OXFORD UNIVERSITY PRESS
Briefings in Bioinformatics

Travelling the world of gene-gene interactions

Journal:	*Briefings in Bioinformatics*
Manuscript ID:	BIB-10-0143.R1
Manuscript Type:	Paper
Date Submitted by the Author:	n/a
Complete List of Authors:	van Steen, Kristel; Université de Liège, Electrical Engineering and Computer Science
Keywords:	gene-gene interaction, variable selection, controlling false-positives, translational medicine

Travelling the world of gene-gene interactions

Kristel Van Steen[1,2]

Affiliations:

[1] Systems and Modeling Unit, Montefiore Institute, University of Liège, Grande Traverse 10, 4000 Liège, Belgium

[2] Bioinformatics and Modeling, GIGA-R, University of Liège, Avenue de l'Hôpital 1, 4000 Liège, Belgium

Bio:

Kristel Van Steen is an associate professor of bioinformatics at the University of Liège and an ambassador for Belgium for the International Genetic Epidemiology Society IGES.

Key points:

- It is clear that epistasis plays a important role in human genetics, but it is less clear how to best bridge the gap between biological/genetical and statistical epistasis
- Over the last few years the field has seen an explosion of methodological developments to either directly or indirectly test for epistasis
- Whatever strategy is chosen, the analyst has yet to find a clever solution to overcome the burden of dimensionality, and to handle a severe multiple testing problem, while adequately controlling the number of false positives.

- The exploitation of several omics data bases, while performing an epistasis analysis, may substantially improve clinical decision making and therefore patient outcome.

Correspondence:

Kristel Van Steen

Department of Electrical Engineering and Computer Science (Montefiore Institute)

Grande Traverse, 10 - BAT. B28 Bioinformatique

4000 Liège 1

Belgium

Tel: +32 4 366 2692

Email : Kristel.VanSteen@ulg.ac.be

Abstract

Over the last few years, main effect genetic association analysis has proven to be a successful tool to unravel genetic risk components to a variety of complex diseases. In the quest for disease susceptibility factors and the search for the "missing heritability", supplementary and complementary efforts have been undertaken. These include the inclusion of several genetic inheritance assumptions in model development, the consideration of different sources of information, and the acknowledgement of disease underlying pathways of networks. The search for epistasis or gene-gene interaction effects on traits of interest is marked by an exponential growth, not only in terms of methodological development, but also in terms of practical applications, translation of statistical epistasis to biological epistasis, and integration of omics information sources. The current popularity of the field, as well as its attraction to interdisciplinary teams, each making valuable contributions with sometimes rather unique viewpoints, renders it impossible to give an exhaustive review of to-date available approaches for epistasis screening. The purpose of this work is to give a perspective view on a selection of currently active analysis strategies and concerns in the context of epistasis detection, and to provide an eye to the future of gene-gene interaction analysis.

Key Words: gene-gene interaction, variable selection, controlling false-positives, translational medicine

INTRODUCTION

That epistasis plays a role in human genetics is without doubt, given the numerous discoveries of significant gene-gene interactions in model organisms, providing evidence for interactions in the presence and absence of important individual effects [1], and the insights gained in cell biology showing complex interactions between different types of biomolecules [2]. Epistasis and genomic complexity are correlated, in the sense that in less complicated genomes mutational effects involved in epistasis tend to cancel each other out, whereas in more complex genomes mutational effects rather strengthen each other, leading to so-called synergetic epistasis [3, 4]. Hence, dependencies among genes in networks, leading to epistasis, naturally arise when believing that the human system guards itself to negative evolutionary effects of mutations via redundancy and robustness [5]. It is therefore not surprising that, with a growing tool-box of analysis techniques and approaches, the number of identified epistatis effects in humans, showing susceptibility to common complex human diseases, follows a steady growth curve [6, 7].

Field Code Changed

Field Code Changed

But what is meant by epistasis? William Bateson [8] defined it from a biological viewpoint as distortions of Mendelian segregation ratios due to one gene masking the effects of another. Statisticians adopt another viewpoint. For them, like Fisher [9], interactions represent departures from a linear model that describes how two or more predictors predict a phenotypic outcome. The presence and magnitude of non-additivity are scale and model dependent, so that in principle, one strategy in the context of an epistasis analysis could be to remove any non-additivity by a transformation prior to data

analysis, followed by a back-transformation to the original scale for easy interpretation [10]. This is the path least travelled by in practice.

Regression-based approaches are still seen as the most natural first-line approach for modeling of and testing for interactions [11], despite many difficulties this approach brings along, whether from a technical, computational, or interpretation point of view. The unability to identify epistasis using statistical tools may simply be due to insufficient sample size and hence inadequate theoretical power to detect statistical epistasis (occurring as the result of differences in genetical and biological epistasis among individuals in a population, [9]). Remarkably, it is perfectly possible for genetical and biological epistasis (both occurring at the individual level, [9]) to exist in the absence of statistical epistasis, simply as an artifact of the sample's characteristics, even with sufficiently large samples [5]. For a comprehensive discussion about the meaning of epistasis and its consequences for analysis, we refer to the recent paper of Wang et al [12]. Notably, over the last few years, many reports of statistical epistasis have been made, involving a variety of study designs, analysis techniques and human diseases. However, so far, only for some of the reported findings additional support could be provided by functional analysis [13], as was the case for multiple sclerosis [14]. The future will reveal whether the latter observation should be seen as a consequence of a possible negligible role of epistatic variance in a population [15], or rather as a consequence of not yet available powerful epistasis detection methods.

The remainder of the paper is organized as follows. We first discuss several strategies to identify epistasis. We structure these strategies according to those who are exploratory in nature, and those who are more targeted, while putting more structure in the (statistical) models used. For the majority of these strategies, computation time can be substantially improved by appropriate variable selection. Second, we highlight some of the most relevant hurdles to take when performing a large-scale epistasis screening and show by means of current state-of-the-art developments how they can be adequately addressed. Finally, we give a perspective view on the importance of epistasis screening for personalized medicine.

IDENTIFICATION STRATEGIES FOR STATISTICAL EPISTASIS

General setting

The space of possible epistasis models is infinitely large, and almost every purely epistatic model occurring in practice is expected to include both incomplete penetrances and phenocopies, in some sense "blurring" the picture [16]. In an attempt to get a handle on the wide variety of possible multi-locus models, Li and Reich [17] drafted a classification of all two-locus, fully penetrant disease models (binary trait – 512 models). These can be further reduced to 50 classes of equivalent models, with varying degrees of epistasis. Via geometric arguments, Hallgrímsdottir and Yuster [18] showed that there are 387 distinct types of two-locus models with continuous penetrance values, which again can be reduced to a much smaller number (in this case, 69) when symmetry in the epistasis models is accounted for.

In addition, the abundance of developed strategies in the context of epistasis detection clearly complicates a rigorous classification. Nevertheless, in the past, several authors have used a variety of criteria, in the attempt to categorize the methodologies used. These include criteria about 1) whether the strategy is exploratory in nature or not, 2) whether modeling is the main aim, or rather testing, 3) whether the approach is parametric or non-parametric, 4) whether the epistatis effect is tested indirectly or directly, 5) whether or not the method is able to distinguish between epistasis and other signals, and 6) whether the strategy uses exhaustive search algorithms or whether screening is based on a reduced set of input-data, that may be derived from prior expert knowledge or some filtering approach. Obviously, there is some overlap between the described classification schemes, and no pair of schemes is mutually exclusive. It is therefore not surprising that already many reviews on the topic exist. A non-exhaustive display of different methods is given in for instance Onkamo and Toivonen [19], Musani et al [20], Cordell [11, 21] and Kilpatrick and Nakhleh [22].

Field Code Changed

Field Code Changed

Variable selection: a must?

One of the problems with high-dimensional datasets is that usually not all the measured variables are important for understanding the underlying phenomena of interest. Hence, a balance needs to be found between making most of hard to acquire data using computationally expensive methods and reducing the dimension of the original data prior to any modeling or detailed analysis. In this context, two concepts play a crucial role: feature extraction and feature selection. Feature extraction [23] aims to reduce dimensionality by aggregation or projection. Feature selection simply involves looking

for optimal subsets of variables, so as to reduce storage requirements during data analysis and to reduce the waiting time for analysis results to be generated. Feature selection methods tend to avoid over-fitting, to improve model performance, and to enhance data understanding. They can be classified as "filter", "wrapper" or "embedded" methods (Table 1). Whereas filters select subsets of variables independently of the chosen subsequent analysis method, as a pre-processing step, wrappers use the particular classifier/discriminator tool to score subsets of variables according to their predictive power. Embedded methods perform variable selection during a training step and are usually specific to the chosen learning machine. Notably, in contrast to dimensionality reduction techniques like those based on projection (e.g., Principal Components Analysis [24]), feature selection techniques do not change the original presentation of the variables, while reducing the burden of multiple testing. More details on variable selection methods can be retrieved from Guon et al [25]. For a thorough review of feature selection methods in bioinformatics applications, we refer to Saeys et al [26].

TABLE 1: about here

Because it improves genetic and biological meaning of epistasis analyses, it is not surprising that a popular concept to filter SNPs for epistasis analysis is "synergy" [28]. In the bivariate case, this quantity represents the additional information that both genetic factors jointly provide about the phenotype, after removing the individual information provided by each genetic factor separately [38]. Bearing in mind this representation, a synergy-based analysis can be performed as a stand-alone method to detect gene-gene

interactions. However, traditionally, information-theoretic measures have mostly been used as a means to select "informative" variables in a variety of fields within and outside the pharmaceutical or health sciences. If significance needs to be assessed, the user needs to turn to permutation strategies, or bootstrapping strategies that involve re-sampling with replacement via random samples of the original data's sample size. Especially when concerns about computational feasibility arise, using proxies for computing relevance and redundancy among variables can provide a way out. In particular, computing $Syn(X_1,X_2;Y)$ for every pair of markers X_1 and X_2, will allow a ranking of pairs of markers according to the gain in mutual information of SNP1 (X_1) and SNP2 (X_2), due to a class variable Y. Chanda et al [39-41] described a more general framework of entropy-based measures for epistasis detection, hereby allowing for higher-order interactions and accommodating scenarios of categorical trait values with more than two classes, as well as markers or environmental factors with varying number of factor levels [40]. Entropy is a measure of randomness or disorder within a system. The lower the entropy, the higher the likelihood that the system is in a more stable state and consequently, the more likely our predictions will be. No matter how popular epistasis screening based on entropy-like measures may be [27], these entropy-based measures are less commonly used in the light of quantitative trait analysis (although Shannon's entropy [42], defined for a discrete random variable, is easily extended to situations when the random variable Y under consideration is continuous, in which case it is then sometimes referred to as "differential entropy").

Field Code Changed

Field Code Changed

Notably, there is a correspondence between mutual information [43] and the coefficient of determination in a regression framework. Indeed, mutual information *I* can be expressed as a Kullback-Leibler directed divergence, of the product of the marginal distributions of two random variables, for instance (a predictor) *X* and (an outcome) *Y*, from the random variables' joint distribution. Although mutual information is symmetric in its components, it is not a symmetric distance between the corresponding aforementioned densities. With a symmetric version of this distance, *J*, the coefficient of determination of *Y* by *X* through μ can be defined as $R^2_J = \frac{J(\mu(X),Y)}{1+J(\mu(X),Y)}$ [44]. Here, μ is a parameter that determines the distribution of the response *Y* as a function of independent variables *X* and regression coefficients. In case *(X,Y)* follow a bivariate Gaussian distribution, it can be shown that $R^2_J = \rho^2$, with ρ^2 the usual correlation coefficient of *Y* with *X*. It can also be shown that the correlation coefficient ρ is related to the mutual information $I(\mu(X), Y)$ as $I(\mu(X), Y) = -\frac{1}{2}\log(1-\rho^2)$ [45]. Hence, in this special case, by again setting $R^2_I = \rho^2$, R^2_I can be derived from the directed divergence by defining $R^2_I = 1 - \exp(-2I(\mu(X), Y))$ [44]. Therefore, the definitions proposed for R^2_I and R^2_J not only generalize the R^2 from classical linear regression, but also apply to generalized regression models with arbitrary link functions, as well as multivariate and non-parametric regression.

The relationship between the well-known concept of coefficient of determination and mutual information opens up some interesting avenues to consider mutual information-based measures of association, such as $J(\mu(X), Y)$ or R^2_J, for variable or model selection in the context of epistasis screening. A growing literature on how to optimally estimate

the aforementioned generalized measures and on how to derive confidence or credibility bounds around them in fast and efficient way, makes them particularly interesting as a stand-alone method to detect gene-gene interactions [46].

Another filtering method is the ReliefF algorithm [29], which is able to acknowledge SNP-group effects. This advantage is also a disadvantage because the presence of many noisy attributes can actually reduce the interaction signal the algorithm is trying to capture. This understanding led to another multivariate filtering technique, which systematically removes attributes of insufficient quality (TuRF [30]). The similarity between the Relief weight and the Gini index (a feature evaluation measure in Random Forests) has been previously discussed by Kononenko and Robnik-Sikonia [47]. Within the same family, Spatially Uniform ReliefF [31] allows computationally efficient filtering of specifically gene-gene interactions. Also Evaporative Cooling filtering (ReliefF combined with entropy) has proven to be a promising filtering approach [32].

Examples of two-stage approaches in which SNPs are selected according to some criterion and subsequently considered for epistasis analysis include the Focused Interaction Testing Framework (FITF) of Marchini et al [48], Model-Based Multifactor Dimensionality Reduction (MB-MDR) after entropy-based feature selection of Calle et al [49], or the “MDR flexible framework” approach of Moore et al [50]. Apart from greedy algorithms that perform filtering based on non-epistatic or lower-order interaction results, stochastic approaches are quite common in the field as well. These approaches also perform a partial search in the interaction space, but select small numbers of loci in an

iterative fashion (e.g., Random Forest (RF)-based prescreening method prior to executing an MDR (Multifactor Dimensionality Reduction) scan [51], Random Jungle [52], SNPharvester [53] or Bayesian Epistasis Association Mapping (BEAM – [54]).

Field Code Changed

To enhance genome-wide analysis of common human diseases with a complex genetic architecture, Moore and White [50] developed and evaluated a simple Genetic Programming wrapper for attribute selection. One of the advantages of genetic programming is that it naturally provides a set of competing models with comparable fits. Also the procedure Genetic Programming for Association Studies (GPAS, [33]) exploits this advantage. Alternatively, Greene et al [34] suggested ant colony optimization (ACO) as a useful wrapper in the presence of complex systems of interactions. The application of ACO to data mining techniques requires the transformation of the optimization problem into the problem of finding the best path on a weighted graph: The field of ACO is a translation of the attempt to develop algorithms inspired by the ability of ants to find shortest paths [55]. Artificial ants incrementally build solutions by moving on the graph, a process that therefore allows incorporating expert (biological) knowledge. An application of its principles is laid out in the AntEpiSeeker epistasis searching tool [35].

Field Code Changed

Examples of embedded variable selection methods are decision-tree based methods (see next section). One of the main advantages of these methods is that they are able to "model" feature dependencies.

Let the data speak for themselves

Simple "exploration" of huge amounts of data is just one step of a so-called data mining process. Data mining techniques are much more comprehensive in that they also involve model building or pattern identification and choosing the best model based on selected criteria, as well as the application of that model to new data in order to generate predictions. Naively, exploratory data analysis techniques can be further grouped in i) data segmentation methods, such as clustering methods [56], ii) tree-based methods [57], such as Recursive Partitioning, Random Forests and Logic Regression [36, 37], iii) pattern recognition methods [58], such as Symbolic Discriminant Analysis, Support Vector Machines, Mining Association rules and Neural Networks, and iv) multidimensional reduction methods (i.e. a form of feature extraction methods in which the data are projected or embedded into a lower dimensional space while retaining as much information as possible), such as Principal or Independent Components, Multidimensional Scaling, Detection of Informative Combined Effects (DICE), Polymorphism Interaction Analysis (PAI, [59]), Multifactor Dimensionality Reduction (MDR, [60]) and Model-Based Multifactor Dimensionality Reduction (MB-MDR).

Field Code Changed

Most of these methods examine the combination effect simultaneously and test the epistatic effect implicitly, while adopting a global null hypothesis (Table 2). Although this strategy is able to alleviate some of the multiple testing problem, a more detailed follow-up analysis is needed when the detection of epistasis (above and beyond main effects) is envisaged. Examples of these methods include the Combinatorial Partitioning Method [61], the Restricted Partitioning Method [62, 63], Multi-locus Penetrance Variance Analysis [64], (MCMC) Logic Regression [65, 66], Backward Genotype-Trait

Field Code Changed

Association [67], Bayesian Epistasis Association Mapping (BEAM – [54]), Genetic Ensemble algorithmic epistasis search (GE – [68]), Logic Forests [69] and Grammatical Evolution Neural Networks (GENN – [70]).

TABLE 2: about here

Especially for large sample sizes, there is a clear benefit of Random Forests algorithms [71-73] over regression-based approaches [80]: The initial algorithms have recently been further adapted for fast and computationally efficient analysis of GWAs and coined Random Jungle [52]. Another beauty of Random Forests methodology is that its principles can be generalized to other methods, such as Random Multinomial Logit [74], Random Naïve Bayes [75], or adapted to accommodate cluster-correlated data [76, 77]. EpiForest [72] combines a Random Forests analysis with a sliding window sequential forward feature selection (SWSFS) algorithm.

Field Code Changed

Field Code Changed

Interestingly, combining information over "ensembles" has also proven to be beneficial in developing methods to separate purely epistatic effects from other signals in the data. Although not in the context of tree-building, Wongseree et al [81] developed an algorithm for ensembles of two-locus non-parametric analyses, leading to an omnibus permutation test for pure epistasis.

As mentioned before, Multifactor-Dimensionality Reduction [60] also belongs to the category of "exploratory methods". Although MDR has been widely used for interaction detection, it suffers from some major drawbacks including that important interactions

could be missed due to pooling too many multi-locus genotype cells together and that it cannot adjust for lower-order genetic effects (that are possibly components of a higher-order interaction of interest). Therefore, a (potentially) model-based version, MB-MDR [49, 78, 79], was developed. Unlike MDR, MB-MDR controls false positives under any configuration of true and false null hypotheses, if the condition of hypothesis subset pivotality is fulfilled, is able to assess joint significance of multiple higher-order interaction models at once, and facilitates distinguishing between epistatic effects and contributing main effects to the multi-locus signal via the "MB" part in MB-MDR [82]. At least for quantitative trait loci it has been shown that increased efficiency can be attained when interacting loci are searched for simultaneously [83]. Because of the Model-Based component in MB-MDR, more structure can be imposed to the modeling of multi-locus effects and epistasis can be tested directly.

Field Code Changed

Imposing assumptions about the functional form of models and the effects being modeled

Perhaps one of the most important lesson learned from thorough investigations for epistatic effects in model organisms, is that multi-factorial traits are driven by complex systems that do not let themselves be described by simple and uniform modes of inheritance, hereby leading to varying levels of epistasis throughout the genome [1]. The necessity to develop tools that are flexible and are able to accommodate variable modes of inheritance when screening for gene-gene interactions is a major motivation for those who advocate the use of non-parametric epistasis detection methods.

However, for genetic association studies (parametric) regression analysis remains the most commonly used paradigm. Here, the disease trait is usually considered as a response variable and the coded genotype(s) as predictor variable(s). Obviously, the validity of analysis conclusions crucially depends on the underlying model assumptions. Despite the wide-spread use of regression-based approaches, these traditional methods often fail due to 1) the large number of genotyped polymorphisms requiring very small p-values for significance assessment, 2) the "curse of dimensionality" [84] or the fact that the convergence of any parametric model estimator to the true value of a smooth function defined on a space of high dimension is very slow, 3) the presence of important interacting loci with relatively small marginal effects, 4) the abundance of rare (or absent) multi-locus genotype combinations with increasing dimensionality.

Nevertheless, one of the artifacts of methods that allow putting more structure on the data compared to classical data exploration techniques is that it easily accommodates testing both the main effect and the epistatic effect explicitly (e.g., [85, 86]), as we have seen with the (semi-parametric) MB-MDR method. On the downside, whenever a direct test for epistasis is the target, one has to realize that different choices of scale may lead to different implications of epistasis … For instance, the additive model defined on the outcome scale as a sum of effects at contributing loci is a non-epistatic model, whereas the multiplicative model is epistatic, yet both formalisms give similar results when used to model familial risks of disease [58]. "Compositional epistasis" is said to be present when the effect of a genetic factor at one locus is masked by a variant at another locus [13] and hence coincides with the original Bateson definition of epistasis. VanderWeele

Field Code Changed

and colleagues [87-89] derived empirical tests for compositional epistasis under models for the joint effect of two genetic factors which place no restrictions on the main effects of each factor but constrain the interactive effects of the two factors so as to be captured by a single parameter in the model. Alternatively, a likelihood method is developed to determine the "best" statistical representation of the epistatic interaction [90].

Many regression-based approaches have been discussed and applied in the context of gene-gene interactions, such as exhaustive methods (envisaging all possible interactions using full interaction models) [48, 91] or focused regression-based interaction screening approaches (thresholding combinations for interaction testing) [48, 91]. Particular regression-based methodologies include (Penalized) Logistic Regression [92-94], Multivariate Adaptive Regression Splines [95], Mnets that are able to select or drop highly correlated predictors together [96], Partial Least Squares [97], BOolean Operation-based Screening and Testing [98], or adaptive group lasso [99]. Alternatively, genotypic values are decomposed into several components including epistasis and test statistics are derived accordingly [100]. Irrespective of whether an automated model selection strategy is implemented or not, proper account should be given to the uncertainty involved in the model selection (e.g., via Bayesian model averaging – [101]).

Notably, gene association networks are an efficient method to summarize dependencies at the gene-level. In these undirected graphs, the association between two "nodes" is measured using Pearson correlation or mutual information [102], or a measure of partial correlation as in Graphical Gaussian Models (GMMs - [103]) using gene expression data.

Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed

The latter models allow making a distinction between direct associations and indirect associations due to bonds within the network. However, GGMs can lead to biased inference regarding statistical interactions, since only linear dependencies are accounted for [104]. GGM analysis has turned useful in attempts to infer gene-SNP networks from gene expression and genotyped SNP data [105] and shows some degree of overlap with so-called Reconstructability Analysis (RA, [106]), a new promising graphical modeling strategy, initially developed in the systems community, that is able to analyze epistatic interactions involving an arbitrary number of genes or SNPs, and can be combined with information theory, when deemed relevant.

A classification of the aforementioned analysis strategies is given in Table 3.

TABLE 3: about here

A note on study design

Although most of the aforementioned methods pertain to population based studies, family studies may also be useful in identifying gene–gene interactions, because affected relatives are more likely to share two nearby epistatic loci in linkage disequilibrium that would be unlinked in unrelated individuals [109]. Cordell and Clayton [110] described a unified approach to performing genetic association analysis with nuclear families (or case/control data) in a regression context. In their approach case/parent trios are analyzed via conditional logistic regression using the case and three pseudo-controls derived from the un-transmitted parental alleles. The beauty of the method is that it can be performed

Field Code Changed

using standard statistical software and that additional effects such as parent-of-origin effects can be included. Apart from the mis-specification problem in regression modeling, the major drawback is that, to date, the technique has not been adapted to include extended pedigrees without splitting them up into simple nuclear families. In addition, all aforementioned cons of working within a classical regression paradigm are taken over. In contrast, De Lobel et al [111] developed a flexible mixed modeling approach that has no problems with extended pedigrees and can easily adjust association signals for the presence of linkage. Alternatively, a multifactor dimensionality reduction method can be considered. Cattaert et al [112] developed such an approach for related individuals (who can belong to pedigrees of any size) as part of the Model-Based Multifactor Dimensionality Reduction framework introduced before. FAM-MDR (FAMily-based Multifactor Dimensionality Reduction) combines properties of GRAMMAR [113] and MB-MDR, while deriving family-free residuals from a polygenic model and submitting these as new traits to a classical MB-MDR run.

PROBLEM IDENTIFICATION AND POSSIBLE SOLUTIONS

Computation time

When genetic markers are believed to be effect modifiers of each other, and the search for epistatic effects is envisaged, it is impossible for most computer facilities to analyze the resulting phenomenal number of all possible combinations. Assuming that 5000 pair-wise combination can be analyzed in 1 second (this is comparable with PLINK epistasis testing performances [114] out of 1 million available variants, it would take over $3 \approx \frac{5*10^{11}}{5000*(60*60*24*365)}$ years to perform an exhaustive search… Graphics Processing Units

Field Code Changed

(GPUs) based implementations of epistasis screening efforts have been shown to be beneficial, especially when adopted algorithms in the screening do not rely on many interdependent operations applied to relatively small amounts of data [115]. However, the number of computations, such as those described before, is further multiplied in gene expression studies of quantitative trait mapping. Therefore, fast and high-performance computing solutions are required, that scale with the number of processors, such as the FastEpistasis algorithm for quantitative trait epistasis screening [116]. However, it is not always straightforward for researchers to adapt existing (in-house) software to allow for parallel processing. Generic tools are on the way, such as the cloud-based epistasis computing (CEO) model of Wang et al [117] to find statistically significant epistatic interactions. The advantages of Graphical Processing Units can be further accelerated in combination with Ant Colony Opimization techniques that use prior knowledge to reduce data complexity [118]. Alternatively, search space pruning can also dramatically speed up the process of epistasis detection without compromising the optimality of the results (e.g., Convex Optimization-based Epistasis detection algorithm [119] and Tree-based Epistasis Association Mapping [120]).

Multiple testing

The interpretation of epistasis screening studies involving a large number or all available polymorphic variants in the human genome is severely hampered by the statistical problem that a large number of genetic markers will be highlighted as significant signals or contributing factors, whereas in reality they are not. To correct for occurrences of false positives typically arising from performing multiple statistical tests, several multiple

testing corrections have been developed and customized to a genome-wide association context, when deemed necessary. There is not a single measure to quantify false positives [121]. In general, false-positive controlling measures either control the family-wise error rate (FWER), known as the overall type I error rate, the generalized family-wise error rate (gFWER), tail probabilities for the proportion of false positives among the rejected null hypotheses (TPPFP) and the false discovery rate (FDR). For discussions about the utility of the aforementioned multiple testing procedures in genomics applications, we refer to other publications [122-125].

Field Code Changed

In either case, it is important to verify the validity of the assumptions that underlie each of these techniques, in order to select the optimal corrective method for the data at hand. For instance, not many approaches adequately account for their dependence on the effective number of tests or dependencies between tests, while correcting for multiple-testing. Several methods have been developed to implement corrective methods for GWAs with genetic markers that are in linkage disequilibrium with each other or in the presence of correlated hypothesis tests. These methods include applying a Bonferroni correction using effective sample size derived from principal components [126], deriving more accurate estimates for the effective number of tests based on an upper bound for the overall type I error probability in the presence of highly correlated markers [127], exploiting haplotype blocking algorithms [128], developing a framework for hidden Markov Model-dependent hypothesis testing [129], and further elaborating on the latter approach, using a pooled local index of significance (PLIS) ranking strategy [130].

The FDR comes in different shapes and flavors that mainly differ in the way the number of true null hypotheses is handled (or estimated) or account is made for dependent hypotheses [131, 132]. Rather than setting a fixed FDR rate to control, Storey and colleagues [133, 134] suggest giving a q-value to each test that indicates what pFDR would result from declaring that test significant. A difficulty with FDR is that it says little about the individual tests. Even the q-values ignore that the most significant tests are most likely to be true positives. This led to the concept of False Positive Report Probability (FPRP – [135, 136], which can be shown to have similarities with the so-called local FDR [137] and the q-value of Storey et al [133]. It is less obvious how to optimally adapt these methods in the context of epistasis screening, in particular, how to best account for underlying genetic networks and hence a complex structure of correlated test statistics when testing for epistasis.

Field Code Changed

Field Code Changed

When tests are not identically distributed, for instance due to inadequate numbers of observations across all combinations of the factors studied, procedures such as FWER controlling maxT adjustments [138] may be highly unbalanced in that not all hypothesis tests will contribute to the adjustment in a comparable fashion (an observation that I also made when analyzing exome sequencing data with a combination of common and extremely rare alleles – Genetic Analysis Workshop (GAW) 17, [139]). Here, "standardized" test statistics may need to be derived prior to correction [140]. On the side, "standardizing" test statistics, i.e., making test statistics more comparable, is not a new idea to genetic analysis. It has also been adopted in GWAs main effects screening

when evidence over different genetic models is combined [141] or in meta-analysis contexts when different study designs are involved [142].

Actually, the maxT corrective method is an example of another strategy to control the number of false positives, namely by means of permutation replicates. For permutation tests (i.e., randomization tests, exact tests) the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under multiple reshuffles of the observed trait labels. An important assumption behind a permutation test is that the observations are exchangeable under the null hypothesis, in which case this procedure will provide exact significance levels. Usually an asymptotically equivalent permutation test is obtained, via Monte Carlo sampling (i.e., random sampling among all possible permutation replicates). Significance assessment can also be based on bootstrap samples that are less stringent in the adopted assumptions [143].

Field Code Changed

Permutation-based strategies are widely considered as the gold standard for accurate multiple testing corrections, but it is often computationally impractical for GWA datasets. Moreover, its validity heavily depends on whether or not the permutation distribution adequately reflects the distribution under the null hypothesis [144]. Fortunately, when limited in the number of permutations a computer environment can handle, an early stopping rule can be imposed [145], as was applied in FAM-MDR while aiming for the best higher-order multi-locus model [112]. Building upon the work of Churchill and Doerge [146] and Doerge and Churchill [147], Carlborg et al [148] developed a randomization technique to derive empirical significance thresholds when mapping

interactive quantitative trait loci. Alternatively, in a permutation-testing framework, fewer replicates are required when noting that for instance the minimum p-value, sum statistic and truncated product can all be regarded as the extreme value of a large number of observations [149]. Tail distributions of observed p-values were successfully approximated by generalized extreme value distributions for genome-wide main effects scenario's [150] and epistasis screening scenario's [151].

Field Code Changed

The field is not yet saturated with time-efficient false-positive controlling methods. New promising tools, even in the presence of millions of correlated markers, are emerging as we speak, claiming to be as accurate as permutation-based testing. One of these methods is SLIDE (a Sliding-window Monte-Carlo approach for Locally Inter-correlated markers with asymptotic Distribution Errors corrected - [152]. Another one is PACT (P values Adjusted for Correlated Tests - [153]).

Finally, adhering to a frequentist paradigm may be the most convenient approach in simple analysis settings. Because these tests, in their simplest form, may be conservative when statistical tests are not independent, may involve omnibus rather than specific null hypotheses, and may have varying interpretations with varying number of considered statistical tests, an open mind and common sense are needed in order not to miss true epistatic associations. Under the Bayesian approach, there is no penalty for analyzing data exhaustively because the prior probability of an association should not be affected by what tests the investigator chooses to carry out.

The curse of dimensionality

The curse of dimensionality has been a difficulty with Bayesian statistics as well, for which the posterior distributions often have many parameters. The problem has been circumvented by the implementation of simulation-based Bayesian inference, especially using Markov chain Monte Carlo.

In the field of neural networks (NNs), the curse of dimensionality expresses itself in several ways. For instance, as dimensionality of the input space grows, the inclusion of many relatively poor performing attributes into the resulting network needs to be avoided. This is a particular concern for unsupervised learning strategies. Also, the higher the dimensionality of the input space, the more data may be needed to separate the good from the bad input signals [154]. To this end, several adjustments have been made to classical NN approaches in the context of epistasis detection, such as the incorporation of genetic programming and grammatical evolution (e.g., [70, 155]).

Several strategies can be adopted to select the number of genetic variants to be used for epistasis screening, hereby downplaying the curse of dimensionality. Strategy I involves performing an exhaustive search, with the associated need to address several computational issues and the need to confront a severe multiple testing problem. An example of an exhaustive epistasis screening method is the earlier introduced (MB-)MDR [79].

Strategy II involves selecting genetic markers based on the statistical significance or strength of their singular main effects [156]. This approach has long been the traditional strategy to select variables from GWAs studies for further epistasis-oriented evaluations.

Field Code Changed

Field Code Changed

Field Code Changed

A weighting or evaluation of singular main effects may have been obtained via non-classical methods, such as those using prior probabilities of disease association [157] or prior belief on the plausibility of obtaining a positive finding [136]. Obviously, finding gene-gene interactions in this way is unlikely to be successful when the underlying disease model is purely epistatic [16].

Strategy III involves data mining type of (multi-)variable selection methods (cfr section "Variable Selection).

Strategy IV involves prioritizing sets of genetic markers based on available biological data base resources, such as pathway information. In the extreme, an example of this strategy is to bin markers according to their reference to genes and to perform subsequent testing of gene-gene co-association [158]. Employing Interaction-Based Gene Set Analysis strategies (IB-GSA – [159]) may be particularly powerful to achieve a biologically meaningful data reduction prior to epistasis modelling. A word of caution is in place though. When prioritization is based on aggregating information from publicly available -omics data bases [160] the caveat is generating findings which may be biased towards "what is already known".

Epistasis in the presence of linked SNPs

Linkage disequilibrium (LD) is the non-random association of alleles at different loci within a randomly mating population assuming Hardy-Weiberg equilibrium at each locus. When this form of allelic association is observed for unlinked markers, it is often referred to as gametic phase disequilibrium. Missing heritability may hide in epistasis between linked markers [161]: In traditional genetic linkage and founder haplotype

mapping studies we expect relatively long stretches of shared chromosome inherited from a relatively recent common ancestor. This is in contrast to what is to be expected in GWAs with (apparently, or assumed) unrelated individuals. Hence, whereas the genetic effect on phenotype involving multiply tightly linked loci may appear in pedigree studies as part of the additive genetic variance, it may actually appear as a gene-gene interaction in a population-based genome-wide screening [161]. This is why usually detected interaction signals between linked loci is coined as "redundant" [162].

In effect, when studying gene-gene interactions for a binary trait, it can be shown that there is complete confounding of interaction with linkage disequilibrium for linked genes and with gametic phase disequilibrium for unlinked genes [10]. Zhao et al [163] investigated generated LD patterns in the presence of gene-gene interactions between two disease-susceptibility loci in Hardy-Weinberg equilibrium and between two unlinked marker loci, each of which is in LD with either of two interacting loci. They noted that LD-based measures can serve as useful statistics to detect gene-gene interaction between two unlinked loci, a note that was further elaborated on in the EPIBLASTER software [107].

When the loci are in Linkage Equilibrium (LE), the total variance can be partitioned into two main variances and one epistatic variance [164]. In the absence of LD, the main effects model, a model for which the epistatic variance is zero and the total variance is equal to the sum of the main variances, is equivalent to the additive model, which describes additivity on the penetrance scale [165, 166]. This is no longer the case when

Field Code Changed

loci are in LD, in which case the main-effects model can be viewed as a special case of the additive model. Since in the event of epistasis the degree of deviation from these models may be significantly different, Zhang and Ji [167] suggest testing statistical epistatic effects as a departure from the main-effects model.

Rare variants

Current disease risk prediction models using results of classical main effects GWAs, relying on an abundance of common variants, are seldom useful in clinical practice. Although hundreds of "genetic signals" have been identified in association with certain complex human diseases, only a handful of causative genes have been discovered in follow-up studies [168]. This understanding of "lost signals" or "missing heritability" [109, 168, 169] paved the way for investigators to perform a quest for rare variants and to further unravel the contribution of rare variants to the multi-factorial inheritance of common diseases [170].

Field Code Changed

Interpretation of GWAs in terms of providing leads for causal variants may indeed be severely hampered when disregarding the possibility that disease may be caused by multiple strong-effects variants, each of which are found in only a few people [171]. Dickson et al [171] pointed towards the potential for so-called "synthetic associations" to SNPs that are quite distant from the (many) true causative (strong-effect) variants. Moreover, whereas it seems unlikely, a priori, that variants with small single-locus effects would give rise to significant interactions, the prospects might be much more optimistic when rare high-impact variants are involved. There is evidence for complex

diseases, such as Type 2 diabetes mellitus, to result from complex genetic interactions between a large number of rare alleles and a small number of common alleles [172]. The so-called "mosaic model" of interactions poses interesting challenges in the context of epistasis detection, given the statistical problems to detect rare variant single effects associations.

Bansal et al [173] give a nice overview of different data analysis methods that can be useful to decipher simple associations between collections of rare variants and a trait of interest. The wide variety of possible settings in which a collection of rare variants might show an association with a trait (whether or not interacting, possibly with more than one common variant) makes it even harder to recommend a single statistical analysis strategy in this context.

When dimensionality increases and higher-order interactions are targeted, an increasing number of multi-locus factor levels will only be present in a few samples or no sampled individual at all will exhibit the particular combination. Large discrepancies in numbers of observations between different combinations of multi-locus factors (such as those generated in the presence of rare variants), may technically cause a problem of confounding among the parameters of interest, and is a point of major concern. Nevertheless, continuing efforts to improve the detection of complex traits associations with rare variants due to both gene main effects and interactions, led to the kernel-based adaptive cluster (KBAC) approach [174]. This method was demonstrated to have superior power compared to other rare variant analysis methods, such as the weight sum statistic [175] and the combined multivariate and collapsing method [176].

Interpretation of results

The study of epistasis poses problems of interpretability. Statistically, epistasis is usually defined in terms of deviation from a model of additive multiple effects, but this might be on either a linear or logarithmic scale, which implies different definitions. Hence, the implication of epistasis may vary due to the choice of scale related to the trait of interest. Despite this conceptual hurdle, recent work has shown that identified epistatic effects *are* able to reveal useful information about gene function [177] and interpretation can be greatly enhanced when incorporating prior knowledge, such as those derived from pathways data bases [178], or omics data bases that offer a wealth of information on cellular processes at the level of molecular biology, biochemistry and systems biology [179].

For instance, Pattin and Moore [180] explored the role of information extraction from protein-protein interaction data bases to enhance the genome-wide analysis of epistasis in complex human diseases. Baranzini et al [181] proposed a protein interaction and network-based analysis (PINBPA) to exploit signals from main effects GWA studies that would have been ignored when strictly adhering to stringent multiple testing criteria. These types of analyses may give new leads to previously unidentified pathways and hence new leads to interactions in GWAs. Lee et al [182] used functional genetic networks or a map of biological interactions between genes to reduce to increase the power to test for the existence of gene-gene interactions throughout the genome. This approach aims to discover (predict) new epistatic interactions by adopting the principle

Field Code Changed

that genes who act in a common pathway or are involved in a common biological process may serve as modifier genes for the same mutation of interest. The authors indicated that using a network of functional interactions is more predictive than using physical networks, such as the popular protein-protein networks [183]. Lin et al [184] were able to identify a large number of human gene-gene interactions, while constructing a human genome-wide map of genetic interactions inferred from radiation hybrid (RH) data. Radiation hybrid mapping is a genetic technique that is based on a statistical method to determine the distances between DNA markers and their order on the chromosomes. The network resulting from testing pair-wise interactions by comparing co-retention frequencies with chance frequencies was shown to give substantial improvements in power to identify potential gene-gene interactions, especially when combining RH data from different species. It also provided unbiased evidence that essential genes are central to network, as both highly connected hubs and as highly trafficked bottlenecks. Despite the potential of the technique, the size of the RH network (it tends to saturation) does not allow rapid experimental validation of interactions. More work is needed to prioritize interactions for further follow-up.

Along the same lines, but specifically targeting the identification of epistasis, Bush et al [160] integrated multiple publicly available databases of gene groupings and sets of disease-related genes in their Biofilter system. It leaves no doubt that using prior biological knowledge in this sense to inform the analysis of epistasis detection is essential. But the end of the tunnel is not yet in sight. Addition of other potentially informative data bases, assessment and incorporation of “optimal” scoring systems to accumulate evidence from these data bases, possibly allowing for uncertainty involved in

the data source entries, acknowledging the complementary characteristics of each of the available data sources, and allowance for different assignment strategies from genetic variants to genes, are only some of the components of such an biology assistant-driven approach that need careful thought.

Clearly, visualization techniques can assist in interpreting analysis results [41, 162]. Not surprisingly, in the context of gene-gene interactions, one of these visualization techniques is adopted from the "clustering" community, i.e. dendrograms. A dendrogram is a tree diagram that illustrates the arrangement of clusters (here, genetic markers) produced by hierarchical clustering. Merges and splits of clusters are decided upon via a measure of dissimilarity, which may or may not be entropy-based.

Field Code Changed

FUTURE CHALLENGES FOR PERSONALIZED MEDICINE

Although GWAs, that classically exploit the common genetic variations in the human genome, have been successful for a variety of human complex traits, their success is less apparent when trying to replicate the findings or when trying to translate the findings to useful risk prediction models. One possible explanation is the not fully exploited ubiquity of epistasis. Our understanding about the role of epistasis during evolution, its biological relevance and its relation to common complex diseases, is only developing and its impact on personalized medicine yet needs to be determined. However, accounting for epistatic effects or modeling epistasis is just one corner stone of the complex human architecture that also involves important networks of gene-environment interactions, such as pharmacogenetic interactions.

It is worthwhile to further explore the potential benefits of integrating systems biology approaches or views into the field concerned with epistasis detection. In particular, more work is needed to investigate the similarities between methodologies used to model cellular systems (exploiting information about the molecular content of a system and interactions within the system) and the efforts the field of systems biology is making towards omics integration and tying all architectural components together [2, 185]. It will be challenging for some time to come though, to design customized charts of individualized risk estimates, using as much of the "complete picture" as possible. A nomogram [186] is a graphical calculating device, a two-dimensional diagram, designed to allow the approximate graphical computation of a possibly complex function. Construction methods of nomograms such as those proposed by Lee et al [187], using genetic algorithm and naïve Bayesian techniques, are promising in the light of using both clinical, genetic and pharmaco-genetic information in patients' risk-factor nomograms. However, simply developing an effective nomogram from clinical data, whether these refer to lab experiments, therapy history or disease progression is not obvious. What does it take to also account for the complexity of each individual's personal genetic blueprint?

Field Code Changed

IN CONCLUSION

Similar to main effects GWA studies, the power of a genome-wide interaction analysis depends on many parameters, such as minor allele frequencies of involved markers and disease susceptibility loci and linkage disequilibrium patterns, but also study design, genetic multi-locus effect size, test size, and last but not least, sample size. Because of the

variety of possible epistasis models and analysis tools, available epistasis genetic power calculators, such as QUANTO [188], only accommodate a fraction of the scenario's an investigator is confronted with in practice. Whether the power of an envisaged epistasis study is computed via available software, or estimated via an extensive simulation study, it leaves no doubt that with sample sizes of the order of only thousands of individuals, there is insufficient power to detect interaction effects unless the underlying epistasis model is extreme. World-wide collaborative efforts should solve this issue.

Ideally, several analysis viewpoints are taken in the search for gene-gene interactions and the performance of different analysis techniques on power and false positive control is formally investigated. Comparing methods is not always an easy task, since the comparability of many methods is complicated by the different ways in which results are reported. In general, regression-based statistical tests for interaction are of limited use in detecting "epistasis" in the sense of masking [11, 189]. Here, the concept of "compositional epistasis" may be more useful. Although there is no single best outstanding analysis strategy in the search for epistatic effects, there is a clear trend towards the development of data reduction techniques and the merging of evidence from "ensembles" of techniques. This was already observed by Musani et al [20], and is expected to be continued for some years to come. To date, it is unclear which role the availability of next generation (whole genome) sequencing data will play in the epistasis story [174]. However, whatever analysis route is taken, replication and validation of (positive) findings in additional independent studies remain essential [190]. Also, visualization techniques may further increase insights into complex epistasis patterns

Field Code Changed

Field Code Changed

[191] and may facilitate the translation of statistical findings into a tool that can improve clinical decision making and therefore patient outcome.

Last, but not least, I strongly believe that the world of interactions will expose itself in greater detail if better use is made of all available bio-data and several pieces of omics-information are glued together in a single analysis work flow.

Acknowledgements

K. Van Steen acknowledges research opportunities offered by the Belgian Network BioMAGNet (Bioinformatics and Modeling: from Genomes to Networks), funded by the Interuniversity Attraction Poles Programme (Phase VI/4), initiated by the Belgian State, Science Policy Office. Her work was also supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence (Pattern Analysis, Statistical Modelling and Computational Learning), IST-2007-216886. The scientific responsibility for this work rests with the author.

References

1. Carlborg O, Haley CS. Epistasis: too often neglected in complex trait studies?, Nature Reviews Genetics 2004;5:618-U614.
2. Joyce AR, Palsson BO. The model organism as a system: integrating 'omics' data sets, Nature Reviews Molecular Cell Biology 2006;7:198-210.
3. Sanjuan R, Elena SF. Epistasis correlates to genomic complexity, Proc Natl Acad Sci U S A 2006;103:14402-14405.
4. Sanjuan R, Nebot MR. A network model for the correlation between epistasis and genomic complexity, PLoS One 2008;3:e2663.
5. Moore JH. A global view of epistasis, Nat Genet 2005;37:13-14.
6. Emily M, Mailund T, Hein J et al. Using biological networks to search for interacting loci in genome-wide association studies, Eur J Hum Genet 2009;17:1231-1240.

7. Wu J, Devlin B, Ringquist S et al. Screen and clean: a tool for identifying interactions in genome-wide association studies, Genet Epidemiol 2010;34:275-285.
8. Bateson W, Mendel G. Mendel's Principles of Heredity. A Defence, with a Translation of Mendel's Original Papers on Hybridisation. Cambridge University Press, 1902.
9. Fisher R. The correlation between relatives on the supposition of Mendelian inheritance, Trans. R. Soc. Edinb. 1918;52:399-433.
10. Wang X, Elston RC, Zhu X. Statistical interaction in human genetics: how should we model it if we are looking for biological interaction?, Nat Rev Genet 2010;12:74.
11. Cordell HJ. Detecting gene-gene interactions that underlie human diseases, Nat Rev Genet 2009;10:392-404.
12. Wang X, Elston RC, Zhu X. The meaning of interaction, Human Heredity 2010;70:269-277.
13. Phillips PC. Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems, Nature Reviews Genetics 2008;9:855-867.
14. Gregersen JW, Kranc KR, Ke X et al. Functional epistasis on a common MHC haplotype associated with multiple sclerosis, Nature 2006;443:574-577.
15. Davierwala AP, Haynes J, Li Z et al. The synthetic genetic interaction spectrum of essential genes, Nat Genet 2005;37:1147-1152.
16. Culverhouse R, Suarez BK, Lin J et al. A perspective on epistasis: limits of models displaying no main effect, Am J Hum Genet 2002;70:461-471.
17. Li W, Reich J. A complete enumeration and classification of two-locus disease models, Hum Hered 2000;50:334-349.
18. Hallgrimsdottir IB, Yuster DS. A complete classification of epistatic two-locus models, BMC Genet 2008;9:17.
19. Onkamo P, Toivonen H. A survey of data mining methods for linkage disequilibrium mapping, Hum Genomics 2006;2:336-340.
20. Musani SK, Shriner D, Liu NJ et al. Detection of gene x gene interactions in genome-wide association studies of human population data, Human Heredity 2007;63:67-84.
21. Moore JH. Mining patterns of epistasis in human genetics. In: Chen J. Y., Lonardi S. eds). Biological data mining. New York: Chapman and Hall 2009.
22. Kilpatrick JR, Nakhleh LK. Classifying and evaluating methods for detecting multi-locus genotype-phenotype association, Bioinformatics 2010.
23. Guyon I, Gunn S, Nikravesh M et al. (2006), 'Feature extraction'.
24. Li J, Tang R, Biernacka JM et al. Identification of gene-gene interaction using principal components, BMC Proc 2009;3 Suppl 7:S78.
25. Guyon I, Elisseeff A. An introduction to variable selection and feature selection, Journal of machine learning research 2003;3:1157-1182.
26. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics, Bioinformatics 2007;23:2507-2517.
27. Dong CZ, Chu X, Wang Y et al. Exploration of gene-gene interaction effects using entropy-based methods, European Journal of Human Genetics 2008;16:229-235.
28. Varadan V, Miller DM, 3rd, Anastassiou D. Computational inference of the molecular logic for synaptic connectivity in C. elegans, Bioinformatics 2006;22:e497-506.

29. Kononenko I. Estimation attributes: analysis and extensions of RELIEF. Springer Berlin / Heidelberg, 1994.
30. Moore JH, White BC. Tuning ReliefF for Genome-Wide Genetic Analysis. 2007, 166-175.
31. Greene CS, Penrod NM, Kiralis J et al. Spatially uniform relieff (SURF) for computationally-efficient filtering of gene-gene interactions, BioData Min 2009;2:5.
32. McKinney BA, Reif DM, White BC et al. Evaporative cooling feature selection for genotypic data involving interactions, Bioinformatics 2007;23:2113-2120.
33. Nunkesser R, Bernholt T, Schwender H et al. Detecting high-order interactions of single nucleotide polymorphisms using genetic programming, Bioinformatics 2007;23:3280-3288.
34. Greene CS, White BC, Moore JH. Ant Colony Optimization for Genome-Wide Genetic Analysis, Lect Notes Comput Sci 2008;5217/2008:37-47.
35. Wang Y, Liu X, Robbins K et al. AntiEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm, BMC Res Notes 2010;3:117.
36. Schwender H, Ickstadt K. Identification of SNP interactions using logic regression, Biostatistics 2008;9:187-198.
37. Schwender H, Ruczinski I, Ickstadt K. Testing SNPs and sets of SNPs for importance in association studies, Biostatistics 2011;12:18-32.
38. Anastassiou D. Computational analysis of the synergy among multiple interacting genes, Mol Syst Biol 2007;3:83.
39. Chanda P, Sucheston L, Zhang A et al. AMBIENCE: A Novel Approach and Efficient Algorithm for Identifying Informative Genetic and Environmental Associations With Complex Phenotypes, Genetics 2008;180:1191-1210.
40. Chanda P, Sucheston L, Zhang AD et al. The interaction index, a novel information-theoretic metric for prioritizing interacting genetic variations and environmental factors, European Journal of Human Genetics 2009;17:1274-1286.
41. Chanda P, Zhang A, Brazeau D et al. Information-theoretic metrics for visualizing gene-environment interactions, American Journal of Human Genetics 2007;81:939-963.
42. Shannon CE. A Mathematical Theory of Communication, Bell Syst. Tech. J. 1948;27, .379-423, 623-656.
43. Jakulin A (2003), 'Attribute interactions in machine learning', *Faculty of Computer and Information Science*, University of Ljubljana.
44. van der Linde A, Tutz G. On association in regression: the coefficient of determination revisited., Statistics in Medicine 2008;42:1-24.
45. Kullback S. *Information Theory and Statistics*. New York Dover Publ.: Mineola, 1968.
46. Sucheston L, Chanda P, Zhang A et al. Comparison of information-theoretic to statistical methods for gene-gene interactions in the presence of genetic heterogeneity, BMC Genomics 2010;3:487.
47. Kononenko I, Robnik-Sikonia M. Relieff for estimation and discretization of attributes in classification, regression and ilp problems. Artificial Intelligence: Methodology, Systems, Applications. Amsterdam, the Netherlands: IOS Press, 1996.
48. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases, Nat Genet 2005;37:413-417.

49. Calle ML, Urrea V, Vellalta G et al. Improving strategies for detecting genetic patterns of disease susceptibility in association studies, Statistics in Medicine 2008;27:632-6546.
50. Moore JH. Genome-wide analysis of epistasis using multifactor dimensionality reduction: feature selection and construction in the domain of human genetics. In: Zhu X., Davidson I. eds). Knowledge discovery and data mining: challenges and realities with real world data. 2007, 17-30.
51. De Lobel L, Geurts P, Baele G et al. A screening methodology based on Random Forests to improve the detection of gene-gene interactions, European Journal of Human Genetics 2010;18:1127-1132.
52. Schwarz DF, Konig IR, Ziegler A. On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data, Bioinformatics 2010;26:1752-1758.
53. Yang C, He Z, Wan X et al. SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies, Bioinformatics 2009;25:504-511.
54. Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies, Nat Genet 2007;39:1167-1173.
55. Dorigo M, Stützle T. Ant Colony Optimization. MIT Press, 2004.
56. Tyron RC. Cluster analysis. New-York: McGraw-Hill, 1939.
57. Breiman L, Friedman JH, Olshen RA et al. Classification and Regression Trees. Taylor & Francis, Inc., 1984.
58. Ripley BD. Pattern recognition and neural networks. Cambridge University Press, 1996.
59. Mechanic LE, Luke BT, Goodman JE et al. Polymorphism Interaction Analysis (PIA): a method for investigating complex gene-gene interactions, Bmc Bioinformatics 2008;9:146.
60. Ritchie MD, Hahn LW, Roodi N et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer, American Journal of Human Genetics 2001;69:138-147.
61. Nelson MR, Kardia SL, Sing CF. The combinatorial partitioning method, Combinatorial Pattern Matching 2000;1848:293-304.
62. Culverhouse R. The use of the restricted partition method with case-control data, Human Heredity 2007;63:93-100.
63. Culverhouse R, Jin W, Jin CH et al. Power and false-positive rates for the restricted partition method (RPM) in a large candidate gene data set, BMC Proc 2009;3 Suppl 7:S74.
64. Sun X, Zhang Z, Zhang Y et al. Multi-locus penetrance variance analysis method for association study in complex diseases, Hum Hered 2005;60:143-149.
65. Kooperberg C, Ruczinski I. Identifying interacting SNPs using Monte Carlo logic regression, Genetic Epidemiology 2005;28:157-170.
66. Ruczinski I, Kooperberg C, LeBlanc ML. Logic Regression, Journal of Computational and Graphical Statistics 2003;12:475-511.
67. Zheng T, Wang H, Lo SH. Backward genotype-trait association (BGTA)-based dissection of complex traits in case-control designs, Human Heredity 2006;62:196-212.

68. Yang P, Ho JW, Zomaya AY et al. A genetic ensemble approach for gene-gene interaction identification, Bmc Bioinformatics 2010;11:524.
69. Wolf BJ, Hill EG, Slate EH. Logic Forest: an ensemble classifier for discovering logical combinations of binary markers, Bioinformatics 2010;26:2183-2189.
70. Turner SD, Dudek SM, Ritchie MD. Grammatical evolution of neural networks for discovering epistasis among quantitative trait loci. 2010, 86-97.
71. Bureau A, Dupuis J, Falls K et al. Identifying SNPs predictive of phenotype using random forests, Genet Epidemiol 2005;28:171-182.
72. Jiang R, Tang WW, Wu XB et al. A random forest approach to the detection of epistatic interactions in case-control studies, Bmc Bioinformatics 2009;10:(Suppl 1) S65.
73. Lunetta KL, Hayward LB, Segal J et al. Screening large-scale association study data: exploiting interactions using random forests, BMC Genet 2004;5:32.
74. Prinzie A, Van den Poel D. Random forests for multiclass classification: Random MultiNomial Logit, Expert Systems with Applications 2008;34:1721-1732.
75. Prinzie A, Van den Poel D. Random Multiclass Classification: Generalizing RF to RMNL and RNB. 2007, 349-358.
76. Karpievitch YV, Hill EG, Leclerc AP et al. An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++, PLoS One 2009;4:e7087.
77. Sela RJ, Simonoff JS (2009), 'Re-Em Trees: A New Data Mining Approach for Longitudinal Data ', *Statistics Working Papers Series*.
78. Calle ML, Urrea V, Malats N et al. mbmdr: an R package for exploring gene-gene interactions associated with binary or quantitative traits, Bioinformatics 2010;26:2198-2199

79. Calle ML, Urrea V, Vellalta G et al. (2008), 'Model-Based Multifactor Dimensionality Reduction for detecting interactions in high-dimensional genomic data.', Department of Systems Biology, Universitat de Vic.
80. Kim Y, Wojciechowski R, Sung H et al. Evaluation of random forests performance for genome-wide association studies in the presence of interaction effects, BMC Proc 2009;3 Suppl 7:S64.
81. Wongseree W, Assawamakin A, Piroonratana T et al. Detecting purely epistatic multi-locus interactions by an omnibus permutation test on ensembles of two-locus analyses, Bmc Bioinformatics 2009;10.
82. Mahachie John JM, Cattaert T, Van Lishout F et al. Model-Based Multifactor Dimensionality Reduction to detect epistasis for quantitative traits in the presence of error-free and noisy data, Eur J Hum Genet 2010.
83. Carlborg O, Andersson L, Kinghorn B. The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci, Genetics 2000;155:2003-2010.
84. Bellman R, Kalaba R. A Mathematical Theory of Adaptive Control Processes, Proc Natl Acad Sci U S A 1959;45:1288-1290.
85. Chapman J, Clayton D. Detecting association using epistatic information, Genet Epidemiol 2007;31:894-909.

86. Chatterjee N, Kalaylioglu Z, Moslehi R et al. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions, Am J Hum Genet 2006;79:1002-1016.
87. VanderWeele TJ. Empirical tests for compositional epistasis, Nat Rev Genet 2010;11:166.
88. VanderWeele TJ. Epistatic interactions, Stat Appl Genet Mol Biol 2010;9:Article 1.
89. VanderWeele TJ, Laird NM. Tests for Compositional Epistasis under Single Interaction-Parameter Models, Ann Hum Genet 2010;75:146-156.
90. Gao H, Granka JM, Feldman MW. On the classification of epistatic interactions, Genetics 2010;184:827-837.
91. Ionita I, Man M. Optimal two-stage strategy for detecting interacting genes in complex diseases, BMC Genet 2006;7:39.
92. North BV, Curtis D, Sham PC. Application of logistic regression to case-control association studies involving two causative loci, Human Heredity 2005;59:79-87.
93. Park M, Hastie T. Penalized logistic regression for detecting gene-gene interactions, Biostatistics 2007;9:30-50.
94. Tanck MW, Jukema JW, Zwinderman AH. Simultaneous estimation of gene-gene and gene-environment interactions for numerous loci using double penalized log-likelihood, Genet Epidemiol 2006;30:645-651.
95. Lin HY, Wang W, Liu YH et al. Comparison of multivariate adaptive regression splines and logistic regression in detecting SNP-SNP interactions and their application in prostate cancer, J Hum Genet 2008;53:802-811.
96. Huang J, Breheny P, Ma S et al. (2010), 'The Mnet method for variable selection', Department of Statistics and Actuarial Science, The University of Iowa.
97. Wang T, Ho G, Ye K et al. A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped, Genet Epidemiol 2009;33:6-15.
98. Wan X, Yang C, Yang Q et al. BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies, Am J Hum Genet 2010;87:325-340.
99. Yang C, Wan X, Yang Q et al. Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group Lasso, Bmc Bioinformatics 2010;11 Suppl 1:S18.
100. Wang Z, Liu T, Lin Z et al. A general model for multilocus epistatic interactions in case-control studies, PLoS One 2010;5:e11384.
101. Viallefont V, Raftery AE, Richardson S. Variable selection and Bayesian model averaging in case-control studies, Stat Med 2001;20:3215-3230.
102. Butte AS, Kohane IS. Relevance networks: a first step towards finding genetic regulatory networks within microarray data. In: Parmigiani G., Garrett E. S., Irizarry R. A. et al. eds). The analysis of gene expression data. New-York: Springer, 2003, 428-446.
103. Schafer J, Strimmer K. An empirical Bayes approach to inferring large-scale gene association networks, Bioinformatics 2005;21:754-764.
104. Andrei A, Kendziorski C. An efficient method for identifying statistical interactors in gene association networks, Biostatistics 2009;10:706-718.

105. Chu JH, Weiss ST, Carey VJ et al. A graphical model approach for inferring large-scale networks integrating gene expression and genetic polymorphism, BMC Syst Biol 2009;3:55.
106. Zwick M. Reconstructability analysis of epistasis, Ann Hum Genet;75:157-171.
107. Kam-Thong T, Czamara D, Tsuda K et al. EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units, Eur J Hum Genet.
108. Ruczinski I, Kooperberg C, LeBlanc ML (2002), 'Logic Regression - methods and software.', in Denison D., Hansen M., Holmes C. et al. Eds., *MSRI workshop on Nonlinear Estimation and Classification*, Springer New York, pp. 333-344.
109. Manolio TA, Collins FS, Cox NJ et al. Finding the missing heritability of complex diseases, Nature 2009;461:747-753.
110. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans, Hum Mol Genet 2002;11:2463-2468.
111. De Lobel L, De Meyer H, Thijs L et al. A Family-Based Association Test to Detect Gene-Gene Interactions in the Presence of Linkage, Genetic Epidemiology 2009;33:771-772.
112. Cattaert T, Urrea V, Naj AC et al. FAM-MDR: a flexible family-based multifactor dimensionality reduction technique to detect epistasis using related individuals, PLoS One 2010;5:e10304.
113. Aulchenko YS, de Koning DJ, Haley C. Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis, Genetics 2007;177:577-585.
114. Purcell S, Neale B, Todd-Brown K et al. PLINK: A tool set for whole-genome association and population-based linkage analyses, American Journal of Human Genetics 2007;81:559-575.
115. Sinnott-Armstrong NA, Greene CS, Cancare F et al. Accelerating epistasis analysis in human genetics with consumer graphics hardware, BMC Res Notes 2009;2:149.
116. Schupbach T, Xenarios I, Bergmann S et al. FastEpistasis: a high performance computing solution for quantitative trait epistasis, Bioinformatics 2010;26:1468-1469.
117. Wang Z, Wang Y, Tan K-L et al. (2010), 'CEO: a cloud epistasis computing model in GWAS', *International Conference on Bioinformatics & Biomedicine*, Hong Kong.
118. Sinnott-Armstrong NA, Greene CS, Moore JH (2010), 'Fast genome-wide epistasis analysis using ant colony optimization for multifactor dimensionality reduction analysis on graphics processing units', *GECCO*, pp. 215-216.
119. Zhang XA, Pan F, Xie YY et al. COE: A General Approach for Efficient Genome-Wide Two-Locus Epistasis Test in Disease Association Study, Journal of Computational Biology 2010;17:401-415.
120. Zhang X, Huang SP, Zou F et al. TEAM: efficient two-locus epistasis tests in human genome-wide association study, Bioinformatics 2010;26:i217-i227.
121. Hochberg Y, Tamhane AC. Multiple comparison procedures. Wiley, 1987.
122. Dudbridge F, Gusnanto A, Koeleman BP. Detecting multiple associations in genome-wide studies, Hum Genomics 2006;2:310-317.
123. Dudoit S, Van der Laan MJ. Multiple testing procedures with applications to genomics. 2008.

124. Manly KF, Nettleton D, Hwang JT. Genomics, prior probability, and statistical tests of multiple hypotheses, Genome Res 2004;14:997-1001.
125. Pollard KS, van der Laan MJ. Choice of a null distribution in resampling-based multiple testing, Journal of Statistical Planning and Inference 2004;125:85-100.
126. Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other, American Journal of Human Genetics 2004;74:765-769.
127. Moskvina V, Schmidt KM. On multiple-testing correction in genome-wide association studies, Genetic Epidemiology 2008;32:567-573.
128. Nicodemus KK, Liu WL, Chase GA et al. Comparison of type I error for multiple test corrections in large single-nucleotide polymorphism studies using principal components versus haplotype blocking algorithms, Bmc Genetics 2005;6:-.
129. Sun W, Cai T. Large-scale multiple testing under dependence. , J. R. Stat. Soc. B 2009;71:393-342.
130. Wei Z, Sun W, Wang K et al. Multiple testing in genome-wide association studies via hidden Markov models, Bioinformatics 2009;25:2802-2808.
131. Kall L, Storey JD, Noble WS. <sc>qvality</sc>: non-parametric estimation of q-values and posterior error probabilities, Bioinformatics 2009;25:964-966.
132. Strimmer K. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates, Bioinformatics 2008;24:1461-1462.
133. Storey JD. A direct approach to false discovery rates, J. R. Stat. Soc. B 2002;64:479-498.
134. Storey JD. The positive false discovery rate: A Bayesian interpretation and the q-value., Annals of Statistics 2003;31:2013-2035.
135. Wacholder S, Chanock S, Garcia-Closas M et al. Re: Assessing the probability that a positive report is false: An approach for molecular epidemiology studies - Response, Journal of the National Cancer Institute 2004;96:1722-1723.
136. Wacholder S, Chanock S, Garcia-Closas M et al. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies, Journal of the National Cancer Institute 2004;96:434-442.
137. Efron B, Tibshirani R, Storey JD et al. Empirical Bayes analysis of a microarray experiment, Journal of the American Statistical Association 2001;96:1151-1160.
138. Westfall PH, Young SS. Resampling-based multiple testing: examples and methods for p-value adjustment. New-York: Wiley, 1993.
139. Mahachie John JM, Cattaert T, De Lobel L et al. Comparison of genetic association strategies in the presence of rare alleles 2010.
140. Nacu S, Critchley-Thorne R, Lee P et al. Gene expression network analysis and applications to immunology, Bioinformatics 2007;23:850-858.
141. Hothorn LA, Hothorn T. Order-restricted Scores Test for the Evaluation of Population-based Case-control Studies when the Genetic Model is Unknown, Biometrical Journal 2009;51:659-669.
142. de Bakker P, Ferreira M, Jia X et al. Practical aspects of imputation-driven meta-analysis of genome-wide association studies**.**, Human Molecular Genetics 2008;17:R122.
143. Good P. Introduction to Statistics Through Resampling Methods and R/S-PLUS.: Wiley, 2005.

144. Dering C, Pugh E, Ziegler A. Statistical analysis of rare sequence variants: An overview of collapsing methods.
145. Nettleton D, Doerge RW. Accounting for variability in the use of permutation testing to detect quantitative trait loci, Biometrics 2000;56:52-58.
146. Churchill GA, Doerge RW. Empirical threshold values for quantitative trait mapping, Genetics 1994;138:963-971.
147. Doerge RW, Churchill GA. Permutation tests for multiple loci affecting a quantitative character, Genetics 1996;142:285-294.
148. Carlborg R, Andersson L. Use of randomization testing to detect multiple epistatic QTLs, Genet. Res., Camb. 2002;79:175-184.
149. Dudbridge F, Koeleman BP. Efficient computation of significance levels for multiple associations in large studies of correlated data, including genomewide association studies, Am J Hum Genet 2004;75:424-435.
150. Knijnenburg TA, Wessels LF, Reinders MJ et al. Fewer permutations, more accurate P-values, Bioinformatics 2009;25:i161-168.
151. Pattin KA, White BC, Barney N et al. A Computationally Efficient Hypothesis Testing Method for Epistasis Analysis Using Multifactor Dimensionality Reduction, Genetic Epidemiology 2009;33:87-94.
152. Han B, Kang HM, Eskin E. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers, PLoS Genet 2009;5:e1000456.
153. Conneely KN, Boehnke M. So Many Correlated Tests, So Little Time! Rapid Adjustment of P Values for Multiple Correlated Tests, Am J Hum Genet 2007;81.
154. Yao X (1999), 'Evolving artificial neural networks', *IEEE*, pp. 1423-1447.
155. Turner SD, Dudek SM, Ritchie MD. ATHENA: A knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait Loci, BioData Min 2010;3:5.
156. Kooperberg C, LeBlanc M. Increasing the power of identifying gene x gene interactions in genome-wide association studies, Genetic Epidemiology 2008;32:255-263.
157. Pe'er I, de Bakker PIW, Maller J et al. Evaluating and improving power in whole-genome association studies using fixed marker sets, Nature Genetics 2006;38:663-667.
158. Peng QQ, Zhao JH, Xue FZ. A gene-based method for detecting gene-gene co-association in a case-control association study, European Journal of Human Genetics 2010;18:582-587.
159. Zhang J, Li J, Deng HW. Identifying gene interaction enrichment for gene expression data, PLoS One 2009;4:e8064.
160. Bush WS, Dudek SM, Ritchie MD. Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies, Pac Symp Biocomput 2009:368-379.
161. Haig D. Does heritability hide in epistasis between linked SNPs?, Eur J Hum Genet 2010.
162. Moore JH, Gilbert JC, Tsai CT et al. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility, J Theor Biol 2006;241:252-261.
163. Zhao JY, Jin L, Xiong MM. Test for interaction between two unlinked loci, American Journal of Human Genetics 2006;79:831-845.

164. Cockerham CC. An Extension of the Concept of Partitioning Hereditary Variance for Analysis of Covariances among Relatives When Epistasis Is Present, Genetics 1954;39:859-882.
165. Cordell HJ, Todd JA, Bennett ST et al. Two-locus maximum lod score analysis of a multifactorial trait: joint consideration of IDDM2 and IDDM4 with IDDM1 in type 1 diabetes, Am J Hum Genet 1995;57:920-934.
166. Farrall M. Reports of the death of the epistasis model are greatly exaggerated, Am J Hum Genet 2003;73:1467-1468; author reply 1471-1463.
167. Zhang YL, Ji L. Main-Effects Model Is a Special Kind of Additive Model in the Presence of Linkage Disequilibrium, Human Heredity 2009;67:13-25.
168. Robinson R. Common disease, multiple rare (and distant) variants, PLoS Biol 2010;8:e1000293.
169. Eichler EE, Flint J, Gibson G et al. Missing heritability and strategies for finding the underlying causes of complex disease, Nat Rev Genet 2010;11:446-450.
170. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases, Nat Genet 2008;40:695-701.
171. Dickson SP, Wang K, Krantz I et al. Rare variants create synthetic genome-wide associations, PLoS Biol 2010;8:e1000294.
172. Sharma A, Chavali S, Mahajan A et al. Genetic association, post-translational modification, and protein-protein interactions in type 2 diabetes mellitus, Molecular & Cellular Proteomics 2005;4:1029-1037.
173. Bansal V, Libiger O, Torkamani A et al. Statistical analysis strategies for association studies involving rare variants, Nat Rev Genet;11:773-785.
174. Liu DJ, Leal SM. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions, PLoS Genet 2010;6:e1001156.
175. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic, PLoS Genet 2009;5:e1000384.
176. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data, Am J Hum Genet 2008;83:311-321.
177. Franke L, Jansen RC. eQTL analysis in humans, Methods Mol Biol 2009;573:311-328.
178. Ritchie MD. Using prior knowledge and genome-wide association to identify pathways involved in multiple sclerosis, Genome Med. 2009;1:65.
179. Thomas DC. The need for a systematic approach to complex pathways in molecular epidemiology, Cancer Epidemiol Biomarkers Prev 2005;14:557-559.
180. Pattin KA, Moore JH. Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases, Human Genetics 2008;124:19-29.
181. Baranzini SE, Galwey NW, Wang J et al. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis, Human Molecular Genetics 2009;18:2078-2090.
182. Lee I, Lehner B, Vavouri T et al. Predicting genetic modifier loci using functional gene networks, Genome Res 2010;20:1143-1153.
183. Kelley R, Ideker T. Systematic interpretation of genetic interactions using protein networks, Nature Biotechnology 2005;23:561-566.

184. Lin A, Wang RT, Ahn S et al. A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes, Genome Res 2010;20:1122-1132.
185. Moore JH, Williams SM. Epistasis and its implications for personal genetics, Am J Hum Genet 2009;85:309-320.
186. d'Ocagne M (1902), 'Sur les divers modes d'application de la méthode graphique à l'art du calcul. Calcul graphique et calcul nomographique', in Duporcq E. Ed., *deuxième Congrès international des mathématiciens*, Paris, pp. 419-424.
187. Lee KM, Kim WJ, Ryu KH et al. (2009), 'A nomogram construction method using genetic algorithm and naive Bayesian technique', *11th WSEAS international conference on Mathematical and computational methods in science and engineering*.
188. Gauderman WJ. Sample size requirements for association studies of gene-gene interaction, Am J Epidemiol 2002;155:478-484.
189. Vermeulen SH, Den Heijer M, Sham P et al. Application of multi-locus analytical methods to identify interacting loci in case-control studies, Ann Hum Genet 2007;71:689-700.
190. Milne RL, Fagerholm R, Nevanlinna H et al. The importance of replication in gene-gene interaction studies: multifactor dimensionality reduction applied to a two-stage breast cancer case-control study, Carcinogenesis 2008;29:1215-1218.
191. Böhringer S, Hardt C, Miterski B et al. Multilocus statistics to uncover epistasis and heterogeneity in complex diseases: revisiting a set of multiple sclerosis data, Eur J Hum Genet 2003;11:573-584.

Table 1: Variable input reduction methods

Type	Example	Note
Variable selection [25] [26]		Selects optimal subsets of variables to improve model "performance". Usually the original presentation of the input variables is maintained. Distinct from reducing dimensionality by aggregation or projection, for which the original presentation of the input variables is often lost (Principal Components Analysis [24])
Filter method	Entropy-based [27], synergy-based [28]	While historically used in variable selection approaches, when combined with for instance permutation or bootstrapping strategies, these methods also serves as stand-alone analysis
	ReliefF [29], TuRF [30], Spatially Uniform ReliefF [31]	Foundation: the closest instance of the same class (nearest hit) and the closest instance of a different class (nearest miss) are selected, through a type of nearest neighbor algorithm
	Evaporative Cooling [32]	ReliefF combined with entropy
Wrapper method	Genetic Programming for Association Studies (GPAS – [33])	
	Ant Colonization Optimization [34], AntEpiSeeker [35]	Transformating the optimization problem into the problem of finding the best path on a weighted graph
Embedded method	Decision-tree based methods: Recursive Partitioning, Random Forests and Logic Regression [36, 37]	

Table 2: Implicit testing of epistasis

Example	Note
Random Forests algorithms [71-73] [52] and generalizations such as Random Multinomial Logit [74], Random Naïve Bayes [75], or adaptations to cluster-correlated data [76, 77], Logic Forests [69]	Decision-tree based methods
EpiForest [72]	Combines a Random Forests analysis with a sliding window sequential forward feature selection (SWSFS) algorithm
Symbolic Discriminant Analysis, Support Vector Machines, Mining Association rules and Neural Networks	Pattern recognition methods [58]
Combinatorial Partitioning Method [61], the Restricted Partitioning Method [62, 63], Genetic Ensemble algorithmic epistasis search (GE – [68]), Bayesian Epistasis Association Mapping (BEAM – [54])	Combinatorial / partitioning methods
Principal or Independent Components, Multidimensional Scaling, Detection of Informative Combined Effects (DICE), Polymorphism Interaction Analysis (PAI, [59]), Multifactor Dimensionality Reduction (MDR, [60]), Model-Based Multifactor Dimensionality Reduction (MB-MDR – [49, 78, 79])	Multidimensional reduction methods
(MCMC) Logic Regression [65, 66]	Regression-based methods
Multi-locus Penetrance Variance Analysis [64], Backward Genotype-Trait Association [67], and Grammatical Evolution Neural Networks (GENN – [70])	Other

Table 3: Epistasis detection methods

Type	Example	Note
Exhaustive Epistasis Analysis Methods		All possible interactions of the input variables When necessary, combined with variable reduction step, which may (cfr. variable selection) or may not involve the phenotype of interest
	Multifactor Dimensionality Reduction (MDR – [60])	Non-parametric data mining method that aggregates multi-locus signals into "risk" groups
	Model-Based Multifactor Dimensionality Reduction (MB-MDR – [49])	Semi-parametric data mining method that aggregates multi-locus signals and orders them according to "severity"
	(Penalized) Logistic Regression [92-94], Multivariate Adaptive Regression Splines [95], adaptive group lasso [99], Mnets [96], Partial Least Squares [97], BOolean Operation-based Screening and Testing [98], Interaction Testing Framework (ITF) [48] Compositional epistasis [87-89], Reconstructability Analysis (RA – [106])	Parametric approach with regression-based foundation or overlap
	EPIBLASTER [107]	Contrasting measure of LD between markers
Non-exhaustive Epistasis Analysis Methods		Partial search among all possible interactions of the input variables
Greedy viewpoint		Pre-select candidate interactions based on evidence for lower-order effects
	Focused regression-based interaction screening approaches (thresholding combinations for interaction testing: Focused Interaction Testing Framework (FITF – [48])	
	Variable selection (filtering) followed-up by an exhaustive epistasis screening method	
Stochastic viewpoint		Iteratively pre-select a subgroup of variables for full-blown epistasis analysis
	SNPHarvester [53]	Interaction detection method merging ideas from k-means clustering and Markov Chain Monte Carlo
	Logic Regression (LR) [36, 66, 108], MCMC Logic Regression [65], Logic Forest [69], Random Forests + MDR [51], Random Jungle (RJ – [52])	Decision-tree based methods
	Bayesian Epistasis Association Mapping (BEAM – [54])	Bayesian partitioning with posterior probabilities for epistatic markers