Published online DD MMMM YYYY *Nucleic Acids Research, 2011, Vol. NN, No. NN* **1–5**
doi:10.1093/nar/gkn000

SNPsyn: detection and exploration of SNP-SNP interactions

Tomaz Curk [1,*], Gregor Rot [1], Antonio Julià [2], Sara Marsal [2], Blaz Zupan [1,3 *]

[1]Faculty of Computer and Information Science, University of Ljubljana, Trzaska cesta 25, SI-1000 Ljubljana, Slovenia, [2]Grup de Recerca de Reumatologia, Institut de Recerca Vall d'Hebron, Passeig Vall d'Hebron, 119-129, 08025 Barcelona, Spain and [3]Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

Received MMMM D, YYYY; Revised MMMM D, YYYY; Accepted MMMM D, YYYY

ABSTRACT

SNPsyn (http://snpsyn.biolab.si) is an interactive software tool for the discovery of synergistic pairs of SNPs in 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 Flex 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 Ontology 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 and thus provide an important insight into complex diseases.

Various approaches to detect synergy have been proposed, which is commonly referred to as positive interaction (4), epistasis (5) or SNP synergy (6). In this paper we use the term *synergy* and present a software tool that implements an information-theoretic based approach to synergistic interaction analysis (4, 6). 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, 6). 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 can be defined (7). XOR is an extreme example of synergistic relation where 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. By definition, it is not possible to predict which SNPs will form a synergistic combination just by observing individual SNPs. Two SNPs must be combined into a new feature, only then can the total information content of 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 with the order of interaction (*i.e.*, number of SNPs forming a combination) and the number of SNPs in data. Given N SNPs, $N(N-1)/2$ pairs and $N(N-1)(N-2)/6$ triplets can be formed. Exploring higher combinations of SNPs may be desired, but is computationally intractable at few ten thousands SNPs. Current GWAS data include over one million SNPs but typically at best include a 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 exploration to pairs of SNPs. Heuristic, non-exhaustive search required shorter run times, but cannot guarantee the detection of all synergistic pairs.

*To whom correspondence should be addressed. Tel: +386 1 4768 267; Email: tomaz.curk@fri.uni-lj.si

METHODS AND IMPLEMENTATION

SNPsyn aims to optimize the computational time and at the same time provide an interaction-rich graphical interface. The computationally intensive data analysis is implemented in C++. The 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, (8)) 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. The C++ library can be used to build custom applications for interaction analysis. A command-line interface to the library is provided and used by SNPsyn's web server to perform interaction analysis.

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

Interaction analysis

Synergy (Syn) for a pair of SNPs (M_1 and M_2) is the difference between the information on phenotype P encoded in the newly derived feature (define by function f) and the sum of information encoded in the two individual features (4, 6):

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

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, 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 (9) or constructive induction by feature decomposition (10). 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 (6).

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. For this reason, data from other formats (PLINK's ped, tped (11) and tab-delimited file, where each row holds genotype information and other annotation on a sample) is first transformed into SNPsyn's compact binary format (see web site for 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 GO enrichment analysis

SNPs are mapped to genes using the mapping in NCBI's dbSNP database. Gene Ontology (GO) term enrichment analysis (12) 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 do not enter the GO term enrichment analysis. SNPsyn uses the hypergeometric distribution to compute the associated significance (p. values) and visualizes the results similarly as in GOAT (13). Currently, only human SNPs can be analyzed. The documentation incudes examples on how to prepare a local installation of SNPsyn for mouse or other species.

Permutation analysis and significance assessment

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), each time computing the scores for all pairs of SNPs. Two random data shuffling approaches are implemented in SNPsyn: permute class labels and permute 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 ($I_{rnd}>I$ and $Syn_{rnd}>Syn$) and N_{all} is the number of all tests performed on randomly permuted data. Obtained significance scores are corrected for multiple testing using the False Discovery Rate method described by Benjamini and Yekutieli (8).

Computational requirements and explorative interactivity

Exhaustive search for SNP synergies requires long processing times and may identify possibly large number of candidate synergistic SNP pairs. Best 5,000 SNP pairs with highest synergy and 5,000 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 30,000 SNPs (450M pairs, which require 9h of CPU time), because of the associated high computational costs and the desire to offer this service to a large number of researches. Although no fast and exact solution is known for this search problem, the XOR model being an example where most heuristics fail, some theoretical studies have demonstrated (14) the applicability of two-stage heuristic approaches. When more than 30,000 SNPs are given, the server will use a heuristic search: SNPs are initially scored based on tests of single SNPs. Only 30,000 SNPs with highest

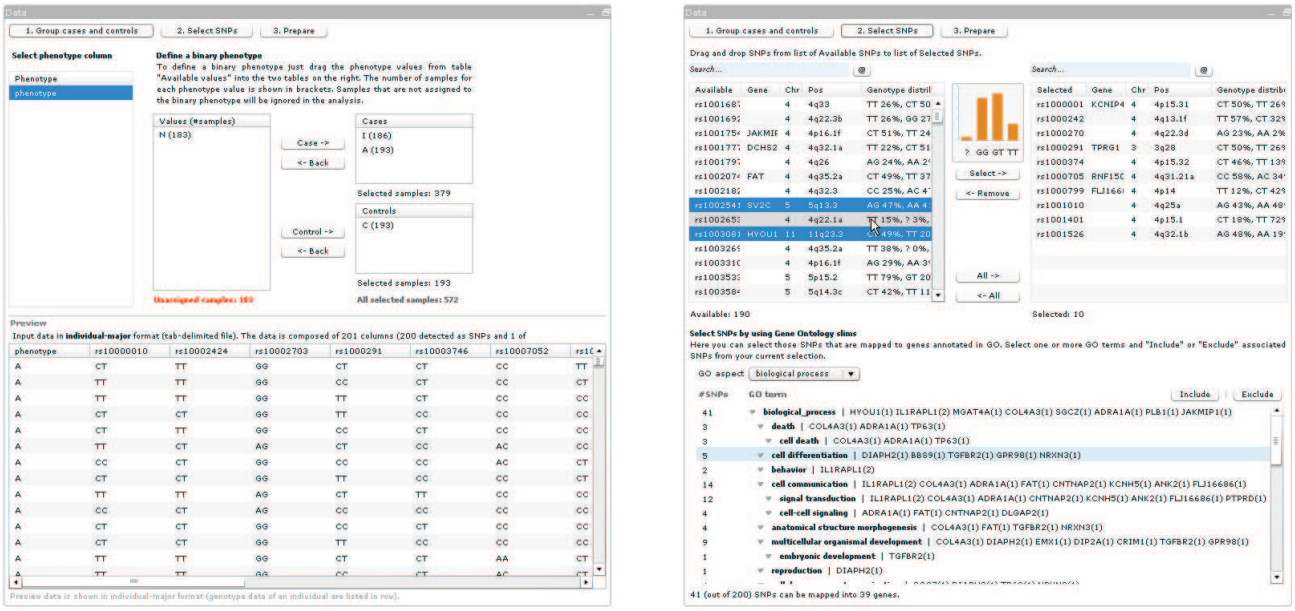

Figure 1. Data preparation. **(a)** Grouping of samples into cases and controls. **(b)** Selection of SNPs for analysis may be aided with Gene Ontology annotation.

information gain ($I(M_i|P)$) are then combined to discover high-order interactions within the amount of computational time allocated for the analysis.

If needed, constraints 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 run the analysis locally using the command-line utility.

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 PLINKs ped, tped formats and a tab-delimited format. PLINKs 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 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 (Fig. 1a)

Next, two approaches to SNP selection for synergy exploration are supported: the hypothesis-free *de novo* 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 for synergy. The latter approach allows to focus on a more specific biological question and also drastically reduces the number of SNP pairs to test. The user can hand-pick individual SNPs or subsets of SNPs associated with genes in specific, biologically relevant annotation terms in Gene Ontology (Fig 1b).

Data is then encoded in SNPsyn's compact binary format file and submitted to SNPsyn web server for analysis or stored locally to be analyzed on local computational facilities.

Visual exploration of results

The main results of data analysis is 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 $Syn(f(M_i,M_j)|P)$, which are presented to the user for exploration and subsequent analyses.

Calculated scores of SNP pairs on true data are plotted in a I *vs.* Syn scatterplot (red dots, Fig. 2a), and the null-distribution is superimposed (blue dots, Fig. 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 selected region is highlighted in blue on the scatterplot and associated SNP pairs are displayed in a table (Fig. 2b). There, the user can bookmark (star) favorite SNP pairs for latter access in a separate window and are included in the report.

Even after filtering by the four measures listed above, the list of best synergistic SNP pair can be quite extensive and may include a large number of false positives. Instead of just examining details on individual pairs (shown in Fig. 2e), one may benefit from exploring and reasoning on commonalities among genes associated with best-ranked pairs. For this, links to detailed information on SNP and gene annotation at NCBI and HapMap (15) are provided throughout the interface wherever a gene or SNP is shown. Additionally, the user can perform Gene Ontology (16) term enrichment analysis (Fig. 2c) and to visualize an interaction network (Fig. 2d).

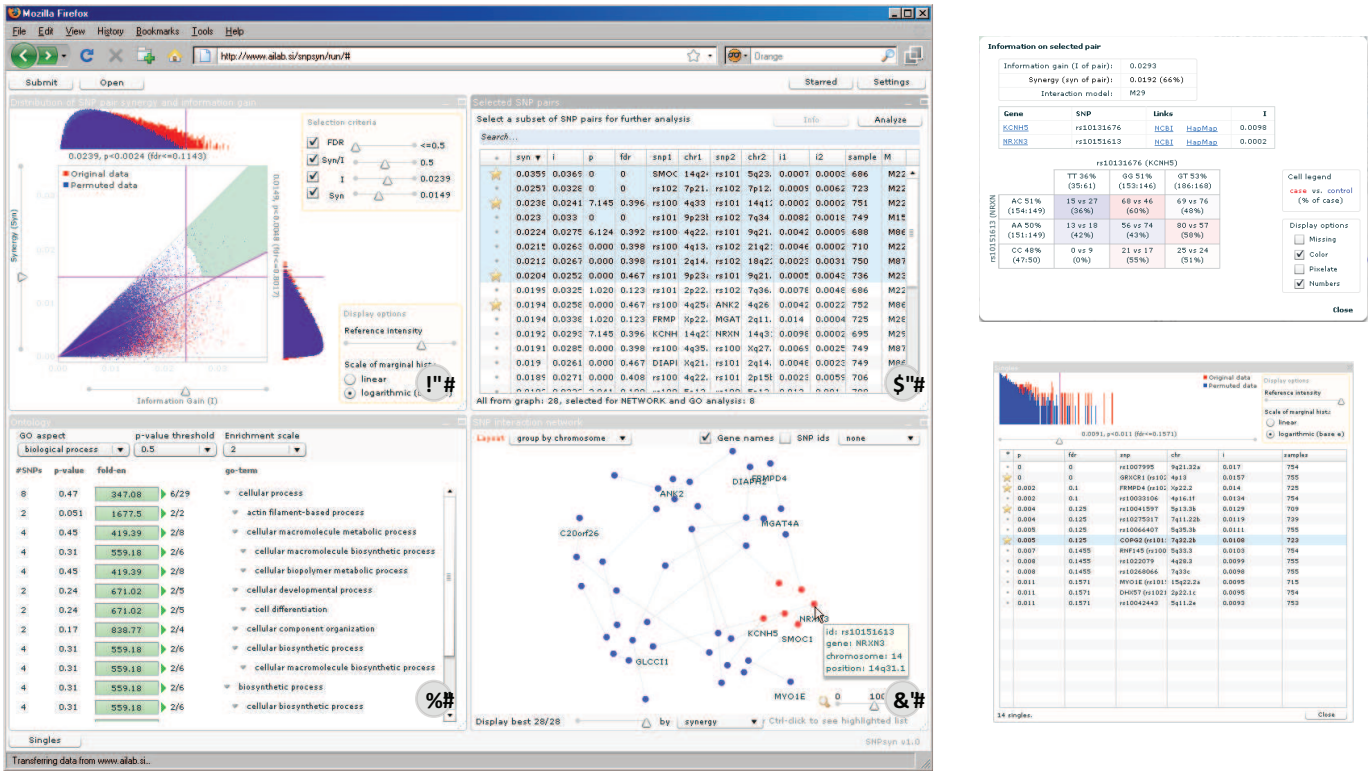

Figure 2. Exploration of results. **(a)** I *vs.* Syn scatterplots. **(b)** List of SNPs selected in (a). **(c)** Gene Ontology enrichment analysis of SNP pairs selected in (b). **(d)** Synergy network from SNP pairs selected in (b). **(e)** Details on a selected pair. **(f)** Results of individual SNP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

Enriched GO terms are drawn in a tree plot (Fig. 2c). Each row represent an enriched term and lists 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 get highlighted in green in the interaction network.

Nodes in a SNP interaction network (Fig. 2d) are connected if the two associated SNPs form a pair that was selected by the user in Fig. 2b. Three layouts of the network are available to survey the overall structure and quickly identify commonalities among interacting genes: 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"). The user can choose what to display on edges: Syn, I or none. 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 same gene or same chromosome get highlighted. 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 synergies, SNPsyn can also display the results of analysis of individual SNPs (Fig. 2f). These are available by clicking on the Single SNPs button (Fig. 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 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 (11), MDR (5), HFCC (17) and PLR (18). Two main features differentiate SNPsyns from other tools. 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 (PLINK, 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. A second distinctive feature of SNPsyns is its highly interactive, graphical user interface that supports all steps of an explorative analysis of synergy and structure of the gene interaction network. A planned improved of the interactivity and the extension to support gene set enrichment analysis (19, 20) will further improve the usability of SNPsyn.

CONCLUSION

With raising number of whole genome-wide case-control association studies of complex diseases, now being facilitated by high-throughput sequencing (21), appropriate bioinformatics and data analytics software tools are needed to support biologists in their search for relations between genotype and phenotype. The endeavor of developing such tools is not an easy one. Critical issues are, on one hand, computational speed and appropriate statistical treatment when dealing with low sample-to-feature rations and, on the other hand, presentation of results that can support data exploration and performing analytics tasks in a format accessible to biologists. 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.

A thorough exploration across sets of readily available GWAS data on many diseases (22) in now possible with SNPsyn, which could help answer the question on the role of gene synergy in complex diseases.

FUNDING

This work was supported by the Slovenian Research Agency [P2-0209, J2-2197, L2-1112, Z7-3665 to TC, GR and BZ] and by the Spanish Ministry of Science and Innovation [PSE-010000-2006-6 to A.J. and S.M.].

Conflict of interest statement. None declared.

REFERENCES

1. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78, 6 2007. [PubMed:17554300 [PubMed Central:PMC2719288] [doi:10.1038/nature05911].
2. MJ Daly and D. Altshuler. Partners in crime. *Nature genetics*, 37(4):337–8, 2005. [PubMed:15800643 [doi:10.1038/ng0405-337].
3. J. Gerke, K. Lorenz, and B. Cohen. Genetic interactions between transcription factors cause natural variation in yeast. *Science*, 323(5913):498–501, 1 2009. [PubMed:19164747 [doi:10.1126/science.1166426].
4. A. Jakulin and I. Bratko. Analyzing attribute dependencies. In *PKDD 2003, volume 2838 of LNAI*, pages 229–240. Springer-Verlag, 2003.
5. L. W. Hahn, M. D. Ritchie, and J. H. Moore. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, 19(3):376–82, 2 2003. [PubMed:12584123 [doi:10.1093/bioinformatics/btf869].
6. A. Dimitris. Computational analysis of the synergy among multiple interacting genes. *Mol Syst Biol*, 3:83, 2007. [PubMed:17299419 [PubMed Central:PMC1828751] [doi:10.1038/msb4100124].
7. W. Li and J. Reich. A complete enumeration and classification of two-locus disease models. *Hum Hered*, 50(6):334–49, 2000. [PubMed:10899752 [doi:10.1159/000022939].
8. Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001. [doi:10.1214/aos/1013699998].
9. S. Kramer. Cn2-mci: A two-step method for constructive induction. In *Proceedings of the ML-COLT-94 Workshop on Constructive Induction and Change of Representation*, 1994.
10. B. Zupan, M. Bohanec, J. Demsar, and I. Bratko. Learning by discovering concept hierarchies. *Artificial Intelligence*, 109(1):211–242, 1999. [doi:10.1016/S0004-3702(99)00008-9].
11. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–75, 9 2007. [PubMed:17701901 [PubMed Central:PMC1950838] [doi:10.1086/519795].
12. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9, 5 2000. [PubMed:10802651 [PubMed Central:PMC3037419] [doi:10.1038/75556].
13. Q. Xu and G. Shaulsky. Goat: An r tool for analysing gene ontology-trade mark term enrichment. *Appl Bioinformatics*, 4(4):281–3, 2005. [PubMed:16309346.
14. D. M. Evans, J. Marchini, A. P. Morris, and L. R. Cardon. Two-stage two-locus models in genome-wide association. *PLoS Genet*, 2(9):e157, 9 2006. [PubMed:17002500 [PubMed Central:PMC1570380] [doi:10.1371/journal.pgen.0020157].
15. K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, A. Boudreau, P. Hardenbolt, S. M. Leal, et al. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–61, 10 2007. [PubMed:17943122 [PubMed Central:PMC2689609] [doi:10.1038/nature06258].
16. K. Wang, M. Li, and M. Bucan. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet*, 81(6), 10 2007. [PubMed:17966091 [doi:10.1086/522374].
17. J. Gayán, A. González-Pérez, F. Bermudo, M. E. Sáez, J. L. Royo, A. Quintas, J. J. Galan, F. J. Morón, R. Ramirez-Lorca, L. M. Real, et al. A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis. *BMC Genomics*, 9:360, 2008. [PubMed:18667089 [PubMed Central:PMC2533022] [doi:10.1186/1471-2164-9-360].
18. M. Y. Park and T. Hastie. Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9(1):30–50, 1 2008. [PubMed:17429103 [doi:10.1093/biostatistics/kxm010].
19. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–50, 10 2005. [PubMed:16199517 [PubMed Central:PMC1239896] [doi:10.1073/pnas.0506580102].
20. M. Holden, S. Deng, L. Wojnowski, and B. Kulle. Gsea-snp: applying gene set enrichment analysis to snp data from genome-wide association studies. *Bioinformatics*, 24(23):2784–5, 12 2008. [PubMed:18854360 [doi:10.1093/bioinformatics/btn516].
21. The 1000 Genomes Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, 10 2010. [PubMed:20981092 [PubMed Central:PMC3042601] [doi:10.1038/nature09534].
22. The database of genotypes and phenotypes (dbgap), http://www.ncbi.nlm.nih.gov/gap.