W444–W449 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* *Published online 16 May 2011*
doi:10.1093/nar/gkr321

SNPsyn: detection and exploration of SNP–SNP interactions

Tomaz Curk[1,*], Gregor Rot[1] and Blaz Zupan[1,2,*]

[1]Faculty of Computer and Information Science, University of Ljubljana, Trzaska cesta 25, SI-1000 Ljubljana, Slovenia and [2]Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

Received March 5, 2011; Revised April 15, 2011; Accepted April 20, 2011

ABSTRACT

SNPsyn (http://snpsyn.biolab.si) is an interactive software tool for the discovery of synergistic pairs of single nucleotide polymorphisms (SNPs) from large genome-wide case-control association studies (GWAS) data on complex diseases. Synergy among SNPs is estimated using an information-theoretic approach called interaction analysis. SNPsyn is both a stand-alone C++/Flash application and a web server. The computationally intensive part is implemented in C++ and can run in parallel on a dedicated cluster or grid. The graphical user interface is written in Adobe Flash Builder 4 and can run in most web browsers or as a stand-alone application. The SNPsyn web server hosts the Flash application, receives GWAS data submissions, invokes the interaction analysis and serves result files. The user can explore details on identified synergistic pairs of SNPs, perform gene set enrichment analysis and interact with the constructed SNP synergy network.

INTRODUCTION

Current genome-wide case-control association studies (GWAS) focus on identifying a set of single nucleotide polymorphisms (SNPs) that are most associated with the disease under study. While individual SNPs are important indicators of main genetic components of complex diseases, they explain only a fraction of the genetic risk (1). Because of the low or at best modest information content of individual SNPs, it has been suggested (2) that uncovering synergy among genes may improve the predictive accuracy of models. A recent report by Gerke *et al.* (3) also suggests that synergistic combinations may carry information about the phenotype that cannot be discovered from observations of individual SNPs alone. An unequivocal proof of existence of SNP synergy would push the modeling efforts from trying to add effects of individual most informative SNPs towards models that include non-additive SNP interactions, in this way providing important insight into complex diseases and underlying molecular mechanisms.

Various approaches to detect synergy have been proposed, which is commonly referred to as positive interaction (4), k-way interaction information (5), epistasis (6,7) or SNP synergy (8). In this article, we use the term 'synergy' and present a software tool that implements an information-theoretic approach to synergistic interaction analysis (4,5,8). Contrary to other approaches, interaction analysis does not require the user to specify which gene interaction models to test, but instead it discovers them from data. It assumes an additive model, where the expected amount of information on the phenotype for a combination of SNPs is equal to the sum of information of individual SNPs. Synergy is said to occur when a combination carries more information than the sum of information provided by individual SNPs (4,8). This difference between the 'whole' and 'sum of parts' cannot be gained from observations of individual SNPs alone, but only by simultaneously observing a combination of SNPs.

Various degrees of synergy are associated with different SNP pair models (9). An extreme case is when the outcome is an XOR function of two SNPs. There, each individual SNP does not carry any information on the phenotype, while a simultaneous consideration of the two SNPs produces a perfect association with disease. This extreme case illustrates that, by definition, it is not possible to predict which SNPs will form a synergistic combination by observing individual SNPs alone. Two SNPs must first be combined into a new feature, and only then can the total information content for that particular combination be computed.

Consequently, to discover a set of best-interacting SNPs we need to test exhaustively all possible combinations. The number of SNP combinations grows exponentially

*To whom correspondence should be addressed. Tel: +386 1 4768 267; Fax: +386 1 4264 647; Email: tomaz.curk@fri.uni-lj.si
Correspondence may also be addressed to Blaz Zupan. Tel: +386 1 4768 402; Fax: +386 1 4264 647; Email: blaz.zupan@fri.uni-lj.si

with the order of interaction (i.e. number of SNPs in combination) and the number of SNPs in data. Given N SNPs, there are $N(N-1)/2$ pairs and $N(N-1)(N-2)/6$ triplets. Exploring higher order combinations of SNPs may be desired, but is computationally intractable with data sets that include more than few ten thousands of SNPs. Current GWAS data include over one million SNPs but typically do not include more than few thousands cases and controls. Low sample-to-feature ratio, which decreases exponentially with number of SNPs, is another limiting factor. It prevents obtaining statistically significant results, increases the opportunity to over-fit and thus limits SNPsyn's exploration to pairs of SNPs. Heuristic, non-exhaustive search require shorter run times, but cannot guarantee the detection of all synergistic pairs.

METHODS AND IMPLEMENTATION

SNPsyn aims to optimize the computational time and at the same time provides an interaction-rich graphical user interface. The computationally intensive data analysis is implemented in C++. This computational library implements functions for calculating mutual information and information gain of individual and pairs of SNPs and synergy of pairs of SNPs. The library also includes functions for random data sampling and shuffling, estimation of probability distribution, calculation of false discovery rate [FDR, (10)] and functions for the subdivision of the analysis into independent subtasks that can run in parallel. Example scripts to perform the analysis in parallel on a cluster or grid are included in the distribution package. SNPsyn's C++ library can be used to build custom applications for interaction analysis. A command-line interface to the library is provided, and is actually used by SNPsyn's web server to perform interaction analysis.

Results of interaction analysis are presented to the user through an interactive web application with a graphical user interface (GUI). The interface has a desktop-like feel and was designed using the Adobe's Flash Builder 4 development framework. The GUI offers a series of effective visualizations for explorative analysis of results generated by the computationally intensive part of the system. The GUI runs as a web application inside web browsers that supports Adobe's Flash player. It sends analysis requests to SNPsyn server and renders the results of analysis. We also provide a stand-alone version of the application that runs in Adobes AIR runtime environment and is completely independent from the web server.

Interaction analysis

Synergy (Syn) of a pair of SNPs (M_1 and M_2) is the difference between the information on the phenotype P encoded in the newly derived feature (defined by cross-product function f) and the sum of information encoded by the two individual features (4,8):

$$\mathrm{Syn}(M_1, M_2|P) = \mathrm{I}(f(M_1, M_2); P) - [I(M_1; P) + I(M_2; P)]$$

Mutual information $I(M; P)$, also called information gain, is based on calculations of entropy and corresponds to the level of association (i.e. shared information) between marker M and phenotype P. Given the value of marker M, mutual information estimates how well can we predict the value of phenotype P. The new feature $f(M_1, M_2)$ may be derived by Cartesian product of values of SNPs M_1 and M_2 or by other methods for feature construction, e.g. Kramers method (11) or constructive induction by feature decomposition (12). For reasons of simplicity and speed, SNPsyn uses Cartesian product. Pairs of SNPs with positive synergy ($Syn > 0$) are called synergistic. Negative synergy ($Syn < 0$) indicates that the two SNPs carry redundant information, an effect typically observed among highly correlated SNPs. For further details on interaction analysis see Jakulin and Bratko (4) and a review by Anastassiou (8).

Compact data format

GWAS data are usually encoded in large human-readable text files exceeding one GB in size, which is not suitable when data is read many times by concurrently running processes on a cluster. SNPsyn accepts GWAS data in various text formats, such as PLINK's ped, tped (13) or tab-delimited files, where each row holds genotype information and other annotation on a sample. For reasons of speed, the data is first transformed into SNPsyn's compact binary format (see web site for detailed specification). The format is similar to PLINK's Binary PED file and allows up to 255 different genotype values for each marker (PLINK can encode only four values: three for genotype, one for missing value). This will allow future extensions of SNPsyn to work with other kinds of markers, such as haplotypes and structural variants data.

SNP-to-gene mapping and gene set enrichment analysis

SNPs are mapped to genes using the mapping in NCBI's dbSNP database. Gene Ontology (GO) term enrichment analysis (14) requires two sets of genes. The 'cluster' set is obtained by mapping the user-selected SNPs to genes. The 'reference' set is obtained by mapping all SNPs in the data to genes. SNPs that cannot be mapped to genes are excluded from enrichment analysis. SNPsyn uses the hypergeometric distribution to compute the associated significance (P-values) and visualizes the results similarly as in GOAT (15). Currently, only human SNPs can be analyzed. SNPsyn's documentation incudes examples on how to prepare a local installation of the software for mouse or other species.

Permutation analysis and assessment of significance

The large number of tests performed when searching for synergy demands a strict assessment of the significance of results. Because the goal is to select SNP pairs with both high information gain and high synergy, we define the null distribution of (I, Syn) scores by randomly shuffling data a number of times (e.g. 100 times, see Supplementary Data), each time computing the scores for all pairs of SNPs. Two random data shuffling approaches are implemented in SNPsyn: permute phenotype labels and permute

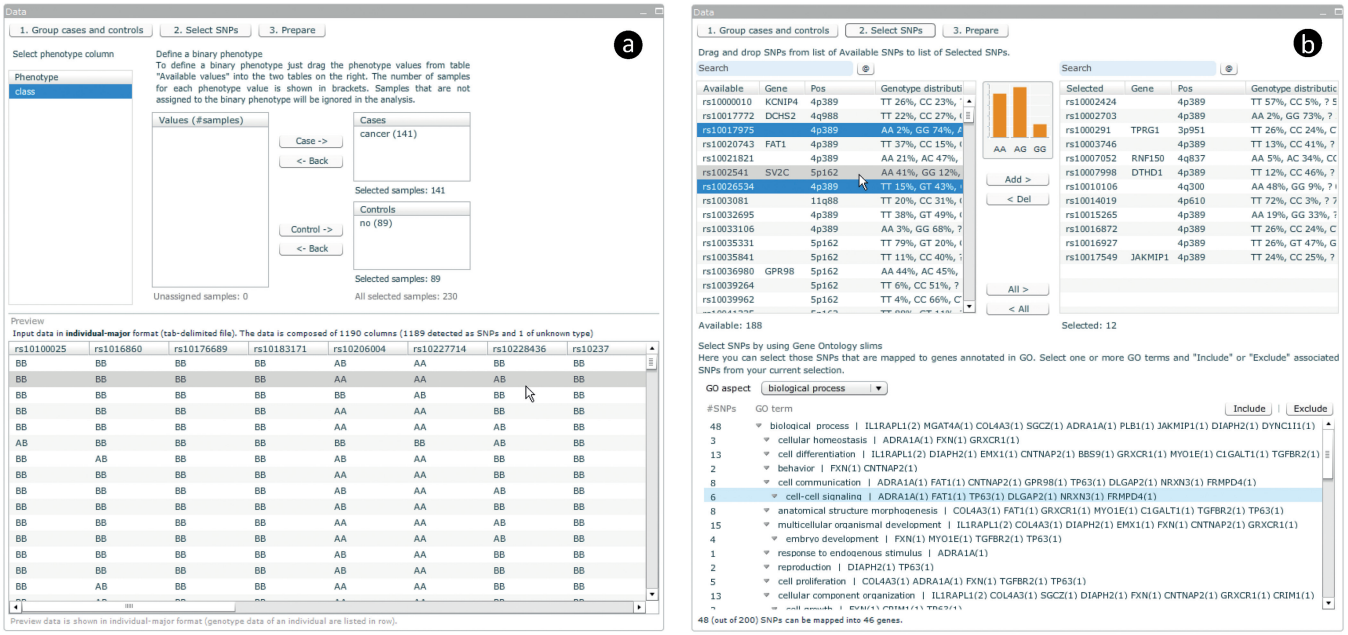

Figure 1. Data preparation. (**a**) Grouping of samples into cases and controls. (**b**) Selection of SNPs for analysis may be aided with Gene Ontology browser.

genotype data across samples (default). The significance of a given SNP pair with score (I, Syn) is determined from the null distribution by calculating N_{ge}/N_{all}, where N_{ge} is the number of equally or better scored pairs obtained on random data ($I_{rnd} \geq I$ and $Syn_{rnd} \geq Syn$) and N_{all} is the number of all tests performed on randomly permuted data. Obtained significance scores are corrected for multiple testing using the FDR method described by Benjamini and Yekutieli (10).

Computational requirements and restrictions on search

Exhaustive search for SNP synergies requires long processing times and may identify possibly large number of candidate synergistic SNP pairs. Best 5000 SNP pairs with highest synergy and 5000 SNP pairs with highest information gain are retained and presented to the user for explorative analysis.

The publicly available SNPsyn web server limits exhaustive search to 22 000 SNPs (242M pairs, which require 5 h of CPU time). These limits are imposed because of the associated high computational costs and the desire to offer this service to a larger number of researches. Although no fast and exact solution is known for this search problem—the XOR SNP interaction model being an example where all heuristics fail—some theoretical studies have demonstrated (16) the applicability of two-stage heuristic approaches. When more than 22 000 SNPs are given on SNPsyn's input, the user must choose between two heuristics: 'synergy among main effects' or 'approximate screening of all pairs'. In the former, SNPsyn examines all pairs among 22 000 SNPs with highest information gain. In the latter, closely following an idea implemented in BOOST (17), an upper approximate bound on Syn is used to quickly screen all pairs. The analysis is then performed on a smaller subset of best candidate pairs. See Supplementary Data for details and comparison of the two heuristics.

If needed, these search restrictions can be lifted in a local installation on the user's computer, dedicated cluster or grid. Instructions how to set up a local SNPsyn server are provided on the web site. Another possibility is to use the stand-alone Adobe AIR version of SNPsyn to prepare the data for analysis and then use the command-line utility to run the analysis locally.

RESULTS

SNPsyn provides a user-friendly interactive graphical interface that supports all steps in the analysis of GWAS data: data preparation, interaction analysis and exploration of results. We briefly describe each of these steps below.

Data preparation and submission

SNPsyn can read GWAS data in PLINK's ped and tped formats and a tab-delimited format. PLINK's formats store assignment of samples into cases and controls. When loading data from tab-delimited files, the user must select an annotation column that is used to assign samples into groups. Groups are usually defined based on the phenotype (e.g. classes or subclasses of a disease with a common genetic component, etc.). Samples from each group can be assigned into either the case or control class. This group-to-class mapping allows easy exploration of synergy in specific subgroups of cases and controls (Figure 1a).

SNPsyn implements two approaches to SNP selection for synergy exploration. The hypothesis-free *de novo*

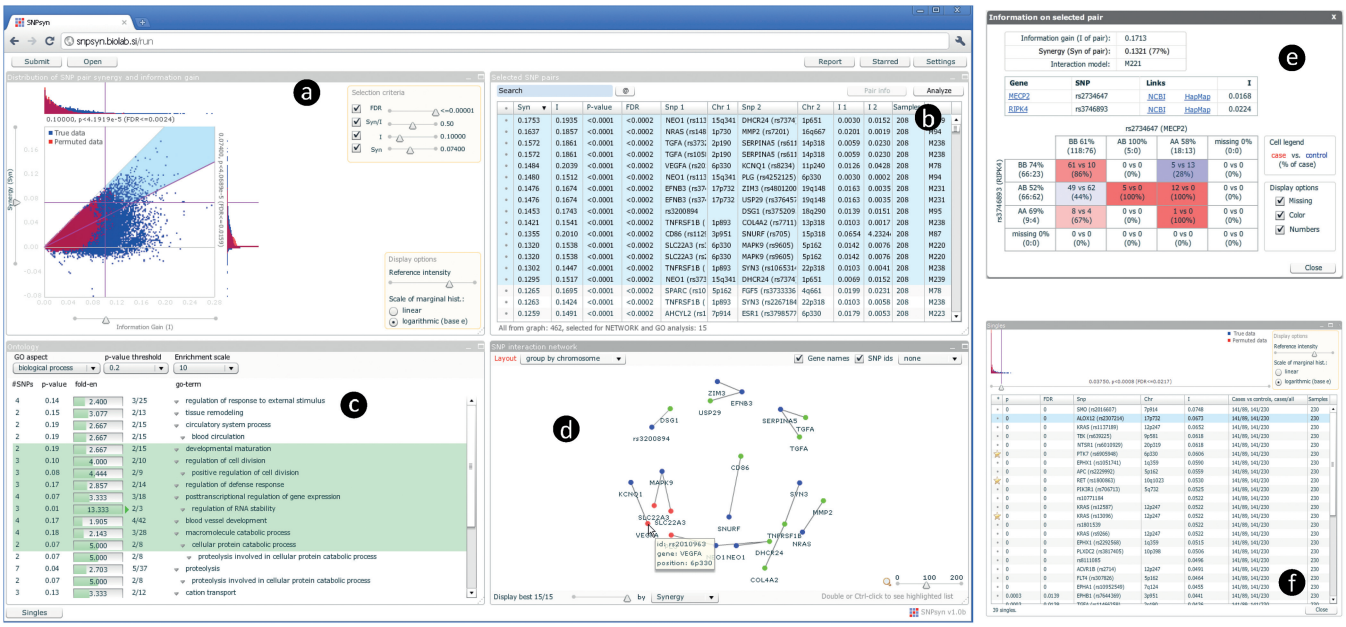

Figure 2. Exploration of results. (**a**) *I* versus *Syn* scatterplots. (**b**) List of SNPs selected in (a). (**c**) Gene Ontology enrichment analysis of SNP pairs selected in (b). (**d**) Synergy network from SNP pairs selected in (b). (**e**) Details on a selected pair. (**f**) Results of individual SNP analysis.

whole genome approach, where all SNPs are used, and the hypothesis-driven investigation, where a user-defined, knowledge-based selection of a subset of SNPs is explored. In the latter approach the user can focus on a more specific biological question, drastically reducing the number of candidate SNP pairs. The user can hand-pick individual SNPs or subsets of SNPs associated with genes in specific, biologically relevant annotation terms in Gene Ontology (Figure 1b).

Data is then encoded in SNPsyn's compact binary format file and is submitted to SNPsyn web server for analysis. The user may also choose to store the data locally and analyze it on local computational facilities.

Visual exploration of results

The results of data analysis include a list of single SNPs with highest information gain $I(M_i; P)$, a list of SNP pairs with highest information gain $I(f(M_i, M_j); P)$ and a list of SNP pairs with highest synergy $Syn(f(M_i, M_j)|P)$.

Calculated scores for SNP pairs on true data are plotted in a *I* versus *Syn* scatterplot (blue dots, Figure 2a), with the superimposed null-distribution (red dots, Figure 2a). Distributions of *Syn* and *I* are plotted in histograms on the sides of the scatterplot. Pairs of SNPs can be selected by user-defined minimum synergy (*Syn*), information gain (*I*), synergy ratio (*Syn*/*I*) and FDR. The scatterplot region defined by these constraints is highlighted in blue and associated SNP pairs are displayed in a table (Figure 2b). There, the user can bookmark (star) favorite SNP pairs for latter access in a separate window. Starred SNPs are included in the detailed report.

Even after filtering by information gain, synergy and significance scores, the list of SNP pairs with highest synergy can be quite extensive and may include a large number of false positives. Instead of just examining details on individual pairs (shown in Figure 2e), one may benefit from exploring and reasoning on commonalities among genes associated with best-ranked pairs. Links to detailed information on SNP and gene annotation at NCBI and HapMap (18) are provided throughout the interface wherever a gene or SNP is shown. Additionally, the user can perform Gene Ontology (19) term enrichment analysis (Figure 2c) and visualize an interaction network (Figure 2d).

Enriched GO terms are drawn in a tree plot browser (Figure 2c). Each row represent an enriched term and provides details on the number of matching genes (and SNPs) in the cluster and reference sets, *P*-value, FDR and fold enrichment. Genes associated with user-selected GO-terms are rendered in green in the interaction network.

SNPs displayed in the interaction network (Figure 2d) are connected if the pair was selected by the user in Figure 2b. Three layouts of the network are available to survey the overall structure and quickly identify commonalities among interacting genes: network nodes (SNPs) are spread out uniformly and connected nodes tend to be displayed closer to each other ('basic' layout), SNPs from same gene are displayed closer to each other ('group by gene'), SNPs from same chromosome are shown closer to each other ('group by chromosome'). Network edges may additionally include labels for *Syn* or *I* scores of a corresponding pair. A sliding bar can be used to reduce or expand the network by selecting the number of best-ranked pairs to draw. When a node is selected on the network, other nodes in the network that are either from the same gene or the same chromosome get highlighted in red. This visual cue allows to quickly identify groups of similar SNPs. Details on the group of highlighted SNPs can be displayed in a separate window.

Besides SNP synergy, SNPsyn can also display the results of analysis of individual SNPs (Figure 2f). These are available by clicking on the Single SNPs button (Figure 2, lower left). Detailed results on individual SNPs are displayed in a separate window. By moving a slide bar below the distribution of information gain scores, the user can select the corresponding most informative SNPs to be displayed in the table. Individual SNPs can be marked with a star for latter access and are included in the report.

A report on the results of the analysis including the current state of the exploration, with details on best-rated individual SNPs and pairs of SNPs, can be generated and downloaded at any time during exploration.

Comparison with other tools

SNPsyn addresses a type of genome-wide analysis of SNP interactions similar to those implemented in PLINK (13), MDR (7), HFCC (20), PLR (21), BEAM (22), etc. Various existing methods and tools are reviewed in Cordell (23). Two main features differentiate SNPsyns from other tools. The first one is its application of information theory to determine synergy. This solves a critical problem in dealing with main effect SNPs that afflicts some of the mentioned tools (MDR, HFCC), which tend to rank highly SNP pairs with low synergy but high information content that is due to a highly informative SNP in the pair. To compensate for this, *ad hoc* filters are applied, e.g. main effect SNPs are removed from the analysis, potentially missing a subset of synergistic pairs. The information-theoretic approach implemented by SNPsyn elegantly solves this by directly calculating the amount of synergy. The second distinctive feature of SNPsyns is its highly interactive, GUI that supports all steps of an explorative analysis of synergy and reveals a structure of the discovered gene interaction network.

CONCLUSION

With increased availability of the experimental technology, decrease of its costs, and emerging techniques such as high-throughput sequencing (24), the number of whole GWAS is on the rise. Efficient and easy-to-use bioinformatics and data analytics tools are needed to support biologists in their search for relations between genotype and phenotype. The development of such software is far from trivial. It needs to address critical issues such as, on one hand, computational speed and appropriate statistical treatment when dealing with low sample-to-feature ratios and, on the other hand, presentation of results that can support interactive data exploration. The latter is crucial as it provides means to biologists to reconnect with their own data in the absence of constant required interventions by computer and software specialists. SNPsyn addresses all these issues with a carefully designed implementation of selected computational and statistical approaches and with its intuitive and easy-to-use interactive graphical interface for explorative analysis of synergistic gene interactions.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the referees for very insightful suggestions.

FUNDING

Slovenian Research Agency (P2-0209, J2-2197, L2-1112, Z7-3665). Funding for open access charge: Slovenian Research Agency.

Conflict of interest statement. None declared.

REFERENCES

1. Wellcome Trust Case Control Consortium. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
2. Daly,M.J. and Altshuler,D. (2005) Partners in crime. *Nat. Genet.*, **37**, 337–338.
3. Gerke,J., Lorenz,K. and Cohen,B. (2009) Genetic interactions between transcription factors cause natural variation in yeast. *Science*, **323**, 498–501.
4. Jakulin,A. and Bratko,I. (2003) Analyzing attribute dependencies. In: *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2003), Lecture Notes in Artificial Intelligence*. Springer, Berlin/Heidelberg, New York, pp. 229–240.
5. Chanda,P., Zhang,A., Brazeau,D., Sucheston,L., Freudenheim,J.L., Ambrosone,C. and Ramanathan,M. (2007) Information-theoretic metrics for visualizing gene-environment interactions. *Am. J. Hum. Genet.*, **81**, 939–963.
6. Cordell,H.J. (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.*, **11**, 2463–2468.
7. Hahn,L.W., Ritchie,M.D. and Moore,J.H. (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, **19**, 376–382.
8. Anastassiou,D. (2007) Computational analysis of the synergy among multiple interacting genes. *Mol. Syst. Biol.*, **3**, 83.
9. Li,W. and Reich,J. (2000) A complete enumeration and classification of two-locus disease models. *Hum. Hered.*, **50**, 334–349.
10. Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, **29**, 1165–1188.
11. Kramer,S. (1994) Cn2-mci: a two-step method for constructive induction. In *Proceedings of the ML-COLT-94 Workshop on Constructive Induction and Change of Representation*. New Brunswick, New Jersey.
12. Zupan,B., Bohanec,M., Demsar,J. and Bratko,I. (1999) Learning by discovering concept hierarchies. *Artif. Int.*, **109**, 211–242.
13. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A.R., Bender,D., Maller,J., Sklar,P., de Bakker,P.I.W., Daly,M.J. *et al.* (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
14. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.
15. Xu,Q. and Shaulsky,G. (2005) Goat: An r tool for analysing gene ontology term enrichment. *Appl. Bioinform.*, **4**, 281–283.
16. Evans,D.M., Marchini,J., Morris,A.P. and Cardon,L.R. (2006) Two-stage two-locus models in genome-wide association. *PLoS Genet.*, **2**, e157.

17. Wan,X., Yang,C., Yang,Q., Xue,H., Fan,X., Tang,N.L.S. and Yu,W. (2010) Boost: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.*, **87**, 325–340.
18. Frazer,K.A., Ballinger,D.G., Cox,D.R., Hinds,D.A., Stuve,L.L., Gibbs,R.A., Belmont,J.W., Boudreau,A., Hardenbolt,P., Leal,S.M. *et al.* (2007) A second generation human haplotype map of over 3.1 million snps. *Nature*, **449**, 851–861.
19. Wang,K., Li,M. and Bucan,M. (2007) Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.*, **81**, 1278–1283.
20. Gayán,J., González-Pérez,A., Bermudo,F., Sáez,M.E., Royo,J.L., Quintas,A., Galan,J.J., Morón,F.J., Ramirez-Lorca,R., Real,L.M. *et al.* (2008) A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis. *BMC Genomics*, **9**, 360.
21. Park,M.Y. and Hastie,T. (2008) Penalized logistic regression for detecting gene interactions. *Biostatistics*, **9**, 30–50.
22. Zhang,Y. and Liu,J.S. (2007) Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet.*, **39**, 1167–1173.
23. Cordell,H.J. (2009) Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **10**, 392–404.
24. The 1000 Genomes Consortium. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.