BIOINFORMATICS APPLICATIONS NOTE
Vol. 27 no. 1 2011, pages 134–136
doi:10.1093/bioinformatics/btq616

Genetics and population analysis
Advance Access publication November 13, 2010

Automated construction and testing of multi-locus gene–gene associations

Ryan Abo[1,*], Stacey Knight[2], Alun Thomas[2] and Nicola J. Camp[2]

[1]Department of Biomedical Informatics and [2]Department of Internal Medicine, University of Utah School of Medicine, UT, USA

Associate Editor: Martin Bishop

ABSTRACT

Summary: It has been argued that the missing heritability in common diseases may be in part due to rare variants and gene–gene effects. Haplotype analyses provide more power for rare variants and joint analyses across genes can address multi-gene effects. Currently, methods are lacking to perform joint multi-locus association analyses across more than one gene/region. Here, we present a haplotype-mining gene–gene analysis method, which considers multi-locus data for two genes/regions simultaneously. This approach extends our single region haplotype-mining algorithm, hapConstructor, to two genes/regions. It allows construction of multi-locus SNP sets at both genes and tests joint gene–gene effects and interactions between single variants or haplotype combinations. A Monte Carlo framework is used to provide statistical significance assessment of the joint and interaction statistics, thus the method can also be used with related individuals. This tool provides a flexible data-mining approach to identifying gene–gene effects that otherwise is currently unavailable.

Availability: http://bioinformatics.med.utah.edu/Genie/hapConstructor.html

Contact: ryan.abo@hsc.utah.edu

Received on March 8, 2010; revised on October 13, 2010; accepted on October 31, 2010

1 INTRODUCTION

Haplotype and gene–gene analyses have been suggested as strategies to identify disease loci that single nucleotide polymorphism (SNP) approaches may have missed (Manolio *et al*., 2009). Haplotypes have the potential for improved characterization of variation across the locus set (Clark, 2004; Schaid, 2004). Yet, it is usually unclear which haplotypes to test and how to model them. Numerous methods consider all haplotypes spanning the entire locus set, with attempts to reduce the degrees of freedom that this approach otherwise confers (Liu *et al.*, 2007; Tzeng and Zhang, 2007). Other techniques have been designed to analyze contiguous and non-contiguous locus subsets (Abo *et al*., 2008; Browning, 2006; Browning and Browning, 2007; Laramie *et al.*, 2007; Lin, 2004).

It has been hypothesized (Moore, 2003), and in some cases shown (Combarros *et al.*, 2009), that genetic factors at one gene can modify the effects of another gene on disease susceptibility. If such biological interaction exists, the association may only be evident by considering both genes simultaneously. Gene–gene studies are complicated by issues surrounding what constitutes a gene–gene interaction. For example, some approaches for testing interactions focus on association between two unlinked loci (Wu *et al*., 2008; Zhao *et al*., 2006), which do not provide any measure of departure from additivity as a statistical interaction is classically defined.

Most often haplotype analyses are performed for a single region and gene–gene studies concentrate on single SNPs in each region. Methods that consider multi-locus data at more than one gene would be desirable to maximize the ability to detect association evidence. One such method exists to test specific haplotype interactions at unlinked regions (Becker *et al*., 2005). However, both haplotype and gene–gene analyses can result in high-dimensionality, and how to combine them is therefore a challenging problem.

To address these challenges, we have extended our single region haplotype-mining approach (Abo *et al*., 2008) to consider multi-locus data at two genes and test for association and interaction. We concentrate on a broad set of tests that considers both joint effects and interaction effects. In our gene–gene-mining process, data considered at each gene can be single or multi-locus. We anticipate that this gene–gene-mining approach will be most useful for hypothesis generation. However, if required, haplotype testing can also be performed using an empirical correction for multiple testing. Case–control and case-only designs are available, in addition to statistics to test joint and interaction effects. The method is implemented in a Monte Carlo (MC) testing framework and empirical construction-wide significance assessment is available for hypothesis testing.

2 METHODS

For both genes/regions considered, maximum likelihood estimates (MLE) for all individuals' haplotype pairs and population haplotype frequencies are determined. All SNPs in each region and all individuals with sufficient data at both regions are considered (based on a user-defined genotype call rate threshold). Full-length MLE haplotypes, or sub-haplotypes extracted from them, are the genetic variables considered in the construction and testing process.

Consider h and k loci in unlinked genes, $G_1 = \{M_1, \ldots, M_h\}$ and $G_2 = \{M_{h+1}, \ldots, M_{h+k}\}$. The full locus set $S = G_1 \cup G_2$. First, all single locus association tests are conducted. These single locus associations are assessed against the first significance threshold, T_1, which is user-defined. For any locus i with P-value $\leq T_1$, all locus pairs $\{M_i,\ M_j | \forall M_j \in \mathsf{S};\ j \neq i\}$ are considered at the second step. The locus pair $\{M_i,\ M_j\}$ is the locus set, L, being considered. When the two loci in L span both genes, gene–gene tests between the loci are performed. When loci in L are all within the same gene, the two loci are tested as a haplotype or composite genotype. Tests at step n are assessed at significance threshold T_n ($\in \{T_1, \ldots, T_{h+k}\}$), which are usually chosen to be increasing in stringency with n. A locus set can be

*To whom correspondence should be addressed.

written as $L=\{g_1\ g_2 | g_1 \subset G_1$ and $g_2 \subset G_2\}$ where g_1 denotes loci that reside in G_1 and g_2 those that reside in G_2. In steps $n>2$, if there are multiple SNPs in both genes, gene–gene tests between haplotypes across g_1 and haplotypes across g_2 will be performed. The steps continue until no further locus sets pass the defined threshold values or the full locus sets have been tested.

To avoid a strict uphill climb algorithm, which is susceptible to identifying local minimums, we have incorporated a backward step. At each backward step, the algorithm considers subsets of size $n-1$ from the current locus set that were not previously tested. Any subsets which pass the significance threshold, T_n, will be retained and the process will continue forward again.

For locus sets where g_1 and/or g_2 are multi-locus, haplotypes or composite genotypes are considered. The algorithm considers each haplotype across g_i as a potential 'risk haplotype', and compares with all other haplotypes grouped together. For any specific haplotype, this reduces the multi-locus data to a biallelic system which can be used for standard allelic, dominant, recessive and additive models for testing both within and across genes. For composite genotype combinations, phase is unimportant, each locus in L is modeled separately as dominant or recessive and the combinations of these considered across loci. Hence, composite genotypes tests can be performed within or across genes.

To reduce the tests performed, at step $n+1$ the algorithm only expands the specific risk haplotypes that passed the significance threshold (i.e. the alleles at loci from step n are fixed). A similar rule is applied to the composite genotypes.

Single locus, haplotype and composite genotype models are tested using odds ratios, chi-square and chi-square trend association statistics. For locus sets containing loci in two genes, $L=\{g_1\, g_2 | g_1 \subset G_1$ and $g_2 \subset G_2\}$, an interaction odds ratio test and a correlation-based statistic are offered to identify gene–gene effects between the two loci sets, g_1 and g_2. As described above, multi-locus sets within genes are considered using biallelic recoding. We refer to specific haplotypes across g_1 and g_2 as h_1 and h_2.

The interaction odds ratio between h_1 and h_2 is calculated using the method described by Thomas (2004), $\mathrm{IOR}_{m,n}$, where m and n denote dominant or recessive models imposed on h_1 and h_2, respectively, and 0 indicates the wildtype.

$$\mathrm{IOR}_{m,n} = \frac{(\mathrm{OR}_{m,n}\mathrm{OR}_{00})}{(\mathrm{OR}_{m,0}\mathrm{OR}_{0,n})}$$

Under the null hypothesis, H_0: $\mathrm{IOR}_{m,n}=1$, the odds of disease given h_1 and h_2 is the product of the odds of disease for each h_i.

We have also implemented interaction tests based on correlation (Wu *et al*., 2008; Zhao *et al*., 2006). Correlation of specific haplotypes, h_1 and h_2, from locus sets g_1 and g_2 are performed. Following Wang *et al*. (2007), the correlation is determined as follows, where each individual$i$ is assigned a value x_{ij} for locus set g_j based on its MLE haplotype pairs:

$$x_{i,j} = \left\{ \begin{array}{c} -1 \text{ for 0 copies of h}_j \\ 0 \quad \text{for 1 copy of h}_j \\ 1 \quad \text{for 2 copies of h}_j \end{array} \right\}$$

The correlation between h_1 and h_2 is estimated by the correlation coefficient:

$$r = \frac{S_{x_{\cdot 1} x_{\cdot 2}}}{\sqrt{S_{x_{\cdot 1}} S_{x_{\cdot 2}}}},$$

where $S_{x_1,x_2} = \sum_{i=1}^{N} \left(x_{i1} - \bar{x}_{\cdot 1}\right)\left(x_{i2} - \bar{x}_{\cdot 2}\right)$ and $S_{x_{\cdot j}} = \sum_{i=1}^{N} \left(x_{ij} - \bar{x}_{\cdot j}\right)$, $j=(1,\ 2)$, and N is the number of individuals.

This correlation coefficient is an estimate of the composite correlation statistic (Zaykin *et al*., 2008) which is robust to Hardy–Weinberg disequilibrium. For a case–control study design, the method tests $H_0: r_{\mathrm{case}} - r_{\mathrm{control}} = 0$. For a case-only $H_0: r_{\mathrm{case}}=0$ and the first step in the automated process considers the correlation between pairs of single SNPs. We also note the availability of meta-statistics for analyzing multiple datasets.

Statistical significances are determined with a MC procedure. The validity of the MC procedure is based on properly matching the null simulations with the observed data with regard to pedigree structure, missing data structure and phasing procedure (Curtis and Sham, 2006). Our MC procedure is based on a two-region multi-locus gene-drop. In both regions, haplotype pairs are assigned to founders and independent individuals based on the estimated full-length haplotype frequencies. Full-length haplotypes for both regions are then assigned to pedigree descendants using gene-dropping techniques based on Mendelian inheritance (MacCluer *et al*., 1986). The missing data structure is then imposed on the simulated multi-locus genotype data and the known phase is ignored. These simulated data are then statistically phased, to match the procedure performed with the observed data. The procedure generates null genotype configurations from which null statistics are calculated and a null empirical distribution created. It must be noted that this MC procedure assumes a null of no linkage and no association. If strong linkage exists (but no association), there is the potential for inflated type 1 errors; although in simulations we find that for reasonable linkage models that the MC procedure remains a good approximation for the null and type 1 errors remain valid.

Correction for the data-mining process is also available and, if selected, will provide construction-wide significance and false discovery rates. Correction for construction is implemented in the same way as for hapConstructor (Abo *et al*., 2008), where the null distribution for a complete construction run is generated by conducting the same search process starting from 1000 null configurations.

3 IMPLEMENTATION

Our method is implemented as a Java-based program. It is an extension of the hapConstructor module (Abo *et al*., 2008) in the Genie software (Allen-Brady *et al*., 2006). The program can be run on Windows, Unix or Linux machines with Java 1.6 and at least 2 GB of RAM. An example dataset consisting of 14 SNPs in one gene and 11 SNPs in the second gene required 7 h and 11 min with 4 GB of memory to complete building to step 3. Parameter options for this example included default critical thresholds, 10 000 null simulations and no construction-wide assessment. It is important to note that this example may not provide useful insight to other implementations of the method because there are many factors that will affect the running time of the program. These include: number of SNPs, number of samples, number of null simulations selected for significance assessment, critical thresholds selected for the steps in the building process, use of the multiple-testing correction procedure and whether or not there is an association signal. Program details, including the example described above, are available at http://bioinformatics.med.utah.edu/Genie/hapConstructor.html.

Funding: R.A. is an NLM fellow (grant T15 LM0724); National Institutes of Health (CA 098364); the Susan G. Komen Foundation and the Avon Foundation Breast Cancer Fund (to N.J.C.).

Conflict of Interest: none declared.

REFERENCES

Abo,R. *et al.* (2008) hapConstructor: automatic construction and testing of haplotypes in a Monte Carlo framework. *Bioinformatics*, **24**, 2105–2107.

Allen-Brady,K. *et al.* (2006) PedGenie: an analysis approach for genetic association testing in extended pedigrees and genealogies of arbitrary size. *BMC Bioinformatics*, **7**, 209.

Becker,T. *et al.* (2005) Haplotype interaction analysis of unlinked regions *Genet. Epidemiol.*, **29**, 313–322.

Browning,B.L. and Browning,S.R. (2007) Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Power*, **375**, 365–375.

Browning,S.R. (2006) Multilocus association mapping using variable-length Markov chains. *Am. J. Hum. Genet.*, **78**, 903–913.

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

Combarros,O. *et al*. (2009) Epistasis in sporadic Alzheimer's disease. *Neurobiol. Aging*, **30**, 1333–1349.

Curtis,D. and Sham,P.C. (2006) Estimated haplotype counts from case-control samples cannot be treated as observed counts *Am. J. Hum. Genet.*, **78**, 729–730; author reply 728–729.

Laramie,J.M. *et al.* (2007) HaploBuild: an algorithm to construct non-contiguous associated haplotypes in family based genetic studies. *Bioinformatics*, **23**, 2190–2192.

Lin,S. (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat. Genet.*, **36**, 1181–1188.

Liu,J. *et al*. (2007) Incorporating single-locus tests into haplotype cladistic analysis in case-control studies. *PLoS Genet.*, **3**, e46.

MacCluer,J.W. *et al.* (1986) Pedigree analysis by computer simulation. *Zoo Biol.*, **5**, 147–160.

Manolio,T.A. *et al*. (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.

Moore,J.H. (2003) The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum. Hered.*, **56**, 73–82.

Schaid,D.J. (2004) Evaluating associations of haplotypes with traits. *Genet. Epidemiol.*, **27**, 348–364.

Thomas,D.C. (2004) *Statistical Methods in Genetic Epidemiology*, Oxford University Press, New York, USA.

Tzeng,J. and Zhang,D. (2007) Haplotype-based association analysis via variance-components score test. *Am. J. Hum. Genet.*, **81**, 927–938.

Wang,T. *et al*. (2007) Improving power in contrasting linkage-disequilibrium patterns between cases and controls. *Am. J. Hum. Genet.*, **80**, 911–920.

Wu,X. *et al*. (2008) Composite measure of linkage disequilibrium for testing interaction between unlinked loci. *Eur. J. Hum. Genet.*, **16**, 644–651.

Zaykin,D.V. *et al*. (2008) Correlation-based inference for linkage disequilibrium with multiple alleles. *Genetics*, **180**, 533–545.

Zhao,J. *et al*. (2006) Test for interaction between two unlinked loci. *Am. J. Hum. Genet.*, **79**, 831–845.