CASMAP: Detection of statistically significant combinations of SNPs in association mapping

Felipe Llinares-López, Laetitia Papaxanthos, Damian Roqueiro,
Dean Bodenham and Karsten Borgwardt

Introduction

In this document, we first provide an overview of existing methods to perform epistasis search in genomic data. We explain where CASMAP fits in this context and describe how to use the software package to perform: 1) region-based association studies and 2) higher-order epistasis analyses.

CASMAP (**C**omputational **AS**ociation **MAP**ping) is implemented in C++ and its core functionality is exported through R and Python packages. The CASMAP package can be cloned from GitHub at `http://github.com/BorgwardtLab/CASMAP`. A detailed, step-by-step guide on how to compile, install and uninstall the software can be found in the file `README.md` located at the root of the package. Additionally, the R package can be downloaded from the standard CRAN repository[1].

S1 Epistasis search, an overview

From a biological perspective, epistasis is defined as an interaction between different genes in which the effects of one gene are masked by the effects of another gene (Cordell, 2002). From a statistical point of view, epistasis is normally considered to be a deviation from additive effects. In a regression equation, and when considering genotype data in the form on single nucleotide polymorphisms (SNPs) this can be modeled as:

$$y = \alpha_0 + \alpha_1 x_i + \alpha_2 x_j + \alpha_3 x_i x_j$$

where y is the phenotype, x_i and x_j are the genotypes of SNPs i and j, respectively (following a specific encoding). The magnitude of the epistatic interaction between the two SNPs is captured by the coefficient α_3 in the last term.

The search for **two-locus epistasis**, as exemplified above, yielded a plethora of methods and tools. Exhaustive enumeration and analysis of all SNP pairs can be performed with PLINK (Purcell *et al.*, 2007), GLIDE (Kam-Thong *et al.*, 2012) and BOOST (Wan *et al.*, 2010), to name a few. To avoid considering all possible pairs of SNPs, methods like MB-MDR (Calle *et al.*, 2010) perform a pre-selection of promising SNPs by performing multifactor dimensionality reduction.

In regards to **higher-order epistasis** analysis, traditional approaches have tried to circumvent the combinatorial explosion that arises when considering all possible subsets of SNPs by analyzing interactions of order k in an iterative manner. For example, the FITF approach (Millstein *et al.*, 2006) filters SNPs (or combinations of k SNPs) based on their marginal effects, where sets of order k are conditioned on significant lower-order effects. Another approach, like the one implemented in BHIT (Wang *et al.*, 2015), analyzes sets of SNPs randomly chosen from a candidate set. SNPs are pre-selected for the analysis of higher-order epistasis based on either of the following stategies: i) selection of individual SNPs based on a LASSO model of additive effects, ii) considering interactions only between SNPs in the same chromosome, and iii) using domain knowledge in the form of user-provided sets of SNPs to analyze. It is important to mention that neither of these approaches account for the multiple-comparison problem and thus their results are not comparable from a statistical perspective to those of CASMAP.

In summary, existing tools suffer from one or more of the following limitations:

1. They cannot search for higher-order interactions, focusing on pairwise interactions only.
2. They make use of "greedy" pruning criteria to reduce runtime, removing "non-promising" candidates from the search space without guaranteeing that no significant interactions will ever be wrongly removed during that process.

[1] `https://cran.r-project.org/web/packages/CASMAP/index.html`

3. They do not account for the multiple hypothesis testing problem and thus cannot adjust the resulting P-values to account for the number of interactions under study.

In contrast, what differentiates our approach from the ones mentioned above is that CASMAP overcomes all these limitations simultaneously. In particular, CASMAP (1) searches for interactions of any order, (2) uses a pruning criteria that can be shown to never remove significant interactions from the search space (Tarone, 1990) and (3) corrects for the multiple hypothesis testing problem by strictly controlling the family-wise error rate (FWER). Thus, we hope CASMAP will be of particular interest to practitioners who wish to exhaustively test *all* higher-order significant interactions for association and require the significance of the results to be quantified using multiplicity-adjusted P-values.

Numerous methods and tools have been developed to detect the association between sets of (interacting) SNPs and a trait of interest. The goal of this section was to present an overview of existing methods and to provide a context for CASMAP while highlighting its strong statistical and computational properties. For a thorough review of epistasis search at a genome-wide scale, we refer the reader to Cordell (2009) and Wei *et al.* (2014).

S2 Details about the input files

The input to the tools consists of: i) sample data, in the form of an $n \times l$ data matrix, for n individuals and l markers per individual, ii) the phenotype and iii) an optional covariate file to correct for confounding factors. The sample data will generally consist of standard PLINK files containing genotype information. For general-purpose mining tasks, CASMAP also allows sample data in the form of tab-separated text files as input. An important requirement is that both the sample data and the phenotype must be binary. For heterozygous genotypes, the data can be binarized following a dominant or recessive encoding. Binary phenotypes, on the other hand, are a standard staple in case-control studies.

S3 Representation of groups of SNPs (meta-marker)

In practice, each group of SNPs will be represented by a *meta-marker* whose association with the phenotype will be tested (see Figure 1 in the main document). The choice of how to compute the meta-marker follows a biological interpretation that depends on the type of analysis. In *region-based association studies* the meta-marker for each sample is encoded as a 1 if the genomic region contains at least one SNP with one minor allele—for the dominant encoding— or at least one SNP with two copies of a minor allele—for the recessive encoding. In other cases, the meta-marker is encoded as 0. The rationale is that single SNPs may have weak effects and, thus, may not reach significance. Nevertheless, the pooling of minor alleles that jointly give rise to the phenotype might be statistically significant. In *higher-order epistasis analysis* the meta-marker follows a multiplicative encoding, i.e., it is encoded as a 1 if all SNPs in the set have a minor allele—for the dominant encoding—or have two copies of the minor allele—for the recessive encoding. The existence of the minor allele(s) here represents the presence of *risk*. Therefore, this model assumes that the interaction between a set of SNPs exists if all of them have risk alleles.

S4 Use case 1: Region-based association study

In this section, we discuss an example of how to use CASMAP to carry out a region-based association study in a real-world dataset from the plant model organism *A. thaliana*. For this study, we downloaded from the easyGWAS online resource (Grimm *et al.*, 2017) the genotype data and the binary phenotype *avrB*. This dataset contains the genotypes of 87 samples measured at 214, 032 homozygous SNPs. The folder examples/data/region_based/avrB in the CASMAP software package includes this dataset. The Python and R scripts to reproduce all steps of this analysis can be found under examples/code. More precisely, we will use the following files:

1. A Jupyter notebook run_region_based.ipynb, containing all Python code necessary to execute the analysis using Python 2.x.

2. A Jupyter notebook run_region_based.Rmd, containing all R code necessary to execute the analysis using R.

3. The folder `examples/data/region_based/avrB`, with files `X.dat` (genotype), `Y.dat` (binary phenotype) and `C.dat` (categorical covariate to correct for population structure).

Each step in the Jupyter notebooks is commented in detail, aiming to provide a comprehensive and guided explanation of how to perform region-based association studies using `CASMAP`.

The output of the scripts includes code profiling information, high-level statistics of the region-based association study method and a list of significantly associated genomic regions, both before and after clustering together (potentially redundant) associated regions that overlap with each other. More precisely, the output is composed of the following files:

1. A **text file with information about code profiling** that indicates the computational resources used throughout the analysis, in terms of total execution time and peak memory usage. In addition, it provides a breakdown of the runtime in terms of: 1) the initialization time, 2) the I/O time, 3) the time to compute the corrected significance threshold and 4) the time required to find the significant genomic regions. Additional details about the computation of the corrected significance threshold can be found in Llinares-López *et al.* (2017).

2. The **summary text file** enumerates the main characteristics of the dataset. Firstly, it indicates the total number of samples and, of these, how many are *cases*. It also details the total number of markers over the whole dataset. If a categorical covariate is used in the analysis, this information is followed by the number of samples for each category of the covariate. Secondly, it gives high-level statistics of the output of the significant pattern mining algorithm such as the total number of genomic regions that were enumerated (which gives an indication of the efficiency of the branch-and-bound pruning criterion), the resulting corrected significance threshold and the total number of significantly associated genomic regions before clustering. Refer to Llinares-López *et al.* (2017) for more technical details.

3. The **list of significantly associated genomic regions before clustering** shows all regions (i.e., sets of contiguous markers) whose corresponding P-values are lower than the corrected significance threshold, with one region per row. If the input dataset is provided in PLINK format, the description of each genomic region complies with the following format:

P-value	score	OR	rsID$_1$;rsID$_2$;...;rsID$_N$	chr:pos$_1$;chr:pos$_2$;...;chr:pos$_N$
9.99	--	--	--	--

   where

   - `score` is the test score
   - `OR` is the odds ratio

   Note: Columns in the file are tab-separated.

   In contrast, if the input dataset is provided as a tab-separated text file, with no additional information about the SNP rsIDs or positions, each genomic region will be described as:

P-value	score	OR	index$_1$;index$_2$;...;index$_N$
9.99	--	--	--

   where index$_i$ refers to the position of the ith. SNP in the tab-separated text file.

4. The **list of most significant genomic regions after clustering** outputs the most significant genomic region in each cluster. As mentioned before, a cluster is a group of statistically significant associated regions that overlap. The most significant regions are output row by row. It can be considered to be a summary of the file previously described (before clustering), where redundant or overlapping regions have been removed. We refer the reader to Section S6 for details about the clustering algorithm and the selection of the representative region of each cluster.

   When the input dataset is provided in PLINK format, the SNPs in each region are identified according to their rsIDs and positions. The most significant regions per cluster are displayed as:

P-value	score	OR	rsID$_1$;...;rsID$_N$	chr:pos$_1$;...;chr:pos$_N$	#n regions	chr:pos L SNP	chr:pos R SNP
9.99	--	--	--	--	--	--	--

where

- `score` is the test score
- `OR` is the odds ratio
- `#n regions` is the number of overlapping regions in the cluster
- `chr:pos` L `SNP` is the chromosome and position of the leftmost SNP in the cluster
- `chr:pos` R `SNP` is the chromosome and position of the rightmost SNP in the cluster

In contrast, when the input dataset is a tab-separated text file, the SNPs are identified by their index in the array. The most significant region per cluster is then represented as:

P-value	score	OR	index$_1$;index$_2$;...;index$_N$	#n regions	index L SNP	index R SNP
9.99	--	--	--	--	--	--

where

- index$_i$ refers to the position of the ith. SNP
- `index` L `SNP` is the index of the leftmost SNP in the cluster
- `index` R `SNP` is the index of the rightmost SNP in the cluster

S5 Use case 2: Higher-order epistasis search

Similarly to the previous section, here we discuss an example of how to use `CASMAP` focused on higher-order epistasis analyses, using a real-world dataset from the plant model organism *A. thaliana*. Since higher-order epistasis search has a considerably higher computational burden than region-based association studies, we subsampled the dataset in the previous use case by keeping 1 every 80 SNPs from chromosome 1. The resulting dataset contains the genotypes of 87 samples, measured at 650 homozygous SNPs. The folder `examples/data/higher_order_epistasis/avrB` in the `CASMAP` software package includes this dataset. The Python and R scripts to reproduce all steps of this analysis can be found under `examples/code`. We will make use of the following files:

1. A Jupyter notebook `run_higher_order_epistasis.ipynb`, containing all Python code necessary to execute the analysis using Python 2.x.

2. A Jupyter notebook `run_higher_horder_epistasis.Rmd`, containing all R code necessary to execute the analysis using R.

3. Folder `examples/data/higher_order_epistasis/avrB`, with files `X.dat` (genotype), `Y.dat` (binary phenotype) and `C.dat` (categorical covariate representing population structure).

Both Jupyter notebooks are commented in as much detail as those in the previous section, providing a self-contained guide on how to employ `CASMAP` for higher-order epistasis search.

As previously, the output of the scripts includes code profiling information, high-level statistics of the higher-order epistasis search algorithm and a list of significantly associated multiplicative interactions between markers. The aim in designing the output files was to maximize consistency between both usages of `CASMAP`: region-based GWAS and higher-order epistasis search. More precisely, the following files will be generated after carrying out a higher-order epistasis analysis with `CASMAP`:

1. A **text file with information about code profiling** identical to the one described in Section S4. Please, refer to Papaxanthos *et al.* (2016) for details about the definition and computation of the corrected significance threshold.

2. The **summary text file** enumerates the main characteristics of the dataset. Following the same structure as for the region-based analysis, the file first lists the number of samples and of cases, in addition to the total number of markers in the whole dataset. If a categorical covariate is provided, this information is followed by the number of samples for each category of the covariate. The file also provides high-level statistics about the output of the significant pattern mining algorithm. For example, it shows the total number of interactions of markers that were enumerated (which gives an indication of the efficiency of the branch-and-bound pruning criterion), the resulting corrected significance threshold and the total number of significantly associated interactions of markers. Refer to Papaxanthos *et al.* (2016) for technical details.

3. The **list of significantly associated multiplicative interactions of markers** shows all interactions whose corresponding P-values are lower than the corrected significance threshold, where each row corresponds to one significant interaction. If the input dataset is provided in PLINK format, the description of each interaction complies with the following format:

`P-value`	`score`	`OR`	`rsID`$_1$`;rsID`$_2$`;...;rsID`$_N$	`chr:pos`$_1$`;chr:pos`$_2$`;...;chr:pos`$_N$
`9.99`	`--`	`--`	`--`	`--`

where

- `score` is the test score
- `OR` is the odds ratio

Note: Columns in the file are tab-separated.

In contrast, if the input dataset is provided as a tab-separated text file, with no additional information about the SNP rsIDs or positions, each significant interaction will be described as:

`P-value`	`score`	`OR`	`index`$_1$`;index`$_2$`;...;index`$_N$
`9.99`	`--`	`--`	`--`

where `index`$_i$ refers to the position of the ith. SNP in the input file.

S6 Clustering as a post-processing step in region-based association studies

The significantly associated genomic regions retrieved by a region-based analysis may overlap. In order to summarize (potentially redundant) overlapping genomic regions, we implemented a clustering algorithm (Llinares-López *et al.*, 2017) that groups together overlapping regions into large contiguous regions, called *clusters*. It is important to note that different clusters are mutually disjoint. Each cluster is then represented by its *cluster representative*, which is the genomic region included in the cluster with the smallest P-value (see Figure S1). If two regions are equally significant, the algorithm returns the longest one.

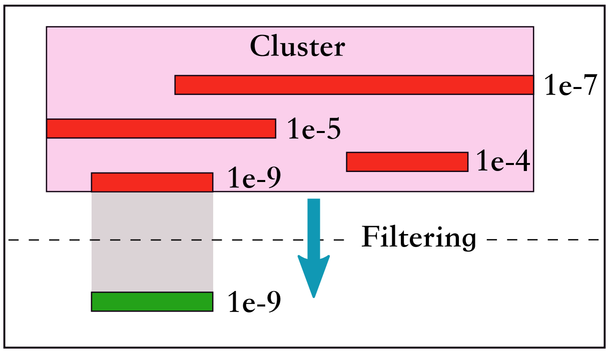

Figure S1: Four overlapping significant regions (in red) form a single cluster (in magenta). The cluster representative is the green region, since it has the smallest P-value, p=1e-9. Source: Supplementary Material of Llinares-López *et al.* (2017)

S7 Additional resources

The GitHub repository for `CASMAP` at `https://github.com/BorgwardtLab/CASMAP` contains step-by-step examples on how to run the tool as well as a detailed tutorial covering different use cases on various types of genomic data.

S8 Citations

If you perform a **region-based association study**, in addition to citing this package, please include the following manuscript among your citations:

> Llinares-Lopez, F., et al. *Genome-wide genetic heterogeneity discovery with categorical covariates*, Bioinformatics 2017.

If you perform a **higher-order epistasis search**, in addition to citing this package, please include the following manuscript among your citations:

> Papaxanthos, L., et al. *Finding significant combinations of features in the presence of categorical covariates*, NIPS 2016.

References

Calle, M. L., Urrea, V., Malats, N., and Van Steen, K. (2010). mbmdr: an R package for exploring gene-gene interactions associated with binary or quantitative traits. *Bioinformatics*, **26**(17), 2198–2199.

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, **11**(20), 2463.

Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **10**(6), 392–404.

Grimm, D. G., Roqueiro, D., Salomé, P. A., Kleeberger, S., Greshake, B., Zhu, W., Liu, C., Lippert, C., Stegle, O., Schölkopf, B., Weigel, D., and Borgwardt, K. M. (2017). easygwas: A cloud-based platform for comparing the results of genome-wide association studies. *The Plant Cell*, **29**(1), 5–19.

Kam-Thong, T., Azencott, C. A., Cayton, L., Putz, B., Altmann, A., Karbalai, N., Samann, P. G., Scholkopf, B., Muller-Myhsok, B., and Borgwardt, K. M. (2012). GLIDE: GPU-based linear regression for detection of epistasis. *Hum. Hered.*, **73**(4), 220–236.

Llinares-López, F., Papaxanthos, L., Bodenham, D., Roqueiro, D., and Borgwardt, K. (2017). Genome-wide genetic heterogeneity discovery with categorical covariates. *Bioinformatics*, **33**(12), 1820–1828.

Millstein, J., Conti, D. V., Gilliland, F. D., and Gauderman, W. J. (2006). A testing framework for identifying susceptibility genes in the presence of epistasis. *Am. J. Hum. Genet.*, **78**(1), 15–27.

Papaxanthos, L., Llinares-Lopez, F., Bodenham, D., and Borgwardt, K. (2016). Finding significant combinations of features in the presence of categorical covariates. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2279–2287. Curran Associates, Inc.

Purcell, S. *et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**(3), 559–575.

Tarone, R. E. (1990). A modified Bonferroni method for discrete data. *Biometrics*, **46**(2), 515–522.

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L., and Yu, W. (2010). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.*, **87**(3), 325–340.

Wang, J., Joshi, T., Valliyodan, B., Shi, H., Liang, Y., Nguyen, H. T., Zhang, J., and Xu, D. (2015). A Bayesian model for detection of high-order interactions among genetic variants in genome-wide association studies. *BMC Genomics*, **16**, 1011.

Wei, W. H., Hemani, G., and Haley, C. S. (2014). Detecting epistasis in human complex traits. *Nat. Rev. Genet.*, **15**(11), 722–733.