Sebastian Okser

Scalable Feature Selection Applications for Genome-Wide Association Studies of Complex Diseases

Turku Centre *for* Computer Science

TUCS Dissertations
No 201, August 2015

Scalable Feature Selection Applications for Genome-Wide Association Studies of Complex Diseases

Sebastian Okser

To be presented, with the permission of the Faculty of Mathematics and Natural Sciences of the University of Turku, for public criticism in Auditorium Lambda on August 19, 2015, at 12 noon.

University of Turku
Department of Information Technology
FI-20014 Turun yliopisto
Finland

2015

Supervisors

Prof. Tero Aittokallio, PhD
Department of Mathematics and Statistics
University of Turku
Finland; and
FIMM-EMBL Group Leader
Institute for Molecular Medicine Finland (FIMM)
University of Helsinki
Finland

Adj. Prof. Tapio Pahikkala, PhD
Department of Information Technology
University of Turku
Finland

Prof. Olli T Raitakari, MD, PhD
Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku and Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital
Finland

Prof. Tapio Salakoski, PhD
Department of Information Technology
University of Turku
Finland

Reviewers

Prof. Harri Lähdesmäki, D.Sc.
Department of Computer Science
Aalto University
Finland

Prof. Garry Wong, PhD
Faculty of Health Sciences
University of Macau
Macau S.A.R, China

Opponent

Prof. Willem Waegeman, PhD
Department of Mathematical Modelling, Statistics and Bioinformatics
Ghent University
Belgium

ISBN 978-952-12-3245-9
ISSN 1239-1883

Abstract

Personalized medicine will revolutionize our capabilities to combat disease. Working toward this goal, a fundamental task is the deciphering of genetic variants that are predictive of complex diseases. Modern studies, in the form of genome-wide association studies (GWAS) have afforded researchers with the opportunity to reveal new genotype-phenotype relationships through the extensive scanning of genetic variants. These studies typically contain over half a million genetic features for thousands of individuals. Examining this with methods other than univariate statistics is a challenging task requiring advanced algorithms that are scalable to the genome-wide level. In the future, next-generation sequencing studies (NGS) will contain an even larger number of common and rare variants.

Machine learning-based feature selection algorithms have been shown to have the ability to effectively create predictive models for various genotype-phenotype relationships. This work explores the problem of selecting genetic variant subsets that are the most predictive of complex disease phenotypes through various feature selection methodologies, including filter, wrapper and embedded algorithms. The examined machine learning algorithms were demonstrated to not only be effective at predicting the disease phenotypes, but also doing so efficiently through the use of computational shortcuts. While much of the work was able to be run on high-end desktops, some work was further extended so that it could be implemented on parallel computers helping to assure that they will also scale to the NGS data sets.

Further, these studies analyzed the relationships between various feature selection methods and demonstrated the need for careful testing when selecting an algorithm. It was shown that there is no universally optimal algorithm for variant selection in GWAS, but rather methodologies need to be selected based on the desired outcome, such as the number of features to be included in the prediction model. It was also demonstrated that without proper model validation, for example using nested cross-validation, the models can result in overly-optimistic prediction accuracies and decreased generalization ability. It is through the implementation and application of machine learning methods that one can extract predictive genotype–phenotype relationships and biological insights from genetic data sets.

Tiivistelmä

Yksilöllistetty lääketiede on mullistamassa mahdollisuutemme ymmärtää ja paremmin hoitaa sairauksia. Yksilöllisen geneettisen vaihtelun vaikutuksen tutkiminen on keskeinen osa tätä tavoitetta, ja keskeisessä roolissa ovat uudet tekniikat, kuten koko genomin laajuinen geenivarianttien assosiaatioanalyysi. Modernit geneettiset analyysit koostuvat tyypillisesti tuhansille yksilöille kartoitetuista miljoonista geneettisestä piirteestä. Massiivisten aineistojen analysoiminen perinteisillä data-analyysimenetelmillä on haastavaa, ja seuraavan sukupolven sekvensointitekniikat tuottavat vielä jopa paljon suurempia datamääriä.

Koneoppimiseen perustuvien piirteidenvalintamenetelmien käyttö on osoittautunut tehokkaaksi tavaksi luoda ennustavia malleja genotyyppien ja fenotyyppien välisten vuorovaikutussuhteiden päättelemiseksi. Tämä väitöskirjatyö tarkastelee piirteidenvalintamenetelmien käyttöä erityisesti geenivarianttialijoukkojen ja monimutkaisten sairauksien välisten riippuvuuksien tutkimisessa. Työssä kehitetyt menetelmät osoittautuivat sekä ennustuskyvyltään hyviksi että laskennallisesti niin tehokkaiksi, että suuri osa algoritmeista voitiin ajaa jopa tavallisilla pöytätietokoneilla. Työssä esitellään lisäksi rinnakkaislaskentaa hyödyntävä algoritmi, joka skaalautuu vielä huomattavasti suuremmille datamäärille.

Tulokset osoittavat, ettei yksikään tarkastelluista piirteidenvalintamenetelmistä ole yleispätevä, vaan sopivin menetelmä pitää valita aina ratkaistavana olevan ongelman yksityiskohtaisten tavoitteiden perusteella. Hyvä esimerkki tästä on mallien ennustustarkkuuden ja valittujen piirteiden lukumäärän välinen kompromissi. Työssä tuodaan lisäksi esiin tarkan koesuunnittelun ja menetelmien testauksen merkitys. Erityisesti mallin validointi esimerkiksi sisäkkäisen ristiinvalidoinnin avulla on tärkeätä, jotta menetelmän ennustuskyky kyetään mittaamaan harhattomasti. Vain koneoppimismenetelmien huolellisen toteuttamisen ja soveltamisen avulla voidaan geneettisestä datasta löytää sellaisia genotyyppien ja fenotyyppien välisiä riippuvuuksia, jotka tuottavat uutta biologista näkemystä monimutkaisten sairauksien syntyyn.

Acknowledgements

I would like to start by thanking my research director and supervisor Professor Tapio Salakoski, who has provided me with the opportunity and support to conduct my research over the course of a number of years. Without this support, much of the work in the included publications would not have been possible. Working with his research group, I have been able to conduct meaningful research in an exciting environment and I am proud to have worked under him while pursuing my degree. My supervisor Adjunct Professor Tapio Pahikkala has provided me with the knowledge and tools necessary to conduct my research. Proving to be an essential mentor to my development as a researcher, he has taken a personal interest in seeing that I succeed in my endeavors. For this I owe him thanks that go above and beyond the words on this page. Professor Tero Aittokallio has been vital in starting my research career and introducing me to the world of genome-wide association studies. It is likely that without him my career would have taken a different path. He has continually pushed me to find solutions to complex problems while the work ethic that he has instilled in me is something that I take pride in and use daily. Professor Olli Raitakari has been instrumental in my early stages as a researcher, involving me in projects through which I have learned to not only produce results but also to question the current work that has been done and to try to develop better solutions to those problems.

Over the years it has been my pleasure to work and be associated with a number of excellent researchers. First and foremost I must thank Antti Airola. While not an official supervisor he has been a mentor, a colleague and a friend. He has been a co-author on the majority of my publications. Additionally, we have worked very closely and his willingness to teach me has been essential to my career. His support has been instrumental in getting me as far as I now am. I would like to thank Pekka Naula for being both a colleague and a friend. The two of us have grown as researchers together, often tackling similar algorithms but in different fields. It is with his support that I have been able to investigate ideas and attain better methodologies. I would also like to thank Jari Björne for being an inspiration to my work and someone that I have always strived to emulate. Jari has provided our group

with advice and insights when we were looking for new methods, helping us to come up with the ideas necessary to conduct this research.

Additionally, I am grateful to Eija Nordlund who runs the Master's program in Bioinformatics at the University of Turku. It was in this program that I began in Finnish academia and with her help I became a part of the research group where I have conducted this work. Marcus Alanen has always been there to provide advice. Though we never worked on projects together, his experience in the TUCS program has been forever helpful in navigating through the system while also encouraging my work and professional development. Providing me with continual support, my friend Rachel Ravens has helped me to expand my ideas and has taught me to evaluate my work critically. Additionally, I owe a great deal of gratitude to my employer Landy Ung who encouraged me to maintain focus on completing my dissertation.

The Turku Centre for Computer Science (TUCS) and the Department of Information Technology at the University of Turku, as well as their associated staff have been vital to my research. My work on this dissertation and the associated papers would not have been possible without their continual support. The Finnish Cultural Foundation and the Turun Yliopistosäätiö have also been critical to this work for their financial support of my projects.

I am grateful to Professor Garry Wong and Professor Harri Lähdesmäki for their excellent criticisms and comments on this dissertation. Additionally, I would like to thank Professor Willem Waegeman for acting as my opponent.

My parents Lewis and Claire have been supportive throughout my entire career. They encouraged me to travel far from my home in New York, which led to opportunities for me in Finland. To my wife Tuulia, I cannot find words that describe my gratitude for the support that I have received from you during the many years of these projects. Your love and encouragement have been essential to this process. And finally, to my young son Lukas, having joined my family at the close of my studies, you have been a delightful diversion.

List of original publications

I Okser, S., Lehtimäki, T., Elo, L. L., Mononen, N., Peltonen, N., Kähönen, M., Juonala, M., Fan, Y.-M., Hernesniemi, J. A., Laitinen, T., Lyytikäinen, L.-P., Rontu, R., Eklund, C., Hutri-Kähönen, N., Taittonen, L., Hurme, M., Viikari, J. S. A., Raitakari, O. T., and Aittokallio, T. (2010). Genetic variants and their interactions in the prediction of increased preclinical carotid atherosclerosis: The cardiovascular risk in young Finns study. *PLoS Genet*, 6(9):e1001146

II Pahikkala, T., Okser, S., Airola, A., Salakoski, T., and Aittokallio, T. (2012). Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations. *Algorithms for Molecular Biology*, 7(1):11

III Okser, S., Airola, A., Aittokallio, T., Salakoski, T., and Pahikkala, T. (2013). Parallel feature selection for regularized least-squares. In Manninen, P. and Öster, P., editors, *Applied Parallel and Scientific Computing*, volume 7782 of *Lecture Notes in Computer Science*, pages 280–294. Springer Berlin Heidelberg

IV Okser, S., Pahikkala, T., and Aittokallio, T. (2013). Genetic variants and their interactions in disease risk prediction - machine learning and network perspectives. *BioData Mining*, 6(1):5

V Okser, S., Pahikkala, T., Airola, A., Salakoski, T., Ripatti, S., and Aittokallio, T. (2014). Regularized machine learning in the genetic prediction of complex traits. *PLoS Genet*, 10(11):e1004754

List of original publications not included in the thesis

- Okser, S., Pahikkala, T., Airola, A., Aittokallio, T., and Salakoski, T. (2011). Fast and parallelized greedy forward selection of genetic variants in genome-wide association studies. In Chen, Y., Huang, Y., and Dougherty, E., editors, *IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS'11)*, pages 214–217. IEEE Signal Processing Society

List of contributions to original publications

I **Genetic Variants and Their Interactions in the Prediction of Increased Pre-Clinical Carotid Atherosclerosis: The Cardiovascular Risk in Young Finns Study**.
Contributed to the algorithm development, analysis of the data and drafting of the manuscript.

II **Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations**.
Design of the experiment, implementation of the algorithm on GWAS, analysis of the results and drafting of the manuscript.

III **Parallel Feature Selection for Regularized Least-Squares**.
Design of the parallel version of greedy RLS, implementation of the algorithm, analysis of the results and drafting of the manuscript.

IV **Genetic variants and their interactions in disease risk prediction - machine learning and network perspectives**.
Conducting the experiments and drafting of the manuscript.

V **Regularized Machine Learning in the Genetic Prediction of Complex Traits**.
Conducting the experiments, analyzing the results and drafting of the manuscript.

Contents

Chapter 1

Introduction

With the advent of modern technology, humanity is experiencing an exponential growth in the technology that affects our daily lives. Thanks to modern medicine and the accompanying research, people are living longer, more fulfilling lives. It is often taken for granted that 150 years ago, during the American Civil War, *modern medicine* meant the amputation of limbs, where the surgery was often more dangerous than the original wound. Since that time, medicine has become revolutionized in a way that no one could have ever imagined. The human genome has been mapped, affording us the opportunity to predict human disorders years before their associated symptoms manifest themselves. This is occurring due to the development of methods of identifying the biological sources of various disorders while requiring a minimal amount of time in the laboratory.

This genetic revolution started in the early 1900's when Thomas Hunt Morgan set up the *Fly* lab at Columbia University. By the 1920's he had determined that genes were actually heritable units and the DNA was located on chromosomes. Refining the genetic model, in 1953 James Watson and Francis Crick first suggested the double-helix model of DNA [93]. From this point forward, the way that biological research is conducted has been developing at an exponential rate, eventually resulting in efficient whole-genome sequencing. Through this expanding technology, coupled with the approximately 3 billion base pairs in the human genome [40], a scenario has emerged in which researchers are unable to exhaustively search the entire span of the genetic variant subsets and are faced with the conundrum of how to effectively analyze this data.

Making the simple assumption that 99.9% of the human genome is shared among any two individuals, this still leaves millions of distinct data points. If one accounts for insertions and deletions in the human genome the rate of variation can be even greater. There exists a wide acceptance of the fact that common disorders are influenced through the differences that are

observed in genetic variants [13, 47, 75]. Aggregation of this information has become invaluable in the modern research on the influence of genetic variations on various diseases. Modern day computational researchers are mining this data in order to find links between the genetic variants and the onset of various disorders and disease phenotypes [18, 59, 84, 96].

This data aggregation, as it is now being analyzed, is commonly constructed into various studies and cohorts that are known as genome-wide association studies (GWAS). GWAS can be defined as a study composed of common genetic variants, collected together and paired with various disease outputs in an attempt to generate a correlation between variant subsets and the phenotypic output [72]. As a result of the high levels of similarity between two individuals, GWAS studies have aimed at drastically decreasing the number of base-pairs that must be subsequently analyzed through the primary scanning of common variants with small effect sizes.

A number of feature selection methodologies exist which can be capable of addressing the problem of identifying meaningful feature subsets in GWAS [27]. These methods work through the identification of those variants which when selected both individually and through epistasis interactions, provide informative feature subsets and high predictive performance. Variant subsets discovered through feature selection can extend beyond those which are univariately significant and can therefore reveal hidden interactions between multiple variants. As this tends to be implemented through automated processes, it allows for the efficient analysis of millions of features and their associated interactions in a finite period of time. Post-selection, through the analysis of the results and their associated molecular pathways, researchers are hoping to improve upon modern medical treatments and eventually extend to personalized medicine.

1.1 Challenges in Genome-Wide Analyses

GWAS have become prominent partially due to the progression of technology which has resulted in a reduction in genotyping costs. Further, a growth in the scale of GWAS is clearly evident, with next-generation sequencing studies now containing millions of genetic variants. This is a far cry from the linkage studies that started in the 1980's and contained only hundreds of candidate genes [72]. GWAS are based on genotyping that is done via the use of SNP arrays that typically genotype approximately 500,000 SNPs for thousands of samples.

The SNP arrays are sourced from various technologies, two of the leading ones being the Affymetrix and Illumina SNP arrays. These arrays are not perfect and their quality is typically defined by the call rate which measures the fraction of genotypes that are reported. Typically, if an array's call

rate fails to surpass a predefined quality control threshold the genotyping will be repeated [14]. The genotypes are then called using various scoring algorithms, such as the Bayesian Robust Linear Model with Mahalanobis distance classifier (BRLMM) and CHIAMO algorithms [14]. Quality control procedures are then run to remove SNPs/samples with too high missing rates, SNPs that depart from Hardy-Weinberg equilibrium and other metrics.

Imputation of GWAS studies is defined as the process of predicting genotypes that are not observed or have been marked as missing in the original study [53]. The imputation process is done through the use of a high-density reference panel of haplotypes such as the 1000 Genomes Project [1] which is used for determining the omitted variants. In silico experiments are the run on these newer data sets which provide a number of advantages to researchers such as: increased power to identify causal variants, high-resolution mapping and the ability to combine studies genotyped with different assays. Population stratification can be further used to standardize the array data for the subsequent analyses [71, 98].

A number of key challenges exist in using GWAS data to study complex phenotypes. These include: the need for learning algorithms that can incorporate both SNPs and conventional risk factors into singular models, the selection of complex variant subsets that are explanatory of the output and the biological interpretation of the resulting models [55]. While these challenges represent the modeling difficulties that are faced when developing models, additional problems are experienced in GWAS due to their limited coverage of the human genome and the interpretation of associated variants that are non-coding.

A particular drawback to GWAS has been their reliance on genotyping common variants which may not hold the power to reveal the hidden heritability in many complex diseases. As a result, a new generation of studies, referred to as Next-Generation Sequencing (NGS) are being developed to take advantage of rare variants that were once thought to lack an association with the disease phenotypes [103]. While these studies are still emerging from their infancy, researchers are starting to explore ways to improve upon the results of standard GWAS to help explain a larger proportion of the variance for these diseases [76]. GWAS have only revealed a proportion of the missing heritability of many disorders and have started to be considered as not robust enough to explain this deficit [22, 81].

NGS studies are eliminating the need to base results off of partial information allowing for more complete analysis between genetic variants and the associated diseases [35]. GWAS screen common variants that are often in intronic chromosomal regions. In contrast NGS offers the ability to look more closely at exonic regions, which more often code for functional proteins. Researchers are thus starting to place a larger emphasis on NGS

studies and have encountered promising results such as new disease-causing mutations [8, 17, 44, 54, 76]. In the case of Belloni et al, the authors were able to uncover a lung cancer mutational profile that could only be revealed using NGS technologies [8]. Results such as these are proving that NGS technologies are helping to bring the goal of personalized medicine closer to reality.

This growth of technology and the associated information boom has generated a new problem, one often referred to as the curse of dimensionality. With too much data at the disposal of analysts, the question remains; how is one supposed to decipher this information in an efficient and intelligent manner? The answer to this would help to develop solutions for some highly complex problems that mankind has even been faced with such as: personalized medicine, disease prediction and medical genetics. Dealing with modern data sets that often range from billions to tens of billions of data points, modern medicine has reached a stage in which it is no longer feasible to conduct physical experiments for all hypotheses and a reliance on technology has become both a necessity and accepted part of the research process. Through the automation that is afforded through the use of machine learning, researchers can help to model a hypothesis' outcome and then only test those that are determined to maximize the likelihood of success.

A modeling challenge when analyzing genome-wide data sets refers to the process of deciphering the complex interactions within the human genome, determining the relationship among these seemingly disjoint genetic variants and the utilization of this information to attempt to explain the missing heritability that exists in many diseases [50, 100]. Interactions among the genetic variants, known as epistasis, are essential to successfully generating accurate models of complex diseases [59, 95]. However, the detection of these interactions remains a difficult task due to the large number of feature combinations that exists in a GWAS. Running an exhaustive search on large data sets can be prohibitive for multi-variant searches as the problem has an exponential number of subsets to examine [95].

Epistasis can be the result of more than two genetic variants, which can create a computationally prohibitive problem. Additionally, a statistical power problem exists in epistasis analyses. Due to the small effect size that these interactions exhibit, their contribution on the outcome phenotype can often be over-shadowed in traditional analyses that tend to make use of only the most significant feature sets. To account for epistasis interactions it is therefore necessary to look for selection algorithms capable of handling such effects in their models. When reporting the algorithm results, the measurement scale on which they are reported can have a significant impact [97] and it is therefore necessary to provide results based on metrics that can minimize the distortion of the results based on effects such as unbalanced classes.

In this thesis, the primary focus is on computational methodologies that assist researchers in making sense of these massive data sets, often in the form of genome-wide association studies, through the use of feature selection, machine learning and model validation. During this process, emphasis was placed on the scalability of the algorithms with respect to all dimensions such as the number of features, samples and selected features. Their ability to make efficient use of modern computing resources helps to ensure that they can scale to the next-generation sequencing studies which can contain millions of rare variants.

1.2 Aims of the Thesis

The overall aims of this thesis are to present both the computational feasibility of running advanced machine learning based, feature selection algorithms on GWAS and to provide a procedural workflow in both an efficient and scalable manner that would prove to be implementable by most research groups, regardless of their resources. Specifically, the main focus is on the implementation of wrapper feature selection methodologies on large scale data sets, in a field where they are seldom used due to their traditionally high computational complexities.

By implementing these complex algorithms on large scale data sets, the aim was to show that not only were these algorithms feasible, but they were able to provide new, biomedical results. These results would warrant their use alongside more traditional techniques of analyzing data sets of genome-wide scale, such as those methods based on univariate statistics. Due to the high computational complexities of the analyzed techniques versus these more traditional implementations, it was necessary to spend time addressing the adaptation of the algorithms to scalable, distributed systems, demonstrating that if the correct resources are available, the methods will scale even to the next-generation of genome analysis studies, assuring the longevity of the work presented here.

In addition to algorithm development and implementation, another crucial part of this thesis was to publish guidelines in peer-reviewed journals to advocate to researchers the reasoning and necessity of exploring algorithms that often falls out of the reach of general purpose analysis toolkits. A common problem observed in publications is that they only make use of methods available in existing implementations, regardless of their suitability for data sets on a GWAS scale. Such use, when not on a pilot study, has the potential to set back the field as it is easy to report that a non-optimally implemented machine learning implementation is not suitable for use on large scale studies. These applications may ignore the fact that there may be existing fields of research in which other groups may have already demon-

strated that they have adapted similar methods for problems of this type. Due to this, it is only through the successful pairing of computer scientists, medical researchers, bioinformaticians and mathematicians that will we be able to solve these problems.

Utilizing only research from key fields will often lead to a lack of the complete knowledge necessary to solve these problems. For that reason, this thesis was conducted as a cross-research center effort with participating researchers from the Department of Information Technology, University of Turku, the Institute for Molecular Medicine Finland (FIMM) and the Cardiovascular Risk in Young Finns Study (YFS). It was through the guidance of researchers in these various institutions that we were able to generate a set of methods that would aim to satisfy the requirements for a durable algorithm that would hopefully prove useful to researchers who continue the work after my completion with the work on the study.

Further, an overall theme of the work is that there is no universal method for determining the variants that are the most associated with a particular disorder. As seen in Publication V [61], depending on the heritability of a particular condition, different methodologies may present alternating results. While wrappers are in a class of more computationally complex methodologies, they have the potential to perform well (see e.g [55]). Their use can result in overfitting and in highly genetic diseases may actually underperform when compared to some more traditional two step implementations as demonstrated in Publication V. This is not to state that their use should be ignored in real-world applications. As such, it is demonstrated that it is not sufficient for researchers to ignore this class as too expensive to run, and rather advocate the use of more in-depth algorithm comparisons before making declarations on the features associated with particular diseases.

The work in Publication I [59] was done under the guidance of the Data Mining and Modeling Group at the Turku Centre for Biotechnology in collaboration with the YFS. All other publications were conducted with the Algorithmics and Computational Intelligence Group at the Department of Information Technology, University of Turku.

The specific aims of the thesis are:

1. Variant Selection Nearly all of the work done in this thesis was focused on the selection of genetic variants that are able to distinguish between the various phenotypic output classes present in the implemented studies. These feature selection methodologies were based on various techniques including: filters, wrappers, embedded and hybrid methodologies. Through the use of these assorted selection algorithms, it was aimed to demonstrate the ability to efficiently and effectively identify those genetic variants that when their

synergistic interactions with other polymorphisms were accounted for, could better predict the class outcome than basic univariate methods.

For the various selection methods, a number of algorithms were explored. These included but were not limited to: Information-Gain combined with a Naive Bayes based wrapper (see Publication I), greedy Regularized Least-Squares (RLS) based wrappers (see Publications II [67], III [58], V and [60, 64]), and various embedded based methodologies, including Lasso and Elastic Net (see Publication V). These methods exhibited a technological progression from a pilot study in Publication I in which widely available tools were used to those that required custom development to efficiently implement computational shortcuts and caching. By following this progression, the evolution of these methods in relation to one another can be observed. Additionally, their continued ability to be implemented on larger studies throughout the course of research demonstrates their scalability.

Considerable resources were spent towards the analysis, implementation and advancement of the greedy RLS algorithm. The reasoning behind this decision was that as a base it proved to be a highly efficient, yet underutilized algorithm which was able to provide results that are comparable to some of the leading methods. This performance was achieved in a fraction of the time of comparable methods and provided interesting variant subsets in the results. Due to these attributes, it was decided that this would be the algorithm that would be implemented on large-scale, distributed systems through the use of both OpenMP and MPI based implementations [58, 60]. In contrast to scaling filter-based feature selection methodologies, in which each processor can calculate independently of the other processors in the system, the use of a method that required caching and computational shortcuts required applying significant attention to the network communications.

2. Machine Learning The use of machine learning in the context of genetic association studies has remained relatively limited until recent years. While variant selection can be implemented through a multitude of methodologies, the aim here was to primarily make use of machine learning algorithms due to their ability to detect unknown variant interactions. Through the automated training process that is wrapped around these techniques during the *variable selection* phase, machine learning provided the ability to conduct an automated training process in which it was possible to rescale the effects of the various variable coefficients depending on the data present. Due to the size of GWAS, this is an increasingly important step to developing these models; since only looking at candidate gene sets has the adverse effect of potentially not allowing the algorithm to maximize the percentage of the variance that could be explained by the genotypes [67].

The implemented algorithms took on a number of forms that range from straightforward methods that simply maintain the ability to generate an

outcome prediction, to advanced methods that are capable of incorporating feature selection within the learning model itself. This analysis on a variety of techniques affords the ability to analyze the effect of variable selection and whether there was an added benefit to various methodologies. It was through the use of machine learning algorithms that we were able to implement the algorithms capable of generating predictive models based on GWAS data.

3. Model Validation Determining whether the generated model, from either a machine learning methodology, or that combined with a feature selection result will yield predictive results on an independent data set is vital for the verification of the results. Due to the expense of generating GWAS and the differences in their development, it is often difficult to obtain independent data sets for model verification. Therefore, a significant effort was paid towards the implementation of model-validation methodologies such as nested cross-validation that would allow for more realistic estimates of the real-world predictive capabilities of the developed models.

As producing these forms of model validation can significantly increase the required computation time, algorithms were commonly utilized based on their capacity for the efficient calculation of parts of the validation internally. An example of this is the case of greedy RLS, which performs an internal leave-one-out cross-validation while selecting the features. By efficiently computing this in a single iteration it is possible to reduce the computational overhead of the feature selection, making them more feasible on large-scale data sets.

4. Scalability Significant effort was paid towards the scalability of the algorithms. Developing algorithms which are only capable of being scaled to the current GWAS would result in the method quickly becoming outdated as the scale of these studies continues to increase. Further, it can currently be observed in two-step methodologies which aim to reduce computational costs by coupling advanced model selection methods with faster univariate ones, that their use can potentially result in a loss of information. This can lead to a failure to maximize the amount of the heritability or predictability which can be explained by the genetic variants. More computationally complex algorithms may provide alternative feature subsets that can potentially explain some of this heritability, but require additional attention to assure that they will scale to the GWAS and beyond. It is vital that new algorithms being released to the GWAS community are capable of scaling to the next-generation sequencing studies.

Adapting greedy RLS to distributed computers yielded results that helped to confirm the algorithms capability to scale to larger studies, with its test application being limited largely by the resources available. Further focus was paid to its scalability in all dimensions, regardless of whether the new

data set had an increased number of features, samples or selecting a larger subset of features. This assured its suitability for a wide variety of use cases. It was also important to demonstrate that before scaling the algorithm to a parallel environment, every effort was made to minimize the running times on a serial machine. The optimization included both computational and memory efficient variation that allowed users to determine the method that best works given their available resources.

1.3 Organization of the Thesis

This thesis' main contribution to the academic community is in the five original publications that were published previously in various journals and and conference proceedings. Chapter 2 provides an overview of the machine learning methodologies that were analyzed and implemented during the research for this thesis. Additionally, it discusses feature selection methodologies, their implementations and applications to genetic studies. Further, it briefly discusses some model-validation techniques. Chapter 3 provides information on further scalability advances such as those based on parallel computing. Chapter 4 summarizes the publications that make up this dissertation. Finally, Chapter 5 provides a conclusion to thesis and a brief insight into potential future applications.

Chapter 2

Methods for Analyzing GWAS data

2.1 Algorithmic Background

The ability to collect massive amounts of data has created an expanding need for algorithms that can handle such studies. With cohorts now containing millions of genetic variants and thousands of individuals, analysts are being faced with the challenge of processing the data sets. While this does not pose significant problems for straightforward techniques such as univariate statistics which can be run in linear time, the use of advanced multivariate algorithms is easily hindered when being implemented on studies of this magnitude.

Older studies have been primarily implemented on either smaller sample sizes, have been pre-filtered for variants that are the most statistically significant with the outcome variable or consist of prior identified loci [21, 96]. In recent years, researchers have been working on the development of new methods that are scalable to larger data sets [32, 67, 92]. The advent of NGS data sets will require methods capable of both scalability and high predictive performance.

2.2 Machine Learning

Machine learning (ML) is a field of study that aims to develop algorithms capable of learning from data without the need to be explicitly programmed. Acting as a cross-disciplinary field, it encompasses the work from mathematics, statistics and computer science. Acting as an automated methodology for learning, ML has allowed for a new realm of problem sets to become viable for analysis. Without the explicit need of programming the models, researchers can be assisted by computers in the development process.

This has allowed ML to be applied to a broad spectrum of projects. From Google's self-driving cars, to image recognition, commercial recommendation engines and even biomedicine, it is becoming rare to find a high-tech field that doesn't encompass at least some aspects of this discipline. Through image recognition, ML is helping cars to interpret whether the object in front of them is a person or vehicle and how to react accordingly. In businesses, recommendation engines are using methods such as collaborative filtering to determine which products and services would most interest a particular consumer, while in biomedicine researchers are able to analyze problem sets that only a couple of years ago were considered not computationally feasible.

With modern GWAS consisting of billions of data points, it remains an impossible task to explicitly define a set of rules that can explain the interrelationships between the genetic variants. Rather researchers are relying on classification, regression and feature selection methodologies to help manage the vast data sets and determine how to handle the individual variants in terms of complex problems. This automated process aims to make sense of the feature subsets, seek out both univariate and epistasis interactions along with helping to group the resulting features into potential subsets that may help researchers to analytically explain the results.

In GWAS, researchers typically deal with problems that can take on one of two states, either regression or classification. Regression can be defined as a statistical process of determining the relationship of an output to one or more inputs and noting how changes in these independent variables will cause subsequent changes in the output. Regression is widely used in studies where the outcome variable can be considered to be continuous, such as the case when considering blood pressure, an individual's weight, cholesterol levels, age, etc. Classification problems occur, for instance, when classifying subjects into discrete categories such as the disease or non-diseased classes based on their genetic profiles

2.2.1 Notation

The articles leading up to this dissertation are based primarily on determining the dichotomous class through the use of machine learning. This form of prediction, in terms of GWAS can be formalized as follows. Let m and n be the total number of individuals and features respectively. In supervised learning the problem contains a set of input-output data pairs known as the training set:

$$S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\} \tag{2.1}$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^n$ and $y_1, \ldots, y_m \in \mathbb{R}$. In addition we define $\boldsymbol{y} = (y_1, \ldots, y_m)^{\mathrm{T}}$, $\boldsymbol{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_m)^{\mathrm{T}} \in \mathbb{R}^{m \times n}$. In the above $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,n})^{\mathrm{T}}$

and $i \in 1, \ldots, m$. It is assumed that the training set is composed of independent identically distributed (i.i.d) samples drawn from the same unknown distribution of the data.

In supervised learning the goal is to infer a prediction function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ such that f can be applied to unseen data. A training algorithm is defined as a mapping from the training set to the hypothesis set.

$$\mathbb{A} : (\mathbb{R}^n \times \mathbb{R})^m \rightarrow \mathcal{F} \tag{2.2}$$

where $\mathcal{F}$ is the set of possible prediction functions. The algorithm in (2.2) is a mapping from all possible training sets of all possible sizes to the set of prediction functions. For relevance to the work that comprises this dissertation, $\mathcal{F}$ is made up primarily of linear models.

2.2.2 Classifiers

When trying to use a classifier for predicting discrete output, the output values are taken from a finite ordered set $C = \{c_1, \ldots, c_k\}$, which can for example represent different stages in disease progression. In the literature this is referred to as ordinal classification [5]. A special case of this is when there exists two classes, known as binary classification. This is typical when one is trying to determine whether an individual is either a case or a control for a particular disease phenotype. In this case we denote $y_i \in \{-1, 1\}$ where -1 indicates subjects who have a negative outcome status and 1 represents individuals who have a positive one.

In classification one can consider probabilistic models where the probability of y being categorized in a particular class is assigned a particular value. For example, the probabilities of an instance being classified as a case or a control can take on the probabilities $P(1|\mathbf{x})$ and $P(-1|\mathbf{x})$ respectively. To determine the classification class, one method of adjusting the prediction functions output to a discrete value is to apply a scoring function to convert the score to a corresponding class.

$$f(\mathbf{x}) = \begin{cases} 1 & P(1|\mathbf{x}) > P(-1|\mathbf{x}) \\ -1 & \text{otherwise} \end{cases} \tag{2.3}$$

The model in (2.3) can be interpreted as if the probability of an instance being classified as a case is greater than the probability of an instance being classified as a control then classify as 1, otherwise it should be -1.

Naïve Bayes

While not a complex predictive function like the others utilized in this dissertation, Naïve Bayes, is described here due to its implementation in Publication I. Naïve Bayes is a classifier assuming feature independence when

conditioned on the class. Some of its main advantages are its simplicity and computational efficiency which leads to its scalability to large data sets.

A Bayes predictor would generate the class predictions based on:

$$\hat{y} = \arg\max_{c \in C} P(c|\mathbf{x}) \tag{2.4}$$

where $P(c|\mathbf{x})$ is the posterior probability of a class c, given a feature vector $\mathbf{x}$. $P(c|\mathbf{x})$ can be computed with the Bayes Theorem:

$$P(c|\mathbf{x}) = \frac{P(\mathbf{x}|c)P(c)}{P(\mathbf{x})} \tag{2.5}$$

Here, $P(\mathbf{x})$ does not affect (2.4) and hence it can be ignored. $P(c)$ is the prior probability of the class c based on the training data, e.g. $P(c) = \frac{|M_c|}{m}$, where $M_c = \{i|y_i = c\}$. In Naïve Bayes it is assumed that the features are mutually independent given the class. Therefore, $P(\mathbf{x}|c) = \Pi_{i=1}^{n} P(x_i|c)$. If the feature values are binary then $P(x_i|c)$ is the number of times the feature value x_i has appeared in the training data belonging to class c divided by the total number of training points in class c.

To determine which class to classify a particular instance, Naïve Bayes makes use of the maximum likelihood for the available classes which can be calculated through the application of Equation 2.6.

$$\hat{y} = \arg\max_{c \in C} P(c)\Pi_{j=1}^{n} P(x_j|c) \tag{2.6}$$

where x_j refers to the j^{th} feature of the data point $\boldsymbol{x}$.

If a feature x is continuous, a commonly used approach is to assume it Gaussianity and to estimate the corresponding density function via [33].

$$P(x_j|c) = \frac{1}{\sqrt{2\pi}\sigma_{j,c}} e^{-\frac{(x_j - \mu_{j,c})^2}{2\sigma_{j,c}^2}} \tag{2.7}$$

where $\mu_{j,c}$ and $\sigma_{j,c}^2$ are the mean and the variance of j^{th} feature of the training set in class c:

$$\mu_{j,c} = \frac{1}{|M_c|} \sum_{i \in M_c} x_{i,j} \tag{2.8}$$

$$\sigma_{j,c}^2 = \frac{1}{|M_c| - 1} \sum_{i \in M_c} (x_{i,j} - \mu_{j,c})^2 \tag{2.9}$$

Equations (2.8) and (2.9) can now be used to define the training algorithm $\mathbb{A}$ in (2.2).

Some of the main advantages of Naïve Bayes are its scalability and relatively good real-world performance, despite the simplicity of the algorithm

[52, 59, 82, 94]. These characteristics lead it to remain a widely explored algorithm that will have a strong potential to be applied to next-generation sequencing studies in which the scalability of algorithms may become a predominant factor in their selection.

2.2.3 Regression

When dealing with $\boldsymbol{y} \in \mathbb{R}^m$ such that y is continuous, the problem of predicting the output y_i, based on the input $\boldsymbol{x}_i$ is know as regression. The relationship between the input and output is generalized by equation 2.10:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon} \tag{2.10}$$

in which $\boldsymbol{w}$ are the coefficients assigned to $x_1, \ldots, x_n$ and the goal is to minimize the error $\boldsymbol{\epsilon}$ such that $\boldsymbol{y} \approx \boldsymbol{X}\boldsymbol{w}$.

To solve for this, the first step is to define an objective function as seen in (2.11)

$$\|\boldsymbol{\epsilon}\|_2^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \tag{2.11}$$

$$= \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y} - 2\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{w} \tag{2.12}$$

The minimum of $\|\boldsymbol{\epsilon}\|_2^2$ can be found from the zero point of the derivative. The derivative with respect to $\boldsymbol{w}$ is:

$$\frac{\partial}{\partial \boldsymbol{w}}\|\boldsymbol{\epsilon}\|_2^2 = -2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} + 2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{w}. \tag{2.13}$$

Setting the derivative $\frac{\partial}{\partial \boldsymbol{w}}\|\boldsymbol{\epsilon}\|_2^2 = 0$ and solving for (2.13) with the assumption of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ being full rank one gets

$$\boldsymbol{w} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{2.14}$$

2.2.4 Regularization

The matrix $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})$ in 2.14 can become singular when there is a linear dependence between the features. This makes it challenging to compute $\boldsymbol{w}$ since the inverse of a singular matrix cannot be calculated. As (2.14) becomes closer to singular, its sensitivity to random errors increases, which results in an increase in the variance. This process of overfitting results in the need to help reduce the overly-optimistic effect that can occur when fitting models to training data.

To help account for this, users can apply regularization to the regression model. Regularization applies a complexity factor to the loss penalty, resulting in larger penalizations for large coefficient values. By penalizing these

coefficients, regularization aims to reduce overfitting and helps to develop an algorithm that will better generalize to unseen data. Least squares can be regularized in several different ways. The most common example is the squared Eucledian norm of the vector $||\boldsymbol{w}||_2^2$ which in the literature is known as the ridge regularizer. The regularized least-squares problem becomes:

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}\in\mathbb{R}^n}{\arg\min}\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) + \lambda||\boldsymbol{w}||_2^2\} \tag{2.15}$$

such that the regularization parameter $\lambda > 0$ and

$$\|\boldsymbol{w}\|_2^2 = \sum_{j=1}^{n} w_j^2 \tag{2.16}$$

Referring back to the matrix algebra solution presented for regression, this can extended to apply to RLS through the closed-form solution presented in equation 2.12.

$$\boldsymbol{w} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\mathbf{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{2.17}$$

Through the regularization applied by λ, regression has increased capabilities for handling multi-collinearity which is a primary concern in GWAS. It is important to note that there exist other regularizers such as the 1-norm which is discussed further in the feature selection subsections.

2.3 Scoring Metrics

A scoring metric is a way of measuring the ability of an algorithm to predict the output based on its input. Identification of the appropriate scoring metric for a particular problem set is an important task when evaluating feature selection methodologies [97]. A number of scoring metrics are commonly applied in learning studies. The final selection is dependent upon the particular problem which is being analyzed. Regression and classification problems often make use of different metrics as the structure of the output has an affect on the appropriate technique. As an example, having single-class output predictions can result in a very high accuracy but still be a poor model. For this reason it is necessary to select models which are suitable for the problem set.

Certain measures have the propensity to misrepresent the predictive power of the model. As an example, a non-stratified data set with 90% controls would be trivial to achieve a high accuracy. Rather, measures that can account for class size differences, such as the Area Under the Receiver Operating Characteristic Curve (AUC) would better represent the actual performance of the model.

Table 2.1 provides a brief overview of various scoring metrics. These metrics are a partial list of those that are commonly applied in research. For a more detailed description of many of these methods one can view the supplementary materials of Publication V.

Metric	Algorithm
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
AUC	$\frac{1}{\lvert m^+\rvert\lvert m^-\rvert}\sum_{j\in m^+}\sum_{k\in m^-} g(\hat{y}_j-\hat{y}_k)$
MSE	$\frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}{m}$
COD	$1-\frac{\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{m}(y_i-\bar{y})^2}$
χ^2	$\sum\frac{(O-E)^2}{E}$
Odds Ratio	$\frac{TP\times TN}{FP\times FN}$
Fisher's Exact Test	$\frac{(TP+FN)!(FP+TN)!(TP+FP)!(FN+TN)!}{TP!FP!FN!TN!(TP+FP+FN+TN)!}$

Table 2.1: In this table TP, TN, FP and FN represent the number of true positives, true negatives, false positives and false negatives respectively. Here, $g(x)$ is the Heaviside step function. O and E are the entries from the observed and expected frequency tables. AUC is the area under the receiver operating characteristic curve, MSE is the mean squared error and COD is the coefficient of determination. m^+ and m^- are the index sets of the data points whose correct output is positive and negative respectively. $\hat{y}$ and $\bar{y}$ are the predicted and average values of y respectively.

2.4 Model Validation

Recent GWAS, such as those based on schizophrenia have demonstrated great strides in the ability to identify over a third of the heritability of the disease in the primary cohort, but the results did not generalize well on a secondary, independent cohort [24]. Predictive models are often capable of achieving good results on the data they were trained with, but the question remains: how does this model perform on independent data that would be reminiscent of real-world conditions? To address this question one can apply so-called model validation techniques. The two main validation methods examined here are the use of independent data sets and cross-validation. These methods are not mutually exclusive and it is not uncommon to see both of these methods combined with one another to simultaneously perform both model selection and validation. For example, cross-validation is often used for feature and/or parameter/model selection, while the use of independent

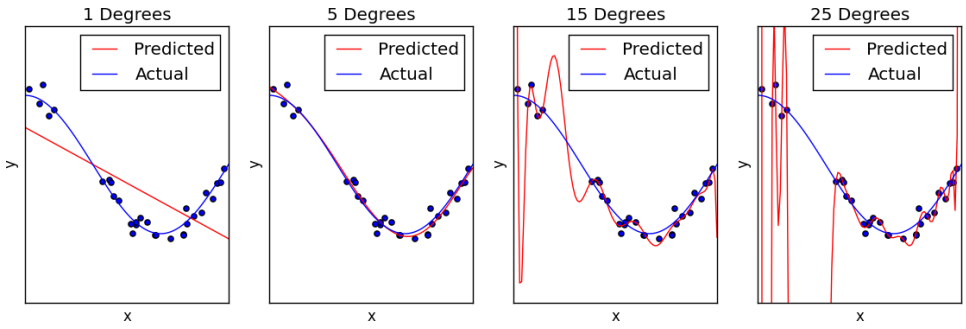

Figure 2.1: An example of how a set of points can be overfit by increasing the degree of the polynomial. In this example, 30 points are generated along a randomly chosen function. A polynomial is then fit to these points using various degrees. It can be observed that as the number of degrees approaches the number of random points the curve demonstrates overfitting. This is a modification of the sample provided by Scikit on their website[1].

data sets are used at the end of the selection process to test whether the model is overfit to the training data. As described later in this section, we chose to omit the use of bootstrapping in our publications.

An example of model overfitting can be seen in Figure 2.1, consisting of a random set of 30 points from a polynomial plus random noise. A polynomial model has been fit to these points, along with showing their true function for varying numbers of degrees. The higher degree of the polynomial that is used to fit the model, the smaller the error is. However, analyzing the fitted model, it is apparent that the higher degree models would poorly generalize to unseen data.

The use of independent data sets is trivial, but is important in determining whether the model has been overfit to the training data, meaning that it will not likely generalize to unseen data. The concept behind this method is to use a set of data that has not been examined during the model construction and apply the trained prediction function from the training set onto the independent data set. The prediction results of the generalization on this new set are then recorded and evaluated. The main advantage to this method is that it makes use of data that has had no influence on the training of the model and thus can be assumed to provide a realistic estimate of the model's performance.

The selection of independent data sets can seem trivial, but it is important to consider that population demographics may partially determine the

[1] `http://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html`

performance on this data [86]. For example, comparing a study that comes from an Asian cohort to a European cohort can reveal underlying genetic differences that can reduce the reproducibility of the results. For this reason, it is common for researchers to examine the population stratification within the studies to assure that similar clusters are being compared. This is meant to help assure that population differences will have minimal impact on the final model.

Due to a limited number of samples in GWAS, it is also possible for researchers to implement a hold-out set method in which a percentage of the data set is removed before the model development and only used during the testing phase. This method has the advantage of allowing researchers to have a higher degree of confidence in the similarity of populations being examined, though it is more prone to being affected by any experimental errors that may exist in the data set. Additionally, as a portion of the population is being removed, there is left a more limited data set in which to train the model, hence affecting its generalizability. If possible it is recommended to make use of independent data for model validation.

When developing models whose predictive performance needs to be assessed, the most straightforward model validation implementation to help alleviate overfitting is the use of training, validation and test sets (see Figure 2.2). In this model the data is initially split into separate subsets, one is used for the training of the model and a validation set which is used to optimize the performance of the training model. The percentage split that is used to partition the data varies, though it is most common that the training set is significantly larger than the validation set. Some common splits include a 67%/33% and a 90%/10% split.

The test set is used for measuring how well the model generalizes to data that is not observed during the training phase. This testing set ideally comes from a data set that is external to the data set implemented in the training/validation stages of the algorithm fitting. In GWAS this would be a study conducted externally to the one from which the training set originates. In the absence of external validation data, the original sample can be used to generate the testing set, so long as this data is extracted prior to any model fitting.

Bootstrapping is another validation method that is commonly used in machine learning problems. It is primarily used for data sets in which the sample size is small. It makes use of uniform random sampling with replacement of the training data to increase the amount of data available. This provides a relatively accurate estimate of the sampling distribution for the algorithm being analyzed. Bootstrapping has the advantage of commonly displaying a lower variance, but has a higher bias when compared to cross-validation [19].

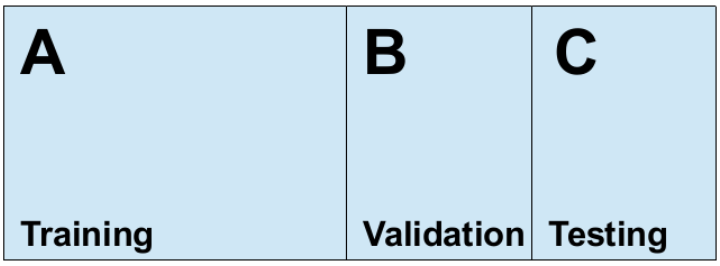

Figure 2.2: An example of splitting a singular data set up into a training, validation and testing data set. In the figure A represents the training data set, B is the validation set and C is the testing set.

As stated by Kohavi in [37], bootstrapping has the significant risk of failing if the learning algorithm tends to memorize the training data such as can potentially happen in the nearest neighbor-based methods. This is due to the memorizer being able to remember the training data and then applying this to a bootstrapped sample which can contain some of the same training instances [37]. Therefore, the performance of bootstrapping can be highly dependent on both the data and learning algorithm selected. As a result, it is commonly used for measuring the uncertainty of a fixed model while cross-validation is more often used for model selection. Since the contained publications focused primarily on model selection, bootstrap validation techniques were not explored.

2.4.1 Cross-Validation

Cross-validation can be described as the following [29]: Given a data set, we split it up into k parts, training the model on $k-1$ parts while the remaining split is used as a testing set. In other words each of these folds are used a single time as a test set, while the other $k-1$ folds are combined together to form a training set. This can interpreted as an indexing function $\kappa : \{1, \ldots, m\} \mapsto \{1, \ldots, k\}$. Next, a fitted function $\hat{f}^{-\kappa(h)}(x)$ can be defined as the function with the h^{th} fold removed from the data. The results of the k models are then averaged together to generate an overall score for the model.

$$CV = \frac{1}{k} \sum_{h=1}^{k} S\Big((\boldsymbol{y}_i)_{i \in K_h}, (\hat{f}^{-\kappa(h)}(\boldsymbol{x}_i))_{i \in K_h} \Big) \tag{2.18}$$

Where S is a scoring metric and K_h is the set of indicies contained in the h^{th} fold. Here $(\boldsymbol{y}_i)_{i \in K_h}$ is a vector of outputs in h^{th} fold. Further, in cross-validation researchers generally deal with two types of splits, stratified and

unstratified. In the case of stratified cross-validation, the splits are arranged in such a fashion so that the ratio of each class of the output in each fold is approximately equal to the ratios that were present in the original data set.

Common implementations of k-fold cross validation are the 10-fold and m folds, where m is the number of examples in the data set. The former is commonly implemented due to its relatively low computational overhead requirements as the model only needs to be trained k times. In the latter, commonly known as leave-one-out cross-validation (LOOCV), when dealing with relatively high sample size data sets a computational limitation is confronted in which the problem may not be feasible to run that many times.

An issue with LOOCV is that due to only a single sample being removed at each step, there exists a very high correlation between the different folds of the data set. As it is generally accepted that averaging the performance of many highly correlated models will result in a relatively high variance, it can be assumed that a LOOCV based estimate will have a higher variance when compared with that of a traditional k-fold cross-validation. This does not mean that the method should be ignored in the case of larger studies. Fortunately, there exists ML methods that are capable of producing an exact value for the LOOCV model through only a single iteration by implementing computational shortcuts [67, 68]. These shortcuts are what made it possible to implement wrapper style feature selection on entire genome-wide association studies.

Nested Cross-Validation

Selection bias [3] often exhibits itself when the cross-validation that is used for the model building is the same CV that is used for calculating the error estimate. To help avoid this pitfall it is recommended to use nested cross-validation. Through this, a relatively unbiased estimate of the actual error of the final model trained with the whole data set can be established [3, 90].

In order for CV to provide an unbiased estimate of the final model, it is necessary that each of the learning phases including feature and parameter selection are done within an inner-cv loop. The inner-cv is performed on the training data during each round of the outer-cv by splitting it into k_2 sub-folds, where k_2 is the number of folds in the inner-cv. After evaluating the performance with nested-cv, this model selection done during the inner-cv can be performed over the entire data set to obtain a final model. An example of nested cross-validation can be seen in Figure 2.3. In this figure a traditional 3-fold cross-validation is being performed and, within each of the three folds, a further 3-fold internal cross-validation is being performed on the corresponding round's training set to select the model parameters.

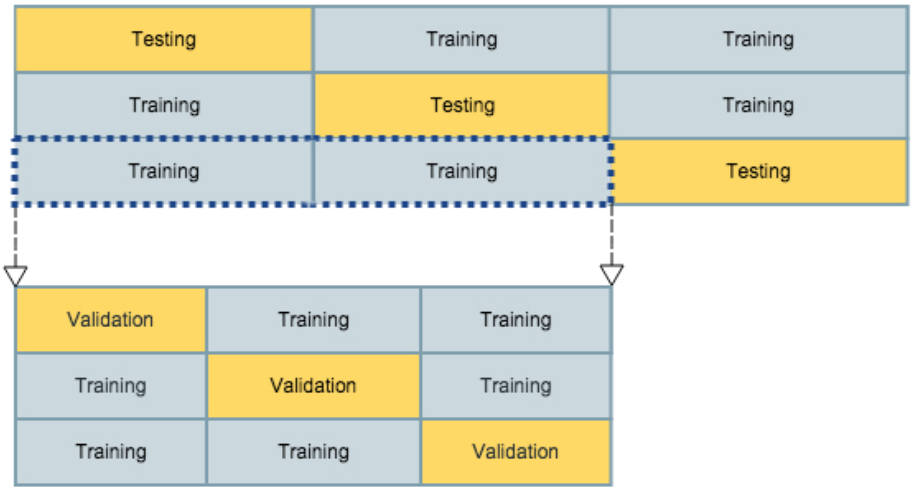

Figure 2.3: An example of 3-fold cross-validation with an extension to demonstrate a 3-fold nested cross-validation. In k-fold cross-validation each fold of the data is used for testing exactly once. When extended by nested cross-validation, each of the training sets has a subsequent cross-validation performed inside of it in order to select the model parameters/features. This nested cross-validation is done on all of the training data in the external fold.

Examples of nested-cv's use in GWAS can be seen in [36, 43, 61, 67]. The requirement of additional cv testing can lead to impractical computation times in studies where the cross-validation is expensive to compute [88]. This computational cost can increase even further as it is not uncommon for researchers to repeat the procedure multiple times [39]. One potential solution demonstrated in this thesis when dealing with data sets of scale such as GWAS and NGS, is the use of scalable algorithm that can be speed up through other techniques such as parallel computing (see Chapter 3).

2.5 Feature Selection

The information age has brought about a wealth of information that would have previously been unthinkable. In research of large-scale scientific and *Big Data* corporate data sets, researchers are clamoring to develop techniques to mine through the data to find the components that are both directly and indirectly related to the output. The understanding of the biological components of disorders is necessary for the development of new and effective treatments and therapies.

It has been shown in prior publications that features that do not necessarily pass the threshold for genome-wide significance, $p < 5 \times 10^{-8}$ [7], can be used to help boost the performance of classifiers [38, 59, 96]. This can lead to a speculation of whether filters are capable of capturing all of the information contained within the human genome [59, 67, 96]. It can be advantageous to also analyze whether the statistically significant features can be complemented by other variants to explain a larger proportion of the variability[59, 100, 101].

GWAS data creates a curse of dimensionality where $n \gg m$, which can result in overly-complex models that fails to generalize to unseen data. To efficiently determine the features which should be further examined, feature selection can be applied. It is through feature selection that a marriage between the methodologies of computational and biological researchers can occur and more efficient development of medical treatments can start to be produced.

Feature selection is the process of selecting those sets of variants which are the most predictive of a particular outcome variable. In the case of GWAS, this is primarily concerned with selecting the subset of variants that are most predictive of the outcome variable, commonly either a case/control qualitative phenotypes, or a quantitative one such as blood pressure.

Feature selection methods are commonly divided into three categories: filter, wrapper and embedded methods (see e.g. [27, 79]). While each of these methods has its own foundation it is very common for researchers to make use of various combinations to help account for the shortcomings of the various methods. However, not all approaches fall neatly into one of these categories, certain algorithms may be considered both wrapper and embedded methods depending on the viewpoint, and many approaches also combine several different selection methods. Of interest to this paper is a focus on algorithm development to start equalizing the playing field between the methodologies, allowing singular algorithms to eliminate shortcomings that were traditionally present.

In this work, the main division between the different approaches is established. Filters are primarily implemented by computing univariate test statistics for individual features, in order to evaluate their predictability for a particular phenotype. The approach is easy to implement, scales to large data sets and the results yield straightforward interpretations. However the resulting predictive performance can be sub-optimal, since the approach misses possible interactions between the features, and does not take into account the properties of the used learning algorithm. Wrapper and embedded methods allow addressing these problems, but at the cost of needing to implement much more complicated algorithms, whose scaling to large data set sizes is a challenging problem.

2.5.1 Filter Methods

Filter feature selection is one of the most widely used methodologies for determining the feature subsets for subsequent analysis and can act as both a standalone selection technique or combined with others to assist in the process of analyzing data sets of scale [59, 83, 96]. Traditional input-output relationships have been identified through the individual analysis of the feature set to identify those that are statistically associated with the output through the application of a univariate statistic, $\mathcal{H}$ (such as the mean squared error, see T able 2.1), as seen in Algorithm 1. Here, $\mathcal{H}$ is iteratively applied to the features-output pairs, $(\boldsymbol{X}_{:,1}, \boldsymbol{y}), \ldots, (\boldsymbol{X}_{:,n}, \boldsymbol{y})$, where $\boldsymbol{X}_{:,j}$ refers to the j^{th} column of the data matrix. They are examined individually to find the one that has the highest association (lowest p-value) with the outcome labels, $\boldsymbol{y}$. The features which are selected for subsequent analysis are based on either those features whose statistic surpass a particular threshold, the top k features, or a combination of the two aforementioned methods. Those features that are selected will then typically have either a learning algorithm applied to them to train a predictive model [96] or may have a subsequent feature selection run on them [59, 83].

Algorithm 1 Filter feature-selection

```
1: S ← ∅
2: for j ∈ {1, ..., n} do
3:     e_j ← H(X_{:,j}, y)
4:     if e_j < e then                ▷ Compare error to pre-defined threshold
5:         S ← S ∪ {j}
6: return S
```

Filters are highly scalable and can normally be done in $O(Dn)$ time where D is the complexity of computing $\mathcal{H}$. The speed of computing filters allows them to be applied to large data sets in a fraction of the time that it takes to complete wrapper and embedded methods. PLINK (see [73]) can calculate the genotypic p-values for all features in a GWAS in a matter of minutes. Their application can be further enhanced through the use of multi-processor computers so that the feature scores can be computed simultaneously. Due to these method's characteristic of the selection process being external to the applied learning model, the feature subset is easily interpretable.

While filters-methods have shown to provide optimal results in some studies in which there is known to be a strong statistical association between the variants and the outcome labels (see Publication V), their application is known to have the potential to fall short of their wrapper and embedded counterparts [55, 59, 67]. For this reason their analysis should be comple-

mented by more complex methodologies on a training set and only further applied if their results remain optimal in the experiment. As they tend to be based on univariate statistics, they often ignore the feature dependencies that commonly exist through epistasis interactions in GWAS. This false assumption of feature independence can lead to sub-optimal results [26]. Further, analyzing continuous-valued output with discrete inputs, applying univariate filter methods is a complex task. A common way to handle this can be to use algorithms such as those from the ReliefF family of methods [77].

2.5.2 Wrapper Methods

Wrapper methods search through the power set of features for the subset of variants which maximizes the estimated predictive power of the model. This method uses the learning algorithm itself to search for the feature subset, allowing the features to be selected based on how well they will work with this algorithm, rather than performing a selection based on potentially different criteria, as is the case with filter methods. In other words, the features selected are classifier dependent.

Wrappers have been shown to be advantageous compared to filters because of their ability to detect interactions between features that normally could not be identified through the use of univariate methods. Further, as has been shown in Guyon et al., a variable that provides no useful information on its own can provide information when taken into account with other features [27]. This means that epistasis interactions among SNPs may cause features that would be ignored during filter methods to create more suitable feature subsets.

While the ability to examine large numbers of feature subsets can afford the opportunity to maximize the performance of the learning algorithm, it comes at a cost, such as high complexities and the strong possibility of overfitting to the training data, potentially limiting the generalizability of the model. Moreover, wrappers do not make any assumption of prior knowledge to the final feature set and it is therefore possible that features which domain experts may be interested in are not included after the selection process. Additionally, while redundant features may be likely to be excluded from the final set (dependent on the algorithm used), this action may cause useful information regarding features that map to particular genes and pathways of interest to be lost during the selection process. When sample sizes are limited, redundancy in the training set does not definitively indicate redundancy in the test set. If certain features are required to be included in the final feature set, then alterations to the wrapper algorithm must be manually made to force their inclusion.

A standard wrapper algorithm works by combining three separate components. The first is the learning algorithm, around which the feature selection is wrapped. This can in principle be any classification algorithm. The method then uses a search algorithm to search over the varying feature subsets, by deciding which features will be considered by the current iteration as potential candidates. The fitness of these features is finally evaluated through the use of a heuristic which estimates how well the analyzed SNP subset is capable of predicting the correct phenotypes.

The use of wrappers requires the implementation of search heuristics, which guide the selection process through the feature space. Search algorithms are necessary for wrapper methods since even when dealing with a seemingly low number of genetic variants there are an extremely large number of feature subsets that can be analyzed. Given the current size of GWAS, often containing hundreds of thousands to millions of variants, and the exponential growth of exhaustive searches, this creates a computationally impossible problem to solve.

To analyze the most basic form of wrapper-methods one can start with a study of greedy forward selection. Algorithm 2 starts with $\mathcal{S} = \emptyset$ and iteratively selects one feature at a time until a predefined number of features, k, has been selected. During this process, new features are only added to the $\mathcal{S}$ and are never removed. In Algorithm 2 it can be observed that the outermost loop inserts an additional feature into $\mathcal{S}$, at each iteration until the subset contains a predetermined number of features, k. During the inner loop, the wrapper examines every variant that has not yet been selected and computes the value of the heuristic $\mathcal{H}$ for the prior-selected features combined with the new feature under consideration. Note that the heuristic is now defined for pairs of feature subsets rather than individual features. This can be represented by $\mathcal{H}(\boldsymbol{X}_{\mathcal{S}\cup\{i\}}, \boldsymbol{y})$, where $\mathcal{S} \cup \{i\}$ is the union of the currently selected features, $\mathcal{S}$, and a new feature i.

Algorithm 2 Greedy, forward feature selection

$\mathcal{S} \leftarrow \emptyset$
while $|\mathcal{S}| < k$ **do**
    $e \leftarrow \infty$
    $b \leftarrow 0$
    for $i \in \{1, \ldots, n\} \backslash \mathcal{S}$ **do**
        $e_i \leftarrow \mathcal{H}(\boldsymbol{X}_{\mathcal{S}\cup\{i\}}, \boldsymbol{y})$
        if $e_i < e$ **then**
            $e \leftarrow e_i$
            $b \leftarrow i$
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{b\}$
return $\mathcal{S}$

An implementation of Algorithm 2 has a complexity of $O(knT(k,m))$, where k is the number of SNPs to be selected, m is the number of subjects, n is the overall number of features and $T(k,m)$ is the time required to train $\mathcal{H}$ on k features and m training examples. While less efficient than filter methods, wrappers are possible to run on the GWAS scale, but in order to do so, they need to be coupled with computationally efficient classifiers and shortcuts [66, 67]. One such example is greedy RLS which incorporates a greedy forward search with leave-one-out cross-validation and regularized least-squares regression, also known as ridge regression [65, 67]. This method performs the feature selection in $O(kmn)$ time.

A problem with forward-selection is that the addition of new features may result in one or more of the prior selected features becoming redundant, but provides no means of eliminating these variants. The advantage of forward searches is that they are ideal in scenarios where the dimensionality of the feature set are high.

A backward selection starts with the feature set $\mathcal{S} = \{1, \ldots, n\}$, iteratively removing a single feature at a time until the criteria for ending the feature selection has been met (see Algorithm 3). While modifications can be made, such as stopping the search once a predefined number of features have been selected, here the simplest form of the algorithm is presented. The drawback of backward elimination is its slowness with large GWAS data sets. When using large feature sets, the algorithm is slow, as it must constantly be retrained using a large feature set until only k features remain, where $k \ll n$. On the other hand, backward selection has been shown to produce better results than forward selection [41].

Algorithm 3 Greedy, backward feature selection

$\mathcal{S} \leftarrow \{1, \ldots, n\}$
while $|\mathcal{S}| > k$ **do**
    $e \leftarrow \infty$
    $b \leftarrow 0$
    for $i \in \mathcal{S}$ **do**
        $e_i \leftarrow \mathcal{H}(\boldsymbol{X}_{\mathcal{S}\setminus\{i\}}, \boldsymbol{y})$
        if $e_i < e$ **then**
            $e \leftarrow e_i$
            $b \leftarrow i$
    $\mathcal{S} \leftarrow \mathcal{S}\setminus\{b\}$ ▷ Remove the feature whose removal leads to the best score
return $\mathcal{S}$

These two methods can be augmented to create alternative search methods that will help to alleviate some of the drawbacks of forward searches. Backtracking can be coupled with forward selection to allow for the algo-

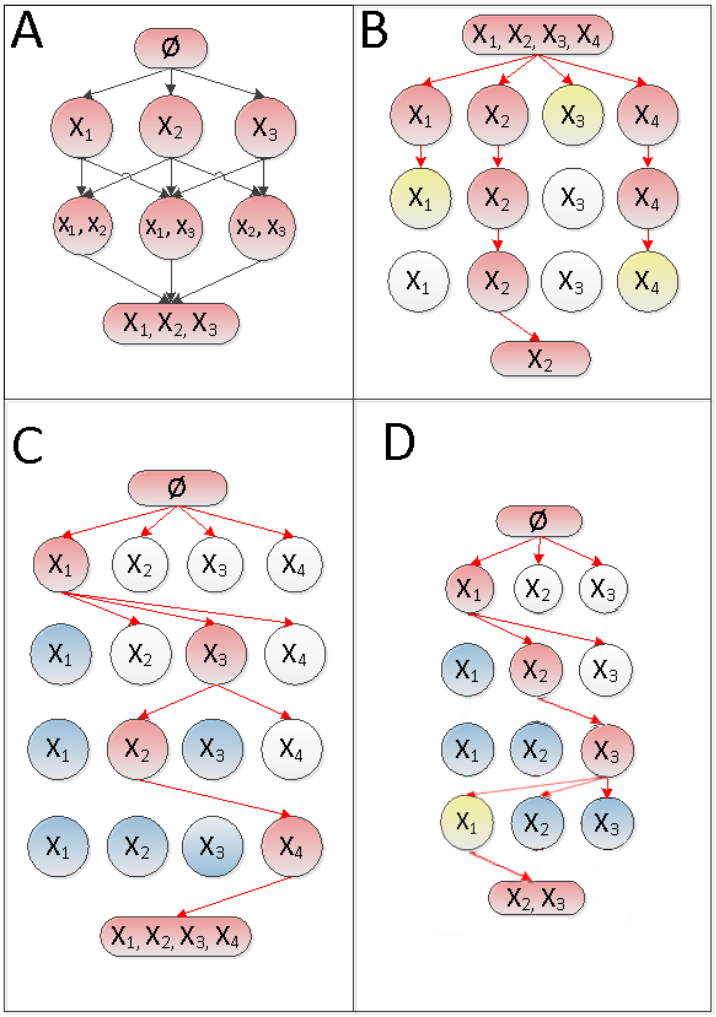

Figure 2.4: In this diagram, the prior selected features in forward selection are blue and the currently selected feature is red. In backward selection and exhaustive search the included features are red. Yellow and white nodes represent the features that were removed and those that were examined but not selected, respectively. *A* represents the exhaustive search algorithm which tests all feature subsets and selects the globally optimal one. *B* represents an example of the backward search algorithm. The selection starts with the complete set of all features and progressively removes the feature which either improves the score the most or decreases the score the least. The selection stops once a predetermined criteria has been satisfied. *C* is an example of the forward, greedy search algorithm. The selection starts with the empty set and progressively selects the feature which optimizes the scoring metric. The selection stops once a predetermined criteria has been met. *D* is an example of the forward search with backtracking. The algorithm starts the same as *C* and removes features if that improves the score.

rithm to go back and identify other potentially better search paths. An example of this can be seen in Algorithm 4

Forward selection with backtrcking works in a similar manner to forward selection, except that at each iteration it allows for the algorithm to search the space of selected features to see if the removal of any variants will not have a negative effect on the value of the scoring metric (see Figure 2.4D, Algorithm 4). This is advantageous when the aforementioned methods are not optimal, but the user is looking for a method capable of escaping local minima. This method, known as backtracking is the process by which the algorithm looks to remove unnecessary features after each selection of a new feature. While backtracking adds computational costs to the algorithm, as it requires a more thorough search of the feature space, it does so while eliminating branches that ultimately will lead to inadequate solutions.

Algorithm 4 Forward selection with backtracking

$\mathcal{S} \leftarrow \emptyset$
$e \leftarrow \infty$
repeat
    $b \leftarrow 0$
    $\gamma \leftarrow$ **FALSE**
    for $i \in \{1, \ldots, n\} \backslash \mathcal{S}$ **do**
        $e_i \leftarrow \mathcal{H}(\boldsymbol{X}_{\mathcal{S} \cup \{i\}}, \boldsymbol{y})$
        if $e_i < e$ **then**
            $e \leftarrow e_i$
            $b \leftarrow i$
            $\gamma \leftarrow$ **TRUE**
    if γ **then**
        $\mathcal{S} \leftarrow \mathcal{S} \cup \{b\}$
    for $j \in \mathcal{S}$ **do**
        $e_j \leftarrow \mathcal{H}(\boldsymbol{X}_{\mathcal{S} \backslash \{j\}}, \boldsymbol{y})$
        if $e_j < e$ **then**
            $\mathcal{S} \leftarrow \mathcal{S} \backslash \{j\}$
            $e \leftarrow e_j$
until $\gamma =$ **FALSE**
return $\mathcal{S}$

2.5.3 Embedded Methods

With the recent popularity of algorithms such as Lasso and Elastic Net [25, 38, 42, 84, 104], the field of feature selection is seeing a rise in the use of embedded methods. Embedded methods often have the speed advantage of filters while maintaining the relatively high predictive accuracies found

in wrappers. The features are selected at the same time that the model is trained [79], allowing for computational savings over wrapper based implementations. Like wrappers, the selected features are algorithm dependent and they allow for feature dependencies to be modeled.

With embedded methods, the user has more limited control over the feature selection process. Alterations such as changing the scoring metric can require a redesign of the algorithm itself while wrappers and filters can be applied with a number of different algorithms, each which may be advantageous to a particular problem set.

Lasso

Known for both its speed and predictive performance, the least absolute shrinkage and selection operator (Lasso) is a prime example of an embedded feature selection algorithm. Similar to other regression-based methodologies, Lasso generates a solution through the minimization of the sum of squares error and augments this with the ℓ_1-norm. In contrast to the ℓ_2-norm, the ℓ_1-norm forces a sparse solution [61, 87]. As it implements regularization, similar to RLS-based methodologies, it helps to penalize overly optimistic coefficients, reducing their affect on the output.

The objective function of Lasso is composed of a sum of squares, whose model is augmented through the use of ℓ_1 as its regularizer:

$$\sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} \boldsymbol{X}_{i,j} \boldsymbol{w_j} \right)^2 + \lambda \|\boldsymbol{w}\|_1 \ . \tag{2.19}$$

In fact, looking at Equation 2.19, the only difference with RLS is that while Lasso implements an ℓ_1 penalty via $||w||_1$, RLS implements an ℓ_2 penalty via $||w||_2^2$. While the overall equations appear similarly, it is important to not that it is this part of the solution that results in a sparse solution in the case of Lasso.

Through adjustments of the regularization parameter λ, the objective function can control the penalty applied to the features. If one were to set the value of $\lambda = 0$, the predicted value from the application of the objective function, the model learned by Lasso is equivalent to the model learned by linear regression. In other words, it can be expected that increasing λ will lead to a smaller number of selected features. The ℓ_1 penalty also regularizes in a similar way to the ℓ_2 regularizer in that it shrinks all coefficient values.

In Lasso, if two identical features exist, only a single one will be selected. The feature that is selected is dependent on the ordering of the features, as only the first one observed by the algorithm will remain. This can be problematic if a researcher's aim is to identify features whose combination can be accounted for by systems such as biological pathways. In this scenario,

one would want to maximize the number of features that appear in any given pathway to increase its enrichment score (see Publication IV [62]). The effect of maintaining multiple correlated features is known as grouping. Grouping can also be advantageous in the case when one of the selected features may not be reproducible in independent data sets. To incorporate grouping it is common to implement the Elastic Net.

Elastic Net

One important constraint that occurs in Lasso models is that due to the nature of the convex optimization problem being solved, selecting a larger number of features than there are training instances will cause a saturation in the predictive performance and will lead to severely overfit models [104]. To help alleviate this issue as well as allowing for grouping, a method known as Elastic Net has been proposed [104] and has been implemented on genetic studies [2, 16].

In a similar manner as Lasso, Elastic Net is a regularization technique which simultaneously applies a Lasso type feature selection with the ℓ_1-norm and ridge regularization with the ℓ_2-norm. Thus, Elastic Net can be considered as a trade-off between Lasso and RLS. The objective function of Elastic Net is the following:

$$\sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{n} \boldsymbol{X}_{i,j} \boldsymbol{w_j} \right)^2 + \lambda_2 \|\boldsymbol{w}\|_2^2 + \lambda_1 \|\boldsymbol{w}\|_1 \tag{2.20}$$

where λ_1 and λ_2 are the ℓ_1 and the ℓ_2 regularization parameters respectively.

The added λ_2 variable adds complexity to parameter selection that must be conducted in order to effectively run the algorithm. This can be problematic in large-scale data sets, where computing power is often already limited in its ability to apply these methodologies to large-scale studies, however the added running times are often not excessively prohibited. Rather, when dealing with larger data sets a more limited evaluation of λ_2 parameters can be conducted. This limited parameter selection in which $0 \leq \lambda_2 \leq 1$ is possible due to Elastic Net's nature to function between both RLS and Lasso that when setting $\lambda_1 = 0$ is equivalent to RLS, while setting $\lambda_2 = 0$ yields a result equivalent to Lasso.

2.5.4 Greedy Regularized Least-Squares

While regularization of least-squares based regression can help to alleviate the affect of large coefficients, greedy RLS, originally introduced in [65], additionally implements a greedy forward feature selection with the LOOCV heuristic (Algorithm 2). The computational efficiency of this is achieved via

matrix algebraic shortcuts and caching of the preliminary results. This allows the algorithm to be able to scale up to GWAS on modern, high-end desktop computers.

The formulation of greedy RLS is lengthy and complex. Without re-iterating much of the details in Publication II, the description would be insufficient. Therefore, a concise overview is provided here and readers are referenced to [65, 67] for a detailed algorithm. If readers are interested in the space-efficient variation then they should specifically see Publication II.

Let us define

$$J(\boldsymbol{Z}, \boldsymbol{u}) = \underset{\boldsymbol{w} \in \mathbb{R}^a}{\arg\min} \{ (\boldsymbol{u} - \boldsymbol{Z}\boldsymbol{w})^{\mathrm{T}} (\boldsymbol{u} - \boldsymbol{Z}\boldsymbol{w}) + \lambda ||\boldsymbol{w}||_2^2 \} \tag{2.21}$$

where $\boldsymbol{Z} \in \mathbb{R}^{a \times b}$ and $\boldsymbol{u} \in \mathbb{R}^a$ for some $a, b \in \mathbb{N}$ with $a \leq m$ and $b \leq n$. Here we use the symbols $\boldsymbol{Z}$ and $\boldsymbol{u}$ instead of $\boldsymbol{X}$ and $\boldsymbol{y}$ in order to stress that the objective $J(\boldsymbol{Z}, \boldsymbol{u})$ is optimized not with respect to the whole training data but with a modified set. That is, $\boldsymbol{Z}$ and $\boldsymbol{u}$ consist of only a subset of the rows of $\boldsymbol{X}$ and $\boldsymbol{y}$ due to the use of cross-validation and $\boldsymbol{Z}$ may consist of only a subset of the columns of $\boldsymbol{X}$ due to the greedy subset selection. Recalling that the values of the J can be expressed as:

$$\boldsymbol{w} = (\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z} + \lambda \boldsymbol{I})^{-1} \boldsymbol{Z}^{\mathrm{T}} \boldsymbol{u} \tag{2.22}$$

$$= \boldsymbol{Z}^{\mathrm{T}} (\boldsymbol{Z} \boldsymbol{Z}^{\mathrm{T}} + \lambda \boldsymbol{I})^{-1} \boldsymbol{u} \tag{2.23}$$

where the second form follows the first due to the matrix inversion identities [30].

A high-level overview of greedy RLS can be seen in Algorithm 5. This does not include the computational shortcuts. In order to calculate both the exact value of the LOOCV and updating the model with new features, greedy RLS makes use of the Sherman-Morrison-Woodbury formula (see [30]):

$$(\boldsymbol{A} + c\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{v}(c^{-1} + \boldsymbol{v}^{\mathrm{T}} \boldsymbol{A}^{-1} \boldsymbol{v})^{-1} \boldsymbol{v}^{\mathrm{T}} \boldsymbol{A}^{-1} \tag{2.24}$$

where $\boldsymbol{A} \in \mathbb{R}^{a \times a}$, $\boldsymbol{v} \in \mathbb{R}^a$, $c \in \mathbb{R}$ for some $a \in \mathbb{N}$.

To accelerate the computation of LOOCV one can set $\boldsymbol{A} = \boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z} + \lambda \boldsymbol{I}$, $\boldsymbol{v}$ is a single input vector $\boldsymbol{x}$ and $c = -1$. That is, it is considerably cheaper to update the previously trained model than to train it from scratch given that (2.22) has already been computed. The Sherman-Morrison-Woodbury formula can be used to remove the effect of one training example from the form 2.22.

To test the effect of an extra feature one can use the Sherman-Morrison-Woodbury formula to update the form 2.23 by setting $\boldsymbol{A} = \boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}} + \lambda \boldsymbol{I}$, $c = 1$ and $\boldsymbol{v} = \boldsymbol{X}_{:,j}$ where $j \in \{1, \ldots, n\} \backslash \mathcal{S}$.

Algorithm 5 Greedy RLS Algorithm

1: $\mathcal{S} \leftarrow \emptyset$
2: **while** $|\mathcal{S}| < k$ **do**
3: $\quad e \leftarrow \infty$
4: $\quad b \leftarrow 0$
5: $\quad$ **for** $j \in \{1, \ldots, n\} \backslash \mathcal{S}$ **do**
6: $\quad\quad$ **for** $i \in \{1, \ldots, m\}$ **do**
7: $\quad\quad\quad \boldsymbol{w} \leftarrow J(\boldsymbol{X}_{\{1,\ldots,m\}\backslash\{i\},\mathcal{S}\cup\{j\}}, \boldsymbol{y}_{\{1,\ldots,m\}\backslash\{i\}})$
8: $\quad\quad\quad e_j \leftarrow e_j + (y_i - \boldsymbol{X}_{i,\mathcal{S}\cup\{j\}}\boldsymbol{w})^2$
9: $\quad\quad$ **if** $e_j < e$ **then**
10: $\quad\quad\quad e \leftarrow e_j$
11: $\quad\quad\quad b \leftarrow j$
12: $\quad \mathcal{S} \leftarrow \mathcal{S} \cup \{b\}$
13: $\quad \boldsymbol{w} \leftarrow J(\boldsymbol{X}_{:,\mathcal{S}}, \boldsymbol{y})$
14: $\quad$ Update the cache matrices
15: **return** $\mathcal{S}$

By taking advantage of (2.24) and the cached preliminary results in both LOOCV and testing the effect of new features, lines 7-8 can be computed in constant time and it takes $O(mn)$ time to select each feature. Therefore, since there are k rounds, the overall complexity of the whole algorithm is only $O(kmn)$.

The primary advantage of greedy RLS in respect to its application to GWAS is its speed to calculate a wrapper-based feature selection in a reasonable time period without the need for any prior filtering. The multi-target variation of the algorithm has proved to be advantageous in other fields, recently performing the best at Sub-challenge 3 of the recent Broad-DREAM Gene Essentiality Prediction Challenge. Additionally, greedy RLS has later been extended to multi-target prediction problems (see [56]).

Chapter 3

Scalability of the Algorithms

3.1 Parallel Computing

While computers continue to progressively becoming more powerful, this power is no longer being added solely into individual processors as it was in the early 2000's. Although the overall processing power of machines has increased, this increase is accomplished through a combination of weaker processors. Consequently, one has to resort to parallel programming in order to take advantage of these new architectures. As of June 2015, with DigitalOcean.com, a shared-memory 4-core machine with 8GB of memory can be spawn for as low as $0.119 per hour or a single-core machine with 512MB of memory for $0.007 per hour. If a researcher is interested in developing a large network of smaller machines, he/she can easily generate a 100 processor machine for under $1.00 per hour.

A precursor to parallel programming is the assurance that the serial algorithm has been optimized. Numerous efficient numerical packages exist, each being capable of effectively performing scientific computations without running into problems with the high overhead. Packages such as Scipy/Numpy[34, 89], Lapack/Scalapack [4, 9] and R [74] all have the capabilities to handle such calculations, minimizing the required overhead and thus decreasing running times. Knowing the advantages of these different methods allows for the efficient scalability of algorithms.

An understanding of memory requirements can further help to optimize the data management in such a manner as to minimize the amount of space required to run operations on entire GWAS. A simple example is that treating the data as type *short int* (assuming one is not using the expected real values of the imputation), can reduce the required memory by 50% and 75% when compared to storing the data as a float and as a double respectively.

Only a limited number of GWAS implementations and packages (see e.g. [49, 70, 85, 102]) make use of any sort of parallel processing, despite

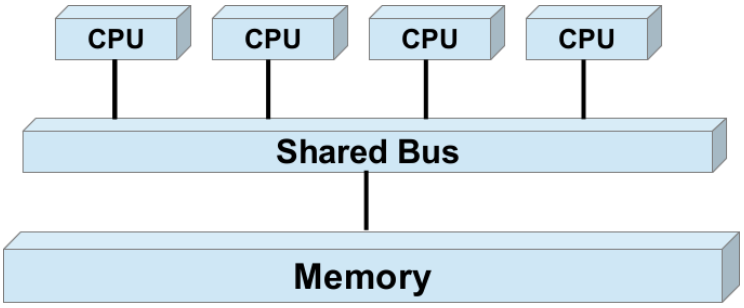

Figure 3.1: The above diagram represents a typical shared memory machine, in which a common set of memory is shared among multiple individual processing units, connected by a shared bus.

the fact that GWAS are naturally paraellizable. Some of the reasons for the lack of implementation likely stem from both a lack of knowledge on the part of the researchers as well as the significant effort that is often required to develop a parallel program when compared to a straight forward, serial implementation. Development of a parallel program requires knowledge of both computer architecture and advanced programming libraries.

Implementation of parallel computing on GWAS comes in a multitude of manners. The most straightforward way would be to split the features among the available processing units and then to have each processor calculate the univariate statistics or other simple calculations for each of the variants/combinations that are made available to it [85]. Those features passing a particular threshold can then be selected.

3.1.1 Architecture

Two popular types of parallel systems are shared memory and distributed memory machines. While both make use of multiple processors, they rely on different architectures and programming paradigms. A shared memory machine is one in which all processing units will have access to the same core memory, as shown in Figure 3.1. The most basic example would be a multi-core home computer, such as those using the Intel i7 processor. On this type of machine cores a and b can both access the same regions of memory and update the information. This type of computing is highly efficient as it generally allows for a lower level of network overhead which is commonly required when sharing data between distributed memory machines. Additionally, programming on a shared memory system tends to be easier than that of distributed memory machines.

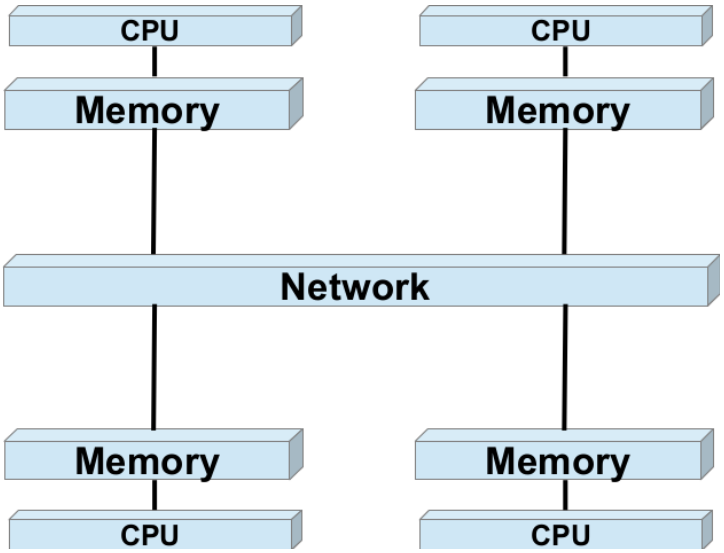

Figure 3.2: The above diagram represents a typical distributed memory machine in which multiple individual processing units and their associated memory units are connected by a common network.

In distributed memory systems, such as those shown in Figure 3.2, each processing unit tends to have its own designated memory that the other units are not able to access. However, their scale far exceeds that of shared-memory architectures and can often range into the petaflops of processing power with hundreds of thousands of processors. Programming for distributed memory machines is often done with languages such as the Message Passing Interface, better known as MPI [6]. Combining both MPI and OpenMP allows programs to make use of the computer's architecture.

3.1.2 Strategy

In parallel computing, each processing unit computes a fraction of the overall number of calculations, namely those that pertain to the features that have been assigned to it. The results of these calculations can then be broadcasted or sent directly to the other processing units if they require them (e.g. Single Program Multiple Data also known as SPMD). Alternatively, this message can be sent to a master processor who uses this data to determine and distribute new tasks to the slaves (e.g. master-slave) [15]. Other paradigms exists, but fall outside of the scope of this work so are not discussed here. The determination of which paradigm to use will often be dependent on a number of factors including the algorithm being implemented, the number

of processors being used and how often communication is required. An example of these paradigms can be seen in Figure 3.3.

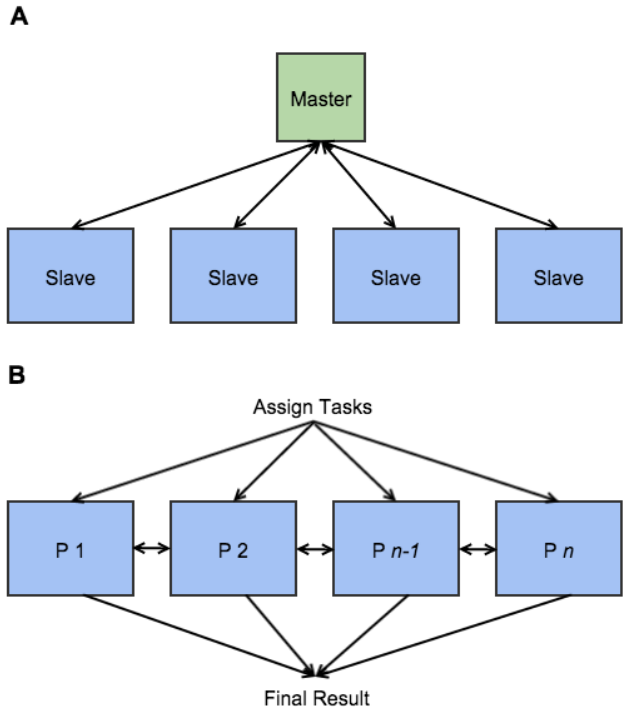

Figure 3.3: *A* is an example of the master-slave paradigm in which a master distributes work loads to the slaves. The slaves complete the tasks and then return the results to the master. The master then can either end the program or distribute new tasks to the slaves. *B* is an example of SPMD in which the problem is decomposed into smaller problems which are simultaneously solved. The data is first distributed among the n processors. After the processors have solved their respective problems they can either communicate their results to other processors to determine their next step or the results can be collected. In the figure, the double-ended arrows indicate communication between neighboring processors.

The efficiency of a parallel program is typically evaluated with metrics that measure the performance with relation to the number of processing units being utilized. The first method known as the speedup is defined for p processors by:

$$S_p = \frac{T_1}{T_p} \tag{3.1}$$

where T_1 and T_p are the execution times on a single processing unit of the fastest known serial algorithm and when applied to the parallel program using p processors respectively. Ideally, $S_p = p$ which would indicate that doubling the number of processors would double the speedup.

The efficiency of the program when using p processors is defined as:

$$E_p = \frac{T_1}{pT_p} \tag{3.2}$$

$$= \frac{S_p}{p} \tag{3.3}$$

In an ideal scenario $E_p = 1$ which indicates perfect scalability of the algorithm. Obtaining a value of $E > 1$ is known as superlinear speedup, which while rare is known to happen. A scenario in which this would be feasible is when a sequential algorithm would not be able to load the entire data set into memory on a single processor. When running on additional processing units, more memory can become available leading to efficient caching of the data resulting in a performance boost.

3.1.3 Application

Wrapper-based feature selections have traditionally been computationally prohibitive which may generate an interest in parallel implementations. The primary challenge in these algorithms is how to share the data between the different processing units, as the value of the previously selected feature will have an affect on the calculated values of all of features at the next iteration. Additionally, the wrapper methods that have been shown to be able to scale to GWAS (e.g. greedy RLS) require excessive caching and recalculation of the cache after each iteration. To parallelize such algorithms, a system would have to be developed that could handle such cache matrices, while keeping them in sync and allowing for communication among the processors.

An overview of parallelized greedy RLS is outlined in Algorithm 6. The algorithm provides identical results to the original form of greedy RLS, presented in Chapter 2, but rather starts on p processing units. Further it has both global ($\mathcal{S}$) and processor specific feature sets ($\mathcal{S}_p$). $\mathcal{S}$ maintains the list of the selected features at each iteration and $\mathcal{S}_p$ are the features on processing unit p which are in $\mathcal{S}$. The algorithm then continues in a similar fashion as greedy RLS, except that each processing unit calculates its locally optimal feature at each iteration (lines 7-13). Once a locally optimal feature has been selected, these are then compared by the master processor

which selects a globally optimal one (lines 14-18). The processing unit which originally selected the optimal feature then broadcasts its part of the cache matrix to all other processors (lines 19-22). All processing units then update their local caches based on the previously broadcasted vector (line 23). A more in-depth overview of the algorithm is described in Publication III.

Algorithm 6 Parallel Greedy RLS Algorithm

1: $\mathcal{S} \leftarrow \emptyset$
2: $\mathcal{S}_p \leftarrow \emptyset$
3: $\mathcal{N}_p \subset \{1, \ldots, n\}$
4: **while** $|\mathcal{S}| < k$ **do**
5: $\quad e \leftarrow \infty$
6: $\quad b_p \leftarrow 0$
7: $\quad$ **for** $j \in \mathcal{N}_p \backslash \mathcal{S}_p$ **do**
8: $\quad\quad$ **for** $i \in \{1, \ldots, m\}$ **do**
9: $\quad\quad\quad \boldsymbol{w} \leftarrow J(\boldsymbol{X}_{\{1,\ldots,m\}\backslash\{i\},\mathcal{S}_p\cup\{j\}}, \boldsymbol{y}_{\{1,\ldots,m\}\backslash\{i\}})$
10: $\quad\quad\quad e_j \leftarrow e_j + (y_i - \boldsymbol{X}_{i,\mathcal{S}_p\cup\{j\}}\boldsymbol{w})^2$
11: $\quad\quad$ **if** $e_j < e$ **then**
12: $\quad\quad\quad e \leftarrow e_j$
13: $\quad\quad\quad b_p \leftarrow j$
14: $\quad$ **if** rank $= 0$ **then**
15: $\quad\quad$ Gather from all processes e_p, b_p
16: $\quad\quad$ $q \leftarrow \text{argmin}_i\ e_i$
17: $\quad\quad$ Broadcast process index q to all processes
18: $\quad\quad$ $\mathcal{S} \leftarrow \mathcal{S} \cup \{b_q\}$
19: $\quad$ **if** rank $= q$ **then**
20: $\quad\quad$ $\mathcal{S}_p \leftarrow \mathcal{S}_p \cup \{b_p\}$
21: $\quad\quad$ Broadcast $\boldsymbol{X}_{:,b_q}$ and the relevant parts of the local cache
22: $\quad\quad$ matrices to all cores
23: $\quad$ Update the cache matrices on all p
24: $\boldsymbol{w}_\mathcal{S} \leftarrow J(\boldsymbol{X}_{:,\mathcal{S}}, \boldsymbol{y})$
25: **return** $\boldsymbol{w}_\mathcal{S}, \mathcal{S}$

Not all feature selection methodologies can be parallelized in their original form. Lasso trained with cyclic coordinate descent is an example in which parallelization is traditionally done over the samples. This can be inefficient when the number of variables exceeds the number of samples. Thus, parallelization would require the data to be split in a non-optimal manner and most implementations make use of stochastic coordinate descent rather than cyclic coordinate descent. This has the potential to create a complex scenario, such as the need for the use of Shotgun (see [11]) or Coloring-based (see [80]) parallel methods. It has been shown that in Shot-

gun, having correlated features can lead to divergence in the scenario where too many features are updated simultaneously [11]. This increases the problem complexity when compared to the exact solution of greedy RLS.

Chapter 4

Summary of the Thesis Work

4.1 Contributions

This dissertation is composed of a total of five original, peer-reviewed research publications that are referred to as Publications I-V. These works follow a general theme of genetic feature selection/complex disease prediction and progress from one work to another. While the papers are not direct extensions of each other, significant effort was made to maintain the consistent theme. The forums in which they were submitted were selected as to maximize their wide accessibility by the research community and to guarantee that the methods and their findings were available freely to the public. This is an essential aspect of assuring that all researchers will have the ability to continue on extensions to both the algorithms and applications of the methods.

4.1.1 Genetic Variants and Their Interactions in the Prediction of Increased Pre-Clinical Carotid Atherosclerosis: The Cardiovascular Risk in Young Finns Study

Publication I in this dissertation was based on a collaboration with the YFS. The study is an on-going, population-based, follow-up study that followed a group of Finnish individuals from childhood until adulthood. It started in 1980 as a multi-center research project which sampled individuals from five university hospitals in Finland. The five locations were Turku, Tampere, Helsinki, Kuopio and Oulu. When the study began in 1980 there were 3,596 subjects who were aged between 3 and 18 years of age. Follow-ups with the patients were conducted at various time points during which time various clinical and/or genetic measurements were taken from the cohort subjects. This particular study utilized the data from the 2001 and 2007 follow-up periods, during which there were 2,283 and 2,204 subjects respectively who remained in the study.

The particular phenotypic trait that was of interest was the carotid artery's intima-media thickness (IMT). The IMT was selected as the outcome variable to be explored due to it's known association with the onset of cardiovascular disease [46, 63]. In addition to the IMT, various clinical measurements including but not limited to HDL cholesterol, LDL cholesterol, total cholesterol, BMI, systolic and diastolic blood pressure, waist circumference, triglycerides, ApoA1, ApoB, age, sex and smoking habits were measured. This was then complemented by a genetic analysis of single-nucleotide polymorphisms that were reasonably expected to have some relationship to cardiovascular disease.

As missing data can have adverse affect on machine learning methodologies, and this project was started without a particular algorithm that was going to be implemented, the data set was first adjusted so that it consisted of a complete set of genetic data. A predefined set of 17 genetic variants was identified by the clinical heads of the study that were of particular interest due to their notation in similar studies as having a potential link to cardiovascular disease. Genetic variants and individuals with missing data were then removed from the study in such a manner so as to attempt to maximize the size of the data set. The resulting data set had 1,027 and 813 individuals in 2001 and 2007 respectively. The final study also contained 108 SNPs and 13 conventional risk factors (CRFs).

The study started through the demonstration that while several CRFs were statistically significant in both the 2001 and 2007 studies, they were not significant in predicting the IMT progression, which was calculated by subtracting the 2001 IMT value from the 2007 value. This was important since determining if an individual currently has a particular prognosis is trivial, but generating a prediction regarding their future disease status can be complex.

The goals of this study were numerous. While the main task was to generate a prediction of the IMT, an equally important task was to examine which features were being selected since this could help lead clinicians towards candidate variants for future studies which may have additional complex phenotypic associations. To do this, various feature combination models were analyzed including the CRFs and the 17 prior identified SNPs, the conventional risk factors alone, the significant SNPs combined with the CRFs and a two-step feature selection based methodology which automatically selected the relevant SNPs and CRFs. Each of these various data sets was tested for 2001, 2007 and the progression.

During the study it quickly became apparent that due to the nature of the IMT, in which a vast majority of the individuals are not considered to be at risk, we were posed with a problem of class distributions that would be heavily skewed towards the controls. Combined with the small size of the data set and even more limited class sizes when taking into

account the various cross-validation folds, a different methodology needed to be implemented to compose the case and control sets. This was done by gradually increasing the class-sizes of both the low-risk and high-risk individuals based on the top and bottom percentage of IMT values. For this study five different quantile points with ranges of 5-25% at 5% intervals were implemented. In other words, anywhere between 10-50% of the data set was used with the cases and controls stratified. The high-risk individuals were adjusted into a unified class, in which they were considered the cases and in a similar fashion the low-risk individuals were pooled into a singular, low-risk class. These classes were then used as a binary classification problem.

As this study was considered to be a pilot study of the feasibility of applying machine learning-based feature selection methodologies to data sets combining both genetic data and CRFs, the goal was for the development of a methodology that would potentially scale to entire GWAS. When initially deciding on which machine learning technique to implement, this was strongly taken into account and due to Naïve Baye's ability to perform well in the test studies, along with its scalability and lack of a need for complex parameter selections it was chosen as the candidate method for the study. To help reduce the computational run-time of the program an Information Gain based filter was first applied to the data to reduce the number of associated variants and CRFs to the top 40. A best-first search strategy, combined with a backwards selection and cross-validation was then used to develop the final set of variants. This process was repeated over all of the various quantile points and for the different outcomes that were being examined. One difference between the outcomes was that when trying to predict both the 2007 and the IMT progression class labels only the CRFs that were measured in 2001 were used. This helped us to examine whether the particular selected variants were indicative of not only the current risk level but also the future risk level based on current characteristics.

Positive 10-fold CV predictions based on the AUC were achieved, with a demonstration that an increase in the AUC could be achieved for the 2001, 2007 and IMT progression studies. It was additionally shown that these scores were capable of surpassing the results that were achievable when using the CRFs alone. This demonstrated that through the use of the genetic variants combined with the CRFs, the predictive performance could be significantly improved. Similar patterns were also noticed in the IMT progression experiments. As expected, increasing the size of the quantile groups generally reduced the performance.

As testing on independent data is the gold standard of model-validation. The originally discarded data was mined to search for additional instances that could be utilized to generate an independent test set. This *recycling* of the discarded data was a valid contribution as it demonstrated that after running a feature selection that utilized only a fraction of the original data

set, an independent data set of a similar scale could be composed of these instances which would otherwise have been lost. The same IMT thresholds that were used in the main study were used for splitting the independent data set into the various quantile groups.

While a decreased predictive performance was noted on the independent data set, this performance loss was relatively limited and helped to confirm the results from the first part. As an additional step, the effect of slight adjustments to the quantile group cutoff points and their affect on the results was examined. This demonstrated that while the arbitrary selected cut-off points for the various cohort sizes could perform well, adjustments could help to further boost the performance. Finally, the selected features were analyzed and an analysis based on the epistatic interactions between the features were examined. This helped to affirm that the selected features were in fact interesting candidates and warranted further investigation into their underlying biology.

This study gave a solid foundation on which the research could be continued and expanded upon. The base methodology was implemented through generic means (in this case Weka [28]), though future methods would make use of more complex and custom programs. A major contribution of this study was the demonstration that machine-learning based feature selection of genetic variants could outperform traditional CRF based analyses. Further, it was shown that machine learning approaches could improve beyond p-value based filter methods. These findings help to support the theory that researchers need to look at rare variants in order to explain the heritability of complex disorders. Through the base developed during this work, the future research, while not as biologically oriented would prove to address many of the computational issues that were discovered during this scaled down feasibility study.

4.1.2 Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations

Publication II was the first study during this dissertation in which an entire GWAS was utilized as a data set. Its main purpose was to examine the feasibility of scaling up wrapper-based machine learning methodologies to GWAS scale without the need for any pre-filtering of the features. This was done while implementing a methodology that made use of valid practices such as implementation of nested cross-validation to help better estimate the true predictive performance of the final trained model. An application was made to the Wellcome Trust Case Control Consortium's (WTCCC) Hypertension data set combined with the UK National Blood Services control set [14]. From this it was concluded that wrappers have the potential to be readily

applied to GWAS and are in fact not too computationally inefficient when applied with intelligent caching and matrix shortcuts. Further, based on the predictive performance, it was shown that wrappers are a viable option to be run simultaneously with other analysis methodologies.

The methodology implemented in Publication I is relatively common in GWAS feature selection papers [45, 96], where a filter is first applied to reduce the sample space, followed by some variation of either a machine-learning or feature selection methodology to provide the final data set and/or predictions. While this is a valid methodology that has proved successful, it is exposed to the potential loss of information due to drawback of the use univariate statistics. As previously discussed, these methods are not always particularly suited towards selecting features in which there may exist complex epistasis interactions among the genetic variants. By demonstrating that it would be computationally feasible to eliminate the univariate statistic filter and then run a wrapper feature selection over an entire GWAS, we aimed to help expose the research field to more advanced methods that may detect relationships that could explain some of the heritability that often remains hidden in these studies. It should be made clear that while one of the primary goals was to demonstrate methods that could provide good predictive power, it was not to meant to imply that the investigated methods could universally outperform other methods.

Tapio Pahikkala et al. designed greedy RLS in [65]. My contribution to this work was in adjustments and application of the algorithm to GWAS. In conjunction with the group we were able to adapt and implement the method greedy RLS [65] to the WTCCC Hypertension control set combined with the UK National Blood Service's (NBS) set of controls. After applying standard QC filters to the study the resulting data set contained 3,410 individuals and 404,452 SNPs.

Analyses on greedy RLS's ability to be implemented as a GWAS-wrapper feature selection methodology was done to not only examine its ability to select features on the current set, but also to test for its scalability to larger sets that would prove to be more computationally intensive. This undertaking resulted in two different versions of greedy RLS being examined, a space-efficient variation which was dependent on the number of features selected with an approximate performance of $O(min(k^2mn, km^2n))$ and the ordinary implementation with a complexity of $O(kmn)$. While it was demonstrated that significant speedups could be obtained using the original implementation, this came at the cost of having to store four copies of the data matrix in memory.

Assuming that a typical GWAS contains 500,000 SNPs and 3,000 subjects, researchers are left analyzing a data set consisting of approximately 1.5 billion data points. Considering that the SNPs can be represented as an integer value consuming 4 bytes, this would require approximately 6GB

of data for a single copy of the data set. This is prohibitive when also accounting for the cache matrices that are used and require other data storage types. This can quickly grow as the size of GWAS increases, resulting in the need for space-efficient variations, even though these methods have a higher computational cost. Despite these requirements, the fast implementation was able to perform a wrapper-based feature selection on a nested, three-fold cross-validation, selecting the top 50 features for each fold in under 26 minutes on a high-end desktop machine. This helped to bring the feasibility of large-scale machine-learning, feature selection on GWAS to all groups, regardless of their access to high-end clusters.

The results of the wrapper-based feature selection was compared to both traditional two-step feature selection methodologies that made use of a Fisher's Exact Test filter combined with a wrapper step and one in which the statistically significant features were analyzed in the order of their significance and RLS applied on top of these in the same order. A performance gain in the AUC of approximately 0.04 was gained through the use of only greedy RLS, helping to enforce the theory that wrappers are capable of outperforming other methodologies, though coming at a higher computational cost. Additionally, through a literature review-based analysis of the selected features, it was shown that greedy RLS was able to select both SNPs that were supported by previous studies as well as a series of new potential candidates.

As commonly seen in GWAS, SVMs tend to be industry leaders in terms of their frequent use. While they have provided good results, their computational complexities often made them prohibitive for running in conjunction with wrapper based feature selection methodologies. This was examined by demonstrating that the linear-kernel in LibSVM was not capable of performing large scale wrappers on GWAS. While the provided implementations were done as an example case of what is often seen (implementations of pre-built packages such as e1071 in R and Weka), it does provide a comparison which was intended to inspire other researchers of the need to examine machine-learning research to identify new methods which may be capable of solving their problems while not limiting the problem search space.

Through this article it was aimed to introduce new methods capable of performing the wrapper-equivalent feature selection on entire GWAS on a high-end, readily available desktop machine. This analysis was not only able to provide both computationally and space-efficient variations, but demonstrate their performance on real-world data sets. It was also useful in demonstrating that a performance gain could be obtained on an entire GWAS data set when implementing a wrapper based methodology compared to p-value based and two stage feature selection. This is not to say that wrappers will always outperform other methodologies, but rather that they have the potential of making them a candidate methodology for modern GWAS.

4.1.3 Parallel Feature Selection for Regularized Least-Squares

Consistent with one of this dissertation's underlying themes, Publication III directly applies to the scalability of machine-learning based feature selection (namely, greedy RLS) toward the next-generation sequencing-based GWAS which will be larger and contain rare variants. These data sets will provide a new host of computational issues, which if left unaddressed will leave even more limited discovery methodologies then are currently implemented on the current GWAS scale. By containing tens of millions of variants and thousands of examples, these cohorts will be so massive that it is not feasible to assume that they can be run on desktop machines. As was demonstrated in Publication II, even data sets of the current GWAS scale require enormous amounts of memory just to read in the data.

To address this concern, researchers will have to start to focus on modern computing technologies such as parallel and cloud computing to gain enough memory and computational power to process these massive data sets. Without these technologies, it is plausible that researchers will fail to apply methods capable of creating complex models that are necessary to explain much of the missing heritability in these studies. In order to move beyond the univariate algorithms, the current set of feature selection methodologies will need to be adapted to the aforementioned technologies so that they remain accurate, scalable and efficient enough for widespread use.

To address this scenario Publication III aimed to demonstrate that greedy RLS is a particularly well suited method for GWAS studies due to its scalability to large number of processors. The algorithm was decomposed to its core elements and rebuilt in a distributed manner so that the fundamental processes of analyzing the genetic variants could be efficiently distributed among the processors. In this implementation, a single processor acted as the master which generated a number of sub-problems that are distributed among the other assigned processors.

In this particular situation, each processor selects a locally optimal feature and its associated performance and feature information are sent back to the master for a determination of the globally optimal feature for this particular iteration. Once a globally optimal feature had been determined by the master, the processor which has the corresponding feature contained in its data partition then extracts and broadcasts the necessary cache data from the particular processor to all other processes. This is done so that they can update their locally stored caches accordingly and continue to select features in accordance with the original algorithm. In this particular setup, static-load balancing was utilized allowing the master processor to contain its own data partition and conducted the same computations as the other processor. While this procedure could have been equally done with

broadcasts between the processes in a purely SPMD manner, initial tests indicated no additional speedup from this alternative implementation.

While feature selection parallelization can be trivial if no intra-process communication is required, a complex step was decomposing greedy RLS into its core computations to determine how to effectively split up the cache matrix. This splitting had to be done in such a manner so that each core was able to not only calculate its respective locally optimal feature, but did so while staying in sync with the cache matrices that were disjoint and stored on the other distributed processes. As a globally optimal feature was required to be selected after each iteration of the loop, all processes had to wait until all other units had completed their feature selection so that the results could be compared. Upon receiving the broadcasted caching matrix slice, each individual process would then update the caches for their local features according to Algorithm 1 in Publication III.

This method was tested on the WTCCC Type 1 Diabetes (T1D) data set combined with the NBS controls, along with an artificial sample data set. It was run under a wide variety of scenarios that were aimed at testing its scalability with respect to the number of features, number of examples and the selected subset size. These scenarios were analyzed for varying number of cores (from 1 to 128), by doubling the number of processing units between each test. It was demonstrated that the parallel greedy RLS algorithm was able to attain high levels of speedup and efficiency, indicating that the algorithm is suitable for larger scale testing. When testing on large numbers of cores, the running times decreased to such a level that they started to become monopolized by the start-up costs, likely explaining the decreases in speedup and efficiency at these points (Figure 3 in Publication III).

On the WTCCC T1D data set, the top 14 features on the entire study were selected. This number was determined after running a nested cross-validation over three external folds, with an inner-LOOCV conducted by greedy RLS and selecting the optimal point in this process. Of the selected features, 11 out of the 14 were able to be shown to have a possible association with the disease or to be located in the MHC region, an area of the human genome known to be associated with T1D [12, 57]

Through the ability to calculate entire wrapper-based GWAS feature selections in only a matter of tens of seconds to minutes (depending on the data set size, number of processors and only including post-preprocessing analysis), this article demonstrated the ability for greedy RLS to act as an implementable tool for both modern day GWAS and the next-generation sequencing studies that are starting to emerge. It provided the algorithms necessary to conduct such experiments while still demonstrating its potential on real-word studies. In addition to the research, it also acted as a tool for promoting the need to analyze these studies in conjunction with computer

scientists who have the know-how to adapt the current technologies to create synergistic research that will help to identify those variants that have thus far remained hidden in current research.

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

Publishing research does not guarantee awareness of the work and due to the often limited scope of an individual's work it is often necessary to analyze the work in relation to follow-up research that may have been conducted. In order to analyze the current field of genetic interactions and to raise awareness of our group's work in machine-learning based feature selection of GWAS, Publication IV, a perspective article in BioData Mining was published with a look into network interactions and their potential to help further explain the genetic basis of T1D.

While traditional studies have often relied on univariate statistical methods to filter down the input data set that would be analyzed, this has the unfortunate effect of potentially filtering out relevant genetic variants. Various epistatic interactions among these features may help to explain some of the missing heritability that is often seen in GWAS and similar studies [50, 51]. In Publications I and II the ability to identify SNPs which were able to help to explain some of the variance found in the phenotypic predictions were examined. However, as seen in Publication II, not all of the SNPs would be able to mapped to references in established research.

Identifying SNPs that have not been previously associated with a particular disease has the potential to act as a positive result since it can be assessed as the algorithms ability to distinguish new, potential epistatic interactions that cannot be determined through univariate studies. Publication IV pointed out that the SNPs themselves are not necessarily the interacting factor with the output, but rather it can be through synergistic interactions with molecular pathways that may lead to the phenotypic outcomes. Additionally, as has been shown in microarray based studies, there may be only a limited overlap between the various studies, while still allowing for a greater union when considering the molecular pathways [48]. To analyze the usability of molecular pathways for explaining feature subsets, we examined the selected features from greedy RLS applied to a well known GWAS.

Further, through the analysis of the selected variants individually, synergistically and via their associated molecular pathways, it was demonstrated in Publication IV that it is possible to verify both the selection of genetic variants that were previously known to be associated with the disease in combination with newly selected features. While the computational model

validation aimed at confirming the predictive power of the results, this was coupled with an analytic justification of the selected variants. This analysis demonstrated that the methods implemented were capable of selecting new feature subsets that could warrant further experimental investigation.

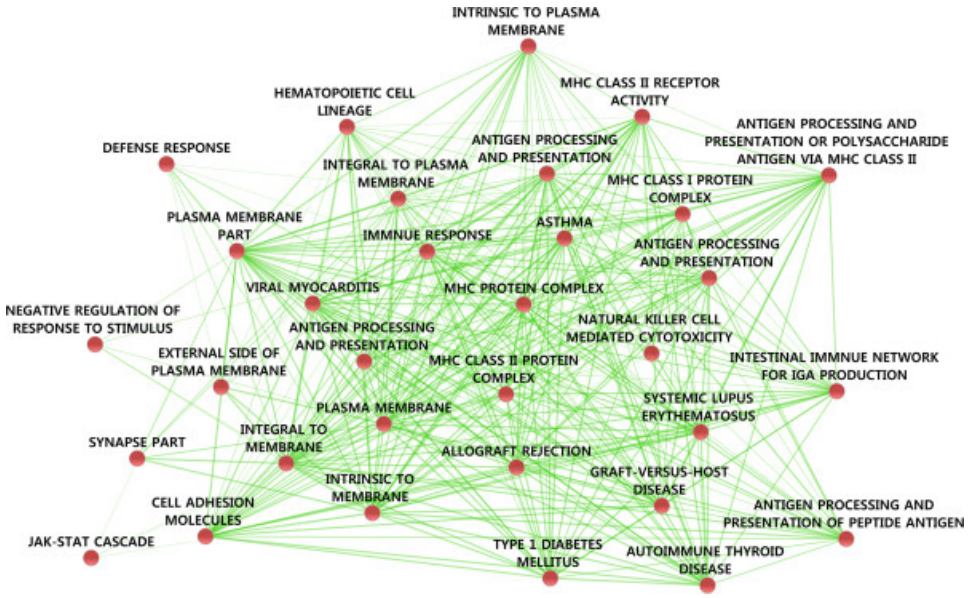

Figure 4.1: An example of the molecular networks that were determined to be involved in the onset of type 1 diabetes via a feature selection conducted with greedy RLS combined with features that were prior selected in another research group's publication [20]. The molecular pathways may help to explain why studies selecting different variants may still provide similar predictive results. The figure is taken from Publication IV.

Using greedy RLS, a set of features selected from the WTCCC's Type 1 Diabetes (T1D) data set combined with the UK National Blood Services' (NBS) controls were selected. The SNPs were mapped to their respective genes using the DAVID web service [31]. DAVID is a software tool that performs gene set enrichment analysis to identify biological pathways and gene ontologies that are significantly overrepresented in the study when compared to a background set. The mapped genes from the web service were extended with a set of previously selected genes from another work [20]. This combined aggregation of genes was analyzed in DAVID to get the resulting set of significant pathways which were mapped in relation to one another based on their overlapping genetic components (Figure 4.1). These molecular interactions can in turn be used as *a priori* information for predictive models. This may help to lead to more effective filtering techniques which can make

more informed decisions regarding which features to remove if analyzing the data via multi-stage feature selection.

Networks such as those seen in Figure 4.1 can help further the understanding of their associated disease phenotypes due to their ability to help explain complex interactions that may occur between larger groups of variants that would not normally be examined during initial analyses. Models may be able to more effectively be applied to the individual pathways, complemented by an ensemble of multiple machine-learning based applications. Additionally, models of this sort may help reveal additional variants not detected by feature selection algorithms, complementing the model and may help to increase its predictive accuracy.

The article continues by discussing principles of machine learning, model selection and model validation and how to avoid pitfalls that can result in information leaks and bias the final results. Through additional analysis on the metric curves from Publication II, it was concluded that simply applying model validation methodologies such as CV can be affected by overfitting and results in overly optimistic conclusions. When testing, it was illustrated that the error in the training set continually decreased as the CV becomes part of the training algorithm itself (see Figure 1 in Publication IV).

4.1.5 Regularized Machine Learning in the Genetic Prediction of Complex Traits

Having spent a significant amount of time and resources toward the analysis of two-step feature selections and greedy RLS, its scalability, parallelization and the application to GWAS, the final survey in this dissertation, Publication V, was meant to address the underlying fact that there is not a universally optimal method. Rather, each method will tend to have specific characteristics that while making it appropriate for certain tasks, will similarly make it sub-optimal when applied to different problem sets.

A growing class of machine learning methodologies is the embedded class, which has the potential to combine the speed of filters with the accuracy of wrappers. Within this class of algorithms, the methods, such as Lasso and Elastic Net, tend to conduct the feature selection within the learning algorithm itself. These techniques have experienced an increase in their implementation on GWAS due to their speed, simplicity and performance [16, 38, 91, 99]. Their performance has demonstrated the ability to outperform standard linear models [38] and their implementation is widely available in numerous publicly available software packages [23, 69].

Publication V made use of the WTCCC T1D and a second yeast cross data set [10]. These were analyzed through the application of a number machine learning and modeling techniques to demonstrate that while numerous methods may be capable of performing well they all contain their

own characteristics that may make them better suited for different studies. Since T1D has been previously shown to have high AUC's when applying machine learning methodologies [96], it is no surprise that many of our methods performed well. However, the characteristics of the selected features is what is of higher interest. It was observed that methods such as greedy RLS actually performed slightly worse than embedded methods (Elastic Net and Lasso) and the top performing method was actually a filter feature selection that combined a χ^2 feature selection with a ℓ_2 regularized Logistic Regression. However, this required a far greater number of features than greedy RLS. Ultimately, this can be interpreted as a possible sign that methods such as greedy RLS, while not always performing optimally in terms of the performance of the scoring metric, may in fact select features, whose interactions with one another may solve some of the missing heritability that has been identified in certain diseases.

It was also interesting to note the relatively good performance that was achieved through the use of embedded methods. This is important since the predictive results, combined with the simplicity of their implementations may make them suitable methods for many researchers who do not necessarily have the advanced training that is necessary to properly tune many of the other algorithms. While tuning for these methods is required, their widespread use has led to the development of algorithmic functions that are capable of performing an analysis on a widespread set of tuning parameters with minimal interaction from the user [69].

The performance of the embedded methods outperformed greedy RLS in both data sets, but did so with a larger number of selected features. However, it was noted that while the performance of the embedded methods was quite good, it is difficult to come up with a prediction of their running times on other data sets due to their reliance on convergence of a solution that is often based on coordinate-descent. This form of running time, that is reliant on the data set being implemented on, may result in impractical running times if implemented on next-generation sequencing studies. While there is the potential that they will arrive at a solution in a relatively small number of steps, this cannot be guaranteed, which may result in researchers having to either rely on approximations of the solutions, and/or examining other methods simultaneously.

It was also noted that the filter feature selection in the yeast-cross underperformed all of the other methods. While it was indicative that it may eventually catch up with the other methods, the spread between the algorithm's performance was so large, that any performance gain would potentially be the result of many false positives. This is an ideal example of how different methods may outperform others depending on the study being analyzed. Their is no universally optimal method and for this reason we need

to adapt the current methods to be both fast enough and accurate enough to warrant their use in a wide variety of studies.

Through the comparison of multiple methods in different studies (both regression and classification problems) we were able to examine the performance of various feature selection methodologies in a multitude of settings. It was interesting to see that in some scenarios, the simplest methods may outperform more complex ones, but do so through the selection of a much higher number of features. In other scenarios, similar filter based selection methodologies, severely underperformed their wrapper and embedded counterparts. Through these results and others that analyzed both the performance and selection characteristics we were able to demonstrate the lack of a universally optimal feature selection methodology and the need to often test a wide variety of techniques on a sample data set in order to select the final methodology that will be implemented. Of similar importance, was the observation that while some methods may have performed slightly worse based on the predictive performance, their ability to do so with a much smaller number of features warrants their further examination. Because of their potentially unique feature sets, their selection may reveal added insight also into the disease biology. This ability of greedy RLS to perform well with relatively small feature sets is highly applicable to many fields such as personalized medicine. This is due to the problem setup in which a minimal set of sufficiently predictive features is being identified.

Chapter 5

Conclusion

Modern science has unleashed a vast amount of information about varying topics from medical science to search engine optimization. This data has come at the cost of high-processing power and is on such a scale that explicit programming of accompanying models is not a feasible task. In the field of bioscience and personalized medicine the advent of genome-wide association studies and next-generation sequencing have provided a prospective insight into the capabilities of this data if processed correctly. Due to both their size and complexity it has become a difficult task to account for the exponential amount of genetic interactions that can occur between the genetic variants and to develop models that will maximize the heritability and predictability explained by these inputs.

To develop predictive algorithms researchers have started to implement machine learning based methodologies to these studies in an attempt to extract relevant data, while improving upon traditional univariate approaches [38, 59, 67, 78, 84, 96]. Through implementations of various classifiers and regressors, a steady improvement upon the predictive accuracy has been noted [59, 96]. This improvement comes at both increased computational costs and complexities of the algorithms involved. Further, as seen in Publications II and III, scaling these methods to the GWAS is a difficult task that often requires the use of optimized algorithms or the use of large-scale parallel machines. By combining the aforementioned methods, it is feasible to scale even these complex algorithms to NGS studies.

The thesis is based primarily on the application and scalability of machine learning based feature selection to both select relevant genetic variants and generate predictions for disease onset in a variety of cardio-metabolic disorders from cardiovascular disease [59] to Type 1 Diabetes [58, 61, 62]. The papers have been published in an approximate progression starting with an exploratory pilot study (see Publication I), to a proof of concept paper demonstrating the scalability of greedy RLS to entire GWAS (see Publica-

tion II), to the development, implementation and analysis of parallel versions (see Publication III), network integration (see Publication IV), finally to a cross-method analysis of machine learning methodologies in genetic disease prediction (see Publication V). Through the course of these studies, ideas were developed, presented, applied, altered and analyzed. While new feasible methods were presented, it was equally important to demonstrate that there is no universally optimal method for analyzing GWAS.

This lack of a universally optimal method is vital to continued development of the field. It is when researchers base a new study's methodologies completely on the work of an alternative researcher without using the same data sets that suboptimal results are likely to present themselves. There are a multitude of factors that are influential on an algorithm's performance and while a method may outperform others in a particular data set, as seen in Publication V, this is not an indication that it will perform the same in other studies. It is only through the careful analysis and model validation that researchers can make determinations on which methods should be applied for the final analysis.

Being useful today, but essential tomorrow has been an underlying theme to this thesis. As GWAS start to dissipate and NGS become the new wave of studies that aim to correct the flaws in GWAS [17, 103] that include having too small sample sizes and inadequate SNP coverage, researchers are approaching problems that are rapidly increasing in computational complexity, especially when attempting analyses more complex than univariate statistics. For this reason, the methods presented here are intended to be a useful resource for these studies. While they are applicable to GWAS, presenting algorithms whose computational complexity is hindered by modern data sets would only yield an unnecessary method in future studies. The work here is aimed at staying relevant for future studies as well.

5.1 Future Directions

While GWAS have revealed numerous meaningful genotype-phenotype relationships, they are readily being replaced by next-generation sequencing studies which aim at including rare variants and correcting other issues in GWAS. These studies will be larger than GWAS, while simultaneously posing a problem of how to decipher the complex relationships between the rare genetic variants and the disease phenotypes. This will require algorithms that are both scalable and capable of detecting complex epistasis interactions among the genetic variants.

The publications contained in this dissertation were meant to provide a framework on which others could build. By developing and adapting algorithms capable of feature selection on the large scale, it was aimed that this

work would remain relevant for researchers continuing in the field. These methods were not meant to be all-inclusive and rather were supposed to act as a base which could be expanded upon. An example would be the incorporation of pathway analysis during the feature selection stage. This would allow the knowledge of systems biologists to complement that of computer scientists. These types of advanced feature selection methodologies that make use of external annotation data would help control the false positive rates commonly seen in these studies and hopefully help to increase the generalization of the work, but come at a large increase in computational costs.

Further, as computer technologies advance, a new wave of computational methods will start to emerge in genetic research. Cloud computing has allowed for the rapid decrease in computational costs while increasing the work flow. With algorithms based on methods similar to the parallel one presented here it is likely that we will see a new wave of programs that aim to use more advanced search methodologies on NGS studies. However, this will need to come with an increased acceptance of commercial applications such as Mahout, through which the use of the Hadoop framework is capable of making use of these cloud technologies. Researchers should contribute their algorithms directly into these types of frameworks assuring that the research community and commercial industries as a whole may help them grow. Building algorithms from scratch is necessary for research, but we should make use of leading technologies to assure that our work will remain utilized and expanded upon.

Eventually, the aggregation of these genetic studies and the associated algorithms will help to create personalized medicine. This assures that medicines can be tailored for a specific individual, reducing side effects and increasing their effectiveness. The treatments for diseases would be specific for the affected individual and based on their individual genetic footprint. Further, it is hoped that disease risks would be identified at much earlier ages so that individuals could be more in control of the environmental factors that affect the onset. These advancements will not come immediately and it is only through the effort of countless individuals that medicine will eventually reach the pinnacle of personalize medicine.

Bibliography

[1] Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

[2] Abraham, G., Kowalczyk, A., Zobel, J., and Inouye, M. (2013). Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. *Genet. Epidemiol.*, 37(2):184–195.

[3] Ambroise, C. and McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10):6562–6566.

[4] Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., and Sorensen, D. (1999). *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition.

[5] Archer, K. J. (2010). rpartordinal: An r package for deriving a classification tree for predicting an ordinal response. *Journal of Statistical Software*, 34(7):1–17.

[6] Balaji, P., Buntinas, D., Goodell, D., Gropp, W., Kumar, S., Lusk, E., Thakur, R., and Träff, J. L. (2009). MPI on a million processors. In *Proceedings of the 16th European PVM/MPI Users' Group Meeting on Recent Advances in Parallel Virtual Machine and Message Passing Interface*, pages 20–30, Berlin, Heidelberg. Springer-Verlag.

[7] Barsh, G. S., Copenhaver, G. P., Gibson, G., and Williams, S. M. (2012). Guidelines for genome-wide association studies. *PLoS Genet*, 8(7):e1002812.

[8] Belloni, E., Veronesi, G., Rotta, L., Volorio, S., Sardella, D., Bernard, L., Pece, S., Di Fiore, P. P., Fumagalli, C., Barberis, M., Spaggiari, L.,

Pelicci, P. G., and Riva, L. (2015). Whole exome sequencing identifies driver mutations in asymptomatic computed tomography-detected lung cancers with normal karyotype. *Cancer Genet*, 208(4):152–155.

[9] Blackford, L. S., Choi, J., Cleary, A., D'Azevedo, E., Demmel, J., Dhillon, I., Dongarra, J., Hammarling, S., Henry, G., Petitet, A., Stanley, K., Walker, D., and Whaley, R. C. (1997). *ScaLAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA.

[10] Bloom, J. S., Ehrenreich, I. M., Loo, W., Lite, T.-L. V., and Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. *Nature*, 494(7436):234–237.

[11] Bradley, J. K., Kyrola, A., Bickson, D., and Guestrin, C. (2011). Parallel coordinate descent for l1-regularized loss minimization. In *International Conference on Machine Learning (ICML 2011)*, Bellevue, Washington.

[12] Brorsson, C., Hansen, N. T., Lage, K., Bergholdt, R., Brunak, S., and Pociot, F. (2009). Identification of T1D susceptibility genes within the MHC region by combining protein interaction networks and SNP genotyping data. *Diabetes Obes Metab*, 11 Suppl 1:60–66.

[13] Burgner, D., Davila, S., Breunis, W. B., Ng, S. B., Li, Y., Bonnard, C., Ling, L., Wright, V. J., Thalamuthu, A., Odam, M., Shimizu, C., Burns, J. C., Levin, M., Kuijpers, T. W., Hibberd, M. L., and International Kawasaki Disease Genetics Consortium (2009). A genome-wide association study identifies novel and functionally related susceptibility Loci for Kawasaki disease. *PLoS Genet*, 5(1):e1000319+.

[14] Burton, P. R., Clayton, D. G., Cardon, L. R., Craddock, N., Deloukas, P., Duncanson, A., Kwiatkowski, D. P., McCarthy, M. I., Ouwehand, W. H., Samani, N. J., Todd, J. A., Donnelly, P., and Et Al (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678.

[15] Buyyaz, R. (1999). *Parallel Programming Models and Paradigms*, chapter 1, pages 16–21. Prentice Hall, Upper Saddle River, NJ, USA, 1st edition.

[16] Cho, S., Kim, H., Oh, S., Kim, K., and Park, T. (2009). Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. *BMC Proc*, 3 Suppl 7:S25.

[17] Christodoulou, K., Wiskin, A. E., Gibson, J., Tapper, W., Willis, C., Afzal, N. A., Upstill-Goddard, R., Holloway, J. W., Simpson, M. A.,

Beattie, R. M., Collins, A., and Ennis, S. (2013). Next generation exome sequencing of paediatric inflammatory bowel disease patients identifies rare and novel variants in candidate genes. *Gut*, 62(7):977–984.

[18] Dudbridge, F. (2013). Power and predictive accuracy of polygenic risk scores. *PLoS Genet*, 9(3):e1003348+.

[19] Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: The .632 bootstrap method. *Journal of the American Statistical Association*, 92(438).

[20] Eleftherohorinou, H., Wright, V., Hoggart, C., Hartikainen, A.-L., Jarvelin, M.-R., Balding, D., Coin, L., and Levin, M. (2009). Pathway analysis of gwas provides new insights into genetic susceptibility to 3 inflammatory diseases. *PLoS ONE*, 4(11):e8068.

[21] Evans, D. M., Visscher, P. M., and Wray, N. R. (2009). Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Human molecular genetics*, 18(18):3525–3531.

[22] Frazer, K. A., Murray, S. S., Schork, N. J., and Topol, E. J. (2009). Human genetic variation and its contribution to complex traits. *Nat Rev Genet*, 10(4):241–251.

[23] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

[24] Gibson, G. (2010). Hints of hidden heritability in GWAS. *Nature genetics*, 42(7):558–560.

[25] González-Recio, O., de Maturana, E. L., Vega, A. T., Engelman, C. D., and Broman, K. W. (2009). Detecting single-nucleotide polymorphism by single-nucleotide polymorphism interactions in rheumatoid arthritis using a two-step approach with machine learning and a bayesian threshold least absolute shrinkage and selection operator (lasso) model. *BMC Proceedings*, 3(Suppl 7):S63.

[26] Greene, C. S., Kiralis, J., and Moore, J. H. (2009). Nature-inspired algorithms for the genetic analysis of epistasis in common human diseases: Theoretical assessment of wrapper vs. filter approaches. In *IEEE Congress on Evolutionary Computation*, pages 800–807. IEEE.

[27] Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182.

[28] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

[29] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

[30] Henderson, H. V. and Searle, S. R. (1981). On deriving the inverse of a sum of matrices. *SIAM Review*, 23(1):53–60.

[31] Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, 4(1):44–57.

[32] Huang, H., Tata, S., and Prill, R. J. (2013). Bluesnp: R package for highly scalable genome-wide association studies using hadoop clusters. *Bioinformatics*, 29(1):135–136.

[33] John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI'95, pages 338–345, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[34] Jones, E., Oliphant, T., Peterson, P., et al. (2001–). SciPy: Open source scientific tools for Python. [Online; accessed 2014-11-01].

[35] Kilpinen, H. and Barrett, J. C. (2013). How next-generation sequencing is transforming complex disease genetics. *Trends Genet*, 29(1):23–30.

[36] Kohannim, O., Hibar, D. P., Stein, J. L., Jahanshad, N., Hua, X., Rajagopalan, P., Toga, A. W., Jack, C. R., Weiner, M. W., de Zubicaray, G. I., McMahon, K. L., Hansell, N. K., Martin, N. G., Wright, M. J., and Thompson, P. M. (2012). Discovery and replication of gene influences on brain structure using LASSO regression. *Front Neurosci*, 6:115.

[37] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[38] Kooperberg, C., LeBlanc, M., and Obenchain, V. (2010). Risk prediction using genome-wide association studies. *Genetic epidemiology*, 34(7):643–652.

[39] Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *J Cheminform*, 6(1):10.

[40] Kruglyak, L. and Nickerson, D. A. (2001). Variation is the spice of life. *Nat Genet*, 27(3):234–236.

[41] Li, G., Ma, J., and Zhang, L. (2010). Greedy selection of species for ancestral state reconstruction on phylogenies: Elimination is better than insertion. *PLoS ONE*, 5(2):e8985+.

[42] Li, J., Das, K., Fu, G., Li, R., and Wu, R. (2011). The bayesian lasso for genome-wide association studies. *Bioinformatics*, 27(4):516–523.

[43] Liu, J., Zhang, C., McCarty, C. A., Peissig, P. L., Burnside, E. S., and Page, D. (2012). High-dimensional structured feature screening using binary markov random fields. In Lawrence, N. D. and Girolami, M., editors, *AISTATS*, volume 22 of *JMLR Proceedings*, pages 712–721. JMLR.org.

[44] Liu, S., Wang, H., Zhang, L., Tang, C., Jones, L., Ye, H., Ban, L., Wang, A., Liu, Z., Lou, F., Zhang, D., Sun, H., Dong, H., Zhang, G., Dong, Z., Guo, B., Yan, H., Yan, C., Wang, L., Su, Z., Li, Y., Huang, X. F., Chen, S. Y., and Zhou, T. (2015). Rapid detection of genetic mutations in individual breast cancer patients by next-generation DNA sequencing. *Hum. Genomics*, 9(1):2.

[45] Long, N., Gianola, D., Rosa, G. J., Weigel, K. A., and Avendano, S. (2009). Comparison of classification methods for detecting associations between SNPs and chick mortality. *Genet. Sel. Evol.*, 41:18.

[46] Lorenz, M. W., Markus, H. S., Bots, M. L., Rosvall, M., and Sitzer, M. (2007). Prediction of clinical cardiovascular events with carotid intima-media thickness: a systematic review and meta-analysis. *Circulation*, 115(4):459–467.

[47] Lücking, C. B., Durr, A., Bonifati, V., Vaughan, J., De Michele, G., Gasser, T., Harhangi, B. S., Meco, G., Denefle, P., Wood, N. W., Agid, Y., Brice, A., and on Genetic Susceptibility in Parkinson, T. E. C. (2000). Association between early-onset parkinson's disease and mutations in the parkin gene. *N Engl J Med*, 342(21):1560–1567.

[48] Luo, L., Peng, G., Zhu, Y., Dong, H., Amos, C. I., and Xiong, M. (2010). Genome-wide gene and pathway analysis. *European journal of human genetics : EJHG*, 18(9):1045–1053.

[49] Ma, L., Runesha, H. B., Dvorkin, D., Garbe, J. R., and Da, Y. (2008). Parallel and serial computing tools for testing single-locus and epistatic snp effects of quantitative traits in genome-wide association studies. *BMC Bioinformatics*, 9:315.

[50] Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature*, 456(7218):18–21.

[51] Makowsky, R., Pajewski, N. M., Klimentidis, Y. C., Vazquez, A. I., Duarte, C. W., Allison, D. B., and de los Campos, G. (2011). Beyond missing heritability: Prediction of complex traits. *PLoS Genet*, 7(4):e1002051.

[52] Malovini, A., Barbarini, N., Bellazzi, R., and Michelis, F. D. (2012). Hierarchical naive bayes for genetic association studies. *BMC Bioinformatics*, 13(S-14):S6.

[53] Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 11(7):499–511.

[54] Meder, B., Haas, J., Keller, A., Heid, C., Just, S., Borries, A., Boisguerin, V., Scharfenberger-Schmeer, M., Stahler, P., Beier, M., Weichenhan, D., Strom, T. M., Pfeufer, A., Korn, B., Katus, H. A., and Rottbauer, W. (2011). Targeted next-generation sequencing for the molecular genetic diagnostics of cardiomyopathies. *Circ Cardiovasc Genet*, 4(2):110–122.

[55] Moore, J. H., Asselbergs, F. W., and Williams, S. M. (2010). Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4):445–455.

[56] Naula, P., Airola, A., Salakoski, T., and Pahikkala, T. (2014). Multi-label learning under feature extraction budgets. *Pattern Recognition Letters*, 40:56–65.

[57] Nejentsev, S., Howson, J. M., Walker, N. M., Szeszko, J., Field, S. F., Stevens, H. E., Reynolds, P., Hardy, M., King, E., Masters, J., Hulme, J., Maier, L. M., Smyth, D., Bailey, R., Cooper, J. D., Ribas, G., Campbell, R. D., Clayton, D. G., and Todd, J. A. (2007). Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A. *Nature*, 450(7171):887–892.

[58] Okser, S., Airola, A., Aittokallio, T., Salakoski, T., and Pahikkala, T. (2013). Parallel feature selection for regularized least-squares. In Manninen, P. and Öster, P., editors, *Applied Parallel and Scientific Computing*, volume 7782 of *Lecture Notes in Computer Science*, pages 280–294. Springer Berlin Heidelberg.

[59] Okser, S., Lehtimäki, T., Elo, L. L., Mononen, N., Peltonen, N., Kähönen, M., Juonala, M., Fan, Y.-M., Hernesniemi, J. A., Laitinen, T., Lyytikäinen, L.-P., Rontu, R., Eklund, C., Hutri-Kähönen, N., Taittonen, L., Hurme, M., Viikari, J. S. A., Raitakari, O. T., and Aittokallio, T. (2010). Genetic variants and their interactions in the prediction of increased pre-clinical carotid atherosclerosis: The cardiovascular risk in young Finns study. *PLoS Genet*, 6(9):e1001146.

[60] Okser, S., Pahikkala, T., Airola, A., Aittokallio, T., and Salakoski, T. (2011). Fast and parallelized greedy forward selection of genetic variants in genome-wide association studies. In Chen, Y., Huang, Y., and Dougherty, E., editors, *IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS'11)*, pages 214–217. IEEE Signal Processing Society.

[61] Okser, S., Pahikkala, T., Airola, A., Salakoski, T., Ripatti, S., and Aittokallio, T. (2014). Regularized machine learning in the genetic prediction of complex traits. *PLoS Genet*, 10(11):e1004754.

[62] Okser, S., Pahikkala, T., and Aittokallio, T. (2013). Genetic variants and their interactions in disease risk prediction - machine learning and network perspectives. *BioData Mining*, 6(1):5.

[63] O'Leary, D. H., Polak, J. F., Kronmal, R. A., Manolio, T. A., Burke, G. L., and Wolfson, S. K. (1999). Carotid-artery intima and media thickness as a risk factor for myocardial infarction and stroke in older adults. Cardiovascular Health Study Collaborative Research Group. *N. Engl. J. Med.*, 340(1):14–22.

[64] Pahikkala, T., Airola, A., Naula, P., and Salakoski, T. (2010). Greedy RankRLS: a algorithm for learning sparse ranking models. In Gabrilovich, E., Smola, A. J., and Tishby, N., editors, *SIGIR 2010 Workshop on Feature Generation and Selection for Information Retrieval*, pages 11–18. ACM.

[65] Pahikkala, T., Airola, A., and Salakoski, T. (2010). Speeding up greedy forward selection for regularized least-squares. In Draghici, S., Khoshgoftaar, T. M., Palade, V., Pedrycz, W., Wani, M. A., and Zhu, X., editors, *Proceedings of The Ninth International Conference on Machine Learning and Applications (ICMLA'10)*. IEEE Computer Society.

[66] Pahikkala, T., Boberg, J., and Salakoski, T. (2006). Fast n-fold cross-validation for regularized least-squares. In Honkela, T., Raiko, T., Kortela, J., and Valpola, H., editors, *Proceedings of the Ninth Scandinavian Conference on Artificial Intelligence (SCAI 2006)*, pages 83–90, Espoo, Finland. Otamedia.

[67] Pahikkala, T., Okser, S., Airola, A., Salakoski, T., and Aittokallio, T. (2012). Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations. *Algorithms for Molecular Biology*, 7(1):11.

[68] Pahikkala, T., Suominen, H., and Boberg, J. (2012). Efficient cross-validation for kernelized least-squares regression with sparse basis expansions. *Machine Learning*, 87(3):381–407.

[69] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[70] Peise, E., Fabregat-Traver, D., and Bientinesi, P. (2015). High performance solutions for big-data gwas. *Parallel Computing*, 42:75 – 87.

[71] Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nat Rev Genet*, 11(7):459–463.

[72] Psychiatric GWAS Consortium Coordinating Committee and Lewis, C. (2009). Genomewide association studies: history, rationale, and prospects for psychiatric disorders. *The American Journal of Psychiatry*, 166(5):540 – 556.

[73] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3):559–575.

[74] R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[75] Rioux, J. D., Xavier, R. J., Taylor, K. D., Silverberg, M. S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M. M., Datta, L. W. W., Shugart, Y. Y. Y., Griffiths, A. M., Targan, S. R., Ippoliti, A. F., Bernard, E.-J. J., Mei, L., Nicolae, D. L., Regueiro, M., Schumm, L. P., Steinhart, A. H., Rotter, J. I., Duerr, R. H., Cho, J. H., Daly, M. J., and Brant, S. R. (2007). Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nature genetics*, 39(5):596–604.

[76] Rivas, M. A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C. K., Boucher, G., Ripke, S., Ellinghaus, D., Burtt, N., Fennell, T., Kirby, A., Latiano, A., Goyette, P., Green, T., Halfvarson, J., Haritunians, T., Korn, J. M., Kuruvilla, F., Lagacé, C., Neale, B., Lo, K. S. S., Schumm, P., Törkvist, L., National Institute of Diabetes and Digestive Kidney Diseases Inflammatory Bowel Disease Genetics Consortium (NIDDK IBDGC), United Kingdom Inflammatory Bowel Disease Genetics Consortium, International Inflammatory Bowel Disease Genetics Consortium, Dubinsky, M. C., Brant, S. R., Silverberg, M. S., Duerr, R. H., Altshuler, D., Gabriel, S., Lettre, G., Franke, A., D'Amato, M., McGovern, D. P., Cho, J. H., Rioux, J. D., Xavier, R. J., and Daly, M. J. (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature genetics*, 43(11):1066–1073.

[77] Robnik-Sikonja, M. and Kononenko, I. (2003). Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, 53(1-2):23–69.

[78] Roshan, U., Chikkagoudar, S., Wei, Z., Wang, K., and Hakonarson, H. (2011). Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest. *Nucleic Acids Res.*, 39(9):e62.

[79] Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517.

[80] Scherrer, C., Halappanavar, M., Tewari, A., and Haglin, D. (2012). Scaling up coordinate descent algorithms for large l1 regularization problems. In *Proceedings of the 29th International Conference on Machine Learning*.

[81] Schork, N. J., Murray, S. S., Frazer, K. A., and Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.*, 19(3):212–219.

[82] Sebastiani, P., Solovieff, N., and Sun, J. (2012). Naive bayesian classifier and genetic risk score for genetic risk prediction of a categorical trait: Not so different after all! *Front Genet*, 3(26).

[83] Sebban, M. and Nock, R. (2002). A hybrid filter/wrapper approach of feature selection using information theory. *Pattern Recognition*, 35(4):835–846+.

[84] Shi, G., Boerwinkle, E., Morrison, A. C., Gu, C. C., Chakravarti, A., and Rao, D. C. (2011). Mining gold dust under the genome wide signifi-

cance level: a two-stage approach to analysis of GWAS. *Genetic epidemiology*, 35(2):111–118.

[85] Sikorska, K., Lesaffre, E., Groenen, P. F. J., and Eilers, P. H. C. (2013). Gwas on your notebook: fast semi-parallel linear and logistic regression for genome-wide association studies. *BMC Bioinformatics*, 14:166.

[86] Tian, C., Gregersen, P. K., and Seldin, M. F. (2008). Accounting for ancestry: population substructure and genome-wide association studies. *Human Molecular Genetics*, 17(2):143–150.

[87] Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

[88] Tibshirani, R. J. and Tibshirani, R. (2009). A bias correction for the minimum error rate in cross-validation. *Ann. Appl. Stat.*, 3(2):822–829.

[89] van der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The numpy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, 13(2):22–30.

[90] Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*, 7(1):91.

[91] Waldmann, P., Meszaros, G., Gredler, B., Fuerst, C., and Solkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Front Genet*, 4:270.

[92] Wang, Y., Goh, W. W. B., Wong, L., and Montana, G. (2013). Random forests on hadoop for genome-wide association studies of multivariate neuroimaging phenotypes. *BMC Bioinformatics*, 14(S-16):S6.

[93] Watson, J. D. and Crick, F. H. C. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

[94] Wei, W., Visweswaran, S., and Cooper, G. F. (2011). The application of naive bayes model averaging to predict alzheimer's disease from genome-wide data. *JAMIA*, 18(4):370–375.

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

[96] Wei, Z., Wang, K., Qu, H.-Q., Zhang, H., Bradfield, J., Kim, C., Frackleton, E., Hou, C., Glessner, J. T., Chiavacci, R., Stanley, C., Monos, D., Grant, S. F. A., Polychronakos, C., and Hakonarson, H. (2009). From disease association to risk assessment: An optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet*, 5(10):e1000678.

[97] Witte, J. S., Visscher, P. M., and Wray, N. R. (2014). The contribution of genetic variants to disease depends on the ruler. *Nat Rev Genet*.

[98] Wu, C., DeWan, A., Hoh, J., and Wang, Z. (2011). A comparison of association methods correcting for population stratification in case–control studies. *Annals of Human Genetics*, 75(3):418–427.

[99] Yang, C., Wan, X., Yang, Q., Xue, H., and Yu, W. (2010). Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group Lasso. *BMC bioinformatics*, 11 Suppl 1.

[100] Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., and Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569.

[101] Yip, W.-K. and Lange, C. (2011). Quantitative trait prediction based on genetic marker-array data, a simulation study. *Bioinformatics*, 27(6):745–8.

[102] Yung, L. S., Yang, C., Wan, X., and Yu, W. (2011). Gboost: a gpu-based tool for detecting gene–gene interactions in genome–wide case control studies. *Bioinformatics*, 27(9):1309–1310.

[103] Zhang, J., Chiodini, R., Badr, A., and Zhang, G. (2011). The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics*, 38(3):95–109.

[104] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2):301–320.

Publication Reprints

Publication I

Genetic Variants and Their Interactions in the Prediction of Increased Pre-Clinical Carotid Atherosclerosis: The Cardiovascular Risk in Young Finns Study

Okser, S., Lehtimäki, T., Elo, L. L., Mononen, N., Peltonen, N., Kähönen, M., Juonala, M., Fan, Y.-M., Hernesniemi, J. A., Laitinen, T., Lyytikäinen, L.-P., Rontu, R., Eklund, C., Hutri-Kähönen, N., Taittonen, L., Hurme, M., Viikari, JS., Raitakari, O. T., Aittokallio, T. (2010). *PLoS Genet*, 6(9):e1001146.

OPEN ACCESS Freely available online

PLoS GENETICS

Genetic Variants and Their Interactions in the Prediction of Increased Pre-Clinical Carotid Atherosclerosis: The Cardiovascular Risk in Young Finns Study

Sebastian Okser[1☯], Terho Lehtimäki[2☯], Laura L. Elo[1,3], Nina Mononen[2], Nina Peltonen[2], Mika Kähönen[4], Markus Juonala[5,6], Yue-Mei Fan[2], Jussi A. Hernesniemi[2], Tomi Laitinen[7], Leo-Pekka Lyytikäinen[2], Riikka Rontu[2], Carita Eklund[8], Nina Hutri-Kähönen[9], Leena Taittonen[10], Mikko Hurme[8], Jorma S. A. Viikari[5,11], Olli T. Raitakari[6,12], Tero Aittokallio[1,3]*

1 Biomathematics Research Group, Department of Mathematics, University of Turku, Turku, Finland, **2** Department of Clinical Chemistry, Tampere University Hospital and University of Tampere, Tampere, Finland, **3** Data Mining and Modeling Group, Turku Centre for Biotechnology, Turku, Finland, **4** Department of Clinical Physiology, Tampere University Hospital and University of Tampere, Tampere, Finland, **5** Department of Medicine, Turku University Central Hospital, Turku, Finland, **6** Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku, Finland, **7** Department of Clinical Physiology and Nuclear Medicine, Kuopio University Hospital and University of Eastern Finland, Kuopio, Finland, **8** Department of Microbiology and Immunology, University of Tampere, Tampere, Finland, **9** Department of Pediatrics, Tampere University Hospital, Tampere, Finland, **10** Department of Pediatrics, University of Oulu, Oulu, Finland, **11** Department of Medicine, University of Turku, Turku, Finland, **12** Department of Clinical Physiology, Turku University Hospital, Turku, Finland

Abstract

The relative contribution of genetic risk factors to the progression of subclinical atherosclerosis is poorly understood. It is likely that multiple variants are implicated in the development of atherosclerosis, but the subtle genotypic and phenotypic differences are beyond the reach of the conventional case-control designs and the statistical significance testing procedures being used in most association studies. Our objective here was to investigate whether an alternative approach—in which common disorders are treated as quantitative phenotypes that are continuously distributed over a population—can reveal predictive insights into the early atherosclerosis, as assessed using ultrasound imaging-based quantitative measurement of carotid artery intima-media thickness (IMT). Using our population-based follow-up study of atherosclerosis precursors as a basis for sampling subjects with gradually increasing IMT levels, we searched for such subsets of genetic variants and their interactions that are the most predictive of the various risk classes, rather than using exclusively those variants meeting a stringent level of statistical significance. The area under the receiver operating characteristic curve (AUC) was used to evaluate the predictive value of the variants, and cross-validation was used to assess how well the predictive models will generalize to other subsets of subjects. By means of our predictive modeling framework with machine learning-based SNP selection, we could improve the prediction of the extreme classes of atherosclerosis risk and progression over a 6-year period (average AUC 0.844 and 0.761), compared to that of using conventional cardiovascular risk factors alone (average AUC 0.741 and 0.629), or when combined with the statistically significant variants (average AUC 0.762 and 0.651). The predictive accuracy remained relatively high in an independent validation set of subjects (average decrease of 0.043). These results demonstrate that the modeling framework can utilize the "gray zone" of genetic variation in the classification of subjects with different degrees of risk of developing atherosclerosis.

Citation: Okser S, Lehtimäki T, Elo LL, Mononen N, Peltonen N, et al. (2010) Genetic Variants and Their Interactions in the Prediction of Increased Pre-Clinical Carotid Atherosclerosis: The Cardiovascular Risk in Young Finns Study. PLoS Genet 6(9): e1001146. doi:10.1371/journal.pgen.1001146

Editor: Nicholas J. Schork, University of California San Diego and The Scripps Research Institute, United States of America

Received December 18, 2009; **Accepted** September 1, 2010; **Published** September 30, 2010

Funding: This work was financially supported by the Academy of Finland (grants 77841, 117832, 117941, 201888, 120569, 133227, 127575, 121584, 126925), the Social Insurance Institution of Finland, Turku University Foundation, Kuopio, Tampere and Turku University Hospital Medical Funds, Emil Aaltonen Foundation, Juho Vainio Foundation, Finnish Foundation of Cardiovascular Research and Finnish Cultural Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: tero.aittokallio@utu.fi

☯ These authors contributed equally to this work.

Introduction

A major challenge of medical genetics is to determine an optimal set of genetic markers, typically in the form of single nucleotide polymorphisms (SNP), which when combined together with conventional risk factors, could be used in individual level risk prediction, classification and clinical decision-making. However, genome-wide association studies (GWAS) have demonstrated that the ubiquitous heritability of most common disorders is due to multiple SNPs of small effect size and even an aggregate of these effects is not yet predictive enough for clinical utility [1]. It has therefore been suggested that the traditional case-control studies, which focus on qualitative phenotypes such as diagnosed cases versus controls, could be complemented by population-based cohort studies, which profile quantitative clinical phenotypes and how they change over time in individuals who are representative of

Author Summary

Although cardiovascular events, such as myocardial infarction and stroke, usually occur at later ages, it is known that the atherogenic process begins much earlier in life. Detection of subclinical atherosclerosis would therefore offer the means to identify individuals who are at increased risk of developing cardiovascular events. What remains unclear is the relative contribution of genetic variation to the development of the early stages of atherosclerosis. To address this question, we searched for combinations of both genetic and clinical determinants that are the most predictive of the progression of subclinical carotid atherosclerosis in a sample of 1,027 young adults, aged between 24–39 years, from the Finnish general population (The Cardiovascular Risk in Young Finns Study). We demonstrate here, for the first time in a population-based follow-up study, a predictive relationship between individual's genotypic variation and early signs of atherosclerosis, which cannot be explained by conventional cardiovascular risk factors, such as obesity and elevated blood pressure levels. The predictive modeling framework facilitates the usability of genetic information by identifying informative panels of variants, along with conventional risk factors, which may prove to be useful in early detection and management of atherosclerosis. The clinical implications of these findings remain to be studied.

the general population. Consequently, certain common disorders may be interpreted as being the extremes of the quantitative phenotypes that are continuously distributed over the population [1]. Comparing various ranges of the low and high extremes of such quantitative traits, rather than dichotomizing the same distribution exclusively into cases and controls, can offer the means to increase the statistical power of the variants [2–5], uncover molecular pathways and networks behind various subtypes and progression stages [6], and eventually even help to improve the early diagnosis, treatment and prevention of the most extreme cases. The objective here was to systematically investigate the potential of this extreme selection strategy to provide predictive insights into the early development of atherosclerosis, using the carotid IMT as a quantitative phenotype and our unique population-based follow-up study of atherosclerosis precursors as a basis for sub-sampling of subjects with increasing disease risk.

Atherosclerosis is a common disorder which develops due to the complex interplay of various genetic and environmental factors, most of which are still poorly understood. It is known that conventional cardiovascular risk factors, such as obesity, elevated blood pressure and high low-density lipoprotein (LDL) cholesterol levels, play an important role in the risk of its progression into severe clinical manifestations, for instance, coronary heart disease (CHD) [7,8]. Recently, a number of genetic risk markers that associate with coronary disease outcomes and serum lipid concentrations have also been identified in case-control settings [9–21]. However, the relative contribution of genetic variation to the early stages of the cardiovascular disease remains unclear. From the experimental design point of view, the subtle inter-individual phenotypic variability makes it difficult to prognosticate clear-cut cases and controls in a pre-clinical setting, thereby limiting the capability of the cross-sectional case-control designs in distinguishing the variants associated with an increased progression risk from the background variability. An additional challenge is that even in the absence of significant single-marker effects, multiple genetic markers from distinct molecular pathways may act synergistically when combined, leading to different atherosclerosis phenotypes. Confounding inter-individual variation and interactions across the genetic and conventional risk factors can also mask the phenotypic variation, especially when studying composite phenotypes such as LDL-cholesterol levels [22]. Therefore, a well-defined quantitative measurement that reflects the full spectrum of the disease progression is needed, together with an efficient computational approach, to systematically explore the genotype-phenotype relationships across different development stages of atherosclerosis.

Measurement of the carotid artery intima-media thickness (IMT) is an established, intermediate phenotype of atherosclerosis that has been used, for instance, to investigate the development of pre-clinical atherosclerosis [23,24], and to predict the onset of future cardiovascular events, such as myocardial infraction and stroke [25–27]. It can be measured non-invasively through the use of ultrasound imaging in large populations of healthy subjects, without the biases related to clinically diagnosed cases and controls [28], making it an ideal quantitative measurement for stratifying subjects into various risk classes. However, comparisons of such risk classes using statistical significance testing procedures that consider only one SNP at a time may yield sub-optimal findings when exploring the genotype-specific effects of large number of SNPs, given that these modest phenotypic effects are likely to be characterized by substantial genetic heterogeneity among multiple variants [29–31]. Accordingly, it has been argued that the statistics being used to identify variants that are significantly associated with the disease risk - typically odds ratios or *p*-values for association - are not the most appropriate means for evaluating the predictive or clinical value of the genetic profiles [32,33]. For example, the individual SNPs with the strongest statistical support in coronary artery disease-related case-control studies seem to have only a minor, if any, role in predicting carotid IMT or its progression, when compared to the conventional risk factors [34,35]. In fact, these susceptibility variants are able to provide only a marginal and inconsistent improvement even in the discrimination of the CHD cases or prediction of cardiovascular events [36–41], thus hindering the value of these 'top hits' for diagnostic prediction. Moreover, additional challenges stem from the identification of gene-gene and gene-environment interactions, which are thought to be profoundly important in the development of many complex diseases [29,30,42].

In the present analysis from the Young Finns Study, we took a more holistic approach towards revealing the contribution of genetic variation to the early progression of atherosclerosis. The approach was based on a stratified sampling and comparison of the increasing risk classes from our longitudinal population cohort. Rather than using the conventional single-SNP statistical significance testing in the identification of risk-modifying variants and their interactions, we explicitly searched for those subsets of SNPs that are the most predictive of the increasing risk classes by means of a predictive modeling framework using a machine learning-based SNP-subset selection procedure. The predictive approach was used here to mine those associations that did not necessary meet the stringent levels of statistical significance at the level of individual SNPs, yet still having significant contribution to the combined predictive power at the level of SNP-subsets. In particular, we addressed the following questions: (*i*) whether the genetic variants can improve the prediction accuracy of IMT-based risk classes beyond that obtained with conventional risk factors; (*ii*) which variants are the most predictive of the subjects that show extreme IMT levels either at the baseline or in the follow-up study, or progression over the 6-year period; (*iii*) whether the predictive SNP-panels also include other variants than those

risk markers identified in the previous case-control association studies; and (*iv*) whether the machine learning-based SNP selection can provide variants with increased predictive power compared to the SNPs with the greatest statistical significance in the present study population. We also illustrate how the predictive modeling framework can be employed to identify epistasis interactions among genetic variants that are related to the disease progression. Finally, as the first step toward elucidating functional mechanisms behind the genetic variants and their interactions, we also mapped the biological pathways and processes that underlie those variants most predictive of the extreme progression cases.

Results

The baseline study cohort in 2001 was comprised of 1,027 subjects from the Finnish general population, aged 24–39 years, with complete data including both the ultrasound-based imaging of the carotid IMT and the blood sample-based genotyping of the candidate SNPs (see Table S1); of these subjects, 813 also participated in the 2007 follow-up study of the IMT progression (see Materials and Methods for details). The relative contribution of the SNPs to the individual IMT levels was evaluated by means of a predictive modeling framework, in which the study subjects were first divided into gradually increasing low-risk and high-risk classes according to the quantile points, say (1-q) and q, of their pooled IMT distribution (q ranges from 5% to 25%; see Figure 1). A non-linear Bayesian classifier was implemented here as the predictive model (see Materials and Methods for details). Using both the genetic and conventional risk factors collected in the baseline study in 2001 as predictor variables, we determined the most predictive risk factor combinations separately for both the 2001 and 2007 IMT risk classes, as well as the IMT progression between 2001 and 2007. For a comparison, the most significant genetic variants were determined using single-SNP statistical testing for the same risk classes. The area under the receiver operating characteristic curve (AUC), with cross-validation, was used to evaluate the predictive value of the different factor combinations, followed by independent validation set-based assessment of how well the predictive models can generalize to independent sets of subjects.

Clinical characteristics of the study subjects

The quantitative distributions of the levels of IMT and its progression over the 6-year period are shown in Figure 1. The IMT levels in the study population showed a slightly positive-skewed enrichment of subjects with higher IMT values indicating an increased risk of atherosclerosis (Figure 1A). There was a significant difference in the IMT distributions between the 2001 and 2007 follow-up studies (Kolmogorov-Smirnov $D = 0.234$, $p < 0.001$). As expected, the majority of the conventional risk factors measured in 2001, including age, sex and BMI, were strongly correlated with the

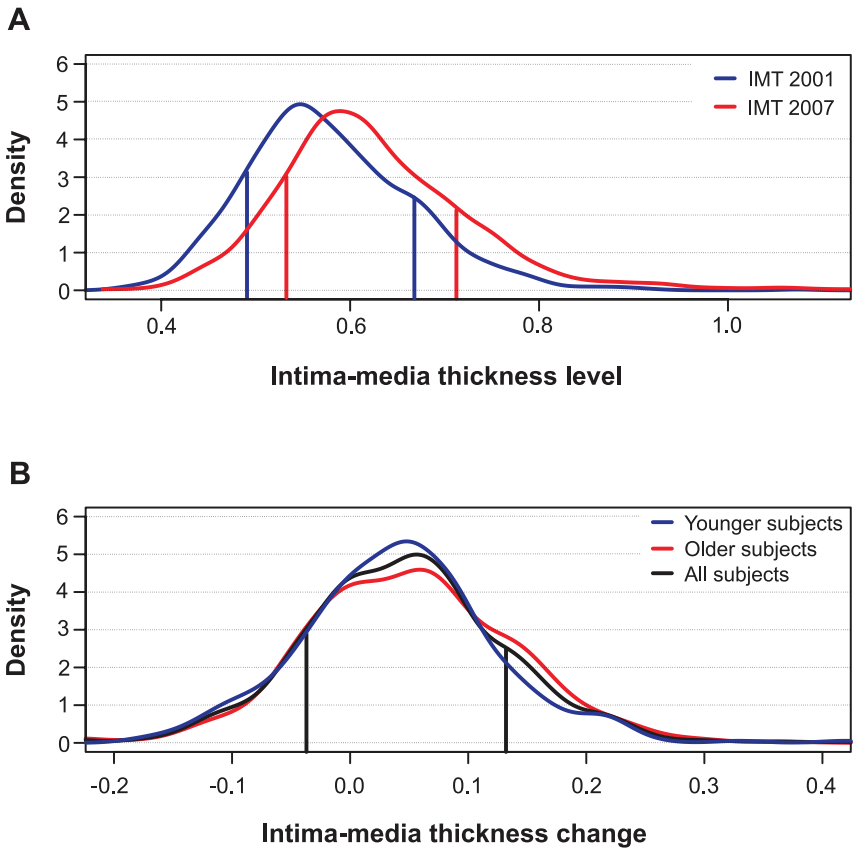

Figure 1. Distributions of intima-media thickness (IMT) of the study subjects. (A) IMT levels in the baseline and follow-up studies in 2001 and 2007, respectively. (B) IMT changes from 2001 to 2007. The age-stratified distributions depict the baseline age groups of 24–30 and 33–39 years (Younger and Older subjects), as well as their combined distribution (All subjects). The vertical lines indicate the representative 15% and 85% quantile points (q) that divide the subjects into two risk groups: the low-risk class (subjects with the lowest q% of IMT levels or changes) and the high-risk class (subjects with the highest q% of IMT levels or changes).
doi:10.1371/journal.pgen.1001146.g001

IMT levels both in the 2001 and 2007 studies (Table 1). However, only two risk factors, waist circumference and apolipoprotein B (ApoB), correlated with the IMT progression (the 2007-2001 change). In particular, even if the age was the most significant correlate of the IMT levels in 2001 and 2007, its linear explanatory power turned out to be insignificant for the IMT progression. Accordingly, the distributions of the IMT progression were similar in the groups of younger and older subjects ($D = 0.0791$, $p > 0.10$; Figure 1B). To keep the non-linear prediction problem as general as possible, the age-groups and sexes were pooled into a single continuous distribution; however, all the predictive models were adjusted for the baseline conventional risk factors (Table 1). This enabled us to examine, for instance, the added contribution of genetic variation to the IMT progression not explained by the variation in the conventional cardiovascular risk factors.

Prediction of baseline IMT using genetic variants

To assess whether the genetic variants can increase the prediction accuracy of the risk classes beyond that obtained with the conventional risk factors alone, we used the predictive modeling framework with a machine learning-based SNP selection. The predictive risk factor combinations selected using this procedure were able to significantly improve the prediction of the subjects across the spectrum of low-risk and high-risk classes in 2001 (Figure 2A), when compared to using the conventional risk factors (CRFs) either alone or combined with those SNPs that were significantly associated with the low- and high-risk differences in the study subjects (the significances of the SNPs are detailed in Table S2). Interestingly, the panel of genetic risk markers established in the previous case-control association studies alone had a predictive power similar to that of a random classifier (average AUC 0.489), and these SNPs could not improve the prediction of the IMT risk classes over and above of the conventional risk factors (Established SNPs and CRFs; Figure 2). As expected, the predictive accuracy gradually decreased when moving from 5% to 25% quantile level, as the risk classes became phenotypically more heterogeneous in terms of the quantitative IMT-levels (see Figure 1A). The variants most predictive of the subjects with 15% of the lowest and highest IMT-levels in 2001 are listed in Table 2, together with their gene annotation information and the single-SNP statistical and predictive powers.

Prediction of follow-up IMT using genetic variants

The predictive power of the genetic variants that were selected using the machine learning-based procedure increased further when predicting the risk classes in the 2007 follow-up, even if the genetic and conventional risk factors collected in only the baseline study were used as predictors (Figure 2B). This result can partly be attributed to the progression of the disease condition over the six years in a part of the study subjects (see Figure 1A). In particular, the classes of the most extreme levels of the IMT could be predicted with reasonably high accuracy also using single-SNP statistical testing, whereas the panel of established SNPs either with or without the conventional risk factors again showed much poorer performance (Figure 2B). These results suggest that the genetic variants, especially those that were identified using the machine learning-based SNP selection (see Table 3), can encode significant information according to which it is possible to predict subjects who will belong to different risk classes later in their lives with accuracies beyond that obtained with the conventional risk factors. We note that the baseline 2001 IMT-level was not used in the reported results when predicting the 2007 risk classes; however, in the case when the baseline IMT-level was used as an additional predictor, the prediction accuracies became very close to perfect discrimination (AUC ranged from 0.920 to 0.999). This shows that the non-linear modeling approach could learn also the significant linear correlation between the 2001 and 2007 IMT-levels ($r = 0.582$; Table S3).

Genetic variants predisposing to IMT progression

We next searched explicitly for those factors that are most predictive of the subjects who show extreme progression in their

Table 1. The baseline characteristics in 2001 along with their correlations with the 2007 level and progression of intima-media thickness (IMT).

Conventional Risk Factor*	Mean (SD)	IMT 2001		IMT 2007		IMT Progression	
		r†	p‡	r†	p‡	r†	p‡
Sex (% women)	55.3	0.132	<0.001	0.195	<0.001	0.086	NS
Age in 2001 (years)	31.7 (4.92)	0.290	<0.001	0.301	<0.001	0.041	NS
BMI (kg/m^2)	25.2 (4.38)	0.152	<0.001	0.188	<0.001	0.094	NS
Waist circumference (mm)	84.0 (12.0)	0.189	<0.001	0.260	<0.001	0.133	0.006
Systolic blood pressure (mmHg)	117 (13.2)	0.180	<0.001	0.158	<0.001	0.044	NS
Diastolic blood pressure (mmHg)	70.6 (10.5)	0.220	<0.001	0.160	<0.001	−0.020	NS
Total cholesterol (mmol/L)	5.17 (0.99)	0.113	0.011	0.155	<0.001	0.082	NS
LDL cholesterol (mmol/L)	3.28 (0.86)	0.126	0.002	0.166	<0.001	0.087	NS
HDL cholesterol (mmol/L)	1.29 (0.32)	−0.037	NS	−0.107	NS	−0.089	NS
Triglycerides (mmol/L)	1.35 (0.86)	0.047	NS	0.131	0.007	0.099	NS
ApoA1 (g/L)	1.49 (0.26)	−0.052	NS	−0.085	NS	−0.039	NS
ApoB (g/L)	1.06 (0.27)	0.110	0.016	0.195	<0.001	0.138	0.003
Smoking (% subjects)	22.8	0.049	NS	0.007	NS	−0.011	NS

*The characteristics in 2001 were used as potential confounding risk factors in predictive models.
†Pearson correlation coefficient (r-value) was calculated using the risk factors collected in 2001.
‡Statistical significance (Bonferroni corrected p-value) is from the t-distribution with n-2 df ($n = 1,027$ in 2001 and $n = 813$ in 2007); NS, non-significant.
doi:10.1371/journal.pgen.1001146.t001

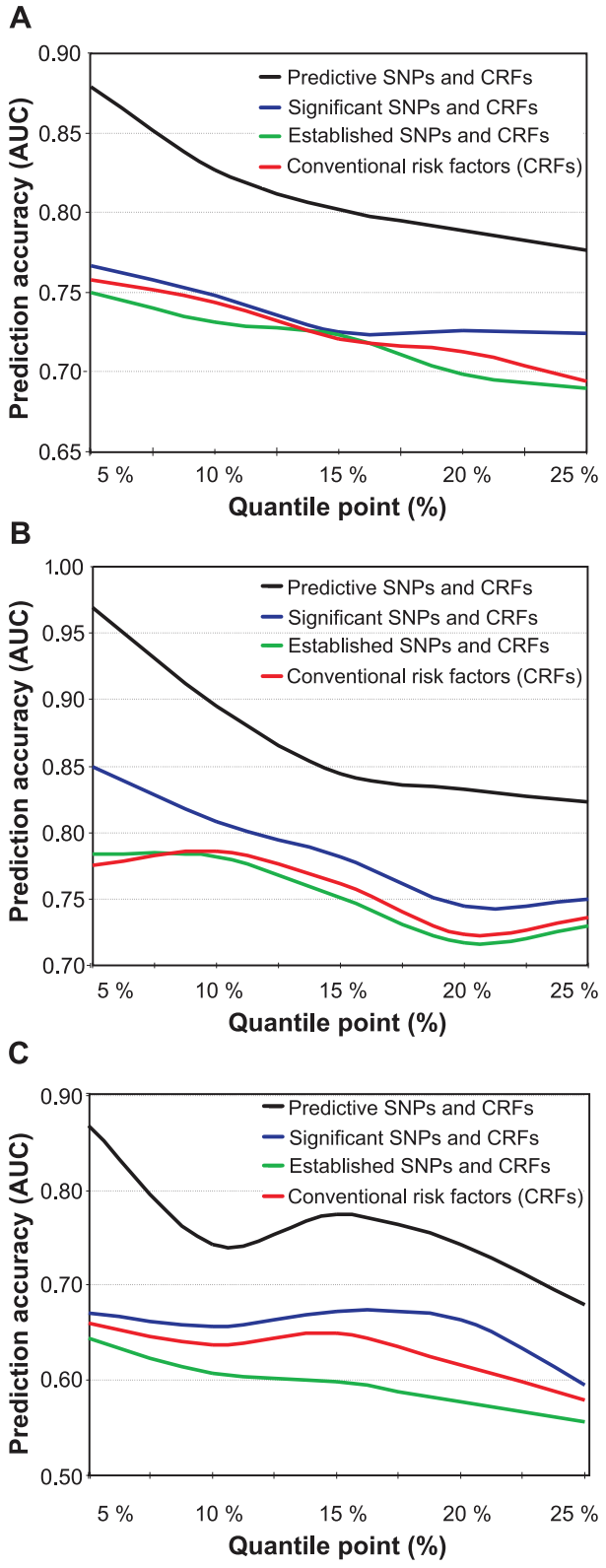

Figure 2. Prediction accuracy as a function of increasing risk classes. The accuracy was defined using the area under the receiver operating characteristic curve (AUC), and the risk classes using the quantile points (5–25%). (A) Prediction of the baseline IMT risk classes in 2001 when using the conventional risk factors either alone, or when combined with the panel of 17 SNPs associated in previous studies with cardiovascular morbidity (Established SNPs), with those SNPs that are significantly associated with the low- and high-risk classes (Significant SNPs), or with the most predictive SNPs identified using the machine learning-based approach (Predictive SNPs). (B) Prediction of the follow-up IMT risk classes in 2007 using the baseline conventional and genetic risk factors measured in 2001. (C) Prediction of the IMT progression risk classes when using the baseline conventional and genetic risk factors measured in 2001 (the same as in (A,B)).
doi:10.1371/journal.pgen.1001146.g002

IMT-levels between the two follow-up studies. When applying the machine learning-based procedure to prediction of the subjects with increasing changes in their IMT-levels between the study years 2001 and 2007, the selected SNPs could again systematically increase the predictive power across all the progression risk classes, compared to the accuracy obtained with the conventional risk factors either alone or when combined with the panels of variants identified in the previous case-control studies or in the present study population using single-SNP statistical testing (Figure 2C). In this case, however, the prediction accuracies were not anymore monotonically decreasing functions of the quantile point (q). In particular, the 10% risk class was found to be problematic, which could be due to the particular IMT cutoff values used in its quantitative definition. Interestingly, the SNP set most predictive of the IMT progression contained a relatively large number of variants with modest contributions to the predictive power; of these variants, only one was among the established markers identified in the previous case-control studies (Table 4). Even if the IMT progression proved relatively difficult to predict, the many novel markers support the potential and added value of genetic variation, especially when evaluating susceptibility to the most extreme progression risk class ($q = 5\%$).

Epistasis interactions between the predictive variants

To identify candidate epistasis (or synergistic) interactions between the genetic risk factors, we searched for such pairs of genetic variants that led to the largest drop in the prediction accuracy when removed together from the set of predictive SNPs, relative to the drop resulting from removing either of the variants separately. As a feasibility study, we explored the particular SNP set which was found to be highly predictive of the subjects with the most extreme IMT progression from 2001 to 2007 (Figure 2C, $q = 5\%$). When investigating a specific variant (rs2516839) in the upstream stimulatory factor 1 (USF1), a known regulator of the transcription of several cardiovascular-related genes, we identified a number of potential genetic interaction partners of USF1 (Figure 3), including formin 2 (FMN2, rs17672135), protein tyrosine phosphatase, non-receptor type 22 (PTPN22, rs2476601), hepatic triglyceride lipase (LIPC, rs1800588), and arachidonate 5-lipoxygenase-activating protein (ALOX5AP, rs17222814). It is interesting to note that each of these candidate gene-gene interactions originated from different biological processes, indicating that the disease progression and phenotypic heterogeneity is likely due to genetic alterations within multiple molecular pathways (Table S4). Such interactions and pathways may serve as basis for more detailed further studies of the molecular mechanisms and disease networks that predispose to such excess levels of the IMT-progression that can lead to clinical cardiovascular events in the future.

Evaluation on independent and randomized subject sets

To further explore the generalization capability of the prediction models estimated and evaluated on the current study subjects, we constructed a separate validation set consisting of those subjects who

Table 2. The single nucleotide polymorphisms (SNPs) predictive of the subjects with 15% lowest and highest IMT levels in 2001.

SNP ID* (rs number)	Gene symbol (HGNC name)	SNP location (Chr region)	Significance† (*p*-value)	Predictive power‡ (%AUC)
rs2073658	USF1	1q23.3	0.70	11.8
rs1205	CRP	1q23.2	0.02	10.6
rs805305	DDAH2	6p21.33	0.38	9.68
rs3890182	ABCA1	9q31.1	0.81	7.53
rs6929137	C6orf97	6q25.1	0.10	7.53
rs4073307	IGSF1	Xq26.1	0.71	6.45
rs693	APOB	2p24.1	0.53	6.45
rs3130340	INTERGENIC	6p21.32	0.11	6.45
rs599839	PSRC1	1p13.3	0.10	6.45
rs754523	INTERGENIC	2p24.1	1.00	5.38
rs1143634	IL1B	2q13	0.51	5.38
rs4404254	ICOS	2q33.2	0.16	4.30
rs2548861	WWOX	16q23.1	0.14	4.30
rs2553268	WRN	8p12	0.15	3.23
rs4937100	IL18	11q23.1	0.22	2.15
rs2516839	USF1	1q23.3	0.13	2.15

*The SNPs identified also in the previous case-control association studies [9–21] are boldfaced.
†The corrected *p*-values larger than one were truncated to unity.
‡The SNPs are arranged according to their contribution to the overall prediction accuracy (AUC).
doi:10.1371/journal.pgen.1001146.t002

were filtered out in the initial subject selection because of missing data, but had a complete set of those SNPs identified for the particular risk class (see Figure S1). These new subjects were then split into the classes of 'low-risk' and 'high-risk' based on the exact same IMT-cutoff values that were used in the original subjects. In general, the results in the independent validation set scaled as expected (Figure 4). Even if the prediction of the new subject classes using those SNPs identified in the original dataset led to decreased prediction accuracies (average decrease in AUC was 0.043), their prediction capability was shown to extend beyond the original subjects, especially for the extreme 5% IMT cases, whereas the 10% risk class again showed poorer performance. A part of the decreased accuracy can be attributed to the sensitivity of the extreme selection strategy to the particular IMT quantile cut-offs being used (the dotted trace). We also repeated the same model building and evaluation framework for randomized datasets, in which subjects

Table 3. The single nucleotide polymorphisms (SNPs) predictive of the subjects with 15% lowest and highest IMT levels in 2007.

SNP ID* (rs number)	Gene symbol (HGNC name)	SNP location (Chr region)	Significance† (*p*-value)	Predictive power‡ (%AUC)
rs17672135	FMN2	1q43	0.41	17.5
rs9941339	CDH13	16q24.2-q24.3	0.75	8.75
rs2548861	WWOX	16q23.1	0.14	8.75
rs9939609	FTO	16q12.2	0.69	7.50
rs693	APOB	2p24.1	0.53	7.50
rs17222814	ALOX5AP	13q12.3	0.89	7.50
rs1041981	LTA	6p21.33	1.00	7.50
rs9551963	ALOX5AP	13q12.3	0.64	6.25
rs7524102	INTERGENIC	1p36.12	0.77	5.00
rs2516839	USF1	1q23.3	0.13	5.00
rs2301880	WNK1	12p13.33	1.00	5.00
rs7759938	INTERGENIC	6q21	0.12	3.75
rs9479055	C6orf97	6q25.1	0.40	3.75
rs3130340	INTERGENIC	6p21.32	0.11	3.75
rs2553268	WRN	8p12	0.15	2.50

*The SNPs identified also in the previous case-control association studies [9–21] are boldfaced.
†The corrected *p*-values larger than one were truncated to unity.
‡The SNPs are arranged according to their contribution to the overall prediction accuracy (AUC).
doi:10.1371/journal.pgen.1001146.t003

Table 4. The single nucleotide polymorphisms (SNPs) predictive of the subjects with 15% lowest and highest IMT changes from 2001 to 2007.

SNP ID* (rs number)	Gene symbol (HGNC name)	SNP location (Chr region)	Significance† (*p*-value)	Predictive power‡ (%AUC)
rs2073658	USF1	1q23.3	0.70	9.40
rs9479055	C6orf97	6qs25.1	0.40	8.55
rs17672135	FMN2	1q43	0.41	8.55
rs9687339	MAST4	5q12.3	0.93	7.69
rs1042713	ADRB2	5q33.1	0.48	7.69
rs2301880	WNK1	12p13.33	1.00	6.84
rs3130340	INTERGENIC	6p21.32	0.11	6.84
rs2476601	PTPN22	1p13.2	0.44	5.13
rs11898505	SPTBN1	2p16.2	0.27	5.13
rs3798220	LPA	6q25.3	1.00	5.13
rs10172036	ICOS	2q33.2	0.52	5.13
rs2820037	INTERGENIC	1q43	0.66	4.27
rs2234693	ESR1	6q25.1	0.74	3.42
rs1800896	IL10	1q32.1	0.71	3.42
rs17222814	ALOX5AP	13q12.3	0.89	3.42
rs1801274	FCGR2A	1q23.3	0.75	2.56
rs854560	PON1	7q21.3	0.81	1.71
rs10246939	TAS2R38	7q34	0.80	1.71
rs9594738	INTERGENIC	13q14.11	0.58	1.71
rs1799983	NOS3	7q36.1	0.06	0.855
rs1256049	ESR2	14q23.2	0.46	0.855

*The SNPs identified also in the previous case-control association studies [9–21] are boldfaced.
†The corrected *p*-values larger than one were truncated to unity.
‡The SNPs are arranged according to their contribution to the overall prediction accuracy (AUC).
doi:10.1371/journal.pgen.1001146.t004

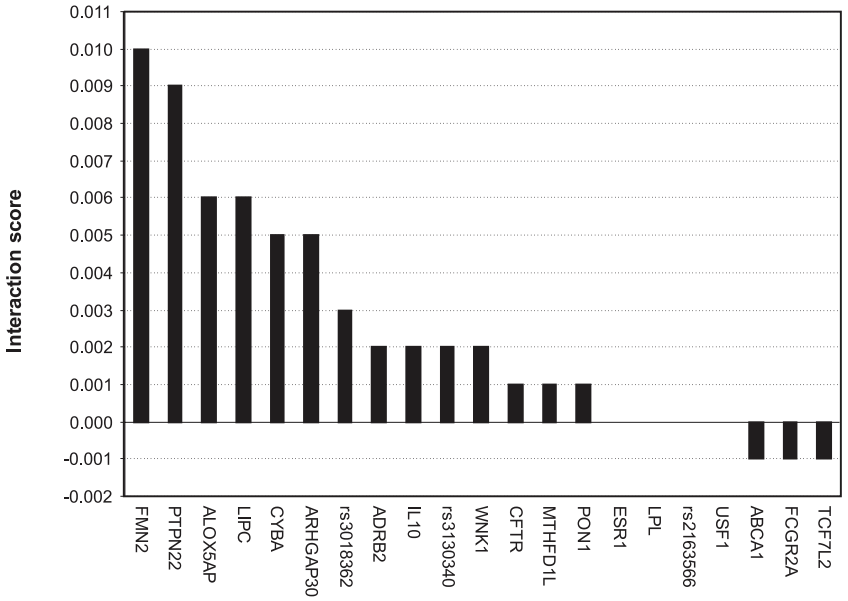

Figure 3. Candidate interaction partners of a variant in USF1 (rs2516839). The candidate SNP-SNP interactions were searched among the variants predictive of the extreme IMT progression (see Table S4). The interaction score for a SNP-pair (*x,y*) is $P_{x,y} - (P_x + P_y)$, depicting the combined contribution of the SNP-pair to the predictive power ($P_{x,y}$), relative to that of the individual SNPs' contributions (P_x and P_y). The predictive power was assessed in terms of how much the AUC value changed when the particular SNP or SNP-pair was deleted from the subset of variants. The Gene ID was used as a SNP identifier, where available; otherwise, the rs ID was used instead.
doi:10.1371/journal.pgen.1001146.g003

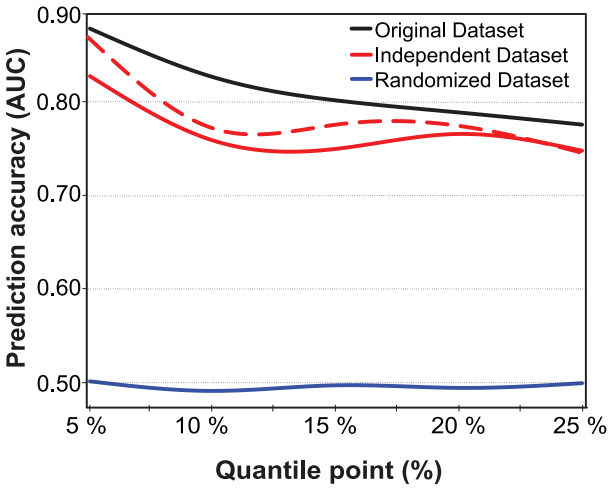

Figure 4. Prediction accuracies on independent and randomized subject sets. The accuracy was defined using the area under the receiver operating characteristic curve (AUC), and the risk classes using the quantile points (5%–25%). The prediction accuracies were evaluated for the baseline IMT risk classes in the independent dataset, in comparison with the cross-validated accuracies obtained in the original dataset using the same IMT thresholds, conventional risk factors and the most predictive SNPs identified with the machine learning-based procedure in the original subject set. The dotted trace shows the effect of deleting those subjects whose IMT level was the same or close to the quantile cut-off value (<0.02 difference in IMT). The randomized datasets were generated by first dividing the original set of subjects into the low- and high-risk classes at random, independent of their IMT-levels, and then repeating the same randomization process 100 times for each of the risk classes. The average AUC level for the various risk classes is reported. None of the 500 randomized datasets produced prediction accuracy higher than that obtained using the most predictive SNPs identified in the original set of subjects.
doi:10.1371/journal.pgen.1001146.g004

were divided into the low- and high-risk classes at random. This resulted in random prediction accuracies (average AUC 0.496), indicating that the high accuracies obtained with the predictive models were not by chance alone (Figure 4). Based on these results, independent and randomized subject sets were found to be useful for controlling the degree of overfitting, even when cross-validation is used in the model building.

Discussion

The present results demonstrate a predictive relationship between an individual's genotypic variation and early signs of atherosclerosis along with its progression over a 6-year period in our population-based longitudinal follow-up study. The relationship was much stronger with the variants identified using the machine learning-based approach compared to the variants identified using single-locus statistical hypothesis testing procedures either in the present study population or in the previous case-control association studies of clinically manifesting CHD [9–21]. This latter finding is in line with a recent observation that the genetic scores, constructed from individual SNPs that met the genome-wide level of statistical significance in earlier GWASs, could not improve the prediction of cardiovascular risk after adjustment for conventional cardiovascular risk factors [41]. Similar observations have been made in the context of other diseases when using such a 'bottom-up' approach to building discrimination models [33]. In the present study, rather than exclusively using only those variants with the lowest *p*-values for association, we took here an alternative 'top-down' approach to predictive modeling by explicitly searching for all of the genetic and conventional risk factors that positively contribute to the prediction power. It was surprising to note that, among the most predictive variants, there was only a single statistically significant SNP in the present cohort (see Table 2, Table 3, Table 4), supporting the idea that many of the predictive associations are detected much lower down on the ranked list of hits compared to the top hits with the highest statistical support [43]. Ignoring such 'gray zone' variants is likely to result in missing an important proportion of the quantitative variation in heritability [44]. The proposed predictive modeling framework therefore complements the statistical class comparison procedures traditionally used during the discovery phase.

We used our longitudinal cohort data of carotid atherosclerosis precursors to implement a class prediction model, with the specific aim to build a multivariate discrimination function, or a classifier [45], which can accurately predict the risk class of a new subject on the basis of a panel of key variants. Sampling of the subjects with increasing carotid IMT levels from our follow-up study provided us with the unique opportunity to investigate the genetic variants contributing to the present and future atherosclerosis risk. Evaluation of the genetic variants predictive of the 2001 IMT risk classes was used here to set a baseline for the prediction accuracies and for the corresponding SNP panels. Medically, it is perhaps most interesting to evaluate the ability to predict the future IMT risk classes as well as the progression of the IMT levels over the time. The determination of the future atherosclerosis risk is analogous to predicting the 2007 IMT risk classes based on the data reflecting the 2001 baseline genetic variants and confounding risk factors. The IMT progression (i.e., difference between the 2007 and 2001 IMT levels) is relevant in that even though an individual may not be considered to be in the risk group in 2007, the rate of change in the IMT levels between the evaluation years is large enough to warrant the subject as still being regarded as being at higher risk. The group with extreme IMT progression therefore represents the set of subjects who would be potential candidates for primary prevention in order to offset their likelihood of developing carotid atherosclerosis in the future. The full set of the SNP-panels predictive of the IMT-levels in the 2001 and 2007 studies, as well as of its relative progression from 2001 to 2007, are listed and characterized in Table S1. The genetic interactions between those variants that were highly predictive of the extreme IMT-progression are further discussed in Text S2.

Those SNPs that were found to be the most predictive of the 15% risk classes of IMT-levels and progression (Table 2, Table 3, Table 4) can be interpreted on the basis of a prior knowledge (Table S5). Most of the SNPs and corresponding genes have earlier been associated with cardiovascular disease risk factors such as low serum HDL-cholesterol and high serum LDL-cholesterol, triglycerides, lipoprotein(a) and apolipoprotein B concentrations (i.e., APOB, LPA, WWOX, ABCA1, USF1, PSRC1, ADRB2), inflammation, inflammatory and immunological factors such as serum CRP and interleukin levels (i.e., CRP, IL18, IL1B, LTA, ALOX5AP, IL10, ICOS, PTPN22), blood pressure, hemodynamics as well as serum asymmetric dimethyl arginine concentrations (DDAH2, WRN, WNK1, CDH13, NOS3), obesity, BMI, metabolic syndrome (FTO, ADRB2), and lipoprotein oxidation (PON1). Most of these SNPs are also linked to different cardiovascular traits, such as coronary artery disease, coronary artery calcification and atherosclerosis plaque areas, myocardial infarction, sudden cardiac death, stroke, as well as having phenotypic relationships with subclinical atherosclerotic traits such as carotid IMT (ESR1, APOB, PON1, USF1, ALOX5AP, ESR2, IL10, FCGR2A). Such associations have been found either alone or by interaction with other genes and clinical or

environmental factors, including diabetes mellitus and use of alcohol or smoking [46,47]. There were also novel IMT-related SNP candidates, earlier associated with bone density (C6orf97 and some intergenic SNPs), revealing possible mechanistic links to bone mineral and calcium metabolism. It is known that morphogenetic proteins and vascular calcification are activated in advanced atherosclerotic plaques [48–50]. On the basis of the present results, the same seems to hold true already in the subclinical stage of carotid atherosclerosis.

Limitations of the study and future developments

As with any association study that evaluates the contribution of a large number of candidate variants to a given phenotype, the question of how well the results will generalize to other study populations remains to be studied. This is a potential limitation in all SNP studies regardless of whether the class comparison or class prediction approach is being applied. It is known that associations identified in one population using the single-SNP statistical hypothesis testing procedures may not be detected in other populations in part due to the *p*-values being affected by the confounding factors [29,51]. Measures which directly evaluate the predictive value of multiple factors, such as AUC-values, can overcome some of these limitations but are not without caveats [32,33,52]. Unlike many other class prediction studies that have used the AUC to assess the discrimination accuracy within the given cases and control subjects only, here we used cross-validation both when selecting coherent subsets of the most predictive variants, through feature selection, as well as when evaluating their prediction accuracy, as compared to the subsets of the most significant SNPs. Cross-validation was necessary to avoid a selection bias, which can lead to over-optimistic prediction results and the reporting of a large number of over-fitted genetic variants [45,53]. The final evaluation of the panels of SNPs was done using an independent subject set to confirm that the reported models also generalize to other sub-populations beyond those used in the initial model estimation and validation. Testing on an independent dataset can also help to resolve any biases that may exist due to the fact that the cross-validation folds are far from independent of one another.

In common with many other SNP-studies, our main objective here was to find out those variants that are the most predictive of the atherosclerosis risk and progression in our follow-up study. When the aim is to obtain high prediction accuracies, the rules for including factors in the discrimination model are different from those when searching for the strongest statistical associations [54]. However, regardless of whether the discoveries come from statistical significance testing or from machine learning-based SNP-selection, the selected variants need to be carefully validated in further studies [55]. These two complementary approaches have also been combined, by building prediction models based exclusively on statistically significant SNPs, but this combined approach has been shown to result in poor classification accuracies [33]. In fact, reasonable increases in the prediction accuracies are often obtained only after including hundreds of top variants, depending on the complexity of the disease phenotype and whether or not cross-validation is utilized [32,38,39]. When the aim is class prediction, we believe it is better to make use of those methods that are specifically designed for optimal prediction, together with stringent feature selection and cross-validation, to output modest number of highly predictive and reliable variants for further study [45]. Further evaluation of the prediction power on independent and randomized subject sets was also found to be useful for controlling the degree of over-fitting, as shown in Figure 4, even when systematic cross-validation schemes are being used in the model building process [56,57].

It was interesting to note here that the simple naïve Bayes classifier performed well in the prediction of the atherosclerosis risk. The conditional independence assumption behind this probabilistic prediction model results in the nominal predictive probabilities that are often unrealistic, in the sense of being very close to either zero or one. Therefore, we followed the standard practice and chose the class with the highest posterior probability. Despite this simplifying assumption, the naïve Bayes classifier generally provided the best prediction results across the various risk classes, compared to other classification models, such as Bayes Nets, Support Vector Machines, or Random Forest (see Text S1 for their comparison). Moreover, because of its simplicity, the naïve Bayes classifier is also computationally more efficient than the other, more complex prediction models, enabling its usage in GWAS meta-analyses as well. These observations are in line with previous works, which have shown that the naïve Bayes classifier can perform well even in the case when there are strong dependencies in the dataset [58–60]. In particular, it has proven to be effective in the context of the IMT-phenotype and in SNP-data [61,62]. Standard filtering procedures, such as those based on the Hardy-Weinberg equilibrium, and other quality control measures implemented during the genotyping can result in severe restrictions on the joint distribution of alleles, enabling them to appear independent of one another, further explaining the good performance of the naïve Bayes classifier. However, other efficient SNP-subset selection methods that go beyond the single-SNP testing, such as those based on penalized maximum-likelihood approach [63], or different filter-wrapper machine learning approaches [31], could be used in the generic modeling framework.

While previous studies have identified sex-related differences in the cardiovascular disease incidence and genetic risk factors [64], the objective of the present study was to demonstrate that a common panel of genetic risk factors can already improve the prediction of subclinical carotid atherosclerosis risk and progression in a general population of young adults. Therefore, we did not stratify the subjects on the basis of any of the conventional risk factors, including sex or age, but the subjects were combined into a single distribution (Figure 1). In the future studies, however, it is possible to divide the heterogeneous population into more homogeneous sub-samples to investigate the relationship between the genetic and conventional risk factors in more controlled settings. Further, pathway and network analyses of such sub-sample-specific genetic variants and their interactions could reveal also underlying similarities or differences in the biological processes and genetic networks [6]. We have previously shown that sub-sampling-based automated procedures can help to detect hidden subject sub-groups that present with similar genetic profiles in genome-wide studies and which may associate with divergent clinical outcomes [65]. An automated subject grouping combined with the predictive modeling framework introduced in the present study could offer possibilities to start developing personalized approaches that make the most of genetic variation together with clinical data to predict individual susceptibility to the initiation and progression of carotid atherosclerosis and other complex diseases. Such experimental-computational approaches may prove to have also clinical utility in the early detection and management of subclinical atherosclerosis and other quantitative disorders.

Materials and Methods

Subject selection

The Cardiovascular Risk in Young Finns Study is an on-going population-based follow-up study of atherosclerosis precursors from childhood to adulthood [66]. The multi-center study has

been carried out in five university hospitals across Finland (Turku, Tampere, Helsinki, Kuopio and Oulu). The baseline cross-sectional study in 1980 included a total of 3,596 children and adolescents, aged between 3–18 years, who were randomly chosen from the national population register [67]. Since then, follow-up studies have been conducted in 1983, 1986, 2001 and 2007, in which the conventional risk factor data have systematically been collected from the individuals participating in those studies. In the two most recent follow-ups in 2001 and 2007, which were used in the present analysis, a total of 2,283 and 2,204 participants were re-examined, comprising the age groups of 24, 27, 30, 33, 36, 39 years and 30, 33, 36, 39, 42, 45 years, respectively; out of these, a total of 1,828 subjects participated both in the 2001 and 2007 follow-up studies [68]. The subjects involved in the cohort provided written consent to be included in the study approved by local ethics committees.

The study cohort for the present analysis was comprised of those subjects who took part in both the ultrasound and the genotyping studies in 2001. The carotid artery intima-media thickness (IMT) was measured from 1,809 subjects in both of the follow-up studies. Genotyping of single nucleotide polymorphisms (SNPs) was based on the DNA collected in 2001. The candidate gene approach was used to explore potentially interesting relationships between several known SNPs and clinical traits. Subjects who had missing values either in their IMT or SNP data in the year 2001 or 2007 were excluded from the present analysis, in order to eliminate their potentially adverse effects on both the reported prediction accuracies and on the selected genetic variants. Due to such stringent subject selection criteria (see Figure S1), the complete data matrices from $n = 1{,}027$ subjects were used in the search of genetic variants (SNP sets) that are predictive of the atherosclerosis (indexed by IMT) at baseline (2001); of these, $n = 813$ had complete data also in the follow-up study (2007), and could be used when searching for variants predictive of IMT progression (the change from 2001 to 2007).

Clinical characteristics

In the present analysis, we used the conventional risk factor data from the 2001 follow-up study. The physical examination consisted of the measurement of height, weight, systolic and diastolic blood pressure, and waist circumference [66]. The body mass index (BMI) was calculated by dividing the patients' weight in kilograms by the square of their height in meters. Waist circumference was recorded as the average of two measurements with an accuracy of 0.1 cm. Blood pressure was measured at least three times with a random zero sphygmomanometer, and the average of the three readouts of systolic and diastolic blood pressure was recorded. Lifestyle risk factors, such as smoking, were examined with questionnaires; the subjects who smoked daily were regarded as smokers. For the assessment of serum lipoprotein levels, venous blood samples were drawn after an overnight fast and the serum was separated, aliquoted and stored at -70°C until analysis. Standard enzymatic methods were used for recording the levels of serum total cholesterol, HDL-cholesterol, and LDL-cholesterol, as well as the concentrations of serum triglycerides, apolipoprotein A1 (ApoA1) and B (ApoB) [67,68].

Genotyping studies

Genomic DNA was extracted from peripheral blood leukocytes with a commercially available kit (Qiagen Inc., Valencia, CA). The DNA samples collected during the 2001 follow-up study were genotyped as described previously [66,69]. In the present analysis, we included the panel of 17 SNPs with the highest single-SNP statistical significance in the recent GWASs identifying variants for CHD outcomes and serum lipids [9–21], as well as a number of other candidate SNPs listed in the first phase of the international pooling project of cardiovascular cohorts [70]. A total of 108 SNPs with complete genotyping data in the selected subjects were considered here in the predictive modeling; these SNPs are generally related to serum lipid and lipoprotein metabolism, oxidation, cellular lipid metabolism, inflammation, immunological system, cell signaling, cell migration, cell growth, homocystein metabolisms, cellular adhesion and blood coagulation (see Table S1 for the full list of SNPs together with information on their gene annotation and chromosomal location, as well as on associated phenotypes available from previous studies).

Ultrasound imaging

Ultrasound studies were performed using Sequoia 512 ultrasound mainframes (Acuson Inc., Mountain View, CA, USA), with 13.0 MHz linear array transducers. Exactly the same scanning protocol was used both in 2001 and 2007 studies, as previously described [23]. Briefly, carotid IMT was measured on the posterior (far) wall of the left carotid artery. At least four measurements were taken 10 mm proximal to the bifurcation, and the average of the readouts was recorded. The digitally stored scans were manually analyzed by the same reader both in 2001 and 2007 blinded to the subjects' characteristics. The between-visit coefficient of variation of such IMT measurements was 6.4%, as estimated between two visits that were three months apart [23]. Since the IMT correlates with the risk of atherosclerosis progression and subsequent cardiovascular events [23–27], it was used here for stratifying the subjects into gradually increasing risk classes. Being non-invasive in its nature, this measurement can be justified in large populations of healthy subjects, without biases related clinically diagnosed cases and controls [28], making it a convenient quantitative phenotype of atherosclerosis in population-based follow-up studies. The quantitative IMT measurement suffers from a degree of measurement error, which can lead to regression to the mean (Figure S2).

Predictive modeling

The relative contribution of the conventional and genetic risk factors to the individual IMT levels was investigated by means of a predictive modeling framework, similar to that which we and others have used before [61,62]. Briefly, the study subjects were first divided into several risk classes according to their IMT levels. Based on the concept of extreme selection strategy [1–3], the quantile points, say (1-q) and q, of the IMT distribution were used to define the low and high risk classes, respectively (see Figure 1). The prediction of whether a subject belongs to the high-risk (H_q) or low-risk (L_q) class was done on the basis of his or her individual SNP data (S_1, …, S_l), whereas clinical characteristics, smoking habits, sex and age were used as confounding risk factors (C_1, …,C_m). A probabilistic prediction model, the so-called naïve Bayes classifier, was used here because of its low computational cost and good performance in previous studies [61,62,71]. Mathematically, the predictive classifier can be formulated as a conditional probability of observing the true class R (either H_q or L_q) given the genetic and confounding risk factors (the predictors P):

$$p(R|P) = Kp(R)\prod_{i=1}^{l} p(S_i|R)\prod_{j=1}^{m} p\left(C_j|R\right), \qquad (1)$$

where K is a scaling factor independent of the risk class R. The *a priori* probabilities $p(R)$ were set to the number of training samples in the low and high classes [71], and for numeric risk factors, the

training algorithm estimates the densities $p(x|R)$ using Gaussian distributions [72] (see Text S1 for more details). The subjects in the test material were then classified by choosing the risk class with the highest *posterior* probability in Eqn (1). The predictive power of different risk factor combinations was assessed with the *k*-fold cross-validation procedure, in which the given sample was divided into *k* distinct subsets of equal sizes, each of which in turn was used as a validation set, to assess how well the results will generalize to new sets of subjects, while the remaining sub-samples were used in the initial training of the prediction model [71]. The final prediction accuracy was reported as the average over the *k* validation rounds (here $k = 10$; see Figure S3).

Selection of predictive variants

The selection of predictive genetic and conventional risk factors was performed in two-steps, with the aim of identifying a minimal set of informative features for predicting the different risk classes (see Figure S3). The SNP selection was done using a machine-learning-based procedure, similar to the 'filter-wrapper' method [73]. The filtering phase starts from the full set of SNPs and uses an entropy-based information gain measure to reduce the high-dimensional search space to the subset of most informative genetic and conventional risk factors (here 40), which could subsequently be traversed thoroughly in the next phase of selection. In the wrapper phase, the best first-based iterative search-and-evaluate algorithm was used to further improve this subset by excluding those factors with least predictive power, using backward search direction, while the backtracking option allows for escaping from local optima [71]. The predictive power of the selected factor combinations was assessed using the naïve Bayes classifier, run with a 5-fold cross-validation to avoid potential selection bias, and the final prediction accuracy was evaluated using external 10-fold cross-validation (see Figure S3). The predictive modeling and risk factor selection was carried out with the Weka data mining platform (version 3.7; University of Waikato, New Zealand) [71].

Assessment of prediction accuracy

The predictive accuracy of the classifiers, constructed using either the *p*-value-based selection of the most significant SNPs or the machine-learning-based selection of the most predictive SNP-sets, was assessed using the receiver operating characteristic (ROC) analyses; ROC curves characterize the relative trade-off between true positive rate (sensitivity) and false positive rate (1 – specificity) of a classifier over the whole range of discrimination thresholds [32,33,71]. The overall accuracy of a classifier was summarized using the area under the ROC curve (AUC) measure; for an ideal classifier, $AUC = 1$, whereas a random classifier obtains an $AUC = 0.5$ on average [52,61,71]. The relative predictive power of each individual SNP or SNP-SNP interaction was assessed in terms of the change in AUC level when the particular SNP (say *x*) or the SNP-pair (*x*,*y*) was deleted from the selected set of variants (denoted by P_x and $P_{x,y}$, respectively). The interaction score for detecting epistasis effects was defined as $P_{x,y} - (P_x + P_y)$, resembling additive definition of genetic interactions based on single and double-deletion experiments in model organisms [74]. The AUC-values were calculated using the Weka platform (version 3.7; University of Waikato, New Zealand) [71].

Statistical procedures

The level of statistical association of single SNPs with the IMT-classes was assessed by determining the genotypic probabilities (*p*-values), on the basis of the 2×3 contingency matrix that contains the counts of the three genotypes among the low-risk and high-risk subjects [75]. Computationally efficient calculation of the exact *p*-values for each individual SNP was carried out with the ExactFDR software [76]. The Pearson correlation coefficient was used to assess the linear association between the various conventional risk factors and IMT-levels or changes. These *p*-values were adjusted for multiple testing using the Bonferroni correction. Although it is known that this correction may be conservative, especially when the test statistics are dependent, it provides an effective means for ensuring that the findings deemed most significant are not by chance alone when many hypotheses are being tested simultaneously. Differences in the distributions of the IMT-levels or changes between sub-populations were assessed using the Kolmogorov-Smirnov *D*-statistic, which is based on the maximal vertical distance between the two distributions. The statistical analyses were carried out with the SPSS Statistics software (version 17.0; SPSS Inc., Chicago, IL, USA) and with the statistical computing platform R (http://www.rproject.org/).

Supporting Information

Figure S1 The selection of the subjects and SNPs for the original dataset and for the independent validation set from the Cardiovascular Risk in Young Finns Study cohort. The white entries represent missing data points and their corresponding SNPs and subjects were removed by the final dataset which is represented by the completely shaded box on the upper left hand corner. The first inclusion criterion for the subjects was that they must have complete data for the set of 17 variants that have previously been associated with cardiovascular events (Established SNPs, the yellow submatrix). After that, the set of SNPs was extended gradually, to incorporate as many subjects as possible with complete SNP data. This selection procedure resulted in a sub-matrix of 1027 subjects and 108 SNPs that were used here when searching for the variants predictive of the severity and progression of sub-clinical atherosclerosis (Candidate SNPs, the blue submatrix). In order to create the independent validation dataset, the set of patients who were not part of the original 1027 subject subset were searched for those individuals who had complete data for all of the SNPs involved in a particular predictive model (Predictive SNPs). The number of patients, *n*, in each of the independent sets varied according to the particular risk class the validation set was created in relation to ($n = 103$, 222, 300, 351 and 423, for the 5%–25% risk classes, respectively).
Found at: doi:10.1371/journal.pgen.1001146.s001 (0.17 MB PDF)

Figure S2 Scatter plots of the IMT levels (A) in 2001 and 2007, and (B) with 2001 and the change in value between 2001 and 2007, both fitted with their respective linear correlation models (black lines). The plots are marked with two sets of vertical lines indicating the numerical IMT cutoff values used to select the 5% (red solid lines) and 15% (blue dashed lines) extreme quantiles and to split the subjects into the low-risk and high-risk classes. Although regression to the mean is observed, as was expected, it can be seen that the 15% extreme value class contains both increasing and decreasing IMT values, making it a unique situation in which the classifier must try to predict different IMT change directions within individual risk classes.
Found at: doi:10.1371/journal.pgen.1001146.s002 (0.53 MB PDF)

Figure S3 Schematic illustration summarizing the model building and evaluation procedure. Implementation and evaluation of the machine learning-based feature selection algorithm, compared to using the single-SNP *p*-values (the right-hand track). The aim of the algorithm was to select the subset of genetic factors (SNPs) and conventional risk factors (CRFs) from the filtered dataset that were the best predictors of the risk classes, determined

separately for 2001 and 2007 IMT levels (two follow-up points), as well as for its progression between 2001 and 2007 (IMT progression). The low-risk and high-risk were defined based on the gradually increasing quantiles of the pooled IMT distribution (q ranges from 5% to 25%). The most significant SNPs, determined using single-SNP statistical testing for the same risk classes, were used as a reference SNP selection approach in the evaluations.
Found at: doi:10.1371/journal.pgen.1001146.s003 (0.12 MB PDF)

Table S1 The SNPs explored in the present study, together with information on their gene annotation and chromosomal location (from the dbSNP database), and on associated phenotypes as available from the existing studies (listed in references). Established SNPs refer to those 17 variants identified in the previous CHD case-control association studies. The other columns indicate whether the SNPs were considered predictive of the various IMT risk classes.
Found at: doi:10.1371/journal.pgen.1001146.s004 (0.10 MB XLS)

Table S2 The statistical significance (p-value) calculated for each of the individual SNPs, depicting their degree of association with the various IMT risk classes in 2001, 2007, and with the IMT changes from 2001 to 2007.
Found at: doi:10.1371/journal.pgen.1001146.s005 (0.08 MB XLS)

Table S3 Pairwise correlations between the conventional risk factors and with the IMT levels in 2001, 2007, and progression from 2001 to 2007.
Found at: doi:10.1371/journal.pgen.1001146.s006 (0.05 MB XLS)

Table S4 Molecular pathways and biological processes of the genetic variants predictive of the most extreme 5% IMT change from 2001 to 2007.
Found at: doi:10.1371/journal.pgen.1001146.s007 (0.04 MB XLS)

Table S5 The interpretation of the SNPs most predictive of the 15% IMT risk classes in 2001, 2007, and of its progression from 2001 to 2007.
Found at: doi:10.1371/journal.pgen.1001146.s008 (0.04 MB XLS)

Text S1 Details of how Weka platform was used in the prediction studies.
Found at: doi:10.1371/journal.pgen.1001146.s009 (0.24 MB PDF)

Text S2 Supporting discussion text.
Found at: doi:10.1371/journal.pgen.1001146.s010 (0.05 MB PDF)

Author Contributions

Conceived and designed the experiments: TL LLE MK LT JSAV OTR TA. Performed the experiments: NM NP YMF JAH TL LPL RR CE NHK MH. Analyzed the data: SO LLE NM MJ JAH LPL TA. Contributed reagents/materials/analysis tools: SO TL LLE MK MJ TL LT JSAV OTR TA. Wrote the paper: SO TL OTR TA.

References

1. Plomin R, Haworth CM, Davis OS (2009) Common disorders are quantitative traits. Opinion. Nat Rev Genet 10: 872–878.
2. Schork NJ, Nath SK, Fallin D, Chakravarti A (2000) Linkage disequilibrium analysis of biallelic DNA markers, human quantitative trait loci, and threshold-defined case and control subjects. Am J Hum Genet 67: 1208–1218.
3. Lanktree MB, Hegele RA, Schork NJ, Spence JD (2010) Extremes of unexplained variation as a phenotype: an efficient approach for genome-wide association studies of cardiovascular disease. Circ Cardiovasc Genet 3: 215–221.
4. Zhang G, Nebert DW, Chakraborty R, Jin L (2006) Statistical power of association using the extreme discordant phenotype design. Pharmacogenet Genomics 16: 401–143.
5. Eguchi T, Maruyama T, Ohno Y, Morii T, Hirao K, et al. (2009) Possible association of tumor necrosis factor receptor 2 gene polymorphism with severe hypertension using the extreme discordant phenotype design. Hypertens Res 32: 775–779.
6. Torkamani A, Schork NJ (2009) Pathway and network analysis with high-density allelic association data. Methods Mol Biol 563: 289–301.
7. Pearson TA (2002) New tools for coronary risk assessment: what are their advantages and limitations? Circulation 105: 886–892.
8. Koskinen J, Kähönen M, Viikari JS, Taittonen L, Laitinen T, et al. (2009) Conventional cardiovascular risk factors and metabolic syndrome in predicting carotid intima-media thickness progression in young adults: the cardiovascular risk in young Finns study. Circulation 120: 229–236.
9. Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, et al. (2007) Genome-wide association analysis of coronary artery disease. N Engl J Med 357: 443–453.
10. McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, et al. (2007) A common allele on chromosome 9 associated with coronary heart disease. Science 316: 1488–1491.
11. Helgadottir A, Thorleifsson G, Manolescu A, Gretarsdottir S, Blondal T, et al. (2007) A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science 316: 1491–1493.
12. Larson MG, Atwood LD, Benjamin EJ, Cupples LA, D'Agostino RB, Sr, et al. (2007) Framingham Heart Study 100K project: genome-wide associations for cardiovascular disease outcomes. BMC Med Genet 8: S5.
13. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14 000 cases of seven common diseases and 3 000 shared control. Nature 447: 661–678.
14. Luke MM, Kane JP, Liu DM, Rowland CM, Shiffman D, et al. (2007) A polymorphism in the protease-like domain of apolipoprotein(a) is associated with severe coronary artery disease. Arterioscler Thromb Vasc Biol 27: 2030–2036.
15. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40: 161–169.
16. Kathiresan S, Melander O, Anevski D, Guiducci C, Burtt NP, et al. (2008) Polymorphisms associated with cholesterol and risk of cardiovascular events. N Engl J Med 358: 1240–1249.
17. Shiffman D, Kane JP, Louie JZ, Arellano AR, Ross DA, et al. (2008) Analysis of 17,576 potentially functional SNPs in three case-control studies of myocardial infarction. PloS ONE 3: e2895. doi:10.1371/journal.pone.0002895.
18. Abdullah KG, Li L, Shen GQ, Hu Y, Yang Y, et al. (2008) Four SNPs on chromosome 9p21 confer risk to premature, familial CAD and MI in an American Caucasian population (GeneQuest). Annals Human Genet 72: 654–657.
19. Sagoo GS, Tatt I, Salanti G, Butterworth AS, Sarwar N, et al. (2008) Seven lipoprotein lipase gene polymorphisms, lipid fractions, and coronary disease: a HuGE association review and meta-analysis. Am J Epidemiol 168: 1233–1246.
20. Anderson JL, Horne BD, Kolek MJ, Muhlestein JB, Mower CP, et al. (2008) Genetic variation at the 9p21 locus predicts angiographic coronary artery disease prevalence but not extent and has clinical utility. Am Heart J 156: 1155–1162.
21. Paynter NP, Chasman DI, Buring JE, Shiffman D, Cook NR, et al. (2009) Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3. Ann Intern Med 150: 65–72.
22. Lusis AJ, Pajukanta P (2008) A treasure trove for lipoprotein biology. Comment. Nat Genet 40: 129–130.
23. Raitakari OT, Juonala M, Kähönen M, Taittonen L, Laitinen T, et al. (2003) Cardiovascular risk factors in childhood and carotid artery intima-media thickness in adulthood: The Cardiovascular Risk in Young Finns Study. JAMA 2003 290: 2277–2283.
24. Li S, Chen W, Srinivasan SR, Bond MG, Tang R, et al. (2003) Childhood cardiovascular risk factors and carotid vascular changes in adulthood: The Bogalusa Heart Study. JAMA 290: 2271–2276.
25. Salonen JT, Salonen R (1991) Ultrasonographically assessed carotid morphology and the risk of coronary heart disease. Arteroscler Thromb 11: 1245–1249.
26. O'Leary DH, Polak JF, Kronmal RA, Manolio TA, Burke GL, et al. (1999) Carotid-artery intima and media thickness as a risk factor for myocardial infarction and stroke in older adults. Cardiovascular Health Study Collaborative Research Group. N Engl J Med 340: 14–22.
27. Lorenz MW, Markus HS, Bots ML, Rosvall M, Sitzer M (2007) Prediction of clinical cardiovascular events with carotid intima-media thickness: a systematic review and meta-analysis. Circulation 115: 459–467.

28. O'Leary DH, Polak JF (2002) Intima-media thickness: a tool for atherosclerosis imaging and event prediction. Am J Cardiol 90: 18L–21L.
29. Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10: 241–251.
30. Moore JH, Williams SM (2009) Epistasis and its implications for personal genetics. Am J Hum Genet 85: 309–320.
31. Moore JH, Asselbergs FW, Williams SM (2010) Bioinformatics challenges for genome-wide association studies. Bioinformatics 26: 445–455.
32. Kraft P, Wacholder S, Cornelis MC, Hu FB, Hayes RB, et al. (2009) Beyond odds ratios: communicating disease risk based on genetic profiles. Perspective. Nat Rev Genet 10: 264–9.
33. Jakobsdottir J, Gorin MB, Conley YP, Ferrell RE, Weeks DE (2009) Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. PLoS Genet 5: e1000337. doi:10.1371/journal.pgen.1000337.
34. Samani NJ, Raitakari OT, Sipilä K, Tobin MD, Schunkert H, et al. (2008) Coronary artery disease-associated locus on chromosome 9p21 and early markers of atherosclerosis. Arterioscler Thromb Vasc Biol 28: 1679–1683.
35. Fan YM, Raitakari OT, Kähönen M, Hutri-Kähönen N, Juonala M, et al. (2009) Hepatic lipase promoter C-480T polymorphism is associated with serum lipids levels, but not subclinical atherosclerosis: The Cardiovascular Risk in Young Finns Study. Clin Genet 76: 46–53.
36. Humphries SE, Cooper JA, Talmud PJ, Miller GJ (2007) Candidate gene genotypes, along with conventional risk factor assessment, improve estimation of coronary heart disease risk in healthy UK men. Clin Chem 53: 8–16.
37. Morrison AC, Bare LA, Chambless LE, Ellis SG, Malloy M, et al. (2007) Prediction of coronary heart disease risk using a genetic risk score: the Atherosclerosis Risk in Communities Study. Am J Epidemiol 166: 28–35.
38. van der Net JB, Janssens AC, Defesche JC, Kastelein JJ, Sijbrands EJ, et al. (2009) Usefulness of genetic polymorphisms and conventional risk factors to predict coronary heart disease in patients with familial hypercholesterolemia. Am J Cardiol 103: 375–380.
39. van der Net JB, Janssens AC, Sijbrands EJ, Steyerberg EW (2009) Value of genetic profiling for the prediction of coronary heart disease. Am Heart J 158: 105–110.
40. Ioannidis JP (2009) Prediction of cardiovascular disease outcomes and established cardiovascular risk factors by genome-wide association markers. Circ Cardiovasc Genet 2: 7–15.
41. Paynter NP, Chasman DI, Paré G, Buring JE, Cook NR, et al. (2010) Association between a literature-based genetic risk score and cardiovascular events in women. JAMA 303: 631–637.
42. Cordell HJ (2009) Genome-wide association studies: Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10: 392–404.
43. Donnelly P (2008) Progress and challenges in genome-wide association studies in humans. Commentary. Nature 456: 728–731.
44. Maher B (2008) Personal genomes: The case of the missing heritability. News Feature. Nature 456: 18–21.
45. Simon R, Radmacher MD, Dobbin K, McShane LM (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 95: 14–18.
46. Rontu R, Karhunen PJ, Ilveskoski E, Mikkelsson J, Kajander O, et al. (2003) Smoking-dependent association between paraoxonase 1 M/L55 genotype and coronary atherosclerosis in males: an autopsy study. Atherosclerosis 171: 31–37.
47. McGeachie M, Ramoni RL, Mychaleckyj JC, Furie KL, Dreyfuss JM, et al. (2009) Integrative predictive model of coronary artery calcification in atherosclerosis. Circulation 120: 2448–2454.
48. Bostrom K, Watson KE, Horn S, Wortham C, Herman IM, et al. (1993) Bone morphogenetic protein expression in human atherosclerotic lesions. J Clin Invest 91: 1800–1809.
49. Bucay N, Sarosi I, Dunstan CR, Morony S, Tarpley J, et al. (1998) Osteoprotegerin-deficient mice develop early onset osteoporosis and arterial calcification. Genes Dev 12: 1260–1268.
50. Collin-Osdoby P (2004) Regulation of vascular calcification by osteoclast regulatory factors RANKL and osteoprotegerin. Review. Circ Res 95: 1046–1057.
51. Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. Nat Rev Genet 10: 681–690.
52. Janssens AC, van Duijn CM (2009) Genome-based prediction of common diseases: methodological considerations for future research. Genome Med 1: 20.
53. Ambroise C, McLachlan GJ (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci USA 99: 6562–6566.
54. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P (2004) Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. Am J Epidemiol 159: 882–890.
55. Ioannidis JP, Thomas G, Daly MJ (2009) Validating, augmenting and refining genome-wide association signals. Nat Rev Genet 10: 318–329.
56. Reunanen J (2003) Overfitting in making comparisons between variable selection methods. J Machine Learn Res 3: 1371–1382.
57. Anderssen E, Dyrstad K, Westad F, Martens H (2006) Reducing over-optimism in variable selection by cross-model validation. Chemometrics Intell Laborat Systems 84: 69–74.
58. Domingos P, Pazzan M (1997) On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning 29: 103–130.
59. Hand DJ, Yu K (2001) Idiot's Bayes – not so stupid after all? International Statistical Rev 69: 385–398.
60. Zhang H (2005) Exploring conditions for the optimality of naïve Bayes. International J Patt Recogn Artif Intelligence 19: 183–198.
61. Aittokallio J, Polo O, Hiissa J, Virkki A, Toikka J, et al. (2008) Overnight variability in transcutaneous carbon dioxide predicts vascular impairment in women. Exp Physiol 93: 880–891.
62. Long N, Gianola D, Rosa GJ, Weigel KA, Avendaño S (2009) Comparison of classification methods for detecting associations between SNPs and chick mortality. Genet Sel Evol 41: 18.
63. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS Genet 4: e1000130. doi:10.1371/journal.pgen.1000130.
64. Silander K, Alanne M, Kristiansson K, Saarela O, Ripatti S, et al. (2008) Gender differences in genetic risk profiles for cardiovascular disease. PLoS ONE 3: e3615. doi:10.1371/journal.pone.0003615.
65. Hiissa J, Elo LL, Huhtinen K, Perheentupa A, Poutanen M, et al. (2009) Resampling reveals sample-level differential expression in clinical genome-wide studies. OMICS 13: 381–396.
66. Raitakari OT, Juonala M, Rönnemaa T, Keltikangas-Järvinen L, Räsänen L, et al. (2008) Cohort profile: the Cardiovascular Risk in Young Finns Study. Int J Epidemiol 37: 1220–6.
67. Åkerblom HK, Viikari J, Uhari M, Räsänen L, Byckling T, et al. (1985) Atherosclerosis precursors in Finnish children and adolescents. I. General description of the cross-sectional study of 1980, and an account of the children's and families' state of health. Acta Paediatr Scand Suppl 318: 49–63.
68. Raiko JR, Viikari JS, Ilmanen A, Hutri-Kähönen N, Taittonen L, et al. (2010) Follow-ups of the Cardiovascular Risk in Young Finns Study in 2001 and 2007: Levels and 6-year changes in risk factors. J Intern Med 267: 370–384.
69. Livak KJ (1999) Allelic discrimination using fluorogenic probes and the 5′ nuclease assay. Genet Anal 14: 143–149.
70. Evans A, Salomaa V, Kulathinal S, Asplund K, Cambien F, et al. (2005) MORGAM (an international pooling of cardiovascular cohorts). Review. Int J Epidemiol 34: 21–27.
71. Witten IH, Frank E (2005) Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. San Francisco: Morgan Kaufmann Publishers.
72. John G, Langley P (1995) Estimating continuous distributions in Bayesian classifiers. In Proceedings of the Eleventh Conference of Uncertainty in Artificial Intelligence. San Mateo: Morgan Kaufmann Publishers. pp 338–345.
73. Long N, Gianola D, Rosa GJ, Weigel KA, Avendaño S (2007) Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. J Anim Breed Genet 124: 377–389.
74. Phillips PC (2008) Epistasis: the essential role of gene interactions in the structure and evolution of genetic systems. Review. Nat Rev Genet 9: 855–867.
75. Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7: 781–791.
76. Wojcik J, Forner K (2008) ExactFDR: exact computation of false discovery rate estimate in case-control association studies. Bioinformatics 24: 2407–2408.

Publication II

Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations

Pahikkala, T., Okser, S., Airola, A., Salakoski, T., Aittokallio, T. (2012). *Algorithms for Molecular Biology*, 7(1):11.

Pahikkala *et al. Algorithms for Molecular Biology* 2012, **7**:11
http://www.almob.org/content/7/1/11

ALGORITHMS FOR MOLECULAR BIOLOGY

RESEARCH Open Access

Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations

Tapio Pahikkala[1,2*], Sebastian Okser[1,2], Antti Airola[1,2], Tapio Salakoski[1,2] and Tero Aittokallio[2,3,4,5]

Abstract

Background: Through the wealth of information contained within them, genome-wide association studies (GWAS) have the potential to provide researchers with a systematic means of associating genetic variants with a wide variety of disease phenotypes. Due to the limitations of approaches that have analyzed single variants one at a time, it has been proposed that the genetic basis of these disorders could be determined through detailed analysis of the genetic variants themselves and in conjunction with one another. The construction of models that account for these subsets of variants requires methodologies that generate predictions based on the total risk of a particular group of polymorphisms. However, due to the excessive number of variants, constructing these types of models has so far been computationally infeasible.

Results: We have implemented an algorithm, known as greedy RLS, that we use to perform the first known wrapper-based feature selection on the genome-wide level. The running time of greedy RLS grows linearly in the number of training examples, the number of features in the original data set, and the number of selected features. This speed is achieved through computational short-cuts based on matrix calculus. Since the memory consumption in present-day computers can form an even tighter bottleneck than running time, we also developed a space efficient variation of greedy RLS which trades running time for memory. These approaches are then compared to traditional wrapper-based feature selection implementations based on support vector machines (SVM) to reveal the relative speed-up and to assess the feasibility of the new algorithm. As a proof of concept, we apply greedy RLS to the Hypertension – UK National Blood Service WTCCC dataset and select the most predictive variants using 3-fold external cross-validation in less than 26 minutes on a high-end desktop. On this dataset, we also show that greedy RLS has a better classification performance on independent test data than a classifier trained using features selected by a statistical p-value-based filter, which is currently the most popular approach for constructing predictive models in GWAS.

Conclusions: Greedy RLS is the first known implementation of a machine learning based method with the capability to conduct a wrapper-based feature selection on an entire GWAS containing several thousand examples and over 400,000 variants. In our experiments, greedy RLS selected a highly predictive subset of genetic variants in a fraction of the time spent by wrapper-based selection methods used together with SVM classifiers. The proposed algorithms are freely available as part of the RLScore software library at http://users.utu.fi/aatapa/RLScore/.

Keywords: GWAS, Genome-wide association study, Machine learning, Feature selection

*Correspondence: tapio.pahikkala@utu.fi
[1]Department of Information Technology, University of Turku, Turku, Finland
[2]Turku Centre for Computer Science, Turku, Finland
Full list of author information is available at the end of the article

Background

The common goal of genome-wide association studies (GWAS) is the identification of genetic loci that can help to discriminate an individual's susceptibility to various common disorders. Identification of genetic features that are highly predictive of an individual's disease status would facilitate the development of methods for determining both an individual's risk of developing a clinical condition along with the possibility of new treatment options such as personalized medicine [1-5]. In the case of GWAS, the common genetic marker of interest is the single nucleotide polymorphism (SNP). It is widely theorized that complex diseases can be predicted before an individual has been found to have a clinical manifestation of a particular disorder [4,6,7]. The creation of more accurate disease risk detection techniques will ideally assist clinicians in the development of new medicines in addition to determining which individuals are in a greater need of receiving expensive preventative treatments, while allowing those who are at a low risk to avoid undergoing potentially superfluous medical care.

While numerous genetic loci have been prior identified through standard SNP analyses, the results of these studies have only provided a limited explanation regarding an individual's disease status [3,5,7-9]. Contrary to the current knowledge of synergistic interactions amongst genetic variants, traditional GWAS, through the use of single-SNP association testing, have implemented analysis methodologies that ignore the epistasis interactions between the genetic loci [3,7,10-12]. While it has been prior demonstrated that the heritability of most disorders is the result of numerous complex interactions between multiple SNPs, the aggregate of the effects of these markers still provides a clinically insufficient prediction of the disease status [10,13]. To account for these variant interactions, association studies have begun to implement various machine learning-based approaches to incorporate the complex epistasis pattern effects [3,7,14-16]. In contrast to conventional statistical methods, machine learning algorithms tend to place a larger emphasis on prediction making and how the values of a particular variant contribute to the effect of other markers, making them ideal for developing predictive strategies in genetic association studies.

In typical GWAS, the problems under study are modeled as binary classification tasks. Examples are labeled either as cases or controls for a particular disease, with the cases representing those individuals who have the disease and the controls those who are free of the disease. In recent years, methods of selecting the most relevant variants to prediction of a disease, known as feature selection, have begun to gain prominence in bioinformatics studies [7,17-20]. Two common feature selection methodologies are commonly presented, filter and wrapper methods [17,18]. In filter methods, the selection is done as a preprocessing step before learning by computing univariate statistics on feature-by-feature basis. The approach is computationally efficient, but the methods are not able to take into account the dependencies between the variants, or the properties of the learning algorithm which is subsequently trained on the features. This can lead to suboptimal predictive performance.

Delving deeper into feature selection, we consider the wrapper model, in which the features are selected through interaction with a classifier training method [21]. The selection consists of a search over the power set of features. For each examined feature set, a classifier is trained, and some scoring measure, which estimates its generalization error, is used to evaluate the quality of the considered feature set. Measuring the feature set quality on the training set is known to have a high risk of overfitting, and hence other estimates, such as those based on cross-validation (CV) [22,23], have been proposed as more reliable alternatives (see e.g. [21]). Since the size of the search space grows exponentially with the number of features, testing all feature subsets is infeasible. Rather, wrapper methods typically use search heuristics, such as greedy forward or backward selection, or genetic algorithms, to find locally optimal solutions. The wrapper methods have been demonstrated to have the potential to achieve better predictive performance than the filter approach [7,18,24,25], but this increase in performance is accompanied by increased computation times. This is due to the property of the wrapper methods that they require re-training a classification algorithm for each search step and each round of CV.

A number of studies related to the use of wrapper-based feature selection and the implementation of classifiers on biological markers haven been published, with the majority of the work dealing with the problem of gene selection from DNA microarray data. One of the most successful classifier learning algorithms in this domain has been the support vector machine (SVM) [26]. Proposed approaches include the combination of SVM classification with prefiltering of features [27,28], wrapper based methods [29-31], as well as embedded methods that incorporate feature selection within the SVM training algorithm, such as the recursive feature selection method [32]. These previous approaches have been mostly proposed for and tested on small scale learning problems, where the number of training examples ranges in at most hundreds, and the number of features in thousands. However, it is not straightforward to extend these methods to GWAS problems, where the training set sizes range in thousands and feature set sizes in hundreds of thousands or even millions. From a scalability perspective, SVMs are actually not a particularly suitable choice as a building block for constructing feature selection methods, since the method

has to be re-trained from scratch for each tested feature set, and for each round of CV. This can lead to unfeasible computational costs on large and high-dimensional data sets. Due to this reason, previous studies on implementing SVMs on GWAS have required pre-filtering of the data [3,20,33]. The same problem naturally applies also to most other classifier training methods.

Regularized least-squares (RLS), also known as the least-squares support vector machine (LS-SVM), and ridge regression, among other names, is a learning algorithm similar to SVMs [34-40]. Numerous comparisons of the SVM and RLS classifier can be found in the literature (see e.g. [37,40-43]), the results showing that typically there is little to no difference in classification performance between the two methods. However, for the purposes of wrapper based feature selection, RLS has one major advantage over SVMs, namely that RLS has a closed form solution that can be expressed in terms of matrix operations. This in turn allows the development of computational shortcuts, which allow re-using the results of previous computations when making minor changes to the learning problem. The existence of a fast leave-one-out (LOO) CV shortcut is a classical result [44], that has recently been extended to arbitrarily sized folds [45]. Similar shortcuts can be developed for operations where features are added to the, or left out of the training set, and the resulting classification model is updated accordingly. Such shortcuts have been used to derive RLS-based wrapper selection methods for gene selection from microarray data [19,46-48]. However, the previously proposed methods did not fully utilize the possibilities of matrix algebra for speeding up the computations, making them still unsuitable for very large data sets, such as those encountered in GWAS.

In the present work, we have developed and implemented the first wrapper-based feature selection method capable of performing feature selection on the entire span of SNPs available in a typical GWAS, without the necessity for pre-filtering to reduce the number of attributes. The method is based on the greedy RLS algorithm [49], which uses computational shortcuts to speed up greedy forward selection with LOO error as the selection criterion. Greedy RLS yields equivalent results to the most efficient of the previously proposed methods for wrapper based feature selection with RLS, called the low-rank updated LS-SVM method [48], while having lower computational complexity. Namely, the running time of greedy RLS grows linearly in the number of training examples, the number of features in the original data set, and the number of selected features. This is in contrast to the low-rank updated LS-SVM that scales quadratically with respect to the number of training data points. Further, we propose a space-efficient variation of greedy RLS that trades speed for decreased memory consumption. The method is efficient enough to perform feature selection on GWAS data with hundreds of thousands of SNPs and thousands of data points on a high-end desktop machine. As a case study, we were able to implement the method on the Wellcome Trust Case-Control Consortium (WTCCC) Hypertension (HT) dataset combined with the National Blood Service (NBS) controls samples, obtaining a highly discriminant classification on independent test data.

Related works

There exists a number of prior works in applying machine learning based method to GWAS studies. For instance, it was demonstrated that when SVMs are applied to the results of filter based feature selection, high area under the curve (AUC) values in the detection of Type 1 Diabetes (T1D) can be obtained [3]. More specifically, it was shown that through the use of a filter method, in which they selected only those features with significance values of less than pre-selected thresholds, they could outperform logistic regression methods. The paper made the discerning observation that using only more statistically significant markers in disease prediction actually causes a loss of information and thus a decrease in AUC [3]. Such statistical p-value based filtering has also been shown to result in sub-optimal prediction performance in other studies [2,50,51].

Previously, we have shown that in a population based candidate SNP study, a combined filter-wrapper approach allowed for an accurate prediction of the onset of carotid atherosclerosis on independent test data [7]. While the accuracy of the wrapper-based methods was demonstrated on a small subsample of available SNPs, the method would not scale to unfiltered SNP sets. Also other approaches, such as dimensionality reduction, have been applied, but they were not able to scale to an entire GWAS either [52]. Moreover, LASSO-based feature selection methods have been used, but only on a filtered-subset rather than an entire GWAS [12,53]. Furthermore, several other works have also addressed the issue of the computational feasibility of implementing machine learning algorithms on entire GWAS but have reported the same conclusion, that at the moment it was not practical to use such methods without extensive pre-filtering [15,54-56].

To conclude, the above mentioned works tend to make use of various filters to initially reduce the total number of features to a number in which computationally non-optimized algorithms can be applied. Most works tend to filter the final number of SNPs being analyzed to the tens of thousands. While such methods are often sufficient for analyzing GWAS datasets, our aim here is to show that it is computationally feasible to implement wrapper methods on entire GWAS scale with a large number of training examples and all of the available features. This, in turn, can lead to discovering models

with increased predictive performance, as is shown in our experiments.

Methods

Preliminaries

Let us start by making the assumption that the task being solved is a binary classification problem. We are supplied with a training set of m examples, each having n real-valued features, as well as a class label denoting whether the example belongs to the positive or to the negative class. In the case of GWAS, the features are representative of the number of minor alleles present in a particular SNP (either 0, 1 or 2, representing the minor allele count), the examples represent each individuals data in a particular study and the class label is the disease status of a particular example, with the positive class representing those who have the disease and the negative class indicative of those without the disease. Our goal is to select an informative subset of the features, based on which we can construct an accurate classifier for predicting the class labels of new, unseen test examples.

Next, we introduce some matrix notation. Let $\mathbb{R}^m$ and $\mathbb{R}^{n\times m}$ denote the sets of real-valued column vectors and $n \times m$-matrices, respectively. To denote real valued matrices and vectors we use bold capital letters and bold lower case letters, respectively. Moreover, index sets are denoted with calligraphic capital letters. By denoting $\mathbf{M}_i$, $\mathbf{M}_{:,j}$, and $\mathbf{M}_{i,j}$, we refer to the ith row, jth column, and (i,j)th entry of the matrix $\mathbf{M} \in \mathbb{R}^{n\times m}$, respectively. Similarly, for index sets $\mathcal{R} \subseteq \{1,\ldots,n\}$ and $\mathcal{L} \subseteq \{1,\ldots,m\}$, we denote the submatrices of $\mathbf{M}$ having their rows indexed by $\mathcal{R}$, the columns by $\mathcal{L}$, and the rows by $\mathcal{R}$ and columns by $\mathcal{L}$ as $\mathbf{M}_{\mathcal{R}}$, $\mathbf{M}_{:,\mathcal{L}}$, and $\mathbf{M}_{\mathcal{R},\mathcal{L}}$, respectively. We use an analogous notation also for column vectors, that is, $\mathbf{v}_i$ refers to the ith entry of the vector $\mathbf{v}$.

Let $\mathbf{X} \in \mathbb{R}^{n\times m}$ be a matrix containing the whole feature representation of the examples in the training set, where n is the total number of features and m is the number of training examples. The (i,j)th entry of $\mathbf{X}$ contains the value of the ith feature in the jth training example. Note that while we here define $\mathbf{X}$ to be real-valued, in GWAS the data can usually be stored in an integer-valued matrix, which is much more memory efficient. The memory issues concerning the data types are discussed more in detail below. Moreover, let $\mathbf{y} \in \mathbb{R}^m$ be a vector containing the labels of the training examples. In binary classification tasks, we restrict the labels to be either 1 or -1, indicating whether the data point belongs to the positive or negative class, respectively.

In this paper, we consider linear predictors of type

$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}}, \tag{1}$$

where $\mathbf{w}$ is the $|\mathcal{S}|$-dimensional vector representation of the learned predictor and $\mathbf{x}_{\mathcal{S}}$ can be considered as a mapping of the data point x into $|\mathcal{S}|$-dimensional feature space.[a] Note that the vector $\mathbf{w}$ only contains entries corresponding to the features indexed by $\mathcal{S}$. The rest of the features of the data points are not used in the prediction phase. The computational complexity of making predictions with (1) and the space complexity of the predictor are both $O(|\mathcal{S}|)$ provided that the feature vector representation $\mathbf{x}_{\mathcal{S}}$ for the data point x is given.

Wrapper-based feature selection

In wrapper-based feature selection, the most commonly used search heuristic is greedy forward selection in which one feature is added at a time to the set of selected features, but features are never removed from the set. A pseudo code of a greedy forward selection that searches feature sets up to size k, is presented in Algorithm 1. In the algorithm description, the outermost loop adds one feature at a time into the set of selected features $\mathcal{S}$ until the size of the set has reached the desired number of selected features k. The inner loop goes through every feature that has not yet been added into the set of selected features and, for each of those, computes the value of the heuristic H for the set including the feature under consideration and the current set of selected features. With $\mathrm{H}(\mathbf{X}_{\mathcal{R}}, \mathbf{y})$, we denote the value of the heuristic obtained with a data matrix $\mathbf{X}_{\mathcal{R}}$ and a label vector $\mathbf{y}$. In the end of the algorithm description, $t(\mathbf{X}_{\mathcal{S}}, \mathbf{y})$ denotes the black-box training procedure which takes a data matrix and a label vector as input and returns a vector representation of the learned predictor $\mathbf{w}$.

Algorithm 1 Wrapper-based feature selection

$\mathcal{S} \leftarrow \emptyset$
while $|\mathcal{S}| < k$ **do**
  $e \leftarrow \infty$
  $b \leftarrow 0$
  for $i \in \{1,\ldots,n\} \setminus \mathcal{S}$ **do**
    $\mathcal{R} \leftarrow \mathcal{S} \cup \{i\}$
    $e_i \leftarrow \mathrm{H}(\mathbf{X}_{\mathcal{R}}, \mathbf{y})$
    if $e_i < e$ **then**
      $e \leftarrow e_i$
      $b \leftarrow i$
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{b\}$
$\mathbf{w} \leftarrow t(\mathbf{X}_{\mathcal{S}}, \mathbf{y})$

Using the training set error as a selection heuristic is known to be unreliable due to overfitting, and therefore it has been proposed to measure the quality of feature sets with CV [57]. The CV approach can be formalized as follows. Let $\mathcal{C} = \{1,\ldots,m\}$ denote the indices of the training instances. In CV, we have a set $\mathcal{H} = \{\mathcal{H}_1,\ldots,\mathcal{H}_N\}$ of hold-out sets, where $N \in \mathbb{N}$ is the number of rounds in CV and $\mathcal{H}_i \subseteq \mathcal{C}$. In the most popular form of N-fold CV, the hold-out sets are mutually disjoint, that is, $\mathcal{H}_i \cap \mathcal{H}_j = \emptyset$ when

$i \neq j$. Now, given a performance measure l, the average performance over the CV rounds is computed as

$$\mathcal{L}(\mathbf{X}, \mathbf{y}) = \sum_{\mathcal{H} \in \mathcal{H}} l\left(f_{\overline{\mathcal{H}}}(\mathbf{X}_{:,\mathcal{H}}), \mathbf{y}_{\mathcal{H}}\right),$$

where $f_{\overline{\mathcal{H}}}$ is a predictor which is trained with the examples indexed by $\overline{\mathcal{H}} = \mathcal{C} \setminus \mathcal{H}$, and $\mathbf{X}_{:,\mathcal{H}}$ and $\mathbf{y}_{\mathcal{H}}$ contain, respectively, the features and the labels of the examples indexed by $\mathcal{H}$. Leave-one-out (LOO) CV is an extreme form of N-fold CV in which every hold-out set is of size one and every training example is held out at a time, that is, $N = m$.

Since the outer and inner loops in Algorithm 1 have k and n rounds, respectively, the computational complexity of the wrapper based greedy forward selection is $O(knH)$, where H is the complexity of calculating the value of the heuristic for feature sets of size up to k. For example, if we use LOO error as a heuristic and the LOO calculation is wrapped around a black-box training algorithm, the time complexity of the heuristic is usually m times the complexity of the training method. This is often infeasible in practice. Fortunately, as it is widely known in literature, computational short-cuts enabling the calculation of the LOO error without needing to retrain the predictor from scratch exist for many machine learning methods (see e.g. [23]).

The selection of the performance measure l used in the CV heuristics may also have an effect on the computation time. The performance measure can be selected to be the same as the one we aim to maximize in the first place but it may also make sense to use approximations in order to speed up the feature selection process. For example, while the computation of AUC requires $O(m \log(m))$ floating point operations, the mean squared error can be computed in a linear time. These complexities are, of course, usually negligible compared to the training complexities of the learning methods. However, this is not the case for the greedy RLS method as we will show below.

Support vector machines and regularized least squares

A large class of machine learning algorithms can be formulated as the following regularized risk minimization problem [58]:

$$\mathbf{w}^* = \underset{\mathbf{w} \in \mathbb{R}^{|\mathcal{S}|}}{\operatorname{argmin}} \left\{ l\left((\mathbf{X}_{\mathcal{S}})^{\mathrm{T}}\mathbf{w}, \mathbf{y}\right) + \lambda \mathbf{w}^{\mathrm{T}}\mathbf{w} \right\}, \quad (2)$$

where the first term is the empirical risk measuring how well $\mathbf{w}$ fits the training data, $\mathbf{w}^{\mathrm{T}}\mathbf{w}$ is the quadratic regularizer measuring the complexity of the considered hypothesis, $\lambda > 0$ is a parameter, and $l : \mathbb{R}^m \times \mathbb{R}^m \mapsto [0, \infty)$ is a convex loss function measuring how well a predicted and true label match. The regularized risk minimization framework (2) can be extended to non-linear learning and structured data by means of the *kernel trick* [59], however this is not necessary for the considerations in this paper.

The hinge loss, defined as

$$l\left((\mathbf{X}_{\mathcal{S}})^{\mathrm{T}}\mathbf{w}, \mathbf{y}\right) = \sum_{i=1}^{m} \max\left(1 - \mathbf{y}_i((\mathbf{X}_{\mathcal{S}})^{\mathrm{T}}\mathbf{w})_i, 0\right), \quad (3)$$

leads to the soft margin Support Vector Machine (SVM) problem[b] [26], when inserted into equation (2).

The squared loss, defined as

$$l\left((\mathbf{X}_{\mathcal{S}})^{\mathrm{T}}\mathbf{w}, \mathbf{y}\right) = \left((\mathbf{X}_{\mathcal{S}})^{\mathrm{T}}\mathbf{w} - \mathbf{y}\right)^{\mathrm{T}} \left((\mathbf{X}_{\mathcal{S}})^{\mathrm{T}}\mathbf{w} - \mathbf{y}\right), \quad (4)$$

leads to the Regularized Least-Squares (RLS) problem [34-40].

Greedy regularized least-squares

We next recall the description of greedy RLS, our linear time algorithm for greedy forward selection for RLS with LOO criterion, which was introduced by us in [49]. A detailed pseudo code of greedy RLS is presented in Algorithm 2.

Algorithm 2 Greedy RLS

```
1: a ← λ^{-1}y
2: a ← λ^{-1}y
3: C ← λ^{-1}X^T
4: S ← ∅
5: while |S| < k do
6:     e ← ∞
7:     b ← 0
8:     for i ∈ {1,...,n} \ S do
9:         u ← C_{:,i}(1 + X_i C_{:,i})^{-1}
10:        ã ← a − u(X_i a)
11:        e_i ← 0
12:        for j ∈ {1,...,m}
13:            d_j ← d_j − u_j C_{j,i}
14:            p ← y_j − (d̃_j)^{-1} ã_j
15:            e_i ← e_i + (p − y_j)^2
16:        if e_i < e then
17:            e ← e_i
18:            b ← i
19:    u ← C_{:,b}(1 + X_b C_{:,b})^{-1}
20:    a ← a − u(X_b a)
21:    for j ∈ {1,...,m} do
22:        d_j ← d_j − u_j C_{j,b}
23:    C ← C − u(X_b C)
24:    S ← S ∪ {b}
25: w ← X_S a
```

First, we consider finding a solution for the regularization problem (2) with the squared loss (4) for a fixed set of

features $\mathcal{S}$:

$$\underset{\mathbf{w}\in\mathbb{R}^{|\mathcal{S}|}}{\operatorname{argmin}} \left\{ \left((\mathbf{X}_{\mathcal{S}})^{\mathrm{T}}\mathbf{w} - \mathbf{y} \right)^{\mathrm{T}} \left((\mathbf{X}_{\mathcal{S}})^{\mathrm{T}}\mathbf{w} - \mathbf{y} \right) + \lambda \mathbf{w}^{\mathrm{T}}\mathbf{w} \right\}. \tag{5}$$

By setting the derivative of (5) with respect to $\mathbf{w}$ to zero, we get

$$\begin{aligned} \mathbf{w} &= (\mathbf{X}_{\mathcal{S}}(\mathbf{X}_{\mathcal{S}})^{\mathrm{T}} + \lambda \mathbf{I})^{-1}\mathbf{X}_{\mathcal{S}}\mathbf{y} && (6)\\ &= \mathbf{X}_{\mathcal{S}}((\mathbf{X}_{\mathcal{S}})^{\mathrm{T}}\mathbf{X}_{\mathcal{S}} + \lambda \mathbf{I})^{-1}\mathbf{y}, && (7) \end{aligned}$$

where $\mathbf{I}$ is the identity matrix and the second equality is due to the well-known matrix inversion identities (see e.g. [60]).

Before continuing, we introduce some extra notation. Let

$$\mathbf{G} = ((\mathbf{X}_{\mathcal{S}})^{\mathrm{T}}\mathbf{X}_{\mathcal{S}} + \lambda \mathbf{I})^{-1}. \tag{8}$$

While the matrix $\mathbf{G}$ is only implicitly used by the algorithms we present below, it is nevertheless a central concept in the following considerations. Moreover, let

$$\begin{aligned} \mathbf{a} &= \mathbf{G}\mathbf{y},\\ \mathbf{d} &= \mathrm{diag}(\mathbf{G}),\\ \mathbf{C} &= \mathbf{G}\mathbf{X}^{\mathrm{T}}, \end{aligned} \tag{9}$$

where $\mathrm{diag}(\mathbf{G})$ denotes a vector that consist of the diagonal entries of $\mathbf{G}$. In the literature, the entries of the vector $\mathbf{a} \in \mathbb{R}^m$ are often called the dual variables, because the solutions of (5) can be equivalently expressed as $\mathbf{w} = \mathbf{X}_{\mathcal{S}}\mathbf{a}$, as can be observed from (7).

Next, we consider a well-known efficient approach for evaluating the LOO performance of a trained RLS predictor (see e.g. [23,61]). Provided that we have the vectors $\mathbf{a}$ and $\mathbf{d}$ available, the LOO prediction for the jth training example can be obtained in constant number of floating point operations from

$$\mathbf{y}_j - (\mathbf{d}_j)^{-1}\mathbf{a}_j. \tag{10}$$

We note that (10) can be further generalized to hold-out sets larger than one (see e.g. [45]).

In order to take advantage of the computational shortcuts, greedy RLS maintains the current set of selected features $\mathcal{S} \subseteq \{1, \ldots, n\}$, the vectors $\mathbf{a}, \mathbf{d} \in \mathbb{R}^m$ and the matrix $\mathbf{C} \in \mathbb{R}^{m\times n}$. In the initialization phase of the greedy RLS algorithm (lines 1-4 in Algorithm 2) the set of selected features is empty, and hence the values of $\mathbf{a}$, $\mathbf{d}$, and $\mathbf{C}$ are initialized to $\lambda^{-1}\mathbf{y}$, $\lambda^{-1}\mathbf{1}$, and $\lambda^{-1}\mathbf{X}^{\mathrm{T}}$, respectively, where $\mathbf{1} \in \mathbb{R}^m$ is a vector having every entry equal to 1.

The middle loop of Algorithm 2 traverses through the set of $n - |\mathcal{S}|$ available features and selects the one whose addition decreases the LOO error the most. The innermost loop computes the LOO error for RLS trained with features $\mathcal{S} \cup \{i\}$ with formula (10). For this purpose, the vectors $\mathbf{a}$ and $\mathbf{d}$ must be modified so that the effect of the ith feature is removed. In addition, when the best feature is found, it is permanently added into $\mathcal{S}$ after which the vectors $\mathbf{a}$ and $\mathbf{d}$ as well as the matrix $\mathbf{C}$ are updated. Since the definitions of $\mathbf{a}$, $\mathbf{d}$, and $\mathbf{C}$ all involve the matrix $\mathbf{G}$, we first consider how the feature additions affect it. We observe that $\mathbf{G}$ corresponding to the feature set $\mathcal{S} \cup \{i\}$ can be written as

$$\begin{aligned} \widetilde{\mathbf{G}} &= ((\mathbf{X}_{\mathcal{S}})^{\mathrm{T}}\mathbf{X}_{\mathcal{S}} + (\mathbf{X}_i)^{\mathrm{T}}\mathbf{X}_i + \lambda \mathbf{I})^{-1} && (11)\\ &= \mathbf{G} - \mathbf{u}\mathbf{X}_i\mathbf{G}, && (12) \end{aligned}$$

where

$$\mathbf{u} = \mathbf{C}_{:,i}(1 + \mathbf{X}_i\mathbf{C}_{:,i})^{-1}. \tag{13}$$

The equality (12) is due to the well-known Sherman-Morrison-Woodbury formula (see e.g. [60]). Accordingly, the vector $\tilde{\mathbf{a}}$ corresponding to $\mathcal{S} \cup \{i\}$ can be written as

$$\begin{aligned} \tilde{\mathbf{a}} &= (\mathbf{G} - \mathbf{u}\mathbf{X}_i\mathbf{G})\mathbf{y}\\ &= \mathbf{a} - \mathbf{u}(\mathbf{X}_i\mathbf{a}), \end{aligned} \tag{14}$$

the jth entry of $\tilde{\mathbf{d}}$ as

$$\begin{aligned} \tilde{\mathbf{d}}_j &= (\mathbf{G} - \mathbf{u}\mathbf{X}_i\mathbf{G})_{j,j}\\ &= (\mathbf{G} - \mathbf{u}(\mathbf{C}_{:,i})^{\mathrm{T}})_{j,j}\\ &= \mathbf{d}_j - \mathbf{u}_j\mathbf{C}_{j,i}, \end{aligned} \tag{15}$$

and the cache matrix $\mathbf{C}$ as

$$\mathbf{C} - \mathbf{u}(\mathbf{X}_i\mathbf{C}).$$

By going through the matrix operations in the pseudo code of greedy RLS in Algorithm 2, it is easy to verify that the computational complexity of the whole algorithm is $O(kmn)$, that is, the complexity is linear in the number of examples, features, and selected features. Considering this in the context of the analysis of wrapper-based feature selection presented above, this means that the time spent for the selection heuristic is $O(m)$, which is far better than the approaches in which a black-box training algorithm is retrained from scratch each time a new feature is selected.

Space efficient variation

The computational efficiency of greedy RLS is sufficient to allow its use on large scale data sets such as those occurring in GWAS. However, the memory consumption may become a bottleneck, because greedy RLS keeps the matrices $\mathbf{X} \in \mathbb{R}^{n\times m}$ and $\mathbf{C} \in \mathbb{R}^{m\times n}$ constantly in memory. In GWAS, the data matrix $\mathbf{X}$ usually contains only integer-valued entries, and one byte per entry is sufficient for storage. In contrast, the matrix $\mathbf{C}$ consists of real numbers which are in most systems stored with at least four bytes per entry.

In this section, we present a variation of greedy RLS which spends less memory when dealing with large data sets. Namely, the proposed variation avoids storing the

cache matrix $\mathbf{C}$ in memory, and hence the memory consumption is dominated by storing the matrix $\mathbf{X}$. The savings can be significant if the training data is integer valued, such as in SNP datasets.

The pseudo code of this variation is given in Algorithm 3. Next, we describe its main differences with Algorithm 2 and analyze its computational complexity and memory consumption in detail. Formally, let

$$r = \min(m, |\mathcal{S}|)$$

and let

$$\mathbf{X}_{\mathcal{S}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$$

be the economy-size (see e.g. [62]) singular value decomposition (SVD) of $\mathbf{X}_{\mathcal{S}}$, where $\mathbf{U} \in \mathbb{R}^{|\mathcal{S}| \times r}$ and $\mathbf{V} \in \mathbb{R}^{m \times r}$ contain the left and the right singular vectors of $\mathbf{X}$, respectively, and $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ is a diagonal matrix containing the corresponding singular values. Note that $\mathbf{X}_{\mathcal{S}}$ has at most r nonzero singular values. Since we use the economy-size SVD, where we only need to store those singular vectors that correspond to the nonzero singular values, the size of the matrices $\mathbf{U}$ and $\mathbf{V}$ is determined by r. The computational complexity of the economy-size SVD of $\mathbf{X}_{\mathcal{S}}$ is $O(\min(m^2|\mathcal{S}|, m|\mathcal{S}|^2))$ (see e.g. [62]). Substituting the decomposed data matrix into (8), we get

$$\begin{aligned}
\mathbf{G} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1} \\
&= (\mathbf{V}\mathbf{\Sigma}^{\mathrm{T}}\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} + \lambda\mathbf{I})^{-1} \\
&= (\mathbf{V}\mathbf{\Sigma}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}} + \lambda\mathbf{I})^{-1} \\
&= \mathbf{V}((\mathbf{\Sigma}^{\mathrm{T}}\mathbf{\Sigma} + \lambda\mathbf{I})^{-1} - \lambda^{-1}\mathbf{I})\mathbf{V}^{\mathrm{T}} + \lambda^{-1}\mathbf{I} \\
&= \mathbf{V}\mathbf{\Omega}\mathbf{V}^{\mathrm{T}} + \lambda^{-1}\mathbf{I},
\end{aligned}$$

where

$$\mathbf{\Omega} = (\mathbf{\Sigma}^{\mathrm{T}}\mathbf{\Sigma} + \lambda\mathbf{I})^{-1} - \lambda^{-1}\mathbf{I}$$

and the dimensions of the identity matrices are either $r \times r$ or $m \times m$ depending on the context. Note that inverting $\mathbf{\Sigma}^{\mathrm{T}}\mathbf{\Sigma} + \lambda\mathbf{I}$ requires only $O(r)$ time, because it is a diagonal matrix. Now, the ith column of the matrix $\mathbf{C}$ can be written as

$$\mathbf{c} = \mathbf{V}(\mathbf{\Omega}(\mathbf{V}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}})) + \lambda^{-1}(\mathbf{X}_i)^{\mathrm{T}} \tag{16}$$

which can be computed in $O(mr)$ time.

Algorithm 3 Space Efficient Greedy RLS

$\mathbf{a} \leftarrow \lambda^{-1}\mathbf{y}$
$\mathbf{d} \leftarrow \lambda^{-1}\mathbf{1}$
$\mathcal{S} \leftarrow \emptyset$
$\mathbf{V} \leftarrow \mathbf{0}$
$\mathbf{\Omega} \leftarrow 0$
while $|\mathcal{S}| < k$ **do**
  $e \leftarrow \infty$
  $b \leftarrow 0$
  for $i \in \{1, \ldots, n\} \setminus \mathcal{S}$ **do**
    $\mathbf{c} \leftarrow \mathbf{V}(\mathbf{\Omega}(\mathbf{V}^{\mathrm{T}}(\mathbf{X}_i)^{\mathrm{T}})) + \lambda^{-1}(\mathbf{X}_i)^{\mathrm{T}}$
    $\mathbf{u} \leftarrow \mathbf{c}(1 + \mathbf{X}_i\mathbf{c})^{-1}$
    $\tilde{\mathbf{a}} \leftarrow \mathbf{a} - \mathbf{u}(\mathbf{X}_i\mathbf{a})$
    $e_i \leftarrow 0$
    for $j \in \{1, \ldots, m\}$ **do**
      $\tilde{\mathbf{d}}_j \leftarrow \mathbf{d}_j - \mathbf{u}_j\mathbf{c}_j$
      $p \leftarrow \mathbf{y}_j - (\tilde{\mathbf{d}}_j)^{-1}\tilde{\mathbf{a}}_j$
      $e_i \leftarrow e_i + (\mathbf{y}_j - p)^2$
    if $e_i < e$ **then**
      $e \leftarrow e_i$
      $b \leftarrow i$
  $\mathbf{c} \leftarrow \mathbf{V}(\mathbf{\Omega}(\mathbf{V}^{\mathrm{T}}(\mathbf{X}_b)^{\mathrm{T}})) + \lambda^{-1}(\mathbf{X}_b)^{\mathrm{T}}$
  $\mathbf{u} \leftarrow \mathbf{c}(1 + \mathbf{X}_b\mathbf{c})^{-1}$
  $\mathbf{a} \leftarrow \mathbf{a} - \mathbf{u}\mathbf{X}_b\mathbf{a})$
  for $j \in \{1, \ldots, m\}$ **do**
    $\mathbf{d}_j \leftarrow \mathbf{d}_j - \mathbf{u}_j\mathbf{c}_j$
  $\mathbf{\Sigma}, \mathbf{V}^{\mathrm{T}} \leftarrow \mathrm{SVD}(\mathbf{X}_{\mathcal{S}})$
  $\mathbf{\Omega} \leftarrow (\mathbf{\Sigma}^{\mathrm{T}}\mathbf{\Sigma} + \lambda\mathbf{I})^{-1} - \lambda^{-1}\mathbf{I}$
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{b\}$
$\mathbf{w} \leftarrow \mathbf{X}_{\mathcal{S}}\mathbf{a}$

If k is the number of features that will be selected, SVD has to be computed k times, resulting in complexity $O(\min(k^3m, k^2m^2))$. The computation of (16) is performed $O(kn)$ times resulting in a complexity $O(\min(k^2mn, km^2n))$, which dominates the overall computational complexity of this variation. Since storing and updating the cache matrix $\mathbf{C}$ is not required in Algorithm 3, the memory consumption is dominated by the data matrix $\mathbf{X}$, which can, in the context of GWAS data, be stored as an array of integers. In addition, computing and storing the right singular vectors requires a real valued matrix of size $m \times r$. However, this has a negligible memory consumption unless both k and m are close to n, which is usually not the case in GWAS. To conclude, in GWAS experiments, the memory consumption of Algorithm 3 is about one fifth of that of Algorithm 2 because it avoids storing $\mathbf{C}$ that requires four bytes of memory per entry whereas $\mathbf{X}$ requires only one. The timing comparison of the space efficient model when compared with the normal greedy RLS can be seen in Figure 1.

Results and discussion

In the experiments, we first demonstrate the scalability of the greedy RLS method to large-scale GWAS learning. As a point of comparison, we present runtimes for a wrapper-based selection for an SVM classifier to which we refer as SVM-wrapper. The greedy RLS algorithm was implemented in C++ to allow for minimal overhead with regards to looping over large datasets and to allow efficient future adaptations of the code, such as parallelization to take advantage of both shared and distributed memory systems. The space-efficient version of the greedy RLS method was implemented in Python, in order to make

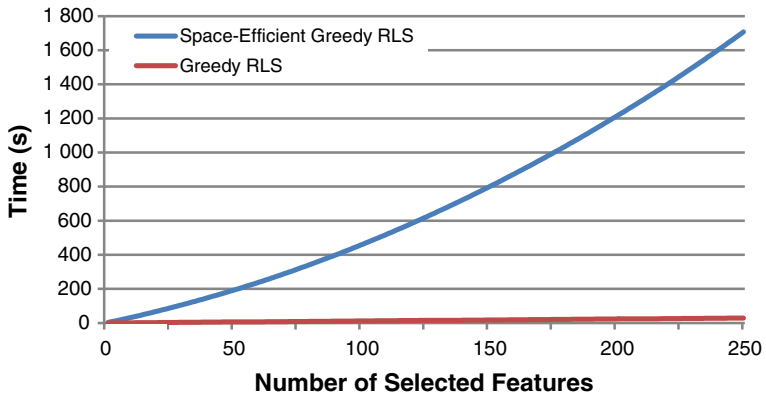

Figure 1 Comparison of the greedy RLS implementations. Plot showing the timing comparison (in seconds) for the two variations of greedy RLS. Note the linearity in the greedy RLS curve compared to the quadratic nature of the space-efficient version with respect to the number of selected features. The run was based on randomly sampled datasets with 1,000 training examples and 10,000 features.

use of its well established numerical analysis packages for computing the required singular value decompositions. For the SVM-wrapper, we chose to use the LibSVM in Weka 3.7.3 [63,64] and LibSVM in the e1071 package in R [65-67]. This choice was made because these environments have been commonly used in other studies that have attempted to solve similar problems, and since the LibSVM package itself is known to be one of the most efficient existing SVM implementations. The scalability experiments were run on randomly sampled subsets of the WTCCC HT-NBS dataset [1]. The predictive performance of greedy RLS is demonstrated on an independent test set, and the biological relevance of the results are briefly analyzed.

Scalability experiments

In the scalability experiment, the number of training examples was held fixed at 1,000, but the number of features was incrementally increased. The considered feature set sizes were 10, 100, 1,000, 10,000, 100,000, 250,000 and 500,000. All methods implemented greedy, wrapper-based selections. The number of selected features was set to 10. By definition, greedy RLS uses LOO-CV as the selection criterion. We used the less computationally demanding 10-fold CV for the SVM-wrappers, because of the high computational costs of performing LOO for SVMs. The selection criterion for the individual features in the dataset was based on the root mean squared error (RMSE). The choice was made for computational reasons, since computing RMSE can be done in linear time, whereas computing the more commonly used AUC measure has $O(m \log(m))$ complexity due to a required sorting operation. RMSE as a selection criterion can be expected to work well as long as the class distributions are not very imbalanced (see e.g. [68]).

In Figure 1, we present the run-time comparisons of the two proposed variations of greedy RLS. As expected from the theory presented in the Methods section, along with the speed advantages of C++ over Python, the fast implementation turned out to be orders of magnitude faster than the space-efficient version. This performance increase comes at a cost requiring higher memory usage, hence making it infeasible to run the basic greedy RLS on the GWAS containing a very large number of training examples. For these scenarios it would be necessary to implement the space-efficient variation.

From the runtimes in Figure 2 it can be ascertained that other than greedy RLS, the current, commonly used algorithms for wrapper-based methods are not computationally efficient enough to scale up to entire GWAS. The R implementation of the SVM-wrapper took over 5 hours to select 10 features out of 10,000 and at 100,000 features the run had to be terminated early since it exceeded the pre-determined cut-off time of 24 hours. In the commonly used Weka environment, the approach scaled worse with the program not being able to complete the selection in a 24 hour period for the dataset consisting of 10,000 features. In contrast, greedy RLS computed the selection process even on 500,000 features in 1 minute while the space-efficient greedy RLS performed the feature selection process on the same dataset in under 24 minutes (see Figure 2).

Generalization Capability

In addition to the run time comparisons, we also conducted a sample run on the entire WTCCC HT-NBS dataset to predict an individual's risk for hypertension and to investigate whether greedy RLS can accurately discriminate between the risk classes on an independent test set. In order to reduce the variance of the results, we adopt

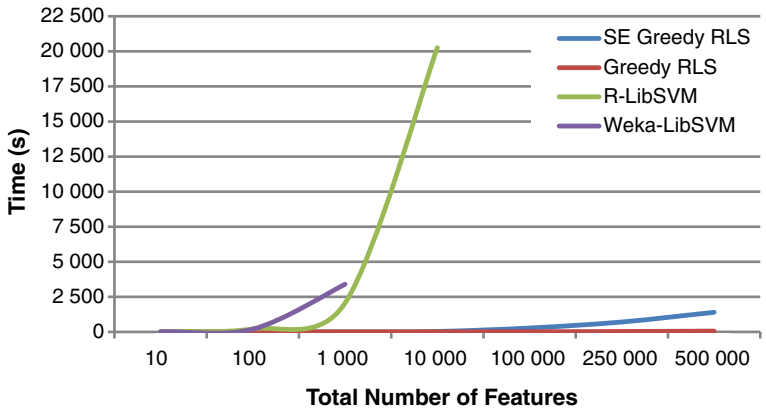

Figure 2 Timing results of various wrapper-based methods. Plot showing the comparisons between the timing results of the different feature selection implementations. Greedy RLS and space-efficient (SE) greedy RLS both used LOO, greedy feature selection and an RLS classifier, while Weka and R implemented LibSVM, used greedy forward selection as the search strategy and 10-fold CV as the selection criterion.

the so-called nested CV approach (see e.g. [69-71]), in which an external CV is used for estimating the generalization capability of the learned models and an internal CV for assessing the quality of feature sets separately during each round of the external CV. First, the whole dataset was divided into three equally sized folds. Each of the three folds were used as a test set one at a time, while the remaining two were used to form a training set. Finally, the results of these three external CV rounds were averaged. The internal selection process itself with the LOO-CV criterion was run on the training sets, and up to 50 features were selected. The test folds were used only for computing the final test results for the models obtained after running the whole feature selection process.

In Figure 3, we present the leave-one-out cross-validated mean squared errors on the training sets in the three external CV rounds, used as the selection criterion by greedy RLS. The three selection criterion curves behave quite similarly, even if the corresponding training sets overlap with each other only by half of their size. The curves are monotonically decreasing, which is to be expected, as it is very likely that the selection criterion overfits due to the excessive number of available features to choose from (see e.g. [69] for further discussion). Clearly, they are not trustworthy in assessing the true prediction performance of the learned models. A separate test fold is thus necessary for this purpose.

During each round of the external CV, after the selection process has been performed for a number of features ranging from 1 to 50, the AUCs of the learned models are evaluated on the independent test fold that was not seen during the selection (a.k.a nested CV). The results averaged over the three test folds are presented in Figure 4. At first the AUC keeps increasing with the number of

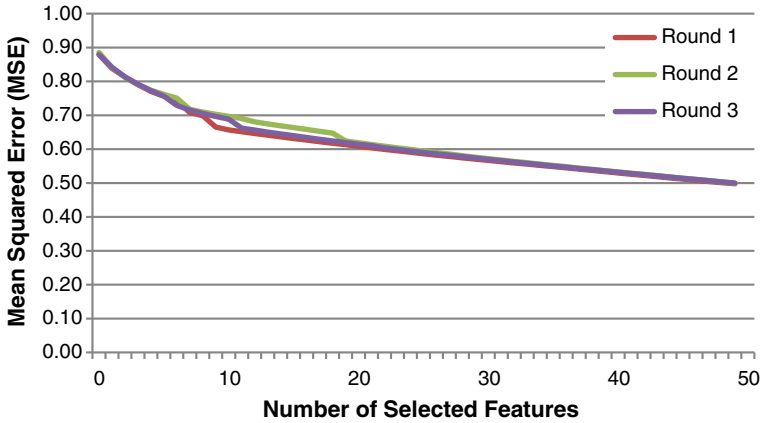

Figure 3 Mean squared error for greedy RLS. The plot displays the mean squared LOOCV errors used as a selection criteria by Greedy RLS during the three rounds of the external CV. It can be observed that as expected, the errors are consistently decreasing since the selection criterion quickly overfits to the training folds during the selection process.

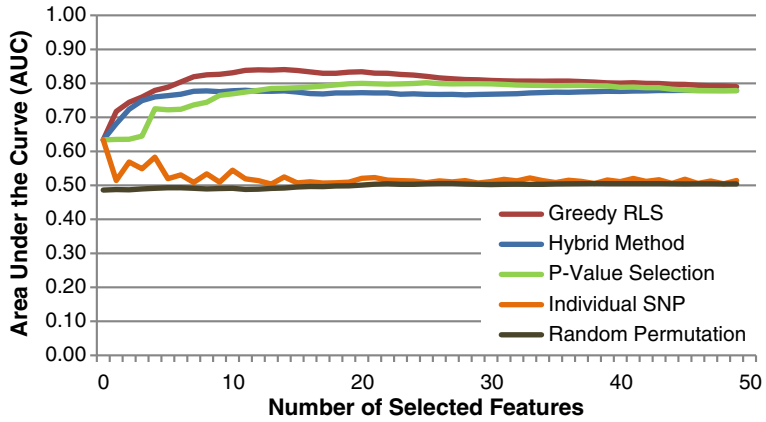

Figure 4 Comparison of feature selection approaches in terms of predictive accuracy. The prediction performances of the models learned by greedy RLS were assessed using area under the ROC curve (AUC), averaged over the three folds of an external CV. On each round of the external CV, the training set on which the features are selected consisted of 2/3, and the independent test set on which the prediction performance is measured, 1/3 of the 3,410 subjects, with a stratified training/test split. The graph also displays the individual SNP AUCs for each of the variants selected by greedy RLS. Further, results are depicted for a p-value based filtering in which the top k most significant features were selected. We also present a curve that displays the results for the hybrid method in which greedy RLS runs on the top k features ranked according to their p-values. Finally, we present a random permutation on the class labels and running greedy RLS on this randomized dataset.

selected features reaching its peak 0.84 AUC at 15 features, after which it starts decreasing. The result demonstrates that the selection process must be stopped early enough in order to avoid overfitting. Note that as observed from the Figure 3, the leave-one-out error does not provide a reliable criterion for determining the stopping point due to its use in the feature selection process. Rather, the AUC observed on an independent test fold not used during selection can be used to determine the number of features to select.

We compared the prediction performance of greedy RLS to that of two commonly used approaches in GWAS, which are both based on training a classifier on feature sets selected through filtering [3,7,18]. The reference methods start by using p-value based filtering to rank the features. The p-values were computed on the training sets using PLINK, based on Fisher's exact test on a 3x2 contingency table of the genotypes. The filter approach is based on training a RLS classifier directly on the top k features having the smallest p-values. The second approach is a hybrid method, where the filters are first used to select 50 features with the smallest p-values and then greedy RLS is used to select k features from this set of pre-filtered features afterwards. The baseline results are based on the same three-fold CV setting as the results of greedy RLS.

As expected, the first feature selected by all of the approaches was the same since the LOO error employed by the greedy RLS as a selection criterion does not considerably differ from the statistical tests when computed for a single feature. Afterwards, however, greedy RLS begins to outperform the baseline methods, as the filter-based and hybrid approaches tend to select features that may be highly correlated with the already selected features. From Figure 4 it can be noted that while the performance of the hybrid method on the test set performs similarly to that of greedy RLS for the first couple of features, it relatively quickly begins to level off around 0.77 AUC, peaking at 0.78, below that of greedy RLS's maximum. In contrast, the filter method requires considerably more features before its prediction performance gets close to 0.77, peaking at 0.80 before beginning to decline. The results indicate that through the use of wrapper based feature selection, it is possible to identify sets of features that have the capacity to outperform those selected by filter or hybrid methods. The total time to select the top features over the three folds of the external CV was approximately 26 minutes.

To measure the performance of the selected variants in a single-feature association analysis, the individual AUC of each of the 50 selected features was computed (see Figure 4). It can be observed that most of the single-feature AUCs are close to a random level. The maximal AUC (0.63) occurs for the first selected variant. This lack of power for the majority of the selected SNPs to distinguish between cases and controls would lead to the conclusion that the selected variants individually are not associated with the disease. On the contrary, when the combined phenotypic effect of these variants is taken into account with the RLS algorithm, much more accurate models can be trained.

To demonstrate that the experimental setup was implemented correctly, so that there is no information leak

between the training and test data, we conducted a feature selection based on a random permutation of the class labels. The data and the training/test set splits were identical to those used in the original run, with the only difference being that the class labels in the dataset were randomly permuted prior to running the experiments. The top fifty features were then selected as before and the resulting AUC of the trained classifier implemented on the test set was recorded. As expected, the randomized class labels run resulted in random AUCs regardless of the number of features that were selected (see Figure 4), indicating that the results of a random labeling can not generalize beyond the original data, whereas the original SNPs have a greater ability to make accurate predictions on independent datasets.

Feature selection results

The application of machine learning algorithms to complex GWAS datasets is not a trivial task, and there are numerous factors that can strongly alter the result in such settings. Without a solid understanding of the methodologies, it is very easy for researchers to come to incorrect conclusions about the results presented to them. Additionally, these methods can be heavily affected by any quality control procedures that are implemented. We show here that a number of the selected features are linked to prior identified factors in other published manuscripts. However, wrapper-based approaches are prone to selecting features that have unforeseen epistatic interactions amongst them and it can therefore be expected that not all of the selected features will be present in the literature. As such, while certain variants with known phenotypes [72-74], such as blood pressure, can be expected to be selected, as with any GWAS, it is likely that previously unidentified SNPs may also demonstrate disease associations.

To study the cellular mechanisms behind the selected variants, we mapped the top selected features identified by greedy RLS run on the entire cohort. To map the phenotypes we conducted a literary review of the SNPs and genes that are located within 20,000 base-pairs based on results from the dbSNP database [75]. The number of features to be analyzed, 15, was determined by the point at which the maximal AUC was obtained from the nested CV, as explained in the previous section. Of the fifteen variants, five have been identified in other publications to have either known or possible links (through gene mapping) to hypertension and related phenotypes (see Table 1): HTR3B (two variants), MIR378D1, rs10771657, SCOC.

Variants with interesting mappings included MIR378D1, HTR3B, SCOC and rs10771657. MIR378D1, better known as microRNA 378d-1, is a gene located on chromosome 4 which is involved in the function of microRNA-378. It has been shown previously that mircoRNA-378 promotes angiogenesis through its overexpression and targeting of Sufu-associated pathways [76]. Angiogenesis, the process of new blood-vessels growing from existing ones, is associated with hypertension in [77]. Also, SCOC (short coiled-coil protein) has been significantly associated with hypertension [78]. HTR3B was previously identified as having a possible link to the control of blood pressure in rats, through its central influence on the sympathoinhibitory mechanism [74,79]. While this study focused on rats, it provides enough evidence to warrant HTR3B as being a candidate for examination in human-based GWAS studies. Similarly, rs10771657 was examined in other studies and identified as having a statistical association towards pulmonary function, a trait related to hypertension [80].

Table 1 Variants selected by the greedy RLS algorithm

SNP	Gene	Chromosome	Position
rs7837736	Intergenic	8	15296703
rs1908465	Intergenic	8	15308433
rs17116117	HTR3B	11	113801591
rs10843660	Intergenic	12	30368457
rs17667894	MIR17HG	13	92014309
rs17116145	HTR3B	11	113804326
rs10771657	Intergenic	12	30359294
rs17459885	Intergenic	12	30360879
rs16837871	MIR378D1	4	5941112
rs7691494	C4orf50	4	5942649
rs6588810	ASMT	X	1753118
rs11005510	Intergenic	10	58532989
rs6840033	SCOC	4	141228861
rs10499044	Intergenic	6	107141295
rs2798360	LOC100422737	6	107148473

The list of the top 15 selected features on the entire cohort. The first column represents the SNP identifier. The second column indicates which gene the particular SNP is mapped to, or if it can not be mapped to any gene then it is marked as an intergenic sequence. The third and fourth columns are the chromosome number and base-pairs location of the SNP, respectively.

Nine out of the top fifteen selected features were also among the fifty features with the lowest p-values. As already discussed in previous section, the filter methods tend to select features that are correlated with each other, and therefore some of the features among the ones with the lowest p-values will not be selected by greedy RLS because of their redundancy with the previously selected features. Moreover, in contrast to the filter methods, all the features selected by greedy RLS may not be very informative individually but will be helpful for constructing a predictor when used together with other genetic features. We therefore believe that there is a strong possibility that the genetic features selected by greedy RLS

are linked to the underlying biology, even if all of their disease-associations have not yet been established.

Materials

Study cohort

For building and testing the model we examined data from the Wellcome Trust Case-Control Consortium's (WTCCC) study cohorts along with the set of controls from the UK National Blood Service Control Group (NBS). WTCCC is a group of 50 UK-based research groups whose aim is to better understand patterns amongst the genetic variants and their relation to disease onset [1].

From the WTCCC data cohorts we chose to examine a single case study, the Hypertension (HT) dataset in conjunction with the NBS controls set [1]. The original dataset consisted of 3,501 individuals and 500,568 SNPs distributed across 23 chromosomes that were originally sequenced with the Affymetrix 500k chip. From this set, 91 individuals and 30,956 SNPs were removed based on the exclusion lists for the associated datasets [1]. This reduced set was further filtered in PLINK based on standard quality control procedures including implementing filters that excluded features that failed the Hardy-Weinberg equilibrium in the controls with a threshold of $P < 1 \times 10^{-3}$, a minor allele frequency of 1%, a missing rate of 5%, along with a filter eliminating individuals who were missing data from more than 5% of SNPs [2,3,81-85]. After this quality control the dataset incorporated 3,410 individuals and 404,452 SNPs. As the aim of this study was to test the feasibility of the proposed algorithm, rather than the suitability of the selected features, we omitted advanced filtering methodologies such as population stratification or the adjustment of call rates to more conservative values.

Data treatment

The RLS and SVM based methods require, that the features are encoded as numerical values. The SNP data that was used in the runs were 0, 1 and 2 corresponding to the minor allele count for the genetic feature, representing the major allele homozygote, the heterozygote and the minor allele homozygote respectively. For the scalability experiment, the runs used 1,000 examples and 10, 100, 1,000, 10,000, 100,000, 250,000 and 500,000 features. The file formats used for the data input were ARFF, binary and binary file formats for Weka, R and greedy RLS respectively.

Conclusions

This paper is a proof-of-concept of wrapper methods being able to scale up to entire GWAS and having the capacity to perform better than the traditional filter or hybrid methods. Thorough consideration of the effects of different quality control procedures on the results, and biological validation of the found feature sets falls outside the scope of this study. The greedy RLS algorithm is the first known method that has been successfully used to perform a wrapper-based feature selection on an entire GWAS. This novel approach created a solution for an important problem, providing highly accurate results. Both the computational complexity analysis and practical scalability experiments demonstrate that the method scales well to large datasets. One critical question that remains is, what is the optimum number of features to select in such as study. While there is no definitive answer, our results indicate that even a small number of features may provide accurate prediction models.

The scalability of greedy RLS was compared to that of SVM-based wrapper methods, namely LibSVM in both the e1071 library in R and through a command line interface with the Weka software package. We demonstrate that unlike the proposed method, the other publicly available methods have too high computational runtimes to be suitable for GWAS data sets. This is not to say that there do not exist other equally valid machine learning algorithms that could handle this task. However, our work is the first known implementation of wrapper based selection that has been demonstrated to scale to entire genome scans in GWAS. Machine learning-based feature selection is a powerful tool, capable of discovering unknown relationships amongst feature subsets. However, researchers need to account for the computational complexities involved in scaling the wrapper-based feature selection methods up to GWAS. Implementation of wrapper approaches through the use of the learning algorithm as a black box inside the wrapper is simply not feasible on GWAS scale. Rather, one needs to know how to optimally implement the procedure in order to re-use computations done at different search steps and round of cross-validation. Embedding of the computations is the central key to allowing greedy RLS to scale to GWAS.

Endnotes

[a]In the literature, the formula of the linear predictors often also contain a bias term. Here, we assume that if such a bias is used, it will be realized by using an extra constant valued feature in the data points.

[b]The method is often presented in the literature as an alternative, but equivalent formulation as a constrained optimization problem.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Tapio Pahikkala and Sebastian Okser made equal contributions to this article. Tapio Pahikkala - Design, implementation, and formalization of the greedy RLS algorithm and drafting of the manuscript. Sebastian Okser - Design of the experiments, implementation of the algorithm on the GWAS, analysis of the results and drafting of the manuscript. Antti Airola - Participated in the

development process of greedy RLS, experimental design and implementation, and drafting of the manuscript. Tapio Salakoski - Supervision of the research and methodology design. Tero Aittokallio - Conceived the study, participated in the design of the experiment and in drafting of the manuscript. All authors read and approved the final manuscript.

Acknowledgements
This project was funded by the Academy of Finland (grants 120569, 133227, 134020, 140880, 218310), the Turku University Foundation and the Finnish Cultural Foundation. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the 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, Turku, Finland. [3]Department of Mathematics, University of Turku, Turku, Finland. [4]Data Mining and Modeling group, Turku Centre for Biotechnology, Turku, Finland. [5]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland.

Received: 8 June 2011 Accepted: 23 April 2012
Published: 2 May 2012

References

1. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ, Todd JA, Donnelly P, et al.: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447:**661–678.
2. Evans DM, Visscher PM, Wray NR: **Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk.** *Human Mol Genet* 2009, **18**(18):3525–3531.
3. Wei Z, Wang K, Qu HQ, Zhang H, Bradfield J, Kim C, Frackleton E, Hou C, Glessner JT, Chiavacci R, Stanley C, Monos D, Grant SFA, Polychronakos C, Hakonarson H: **From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes.** *PLoS Genet* 2009, **5**(10):e1000678.
4. Holmes MV, Harrison S, Talmud PJ, Hingorani AD, Humphries SE: **Utility of genetic determinants of lipids and cardiovascular events in assessing risk.** *Nat Rev Cardiology* 2011, **8**(4):207–221.
5. Krawczyk M, Müllenbach R, Weber SN, Zimmer V, Lammert F: **Genome-wide association studies and genetic risk assessment of liver diseases.** *Nat Rev Gastroenterol Hepatol* 2010, **7**(12):669–681.
6. Juonala M, Viikari JS, Kahonen M, Taittonen L, Ronnemaa T, Laitinen T, Maki-Torkko N, Mikkila V, Rasanen L, Akerblom HK, Pesonen E, Raitakari OT: **Origin as a determinant of carotid artery intima-media thickness and brachial artery flow-mediated dilation: the cardiovascular risk in young finns study.** *Arterioscler Thromb Vasc Biol* 2005, **25**(2):392–398.
7. Okser S, Lehtimäki T, Elo LL, Mononen N, Peltonen N, Kähönen M, Juonala M, Fan YM, Hernesniemi JA, Laitinen T, Lyytikäinen LP, Rontu R, Eklund C, Hutri-Kähönen N, Taittonen L, Hurme M, Viikari JSA, Raitakari OT, Aittokallio T: **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.
8. Bleumink GS, Schut AF, Sturkenboom MC, Deckers JW, van Duijn, C M, Stricker BH: **Genetic polymorphisms and heart failure.** *Genet Med* 2004, **6**(6):465–474.
9. Levy D, Ehret GBB, Rice K, Verwoert GCC, Launer LJJ, Dehghan A, Glazer NLL, Morrison ACC, Johnson ADD, Aspelund T, Aulchenko Y, Lumley T, Köttgen A, Vasan RSS, Rivadeneira F, Eiriksdottir G, Guo X, Arking DEE, Mitchell GFF, Mattace-Raso FUSU, Smith AVV, Taylor K, Scharpf RBB, Hwang SJJ, Sijbrands EJGJ, Bis J, Harris TBB, Ganesh SKK, O'Donnell CJJ, Hofman A, Rotter JII, Coresh J, Benjamin EJJ, Uitterlinden AGG, Heiss G, Fox CSS, Witteman JCMC, Boerwinkle E, Wang TJJ, Gudnason V, Larson MGG, Chakravarti A, Psaty BMM, van Duijn CMM: **Genome-wide association study of blood pressure and hypertension.** *Nat Genet* 2009, **41:**677–687.
10. Moore JH, Williams SM: **Epistasis and its implications for personal genetics.** *Am J Human Genet* 2009, **85**(3):309–320.
11. Pattin K, Moore J: **Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases.** *Human Genet* 2008, **124:**19–29.
12. Li M, Romero R, Fu WJ, Cui Y: **Mapping Haplotype-haplotype interactions with adaptive LASSO.** *BMC Genet* 2010, **11:**79.
13. Plomin R, Haworth CMA, Davis OSP: **Common disorders are quantitative traits.** *Nat Rev Genet* 2009, **10**(12):872–878.
14. Mckinney BA, Reif DM, Ritchie MD, Moore JH: **Machine learning for detecting gene-gene interactions: a review.** *Appl Bioinf* 2006, **5**(2):77–88 .
15. Szymczak S, Biernacka JM, Cordell HJ, González-Recio O, König IR, Zhang H, Sun YV: **Machine learning in genome-wide association studies.** *Genet Epidemiol* 2009, **33**(Suppl 1):S51–S57.
16. Ban HJ, Heo JY, Oh KS, Park KJ: **Identification of Type 2 diabetes-associated combination of SNPs using support vector machine.** *BMC Genet* 2010, **11:**26.
17. Saeys Y, Inza I, Larrañaga P: **A review of feature selection techniques in bioinformatics.** *Bioinformatics* 2007, **23**(19):2507–2517.
18. Long N, Gianola D, Rosa G, Weigel K, Avendaño S: **Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers.** *J Animal Breeding Genet* 2007, **124**(6):377–389.
19. Tang EK, Suganthan PN, Yao X: **Gene selection algorithms for microarray data based on least squares support vector machine.** *BMC Bioinf* 2006, **7:**95.
20. Roshan U, Chikkagoudar S, Wei Z, Wang K, Hakonarson H: **Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest.** *Nucleic Acids Res* 2011, **39**(9):e62.
21. Kohavi R, John GH: **Wrappers for feature subset selection.** *Artif Intell* 1997, **97**(1-2):273–324.
22. Lachenbruch PA: **An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis.** *Biometrics* 1967, **23**(4):639–645.
23. Elisseeff A, Pontil M: **Leave-one-out error and stability of learning algorithms with applications.** In *Advances in Learning Theory: Methods, Models and Applications, Volume 190 of NATO Science Series III: Computer and Systems Sciences*. Edited by Suykens J, Horvath G, Basu S, Micchelli C, Vandewalle J. Amsterdam: IOS Press; 2003:111–130.
24. Inza I, Larrañaga P, Blanco R, Cerrolaza AJ: **Filter versus wrapper gene selection approaches in DNA microarray domains.** *Artif Intell Med* 2004, **31**(2):91–103.
25. Moore JH, Asselbergs FW, Williams SM: **Bioinformatics challenges for genome-wide association studies.** *Bioinformatics* 2010, **26**(4):445–455.
26. Vapnik VN: *The Nature of Statistical Learning Theory*. New York: Springer-Verlag New York Inc.; 1995.
27. Ben-Dor A, Bruhn L, Friedman N, Nachman I, Schummer M, Yakhini Z: **Tissue classification with gene expression profiles.** *J Comput Biol* 2000, **7**(3-4):559–583.
28. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: **Support vector machine classification and validation of cancer tissue samples using microarray expression data.** *Bioinformatics* 2000, **16**(10):906–914.
29. Peng S, Xu Q, Ling XB, Peng X, Du W, Chen L: **Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines.** *FEBS Lett* 2003, **555**(2):358–362 .
30. Huerta EB, Duval B, Hao JK: **A hybrid GA/SVM approach for gene selection and classification of microarray data.** In *EvoWorkshops 2006, LNCS 3907*. Berlin, Heidelberg, Germany: Springer; 2006:34–44.
31. Duval B, Hao JK: **Advances in metaheuristics for gene selection and classification of microarray data.** *Brief Bioinf* 2010, **11:**127–141.
32. Guyon I, Weston J, Barnhill S, Vapnik V: **Gene selection for cancer classification using support vector machines.** *Mach Learn* 2002, **46**(1-3):389–422.
33. Liu Q, Yang J, Chen Z, Yang MQ, Sung A, Huang X: **Supervised learning-based tagSNP selection for genome-wide disease classifications.** *BMC Genomics* 2008, **9**(Suppl 1):S6.
34. Hoerl AE, Kennard RW: **Ridge regression: biased estimation for nonorthogonal problems.** *Technometrics* 1970, **12:**55–67.

35. Poggio T, Girosi F: **Networks for approximation and learning.** *Proc IEEE* 1990, **78**(9).
36. Saunders C, Gammerman A, Vovk V: **Ridge regression learning algorithm in dual variables.** In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998)*. San Francisco: Morgan Kaufmann Publishers Inc.; 1998:515–521.
37. Suykens JAK, Vandewalle J: **Least squares support vector machine classifiers.** *Neural Process Lett* 1999, **9**(3):293–300.
38. Suykens J, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J: *Least Squares Support Vector Machines*. Singapore: World Scientific Pub Co.; 2002.
39. Rifkin R, Yeo G, Poggio T: **Regularized least-squares classification.** In *Advances in Learning Theory: Methods, Model and Applications, Volume 190 of NATO Science Series III: Computer and System Sciences*. Edited by Suykens J, Horvath G, Basu S, Micchelli C, Vandewalle J. Amsterdam: IOS Press; 2003:131–154.
40. Poggio T, Smale S: **The mathematics of learning: dealing with data.** *Not Am Math Soc (AMS)* 2003, **50**(5):537–544.
41. Fung G, Mangasarian OL: **Proximal support vector machine classifiers.** In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2001)*. New York: ACM; 2001:77–86.
42. Rifkin R: **Everything Old Is New Again: A Fresh Look at Historical Approaches in Machine Learning.** *PhD thesis*, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA 2002.
43. Zhang P, Peng J: **SVM vs regularized least squares classification.** In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR 2004)*. Edited by Kittler J, Petrou M, Nixon M. Washington: IEEE Computer Society; 2004:176–179.
44. Vapnik V: *Estimation of Dependences Based on Empirical Data*. New York: Springer; 1982.
45. Pahikkala T, Boberg J, Salakoski T: **Fast n-Fold cross-validation for regularized least-squares.** In *Proceedings of the Ninth Scandinavian Conference on Artificial Intelligence (SCAI 2006)*. Edited by Honkela T, Raiko T, Kortela J, Valpola H. Otamedia: Espoo; 2006:83–90.
46. Daemen A, Gevaert O, Ojeda F, Debucquoy A, Suykens J, Sempoux C, Machiels JP, Haustermans K, De Moor B: **A kernel-based integration of genome-wide data for clinical decision support.** *Genome Med* 2009, **1**(4):39.
47. Chen PC, Huang SY, Chen W, Hsiao C: **A new regularized least squares support vector regression for gene selection.** *BMC Bioinf* 2009, **10:**44.
48. Ojeda F, Suykens JA, Moor BD: **Low rank updated LS-SVM classifiers for fast variable selection.** *Neural Networks* 2008, **21**(2–3):437–449.
49. Pahikkala T, Airola A, Salakoski T: **Speeding up greedy forward selection for regularized least-squares.** In *Proceedings of The Ninth International Conference on Machine Learning and Applications (ICMLA 2010)*. Edited by Draghici S, Khoshgoftaar TM, Palade V, Pedrycz W, Wani MA, Zhu X: IEEE Computer Society; 2010.
50. Paynter NP, Chasman DI, Paré G, Buring JE, Cook NR, Miletich JP, Ridker PM: **Association between a literature-based genetic risk score and cardiovascular events in women.** *J Am Med Assoc* 2010, **303**(7): 631–637.
51. 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.
52. Kwon S, Cui J, Rhodes SL, Tsiang D, Rotter JI, Guo X: **Application of Bayesian classification with singular value decomposition method in genome-wide association studies.** *BMC proc* 2009, **3**(Suppl 7):S9.
53. D'Angelo GM, Rao D, Gu CC: **Combining least absolute shrinkage and selection operator (LASSO) and principal-components analysis for detection of gene-gene interactions in genome-wide association studies.** *BMC Proc* 2009, **3**(Supp 7):S62.
54. He Q, Lin DYY: **A variable selection method for genome-wide association studies.** *Bioinformatics* 2011, **27:**1–8.
55. Rodin AS, Litvinenko A, Klos K, Morrison AC, Woodage T, Coresh J, Boerwinkle E: **Use of wrapper algorithms coupled with a random forests classifier for variable selection in large-scale genomic association studies.** *J Comput Biol* 2009, **16**(12):1705–1718.
56. Shi G, Boerwinkle E, Morrison AC, Gu CC, Chakravarti A, Rao DC: **Mining gold dust under the genome wide significance level: a two-stage approach to analysis of GWAS.** *Genet Epidemiol* 2011, **35**(2):111–118.
57. John GH, Kohavi R, Pfleger K: **Irrelevant features and the subset selection problem.** In *Proceedings of the Eleventh International Conference on Machine Learning (ICML 1994)*. Edited by Cohen WW, Hirsch H. San Fransisco: Morgan Kaufmann Publishers; 1994:121–129.
58. Evgeniou T, Pontil M, Poggio T: **Regularization networks and support vector machines.** *Adv Comput Math* 2000, **13:**1–50.
59. Shawe-Taylor J, Cristianini N: *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University Press; 2004.
60. Henderson HV, Searle SR: **On deriving the inverse of a sum of matrices.** *SIAM Rev* 1981, **23:**53–60.
61. Rifkin R, Lippert R: **Notes on Regularized Least Squares.** Tech. Rep. MIT-CSAIL-TR-2007-025, Massachusetts Institute of Technology 2007.
62. Golub GH, Van Loan C: *Matrix Computations,* second edition. Baltimore and London: Johns Hopkins University Press; 1989.
63. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH: **The WEKA data mining software: an update.** *ACM SIGKDD Explorations Newsletter* 2009, **11:**10–18.
64. **Weka3: Data Mining Software in Java** [http://www.cs.waikato.ac.nz/ml/weka/]
65. R Development Core Team: *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing; 2008. [http://www.R-project.org]. [ISBN 3-900051-07-0]
66. Meyer D: *Support Vector Machines: The Interface to Libsvm in Package e1071*. Techische Universitat Wien, Austria 2004.
67. **Misc Functions of the Department of Statistics (e1071).** [http://cran.r-project.org/web/packages/e1071/index.html]
68. Pahikkala T, Tsivtsivadze E, Airola A, Järvinen J, Boberg J: **An efficient algorithm for learning to rank from preference graphs.** *Mach Learn* 2009, **75:**129–165.
69. Ambroise C, McLachlan GJ: **Selection bias in gene extraction on the basis of microarray gene-expression data.** *Proc Nat Acad Sci USA* 2002, **99**(10):6562–6566.
70. Varma S, Simon R: **Bias in error estimation when using cross-validation for model selection.** *BMC Bioinf* 2006, **7:**91.
71. Braga-Neto U, Hashimoto R, Dougherty ER, Nguyen DV, Carroll RJ: **Is cross-validation better than resubstitution for ranking genes?** *Bioinformatics* 2004, **20**(2):253–258.
72. Franceschini N, Reiner AP, Heiss G: **Recent findings in the genetics of blood pressure and hypertension traits.** *Am J Hypertens* 2010, **24**(4):392–400.
73. Laramie JM, Wilk JB, Williamson SL, Nagle MW, Latourelle JC, Tobin JE, Province MA, Borecki IB, Myers RH: **Multiple genes influence BMI on chromosome 7q31-34: the NHLBI Family Heart Study.** *Obesity* 2009, **17**(12):2182–2189.
74. Seda O, Liska F, Sedová L, Kazdová L, Krenová D, Kren V: **A 14-gene region of rat chromosome 8 in SHR-derived polydactylous congenic substrain affects muscle-specific insulin resistance, dyslipidaemia and visceral adiposity.** *Folia Biologica* 2005, **51**(3):53–61.
75. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29:**308–311.
76. Lee DY, Deng Z, Wang CH, Yang BB: **MicroRNA-378 promotes cell survival, tumor growth, and angiogenesis by targeting SuFu and Fus-1 expression.** *Proc Nat Acad Sci* 2007, **104**(51):20350–20355.
77. Humar R, Zimmerli L, Battegay E: **Angiogenesis and hypertension: an update.** *J Human Hypertens* 2009, **23**(12):773–82.
78. Corona E, Dudley JT, Butte AJ: **Extreme evolutionary disparities seen in positive selection across seven complex diseases.** *PLoS ONE* 2010, **5**(8):e12236.
79. Ferreira HS, de Castro e Silva E, Cointeiro C, Oliveira E, Faustino TN, Fregoneze JB: **Role of central 5-HT3 receptors in the control of blood pressure in stressed and non-stressed rats.** *Brain Res* 2004, **1028:**48–58.
80. Wilk JB, Chen Th, Gottlieb DJ, Walter RE, Nagle MW, Brandler BJ, Myers RH, Borecki IB, Silverman EK, Weiss ST, O'Connor GT: **A genome-wide association study of pulmonary function measures in the Framingham Heart Study.** *PLoS Genet* 2009, **5**(3):e1000429.
81. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker, P I, Daly MJ, Sham PC: **PLINK: a tool set for**

whole-genome association and population-based linkage analyses. *Am j human genet* 2007, **81**(3):559–575.

82. Rich SS, Goodarzi MO, Palmer ND, Langefeld CD, Ziegler J, Haffner SM, Bryer-Ash M, Norris JM, Taylor KD, Haritunians T, Rotter JI, Chen YDD, Wagenknecht LE, Bowden DW, Bergman RN: **A genome-wide association scan for acute insulin response to glucose in Hispanic-Americans: the Insulin Resistance Atherosclerosis Family Study (IRAS FS).** *Diabetologia* 2009, **52**(7):1326–1333.
83. Sun LD, Xiao FL, Li Y, Zhou WM, Tang HY, et al T: **Genome-wide association study identifies two new susceptibility loci for atopic dermatitis in the Chinese Han population.** *Nat Genet* 2011, **43**(7):690–694.
84. Michel S, Liang L, Depner M, Klopp N, Ruether A, Kumar A, Schedel M, Vogelberg C, von Mutius E, von Berg A, Bufe A, Rietschel E, Heinzmann A, Laub O, Simma B, Frischer T, Genuneit J, Gut I, Schreiber S, Lathrop M, Illig T, Kabesch M: **Unifying candidate gene and GWAS approaches in asthma.** *PLoS ONE* 2010, **5**(11):e13894.
85. Kang G, Childers D, Liu N, Zhang K, Gao G: **Genome-wide association studies of rheumatoid arthritis data via multiple hypothesis testing methods for correlated tests.** *BMC Proc* 2009, **3**(Suppl 7):S38.

doi:10.1186/1748-7188-7-11
Cite this article as: Pahikkala *et al.*: Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations. *Algorithms for Molecular Biology* 2012 **7**:11.

Publication III

Parallel Feature Selection for Regularized Least-Squares

Okser, S., Airola, A., Salakoski, T., Aittokallio, T., Pahikkala, T. (2013). In Manninen, P. and Öster, P., editors, *Applied Parallel and Scientific Computing*, volume 7782 of *Lecture Notes in Computer Science*, pages 280–294, Springer Berlin Heidelberg.

Publication IV

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

Okser, S., Pahikkala, T., Aittokallio, T. (2013). *BioData Mining*, 6(1):5.

Okser et al. BioData Mining 2013, 6:5
http://www.biodatamining.org/content/6/1/5

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

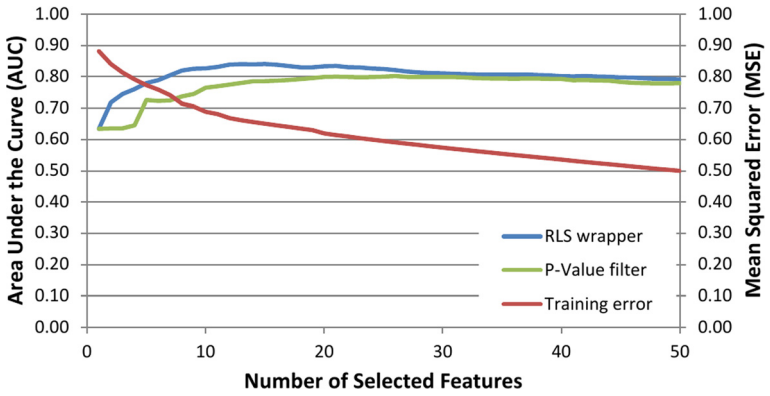

Figure 1 The figure illustrates how the external and internal cross-validation results behave as functions of the number of selected features. The external-cross validation consists of three training/test splits. The wrapper-based feature selection method, greedy RLS [23], is separately run during each 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

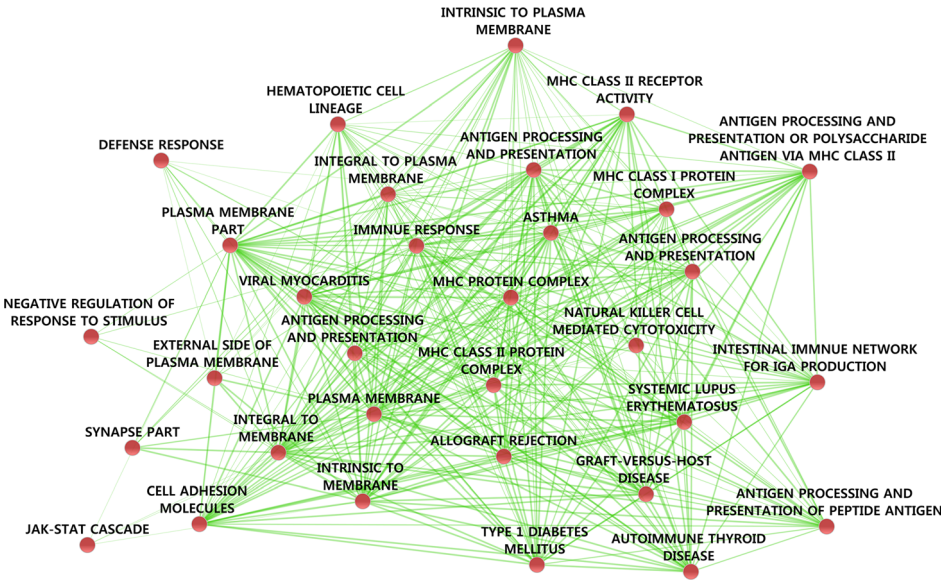

Figure 2 Sample network visualization constructed for type 1 diabetes. The risk variants were selected using the greedy RLS on the WTCCC type 1 diabetes GWAS data and the UK National Blood Services' controls, extended with those genes selected in another work [62]. The biological processes and pathways 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.

Publication V

Regularized Machine Learning in the Genetic Prediction of Complex Traits

Okser, S., Pahikkala, T., Airola, A., Salakoski, T., Ripatti, S., Aittokallio, T. (2014). *PLoS Genet*, 10(11):e1004754.

OPEN ACCESS Freely available online

PLOS | GENETICS

Review

Regularized Machine Learning in the Genetic Prediction of Complex Traits

Sebastian Okser[1,2], Tapio Pahikkala[1,2], Antti Airola[1,2], Tapio Salakoski[1,2], Samuli Ripatti[3,4,5], Tero Aittokallio[2,4]*

1 Department of Information Technology, University of Turku, Turku, Finland, **2** Turku Centre for Computer Science (TUCS), University of Turku and Åbo Akademi University, Turku, Finland, **3** Hjelt Institute, University of Helsinki, Helsinki, Finland, **4** Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland, **5** Wellcome Trust Sanger Institute, Hinxton, United Kingdom

Overview

Compared to univariate analysis of genome-wide association (GWA) studies, machine learning–based models have been shown to provide improved means of learning such multilocus panels of genetic variants and their interactions that are most predictive of complex phenotypic traits. Many applications of predictive modeling rely on effective variable selection, often implemented through model regularization, which penalizes the model complexity and enables predictions in individuals outside of the training dataset. However, the different regularization approaches may also lead to considerable differences, especially in the number of genetic variants needed for maximal predictive accuracy, as illustrated here in examples from both disease classification and quantitative trait prediction. We also highlight the potential pitfalls of the regularized machine learning models, related to issues such as model overfitting to the training data, which may lead to over-optimistic prediction results, as well as identifiability of the predictive variants, which is important in many medical applications. While genetic risk prediction for human diseases is used as a motivating use case, we argue that these models are also widely applicable in nonhuman applications, such as animal and plant breeding, where accurate genotype-to-phenotype modeling is needed. Finally, we discuss some key future advances, open questions and challenges in this developing field, when moving toward low-frequency variants and cross-phenotype interactions.

Introduction

Supervised machine learning aims at constructing a genotype–phenotype model by learning such genetic patterns from a labeled set of training examples that will also provide accurate phenotypic predictions in new cases with similar genetic background. Such predictive models are increasingly being applied to the mining of panels of genetic variants, environmental, or other nongenetic factors in the prediction of various complex traits and disease phenotypes [1–8]. These studies are providing increasing evidence in support of the idea that machine learning provides a complementary view into the analysis of high-dimensional genetic datasets as compared to standard statistical association testing approaches. In contrast to identifying variants explaining most of the phenotypic variation at the population level, supervised machine learning models aim to maximize the predictive (or generalization) power at the level of individuals, hence providing exciting opportunities for e.g., individualized risk prediction based on personal genetic profiles [9–11]. Machine learning models can also deal with genetic interactions, which are known to play an important role in the development and treatment of many complex diseases [12–16], but are often missed by single-locus association tests [17]. Even in the absence of significant single-loci marginal effects, multilocus panels from distinct molecular pathways may provide synergistic contribution to the prediction power, thereby revealing part of such *hidden heritability* component that has remained missing because of too small marginal effects to pass the stringent genome-wide significance filters [18]. Multivariate modeling approaches have already been shown to provide improved insights into genetic mechanisms and the interaction networks behind many complex traits, including atherosclerosis, coronary heart disease, and lipid levels, which would have gone undetected using the standard univariate modeling [2,19–22]. However, machine learning models also come with inherent pitfalls, such as increased computational complexity and the risk for model overfitting, which must be understood in order to avoid reporting unrealistic prediction models or over-optimistic prediction results.

We argue here that many medical applications of machine learning models in genetic disease risk prediction rely essentially on two factors: effective model regularization and rigorous model validation. We demonstrate the effects of these factors using representative examples from the literature as well as illustrative case examples. This review is not meant to be a comprehensive survey of all predictive modeling approaches, but we focus on *regularized machine learning models*, which enforces constraints on the complexity of the learned models so that they would ignore irrelevant patterns in the training examples. Simple risk allele counting or other multilocus risk models that do not incorporate any model parameters to be learned are outside the scope of this review; in fact, such simplistic models that assume independent variants may lead to suboptimal prediction performance in the presence of either direct or indirect interactions through epistasis effects or linkage disequilibrium, respectively [23,24]. Perhaps the simplest models considered here as learning approaches are those based on weighted risk allele summaries [23,25]. However, even with such basic risk models intended for predictive purposes, it is

Citation: Okser S, Pahikkala T, Airola A, Salakoski T, Ripatti S, et al. (2014) Regularized Machine Learning in the Genetic Prediction of Complex Traits. PLoS Genet 10(11): e1004754. doi:10.1371/journal.pgen.1004754

Editor: Nicholas J. Schork, University of California San Diego and The Scripps Research Institute, United States of America

Published November 13, 2014

Funding: Academy of Finland (grants 133227, 134020, 218310, 269862, 213506, 251217 and 129680), Turun Yliopistosäätiö, the Finnish Cultural Foundation, the Finnish foundation for Cardiovascular Research and the Sigrid Juselius Foundation. The funders had no role in the preparation of the article.

Competing Interests: The authors have declared that no competing interests exist.

* Email: tero.aittokallio@fimm.fi

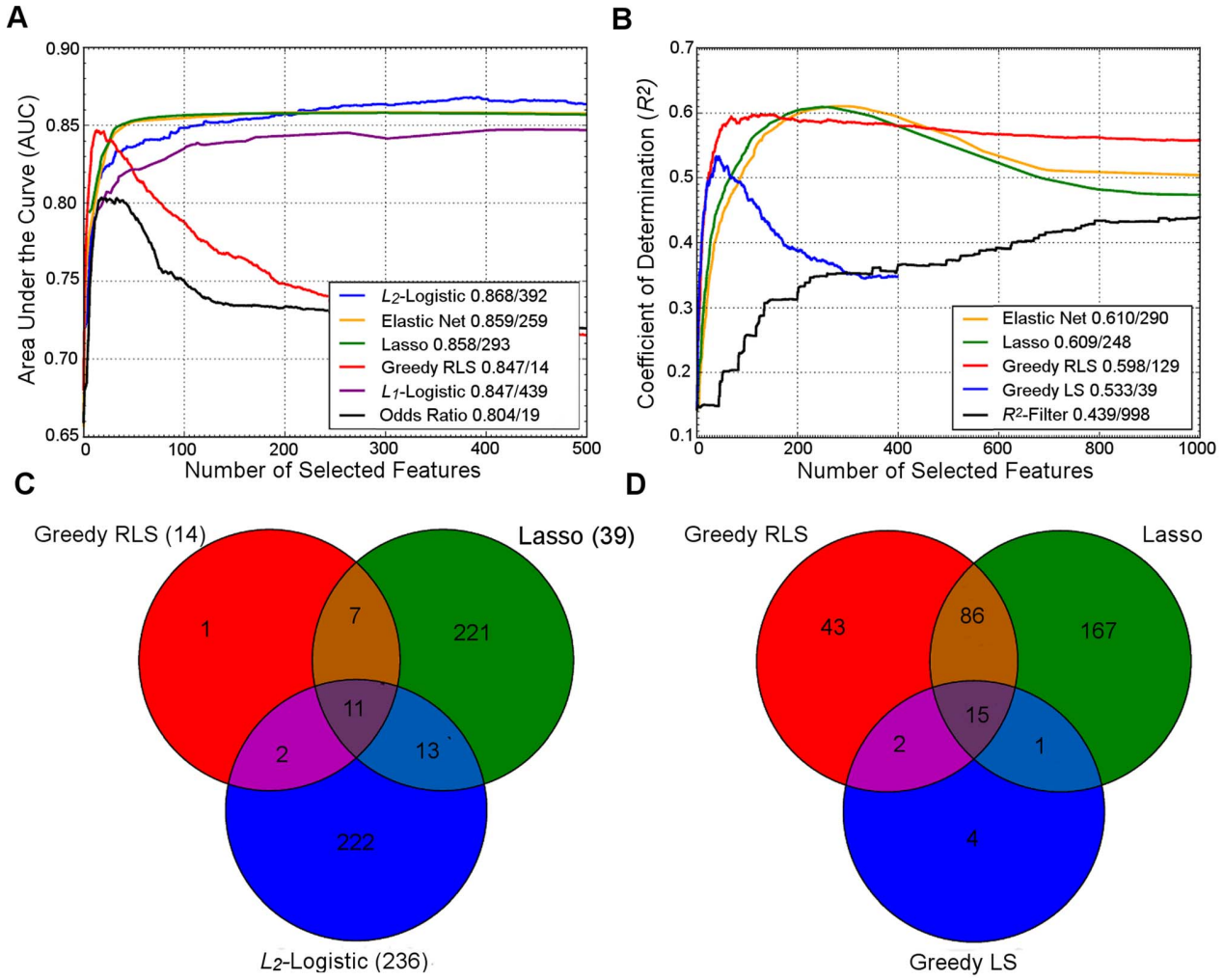

Figure 1. Performance of regularized machine learning models. Upper panel: Behavior of the learning approaches in terms of their predictive accuracy (*y*-axis) as a function of the number of selected variants (*x*-axis). Differences can be attributed to the genotypic and phenotypic heterogeneity as well as genotyping density and quality. (A) The area under the receiver operating characteristic curve (AUC) for the prediction of Type 1 diabetes (T1D) cases in SNP data from WTCCC [118], representing ca. one million genetic features and ca. 5,000 individuals in a case-control setup. (B) Coefficient of determination (R^2) for the prediction of a continuous trait (Tunicamycin) in SNP data from a cross between two yeast strains (Y2C) [44], representing ca. 12,000 variants and ca. 1,000 segregants in a controlled laboratory setup. The peak prediction accuracy/number of most predictive variants are listed in the legend. The model validation was implemented using nested 3-fold cross-validation (CV) [5]. Prior to any analysis being done, the data was split into three folds. On each outer round of CV, two of the folds were combined forming a training set, and the remaining one was used as an independent test set. On each round, all feature and parameter selection was done using a further internal 3-fold CV on the training set, and the predictive performance of the learned models was evaluated on the independent test set. The final performance estimates were calculated as the average over these three iterations of the experiment. In learning approaches where internal CV was not needed to select model parameters (e.g., log odds), this is equivalent to a standard 3-fold CV. T1D data: the L_2-regularized (ridge) regression was based on selecting the top 500 variants according to the χ^2 filter. For wrappers, we used our greedy L_2-regularized least squares (RLS) implementation [30], while the embedded methods, Lasso, Elastic Net and L_1-logistic regression, were implemented through the Scikit-Learn [119], interpolated across various regularization parameters up to the maximal number of variants (500 or 1,000). As a baseline model, we implemented a log odds-ratio weighted sum of the minor allele dosage in the 500 selected variants within each individual [25]. Y2C: the filter method was based on the top 1,000 variants selected according to R^2, followed by L_2-regularization within greedy RLS using nested CV. As a baseline model, we implemented a greedy version of least squares (LS), which is similar to the stepwise forward regression used in the original work [44]; the greedy LS differs from the greedy RLS in terms that it implements regularization through optimization of L_0 norm instead of L_2. It was noted that the greedy LS method drops around the point where the number of selected variants exceeds the number training examples (here, 400). Lower panel: Overlap in the genetic features selected by the different approaches. (C) The numbers of selected variants within the major histocompatibility complex (MHC) are shown in parentheses for the T1D data. (D) The overlap among then maximally predictive variants in the Y2C data. Note: these results should be considered merely as illustrative examples. Differing results may be obtained when other prediction models are implemented in other genetic datasets or other prediction applications.
doi:10.1371/journal.pgen.1004754.g001

important to learn the model parameters (e.g., select the variants and determine their weights) based on training data only; otherwise there is a severe risk of *model overfitting*, i.e., models not being capable of generalizing to new samples [5]. Representative examples of how model learning and regularization approaches address the overfitting problem are briefly summarized in Box 1, while those readers interested in their implementation details are referred to the accompanying Text S1. We

Box 1. Synthesis of Learning Models for Genetic Risk Prediction

The aim of risk models is to capture in a mathematical form the patterns in the genetic and non-genetic data most important for the prediction of disease susceptibility. The first step in model building involves choosing the functional form of the model (e.g., linear or nonlinear), and then making use of a given training data to determine the adjustable parameters of the model (e.g., a subset of variants, their weights, and other model parameters). While it is often sufficient for a statistical model to enable high enough explanatory power in the discovery material, without being overly complicated, a predictive model is also required to generalize to unseen cases.

One consideration in the model construction is how to encode the genotypic measurements using genotype models, such as the dominant, recessive, multiplicative, or additive model, each implying different assumptions about the genetic effects in the data [79]. Categorical variables 0, 1, and 2 are typically used for treating genetic predictor variables (e.g., minor allele dosage), while numeric values are required for continuous risk factors (e.g., blood pressure). Expected posterior probabilities of the genotypes can also be used, especially for imputed genotypes. Transforming the genotype categories into three binary features is an alternative way to deal with missing values without imputation (used in the T1D example; see Text S1 for details).

Statistical or machine learning models identify statistical or predictive interactions, respectively, rather than biological interactions between or within variants [12,80]. While nonlinear models may better capture complex genetic interactions [7,81], linear models are easier to interpret and provide a scalable option for performing supervised selection of multilocus variant panels at the genome-wide scale [3]. In linear models, genetic interactions are modeled implicitly by selecting such variant combinations that together are predictive of the phenotype, rather than considering pairwise gene–gene relationships explicitly. Formally, trait y_i to be predicted for an individual i is modeled as a linear combination of the individual's predictor variables x_{ij}:

$$y_i = w_0 + \sum_{j=1}^{p} w_j x_{ij} \quad i = 1,2,\ldots,n. \tag{1}$$

Here, the weights w_j are assumed constant across the n individuals, w_0 is the bias offset term and p indicates the number of predictors discovered in the training data. In its basic form, Eq. 1 can be used for modeling continuous traits y (linear regression). For case-control classification, the binary dependent variable y is often transformed using a logistic loss function, which models the probability of a case class given a genotype profile and other risk factor covariates x (logistic regression). It has been shown that the logistic regression and naïve Bayes risk models are mathematically very closely related in the context of genetic risk prediction [81].

Model regularization refers to the technique of controlling the model complexity, with the aim of preventing overfitting the model to the training data, and hence to improve its generalization capability to new samples. Classical regularization approaches rely on explicit penalization of the model complexity through penalty terms such as L_1 and L_2 norms for model weights (Figure 2A). Together with the squared loss function (Figure 2B), which is often used to measure the fit between the observed y_i and estimated $\hat{y}_i$ phenotypes (Eq.1), these functional norms give rise to the optimization problem used in various types of linear genetic risk prediction models:

$$\underset{\text{Squared loss}}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} + \underset{L_1 \text{ penalty}}{\lambda_1 \sum_{j=1}^{p}|w_j|} + \underset{L_2 \text{ penalty}}{\lambda_2 \sum_{j=1}^{p} w_j^2}. \tag{2}$$

Ridge regression is the special case of Eq. 2, in which $\lambda_1 = 0$, and the regularization parameter λ_2 is used to shrink the variable weights toward zero to prevent any particular variable from having too large effect on the model. However, the use of L_2 penalty alone tends to favor models that depend on all the variables. In Lasso, $\lambda_2 = 0$, and through adjusting the regularization parameter λ_1, it is possible to favor sparse models with only a few nonzero weights, leading to variable selection within the model fitting [82]. The Elastic Net model makes use of both penalty terms L_1 and L_2 to select also correlated features [83]; for instance, groups of variants within a pathway that together contribute to the predictive accuracy.

Methods such as Lasso and Elastic Net are traditionally known as *embedded models*, since the feature selection is embedded into the learning algorithm itself [5]. These methods select the features simultaneously and therefore do not provide the user with a direct control over the number of variables to be selected in the final prediction model, although heuristics based on absolute weights and other tuning criterion can be used for ranking the variables [24,84]. In contrast, *wrapper models* enable the user to preset the number of features in the final model. However, due to the exponentially increasing size of the genetic search spaces, in practice one must resort to local search methods, such as greedy feature selection implemented e.g., in L_2-RLS wrappers [30].

The wrapper and embedded methods are not distinct classes of algorithms. Scalable wrappers often incorporate elements of embedded methods to guarantee computational efficacy. For instance, RLS shares similar properties with Lasso and linear variants of SVMs. The accompanying Text S1 describes interrelationships between different learning models in terms of their norms and loss functions (Figure 2), including squared loss (RLS, Lasso and Elastic Net), logistic loss (logistic regression) and hinge loss (SVMs). It also presents a generic optimization framework that implements some of the most efficient methods currently available for genome-wide data. There are also other implementations available, including Mendel [85], HyperLasso [86] and SparSNP [87], gpu-lasso [88], and PUMA [89].

In addition to the classical regularization approaches, where an explicit model complexity penalization term is included in the optimization problem (Eq. 2), alternative strategies have been developed for avoiding overfitting. Among the most popular ones are *ensemble learning*, implemented e.g., in the popular Random Forests (RF) algorithm [90–92], as well as in the Bayesian modeling approaches, where probabilistic prior distributions on the model parameters are used for the shrinkage and regularization purposes [93–95]. Other approaches are based on the ensemble of models composed of varying number of features [96], bagging or boosting and various

search-based algorithms [3]. From the theoretical viewpoint, however, all of these learning approaches can be considered as different types of regularization approaches [97–100].

Whereas classical, univariate filter methods evaluate the relevance of each genetic feature independently of the others, more advanced *multivariate filters* have also been proposed, including the Relief family of approaches [101]. The main advantage of the multivariate filters over the univariate ones is that they can detect complex relationships between multiple genetic features and also yield smaller feature sets with less redundancy. Results from the ReliefF runs can also be aggregated, similar to ensemble learning, to yield more robust variant rankings and identification of gene–gene interactions [102]. However, multivariate filters also have specific limitations, such that their selection criteria are not directly connected to the generalization capability of the final prediction model, which may lead to suboptimal results [103].

Even advanced machine learning methodologies have been shown to be negatively affected by the presence of *population stratification*, leading to either false positives or false negative detections. To avoid the need to cluster the data into smaller substrata according to population structures, learning machines can be complemented by information of such substructures extracted using feature extraction methods, such as EIGENSTRAT, PCA, or MDS [104]. Lasso has been extended to account for population structures through linear mixed models [105], which are gaining much popularity in association studies [106]. Machine learning methods enable also the detection of population substructures, for instance, by learning ensembles of decision trees that are capable of accurately predicting individual's subcontinental ancestry [107].

Linkage disequilibrium (LD) tends to lead to the selection of highly correlated genetic features when using unpenalized modeling approaches [24]. A simple strategy is to select SNPs in linkage equilibrium, but this cannot distinguish the functionally relevant variants from the nonfunctional ones. Alternative approaches have revised, for instance, the tree-building process or importance measure calculation in RF [108], or replaced the univariate split functions by nonlinear multivariate split functions of contiguous SNPs, modeled as decision trees, to better account for SNP correlations [109]. Penalization strategies, such as ridge regression, Lasso and RLS, allow the model to avoid placing too much weight on potentially overfit variables in the presence of LD, which can lead to improved selection of causal variants [110,111].

Finally, *whole-genome prediction* (WGP) models fit all of the genotyped variants of the genetic data onto ridge regression type of linear models, such as genomic best linear unbiased prediction (GBLUP) or its variants [34,112]. WGP approach has been widely used in animal and plant breeding applications [113–115] and, with recent improvements, increasingly also in human genetics [116,117]. However, imperfect LD between markers and the causal loci can impose suboptimal prediction accuracy of WGP, especially when analyzing unrelated individuals, but this can be improved through variable selection or other model regularization approaches [61]. Moreover, due to the lack of direct control for the number of variants, WGP approaches are not optimal for those applications in which the size of the genotyped variant panel is limited.

specifically promote here the use of such regularized machine learning models that are scalable to the entire genome-wide scale, often based on linear models, which are easy to interpret and also enable straightforward variable selection. Genome-scale approaches avoid the need of relying on *two-stage approaches* [26], which apply standard statistical procedures to reduce the number of variants, since such prefiltering may miss predictive interactions across loci and therefore lead to reduced predictive performance [8,24,25,27,28].

Preview: Selection of Genetic Variants into the Predictive Models

A recent perspective article gave an excellent overview of the common concepts and potential pitfalls when making predictions of complex phenotypes using genotypic data [28]; however, one of the key components in the construction of predictive models—variant selection—was ignored in this and many other previous works. In the context of machine learning, a method known as *feature selection* is commonly implemented to identify the subset of variants having most predictive power for the particular phenotypic trait. The aims of feature selection include the reduction of the dimensionality of the genetic search space, excluding correlated variants without independent contribution to the prediction, and facilitating the implementation of the final prediction model, for instance, in clinical setup. Three main types of feature selection methods have traditionally been considered in the context of genetic predictors: filters, wrappers, and embedded methods (Box 1). These methods have different characteristics in terms of their computational complexities, potential to detect joint effects between variants, and whether the feature selection is done explicitly in the optimization process or implicitly through model regularization, which make them more or less suitable for different application cases [5–8].

A class of widely used filter approaches includes the standard multilocus genetic risk models, where the risk alleles and their weights are determined through single-locus statistical tests, such as odds–ratio, χ^2, or Fisher's exact text (so-called weighted risk scores). While such standard models have provided relatively good predictive accuracies, as assessed using simulation studies or hypothetical effect size distributions [29], we argue here that it makes sense to use machine learning both for selecting the subsets of the most predictive genetic features as well as training the final prediction model using regularized learning approaches [5,30]. The recent work of Chatterjee et al., where they estimated the effect size distributions for various quantitative and disease traits, highlighted the benefits gained from more holistic models that make use of the whole spectrum of genetic variation toward improving the predictive power of the genetic risk prediction models [31]. By design, the performance of any prediction model will depend on the sample size of the training set, as well as heritability of the disease trait, its underlying genetic architecture, and whether there is additional information available such as family history [29–33].

Representative Examples of Supervised Predictive Modeling Studies

Predictive modeling can be treated either as a classification problem (e.g., disease prediction in a case-control setting) or as a regression formulation (e.g., prediction of height in a general population cohort). Regardless of the problem formulation, however, the critical issue is how to guarantee that the model estimated in the training sample enables generalization power on

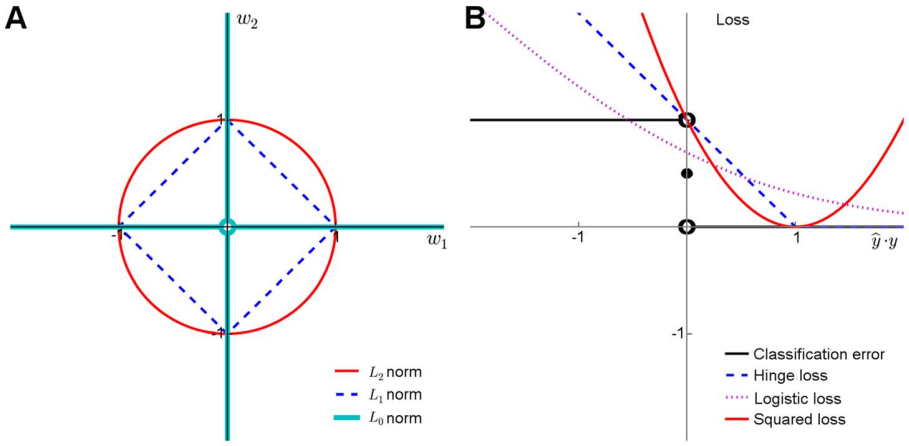

Figure 2. Penalty terms and loss functions. (A) Penalty terms: L_0-norm imposes the most explicit constraint on the model complexity as it effectively counts the number of nonzero entries in the model parameter vector. While it is possible to train prediction models with L_0-penalty using, e.g., greedy or other types of discrete optimization methods, the problem becomes mathematically challenging due to the nonconvexity of the constraint, especially when other than the squared loss function is used. The convexity of the L_1 and L_2 norms makes them easier for the optimization. While the L_2 norm has good regularization properties, it must be used together with either L_0 or L_1 norms to perform feature selection. (B) Loss functions: The plain classification error is difficult to minimize due to its nonconvex and discontinuous nature, and therefore one often resorts to its better behaving surrogates, including the hinge loss used with SVMs, the cross-entropy used with logistic regression, or the squared error used with regularized least-squares classification and regression. These surrogates in turn differ both in their quality of approximating the classification error and in terms of the optimization machinery they can be minimized with (Text S1).
doi:10.1371/journal.pgen.1004754.g002

new sets of individuals using appropriate learning models and regularization approaches. Another important issue is how to evaluate and quantify the predictive performance of these models using procedures such as cross validation (CV) and statistics such as the area under the curve (AUC) or coefficient of determination (R^2) (Text S1). These factors are next highlighted using representative examples from the recent literature [1–4,34,35], where various machine learning models have been implemented to gain insights and prediction capability beyond that obtained using standard statistical analyses of single nucleotide polymorphism (SNP) data.

In one of the first machine learning applications, Wei et al. showed that support vector machines (SVM) and L_2-regularized (ridge) logistic regression enabled construction of a highly predictive risk model for type 1 diabetes (T1D) using less than 500 variants that passed a relatively stringent prefiltering threshold ($p<10^{-5}$) on a case-control GWA dataset [1]. In contrast, relying merely on a collection of known T1D susceptibility loci led to poor performance in the predictive setting. More specifically, when the predictive accuracy was evaluated in terms of within-study 5-fold CV, they obtained extremely good prediction power (AUC close to 0.9). However, it is known that simple CV may lead to over-optimistic results due to information leakage between the two stages of the feature selection process [5]. Indeed, when the predictive models were evaluated using totally independent validation cohort, the between-study performance dropped drastically (AUC 0.84 for SVM) [1], highlighting the importance of independent samples in the model validation.

Recently, Wei et al. made use of larger sample sizes (>10,000 individuals), using variant data from 15 European countries for risk prediction of Crohn's disease (CD) and ulcerative colitis (UC) [4]. They applied a custom Immunochip that provides a more comprehensive catalog of both common variants and certain rare variants missed in the first generation of GWA studies. Using a relatively liberal threshold ($p<10^{-4}$), they preselected around 10,000 variants and applied regularized logistic regression with L_1 penalty for sparse genetic risk modeling. In an independent validation set from the meta-analysis cohort, the predictive models achieved the best prediction performance reported for CD and UC (AUCs of 0.86 and 0.83, respectively) so far. In contrast, the simple odds–ratio-weighted genetic risk model showed relatively poor results (AUC of 0.730 and 0.685, respectively). The study also confirmed the projections from previous works [31–33], suggesting that predictive accuracy is highly dependent on the sample sizes and the spectrum of variants included in the model, in addition to the heritability of the disease trait.

The final example comes from the regression formulation. With the aim to explain a part of the missing heritability of height, Yang et al. [34] went beyond the two-stage approach and fit a simple linear regression model to all directly genotyped 294,831 variants that passed their quality control. Using such a whole genome prediction (WGP) approach, without any variant selection, the authors were able to explain 45% of the phenotypic variation in height in a cohort of approximately 4,000 European descents. Similarly high R^2 values were also confirmed in another study [35] where the WGP approach was trained in an European cohort; however, R^2 values dropped dramatically when the fitted model was applied to an independent validation dataset using 10-fold CV (R^2 ranging around 0.2, depending on the number of variants and whether familial information was used) [35]. These studies highlight the risk of overfitting to the training sample when no feature selection or model regularization is used in the model construction.

Prediction Performance Using Examples of Model Regularization

To illustrate the similarities and differences in their behavior, we ran a number of common regularization approaches on two example datasets (Figure 1). In both datasets, the two embedded methods, Lasso and Elastic Net, showed strikingly similar prediction behavior, but needed a larger number of variants for their peak performance, compared to the greedy regularized least-squares (RLS) wrapper, which peaked much earlier but resulted in

lower prediction accuracy. As was expected, the top performance of the L_2-regularized logistic (ridge) regression required a very large number of features, while showing reduced accuracy at a lower number of variants. Surprisingly, the popular L_1-penalized logistic model showed slightly suboptimal performance; although its peak performance was similar to that of greedy RLS, it required a much larger number of variants in these datasets. We note that the relative behavior of these methods may well change in other genetic datasets and applications. In line with the previous results in CD and UC cases [4], the simple log odds-weighted risk model also showed poor results in the T1D case. While for some other traits such accuracies would be considered excellent, the high heritability and dependence on the human leukocyte antigen (HLA) region often leads to higher predictive performance for T1D [1]. However, these accuracies are better than expected for a sample of this size if the standard, nonmachine learning, multilocus genetic models were utilized in the risk prediction [28].

The relatively small overlap in the selected features highlights an interesting point that the models tend to select different panels of variants while achieving rather similar prediction performance (Figure 1C, D), suggesting that the selected variants may provide complementary views of the genetic mechanisms behind the phenotypes. In the T1D case, for instance, most of the variants selected by the L_2-logistic and greedy RLS were from the major histocompatibility complex (MHC) region (95% and 67%, respectively), in line with the previous studies [1,4], whereas Lasso also selected novel variants mostly outside the MHC region (15%), which may provide complementary information for the risk assessment. This difference is likely due to its embedded nature; Lasso selects variants simultaneously, rather than one at a time, which often requires further optimization in applications where the size of the variant panel is limited. As expected, the univariate filters tend to select larger numbers of correlated features, since they cannot consider interactions with already selected variants. At the other extreme, greedy RLS selects relatively uncorrelated variants while the embedded methods lie in between. These example cases suggest that there is no golden rule for feature selection, but that the model should be selected based on the characteristics of the data and goals of the genetic application (e.g., whether small number of variants is preferred over the overall predictive accuracy).

Perspective: Current Challenges and Emerging Developments

While rare variants have been proposed as one explanation for the missing heritability [36,37], there has been a divergence of opinion over whether rare variants of large effect or common variants of small effect are contributing most to the phenotypic variability [38]. It has been suggested that incorporating low-frequency or rare variants will make the disease risk prediction increasingly more accurate [4,28,29,31]. However, recent reports have shown only incremental impact of rare variants on disease susceptibility and prediction of complex diseases, as evaluated at the population level using either simulated data [39] or by sequencing of known risk variants for autoimmune disease traits [40]. We believe that a more systematic investigation of the variants across portions of the allelic spectrum will likely contribute to explaining more of the missing heritability. While the presented machine learning algorithms easily scale to a GWA level, the emerging sequencing data, either from genotype imputation or whole-exome and genome profiling, are posing new technical challenges, where parallelization and cloud technologies for distributed memory and high-performance computing will become increasingly important. Placing the focus on individual-level predictions should help also with the low-frequency variants shared only by a small portion of the individuals. For instance, selection of the most robust variants was shown to improve various prediction models, especially when the variants are poorly tagged or have low minor allele frequency (MAF) [41]. Since most rare variants are highly population-specific, it may be necessary to borrow prior biological information from shared regulatory regions, genes, or pathways, similar to the recent collapsing methods for rare association analyses [42]. However, improved model regularization options that allow more flexibility and sparsity in the selected panels of variants across various subgroups of individuals will likely be needed to deal with the rare variants and to account for population stratification. Regularization methods based on sparse group Lasso, for instance, can be extended to rare variants and pathway-driven variant selection [22,43].

It has been argued that, even with increasingly large-scale and dense genomic data, genetic prediction alone may still not reach the accuracy regarded as clinically informative for the population at large [18]. High-quality and controlled genetic data from model organisms will likely give the first estimates on how much sequencing data can really add to the predictive accuracy of complex phenotypes [44,45]. Lessons from model organisms have already shown that additional information originating from environmental and stochastic factors, as well as from phenotypic robustness and transgenerational effects, will be necessary for accurate predictions at an individual level [46–48]. In particular, gene expression should prove especially useful, since such intermediate phenotype captures both genetic and nongenetic contributions to phenotypic variation [49]. For instance, epigenetic gene expression variability of genetic interaction partners plays an important role in explaining complex regulatory relationships, characterized using concepts such as "epigenetic epistasis" [50] or "eQTL epistasis" [51]. Although modeling of gene expression variability poses some technical challenges, similar to those already encountered when modeling GWA datasets [52,53], incorporating such continuous features into the disease prediction models should be relatively straightforward. Adding the nongenetic information will likely be instrumental when going toward less heritable diseases, such as some cancer subtypes, which traditionally have been challenging to predict using standard GWA approaches [29,32,33,54–56]. Finally, including family medical history and other clinical data from electronic health records should improve the personal risk assessment models, as well as provide guidance on lifestyle changes for those currently healthy individuals that have increased genetic risk for the disease susceptibility [57,58].

An interesting question under debate is how many genetic features should be incorporated into the prediction models [3,28,31,59,60]. Although the WGP methods have been successfully applied in animal and plant breeding applications [61], these are not suitable for applications in which the number of genetic markers is constrained. In embedded models, the number of features to be selected is often dependent on the regularization parameter. However, in the current Lasso and Elastic Net implementations, the user cannot explicitly specify the number of variants to be included in the final model, but the selection of final predictors often requires further grid searches or other tuning options. Such lack of direct control over the size of the variant panel may be an important practical consideration in medical applications, where the size of the variant panel is often associated with an additional cost, for instance, in disease screening applications, or when the goal is to select a few of the variants for follow-up experimentation, for instance, using functional assays. Greedy feature selection offers full control to the user and often leads to smaller panels of predictive, uncorrelated

variants, which may be beneficial when the size of clinical assay is limited. However, the trade-off is a slight drop in the overall predictive accuracy (Figure 1), indicating that more in-depth and effective wrapper selection strategies need to be implemented. There are also other strategies to reduce the dimensionality of genetic feature spaces using data transformations, such as principal components analysis (PCA), multidimensional scaling (MDS), partial least squares (PLS), or discrete wavelet transformation (DWT), which may in some cases lead to improved predictive accuracy [62]. However, rather than selecting combinations of transformed features, feature selection on the original variant space offers directly actionable modeling outcomes, such as a selected set of predictive genetic loci for follow-up applications and experimentation.

We envision a number of future directions for improvements in disease risk prediction. One exciting development involves modeling of cross-phenotype interactions (pleiotropy). Many genetic variants are associated with multiple disease phenotypes, particularly across autoimmune diseases, cancers, and neuropsychiatric disorders [63]. Statistical approaches have been suggested for making use of the complementary information from multiple phenotypes to gain power to detect small effects that would have been missed if tested individually [64–65]. Bayesian learning approaches seem particularly fitting for multivariate modeling of pleiotropic associations, especially for the lower-frequency variants where shared genetic features across individuals for any single phenotype become increasingly rare [66–71]. We expect that regularized machine learning models will also prove useful when translating the subtle multivariant–multiphenotype relationships into genetic risk prediction models. Modeling studies in yeast have already shown that multiple phenotypic measurements enable mapping of genetic interaction networks with distinct biological processes across pathways [72]. Networks of genetic and/or physical interactions may therefore serve as useful prior information for the prediction models to move from variant-level features towards pathway-level features [5,73–75]. Using such functional relationships to assemble or collapse higher-level predictive features might better account for the interindividual genetic variation at the lower end of variant frequency. For instance, predictive subnetwork modules could enable more robust personalized medicine strategies by allowing that individuals with the same disease phenotype may show interindividual genetic heterogeneity in the sense that their disease predisposing variants may lie in distinct loci within the shared pathways. Such advances will rely on the next generation of machine learning models that can effectively deal with the complexity arising from massive number of interactions between rare and common genetic and nongenetic factors [76–78].

Conclusions

The current evidence contradicts the idea of a universally optimal model across datasets and prediction applications; rather, the model should be selected based on whether one is trying to achieve a maximally predictive model without restricting the number or type of variants, or whether the goal is to build a sufficiently predictive model with a limited number of genetic and nongenetic features. This highlights the importance of feature selection as a key component in the construction of prediction models, whether it is done explicitly in the optimization process (e.g., wrappers) or implicitly through the model regularization (embedded models). One common finding is that those variants not meeting the stringent genome-wide significance levels may also contribute to the predictive signals when combined in the multilocus prediction modes [2,4,24,25,27,28, 31,33]. Another consensus point is that regularized models often outperform their unregularized counterparts [24], which was also supported by our example results (Figure 1).

Regardless of the model used, however, careful evaluation of its generalizability is critical for prediction applications. We encourage using systematic and unbiased procedures, such as nested CV, for the selection of genetic variables and other model parameters and for the evaluation of the generalization performance of the model. The final model construction and feature selection should be performed on the complete set of samples using standard CV options. However, the eventual predictive power must be assessed by implementing the final model on a sufficiently large, representative, and independent test set in order to avoid reporting over-optimistic prediction results. The model evaluation also depends on the application case; for instance, if the aim is to carry out disease screening in Finland, then a relatively large Finnish population sample should be used both in the model construction and validation.

Genetic risk prediction through supervised machine learning models goes beyond the single-locus association testing with the complex disease phenotypes. The main objective of regularized learning approaches is to find the most predictive combinations of variants, the functional roles of which must to be validated using follow-up experimentation. However, it is likely that predictive power is linked to the underlying biological mechanisms and even causality, but whether this comes through the selected variants and their interactions, or via synthetic associations or other nondirect relationships needs to be evaluated mechanistically. Genotype–phenotype modeling is a highly challenging problem, but we believe that through appropriate implementation and application of the supervised machine learning methods, such as those presented here, increasingly predictive relationships and biological understanding will be extracted from the current and emerging genetic and phenotypic datasets.

Supporting Information

Text S1 Implementation details for a range of regularized machine learning models.
(PDF)

Acknowledgments

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 data is available from www.wtccc.org.uk. The authors also thank CSC, the Finnish IT center for science, for providing us with extensive computational resources for the experiments.

References

1. Wei Z, Wang K, Qu HQ, Zhang H, Bradfield J, et al. (2009) From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet 5: e1000678.
2. Okser S, Lehtimäki T, Elo LL, Mononen N, Peltonen N, et al. (2009) Genetic variants and their interactions in the prediction of increased pre-clinical carotid atherosclerosis: the cardiovascular risk in young Finns study. PLoS Genet 6: e1001146.
3. Kruppa J, Ziegler A, König IR (2012) Risk estimation and risk prediction using machine-learning methods. Hum Genet 131: 1639–1654.
4. Wei Z1, Wang W, Bradfield J, Li J, Cardinale C, et al. (2013) Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. Am J Hum Genetics 92: 1008–1012.
5. Okser S, Pahikkala T, Aittokallio T (2013) Genetic variants and their interactions in disease risk prediction - machine learning and network perspectives. BioData Min 6: 5. doi:10.1186/1756-0381-6-5
6. Szymczak S, Biernacka JM, Cordell HJ, González-Recio O, König IR, et al. (2009) Machine learning in genome-wide association studies. Genet Epidemiol 33 Suppl 1: S51–S57.

7. Moore JH, Asselbergs FW, Williams SM (2010) Bioinformatics challenges for genome-wide association studies. Bioinformatics 26: 445–455.
8. Kooperberg C, LeBlanc M, Obenchain V (2010) Risk prediction using genome-wide association studies. Genet Epidemiol 34: 643–652.
9. Kraft P, Wacholder S, Cornelis MC, Hu FB, Hayes RB, et al. (2009) Beyond odds ratios: communicating disease risk based on genetic profiles. Nat Rev Genet 10: 264–269.
10. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, et al. (2010) Clinical assessment incorporating a personal genome. Lancet 375: 1525–1535.
11. Manolio TA (2013) Bringing genome-wide association findings into clinical use. Nat Rev Genet 14: 549–558.
12. Lehner B (2011). Molecular mechanisms of epistasis within and between genes. Trends Genet 27: 323–331.
13. Lehner B (2007) Modelling genotype-phenotype relationships and human disease with genetic interaction networks. J Exp Biol 210: 1559–1566.
14. Moore JH, Williams SM (2009) Epistasis and its implications for personal genetics. Am J Hum Genet 85: 309–320.
15. Ashworth A, Lord CJ, Reis-Filho JS (2011) Genetic interactions in cancer progression and treatment. Cell 145: 30–38.
16. Brough R, Frankum JR, Costa-Cabral S, Lord CJ, Ashworth A (2011) Searching for synthetic lethality in cancer. Curr Opin Genet Dev 21: 34–41.
17. Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10: 392–404.
18. Gibson G (2010) Hints of hidden heritability in GWAS. Nat Genet 42: 558–60.
19. Inouye M, Ripatti S, Kettunen J, Lyytikäinen LP, Oksala N, et al. (2012) Novel Loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. PLoS Genet 8: e1002907.
20. Ripatti S, Tikkanen E, Orho-Melander M, Havulinna AS, Silander K, et al. (2012) A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. Lancet 376: 1393–400.
21. Wineinger NE, Harper A, Libiger O, Srinivasan SR, Chen W, et al. (2013) Front Genet 4: 86.
22. Silver M, Chen P, Li R, Cheng CY, Wong TY, et al. (2013) Pathways-driven sparse regression identifies pathways and genes associated with high-density lipoprotein cholesterol in two Asian cohorts. PLoS Genet 9: e1003939.
23. Che R, Motsinger-Reif AA (2013) Evaluation of genetic risk score models in the presence of interaction and linkage disequilibrium. Front Genet 4: 138.
24. Abraham G, Kowalczyk A, Zobel J, Inouye M (2013) Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. Genet Epidemiol 37: 184–195.
25. Evans DM, Visscher PM, Wray NR (2009) Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum Mol Genet 18: 3525–3531.
26. Shi G, Boerwinkle E, Morrison AC, Gu CC, Chakravarti A, et al. (2011) Mining gold dust under the genome wide significance level: a two-stage approach to analysis of GWAS. Genetic Epidemiol 35: 111–118.
27. Jakobsdottir J, Gorin MB, Conley YP, Ferrell RE, Weeks DE (2009) Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. PLoS Genet 5: e1000337.
28. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, et al. (2013) Pitfalls of predicting complex traits from SNPs. Nat Rev Genet 14: 507–515.
29. Jostins L, Barrett JC (2011) Genetic risk prediction in complex disease. Hum Mol Genet 20: R182–188.
30. Pahikkala T, Okser S, Airola A, Salakoski T, Aittokallio T (2012) Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations. Algorithms Mol Biol 7: 11. doi:10.1186/1748-7188-7-11
31. Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock S, et al. (2013) Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat Genet 45: 400–405.
32. Dudbridge F (2013) Power and Predictive Accuracy of Polygenic Risk Scores. PLoS Genet 9: e1003348.
33. Do CB, Hinds DA, Francke U, Eriksson N (2012) Comparison of family history and SNPs for predicting risk of complex disease. PLoS Genet 8: e1002973.
34. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42: 565–569.
35. Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, et al. (2011) Beyond missing heritability: prediction of complex traits. PLoS Genet 7: e1002051.
36. Maher B (2008) Personal genomes: The case of the missing heritability. Nature 456: 18–21.
37. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, et al. (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11: 446–450.
38. Gibson G (2012) Rare and common variants: twenty arguments. Nat Rev Genet 13: 135–145.
39. Mihaescu R, Pencina MJ, Alonso A, Lunetta KL, Heckbert SR, et al. (2013) Incremental value of rare genetic variants for the prediction of multifactorial diseases. Genome Med 20: 76.
40. Hunt KA, Mistry V, Bockett NA, Ahmad T, Ban M, et al. (2013) Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature 498: 232–235.
41. Manor O, Segal E (2013) Predicting disease risk using bootstrap ranking and classification algorithms. PLoS Comput Biol 9: e1003200.
42. Moore CB, Wallace JR, Wolfe DJ, Frase AT, Pendergrass SA, et al. (2013) Low frequency variants, collapsed based on biological knowledge, uncover complexity of population stratification in 1000 genomes project data. PLoS Genet 9: e1003959.
43. Zhou H, Sehl ME, Sinsheimer JS, Lange K (2010) Association screening of common and rare genetic variants by penalized regression. Bioinformatics 26: 2375–2382.
44. Bloom JS1, Ehrenreich IM, Loo WT, Lite TL, Kruglyak L (2013) Finding the sources of missing heritability in a yeast cross. Nature 494: 234–237.
45. Rat Genome Sequencing and Mapping Consortium (2013) Combined sequence-based and genetic mapping analysis of complex traits in outbred rats. Nat Genet 45: 767–775.
46. Burga A, Lehner B (2012) Beyond genotype to phenotype: why the phenotype of an individual cannot always be predicted from their genome sequence and the environment that they experience. FEBS J 279: 3765–3775.
47. Lehner B (2013) Genotype to phenotype: lessons from model organisms for human genetics. Nat Rev Genet 14: 168–178.
48. Queitsch C, Carlson KD, Girirajan S (2012) Lessons from model organisms: phenotypic robustness and missing heritability in complex disease. PLoS Genet 8: e1003041.
49. Burga A, Lehner B (2013) Predicting phenotypic variation from genotypes, phenotypes and a combination of the two. Curr Opin Biotechnol 24: 803–809.
50. Park S, Lehner B (2013) Epigenetic epistatic interactions constrain the evolution of gene expression. Mol Syst Biol 9: 645.
51. Huang Y, Wuchty S, Przytycka TM (2013) eQTL epistasis - challenges and computational approaches. Front Genet 4: 51.
52. Manor O, Segal E (2013) Robust prediction of expression differences among human individuals using only genotype information. PLoS Genet 9: e1003396.
53. Goldinger A, Henders AK, McRae AF, Martin NG, Gibson G, et al. (2013) Genetic and Non-Genetic Variation Revealed for the Principal Components of Human Gene Expression. Genetics 195: 1117–1128.
54. Galvan A, Ioannidis JP, Dragani TA (2010) Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. Trends Genet 26: 132–141.
55. Machiela MJ, Chen CY, Chen C, Chanock SJ, Hunter DJ, et al. (2011) Evaluation of polygenic risk scores for predicting breast and prostate cancer risk. Genet Epidemiol 35: 506–514.
56. Urbach D, Lupien M, Karagas MR, Moore JH (2012) Cancer heterogeneity: origins and implications for genetic association studies. Trends Genet 28: 538–543.
57. Gibson G, Visscher PM (2013) From personalized to public health genomics. Genome Med 5: 60.
58. Bromberg Y (2013) Building a genome analysis pipeline to predict disease risk and prevent disease. J Mol Biol 425: 3993–4005.
59. Wu J, Pfeiffer RM, Gail MH (2013) Strategies for developing prediction models from genome-wide association studies. Genet Epidemiol 37: 768–777.
60. Warren H, Casas JP, Hingorani A, Dudbridge F, Whittaker J (2014) Genetic prediction of quantitative lipid traits: comparing shrinkage models to gene scores. Genet Epidemiol 38: 72–83.
61. de Los Campos G, Vazquez AI, Fernando R, Klimentidis YC, Sorensen D (2013) Prediction of complex human traits using the genomic best linear unbiased predictor. PLoS Genet 9: e1003608.
62. Hennings-Yeomans PH, Cooper GF (2012) Improving the prediction of clinical outcomes from genomic data using multiresolution analysis. IEEE/ACM Trans Comput Biol Bioinform 9: 1442–1450.
63. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW (2013) Pleiotropy in complex traits: challenges and strategies. Nat Rev Genet 14: 483–495.
64. Silver M, Janousova E, Hua X, Thompson PM, Montana G, et al. (2012) Identification of gene pathways implicated in Alzheimer's disease using longitudinal imaging phenotypes with sparse regression. Neuroimage 63:1681–1694.
65. Schifano ED, Li L, Christiani DC, Lin X (2013) Genome-wide association analysis for multiple continuous secondary phenotypes. Am J Hum Genet 92: 744–759.
66. Marttinen P, Gillberg J, Havulinna A, Corander J, Kaski S (2013) Genome-wide association studies with high-dimensional phenotypes. Stat Appl Genet Mol Biol 12: 413–431.
67. Mutshinda CM, Noykova N, Sillanpää MJ (2012) A hierarchical Bayesian approach to multi-trait clinical quantitative trait locus modeling. Front Genet 3: 97.
68. Hartley SW, Monti S, Liu CT, Steinberg MH, Sebastiani P (2012) Bayesian methods for multivariate modeling of pleiotropic SNP associations and genetic risk prediction. Front Genet 3: 176.
69. Hartley SW, Sebastiani P (2013) PleioGRiP: genetic risk prediction with pleiotropy. Bioinformatics 29: 1086–1088.
70. Bottolo L, Chadeau-Hyam M, Hastie DI, Zeller T, Liquet B, et al. (2013) GUESS-ing polygenic associations with multiple phenotypes using a GPU-based evolutionary stochastic search algorithm. PLoS Genet 9: e1003657.
71. Marttinen P, Pirinen M, Sarin AP, Gillberg J, Kettunen J, et al. (2014) Assessing multivariate gene-metabolome associations with rare variants using Bayesian reduced rank regression. Bioinformatics 30: 2026–2034.

72. Carter GW, Hays M, Sherman A, Galitski T (2012) Use of pleiotropy to model genetic interactions in a population. PLoS Genet 8: e1003010.
73. Kim YA, Przytycka TM (2012) Bridging the gap between genotype and phenotype via network approaches. Front Genet 3: 227.
74. Bebek G, Koyutürk M, Price ND, Chance MR (2012) Network biology methods integrating biological data for translational science. Brief Bioinform 13: 446–459.
75. Mitra K, Carvunis AR, Ramesh SK, Ideker T (2013) Integrative approaches for finding modular structure in biological networks. Nat Rev Genet 14: 719–732.
76. Upstill-Goddard R, Eccles D, Fliege J, Collins A (2013) Machine learning approaches for the discovery of gene-gene interactions in disease data. Brief Bioinform 14: 251–260.
77. Lu C, Latourelle J, O'Connor GT, Dupuis J, Kolaczyk ED (2013) Network-guided sparse regression modeling for detection of gene-by-gene interactions. Bioinformatics 29: 1241–1249.
78. Su C, Andrew A, Karagas MR, Borsuk ME (2013) Using Bayesian networks to discover relations between genes, environment, and disease. BioData Min 6: 6.
79. Bush WS, Moore JH (2012) Chapter 11: Genome-wide association studies. PLoS Comput Biol 8: e1002822
80. Sun X, Lu Q, Mukheerjee S, Crane PK, Elston R, et al. (2014) Analysis pipeline for the epistasis search - statistical versus biological filtering. Front Genet 5: 106.
81. Sebastiani P, Solovie N, Sun J (2012) Naive Bayesian classifier and genetic risk score for genetic risk prediction of a categorical trait: not so different after all! Front Genet 3: 26.
82. Tibshirani R (1994) Regression shrinkage and selection via the Lasso. J Royal Stat Soc B 58: 267–288.
83. Zou H, Hastie T (2003) Regularization and variable selection via the elastic net. J Royal Stat Soc B 67: 301–320.
84. Waldmann P, Mészáros G, Gredler B, Fuerst C, Sölkner J (2013) Evaluation of the lasso and the elastic net in genome-wide association studies. Front Genet 4: 270.
85. Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009) Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25: 714–721.
86. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS Genet 4: e1000130.
87. Abraham G, Kowalczyk A, Zobel J, Inouye M (2012) SparSNP: fast and memory-efficient analysis of all SNPs for phenotype prediction. BMC Bioinformatics 13: 88.
88. Chen GK (2012) A scalable and portable framework for massively parallel variable selection in genetic association studies. Bioinformatics 28: 719–720.
89. Hoffman GE, Logsdon BA, Mezey JG (2013) PUMA: a unified framework for penalized multiple regression analysis of GWAS data. PLoS Comput Biol 9: e1003101.
90. Breiman L (2001) Random Forests. Machine Learning 45: 5–32.
91. Goldstein BA, Hubbard AE, Cutler A, Barcellos LF (2010) An application of Random Forests to a genome-wide association dataset: methodological considerations and new findings. BMC Genet 11: 49.
92. Boulesteix AL, Bender A, Lorenzo Bermejo J, Strobl C (2012) Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations. Brief Bioinform 13: 292–304.
93. Li J, Das K, Fu G, Li R, Wu R (2011) The Bayesian lasso for genome-wide association studies. Bioinformatics 27: 516–523.
94. Peltola T, Marttinen P, Jula A, Salomaa V, Perola M, et al. (2012) Bayesian variable selection in searching for additive and dominant effects in genome-wide data. PLoS ONE 7: e29115.
95. Zhou X, Carbonetto P, Stephens M (2013) Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet 9: e1003264.
96. Milton JN, Gordeuk VR, Taylor JG, Gladwin MT, Steinberg MH, et al. (2014) Prediction of fetal hemoglobin in sickle cell anemia using an ensemble of genetic risk prediction models. Circ Cardiovasc Genet 7: 110–115.
97. Brown G,Wyatt JL, Tino P (2005) Managing diversity in regression Ensembles. J Mach Learn Res 6: 1621–1650.
98. Poggio T, Rifkin R, Mukherjee S, Rakhlin A (2002) Bagging regularizes. CBCL Memo 214. MIT AI lab. Available: http://cbcl.mit.edu/publications/ai-publications/2002/AIM-2002-003.pdf. Accessed 24 June 2014.
99. Gerfo LL, Rosasco L, Odone F, De Vito E, Verri A (2008) Spectral algorithms for supervised learning. Neural Comput 20: 1873–1897.
100. Mitchell TJ, Beauchamp JJ (1998) Bayesian variable selection in linear regression. J Am Stat Assoc 83: 1023–1036.
101. Robnik-Sikonja M, Kononenko I (2003) Theoretical and empirical analysis of ReliefF and RReliefF. Machine Learning 53: 23–69.
102. Yang P, Ho JW, Yang YH, Zhou BB (2011) Gene-gene interaction filtering with ensemble of filters. BMC Bioinformatics 12 Suppl 1: S10.
103. McKinney BA, Crowe JE, Guo J, Tian D (2009) Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. PLoS Genet 5: e1000432.
104. Zhao Y, Chen F, Zhai R, Lin X, Wang Z, et al. (2012) Correction for population stratification in random forest analysis. Int J Epidemiol 41: 1798–1806.
105. Rakitsch B, Lippert C, Stegle O, Borgwardt K (2013) A Lasso multi-marker mixed model for association mapping with population structure correction. Bioinformatics 29: 206–214.
106. Yang J, Zaitlen NA, Goddard ME, Visscher PM, Price AL (2014) Advantages and pitfalls in the application of mixed-model association methods. Nat Genet 46: 100–106.
107. Hajiloo M, Sapkota Y, Mackey JR, Robson P, Greiner R, et al. (2013) ETHNOPRED: a novel machine learning method for accurate continental and sub-continental ancestry identification and population stratification correction. BMC Bioinformatics 14: 61.
108. Meng YA, Yu Y, Cupples LA, Farrer LA, Lunetta KL (2009) Performance of random forest when SNPs are in linkage disequilibrium. BMC Bioinformatics, 10: 78.
109. Botta V, Louppe G, Geurts P, Wehenkel L (2014) Exploiting SNP correlations within random forest for genome-wide association studies. PloS ONE 9: e93379.
110. Malo N, Libiger O, Schork NJ (2008) Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. Am J Hum Genet 82: 375–385.
111. He Q, Lin DY (2011) A variable selection method for genome-wide association studies. Bioinformatics 27: 1–8.
112. Ober U, Erbe M, Long N, Porcu E, Schlather M, et al. (2011) Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. Genetics 188: 695–708.
113. Wimmer V, Albrecht T, Auinger HJ, Schön CC (2012) Synbreed: a framework for the analysis of genomic prediction data using R. Bioinformatics 28: 2086–2087.
114. Ober U, Ayroles JF, Stone EA, Richards S, Zhu D, et al. (2012) Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. PLoS Genet 8: e1002685.
115. Wimmer V, Lehermeier C, Albrecht T, Auinger HJ, Wang Y, et al. (2013) Genome-wide prediction of traits with different genetic architecture through efficient variable selection. Genetics 195: 573–587.
116. Zhang Z, Ober U, Erbe M, Zhang H, Gao N, et al. (2014) Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. PLoS ONE 9: e93017.
117. Speed D, Balding DJ (2014) MultiBLUP: improved SNP-based prediction for complex traits. Genome Res 24: 1550–1557.
118. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661–678.
119. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et al. (2011) Scikit-learn: machine learning in Python. J Machine Learn Res 12: 2825–2830.

Supporting Text S1

Implementation details of regularized machine learning models

This work aims to infer a model for making predictions for the disease class status through the use of supervised machine learning and feature selection techniques [1–4]. We assume that we are given a training set of m examples, each consisting of an n-dimensional feature vector, representing the SNP data and an associated outcome label. The aim of supervised learning is to infer a model from the training data that would allow predicting the labels for new data, for which only the features data is provided. Formally, we define the training set $\mathcal{T} = \{(\boldsymbol{X}_{1,:}, \boldsymbol{y}_1), \ldots, (\boldsymbol{X}_{m,:}, \boldsymbol{y}_m)\}$, where $\boldsymbol{y}_i \in \mathbb{R}^m$. By $\boldsymbol{X} \in \mathbb{R}^{m \times n}$ we denote a data matrix containing the feature vectors as rows and by $\boldsymbol{y}$ we denote the m-dimensional vector containing all the training set labels. By $\boldsymbol{X}_{i,:}$ we denote the i:th row, and by $\boldsymbol{X}_{:,i}$ the i:th column of $\boldsymbol{X}$. Our aim is to learn a prediction function $h : \mathbb{R}^n \to \boldsymbol{y}$ such that for new input-output pairs $(\mathbf{x}, y)$ it holds true that $h(\mathbf{x}) \approx y$.

Based on the choice of y we can recover a variety of different supervised learning settings. In binary classification, typically encoded as $y \in [-1, 1]$, we have two possible classes called the positive and the negative class. This is the standard way for modeling case-control studies where individuals belong to one of two classes based on whether they have a disease or not. In settings where there is a natural ordering over the classes, for example when the classes represent different stages of a disease in progression, the problem is known as ordinal regression. Finally, when $y \in \mathbb{R}$, we are presented with the problem of regression, where the aim is to learn to predict a real-valued variable, such as individual's blood pressure or weight.

In practice the model h is often implemented using one or several real-valued prediction functions $f : \mathbb{R}^n \to \mathbb{R}$. This is natural for regression, while binary classification may be implemented as a decision rule, for example it can defined by a transformation $h : \mathbb{R} \to \{-1, 1\}$ such that

$$h(q) = \begin{cases} 1, & \text{if } q > t \\ -1, & \text{if } q \leq t \end{cases} .$$

Here, one may as a default set $t = 0$ when aiming to minimize the misclassification cost, or choose some other threshold value when the misclassification costs between the two classes are known to be asymmetric.

Filter Methods for Genetic Feature Selection

The simplest form of feature selection that is commonly applied to GWAS is known as filter methods. Filters make use of the intrinsic properties of an individual's genetic variants to determine the bearing of the feature on the phenotype [5]. When features are selected by filter methods, the set of genetic variants are evaluated individually with a test statistic to determine their predictability for a particular phenotype. Let $\mathcal{S}$ denote the set of selected features, and $\mathcal{P}(\boldsymbol{X}_{:,i}, \boldsymbol{y})$ the value of some univariate statistic, that computes the error obtained when using the i^{th} feature, whose value for all training examples is contained in $\boldsymbol{X}_{:,i}$, for predicting the corresponding labels contained in $\boldsymbol{y}$. It can be assumed that a lower value for the statistic means better predictive power, as we can usually define simple transformations for statistics that do not behave this way (i.e. use 1-accuracy as an error measure instead of using accuracy). A number of studies have shown that filters were able to provide predictive results for their respective datasets [3, 6]. Some studies have coupled these filters with more advanced methods such as wrappers [2, 7]. These studies collectively argue that through the intelligent use of these algorithms researchers can help to explain a larger portion of the heritability of complex diseases and help to find more causal variants for these phenotypes.

Algorithm 1 presents the general pseudocode for the filter method. Those SNPs for which the value of the statistic is below a certain threshold are selected. Further, one may restrict the method to selecting only the top k variants, in which case the features need to be ranked by sorting the computed values. Running algorithm 1 can be implemented highly efficiently, as they require only a single pass through the data. In practice, standard GWAS analysis software such as PLINK [8] are capable of calculating single-locus statistical associations for entire GWAS in only a matter of minutes.

Algorithm 1 Filter-based feature selection

$\mathcal{S} \leftarrow \emptyset$ ▷ The set of selected features
for $i \in \{1, \ldots, n\}$ **do**
  $\epsilon_i \leftarrow \mathcal{P}(\boldsymbol{X}_{:,i}, \boldsymbol{y})$ ▷ Calculate the value of the statistic $\mathcal{P}$
  if $\epsilon_i < \epsilon_t$ **then** ▷ Select the feature if the value is below threshold ϵ_t
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{i\}$
return $\mathcal{S}$

There are a number of commonly used filters and which method to use is a subject of determination as to what provides the most meaningful results for a particular study. In case-control studies, a common analysis is to test the hypothesis of no-association between the SNP's values in cases and controls, represented in a 2×3 matrix (see Table 1). This matrix contains the counts of the three genotypes in the different subgroups. A less computationally intensive method is to treat the contingency table as 2×2 in which the entries represent the count of the total number of the allele possibilities in cases and controls (see Table 2).

	AA	**Aa**	**aa**	**Total**
Cases	z_1	z_2	z_3	R_1
Controls	z_4	z_5	z_6	R_2
Total	C_0	C_1	C_2	N

Table 1: Example of 2x3 genotypic table

	A	**a**	**Total**
Cases	$2z_1 + z_2$	$2z_3 + z_2$	$2R_1$
Controls	$2z_4 + z_5$	$2z_6 + z_5$	$2R_2$
Total	$2C_0 + C1$	$C_1 + 2C_2$	$2N$

Table 2: Example of 2x2 observed genotypic table

	A	**a**
Cases	$R_1(2C_0 + C_1)/N$	$R_1(C_1 + 2C_2)/N$
Controls	$R_2(2C_0 + C_1)/N$	$R_2(C_1 + 2C_2)/N$

Table 3: Example of 2x2 expected genotypic table

Given that the tables contain r rows and c columns, with R_i and C_j denoting the sums of the entries of row i and column j, respectively, the p-value for Fisher's Exact Test can be calculated with an application of the following equation [9].

$$p = \frac{\prod_{b=1}^{r} R_b! \prod_{d=1}^{c} C_d!}{N! \prod_{i=1}^{c \times r} z_i}. \tag{1}$$

Similar to Fisher's Exact test, one can also use the χ^2 test statistic to calculate the association between individual SNPs and the correct class labels. Based on the observed genotypes in Table 2 and the expected ones in Table 3, the test statistic can be computed with $(r-1)(c-1)$ degrees of freedom:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}}$$

Another commonly used metric is known as the odds ratio (OR) which measures the association of a variable on the class labels. This measure indicates how likely a given class label will be based on an

observed genotype compared to the likelihood of that same label without the appearance of the particular genotype [10]. Using Table 2 we can define the allele count based OR:

$$OR = \frac{(2z_1 + z_2)(2z_6 + z_5)}{(2z_3 + z_2)(2z_4 + z_5)}$$

If the OR = 1, it can be assumed that there is no association between the genotype and the disease. A value of OR > 1 represents that allele **A** increases the risk of the disease and an OR < 1 means that the occurrence of the disease is less likely.

Coordinate Descent Optimization for Feature Selection

With wrapper and embedded feature selection, we refer to methods for which the selection process is optimized for a particular learning algorithm. In traditional wrapper methods, feature selection is implemented as a meta-algorithm that performs feature selection as a search over the power set of features, and the learning algorithm is used as a black-box subroutine that evaluates the quality of different feature sets. Embedded methods on the other hand incorporate feature selection within a particular machine learning algorithm, for example by changing the objective function optimized so that it favors sparsity in addition to prediction performance. Wrapper methods can be considered as more general, whereas embedded methods are typically more efficient as they allow algorithm specific optimizations in implementing the methods. It is not always possible to draw clear distinctions between these two classes of approaches, as a specific optimized realizations of the general wrapper framework may in many cases be considered as embedded algorithms, as is the case for example the greedy RLS method considered later in this section.

We next present an optimization framework for wrapper and embedded methods, under which various types of feature selection methods can be conveniently considered. The linear models considered in the framework can be written as

$$f(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w} + b \tag{2}$$

where $\boldsymbol{w} = (w_1, \ldots, w_n)^{\mathrm{T}}$ is a vector of model parameters, b is the bias and $\mathbf{x} = (x_1, \ldots, x_n)^{\mathrm{T}}$ is a vector containing the feature values of a datum. Typically the bias term b is implemented by appending to each feature vector $\boldsymbol{x}$ a constant valued feature $\boldsymbol{x}_0 = 1$, so that we may define $\boldsymbol{w}_0 = b$, as this simplifies the notation. When dealing with this data representation, we assume that the feature selection algorithms always automatically select the feature corresponding to the bias term. Thus, the model (2) will be subsequently written without the bias term.

The training algorithms for learning the above considered types of linear models can be expressed as a following optimization problem:

$$\begin{aligned} &\underset{\boldsymbol{w}\in\mathbb{R}^n}{\operatorname{argmin}} \textstyle\sum_{i=1}^{m} L\left(f(\boldsymbol{X}_{i,:}), \boldsymbol{y}_i\right) \\ &\text{subject to } C(\boldsymbol{w}) \ , \end{aligned} \tag{3}$$

where L is a loss function indicating how the prediction obtained for the i^{th} training example fits to its label, and C is a constraint function. One of the most well-known and most widely used constraint functions is the so-called quadratic (or ridge) regularizer

$$C(\boldsymbol{w}) \equiv \|\boldsymbol{w}\|_2^2 < r, \tag{4}$$

where $r \in \mathbb{R}^+$. Constraining the norms of the models is traditionally used for finding a balance between fitting the model to the training data and the complexity of the model. However, this constraint alone tends to favor models that depend on all features.

One of the simplest sparsity enforcing constraints is the one setting a hard limit on the number of features the model can depend on, that is, on the number of nonzero entries in the model vector $\boldsymbol{w}$, known in the

literature as the zero norm $\|\boldsymbol{w}\|_0 = |\{i \mid w_i \neq 0\}|$ of the model vector. This can be formally expressed as follows:

$$C(\boldsymbol{w}) \equiv \|\boldsymbol{w}\|_0 \leq k, \tag{5}$$

where $k \in \mathbb{N}$ is a user given limit which the number of features must not exceed. The discrete and non-convex nature of the constraint (5) makes its direct optimization challenging, and one must resort to combinatorial optimization techniques, such as discrete searches over the space of all feature subsets. In the literature, the methods are often referred to as wrapper-based feature selection methods, which originate from the idea of retraining a model from scratch for each step of the search process. Namely, the search algorithm is "wrapped" around a base training algorithm that outputs a new model for each tested subset of features. This can be very slow in practice due to the size of the search space growing exponentially in the number of features and due to the slow training speed of the base learning algorithms. As we will show below in more detail, the large search space can be countered by designing smart search heuristics and the training speed can be accelerated by taking advantage of the models trained during the previous iterations of the search algorithm. An extra benefit of the direct search of feature subsets is that, instead of optimizing the training error, one can also optimize more sophisticated objectives, such as the leave-one-out cross-validation error for which some learning algorithms have easy to optimize closed-form solutions.

One of the oldest computationally efficient algorithms for directly optimizing the least-squares loss with the discrete constraint (5) are the so-called greedy least squares (GLS) methods, also known as orthogonal matching pursuit [11]. Another typical example of a direct optimization approach is the greedy RLS algorithm proposed by us [12]. It uses both (4) and (5) simultaneously to constrain the space of the models, leave-one-out cross-validation as a search heuristic and combinatorial search as an optimization method.

Algorithm 2 Greedy coordinate descent

1: $\mathcal{S} \leftarrow \emptyset$
2: **while** $|\mathcal{S}| < k$ **do**
3: $\quad b, r \leftarrow \text{argmin}_{b \in \{1,\ldots,n\} \setminus \mathcal{S}, r \in \mathbb{R}} J(\boldsymbol{w} + r\boldsymbol{e}_b)$
4: $\quad \mathcal{S} \leftarrow \mathcal{S} \cup \{b\}$
5: $\quad \boldsymbol{w} \leftarrow \boldsymbol{w} + r\boldsymbol{e}_b$
6: **return** $\mathcal{S}$

This type of approaches can be conveniently considered under the framework of coordinate descent methods. Coordinate descent (see e.g. [13]) fixes all entries of the vector except one and updates its value so that the value of the objective function will be reduced. A greedy coordinate descent is illustrated in Algorithm 2. The algorithm starts from an initial set of features that can be an empty set as in greedy forward selection and other similar approaches. During each iteration the algorithm prepares a set of candidate feature sets that usually differ only slightly from the current set of selected features. The candidates cover all subsets that have one extra feature in addition to the features already selected. Next, the algorithm selects from the candidates the one that is optimal with respect to the function to be optimized and subsequently updates the model according to the new set of selected features. The algorithm stops when a predefined number of features k have been selected, or based on other varying criteria, such as the identification of an optima.

Instead of directly optimizing the hard constraint of type (5), an alternative approach is to approximate it with an easier to optimize proxy function, such as the so-called 1-norm of the model vector

$$C(\boldsymbol{w}) \equiv \|\boldsymbol{w}\|_1 \leq r. \tag{6}$$

Unlike the 2-norm based constraint (4) this tends to favor sparse models depending only on a subset of the original features, and unlike (5), this is a convex and continuous constraint. Combined with a convex loss function, the corresponding optimization problem is considerably easier to solve than those with discrete non-convex constraints due to the objective function having a global optimum easily searchable with the powerful family of convex optimization methods.

Algorithm 3 Cyclic coordinate descent

```
while not converged do
    for j = 1, ..., n do
        w_j ← w_j + argmin_{r∈ℝ} J(w + r e_j)
```

The methods resorting to the convex approximation constraint (6), are usually called the embedded methods, since the feature selection mechanism can be considered to be built into the training algorithm itself. While there is a long history of various convex optimization methods being applied for training the embedded methods, currently the most popular ones are coordinate descent algorithms due to their simplicity and computational efficiency.

In this case, cyclic coordinate descent may be applied for minimizing the objective function (see Figure 1). The method repeats coordinate descent steps for each coefficient in the model vector at a time in a cyclic fashion, until the solution has converged or if a pre-defined maximum number of passes through the whole data has been performed. The idea is illustrated in Algorithm 3, where J denotes the constrained objective function and $\boldsymbol{e}_j$ is the jth standard basis vector of $\mathbb{R}^n$ (e.g. the jth element of $\boldsymbol{e}_j$ is 1 while the other entries are zero).

The most well-known algorithm involving the constraint (6) is known as Lasso or basis pursuit in the fields of machine learning and signal processing, respectively. Elastic Net is a variation of Lasso that simultaneously uses both (4) and (6). Lasso and Elastic Net are both least-squares regression methods but ℓ_1-regularization has also been employed together with other loss functions such as the logistic loss for binary classification. There has recently been growing sectors of research that have made use of embedded methods, primarily Lasso and similar ℓ_1-based methods, for the development of predictive models [4, 14–18].

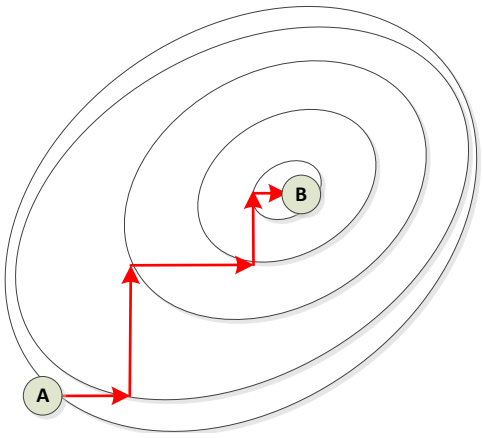

Figure 1: Example of coordinate descent starting from A and searching for the minimum B of a convex function.

Learning Algorithms

Let us denote $p = f(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{w} \rangle$, thus

$$\frac{\partial p}{\partial w_j} = x_j.$$

Then by chain rule for any loss function

$$\frac{\partial L(p,y)}{\partial w_j} = \frac{\partial L(p,y)}{\partial p}\frac{\partial p}{\partial w_j} = x_j \frac{\partial L(p,y)}{\partial p}.$$

From these we can define the derivatives needed for finding the coordinate descent step directions for a number of loss functions (see table 4)[1]. Based on these and the various regularizers listed in Table 5 we can construct a number of well-known machine learning algorithms, as seen in Table 6. Note that, while the zero-norm is not differentiable, the corresponding constraint is still satisfied with the greedy coordinate descent based training methods. Different algorithms vary on their respective computational complexities (see table 7).

Name	$L(p, y)$	$\frac{\partial L(p,y)}{\partial p}$
Squared	$(p-y)^2$	$2(p-y)$
Logistic	$\log(1+e^{-yp})$	$\frac{-y}{1+e^{yp}}$
Hinge	$\max(0, 1-yp)$	0 if $y \geq 1$, $-y$ else

Table 4: Common loss functions

Name	$\Omega(\boldsymbol{w})$	$\frac{\partial \Omega(\boldsymbol{w})}{\partial w_i}$
ℓ_0	$\|\boldsymbol{w}\|_0$	–
ℓ_1	$\|\boldsymbol{w}\|_1$	$\text{sign}(w_j)$
ℓ_2	$\|\boldsymbol{w}\|_2^2$	$2w_j$

Table 5: Common regularizers

Method	S	L	H	ℓ_1	ℓ_2	ℓ_0	R/C	Ref.
Lasso	•			•			R	[19]
Elastic Net	•			•	•		R	[20]
ℓ_1 Logistic		•		•			C	[21]
ℓ_2 Logistic		•			•		C	[22]
ℓ_1 SVM			•	•			C	[23]
SVM			•		•		C	[24]
OLS	•						R	[25]
Greedy RLS	•				•	•	R	[12]
Ridge Reg	•				•		R	[26]
GLS	•					•	R	[11]

Table 6: Construction of various methods based on different loss functions and regularizers. *S*, *L* and *H* stand for squared loss, logistic loss and hinge loss, respectively. R/C denotes whether the method is a (R)egression or a (C)lassification method only, all the regression methods can also be used for classification. OLS stands for ordinary least squares and *Ridge Reg* for ridge regression. *GLS* represents greedy least squares. These methods may also be known by other names in the literature, for example ridge regression is also known as regularized least squares.

[1] To be exact, the hinge loss and ℓ_1 norm are not differentiable everywhere, for these we provide subderivatives.

Method	Complexity
Greedy RLS	$O(kmn)$
Filter	$O(mn)$
Cyclic Coordinate Descent	$O(mnD)$
Greedy Coordinate Descent	$O(kmn)$

Table 7: Computational complexities of various feature selection methodologies. Here D represents the number of iterations necessary for the algorithm to cycle through until convergence. This is a data dependent variable and will vary depending on the study examined.

Examples in genetic prediction of complex traits

Next, we consider representative examples of implementations of the previously considered optimization framework, such that allow feature selection and can scale to entire GWAS without the need for pre-filtering. Greedy least squares (GLS) [11], minimizes the squared loss with a zero-norm constraint, e.g. the number of nonzero features is restricted. A straightforward approach for training GLS is to use the greedy coordinate descent (see Algorithm 2), which greedily selects one feature at a time and updates the model accordingly, until the number of features determined by the constraint is selected. The method is very simple to implement and the greedy search steps can be accelerated by caching the results from previous iterations. Accordingly, the computational complexity is only linear with respect to the number of features, data and the constraint (Table 7). The squared loss for m data will become zero at the latest after m linearly independent features has been selected, and hence GLS cannot be used to select more than that.

Greedy RLS [12] is similar to GLS except that it uses a combination of zero- and two-norm constraints and the it is trained with greedy coordinate descent that optimizes the leave-one-out cross-validation error rather than a traditional type of objective function. That is, the method yields identical results to running a traditional greedy forward feature selection wrapper on quadratically regularized least-squares (RLS) (e.g. ridge regression), but does so with computational shortcuts that are similar to those used in embedded methods. Formally, the selection heuristic $\mathcal{H}$ is the leave-one-out cross-validation error measured on RLS trained with features $\mathcal{S} \cup \{j\}$. Formally, it can be expressed as

$$\mathcal{H}(\mathcal{S} \cup \{j\}) = \sum_{i=1}^{m} \left(\boldsymbol{X}_{i,:} \boldsymbol{w}^{(i)} - \boldsymbol{y}_i \right)^2 , \tag{7}$$

where $\boldsymbol{w}^{(i)}$ is the RLS model trained with the whole training dataset except the ith datum and using the features indexed by $\mathcal{S} \cup \{j\}$, that is, the minimizer of

$$\sum_{h \neq i} \left(\boldsymbol{X}_{h,\mathcal{S} \cup \{j\}} \boldsymbol{w} - \boldsymbol{y}_h \right)^2 + \lambda \|\boldsymbol{w}\|_2^2.$$

Despite the use of leave-one-out based selection heuristic, the running time of Greedy RLS is analogous to that of GLS, it scales linearly with respect to the number of features selected, the total number of features and the number of examples.

The class of feature selection methods based on ℓ_1-norm regularization incorporates a wide variety of methods. Here we focus primarily on Lasso and the Elastic Net method, but other variations can be obtained simply by changing the loss function. The objective function that Lasso minimizes is [20]:

$$\sum_{i=1}^{m} \left(\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{w} \right)^2 + \lambda \|\boldsymbol{w}\|_1 . \tag{8}$$

The number of features selected by Lasso and the running time of the algorithm are dependent on the value of the regularization parameter λ. The larger is the value, the smaller both the number of features being

selected and the number of iterations until convergence tend to be. However, unlike the methods directly optimizing the ℓ_0-norm, the dependence for Lasso is indirect in the sense that one can not tell exactly how many features will get selected with a given parameter value before training the model or how many iterations will be required, as both of them are very much dependent on the data. Similarly to GLS, Lasso is only good for selecting less than m features [20], which may be problematic in GWAS where there is often a small effect size for the individual variants and large numbers of SNPs may be necessary to produce suitable models. The use of cross-validation for selecting the value of λ can make the problem even worse, since part of the data must be reserved for validating the parameter values.

The method known as Elastic Net avoids the above-described drawback of Lasso to some extent. It represents a continuum between the Lasso and ridge regression methods, with the method acting as Lasso when $\lambda_1 > 0$ and $\lambda_2 = 0$ and as ridge regression when $\lambda_1 = 0$ and $\lambda_2 > 0$. In (9), the quadratic component removes the limitation of the number of selected variables with Lasso, it encourages grouping and helps to stabilize the ℓ_1 regularization [20]. As mentioned in [20], this means that the Elastic Net is a more viable solution for methods making use of grouping effects. By grouping we mean the effect of correlated features having a similar effect on the model, which might be a useful method in applications such as in pathway analysis. Zou et al. calls the equation $(1-\alpha)\|\boldsymbol{w}\|_1 + \alpha\|\boldsymbol{w}\|_2^2$ the Elastic Net penalty, where $\alpha = \lambda_2/(\lambda_2 + \lambda_1)$ [20].

$$\sum_{i=1}^{m} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{w})^2 + \lambda_2\|\boldsymbol{w}\|_2^2 + \lambda_1\|\boldsymbol{w}\|_1 \tag{9}$$

Having the capability to fluctuate between Lasso and ridge regression affords the opportunity to provide both sparse solutions while, allowing for the shrinkage necessary for the model to not overfit the training data. Additionally, as it can be trained through the use of efficient optimization techniques such as coordinate descent, it potentially does not suffer from as high run times as wrapper methods. This does not mean that Elastic Net will not lead to universally better results than other leading algorithms, as the suitability of the method can depend on the applied search function and the dataset used.

Method Implementations for the Experiments

In order to examine their practical performance on common datasets, representative examples of machine learning methods were implemented on two SNP studies. The first experiment used WTCCC-T1D cases, combined with the UK National Blood Service and the 1958 Birth Cohort's control sets [27]. We implemented quality control procedures to filter based on having a MAF (5%), missing rate (1%), HWE (0.001), the genotype quality score (Chiamo value of 0.9) and exclusion lists provided by the WTCCC. This resulted in a set of 1,915 cases and 2,871 controls. The genotypes were encoded in the prediction models as binary genetic features through a genotype model which transforms each minor allele dosage (represented by 0,1,2) into three binary features: 0→(1,0,0), 1→ (0,1,0), 2→ (0,0,1), NA→ (0,0,0). This resulted in 986,331 binary genetic features in the T1D case, tripling the dataset size and memory requirements, while allowing the models to deal with missing genotype data (NA).

The second experiment involved a genetic cross between two yeast strains (Y2C), originally used to estimate the effect of epistasis on missing heritability of various quantitative yeast traits under highly-controlled laboratory conditions [28]. This dataset consisted of 1,008 segregants, 30,594 SNPs and 46 traits. We randomly selected one of the traits (Tunicamycin), but note that the Y2C data could be used to model genetic interactions also across multiple quantitative traits (pleiotropy, see Discussion). The only quality control procedure that we implemented was to remove those individuals with missing values for the particular trait. The original genotype data was adjusted so that variants sharing the same genotype across all of the segregants were merged, resulting in a set of 11,623 genetic features. Since the haploid dataset contains only two genotypic values (R and B), the genotypes were encoded by binary features (1 and 0).

For the case-control T1D data, we implemented a standard χ^2–based filter, by first selecting the top 500 variants according to their association p-values for each of the external folds, followed by L_2-regularized (ridge) regression. For wrappers, we used our greedy L_2-regularized least squares (RLS) implementation [1],

while for embedded methods, Lasso, Elastic Net and L_1-penalized logistic regression, were implemented through the Scikit-Learn package [29]. As a baseline reference method, we used the log odds-ratio weighted polygenic model [30], implemented as a weighted sum of the minor allele dosage in the 500 selected variants within each individual. For the quantitative Y2C data, we compared the performance of the greedy RLS, Lasso and Elastic Nets to a filter method, which selected the top 1,000 variants based on R^2, and then optimized the L_2-regularization parameter for RLS using nested CV. As a baseline method, we implemented a greedy version of least squares (LS), as this represents a model that is theoretically similar to the stepwise forward regression used in the original work [28]; greedy LS differs from the greedy RLS in terms that it implements regularization through optimization of L_0 norm instead of L_2 norm used in RLS.

Due to the large amount of processing power and memory needed for performing GWAS-scale experiments, a supercomputer at CSC - Finland's IT Center for Computer Science was used. The Hippu server is composed of a pair of HP ProLiant DL580 G7's and a pair of HP ProLiant DL785 G5s. The machine has an Rpeak of 1280 Gflop/s, and the two G7 servers have a total of 2 TB of memory while the two G5 servers have a total of 1 TB of memory. Additionally, the G7s are equipped with a total of eight 8-core Intel Xeon processors and the G5's have a total of 16 quad-core AMD Opteron processors.

As expected, filters have the lowest running times as they simply calculate a test statistic over all features. While greedy RLS tends to be fast when selecting a small amount of features, the cyclic coordinate descent based Lasso and Elastic Net implementations are more efficient when selecting a large amount of features. The main computational bottleneck, however, results from the need to select the hyperparameters of the learners. Selecting the ℓ_2-regularization parameter for greedy RLS and Elastic Net requires training the methods K-times for each tested parameter value, when K-fold cross-validation is used to select the parameter value. Further, when using a ℓ_1-regularizer for controlling the amount of selected features, such as is the case for Lasso and Elastic Net, a grid of parameters needs to be tested. Finally, if we do not have separate test sets, selecting parameters and evaluating test performance requires nested cross-validation, where inner cross-validation is used for parameter selection and outer-cv for performance evaluation. Combining nested cross-validation with parameter grid searches results in a combinatorial explosion that results in running times that are measured in days (e.g. Lasso), or weeks (Elastic Net and greedy RLS). This problem can be alleviated by using smaller or sparser parameter grids, small amount of folds or simpler heuristics for parameter selection. For example, for greedy RLS one may estimate the regularization parameters based on a filtered subset of the data and still provide a reasonable estimate. This allows selecting the hyperparameters in a matter of minutes.

Computational Validation of Predictive Accuracy

As the models become increasingly complex, their prediction errors decrease with the number of selected variants and other model parameters, which capture increasing details of the training data. However, this is true only to a certain extent for independent test data; while increasing complexity first allows for more accurate modeling, the test set error begins later to increase as the complexity of the model is no longer improving the generalization power [31]. Such model overfitting necessitates the use of a careful model validation, even after model regularization. Since the use of the same dataset during both the model construction and model validation may lead to a severe selection bias [32] resulting in overoptimistic estimates of predictive accuracy, separate validation data is needed.

A straightforward validation option is to apply the trained model onto an independent set of samples, which has not been examined during the whole model construction process. However, in addition to leading to a smaller proportion of training data, perhaps affecting the model generalizability, a within-study hold-out validation is prone to being effected by any experimental errors that may exist in the particular study. Between-study evaluation is a valid option in case such replication samples are available, especially if the model is intended to generalize beyond the genetic background of the training subjects; otherwise, population stratification methods may be needed to make the population structures more comparable [33].

Especially when limited numbers of samples are available, some type of cross-validation (CV) is frequently used to evaluate the predictive performance [1, 34]. In the simple K-fold CV, the sample is randomly

partitioned into K subsamples of equal size; the model is first trained on K-1 subsamples and then validated on the remaining sample (Figure 2). This process is repeated K times and an average over the K folds is used as an estimate of the predictive performance. Stratified CV guarantees that the phenotypic effect is similar in each fold; in disease classification, for instance, each fold contains approximately equal proportions of cases and controls. Further, when one needs to use CV both for parameter selection (including feature selection) and for estimating the accuracy of the learned model, the CV procedure should be nested. That is, on each round of CV (outer CV), where the data is split into a training set consisting of $K-1$ folds and the test set formed from the remaining fold, one performs also CV on this training set (inner CV) in order to select the learner parameters (see Figure 2). Such procedures can provide performance estimates free of a selection bias [31, 35]. After this estimate has been computed, the final model construction or feature selection can be performed on all the available data combined in order to use all the information available. In the two examples cases considered here the performance evaluation was implemented using nested 3-fold CV.

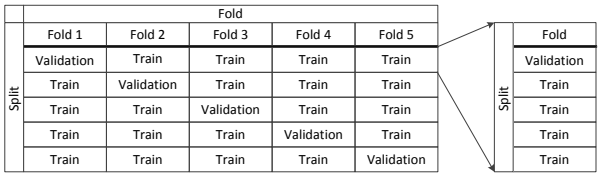

Figure 2: Organization of the standard cross-validation (left) and nested cross-validation (right) in terms of splitting the genetic data into the training and validation folds.

One of the most commonly reported metrics for quantifying the performance in the case-control setting is the area under the receiver operating characteristic curve (AUC). While having some specific caveats, the AUC has the advantage over many other metrics of being invariant to unbalanced settings, where the number of cases (m^+) and controls (m^-) is drastically different [1, 33, 36–38]. The AUC corresponds to the probability that the predicted phenotype of a randomly selected case ($\hat{y}_i^+$) will be ranked higher than that of a randomly selected control ($\hat{y}_j^-$). In its most basic formulation [39], the AUC can be summarized as

$$AUC = \frac{1}{m^+ m^-} \sum_{j=1}^{m^+} \sum_{k=1}^{m^-} g(\hat{\boldsymbol{y}}_j^+ - \hat{\boldsymbol{y}}_k^-) \tag{10}$$

where $g(x) = 0$, 0.5 and 1 if $x < 0$, $x = 0$ and $x > 0$, respectively. For an ideal classifier, AUC = 1, whereas a random classifier obtains an AUC = 0.5 on average. The AUC is closely related to the Mann-Whitney test statistic. While being useful in many applications, any single summary metric cannot capture all of the different tradeoffs in the predictive modeling. For instance, the true positive rate (sensitivity) is often more important in clinical applications than the false positive rate (1-specificity). The partial AUC can be used in such applications to integrate the sensitivity levels of a model up to a specified specificity cut-off.

In regression problems, the predictive accuracy for a continuous trait is often evaluated in terms of the coefficient of determination (R^2). This metric corresponds to the proportion of the phenotypic variance explained by the genetic model. Using squared errors between the observed y_i and predicted $\hat{y}_i$ phenotypes, R^2s is formally defined by the ratio of the variance accounted for by the model fitted to the training set relative to the variance of the phenotypic trait in the validation sample:

$$R^2 = 1 - \frac{\sum_{i=1}^{m} (\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i)^2}{\sum_{i=1}^{m} (\boldsymbol{y}_i - \bar{\boldsymbol{y}})^2} \tag{11}$$

Here, $\bar{y}$ is the mean of the phenotypic trait over all the m individuals. Higher values of R^2 indicate larger portion of explained variance and hence a more predictive model. R^2 is also related to the estimated heritability (h^2), which corresponds to the proportion of phenotypic variance explained by true genetic values in the base population; however, since R^2 ignores inbreeding, relationships between individuals and estimation errors, it cannot be used as a consistent estimate of heritability [40].

References

1. 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.** *Algorithms for Molecular Biology* 2012, **7**:11.
2. Okser S, Lehtimäki T, Elo LL, Mononen N, Peltonen N, Kähönen M, Juonala M, Fan YM, Hernesniemi JA, Laitinen T, Lyytikäinen LP, Rontu R, Eklund C, Hutri-Kähönen N, Taittonen L, Hurme M, Viikari JSA, Raitakari OT, Aittokallio T: **Genetic Variants and Their Interactions in the Prediction of Increased Pre-Clinical Carotid Atherosclerosis: The Cardiovascular Risk in Young Finns Study**. *PLoS Genetics* 2010, **6**(9):e1001146.
3. Wei Z, Wang K, Qu HQ, Zhang H, Bradfield J, Kim C, Frackleton E, Hou C, Glessner JT, Chiavacci R, Stanley C, Monos D, Grant SFA, Polychronakos C, Hakonarson H: **From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes**. *PLoS Genetics* 2009, **5**(10):e1000678.
4. Kooperberg C, LeBlanc M, Obenchain V: **Risk prediction using genome-wide association studies**. *Genetic epidemiology* 2010, **34**(7):643–652.
5. Saeys Y, Inza I, Larrañaga P: **A review of feature selection techniques in bioinformatics**. *Bioinformatics* 2007, **23**(19):2507–2517.
6. Roshan U, Chikkagoudar S, Wei Z, Wang K, Hakonarson H: **Ranking causal variants and associated regions in genome-wide association studies by the support vector machine and random forest.** *Nucleic acids research* 2011, **39**(9).
7. Long N, Gianola D, Rosa GJM, Weigel KA, Avendano S: **Machine learning classification procedure for selecting SNPs genomic selection: application to early mortality in broilers**. *Journal of Animal Breeding and Genetics* 2007, **124**(6):377–389.
8. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC: **PLINK: a tool set for whole-genome association and population-based linkage analyses**. *American journal of human genetics* 2007, **81**(3):559–575.
9. Raymond M, Rousset F: **An Exact Test for Population Differentiation**. *Evolution* 1995, **49**(6):1280–1283.
10. Szumilas M: **Explaining odds ratios**. *Journal of the Canadian Academy of Child and Adolescent Psychiatry* 2010, **19**(3):227–229.
11. Pati Y, Rezaiifar R, Krishnaprasad PS: **Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition**. In *Proceedings of the 27th Annual Asilomar Conference on Signals, Systems, and Computers* 1993:40–44.
12. Pahikkala T, Airola A, Salakoski T: **Speeding up Greedy Forward Selection for Regularized Least-Squares**. In *Proceedings of The Ninth International Conference on Machine Learning and Applications (ICMLA'10)*. Edited by Draghici S, Khoshgoftaar TM, Palade V, Pedrycz W, Wani MA, Zhu X, IEEE Computer Society 2010.
13. Nocedal J, Wright SJ: *Numerical Optimization*. Springer 2000.
14. Shi G, Boerwinkle E, Morrison AC, Gu CC, Chakravarti A, Rao DC: **Mining gold dust under the genome wide significance level: a two-stage approach to analysis of GWAS.** *Genetic epidemiology* 2011, **35**(2):111–118.
15. Srivastava S, Chen L: **Comparison between the stochastic search variable selection and the least absolute shrinkage and selection operator for genome-wide association studies of rheumatoid arthritis.** *BMC Proceedings* 2009, **3**(Suppl 7):S21.
16. González-Recio O, de Maturana EL, Vega AT, Engelman CD, Broman KW: **Detecting single-nucleotide polymorphism by single-nucleotide polymorphism interactions in rheumatoid arthritis using a two-step approach with machine learning and a Bayesian threshold least absolute shrinkage and selection operator (LASSO) model.** *BMC Proceedings* 2009, **3**(Suppl 7):S63.
17. Wei Z, Wang W, Bradfield J, Li J, Cardinale C, Frackelton E, Kim C, Mentch F, Van Steen K, Visscher PM, Baldassano RN, Hakonarson H: **Large Sample Size, Wide Variant Spectrum, and Advanced Machine-Learning Technique Boost Risk Prediction for Inflammatory Bowel Disease**. *American journal of human genetics* 2013, **92**(6):1008–1012.
18. Li J, Das K, Fu G, Li R, Wu R: **The Bayesian lasso for genome-wide association studies**. *Bioinformatics* 2011, **27**(4):516–523.

19. Tibshirani R: **Regression Shrinkage and Selection Via the Lasso**. *Journal of the Royal Statistical Society, Series B* 1994, **58**:267–288.

20. Zou H, Hastie T: **Regularization and Variable Selection via the Elastic Net**. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2003, **67**(2):301–320.

21. Ng AY: **Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance**. In *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04 2004:78–85.

22. Park MY, Hastie T: **Penalized logistic regression for detecting gene interactions**. *Biostatistics* 2008, **9**:30–50.

23. Zhu J, Rosset S, Tibshirani R, Hastie TJ: **1-norm Support Vector Machines**. In *Advances in Neural Information Processing Systems 16*. Edited by Thrun S, Saul L, Schölkopf B, MIT Press 2004:49–56.

24. Vapnik VN: *The nature of statistical learning theory*. New York, NY, USA: Springer-Verlag New York, Inc. 1995.

25. Chetty VK: **Ordinary Least Squares Regression**. In *Encyclopedia of Medical Decision Making*. Edited by Kattan MW, SAGE Publications, Inc. 2009:838–844.

26. Hoerl AE, Kennard RW: **Ridge Regression: Biased Estimation for Nonorthogonal Problems**. *Technometrics* 1970, **12**:55–67.

27. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, Kwiatkowski DP, McCarthy MI, Ouwehand WH, Samani NJ, Todd JA, Donnelly P, Et Al: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls**. *Nature* 2007, **447**:661–678.

28. Bloom JS, Ehrenreich IM, Loo W, Lite TLV, Kruglyak L: **Finding the sources of missing heritability in a yeast cross**. *Nature* 2013, **494**(7436):234–237.

29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E: **Scikit-learn: Machine Learning in Python**. *Journal of Machine Learning Research* 2011, **12**:2825–2830.

30. Evans DM, Visscher PM, Wray NR: **Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk.** *Human molecular genetics* 2009, **18**(18):3525–3531.

31. Okser S, Pahikkala T, Aittokallio T: **Genetic variants and their interactions in disease risk prediction - machine learning and network perspectives.** *BioData mining* 2013, **6**:5.

32. Ambroise C, McLachlan GJ: **Selection bias in gene extraction on the basis of microarray gene-expression data**. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(10):6562–6566.

33. Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM: **Pitfalls of predicting complex traits from SNPs.** *Nature reviews. Genetics* 2013, **14**(7):507–515.

34. Kruppa J, Ziegler A, König I: **Risk estimation and risk prediction using machine-learning methods.** *Human genetics* 2012, **131**(10):1639–1654.

35. Varma S, Simon R: **Bias in error estimation when using cross-validation for model selection**. *BMC Bioinformatics* 2006, **7**(91).

36. Jostins L, Barrett JC: **Genetic risk prediction in complex disease.** *Human Molecular Genetics* 2011, **20**(R2):R182–8.

37. 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 Genetics* 2009, **5**(2):e1000337.

38. Janssens A, van Duijn CM: **Genome-based prediction of common diseases: methodological considerations for future research.** *Genome Medicine* 2009, **1**(20).

39. Hanley JA, McNeil BJ: **The meaning and use of the area under a receiver operating characteristic (ROC) curve**. *Radiology* 1982, **143**:29–36.

40. Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, de los Campos G: **Beyond Missing Heritability: Prediction of Complex Traits**. *PLoS Genetics* 2011, **7**(4):e1002051.

Turku Centre for Computer Science
TUCS Dissertations

1. **Marjo Lipponen**, On Primitive Solutions of the Post Correspondence Problem
2. **Timo Käkölä**, Dual Information Systems in Hyperknowledge Organizations
3. **Ville Leppänen**, Studies on the Realization of PRAM
4. **Cunsheng Ding**, Cryptographic Counter Generators
5. **Sami Viitanen**, Some New Global Optimization Algorithms
6. **Tapio Salakoski**, Representative Classification of Protein Structures
7. **Thomas Långbacka**, An Interactive Environment Supporting the Development of Formally Correct Programs
8. **Thomas Finne**, A Decision Support System for Improving Information Security
9. **Valeria Mihalache**, Cooperation, Communication, Control. Investigations on Grammar Systems.
10. **Marina Waldén**, Formal Reasoning About Distributed Algorithms
11. **Tero Laihonen**, Estimates on the Covering Radius When the Dual Distance is Known
12. **Lucian Ilie**, Decision Problems on Orders of Words
13. **Jukkapekka Hekanaho**, An Evolutionary Approach to Concept Learning
14. **Jouni Järvinen**, Knowledge Representation and Rough Sets
15. **Tomi Pasanen**, In-Place Algorithms for Sorting Problems
16. **Mika Johnsson**, Operational and Tactical Level Optimization in Printed Circuit Board Assembly
17. **Mats Aspnäs**, Multiprocessor Architecture and Programming: The Hathi-2 System
18. **Anna Mikhajlova**, Ensuring Correctness of Object and Component Systems
19. **Vesa Torvinen**, Construction and Evaluation of the Labour Game Method
20. **Jorma Boberg**, Cluster Analysis. A Mathematical Approach with Applications to Protein Structures
21. **Leonid Mikhajlov**, Software Reuse Mechanisms and Techniques: Safety Versus Flexibility
22. **Timo Kaukoranta**, Iterative and Hierarchical Methods for Codebook Generation in Vector Quantization
23. **Gábor Magyar**, On Solution Approaches for Some Industrially Motivated Combinatorial Optimization Problems
24. **Linas Laibinis**, Mechanised Formal Reasoning About Modular Programs
25. **Shuhua Liu**, Improving Executive Support in Strategic Scanning with Software Agent Systems
26. **Jaakko Järvi**, New Techniques in Generic Programming – C++ is more Intentional than Intended
27. **Jan-Christian Lehtinen**, Reproducing Kernel Splines in the Analysis of Medical Data
28. **Martin Büchi**, Safe Language Mechanisms for Modularization and Concurrency
29. **Elena Troubitsyna**, Stepwise Development of Dependable Systems
30. **Janne Näppi**, Computer-Assisted Diagnosis of Breast Calcifications
31. **Jianming Liang**, Dynamic Chest Images Analysis
32. **Tiberiu Seceleanu**, Systematic Design of Synchronous Digital Circuits
33. **Tero Aittokallio**, Characterization and Modelling of the Cardiorespiratory System in Sleep-Disordered Breathing
34. **Ivan Porres**, Modeling and Analyzing Software Behavior in UML
35. **Mauno Rönkkö**, Stepwise Development of Hybrid Systems
36. **Jouni Smed**, Production Planning in Printed Circuit Board Assembly
37. **Vesa Halava**, The Post Correspondence Problem for Market Morphisms
38. **Ion Petre**, Commutation Problems on Sets of Words and Formal Power Series
39. **Vladimir Kvassov**, Information Technology and the Productivity of Managerial Work
40. **Frank Tétard**, Managers, Fragmentation of Working Time, and Information Systems

41. **Jan Manuch**, Defect Theorems and Infinite Words
42. **Kalle Ranto**, Z_4-Goethals Codes, Decoding and Designs
43. **Arto Lepistö**, On Relations Between Local and Global Periodicity
44. **Mika Hirvensalo**, Studies on Boolean Functions Related to Quantum Computing
45. **Pentti Virtanen**, Measuring and Improving Component-Based Software Development
46. **Adekunle Okunoye**, Knowledge Management and Global Diversity – A Framework to Support Organisations in Developing Countries
47. **Antonina Kloptchenko**, Text Mining Based on the Prototype Matching Method
48. **Juha Kivijärvi**, Optimization Methods for Clustering
49. **Rimvydas Rukšėnas**, Formal Development of Concurrent Components
50. **Dirk Nowotka**, Periodicity and Unbordered Factors of Words
51. **Attila Gyenesei**, Discovering Frequent Fuzzy Patterns in Relations of Quantitative Attributes
52. **Petteri Kaitovaara**, Packaging of IT Services – Conceptual and Empirical Studies
53. **Petri Rosendahl**, Niho Type Cross-Correlation Functions and Related Equations
54. **Péter Majlender**, A Normative Approach to Possibility Theory and Soft Decision Support
55. **Seppo Virtanen**, A Framework for Rapid Design and Evaluation of Protocol Processors
56. **Tomas Eklund**, The Self-Organizing Map in Financial Benchmarking
57. **Mikael Collan**, Giga-Investments: Modelling the Valuation of Very Large Industrial Real Investments
58. **Dag Björklund**, A Kernel Language for Unified Code Synthesis
59. **Shengnan Han**, Understanding User Adoption of Mobile Technology: Focusing on Physicians in Finland
60. **Irina Georgescu**, Rational Choice and Revealed Preference: A Fuzzy Approach
61. **Ping Yan**, Limit Cycles for Generalized Liénard-Type and Lotka-Volterra Systems
62. **Joonas Lehtinen**, Coding of Wavelet-Transformed Images
63. **Tommi Meskanen,** On the NTRU Cryptosystem
64. **Saeed Salehi**, Varieties of Tree Languages
65. **Jukka Arvo**, Efficient Algorithms for Hardware-Accelerated Shadow Computation
66. **Mika Hirvikorpi**, On the Tactical Level Production Planning in Flexible Manufacturing Systems
67. **Adrian Costea**, Computational Intelligence Methods for Quantitative Data Mining
68. **Cristina Seceleanu**, A Methodology for Constructing Correct Reactive Systems
69. **Luigia Petre**, Modeling with Action Systems
70. **Lu Yan**, Systematic Design of Ubiquitous Systems
71. **Mehran Gomari**, On the Generalization Ability of Bayesian Neural Networks
72. **Ville Harkke**, Knowledge Freedom for Medical Professionals – An Evaluation Study of a Mobile Information System for Physicians in Finland
73. **Marius Cosmin Codrea**, Pattern Analysis of Chlorophyll Fluorescence Signals
74. **Aiying Rong**, Cogeneration Planning Under the Deregulated Power Market and Emissions Trading Scheme
75. **Chihab BenMoussa**, Supporting the Sales Force through Mobile Information and Communication Technologies: Focusing on the Pharmaceutical Sales Force
76. **Jussi Salmi**, Improving Data Analysis in Proteomics
77. **Orieta Celiku**, Mechanized Reasoning for Dually-Nondeterministic and Probabilistic Programs
78. **Kaj-Mikael Björk**, Supply Chain Efficiency with Some Forest Industry Improvements
79. **Viorel Preoteasa**, Program Variables – The Core of Mechanical Reasoning about Imperative Programs
80. **Jonne Poikonen**, Absolute Value Extraction and Order Statistic Filtering for a Mixed-Mode Array Image Processor
81. **Luka Milovanov**, Agile Software Development in an Academic Environment
82. **Francisco Augusto Alcaraz Garcia**, Real Options, Default Risk and Soft Applications
83. **Kai K. Kimppa**, Problems with the Justification of Intellectual Property Rights in Relation to Software and Other Digitally Distributable Media
84. **Dragoş Truşcan**, Model Driven Development of Programmable Architectures
85. **Eugen Czeizler**, The Inverse Neighborhood Problem and Applications of Welch Sets in Automata Theory

86. **Sanna Ranto**, Identifying and Locating-Dominating Codes in Binary Hamming Spaces
87. **Tuomas Hakkarainen**, On the Computation of the Class Numbers of Real Abelian Fields
88. **Elena Czeizler**, Intricacies of Word Equations
89. **Marcus Alanen**, A Metamodeling Framework for Software Engineering
90. **Filip Ginter**, Towards Information Extraction in the Biomedical Domain: Methods and Resources
91. **Jarkko Paavola**, Signature Ensembles and Receiver Structures for Oversaturated Synchronous DS-CDMA Systems
92. **Arho Virkki**, The Human Respiratory System: Modelling, Analysis and Control
93. **Olli Luoma**, Efficient Methods for Storing and Querying XML Data with Relational Databases
94. **Dubravka Ilić**, Formal Reasoning about Dependability in Model-Driven Development
95. **Kim Solin**, Abstract Algebra of Program Refinement
96. **Tomi Westerlund**, Time Aware Modelling and Analysis of Systems-on-Chip
97. **Kalle Saari**, On the Frequency and Periodicity of Infinite Words
98. **Tomi Kärki**, Similarity Relations on Words: Relational Codes and Periods
99. **Markus M. Mäkelä**, Essays on Software Product Development: A Strategic Management Viewpoint
100. **Roope Vehkalahti**, Class Field Theoretic Methods in the Design of Lattice Signal Constellations
101. **Anne-Maria Ernvall-Hytönen**, On Short Exponential Sums Involving Fourier Coefficients of Holomorphic Cusp Forms
102. **Chang Li**, Parallelism and Complexity in Gene Assembly
103. **Tapio Pahikkala**, New Kernel Functions and Learning Methods for Text and Data Mining
104. **Denis Shestakov**, Search Interfaces on the Web: Querying and Characterizing
105. **Sampo Pyysalo**, A Dependency Parsing Approach to Biomedical Text Mining
106. **Anna Sell**, Mobile Digital Calendars in Knowledge Work
107. **Dorina Marghescu**, Evaluating Multidimensional Visualization Techniques in Data Mining Tasks
108. **Tero Säntti**, A Co-Processor Approach for Efficient Java Execution in Embedded Systems
109. **Kari Salonen**, Setup Optimization in High-Mix Surface Mount PCB Assembly
110. **Pontus Boström**, Formal Design and Verification of Systems Using Domain-Specific Languages
111. **Camilla J. Hollanti**, Order-Theoretic Mehtods for Space-Time Coding: Symmetric and Asymmetric Designs
112. **Heidi Himmanen**, On Transmission System Design for Wireless Broadcasting
113. **Sébastien Lafond**, Simulation of Embedded Systems for Energy Consumption Estimation
114. **Evgeni Tsivtsivadze**, Learning Preferences with Kernel-Based Methods
115. **Petri Salmela**, On Commutation and Conjugacy of Rational Languages and the Fixed Point Method
116. **Siamak Taati**, Conservation Laws in Cellular Automata
117. **Vladimir Rogojin**, Gene Assembly in Stichotrichous Ciliates: Elementary Operations, Parallelism and Computation
118. **Alexey Dudkov**, Chip and Signature Interleaving in DS CDMA Systems
119. **Janne Savela**, Role of Selected Spectral Attributes in the Perception of Synthetic Vowels
120. **Kristian Nybom**, Low-Density Parity-Check Codes for Wireless Datacast Networks
121. **Johanna Tuominen**, Formal Power Analysis of Systems-on-Chip
122. **Teijo Lehtonen**, On Fault Tolerance Methods for Networks-on-Chip
123. **Eeva Suvitie**, On Inner Products Involving Holomorphic Cusp Forms and Maass Forms
124. **Linda Mannila**, Teaching Mathematics and Programming – New Approaches with Empirical Evaluation
125. **Hanna Suominen**, Machine Learning and Clinical Text: Supporting Health Information Flow
126. **Tuomo Saarni**, Segmental Durations of Speech
127. **Johannes Eriksson**, Tool-Supported Invariant-Based Programming

128. **Tero Jokela**, Design and Analysis of Forward Error Control Coding and Signaling for Guaranteeing QoS in Wireless Broadcast Systems
129. **Ville Lukkarila**, On Undecidable Dynamical Properties of Reversible One-Dimensional Cellular Automata
130. **Qaisar Ahmad Malik**, Combining Model-Based Testing and Stepwise Formal Development
131. **Mikko-Jussi Laakso**, Promoting Programming Learning: Engagement, Automatic Assessment with Immediate Feedback in Visualizations
132. **Riikka Vuokko**, A Practice Perspective on Organizational Implementation of Information Technology
133. **Jeanette Heidenberg**, Towards Increased Productivity and Quality in Software Development Using Agile, Lean and Collaborative Approaches
134. **Yong Liu**, Solving the Puzzle of Mobile Learning Adoption
135. **Stina Ojala**, Towards an Integrative Information Society: Studies on Individuality in Speech and Sign
136. **Matteo Brunelli**, Some Advances in Mathematical Models for Preference Relations
137. **Ville Junnila**, On Identifying and Locating-Dominating Codes
138. **Andrzej Mizera**, Methods for Construction and Analysis of Computational Models in Systems Biology. Applications to the Modelling of the Heat Shock Response and the Self-Assembly of Intermediate Filaments.
139. **Csaba Ráduly-Baka**, Algorithmic Solutions for Combinatorial Problems in Resource Management of Manufacturing Environments
140. **Jari Kyngäs**, Solving Challenging Real-World Scheduling Problems
141. **Arho Suominen**, Notes on Emerging Technologies
142. **József Mezei**, A Quantitative View on Fuzzy Numbers
143. **Marta Olszewska**, On the Impact of Rigorous Approaches on the Quality of Development
144. **Antti Airola**, Kernel-Based Ranking: Methods for Learning and Performace Estimation
145. **Aleksi Saarela**, Word Equations and Related Topics: Independence, Decidability and Characterizations
146. **Lasse Bergroth**, Kahden merkkijonon pisimmän yhteisen alijonon ongelma ja sen ratkaiseminen
147. **Thomas Canhao Xu**, Hardware/Software Co-Design for Multicore Architectures
148. **Tuomas Mäkilä**, Software Development Process Modeling – Developers Perspective to Contemporary Modeling Techniques
149. **Shahrokh Nikou**, Opening the Black-Box of IT Artifacts: Looking into Mobile Service Characteristics and Individual Perception
150. **Alessandro Buoni**, Fraud Detection in the Banking Sector: A Multi-Agent Approach
151. **Mats Neovius**, Trustworthy Context Dependency in Ubiquitous Systems
152. **Fredrik Degerlund**, Scheduling of Guarded Command Based Models
153. **Amir-Mohammad Rahmani-Sane**, Exploration and Design of Power-Efficient Networked Many-Core Systems
154. **Ville Rantala**, On Dynamic Monitoring Methods for Networks-on-Chip
155. **Mikko Pelto**, On Identifying and Locating-Dominating Codes in the Infinite King Grid
156. **Anton Tarasyuk**, Formal Development and Quantitative Verification of Dependable Systems
157. **Muhammad Mohsin Saleemi**, Towards Combining Interactive Mobile TV and Smart Spaces: Architectures, Tools and Application Development
158. **Tommi J. M. Lehtinen**, Numbers and Languages
159. **Peter Sarlin**, Mapping Financial Stability
160. **Alexander Wei Yin**, On Energy Efficient Computing Platforms
161. **Mikołaj Olszewski**, Scaling Up Stepwise Feature Introduction to Construction of Large Software Systems
162. **Maryam Kamali**, Reusable Formal Architectures for Networked Systems
163. **Zhiyuan Yao**, Visual Customer Segmentation and Behavior Analysis – A SOM-Based Approach
164. **Timo Jolivet**, Combinatorics of Pisot Substitutions
165. **Rajeev Kumar Kanth**, Analysis and Life Cycle Assessment of Printed Antennas for Sustainable Wireless Systems
166. **Khalid Latif**, Design Space Exploration for MPSoC Architectures

167. **Bo Yang**, Towards Optimal Application Mapping for Energy-Efficient Many-Core Platforms
168. **Ali Hanzala Khan**, Consistency of UML Based Designs Using Ontology Reasoners
169. **Sonja Leskinen**, m-Equine: IS Support for the Horse Industry
170. **Fareed Ahmed Jokhio**, Video Transcoding in a Distributed Cloud Computing Environment
171. **Moazzam Fareed Niazi**, A Model-Based Development and Verification Framework for Distributed System-on-Chip Architecture
172. **Mari Huova**, Combinatorics on Words: New Aspects on Avoidability, Defect Effect, Equations and Palindromes
173. **Ville Timonen**, Scalable Algorithms for Height Field Illumination
174. **Henri Korvela**, Virtual Communities – A Virtual Treasure Trove for End-User Developers
175. **Kameswar Rao Vaddina**, Thermal-Aware Networked Many-Core Systems
176. **Janne Lahtiranta**, New and Emerging Challenges of the ICT-Mediated Health and Well-Being Services
177. **Irum Rauf**, Design and Validation of Stateful Composite RESTful Web Services
178. **Jari Björne**, Biomedical Event Extraction with Machine Learning
179. **Katri Haverinen**, Natural Language Processing Resources for Finnish: Corpus Development in the General and Clinical Domains
180. **Ville Salo**, Subshifts with Simple Cellular Automata
181. **Johan Ersfolk**, Scheduling Dynamic Dataflow Graphs
182. **Hongyan Liu**, On Advancing Business Intelligence in the Electricity Retail Market
183. **Adnan Ashraf**, Cost-Efficient Virtual Machine Management: Provisioning, Admission Control, and Consolidation
184. **Muhammad Nazrul Islam**, Design and Evaluation of Web Interface Signs to Improve Web Usability: A Semiotic Framework
185. **Johannes Tuikkala**, Algorithmic Techniques in Gene Expression Processing: From Imputation to Visualization
186. **Natalia Díaz Rodríguez**, Semantic and Fuzzy Modelling for Human Behaviour Recognition in Smart Spaces. A Case Study on Ambient Assisted Living
187. **Mikko Pänkäälä**, Potential and Challenges of Analog Reconfigurable Computation in Modern and Future CMOS
188. **Sami Hyrynsalmi**, Letters from the War of Ecosystems – An Analysis of Independent Software Vendors in Mobile Application Marketplaces
189. **Seppo Pulkkinen**, Efficient Optimization Algorithms for Nonlinear Data Analysis
190. **Sami Pyöttiälä**, Optimization and Measuring Techniques for Collect-and-Place Machines in Printed Circuit Board Industry
191. **Syed Mohammad Asad Hassan Jafri**, Virtual Runtime Application Partitions for Resource Management in Massively Parallel Architectures
192. **Toni Ernvall**, On Distributed Storage Codes
193. **Yuliya Prokhorova**, Rigorous Development of Safety-Critical Systems
194. **Olli Lahdenoja**, Local Binary Patterns in Focal-Plane Processing – Analysis and Applications
195. **Annika H. Holmbom**, Visual Analytics for Behavioral and Niche Market Segmentation
196. **Sergey Ostroumov**, Agent-Based Management System for Many-Core Platforms: Rigorous Design and Efficient Implementation
197. **Espen Suenson**, How Computer Programmers Work – Understanding Software Development in Practise
198. **Tuomas Poikela**, Readout Architectures for Hybrid Pixel Detector Readout Chips
199. **Bogdan Iancu**, Quantitative Refinement of Reaction-Based Biomodels
200. **Ilkka Törmä**, Structural and Computational Existence Results for Multidimensional Subshifts
201. **Sebastian Okser**, Scalable Feature Selection Applications for Genome-Wide Association Studies of Complex Diseases

Turku Centre *for* Computer Science

Joukahaisenkatu 3-5 B, 20520 Turku, Finland | www. tucs.fi

University of Turku
Faculty of Mathematics and Natural Sciences
- Department of Information Technology
- Department of Mathematics and Statistics

Turku School of Economics
- Institute of Information Systems Science

Åbo Akademi University
Faculty of Science and Engineering
- Computer Engineering
- Computer Science

Faculty of Social Sciences, Business and Economics
- Information Systems

ISBN 978-952-12-3245-9
ISSN 1239-1883

Sebastian Okser

Scalable Feature Selection Applications for Genome-Wide Association Studies of Complex Diseases

Sebastian Okser

Scalable Feature Selection Applications for Genome-Wide Association Studies of Complex Diseases