BioData Mining

REVIEW Open Access

Genetic variants and their interactions in disease risk prediction – machine learning and network perspectives

Sebastian Okser[1,2], Tapio Pahikkala[1,2] and Tero Aittokallio[2,3]*

* Correspondence:
tero.aittokallio@helsinki.fi
[2]Turku Centre for Computer Science (TUCS), Turku, Finland
[3]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland
Full list of author information is available at the end of the article

Abstract

A central challenge in systems biology and medical genetics is to understand how interactions among genetic loci contribute to complex phenotypic traits and human diseases. While most studies have so far relied on statistical modeling and association testing procedures, machine learning and predictive modeling approaches are increasingly being applied to mining genotype-phenotype relationships, also among those associations that do not necessarily meet statistical significance at the level of individual variants, yet still contributing to the combined predictive power at the level of variant panels. Network-based analysis of genetic variants and their interaction partners is another emerging trend by which to explore how sub-network level features contribute to complex disease processes and related phenotypes. In this review, we describe the basic concepts and algorithms behind machine learning-based genetic feature selection approaches, their potential benefits and limitations in genome-wide setting, and how physical or genetic interaction networks could be used as *a priori* information for providing improved predictive power and mechanistic insights into the disease networks. These developments are geared toward explaining a part of the missing heritability, and when combined with individual genomic profiling, such systems medicine approaches may also provide a principled means for tailoring personalized treatment strategies in the future.

Introduction

Most disease phenotypes are genetically complex, with contributions from combinations of genetic variation in different loci. A major challenge of medical genetics is to determine a set of genetic markers, which when combined together with conventional risk factors could be used in predicting an individual's susceptibility to developing various complex disorders. The recent advances and wide availability of genetic technologies, such as those based on genome-wide association (GWA) and next-generation sequencing (NGS), have allowed for the in-depth analysis of the variation contained in the human genome. In particular, these technologies are enabling the investigation of the genetic architecture of complex diseases, with the aim of constructing more accurate disease risk prediction models that would eventually facilitate effective approaches to personalized prevention and treatment alternatives for many diseases [1,2]. While GWA studies have successfully identified hundreds of genetic variants that are associated with complex human diseases and other traits [3-6], most variants identified so far using mainly statistical association testing approaches only capture a small portion of the heritability and even an aggregate of these effects is often not predictive enough for clinical utility, leaving open the question of what may explain the remaining or ***'missing heritability'*** [7]. Suggested explanations include, for instance, contributions from rare and structural variants, genotype–environment and gene-gene interactions and sample stratification, or simply that complex traits truly are affected by thousands of variants of small effect size [8,9]. The relative contributions of these and other factors remain poorly understood, which is hindering the development of improved models for disease risk assessment.

Given the multi-factorial nature of complex diseases, many authors have reiterated the concept of interactions among genetic loci, so-called ***epistatic interactions***, as one of the major factors contributing to the missing heritability [9,10]. Epistatic genetic interactions between or within genes are thought to be profoundly important in the development of many complex diseases, but these interactions are often beyond the reach of the conventional single-variant association testing procedures [11-14]. There exist also increasingly complex interactions between genetic variants and environmental factors that may contribute to the disease risk on an individualized basis. Consequently, it has been argued that we should move away from the traditional 'one variant at a time' approach toward a more holistic, network-centric approaches, which take into account the complexity of the genotype-phenotype relationships characterized by multiple gene-gene and gene-environment interactions [15,16]. Although the conventional statistical significance testing procedures have successfully identified several susceptibility loci, it has become clear that many of the true disease associations may be much lower down on the ranked list of hits, compared to the top hits with the most statistical support [4,17,18]. Ignoring the potential risk variants in this '*gray zone*' of genetic information is likely to result in models that are missing an important proportion of the quantitative variation in heritability. Therefore, it may be that most of the heritability is hidden rather than missing, but has not previously been detected because the individual effects are too small to pass the stringent significance filters used in many studies, yet still having significant contribution to the predictive power at the level of variant or subject subsets, or when combined with non-genetic risk factors.

Here, we discuss how computational machine learning approaches can utilize hidden interactions among panels of the genetic and other risk factors, predictive of the individual disease risk by means of implementing genetic feature selection procedures and network-guided predictive models. In contrast to the conventional population-level association testing, which often detect only a few variants with statistical support beyond the genome-wide significance level (e.g. $p < 10^{-8}$), machine learning algorithms place special emphasis on maximizing the predictive accuracy at the level of individual subjects. The goal of feature selection is to identify such a panel of genetic and other risk factors, which result in a model that optimally predicts the phenotypic response variables, either the class labels in case-control classification (e.g. disease *vs.* healthy), or quantitative phenotypes in regression problems (e.g. height prediction). While epistatic genetic interactions may easily end up being averaged out in statistical association models, machine learning-based predictive modeling can also take into account those individual effects that are dependent on interactions with other variants or environmental exposures, making these models convenient for developing predictive strategies for multi-factorial diseases. Indeed, it has been shown that single-locus *p*-value-based selection strategies for constructing prediction models may lead to sub-optimal prediction accuracies [17]. In another example, hundreds of genetic markers, many of which did not originally meet the genome-wide level of statistical significance, were combined into a predictive model of type 1 diabetes risk [18]. Even though diabetes is known to involve many biological pathways, the large number of variants required may partly be attributed also to the selection of variants based solely on their individual *p*-values, which does not take into account any gene-gene interactions.

While machine learning-based computational approaches may provide a convenient framework for making use of the whole spectrum of genetic information when predicting an individual's risk of developing a disease, these developments are still in their very early stages. Implementation of highly scalable computational algorithms for genetic feature selection is a key for making these frameworks effective enough for mining data from current GWA studies, in which more than a million genetic variants are assayed in thousands of individuals, not to mention the emerging data from NGS studies, such as the 1000 Genomes project [19]. Recent improvements in constructing accurate and scalable machine learning-based predictive models will be discussed in Section 2. Another pressing problem inherent in every machine learning application is the challenge of how to evaluate the predictive capability of the constructed models, in order to avoid stating over-optimistic prediction results [20]. Model validation approaches are described in Section 3. One approach to reducing the massive search spaces and computational complexities is to use additional biological information in the model construction process. There are already several successful examples of how to make use of physical protein interaction networks when mining data from GWA studies in the search of, for instance, regulatory models [16], epistatic interactions [21], or disease genes [22]. In Section 4, we take the next step of network level analysis of genetic variants and review recent data mining solutions capable of systematically utilizing functional information from the interaction networks as *a priori* information when building disease prediction models. Finally, in Section 5, we will list some current challenges and possibilities as future directions toward improved understanding of individual predisposition to genetically complex diseases such as cancers.

Selection of genetic risk factors for machine learning-based prediction models

Rather surprisingly, the use of machine learning method in the context of genome-wide data on genetic variants has yielded a relatively limited number of studies until the very recent years (for a systematic literature review, see [20]), compared to the large number of machine learning studies on other types of genomic datasets, especially genome-wide gene expression profiles. Further, the combination of predictive modeling and advanced feature selection algorithms have been implemented in an even more restricted set of studies, even though these have generally yielded quite positive results [15,23]. Indeed, many studies have demonstrated that the use of feature selection approaches are capable of improving the prediction results beyond that when the same model is implemented on features selected solely through prior knowledge of the disease or on those genetic variants which reach genome-wide statistical significance

[18,23-25]. However, it is relatively challenging to extract the predictive signal from the high-dimensional datasets originating from GWA or NGS studies, due to a number of experimental and computational issues, many of which are different from those faced when using data from microarray gene expression profiling. Further, in order to construct accurate and reliable predictive models of complex phenotypes based on genome-wide profiles of genetic variants, it is essential to have an understanding of how to identify predictive features both individually and in groups of variant subsets, and how different feature selection approaches can deal with issues such as epistatic interactions and high-dimensional datasets [15]. Feature selection methods in machine learning can broadly be divided into filter, wrapper and embedded methods. This categorization is not strict, and each of the approaches has its own advantages and disadvantages which are, in turn, very problem dependent. Next, we briefly describe each feature selection category and consider some representative examples of each.

Filter methods

Filter methods for genetic feature selection are the most common in GWA studies due to the simplicity of their implementation, low computational complexity, and the human interpretability of the results. In their simplest form, filter methods calculate a univariate test statistic separately for each genetic feature, and the features are then ranked based on the observed statistic values. The highest ranked features form the final set of selected features, on which a predictive model may be subsequently trained. The number of features to be selected is either decided in advance or determined by a pre-defined significance threshold for the test statistic. Several well-known statistical tests have been used in GWA studies, including the Fisher's exact test and Armitage trend test [26-28], and an increasing number of statistical approaches are being developed for rare variants and the NGS data [29-31]. Since this feature selection approach requires only a single pass through the whole data, single-locus filters can be straightforwardly applied to even the largest genome sequencing datasets. Along with the multiple testing problem, the primary drawback of the single-locus filter methods is that they do not take account of the interactions between the features, which may lead to selection of both false positives, such as redundant loci, and false negatives due to epistasis interactions between or within loci [12,13,15]. More advanced filter methods can also select specific risk variant combinations that are associated with a disease risk. For instance, multifactor dimensionality reduction (MDR) is a non-parametric method that can detect statistically significant genetic interactions among two or more loci in the absence of any marginal effects, even in relatively small sample sizes [32]. While proved to be useful for association testing, however, it has been argued that the statistics being used to identify variants or their combinations, typically *p*-values for disease risk association, are perhaps not the most appropriate means for evaluating the predictive or clinical value of the genetic profiles [33].

Wrapper methods

Wrapper methods consist of three components: a search algorithm for systematically traversing through the space of all possible feature subsets, a scoring function for evaluating the predictive accuracy of the feature subsets, and the learning algorithm around which the feature selection procedure is wrapped [34]. Since the size of the power set of the features grows exponentially with the number of genetic variants screened (say n), testing all the feature subsets (2^n) is computationally infeasible (n is on the order of a million in a typical GWA study and much larger in NGS studies). Therefore, one must resort in practice to local search methods that do not guarantee finding the optimal subset but, nevertheless, usually lead to good local optima. For example, the greedy forward selection adds one feature at a time to the set of selected features after checking which of the remaining features would improve the value of the scoring function the most. Thus, the whole data set is traversed through once for each selected feature. To avoid getting trapped in poor local optima during the search in the complex and high-dimensional genetic landscapes, modified local search strategies can be utilized, including the backtracking option or several variations of evolutionary algorithms. The most popular scoring functions used with wrapper methods are the prediction error on the training set, a separate validation set, or cross-validation error. The feature selection can be in principle wrapped around any learning method, but it is beneficial if the method can be efficiently trained or if the already learned model can be efficiently updated. Indeed, for some learning methods, such as regularized least-squares (RLS), the search process can be considerably accelerated with computational short-cuts for scoring function evaluation [23]. These inbuilt short-cuts bring the methods closer to the next category of the selection methods, namely the embedded ones.

Embedded methods

Embedded methods have the feature selection mechanism built into the training algorithm itself [35], that is, the predictive models they produce tend to depend only on a subset of the original features. Perhaps the most well-known embedded method is LASSO (least absolute shrinkage and selection operator), which is also recently being applied to a larger number of GWA studies [20,25,36-38]. While only a few machine learning approaches, in fact, allow for scaling-up to the genome-wide level, this has been made possible in LASSO by the recently developed model training algorithms, such as those based on the coordinate descent methods, which are computationally very efficient. The problem setup resembles the wrapper approach in the sense that there is an objective function for which one performs a stochastic search, such as cyclic or stochastic coordinate descent, in order to find a global optimum. Basically, the search algorithm goes through each feature at a time, and updates the corresponding coefficient in the linear model under construction. The objective function consists of a scoring metric such as the mean squared error (MSE) on the training data and a regularization term that favors sparse linear models, that is, it tends to push the search algorithm towards such models that have only a few nonzero parameters. Typically, coordinate descent passes through the whole data set only a couple of times before convergence, but the number of passes depends on the properties of the data, the desired sparsity level, and the other possible hyperparameters. Wrappers and embedded methods are known to have the ability to produce better results than filter methods in many applications [23,25,39], but if not implemented correctly, they can easily lead to the models failing to generalize beyond the training data, underscoring the importance of rigid evaluation of the prediction models.

The importance of evaluation of the predictive models for complex phenotypes

One of the main challenges in feature selection is the accurate estimation of the prediction performance of the machine learning models on new samples unseen at the training phase, especially in settings in which the data is high-dimensional and the number of labeled training data is relatively small. Given the massive dimensionality of modern GWAS and NGS studies, it is in fact not very hard to find genetic features that can almost perfectly fit to a small training set but fail to generalize to unseen data, a phenomenon known as ***model overfitting***. Therefore, the models learned from genetic data should always be tested on independent data not used for training the model. In case the number of labeled data is small, one must resort to ***cross-validation*** techniques that repeatedly split the data into training and test sets, and the predictive accuracy is reported as an average over the test folds. In many applications of genomic predictors, there are a number of examples of the so-called ***selection bias*** [40], meaning that the cross-validation is used to estimate the performance of the learning algorithm only, but not the preliminary feature selection done on the whole data, therefore leading to information leak and grossly over-optimistic results. Further, if cross-validation is used for selecting the hyper-parameters of the learning algorithm or for feature selection, this needs to be done within an internal cross-validation loop, separately during each round of an outer cross-validation loop [40-43]. This two-level technique is sometimes referred to as the ***nested cross-validation*** [42,44]. An example demonstrating the behavior of a cross-validation error when it is used as a selection criterion with greedy forward selection is presented in Figure 1. The error curve that constantly decreases as a function of the number of selected features clearly indicates that the cross-validation becomes a part of the training algorithm itself in the inner loop, and therefore it cannot be trusted as a measure of true prediction performance for unseen data.

The evaluation of the predictive power is important also when considering predictive models constructed on the basis on statistical significant variants. For instance, there are numerous observations showing that the increases in the proportion of variance explained by significant variants does not go hand in hand with improved genetic prediction of disease risk. For instance, when using statistical modeling on the single training sample only, a panel of thousands of non-significant variants collectively could capture over one-third of the heritability for schizophrenia, but the same panel only explained a few percent of disease susceptibility in another replication cohort [8]. Similarly, while the statistical explanation power of the genetic variation in human height could be substantially increased by considering increasing number of common variants in a single population sample [45], the proportion of variance accounted for in other independent samples was much smaller [46]. These examples underscore the importance of rigid validation of the predictive accuracy of the models based on genetic profiles. While external cross-validation is a valid option, it is not free of any study-specific factors. For example, if there is a problem during the genotyping phase, it will appear also in any training and test data splits. These errors, stemming from problems during the experimental design and/or quality control have led for the need to re-evaluate the established methods and use caution when claiming replication [47]. The recommended option for truly validating the generalizability of predictive risk models

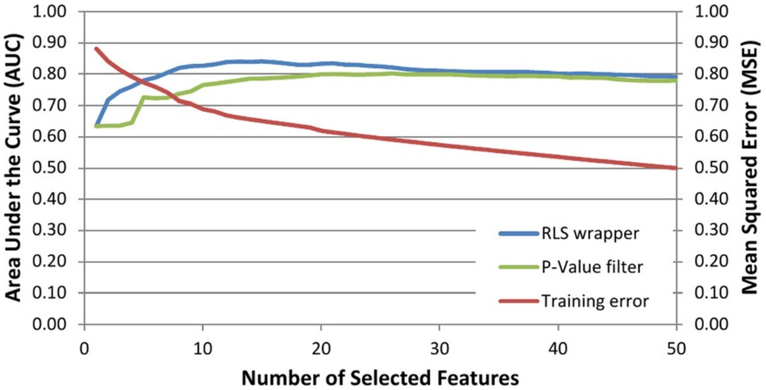

Figure 1 The figure illustrates how the external and internal cross-validation results behave as functions of the number of selected features. The external-cross validation consists of three training/test splits. The wrapper-based feature selection method, greedy RLS [23], is separately run during each round of the external cross-validation. Greedy RLS, in turn, employs an internal leave-one-out cross-validation on the training set for scoring the feature set candidates. The red curve depicts the mean values over these internal cross-validation errors. As can be easily observed from the blue curve, this internal cross-validation MSE used for the model training keeps constantly improving, which is expected, because the internal cross-validation quickly overfits to the training data when it is used as a selection measure. The blue curve depicts the area under curve (AUC) on the test data, held out during the external cross-validation round, that is, data completely unseen during the internal cross-validation and feature selection process. In contrast to the red curve, the blue curve starts to level off soon after the number of selected variants reaches around 10, indicating that adding extra features is not beneficial anymore even if the internal scoring function keeps improving. The green curve depicts the AUC of the RLS model trained using features selected by single-locus *p*-value based filter method, Fisher's exact test, which is run with the same external training/test split as the greedy selection method. Similarly to the blue curve, the green one also stops improving soon after a relatively small set of features has been selected. The data used in the experiments is the Wellcome Trust Case Controls Consortium (WTCCC) Hypertension dataset combined with the UK National Blood Services' controls.

is to make use of a large enough set of independent samples in which there is no overlap between the examined cohorts [48]. However, here one should consider whether the aim is to validate the predictive model itself (e.g. using external cross-validation or independent validation samples), or the predictive variants selected by the model (replication of the model construction or its application to separate cohorts) [49].

Through the development of better model validation techniques and unbiased examination of all feature subsets in genome-wide scale, we are likely to continuously improve the accuracy of the predictive models and increase their reproducibility on independent population samples. A challenge here is that differences in the population genetic structure, attributable to confounding factors such as the ethnicity or ancestry of the subjects, may result in highly heterogeneous datasets with a number of hidden subject sub-groups, which may associate with divergent disease phenotypes and therefore cause an increased false-positive rates [50]. Related to this, while there are comparisons among various feature selection methods and predictive modeling frameworks on individual cohorts [23,24,27], there is not yet any definitive results whether one method will universally lead to optimal results in other subject cohorts or populations. Such confounding variability should also be taken into account in the model construction and evaluation, perhaps in some form of population stratified cross-validation. Failure to replicate a genetic association should not only be considered as a negative result, as it may also provide important clues about genetic architecture among study populations or genetic interactions among risk variants [51]. When epistasis interactions are involved, then it is likely that simple methods, such as single-locus filters, will not alone be able to provide most optimal results, while in extremely large datasets, wrapper methods may pose computational limitations if combined with complex prediction models. Finally, even though the improvements obtained by the machine learning wrappers, compared to those from the traditional *p*-value based filters, may seem quite modest (e.g. Figure 1), it may turn out that even slight improvements in the predictive accuracy can result in significant clinical benefits. Moreover, it is argued that the modest predictive improvements may be further aggregated through pathway and network-level analyses of the selected variants.

Molecular networks as a prior information for constructing predictive models

Even in the absence of significant single-locus marginal effects, multiple genetic loci from a number of molecular pathways may act synergistically and lead to disease phenotype when combined. Therefore, it has become popular to map the genetic loci identified in GWA or NGS studies to established biological pathways in order to elucidate the potential cellular mechanisms behind the observed genetic and phenotypic variation. There exist a wide variety of tools and guidelines on how to implement such pathway analyses in the context of genetic association studies [52-56]. Building on approaches originally developed in the context of microarray gene expression experiments, the common theme in the pathway analysis approaches is that they examine whether a group of related loci in the same biological pathway are jointly associated with a trait of interest. In line with the observations in microarray gene expression studies, it has been shown that in those cases where there is only a modest overlap in the variant or gene-level findings between different studies, due to factors such as differences in the genetic structure, the pathway-level associations may be much more reproducible even between different study populations [57-60]. These findings support the concept that individuals with the same disease phenotype may have marked inter-individual genetic heterogeneity in the sense that their disease predisposing variants may lie in distinct loci within the same or related pathways [14]. Machine learning-based predictive models constructed upon gene expression profiling have already shown the benefits of using pathway activities as features in terms of improved classification accuracy, compared to those models that consider merely individual gene expression levels [61]. It has also been demonstrated in the context of GWA datasets that pathway analysis can provide not only mechanistic insights but also improved discrimination power using tailored statistical data mining techniques, such as HyperLasso [62] or so-called pathways of distinction analysis (PoDA) [63].

A limitation of constructing predictive models for a disease merely on the basis of established pathways is that these models may become biased toward already known biological processes, thereby potentially missing novel yet causal mechanisms predictive of the disease risk [64]. It may also not be so straightforward to infer the set of pathways that should be included in the model building process, in the absence of any *a priori* knowledge. Perhaps more importantly, statistical analysis of separate biological pathways or distinct gene sets undermines the effect of pathway cross-talk behind disease development, in which multiple genetic variants from distinct molecular pathways show synergistic contribution to the disease phenotype. In practice, the regulatory relationships behind many phenotypes are determined by complex and highly interconnected networks of physical and functional interplay between a multitude of pathway components [16]. As an example, we constructed a network representation for variants predictive of type 1 diabetes risk, which illustrate a selected portion of the number of pathways and their relationships that may be predictive of the disease onset (Figure 2). Given such high degree of interconnectivity, not only between the genetic variants but also among the implicated pathways, it is not surprising that the first machine learning frameworks for explicitly accounting epistatic gene-gene interactions have focused mostly on measures from information theory, such as those based on additive models, information gain, conditional entropy, or mutual information [24,65-67]. These models treat pairwise genetic interactions in a way that closely resembles the classic definition of epistasis, involving single and double-deletion experiments in model organisms [68]. However, even if allowing computationally efficient exploration of genetic interactions, *a posteriori* detection and heuristic search schemes cannot guarantee that the detected pairs of genetic risk factors will eventually be the most essential ones for the improved predictive power among all the possible variant combinations.

Toward more systematic network-centric analysis of genetic variants on a genome-wide scale, molecular interaction networks can be used as *a priori* information in the predictive models, in the form e.g. filters or integrators, with the aim of either reducing

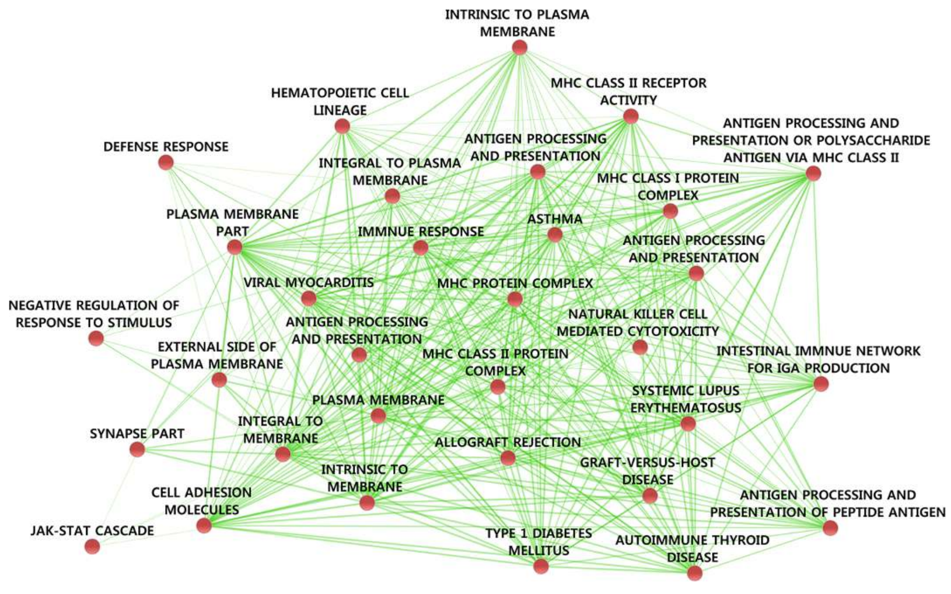
INTRINSIC TO PLASMA MEMBRANE
MHC CLASS II RECEPTOR ACTIVITY
HEMATOPOIETIC CELL LINEAGE
ANTIGEN PROCESSING AND PRESENTATION OR POLYSACCHARIDE ANTIGEN VIA MHC CLASS II
DEFENSE RESPONSE
ANTIGEN PROCESSING AND PRESENTATION
INTEGRAL TO PLASMA MEMBRANE
MHC CLASS I PROTEIN COMPLEX
PLASMA MEMBRANE PART
ASTHMA
IMMNUE RESPONSE
ANTIGEN PROCESSING AND PRESENTATION
VIRAL MYOCARDITIS
MHC PROTEIN COMPLEX
NEGATIVE REGULATION OF RESPONSE TO STIMULUS
NATURAL KILLER CELL MEDIATED CYTOTOXICITY
ANTIGEN PROCESSING AND PRESENTATION
EXTERNAL SIDE OF PLASMA MEMBRANE
MHC CLASS II PROTEIN COMPLEX
INTESTINAL IMMNUE NETWORK FOR IGA PRODUCTION
SYSTEMIC LUPUS ERYTHEMATOSUS
PLASMA MEMBRANE
SYNAPSE PART
INTEGRAL TO MEMBRANE
ALLOGRAFT REJECTION
GRAFT-VERSUS-HOST DISEASE
INTRINSIC TO MEMBRANE
CELL ADHESION MOLECULES
ANTIGEN PROCESSING AND PRESENTATION OF PEPTIDE ANTIGEN
JAK-STAT CASCADE
TYPE 1 DIABETES MELLITUS
AUTOIMMUNE THYROID DISEASE

Figure 2 Sample network visualization constructed for type 1 diabetes. The risk variants were selected using the greedy RLS on the WTCCC type 1 diabetes GWAS data and the UK National Blood Services' controls, extended with those genes selected in another work [62]. The biological processes and pathways were then mapped using DAVID [112,113], and the network visualization was done with the Enrichment Map plugin for Cytoscape [114,115]. The nodes represent pathways and the edges are the amount of overlap between the members of the pathways. The visualized network represents a selected sub-network of complex interconnections and cross-talks between a number of pathways, including MHC-related processes and other biological pathways associated with diabetes phenotypes. The pathways were identified initially using DAVID, with the criteria that they demonstrate enrichment when compared to the genome-wide background. The retrieved pathways were subsequently filtered in Cytoscape through the Enrichment Map plugin using the false-discovery rate and overlap coefficient to filter out non-significant pathways.

the massive size of the search space in the variant selection process or boosting the signal-to-noise ratio through external knowledge incorporated in terms of physical or functional molecular networks [69,70]. Network graphs provide a convenient mathematical framework for modeling, integrating and mining high-dimensional genomic datasets, in which to present the relationships among genetic loci, genes and diseases [64,69-72]. Successful examples of combining individual-level gene expression measurements with background networks of physical interactions between proteins and transcription factor targets have demonstrated that it is possible to identify and make use of disease-specific sub-networks, so-called ***modules***, in order to reduce both the number of false positives and negatives, caused by factors such as technical variability and genetic heterogeneity, respectively, as well as to improve individual-level prediction of clinical outcomes, such as cancer metastasis or survival time [64,73-75]. There are also studies in the context of GWA datasets, which motivate the use of network connectivity structures, such as sub-network modules or highly-connected network hubs [22,64,76-78], as aggregate features in the disease prediction models. However, what has been largely missing is a systematic approach that could combine network topology as *a priori* information when constructing predictive models. Recently, a particularly interesting approach was introduced as a principled method that uses genetic algorithms guided by the structure of a given gene interaction network to discover small groups of connected variants, which are jointly associated with a disease outcome on a genome-wide scale [79]. Combined with more efficient, wrapper-type of search algorithms, such network-guides feature selection approaches could be scaled-up in the future to enable extracting also larger sub-networks with improved predictive capability.

Future directions: lessons from model organisms and individualized medicine

Given the rather modest progress made so far in pursuing the expensive and suboptimal route of current drug discovery, there has been much interest lately in moving towards ***personalized medicine*** strategies [80,81]. Another major paradigm shift in disease treatment is moving away from the traditional 'one target, one drug' strategy towards the so-called ***network pharmacology***, a novel paradigm which provides more global understanding of the mechanisms behind disease processed and drug action by considering drug targets in their context of biological networks and pathways [82]. These emerging paradigms can offer holistic information on disease networks and drug responses, with the aim of identifying more effective drug targets and their combinations tailored for individualized treatment strategies. A prime challenge in developing such strategies is to understand how genes function as interaction networks to carry out and regulate cellular processes, and how perturbations in these cellular networks cause certain phenotypes, such as human diseases, in some individuals, but not in the others. There has been active research in model organisms addressing the question why disease causing mutations do not cause the disease in all individuals [14]. Recent studies in yeast *Saccharomyces cerevisiae*, worm *Caenorhabditis elegans*, and fly *Drosophila melanogaster* have demonstrated the importance of incorporating functional genetic interaction partners of the mutated genes in the prediction of phenotypic variation and mutational outcomes at an individual level [83-85]. Pilot studies in human trials have also suggested that personal genomic approaches, such as those based on GWA or NGS studies, may indeed yield useful and clinically relevant information for individual patients [1,2]. However, a number of experimental, modeling and computational challenges have to be solved before the promises of personalized medicine can be translated into routine clinical practice [5,81,86].

From the experimental point of view, the whole-genome sequencing efforts will enable us to delve deeper into the individual genomes by elucidating the role of low-frequency variants in the genetic architecture of complex diseases. The sequencing efforts, such as the 1000 Genomes project [10], are also being used to subsequently extend the coverage of the existing GWA datasets by means of imputation methods and population-specific reference haplotypes [87,88]. However, while the emerging shift from population-level common variants toward individual-level rare or even personal variants holds great promise for medical research, it also represents with unique modeling challenges; in particular, the traditional statistical modeling frameworks that were developed under settings where the number of study samples greatly exceeds the number of study variables may not to be ideally suited for the personalized medicine settings, in which the individuals and disease subtypes are stratified into increasingly smaller subgroups [89]. Although machine learning methods are better targeted at individual-level prediction making, the feature selection methods would also benefit from more stratified options, for instance, in terms of enabling phenotype-specific genetic features, rather than assuming that all subjects share the same panel of predictive genotypes. Also, since the binary disease outcomes, typically in the form of case or control dichotomy, may not provide the most reliable study phenotypes, the predictive modeling frameworks might become more successful for predicting quantitative phenotypic traits [90-92]. This also raises related modeling questions, such as how to encode imputed variants (e.g. expected or most likely genotype), how to treat missing data (exclude or impute), or how to model the variants and their interactions (multiplicative, additive, recessive or dominant models) [90-94]; these all may have an important effect on the prediction performance, especially in the presence of epistatic interactions at an individual level.

From the computational perspective, the ever increasing sizes of the raw NGS and imputed GWA datasets pose great challenges to the computational algorithms. For instance, while systematic genetic mappings in model organisms have revealed widespread genetic interactions within individual species [85,95-97], epistasis interactions have remained extremely difficult to identify on a global scale in human populations. This can be attributed to the vast number of potential interaction partners, along with complex genotype-phenotype relationships and their individual-level differences. Improvements in computational performance have recently been obtained through effective usage of computer hardware, for instance, through graphics processing units, Cloud-based computing environments, or multithread parallelization, when exploring genetic variants or their interactions in GWA studies [98-101]. Furthermore, since the memory consumption in the high-dimensional NGS applications can form even a tighter bottleneck than the running time, there is also a need to develop space-efficient implementations, which trade running time for decreased memory consumption [23]. Lessons from model organisms, such as yeast, have also demonstrated that data integration between complementary screening approaches, either functional or physical assays, can reveal novel genetic interactions and their modular organization which have gone undetected by any of the individual approaches alone [95,96,102]. Also, integrating diverse phenotypic readouts facilitates genetic interaction screens [103], and Bayesian models have been shown especially useful for making use of multiple traits, gene-gene or gene-environment interactions in disease risk prediction [104]. Finally, visualization algorithms that can capture the hierarchical modularity of the physical and functional interaction networks may help reveal interesting biological patterns and relationships within the data, such as pathway components and biological processes, which can be further investigated by follow-up computational and/or experimental analyses [105].

Better understanding of the general design principles underlying genetic interaction networks in model organisms can provide important insights into the relationships between genotype and phenotype, toward better understanding and treating also complex human diseases, such as cancers. Cancer phenotypes are known to arise and develop from various genetic alterations, and therefore the same therapy often results in different treatment responses. Moreover, the underlying genetic heterogeneity results in alterations within multiple molecular pathways, which lead to various cancer phenotypes and make most tumors resistant to single agents. Cancer sequencing efforts, such as The Cancer Genome Atlas (TCGA), are systematically characterizing the structural basis of cancer, by identifying the genomic mutations associated with each cancer type. These efforts have revealed tremendous inter-individual mutational and phenotypic heterogeneity, which renders it difficult to translate the genetic information into clinically actionable individualized treatment strategies [106-108]. Therefore, integrating the structural genomic information with systematic functional assessment of genes for their contribution to genetic dependencies and cancer vulnerabilities, such as oncogenic addictions or synthetic lethalities [109,110], is likely needed for providing more comprehensive insight into the molecular mechanisms and pathways behind specific cancer types and for improving their prevention, diagnosis and treatment [106,111]. Machine learning-based predictive modeling approaches are well-powered to make the most of the exciting functional and genetic screens toward revealing hidden genetic variants and their interactions behind cancer and other complex phenotypes. When combined with network analyses, these integrated systems medicine approaches may offer the possibility to identify key players and their relationships responsible for multi-factorial behavior in disease networks, with many diagnostic, prognostic and pharmaceutical applications.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SO contributed to the drafting of the manuscript and conducting experiments for the illustrations. TP contributed to the drafting of the manuscript. TA conceived the study, participated in the design of the experiments and contributed to the drafting of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Dr. Antti Airola for his help with the RLS algorithm, the two expert reviewers, Prof. Greg Gibson and Prof. Jason Moore, for their constructive comments, and CSC, the Finnish IT center for science, for providing us with extensive computational resources. The work was supported by the Academy of Finland (grants 120 569, 133 227 and 140 880 to T.A. and 134020 to T.P.), the Turku Centre for Computer Science (TUCS), Turun Yliopistosäätiö and the Finnish Cultural Foundation. This study makes use of data generated by the Wellcome Trust Case-Control Consortium (WTCCC). A full list of the investigators who contributed to the generation of the WTCCC data is available from www.wtccc.org.uk. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113.

Author details
[1]Department of Information Technology, University of Turku, Turku, Finland. [2]Turku Centre for Computer Science (TUCS), Turku, Finland. [3]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland.

Received: 5 October 2012 Accepted: 11 February 2013
Published: 1 March 2013

References

1. Ashley EA, *et al*: **Clinical assessment incorporating a personal genome.** *Lancet* 2010, **375**(9725):1525–1535.
2. Ripatti S, *et al*: **A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses.** *Lancet* 2010, **376**(9750):1393–1400.
3. Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447**(7145):661–678.
4. Donnelly P: **Progress and challenges in genome-wide association studies in humans.** *Nature* 2008, **456**(7223):728–731.
5. Manolio TA: **Genomewide association studies and assessment of the risk of disease.** *N Engl J Med* 2010, **363**(2):166–176.
6. Lander ES: **Initial impact of the sequencing of the human genome.** *Nature* 2011, **470**(7333):187–197.
7. Maher B: **Personal genomes: The case of the missing heritability.** *Nature* 2008, **456**(7218):18–21.
8. Gibson G: **Hints of hidden heritability in GWAS.** *Nat Genetics* 2010, **42**(7):558–560.
9. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH: **Missing heritability and strategies for finding the underlying causes of complex disease.** *Nat Rev Genetics* 2010, **11**(6):446–450.
10. Zuk O, Hechter E, Sunyaev SR, Lander ES: **The mystery of missing heritability: Genetic interactions create phantom heritability.** *Proc Natl Acad Sci U S A* 2012, **109**(4):1193–1198.
11. Lehner B: **Modelling genotype-phenotype relationships and human disease with genetic interaction networks.** *J Exp Biol* 2007, **210**(Pt 9):1559–1566.
12. Moore JH, Williams SM: **Epistasis and its implications for personal genetics.** *Am J Hum Genet* 2009, **85**(3):309–320.
13. Cordell HJ: **Detecting gene-gene interactions that underlie human diseases.** *Nat Rev Genet* 2009, **10**(6):392–404.
14. Lehner B: **Molecular mechanisms of epistasis within and between genes.** *Trends Genet* 2011, **27**(8):323–331.
15. Moore JH, Asselbergs FW, Williams SM: **Bioinformatics challenges for genome-wide association studies.** *Bioinformatics* 2010, **26**(4):445–455.
16. Califano A, Butte AJ, Friend S, Ideker T, Schadt E: **Leveraging models of cell regulation and GWAS data in integrative network-based association studies.** *Nat Genet* 2012, **44**(8):841–847.
17. Jakobsdottir J, Gorin MB, Conley YP, Ferrell RE, Weeks DE: **Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers.** *PLoS Genet* 2009, **5**(2):e1000337.
18. Wei Z, Wang K, Qu H-Q, Zhang H, Bradfield J, *et al*: **From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes.** *PLoS Genet* 2009, **5**(10):e1000678.
19. 1000 Genomes Project: **A map of genome variation from population-scale sequencing.** *Nature* 2010, **467**(7319):1061–1073.
20. Kruppa J, Ziegler A, König IR: **Risk estimation and risk prediction using machine-learning methods.** *Hum Genet* 2012, **131**(10):1639–1654.
21. Pattin KA, Moore JH: **Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases.** *Hum Genet* 2008, **124**(1):19–29.
22. Barrenäs F, Chavali S, Alves AC, Coin L, Jarvelin MR, Jörnsten R, Langston MA, Ramasamy A, Rogers G, Wang H, Benson M: **Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms.** *Genome Biol* 2012, **13**(6):R46.
23. Pahikkala T, Okser S, Airola A, Salakoski T, Aittokallio T: **Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations.** *Algorithm Mol Biol* 2012, **7**(1):11.
24. Okser S, Lehtimäki T, Elo LL, Mononen N, Peltonen N, *et al*: **Genetic Variants and Their Interactions in the Prediction of Increased Pre-Clinical Carotid Atherosclerosis: The Cardiovascular Risk in Young Finns Study.** *PLoS Genet* 2010, **6**(9):e1001146.
25. Kooperberg C, LeBlanc M, Obenchain V: **Risk prediction using genome-wide association studies.** *Genet Epidemiol* 2010, **34**(7):643–652.
26. Balding DJ: **A tutorial on statistical methods for population association studies.** *Nat Rev Genet* 2006, **7**(10):781–791.
27. Evans DM, Visscher PM, Wray NR: **Harnessing the Information Contained Within Genome-wide Association Studies to Improve Individual Prediction of Complex Disease Risk.** *Hum Mol Genet* 2009, **18**(18):3525–3531.
28. Clarke GM, Anderson CA, Pettersson FH, Cardon LR, Morris AP, Zondervan KT: **Basic statistical analysis in genetic case-control studies.** *Nat Protoc* 2011, **6**(2):121–133.
29. Bansal V, Libiger O, Torkamani A, Schork NJ: **Statistical analysis strategies for association studies involving rare variants.** *Nat Rev Genet* 2010, **11**(11):773–785.
30. Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CM, Richards JB: **The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals.** *PLoS Genet* 2012, **8**(2):e1002496.
31. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, NHLBI GO Exome Sequencing Project—ESP Lung Project Team, Christiani DC, Wurfel MM, Lin X: **Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies.** *Am J Hum Genet* 2012, **91**(2):224–237.
32. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, Moore JH: **Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer.** *Am J Hum Genet* 2001, **69**(1):138–147.

33. Kraft P, Wacholder S, Cornelis MC, Hu FB, Hayes RB, Thomas G, Hoover R, Hunter DJ, Chanock S: **Beyond odds ratios: communicating disease risk based on genetic profiles. Perspective.** *Nat Rev Genetics* 2009, **10:**264–269.
34. Saeys Y, Inza I, Larrañaga P: **A review of feature selection techniques in bioinformatics.** *Bioinformatics* 2007, **23**(19):2507–2517.
35. Guyon I, Elisseeff A: **An introduction to variable and feature selection.** *J Mach Learn Res* 2003, **3:**1157–1182.
36. Wu TT, Chen YF, Hastie T, Sobel E, Lange K: **Genome-wide association analysis by lasso penalized logistic regression.** *Bioinformatics* 2009, **25**(6):714–721.
37. He Q, Lin DY: **A variable selection method for genome-wide association studies.** *Bioinformatics* 2011, **27**(1):1–8.
38. Rakitsch B, Lippert C, Stegle O, Borgwardt K: **A Lasso multi-marker mixed model for association mapping with population structure correction.** *Bioinformatics* 2013, **29**(2):206–214.
39. Aha DW, Bankert RL: **A comparative evaluation of sequential feature selection algorithms.** In *Learning from Data: Artificial Intelligence and Statistics V, Lecture Notes in Statistics*. Edited by Fisher DH, Lenz HJ. New York: Springer-Verlag; 1996:199–206.
40. Ambroise C, McLachlan GJ: **Selection bias in gene extraction on the basis of microarray gene-expression data.** *Proc Natl Acad Sci U S A* 2002, **99**(10):6562–6566.
41. Simon R, Radmacher MD, Dobbin K, McShane LM: **Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification.** *J Natl Cancer Inst* 2003, **95**(1):14–18.
42. Varma S, Simon R: **Bias in error estimation when using cross-validation for model selection.** *BMC Bioinformatics* 2006, **7:**91.
43. Smialowski P, Frishman D, Kramer S: **Pitfalls of supervised feature selection.** *Bioinformatics* 2010, **26**(3):440–443.
44. Statnikov A, Wang L, Aliferis C: **A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification.** *BMC Bioinformatics* 2008, **9**(1):319.
45. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM: **Common SNPs explain a large proportion of the heritability for human height.** *Nat Genet* 2010, **42**(7):565–569.
46. Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, de los Campos G: **Beyond missing heritability: prediction of complex traits.** *PLoS Genet* 2011, **7**(4):e1002051.
47. Lambert CG, Black LJ: **Learning from our GWAS mistakes: from experimental design to scientific method.** *Biostatistics* 2012, **13**(2):195–203.
48. Castaldi PJ, Dahabreh IJ, Ioannidis JP: **An empirical assessment of validation practices for molecular classifiers.** *Brief Bioinform* 2011, **12**(3):189–202.
49. König I: **Validation in genetic association studies.** *Brief Bioinform* 2011, **12**(3):253–258.
50. Tian C, Gregersen PK, Seldin MF: **Accounting for ancestry: population substructure and genome-wide association studies.** *Hum Mol Genet* 2008, **17**(R2):R143–R150.
51. Greene CS, Penrod NM, Williams SM, Moore JH: **Failure to replicate a genetic association may provide important clues about genetic architecture.** *PLoS One* 2009, **4**(6):e5639.
52. Torkamani A, Topol EJ, Schork NJ: **Pathway analysis of seven common diseases assessed by genome-wide association.** *Genomics* 2008, **92**(5):265–272.
53. Torkamani A, Schork NJ: **Pathway and network analysis with high-density allelic association data.** *Methods Mol Biol* 2009, **563:**289–301.
54. Zhong H, Yang X, Kaplan LM, Molony C, Schadt EE: **Integrating pathway analysis and genetics of gene expression for genome-wide association studies.** *Am J Hum Genet* 2010, **86**(4):581–591.
55. Wang K, Li M, Hakonarson H: **Analysing biological pathways in genome-wide association studies.** *Nat Rev Genet* 2010, **11**(12):843–854.
56. Ramanan VK, Shen L, Moore JH, Saykin AJ: **Pathway analysis of genomic data: concepts, methods, and prospects for future development.** *Trends Genet* 2012, **28**(7):323–332.
57. Srinivasan BS, Doostzadeh J, Absalan F, Mohandessi S, Jalili R, Bigdeli S, Wang J, Mahadevan J, Lee CL, Davis RW, William Langston J, Ronaghi M: **Whole genome survey of coding SNPs reveals a reproducible pathway determinant of Parkinson disease.** *Hum Mutat* 2009, **30**(2):228–238.
58. Askland K, Read C, Moore J: **Pathways-based analyses of whole-genome association study data in bipolar disorder reveal genes mediating ion channel activity and synaptic neurotransmission.** *Hum Genet* 2009, **125**(1):63–79.
59. Luo L, Peng G, Zhu Y, Dong H, Amos CI, Xiong M: **Genome-wide gene and pathway analysis.** *Eur J Hum Genet* 2010, **18**(9):1045–1053.
60. Peng G, Luo L, Siu H, Zhu Y, Hu P, Hong S, Zhao J, Zhou X, Reveille JD, Jin L, Amos CI, Xiong M: **Gene and pathway-based second-wave analysis of genome-wide association studies.** *Eur J Hum Genet* 2010, **18**(1):111–117.
61. Lee E, Chuang HY, Kim JW, Ideker T, Lee D: **Inferring pathway activity toward precise disease classification.** *PLoS Comput Biol* 2008, **4**(11):e1000217.
62. Eleftherohorinou H, Wright V, Hoggart C, Hartikainen AL, Jarvelin MR, Balding D, Coin L, Levin M: **Pathway Analysis of GWAS Provides New Insights into Genetic Susceptibility to 3 Inflammatory Diseases.** *PLoS One* 2009, **4**(11):e8068.
63. Braun R, Buetow K: **Pathways of distinction analysis: a new technique for multi-SNP analysis of GWAS data.** *PLoS Genet* 2011, **7**(6):e1002101.
64. Bebek G, Koyutürk M, Price ND, Chance MR: **Network biology methods integrating biological data for translational science.** *Brief Bioinform* 2012, **13**(4):446–459.
65. McKinney BA, Crowe JE, Guo J, Tian D: **Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis.** *PLoS Genet* 2009, **5**(3):e1000432.
66. Lavender NA, Rogers EN, Yeyeodu S, Rudd J, Hu T, Zhang J, Brock GN, Kimbro KS, Moore JH, Hein DW, Kidd LC: **Interaction among apoptosis-associated sequence variants and joint effects on aggressive prostate cancer.** *BMC Med Genomics* 2012, **5:**11.
67. Hu T, Sinnott-Armstrong NA, Kiralis JW, Andrew AS, Karagas MR, Moore JH: **Characterizing genetic interactions in human disease association studies using statistical epistasis networks.** *BMC Bioinformatics* 2011, **12:**364.

68. Phillips PC: **Epistasis: the essential role of gene interactions in the structure and evolution of genetic systems.** *Nat Rev Genet* 2008, **9**(11):855–867.
69. Schadt EE: **Molecular networks as sensors and drivers of common human diseases.** *Nature* 2009, **461**(7261):218–223.
70. Ideker T, Dutkowski J, Hood L: **Boosting signal-to-noise in complex biology: prior knowledge is power.** *Cell* 2011, **144**(6):860–863.
71. Vidal M, Cusick ME, Barabási AL: **Interactome networks and human disease.** *Cell* 2011, **144**(6):986–998.
72. Barabási AL, Gulbahce N, Loscalzo J: **Network medicine: a network-based approach to human disease.** *Nat Rev Genet* 2011, **12**(1):56–68.
73. Chuang HY, Lee E, Liu YT, Lee D, Ideker T: **Network-based classification of breast cancer metastasis.** *Mol Syst Biol* 2007, **3**:140.
74. Winter C, Kristiansen G, Kersting S, Roy J, Aust D, Knösel T, Rümmele P, Jahnke B, Hentrich V, Rückert F, Niedergethmann M, Weichert W, Bahra M, Schlitt HJ, Settmacher U, Friess H, Büchler M, Saeger HD, Schroeder M, Pilarsky C, Grützmann R: **Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes.** *PLoS Comput Biol* 2012, **8**(5):e1002511.
75. Lavi O, Dror G, Shamir R: **Network-induced classification kernels for gene expression profile analysis.** *J Comput Biol* 2012, **19**(6):694–709.
76. Feldman I, Rzhetsky A, Vitkup D: **Network properties of genes harboring inherited disease mutations.** *Proc Natl Acad Sci U S A* 2008, **105**(11):4323–4328.
77. Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, Wu W, Uitdehaag BM, Kappos L, GeneMSA Consortium, Polman CH, Matthews PM, Hauser SL, Gibson RA, Oksenberg JR, Barnes MR: **Pathway and network-based analysis of genome-wide association studies in multiple sclerosis.** *Hum Mol Genet* 2009, **18**(11):2078–2090.
78. McKinney BA, Pajewski NM: **Six Degrees of Epistasis: Statistical Network Models for GWAS.** *Front Genet* 2012, **2**:109.
79. Mooney M, Wilmot B, The Bipolar Genome Study, McWeeney S: **The GA and the GWAS: Using Genetic Algorithms to Search for Multi-locus Associations.** *IEEE/ACM Trans Comput Biol Bioinform* 2012, **9**(3):899–910.
80. Deisboeck TS: **Personalizing medicine: a systems biology perspective.** *Mol Syst Biol* 2009, **5**:249.
81. Reynolds KS: **Achieving the promise of personalized medicine.** *Clin Pharmacol Ther* 2012, **92**(4):401–405.
82. Hopkins AL: **Network pharmacology: the next paradigm in drug discovery.** *Nat Chem Biol* 2008, **4**:682–690.
83. Jelier R, Semple JI, Garcia-Verdugo R, Lehner B: **Predicting phenotypic variation in yeast from individual genome sequences.** *Nat Genet* 2011, **43**(12):1270–1274.
84. Burga A, Casanueva MO, Lehner B: **Predicting mutation outcome from early stochastic variation in genetic interaction partners.** *Nature* 2011, **480**(7376):250–253.
85. Huang W, Richards S, Carbone MA, Zhu D, Anholt RR, Ayroles JF, Duncan L, Jordan KW, Lawrence F, Magwire MM, Warner CB, Blankenburg K, Han Y, Javaid M, Jayaseelan J, Jhangiani SN, Muzny D, Ongeri F, Perales L, Wu YQ, Zhang Y, Zou X, Stone EA, Gibbs RA, Mackay TF: **Epistasis dominates the genetic architecture of Drosophila quantitative traits.** *Proc Natl Acad Sci USA* 2012, **109**(39):15553–15559.
86. Corander J, Aittokallio T, Ripatti S, Kaski S: **The rocky road to personalized medicine: computational and statistical challenges.** *Personalized Med* 2012, **9**(2):109–114.
87. Surakka I, Kristiansson K, Anttila V, Inouye M, Barnes C, Moutsianas L, Salomaa V, Daly M, Palotie A, Peltonen L, Ripatti S: **Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging.** *Genome Res* 2010, **20**(10):1344–1351.
88. Holm H, Gudbjartsson DF, Sulem P, Masson G, Helgadottir HT, Zanon C, Magnusson OT, Helgason A, Saemundsdottir J, Gylfason A, Stefansdottir H, Gretarsdottir S, Matthiasson SE, Thorgeirsson GM, Jonasdottir A, Sigurdsson A, Stefansson H, Werge T, Rafnar T, Kiemeney LA, Parvez B, Muhammad R, Roden DM, Darbar D, Thorleifsson G, Walters GB, Kong A, Thorsteinsdottir U, Arnar DO, Stefansson K: **A rare variant in MYH6 is associated with high risk of sick sinus syndrome.** *Nat Genet* 2011, **43**(4):316–320.
89. Marko NF, Weil RJ: **Mathematical modeling of molecular data in translational medicine: theoretical considerations.** *Sci Transl Med* 2010, **2**(56):56rv4.
90. Peltola T, Marttinen P, Jula A, Salomaa V, Perola M, Vehtari A: **Bayesian variable selection in searching for additive and dominant effects in genome-wide data.** *PLoS One* 2012, **7**(1):e29115.
91. Sebastiani P, Solovieff N, Dewan AT, Walsh KM, Puca A, Hartley SW, Melista E, Andersen S, Dworkis DA, Wilk JB, Myers RH, Steinberg MH, Montano M, Baldwin CT, Hoh J, Perls TT: **Genetic signatures of exceptional longevity in humans.** *PLoS One* 2012, **7**(1):e29848.
92. Ober U, Ayroles JF, Stone EA, Richards S, Zhu D, Gibbs RA, Stricker C, Gianola D, Schlather M, Mackay TF, Simianer H: **Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster.** *PLoS Genet* 2012, **8**(5):e1002685.
93. Sillanpää MJ: **Detecting interactions in association studies by using simple allele recoding.** *Hum Hered* 2009, **67**(1):69–75.
94. Ober U, Erbe M, Long N, Porcu E, Schlather M, Simianer H: **Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data.** *Genetics* 2011, **188**(3):695–708.
95. Beltrao P, Cagney G, Krogan NJ: **Quantitative genetic interactions reveal biological modularity.** *Cell* 2010, **141**(5):739–745.
96. Lindén RO, Eronen VP, Aittokallio T: **Quantitative maps of genetic interactions in yeast - comparative evaluation and integrative analysis.** *BMC Syst Biol* 2011, **5**:45.
97. Dixon SJ, Costanzo M, Baryshnikova A, Andrews B, Boone C: **Systematic mapping of genetic interaction networks.** *Annu Rev Genet* 2009, **43**:601–625.
98. Wang Z, Wang Y, Tan KL, Wong L, Agrawal D: **eCEO: an efficient Cloud Epistasis cOmputing model in genome-wide association study.** *Bioinformatics* 2011, **27**(8):1045–1051.
99. Chen GK: **A scalable and portable framework for massively parallel variable selection in genetic association studies.** *Bioinformatics* 2012, **28**(5):719–720.

100. Gyenesei A, Moody J, Laiho A, Semple CA, Haley CS, Wei WH: **BiForce Toolbox: powerful high-throughput computational analysis of gene-gene interactions in genome-wide association studies.** *Nucleic Acids Res* 2012, **40**(Web Server issue):W628–W632.
101. Schupbach T, Xenarios I, Bergmann S, Kapur K: **FastEpistasis: a high performance computing solution for quantitative trait epistasis.** *Bioinformatics* 2010, **26**(11):1468–1469.
102. Hannum G, Srivas R, Guénolé A, van Attikum H, Krogan NJ, Karp RM, Ideker T: **Genome-wide association data reveal a global map of genetic interactions among protein complexes.** *PLoS Genet* 2009, **5**(12):e1000782.
103. Michaut M, Bader GD: **Multiple genetic interaction experiments provide complementary information useful for gene function prediction.** *PLoS Comput Biol* 2012, **8**(6):e1002559.
104. Hartley SW, Monti S, Liu CT, Steinberg MH, Sebastiani P: **Bayesian methods for multivariate modeling of pleiotropic SNP associations and genetic risk prediction.** *Front Genet* 2012, **3**:176.
105. Tuikkala J, Vähämaa H, Salmela P, Nevalainen OS, Aittokallio T: **A multilevel layout algorithm for visualizing physical and genetic interaction networks, with emphasis on their modular organization.** *BioData Min* 2012, **26**(5):2.
106. Ashworth A, Lord CJ, Reis-Filho JS: **Genetic interactions in cancer progression and treatment.** *Cell* 2011, **145**(1):30–38.
107. Urbach D, Lupien M, Karagas MR, Moore JH: **Cancer heterogeneity: origins and implications for genetic association studies.** *Trends Genet* 2012, **28**(11):538–543.
108. Galvan A, Ioannidis JP, Dragani TA: **Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer.** *Trends Genet* 2010, **26**(3):132–141.
109. Kaelin WG Jr: **The concept of synthetic lethality in the context of anticancer therapy.** *Nat Rev Cancer* 2005, **5**(9):689–698.
110. Iglehart JD, Silver DP: **Synthetic lethality-a new direction in cancer-drug development.** *N Engl J Med* 2009, **361**(2):189–191.
111. Heiskanen MA, Aittokallio T: **Mining high-throughput screens for cancer drug targets—lessons from yeast chemical-genomic profiling and synthetic lethality.** *Wiley Interdisciplinary Rev: Data Min Knowl Discov* 2012, **2**(3):263–272.
112. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources.** *Nat Protocol* 2009, **4**(1):44–57.
113. Huang DW, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucleic Acids Res* 2009, **37**(1):1–13.
114. Smoot M, Ono K, Ruscheinski J, Wang P-L, Ideker T: **Cytoscape 2.8: new features for data integration and network visualization.** *Bioinformatics* 2011, **27**(3):431–432.
115. Merico D, Isserlin R, Stueker O, Emili A, Bader GD: **Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation.** *PLoS One* 2010, **5**(11):e13984.

doi:10.1186/1756-0381-6-5
Cite this article as: Okser *et al.*: Genetic variants and their interactions in disease risk prediction – machine learning and network perspectives. *BioData Mining* 2013 **6**:5.