UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Biomediche

CORSO DI DOTTORATO DI RICERCA IN : SCIENZE BIOMEDICHE SPERIMENTALI

CICLO: XXX

# Development of bioinformatics tools to predict disease predisposition from Next Generation Sequencing (NGS) data.

**Coordinatore:** Ch.mo Prof. Paolo Bernardi

**Supervisore:** Ch.mo Prof. Silvio C.E. Tosatto

**Dottorando:** Marco Carraro

# Index

# List of publications

## Journal articles

**Carraro, M.,** Minervini, G., Giollo, M., Bromberg, Y., Capriotti, E., Casadio, R., Dunbrack, R., Elefanti, L., Fariselli, P., Ferrari, C., Gough, J., Katsonis, P., Leonardi, E., Lichtarge, O., Menin, C., Martelli, P.L., Niroula, A., Pal, L.R., Repo, S., Scaini, M.C., Vihinen, M., Wei, Q., Xu, Q., Yang, Y., Yin, Y., Zaucha, J., Zhao, H., Zhou, Y., Brenner, S.E., Moult, J., Tosatto, S.C.E., 2017. *Performance of in silico tools for the evaluation of p16INK4a (CDKN2A) variants in CAGI*. Hum. Mutat. doi:10.1002/humu.23235

Daneshjou, R., Wang, Y., Bromberg, Y., Bovo, S., Martelli, P.L., Babbi, G., Lena, P.D., Casadio, R., Edwards, M., Gifford, D., Jones, D.T., Sundaram, L., Bhat, R., Li, X., Pal, L.R., Kundu, K., Yin, Y., Moult, J., Jiang, Y., Pejaver, V., Pagel, K.A., Li, B., Mooney, S.D., Radivojac, P., Shah, S., **Carraro, M**., Gasparini, A., Leonardi, E., Giollo, M., Ferrari, C., Tosatto, S.C.E., Bachar, E., Azaria, J.R., Ofran, Y., Unger, R., Niroula, A., Vihinen, M., Chang, B., Wang, M.H., Franke, A., Petersen, B.-S., Pirooznia, M., Zandi, P., McCombie, R., Potash, J.B., Altman, R.B., Klein, T.E., Hoskins, R.A., Repo, S., Brenner, S.E., Morgan, A.A., 2017. *Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges.* Hum. Mutat. doi:10.1002/humu.23280

Giollo, M., Jones, D.T., **Carraro, M**., Leonardi, E., Ferrari, C., Tosatto, S.C.E., 2017. *Crohn disease risk prediction-Best practices and pitfalls with exome data.* Hum. Mutat. doi:10.1002/humu.23177

Chandonia, J.-M., Adhikari, A., **Carraro, M.,** Chhibber, A., Cutting, G.R., Fu, Y., Gasparini, A., Jones, D.T., Kramer, A., Kundu, K., Lam, H.Y.K., Leonardi, E., Moult, J., Pal, L.R., Searls, D.B., Shah, S., Sunyaev, S., Tosatto, S.C.E., Yin, Y., Buckley, B.A., 2017. *Lessons from the CAGI-4 Hopkins clinical panel challenge.* Hum. Mutat. doi:10.1002/humu.23225

Cai, B., Li, B., Kiga, N., Thusberg, J., Bergquist, T., Chen, Y.-C., Niknafs, N., Carter, H., Tokheim, C., Beleva-Guthrie, V., Douville, C., Bhattacharya, R., Yeo, H.T.G., Fan, J., Sengupta, S., Kim, D., Cline, M., Turner, T., Diekhans, M., Zaucha, J., Pal, L.R., Cao, C., Yu, C.-H., Yin, Y., **Carraro, M**., Giollo, M., Ferrari, C., Leonardi, E., Tosatto, S.C.E., Bobe, J., Ball, M., Hoskins, R.A., Repo, S., Church, G., Brenner, S.E., Moult, J., Gough, J., Stanke, M., Karchin, R., Mooney, S.D., 2017. *Matching phenotypes to whole genomes: Lessons learned from four iterations of the personal genome project community challenges.* Hum. Mutat. doi:10.1002/humu.23265

## Book Chapters

Carraro M., Tosatto S.C.E, Rizzuto R., *The origin of personalized medicine and the systems biology revolution,* Chapter of the book "P5 Medicine & Justice" (accepted)

# Conference Abstracts/Posters/Oral presentations

**Carraro M.,** Gasparini A., Leonardi E. and Tosatto S.C.E., CAGI -4 Bipolar disorder challenge, CAGI 4 2016, Critical Assessment of Genome Interpretation (March 25-27 2016, San Francisco, USA).

Aspromonte M.C., Gasparini A, Polli R., Bettella E., Cesca F., Sartori S., Bigoni S., Mammi I., **Carraro M**., Tosatto S.C.E, Murgia A., Leonardi E. TRIO variants in individuals with variable intellectual deficits. Poster at The European Human Genetics Conference (May 27-30 2017, Copenhagen, Denmarrk)

Aspromonte M.C., Gasparini A, **Carraro M**., Bettella E., Polli R., Cesca F., Sartori S., Toldo I., Bigoni S., Peron A., Stanzial F., Tosatto S.C.E., Murgia A. and Leonardi E. Targeted gene panel for comorbid neurological disorders. Poster at "XIX Congresso Nazionale SIGU" (November 23-26 2016, Turin, Italy.)

Leonardi E., Gasparini A, **Carraro M.,** Bettella E., Polli R., Cesca F., Sartori S., Tosatto S.C.E. and Murgia A. Targeted gene panel to investigate intellectual disability and autism spectrum disorder comorbidity. Poster at The European Human Genetics Conference (May 21-24 2016, Barcelona, Spain).

# Sommario

Il completamento del progetto genoma umano ha aperto numerosi nuovi orizzonti di ricerca. Tra questi, la possibilità di conoscere le basi genetiche che rendono ogni individuo suscettibile alle diverse malattie ha aperto la strada ad una nuova rivoluzione: l'avvento della medicina personalizzata. Le tecnologie di sequenziamento del DNA hanno subito una notevole evoluzione, ed oggi il prezzo per sequenziare un genoma è ormai prossimo alla soglia psicologica dei $ 1 000. La promessa di identificare varianti genetiche che influenzano il nostro stile di vita e che ci rendono suscettibili alle malattie sta quindi diventando realtà. Tuttavia, molto lavoro è ancora necessario perché questo nuovo tipo di medicina possa trasformarsi in realtà. In particolare la sfida oggi non è più data dalla generazione dei dati di sequenziamento, ma è rappresentata invece dalla loro interpretazione. L'obiettivo del mio progetto di dottorato è lo sviluppo di metodi bioinformatici per predire la predisposizione a patologie, a partire da dati di sequenziamento. Molti di questi metodi sono stati testati nel contesto del Critical Assessment of Genome Interpretation (CAGI), una competizione internazionale focalizzata nel definire lo stato dell'arte per l'interpretazione del genoma, ottenendo sempre buoni risultati. Durante il mio progetto di dottorato ho avuto l'opportunità di affrontare l'intero spettro delle sfide che devono essere gestite per tradurre le nuove capacità di sequenziamento del genoma in pratica clinica. Uno dei problemi principali che si devono gestire quando si ha a che fare con dati di sequenziamento è l'interpretazione della patogenicità delle mutazioni. Decine di predittori sono stati creati per separare varianti neutrali dalle mutazioni che possono essere causa di un fenotipo patologico. In questo contesto il problema del benchmarking è fondamentale, in quanto le prestazioni di questi tool sono di solito testate su diversi dataset di varianti, rendendo impossibile un confronto di performance. Per affrontare questo problema, una comparazione dell'accuratezza di questi predittori è stata effettuata su un set di mutazioni con fenotipo ignoto nel contesto del CAGI, realizzando la valutazione per predittori di patogenicità più completa tra tutte le edizioni di questo esperimento collaborativo. La previsione di fenotipi a partire da dati di sequenziamento è un'altra sfida che deve essere affrontata per realizzare le promesse della medicina personalizzata. Durante il mio dottorato ho avuto l'opportunità di sviluppare diversi predittori per fenotipi complessi utilizzando dati provenienti da pannelli genici ed esomi. In questo contesto sono stati affrontati

problemi come errori di interpretazione o la sovra interpretazione della patogenicità della varianti, come nel caso della sfida focalizzata sulla predizione di fenotipi a partire dall'Hopkins Clinical Panel. Sono inoltre emersi altri problemi complementari alla previsione di fenotipo, come per esempio la possibile presenza di risultati accidentali. Specifiche strategie di predizione sono state definite lavorando con diversi tipi di dati di sequenziamento. Un esempio è dato dal morbo di Crohn. Tre edizioni del CAGI hanno proposto la sfida di identificare individui sani o affetti da questa patologia infiammatoria utilizzando unicamente dati di sequenziamento dell'esoma. L'analisi dei dataset ha rivelato come la presenza di struttura di popolazione e problemi nella preparazione e sequenziamento degli esomi abbiano compromesso le predizioni per questo fenotipo, generando una sovrastima delle performance di predizione. Tenendo in considerazione questo dato è stata definita una strategia di predizione completamente nuova per questo fenotipo, testata in occasione dell'ultima edizione del CAGI. Dati provenienti da studi di associazione GWAS e l'analisi delle reti di interazione proteica sono stati utilizzati per definire liste di geni coinvolti nell'insorgenza della malattia. Buone performance di predizione sono state ottenute in particolare per gli individui a cui era stata assegnata una elevata probabilità di essere affetti. In ultima istanza, il mio lavoro è stato focalizzato sulla predizione di gruppi sanguigni, sempre a partire da dati di sequenziamento. L'accuratezza dei test sierologici, infatti, è ridotta in caso di gruppi di sangue minori o fenotipi deboli. Incompatibilità per tali gruppi sanguigni possono essere critiche per alcune classi di individui, come nel caso dei pazienti oncoematologici. La nostra strategia di predizione ha sfruttato i dati genotipici per geni che codificano per gruppi sanguigni, presenti in database dedicati, e il principio di nearest neighbour per effettuare le predizioni. L'accuratezza del nostro metodo è stata testata sui sistemi ABO e RhD ottenendo buone performance di predizione. Inoltre le nostre analisi hanno aperto la strada ad un ulteriore aumento delle prestazioni per questo tool.

# Abstract

The sequencing of the human genome has opened up completely new avenues in research and the notion of personalized medicine has become common. DNA Sequencing technology has evolved by several orders of magnitude, coming into the range of $1,000 for a complete human genome. The promise of identifying genetic variants that influence our lifestyles and make us susceptible to diseases is now becoming reality. However, genome interpretation remains one the most challenging problems of modern biology. The focus of my PhD project is the development of bioinformatics tools to predict diseases predisposition from sequencing data. Several of these methods have been tested in the context of the Critical Assessment of Genome Interpretation (CAGI), always achieving good prediction performances. During my PhD project I faced the complete spectrum of challenges to be address in order to translate the sequencing revolution into clinical practice. One of the biggest problem when dealing with sequencing data is the interpretation of variants pathogenic effect. Dozens of bioinformatics tools have been created to separate mutations that could be involved in a pathogenic phenotype from neutral variants. In this context the problem of benchmarking is critical, as prediction performance are usually tested on different sets of variants, making the comparison among these tools impossible. To address this problem I performed a blinded comparison of pathogenicity predictors in the context of CAGI, realizing the most complete performance assessment among all the iterations of this collaborative experiment. Another challenge that needs to be address to realize the personalized medicine revolution is the phenotype prediction. During my PhD I had the opportunity to develop several methods for the complex phenotype prediction from targeted enrichment and exome sequencing data. In this context challenges like misinterpretation or overinterpretation of variants pathogenicity have emerged, like in the case of phenotype prediction from the Hopkins Clinical Panel. In addition, other complementary issues of phenotype predictions, like the possible presence of incidental findings have to be considered. *Ad hoc* prediction strategies have been defined while facing with different kinds of sequencing data. A clear example is the case of Crohn's disease risk prediction. Always in the context of the CAGI experiment, three iterations of this prediction challenge have been run so far. Analysis of datasets revealed how population structure and bias in data preparation and sequencing could affect prediction performance, leading to inflated

results. For this reason a completely new prediction strategy has been defined for the last edition of the Crohn's disease challenge, exploiting data from Genome Wide Association Studies and Protein Protein Interaction network, to address the problem of missing heritability. Good prediction performance have been achieved, especially for individuals with an extreme predicted risk score. Last, my work has been focused on the prediction of a health related trait: the blood group phenotype. The accuracy of serological tests is very poor for minor blood groups or weak phenotypes. Blood groups incompatibilities can be harmful for critical individuals like oncohematological patients. BOOGIE exploits haplotype tables, and the nearest neighbor algorithm to identify the correct phenotype of a patient. The accuracy of our method has been tested in ABO and RhD systems achieving good results. In addition, our analyses paved the way for a further increase in performance, moving towards a prediction system that in the future could become a real alternative to wet lab experiments.

# Introduction

The technological advances achieved after the conclusion of the Human Genome Project, has opened the possibility to easily sequence individuals genome. Thanks to the introduction of high throughput technologies, for the first time is possible to investigate patients genome, looking for variants responsible for disease onset and on this basis, define personalized therapies. Capability to obtain patients genetic data is now easier than ever. These achievements in the sequencing technologies lead to the definition of new promises like, the possibility to prevent disease onset and the possibility to predict adverse effects of pharmacological treatments, moving the clinical practice towards a new kind of personalized and preventive medicine. Wherever for monogenic diseases big progresses have been achieved, at the moment the realization of such promises is still far away from being realized for complex phenotypic traits. In this context, the aim of my research has been to develop bioinformatics tools useful to predict individual risk for complex diseases and to predict health-related genetic phenotypes.

In this thesis I will try to summarize my contribution in answering demands of translating the sequencing revolution in real advantages for patients of tomorrow.

A comprehensive assessment of bioinformatics tools predicting variants pathogenicity was performed to address the problem of interpretation for the thousands of variants identified in genome or exome sequencing experiments. In addition several different strategies has been applied to the problem of phenotype prediction from sequencing data. Both data coming from targeted enrichment sequencing and exome sequencing have been used to predict the onset of disease phenotypes or to define relevant phenotypes for human health like blood types.

# 1 The personalized medicine revolution

## 1.1 The advent of personalized medicine

The completed sequencing of the human genome in 2003 has been a scientific watershed with great potential to improve medicine. The resulting technological advance has opened the possibility to sequence individual genomes in a short amount of time and at a reasonable price. The promise of identifying genetic variants that influence our lifestyles and make us susceptible to diseases is now becoming reality. A new era for

healthcare is beginning, the era of personalized treatments. This has been anticipated since the end of the 19th century by Sir William Osler, a Canadian Physician, who said that "If it were not for the great variability among individuals, medicine might as well be a science and not an art"[1].

Until the last decade, the prevalent idea was that susceptibility to diseases could be described as a normal distribution, considering the incidence of a specific phenotype in the general population. Influenced by this idea, for many years, pharmaceutical research focused its attention on the discovery of drugs that could be effective on the general population. Increasing cases of individuals with reduced or toxic effects, makes evident that personal genetic variations in the normal phenotypical distribution have to be considered carefully.

Two pioneers of the genetic era were among the first to have their genome sequenced: James Watson (Nobel Prize for discovering the DNA structure) and Craig Venter (lead scientist of Celera genomics at the time of Human Genome Project). This fact created great expectations among the general public, especially with the publication of their genetic code. Venter published his entire genome sequence, revealing the presence of polymorphisms that make him potentially susceptible to antisocial behavior, alcoholism, obesity, stroke and Alzheimer's disease[2]. Interestingly, Watson decided not to publish a short part of his genome containing the APOE gene, which is linked to Alzheimer's disease onset[3], saying something very inspiring "Since we can't do much about Alzheimer's disease, I didn't want to know if I was at risk"[4]. The hard reality is that great advance in diagnosis often has no reflection on our ability to treat genetic diseases.

The advent of personalized medicine promises to achieve a shift in future healthcare not only with a predictive, but mainly a proactive approach to medicine, where emphasis should be placed more on disease prevention than treatment. A change of paradigm in research is also needed to achieve this goal. Different disciplines cannot be considered separated anymore and patient data obtained by different high-throughput techniques has to be necessarily integrated in a conceptual data cloud. What over a dozen years ago could only be considered fiction will only become reality in this way. The time is ripe for the next revolution in healthcare: personalized medicine.

## 1.2 Evolution and perspectives of personalized medicine

New achievements of high throughput sequencing did not stop with the end of the Human Genome Project. Year after year, this technology continues to evolve, chasing the goal of the $1 000 genome. Pushed by public research and private companies, this new challenge lead to a rapid decrease of DNA sequencing costs. Over the last 15 years, the sequencing cost for a human genome dropped from $100 000 000 (estimated cost of the Human Genome Project) to less than $10 000 for a genome. See Table 1 for a comparison of technology before and after the Human Genome Project. During the last few years, several companies claim to have reached the goal of $1 000 dollars. Illumina HiSeqX Ten in particular seems to have found this "holy grail".

At present, state-of-the-art technology allows to genotype any sample of interest at an affordable price in a short period of time, sometimes less than a couple of weeks. Whole Genome Sequencing (WGS) and Whole Exome Sequencing (WES) have become affordable tools for understanding the genetic bases of human phenotypes and diseases[5]. Analysis of genetic variants in an individual genome has allowed to examine the genetic bases of disease with an unprecedented level of detail. The huge amount of data generated for both healthy and diseased individuals, has not only helped in the definition of the molecular bases of genetic diseases, but is also transforming future healthcare. The overlay of genomic data with medical patient records will soon allow to predict and prevent disease onset, enabling new pro-active therapeutic strategies.

In this context, National Cancer Institute (USA) in 2011 coined the term "personalized medicine" for healthcare considering information about individual genome, proteins and environment for diagnosis and to treat diseases[6]. This will transform the perspective of future healthcare from disease diagnosis and treatment to personalized health monitoring and preventive medicine. First examples of personalized healthcare are based on the analysis of patient genomic markers to define whether a person is likely to respond to a given pharmacological therapy, adjusting dosage to optimize drug efficacy and safety.

| | HGP begins | HGP ends | 10 years after HGP |
|---|---|---|---|
| **Cost to generate a human genome sequence** | $1 billion | $10-50 million | $3-5 thousand |
| **Time to generate a human genome sequence** | 6-8 years | 3-4 months | 1-2 days |
| **Vertebrate genome sequences** | 0 | 3 | 112 |
| **Prokaryotic genome sequences** | 0 | 167 | 8760 |
| **Human single nucleotide polymorphisms** | 4.4 thousand | 3.4 million | 53.6 million |
| **N° genes with known phenotype/disease causing mutation** | 53 | 1474 | 2972 |
| **Drugs with pharmacogenomics information on label** | 4 | 46 | 104 |

HGP human genome project

**Table 1. Quantitative advances since the Human Genome Project.** From the beginning of the human genome project, both price and time needed to sequence a genome have dramatically decreased. Modified from data of the National Human Genome Research Institute.

One of the most famous attempts of personalized treatment based on the analysis of genetic variants is warfarin, the most commonly used anticoagulant worldwide. Warfarin targets the vitamin K epoxide reductase complex subunit 1 (VKORC1) enzyme. Inhibition of VKORC1 by warfarin leads to the production of coagulation factors with reduced activity. Several VKORC1 mutations have been identified and most are common variants affecting VKORC1 expression influencing warfarin dosage within the normal range. Rare mutations have been associated with warfarin resistance, requiring an increase in drug dosage. Without knowing the personal characteristics of patients and their genetic

background, it could take months of trial-and-error testing to find the right drug dose. The problem could be more widespread than expected. Mutations can also affect enzymes involved in the process of drug metabolization such as the Cytochrome P450 family members. E.g. for Cytochrome P450 2D6 (CYP2D6), one of the best studied drug-metabolizing enzymes, about 10% of the general population has a slow-acting form, while another 7% have a super-fast-acting form[6]. Some subjects may therefore process drugs too rapidly (ultra-metabolizers), rendering them ineffective, or too slowly (poor metabolizers), causing an increase in blood concentration and potentially leading to toxic effects. At the moment, several studies are trying to optimize drug dosage after Single Nucleotide Polymorphism (SNP) genotyping and the first on-line tools become available to identify and suggest if patients need a more specific drug dosage based on their genetic background. A growing number of drugs It is expected to have companion diagnostics, as about 10% of marketed medications will propose or recommend genetic testing for treatment optimization in the future.

Another field of personalized medicine that has greatly advanced is precision disease diagnosis (see Figure 1). The decrease in WGS cost allow causal gene identification for diseases and complications at a personalized level. E.g. Bainbridge and colleagues sequenced the complete genomes of a twin pair and identified compound heterozygous mutations in the SPR gene responsible for the dopa (3,4-dihydroxyphenylalanine)-responsive dystonia in both twins[5]. Precise identification of the specific causal variants open the possibility of improving child health by supplementing L-dopa therapy with 5-hydroxytryptophan, the serotonin precursor whose synthesis depends on SPR. Precise diagnosis in cancer research has markedly benefited from WGS/WES. Hundreds of cancer genomes have been sequenced, allowing previously unimaginable collaborative efforts that led to the creation of fundamental resources such as the Cancer Genome Atlas. In addition to bulk cancer sequencing, single-cell cancer exomes have also been examined. When compared to normal tissues, somatic mutations for specific cancer genomes, as well as molecular markers for cancer subtyping, could be identified. Børresen-Dale propose to classify breast carcinomas on the basis of different gene expression patterns to link tumor characteristics with clinical outcome. For patients that had received the same therapy, estrogen receptor positive tumors could be divided into at least two groups, each with its specific gene expression profile and different prognosis[7]. These data could provide potential targets for personalized cancer treatment in the future. WGS

could also help identify spontaneous mutations in the 'normal' genome of cancer patients that may lead to carcinogenesis. Sequencing has already been applied to patients with suspected increased cancer susceptibility such as those with multiple primary tumors. E.g. a germline de novo p53 deletion was identified in a patient who developed 3 different cancer types in 5 years[5].

So far, we described how disease diagnosis and treatment are changing in the "omics" age, but something is still missing. The personalized medicine revolution seems to be dramatically deeper than could be expected. Knowing patient biological background, preventive risk assessment will become possible, defining individuals as "not-yet patients". Not-yet patients are individuals with a genetic background predisposing for a specific disease, who do not present any symptoms (yet). This situation could be very dramatic in the case of patients at risk of a lethal or disabling disease that may never develop. Examples, are cases of individuals found to be positive for a BRCA gene mutation in genetic testing. BRCA1 and BRCA2 tumor suppressors ensuring the stability of genetic material by helping in DNA damage repair. When either is mutated, DNA damage may not be properly repaired and cells are more likely to develop additional genetic alterations that can lead to cancer. Specific BRCA1 and BRCA2 variants increase the risk of several cancer types[8,9]. In particular, BRCA gene mutations account for about 5 to 25% of hereditary breast cancers and around 15% of ovarian cancers[10]. Individuals positive to BRCA mutations and with a strong familiarity for cancer, could even decide to undergo to preventive mastectomy and removal of ovaries and fallopian tubes, to prevent a pathology that they may have never developed. An additional a chapter could be dedicated to incidental or secondary findings. These are unexpected results, not related with the clinical condition for which sequencing has been performed. New specific regulations are needed to manage these situations, ranging from the uncommon "misattributed paternity" to findings of critical medical value like possible predisposition to degenerative diseases[11].

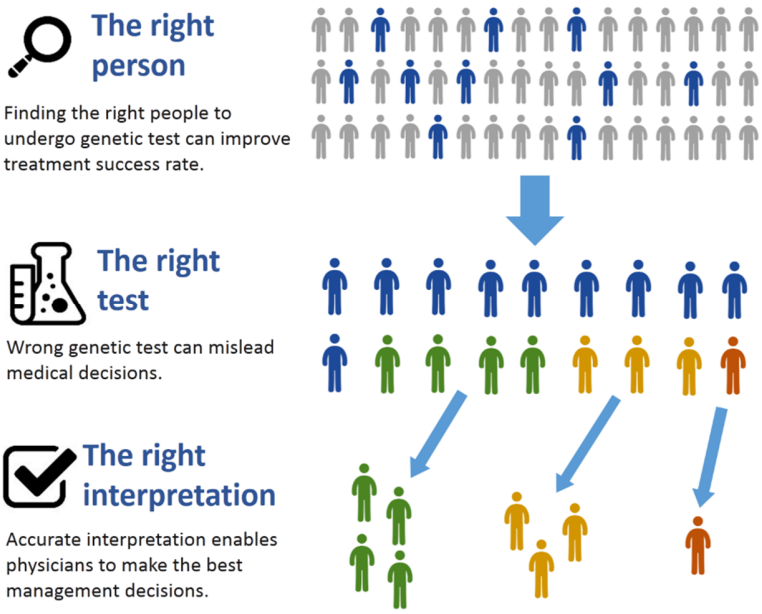


**Figure 1. The three main steps of precision disease diagnosis.** Genetic tests could lead to more precise diagnosis and more effective treatments.

## 1.3 Omic sciences and their interaction

New technologies and knowledge developed in the context of the Human Genome Project opened the field to the so-called omics revolution during the beginning of the 21st century[12]. The technological effort needed to sequence the human genome lead to the definition of new protocols and technologies, suitable for the production and analysis of an enormous quantity of scientific data. These technologies generating a previously unbelievable amount of data thanks to the high processivity of the new approaches, defined new "high-throughput" standards of performance. In fact, high-throughput technologies were essential to reach the ambitious aim of the Human Genome Project. Without the ability to rapidly and accurately measure thousands of data in a short time period, there would be no way to sequence an entire genome.

Like the new approaches developed in the context of Human Genome Project, all disciplines focused on the development of new techniques and on their data analysis have been call "omics", from "-ome", a term derived from the word genome. Genomics has been the first "omic" discipline to be defined in the context of the Human Genome Project. Genomics focuses on the sequencing and analysis of genomes and exomes. E.g. single

nucleotide polymorphism genotyping (SNP genotyping) measures individual genotypes for several hundred thousand SNPs in the genome. Approaches like WGS and WES were not anticipated. Other assays to sequence and analyze a small amount of the genome have been proposed to focus only on positions that could be causal of disease onset. After more than 15 years from the conclusion of the Human Genome Project, genotyping technologies are now accurate and affordable, but analysis at DNA level sometimes presents limitations. DNA sequence variations tend to be very common, generating a lot of noisy signals that can be hard to decipher. In addition, other epigenetic modifications and environmental factors may modify gene expression in a non-predictable way. Even so, SNP genotyping is currently considered among the most useful techniques to predict disease risk.

Use of high-throughput technologies is not limited to genome analysis but at least three others omics disciplines can be identified: Transcriptomics, Proteomics and Metabolomics. Transcriptomics is the simultaneous measurement of gene expression levels in a cell or tissue by oligonucleotide arrays in which hundred thousands of probes capture RNA molecules. Proteomics, instead, focuses directly on protein levels in a tissue as mainly obtained by mass spectrometry. The size of each peptide is defined after protein extraction and digestion. Proteins can be identified by comparing the size of the peptides extracted from the tissue with a database containing the digest of all known proteins. Last but not least, metabolomics is the high-throughput measure of metabolites present in a cell or a tissue. In general, each discipline offers a different perspective on the molecular mechanisms underlying disease initiation and progression.

The omics revolution also opens new challenges, as laboratories usually do not have sufficient computational resources and storage to process this large amount of data. Since storage and analysis costs are not falling as fast as data generation, this represents a new bottleneck for advancing the field. New kind of scientists and technical infrastructures are needed. One way to address these challenges is the training of an ever-increasing number of bioinformaticians. Cloud computing is a promising technology to fill the gap between data generation and storage, such as the Embassy cloud which is part of the European ELIXIR bioinformatics infrastructure[12].

## 1.4 Systems biology: a further starting point

Despite WES/WGS genotyping is the most useful techniques to predict risk for genetic diseases, genomic information may not always be sufficient to predict a person's health. Environmental factors in fact, can contribute to disease development or even trigger disease onset in susceptible individuals. For complex and multifactorial diseases, many authors consider the analysis of WES/WGS data like the starting point to predic disease outcome. E.g. Baranzini and colleagues failed to find evident genomic or transcriptomic differences in monozygotic twin pairs discordant in multiple sclerosis. Despite a strong genetic component having been postulated for this disorder, it is likely that other factors contribute to disease onset. Access to large omics data could provide new insights for the treatment of human diseases. Transcriptomic, proteomic and metabolomic information, could be considered a more precise index of human health than genomic sequence alone. Combining genomic information with a scheduled monitoring of these omics parameters should serve to obtain real-time information of an individual's condition.

Integration of different omics data has led to a new discipline called systems biology, aiming to model complex biological interactions integrating information in a holistic manner. In contrast to treating a mixture of factors as single entities, systems biology relies on experimental and computational approaches to provide mechanistic insights[12]. In systems biology, data are often elements integrated into networks. E.g. consider information coded in our genome and environmental signals. In systems biology, these two information types have to be considered together, integrated into the individual organism to produce its phenotype – normal or diseased. These two information types and the phenotypes they produce are considered part of biological networks that capture, integrate and transmit the information to molecular machines. A fundamental postulate in systems biology is that disease arises from networks perturbed by genetic changes and/or environmental signals. The resulting altered molecular machinery encoded by the perturbed network leads to the disease pathophysiology[13]. Integrated omics data analysis could monitor molecular profiles and detect subtle changes that may indicate network perturbation. E.g. Snyder and colleagues studied the omics profile of healthy volunteers monitored for 14 months with a so-called integrative Personal Omics Profile (iPOP) analysis[5]. The individual's genome was sequenced at high accuracy with WGS/WES and genetic predispositions for diseases and drug responses were identified.

The physiological state changes occurring during two viral infections and onset of type 2 Diabetes were monitored with information from the transcriptome, proteome and metabolome. The generated integrative profile could observe both trend changes, associated with more gradual changes, and spikes of particularly enriched genes and pathways, especially at the beginning of each physiological event. This integrative analysis provided a much more comprehensive view of the biological pathways changing during disease onset. Importantly, thanks to the previous genome sequencing and active monitoring, Diabetes onset was detected in its early stage and could be effectively controlled by proactive interventions such as diet change and physical exercise. It is clear that neither "genomic medicine" in the future will be sufficient to describe the new horizons offered by system biology. Genomic medicine is one-dimensional in nature, considering only nucleic acid information. In contrast, this new medicine will be holistic, using all types of biological information from DNA and RNA to proteins, metabolites, interactions, cells, organs, and external environmental signals, integrating them in predictive models for health and disease. A new term for future systems biology-based medicine has been coined: systems medicine.

# 2 Next Generation Sequencing

During the development of the Human Genome Project sequencing technologies undergone a rapid development. The end of this big scientific and technological effort is universally recognized as a watershed that separates two scientific eras: the era of Sanger sequencing and the so-called "Next Generation Sequencing" (NGS) era. Sanger sequencing was used for the Human Genome Project, however, from the beginning, it was clear that a faster, high throughput and cheaper technology was required to make genetic sequencing available to the scientific community. The huge amount of time and resources required for sequencing the first human genome was in fact unsustainable if applied to a large number of patients. For this reasons, in 2004 the National Human Genome Research Institute (NHGRI) started a funding program with the goal of reducing the cost of single genome sequencing to $1,000 in ten years[14]. This, stimulated the development of the next-generation sequencing (NGS) technology. This new method presents several improvements respect to Sanger sequencing. In particular the enormous numbers of reads generated by NGS in parallel enabled the sequencing of entire genomes at an unprecedented speed.

Thanks to high performance achieved by NGS technologies, for the first time the role of genetic variants in disease onset could be investigated on a large scale. Three main approaches are nowadays used to perform this kind of analysis: Whole Genome Sequencing (WGS), Whole Exome Sequencing (WES) and targeted enrichment sequencing[15]. Currently WES represents the most efficient approach to investigate the whole spectrum of genetic variants carried by an individual. Thanks to WES, it is usually possible to identify 3 to 4 millions Single Nucleotide Variants (SNVs) per genome. In general, 80-90% of the retrieved variants are single nucleotide polymorphisms already present in the dbSNP database, while 0.5 millions are novel[16]. Compared to WGS, which requires the analysis of the whole haploid human genome ($3.2*10^9$ bases), WES focuses only on protein coding regions which correspond approximately to 1% of the genome. In this way WES allows a strong reduction of both time and resources in respect to WGS. An even more effective approach is the targeted enrichment sequencing which is at the basis of gene panels development. Targeted enrichment sequencing focuses on the analysis of a small set of specific genes or genetic sequences usually associated with a disease or a molecular pathway of interest. This kind of technology is particularly effective while studying monogenic or oligogenic diseases associated to specific phenotypes. In this work we mainly used WES and targeted enrichment sequencing data for diagnosis and phenotype prediction.

## 2.1 Whole Exome Sequencing

During the last years, several groups have demonstrated the power of WES in the discovery of variants involved in human diseases. This approach is characterized by two distinct phases. The first phase is mainly experimental and consists in the preparation of genomic DNA libraries, exome hybridization by means of capture probes and sequencing (NGS) of the captured sequences. The second phase instead, is based on a series of computational analyzes that allow the analysis of the sequences, the identification of all the variants and the detection of those associated with the phenotype of interest (see Figure 2).

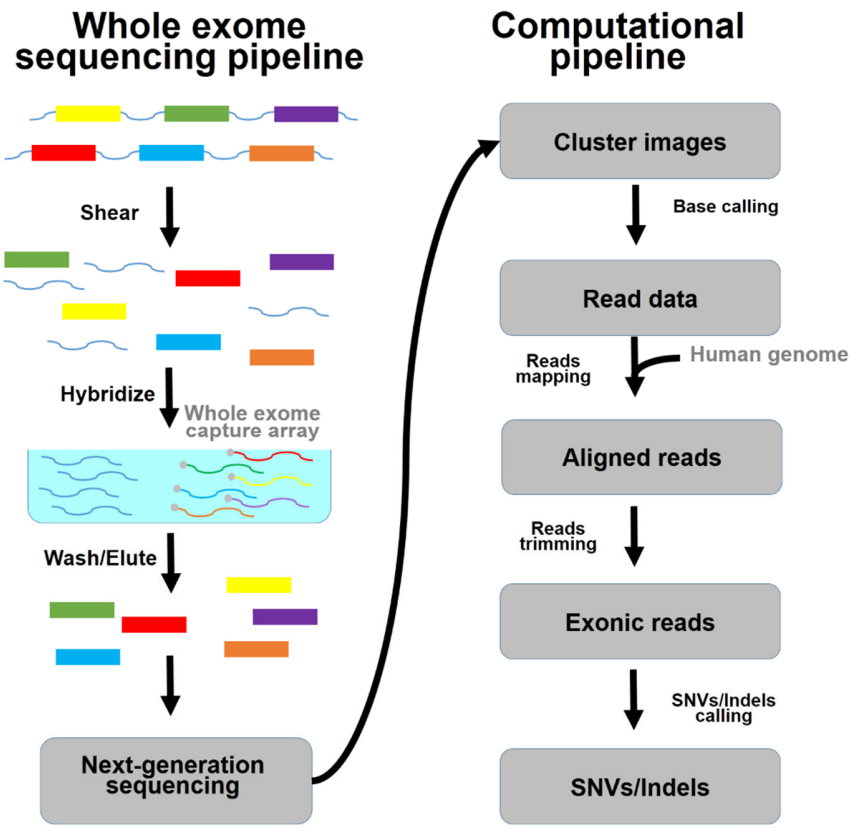


**Figure 2. Whole Exome Sequencing experimental and computational pipeline.** WES is composed by two phases: the experimental phase and the computational analysis. Figure modified from[17].

Focusing on the first phase, despite various exonic DNA sequencing technologies exist, most of them follow a similar procedure. They typically differ only for the capture method and for the NGS platform used for sequencing. The first step is the extraction of genomic DNA from the samples of interest. Once the extraction has been completed it is possible to proceed with exons capture, which filters out all non-coding sequences. At this stage, the extracted DNA is fragmented by sonication or enzymatic digestion. DNA fragments are then ligated to specific adapters and amplified to obtain a library containing the entire sequence. At this point the exome is captured through a hybridization process with specific probes for protein-coding sequences. This hybridization process can be carried out using two main technologies: liquid phase hybridization or solid phase hybridization.

Liquid phase methods use biotinylated DNA or RNAs complementary to the coding regions that. Exonic sequences, once hybridized, are captured by means of streptavidin-coated magnetic beads. In solid-phase protocols, DNA probes are bound to a solid support such as such as microarray slides or paper filters. Several kits could be used in this phase. Commercial kits differ mainly for probes features, which typically may differ by length and/or overlap with target sequences. For this reason, although all capture kits use probes to hybridize the coding regions, the captured sequences may actually be slightly different allowing for examples the capture of regulative or spicing sequences. Once hybridization of the coding regions is completed, sequences not bound to probes are removed by repeated washes. Captured sequences are then eluted and the whole procedure is checked by quantitative PCR at control loci. At this point, the captured exon fragments are sequenced by NGS, which generates short reads of 25-100bp covering the whole exome.

The second phase of WES consists in a series of bioinformatics analyzes. The first step consists in the alignment and mapping of the reads obtained from the NGS to the "reference sequence" of the human genome. The number of differences between the reference and the sequenced genome is generally very high. Often it contains many false positives due to sequencing errors or errors generated in the alignment phase. They are mainly generated by the presence of repeated sequences. In order to clean up the data, different quality control tools and filters can be used. These tools allow to obtain a list of reliable variants representing the starting point for investigating the role of mutations in the onset of genetic pathologies.

## 2.2 Strengths and weaknesses of Whole Exome Sequencing

Before new sequencing technologies emerged in the last decade, all genetic studies on hereditary pathologies were based on linkage disequilibrium analyses, followed by positional cloning of candidate regions. Very often, however, these regions are characterized by the presence of numerous genes, which makes necessary to clone and sequence all genes separately, causing a considerable waste of time and resources. Sequencing the entire genome by means of current NGS techniques instead, allows to quickly define the entire variability of an individual. Unfortunately, costs related to WGS

are not yet accessible for large-scale analysis and generates a huge amount of data difficult to interpret.

In this context, i.e. where the focus is exclusively on coding sequences, WES allows to obtain the advantages of NGS and with reduced costs and time. By limiting the sequencing to a small number of sequences, it is possible to obtain a higher quality and simplify the analysis of the results, as the number of variants is considerably lower in respect to WGS. The loss of information is minimal, since the 85% of mutations involved in genetic pathologies fall within protein coding sequences[18].

The weaknesses of WES are the following. First, in some cases some of the coding regions may not be sequenced. This problem is mainly due to the GC content of the region to be sequenced (which may affect the sequence capture process). A second disadvantage of WES is the limited ability to detect structural variations, like Copy Number Variations (CNVs), translocations and inversions. To date, WES is the most popular method for the genetic study of human phenotypes.

## 2.3 Targeted enrichment sequencing

Despite the power of WGS and WES approaches, single-gene testing and targeted enrichment sequencing still holds great value for many types of disorders (see Table 2), in particular for molecular diagnosis. Due to resource limitations, researchers can focus on a subsets of genes strongly related to the phenotype of interest with a strong cost reduction. It also enables multiplexing, i.e. parallel sequencing.

Gene panels are also useful for monogenic and oligogenic phenotypes, where approaches like WES are considered an overkill. Experimental procedures to produce gene panels are very similar to those used for WES. A crucial phase for gene panel is the definition of target genes. Since the introduction of NGS into clinical practice, the number of disorders for which gene panel are offered is increased dramatically. The number of genes for diagnosis of the same disease may vary significantly among different clinical laboratories. As an example, there are several epilepsy gene panels, with the number of testing genes ranging from 70 to 377[19]. These differences depend on the available previous knowledge, i.e. association studies, for a given disease or just on different confidence thresholds applied by different groups .

Several considerations have to be taken into account when deciding which genes will be included in a panel. Definitely, genes with a strong disease association are worth to be included. Genes associated with disorders that have overlapping phenotypes with the studied pathology, could be included for the purpose of differential diagnosis. For example, the SLC2A2 gene for Fanconi–Bickel syndrome could be included in a glycogen storage disease gene panel because when a patient shows fasting hypoglycemia, both Fanconi–Bickel syndrome and glycogen storage disease are considered[19]. In addition even genes for phenotypes associated with syndromic and nonsyndromic forms could be considered depending on the purpose of the gene panel. In such context of different possible choices, a strong partnership between clinicians and geneticists is required.

| | Single-gene test | Gene panel | Exome sequencing |
|---|---|---|---|
| **Phenotype level** | Specific features point to one disorder associated with one gene | Genetically heterogeneous disorders | Multiple non specific features<br>Extreme heterogeneity |
| **Gene level** | Disease-causing genes | Well-defined disease associated genes | All 20,000 genes with 4,600 medically well defined genes |
| **Variant level** | Minimal VUS<br>No IFs | Fewer VUS than exome sequencing<br>Less likely to find IFs | Large number of VUS<br>Potential to find IFs |
| **Technical issues** | Traditional Sanger: gold standard for sequencing | Need Sanger confirmation<br>Overall higher coverage than exome sequencing | Need Sanger confirmation |

IF incidental finding, VUS variants of unknown significance

**Table 2. Comparison between single-gene, gene panel and exome sequencing test.** Different kinds of genetic test could target different clinical needs. Modified from[19]**.**

# 3 Genome-Wide Association Studies

Genome-Wide Association Studies (GWAS) are based on the comparison of allelic frequencies in a sample of patients with a specific phenotype or disease (defined as cases), and a group of healthy individuals (defined as controls). By means of specific statistical tests, it is possible to detect whether some variants could predispose to a disease onset (statistically significant in cases), or have a protective role (more common in controls)[20].

GWAS represent one of the principal methods to investigate etiology of complex pathologies. The first GWA study was published in Science in 2006 with the aim of identifying susceptibility variables for age-related macular degeneration[21]. Since that, GWAS have been applied to the most common complex pathologies (rheumatoid arthritis, bipolar disorder, hypertension, coronary artery disease, psoriasis, etc.). To date, more than 2,000 loci associated with multifactorial phenotypic traits have been discovered thanks to GWAS[21]. Many of them never hypothesized before (see Figure 3). The total number of associated loci varies considerably for different phenotypic traits, from dozens for psychiatric disorders, to hundreds for chronic intestinal inflammatory diseases like the Ulcerative Colitis and Crohn's disease[21].

Although association studies have made significant progress in defining the genetic risk, they have to be interpreted carefully. The first problem is the high rate of false positives due to the extremely high number of statistical tests necessary to infer an association. The Bonferroni correction solves this problem. For a typical GWAS study, the statistical significance of an association is defined by a corrected *p-value* threshold of $10^{-7}$ [20]. Second limitation of GWAS studies is that, despite the high number of loci identified, very often these are not able to fully explain the estimated inheritance for the specific trajectory, resulting in a phenomenon called "missing heritability"[22]. In conclusion, despite GWAS is essential to investigate genotype-phenotype relationships, the results should be integrated with alternative analysis.

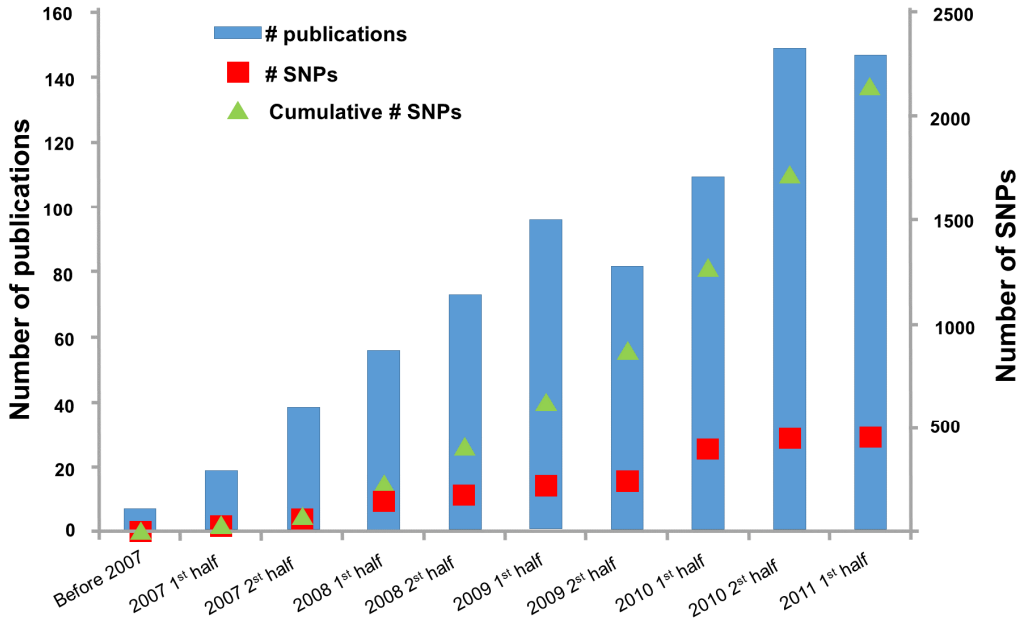


**Figure 3. The exponential growth of GWAS.** Data obtained from GWAS studies published in the GWAS Catalog[23]. Only SNPs with a *p-value* lower than $5*10^{-8}$ are reported. Figure modified from[21].

# 4 Rare and common variants in complex diseases

An interesting debate in the scientific community is about the contribution of high frequency and rare variants on the onset of complex pathologies. Various hypotheses have been formulated on this topic, most of which could be summarized in the so-called "Common Disease, Common Variant" (CDCV) *vs.* "Common Disease, Rare Variant" (CDRV) controversy.

The CDCV hypothesis claims that genetic variants with a relatively high frequency in the population, but characterized by low penetration, are the main responsible for individuals susceptibility to pathologies, and particularly to complex disorders[24]. This hypothesis is supported by the consideration that mutations involved in complex pathologies are subjected to a very weak purifying selection. This assumption can be considered true in particular for the low penetrance mutations involved in complex phenotypes.

Recent studies, however, hypotheses that even a weak purifying selective pressure could be sufficient to maintain variants involved in complex pathologies at low frequencies.

These considerations could be realistic in particular for cases where purifying selection acts on a very long time scale (Pritchard, 2001). On the basis of this kind of hypothesis, the CDRV thesis was defined. This theory claims that genetic variants that have a low frequency within the population can be considered as the main responsible for susceptibility of individuals to complex pathologies[24]. This hypothesis is supported by numerous studies on multifactorial phenotypes such as lipid metabolism, immune system and blood pressure regulation. For these kind of phenotypes in fact, it has been shown that key role is played by genetic variants with very low frequency[25]. Considering the existence of supporting evidence for both CDCV and CDRV hypothesis, it is important to considered that the applicability of both theory is probably dependent on the phenotypic trait analyzed. It can therefore be concluded that the two hypotheses does not have to be considered mutually exclusive, and indeed the debate should probably be more focused on how the interplay between common and rare variants contributes to the onset of complex phenotype.

# 5 Interaction networks in the study of complex phenotypes

All cell components perform their functions interacting with other cellular structure. In some cases both actors are located within the same cell, in some other cases instead, it is possible that these elements interact with structures located at considerable distance, even in different organs. The overall set of interactions between the various cellular components constitutes a network that takes the name of "interactome". In the human being it is estimated that the number of non-coding proteins, metabolites and non-coding RNAs present in the cell is abundantly greater than 100,000 units, and that the number of interactions between these units should be even much higher. In a system characterized by such degree of interdependence, a pathology cannot be considered as the simple consequence of an alteration of a single component. Diseases in fact should rather be investigated considering how variations in single cell components are reflected and propagated within the network interaction. Just considering protein interactions only, it is evident that the impact of a mutation is not limited to the activity of the protein encoded by the mutated gene. It is possible to imagine in fact, that an alteration in the activity of a single protein may potentially be spread along the meshes of the network and affect the activity of other proteins that actually do not present any alteration. In this

perspective, in order to estimate the effect of a single mutation on a phenotype, it is necessary to consider the role of altered components within the interaction network.

Nowadays, thanks to the availability of several interaction databases and network analysis tools, pathologies can be finally investigated under this new lens. This network-based approach has numerous biological and clinical applications. By mean of this kind of analysis a deeper understanding of the pathophysiological role of mutations identified by association studies cold be achieved, maybe leading to the discovery of new genes involved in the onset of the same phenotype.

As anticipated, the increasing number of GWAS studies has led to the identification of many genes involved in human diseases. Unfortunately, these genes usually are not sufficient to explain the entire estimated inheritance for the investigated phenotype[22]. However, taking in account the role of these genes in the interaction network, it is potentially possible to identify new ones playing a role in disease onset. In support of this hypothesis, numerous studies have shown that proteins involved in the same pathology have a high propensity to interact with each other[26]. These observations allow hypothesizing that, if an element associated with the pathology has been identified, other elements could be discovered by analyzing elements neighborhood within the interaction network.

Once genes associated with a pathology have been identified (for example, by mean of GWAS studies), three main types of methods can be used to expand the list of candidate genes by exploiting the interaction network: linkage methods, module or cluster-based methods, and diffusion-based methods[26] (See Figure 4).

In linkage methods it is assumed that genes located in the linkage interval of a disease whose protein interact with a known disease-associated protein are considered likely candidate disease genes (See Figure 4 -1-).

In the second type of methods, it is assumed that for each pathology an interaction subset made by elements involved in the onset of disease could be defined. These portions of the network are defined like "modules" or "clusters" of the disease. To define a cluster, all the proteins encoded by the genes associated with the pathology are initially identified on the network. At this point, clustering tools could be used to test the existence of topological or functional modules within the network that contain most of the associated proteins. The genes coding for members of such modules are therefore considered as candidate genes (See Figure 4 -2-).

In the third type of methods, “random walkers” are released from the protein products of the known disease genes. These explorers are then allowed to diffuse along the links of the interactome, moving to any node with the same probability.

During this exploration, proteins that are closer to ones involved in pathology will be "visited" more often. To these proteins, a higher probability score will be assigned respect to other interactors. In this way, direct interacters and indirect interacters that are close to the associated proteins will have the highest score. Genes encoding for these proteins can therefore be considered as candidate genes to play a role in disease onset (See Figure 4 –3- ).

Each of the three methodologies described above exploits, to an increasing degree, the topological and functional information encoded by the interactome. Linkage methods exploit only interactions between protein couples (direct interactors), while cluster-based methods consider a small portion of the interaction network. Finally, dissemination methods exploit the entire information contained in the network. In this respect, it should not be surprising that results of a comparative analysis have shown that using the same data set, diffusion-based methods are those that provide the best prediction performance[26]. Overall, it is clear that protein network analysis is crucial to expand knowledge on the genetic basis of human disease. In the future, thanks to the increasing spread of GWAS studies and the increase of data stored in interaction databases, it is possible to assume that the contribution provided by the study of interaction networks will become fundamental to investigate the molecular background of complex phenotypes.

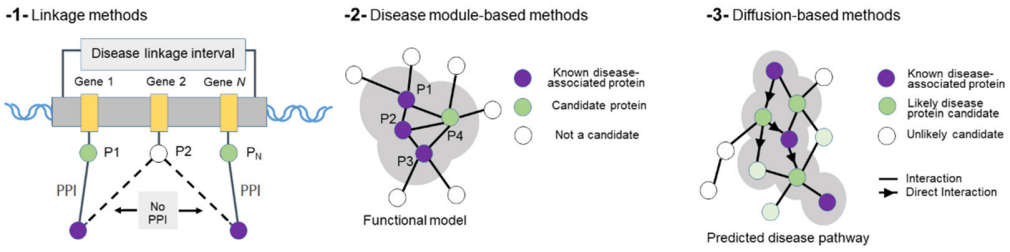


**Figure 4. Methods to identify disease candidate genes exploiting information of interaction networks.** -1- Linkage methods -2- Disease module-based methods -3- Diffusion-based methods. Figure modified from[26].

# 6 Development of a prediction algorithm

Advances achieved thanks to the introduction of high-throughput sequencing technologies have made large biological datasets available to the scientific community. In this context, it is not only possible to interrogate and analyze big datasets, but it is also possible to infer knowledge from this huge amount of data. To this end, dozens of algorithms have been developed to detect hidden patterns. This kind of predictors have been applied to different biological problems, in particular when knowledge is incomplete or when the amount of available data is too large to be handled manually. The development of these prediction algorithms could be divided in three main processes: i) the analysis of the dataset and the underling features, ii) the selection (or development) of the learning algorithm and iii) the analysis of prediction performance.

## 6.1 Analysis of the dataset

The first important step for the definition of a reliable predictor is the analysis of dataset properties or features. First, before beginning with the definition of the predictor itself, is crucial to identify if enough data are available to solve the specific biological problem. Nowadays, in the *Big Data* era, with very large biological datasets available online, this point might appear irrelevant, but in reality it raises an important problem that has to be address before moving forward with the definition of a predictive method. The ideal situation would be having at least ten times as many data instances as the number of measurable property or characteristic of the observed phenomenon (data features)[27].
Different datasets have specific features and often contain errors hard to be identified. Given the uniqueness of each dataset, a reliable predictor could be defined only if we are able to clearly understand dataset strengths and weaknesses, and we are able to arrange them properly. This phase is usually made by several steps, often grouped together under the name of *data pre-processing*. First, an initial useful practice is to randomly shuffle the dataset. This operation is crucial to remove any possible bias related to the order of the data instances. Another step is *data cleaning.* The aim of this process is to reduce redundancy and bias present in the dataset. Analysis based on non-redundant dataset for example, will be more representative of all the items in the dataset, rather than just the largest dominant group. In addition, in this phase all the data which have corrupt, inaccurate, inconsistent values should be removed[28]. When dataset is too small instead,

outliers can be rounded to the maximum (or minimum) accepted limit. Finally for numerical datasets, the *normalization* of values into the [0, 1] interval is often necessary to improve prediction performance.

### *6.1.1 Definition of data subsets*

Many textbooks state that in order to refine a reliable model, data have to be split in two subsets: *training* and *testing*. In practice, this approach is wrong as *training* and *testing* dataset may contain the same bias. This is a common mistake and can lead to inflated prediction performance[29]. To avoid this kind of problem, data should always be split into three independent subsets: *training*, *validation* and *testing* sets. Typically, the suggested ratio is 50% for the *training*, 30% for the *validation*, and 20% for the *test set*. When the dataset is small, alternative techniques such as *cross-validation* could be used[30]. The *training* and *validation* sets are used to refine the predictor model and to optimize hyper-parameter values. The *test* set is used to evaluate prediction performance. This three steps approach is commonly defined as the "lock box approach" and constitute the best practices to be use while developing prediction algorithms[31].

## 6.2 Choice of the prediction method

Many prediction algorithms have been already developed and are available as open source libraries. The first important step to choose the most suitable algorithm for prediction purpose is to clearly define the problem that has to be addressed. Tasks where training data comprises examples along with their corresponding target value are known as *supervised learning* problems. Among these, cases in which the aim is to assign each input to one of a finite number of discrete categories, are called *classification* problems. If the desired output consists of one or more continuous variables instead, then the task is called *regression*. In other cases the training data consists of a set of input data without any corresponding target values. These cases are called *unsupervised learning* problems and in these cases the task may be to discover groups of similar examples within the data (*clustering*), or to determine the distribution of data within the input space (*density estimation*).

To address these different problems several specific algorithms have been developed. A general suggestion for algorithm selection is to start with the simplest one[32]. Using a simple algorithm will make possible to better understand what is happening during the

application of the method. In addition, a simple algorithm will provide better generalization and present less chance of overfitting respect to more complex methods. Some examples of simple algorithms are the *k*-means and the *k*-nearest neighbors clustering algorithms. More complex models such as Bayesian classifier and neural networks should be employed only if the dataset features provide some reasonable justification for their usage[32]. As algorithm selection could be a non-trivial task, a general advice is to use multiple techniques and compare their results[27].

Another main task to be considered while dealing with a prediction algorithm, is to avoid that the model memorize training set properties instead of learning hidden relationship among the data. This phenomenon is called overfitting. Several powerful strategies exist to minimize this phenomenon, *cross-validation* is among the most used. In 10-fold cross-validation for example, the algorithm defines 10 different portions of the dataset as training set and validation set. After shuffling the input dataset instances and setting apart the test set, the algorithm takes the remaining data and divides it into ten shares. The model is then evaluated on each share while being trained on the remaining data[27].

## 6.3 Evaluation of predictor performance

Several measures describe strength and weakness of a prediction method by highlighting different performance aspects. Predictors can be classified as discrete or probabilistic, depending on whether they provide a score for predictions or not. They are evaluated by different metrics. In this chapter I will focus my attention on the description of statistics for the evaluation of discrete predictors, as most of the algorithm presented in this manuscript are binary classifiers. To evaluate the outcome of binary predictors, results are often presented in a 2x2 contingency table. The number of correctly predicted cases are indicated by TP (True Positives) and TN (True Negatives), and the number of incorrectly predicted cases are FN (False Negatives) and FP (False Positives), respectively. Based on these four metrics evaluation measures can be derived and listed in the following. The sensitivity, also called True Positive Rate (TPR) or recall. The specificity (true negative rate, TNR). The positive predictive value (PPV), also called precision. The negative predictive value (NPV).

$$TPR = \frac{TP}{TP + FN}$$

$$TNR = \frac{TN}{FP + TN}$$

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{FN + TN}$$

These statistics are useful to evaluate prediction performances only when positive and negative cases are balanced. Instead, for unbalanced datasets, Accuracy and the Matthew's Correlation Coefficient (MCC) can give better estimation of the real performance[33].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

For all the measures here presented the higher the value the better. Except for MCC, all these values range from 0 to 1. MCC ranges from -1 to 1, where -1 indicates a perfect negative correlation and 1 perfect positive correlation, 0 represents a random predictor. To visually compare different classifier (tested upon the same test set) Receiver Operating Characteristics (ROC) analysis that can be useful. To draw a ROC curve data have to be ranked based on the prediction score. Data are than divided to intervals of equal size in a graph where the x-axis represents 1-specificity (also called FPR) and the y-axis represents sensitivity (TPR). In an ideal case all the true positive cases should be on the first half of the ranked list. In this case the plot will rise to (0, 1) and then continues straight to the right with all the true negative cases[33]. The faster the curve rises the better the method is. A random classification would be on the diagonal. Area under the ROC curve (AUC) could been used as a numerical measure of goodness for predictions starting from the ROC curve. For this index, a value of 0.5 indicates random classification while 1 would indicate a perfect classifier. This analysis is particularly useful also to highlight the presence of tradeoffs between sensitivity and specificity.

When dealing with probabilistic predictors different indices are calculated[33]. For example the Humming distance, the Pearson Correlation Coefficient (PCC) and the Kendall

Correlation Coefficient (KCC). In addition, results can still be presented in a contingency table, dividing the date in several partition of two categories.

# 7 Genome-based prediction of complex phenotypes

The availability of NGS techniques allowed an exponential increase in the number of sequenced individuals. This huge amount of data is revolutionizing the study of genetic pathologies. The future perspective is to exploit the considerable amount of available sequences to develop methods that will allow us to predict the risk of developing genetic diseases and possibly prevent their onset. Thanks of these methods, it will be possible to obtain important benefits both in diagnostic and in preventive medicine. Focusing to the diagnostic side, it would be possible to reduce the use of costly and invasive instrumental investigations, e.g. the use of painful colonoscopy in subjects who are predisposed to Crohn's disease, knowing the genetic characteristics of an individual. Interesting is also the idea of undergoing newborns to NGS screenings that will help to define the genetic predispositions for inheritable disorders[34]. Knowing genetic predispositions in childhood will make possible to suggest lifestyles, or to begin preventative medical therapies, which could reduce disease onset probability.

Initial models to predict disease predisposition from NGS data were called risk prediction models. These models were based on small numbers of SNPs, typically considering only statistically significant variations identified in GWAS. Subsequent studies shown that predictive ability of these methods could be further increased by considering in the model all SNPs including also the ones that did not reach the genome-wide significance threshold[35]. These methods were based on polygenic scoring, summing the estimated effects of a limited number of known risk alleles, sometimes by simply counting the number of risk alleles each sample carries.

Different kind of methods are the one based on machine learning algorithms like neural networks, logistic regression, support-vector machines (SVM) and Bayesian models employing. Regardless of the used algorithm, each method built a mathematical mapping from the SNP data to the phenotype. Main advantages of these sophisticated approaches are that they can account for inter-SNP correlations rather than assuming SNPs independence like in risk prediction models. For these kind of methods, care must be

taken to minimize the issue of overfitting, when a model mistakes noise for signal. This kind of issue could typically be identified when good performance are present in training dataset but poor performance are achieved in independent datasets instead. In this context it should be noted that population stratification, usually considered as an unwanted noisy signal, may be useful instead, as SNPs sometimes can also serve as proxies for shared environmental conditions, such as proximity to a pollutant or adherence to a particular diet[35]. Available tools in literature analyze variants present in the genomes, identifying the presence of potentially pathogenetic mutations such as, VAAST 2.0[36] focus on the identification of variants that cause amino acids substitutions and on the frequency of mutated alleles. Other methods such as VEST[37] use a different approach exploiting a machine learning based algorithm.

Despite the interest in predicting individuals phenotype, it is evident that at the moment universally valid standard methods does not yet exist. Whether polygenic scoring methods or other more sophisticated models are more suitable for a given disease depends on several factors. Among these factors, the most important are probably: the genetic architecture of the disease, the availability of training data, and the available sample size. Existing methods in fact, often need to be adapted to the phenotype or pathology analyzed, with results that sometimes are far away from being considered satisfactory. An example is the case of the 23andme company, which offered for $ 99 the possibility of sequencing DNA and getting information on the presence of mutations involved in pathogenesis. As 22 November 2013, FDA (Food and Drug Administration) suspended the sale of the kit, since the company was unable to provide sufficient evidence to ensure that the genetic test was validated for clinical[38]. In this context, in 2010 the first prototype of Critical Assessment of Genome Interpretation was proposed, an international experiment in which the most recent computational methods are tested to predict effects of genetic variants on phenotypes.

# 8 Critical Assessment of Genome Interpretation

The Critical Assessment of Genome Interpretation (CAGI) is a community effort to objectively evaluate the state-of-the-art in relating genetic information to phenotype, particularly the relationship between human genetic variation and disease. The primary goals of CAGI are to assess current computational methods for interpreting genomic data, highlight innovations & progress, and broadly disseminate the results. CAGI aims to guide

future research efforts in computational genome interpretation and build a strong community for collaboration and interaction. In order to achieve these goals, CAGI conducts experiments in which participating analysts are provided genetic variants and asked to make predictions of corresponding molecular, cellular, or organismal phenotypes.

Many successful genome interpretation studies have been published[39–43], and in the clinic, exome and genome sequencing are increasingly being used to improve prevention, diagnosis, treatment and understanding of human diseases. Variants of uncertain significance are perhaps the greatest current challenge in clinical genetics, and the availability of individuals' whole genomes has vastly increased the ascertainment of such variants without comparably aiding their interpretation. Yet, the field lacks a clear consensus on what kind of methods provide useful tools to interpret the data. For example, although there are now dozens of techniques for assessing the impact of missense single base variants on *in vivo* protein function, the accuracy and robustness of these methods are generally not known, and newer methods are often overlooked because of uncertainty about their performance. Critically, it is almost unclear what is the appropriate use of these and other methods for informing clinical decisions. CAGI aims to address these gaps, in order to help the broader community, understand the appropriate level of confidence they should have in variant prediction methods, and which classes of approaches are most suitable to a particular application.

To date, four CAGI experiments were conducted: a pilot experiment to test the methodology in 2010 (CAGI 1), and three full-scale events in 2011 (CAGI 2), 2013 (CAGI 3), and 2016 (CAGI 4). Participations in CAGI has increased with each experiment, including participants from academic, clinical, and commercial laboratories (See Table 3). The principles for the conduct of CAGI experiments are similar to those in use in other community experiments that evaluate the state-of-the-art in areas of computational biology, particularly those established by Critical Assessment of protein Structure Prediction (CASP)[44–46]. Challenges are constructed from unpublished datasets generously provided by collaborating academic, commercial and clinical laboratories. Phenotype predictions are made by academic and commercial groups without knowledge of the experimental answers. Performance of methods is evaluated in terms of agreement between the predictions and corresponding experimental data; and independent assessors, who do not participate as predictor, judge the significance of the results.

Addressing the challenges requires the development of new computational approaches that build on a common core of existing computational methods and knowledge, thus providing a community of potential participants with shared interests. CAGI challenges are chosen on the basis of two primary criteria: first, to probe the performance of methods as effectively as possible, over a broad range of genome interpretation scenarios. Second, to provide continuity over the CAGI experiments, so that it is possible to evaluate progress.

CAGI datasets are selected to reflect the range of challenges pertinent to assessing health-related phenotype prediction. Conditions include rare diseases, common traits and diseases, and germline and somatic cancer. The type of variation data used mirrors that encountered in current and imminent clinical practice, with a focus on genomes, exomes, SNPs, eQTL, splice-affecting SNPs, and CNVs as well as additional data such as transcriptomics. In general, challenges can only be finalized immediately before the prediction season. This is because datasets must be robust enough to share with predictors, but must not be publicly released before the conclusion of the prediction season.

| Edition | Number of challenges | Number of prediction submitted | Number of groups participating | Participating Countries | Conference |
|---|---|---|---|---|---|
| CAGI 4 (2016) | 11 | 191 | 37 | 13 | 25-27 March 2016, San Francisco (USA) |
| CAGI 3 (2012/2013) | 10 | 188 | 33 | 15 | 17-18 July 2013, Berlin (DE) |
| CAGI 2 (2011) | 11 | 114 | 21 | 16 | 9-10 December 2011, San Francisco (USA) |
| CAGI pilot (2010) | 6 | 108 | 17 | 8 | 10 December 2010, Berkeley (USA) |

**Table 3. Summary of the CAGI experiments.** Participation to the several editions of the CAGI experiments has always increased since the first pilot edition.

In several editions of CAGI, our group focused its attention on two main kind of challenges: genome oriented predictions and nonsynonymous variant challenges. Research exome and genomes challenges assess methods for interpretation of whole exome or whole genome sequence data, generally collected in case-control studies of complex diseases. Predictors are asked to provide a probability that each individual in the provided dataset displays the phenotype in question (i.e. probability that the individual has been diagnosed with the disorder). In the second kind of challenges, the ability to predict the functional impacts of variants in single protein in targeted assay is assessed. In general, quantitative prediction of the impact of a set of missense or nonsense single base variants on gene function are requested, reflecting measurements in targeted *in vitro* or *in vivo* assay. These challenges objectively compare published and unpublished methods for quantitative prediction of variant impact on phenotype. There are many such methods available, and CAGI seeks to assess their relative and absolute performance, albeit in the context of laboratory assays which may not accurately reflect clinical impact.

When prediction session is over, rigorous assessments are performed for each challenges. Assessor are instructed to use a variety of measures to highlight the different goals one might have in predicting the impact of genetic variation and also to reveal deeper insights into a method's strengths and weaknesses. In the end each CAGI experiment culminates in a conference at which participants report on and discuss their results. Predictors make presentations on their approaches, assessors make presentation on their evaluations of the submitted predictions and the state of the field is discussed.

Our group has a solid track record, participating in CAGI since its pilot edition in 2010 and having a particular focus especially on genome oriented challenges. We performed predictions for Crohn's disease, a chronic intestinal inflammation, for three editions since 2011, always with statistically significant results and ranking among best scoring groups for most of the times. Remarkable results has been even achieved during last edition for the Hopkins Clinical Panel (HCP acronym) challenge where genetic information were available only for some dozens of genes. Regarding nonsynonymous variant predictions, we participated with the role of assessor in the p16INK4A challenge dealing with the problem of giving back to the community a rigorous evaluation of predictor's performance and trying to pave the way for the definition of the state of the art methods in this field.

# 9 Thesis outline

This manuscript is organized in seven chapters. Chapter 2 is based on *Carraro et al., Performance of in silico tools for the evaluation of p16INK4a (CDKN2A) variants in CAGI. Hum. Mutat. doi:10.1002/humu.23235.* In this work I had the opportunity to face with one of the main critical issue of the NGS-revolution: interpretation of Variants of Unknown Significance (VUS). Most of the variants identified in genome and exome sequencing experiments are rare or even personal. For these kind of variants evidences of pathogenicity are predicted by bioinformatics tools, usually exploiting sequence conservation and secondary structure information. The focus of this work was to perform an extensive assessment of VUS pathogenicity predictors over an unpublished set of variants. In this work we exploited all possible benchmarking techniques producing one of the most complete assessment performed over the fourth editions of the CAGI experiment. All the evaluation scripts used in this work have been reimplemented by myself starting from a draft produced for the final conference of the third CAGI edition. Final discussion on predictors performance have been defined in association with the other authors of the paper.

Chapter 3 is based on *Chandonia et al., Lessons from the CAGI-4 Hopkins clinical panel challenge. Hum. Mutat. doi:10.1002/humu.23225.* In this work we dealt with the prediction of disease phenotypes from targeted enrichment sequences. Good prediction performance were achieved for this challenge, with our group ranking among best performers. Interesting for this challenge was the fact that several groups predicted individuals to be affected by pathologies that were not being diagnosed. In addition, some groups based right prediction on variants that were not considered to be causative by clinicians. In this way, central for this challenge was the emerging of another of the NGS-based personalized medicine crucial issues: the overinterpretation of variants and the presence of incidental findings. All the scripts used to define our group submissions have been defined and realized by myself. Variants selection and scoring system used for phenotype prediction has been defined in association with other colleagues from our group. Final performance assessment has been performed by the main authors of the paper.

Chapter 4 is based on *Giollo et al., Crohn disease risk prediction-Best practices and pitfalls with exome data. Hum. Mutat. doi:10.1002/humu.23177.* In this work we analyzed

performance of main methods proposed in the first three CAGI editions, tested on CAGI 4 Crohn's dataset. Critical aspect that biased performance evaluation in the previous CAGI editions of this challenge was found to be population structure. Due to the presence of such bias in datasets, performance of prediction methods were found to be overinflated and reliable performance assessment was possible only with test set of last CAGI edition. Thanks to this compared analysis performed over several editions, best practices in Crohn's disease risk prediction could be identified. In this work, my contribution was limited to the definition of some prediction methods and to results interpretation and discussion.

Chapter 5 is focused on the description of the Crohn's disease challenge in the fourth edition of CAGI experiment. In this part, the approach used to predict individuals phenotype for this complex disease using exome sequencing data, is described. Initially, analysis of population structure followed by identification of disease predisposing variants was performed. Trying to uncover genetic basis of missing heritability, a system biology approach was exploited, leading to the definition of new candidate genes involved in disease onset. All the analyses presented in this chapter have been planned and realized by myself (clustering analyses have been taken from[47], only for graphical reasons). Variants selection has been performed in association with other colleagues from our group. Assessment part of this chapter is based on *Daneshjouet al., Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. Hum. Mutat. doi:10.1002/humu.23280.*

Chapter 6 describes the upgrade of BOOGIE, a Java tool to predict blood phenotypes from sequencing data. The main update was performed on the prediction algorithm. In this project I had the opportunity to deal with the prediction of a completely genetic phenotype for which rules that link genetic variants with phenotypes have been only partially defined. Knowledge about blood groups definition has been extracted from on line resources and a great effort was put on the definition of the algorithm responsible for phenotype prediction. To this aim, trade-off between performance and computational resources has been managed to realize a prediction tool usable both in bioinformatics facilities and in clinics. All the analyses presented in this chapter have been planned and realized in association with other colleagues from our group.

In chapter 7 I summarized the main findings obtained in the previous chapters, describing their relevance in the research and clinical context.

# Performance assessment of *in silico* tools for the evaluation of p16INK4a (CDKN2A) variants in CAGI

This chapter is based on "Carraro, M. *et al.* Performance of in silico tools for the evaluation of p16INK4a (CDKN2A) variants in CAGI. *Hum. Mutat.* (2017). doi:10.1002/humu.23235".

# 1 Introduction

Genetic tests are nowadays become routinely applied to the investigation of disease-associated variants and relevant efforts are made by the scientific community to develop computational tools for genetic variant evaluation[48]. A number of methods presenting different strategies have been presented, and their application is becoming a common routine in cancer research[49,50]. *In silico* predictors are generally designed to provide a fast simplified response when compared to experimental screening protocols. However, lack of properly validated benchmarking represents the main limiting factor hampering wider application in a clinical scenario[51]. Variants affecting tumor suppressor genes, such as *TP53*[52], *VHL*[53] and *CDKN2A*[54] are actively investigated and collected in freely accessible databases[55–57]. However, the correct interpretation of their pathogenic significance is far being from definitively addressed. One relevant issue remains our ability to correctly predict disease-causing gene variants among variants of unknown significance (VUS)[58]. Correct prediction of susceptibility variants can foster the identification of molecular pathways causative of human diseases, particularly when variants affect well-understood genes previously validated by functional studies[59]. Since 2010, the Critical Assessment of Genome Interpretation (CAGI) experiment tries to objectively assess the state of the art of computational tools developed for genotype-phenotype determination. Here I will present the critical assessment of pathogenicity predictors applied to variants from the *CDKN2A* (OMIM ID: 600160) tumor suppressor also known as p16. *CDKN2A* is the major susceptibility gene identified in familial malignant melanoma. Approximately 40% of melanoma prone families worldwide have *CDKN2A* germline variants[60]. The *CDKN2A* locus maps to chromosome 9p21 and its regulation is particularly complex, involving alternative promoters, splicing and reading frames of shared coding regions. Two structurally unrelated tumor suppressors, p16INK4a and p14ARF, involved in cell cycle

regulation, are coded by alternative splicing of different first exons (1-α and 1-β). p16INK4a is a cyclin-dependent kinase (CDK4/6) inhibitor and p14ARF acts in *TP53* stabilization, binding and sequestering the MDM2 proto-oncogene[61,62]. Thus, alterations of this single locus compromises two important tumor suppressor pathways at the same time[63,64]. When associated with D-type cyclins, CDK4/6 promotes cell cycle progression through the G1 phase by contributing to the phosphorylation and functional inactivation of Retinoblastoma-associated protein[65,66]. Structurally, p16INK4a consists of four repeated ankyrin-type motifs, composed of two anti-parallel helices and a loop forming the CDK4/6 binding interface (See Figure 5).

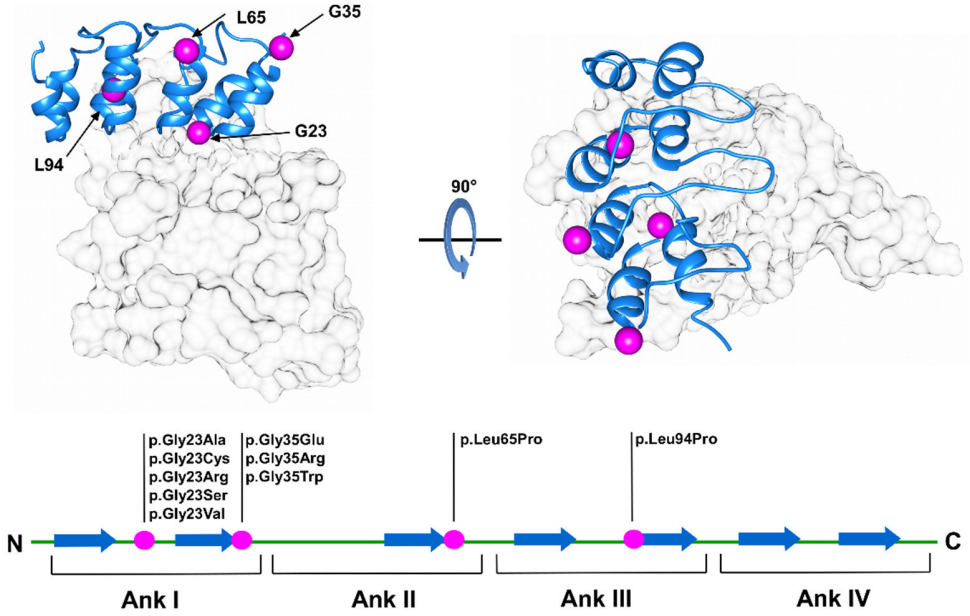


**Figure 5. Overview of CDK6-P16INK4A tumor suppressor complex.** Cartoon representations of the p16INK4a 3D structure (PDB code 1BI7) colored blue, while CDK6 is presented as full surface (light grey). Magenta spheres represent positions of variants considered for the challenge mapped on its surface. The ankyrin repeats composing p16INK4a structure are presented below with a schematic representation of mutated amino acid positions (magenta spots). Variant nomenclature refers to CDKN2A mRNA isoform1 (GenBank identifier: NM_000077.4), nucleotide numbering starts with the A of the ATG translation initiation site.

In the context of pathogenicity prediction, the ankyrin fold is challenging. Ankyrin repeats stack against one another to form a unique elongated single domain, with a multistate folding pathway conferring high structural plasticity. This highly modular nature confers

unique characteristics such as a high affinity for protein-protein interactions[67]. However, stack modularity can also be seen as a gradient of transiently folded states, where a single amino acid substitution may be able to interrupt p16INK4a-specific periodicity, causing a severe perturbation of the entire protein structure[68]. For this CAGI challenge, participants were asked to predict the effect of 10 CDKN2A variants in the p16-challenge, previously validated in cell proliferation rate assays. Twenty-two predictions using different strategies, e.g. scoring functions based on sequence conservation, or machine learning predictors, were assessed. The results allow us to propose where pathogenicity prediction might be improved, as methods combining information from different strategies were found to have the most promising results.

# 2 Materials and Methods

## 2.1 Dataset and classifications

The challenge includes 10 nucleotide variants affecting only the CDKN2A gene coding region without interfering with p14ARF. Each variant codes for a single amino acid substitution, with no insertions or deletions. The variant nomenclature used in this chapter refers to CDKN2A mRNA isoform1 (GenBank identifier: NM_000077.4). Participants were requested to perform predictions of the cellular proliferation rate for each of the 10 mutant protein as a percentage of the proliferation rate relative to pathogenic mutants (See Table 4). A proliferation rate of 100% is used for pathogenic variants (positive controls), and 50% for wild-type-like variants (negative controls). Predictors were also allowed to specify a prediction confidence (standard deviation) for each variant, with a maximum of six alternative submissions per group. The standard deviation was only reported for 14 submissions and the same confidence value was used for all predictions in 5 submissions. In a few cases, predictions have been manually rescaled during assessment as proliferation levels were wrongly reported as a fraction of 1 rather than 100 (where 100 represents the 100% positive control proliferation rate). A training set composed of 19 CDKN2A variants from[49,50] was also provided to the participants for training the prediction methods. This choice was justified based on the similar use of bioinformatics tools to predict CDKN2A variant effects on cell proliferation as verified by experimental assays. Bioinformatics predictions were described to be comparable with verified real values for most variants[49,50]. Real proliferation levels

obtained from the literature were rescaled between 0.5 and 1 (proliferation level of wild-type and disease-like phenotypes respectively).

| | | Proliferation rate | |
|---|---|---|---|
| Nucleotide variant | Protein variant | Average | Standard Deviation |
| c.67G>A | p.Gly23Ser | 0.69 | 0.04 |
| c.67G>C | p.Gly23Arg | 0.91 | 0.14 |
| c.67G>T | p.Gly23Cys | 0.86 | 0.13 |
| c.68G>C | p.Gly23Ala | 0.53 | 0.09 |
| c.68G>T | p.Gly23Val | 0.90 | 0.1 |
| c.103G>A; c.103G>C | p.Gly35Arg | 0.53 | 0.02 |
| c.103G>T | p.Gly35Trp | 0.86 | 0.09 |
| c.104G>A | p.Gly35Glu | 0.60 | 0.11 |
| c.194T>C | p.Leu65Pro | 0.66 | 0.1 |
| c.281T>C | p.Leu94Pro | 0.93 | 0.13 |

**Table 4. p16INK4a proliferation rate test set.** Identifiers of variants affecting cell proliferation and relative proliferation level. Variant nomenclature refers to CDKN2A mRNA isoform1 (GenBank identifier: NM_000077.4), nucleotide numbering starts with the A of the ATG translation initiation site**.** Proliferation levels were rescaled between 0.5 (WT-like phenotypes) and 1 (tumor-like phenotypes).

## 2.2 In vitro proliferation assay of CDKN2A variants and data normalization

The experimental validation of the pathogenic effect of the variants used in CAGI is described in detail in[54]. Briefly, the full-length *CDKN2A* cDNA was cloned in the pcDNATM3.1 D/V5-His-TOPO®_expression vector (Invitrogen, Life Technologies Corporation, Carlsbad, CA), engineered by site-specific mutagenesis (QuikChange® II XL Site-Directed Mutagenesis Kit; Stratagene, CA), and finally transfected in U2-OS human osteosarcoma cells (p16INK4a and ARF null, p53 and pRb wild type), as previously described[54,69]. Three controls, no vector (G418 selection control), pcDNA3.1–EGFP (positive, variant-like control), and pcDNA3.1–p16INK4a wild-type (negative control),

were included in each experiment. All variants were independently tested at least three times. The proliferation rate (PR) was calculated as a percentage of the proliferation of variant transfected-cells (average of all replicates) at day 8 relative to the proliferation of EGFP-transfected cells, which was set as 100%. Transfection with wild-type *CDKN2A* induced a detectable, substantial growth inhibition (proliferation rate 50%), whereas various p16INK4a variants had different effects on cell proliferation, from wild-type-like to loss-of-function. The proliferation rates used for CAGI are shown in Table 4.

## 2.2 Performance assessment

Evaluating the performance of bioinformatics tools in predicting VUS impact is a non-trivial task. The assessment should not be seen as a mere discrimination of winners/losers, but rather aim at identifying which tool generated the most reliable prediction. A considerable number of performance measures were considered in order to perform a thorough assessment. The final goal was to generate a global overview of the strengths and weaknesses of each method. Correlation indices were considered first, as predictions are in a continuous range (cell proliferation rate). Both the Pearson (PCC) and Kendall's Tau correlation coefficients (KCC) were calculated. Both range from +1 (perfect positive correlation) to -1 (perfect inverse correlation) with 0 representing a random performance. Root Mean Square Error (RMSE) was calculated to better estimate the difference between predicted and real values. To further assess the prediction reliability in a medical setting, a binary classification was used. Proliferation levels were divided in two classes, benign and pathogenic, with three different proliferation thresholds suggested by the data provider, i.e. potentially pathogenic (>65%), probably pathogenic (>75%) and likely pathogenic (>90%). The Area under the ROC curve (AUC) for each classification threshold was also calculated. The standard deviation of the predicted proliferation rate was used to calculate the fraction of Predictions Within Standard Deviation (PWSD). To address issue related to missing and very large confidence range, PWSD was calculated assuming a standard deviation of 10% for all submissions (PSWD10). All performance indices are presented in Table 6. To assess the statistical significance of each performance index, 10,000 random predictions were generated and used to calculate an empirical continuous probability (score *s*), with a p-value defining the proportion of random predictions scoring > *s*.

# 3 Results

## 3.1 Participation and similarity between predictions

In the p16INK4a CAGI challenge, participants were requested to predict the effects of ten p16INK4a VUS potentially causing malignant proliferation validated with cellular proliferation assays[54]. This challenge attracted 22 submissions from ten participating groups, which were assessed without knowing the identity of the predictors. After the assessment was completed, only one group remained anonymous. Table 5 lists the participating groups, their submission IDs and main features used for prediction. The majority of methods used evolutionary information derived from multiple-sequence alignments for prediction. Several methods also used the available crystal structure of p16INK4a bound to CDK6 (See Figure 5) to calculate folding energies. Combinations of both approaches or of different predictors were also submitted. A summary for each method could be found in the Supplementary Material of the on-line version of this work. Of the ten participating groups, four contributed one prediction, one submitted two, four submitted three and only one group submitted four different submissions.

An analysis of prediction similarity was performed to better highlight the peculiarity of each submission. Almost all groups performing multiple submissions made very similar predictions (See Figure 6). This is particularly evident for predictions from the Bromberg group, which were *de facto* mostly identical for many variants. A similar situation can be drawn for the Moult group, where a different fitting of two linear models (submissions 9, 15) produced identical predictions for most variants. The third prediction (submission 20) was defined by a different rescaling process of submission 15. Submissions 9 and 15 both predicted a majority of variants between 0.88 and 1. Predictions from the Gough and BioFolD groups are also quite strongly correlated among each other. Interestingly, submissions 5 and 3 (BioFolD and Casadio lab, respectively) are also highly correlated as both are based on two versions of the SNPs&GO method[70,71]. The Vihinen lab (submissions 6, 13) presents a weak anticorrelation among its predictions, probably due to fact that predictions for all except one variant were very high (>=0.85). The four submissions from Yang&Zhou lab (10, 16, 21, 22) present almost no correlation, possibly also due to a sign error affecting three submissions.

| Submission ID | Group ID | Prediction features |
|---|---|---|
| Submission 1 | Anonymous | / |
| Submission 2 | Bromberg Lab. | conservation, annotation |
| Submission 3 | Casadio Lab. | conservation, Gene Ontology |
| Submission 4 | Lichtarge Lab. | conservation |
| Submission 5 | BioFolD Lab. | conservation, Gene Ontology |
| Submission 6 | Vihinen Lab. | meta-predictor |
| Submission 7 | Dunbrack Lab. | protein structure |
| Submission 8 | Gough Lab. | conservation |
| Submission 9 | Moult Lab. | meta-prediction |
| Submission 10 | Yang&Zhou Lab. | conservation, folding energy |
| Submission 11 | Bromberg Lab. | conservation, annotation |
| Submission 12 | BioFolD Lab. | conservation, Gene Ontology |
| Submission 13 | Vihinen Lab. | conservation, amino acid features, Gene Ontology |
| Submission 14 | Gough Lab. | conservation |
| Submission 15 | Moult Lab. | meta-prediction |
| Submission 16 | Yang&Zhou Lab. | conservation |
| Submission 17 | Bromberg Lab. | conservation, annotation |
| Submission 18 | BioFolD Lab. | meta-prediction |
| Submission 19 | Gough Lab. | conservation |
| Submission 20 | Moult Lab. | meta-prediction |
| Submission 21 | Yang&Zhou Lab. | folding energy |
| Submission 22 | Yang&Zhou Lab. | folding energy |

**Table 5. Predictor overview.** For each submission, predictor and a summary of features used for prediction are indicated.

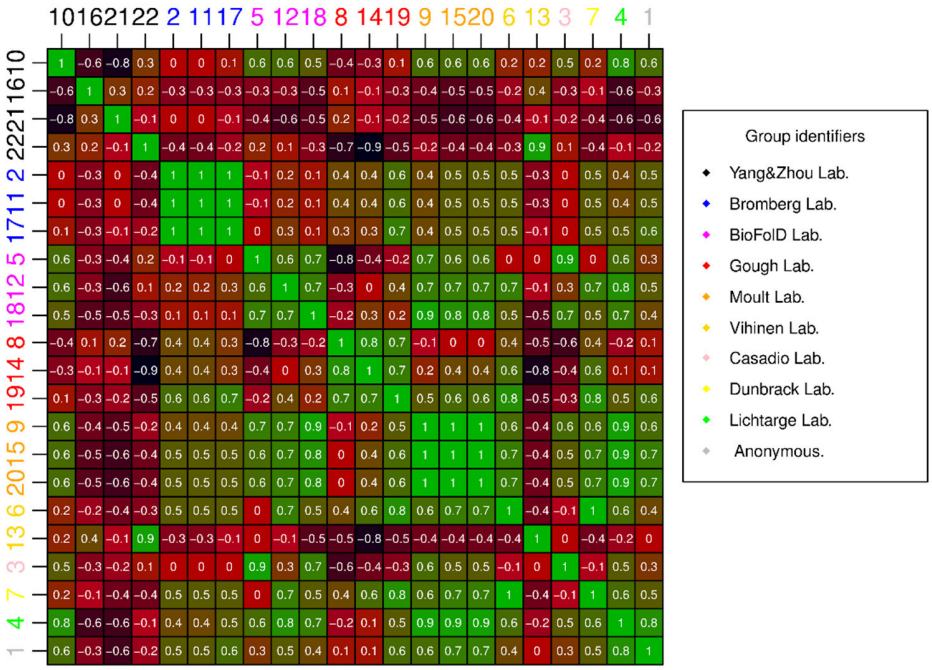


**Figure 6. Correlation among submissions.** Each cell shows the Pearson correlation coefficient between two submissions, with a color scale ranging from green (+1, perfect correlation) to red (0, no correlation) and black (-1, perfect anti-correlation). Submissions are clustered by group.

## 3.2 Assessment criteria and performance measures

The type of insights to be gained from assessing a CAGI challenge depends strongly on the criteria used for evaluation. As this is a relatively novel field, extra care was given to this point. Ideally, the criteria should reflect the true performance of the methods, highlighting submissions which are of practical relevance. The simplest measures, binary classification and derived measures such as AUC, suffer from the choice of an arbitrary threshold that may obfuscate interesting results. Correlation measures are good to indicate overall trends, but of little use to guide the selection of pathogenic cases as no threshold is used. At the other numerical extreme, RMSE is very clear, but can result in poor performance for all submissions. For an inherently continuous prediction challenge such as p16, determining the number of predictions within a fixed distance can arguably provide a measure combining features of binary classification and correlation. In order to understand better how related the assessment criteria are among each other, their correlation was plotted (See Figure 7). The PCC and KCC correlation coefficients are

highly correlated with each other and with the three AUC measures. RMSE and two PWSD variants are less correlated and offer two alternative views of the data.

Using a reduced set of measures for the final ranking is suggested by the high pairwise correlation coefficients, suggesting they are measuring very similar features (see Figure 7). A ranking including largely orthogonal measures should prove more robust and informative. For this reason, only four measures (one for each group) with low pairwise correlation were considered for the final ranking, i.e. Kendall Correlation Coefficient (KCC), Root Mean Square Error (RMSE), Area Under the Curve considering a 75% of proliferation threshold (AUC75) and Prediction Within Standard Deviation considering a standard deviation of 10% for all submission (PWSD10). In particular, KCC was chosen as it is a rank-based measure appropriate when targets are continuous and their relative order is critical. The data provider recommended to use AUC75, as the corresponding proliferation level appeared to be the best threshold to separate pathogenic and neutral phenotypes. Finally, PSWD10 was preferred over PSWD as many predictors did not report standard deviation for their submissions.

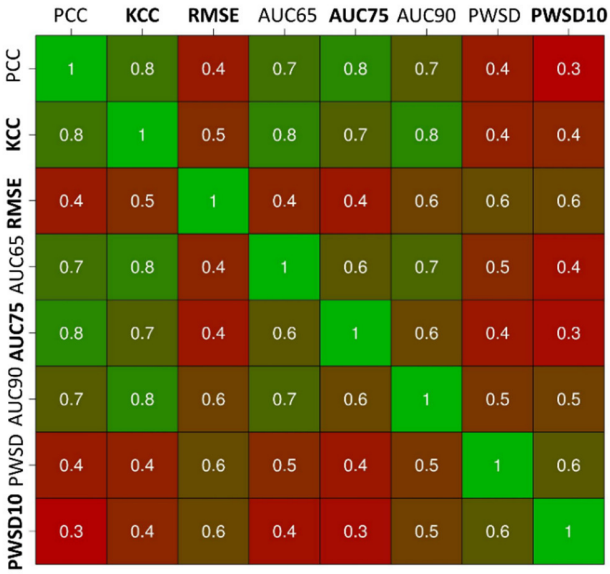


**Figure 7. Correlation among performance indices.** Each cell shows the Kendall correlation coefficient between the two corresponding measures, with a color scale ranging from green (+1, perfect correlation) to red (-1, perfect anti-correlation). Notice how similar measures tend to cluster together. The four selected measures are highlighted in bold face.

## 3.3 Performance evaluation

The assessment of performance achieved by the 22 methods showed many predictions to have good results on average. This is particularly true considering AUC75, where most of the submissions achieved values between 0.7 and 1. For KCC, the average of the submissions shows a moderate to strong correlation with real data (See Table 6). Good results were however not sufficient for most predictions to be statistically significant. Very demanding thresholds emerged to separate significant results from random for this challenge, with only the top ranking methods being significant for most of the 4 performance indices. This is probably due to the limited number of variants present in the test set, where wrong prediction of one variant corresponds to 10% of the dataset. Small variations in predictions could be reflected in remarkable fluctuation of performance indices due to the small number of variants considered. To perform a global assessment of predictor performance we therefore decided to focus more on ranking than on numerical values achieved for each measure. Ranking variations not only may better reflect the magnitude of performance variation, but can also be considered more intuitive for non-specialist readers. The Yang&Zhou lab (submission 10) performed best, ranking first in all performance indices except AUC75, where it is fifth (See Table 7). Higher AUC75 values were obtained by the Lichtarge lab (submission 4), an anonymous prediction (submission 1) and the Moult lab (submissions 15, 20). The Lichtarge lab also obtained good results considering KCC, where it ranked second. BioFolD (submission 5) also achieved good results, ranking second for both PSWD10 and RMSD and third for KCC. Furthermore, the BioFolD lab also performed well with submission 12, being second and third for PSWD10 and RMSD, respectively. Among the lower ranked predictions, an inverse correlation is found for Submission 8 (-0.40), mainly resulting from low proliferation levels being predicted when real proliferation levels were high. Submissions 16 and 21 rank poorly, achieving an inverse KCC correlation (-0.56, -0.6). Notably, while all three submissions perform poorly, they probably followed opposed strategies. Submission 8 tends to be very conservative, with most of the predicted values close to a wild-type phenotype. Submissions 16 and 21 tend to be more biased towards the prediction of malignant phenotypes, with only one predicted value close to a milder phenotype. This trend seems to be shared among lower ranking predictions.

A statistical test of the average ranking over all four performance measures, confirmed submission 10 (Yang&Zhou lab) as the best performer. No statistically significant difference can be identified between submissions 4 and 5 (Lichtarge, BioFolD; See Figure 8) ranked second and third, respectively. A bootstrap simulation with 10,000 replicas was used to test whether the performance achieved by the three best submissions could be achieved by chance. Submission 10 performs better than random (p-value < 0.05) for three out of four measures, the only exception being PSWD10. Submissions 4 and 5 perform better than random only considering KCC and AUC75 (See Table 8).

| Submission | PCC | KCC | RMSE | AUC65 | AUC75 | AUC90 | PWSD | PWSD10 |
|---|---|---|---|---|---|---|---|---|
| S1 | 0.83 | 0.45 | 23.51 | 0.81 | **1** | 0.76 | 5 | 3 |
| S2 | 0.33 | 0.02 | 21.29 | 0.57 | 0.62 | 0.55 | 3 | 2 |
| S3 | 0.53 | 0.47 | 25.5 | 0.83 | 0.7 | 0.64 | 2 | 2 |
| S4 | **0.84** | 0.63 | 16.48 | 0.81 | **1** | **1** | 4 | 5 |
| S5 | 0.66 | 0.6 | 15.81 | 0.9 | 0.88 | 0.9 | **7** | 6 |
| S6 | 0.23 | 0.34 | 25.67 | 0.57 | 0.58 | 0.79 | 2 | 3 |
| S7 | 0.22 | 0.2 | 18.2 | 0.57 | 0.68 | 0.62 | 3 | 4 |
| S8 | -0.34 | -0.4 | 39.21 | 0.19 | 0.42 | 0.26 | 1 | 1 |
| S9 | 0.7 | 0.38 | 20.18 | 0.86 | 0.88 | 0.71 | 3 | 3 |
| S10 | 0.83 | **0.69** | **9.24** | **1** | 0.92 | **1** | **7** | **7** |
| S11 | 0.33 | 0.02 | 21.29 | 0.57 | 0.62 | 0.55 | 2 | 2 |
| S12 | 0.57 | 0.47 | 15.93 | 0.67 | 0.84 | 0.9 | 4 | 6 |
| S13 | 0.11 | 0.05 | 20.08 | 0.57 | 0.42 | 0.64 | 5 | 5 |
| S14 | -0.22 | -0.4 | 23.29 | 0.19 | 0.42 | 0.26 | 5 | 5 |
| S15 | 0.76 | 0.51 | 18.83 | 0.86 | 0.96 | 0.81 | 4 | 3 |
| S16 | -0.45 | -0.56 | 22.48 | 0.12 | 0.08 | 0.14 | 2 | 2 |
| S17 | 0.43 | 0.25 | 21.8 | 0.67 | 0.72 | 0.57 | 2 | 2 |
| S18 | 0.46 | 0.28 | 16.35 | 0.67 | 0.72 | 0.76 | 6 | 2 |
| S19 | 0.3 | 0.07 | 20.3 | 0.45 | 0.76 | 0.55 | 2 | 3 |
| S20 | 0.76 | 0.51 | 17.7 | 0.86 | 0.96 | 0.81 | 4 | 4 |
| S21 | -0.62 | -0.6 | 23.71 | 0.19 | 0.12 | 0 | 2 | 2 |
| S22 | 0.15 | 0.2 | 18.45 | 0.6 | 0.4 | 0.76 | 3 | 3 |

**Table 6. Performance indices.** Results are shown for the main performance indices considered in the assessment. The top performing submission in each category is shown in bold and the second best is underlined.

| | Rank | | | | | |
|---|---|---|---|---|---|---|
| **Submission** | **KCC** | **RMSE** | **AUC75** | **PWSD10** | **Average** | **Overall** |
| S1 | 8 | 18 | **1** | 9 | 9 | 8 |
| S2 | 17 | 13 | 14 | 15 | 14.75 | 18 |
| S3 | 6 | 20 | 12 | 15 | 13.25 | 15 |
| S4 | 2 | 5 | **1** | 4 | 3 | 2 |
| S5 | 3 | 2 | 6 | 2 | 3.25 | 3 |
| S6 | 10 | 21 | 16 | 9 | 14 | 16 |
| S7 | 14 | 7 | 13 | 7 | 10.25 | 10 |
| S8 | 19 | 22 | 17 | 22 | 20 | 22 |
| S9 | 9 | 11 | 6 | 9 | 8.75 | 7 |
| S10 | **1** | **1** | 5 | **1** | **2** | **1** |
| S11 | 17 | 13 | 14 | 15 | 14.75 | 18 |
| S12 | 7 | 3 | 8 | 2 | 5 | 4 |
| S13 | 16 | 10 | 17 | 4 | 11.75 | 12 |
| S14 | 19 | 17 | 17 | 4 | 14.25 | 17 |
| S15 | 4 | 9 | 3 | 9 | 6.25 | 6 |
| S16 | 21 | 16 | 22 | 15 | 18.5 | 20 |
| S17 | 12 | 15 | 10 | 15 | 13 | 14 |
| S18 | 11 | 4 | 10 | 15 | 10 | 9 |
| S19 | 15 | 12 | 9 | 9 | 11.25 | 11 |
| S20 | 4 | 6 | 3 | 7 | 5 | 4 |
| S21 | 22 | 19 | 21 | 15 | 19.25 | 21 |
| S22 | 13 | 8 | 20 | 9 | 12.5 | 13 |

**Table 7. Submission ranking.** Ranking of the different prediction methods based on performance indices in Table 1. To define the final ranking average of ranking position for each performance index was used. The top performing submission in each category is shown as bold, while underlined is for the second best performance.

| | **S10** | **S4** | **S5** |
|---|---|---|---|
| **KCC** | **0.015** | **0.015** | **0.015** |
| **AUC75** | **0.029** | **0.004** | **0.048** |
| **RMSE** | **0.006** | 0.222 | 0.151 |
| **PWSD10** | 0.059 | 0.389 | 0.183 |

**Table 8. Statistical significance test for top three submissions**. The p-value for random predictions scoring better using each assessment metric is shown over 10,000 simulations. P-values < 0.05 are shown as bold.

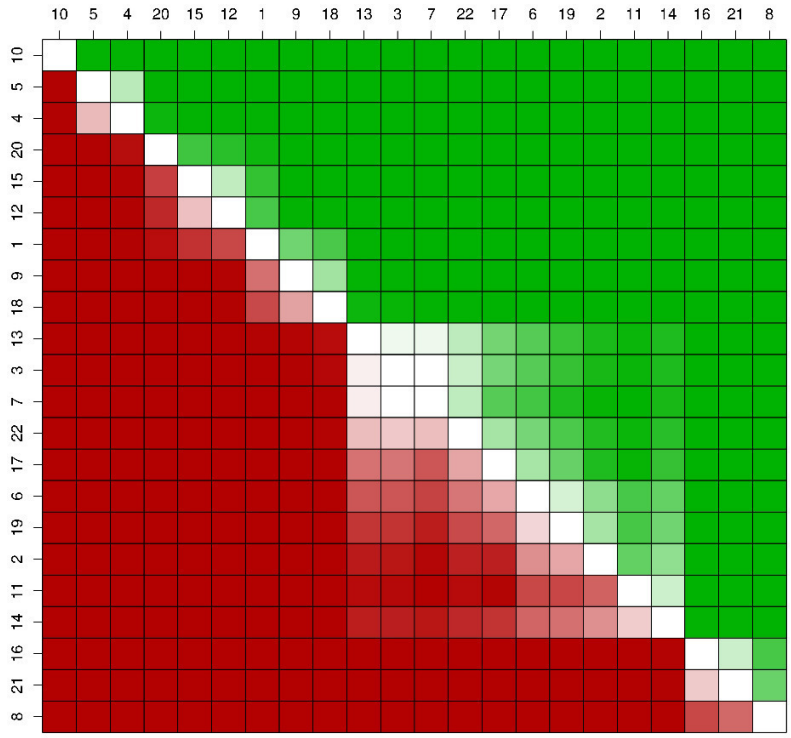


**Figure 8. Pairwise difference between submissions.** Statistical differences between submissions based on the overall ranking achieved by each submission. Submissions were sorted in agreement with the final ranking. White squares are indices of tied predictions (P-values > 0.05) meaning that performances are similar and the difference between two predictors is not statistically significant.

## 3.4 Difficult variants

An analysis of submissions shows prediction reliability to depend on position, with p.Gly23Ser, p.Gly35Glu and p.Gly35Arg being particularly complex to address (See Table 9). p.Gly23Ser and p.Gly35Arg are the most mispredicted variants using PWSD10, with only two correct predictions. Both variants affect conserved positions that are known to have role in correct p16INK4a folding and CDK inhibition. A previous study[54] addressing the same genetic changes showed p.Gly23Ser to introduce a weak interaction with S56. Although weak, this is thought to stabilize the overall fold, inducing a small local rearrangement of the p16-CDK4/6 binding interface. Predictions seem to miss this twofold effect. The p.Gly23Ser variant is mainly predicted as damaging, suggesting that current methods over-predict a pathogenic effect. A similar scenario can be seen for

p.Gly35Glu and p.Gly35Arg. The G35 is a solvent-exposed residue, which localizes at the end of the first α-helix in the p16INK4a structure. Substitution of G35 with charged residues can be accommodated in the Ankyrin fold, likely yielding neutral phenotypes[54] mispredicted in this case. The only notable exception is submission 20, which shows the best accuracy with these difficult variants but misses most of the other variants. The p16INK4a challenge shows how different variants on the same residue can have widely diverging effects which are not well predicted by many submissions.

| | G23S | G23R | G23C | G23A | G23V | G35R | G35W | G35E | L65P | L94P |
|---|---|---|---|---|---|---|---|---|---|---|
| **S1** | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| **S2** | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **S3** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **S4** | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| **S5** | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| **S6** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| **S7** | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| **S8** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **S9** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| **S10** | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 |
| **S11** | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **S12** | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| **S13** | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 |
| **S14** | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| **S15** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| **S16** | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **S17** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| **S18** | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **S19** | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| **S20** | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| **S21** | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **S22** | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **Total** | 2 | 13 | 12 | 5 | 9 | 2 | 10 | 4 | 6 | 12 |

**Table 9. Correct predictions per variant.** Submissions are shown as rows, followed by the total count of correct predictions. Columns list each variant of the p16INK4a challenge and whether the corresponding submission correctly predicted (1, grey background) the effect according to PWSD10. Notice how certain substitutions at the same position were more difficult to predict than others.

# 4 Conclusions

Pathogenicity prediction of VUS is a challenging problem. It can manifest at different levels, such as protein function, sub-cellular localization and pathways, as well as impairing multiple interactions a specific protein can exert with different partners[72]. Pathogenicity predictions are frequently performed through *a priori* knowledge of the biological problem, in most cases from an experimental characterization of disease-associated variants. *In silico* prediction can be considered a realistic benchmark of our understanding of these biological problems. In this chapter I presented, we presented results from the critical assessment of 22 different predictions in the CAGI p16INK4a challenge. Different submissions were compared to highlight the strengths and weaknesses of prediction strategies as applied to the human tumor suppressor p16INK4a. The challenge had several peculiar characteristics. p16INK4a is a cancer-associated kinase inhibitor whose main function is protein-protein binding. It is also an Ankyrin repeat protein, characterized by repetitive local short-range interactions[54,68]. In an ideal scenario, a reliable pathogenicity predictor should discriminate variations affecting both p16INK4a features. From a computational point of view, most predictors use Position-Specific Scoring Matrices (PSSM) and machine learning. The assessment suggests that our knowledge is sufficient to perform reliable predictions for most of the analyzed variants. However, relevant differences emerged among predictions. These differences stem in part from the strategy used for pathogenicity assessment. Others arise from expert knowledge, with similar approaches generating discordant predictions. Groups combining different strategies seem more robust when predicting CDKN2A variants. Predictions supplied from the Yang&Zhou lab are emblematic of this phenomenon. This group contributed four different submissions, rescaling PSSM value differences between wild type and variants, computing ΔΔG variation with ROSETTA3[73], computing ΔΔG with Dmutant[74] or combining them in a support vector machine using a linear kernel. Our assessment showed the Yang&Zhou lab reliability improving with prediction complexity (See Tables 6 and 7), peaking with the most complex submission 10. A similar reliability gradient was observed for other groups using different strategies, suggesting how a single method may be insufficient for pathogenicity prediction. Submission 10 presents the best fit with experimental data. On the other hand, a sub-optimal AUC75 suggests the submission is not the most convenient for discriminating

pathogenic from a wild-type-like phenotype. Conversely, submission 4 (Lichtarge group) presents the best AUC75 value, which may make it useful in a clinical setting. However, submission 4 predicts all variants as pathogenic at this threshold, which probably renders this method unreliable for clinical practice. Prediction performance seems to be also influenced by variant type. For example variants affecting glycine 35 are on average easier to predict than glycine 23. The latter is known to be relevant for the correct Ankyrin fold[68], as well as to localize at the p16INK4a/CDK4/6 binding interface[50,54]. For a generic pathogenicity predictor this may be the worst case scenario. Sequence conservation analysis highlights the residue as conserved and relevant for protein structure, but may miss the pathogenic effect caused by interference at the protein-protein interaction interface. More advanced approaches, such as Hidden Markov Models and neural networks, turned out to be the best strategies for this specific problem. It can be argued that the limited number of variants composing the dataset may limit generalization of the results and a larger set of variants might produce a different ranking. The dataset was chosen to represent a balanced ratio between pathogenic and neutral variants. Despite these intrinsic limitations, we believe this challenge may be representative of a clinical setting, where disease-associated genes are poorly described when it comes to variants found in patients. It is evident from this assessment that no method is able to perform errorless predictions. We expect the CAGI results to provide a starting point to further improve the available methods for VUS pathogenicity predictions.

# Phenotype prediction in the CAGI 4 Hopkins clinical panel challenge.

This chapter is based on "Chandonia, J.-M., Adhikari, A., Carraro, M., Chhibber, A., Cutting, G.R., Fu, Y., Gasparini, A., Jones, D.T., Kramer, A., Kundu, K., Lam, H.Y.K., Leonardi, E., Moult, J., Pal, L.R., Searls, D.B., Shah, S., Sunyaev, S., Tosatto, S.C.E., Yin, Y., Buckley, B.A., 2017. Lessons from the CAGI 4 Hopkins clinical panel challenge. Hum. Mutat. doi:10.1002/humu.23225".

# 1 Introduction

DNA sequencing tests are increasingly used in medical practice to confirm or assign clinical diagnoses[75]. However, the interpretation and classification of novel sequence variants identified in a patient remains difficult, even for well-studied disorders like cystic fibrosis[76]. Improved computational methods may aid in the interpretation of sequence variants and, when used in conjunction with clinical data, could increase the confidence of a diagnosis[77]. Until recently, genetic testing was limited to genes associated with a specific clinical phenotype. However, recent technological advances have made it feasible to sequence large gene panels, exomes, and genomes[78–80]. As the number of genes sequenced per patient increases, the number of novel, rare, and unclassified variants also increases. Clinical molecular geneticists must determine which variants, if any, are likely to contribute to the patient's clinical presentation. The current gold standards for assessing a variant's pathogenicity are segregation of the variant with the clinical phenotype in multiple pedigrees, and functional assays demonstrating a detrimental effect of that specific nucleotide change. In most instances, when a novel genetic variant is identified there is no rapid and reliable method to assess its pathogenicity. Predictive software tools are interrogated, but none are considered strong evidence to assert a novel variant's pathogenicity[81]. The shift towards analyzing large datasets has led to a need for high-throughput methods to aid in variant classification and also for computation tools to help better interrogate the increasing number of variants of uncertain clinical significance.

Crowd sourced data analysis challenges such as the 4th Critical Assessment of Genome Interpretation (CAGI 4) have emerged as a framework to compare predictive methods and assess the overall state of particular analysis areas [82]. In the CAGI 4 Hopkins Clinical

Panel challenge, participants were asked to develop or use existing computational methods to analyze data from a next generation sequencing (NGS) panel in order to match a patient's genotype to their clinical phenotype in the absence of additional clinical information. The Johns Hopkins DNA Diagnostic Laboratory (henceforth, Hopkins), a CLIA and CAP certified lab that specializes in clinical molecular testing for rare, inherited disorders, provided data for this challenge. The Hopkins lab offers testing for approximately 50 phenotypes and disorders totaling 3,500 tests annually. They offer NGS-based tests targeted for ~20 specific phenotypes. The same NGS capture probe set is used for all panels and only the requested genes are analyzed in each patient. Hopkins provided CAGI 4 organizers with the VCF files for the entire NGS panel for 106 patients with a range of clinical presentations. The genetic disorders associated with variants in the 83 genes on the panel were grouped into 14 'disease classes' which include lung disorders, peroxisomal disorders, aneurysm disorders and craniofacial disorders (See Table 10). The goal of the challenge was for the participants to match each patient to a disease class based on informatics analysis of the sequence data. A further part of the challenge was to predict the specific gene and variant(s) that is/are the underlying cause of disease.

# 2 Materials and methods

## 2.1 Sequencing, variant calling, and analysis by the Hopkins lab

Gene sequences were captured using one of two custom probe sets (Agilent SureSelectXT Target Enrichment Kit) and sequenced by a NGS platform (Illumina MiSeq, 2x100 nt reads). The NGS panels used to test assessed exons and exon-adjacent sequences for 64 or 83 loci. Sequences were aligned to the human reference genome (GRCh37/hg19) using the Burrows-Wheeler Aligner (bwa). Sequence variants were called individually for each patient to produce two Variant Call Format (VCF) files, one for single nucleotide variants (SNVs; GATK UnifiedGenotyper, v2.7-4) and one for insertion-deletion variants (InDels; GATK HaplotypeCaller, v2.7-4). Deidentified VCF files were provided to the CAGI 4 organizers. Note that the CAGI 4 organizers combined individual VCF files for each patient into a single VCF, resulting in potentially misleading data in the INFO and FILTER fields of the file. The panel of 83 genes was sequenced in 96 of the 106 patients; for the other

10 patients, a partially overlapping list of 64 genes were sequenced. Although the whole NGS panel was sequenced in all patients, only the genes selected on the patient's test requisition form were analyzed by the lab (n=1-24 genes/patient).

For more information on the specific NGS tests offered by the lab refer to the Hopkins lab website (http://www.hopkinsmedicine.org/dnadiagnostic/tests/).

The Hopkins lab included variants in the genes they analyzed that were classified as Variants of Uncertain Significance (VUS), Likely Pathogenic, and Pathogenic as an answer key. The disease class of each patient was also provided in the answer key and reflects the test selected by the patient's physician on the test requisition form. The ~20 phenotypes that Hopkins tests for were narrowed down to 14 disease classes in order to simplify the challenge. Some disease classes were not represented by any patients and were included as red herrings.

## 2.2 Challenge format

Participants in the Hopkins clinical panel challenge were provided with the two VCF files above, a detailed description of the 14 disease classes given in Table 10, a submission template, a submission validation script, and the gene capture regions used in sequencing the patients (in Browser Extensible Data, or BED format). Participants were also instructed that every patient matched exactly one disease class.

Participants were asked to submit predictions of each patient's disease class based on their gene panel sequences, along with predicted causal variant(s). Each participant was allowed to submit up to six distinct submissions, in which each submission contained predictions for each patient. For each submission, participants were required to predict the probability that the patient has a referring disease in each of the 14 disease classes in the provided list, 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 probability of disease class also included a mandatory standard deviation (SD) field indicating confidence in the prediction, with low SD indicating high confidence, and high SD indicating low confidence.

| Disease class | Description |
|---|---|
| **Cystic fibrosis and CF-related disorders** | Classic cystic fibrosis consists of progressive lung disease, exocrine pancreatic insufficiency, and male infertility. |
| **Diffuse lung disease** | Diffuse lung disease is an umbrella term encompassing multiple lung disease phenotypes. |
| **Primary ciliary dyskinesia** | Primary ciliary dyskinesia is a genetically heterogeneous group of disorders resulting from dysfunction in different parts of the cilia. |
| **Peroxisomal beta-oxidation defects** | The majority of patients with peroxisomal beta-oxidation defects have liver disease, brain malformations, developmental retardation, sensory deficits, and dysmorphic craniofacial features. |
| **Rhizomelic chondrodysplasia punctata** | Symptoms of rhizomelic chondroplasia punctata include proximal shortening of the limbs, cataracts, severe intellectual disability, seizures, and calcific stippling of cartilage. |
| **Zellweger spectrum disorders** | Zellweger spectrum disorders consist of Zellweger syndrome (cerebro-hepato-renal syndrome; most severe phenotype), neonatal adrenoleukodystrophy (intermediate phenotype), and infantile Refsum disease (mildest phenotype). |
| **Loeys-Dietz syndrome** | Loeys-Dietz syndrome is a connective tissue disorder that predisposes individuals to aortic aneurysms. |
| **Marfan syndrome** | Marfan syndrome is an inherited connective tissue disorder that affects the skeletal, ocular, and cardiovascular systems. |
| **Thoracic aortic aneurysm and dissection** | Thoracic aortic aneurysm and dissection is a cardiovascular disease characterized by dilation of the aorta, which leads to aortic aneurysms (most commonly in the ascending aorta) and aortic dissection. |
| **Ataxia telangiectasia** | Ataxia-telangiectasia is a disorder of childhood onset progressive cerebellar ataxia and occulocutaneous telangiectasias. |
| **Liddle syndrome** | Liddle syndrome is a rare genetic disorder characterized by early onset high blood pressure (hypertension) and low blood potassium (hypokalemia). |
| **Pseudohypoaldosteronism type 1** | Pseudohypoaldosteronism type 1 is a salt-wasting disease with onset during infancy. |
| **Telomere shortening disorders** | Telomere shortening disorders represent a spectrum of phenotypes that result from mutations in genes involved in telomere maintenance protein complexes. |
| **Treacher Collins and related syndromes** | Treacher Collins syndrome is a rare disorder affecting craniofacial development. |

**Table 10. Summary of the 14 disease classes.** Disease classes and description for the CAGI 4 Hopkins clinical panel challenge are provided in the table.

## 2.3 Assessment

Formatting errors in all submissions were corrected to the best of the assessor's ability, and redundant submissions were removed. Predicted disease classes made in each submission for each patient were assessed against the correct disease class given in the Hopkins answer key, using the metrics described below. The predicted causal variant(s) were also compared to interpretations from the clinical laboratory, but because these are not known with certainty, such predictions cannot be rigorously assessed. In their answer key, Hopkins noted which variants they regarded as Variants of Uncertain Significance (VUS), Likely Pathogenic, and Pathogenic; however, for purposes of matching participants' predictions to the answer key, all variants noted by Hopkins for each patient were treated equivalently. Assessors first calculated the number of correct predictions of disease class made in each submission. For each patient, the predicted disease class was the one assigned the highest probability among all 14 disease classes. Ties (i.e., cases where multiple disease classes were all assigned the highest probability) were handled as described below.

1. If all 14 probabilities for a patient were equal (e.g., all zeroes), those predictions were not counted in the following three metrics.
2. In other cases, assessors calculated one metric (nCorrect) in which the number of correct predictions was counted, giving ties full credit; another metric ($nCorrect_{tie}$) was calculated in which N-way ties were given 1/N credit.
3. Finally, assessors calculated a third metric ($nCorrect_{var}$) in which they counted the number of predictions for which the disease class was correct (giving ties full credit) AND for which at least one of the variants submitted in the corresponding column for that disease class matched one of the variants noted by Hopkins.

Assessors also calculated the following metrics for each submission:

1. avgPCorrect – the average probability assigned by the predictor to the correct disease class. This statistic provides an assessment of predictions that is not dependent on whether the submitter's highest probability prediction was correct.
2. $avgPCorrect_{norm}$– the average probability assigned by the predictor to the correct disease class, after normalizing all probabilities predicted in each submission for each patient to sum to 1.0. (Exception: if all probabilities for a patient were zero, they were not normalized).

3. avgRank – the average rank assigned by the predictor to the correct disease class. Ties were assigned the average rank of each set of tied predictions; e.g., if the two highest probability disease classes had equal rank, both were assigned a rank of 1.5; a 3-way tie for 2nd highest probability would be assigned a rank of 3. Note that because there were 14 disease classes, an all-zero prediction would have an avgRank score of 7.5 (i.e., was scored as a 14-way tie).
4. avgError – the average error in predictions, where the error was measured as the absolute difference between the probability assigned each disease class and zero (if not the correct disease class) or one (if the correct disease class). Like avgPcorrect, avgError assesses predictions independent of their rank, but also includes correct negative predictions.

## 2.4 Prediction Methodology

A summary of each group's prediction methods is given below.

**Group 57** (Jones)**:** The Jones-UCL group made use of one-class Support Vector Machine (SVM) classifiers to automatically assign disease classes according to the supplied exome data. In a normal machine learning experiment, sufficient positive and negative cases are needed to define a hypersurface which separates the two classes. Standard SVMs attempt to define this hypersurface such that the chance of misclassifying new cases is minimized. In some applications, however, only positive or negative cases are readily available, but not both. One-class SVMs [83] have been proposed for problems where either negative or positive case data is unavailable. In this situation, the SVM attempts to identify outliers from a distribution modeled on the available single class of data, and it is assumed that the outliers belong to the alternative class. In this CAGI challenge, of course, neither negative nor positive training data was readily available. However, the assumption was made that the 1000 Genomes data set[84] could be used as a proxy for negative case data. This is a reasonable assumption if we assume that the diseases in question are relatively rare. To start with, gene variants relating to each disease class were collated using ClinVar[85]. Feature sets were generated for each disease class by encoding variant 0/0, 0/1 and 1/1 calls as 0, 1 and 2 respectively, and for each disease-specific feature set, a one class $\nu$-SVM (using a RBF kernel) was trained. The single parameter $\nu$, which controls both the number of support vectors and the misclassification cost, was optimized for each

disease class so as to minimize the number of outliers detected in the 1000 Genome training data. Once trained, the SVM was then applied to the test sample data, and the distance to decision boundary was used as a proxy for classification confidence. The most important variant was identified in each case by systematically removing each variant from the feature set and recalculating the confidence scores.

**Group 58** (Tosatto): The analysis started with a manually curated association between the genes of the panel and the 14 clinical phenotypes of interest based on literature review. Sequencing data was annotated with ANNOVAR[86], considering for each variant the corresponding affected gene, frequency estimated from the 1000 Genomes Project[87] and predicted pathogenicity score from SIFT[88] and PolyPhen2[89]. The method to define association between genetic data and phenotypes was based mainly on two phases. For each individual, variations that are less probable to be disease causing were filtered out and a probability to be affected based on the analysis of variants defined. Only coding and splice-site variants which can affect protein function were considered according to the Common Disease-Rare Variant Hypothesis (CDRVH)[90]. Common (Minor Allele Frequency MAF > 5%) and/or synonymous single nucleotide variations (SNVs) were filtered out. Insertion and deletions were excluded as their impact on protein function is difficult to predict compared to SNVs. Only insertions and deletions (indels) affecting the coding part of a gene and predicted to be "damaging" or known to be pathogenic were considered. Heterozygous indels in genes with autosomal recessive inheritance, occurring in GC-rich or repeated regions were filtered out from the disease candidate mutation pool. An empirically derived scoring scheme was implemented to define association between patients and phenotypes, considering both disease inheritance and predicted SNV pathogenicity. Different weights were assigned to different mutation types, i.e. a high score for known variants associated with a specific disease (mainly by literature review) and a lower score for mutations not affecting protein function according to predictor output (i.e. tolerated, benign and unknown). For autosomal dominant (AD) pathologies, only heterozygous variants plus few manually curated homozygous mutations were considered (i.e. the one with the highest probability score). The disease cutoffs were set at different values between submissions, allowing the stringency of the analysis to vary. Both homozygous and compound heterozygous variants were considered for autosomal recessive (AR) conditions. When more than one match per patient occurred, only the most likely was considered (e.g. the one with higher probability score). Different

submissions correspond to different sets of weights. In particular, for submission 58.1 a slightly lower weight was assigned to variants whose effect is more difficult to assess (i.e. compound heterozygous, homozygous variants with uncertain significance, variants affecting different genes coding for subunits of the same complex) with respect to submission 58.1.

**Group 59** (Qiagen Bioinformatics): All 106 samples were uploaded to Ingenuity Variant Analysis (QIAGEN- Hereditary Disease Solution) and set up an analysis with all samples to filter low quality (call quality < 20) and common variants (>0.5% MAF in 1000 Genomes[84], NHLBI-EVS (http://evs.gs.washington.edu/EVS/), ExAC[91], and Allele Frequency Community (www.allelefrequencycommunity.org), using the Confidence and Common Variants filters, respectively. The Allele Frequency Community is a QIAGEN hosted allele frequency database, founded by QIAGEN and participating members in 2014. It is a freely accessible "opt-in" community resource designed to facilitate sharing of anonymized, pooled allele frequency statistics among community members. The Predicted Deleterious filter was used to keep only those variants that are previously published and classified Pathogenic or Likely Pathogenic, using ACMG guidelines, DM variants (pathological mutations reported to be disease causing in the original literature report) present in HGMD, along with other loss of function (frameshift, start/stop loss or gain, splice site) and missense variants. Finally, the biological context filter was applied to find variants linked to each one of the 14 categories and patient disease category was predicted based variant-disease connection, using path-to-phenotype evidence.

**Group 60** (RSS): Gene phenotype associations were mined from the Hopkins diagnostic panels, OMIM[92], and GeneReviews[93]. Inheritance mode and penetrance information were extracted from online resources for each gene-phenotype pair. Variants with low quality or high population allele frequencies were filtered out and the functional impact was annotated with Variant Effect Predictor[94]. To estimate the probability that a variant is damaging to protein function, we integrated multiple prediction methods to score all types of variants, e.g. missense, nonsense, indels and intronic variants. The damaging scores were scaled and normalized to reflect the relative deleteriousness, e.g. frame-shift / nonsense variants would have higher scores than missense variants. We then used the damaging scores to estimate the probability that each individual has a particular phenotype with a probabilistic model, i.e. calculated as the probability that at least one associated gene in the individual causes the phenotype. For a particular gene, the

probability the gene causes the phenotype was calculated as the probability that the gene is disrupted (taking into account inheritance mode) multiplied by its penetrance score. The confidence level of the prediction was calculated from the distribution of the estimated probabilities across phenotypes and across individuals. Considering the 14 phenotypes are Mendelian like diseases, if one individual has high prediction scores across phenotypes, it is more likely to be false positive. Thus high confidence was assigned to individuals with high variability across phenotypes.

**Group 61** (Moult): The method (implemented in Python) has four modules – Variant annotation, QC (quality check), Variant Prioritization, and Probability scoring for the disease. The modules were executed sequentially. Inputs were the two gVCF files and a gene configuration file containing the genes associated with each disease class and their inheritance pattern. The Varant tool (http://compbio.berkeley.edu/proj/varant) was used to annotate variants with: region of occurrence in the genome, allele frequency from ExAC[91], predicted pathogenicity based on four methods[89,95–97] (for missense), and previously reported disease associations in database[85,98]. Three QC analyses were run: (1) Variant counts (common vs. rare vs. novel & homozygous vs. heterozygous) per sample, (2) Read depth for each gene in each sample was obtained by averaging DP values over all bases in a gene recorded in the gVCF file, and (3) Exons with relatively low or no coverage compared to other exons in a gene. The QC qualified variants per sample were prioritized by first assigning them to one of three classes, ranked by the likelihood that the variant is causative and further grouping the variants in each class by frequency based on its ExAC MAF (group 1 – novel, 2 - very rare (MAF <= 0.005), or 3 – rare (MAF <= 0.01).. Class-1 identified variants previously reported in disease databases as pathogenic, Class-2 identified loss of function, splice and missense variants predicted damaging by in-silico prediction tools, and Class-3 identified missense variants (not predicted damaging), UTR, and intronic variants. Variants were further filtered for inheritance model. For each sample, once putative causative variants were found, the process was terminated (e.g. if a suitable variant or variants were found using Class-1, Class-2 and Class-3 were not executed). Finally, a probability score for a sample to have a particular disease was computed based on the type of prioritized variant(s) and inheritance pattern. For the missense variants, the probability model was based on the extent of consensus among the four prediction methods, using a previous HGMD derived calibration. For other variant types, subjective probability rules were used.

# 3 Results

## 3.1 Summary of submissions

Five groups submitted predictions (with 4, 2, 2, 2, and 1 distinct predictions per group). An overview of the challenge and results is shown in Figure 1. The 106 patients in the challenge can be roughly grouped into two difficulty classes: 1) patients for whom Hopkins noted a potentially causal variant in the answer key (43 patients) and 2) patients for whom Hopkins did not note any variants (63 patients) (Figure 9A). At least one CAGI 4 predicting group correctly predicted the disease class for 36 of the 43 patients who had a reported variant (Figure 9B). Fewer groups correctly predicted both the disease class and at least one of the variant(s) that Hopkins reported (Figure 9C). CAGI 4 predictors were not as accurate at predicting disease classes for the remaining 63 patients for whom Hopkins did not note a variant, although at least one group correctly predicted the disease class for the majority of these patients (Figure 9D). The lower prediction accuracy is perhaps unsurprising given the negative test results for these 63 patients.

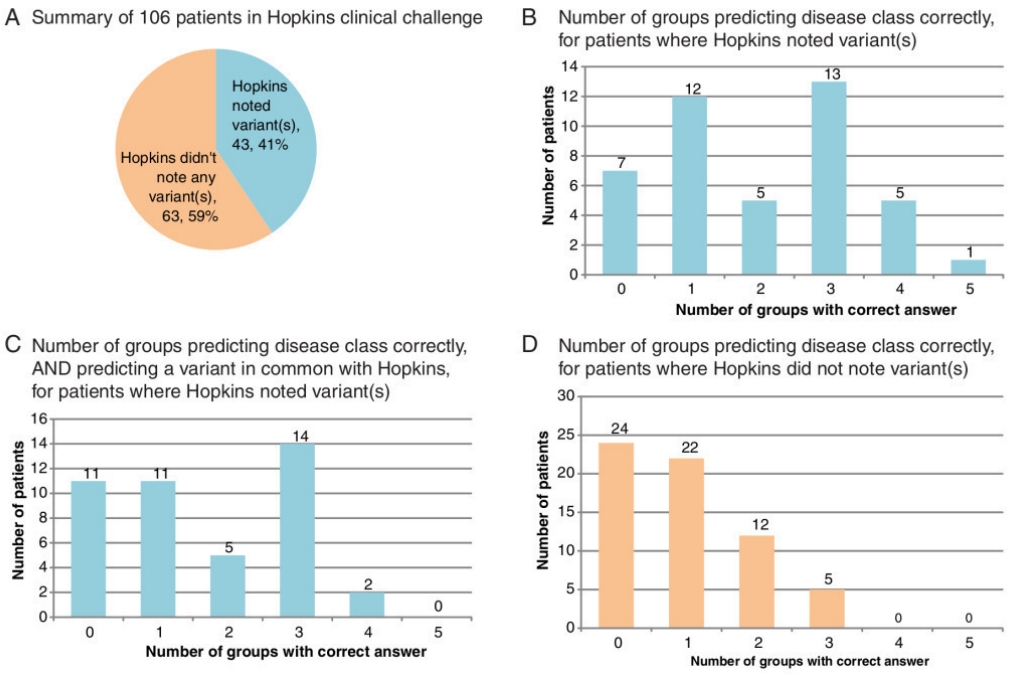


**Figure 9. Summary of CAGI 4 Hopkins clinical panel challenge and results.** **A**: One-hundred six patients were included in the study. Hopkins noted at least one variant relevant to the disease class for which the patient was referred in 43 cases, and

did not note a variant for the remaining 63 cases. Hopkins noted variants of the following classes: variant of uncertain significance, likely pathogenic, or pathogenic. Clinically, Hopkins would have reported 25/43 as positive and 18/43 as uncertain. **B**: Among the 43 patients for whom Hopkins had noted a variant, at least one CAGI 4 prediction group predicted the correct disease class in 36 cases, and one patient's disease class was predicted correctly by all five groups. **C**: Among the 43 patients for whom Hopkins had noted a variant, at least one CAGI 4 prediction group predicted both the correct disease class and a causal variant noted by Hopkins in 32 cases. **D**: Sixty-three patients for whom Hopkins did not note a variant were more difficult for CAGI 4 groups to predict: 24were not predicted correctly by any group, and only five patients' disease class was predicted correctly by three groups (none were predicted correctly by four or more groups).

## 3.2 Numeric assessment summary

Table 11 summarizes our numeric assessment metrics for each non-redundant, submitted prediction, for all patients. Table 12 shows the same statistics for only the 43 patients for which Hopkins noted at least one potentially causal variant. The best values for each metric in each table are indicated in bold. Each group's overall performance is briefly discussed below.

Table 13 shows a summary of the performance of all predicting groups on each patient. Tables 5 and 6 summarize the most frequent combinations of groups that predicted the correct disease class for patients (Table 14 ignores causal variant predictions, while Table 15 requires each group to predict one of the variants noted by Hopkins).

**Group 57** (Jones): Group 57's primary submission (57.1) scored much higher than their other submissions by our metrics. Their method was less accurate than other groups in cases where Hopkins reported a potential causal variant, but it was more accurate at predicting the correct disease class in cases where Hopkins didn't report a variant. Group 57's primary submission was also the most accurate among all submissions at rank-ordering the disease classes. As seen in Table 14, Group 57 predicted disease classes correctly for 18 patients that no other group predicted correctly, with seven of these cases in their primary submission. This method was unique in that it did not attempt to mimic a traditional clinical genetics approach. No attempt was made to independently predict the pathogenicity of the ClinVar variants used as features or to correct for linkage disequilibrium, which may explain why the method was able to make correct inferences where no causal variants were reported and why correct inference can arise without

reporting the correct variants. A possibility is that some or even a majority of the variants relied on by the classifiers were non-causal variants which simply happen to be in linkage disequilibrium with one or more true causal variants. Thus the occurrence of these variants were sufficient to identify the sample as a genetic outlier, though not indicating true causation. It is possible that by addressing these issues, the method might be further enhanced to make more accurate predictions relating to true causal variants. It would be interesting to test this method on a larger dataset to rule out the possibility that there is some underlying structure in this dataset that the algorithm is detecting.

**Group 58** (Tosatto): As seen in Table 14, most cases that Group 58 predicted correctly were also predicted by at least one other group. However, Group 58 predicted the disease class for one patient (P81) that no other groups predicted; they also assigned 100% probability of the correct disease to that patient, and predicted exactly the same causal variants as noted by Hopkins. Many of the diseases in this challenge result from loss of function variants in a given gene, thus by excluding frameshift variants (out of frame deletions and/or insertions within an exon) Group 58 missed these cases. The genes and molecular mechanisms associated with each of the 14 disease classes were not provided as part of the dataset, which increased the difficulty of the matching exercise.

**Group 59** (Qiagen): Group 59 had the highest average P values for the correct disease classes, after normalization; they also had some of the best scores in the avgError metric. Group 59 correctly predicted the disease class for five patients that no other groups predicted. Among all the groups, they were the only group for which both P values and SD values were independent and positively correlated with the values they were expected to correlate with (see discussion of P and SD, below). This challenge was well-suited for the Qiagen group, as they specialize in large scale variant interpretation[99].

**Group 60** (RSS: Due to the misleading fields in the combined VCF files (see the Methods section on sequencing and variant calling), Group 60 made only 11 high-confidence (P > 0.6) predictions, of which 9 were correct. Interestingly, four of these nine cases were not predicted correctly by any other group. Because of the small number of high-confidence predictions, Group 60 had the lowest avgError score among all groups, and the best correlation between assigned P values and correct answers (see discussion of P and SD, below). After the challenge closed, Group 60 provided the CAGI organizers with a corrected submission, in which the misleading VCF fields were ignored. In this corrected submission (which arrived late and therefore was not formally assessed), Group 60

correctly predicted 38 disease classes. Group 60 adeptly used a series of online clinical genetics resources in their analysis pipeline.

| Group | Prediction | nCorrect | $nCorrect_{tie}$ | avgPCorrect | $avgPCorrect_{norm}$ | avgRank | avgError |
|---|---|---|---|---|---|---|---|
| **Jones** | 57.1 | 24 | 24 | 0.305 | 0.098 | 5.32 | 0.251 |
| | 57.2 | 9 | 9 | 0.239 | 0.068 | 7.66 | 0.287 |
| | 57.3 | 7 | 7 | 0.236 | 0.068 | 7.78 | 0.289 |
| | 57.4 | 7 | 6.5 | 0.426 | 0.074 | 7.1 | 0.42 |
| **Tosatto** | 58.1 | 23 | 23 | 0.178 | 0.217 | 6.48 | 0.105 |
| | 58.4 | 26 | 25 | 0.223 | 0.227 | 6.15 | 0.107 |
| **Qiagen** | 59.1 | 32 | 29.5 | 0.302 | 0.278 | 5.82 | 0.09 |
| | 59.2 | 31 | 28.5 | 0.292 | 0.269 | 5.88 | 0.091 |
| **RSS** | 60.1 | 12 | 12 | 0.072 | 0.102 | 7.14 | 0.08 |
| | 60.2 | 12 | 12 | 0.068 | 0.094 | 7.15 | 0.082 |
| **Moult** | 61.1 | 38 | 34.99 | 0.261 | 0.265 | 5.65 | 0.105 |

*Note:* Predictions are numbered according to the group's (formerly anonymized) group number (57, Jones; 58, Tosatto; 59, Qiagen Bioinformatics; 60, RSS; 61, Moult).

**Table 11. Performance assessment, all patients.** Summary of assessment metrics for each nonredundant, submitted prediction, for all patients

| Group | Prediction | nCorrect | $nCorrect_{tie}$ | $nCorrect_{var}$ | avgPCorrect | $avgPCorrect_{norm}$ | avgRank | avgError |
|---|---|---|---|---|---|---|---|---|
| **Jones** | 57.1 | 5 | 5 | 2 | 0.255 | 0.082 | 6.53 | 0.257 |
| | 57.2 | 5 | 5 | 2 | 0.325 | 0.091 | 6.29 | 0.274 |
| | 57.3 | 2 | 2 | 0 | 0.22 | 0.063 | 8.49 | 0.296 |
| | 57.4 | 1 | 1 | 0 | 0.394 | 0.07 | 7.5 | 0.421 |
| **Tosatto** | 58.1 | 15 | 15 | 13 | 0.32 | 0.349 | 5.56 | 0.087 |
| | 58.4 | 17 | 16 | 16 | 0.38 | 0.339 | 5.16 | 0.094 |
| **Qiagen** | 59.1 | 23 | 21 | 19 | 0.535 | 0.488 | 4.24 | 0.065 |
| | 59.2 | 22 | 20 | 19 | 0.512 | 0.465 | 4.4 | 0.066 |
| **RSS** | 60.1 | 9 | 9 | 8 | 0.149 | 0.193 | 6.41 | 0.073 |
| | 60.2 | 9 | 9 | 8 | 0.145 | 0.181 | 6.4 | 0.075 |
| **Moult** | 61.1 | 26 | 26 | 25 | 0.5 | 0.512 | 3.78 | 0.07 |

**Table 12. Performance assessment, 43 patients dataset.** Summary of assessment metrics for each nonredundant, submitted prediction, for the 43 patients for which Hopkins noted at least one potentially causal variant.

**Group 60** (RSS: Due to the misleading fields in the combined VCF files (see the Methods section on sequencing and variant calling), Group 60 made only 11 high-confidence (P > 0.6) predictions, of which 9 were correct. Interestingly, four of these nine cases were not predicted correctly by any other group. Because of the small number of high-confidence predictions, Group 60 had the lowest avgError score among all groups, and the best correlation between assigned P values and correct answers (see discussion of P and SD, below). After the challenge closed, Group 60 provided the CAGI organizers with a corrected submission, in which the misleading VCF fields were ignored. In this corrected submission (which arrived late and therefore was not formally assessed), Group 60 correctly predicted 38 disease classes. Group 60 adeptly used a series of online clinical genetics resources in their analysis pipeline.

**Group 61** (Moult): Group 61 made more correct predictions of both disease class and Hopkins-annotated variants than any other group. For the 43 cases where Hopkins noted variants, Group 61 did especially well, getting 26 disease classes correct, and predicting the best average rank for the correct disease. In 25 of these cases, Group 61 also predicted at least one causal variant that was noted by Hopkins. Group 61 correctly predicted the disease class for six patients that no other groups predicted correctly, and also predicted at least one of the potentially causal variants noted by Hopkins in four of these six cases.

## 3.3 Accuracy of P and SD values

We expected that predictors' submitted probabilities for each patient and disease should correlate with the correct disease class for each patient, and we also expected that their submitted standard deviations on each prediction should correlate with the error in each prediction (i.e., the absolute difference between the P value and either 1 or 0, for cases where the patient does or does not have the disease, respectively). Overall, predictors did better in the first case, and not as well in the second. Only one group (59; Qiagen) had an independent SD model that correlated positively with error.

| Patient | nC | nCV | Correct groups | Correct groups, with variant | Correct predictions | Correct predictions, with variant |
|---|---|---|---|---|---|---|
| P1 | 4 | 4 | 57, 59, 60, 61 | 57, 59, 60, 61 | 59.2, 60.1, 60.2, 61.1, 57.1, 57.3, 59.1 | 59.2, 60.1, 60.2, 61.1, 57.1,, 59.1 |
| P2 | 1 | N/A | 57 | N/A | 57.2 | N/A |
| P3 | 0 | N/A | None | N/A | None | N/A |
| P4 | 5 | 3 | 57, 58, 59, 60, 61 | 58, 59, 61 | 59.2, 58.4, 60.1, 60.2, 61.1, 57.1, 58.1, 59.1 | 59.2, 58.4, 61.1, 58.1, 59.1 |
| P5 | 2 | 2 | 60, 61 | 60, 61 | 60.1, 60.2, 61.1 | 60.1, 60.2, 61.1 |
| P6 | 3 | N/A | 57, 59, 61 | N/A | 59.2, 61.1, 57.1, 59.1 | N/A |
| P7 | 0 | N/A | None | N/A | None | N/A |
| P8 | 1 | 1 | 60 | 60 | 60.1, 60.2 | 60.1, 60.2 |
| P9 | 1 | 0 | 57 | None | 57.4, 57.1 | None |
| P10 | 2 | N/A | 57, 58 | N/A | 58.4, 57.1, 58.1 | N/A |
| P11 | 1 | 1 | 61 | 61 | 61.1 | 61.1 |
| P12 | 0 | N/A | None | N/A | None | N/A |
| P13 | 3 | N/A | 57, 58, 60 | N/A | 57.4, 58.4, 60.1, 60.2, 57.2, 58.1 | N/A |
| P14 | 0 | N/A | None | N/A | None | N/A |
| P15 | 0 | N/A | None | N/A | None | N/A |
| P16 | 2 | N/A | 57, 58 | N/A | 58.4, 57.1, 58.1 | N/A |
| P17 | 3 | 3 | 58, 59, 61 | 58, 59, 61 | 59.2, 58.4, 61.1, 58.1, 59.1 | 59.2, 58.4, 61.1, 58.1, 59.1 |
| P18 | 2 | N/A | 57, 58 | N/A | 58.4, 57.1, 58.1 | N/A |
| P19 | 1 | 1 | 60 | 60 | 60.1, 60.2 | 60.1, 60.2 |
| P20 | 1 | N/A | 57 | N/A | 57.1 | N/A |
| P21 | 1 | N/A | 57 | N/A | 57.1 | N/A |
| P22 | 1 | N/A | 59 | N/A | 59.2, 59.1 | N/A |
| P23 | 0 | 0 | None | None | None | None |
| P24 | 4 | 3 | 57, 58, 59, 61 | 58, 59, 61 | 59.2, 58.4, 61.1, 57.2, 58.1, 59.1 | 59.2, 58.4, 61.1, 58.1, 59.1 |
| P25 | 1 | 0 | 57 | None | 57.2 | None |
| P26 | 3 | 3 | 58, 59, 61 | 58, 59, 61 | 59.2, 58.4, 61.1, 59.1 | 59.2, 58.4, 61.1, 59.1 |
| P27 | 3 | 1 | 58, 59, 61 | 59 | 59.2, 58.4, 61.1, 58.1, 59.1 | 59.2, 59.1 |
| P28 | 2 | N/A | 57, 59 | N/A | 59.2, 57.1, 59.1 | N/A |
| P29 | 1 | N/A | 57 | N/A | 57.1 | N/A |
| P30 | 3 | 2 | 58, 59, 61 | 58, 61 | 58.4, 61.1, 58.1, 59.1 | 58.4, 61.1, 58.1 |
| P31 | 1 | N/A | 57 | N/A | 57.1 | N/A |
| P32 | 4 | 3 | 58, 59, 60, 61 | 58, 60, 61 | 59.2, 58.4, 60.1, 60.2, 61.1, 58.1, 59.1 | 58.4, 60.1, 60.2, 61.1, 58.1 |
| P33 | 0 | N/A | | N/A | None | N/A |
| P34 | 3 | 3 | 58, 59, 61 | 58, 59, 61 | 59.2, 58.4, 61.1, 58.1, 59.1 | 59.2, 58.4, 61.1, 58.1, 59.1 |
| P35 | 0 | N/A | None | N/A | None | N/A |
| P36 | 0 | 0 | None | None | None | None |

| | | | | | | |
|---|---|---|---|---|---|---|
| P37 | 0 | 0 | None | N/A | None | N/A |
| P38 | 3 | 3 | 58, 60, 61 | 58, 60, 61 | 58.4, 60.1, 60.2, 61.1, 58.1 | 58.4, 60.1, 60.2, 61.1, 58.1 |
| P39 | 1 | 0 | 59 | None | 59.2, 59.1 | None |
| P40 | 0 | N/A | None | N/A | None | N/A |
| P41 | 0 | N/A | None | N/A | None | N/A |
| P42 | 2 | 1 | 59, 61 | 61 | 59.2, 61.1, 59.1 | 61.1 |
| P43 | 2 | N/A | 57, 61 | N/A | 61.1, 57.1 | N/A |
| P44 | 1 | N/A | 59 | N/A | 59.2, 59.1 | N/A |
| P45 | 1 | N/A | 57 | N/A | 57.1 | N/A |
| P46 | 0 | N/A | None | N/A | None | N/A |
| P47 | 1 | 1 | 61 | 61 | 61.1 | 61.1 |
| P48 | 0 | 0 | None | None | None | None |
| P49 | 1 | N/A | 57 | N/A | 57.1 | N/A |
| P50 | 0 | N/A | None | N/A | None | N/A |
| P51 | 1 | N/A | 61 | N/A | 61.1 | N/A |
| P52 | 3 | 3 | 58, 59, 61 | 58, 59, 61 | 59.2, 58.4, 61.1, 59.1 | 59.2, 58.4, 61.1, 59.1 |
| P53 | 2 | N/A | 58, 59 | N/A | 59.2, 58.4, 58.1, 59.1 | N/A |
| P54 | 1 | N/A | 57 | N/A | 57.3 | N/A |
| P55 | 0 | 0 | None | None | None | None |
| P56 | 2 | 1 | 58, 59 | 59 | 59.2, 58.1, 59.1 | 59.2, 59.1 |
| P57 | 1 | 1 | 61 | 61 | 61.1 | 61.1 |
| P58 | 0 | N/A | None | N/A | None | N/A |
| P59 | 0 | 0 | None | None | None | None |
| P60 | 2 | 2 | 59, 61 | 59, 61 | 59.2, 61.1, 59.1 | 59.2, 61.1, 59.1 |
| P61 | 1 | N/A | 57 | N/A | 57.4, 57.3 | N/A |
| P62 | 2 | N/A | 57, 60 | N/A | 60.1, 60.2, 57.1 | N/A |
| P63 | 3 | N/A | 57, 59, 61 | N/A | 59.2, 61.1, 57.1, 59.1 | N/A |
| P64 | 1 | 1 | 60 | 60 | 60.1, 60.2 | 60.1, 60.2 |
| P65 | 3 | N/A | 58, 60, 61 | N/A | 58.4, 60.1, 60.2, 61.1 | N/A |
| P66 | 0 | N/A | None | N/A | None | N/A |
| P67 | 0 | 0 | None | None | None | None |
| P68 | 1 | N/A | 59 | N/A | 59.2, 59.1 | N/A |
| P69 | 0 | 0 | None | None | None | None |
| P70 | 0 | N/A | None | N/A | None | N/A |
| P71 | 0 | N/A | None | N/A | None | N/A |
| P72 | 3 | 2 | 57, 59, 61 | 59, 61 | 59.2, 61.1, 57.2, 59.1 | 59.2, 61.1, 59.1 |
| P73 | 4 | 3 | 57, 58, 59, 61 | 58, 59, 61 | 59.2, 58.4, 61.1, 57.1, 58.1, 59.1 | 59.2, 58.4, 61.1, 58.1, 59.1 |
| P74 | 0 | N/A | None | N/A | None | N/A |
| P75 | 0 | N/A | None | N/A | None | N/A |
| P76 | 0 | N/A | None | N/A | None | N/A |
| P77 | 0 | N/A | None | N/A | None | N/A |

| | | | | | | |
|---|---|---|---|---|---|---|
| P78 | 1 | N/A | 61 | N/A | 61.1 | N/A |
| P79 | 0 | N/A | None | N/A | None | N/A |
| P80 | 3 | 3 | 57, 59, 61 | 57, 59, 61 | 59.2, 61.1, 57.1, 57.2, 59.1 | 59.2, 61.1, 57.1, 57.2, 59.1 |
| P81 | 1 | 1 | 58 | 58 | 58.4, 58.1 | 58.4, 58.1 |
| P82 | 1 | N/A | 57 | N/A | 57.2 | N/A |
| P83 | 1 | N/A | 57 | N/A | 57.4, 57.3 | N/A |
| P84 | 4 | 4 | 57, 58, 59,61 | 57, 58, 59, 61 | 59.2, 58.4, 61.1, 57.2, 58.1, 59.1 | 59.2, 58.4, 61.1, 57.2, 58.1,59.1 |
| P85 | 2 | N/A | 57, 61 | N/A | 57.4, 61.1 | N/A |
| P86 | 2 | N/A | 58, 59 | N/A | 59.2, 58.4, 58.1, 59.1 | N/A |
| P87 | 3 | N/A | 57, 58, 61 | N/A | 57.4, 58.4, 61.1, 57.1, 58.1, 57.3 | N/A |
| P88 | 1 | N/A | 57 | N/A | 57.1 | N/A |
| P89 | 1 | N/A | 57 | N/A | 57.4, 57.1 | N/A |
| P90 | 2 | N/A | 58, 61 | N/A | 58.4, 61.1, 58.1 | N/A |
| P91 | 1 | N/A | 57 | N/A | 57.2 | N/A |
| P92 | 3 | 3 | 58, 59, 61 | 58, 59, 61 | 59.2, 58.4, 61.1, 59.1 | 59.2, 58.4, 61.1, 59.1 |
| P93 | 1 | 1 | 60 | 60 | 60.1, 60.2 | 60.1, 60.2 |
| P94 | 2 | 2 | 59, 61 | 59, 61 | 59.2, 61.1, 59.1 | 59.2, 61.1, 59.1 |
| P95 | 3 | 3 | 58, 59, 61 | 58, 59, 61 | 59.2, 58.4, 61.1, 58.1, 59.1 | 59.2, 58.4, 61.1, 58.1, 59.1 |
| P96 | 1 | 0 | 57 | None | 57.3 | None |
| P97 | 0 | N/A | None | N/A | None | N/A |
| P98 | 2 | N/A | 57, 61 | N/A | 61.1, 57.1 | N/A |
| P99 | 1 | N/A | 59 | N/A | 59.2, 59.1 | N/A |
| P100 | 0 | N/A | None | N/A | None | N/A |
| P101 | 2 | N/A | 57, 61 | N/A | 61.1, 57.1 | N/A |
| P102 | 1 | N/A | 57 | N/A | 57.3 | N/A |
| P103 | 0 | N/A | None | N/A | None | N/A |
| P104 | 3 | 3 | 58, 59, 61 | 58, 59, 61 | 59.2, 58.4, 61.1, 58.1, 59.1 | 59.2, 58.4, 61.1, 58.1, 59.1 |
| P105 | 3 | 3 | 58, 59, 61 | 58, 59, 61 | 59.2, 58.4, 61.1, 58.1, 59.1 | 59.2, 58.4, 61.1, 58.1, 59.1 |
| P106 | 1 | N/A | 61 | N/A | 61.1 | N/A |

*Note:* **nC**, number of groups predicting the disease class correctly, among all submissions from each group (counting ties, except in cases where all 14 disease classes were assigned equal probability; **nCV**, number of groups predicting both the correct disease class and at least one variant noted by Hopkins; correct groups, a list of groups in which the disease class was predicted correctly in at least one submission (counting ties, except in cases where all 14 disease classes were assigned equal probability); **correct groups**, with variant, a list of groups with at least one prediction of the correct disease class, and also at least one variant noted by Hopkins (N/A in this field indicates that Hopkins did not note any variants); **correct predictions, with variant**, same as above, but indicating individual submission numbers that were correct.

**Table 13. Performance assessment for each patient.** Summary of the performance of all predicting groups on each patient.

## 3.4 Commentary on novel variant predictions

One large limitation in the design of this challenge is that only a subset of the sequence data were clinically analyzed in each patient. This allowed for the possibility of false negatives, where true pathogenic variants may have been present in genes that were not analyzed by the lab. Further, Internal Review Board (IRB) restrictions prevented the data provider from acting as an assessor for the challenge or providing detailed feedback on variant predictions in genes that were not clinically analyzed. In addition, specific variants cannot be listed in the following discussion. In the future, advanced planning is needed to ensure that the appropriate consents and approvals are in place to maximize the use of clinical data. Ideally, a dataset should be fully analyzed by a clinical lab and patients should be specifically asked for consent that their data be used for research purposes such as the CAGI challenge. This would allow a more critical analysis of the challenge data, would eliminate the possibility of unwanted incidental findings, and would allow more in-depth discussion of challenge results. Clinical data from human patients makes an interesting challenge set, but data from human subjects involve privacy concerns vastly different from that of laboratory model organisms.

The CAGI 4 Hopkins clinical panel challenge gives us an opportunity to test state-of-the-art genetic analysis pipelines on a subset of the data that would be obtained from complete exome sequencing of patients, and to explore potential advantages and disadvantages of genomics-driven approaches to clinical testing versus the phenotype-driven approach currently employed by Hopkins.

In some cases multiple groups reported the same causal variant for a case where Hopkins did not identify a variant. Since Hopkins only analyzed the genes ordered by the physician, it is possible that there were true pathogenic variants identified in the challenge that were not included on the answer key, such cases are elaborated on below. In order to explore the potential complication of false positives in the genomics-driven approach, we also examined cases in which CAGI 4 predictors consistently predicted the wrong disease class along with the same causal variants. Several of these cases are described below.

| Number of patients | Groups predicting correct disease class |
|---|---|
| 31 | No group predicted correct disease |
| 18 | 57 (Note: 7 from 57.1) |
| 10 | 58, 59, 61 |
| 6 | 61 |
| 5 | 59 |
| 4 | 60 |
| 4 | 57, 61 |
| 4 | 57, 59, 61 |
| 3 | 59, 61 |
| 3 | 58, 59 |
| 3 | 57, 58, 59, 61 |
| 3 | 57, 58 |
| 2 | 58, 60, 61 |
| 1 | 60, 61 |
| 1 | 58, 59, 60, 61 |
| 1 | 58 |
| 1 | 57, 60 |
| 1 | 57, 59, 60, 61 |
| 1 | 57, 59 |
| 1 | 57, 58, 61 |
| 1 | 57, 58, 60 |
| 1 | 57, 58, 59, 60, 61 |

**Table 14.** Frequency with which each combination of groups correctly diagnosed patients.

| Number of patients | Groups predicting correct disease & variant |
|---|---|
| 63 | (Hopkins did not note any variants) |
| 11 | 58, 59, 61 |
| 11 | (No group predicted disease and variant correctly) |
| 4 | 61 |
| 4 | 60 |
| 3 | 59, 61 |
| 2 | 59 |
| 2 | 58, 60, 61 |
| 1 | 60, 61 |
| 1 | 58, 61 |
| 1 | 58 |
| 1 | 57, 59, 61 |
| 1 | 57, 59, 60, 61 |
| 1 | 57, 58, 59, 61 |

**Table 15:** Frequency with which each combination of groups correctly diagnosed patients, and also noted a Hopkins variant.

Patient P7 – Groups 57 (submission 4), 58, 59, and 61 all predicted Telomere Shortening Disorders, and the latter 3 groups consistently noted a missense variant in *TERT*. The patient's diagnosis was Cystic Fibrosis and CF-Related disorders, and Hopkins did not note any reportable variants and did not analyze the *TERT* gene. The *TERT* variant is described in the literature; it leads to telomere shortening and is involved in bone marrow failure. Telomere shortening due to mutations in *TERT* is known to be involved in pulmonary fibrosis. Clinical presentation of pulmonary fibrosis is very different from cystic fibrosis. This *TERT* variant is annotated in ClinVar as involved in pulmonary fibrosis, but literature support for this phenotype is unclear. The variant is found in 120 ExAC participants including 2 homozygotes.

Patient P36 – Groups 57 (submission 2), 58, 59, and 61 all predicted Liddle syndrome, with the same missense variant in *SCNN1G*. The patient's diagnosis was Diffuse Lung Disease. The *SCNN1G* variant is a known pathogenic variant observed in two independent patients with bronchiectasis. The predictors presumably predicted Liddle syndrome because the same gene is involved in that disorder. This is likely an example of another false positive prediction common to multiple groups. Hopkins did not note a reportable variant for this patient and the *SCNN1G* gene was not analyzed.

Patient P37 – Groups 57 (submission 2), 58, 59, and 61 all predicted Marfan syndrome with the same variant, a missense variant in *FBN1*. The patient's diagnosis was Diffuse Lung Disease. *FBN1* is involved in Marfan syndrome and in other cardiac phenotypes. A subgroup of Marfan patients develop lung emphysema, which is possibly a reason for the predictions. The missense variant is a known low frequency polymorphism annotated as "benign" in ClinVar, so this is likely a false positive prediction. Hopkins did not note any variants for this patient and did not analyze the *FBN1* gene.

Patient P14 – Groups 57 (submissions 3 and 4), 58, 59, and 61 all predicted Cystic Fibrosis and CF-Related disorders, along with one to two out of four variants in *CFTR*. The patient's diagnosis was Diffuse Lung Disease, and Hopkins did not analyze the *CFTR* gene. All the predicted *CFTR* variants have previously been reported. One is a common polymorphism, and unlikely to contribute to disease. Another is intronic, and it is not clear whether it may be involved in splicing. The remaining two *CFTR* variants were rare missense variants. One missense variant is seen in ExAC 739 times including once in the homozygous state, and there is no information on its pathogenicity reported in the literature or public databases. The second missense variant is seen in ExAC 623 times including once in the homozygous state, and there is conflicting evidence reported in the literature regarding its pathogenicity. The latter two variants appear to be too common to be causal in this case, but as mentioned above, CF studies may be included in ExAC. It would be prudent to study the background frequencies of these two variants in further detail, in order to decide whether they are likely to be causative.

# 4 Conclusions

Overall, we found that current state of the art computational prediction methods do a reasonable job of predicting clinical phenotype from genotype, even when blinded to clinical diagnoses. At the same time, current genotype-driven prediction methodologies generate false positives and false negatives at a rate unacceptable for clinical use. In cases where the Hopkins lab reported a variant, predictors did relatively well, with at least one group correctly identifying the disease class in 36 of 43 patients (84%), and at least one group identifying the correct disease class and variant in 33 of 43 cases (77%). In cases where the Hopkins lab did not find a reportable variant in the genes they analyzed, at least one group correctly matching the disease class in 39 of 63 patients (62%). In the latter cases, methods based on machine learning (SVM) technology appeared to be most effective at correctly identifying the disease. Interestingly, despite the ability to correctly match genotype to phenotype, the SVM-based method could not correctly identify the pathogenic variant. It is unclear what is happening in cases where groups correctly identify the disease class, but not the causal variant. In retrospect, it would have been prudent to include a list of gene-disease associations as well as modes of inheritance to the predictors to aid in the matching process.

Different groups performed better depending on which metric was used; there was no clear "winner" that dominated performance across all metrics. Indeed, every group predicted at least one patient's disease class correctly that no other group predicted correctly. This result suggests that a "meta-predictor" or a human clinical expert with access to all groups' results might improve on the performance of each individual group.

Currently, clinical genetic testing is almost entirely phenotype-driven: given a clinical diagnosis, laboratories analyze variants in genes known to be relevant to the diagnosed disease. This is partially due to the historic technical limitations on genetic testing, e.g., sequencing costs limited the number of genes for which data could be obtained. The standards for reporting variants to the patient are also currently conservative, in part because common, benign polymorphic variants have caused many false positives in past genetic analyses[100,101]. However, as whole-exome and whole-genome sequencing become more economical, the phenotype-driven paradigm may be replaced by a genomics-driven approach, in which all rare, putatively functional variants in a patient's genome are first identified, then evaluated based on the plausibility that they may be pathogenic. The

genomics-driven approach has the potential for higher sensitivity, due to more genes being analyzed, and also has the potential to diagnose diseases not identified by the referring physician. However, the main tradeoff compared to phenotype-driven approaches is a potentially higher false positive rate.

Multiple CAGI 4 groups in the Hopkins challenge were in consensus in identifying several possible causative variants that were not identified by the current panel testing paradigm. They also identified several other variants that were likely to be false positives. Distinguishing these two possibilities, and identifying which variants to report to the patient, is a topic that requires further research. The American College of Medical Genetics and Genomics has published guidelines for the interpretation of sequence variants in order to help codify variant assessment[81]. However, even when adhering to these guidelines there are still elements of variant interpretation that are subjective and vary between labs[102,103]. Given large databases of "control" exomes (i.e., without a known phenotype), researchers could develop statistical models to predict whether particular variants are in fact causative[91]. Such models could inform the development of new statistically justified reporting standards based on, for example, particular thresholds on the probability that the prediction of a causal variant is a false positive.

This challenge was designed to reflect the range of cases seen in the Hopkins diagnostic lab (Figure 9A). This includes a high percentage of cases for which no likely pathogenic variant was identified, despite the patient presenting with a clinical phenotype. Even for clinical exome sequencing, nearly 75% of cases are negative[78,79]. Negative cases proved especially challenging to participants, as 'phenotype not discernable' was not listed as a matching option. Despite the fact that no pathogenic variants were identified by the Hopkins lab, most groups were able to make a disease prediction and to identify putative pathogenic alleles in these negative cases. Indeed, the reason data from all 83 genes was included in the challenge was to highlight the difficulty in interpreting a large data set of rare variants that are unrelated to the patient's phenotype. The presence of negative cases in the data set reflects clinical practice and cautions on the overinterpretation of rare variants. Unlike prior prediction challenges, where the activity of an enzyme had been quantitatively measured in the laboratory, there was no definitive answer key for this challenge. The predictors were asked to match sequencing data to a phenotype, and many groups did so by first identifying a causative variant. Only in a minority of cases (~23% in this dataset) could it be said with high confidence that a variant was likely

contributing to disease in a patient. When a clinical laboratory reports a variant as Pathogenic, this is often because the variant has previously been reported in patients with the same phenotype or the nucleotide change introduces a premature termination codon in a gene where loss-of-function variants cause disease[81]. Thus, with a foundation in clinical genetics and access to online resources one could identify a large proportion of the 'Pathogenic' variants in this dataset. However, many of the variants detected in the clinical laboratory are rare missense or synonymous variants that have not previously been reported in the literature; these are almost always classified as variants of uncertain clinical significance. It is for these variants of uncertain significance, that are difficult to interpret and for which there is no answer key, that better assessment tools are needed.

A CAGI challenge focused on the interpretation of variants of uncertain clinical significance would be more relevant to current clinical genetics practice. A clinical lab may upgrade a variant's classification from 'Uncertain' to 'Pathogenic' based on new clinical information, segregation of a variant within a family, or identification of the variant in multiple unrelated individuals. Many molecular diagnostic labs maintain internal variant databases; such databases could be mined to curate a challenge set of 'Uncertain' variants for which there is unpublished data to support pathogenicity. In this proposed challenge, participants would have to correctly identify these 'Pathogenic' variants from a set of 'Uncertain' variants (for which there was unpublished data that they were NOT likely to contribute to disease). This would more directly test the challengers' ability to predict pathogenicity without relying on allele frequency or online databases and without requiring knowledge of gene-disease associations. Assessment of the challenge would benefit from having fully vetted data and a clear answer key. This type of challenge, while still lacking a phenotype component, would more accurately mirror the clinical challenge of interpreting rare variants. Obtaining this data set would also invite communication between clinical testing labs (both academic and commercial) and the research community.

In this vein, the development of a clinically useful variant assessment tool will require collaboration between clinical geneticists and data scientists. Discussions resulting from the Hopkins Clinical challenge demonstrated that although most participants incorporated genetic principles into their pipelines, they approached variant interpretation in a very different manner than a clinical laboratory. In future challenges, it would be interesting to pair an informatics group with a clinical group as a challenge

team, particularly for whole exome sequencing challenges. Ideally, the back-and-forth between clinical and informatics groups would produce a method that could outperform that of either group alone. Diverse collaborations at CAGI could help bridge the communication gap between fields and pave the way for development of better tools.

# Crohn's disease risk prediction - Best practices and pitfalls with exome data.

This chapter is based on "Giollo, M., Jones, D.T., Carraro, M., Leonardi, E., Ferrari, C., Tosatto, S.C.E., 2017. Crohn disease risk prediction-Best practices and pitfalls with exome data. Hum. Mutat. doi:10.1002/humu.23177".

## 1 Introduction

One of the main applications of next-generation sequencing is related to human health diagnostics. Although there are already solid demonstrations that disease causal variants could be identified[104], only a few studies have tried to build predictive models for disease risk and phenotype prediction[105–107]. The Critical Assessment of Genome Interpretation (CAGI) is the first effort aimed at objectively assessing the state of the art for genome interpretation. As already introduced in this manuscript, in the literature there are a vast number of bioinformatics tools available to perform predictions and risk assessments, mostly based on statistical methods and machine learning[108]. It is possible to build a disease risk estimation tool using the same principles, but the *curse of dimensionality*[109] and limited sample size together represent a huge challenge in CAGI. The former issue is due to the high number of variants that can be observed in each sample, on the order of several thousands. Just a few of them are likely to be important for human health, but in most situations the key variants for a disease are unknown. Ideally, one should first perform *feature selection*[110] in CAGI, with the aim to discard irrelevant variants for disease onset. This step is the main result of Genome-Wide Association Studies (GWAS)[111] and linkage analysis[112], but there is still a huge number of variants that need to be annotated. Tools for pathogenicity prediction like SIFT[113] and PolyPhen2[114] can mitigate the problem just partially. In fact, these tools (1) only work on Single-Nucleotide Polymorphisms (SNPs), (2) have a limited accuracy[115] and (3) predict protein loss-of-function, which is not the same as predicting disease risk. The interaction among different variants[116] and environmental relationships[117] are even harder to assess for a proper disease risk prediction. These problems must be considered in CAGI, but their solutions require a large sample size which is not available. We participated successfully in all CD challenges and in this work we tested all the best methods ever proposed for disease risk

prediction. Here, we report the state-of-the-art in this challenge, and emphasize the key features of the most effective methods. We also highlight some issues related with all datasets and the proper evaluation of algorithm performance.

# 2 Materials and Methods

## 2.1 Datasets

CAGI published three different CD datasets over the last three editions. For each of them, the task was always the same. Prediction of a disease risk indicator based on exome data. As introduced in the previous chapter, genotype sequences were collected from German patients, and part of them lead to the association of PRDM1 and NDP52 variants to CD[118]. It is therefore clear that careful study can extract valuable knowledge from CAGI exomes. From the experimental point of view, Illumina instruments were used for sequencing. In 2011, reads were aligned with respect to the human genome build 18 (hg18), and base calling was obtained by a combination of BWA[119], Picard and SAMtools[120]. The main differences in the 2013 and 2016 editions were the introduction of GATK[121] and the use of hg19. Finally, data was provided to CAGI participants as a VCF formatted file (See Table 16). An interesting peculiarity of CAGI 2013 was the presence of exomes from 28 pedigrees, which included a pair of monozygous discordant twins. The number of cases and controls was declared during the prediction season. In addition, during the last two editions data from the previous challenges could be used for training. This idea will be explored heavily over the next sections.

| Edition | Cases | Controls | Ref. Genome |
|---|---|---|---|
| **CAGI 2011** | 42 | 14 | hg18 |
| **CAGI 2013** | 51 | 15 | hg19 |
| **CAGI 2016** | 64 | 47 | hg19 |

**Table 16. Summary of the Crohn's disease challenge data.** Over time, the number of samples increased significantly, with special attention for controls in the latest edition. All exomes data is provided by Andre Franke and Britt-Sabina Petersen.

## 2.2 Algorithms

This section explains the ten algorithms used to prioritize exome variants and predict disease risk. These were implemented because they proved to be among the most effective methods in CAGI 2011 and 13. The goal was to validate them on the in CAGI 2016 dataset. All implementations were written in R, with the intention of obtaining a fully automatic prediction. At first, coding variants with Minor Allele Frequency (MAF) < 0.04 were selected using information from dbSNP[122] and a collection of BioConductor packages[123–125]. Let $S \in \{0, 1, 2, NA\}^{n \times m}$ be the resulting matrix of exome variants, where *n* is the sample size and *m* represents the observed Single-Nucleotide Variants (SNVs). By construction, $s_{ij}$ represents the number of variants at genomic position *j* for sample $s_i$. In other words, 0 and 2 are equivalent to observing twice the nucleotide with the major and minor allele frequency, respectively. A heterozygous variant is encoded with 1. Finally, NA is used to denote unobserved nucleotides in a sample, e.g. due to technical issues or different experimental setup.

Algorithms developed for CD risk prediction exploit either a *weighting scheme* or *machine learning* (ML). A weighting scheme *w* is used as a linear model. Positive weights correspond to pathogenic mutations, whereas negative coefficients are protective variants. By computing the dot product between a genotype $s_i$ and *w* a disease risk can be computed. Machine Learning instead assumes that one can identify patterns to predict a disease risk from a training set, i.e. CAGI 2011 and 2013 CD data. Based on these two ideas, the following methods were tested.

### *2.1.1 Key variants weighting*

This is the simplest form of weighting, which looks at the presence of a predetermined set of important SNVs to predict disease risk. These variants are given a weight of 1, while all other variants are set to 0. This is the typical model used for Mendelian diseases, when a single SNV is evaluated.

### *2.2.2 Odds Ratio weighting*

GWAS estimated odds ratios for variants related to a disease[126]. This is a risk measure of developing a disease based on direct associations on real data. Given this information, let define the weight $w_g$ as follows:

$$w_g = \begin{cases} \frac{\sum(or_i - 1)}{|OR_g|}, or_i \in OR_g \\ 0, |OR_g| = 0 \end{cases}$$

where $OR_g$ is the set of CD associated variants with known odds ratios in gene $g$. In other words, all variants $s_{ij}$ within the same gene $g$ will share the same average weight $w_g$.

### 2.2.3 Publication weighting

Genes related to CD are reported in the literature. Independent studies also corroborate some associations, providing additional belief in previous findings. As an example, 623 genes were linked CD[127] so far. 67% of them were reported just once, while NOD2 alone appears in 356 publications. Phenopedia[127] is a public database that stores the number of times $c_g$ that a gene $g$ was linked to a disease in a scientific publication. In this case, the weighting scheme $w_g$ is defined as follows:

$$w_g = \begin{cases} \log c_g, c_g > 1 \\ 0, otherwise \end{cases}$$

Once again, variants within the same gene share the same weight.

### 2.2.4 NA weighting

DNA sequencing techniques can measure a wide range of information, e.g. single SNVs, exomes and full genomes. Based on the experimental setup, some DNA regions are accessible or not detectable. On top of this, sequencing errors and computational limitations might lead to the impossibility of observing the nucleotides in some DNA regions. These are called Not Available (NA) variables, and pose big issues in data analysis. In this method, disease risk $r$ for sample $s_i$ is defined as

$$r = |NA(s_i)|$$

where $NA(s_i)$ is the set of not observable variants in $s_i$.

### 2.2.5 Overrepresented weighting

More than 50% of the samples in CD challenges are expected to be cases. This proportion is much larger than the disease prevalence in Europe[128]. Thus, variants that are overrepresented with respect to these in a reference populations (e.g. 1000 genomes[129]) might be the causal CD mutations. In this method, dbSNP[122] was used to obtain the minor allele frequency of variants. The overrepresented ones were identified using a binomial test. Bonferroni correction was

applied to correct for the standard p-value threshold of 0.05. The variants selected in this way were given a weight of 1, and 0 for the others.

### 2.2.6 Bi-clustering

This is a type of unsupervised learning techniques that can divide automatically samples in two groups by looking at their reciprocal similarity. We used both k-means (with k = 2) and Hierarchical Clustering in order to solve this problem[130]. We assumed that the smaller cluster is the one with healthy samples.

### 2.2.7 Transductive clustering

This method is based on the Transductive Learning principle[131]. Similarity among samples was estimated using the cophenetic distance[130]. Small clusters ($n \leq 5$) of highly similar samples were identified. Within such groups, the average disease onset age of CD was estimated using knowledge from past CD editions (where available) and transferred to the CAGI 2016 samples. Healthy samples were assumed to have a disease onset age of 1000, just to allow the computation of a trend within the group.

### 2.2.8 Manual prediction

By using the evidence gained from the previous methods, a manual assessment of each sample was performed. Clearly, this is not an algorithm.

### 2.2.9 Transductive SVM

To construct a suitable feature set for classification, an initial assessment of each variant sequence element was carried out against the CAGI 2013 training data. The Fisher exact test was used to select variants associated with disease onset. Overall, 43 variants had a p-value lower than $5x10^{-6}$ and were used to build a machine learning classifier. CAGI 2013 labels were attached to CAGI 2015 samples using transductive learning[131]. The fully labeled dataset was used to train an SVM classifier with RBF kernel.

### 2.2.10 Logistic Regression

Variant relevance was estimated using log-odds ratios on the CAGI 2013 dataset to build a classifier. Only protective SNVs were selected by looking at negative log-odds scores and logistic regression algorithm was used train a classifier.

### *2.2.11 Ensemble*

The disease risk was estimated by combining the previous techniques with bootstrap[132]. In this case, the disease indicator of a given sample was equal to the proportion of methods that ranked it higher than the 67$^{th}$ percentile of all samples.

## 2.3 Performance Measures

The main task in CAGI was the definition of a method that could estimate a disease risk probability given the exome data. The CAGI assessors used Receiver Operating Characteristic (ROC) curves to identify the best submissions[133]. ROC represents the relationship between True Positive Rate (TPR) and False Positive Rate (FPR):

$$TPR = \frac{TP}{P} \quad FPR = \frac{FP}{N}$$

where TP and FP are the number of True Positives and False Positives, while P and N are the total number of Positive and Negative examples in the test set. The ROC curve integral provides the well-known Area Under the Curve (AUC) metric[133]. Accuracy (ACC) and Pearson correlations (COR) are also used to indicate the association between variables:

$$ACC = \frac{TP + TN}{P + N} \quad COR(x, y) = \frac{covariance(X, Y)}{\sigma_X \sigma_Y}$$

where TN are True Negatives, while σ is the standard deviation of a population.

# 3 Results

Each CD challenge proposed during CAGI has peculiarities that should be addressed properly to maximize performance. In this section a description for all three CD datasets is presented, with considerations about the best method to use in each case.

## 3.1 CAGI 2011

CAGI 2011 proposed a disease risk challenge based on exome data for the first time. It represents an important event in establishing best practice that should be taken into account in similar tasks. Unfortunately, there was a critical issue in this challenge, which hampered a proper evaluation of submissions. Cases and controls in the dataset were collected using different experimental setups, which produced incomparable sequencing results. Cases appeared to be sequenced in a limited amount of exome regions, whereas controls were covered in a much wider range. As a result, one could observe a very large

variation in the number of genetic variants reported depending on the two groups, which influenced most prediction algorithms and the submissions. Interestingly, the *NA weighting* strategy would pick-up this signal, and achieves a 95% classification accuracy. Given this huge bias, it is clear that *bi-clustering* can achieve the same results. An implementation of the *publication weighting* strategy by Yana Bromberg proved to be the most effective CAGI 2011 submission among all participants, leading to a nearly perfect ranking. However, it is very likely that this submission converged implicitly to NA weighting due to the dataset bias.

In this edition, only a manual strategy was implemented by our group, which was based on ANNOVAR[134] variants selection of rare SNVs (MAF < 0.04) for genes related to CD according to PheGenI[135] and String[136]. Samples with fewer variants were assumed to be controls (See Table 17). The five submissions were very similar in terms of reciprocal correlation, mainly because they were based on the same variant set. However, rs76982592 was the one that dominated all submissions. By just looking at its presence, according to the *key variants weighting*, one would achieve 89% accuracy.

| Method | Ranking Details | AUC |
|---|---|---|
| Uniform weight | Count variants in CD genes. | 0.66 |
| SNV co-occurrence | Count twice variants that occurs frequently paired with others, according to association rules. | 0.678 |
| *Bi-clustering* | K-means on the variants. Rank according to the number of variants in CD genes. | 0.666 |
| Ensemble | Average of the previous | 0.626 |
| Manual | A manual evaluation of variants | 0.678 |

**Table 17. Performance on the CAGI 11 dataset.** The five methods rely on 9 variants identified in PTPN11, DSPP, TDG, NCOA3, RBMX, PRKRA, ZFHX3, RUNX2 and CELA1 genes. Methods have a very high correlation, which ranged between 0.76 and 0.96.

## 3.2 CAGI 2013

Given the experience from CAGI 2011, one might try to use the same idea on the CAGI 2013 challenge. However, this was quite a unique dataset, with knowledge about (1) of the number of controls (See Table 16), (2) 28 clear pedigrees, and (3) a pair of discordant twins. The CAGI 2011 data was available for training as well. Given this specific setting, any CAGI 2011 strategy should be refined to achieve better performance. In fact, the plain *publication weighting* strategy was not as effective as in the previous challenge, probably due to the complexity of this structured dataset. Clustering is a key technique to highlight the 28 pedigrees and the twins (See Figure 10). The overall best submissions used bi-clustering and CAGI 2011 as training data. Using this, 94% accuracy could be achieved, with 13 out of 15 controls detected. *Overrepresented weighting* implemented by Rita Casadio's group also proved to be very effective, even though this technique would be strongly biased by the presence of pedigrees.

In this challenge, we implemented the most effective strategy. *Transductive clustering* managed to classify effectively 8 samples (blue samples in Figure 10), as they proved to be very similar to CAGI 2011 controls. The twins had mostly an identical set of variants, except for a variant in the MOC2 gene. This was assumed to be the causal mutation for CD in all submissions. After this initial screening, 9 out of 15 controls were detected. Interestingly, the majority of healthy samples are part of an *outlier* group, marked orange in Figure 10. In the best submission (*Bi-clustering* in Table 18), it was assumed that the blue and orange groups contained all controls. Within each group, disease risk was proportional to the number of variants – the same principle used in CAGI 2011. The bi-clustering submission was somehow related to the same result seen in the CAGI 2011 dataset. Healthy samples were very dissimilar with respect to the CD ones, probably due to a different experimental setup. Transductive clustering confirmed that this group was largely composed of controls, boosting significantly belief in this submission. It is still unclear if the controls were actually reused in the two challenges. However, CAGI 2013 confirmed once again the need for better controls in a proper challenge setup.

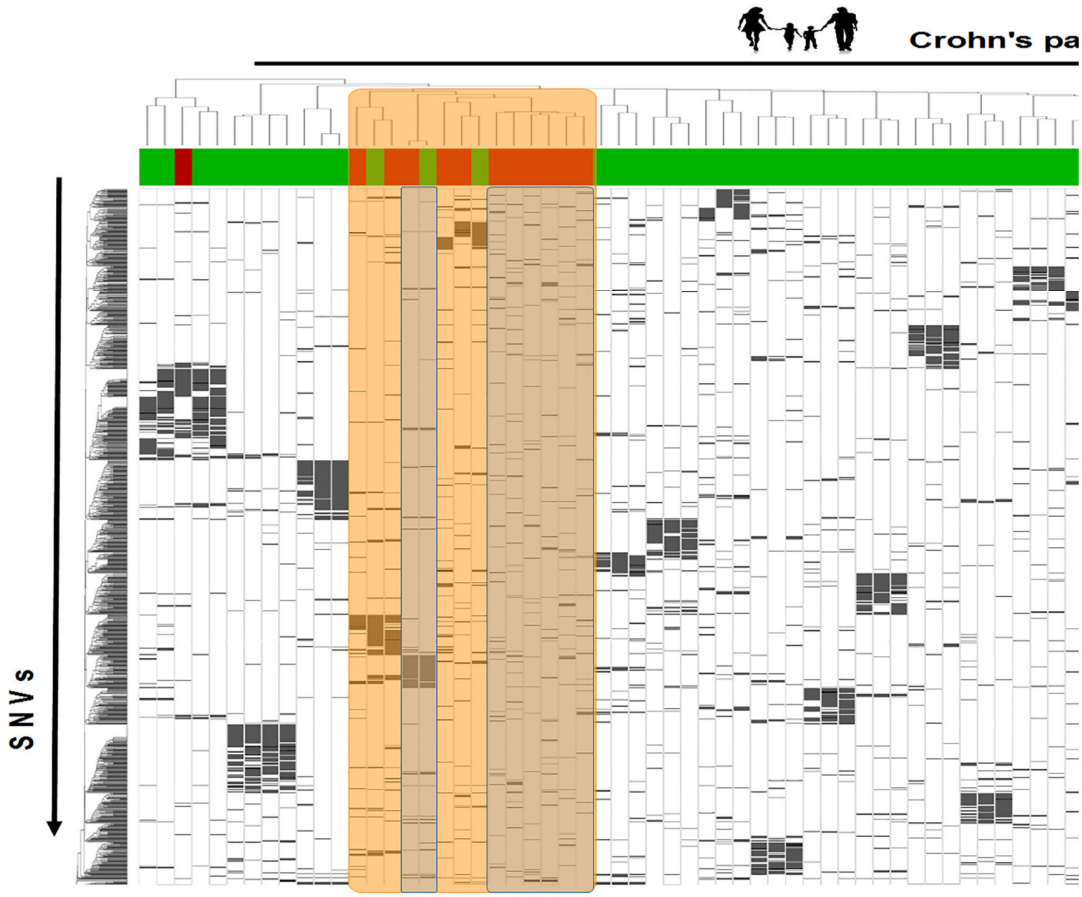


**Figure 10. Heatmap of CAGI 2013 data.** Columns represent the 66 dataset samples (red: controls, green: cases), grouped by genetic similarity using hierarchical clustering. Rows contain mutations clustered by sample similarity. The orange columns are outliers forming a sub-cluster with most of the controls. Prior knowledge is available for samples blue due to control-group membership in CAGI 2011.

| Method | Ranking Details | AUC |
|---|---|---|
| Uniform weight 1 | Count SNVs in CD genes. | 0.743 |
| Uniform weight 2 | Count SNVs in CD genes and their STRINGdb interactors. | 0.736 |
| Mixed pedigree 1 | Max one control per family. Count SNVs in CD genes. | 0.844 |
| Mixed pedigree 2 | Max one control per family. Count pathogenic SNVs (according to SIFT) in CD genes. | 0.688 |
| Mixed pedigree 3 | Max one control per family. Look for families with large difference in SNVs count in CD genes. | 0.798 |
| Bi-clustering | Bi-clustering on the dataset. Rank according to the number of SNVs in CD genes. | 0.866 |

**Table 18. Performance on the CAGI 2013 dataset**. The submissions have a high correlation, ranging between 0.67 and 0.89. This is mainly due to (1) a similar set of variants selected for all methods and (2) the use of transductive clustering from CAGI 2011. The use of clustering is critical to maximize performance

## 3.3 CAGI 2016

This last challenge is probably the hardest ever proposed in CAGI. Samples are apparently quite uniform and there are no obvious issues with experimental settings, like the one described for CAGI 2011 or any prior information available like in CAGI 2013. The dataset size is also much larger compared to previous editions (See Table 16). Transductive clustering was not very effective, since it could match just a single sample as a control. In addition, just two pedigrees could be detected in the entire dataset.

CAGI 2016 is hence composed of independent cases and controls, with basically no relationship to past challenges. This is a good dataset to validate the methods that proved to be most effective in previous challenges. The simplest approach to look for data bias is *NA weighting*. As can be seen in Table 19, it would be probably one of the best methods, with an AUC of 0.7. Methods dealing explicitly with missing data, like variable imputation, are likely to exploit the same source of information. The three methods using published SNVs associated to CD are significantly effective, with an AUC ranging between 0.59 and

0.61. *Key variants weighting* worked well with rs2066844, a SNP known to increase significantly CD risk[137]. On the other hand, *Odds Ratio weighting* proved that published GWAS variants are informative. *Publication weighting* and *Overrepresented weighting* results were not statistically significant, suggesting that ad-hoc weighting strategies are not effective.

Importantly, tested methods rely on very different assumptions. *Key variants weighting* and *Odds Ratio weighting* are the most similar by design, with a correlation of 0.37. Nevertheless, this value is much lower than any method tested in the previous challenges. All other methods have a correlation between -0.1 and 0.26, proving that these strategies explore more divergent hypotheses than in the former CD challenges. Finally, NA weighting is apparently strongly related to bi-clustering on the dataset with a 0.74 correlation between both.

An ensemble prediction was realized with bootstrap, where the disease risk of a given sample was estimated as the amount of methods in agreement for a high risk.

| Method | AUC | Pearson Correlation |
|---|---|---|
| NA weighting | 0.7 * | 0.36 * |
| Publication weighting | 0.56 | 0.05 |
| Key variants weighting (rs2066844) | 0.59 * | 0.23 * |
| Overrepresented weighting | 0.47 | -0.06 |
| Transductive clustering | 0.52 | 0.06 |
| Odds Ratio weighting | 0.59 * | 0.14 |
| Manual prediction | 0.63 * | 0.2 * |
| Transductive SVM | 0.6 * | 0.2 * |
| Logistic regression | 0.57 | 0.16 |
| Ensemble | 0.66 * | 0.29 * |
| Key variants weighting (clinical studies SNVs) | 0.61 * | 0.21 * |

**Table 19. Results of the methods on CAGI 2016 dataset**. Performances marked with a star are statistically significant.

# 4 Conclusions

Present over the last three editions, the Crohn's disease challenge represents well the idea behind CAGI: developing novel tools for genetic interpretation. During these years, the organizers proposed ever larger datasets where methods could be tested more accurately. The prior knowledge provided to participants and the sequencing pipelines varied slightly over time, but it is still obvious that the CD challenge is the only one enabling an analysis over time in this context. There are a few lessons that are clear from the reported results.

In this work we test thoroughly all top performing approaches previously proposed in CAGI, implementing from scratch 10 methods. Many of these were previously proposed by other participants, so we could implement the core ideas based on our understanding. From the results, weighting schemes managed to assign proper importance to exome variants only when numerical coefficients were based on solid studies like GWAS. SNP rs2066844 is a clear example, as it is well known to be associated to CD and it was indeed a powerful discriminative variable in CAGI 2016. No heuristic could address effectively this problem in a similar way. From a perspective of machine learning methods, use of CAGI 2011 data was critically important in CAGI 2013 for the best submissions. During CAGI 16, past training sets were not as useful as in CAGI 2013, probably due to the high dissimilarity of new samples. It is clear that ML is very effective in simple scenarios where samples share high similarity. However, new methods for improved sample comparison are needed to boost the performance of ML algorithms. Overall, both weighting schemes and machine learning performance are limited in the same way, due to the curse of dimensionality. To deal with this issue it is critical to use training sets with a large number of samples. This would help with the selection of key variants and proper estimation of their effect on disease onset. This idea was already explored using data collected during the International IBD Genetics Consortium's Immunochip project[138], where a simple regularized logistic model achieved an AUC of 0.86 on a very large CD test set. Interestingly, models based on Support Vector Machines and Gradient Boosted Trees decreased the predictive performance, suggesting that more complex algorithms are overfitting the data. Most of CAGI methods do not use any training set at all, mainly due to the efforts needed to request such controlled datasets to Data Access Committees. As a result, a significant number of CAGI submissions were no better than random. We

believe that the main improvements in CD prediction will be enabled by just using regularized additive data-driven models We therefore hope to see a simplified access to large datasets[138] for all CAGI participants in the future.

The huge number of variables in all datasets is the main motivation for the use variant selection tools like ANNOVAR. This tool was heavily used in CAGI 2011 as a black box, where its filtering process lead to the variants reported in Table 17. Variant rs76982592 was the one with the highest impact given our weighting strategy. Just at the time of the first CAGI conference, it became clear that the variant was strongly associated with the different experimental setup. This is surely an important contribution of CAGI highlighting a bi-clustering pattern in the data. ANNOVAR was also used in CAGI 2013 for the annotation step. In that edition the tool was used carefully, as its blind usage might filter away important disease variants. Bi-clustering was once again an effective method. ANNOVAR filtering was finally replaced completely by a collection of R packages in this work. Having an in-house tool for variant selection helped to move from manual predictions, like in CAGI 2011, to a fully automated pipeline where human decisions are limited. This is a key step for reproducibility, which is achievable with a full control and understanding of the tools at hand. In this implementation, it was much simpler to identify the relationship between NAs and bi-clustering, ANNOVAR did not emphasize at all the unobserved variables. BioConductor packages allowed controlling in detail the entire filtering step, and learning more about the samples.

Over these three editions, best method performances kept decreasing from an AUC of ~0.9 to an AUC of ~0.7. This negative trend is probably unexpected, but we believe that this is the result of an improper performance evaluation, especially in the first CAGI editions. We believe that all submission results reported so far are not truly representative of our current ability to predict disease risk, but they are inflated due to the strong bias in the datasets. Bi-clustering appeared to be an effective method for predicting disease risk in CAGI datasets, but it is unlikely that this approach would work well in a real-world context. Case-control separation is in fact induced largely by NA variants and experimental issues. In well-designed genetic studies, a data normalization step typically adjusts the dataset for population stratification issues. NA variants are also addressed in this step, as they may lead to spurious associations. By evaluating submissions through a crude ranking of samples with AUC, CD submissions that exploit patient stratification (intentionally or not) will be the most effective. Therefore, it is

important to improve CD challenge either by normalizing the dataset provided to participants or by asking for a set of causal SNVs in the submission. The former step is probably easier to implement, so we hope to see this improvement in future CAGI editions. By doing so, we believe that submission performance will reduce even further, leading to a truly effective validation of our current ability to predict disease risk. A well-structured dataset would also be a step forward to promote automated methods for disease risk prediction, as this should remove the manual analysis used to detect irrelevant facts like twins or pedigrees.

Overall, CAGI was successful in increasing the attention toward genome interpretation. In CAGI 2011, many participants tested a number of approaches. Many submissions were much worse than random (AUC < 0.5), suggesting that there was a substantial lack of expertise in the field. Over time, the assessor and participant talks shed light on the common pitfalls and shared the most effective ideas. This has increased remarkably our understanding of this field, and raised the interest of many researchers. Even though submission results are currently inflated, we believe that CAGI organizers are also in the process of learning how to evaluate properly disease risk predictions, and there are positive signals of improvement in this direction. Therefore, we believe that with the next editions, and with bigger and well normalized datasets, CAGI will play a role in evaluating and testing innovative methods and provide novel ideas for disease risk evaluation.

# Predicting Crohn's disease phenotypes from exome data in the CAGI 4 experiment.

The assessment part of this chapter comes "Daneshjou, R., Wang, Y., Bromberg, Y., Bovo, S., Martelli, P.L., Babbi, G., Lena, P.D., Casadio, R., Edwards, M., Gifford, D., Jones, D.T., Sundaram, L., Bhat, R., Li, X., Pal, L.R., Kundu, K., Yin, Y., Moult, J., Jiang, Y., Pejaver, V., Pagel, K.A., Li, B., Mooney, S.D., Radivojac, P., Shah, S., Carraro, M., Gasparini, A., Leonardi, E., Giollo, M., Ferrari, C., Tosatto, S.C.E., Bachar, E., Azaria, J.R., Ofran, Y., Unger, R., Niroula, A., Vihinen, M., Chang, B., Wang, M.H., Franke, A., Petersen, B.-S., Pirooznia, M., Zandi, P., McCombie, R., Potash, J.B., Altman, R.B., Klein, T.E., Hoskins, R.A., Repo, S., Brenner, S.E., Morgan, A.A., 2017. Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. Hum. Mutat. doi:10.1002/humu.23280".

# 1 Introduction

Precision medicine aims to use a patient's genomic and clinical data to make predictions about medically relevant phenotypes such as disease risk or drug efficacy[139,140]. Exome sequencing data, which captures exons and nearby flanking regulatory regions, is already being used clinically to solve medical mysteries with well-defined symptoms[141]. However, in order to advance precision medicine, clinicians and scientists will need to be able to make inferences about disease risk or drug efficacy from genetic data. Interpretation of genetic data is one of the major difficulties in the implementation of precision medicine[142]. At CAGI 4 (2016), three challenges involved making predictions using exome sequence. In particular the Crohn's disease challenge has been a part of previous CAGI iterations. In this chapter I will discuss the approach used by our group for Crohn's phenotype prediction. In addition, I will present the assessment of proposed predictive models, always in the context of the Crohn's disease challenge.

## 1.1 Crohn's disease

Crohn's disease is a chronic inflammatory pathology that can affect the entire gastrointestinal system, but most frequently affects the terminal part of the small intestine (ileum), and colon. Disease may be manifested at any age, but the incidence is higher between 15 and 30 years. In recent years, an increase has been observed in both the incidence and prevalence of the disease in all ethnic groups[143].

Main symptom is the presence of severe abdominal pain often accompanied by diarrhea. Symptoms can differ depending on the intestinal portion involved by the inflammation. In case of ileal localization, high volumes of blood-free aqueous feces are usually present, in the case of colitis instead, bloody diarrhea could be typically identified. In most of the cases fever, weight loss, and appetite decrease are often present. A number of extra-intestinal manifestations of the disease could be also identified. In these cases, inflammation may affect different organs and apparatus, including liver (e.g. primary sclerosing cholangitis), cute (e.g. erythema nodosum), eyes (e.g. uveitis) and joints (e.g. peripheral arthritis, ankylosing spondylitis).
Disease diagnosis is typically based on the presence of the symptoms described above. Unfortunately, most these signals are shared with other diseases and for this reason, patients affected by Crohn's disease may suffer symptoms for years before correct diagnosis is performed. In order to confirm diagnosis of Crohn's disease, several clinical analysis have to be carried out. Blood tests have little specificity as they may only indicate the presence of an inflammatory state (increased ESR, C- reactive protein, leukocytosis). More useful is the dosage of fecal calprotectin that can confirm the presence of intestinal inflammation. Useful analysis to achieve a more precise diagnosis are also: intestinal echography, which can detect wall thickening of the affected intestinal tracts, intestinal magnetic resonance tomography, and colon biopsy which is certainly the most reliable test for diagnosis of this inflammatory disease. As it is possible to understand, unfortunately, Crohn's disease diagnosis never relies on the results of a single investigation, but requires an overall evaluation of the patient's clinical status and the execution of several instrumental investigations.

Causes underlying this pathology have not been fully clarified yet, however the prevailing hypothesis is that Crohn's disease is an immune-related pathology due to an abnormal immune reaction of the intestinal tissues against antigens[144]. The onset of the inflammatory process seems to be due to a misregulated interaction between genetic factors and environmental factors.

Particular cases are the Very Early Onset (VEO) forms of this pathology. Early onset is defined by the onset of the disease within the 6$^{th}$ year of age[145].These forms of intestinal inflammation tend to be much more severe and much more difficult to control with conventional therapies, compared with adult-onset Crohn's disease. Increasing evidences

suggest a stronger genetic contribution to these forms, compared to "classical" manifestations[145]. Distinguish such forms of the pathology has a crucial importance to provide better clinical treatments. In this context the identification of pathogenic genes by means of NGS technologies could lead to more rapid diagnosis and the definition of specific treatments for these severe forms of the pathology.

### *1.1.1 Genetic factors*

Genetic predisposition to the onset of this inflammatory disease is confirmed by the presence of 35% of concordance among monozygotic twins for this pathological phenotype, whereas in dizygotic twins phenotype concordance is present in only 3% of cases[146]. Current knowledge on the molecular basis this pathology has been mainly achieved thanks to gene association studies, followed by experimental validation and clinical investigations. Unfortunately, these studies have allowed to understand only part of the genetic background underlying the abnormal inflammatory response.

The main gene associated with Crohn's disease is NOD2, which encodes for a protein responsible for the recognizing of muramyl dipeptide, a component of the bacterial wall. This protein, expressed primarily in epithelial cells and in the Antigen Presenting Cells (APC), once activated by the binding to muramyl dipeptide, activates in turn the transcription factor NF-kB and the MAP kinases pathway, leading to production cytokines and antimicrobial peptides. Experimental analysis confirms that subjects carrying mutations in NOD2, after stimulation with bacterial peptidoglycan, exhibit a reduced release of proinflammatory cytokines and a reduction in the activation of NF-kb complex[144].

Another gene playing a role in disease onset is ATG16L1. This gene is involved in the autophagy process, a mechanism by which cells can eliminate cytoplasmic components of degraded organelles and microbial fragments.

In addition to the cellular processes involved in the innate immunity, such as the ones listed above, also pathways of adaptive immune response are involved in Crohn's disease onset. An example is the case of signaling activated by interleukin-23 (IL-23). The IL23R gene, one among the most associate to the pathogenesis, encodes for one of the two components of the heterodimeric receptor for the IL-23. Main function of the receptor, once binding with the interleukin has been established, is to determine the activation of the JAK-STAT signaling pathway, which regulates the transcription of several genes,

including IL-23 (Abraham et al., 2009). IL-23, in turn, contributes to regulating the proliferation and survival processes of Th-17 lymphocytes a typical misregulated process in autoimmune inflammatory diseases. Finally, also variants in the interleukin-10 receptor (IL-10R), master regulator of intestinal mucosal homeostasis[147], seem to be involved in the onset of this inflammatory pathology.

### *1.1.2 Environmental factors*

Genetic predisposition is not sufficient to explain the onset of the disease by itself. The importance of environmental factors is mainly suggested by the relatively low concordance of the phenotype in monozygotic twins. A second aspect suggesting a role for the environmental component is the increase in incidence of the phenotype in ethnic groups which, as a result of migrator process, have shifted from low-incidence regions to areas characterized by a higher incidence[143]. Many aspects of life-style in industrialized societies have been linked to the onset of Crohn's disease such as improved hygienic conditions, sedentary lifestyle, junk food diet, exposure to pollutants and antibiotics consumption. Above all these factors, however, the most important seems to be cigarette smoking. It has been highlighted that early use of tobacco can significantly increase the likelihood of developing the disease. In addition, smokers generally present more severe phenotypic manifestations of the pathology if compared with non-smoking subjects[148]. Crohn's disease has also been linked with alterations in the intestinal microbiota. In particular, thanks to metagenomics studies, it has been highlighted that affected subjects present biodiversity reduction in the intestinal microbiota, in particular within the Firmicutes and Bacteroidetes phyla[143]. It has to be pointed out that to date it has not yet been clarified how this decrease in intestinal biodiversity can contribute to disease onset. Experiments on murine model have however shown that these alterations, alone, are not sufficient to be causative of disease onset that necessarily requires the presence of a genetic predisposition[149].

## 1.2 Crohn’s Disease Challenge in CAGI 4

The CAGI 4 dataset was made by 111 unrelated German ancestry exomes. For CAGI 4, submitting groups were allowed to use the data from the Crohn’s disease CAGI challenges of 2011 and 2013 as training sets. As in all iterations of the challenge, groups were asked to report for each individual a probability to be affected by Crohn’s disease between 0

and 1 (0 healthy, 1 affected) and a standard deviation representing confidence in that prediction.

# 2 Materials and methods

## 2.1 CAGI 4 Datasets

### *2.1.1 Test set*

The CAGI 4 dataset was made by 111 exomes (64 cases, 47 controls) sequence from unrelated German individuals. Proportion of healthy and affected individuals was not revealed during the prediction season. Exome sequencing was performed using the TruSeq exome enrichment kit (Illumina) and the Illumina HiSeq2000 instrument. Reads were mapped to the human genome build hg19, and variants were called for all 111 exomes together using the Genome Analysis Toolkit (GATK version 3.3-0) Haplotype Caller. Variant calls were restricted to the TruSeq exome target. GATK was also used for variant quality score recalibration, and only high quality variants passing the filters were retained. Further information on data processing could be retrieved in the header of the provided *.vcf* file. In CAGI 4 predictors could use data from the previous iteration of the Crohn's disease challenge as training sets. To further improve the possibility to use old data for training , both CAGI 3 and CAGI 2 datasets have been supplied with onset age for every individual.

Being aware of possible bias in data distribution due to batch effects or population stratification, a clustering analysis of the CAGI 4 dataset was performed (See Figure 15). To further investigate the presence of artifacts that could affect the prediction process, the distribution of number of non sequenced positions has been investigated (See Figure 16).

### *2.1.2 Training sets*

### *CAGI 2 (2011) dataset*

The CAGI 2 dataset was composed by 56 exomes (42 cases, 14 controls), all of German ancestry. As reported by data providers, sequencing was performed for all the samples using an Illumina sequencing platform. Base calling was performed with the Burrows–Wheeler Aligner (BWA) followed by duplicate removal by mean of Picard tools and

SAMtools Pileup. The chromosomal positions in the final *.vcf* files were according to hg18 human build. No information were provided about the possible presence of population structure in the dataset.

Looking for the presence of possible batch effects, clustering analysis (See Figure 11) and investigation of total number non sequenced positions have been performed (See Figure 12).

### *CAGI 3 (2013) dataset*

The CAGI 3 dataset was made by 66 exomes (51 cases, 15 controls). Again, the TruSeq exome enrichment kit was used for exomes capture and all the samples were sequenced using the same protocol. Variant calls was made for all 66 samples together, providing better quality calls. A TruSeq exome bed file was used for combined variant calling for all the 66 exomes using the GATK program. The Variant Quality Score Recalibration (VQSR) method was employed to identify true polymorphisms in the samples rather than those due to sequencing, alignment, or data processing artifacts. The chromosomal positions in the final *.vcf* files were according to hg19 human build.

Out of 66 exomes, 51 are Crohn's patient and rest are healthy. Samples were of German ancestry, cases were selected from 28 pedigrees of families with multiple cases of affected individuals, including 1 monozygous discordant twin pair. For this reason, some of these cases were obviously related. Controls instead, were unrelated healthy individuals. Exceptions were the unaffected parents of three cases and the unaffected twin of one case. Even in this case a clustering analysis (See Figure 13) followed by investigation of total number of non sequenced positions was performed (See Figure 14).

## 2.2 Annovar

ANNOVAR (ANNOtate VARiation)[150] is a bioinformatics tool that allows annotation of high-throughput sequencing data. This tool consists of a command-line application that can be used on a standard PC or cluster platform that has Perl modules installed. Since 2012, even a web server called wANNOVAR[151] exists, providing a simple and intuitive interface to help users in the definition of the functional significance of variants identified in sequencing experiments. In the context of the Crohn's disease challenge, the stand-alone version of the tool was used to annotate the 111 exomes. To perform the annotation, many databases have to be previously downloaded on your local machine. An essential database for the annotation was certainly the RefSeqGene[152] database that has

been crucial to identify genes positions. Thanks to the information present in the RefSeqGene database we have been able to identify for each variant, if it is intronic, exonic, intergenic, or present in the 5 '/ 3' UTR regions of the gene. Other crucial information has been retrieved from the 1000 Genomes Project (1kGP)[153] database from which, information on the Minor Allele Frequency (MAF) of each variant was obtained.
The efficiency of this software allow us to annotate the exomes of all the 111 individuals in about one hour of calculation using a Linux desktop computer, equipped with an AMD 4300 series, quad core CPU (3.8 GHz) and 8 GB of RAM.

## 2.3 PheGenI

Among the available resources that allow access to data obtained by GWAS studies, we chose the PheGenI (Phenotype-Genotype Integrator)[154] database. This resource integrates content from numerous NIH databases (National Institutes of Health), allowing access to more than 66,000 associations between nucleotide variants and phenotypic traits. Associations are derived from two databases: dbGaP[155] a database containing data of studies investigating associations between genotypes and phenotypes, and the NHGRI GWAS Catalog[156] which contains GWAS data derived from the literature. These data are integrated with information obtained from other databases such as dbSNP[157], which collects information on SNPs, and the NCBI Gene database which contains gene names, chromosomal localization and protein products encoded by each gene[154].
Two kinds of queries can be carried out by mean of graphic interface: genotype and phenotype-oriented query.
Phenotype-oriented queries exploits data in dbGaP and GWAS Catalog, variants are assigned to a specific phenotype using the Medical Subject Headings (MeSH)[158] categories. Genotype-oriented queries can be performed by searching for a specific gene, SNP, or chromosomal region, resulting with all associations identified with that specific target.

## 2.4 STRING

STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a PPI database. This resource contains both direct interactions such as physical link between proteins, as well as indirect associations e.g. the ones derived by the fact that 2 proteins are involved in the same cellular process[159].

This interaction database is constantly updated and to date it contains interactions between approximately 9.6 millions proteins belonging to 2031 different organisms[160].
Interaction data can be divided into three types, depending on how the interaction has been defined: interactions detected by direct PPI experiments, associations defined by the fact that proteins that are involved in the same metabolic or signaling process and *de novo* interactions predicted by means of computational techniques. Predicted interactions are defined in turns, in three different ways: text mining of scientific literature, predicted interactions based on shared genomic characteristics, and predicted interactions based on orthology relationships between organisms. As interaction data can be obtained in very different ways, a score representing data reliability is assigned to each interaction. For predicted interactions and associations obtained by mean of high throughput experiments, interaction score is assigned by comparing predicated data with data present in the KEGG database[161], used as reference database for interactions. In particular, each association between 2 proteins that are assigned to the same metabolic or biochemical pathway in KEGG is defined as a true positive interaction. For the interactions defined by mean of low throughput experiments, associations in known protein complexes and manually curated metabolic pathways, the confidence score is assigned according to the source from which the information is derived[162].

## 2.5 KEGG

The KEGG (Kyoto Encyclopedia of Genes and Genomes) is an on-line resource made up of numerous databases, developed to allow a systematic analysis of genes functions. The purpose of this database is to enable the understanding of the mechanisms underlying cells and organisms functions, focusing on genomic data[163].
Databases in KEGG can be categorized into three main categories.
The catalog of all genes found in the genomes of fully sequenced organisms and some partially sequenced genomes is contained in the GENES[164]. In this database, each gene is defined by a specific identifier that allows links to orthologs present in other species. The second type of database contains information about chemical reactions that occur in the cell: the LIGAND database. This resource contains information about chemical compounds and relevant cell reactions. Among these chemical compounds is possible to find cellular metabolites, drugs, and other environmental compounds. In this database

chemical reactions are typically enzymatic reactions. Finally, the third type of database contains information on protein interaction networks, such as signaling pathways and protein complexes involved in various cellular processes. This type of information is present in the PATHWAY database. This database can be considered as an attempt to compute current knowledge on PPI networks by means of a representation based on the graph theory. Within this database, each cellular process is described by mean of a graph in which gene products (proteins) correspond to nodes of the network that could be connected by means of three types of connections. For metabolic processes, the connections define enzyme-enzyme relationships, in which two enzymes catalyze metabolic reactions occurring in succession along the same metabolic pathway. The second type of connection defines direct protein-protein interactions, such as binding between two proteins, phosphorylation and ubiquitination. The third type of connection are gene expression interactions that describe relationship between transcription factors and protein products. Graphs describing each cellular process are typically manually produced exploiting literature data. In addition, interaction networks are expanded by exploiting information derived from high throughput experiments, such as two-hybrid assays to infer protein interactions, microarray to infer relations between co-expressed genes, and comparing graphs describing same cellular process in different organisms.

## 2.6 Prediction Strategy

Prediction strategy was planned taking in strong consideration lessons learned from the previous iteration of this CAGI challenge.

In the previous editions, our predictions were used to be based on different implementation of agglomerative clustering to uncovered hidden relationships between samples. To test if this kind of approach could be useful to achieve good prediction performance even in the last iteration of this challenge a clustering analysis followed by investigation of total number of non sequenced positions was performed over all the available datasets. For the CAGI 2 dataset a strong bias in samples distribution was identified, with most of the controls clustering together (See Figure 11). In addition it seems clear that a discrimination of patients based on the simple count of variants that did not match the reference genome could achieve great results (See Figure 12). In the case of CAGI 3 dataset, is possible to see that once again a substantial difference in

clustering between cases and controls (See Figure 13 and Figure 14). However this biased is reduced if compared with the CAGI 2 dataset and a substantially more homogeneity cold be observed among the cases. Considering this result we decided to focus only on the CAGI 3 datasets for analyses aimed to increase prediction performance for this last edition of the Crohn's disease challenge. For the current CAGI 4 dataset, the analysis demonstrates that no trivial bias in data distribution (See Figure 15 and Figure 16). It has to be noted that this was the first time, among all the iteration of the Crohn's disease challenge, where predictions cold be performed on a homogenous dataset.

In this context, seemed clear that a strong reduction of prediction performance could be recorded if our clustering prediction approach would be tested in a homogenous dataset. For this reason a completely new strategy was designed to approach the CAGI 4 Crohn's disease challenge. Prediction strategy for this last CAGI edition was based on a simple working hypothesis: the higher is the mutation burden for Crohn's related variants, the higher should be the probability for an individual to be affected. This working hypothesis has 2 advantages respect to our previous prediction strategy. First, no explicit influence of population structure is assumed. In this way the presence of bias or batch effects should not affect our prediction performance. Second, as predictions will be based on a pre-defined set of variants related to Crohn's disease, the biological background that guided our prediction would be clear, leading to a better comprehension of the molecular basis that triggered disease onset in affected patients.

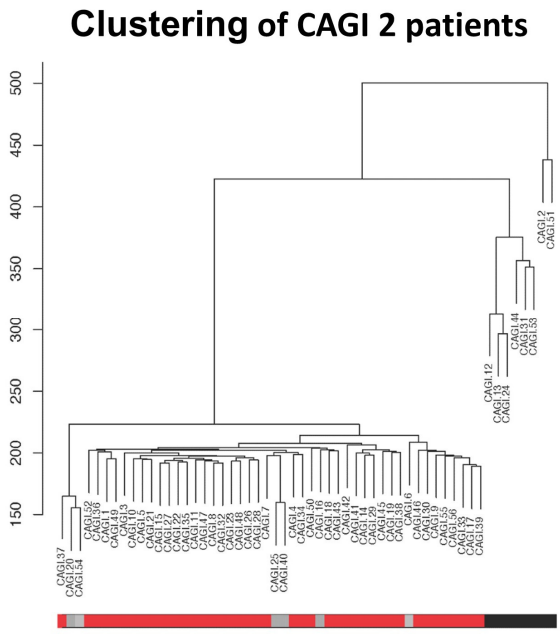


**Figure 11. Clustering of patients from the CAGI 2 Crohn's disease challenge.** Black and grey bars at the bottoms are representative of controls, red represents the cases. As it is possible to see, strong bias in samples distribution could be identified with most of the controls clustering together.

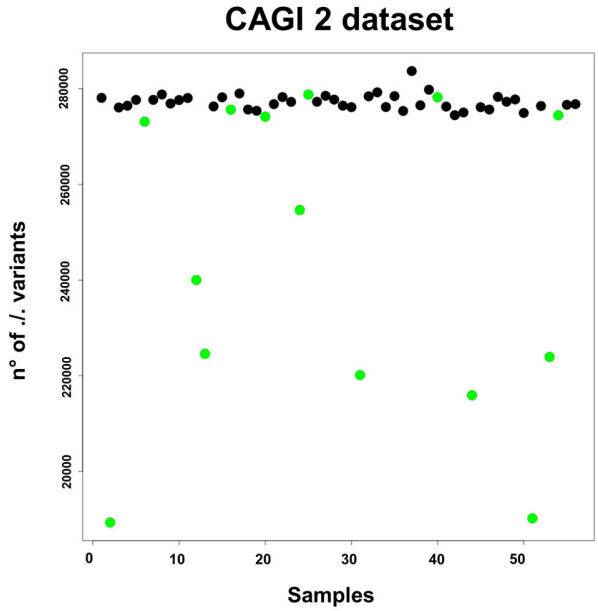


**Figure 12. Distribution of non sequenced positions in the CAGI 2 dataset.** Black represents cases, green represent controls. Evident differences could be found between cases and controls, with most of the controls what could be clearly classified considering this feature.

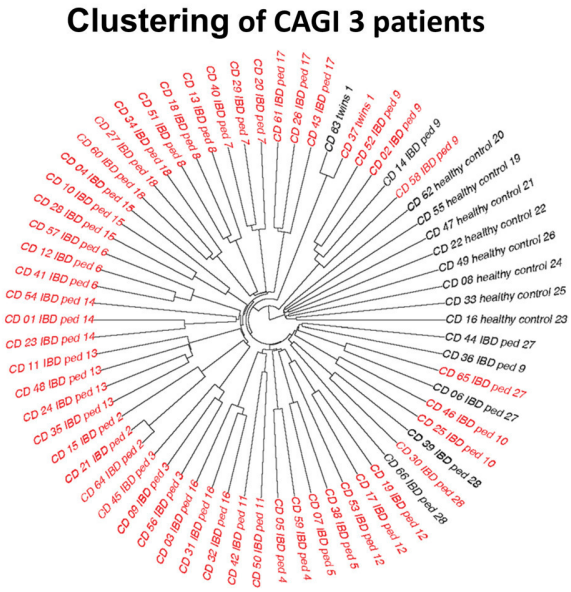


**Figure 13. Clustering of patients from the CAGI 3 Crohn's disease challenge**. Black represents controls, red represent cases. As it is possible to see even in this dataset strong bias in samples distribution could be identified.

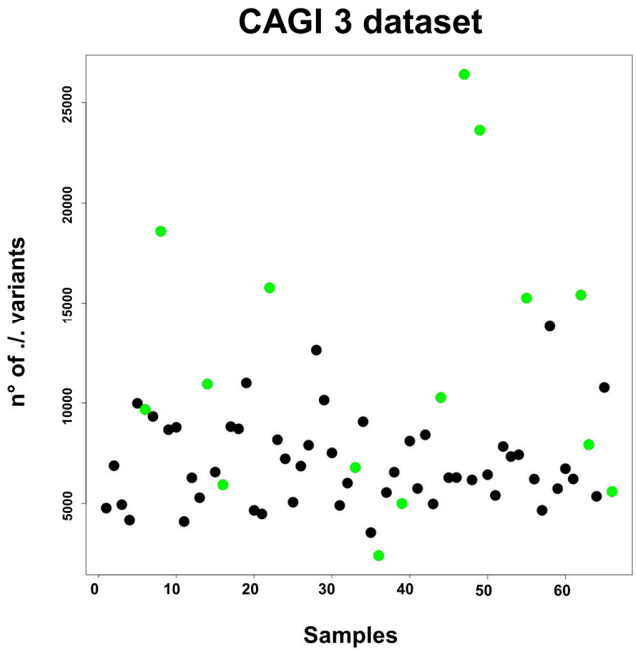


**Figure 14. Distribution of non sequenced positions in the CAGI 3 dataset.** Black represents cases, green represent controls. In this case only few individuals could be classified exploiting this feature.

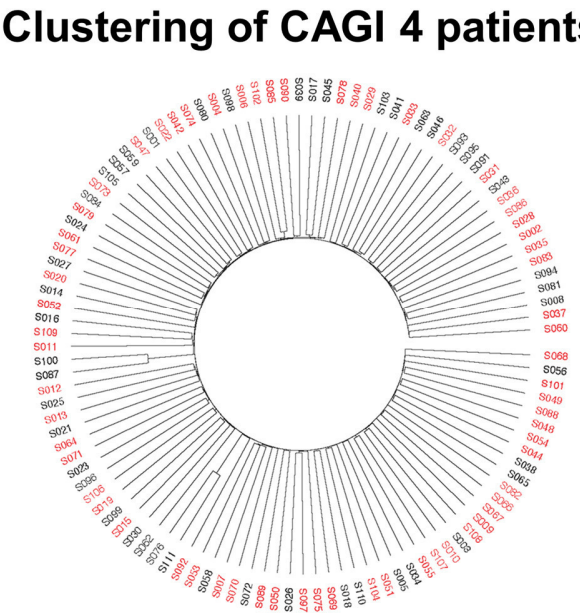


**Figure 15. Clustering of patients from the CAGI 4 Crohn's disease challenge.** No trivial dataset stratification could be found.

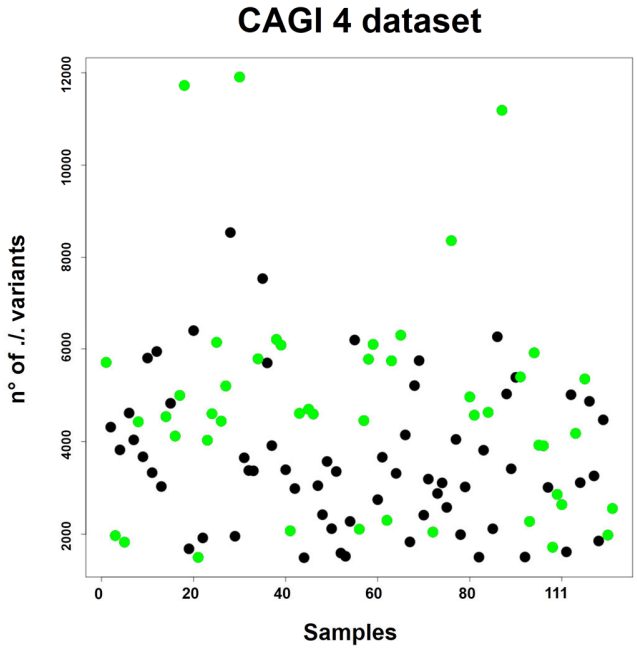


**Figure 16. Distribution of non sequenced positions in the CAGI 4 dataset.** Unless for a little subset of individuals, no evident population stratification could be identified. Healthy control are in green, black is for affected individuals.

Predicting disease status of unknown individuals considering only exome data is a rather complex challenge. In each exome in fact, dozens of thousands of variants could be identified. In order to discriminate between healthy and affected patients, it is necessary to reduce the number of variants to be analyzed, eliminating neutral variability and focusing only on variants involved in the onset of the phenotype of interest. In this context, prediction strategy for CAGI 4 Crohn's disease challenge was composed by three phases: annotation and first variants reduction, definition of genes involved in pathology and second variants reduction, and finally, discrimination of healthy and affected patients (See Figure 17).

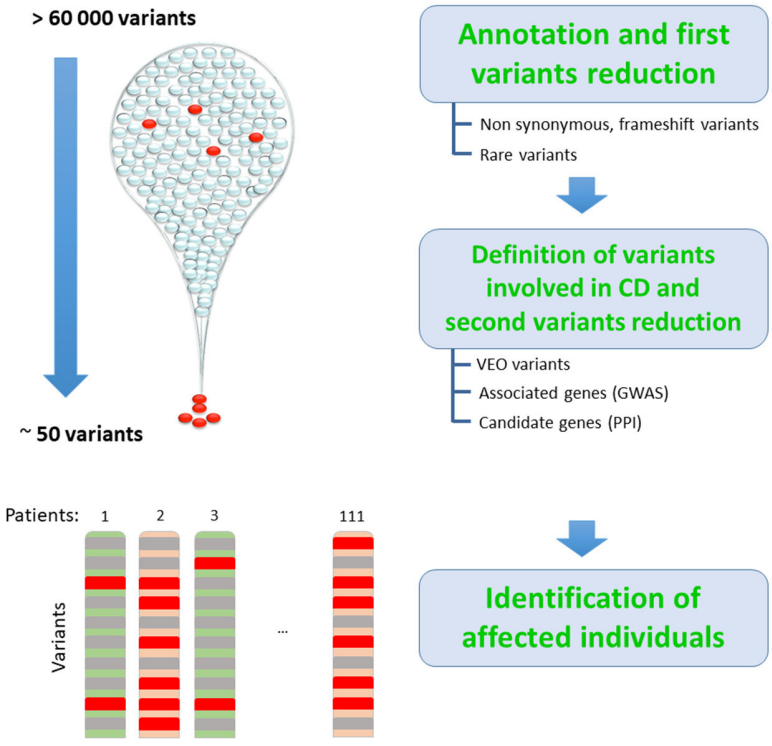


**Figure 17. Representation of our prediction workflow.** On the left side, the bottleneck figure represents the need to reduce the number of variants to be considered for phenotype prediction. On the left side the three steps that compose our prediction strategy are listed.

## *2.6.1 Annotation and first variants reduction*

Before starting with the analysis of the genetic data, a first annotation phase was performed. In this phase all variants have been mapped on the reference genome. For this purpose we decided to use the ANNOVAR[151] annotation tool.

Once the annotation of all variants present in exomes of the 111 individuals was completed, a first variant reduction phase was carried out. Thanks to this first filter, we started to reduce the amount of data to be analyzed for classification purpose. First, all variants that are less likely to influence human phenotypes have been removed. In this way, all synonym mutations has been sifted out, while all the mutations that caused non-synonymous and frameshift variations has been retained.

We further decide to eliminate also all common variants. This choice was based on the hypothesis that the major contribution to susceptibility for complex pathologies is due to mutations present at low frequency in the population[24]. Based on this consideration, we filtered out all variants with MAF >5%[165] in the 1kGP project.

Resuming this first phase, in order to reduce the number of variants to be analyzed, we sifted out all mutations that do not affect the amino acidic sequence, as well as all variations that, by their frequency, are less likely to be involved in a pathological phenotype. Thanks to this phase, variants to be considered for prediction purpose has been reduced by about one third, from 60 000 to less than 20 000 for each patient.

## *2.6.2 Definition of variants involved in pathology onset and second variants reduction*

In this second phase our attention was focused on variants and genes that if mutated could increase the probability to be affected by Crohn's disease. However, it is necessary to remember that the molecular etiology of this inflammatory disease has not be fully understood yet. For this reason, in order to consider all possible variants involved in the onset of this pathology, we decided to define three variants lists. In a first list we store all variants associated to the VEO forms of the pathology. A second list was made considering genes associated to the phenotype by means of GWAS studies. Last, a third list was made performing an expansion of the GWAS genes list, exploiting information retrieved from the PPI network.

## *Definition of a list of variants associated to the VEO forms of the pathology*

As already introduced, some particular forms of Crohn's disease have a very early phenotypic manifestation. These forms are not only characterized by the precocious onset of the pathology, but also present a more severe manifestation and a reduced responsiveness to pharmacological treatments.

Increasing evidences suggest a stronger genetic contribution to these forms of the pathology, respect to adult onset manifestations[145]. For this reason we decided to start the collection of variants to be used for phenotype prediction, looking for mutations associated to these severe forms of disease. As genetic investigation of VEO manifestations is still a young field of research, dedicated databases to automatically retrieve variants involved in this phenotype have not been identified. For this reason a manual curation of the recent scientific literature has been performed. In particular, we focus our attention on case reports and works were VEO variants have been detected by means of WES or WGS experiments. Thanks to this analysis we retrieved a list of 77 variants associated to VEO manifestations of Crohn's disease. In order avoid possible misinterpretation of variants position, only variants defined by a Reference SNP "rs" ID have been further considered. In this way only 48 mutations have been retained. Finally, only 20 of these variants were present in at least one individual in the CAGI 4 datasets, and have been considered for phenotype prediction (See Table 20).

| Variant ID | Affected gene | Variant ID | Affected gene |
|---|---|---|---|
| rs41313262 | IL23R | rs5743266 | NOD2 |
| rs11209026 | IL23R | rs2066842 | NOD2 |
| rs2241880 | ATG16L1 | rs2066843 | NOD2 |
| rs72553867 | IRGM | rs2066844 | NOD2 |
| rs10065172 | IRGM | rs5743277 | NOD2 |
| rs4252249 | IL10RA | rs2066845 | NOD2 |
| rs3135932 | IL10RA | rs2303015 | NDP52/CALCOCO2 |
| rs2228054 | IL10RA | rs8178561 | IL10RB |
| rs2228055 | IL10RA | rs1058867 | IL10RB |
| rs2229113 | IL10RA | rs34688635 | NOX1 |

**Table 20. Variants associated to VEO form of Crohn's disease.** Variants associated to VEO manifestation of the pathology retrieved by manual curation of the academic literature.

To investigate if these 20 variants could be useful to classify affected individuals and healthy controls, we defined the following working hypothesis: the highest is the mutation burden for these VEO SNPs, the earlier should be the onset of the pathology. To test this hypothesis we exploit exomes and onset ages retrieved from the CAGI 3 Crohn's disease dataset. The Pearson correlation coefficient was calculated between the counts of VEO variants and the onset age for all the 66 individuals present in the dataset. A correlation of -0.37 was found to link mutation burden and onset age, confirming our working hypothesis and suggesting that these variants could be useful to distinguish healthy and affected patients. The presence of an only moderate correlation could be justified by the fact that a simple mutation count, could not account for the different impact of heterozygous and homozygous variants. In addition, due to the phenotype heterogeneity present in the CAGI 3 dataset, it is possible to imagine that these variants alone, could not be sufficient to correctly separate healthy and affected individuals. For this reason we expand this short list of variants, considering also variants identified in GWAS studies and information retrieve from PPI network.

## *Definition of a list of genes associated to the pathology by means of GWAS studies*

The underlying hypothesis of a Genome Wide Association Study is that the presence of genetic polymorphisms could be related with an increased or decreased probability to develop a phenotypic trait. By means of these studies in fact, it is possible to define if some variants could predispose to the onset of a pathology or instead could have a protective role.

Variants associated to Crohn's disease by means of GWAS studies, have been retrieved from the PheGenI[154] database. A phenotype-oriented query was performed according to the MeSH term "Crohn Disease". In this way a list of 138 variants associated to the phenotype of interest has been retrieved. We than focus our attention on the genes affected by these variants: some variants in fact, could affect the same gene, while others could fall in intergenic regions. In this second case, both the upstream and downstream genes have been considered. In this way, a list of 133 genes associated to Crohn's Disease was obtained (See Table 21).

| Genes associated to Crohn's disease - PheGenI | | |
|---|---|---|
| IL23R | RPL12P7 | ALDH7A1P4 |
| ATG16L1 | RPL30P13 | ZNF300 |
| NOD2 | FABP3P2 | GSDMC |
| CARD9 | PRDX5 | PTPN2 |
| DAB2 | DNMT3A | IPMK |
| ZMIZ1 | CDKAL1 | IL10 |
| PDGFB | ITLN1 | CCL7 |
| ZNF365 | C11orf30 | ADRA1B |
| C5orf56 | C21orf33 | CPEB4 |
| GOT1 | FASLG | FOXD1 |
| SMAD3 | MRPL11P2 | MAMSTR |
| IRGM | RPS12P16 | LIF |
| PVT1 | RCL1 | KIF21B |
| GPR65 | IL2RA | FNDC1 |
| PSMG2 | PLCL1 | 3.8-1.5 |
| MST1 | PUS10 | PRDM1 |
| MRPS35P3 | BACH2 | ZFP36L1 |
| NKX2-3 | C7orf72 | LRRK2 |
| UBE2L3 | TRIB1 | TNFSF11 |
| TNFSF15 | ZPBP2 | CCDC88B |
| ZGPAT | PER3 | LRRC32 |
| RPS14P1 | B3GNT2 | ICOSLG |
| THADA | SATB1 | TNFSF18 |
| SCAMP3 | SLC22A23 | NDFIP1 |
| C13orf31 | NRIP1 | CUL2 |
| CCL2 | SLC43A3 | JAK2 |
| SP140 | DENND1B | IKZF1 |
| IL12B | SLC7A10 | FAM84B |
| FGFR1OP | PTPN22 | GSDMB |
| TYK2 | SLCO6A1 | TMEM17 |
| IL18RAP | C8orf84 | KCNH8 |
| BOD1 | FAM5C | CYCSP42 |
| TMEM174 | TCERG1L | RTN4RL2 |
| STAT3 | BSN | CEBPA |
| FUT2 | IL3 | PAM |
| RPS3AP51 | PSMB10 | RGS18 |
| C1orf81 | FLJ45139 | FLJ46300 |
| TAGAP | HLA-DQB1 | CSF2 |
| C11orf10 | SLC22A5 | RPL23AP12 |
| CLN3 | NELL1 | HLA-DQA2 |
| GCKR | CNTNAP2 | RPL32P17 |
| ERAP2 | C10orf67 | OTUD1 |
| C5orf62 | KLF6 | AKR1E2 |
| IFITM4P | PTGER4 | |
| RPL35P3 | RPL3 | |

**Table 21. Genes associated to Crohn's disease.** Genes associated to Crohn's disease by means of GWAS studies. Data have been retrieved from the PheGenI database.

Unfortunately, it is well established that these genes are not sufficient to explain the entire inheritance of this inflammatory pathology. Recent studies demonstrate that the 71 genes with the highest association score are able to explain only 21.5% of the Crohn's disease estimated inheritability[166]. In addition, lessons learned from the previous edition of this challenge suggest that even these genes, alone, are not sufficient to clearly separate affected and healthy individuals. In order to effectively identify the phenotype of the unknown individuals, we decided to further expand this gene list exploiting the PPI network.

## *Expansion of the candidate genes list, exploiting the PPI network*

In order to increase our ability to discriminate healthy and affected patients, we expanded the list of 133 genes obtained from PheGenI database, considering the interaction partners of proteins associated to Crohn's disease. It is well established in fact, that mutated proteins could produce perturbations of the PPI network, triggering disease onset. In several cases have been demonstrated that proteins involved in diseases could interact directly with each other or through common interactors[167]. Our aim was to identify these interactors and consider them as potential candidates, involved in Crohn's disease onset. The strategy used to expand the list of genes associated to the pathology was based on two steps.

First, exploiting the STRING database, we tried to define an interaction distance that could separates random interactors and interactors potentially involved in the pathology. To this aim, we tested if the proteins coded by genes associated to Crohn's disease are closer than random within the PPI network, assuming the existence of a cluster of genes involved in the onset of the disease.

Second, using the KEGG database, we expanded the list of candidate genes associated to the pathology considering the threshold distance defined in the previous analysis.

## Definition of an interaction distance to separates Crohn's genes and random interactors in the STRING database

To expand the list of candidate genes, we initially used the STRING database[159]. This database was chosen as the number of interactions present in this resource is extremely high. Due to the presence of both validated and predicted interactions, crucial, is the presence of a confidence score representing edges reliability.

As STRING is a protein-protein interaction database, we first identified proteins encoded by the 133 genes associated to Crohn's disease. To this aim, we exploited a STING mapping sub-database, which allowed us to link each gene to the encoded proteins. Thanks to this resource we identified 155 proteins. Of these, only 151 were present in the STRING PPI network.

We then tested if these proteins, are actually "closer" than random interactors, assuming the involvement of Crohn's associated proteins in a small subset of cellular process. Manipulations and analysis of interaction data has been performed by means of several R packages.

Initially, the STRING human interaction network was imported and converted in a graph object using the *igraph* package. In this way, a direct graph consisting of 20 136 proteins connected by 4 442 852 edges was obtained.

In order to verify our working hypothesis, the distribution of shortest paths between the Crohn's associated proteins and a random set proteins was compared. The first distribution was obtained by computing the minimum distance between all possible couples of the 151 associated proteins. Distribution of shortest path between random proteins, instead, was obtained calculating the minimum distance between 10 000 couples of proteins randomly selected within the network.

The shortest paths have been calculated exploiting the *shortest.paths* function of the *igraph* package (*Dijkstra algorithm)*.

Comparison of the two distributions confirmed our working hypothesis that proteins associated to the disease are closer than random, forming a cluster inside the PPI network (*Wilcoxon-Mann-Whitney* test, $p$-value < 0.01)

We than tried to define an interaction distance that will allow us to distinguish a random interactor from an interactor present within this cluster. To this aim, the mean of the shortest paths connecting the couples of pathogenic proteins, and the mean of the

shortest paths that connecting the couples of random proteins have been compared. The two mean values were respectively: 2.18 and 2.48.

To further confirm our hypothesis we than verified if these findings were artifacts caused by the presence of a high number of low-confidence connections, or due to the presence of proteins with a very high degree. In particular, it is possible that low-confidence edges, derived from predicted interactions and high-throughput experiments could be false positive. On the other hand, proteins with an extremely high degree, such as ubiquitin, could bias our analysis as, most of the shortest paths could be of length 2 (e.g. protein A-ubiquitin-protein B). To exclude these problems, our analysis has been repeated, removing from the network connections and nodes, considering different threshold of confidence scores and degree. For most of the thresholds, results of the comparison between the shortest paths distributions were in agreement with our working hypothesis (See Table 22)

It has to be noted that in some rare cases the two distributions are not significantly different. However, it has to be consider that these situations are limited to the combinations of very high thresholds, reflecting a dramatic resizing of the PPI network (See Table 22).

Resuming this part we can state that:

1. analyzing differences between the shortest path distributions (See Table 22, column 5), the hypothesis that proteins involved in Crohn's disease are "closer" respect to random interactors has been confirmed.
2. analyzing the average values of the two distributions (See Table CR 22, columns 3 and 4), we can conservatively define that an interaction distance that can be used to distinguish between interactors within the Crohn cluster and random interactors, is equal to one step (a direct interaction).

Considering these results only direct interactors of genes associated to Crohn's disease will be considered to expand the list of candidate genes.

| Confidence threshold | Degree threshold | Avg. Shortest Path Crohn | Avg. Shortest Path random | *p*-value | Biggest Connected Comp. dimension | Connected Comp. count | Connected Comp. > 10 nodes count |
|---|---|---|---|---|---|---|---|
| 150 | 2 050 | 2.12 | 2.42 | 0.001 | 20 038 | 39 | 38 |
| 150 | 1 550 | 2.13 | 2.44 | 0.001 | 19 986 | 41 | 40 |
| 150 | 1 050 | 2.17 | 2.47 | 0.001 | 19 657 | 57 | 56 |
| 150 | 550 | 2.28 | 2.71 | 0.001 | 17 994 | 106 | 105 |
| 150 | 50 | 6.00 | 4.33 | 0.157 | 3 655 | 1 221 | 1 219 |
| 362 | 2 050 | 2.94 | 3.32 | 0.001 | 19 143 | 932 | 931 |
| 362 | 1 550 | 2.98 | 3.34 | 0.001 | 19 085 | 940 | 939 |
| 362 | 1 050 | 3.09 | 3.44 | 0.003 | 18 739 | 973 | 972 |
| 362 | 550 | 3.34 | 3.82 | 0.001 | 16 931 | 1 166 | 1 165 |
| 362 | 50 | NA | NA | 0.000 | 316 | 3 320 | 3 295 |
| 574 | 2 050 | 3.32 | 3.82 | 0.011 | 15 878 | 3 954 | 3 953 |
| 574 | 1 550 | 3.33 | 3.84 | 0.019 | 15 796 | 3 985 | 3 984 |
| 574 | 1 050 | 3.45 | 3.98 | 0.010 | 15 354 | 4 101 | 4 100 |
| 574 | 550 | 4.04 | 3.34 | 0.227 | 13 202 | 4 596 | 4 595 |
| 574 | 50 | NA | NA | 0.000 | 95 | 4 307 | 4 295 |
| 786 | 2 050 | 3.32 | 4.08 | 0.001 | 12 314 | 7 473 | 7 472 |
| 786 | 1 550 | 3.30 | 4.08 | 0.001 | 12 213 | 7 533 | 7 532 |
| 786 | 1 050 | 3.62 | 4.34 | 0.001 | 11 694 | 7 713 | 7 712 |
| 786 | 550 | 4.56 | 5 | 0.116 | 9 385 | 8 294 | 8 293 |
| 786 | 50 | NA | NA | 0.000 | 60 | 4746 | 4 744 |

**Table 22. Summary of analysis performed on the STRING PPI network.** In this table are summarized results performed to confirm that Crohn's disease genes form a kind of cluster within the PPI network. The analyses have been repeated considering different thresholds for node degree and edge confidence score. NA values are due to the use of very demanding thresholds that cause the removal of nodes used to perform calculations.

## Definition of new candidate genes in the KEGG database

Confirmed that genes associated to Crohn's disease are involved in a kind of "cellular module" and defined the threshold distance characterizing interactors present within this cluster, we expanded the list of candidate genes involved in the disease.

To select these new putative genes we used the KEGG[163] database. We decided to focus on this resource because interactions, even if less in number respect to STRING database, are more reliable thanks to the manual curation process.

Using the *KEGGREST* package we searched for the 133 genes involved in Crohn's disease, inside the 281 human cellular processes (KEGG PATHWAYS sub-database). 16 cellular processes contained at least 5 (arbitrarily threshold) genes involved in Crohn's disease. These pathways have been merged in single graph considered only nodes representing genes and removing from the network all nodes that represent chemical compounds or links to other pathways.

A final network composed by 1 298 genes and 8 041 edges was created.

Within this graph we select all the direct interactors of genes involved in Crohn's disease, obtaining a list of 210 new candidate genes (See Figure 18 and Table 23).

At this point, the list of 20 variants associated to VEO forms of the pathology, the 133 genes retrieved from PheGenI and the 210 putative genes identified thanks to network expansion were used to further filter the exomes of the 111 individual. Thanks to this second phase of variants reduction, it was possible to reduce the number of mutations to be considered for phenotype prediction from around 20 000 to less 50 per patient.

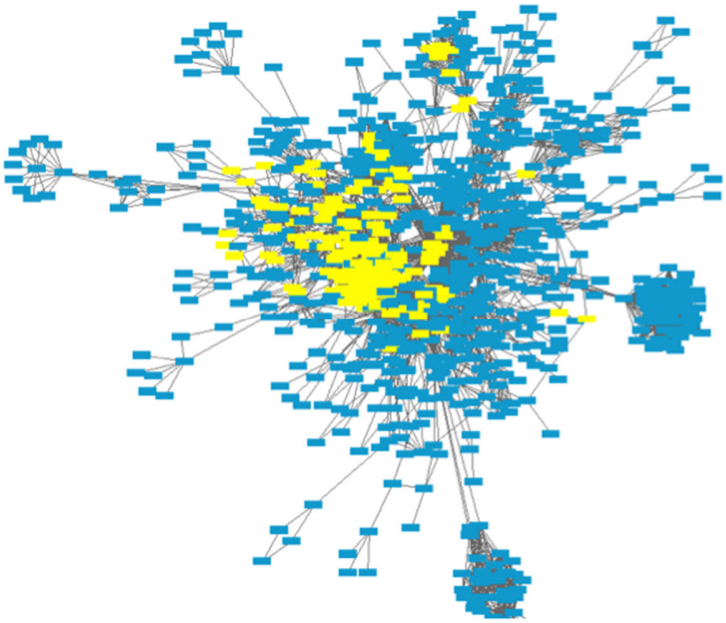

**Figure 18. Selection of candidate genes in the KEGG network.** Image representing the custom network created by the combination of 16 KEGG pathways. Highlighted nodes represent GWAS associated genes and their direct interactors.

| Candidate gene | n° of GWAS interactors | GWAS interactors |
|---|---|---|
| IL12RB1 | 9 | CSF2 IL3 IL10 IL12B JAK2 LIF STAT3 TYK2 IL18RAP |
| IL21R | 8 | CSF2 IL3 IL10 IL12B JAK2 LIF STAT3 TYK2 |
| IL12RB2 | 8 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 IL18RAP |
| CSF3R | 8 | CEBPA CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL20RB | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL10RB | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL22RA1 | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL2RG | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL2RB | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| CSF2RB | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| LIFR | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL6ST | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IFNLR1 | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL22RA2 | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL20RA | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL10RA | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IFNGR2 | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IFNGR1 | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IFNAR2 | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IFNAR1 | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| MPL | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| PRLR | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| GHR | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| EPOR | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| CRLF2 | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL7R | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL15RA | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL9R | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL4R | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL5RA | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL3RA | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| CSF2RA | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL13RA1 | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| LEPR | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| CNTFR | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| OSMR | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL11RA | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL6R | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |

| Candidate gene | n° of GWAS interactors | GWAS interactors |
|---|---|---|
| IL13RA2 | 7 | CSF2 IL3 IL10 IL12B JAK2 LIF TYK2 |
| IL6 | 5 | CEBPA IL23R IL2RA JAK2 STAT3 |
| SOCS4 | 5 | IL23R IL2RA JAK2 STAT3 TYK2 |
| SOCS7 | 5 | IL23R IL2RA JAK2 STAT3 TYK2 |
| SOCS1 | 5 | IL23R IL2RA JAK2 STAT3 TYK2 |
| SOCS2 | 5 | IL23R IL2RA JAK2 STAT3 TYK2 |
| SOCS3 | 5 | IL23R IL2RA JAK2 STAT3 TYK2 |
| SOCS5 | 5 | IL23R IL2RA JAK2 STAT3 TYK2 |
| PTPN6 | 4 | IL23R IL2RA JAK2 TYK2 |
| CBLC | 4 | IL23R IL2RA JAK2 TYK2 |
| CBL | 4 | IL23R IL2RA JAK2 TYK2 |
| CBLB | 4 | IL23R IL2RA JAK2 TYK2 |
| JAK1 | 4 | IL23R IL2RA IL3 STAT3 |
| JAK3 | 4 | IL23R IL2RA IL3 STAT3 |
| IL4 | 3 | IL23R IL2RA JAK2 |
| IFNB1 | 3 | IL23R IL2RA JAK2 |
| IFNA1 | 3 | IL23R IL2RA JAK2 |
| IFNA2 | 3 | IL23R IL2RA JAK2 |
| IFNA4 | 3 | IL23R IL2RA JAK2 |
| IFNA5 | 3 | IL23R IL2RA JAK2 |
| IFNA6 | 3 | IL23R IL2RA JAK2 |
| IFNA7 | 3 | IL23R IL2RA JAK2 |
| IFNA8 | 3 | IL23R IL2RA JAK2 |
| IFNA10 | 3 | IL23R IL2RA JAK2 |
| IFNA13 | 3 | IL23R IL2RA JAK2 |
| IFNA14 | 3 | IL23R IL2RA JAK2 |
| IFNA16 | 3 | IL23R IL2RA JAK2 |
| IFNA17 | 3 | IL23R IL2RA JAK2 |
| IFNA21 | 3 | IL23R IL2RA JAK2 |
| PRL | 3 | IL23R IL2RA JAK2 |
| GH2 | 3 | IL23R IL2RA JAK2 |
| CSH1 | 3 | IL23R IL2RA JAK2 |
| EPO | 3 | IL23R IL2RA JAK2 |
| IL21 | 3 | IL23R IL2RA STAT3 |
| IL7 | 3 | IL23R IL2RA JAK2 |
| IL2 | 3 | IL23R IL2RA JAK2 |
| CSF3 | 3 | IL23R IL2RA JAK2 |
| OSM | 3 | IL23R IL2RA JAK2 |
| CSH2 | 3 | IL23R IL2RA JAK2 |

| Candidate gene | n° of GWAS interactors | GWAS interactors |
|---|---|---|
| CSH2 | 3 | IL23R IL2RA JAK2 |
| NFKB1 | 3 | IL12B CARD9 IL18RAP |
| RELA | 3 | IL12B CARD9 IL18RAP |
| PIK3R5 | 3 | IL2RA JAK2 TYK2 |
| PIK3CA | 3 | IL2RA JAK2 TYK2 |
| PIK3CB | 3 | IL2RA JAK2 TYK2 |
| PIK3CD | 3 | IL2RA JAK2 TYK2 |
| PIK3CG | 3 | IL2RA JAK2 TYK2 |
| PIK3R1 | 3 | IL2RA JAK2 TYK2 |
| PIK3R2 | 3 | IL2RA JAK2 TYK2 |
| PIK3R3 | 3 | IL2RA JAK2 TYK2 |
| STAT6 | 3 | IL10 JAK2 TYK2 |
| IL26 | 2 | IL23R IL2RA |
| IFNL1 | 2 | IL23R IL2RA |
| IFNL3 | 2 | IL23R IL2RA |
| IFNL2 | 2 | IL23R IL2RA |
| IL22 | 2 | IL23R IL2RA |
| IL24 | 2 | IL23R IL2RA |
| IL20 | 2 | IL23R IL2RA |
| IL19 | 2 | IL23R IL2RA |
| IFNG | 2 | IL23R IL2RA |
| IFNE | 2 | IL23R IL2RA |
| IFNK | 2 | IL23R IL2RA |
| IFNW1 | 2 | IL23R IL2RA |
| TPO | 2 | IL23R IL2RA |
| TSLP | 2 | IL23R IL2RA |
| IL15 | 2 | IL23R IL2RA |
| IL9 | 2 | IL23R IL2RA |
| IL5 | 2 | IL23R IL2RA |
| IL23A | 2 | IL23R IL2RA |
| IL12A | 2 | IL23R IL2RA |
| IL13 | 2 | IL23R IL2RA |
| LEP | 2 | IL23R IL2RA |
| CTF1 | 2 | IL23R IL2RA |
| CLCF1 | 2 | IL23R IL2RA |
| CNTF | 2 | IL23R IL2RA |
| IL11 | 2 | IL23R IL2RA |
| EPAS1 | 2 | PDGFB CUL2 |
| HIF1A | 2 | PDGFB CUL2 |
| CREBBP | 2 | SMAD3 STAT3 |
| EP300 | 2 | SMAD3 STAT3 |
| CCR2 | 2 | CCL2 CCL7 |
| PTPN11 | 2 | JAK2 TYK2 |
| STAT1 | 2 | JAK2 TYK2 |
| STAT2 | 2 | JAK2 TYK2 |
| STAT4 | 2 | JAK2 TYK2 |
| STAT5A | 2 | JAK2 TYK2 |
| STAT5B | 2 | JAK2 TYK2 |
| STAM2 | 2 | JAK2 TYK2 |
| STAM | 2 | JAK2 TYK2 |
| MYC | 2 | CEBPA STAT3 |
| RFX5 | 2 | HLA-DQA2 HLA-DQB1 |
| RFXAP | 2 | HLA-DQA2 HLA-DQB1 |
| RFXANK | 2 | HLA-DQA2 HLA-DQB1 |
| CREB1 | 2 | HLA-DQA2 HLA-DQB1 |
| NFYA | 2 | HLA-DQA2 HLA-DQB1 |
| NFYB | 2 | HLA-DQA2 HLA-DQB1 |
| NFYC | 2 | HLA-DQA2 HLA-DQB1 |
| RIPK2 | 1 | NOD2 |
| SYK | 1 | CARD9 |
| KIT | 1 | PDGFB |
| CSF1R | 1 | PDGFB |
| EGFR | 1 | PDGFB |
| MET | 1 | PDGFB |
| FLT4 | 1 | PDGFB |
| KDR | 1 | PDGFB |
| FLT1 | 1 | PDGFB |
| PDGFRB | 1 | PDGFB |
| PDGFRA | 1 | PDGFB |
| NGFR | 1 | PDGFB |
| IGF1R | 1 | PDGFB |
| FGFR1 | 1 | PDGFB |
| FGFR3 | 1 | PDGFB |
| FGFR2 | 1 | PDGFB |
| EPHA2 | 1 | PDGFB |
| FGFR4 | 1 | PDGFB |
| INSR | 1 | PDGFB |
| TEK | 1 | PDGFB |
| TGFBR1 | 1 | SMAD3 |
| TGFBR2 | 1 | SMAD3 |

| Candidate gene | n° of GWAS interactors | GWAS interactors |
|---|---|---|
| TGFB3 | 1 | SMAD3 |
| TGFB2 | 1 | SMAD3 |
| TGFB1 | 1 | SMAD3 |
| MAPK1 | 1 | IL12B |
| MAPK3 | 1 | IL12B |
| MAPK8 | 1 | IL12B |
| MAPK9 | 1 | IL12B |
| MAPK10 | 1 | IL12B |
| MAPK14 | 1 | IL12B |
| MAPK11 | 1 | IL12B |
| MAPK13 | 1 | IL12B |
| MAPK12 | 1 | IL12B |
| FCGR2B | 1 | IL12B |
| IL18 | 1 | IL18RAP |
| JUN | 1 | IL18RAP |
| FIGF | 1 | STAT3 |
| VEGFC | 1 | STAT3 |
| VEGFA | 1 | STAT3 |
| VEGFB | 1 | STAT3 |
| CCND1 | 1 | STAT3 |
| CCND2 | 1 | STAT3 |
| CCND3 | 1 | STAT3 |
| CISH | 1 | STAT3 |
| PIM1 | 1 | STAT3 |
| IRF9 | 1 | STAT3 |
| PIAS3 | 1 | STAT3 |
| PIAS4 | 1 | STAT3 |
| PIAS1 | 1 | STAT3 |
| PIAS2 | 1 | STAT3 |
| RORC | 1 | STAT3 |
| RORA | 1 | STAT3 |
| PGF | 1 | STAT3 |
| FAS | 1 | FASLG |
| TNFRSF6B | 1 | FASLG |
| FOXO3 | 1 | FASLG |
| NOS2 | 1 | IL10 |
| PLA2R1 | 1 | IL10 |
| MRC1 | 1 | IL10 |
| MRC2 | 1 | IL10 |
| CLEC4M | 1 | IL10 |
| CD209 | 1 | IL10 |
| CCR1 | 1 | CCL7 |
| TNFRSF11A | 1 | TNFSF11 |
| TNFRSF11B | 1 | TNFSF11 |
| TNFRSF18 | 1 | TNFSF18 |
| E2F1 | 1 | CEBPA |
| E2F2 | 1 | CEBPA |
| E2F3 | 1 | CEBPA |
| RUNX1 | 1 | CEBPA |
| RUNX1T1 | 1 | CEBPA |
| SLC2A1 | 1 | SMAD3 |
| SMAD4 | 1 | SMAD3 |

**Table 23. Candidate genes derived from the analysis of the PPI network.** Direct interactors of GWAS genes in the KEGG network have been considered like candidate genes involved in Crohn's disease onset. For each candidate gene, the GWAS interactors are reported.

### *2.6.3 Matching of patients phenotype*

After the application of our 2 steps-protocol of variants reduction, we moved to the definition of disease risk scores for the 111 individuals. As already introduced, our working hypothesis for this edition of the Crohn's disease challenge was based on the following assumption: the higher is the mutation burden for Crohn's related variants, the higher should be the probability for an individual to be affected. This hypothesis was already partially confirmed by the presence of a moderate negative correlation between mutation burden for VEO SNPs and onset age, in the CAGI 3 dataset. A maximum of 6 prediction could be submitted to the assessors. For this reason, we decided to slightly differentiate our predictions considering different sets of weight for each category of variants. In particular, in the first three submissions, higher scores have been assigned to VEO SNPs respect to mutations present in the other lists of variants. In addition, scores for heterozygous and homozygous variants have been differentiated trying to maximize prediction performance in the CAGI 3 dataset. In the second set of three predictions instead, weights assigned to the three set of variants were more homogeneous. Also in these last three submissions, scores for heterozygous and homozygous variants have been differentiated maximizing prediction performance in the CAGI 3 dataset.

# 3 Performance assessment and conclusions

## 3.1 Performance assessment and comparison with other participants

Performance assessment of methods predicting complex phenotypes is a rather complex task. Simple accuracy, calculated by setting a fixed threshold for prediction (e.g. at 0.5), neither supports the goals of CAGI nor is it representative of a clinically relevant scenario. Datasets from CAGI are derived from case-control studies, as well as pedigree studies in families with a strong history of disease, and so, not representing a random sampling of the population. The definition of a fixed threshold for evaluation, and the report of a basic accuracy score in such datasets, would obscure results interpretation. The use of Receiver Operator Characteristics (ROC) curves for genomic test evaluation has been investigated in several studies[168]. ROC offers many advantages for the evaluation of clinical tests. The shape of a ROC curve can help differentiate between highly sensitive tests, which could

rule in a possible diagnosis, and highly specific tests. The prediction of Crohn's disease status from sequencing data might be used in either of those situations depending on clinical manifestations, risk factors, or stage of patient evaluation. Additionally, ROC curves allow easy selection of a classification threshold (based on selecting a position on the curve).

Assessors evaluated also the robustness of prediction accuracy making predictions on different subsamples of exomes and assessed the confidence intervals reported by the participants. To capture confidence intervals on the predictions, multiple samples with replacement were selected. Each prediction was then modified by adding a random amount taken from a normal distribution with a mean of zero and a standard deviation equivalent to the standard deviation reported for the original prediction. The average area under the ROC curves from the bootstrap sampling was used, accompanied by the bootstrapped confidence interval, to estimate the robustness of differences between prediction performances.

Finally a cross-validated logistic regression based meta-classifier was trained on the submissions for this challenge. This step allowed the assessor to identify whether combining the features selected across the different groups would improve prediction performance over a single method.

As expected performance of prediction methods dropped from previous CAGI iterations with an Area Under the Curve (AUC) of 0.72 for the best performer[169]. The top approach used a compiled list of genes associated with Crohn's disease. Imputation was used to evaluate risk contribution from known regions associated with Crohn's disease but not covered by exome sequencing. Finally the Welcome Trust Case Control Consortium (WTCCC) Crohn's disease genotyping array data were used to train the disease classifier.

Regarding our submissions, overall good results have been achieved, with a maximum AUC of 0.609 (See Figure 19). Considering this result we ranked fourth among all the groups, confirming the good track record of our laboratory in this challenge. Definitely, an evident reduction of performance characterized our participation in this last edition of the Crohn's disease challenge. In addition, considering the range of AUCs achieved by our submissions ($AUC_{min}$ 0.587 - $AUC_{max}$ 0.609), seems clear that the strategy used to differentiate our predictions has to be completely redefined.

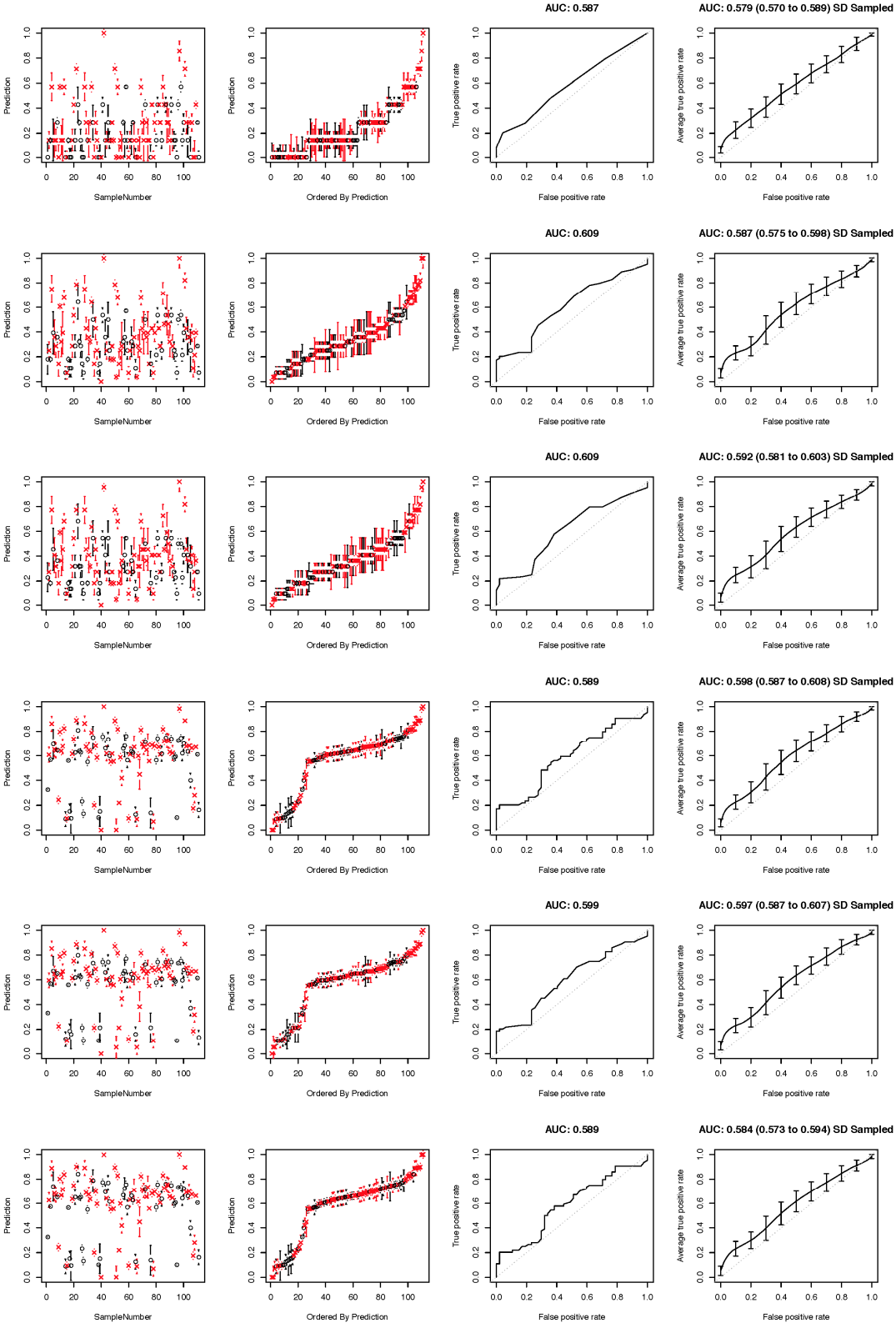
AUC: 0.587
AUC: 0.579 (0.570 to 0.589) SD Sampled
AUC: 0.609
AUC: 0.587 (0.575 to 0.598) SD Sampled
AUC: 0.609
AUC: 0.592 (0.581 to 0.603) SD Sampled
AUC: 0.589
AUC: 0.598 (0.587 to 0.608) SD Sampled
AUC: 0.599
AUC: 0.597 (0.587 to 0.607) SD Sampled
AUC: 0.589
AUC: 0.584 (0.573 to 0.594) SD Sampled
Prediction
SampleNumber
Ordered By Prediction
True positive rate
Average true positive rate
False positive rate

**Figure 19. Performance assessment of CAGI 4 Crohn's disease challenge.** The first column shows a scatterplot of predictions, with the sample number as the abscissa and the predicted value as the ordinate. Affected samples are shown in red, health individuals are black. The second column shows the same thing as the first, but with the predictions rank ordered. The third column shows the ROC curve for the predictions, and AUC is reported above the plot. The fourth column shows the ROC again, but with confidence intervals derived from bootstrap sampling with replacement over the submitted values. Modified from[169].

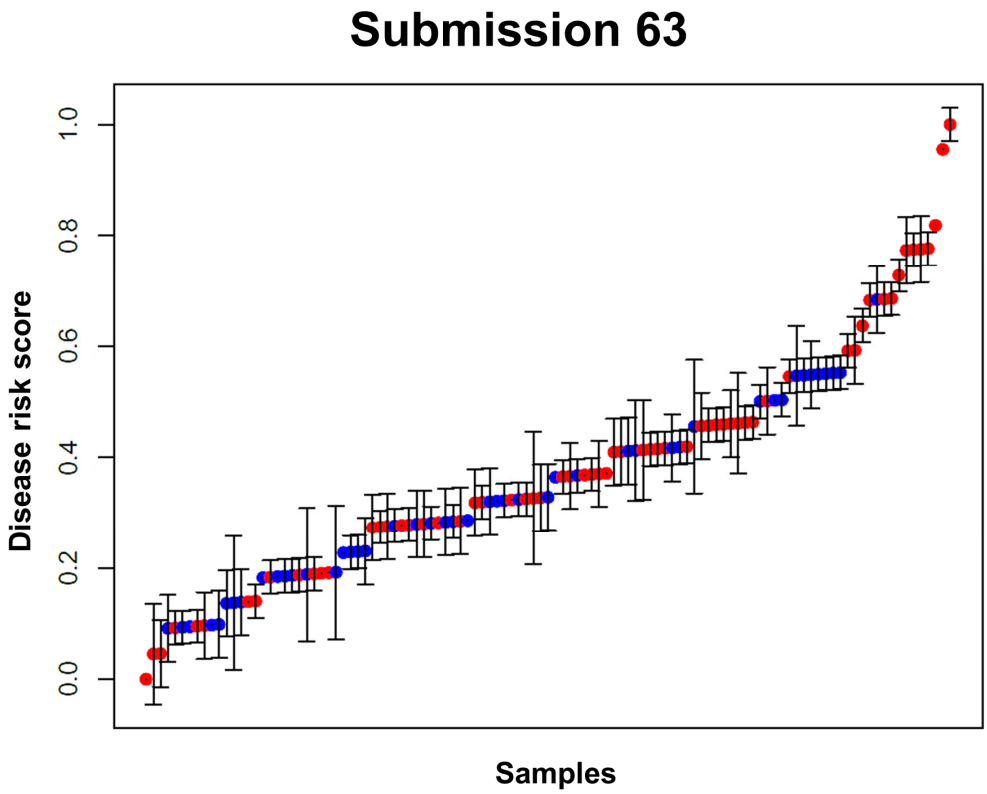


**Figure 20. Matched and mismatched individuals, sorted by disease risk score.** Submission 63, samples are ordered by disease risk score (magnification from Figure CR9). Affected individuals are marked in red, healthy samples are blue. As it possible to see on the right side, all patients with an extreme score are effectively affected.

However, analyzing our best submission, an interesting pattern could be identified. Focusing on the right tail of the risk scores distribution, all the patients with an extremely high risk score, were effectively affected by Crohn's disease (See Figure 20), suggesting that the right signal was caught by our method. This long tail of 14 matched individuals

(with only one exception) than reached a plateau, interrupted by the presence of a block of 10 mismatched individuals, and finally it starts again with another series of 8 matched affected individuals. Interesting is also the fact that all the 10 mismatched individuals present substantially the same disease risk score, underlining the fact that probably all these individuals present a shared pattern of mutations.

The understanding of this phenomenon could be useful to improve our prediction performance in the next edition of the Crohn's disease challenge. For this reason an *a posteriori* investigation of variants carried by these individuals was performed. Our attention was particularly focused on VEO SNPs (See Figure 21). Surprisingly these individuals presented a considerable number of variants involved in the precocious onset of the pathology. Almost all these healthy individuals, presented one or more heterozygous variants affecting the NOD2 gene, together with at least one homozygous mutation in one of the two subunits of the interleukin 10 receptor (IL10RA, IL10RB) genes or in the ATG16L1 gene. Particularly interesting are the S091, S087, S81 mismatched individuals, which in addition to the several heterozygous VEO variants in the NOD2 gene, presented an homozygous variants for each of the three other genes (IL10RA, IL10RB, ATG16L1).

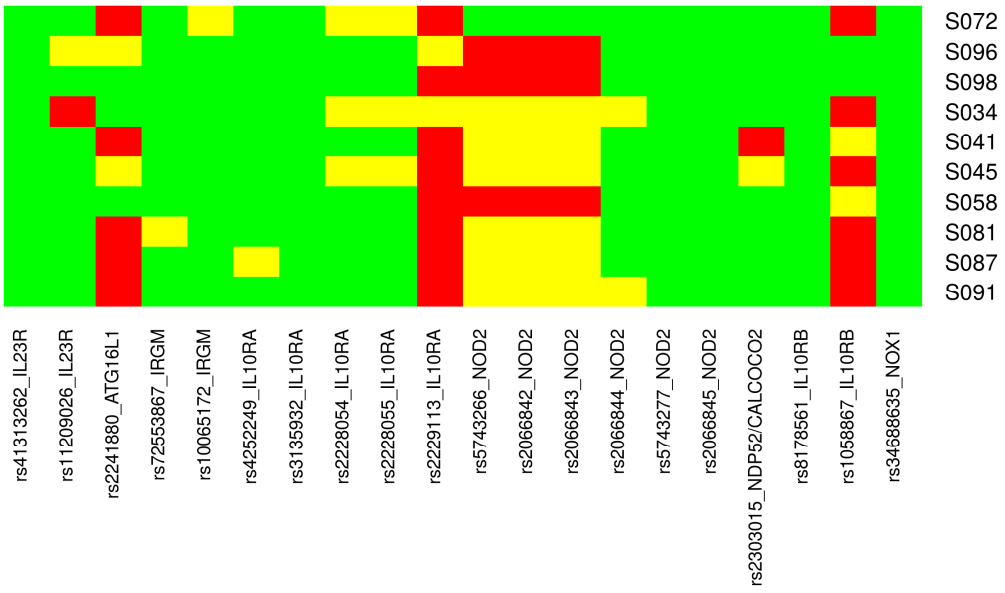


**Figure 21. Heatmap of VEO variants found in healthy individuals with high disease risk score.** VEO variants found in the exomes of healthy individuals with a

high predicted risk score. Homozygous variants are represented in red, in yellow are heterozygous mutations. In green are wild-type positions.

Several hypothesis could be formulated to explain this situation of healthy individuals with such a high mutation burden.

One first hypothesis is based on the analysis and interpretation of the genotype data. As already anticipated most of these healthy patients presented at least three heterozygous variants affecting the NOD2 gene. Unfortunately, as typical of NGS data, no phasing information could be inferred from the *.vcf* file. In this way, the identification of cases of composed heterozygosis is impossible.

It is possible in fact, that all of these variants could be inherited from one of the two parents, defining a situation where only one of the two copy of the gene is defective.

Another possible explanation for these mismatched prediction, is related with the interpretation of the clinical significance of the NOD2 variants. As already presented, these variants have been manually selected from papers investigating the genetic background of the VEO forms of Crohn's disease, focusing on case reports and WES studies. However, the investigation of the molecular background of these severe manifestations is still an open field of research. Several months after the conclusion of the Crohn's disease challenge, the NOD2 variants selected for prediction purpose, have been annotated like benign in the Clinvar[85] database. Some doubts could be raised for this classification as it seems that the benign tag for these variants comes from a kind of automatic annotation based on variant features like, the presence in other databases, results of pathogenicity predictors and frequency in the population. Despite these considerations on the annotation methods, seems clear than the definition of variants useful for prediction purpose is still far away from being a solved problem.

Other considerations cold be raised to explain the presence of such row of healthy individuals with a high risk score. From the analysis of Figure 21 it is clear that these patients present a compromised genetic background for genes involved in the onset and regulation of the inflammatory process. Unfortunately, no phenotypic information, others than disease status for Crohn's disease, have been released by data provider. This lack of complete phenotype description is a crucial limit of CAGI challenges where complex phenotype have to be predicted. It is well established in fact, that for most complex traits, the phenotype distribution could be described by a Gaussian curve. In this way the

classification of individuals in only two classes, healthy - affected, could not be sufficient to describe the spectrum of possible clinical manifestations. In our case, interesting would be the possibility to know if the healthy individuals with a predicted high risk score, are in reality affected by other auto inflammatory diseases like one among the typical extra intestinal manifestations associated with Crohn's disease. Another interesting information that could help to better explain these mismatched predictions, is the presence of family within the dataset. It is possible in fact, that these individuals could be parents of affected individuals. Unfortunately, without a deeper phenotypic characterization of cases and controls, none of the above hypothesis can be confirmed.

Regarding other groups, a plethora of different methods have been used to perform predictions. Several strategies have been used for variants selection, highlighting the many different approaches to building a Crohn's disease classifier. Similar to our approach, many groups used variants previously found to be associated in GWAS studies; the NHGRI catalog was a popular choice to identify these associated variants[170]. Other approaches relied on gene lists of associated and "predicted" Crohn's disease genes. Other examples include the usage of tools like Phenolyzer[171], the creation of gene lists based on GO pathways enriched with Crohn disease associated variants and the use of natural language processing to identify genes of interest from Pubmed abstracts[171,172]. In addition, some groups used population level frequency data to help distinguish variants more likely to be pathogenic. Other methods relied on pathogenicity prediction tools such as SNAP, PON-P2, SNPs&GO, and Variant Effect Predictor to inform variant selection and weighting [173–176] . A range of machine learning approaches were used to build the classifiers, such as naïve Bayes, logistic regression, neural networks, and random forests. Additionally, some groups purposed metaclassifiers based on combinations of methods used in previous editions of these challenge.

Finally, a metaclassifier was created by the assessors using all the submitted predictions for this challenge. The combination of features selected by predictors produced an improvement of prediction performance over the top method, with AUC of 0.78 achieved by the metaclassifier[169].

## 3.2 Conclusions

Disease risk prediction from genetic data is still one of the biggest missed promises of the NGS-era. The Critical Assessment of Genome Interpretation is a community experiment, which aims to define the state of the art for phenotype prediction through a series of challenges with data coming from unpublished WES and targeted enrichment sequencing studies. For CAGI 4 (2016), three challenges proposed the use of exome sequencing data: warfarin dosing, bipolar disorder and Crohn's disease. The Crohn's disease challenge has been part of other two previous iterations of the CAGI experiment, giving the possibility to assess the evolution of methods predicting complex phenotypic traits over the last six years. The several iterations of this challenge gave also the possibility to understand how caveats in genetic data generation and processing could affect prediction performance. In addition, challenge assessments raised questions about how to correctly evaluate methods predicting disease risk score from genotype data. For these reasons, the Crohn's disease challenge represents the best story of success of the CAGI experiment, pushing forward the field of genomic interpretation.

Our group participated to all the previous editions of these challenge, always achieving remarkable results. Best results have been obtained using different implementations of agglomerative clustering, which allowed us to discover hidden relationships between the unknown individuals.

Unfortunately, both in CAGI 2 and CAGI 3 evident bias in sample distribution have been found by challenge assessors. In the first iteration of this challenge a strong batch effect was discovered as consequence of sample preparation and sequencing. In the second edition, a great effort was made to avoid bias in data processing. In this case, samples were collected from German families characterized by strong history of Crohn's disease. Additional healthy controls were selected always among the German population, in order to avoid the presence of population structure in the dataset. Unfortunately, even in this case, challenge assessors revealed that affected individuals from different families clustered much more closely with other affected patients. In both the iterations of this challenge, predictors were not aware that such bias in the datasets could be used to achieve great prediction performance. Nevertheless it is not unlikely that several methods could have accidentally exploited such dataset structure, achieving great results.

For the first and second iteration of the Crohn's disease challenge, amazing prediction performance have been recorded: AUC 0.92 and AUC 0.87 respectively.
Nevertheless it is more likely that the problem of predicting Crohn's disease status from exome data was far away from being solved. Last iteration of the Crohn's disease challenge has been a kind of testing ground to understand if great results achieved in 2011 and 2013 represent the real capacity of differentiate healthy and affected individuals. The 2016 dataset was composed by 111 individuals with 64 Crohn's disease patients and 47 healthy controls, all taken from the German population.
Our investigation started with a clustering analysis aimed to detect the presence of possible bias in the dataset. An analysis of the distribution of non sequenced variants was even performed as this feature was helpful to reveal the presence of batch effects in previous datasets. These analyses revealed that in the 2016 dataset, clear separation of cases and controls based on genetic structure was not present, suggesting that problems with batch effects and sampling bias were no longer present. Dealing with a such unbiased dataset required a complete redefinition of our prediction strategy. Our old approach, based on the identification of hidden relationship between samples in fact, would be less effective to identify affected individuals in a homogeneous dataset. The new strategy defined to approach the 2016 edition of this challenge was based on the hypothesis that individuals with the highest mutation burden, for variants involved in Crohn's disease, would be the one with the highest probability to be affected. On the base of this working hypothesis, a two step prediction approach has been defined. First, a great effort was made to define variants that are more likely to increase probability to be affected by Crohn's disease. To this aim, all common and synonymous variants have been sifted out from our analysis, considering this variants less likely to be involved in the onset of a pathogenic phenotype. We than focus our attention on the definition of lists of variants involved in the onset of the pathology. To this aim three list of variants have been defined. A manual curation of the academic literature lead to the definition of a short list of mutations associated with the very early onset forms of the disease. A second list was defined considering genes associated to the pathology by means of GWAS studies. In the end, trying to address the problem of missing heritability, we expanded the list of GWAS genes exploiting information extracted from the PPI network.
Interesting is to highlight the wide variety of methods used by other groups trying to identify variants useful for prediction purpose. Similar to our approach, many groups

used variants associated to the pathology by means of GWAS studies, other group created gene lists based on GO pathways, enriched with Crohn disease associated variants. Some other groups used variants frequency to identify pathogenic variants or pathogenicity prediction tools such as SNAP, PON-P2, SNPs&GO, and Variant Effect Predictor to guide variant selection and weighting. Considering the wide spectrum of methods used by the different group, is clear that the definition of variant involved in the disease is a problem that is still far away from being solved. This problem was even more clear considering that the clinical significance of some variants associated with the VEO forms of the pathology, in less than one year from the prediction season, changed their status to benign variants associated to the pathology. From this situation is clear that the great effort provided in the classification of variants collected in databases like Clinvar, HGMD[85,177], has to be further improved to allow the achievement of better prediction performance.

The second step of our prediction strategy was the scoring of variants present in the exomes, to define a personalized disease score. To this aim, critical was the usage of genetic data coming from the previous editions as training set. Unfortunately, as already anticipated, the 2011 dataset was useless for this purpose as strong batch effects could affect the training of our method. In addition, population structure was present also in the 2013 dataset. In this context we decided to avoid usage of machine learning or SVM algorithms, being aware that these methods could learn more about dataset bias than useful differences between healthy and affected individuals. It was clear that the 2016 editions of Crohn disease challenge suffers the absence of a good training set. This weakness is even more clear considering that the top performing group exploited a different dataset coming from the WTCCC project, to train its method.

However, considering the elimination of biases in the 2016 dataset, this incarnation of the Crohn's disease challenge is likely to be the best reflection of how the prediction methods perform. In this edition an expected drop of prediction performance has been registered, with the best performing method achieving an AUC of 0.72. Good results have been even achieved by our method with an AUC of 0.609, registering anyway a strong performance reduction respect to previous editions. Interesting of our best performing submission is the fact the all the 14 individuals with the highest risk disease score were effectively affected by Crohn's disease. A series of 10 healthy individuals with mostly the same risk disease score than interrupt the row of right predictions. An investigation of

the genetic background of these individuals revealed that most of them present several mutations associated to the VEO forms of the pathology. A further phenotypic characterization would be useful to understand reasons behind these mismatched predictions.

Despite limitations, the three iterations of the Crohn's disease challenge, has been the biggest story of success of the CAGI experiment so far. Not so much for performance achieved by predictors, but for lessons learned both by predictors, assessors and data providers, reflecting the real purpose of this community experiment. Even if the road to translate phenotype prediction of complex traits in clinical practice is still long, seems clear that we are moving in the right direction.

# BOOGIE 2: Predict blood groups from high throughput sequencing data

## 1 Introduction

Blood groups are a classification of blood based on the presence or absence of inherited antigenic substances on the surface of Red Blood Cells (RBCs) and on the presence antibodies freely running into the blood flow. Before the 1900s, it was thought that all blood was the same, a misunderstanding that led to frequently fatal transfusions of blood between people and hazardous transfusions of animal blood into humans. Over the time, our understanding on blood groups has evolved to encompass not only transfusion-related problems but also specific disease association with RBC surface antigens. Blood group antigens are either sugars or proteins, and in most of the cases they are attached to various components of the RBCs membrane. Antigens of the ABO blood group for example are sugars. They are produced by a series of reactions in which enzymes catalyze the transfer of sugar units. In this case, person's DNA determines the type of enzymes they have, and, therefore, the type of sugar antigens that end up on their red blood cells. In contrast, the antigens of the Rh blood group are proteins. A person's DNA holds the information for producing the protein antigens. The RhD gene encodes the D antigen, which is a large protein on the RBCs membrane. Some people have a version of the gene that does not produce D antigen, and therefore the RhD protein is absent from their red RBCs. Blood phenotype is long known to be dependent on the sole genotype. Inheritance is mostly Mendelian with only few exceptionalities reported so far[178]. Other than the most studied “major blood groups” ABO and RhD, at present more than thirty blood groups have been documented in literature[179]. Like for ABO and RhD, these “minor blood groups” could be defined by their carbohydrate structures on RBCs (H, P1Pk, I, GLOB); 23 are characterized by the protein sequence of the RBCs membrane protein, while two are obtained from the plasma (LE, CH/RG). Regarding the function of the proteins

responsible for blood group definition, some are expressed at higher levels and function as membrane transporters, whereas the functional importance of other antigens has not been well defined yet. Proposed functions of other antigens are: receptors involved in ligand signaling, enzymes, and protein/carbohydrates involved in glycocalyx formation[180] (See Table 24).

| Blood group system | Protein products | Functions |
|---|---|---|
| DI | Band-3 | Anion transport |
| MNS | Gliycophorin A<br>Gliycophorin B | Facilitates membrane assebly of band-3 |
| Rh | D polypeptide<br>CE polypeptide | Facilitates band-3/RhAG complex assembly |
| RhAG | Rh associated glycoprotein | Neutral gas transport |
| GE | Gliycophorin C<br>Gliycophorin D | Maintains red cell shape through interaction with protein 4.1 |
| CO | Aquaporin-1 | Water/$CO_2$ transport |
| FY | Duffy glycoprotein (DARC) | Chemokine receptor for proinflammatory cytokines |
| Kell | Kell glycoprotein | Zinc endopeptidase |
| Jk | Urea transporter | Urea transporter |
| Cromer | Decay accelerating factor (CD55) | C3 convertase inhibitor |
| LU | Lutheran glycoprotein | Ligand for laminin 511/512 |
| XK | Kx glycoprotein | Amino acid tramsport |

**Table 24. Summary of blood systems with protein determinants and relative function.** Table modified from[180].

As anticipated, blood groups typing is crucial in transfusional medicine. In cases where patients undergo to incompatible blood transfusion a massive activation of the immune and clotting system can cause shock, kidney failure, circulatory collapse even leading to death. To avoid these kind of adverse reactions, blood compatibility test are always performed before transfusion. Both investigations at phenotype level and genotype level could be performed to test blood compatibility in clinical practice. Investigation of blood

system phenotype is routinely performed by mean of antiglobulin test (Coombs test), direct and indirect[181]. First phase of serologic test is the direct antiglobulin test where monoclonal antihuman globulin are used to detect either erythrocyte-directed IgG in plasma or IgG or complement coating on the surface of circulating erythrocytes. If agglutination occurs, the screen is considered positive. Indirect antiglobulin test instead, is used by the blood bank to detect unexpected erythrocyte antibodies in the patient's serum or plasma[181]. The indirect test is the second and final phase of the antibody screen and serologic crossmatch procedures. In an indirect antibody screen the recipient's serum is incubated with 2 or 3 different type O erythrocytes that express clinically significant antigens. Erythrocyte and serum mixture is then incubated with anti-IgG monoclonal antihuman globulin and observed for agglutination. Again, if agglutination occurs, the screen is considered positive[181]. Antiglobulin screening normally requires no more than a dozen of minutes to be performed so, blood compatibility could be guaranteed in emergency. Even thanks to its cost-effectiveness (approximately a dozen of US dollars) antiglobuline test is the state-of-the-art procedure for blood compatibility investigation. However, accuracy of Coombs test could be reduced in several cases. Many pharmacologic agents have been associated with positive direct antiglobulin test result, even after the drug has been cleared[181]. Many drugs indeed bind to the membranes of circulating cells. Antibodies elicited by these medications may be directed either against a combination of the drug and certain membrane components or against epitopes of the drug molecule that are bound tightly to the erythrocyte surface. Other conditions of reduced accuracy for antiglobulin test are cases where patients have had recent transfusions[182], in this condition often unreliable results can occur. In particular if a patient has received a transfusion within the previous 3 months, a positive direct antiglobuline test may indicate alloimmunization to an antigen on the transfused cells that is not present on the recipient's own erythrocytes[181]. Cases of unreliable results to antiglobuline test con also occur with the presence of certain variant antigens that can cause false-negative reactions. This conditions are often related to the presence of specific minor blood groups (e.g. Duffy) or, so-called “weak blood groups” (that present weaker reactions to antiglobuline test respect to the population average), leading to inaccurate results[183]. Although for most of the patients minor blood or weak incompatibles are substantially armless, in some particular conditions these typing mismatches could be critical. Particular attention has to be a paid in case of sickle-cell and

Mediterranean anemia, thalassemia and cancer, where blood mismatch could have severe consequences[183]. Other critical cases are the transfusion dependent patients where a fine blood matching could extend period between transfusion, reduce hospitalization time, improve life quality and even increase life expectancy[183]. In this particular conditions an investigation of blood type even at a genotype level is needed. At present some solutions involving multiplex-PCR combined with flow cytometry already exists in the markets[184] but these techniques are over an order of magnitude more expensive than serological test. As extensively described in this manuscript, the advent of NGS technology have led to a widespread availability of sequencing data thanks exponential reduction in sequencing cost and time. In such context were the possibility of having children sequenced at birth[185] is closer than ever, blood groups typing from NGS data seems to be an appealing aim for such a genotype dependent trait. While trying to address this challenge, several problems has to be considered. First, for most of the blood groups, deep knowledge of genetic background reveals complicated relationship that links genotypes and blood types. Rh system is just an example with good genotype knowledge and a complicated basis, since it is encoded by two different genes resulting in the two proteins RhD and RhCE[186]. The former is the determinant of the most common Rh antigen while the latter is responsible for a large part of weak Rh phenotypes. To further complicate this scenario, more than 40 Rh antigens are known so far. Situations like this are common, even for other blood systems (See Table 25).

| Name | Symbol | Number of antigens | Gene name |
|---|---|---|---|
| ABO | ABO | 4 | ABO |
| MNS | MNS | 43 | GYPA,GYPB,GYPE |
| P | P1 | 1 | P1 |
| Rhesus | Rh | 49 | RhD,RhCE |
| Lutheran | LU | 20 | LU |
| Kell | KEL | 25 | KEL |
| Lewis | LE | 6 | FUT3 |

| | | | |
|---|---|---|---|
| Duffy | FY | 6 | FY |
| Kidd | Jk | 3 | SLC14A1 |

**Table 25. Summary of blood systems with specified the number of antigens.** It is possible to see that for most of the blood systems, prediction of phenotypes is complicated by the high number of different antigens coded by blood system genes. Table modified from[180].

A second level of complexity has also to be considered as the one mutation-one phenotype paradigm is clearly unable to explain many traits of clinical relevance like blood systems. For these kind of phenotypes in fact multiple co-occurring variants are involved in the definition of the trait[187]. The situation is finally complicated by heterozygous variants. For these kind of variants, inference of the correct phased haplotype is required to identify if alleles are co-located on the same chromosome or not[188] (so-called "phasing problem"). Genetic data coming from NGS platforms generally take the form of unphased genotypes in which is not possible to define on which of the two chromosomes, or haplotypes, a particular allele falls on.

Despite these complications, in 2015 we published the first version of BOOGIE, a Java tool to predict blood group phenotypes from NGS data. After more than two years, this tool need to be updated and upgraded to face several issue identified thanks to the increasing knowledge about genetic determination of blood phenotypes. First improvements regard the BOOGIE algorithm in case of complex heterozygous conditions, leading to the achievement of a less CPU time-consuming prediction phase. This is a critical improvement in cases where predictor runs on a local machine or predictions over a large set of patients are requested. Another algorithm upgrade is the capability to manage the presence of single variants with complete penetrance. The absence of this specific feature was one of the main reasons causing several mismatch in predictions for the first version of the tool. Other updates regard the possibility to perform prediction on patients sequenced with the last version of the human reference genome (hg38) and the update of genotype data for blood group prediction. In addition, the possibility to use the standard Variant Call Format (*.vcf*) like input for patients genetic data was introduced. Last, predictions for mutations with damaging phenotypic effects are now reported in BOOGIE output, providing a possible biological explanation for cases where mismatch between real and predicted phenotypes occurs. Performance against the first version of

the tool were tested for ABO and Rh blood groups, achieving slightly better performance for the former while same performance were recorded for the latter.

# 2 Materials and methods

From a conceptual point of view, the BOOGIE tool could be divided in two main components: knowledge about blood groups and the algorithm that perform prediction. Knowledge about blood groups genotype-phenotype relationships comes mainly from the BGMUT database[189], while predictions are performed by the so-called "BOOGIE prediction framework".

## 2.1 The BGMUT database

BGMUT was developed in 1999 as a locus-specific gene mutation database for blood groups. As of March 2014, 1,545 alleles representing 34 different blood systems were present with all except one recognized by the International Society Blood Transfusion (ISBT)[190]. Alleles in BGMUT are grouped by blood group system. For each allele in the database, BGMUT provides details on the nucleotide variants and the deduced amino acid changes in the protein encoded by the gene the allele belongs to. Obliviously the associated blood group phenotype is reported too. All loci of interest for genome classification have been grouped in haplotype tables for each blood group. These tables define all expected SNVs that should be observed for determination of a specific blood group allele (See Table 26). Whenever BGMUT reports no data about a SNV, that specific position is assumed to be the same of the reference hg19. Only exonic mutations were used, as we assumed that these are sufficient to cover the largest part of crucial positions involved in the definition of blood phenotypes. In the new version of the tool, haplotype tables have been updated introducing all the haplotypes uploaded in the database during the last three years. In this way, knowledge in particular for ABO and RhD systems has been strongly improved. Another huge contribution in increasing the knowledge about blood haplotypes comes from a recent paper[191] where more than 110,000 individuals of German origin have been sequenced for the exon 6 and 7 of the ABO gene. In the work of Lange and colleagues, new 287 distinct so far not described alleles for the ABO system have been identified. To note that this new knowledge doesn't seems to be related to low frequency, rare alleles as for example one has been identified in more than 150 individuals. Even these new data have been added to the new set of BOOGIE haplotype

tables. Obtaining haplotype tables from such different sources required meticulous manual curation as many blood groups and traits assumed old reference genes. This is an important issue related to recent improvement of sequencing techniques, as they provided a number of different DNA sequence versions of increasing quality that cannot be easily combined. E.g. the ABO reference gene in BGMUT corresponds to the A group, while the reference gene in hg19 corresponds to the O group. This leads to a shift and re-labelling of most mutations. Another problem due to the presence of different releases of the human genome was that the first BOOGIE version was not able to manage genetic data of patients sequenced with the last version of the human genome hg38 (GRCh38). As haplotypes tables are reference system specific, in the new version of the tool we address this problem providing tables compiled for last two reference systems: hg19 (GRCh37), and hg38 (GRCh38). Coordinates conversion was performed using CrossMap[192].

| Haplotype | Chr9:136132908 | Chr9:136131650 | Chr9:136131414 |
|---|---|---|---|
| **A101** | GG | C | G |
| **A102** | GG | T | G |
| **O02** | G | C | G |
| **B101** | GG | C | A |

**Table 26. Sample of haplotype table for the ABO system.** For each possible haplotype, positions and variants that differs from the reference are reported.

## 2.2 The BOOGIE prediction framework

BOOGIE 2 is written in Java. It requires two input files (See Figure 22), the haplotype table for the phenotype of interest and the target genotype file, i.e. the patient genome. Among all variants contained in the genotype file only those matching the haplotype tables are considered for the prediction. Once variants have been selected, all the possible assignments to the two chromatids are enumerated. The phenotype of each permutation is predicted by means of the 1-nearest neighbour algorithm, and the corresponding score stored. The haplotype assignment with the highest score becomes the predicted phenotypes. Note that predictions are performed for each chromosome, so it is up to the

user to determine the final phenotype. E.g., if the two copy of the ABO gene code for the A and 0 alleles respectively, the user should infer that the A group will be the resulting blood phenotype of the individual. In the second version of the predictor two main improvements were introduced. First, the possibility to manage standard ".vcf" files, in order to make possible to run the predictor directly from the output of NGS experiments. The second critical improvement regards the algorithm. In particular a sensible reduction of computing time for complex cases of heterozygosis without a reduction of prediction accuracy was achieved. This improvement of efficiency is crucial and leaves the door open to run BOOGIE on a local PC without the need of high performance computing units.

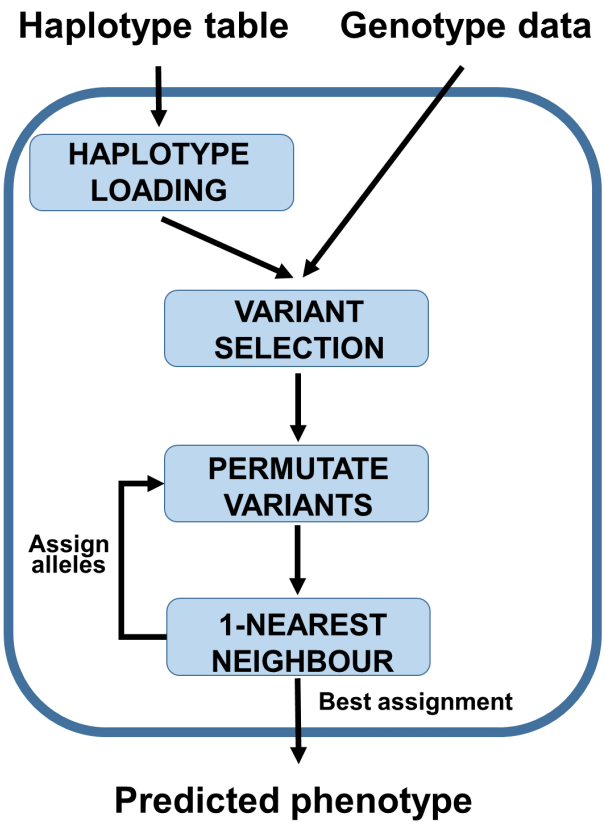


**Figure 22. Schematic BOOGIE overview.** Two files are required for prediction: the haplotype table and target genotype file. Variants in the genotype file are filtered and only the ones present in the haplotype table used for prediction. Phenotype is predicted by means of the 1-nearest neighbour algorithm, and the assignment with maximal score become the predicted phenotypes. Figure modified from[193].

## *2.2.1 Updated prediction system*

Prediction of phenotypes from genetic data is further complicated by the fact that NGS platforms usually provide unphased genotype data. With this kind of data is impossible to define on which of the two chromosomes a particular allele falls on in case of heterozygosity. This problem has been address in literature using information from HapMap[194] and expectation maximization approaches with Hidden Markov Models[195] but these methods are computationally expensive and usually cannot distinguish rare haplotypes or uncommon SNVs. Need of better computational performance in case of heterozygosity made the development of a new prediction algorithm necessary. To better

understand the problematic and the strategy used to address it, a comparison between the old and the new prediction system will be presented.

As already introduced, variants present in the target genotype are parsed and only the ones present in the haplotype tables will be used for prediction (See Figure 23 -1-). In BOOGIE 1 each haploid allele in the genotype file is scored singularly against each entry in the haplotype tables using as scoring function the inverse Hamming distance. The genetic status for every position is compared and a positional score is given: 1 for match, 0 for mismatch. Strong penalty is given to Indels accounting for frameshift variants. The overall score for each comparison can be expressed as the sum of positional scores, with the highest possible score equal to the number of considered positions. At the end of the process, an allele set will have a similarity score for each haplotype. The allele sets are rearranged for every possible combination of the allele status. The score is recalculated and only the highest combination score is taken for. Here emerges a critical issue of BOOGIE 1: the number of possible allele configurations is $2^{het}$ where *het* is the number of positions in heterozygosis. This cause an exponential growth in the number of comparisons, proportional to the number of heterozygous mutations. Such situation is critical especially in cases where the number of variants that differ from the reference is high, as for example RhD negative patients, leading to a huge increase of computational time to perform phenotype prediction.

BOOGIE 2 address this problem with an updated scoring algorithm and new heterozygous management. The new scoring function takes both allele sets at the same time as input, and compares them with a random couple of haplotypes (See Figure Figure 23 -2a-). The positional score then becomes: 2 if both variants in the allele set match the ones in the couple of haplotype, 1 if just one match, 0 if no matches are present. If a heterozygous variant is present, it is considered in the fittest position for scoring purposes (See Figure Figure 23 -2b-). The overall score is again the sum of positional scores but this time it represents the similarity of the couple of haplotypes with the best possible arrangement of the alleles. The comparison is than repeated for all the possible couple of haplotypes in the haplotype table and the couple with highest score is selected as prediction (See Figure Figure 23 -3-).

Concerning complexity, the new scoring function scales linearly with the length of the alleles array. Note that, in this case, time complexity does not depend on the number of

heterozygous positions in the input genotype file but is fixed, depending on the number of haplotypes present the haplotype tables. In this way, phasing problem is addressed with the same assumption of BOOGIE 1: the phase state of alleles is the one identical or most similar to a state already observed in literature (and therefore present in the haplotype table). The difference stands on the fact that in BOOGIE 2 the state of both alleles is assessed at the same time, while in BOOGIE 1 all the possible permutations of heterozygous alleles are score independently for both the alleles.

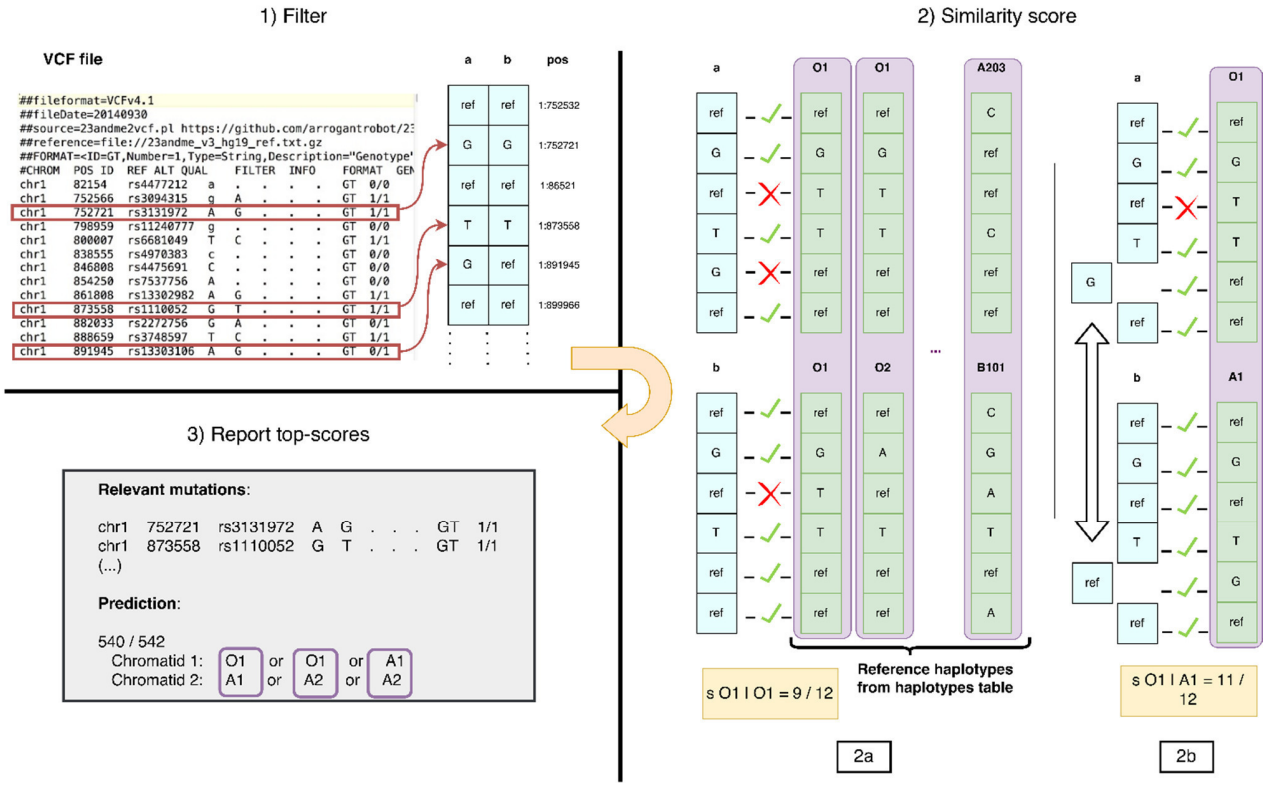


**Figure 23. BOOGIE 2 pipeline.** -1- Variants present in the target genotype filtered and only the ones present in the haplotype tables will be considered for prediction. -2a- The new scoring function takes both allele sets at the same time as input, and compares them with a random couple of haplotypes. -2b- If a heterozygous variant is present, it is considered in the fittest position for scoring purposes. The comparison is than repeated for all the possible couple of haplotypes in the haplotype table. -3- The couple of haplotypes from BGMUT with highest score is selected as prediction.

## 2.2.2 Dominant variants management

A further advancement has been introduced in the prediction system of BOOGIE 2. As it was reported in the BOOGIE 1 paper, some variants showed a greater effect on the resulting phenotype. In particular it was noticed that the ABO c.53G>T mutation had a complete penetrance for the determination of the O group[196]. Even if this position was reported in the ABO haplotype table, the BOOGIE 1 scoring system used the same weight for all SNVs leading to the definition of wrong phenotypes. As an example, for several profiles with the ABO c.53G>T variant, prediction of the A haplotype was supported by 12–14 variants, real phenotype was O instead. To address this problem, a polyphasic scoring system has been defined, prioritizing the mutations with a determinant effect. A list of such dominant variants has been manually curated. During the initial step of the prediction phase, the genomic file is parsed, looking for such dominant variants. In case of match, prediction is straightforward in favour of the corresponding dominant phenotype.

## 2.2.3 Variants of Unknown Significance management

An additional feature has been introduced in BOOGIE 2, the interpretation of Variants of Unknown Significance (VUS) that affect blood group genes. Rational for this improvement is the inability for BOOGIE 1 to consider variants not present in the BGMUT but having a crucial role in blood group definition. SIFT[197] and PolyPhen-2[198] are tools that predict the possible impact of an amino acid substitution on the protein's structure and function. SIFT is an application developed to predict the likelihood that a variant is damaging based on sequence conservation. PolyPhen-2 uses instead both sequence alignment and structural predictions, when available, and generates a final prediction based on an underlying machine-learning algorithm[199]. Considering predictions from both the tools we could expect to have rather good prediction performance, both for protein with known three-dimensional structure and also for the large number of proteins for which this information in not yet available. Both tools generate a score of deleteriousness that can be used to infer mutation effects.

Exonic sequence of all BGMUT-reported genes has been extracted from Ensembl's BioMart[200]. Then, every possible point mutation in each possible position has been

simulated and scored by means of the previously mentioned software. Results has been stored in a prediction file included in BOOGIE 2 package.

While reading a genomic file, all variants with the following features are selected:

1. they must not be included in the haplotype table
2. they must lie between the first and the last known mutation present in the haplotype table

Reason for the first point is straight forward, as we want to consider only non-literature-validated mutations that are therefore not present in the haplotype table. Rational for the second point explained by the assumption that a damaging mutation can more probably appears in the window between two observed critical variants. Of course, this is an oversimplification of the biological problem, which could underestimate the power of upstream and downstream variants. However for most of the of cases, intervals selected in this way contain most of the relevant part of the protein involved in blood group definition. Eventually it is important to underline that BOOGIE 2 predictions are not influenced by this additional procedure but a report of the unknown mutations found to be "deleterious" or "probably damaging" is given to the user as a warning. In this way, the predictions are still based on literature-validated data only but such warning could be useful to understand reasons of possible wrong predictions.

### *2.2.4 Test set*

Test set for BOOGIE 2 was doubled compared to the test set for the former version. New dataset is composed by 133 phenotyped genomes, gathered from the Personal Genome Project (PGP). PGP is a scientific project in which genomic information of willing candidates, along with their specific phenotypic data over a number of medical traits, are freely shared. Types of genomic assays and phenotypic information varies wildly among candidates.

For our research purposes, only candidates having blood serotyping and whole genome sequencing data (Complete Genomics data) available have been selected. Distribution of the blood types in this subgroup is represented in Figure 24.

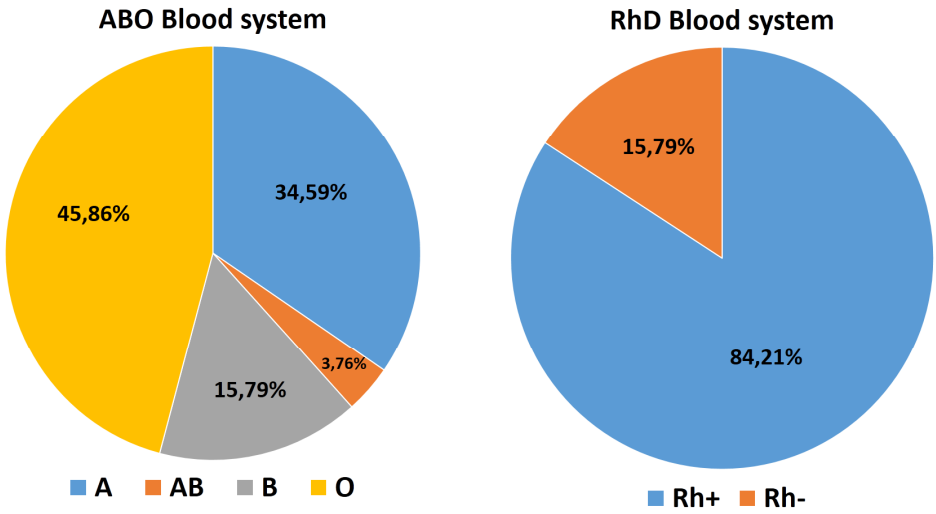


**Figure 24. Distribution of blood types for major blood groups in test set.** Biased in assessment should be mitigated by the presence of all phenotype classes.

A notable characteristic of this dataset is ancestry heterogeneity. Grandparent's country of origin for test set individuals is reported in Figure 25. Most of candidates have North American (non-native) ancestry, which is heterogeneous *per se*. Other samples comes from Europe, Asia and South America. Prediction in such dataset is more challenging due to the differences in the genomic background and geographic blood types distribution[201]. Data gathering was complicated by the fact that some of the phenotypes could be self-assessed by the patient as questionnaire. This leads to a degree of uncertainty for the data where this table is the only source of information regarding a patient blood group. E.G. for one sample (patient huC92BC9), the self-assessed blood group was in contrast with the blood group assessed by serological test.

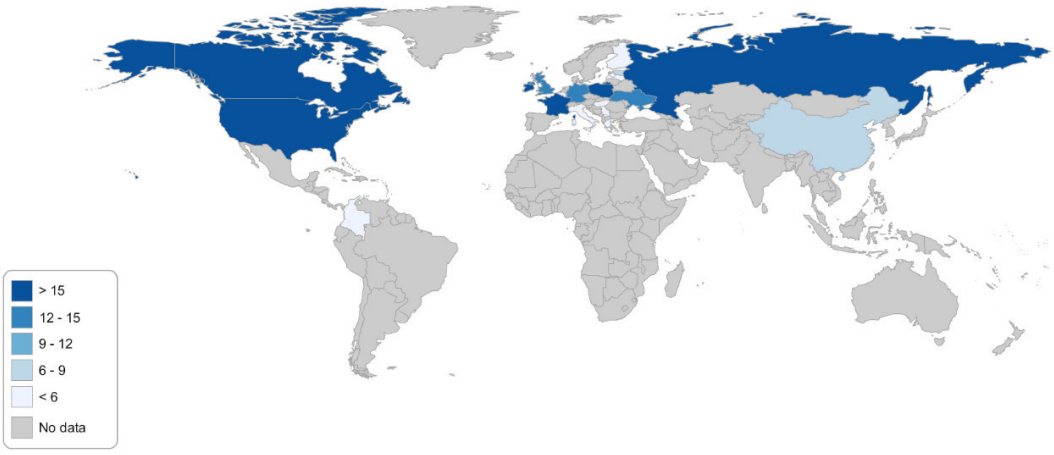


**Figure 25. Ancestry of individuals from the PGP dataset.** In the map, grandparents origin of individuals in the test set is represented. Presence of individuals with ancestry from all over the word is crucial to mitigate effects of blood group alleles stratification.

## *2.2.5 Bootstrap*

A bootstrap sampling of 10,000 simulated genotypes has been performed to further test BOOGIE 2 performance. In particular bootstrap simulation was used to test how well the phasing problem is addressed by the tool. Each sample has been generated from the haplotype tables in the following way: two haplotypes are randomly chosen, then variants are assembled to form a virtual diploid individual losing the phase information. The generated combination of alleles are chosen regardless of real world blood groups frequency, therefore exacerbating allele couples of rare haplotypes. It has to be noted for example that in ABO bootstrap test, the number of virtual individuals with AB blood group is 28%, in contrast to the 5% of cases recorded worldwide.

## *2.2.6 Haplotypes repetition*

To test quality of the haplotype data retrieved from the BGMUT database, a simple test was performed: BGMUT haplotypes were tested for overlap considering only exonic variants. It is possible in fact that several blood haplotype could be rather similar, or even identical as they differ only for intronic variants. Considering that the BOOGIE 2 prediction algorithm does not consider such non-coding variants, the presence of

identical haplotypes could produce equally scoring predictions, reducing BOOGIE 2 prediction performance. Only the presence of identical haplotypes, after introns removal, was reported in our analysis (See Table BOOGIE 5).

### *2.2.7 Analysis of ratio between BGMUT known and unknown variants*

An assessment of the number of variants occurring in the major blood groups genotype has been carried out. Every patient of the PGP dataset was divided in respect of its blood group, and the number of mutations on the exons of ABO and RhD genes was counted, resulting in Figure 29. As anticipated in the BOOGIE 1 manuscript[193] the number of mutations occurring in patients with negative RhD is evidently higher compared to other phenotypes. Ratio between known and unknown mutations in BGMUT was even assessed for correctly and wrongly predicted individuals. Rational for this analysis was to test if wrong BOOGIE 2 predictions could be related to a high number of BGMUT unknown variants (See Figure 30).

# 3 Results

BOOGIE 2 is a tool to predict phenotypes from NGS data using explicit truth tables that link specific SNPs sets to phenotypic traits. Variants sets are extracted from the BGMUT database[190] which stores information about experimentally validated mutations known to be relevant for the determination of 34 blood groups. Prediction performance have been tested on a test set that has been double from the publication of the first version of the tool (69 vs. 133 individuals of current test set). Prediction performance on the major blood groups will be here presented like case study.

## 3.1 ABO blood group performance

On the new PGP dataset, BOOGIE 2 accuracy is 94%, outperforming the first version of the software by 3.1 points (See Figure 26). A physiological reduction of accuracy has to be noted for both BOOGIE versions, respect to the older test set where accuracy is 97.1% and 94.2%, respectively for BOOGIE 2 and BOOGIE 1. Common mistakes could be identified between the two version of the tool, underlining a possible shared issue. Nevertheless, half of the wrong prediction of BOOGIE 1 have been recovered in the new

version. In particular better performance has been achieved for the O blood group were accuracy moved from 88.5% to 96.7%. In addition, for half of the mistakes, BOOGIE 2 is able to report along with the wrong prediction an equally scoring correct phenotype.

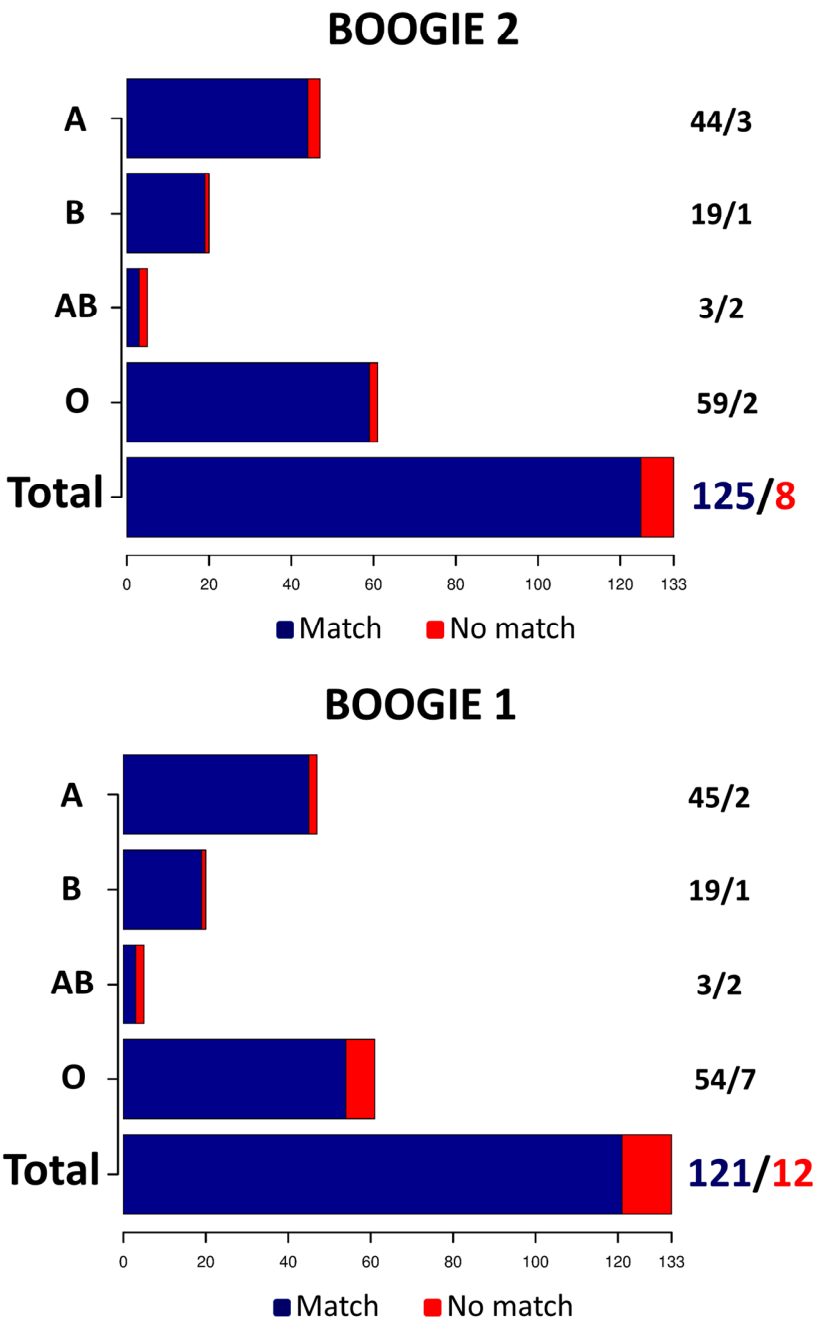


**Figure 26. Accuracy for ABO blood system is compared between BOOGIE 2 and BOOGIE 1.** Overall performance are slightly improved, while for the O phenotype, accuracy is finally comparable to other classes in comparison with BOOGIE 1.

## 3.2 RhD blood group performance

Performance for RhD blood system in the new test set are tied, but a strategic difference could be identified. Both BOOGIE versions correctly predict a total of 118 samples (See Figure 27). Although, in the case of six RhD negative patients mistakenly predicted as RhD positive, BOOGIE 2 report a warning about an unknown deleterious mutation detected with SIFT/PolyPhen-2. Analyzing wrong predictions, is possible to identify another particular situation. Sample hu627574 is predicted as RhD negative being instead RhD positive. For this individual, BOOGIE 2 reports a total of 41 mutations, 28 of which in homozygosis, a status that is typical of RhD negative samples (See Figure 29). In addition, also in this sample the deleterious variants detected with SIFT/PolyPhen-2 is present. A possible explanation for this wrong prediction was found by carefully examining the patient profile on the PGP website. In fact, the blood group phenotype is reported in a self-assessed survey and not by serological test, casting doubts on the reliability of this particular case. To test if cases of self-assessed phenotype could affect the prediction performance assessment, BOOGIE 2 was tested on the dataset resulting after the exclusion of blood group survey information (99 samples). As prediction accuracy for both ABO and RhD systems was exactly the same as in the full dataset, we can assume the performance assessment was only slightly affected by these cases of self-reported phenotypes.

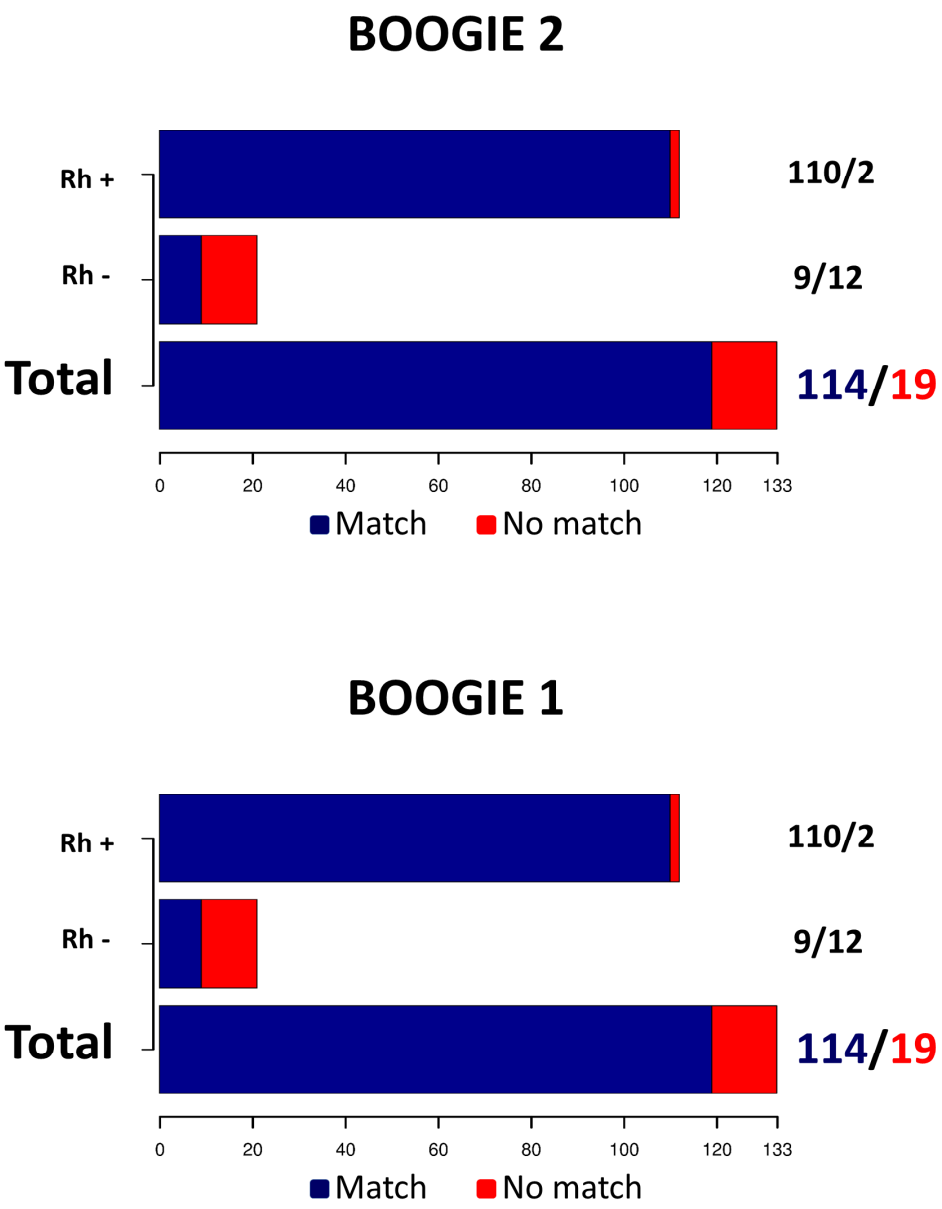


**Figure 27. Accuracy for RhD blood system is compared between BOOGIE 2 and BOOGIE 1.** Unfortunately no improvements have been achieved for this blood system.

## 3.3 Bootstrap

Bootstrap analysis has been conducted to test the effect of the haplotype phasing problem on BOOGIE 2 performance. The blood systems taken into consideration are: ABO, RhD, Kell, Duffy and Lewis. For all blood systems, the identification of haplotype was tested, and for ABO and RhD also phenotype prediction was assessed. Results are shown in Table 27.

For both ABO and RhD blood systems real phenotype is reported in the 99% of the cases, while the haplotype is correctly assigned in 100% of the times (See Table 27). Same situation is present for haplotype detection even for minor blood groups (See Table 27). Considering results of this analysis is possible to assume that BOOGIE 2 management of haplotype phasing is pretty effective and only a marginal number of wrong prediction in real patients could be due to incorrect haplotype management.

It has to be noted that that in some case, equally scoring predictions could be identified. In these conditions the most likely result is predicted by majority rule. This could be considered as a weak point of the BOOGIE 2 prediction algorithm leading to an increase level of uncertainty for blood group prediction. To test how common equally scoring predictions are, we decided to test how often this situation occurs in bootstrap simulation. We test the degree of certainty in the bootstrap predictions considering the number of times an haplotype is reported not univocally. Results are shown in Table 27. From this analysis it is possible to identify that for the ABO system, level of uncertainty is limited (8% of samples, 1552/20000 cases). A similar situation could be identified for most of the other blood systems (between 16% and 25% of the samples). A most complex case has been identified for the Kell system where for 87% (17372/20000 cases) of the samples, an equally scoring prediction was present defining a high degree of uncertainty for such predictions. The same analysis was performed over the haplotypes of the 133 samples from the PGP project. Results are shown in Figure 28. Also in this case, degree of uncertainty is limited for the ABO system with 67% of cases (180/266 haplotypes) predicted with no or little degree of uncertainty (max 3 equally scoring haplotypes). Even better situation is present for the RhD system with 97% of the cases (258/266 haplotypes) with no or little degree of uncertainty (max 3 equally scoring haplotypes).

| | ABO | RhD | FY | KEL | LE | Overall trials |
|---|---|---|---|---|---|---|
| Correctly predicted **haplotypes** | 20000 | 20000 | 20000 | 20000 | 20000 | 20000 |
| Correctly predicted **phenotypes** | 9900 | 9942 | - | - | - | 10000 |
| Uncertain predictions | 1552 | 4396 | 5064 | 17372 | 3176 | 20000 |

**Table 27. Results of bootstrap analysis**. All haplotypes have been correctly predicted and in more than 99% of cases even phenotypes have been matched. An uncertain prediction is counted every time an equally scoring prediction is recorded.

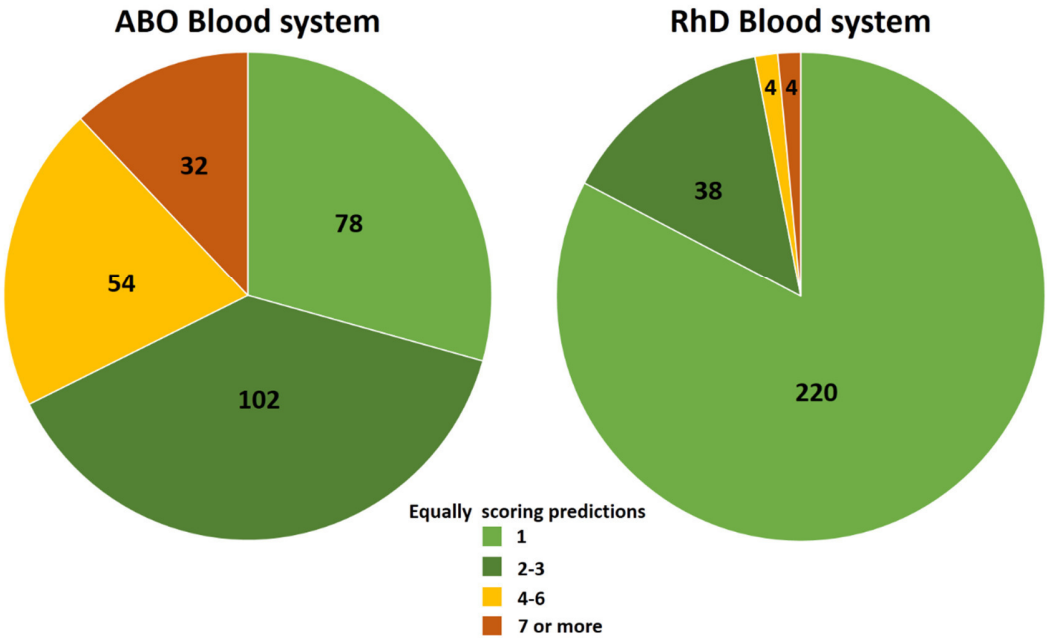


**Figure 28. Uncertain prediction count on the PGP test set.** An uncertain prediction is counted every time an equally scoring prediction is recorded.

## 3.4 Haplotypes repetitions

Haplotypes in the truth tables have been checked for repetitions. As shown in Table 28, ABO haplotype table shows a high absolute number of repeated haplotypes (77). This situation seems to be due to a little number of haplotypes (24) that seems to be repeated few times. Different and more complex situation could be reported (again) for the Kell system where repetitions are due to 41 haplotypes, most of which seems to repeated exactly two times. The explanation of this phenomenon lies in BGMUT database, from where data was gathered. In the ABO case, several alleles have the same exonic genotype but are annotated because they differ by the intronic sequence, which is not considered in BOOGIE 2 predictions. For this reason, such haplotypes in the input haplotypes table, appear the same. This happens few times for the most of the blood phenotypes (See Table 28), resulting in many repetitions of few genotypes. The landscape is different for the Kell blood system. In this case, on many occasions, BGMUT has two identical genotypes with two different names, coming from two different papers. The supposed reason for this phenomenon could be due do the need of link the genotype to the names of both papers, but this obviously generates an issue that could lead to confusion.

| | ABO | RhD | FY | KEL | LE |
|---|---|---|---|---|---|
| Haplotypes repeated 1 or more times | 24 | 8 | 1 | 41 | 2 |
| Total number of repetitions | 77 | 16 | 2 | 84 | 4 |

**Table 28. Haplotypes repetitions.** To test quality data retrieved from the BGMUT database haplotypes were tested for overlap considering only exonic variants**.**

## 3.5 Analysis of ratio between BGMUT known and unknown variants

The 133 PGP patients have been clustered by blood group and plotted by the total number of exonic mutations in Figure 29. As it was hinted in the previous BOOGIE paper[193], RhD negative samples show a very high number of mutations in respect of the reference. Considering that RhD negative samples are also the most difficult to predict (43% of accuracy, see Figure 27) we speculate that the reason for this low performance could be due to a possible high number of variants which are not reported in the BGMUT database. To confirm this hypothesis we tested if wrongly predicted samples presents a higher number of unknown variants respect to guessed individuals. To this aim for each of the 133 PGP genomic sample, the total number of exonic variants in ABO and RhD genes have been compared with the number of variants taken into account by BOOGIE 2 predictions (i.e. variants present in BGMUT). Results have been divided into two subgroups: correctly predicted and wrongly predicted samples (See Figure 30). Differences of two samples distributions have been tested by mean of Student's *t*-test. For both ABO and RhD blood systems statistical test fails to support our hypothesis that wrong predictions correlates with a high number of variants not reported in the BGMUT database.

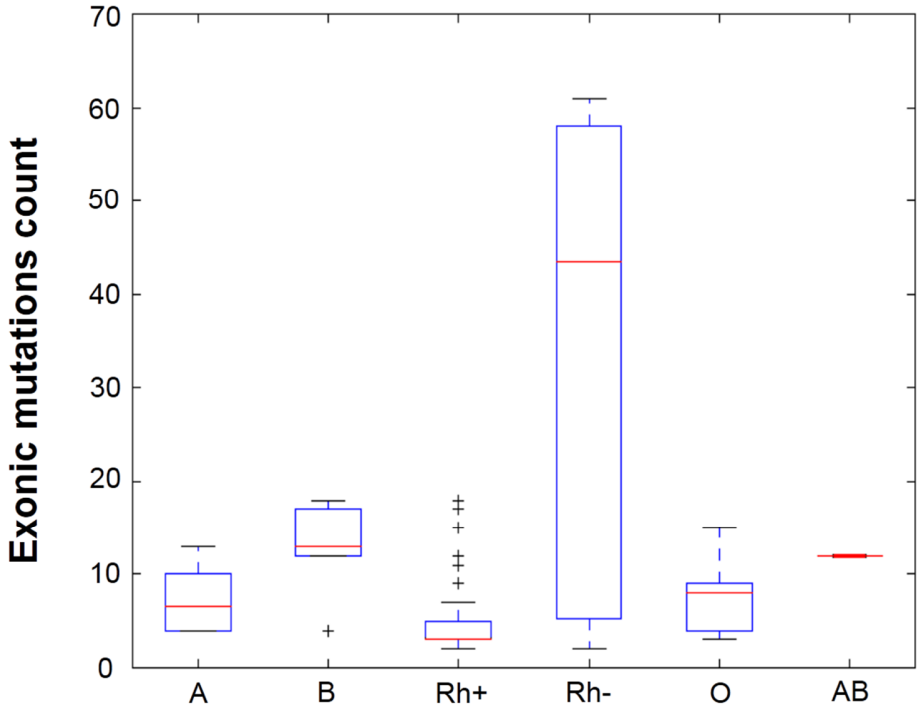


**Figure 29. Exonic variants for major blood groups.** The distribution of exonic variants for major blood groups in PGP individuals is reported.

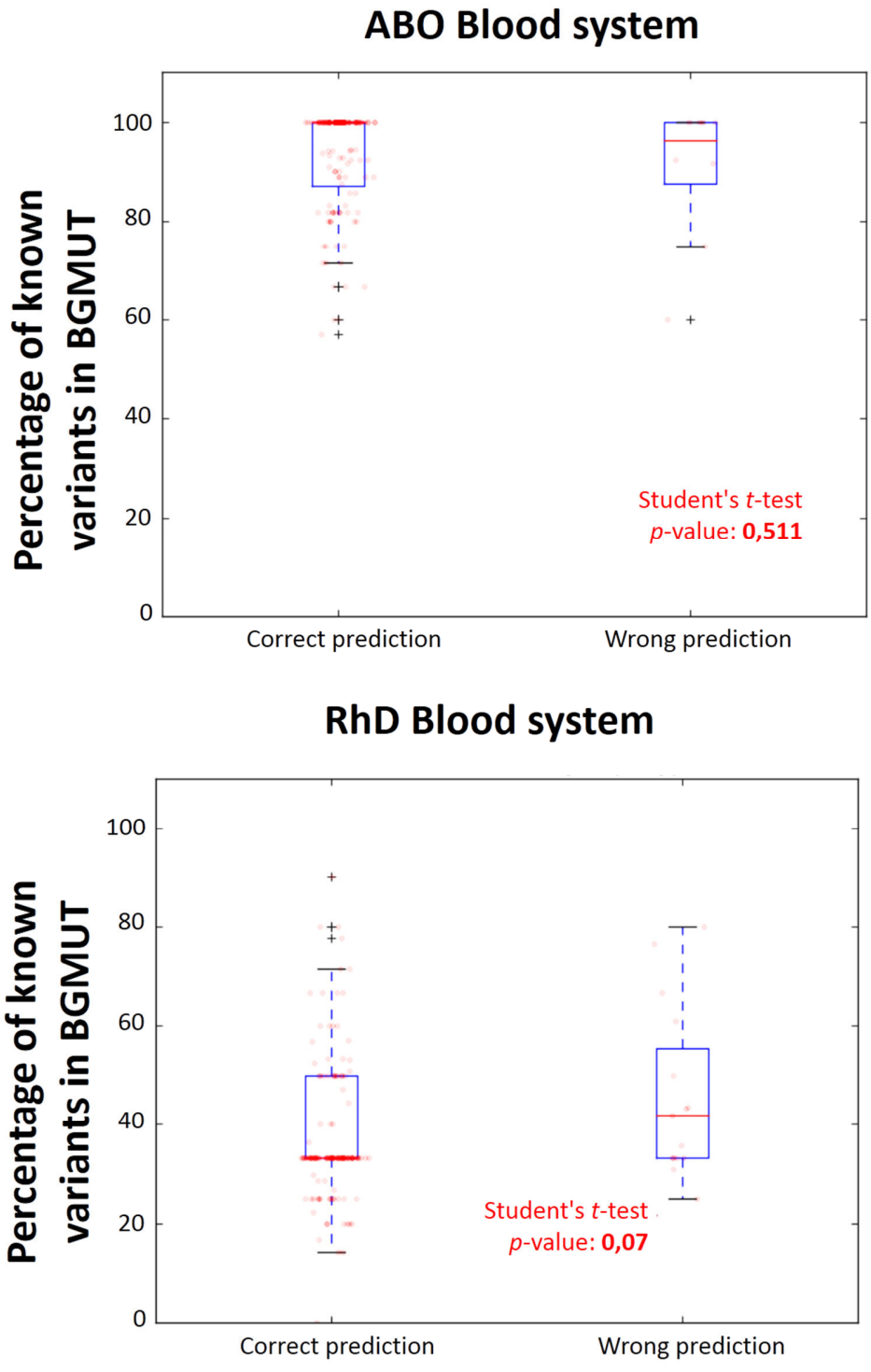


**Figure 30. Percentage of known variants in BGMUT for matched and wrong predictions.**

# 4 Conclusions

Blood typing has a crucial relevance in transfusional medicine. Incompatible transfusion for major blood groups can cause activation of the clotting system, leading to kidney failure, circulatory collapse and in the most severe cases causing the death of the patient. Incompatibilities for minor blood groups or weak phenotypes are substantially armless but in some particular conditions these mismatch could be critical too. In particular for patients affected by anemia, thalassemia and cancer, incompatible transfusions could have severe consequences[183]. Even for transfusion dependent patients, a fine blood matching could be crucial extending period between transfusion and increasing life expectancy[183]. To avoid these kind of reactions, blood compatibility test are always performed before transfusion. Both investigations at phenotype level and genotype level are performed to test blood compatibility in clinical practice. Investigation of blood system phenotype is routinely performed by mean of antiglobulin test but accuracy could be reduced in several cases like weak blood phenotypes, recent transfusion or drug assumption. In these cases an investigation of blood type even at a genotype level is performed thanks to dedicated platforms involving multiplex-PCR combined with flow cytometry. Unfortunately this level of analysis is ten times more expensive than serological test. Nowadays tanks to NGS technology a widespread availability of sequencing data has been achieved thanks exponential reduction in sequencing cost and time.

Thanks to this NGS data deluge, blood groups typing from NGS data is becoming an appealing alternatives. In this context we proposed BOOGIE: is a java tool to predict blood group phenotypes from NGS data. First version of this predictor was published in 2015 and after 2 years, an update and upgrade of this tool was performed.

BOOGIE 2 performance were tested on a PGP dataset that has been doubled respect to test set for the first version of the tool. This dataset could be considered strongly representative of general population as samples cover all possible classes of major blood systems and heterogeneous ethnicity of samples should reduce effects of blood alleles stratification on performance assessment. In addition a complete separation of test set and training set (i.e BGMUT haplotypes) guarantees the presence of no overfitting of the prediction algorithm. Thanks to this independence of training and test sets, it is

reasonable to believe that prediction performance cold been maintained even in real clinical use.

Despite improvements on both the knowledge about blood groups and the algorithm, only slightly improvements in predictions accuracy have been achieved. For the ABO system accuracy is 94%, outperforming the first version of the software by 3.1 points (See Figure 26). Better performance have been achieved on the older PGP test set where accuracy is 97.1% and 94.2%, respectively for BOOGIE 2 and BOOGIE 1. Unfortunately, such inflated accuracy could be probably explained by reduced heterogeneity of samples due to the limited dimension of the old test set (69 individuals). For this reason, an even bigger test set would be required to test BOOGIE 2 performance. Despite PGP samples heterogeneity, it is possible that rare blood phenotypes coded in BGMUT haplotype tables could be underrepresented or absent in such little test sets. Despite these limitations, for the ABO system half of the wrong prediction of BOOGIE 1 have been recovered in the new version. In addition, for half of the remaining mistakes, BOOGIE 2 was able to report along with the wrong prediction an equally scoring correct phenotype. Better predictions performance could be identified in particular for individuals with O phenotype. Probably these improvements are due to introduction of the dominant mutations framework as the ABO c.53G>T dominant variant, introduced in the dominant mutations list, code exactly for the O phenotype. For the RhD system, no improvements in prediction accuracy could been recorded instead. In particular prediction of RhD negative patients still remain a difficult task, probably due to the high rate of variants that differs from reference genome respect to other blood phenotypes (See Figure 29).

To investigate reasons for such limited increase of performance, we focus again our attention on the algorithm and on the knowledge about blood groups.

Our first analysis was focused on the algorithm responsible of heterozygous mutations management, addressing the haplotype phasing problem. To this aim an extensive bootstrap analysis was performed. In simulating patients, each sample was generated selecting two random haplotypes and variants are assembled to form a virtual diploid individual, losing the phase information. For all tested blood systems, real phenotype was reported in the 99% of the cases and haplotypes were correctly assigned in 100% of the times (See Table 27). Considering these results, is reasonable to assume that BOOGIE 2 is able to address effectively the haplotype phasing problem, with only a marginal number of wrong prediction due to incorrect haplotype management.

On the side of knowledge about blood groups, we tried to mitigate a crucial weakness of the BOOGIE framework: the inability to manage presence of VUS that could affect blood group genes. As anticipated, BOOGIE can only consider variants present in the BGMUT tables, while other deleterious mutations are ignored in the prediction process. To address this weakness, predictions of possible mutation impact on the protein's structure and function have been introduced. In the new version of the tool, a report on the unknown mutations found to be "deleterious" or "probably damaging" by SIFT[197] and PolyPhen-2[198] is given to the user as a warning. In this way, BOOGIE 2 predictions are still literature-based, but these warnings could be useful to understand reasons of possible mistaken predictions. Another hint that knowledge on blood genotypes could be responsible for missed prediction comes from the observation that in half of wrong prediction, the correct phenotype was present as an equally scoring predictions. As already introduced, in case of multiple prediction with the same similarity score, final phenotype is defined by mean of majority rule. As BOOGIE considers in its predictions only exonic variants we speculate if haplotypes tables representing different blood haplotypes could be identical after intronic mutations were sifted out. Analysis of haplotypes repetitions confirmed that this problem seems to affect mainly to the ABO system with a high absolute number of repeated haplotypes (77) due to a little number of haplotypes (24) few times. Different situation could be identified for the Kell system where repetitions are due to 41 haplotypes, most of which repeated twice. Similar to this situation is the case of the RhD system. For this blood system repetitions are limited, respect to Kell, but also in this case same haplotypes are repeated twice with different names. In a context like this, it is possible that the presence of several identical haplotypes, coding all for the same blood phenotype could influence the majority rule based phenotype prediction leading to samples misclassification.

Despite these flaws, test on the PGP dataset confirmed very good prediction performance. Even if BOOGIE 2 is not yet suitable for direct medical application, results of our analysis suggests that an improvement of prediction performance is still possible. In particular working on haplotype tables curation could probably reduce cases of equally scoring predictions and also use of majority rule for final phenotype prediction could be revised. As final consideration a crucial, still missing analysis, is test of BOOGIE 2 performance on minor blood systems where performance of serological test is reduced in respect to ABO

and Rh. Despite this deficiency, thanks to our bootstrap analysis (See Table27) good prediction performance are expected also for these blood groups phenotypes.

# Conclusions

The Human Genome Project leaded to new extraordinary technical achievements. DNA Sequencing technology has evolved by several orders of magnitude and nowadays sequencing the human genome is approaching the psychological threshold of $1,000. High-throughput data has never been so accessible, opening many new ways in both research and clinical practice. In this context a new trend in medicine has emerged: the personalized medicine. The central paradigm of personalized medicine is the use of the genetic data to find specific disease mechanisms that can be treated with a specific personalized therapy. Thanks to the combination of genome sequencing with other "omics" data, in the near future will be possible to define personalized disease risk and potentially apply specific prevention strategies to delay or avoid disease onset. In addition the possibility to achieve precise diagnosis early after disease onset, could allow more effective treatments. Several examples of personalized medicine have been already translated in clinical routine[202], however these cases are usually limited to monogenic diseases and a greater effort in order to realize the potential of personalized medicine is required, especially for complex phenotypes. My PhD project focused on the development of bioinformatics tools to predict phenotypes from NGS-data. To achieve this aim I had the opportunity to deal with all the typical challenges that have to be address in the context of the NGS revolution. The problem of evaluating methods developed for the interpretation of variants of unknown significance, was addressed like assessor in the CAGI p16INK4a challenge. Our attention focused on the identification of methods that generated the most reliable prediction. To this aim, a plethora of different metrics was considered in order to perform a fair assessment. The results of this challenge suggest that methods combining different strategies seem to perform better than simpler approaches. Unfortunately, this trend needs to be further confirmed, considering a larger number of variants and different proteins. In addition, despite some methods scored reasonably well, their performance are far from making these predictors reliable resources to be used in the clinical practice.

Several predictors have been developed during my PhD project, dealing both with data coming from targeted enrichment and exomes sequencing experiments. Most of these predictors have been tested in the context of the CAGI experiment. In the case of the Hopkins Clinical Panel, several disease phenotype have to be predicted from the few

genes present in a diagnostic panel developed for multiple pathology testing. Our group did a reasonable job in predicting clinical phenotypes but at the same time, false positives and false negatives have been predicted ad an unacceptable rate for clinical purposes. Interestingly about this challenge was the fact that several groups were able to make prediction and to identify putative pathogenic alleles even for patients where no pathogenic variants were identified by data provider. In addition, several individuals have been predicted by several groups like affected by pathologies that they didn't develop, reporting even the same putative casual variants. These findings well explain two emerging issues in the personalized medicine, the misinterpretation and the overinterpretation of variants with unknown significance.

Other kinds of predictors have been developed starting from exome sequencing data, like in the case of Crohn's disease, a complex pathology characterized by a spectrum of clinical traits. The main phenotype corresponds to the misregulation of intestinal inflammation, sometimes extended to the whole digestive system and in some critical cases also affecting skin and joints. The onset age ranges from birth, with the most severe manifestation, to adulthood. We developed a Crohn's predictor to identify healthy and affected individuals from exome sequencing data. Prediction performance have been tested in the context of the CAGI 4 experiment. Our prediction strategy is based on the hypothesis that the amount of Crohn's related mutation correlates directly with the probability of being affected. To this aim, all common and synonymous SNPs have been sifted out, since less likely to be involved. In addition, a big effort was made to define the list of relevant variants by manually curating literature in order to identify a list of mutations associated with the very early disease onset. Another list was defined considering genes associated to the pathology from GWAS studies. At the end, trying to address the problem of missing heritability, we expanded the list of associated genes exploiting information extracted from protein-protein interaction networks. Our method scored an AUC of 0.61. Particularly interesting is the fact that all top scoring individuals are effectively affected by Crohn's disease, indicating that by applying an appropriate confidence threshold our method is able to detect the disease phenotype unambiguously. Comparing the results of previous CAGI experiments, the best methods performance decreased from an AUC of ca. 0.9 to ca. 0.7. This can be explained by a bias in the datasets of previous CAGI experiments, i.e. the presence of a population structure. Previous results were inflated by the presence of strong bias in both training sets and test sets. In this

context, it will be crucial for the next CAGI Crohn's disease challenges to provide normalized datasets.

Finally, a predictor for blood types from exome sequencing data has been proposed. Blood type is an example of a phenotype fully determined by the genotype. Blood group typing is routinely performed in transfusional medicine. Incompatibilities for major blood groups are tested by means of cost effective antiglobuline tests. Unfortunately, the accuracy of these investigations for minor and weak phenotypes is very limited. Blood incompatibilities might have severe consequences for critical individuals as those affected by anemia, cancer or bounded to routinely transfusions. To improve blood typing we developed an improved version of BOOGIE (BOOGIE 2), a bioinformatics tool to predict blood types from NGS-data. The tool exploits haplotype tables storing genotype information for all known blood groups. Predictions are based on a nearest neighbor search to find the best match between a given genotype and the registered haplotypes. BOOGIE 2 for the ABO system on a dataset of 133 individuals, reaches 94% accuracy. For half of the wrong predictions, BOOGIE 2 is even able to identify the correct phenotype but two or more phenotypes with same score are returned. Predictions for the RhD system achieved an accuracy of 85.7%. In particular, the RhD negative type is particularly difficult to predict, probably due to the high rate of variants compared to other groups. A bootstrap analysis statistically validates the ability of BOOGIE to effectively address the haplotype phasing problem. Poor performances seem to be related to the limited knowledge about some groups genotypes. However, it is likely that in the near future, with more genotype data coming, further performance improvements will be possible.

Personalized medicine is improving fast but genotype interpretation seems to be the real problem to solve in order to translate the benefits to clinics. Big companies are moving their business from genetic testing to genomic accumulation and data reselling[203], suggesting the problem is not the cost but the scientific knowledge.

# Bibliography


1. Roses, A. D. Pharmacogenetics and the practice of medicine. *Nature* **405,** 857–865 (2000).
2. Mayor, S. Genome sequence of one individual is published for first time. *BMJ* **335,** 530–531 (2007).
3. Nyholt, D. R., Yu, C.-E. & Visscher, P. M. On Jim Watson's APOE status: genetic information is hard to hide. *Eur. J. Hum. Genet.* **17,** 147–149 (2009).
4. Davies, K. *The $1,000 Genome: The Revolution in DNA Sequencing and the New Era of Personalized Medicine*. (Simon and Schuster, 2015).
5. Chen, R. & Snyder, M. Systems biology: personalized medicine for the future? *Curr. Opin. Pharmacol.* **12,** 623–628 (2012).
6. Tremblay, J. & Hamet, P. Role of genomics on the path to personalized medicine. *Metabolism.* **62 Suppl 1,** S2-5 (2013).
7. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 10869–10874 (2001).
8. Easton, D. F. How many more breast cancer predisposition genes are there? *Breast Cancer Res. BCR* **1,** 14–17 (1999).
9. Campeau, P. M., Foulkes, W. D. & Tischkowitz, M. D. Hereditary breast cancer: new genetic developments, new therapeutic avenues. *Hum. Genet.* **124,** 31–42 (2008).
10. Pal, T. *et al.* BRCA1 and BRCA2 mutations account for a large proportion of ovarian carcinoma cases. *Cancer* **104,** 2807–2816 (2005).
11. Green, R. C. *et al.* ACMG Recommendations for Reporting of Incidental Findings in Clinical Exome and Genome Sequencing. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **15,** 565–574 (2013).
12. Alyass, A., Turcotte, M. & Meyre, D. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med. Genomics* **8,** 33 (2015).
13. Hood, L. & Flores, M. A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory. *New Biotechnol.* **29,** 613–624 (2012).
14. Schloss, J. A. How to get genomes at one ten-thousandth the cost. *Nat. Biotechnol.* **26,** 1113–1115 (2008).
15. Sun, Y. *et al.* Next-Generation Diagnostics: Gene Panel, Exome, or Whole Genome? *Hum. Mutat.* **36,** 648–655 (2015).
16. Marian, A. J. Challenges in medical applications of whole exome/genome sequencing discoveries. *Trends Cardiovasc. Med.* **22,** 219–223 (2012).
17. Goh, G. & Choi, M. Application of Whole Exome Sequencing to Identify Disease-Causing Variants in Inherited Human Diseases. *Genomics Inform.* **10,** 214–219 (2012).
18. Rabbani, B., Tekin, M. & Mahdieh, N. The promise of whole-exome sequencing in medical genetics. *J. Hum. Genet.* **59,** 5–15 (2014).
19. Xue, Y., Ankala, A., Wilcox, W. R. & Hegde, M. R. Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation

sequencing: single-gene, gene panel, or exome/genome sequencing. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **17,** 444–451 (2015).
20. Bush, W. S. & Moore, J. H. Chapter 11: Genome-Wide Association Studies. *PLOS Comput. Biol.* **8,** e1002822 (2012).
21. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90,** 7–24 (2012).
22. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
23. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001–D1006 (2014).
24. Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. Common vs. Rare Allele Hypotheses for Complex Diseases. *Curr. Opin. Genet. Dev.* **19,** 212–219 (2009).
25. Crawford, D. C. *et al.* Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am. J. Hum. Genet.* **74,** 610–622 (2004).
26. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12,** 56–68 (2011).
27. Chicco, D. Ten quick tips for machine learning in computational biology. *BioData Min.* **10,** 35 (2017).
28. Apiletti, D., Bruno, G., Ficarra, E. & Baralis, E. Data Cleaning and Semantic Improvement in Biological Databases. *J. Integr. Bioinforma.* **3,** 219–229 (2006).
29. Boulesteix, A.-L. Ten Simple Rules for Reducing Overoptimistic Reporting in Methodological Computational Research. *PLOS Comput. Biol.* **11,** e1004191 (2015).
30. Refaeilzadeh, P., Tang, L. & Liu, H. Cross-Validation. in *Encyclopedia of Database Systems* 532–538 (Springer, Boston, MA, 2009). doi:10.1007/978-0-387-39940-9_565
31. Skocik, M., Collins, J., Callahan-Flintoft, C., Bowman, H. & Wyble, B. I Tried a Bunch of Things: The Dangers of Unexpected Overfitting in Classification. *bioRxiv* 078816 (2016). doi:10.1101/078816
32. Hand, D. J. Classifier Technology and the Illusion of Progress. *Stat. Sci.* **21,** 1–14 (2006).
33. Vihinen, M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics* **13 Suppl 4,** S2 (2012).
34. Bromberg, Y. Building a genome analysis pipeline to predict disease risk and prevent disease. *J. Mol. Biol.* **425,** 3993–4005 (2013).
35. Abraham, G. & Inouye, M. Genomic risk prediction of complex human disease and its clinical application. *Curr. Opin. Genet. Dev.* **33,** 10–16 (2015).
36. Hu, H. *et al.* VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet. Epidemiol.* **37,** 622–634 (2013).
37. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3,** S3 (2013).
38. Annas, G. J. & Elias, S. 23andMe and the FDA. *N. Engl. J. Med.* **370,** 985–

988 (2014).
39. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/20711175. (Accessed: 18th August 2017)
40. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.* **42,** 30–35 (2010).
41. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 19096–19101 (2009).
42. Worthey, E. A. *et al.* Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **13,** 255–262 (2011).
43. Bilgüvar, K. *et al.* Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* **467,** 207–210 (2010).
44. Moult, J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.* **15,** 285–289 (2005).
45. Moult, J. Rigorous performance evaluation in protein structure modelling and implications for computational biology. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **361,** 453–458 (2006).
46. Moult, J., Fidelis, K., Kryshtafovych, A. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)--round IX. *Proteins* **79 Suppl 10,** 1–5 (2011).
47. Daneshjou, R. *et al.* Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. *Hum. Mutat.* (2017). doi:10.1002/humu.23280
48. Niroula, A. & Vihinen, M. Variation Interpretation Predictors: Principles, Types, Performance, and Choice. *Hum. Mutat.* **37,** 579–597 (2016).
49. Kannengiesser, C. *et al.* Functional, structural, and genetic evaluation of 20 CDKN2A germ line mutations identified in melanoma-prone families or patients. *Hum. Mutat.* **30,** 564–74 (2009).
50. Miller, P. J. *et al.* Classifying variants of CDKN2A using computational and laboratory studies. *Hum. Mutat.* **32,** 900–11 (2011).
51. Walsh, I., Pollastri, G. & Tosatto, S. C. E. Correct machine learning on protein sequences: a peer-reviewing perspective. *Brief. Bioinform.* **17,** 831–840 (2016).
52. Liu, Y. & Bodmer, W. F. Analysis of P53 mutations and their expression in 56 colorectal cancer cell lines. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 976–981 (2006).
53. Leonardi, E., Martella, M., Tosatto, S. C. E. & Murgia, A. Identification and in silico analysis of novel von Hippel-Lindau (VHL) gene variants from a large population. *Ann. Hum. Genet.* **75,** 483–496 (2011).
54. Scaini, M. C. *et al.* CDKN2A unclassified variants in familial malignant melanoma: combining functional and computational approaches for their assessment. *Hum. Mutat.* **35,** 828–840 (2014).
55. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43,** D805-811 (2015).
56. Wang, C. *et al.* DBGC: A Database of Human Gastric Cancer. *PloS One* **10,** e0142591 (2015).

57. Tabaro, F. *et al.* VHLdb: A database of von Hippel-Lindau protein interactors and mutations. *Sci. Rep.* **6,** 31128 (2016).
58. Wang, J. & Shen, Y. When a "Disease-Causing Mutation" Is Not a Pathogenic Variant. *Clin. Chem.* **60,** 711–713 (2014).
59. Manolio, T. A. Genomewide Association Studies and Assessment of the Risk of Disease. *N. Engl. J. Med.* **363,** 166–176 (2010).
60. Hussussian, C. J. *et al.* Germline p16 mutations in familial melanoma. *Nat. Genet.* **8,** 15–21 (1994).
61. Serrano, M., Hannon, G. J. & Beach, D. A new regulatory motif in cell-cycle control causing specific inhibition of cyclin D/CDK4. *Nature* **366,** 704–707 (1993).
62. Zhang, Y., Xiong, Y. & Yarbrough, W. G. ARF promotes MDM2 degradation and stabilizes p53: ARF-INK4a locus deletion impairs both the Rb and p53 tumor suppression pathways. *Cell* **92,** 725–734 (1998).
63. Aoude, L. G., Wadt, K. A. W., Pritchard, A. L. & Hayward, N. K. Genetics of familial melanoma: 20 years after CDKN2A. *Pigment Cell Melanoma Res.* **28,** 148–160 (2015).
64. Andreotti, V. *et al.* The CDKN2A/p16(INK) (4a) 5'UTR sequence and translational regulation: impact of novel variants predisposing to melanoma. *Pigment Cell Melanoma Res.* **29,** 210–221 (2016).
65. Sherr, C. J. G1 phase progression: cycling on cue. *Cell* **79,** 551–555 (1994).
66. Weinberg, R. A. The retinoblastoma protein and cell cycle control. *Cell* **81,** 323–330 (1995).
67. Tang, K. S., Guralnick, B. J., Wang, W. K., Fersht, A. R. & Itzhaki, L. S. Stability and folding of the tumour suppressor protein p16. *J. Mol. Biol.* **285,** 1869–1886 (1999).
68. Peng, Z. The ankyrin repeat as molecular architecture for protein recognition. 1435–1448 (2004). doi:10.1110/ps.03554604.ity
69. Scaini, M. C. *et al.* Functional impairment of p16INK4A due to CDKN2A p.Gly23Asp missense mutation. *Mutat. Res. Mol. Mech. Mutagen.* **671,** 26–32 (2009).
70. Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L. & Casadio, R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.* **30,** 1237–1244 (2009).
71. Capriotti, E. *et al.* WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. *BMC Genomics* **14,** 1–7 (2013).
72. Hamp, T. & Rost, B. Alternative Protein-Protein Interfaces Are Frequent Exceptions. *PLoS Comput. Biol.* **8,** e1002623 (2012).
73. Dimaio, F., Leaver-Fay, A., Bradley, P., Baker, D. & André, I. Modeling symmetric macromolecular structures in rosetta3. *PloS One* **6,** e20450 (2011).
74. Zhou, H. & Zhou, Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci. Publ. Protein Soc.* **11,** 2714–2726 (2002).
75. Katsanis, S. H. & Katsanis, N. Molecular genetic testing and the future of clinical genomics. *Nat. Rev. Genet.* **14,** 415–426 (2013).
76. Sosnay, P. R. *et al.* Applying Cystic Fibrosis Transmembrane Conductance Regulator Genetics and CFTR2 Data to Facilitate Diagnoses. *Cyst. Fibros. Found.*

*Consens. Guidel. Diagn. Cyst. Fibros.* **181, Supplement,** S27–S32.e1 (2017).
77. Schulz, W. L., Tormey, C. A. & Torres, R. Computational Approach to Annotating Variants of Unknown Significance in Clinical Next Generation Sequencing. *Lab. Med.* **46,** 285–289 (2015).
78. Lee, H. *et al.* Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders. *JAMA* **312,** 1880–1887 (2014).
79. Posey, J. E. *et al.* Molecular Diagnostic Experience of Whole-Exome Sequencing in Adult Patients. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **18,** 678–685 (2016).
80. Vassy, J. L. *et al.* The MedSeq Project: a randomized trial of integrating whole genome sequencing into clinical medicine. *Trials* **15,** 85 (2014).
81. Richards, S. *et al.* 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 (2015).
82. Hill, S. M. *et al.* Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat. Methods* **advance online publication,** (2016).
83. Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J. & Williamson, R. C. Estimating the Support of a High-Dimensional Distribution. *Neural Comput.* **13,** 1443–1471 (2001).
84. The 1000 Genomes Project Consortium. A map of human genome variation from population scale sequencing. *Nature* **467,** 1061–1073 (2010).
85. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44,** D862–D868 (2016).
86. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).
87. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
88. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–3814 (2003).
89. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr. Protoc. Hum. Genet. Editor. Board Jonathan Haines Al* **0 7,** Unit7.20-Unit7.20 (2013).
90. El-Fishawy, P. Common Disease-Rare Variant Hypothesis. in *Encyclopedia of Autism Spectrum Disorders* (ed. Volkmar, F. R.) 720–722 (Springer New York, 2013). doi:10.1007/978-1-4419-1698-3_1997
91. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).
92. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33,** D514–D517 (2005).
93. Kirmani, S. & Young, W. F. Hereditary Paraganglioma-Pheochromocytoma Syndromes. in *GeneReviews* (eds. Pagon, R. A. et al.) (University of Washington, Seattle, 1993).

94. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17,** 122 (2016).
95. Yue, P., Melamud, E. & Moult, J. SNPs3D: Candidate gene and SNP selection for association studies. *BMC Bioinformatics* **7,** 166–166 (2006).
96. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc.* **4,** 1073–1081 (2009).
97. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).
98. Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD®): 2003 update. *Hum. Mutat.* **21,** 577–581 (2003).
99. Tricarico, R. *et al.* Assessment of the InSiGHT Interpretation Criteria for the Clinical Classification of 24 MLH1 and MSH2 Gene Variants. *Hum. Mutat.* **38,** 64–77 (2017).
100. Manrai, A. K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *N. Engl. J. Med.* **375,** 655–665 (2016).
101. Walsh, R. *et al.* Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet. Med.* **19,** 192–203 (2017).
102. Amendola, L. M. *et al.* Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. *Am. J. Hum. Genet.* **98,** 1067–1076 (2016).
103. Garber, K. B. *et al.* Reassessment of Genomic Sequence Variation to Harmonize Interpretation for Personalized Medicine. *Am. J. Hum. Genet.* **99,** 1140–1149 (2016).
104. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508,** 469–476 (2014).
105. Weedon, M. N. *et al.* Combining Information from Common Type 2 Diabetes Risk Polymorphisms Improves Disease Prediction. *PLOS Med* **3,** e374 (2006).
106. Morrison, A. C. *et al.* Prediction of Coronary Heart Disease Risk using a Genetic Risk Score: The Atherosclerosis Risk in Communities Study. *Am. J. Epidemiol.* **166,** 28–35 (2007).
107. Giollo, M. *et al.* BOOGIE: Predicting Blood Groups from High Throughput Sequencing Data. *PloS One* **10,** e0124579 (2015).
108. Giollo, M., Martin, A. J., Walsh, I., Ferrari, C. & Tosatto, S. C. NeEMO: a method using residue interaction networks to improve prediction of protein stability upon mutation. *BMC Genomics* **15,** 1–11 (2014).
109. Friedman, J. H. On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality. *Data Min. Knowl. Discov.* **1,** 55–77
110. Saeys, Y., Inza, I. & Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **23,** 2507–2517 (2007).
111. WTCC Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–678 (2007).
112. Easton, D. F., Bishop, D. T., Ford, D. & Crockford, G. P. Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. *Am. J. Hum. Genet.* **52,** 678–701 (1993).

113. Sim, N.-L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40,** W452–W457 (2012).
114. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).
115. Thusberg, J., Olatubosun, A. & Vihinen, M. Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.* **32,** 358–368 (2011).
116. GAP & WTCC Consortium. A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nat. Genet.* **42,** 985–990 (2010).
117. Molodecky, N. A. & Kaplan, G. G. Environmental Risk Factors for Inflammatory Bowel Disease. *Gastroenterol. Hepatol.* **6,** 339–346 (2010).
118. Ellinghaus, D. *et al.* Association between variants of PRDM1 and NDP52 and Crohn's disease, based on exome sequencing and functional studies. *Gastroenterology* **145,** 339–347 (2013).
119. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
120. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
121. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).
122. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).
123. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).
124. Obenchain, V. *et al.* VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics* **30,** 2076–2078 (2014).
125. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4,** 1184–1191 (2009).
126. Beck, T., Hastings, R. K., Gollapudi, S., Free, R. C. & Brookes, A. J. GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur. J. Hum. Genet.* **22,** 949–952 (2014).
127. Yu, W., Clyne, M., Khoury, M. J. & Gwinn, M. Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. *Bioinformatics* **26,** 145–146 (2010).
128. Shivananda, S. *et al.* Incidence of inflammatory bowel disease across Europe: is there a difference between north and south? Results of the European Collaborative Study on Inflammatory Bowel Disease (EC-IBD). *Gut* **39,** 690–697 (1996).
129. 1000GP Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
130. Jain, A. K. Data clustering: 50 years beyond K-means. *Pattern Recognit. Lett.* **31,** 651–666 (2010).
131. Chapelle, O., Scholkopf, B. & Eds, A. Z. Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006) [Book reviews]. *IEEE Trans. Neural Netw.* **20,** 542–542 (2009).

132. Efron, B. Bootstrap Methods: Another Look at the Jackknife. in *Breakthroughs in Statistics* (eds. Kotz, S. & Johnson, N. L.) 569–593 (Springer New York, 1992). doi:10.1007/978-1-4612-4380-9_41
133. Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **30,** 1145–1159 (1997).
134. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164–e164 (2010).
135. Ramos, E. M. *et al.* Phenotype–Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur. J. Hum. Genet.* **22,** 144–147 (2014).
136. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39,** D561–D568 (2011).
137. Ningappa, M. *et al.* NOD2 Gene Polymorphism rs2066844 Associates With Need for Combined Liver–Intestine Transplantation in Children With Short-Gut Syndrome. *Am. J. Gastroenterol.* **106,** 157–165 (2011).
138. Wei, Z. *et al.* Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am. J. Hum. Genet.* **92,** 1008–1012 (2013).
139. Ashley, E. A. The precision medicine initiative: a new national effort. *Jama* **313,** 2119–2120 (2015).
140. Ashley, E. A. *et al.* Clinical evaluation incorporating a personal genome. *Lancet* **375,** 1525–1535 (2010).
141. Brown, T. L. & Meloche, T. M. Exome sequencing a review of new strategies for rare genomic disease research. *Genomics* **108,** 109–114 (2016).
142. Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K. J. & Altman, R. B. Bioinformatics challenges for personalized medicine. *Bioinformatics* **27,** 1741–1748 (2011).
143. Baumgart, D. C. & Sandborn, W. J. Crohn's disease. *The Lancet* **380,** 1590–1605 (2012).
144. Abraham, C. & Cho, J. H. Inflammatory bowel disease. *N. Engl. J. Med.* **361,** 2066–2078 (2009).
145. Bianco, A. M., Girardelli, M. & Tommasini, A. Genetics of inflammatory bowel disease from multifactorial to monogenic forms. *World J. Gastroenterol.* **21,** 12296–12310 (2015).
146. Spehlmann, M. E. *et al.* Epidemiology of inflammatory bowel disease in a German twin cohort: Results of a nationwide study. *Inflamm. Bowel Dis.* **14,** 968–976 (2008).
147. Shouval, D. S. *et al.* Interleukin 10 Receptor Signaling: Master Regulator of Intestinal Mucosal Homeostasis in Mice and Humans. *Adv. Immunol.* **122,** 177–210 (2014).
148. Tuvlin, J. A. *et al.* Smoking and inflammatory bowel disease: trends in familial and sporadic cohorts. *Inflamm. Bowel Dis.* **13,** 573–579 (2007).
149. Bloom, S. M. *et al.* Commensal Bacteroides species induce colitis in host-genotype-specific fashion in a mouse model of inflammatory bowel disease. *Cell*

*Host Microbe* **9,** 390–403 (2011).
150. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).
151. Chang, X. & Wang, K. wANNOVAR: annotating genetic variants for personal genomes via the web. *J. Med. Genet.* **49,** 433–436 (2012).
152. Pruitt, K. D., Tatusova, T., Klimke, W. & Maglott, D. R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.* **37,** D32-36 (2009).
153. Consortium, T. 1000 G. P. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
154. Ramos, E. M. *et al.* Phenotype–Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur. J. Hum. Genet.* **22,** 144–147 (2014).
155. Tryka, K. A. *et al.* NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res.* **42,** D975–D979 (2014).
156. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901 (2017).
157. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).
158. Baumann, N. How to use the medical subject headings (MeSH). *Int. J. Clin. Pract.* **70,** 171–174 (2016).
159. von Mering, C. *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **31,** 258–261 (2003).
160. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* **45,** D362–D368 (2017).
161. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw1092
162. von Mering, C. *et al.* STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33,** D433-437 (2005).
163. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).
164. Kanehisa, M. *et al.* From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34,** D354-357 (2006).
165. Raychaudhuri, S. Mapping rare and common causal alleles for complex human diseases. *Cell* **147,** 57–69 (2011).
166. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci. U. S. A.* **109,** 1193–1198 (2012).
167. Rossin, E. J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.* **7,** e1001273 (2011).
168. Wray, N. R., Yang, J., Goddard, M. E. & Visscher, P. M. The genetic

interpretation of area under the ROC curve in genomic profiling. *PLoS Genet.* **6,** e1000864 (2010).
169. Daneshjou, R. *et al.* Working towards precision medicine: predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. *Hum. Mutat.* (2017). doi:10.1002/humu.23280
170. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001–D1006 (2013).
171. Yang, H., Robinson, P. N. & Wang, K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods* **12,** 841–843 (2015).
172. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29 (2000).
173. Bromberg, Y. & Rost, B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* **35,** 3823–3835 (2007).
174. Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L. & Casadio, R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.* **30,** 1237–1244 (2009).
175. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).
176. Niroula, A., Urolagin, S. & Vihinen, M. PON-P2: Prediction Method for Fast and Reliable Identification of Harmful Variants. *PLoS ONE* **10,** e0117380 (2015).
177. Cooper, D. N., Stenson, P. D. & Chuzhanova, N. A. The human gene mutation database (HGMD) and its exploitation in the study of mutational mechanisms. *Curr Protoc Bioinforma.* **Chapter 1,** (2006).
178. Yazer, M. H., Olsson, M. L. & Palcic, M. M. The cis-AB Blood Group Phenotype: Fundamental Lessons in Glycobiology. *Transfus. Med. Rev.* **20,** 207–217 (2006).
179. Cooling, L. Blood Groups in Infection and Host Susceptibility. *Clin. Microbiol. Rev.* **28,** 801–870 (2015).
180. Mitra, R., Mishra, N. & Rath, G. P. Blood groups systems. *Indian J. Anaesth.* **58,** 524–528 (2014).
181. Zarandona, J. M. & Yazer, M. H. The role of the Coombs test in evaluating hemolysis in adults. *CMAJ Can. Med. Assoc. J.* **174,** 305–307 (2006).
182. Jungbauer, C. Blood group molecular genotyping. *ISBT Sci. Ser.* **6,** 399–403 (2011).
183. Matteocci, A. & Pierelli, L. Red blood cell alloimmunization in sickle cell disease and in thalassaemia: current status, future perspectives and potential role of molecular typing. *Vox Sang.* **106,** 197–208 (2014).
184. Finning, K. *et al.* Evaluation of red blood cell and platelet antigen genotyping platforms (ID CORE XT/ID HPA XT) in routine clinical practice. *Blood Transfus.* **14,** 160–167 (2016).
185. Thomson, H. Baby's genes mapped at birth. *New Sci.* **226,** 8–9 (2015).
186. Flegel, W. A. Molecular genetics and clinical applications for RH. *Transfus. Apher. Sci.* **44,** 81–91 (2011).
187. Yip, S. P. Sequence variation at the human ABO locus. *Ann. Hum. Genet.* **66,** 1–27 (2002).

188. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12,** 703–714 (2011).
189. BGMUT: NCBI dbRBC database of allelic variations of genes encoding antigens of blood group systems | Nucleic Acids Research | Oxford Academic. Available at: https://academic.oup.com/nar/article/40/D1/D1023/2903604/BGMUT-NCBI-dbRBC-database-of-allelic-variations-of. (Accessed: 22nd September 2017)
190. Patnaik, S. K., Helmberg, W. & Blumenfeld, O. O. BGMUT Database of Allelic Variants of Genes Encoding Human Blood Group Antigens. *Transfus. Med. Hemotherapy* **41,** 2–2 (2014).
191. Lange, L. A. *et al.* Whole-Exome Sequencing Identifies Rare and Low-Frequency Coding Variants Associated with LDL Cholesterol. *Am. J. Hum. Genet.* **94,** 233–245 (2014).
192. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30,** 1006–1007 (2014).
193. Giollo, M. *et al.* BOOGIE: Predicting Blood Groups from High Throughput Sequencing Data. *PLOS ONE* **10,** e0124579 (2015).
194. Thorisson, G. A., Smith, A. V., Krishnan, L. & Stein, L. D. The International HapMap Project Web site. *Genome Res.* **15,** 1592–1593 (2005).
195. Browning, S. R. & Browning, B. L. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *Am. J. Hum. Genet.* **81,** 1084–1097 (2007).
196. Amado, M., Bennett, E. P., Carneiro, F. & Clausen, H. Characterization of the Histo-Blood Group O2 Gene and Its Protein Product. *Vox Sang.* **79,** 219–226 (2000).
197. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–1081 (2009).
198. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr. Protoc. Hum. Genet. Editor. Board Jonathan Haines Al* **0 7,** Unit7.20 (2013).
199. Schulz, W. L., Tormey, C. A. & Torres, R. Computational Approach to Annotating Variants of Unknown Significance in Clinical Next Generation Sequencing. *Lab. Med.* **46,** 285–289 (2015).
200. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21,** 3439–3440 (2005).
201. Aminov, R. I. A Brief History of the Antibiotic Era: Lessons Learned and Challenges for the Future. *Front. Microbiol.* **1,** (2010).
202. Dammann, M. & Weber, F. Personalized medicine: caught between hope, hype and the real world. *Clinics* **67,** 91–97 (2012).
203. Roberts, J. L., Pereira, S. & McGuire, A. L. Should you profit from your genome? *Nat. Biotechnol.* **35,** nbt.3757 (2017).