A computational framework for prioritization of disease-causing mutations

Dusan POPOVIC

Examination committee:
Prof. dr. ir. Hendrik Van Brussel, chair
Prof. dr. ir. Bart De Moor, supervisor
Prof. dr. ir. Yves Moreau
Prof. dr. ir. Johan Suykens
Prof. dr. Jesse Davis
Prof. dr. ir. Jan Aerts
Prof. dr. ir. Jos Van Pelt
Prof. dr. ir. Pierre Geurts
(University of Liege, Belgium)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor in Engineering Science

October 2014

© 2014 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Dusan Popovic, Kasteelpark Arenberg 10 postbus 2440, B-3001 Heverlee (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

ISBN 978-94-6018-905-0

D/2014/7515/129

Preface

A proverb says that even the thousand-mile journey begins with the first step. I believe this is how the most of us foresee research when we dive into it for the first time – as a structured, step-by-step endeavor. Well, in my case it was not like that at all. Wandering through these novel and exciting problems in fields of machine learning and bioinformatics was more like trying to dance waltz on acid. Luckily, after all I have finally arrived... well... somewhere. Even better, the path to that "somewhere" turned out to be more interesting and reaching the final destination more fruitful than I initially expected. For this reason solely it seems impossible to mention all the people who helped or, at least, inspired me during this journey. Regardless, I will at least try to do so...

I will start with my mentor, Professor Bart De Moor. He has made sure that I do not face any unnecessary obstacles to my scientific work, including these of financial and administrative nature. He has also trusted me enough to give me all the independence I needed, but at the same time he was always there for me when necessary. Bart, thank you for that!

Besides my supervisor, I feel that I have to mention three other professors from the ESAT (Prof. Yves Moreau, Prof. Johan Suykens and Prof. Jan Aerts) and one from the CS Department (Prof. Jesse Davis) - with whom I have realized several successful joint projects. I am grateful for all constructive discussions and scientific "confrontations" that we had, whose helped me in developing deeper understanding of our field. In addition, I deeply appreciate the voluntary effort that Prof. Davis invested in correcting grammatical errors in this book.

I also want to acknowledge two professors from the University Hospital Leuven, namely Prof. Jos Van Pelt and Prof. Tomas D'Hooghe (together with their co-workers Hannah van Malenstein, Jeroen Dekervel and Alexandra Vodolazkaia), for a very rewarding collaboration. Likewise, I am thankful to the two remaining members of my examination committee, Prof. Pierre Geurts and Prof. Hendrik Van Brussel, for their valuable suggestions for improving this manuscript and

the accompanying presentation, as well as for agreeing to take part in my PhD defense in the first place.

I would like to thank to my current and former colleagues from the STADIUS research group - Nico, Georgios, Charalampos, Yousef, Rocco, Leo, Amin, Peter(s), Raf(s), Anneleen, Marc, Jaak, Toni, Inge, Pooya, Ryo, Sarah, Griet, Gorana, Tunde, Jiqiu, Joana, Daniela - for being the constant source of inspiration and fun, as well as for being a rich "data-base" of expertize. Among them I have to single out Alejandro Sifrim, who was the "biological" brain behind the research project which results are presented in this text. Working with him was truly enjoyable and fulfilling experience! Furthermore, I definitely have to indicate all the technical assistance that my colleague Arnaud provided me with during the final phase of my PhD. In addition, I am thankful to the administrative staff at the STADIUS, in particular to Ida, John, Mimi and Elsy, for the efficient management of all bureaucratic issues that I was facing or creating along the way.

I owe special gratitude to my family – to my father Ilija, mother Rada, brother Milan (with his wife Ljubica and two little devils – Mateja and Dunja), as well as to my wife Ivana, for all their love and support. At the end, I would like to thank to my yet-unborn son Oliver, whose imminent appearance in our lives made me more agile in finishing my PhD.

Thank you all! Thank you for dancing this waltz with me during all these years!

U sećanje na mog pokojnog dedu Milovana, koji bi sada bio jos ponosniji na mene,

Dušan Popović
Leuven, October 2014.

Abstract

Approximately eight percent of total population is affected by one of more than seven thousand identified genetic disorders. Causes of many of these disorders are poorly understood, which complicates disease management and, in some cases, increases morbidity and mortality. At the same time, rapid development of high-throughput technologies in the past few decades gave a considerable boost to the biomarker discovery in general. Among these techniques, the exome sequencing appears to be especially promising approach for identification of novel genes causing inheritable diseases. However, each individual genome typically harbors thousands of mutations, hence detecting the disease-causing ones remains a challenging task, even when the majority of the putatively neutral variation is filtered-out beforehand. Several computational methods have been proposed to assist this process, but most of them do not display satisfactory precision to be used in real-life environment.

We propose a novel, genomic data fusion based method for prioritization of single nucleotide variants that cause rare genetic disorders. It implements several key innovations that resulted in approximately 10-fold increase in the prioritization performance compared to the rest of state-of-the-art. First, it blends together conservation scores, haploinsufficiency and various impact prediction scores, practically subsuming all the other major algorithms. Second, it is the first of its kind to fully exploit phenotype-specific information. Third, it is directly trained to distinguish rare disease-causing from rare neutral variants, instead of using common polymorphisms as a proxy. We also describe several strategies for aggregation of predictions across multiple phenotypes and explore how each of them affects the prioritization under different levels of noise. In addition, we formulate a simplified version of the model to increase the interpretability of the decision-making process, as well as to reduce a storage demand and a computational burden induced by the system. Finally, we identify a bias originating from the hierarchically granular nature of the problem's data domain and develop a sampling-based way to bypass it, which translates to a considerable additional increase of the system's performance.

Beknopte samenvatting

Naar schatting acht procent van de totale bevolking heeft een van zevenduizend geïdentificeerde genetische aandoeningen. De oorzaak van deze aandoeningen is vaak slecht gekend, wat ziektemanagement bemoeilijkt, en in sommige gevallen leidt tot verhoogde morbiditeit en mortaliteit. Tegelijkertijd is de ontdekking van biomarkers in een stroomversnelling gekomen, dank zij de ontwikkeling van hoge-doorvoer technologieën. Hiervan biedt de exoom sequencing technologie een veelbelovende manier om nieuwe genen te identificeren die erfelijke ziektes veroorzaken. Elk individueel genoom bevat echter typisch duizenden mutaties, zodat het achterhalen van welke de oorzaak zijn van ziektes een uitdaging blijft, zelfs wanneer het merendeel van de vermoedelijk neutrale variatie vooraf weggefilterd werd. Verscheidene rekenkundige methodes werden voorgesteld om dit proces te ondersteunen, maar de meeste onder hen zijn onvoldoende precies voor praktisch gebruik.

We stellen een nieuwe methode voor voor de prioritering van enkel-nucleotide varianten die genetische aandoeningen veroorzaken, gebaseerd op genomische data fusie. Deze methode implementeert verschillende belangrijke vernieuwingen die leiden tot een ongeveer tienvoudige verhoging van de prioriteringsperformantie vergeleken met huidig gangbare methodes. Ten eerste vermengt deze methode conservatie, haploinsufficiëntie en verscheidene impact predictie scores, zodat het alle bestaande belangrijkste algoritmes vervangt. Ten tweede is het de eerste methode die fenotype-specifieke informatie ten volle benut. Ten derde is het rechtstreeks getraind om zeldzame ziekte-veroorzakende varianten te onderscheiden van zeldzame neutrale varianten, in plaats van vaak voorkomende polymorfismen te gebruiken als proxy.

Verder beschrijven we verschillende strategieën voor de aggregatie van predicties over meerdere fenotypes en verkennen we hoe deze de prioritering beïnvloeden in de aanwezigheid van verschillende ruisniveaus. Bovendien formuleren we een vereenvoudigde versie van het model om de interpreteerbaarheid van het beslissingsproces te verhogen, alsook de benodigde opslag en de rekenkundige

vereisten te verlagen. Tenslotte identificeren we een bias veroorzaakt door de hiërarchische, granulaire aard van het gegevensdomein en ontwikkelen we een bemonsteringsgebaseerde methode om deze te omzeilen, welke resulteert in een aanzienlijke bijkomende verhoging van de performantie van het systeem.

Abbreviations

1000G	1000 Genomes project database
AUC	area under the curve
CCHS	Colon Cancer Hypoxia Score
ECDF	Empirical Cumulative Distribution Function
FPR	False positive rate
GA	Genetic algorithm
GWAS	Genome-wide association study
HDSS	Health Decision Support System
HGMD	Human Gene Mutation Database
HPO	Human Phenotype Ontology
KNN	k-nearest neighbors
LDA	Linear discriminant analysis
LS-SVM	Least-Squares Support Vector Machine
MAF	minor allele frequency
MCC	Matthews correlation coefficient
NB	Naïve Bayes
NPV	Negative predictive value
nSNV	nonsynonymous single nucleotide variant
OMIM	Online Mendelian Inheritance in Man
OOB	out-of-bag
PPV	Positive predictive value
PR	Precision-Recall
QDA	Quadratic discriminant analysis
RF	Random Forest
RF-FI	Random Forest Feature Importance
ROC	Receiver Operator Characteristics
RRA	Robust Rank Aggregation
SVM	Support Vector Machine
TPR	True positive rate

Contents

Abstract iii

Abbreviations vii

Contents ix

List of Figures xiii

List of Tables xv

1 Introduction 1

1.1 Two views on the in-silico biomarker discovery 5

1.1.1 Examples of biomarker discovery by feature selection . . 7

1.1.2 Examples of biomarker discovery by classification 10

1.1.3 Strengths and weaknesses of two paradigms 11

1.2 Prioritization of the disease-causing mutations 13

1.2.1 Limitations of the existing methods 14

1.2.2 Main research goals addressed in the thesis 16

1.3 Organizational context of the research project 20

1.4 Structure of the thesis text 24

2 Variant prioritization by genomic data fusion 29

2.1 Main . . . 30

2.2 Methods . . . 34

2.2.1 Data generation . . . 34

2.2.2 Classifier benchmarks . . . 35

2.2.3 Control-set benchmarks . . . 37

2.2.4 Temporal stratification analysis . . . 39

2.2.5 Feature importance analysis . . . 41

2.3 Supplementary material . . . 42

2.3.1 Supplementary Note 1 – Data generation . . . 42

2.3.2 Supplementary Note 2 - Performance measures and their interpretation . . . 44

2.3.3 Supplementary figures and tables . . . 46

3 Aggregation of prioritization scores across phenotypes 55

3.1 Introduction . . . 56

3.2 Material and Methods . . . 58

3.2.1 Benchmark . . . 58

3.2.2 Non-parametric Order Statistics . . . 59

3.2.3 Parametric Modeling . . . 60

3.3 Results . . . 61

3.4 Discussion . . . 63

4 Interpretation of the model 67

4.1 Introduction . . . 68

4.2 Materials and methods . . . 69

4.2.1 The method . . . 69

4.2.2 The data . . . 72

4.3 Results and discussion 73
4.4 Concludions 76

5 Model improvements **77**

5.1 Background 78
5.2 Methods 81
5.2.1 Hierarchical sampling 81
5.2.2 Experiments with synthetic data 83
5.2.3 Experiments with eXtasy data 85
5.3 Results and Discussion 86
5.4 Conclusions 91

6 Conclusions **93**

Bibliography **99**

Curriculum vitae **113**

List of publications **115**

List of Figures

1.1 The evolution of the whole-genome sequencing price through years from 2001 to 2014 2

1.2 An illustration of feature selection and classification based biomarker discovery paradigms 6

1.3 Distribution of deleteriousness prediction methods scores for the two classes of mutations 15

1.4 An illustration of the difference in the training procedure between eXtasy and the other mutation prioritization methods 17

1.5 Organizational context of the thesis 22

1.6 Structure of the thesis text 27

2.1 Classification scenarios performance 32

2.2 Temporal stratification of performance 33

2.3 Feature importance measures 34

2.4 Workflow of the classifier benchmark 36

2.5 Receiver Operating Curves (ROC) of classifier performance on benchmark data 38

2.6 Precision-Recall curves of classifier performance on benchmark data 38

2.7 Creating data sets for testing classification scenarios 39

2.8 Stratification of data for temporal analysis 40

2.9 Feature distributions 50

3.1 Schematic representation of the benchmarking scheme 59

3.2 Empirical Cumulative Distribution Function (ECDF) of ranks comparing different aggregation schemes 62

3.3 Empirical Cumulative Distribution Function (ECDF) of ranks by fraction of informative phenotypes comparing different aggregation schemes 63

4.1 Sequential algorithm for training the simplified eXtasy 70

4.2 An example of two different initial decision tree splits for the same classification problem 71

4.3 ROC curves obtained by applying the three classifiers : standard eXtasy (Random Forest), simplified eXtasy and the decision tree built on the eXtasy data 74

4.4 Decision tree constructed on the data generated by Random Forest 75

5.1 A part of the eXtasy data 79

5.2 Features of the synthetic data. 84

5.3 Results of the first experiment on the synthetic data. 87

5.4 Results of the second experiment on the synthetic data. 88

5.5 PR curves from eXtasy-based experiment. 89

List of Tables

2.1 Performance of the classifiers on the benchmark data 51

2.2 Performance comparison of different control sets 51

2.3 Number of disease-causing variants and corresponding genes by year 52

2.4 Performance comparison by year of discovery 52

2.5 Classifier implementations and parameters 53

2.6 Descriptions of the Endevour data sources 54

3.1 Summary statistics for the overall comparison of different aggregation methods 62

3.2 Summary statistics for the comparison of different aggregation schemes by fraction of informative phenotypes 64

4.1 Average values of six performance measures obtained by testing the three methods 75

5.1 Results of the eXtasy-based benchmark. 90

Chapter 1

Introduction

The past few decades have been marked by the unprecedented progress of biomedical technologies. The scale and speed with which these can measure biological phenomena have completely transformed the landscape of biomedical research. High-throughput techniques have become an irreplaceable tool for studying complex [1, 2] and heritable diseases [3]. Meanwhile, the spectrum of their applications has slowly extended to other, more practically-oriented domains, such as molecular diagnostics [4, 5] and drug discovery [6, 7].

Consequently, the price of high-throughput experiments has dropped significantly, allowing many small research centers to participate in the race. A notable example of this is the rapid decline of the cost for sequencing the whole genome. In 2001, it cost one hundred million dollars and in 2014 it finally reached "the holy grail" of genomics technology - the one thousand dollar genome [8].This drop in price has tremendously outpaced the famous Moore's Law (see Figure 1.1).

The reduced costs have naturally led to a phenomenon often described as a *tsunami of data*. In the biomedical domain this not only reflects the shear volume of data resulting from massively parallelised experiments, but also the inherent complexity of it. For example, next-generation sequencing technologies associate up to three billion features (base-pairs) with a *single sample*. Mass spectrometry imaging is even a more extreme instance of this effect [9], as it routinely produces one mass spectrum for each cell of a dense, two-dimensional spatial grid and this for each slice of a tissue.

Thus today the focus is shifting from the question of how to generate the data to the problem of what to do with all these data. Some authors even argue that

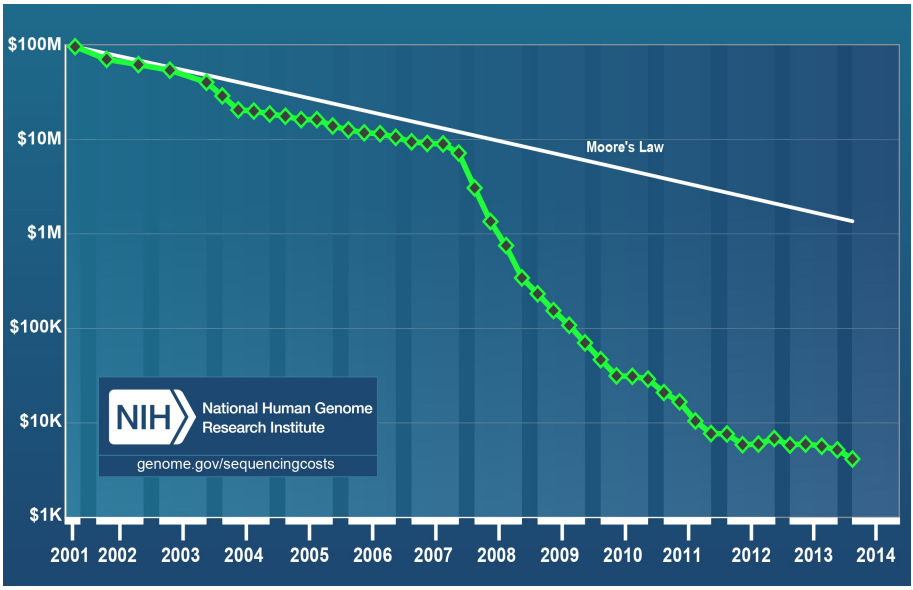


Figure 1.1: **The evolution of the whole-genome sequencing price through years from 2001 to 2014.** Source : National Human Genome Research Institute [10]

the development of analytical tools has lagged way behind that of hardware, so the price of *analysis* per sample has become the financial bottleneck of research pipelines that rely on high-throughput technologies [11]. Sharing this position or not, most of the biomedical researchers would still agree that these amounts of data can not be processed efficiently using standard quantitative approaches that predate the -omics era. Clearly, a constant improvement of analytics is necessary to get the best out of the rapidly developing hardware technology.

One might wonder what makes biomedical data so unique and difficult to analyze? First, it can rarely be used in its raw form, thus some sort of preprocessing has to be performed in almost every type of analysis. For instance, the data resulting from a sequencing experiment have to be aligned to the reference genome before their use [12]. Similarly, microarray data in the form of raw intensities has to be normalized to be amenable for study [13]. Hence, the preprocessing step is always technology dependent and it has a tremendous impact on the final result.

Second, many of the high-throughput technologies produce data characterized by high levels of technical noise due to the variable quality in the measurements [14]. Third, there is typically a large number of features in contrast to the

usually small number of samples available in a single study. This renders these data unsuitable for the standard statistical apparatus. Finally, the diversity of the technological platforms together with the frequent batch effects and substantial systematic differences poses a serious challenge in terms of data integration.

These enumerated issues eventually led to the birth of the field of bioinformatics as a separate discipline that is based on a crossroad of biology, databases, optimization, statistics and data mining. Also, many techniques in the field were borrowed from machine learning where these types of non-standard problems have been tackled already for quite some time. For instance, the methodology of Support Vector Machnines (SVM) [15, 16] has become one of the most popular classification paradigms within the bioinformatics community due to its ability to handle data of extremely high dimensionality. In addition, many new machine learning methods were developed in response to the needs of biomedical research. For example, the recent improvements [17, 18, 19] of the random forest feature importance measures [20] are mostly motivated by particular applications in bioinformatics.

A large portion of current bioinformatics efforts is devoted to facilitating the discovery of biomarkers, i.e., of "characteristics which are objectively measured and evaluated as indicators of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" [21, 22]. Biomarkers can come in various forms - metabolites, lipides, proteins, RNA, mRNA, genes and such - whose altered state (ex. concentration, expression, mutation) is related to the clinical outcome or the biological process of interest [23, 24]. Most of the biomarker studies are now done *in-silico*, meaning that they utilize data available from previously conducted experiments, executed by means of computer *simulation*. This approach offers a great potential to increase the discovery rate by improving the efficiency of biomedical research.

However, *in-silico* biomarker discovery also has a few serious shortcomings, including the high heterogeneity of patient cohorts, limited clinical utility and low reliability of public data sources, just to name a few [25]. But perhaps the biggest problem with this methodology is its observational nature. By not relying on designed and controlled experiments, *in-silico* biomarker mining can aim at best on the characterization of correlation structure that is present in the data, without providing *any* evidence of causal relationships.

For these reasons, *in-silico* biomarker mining can not be treated as a sole lane in the development of clinically applicable diagnostic tools, but rather as part of a pipeline that also includes external validation by means of a controlled clinical trial. In a wider sense, the *in-silico* biomarker discovery methodology can never provide definite mechanistic explanations for biological processes

under examination, so it should be treated more as an exploratory than an explanatory research device. That is, maximal caution should be exercised during the interpretation of its outcomes. Nevertheless, this approach is still a valuable resource for quick and relatively inexpensive prototyping of biomarker panels by reducing the total number of targets for confirmatory experiments.

Having this perspective in mind, we can conclude that a very important component of any computational method that is used for *in-silico* biomarker discovery, is its ability to *prioritize* targets with respect to their *likelihood* of being involved in certain biological process. If this requirement is fulfilled, the external validation can start from the top of a prioritized list containing biological indicators, saving time and resources for the experimentation.

This work focuses on one particular facet of biomarker discovery research, namely the computational prioritization of mutations that cause rare genetic disorders. From the biomedical standpoint, this is important problem as a significant proportion of general population (i.e., approximately 8% [26]) suffers from at least one of these inheritable disorders. At the same time, each rare genetic disease typically affects a relatively small number of individuals, which renders characterizing its triggers extremely difficult. That is, it is often very hard to collect a sufficient number of samples for conducting large-scale studies.

The main role of computational variant prioritization in the discovery of disease-causing mutations is the reduction of the total number of targets for confirmatory experiments. To fulfill this role and, consequently, to be actually *used* in practice, any algorithm that is proposed in this context has to display sufficient performance. Although several promising methods have been developed to address this task, many of them still display limited utility in real-life usage scenarios, as we shall show later. Our main aim is to significantly advance the current state of the art in computational mutation prioritization, such that progress in discovering causes of rare genetic disorders could be accelerated.

To fulfill this aim, we have formulated a novel disease-causing variant prioritization method that displays considerable improvement over the existing approaches in terms of the practical performance. The algorithm is based on the genomic data fusion and it is first of its kind to directly utilize information on phenotypic representations of genetic disorders. We have also proposed several algorithmic modifications of the standard machine-learning methods in order to tackle additional computational sub-problems that typically arise in this domain. Finally, we have proven the utility of various data sources (including disease-specific ones) for the mutation prioritization task.

In the following section (1.1), we treat the general computational biomarker mining problem in greater detail. We define two global approaches for tackling

the problem, discuss their relative strengths and weaknesses (subsection 1.1.3), and illustrate this duality on a number of studies that we have conducted in the field (subsections 1.1.1 and 1.1.2). These studies have not been included in the main thesis text as they do not specifically address mutation prioritization, yet they are important for placing the contributions of this research project in the wider context of our previous work. Also, they serve as a justification of our choice of the biomarker mining paradigm that we follow in the development of our mutation prioritization methodology. For the complete list of corresponding bibliographical references, we refer the interested reader to the section "List of publications" that is available at the end of this manuscript.

In section 1.2, we further focus on the problem of prioritizing disease-causing variants. We discuss present approaches together with their limitations (subsection 1.2.1), and show how these limitation translate to the main research questions that have been addressed during development of our methodological framework (subsection 1.2.2). We also briefly introduce the approaches that have been undertaken for solving each of defined sub-problems.

In section 1.3, we describe the placement of the project within the organizational context in which it has been conducted, and discuss how it relates to the overall strategic goals of the research group. In addition, the major deliverables of the completed work are summarized. Furthermore, the most important personal contributions of this author are enumerated in detail, together with full bibliographical references. Finally, in section 1.4, we outline the structure of the thesis in relation to the main research questions that it addresses.

1.1 Two views on the in-silico biomarker discovery

With no ambition to exhaustively categorize all available developments in the field, we think that most of the computational biomarker mining techniques can, in essence, be reduced to instances of two general approaches, which we will refer to as *feature selection* and *classification*. In this context, *feature selection* would be any method that utilizes information about the *outcome* to select or to prioritize biological indicators of interest. For example, measuring differential expression of genes between two conditions (ex. disease vs. healthy) with the aim to identify those involved in the process under study.

Note that the notion of outcome can be sometimes implicit. One example is the identification of mutations that are common to the genomes of several people suffering from the same genetic disorder, as is often done in GWAS (Genome Wide Association Study) studies. The outcome is not directly used in those

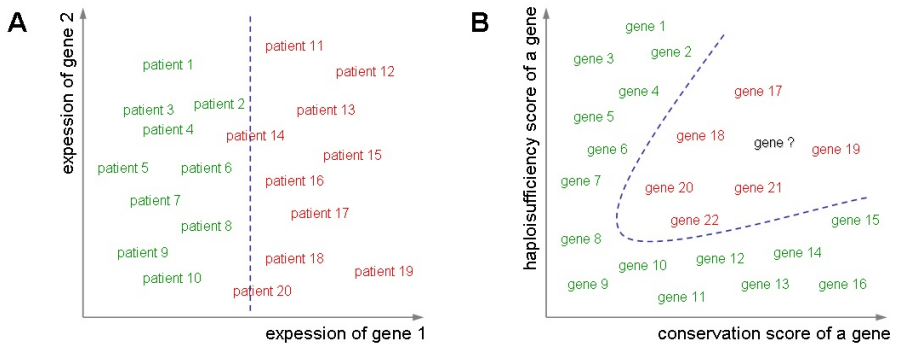


Figure 1.2: **An illustration of the feature selection and classification based biomarker discovery paradigms.** A) The expression of gene 1 is more predictive than that of the gene 2, so under the feature selection approach it would be considered better candidate for a biomarker. B) Two meta-features are used to classify genes into two groups. The tested gene is classified as being associated with the "red outcome", so it can be used as biomarker for whatever biological state "red outcome" represents (ex. presence of a disease), together with genes 17-22. Additionally, it probably also plays a role in the underlying process causing the "red outcome" itself as it is very similar to the known genes involved in it.

studies, but it is known and it guides the search, i.e., all the people in a group have a disease and the goal is to pinpoint a mutual factor.

In contrast, *classification*-based methods are those which do not treat biomarkers as predictors, but rather as instances, i.e., the *subjects* of the prediction. Thus, biomarkers are *classified* or, more narrowly speaking, *prioritized* with respect to the likelihood of being associated with a certain biological state such as a disease or some other bodily process. The features used in classification are usually not directly related to the indicators individually, but instead they capture some of their general characteristics. For example, when the mutations are prioritized with regard to their potential to disrupt a function, one can use their structural or evolutionary properties to assess it.

Alternatively, the information captured by features can also come in the form of similarities to biological agents known to be involved in a certain process. For instance, if the goal is to rank (i.e., *to prioritize*) a group of genes according to their likelihood of being implicated in a disease, sequence similarity to the known disease genes can be used. The difference between the two introduced paradigms is illustrated in Figure 1.2. We now further delineate it on a number of studies that we conducted in the context of biomarker discovery, starting from the feature selection-based cases.

1.1.1 Examples of biomarker discovery by feature selection

In Dekervel *et al.* [27] we aimed to prove the link between the differential expression of certain genes and the recurrence of colorectal cancer and to isolate a relatively small subset of these genes that can be used for survival prediction. We started from the hypothesis that genes that are functionally related to the cancer under study will be also differentially expressed under chronic hypoxia. Thus, initially a panel of 923 genes has been experimentally identified as differentially expressed under hypoxia. In the next step, several publicly available microarray data sets on colorectal cancer have been used to further reduce this list.

In particular, only the genes having high corresponding Z-values of Goeman's global test [28] on the three data sets [29, 30, 31] simultaneously and whose direction of expression matched that observed in-vivo were retained for further modeling. The fourth data set (also from [31]) has been subsequently used to optimize this genetic signature by a backward regression procedure. When applied on an external cohort of 90 patients, the resulting six gene Colon Cancer Hypoxia Score (CCHS) was found to be an independent prognostic factor for disease-free survival time.

In Vodolazkaia *et al.* [32] we addressed the problem of non-invasive diagnostics of endometriosis. At present, the only way of establishing a conclusive diagnosis for this disease is by laparoscopic inspection, so in many patients endometriosis remains undetected for several years after its initial onset. Hence, morbidity associated with the disease could be greatly reduced if a screening tool based on a blood sample was available. In this study, we utilized a univariate method and a two step-wise logistic regression based approaches to select biomarkers that display sufficient discriminative power to be used for diagnostics, from a panel of 28 plasma proteins.

All feature selection schemes were repeatedly applied on random subsets extracted from the training set, keeping track of biomarkers that were found to be significantly related to the outcome. When the Mann-Whitney test [33] was used (i.e., the univariate approach), we retained in the final model only these proteins for which the difference in their blood concentrations between healthy and diseased individuals was statistically significant in more than 70% of the randomizations. Similarly, in the first multivariate case, variables present in more than 70% of the fitted regression models were considered. In the second approach, all models *containing* biomarkers found by the first multivariate approach have been selected, after which all features figuring in the best among these were deemed informative. Subsequently, logistic regression and Least-Squares Support Vector Machine (LS-SVM [34]) classifiers were trained on

the two feature panels containing four predictors each (Namely, in Model 1 - annexin V, VEGF, CA-125 and glycodelin; in Model 2 - annexin V, VEGF, CA-125 and sICAM-1), and applied on the independent test set displaying with a high sensitivity (82%) and an acceptable specificity (63–75%).

The feature selection approach utilized in Dekervel *et al.* does not interact with a classifier, thus it is basically a *filter* [35]. In addition, in both the studies discussed so far, the methods used are in essence based on a greedy search, hence failing to fully account for correlations that are typically present among features. At the same time, measurements of biological processes (i.e., features of biomedical data) usually exhibit highly complex correlation structure. However, finding the optimal set of predictors while this structure is completely accounted for, requires exhaustive testing of all possible combinations of features, i.e., it is an NP-hard problem [36, 37]. This type of optimization problems is often tackled using various meta-heuristics, such as genetic algorithms, simulated annealing, particle swarm optimization, direct search, memetic algorithms and others. They often produce a near optimal solution in a fraction of the time that is necessary for the full search. Among them, genetic algorithms are the most suitable for the feature selection task, as their binary representation directly matches the binary nature of the problem.

Hence, we proposed a solution based on a simple genetic algorithm (GA) and demonstrated it on the colon cancer recurrence prediction task, achieving better performance than that reported in the literature for the same problem [38]. In essence, the method optimizes the choice of features to obtain the maximal accuracy on a validation set with the minimal possible number of predictors. Inclusion or exclusion of each feature from model have been encoded by a binary variable that corresponds to a gene in a genetic algorithm, while each solution (i.e., panel of selected features) have been represented as a chromosome. The fitness function has been designed as a combination of balanced accuracy and size penalty term. This term depends on a relative increase in the number of selected features compared to the average length of the solution in the initial generation.

Due to the stochastic nature of genetic algorithms, each run of the aforementioned procedure produces a slightly different selection of genes. Therefore, to construct the final genetic signature, the results of many iterations have been harvested and the total gene counts were modeled using a negative binomial distribution. We have assumed that genes that are truly informative for predicting the colon cancer would be selected significantly more times than would be expected under this distribution, thus we used a threshold of 1% ($p=0.01$) for the final selection. We later successfully utilized a similar approach that also incorporates the tuning of the support vector machine parameters for predicting the occurrence of pancreatic cancer [39].

However, genetic algorithms are characterized by several parameters that, in general, have to be tuned for each new type of problem. So, in order to make the method more widely applicable, we further generalized it by establishing mechanisms for self-adaptation [40]. In particular, mutation and crossover rates are represented as additional chromosomes in a single individual, and are thus subjected to the evolutionary pressure as well. The intuition behind this approach is that those individuals having a too low or a too high affinity for some operator will be ultimately suppressed by the other, more flexible, solutions or destroyed prematurely in the process.

Additionally, we based the algorithm on the island model (several interacting sub-populations) with a dynamic control of migration. The rate of migration has been formulated as an increasing function of a number of expired generations. This helps in preventing the homogenization of populations across islands during early phases of the algorithm's execution. The described setup resulted in a slightly better performance than the one obtained with the problem-tailored genetic algorithm, regardless of the classifier used in the subsequent classification step.

The genetic algorithm approach can be also combined with other multivariate feature selection methods. In Popovic *et al.* [19, 41] we have used it in conjunction with the Random Forest Feature Importance (RF-FI) measures [20], aiming to mitigate a bias that typically arises when these measures are estimated from data characterized by the high correlation among the features. Specifically, the importance of a group of inter-correlated features tends to be systematically underestimated by the RF-FI [42], so the truly informative predictors might be ranked lower than their less informative, but uncorrelated, counterparts.

A genetic algorithm helps in decoupling the dependencies between the importance estimates of discriminative correlated predictors by constructing and evaluating smaller, yet still informative variable subsets. In particular, the algorithm tries to find a compact subset of features that yields the highest out-of-bag balanced accuracy of the Random forest classifier for a given classification problem, while keeping track of the RF-FI measures for each generated solution. During the execution of the algorithm, these importance measures are added to the total counts of the corresponding features that are subsequently used for the final ranking of the predictors.

The intuition behind this approach is that GA tends to eliminate bad solutions, thus the features figuring in these solutions would receive small final scores. In parallel, correlated informative features are rarely a part of the same solution due to the explicitly encoded (i.e., by fitness function) parsimony pressure, so their importance can be estimated more reliably. The study [19, 41] demonstrated

that this strategy yields a significant improvement over the standard RF-FI method.

1.1.2 Examples of biomarker discovery by classification

The first example of a classification-based biomarker discovery that we present here is the study by Iacucci *et al.* [43], which tackled the problem of protein-ligant pairing prediction. In this case we have defined the data instances as *pairs* of proteins which may be assigned to one of two classes, the interacting and the non-interacting class. All proteins have associated measurements from various data sources such as expression profiles or sequence information, hence the association between the measurements of two proteins can be represented by similarity. These similarities were used as predictors, implicitly assuming that proteins that interact also share many other common properties. The method displayed a competitive classification performance (sensitivity of 0.87 and a specificity of 0.79), especially for certain protein families (ex. GPCR family with 0.96 of the balanced accuracy).

The second example of a study where we used the classification approach, is one addressing the problem of the epigenetic modification prediction [44], where the utility of various derived features in detecting modified regions was assessed. We propose two novel features - the distance from the nearest transcription site on the one hand and the histone coverage that is computed from the size of the affected region on the other hand. We have included these together with the current standard (i.e., the enrichment score obtained directly from a chip-chip experiment) to the classification model. We have shown that using these additional features improves the prediction results compared to the case where only the enrichment score is present. In addition, the histone coverage turned out to be more informative then the widely established enrichment score.

Finally, in [45] we have tried to combine *feature selection* and *classification* approaches to biomarker discovery. We have proposed a modification of the simple genetic algorithm (previously presented in [38]) which integrates the prior knowledge on gene interactions in the form of an additional genetic operator. Specifically, this new operator is triggered after the crossover and the mutation, and it activates genes that interact with genes that are already present in a solution. Probability of including an additional gene to a solution is defined in relation to its interaction "strength", i.e., to candidate gene's likelihood of being somehow related to one or more of previously selected genes. In this work we have used String 9.1 network [46] weights as a substitute for the interaction "strength", but the same idea can be easily generalized to any gene-gene similarity

matrix. The described modification led to a performance improvement over simple genetic algorithm.

1.1.3 Strengths and weaknesses of two paradigms

Both of the discussed *in-silico* biomarker mining paradigms have certain distinctive strengths, yet these come with the price of associated caveats. The feature selection-based approaches can more or less successfully identify predictors that are correlated with the outcome, but being purely observational they can not possibly reveal causal relations. For example, if the expression of a gene is indicative of a disease status, that does not necessary mean that this gene is actually involved in the development of that disease. It might be that it resides downstream on the pathway that is triggered by the genuine disease-causing gene and that it only amplifies its signal.

In the case of studies with a small number of examples and high dimensionality (ex. microarray analysis), this may be further complicated by the existence of many spurious correlations, to the point that even randomly selected predictors may seem informative [47]. However, the feature selection approach is still a valuable tool for obtaining *diagnostic* biomarkers, as a feature can be predictive without being a determinant of the biological process. In addition, it might be the only way of obtaining extraordinary or unsuspected findings, for the reasons we will exemplify right away.

In parallel, the classification-based approaches always rely on the notion of similarity, regardless on how implicit it might be. This implies that they can not be used for discovering something that is far away from the existing examples or outside of the training data domain, i.e., something *very different* from what we already knew. This is in fact a corollary of one of the basic machine learning assumptions - the equality of the training and testing data distributions [48, 49]. Still, this is an important philosophical aspect of data analysis that is often neglected in the field of bioinformatics and that has to be taken into account for a proper interpretation of the findings.

Think of predicting a gene function based on the data derived from article abstracts as an example. One might construct a predictive score by calculating similarity between vectors (ex. bag-of-words representation) corresponding to the articles mentioning the query gene and vectors corresponding to the articles describing genes proven to have the function under consideration. If a candidate gene has been recently discovered, and consequently described in just one or very few studies focusing on aspects different from gene function, its similarity to the known genes would be artificially small. Thus, a learning algorithm

based on text-mining features constructed in the described way would be biased toward *already well-characterized* genes.

Similarly, if some structural properties of candidate genes are compared with those of known cancer-related genes in order to identify ones that are potentially implicated in the disease, genes that trigger the previously undescribed mechanism of oncogenesis might stay undetected, simply for not sharing the same properties with genes involved in previously discovered processes of cancer development. Conversely, under the feature selection-based biomarker mining paradigm, these genes might still be found. For example, their expression levels can be altered in affected individuals compared to those in healthy subjects.

Hence, with *classification-based* biomarker mining one can hope for incremental advances at best, i.e., for these that already reside somewhere within the information *domain*. In other words, findings that considerably diverge from the current corpus of knowledge in principle can not be detected through mere similarity to already known facts. In these situations, *feature selection* approaches are clearly superior, as they rely on the existence of *examples* whose characteristics can go beyond the information domain. Thus, from the feature selection paradigm perspective, *similarity* to known concepts can even work against discovery.

However, the *classification* approach is more appropriate than the feature selection approach in situations where there are many candidate biomarkers and only a few specimens of a biological process under study, as is the case with the prioritization of rare disease-causing mutations. By defining biological indicators as *instances* rather than *dimensions* one can escape the curse of dimensionality in these situations. Naturally, meta-features that are informative enough in the context of a biological concept that is studied should be available or possible to construct.

An additional advantage of the *classification* paradigm in the context of prioritization is the straightforward way of obtaining *scores*. That is, scores are often routinely assigned during classification, as a by-product of the process. For example, if a logistic regression model is used to classify mutations as benign or disease causing, it would typically output the class probability together with a hard-thresholded decision. This class-probability can be used to rank mutations, such that the most probably disease-causing ones can be analyzed first. In contrast, a number of feature selection-based methods (ex. genetic algorithms) only provide a closed set of features as an outcome, which limits their utility for prioritization.

Finally, by using the meta-features, *classification-based* biomarker mining is less susceptible to batch effects. Usually, meta-features are defined in a way that

does not rely on a particular cohort of patients. Take any mutation conservation score for example. A conservation score can be used as a feature when predicting if a certain mutation is damaging or not, under the assumption that damaging variants will be, in principle, selected against during the evolution. A value of this score does not depend on the presence of a given mutation in any individual patient, but on the presence of that mutation in different species. At the same time, if the *feature-selection* paradigm is followed when assessing the damaging potential of the same mutation, one would typically somehow compare the group of affected individuals to the group of healthy subjects. The result of such a comparison is usually strongly dependent on the characteristics of the given groups.

1.2 Prioritization of the disease-causing mutations

The advent of DNA sequencing technologies has tremendously accelerated the rate of discovery of disease-causing mutations. However, each genome typically contains thousands of neutral mutations, i.e., those having no observable effect on the organism's fitness. In fact, benign mutations account for most of the variation in the genetic makeup between individuals. This variability poses a large problem for *feature selection* based computational biomarker discovery methods, especially in the case of studies focusing on rare monogenic disorders. Typically only a handful of whole-genome samples are available in the context of a single study, with each of them harboring numerous mutations. Even after selecting only nonsynonymous single nucleotide variants (nSNVs) and loss-of-function mutations that are not present in healthy populations, several thousand candidate variants remain, with many of them common to all of the affected individuals. Thus, finding the causative ones under the feature selection paradigm often turns out to be a futile exercise.

It seems that this is the reason why many of the state-of-the-art methods are classification-based. They often use certain external characteristics of the mutations to predict their relation to a disease, including structural, evolutionary and biochemical properties. For example, a popular deleteriousness prediction method SIFT [50] relies on the assumption that changes at well-conserved positions are more likely to be damaging, as amino acids that are important for the protein functioning usually reside in these. PolyPhen-2 [51] also uses several sequence and structure-based features to predict if a mutation is damaging or not.

1.2.1 Limitations of the existing methods

Most of the state-of-the-art mutation prioritization algorithms suffer from two major issues when facing a practical application, that both originate from an initial inadequate evaluation. Firstly, these methods are usually trained and benchmarked on the data sets composed of disease-causing variants and common polymorphisms. That makes them less sensitive to the difference between rare benign and rare disease-causing mutations. We even hypothesize that some of them actually (implicitly) model only the difference between rare and common variants, ignoring information related to disease status altogether. In other words, they classify most of the rare mutations, including benign ones, as disease-causing only because they are rare (and thus also more likely to be related to disease).

We do not really go deeply into the quantification or even proof of this effect, but distributions of scores of several methods for different types of mutations are strongly suggestive. Figure 1.3 illustrates this issue in PolyPhen-2 and SIFT, while Figure 2.9 exposes it on the wider group of methods. In particular, it is apparent in Figure 1.3 that SIFT assigns very high scores to almost all rare benign mutations, implying that they might be associated with a disease. Conversely, only about a half of common polymorphisms receive high scores from SIFT. Similarly, PolyPhen-2 scores a bigger proportion of rare benign variants high, than is the case with common polymorphisms. That is, approximately one third of rare mutations in contrast to one tenth of common variants occupy the right end of PolyPhen-2 scores distribution.

Second, many of these algorithms in fact do not have satisfactory performance for useful prioritization under a real-life class distribution. This issue is obscured by the utilization of fairly balanced benchmark data sets in corresponding studies. For example, estimates of precision obtained from data where the class balance differs from that typically encountered in the wild are not valid or even indicative of performance that can be obtained in a practical application. In other words, as precision strongly depends on the class balance, its estimate might be artificially high when it comes from a typical benchmarking data set.

The problem of realistic performance estimation is further complicated by the fact that the studies describing prioritization algorithms rarely (if ever) directly report on the performance metrics of major interest for the prioritization task, namely the area under the ROC curve, precision-recall curve, or precision/sensitivity pair in conjunction with the prevalence. In general, the principles of a proper evaluation of learning algorithms [52] are not strictly followed in those.

For example, the PolyPhen-2 study [51] utilizes two data sets to train and

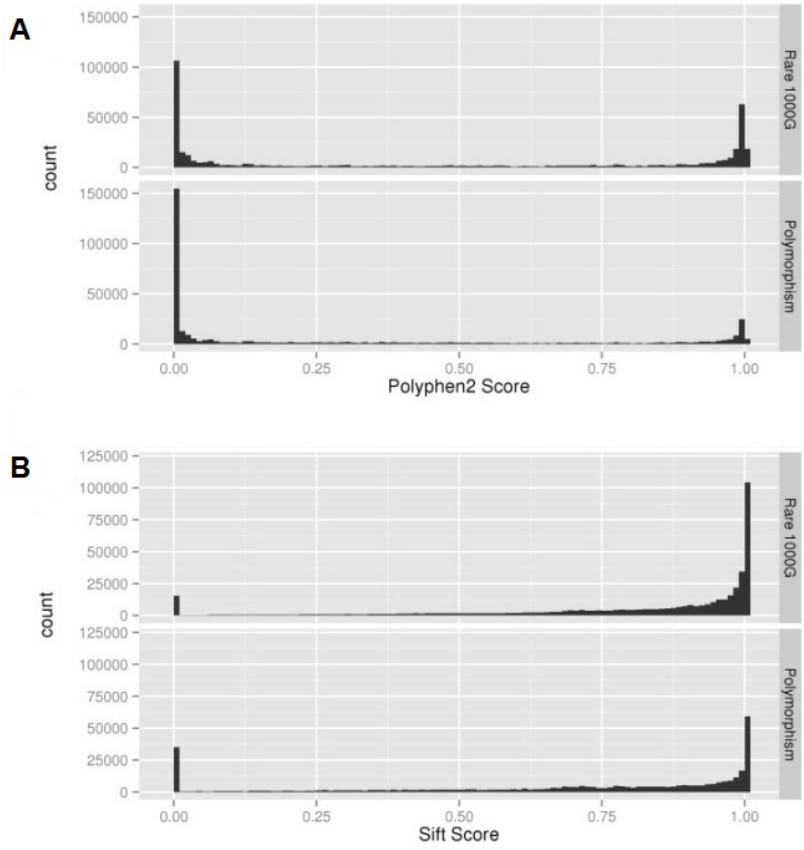


Figure 1.3: **Distribution of deleteriousness prediction methods scores for the two classes of mutations.** The distributions of Polyphen-2 (panel A) and SIFT (panel B) scores for common polymorphisms and rare benign variants obtained from the 1000Genome project. Both methods consistently assign high scores to a big proportion of the rare benign variants, implying that these are likely to be disease-causing (which increases the number of false calls). Conversely, the proportion of common polymorphisms that are scored high is considerably smaller.

validate the model, where the first one contains 3155 damaging and 6321 benign mutations. The second one consists of 13032 damaging and 8946 benign variants. The decision threshold of the method yields true positive rates (TPR) of 78% and 71% accompanied by false positive rates (FPR) of 10% and 19% on the two data sets, respectively. It is easy to calculate that under a realistic scenario, where just one out of nine thousands non-synonymous mutations is *possibly* damaging [53], these false-positive rates translate to precisions of approximately

0.09% and 0.04% for the given levels of the sensitivity (i.e., TPR). This implies that, under the real-world class distribution, PolyPhen-2 classifies 900 or 1710 variants as positive (depending on which test set is used), among which only one is really disease-causing with the associated probabilities of 78% and 71%.

Similarly, authors of SIFT [50] report 20% of false positive errors, which means that just one variant out of 1800 called positive might be in fact correctly identified (with 69% probability). The same caveat persists in a number of studies focusing on the other mutation prioritization methods, including the popular CAROL [54], MutationTaster [55] and LRT [56] prediction scores. These figures are clearly insufficient, as they suggest that in practice hundreds of confirmatory experiments must be performed in order to find the one genuinely disease-causing mutation.

1.2.2 Main research goals addressed in the thesis

In the previous subsection we have discussed a number of issues that limit the utility of the current mutation prioritization algorithms in realistic usage scenarios. Therefore, the central goal of our work has been to develop a computational variant prioritization method whose performance would be significantly better than the rest of state-of-the-art, and to prove that this result would still hold in practical applications. In short, we aim to make a system that is *good enough* for real-life use, which is anything but trivial, as we saw from the previous examples. In our opinion, there are three main aspects of this goal that had to be taken into account for its successful realization.

First, we stand on a viewpoint that the validation schemes and evaluation metrics that have been used during the assessment of current methods are inadequate, and that better approaches to evaluation can be designed for this purpose. These approaches should rely on performance aspects that facilitate insight into potential behavior of algorithms under the intended *realistic* usage scenarios, and different methods should be compared on these grounds.

Second, we hypothesize that combining various genomic data sources can dramatically increase performance of mutation prioritization. Specifically, we believe that phenotypic representations of rare genetic disorders convey important information, and that incorporating this information into computational method can result in significant improvement over the standard approaches. Third, we think that right choice of algorithmic strategies and extensive customization can give an additional edge to the method. Later in this chapter, we further decompose these claims to the concrete research questions that we have addressed to fulfill defined main goals.

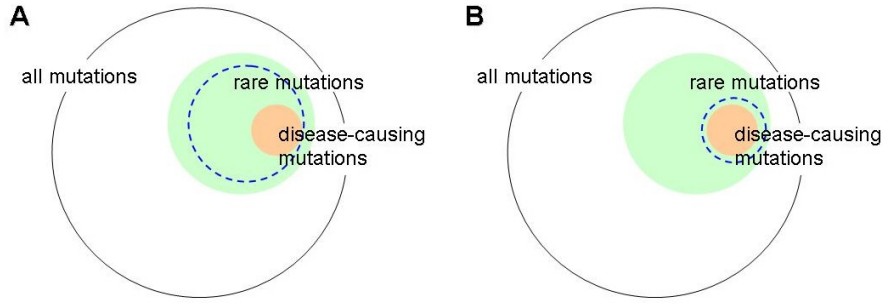


Figure 1.4: **An illustration of the difference in the training procedure between the eXtasy and the other mutation prioritization methods.** Each circle represents the space of all mutations, while the white surface within the circle covers common polymorphisms. Orange and green shaded areas represent rare benign and rare disease-causing variants, respectively. The dotted blue line denotes a hypothesized classification boundary. A) When trained on the data composed of the common polymorphisms and the disease-causing variants a method can not always delineate the rare disease-causing from the rare benign mutations. This is the typical training and validation scheme used for many state-of-the-art prediction scores. B) eXtasy has been trained using only a subset of the all benign variants (the rare ones), which renders its decision boundary more conservative.

As a result, we proposed a novel prioritization method for detecting disease-causing variants, named eXtasy [57], and released the resulting actionable algorithm as an open access software platform. The system displays **ten-fold increase of precision** compared to the competing methods, without degraded sensitivity. This translates to, in average, ten times higher ranks of tested disease-causing variants, and consequently, ten-fold reduction in number of confirmatory experiments to be preformed on a prioritized mutation list.

eXtasy implements the three key performance-boosting innovations. Firstly, it is directly trained to distinguish rare disease-causing from rare neutral variants, which implicitly extends to the common polymorphisms as well (see Figure 1.4). Secondly, it integrates phenotype-specific information in the form of similarities to known disease genes [58] by fusing heterogeneous genomic data. Thirdly, it also incorporates several additional information sources, including conservation and haploinsufficiency scores together with various state-of-the-art impact predictions scores discussed before. A full treatment of the method with all the necessary details is provided in Chaper 2.

As we have briefly outlined in the previous paragraph, an additional advantage of eXtasy over other methods comes from the fact that it scores mutations not only based on their potential functional impact but also based on their relevance

in the light of available information on a disease. As genetic disorders typically initiate several distinctive phenotypic presentations, the system generates a separate prioritization score for each of them given a single mutation. This leads to the problem of *rank aggregation*. That is, the process of constructing a global score for a mutation is severely complicated by the varying reliability of the assigned phenotypes, their changing number and amount of available information that is linked to the each of them.

In the follow-up study [59] we compare several aggregation methods of increasing complexity to address this issue, including mean, maximum, two types of non-parametric order statistics and parametric modeling. We show that eXtasy is quite robust with respect to the choice of aggregation scheme, even the selection of a fully optimal method depends strongly on the fraction of uninformative phenotypes. However, we also show that if a sufficient prior knowledge on phenotypes is available, it can be easily exploited for boosting overall performance of eXtasy. For example, if the most of phenotypes are informative (i.e., more than two thirds of their total number), aggregation by parametric modeling leads to in average **5.5 times higher rank** of a disease-causing mutation compared to the default aggregation by taking maximal obtained score across phenotypes. For this reason, we provide general guidelines for selecting the appropriate technique based on knowledge of the phenotypes under study. The problem of aggregating eXtasy scores is further detailed in Chapter 3.

One issue associated with eXtasy is the low interpretability accompanied by a relatively prohibitive size and time complexity of its core classification algorithm. The system is based on an instance of the Random Forest classifier [20] that consists of thousand trees with average depth expressed in tens of thousands. The black-box nature of this algorithm prevents the gaining of any insight into the decision making process, while the very size of the model restricts its offline usage. For these reasons we propose a straightforward simplification of eXtasy that relies on a two-step modeling approach [60] and that retains the advantageous performance of the former. As a by-product we demonstrate the importance of phenotype-based scores in classification performed by the system. Chapter 4 fully addresses interpretation issues with eXtasy.

Finally, the features used in the eXtasy model are defined over domains that have different granularity. In particular, each predictor describe either gene, mutation or mutation/phenotype combination. At the same time, each data instance (row in data matrix) represents one gene/phenotype combination. Consequently, it is often the case that several instances share the same value of some feature. Also, the instances (examples) are often interdependent given a granularity level, i.e., all examples have the same value for a specific feature as well as for the outcome. For example, several data records that correspond to a

single mutation normally share the same outcome (disease-causing mutation or not) and the same value of some feature that describes mutation (ex. SIFT score).

This induces a bias during the training phase that results in decreased classifier performance on unseen examples. The bias arises because the feature value can appear correlated with the target variable, whereas it is in reality correlated with the hierarchical structure present in the data. In other words, as many data instances share the same outcome and the same value of specific feature, the algorithm can be prompted to *memorize* this particular value of the feature instead of generalizing from it.

A typical example is the haploinsufficiency score which is defined for a gene, so all the mutations within a given gene share the same value of it. If many mutations associated with this gene also have the same outcome (which is often the case), the algorithm has a sufficient number of examples to learn that given value of haploinsufficiency implies that specific outcome. Clearly, a rule of this form can not say anything about a new example with a previously unobserved value of haploinsufficiency.

Hence, in eXtasy, this bias can materialize as learning to identify genes which constitute the training set, instead of extracting general characteristics of the disease causing mutations. In our last study [61, 62] we assess the extent of this bias and propose a sampling-based remedy (Chapter 5). This modification has resulted in the additional **three-fold** increase of system's precision under the class distribution as observed in wild.

Important points of previous discussion in this section can be summarized by three major research directions that have been followed to completely fulfill the goal of improved disease-causing variant prioritization. These include modeling aspects (How can prioritization model be improved algorithmically?), data aspects (Which sources of information can be utilized for variant prioritization?) and validation aspects (What is correct way to validate a model in order to prove its practical utility?). Each of these directions translates to several research questions that we individually addressed in this work :

- **How can a prioritization model be improved algorithmically?**
  - Which computational method should be used for prioritization?
  - How to combine predictions for several mutation/phenotype pairs into one score for a variant?
  - How to treat hierarchical granularity of feature domains?
- **Which sources of information can be utilized for variant prioritization?**

  - How to choose the control cases?
  - Are all features informative for the task at hand?
  - Does phenotypic information contribute to prioritization performance and how important is it?

- **Which is the correct way to validate a model in order to prove its practical utility?**
  - Which performance measures are appropriate for evaluating variant prioritization algorithms?
  - Which validation scheme is appropriate in the presence of hierarchical granularity of the feature domains?
  - What performance can be expected from the method in real-life applications?

The main chapters of this manuscript (Chapters 2-5) provide a detailed treatment of the enumerated research questions. By delivering solutions to each of them individually, we manage to utilize several avenues for improvement of variant prioritization and, even more important, to prove that the resulting algorithm has sufficient performance to be really *used* in practice. Having in mind the translational focus of this work, we strongly believe that this is the outcome that matters the most.

1.3 Organizational context of the research project

The work presented in this manuscript has been conducted in the STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics. STADIUS is a group within the Department of Electrical Engineering (ESAT) of KU Leuven, with research focus on digital systems and mathematical modeling. Fundamental research efforts of the group concentrate on signal theory, control theory, circuit theory and linear and non-linear systems (including neural networks, Support Vector Machines and complex systems). In addition, STADIUS performs applications driven algorithm design and implementation, in which these theoretical concepts translate to three main application domains - digital communications, biomedicine and modeling and control of industrial processes. The group also maintains extensive collaboration with the University Hospital Leuven (UZ Leuven) that resulted in numerous successful joint projects in biomedical and clinical fields.

STADIUS is also a research group in the Department of Medical IT of the Flemish strategic research center iMinds, where it facilitates the utilization of data for

providing better health care, via development of ICT-based Health Decision Support Systems (HDSS) for clinics, patients and policy makers. Decision support for health care professionals primarily concentrates on modeling and interpretation of *big* medical data (such as imaging, signals, -omics data) for better characterization of pathologies, improved diagnostics and development of patient-tailored therapies. Patient-centered decision support covers solutions for disease management and remote monitoring of health parameters. Decision support for policy makers focuses on data mining of data bases in the overall social security system. These three main objectives of the iMinds Medical IT Department are materialized through five research lanes in STADIUS - signal processing, bio-informatics, medical imaging, big data and analytics decision-support.

The Bioinformatics research group within STADIUS concentrates on development of computational methods to address challenges in two major subfields, i.e., System Biomedicine and Top-down System Biology. Within system biomedicine, the group develops statistical and machine-learning methods for integration, analysis and mining of clinical and high-throughput data to facilitate identification of diagnostics, prognostics or drug-target biomarkers. Top-down systems biology research in the group mostly addresses the reconstruction of networks from high-throughput molecular data.

Specifically, the Bioinformatics group is also one of leading players in development of gene and mutation prioritization methods. Contributions of the group in this domain include several widely accepted prioritization tools on one hand and important theoretical breakthroughs on the other hand. For example, the Endeavour framework [58, 63] was the first method to implement genomic data fusion, and it is still *de-facto* standard in the field of gene prioritization [64, 65]. Other notable algorithmic developments in the group include that for kernel [66], networks [67, 68], and text-mining based [69] gene prioritization. Finally, researches from the group were frequently invited to disseminate their visions of the current state and prospects of the field [70, 71].

This work is a natural extension of the established tradition of gene prioritization research in STADIUS. Setting foot in the disease-causing mutation prioritization domain represents an additional step toward completing the competence profile of the Bioinformatics group and maintaining the excellence status of the group within the field. Furthermore, research questions addressed in this project are perfectly aligned with one of the top-level strategical objectives of the iMinds Medical IT department, namely development of Clinical HDSSs (Figure 1.5). In particular, the theoretical insights and ready-for-use software platforms that have resulted from this work have the potential to aid characterization and diagnostics of rare genetic disorders, which is clearly a beneficial translational outcome. Finally, the complete project has been developed in context of STADIUS applied

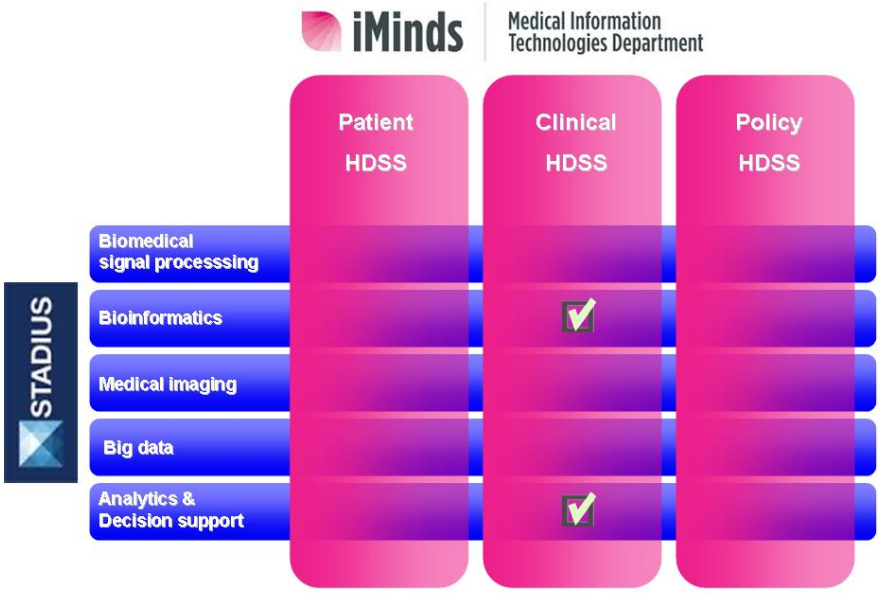


Figure 1.5: **Organizational context of the thesis.** Organizational placement of research conducted in this work in terms of iMinds Medical IT Department strategic objectives and STADIUS research tracks.

biomedical research and it is strongly backed by the corpus of fundamental knowledge in domain of non-linear systems theory that is accumulated and extended by the group over the years.

The major deliverables of this research project include scientific publications, software tools and theoretical insights. Each phase of its development has been individually documented and reported, resulting in three journal (Nature Methods, BMC Bioinformatics and PLOS One) and two conference (IEEE BIBM 2013, PRIB 2014) publications. An actionable prioritization model has been deployed via two freely accessible software platforms - one Web-based and one stand-alone computer application. Furthermore, the Matlab implementation of hierarchical sampling for public use will be released together with the corresponding manuscript as a supplementary material.

Theoretical insights resulting from this work cover several diagnosed shortcomings in the current practice of training and evaluation of variant prioritization

algorithms, as well as recommendations for their improvement. They also include detailed description, analysis and solution for relatively unrecognized training and validation bias. In addition, the utility of phenotypic information for variant prioritization has been immediately recognized by the community, resulting in a new generation of algorithms that incorporate this type of information into their design [72]. Finally, successful completion of the project attracted modest media attention (articles in two local newspapers, interview with prof. Moreau on ROB-TV, TEDx Brussel with prof. Moreau on October 28 2013), potentially improving visibility of our research group.

The complete eXtasy framework is the result of a team effort of several researchers form STADIUS, UZ Leuven and the Department of Computer Science of the KU Leuven. Therefore, as a majority of this thesis is based on it, it seems appropriate to outline my personal contributions to the project. Together with my colleague Alejandro Sifrim, I assumed one of the major roles in the eXtasy development. My personal contributions (organized by resulting manuscripts) include :

- Sifrim A.*, Popovic D.*, Tranchevent L.-C., Ardeshirdavani A., Sakai R., Konings P., Vermeesch J., Aerts J., De Moor B., Moreau Y., **"eXtasy: Variant Prioritization By Genomic Data Fusion"**, Nature Methods, vol. 10, no. 11, Sep. 2013, pp. 1083-1084.
  ** Equally contributing first authors*
  - Conceptual definition of the project
  - Design, implementation and execution of benchmarking pipelines
  - Analysis of experimental results
  - Model building and deployment for inclusion in web site and standalone tool
  - Application testing
  - Initial draft of the report
- Sifrim A., Popovic D., Vermeesch J., Aerts J., De Moor B., Moreau Y.,**"Comparison of aggregation methods for multiphenotype exomic variant prioritization."**, submitted for publication in PLOS One
  - Method design
  - Analysis of experimental results
  - Proofreading and critical discussion on report

- Popovic D.*, Sifrim A.*, Moreau Y., De Moor B., **"eXtasy simplified - towards opening the black box"**, in Proc. of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shanghai, China, Dec. 2013, pp. 24-28
  * *Equally contributing first authors*

  - Conceptual definition of the project
  - Method design
  - Design, implementation and execution of experiments
  - Analysis of experimental results
  - Initial draft of the report

- Popovic D., Sifrim A., Davis J., Moreau Y., De Moor B., **"Problems with nested granularity of feature domains in bioinformatics - the eXtasy case."**, accepted for publication in BMC Bioinformatics
  &
  Popovic, D., Sifrim, A., Davis, J., Moreau, Y., De Moor, B. (2014). **"Improving performance of the eXtasy model by hierarchical sampling."**, in Proc. of the Pattern Recognnition in Bioinformatics - 9th IAPR International Conference PRIB 2014, Stockholm, Sweden, Aug 2014, pp. 125-128.

  - Conceptual definition of the project
  - Description, clarification and analysis of previously not addressed learning bias
  - Novel computational method for mitigating the training bias resulting from nested granularity of feature domains
  - Design, implementation and execution of experiments
  - Analysis of experimental results
  - Initial draft of the report
  - Software tool

1.4 Structure of the thesis text

The rest of this manuscript is based on the enumerated articles and is structured as follows. In Chapter 2 we introduce the eXtasy variant prioritization method, including details on how it was constructed and benchmarked. In section 2.1 we motivate the need for computational variant prioritization, review existing methods and their limitations, outline the design of our algorithm and

summarize the results of the experiments. We also discuss temporal performance-deteriorating effects and importance of different information sources for the model.

In subsection 2.2.1 we describe the data generation process and motivate our choice of outcomes used in training. Hence, together with subsection 2.2.3 where we test different prioritization scenarios, subsection 2.2.1 provides an answer to the question how to optimally choose control cases. Subsection 2.2.2 describes how the classifier benchmark was designed and conducted, addressing the choice of the learning algorithm. Here we also discuss the proper way to validate classifiers trained on features that are defined over domains of uneven granularity. Subsection 2.2.4 introduces a scheme for *semi-prospective* model validation. This experiment has facilitated the assessment of future performance of eXtasy, and revealed potential sources of bias. In subsection 2.2.5 we analyze the importance of individual features in prioritization, proving the viability of the genomic data fusion strategy.

Subsection 2.3.1 zooms-in on the data generation process, describing it in greater detail. An appropriate way of interpreting the benchmark results is discussed in subsection 2.3.2, together with the choice of the correct performance metrics for mutation prioritization algorithms assessment. The final part of this chapter (2.3.3) contains a collection of figures and tables that provide extensive information on all aspects of benchmark data, implementation of the method and experimental results.

In chapter 3 we suggest and analyze several algorithms for combining multiple phenotype-based predictions into a single score. Section 3.1 introduces the problem of aggregating across different prioritizations and discuss how does this problem relate to eXtasy data sources. In subsection 3.2.1 we define the benchmarking approach and outline the process of generating quasi-synthetic data. Subsections 3.2.2 and 3.2.3 describe a number of prediction aggregation schemes based on non-parametric order statistics and parametric modeling respectively. In section 3.3 we present results of experiments, while section 3.4 discusses major findings of the study and summarizes recommendations for application-tailored choice of aggregating scheme.

In chapter 4 we propose a decision tree-based simplification of the core eXtasy model, hoping to facilitate the interpretability of the decision-making process. This chapter as a whole provides a proof that phenotype-specific information can aid variant prioritization and that this type of information in fact plays major role in superior performance of eXtasy. Section 4.1 introduces the problem of eXtasy's black-box nature and describes shortcomings in using RF-FI to circumvent it. Subsections 4.2.1 and 4.2.2 describe solution to the problem and data used for benhmarking, respectively. In section 4.3 we present and

discuss results of the experiments. In section 4.4 we enumerate the important lessons-learned and suggest how caveats of the proposed approach can be addressed.

In chapter 5 we discuss a hierarchical bias that is present in the eXtasy training scheme and propose a sampling-based solution to it. In section 5.1 we further develop discussion from 2.2.2 on validation problems that originate from hierarchical granularity of data. In addition, we illustrate how the existence of these structures in training data can degrade generalization performance of a model, even if the validation is correctly preformed.

Subsection 5.2.1 describes a method that eliminates this bias from the training phase, which improves performance on unseen examples. In subsection 5.2.2 we describe synthetic data that we use to characterize the source and extent of the problem in detail. Subsection 5.2.3 outlines the design of the eXtasy-based benchmark. This benchmark relies on data that mimic the class distribution that is typically encountered in practical applications, so it provides complementary insight into the real-life utility of eXtasy. The results of both benchmarks are presented and discussed in section 5.3, while section 5.4 summarizes the major findings.

Chapter 6 enumerates the most important findings that have resulted from the complete research project, discusses the limitations of the proposed methodology and outlines possible future directions of exploration. The organization of thesis text in terms of structure and relations to the previously formulated main research questions and sub-questions, that we briefly introduced here, is outlined on Figure 1.6.

As apparent from the figure, the algorithmic aspects of the method development are tackled in section 2.2.2 of the chapter 2, chapter 3 and chapter 5. These sections of the text address choice of the appropriate classification algorithm, aggregation of multiple decisions and approach for de-biasing the model, respectively.

Design of benchmarks in terms of data used (i.e., How to choose the control cases?) is described in sections 2.2.1 and 2.2.3. In section 2.2.5 we prove viability of the genomic data fusion strategy (Are all features informative for the task at hand?), while in chapter 4 we further explore the utility of phenotypic information in variant prioritization.

Finally, the research questions related to the validation are answered in sections 2.3.2 (Which performance measures are appropriate for evaluating variant prioritization algorithms?), 5.1 (Which validation scheme is appropriate in the presence of hierarchical granularity of the feature domains?) and 2.2.4 and 5.2.3 (What performance can be expected from the method in real-life applications?).

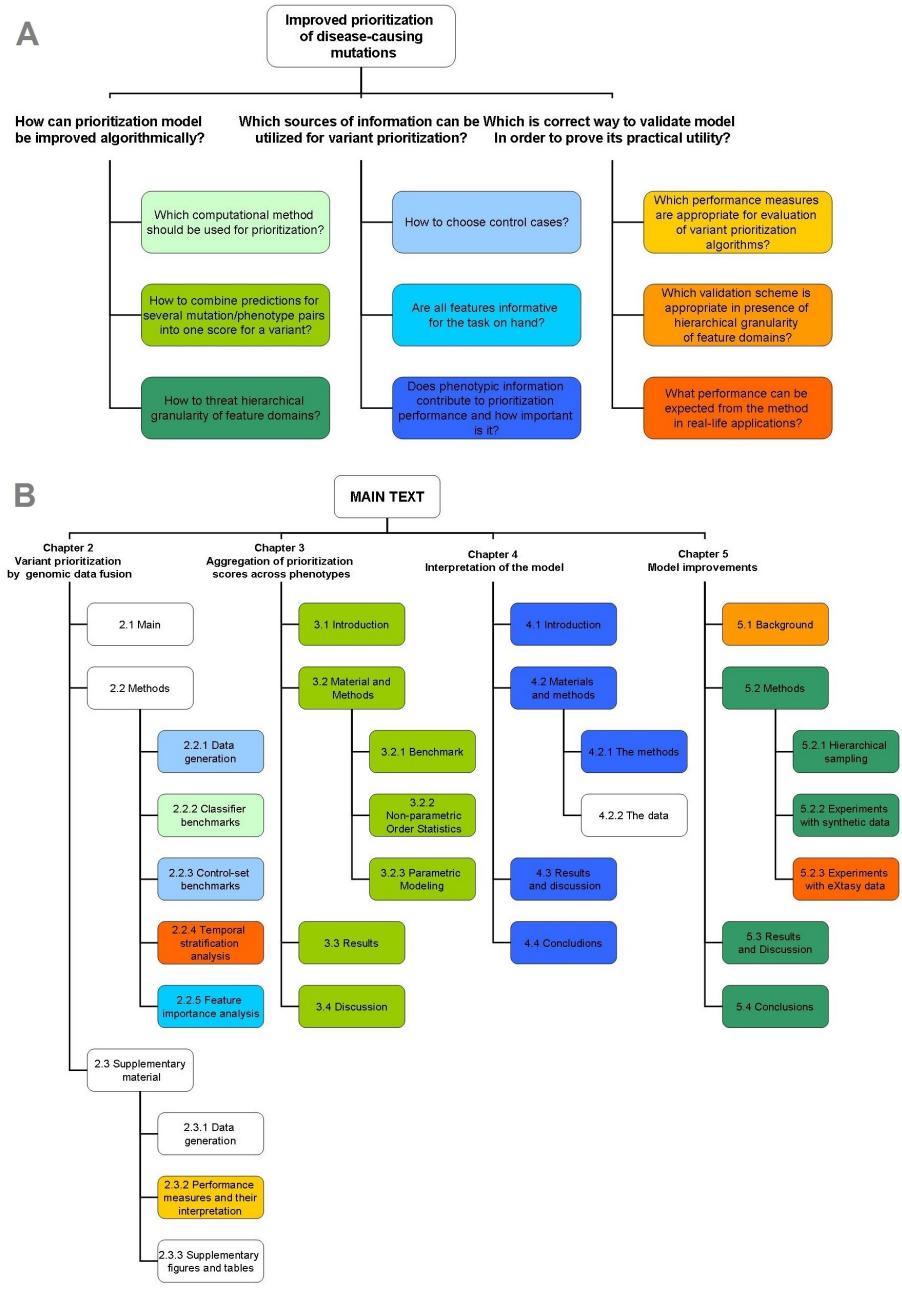


Figure 1.6: **Structure of the thesis text** A) Hierarchy of main research questions that are tackled in thesis. B) Structure of main thesis text. Chapters and sub-chapters are shaded in colors corresponding to research questions they address. White boxes denote technical sections (data acquisition procedures, general chapters and such).

Chapter 2

Variant prioritization by genomic data fusion

This chapter has been previously published as :

Sifrim A.*, Popovic D.*, Tranchevent L.-C., Ardeshirdavani A., Sakai R., Konings P., Vermeesch J., Aerts J., De Moor B., Moreau Y., **"eXtasy: Variant Prioritization By Genomic Data Fusion"**, Nature Methods, vol. 10, no. 11, Sep. 2013, pp. 1083-1084.
* *Equally contributing first authors*

ABSTRACT

Massive parallel sequencing greatly facilitates the discovery of novel disease genes causing Mendelian and oligogenic disorders. However, many mutations are present in any individual genome, and identifying which ones are disease causing remains a largely open problem. We introduce a new approach to prioritize nonsynonymous single nucleotide variants (nSNVs) that substantially improves prediction of disease-causing variants in exome sequencing data by integrating variant impact prediction, haploinsufficiency prediction and phenotype-specific gene prioritization.

2.1 Main

Rare variation identified by exome sequencing is particularly useful for discovering the cause of rare monogenic disorders. By selecting only nSNVs and loss-of-function mutations that are not present in healthy populations or unaffected samples, we can discard a large proportion of exomic variation as likely neutral. However, despite such aggressive filtering, several thousand candidate causal mutations remain, and we need predictive methods to prioritize variants for further validation. Several computational methods take into account biochemical, evolutionary and structural properties of mutations to assess their potential deleteriousness [51, 50, 55, 73, 56] but most of these methods suffer from high false positive rates when predicting the impact of rare nSNVs [73]. A plausible explanation for this poor performance is that many of the scrutinized variants are mildly deleterious and subject to weak purifying selection [74, 75] but are not specific to the disease of interest.

To assess this hypothesis and to improve variant prioritization, we propose a genomic data fusion methodology [71] that integrates multiple strategies to detect the deleteriousness of mutations and prioritizes them in a phenotype-specific manner. A key innovation included in our strategy is a method for gene prioritization [58], which scores mutated genes according to their similarity to known disease genes by fusing heterogeneous genomic information. We also integrate haploinsufficiency prediction scores [76] that predict the probability that the function of a gene is affected if it is present in a functionally haploid state. To integrate these data sources, we use Random Forests learning [20] and train our model on the Human Gene Mutation Database (HGMD) of human disease-causing mutations [77] compared against three control sets: common polymorphisms and two independent sets of rare variation (Supplementary Note 1).

After generating and annotating the different data sets, we inspected the distribution of deleteriousness prediction scores across the positive and control sets (Figure 2.9, Supplementary figures and tables). All predictions seemed to score the positive set high and thus showed a high sensitivity. In the control sets, most methods correctly classified common polymorphisms as benign, yet they classified a substantial proportion of rare variation as deleterious, leading to low precision. Control variants seemed to occur more often in genes predicted to maintain functionality in a haploid state (haplosufficient), whereas disease-causing variants showed no clear pattern. Disease-causing variants were primarily found in top-ranked genes after gene prioritization. By contrast, control variants showed a homogeneous distribution of gene prioritization ranks, which is to be expected under the assumption that they are prioritized for randomly selected phenotypes.

By integrating these different scores, we aimed to boost our ability to discriminate between putatively mildly deleterious rare variants and actual disease-causing variants. We evaluated several commonly used classification approaches and chose a Random Forests algorithm because it outperformed all other classification algorithms on this task (Table 2.1). We trained this classifier by comparing our positive set of disease-causing variants to the rare-variant control sets and observed a considerable improvement in all performance measures over classical deleteriousness prediction methods. This was the case when distinguishing between disease-causing and rare control variants as well as between disease-causing variants and common polymorphisms (Figure 2.1 and Table 2.2).

The performance against common polymorphisms was in line with published results for deleteriousness prediction tools because these tools were trained using common polymorphisms as controls. The performance of these tools against rare non–disease-causing variants was much poorer than that against common polymorphisms. Precision was the most improved performance measure, which is important when dealing with a prioritization task (Table 2.1 and Supplementary Note 2). However, performance measures obtained in retrospective benchmarks such as ours are usually optimistic estimates because of the bias of prior information for gene-prioritization predictions [71, 78]. This problem is difficult to address in an initial study and can be truly resolved only by long-term prospective benchmarks wherein predictions are made on the current state of knowledge and validated in future studies.

To estimate the extent of this bias in our benchmark, we assessed classification performance on the basis of the discovery date of the causative variation using data for gene prioritization before the year of discovery (Figure 2.2 and Table 2.3). If a bias was present, it would be less prominent in recent discoveries, as these gene-disease associations would be less likely to be directly or indirectly incorporated in the gene-prioritization data sources. Even though we saw a slight decline in performance for more recent discoveries, the method still performed considerably better than classical deleteriousness prediction scores (Figure 2.2 and Table 2.4). The performance of deleteriousness prediction scores was not affected by the year of discovery of the association, which is expected because these scores do not integrate time-dependent information. Finally, we looked at feature importance by shuffling each feature across disease-causing and control variants, evaluating the increase in classification error (Figure 2.3). This analysis showed that all included features were informative and improved classification to some degree.

We have demonstrated an approach that integrates biochemical, evolutionary and phenotype-specific information to prioritize mutations for follow-up validation studies. This approach helps to distinguish disease-causing mutations

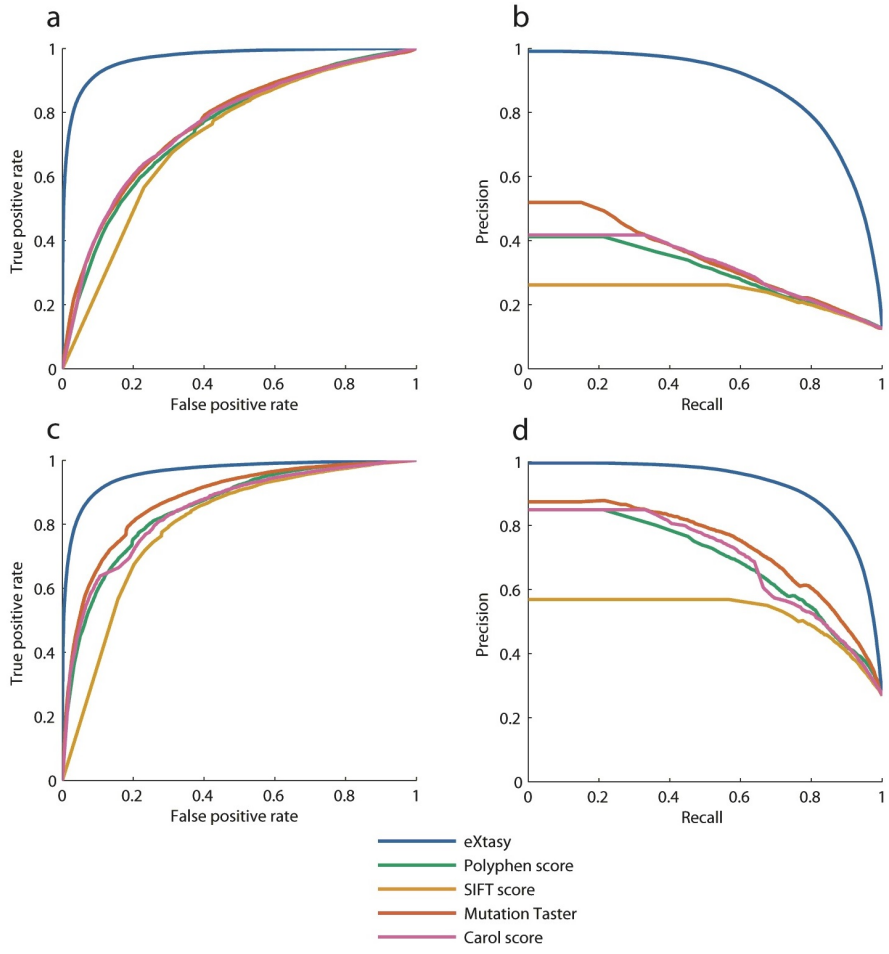


Figure 2.1: **Classification scenarios performance.** Receiver Operator Characteristics (ROC, panels A and C) and Precision-Recall (PR, panels B and D) curves for five classifiers tested against test sets comparing disease-causing variants to either rare non-disease-causing variants (A and B) or common polymorphisms (C and D). All curves represent the averages of 100 iterations comprising of different random splits of training/testing sets.

from neutral common polymorphisms as well as rare, potentially deleterious but phenotypically unrelated variation in the coding genome. We acknowledge that performance measures are likely overestimated because of the biased nature of retrospective benchmarks, yet we see a marked improvement in prediction performance over frequently used deleteriousness prediction scores in recently published mutations. We envision that in the near future, initiatives such as

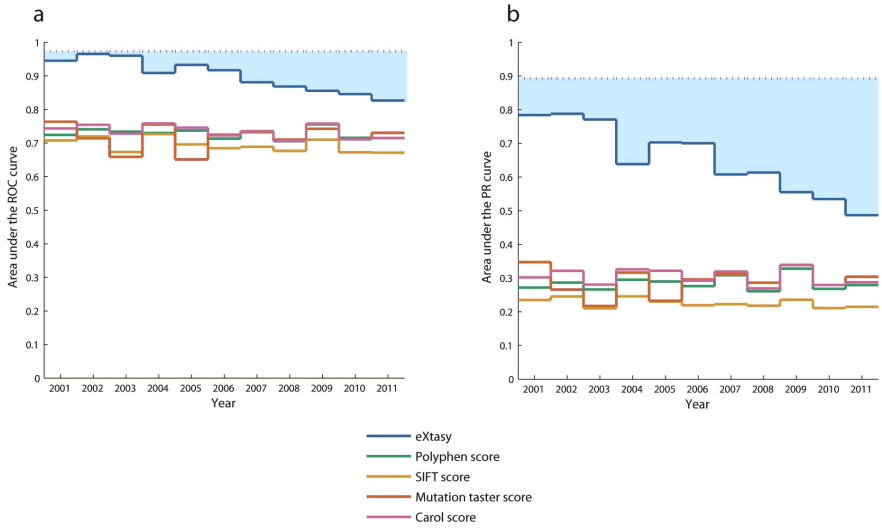


Figure 2.2: **Temporal stratification of performance.** Values of the area under the ROC (A) and PR (B) curves for the five classifiers obtained during testing by stratifying the positive testing set according to the year of publication. The dotted blue line indicates performance of eXtasy obtained on the full classifier benchmark (Figure 2.1), while shaded area covers space of expected performances for different use-case scenarios. eXtasy was trained only on disease-causing mutations published prior to 2001 and all variants corresponding to genes present in both training and test sets were removed from the test set.

the Critical Assessment of Genome Interpretation, although currently focused on single-phenotype benchmarks, will play a major role in providing unbiased prospective benchmarks to optimally assess the performance of methods such as the one described in this study.

Future research and the availability of larger public disease-causing variation data sets will likely widen the scope of the method to other types of mutations (mitochondrial, noncoding, synonymous, nonsense and splice-site mutations, as well as indels). Also, the addition of other data sources, such as locus-specific information (for example, copy number–variant prevalence and genome-wide association study–related loci), to our method and its integration into genetic association tests across multiple samples [79, 80] will likely increase its power to discover the cause of genetic disease.

A web tool and stand-alone version implementing our approach are available at http://homes.esat.kuleuven.be/~bioiuser/eXtasy/ (source code is at http://github.com/asifrim/eXtasy/).

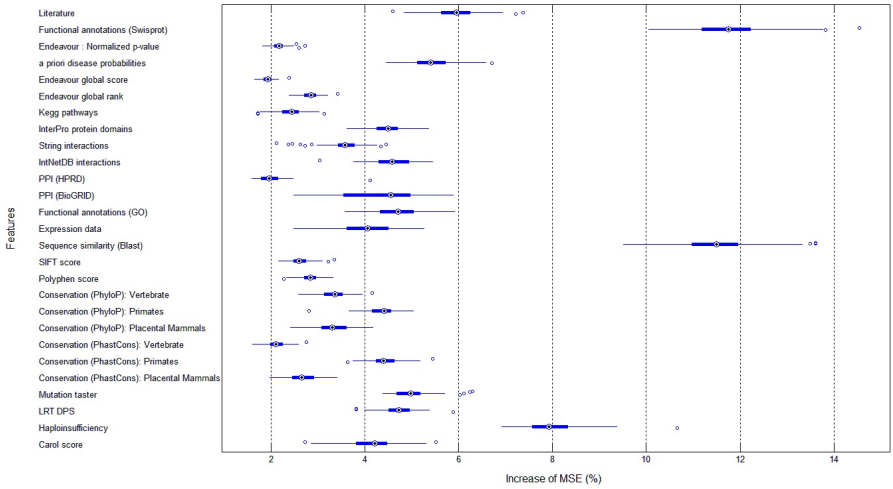


Figure 2.3: **Feature importance measures.** The box plots depict the range of mean square error increase when an individual feature is shuffled across classes for 100 randomizations. The central mark (dot in circle) corresponds to the median, the box represents the range between the 25th and 75th percentiles, the line depicts the range of the error values and small circles correspond to potential outliers. Non-zero increases in mean square error mean that the feature is informative. However, the relative informativeness of each feature is hard to interpret because some features are not independent of each other (and thus removal/shuffling of the feature is in part compensated by the dependent features).

2.2 Methods

2.2.1 Data generation

To determine the effectiveness of our approach, we assembled and evaluated a data set consisting of disease-causing variants and different types of control variants. We used a positive data set consisting of 24,454 disease-causing nSNVs from the HGMD associated with 1,142 different Human Phenotype Ontology (HPO) terms [81]. Mapping between HGMD disease descriptors and HPO terms was performed using the Phenomizer tool [82]. Phenotypic terms for which fewer than three implicated genes were known were discarded (excluding the gene in which the variant was located) so as to allow meaningful subsequent gene prioritization. For the control data sets, we considered two classes of variants on the basis of their minor allele frequency (MAF) in the 1000 Genomes Project data set: common polymorphisms (MAF $>$ 1%) and rare variation (MAF 1%). Additionally, we compiled a third control set of rare variants

using in-house–sequenced exomes of healthy individuals and retaining only high-quality calls (coverage $> 30\times$) not present in any publicly available variant database (NHLBI EVS, dbSNP or 1000G). After randomly assigning groups of 500 control variants from each control set to each HPO term represented in the disease-causing variants, we annotated all variant-phenotype combinations with functional information.

For every variant, we extracted PolyPhen-2 [51], sorting intolerant from tolerant (SIFT) [50], MutationTaster [55] and likelihood-ratio test (LRT) [56] deleteriousness prediction scores from the dbNSFP database [83]. We calculated combined annotation scoring tool (CAROL) aggregate deleteriousness scores [54] and added phyloP [84] and PhastCons [85] evolutionary conservation scores across vertebrates, primates and placental mammal subsets from the UCSC Genome Browser database. We also included precalculated gene haploinsufficiency prediction scores [76] where available. Finally, we computed gene prioritization scores with Endeavor [58] using gene-phenotype associations obtained from HPO. (For additional information on the data generation and annotation process see Supplementary Note 1).

2.2.2 Classifier benchmarks

We set up a benchmark to determine the optimal classifier for genomic data fusion. To do this, we followed published guidelines [86] applicable to these types of studies. Initially, we removed all records containing missing values and split the disease-causing variants into training (two-thirds of the total number of genes) and testing sets (one-third of the genes) (Figure 2.4). The data records were stratified at the gene level (the highest level of granularity) to assure that the algorithm is not overfitting the gene level information and thus does not overestimate performance. Subsequently, we randomly subsampled the negative examples in the training set so as to balance their number with the number of positive examples. We then trained eight different classifiers using the same combinations of training and test subsets for all classifiers.

In addition, we applied state-of-the-art deleteriousness prediction methods (PolyPhen, SIFT, MutationTaster and CAROL scores) using their respective published or documented decision thresholds (0.85, 0.95, 0.5 and 0.98, respectively). As there were several records per single variant in the data set (because of the different phenotypes per variant), we used the maximum score across phenotypes in testing to predict the outcome for each variant. We chose the maximum as this is robust to noninformative phenotypes (because of phenotypically variable diseases), which might be present in our benchmark data set of disease-causing variants. This did not apply for methods that produce

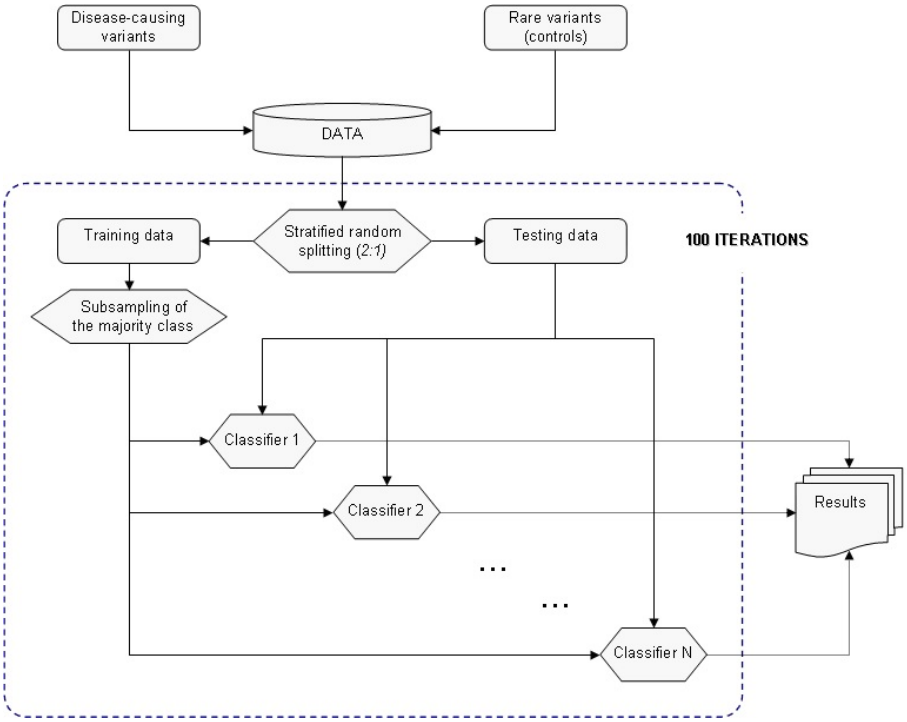


Figure 2.4: **Workflow of the classifier benchmark.** The data set composed of disease causing and rare variants is randomly divided into two parts, where two-third of the examples are assigned to the training set and one third to the test set (stratified by outcome and divided along gene identifiers). The majority class in the training data set is subsequently sub-sampled to equalize the number of outcomes in each class. The eight classifiers are trained on this data and tested against test sets not containing genes or variants present in the training data. The whole process is repeated 100 times, with different random data splits in each iteration.

a single score per variant because they are not phenotype specific (PolyPhen, SIFT, MutationTaster and CAROL).

The procedure was repeated 100 times on different random splits of data to obtain an estimate of the variance of the resulting performance measures. The results were computed in terms of accuracy, sensitivity, specificity, positive and negative predictive value, Matthews correlation coefficient, area under the receiver operating characteristic (ROC) curve and precision-recall (PR) curves. We observed that all the classifiers built using all heterogeneous data sources performed better than state-of-the-art deleteriousness prediction methods, proving that these data sources contain additional information that facilitates better predictions.

Among all classification algorithms, Random Forests (RF) outperformed all others in terms of all performance metrics, with the exception of sensitivity and negative predictive value (NPV), for which linear discriminant analysis (LDA) performed marginally better (Table 2.1 and Figures 2.5 and 2.6). However, precision (positive predictive value or PPV) of the LDA was very low (0.35), indicating that the classifier is overly optimistic. The same observation holds for all of the state-of-the-art methods (precision between 0.20 and 0.23).

Furthermore, the PR curve for the Random Forests shows that changing the default threshold (0.5) of the classifier results in a sharp increase in precision but a small loss in sensitivity. This suggests that most of the true positives are highly ranked by the method. Also, the reported accuracy and other aggregate performance measures of the state-of-the-art tools depend greatly on where the decision threshold is set and the skewness of the class distribution. Even though the AUC for some of the ROC curves is higher on one tool than on another, sometimes a point can be found where a certain measure is higher for the tool with the lower AUC.

For all classification algorithms, we use their respective Matlab 7.10 implementations: classregtree class for the decision trees, TreeBagger class for the Random Forests, function knnclassify for k nearest neighbors, function classify (with argument "linear" or "quadratic") for the LDA and QDA, respectively, class NaiveBayes for naive Bayes classification, and functions svmtrain and svmclassify for the feed-forward neural networks. Most of the used functions and classes are part of the Statistical Toolbox, except for KNN and SVMs (Bioinformatics Toolbox). The details on parameter settings of particular classifiers are provided in the Table 2.5

2.2.3 Control-set benchmarks

We set up an additional benchmark to assess the behavior of the final model under different classification schemes with regard to the different negative outcomes. We considered two different scenarios for training and testing. In a first scenario, the Random Forests model is trained using a subset of the rare non–disease-causing variants as negatives and is tested against the rest of them (as in our standard setup). In a second scenario, the Random Forests model is trained using all of the rare non–disease-causing variants available as negatives and is tested against the data set containing common polymorphisms.

In the case of the first scenario, the validation scheme is identical to the classifier benchmark: all data, including all positive and the negative examples, have been repeatedly (100 iterations) divided at random into training and testing subsets, grouped gene by gene. In the case of the second scenario, only the

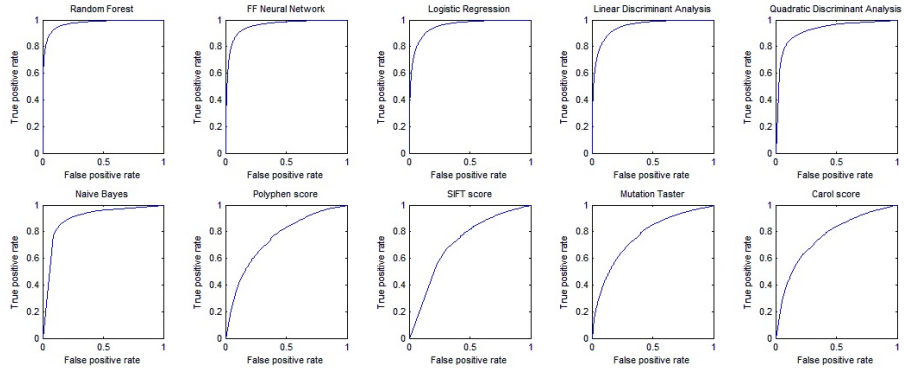


Figure 2.5: **Receiver Operating Curves (ROC) of classifier performance on benchmark data.** The averaged ROC curves of 6 classifiers based on genomic data fusion (from Random Forest to Naive Bayes) and 4 deleteriousness prediction methods (from Polyphen to Carol score). The curves are averaged across 100 training vs. test splits of the benchmark data using threshold averaging. The genomic data fusion classifiers clearly outperform four deleteriousness prediction scores across all operating points.

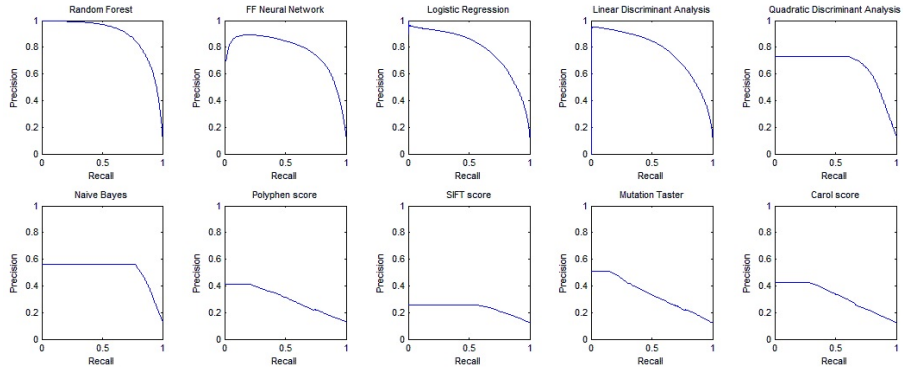


Figure 2.6: **Precision-Recall curves of classifier performance on benchmark data.** The averaged PR curves of 6 classifiers based on genomic data fusion (from Random Forest to Naive Bayes) and 4 state-of-the-art methods (from Polyphen to Carol score). The curves are averaged across 100 training vs. test splits of the benchmark data using threshold averaging. The genomic data fusion classifiers clearly outperform four deleteriousness prediction scores across all operating points.

positive examples were randomly assigned, whereas the two distinct groups of negatives, rare and polymorphisms, were used respectively for training and testing (Figure 2.7). We made sure that during a single iteration, a split of positives stayed the same across the two scenarios.

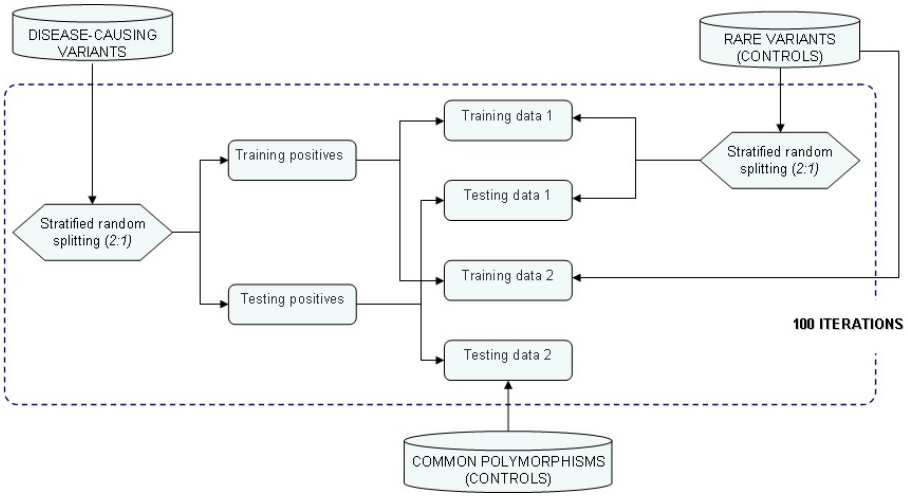


Figure 2.7: **Creating data sets for testing classification scenarios.** The data set of disease causing variants is randomly divided into two parts with variants grouped gene by gene, where two-thirds of the examples are assigned to the training set and one third to the test set. The first training/test pair is augmented with negative examples from the rare variants data set using the same splitting procedure. The second pair is augmented with two distinct groups of negatives, rare and polymorphisms. Prior to classification, both training sets are balanced in terms of outcome distribution (not shown on the diagram for clarity). The whole process is repeated 100 times, with different random data splits in each iteration.

2.2.4 Temporal stratification analysis

To analyze the sensitivity of the method with regard to a priori gene or disease association biases, we set up an additional benchmark to estimate the effect size of such biases. Under the assumption that recently published genes would be less likely to be biased by our gene prioritization step, we stratified our positive testing set of disease-causing variants by year of publication (2000–2011) while training the model only on data published before 2000. In this way we measure performance before and after Endeavor's data sources were last updated (the last update occurred in 2008).

This threshold applies to both variant and gene level of data granularity, so variants discovered after 2000 but that are associated with genes that are part of the training (i.e., for which the gene-phenotype association was discovered before 2000) are also removed from test sets. The negative examples (non–disease-causing variants) were randomly assigned to one given year between 2000 and 2011, in numbers matching the class distribution of the whole data (given the number of positives in a particular year) and with no overlap between training

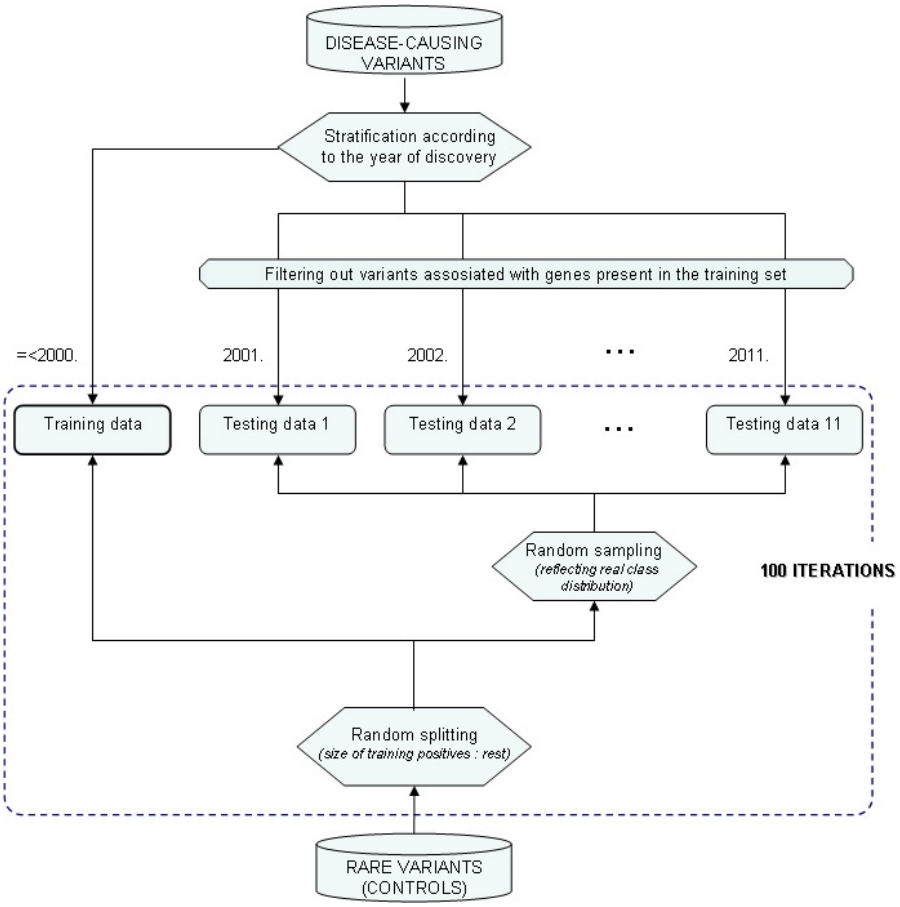


Figure 2.8: **Stratification of data for temporal analysis.** The data set of disease causing variants is divided into twelve parts (one for training and eleven for testing) given the year of discovery. Variants associated with genes present in the training set are filtered out from all of the testing sets. The training data is then enriched with negative examples by random sampling from the rare variants data set (same amount as positives, grouped gene by gene). The rest of the rare variants are randomly assigned to one of the eleven testing sets, in amount that corresponds to the class distribution of the whole data given the number of positives in a year. The whole process of assigning negatives is repeated 100 times with different random splits in each iteration.

and test sets (see Figure 2.8). As before, the splits were performed genewise. After the training phase, the classifier was used on cases from the subsequent years (2000–2011). The whole procedure of randomly assigning negatives was repeated 100 times to get stable estimates of performance metrics.

We observed a slight temporal decline in performance throughout all testing years. Nevertheless, eXtasy still significantly outperformed all classical deleteriousness prediction methods across all years. We attribute this decline in performance to the fact that over time some disease-causing genes are better described in the gene-prioritization sources (e.g., literature mining) and are therefore easier to classify. Although such effects are likely present and point to the fact that our main benchmark is an overoptimistic estimate of the real performance, this benchmark setting is a pessimistic estimate of the real performance because of various properties of the training/validation scheme.

First, only a fraction of the positive training data can be used (data before 2000), leading to suboptimal learning of the features of disease-causing variants. Second, the training data themselves contain well-described genes and are thus themselves biased toward easier-to-classify examples. This leads to overly optimistic decision thresholds in the classifier and thus degrades the performance when faced with more difficult examples of less well–described genes in the testing set. Finally, owing to the gene-wise stratification of training and testing sets (to avoid overfitting specific genes), if gene data were published before 2000 but then later published in the light of a new phenotype, they were omitted from the test set. Well-described genes are often discovered to play important roles in new (and often related) physiological processes. This type of discovery can greatly benefit from gene-prioritization approaches but are excluded from this benchmark.

The real performance of the classifier depends greatly on the use case, which is unknown to the researcher applying the method. In the case where the cause of the phenotype is a novel mutation in a previously described gene, the performance is likely to resemble, or even exceed, that of the cross-validation benchmark. When the phenotype is caused by a novel mutation in a gene not previously associated with the phenotype, or in a gene associated with a different phenotype, the real performance likely lies between the overoptimistic cross-validation benchmark and the pessimistic temporal-analysis benchmark (the shaded light blue area in Figure 2.2).

2.2.5 Feature importance analysis

The Random Forests algorithm has the intrinsic ability to estimate the importance of features used for training [20]. This is achieved by measuring the difference between the mean-square error of the prediction on out-of-bag (OOB) examples when values of a given feature are shuffled compared to the error on undisturbed OOB data. This procedure is repeated for each and every

tree in the ensemble with its corresponding OOB examples, providing a global measure of feature importance.

Here we analyzed how the different features in the data contributed to the overall classification. In particular, we ran 100 simulations (in line with previously described benchmarks) per feature, during which ensembles are built using different random subsamples of the negative data. The result in terms of mean-square error increase when shuffling the feature is displayed in Figure 2.3. From the plot, it appears that all features contribute to the classification to some degree. The increase of total mean-square error when one of them is randomized ranges from around 2% for PPI-HPRD to 12% for sequence similarity and functional annotations.

Second, highly correlated features, such as various Endeavor scores or state-of-the-art methods (CAROL with SIFT or PolyPhen) usually form clusters of seemingly less important features with a low yet nonzero increase in mean-square error. This is expected, as it has been shown that feature importance measures for Random Forests are strongly affected by the presence of correlation between features [18]. In the absence of a particular feature, other correlated features partially 'take over' the role of former, reducing the impact of the shuffling on the classification error. Hence, these variables are still individually very important—especially if data records contain missing values.

2.3 Supplementary material

2.3.1 Supplementary Note 1 – Data generation

Data sets of variants

Disease-causing variants

We acquired 24,454 nonsynonymous variants from the Human Gene Mutation Database (HGMD) [77] Professional version 2012.1. We manually mapped HGMD disease descriptors to OMIM disease terms. Using the Phenomizer tool [82], we mapped the OMIM disease descriptors to Human Phenotype Ontology (HPO) terms. Diseases that were mapped to multiple HPO terms were split into one record per HPO term. For example, "Miller syndrome" would be mapped to OMIM:263750 by searching OMIM, the Phenomizer tool would then translate this into individual phenotypes such as: cleft palate (HP:0000175), micrognathia (HP:0000347), hypoplasia of the radius (HP:0002984) and other HPO terms. A variant associated with Miller syndrome would then be associated to multiple records associated with the individual HPO terms. Records with HPO terms

with less than 3 known associated genes (not including the gene where the variant was located in) were discarded. Cancer-related HPO term containing records were also removed from the data set. In total, we retained records for 1,142 distinct HPO terms.

Common polymorphisms For every HPO term present in the disease-causing variant data set, we randomly selected 500 variants genomewide with a global minor allele frequency (MAF) greater than 1% in the 1000 Genomes Project [87] (n=43,724).

Rare genomic variation Analogous to the common polymorphism set, we selected 500 variants per HPO term but with a 1000 Genomes Project MAF lesser than 1% (n = 257,556). Using the same procedure, we generated a second rare variation data set based on an in-house data set of 68 in-house sequenced healthy control exomes of which the variants were unique to one individual, not present in any of the public variation databases (dbSNP, 1000Genomes, NHLBI Exome Variant Server) and sequenced at a coverage higher than 20-fold (n=25,429).

Variant features

Variant level features To every record, we appended Polyphen2 [51], SIFT [50], MutationTaster [55], and LRT [56] precomputed deleteriousness prediction scores extracted from dbNSFP [83] (v 1.3). Because of differences in the underlying tools, dbNSFP handles alternative splicing differently for each tool. SIFT and LRT do not use transcript information and are thus not insensitive to alternative splicing. For Polyphen2, the canonical transcript was chosen. In the case of MutationTaster, ANNOVAR was used to annotate the variant with transcript information from Ensembl. The first transcript ID was then submitted to MutationTaster. If the first transcript could not be processed by MutationTaster, the second ID was used. This process was repeated until MutationTaster was successfully run. Additionally, we computed an aggregated deleteriousness prediction score using CAROL [54]. We further added PhyloP [84] and PhastCons [85] 46-way conservation scores for vertebrate, placental mammals, and primate groups acquired from the UCSC genome browser [88].

Gene level features We added publicly available gene-wise haploinsufficiency prediction scores [76] where available. Finally, we computed genome-wide prioritization scores using Endeavour [58, 63] for each of the phenotypes using gene-phenotype associations obtained from the HPO. If the gene in which the variant was located was present amongst these known associations, this gene was excluded from the Endeavour training genes. This mimics the case where the gene itself is not known to be involved. The Endeavour data sources included sequence and protein domain similarity, molecular function, expression

data, protein-protein interactions, text mining, and pathway information. We appended a global prioritization score using order statistics, as well as a score per data source to each variant according to the phenotype-specific scores for each gene. A short description of the individual Endeavour data sources can be found in the Table 2.6.

2.3.2 Supplementary Note 2 - Performance measures and their interpretation

We use eight performance measures to compare classifiers based on genomic data fusion against state-of-the art variant prioritization scores. These are standard metrics that are routinely reported in classification benchmarking, and we provide them all to facilitate complete evaluation of the performance. However, we strongly believe that in context of variant prioritization some of them are more interesting than others, simply because of the nature of the problem.

Firstly, there are usually big costs associated with experimental verification of highly prioritized variants, so the good precision of a prioritization method is a desirable feature. Precision (or positive predictive value) is the proportion of truly positive variants among all classified as positive. At the same time, one wants to capture as much as possible of the real disease-causing variants, which renders the sensitivity of the method important as well. Sensitivity (or recall) captures the proportion of the real positives that are classified positive by an algorithm, so it essentially gives information on the expected number of the real disease-causing variants that cannot be detected. Note that eXtasy outperforms all standard methods with regards to these two measures.

Secondly, one might be interested in the relative scores produced by prioritization method, rather than in a binary classification based on a threshold of these scores. Downstream analyses are usually performed on the higher prioritized variants first, so the behavior of the scores that reside in vicinity of the top of the list is of great importance. Examination of the graphical measures, such are ROC and PR curves, can provide quantified insights into this behavior.

For example, from Figure 2.1 (B) it appears that a big proportion of the real positives in case of the eXtasy prioritizations reside in region of highly prioritized variants, in contrast with the state-of-the-art methods. The precision across different sensitivity values is not just bigger in absolute terms, it also declines later. This is an indicator of increased density of the real positives while moving towards the top of the list.

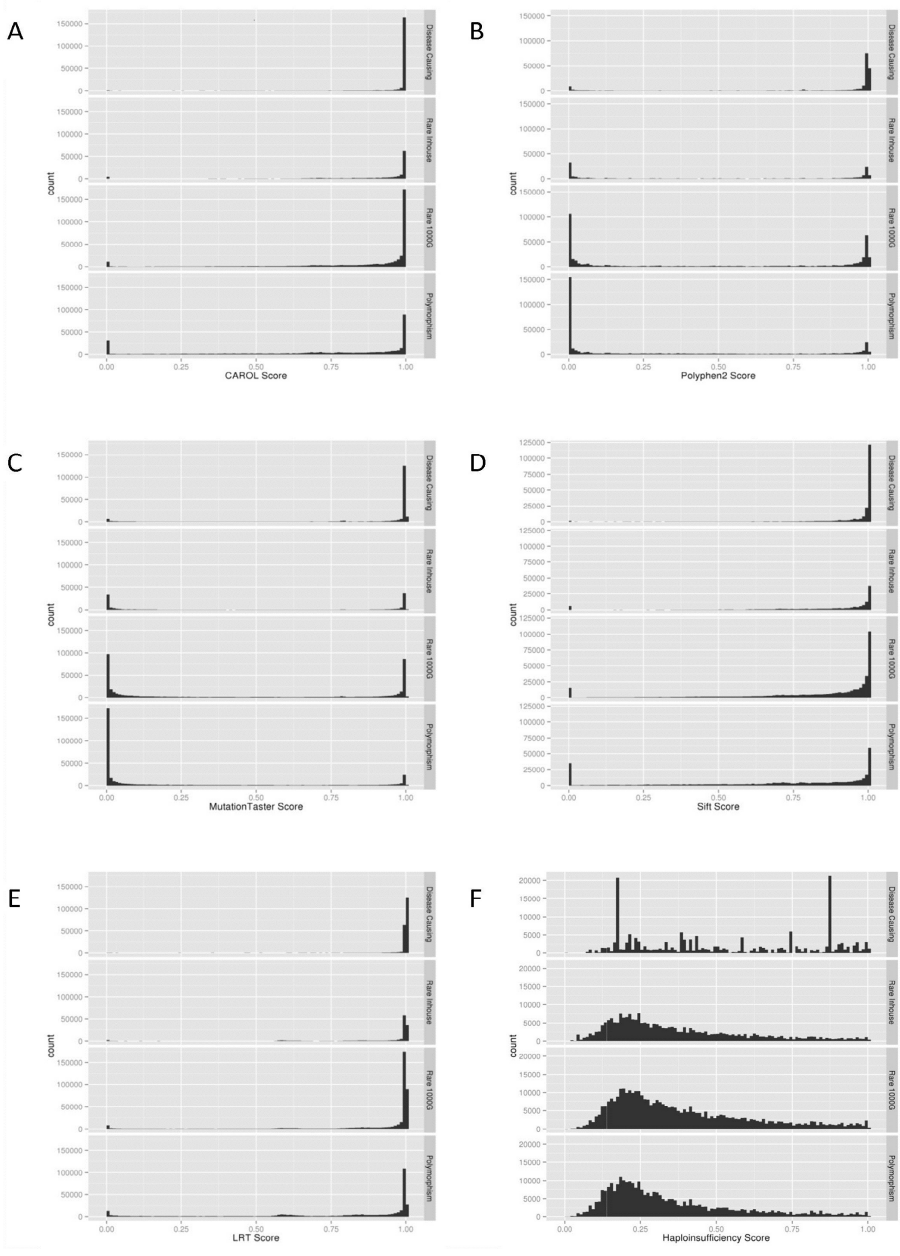
In our example, this means that one can expect better precision/recall than that reported for the hard-threshold classification, provided that variants are verified in descending order of prioritization scores. Also, this result suggests that the method is less sensitive than previous methods to increase in class imbalance. The latter is desirable property for practical application as the real class inbalance across the whole genome is much more severe than in any benchmarking data set.

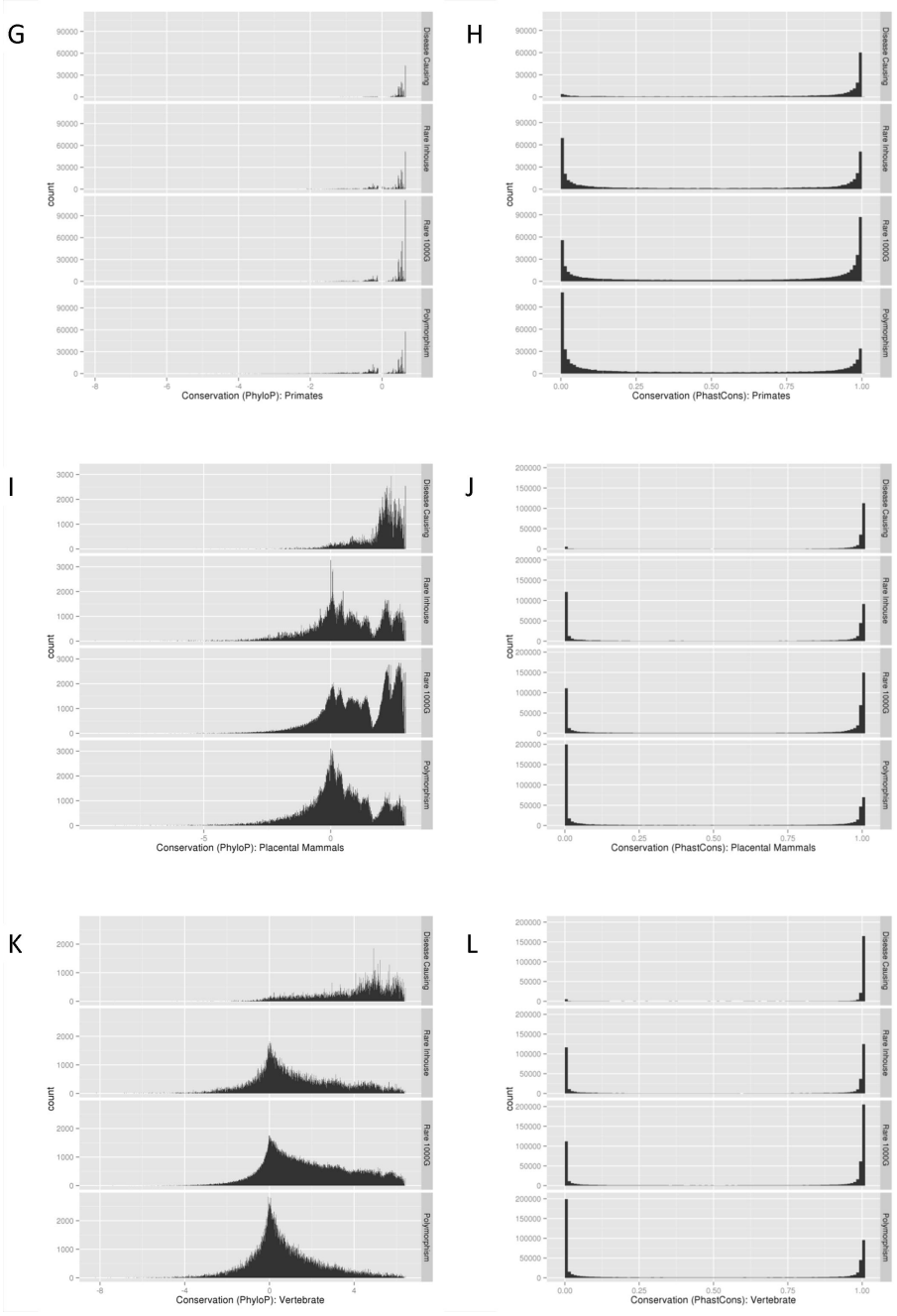
Finally, the aggregates of graphical measures also provide additional views on prioritization performance. For example, the area under the ROC curve could be, and often is, interpreted as the probability that a real positive example will be ranked higher than a negative. That is, given two examples, eXtasy will prioritize a real positive over a real negative in 97% of the cases.

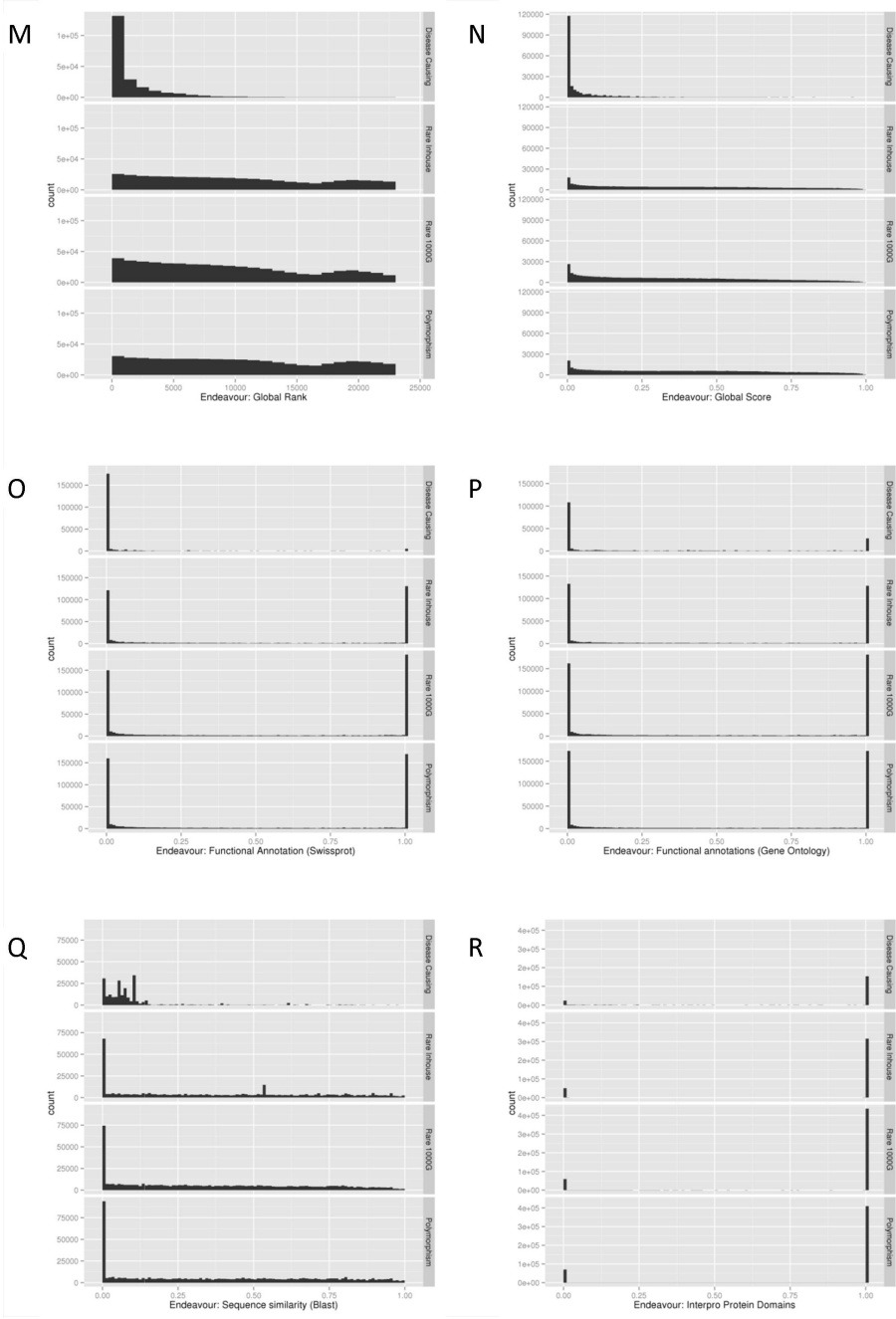
Given the performance measures reported in Supplementary Table 1 and that an average human genome harbors roughly 9000 nonsynonymous mutations [53] and assuming the disease of interest is caused by a single mutation we expect the real mutation to be found in 87% of the times using eXtasy (given the sensitivity. Out of the 9000 mutations and false positive rate of about 5% (FPR = 1-specificity) we expect to call roughly 450 non-disease causing variants as being causative. These numbers correspond to a single point in the ROC curve where we set the decision threshold for eXtasy scores at 0.5, one can easily reduce the FPR by shifting this decision threshold with only a moderate loss in sensitivity (due to the steepness of the ROC curve).

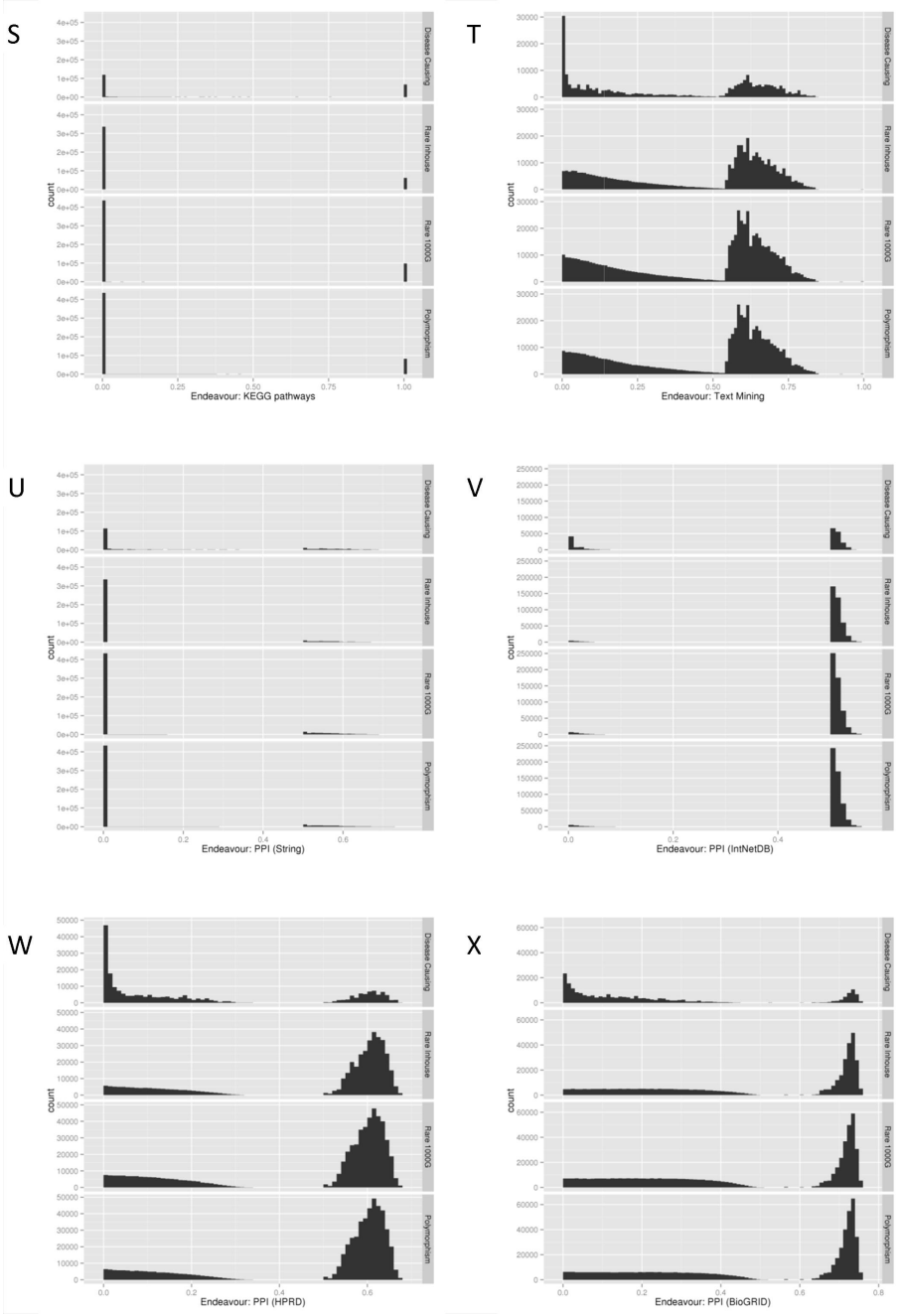
Using the published thresholds for SIFT, Polyphen2, MutationTaster and CAROL we would correctly call the disease-causing variant respectively 71%, 78%, 84% and 80% of the time while facing 3150, 4050, 4500, 4050 false positive calls. Compared to MutationTaster, which is the best performing deleteriousness prediction tool, eXtasy achieves a 10-fold reduction of false positives. Although the candidate variant set will be ranked according to their probability of being disease-causing, additional manual inspection and filtering of population databases, expected zygosity, or other samples of the same phenotype can provide further means of prioritization.

2.3.3 Supplementary figures and tables




G
H
I
J
K
L
count
Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
Conservation (PhyloP): Primates
Conservation (PhastCons): Primates
Conservation (PhyloP): Placental Mammals
Conservation (PhastCons): Placental Mammals
Conservation (PhyloP): Vertebrate
Conservation (PhastCons): Vertebrate


M
N
O
P
Q
R
Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
Endeavour: Global Rank
Endeavour: Global Score
Endeavour: Functional Annotation (Swissprot)
Endeavour: Functional annotations (Gene Ontology)
Endeavour: Sequence similarity (Blast)
Endeavour: Interpro Protein Domains


S
T
U
V
W
X
Disease Causing
Rare Inhouse
Rare 1000G
Polymorphism
count
Endeavour: KEGG pathways
Endeavour: Text Mining
Endeavour: PPI (String)
Endeavour: PPI (IntNetDB)
Endeavour: PPI (HPRD)
Endeavour: PPI (BioGRID)

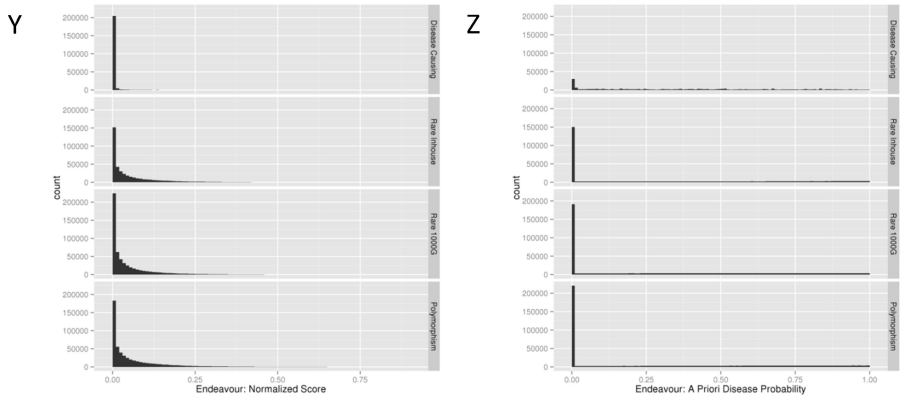


Figure 2.9: **Feature distributions.** Plots for the distributions of each of the used features are given for each of the data sets (the positive disease-causing variant set and the 3 control sets). Variables plotted are :
A) Carol score
B) Polyphen2 score
C) MutationTaster score
D) Sift score
E) LRT score
F) Haploinsufficiency scores
G) Conservation (PhyloP) for primates
H) Conservation (PhyloP) for placental mammals
I) Conservation (PhyloP) for vertebrate
J) Conservation (PhastCons) for primates
K) Conservation (PhastCons) for placental mammals
L) Conservation (PhastCons) for vertebrate
M) Endeavour: Global rank
N) Endeavour: Global score
O) Endeavour: Functional Annotation: Swissprot
P) Endeavour: Functional Annotation: Gene Ontology
Q) Endeavour: Sequence similarity (Blast)
R) Endeavour: Interpro Protein Domains
S) Endeavour: KEGG Pathways
T) Endeavour: Text Mining
U) Endeavour: Protein-Protein Interactions: String
V) Endeavour: Protein-Protein Interactions: IntNetDB
W) Endeavour: Protein-Protein Interactions: HPRD
X) Endeavour: Protein-Protein Interactions: BioGRID
Y) Endeavour: Normalized score
Z) Endeavour: A Priori Disease Probability score

Classifier/score	**Accuracy**		**Sensitivity**		**Specificity**		**PPV**		**NPV**		**MCC**		**aROC**	**aPR**
	mean	*std.*	*mean*	*std.*	*mean*	*std.*	*mean*	*std.*	*mean*	*std.*	*mean*	*std.*		
Random forest	**0.9402**	0.0045	0.8725	0.0298	**0.9496**	0.0060	**0.7112**	0.0309	0.9815	0.0033	**0.7545**	0.0260	**0.9729**	**0.8929**
Decision trees	0.8745	0.0075	0.8695	0.0311	0.8749	0.0103	0.4977	0.0341	0.9794	0.0042	0.5957	0.0293	N/A	N/A
FF Neural Network	0.8547	0.0088	0.9291	0.0163	0.8439	0.0115	0.4590	0.0326	0.9883	0.0021	0.5877	0.0251	0.9541	0.7711
Logistic regression	0.8118	0.0090	0.9410	0.0145	0.7931	0.0120	0.3934	0.0304	0.9896	0.0022	0.5300	0.0233	0.9509	0.7856
k-Nearest neighbours	0.7882	0.0098	0.9272	0.0207	0.7681	0.0133	0.3631	0.0285	0.9868	0.0033	0.4929	0.0238	N/A	N/A
LDA	0.7718	0.0098	**0.9526**	0.0134	0.7458	0.0126	0.3483	0.0289	**0.9911**	0.0021	0.4865	0.0229	0.9474	0.7707
QDA	0.7814	0.0111	0.9106	0.0170	0.7627	0.0148	0.3539	0.0289	0.9837	0.0026	0.4764	0.0229	0.9216	0.6507
Naive Bayes	0.7565	0.0118	0.9141	0.0152	0.7338	0.0156	0.3289	0.0269	0.9837	0.0025	0.4497	0.0214	0.8988	0.5169
Polyphen score	0.6578	0.0041	0.7119	0.0207	0.6499	0.0029	0.2251	0.0251	0.9406	0.0061	0.2446	0.0218	0.7499	0.3013
SIFT score	0.5779	0.0049	0.7855	0.0165	0.5482	0.0043	0.1989	0.0203	0.9470	0.0075	0.2204	0.0145	0.7180	0.2323
Mutation Taster	0.5457	0.0076	0.8482	0.0196	0.5024	0.0051	0.1959	0.0215	0.9588	0.0059	0.2326	0.0200	0.7647	0.3390
Carol score	0.5863	0.0052	0.8080	0.0112	0.5545	0.0037	0.2058	0.0215	0.9529	0.0055	0.2396	0.0154	0.7602	0.3187

Table 2.1: **Performance of the classifiers on the benchmark data.** Accuracy, sensitivity, specificity, positive predictive values (PPV also referred to as precision), negative predictive values (NPV), Matthews correlation coefficients (MCC), area under the ROC (aROC) and PR (aPR) curves of 8 classifiers based on genomic data fusion and 4 state-of-the-art methods expressed in terms of the corresponding means and standard deviations across 100 training vs. test splits of the benchmark data. The best values of the performance measures across classifiers are indicated in bold

Method	**Area under the ROC curve**		**Area under the PR curve**	
	Rare non-disease causing	Common polymorphisms	Rare non-disease causing	Common polymorphisms
eXtasy	**0.9691**	**0.9616**	**0.8744**	**0.9244**
Polyphen score	0.7516	0.8504	0.3040	0.6888
SIFT score	0.7207	0.7974	0.2351	0.5275
Mutation Taster	0.7659	0.8774	0.3433	0.7350
Carol score	0.7627	0.8476	0.3199	0.6956

Table 2.2: **Performance comparison of different control sets.** The performance is expressed in terms of area under the ROC and PR curves. The best values of a performance measure across scores and scenarios are indicated in bold.

Data granularity level	**Training years**	**Testing years**										
	1980-2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Variants	5658	156	164	294	261	400	304	339	357	465	533	505
Genes	292	46	55	87	82	98	117	115	122	148	174	194

Table 2.3: **Number of disease-causing variants and corresponding genes by year.**Numbers of disease-causing variants and corresponding unique genes in training and testing sets, after removing these associated with genes present in the training. The number of control variants per year is proportional to the number of disease-causing variants.

Method	**Year**										
	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
eXtasy	**0.9456**	**0.9656**	**0.9601**	**0.9088**	**0.9332**	**0.9170**	**0.8813**	**0.8685**	**0.8559**	**0.8460**	**0.8268**
	(0.0084)	(0.005)	(0.0039)	(0.0096)	(0.0059)	(0.0055)	(0.0069)	(0.0088)	(0.0074)	(0.0083)	(0.009)
Polyphen score	0.7243	0.7408	0.7342	0.7302	0.7374	0.7134	0.7325	0.7057	0.7561	0.7155	0.7146
	(0.0081)	(0.0088)	(0.0066)	(0.0068)	(0.0057)	(0.0067)	(0.0063)	(0.0063)	(0.0046)	(0.0046)	(0.0051)
SIFT score	0.7079	0.7195	0.6730	0.7265	0.6961	0.6848	0.6890	0.6768	0.7103	0.6727	0.6714
	(0.0099)	(0.0094)	(0.0071)	(0.0073)	(0.0055)	(0.0076)	(0.0059)	(0.0069)	(0.0054)	(0.0047)	(0.0053)
Mutation Taster	0.7634	0.7141	0.6590	0.7554	0.6514	0.7250	0.7334	0.7103	0.7425	0.7107	0.7303
	(0.0087)	(0.0091)	(0.0066)	(0.0074)	(0.005)	(0.0067)	(0.0068)	(0.0055)	(0.0043)	(0.0048)	(0.005)
Carol score	0.7436	0.7543	0.7280	0.7586	0.7459	0.7230	0.7356	0.7076	0.7579	0.7108	0.7150
	(0.0086)	(0.0091)	(0.0064)	(0.0066)	(0.0056)	(0.007)	(0.0058)	(0.0063)	(0.0047)	(0.0046)	(0.005)
eXtasy	**0.7840**	**0.7879**	**0.7712**	**0.6386**	**0.7027**	**0.7007**	**0.6077**	**0.6133**	**0.5553**	**0.5347**	**0.4869**
	(0.0276)	(0.0304)	(0.0233)	(0.0237)	(0.0258)	(0.0166)	(0.0201)	(0.0173)	(0.0166)	(0.0187)	(0.0136)
Polyphen score	0.2717	0.2868	0.2663	0.2951	0.2900	0.2760	0.3085	0.2610	0.3278	0.2678	0.2793
	(0.0116)	(0.012)	(0.0092)	(0.0094)	(0.0092)	(0.0102)	(0.0098)	(0.0076)	(0.0092)	(0.0064)	(0.0073)
SIFT score	0.2347	0.2451	0.2099	0.2460	0.2301	0.2191	0.2223	0.2177	0.2355	0.2108	0.2147
	(0.0092)	(0.0087)	(0.0057)	(0.0063)	(0.0047)	(0.0063)	(0.0048)	(0.0061)	(0.0047)	(0.0038)	(0.0044)
Mutation Taster	0.3475	0.2659	0.2173	0.3160	0.2327	0.2963	0.3114	0.2863	0.3392	0.2794	0.3040
	(0.0179)	(0.0144)	(0.0071)	(0.0142)	(0.0072)	(0.0117)	(0.0121)	(0.0103)	(0.0117)	(0.0076)	(0.0079)
Carol score	0.3020	0.3216	0.2804	0.3262	0.3220	0.2922	0.3195	0.2694	0.3387	0.2794	0.2877
	(0.0126)	(0.015)	(0.0091)	(0.0102)	(0.009)	(0.0102)	(0.0105)	(0.0082)	(0.0082)	(0.007)	(0.0077)

Table 2.4: **Performance comparison by year of discovery.** Area under the ROC (upper part of the table) and PR (lower part of the table) curves for five prediction scores for disease causing variants published during eleven consequent years (2001-2011); with standard deviations (in brackets). The best value of a performance measure given a score and a year is indicated in bold.

Classifier name	**Matlab R2010a function/class**	**Parameters**
Random forest	TreeBagger class method predict	Number of trees in ensemble : 100 Number of variables to select at random for each decision split. : square root of the number of variables (default) Minimal number of observations in a leaf : 1 (default)
Decision trees	classregtree class, method eval	Pruning : on (default) Minimal number of observations in a leaf : 1 (default)
Feedforward Neural Network	Functions newff, train and sim	Number of hidden layers : 1 Number of neurons in the hidden layer: 200 Number of training epochs : 100
Logistic regression	Functions glmfit and glmval	Distribution : binomial Link : logit (default for binomial)
k-Nearest neighbours	Function knnclassify	Distance measure : Euclidean (default) Number of neighbours considered : 15
Quadratic Discriminant Analysis	Function classify	Type : quadratic
Linear discriminant analysis	Function classify	Type : linear
Naive Bayes	NaiveBayes class, Methods fit, predict and posterior	/

Table 2.5: **Classifier implementations and parameters.** Matlab functions or classes used for implementation of classifiers with corresponding parameters. Only the implementation-independent parameters of the classification methods are displayed, with the exception of logistic regression, LDA and QDA where Matlab-specific parameters are used to choose the classifier. All implementation-dependent parameters that are not enumerated in this table have been set to default values.

GO	This data source is based on associations between genes and functional attributes that describe molecular processes, biological functions and cellular compartments. Data comes from Gene Ontology.
SWISSPROT	This data source is extracted from the UniProt -TrEMBL/SwissProt database. In particular, the keywords that are associated to the genes by UniProt are collected and organized in a pseudo-ontology. These keywords describe in general the main function of the genes and their products.
INTERPRO	This data source is based on associations between genes and domains identified from the sequences of the associated gene products. These domains describe the functional elements of the proteins and are coming from the InterPro database.
KEGG	This data source is based on biomolecular pathways from the KEGG database. These pathways describe common molecular processes and list the genes playing a role in these pathways.
STRING	This data source comes from the STRING database and is organized as a gene network with edges connecting the genes that are functionally related (or predicted to be). STRING integrates several types of information including Protein-Protein Interaction (PPI) networks, gene expression data sets and text mining of the scientific literature.
BIOGRID	This data source is a PPI network and is based on the integration of several large-scale experimental data sets. It contains both direct physical interactions as well as indirect interactions, such as genetic interactions. This data comes from the BioGRID database.
INTNETDB	The IntNetDB database integrates various gene networks including PPI and functional networks into one global gene network that contains both known and predicted gene interactions.
HPRD	This data source is a manually verified human PPI network and is based on the integration of several large-scale experimental datasets. It contains both direct physical interactions as well as indirect interactions such as colocalization. This data comes from the BioGRID database.
SU et. al	This data set comes from a large scale analysis of the human transcriptome (Su et al., PNAS 2002). It was obtained by profiling gene expression from 91 human tissues, organs, or cell lines.
OUZOUNIS	This data set represents a priori disease probabilities (Lopez-Bigas et al., NAR 2004). It is based on the use of sequence features (e.g., sequence length, UTR length, number of introns, and intron length) and a statistical framework to discriminate the human disease causing genes from the rest of the genome.
BLAST	This data source is based on the sequence similarities between all human proteins. It is obtained by using NCBI Blast on the protein sequences from Ensembl.
TEXT	This data source associates genes with keywords found in publications associated with these genes (from GeneRIFs). This text-mining data was obtained using a framework derived from the TxtGate tool starting from MEDLINE abstracts.

Table 2.6: **Descriptions of the Endevour data sources.**

Chapter 3

Aggregation of prioritization scores across phenotypes

This chapter is submitted for publication in PLOS One as :

Sifrim A., Popovic D., Vermeesch J., Aerts J., De Moor B., Moreau Y.,**"Comparison of aggregation methods for multiphenotype exomic variant prioritization"**

Abstract

The identification of disease-causing genes in Mendelian disorders has been facilitated by the detection of rare disease-causing variation through exome sequencing experiments. These studies rely on population databases to filter a majority of the putatively neutral variation in the genome and additional filtering steps using either cohorts of diseased individuals or familial information to narrow down the list of candidate variants. Recently, new computational methods have been proposed to prioritize variants by scoring them not only based on their potential impact on protein function but also on their relevance given the available information on the disease under study. Usually these diseases comprise several phenotypic presentations, which are separately prioritized and then aggregated into a global score. In this study we compare several simple (e.g. maximum and mean score) and more complex aggregation methods (e.g. order statistics, parametric modeling) in order to obtain the best possible prioritization performance. We show that all methods perform reasonably well (median rank below 20 out of more than 8000 variants) and that the

selection of an optimal aggregation method depends strongly on the fraction of uninformative phenotypes. Finally, we propose guidelines as to how to select an appropriate aggregation method based on knowledge of the phenotype under study.

3.1 Introduction

In recent years, whole-exome sequencing has facilitated the discovery of genes underlying rare Mendelian disease by allowing the identification of disease-causing genomic variation [89, 90, 91, 92, 93, 94]. Several different strategies have been proposed to sift through the thousands of variants identified in a single human exome in order to retain a handful of meaningful candidate variants for further downstream functional validation studies [95].

These strategies usually start by filtering variants with a higher potential for disturbing gene function, such as gaining a stop-codon, resulting in a frameshift, changing the amino-acid sequence of the protein or interfering with a splice site. In this first filtering step intergenic, intronic and synonymous mutations are discarded. A second filtering step removes all variants present under a certain minor allele frequency (e.g. the variant occurs in less than 1% of the population) in commonly used population databases such as the 1000 Genomes Project [87] or the NHLBI Exome Sequencing Project [75]. This step relies on the assumption that variants causing rare disease, following either dominant or recessive inheritance patterns, would not be present in high frequencies in populations of healthy individuals.

Even after such filtering several hundreds of potential candidate variants remain and additional filtering is required. At this point strategies diverge depending on the available data. If familial samples are available, linkage analysis or searching for de novo variants can sufficiently narrow down the list to only a couple of variants [96]. Another strategy is to sequence a number of unrelated individuals exhibiting very similar phenotypical presentations and looking for common genes struck by variants in several of these individuals [90, 91]. This strategy usually assumes a monogenic inheritance pattern in order to have sufficient signal to overcome the noise of neutral genomic variation. Additionally, acquiring sufficient samples is difficult as patient recruitment usually has to happen across several clinical institutions due to the rarity of the disorder in any given population.

In order to assist in this process, several computational methods have been developed to predict the impact of a given variant based on various biochemical, structural and evolutionary properties [50, 55, 51]. Although these methods

provide reasonable performance in predicting variants affecting protein function, they lack the specificity to identify true disease-causing variants as the human genome is riddled with mildly deleterious variation unrelated to the specific phenotype under study [73, 97, 98]. In order to remedy this, new computational approaches were proposed which consider not only functional impact but also phenotypic relevance [71, 99].

In a recent study we proposed a method, named eXtasy, that integrates deleteriousness and haploinsufficiency prediction with phenotype-specific gene prioritization through genomic data fusion [57]. This approach relies on the similarity across several data sources (e.g. gene expression, protein-protein interactions, sequence similarity) between known phenotype-associated genes and the mutated gene [58, 63]. This process aids in discriminating variants in genes likely to be involved in the phenotype and variants in genes which might contain functional variants but for which there is no evidence of the genes being involved in the disease.

In this study we elaborate on the problem of aggregating across different prioritizations when the observed clinical presentation is an agglomeration of different phenotypes, as is often the case in rare Mendelian disorders. For example, Miller syndrome, one of the first Mendelian diseases elucidated by massively parallel exome sequencing [91], comprises a combination of craniofacial abnormalities, postaxial limb deformities and sometimes internal malformations.

It would be reasonable to assume that some of these phenotypes would be more informative than others when considering them for variant prioritization, as some of them might be better understood in the current literature, they might show less complex known underlying molecular mechanisms or due to the inherent phenotypic variability of the disease. Yet to estimate their relative informativeness would require extremely in-depth knowledge of these properties.

To circumvent this problem in eXtasy, we opted to use the maximum score obtained from any single variant prioritization for a given variant. This guaranteed that if one of the phenotypes delivered an informative prioritization this would be taken into account (increasing sensitivity of the method), yet potentially also inflated our scores of non-disease causing variants, leading to increased false positive rates. Using the maximum essentially discards information when more than one informative phenotype is used for prioritization as only the best phenotype is considered.

In this study we propose 3 alternative ways of aggregating over multiple phenotype-specific prioritizations using either classical order statistics [58, 100], robust rank aggregation [101] and statistical modeling of score distributions. To estimate and compare their performance we also adapt our benchmarking

scheme to resemble a more real-life application of variant prioritization. We conclude by showing that although in most cases the maximum outperforms the other aggregation methods, using more complex aggregation methods is usually the best when phenotypes are carefully selected. These methods also offer the added benefit of obtaining probabilistic outcomes which could be integrated further in other statistical frameworks, such as case/control association studies.

3.2 Material and Methods

3.2.1 Benchmark

In order to compare the performance of the different aggregation methods we set up a benchmark which closely resembles a real-life application of variant prioritization using whole exome sequencing. We first generated 10000 synthetic exomes by randomly sampling nonsynonymous variants from the 1000Genomes Project. For every nonsynonymous variant in the 1000Genomes Project (October 2012 release) a random number between 0 and 2184 (the maximum number of times an allele could be present in the 1092 diploid individuals) was generated and divided by 2184. If this randomly generated number was smaller than the observed global minor allele frequency for that variant, then this variant was included in the synthetic exome. We chose the Human Gene Mutation Database [77] as a set of disease-causing variants, and selected 16272 nonsynonymous variants which were suitable for variant prioritization and studying phenotype aggregation (having at least 3 different phenotypes with at least 3 known gene-phenotype associations each).

For each of these variants we randomly selected one of the 10000 synthetic exomes and injected that disease-causing variant into the exome. We then ran these synthetic exomes containing one disease-causing nonsynonymous mutation through eXtasy for each phenotype describing the disease. For each exome we aggregated the resulting scores of the different prioritization either using the maximum, mean, the median, classic and robust order statistics and a statistical modelling approach. We then obtained the rank of the injected disease-causing variant in the synthetic exome for each aggregation method (Figure 3.1). We additionally computed the fraction of informative phenotypes for a disease-causing variant as the number of phenotypes with an eXtasy score > 0.5 over the the total number of phenotypes prioritized. This threshold is the same at which a variant is considered disease-causing in the original eXtasy paper and corresponds to the majority vote of trees in the Random Forest model.

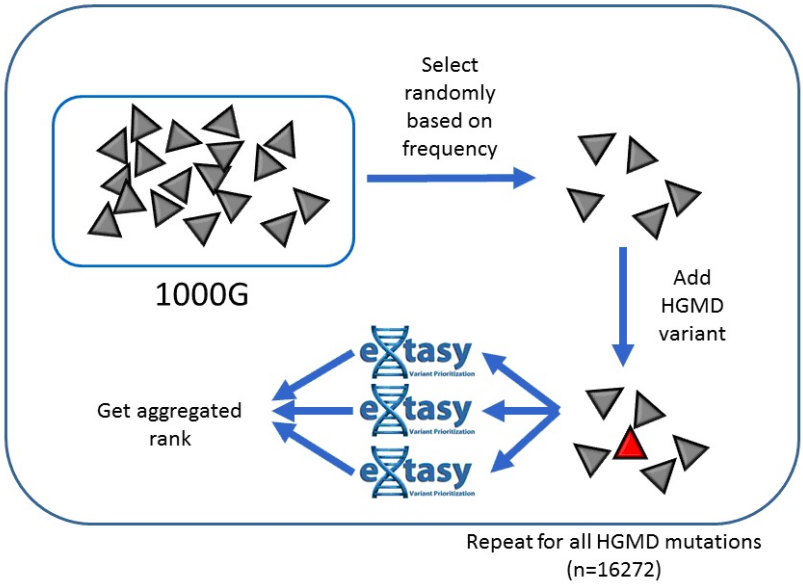


Figure 3.1: **Schematic representation of the benchmarking scheme.** Nonsynonymous variants are first sampled from 1000Genomes Project data based on their respective minor allele frequency, generating a synthetic control exome. Then, one disease-causing variant from HGMD is added to a single synthetic control exome and eXtasy prioritizations are performed for each phenotype annotated to the disease-causing variant in that exome. The different phenotype-specific prioritizations are then combined into a single score using different aggregation schemes and the ranks for the injected disease-causing variant is compared across the different methods. This process is repeated for each of the 16272 nonsynonymous HGMD variants.

3.2.2 Non-parametric Order Statistics

Nonparametric order statistics have been previously used to aggregate rankings of genes obtained from different experiments or data sources [58, 100]. These methods estimate the probability of a given combination of ranks under the null hypothesis that the ranks were randomly drawn. These methods have the benefit if being robust to non-informative rankings, for example adding an erroneous phenotype in the variant prioritization, without losing the added power of additional informative rankings (which are lost by aggregating using the maximum score). In an initial formulation of a gene rank aggregation algorithm, an algorithm was proposed which estimates the cumulative density function of the order statistic under the null hypothesis using a recursive formulation. This algorithm was computationally prohibitive for moderate numbers of ranked

lists and was further refined by Aerts et al. (2006) decreasing the algorithmic complexity and thus computational time. Both formulations assume the null hypothesis that all ranked lists are potentially informative.

Another computationally efficient formulation called Robust Rank Aggregation (RRA) was proposed [101] where a fraction of the ranked lists is allowed to be uninformative. In their benchmarks they showed that both RRA and the Aerts et al. formulation perform similarly but that the latter algorithm becomes unstable for large numbers of ranked lists (N > 40), but it is unlikely that such numbers of phenotypes would need to be integrated in variant prioritization experiments. In this study we use the R package (RobustRankAggreg) implemented by Kolde et al. which implements both algorithmic order statistic formulations and apply both of them to our benchmark to rank variants according to their estimated significance.

3.2.3 Parametric Modeling

Order statistic approaches are extremely generalizable as they do not make any assumptions of the underlying score distributions by transforming the data into rankings. Although this is extremely useful in many situations, such nonparametric approaches are usually less powerful than their parametric counterparts that take these distributions into account. While studying the eXtasy output score distributions we observed highly reproducible non-Gaussian density functions skewed towards zero and with long tails towards higher scores, which can be best fitted using a Gamma distribution model. In order to come up with an aggregate score across different eXtasy prioritizations we developed a parametrical statistical approach and applied it to our benchmark.

Assume n prioritizations N_i for $i = 1..n$ each containing a set of scores $s_{i,k}$ with $k = 1..K$ where K is the number of variants under study. For each N_i we robustly fitted a Gamma distribution Γ_i with parameters k_i and θ_i representing the shape and scale of the distribution using the robust package in R. We then computed the probability $p_{i,k}$ of seeing a score equally or more extreme as $s_{i,k}$ given the cumulative density function of Γ_i.

To aggregate these individual probabilities into a single probability of seeing a score as extreme as $s_{i,k}$ across the n different distributions we compute the Fisher's omnibus meta-analysis statistic $r_k = -2\sum_{i=1}^{n} \ln p_{i,k}$ [102]. This allows us to compute the global probability of seeing such a combination of eXtasy scores by computing the probability of r_k given the cumulative density function of a χ^2 distribution with $2n$ degrees of freedom. Because of the large number of variants tested we correct these probabilities using Benjamini-Hochberg False

Discovery Rate correction. We then rank the variants for each synthetic exome in our benchmark given the aggregated multiple testing corrected p-value.

3.3 Results

In contrast with the previously reported eXtasy case-control classification benchmark, our goal with the synthetic exome benchmark was to simulate a real-life situations where variant prioritization would be used and resembles the validation benchmark proposed by Robinson et al. [99]. Our exome generation algorithm generated exomes containing an average of 8399 nonsynonymous variants (n=10000, sd=242).

This number is higher than the roughly 6000 variants reported from large-scale exome sequencing projects [75], but is in line with whole-genome sequencing studies [87, 53] which do not face the same difficulties that exome-capturing techniques due to coverage variability. As an indication of real-life performance we describe the rank of the injected disease-causing variant as a performance metric, as opposed to classical metrics (e.g. precision/recall), to compare aggregation schemes for multiphenotype variant prioritization using eXtasy.

Overall, we observe that all aggregation methods perform reasonably well (Table 3.1), having a median rank for the disease-causing mutation in the top 20 and a below the top 100 in 75% of the cases (with the exception of the median eXtasy score and RRA). Looking at the cumulative distribution functions of the ranks of the different methods (Figure 3.2), we can see that using the maximum eXtasy score (as used in the original eXtasy publication) delivers the best rankings compared to the other methods. Aggregation using the average eXtasy score or the more involved parametric and nonparametric methods seem to show similar performance.

Finally, RRA and the median eXtasy score provide the worst performance out of all the aggregation schemes. These results indicate that uninformative or incorrectly prioritized phenotypes might affect the global ranking significantly, which might explain why using the maximum score (being the most robust method) generally outperforms other aggregation schemes. In order to further investigate this hypothesis, we looked at how varying degrees of uninformative phenotypes affect the different aggregation schemes by stratifying our benchmark according to the fraction of individual phenotypes correctly prioritized.

After stratification into low (n=5862), moderate (n=8559) and high (n=1837) fractions of informative phenotypes we observe that no single aggregation method performs best across the three classes (Figure 3.3). When only less than

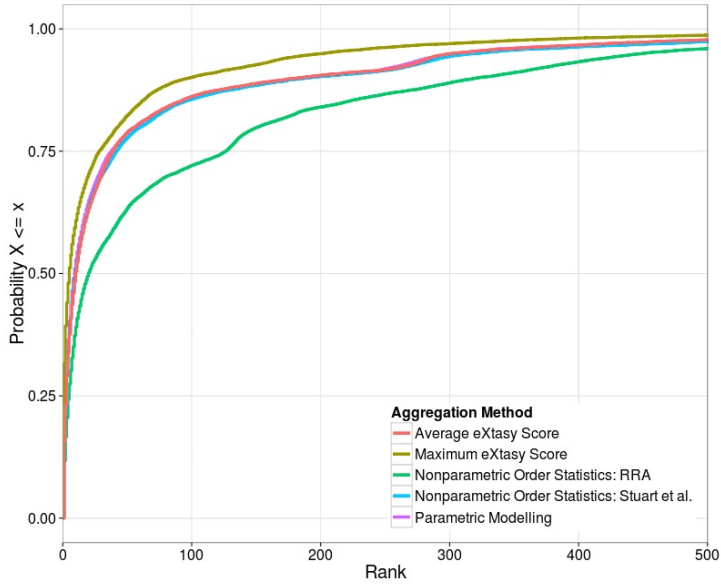


Figure 3.2: **Empirical Cumulative Distribution Function (ECDF) of ranks comparing different aggregation schemes.** This chart shows the proportion of variants out of all disease-causing variants (n=16272) with global ranks after aggregation lower or equal than the rank shown on the x-axis.

Aggregation Method	Q1	Median	Mean	Q3
Average eXtasy Score	3	10	59	39
Maximum eXtasy Score	1	5	41	28
Nonparametric Order Statistics: Stuart et al.	3	10	64	42
Nonparametric Order Statistics: RRA	5	20	102	126
Parametric Modeling	2	9	111	37

Table 3.1: **Summary statistics for the overall comparison of different aggregation methods.** This table shows the mean, median, 25%-quantile (Q1) and 75%-quantile of the global ranks of disease-causing variants (n=16272) using different aggregation schemes.

a third of phenotypes are informative, the average rank of the disease-causing variants degrades significantly independently of the aggregation method (Table 3.2). In this scenario, aggregation using the maximum eXtasy score delivers the best ranking compared to all other methods, while all other methods show nearly equal performance (with the exception of RRA, showing significantly lower performance).

When roughly halve the phenotypes are informative, all aggregation methods (except RRA) perform reasonably well (average rank lower in 10, and 75% of

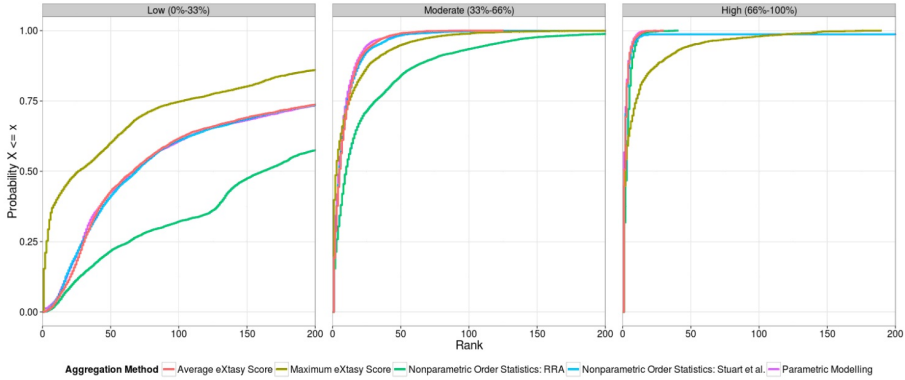


Figure 3.3: **Empirical Cumulative Distribution Function (ECDF) of ranks by fraction of informative phenotypes comparing different aggregation schemes.** This chart shows the proportion of variants out of all disease-causing variants for fractions of informative phenotypes divided into 3 classes: Low (n=5862), Moderate (n=8559) and High (n=1837) with respectively 0%-33%, 33%-66% and 66%-100% of phenotypes showing an eXtasy score > 0.5 for the prioritized disease-causing variant.

variants ranked below 12). Additionally, methods which leverage information across different phenotypes (mean score, order statistics, parametric modelling) perform slightly better than aggregation using the maximum score. In the case where more than two-thirds of the phenotypes are informative 75% of the disease-causing variants are in the top 3 ranked mutations for all methods except aggregation by maximum score (75% of variants in the top 10).

3.4 Discussion

In this study we investigate the performance of different aggregation schemes when dealing with multiphenotype exomic variant prioritization. We benchmark these methods mimicking a real-life application by generating synthetic healthy exomes and injecting single known disease-causing nonsynonymous mutations from HGMD. By comparing different aggregation schemes, ranging from simple metrics to more involved nonparametric and parametric methods, we show that no single method is optimal for all cases and that performance depends on the expected fraction of informative phenotypes.

Not unsurprisingly, the optimal aggregation method is dependent on the signal-to-noise ratio originating from the underlying individual prioritizations. Any prioritization for a specific phenotype can add noise to the global prioritization

Class	Aggregation Method	Q1	Mdn.	Mean	Q3
Low (0-33%)	Average eXtasy Score	31	66	152	218
	Maximum eXtasy Score	3	26	95	102
	Nonparametric Order Statistics: Stuart et al.	30	69	158	224
	Nonparametric Order Statistics: RRA	64	162	243	347
	Parametric Modelling	29	67	295	222
Moderate (33-66%)	Average eXtasy Score	2	5	8	12
	Maximum eXtasy Score	1	3	11	12
	Nonparametric Order Statistics: Stuart et al.	2	5	10	11
	Nonparametric Order Statistics: RRA	3	10	26	31
	Parametric Modelling	1	5	8	10
High (66-100%)	Average eXtasy Score	1	1	2	3
	Maximum eXtasy Score	1	2	11	10
	Nonparametric Order Statistics: Stuart et al.	1	1	11	3
	Nonparametric Order Statistics: RRA	1	3	3	5
	Parametric Modelling	1	1	2	3

Table 3.2: **Summary statistics for the comparison of different aggregation schemes by fraction of informative phenotypes.** Analogous to Table 1, this table shows summary statistics for the ranks of disease-causing mutation but after stratification into 3 classes representing different fractions of informative phenotypes: low (0%-33%), moderate (33%-66%) and high (66%-100%) number of informative phenotypes showing an eXtasy score > 0.5 for the prioritized disease-causing variant.

for two reasons. First, the specific phenotype might not be biologically well-suited to describe the disease under study. This can for example be attributed to phenotypic variability of the disease (e.g. some of the phenotypes might not always be present in all patients showing the disease). This problem can relatively easily be managed by clinical knowledge of the disease, and thus be avoided in real-life applications. In our benchmark this is more difficult to address due to the sheer number of included diseases/phenotypes where disease terms are linked to phenotypic ontologies automatically.

Secondly, the gene prioritization step can be noisy due to the selected set of training genes. For example, certain phenotypes could be overly general (e.g. intellectual disability, abnormality of the face) and have many heterogeneous underlying molecular mechanisms. Using all known genes in the gene prioritization step for such phenotypes results in uninformative prioritizations due to a dilution of the signal where similarity is estimated to an extremely diverse group of genes and thus leads to reduced performance.

The inverse problem also exists where too little is known about the phenotype, biasing the prioritization towards certain training genes which might not necessarily be similar to the as-of-yet undiscovered causative gene. Predicting a priori the performance of the gene prioritization is difficult and requires a good understanding of the underlying data being used.

Because of these problems, it is important to use aggregation methods which can handle these uninformative or erroneous prioritizations robustly. Aggregating using the maximum score is a simple and intuitive robust solution and performs well in our previous and current benchmark, especially in low signal-to-noise situations.

The drawback of using the maximum is that it ignores information from additional phenotypes, which could be quite powerful in increasing the performance of a global prioritization. This is indicated by the better performance of methods which take advantage of these phenotypes (mean, order statistics and parametric modeling) when better signal-to-noise ratios are observed.

It is important to remark that uninformative phenotypes are likely overrepresented in our benchmark, compared to real-life applications, because of the automatic mappings generated between disease-variants and phenotypic ontologies and the lack of clinical expertise for each evaluated disease. Considering this, it is likely that using these methods will result in better variant prioritizations if the phenotypes are carefully selected.

Here we show that choosing a suitable aggregation method in multiphenotype variant prioritization can affect the results of the resulting global prioritization. We demonstrate this by using simple aggregation methods such as the maximum score and the mean score. Additionally we show that more involved nonparametric and parametric statistical methods offer similar performance with the added benefit of providing estimates of significance which could in future studies could easily be used in broader statistical frameworks such as those used in weighted association or familial studies [103, 80].

We do remark that these methods assume independence of the individual prioritizations, an assumption which is likely not true due to the usage of phenotype-aspecific data sources in eXtasy. This might not greatly influence their performance in a ranking scenario but result in biased estimates of significance. Several methods have been proposed which handle these dependencies [104, 105] but were not considered in this study due to the unavailability of usable implementations, but should be considered when true statistical significance is of importance.

Although currently we rely on the user to select an appropriate aggregation method given his experience on how well-characterized the phenotype under study is. In future work we hope to partially or fully automate this process by trying to evaluate training gene set heterogeneity and gene prioritization performance through cross-validation. By gradually improving variant prioritization performance and usability we hope to aid clinical geneticists

in discovering disease-causing mutations for their respective phenotypes of interest.

We currently offer all the proposed aggregation schemes in the current version of eXtasy (http://homes.esat.kuleuven.be/~bioiuser/eXtasy/). We also publicly provide the benchmark data for further development or benchmarking of other score/rank aggregation methodology (file aggregation.tar.gz, available on the same website).

Chapter 4

Interpretation of the model

This chapter has been previously published as :

Popovic D.*, Sifrim A.*, Moreau Y., De Moor B., **"eXtasy simplified - towards opening the black box"**, in Proc. of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shanghai, China, Dec. 2013, pp. 24-28
* *Equally contributing first authors*

Abstract—Exome sequencing remarkably simplifies the search for mutations causing rare monogenic disorders. Still, due to a big number of potential candidate variants, computational methods are needed to facilitate this process. Recently, an algorithm based on genomic data fusion has been proposed in this context (eXtasy), which exhibits highly competitive performances among the state of the art methods. Nonetheless, being based on a Random Forest classifier, its core model is characterized by a prohibitive size, slow execution speed and difficulties associated with gaining insights in the decision-making process. Here we propose a simplification of the original eXtasy algorithm that retains superior ranking capability of former without suffering from the both high complexity and low interpretability.

Keywords—eXtasy, variant prioritization, genomic data fusion, rare genetic disorders, interpretable model, random forest, decision trees, hybrid sequential system

4.1 Introduction

The discovery of mutations involved in the etiology of genetic disorders has been significantly accelerated in the recent years due to advances in massively parallel sequencing technologies. Amongst other approaches, sequencing of the exome (i.e. protein-coding region of a gene) in particular has been shown to be an efficient strategy for identification of causes of rare monogenic disorders [94]. However, a single exome typically harbors approximately 8000 non-synonimus variants, while DNA samples from affected individuals are often scarce.

Even after aggressive filtering of the exome against nSNVs (nonsynonymous single nucleotide variants) and loss-of-function mutations that are present in healthy individuals, roughly 200 candidate mutations still remain [53, 96]. To overcome this problem, several computational methods for variant prioritization have been developed so far [56, 50, 55, 51, 54, 73, 98]. These methods mostly rely on evolutionary, biochemical and structural properties of the mutation in question. Recently, a new approach based on genomic data fusion has been proposed (eXtasy [57]) that displays vastly superior performance compared to the rest of the current state-of-the art methods.

eXtasy combines phenotypic-specific information, haploinsufficiency prediction and deleteriousness prediction scores of mutations to deliver variants ranking. Firstly, given the list of phenotypes associated with a disease of interest, the mutated genes are scored based on their similarity with known disease genes [58]. Secondly, resulting vectors of scores obtained per each phenotype/variant combination are augmented with several additional features : haploinsufficiency prediction [76], state-of-the art deleteriousness prediction scores [56, 50, 55, 51, 54] and conservation scores across several species [84, 85]. Finally, this information is used to classify a mutation as disease-causing and to assign a score to it. As several values are obtained for each mutation (in context of associated phenotypes), only the maximum score is retained.

The core of the eXtasy algorithm is the Random Forest classifier [20] trained using the data corresponding to known disease-causing mutations (positives) and to rare mutations found in healthy individuals (negatives). The Random Forest is an ensemble method, so it combines prediction of many base models to reach a final decision. By doing so, it stabilizes the variance of low-bias/high-variance base learners (i.e. decision trees), improving overall performance. Randomness is injected into the algorithm by building each tree on a different bootstrap from data and by choosing each decision tree split from a random subset of all variables. This procedure increases the diversity of an ensemble, which is crucial for triggering the previously discussed effect. In addition, it has been shown that ensemble approaches have a smaller number of generalized degrees of freedom

than some supposedly simple algorithms [106], making their predictions quite robust against small perturbations in the data.

However, these advantages come with the price of increased model complexity and reduced interpretability. The complexity translates to limitations associated with memory constraints and execution speed in practical applications. For example, the eXtasy stand-alone application comes with the model that is 100 Mb large (for 500 trees) and takes approximately 2 minutes to score all mutations from a single exome using a standard PC. Furthermore, being based on a Random Forest algorithm, the decision process of eXtasy takes place in a "black box", which prevents users from gaining insight into it.

An attempt has been made to address this issue in the original study by using the Random Forest feature importance measures [20]. Yet, these are extensively criticized for displaying strange statistical properties [107]; and for being highly biased in presence of correlation between variables [108, 109] or when data of mixed types is used [17]. Therefore, in this work we propose a straightforward simplification of the eXtasy model, aiming to address previously risen concerns. In addition, we demonstrate that the described model achieves similar performance as the original formulation, while being more interpretable, having a smaller memory footprint and faster prediction speed.

4.2 Materials and methods

4.2.1 The method

The basic idea behind our approach is usage of a simple classifier to model the decision boundary that is implicit in a more complex one (i.e. Random Forest). In this setup, two algorithms are organized sequentially, where the first learns from the data and then acts as a "teacher" to the second. Provided that an inductive bias of a simple method allows the construction of a frontier of the same type as that generated by the complex method, this procedure ideally results in preserving advantageous performances of the later in a more compact and comprehensible form. The described paradigm was already successfully employed in several different settings; amongst which to extract symbolic rules from neural network [110, 111] and decision trees from bagged ensemble [112]. Our approach is somewhat similar to that in [112], with the exception of a different synthetic data generation scheme.

The general outlook of the algorithm is provided on Figure 4.1. First, a Random Forest classifier is trained using class-balanced data, identical to the way the eXtasy model was built [57]. Then, artificial data is generated by sampling from

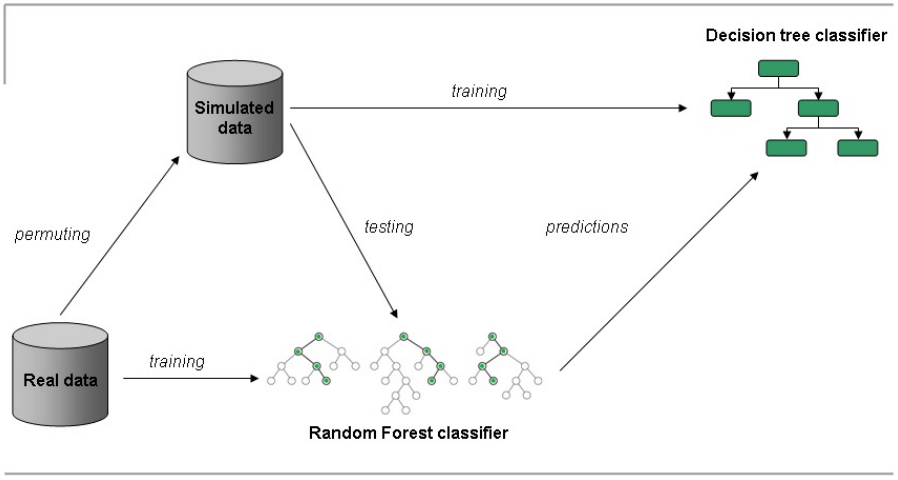


Figure 4.1: **Sequential algorithm for training the simplified eXtasy.** The real data is used as an input for training the Random Forest classifier (as in the classical eXtasy) and to generate artificial data that follows empirical distribution of the former. The output of the first classifier together with corresponding simulated examples is passed to the decision tree for the second round of training

the empirical distributions of the features. In particular, values of predictors are shuffled across examples. This randomization is repeated ten times to create a large data set with distributional properties similar to that of the real data. Empirical sampling has been chosen over uniform and model-based sampling due to a high skew of predictor values and a lack of fit to any standard distribution. Subsequently, simulated data is subjected to classification by the previously trained Random Forest, assigning labels to each generated example. Finally, the new data set is used to build a fully-grown decision tree. This model is deliberately not pruned, as the goal is to "overfit" the already generalized decision boundary of the Random Forest.

A single decision tree is used as it can represent function of arbitrary complexity over the instance space, analogously to the Random Forest. Also, the interpretation of a single tree decision process is quite straightforward compared to that of an ensemble system. However, it might not be immediately obvious how the splits of such a derived tree differ from that inferred directly from real data, and consequently, how positions of these in a tree indicate importance of corresponding variables. A graphical illustration of this difference on a toy example is displayed on Figure 4.2.

Briefly, decision tree induction algorithms are based on greedy, top-down

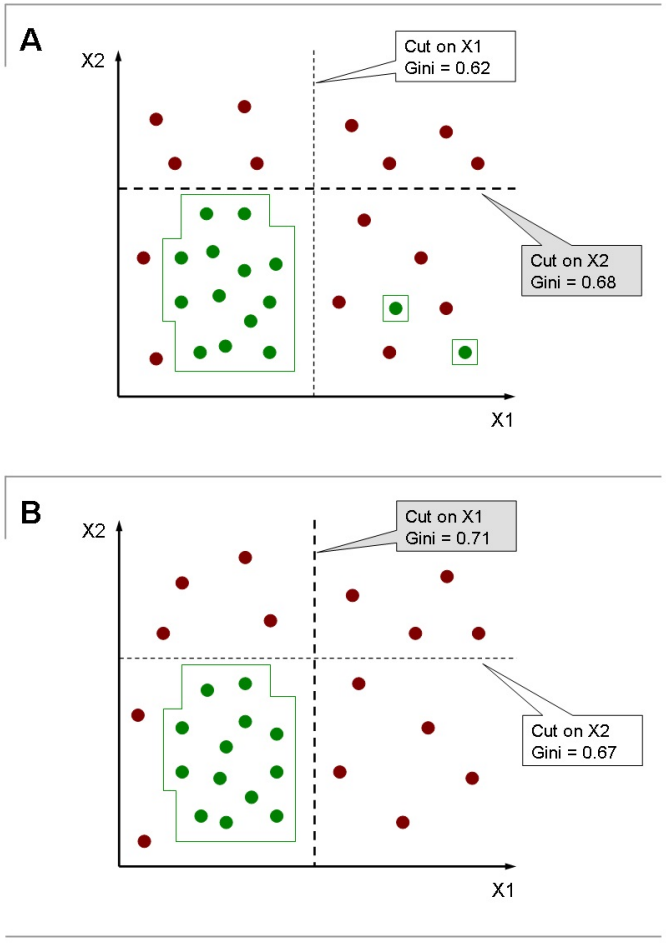


Figure 4.2: **An example of two different initial decision tree splits for the same classification problem.** Squares and dots represent two classes; solid lines enclose implicit general concept while two empty squares indicate outliers. A) Variable X2 is chosen for the first split due to a higher Gini index. B) Data from the same distribution, but without outliers. Now the X1 is chosen for the split. In both cases the decision boundary can be learned by a single decision tree, but the ordering of variables changes as well as their relative importance in describing the underlying concept.

recursive partitioning, thus their first (or any other) split might not be the most optimal with respect to the boundary constructed by the Random Forest. For example, a few outliers can render one variable more informative than others at a moment of choosing the optimal (current) split; even subsequent

nodes generated using these instances would be pruned off later in a process. This effectively disregards them as non-informative during the validation, yet they bias the choice of variables during the tree induction phase. Conversely, outliers are routinely averaged out by bagging when the Random Forest decision boundary is build; thus they can not influence order in which variables are selected when a decision tree is used to model this already learned underlying concept.

The proposed algorithm is tested against the original eXtasy formulation and a classical decision tree classifier in the following manner – the whole dataset has been divided into training and testing partitions, with two-thirds of the total number of genes in the first, and one-third in the second part. Note that one gene contains many variants, while each variant has several corresponding data records (i.e. phenotypes). The validation scheme relies on stratification of the data on the highest level of granularity to prevent algorithms from overfitting gene-level information (i.e. to learn to recognize particular genes). After splitting, the class distribution of a training part is balanced, followed by training of all three methods on the same examples. In addition, the identical Random Forest classifier has been used both to access performance of the standard eXtasy and to generate synthetic data for the simplified version. To stabilize performance metrics, this process has been repeated one hundred times, randomizing the data division each time.

4.2.2 The data

The data set consists of two classes of mutations: Mendelian disease-causing variants and rare mutations present in healthy individuals as controls. For the disease-causing variants we obtained 24,454 nSNVs from the Human Gene Mutation Database (HGMD) associated in 1142 different Human Phenotype Ontology (HPO) terms. The HGMD is a database of expertly curated disease mutations published in scientific literature. Control variants were obtained from two different sources: the 1000 Genomes Project [87] and inhouse sequenced exomes (n=68) of healthy individuals. We selected nonsynonymous variants from the 1000 Genomes Project with a minor allele frequency lower than 1% (n=257556).

For the inhouse variants we selected variants not present in any publicly available variation repository (1000Genomes, dbSNP, NHLBI Exome Variant Server) and imposed a quality criterium of being sequenced at a depth of at least 20 (n=25429). For each of the phenotypes in the disease-variant set we sampled 500 variants from the pool of control variants and assigned them that given phenotype. Although the rare variants in our control sets could potentially

have a functional impact (and thus have a lower frequency due to evolutionary negative selective pressure), it is safe to assume that it is unlikely that they would contribute to the randomly selected phenotype.

For each of the phenotypes in our data set we performed an Endeavour gene prioritization [58]. Training of Endeavour has been preformed using known disease-associated genes from the Online Mendelian Inheritance in Man database. We appended the predicted Endeavour scores to each variant-phenotype combination based on their respective gene and phenotype. Haploinsufficiency [76], conservation, deleteriousness prediction scores (SIFT, Polyphen2, MutationTaster, LRT obtained from dbNSFP v1.3 [83] and CAROL score [54] were also appended.

4.3 Results and discussion

Table 4.1. lists values of various performance metrics, together with standard deviations, for three methods under the study; as recommended in [86] for this type of benchmarks. Figure 4.3 displays corresponding Receiver operating characteristic (ROC) curves. It is immediately apparent that the simplified version of eXtasy reaches almost the same performances as the original one across all indicators of interest; while the simple decision tree does not - analogously to what was observed during the initial benchmark [57]. In ROC space though, the simplified version exhibits somewhat reduced discriminatory capability in certain regions.

However, in its operating point (i.e. a decision tree threshold) it behaves almost the same as the original eXtasy model. This distinction might be an artifact of a method used to estimate posterior probabilities of the decision tree outputs. Indeed, the ROC curve of the simplified eXtasy is rather straight in the region of low false positive rate, indicating that the corresponding scores might be too rough compared to that of the Random Forest. Thus, a proper calibration might help in overcoming this effect. Alternatively, if only the importance of variables for prediction is of interest, different trees can be constructed using the simulated data corresponding to various thresholds of the Random Forest. In this way, contribution of predictors to certain performance metrics (ex. precision) can be analyzed along the operating range of the classifier.

In terms of size, the simplified eXtasy considerably exceeds the decision tree constructed using the real data, as it contains approximately nine times more nodes than the latter. This is not surprising, as no pruning took place in this case, in contrast to a usual decision tree induction. Additionally, the simulated sample is ten times larger than the real one. Yet, the Random Forest

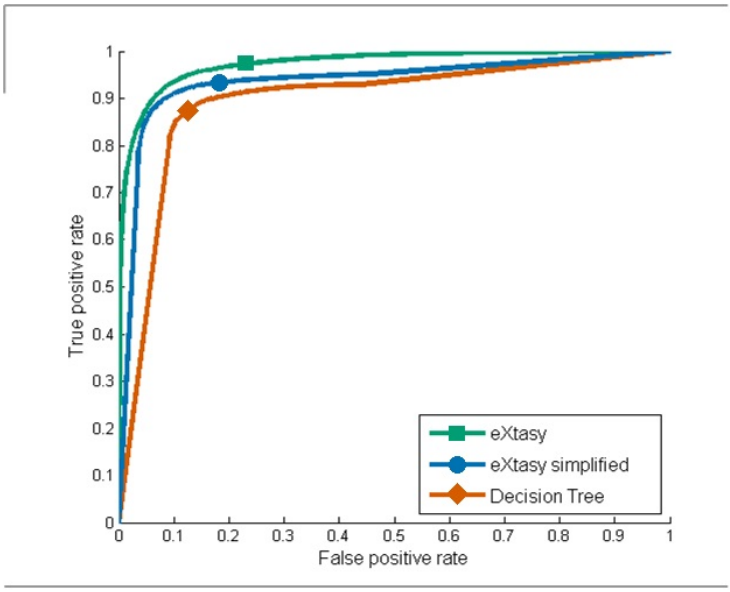


Figure 4.3: **ROC curves obtained by applying the three classifiers : standard eXtasy (Random Forest), simplified eXtasy and the decision tree built on the eXtasy data.**

classifier consists of unpruned trees too, thus the size of the new model is reduced compared to the old one by the total number of trees in the ensemble (originally 500). That is, in general, the size complexity of an unpruned decision tree is N times smaller than the complexity of a Random Forest that is built using the same data, where N stands for the number of trees in the ensemble. Consequently, as the execution speed scales linearly with the depth of a tree, the gain achieved in that aspect is also quite remarkable.

Figure 4.4 depicts first few levels of the decision tree induced from the simulated data. The Random Forest classifier has been previously trained using the whole dataset. Sequence similarity (as determined by BLAST) and gene function annotation (as annotated in the Swissprot database) compared to the known phenotype-associated genes are among the most discriminatory features. Also the global Endeavour scores seems to play an important role in the decision-making process be it in the form of the rank, the p-value or the normalized p-value of the gene prioritization.

Interestingly, not only phenotype-specific information but also prediction of the deleterious impact of the mutation (here shown by the MutationTaster node)

	Method		
Metrics	***eXtasy***	***eXtasy simplified***	***decision tree***
Accuracy	0.9404 (0.0045)	0.9348 (0.0052)	0.8743 (0.0076)
Sensitivity	0.8758 (0.0268)	0.8617 (0.0329)	0.8688 (0.0328)
Specificity	0.9496 (0.0057)	0.9450 (0.0067)	0.8748 (0.0104)
PPV	0.7158 (0.0298)	0.6946 (0.0328)	0.5019 (0.0348)
NPV	0.9816 (0.0031)	0.9795 (0.0035)	0.9789 (0.0048)
MCC	0.7586 (0.0245)	0.7373 (0.0293)	0.5977 (0.0299)

Table 4.1: **Average values of six performance measures obtained by testing the three methods** : standard eXtasy (Random forest), simplified eXtasy and the decision tree built on the eXtasy data. PPV stands for the positive predictive value (precision), NPV for the negative predictive value and MCC for the Matthews correlation coefficient. The corresponding values of standard deviations are enclosed between brackets.

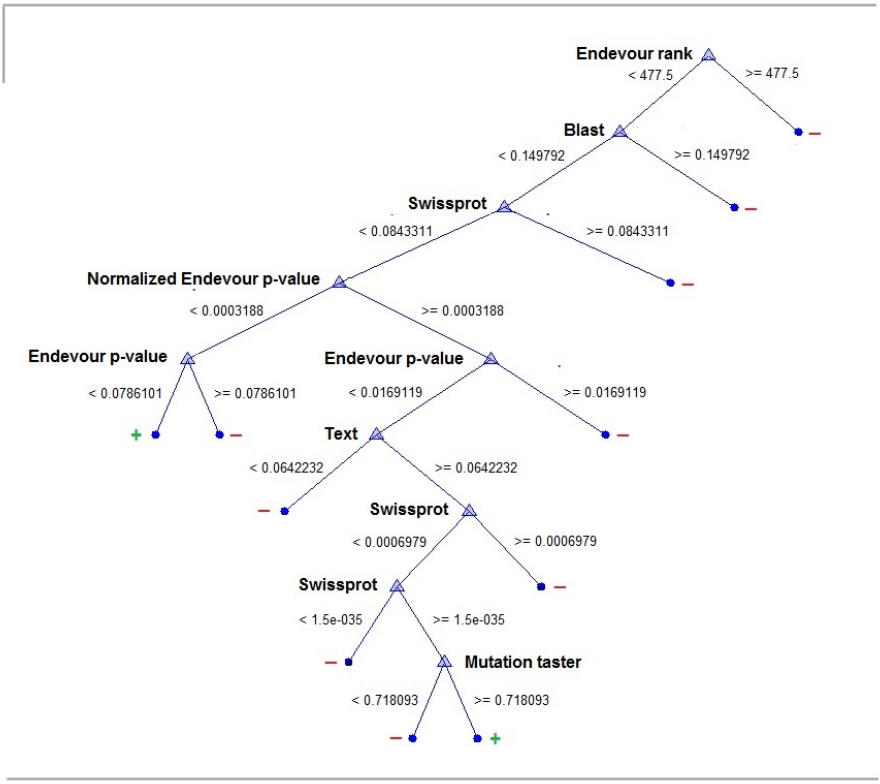


Figure 4.4: **Decision tree constructed on the data generated by Random Forest.** For clarity of the display tree has been pruned such that only ten the most discriminative nodes remain. Note that, due to pruning, the labels (signs) associated with "terminal" nodes only indicate which class prevails in a node and should not be confused with the final decision provided by the model.

appears informative. This seems logical as not any mutation in a potentially phenotypically-associated gene might be disease-causing if it does not perturb protein function significantly. This demonstrates that the decision boundary of the Random Forest clearly takes into account both aspects of the mutation: the damaging functional impact and the phenotypical relevance of the gene in which it lies.

When the structure of this tree is compared with Random Forest feature importance measures from the initial study [57], two major observations can be made. Firstly, importance of sequence similarity and gene function annotation for discriminating between two classes seems to be consistent, regardless to a method used to estimate it. This is the strong indication of the biological relevance of these features. Secondly, the Endevour scores apparently characterize this classification problem much more than previously implied. For being highly correlated, their importance was heavily underestimated by RF-FI; while here they constitute the most discriminative nodes in the tree corresponding to the simplified eXtasy. This suggests that importance of a variable for a particular classification problem can be indirectly assessed by modeling decision boundary of a complex classifier by a more interpretable one.

4.4 Concludions

We presented a simple method for extracting a comprehensible model from a Random Forest classifier and applied it to reduce the complexity of the eXtasy variant prioritization algorithm. The achieved performances are in line with that previously established in the original study, while the size (and consequently execution speed) of the model has been reduced considerably. In addition, we provide insights into the working mechanism behind eXtasy using an improved explanatory capacity of the new formulation. In the near future, we plan to further simplify the model through feature selection. Also, we plan to experimentally determine an appropriate method for calibrating posterior probabilities provided by a decision tree induction algorithm implementation(s), such that they can be used for ranking variants without losing performance along certain ranges of the operating space.

Chapter 5

Model improvements

This chapter has been accepted for publication in BMC Bioinformatics as :

Popovic D., Sifrim A., Davis J., Moreau Y., De Moor B., **"Problems with nested granularity of feature domains in bioinformatics - the eXtasy case."**

Preliminary results of research project corresponding to this chapter have been previously published as :

Popovic, D., Sifrim, A., Davis, J., Moreau, Y., De Moor, B. (2014). **"Improving performance of the eXtasy model by hierarchical sampling."**, in Proc. of the Pattern Recognnition in Bioinformatics - 9th IAPR International Conference PRIB 2014, Stockholm, Sweden, Aug 2014, pp. 125-128.

ABSTRACT

Background. Data from biomedical domains often have an inherit hierarchical structure. As this structure is usually implicit, its existence can be overlooked by practitioners interested in constructing and evaluating predictive models from such data. Ignoring these constructs leads to potentially problematic and the routinely unrecognized bias in the models and results. In this work, we discuss this bias in detail and propose a simple, sampling-based solution for it. Next, we explore its sources and extent on synthetic data. Finally, we demonstrate how the state-of-the-art variant prioritization framework, the eXtasy, benefits from the application of the described approach on its Random forest-based core classification model.

Results and Conclusions. The conducted simulations clearly indicate that the heterogeneous granularity of feature domains poses significant problems for both the standard Random forest classifier and a modification that relies on stratified bootstrapping. Conversely, using the proposed sampling scheme when training the classifier mitigates the described bias. Furthermore, when applied to the eXtasy data under a realistic class distribution scenario, a Random forest learned using the proposed sampling scheme displays much better precision that its standard version, without degrading recall. Moreover, the largest performance gains are achieved in the most important part of the operating range: the top of prioritized gene list.

Keywords : data granularity, validation bias, learning bias, hierarchical sampling, bootstrapping, eXtasy, Random forest, ensemble classifiers

5.1 Background

The data resulting from biomedical experiments often forms an implicit hierarchy in terms of granularity. A typical example is genomic data, where a single gene can harbor many mutations while it is at the same time a part of higher-order constructs (e.g., chromosome, patient). Likewise, a modeled target variable reflects this intrinsic property of the data. These structures pose no significant problem if the granularity of a target matches that of a data record. However, very often this is not the case as the values of features from different levels of the hierarchy could be mixed together to form data records, while the outcome might remain implicitly defined over coarser domain. In these situations, a flexible enough learning algorithm can be trained to "recognize" higher order structure by using only its corresponding features. This deficiency of learning can affect both performance of the prediction and correctness of the validation procedure.

We illustrate the described issue on the case of the state-of-the-art variant prioritization algorithm called eXtasy [57]. This method is based on fusing genomic data and it incorporates predictors defined over three distinct levels of data granularity. On the gene level (the coarsest grain) it integrates the haploinsufficiency scores [76]. On the intermediate level, each mutation within a single gene is characterized by several additional mutation-level features including various deleteriousness prediction scores (Polyphen [51], SIFT [50], MutationTaster [55], LRT [56], CAROL [54], which are all extracted from the dbNSFP database [83]) and conservation scores for vertebrate, placental mammals, and primate groups [84, 85]. Finally, a data record (the finest grain) is defined over mutation and a disease phenotype pair. It contains

	A	B	C	D	E	F	G	H		AA	AB	AC	AD
1	GENE_ID	PHENO_ID	VARIANT_ID	carol_score	haploinsufficiency	lrt_score	mutationtaster_score	phastconsprimates	pha	q_int	swissprot	text	outcome
2	216128	2710	64014633	0.999988	0.701	1	0.999618	0.143	5917669	3.90E+09	0.00390126	0.011345148	1
3	216128	2777	64014633	0.999988	0.701	1	0.999618	0.143	0256063	1.09E+06	0.0406819	0.0177213	1
4	216128	2805	64014633	0.999988	0.701	1	0.999618	0.143	0715489	7.99E+08	1	0.00861879	1
5	216128	2881	64014633	0.999988	0.701	1	0.999618	0.143	509299	0.000312544	1	0.67187019	1
6	216128	2950	64014633	0.999988	0.701	1	0.999618	0.143	787668	4.86E+09	0.204193898	0.073435645	1
7	216128	2962	64014633	0.999988	0.701	1	0.999618	0.143	0867674	1.22E+07	3.01E+05	0.000659602	1
8	216128	3027	64014633	0.999988	0.701	1	0.999618	0.143	513319	7.13E+07	1	0.007299591	1
9	216128	3172	64014633	0.999988	0.701	1	0.999618	0.143	24309	2.12E+08	0.697565055	0.026911745	1
10	216128	3200	64014633	0.999988	0.701	1	0.999618	0.143	601443	0.000225493	0.993871154	0.030429621	1
11	216128	3810	64014633	0.999988	0.701	1	0.999618	0.143	27345	0.000169521	0.576543161	0.370036498	1
12	216128	4683	64014633	0.999988	0.701	1	0.999618	0.143	4023	0.000779742	0.822516525	0.186183545	1
13	216128	8780	64014633	0.999988	0.701	1	0.999618	0.143	178741	1.31E+09	0.329630384	0.005892441	1
14	216128	9939	64014633	0.999988	0.701	1	0.999618	0.143	175718	0.006691543	1	0.104085133	1
15	216128	10658	64014633	0.999988	0.701	1	0.999618	0.143	101496	0.00153806	1	0.0400158	1
16	216128	11750	64014633	0.999988	0.701	1	0.999618	0.143	34735	0.000146783	0.000410744	0.008530847	1
17	216130	2772	30542221	0.909975	0.616	0.999923	0.963902	0.96	660913	0.0931664	1	0.609164	0
18	216130	3654	53818553	0.988663	0.616	0.999887	0.989713	0.981	195488	0.00694045	1	0.0423024	0
19	216130	11057	53818553	0.988663	0.616	0.999887	0.989713	0.981	518522	0.000542245	0.238437	0.0234818	0
20	216130	12327	53818554	0.889736	0.616	0.997731	0.242698	0.711	599532	0.108767	0.000454737	0.811486	0
21	216130	5404	53818554	0.889736	0.616	0.997731	0.242698	0.711	2	88	0.0138547	0.74324	0.036454	0
22	216130	4868	53818554	0.889736	0.616	0.997731	0.242698	0.711	8675	0.0539004	1	0.115562	0
23	216151	4823	30548241	0.883272	0.957	1	0.99995	0.936	61184	0.00103349	0.000509863	0.631854	0
24	216151	4405	30548241	0.883272	0.957	1	0.99995	0.936	2511	0.00032273	1.63E-02	0.395321	0
25	216151	3767	30550432	0.775791	0.957	1	0.99636	0.994	9167	0.00790289	6.62E-01	0.708676	0
26	216151	4987	30550432	0.775791	0.957	1	0.99636	0.994	61292	0.000187058	4.01E+02	0.713469	0
27	216151	5078	30551517	0.999944	0.957	1	1	0.943	59522	0.000318606	3.31E+06	0.639198	0

Figure 5.1: **A part of the eXtasy data.** Rectangles enclose values of haploinsufficiency (gene-level feature), mutation taster score (mutation-level feature) and a text mining based score from the Endevour (gene+phenotype level feature). Note that all mutations within the gene associated with these data records are non disease-causing.

additional features that corresponding to diverse metrics obtained from the Endeavor variant prioritization tool [58]. The fact that Endeavor provides scores for gene/phenotype pairs makes the problem even more complex. Figure 5.1 displays a part of the eXtasy data where the described hierarchy of granularity is clearly visible.

One issue is how to evaluate predictive models learned from this kind of data. In this particular case all data records that correspond to the same mutation have an identical outcome: disease causing or not. If the data were randomly divided into training and validation sets on the instance level (i.e., without considering the hierarchical structure), an algorithm could achieve high accuracy by simply learning the SIFT score value for each mutation. In other words, the SIFT value acts as an unique identifier. This clearly distorts a validation scheme's reliability that leads to optimistically biased performance estimates. Fortunately, the aforementioned problem is relatively easy to alleviate: the partition scheme must ensure that all data instances sharing the same mutation ID either only appear in the training set or only appear in the validation set.

However, the gene level information is also present in the data records while most (but not always all!) mutations within a single gene share the same outcome. If the percentage of mutations that have different labels within a gene is smaller than the theoretically achievable classification accuracy, the validation could be again optimistically biased due to the previously described issue. If this is the case, the previous reasoning implies that the partition should be based on gene as opposed to the mutation ID. This had been done during

the initial eXtasy evaluation. However, sometimes the higher-order structures are not obvious or are not known beforehand. In these cases, only a careful examination of the feature values can reveal potential problems.

The problem of inflated performance estimates in presence of the hierarchical feature granularity has been implicitly recognized before. One example is in the multicentric study that addresses the prediction of stroke patients restitution [113]. In this work, the authors noticed that when the data from several clinics are mixed together and then divided for the purpose of validation, the resulting performance is an overestimate compared to a genuinely external data set. To circumvent this, they advocate a "one-center-out" validation approach. Note that in this case there was not a single center-level variable in the data, yet the center-level *information* may be still implicit in some features, such as demographics. Similarly, a "leave-one-drug-subclass-out" approach was used to assess the prediction of the functional class of an unrepresented drug type [114]. There each model has been trained on all but one subclass of antidepressant drugs and then used to classify members of the subclass that was omitted from training. Furthermore, a "leave-one-cow-out" strategy was proposed in the context of predicting the somatic cell count in whole milk from near infrared spectroscopy measurements [115]. Surprisingly, this validation problem is not always fully recognized.

The second issue associated with hierarchical granularity is much more latent and therefore often neglected in the data modeling. That is, even if the validation scheme is appropriately set with the respect to the highest level of data granularity on which the decision values are (mostly) uniform, the discussed bias still exists. In such a situation, performance estimates will be correct but the performance itself might decline. This happens if the rows (i.e., examples) are inter-independent, even though the data consists of a single table where each example is described as a fixed-length feature vector. The interdependencies exhibit themselves on different levels of granularity, where all inter-dependent examples have an identical value for a specific feature as well as the same value for the target variable. Thus the feature value appears correlated with the target variable whereas in reality the feature value is correlated with the hierarchical structure of the data. Failing to consider the interdependencies during learning could cause the algorithm to produce a model that simply identifies a pattern that is correlated with the hierarchical structure of the data as opposed to a pattern that is correlated with the target variable. In other words, the algorithm can be still overfitting on coarse grained features of the training data partition, failing to actually generalize from it.

In the context of the eXtasy data, the described bias materializes as learning, to a certain degree, to recognize genes which constitute the training set, rather than extracting general characteristics of disease causing mutations; which altogether

leads to a reduced performance on the test set. This happens as many data instances share the same values of higher order features. In other words, instead of learning that "mutation A that is a part of gene B is probably disease causing due to high haploisufficiency score of gene B (which implies correlation of haploisufficiency with outcome)", algorithm may infer that "mutation A is probably disease causing as haploisufficiency score of gene B is exactly 0.998 (which uniquely identifies given gene)"; which in turn does not tell anything about a new example with haploissuficiency of 0.999.

This situation is radically different from having one or more categorical variables in the data set, although sometimes it might be hard to distinguish between these two. While categorical features can also have identical values for several data instances that can be in same time correlated with the outcome, their values are *not ordered*. Thus, in that case might be actually preferred that the algorithm learn the correlation between particular categories and the outcome, instead of extracting some trend that involves the ranges of the variable values. In addition, when the trained model is applied on an unseen data the number of possible categories typically stays the same. This is not a case with coarse-grained features, as previously unobserved values of these are often present in new data (ex. a previously undiscovered mutation that did not participate in training).

In the remaining text, we describe a simple, sampling-based method for dealing with the described bias. First, we further illustrate the problem using a conveniently generated synthetic data set and analyze robustness of the solution under different factors, including the grain size and the level of label noise. Second, we demonstrate how the performance of the core eXtasy model can be improved by the application of the hierarchical sampling.

5.2 Methods

5.2.1 Hierarchical sampling

The core model of eXtasy is based on the Random forest classifier [20]. Random forests are an ensemble method that constructs a set of unpruned decision trees using different bootstrap samples [116] of the data. Furthermore, randomness is injected during model construction by only considering a randomly sampled subset of the candidate variables at each split point, as opposed to scanning all candidate variables as is done in traditional decision tree learning. The final label of an unseen example is an unweighted vote of the predictions by each model in the ensemble (as done in bagging [117]). Hence, Random forests

Algorithm 1 Random Forest with hierarchical sampling

Inputs:
$D = \{P_1 = \{e_{1,1}, \ldots, e_{1,n}\}, \ldots, P_p = \{e_{p,1}, \ldots, e_{p,m}\}\}$ is the set of training examples, divided into p partitions, where each partition corresponds to examples sharing the same value of the feature that is defined over the coarsest domain
t is the number of trees to learn

Let $T = \emptyset$
for $i = 0$ to t **do**
  Let $BS = \emptyset$
  for $j = 0$ to p **do**
    Let $k \sim \text{Uniform}[0, |P_j|]$
    $BS = BS \cup \{e_{j,k}\}$
  end for
  $t_i = \text{LearnTree}(BS)$
  $T = T \cup \{t_i\}$
end for
return T

may be particularly susceptible to the described problem due to the fact that they both sample examples and attributes. That is, failing to account for the correlations during this sampling could produce a biased sample.

We propose a straightforward modification to the Random forest framework, named hierarchical sampling. Instead of extracting a bootstrap from the complete training set to build a single tree on, we first stratify the training examples according to the distinct values of the feature over which the coarsest granularity level is defined. In the case of the eXtasy data, that would be the gene identifier. Note that a feature could implicitly define the grain while not being a part of a data set at all (a "latent factor" of the outcome grouping), suggesting that domain knowledge is needed to perform this procedure. After stratification, we randomly select just one data instance from each partition to form the in-bag sample for one tree. This helps in preventing single trees from learning to recognize a particular value of the higher-order feature, as only one example having it will be present in each sample. At the same time, each partition will be well represented in the ensemble as a whole, provided that a sufficient number of trees is created. Algorithm 1 formally describes this procedure.

It is important to distinguish the described procedure from stratification in the classical sense as outlined in Algorithm 2. While the later is often used with the Random forest classifier and seems very similar to the hierarchical

Algorithm 2 Random Forest with stratified bootstrapping

Inputs:
$D = \{P_1 = \{e_{1,1}, \ldots, e_{1,n}\}, \ldots, P_p = \{e_{p,1}, \ldots, e_{p,m}\}\}$ is the set of training examples, divided into into p partitions, where each partition corresponds to examples sharing the same value of the feature that is defined over the coarsest domain
t is the number of trees to learn

Let $T = \emptyset$
for $i = 0$ to t **do**
  Let $BS = \emptyset$
  for $j = 0$ to p **do**
    for $k = 0$ to $|P_j|$ **do**
      Let $l \sim \text{Uniform}[0, |P_j|]$
      $BS = BS \cup \{e_{j,l}\}$
    end for
  end for
  $t_i = \text{LearnTree}(BS)$
  $T = T \cup \{t_i\}$
end for
return T

sampling, its purpose is diametrically different. The stratification is typically used for making sure that each of the (latent) classes (in our case examples having the same value of the coarsest feature) is well represented. This is usually achieved through bootstrapping each stratum separately, in contrast to taking just one example from it. Hence, even it can be quite desirable when dealing with categorical variables for example, in this given context classical stratification can promote overfitting by guarantying wide presence of each gene in an in-bag data partition of any given tree forming the ensemble. Conversely, taking one example per stratum assures that no gene-level information will be over-represented in a single tree. Repeating the sampling before learning each new tree helps to protect the model from under representing the data as *a whole* in the *ensemble*. To further underline this distinction, we include Random Forest trained with stratification in the following synthetic data analyses.

5.2.2 Experiments with synthetic data

To characterize the discussed bias in greater detail and to analyze its sources and extent, we constructed synthetic data where the phenomena is isolated from other sources of variation and is relatively easy to control. This data set

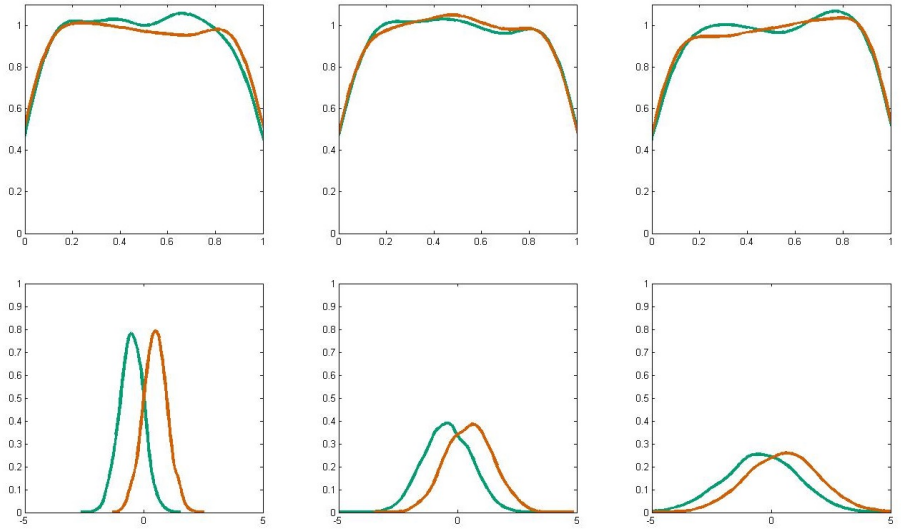


Figure 5.2: **Features of the synthetic data.** Kernel density estimates of the synthetic data distributions. Panels in the first row display distributions of three *non-informative* features for two classes separately. Note that decreased density in the vicinity of sampling interval borders is in fact an artifact of the kernel density estimation procedure and not of the uniform sampling itself, thus it affects only this visualization. The bottom three panels depict *informative* features.

is described by seven variables and it has a binary class label. We generated four thousand examples such that the class distribution is balanced (i.e., that data have two thousands examples for each class).The first three features are non-informative and are simply sampled uniformly from the interval (0,1). The next three features are *informative* and each one is represented by a conditional Gaussian distribution. The mean depends on the class and the variance, and hence the overlap, is different for each feature (see Figure 5.2). The first six features remain unchanged in all experiments. The seventh feature has been created to represent the hierarchical structure in the data. It will be manipulated in the two experiments.

The first one has been designed such to reveal the relation between the size of the coarsest variable partitions (grain) and magnitude of the described bias. In particular, the additional *non-informative* feature has been generated by the uniform sampling and appended to the rest of the data, similarly as before. However, in this case the number of distinctive values that this new feature can take has been varied from two to two thousands per class, which led to a change in granularity of the feature domain. In total, ten possible sizes of

partition have been considered ($|P| \in \{1, 2, 5, 10, 20, 50, 100, 200, 500, 1000\}$).

Given each of these, the appropriate feature has been created, after which the whole data set has been divided into equally sized training and testing partitions along its distinct values. All three classification models are then learned using the training data and evaluated on the test data. The total number of trees in each model ensemble has been set to 1000 and the size of the random subset of variables from which each decision tree split is chosen (parameter M) to its recommended value for classification, that is, to the square root of the total number of features [20]. This whole procedure has been repeated one hundred times for every bin size. In each repetition, the auxiliary feature is regenerated and a new, random training-test split is made. In addition, in order to facilitate complementary insight into the learning process, the out-of-bag feature importance measures [20] obtained from the classical Random forest model have been harvested along the way.

The second experiment will assess the effect of label noise on the extent of the bias. The label noise was introduced by swapping certain percentages of the values from each partition of the auxiliary feature belonging to the one class with values of its counterpart from the other class. This percentage varied from 5 to 50% in steps of 5, resulting in ten possible noise levels in total. The partition size was again increasing, but in this case starting from 20 to allow for meaningful noise injection. Finally, training and evaluation of the models have been performed in the same way and under the same parameter setting as in the previous experiment.

5.2.3 Experiments with eXtasy data

To ensure a fair comparison, we test the method on the original eXtasy benchmark data. Briefly, this data set consists of two classes of mutations: disease causing variants and rare mutations present in healthy individuals. There are 24,454 disease causing variants in the data set, which were obtained from the Human Gene Mutation Database (HGMD [77]). The controls come from 68 in-house sequenced exomes of healthy individuals and the 1000 Genomes Project [87]. For each of the 1142 Human Phenotype Ontology (HPO) terms associated with the disease-causing variants, 500 mutations were randomly sampled from the pool of controls and assigned to a given phenotype. Endeavor scores [58] have been appended to each variant-phenotype combination, together with haploinsuficiency [76], conservation [84, 85] and deleteriousness prediction scores [51, 50, 55, 56].

To compare hierarchical sampling to the standard bootstrapping, we use the same evaluation scheme as in original study [57]. That is, we randomly divide

the complete benchmark data set on the gene-level such that two-thirds of the genes constitute the training set and one-third are in the test set. We consider two test scenarios. In the first one we compare two sampling schemes on the unaltered test set, effectively repeating the eXtasy benchmark. In the second one we randomly undersample the positives from the test set in order to mimic the class distributions we would expect to see in the wild; where only one out of 9000 non-synonymous mutation in a genome is potentially disease-causing [53]. In both scenarios and for both sampling schemes we use the same setting of Random forest parameters as in experiments with the synthetic data. We repeat the aforementioned procedure 100 times to stabilize the values of the performance metrics.

Finally, we test significance of observed differences between the two methods by using Wilcoxon signed ranks test after which we apply Bonferroni multiple testing correction. For a comparison of two classifiers, the Wilcoxon signed ranks test is preferred over standard t-test [118], as later exhibits high probability of type I error when used in this context [119]. Similarly, the Bonferroni correction is very conservative [120], so we use it here to stay on a "safe side"when making inference about performance differences.

5.3 Results and Discussion

Results of the synthetic data benchmarks are provided on Figures 5.3 and 5.4. Due to text space limitations and class balanced design of the synthetic data, we restrict the presentation of results on aggregate measures of the classification performance. From the Figure 5.3, it is immediately apparent that the increase in partition size directly leads to a decrease in the performance of the classical Random Forest formulation, starting from bin sizes as small as twenty. This trend is evident even in this setup, where we do not make smooth transition between numbers of instances in partitions.

The evolution of associated feature importance metrics suggests that the reason for decreased performance is indeed progressive overfitting on the additional non-informative feature. That is, in parallel with bin enlargement this feature becomes increasingly important for the classification, while in the same time the importance of genuinely informative features decline. Furthermore, training the same classifier with stratification does not seem to help at all, as its performance tightly follows that of the regular version.

Conversely, the Random Forest trained with hierarchical sampling stays relatively insensitive to change in the bin size, with only exception being the last step. However, this is a degenerate case in which the method has only

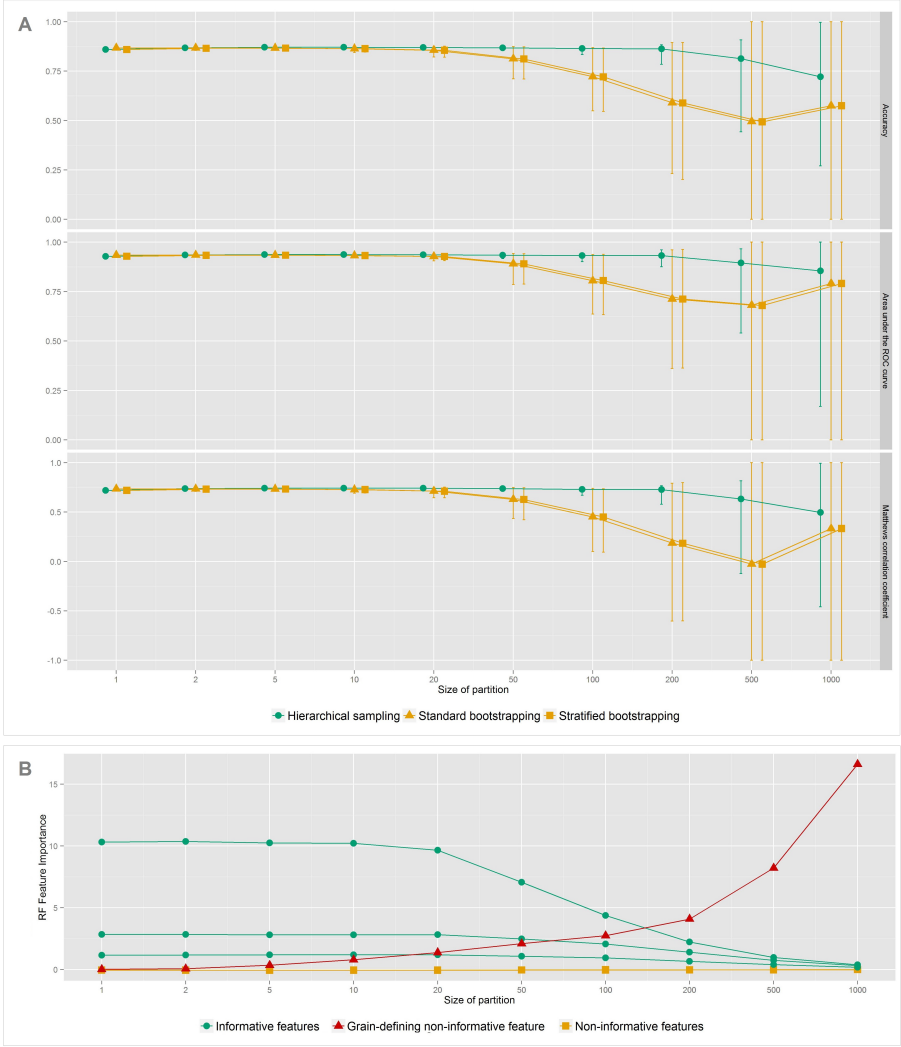


Figure 5.3: **Results of the first experiment on the synthetic data.** Evolution of testing accuracy, area under the ROC curve and Mathews correlation coefficient of Random forest classifiers trained with three sampling schemes and under the increasing size of partitions of the grain-defining feature (Panel A). Vertical bars indicate empirical 95% confidence intervals. Panel B displays corresponding change in Random forest feature importance metrics for all features of the synthetic data set under the standard bootstrapping. This metric captures an increase in out-of-bag classification error when the values of given feature are shuffled.

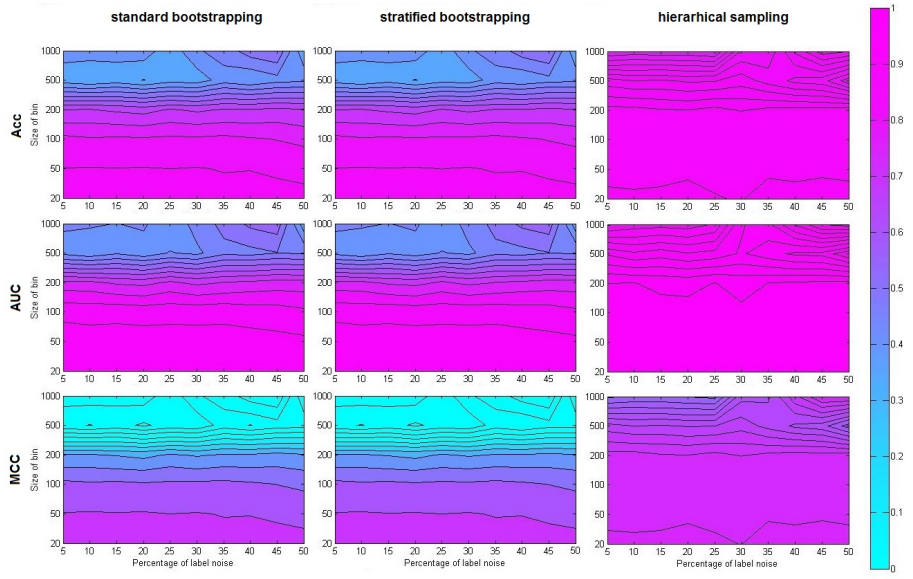


Figure 5.4: **Results of the second experiment on the synthetic data.** Heat-maps of mean testing accuracy (the first row), area under the ROC curve (the second row) and Mathews correlation coefficient (the third row) of Random forest classifiers trained with either standard bootstrapping (the first column), stratified bootstrapping (the second column) or hierarchical sampling (the third column). Panels capture relation between given performance metrics and bin size/label noise level combinations. Note that values of Mathews correlation coefficient can be as low as -1, which is the reason why upper parts of corresponding plots have uniform coloring (i.e. all values in this region are smaller than zero).

one example per class on disposal for training of each tree, which is clearly insufficient for good generalization. Also, this situation is not likely to be encountered in practice. The described effect persists even in the presence of label noise, as it is apparent in the Figure 5.4. Yet in the case of classical Random forest or Random forest trained with stratified bootstrapping, the bias obviously declines with the increase in the amount of noise added. This is expected, as addition of label noise translates to an implicit reduction of the partition size.

Results of the eXtasy-based benchmark are provided in Table 5.1 and Figure 5.5. Figure 5.5 shows the PR (Precision-Recall) curves only, as there the difference in performance between two methods is graphically evident. These curves are constructed using the threshold averaging [121] over the 100 runs. Table 5.1 reports the corresponding area under the curve values. Following the recommendations for reporting on classification benchmarks [86] and to

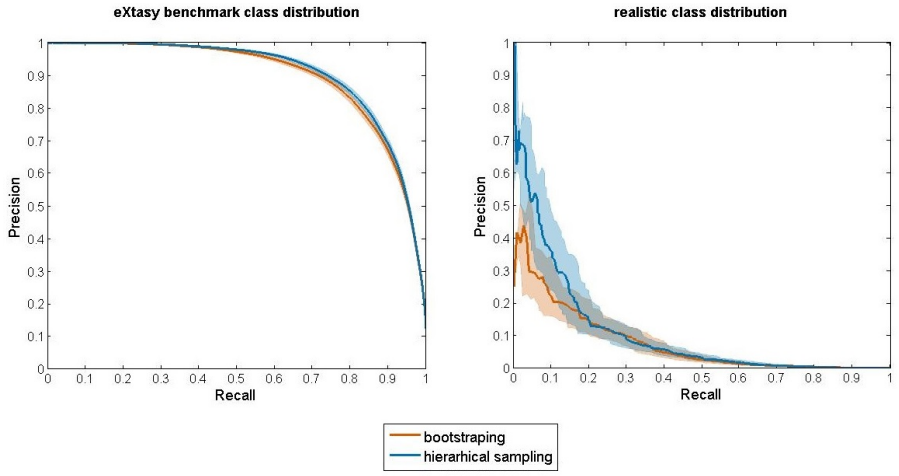


Figure 5.5: **PR curves from eXtasy-based experiment.** Precision-recall (panel B) curves obtained by application of the eXtasy on the data with class balance as in original eXtasy benchmark (left panel) and on the data with realistic class distribution (right panel). Each panel display two solid curves – the one corresponding to the standard Random forest classifier training with bootstrapping and the one corresponding to hierarchical sampling based training. The shaded areas represent empirical 95% confidence intervals obtained as a by-product of threshold averaging.

provide insight into different aspects of the classification performance, Table 5.1 also provides average values of Precision (Positive predictive value), Recall (Sensitivity), Negative predictive value, Specificity and Matthews correlation coefficient for both scenarios along with the indication of statistical significance of differences. In addition, it provides values of area under the ROC (Receiver Operating Characteristics) curves obtained by threshold averaging.

The difference between bootstrapping and hierarchical sampling in terms of achieved area under the ROC and PR curve might seem marginal when only the result on the eXtasy benchmark data is considered. However, the realistic class distribution scenario highlights the benefit of hierarchical sampling as there are large improvements in PR space, especially in the low-recall, high-precision region of curve which is of most interest for prioritization algorithms. Due to the associated financial costs of confirmatory experiments, analysis is usually only conducted on highly ranked candidate variants (i.e., those at the top of the ordered list) and not on all mutations that are classified as positive. Thus the precision of the prioritization method for the highest ranked genes is of critical importance, especially as most state-of-the art deleteriousness prediction methods suffer from high false positive rates when predicting the impact of rare

	eXtasy benchmark class distribution		realistic class distribution	
Metric	***bootstraping***	***hierarchical sampling***	***bootstraping***	***hierarchical sampling***
Sensitivity	**0.863751***	0.783985	**0.888667***	0.816667
Specificity	0.951801	**0.980001***	0.951809	**0.980030***
Precision	0.711571	**0.842510***	0.002423	**0.005313***
NPV	**0.980934***	0.970764	**0.999985***	0.999976
MCC	0.751416	**0.788129***	0.044896	**0.064834***
AUC-ROC	0.972466	**0.973497**	0.979336	**0.981345**
AUC-PR	0.888809	**0.895255**	0.095083	**0.128191**

Table 5.1: **Results of the eXtasy-based benchmark.** Results of the benchmark in terms of average Sensitivity, Specificity, Precision, Negative predictive value (NPV) and Matthews correlation coefficient (MCC). Bold typing indicates which method of generating in-bag sample is better given a performance measure and a classification scenario, while asterisk (*) points on statistically significant differences (p=0.05 level). The values of the area under the ROC and PR curve are obtained from the average curves constructed by threshold averaging; hence no statistical testing has been applied on these.

disease-causing mutations [73].

Discussed performance differences are difficult to detect when using ROC analysis alone, as ROC curves are invariant to changes in the class distribution. This property of ROC curves is also the reason for which they should, in principle, be identical for both *scenarios*, given a classifier. Naturally, they are not, due to variation in performance estimates. This variance is large enough to also cast doubt on systematic cause of any observed difference in ROC curves between *classifiers* in the same scenario. However, dominance of one curve over another in PR space implies dominance in ROC space as well [122], while *size* of the effect can differ. Hence, the extent of difference between two classifiers in PR space for realistic class distribution scenario constitutes indirect prof of existence of systematic effect in ROC space.

As the result table show, the hierarchical sampling RF performs worse than the standard RF in terms of sensitivity in both test scenarios. Conversely, this is offset by the higher precision it is able to achieve in both settings. Being complementary to measures already mentioned, specificity and the negative predictive value also reflect the aforementioned trade-off, albeit to a smaller degree (as more abundant negatives play major role in these). However, as apparent from PR curves, setting the decision threshold of the RF trained using the hierarchical sampling so that it achieves the same sensitivity of the standard RF still results in the higher precision of the first RF. These results are in line with previous discussions on differences in performance at the top of prioritized lists. Additionally, even with the default RF threshold (0.5), the improved version of eXtasy classifies approximately 188 out of 9000 variants as disease causing, with the probability of capturing the real one equal to 0.81 (i.e.

sensitivity). In contrast, the standard eXtasy calls 417 out of 9000 variants, with the probability of hit being 0.88.

We acknowledge that presented results may vary with information content of the grain-defining feature and the rest of predictors, correlations among them and the other factors. However, due to difficulties with quantification of these aspects and the exponential number of potential relations among them, we omit their detailed treatment from this work. We also expect that further improvements in performance could be obtained by fine tuning the parameter M of the Random forest algorithm. That is, the application of hierarchical sampling increases the diversity of an ensemble compared to the standard bootstrapping, possibly allowing for larger values of M to be used [123]. Bigger values of M might translate to an improved accuracy for the base classifiers, resulting in an overall boost in performance.

5.4 Conclusions

We described the bias arising in learning from data characterized by nested granularity of feature domains and proposed a simple solution for it, named hierarchical sampling. Using the synthetic data, we demonstrated that the approach efficiently mitigates the discussed bias, regardless to the grain size and the amount of the label noise that is present in the data. The simulation study also indicated that the gain in performance can vary substantially for different classes of problems. That is, if the number of distinctive values of the coarsest grain concept is much smaller than total number of data records, overfitting on these concepts is more likely to occur. Furthermore, in the case of the eXtasy variant prioritization algorithm, the hierarchical sampling led to a notable improvement in the model performance in terms of the precision, especially in the most important operating regions for this particular application. Also, as it uses less data (per single tree) than standard Random forests bootstrapping, it typically results in a more parsimonious model in terms of the average tree depth.

However, an improvement in performance that might be obtained by an application of the hierarchical sampling is only visible under proper evaluation. In particular, if a hierarchy of feature domains (even implicitly) exists in the data and it is neglected during the evaluation, standard Random forest will falsely perform better on the test data due to the evaluation problems described in the "Background" section of this manuscript. Conversely, a Random forest trained with the hierarchical sampling would not be able to memorize values of the grain-defining feature, and thus to recognize them in the test set. Hence

the situations where the proposed approach can be useful are the same ones where the data partitioning for validation purpose must be performed along values of an implicit identifier having the lowest number of distinctive values.

Finally, even the application of hierarchical sampling is not in principle restricted to Random forest, the method can be efficiently used only in conjunction with an ensemble classifier that relies on some sort of bagging scheme. If sampling from the highest granularity level would be applied in the one-shot fashion (e.g., for training an SVM), much data would be ignored during training which could negatively affect performance. That is, even in cases where the sample itself is reasonably large, general information embedded in the fine-grained features might not be sufficiently represented. In contrast, bagging allows for all training data to be fully exploited, while the proposed sampling ensures that no base classifier can learn to recognize higher order structures in data. In the near future, we plan to integrate the Random forest model learned using the hierarchical sampling into the next version of the eXtasy framework.

Competing interests

The authors declare no competing interests.

Authors' contributions

D.P. and A.S. conceptually defined the problem. D.P. proposed a solution, developed the benchmarks and performed the analyzes. D.P. and J.D. wrote the initial draft of the manuscript. A.S. generated the data sets. J.D. advised on general machine learning concerns. B.D.M. supervised the project. Y.M. co-supervised the project. All authors revised and proofread the manuscript.

Additional Files

Additional file 1 — rf_hierarchical.mat, a Matlab function for training Random Forest with hierarchical sampling : This Matlab function accepts training data, training outcomes, number of trees in the ensemble and the ordinal number of the feature over which the hierarchical sampling has to be performed and returns a trained Random forest classifier.

Chapter 6

Conclusions

Many genetic disorders have a small prevalence within a general population when considered individually. However, approximately 8% of all people suffer from *some* of the more than 7000 described inheritable diseases [26]. Often the right diagnosis is established only later in life for these patients, which complicates the treatment and increases the morbidity and mortality. One of the reasons for this situation is certainly a low awareness of these conditions among physicians, but also the insufficiency of the current diagnostic tools. At the same time, recent advances in exome sequencing have the potential to bridge the gap between fundamental research on the one hand and ready-to-use, clinical genetic tests on the other hand. Yet, finding a single causative mutation among numerous candidate variants is still a formidable problem.

Many variant prioritization algorithms have been proposed to facilitate this process by the identification of mutations that are *susceptible* to being involved in a disease development. We reviewed some of them belonging to the state-of-the-art and pointed out several serious issues that limit their practical applications. Most of these issues originate from an inadequate training or validation, resulting in either hampered generalization properties or problematic performance estimates.

We have developed a complete computational framework for prioritizing mutations that cause rare genetic disorders while being focused on avoiding the pitfalls, as described in Chapter 2. As a result the system, named eXtasy, achieves an approximately 10-fold increase in precision compared to the state-of-the-art, without an associated loss of sensitivity.

Our strategy relies on three key innovations that significantly improved its prioritization capabilities. First, it incorporates phenotype-specific information in the form of similarities to known disease genes. Second, it subsumes most of the current deleteriousness prediction scores and integrates other sources of biochemical, structural and evolutionary data. Finally, to our best knowledge, it is the only method specifically trained to distinguish between rare benign and rare disease-causing mutations. Surprisingly, this issue is seldom taken into account, hence many algorithms encounter difficulties in the correct classification of rare non-causative variants.

In chapter 2, we have shown that the method performs significantly better than the currently popular deleteriousness prediction tools, i.e., SIFT, PolyPhen, Mutation Taster and Carol scores. This result holds both for our composition of the benchmark data and for the typical setup in which solely the common polymorphisms play the role of benign outcome. In addition, we tested our algorithm in a *semi-prospective* fashion, where the classifier has been trained on data predating the year 2001 and tested on several test sets that were composed of variants discovered in the following years. Even with a slight temporal decline in performance throughout the testing years, the method still outperforms the competing algorithms by a considerable margin.

Finally, compared to typical, relatively balanced benchmarking setups, the observed difference in performance is even more evident when the algorithms are applied on data that follow a class distribution that is normally encountered in practice and when the metrics appropriate for the prioritization algorithms evaluation are used (i.e., precision and recall).

One subtle, but important aspect of the functioning of eXtasy is reflected in the behavior of its corresponding precision-recall curves under a changing class distribution. A careful examination of the figures depicting different classification scenarios suggests that the method is relatively insensitive to these shifts compared to the other considered algorithms. In particular, class imbalance is more severe in the "rare benign vs. rare disease-causing mutations" scenario than it is in the "common polymorphisms vs. rare-disease causing" one, which naturally translates to reduced values of the prevalence-sensitive classification metrics.

Yet, the difference in area under the PR curve between the two setups is much lower for eXtasy than for the other methods, both in relative and absolute terms. In addition, its precision seems to drop slower with increase in recall. These two observations indicate that most of the instances classified as disease-causing by eXtasy actually have high associated scores, i.e., they reside near the top of a prioritized variant list. This property is desirable as it matches the typical usage of a prioritization output where the downstream analysis starts from the

highest scored mutation.

To evaluate the informativeness of each data source that was used in the model, we conducted two complementary analyzes. Firstly, Random forest feature importance estimation was applied, showing that all of the features contributed to the observed classification performance to some degree (subsection 2.2.5). As a corollary, this implies that the superior ranking capabilities of eXtasy are not a mere consequence of the intelligent combining of existing deleteriousness prediction scores but also of the inclusion of phenotype specific information. However, the Random forest feature importance measures are notorious for being highly biased in the presence of correlations between predictors, hence the *extent* of contribution from individual features remained undetermined.

Therefore, as detailed in Chapter 4, we ran a second analysis where we described the decision boundary that is implicit in the fitted eXtasy algorithm by the more interpretable, decision tree based model. Comparing the structure of this tree with the previously generated importance estimates revealed that the Endevour-based features are among the most discriminative ones for this particular prioritization task. At the same time, they exhibit severe inter-dependency, thus they can easily assume the "roles" of each other in a classification. As a by-product of this analysis, we deployed a simple formulation of the eXtasy core model that is characterized by a smaller size and a higher execution speed than the original one.

Moreover, in Chapter 3 we investigated the behavior of the method when different aggregation schemes are used to integrate predictions for multiple phenotypes. The results confirmed the intuition that taking the maximal score obtained as the final prediction for a mutation, is the most robust strategy when the quality of the phenotypic information is unknown. Specifically, a few of all phenotypes that are associated to a variant can be either under or over described, i.e., their Endeavor training gene sets are either too small to learn from or too general. In addition, some of the phenotypes might not be biologically well-suited to describe the disease under study.

However, we also demonstrated that if the selection of phenotypes is guided by domain knowledge, the prioritization can benefit from using more involved statistical methods which we adapted for this application. These methods also provide significance estimates, which facilitates the integration of a prioritization result into wider statistical frameworks.

Finally, in Chapter 5 we characterized a bias arising in learning from data defined over several nested levels of domain granularity and described how and why it affects the eXtasy performance. We also proposed a simple sampling-based solution for this bias, the usefulness of which extends to many similar

problems. The inclusion of this strategy to the eXtasy training scheme resulted in a noteworthy improvement of the model's precision, especially in the vicinity of the most highly scored variants. Furthermore, it also contributed toward the construction of a more parsimonious model.

Unfortunately, this is not the only bias that affects eXtasy. As mentioned before, we observed a decline in the performance with years passing. The reason behind this is the non-stationary nature of the data distribution that violates the basic machine learning assumption which was introduced previously. In particular, new mutations and genes are constantly discovered, sometimes having different properties than could be expected from the existing knowledge. This renders eXtasy biased towards well described genes, phenotypes and mutations. Fortunately, it seems that the data distribution changes rather slowly and does not seriously hamper the method's performance. However, this notion underlines the importance of frequent updates to the system; on which we will further elaborate in turn.

It is important to note that some of the other deleteriousness prediction methods do not suffer from the discussed issue, as suggested by the result presented in Figure 2.2. This is expected as many of them do not use prior knowledge in any form but rather rely on some general properties that do not change depending on how much information about the target is available. For example, SIFT [50] assumes that important amino acids tend to be conserved in the protein family, hence it considers the position and the type of amino acid change for the calculation of the score.

The additional bias that influences the system is much more subtle, and therefore much harder to analyze, quantify and account for. It is in fact a consequence of several technical issues with the data sources that are beyond our control. For example, if a gene belongs to the test set, we typically do not use it as a training gene in the Endevour tool either. However, sometimes the year of a gene discovery has not been correctly assigned in some of the biological databases used, leading to the situation where a gene figures in a testing set of the temporal benchmark and at the same time remains in the Endevour training pool.

Even this might be a way too indirect connection to cause any observable difference in the tool's performance, the only way to fully circumvent the problem would be a manual curation of *all* external data sources. Otherwise, the complexity of individual features and the subtle links among them may obscure potential information leakage. For this and the previous reason, retrospective benchmarks can not be expected to accurately predict how the method will behave when used in practice, but rather to give a rough indication of it. As before, the other tools are not affected by this bias as they do not rely on

heterogeneous data fusion.

In the light of the importance of continuous system updating to address the discussed biases, in the future we plan to systematically react to changes in several critical aspects of the training data by refitting the base eXtasy model. First, each new discovery of a disease-causing mutation expands the corpus of available information, hence batches of novel variants have to be added to the model-building data set at regular time intervals. Second, some of the data sources also get updated periodically, thus the appropriate eXtasy feature has to be replaced. For example, while writing this text, a new version of Mutation Taster appeared [124]. So, we will include it in the next incarnation of the system instead of the old one. Third, some tools that we incorporate in our framework may expand their panel of features, as will be the case with a new version of Endeavor. Some of these that convey novel information for the variant prioritization task will be added to eXtasy too. Finally, completely new data sources, such as copy number–variant prevalence or genome-wide association study–related loci, can be appended to the existing ones.

In addition, we plan to further improve the model itself. Next to the alternative sampling scheme that was discussed in Chapter 5, we will add a module for the intelligent assignment of the appropriate aggregation scheme according to the available information about phenotypes. In particular, the synthetic data described in Chapter 5, together with several phenotypic meta-features (ex. number of Endeavour training genes for a given phenotype), will be used to train a multiclassification model to decide which aggregation method should be used for each newly tested variant. The initial tests have showed that this approach can boost performance of the complete system.

Furthermore, the *scope* of the tool might also be expanded. Currently, eXtasy can be used only for the prioritization of nonsynonymous single nucleotide variants, but we aim to spread the spectra of mutations which can be analyzed on other types as well (noncoding, synonymous, nonsense...) in the near future. Similarly, the ideas applied in the development of eXtasy can be readily utilized for the "vertical" expansion to a different class of diseases, such as cancer. However, in that case we expect that radically different data sources have to be considered.

An auxiliary dimension of further development might be an integration of the method to multiphase data analysis pipelines. For example, it could be fused into genetic association tests across multiple samples, which would probably increase its power. Alternatively, it can be hybridized with some feature selection-based biomarker discovery algorithms, in which case its bias toward well described genes would be partially alleviated. In particular, it can be used to further prioritize candidate variants that are selected by the feature selection-based

method.

At the end, it seems appropriate to emphasize that currently no single computational method for biomarker discovery can function in the vacuum of self-sufficiency. They are rather a mere piece in a jigsaw puzzle that this multidisciplinary effort is. Hence, we will close this discussion with a bit of adapted advice that Leo Brieman originally gave on the proper interpretation of a Random forest output [125]:

"The cleverest algorithms are no substitute for human intelligence and knowledge of the data in the problem. Therefore, take the output of eXtasy not as absolute truth, but as smart computer generated guesses that may be helpful in leading to a deeper understanding of the problem"

Bibliography

[1] S. F. Idris, S. S. Ahmad, M. A. Scott, G. S. Vassiliou, and J. Hadfield, "The role of high-throughput technologies in clinical cancer genomics," *Expert review of molecular diagnostics*, vol. 13, pp. 167–181, Mar. 2013.

[2] M. Tuna and C. I. Amos, "Genomic sequencing in cancer," *Cancer letters*, vol. 340, pp. 161–170, Nov. 2013.

[3] J. J. McCarthy, H. L. McLeod, and G. S. Ginsburg, "Genomic medicine: A decade of successes, challenges, and opportunities," *Science Translational Medicine*, vol. 5, pp. 189sr4–189sr4, June 2013.

[4] G. Jannes and D. De Vos, "A review of current and future molecular diagnostic tests for use in the microbiology laboratory," *Methods in molecular biology (Clifton, N.J.)*, vol. 345, pp. 1–21, 2006.

[5] N. Normanno, A. M. Rachiglio, C. Roma, F. Fenizia, C. Esposito, R. Pasquale, M. L. La Porta, A. Iannaccone, F. Micheli, M. Santangelo, F. Bergantino, S. Costantini, and A. De Luca, "Molecular diagnostics and personalized medicine in oncology: challenges and opportunities," *Journal of cellular biochemistry*, vol. 114, pp. 514–524, Mar. 2013.

[6] X. Sun, S. Vilar, and N. P. Tatonetti, "High-throughput methods for combinatorial drug discovery," *Science Translational Medicine*, vol. 5, pp. 205rv1–205rv1, Oct. 2013.

[7] R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. S. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, U. Schopfer, and G. S. Sittampalam, "Impact of high-throughput screening in biomedical research," *Nature Reviews Drug Discovery*, vol. 10, pp. 188–195, Mar. 2011.

[8] E. Check Hayden, "Is the $1,000 genome for real?," *Nature*, Jan. 2014.

[9] L. A. McDonnell and R. M. A. Heeren, "Imaging mass spectrometry," *Mass Spectrometry Reviews*, vol. 26, pp. 606–643, Aug. 2007.

[10] K. Wetterstand, M.S., "DNA sequencing costs: Data from the NHGRI genome sequencing program (GSP)."

[11] E. R. Mardis, "The $1,000 genome, the $100,000 analysis?," *Genome Medicine*, vol. 2, no. 11, p. 84, 2010.

[12] D. W. Mount, *Bioinformatics: sequence and genome analysis*. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory Press, 2nd ed ed., 2004.

[13] J. Quackenbush, "Microarray data normalization and transformation," *Nature Genetics*, vol. 32 Suppl, pp. 496–501, Dec. 2002.

[14] D. Venet, V. Detours, and H. Bersini, "A measure of the signal-to-noise ratio of microarray samples and studies using gene correlations," *PLoS ONE*, vol. 7, p. e51013, Dec. 2012.

[15] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc., 1995.

[16] V. Vapnik, S. E. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation, and signal processing," in *Advances in Neural Information Processing Systems 9*, pp. 281–287, MIT Press, 1996.

[17] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: illustrations, sources and a solution," *BMC bioinformatics*, vol. 8, p. 25, 2007.

[18] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, no. 1, p. 307, 2008.

[19] D. Popovic, A. Sifrim, C. Moschopoulos, Y. Moreau, and B. De Moor, "A hybrid approach to feature ranking for microarray data classification," in *Engineering Applications of Neural Networks* (L. Iliadis, H. Papadopoulos, and C. Jayne, eds.), vol. 384 of *Communications in Computer and Information Science*, pp. 241–248, Springer Berlin Heidelberg, 2013.

[20] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, Oct 2001.

[21] K. Strimbu and J. A. Tavel, "What are biomarkers?," *Current opinion in HIV and AIDS*, vol. 5, pp. 463–466, Nov. 2010.

[22] Biomarkers Definitions Working Group., "Biomarkers and surrogate endpoints: preferred definitions and conceptual framework," *Clinical Pharmacology and Therapeutics*, vol. 69, pp. 89–95, Mar. 2001.

[23] L. Liotta and E. Petricoin, "Molecular profiling of human cancer," *Nature Reviews. Genetics*, vol. 1, pp. 48–56, Oct. 2000.

[24] E. F. Petricoin, K. C. Zoon, E. C. Kohn, J. C. Barrett, and L. A. Liotta, "Clinical proteomics: translating benchside promise into bedside reality," *Nature Reviews. Drug Discovery*, vol. 1, pp. 683–695, Sept. 2002.

[25] L. Pusztai and B. Leyland-Jones, "Promises and caveats of in silico biomarker discovery," *British Journal of Cancer*, vol. 99, pp. 385–386, Aug. 2008.

[26] Q. Ferry, J. Steinberg, C. Webber, D. R. FitzPatrick, C. P. Ponting, A. Zisserman, and C. Nellåker, "Diagnostically relevant facial gestalt information from ordinary photos," *eLife*, vol. 3, 2014.

[27] J. Dekervel, D. Hompes, H. van Malenstein, D. Popovic, X. Sagaert, B. De Moor, E. Van Cutsem, A. D'Hoore, C. Verslype, and J. van Pelt, "Hypoxia-driven gene expression is an independent prognostic factor in stage II and III colon cancer patients," *Clinical Cancer Research*, vol. 20, pp. 2159–2168, Apr. 2014.

[28] J. J. Goeman, S. A. van de Geer, F. de Kort, and H. C. van Houwelingen, "A global test for groups of genes: testing association with a clinical outcome," *Bioinformatics*, vol. 20, pp. 93–99, Jan. 2004.

[29] R. N. Jorissen, P. Gibbs, M. Christie, S. Prakash, L. Lipton, J. Desai, D. Kerr, L. A. Aaltonen, D. Arango, M. Kruhøffer, *et al.*, "Metastasis-associated gene expression changes predict poor outcomes in patients with dukes stage b and c colorectal cancer," *Clinical Cancer Research*, vol. 15, no. 24, pp. 7642–7651, 2009.

[30] S. Kaiser, Y.-K. Park, J. L. Franklin, R. B. Halberg, M. Yu, W. J. Jessen, J. Freudenberg, X. Chen, K. Haigis, A. G. Jegga, *et al.*, "Transcriptional recapitulation and subversion of embryonic colon development by mouse colon tumor models and human colon cancer," *Genome biology*, vol. 8, no. 7, p. R131, 2007.

[31] J. J. Smith, N. G. Deane, F. Wu, N. B. Merchant, B. Zhang, A. Jiang, P. Lu, J. C. Johnson, C. Schmidt, C. E. Bailey, *et al.*, "Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer," *Gastroenterology*, vol. 138, no. 3, pp. 958–968, 2010.

[32] A. Vodolazkaia, Y. El-Aalamat, D. Popovic, A. Mihalyi, X. Bossuyt, C. M. Kyama, A. Fassbender, A. Bokor, D. Schols, D. Huskens, C. Meuleman, K. Peeraer, C. Tomassetti, O. Gevaert, E. Waelkens, A. Kasran, B. De Moor, and T. M. D'Hooghe, "Evaluation of a panel of 28 biomarkers for the non-invasive diagnosis of endometriosis," *Human Reproduction*, vol. 27, pp. 2698–2711, Sept. 2012.

[33] Z. Birnbaum *et al.*, "On a use of the mann-whitney statistic," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, The Regents of the University of California, 1956.

[34] J. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

[35] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[36] S. Davies and S. Russell, "Np-completeness of searches for smallest possible feature sets," in *Proceedings of the 1994 AAAI fall symposium on relevance*, pp. 37–39, 1994.

[37] E. Amaldi and V. Kann, "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems," *Theoretical Computer Science*, vol. 209, no. 1–2, pp. 237 – 260, 1998.

[38] D. Popovic, A. Sifrim, G. Pavlopoulos, Y. Moreau, and B. Moor, "A simple genetic algorithm for biomarker mining," in *Pattern Recognition in Bioinformatics* (T. Shibuya, H. Kashima, J. Sese, and S. Ahmad, eds.), vol. 7632 of *Lecture Notes in Computer Science*, pp. 222–232, Springer Berlin Heidelberg, Jan. 2012.

[39] C. Moschopoulos, D. Popovic, A. Sifrim, G. Beligiannis, B. Moor, and Y. Moreau, "A genetic algorithm for pancreatic cancer diagnosis," in *Engineering Applications of Neural Networks* (L. Iliadis, H. Papadopoulos, and C. Jayne, eds.), vol. 384 of *Communications in Computer and Information Science*, pp. 222–230, Springer Berlin Heidelberg, Jan. 2013.

[40] D. Popovic, C. Moschopoulos, R. Sakai, A. Sifrim, J. Aerts, Y. Moreau, and B. D. Moor, "A self-tuning genetic algorithm with applications in biomarker discovery," in *Computer-Based Medical Systems (CBMS), 2014 IEEE 27th International Symposium on*, pp. 233–238, May 2014.

[41] D. Popovic, A. Sifrim, C. Moschopoulos, Y. Moreau, and B. De Moor, "Estsimating random forest feature importance using a genetic algorithm

approach," *Accepted for publication in Artificial Intelligence in Medicine with major revision*, 2014.

[42] Y. A. Meng, Y. Yu, L. A. Cupples, L. A. Farrer, and K. L. Lunetta, "Performance of random forest when SNPs are in linkage disequilibrium," *BMC Bioinformatics*, vol. 10, no. 1, p. 78, 2009.

[43] E. Iacucci, L.-C. Tranchevent, D. Popovic, G. A. Pavlopoulos, B. De Moor, R. Schneider, and Y. Moreau, "ReLiance: a machine learning and literature-based prioritization of receptor–ligand pairings," *Bioinformatics*, vol. 28, pp. i569–i574, Sept. 2012.

[44] E. Iacucci, D. Popovic, G. A. Pavlopoulos, L.-C. Tranchevent, M. Bauters, B. De Moor, and Y. Moreau, "Towards better prioritization of epigenetically modified dna regions," in *Artificial Intelligence: Theories and Applications*, pp. 270–277, Springer Berlin Heidelberg, 2012.

[45] C. Moschopoulos, D. Popovic, R. Langone, J. Suykens, Y. Moreau, and B. De Moor, "Gene interaction networks boost genetic algorithm perfomance in biomarker discovery," *Accepted for Proceedings of IEEE Symposium Series on Computational Intelligence (IEEE SSCI 2014)*, 2014.

[46] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering, *et al.*, "String v9. 1: protein-protein interaction networks, with increased coverage and integration," *Nucleic acids research*, vol. 41, no. D1, pp. D808–D815, 2013.

[47] D. Venet, J. E. Dumont, and V. Detours, "Most random gene expression signatures are significantly associated with breast cancer outcome," *PLoS Computational Biology*, vol. 7, p. e1002240, Oct. 2011.

[48] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[49] B. Taskar, P. Abbeel, and D. Koller, "Discriminative probabilistic models for relational data," in *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pp. 485–492, Morgan Kaufmann Publishers Inc., 2002.

[50] P. C. Ng and S. Henikoff, "SIFT: predicting amino acid changes that affect protein function," *Nucleic acids research*, vol. 31, pp. 3812–3814, July 2003.

[51] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nature methods*, vol. 7, pp. 248–249, Apr. 2010.

[52] N. Japkowicz, *Evaluating Learning Algorithms: a classification perspective*. Cambridge ; New York: Cambridge University Press, 2011.

[53] J. R. Lupski, J. G. Reid, C. Gonzaga-Jauregui, D. Rio Deiros, D. C. Y. Chen, L. Nazareth, M. Bainbridge, H. Dinh, C. Jing, D. A. Wheeler, A. L. McGuire, F. Zhang, P. Stankiewicz, J. J. Halperin, C. Yang, C. Gehman, D. Guo, R. K. Irikat, W. Tom, N. J. Fantin, D. M. Muzny, and R. A. Gibbs, "Whole-genome sequencing in a patient with charcot-marie-tooth neuropathy," *The New England journal of medicine*, vol. 362, pp. 1181–1191, Apr. 2010.

[54] M. C. Lopes, C. Joyce, G. R. S. Ritchie, S. L. John, F. Cunningham, J. Asimit, and E. Zeggini, "A combined functional annotation score for non-synonymous variants," *Human heredity*, vol. 73, no. 1, pp. 47–51, 2012.

[55] J. M. Schwarz, C. Rödelsperger, M. Schuelke, and D. Seelow, "MutationTaster evaluates disease-causing potential of sequence alterations," *Nature Methods*, vol. 7, pp. 575–576, Aug. 2010.

[56] S. Chun and J. C. Fay, "Identification of deleterious mutations within three human genomes," *Genome Research*, vol. 19, pp. 1553–1561, Sept. 2009.

[57] A. Sifrim, D. Popovic, L.-C. Tranchevent, A. Ardeshirdavani, R. Sakai, P. Konings, J. R. Vermeesch, J. Aerts, B. De Moor, and Y. Moreau, "eXtasy: variant prioritization by genomic data fusion," *Nature Methods*, vol. 10, pp. 1083–1084, Sept. 2013.

[58] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau, "Gene prioritization through genomic data fusion," *Nature Biotechnology*, vol. 24, pp. 537–544, May 2006.

[59] A. Sifrim, D. Popovic, J. R. Vermeesch, J. Aerts, B. De Moor, and Y. Moreau, "Comparison of aggregation methods for multiphenotype exomic variant prioritization," *Submitted for publication in Plos ONE*, 2014.

[60] D. Popovic, A. Sifrim, Y. Moreau, and B. De Moor, "extasy simplified-towards opening the black box," in *Proceedings of 2013 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 24–28, 2013.

[61] D. Popovic, A. Sifrim, J. Davis, Y. Moreau, and B. De Moor, "Improving performance of the extasy model by hierarchical sampling," in *Pattern Recognition in Bioinformatics*, vol. 8626 of *Lecture Notes in Computer Science*, pp. 125–128, Springer Berlin Heidelberg, Aug. 2014.

[62] D. Popovic, A. Sifrim, J. Davis, Y. Moreau, and B. De Moor, "Problems with nested granularity of feature domains in bioinformatics - the extasy case.," *Accepted for publication in BMC Bioinformatics*, 2014.

[63] L.-C. Tranchevent, R. Barriot, S. Yu, S. Van Vooren, P. Van Loo, B. Coessens, B. De Moor, S. Aerts, and Y. Moreau, "ENDEAVOUR update: a web resource for gene prioritization in multiple species," *Nucleic acids research*, vol. 36, pp. W377–384, July 2008.

[64] S. Schuierer, L.-C. Tranchevent, U. Dengler, and Y. Moreau, "Large-scale benchmark of endeavour using metacore maps," *Bioinformatics*, vol. 26, no. 15, pp. 1922–1923, 2010.

[65] D. Börnigen, L.-C. Tranchevent, F. Bonachela-Capdevila, K. Devriendt, B. De Moor, P. De Causmaecker, and Y. Moreau, "An unbiased evaluation of gene prioritization tools," *Bioinformatics*, vol. 28, no. 23, pp. 3081–3088, 2012.

[66] T. De Bie, L.-C. Tranchevent, L. M. Van Oeffelen, and Y. Moreau, "Kernel-based data fusion for gene prioritization," *Bioinformatics*, vol. 23, no. 13, pp. i125–i132, 2007.

[67] D. Nitsch, L.-C. Tranchevent, B. Thienpont, L. Thorrez, H. Van Esch, K. Devriendt, and Y. Moreau, "Network analysis of differential expression for the identification of disease-causing genes," *PloS one*, vol. 4, no. 5, p. e5526, 2009.

[68] D. Nitsch, L.-C. Tranchevent, J. P. Goncalves, J. K. Vogt, S. C. Madeira, and Y. Moreau, "Pinta: a web server for network-based gene prioritization from expression data," *Nucleic acids research*, vol. 39, no. suppl 2, pp. W334–W338, 2011.

[69] S. Yu, L.-C. Tranchevent, B. De Moor, and Y. Moreau, "Gene prioritization and clustering by multi-view text mining," *BMC bioinformatics*, vol. 11, no. 1, p. 28, 2010.

[70] L.-C. Tranchevent, F. B. Capdevila, D. Nitsch, B. De Moor, P. De Causmaecker, and Y. Moreau, "A guide to web tools to prioritize candidate genes," *Briefings in bioinformatics*, vol. 12, no. 1, pp. 22–32, 2011.

[71] Y. Moreau and L.-C. Tranchevent, "Computational tools for prioritizing candidate genes: boosting disease gene discovery," *Nature Reviews Genetics*, vol. 13, pp. 523–536, July 2012.

[72] A. Javed, S. Agrawal, and P. C. Ng, "Phen-gen: combining phenotype and genotype to analyze rare disorders," *Nature methods*, 2014.

[73] S. Kumar, M. Sanderford, V. E. Gray, J. Ye, and L. Liu, "Evolutionary diagnosis method for variants in personal exomes," *Nature Methods*, vol. 9, pp. 855–856, Aug. 2012.

[74] S. Asthana, W. S. Noble, G. Kryukov, C. E. Grant, S. Sunyaev, and J. A. Stamatoyannopoulos, "Widely distributed noncoding purifying selection in the human genome," *Proceedings of the National Academy of Sciences*, vol. 104, pp. 12410–12415, July 2007.

[75] J. A. Tennessen, A. W. Bigham, T. D. O'Connor, W. Fu, E. E. Kenny, S. Gravel, S. McGee, R. Do, X. Liu, G. Jun, H. M. Kang, D. Jordan, S. M. Leal, S. Gabriel, M. J. Rieder, G. Abecasis, D. Altshuler, D. A. Nickerson, E. Boerwinkle, S. Sunyaev, C. D. Bustamante, M. J. Bamshad, J. M. Akey, Broad GO, Seattle GO, and on behalf of the NHLBI Exome Sequencing Project, "Evolution and functional impact of rare coding variation from deep sequencing of human exomes," *Science*, vol. 337, pp. 64–69, July 2012.

[76] N. Huang, I. Lee, E. M. Marcotte, and M. E. Hurles, "Characterising and predicting haploinsufficiency in the human genome," *PLoS Genetics*, vol. 6, p. e1001154, Oct. 2010.

[77] P. D. Stenson, M. Mort, E. V. Ball, K. Howells, A. D. Phillips, N. S. Thomas, and D. N. Cooper, "The human gene mutation database: 2008 update," *Genome Medicine*, vol. 1, no. 1, p. 13, 2009.

[78] C. L. Myers, D. R. Barrett, M. A. Hibbs, C. Huttenhower, and O. G. Troyanskaya, "Finding function: evaluation methods for functional genomic data," *BMC genomics*, vol. 7, p. 187, 2006.

[79] M. Yandell, C. Huff, H. Hu, M. Singleton, B. Moore, J. Xing, L. B. Jorde, and M. G. Reese, "A probabilistic disease-gene finder for personal genomes," *Genome research*, vol. 21, pp. 1529–1542, Sept. 2011.

[80] I. Ionita-Laza, V. Makarov, S. Yoon, B. Raby, J. Buxbaum, D. L. Nicolae, and X. Lin, "Finding disease variants in mendelian disorders by using sequence data: methods and applications," *American journal of human genetics*, vol. 89, pp. 701–712, Dec. 2011.

[81] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos, "The human phenotype ontology: a tool for annotating and analyzing human hereditary disease," *American journal of human genetics*, vol. 83, pp. 610–615, Nov. 2008.

[82] S. Köhler, M. H. Schulz, P. Krawitz, S. Bauer, S. Dölken, C. E. Ott, C. Mundlos, D. Horn, S. Mundlos, and P. N. Robinson, "Clinical

diagnostics in human genetics with semantic similarity searches in ontologies," *American journal of human genetics*, vol. 85, pp. 457–464, Oct. 2009.

[83] X. Liu, X. Jian, and E. Boerwinkle, "dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions," *Human mutation*, vol. 32, pp. 894–899, Aug. 2011.

[84] M. Pertea, G. M. Pertea, and S. L. Salzberg, "Detection of lineage-specific evolutionary changes among primate species," *BMC bioinformatics*, vol. 12, p. 274, 2011.

[85] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," *Genome research*, vol. 15, pp. 1034–1050, Aug. 2005.

[86] M. Vihinen, "How to evaluate performance of prediction methods? measures and their interpretation in variation effect analysis," *BMC genomics*, vol. 13 Suppl 4, p. S2, 2012.

[87] R. M. Durbin, D. L. Altshuler, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, F. S. Collins, F. M. De La Vega, P. Donnelly, M. Egholm, P. Flicek, S. B. Gabriel, R. A. Gibbs, B. M. Knoppers, E. S. Lander, H. Lehrach, E. R. Mardis, G. A. McVean, D. A. Nickerson, L. Peltonen, A. J. Schafer, S. T. Sherry, J. Wang, R. K. Wilson, D. Deiros, M. Metzker, D. Muzny, J. Reid, D. Wheeler, J. Li, M. Jian, G. Li, R. Li, H. Liang, G. Tian, B. Wang, J. Wang, W. Wang, H. Yang, X. Zhang, H. Zheng, L. Ambrogio, T. Bloom, K. Cibulskis, T. J. Fennell, D. B. Jaffe, E. Shefler, C. L. Sougnez, N. Gormley, S. Humphray, Z. Kingsbury, P. Koko-Gonzales, J. Stone, K. J. McKernan, G. L. Costa, J. K. Ichikawa, C. C. Lee, R. Sudbrak, T. A. Borodina, A. Dahl, A. N. Davydov, P. Marquardt, F. Mertes, W. Nietfeld, P. Rosenstiel, S. Schreiber, A. V. Soldatov, B. Timmermann, M. Tolzmann, J. Affourtit, D. Ashworth, S. Attiya, M. Bachorski, E. Buglione, A. Burke, A. Caprio, C. Celone, S. Clark, D. Conners, B. Desany, L. Gu, L. Guccione, K. Kao, A. Kebbel, J. Knowlton, M. Labrecque, L. McDade, C. Mealmaker, M. Minderman, A. Nawrocki, F. Niazi, K. Pareja, R. Ramenani, D. Riches, W. Song, C. Turcotte, S. Wang, D. Dooling, L. Fulton, R. Fulton, G. Weinstock, J. Burton, D. M. Carter, C. Churcher, A. Coffey, A. Cox, A. Palotie, M. Quail, T. Skelly, J. Stalker, H. P. Swerdlow, D. Turner, A. De Witte, S. Giles, M. Bainbridge, D. Challis, A. Sabo, F. Yu, J. Yu, X. Fang,

X. Guo, Y. Li, R. Luo, S. Tai, H. Wu, H. Zheng, X. Zheng, Y. Zhou, G. T. Marth, E. P. Garrison, W. Huang, A. Indap, D. Kural, W.-P. Lee, W. Fung Leong, A. R. Quinlan, C. Stewart, M. P. Stromberg, A. N. Ward, J. Wu, C. Lee, R. E. Mills, X. Shi, M. J. Daly, M. A. DePristo, A. D. Ball, E. Banks, B. L. Browning, K. V. Garimella, S. R. Grossman, R. E. Handsaker, M. Hanna, C. Hartl, A. M. Kernytsky, J. M. Korn, H. Li, J. R. Maguire, S. A. McCarroll, A. McKenna, J. C. Nemesh, A. A. Philippakis, R. E. Poplin, A. Price, M. A. Rivas, P. C. Sabeti, S. F. Schaffner, I. A. Shlyakhter, D. N. Cooper, E. V. Ball, M. Mort, A. D. Phillips, P. D. Stenson, J. Sebat, V. Makarov, K. Ye, S. C. Yoon, C. D. Bustamante, A. Boyko, J. Degenhardt, S. Gravel, R. N. Gutenkunst, M. Kaganovich, A. Keinan, P. Lacroute, X. Ma, A. Reynolds, L. Clarke, F. Cunningham, J. Herrero, S. Keenen, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, R. E. Smith, V. Zalunin, X. Zheng-Bradley, J. O. Korbel, A. M. Stütz, M. Bauer, R. Keira Cheetham, T. Cox, M. Eberle, T. James, S. Kahn, L. Murray, K. Ye, Y. Fu, F. C. L. Hyland, J. M. Manning, S. F. McLaughlin, H. E. Peckham, O. Sakarya, Y. A. Sun, E. F. Tsung, M. A. Batzer, M. K. Konkel, J. A. Walker, M. W. Albrecht, V. S. Amstislavskiy, R. Herwig, D. V. Parkhomchuk, R. Agarwala, H. M. Khouri, A. O. Morgulis, J. E. Paschall, L. D. Phan, K. E. Rotmistrovsky, R. D. Sanders, M. F. Shumway, C. Xiao, A. Auton, Z. Iqbal, G. Lunter, J. L. Marchini, L. Moutsianas, S. Myers, A. Tumian, J. Knight, R. Winer, D. W. Craig, S. M. Beckstrom-Sternberg, A. Christoforides, A. A. Kurdoglu, J. V. Pearson, S. A. Sinari, W. D. Tembe, D. Haussler, A. S. Hinrichs, S. J. Katzman, A. Kern, R. M. Kuhn, M. Przeworski, R. D. Hernandez, B. Howie, J. L. Kelley, S. Cord Melton, Y. Li, P. Anderson, T. Blackwell, W. Chen, W. O. Cookson, J. Ding, H. Min Kang, M. Lathrop, L. Liang, M. F. Moffatt, P. Scheet, C. Sidore, M. Snyder, X. Zhan, S. Zöllner, P. Awadalla, F. Casals, Y. Idaghdour, J. Keebler, E. A. Stone, M. Zilversmit, L. Jorde, J. Xing, E. E. Eichler, G. Aksay, C. Alkan, I. Hajirasouliha, F. Hormozdiari, J. M. Kidd, S. Cenk Sahinalp, P. H. Sudmant, K. Chen, A. Chinwalla, L. Ding, D. C. Koboldt, M. D. McLellan, J. W. Wallis, M. C. Wendl, Q. Zhang, C. A. Albers, Q. Ayub, S. Balasubramaniam, J. C. Barrett, Y. Chen, D. F. Conrad, P. Danecek, E. T. Dermitzakis, M. Hu, N. Huang, M. E. Hurles, H. Jin, L. Jostins, T. M. Keane, S. Quang Le, S. Lindsay, Q. Long, D. G. MacArthur, S. B. Montgomery, L. Parts, C. Tyler-Smith, K. Walter, Y. Zhang, M. B. Gerstein, M. Snyder, A. Abyzov, S. Balasubramanian, R. Bjornson, J. Du, F. Grubert, L. Habegger, R. Haraksingh, J. Jee, E. Khurana, H. Y. K. Lam, J. Leng, X. Jasmine Mu, A. E. Urban, Z. Zhang, C. Coafra, H. Dinh, C. Kovar, S. Lee, L. Nazareth, J. Wilkinson, A. Coffey, C. Scott, N. Gharani, J. S. Kaye, A. Kent, T. Li, A. L. McGuire, P. N.

Ossorio, C. N. Rotimi, Y. Su, L. H. Toji, L. D. Brooks, A. L. Felsenfeld, J. E. McEwen, A. Abdallah, C. R. Juenger, N. C. Clemm, A. Duncanson, E. D. Green, M. S. Guyer, and J. L. Peterson, "A map of human genome variation from population-scale sequencing," *Nature*, vol. 467, pp. 1061–1073, Oct. 2010.

[88] R. M. Kuhn, D. Karolchik, A. S. Zweig, T. Wang, K. E. Smith, K. R. Rosenbloom, B. Rhead, B. J. Raney, A. Pohl, M. Pheasant, L. Meyer, F. Hsu, A. S. Hinrichs, R. A. Harte, B. Giardine, P. Fujita, M. Diekhans, T. Dreszer, H. Clawson, G. P. Barber, D. Haussler, and W. J. Kent, "The UCSC genome browser database: update 2009," *Nucleic acids research*, vol. 37, pp. D755–761, Jan. 2009.

[89] A. Hoischen, B. W. M. van Bon, C. Gilissen, P. Arts, B. van Lier, M. Steehouwer, P. de Vries, R. de Reuver, N. Wieskamp, G. Mortier, K. Devriendt, M. Z. Amorim, N. Revencu, A. Kidd, M. Barbosa, A. Turner, J. Smith, C. Oley, A. Henderson, I. M. Hayes, E. M. Thompson, H. G. Brunner, B. B. A. de Vries, and J. A. Veltman, "De novo mutations of SETBP1 cause schinzel-giedion syndrome," *Nature genetics*, vol. 42, pp. 483–485, June 2010.

[90] J. K. J. Van Houdt, B. A. Nowakowska, S. B. Sousa, B. D. C. van Schaik, E. Seuntjens, N. Avonce, A. Sifrim, O. A. Abdul-Rahman, M.-J. H. van den Boogaard, A. Bottani, M. Castori, V. Cormier-Daire, M. A. Deardorff, I. Filges, A. Fryer, J.-P. Fryns, S. Gana, L. Garavelli, G. Gillessen-Kaesbach, B. D. Hall, D. Horn, D. Huylebroeck, J. Klapecki, M. Krajewska-Walasek, A. Kuechler, M. A. Lines, S. Maas, K. D. Macdermot, S. McKee, A. Magee, S. A. de Man, Y. Moreau, F. Morice-Picard, E. Obersztyn, J. Pilch, E. Rosser, N. Shannon, I. Stolte-Dijkstra, P. Van Dijck, C. Vilain, A. Vogels, E. Wakeling, D. Wieczorek, L. Wilson, O. Zuffardi, A. H. C. van Kampen, K. Devriendt, R. Hennekam, and J. R. Vermeesch, "Heterozygous missense mutations in SMARCA2 cause nicolaides-baraitser syndrome," *Nature genetics*, vol. 44, pp. 445–449, S1, Apr. 2012.

[91] S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure, and M. J. Bamshad, "Exome sequencing identifies the cause of a mendelian disorder," *Nature Genetics*, vol. 42, pp. 30–35, Jan. 2010.

[92] G. J. Lyon and K. Wang, "Identifying disease mutations in genomic medicine settings: current challenges and how to accelerate progress," *Genome Medicine*, vol. 4, no. 7, p. 58, 2012.

[93] M. Choi, U. I. Scholl, W. Ji, T. Liu, I. R. Tikhonova, P. Zumbo, A. Nayir, A. Bakkaloglu, S. Ozen, S. Sanjad, C. Nelson-Williams, A. Farhi, S. Mane, and R. P. Lifton, "Genetic diagnosis by whole exome capture and massively parallel DNA sequencing," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 19096–19101, Nov. 2009.

[94] S. B. Ng, A. W. Bigham, K. J. Buckingham, M. C. Hannibal, M. J. McMillin, H. I. Gildersleeve, A. E. Beck, H. K. Tabor, G. M. Cooper, H. C. Mefford, C. Lee, E. H. Turner, J. D. Smith, M. J. Rieder, K.-i. Yoshiura, N. Matsumoto, T. Ohta, N. Niikawa, D. A. Nickerson, M. J. Bamshad, and J. Shendure, "Exome sequencing identifies MLL2 mutations as a cause of kabuki syndrome," *Nature Genetics*, vol. 42, pp. 790–793, Sept. 2010.

[95] C. Gilissen, A. Hoischen, H. G. Brunner, and J. A. Veltman, "Disease gene identification strategies for exome sequencing," *European journal of human genetics: EJHG*, vol. 20, pp. 490–497, May 2012.

[96] L. E. L. M. Vissers, J. de Ligt, C. Gilissen, I. Janssen, M. Steehouwer, P. de Vries, B. van Lier, P. Arts, N. Wieskamp, M. del Rosario, B. W. M. van Bon, A. Hoischen, B. B. A. de Vries, H. G. Brunner, and J. A. Veltman, "A de novo paradigm for mental retardation," *Nature genetics*, vol. 42, pp. 1109–1112, Dec. 2010.

[97] G. V. Kryukov, L. A. Pennacchio, and S. R. Sunyaev, "Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies," *The American Journal of Human Genetics*, vol. 80, pp. 727–739, Apr. 2007.

[98] B. D. O'Fallon, W. Wooderchak-Donahue, P. Bayrak-Toydemir, and D. Crockett, "VarRanker: rapid prioritization of sequence variations associated with human disease," *BMC bioinformatics*, vol. 14 Suppl 13, p. S1, 2013.

[99] P. N. Robinson, S. Kohler, A. Oellrich, Sanger Mouse Genetics Project, K. Wang, C. J. Mungall, S. E. Lewis, N. Washington, S. Bauer, D. Seelow, P. Krawitz, C. Gilissen, M. Haendel, and D. Smedley, "Improved exome prioritization of disease genes through cross-species phenotype comparison," *Genome Research*, vol. 24, pp. 340–348, Feb. 2014.

[100] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *Science (New York, N.Y.)*, vol. 302, pp. 249–255, Oct. 2003.

[101] R. Kolde, S. Laur, P. Adler, and J. Vilo, "Robust rank aggregation for gene list integration and meta-analysis," *Bioinformatics*, vol. 28, pp. 573–580, Feb. 2012.

[102] R. A. Fisher, *Statistical methods for research workers.* Edinburgh: Oliver and Boyd, 14th ed., revised and enlarged ed., 1970.

[103] A. Rope, K. Wang, R. Evjenth, J. Xing, J. Johnston, J. Swensen, W. Johnson, B. Moore, C. Huff, L. Bird, J. Carey, J. Opitz, C. Stevens, T. Jiang, C. Schank, H. Fain, R. Robison, B. Dalley, S. Chin, S. South, T. Pysher, L. Jorde, H. Hakonarson, J. Lillehaug, L. Biesecker, M. Yandell, T. Arnesen, and G. Lyon, "Using VAAST to identify an x-linked disorder resulting in lethality in male infants due to n-terminal acetyltransferase deficiency," *The American Journal of Human Genetics*, vol. 89, pp. 28–43, July 2011.

[104] J. T. Kost and M. P. McDermott, "Combining dependent p-values," *Statistics & Probability Letters*, vol. 60, pp. 183–190, Nov. 2002.

[105] M. B. Brown, "400: A method for combining non-independent, one-sided tests of significance," *Biometrics*, vol. 31, p. 987, Dec. 1975.

[106] J. F. Elder, "The generalization paradox of ensembles," *Journal of Computational and Graphical Statistics*, vol. 12, pp. 853–864, Dec. 2003.

[107] C. Strobl and A. Zeileis, "Danger: High power! exploring the statistical properties of a test for random forest variable importance," in *Proceedings of the 18th International Conference on Computational Statistics*, (Porto, Portugal), 2008.

[108] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational Statistics & Data Analysis*, vol. 52, pp. 2249–2260, Jan. 2008.

[109] L. Tolosi and T. Lengauer, "Classification with correlated features: unreliability of feature ranking and solutions," *Bioinformatics*, vol. 27, pp. 1986–1994, July 2011.

[110] Z.-H. Zhou, Y. Jiang, and S.-F. Chen, "Extracting symbolic rules from trained neural network ensembles," *AI Commun.*, vol. 16, pp. 3–15, Jan. 2003.

[111] L. Fu, "Rule generation from neural networks," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 24, pp. 1114–1124, Aug. 1994.

[112] P. Domingos, "Knowledge discovery via multiple models," *Intelligent Data Analysis*, vol. 2, pp. 187–202, 1998.

[113] I. R. König, J. Malley, C. Weimar, H.-C. Diener, and A. Ziegler, "Practical experiences on the necessity of external validation," *Statistics in medicine*, vol. 26, no. 30, pp. 5499–5511, 2007.

[114] E. C. Gunther, D. J. Stone, R. W. Gerwien, P. Bento, and M. P. Heyes, "Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro," *Proceedings of the national academy of sciences*, vol. 100, no. 16, pp. 9608–9613, 2003.

[115] R. Tsenkova, H. Meilina, S. Kuroki, and D. H. Burns, "Near infrared spectroscopy using short wavelengths and leave-one-cow-out cross-validation for quantification of somatic cells in milk," *Journal of Near Infrared Spectroscopy*, vol. 17, no. 6, pp. 345–351, 2010.

[116] B. Efron, "Bootstrap methods: Another look at the jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[117] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, pp. 123–140, Aug. 1996.

[118] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[119] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[120] J. P. Shaffer, "Multiple hypothesis testing," *Annual review of psychology*, vol. 46, no. 1, pp. 561–584, 1995.

[121] T. Fawcett, "An introduction to roc analysis," *Pattern Recogn. Lett.*, vol. 27, pp. 861–874, June 2006.

[122] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, ACM, 2006.

[123] R. Polikar, "Ensemble based systems in decision making," *Circuits and Systems Magazine, IEEE*, vol. 6, no. 3, pp. 21–45, 2006.

[124] J. M. Schwarz, D. N. Cooper, M. Schuelke, and D. Seelow, "MutationTaster2: mutation prediction for the deep-sequencing age," *Nature Methods*, vol. 11, pp. 361–362, Mar. 2014.

[125] "Random forests." `http://www.stat.berkeley.edu/~breiman/RandomForests/cc_philosophy.htm`. Accessed: 15-08-2014.

Curriculum vitae

Dusan Popovic was born on April 17th 1981 in Belgrade, Serbia. He was attending "Majka Jugovic" primary school and "Mathematical Gymnasium" secondary school. In parallel with secondary education, he attended primary musical school "Kosta Manojlovic", section for double bass.

He obtained the degree of master of science in engineering at the University of Belgrade Faculty of Organizational sciences, option Information systems. In addition, he obtained advanced master degree in Engineering (Artificial intelligence) at the KU Leuven, Belgium.

In parallel with the studies, he was working in family-owned construction company and professionally played bass guitar. He started his PhD in Engineering at the KU Leuven in year of 2010 under supervision of Prof. Bart De Moor as a member of the STADIUS research group.

During the PhD research mandate in STADIUS, Dusan have authored and co-authored 20 scientific manuscripts (from which 16 published or accepted for publication at the moment), numerous conference abstracts and posters, as well as several grant proposals.

He also assumed a number of professional duties, including teaching, supervision of master students, reviewer work for several journals and conference proceedings and participation in programme committees of various research meetings and symposiums.

List of publications

International Journal Papers:

- Dekervel J., Hompes D., Van Malenstein H., Popovic D., Sagaert X., De Moor B., Van Cutsem E., D'hoore A., Verslype C., Van Pelt J., "Hypoxia driven gene expression is an independent prognostic factor in stage II and III colon cancer patients", Clinical Cancer Research, vol. 20, no. 8, Apr. 2014, pp. 2159-2168.
- Sifrim A.*, Popovic D.*, Tranchevent L.-C., Ardeshirdavani A., Sakai R., Konings P., Vermeesch J., Aerts J., De Moor B., Moreau Y., "eXtasy: Variant Prioritization By Genomic Data Fusion", Nature Methods, 10, 1083–1084, Sep.2013
  *Equally contributing first authors
- Vodolazkaia A., El Aalamat Y., Popovic D., Mihalyi A., Bossuyt X., Kyama C., Fassbender A., Bokor A., Schols D., Huskens D., Meuleman C., Peeraer K., Tomassetti C., Gevaert O., Waelkens E., Kasran A., De Moor B., D'Hooghe T., "Evaluation of a panel of 28 biomarkers for a non-invasive diagnosis of endometriosis", Human Reproduction, vol. 27, no. 9, Sep. 2012, pp. 2698-2711.
- Iacucci E., Tranchevent LC., Popovic D., Pavlopoulos GA., De Moor B., Schneider R., Moreau Y., "ReLiance: a machine learning and literature-based prioritization of receptor—ligand pairings", Bioinformatics, vol. 28, no. 18, Sep. 2012, pp. i569-i574.

International Conference Papers :

- Popovic, D., Sifrim, A., Davis, J., Moreau, Y., De Moor, B. (2014). "Improving performance of the eXtasy model by hierarchical sampling.", in Proc. of the Pattern Recognnition in Bioinformatics - 9th IAPR International Conference PRIB 2014, Stockholm, Sweden, Aug 2014, pp. 125-128.

- Popovic D., Moschopoulos C., Sakai R., Sifrim A., Aerts J., Moreau Y., De Moor B.: "A self-tuning genetic algorithm with applications in biomarker discovery," in Computer-Based Medical Systems (CBMS), 2014 IEEE 27th International Symposium on, pp. 233–238, May 2014.

- Popovic D.*, Sifrim A.*, Moreau Y., De Moor B., "eXtasy simplified - towards opening the black box", in Proc. of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shanghai, China, Dec. 2013, pp. 24-28.
  *Equally contributing first authors

- Moschopoulos C., Popovic D., Sifrim A., Beligiannis G., De Moor B., Moreau Y., "A genetic algorithm for pancreatic cancer diagnosis", in Engineering Applications of Neural Networks (Communications in Computer and Information Science) 384 (EANN2013), Chalkidiki, Greece, Sep. 2013, pp. 222-230.

- Popovic D., Sifrim A., Moschopoulos C., Moreau Y., De Moor B., "A Hybrid Approach to Feature Ranking for Microarray Data Classification", in Engineering Applications of Neural Networks (Communications in Computer and Information Science) Volume 384 (EANN 2013), Chalkidiki, Greece, Sep. 2013, pp. 241-248.

- Popovic D., Sifrim A., Pavlopoulos G.A., Moreau Y., De Moor B., "A Simple Genetic Algorithm for Biomarker Mining", in Springer's Lecture Notes in Computer Science / Lecture Notes in Bioinformatics (PRIB 2012), Tokyo, Japan, Nov. 2012, pp. 222-232.

- Iacucci E., Tranchevent L.-C., Popovic D., Pavlopoulos G. A., De Moor B., Schneider R., Moreau Y., "A Bioinformatics E-dating Story: Computational Prediction and Prioritization of Receptor-Ligand Pairs", in Proc. of the International Conference on Intelligent Systems for

Molecular Biology (ISMB SCS 2012) published in special issue of BMC Bioinformatics, vol. 13, suppl. 18, Long Beach, California, Jul. 2012, A7 p.

- Iacucci E., Popovic D., Pavlopoulos G., Tranchevent L-C., De Moor B., Moreau Y., "Towards Better Prioritization of Epigenetically Modified DNA Regions", in Springer's Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence (SETN 2012), Lamia, Greece, May 2012.

Forthcoming publications :

- Moschopoulos C.*, Popovic D.*, Langone R., Suykens JAK., De Moor B., Moreau Y., "Gene interaction networks boost genetic algorithm perfomance in biomarker discovery", Accepted for publication in Proceedings of IEEE Symposium Series on Computational Intelligence (IEEE SSCI 2014)
  *Equally contributing first authors
- Popovic, D., Sifrim, A., Davis, J., Moreau, Y., De Moor, B. (2014). "Problems with nested granularity of feature domains in bioinformatics - the eXtasy case." Accepted for publication in BMC Bioinformatics
- Popovic D., Sifrim A., Moschopoulos C., Moreau Y., De Moor B.: "Estimating Random Forest feature importance using a genetic algorithm approach", Accepted for publication in Artificial Intelligence in Medicine with major revision
- Moschopoulos C., Popovic D., Sifrim A., Beligiannis G., De Moor B., Moreau Y.: "Applying a genetic algorithm on expression data to reveal the best set of biomarkers", Accepted for publication in Artificial Intelligence in Medicine with major revision
- Sifrim A., Popovic D., Vermeesch J., Aerts J., De Moor B., Moreau Y.,"Comparison of aggregation methods for multiphenotype exomic variant prioritization", Submitted for publication in PLOS One
- Dekervel J., Popovic D., van Malenstein H., Windmolders P., Libbrecht L., Bulle A., De Moor B., Van Cutsem E., Nevens F., Verslype C., van Pelt J.:"A gene signature predicts early and late recurrence after resection of hepatocellular carcinoma", Submitted for publication in Journal of Hepatology