BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 18 2008, pages 2105–2107
doi:10.1093/bioinformatics/btn359

Genetics and population analysis

hapConstructor: automatic construction and testing of haplotypes in a Monte Carlo framework

Ryan Abo[1,*], Stacey Knight[1], Jathine Wong[1], Angela Cox[2] and Nicola J. Camp[1]

[1]Department of Biomedical Informatics, University of Utah, UT, USA and [2]Institute for Cancer Studies, University of Sheffield Medical School, Sheffield, UK

Received on February 20, 2008; revised on May 22, 2008; accepted on July 14, 2008

Advance Access publication July 23, 2008

Associate editor: Alex Bateman

ABSTRACT

Summary: Haplotypes carry important information that can direct investigators towards underlying susceptibility variants, and hence multiple tagging single nucleotide polymorphisms (tSNPs) are usually studied in candidate gene association studies. However, it is often unknown which SNPs should be included in haplotype analyses, or which tests should be performed for maximum power. We have developed a program, hapConstructor, which automatically builds multi-locus SNP sets to test for association in a case-control framework. The multi-SNP sets considered need not be contiguous; they are built based on significance. An important feature is that the missing data imputation is carried out based on the full data, for maximal information and consistency. HapConstructor is implemented in a Monte Carlo framework and naturally extends to allow for significance testing and false discovery rates that account for the construction process and to related individuals. HapConstructor is a useful tool for exploring multi-locus associations in candidate genes and regions.

Availability: http://www-genepi.med.utah.edu/Genie

Contact: ryan.abo@hsc.utah.edu

1 INTRODUCTION

Multiple tagging-SNPs (tSNPs) are widely used in candidate gene association studies. It has been shown that there is increased power to detect disease variants with low frequency by performing both haplotype and single-locus analyses even with the multiple testing correction (Becker and Knapp, 2004). In new studies, tSNPs are usually analyzed independently and in multi-SNP combinations. Even when associations are considered established (Cox *et al.*, 2007), comprehensive SNP-set haplotype analyses can be performed to more accurately define the haplotype/s on which the susceptibility variants lie. One avenue that may effectively guide such searches is a more systematic haplotype-mining analysis.

Multi-locus analyses are of high dimension leading to reduced power when testing association. Haplotype similarity, cladistic and phylogenetic techniques can be used to reduce dimensionality (Bardel *et al.*, 2006; Camp *et al.*, 2005; Jannot *et al.*, 2004; Liu *et al.*, 2007; Molitor *et al.*, 2003; Tzeng and Zhang, 2007; Waldron *et al.*, 2006; Yu *et al.*, 2005). However, these methods require a priori determination of which SNPs to include; and there remains the question of whether to analyze monotype or diplotype data and the mode of expression.

Studying SNP subsets may be optimal and reduce dimension. Sliding windows (Lin *et al.*, 2004) and haplotype clustering using variable-length Markov chain models (Browning and Browning, 2007; Browning, 2006) have been proposed for traditional case-control data and contiguous subsets of SNPs. An approach for non-contiguous SNP subsets exists; constructing haplotypes by starting from SNP pairs and iteratively adding SNPs based on significance and base pair distance (haploBuild; Laramie *et al.*, 2007). While this latter approach is flexible in haplotype construction, it is limited to transmission statistics in the FBAT software (Horvath *et al.*, 2001) and lacks a valid significance assessment that accounts for all the multiple testing inherent in a data-mining technique.

We present hapConstructor, software to construct and test multi-locus data, allowing for non-contiguous SNP subsets. Tests for non-independence and effect size are incorporated. Monotype (alleles or haplotypes); diplotype (genotypes or haplotype pairs); and composite genotype (unphased genotypes across multiple loci) tests are included. Standard reductions of dimensionality are incorporated, such as specific haplotype tests for monotype data, and dominant, recessive and additive tests for specific haplotypes for diplotype data. Multi-locus SNP sets are constructed through a forward–backward stepwise process. HapConstructor operates in a Monte Carlo (MC) framework which offers two advantages. First, it naturally extends to testing related individuals. Second, the null distribution for the full SNP set is simulated once, and can be used to assess both empirical significance of individual tests and *construction-wide P*-values and false discovery rates (FDRs) that account for the construction process. HapConstructor is a Java-based extension of Genie (Allen-Brady *et al.*, 2006).

2 METHODS

The MC framework is provided by Genie, with imputation of missing data, estimation of population haplotype frequencies and maximum likelihood estimates (MLE) of individuals' haplotype pairs provided by the hapMC component.

First, all single SNPs $\{s1, s2, \ldots, sn\}$ are tested. In each forward step, a SNP is added to SNP sets whose *P*-value surpassed the user-defined threshold at the previous step. The thresholds can be constant or may vary by step. For example, if s1 surpassed the first threshold, the next step would consider two-locus SNP sets $\{s1\text{-}s2, s1\text{-}s3, \ldots, s1\text{-}sn\}$. An optional backward process starts at the third step and consists of testing all $(n-1)$-locus subsets

*To whom correspondence should be addressed.

not previously considered. To maintain efficiency and speed and reduce redundancy, each subsequent step in the build process extends the haplotypes with the specific alleles that previously met the threshold at the prior step rather than considering all haplotypes spanning the new loci set.

Test statistics available are χ^2, χ^2-trend and odds ratio. The data can be considered as diplotype or monotype or both. For diplotype data, haplotype and composite genotype tests are performed. Haplotype models are dominant, recessive and additive models for each haplotype. Composite genotypes include each of the dominant and recessive combinations across loci. For monotype data, each specific haplotype is compared to all others.

Summaries for all tests performed are stored. A user interface allows these to be sorted by step, SNP, test-type and significance. If required, a *construction-wide* assessment that accounts for the building process can be made. A valid global *P*-value and FDR is generated; the latter is more appropriate for data mining (Benjamini and Hochberg, 1995). These are achieved by reusing the null configurations generated for the MC procedure. Each null configuration is considered as the 'observed data' and the construction algorithm is used with significances determined from the remaining $N-1$ null configurations. This is repeated to generate a set of null 'constructions' from which valid empirical *construction-wide P*-values and FDRs are determined.

3 RESULTS

We illustrate hapConstructor using a sample of 1128 independent breast cancer cases and 1149 independent controls from Sheffield, UK and 14 tSNPs in the CASP8 gene. Single SNP tests results yielded three SNPs with *P*-values below 0.05 (0.010–0.047). The construction process continued to the fifth step (five-locus haplotypes). A four-locus haplotype was identified as the most significantly associated haplotype with an empirical *P*-value of 8.0×10^{-5} and a construction-wide FDR of 0.044, a result which is consistent with the established association between breast cancer and CASP8 (Cox *et al.*, 2007). This four-locus haplotype contained only one of the three SNPs that had obtained significant single test results.

HapConstructor completed the building process for the real data in 96 h with 100 000 MC simulations, on a machine with an Intel Pentium core 2 duo with 3.0 Ghz per processor and 2 GB of memory. It required 7 days using 10 server nodes to complete 1000 simulated builds for the *construction-wide* significance assessment.

To assess the potential value-added of the construction process in our illustrative example, we analyzed all 14-SNP haplotypes with frequencies over 1% and also performed exhaustive sliding window analyses for window sizes of 2- to 6-SNP haplotypes. Of the 15 14-SNP haplotypes analyzed, only one obtained nominal significance ($P = 0.0357$). For the sliding windows, 2351 tests were conducted and 314 were found to be nominally significant (0.0021–0.05, not accounting for multiple testing). The most significantly associated haplotypes were found in the four-, five- and six-locus window sizes. The results from both of these more standard approaches were inferior to the haplotype building in terms of significance and indicate that hapConstructor was a valuable approach and that exhaustive searches using contiguous multi-SNP sets are not the optimal solution in this situation.

4 CONCLUSIONS

HapConstructor offers a data-mining approach to association analyses, allowing automatic and comprehensive construction of multi-locus SNP-set tests. It improves upon other methods in the variety of analyses and statistics performed, and the ability to appropriately assess global significance. Additional features are the immediate extension to mixtures of independent and related individuals, a virtue of the method being nested in Genie (Allen-Brady *et al.*, 2006), and the ability to impute missing data. It should be noted, however, that the extension to related individuals is limited to an assumption of no recombination, as only under these conditions are MLE haplotype estimates using relatives unbiased.

A limitation of hapConstructor, and MC testing in general, is computational burden. This is dependent upon the number of simulations (especially *construction-wide* assessment), sample size, number of SNPs considered and threshold values. Depending on the dataset being analyzed, hapConstructor may require significant time and computational resources to complete both the build process and construction-wide assessment. Construction-wide assessment may be intractable for large datasets due to time or resources. Despite the computational intensity, hapConstructor is a useful tool for exploring multi-locus associations in candidate genes and regions, and fulfills a current need of many investigators. Our future work will include more sophisticated heuristics for the construction process and extensions to interaction models.

ACKNOWLEDGEMENTS

Funding: This work was supported by NCI (CA098364) and the Komen Foundation (BCTR0706911). R.A. is an NLM fellow (1T15LM007124). Genotypes were funded by the Breast Cancer Campaign and Yorkshire Cancer Research and generated by Ian Brock.

Conflict of Interest: none declared.

REFERENCES

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.

Bardel,C. *et al*. (2006) Clustering of haplotypes based on phylogeny: how good a strategy for association testing? *Eur. J. Hum. Genet.*, **14**, 202–206.

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

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B. Met.*, **57**, 289–300.

Browning,B.L. and Browning,S.R. (2007) Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genet. Epidemiol.*, **31**, 365–375.

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

Camp,N.J. *et al*. (2005) Characterization of linkage disequilibrium structure, mutation history, and tagging SNPs, and their use in association analyses: ELAC2 and familial early-onset prostate cancer. *Genet. Epidemiol.*, **28**, 232–243.

Cox,A. *et al*. (2007) A common coding variant in CASP8 is associated with breast cancer risk. *Nat. Genet.*, **39**, 352–358.

Horvath,S. *et al*. (2001) The family based association test method: strategies for studying general genotype–phenotype associations. *Eur. J. Hum. Genet.*, **9**, 301–306.

Jannot,A.S. *et al*. (2004) Association in multifactorial traits: how to deal with rare observations? *Hum. Hered.*, **58**, 73–81.

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. *et al*. (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.

Molitor,J. *et al*. (2003) Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. *Am. J. Hum. Genet.*, **73**, 1368–1384.

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

Waldron,E.R. *et al*. (2006) Fine mapping of disease genes via haplotype clustering. *Genet. Epidemiol.*, **30**, 170–179.

Yu,K. *et al*. (2005) Using tree-based recursive partitioning methods to group haplotypes for increased power in association studies. *Ann. Hum. Genet.*, **69**, 577–589.