1m41	22m51
	5		1m50	31m42
	6		2m27	76m45
VII	9	632 710	14m20	25h24
	10		17m	29h27
	11		20m38	35h50
VIII	9	150 128	8m55	5h39
	10		8m49	5h42
	11		9m53	7h25

[a]Strategy I: number of single-marker tests. Other strategies: number of multi-marker tests, number of single-marker tests is not counted.
[b]Armitage trend test.

working storage is available. In particular, different jobs can be run in parallel.

We also checked the performance of INTERSNP with datasets with more individuals (data not shown). When test 1 or 2 are used, running time increases roughly by a factor equal to the relative increase in individuals. With the logistic regression models, the increase is higher, because the matrices involved in the computation have size (*number of individuals*) × (*d.f.*). We believe that several weeks of running time are an effort that has to be invested. Keeping in mind that study planning, patient recruiting and genotyping typically take several years, moderately long running time should not be a hindrance for extensive data analysis.

Although the focus of our article is on introducing GWIA strategies, we also wish to briefly present the most interesting association results, since they exemplify potential advantages of GWIA. Single-marker analysis is presented in Table 4 as background information. MC-corrected *P*-values are shown for 10 000 simulations. The first eight lines contain markers from the confirmed X-chromosomal locus (Hillmer *et al.*, 2008). The SNPs reach genome-wide significance after Bonferroni correction already in the initial GWAS. As expected, the *P*-values obtained with MC-simulations are lower since the dependency of the tests is accounted for. Also, SNP rs1998076 in line 9 is genome-wide significant after Bonferroni correction ($P = 0.039$) and MC-simulations improve the level of significance to $P = 0.0248$. Notably, the SNP belongs to the locus on chr 20 that was replicated in an independent sample (Hillmer *et al.*, 2008). Finally, the last SNP of Table 4 was not significant, neither after Bonferroni correction nor with MC-simulations. Consistently, it was not replicated in the independent sample.

Results from strategy VI (analysis of all pairs of SNPs which are among the top 5000 single-marker results and which lie in a common pathway) are found in Table 5. We show the results from the test for genotypic interaction obtained with the log-linear model (test 2). The top pair is defined by SNPs rs608139 (chr 2) and rs4678398 (chr 3). Both SNPs show single-marker association at a moderate level ($P = 0.0080$ and 0.0091) and lie in genes that are annotated to pathway hsa04530 from the KEGG database. According to the database, the pathway is responsible for tight junction.

There is strong evidence for genotypic interaction between the two SNPs, $P = 1.249 \times 10^{-6}$. The result is driven by an excess of double heterozygotes in controls (16.2% versus 3.2% in cases, Table 6).

The interaction *P*-value withstands Bonferroni correction with the number of two-marker tests (18 458 tests, corrected $P = 0.023$). Note that although the two SNPs are selected as the members of the top 5000 single-marker list, Bonferroni correction with the number of actually conducted tests is sufficient since the test statistic of test 2 does not include marginal effects. In order to account for the dependency of the tests caused by LD, we ran 10 000 permutation replicates and obtained a corrected *P*-value of 0.0091. Thus, we were able confirm genome-wide significance for strategy VI. Still, we wish to emphasize that this result requires replication in independent studies, since we carried out all strategies shown in Table 3. Nevertheless, our example demonstrates the potential of the application of priors to GWIA. With a complete GWIA of all SNP pairs (4.5×10^{10} tests), we would have expected 56 250 SNP pairs with a *P*-value smaller than that observed for our top SNP pair from strategy VI. Thus, the pair most probably never would have got into focus.

4 DISCUSSION

We have successfully implemented a software product that allows conduction of a variety of promising and meaningful GWIA strategies. Our running time estimates show that a single computer is sufficient to make those strategies feasible. Thus, researchers are enabled to conduct GWIA even when they do not have access to computing resources that allow massive parallelization. Moreover, by combining a priori information and MC-simulations, our approach could also be more powerful than complete GWIA strategies.

Application to the GWAS on male pattern baldness (Hillmer *et al.*, 2008) revealed an interesting result involving two SNPs lying in genes from a joint pathway. The top hit obtained with this strategy (corrected $P = 0.0086$) pointed to two SNPs that would have been overlooked with a single-marker approach, but also with complete GWIA in which the hit would not have ranked among the top 50 000 hits. In view of the fact that numerous different analysis strategies were conducted, the result warrants replication. Respective efforts are ongoing.

Table 4. Single-marker analysis

No.	Chr_No_1	rs_No_1	Pos_No_1	*P*-value	BONF *P*-value	MC *P*-value	Comment
1	23	rs4548330	496582	6.85333e-10	0.000205899	0.0000	Replicated
2	23	rs5919235	496583	1.03681e-09	0.000311495	0.0001	Replicated
3	23	rs2497938	496629	2.93098e-09	0.000880572	0.0005	Replicated
4	23	rs1041668	496581	3.02748e-09	0.000909565	0.0006	Replicated
5	23	rs5919200	496579	5.92054e-09	0.00177874	0.0012	Replicated
6	23	rs775358	496577	8.35546e-09	0.00251028	0.0016	Replicated
7	23	rs12396249	496648	5.47348e-08	0.0164443	0.0095	Replicated
8	23	rs5919393	496649	6.1065e-08	0.0183461	0.0108	Replicated
9	20	rs1998076	468034	1.30027e-07	0.0390648	0.0248	Replicated
10	13	rs4976846	388824	2.9238e-07	0.0878413	0.0549	Not replicated

Table 5. Results strategy VI

rs_1	*P*_1[a]	rs_2	*P*_2[b]	*P*[c]	BONF *P*-value	MC *P*-value	Pathway
rs608139	0.008	rs4678398	0.009	1.25e-06	0.023	0.0091	hsa04530
rs9436297	0.011	rs17863168	0.012	7.87e-05	1.000	0.4788	hsa04080
rs2892805	0.009	rs348458	0.013	0.00015	1.000	0.7204	hsa00830
rs2892805	0.009	rs610529	0.014	0.00025	1.000	0.8667	hsa00830
rs1199333	0.009	rs5750854	0.009	0.00038	1.000	0.9533	hsa04010
rs9816982	0.001	rs2186598	0.015	0.00038	1.000	0.9535	hsa04080
rs1464443	0.003	rs10487888	0.008	0.00050	1.000	0.9810	hsa04012
rs2575357	0.007	rs3741049	0.012	0.00052	1.000	0.9828	hsa00620
rs918938	0.011	rs7789059	0.014	0.00057	1.000	0.9886	hsa04514
rs11851957	0.002	rs1938958	0.009	0.00062	1.000	0.9917	hsa04080

[a] *P*-value Armitage trend test SNP 1.
[b] *P*-value Armitage trend test SNP 2.
[c] *P*-value test for interaction (4 d.f.).

Table 6. Two-marker genotype frequencies (rs608139 and rs4678398)

	Cases			Controls		
	AA	AB	BB	AA	AB	BB
AA	0.003	0.016	0.003	0.017	0.003	0.003
AB	0.152	0.032	0.045	0.116	0.162	0.023
BB	0.423	0.281	0.045	0.367	0.263	0.046

Recently, Evans *et al*. (2006) have shown that using two-stage two-locus models in GWIA can be less powerful than an exhaustive search. They come to the conclusion that in an approach that requires both SNPs to reach a particular single-marker *P*-value cutoff, power is reduced because it is not unlikely that at least one of the SNPs does not reach the cutoff when marginal effects are small. Since an exhaustive search is typically not feasible, a sensible strategy can, therefore, be to require only one SNP to reach the single-marker cutoff. With the INTERSNP program options it is possible to organize the analysis in this way.

In general, we have to answer the crucial question: 'How important is interaction for the etiology of complex diseases?' We think that this is an open question that cannot be answered at the moment. Of note, according to a literature review containing publications on dichotomous phenotypes until 6/2008, interaction is absent virtually among all the genes found with the standard GWAS single-marker approach (Becker and Herold, 2009). However, there is obviously a bias toward simple disease models in those genes identified via a single-marker strategy. Thus, the phenomenon of missing interaction between genes identified so far need not be representative of the genetic basis of complex diseases in general. Indeed, there are strong arguments that epistasis and genetic interaction are inevitable consequences of the evolutionary process (Frankel and Schork, 1996; Phillips, 2008; Wolf *et al.*, 2000). In view of these reasons and in view of the small fraction of genetic heritability explained so far, it is a reasonable assumption that interaction plays a role in the etiology of at least some complex diseases. The search for those is obligatory. In conclusion, we believe that by application to further GWAS, our software can help to elucidate the actual relevance of interaction for complex diseases.

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

Conflict of Interest: none declared.

REFERENCES

Altshuler,D. *et al*. (2008) Genetic mapping in human disease. *Science*; **322**, 881–888.
Armitage,P. (1955) Tests for linear trends in proportions and frequencies. *Biometrics*, **11**, 375–386.

Becker,T. and Herold,C. (2009) Joint analysis of tightly linked SNPs in screening step of genome-wide association studies leads to increased power. *Eur. J. Hum. Genet.*, **17**, 1043–1049.

Bishop,Y. *et al*. (2007) *Discrete Multivariate Analysis - Theory and Application*. Springer.

Clayton,D. (2008) Testing for association on the X chromosome. *Biostatistics*, **9**, 593–600.

Cordell,H. and Clayton,D. (2002) A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am. J. Hum. Genet.*, **70**, 124–141.

Devlin,B. *et al*. (2004) Genomic control to the extreme. *Nat. Genet.*, **36**, 1129–1130.

O'Dushlaine,C. *et al*. (2009) The SNP ratio test: pathway analysis of genome-wide association datasets. *Bioinformatics*, **25**, 2762–2763.

Evans,D.M. *et al*. (2006) Two-stage two-locus models in genome-wide association. *PLoS Genet.*, **2**, 1424–1432.

Frankel,W. and Schork,N. (1996) Who's afraid of epistasis? *Nat. Genet.*, **14**, 371–373.

Hillmer,A. *et al*. (2008) Susceptibility variants for male-pattern baldness on chromosome 20p11. *Nat. Genet.*, **40**, 1279–1281.

Kanehisa,M. *et al*. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, 354–357.

Maher,B. (2008) Personal genomes: the case of the missing heritability. *Nature*, **456**, 18–21.

Marchini,J. *et al.* (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.,* **37**, 413–417.

Patterson,N. *et al*. (2006) Population structure and eigenanalysis. *PLoS Genet.,* **2**, 2074–2093.

Phillips,P. (2008) Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet*., **9**, 855–867.

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.

Risch,N. and Merikangas,K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.

Tregouet,D. *et al.* (2009) Genome-wide haplotype association study identifies the slc22a3-lpal2-lpa gene cluster as a risk locus for coronary artery disease. *Nat. Genet.*, **41**, 283–285.

Wolf,J. *et al.* (2000) *Epistasis and the Evolutionary Process*. Oxford University Press, Oxford and New York.