UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

CORSO DI DOTTORATO DI RICERCA IN: INGEGNERIA DELL' INFORMAZIONE
CURRICOLO: BIOINGEGNERIA
CICLO: XXXII

DEVELOPMENT AND ASSESSMENT OF BIOINFORMATICS METHODS FOR PERSONALIZED MEDICINE

Coordinatore: Ch.mo Prof. Andrea Neviani
Supervisore: Ch.mo Prof. Carlo Ferrari

Dottorando : Francesco Reggiani

TABLE OF CONTENTS

Sommario

Il genoma umano è una risorsa ricca di informazioni per i ricercatori che si dedicano allo studio delle patologie complesse. L'obiettivo di questo genere di ricerche è giungere ad una migliore comprensione di queste malattie e quindi sviluppare nuove strategie terapeutiche per la cura dei pazienti affetti. Dall'inizio di questo secolo, un numero crescente di tecnologie per il sequenziamento del DNA sono state sviluppate, sono conosciute come tecnologie "*Next Generation Sequencing*" (NGS). Le tecnologie NGS hanno gradualmente diminuito il costo del sequenziamento di un genoma umano fino a circa 1000 dollari, ciò ha consentito l'utilizzo di questi strumenti nella pratica clinica e nella ricerca, in particolare negli studi di associazione *genome-wide* o "*Genome-wide association studies*" (GWAS). Questi lavori hanno portato alla luce l'associazione di alcune varianti con alcune patologie o caratteri complessi. Queste varianti potrebbero essere utilizzate per valutare il rischio che un individuo sviluppi una particolare patologia. Sfortunatamente diverse sorgenti di errore sono in grado di ostacolare l'uso e l'interpretazione dei dati genomici: da una parte abbiamo il rumore legato al processo di sequenziamento e gli errori di allineamento delle *reads*. Dall'altra parte gli SNP non sempre possono essere utilizzati in modo affidabile per predire l'insorgenza della malattia a cui sono stati associati. Il *Critical Assessment of Genome Interpretation* è stato organizzato con l'obiettivo di definire lo stato dell'arte nei metodi che stimano l'effetto di variazioni genetiche a livello molecolare o fenotipico. Negli anni il CAGI ha dato vita a più competizioni in cui diversi gruppi di ricerca hanno testato i loro metodi di predizione su diversi *dataset* condivisi. L'assenza di linee generali su come condurre la valutazione delle performance dei predittori, ha reso difficile un confronto fra metodi sviluppati in edizioni diverse del CAGI.

In questo contesto, il progetto di dottorato si è focalizzato nello sviluppo di un software per la valutazione di metodi di apprendimento automatici basati sulla regressione o la predizione di fenotipi multipli. Questo strumento si fonda su criteri di analisi della performance, derivanti dalla letteratura e da precedenti esperimenti del CAGI. Questo software è stato sviluppato in R ed utilizzato per ripetere o valutare ex novo la qualità dei predittori in un gran numero di esperimenti del CAGI. Le conoscenze acquisite durante lo

sviluppo di questo progetto, sono state utilizzate per valutare due competizioni del CAGI 5: la *Pericentriolar Material 1* (PCM1) e il Pannello per le Disabilità Intellettive (ID). L'esperienza derivante dal completamento dei lavori precedentemente elencati, ha guidato lo sviluppo e il miglioramento delle prestazioni di un metodo predittivo. In particolare è stato sviluppato un software per la predizione dei livelli di colesterolo, basato su dati genotipici, di cui è stata testata la validità con criteri matematici allo stato dell'arte. Questo strumento è stato la pietra portante di un progetto fondato dal Ministero della Salute Italiano.

Abstract

The human genome is a source of information for researchers that study complex diseases with the perspective of a better understanding of the pathologies and the development of new therapeutic strategies. Starting from the beginning of the current century, a growing number of technologies devoted to DNA sequencing have emerged, generally referred to as Next Generation Sequencing (NGS) technologies. NGS gradually decreased the cost of sequencing a human genome to around US$1000, enabling the use of these technologies for clinical and research purposes, such as Genome-wide association studies (GWAS). GWAS studies have enlightened the presence of disease-associated loci, in particular variants that could be used to evaluate the risk of an individual to develop a disease.

Unfortunately, different sources of errors are able to impair the interpretation and use of NGS data: on the one hand, we have noise related to the process of DNA sequencing and read alignment errors, which could lead to false positive calls or artifacts. On the other hand, variants could be poor predictors for the manifestation of their associated disease. Nowadays the challenge of genomic data interpretation has driven the research towards the development of methods for the analysis and interpretation of genomic variations, eventually predicting the probability of a patient to develop a definite disease. A fair evaluation of these tools is essential to understand the applicability of the presented methods in clinical practice. The Critical Assessment of Genome Interpretation (CAGI) has been developed with the aim of defining the current state of art in terms of methods for predicting the impact of genomic changes at molecular and phenotype levels. CAGI is a community-driven experiment in which different prediction methods, developed by a set of invited groups, are evaluated on a common dataset. Unfortunately, no common guidelines were given to evaluate the tools presented in CAGI experiments, this has made the comparison between different CAGI experiments cumbersome, since different mathematical indexes and scripts have been used to evaluate the involved methods.

My PhD project has been focused on the development of software for the assessment of machine learning methods in regression and multiple phenotype challenges. This tool is based on state of the art assessment principles, derived from literature or previous CAGI

experiments. This software is available as an R package and has been used to repeat or perform new assessments on a wide range of CAGI experiments. The knowledge acquired during the development of this project was used to evaluate two CAGI 5 challenges: Pericentriolar Material 1 (PCM1) and Intellectual Disability (ID) panel. The experience I have acquired, through the development of all previously mentioned works, has led the improvement and assessment of a machine learning method. In particular, I have developed a software for the prediction of cholesterol levels, based on genotype data. Eventually I have tested the reliability of this method. This tool was the milestone in a project founded by the Italian Ministry of Health.

Introduction

Nowadays an individual whole genome sequence can be easily obtained, compared to the economic and technical efforts of some decades ago. This means that we could have access to all genomic variants of an individual and use this information to predict the development of a particular disease. This is the major aim of precision medicine: to detect, prevent or treat specific pathologies with a strategy based on the individual genome. This process is still far from being current practice. The fact is that the ideal framework for personalized medicine is based on accurate methods that are able to map variants to phenotypes, unfortunately this is far from true. The aim of my thesis project was to investigate the issue of model quality assessment, this problem has been addressed in the evaluation of different CAGI challenges, with the aim of defining the state of the art in methods for variant effect or related phenotype prediction. Finally I used the assessment methods developed in the previous part of my thesis project to understand the feasibility of a bioinformatic tool for cholesterol level prediction in a clinical environment.

Information theory and biological data analysis

The model of a general communication system presented by Shannon (Shannon, 1948), could be helpful to understand the problem of data analysis in a biomedical context. The sequencing of a molecule of DNA can be considered as a process in which the genetic information (information source) is converted in data (message) by a software (the transmitter). The produced data can be interpreted by an amount of different algorithms (the receiver), that produces an output for the researcher (destination). While in the original diagram purposed by Shannon a source of noise was affecting the passage of information from the transmitter to the receiver, the interpretation of the information related to DNA or protein sequence is the step that is most affected by noise. In principle we can have noise due to the read alignment procedure or technical limitations of the sequencing platform (Goodwin et al., 2016), that could lead to errors during the analysis of genomic data. The final step of the process represented in Figure 1 implies the use of a set of algorithms for variant detection and/or evaluation in terms of pathogenicity. In

order to understand the reliability of these methods, different assessment methodologies have been developed. In the following paragraphs the different steps of the workflow proposed in Figure 1 will be described. In the first section: “NGS: decoding biological information”, the basis of Next Generation Sequences techniques will be presented. The following paragraphs will be more focused on the analysis of NGS data, first the properties of the human genomes and the resources available to analyze genetic data will be exposed. Then one paragraph will be dedicated to show the mathematical basis of predictive models able to assign a particular phenotype to an individual given the corresponding genotype. Different prediction methods of this type have been developed for CAGI challenges, as described in the subparagraph “Complex phenotype predictions and CAGI experiments”. The following chapters will be focused on *in silico* models, their use to simulate biological process and predict the response of the system to a perturbation. In particular one paragraph will expose the development of an *in silico* model for human cholesterol metabolism, able to predict blood lipids levels modifications due to genetic mutations. Then a paragraph will contain a description of precision medicine and real case applications, in which disease treatment has been based on patient sequencing data. The last paragraph will present an example of data analysis techniques applied to a biological problem.

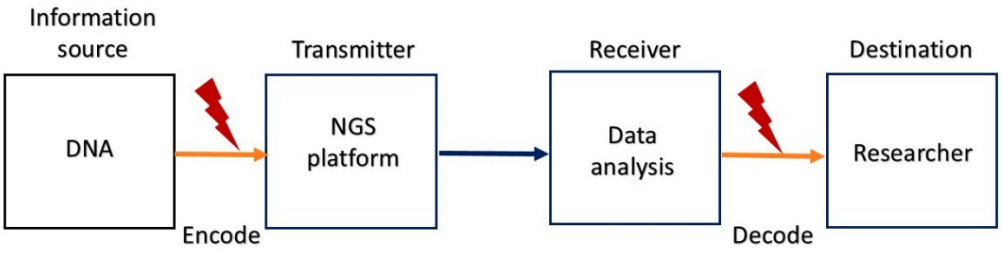


Figure 1: Schematic diagram of a general communication system adapted to NGS data analysis. Bolts represent processes affected by noise.

NGS: decoding biological information

Next generation sequencing technologies have arisen in the mid-2000s, from that time the cost of sequencing one human genome has decreased from millions of dollars to

around US$1000 (Wetterstrand). Basically a sample genome is sequenced producing a set of reads, a sequence of bases from a single molecule of DNA, with variable lengths, depending on the platform used, from 35-700 for short-read to several kilobases for long read approaches (Goodwin et al., 2016). The produced reads are then aligned to the reference genome with specific software, like the Burrows Wheeler Aligner tool (Li and Durbin, 2009). The presence of variants (mismatches) is assessed with tools like the Genome Analysis toolkit, GATK (McKenna et al., 2010), a variant calling file (vcf) will store all detected variants (Figure 2). Read size can directly affect the quality of genome analysis, in particular short reads are not able to span whole complex or repetitive regions of the genome, generating ambiguity in terms of position and size of genomic elements, while long reads are able to solve that problem but are more expensive. Different platforms have limitations in sequencing specific regions of the genomes (AT-rich, GC-regions, homopolimeric regions), for instance related reads are under-represented or present errors. The main consequence is a number of false positive calls for structural variations (SNP or CNVs), that could be a source of bias in following analysis. The coverage of a specific region could be a way to decrease the error, especially with long reads that have an error of 15% for read (Goodwin et al., 2016). In principle it's possible to sequence the whole genome (WGS) or only the exome, in this case we talk about whole exome sequencing. The aim of this approach is to sequence only protein-coding exons, avoiding the less valuable part of the genome, approximately the 98%, composed of repeat, intergenic and non-protein-coding sequences (Hodges et al., 2007). An alternative to whole exome sequencing could be target sequencing, in particular a gene panel developed in order to amplify only specific target regions that are involved in the onset of a disease of interest, for example cancer (Buys et al., 2017). One of the major issues affecting NGS data analysis is the definition of the haplotype, or allelic sequence along the same chromosomes (He et al., 2018). This process is called haplotype phasing and it can hardly affect phenotype prediction from genotype data (Tewhey et al., 2011). One example is blood group identification. In this case the process of haplotype phasing is crucial to predict the blood group antigens coded by each of the two copy of the ABO gene (Giollo et al., 2015). Different methods have been developed to perform haplotype phasing from sequencing reads (He et al., 2018). In a recent study (Koren et al., 2018),

the haplotype of an individual has been reconstructed using NGS data, obtained by the sequencing of the two parents with short reads, used to guide the haplotype phasing of the offspring, sequenced with long reads.

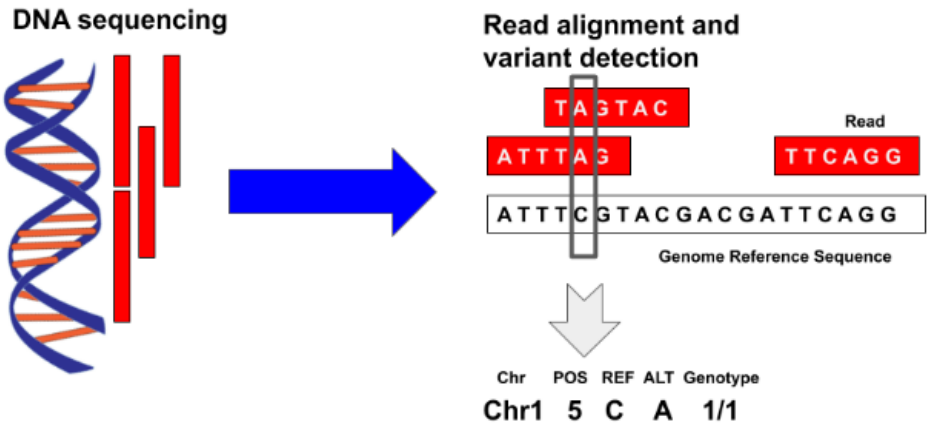


Figure 2. Genome sequencing and variant detection pipeline. The sequencing process starts from the synthesis of a set of reads, each one is a reproduction of a sequence of n DNA bases. Each read is then aligned to the genome reference sequence. Different algorithms will detect the presence of a variant by checking the match between the aligned reads and the genome reference sequence. This information can be stored inside a variant calling file (vcf).

The human genome as a source of variation and interpretation

The human genome project had shed light on the structural complexity of the human genome: of around 3 Giga base pairs, only 2% of it is involved in protein synthesis, representing around 20000 genes (Moraes and Góes, 2016). If we consider one individual sequenced genome, we expect around 3 million nucleotide substitutions compared to the reference genome, a few of them possibly disease related (Niroula and Vihinen, 2016). Among them are single nucleotide polymorphisms (SNPs): genetic events in which different nucleotides can be mapped, with different frequencies in the population, on the same position. More complex variants are insertions or deletions of multiple nucleotides

(Strachan et al., 2011). Different public databases and tools are available for the evaluation of the pathogenicity of genetic variants (Figure 3). One example could be dbSNP, with millions of reported genetic variants, of which a consistent amount were collected from the analysis of the genomes sequenced in the 1000 Genomes project (1000 Genomes Project Consortium et al., 2012). Other databases are more focused on collecting information on phenotypes related to genetic variants. In particular Clinvar collects variants observed in individuals or families, in clinical or research works and reports the clinical significance related to disease, clinical features and mode of inheritance (Landrum et al., 2016). SNPs associated to complex disorders, according to published GWAS are hosted by the GWAS Catalog (MacArthur et al., 2017). In order to evaluate the pathogenicity of variants, different tools have been developed for the prediction of the impact of amino acid changes on protein activity, they are based on three main types of features: Energy functions, Conservation scores and Hybrid methods. Energy functions based methods evaluate the effect of a variant on protein stability, in particular they compute the free energy change (ΔΔG), defined as the difference between the free energy in the normal and variant protein (Niroula and Vihinen, 2016). Conservation scores are instead based on evolutionary principles: given a protein of interest (query), a set of protein sequences with different degrees of similarity is aligned to the query. From the multiple sequence alignment, it is possible to check if the amino acid substitution was present at a conserved position and classify it as deleterious, one example could be Provean (Choi et al., 2012). Hybrid methods are based on different features, as evidence from experimental tests, clinical features and scores from predictors. However most predictors estimate the pathogenicity of a variant, but don't specify the effect on gene or protein function. Different tools have been developed to address specific topics related to this question, in particular: protein stability, localization, splice-site effect, protein disorder regions, protein aggregation (Niroula and Vihinen, 2016). Despite all the resources for variant classification, the analysis of the human genome remains a complex operation that could lead to False Positive results. In 2008 Nature published a work (Wheeler et al., 2008) based on the analysis of James D. Watson genome: the authors compared found non-synonymous known SNPs to the Human Gene Mutation Database (Stenson et al., 2003). They discovered that the sequenced genome

had two homozygous variants previously reported as causative of diseases generally developed at birth or early childhood, while the subject was not affected by any of them (Wheeler et al., 2008). Errors like the one encountered during the analysis of James D. Watson genome and several more could hamper the utilization and interpretation of sequencing information. In principle it's possible to define multiple sources of error: mutations associated to a disease according to clinical or wet lab evidences but wrongly annotated in a database, technical or measurement error rate related to the sequencing procedure, the problem of multiple testing on millions of variants. Most of all, also the penetrance of a variant could drive false positive calls: a variant reported as pathogenic in literature, could activate the disease only in subjects with a particular genetic background and exposed to the environmental effect present only in the patients of the original work (Kohane et al., 2012). In order to reduce the number of false positive calls in the analysis of genome the American College of Medical Genetics and Genomics and the Association for Molecular Pathology published a set of guidelines for the interpretation of genome variants (Richards et al., 2015). This work has presented a precise nomenclature to be used when calling (HGVS nomenclature) and describing the effect of a variant (e.g. pathogenic, likely pathogenic, etc.), a set of computational tools, database and precise criteria for the interpretation of sequence variants. In a recent work (Niroula and Vihinen, 2019), state of the art methods for variant effect prediction were tested on a dataset of benign variants, extracted from the Exome Aggregation Consortium (ExAC) database (Lek et al., 2016). The specificity of the assessed tools ranged from 0.95 to 0.64, with different performances related to variant frequencies on different populations. This kind of information could be useful to understand the reliability of a prediction tool in the interpretation of benign variants (Niroula and Vihinen, 2019).

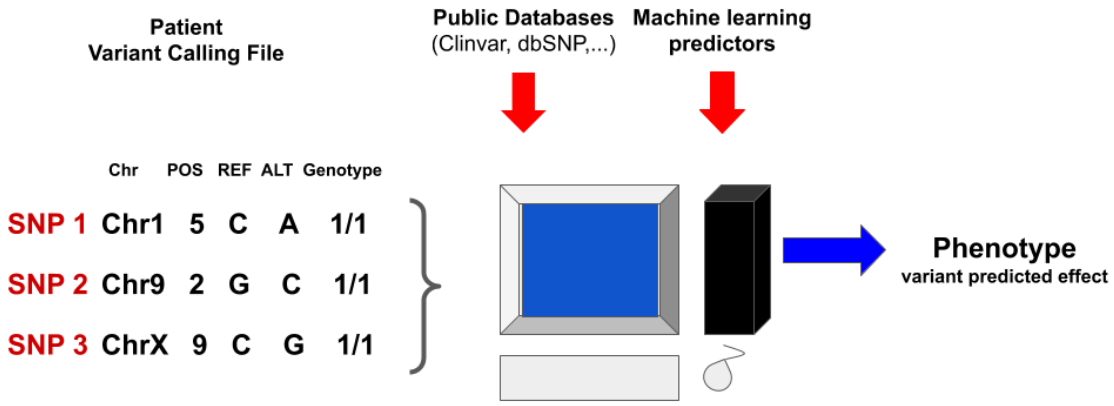


Figure 3. Variant effect assessment. The pathogenicity of variants present in an individual vcf is assessed with different methods. Online databases contains detailed information, based on literature or clinical cases, about the effect of specific variants. Different machine learning tools, freely available online, are able to predict if a specific variant could be pathogenic, or in other words if a specific protein or gene function is impaired.

Methods for phenotype prediction from genomic data

The human genome contains several structural modifications, from single base variants like SNPs to insertions, deletions and Copy Number Variants events, that can cross multiple bases. In principle this variability can be used to predict the probability of developing a complex disease in a subject (Abraham and Inouye, 2015). While in some cases the presence or absence of a single genotype determines the expression of a character (defined Mendelian), in most cases multiple genes and the environment are involved in the expression of a human character (Strachan et al., 2011). One example of a single mutation causing the development of a disease, in a Mendelian manner, could be Sickle cell anemia, caused by a mutation on the hemoglobin subunit beta gene (OMIM MIM Number: 603903 : 02/05/2019). The opposite is true for complex disease like Crohn's Disease (CD) where multiple loci are involved in the susceptibility to the disease (Barrett et al., 2008). When a human trait is affected by a large number of genes, each one with small-effect and interacting with the others, it can be described by a quantitative genetic

model. In principle we can describe a complex phenotype (y_i) as the sum of a genetic signal (g_i, the genetic value) and the model residual (ε, the remaining sources of variation). A practical example could be a model on all genotyped markers (for example SNPs) of an individual, which is able to compute:

$$y_i = g(x_i, \Theta) + \varepsilon \quad (1)$$

$g(x_i, \Theta)$ is a function mapping from markers ($x_i = (x_{i1}, \ldots, x_{ip})'$) onto genetic values, Θ is a vector of parameters that have to be estimated from the data. In practice one has to estimate Θ on a training sample of individuals with phenotype y and then test the model on a blind set of patients (Figure 4). A simple genetic model based on regression could be:

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \quad (2)$$

where, $x_{ij} \in \{0,1,2\}$, is a SNP, and β_j is the additive effect of marker j (allele). The use of this kind of models with human data presents two main challenges, the first is related to the heritability of a trait, which means that not the whole phenotypic variance can be associated to the genotype. The second point is related to the low Linkage Disequilibrium between markers in human populations compared to agricultural species for which whole genome prediction methods have been developed: while this models are evaluated in within-breed predictions in animal science, the capability of these methods to predict genetic value in distantly related individuals is not clear (de los Campos et al., 2010). Another interesting issue of genome-wide predictors is that the number of parameters that have to be fitted (one for each marker) greatly exceeds the number of samples of the dataset, so it is crucial to choose the appropriate training procedure to avoid overfitting (Gianola, 2013).

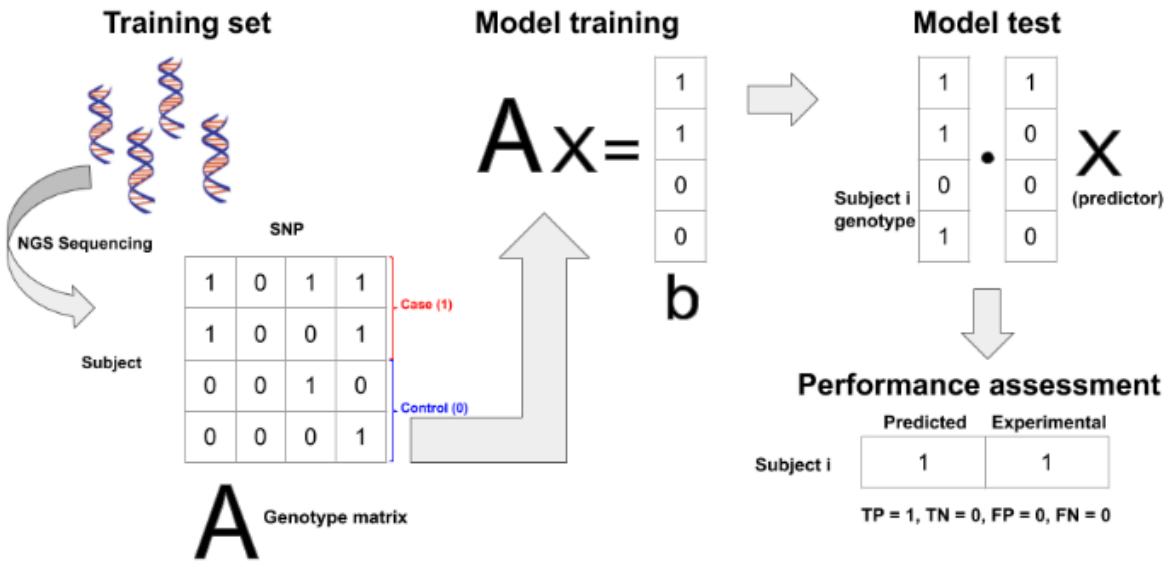


Figure 4. The development of a machine learning tool for genotype data. The process starts with the sequencing of a training set of patients, either case or controls. The genotype is then stored inside a matrix (A) with one row for each subject of the set and one column for any genotyped SNP. During the training phase the model will be trained to correctly classify patients (b), given the genotype (A). The result of the training procedure is a predictor x, which is a vector, in this example with all zeros except the first element, that is the only SNP that is always present in case but not in controls. Now the performance of the trained predictor is assessed on a test set. In this example the class of each element is computed by the dot product between the genotype of a subject and x. Finally the output of the predictor is compared to the real (experimental) class and a set of performance indices are computed (for example True Positive, True Negative, False Positive, False Negative). In this case the developed method has correctly classified the only element of the test set as case. A detailed description of assessment measures is present in chapter 1.

Complex phenotypes prediction and CAGI experiments

The Critical Assessment of Genome Interpretation (CAGI) is an experiment in which the state of the art of variant effect prediction methods is assessed on a dataset, for which

the experimental phenotype is known (Hoskins et al., 2017). Different challenges were part of past CAGI editions: we can classify them in two types of experiments. In the first type participants have to predict, with *in silico* methods, the effect of variants in terms of protein function or a related phenotype, one example could be the p16 challenge (Carraro et al., 2017). In the other type of experiments participants have to predict the phenotype of a patient, from the related vcf file, presented as the probability of a patient to be case of one or multiple disease classes (Chandonia et al.; Daneshjou et al., 2017a).

Different methodologies for complex disease or phenotype prediction on exome sequence data, reported as vcf files, have been tested in three challenges purposed by the fourth edition of CAGI in 2016 (Daneshjou et al., 2017a). Crohn's disease and bipolar disorder challenge were focused on predicting the probability of having the disease, a discrete phenotype based on pathological or clinical diagnosis. While the Warfarin dosing challenge required the prediction of a continuous phenotype: the therapeutic Warfarin dose. In Crohn's disease challenge participants had the possibility to train their method with German ancestry exome data from previous challenges, a total of 93 cases and 29 controls, with relatives among cases. The trained methods had to predict the probability of a patient to be a case and eventually the age of onset of the disease on a blind dataset of 111 unrelated German descent genomes (64 cases, 47 controls). In a first phase the pathogenicity of exome variants has been assessed through population frequency or published prediction tools, like SNP&GO (Calabrese et al., 2009). This data was then used to train different machine learning algorithms like neural networks, logistic regression, naïve Bayes and random forests. The assessment was performed by computing the Area Under the Curve (AUC) for different methods, with a maximum equal to 0.72. The top scoring groups developed their methods taking into account previously published works that related genes and genomic regions to Crohn's Disease. In the bipolar disorder challenge the test and training set had the same size, 500 exomes of unrelated bipolar disorder cases and matched controls. Different groups used methods similar to the ones developed for Crohn's disease challenge, performing with an AUC that in most of cases was below 0.55. The exception was a method with an AUC of 0.64, in this case the prediction algorithm was based on a Neural Network developed by a group with little background in genetics and approaching the problem as a data classification

challenge. In the Warfarin dosing challenge participants had to predict the optimal dose of anticoagulant for a patient, given the exome. The training set was composed of 50 exomes of African Americans, with a prescription corresponding to the tails of Warfarin dose distribution. The developed methods, tested on a blind set of 53 individuals, had a maximum R^2 of 0.35, previously developed algorithms, trained on European populations had a max performance of 0.25 R^2.

In general developing a predictor for this kind of challenge is cumbersome, the main reasons are related to the structure of the challenge's set, which generally comes from research studies, hence it couldn't represent all human populations, since more extreme phenotypes are abundant, for the sake of statistical significance (Daneshjou et al., 2017a).

Guidelines for the assessment of bioinformatics methods

The abundance of NGS data has driven the development of tools that use this information to predict the effect of genetic variations on phenotype. The reliability of these methods should be evaluated with a rigorous assessment, able to show weak and strong points of each algorithm. A gold standard assessment procedure would require to test a set of methods on a benchmark dataset, excluded from the training procedure (e.g., blind set). This blind set should be public, experimentally validated and non-redundant, in particular it should be balanced, with an adequate number of positive and negative results (Vihinen, 2012).

Methods involved in the assessment should be evaluated with measures related to different features of predictions compared to real values: raw percentages or metrics based on a contingency matrix (e.g. sensitivity, see Table 1), distance (e.g. Euclidean distance) and correlation (e.g. Matthews' Correlation Coefficient) (Baldi et al., 2000). Performance measures reliability is influenced by test set class unbalance, in some cases their outcome could be biased as for scores like Positive and Negative predictive values (PPV, NPV), while MCC is not affected by the number of positive and negative examples in the set (Boughorbel et al., 2017; Vihinen, 2012). Different works have performed a rigorous assessment on published methods regarding specific classification or regression

problems. In principle it's possible to fix three steps for an assessment procedure: the first one requires the calculation of a set of performance indexes on all methods, followed by a ranking procedure. The second step is to evaluate if the differences between predictor performances are statistically significant. Finally a bootstrap could be used to estimate if the scores of a given prediction could have been obtained by chance (Figure 5), this step is particularly suggested, in literature, when blind set size is limited (Carraro et al., 2017). Two published examples will be used to illustrate the proposed pipeline for assessment: the CAGI p16 challenge (Carraro et al., 2017) and the assessment of long protein intrinsic disorder (Necci et al., 2017).

In this work the authors have collected a set of published methods, predicting protein disorder from sequence data, and tested them on a curated blind set of disordered proteins, the Disprot dataset (Piovesan et al., 2017a). The authors have assessed the performance of the selected tools at two levels: per residue and protein. MCC and AUC were used in the first level, while on a per protein basis, Root Mean Square Error (RMSE) and Pearson Correlation Coefficient (PCC) has been computed on predicted and observed disorder content normalized by the number of annotated residues. Finally the different predictor has been ranked on the basis of the mean score rank obtained on all indices. A Welch's t-test has been used to determine if the difference between the performance of predictors was statistically significant or not (Necci et al., 2017). A similar workflow has been used to evaluate models involved in the CAGI p16 challenge, but a final bootstrap step has been added in order to evaluate if the performance of the best methods was better than random, considering the limited size of the dataset (10 variants). This procedure is based on a distribution of 10000 random scores, which is used to compute the probability of obtaining an index better than the real one by chance, given a significance threshold of 0.05 (Carraro et al., 2017).

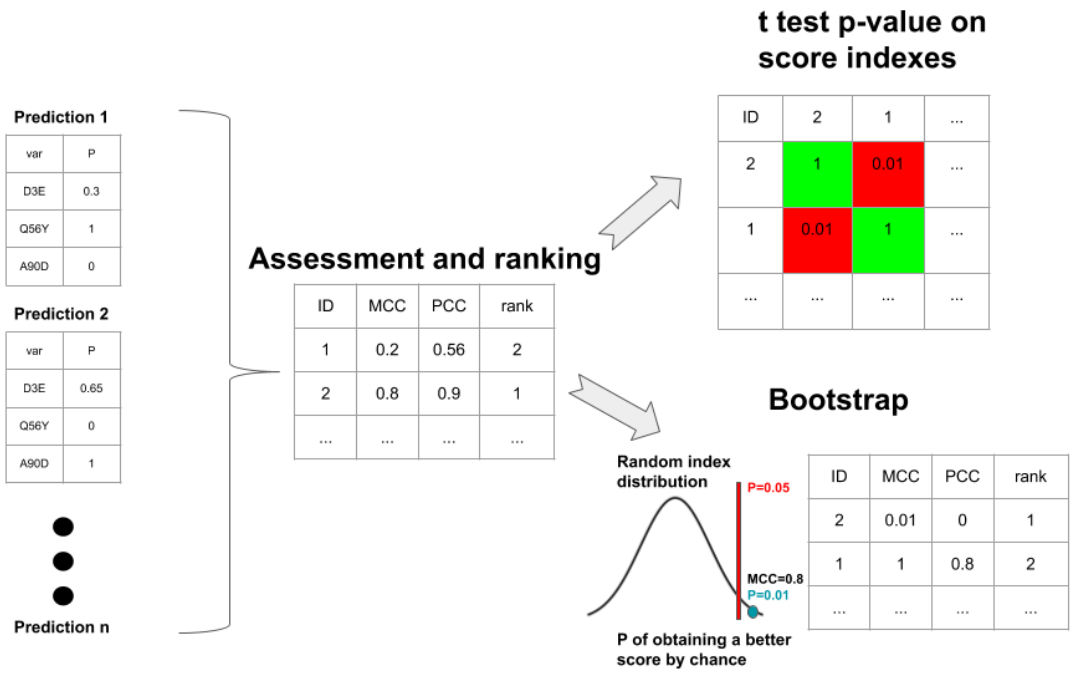


Figure 5. A standard predictors assessment procedure. The first step focuses on methods' assessment and ranking, followed by a t-test on the computed indices, to check if the difference between predictors' performance is statistically significant. Finally a bootstrap procedure will produce a distribution of random scores, in order to compute the probability (P) of getting by chance an index greater or equal to the real one.

Index	Formula
Accuracy (ACC):	$ACC = \frac{TP + TN}{TP + TN + FP + FN}$ (3)
Balanced Accuracy (BACC):	$BACC = \frac{1}{2}(TPR + TNR)$ (4)
Matthew Correlation Coefficient (MCC):	$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$ (5)
True Positive Rate (TPR):	$TPR = \frac{TP}{TP + FN}$ (6)
True Negative Rate (TNR):	$TNR = \frac{TN}{TN + FP}$ (7)
Positive Predictive Value (PPV):	$PPV = \frac{TP}{TP + FP}$ (8)
F_1 score (F_1):	$F_1 = 2\frac{TPR \times PPV}{TPR + PPV}$ (9)
False Positive Rate (FPR):	$FPR = \frac{FP}{TN + FP}$ (10)
Area Under the Curve (AUC):	AUC of TPR and FPR computed at different thresholds

Table 1. Set of performance indexes used in this work.

Models in systems biology

Biological systems are regulated by a huge set of metabolic reactions involved in the synthesis (anabolism) or degradation (catabolism) of complex molecules. Systems biology describes with mathematical models the biological system of interest, with

different degrees of complexity: from thousands of reactions and metabolites considered by genome-wide models to only few reactions considered by toy models (Cazzaniga et al., 2014). A mathematical model describes the interactions and dynamics of a biological system with a set of equations, representing the interaction between biological entities and their development in time. Once model's parameters have been computed on the basis of experimental data, this tool can be used to test different hypotheses in silico and predict the behavior of biological systems in different conditions (Potter and Tobin, 2007). Models are based on mass conservation of chemical reactions: for each metabolite involved in the process we can compute the derivative of the concentration in time as the difference between the amount of metabolite produced and removed.

$$\frac{\Delta x}{dt} = V_{produced} - V_{transported} - V_{consumed} \quad (11)$$

If we know the relation between metabolites, we can generate an mxn stoichiometric matrix S, with m metabolites and n different reactions in the network. It's possible to define a steady state in which the concentration of metabolites doesn't change in time.

$$Sv = 0 \quad (12)$$

We define v as the vector of n fluxes of metabolites in the system, computed from the equation above once the stoichiometric matrix is defined (Helms, 2008).

Once the mathematical model of a biological system is defined, it should be validated (Figure 6), this procedure requires that the computed model is able to reproduce the same behavior of the biological system under the same conditions (Cobelli et al., 2012). One way could be to perform a sensitivity analysis on the developed model. With this method each parameter of the system is reduced, by multiplying its value to a set of parameters ([0, 1]) and the response of the system is computed. This analysis will show if small changes in the input drives major modifications in the output (if there is sensitivity) and if the behavior of the in vivo system is reproduced (decrease or increase of a metabolite) (Cazzaniga et al., 2014). In general models should be optimized by reducing the difference between predicted and experimental values, while keeping low the number of model parameters (parsimony) to avoid overfitting (White et al., 2016). The Akaike's information Criteria index has been developed to evaluate both factors: it computes the reduction of the error between predicted and experimental values (RSS: residual sum of

squares) but it penalizes an increase in the number of model parameters (d) (Marcuzzi, 2011).

$$RSS = \frac{\sum_{k=1}^{N}(y(k) - y(\hat{k}))^2}{N} \quad (13)$$

$$AIC = log\left(RSS \cdot \left(1 + \frac{2d}{N}\right)\right) \quad (14)$$

Different strategies have been developed for parameter estimation from experimental data. Basically a parameter estimation algorithm is used to produce a set of model parameters that will be used to perform the prediction. Then the difference between experimental and predicted values is computed, in order to evaluate if the error has been reduced by the new set of parameters. This iteration will continue until the best fit between model output and experimental values is found (Cobelli et al., 2012). Different global optimization methods have been developed to efficiently perform the research of a parameter set that minimizes the error, by computing a set of derivatives. Basically in each iteration the algorithm computes the gradient of a scalar function, based on the error between predicted and experimental data: parameters are modified towards the steepest descent. One example is the Levenberg-Marquardt algorithm, used for parameter estimation in physiological models (Cobelli et al., 2012; Marcuzzi, 2011). In the following paragraph the development of an in silico model will be presented through an example taken from literature (van de Pas et al., 2012).

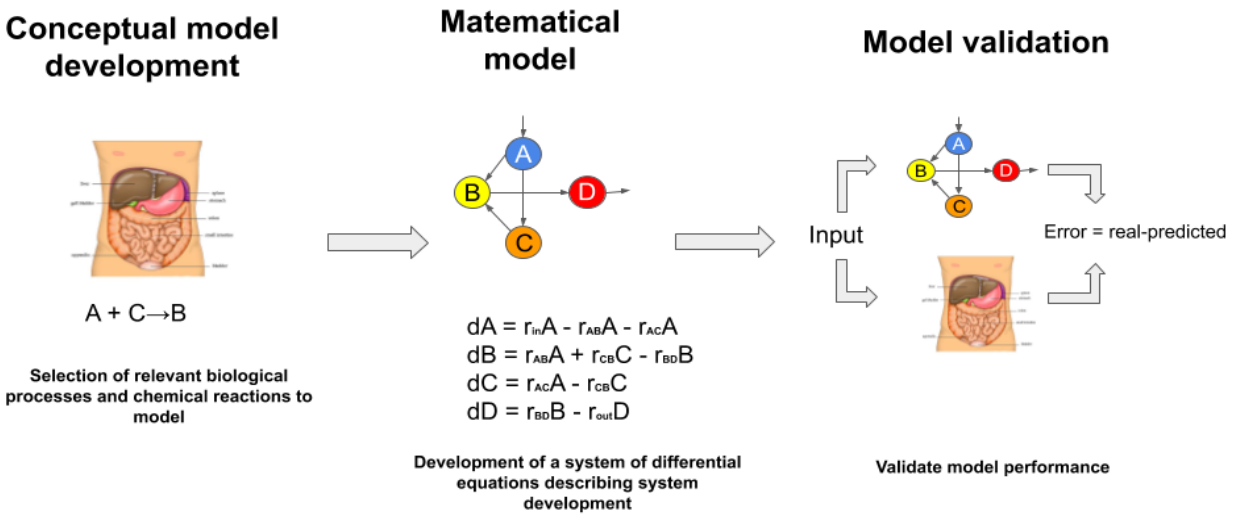


Figure 6. The development of an *in silico* model of a biological process. First step is the development of a conceptual model that includes the most relevant reactions of a biological system. The second step is to add mathematical formalism to the developed model, that is converted into a system of differential equations. System parameters are estimated from literature or custom wet lab experiments. The final step is model validation, a comparison between model predictions and real biological system response to the same input.

An *in silico* model for prediction of lipid profiles based on genetic predisposition

From 2010 Niek C. A. van de Pas and coauthors have published a set of works (van de Pas et al., 2010, 2011, 2012) in which they reported the development and validation of an *in silico* model able to simulate cholesterol metabolism in a human body and predict the effect of genetic mutations on blood cholesterol levels. The first task was to develop a conceptual model, that is to define the most relevant genes that regulate cholesterol metabolism and the corresponding biochemical reactions that should be included in the model. This objective was fulfilled after an extensive research of knockout mice for cholesterol genes in literature. 120 genes were evaluated, 36 were considered as key genes for their effect in the regulation of plasma cholesterol concentration. Finally only 12 genes, directly related to metabolic or transport process, were incorporated into the model. The conceptual model was composed of four compartments (or pools), representing organs involved in cholesterol homeostasis: Intestine, Liver, Plasma (HDL, LDL) and periphery. All compartments were connected by arrows, representing genes involved in the production, removal or movement of cholesterol in the system (van de Pas et al., 2010).

The following task was to add a formal mathematical structure to the conceptual model of cholesterol metabolism developed for mouse. In particular they defined the order of reaction to describe cholesterol flux in a stable way, this was done by selecting a number of submodels able to predict the shift in cholesterol concentration observed in knockout mouse strains. After this phase model parameters were computed as solutions of the steady state, with fluxes based on literature data. The model was later validated by simulating the effect of five knockout mutations and comparing predicted levels of Total cholesterol, HDL and LDL cholesterol to the ones deriving from experimental data (van de Pas et al., 2011). The final task was to convert the *in silico* kinetic model developed for mouse to one able to simulate cholesterol metabolism in man. This required a calibration of the parameters to human data and to add a reaction for Cholesterol ester

transfer protein, an enzyme that is absent in mouse. The model was validated on a dataset of ten gene mutations: the predicted effect by the model, computed by reducing the rate of cholesterol of the affected gene, was compared to the experimental value. The average deviations between model predictions and experimental data (computed as the cholesterol level in case divided by control) were 36% for TC, 49% for HDL-C and 43% for LDL-C. This was considered successful compared to the models available at that time (van de Pas et al., 2012).

Personalized medicine

The term personalized medicine was defined by the National Cancer Institute of the USA, this defines a healthcare approach in which diagnosis and disease treatment is based on individual information: the genome, proteins and environment (Tremblay and Hamet, 2013). This new approach to healthcare is based on biological data retrieval and analysis. While the sequencing of a single human genome is relatively cheap, around 1000$ USD, the storage and following analysis are less easy to afford. For instance a basic framework of personalized medicine, could start from the omics profile of the subject, collecting information on: genomics, transcriptomics, proteomics, epigenomics, metagenomics, metabolomics and nutriomics. The following step would require the analysis and integration of this huge amount of data with bioinformatics approaches. The result would be a personalized clinical care based on patient disease networks, obtained from the analysis of omics data (Alyass et al., 2015). Genotype data is currently driving precision medicine and pharmacogenomics approach in different diseases. In particular cancer treatment can be adapted to variants detected on specific genes of the patient. For instance a personalized therapy could be modeled on the basis of specific SNPs, related to drug adverse effects or weaknesses of cancer against selected treatments (Low et al., 2018). Genetic testing for the presence of variants inside genes related to breast cancer development could influence patient tumor prevention or treatment. In particular patients with variants on specific genes like BRCA1 or BRCA2, can decide to undergo breast mastectomy to prevent the possible development of breast cancer (Welsh et al., 2017). Despite all previous considerations, a clinical test should be proven to be useful: given its cost there should be no better alternatives at the same price or cheaper ones. Usually the

ACCE scheme is used to evaluate the validity of a test. First of all the ability of a method to measure a selected quantity is assessed (Analytical validity) (Strachan et al., 2011), as SNP genotyping (i.e. the detection of the presence or absence of one or more SNPs). The analytical validity of a test based on sequencing data could be limited, especially if variants are located in complex regions of the genome, where many Next Generation sequencers have limited sequencing performance (Goodwin et al., 2016). Therefore the same sample could present different genotypes, according to the technologies involved in the sequencing procedure. Another relevant point considered by the ACCE framework is the ability of the assessed method to predict the health outcome of a subject (Clinical validity). This feature is cumbersome in the case of genetic tests for the prediction of complex diseases development, while some rare variants are strong predictors of the onset of a disease (Giollo et al., 2017), other variants could increase the risk of a subject only if they are present together. As multiple genes or loci are involved and interact in a nonlinear way, it is not clear their effect on patient risk to develop a disease, moreover other clinical data should be taken into account to have a reliable estimate (Strachan et al., 2011). Another important point that should be taken into account, in the evaluation of a clinical test, regards the possible uses of a test (Clinical utility). Some genetic tests could accurately estimate the risk of one individual of developing a genetic disease (e.g. Huntington disease) but in case of positive results no therapeutic strategy will be able to block or postpone the onset of the predicted malady. Eventually the ACCE framework also considers the ethical issues related to the clinical test under analysis (Ethical aspects) (Strachan et al., 2011). This point is complex and still matter of debate in the field of personalized medicine, since genomic information could be misused (Garver and Garver, 1994). Few years ago, a private company offered a service of DNA sequencing, variant detection and related effect description. Hence, customers could be informed about the presence of risk alleles, for specific diseases or variants influencing drug absorption, in their genome. Since this kind of information could lead to drastic medical measures and potentially harm the subject, the U.S. Food and Drug administration had fixed limitations to the use of this test (Check Haiden, 2017). The possibility to perform massive sequencing of genome data offers the possibility to detect harmful variants in the population. This kind of analysis could be useful for heterozygous carriers of mutations

on genes associated to genetic diseases like Cystic Fibrosis, since this disease could be transmitted to the progeny by healthy carriers (Farrell et al., 2017). This kind of test presents ethical and technical issues. From a strictly practical point of view, there should be a fixed threshold between sensitivity and specificity of our test. For instance we could consider a limited amount of variants as pathogenic and therefore have an increase in false negative results, or do the opposite and increase the sensitivity, that would result in an increment of false positives (Strachan et al., 2011). Wrong predictions could negatively influence clinical practice or individual choices, as the decision to stop the pregnancy of a possibly affected child. The massive use of genome information for prenatal screening have raised several ethical concerns, as it could be potentially used for eugenics (Garver and Garver, 1994) or to discriminate carriers of harmful mutations (Strachan et al., 2011). Despite all ethical and technical issues that are limiting the spread of personalized medicine, this new approach is becoming more commonly used in different fields, as disease risk prediction or cancer treatment, in private companies as well as clinical practice (Check Haiden, 2017; Welsh et al., 2017).

Data analysis and bioinformatics

Bioinformaticians have to manage massive amounts of data deriving from wet lab experiments and extract useful information to understand the biological problem they are investigating. One example could be an association study on a population of affected patients and related controls. The objective of this study is to find genetic variants or loci associated to a disease. Given a dataset of patients, their genetic information is stored in a matrix of N individuals with M genotyped SNPs, and their phenotype represented by a binary vector b, representing elements' class (either control or case, figure 4). Before performing any kind of analysis, a rigorous data cleaning procedure must be performed. The aims of this step are to remove missing data and have a reduced but complete dataset (Zhao and Cen, 2014), hence to avoid publishing incorrect results (Sebastiani et al., 2011). Therefore we have to exclude from the analysis features (SNPs) whose state is unknown in a large number of individuals (e.g. 20%) and remove samples (individuals) that have more than 20% of SNPs not genotyped (i.e. we don't know if that individual has that SNP or not). Others criteria applied for filtering datasets are relatedness among

samples and population stratification effects (Marees et al., 2018). Population stratification (as relatedness) could impair the result of an association study (Purcell et al., 2007): variants shared among a group of individuals of the same ethnic origin, could be wrongly associated to a disease (i.e. false positive associations) (Marees et al., 2018). In data mining, outlier detection is a procedure in which elements that could deflect the results are excluded from the analysis (Zhao and Cen, 2014). In the case of a dataset of genotyped individuals, elements with high or low heterozygosity rates or with wrong sex labels (i.e. X chromosome homozygosity estimate is not compatible with the sex assigned to the sample) are excluded as possibly related to sample mix-ups or contaminations (Marees et al., 2018). At this point data could be visualized with standard data-reduction techniques, as Multidimensional scaling (MDS). In this case the resulting plot will show if clusters of individuals are present, as well as outliers, or samples that are distant from the cluster of elements of the same ethnic origin (Purcell et al., 2007). Once data has been cleaned by any outlier, it is possible to compute the association between disease and genotype with a standard test like X^2 (Marees et al., 2018). Since this test is repeated for each SNP considered in the analysis on all individuals, a lot of false positive errors are expected (i.e. p-value multiplied by the number of independent tests or SNPs), therefore a Bonferroni correction could be adopted to avoid an increase in type I errors (i.e. the original p-value divided by the number of independent tests) (Strachan et al., 2011). Eventually we will consider as associated to the disease all SNPs with a statistical test (e.g. X^2) p-value lower than the Bonferroni corrected threshold.

Thesis outline

The manuscript is organized in 4 chapters. In the first chapter I will present my research work in the context of machine learning methods evaluation. In particular, basic guidelines for the assessment of methods for phenotype prediction, from genotype data, will be exposed. Issues as data visualization, outlier removal and machine learning methods evaluation will be treated. In particular I have strongly contributed to both the literature research work and development of an R package for the evaluation of CAGI challenges. The aim of this tool was to create a “common framework” for the assessment of different challenges, to support comparison between different predictors and assure reproducibility

of the CAGI experiments. In the following two chapters I will describe my work in the assessment of two CAGI challenges, one based on complex phenotype predictions, the other on the classification of variants affecting one single protein. The first one is based on “Assessment of patient clinical descriptions and pathogenic variants from gene panel sequences in the CAGI-5 intellectual disability challenge. *Human Mutation* (2019). doi:10.1002/humu.23823”. I had contributed to the assessment of the ID challenge, that is based on the principles under the development of the package for the assessment of CAGI challenges. In particular I have curated the numerical assessment of the presented methods. The third chapter is based on “Performance of computational methods for the evaluation of Pericentriolar Material 1 missense variants in CAGI-5. *Human Mutation* (2019). doi:10.1002/humu.23856”. In this section the assessment of the CAGI-5 PCM1 challenge will be presented. I was involved in the development of the algorithm and workflows that were the backbone of this paper. This assessment is a different application of the work presented in the first chapter, applied to an interesting case: a challenge that was a mixture of a regression and a classification one. In fact predictors had to estimate p-values and the class of each blind set variant. Moreover, a bootstrap procedure was added to evaluate the statistical significance of the computed assessment measures. The last chapter is based on a revised version of “In silico prediction of blood cholesterol levels from genotype data. *Biorxiv* (2019). doi:10.1101/503003”, now under revision in a peer reviewed journal. This work presents the development and assessment of a previously published in silico mathematical model for cholesterol level prediction (van de Pas et al., 2012). I have developed this mathematical model in R language and optimized the performance by adding a training phase. These newly implemented features allowed cholesterol levels prediction from genotype data, while improving the performance of this tool, as I have assessed using the original test set (van de Pas et al., 2012) and the principles exposed in the first chapter of this thesis.

Inside the Conclusions I summarized the main findings and issues presented in the previous chapters. I described possible applications and limitations of previously exposed bioinformatics methods in a research or clinical context

Chapter 1

R-tool for CAGI challenges assessment

In this chapter, I will describe the assessment of different CAGI experiments through a set of guidelines and best practices I have developed after an extensive research work. CAGI experiments were organized to evaluate the state of the art of different methods, estimating the effect of genomic changes on phenotype. The performance was assessed on blind datasets: only data providers and CAGI organizers knew the experimental phenotype. The evaluation of these methods was originally based on a set of performance measures derived from literature or created ad hoc for the challenge. The lack of specific guidelines for the assessment impaired performance comparison between predictors developed for different challenges. In this chapter I will present the R package I have developed for CAGI challenges assessment. This algorithm is able to assist the process of predictors' evaluation on the basis of workflows and principles derived from literature and previous CAGI experiments. I have used the developed algorithm to repeat the assessment of four previously published CAGI datasets and evaluated machine learning methods with the proposed workflow.

This chapter is based on the manuscript "R-Tool for CAGI challenges assessment", authors are: Francesco Reggiani, Emidio Capriotti, Marco Carraro, Alexander Monzon, Carlo Ferrari, Silvio C. E. Tosatto.

Introduction

Large-scale sequencing experiments have allowed to characterize a huge amount of genomic variations that may have functional impact and be potentially associated to human disease (Abraham and Inouye, 2015; Capriotti et al., 2012, 2018). Despite this information is becoming more relevant in the medical practice, no standard procedures are available for matching the novel variants with clinical data. In this context, the Critical Assessment of Genome Interpretation (CAGI) was established to estimate the quality of

the methods for predicting the impact of genomic changes at molecular and phenotype levels. CAGI is a community-driven experiment in which about 10 challenges are proposed in each edition. Groups around the world are invited to submit predictions before a deadline. Then independent assessors evaluate prediction performances and show the results in a final meeting (Hoskins et al., 2017). Two main types of challenges have been considered by CAGI so far: single variant analysis and whole-genome, exome or gene panel analysis. In the first type of challenge, predictors have to evaluate the effect of single-base variants, like single amino-acids modifications on protein function (Carraro et al., 2017; Zhang et al., 2017). The other type of challenge is based on the prediction of individual phenotype and eventually the associated variant, from whole genome, exome (Giollo et al., 2017), or gene panel genotype data (Chandonia et al., 2017). Each experiment is composed of two phases, the former starts with the release of blind test data and is followed by the corresponding predictions, produced by different groups, and generally based on different methods. During the latter phase the quality of the predictions is assessed on related experimental or clinical data (Hoskins et al., 2017). The lack of stringent guidelines for the assessment, is reflected by the different metrics used so far for the evaluation of predictors performance, derived from common indices used in literature (Baldi et al., 2000), or created ad hoc for the challenge (Chandonia et al., 2017). Assessment measures of predictors performance usually evaluate different features of the involved computational methods, in particular: raw percentages or metrics based on a contingency matrix (e.g. sensitivity), the distance (e.g. Euclidean distance) and the correlation (e.g. Matthews' Correlation Coefficient) (Baldi et al., 2000). The absence of specific guidelines for the assessment, in terms of performance measures or statistical methods employed in the analysis, makes comparison between different challenges problematic as repeating the experiment. The use of a subset of evaluation parameters could lead to problems in understanding strong points and limitations of a method, while some assessment measures, like Positive and Negative predictive values (PPV, NPV), could be biased by unbalanced datasets (Vihinen, 2012). One example is the blind test of Hopkins clinical panel challenge, where the different disease classes were not equally represented among patients and healthy controls were absent (Chandonia et al., 2017). In this work we propose the R-tool "CAGI.ASSESSOR" which have been developed on

the basis of previous CAGI challenges to perform regression and multiple phenotype challenge assessments. This tool is based on a framework for visualization, assessment, statistical evaluation of multiple predictors' performance and outlier detection.

Materials and Methods

Challenge evaluation strategies

In general all CAGI challenges proposed in the past can be classified in two main groups indicated as regression and classification experiments. The first group includes all the challenges requiring the prediction of a numerical value resulting from an experimental measure. The second group consists of challenges that classify the effect of a genetic variant or assign patients to a specific phenotype. Examples of regression challenges organized in the previous editions of the CAGI are the p16INK4a (Carraro et al., 2017) and Human SUMO ligase (Yin et al., 2017) which asked to predict the change of cell proliferation rate and competitive growth upon nonsynonymous variant, respectively. While examples of classification experiments are the Crohn (Daneshjou et al., 2017a) and the Hopkins (Chandonia et al., 2017) challenges that asked to assign a phenotype to a patient starting a list of personal variants in the whole exome or in a panel of genes. In this work we propose a methodology for the assessment of both types of CAGI experiment discussing the specific cases of the p16INK4a (Carraro et al., 2017) and Hopkins (Chandonia et al., 2017) challenges. Dataset are available here: https://genomeinterpretation.org/content/cagi-4-2016

p16INK4a challenge dataset

The p16INK4a challenge was focused on the prediction of the effect of a set variants, affecting the tumor suppressor p16, on cell proliferation. The dataset was composed of 10 nucleotide variants, influencing exclusively p16 coding region, and coding for single amino acid substitutions. The effect of each variant was previously validated in distinct cell proliferation rate assays and scaled between 0.5 (wild type variants) and 1 (pathogenic mutants). Participants had to predict the effect of each single variant on cellular proliferation, indicating the percentage of proliferation rate relative to pathogenic

mutants (with values bounded between 50% and 100%) and eventually adding a standard deviation value. Moreover, a set of 19 variants, from a previous publication (Kannengiesser et al., 2009; Miller et al., 2011), was available for the training phase of participants' predictors.

Crohn's disease challenge dataset

The Crohn's disease challenge (Giollo et al., 2017), presented in CAGI 4, was a binary class challenge, in which each sample belongs either to control or case (disease) class. Participants had to predict the probability of a patient to have the disease and specify the level of confidence of their prediction in terms of standard deviation. For each individual predictors had to inform if the age of onset of Crohn's disease was before age of 10. The test set was composed of 111 German ancestry exomes, 64 cases and 47 controls. For each patient the exome was sequenced using the Truseq exome enrichment kit (Illumina) and the Illumina Hiseq2000. Produced reads were mapped on the human genome build hg19, variants were called using the Genome Analysis Toolkit (GATK version 3.3-0) Haplotype Caller. Only high quality variants were retained. The methods developed for the Crohn's Disease CAGI 4 challenge could be trained on test sets published in 2 previous challenges. The first was Crohn's disease challenge dataset of CAGI 2 with 56 exomes of german ancestry, 42 cases and 14 controls. The other was the test set of the Crohn's disease challenge of the third edition of CAGI, with 66 exomes of German ancestry,where 51 disease cases and 15 controls. In this dataset some individuals were related, since cases were collected from families with multiple occurrences of Crohn's disease (Daneshjou et al., 2017a).

Human SUMO ligase (UBE2I) challenge dataset

The SUMO conjugase challenge was a regression challenge of the fourth edition of CAGI. Participants had to predict the fitness effect of missense mutations on the human SUMO-conjugating protein (SUMO E2 ligase coded by the UBE2I gene) (Zhang et al., 2017). SUMO, small ubiquitin related modifiers, is a reversible post-translational protein modifier. In particular sumoylation can modify protein interactions of its target, that results in altered localization. The SUMO E2 ligase catalyzes covalent attachment of SUMO to a range of

target proteins (Geiss-Friedlander and Melchior, 2007). A competitive yeast complementation growth assay was used to quantify the effect of each mutation represented in the challenge. The test set was split in three datasets: the first was composed of 219 single missense mutations and was considered reliable. The second datasets was composed of 463 single missense mutations but less reliable than the first one. The last dataset comprehended a set of 4427 elements with two or more mutations per clone (Yin et al., 2017). For each mutation present in a dataset, submitters had to specify the growth score of each variant and the confidence in their prediction (standard deviation). The effect of each mutation was defined according to growth scores: values lower or equal to 0.3 were considered as deleterious, between 0.3 and 0.7 as intermediate, from 0.7 to 1.3 as wild type and greater than 1.3 for advantageous (Zhang et al., 2017).

Hopkins challenge dataset

The Hopkins clinical panel challenge is a multiple class challenge, each sample belongs to a single class, in which submitters had to predict the probability of each patient to be affected by a disease and the causative mutation/s. Participants had to predict the correct class out from a set of 14 possible diseases, in each one of the 106 patients. For each patient, exon sequence data of a set of 83 genes, sequenced by a Illumina Next Generation Sequencing (NGS) platform, was provided in a single Variant Call Format (VCF) file, containing single nucleotide variants (SNV) and insertion-deletion (InDels) as reported by GATK (GATK UnifiedGenotyper, GATK HaplotypeCaller, v2.7-4, Broad Institute, Cambridge, MA). The submitted disease class was the one with the maximum predicted probability value for each patient (predictions with all equal disease class probabilities were not taken into account) , and only the associated variants were taken into account. Not all disease classes were equally represented in the dataset, in particular 56 patients had “Diffuse lung disease”, while no patient was affected by 5 of the possible disease classes. For only 43 patients Hopkins noted at least one variant related to subject’s disease class, only for these patients matches between noted and predicted causative variants were evaluated.

Assessment of the regression challenge

The evaluation of a regression experiment requires to estimate the ability of the tools to predict a numeric value generally derived from an experimental measure. Thus, for the assessment of the regression challenge we calculated scores including different correlation and error measures. In detail, the regression performance are scored using three correlation coefficients namely Pearson (r_P), Spearman (r_S) and Kendall's Tau (τ), indices based on a confusion matrix or TPR, TNR, BACC, MCC, AUC and two error measures: the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE).
In order to evaluate TPR, TNR, BACC, MCC and AUC of the different methods involved in the challenge, predicted and experimental values were converted in binary data: all values greater than a threshold of 0.75 were converted to one. After the conversion half of the variants in the blind set were marked as positive cases, while the remaining as negatives. According to the selected measures we ranked the methods from the highest to the lowest correlation coefficients (r_P, r_S and τ) and binary classifiers evaluation indices (TPR, TNR, BACC, MCC, AUC), while from the lowest to the highest for the error measures (RMSE and MAE). Each predictor was classified according to the median above all the column wise ranks of the computed indices. The statistical significance of differences between predictors' performance was assessed with a paired student t-test on all scoring indices, using negative RMSE and MAE (the half of the p values are plotted as a heatmap, Figure 7. The presence of outliers, among predicted values, has been assessed by a scatterplot of the error between submitted and experimental proliferation rates (Figure 8). This analysis is useful to understand if the error in the predicted proliferation rates was homogeneous among all variants. Or if in some cases most of submitted values were outside the experimental standard deviation.

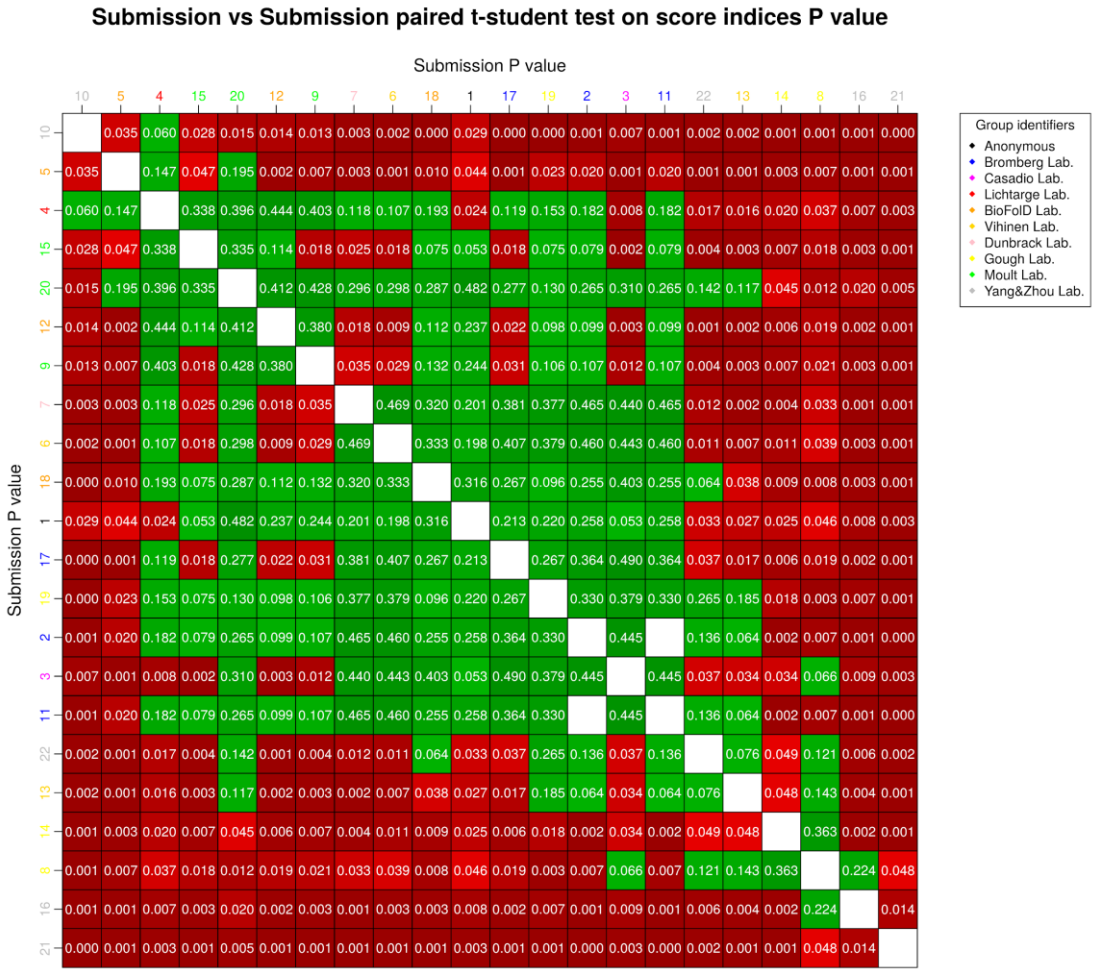


Figure 7. Pairwise difference between submissions for p16 challenge. Paired t-test statistical differences between submissions based on the mean score obtained by each submission over all indices, sorted according to the final ranking. Green squares are indices of tied predictions (P values >= 0.05) meaning that according to the performance indices used, the difference between two predictors is not statistically significant. White squares represents tied predictions with P equal to 1.

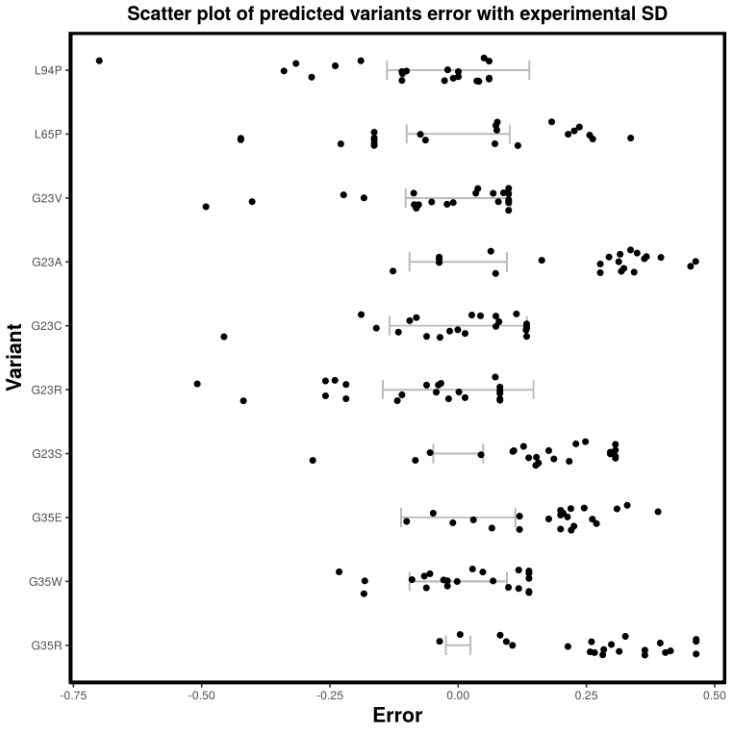


Figure 8. Outlier plot with experimental error bar for p16 challenge. Scatterplot of the errors of all submissions on each variant, computed as the difference between predicted and experimental value. Grey bars represent the experimental error.

Regression measure of performance

Predictors' performance has been evaluated according to three different kinds of metrics: the correlation between predicted and experimental proliferation rates, the error and measures based on a contingency matrix.

The correlation has been assessed on ranks with Spearman's correlation coefficient (SCC or r_s) and on the continuous measures with Pearson correlation coefficient (PCC or r) as follows:

$$r = \frac{n\sum_{i=1}^{n} y_i\bar{y}_i - (\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} \bar{y}_i)}{\sqrt{\left[n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2\right]\left[n\sum_{i=1}^{n} \bar{y}_i^2 - (\sum_{i=1}^{n} \bar{y}_i)^2\right]}} \quad (15)$$

The Kendall's Tau correlation coefficient (KCC or τ) has been used to evaluate the conservation of the order of magnitude between proliferation rates of the experimental set in predicted data. This index was calculated as follows:

$$\tau = \frac{2}{n(n-1)} n \sum_{i<j} sgn(x_i - x_j) sgn(y_i - y_j) \quad (16)$$

The deviation between real and predicted proliferation rates has been evaluated with two different metrics: the Root Mean Square Error (RMSE) and the Mean Absolute Error as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}{n}} \quad (17)$$

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \bar{y}_i|}{n} \quad (18)$$

The first error measure penalize large deviation from real data, while the second computes the mean of the absolute error. The performance of regression methods presented in the p16INK4a challenge has been evaluated using metrics for binary classification algorithm assessment measures. Predictor performance was evaluated using the following metrics: true positive and negative rates (TPR, TNR), balanced accuracy (BACC) and Matthew's correlation coefficient (MCC).

We also calculated the area under the receiver operating characteristic (ROC) curve (AUC), by plotting the True Positive Rate as a function of the False Positive Rate at different probability thresholds.

Assessment of the classification challenge

The evaluation of the classification challenge estimated the ability of the different methods to correctly predict the class and the variants associated to the disease. In a first phase each submission probability matrix, with number of rows equal to the number of patients and the number of columns equal to the number of patient classes, was transformed in a sparse matrix, by converting the values equal to the maximum value of probability of each row to one and the remaining to zero. At this point different metrics, based on a contingency matrix, have been used to score the predictions: in particular BACC, MCC, AUC, TNR, TPR. In order to evaluate correct variant predictions, Jaccard index and FPCV were computed. In a first phase these indices were computed on single phenotype columns, in order to assess predictions' quality on each phenotype, then each score was

computed above all phenotypes through a weighted sum. Predictors were classified according to the rank of the overall median, computed on column wise rank over all indices (computed above all phenotypes, shown in Table 2 and suppl. Table S1, Appendix 1). The statistical significance of the difference between the performance of the different submissions has been assessed with a paired Student's t-test on related scoring indices (the half of the p values are shown in Figure 9 and Suppl. Figure S1).

An outlier analysis has been carried on the elements of the blind set, for each patient the amount of total correct predictions divided by the total amount of submissions was computed, the results have been reported with a bar plot ordered by phenotype (Figure 10, Suppl. Figure S2). The whole assessment has been performed considering the whole test set or only the patients that have at least one variant associated to the disease.

Prediction	BACC	MCC	AUC	TNR	TPR	FCPV	Jaccard	Overall Rank
61.1	0.79	0.594	0.83	0.96	0.628	0.570	0.540	1
59.1	0.75	0.544	0.75	0.98	0.535	0.430	0.394	2
59.2	0.74	0.532	0.73	0.98	0.512	0.430	0.405	3
58.2	0.68	0.404	0.69	0.97	0.395	0.360	0.349	4
58.1	0.66	0.380	0.66	0.97	0.349	0.302	0.295	5
60.1	0.58	0.216	0.60	0.97	0.209	0.178	0.147	6
60.2	0.59	0.236	0.60	0.97	0.209	0.178	0.147	6
57.2	0.54	0.130	0.63	0.95	0.116	0.047	0.016	8
57.1	0.49	-0.012	0.50	0.92	0.116	0.047	0.012	9
57.3	0.50	-0.024	0.46	0.93	0.047	0.000	0.000	10
57.4	0.48	-0.040	0.51	0.94	0.023	0.000	0.000	10

Table 2. Performance indices over all phenotypes of Hopkins clinical panel challenge for those patients with at least one causative variant. Seven performance scores and the median overall rank are shown. Predictions are sorted by the rank of the median among all indices.

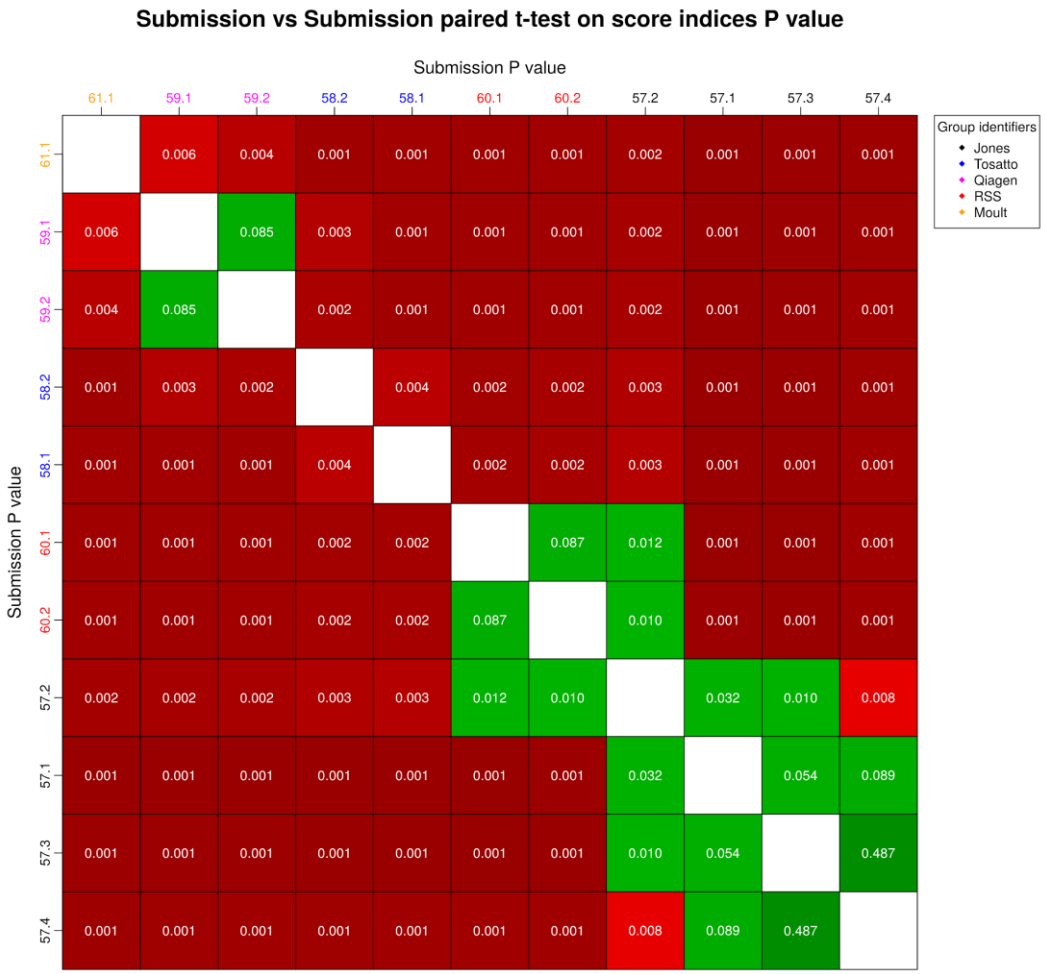


Figure 9. Pairwise difference between submissions (patients with one variant associated to the disease) for Hopkins challenge. Statistical differences between submissions based on the mean score obtained by each submission over all indices, sorted according to the final ranking. Green squares are indices of tied predictions (P values >= 0.01) meaning that according to the performance indices used, the difference between two predictors is not statistically significant. White squares represents tied predictions with P equal to 1.

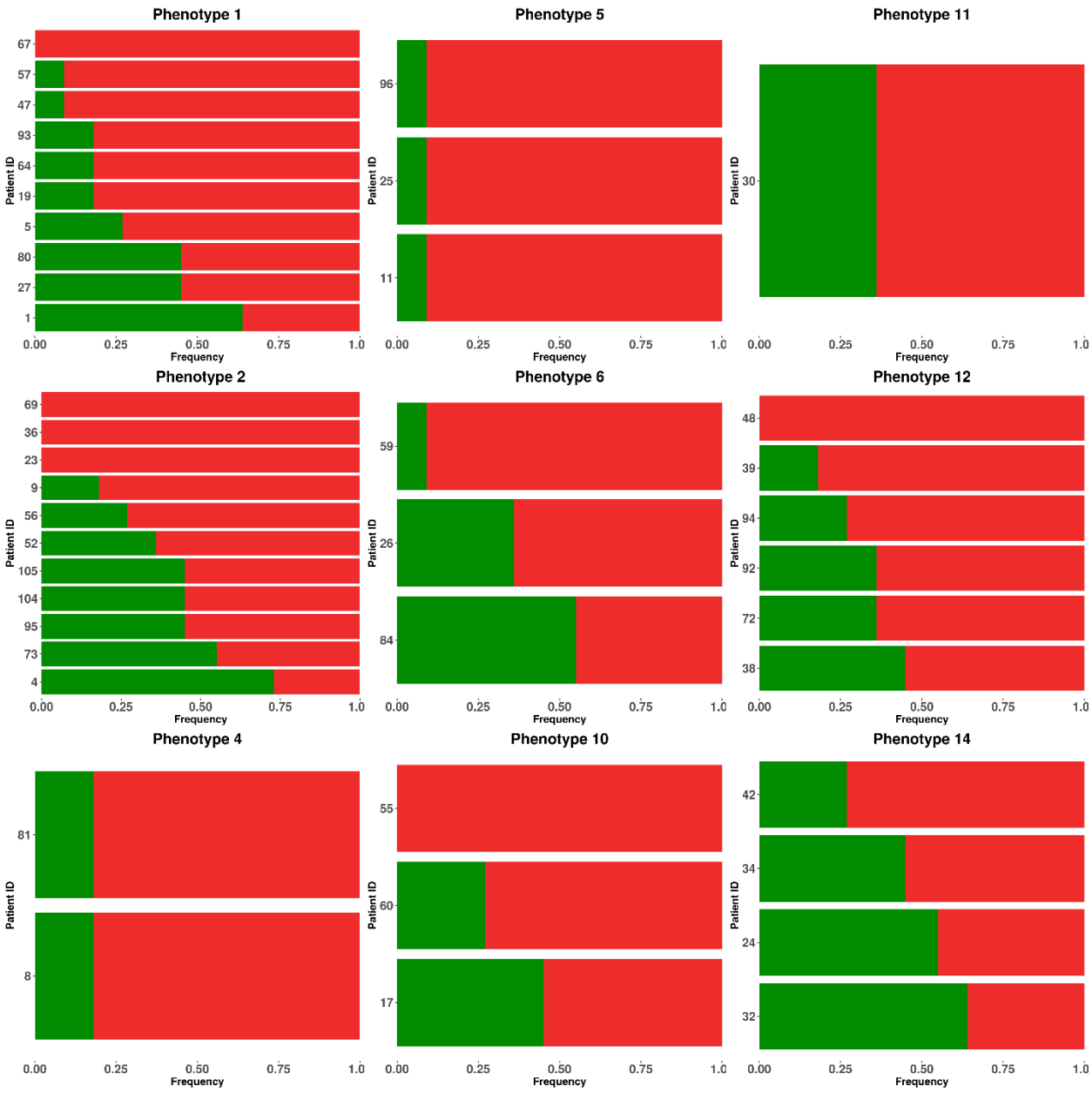


Figure 10. **Outliers barplot for patients with noted variants in Hopkins challenge.** For each patient, the frequency of correct predictions [0, 1] by all the methods is reported in green, while red represents frequency of wrong predictions. Patients are ordered by phenotype and frequency of correct predictions. Red bars are patients for which no method was able to correctly detect the phenotype.

Classification measures of performance

Predictions of multiple phenotype challenges were evaluated on class and variant prediction. Assessment of correct class prediction was computed on single phenotype by using BACC, MCC, AUC, TNR, TPR. The sum of each index multiplied by the weight of each class (the number of patients with that phenotype divided by all patients) was used to compute the total value of that assessment score on all phenotypes as follows:

$$W_{Index} = \sum_{i=0}^{n} Index_i W_i \quad (19)$$

A special case was TNR, in this case the weight has been computed as the rate between the total number of patients minus the number of patients in class i and the number of different classes minus one, multiplied by the total number of elements in the test set, as follows:

$$TNR_{all} = \sum_{i=0}^{n} TNR_i \frac{(n_p - n_{p_i})}{(n_d - 1)n_p} \quad (20)$$

Variants predicted to be associated to the correct class of disease were evaluated with two specific metrics:

FCPV (fraction of correct predicted variants) Calculated as the number of correct variants in a correctly classified patient divided by all submitted variants (for that patient), as follows:

$$FCPV = \frac{\sum_{i=1}^{N} \frac{\left|V_{exp_i} \cap V_{pred_i}\right|}{\left|V_{pred_i}\right|}}{N} \quad (21)$$

The Jaccard index or the number of correct variants in a correctly classified patient divided by all variants (predicted and experimental variants for that patient), calculated as following:

$$Jaccard\ Index = \frac{\sum_{i=1}^{N} \frac{\left|V_{exp_i} \cap V_{pred_i}\right|}{\left|V_{exp_i} \cup V_{pred_i}\right|}}{N} \quad (22)$$

These indices were computed on each correctly classified patient of the dataset and divided by the total number of patients in the test set with at least one variant associated

to the disease. The total value of these indices among all phenotypes, was computed as the sum of each index multiplied by the relative abundance of that class, considering only patients with at least one variant associated to the disease.

Results

Package description

The developed R-tool “CAGI.ASSESSOR” was implemented with the aim to facilitate and standardize the evaluation process of CAGI challenges. The algorithm is able to perform the assessment of regression and classification challenges in a straightforward way. The functionality of the package could be divided in three main blocks: submission format and quality control, data visualization and performance evaluation and ranking.

This first step of the analysis is the quality control on each submission file. Errors in prediction file format, like incorrect row or column order and value format, can affect the appropriate submission assessment. The algorithm performs a quality check, specifying error type and position in order to help the assessor to check submitted files.

After loading all group submissions and experimental value data, the package allows the assessor to produce different plots to visualize and analyze group submissions. The correlation of predicted probability values between submissions could be visualize in a heatmap. For this end the package computes the KCC between all possible pairs of submissions and produce the plot. This analysis allows the assessor to easily identify groups prediction similarity in terms of the relative order between elements inside each probability vector. For regression challenges, the relationship between predicted probabilities and related experimental values is shown by a set of two dimensional scatterplots, one for each prediction. The x axis represents experimental values and the y axis predicted values. Prediction quality is represented by the distance of points from the main diagonal. Moreover, for multiple phenotype challenges the algorithm computes test set composition and general performance by group or by patient, producing pie chart, barplots and tables as present in the Hopkins Challenge assessment paper (Chandonia et al., 2017).

Finally, the package calculates the performance measures as was explained in Material and Methods section. The selected set of measures were chosen according to the acquired experience of past CAGI challenge assessments (Carraro et al., 2017; Chandonia et al., 2017). The output of this analysis is a table presenting the performance measures and overall rank of all submissions, ordered according to overall rank (the median of the ranks of all indices of each row) (Table 2). Another table is also produced with the rank number of each measure instead of the original value. Then the algorithm produces a heatmap which shows correlation between submission scoring indices. At this point users could decide to repeat the analysis excluding highly correlated indices, as reported in the assessment of the p16INK4a challenge (Carraro et al., 2017). Once the performance table has been computed, the algorithm produces a heatmap of the paired t-test p-values between predictions scores (Figure 7). According to the assessment table, the first three predictors of the regression challenge are presented as ROC curves in a plot. Multiple phenotype challenge assessments can be performed using the whole test set or only patients with at least one variant associated to the disease.

Other feature present in this package is the outlier analysis. The algorithm produces outlier plots, giving the possibility to evaluate predictors performance at the level of single elements of the test set. For regression analysis a scatterplot of the difference between predicted and experimental value for each variant of the test set is produced (Figure 8). Behind these points, representing different predictions of the same element, a bar representing experimental SD is reported. In the case of multiple phenotype analysis the outlier plot is composed by a set of single phenotype barplots, representing the frequency of correct positive predictions in each patient.

Practical cases of CAGI regression challenges

CAGI-3 p16 challenge

The package was used to reproduce the assessment of p16 challenge, previously reported by Carraro et al. (Carraro et al., 2017). Submission 10 shows high correlation between predicted and real data: the small difference between RMSE and MAE is explained by the absence of variants that have a predicted proliferation rate far from the

experimental one (Table 3). Balanced accuracy and MCC of submission 10 outperform the score obtained by the other methods. Submission 5 and 4 (second and third in the ranking, respectively) and other submissions have obtained higher TPR values compared to submission 10. In particular, submission 4 had a perfect AUC. However if we look at TPR and TNR values, we can see that submission 4 is classifying all variants as TP at a threshold of 0.75. Indeed, it is confirmed by the null MCC value. Submission 4 also obtained a very low TNR (0.4) compared to the best predictor (1.00). The other submissions have a range of correlation values that goes from positive to negative values. Sometimes prediction errors (MAE and RMSE) of some groups are in the range of the top ranked submissions, like submission 12 and 18. However if we take into account the MCC, TPR and TNR we can see that in most of cases methods are biased to predict most or all variants as positive or negative. Finally we can say that submission 10 outperformed the other methods in terms of correlation and error measures and showed a pretty good balance between correct negative and positive predictions as shown by high BACC, MCC, TNR and TPR values. This situation is unique of submission 10, since in the other cases at least one of this indices is lower than 0.5. A statistical significance test over the distribution of performance indices, showed that submission 10 is different from all the other predictors, except 4 (Figure 7). The difference between submission 5 and 4 was not statistically significant, as the difference between these predictions and others following the third rank. The outlier plot shows that in most cases the error distribution of variants L65P, G23A, G23S, G35E and G35R is outside the experimental standard deviation, showing that in general submissions predicted a value of proliferation rate far from the experimental one. Variant G23C has the largest number of predictions within the experimental standard deviation (Figure 8).

Submission	PCC	SCC	KCC	RMSE	MAE	BCC 0.75	MCC 0.75	AUC 0.75	TNR 0.75	TPR 0.75	Overall Rank
10	0.83	0.867	0.689	0.092	0.08	0.90	0.82	0.92	1.00	0.80	1
5	0.66	0.806	0.600	0.158	0.11	0.70	0.50	0.88	0.40	1.00	2
4	0.84	0.815	0.629	0.165	0.12	0.50	0.00	1.00	0.00	1.00	3
15	0.76	0.693	0.506	0.188	0.16	0.60	0.33	0.96	0.20	1.00	4
20	0.76	0.693	0.506	0.177	0.15	0.50	0.00	0.96	1.00	0.00	4
12	0.57	0.673	0.467	0.159	0.12	0.60	0.33	0.84	0.20	1.00	6
9	0.70	0.627	0.378	0.202	0.17	0.60	0.33	0.88	0.20	1.00	7
7	0.22	0.297	0.200	0.182	0.15	0.60	0.33	0.68	0.20	1.00	8
6	0.23	0.401	0.339	0.257	0.21	0.60	0.33	0.58	0.20	1.00	9
18	0.46	0.374	0.276	0.164	0.14	0.60	0.20	0.72	0.60	0.60	9
1	0.83	0.675	0.454	0.235	0.20	0.50	0.00	1.00	0.00	1.00	11
17	0.43	0.315	0.249	0.218	0.18	0.60	0.22	0.72	0.40	0.80	12
19	0.30	0.123	0.070	0.203	0.17	0.60	0.22	0.76	0.80	0.40	13
2	0.33	0.061	0.023	0.213	0.17	0.70	0.41	0.62	0.60	0.80	14
3	0.53	0.588	0.470	0.255	0.22	0.50	0.00	0.70	0.00	1.00	14
11	0.33	0.061	0.023	0.213	0.17	0.70	0.41	0.62	0.60	0.80	14
22	0.15	0.244	0.205	0.185	0.16	0.50	0.00	0.40	0.00	1.00	17
13	0.11	0.130	0.047	0.201	0.15	0.50	0.00	0.42	0.00	1.00	18
14	-0.22	-0.480	-0.402	0.233	0.18	0.50	0.00	0.42	0.20	0.80	19
8	-0.34	-0.480	-0.402	0.392	0.33	0.50	0.00	0.42	1.00	0.00	20
16	-0.45	-0.692	-0.564	0.225	0.19	0.40	-0.33	0.08	0.00	0.80	21
21	-0.62	-0.806	-0.600	0.237	0.21	0.20	-0.65	0.12	0.00	0.40	22

Table 3. Performance indices of p16 Challenge. Results for 10 performance scores with median rank. Predictions are sorted by the rank of the median among all indices. MCC values equal to infinite (∞) are reported as 0.

CAGI-4 SUMO ligase challenge

The previously reported assessment of Human SUMO ligase challenge (Zhang et al., 2017) was also performed, only for the subset 1, and the results are shown in Suppl. Table S2. Considering the overall ranking, the best group was 43.1. They obtained the best TNR and AUC values and showed slightly better values of RMSE and MAE compared to the other groups. We have also to highlight the performance of submissions 47.1 and 47.2 which ranked second in the overall ranking and obtained the best values of BACC and MCC (submission 47.2). These three top ranked submissions don't present significant differences among their performance measures (Suppl. Figure S3). Indeed, submission 43.1 doesn't show significant differences with submissions 46.1, 44.3, 39.2, 40.1 and 44.4, according to the t-test (Suppl. Figure S3).

Despite the good AUC and BCC values obtained by some groups, the predictions have high values of TNR and TPR. Observing the low MCC values reached by most predictors, we can conclude that they poorly performed in this challenge.

Practical cases of CAGI classification challenge

Hopkins gene panel assessment

The assessment results based on patients where Hopkins noted variants are shown in (Table 2). It shows that most predictors had in general a good specificity, while sensitivity rapidly decrease from the best submissions towards the worst ones. This behavior is expected since the aim of the challenge was to predict one disease class out of 14 for each patient. All groups were able to predict at least one variant associated to the disease. In particular group 61, 59, 58 obtained Jaccard index and FCPV scores significantly greater that the ones obtained by group 60 and 57. Group 61 obtained the best values of Balanced accuracy, MCC and AUC, followed by group 59 and 58.

The pairwise paired student's t-test over performance indices shows that group 61 prediction is different from all the other submissions, while for 59, 60 and 57 predictions from same groups are not significantly different (Figure 9). Additionally, the amount of patients with the phenotype correctly predicted was assessed (Figure 10). No patient was affected by phenotypes 3, 7, 8, 9 and 13, as previously reported (Chandonia et al., 2017). For the phenotypes 4, 5, 6, 11 and 14 all patients were correctly classified by at least one method. In the other cases each phenotype had at least one patient who was not detected by any predictor.

The assessment was repeated on the whole dataset to address the order of groups in terms of performance rank (Suppl. Table S1). In this challenge, generally the groups obtained a worse performance when all patients were taken into account. In particular the top ranking predictor 61.1 has an MCC of 0.29, almost the half of the one computed on the subset of patients for which Hopkins noted a causative variant (MCC = 0.59). The same behaviour is observable in the other groups, except group 57. Specificity and sensitivity computed on the whole dataset showed that in general predictors were able to discriminate Negative results, like in the dataset considering only patients with variants, but their ability to detect true positive results decreased. Paired student t-test among performance indices showed that the difference among the first groups (61, 59) was not statistically significant, the same is true for group 60 and 57 (Suppl. Figure S1). Outliers analysis showed that only for three phenotypes predictors were able to select the correct disease class for each patient (4, 6, 11) (Suppl. Figure S2), while an increase of not correctly predicted patients is observable in the remaining phenotypes, compared to the dataset of patients with at least one causative variant (Figure 10).

CAGI-4 Crohn's Disease challenge

The assessment of the methods of the Crohn's Disease challenge was based on classification challenge metrics. The results are shown in Supp. Table S3. Several methods obtained an AUC greater than 0.6, with a maximum of 0.72 of the first rank prediction(10.1). In particular submission 10.1 has also the highest value of MCC (0.36), BACC (0.68) and similar values of TPR (0.77) and TNR (0.59), computed with a threshold

of 0.5. Following the rank of the different methods, on all classification indices, the value of MCC is above 0.2 for submissions from five groups (10, 1, 3, 2, 6), with AUC values near or over 0.6. In most of following submissions there was a relevant difference between TPR and TNR rate of the same methods, showing that predictions were more skewed towards positive or negative classification. Despite the different range of scoring indices obtained by the predictors, the student t-test computed on the performance measures showed that the difference between most submissions was not statistically significant (Suppl. Figure S4).

Discussion

In this work we presented a framework for the assessment of CAGI challenges. CAGI challenges have been developed to evaluate the current state of art in machine learning prediction algorithms applied to biological problems. In principle method developers could test their methods on a set they have collected and report an arbitrary set of performance parameters. An alternative is a systematic analysis in which the performance is evaluated on an accepted benchmark dataset with suitable metrics (Vihinen, 2012). In this work we evaluated predictors performance of four previously published CAGI challenges on their original dataset with a set of indices used in the assessment of machine learning methods, except FCPV. With the presented methodology strong and weak points of each predictor are exposed in an extensive and clear way, considering both the ability of a predictor to detect positive examples and discriminate them from negative ones in an unbiased way. The assessment of two previous CAGI challenges has shown that the analysis of predictors in terms of correlation, distance and contingency matrix metrics are essential to have an estimation of the reliability of a method. In particular in p16 challenge (Table 3) different predictors had an high level of correlation and low error, but with null MCC, TPR or TNR, showing when methods were basically a perfect rejecter or acceptor.

The developed workflow, has evaluated the methods of two classification challenges presented in CAGI 4 (Chandonia et al., 2017; Giollo et al., 2017). The first was Crohn's disease challenge, a single phenotype challenge presented in CAGI 4. The previous published assessment was based only on AUC (Daneshjou et al., 2017a). The assessment performed with the proposed workflow has enriched the knowledge related

to the properties of methods involved in the challenge with more indices. In particular MCC has been able to evaluate the performance of the methods without being affected by an unbalanced test set, while TPR and TNR has shown predictors ability to detect patients affected by Crohn's Disease without misclassifying healthy controls. In this work we addressed the problem of assessment in multiple phenotype challenges with a new perspective: methods have been evaluated at the level of single class prediction and generalized, through a weighted sum of the results obtained on each phenotype present in the blind set. This procedure has shown predictor performances on the basis of patient phenotype: a novel approach to understand model strengths or weaknesses on definite group of elements of the test set, while preserving the ability to evaluate methods on the whole dataset and rank them. New interesting metrics have been added for the evaluation of predicted disease causing variants, in particular we considered predictor's capability to select the correct disease causing variants while avoiding False Positive calls. The assessment has been performed considering only patients with variants known to be associated to the disease, or the whole dataset. The overall assessment have shown that predictors performance was higher in the first case. A possible explanation is that model predictions were driven by the detection of causative mutations related to the phenotype observed in the patient. At the end of the assessment procedure we evaluated the presence of outliers among the elements of the blind set, in both challenges. Among the different variants of p16 challenge dataset, it has been shown that in some cases predictors were able to predict the proliferation rate with an error inside the experimental one, while in other cases most of predictors have huge errors (see for instance Figure 8, variant G23V and G23S). For the Hopkins' Challenge dataset we conducted an outlier analysis and evaluated the number of correct predictions for each element of the dataset. This way it was possible to understand which phenotype and patients were correctly predicted by most of the submissions and which ones were problematic (Figure 10 and Suppl. Figure S2).

The methodology proposed in this work has shown in a comprehensible way the characteristics of the machine learning methods that had taken part to the p16, SUMO, Crohn's disease and Hopkins' challenge. The R library presented in this work is available

for future challenge assessments in order to simplify and increase the reliability and reproducibility of future CAGI assessments.

Chapter 2

Intellectual Disability challenge assessment

In this chapter the assessment of the methods present in the CAGI 5 intellectual disability challenge will be presented under the concepts that have been exposed in the previous chapter. The test set of the ID challenge is the result of an innovative project, in which state of art NGS and bioinformatics methods have been applied to support neuro-developmental disorders diagnosis in a clinical environment. Neuro-developmental disorders (NDDs) comprehends a set of diseases, genetically heterogeneous and clinically different. Among them the most common is the Intellectually disability (ID), a disorder in which both intellectual and adaptive functioning capabilities are impaired. Co-occurrence between ID and other NDD disorders (in particular Autism Spectrum Disorders, ASD) makes the differential diagnosis complex. To support the clinical diagnosis of the NDD diseases, a panel of 74 genes, with a role in the development of ID or ASD disorders, has been developed. A total of 150 patients affected by ID and ASD have been sequenced with the proposed panel. The phenotype and genetic information of this data set could be used as a resource for the development of methods for detection of disease causing variants (Aspromonte et al., 2019). In the Intellectual Disability Challenge, a set of groups has developed methods for the prediction of NDDs, based on variant calling files. In the following paragraphs a description of the tools and strategies adapted to assess the reliability of the developed predictors will be presented.
This chapter is based on “Assessment of patient clinical descriptions and pathogenic variants from gene panel sequences in the CAGI-5 intellectual disability challenge. *Human Mutation* (2019). doi:10.1002/humu.23823”

Introduction

Neurodevelopmental disorders (NDDs) are a spectrum of disease conditions affecting brain development. Affected patients have increased manifestations as their childhood progresses, as the pathogenic conditions disturb normal brain development.

Manifestations usually start with a non-specific form of intellectual disability (ID), characterized by limitations both in intellectual functioning (reasoning, learning, problem solving) and in adaptive behavior, which covers a range of everyday social and practical skills. However, additional manifestations, such as autistic spectrum disorders (ASD) and epileptic seizures, can arise (Bowley and Kerr, 2000; Tonnsen et al., 2016). Structural abnormalities of the cranium (i.e. microcephaly, macrocephaly) may also be present at birth or appear postnatally. People with ID show also delayed motor development, which become evident with abnormalities in gait, such as ataxic gait (i.e. a lack of coordination in movement with a tendency to fall), hypotonia (general muscle weakness), or with 'unconscious' active motor behaviors (e.g. dyskinetic–dystonic movements or stereotypies) (Almuhtaseb et al., 2014). NDDs are clinically and phenotypically diverse, but driven by a substantial and overlapping genetic component, with numerous shared risk genes underlying these conditions (Mitchell, 2011). In particular, complex conditions such as ID and ASD have already been associated to hundreds of different genes. Next Generation sequencing (NGS) has led to the identification of many new NDDs genes with an excess of *de novo* mutations when compared to controls (Iossifov et al., 2014). Despite remarkable genetic heterogeneity, the findings from NGS and improvement in systems biology approaches, unraveled convergent biological pathways involved in brain development and help our understanding of disease pathophysiology (An et al., 2014; Barabási et al., 2011; Krumm et al., 2014; Pinto et al., 2014).

As NDDs can in principle be diagnosed even before birth by genetic tests, this has led to an increasing application of next-generation sequencing in clinical practice. Medical laboratories are routinely asked to screen hundreds of patients, which are either affected by NDDs or at risk of developing the condition. The limiting factor for successful diagnosis has therefore become the identification of causative mutations to associate to given pathogenic phenotypes. As most of these mutations are extremely rare or *private*, the problem is one of interpreting the effects of scores of variants of unknown significance on a wide range of candidate genes. This background fits well into the framework of the Critical Assessment of Genome Interpretation (CAGI) experiment, which has a declared goal of assessing methods to help interpret the effects of variants of unknown significance. A similar challenge was present in the CAGI-4 experiment with the Hopkins

gene panel, where predictors were asked to predict phenotypes based on the results of a genetic screening performed on a set of 83 genes associated to 14 different conditions (Chandonia et al., 2017).

The setup of the CAGI-5 ID challenge starts from a similar background. The Padua Genetics of Neurodevelopmental Disorders Lab at the Department of Woman and Child Health, University of Padua (henceforth, Padua NDD lab) has been using a gene panel to diagnose different NDD subtypes for the past couple of years. For the purpose of the CAGI-5 challenge, a dataset of 150 unpublished pediatric patients was released. Starting from the gene panel sequencing data, predictors were asked to predict (a) the phenotypes and (b) their causative or potentially causative variations for each patient. Phenotypes have been derived from the clinical notes collected by geneticists visiting the patients. Candidate variants have been validated by segregation analysis, i.e. verifying their absence in the parents according to the *de novo* paradigm, inherited from affected parents. It should be noted that this is a difficult “open world” CAGI challenge, as clinical notes may be somewhat subjective and only a subset of genes have been screened. Furthermore, the phenotypic traits to predict are pathophysiology conditions that can be present in different NDDs, thus, in contrast to the CAGI-4 Hopkins challenge, patients may manifest more than one of these phenotypes, in different combinations.

The challenge is realistic as it represents the issue of assigning causative mutations to complex neurological diseases in clinical practice. In a few selected cases, consistent predictions were used to challenge previous assumptions and have led to a revised molecular diagnosis.

Materials and Methods

Sequencing, variant nomenclature and analysis by the Padua NDD lab

Coding sequences and nearest flanking regions of 74 genes were targeted for deep sequencing with a custom Ampliseq panel assay using a mixture of oligonucleotides generating 1,834 amplicons covering 520 kb. Multiple indexed libraries were pooled and sequenced on the Ion PGM platform (Thermo Fisher Scientific). Alignment and variant calling were performed with the Ion Torrent Suite Software v 5.02 (Thermo Fisher

Scientific). The panel of 74 genes was sequenced in 150 individuals referred to the Padua NDD lab for intellectual disability with or without autistic features. VCF files of the 150 patients were provided to the CAGI-5 organizers with clinical information regarding the presence of seven ‘phenotypic traits’ for each patient (Suppl. Table S4, Appendix 2). The clinical information was provided by the patient’s physician, which were asked to fill a clinical record for each patient. When the clinician leaved a field empty, we indicated information about the specific trait as not available, although we cannot exclude that some patients may present it. The Padua NDD lab also indicated the identified variants of the sequenced genes that have been classified as causative, putative, or contributing factors (Suppl. Table S5). Causative variants are supported by segregation analysis and genotype-phenotype correlation, while "putative" ones are rare or novel variants predicted as pathogenic for which segregation analysis is not available. Contributing factors are rare or novel variants predicted as pathogenic, inherited from apparently healthy parents, mapping on genes that confer a risk but are not sufficient to cause the disease, mapping on genes causing ASD susceptibility, or found mutated in individuals with very mild phenotypes. Table 4 summarizes the amount of patients with variants associated to each phenotype.

To evaluate the putative clinical impact of the variants, the following criteria were applied: 1) allele frequency <0.002% in the Gnomad database, or <0.45% for variants in autosomal-recessive genes, as indicated by(Piton et al., 2013; Whiffin et al., 2019) 2) absence of the variant in other samples (in-house database), 3) stop gain, frameshift and splicing variants were a priori considered to be most likely pathogenic, 4) for missense mutations, amino acid conservation and consensus of pathogenicity predictions were evaluated, 5) inheritance mode, 6) phenotypic consistency with the clinical signs associated to mutations in the same gene.

It is important to note, that for a diagnostic purpose, the thresholds used by the Padua NDD lab to filter candidate variants, have been calculated based on the assumption that the patient phenotype follow a Mendelian transmission.

Whiffin and colleagues demonstrated that for human Mendelian disease clinical genome interpretation is empowered by using high-resolution variant frequencies (Whiffin et al., 2019). To select candidate variants responsible for ID, Piton and colleagues suggested

to filter variants with a frequency compatible with the incidence of the disease (i=2% in the general population) (Piton et al., 2013). Since the repeat expansion on FMR1 gene remains the most frequent cause of X-linked forms of ID and given the genetic heterogeneity of NDDs, we expect that mutations in other genes account for less than 0.1% of all ID cases, resulting in a disease frequency <0.002% (i= 0.02 x 0.001). Variants in genes associated with recessive disorders should not exceed the threshold of 0.45% (√0.002%).

Phenotype	Patients	Disease causing	Putative	Contributing factor	All variants
ID	49	25	18	12	55
ASD Autistic traits	31	14	12	10	36
Epilepsy	18	9	8	2	19
Microcephaly	8	5	2	1	8
Macrocephaly	4	4	0	0	4
Hypotonia	6	4	1	1	6
Ataxia	3	1	2	1	4

Table 4. Patients for whom the Padua NDD lab identified at least one causative or potentially disease variant in the answer key, summarized by phenotype. Each variant is specific for each patient and one patient can be associated to more than one phenotype.

Challenge format

Participants were provided with 150 VCF files, one per patient, a detailed description of the seven disease phenotypes given in Suppl. Table S4, the 74 gene identifiers, the gene captured regions used in sequencing the patients in Browser Extensible Data (BED)

format, a submission template, and a submission validation script. Furthermore, participants were informed that each patient may have more than one phenotypic trait, and all have at least one.

Participants were asked to submit the predictions of phenotypic traits and causative variants for each patient, based on their gene panel sequences. For each submission, participants were required to predict the probability that a patient has a referring phenotypic trait in each of the 7 phenotypic classes provided, as well as the predicted causal variant(s) from the gene panel sequence dataset for every disease class with a non-zero probability. Each predicted disease class probability also included a mandatory standard deviation (SD) field indicating the confidence prediction, with low SD indicating high confidence and high SD indicating low confidence.

Assessment

The prediction assessment was focused on evaluating the predictive ability of the different submissions, considering their performance on each disease phenotype. This approach has been successfully used for the analysis of multilabel classifier performance, since it focuses on a set of two-class prediction problems (Fawcett, 2006). It also simplifies the assessment procedure, allowing to compare and highlight different method performances on each single phenotype, instead of evaluating them considering the whole predicted class matrix (150 x 7, one prediction for each patient and phenotype).

Predicted disease classes for each submission were assessed against the clinical phenotype given in the Padua NDD lab answer key, using the procedure described below. When the predictors did not provide a probability value leaving the asterisk on the template file, it was treated as probability zero in the assessment.

The first phase of the assessment procedure was the conversion of submitted probability values to positive (1) or negative (0) classes. The conversion was done by each phenotype column, considering as threshold the probability value which maximizes the Matthew correlation coefficient (MCC) for that phenotype. We compared all probability values of each phenotype with the corresponding threshold and assign 0 or 1 if the value is lower or higher, respectively. In addition, different performance measures were used to assess the predictions for each phenotype (Table 5). Sensitivity and specificity have been

used to evaluate model capability to detect positive cases and discriminate between positive and negative classes. The MCC, accuracy (ACC) and F1 measures have been used to evaluate both negative and positive predictions at the same time. Particularly, MCC has been proven to be less influenced by an unbalanced dataset (Vihinen, 2012), as is the case of this challenge where some phenotypes are completely unbalanced (Figure 11). The Area Under the Curve (AUC) was calculated for these. The final ranking of predictors has been based on AUC, since this measure considers the submitted probability values and not the previously described optimized conversion purposed for indexes based on a contingency matrix as MCC.

The R scripts used to perform the assessment are publicly available from the GitHub repository at URL: https://github.com/BioComputingUP/CAGI-ID-assessment.

Submission	ID				ASD				Epilepsy				Microcephaly				Macrocephaly				Hypotonia				Ataxia			
	AUC	**MCC**	**ACC**	**F1**	**AUC**	**MCC**	**ACC**	**F1**	**AUC**	**MCC**	**ACC**	**F1**	**AUC**	**MCC**	**ACC**	**F1**	**AUC**	**MCC**	**ACC**	**F1**	**AUC**	**MCC**	**ACC**	**F1**	**AUC**	**MCC**	**ACC**	**F1**
1.1	0.57	0.20	0.95	0.98	0.51	**0.20**	**0.73**	**0.84**	0.53	**0.24**	**0.61**	**0.71**	0.57	0.16	0.46	0.41	0.64	0.27	0.86	0.15	0.49	0.20	0.63	0.36	0.46	0.08	0.78	0.14
2.1	0.70	0.16	0.71	0.82	0.55	0.10	0.36	0.22	0.38	-0.01	0.48	0.07	0.51	**0.37**	**0.81**	0.29	0.55	0.18	0.81	0.29	0.52	0.06	0.51	**0.51**	0.61	0.27	0.81	0.17
2.2	0.75	0.22	0.81	0.89	0.5	0.09	0.36	0.25	0.41	0.03	0.49	0.10	0.55	0.36	**0.81**	0.40	0.50	0.09	0.75	0.23	0.47	0.11	0.60	0.23	0.66	0.28	0.81	0.29
2.3	0.71	0.16	0.71	0.82	**0.56**	0.11	0.59	0.68	0.39	-0.01	0.48	0.07	0.49	**0.37**	**0.81**	0.29	0.56	0.18	0.81	0.29	0.51	0.07	0.53	0.50	0.66	0.33	0.72	0.48
2.4	**0.78**	0.21	0.79	0.88	0.49	0.09	0.36	0.25	0.41	0.03	0.49	0.10	0.48	0.36	**0.81**	0.40	0.56	0.14	0.75	0.29	0.49	0.11	0.60	0.23	**0.72**	**0.37**	0.72	**0.52**
2.5	0.64	0.09	0.55	0.70	0.55	0.10	0.36	0.22	0.40	-0.01	0.48	0.07	0.53	**0.37**	**0.81**	0.29	0.57	0.18	0.81	0.29	0.49	0.06	0.59	0.18	0.56	0.27	**0.81**	0.17
2.6	0.74	0.18	0.75	0.86	0.46	0.09	0.36	0.25	0.41	0.03	0.49	0.10	0.49	0.36	**0.81**	0.40	0.55	0.14	0.75	0.29	0.5	0.11	0.60	0.23	0.66	0.32	0.78	0.45
3.1	0.51	0.12	0.38	0.53	0.52	0.10	0.36	0.22	0.43	-0.10	0.47	0	0.50	0.11	0.78	0.10	0.62	0.18	0.65	0.33	0.54	0.21	0.62	0.13	0.49	-0.07	0.78	0
3.2	0.55	0.13	0.42	0.58	0.52	0.12	0.36	0.23	0.43	-0.07	0.46	0.17	0.50	0.11	0.78	0.10	0.62	0.18	0.65	0.33	0.52	0.15	0.60	0.07	0.49	-0.07	0.78	0
3.3	0.68	**0.34**	**0.97**	**0.99**	0.48	0.05	0.36	0.24	0.44	-0.05	0.47	0.20	0.50	0.11	0.78	0.10	0.62	0.18	0.65	0.33	0.52	0.15	0.60	0.07	0.49	-0.07	0.78	0
4.1	0.61	0.15	0.83	0.91	**0.56**	0.18	0.36	0.19	0.54	0.19	0.51	0.14	0.57	0.23	0.56	0.44	**0.70**	**0.39**	0.79	**0.48**	0.51	0.17	0.62	0.19	0.45	0.27	0.81	0.17

4.2	0.61	0.11	0.78	0.87	**0.56**	0.18	0.36	0.19	0.53	0.19	0.51	0.14	0.56	0.23	0.56	0.44	0.69	**0.39**	0.79	**0.48**	0.52	0.17	0.62	0.19	0.47	0.27	0.81	0.17
4.3	0.68	0.19	0.88	0.94	**0.56**	**0.20**	**0.73**	**0.84**	**0.56**	0.22	0.59	0.51	**0.63**	0.30	0.52	**0.47**	0.67	**0.39**	**0.88**	0.38	**0.56**	**0.25**	**0.65**	0.33	0.46	0.28	0.81	0.29

Table 5. Summary of performance measures for all submissions and phenotypes.

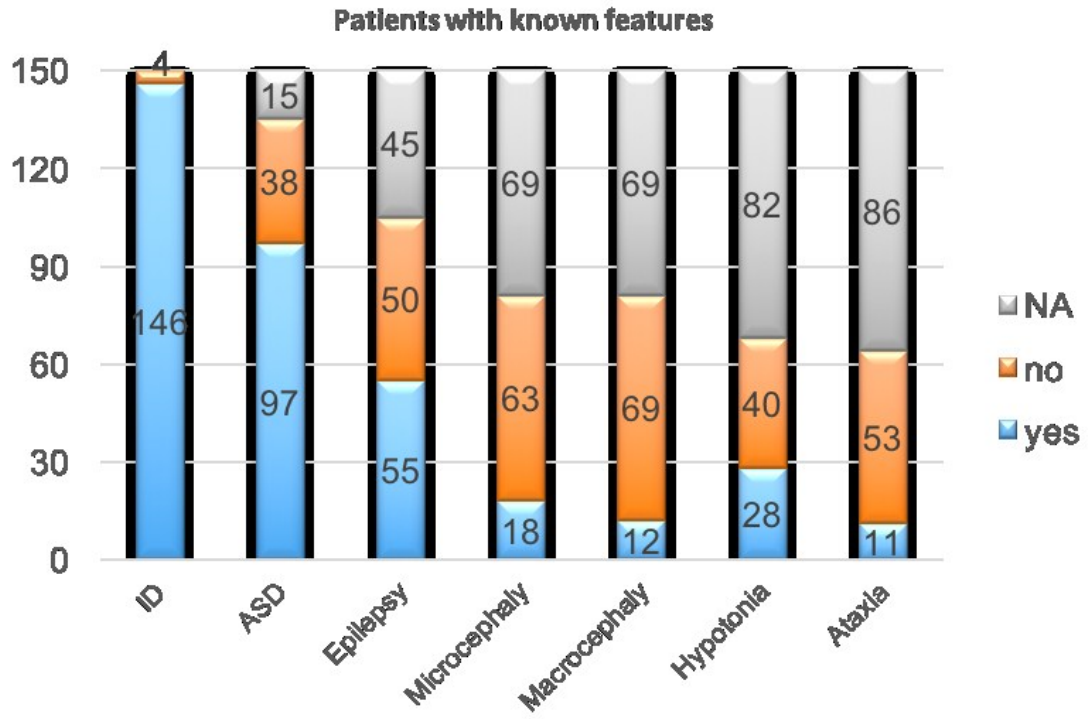


Figure 11. Summary of CAGI-5 intellectual disability challenge experimental data. Number of patients where the presence or absence of the phenotype was ascertained by a clinician.

Prediction methodology

A total of four groups, plus a late prediction (which can be found in the Supplementary Material), submitted predictions for the ID challenge. The group prediction approaches are summarized in Table 6 and described in detail below.

Group 1 (Mooney - Radivojac Lab): Annotation of the protein coding variant in the raw VCF files was performed using ANNOVAR, including extraction of wild type and mutant protein sequences (Wang et al., 2010). Pathogenicity prediction scores were assigned to missense, stop gain, and frameshifting indel variants with Mutpred2 (Pejaver et al., 2017) and Mutpred-LOF (Pagel et al., 2017). In each individual, phenotypic trait risk was determined based only upon the variant with the highest pathogenicity prediction score across a set of phenotype-specific risk genes. For each phenotypic trait, a list of risk

genes that are known to harbor disease-causing variants associated with that phenotypic trait was compiled from the Human Gene Mutation Database (HGMD) (Stenson et al., 2003). Gene lists were extended, particularly those with fewer known risk genes (macrocephaly, hypotonia and ataxic gait), with the PhenoPred web tool (Radivojac et al., 2008) and a gene prioritization algorithm. Confirmed risk genes have been used as “seed" genes on the human protein-protein interaction network for running a network propagation algorithm (Nabieva et al., 2005). The propagation algorithm was performed in a 5-fold cross validation manner so as to get an initial score between [0, 1] for all the genes. The AlphaMax algorithm (Jain et al., 2016) was used to estimate the positive proportion of the risk genes and calibrate those initial scores to be proper probability scores measuring the likelihood of a gene being associated with the disease. For each phenotypic trait, the probability was MutPred2 or MutPred-LOF score of the highest scoring variant in the associated risk genes.

Group 2 (Moult Lab): The 150 VCF files (one VCF file per patient) provided for the challenge were annotated using the Varant tool [http://compbio.berkeley.edu/proj/varant/Home.html], including region of occurrence (intron, exon, splice site or intergenic), observed minor allele frequencies (MAF), mutation type, predicted impact on protein function, and previously established associated phenotypes reported in ClinVar (Landrum et al., 2016). The RefGene (Pruitt et al., 2014) gene definition file was used for gene and transcript annotations in Varant. In addition, in-house scripts were written to further annotate the VCF files with HGMD (Stenson et al., 2003) disease-related variants, with dbscSNV (Jian et al., 2014) and SPIDEX (Xiong et al., 2015) variants that potentially alter splicing, and with REVEL (Ioannidis et al., 2016) scores for missense variants. A quality control (QC) analysis were performed to exclude outlier samples (see Suppl. Material). The transition/transversion ratio (Ts/Tv) and heterozygous/homozygous ratio were compared to the 1000 Genomes dataset for the genomic regions captured for sequencing in the challenge dataset. Comparison of common, rare, and novel variant counts across samples was also performed. The 74 genes were mapped to one or more of the seven phenotype traits using two independent approaches generating two different gene-phenotype mapped files. In addition to the OMIM database, the Genetic Home Reference (https://ghr.nlm.nih.gov/) or Human

Phenotype Ontology (https://hpo.jax.org/app/) databases, respectively, were used to map the phenotypes to the genes. The variant prioritization procedure was performed on each of these phenotype lists. Only rare variants (MAF less than or equal to 1% in Exac (http://exac.broadinstitute.org) or novel variants (not reported in ExAC), flagged as PASS in the VCF files, were considered. Indels in low complexity regions (LCR) were excluded from the analysis, based on the LCR dataset pre-computed for the human genome by Heng Li (Li, 2014). A strand bias filter was used to remove variants whose alternate allele was present only on one strand of the reads mapped to the variant position. Variant prioritization was based on two main criteria, variant quality and variant impact, that were applied in a sequential manner to each sample. For each criterion, five different levels of variant quality and 13 different types of variant impact were defined respectively. Putative causative variants identified were further filtered for inheritance model associated with the gene, according to the available information for the gene concerned in OMIM and Genetic Home Reference database. To compute a probability score, i.e. the probability of a variant causing a disease phenotype, a number of ad hoc procedures were used. An exception was for missense variants, where the probability was assigned using the extent of consensus among the four missense-analysis methods, previously calibrated from HGMD data and a control set of inter-species variants. Other variant types were subjectively assigned probabilities depending on the severity of the impact. Furthermore, depending on the considered mode of inheritance, the probability score was adjusted. Ad hoc probabilities of a correct variant call were also assigned to each variant based on the variant quality filters. Six different predictions were performed based on the two different gene-phenotype lists and different combination of probabilities.

Group 3 (Lichtarge Lab): Variants of poor sequencing quality (QUAL<80) were excluded from the analysis and the rest variants were annotated with ANNOVAR (Wang et al., 2010). There were three submissions that used i) only missense, ii) missense and nonsense, and iii) all variations. The effect of each variant was estimated with the Evolutionary Action (EA) equation (Katsonis and Lichtarge, 2014) and the function loss of each gene was calculated as:

$$LOF_g = 1 - \prod \left(1 - \frac{EA_i}{100}\right) \quad (23)$$

where ∏ indicates the product for all mutations ***i*** in that gene. Nonsense and fs-indel variants were given EA of 100, while silent variants were given EA of 0. Genes were also weighted for their ability to tolerate mutations (w_g), calculated as the fractional rank of the average EA score of mutations seen in the gnomAD data (Lek et al., 2016). The weighted loss of function of each gene (w_g*LOF_g) was used as starting value for diffusion across the CTD gene-disease network (Mattingly et al., 2003). Diffusion scores were calculated for each disease (Lin et al., 2019) and a collective burden was calculated for each of the seven disease categories (normalized between 0-1). The relative ratios of the collective burden of the disease categories was used as the probability that a patient belongs to that disease category. The variants that contributed most to the collective burden of each disease category were reported as the causal variants.

Group 4 (Brenner Lab): This group used their software CHESS v0.1 adjusting some parameters to perform predictions for the CAGI-5 ID challenge. Public data used on CHESS are variant frequency data from GNOMAD v2.0.2 (Lek et al. 2016), pre-calculated variant deleterious scores by REVEL (Ioannidis et al., 2016), and clinical evidence data from ClinVar (Landrum et al., 2016) (Landrum et al. 2016) (downloaded on 2017-10-02). Phenotype matching scores for all genes were calculated using Phenolyzer (Yang et al., 2015). Pre-called variants from the case exome were annotated with data using VEP (McLaren et al., 2016), GNOMAD variant frequency data, ClinVar evidence, and the pre-calculated REVEL scores. To reduce the computing burden, common (variants with MAF >= 5%) and non protein-altering variants have been excluded from the analysis. The selected variants were scored based on quality of data, impact severity, phenotype-match score. Different scoring adjustments were also performed based on the inheritance mode considered. The three submissions correspond to three models with different stringency in the final decision, based on variant frequency in the 150 patient cohort and the probability score threshold used for each prediction. Phenotypic features were associated to each patient by a clinician. Although all patients have at least one feature assigned, the phenotypes were not equally represented in all patients. Figure 11 shows that most of the patients have ID, ASD, or Epilepsy. Other phenotypes (Microcephaly, Macrocephaly, Hypotonia and Ataxia) were less frequently observed in these patients.

Nevertheless, for many patients no information was available about the presence or absence of a phenotype.

Group						Filters		
ID	**Submission**	**Name**	**Annotation**	**Gene-Phenotype**	**Variant impact**	**frequency**	**low quality**	**Inheritance model**
1	1.1	Mooney-Radivojac	ANNOVAR	HGMD, PhenoPred, and PPI for network propagation	MutPred2 and MutPredLOF	-	-	-
2	2.1, 2.2, 2.3, 2.4, 2.5, 2.6	Moult Lab	Varant	OMIM+GHR, OMIM+HPO	13 levels of variant impact	SNVs >1%, SNVs in LCR low complexity region	yes	yes
3	3.1, 3.2, 3.3	Lichtarge Lab	ANNOVAR	Diffusion on CTD (Comparative Toxicogenomics Database) associations	Evolutionary Action	No	yes	no
4	4.1, 4.2, 4.3	Brenner Lab	CHESS v0.1	Phenolyzer	VEP, REVEL score	SNVs MAF>5%	yes	yes

Table 6. Computational approaches adopted by different groups.

Variant prediction assessment

Predictors have been also assessed for their ability to detect variants in 50 patients where clinicians have noted at least one variant probably associated to the phenotype. Figure 12 shows variant predictions for all patients and phenotypes by the different submissions. The amount of experimental variants with their corresponding classification are shown in the first three bars on the plot. Submissions of group 2 show the highest amount of predicted variants associated to the different patient phenotypes (37 out of 56). Indeed, Group 2 outperformed other groups for causative (16 out of 25), putative causative (12 out of 18) and contributing factor (9 out of 13) variants. Submission 3 of group 4 was the second group predicting most of the variants. They correctly predicted 29 variants (11 causative, 9 putative causative and 9 contributing factor variants). In addition, Figure 13 shows the fraction of each mutation type correctly predicted by the different groups. It is possible to see that just a small amount of variants were correctly predicted by all groups. The 28% of causative and 15% of contributing variants were correctly identified by at least 3 groups. On the other hand, 17% of putative variants were predicted by at least 3 groups. Table 7 contains the fraction of correctly predicted variants by each group submission. Group 2 did not only predict most variants but also obtained the highest fraction of correctly predicted variants, calculated as the amount of variants correctly predicted divided by all the predicted variants for all patients and phenotypes

Suppl. Table 5 summarizes all variants noted by the Padua NDD lab and the groups which predicted them correctly. All 25 causative variants, except the *SHANK3* frameshift indel chr22:51159830:A:TTC in patient MR1970.01, were correctly predicted by at least one group. After the initial assessment, we realized that this complex genetic event (nucleotide substitution chr22:51159830:A:C plus a TT insertion) was molecularly characterized by Sanger validation of the chr22:51159830:A:C variant, but the variant caller plugin failed to call the insertion at near position of the same reads. However, group 4 correctly predicted chr22:51159830:A:C as a potentially pathogenic variant.

The Padua NDD lab considered some causative missense variants difficult to interpret (*ATRX*: p.N1377S; *RAB39*B p.F193L; *GRIA3* p.R216Q; *MED13L* p.G706E), since pathogenicity predictions were discordant, allele frequency in control cohorts higher than expected, or proband phenotype partially consistent with those associated to the gene. However, the majority of the groups was able to predict these correctly. One example is the maternally inherited X-linked p.F193L of the *RAB39B* gene associated to recessive X-linked Mental Retardation syndrome (MR-XL72, OMIM 300271) or to Waissman syndrome, which is characterized by ID and early-onset Parkinson disease (OMIM 311510). This variant is predicted damaging by three out of twelve computational tools provided by ANNOVAR (LRT, Mutation Taster, and fathmm MKL), is moderately conserved during evolution, and present in a hemizygous state in two control cohort individuals. However, the variant maps to the C-terminal hypervariable tail of *RAB39B* which is relevant for protein interactions involved in protein targeting. The mother transmitting the p.F193L variant has a mild phenotype, consistent with those reported in the literature associated to a missense mutation at the close p.Gly192Arg position (Mata et al., 2015).

At least one group correctly predicted 16 out of 18 putative mutations. In particular, 7 variants were indicated by the majority of the groups. Three of these 7 variants were inherited and suspected to contribute to the disease together with other genetic or environmental factors. For the other four cases, after the CAGI-5 assessment, we contacted the families to follow up the molecular finding carrying out segregation analysis of the identified variants. Only one family answered our call, which allowed us to characterize the de novo status of the p.Y381H variant in the *CASK* gene. Even if the pathogenicity predictions were discordant, this variant was absent from control cohorts and in silico analysis suggested a structural role of this residue in the homo and hetero-dimerization of the *CASK* protein (Aspromonte et al., 2019). The proband phenotype is also consistent with those associated with *CASK*-related disorders.

In addition, at least one group correctly predicted the 13 variants classified as contributing factor, of which seven were indicated by the majority of the groups. This variant class is particularly relevant for autism susceptibility.

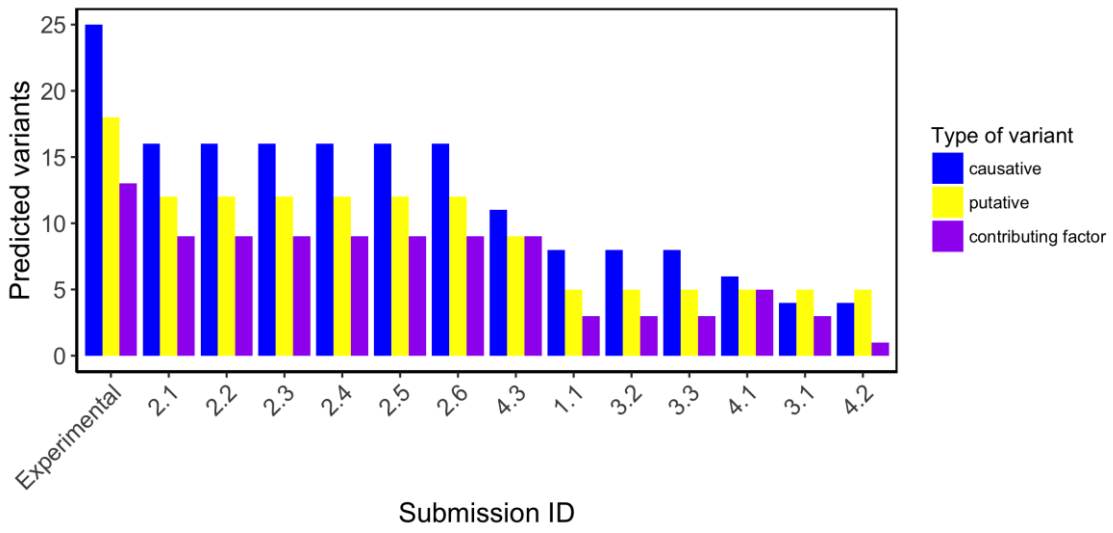


Figure 12. Predicted variants distribution. Category Experimental is the amount of variants which were identified and classified by the Padua NDD lab. Each bar represents the amount of variants and type predicted by each submission.

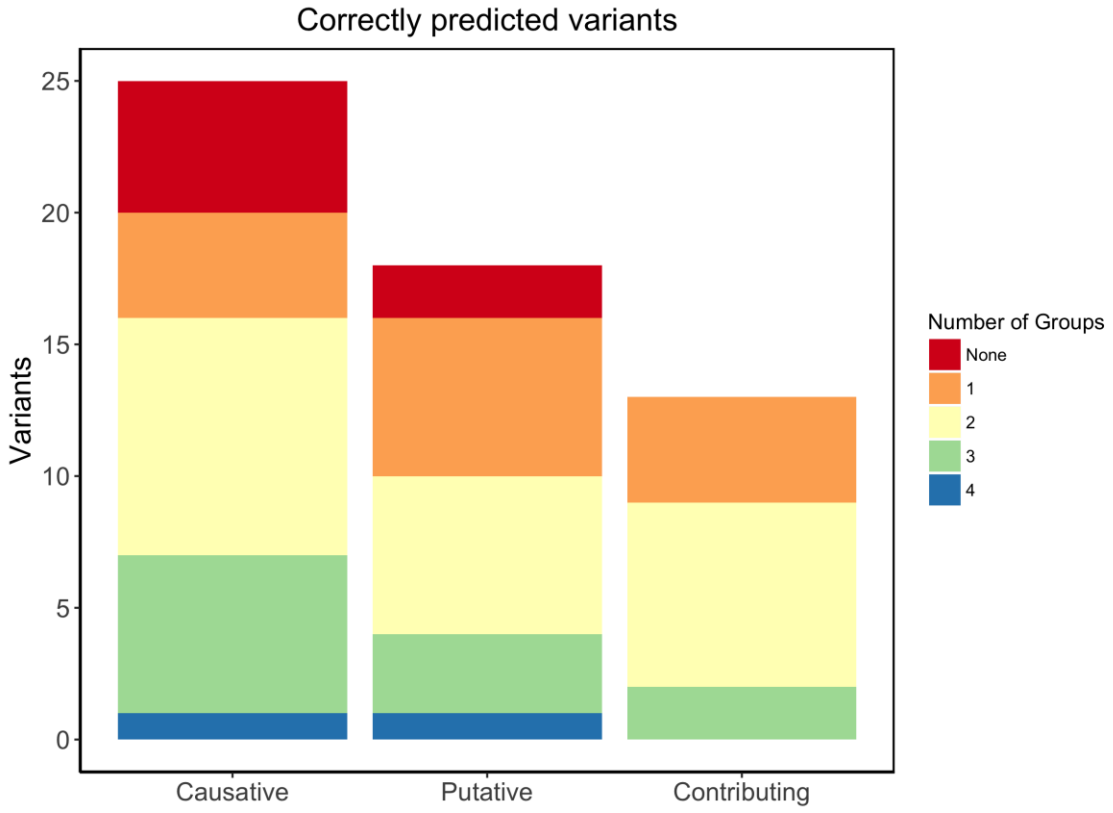


Figure 13. Amount of variants classified by their effect. Colors indicate the proportion and number of groups which correctly predicted those variants.

Submission	Corr. pred. Variants	Total pred. Variants	Corr. pred. Variants / Exp. Variants	Corr. pred. Variants / Total pred. Variants
1.1	16	228	0.29	0.07
2.1	**37**	174	**0.66**	**0.21**
2.2	**37**	171	**0.66**	**0.21**
2.3	**37**	174	**0.66**	**0.21**
2.4	**37**	171	**0.66**	**0.21**
2.5	**37**	174	**0.66**	**0.21**
2.6	**37**	171	**0.66**	**0.21**
3.1	12	129	0.21	0.09
3.2	16	135	0.29	0.12
3.3	16	148	0.29	0.11
4.1	16	157	0.29	0.10
4.2	10	113	0.18	0.09
4.3	29	290	0.52	0.10

Table 7: Summary of variants prediction assessment by each submission.

Discussion

We have described the assessment of the CAGI-5 ID challenge. This challenge is based on the phenotype evaluation of patients using gene panel sequences, in analogy to the CAGI-4 Hopkins panel challenge (Chandonia et al., 2017). Where the Hopkins panel was testing for different monogenic diseases with Mendelian inheritance, the ID challenge focuses on complex disorders. Neurodevelopmental conditions are characterized by strong clinical comorbidity and a complex genetic architecture (Mitchell, 2011)). The genetic information for each patient can at best be considered partial, as compounded by the rather limited fraction of patients (33%) where a putative or causative variant has been detected by the Padua NDD lab. As such, the CAGI-5 ID challenge can be expected to be more difficult that the CAGI-4

Hopkins panel. However, due to the genetic heterogeneity seen in NDDs, the presence of negative cases in the data set reflects the clinical practice, where the sequenced genes cannot explain the phenotype of all tested individuals. This implies that the identified rare variants should be interpreted with caution.

The phenotype prediction component of the ID challenge makes it also similar to the Personal Genome Project (PGP) challenge from previous rounds of CAGI (Cai et al., 2017). In the CAGI-2 PGP challenge, participants were initially asked to predict the presence of a set of phenotypic traits. Later CAGI editions turned the challenge into a matching game between sets of phenotypic profiles and genetic data. The ID challenge is similar to the original PGP challenge, but with a narrower focus on NDDs. Like PGP, it emphasizes complex disease conditions whose genetic bases are not fully understood. It is indeed increasingly accepted that the genetic architecture of NDDs involves the interplay of *de novo*, rare, and many common (>1% frequency) variants, which have a potential role in phenotype variability and severity of the disease. Furthermore, besides some well known monogenic conditions there are oligo- or polygenic forms with multiple gene-gene or gene-environment interactions (Lesch, 2016; Mitchell, 2011).

Despite these difficulties, several predictors participating in the CAGI-5 ID challenge were able to achieve AUC > 0.6 for three non-trivial phenotypes (microcephaly, macrocephaly and ataxia) and also for the ID phenotype which was heavily biased to the positive case. Intriguingly, group 4 (Brenner lab) has been able to make acceptable predictions for most of the individual phenotypic traits, except ataxia (Figure 14). Furthermore, considering the overall clinical manifestations of each patient, for 93 individuals (62%) the correct phenotype has been predicted by at least one group. In particular, group 1 predicted 49 of them (52%), group 4 (submission 3) predicted 46 of them (50%) and 57 (61%) considering their three submissions (Table 8). Finally, group 2 correctly predicted 43 of them (46%) considering all their six submissions. Group 2 in particular accurately predicted each of the seven phenotypic traits in 8 individuals, and the overall phenotype in 12 patients that were not correctly predicted by other groups. Even though this performance is not promising, we have

to consider the extreme difficulty to predict a combination of several pathological conditions that often occur in comorbidity with variable expression and severity.

The assessment on phenotype prediction has also been performed considering only the patients with variants noted by the Padua NDD Lab, both considering each phenotypic trait individually and for the combination of the seven traits. We hypothesized that the phenotype of the individuals carrying a disease mutation must be easier to predict. Furthermore, the Hopkins challenge in CAGI-4 noted a higher performance of the prediction methods in phenotype prediction of cases where the Hopkins lab reported a variant, with at least one group correctly identifying the disease class in 84% of these patients. However, in the CAGI-5 ID challenge there were no improvements in the performance of methods (for instance see the number of patients for which correct variants and phenotypes were predicted by each group: the nC-nCV index on Table 8). Surprisingly, group 2, which performed better in the causative or putative variants prediction, was less accurate in predicting phenotypic traits. Something similar occurs when we tried to remove patients for whom no method was able to correctly predict the phenotype (e.g. correctly predict the presence or absence of each class). We again observed that while some methods improved their performance, others decreased it.

In contrast to the Hopkins challenge, the Padua NDDs lab participated in the assessment of the challenge and provided feedback on predicted variants by the groups. This allowed us to observe that variants supporting the predictions of some groups, in particular group 1 and group 3, are rare or common variants with weak pathogenic predictions. Some of these variants were previously excluded by the Padua NDD lab as inherited from healthy parents (Aspromonte et al., 2019). However, it seems that taking into account the contribution of these inherited rare or common variants may help in the phenotype prediction. This can be explained by the complex genetic architecture of NDDs and the recent findings that different variants cluster in common pathways to determine the expression of the disease (Mitchell, 2011). Thus, particularly for phenotypic traits with little genetic information, group 1 used protein-protein interaction networks to expand the gene-phenotype association, which has been useful to select relatively low frequency variants with less functional impact that

may contribute to the expression of the phenotypic trait. Moreover, group 4 created their gene-phenotype association list using a well establish tool for the prioritization of risk genes in complex diseases.

However, no less important is that some groups made correct predictions based on variants that were excluded by the Padua NDD lab as sequencing errors. Methods using good quality filters, such as groups 2 and 4, are more reliable than others. Nonetheless, the Padua NDD lab reconsidered some of these predicted variants and validated them with Sanger sequencing and segregation analysis. Even if many of the reconsidered variants did not change the molecular diagnosis of the tested patients, the re-assessment of the interpreted data allowed to fix some rules in filtering sequencing errors and interpretation of variants, such as synonymous variants, that can be missed as causative. In particular, re-assessing putative variants that were predicted by the majority of groups as pathogenic, allowed us to select a limited set of putative variants for further investigation. The re-evaluation by segregation analysis was possible only for one family that answered our call. The variant resulted *de novo*, supporting the causative role of a probably hypomorphic *CASK* mutation in a male with a phenotype consistent with a *CASK*-related disorder.

This CAGI-5 challenge has provided a realistic framework to assess the performance of prediction methods in clinical practice. Despite all its inherent limitations, we believe it has demonstrated promising results and avenues for possible future improvements. We will hopefully be able to measure improvement over the next edition of the CAGI experiment.

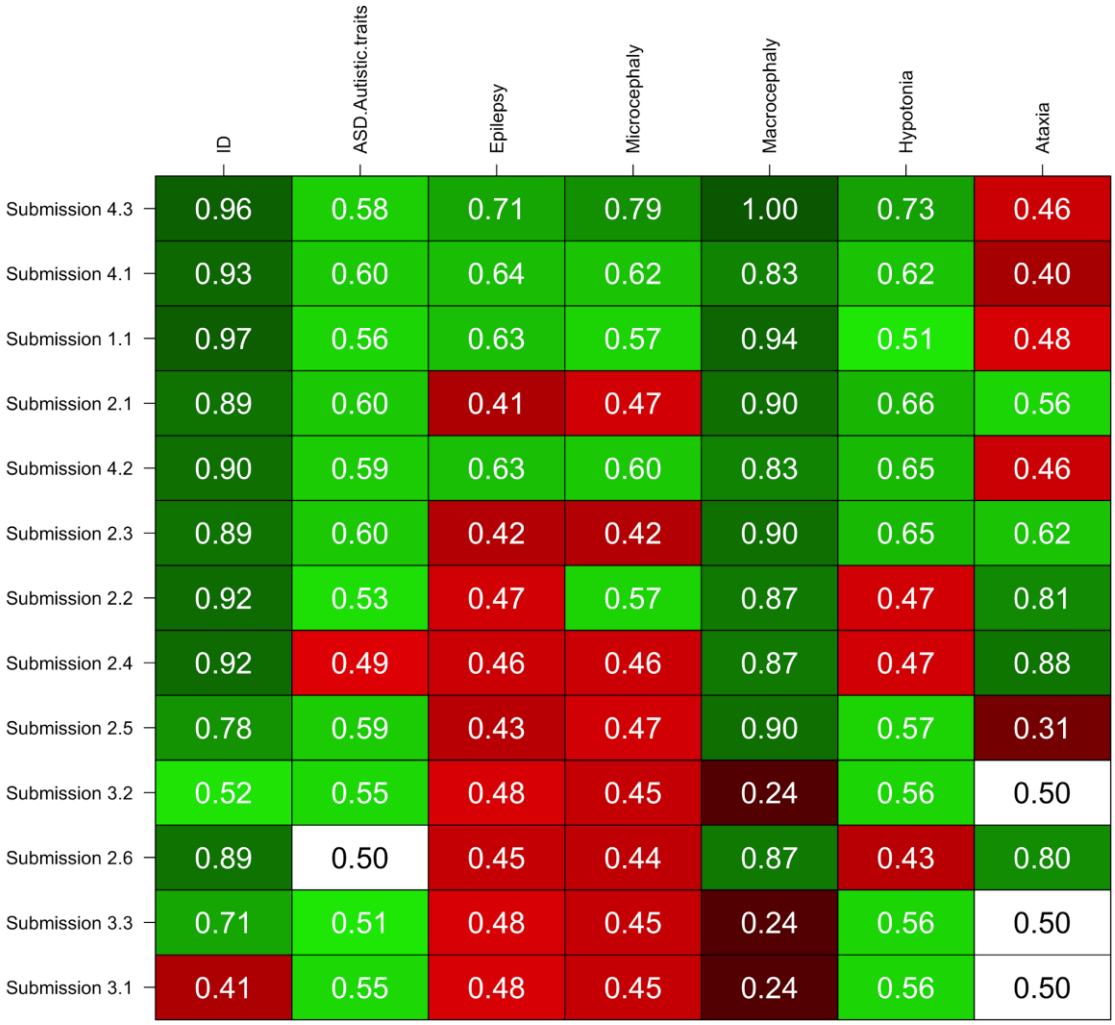


Figure 14. Overall performance for each submission on phenotype prediction. Each cell represents the AUC values. Submissions are ordered by the AUC average rank among all the phenotypes. The color scale ranges from dark green (+1, perfect performance) to dark red 0, bad performance). White means random performance.

Submission	nC	nCV	nC-CV	nC1	nC2	nC3	nC4	nC5	nC6	nC7
1.1	49	16	6	3	22	14	2	5	2	1
2.1	21	37	3	2	8	1	6	1	1	2
2.2	24	37	7	3	7	1	4	2	1	6
2.3	25	37	4	2	11	2	5	3	0	2
2.4	23	37	7	3	7	1	4	1	0	7
2.5	18	37	3	2	7	1	2	2	1	3
2.6	24	37	7	3	7	1	4	2	1	6
3.1	16	12	2	3	7	0	0	3	0	3
3.2	18	16	3	3	7	0	0	4	0	4
3.3	26	16	3	3	10	1	3	3	1	5
4.1	19	16	1	3	7	1	2	2	1	3
4.2	17	10	1	3	7	1	2	2	1	1
4.3	46	29	11	3	19	11	7	4	1	1

Table 8. Summary of phenotype and variant prediction for all patients. nC is the number of patients where their phenotypic trait/s was/were correctly predicted. nC1, nC2, nC3, nC4, nC5, nC6 and nC7 are similar to nC but considering the number of phenotypic traits (from 1 to 7) provided for that patients by Padua NDD Lab. nCV is the number of patients for whom variants were correctly predicted. nC-CV mean the number of patients for whom their phenotypes and variant/s were correctly predicted.

Chapter 3

PCM1 challenge assessment

In this chapter the assessment of the tools presented in CAGI 5 PCM1 challenge will be described. In this experiment, different methods have been developed to predict, in silico, the effect of a set of variants on PCM1 function. Different works has proven the association between this gene, orbitofrontal gray matter volumetric deficits and schizophrenia, for which causing variants on PCM1 have been reported. The Katsanis lab has experimentally determined the effect of 38 variants on this gene in zebrafish, in terms of brain ventricle formation. In particular the effect of each variant, tested in one sample, was compared to two other samples, one representing the fish in normal condition (MO + WT), the other the animal model with the PCM1 gene inactivated (MO). A student t-test was computed to estimate the p-value of the difference between the first group and each one of the remaining. Each variant was then classified as loss of function, hypomorphic or benign on the basis of the obtained p-values (Figure 15). Therefore the participants of the CAGI 5 PCM1 challenge were required to submit the two p-values, (MO, MO+WT) and the predicted effect for each variant. We evaluated the different models taking in account only the predicted class, since the p-values are only indicators of statistically significance and had a weak correlation with experimental ones.

This chapter is based on “Performance of computational methods for the evaluation of Pericentriolar Material 1 missense variants in CAGI-5. *Human Mutation* (2019). doi:10.1002/humu.23856”

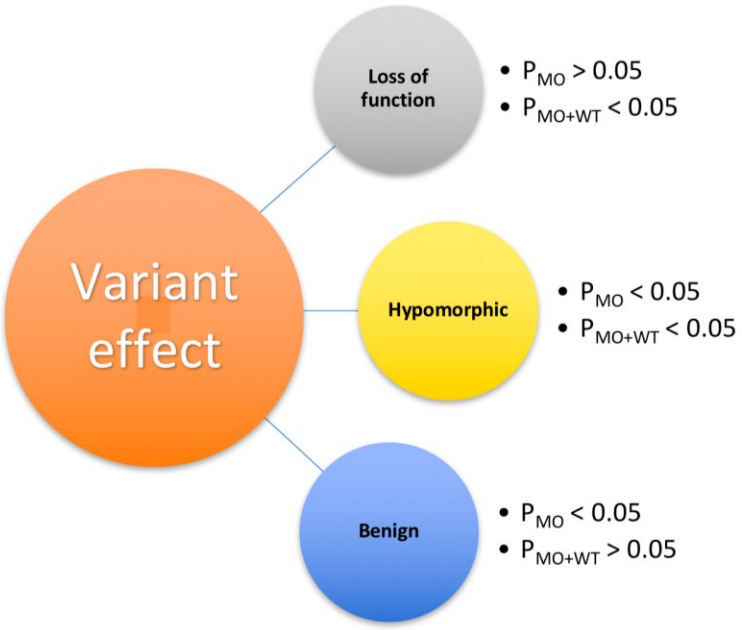


Figure 15. PCM1 variants' classification. For each variant the effect is determined by the t-test p-value between the distribution of brain volumes of the positive control (P_{MO}), individuals that have PCM1 gene inactivated, and the distribution of brain volumes of subjects that express the PCM1 gene with one variant. The other p-value (P_{MO+WT}) is obtained by the same procedure previously described, but considering the distribution of brain volumes in the negative control, expressing wild type (normal) PCM1 gene, in place of the positive control.

Introduction

Next generation sequence techniques produce new gene and genome sequences every day, providing lots of genetic information that is still unanalyzed (Niroula and Vihinen, 2016). Furthermore, genetic analysis is performed more frequently to study human diseases and consequently thousands of variants of unknown significance (VUS) appear. The scientific community has been making a big effort in developing computational tools that allow a better interpretation of VUS and genomic information. However, there is still plenty of work which has to be done to improve the current state of the art. Critical Assessment of Genome Interpretation (CAGI) experiment has been running since 2010 with the aim to assess the state of the art of computational methods which try to predict the phenotypic impact of genomic variations.

Here, we present the assessment of the CAGI-5 Pericentriolar Material 1 (PCM1) challenge. Predictors were asked to predict the pathogenicity of 38 transgenic human missense mutations in the PCM1 gene. The PCM1 gene maps to the human chromosome 8p22. The protein encoded by this gene is localized on centriolar satellites and has an important role in the radial organization of microtubules and the recruitment of proteins to the centrosome (Dammermann and Merdes, 2002; Villumsen et al., 2013). PCM1 is recruited to the centrosome to form a complex with the Bardet-Biedl syndrome 4 (BBS4) and Disrupted in Schizophrenia-1 (DISC1) proteins (Ansley et al., 2003; Guo et al., 2006). Suppression of one of these proteins could lead to neuronal migration defects (Kamiya et al., 2008). PCM1 is a large protein of 2,024 amino acids without known crystal structures. Database annotations in UniProt (The UniProt Consortium, 2017) show several coiled coil regions, while MobiDB (Piovesan et al., 2018) predicts regions of intrinsic disorder accounting for about 40% of the sequence. Linkage analysis has shown that the PCM1 gene has a role in susceptibility to schizophrenia in humans and is associated with orbitofrontal gray matter volumetric deficits (Gurling et al., 2006). Indeed, a candidate pathogenic mutation on this gene has been reported in an affected family (Kamiya et al., 2008). The effects of PCM1 haploinsufficiency have been studied on model animals, whereas affected mice show a significant reduction in brain volume and behavioral alterations (Zoubovsky et al., 2015). In addition to being risk factors for schizophrenia, several studies have also implicated some PCM1 component in genetic susceptibility to cancers and other mental diseases (Kamiya et al., 2008; Zoubovsky et al., 2015).

Ventricular enlargement is one of the most consistent abnormal structural brain findings in schizophrenia. A set of 38 transgenic human PCM1 missense mutations implicated in schizophrenia were assayed in a zebrafish model to determine their impact on the posterior ventricle area. The CAGI challenge aims to predict whether variants implicated in schizophrenia impact zebrafish brain development determining a reduction in the ventricular area of the brain. In particular, in addition to classifying benign variants, predictors have to distinguish between loss of function and hypomorphic variants. This challenge presents new difficulties for current state of the art predictors using different strategies to predict variant effects, while the variability

of results suggests that we are far from a general pathogenicity predictor, some groups have promising results in this challenge.

Materials and methods

Experimental data

The Katsanis lab assessed 38 PCM1 missense mutations in a zebrafish model. The native zebrafish embryo PCM1 protein was suppressed by injecting morpholino (MO) antisense oligonucleotides to inhibit translation of mRNA of the PCM1 gene. MOs are stable molecules consisting of a large, non ribose morpholine backbone with four DNA bases pairing stably with mRNA at either the translation start site (to disrupt protein synthesis) or at intron-exon boundaries (to disrupt mRNA splicing) (Summerton and Weller, 1997). Morpholinos have been shown to bind and block translation of mRNA *in vitro*, in tissue culture cells, and *in vivo* (Davis et al., 2014). Embryos deficient in PCM1 function show an absence of brain ventricle formation.

For each mutation, the Katsanis lab injected a group of embryos with MO and the mRNA of the human gene carrying the mutation (MO+VAR). Brain ventricle formation of the group of (MO+VAR) animals was compared to brain ventricle formation measured in a group of animals with MO alone and a group with MO+WT. The ventricle space is filled with a fluorescent dye and imaged by brightfield and fluorescence microscopy to access the effect on mutations on ventricle size (Gutzman and Sive, 2009; Niederriter et al., 2013). Each image was processed with an automated image processing tool to quantify the ventricle structure volume (Mikut et al., 2013; Näslund and Johnsson, 2016). P-values for statistically significantly different brain ventricle volumes between pairs of conditions (Lowery et al., 2009) were obtained using Student's t-test with a confidence level of 95%. The functional effect of each variant was then assigned as follows. When the p-value for (MO+VAR) is not significantly different from MO (p-value > 0.05), but significantly different from MO+WT (p-value < 0.05), the variant is pathogenic or loss of function. If the p-value (MO+VAR) is significantly different from MO, but not from MO+WT, the variant is benign. When

the p-value for (MO+VAR) is significantly different from MO, and also significantly different from MO+WT, the variant is hypomorphic or partial loss of function.

The experiment was performed in duplicate, blind to injection and the experimental data provided by the Katsanis lab is shown in Table 9. The dataset is composed of 16 benign variants (negative controls), 10 hypomorphic, and 12 loss of function variants. In percentages, 42% of the variants are benign and 58% have some functional effect (~32% loss of function and ~26% hypomorphic).

Nucleotide variant	Protein variant	p-value from MO	p-value from MO+WT	Functional effect (class)	Functional effect (description)
G17A	G6D	0.067	0.0001	2	loss of function
G69C	E23D	0.0004	0.0007	1	hypomorph
A229C	T77A	0.57	0.0001	2	loss of function
C436G	M146V	0.0001	0.13	0	benign
C467T	A156V	0.0001	0.0099	1	hypomorph
T599C	M200T	0.28	0.0001	2	loss of function
G600A	M200I	0.0022	0.0049	1	hypomorph
A641G	D214G	0.0005	0.0013	1	hypomorph
G742C	E248Q	0.53	0.0001	2	loss of function
G931C	E311Q	0.0012	0.0036	1	hypomorph
A1106T	E369G	0.059	0.0001	2	loss of function

C1168T	P390S	0.38	0.0001	2	loss of function
C1414G	L472V	0.039	0.0003	1	hypomorph
G1445T	G482V	0.0002	0.0012	1	hypomorph
G1627A	E543K	0.0001	0.64	0	benign
A1721G	D574G	0.0044	0.0021	1	hypomorph
G1811T	R604L	0.0001	0.55	0	benign
G1870A	E624K	0.0001	0.58	0	benign
C1977G	I659M	0.0001	0.62	0	benign
A2410C	S804R	0.0001	0.69	0	benign
G2498C	R833T	0.0001	0.71	0	benign
T2626C	C876R	0.0033	0.59	0	benign
G2674A	G892W	0.16	0.0007	2	loss of function
A2750G	E917G	0.19	0.0001	2	loss of function
G2862C	K954N	0.0001	0.92	0	benign
A3374G	N1125S	0.0001	0.11	0	benign
A3823G	K1275E	0.0001	0.32	0	benign
A4055T	H1352Y	0.012	0.0045	1	hypomorph
G4082T	C1361Y	0.0003	0.61	0	benign

C4469G	A1490G	0.0001	0.55	0	benign
G4603A	E1535K	0.0001	0.59	0	benign
C4658G	A1553G	0.0015	0.0034	1	hypomorph
G4667A	G1556D	0.36	0.0001	2	loss of function
A5583C	K1861N	0.0001	0.13	0	benign
T5625G	N1875K	0.087	0.0001	2	loss of function
G5720A	R1907H	0.0001	0.12	0	benign
C5738T	P1913L	0.75	0.0027	2	loss of function
G5935T	A1979S	0.72	0.0027	2	loss of function

Table 9: PCM1 experimental data. Variant nomenclature refers to PCM1 mRNA (GenBank identifier: NM_001315507). Each variant is associated with the corresponding p-values in the two evaluated experimental conditions (MO and MO+WT) and the resulting functional effect. Loss of function and hypermorphic variantes were evaluated together as a single category.

Dataset and classifications

The challenge presents 38 transgenic human PCM1 missense mutations implicated in schizophrenia (Experimental data URL: https://genomeinterpretation.org/content/PCM1). These variants were assayed in a zebrafish model to determine their impact on the posterior ventricle area as previously explained. Each variant codes for a single amino acid substitution, showing no insertions or deletions. The variant number used in this work refers to PCM1 mRNA (GenBank identifier: NM_001315507). Participants were asked to predict the probability (p-value) of the effect of the variants on zebrafish brain development.

These p-values were predicted considering the two different case scenarios: the probability that the variant (MO+VAR) is significantly different from MO and the probability that the variant is significantly different from MO+WT. In addition, predictors were also allowed to specify the standard deviation (SD) which defines the confidence of each prediction. Large SD means low confidence, while small SD means that the predictor is confident about the submitted prediction. According to the predicted probabilities and their interpretation, the participants had to inform the functional effect of the variant which could be: pathogenic, hypomorphic or benign. Six out of seven submissions reported for all the variants the p-values, SD and functional effect.

Performance assessment

The performance evaluation of bioinformatics tools aiming to predict VUS is a non-trivial problem, as the assessment should be more than a discrimination between good and bad predictions. In this challenge participants were requested to predict the p-values associated to each variant under two different conditions. According to the data provider results, the functional effects of each variant could be: benign, pathogenic (loss of function) and hypomorphic (partial loss of function). Even though one part of the challenge was to predict the p-values relative to the changes from MO and MO+WT, it was a very difficult task to begin with. After analyzing the correlation between experimental and predicted p-values in the two experimental conditions, we found that Pearson correlation coefficients range between -0.29 and 0.23 for different submissions, showing that there is no relationship between experimental and predicted p-values. Predicted p-values were therefore not taken into account to perform the assessment and consequently the use of global evaluation metrics as ROC or precision-recall curves was not possible. This is why we only used the predicted variant effect informed by the authors to address the final ranking.

To assess further the prediction reliability in a medical setting, a binary classification was used based on the variant predicted effects. The three variant effects mentioned above were reorganized as a binary classification, benign and pathogenic (loss of function and hypomorphic were considered together). A set of measures were implemented in order to perform a thorough assessment and to obtain a better

description about predictor performance (Vihinen, 2012). The aim was to produce a global overview of the strengths and weaknesses of each method. For each submission we calculate five different scores to assess the quality of the binary prediction: Balanced Accuracy (BACC), Matthews Correlation Coefficient (MCC), F1 score (F1), True Positive Rate (TPR) and True Negative Rate (TNR). The final ranking of predictor performances was the average of the individual rankings produced by each measure. To assess the statistical significance of each performance index, we generated 10,000 random predictions and used these data to estimate an empirical continuous score probability distribution (s). The p-value is then calculated by defining the proportion of random predictions scoring greater than s.
The R scripts used to perform the assessment are publicly available from the GitHub repository at URL: https://github.com/BioComputingUP/CAGI-PCM1-assessment.

Groups description

This challenge received 7 submissions from 6 different groups which were assessed blindly. Only one group (Bromberg lab) contributed with two submissions. Group 3 submitted an empty template and method description and consequently was not considered in the assessment. After completing the assessment, all groups provided their name and affiliations. Table 10 lists the participating groups, ID, name, and method used. Group 1 (Casadio lab) based their predictions on the Disease Index matrix (Casadio et al., 2011), which measures how protein stability is affected by mutations. Group 2 (Lichtarge lab) uses their Evolutionary Action approach (Katsonis and Lichtarge, 2014) to relate the variant effect with the evolutionary fitness effect. Group 4 (Bromberg lab) performed the predictions for their first submission using SNAP (Bromberg and Rost, 2007; Bromberg et al., 2008), a neural network-based method for the prediction of the functional effects of non-synonymous SNPs. In their second submission, predictions were depending on fuNTRp (Miller et al., 2019a), a Random Forest-based method to classify protein positions based on the expected range of possible mutational impacts per position (Neutral positions = no or weak effects; Rheostat positions = range of effects, i.e. functional tuning; Toggle positions

= mostly strong effects). Group 5 (Carter lab) analyzed each variant with VEST (Carter et al., 2013), assigning to each mutation a score indicating confidence in a functional mutation. Group 6 (BioFolD unit) used the SNPs&GO (Capriotti et al., 2013) and PhD-SNP (Capriotti et al., 2006) methods..

Submission ID	Group ID	Prediction features
Submission 1.1	Group 1 (Casadio lab)	Protein stability
Submission 2.1	Group 2 (Lichtarge lab)	Evolutionary action
Submission 3.1	Group 3	No predictions made
Submission 4.1	Group 4 (Bromberg lab)	Conservation, annotation
Submission 4.2	Group 4 (Bromberg lab)	Conservation, annotation
Submission 5.1	Group 5 (Carter lab)	Annotation
Submission 6.1	Group 6 (BioFolD unit)	Metaprediction

Table 10. Predictions overview. Each submission is associated to the predictor group and a summary of the features used for the prediction.

Results

Participation and similarity between predictions

In the PCM1 CAGI-5 challenge, participants were requested to predict the probability of the effect caused by 38 variants on zebrafish brain development. Essentially, the predicted probability allowed to infer three kinds of functional effects associated to each variant: benign, hypomorphic (partial loss of function), and loss of function. We performed a correlation analysis between submissions to address the similarity. Then we divided the predictions in two subsets: variants predicted as loss of function and

predicted as hypomorphic. Figure 16 shows the two predictions submitted by Bromberg lab obtained the same probability values for each variant. Both predictions used SNAP (Bromberg and Rost, 2007; Bromberg et al., 2008) to predict the p-values but differed in the way the variant is classified. Their submission 2 used fuNTRp (Miller et al., 2019a), a tool based on random forest that predicts position types (*i.e.* expected range of variant effects per position). Another observation from this analysis is that most groups predicted very different p-values, highlighting difficult of this challenge. We can also observe some weak positive and negative correlations between groups. On one hand we have a weak positive correlation between groups 4 and 6, possibly because predicted p-values are quite similar in some variants. Groups 2 and 5 also show a positive weak correlation possibly because predicted p-values in both groups are close to zero. On the other hand, we have some weak negative correlations between groups which have predicted opposite probability values for some variants, such as groups 2 and 5 versus groups 4 and 6.

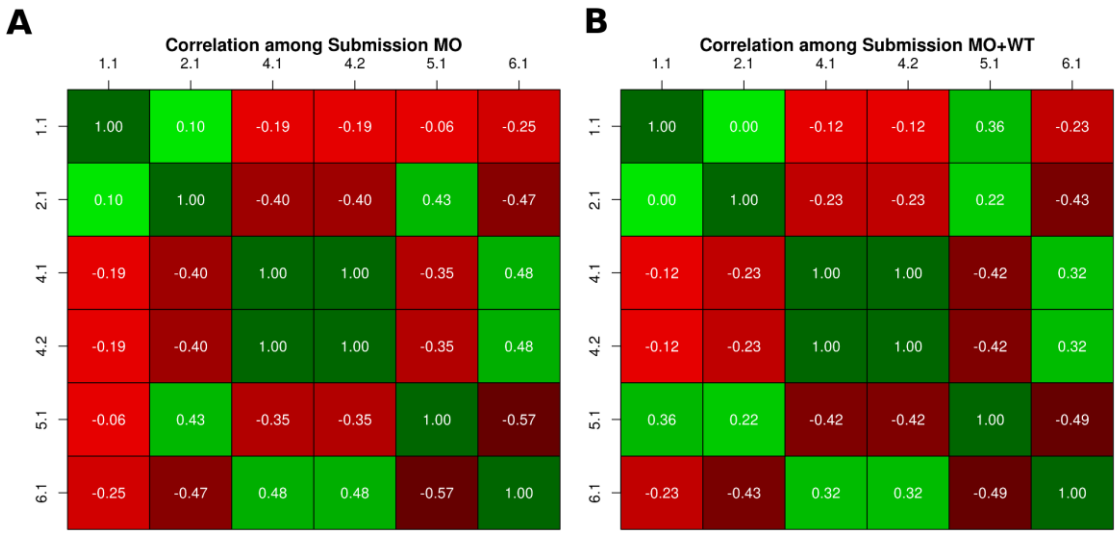


Figure 16. Similarity between predicted p-values. A-B) Each cell shows the Pearson correlation coefficient between two submissions, with a color scale ranging from green (+1, perfect correlation) to red (0, no correlation) and black (−1, perfect anti-correlation).

Assessment criteria and performance evaluation

The evaluation criteria used to assess a CAGI challenge directly influence perceptions gained from the test. In order to highlight predictor performance and their practical relevance, we performed the evaluation only considering the predicted functional effect of each variant provided by the participants. As most submissions reported the predicted p-values, we tried first to perform the assessment as an inherently continuous prediction challenge. After some exploratory analysis, we concluded that predicted p-values among all submissions did not agree at all with the experimental p-values and also with the interpretation of the p-values to infer the functional classes (Figure 17 and Figure 18). For this reason, we decided to perform the assessment using only the predicted functional class of each mutation.

The performance was evaluated using five standard measures as described above. Our assessment shows that the six submissions achieved in general a poor performance. This is highlighted by the MCC values (Figure 19), where most of submissions have values close to or below zero. As the average among all submissions is -0.06, this means that the correlation between the experimental and predicted variant functional effect is no better than random predictions in most of the cases. The highest MCC value is 0.35 and was reached by submission 4.1 (Bromberg lab). This submission correctly predicted 10 out of 22 pathogenic variants and 14 out of 16 benign variants (Table 11). Then, submission 5.1 (Carter lab) obtained the lowest MCC value (-0.35), correctly predicting 12 disease mutations but only 2 benign (Table 11).

For BACC, we can observe that submissions 4.1 (Bromberg lab) and 6.1 (BioFolD), performed better than other methods, also considering their MCC values (Figure 19). Since a method could be biased to predict the more frequent class, BACC is a good way to calculate the accuracy evaluating if the predictor takes advantage or not of an imbalanced test set. Consequently, F1 shows values higher than 0.50 for three out seven submissions. F1 measure considers the precision and recall of the test, submission 1.1 (Casadio lab) obtained the highest F1 value of 0.67, followed by submissions 4.1 and 6.1 with 0.59 and 0.53 respectively. However, if we observe the

TNR and confusion matrix of submission 1.1 (Table 11), this predictor presents a biased confusion matrix and was not able to identify any benign variants.

To perform a global assessment of each predictor performance we need to take into account all performance measures together instead of just comparing them separately. We decided to observe the ranking achieved for each submission on each considered measure. Moreover, this allows non-expert users to better understand the results of the assessment. The Bromberg lab (Submission 4.1) achieved the best overall performance comparing with all other predictors, ranking first in BACC and MCC measures, second in F1 and TNR and sharing the third place in TNR (Table 12). BioFolD (submission 6.1) ranked second in overall performance, second in BACC and MCC, and third in the other measures. The Casadio lab achieved the best rank in F1 and TNR measures and ranking third in overall performance. However, their prediction was biased toward diseases phenotypes, with no benign variant correctly predicted (Table 11). Something similar but opposite happened with the Bromberg lab (Submission 4.2), where the prediction was biased towards benign variants and only one disease variant predicted correctly (Table 11). In addition, we can observe that MCC values for the two submissions mentioned above are negative (i.e. negatively correlated) and almost zero (i.e. close to random). Observing the confusion matrices, we can conclude that most submissions produced unbalanced predictions biased towards the prediction of disease phenotypes.

Considering the poor performance of most predictors, we only calculated the statistical significance of submission 4.1 (Bromberg lab) for the BACC, MCC and F1 measures. A bootstrap with 10,000 replicas was used to test whether the performance of submission 4.1 could be achieved by chance. We can conclude that it performs better than random (p-value < 0.05) for MCC and BACC measures (Suppl. Figure 5, Appendix 3). The only exception is F1, denoting unbalanced predicted classes from the real data.

Another interesting aspect of this challenge is to see how each group correctly predicted the real disease effect, loss of function and hypomorph. In Suppl. Table S6 we can see the contingency matrices split into three categories. Most of the groups had difficulties identifying the correct disease class. Submission 4.1 correctly

predicted 4 hypomorph variants and no loss of function one. Submission 6.1 correctly identified one loss of function and one hypomorph variant. On the other hand, submission 1.1, which was biased to predict disease variants, correctly predicted 6 hypomorph and 4 loss of function.

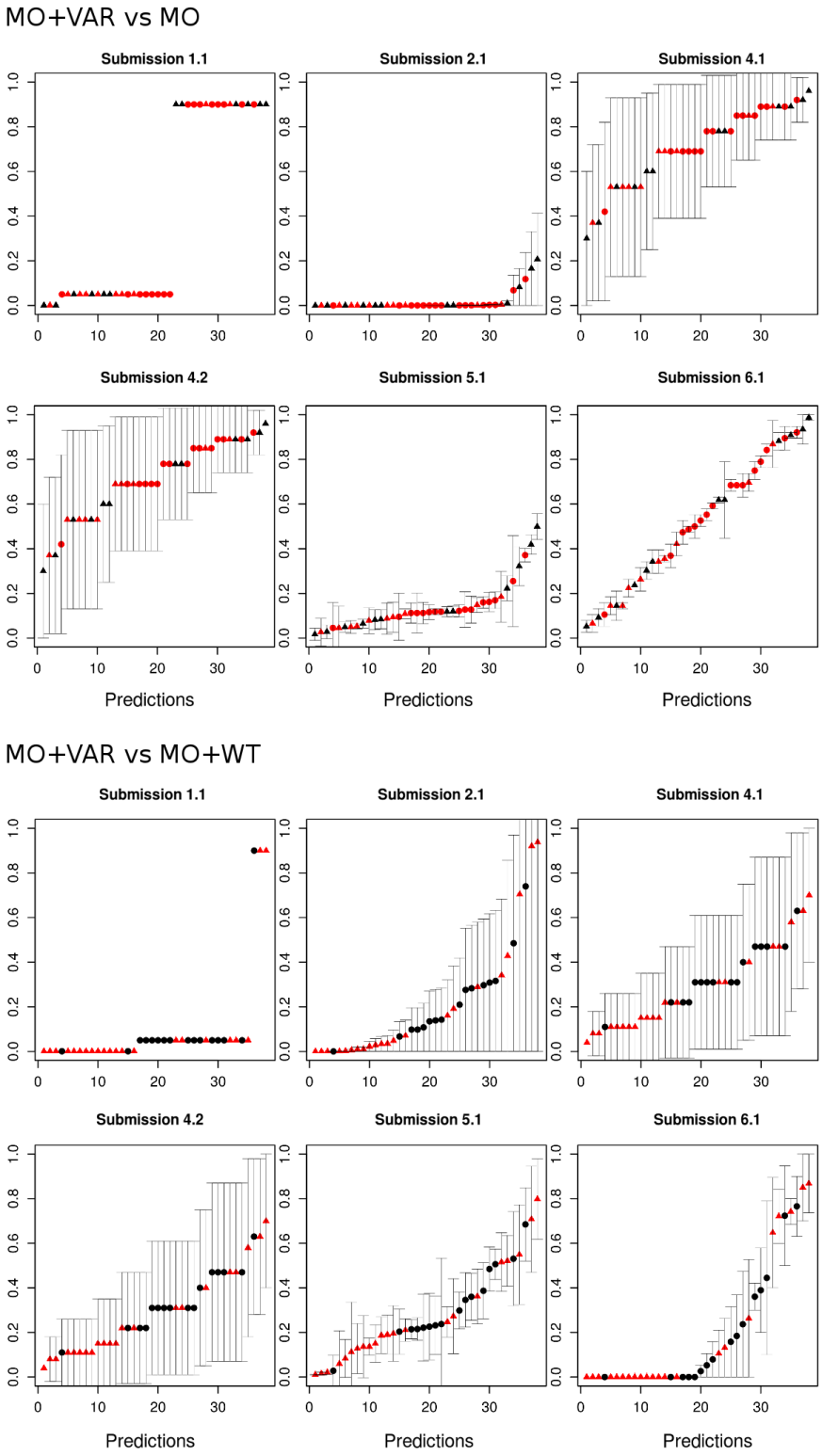


Figure 17. Predicted p-values with their corresponding standard deviation for each experimental condition by group. The x-axis is from 1 to 38 and represents

the predicted p-values for a particular position (sequentially ordered by the position on the sequence). The y-axis is the value of the predicted p-value. Dot shapes represent the variant effect, with triangles for pathogenic and circles for benign. The color indicates the experimental p-value, red for p-value < 0.05 and black for p-value ≥ 0.05.

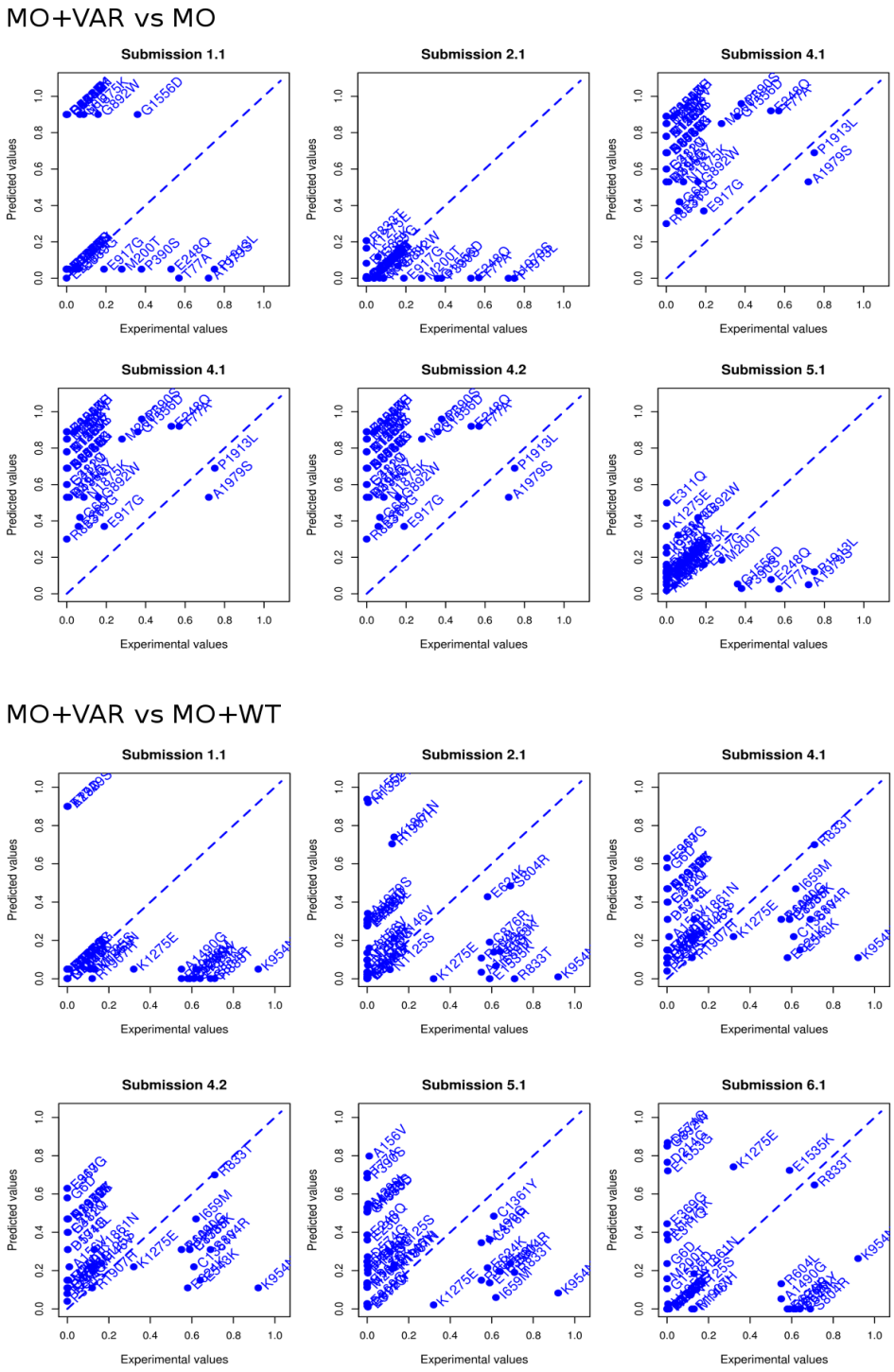


Figure 18. Predicted vs. experimental p-values for all submissions. The predicted value (y-axis) is plotted against the experimental value (x-axis) for all variants (in the two experimental conditions) in each of the 6 submissions.

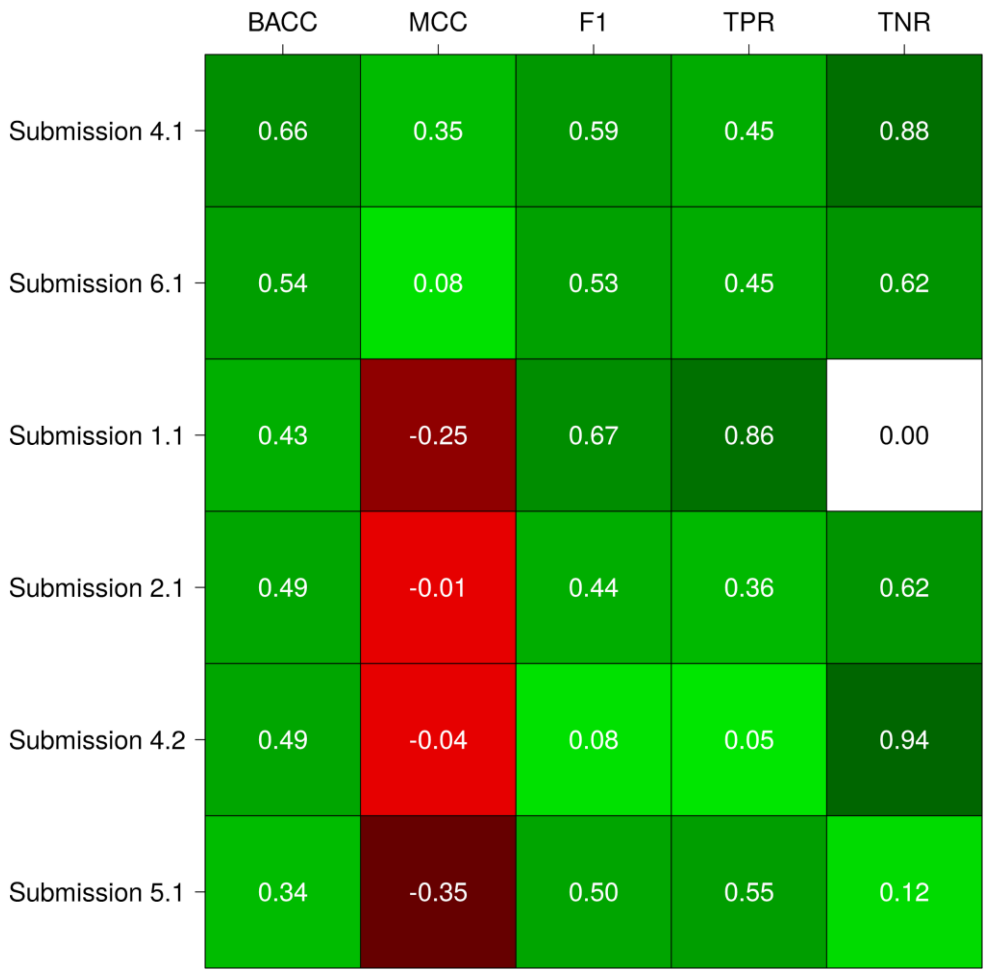


Figure 19. Submissions performance evaluation. Each cell represents the value of a measure for a specific submission. The color scale ranges from dark green (+1, perfect performance) to red (-1, perfect anticorrelation just for MCC). White means zero in terms of performance.

Submission 1.1			Submission 2.1			Submission 4.1		
	Obs. Disease	Obs. Benign		Obs. Disease	Obs. Benign		Obs. Disease	Obs. Benign
Pred. Disease	19	16	Pred. Disease	8	6	Pred. Disease	10	2
Pred. Benign	3	0	Pred. Benign	14	10	Pred. Benign	12	14
Submission 4.2			**Submission 5.1**			**Submission 6.1**		
	Obs. Disease	Obs. Benign		Obs. Disease	Obs. Benign		Obs. Disease	Obs. Benign
Pred. Disease	1	1	Pred. Disease	12	14	Pred. Disease	10	6
Pred. Benign	21	15	Pred. Benign	10	2	Pred. Benign	12	10

Table 11. Confusion matrices for all submissions.

Submission ID	BACC	MCC	F1	TPR	TNR	Avg. Ranking	Final rank
Submission 4.1	**1 (0.67)**	**1 (0.35)**	2 (0.59)	3.5 (0.46)	2 (0.88)	**1.9**	**1**
Submission 6.1	2 (0.54)	2 (0.08)	3 (0.53)	3.5 (0.46)	3.5 (0.63)	2.8	2
Submission 1.1	5 (0.43)	5 (-0.25)	**1 (0.67)**	**1 (0.86)**	6 (0)	3.6	3
Submission 2.1	3 (0.49)	3 (-0.01)	5 (0.44)	5 (0.36)	3.5 (0.63)	3.9	4
Submission 4.2	4 (0.49)	4 (-0.04)	6 (0.08)	6 (0.05)	**1 (0.94)**	4.2	5
Submission 5.1	6 (0.34)	6 (-0.35)	4 (0.5)	2 (0.55)	5 (0.13)	4.6	6

Table 12. Submissions ranking. Individual and overall rankings among all submissions based on the performance measures considered. Each cell contains the ranking of a submission for a specific performance measure and in brackets the performance value. The overall final ranking is obtained by the average rank achieved for each submission considering all the performance measures.

Difficult variants

Looking at the predicted functional effects for each variant, we can see that some were particularly complex to be predicted (Figure 20). Due to the limited structural characterization of PCM1 is difficult to analyze the structural properties of each residue. We tried to explore further some properties of PCM1 using FELLS (Piovesan et al., 2017b). Disease variant p.G892W was correctly predicted by all submissions and that position presents high propensity to be coil and disordered. On the other

hand, disease variant p.E23D was not identified by any predictor and shows high propensity to be disordered.

There are 15 variants where most groups failed to correctly predict their effect (<50% correctly predicted): 4 benign, 5 hypomorphs (disease) and 6 loss-of-function (disease). Interestingly, submission 1.1 (Casadio lab) predicted correctly five of these 15 disease mutations. However, this predictor was biased towards pathogenic variants and not able to identify any benign. The PCM1 challenge highlights how some variants are really hard targets for most of the methods.

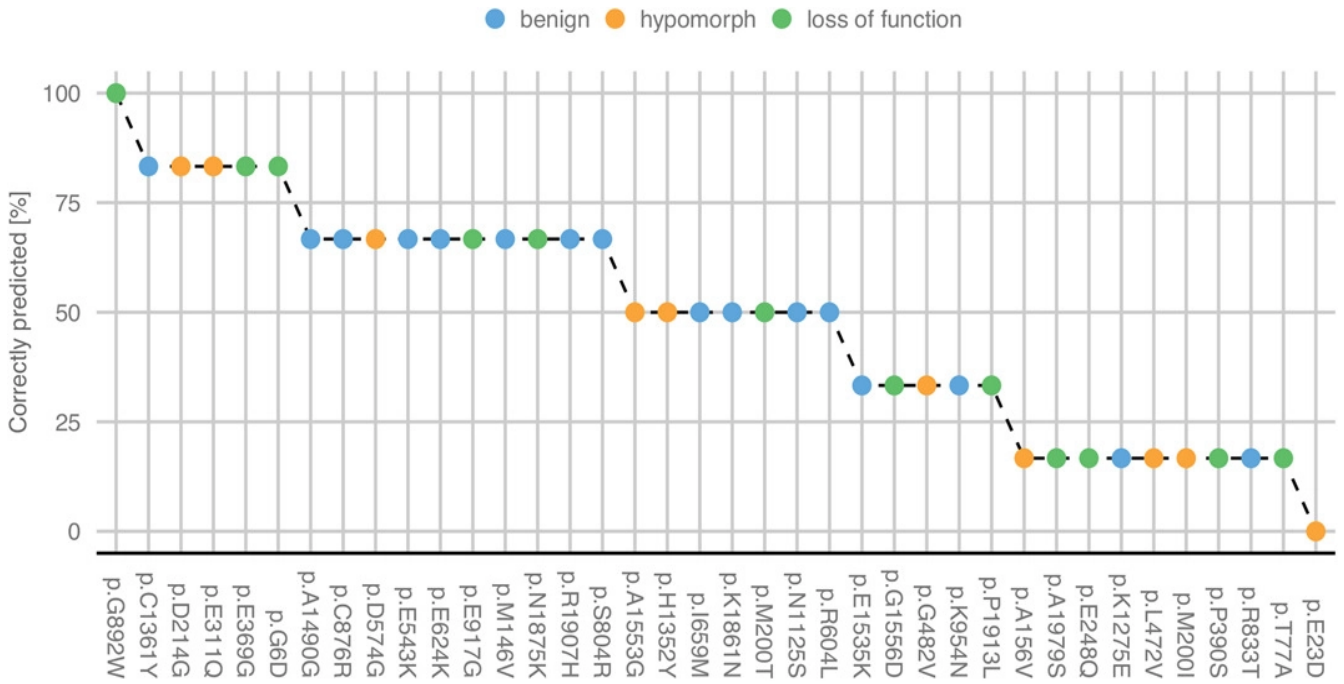


Figure 20: Percentage of groups which correctly predicted the effect of each variant. The variants are colored by their effect.

Discussion

The determination of novel variant effects is a key challenge of great value for clinicians. Due to the diversity and complexity of the biological systems, a variant could impact at different levels such as protein function, subcellular localization, metabolic pathways, among others (Hamp and Rost, 2012). The best predictor should be able to discriminate between pathogenic and benign variants. Here, we presented the assessment of the CAGI-5 PCM1 challenge. This challenge is based on the prediction of the probability of missense variant effects, in analogy to the CAGI-3 p16 challenge (Carraro et al., 2017). While the p16 challenge was testing the ability to predict cell proliferation rate, the PCM1 challenge is focused on predicting the probable variant effect on zebrafish brain development. PCM1 is a component of centriolar satellites occurring around centrosomes in vertebrate cells (Dammermann and Merdes, 2002; Kubo and Tsukita, 2003). It also interacts with BBS4 and DISC1 (Kamiya et al., 2008; Miyoshi et al., 2004) and has an important role in centrosome formation, which is needed for proper neurodevelopment (Ayala et al., 2007; Gupta et al., 2002; Mochida and Walsh, 2004; Solecki et al., 2006; Tsai and Gleeson, 2005) . The Katsanis lab provided experimental data for 38 missense mutations in PCM1 in a zebrafish model. The experimental effect determined by the data providers is unambiguous and resulted of brain brain ventricle volumes between MO and MO+WT. This kind of comparison studies have been performed in the past and the specificity/sensitivity metrics have been reported to be high (Zaghloul et al., 2010).

Submissions were compared with experimental data to evaluate their prediction performance. Using a set of performance measures highlighting strengths and weaknesses of each predictor similar to previous CAGI assessments (Carraro et al., 2017).

From a technical point of view, the groups used different approaches to predict p-values and variant effect, ranging from machine learning to position-specific scoring matrices. The assessment suggests that most state-of-art predictors participating in this challenge were not sufficient to perform reliable variant effect predictions. The absence of structural information and high disorder content make this protein challenging, especially for predictors based on structural information. The MCC values reached by different

submissions are subpar, close to random prediction. MCC is one of the best measures to handle unbalanced data, since some predictions were biased to identify disease or benign phenotypes (Boughorbel et al., 2017). The best MCC and BACC values were reached by submission 4.1 (Bromberg lab, Table 12), showing also the best overall ranking. They correctly predicted 10 out of 22 disease variants and 14 out of 16 benign variants (Table 9). However, if we look at the disease class considering loss of function and hypomorph, submission 4.1 correctly predicted only 4 hypomorph variants. Showing again the difficulty in p-values interpretation (Suppl. Table S6). Anyway, these results suggest that SNAP (Bromberg and Rost, 2007; Bromberg et al., 2008), a neural network-based method, may be a useful method to screen big datasets for pathogenic variants in a similar context.

Interestingly, group 1 (Casadio Lab) obtained a TPR of 0.86 and predicted 19 out 22 disease variants but they could not identify any benign variants. Nevertheless, they identified the highest number of loss of function variants (Suppl. Table S6). Conversely, group 4 submission 2 reached a high TNR score and predicted 15 out 16 benign variants but identified only one disease variant. Group 6 (BioFolD unit) well predicted 10 out 16 benign variants and 10 out 22 disease, scoring second considering the overall rank and MCC value. We should emphasize here that data imbalance frequently occurs in biomedical applications and the use of inadequate performance metrics could lead to misinterpretation of predictors performance (Boughorbel et al., 2017).

This CAGI-5 PCM1 challenge evidences that there is still plenty of work to improve the pathogenicity prediction of VUS. Despite the generally low performance of predictors, some identified a good number of disease and benign variants. However, we still have to improve our prediction methods if we want a generic pathogenicity predictor. We expect that the CAGI challenges which help motivate research, improving the current methods and generating new ideas.

Chapter 4

In silico prediction of blood cholesterol levels from genotype data

In this chapter the development and assessment of an in silico model for blood cholesterol levels from genotype data will be exposed. This work was part of a project founded by the Italian Ministry of Health: “Towards the prediction of dyslipidemia from next generation sequencing data” (grant GR-2011-02346845). Dyslipidemia is a metabolic disorder characterized by abnormal level of lipids in the blood, influenced by genetic and lifestyle factors (Ramasamy, 2016). The aim of the project was to develop a personalized medicine pipeline for dyslipidemia early detection and treatment. I contributed to the bioinformatic part of the project, in particular, the development of a predictor for blood cholesterol levels from genotype data. The in silico cholesterol model I have developed could be used to simulate the effects of damaging mutations on an affected patient and predict blood cholesterol levels. This could be considered as a tentative personalized medicine approach in which patient predisposition to dyslipidemia is genetically determined before the physiological development of the disease.

This chapter is based on a revised version of “In silico prediction of blood cholesterol levels from genotype data. *Biorxiv* (2019). doi:10.1101/503003”, submitted to a peer reviewed journal.

Introduction

Recent exome-wide association studies (Liu et al., 2017) started to shed light on the complex genomic architecture behind the regulation of blood cholesterol levels in humans. Reliable tools to predict human cholesterol levels from genotype are not available yet. The huge number of genes involved in the regulation of this trait and the complex interaction with environmental factors as diet, gender and age make modelling

cholesterol levels a difficult task. However, particular situations exist where a single mutation is related to significant variations of cholesterol levels. Example are damaging mutations on genes involved in hepatic uptake of Low Density Lipoprotein (LDL), as the Low Density Lipoprotein Receptor (LDLR) gene, causing familial hypercholesterolemia characterized by elevated levels of LDL and total plasma cholesterol but with normal concentrations of triglycerides (2012). Other processes involved in cholesterol metabolism are affected by genetic mutations, with a wide range of phenotypes depending on the gene involved, like marked High Density Lipoprotein (HDL) cholesterol levels deficiency as seen in patients affected by Tangier disease (Shapiro, 2000). The aim of this work is to test the reliability of a modelling approach aimed to predict cholesterol levels relying on patient's genotype data only. Different tools have been developed for blood lipid levels prediction, some of them are regression methods based on a set of variables representing patient genotypes (e.g. presence or absence of SNPs associated to lipid traits) (Spiliopoulou et al., 2015) and phenotype (e.g. Body Mass Index, gender, age, etc.). These methods require a huge amount of data for training and test, with predictions having low correlation to lipid profiles (Ramos-Lopez et al., 2018). Other research groups have developed tools that are able to predict a familial hypercholesterolemia phenotype from LDLR missense mutations, but not the range of blood lipid values (Guo et al., 2019). A different strategy is to develop an *in silico* mathematical model, that represents human cholesterol metabolism, simulate the effect of a mutation and take the response of the model as predicted levels of cholesterol. Effective way to simulate *in silico* metabolism are dynamic models. In this kind of simulations, the development of the system in time is computed through a set of ordinary differential equations, able to simulate the variations of chemical species concentration. Several information are required for the development of these models: interactions between the chemical species involved in the biological process, kinetic parameters associated to chemical reactions occurring in the system and its initial state. The simulation of a biological perturbation could be obtained by modifying model parameters (e.g. decreasing kinetic rates) and observing variations occurring in the system (Cazzaniga et al., 2014). Several *in silico* models simulating cholesterol metabolism have been proposed so far, both for human and animal models (Paalvast et al., 2015). A recent

review (Paalvast et al., 2015) has described a set of published mathematical models, based on differential equations, which simulate cholesterol metabolism at different levels. Some of the presented methods were focused on specific reactions, as endocytosis or excretion of lipoproteins by hepatocytes, other attempted to model cholesterol metabolism at a whole body scale. One of these models, published in literature by van de Pas and colleagues in 2012 (van de Pas et al., 2012), was developed on the basis of genes and related metabolic reactions that have a relevant role on the control of human cholesterol homeostasis. In this work we decided to adopt an algorithm based on this mathematical model to predict cholesterol levels. This choice was motivated by different factors, from one hand this method has passed a validation process both in the original publication (van de Pas et al., 2012) and in a following review by different authors (Paalvast et al., 2015) On the other hand this model is gene based and computes levels of LDL and HDL cholesterol induced by a mutation, making it suitable for the prediction of blood lipid levels from genotype data. This physiologically based kinetic model (van de Pas et al., 2012) is based on differential equations, computing the flow of cholesterol in different body organs. The whole process is regulated by a set of rates, each one related to a gene that has a key role in cholesterol metabolism. Simulation of mutations effects depends on reducing rates (f_{mut}) estimated from wet lab experiments. This kind of information is usually not easily accessible, strongly limiting the usability of the model. In this work we implemented and optimized the framework for blood cholesterol levels prediction making it able to perform reliable predictions when only patient's genotype data are available. The model has been improved through a training phase, in which reducing rates (f_{mut}) were estimated from phenotype data of patients affected by mutations on key regulatory genes of cholesterol metabolism. Assessment measures confirmed how the optimized model presents improved performance, reducing the error between experimental and predicted data, compared to the original version available in literature (van de Pas et al., 2012).

Materials and methods

In silico kinetic model for cholesterol levels prediction

An available *in silico* kinetic model (van de Pas et al., 2012) has been used as basis for predicting plasma cholesterol concentrations in humans. The kinetic model was developed to simulate cholesterol levels for a reference man of 70 kg. The model is composed of 8 pools, representing main sites of cholesterol storage in the human body (Figure 21). These pools can be grouped in 4 main entities corresponding to plasma, intestine, liver and periphery. Each cholesterol pool is modeled by a differential equation, composed by a set of rates moving cholesterol from or to a different one. These pools are connected by 21 kinetic rates, each one representing the main gene regulating that specific biochemical reaction (Table 13).

Rates depend on kinetic constants, organ volumes, body weight and pool cholesterol concentrations. In the original model, all parameters have been computed from data published in literature (van de Pas et al., 2012). The model was calibrated to immediately reach a steady state, a stable equilibrium in which each compartment has a constant cholesterol concentration in time. To simulate a mutation affecting the activity of a gene, a set of rate reduction parameters (f_{mut}), each one in the interval [0, 1], multiplies the standard rates to represent the effect of the mutated genes. These values were computed on the basis of experimental data available in literature. Example is the value of the f_{mut} related to mutations affecting the gene CYP7A1 involved in bile acid synthesis, where the rate reduction parameter was computed as the ratio of bile acids contents in the stools of patients carrying the mutation over controls (van de Pas et al., 2012).

These kind of perturbations force a re-tuning of the system, moving from the original steady state to a new one, where blood cholesterol profiles were comparable to the real values detected in patients affected by that particular mutation.

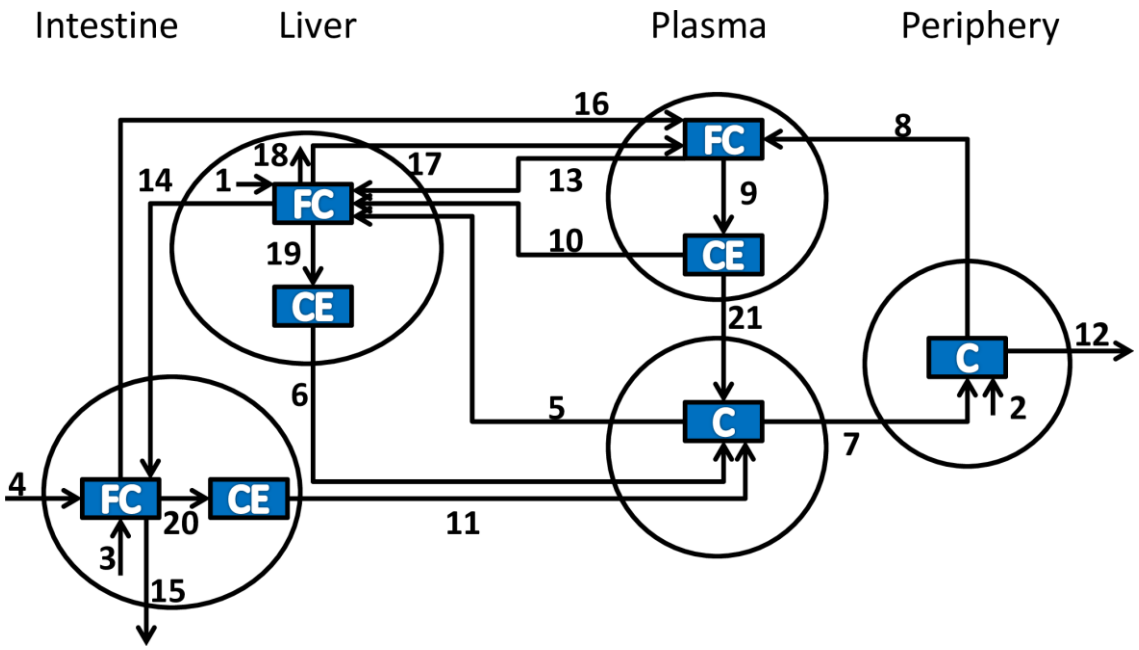


Fig 21. Conceptual model for pathways and genes determining cholesterol plasma levels used van de Pas and colleagues (van de Pas et al., 2010, 2012)**.** Process numbers stand for: 1, hepatic cholesterol synthesis (DHCR7); 2, peripheral cholesterol synthesis(DHCR7); 3, intestinal cholesterol synthesis (DHCR7); 4, dietary cholesterol intake (NPC1L1); 5, hepatic uptake of cholesterol from LDL (LDLR,APOB,APOE); 6, VLDL-C secretion (MTTP); 7, peripheral uptake of cholesterol from LDL (LDLR,APOB,APOE); 8, peripheral cholesterol transport to HDL (ABCA1); 9, HDL-associated cholesterol esterification (LCAT); 10, hepatic HDL-CE uptake (SCARB1); 11, intestinal chylomicron cholesterol secretion (MTTP); 12, peripheral cholesterol loss; 13, hepatic HDL-FC uptake (MTTP); 14, biliary cholesterol excretion (ABCG8,NPC1L1); 15, fecal cholesterol excretion; 16, intestinal cholesterol transport to HDL (ABCA1); 17, hepatic cholesterol transport to HDL (ABCA1); 18, hepatic cholesterol catabolism (CYP7A1); 19, hepatic cholesterol esterification (SOAT2); 20, intestinal cholesterol esterification (SOAT2); and 21, CE transfer from HDL to LDL (CETP).

rate	Biological process	Gene
1	hepatic cholesterol synthesis	DHCR7
2	peripheral cholesterol synthesis	DHCR7
3	intestinal cholesterol synthesis	DHCR7
4	dietary cholesterol intake	NPC1L1
5	hepatic uptake of cholesterol from LDL	LDLR, APOB, APOE
6	VLDL-C secretion	MTTP
7	peripheral uptake of cholesterol from LDL	LDLR, APOB, APOE
8	peripheral cholesterol transport to HDL	ABCA1
9	HDL-associated cholesterol esterification	LCAT
10	hepatic HDL-CE uptake	SCARB1
11	intestinal chylomicron cholesterol secretion	MTTP
12	peripheral cholesterol loss	
13	hepatic HDL-FC uptake	MTTP
14	biliary cholesterol excretion	ABCG8, NPC1L1
15	fecal cholesterol excretion	
16	intestinal cholesterol transport to HDL	ABCA1
17	hepatic cholesterol transport to HDL	ABCA1
18	hepatic cholesterol catabolism	CYP7A1
19	hepatic cholesterol esterification	SOAT2
20	intestinal cholesterol esterification	SOAT2
21	CE transfer from HDL to LDL	CETP

Table 13. Biological process and genes associated to each rate of the model. Reaction rates present in the model and the associated biological process they represent, also the main genes involved in the process are reported (van de Pas et al., 2010, 2011, 2012).

Model implementation

The algorithm of the available physiologically based kinetic model (Paalvast et al., 2015), was implemented in R language (R Core Team, 2015). The *deSolve* package (Soetaert et al., 2010) was used for solving differential equations.

New f_{mut} values have been obtained thanks to a training procedure exploiting a dataset composed of cholesterol levels and genotypes of mutated patients (Suppl Table S7, S8, Appendix 4). This operation required the usage of the Levenberg-Marquardt algorithm as implemented in the *Minpack.lm* package (Timur V. Elzhov, Katharine M. Mullen, Andrej-Nikolai Spiess and Ben Bolker, 2016).

The R scripts are publicly available from the GitHub repository at URL: https://github.com/BioComputingUP/Cholesterol-model

Training phase

To improve performance in predicting genetic mutations' effect on cholesterol levels, f_{mut} parameters, each one related to a particular gene mutation and rates of the model, have been trained on phenotype data of a dataset of patients, retrieved from literature. The Levenberg-Marquardt minimization method has been used to estimate the f_{mut} parameters able to minimize the difference between predicted and experimentally measured levels of HDL and LDL, divided by the control, intended as level of cholesterol of the model when no mutation is present.

$$\Delta HDL = \frac{HDL_{experimental}}{HDL_{control}} - \frac{HDL_{predicted}}{HDL_{control}} \quad (24)$$

$$\Delta LDL = \frac{LDL_{experimental}}{LDL_{control}} - \frac{LDL_{predicted}}{LDL_{control}} \quad (25)$$

Exceptions are patients affected by mutations on the DHCR7 genes where only total cholesterol (TC) levels were found in literature. In this case the difference between real and predicted total cholesterol rate was taken into account.

$$\Delta TC = \frac{TC_{experimental}}{TC_{control}} - \frac{TC_{predicted}}{TC_{control}} \quad (26)$$

The optimized f_{mut} parameters are reported in Table 14, each value represents the mean computed on all parameters estimated during the training phase, representing either heterozygous or homozygous mutations, that affect the same gene. The difference

between optimized f_{mut} parameters and the corresponding ones computed by van de Pas and coauthors is statistically significant for a 0.05 significance level (Welch's t-test p-value 0.027). The values of the first column (Table 14) are the result of a training procedure, which is based on a dataset of patients and regulated by the sensitivity of the rates involved. This is not true for the f_{mut} based on experimentally determined variables, since they were computed on the basis of specific molecule concentrations (e.g. the plasma level of CETP), rates (e.g. cholesterol efflux rate to APOA1), or enzyme activity (e.g. in vitro determined activity of LCAT) (van de Pas et al., 2012). The consequence of these two different strategies of parameter estimation is reflected by the difference between f_{mut} associated to the same gene in the two columns. Example is the CYP7A1 gene: the f_{mut} estimated by van de Pas and coauthors is equal to 0.05, as the value of bile acids in the stools of patients compared to controls (van de Pas et al., 2012). On the contrary, the corresponding value is higher after an optimization procedure, which has been influenced by the sensitivity of that rate and the cholesterol levels of the training set elements.

Gene	Reggiani et al f_{mut}	van de Pas et al 2012 f_{mut}
LDLR	0.58	0.38
APOB	0.9	0.31
APOB (hom.)	0.55	0.32
ABCA1	0.53	0.41
APOE	0.72	0.45
CETP	0.43	0.65
LCAT	0.48	0.62
LCAT (hom.)	1	0
DHCR7	0	0
CYP7A1	0.81	0.05

Table 14. Optimized f_{mut} parameters and related genes. Genes represented in the training, test set and related f_{mut} as computed by the optimization procedure or by using experimental variables, as reported by van de Pas and colleagues (van de Pas et al., 2012)

Training set

The training set is represented by a custom dataset of patients affected by single mutations (either in homozygous or heterozygous form), in one of the key genes regulating cholesterol metabolism (Figure 22). For each patient the levels of HDL, LDL or total cholesterol and the causative mutation were extracted from literature (Suppl Table S7, S8). Each gene is covered by a different number of individuals due to the relative abundance of works in literature (Table 15). Special cases are the CETP gene, where only information regarding the mean levels of blood cholesterol were found in literature and the DHCR7 gene, where only the levels of blood total cholesterol were found. The training set has been divided in two sections (Table 15). The first group is represented by hypercholesterolemic patients with mutations affecting a set of genes involved in the development of Autosomal Dominant Hypercholesterolemia (Marduel et al., 2013): LDLR,

APOB and APOE genes, represented by reaction 5 and 7 of the model (Table 13). The second part of the dataset is composed of patients with damaging mutations on 5 different genes: ABCA1, CETP, LCAT, DHCR7 and CYP7A1 (affected rates are shown in Table 13). Patients of the Autosomal Dominant Hypercholesterolemia dataset are characterized by high levels of LDL cholesterol, while the second part of the dataset is composed by different ranges of HDL and LDL, depending on the gene affected by the mutation (Figure 22).

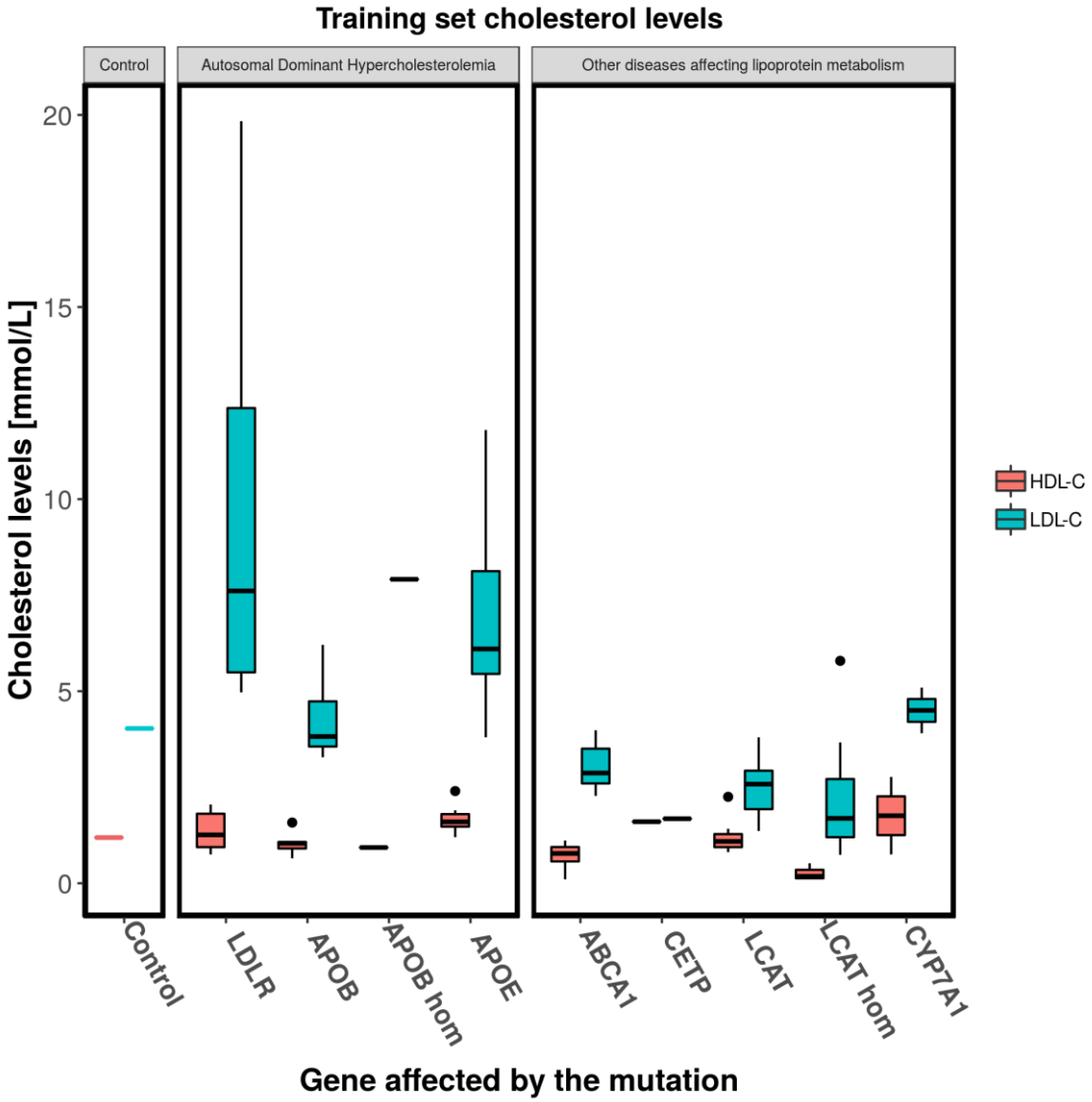


Fig 22. Training set patients cholesterol levels. Boxplot of HDL and LDL cholesterol levels of the patients composing the training set. From left to right: cholesterol levels of the model at the steady state, patients affected by Autosomal Dominant

Hypercholesterolemia (with high levels of LDL and low HDL) and patients affected by other disease altering lipoprotein metabolism.

Dataset	GENE	Patients	Mutations	type	rate
Autosomal Dominant Hypercholesterolemia	*LDLR*	13	9	heterozygous	5, 7
	APOB	7	1	heterozygous	5, 7
	APOB (hom)	1	1	homozygous	5, 7
	APOE	12	2	heterozygous	5, 7
Other disease altering lipoprotein metabolism	*ABCA1*	7	3	6 heterozygous, 1 compound heterozygous	8, 16, 17
	CETP	1	1	heterozygous	21
	LCAT	17	2	heterozygous	9
	LCAT (hom)	7	4	homozygous	9
	CYP7A1	2	1	heterozygous	18

Table 15. Training set composition. Disease, gene, number of patients with a mutation in that gene, number of different mutations, type of mutation (heterozygous, homozygous or compound heterozygous), rates representing that gene in the model

Test phase

Prediction performance was tested on a dataset retrieved from literature (van de Pas et al., 2012). The dataset is the same one used to test performance of the former version of the model. This test set has been used in order to highlight performance comparison between the versions of the algorithm. The effect of a genetic mutation was simulated for each individual of the dataset until a steady state was reached (fixed threshold: 1000 days). Predicted HDL, LDL and total cholesterol were compared to experimental data.

Test set

Test set is composed by patients affected by 10 mutations. All mutations affect genes present in the training set of this work. The first group of mutations maps on the LDLR, APOB and APOE genes, involved in hepatic cholesterol uptake. Patients affected by this kind of mutations have high levels of LDL and total cholesterol. Genetic mutations affecting the other genes of the dataset have different effects on lipid profiles. Mutations on the ABCA1 gene can cause marked HDL cholesterol levels deficiency as reported for different diseases like hypoalphaliproproteinemia or Tangier disease (Shapiro, 2000). CETP is a protein involved in the transport of cholesterol esters from HDL to LDL, deficiency of this protein can cause a marked increase of HDL levels (Shapiro, 2000). LCAT is a gene involved in cholesterol esterification in HDL particles, mutation on this gene can cause LCAT deficiency, characterized by low levels of HDL and LDL cholesterol (Shapiro, 2000). Patients with mutations in heterozygous or homozygous form has been included in the training set. DHCR7 gene is responsible for the last step of the cholesterol biosynthesis pathway. Reduced enzyme activity cause low levels of blood cholesterol, as reported in patients affected by the Smith-Lemli-Opitz syndrome (Yu et al., 2000). CYP7A1 gene is involved in cholesterol catabolism and bile acids synthesis, mutations affecting this gene cause an increase of total, hepatic cholesterol and a decrease in bile acids secretion (Pullinger et al., 2002).

Results and discussion

Performance assessment

The assessment approach used in this work was influenced by the methods used for the evaluation of tools predicting the effect of variants on continuous phenotypes (Carraro et al., 2017). Model performance has been evaluated in terms of distance and correlation, measuring the deviation from experimental values while assessing model capability to predict a decrease or increase of cholesterol levels. The analysis has been conducted at two levels. In the first part of the assessment, predictions were evaluated at the level of the single gene to understand if prediction error was homogeneous or significantly different for some of the mutations. The second part of the assessment focused on the overall performance of the predictor. In the first phase, the analysis was focused on assessing model performance in terms of prediction error computed on each element (i) of the test set: the deviation was evaluated by computing the difference between predicted and experimental data, in terms of rate of cholesterol levels (CL), for TC, HDL or LDL, in case and control.

$$Error(i) = \frac{CL_{predicted}(i)}{CL_{control}} - \frac{CL_{experimental}(i)}{CL_{experimental\ control}(i)} \quad (27)$$

To evaluate the magnitude of the error, compared to real values, this measure was divided by the corresponding experimental value and multiplied by 100.

$$Error(i)\% = \frac{Error(i)}{\frac{CL_{experimental}(i)}{CL_{experimental\ control}(i)}} \cdot 100 \quad (28)$$

This analysis was aimed to highlight mutation effects that where under or over-predicted.

In the second part of the assessment, model performance has been evaluated in terms of correlation and error measures on the whole dataset. Correlation measures used for the assessment were Pearson (r or PCC) and Kendall's tau (τ or KCC) correlation coefficients (equation 15 and 16). The PCC has been used to evaluate the correlation between real and predicted data as continuous measures, while KCC estimated the

conservation of the order of magnitude of the experimental cholesterol levels in predicted ones.

To better understand the amount of variability described by the model compared to the variability inside the data, the R^2 index was used.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (29)$$

RMSE (Root Mean Squared Error) has been used to evaluate if the method predicted cholesterol levels with huge deviation from real ones.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Error(i))^2}{n}} \quad (30)$$

The MAE (Mean Absolute Error) has been computed as the mean absolute error between model predictions and experimental values.

$$MAE = \frac{\sum_{i=1}^{n}|Error(i)|}{n} \quad (31)$$

A bootstrap procedure has been used to evaluate the robustness of the performance measures presented in this work: the probability of obtaining the same or better scores with a random shuffle of model predictions, as seen in (Carraro et al., 2017; Monzon et al., 2019). In particular each index has been computed 10000 times, considering either the rate of HDL, LDL or total cholesterol of all the elements of the test set each time, on the vector of model predictions in a random order and the corresponding vector of experimental values. From the resulting distribution of scores, the probability (p-value) of obtaining a score greater or equal to the real one was computed. The only exception was the RMSE index, in this case the probability of obtaining a value lower or equal to the score computed in the original assessment was calculated. All indices with a p-value lower than 0.05 were considered as statistically significant.

A sensitivity analysis was performed on a set of rates, corresponding to genes represented in the test set. The aim of this analysis was to understand the effect of a perturbation of specific model parameters on the output (Cazzaniga et al., 2014). In this

case, we decreased rates associated to genes represented in the test set, using a reducing factor [0.1, 1] and measured model cholesterol levels when a steady state was reached.

Performance assessment on single gene mutations

The first part of the assessment was aimed to understand how the model performs on the single mutations represented in the test set. This type of analysis highlighted cases where the model overestimated or underestimated cholesterol levels, respectively called positive or negative errors. The error represents the increase or decrease of cholesterol in case relative to control (ΔHDL, ΔLDL, ΔTC), which is not observed in experimental data. The errors were divided by real data and converted to percentages as reported in Table 16. The standard deviation of model predictions, computed as the standard deviation of the predicted cholesterol levels for the elements of the training set given a mutated gene, has been reported in Table 17. As already introduced, the datasets of patients have been divided in two sections. The first group of elements of the test set is composed by patients affected by damaging mutations on genes that have a role in the onset of the Autosomal Dominant Hypercholesterolemia: LDLR, APOB and APOE. The main effect of simulating these mutations is an increase of blood cholesterol levels of LDL and decrease of HDL (Table 16), as observed in real cases (Gidding et al., 2015). The algorithm predicted cholesterol levels caused by mutations in LDLR and APOB with a reduced error intervals: [-35.3%, 11.5%] for HDL, [-26.2%, -12.9%] for LDL and [-20.5%, -13.7%] for total cholesterol, respect to the former version of the model. The original model in fact, has shown to drive predictions toward an overestimation of the mutation effect, as shown by the prediction errors of HDL [-52%, -30.8%], LDL [35%, 139.7%], and total cholesterol [29.7%, 115.9%]. A particular case is the one regarding mutations in APOE, where the algorithm strongly underestimated the effect of damaging mutations on total cholesterol levels. In this case, a higher error has been registered for our optimized model (-53.1%) compared with the former version (-27.4%). This situation is mainly related to the fact that the average levels of total cholesterol of patients in the training set was lower than the one of the test set.

The effect of damaging mutations on the other genes of the test set have been simulated by reducing different set of rates of the model. The model predicted the effect of ABCA1

mutations as a decrease in HDL levels, but also produced overestimated decrease in LDL levels (Table 16), which is not usually observed in patients affected by related disease like Hypoalphalipoproteinemia (Shapiro, 2000). CETP is a protein involved in the transport of cholesterol esters from HDL to LDL, deficiency of this protein can cause a marked increase of HDL levels (Shapiro, 2000). In this case, the model correctly predicted an increase in HDL cholesterol levels, with a bigger error when optimized f_{mut} was used (Table 16). The LCAT gene is involved in cholesterol esterification in HDL particles, patients with mutations on this gene generally have low levels of HDL and LDL cholesterol (Shapiro, 2000). In this case the model was not able to accurately simulate HDL and LDL levels in all cases (Table 16). Explanation could be that it was not possible to train the parameter for patients with a homozygous mutation on the LCAT gene (f_{mut} has been assumed to be equal to 1). DHCR7 gene is involved in cholesterol biosynthesis pathway, mutations reducing related enzymatic activity cause low levels of blood cholesterol (Yu et al., 2000). In all cases the model predicted a bigger decrease in total cholesterol levels with an error of -171.5%. CYP7A1 gene is involved in cholesterol catabolism and bile acids synthesis, mutations affecting this gene can cause an increase of total and hepatic cholesterol (Pullinger et al., 2002). In this case the model generally predicted an increase of LDL and total cholesterol levels and a decrease in HDL cholesterol. Nevertheless, CYP7A1 simulations showed an underestimation of LDL and total cholesterol levels (Table 16).

Mutation	Gene	Predicted van de Pas et al			Predicted Reggiani et al		
		HDL	LDL	TC	HDL	LDL	TC
1	LDLR	-30.83	35.05	29.65	-13.52	-14.82	-13.69
2	APOB	-36.7	139.71	115.91	11.53	-26.23	-20.45
3	APOB *(hom)*	-51.96	62.66	49.05	-35.34	-12.89	-15.14
4	ABCA1	147.99	-57.44	-44.77	200.05	-51.25	-35.99
5	APOE	NA	NA	-27.4	NA	NA	-53.15
6	CETP	12.7	-4.82	-0.72	34.04	-11.53	-0.46
7	LCAT	34.3	-0.66	21.7	39.18	-3.08	20.55
8	LCAT *(hom)*	679.37	-19.37	10.11	426.32	21.95	29.87
9	DHCR7	NA	NA	171.51	NA	NA	171.51
10	CYP7A1	-4.42	-42.09	-34.15	1.37	-50.07	-40.81

Table 16. Models predictions percentage error on elements of the test set. Mutation numeric ID, gene, HDL, LDL and total cholesterol error (as percentage of experimental value), of predictions based on f_{mut} as reported by van de Pas and colleagues (van de Pas et al., 2012), or trained f_{mut}**.**

Mutation	Gene	Experimental value			Predicted van de Pas et al			Predicted Reggiani et al		
		HDL	LDL	TC	HDL	LDL	TC	HDL	LDL	TC
1	LDLR	0.86	2.17	1.85	0.59	2.93	2.4	0.74±0.17	1.85±1.31	1.6±0.97
2	APOB	0.85	1.52	1.36	0.54	3.64	2.94	0.95±0.07	1.12±0.21	1.08±0.14
3	APOB *(hom)*	1.12	2.24	1.97	0.54	3.64	2.94	0.72	1.95	1.67
4	ABCA1	0.22	1.42	1.07	0.55	0.6	0.59	0.66±0.19	0.69±0.15	0.68±0.16
5	APOE	NA	NA	2.8	0.65	2.44	2.03	0.84±0.13	1.45±0.57	1.31±0.41
6	CETP	1.1	0.98	1.01	1.24	0.93	1	1.47	0.87	1.01
7	LCAT	0.79	0.97	0.81	1.06	0.96	0.99	1.1±0.16	0.94±0.11	0.98±0.05
8	LCAT *(hom)*	0.19	0.82	0.77	1.48	0.66	0.85	1	1	1
9	DHCR7	NA	NA	0.2	1.13	0.37	0.54	1.13±0.01	0.37±0.04	0.54±0.03
10	CYP7A1	0.97	2.09	1.74	0.93	1.21	1.15	0.98±0.02	1.04±0.06	1.03±0.04

Table 17. Experimental and predicted cholesterol levels of the test set. Mutation numeric ID, gene, HDL, LDL and total cholesterol, from wet lab experiments (van de Pas et al., 2012), from predictions based on f_{mut} as reported by van de Pas and colleagues (van de Pas et al., 2012), or trained f_{mut} with standard deviation.

Performance assessment on the overall dataset

The overall assessment highlighted that the training phase increased model performance (Table 18). Both Pearson and Kendall correlation coefficients show that the use of trained f_{mut} increased algorithm capability to predict variations on cholesterol levels caused by gene mutations. In particular, HDL levels predicted by the former version of the model have shown negative correlation with experimental values. The MAE and RMSE index computed on HDL and total cholesterol levels have been decreased thanks to the training procedure, and the second index on predicted LDL is one half of the one obtained with the original version of the model. R^2 indices on blood cholesterol levels show an increase in the amount of variability explained by the model when a training phase is added. The bootstrap procedure has shown that for all indices computed on HDL levels predictions, model performance was not better than random.

Prediction	PCC	KCC	MAE	RMSD	R^2
van de Pas et al predicted HDL ratio	-0.22	-0.18	0.4	0.54	0.05
Reggiani et al predicted HDL ratio	0.32	0	0.32	0.4	0.11
van de Pas et al predicted LDL ratio	**0.65**	**0.55**	**0.77**	**1.03**	0.43
Reggiani et al predicted LDL ratio	**0.74**	0.5	**0.39**	**0.5**	**0.55**
van de Pas et al predicted TC ratio	**0.66**	**0.63**	**0.55**	**0.71**	**0.43**
Reggiani et al predicted TC ratio	**0.75**	**0.69**	**0.42**	**0.57**	**0.56**

Table 18. Models performances on the whole test set. Cholesterol level and predictor: Pearson Correlation Coefficient, Kendall rank Correlation Coefficient, Root Mean Squared Error, Mean Absolute Error and R-squared index computed on the test set. Values in bold have a p-value lower than 0.05, computed as the probability of obtaining an index better than the original one in a distribution of 10000 random scores, generated by a bootstrap procedure.

Conclusions

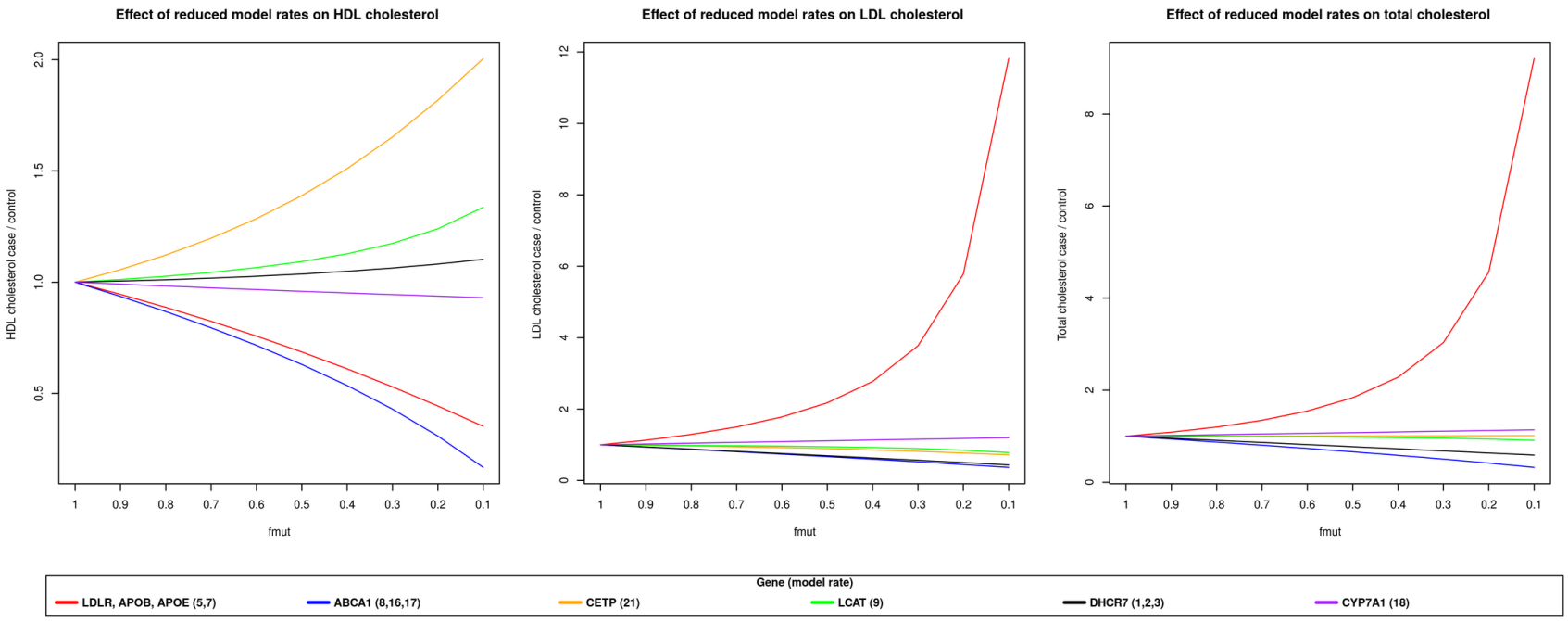
In this work we improved and assessed the performance of an *in silico* prediction method for blood cholesterol levels. The addition of a training phase has generally improved model performance, as shown in Table 18. Our training phase overcomes the problem of model usability when no experimental data is available for f_{mut} parameters estimation. The reducing parameters presented by van de Pas and colleagues were computed from variables obtained in wet lab experiments(van de Pas et al., 2012). This procedure, in contrast with the training methodology we applied in this work, did not take in account that decreasing different rates by the same factor can lead to modification of cholesterol profiles with different magnitude. To better understand model responses to different simulations, we performed a sensitivity analysis on the rates involved in the test set (Figure 23). This analysis showed that reduction of rate 5 and 7 produced a consistent decrease of model predicted HDL, while increasing LDL and total cholesterol. The training procedure has computed f_{mut} on the basis of the difference between experimental levels and model response to the reduction of selected rates, as previously explained. This procedure avoid an overestimation of the effect of mutations on the LDLR and APOB genes, as observed when f_{mut} based on experimental variables were used (Table 16). The use of trained parameters has decreased prediction error when model was not able to correctly simulate the effect of a mutated gene on cholesterol levels. In particular rate 9 regulates the flow of cholesterol from free to esterified form in HDL particles, LCAT gene product activity. The effect of a mutation on this gene is predicted by the model as an increase of HDL cholesterol while the opposite is observed in real data (Table 17). In this case the training procedures had hampered in part model inability to correctly predict HDL and LDL deviations caused by mutations on this gene by fixing the f_{mut} to 1 in the homozygous case, since the reduction of this parameter was not able to reduce the difference between experimental and predicted values. A similar behavior has been observed for the estimation of the reducing CYP7A1 f_{mut}: in this case the trained parameter (0.81) was greater than the value computed by van de Pas and coauthors (0, Table 14), while the difference in predicted cholesterol levels was relatively small (Table 17). This difference is related both to the low sensitivity of the rate (Figure 23) and the

inability to produce a consistent increase of LDL and Total cholesterol levels, while producing a limited decrease of HDL as experimentally observed (Table 17).

This model can be considered a valid tool for the study of cholesterol metabolism *in silico*, considering the other models currently available (Paalvast et al., 2015) and the prediction error: the average relative deviations between model predictions and experimental data were 49% for HDL-C, 43% for LDL-C and 36% for total cholesterol (van de Pas et al., 2012). Mathematical models are a simplified representation of the original system, this from one hand results in a relatively simple tool for making inference and simulate different experimental conditions *in silico*. From the other hand, they don't represent the selected system completely, hence deviation from real data are expected. Prediction error in principle could be decreased by increasing the number of parameters, however this process will increase model complexity and present problems related to parameter identifiability and fitting to experimental error (White et al., 2016). Prediction of *in silico* cholesterol levels is a complex procedure, the physiologically based *in silico* cholesterol model optimized in this review has proven its ability to predict cholesterol levels behavior with reduced error when only genotype data is available. Given the huge number of genomic loci controlling cholesterol homeostasis, much of that still unknown, gender-related effects and environmental factors that affects blood cholesterol levels, the possibility of developing a software able to accurately predict cholesterol levels seems far from true. Despite these critical points, in this work we considered patients affected by monogenic dominant diseases only, so we expect that ethnicity and other factors will have a relatively low contribution on the onset of the phenotype. Nowadays genetic assays are increasingly used to support the diagnosis of monogenic diseases affecting blood lipid levels, as Familial Hypercholesterolemia (FH) (Gidding et al., 2015). Studies have shown that coronary artery disease risk is higher in carrier of FH mutations compared to those without, this is likely the consequence of a higher life-long exposure to LDL (Khera et al., 2016). In this context our work could be considered a further step in the process of using genetic tests for the detection and treatment of patients affected by FH and other genetic disorders affecting blood cholesterol levels. Under this perspective we think that our work could be useful to simulate and study the effect of genetic variants on human cholesterol metabolism, in particular for variants affecting genes involved in hepatic cholesterol

uptake (model rate 5, 7) and FH, as LDLR and APOB, where the trained model has predicted blood cholesterol levels with little error (Table 16, 17). Furthermore, given the newly developed therapies against molecular targets (such as the anti PCSK9 monoclonal antibodies) (Verbeek et al., 2015) the model could be useful to identify the patients that are best candidates for treatment. Simulation of drug actions could be another possible application of the proposed model. It is well known that different genetic backgrounds have a strong effect on drug activity and *in silico* prediction methods can have poor performance with patients that don't have the same ethnicity of individuals used during the training procedure, as seen in CAGI Warfarin dosing challenge (Daneshjou et al., 2017b).

In light of these considerations, the physiologically based *in silico* model of human cholesterol metabolism, optimized in this work, can be a useful tool for studying the effect of damaging mutation on genes involved in cholesterol homeostasis



.

Figure 23. Model response in terms of HDL, LDL and total blood cholesterol at different values of f_{mut}. The effect of reducing model rates, involved in the test procedure, on HDL, LDL or total cholesterol levels

Conclusions

The development of next generation sequencing technologies has reduced the price of sequencing a human genome to around 1000US$. This has improved the study of human genetics, through mass sequencing of multiple individuals. This data shed light on a landscape of loci associated with complex diseases or quantitative traits, as blood lipid levels, a task that was impossible to accomplish with the technologies of the previous century. This huge amount of data has driven the scientific community to develop methods able to extract information from newly sequenced genomes and interpret them under the light of what has been published from previous NGS or biomedical experiments (Goodwin et al., 2016). Among all, the most intriguing application is personalized medicine. In this case we can adapt patient treatment to genomic information. In practice clinicians could optimize pharmaceutical or surgical treatments on the basis of the effect associated to detected genomic variants, predicting drug adverse effects or disease development (Welsh et al., 2017). Unfortunately these promising techniques requires affordable tools to extract and interpret the information inside NGS data. One way to test the quality of a method is to perform a fair assessment. Different machine learning tools has been published during the last decades, with different training set and performance measures. These fact makes comparison between different methods infeasible, while unbalanced test sets or the use of same data for model training and validation can lead to a biased evaluation of model performance (Vihinen, 2012). The CAGI challenge experiment has been established as a way to assess, in a fair way, the state of the art in methods for phenotype or variant effect prediction. In each challenge different predictors are assessed on a public genotype dataset, with known unpublished phenotype. Unfortunately the assessment process has no general guidelines or software, which makes difficult to compare different CAGI challenges, and identify an improvement between different CAGI editions. In my work I have developed an R package, based on a solid literature background, that is able to perform a mathematical assessment of regression and multiple phenotype prediction methods. This tool simplify the assessment procedure, enabling a complete and standardized evaluation procedure. In particular the

outlier analysis and single phenotype assessment, introduced with this work, have been the basis for the evaluation of methods that were part of the last, fifth edition of CAGI. In this experiment I took part to the assessment of two challenges: the ID and PCM1 challenge. The first one was a multiple label challenge, where predictors had to identify healthy or affected individuals for one or more of 7 phenotype classes from patients genotype, i.e. their vcf. This task was cumbersome compared to previous CAGI experiments, since the correlation between genotype and phenotype was lower compared to previous challenges. This issue has been reflected by the low performance in phenotype predictions achieved by most groups. Nevertheless, predictors were able to detect most of the variants that were related to the disease, according to data providers. In some cases variants that were considered pathogenic by most groups, but initially discarded by the Padua NDD lab, have been later validated in wet lab, adding information to the original analysis (Carraro et al., 2019). The evaluation of the PCM1 challenge highlighted the limits of variant effect predictions on proteins with unknown structure. Different groups obtained poor performance, with the best predictor achieving an MCC of 0.35 (Monzon et al., 2019). Nevertheless, it should be taken into account that some of the methods presented in this challenge have outperformed previously published state of the art tools, (Miller et al., 2019b). This assessment has driven to the conclusion that actual methods have limited performance on some proteins, therefore the scientific community should improve available tools and develop more general ones.

My work on the assessment of predictors based on genotype information has given me the knowledge to develop an in silico tool for cholesterol levels prediction for a personalized medicine project. The developed method is able to predict cholesterol levels from patient genotype information. This kind of data could be useful to adapt patient treatment according to the effect of genomic variations, before the onset of the disease by the subject. The mathematical model that is behind the prediction algorithm, has been published in 2012 (van de Pas et al., 2012). Unfortunately no software was made available and the reliability of the tool was not clearly assessed. Therefore I have to develop a script in R language to run the original model and a pipeline to predict cholesterol levels when only genotype information is available. This task has been achieved by adding a training phase of model f_{mut} parameters, representing the

decreasing effect on gene activity caused by mutations (Reggiani et al., 2018). The training set of patients was based on data retrieved from previously published works. The developed tool has been assessed on the same set of the original publication (van de Pas et al., 2012) and showed a substantial improvement in the quality of predictions, compared to the previously published version of the model (van de Pas et al., 2012). With this work I have proved that the prediction of blood cholesterol levels from genotype information is a feasible task, at least in the case of monogenic dominant hypercholesterolemia and with a certain degree of uncertainty.

The development of the whole thesis process has been devoted to develop and assess tools that could be used to predict a phenotype from genotype information. The scientific community has been struggling against the problem of methods availability and their fair assessment. This problem is evident considering published mathematical models simulating biological processes. While in the original publications these methods are presented with good performances, they may show lower performances on different test set, due to lack of generalization or over-fitting during the training phase (White et al., 2016). In other cases the software used to implement the published mathematical model is not available (van de Pas et al., 2012) and has to be written again in order to be used. This requires both a careful interpretation and error checking on model's equations and parameters. Unfortunately in some cases the information published in the paper is still not enough. Therefore the researcher is forced to directly ask authors for a working algorithm able to reproduce the results of the mathematical model presented in the paper (Paalvast et al., 2015). Fortunately new editorial guidelines are forcing research groups to develop reproducible and transparent research (PLOS ONE), while new NGS data is becoming available for the researchers community (PGP-UK Consortium, 2018; Silvester et al., 2018). These factors will probably drive research towards the development of more accurate methods, that will make personalized medicine a practical approach, at least in the long term.

References

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

Abraham, G., and Inouye, M. (2015). Genomic risk prediction of complex human disease and its clinical application. Curr. Opin. Genet. Dev. *33*, 10–16.

Almuhtaseb, S., Oppewal, A., and Hilgenkamp, T.I.M. (2014). Gait characteristics in individuals with intellectual disabilities: a literature review. Res. Dev. Disabil. *35*, 2858–2883.

Alyass, A., Turcotte, M., and Meyre, D. (2015). From big data analysis to personalized medicine for all: challenges and opportunities. BMC Med. Genomics *8*, 33.

An, J.Y., Cristino, A.S., Zhao, Q., Edson, J., Williams, S.M., Ravine, D., Wray, J., Marshall, V.M., Hunt, A., Whitehouse, A.J.O., et al. (2014). Towards a molecular characterization of autism spectrum disorders: an exome sequencing and systems approach. Transl. Psychiatry *4*, e394.

Ansley, S.J., Badano, J.L., Blacque, O.E., Hill, J., Hoskins, B.E., Leitch, C.C., Kim, J.C., Ross, A.J., Eichers, E.R., Teslovich, T.M., et al. (2003). Basal body dysfunction is a likely cause of pleiotropic Bardet-Biedl syndrome. Nature *425*, 628–633.

Aspromonte, M.C., Bellini, M., Gasparini, A., Carraro, M., Bettella, E., Polli, R., Cesca, F., Bigoni, S., Boni, S., Carlet, O., et al. (2019). Characterization of Intellectual disability and Autism comorbidity through gene panel sequencing. Hum. Mutat.

Ayala, R., Shu, T., and Tsai, L.-H. (2007). Trekking across the brain: the journey of neuronal migration. Cell *128*, 29–43.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. Bioinforma. Oxf. Engl. *16*, 412–424.

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nat. Rev. Genet. *12*, 56–68.

Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M., et al. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. Nat. Genet. *40*, 955–962.

Boughorbel, S., Jarray, F., and El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PloS One *12*, e0177678.

Bowley, C., and Kerr, M. (2000). Epilepsy and intellectual disability. J. Intellect. Disabil. Res. JIDR *44 (Pt 5)*, 529–543.
Bromberg, Y., and Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res. *35*, 3823–3835.
Bromberg, Y., Yachdav, G., and Rost, B. (2008). SNAP predicts effect of mutations on protein function. Bioinforma. Oxf. Engl. *24*, 2397–2398.
Buys, S.S., Sandbach, J.F., Gammon, A., Patel, G., Kidd, J., Brown, K.L., Sharma, L., Saam, J., Lancaster, J., and Daly, M.B. (2017). A study of over 35,000 women with breast cancer tested with a 25-gene panel of hereditary cancer genes: Panel Testing in Women With Breast Ca. Cancer *123*, 1721–1730.
Cai, B., Li, B., Kiga, N., Thusberg, J., Bergquist, T., Chen, Y.-C., Niknafs, N., Carter, H., Tokheim, C., Beleva-Guthrie, V., et al. (2017). Matching phenotypes to whole genomes: Lessons learned from four iterations of the personal genome project community challenges. Hum. Mutat. *38*, 1266–1276.
Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P.L., and Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum. Mutat. *30*, 1237–1244.
de los Campos, G., Gianola, D., and Allison, D.B. (2010). Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat. Rev. Genet. *11*, 880–886.
Capriotti, E., Calabrese, R., and Casadio, R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinforma. Oxf. Engl. *22*, 2729–2734.
Capriotti, E., Nehrt, N.L., Kann, M.G., and Bromberg, Y. (2012). Bioinformatics for personal genome interpretation. Brief. Bioinform. *13*, 495–512.
Capriotti, E., Calabrese, R., Fariselli, P., Martelli, P.L., Altman, R.B., and Casadio, R. (2013). WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. BMC Genomics *14 Suppl 3*, S6.
Capriotti, E., Ozturk, K., and Carter, H. (2018). Integrating molecular networks with genetic variant interpretation for precision medicine. Wiley Interdiscip. Rev. Syst. Biol. Med. e1443.
Carraro, M., Minervini, G., Giollo, M., Bromberg, Y., Capriotti, E., Casadio, R., Dunbrack, R., Elefanti, L., Fariselli, P., Ferrari, C., et al. (2017). Performance of in silico tools for the evaluation of p16INK4a (CDKN2A) variants in CAGI. Hum. Mutat. *38*, 1042–1050.
Carraro, M., Monzon, A.M., Chiricosta, L., Reggiani, F., Aspromonte, M.C., Bellini, M., Pagel, K., Jiang, Y., Radivojac, P., Kundu, K., et al. (2019). Assessment of patient clinical descriptions and

pathogenic variants from gene panel sequences in the CAGI-5 intellectual disability challenge. Hum. Mutat.
Carter, H., Douville, C., Stenson, P.D., Cooper, D.N., and Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics *14 Suppl 3*, S3.
Casadio, R., Vassura, M., Tiwari, S., Fariselli, P., and Luigi Martelli, P. (2011). Correlating disease-related mutations to their effect on protein stability: a large-scale analysis of the human proteome. Hum. Mutat. *32*, 1161–1170.
Cazzaniga, P., Damiani, C., Besozzi, D., Colombo, R., Nobile, M.S., Gaglio, D., Pescini, D., Molinari, S., Mauri, G., Alberghina, L., et al. (2014). Computational strategies for a system-level understanding of metabolism. Metabolites *4*, 1034–1087.
Chandonia, J.-M., Adhikari, A., Carraro, M., Chhibber, A., Cutting, G.R., Fu, Y., Gasparini, A., Jones, D.T., Kramer, A., Kundu, K., et al. (2017). Lessons from the CAGI-4 Hopkins clinical panel challenge. Hum. Mutat. *38*, 1155–1168.
Chandonia, J.-M., Adhikari, A., Carraro, M., Chhibber, A., Cutting, G.R., Fu, Y., Gasparini, A., Jones, D.T., Kramer, A., Kundu, K., et al. Lessons from the CAGI-4 Hopkins clinical panel challenge. Hum. Mutat. n/a-n/a.
Check Hayden, E. (2017). The rise and fall and rise again of 23andMe. Nature *550*, 174-177
Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., and Chan, A.P. (2012). Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS ONE *7*, e46688.
Cobelli, C., Carson, E.R., Facchinetti, A., and Dalla Man, C. (2012). Introduzione alla modellistica in fisiologia e medicina (Bologna: Pàtron).
Dammermann, A., and Merdes, A. (2002). Assembly of centrosomal proteins and microtubule organization depends on PCM-1. J. Cell Biol. *159*, 255–266.
Daneshjou, R., Wang, Y., Bromberg, Y., Bovo, S., Martelli, P.L., Babbi, G., Lena, P.D., Casadio, R., Edwards, M., Gifford, D., et al. (2017a). Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. Hum. Mutat. *38*, 1182–1192.
Daneshjou, R., Wang, Y., Bromberg, Y., Bovo, S., Martelli, P.L., Babbi, G., Lena, P.D., Casadio, R., Edwards, M., Gifford, D., et al. (2017b). Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. Hum. Mutat. *38*, 1182–1192.
Davis, E.E., Frangakis, S., and Katsanis, N. (2014). Interpreting human genetic variation with in vivo zebrafish assays. Biochim. Biophys. Acta *1842*, 1960–1970.
Farrell, P. M., White, T. B., Ren, C. L., Hempstead, S. E., Accurso, F., Derichs, N., … Sosnay,

P. R. (2017). Diagnosis of Cystic Fibrosis: Consensus Guidelines from the Cystic Fibrosis Foundation. The Journal of Pediatrics, *181*, S4-S15.e1.
Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognit. Lett. *27*, 861–874.
Garver, K. L., and Garver, B. (1994). The Human Genome Project and eugenic concerns. American Journal of Human Genetics *54*(1), 148-158.
Geiss-Friedlander, R., and Melchior, F. (2007). Concepts in sumoylation: a decade on. Nat. Rev. Mol. Cell Biol. *8*, 947–956.
Gianola, D. (2013). Priors in whole-genome regression: the bayesian alphabet returns. Genetics *194*, 573–596.
Gidding, S.S., Ann Champagne, M., de Ferranti, S.D., Defesche, J., Ito, M.K., Knowles, J.W., McCrindle, B., Raal, F., Rader, D., Santos, R.D., et al. (2015). The Agenda for Familial Hypercholesterolemia: A Scientific Statement From the American Heart Association. Circulation *132*, 2167–2192.
Giollo, M., Minervini, G., Scalzotto, M., Leonardi, E., Ferrari, C., and Tosatto, S.C.E. (2015). BOOGIE: Predicting Blood Groups from High Throughput Sequencing Data. PloS One *10*, e0124579.
Giollo, M., Jones, D.T., Carraro, M., Leonardi, E., Ferrari, C., and Tosatto, S.C.E. (2017). Crohn disease risk prediction-Best practices and pitfalls with exome data. Hum. Mutat. *38*, 1193–1200.
Goodwin, S., McPherson, J.D., and McCombie, W.R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. *17*, 333–351.
Guo, J., Yang, Z., Song, W., Chen, Q., Wang, F., Zhang, Q., and Zhu, X. (2006). Nudel contributes to microtubule anchoring at the mother centriole and is involved in both dynein-dependent and -independent centrosomal protein assembly. Mol. Biol. Cell *17*, 680–689.
Guo, J., Gao, Y., Li, X., He, Y., Zheng, X., Bi, J., Hou, L., Sa, Y., Zhang, M., Yin, H., et al. (2019). Systematic prediction of familial hypercholesterolemia caused by low-density lipoprotein receptor missense mutations. Atherosclerosis *281*, 1–8.
Gupta, A., Tsai, L.-H., and Wynshaw-Boris, A. (2002). Life is a journey: a genetic look at neocortical development. Nat. Rev. Genet. *3*, 342–355.
Gurling, H.M.D., Critchley, H., Datta, S.R., McQuillin, A., Blaveri, E., Thirumalai, S., Pimm, J., Krasucki, R., Kalsi, G., Quested, D., et al. (2006). Genetic association and brain morphology studies and the chromosome 8p22 pericentriolar material 1 (PCM1) gene in susceptibility to schizophrenia. Arch. Gen. Psychiatry *63*, 844–854.
Gutzman, J.H., and Sive, H. (2009). Zebrafish brain ventricle injection. J. Vis. Exp. JoVE.
Hamp, T., and Rost, B. (2012). Alternative protein-protein interfaces are frequent exceptions.

PLoS Comput. Biol. *8*, e1002623.
He, D., Saha, S., Finkers, R., and Parida, L. (2018). Efficient algorithms for polyploid haplotype phasing. BMC Genomics *19*.
Helms, V. (2008). Principles of computational cell biology: from protein complexes to cellular networks (Weinheim: Wiley-VCH).
Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J., et al. (2007). Genome-wide in situ exon capture for selective resequencing. Nat. Genet. *39*, 1522–1527.
Hoskins, R.A., Repo, S., Barsky, D., Andreoletti, G., Moult, J., and Brenner, S.E. (2017). Reports from CAGI: The Critical Assessment of Genome Interpretation. Hum. Mutat. *38*, 1039–1041.
Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D., et al. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am. J. Hum. Genet. *99*, 877–885.
Iossifov, I., O'Roak, B.J., Sanders, S.J., Ronemus, M., Krumm, N., Levy, D., Stessman, H.A., Witherspoon, K.T., Vives, L., Patterson, K.E., et al. (2014). The contribution of de novo coding mutations to autism spectrum disorder. Nature *515*, 216–221.
Jain, S., White, M., and Radivojac, P. (2016). Estimating the class prior and posterior from noisy positives and unlabeled data. ArXiv160608561 Cs Stat.
Jian, X., Boerwinkle, E., and Liu, X. (2014). In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. *42*, 13534–13544.
Kamiya, A., Tan, P.L., Kubo, K., Engelhard, C., Ishizuka, K., Kubo, A., Tsukita, S., Pulver, A.E., Nakajima, K., Cascella, N.G., et al. (2008). Recruitment of PCM1 to the centrosome by the cooperative action of DISC1 and BBS4: a candidate for psychiatric illnesses. Arch. Gen. Psychiatry *65*, 996–1006.
Kannengiesser, C., Brookes, S., del Arroyo, A.G., Pham, D., Bombled, J., Barrois, M., Mauffret, O., Avril M, M.-F., Chompret, A., Lenoir, G.M., et al. (2009). Functional, structural, and genetic evaluation of 20 *CDKN2A* germ line mutations identified in melanoma-prone families or patients. Hum. Mutat. *30*, 564–574.
Katsonis, P., and Lichtarge, O. (2014). A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. Genome Res. *24*, 2050–2058.
Khera, A.V., Won, H.-H., Peloso, G.M., Lawson, K.S., Bartz, T.M., Deng, X., van Leeuwen, E.M., Natarajan, P., Emdin, C.A., Bick, A.G., et al. (2016). Diagnostic Yield and Clinical Utility of

Sequencing Familial Hypercholesterolemia Genes in Patients With Severe Hypercholesterolemia. J. Am. Coll. Cardiol. *67*, 2578–2589.
Kohane, I.S., Hsing, M., and Kong, S.W. (2012). Taxonomizing, sizing, and overcoming the incidentalome. Genet. Med. Off. J. Am. Coll. Med. Genet. *14*, 399–404.
Koren, S., Rhie, A., Walenz, B.P., Dilthey, A.T., Bickhart, D.M., Kingan, S.B., Hiendleder, S., Williams, J.L., Smith, T.P.L., and Phillippy, A.M. (2018). De novo assembly of haplotype-resolved genomes with trio binning. Nat. Biotechnol.
Krumm, N., O'Roak, B.J., Shendure, J., and Eichler, E.E. (2014). A de novo convergence of autism genetics and molecular neuroscience. Trends Neurosci. *37*, 95–105.
Kubo, A., and Tsukita, S. (2003). Non-membranous granular organelle consisting of PCM-1: subcellular distribution and cell-cycle-dependent assembly/disassembly. J. Cell Sci. *116*, 919–928.
Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res. *44*, D862-868.
Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
Lesch, K.-P. (2016). Maturing insights into the genetic architecture of neurodevelopmental disorders - from common and rare variant interplay to precision psychiatry. J. Child Psychol. Psychiatry *57*, 659–661.
Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinforma. Oxf. Engl. *30*, 2843–2851.
Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. *25*, 1754–1760.
Lin, C.-H., Konecki, D.M., Liu, M., Wilson, S.J., Nassar, H., Wilkins, A.D., Gleich, D.F., and Lichtarge, O. (2019). Multimodal network diffusion predicts future disease-gene-chemical associations. Bioinforma. Oxf. Engl. *35*, 1536–1543.
Liu, D.J., Peloso, G.M., Yu, H., Butterworth, A.S., Wang, X., Mahajan, A., Saleheen, D., Emdin, C., Alam, D., Alves, A.C., et al. (2017). Exome-wide association study of plasma lipids in >300,000 individuals. Nat. Genet. *49*, 1758–1766.
Low, S.-K., Zembutsu, H., and Nakamura, Y. (2018). Breast cancer: The translation of big genomic data to cancer precision medicine. Cancer Sci. *109*, 497–506.
Lowery, L.A., De Rienzo, G., Gutzman, J.H., and Sive, H. (2009). Characterization and

classification of zebrafish brain morphology mutants. Anat. Rec. Hoboken NJ 2007 *292*, 94–106.
MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., et al. (2017). The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. *45*, D896–D901.
Marcuzzi, F. (2011). Analisi dei dati mediante modelli matematici: metodi numerici ed applicazioni a problemi di identificazione di modelli dai dati sperimentali, misura indiretta, predizione e stima dello stato (Padova: Libreria internazionale Cortina).
Marduel, M., Ouguerram, K., Serre, V., Bonnefont-Rousselot, D., Marques-Pinheiro, A., Erik Berge, K., Devillers, M., Luc, G., Lecerf, J.-M., Tosolini, L., et al. (2013). Description of a Large Family with Autosomal Dominant Hypercholesterolemia Associated with the *APOE* p.Leu167del Mutation. Hum. Mutat. *34*, 83–87
Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., & Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, *27*(2), e1608. https://doi.org/10.1002/mpr.1608
Mata, I.F., Jang, Y., Kim, C.-H., Hanna, D.S., Dorschner, M.O., Samii, A., Agarwal, P., Roberts, J.W., Klepitskaya, O., Shprecher, D.R., et al. (2015). The RAB39B p.G192R mutation causes X-linked dominant Parkinson's disease. Mol. Neurodegener. *10*, 50.
Mattingly, C.J., Colby, G.T., Forrest, J.N., and Boyer, J.L. (2003). The Comparative Toxicogenomics Database (CTD). Environ. Health Perspect. *111*, 793–795.
McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.
McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol. *17*.
Mikut, R., Dickmeis, T., Driever, W., Geurts, P., Hamprecht, F.A., Kausler, B.X., Ledesma-Carbayo, M.J., Marée, R., Mikula, K., Pantazis, P., et al. (2013). Automated processing of zebrafish imaging data: a survey. Zebrafish *10*, 401–421.
Miller, M., Vitale, D., Kahn, P., Rost, B., and Bromberg, Y. (2019a). fuNTRp: Identifying protein positions for variation driven functional tuning. BioRxiv.
Miller, M., Wang, Y., and Bromberg, Y. (2019b). What went wrong with variant effect predictor performance for the PCM1 challenge. Hum. Mutat.
Miller, P.J., Duraisamy, S., Newell, J.A., Chan, P.A., Tie, M.M., Rogers, A.E., Ankuda, C.K., von

Walstrom, G.M., Bond, J.P., and Greenblatt, M.S. (2011). Classifying variants of CDKN2A using computational and laboratory studies. Hum. Mutat. *32*, 900–911.
Mitchell, K.J. (2011). The genetics of neurodevelopmental disease. Curr. Opin. Neurobiol. *21*, 197–203.
Miyoshi, K., Asanuma, M., Miyazaki, I., Diaz-Corrales, F.J., Katayama, T., Tohyama, M., and Ogawa, N. (2004). DISC1 localizes to the centrosome by binding to kendrin. Biochem. Biophys. Res. Commun. *317*, 1195–1199.
Mochida, G.H., and Walsh, C.A. (2004). Genetic basis of developmental malformations of the cerebral cortex. Arch. Neurol. *61*, 637–640.
Monzon, A.M., Carraro, M., Chiricosta, L., Reggiani, F., Han, J., Ozturk, K., Wang, Y., Miller, M., Bromberg, Y., Capriotti, E., et al. (2019). Performance of computational methods for the evaluation of Pericentriolar Material 1 missense variants in CAGI-5. Hum. Mutat.
Moraes, F., and Góes, A. (2016). A decade of human genome project conclusion: Scientific diffusion about our genome knowledge. Biochem. Mol. Biol. Educ. Bimon. Publ. Int. Union Biochem. Mol. Biol. *44*, 215–223.
Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., and Singh, M. (2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics *21 Suppl 1*, i302-310.
Näslund, J., and Johnsson, J.I. (2016). Environmental enrichment for fish in captive environments: effects of physical structures and substrates. Fish Fish. *17*, 1–30.
Necci, M., Piovesan, D., Dosztanyi, Z., Tompa, P., and Tosatto, S.C.E. (2017). A comprehensive assessment of long intrinsic protein disorder from the DisProt database. Bioinformatics.
Niederriter, A.R., Davis, E.E., Golzio, C., Oh, E.C., Tsai, I.-C., and Katsanis, N. (2013). In vivo modeling of the morbid human genome using Danio rerio. J. Vis. Exp. JoVE e50338.
Niroula, A., and Vihinen, M. (2016). Variation Interpretation Predictors: Principles, Types, Performance, and Choice. Hum. Mutat. *37*, 579–597.
Niroula, A., and Vihinen, M. (2019). How good are pathogenicity predictors in detecting benign variants? PLoS Comput. Biol. *15*, e1006481.
OMIM MIM Number: 603903 : 02/05/2019 Online Mendelian Inheritance in Man, OMIM (TM). Johns Hopkins University, Baltimore, MD. MIM Number: 603903 : 02/05/2019: . World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/.
Paalvast, Y., Kuivenhoven, J.A., and Groen, A.K. (2015). Evaluating computational models of cholesterol metabolism. Biochim. Biophys. Acta *1851*, 1360–1376.

Pagel, K.A., Pejaver, V., Lin, G.N., Nam, H.-J., Mort, M., Cooper, D.N., Sebat, J., Iakoucheva, L.M., Mooney, S.D., and Radivojac, P. (2017). When loss-of-function is loss of function: assessing mutational signatures and impact of loss-of-function genetic variants. Bioinforma. Oxf. Engl. *33*, i389–i398.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

van de Pas, N.C.A., Soffers, A.E.M.F., Freidig, A.P., van Ommen, B., Woutersen, R.A., Rietjens, I.M.C.M., and de Graaf, A.A. (2010). Systematic construction of a conceptual minimal model of plasma cholesterol levels based on knockout mouse phenotypes. Biochim. Biophys. Acta *1801*, 646–654.

van de Pas, N.C.A., Woutersen, R.A., van Ommen, B., Rietjens, I.M.C.M., and de Graaf, A.A. (2011). A physiologically-based kinetic model for the prediction of plasma cholesterol concentrations in the mouse. Biochim. Biophys. Acta *1811*, 333–342.

van de Pas, N.C.A., Woutersen, R.A., van Ommen, B., Rietjens, I.M.C.M., and de Graaf, A.A. (2012). A physiologically based in silico kinetic model predicting plasma cholesterol concentrations in humans. J. Lipid Res. *53*, 2734–2746.

Pejaver, V., Urresti, J., Lugo-Martinez, J., Pagel, K.A., Lin, G.N., Nam, H.-J., Mort, M., Cooper, D.N., Sebat, J., Iakoucheva, L.M., et al. (2017). MutPred2: inferring the molecular and phenotypic impact of amino acid variants. BioRxiv.

PGP-UK Consortium (2018). Personal Genome Project UK (PGP-UK): a research and citizen science hybrid project in support of personalized medicine. BMC Med. Genomics *11*, 108.

Pinto, D., Delaby, E., Merico, D., Barbosa, M., Merikangas, A., Klei, L., Thiruvahindrapuram, B., Xu, X., Ziman, R., Wang, Z., et al. (2014). Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. Am. J. Hum. Genet. *94*, 677–694.

Piovesan, D., Tabaro, F., Mičetić, I., Necci, M., Quaglia, F., Oldfield, C.J., Aspromonte, M.C., Davey, N.E., Davidović, R., Dosztányi, Z., et al. (2017a). DisProt 7.0: a major update of the database of disordered proteins. Nucleic Acids Res. *45*, D219–D227.

Piovesan, D., Walsh, I., Minervini, G., and Tosatto, S.C.E. (2017b). FELLS: fast estimator of latent local structure. Bioinforma. Oxf. Engl. *33*, 1889–1891.

Piovesan, D., Tabaro, F., Paladin, L., Necci, M., Micetic, I., Camilloni, C., Davey, N., Dosztányi, Z., Mészáros, B., Monzon, A.M., et al. (2018). MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. Nucleic Acids Res. *46*, D471–D476.

Piton, A., Redin, C., and Mandel, J.-L. (2013). XLID-causing mutations and associated genes challenged in light of data from large-scale human exome sequencing. Am. J. Hum. Genet. *93*, 368–383.
PLOS ONE Criteria for publication.
Potter, L.K., and Tobin, F.L. (2007). Perspectives on mathematical modeling for receptor-mediated processes. J. Recept. Signal Transduct. Res. *27*, 1–25.
Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M., et al. (2014). RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. *42*, D756-763.
Pullinger, C.R., Eng, C., Salen, G., Shefer, S., Batta, A.K., Erickson, S.K., Verhagen, A., Rivera, C.R., Mulvihill, S.J., Malloy, M.J., et al. (2002). Human cholesterol 7alpha-hydroxylase (CYP7A1) deficiency has a hypercholesterolemic phenotype. J. Clin. Invest. *110*, 109–117.
R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
Radivojac, P., Peng, K., Clark, W.T., Peters, B.J., Mohan, A., Boyle, S.M., and Mooney, S.D. (2008). An integrated approach to inferring gene-disease associations in humans. Proteins *72*, 1030–1037.
Ramasamy, I. (2016). Update on the molecular biology of dyslipidemias. Clin. Chim. Acta *454*, 143–185.
Ramos-Lopez, O., Riezu-Boj, J.I., Milagro, F.I., Cuervo, M., Goni, L., and Martinez, J.A. (2018). Prediction of Blood Lipid Phenotypes Using Obesity-Related Genetic Polymorphisms and Lifestyle Data in Subjects with Excessive Body Weight. Int. J. Genomics *2018*, 4283078.
Reggiani, F., Carraro, M., Belligoli, A., Sanna, M., dal Pra', C., Favaretto, F., Ferrari, C., Vettor, R., and Tosatto, S.C.E. (2018). In silico prediction of blood cholesterol levels from genotype data. BioRxiv 503003.
Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. Off. J. Am. Coll. Med. Genet. *17*, 405–424
Sebastiani, P., Solovieff, N., Puca, A., Hartley, S. W., Melista, E., Andersen, S., … Perls, T. T. (2011). Retraction. Science, 333(6041), 404–404.
https://doi.org/10.1126/science.333.6041.404-a.
Shannon, C.E. (1948). A Mathematical Theory of Communication. Bell Syst. Tech. J. *27*, 379–

423.
Shapiro, M.D. (2000). Rare Genetic Disorders Altering Lipoproteins. In Endotext, L.J. De Groot, G. Chrousos, K. Dungan, K.R. Feingold, A. Grossman, J.M. Hershman, C. Koch, M. Korbonits, R. McLachlan, M. New, et al., eds. (South Dartmouth (MA): MDText.com, Inc.), p.
Silvester, N., Alako, B., Amid, C., Cerdeño-Tarrága, A., Clarke, L., Cleland, I., Harrison, P.W., Jayathilaka, S., Kay, S., Keane, T., et al. (2018). The European Nucleotide Archive in 2017. Nucleic Acids Res. *46*, D36–D40.
Soetaert, K., Petzoldt, T., and Setzer, R.W. (2010). Solving Differential Equations in *R*: Package **deSolve**. J. Stat. Softw. *33*.
Solecki, D.J., Govek, E.-E., Tomoda, T., and Hatten, M.E. (2006). Neuronal polarity in CNS development. Genes Dev. *20*, 2639–2647.
Spiliopoulou, A., Nagy, R., Bermingham, M.L., Huffman, J.E., Hayward, C., Vitart, V., Rudan, I., Campbell, H., Wright, A.F., Wilson, J.F., et al. (2015). Genomic prediction of complex human traits: relatedness, trait architecture and predictive meta-models. Hum. Mol. Genet. *24*, 4167–4182.
Stenson, P.D., Ball, E.V., Mort, M., Phillips, A.D., Shiel, J.A., Thomas, N.S.T., Abeysinghe, S., Krawczak, M., and Cooper, D.N. (2003). Human Gene Mutation Database (HGMD): 2003 update. Hum. Mutat. *21*, 577–581.
Strachan, T., Read, A.P., and Strachan, T. (2011). Human molecular genetics (New York: Garland Science).
Summerton, J., and Weller, D. (1997). Morpholino antisense oligomers: design, preparation, and properties. Antisense Nucleic Acid Drug Dev. *7*, 187–195.
Tewhey, R., Bansal, V., Torkamani, A., Topol, E.J., and Schork, N.J. (2011). The importance of phase information for human genomics. Nat. Rev. Genet. *12*, 215–223.
The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. Nucleic Acids Res. *45*, D158–D169.
Timur V. Elzhov, Katharine M. Mullen, Andrej-Nikolai Spiess and Ben Bolker (2016). minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds.
Tonnsen, B.L., Boan, A.D., Bradley, C.C., Charles, J., Cohen, A., and Carpenter, L.A. (2016). Prevalence of Autism Spectrum Disorders Among Children With Intellectual Disability. Am. J. Intellect. Dev. Disabil. *121*, 487–500.
Tremblay, J., and Hamet, P. (2013). Role of genomics on the path to personalized medicine. Metabolism. *62 Suppl 1*, S2-5.

Tsai, L.-H., and Gleeson, J.G. (2005). Nucleokinesis in neuronal migration. Neuron *46*, 383–388.
Verbeek, R., Stoekenbroek, R.M., and Hovingh, G.K. (2015). PCSK9 inhibitors: Novel therapeutic agents for the treatment of hypercholesterolemia. Eur. J. Pharmacol. *763*, 38–47.
Vihinen, M. (2012). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics *13 Suppl 4*, S2.
Villumsen, B.H., Danielsen, J.R., Povlsen, L., Sylvestersen, K.B., Merdes, A., Beli, P., Yang, Y.-G., Choudhary, C., Nielsen, M.L., Mailand, N., et al. (2013). A new cellular stress response that triggers centriolar satellite reorganization and ciliogenesis. EMBO J. *32*, 3029–3040.
Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.
Welsh, J.L., Hoskin, T.L., Day, C.N., Thomas, A.S., Cogswell, J.A., Couch, F.J., and Boughey, J.C. (2017). Clinical Decision-Making in Patients with Variant of Uncertain Significance in BRCA1 or BRCA2 Genes. Ann. Surg. Oncol. *24*, 3067–3072.
Wetterstrand, K.A. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed [22/09/2019].
Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G.T., et al. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature *452*, 872–876.
Whiffin, N., Roberts, A.M., Minikel, E., Zappala, Z., Walsh, R., O'Donnell-Luria, A.H., Karczewski, K.J., Harrison, S.M., Thomson, K.L., Sage, H., et al. (2019). Using High-Resolution Variant Frequencies Empowers Clinical Genome Interpretation and Enables Investigation of Genetic Architecture. Am. J. Hum. Genet. *104*, 187–190.
White, A., Tolman, M., Thames, H.D., Withers, H.R., Mason, K.A., and Transtrum, M.K. (2016). The Limitations of Model-Based Experimental Design and Parameter Estimation in Sloppy Systems. PLoS Comput. Biol. *12*, e1005227.
Xiong, H.Y., Alipanahi, B., Lee, L.J., Bretschneider, H., Merico, D., Yuen, R.K.C., Hua, Y., Gueroussov, S., Najafabadi, H.S., Hughes, T.R., et al. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science *347*, 1254806.
Yang, H., Robinson, P.N., and Wang, K. (2015). Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. Nat. Methods *12*, 841–843.
Yin, Y., Kundu, K., Pal, L.R., and Moult, J. (2017). Ensemble variant interpretation methods to predict enzyme activity and assign pathogenicity in the CAGI4 NAGLU (Human N-acetyl-

glucosaminidase) and UBE2I (Human SUMO-ligase) challenges. Hum. Mutat. *38*, 1109–1122.
Yu, H., Lee, M.H., Starck, L., Elias, E.R., Irons, M., Salen, G., Patel, S.B., and Tint, G.S. (2000). Spectrum of Delta(7)-dehydrocholesterol reductase mutations in patients with the Smith-Lemli-Opitz (RSH) syndrome. Hum. Mol. Genet. *9*, 1385–1391.
Zaghloul, N.A., Liu, Y., Gerdes, J.M., Gascue, C., Oh, E.C., Leitch, C.C., Bromberg, Y., Binkley, J., Leibel, R.L., Sidow, A., et al. (2010). Functional analyses of variants reveal a significant role for dominant negative and common alleles in oligogenic Bardet-Biedl syndrome. Proc. Natl. Acad. Sci. U. S. A. *107*, 10602–10607.
Zhang, J., Kinch, L.N., Cong, Q., Weile, J., Sun, S., Cote, A.G., Roth, F.P., and Grishin, N.V. (2017). Assessing predictions of fitness effects of missense mutations in SUMO-conjugating enzyme UBE2I. Hum. Mutat. *38*, 1051–1063.
Zhao, Y., & Cen, Y. (2014). Data mining applications with R. Amsterdam ; Boston: Elsevier.
Zoubovsky, S., Oh, E.C., Cash-Padgett, T., Johnson, A.W., Hou, Z., Mori, S., Gallagher, M., Katsanis, N., Sawa, A., and Jaaro-Peled, H. (2015). Neuroanatomical and behavioral deficits in mice haploinsufficient for Pericentriolar material 1 (Pcm1). Neurosci. Res. *98*, 45–49.
(2012). Principles of pharmacology: the pathophysiologic basis of drug therapy (Philadelphia: Wolters Kluwer Health, Lippincott Williams & Wilkins).

Supplementary materials

Appendix 1

Supplementary tables

Submission	BACC	MCC	AUC	TNR	TPR	FCPV	Jaccard	Overall Rank
61.1	0.79	0.594	0.83	0.96	0.628	0.570	0.540	1
59.1	0.75	0.544	0.75	0.98	0.535	0.430	0.394	2
59.2	0.74	0.532	0.73	0.98	0.512	0.430	0.405	3
58.2	0.68	0.404	0.69	0.97	0.395	0.360	0.349	4
58.1	0.66	0.380	0.66	0.97	0.349	0.302	0.295	5
60.1	0.58	0.216	0.60	0.97	0.209	0.178	0.147	6
60.2	0.59	0.236	0.60	0.97	0.209	0.178	0.147	6
57.2	0.54	0.130	0.63	0.95	0.116	0.047	0.016	8
57.1	0.49	-0.012	0.50	0.92	0.116	0.047	0.012	9
57.3	0.50	-0.024	0.46	0.93	0.047	0.000	0.000	10
57.4	0.48	-0.040	0.51	0.94	0.023	0.000	0.000	10

Table S1. Hopkins clinical panel challenge. Performance indices over all phenotypes for all patients. Seven performance scores and the median overall rank are shown. Predictions are sorted by the rank of the median among all indices.

Submission	PCC	SCC	KCC	RMSE	MAE	BACC 0.3	MCC 0.3	AUC 0.3	TNR 0.3	TPR 0.3	Overall Rank
43.1	0.39	0.45	0.31	0.50	0.39	0.65	0.36	0.74	0.91	0.38	1
47.1	0.39	0.45	0.31	0.57	0.43	0.68	0.36	0.74	0.75	0.61	2
47.2	0.39	0.45	0.31	0.57	0.43	0.68	0.37	0.74	0.71	0.66	2
46.1	0.35	0.41	0.28	0.53	0.40	0.67	0.34	0.72	0.74	0.60	4
41.1	0.37	0.38	0.27	0.58	0.47	0.60	0.30	0.70	0.97	0.23	5
41.2	0.37	0.38	0.27	0.58	0.47	0.60	0.30	0.70	0.97	0.23	5
44.3	0.25	0.37	0.26	0.57	0.43	0.60	0.20	0.69	0.44	0.75	7
39.2	0.31	0.37	0.26	0.73	0.54	0.67	0.34	0.68	0.77	0.56	8
40.1	0.28	0.36	0.25	0.63	0.46	0.68	0.35	0.72	0.71	0.64	8
44.4	0.27	0.32	0.22	0.53	0.41	0.67	0.34	0.66	0.77	0.56	8
44.2	0.21	0.37	0.26	0.73	0.51	0.64	0.30	0.69	0.78	0.51	11
39.1	0.32	0.36	0.25	0.77	0.59	0.63	0.30	0.67	0.85	0.42	12
44.1	0.26	0.32	0.22	0.65	0.48	0.62	0.27	0.66	0.81	0.44	12
42.1	0.13	0.18	0.13	0.65	0.48	0.58	0.17	0.60	0.46	0.71	14
42.2	0.14	0.21	0.15	0.65	0.48	0.58	0.17	0.61	0.44	0.72	14

Table S2. SUMO ligase challenge. Performance indices over all phenotypes. Seven performance scores and the median overall rank are shown. Predictions are sorted by the rank of the median among all indexes.

Submission	BACC 0.5	MCC 0.5	AUC 0.5	TNR 0.5	TPR 0.5	Overall Rank
10.3	0.68	0.358	0.72	0.77	0.59	1
10.1	0.68	0.358	0.72	0.77	0.59	2
10.5	0.67	0.33	0.71	0.72	0.61	3
10.4	0.63	0.256	0.66	0.68	0.58	4
1.4	0.62	0.24	0.6	0.77	0.47	5
10.2	0.62	0.24	0.63	0.77	0.47	5
3.2	0.62	0.235	0.64	0.66	0.58	7
1.1	0.58	0.224	0.66	0.21	0.94	8
2.1	0.59	0.243	0.59	0.26	0.92	8
6.1	0.58	0.232	0.59	0.96	0.2	10
10.6	0.61	0.217	0.65	0.74	0.47	10
1.6	0.59	0.201	0.57	0.4	0.78	12
3.5	0.57	0.15	0.62	0.43	0.72	13
1.5	0.57	0.157	0.6	0.36	0.78	14
3.3	0.56	0.125	0.62	0.38	0.73	15
11.3	0.56	0.121	0.55	0.36	0.75	16
6.2	0.54	0.106	0.61	0.85	0.23	17
3.4	0.55	0.105	0.62	0.45	0.66	18
11.2	0.54	0.088	0.54	0.51	0.58	19
6.4	0.54	0.086	0.59	0.28	0.8	20
6.5	0.54	0.086	0.6	0.28	0.8	20
6.6	0.54	0.086	0.59	0.28	0.8	20
3.1	0.53	0.056	0.6	0.45	0.61	23
8.3	0.54	0.077	0.54	0.47	0.61	23
6.3	0.53	0.078	0.61	0.83	0.23	25
8.1	0.52	0.052	0.48	0.36	0.69	26
3.6	0.51	0.029	0.51	0.34	0.69	27
7.1	0.51	0.025	0.5	0.45	0.58	27
1.3	0.51	0.035	0.52	0.15	0.88	29
11.4	0.51	0.013	0.51	0.34	0.67	30
8.4	0.51	0.012	0.49	0.57	0.44	31
5.1	0.5	0	0.45	1	0	32
5.2	0.5	0	0.44	1	0	32
5.3	0.5	0	0.41	0	1	32

5.5	0.5	0	0.5	1	0	32
8.5	0.5	-0.002	0.5	0.4	0.59	36
12.1	0.48	-0.037	0.5	0.45	0.52	37
9.1	0.49	-0.012	0.5	0.72	0.27	38
11.1	0.48	-0.033	0.44	0.4	0.56	39
1.2	0.46	-0.072	0.47	0.55	0.38	40
4.1	0.45	-0.098	0.46	0.34	0.56	41
5.4	0.48	-0.048	0.44	0.3	0.66	41
7.2	0.48	-0.143	0.52	0	0.95	43
8.2	0.47	-0.078	0.54	0.21	0.72	43
13.1	0.47	-0.078	0.55	0.21	0.72	43
14.1	0.35	-0.297	0.32	0.34	0.36	46

Table S3. Crohn's Disease challenge. Performance indices over all phenotypes. Seven performance scores and the median overall rank are shown. Predictions are sorted by the rank of the median among all indices.

Supplementary figures

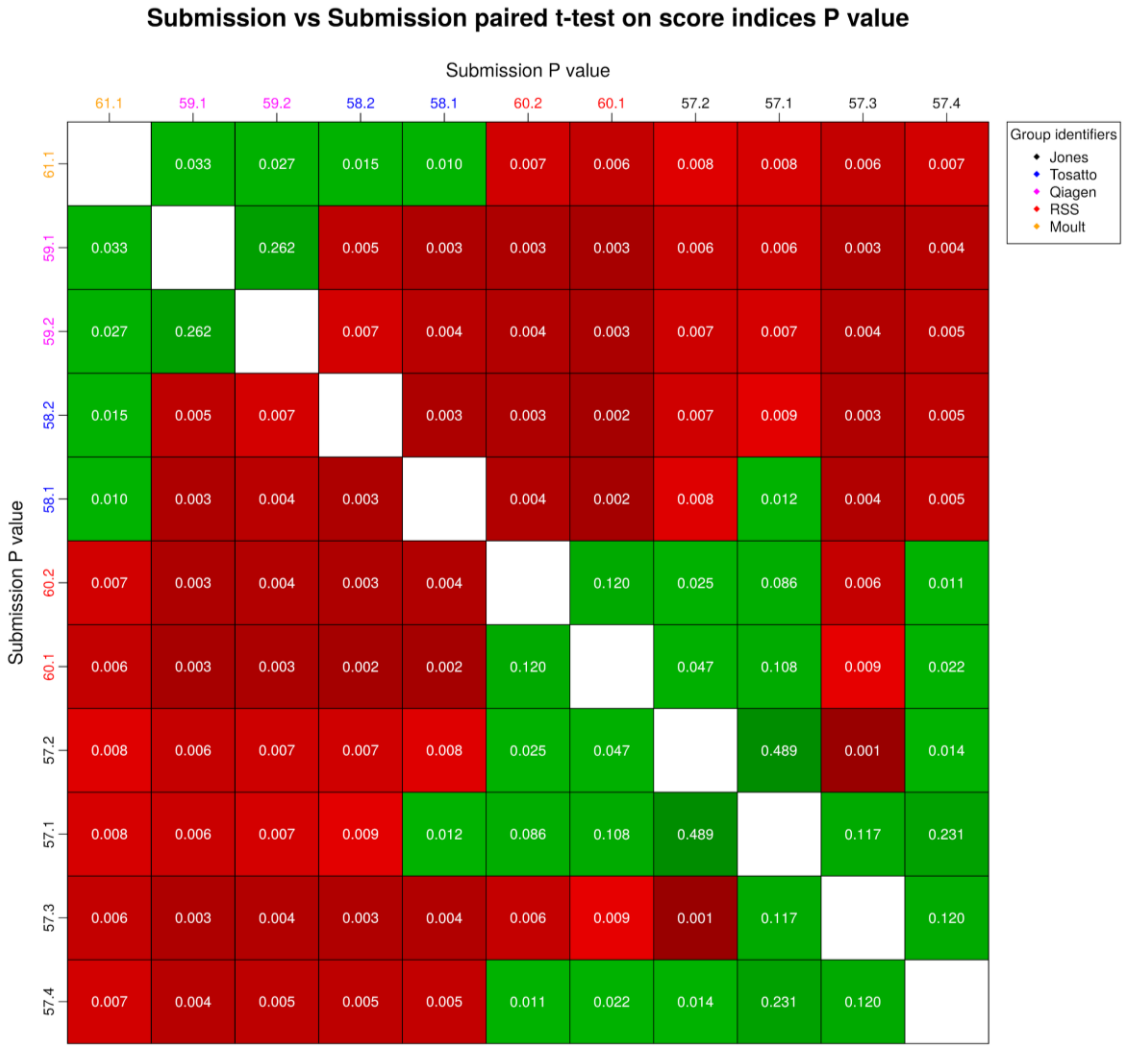


Figure S1. Pairwise difference between submissions for Hopkins challenge. Statistical differences between submissions based on the mean score obtained by each submission over all indices, sorted according to the final ranking. Green squares are indices of tied predictions (P values >= 0.01) meaning that according to the performance indices used, the difference between two predictors is not statistically significant. White squares represents tied predictions with P equal to 1.

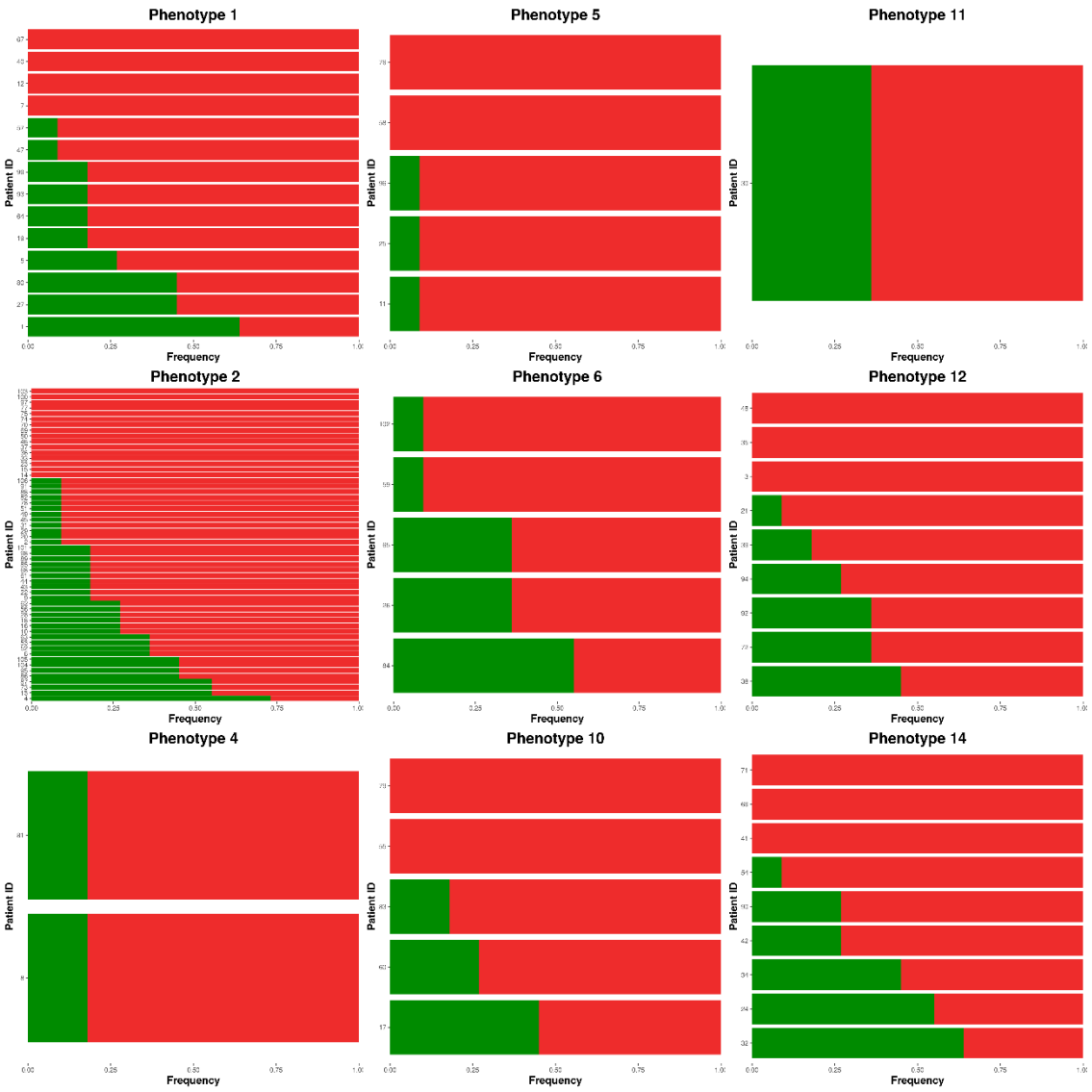


Figure S2. Outliers barplot for all patents in Hopkins challenge. For each patient, the frequency of correct predictions [0, 1] by all the methods is reported in green, while red represents frequency of wrong predictions. Patients are ordered by phenotype and frequency of correct predictions. Red bars are patients for which no method was able to correctly detect the phenotype.

Figure S3. Pairwise difference between submissions for SUMO ligase challenge, subset 1. Statistical differences between submissions based on the mean score obtained by each submission over all indices, sorted according to the final ranking. Green squares are indices of tied predictions (P values >= 0.01) meaning that according to the performance indices used, the difference between two predictors is not statistically significant. White squares represents tied predictions with P equal to 1.

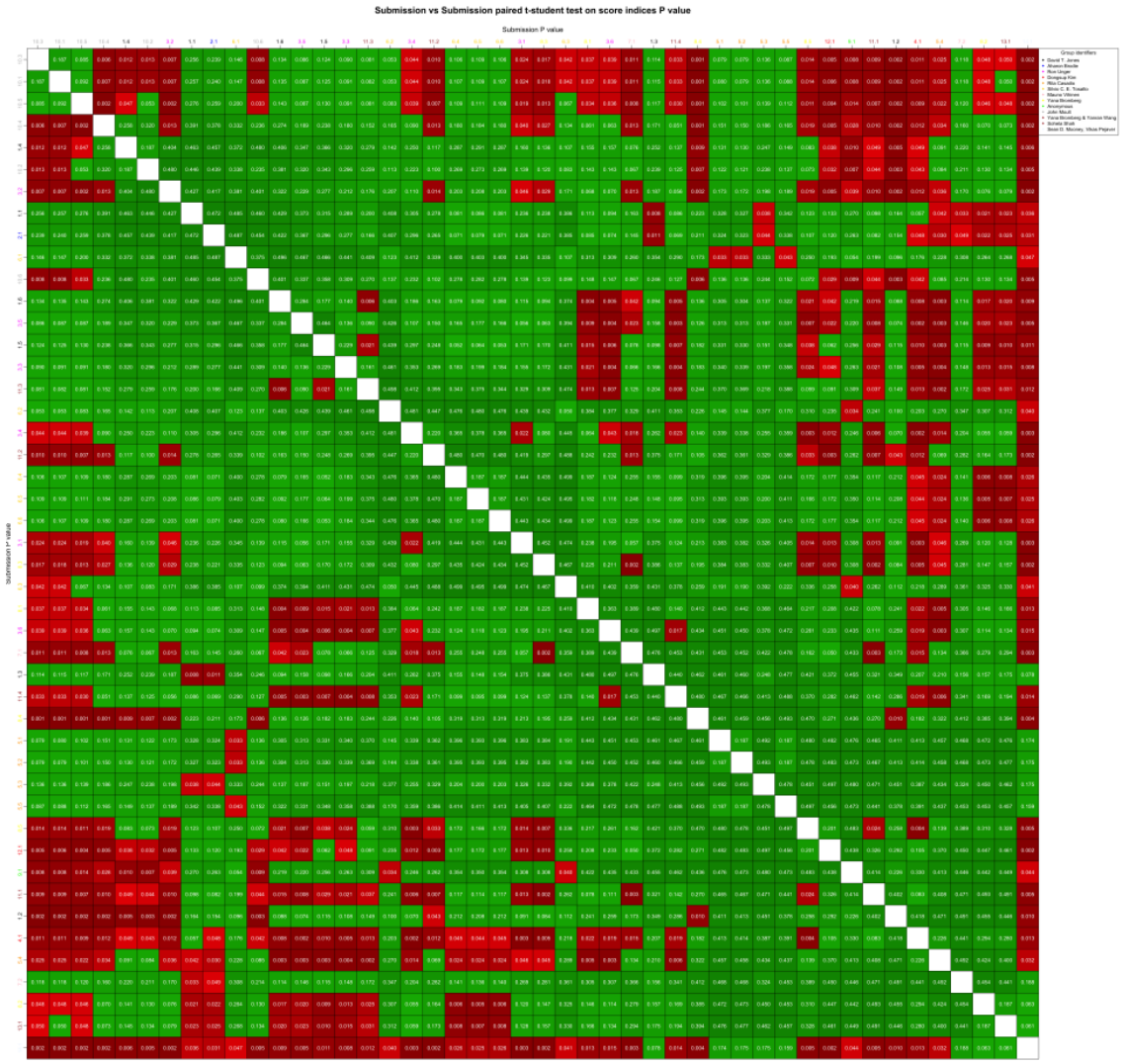


Figure S4. Pairwise difference between submissions for Crohn's disease challenge. Statistical differences between submissions based on the mean score obtained by each submission over all indices, sorted according to the final ranking. Green squares are indices of tied predictions (P values >= 0.01) meaning that according to the performance indices used, the difference between two predictors is not statistically significant. White squares represents tied predictions with P equal to 1.

Appendix 2

Supplementary tables

Phenotypic traits	Description
Intellectual disability (ID)	Intellectual disability is characterized by significant limitations in both intellectual functioning and in adaptive behavior, which covers many everyday social and practical skills. This disability originates during development. Intellectual functioning (also called intelligence) refers to general mental capacity, such as learning, reasoning, problem solving, and so on. Adaptive behavior is the collection of conceptual, social, and practical skills that are learned and performed by people in their everyday lives.
Autism spectrum disorder (ASD)	This term can be used to refer to autism spectrum disorder as a phenotypic feature that can be a component of a disease. A disorder beginning in childhood, it is marked by the presence of markedly abnormal or impaired development in social interaction and communication and a markedly restricted repertoire of activity and interest. Manifestations of the disorder vary greatly depending on the developmental level and chronological age of the individual (DSM-IV). Autism spectrum disorders range from a severe form, called autistic disorder, to a milder form, Asperger syndrome.
Epilepsy	Seizures are an intermittent abnormality of the central nervous system due to a sudden, excessive, disorderly discharge of cerebral neurons characterized clinically by some combination of disturbance of sensation, loss of consciousness, impairment of psychic function, or convulsive movements. The term epilepsy is used to describe chronic, recurrent seizures.
Microcephaly	Occipitofrontal (head) circumference less than -3 standard deviations compared to appropriate, age matched, normal standards (Potter 1978). Alternatively, decreased size of the cranium.

Macrocephaly	Occipitofrontal (head) circumference greater than the 97th centile compared to appropriate, age matched, sex-matched normal standards. Alternatively, an apparently increased size of the cranium.
Hypotonia	Muscular hypotonia (abnormally low muscle tone) manifesting in infancy.
Ataxic gait	A type of ataxia characterized by impairment of the ability to coordinate movements required for normal walking. Gait ataxia is characterized by a wide-based staggering gait with a tendency to fall.

Table S4. Summary of the seven diseases classes in the CAGI-5 intellectual disability challenge.

CAGI-ID	Name	Sex	Class of variant	Chr:POS:REF:ALT	Exonic Function	AAChange	Mode of inheritance	Inheritance	Number of Groups	Number of predictions	Group ID	Predictions ID
135	MR2409	F	Contributing factor	chr22:51117766:T:G	nonsynonymous	SHANK3:NM_033517:exon7:c.T795G:p.H265Q	AD	paternal	4	11	1, 2, 4, 5	1.1, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 4.1, 4.2, 4.3, 5.1
97	MR2250	M	Contributing factor	chr14:21863113:G:C	nonsynonymous	CHD8:NM_001170629:exon29:c.C5348G:p.A1783G	AD	paternal	3	9	2, 4, 5	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 4.1, 4.3, 5.1
125	MR2368	M	Contributing factor	chr3:11078549:G:A	nonsynonymous	SLC6A1:NM_003042:exon16:c.G1697A:p.R566H	AD	paternal	3	9	2, 4, 5	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 4.1, 4.3, 5.1
83	MR2189	M	Contributing factor	chr17:8402701:C:G	nonsynonymous	MYH10:NM_001256012:exon30:c.G3838C:p.E1280Q	AD	paternal	3	8	2, 4, 5	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 4.3, 5.1

17	MR1635	M	Contributing factor	chr11:70653140:C:T	unknown	SHANK2:NM_012309.3:c.G1484A:p.E544K	AD	materna	3	8	2, 4, 5	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 4.3, 5.1
112	MR2337	F	Contributing factor	chr11:70653124:C:T	nonsynonymous	SHANK2:NM_012309.4:c.G1646A:p.R549H	AD	girl adopted	2	8	2, 4	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 4.1, 4.3
6	MR1289	M	Contributing factor	chr14:21860898:C:T	nonsynonymous	CHD8:NM_001170629:exon33:c.G6539A:p.R2180H	AD	maternal	1	6	2	2.1, 2.2, 2.3, 2.4, 2.5, 2.6
47	MR2033	M	Contributing factor	chr7:146829601:G:A	nonsynonymous	CNTNAP2:NM_014141:exon8:c.G1348A:p.G450S	AR, AD	maternal	3	6	1, 3, 4	1.1, 3.1, 3.2, 3.3, 4.1, 4.3
34	MR1980	F	Contributing factor	chr7:103243828:C:A	nonsynonymous	RELN:NM_005045:exon24:c.G3256T:p.V1086F	AR, AD	paternal, paternal grandmother	1	6	2	2.1, 2.2, 2.3, 2.4, 2.5, 2.6
94	MR2241	M	Contributing factor	chr7:103130201:C:T	nonsynonymous	RELN:NM_005045:exon60:c.G9751A:p.E3251K	AR, AD	paternal	1	6	2	2.1, 2.2, 2.3, 2.4, 2.5, 2.6
127	MR2375	F	Contributing factor	chr11:70319321:G:A	nonsynonymous	SHANK2:NM_133266:exon11:c.C3439T:p.P1147S	AD	n.d.	3	5	3, 4, 5	3.1, 3.2, 3.3, 4.3, 5.1
12	MR1543	M	Contributing factor	chr7:148112649:A:C	nonsynonymous	CNTNAP2:NM_014141:exon24:c.A3937C:p.N1313H	AR, AD	maternal	2	4	3, 4	3.1, 3.2, 3.3, 4.3
17	MR1635	M	Contributing factor	chr5:14471497:C:T	nonsynonymous	TRIO:NM_007118:exon38:c.C5834T:p.S1945L	AD	maternal	1	1	1	1,1

105	MR2276	M	Disease causing	chrX:76909661:T:C	nonsynonymous	ATRX:NM_138270:exon13:c.A4130G:p.N1377S	XLD, XLR	maternal; X-inactivation: balanced	5	14	1, 2, 3, 4, 5	1.1, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 3.1, 3.2, 3.3, 4.1, 4.2, 4.3, 5.1
140	MR414	F	Disease causing	chrX:153296777:G:A	stopgain	MECP2:NM_001110792:exon3:c.C538T:p.R180X	XLD	n.d.	4	12	2, 3, 4, 5	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 3.2, 3.3, 4.1, 4.2, 4.3, 5.1
79	MR2166	M	Disease causing	chr9:140728837:G:C	nonsynonymous	EHMT1:NM_024757:exon26:c.G3577C:p.G1193R	AD	de novo	4	11	1, 2, 3, 5	1.1, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 3.1, 3.2, 3.3, 5.1
142	MR602	F	Disease causing	chrX:41401980:G:A	stopgain	CASK:NM_003688:exon22:c.C2119T:p.Q707X	XL	de novo	3	10	2, 4, 5	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 4.1, 4.2, 4.3, 5.1
64	MR2113	M	Disease causing	chr12:116445337:C:T	nonsynonymous	MED13L:NM_015335:exon11:c.G2117A:p.G706E	AD	De novo	3	10	1, 2, 3	1.1, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 3.1, 3.2, 3.3
90	MR2233	M	Disease causing	chr6:33411228:C:T	stopgain	SYNGAP1:NM_006772:exon15:c.C2899T:p.R967X	AD	de novo	4	10	2, 3, 4, 5	2.1, 2.2, 2.3, 2.4,

												2.5, 2.6, 3.2, 3.3, 4.3, 5.1
69	MR2127	F	Disease causing	chr6:157528165:G:T	stopgain	ARID1B:NM_001346813:exon20:c.G6010T:p.E2004X	AD	de novo	3	9	2, 3, 5	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 3.2, 3.3, 5.1
72	MR2140	M	Disease causing	chrX:122460015:G:A	nonsynonymous	GRIA3:NM_000828:exon4:c.G647A:p.R216Q	XL	maternal, X-inactivation: mutated allele 30%	4	9	1, 2, 4, 5	1.1, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 4.3, 5.1
35	MR1985	M	Disease causing	chr2:200213882:G:A	stopgain	SATB2:NM_001172509:exon7:c.C715T:p.R239X	AD	de novo	3	9	2, 3, 5	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 3.2, 3.3, 5.1
104	MR2274	M	Disease causing	chr18:42531498:AAGAGC:A	frameshift deletion	SETBP1:NM_015559:exon4:c.2194_2198del:p.R732fs	AD	de novo	3	9	2, 4, 5	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 4.1, 4.3, 5.1
23	MR1749	M	Disease causing	chr22:51160432:GA:G	frameshift deletion	SHANK3:NM_033517:exon21:c.4130delA:p.E1377fs	AD	de novo	3	9	2, 4, 5	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 4.1, 4.2, 5.1
106	MR2278	M	Disease causing	chr12:13761626:T:G	nonsynonymous	GRIN2B:NM_000834:exon9:c.A1921C:p.I641L	AD	de novo	3	8	2, 4, 5	2.1, 2.2, 2.3, 2.4,

												2.5, 2.6, 4.3, 5.1
89	MR2230	F	Disease causing	chr22:51153476:G:A	splicing	SHANK3:NM_033517:exon19:c.2223+1G>A	AD	de novo	3	8	2, 4, 5	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 4.3, 5.1
150	MR984	M	Disease causing	chr5:14390392:C:T	nonsynonymous	TRIO:NM_007118:exon26:c.C4111T:p.H1371Y	AD	de novo	3	8	1, 2, 5	1.1, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 5.1
96	MR2243	M	Disease causing	chr9:140657209:GA:G	frameshift deletion	EHMT1:NM_024757:exon10:c.1585delA:p.S529fs	AD	de novo	2	7	2, 5	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 5.1
41	MR2019	M	Disease causing	chr12:13724822:C:T	nonsynonymous	GRIN2B:NM_000834:exon10:c.G2087A:p.R696H	AD	de novo	2	7	2, 5	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 5.1
21	MR1730	F	Disease causing	chrX:67273488:C:T	nonsynonymous	OPHN1:NM_002547:exon22:c.G2323A:p.V775M	XLR	De novo	4	6	1, 3, 4, 5	1.1, 3.1, 3.2, 3.3, 4.3, 5.1
32	MR1974	M	Disease causing	chrX:154490151:A:C	nonsynonymous	RAB39B:NM_171998:exon2:c.T579G:p.F193L	XLR	maternal, mother affected	3	3	1, 4, 5	1.1, 4.3, 5.1
64	MR2113	M	Disease causing	chr1:155449342:T:C	nonsynonymous	ASH1L:NM_018489:exon3:c.A3319G:p.I1107V	AD	de novo	1	2	4	4.1, 4.3
87	MR2222	M	Disease causing	chr21:38858777:G:C	nonsynonymous	DYRK1A:NM_101395:exon7:c.G525C:p.K175N	AD	de novo	2	2	1, 5	1.1, 5.1
113	MR2338	F	Disease causing	chr16:89346136:CAG:C	frameshift deletion	ANKRD11:NM_013275:exon9:c.6812_6813del:p.P2271fs	AD	n.d.	1	1	5	5,1

47	MR2033	M	Disease causing	chr16:89345974:CCTTCGGGG:C	frameshift deletion	ANKRD11:NM_013275:exon9:c.6968_6975del:p.A2323fs	AD	de novo	1	1	5	5,1
102	MR2271	M	Disease causing	chr22:51159718:C:T	stopgain	SHANK3:NM_033517:exon21:c.C3415T:p.R1139X	AD	de novo	1	1	5	5,1
78	MR2165	M	Disease causing	chr5:14394159:C:T	stopgain	TRIO:NM_007118:exon28:c.C4231T:p.R1411X	AD	maternal, mother affected	1	1	5	5,1
31	MR1970	F	Disease causing	chr22:51159830:A:TTC	frameshift delins	SHANK3:NM_033517:exon21:c.3527delinsTTC:p.D1176fs	AD	de novo	0	0		
48	MR2039	M	Putative	chr16:3788561:C:T	nonsynonymous	CREBBP:NM_004380:exon26:c.G4393A:p.G1465R	AD	n.d.	5	14	1, 2, 3, 4, 5	1.1, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 3.1, 3.2, 3.3, 4.1, 4.2, 4.3, 5.1
109	MR2322	M	Putative	chrX:76764055:T:A	nonsynonymous	ATRX:NM_000489:exon35:c.A7253T:p.Y2418F	XLD, XLR	n.d.	4	13	2, 3, 4, 5	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 3.1, 3.2, 3.3, 4.1, 4.2, 4.3, 5.1
103	MR2272	M	Putative	chr10:89690828:G:A	nonsynonymous	PTEN:NM_000314:exon4:c.G235A:p.A79T	AD	maternal	3	12	2, 3, 4	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 3.1, 3.2, 3.3, 4.1, 4.2, 4.3

24	MR1769	M	Putative	chr3:71026867:A:C	nonsynonymous	FOXP1:NM_032682:exon16:c.T1355G:p.I452S	AD	paternal	3	8	1, 2, 5	1.1, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 5.1
24	MR1769	M	Putative	chr7:146829502:G:T	nonsynonymous	CNTNAP2:NM_014141:exon8:c.G1249T:p.D417Y	AR, AD	maternal	1	3	3	3.1, 3.2, 3.3
114	MR2340	M	Putative	chr2:166165900:C:T	nonsynonymous	SCN2A:NM_021007:exon6:c.C644T:p.A215V	AD	maternal, familial epilepsy	3	8	1, 2, 4	1.1, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 4.3
40	MR2007	M	Putative	chr11:70644598:G:A	nonsynonymous	SHANK2:NM_012309:exon13:c.C1727T:p.P576Q	AD	n.d.	3	8	2, 4, 5	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 4.3, 5.1
30	MR196	M	Putative	chrX:41448842:A:G	nonsynonymous	CASK:NM_003688:exon13:c.T1159C:p.Y387H	XL	n.d.	3	5	1, 4, 5	1.1, 4.1, 4.2, 4.3, 5.1
99	MR2264	M	Putative	chr16:89349967:T:C	nonsynonymous	ANKRD11:NM_013275:exon9:c.A2983G:p.K995E	AD	n.d.	2	7	1, 2	1.1, 2.1, 2.2, 2.3, 2.4, 2.5, 2.6
73	MR2141	M	Putative	chr14:21876977:G:A	nonsynonymous	CHD8:NM_001170629:exon11:c.C2372T:p.P791L	AD	Not in the mother	2	7	2, 4	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 4.3
130	MR2389	F	Putative	chr14:21882498:T:C	nonsynonymous	CHD8:NM_001170629:exon8:c.A2104G:p.K702E	AD	n.d.	2	7	2, 5	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 5.1

127	MR2375	F	Putative	chr11:684897:C:T	splicing	DEAF1:NM_021008:exon6:c.870+1G>A	AR, AD	n.d.	2	7	2, 5	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 5.1
5	MR1192	F	Putative	chr12:13720096:C:G	nonsynonymous	GRIN2B:NM_000834:exon12:c.G2461C:p.V821L	AD	n.d.	2	7	2, 5	2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 5.1
33	MR1975	M	Putative	chr18:44595922:C:T	nonsynonymous	KATNAL2:NM_031303:exon10:c.C743T:p.A248V	AD	maternal, gene with low penetrance	1	6	2	2.1, 2.2, 2.3, 2.4, 2.5, 2.6
126	MR2374	F	Putative	chr7:148080864:C:T	nonsynonymous SNV	CNTNAP2:NM_014141:exon22:c.C3599T:p.S1200L	AR, AD	n.d.	2	4	3, 4	3.1, 3.2, 3.3, 4.3
116	MR2344	M	Putative	chrX:53964467:A:G	nonsynonymous SNV	PHF8:NM_001184897:exon22:c.T2794C:p.C932R	XLR	maternal, X-inactivation: mutated allele 70%	1	3	4	4.1, 4.2, 4.3
56	MR2053	F	Putative	chr2:171702114:C:T	nonsynonymous SNV	GAD1:NM_000817:exon8:c.C850T:p.L284F	AR	paternal	0	0		
56	MR2053	F	Putative	chr2:171678594:T:C	splice region	GAD1:NM_013445.3:c.83-3T>C	AR	maternal	0	0		

Table S5. **Causative experimentally identified variants and groups predictions**

Appendix 3

Supplementary tables

Submission 1.1			
	Obs. Loss of function	Obs. Hypomorphic	Obs. Benign
Pred. Loss of function	4	3	9
Pred. Hypomorphic	6	6	7
Pred. Benign	2	1	0
Submission 2.1			
	Obs. Loss of function	Obs. Hypomorphic	Obs. Benign
Pred. Loss of function	3	2	3
Pred. Hypomorphic	2	1	3
Pred. Benign	7	7	10
Submission 4.1			
	Obs. Loss of function	Obs. Hypomorphic	Obs. Benign
Pred. Loss of function	0	0	1
Pred. Hypomorphic	6	4	1

Pred. Benign	6	6	14
Submission 4.2			
	Obs. Loss of function	Obs. Hypomorphic	Obs. Benign
Pred. Loss of function	0	0	0
Pred. Hypomorphic	1	0	1
Pred. Benign	11	10	15
Submission 5.1			
	Obs. Loss of function	Obs. Hypomorphic	Obs. Benign
Pred. Loss of function	2	1	1
Pred. Hypomorphic	5	4	13
Pred. Benign	5	5	2
Submission 6.1			
	Obs. Loss of function	Obs. Hypomorphic	Obs. Benign
Pred. Loss of function	1	3	3
Pred. Hypomorphic	5	1	3
Pred. Benign	6	6	10

Table S6: Confusion matrices for all submissions considering the three categories of variant.

Supplementary figures

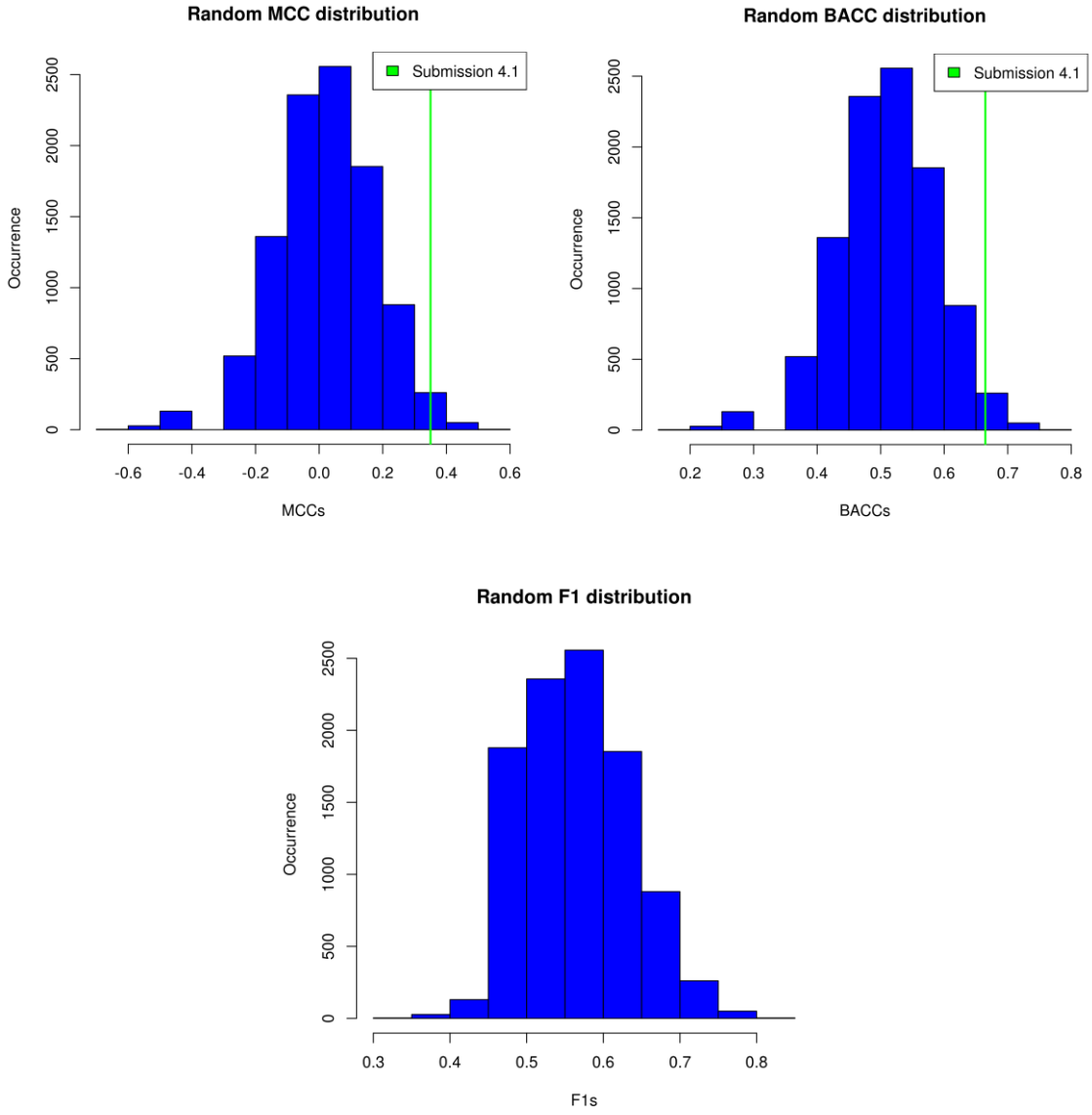


Figure S5: Random distributions of MCC, BACC and F_1 scores. Submission 4.1 score is highlighted in green.

Appendix 4

Supplementary tables

ID	Sex	C-HDL	C-LDL	Unit	Model reaction	Gene	Allele	Mutation	PubMed ID
1	F	1.51	7.53	mmol/L	5,7	LDLR	heterozygous	Cys667Leufs*6	25846081
2	F	0.94	19.84	mmol/L	5,7	LDLR	heterozygous	Trp419*	25846081
3	M	1.81	12.62	mmol/L	5,7	LDLR	heterozygous	Trp483*	25846081
4	M	2.05	7.83	mmol/L	5,7	LDLR	heterozygous	Trp483*	25846081
5	F	0.77	11.2	mmol/L	5,7	LDLR	heterozygous	Cys329Tyr	25846081
6	F	0.86	12.37	mmol/L	5,7	LDLR	heterozygous	His583Tyr	25846081
7	M	0.75	19.28	mmol/L	5,7	LDLR	heterozygous	Val806Glyfs*11	25846081
8	F	1.91	7.61	mmol/L	5,7	LDLR	heterozygous	His583Tyr	23155708
9	M	1.29	7.15	mmol/L	5,7	LDLR	heterozygous	Pro658Leu	23155708
10	M	1.26	4.98	mmol/L	5,7	LDLR	heterozygous	His583Tyr	23155708
11	F	1.96	4.97	mmo	5,7	LDLR	heterozy	Pro658Leu	2315570

				l/L			gous		8
12	F	1.07	5.49	mmol/L	5,7	LDLR	heterozygous	Cys184Phefs*21	23155708
13	M	1.09	4.99	mmol/L	5,7	LDLR	heterozygous	Cys184Phefs*21	23155708
14	F	0.83	3.38	mmol/L	5,7	APOB	heterozygous	Arg3500Gln	18222178
15	M	1.04	5.45	mmol/L	5,7	APOB	heterozygous	Arg3500Gln	18222178
16	M	0.98	4.02	mmol/L	5,7	APOB	heterozygous	Arg3500Gln	18222178
17	M	1.58	3.82	mmol/L	5,7	APOB	heterozygous	Arg3500Gln	10998466
18	M	1.09	3.28	mmol/L	5,7	APOB	heterozygous	Arg3500Gln	10998466
19	M	0.65	3.74	mmol/L	5,7	APOB	heterozygous	Arg3500Gln	10998466
20	F	41	240	mg/dL	5,7	APOB	heterozygous	Arg3500Gln	15135245
21	M	36	306	mg/dL	5,7	APOB	homozygous	Arg3500Gln	28988723
22	M	4	154	mg/dL	8,16,17	ABCA1	heterozygous	Asp1009Tyr, Phe2009Ser	12009425
23	F	38	88	mg/dL	8,16,17	ABCA1	heterozygous	Asp1009Tyr	12009425
24	F	17	111	mg/d	8,16,17	ABCA	heterozy	Asp1009T	1200942

				L		1	gous	yr	5
25	M	30	109	mg/dL	8,16,17	ABCA1	heterozygous	Phe2009Ser	12009425
26	M	35	117	mg/dL	8,16,17	ABCA1	heterozygous	Asp1009Tyr	12009425
27	F	27	92	mg/dL	8,16,17	ABCA1	heterozygous	Asp1009Tyr	12009425
28	F	43	154	mg/dL	8,16,17	ABCA1	heterozygous	Asp1009Tyr	12009425
29	M	1.2	5.5	mmol/L	5,7	APOE	heterozygous	Arg269Gly	22949395
30	M	1.5	7.1	mmol/L	5,7	APOE	heterozygous	pLeu167del	22949395
31	M	1.8	5.7	mmol/L	5,7	APOE	heterozygous	pLeu167del	22949395
32	F	2.4	9.4	mmol/L	5,7	APOE	heterozygous	pLeu167del	22949395
33	F	1.5	5.6	mmol/L	5,7	APOE	heterozygous	pLeu167del	22949395
34	M	1.8	11.8	mmol/L	5,7	APOE	heterozygous	pLeu167del	22949395
35	F	1.4	6.5	mmol/L	5,7	APOE	heterozygous	pLeu167del	22949395
36	M	1.6	7.7	mmol/L	5,7	APOE	heterozygous	pLeu167del	22949395
37	F	1.9	4	mmol/L	5,7	APOE	heterozygous	pLeu167del	22949395

38	M	1.6	9.7	mmol/L	5,7	APOE	heterozygous	pLeu167del	24267230
39	F	1.6	5.3	mmol/L	5,7	APOE	heterozygous	pLeu167del	24267230
40	F	1.2	3.8	mmol/L	5,7	APOE	heterozygous	pLeu167del	24267230
41	NA	62	65	mg/dL	21	CETP	heterozygous	Asp442Gly	19463799
42	F	0.52	1.34	mmol/L	9	LCAT	homozygous	Val309Met	16051254
43	M	0.48	1.76	mmol/L	9	LCAT	homozygous	Val309Met	16051254
44	M	0.13	0.74	mmol/L	9	LCAT	homozygous	His35Thrfs*26	19515369
45	M	0.22	1.69	mmol/L	9	LCAT	homozygous	His35Thrfs*26	19515369
46	M	7	41	mg/dL	9	LCAT	homozygous	Met293Arg	22108153
47	F	5	223.9	mg/dL	9	LCAT	homozygous	Gly119Asp	28942093
48	F	5.03	141.76	mg/dL	9	LCAT	homozygous	Gly119Asp	28942093
49	F	0.94	3.8	mmol/L	9	LCAT	heterozygous	Val309Met	16051254
50	F	1.09	2.42	mmol/L	9	LCAT	heterozygous	Val309Met	16051254
51	F	2.25	2.93	mmol/L	9	LCAT	heterozygous	Val309Met	16051254

52	M	0.84	3.13	mmol/L	9	LCAT	heterozygous	Val309Met	16051254
53	F	1.42	3.12	mmol/L	9	LCAT	heterozygous	Val309Met	16051254
54	M	1.11	3.12	mmol/L	9	LCAT	heterozygous	Val309Met	16051254
55	F	1.12	2.58	mmol/L	9	LCAT	heterozygous	Val309Met	16051254
56	F	0.86	1.51	mmol/L	9	LCAT	heterozygous	Val309Met	16051254
57	F	0.96	2.8	mmol/L	9	LCAT	heterozygous	Val309Met	16051254
58	F	0.81	1.47	mmol/L	9	LCAT	heterozygous	Val309Met	16051254
59	M	1.28	1.36	mmol/L	9	LCAT	heterozygous	Val309Met	16051254
60	F	1.42	2.47	mmol/L	9	LCAT	heterozygous	His35Thrfs*26	19515369
61	M	1.07	2.76	mmol/L	9	LCAT	heterozygous	His35Thrfs*26	19515369
62	M	1.09	1.93	mmol/L	9	LCAT	heterozygous	His35Thrfs*26	19515369
63	F	1.38	2.1	mmol/L	9	LCAT	heterozygous	His35Thrfs*26	19515369
64	M	0.93	2.85	mmol/L	9	LCAT	heterozygous	His35Thrfs*26	19515369
65	F	1.28	1.38	mmol/L	9	LCAT	heterozygous	His35Thrfs*26	19515369

66	F	107	197	mg/dL	18	CYP7A1	homozygous	Leu413fs*1	12093894
67	M	29	151	mg/dL	18	CYP7A1	homozygous	Leu413fs*1	12093894

Table S7. Training set of (Reggiani et al., 2018). Patient ID, sex, HDL and LDL cholesterol, model rates involved, affected gene, variant and PubMed ID of the related publication.

ID	Sex	TC (mg/dL)	Model reaction	Gene	Mutation 1	Mutation 2	PubMed ID
1	F	28	1,2,3	DHCR7	Thr93Met	IVS8-1GC	10814720
2	F	6	1,2,3	DHCR7	Arg469Pro	Glu448Lys	10814720
3	M	15	1,2,3	DHCR7	Arg469Pro	NA	10814720
4	F	29	1,2,3	DHCR7	Thr93Met	IVS8-1GC	10814720
5	F	80	1,2,3	DHCR7	Phe168Ile	IVS8-1GC	10814720
6	M	29	1,2,3	DHCR7	Leu148Arg	Pro179Leu	10814720
7	F	132	1,2,3	DHCR7	Tyr324His	IVS8-1GC	10814720
8	M	38	1,2,3	DHCR7	Thr93Met	IVS8-1GC	10814720
9	F	58	1,2,3	DHCR7	Tyr462His	IVS8-1GC	10814720

10	F	28.7	1,2,3	DHCR 7	Val326 Leu	IVS8-1GC	10814720
11	F	38.7	1,2,3	DHCR 7	Pro243 Arg	Val326 Leu	10814720
12	F	87	1,2,3	DHCR 7	Arg404 Ser	IVS8-1GC	10814720
13	F	88	1,2,3	DHCR 7	Arg352 Trp	IVS8-1GC	10814720
14	F	65	1,2,3	DHCR 7	Arg352 Trp	IVS8-1GC	10814720
15	M	14	1,2,3	DHCR 7	Val326 Leu	IVS8-1GC	10814720
16	F	77.6	1,2,3	DHCR 7	Thr93M et	Thr93 Met	10814720
17	F	120	1,2,3	DHCR 7	Thr93M et	Val326 Leu	10814720
18	F	46	1,2,3	DHCR 7	Thr93M et	IVS8-1GC	10814720
19	M	34.4	1,2,3	DHCR 7	Arg352 Trp	IVS8-1GC	10814720
20	M	65.5	1,2,3	DHCR 7	Asp175 His	IVS8-1GC	10814720
21	F	90	1,2,3	DHCR 7	Ser169 Leu	IVS8-1GC	10814720
22	M	46	1,2,3	DHCR 7	Phe302 Leu	IVS8-1GC	10814720
23	M	9	1,2,3	DHCR 7	Glu37*	NA	10814720

24	F	125	1,2,3	DHCR7	Phe284Leu	Val326Leu	10814720
25	M	88	1,2,3	DHCR7	Glu448Lys	Pro51Ser	10814720
26	F	24	1,2,3	DHCR7	Thr93Met	Gly410Ser	10814720
27	F	26	1,2,3	DHCR7	Asn287Lys	IVS8-1GC	10814720
28	M	65	1,2,3	DHCR7	Asn287Lys	IVS8-1GC	10814720
29	M	8.5	1,2,3	DHCR7	Leu99Pro	IVS8-1GC	10814720
30	M	12	1,2,3	DHCR7	Asn287Lys	Pro243Arg	10814720

Table S8. Training set of (Reggiani et al., 2018). Patient ID, sex, total cholesterol (mg/dL), model rates, affected gene, variants and related PubMed ID.