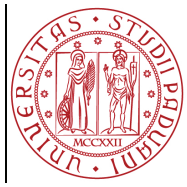


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova
Dipartimento di Biologia

SCUOLA DI DOTTORATO DI RICERCA IN
BIOSCIENZE E BIOTECNOLOGIE

INDIRIZZO: BIOCHIMICA E BIOFISICA
CICLO: XXIV

BIOINFORMATIC ANALYSIS OF PROTEIN MUTATIONS

Direttore della Scuola: Ch.mo Prof. Giuseppe Zanotti

Coordinatore d'indirizzo: Ch.mo Prof. Maria Catia Sorgato

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

Dottorando: Emanuela Leonardi

Index

Figure Index

Table Index

List of Publications

Journal articles

Leonardi E, Andreazza S, Vanin S, Busolin G, Nobile C, Tosatto SC. A computational model of the LGI1 protein suggests a common binding site for ADAM proteins. PLoS One. 2011 Mar 29;6(3):e18142.

Leonardi E, Martella M, Tosatto SC, Murgia A. Identification and in silico analysis of novel von Hippel-Lindau (VHL) gene variants from a large population. Ann Hum Genet. 2011 Jul;75(4):483-96.

Striano P, Busolin G, Santulli L, Leonardi E, Coppola A, Vitiello L, Rigon L, Michelucci R, Tosatto SC, Striano S, Nobile C. Familial temporal lobe epilepsy with psychic auras associated with a novel LGI1 mutation. Neurology. 2011 Mar 29;76(13):1173-6.

Brini M, Di Leva F, Ortega CK, Domi T, Ottolini D, Leonardi E, Tosatto SC, Carafoli E. Deletions and mutations in the acidic lipid-binding region of the plasma membrane Ca2+ pump: a study on different splicing variants of isoform 2. J Biol Chem. 2010 Oct 1;285(40):30779-91.

Benetti E, Caridi G, Malaventura C, Dagnino M, Leonardi E, Artifoni L, Ghiggeri GM, Tosatto SC, Murer L. A novel WT1 gene mutation in a three-generation family with progressive isolated focal segmental glomerulosclerosis. Clin J Am Soc Nephrol. 2010 Apr;5(4):698-702.

Leonardi E, Murgia A, Tosatto SC. Adding structural information to the von Hippel-Lindau (VHL) tumor suppressor interaction network. FEBS Lett. 2009 Nov 19;583(22):3704-10.

Conference Abstracts/Posters

Leonardi E., Giollo M., Tosatto S.C.E.. Prediction of Disease Phenotypes from Genome Sequencing Data Using a Network Approach. Critical Assessment of Genome Interpretation (CAGI) Meeting, San Francisco (CA), December 9-10, 2011

Giollo M., Leonardi E., Tosatto S.C.E.. A Residue Interaction Network Approach to Predict Mutation Effects. Critical Assessment of Genome Interpretation (CAGI) Meeting, San Francisco (CA), December 9-10, 2011

Leonardi E., Cinelli M., Vanin S., Nobile C., Tosatto S.C.E. Shaping the LGI1 protein interaction network. 36th FEBS International Congress. Biochemistry for tomorrow Medicine. Torino, Italy, 25-30 June, 2011

Leonardi E**.**, Murgia M., Tosatto S.C.E.. Adding structural information to the von Hippel-Lindau (VHL) tumor suppressor interaction network. European Conference of Computational Biology (ECCB10) Stokholm, Sweden, September 26-29, 2010

Sommario

Alterazioni genetiche sono state identificate per molte malattie di natura genetica, ma in molti casi i meccanismi molecolari che contribuiscono all'insorgere della malattia non sono ancora chiari. Lo studio degli effetti delle mutazioni a livello della proteina permette di chiarire i processi biologici coinvolti nella malattia e il ruolo della proteina in essa. La bioinformatica può aiutare a affrontare questo problema rappresentando il punto di connessione tra diverse discipline quali la clinica, la genetica, la biologia strutturale e la biochimica.

In questa tesi ho impiegato un approccio computazionale per affrontare l'analisi di alcuni esempi di proteine di interesse biomedico, integrando diverse risorse di dati e indirizzando la ricerca sperimentale e clinica. Strutture proteiche determinate sperimentalmente o mediante il modelling molecolare sono state utilizzate come base per determinare la relazione tra struttura e funzione, essenziale per ottenere informazioni sulla correlazione genotipo-fenotipo. Le proteine prese in esame sono state inoltre analizzate nel loro contesto, considerando le interazioni che avvengono con altre proteine o ligandi nei diversi compartimenti cellulari. I risultati dell'analisi bioinformatica sono stati poi utilizzati per formulare ipotesi funzionali che in alcuni casi sono state verificate e confermate sperimentalmente da altri gruppi di ricerca. Le mutazioni identificate nei geni codificanti per le proteine in esame sono state valutate per il loro impatto sulla struttura e funzione della proteina utilizzando numerosi metodi di predizione disponibili online. Le diverse applicazioni descritte in questa tesi hanno fornito l'idea per lo sviluppo di nuovi approcci computazionali per lo caratterizzazione strutturale e funzionale di proteine e dei loro mutanti. Si è visto che la predizione migliora utilizzando un ensemble dei diversi metodi di predizione disponibili. Inoltre, per la predizione degli effetti di mutazioni è stato ideato un nuovo approccio computazionale che utilizza le reti di interazione tra residui per rappresentare la struttura proteica. Questi metodi sono stati utilizzati anche nell'analisi di dati genomici originati da nuove tecnologie di sequenziamento. Questo ambito necessita di nuove strategie di indagine per l'individuazione di poche varianti causative in un'enorme quantità di

varianti identificate di dubbio significato. A questo scopo viene proposta una strategia di analisi che utilizza informazioni derivanti dalle reti di interazioni proteiche.
I nuovi approcci formulati in questa tesi sono stati applicati e valutati ad un nuovo esperimento internazionale, chiamato Critical Assessment of Genome Interpretation (CAGI), fornendo in alcuni casi ottimi risultati.

Abstract

Many gene defects have been associated to genetic disorders, but the details of molecular mechanisms by which they contribute to the disease are often unclear. The study of mutation effects at the protein level can help elucidate the biological processes involved in the disease and the role of the protein in it. Bioinformatics can help to address this problem, being the connection between different disciplines including clinical, genetics, structural biology, and biochemistry.

By using a computational approach I tackled the analysis of some examples of biomedical interesting proteins integrating various sources of data and addressing experimental and clinical investigations. Experimentally defined structures and molecular modelling were used as a basis to determine the protein structure-function relationship, which is essential to gain insights into disease genotype-phenotype correlation. Proteins have been further analyzed in their context, considering interactions that they take in specific cellular compartments. The results have been used to formulate functional hypotheses, which in some cases have been tested and confirmed by further investigations performed by cooperation groups. Mutations found in genes encoding these proteins have been evaluated for their impact on the protein structure and function by using several available prediction methods. These studies provided the idea for developing novel approaches, using residue interaction networks and an ensemble of methods. A novel strategy has been also designed to evaluate genomic data obtained by next generation sequencing technology. This consists in using available resources and software to prioritize rare functional variants and estimate their contribution to the disease. The novel approaches developed in this thesis have been applied and assessed at the Critical Assessment of Genome Interpretation (CAGI) experiment in 2011, providing in some cases very successful results.

1. Introduction

The identification of genetic variations determining human phenotypic variations, especially causing diseases, is a fundamental goal in human genetics. Genetic disorders associated with the functional disruption of single genes by a variety of genomic alterations have been recognized for some time. Currently, the connection between genotype and phenotype have been reported for approximately 3,000 Mendelian disorders (Online Mendelian Inheritance in Man). On the other hand, Genome-Wide Association Studies (GWAS) have been extensively applied to discover the genetic basis of common multigenic, complex diseases identifying associations between ~1,300 loci and ~200 diseases or traits (Catalog of Published Genome-Wide Association Studies at US National Human Genome Research Institute). However, causal variants, which account for the associations with the trait under study, have been identified only for a small fraction of these loci. Even for many of rare Mendelian diseases, the causal variants remain to be discovered.

The recent advent of next generation sequencing and high-throughput technologies has added a new dimension to genome research by generating a massive amount of data. Research is revealing the spectrum of extensive genotypic variation among human genomes and its association with a broad range of human phenotypes. The study of human genetic disorders is changing and in the near future exome and genome sequencing will probably replace the traditional approaches for gene discovery and clinical testing [1-3]. Exome sequencing has been proven a promising approach to discover rare causal variants and candidate genes for many undiagnosed rare Mendelian diseases [4-8]. Additionally, it is being tailored to investigate the contribution of rare alleles on the hereditability of complex diseases and health-related traits [6, 9].

These advances can be translated into improved clinical management. Characterization of novel rare variants may assist in the discovery of novel disease genes. Studying pathways in which these are involved may explain the pathogenic mechanisms underlying the disease. This provides new opportunities to identify novel therapeutic targets leading to therapeutic drugs and eventually novel biomarkers to improve disease

predictions. Furthermore, the knowledge of individual predisposition to diseases (e.g., through genetic profiling) allows the development of personalized approaches for diagnostics and therapeutic optimization (Fig. 1). However, to gain benefit from the interpretation of genomic data for health care, we need to know how these variants contribute to the phenotype of the individual [10].

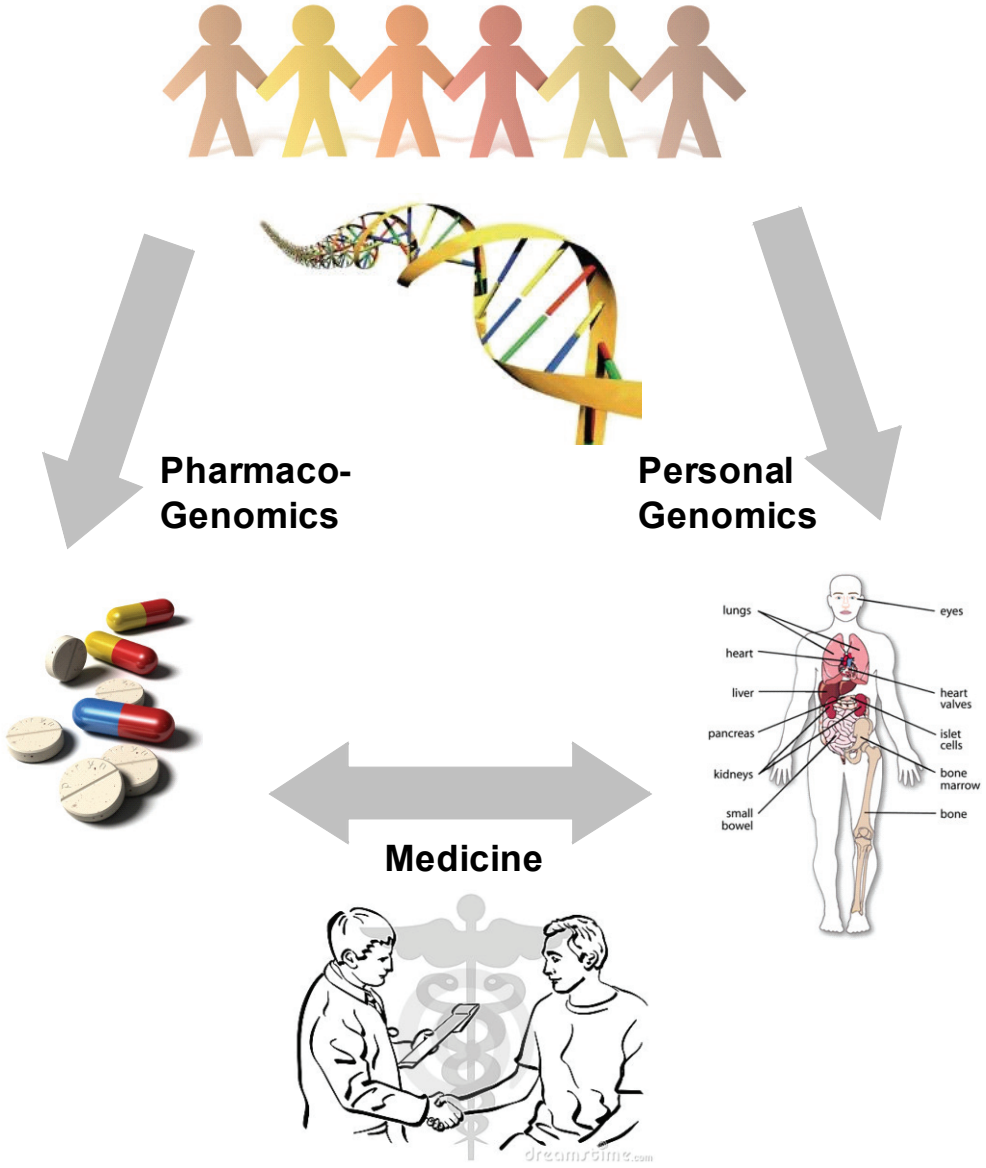


Figure 0.1. Personalized medicine.

Structural genomic alterations such as large deletions/insertions or frame shift, non sense, and splicing mutations are thought to result in non-functional proteins. More difficult to explain is the effect of single nucleotide polymorphisms (SNPs). The effect

of most of these genetic variations is still unknown, but SNPs occurring in or close to a gene can affect its expression or the function of its protein product. In particular, non-synonymous or missense variants (nsSNP) alter the coding sequence causing a change in the corresponding protein that may have drastic phenotypic effects if the structure or the function of the protein is affected [11-12]. The importance of nsSNPs in human is described by the fact that half of these genetic changes are known to cause human monogenic Mendelian diseases [13-14], representing a great resource for understanding disease mechanisms. To characterize the phenotypic effect of genetic variants, a detailed analysis of the structure and function of the protein is essential. Indeed, one of the major molecular pathogenic effect of nsSNP is the alteration of the protein structure which affects protein stability [15-18]. The mutant protein can lose the ability to fold and, recognized as non-native in the endoplasmic reticulum, will be removed and degraded by the quality control machinery. Alternatively, the variant may cause only local fold instability. This may have a direct impact on the functional elements of the protein such as the active site, modification sites, nucleic acid, protein or ligand binding sites. These mutations can also affect other functional elements located in unstructured regions of the protein including cellular localization signals, linear motifs for globular domain binding, or post-translational modification sites. It is expected that many different effects on protein structure and function can result from an amino acid substitution [19]. Experimental characterization of the impact of each nsSNP could therefore be laborious and becoming unfeasible, as rapidly improving sequencing and genotyping technologies continue to generate increasingly large number of genetic variations. The major repository of human SNPs, dbSNP build 132 from October 2010 contained 143 million human SNPs, 30 times higher than present in dbSNP build 106 from 2002 [20]. The 1000 Genomes Project intends to sequence 1000 genomes, hence the volume of genetic data is rapidly growing [21]. However, many computational approaches have been developed to support the study of proteins and to understand molecular effects of genetic variations. Bioinformatics is now needed in most biomedical research fields. Its support is essential to address hypothesis-driven experiments or on the prioritization of multiple hypothesis testing.

The discovery of a novel disease-related gene is accompanied by the characterization of the protein structure-function relationship, which is essential to understand its

contribution to the illness. *In silico* analysis of proteins of biomedical interest can be used to define protein structure and to detect residues or regions crucial for protein function. Knowledge of the protein structure, either experimental or through modelling, can be used to pinpoint finer details, such as the protein domain or segment that mediates interactions. These insights may also be helpful in guiding the design of further experiments to investigate protein function [22-23]. Detailed analysis of known proteins will also serve to elucidate the single pieces involved in regulatory networks at the molecular level, to formulate hypotheses that may explain genotype to phenotype correlation of involved genes. In particular, as mentioned above, the effects of single missense mutations can be predicted in terms of protein stability changes and their impact on interaction partners [24]. In this context, a single protein can be considered as a component in a network. Specific alterations of a protein will then be associated to effects of particular network modules. By studying protein-protein interactions, it will be possible to highlight similarities and differences between apparently unrelated mutations which can be correlated to specific clinical features on a higher level [25-27].

One useful application of bioinformatics research to genetics is the development of computational methods to predict the functional effects of genetic variants. There is a growing body of literature focused on the identification of potentially deleterious mutations and how to distinguish them from neutral substitutions [28-31]. Current prediction methods are based on evolutionary information or combine phylogenetic information with sequence properties and annotations from biological databases. Some methods use structural data which can improve the accuracy of the prediction but their application is limited due to the small number of available structures [32]. In the medical field, these methods can assist in the interpretation of uncharacterized mutations in genes involved in both monogenic and multigenic disorders [4, 33]. Further, exome and genome sequencing of human individuals will lead to the discovery of many previously unknown sequence variants. The main challenge of computational approaches is to predict few deleterious variants among the extensive background of non-pathogenic polymorphisms and help researchers to prioritize SNPs for additional investigations [34]. In this context, there is a strong demand for efficient and accurate bioinformatics tools to classify disease mutations. For available variant prediction methods, the community needs to understand the appropriate confidence level they

should have, and which approaches are most suitable to a particular application. To assess computational methods predicting the functional impact of genome variations, the international community in 2011 organized the experiment called *Critical Assessment of Genome Interpretation* (CAGI, http://www.genomeinterpretation.org/). This is a blind test similar to another competition started in 1994 named the Critical Assessment of protein Structure Prediction (CASP), which had the aim to improve ability to predict protein structures from their amino acid sequences. The goal of CAGI is to accelerate the progress on computational methods for the interpretation of genetic variations. It also wants to test the usefulness of current mutation prediction methods for diverse applications such as predicting the level of enzyme activity from genetic data or the probability of a variant in an intermediate-risk cancer gene to belong to a patient or a control [35]. Other challenges required a major effort and aimed to explore new computational approaches to manage and interpret the large amount of genomic data accumulating with the advent of next generation sequencing technology. In particular, the ability to sequence entire genomes introduces new challenges that the community needs to address. These include the prediction of some traits such as the blood cell types, the predisposition to common diseases such as cancer and diabetes, or knowing the individual response to drugs. Although genome or exome sequences are not yet used in medical practice, we are working to make personalized medicine a reality.

1.1. Objectives

The research conducted in this thesis has involved the application of computational approaches targeted mainly to the study of proteins of biomedical interest. The choice of proteins to study was made on the basis of the presence of an experimental group working or interested in investigating the biological or medical problems. This allowed to test hypothesis that emerged from the *in silico* analysis of the proteins by experiment or clinical investigation. In some cases, the bioinformatics findings were used to formulate hypotheses to interpret data produced by biologists and medical researchers. The various subjects I studied gave me the opportunity to apply diverse computational approaches for various types of proteins with different structural (globular,

transmembrane, and repeat proteins) and functional characteristics (ubiquitin ligase, transcription factors, enzymes, channels). In particular, the approach I adopted in the analysis of each protein followed a workflow that explores the sequence-structure-function relationship up the level of the protein interaction networks.

A set of methods or workflow was designed with the intent to incorporate the use of computational methods in the regular practice of a laboratory studying the molecular basis of genetic diseases. Thus, analysis of the proteins was addressed from the characterization of variants identified by genetic testing whose clinical significant had to be established. For some variants, case-control, segregation, and family history can provide strong evidence of direct association with the disease. However, when genetic data is incomplete, *in silico* analysis of the protein or gene sequence can provide further evidence. In order to classify these variants, a large number of different computational methods have be employed.

The aim was also to evaluate the power of available software on protein structure function prediction and on the interpretation of human genetic variations. Often authors present their software as the best solution demonstrating good accuracy in the prediction of a limited set of experimentally provided data. However, the performance on a specific protein can reveal some weakness of the method. To this end, I participated in the international CAGI experiment which aims to assess computational methods for genome interpretation. This allowed evaluating the different approaches I explored for mutation prediction in diverse applications. Furthermore, I contributed to the development of an approach to predict phenotypes from exome sequencing data.

1.2. Outline

This thesis is organized in 10 chapters followed by a summary. Chapter 2 describes computational methods that can be applied to study the protein structure-function relationship and for the interpretation of genetic variations. The following chapters are divided in two main sections. The first section is composed of chapters 3 to 8, presenting applications of bioinformatics on different proteins of biomedical interest. The second section, corresponding to chapter 9, describes the CAGI experiment and the

results from the participation in different challenges proposed by CAGI in 2011. The last chapter summarizes the work accomplished, discussing the contribution of each chapter.

It should be noted that the chapters of the first section are based on published work that required experimental and clinical data from other research groups. For some part of the work conducted during the CAGI experiment, I required the expertise of my colleagues in the BioComputing group. Thus, for part of the thesis I have used both "I" and "we" to distinguished between my own and shared work. The contributions of each chapter are briefly summarized in the following:

Chapter 3 is based on *Benetti E, Caridi G, Malaventura C, Dagnino M, Leonardi E, Artifoni L, Ghiggeri GM, Tosatto SCE, Murer L. A novel WT1 gene mutation in a three-generation family with progressive isolated focal segmental glomerulosclerosis. Clin J Am Soc Nephrol. 2010 Apr;5(4):698-702*. In this work I analyzed by bioinformatics tools the effect of a novel amino acid substitution on the WT1 protein structure. The effect of the mutation was evaluated through the ability of WT1 to regulate gene expression. The work also allowed to formulate a hypothesis about WT1 function in maintenance of the correct cytoskeletal architecture, which is important for integrity of the filtration barrier in kidneys.

Chapter 4 is based on *Leonardi E, Murgia A, Tosatto SCE. Adding structural information to the von Hippel-Lindau (VHL) tumor suppressor interaction network. FEBS Lett. 2009 Nov 19;583(22):3704-10.* In this work I presented the structural characterization of known interactions of the VHL protein. This allowed better understanding of VHL function in several pathways involved in tumor formation.

Chapter 5 is based on *Leonardi E, Martella M, Tosatto SCE, Murgia A. Identification and in silico analysis of novel von Hippel-Lindau (VHL) gene variants from a large population. Ann Hum Genet. 2011 Jul;75(4):483-96.* In this work I analyzed the effect of novel VHL variants identified in individuals with a clinical diagnosis ranging from von Hippel-Lindau syndrome to sporadic potentially VHL-related tumors. The impact of VHL mutations on the ubiquitin-mediated degradation process is also discussed.

Chapter 6 is based on two works: *Leonardi E, Andreazza N, Vanin S, Busolin G, Nobile C and Tosatto SCE. A computational model of the LGI1 protein suggests a common binding site for ADAM proteins. PLos ONE 6(3): 2011 March 29,;6(3):e18142,* and

Striano P, Busolin G, Santulli L, Leonardi E, Coppola A, Vitiello L, Rigon L, Michelucci R, Tosatto SCE, Striano S, Nobile C. Familial temporal lobe epilepsy with psychic auras associated with a novel LGI1 mutation. Neurology. 2011 Mar 29;76(13):1173-6. This work had two goals: studying the structure-function relationship in the LGI1 protein and studying known LGI1 mutations on protein structure and function. I used the predictions to formulate a hypothesis about the protein function in the synaptic transmission of neural signals. Furthermore, I found a genotype-phenotype correlation for some mutations which is being confirmed by experimental and clinical findings.

Chapter 7 is based on *Bello L, Melacini P, Pezzani R, D'Amico A, Piva L, Leonardi E, Soraru' G, Palmieri A, Smaniotto G, Gavassini B, Vianello A, Bertini E, Angelini C, Tosatto SCE, Torella A, Nigro V, Pegoraro E. Cardiomyopathy in patients with POMT1-related congenital and limb-girdle muscular dystrophy* which has been submitted for publication to the *European Journal of Human Genetics* and is still under review at the time of writing. In this work I analyzed the effect of POMT1 mutations on the enzymatic activity of the protein.

Chapter 8 is based on *Brini M, Di Leva F, Ortega CK, Domi T, Ottolini D, Leonardi E, Tosatto SCE, Carafoli E. Deletions and mutations in the acidic lipid-binding region of the plasma membrane Ca2+ pump: a study on different splicing variants of isoform 2. J Biol Chem. 2010 Oct 1;285(40):30779-91*. In this work I analyzed the structure-function relationship of the plasma membrane calcium ATPase (PMCA) protein. In particular, I studied the role of structural regions in the regulation of pump activity mediated by acidic membrane phospholipids (PL) and Calmodulin.

Chapter 9 describes the participation in the CAGI experiment in 2011. This chapter explains the use of novel computational approaches for mutation effect prediction in different biological applications. In particular, I present the applications of a method using residue interaction networks and an ensemble method. The main idea of the novel approaches arises from the studies presented in the previous chapters. In the first step of the work, I contributed the integration of biological information in the development of new prediction methods. This chapter also describes a new approach to predict phenotypes from genomic data. In particular, I designed a model which can be used in the identification of candidate genes causing Mendelian and complex diseases. Chapter

10 delineates the main findings obtained from the previous chapters, describing their relevance in the biological and medical field.

2. Predicting Protein Function from Sequence and Structure

Protein function can be defined at different interdependent levels and may be classified in three main categories: molecular function, biological process and cellular component. Molecular function indicates the activity of the protein at a molecular level, such as enzymatic activity. Biological process describes a set of molecular functions which jointly operate in a living units (cells, tissues, organs, and organisms) such as apoptosis. Cellular component is the compartment of the cell where the protein exerts its function, which may be an anatomical structure such as the nucleus or endoplasmic reticulum or a protein complex such as the proteasome. To perform these functions, the protein uses functionally distinct regions that can be recognized at the sequence or structural level. Several computational methods have been developed to characterize the structure or function of these regions.

In this chapter I present a workflow that shows how to apply computational tools to retrieve functional information from protein sequence and structure (Fig. 2.1). Since structure determines function and we can use it to predict function, part of this chapter covers generation and inference of protein structure. A separate section is also dedicated to the functional characterization of an intriguing class of proteins presenting intrinsic structural disorder. All information derived from this approach can be used for the interpretation of genetic variants. Beyond the description of how non-synonymous SNPs affect protein function, I provide an overview of computational methods to predict the effect of SNPs on protein stability and distinguish pathogenic from neutral nsSNPs. The characterization of protein functional regions provided a rationale for designing experiments aimed to understand the molecular basis of protein function. Furthermore, for mutations mapping to these regions, interpretation of their impact on protein function allowed to formulate hypotheses, which can be verified by experimental studies.

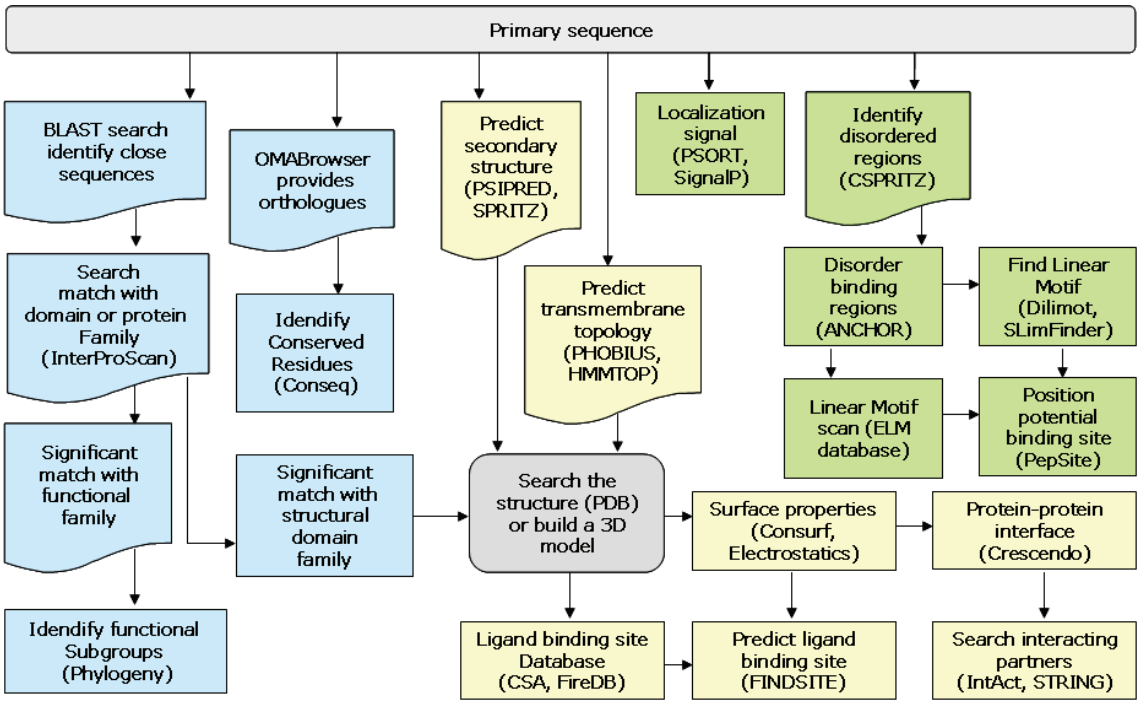


Figure 0.1. Schematic workflow for the analysis of protein functions.
Information can be predicted from sequence or structure.

2.1. Sequence analysis

Sequence database searches

Analysis of a protein starts by obtaining its reference amino acid sequence. Protein sequences can be retrieved from **GenBank** at the NCBI, **Ensembl**, or **UniProt** [36]. The UniprotKB database (protein knowledgebase) collects protein sequences in two sections: SwissProt and TrEMBL. The latter provides automatic annotation for protein sequences derived from the translation of the corresponding nucleic acid sequences deposited in GenBank. The SwissProt section is instead continuously updated with manually curated sequences. Proteins are carefully annotated with functional information including biological processes in which they are involved, cellular localization, protein family and evolutionary information. UniProt provides access to many links to other sources and facilitates the collection of knowledge and classification of an interesting protein.

The simplest way to obtain the identity of a protein and its functional annotation is to use homology information. The general assumption is that sequences with high similarity evolved from a common ancestor and thus share the same function. **BLAST** or **PSI-BLAST** [37] are widely used to search sequence databases. The best hits (lowest E-value) give us the most similar proteins, but the similarity may sometimes refer only to a part of the sequence, usually corresponding to a specific protein domain.

Multiple sequence alignments

To identify relevant functional regions, alignment of the target protein with homologous sequences is very informative. Furthermore, distinguishing between orthologous and paralogous sequences allows the identification of function-discriminating residues. However, the recent explosion of large scale sequencing projects results in an increasing numbers of automatically annotated sequences, with the corresponding possibility of errors [38]. Hence, selection of homologous sequences can be facilitated using curated databases. The Orthologous MAtrix (**OMA**) browser is a web interface offering the search of protein sequences from 1,000 species. Orthologs of a given protein can be download as a group of related sequences or as multiple sequence alignments [39]. Other databases are family specific such as the **KinBase** database, where kinase sequences can be retrieved for different species [40]. Once a set of sequences is obtained, the alignment is built using **CLUSTALW** [41] or **MAFFT** [42] and manually curated using a sequence editor like **Jalview** [43]. The alignment can be modified on the basis of structural information or experimental findings, avoiding gaps in conserved secondary structure elements.

Phylogenetic analysis

The evolutionary relationship between different proteins is further investigated by phylogenetic analysis. Neighbor joining is a fast method based on clustering of distances, while a maximum likelihood approach can be used when we need to consider information from each position. To build phylogenetic trees, a widely used software based on the maximum likelihood approach is PHYLM [44]. The most widely used substitution matrix is the JTT matrix and robustness of the tree topology is usually estimated by nonparametric bootstrap resampling (BT).

Functionally relevant residues

In order to visualize evolutionary conservation for each amino acid position in the sequence or structure, we can use the **ConSeq** or **Consurf** web servers respectively [45]. The graphical view helps to explore functionally or structurally important regions in the protein, using as input a curated alignment and a phylogenetic tree.

Functional classification systems

Functional classification of proteins have been mainly derived with the Gene Ontology (**GO**) [46]. This describes molecular function, biological process, and cellular component of a protein using standard terms, coded by numbers. Alternatively, the Enzyme Commission (**EC**) classification system assigns a number defining classes and subclasses of enzymes. These systems provide a general classification which can be very useful in processing large numbers of proteins, but requiring careful interpretation in specific cases. Some enzymes belonging to the same class have significant differences in reaction mechanisms. This is emerging in a new classification of enzymes based on ligand and mechanistic similarities [47].

Domain architecture

The web interface of BLAST provides a graphical alignment view with homologous sequences integrated from the Conserved Domain Database (**CDD**). These annotate the protein sequence with the location of conserved domains and functional sites [48]. CDD reproduces a pre-computed conserved domain annotation calculated by the Reverse Position Specific BLAST (RPS-BLAST) algorithm importing domain and protein family alignments from **Pfam** [49], **SMART** [50], **COG** [51], **TIGRFAM** and the NCBI protein clusters database. While COG collect orthologous family, the other databases are example of Family-based resources. These database classify proteins with multidomains or individual protein domains in evolutionary families. Pfam is particularly useful to retrieve domain boundaries in a protein and the assignment of the function to the homologous of the same family is manually curated. **InterProScan** interface provides a easy way to obtain signatures, domain family, and functional sites from eleven different databases [52].

Sequence motifs search

Rather than using the whole sequence, protein function can be predicted by similarity with short stretches of conserved protein sequences, of 10-20 residues, referred to as sequence motif or signature. These sequences are important for the biological function of a group of proteins and contain enzyme catalytic sites, prosthetic group attachment sites, residues coordinating metal ions, or cysteines involved in disulfide bonds. In the **PROSITE** database, these signatures are defined as regular expressions or pattern, rules, or profiles on the basis of the prediction methods used. They are manually curated, selecting from relevant biological examples [53]. Other databases dedicated to motif searching are **BLOCKS** [54] and **PRINTS**[55]. Sites of posttranslational modification like phosphorylation, acetylation, and glycosylation sites identified by high-resolution mass spectrometry are collected in the **PHOSIDA** database [56]. The web interface also offers a wide range of analysis tools to predict modification sites. The **ELM** database, in addition to modification sites and sequence motifs from PROSITE, contains linear motifs for protein binding sites [57]. Other resources can be protein specific, such as the **Calmodulin target database** (http://calcium.uhnres.utoronto.ca/ctdb/) collecting four classes of calmodulin binding motifs (IQ, 1-10, 1-14, and 1-16) and other motifs identified from available complex structures involving calmodulin. Several motifs are also defined by various investigators through sequence homology with existing calmodulin binding motifs. Cellular localization of a protein can be predicted with **PSORT II** [58], which searches for potential ER retention or nuclear localization signals. **SignalP** [59] predicts potential peptide cleavage sites in the N-terminal sequence, which are secretory pathway signals.

Secondary structure

When an experimental protein structure is not available, predicting the secondary structure can be the first step to classify its structural components or to model the globular domains. **PSIPRED** [60] (http://bioinf.cs.ucl.ac.uk/psipred/) uses profiles calculated by PSI-BLAST. It is implemented with neural networks to calculate the propensity for secondary structure of each residue in a window of 15 amino acids. **Porter** is another server for secondary structure prediction using bidirectional recurrent neural networks [61]. This approach allows for dynamic window extension during the

assessment process, which is particularly important for prediction of distant beta-strand forming residues. Each predictor divides secondary structure in three classes: alpha-helix (H), extended (E) or beta-strand, and coil (C). The average prediction accuracy of these methods is around 80%, and using a consensus of prediction methods can improve the detection of conserved secondary structure elements.

Transmembrane prediction

Membrane proteins are involved in several vital biological processes, with the main categories being cell adhesion proteins, membrane receptors, transport proteins and enzymes. The presence of a transmembrane segment can be predicted and is indicative of integral membrane proteins. These present two main folds as either alpha-helical bundles or β-barrels with similar amino acid composition. Traditional methods used information about amino acid composition and hydrophobic patterns to predict transmembrane segments. More recent methods, such as **TMHMM** [62] and **HMMTOP** [63]**,** apply Hidden Markov Models as machine learning methods to deduce rules on transmembrane structure. The advantage of these is the possibility to restrict the length of the putative transmembrane segments. Another method based on HMMs is **Phobius** [64], which distinguishes between N-terminal signal peptides and transmembrane segments. These methods have been extended with evolutionary information in the recently developed versions **Prodiv-TMHMM** and **Poli-Phobius** [65]. Consensus prediction of membrane topology is further a useful strategy to improve the reliability of the prediction. Several web servers have been developed to this aim. **TOPCONS** [66] combines the prediction of five methods: **OCTOPUS** [67], **pro-TMHMM** and **Prodiv-TMHMM** [68], **SCAMPI-single** and **SCAMPI-multi** [69]. **SCAMPI** predicts topologies for single and multiple-sequences using a position-specific membrane insertion propensity scale [70]. **OCTOPUS** uses a combination of HMMs and neural networks and is able to predict re-entrant regions and transmembrane hairpins.

2.2. Generating and inferring structure

Protein structure databank (PDB)

The major database of experimentally determined protein structure is the RCSB Protein Data Bank (**PDB**) [71], which is a member of WorldWide PDB. It consists of three main organizations, PDBe (UK), PDBj (Japan), and RCSB (USA). It aims to maintain a global and uniform repository of large biological molecules, including protein and nucleic acid structures. These are deposited as files containing the atomic coordinates determined by X-ray crystallography or nuclear magnetic resonance (NMR). The file can contain more than one chain if the structure represents a protein complex or homopolymer. Each record is integrated with links to other databases such as Pfam [49], SCOP [72], and CATH [73-74] (for SCOP and CATH description see paragraph 2.4). These are useful to retrieve structures similar to the target PDB structure on the basis of structural or evolutionary relationships with the PDB fold. Another integrated resource is the **DSSP** database, containing secondary structure assignments for all PDB protein entries [75].

Visualization tools for 3D structure

Rasmol [76], **Jmol** [77], and **Pymol** (Schrödinger LLC) are the most widely used open source programs for interactive molecular visualization of protein structures. The PDB database allows visualization of the structure using Jmol. Both Jmol and Rasmol are integrated in many web services dedicated to protein structure analysis through an applet. **Pymol** is a widely used molecular visualization system written in Python that can run on different operating systems. It can be used to explore structures in details and produces high quality images for publication. Furthermore, it can be used to visualize electrostatic potential surfaces calculated by the Adaptive Poisson-Boltzmann Solver (APBS) program [78]. Another interesting visualization system is **UCSF Chimera** [79]. One of most important advantage of this tool is the use of **structureViz**, a Cytoscape (see paragraph 2.4) plug-in linking the visualization of molecular structures with biological networks such as residue interaction networks [80] (see paragraph 2.6).

3D structure superposition

The superposition of protein structures with similar folds can be performed in order to explore conserved and variable regions between them. There are several tools for this aim. **CE** (Combinatorial Extension) [81] calculates the best possible alignment between two structures using sequence fragments in a way similar to contact maps to establish structure similarity. Recently, a new version of the software has been published, referred to as **CE-MC**, which aligns multiple structures based on Cα-coordinate distances. The alignment performed by CE algorithm is further optimized by Monte Carlo optimization [82]. Another used approach, called Multiple Structural AligNment AlGorithm (**MUSTANG**), is based on a progressive pairwise alignment heuristic [83]. This software has been demonstrated to perform better with distantly related proteins or with proteins that undergo conformational changes.

Structure annotation tools

To integrate and publish information deriving from different analyses we can use **ESPript** [84-85]. This tool allows to represent aligned sequences with annotation for secondary structure, solvent accessibility, intermolecular contacts, modification sites and other user-supplied markers. The overall representation of annotated sequences is especially appreciated for publication.

Homology modeling

Homology or comparative modeling is used for the construction of three-dimensional models of a target protein starting from its amino acid sequence and a structure of a protein, called template, with at least 30% sequence identity. Building a homology model comprises three main steps: identification of structural template(s), alignment of target and template sequence(s), and model building. The template search can be performed using the **PDB-BLAST** protocol (http://protein.bio.unipd.it/pdbblast/). Here, a profile built from the sequence of the target is used to search possible templates in the PDB database, allowing identification of even distantly related protein structures. The protocol is integrated in the **HOMER** model building server (http://protein.bio.unipd.it/homer/). If more than one template is proposed, it is possible to choose the structure with the best resolution or, after superposition of the structures,

the one better representing the domain we want to predict. The structural alignment between different templates can also be useful to highlight conserved and variable regions and improve the target-template alignment. The critical modeling step consists in the construction of an accurate alignment between the target and template proteins. In this phase it is possible to use several previously described alignment tools. The manual refinement of the alignment is recommended especially if further information is available from structural analysis and experiments.

Finally, alignment and structure template are used as input for specific homology modeling software. **Modeller** is widely used to this aim [86] and uses comparative modeling by satisfaction of spatial restrains. Many other tools exist as free web servers such as **SWISS-MODEL**, accessible from the Expasy web server [87]. HOMER is an in house produced software used for most of the studies presented here. A manual and an automatic protocol for template selection are available, with automatic selection using PDB-BLAST protocol. HOMER can perform additional tasks such as de novo loop modeling and side chain optimization. The loops are modeled using an algorithm based on a divide and conquer approach, named **LOBO**. The method generates a ranked set of possible conformations of the loop with predicted quality measured in terms of RMSD [88]. Next, the **side chain placement** use a rotamer-based method **SCWRL,** which chooses the energetically favored amino acid conformation. The model obtained from HOMER is accompanied by a per-residue energy profile calculated with FRST [89]. This gives a first indication of the model quality since regions with high energy may contain errors. The model can be energetically minimized using **GROMACS**, a molecular dynamics simulation software [90].

Fold recognition

Fold recognition methods identify similar structures with low sequence similarity using secondary structure or accessibility predictions. The **BioInfoBank Meta Server** (http://meta.bioinfo.pl/) allows identification of templates belonging to the same fold class of the target protein. The server uses the prediction of different fold recognition servers. The predicted models are evaluated by **3D-Jury**, scoring them on the basis of their similarity to other models [91]. Output includes the PDB code of each hit, the alignment and the similarity score calculated from every server. The template sequences

are annotated with their SCOP [72] and FSSP [92] classification (see paragraph 2.4) and compared with the target sequence predictions. A better prediction is obtained by using only the sequence of the domain to be modeled. The target-template alignment can be further used to build the model with other software, possibly after manual modifications derived by further analysis. Meta Server is directly linked to Modeller [86] for model construction.

Membrane protein structure prediction

Determining the structure of transmembrane proteins has long represented a difficult problem since it requires very special experimental conditions for the crystallizing process to succeed. Over the last few years, the number of resolved structures increased considerably, with currently 1550 deposited in **PDBTM**, a database for transmembrane protein structures [93]. However, since the sequence-structure gap for transmembrane proteins is still large, modeling remains a important task. Many protein structure prediction algorithms have been developed for soluble proteins, but they are not designed specifically for membrane proteins. The physical differences between soluble and membrane proteins concern the environment in which the proteins are embedded, implying a different strategy should be adopted for modeling. However, it has been demonstrated that template-based approaches can be successfully applied to membrane proteins. New strategies are based on approaching diverse parts of the protein in different ways, especially to identify the core shared by target and template [94]. Given a template with 30% identity, the target-template alignment should consider both sequence conservation and topology of the transmembrane regions. A consensus approach for topology prediction is recommended since this improves the accuracy for boundary prediction of transmembrane segments. The other steps follow the usual homology modeling approach.

Modeling of repeat proteins

Repeat proteins, such as solenoid-like proteins, are composed of several repeated structural units. Although the single repeats can be highly degenerated in sequence, it is possible to recognize a conserved pattern of residues with specific biochemical characteristics which is responsible for the repeated fold. Proteins containing repeat

domains, such as leucine-rich repeats (LRRs) or β-propeller domains, have very low internal sequence similarity. However, repeat protein modeling has been successfully performed for LRR domains using an approach combining homology modeling and structure-based sequence alignments [95-96]. First, the repeat units can be predicted using **REPETITA** [97]. This program uses the discrete Fourrier transform (DFT) to identify the sequence periodicity of a repeat domain using a sequence profile defined by five numeric scales [98]. The Atchley scales reflect polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge. Usually, since sequence similarity is very low, template selection is performed with a fold recognition approach, making secondary structure prediction very important. A consensus approach is recommended since it improves the accuracy of the prediction [99]. The best template will be selected considering also the corresponding number of repeats. In the case of LRR domains this can affect the curvature of the arch. The conserved pattern defined by each repeat in the template structure is compared with those of the target protein and manually adjusted. Finally, the target-template alignment is used as input for HOMER or other homology modeling software. The following steps are the same for homology modeling including loop modeling, side chain placement, and energetic minimization.

Model Quality

The reliability of the structure models has to be evaluated by model quality assessment tools. The **Qmean** server can be used both for the selection of the best models or to evaluate the absolute quality of a protein model in order to know the reliability of each part of the protein [100-101]. The high quality regions may be further investigated to formulate new biological hypotheses (e.g. prediction of binding sites). Qmean scores range from 0 to 1, with higher values representing more reliable models. Qmean uses the combination of six scoring function terms, with the composite score being based on normalized statistical potential terms describing the major geometrical aspects of protein structures. The output presents the Qmean score, estimated absolute model quality, and the z-score of each terms. Good structure are expected to have z-score around zero (light red to blue regions in the plot). Furthermore, the estimate of residue error is mapped on the structure using a color gradient from blue (more reliable regions) to red (poor quality regions).

2.3. Intrinsically disorder proteins (IDPs)

A large fraction of proteins in both prokaryotes and eukaryotes contain disordered regions. A disordered region can be defined as a highly flexible part found partially or completely extended in solution. Globular structures may present some disordered regions, such as extended random coils or secondary structure elements that are not condensed into a stable globular fold [102]. Recently, a more complete definition has been formulated where intrinsically disordered, or unstructured, proteins (IDP/IUPs) or regions of proteins (IDRs) occupy different conformational states from fully disordered (random coil) and compact states [103-105].

The prevalence of disorder in protein function has been studied by different groups based on the GO annotation, addressing the correlation between disorder and the three classes molecular function, biological process and cellular localization. The studies agree on the identification of high prevalence of disorder in transcription regulation, protein kinases, transcription factor, and DNA binding proteins. Proteins with high levels of disorder are also involved in development, protein phosphorylation, regulation of transcription, signal transduction. The majority localize in the nucleus, with most of these being ribonucleoproteins or forming part of the cytoskeleton [106-109].

Disorder prediction

Disordered regions can be distinguished from ordered ones on the basis of their amino acid composition, as IDPs show low hydrophobicity and high net charge. Several methods have been developed for disorder prediction using both neural networks or support vector machines (SVMs). **DISOPRED2** [108] is trained to identify disordered residues that fail to be crystallized. The SVM uses a sequence profile generated by PSI-BLAST and evaluates the disorder propensity of each residue in a symmetric window of 15 positions. **PONDR** [110] (predictor of naturally disordered regions) is a neural network based on amino acid composition, flexibility and other features derived from the sequence. **SPRITZ** [111] is a web server for the prediction of disordered regions also providing the prediction of secondary structure elements performed by Porter [61]. It uses two specialized binary classifiers for long and short disorder which use both

support vector machines. A recent new version of this software is **CSPRITZ** [112], combining the prediction of three different disorder predictors. These use homology, sequence-only, or structure information. An interesting implementation is that the output of CSPRITZ also indicates ELM motifs mapping to the predicted disordered regions. A different approach is used by **IUPred [113]**, which estimates the total pairwise interaction energy created by a polypeptide chain. The idea is that proteins cannot fold because their amino acids are not able to form stabilizing inter-residue interactions. Thus this method takes into account amino acid composition and local neighbors.

Functional Classification of IDPs

On the basis of the molecular mechanisms involved in IDPs, disordered proteins can be divided into six categories. The first unique category for disordered proteins is the entopic chain. These act by either influencing the localization of attached domains, or generating force against conformational changes [114], as has been demonstrated for the entropic gating in nuclear pore complex [115]. The other categories involve molecular recognition, with disordered regions binding transiently or permanently other proteins or ligands. Transient binding of disordered regions is well demonstrated for linear motifs mediating phosphorylation [116], ubiquitination [117], and acetylation [118]. Chaperones showing a very high proportion of disorder also use transient binding of disordered regions to perform their functions [119]. Three other categories identify disordered proteins as effectors, assemblers, or scavengers. Permanent binding modifies the activity of the partner, or assists in protein complex formation, or stores and/or neutralizes small ligands. The last category includes prions, proteins in which disorder is responsible for their autocatalytic conformational transition [104].

Prediction of function-related structural elements in IDPs

IDPs seem to use transient structural elements to interact with their partners. Preformed structural elements can be predicted by usual secondary structure prediction algorithms [102] with higher accuracy than for ordered proteins and directly correlated with molecular recognition elements (MOREs) or molecular recognition features (MORFs). These elements have been identified by studying complex structures containing one

partner which is shorter than the other. MORFs show local structural preferences and correlate with disorder in the unbound state [120-122]. Iakoucheva and colleagues [107] showed that a decreasing **PONDR VL-XT** [123] disorder prediction score may indicate a functionally important recognition elements. **ANCHOR** [124] (http://anchor.enzim.hu/) is a software dedicated to the prediction of protein binding regions in IDPs. It uses IUPred [113] as disorder predictor and the same pairwise energy estimation approach to predict binding regions. The assumption is that parts of the disordered regions might form stabilizing contacts by interacting with globular protein partners. Since the recognition element could be simply represented by a linear motif, the web server also offers the possibility to complement the search of disordered regions with ELM [57] motif searches. Alternatively, these can be provided by the user or from the Calmodulin Target Database [125] (http://calcium.uhnres.utoronto.ca/ctdb/).

Prediction of short recognition motifs in IDRs

A widely used approach to infer function from IDP regions consists in predicting short linear motifs which are directly related to specific function, such as post-translational modification or binding specific protein domain (e.g. SH3 domain). As mentioned before, CSPRITZ [112] and ANCHOR [124] allow the prediction of disorder in concert with known linear motifs deposited in the ELM database [57]. The presence of unidentified short linear motifs (LMs, ELMs, Slims) possibly involved in protein-protein interactions can be inferred. **DILIMOT** [126] is based on the expectation that a set of proteins with common functional feature (e.g. localization or binding of the same protein) may contain a linear motif in their sequences. The identified motifs are ranked on the basis of over-representation among proteins and conservation across homologous. **SLiMFinder** is another approach predicting shared motifs in a set of protein with common attributes. For best performance, the proteins should have little or no similarity [127-128].

PepSite

When a protein is predicted to bind to a candidate peptide, the potential binding site on its surface can be predicted using **PepSite** [129]. This method uses a position specific scoring matrix (PSSM) derived from known peptide-protein complexes describing the

binding site preferences for each amino acid of the interacting peptide. The surface of a protein is scanned to find candidate binding sites for each residue of the target peptide. The resulting prediction may be useful in the functional characterization of many protein interactions. The discover of protein-peptide binding details provides a guide to better understand cellular mechanisms. Furthermore, since transient interactions are easier to modify chemically, protein-peptide interactions are a promising target for new class of drugs.

2.4. Function prediction for globular proteins

Structural classification

Three main resources are available for the classification of protein structures derived from PDB [71] database: **SCOP** [72], **CATH** [73], and **FSSP** [92]. These classification systems are based on structural, functional, or evolutionary features of the proteins. SCOP [72] (Structural classification of proteins) manually annotates proteins in four different levels: Class, Fold, Superfamily, and Family. While the Class describes the secondary structure composition of the protein, the Fold clusters proteins on the basis of the arrangement and topology of their secondary structure elements. The Superfamily level contains proteins with low identity but strong functional relationships, while proteins with very similar function or structure and at least 30% sequence identity have been clustered together at the family level.

CATH [73] (Class, Architecture, Topology, and Homology) classifies proteins with a semi-automated method, called SSAP (sequential structure alignment program) which searches for structural similarity comparing vectors of C_β atoms between two proteins. Among the four classification levels, class and topology are similar to those in SCOP, while topology identifies structural clusters on the number and spatial connections between secondary structure elements. The homology level contains homologous proteins with similar structure and function.

FSSP [92] is a database for fold classification based on a continuously updated exhaustive pairwise structural alignments of PDB proteins. The resulting classification

is reported in a fold tree generated by hierarchical clustering. Each level represents a unique protein family. FSSP is used by the Meta Server to annotate the template with structural classifications.

Flexibility

Proteins are not static and may undergo conformational changes upon binding to other molecules. Furthermore, a certain flexibility in the structure allows allosteric communication and the correct positioning of domains for substrate binding. Thus, analysis of protein flexibility should be important for function prediction. It can be used to identify conserved deformation patterns in functional mechanisms involved in catalysis, binding, and allostery. The most powerful method to study protein flexibility is molecular dynamics. Since this approach is complex and computationally expensive, coarse-grained methods coupled with simple potentials have been developed [130]. To represent protein flexibility, **FlexServ** (http://mmb.pcb.ub.es/FlexServ/) incorporates three coarse-grained algorithms and is integrated with structural databases. The results can be visualized in 3D models using a JMol applet or 2D plots [131].

Protein surface representation

The surface of a protein can be described as either van-der-Waals surface or as solvent accessible surface. In the van-der-Waals representation, protein atoms are represented as spheres with radius equal to their representative van-der-Waals radius. This definition is used in the space-filling model. However, the most commonly used definition is the solvent accessible surface representing a continuous functional surface of the molecule. It is obtained by rolling a water molecule over the van-der-Waals surface, using the centre of the solvent probe as reference. This representation implies that several residues contribute to the properties of the molecular surface. It is widely adopted to infer biological characteristics of the protein surface. The solvent accessible surface is extracted from the PDB file of experimental structures using DSSP.

Surface conservation

The ConSurf server combines two methods, ConSeq and ConSurf, to calculate evolutionary conservation starting from protein sequence or structure respectively [45].

Using only primary sequence as input, the tool discriminates between exposed (e) and buried (b) residues in globular proteins based on evolutionary information. The assumption is that slowly evolving residues should have relevant structural or functional roles depending on their localization with respect to the protein surface. Slowly evolving residues buried in the protein core have a structural role (s), while those solvent exposed have a functional role (f), e.g. protein binding [132]. The ConSurf protocol can instead be used when a protein structure or model is available. In this case, the functionally conserved regions on the protein surface are automatically represented with magenta color and may indicate the presence of a binding or active site. However, the interpretation of conservation is often difficult and extensive similarity does not imply similar function, e.g. in the TIM barrel family. The selection of sequences used to generate the multiple sequence alignment is crucial, and the exclusion of paralogs can be useful to correctly predict specific functional regions [133].

Hydrophobic surface

Non-polar atoms tend to minimize the contacts with surrounding aqueous solvent, representing the so-called hydrophobic effect. Globular proteins present a hydrophobic core and polar surface, and the hydrophobic effect is the driving force determining their structure. This is also the driving force in the stable association of molecules. Therefore, the characterization of hydrophobic patches on the protein surface is indicative of obligate interfaces, such as oligomeric interactions [134]. The identification of surface hydrophobic content has been also used to predict the structure of protein-protein complexes [135].

Electrostatic surface

Regions with high electrostatic potential can be used to predict the location of DNA/RNA binding sites in proteins. Enzyme active sites are also characterized by electrostatic strain which seems to facilitate enzyme catalysis. Therefore, electrostatic surface analysis can be used to predict functional sites. A way to calculate and visualize electrostatic surfaces is provided by Pymol (Schrödinger LLC) and the plugin **APBS** [78]. First, the protein structure has to be prepared for the electrostatic calculation adding atomic charge and radius information. The **PDB2PQR** web server

(http://www.poissonboltzmann.org/pdb2pqr/) can be used to convert PDB files into the PQR format. Different force fields can be selected for calculation. The file PQR is used as input for the Pymol plugin and the electrostatic potential will be plotted on the solvent accessibility surface. The visualization system provided by Pymol is color coded, blue indicating positive and red negative electrostatic potential.

Ligand binding site databases

Protein-ligand binding sites are characterized by the presence of a pocket on the protein surface. Its properties depend on the small molecules that are bound by different family members. The Catalytic Site Atlas (**CSA**) database [136] (http://www.ebi.ac.uk/thornton-srv/databases/CSA/) reports residues that are directly involved in the reaction catalyzed by enzymes. In this database we can find manually annotated catalytic sites derived from the literature and catalytic residues found by homology with annotated sequences. CSA is also used in **FireDB**, a database containing PDB structures and their associated ligands, where residues involved in ligand binding are annotated [137].

Ligand binding site prediction

A different type of methods uses only geometric characteristics to identify cavities on protein surface *de novo*, such as **LIGSITE** [138], **PASS** [139], **SURFNET** [140]. **Q-SiteFinder** uses an energetic approach calculating the higher van-der-Waals interaction potential energy of an interacting probe [141]. The consensus method **metaPocket** combines these methods to improve prediction quality [142]. Ligand binding sites can be predicted with ***firestar*** **[143]**, which uses PSI_BLAST and homology detection by iterative HMM-HMM comparison to retrieve homologous sequences from the FireDB database [137] and MUSCLE [144] to build multiple sequence alignments. The program transfers functional information of ligand binding residues in FireDB to the query sequence on the basis of the conservation between the two sequences [143]. Another predictor using homology information is **FINDSITE** [145]. In this case the software extracts structural information about conserved anchor functional groups rather than for residues accounting for binding specificity. Indeed, among evolutionarily

related but distant protein families, sequence and structure conservation is higher for residues contacting anchor functional groups [145].

Integrated server for structure-informed function prediction (PDBsum)

PDBsum is a web server (http://www.ebi.ac.uk/pdbsum/) providing several structural analyses of PDB entries presented with graphical schemes. It was recently updated allowing users to generate structural analysis for own structure. The sequence of the target protein can be compared with a domain diagram reporting information derived from the available structure and various databases, such as CATH [73]. Catalytic residues, ligand binding residues, PROSITE patterns and disulfide bonds are all indicated in the diagram. In addition to a topology diagram showing the connections between different secondary structure elements, the web page provides a schematic diagram of the protein interface for structures containing more than one chain. The details of interactions between residues across the interface are also represented [146].

Protein-protein interfaces

The analysis and prediction of protein interfaces focuses on non-obligate protein-protein interactions. These involve proteins that can be found in a stable conformation independently in solution. Obligate interactions are those found in oligomeric complexes forming stable interactions through interfaces showing different surface properties. Structural analysis of experimentally determined protein complexes allows to characterize interface properties. These consist of buried cores surrounded by partially accessible specific residues. Few of these residues significantly contribute to the binding affinity and are called “hot spots”. These residues seems to be more frequently represented by aromatic or charged, rather than amphiphylic or hydrophobic, residues [147]. Interaction specificity is determined by the physico-chemical properties of the interacting surfaces which have to be complementary in terms of hydrophobic, charged or polar residues in the surface and between hydrogen bond groups. A protein interface is characterized by several properties and their combination has been used to developed various interface prediction methods. **PPI-pred** applies machine learning to six properties: hydrophobicity, salvation, geometry of interface residues, patch planarity, patch roughness, and solvent accessible surface area of the patch [148]. **Promate** [149]

(http://bioportal.weizmann.ac.il/promate/) in addition to hydrophobicity uses other properties such as atomic distribution, residue conservation, and secondary structure. The **Cons-PPISP** (http://pipe.scs.fsu.edu/ppisp.html) consensus protein-protein interaction surface predictor was trained on native interfaces collected from PDB structure complexes [150]. Sequence profiles and solvent accessibility of each residue and their neighbors in space are used as input for a consensus neural network. The same properties are also used for an empirical scoring function by **PINUP** (http://sparks.informatics.iupui.edu/PINUP/) [151]. The predictions of some of the methods mentioned above are combined in **meta-PPISP** (http://pipe.scs.fsu.edu/meta-ppisp.html) [152]. **Crescendo** (http://www.bioinf.manchester.ac.uk/crescendo/) [153] predicts functional sites on protein surfaces based on a conservation score calculated using amino acid substitution tables derived from particular local environments. This method identifies functional restrains from evolutionary information indicating interaction sites involved in various functions.

Protein-protein interaction databases

Recent advances in high-throughput experimental methods for protein interactions produced an increasing amounts of data and various databases have been developed for their classification. In addition to physical interactions, some databases also provide functional associations or interactions at the domain level. The database of interacting proteins (**DIP**) contains experimentally determined protein interactions obtained from the literature, PDB, TAP-mass spectrometry analysis and other high-throughput methods. The biomolecular interaction network database (**BIND**) includes high-throughput data and protein complex from PDB [154] and can also distinguish different types of interactions. The molecular interactions database (**MINT**) focuses on experimentally verified protein interactions from mammalian organisms [155]. **IntAct** contains data retrieved from the literature manually annotated by experts. The annotations also include experimental methods, conditions and interacting domains [156]. **HPRD** is another database containing data extracted from the literature [157]. The most interesting database is **STRING** [158] which incorporates protein interaction associations for all databases mentioned above and from various other resources including **KEGG** [159], GO [160] and **OMIM** [13] (see paragraph 2.5). Predicted

functional interactions are also included and each interaction is accompanied by confidence scores on the basis of the methods used to find the association. Higher combined scores indicate that more than one method supports the association [161]. However, the database contains many errors because associations are derived by text-mining and need to be verified.

Protein-protein interaction network analysis

The protein interaction databases contain a list of many binary interactions. Graph theory is extensively used to represent these interactions in the context of a particular pathway, tissue, cell, or organism. Some databases such as **IntAct** [156] and **STRING** [161] have incorporated graphs for visualizing dynamically generated network maps. **Cytoscape** [162] is a powerful platform developed to visualize and handle biological networks. Here, protein interactions are represented with nodes and edges as a two dimensional network.

Protein-protein interaction prediction

Experimental approaches studying protein interactions can be complemented by computational methods for their prediction. These methods can be used to choose potential targets for experimental screening or to validate experimental data. They can provide information regarding interaction details which might not be apparent from the experimental techniques. Several prediction methods have been developed using different approaches. Gene neighbor and gene cluster methods predict protein interactions comparing gene order between different genomes. Other methods are based on the hypothesis that interacting proteins have a similar co-evolution pattern or co-expression of genes. Several approaches have also been developed to predict which protein domains are involved in an experimentally determined interaction. Co-evolution and phylogenetic profiles are strategies that could be also applied in these methods [163]. The enrichment of networks with structural information can be used to validate experimentally determined interactions and to predict new protein interactions [24, 164-167]. Proteins can use different surface regions to interact with various domain types. However, proteins belonging to the same family, if they interact, normally do it using similar positions [168]. Furthemore, structural knowledge of network components has

been used to define the interface between two interacting proteins in order to determine compatible and exclusive interactions [169]. Structural analysis of the network can be performed directly using the **STRING** database. This provides information such as protein domains and, where available, 3D structures.

2.5. Mutation analysis

Protein sequence and structure analysis allows to obtain information that can be useful to derive functional effects of genetic variants. In the near future, thousands of genomes or exomes (protein-coding regions of genomes) will be sequenced by ongoing projects thanks to the rapidly evolving of sequencing technologies. These projects consist of collaborative efforts to generate a catalog of single nucleotide polymorphisms (SNPs) occurring in human, annotating them for their genomic position and their distribution within population from different nationalities [21]. These SNPs frequently vary from one person to another and the study of their location has been used to asses disease risk and to identify disease associated mutations. If a SNP appears to be segregating with a disease, or if it is more prevalent in affected versus unaffected subjects, this may indicate that the SNP is physically close to the disease-causing mutation [170]. The identification of rare variants occurring in a coding gene sequence conferring disease susceptibility requires a strategy to interpret their functional role and calculate the probability of these variants to have a phenotypic effect (Fig. 2.2). In particular, a major effort is to distinguish between functionally significant variations and likely neutral variants. A growing amount of methods have been developed to analyze non-synonymous SNPs using protein function, structure, and evolutionary information [29]. Current prediction methods still show relatively low accuracy, but are successfully employed for various biological and medical applications. The prediction of functional effects can be based on evolutionary information or combining phylogenetic information with structural analysis and sequence properties, while some methods use annotation derived from protein databases.

Databases

Single nucleotide polymorphism are collected in the **dbSNP** [171] database at the NCBI. Here we can find all SNPs identified in genome sequencing projects together with their location in the genome and even the frequency at which they are found in control populations. Some variants are annotated with the minor allele frequency (MAF) value representing the frequency at which a variant has been observed in the **1000 Genome Project** [21]. A MAF higher than 0.5 indicates possible common variants in the population. Although this database contains variants frequently found in various populations, other variants have been associated to some phenotypes.

Mutation databases are used as starting point to verify the novelty of the detected variants. These databases contain information about genes and proteins and their associated phenotype. One of the most widely used databases is the Human Gene Mutation Database (**HGMD**) [13] mainly collecting data from the literature. Online Mendelian Inheritance in Man (**OMIM**) [172] is a catalogue of traits and disorders focusing on genotype-phenotype relationships containing variants reported in the literature. A partial list of known mutations reported in the literature or in dbSNP can also be found also in UniprotKB/Swissprot [173]. The Leiden Open Variation Database (**LOVD**) [174] is a gene-centered database collecting DNA variations. Information about variants are submitted manually and their reliability evaluated by volunteer curators. The complete list of Locus Specific Databases (LSDBs) and central mutation databases can be found at the human genome variation society website.

Predicting functional effects from sequence

Amino acid positions important for protein structure or function usually involve evolutionarily conserved residues and disease-causing mutations frequently occur at these positions. The pathogenicity of a single nucleotide polymorphism can be derived from the multiple sequence alignment (MSA). Conservation of mutated residues can be visualized and mapped on structure with Consurf using a color-coding [45]. MSA analysis can be used to evaluate the type of allowed residues on the mutated position, across different species, in terms of physico-chemical properties (e.g. size, charge, hydrophobicity, polarity). Jalview provides different color schemes for MSA

visualization assigning different colors to residues belonging to different physico-chemical classes.

The conservation of the type and properties of each position is an index of the role of the residue in protein structure or function. Some of these characteristic roles can be investigated directly from previously performed sequence and functional analysis. Amino acid substitutions can alter protein function introducing disorder into structured parts of the protein. The mutated sequence can be analyzed using SPRITZ [111] or CSPRIZ [112] and compared to the wild-type protein prediction. This analysis can also highlight functional alterations due to the introduction of the mutated residue in conserved sequence patterns recognized by ELM [57]. These alterations can involve both post-translational protein modifications and interactions mediated by linear motifs. Mutations can also alter localization signals preventing the correct sub-cellular location of the protein important for its function. Proteins can have specific functions for different sub-cellular compartments based on the different combination of protein partners or substrates.

Deriving mutation impact from the structure

When available, experimentally determined structures can be used as templates to infer information about possible structural alterations induced by mutations. The new side chain can be modeled using homology modeling followed by side chain replacement, rotating each side chain to find the new optimal conformation. When the amino acid substitution occurs at residues mapping on a structured domain, the structure of the mutant protein can be modeled on a 3D predicted model. Even in this case interpretation can be quite accurate [175]. Mapping mutated positions on the structure allows distinguishing two classes, core or surface mutations. Solvent accessibility can also be used to distinguish the two classes. In the two situations the different residue properties may cause different structural impacts, e.g. a large side chain in the core causes large structural re-arrangements. Thus mutations destabilizing the protein core have a higher probability to cause protein unfolding, while alterations at the protein surface can alter crucial ligand or protein binding residues. Mutations may also alter specific structural motifs involved in the molecular mechanisms underlying correct protein function. Such motifs are composed by a conserved arrangement of secondary structural elements and

their correct positioning is dictated by few inter-residue interactions. Residues forming α-helices or β-strands have more constrains that those mapping to loops. Residues have specific propensities for different secondary structural elements, e.g. proline is a known secondary structure breaker [176]. Furthermore, the 2D and 3D protein conformations are the result of chemical bonds and interactions between amino acid side chains in the space.

Pathogenicity prediction

A nsSNP can be classified as pathogenic when the amino acid substitution alters functional protein residues (e.g. active site, protein-protein interaction site) affecting the protein's ability to exert its function and disturbing the molecular pathway in which it works. Several methods have been developed in order to classify mutations on the basis of their pathogenicity [19]. These methods use evolutionary, structural, or biochemical information, or a combination of these features to calculate the functional importance of a specific residue position and can use machine learning approach for the classification process [30]. Prediction methods based on phylogenetic information can incorporate a amino acid substitution matrix, but usually consist of two prediction steps. The first step is to build an accurate multiple sequence alignment and the second is to evaluate how well a genetic variant fits the pattern observed in the phylogeny. The choice of the sequences is crucial, with the best choice being to select only orthologs as inclusion of distant sequences may lead to a uninformative MSA. To calculate the probability of the variant to be damaging, prediction methods can use positional conservation measures or probabilistic scoring functions. **Align-GVGD** [177] and **SNAP** [178] are instead based on amino acid physico-chemical properties. SNAP combines many sequence analysis tools using neural networks. It takes into account information derived from solvent accessibility and secondary structure, flexibility, conservation, and PFAM annotations. It also uses the functional effect assigned by **SIFT** [179], a mutation prediction method based only on evolutionary information. SIFT and also **PMut** [180] (http://mmb2.pcb.ub.es:8080/PMut/) assign a weight to the sequences on the basis of their phylogenetic relationship, assuming that the majority of variations observed in human and their orthologs are functionally neutral [180]. Pmut is based on the use of neural networks to process different kinds of sequence information. It also provides a

pre-calculated PMut database containing the results of a mutagenesis using the PMut method, for all positions in all PDB database proteins.

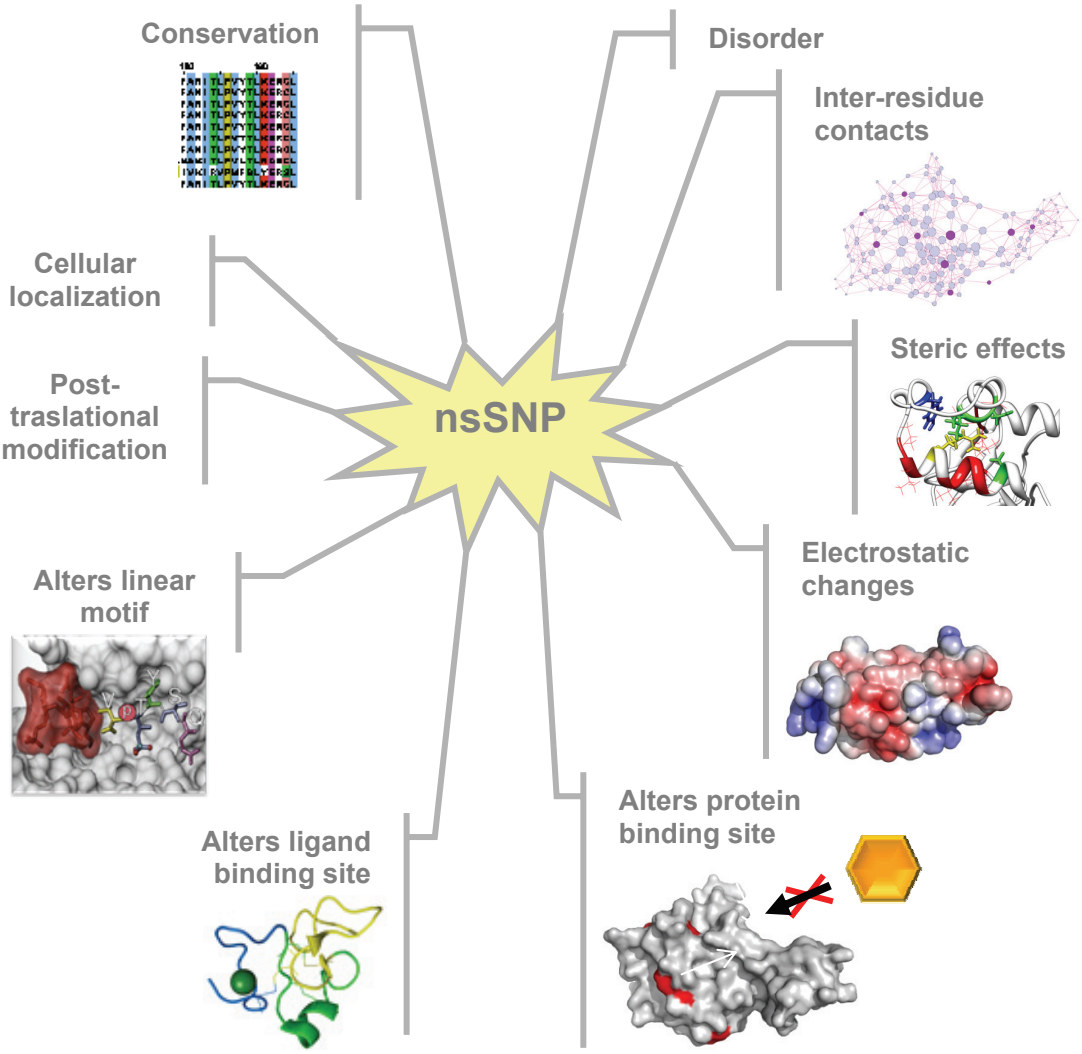


Figure 0.2. **Different aspects to consider in the analysis of non-synonymous SNPs.**

The **PHD-SNP** [181] method (http://gpcr2.biocomp.unibo.it/~emidio/PhD-SNP/) combines conservation with sequence environment using SVM classifiers. Sequence conservation is also used by **Polyphen** [182] (http://genetics.bwh.harvard.edu/pph/) and **SNPs3D** [183] (http://www.snps3d.org/). These methods can use available structural information in addition to reach more accurate results. In Polyphen a profile-matrix is calculated using the PSIC score (position-specific independent counts). Elements of the

matrix are logarithmic ratios of the likelihood of a given amino acid occurring at a particular position to the likelihood of this amino acid occurring at any position (background frequency). It also uses sequence annotation from SWISS-PROT in order to check if the substitution occurs at a specific site (e.g. active or binding site) or in non-globular regions (e.g. transmembrane). Polyphen models the mutant protein to predict the impact of the variant in the hydrophobic core, electrostatic or ligand interactions. SNPs3D, also combining conservation and structure information, uses SVMs to distinguish between disease and non-deleterious SNPs based on features that may affect protein stability. It operates on the hypothesis that disease causing mutations affect protein function, thereby decreasing protein stability.

Stability change prediction

A SNP may affect protein function by altering its stability, either destabilizing or stabilizing it excessively. Protein conformations are in a balance between the folded and unfolded states, and the introduction of an amino acid substitution can shift the balance in either direction. Protein stability differences between wild type and mutant proteins can be calculated using the thermodynamic cycle. The difference in free energy (ΔΔG) between wild type and mutant states can be calculated as:

$$\Delta\Delta G = \Delta G^{U\text{-}F}{}_{wt} - \Delta G^{U\text{-}F}{}_{mut}$$

Where $\Delta G^{U\text{-}F}{}_{wt}$ and $\Delta G^{U\text{-}F}{}_{mut}$ represent the free energy change from the unfolded (U) to the folded (F) state for wild type and mutant proteins respectively. Various methods have been developed to predict stability changes due to protein mutations [30]. We can distinguish two main categories on the basis of the approach used in the calculation: methods calculating stability change by energy functions and machine learning methods. Physical effective energy functions simulate the atomic force field of a structure and can be used to test only small sets of mutants because they are computationally intensive [184-185]. Some methods use empirical potential energy functions combining experimental data, weighted physical and statistical terms, and structural knowledge [186-187]. **FoldX** [186] (http://foldx.crg.es/) is an example of a method using empirical potentials. This program calculates free energy changes based on a 3D protein structure, returning negative ΔΔG values for stabilizing and positive values for destabilizing mutations. Other methods such as **Eris** [188] and **PopMuSic** [189] use 3D structure

information. Eris [188] (http://eris.dokhlab.org/) uses physical force fields in combination with atomic modeling and fast side chain packing. PopMuSic [189] (http://babylone.ulb.ac.be/popmusic/) is based on potential functions derived from statistical analysis on data collected from protein databases such as substitution frequencies, distance potentials and amino acid environmental properties. Stability change prediction methods can also use machine learning approaches trained in protein mutants for which the ΔΔGs have been experimentally determined. **I-Mutant2.0** [190] and **I-Mutant3.0** [191] are based on support vector machines (SVMs) using either protein sequence or structure as input. I-Mutant3.0 classifies mutations in three classes: stabilizing ($\Delta\Delta G>0.5$), destabilizing ($\Delta\Delta G<-0.5$), and neutral ($-0.5<\Delta\Delta G<0.5$). **MUpro** [187] (http://www.igb.uci.edu/server/servers.html) uses the two machine learning approaches SVMs and neural networks and does not require tertiary structure. **Auto-Mute** [192] (http://proteins.gmu.edu/automute/) combines machine learning methods with knowledge-based statistical potentials. Here, the protein residues are represented as points in 3D space and the effect of mutation is estimated as the spatial perturbation to its neighbor residues. Finally, the meta server PON-P [19] (http://bioinf.uta.fi/PON-P/) contains several predictors for disorder, aggregation, stability, and tolerance.

2.6. Residue interaction network analysis

Over the last few decades, network representations has been used to analyze many complex biological problems. Protein interaction networks are the best known example, where nodes represent proteins and connections between nodes their functional or physical interactions. The same approach has been adopted to represent protein structures, where interconnections between nodes (amino acid residues) represent physico-chemical interactions. Thus, the protein structure can be visualized in two dimensions as a residue interaction network (RIN), reducing the visual complexity of a three dimensional model. In addition, the advantage of using a network representation is that there are several efficient and robust algorithms that may be used to manipulate it, since theoretical computer science has studied such structures in detail. The exploration

and analysis of RINs have been applied to study the structural and functional role of residues in protein structures [193]. In particular the RINs have been used to identify key residues determining protein stability, allosteric communication, enzyme catalysis or structural impact of amino acid substitutions [194-199].

Generating residue interaction network

Recently, our laboratory has developed a novel tool named RING [200], to generate RINs for use in Cytoscape [162]. The tool builds a structure network starting from the PDB file and uses different rules to define residue interactions, with closest atom as default. A more realistic connectivity between residues can be obtained by using van-der-Waals surface to sample contacts. The interactions have been defined considering the physico-chemical properties of the interacting residues. For example, RING determines the presence of a salt bridge at physiological pH when a negatively charged residue (Asp or Glu) is in contact with a positively charged residue (Arg, Lys or His). These amino acid types are considered involved in a salt bridge if the distance between the mass centers of the charged groups in their side chains is less that an empirical distance threshold (4 Å default), empirically derived from a large set of protein structures. The RING tool further annotates nodes with structural features, such as secondary structure, solvent accessibility, and data derived from the PDB files (e.g. B factor, occupancy). Evolutionary information is additionally retrieved from the multiple sequence alignment automatically built with PSI-BLAST. The user can alternatively provide his own curated alignment. Conservation scores determined by Consurf is added to each node. Recently, the tool has been improved adding information retrieved from the CSA ligand binding site database in order to annotate known functional relevant residues.

Visual analysis of RINs

Once generated, the RIN can be visualized using Cytoscape [162]. The RING server also produces a VIZ-Mapper property file for Cytoscape including different visual styles (e.g. Consurf, structure conservation, and strong interaction) to color and shape nodes and edges corresponding to different structural features. Nodes or edges can be selected with a target feature to create relevant sub-networks. A useful approach to

analyze the protein is to combine the interactive 2D RIN visualization with the corresponding 3D protein structure using the Rinalyzer [201] plugin for Cytoscape. This software, also providing a database of pre-computed RINs, enables active selection of residues in the RIN while automatically highlighting them on the protein structure. The molecular modeling system UCSF Chimera (http://www.cgl.ucsf.edu/chimera/) is linked to Cytoscape by Rinalyzer, allowing the graphical visualization of the 3D protein structure. Rinalyzer can also be used to compute a set of topological centrality measures characteristic of the networks. In particular, the tool can be used to calculate the simple measure of the node degree, representing the number of connections with neighbor nodes.

Mutation analysis using RINs

The substitution of a residue in a protein structure perturbs the interactions between neighbor residues. Thus, the structural impact of a mutation can be predicted on the basis of the structural role of the substituted residue. A mutant model can be built using an homology modeling approach and the corresponding RINs can be compared to search differences in the presence and type of interactions. The final considerations are only simplistic, since we are analyzing a static representation of the protein in this way. However, knowledge about interactions lost can be used to predict the possible local or global impact caused by the mutation on the protein structure.

3. A novel *WT1* gene mutation in three generations of an Italian family

This chapter has been published in "Benetti E, Caridi G, Malaventura C, Dagnino M, Leonardi E, Artifoni L, Ghiggeri GM, Tosatto SC, Murer L. A novel WT1 gene mutation in a three-generation family with progressive isolated focal segmental glomerulosclerosis. Clin J Am Soc Nephrol. 2010 Apr;5(4):698-702."

3.1. Summary

Wilms tumor-suppressor gene-1 (WT1) plays a key role in kidney development and function. WT1 mutations usually occur in exons 8 and 9 and are associated with Denys-Drash, or in intron 9 and are associated with Frasier syndrome. However, overlapping clinical and molecular features have been reported. Few familial cases have been described, with intrafamilial variability. Sporadic cases of WT1 mutations in isolated diffuse mesangial sclerosis or focal segmental glomerulosclerosis have also been reported.

Molecular analysis of WT1 exons 8 and 9 was carried out in five members on three generations of a family with late-onset isolated proteinuria. The effect of the detected amino acid substitution on WT1 protein's structure was studied by bioinformatics tools.

Three family members reached end-stage renal disease in full adulthood. None had genital abnormalities or Wilms tumor. Histologic analysis in two subjects revealed focal segmental glomerulosclerosis. The novel sequence variant c.1208G>A in WT1 exon 9 was identified in all of the affected members of the family.

The lack of Wilms tumor or other related phenotypes suggests the expansion of WT1 gene analysis in patients with focal segmental glomerulosclerosis, regardless of age or presence of typical Denys-Drash or Frasier syndrome clinical features. Structural analysis of the mutated protein revealed that the mutation hampers zinc finger-DNA

interactions, impairing target gene transcription. This finding opens up new issues about WT1 function in the maintenance of the complex gene network that regulates normal podocyte function.

3.2. Introduction

Wilms tumor-suppressor gene-1 (WT1) encodes a transcription factor that plays a crucial role in kidney and genital tract development. In the developing kidney, *WT1* is predominantly expressed in maturing podocytes, but its expression persists after birth in glomerular visceral epithelial cells, suggesting a role for *WT1* in the function of the differentiated podocyte [202]. *WT1* gene maps on chromosome 11p13, is composed of 10 exons, and encodes a 449-amino acid zinc finger protein. Each zinc finger (Zf) consists of cysteine and histidine residues linked to a zinc atom. A basic amino acid, often an arginine, is located at the top of the finger. Alternative splicing occurs at exon 5 (±17 amino acids) and exon 9 (+3 amino acids; KTS, *i.e.*, Lys-Thr-Ser). The correct ratio of the resulting four isoforms is required for normal gene function during both nephrogenesis and adult life. Depending on splice isoform and the cellular context, *WT1* may indeed act as a transcriptional factor, transcriptional cofactor, or posttranscriptional regulator [203].

Constitutional missense and splice-site mutations of *WT1* gene are the cause of Denys-Drash syndrome (DDS) and Frasier syndrome (FS). DDS (MIM 194080) is characterized by diffuse mesangial sclerosis (DMS) and renal failure with early onset, XY pseudohermaphroditism, and a high risk of developing Wilms tumor [204]. DDS is caused by heterozygous missense mutations in the Zf-encoding exons of the *WT1* gene. These mutations seem to act in a dominant-negative manner, hampering *WT1* activity in cells [205]. FS (MIM 136680) is characterized by focal segmental glomerulosclerosis (FSGS), XY pseudohermaphroditism, and gonadoblastoma. Donor splice site mutations in *WT1* intron 9 have been described as the molecular defect of FS. These mutations result in a deficiency of the usually more abundant KTS-positive isoforms and a reversal of the normal KTS positive-tonegative ratio [206]. Nevertheless, increasing evidence seems to suggest that DDS and FS may represent two facets of the same disease, with overlapping clinical and molecular features [207-211]. In the literature,

sporadic cases of *WT1* mutations in isolated DMS or FSGS have also been reported [207, 212]. We report a novel sequence variant of *WT1* gene, identified in five members on three generations of an Italian family with isolated non-nephrotic proteinuria. The reported clinical and molecular picture raises the hypothesis that *WT1* is associated with a wider spectrum of phenotypes, and WT1 gene may play a more complex role in podocyte function than previously reported.

3.3 Materials and Methods

Patients

Five members of three generations of an Italian family were ascertained. The proband was a 16-year-old boy who underwent clinical assessment and renal biopsy for persistent, isolated non-nephrotic proteinuria, occasionally discovered at the age of 15 years. The other four investigated family members had non-nephrotic proteinuria, with progression to chronic kidney disease in three. None had genital abnormalities or Wilms tumor. All participants provided informed consent to molecular analysis. The study was also approved by our Institutional Review Board.

Molecular Analysis of WT1 Gene

Blood samples from the proband and his relatives were collected. Genomic DNA from fresh whole blood was extracted, and PCR amplification and direct sequencing reaction of coding exons 8 and 9 of the WT1 gene and their intron-exon junctions was carried out. Sequencing data were analyzed using the Sequencher software v.4.9 (Genecodes Corp., Ann Arbor, MI).

Structural Analysis for R403K Mutation

The crystal structure of *WT1* was downloaded from the Protein Data Bank with code 2PRT and was visualized with PyMol.

3.4. Results

Clinical Data

The proband is an Italian 16-year-old boy (III.2 in Fig. 3.1) who was referred to our unit for persistent non-nephrotic proteinuria. His personal and past medical histories were negative: he was born at term after an uncomplicated pregnancy to unrelated parents and had always been healthy. At the age of 15 years, the boy was discovered with isolated proteinuria (75 mg/dl) during regular annual physical examination. Further standard urinalysis confirmed a proteinuria of approximately 50 mg/dl. He was thus referred to our unit for a full nephrologic evaluation. On admission, physical examination was completely normal: weight and stature were at the 50th and 90th percentiles for age, respectively, BP was in the normal range for sex and age (126/69 mmHg), the cardiothoracic and abdominal examination was normal, and there were no abnormalities of genital apparatus. Blood laboratory investigations showed normal hemoglobin (14.9 g/dl), blood urea nitrogen of 45.6 mg/dl, and serum creatinine level of 1.1 mg/dl (clearance according to Schwartz formula 85 ml/min per 1.73 m2). Serum electrolytes were within the normal range, serum albumin was normal (44 g/L), and there were no abnormalities in cholesterol and triglyceride levels (184 mg/dl and 94 mg/dl, respectively). Immunoglobulins and the complement components were normal, and autoantibodies (anti-neutrophil cytoplasmic antibody, antinuclear antibody, anti-dsDNA antibody, anti-myeloperoxidase antibody) were negative. Diuresis was 1800 ml/24 h, with urinary-specific gravity of 1017, urinary pH of 7, and proteinuria of 1.46 g/24 h (corresponding to 33 mg/kg per day). Proteinuria was present and approximately the same in both orthostatic and clinostatic urinary collection, excluding orthostatic proteinuria. Renal ultrasound showed normal-sized kidneys (11.3 cm left and 10.4 cm right), with normal corticomedullary differentiation and no anomalies of the urinary tract.

A renal biopsy was performed (Fig. 3.2). On light microscopy, 30% of sampled glomeruli showed adhesion of glomerular tuft to Bowman's capsule and 10% presented sclerotic lesions, whereas tubuli and interstitium were normal. Immunofluorescence stain testing for IgG, IgA, C3, C4, C1q, fibrinogen, and HBsAg was negative. Electron microscopy showed extensive foot process effacement and mesangial matrix expansion

in the involved glomeruli, consistent with the diagnosis of FSGS. Considering that the markers of autoimmunity were negative, that renal biopsy showed an FSGS with negative immunostaining, and the boy's family history was positive for a still undefined progressive renal disease in several members (see below), we accounted this pattern as more compatible with a genetic form of proteinuria than with an immune-mediated one. Therefore, we found no indications for immunosuppressive therapy in this patient, and angiotensin-converting enzyme inhibitor therapy (Ramipril, 5 mg/d) was undertaken to reduce proteinuria and preserve renal function. At last followup, conducted at the age of 17 years, proteinuria was approximately 1 g/24 h, and renal function was still preserved. The boy's family history was indeed very considerable because his father (II.2), born in 1961, was diagnosed with proteinuria, hypertension, and chronic renal failure at the age of 43 years. Renal ultrasound showed small hyperechoic kidneys, with loss of corticomedullary differentiation compatible with chronic kidney disease stage, but no other peculiar anomalies. Renal biopsy was not performed. After angiotensin-converting enzyme inhibitor therapy, he reached ESRD and underwent hemodialysis at the age of 46 years. The proband's aunt (II.1), born in 1963, developed hypertension at the age of 40 years. Laboratory investigations showed proteinuria and chronic renal failure, but renal biopsy was not performed. By age 44 years, ESRD was reached and emodialysis was undertaken. The proband's grandfather (I.1), born in 1934, developed proteinuria when he was 59 years old. At the age of 64 years, he underwent a renal biopsy, which showed focal glomerular sclerosis, obliteration of capillary lumina with yalinosis, increased matrix, and areas of tubular atrophy. By age 69 years, the developed ESRD, and he started peritoneal dialysis and received a renal transplant 1 year after this. Considering the complex family history, we suggested the uninvestigated family members undergo laboratory tests and ultrasound examination, which revealed isolated non-nephrotic proteinuria in the 18-year-old cousin (III.1) of the index patient.

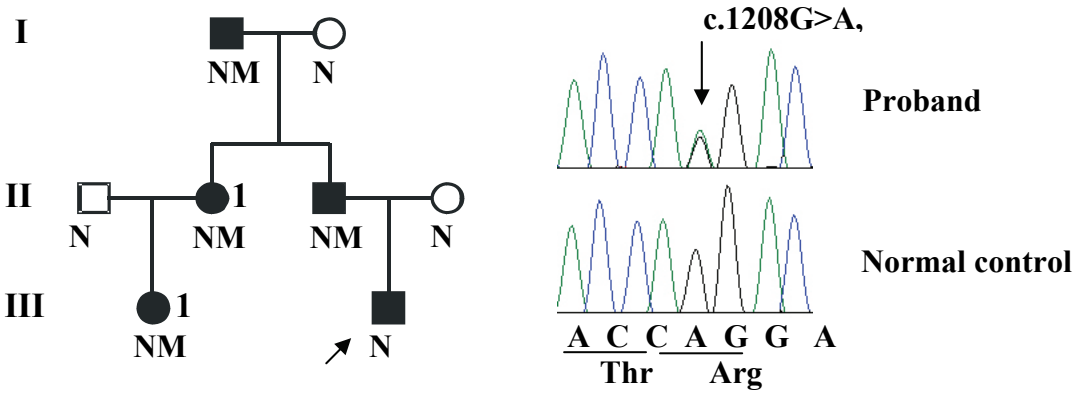


Figure 0.1. Pedigree of the family.

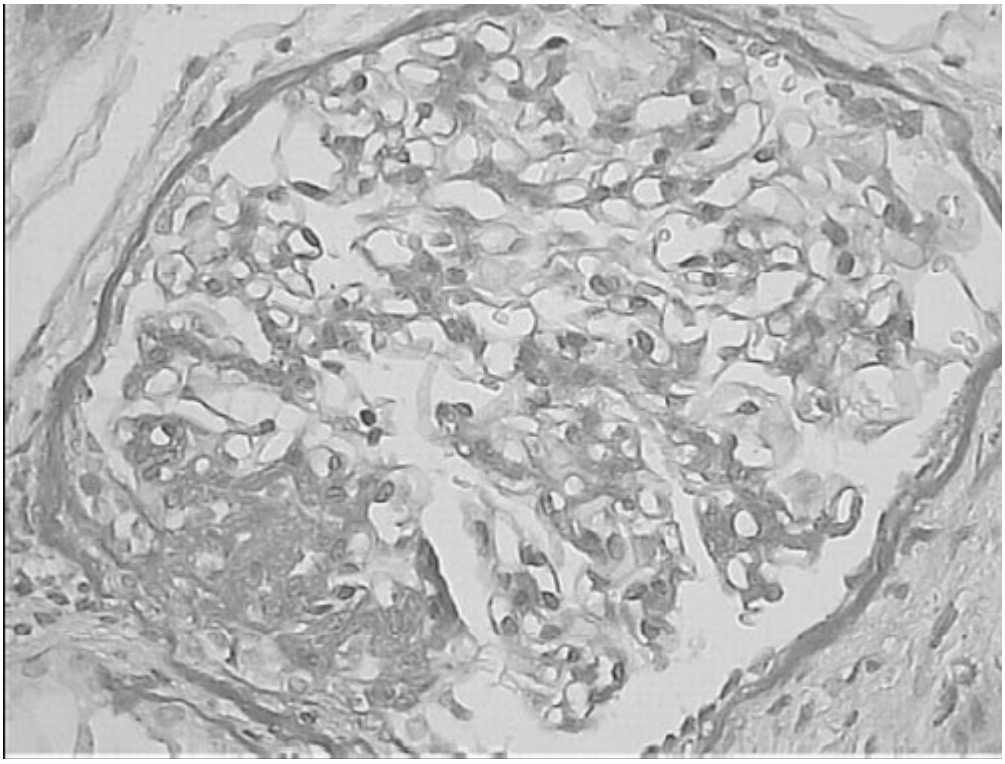


Figure 0.2. Light microscopy of the proband.
The image shows focal segmental glomerulosclerosis (periodic acid-Schiff; magnification, X40).

Molecular Analysis of WT1 Gene

We carried out *WT1* gene exon 8 and 9 analysis by direct sequencing of blood DNA of the proband. *WT1* sequencing revealed nucleotide substitution in position c.1208G>A in exon 9 (GenBank no. M74917), resulting in the substitution of a highly conserved arginine residue with a lysine in the 403 position (p.R403K) of the third Zf domain of the protein. This sequence variant was not observed in 336 control chromosomes.

Molecular analysis was then extended to the other family members, and c.1208G_A variant was detected in the father, aunt, grandfather, and cousin (Fig. 3.1).

Structural Analysis for R403K Mutation

From a structural point of view, replacing Arg with Lys has two effects: the position of the charged residue is shifted, and the charge is somewhat more concentrated (it is spread over three nitrogen atoms in Arg and concentrated at the "tip" on Lys) (Fig. 3.3).

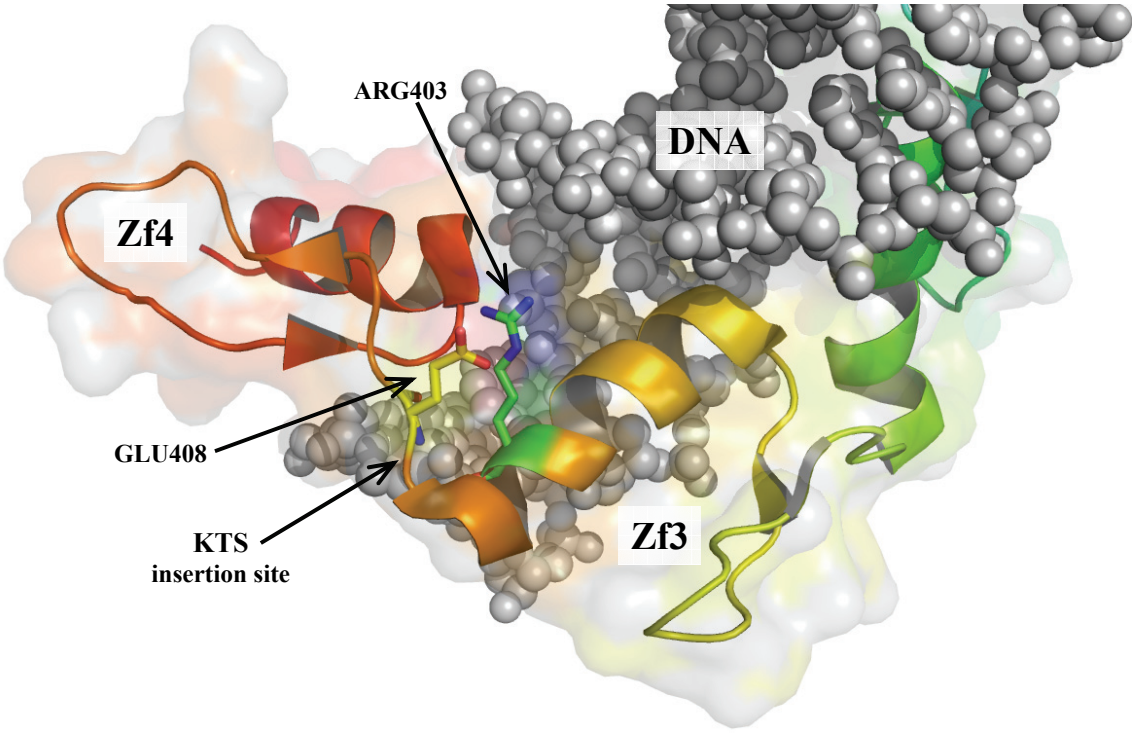


Figure 0.3. Crystal structure of Zf3 of WT1.

3.5. Discussion

We describe an Italian family with isolated FSGS associated with a novel sequence variant in *WT1* gene exon 9. In our study, we tested the detected sequence variant in 336

chromosomes and did not detect it (0 of 168 subjects). According to the literature, a sequence variant is regarded as a polymorphism if minor allele frequency is >1% in normal population. It is universally accepted that a control population of at least 100 subjects is enough to define whether a variant is or is not a polymorphism. In addition, the detected variant has never been observed in the cases of FSGS/DMS or in the somatic mutations associated with Wilms tumor reported in the literature. However, segregation in all the affected members of the family and its absence in a control population suggest that it may be a disease-causing mutation. We applied bioinformatics tools to predict the effect of the amino acid substitution on *WT1* protein structure. c.1208G_A substitution changes an arginine residue located in a strategic position of Zf3 to lysine (p.R403W). Although conservative, this kind of amino acid substitution in critical residues is hypothesized to be of functional significance [212-213]. Zf4 is important for binding, and the presence/absence of the "KTS" insertion seems to switch between DNA binding (-KTS) and RNA binding (+KTS) [214-215]. The residue Arg403 is located in the α-helix of Zf3. It points toward the DNA but is not in direct contact with it (Fig. 3.2). Analysis of the residue's surroundings reveals that it is sterically largely unhindered but forms a salt bridge interaction with Glu408, which is in the linker region between Zf3 and Zf4. Arg403 appears to be anchoring the linker between Zf3 and Zf4 to position Zf4, close to the DNA molecule. Small movements in this conformation would probably make Zf4 swivel out of position. Interestingly, the position of the KTS insertion is exactly between Gly407 and Glu408. Because this insertion is known to affect the position of Zf4 and significantly modify the interactions of WT-1 (DNA *versus* RNA preference), it can be speculated that even small variations of the salt bridge geometry may affect the DNA-binding affinity. Therefore, the sequence alteration would be expected to hamper Zf-DNA interactions, resulting in target gene transcription impairment. Further studies will be requested to confirm the functional effect of the detected variant.

In the literature, *WT1* alterations associated with nephropathy are generally of two types: mutations occurring in exon 8 or 9 (often missense), with patients showing DMS in the context of DDS, and mutations at the intron 9 splice donor site, associated with FSGS in the context of FS phenotype. However, intron 9 splice donor site mutations have also been reported in patients with DMS, pseudohermaphroditism, or gonadic

dysgenesia, with or without gonadoblastoma, and exon 8 or 9 mutations have been described in association with FSGS and gonadal dysgenesis [207, 209-211]. Our patients carried an exon 9 variant and presented FSGS, but they lacked Wilms tumor or genitourinary anomalies. This phenotype is very unusual, especially in male patients who usually show genital abnormalities, but it does actually agree with previously reported cases of isolated FSGS or DMS associated with *WT1* intron 9 mutations, as well as isolated DMS or FSGS associated with exon 8 or 9 mutations [207-208, 215-220]. Although few, taken together these cases highlight that phenotypic variability in *WT1* alterations is probably higher than previously described, suggesting the need for reconsidering and expanding genotype-phenotype correlation in *WT1* alterations. Another peculiar aspect is that the sequence variant was transmitted among three generations of a family in which all members had proteinuria (with eventual progression to chronic kidney disease) and no associated genital anomalies or tumor. In the literature, four cases of familial transmission of *WT1* mutation are reported. Denamur *et al.* [208] described a splice site mutation in *WT1* exon 9 in a 9-year-old girl (karyotype 46, XY) with nephrotic syndrome and DMS, and in her mother, who had proteinuria since the age of 6 years and FSGS. A novel familial read-through mutation in *WT1* exon 10 was detected by Zirn *et al.*[221] in a 22-year-old woman with Wilms tumor and ureter duplex in infancy, as well as slow progressive nephropathy; in her younger brother, who had hypertension but normal renal function; and in their mother, with late-onset nephropathy and ESRD. Regev *et al.* [222] recently reported the transmission of a mutation in exon 1 from a mother with Wilms tumor in infancy to her son with genitourinary anomalies and gonadal dysgenesis with gonadoblastoma foci. Transmission of a substitution in exon 9 from a mother with ESRD to her two daughters (one with nephrotic syndrome and the other healthy) was also reported by Mucha *et al.* [220]. These observations suggested that *WT1* alterations may be associated not only with interindividual but also with intrafamilial variability. Differently from these reports, all members of our family displayed the same phenotype of isolated proteinuria due to FSGS. Furthermore, in our patients the onset of proteinuria was not in early life, and ESRD developed in full adulthood, differently from most cases of the literature, in which clinical manifestations commonly occur in infancy. These peculiarities suggest that *WT1* gene analysis is to be taken into consideration in the assessment of patients

with FSGS-associated proteinuria, regardless of age or presence of typical DDS or FS clinical features. Several studies have shown that the target genes potentially regulated by *WT1* include genes that code for transcription factors (such as PAX2), growth factors or their receptors (EGR1, EGFR, IGFR1R, TGF-_1, IGF2, IGFR, PDGF-A, VEGF), as well as podocyte proteins, such as nephrin and podocalyxin [203]. Because the filtration barrier's function requires the integration of multiple signaling pathways between endothelial, mesangial, and podocyte cells, correct *WT1* interaction with target genes seems to be crucial to the maintenance of such a complex and dynamic structure. Furthermore, a proteomic study of DDS podocytes showed that they misexpress proteins associated with cytoskeletal architecture (including cofilin, calponin, elfin, hsp27, and vinculin), and total levels of filamentous actin were also reduced [223]. *WT1* has also been demonstrated to regulate the intermediate filament protein nestin, whose reduced expression was associated with podocyte dysfunction [224-225]. These findings suggested that in addition to its traditional role in regulation of proliferation, *WT1* can also influence cytoskeletal architecture, accounting for the development of proteinuria and the lack of genitourinary abnormalities or Wilms tumor in some patients. The maintenance of regularly spaced and interdigitated podocyte foot processes with their associated slit diaphragms is indeed essential to filtration barrier integrity, and the loss of podocyte cytoskeletal architecture and slit diaphragms results in its dysfunction. In summary, normal podocyte function is maintained by a complex and dynamic gene network in which *WT1* seems to exert a crucial role, so that its mutations may result in a broad range of phenotypic alterations. Furthermore, our finding of a novel *WT1* mutation in a family with isolated proteinuria suggests extending *WT1* gene mutational screening to patients with FSGS, which will contribute to a better understanding of *WT1* functions in podocytes.

4. Adding structural information to the von Hippel-Lindau (VHL) tumor suppressor interaction network

This chapter has been published in "Leonardi E, Murgia A, Tosatto SC. Adding structural information to the von Hippel-Lindau (VHL) tumor suppressor interaction network. FEBS Lett. 2009 Nov 19;583(22):3704-10".

4.1. Abstract

In this chapter I present a work explaining the function of the crucial tumour suppressor gene von Hippel-Lindau (VHL) on a large scale using advanced bioinformatics methods. The von Hippel-Lindau (VHL) tumor suppressor gene is a protein interaction hub, controlling numerous genes implicated in tumor progression. Here, I show how to systematically apply structural information to enhance our understanding of complex proteins with many interactions.

This work focus on structural aspects of protein interactions for a list of 35 experimentally verified protein VHL (pVHL) interactors. Using structural information and computational analysis I have located three distinct interaction interfaces. lnterface B is the most versatile, recognizing a refined linear motif presents in a number of otherwise non-related proteins. It has been possible to distinguish compatible and exclusive interactions by relating pVHL function to interaction interfaces and subcellular localization. A novel hypothesis is presented regarding the possible function of the N-terminus as an inhibitor of pVHL function.

4.2. Introduction

Von Hippel-Lindau (VHL) syndrome is a dominantly inherited familial cancer syndrome with variable expression and age-dependent penetrance, characterized by a predisposition to develop various tumors, including among others hemangioblastomas, clear cell renal carcinomas and pheochromocytomas. Predisposition to develop this variety of tumors is linked to germ line inactivation of the VHL tumor suppressor gene, coding for VHL protein (pVHL). Pathology development in VHL disease occurs after somatic inactivation of the remaining wild-type allele in a susceptible cell [226]. Certain classes of pVHL mutations confer different site-specific risks of cancer, suggesting pVHL to have multiple tissue-specific tumor suppressor functions [227]. How pVHL mutations cause different disease phenotypes remains incompletely understood.

It is widely accepted that pVHL functions as target recognition component of the E3 ubiquitin ligase complex targeting the α-subunits of hypoxia-inducible factors (HIF) for oxygen-dependent proteolytic degradation [228-230]. pVHL inactivation leads to stabilization of HIF1α and HIF2α and activation of their downstream target genes implicated in angiogenesis, cell growth, and metabolism, e.g. vascular endothelial growth factor (VEGF), platelet derived growth factor (PDGFB), transforming growth factor (TGF) and erythropoietin [231]. However, distinct pVHL-containing complexes identified indicate pVHL involvement in different signaling pathways, including microtubule dynamics, primary cilium maintenance, cell proliferation, neuronal apoptosis and extracellular matrix deposition. Several studies have recently investigated the numerous pVHL functions, providing insight into pVHL mediated signaling networks involved in tumor formation (for recent reviews see [226, 232-233]). Experimental determination of protein interactions provides growing data about new pVHL partners and, partially, the protein regions involved. However, a large amount of data remains to be confirmed.

I address the problem by combining information from experimental data with structural analysis [24, 167-168]. Defining the pVHL interaction interfaces was the primary. As pVHL has more interaction partners than available surface, some interactions must be mutually exclusive, while others interact simultaneously [169, 234]. Structural

information also offers the possibility to distinguish domain-domain from peptide-domain interactions [164]. Hypotheses are drawn regarding the expected interaction type in each interface and interaction partners are classified. Augmented with functional information I contribute to define functional sub-networks existing in different time and space conditions, adding a dynamic component to pVHL function.

4.3. Materials and Methods

Protein interaction data were extracted from the literature, sequences from UniProt [235], protein domain architectures from Pfam [236] and structures from PDB [237]. SPRITZ [111] was used to predict disorder, the consensus method for secondary structure prediction [99], REPETITA [97] to validate the N-terminal repeat and FlexServ [238] to estimate flexibility. Fold recognition was performed using MANIFOLD [239] to identify a structural template with sufficient sequence identity in cases with no experimental structure. Structural and functional classification is based on CATH [240] and GO [241] respectively.

To map pVHL evolutionary sequence conservation, 35 closely related sequences identified with PSI-BLAST [37] (default parameters) were realigned with ClustalW [41]. The multiple sequence alignment was drawn using ESPript [84]. Consurf [242] was used for the analysis of conserved residues and Crescendo [153] to identify functionally conserved residues. Structures were visualized in PyMol (URL: http://www.pymol.org/). Protein sequences were analyzed for linear motifs with Jalview [243] and PepSite [129] was used to score the pVHL surface for candidate interaction motifs.

4.4. Results and discussion

Data retrieval

Figure 4.1 shows an overview of the sequence features of pVHL. Information about sequence, domain architecture, structure and function of interactors was collected in a list containing 35 experimentally determined proteins (Supplementary Material S1).

Structural details are known for Elongin C (EloC), Elongin B (EloB) and HIF1α which are co-crystallized with pVHL [228-230]. For other interacting proteins the pVHL binding regions is a particular domain, e.g. the p53 DNA binding domain [244]. Some pVHL interacting regions are located on linkers between two domains or in regions of unknown structure. The full list is reported in Supplementary Table S.4.1 (see Appendix).

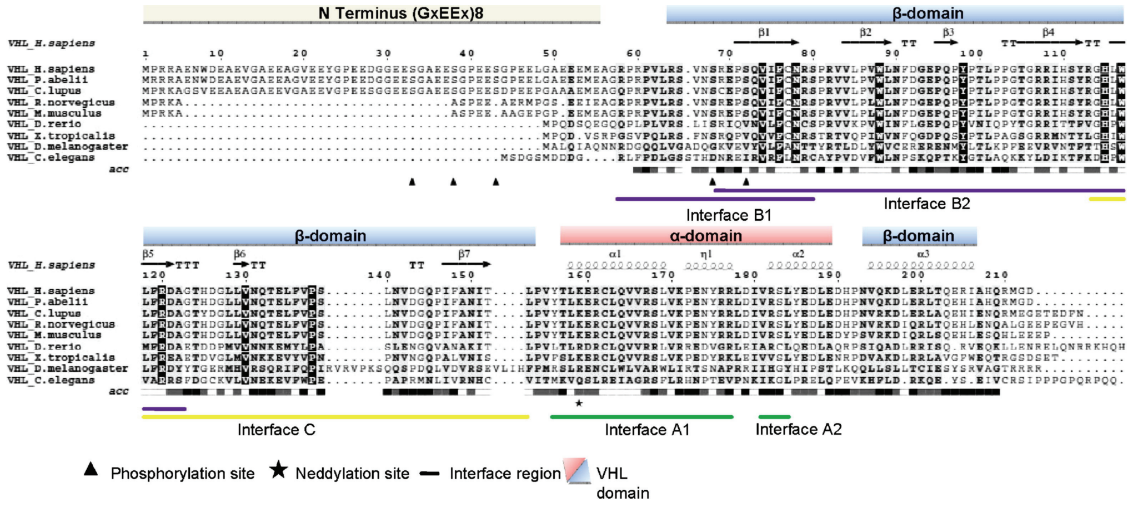


Figure 0.1. pVHL Sequence Features.

Domain architecture and protein sequence alignment of pVHL. The pVHL domains are shown on the top part. Secondary structure and homologous sequences of pVHL are shown in the center. *acc*: accessibility level from DSSP (black=high and white=low). The protein-interaction regions are represented as lines at the bottom of the sequence alignment.

Interface definition

Figure 4.2 summarizes the pVHL structural features with its α- and β-domains. Comparing the results obtained by Consurf (Fig. 4.2A) and Crescendo (Fig. 4.2B) it is noted that surface regions containing relevant functional residues are located in two interfaces known to interact with HIF1α, EloC and EloB. In addition a third interface appears on the back of pVHL. Protein structures frequently interact with various domains using different surface areas [169]. Mapping the putative protein interacting regions on the pVHL structure for interactors listed in Supplementary Table S4.1, the majority of interactions occur in three specific areas of pVHL that we define interfaces A, B and C (see Fig. 4.2C). No structure is available for the pVHL N-terminus

containing putative phosphorylation sites. In the following, we describe the characteristics of each interface separately, which are also summarized in Figure 4.3.

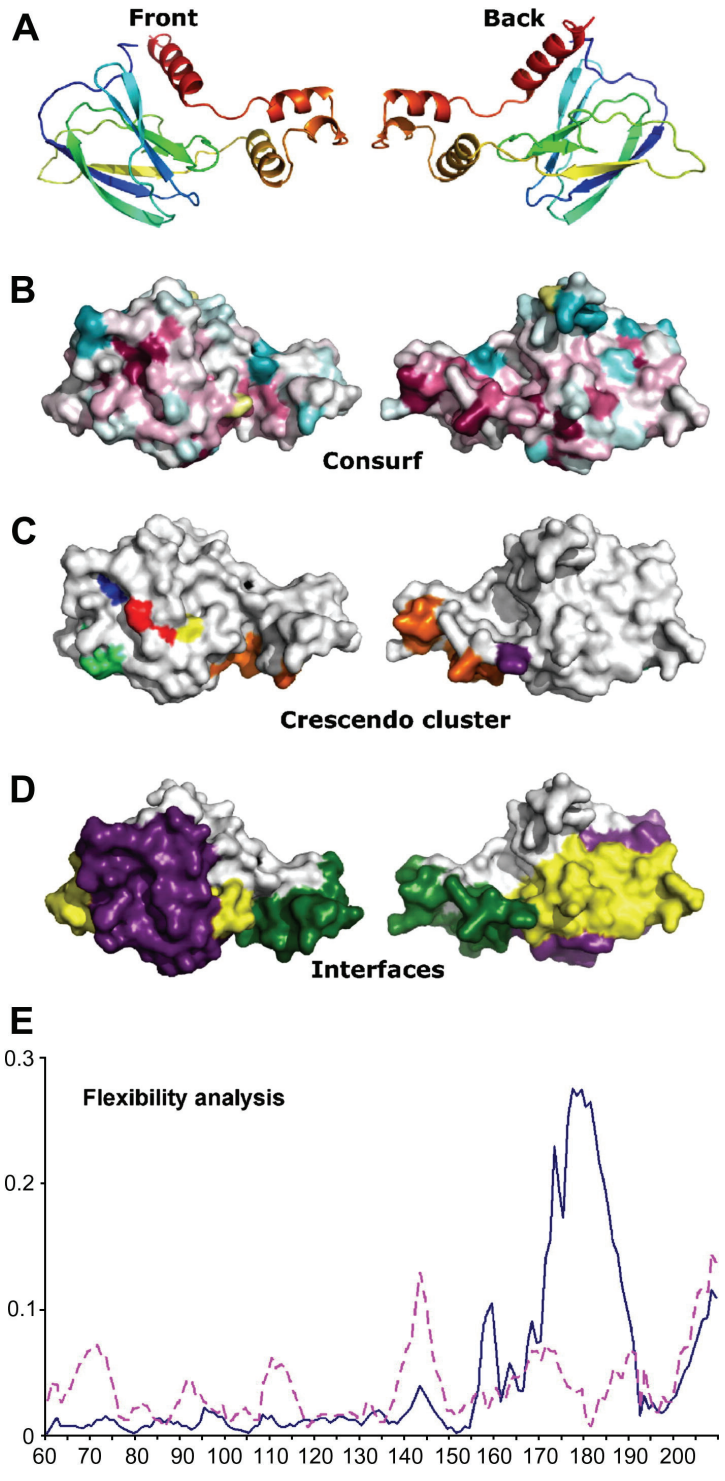


Figure 0.2. pVHL Structure Overview.
A. Cartoon representations of the pVHL structure (pdb code 1LM8 chain V) colored blue (N-terminal) to red (C-terminal). The structure, shown in standard view (front) and after a 180° rotation around the z-axis (back), is maintained throughout the following panels. B. Surface of pVHL projecting phylogenetic conservation by Consurf. The Consurf prediction is presented with magenta shading for highly conserved residues and cyan shading showing for variable residues. C. Surface of pVHL where Crescendo residues clusters are coloured: lime green, cluster A; orange, cluster B; yellow, cluster C; red, cluster D; blue, cluster E; violet, cluster F. D. Three pVHL interfaces mapped on structure surface of pVHL: green, interface A; violet, interface B; yellow, interface C. E. Analysis of pVHL flexibility of each residue (x axis) in terms of relative displacement (y axis, in Angstroms).

Interface A – Processing

It has been demonstrated that interaction of EloB and EloC with pVHL depends on the binding of EloC to a 10-amino acid α-helical sequence motif xLxxxCxxx[AILV], referred to as the BC box [245]. pVHL also interacts with Cullin-2 at the specific Cul2 box, located C-terminal to the BC box, forming the E3 ubiquitin ligase complex responsible for recognition and recruitment of target proteins to be degraded by the 26S proteasome [246]. Four proteins, other than EloB/C and Cul2, have been experimentally determined to interact with the pVHL α-domain through a specific domain. The p53, Nur77, HuR and VBP1 interacting regions overlap with the EloC interaction. This supports the idea that the pVHL α-domain mediates domain-domain interactions. The EloC, p53, HuR and Nur77 domains interacting with VHL have a similar 2-layer sandwich architecture, with different CATH classifications. Their structures are mainly composed of β-sheets with an α-helix that might mediate the interaction in analogy to EloC. The interaction interface between pVHL and EloC is composed of three pVHL α-helices (H1, H2 and H3) and the H4 helix of EloC, forming hydrophobic contacts. Flexibility analysis of the VHL structure indicates the α-domain is flexible (Fig. 4.2C) and suggests a conformational change upon interactor binding. Comparing helices found in other pVHL interactors leads to the hypothesis that interactions with different domains occur in different ways.

EloC, p53, HuR, Nur77 and VBP1 are mutually exclusive and compete for binding to the α-domain (Fig. 4.3). p53, HuR and Nur77 are prevalently expressed in the nucleus, and EloC in the cytoplasm. The α-domain also contains the neddylation site, K159, which appears to be necessary for fibronectin binding. When neddylated, pVHL cannot interact with Cul2 and stabilizes fibronectin, showing processing functions depending on localization and physiological cell status [247].

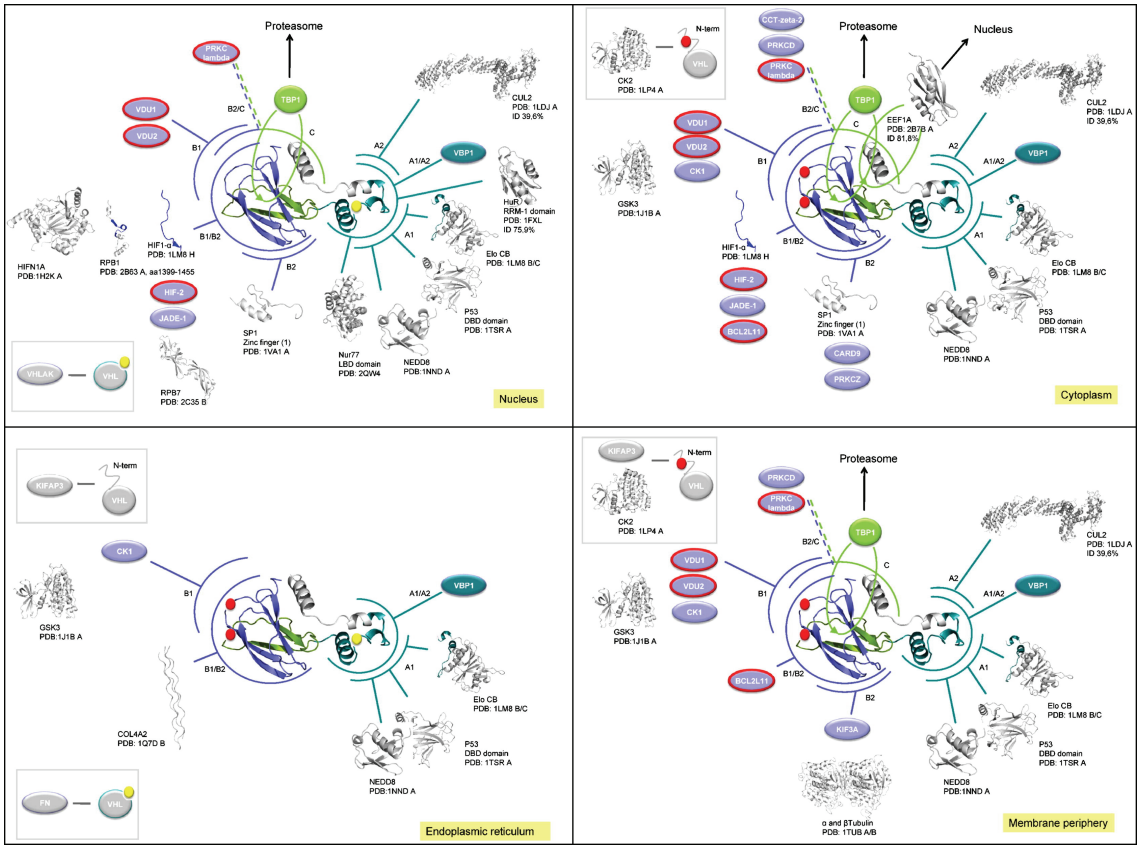


Figure 0.3. Overview of pVHL Interactors.
Pictures represent pVHL interacting partner in four different cellular compartments: Nucleus, cytoplasm, endoplasmic reticulum and membrane periphery. At the center of each picture there is a cartoon representation of the crystal structure of pVHL (PDB code 1LM8, chain V) highlighting the protein interface as circles: interface A in cyan, interface B in blue, interface C in green. When the structure of a pVHL interaction partner is known, it is represented as cartoon, with their acronym and PDB code. A percentage sequence identity is shown if the structure refers to a homologous protein. Light blue circles are used if the interaction region is unknown. A red border indicates proteins for which information about pVHL interaction regions exists but it is located in an unstructured region or in domains for which no structure is available. A grey circle indicates proteins interacting with an unknown pVHL region. KIFAP3 and CK2 interact with the VHL N-terminal sequence which is not included in the crystal structure. Phosphorylation (red balls) and neddylation sites (yellow ball) are also shown.

Interface B – Substrate recognition

The pVHL β-domain, interface B, containing the well studied HIF1α binding site, was found to be essential for binding many different proteins, with variations termed B1 and B2. Some interaction sites were shown to correspond to the HIF1α binding site. The region essential for VDU1 binding overlaps only partially and e.g. for Jade-1 and HIF1AN, the specific pVHL β-domain interaction site is unknown. Only five proteins have experimentally determined VHL interacting regions, with a prevalence of

unstructured linkers. For Sp1 the interaction is known to occur through one of three Zn Finger domains and pVHL interactions with other metalloproteins have been hypothesized [248]. Other Zn finger domain containing proteins are in the list, although in VDU1 and RPB1 the experimentally determined interacting region does not contain the Zn finger domain. In protein Kinase C zeta (PRKCZ) and Jade-1 the binding region is unknown.

Proteins of unknown pVHL interacting region are likely to use disordered sequence stretches. The HIF1α binding sites, named ODD domains, are disordered and contain a hydroxylated proline. It has been demonstrated that RPB1 interacts with pVHL through a proline. Sharing a conserved motif containing this proline, it seems interface B mediates interactions occurring between the pVHL β-domain and different peptides.

The known interface B interactors were systematically searched for linear motifs resembling the HIF1α ODD sequence and structurally validated with PepSite. The results (shown in Figure 4.4) point to the presence of a proline box followed by a hydrophobic box with the consensus pattern [LIV]xPx(6,9)δxδ, where δ is prevalently a hydrophobic residue, in a likely disordered region. This pattern agrees with the location of known pVHL interactions for HIF1α[229] and RPB1 [249]. With few exceptions (PRKCZ and perhaps VDU1, VDU2) the predicted motifs fit well into the respective structures. Although experimental binding assays will be necessary to verify the prediction, these results summarize well the plasticity of the pVHL B interface.

As many interface B interactors result to be mutually exclusive, the selection of interaction occurs in different ways. Domain-domain interactions are more specific or have higher binding affinities than domain-peptide interactions. The different interactors have specific localization, e.g. RPB1 is nuclear while HIF1α is nuclear and cytoplasmatic. Another selection criterion may be the time at which some proteins are expressed with higher concentration. It appears that VHL through interface B can bind proteins which will be ubiquitinylated and then degraded, yet it can also bind other proteins which will be stabilized, e.g. BCL2L11 and microtubules (MT). Interactions with other proteins play a role in transcriptional regulation, e.g. Sp1, HIF1AN and perhaps VHLaK for which an interaction pVHL interface remains unassigned.

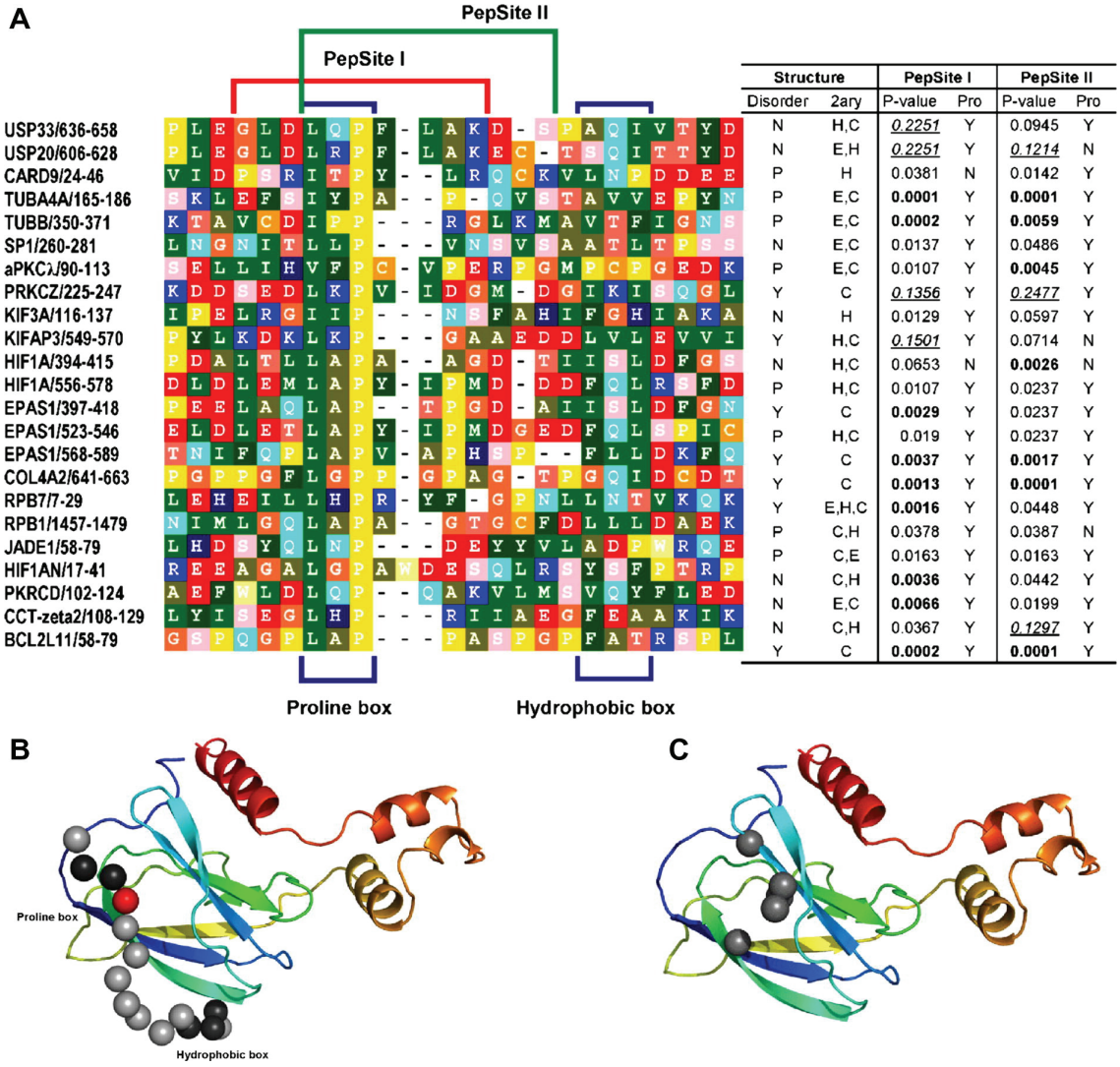


Figure 0.4. pVHL Interface B Linear Motifs.
The putative linear motifs defining pVHL interactions at interface B are shown. (A) Multiple sequence alignment of identified linear motifs with the sequence identifier and positions on the left side. The right side contains information on the disorder (Y = yes, N = no, P = partial) and secondary structure (H = helix, E = extended/beta, C = coil/loop) and two Pepsite predictions. Pepsite P-values, estimating the probability of achieving a similar result by chance, are complemented with the presence of the central proline in the predicted motif. Extremely low (bold) and high (italics, underlined) P-values are highlighted. The two regions corresponding to the Pepsite predictions are drawn above and two motif-determinant boxes are drawn below the sequence alignment. In (B) the pVHL structure (PDB code 1LM8, chain V) is shown oriented as in Figure 4.1 with the bound HIF1α peptide in grey spheres. The hydroxylated Proline residue is shown in red and the two motif-determinant boxes are shown in darker grey. A typical Pepsite prediction (BCL2L11, Pepsite P-value 0.0001) is shown in (C). Note the similar binding site with proximity to the hydroxylated Proline location.

Interface C – Localization

The back view of pVHL β-domain (Fig. 4.1) presents two regions for known interactors. TBP1 interacts with pVHL residues 136-154 containing a β-sheet and a linker between

the VHL α and β domains [250]. It is difficult to predict how TBP1 can interact with this extended region. Residues 114-138 of pVHL interact with eEF1A [251] and contain the nuclear export signal (NES), DxGxxDxxL [252], in a loop between two β-strands forming a polar interface between pVHL α and β domains. Surprisingly, the NES motif is part of the polar interface partially located in front of the α-domain. This may be explained by the flexibility of α-domain which likely changes conformation and increases accessibility for other proteins. During transport from nucleus to cytoplasm, pVHL is captured by a nuclear export receptor component. In the cytoplasm, TBP1 (component of the 26S proteasome) interacts with pVHL in complex with EloC, anchoring it to the proteasome where the ubiquitinylated substrates are degraded. This interface appears to determine pVHL localization. Interestingly, the areas interacting with TBP1 and eEF1A forming interface C are coded by exon 2 of the VHL gene. However, amino acids 114-177 appear to be required for perinuclear (ER) pVHL localization [253].

pVHL dynamics

pVHL contains several posttranslational modification sites, determining its dynamic properties. Reversible pVHL neddylation distinguishes between HIF-related functions and stabilization of the extracellular matrix [247]. Together with the localization signals in interface C this determines to a great extent the precise function pVHL may be carrying out at any given point in time.

A more elusive pVHL regulatory mechanism is phosphorylation. Several phosphorylation sites present in or near the pVHL N-terminus are responsible for proper fibronectin deposition [254] and microtubule dynamics [255]. In addition, the N-terminus is entirely missing in other mammalian orthologs, suggesting a human specific mechanism. Eight tandem repeats with a GxEEx pattern are contained in the N-terminus (see Fig. 4.1). REPETITA periodicity analysis strongly suggests that the repeat is a solenoid and covers the entire pVHL N-terminus, with two degenerate units at the N- and one at the C-terminal flanks of the GxEEx pattern. All known six-residue solenoid repeats form α-helix structures [256], where each repeat corresponds to a short β-strand and a turn typically centered on a glycine. Whether the pVHL N-terminus can form the

same β-solenoid remains however unclear due to the occurrence of adjacent charged residues which may cause electrostatic repulsion.

In any case, the putative phosphorylation sites located in the repeat region are likely to cause a conformational switch, creating a pattern of three charged residues known to be highly correlated with intrinsic disorder [257]. The serines located close to two prolines (see Fig. 4.2) have prompted us to use Pepsite to determine hypothetical pVHL-peptide interactions (Fig. 4.4). Intriguingly, the results place the interaction site on interface B in correspondence with the HIF-1α peptide (P-value = 0.024). This opens the possibility that human pVHL may have evolved a signaling mechanism to deactivate binding of specific interactors through phosphorylation and conformational rearrangement. Experiments will be necessary to verify this hypothesis.

4.5. Conclusions

I have presented a characterization of pVHL which attempts to summarize the known interactions and address them from a structural point of view (see Figure 4.3). The modular nature of pVHL becomes apparent from the subdivision in three interaction interfaces corresponding to processing, substrate recognition and localization. These highlight various protein interaction types, namely domain-domain (interface A) and domain-peptide (interface B), with interface C being less clear. Structural characterization of the putative interaction peptides yielded both a complete list of hypothetical interaction motifs and the intriguing possibility for the pVHL N-terminus to auto-inhibit substrate recognition after phosphorylation.

5. Identification and in silico analysis of novel von Hippel-Lindau (VHL) gene variants from a large population

This chapter has been published in Leonardi E, Martella M, Tosatto SC, Murgia A. Identification and in silico analysis of novel von Hippel-Lindau (VHL) gene variants from a large population. Ann Hum Genet. 2011 Jul;75(4):483-96.

5.1. Summary

In this chapter I present the computational approach adopted to analyse novel variants identified in the VHL gene. Mutational inactivation of the *VHL* gene is the cause of von Hippel-Lindau disease, autosomal dominant hereditary cancer syndrome predisposing to hemangioblastomas, pheochromocytomas and clear-cell renal carcinomas. The gene product (pVHL) functions as an adapter in cellular processes including cell growth and apoptosis.

The mutation data presented in this work was obtained by the Rare Disease Laboratory at the Department of Pediatric, University of Padova, which is the national reference group for the analysis of VHL gene. VHL mutation analysis was carried in 426 unrelated subjects with phenotypes ranging from VHL syndrome, to isolated *VHL-related* tumors that could represent the first manifestation of the disease. 111 individuals were found to carry alterations, with large deletions representing 40% of the variants. 18 of the 95 detected variants were novel, seemingly disease-causing mutations; their pathogenic role has been evaluated *in silico* for effects on protein folding and interactions. Putative regions of interaction between pVHL and proteins involved in common pathways have been identified previously and described in the chapter 2, assessing possible implications for the presence of mutations in these regions. All new variants predicted to truncate or cause complete pVHL loss of structure were associated with phenotypes consistent with VHL type 1. Seven of the new amino acid

substitutions are disease-causing mutations, one is a neutral variant, whereas the results for two remain ambiguous. The computational approach I adopted contributes to the interpretation of the potential pathogenicity of these novel variants.

5.2. Introduction

von Hippel-Lindau disease (VHL; MIM #193300) is a familial cancer syndrome due to mutations of the *VHL* gene [258]. It is characterized by predisposition to the development of highly vascularized tumors such as retinal and central nervous system, hemangiobalstomas (RHB, CHB), pheochromocytomas (PH), and clear-cell renal carcinoma (RCC) [259]. VHL is clinically classified in type 1 and type 2 based on the absence or presence of PH, one of the early onset features of the disease. Type 2 VHL is sub-classified based on lower (type 2A) or higher (type 2B) susceptibility to RCC. Despite extreme phenotypic variability between and within families, important genotype-phenotype correlations have emerged for different classes of pathogenic mutations [227, 260-263]. Mutant copies of the *VHL* gene that completely abolish its normal function are found virtually only in VHL type 1 disease, with low risk of PH. VHL type 2 is almost invariably associated with missense mutations, and a limited number of these mutations have been specifically associated with a PH-only subtype, VHL type 2C [264-265].

pVHL is a substrate recognition component of an E3 ubiquitin ligase complex (VCB), including Elongin C, Elongin B, Cullin2 and Rbx1/Roc1, targeting proteins for ubiquitin-mediated degradation [266]. It contains two functional domains: the prevalently C-terminal α-domain allows the protein to adopt its native 3D conformation after binding to Elongin C. The β-domain forms a substrate docking interface for target proteins. The best known target of this complex is HIF1α (hypoxia-inducible factor-1α) [230], but several other substrates have also been identified[226]. Multiple HIF-dependent and HIF-independent functions are known, all contributing to the VHL-defective oxygen sensing response and tumorigenesis [231]. The growing body of data about pVHL interactions attributes different functions to specific discrete regions of the molecule [267]. Furthermore, in some cases, specific functions have been directly

related to a particular clinical manifestation of VHL syndrome. E.g. HIF deregulation plays an important role in the HB development and cytoskeletal architecture is defective in RCC, while regulation of apoptosis seems to be crucial in prevention of PH during embryological development [226].

Distinguishing between pathogenic and non-pathogenic mutations in carriers of VHL variants is crucial for early diagnosis of a disease with age-related and variable clinical profile. *In vitro* characterization of the pathogenicity of sequence variants can be difficult, especially when a large number of different and often private mutations are detected. In light of this, structural data may be very important. The disease phenotype may in fact be caused by amino acid substitutions affecting residues involved in crucial interactions, or crucial for maintaining protein folding and structural stability. Several computational methods that predict potentially deleterious effects of missense mutations can be used to prioritize the most likely disease causing variants and gain insight into molecular disease mechanisms [19].

In this work I presented the analysis of potential structural and functional effects of mutations found in individuals with different VHL-related phenotypes, with phenotypes ranging from full clinical von Hippel-Lindau disease to isolated VHL-related tumors that could represent the first manifestation of the disease. Indeed, VHL mutation analysis is recommended for cases initially presenting with isolated retinal HB, sporadic CNS hemangioblastomas or seemingly sporadic pheochromocytomas [268]. Novel variants are established and distinguished by type. The role of novel variants has been predicted *in silico* for effects on protein folding and interactions. Known regions of interaction between pVHL and proteins involved in common regulatory pathways have been assessed for possible changes due to the presence of mutations.

5.3. Materials and Methods

Study population

The studied population comprises 426 unrelated individuals presenting phenotypes ranging from von Hippel-Lindau disease to single, apparently sporadic VHL-related tumors, sent in the last 14 years for VHL genetic testing to the referral laboratory of the

Department of Pediatrics, University of Padua. The age of tested subjects ranged from 8 to 62 years (mean age 32). Regular signed informed consent for molecular analysis was obtained for each tested individual. The phenotypes of patients described in the text refer to the clinical conditions ascertained at the time of molecular diagnosis. While the presence of PH allowed to define VHL type 2 phenotypes, the likely attribution to the VHL type 1 category in its absence has to be taken with caution given the possibility of a later occurrence of the tumor.

Molecular analysis

High molecular weight genomic DNA was extracted from peripheral blood leukocytes by standard protocols. Mutation scanning of the *VHL* gene for identification of point or small size mutations was conducted on the entire coding sequence and intron-exon boundaries by PCR amplification, DHPLC and direct sequence analysis, as follows.

PCR amplification was performed with the use of previously reported primers and optimized reaction conditions [269]. All the amplicons were subjected to DHPLC analysis. The temperature for heteroduplex detection was determined using the Navigator™ Software v.1.7.0 (Transgenomic™), and at least 2-3 different temperatures were chosen for distinct melting domains in each fragment to be analyzed (optimized elution profiles and melting temperatures of the entire coding sequence of *VHL* gene are available upon request). All fragments showing altered melting curves were sequenced after purification (Microcon Y100), with the use of the same primers and fluorescently labeled dideoxy chain terminators from ABI Prism kit (Big Dye Terminators 3.1), on an ABI 3100 automated sequencer. Quantitative Real Time PCR for the identification of deletions of part or the entire gene, was performed on genomic DNA fragments representing each VHL exon. Primer pairs and reaction conditions as in [270]. Segregation analysis in families of individuals carrying VHL variants was performed whenever possible and, after extensive information and proper counseling, by targeted sequencing in parents and/or siblings of the probands and in other relatives at risk. Mutation nomenclature follows codon numbering as by [271]. RefSeq: NM_000551, protein ID: NP_000542.

In silico analysis

Protein sequences were retrieved from UniProt [235] and the VHL protein structure was obtained from the PDB database [237] (PDB code: 1LM8). A PSI-BLAST [37] search with the pVHL1-213 sequence as query on a non redundant protein database was performed to collect homologous sequences. The multiple sequence alignment (MSA) was built with CLUSTALW [41], annotated with secondary structure and accessibility values assigned with DSSP [75], and drawn using ESPript [84]. The MSA was used as input for Consurf 3.0 [242], which calculates the conservation score and visualized in PyMol (De Lano Scientific; URL: http://www.pymol.org/). The Universal Mutation Database [272] (UMD; http:/www.umd.be/VHL/) and Human Gene Mutation Database [14] (HGMD; http://www.hgmd.cf.ac.uk/ac/search.html) were used to obtain information about identified variants. Two different splice-site algorithms were used to predict a potential splicing effect: NNSplice [273] and NetGene2 [274].

Amino acid substitutions were mapped onto the pVHL structure and visually evaluated for their structural effects. Stability changes upon single site variants were estimated using I-Mutant 3.0 [191], Eris [188] and Auto-Mute[275]. Polyphen [182], SNPs3D[183], PMut [180] and SNAP [276] were applied to predict potentially deleterious effects of the new variants. The mutant models for 8 new variants were built with ClustAlign [277] and HOMER (http://protein.bio.unipd.it/homer/). GROMACS [90] was used for 1000 steps of steepest descent minimization to relax the mutant structures. The RING server (Martin A.J.M. et al., submitted; http://protein.bio.unipd.it/ring/) was used to generate the residue interaction network useful for evaluation of structural changes induced by amino acid substitutions. Nodes represent single amino acids of the protein structure, while links represent the non-covalent interactions between them. Default minimum distances were used to define interaction types: 3.0 Å for disulfide bridges; 4.0 Å for salt bridges; 6.0 Å for π-π interaction; 7.0 Å for π-cation interaction. Connectivity, i.e. number of contacts to other residues, and interaction types for each amino acid position found altered were recorded.

5.4. Results

Mutation analysis of the VHL gene has been carried out by the VHL referral laboratory, Department of Pediatrics at the University of Padua, in 426 unrelated subjects with a clinical diagnosis ranging from von Hippel-Lindau syndrome to sporadic potentially VHL-related tumors. A VHL alteration was found in 111 unrelated probands (26% of total unrelated individuals tested): 89 presenting a classic VHL syndrome or a VHL-related tumor and/or a family history of VHL disease, and 22 apparently sporadic cases with isolated lesions: 4 with retinal hemangioblastomas; 12 with CNS hemangioblastomas; 6 with pheochromocytomas. 95 different germline VHL alterations were identified: 4 known polymorphic variants (one small duplication and three SNPs), 38 large rearrangements, 15 frameshift or non sense mutations, 1 in frame deletion, 4 splicing alterations, and 33 missense mutations. Among the 53 small/point mutations, 35 were reported as pathogenic and already listed in the UMD-VHL or HGMD databases (Supplementary Table S.5.1), while the other 18 variants were not found in subjects with VHL Syndrome (Table 5.1). One of the novel variants, p.Arg167Leu, was found in two distinct subjects with VHL syndrome. None of the novel mutations were present in 200 normal control individuals. An apparently novel frame shift mutation consisting in deletion of one of the two cytosines in positions 175 and 176 of the VHL cDNA was also detected. This mutation, c.176delC according to current nomenclature recommendations, is listed in the VHL UMD database as c.175delC (p.Pro59ArgfsX8) [260]. Familial segregation analysis was possible for 10 of the 19 families of individuals carrying novel VHL variants. In five cases the variants were inherited, in the other five cases they were not detected in the parents and therefore considered *de novo*. The VHL disease was also considered likely due to *de novo* germline mutations in other five cases with negative family history, even though not available for segregation studies. In two cases of this latter group (index cases 109 and 196) the variants were transmitted to individuals who eventually developed VHL disease. In the other cases neither parents nor other at risk relatives were available for testing. The novel sequence variants were categorized as inactivating (i.e. stop mutations, frame shifts, or splice site alterations) and non-inactivating (i.e. missense mutations, in frame deletions) (Table 5.1).

Table 5.1 – Clinical impact and segregation of novel VHL mutations.

Mutation classes	Index case	Age	DNA change	Protein change	CNS HB	RHB	PL	PH	RCC	RL	Transmission Familial / de novo	Relatives analyzed	Segregation with phenotype
Inactivating	176	46	c.156G>T	p.Glu52X	x		x				Likely de novo	n/a	-
	258	50	c.253_254ins38	p.Leu85fs	x		x			x	n/d	offspring	-
	225	16	c.314_315insC	p.Arg107ProfsX25	x						Familial	sibling	Yes
	107	24	c.375_376insC	p.Asp126ArgfsX6	x	x					De novo	parents and sibling	-
	80	23	c.422_440del19	p.Asn141IlefsX12	x					x	De novo	parents	-
	196	48	c.463+1G>T	Splicing alteration	x						Likely de novo	grand children	yes
	1	15	c.525C>A	p.Tyr175X	x					x	Likely de novo	father and siblings	-
Non inactivating	117	31	c.465_470del6	p.156_157delYT	x	x	x				De novo	parents and siblings	-
	263	50	c.36G>C	p.Glu12Asp	x						n/d	n/a	-
	364	62	c.175C>T	p.Pro59Ser				x			n/d	n/a	-
	379	38	c.197T>G	p.Val66Gly	x		x	x		x	n/d	n/a	-
	141	24	c.250G>C	p.Val84Leu #				x			De novo	parents and siblings	-
	43	36	c.277G>T	p.Gly93Phe		x		x			Familial	mother	yes
	55	16	c.307C>G	p.Pro103Ala				x			Familial	parents	no
	115	21	c.412C>A	p.Pro138Thr		x		x			Familial	father, siblings, and cousins	yes
	2	43	c.464T>G	p.Val155Gly		x		x			De novo	parents, sibling, and nephew	-
	24	57	c.500G>T	p.Arg167Leu	x	x		x			Familial	parents, sibling, and nephew	yes
	232	31	c.500G>T	p.Arg167Leu	x	x		x			Likely de novo	mother and sibling	-
	109	48	c.563T>G	p.Leu188Arg	x					x	Likely de novo	offspring	yes

Table 0.1. Clinical impact and segregation of novel VHL mutations.

DNA mutation numbering is based on cDNA reference sequence (GeneBank Accession number NM_000551) considering nucleotide +1 as the A of the first ATG translation initiation codon. CNS HB, haemangioblastoma of the central nervous system and spinal cord; RHB, retinal haemangioblastoma; PL, pancreatic cyst or tumour; PH, pheocromocytoma; RCC, clear-cell renal carcinoma; RL, renal cystic lesion. *De novo* indicates that the variant has not been found in parents. Likely *de novo* means that the parents were unaffected even if one or both parents were not available (n/a) for genetic testing. Familial indicates transmission of the variant from a parent. In case index 55, the variant was transmitted from the unaffected father. In some cases, transmission of the variant could not be determined (n/d) since both parents and family history were not available. #The novel G>C transition at nucleotide 250 leads to the previously known amino acid substitution p.Val84Leu.

DNA mutation numbering is based on cDNA reference sequence (GeneBank Accession number NM_000551) considering nucleotide +1 as the A of the first ATG translation initiation codon. Abbreviations: CNS HB: hemangioblastoma of the central nervous system and spinal cord; RHB: retinal haemangioblastoma; PL: pancreatic cyst or tumor; PH: Pheocromocytoma; RCC: clear cell renal carcinoma; RL: renal cystic lesion. *De novo* indicates that the variant has not been found in parents. Likely *de novo* means that the parents were unaffected even if one or both parents were not available (n/a) for genetic testing. Familial indicates transmission of the variant from a parent. In case index 55, the variant was transmitted from the unaffected father. In some cases transmission of the variant could not be determined (n/d) since both parents and family history were not available. # - The novel G>C transition at nucleotide 250 leads to the previously known amino acid substitution p.Val84Leu.

Inactivating mutations

Seven new inactivating mutations were identified: a splice-site variant, two nonsense and four frame shift mutations (Table 5.1). All these (inactivating) mutations were found in individuals with a likely VHL type 1 phenotype (Table 5.1).

The c.463+1T>G variant abolishes the normal donor splice site of intron 2, as predicted by computational analysis (Table 5.2), leading either to alternative splicing or complete exon skipping. The same variant was present in the patient's daughter, who also presented isolated HB. Other nucleotide variants at this position have been previously reported in subjects with VHL Syndrome (Supplementary Table S.5.1).

The nonsense mutations generating premature stop codons at positions 52 and 175 and the three frame shift mutations introducing premature termination codon (PTC) at positions 131 and 152 are all expected to produce VHL proteins largely lacking the C-terminal region (Fig. 5.1). Mutations that interrupt pVHL are predicted to exert a severe pathogenic role due to loss of the α-domain which, by interacting with Elongin C and B, stabilizes the 3D conformation of the protein [278].

In frame deletion

Among the non-inactivating mutations we found the novel in-frame deletion c.465_470del6 (p.156-157delYT), determining loss of the Tyr156 and Thr157 linker loop residues connecting the α and β-domains. This deletion in the pVHL linker region can influence correct orientation of the β-domain and therefore alters substrate positioning for ubiquitin transfer [279]. An alteration of the ubiquitin mediated degradation pathway involves many processes and this could explain the pathogenicity of such a mutation, detected in a patient with a likely VHL type 1 phenotype.

Sequence analysis of missense variants

Ten novel putative missense mutations were identified. Five of these lead to new amino acid substitutions at residues previously found mutated: Gly93, Pro138, Val155, Arg167, Leu188 (Fig. 5.1). The new G>C transition at nucleotide 250 results in the typical VHL type 2C substitution p.Val84Leu [280], which, also in this study, was found in a subject with bilateral PH and no other clinical features of the disease. Three other subjects with isolated clinical manifestations carried variants involving residues previously never found mutated: Glu12, Pro59 and Pro103. The clinical profile of individuals carrying novel missense variants is reported in Table 5.1.

We used two computational methods for splicing prediction in order to exclude that the novel variants disrupt the normal splicing pattern, especially those altering exonic regions close to consensus acceptor and donor splice sites. We applied the methods for a list of 23 known mutations found in this study and three other known mutations (p.Pro154Pro, p.Val155Met, p.Val155Leu) which map on exonic regions close to the donor site of intron 2 (Supplementary Table S.5.2). The latter three variants are all predicted by computational methods to disrupt the donor site of intron 2. For

p.Pro154Pro there is experimental evidence showing the variant to result in a splicing aberration [269] (Supplementary Table S.5.2). However, all novel missense variants are predicted to either maintain the normal splicing pattern (Table 5.2) or, in some cases, the novel sequence was recognized as splice site with a slightly increased score (Supplementary Table S.5.3) .

The combined application of 7 different bioinformatic programs was chosen as a strategy to obtain the most reliable prediction of the functional impact of these mutations. I chosed methods that predict effects on protein stability and those sorting mutations according to their overall pathogenicity. Since computational methods perform with moderate accuracy [281], I calculated a prediction score based on a number of methods that define a variant as deleterious testing 23 pathogenic missense mutations found in this study and previously associated to the VHL disease (Supplementary Table S.5.2). The tested known missense mutations have been predicted to be deleterious by at least 50% of the methods. As negative control we used three known polymorphic VHL gene variants (p.Pro25Leu, p.Ala50Ala, p.Pro61Pro) and the p.His110Tyr variant (rs17855706) found by the NIH Mammalian Gene Collection (MGC) project, which has not been associated to VHL disease. Unfortunately polymorphic variants, such as p.Ala50Ala or p.Pro61Pro, resulting in same sense substitutions could not be used as input for most computational methods. Indeed, in order to calculate an energy difference these methods require a modified residue with respect to the wild type protein. A prediction score of 4/7 was chosen as threshold for novel variants to be potentially deleterious (Supplementary Table S.5.2). For 7 out of 10 missense variants the in *silico* analysis unambiguously predicted deleterious effects (Table 5.2 and Supplementary Table S.5.3). The variant p.Pro103Ala scored below threshold (score: 2/7) and was classified as non deleterious. Transmission from an unaffected father further supported the non pathogenic significance of this variant. The variants pGlu12Asp and p.Pro59Ser, located at the pVHL N-terminus, could not be evaluated with the whole array of methods, as this region is outside the protein crystal structure and is poorly conserved. This is also the case for the N-terminal p.Pro25Leu variant used as negative control (Supplementary Table S.5.2).

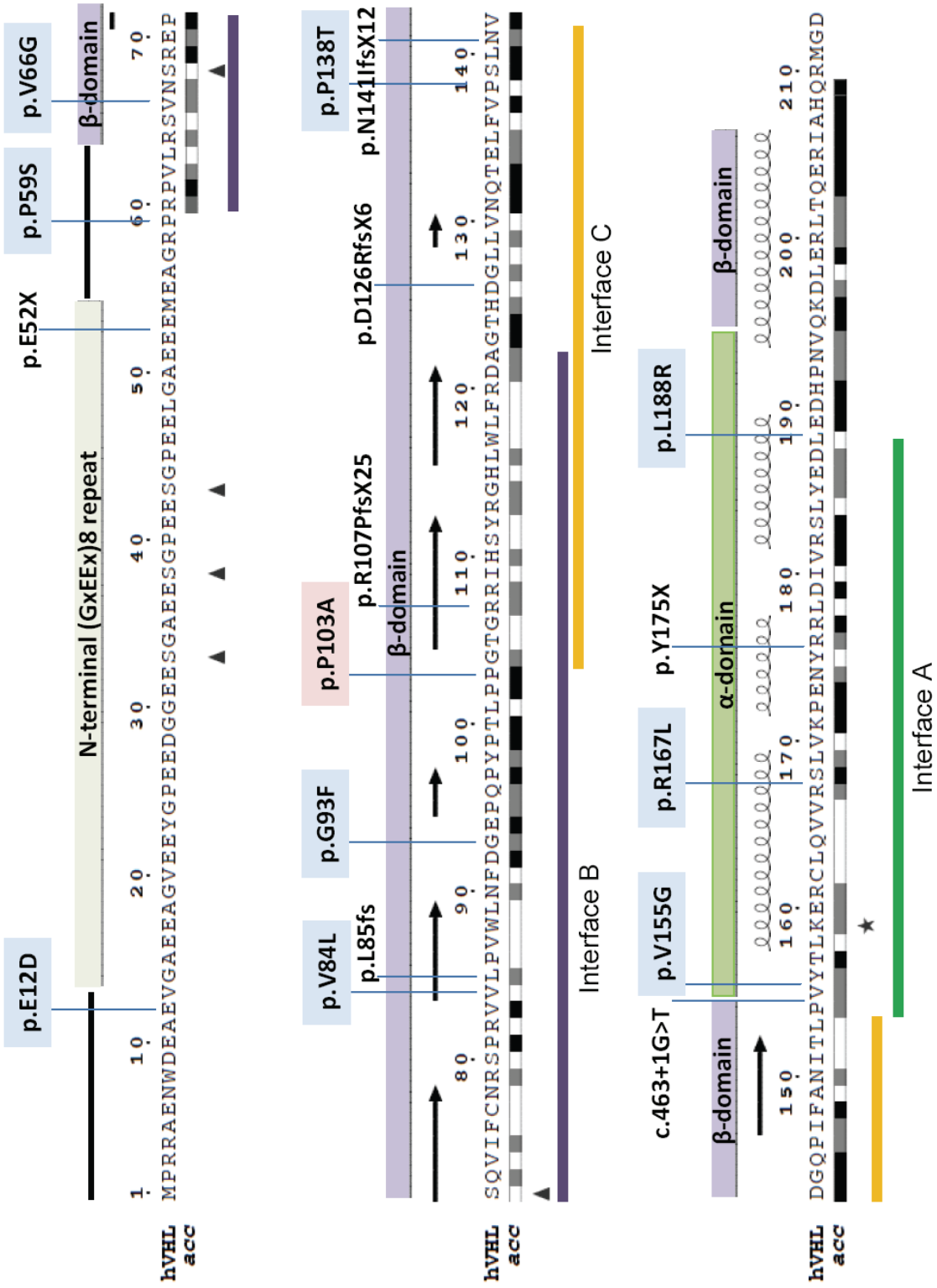


Figure 0.1. Overview of VHL sequence architecture.
The domain organization and secondary structure are shown on the top part. The DSSP accessibility level (*acc*, black = high and white = low), phosphorylation (triangle) and neddylation (star) sites are indicated. Putative protein interaction interfaces are represented as lines at the bottom of the protein

sequence. Novel variants are reported above the sequence. Novel missense variants are indicated with colored boxes with the pink box indicating the only variant predicted to have a neutral effect.

Predicted structural effects of missense variants

To predict the molecular mechanisms of pathogenicity in VHL disease, we evaluated the impact of each new missense variant and studied the role of mutated amino acids on protein stability. We used residue interaction network analysis (Martin A.J.M. et al., submitted) to assess whether the structural effects would cause unfolding and subsequent degradation or local instability which might interfere with protein-protein interactions.

The crystal structure of human pVHL complexed with Elongin C and B (PDB code 1LM8) was chosen for *in silico* evaluation of the structural and functional effects of the 8 novel amino acid substitutions altering residues between the positions 60 and 209 (Fig. 5.2). All detected variants alter the β-domain of the protein, except p.Arg167Leu and p.Leu188Arg (located in the pVHL α-domain).

Two novel missense variants (p.Pro138Thr and p.Leu188Arg) alter residues of the protein core (Fig. 5.2 and Fig. 5.3), suggesting a pathogenic role. These variants are predicted to introduce polar/charged residues in the hydrophobic core, thereby destabilizing protein folding. Residue interaction network analysis with RING revealed these amino acids to have high connectivity, i.e. number of contacts with nearby residues. This is indicative of their centrality in the protein fold and suggests an important structural/functional role (Supplementary Table S.5.3). It is interesting to note that the p.Pro138Thr substitution should have less impact on protein stability because it maintains the only hydrogen bond with Trp117.

The role of Arg167, located in the α-domain, is crucial in forming the charged interface between two domains. Substitution with leucine alters the electrostatic surface and could influence correct folding of the α-domain. Furthermore, given the central position between the two domains, a substitution at that position could cause a conformational change transmitted to other parts of the protein (Fig. 5.2).

Mutation	Location	Conservation	Splicing prediction	Pathogenicity prediction	Predicted effect
c.156G>T; p.Glu52X	N-terminus	Non conserved	unchanged	pathogenic	Reduces protein stability
c.253_254ins38; p.Leu85fs	β-domain	Conserved	nd	pathogenic	Reduces protein stability
c.314_315insC; p.Arg107ProfsX25	β-domain	Conserved	unchanged	pathogenic	Reduces protein stability
c.375_376insC; p.Asp126ArgfsX6	β-domain	Conserved	unchanged	pathogenic	Reduces protein stability
c.422_440del19; p.Asn141IlefsX12	β-domain	Conserved	unchanged	pathogenic	Reduces protein stability
c.463+1G>T	β-domain	Conserved	Disrupts donor splice site intron 2	pathogenic	Abnormal protein folding
c.525C>A; p.Tyr175X	α-domain	Conserved	unchanged	pathogenic	Reduces protein stability
c.465_470del6; p.156_157delYT	linker region between α- and β-domains	Conserved	Increased score acceptor splice site intron 2 (0.84>0.93)	pathogenic	Interface B
c.36G>C; p.Glu12Asp	N-terminus	Non conserved	unchanged	ambiguous	Inconclusive
c.175C>T; p.Pro59Ser	N-terminus	Non conserved	unchanged	ambiguous	Likely affects N-terminal VHL functions
c.197T>G; pVal66Gly	β-domain surface	Non conserved	unchanged	pathogenic	Compromises GSK3 mediated phosphorylation
c.250G>C; p.Val84Leu	β-domain core	Conserved	unchanged	pathogenic	Interface B
c.277G>T; p.Gly93Phe	β-domain surface	Conserved	unchanged	pathogenic	Interface B
c.307C>G; p.Pro103Ala	β-domain surface	Conserved	unchanged	neutral	None
c.412C>A; p.Pro138Thr	β-domain core	Conserved	unchanged	pathogenic	Reduces protein stability
c.464T>G; p.Val155Gly	linker region between α and β-domains surface	Conserved	Increased score acceptor splice site intron 2 (0.84>0.95)	pathogenic	Interface A
c.500G>T; p.Arg167Leu	linker region between α and β-domains	Conserved	unchanged	pathogenic	Interfaces A and B
c.536T>G; p.Leu188Arg	α-domain core	Conserved	unchanged	pathogenic	Reduces protein stability

Table 0.2. Molecular effect prediction of novel VHL mutations.
The table summarizes results obtained by different computational methods used to predict possible splicing aberrations, stability changes on protein and pathogenic structural/functional effects. Conservation is derived from ConSurf which classifies each residue as variable (value 1-3), average (value 4-6), or conserved (value 7-9). For missense variants the values are reported on Supplementary Table S.5.3. Possible splicing aberrations were predicted using the splice site prediction methods NNSplice and NetGene2. We classified a variant as pathogenic when it impairs protein function by creating a premature truncation codon (PTC) or by altering the transcript. Pathogenicity prediction for missense variants was obtained comparing the results from 7 different methods: I-Mutant 3.0, AutoMute, Polyphen, SNPs3D, Pmut and SNAP (Supplementary Table S.5.2). A missense variant was classified as

deleterious when more than 4 methods over 7 predict it as deleterious. Two variants remain ambiguous since most of the methods fail to get a result for them. Surface variants are predicted to alter the three interfaces (A, B, C) of VHL protein which have roles, respectively, in VCB complex formation, substrate recognition, and localization. Abbreviation: n/d not determined.

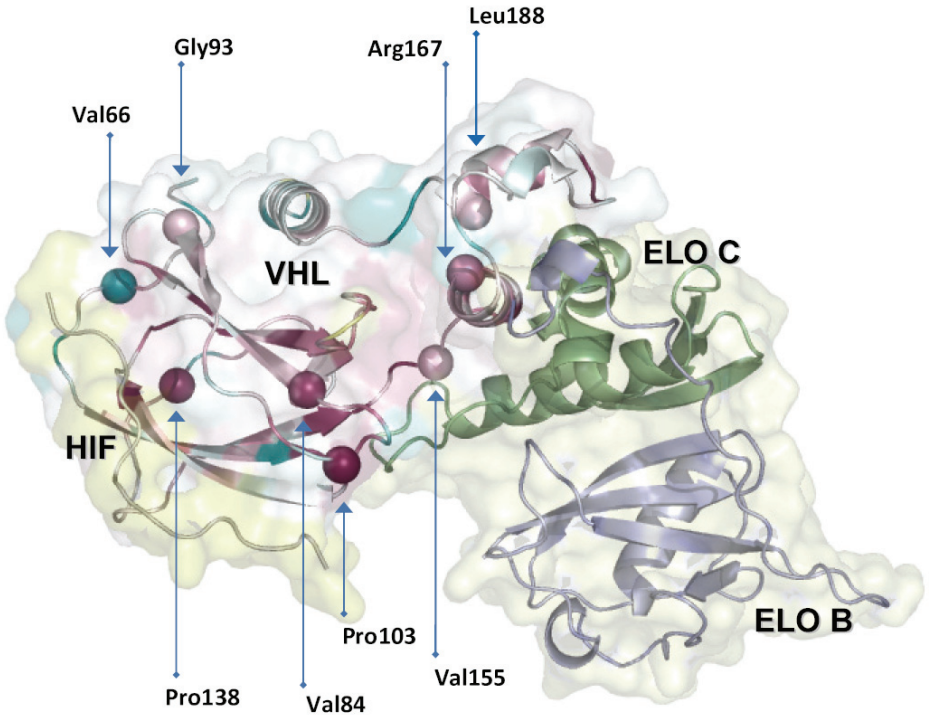


Figure 0.2. Mapping of missense mutations on the VHL structure.
The VHL structure (PDB code: 1lm8) is shown as cartoons, with semi-transparent surface, together with Elongin B (ELOB) and C (ELOC) and the HIF peptide. Mutated pVHL residues are shown as spheres and the degree of conservation is mapped on the structure from magenta (highly conserved) to cyan (unconserved).

Surface variants (pVal66Gly, p.Gly93Phe and p.Val155Gly) were all predicted to be pathogenic (Supplementary Table S.5.3) and to have the potential to disrupt protein binding (Fig. 5.4). Comparing interaction networks of wild-type versus mutant pVHL, we observed how p.Val155Gly lost only one interatomic contact with Arg161 while p.Gly93Phe variant forms a new contact and a new π-cation interaction with Arg64. This induces a new local conformation with the phenylalanine side chain protruding into a hydrophobic pocket and no structure destabilization (Fig. 5.3).

Finally, p.Val66Gly forms a new hydrogen bond with Ser68 (Supplementary Table S.5.3). All these modifications lead to local changes in protein structure. Among missense mutations, only p.Pro103Ala was predicted to have a neutral effect on protein structure. Despite the high conservation of Pro103 it was not possible to identify any

structural or functional role of this residue, neither by visual inspection nor by residue interaction network analysis (Fig 5.3).

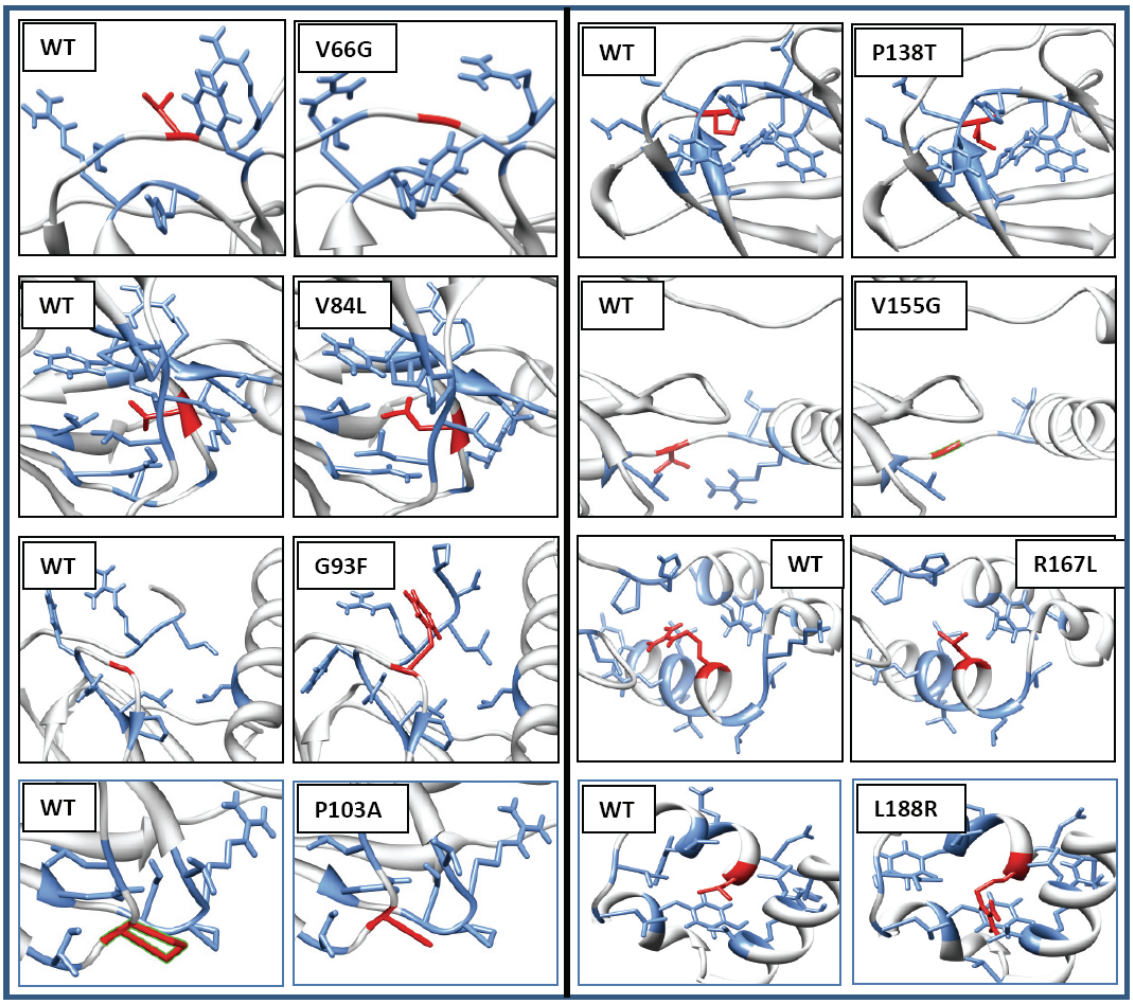


Figure 0.3. Structural effect of novel missense variants.
The impact on VHL structure of each missense variant is shown with respect of the wild type (WT) protein. The mutated residue is highlighted in red and contacting residues in blue as sticks.

Predicted functional effects of missense variants

To predict the functional effects of new missense variants, we have considered the interactions of altered residues with known pVHL interactors (Fig. 5.4). For many of these proteins the pVHL interacting region has been experimentally determined and I previously proposed a structural hypothesis distinguishing at least three different interfaces (termed A, B and C) corresponding to different functions [267]. The basic hypothesis for this functional evaluation is that substitutions occurring at regions essential for protein binding may affect this interaction. Our findings have been

compared with experimentally verified functional alterations of variants occurring at the same position or at neighboring residues (Fig. 5.4) and the results are summarized in Table 5.2.

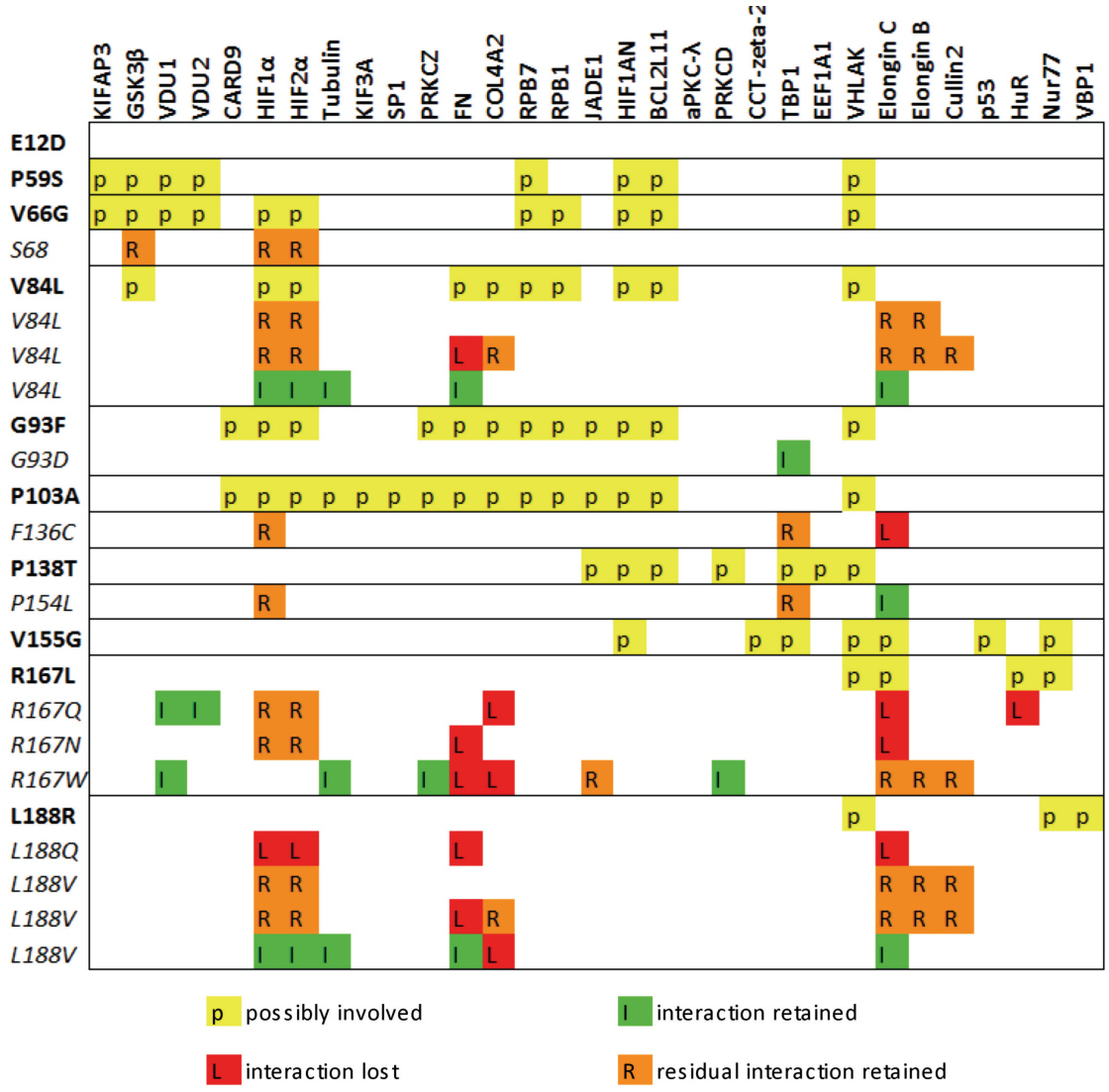


Figure 0.4. pVHL interactions of known and new variants at similar positions.
The list of proteins interacting with pVHL was taken from [267]. A yellow box (p) indicates the possible involvement in pVHL interactions for positions found newly altered. Colored boxes indicate results from experimental studies of variants occurring at the same or neighboring residues, as follows: interaction retained (I, in green), residual interaction retained (R, in orange), interaction lost (L, in red). Different experimental results are reported for the same variant in two cases: p.Val84Leu, p.Leu188Val.

5.5. Discussion

This study reports the results of bioinformatic analysis of VHL mutations found in individuals with phenotypes ranging from full clinical von Hippel-Lindau disease to isolated tumors that could represent the first manifestation of this condition or just isolated lesions. The availability of a highly dependable genetic test, characterized by a mutation detection rate close to 100%, allows to reliably identify VHL affected individuals or virtually exclude this diagnosis. A known or putative VHL disease-causing alteration was found in 89/89 of the cases referred with a clinical diagnosis of von-Hippel-Lindau disease and in 22 subjects presenting with apparently sporadic VHL-related tumors at the time of molecular diagnosis (Supplementary Table S.5.1).

Once a VHL sequence variant is detected for the first time in an individual with negative family history, the interpretation of its pathogenic significance is an important step on which a final diagnosis and subsequent clinical follow-up depend. This interpretation, often not easy, is particularly critical if the variant does not obviously impair the protein function by creating a premature truncation codon (PTC) or by altering the transcript. This work was specifically intended to deal with this interpretative issue in characterizing the 18 newly detected VHL variants (Table 5.2). Besides explaining the pathogenic role of inactivating mutations, I focused on the characterization of the in frame deletion of two amino acids and on the 10 single nucleotide substitutions believed to be putative missense mutations.

I evaluated the impact of each new sequence variant and studied the role of mutated amino acid residues on overall protein stability with the aim of predicting molecular mechanisms of pathogenicity in VHL disease. By residue interaction network analysis I predicted whether the structural effects determine unfolding and subsequent degradation, or local instability which might interfere with protein-protein interactions (Table 5.2).

Inactivating mutations

Mutations that interrupt pVHL, even though not early truncations, are all predicted to exert a very severe pathogenic role due to the loss of the structurally crucial α-domain, which interacting with Elongin C and B allows the 3D conformation of the protein [278]. Furthermore, mutations resulting in a PTC should anyway impact on the mRNA

stability since normally subjected to mRNA surveillance and nonsense-mediated mRNA decay (NMD) [282].

Seven new inactivating mutations have been identified in subjects with a suspected clinical diagnosis of VHL Syndrome (Table 5.1). One of the inactivating mutations (p.Tyr175X) was previously described as a somatic alteration in sporadic RCC [283]. It might escape the NMD machinery since the resulting PTC is located in the last VHL exon and NMD typically degrades transcripts containing nonsense codons followed by at least one intron [284-285]. Even if processed, this transcript would still result in a pVHL molecule lacking an essential part of the α-domain (Fig. 5.1), the Cullin2 interaction region [246], explaining the pathogenic relevance of such a mutation.

The subject carrying the p.Tyr175X variant presented multiple central nervous system HB without PH, which would be consistent with a mutation abolishing VHL gene function [226]. It is worth mentioning that an individual from a VHL type 1 family with the same pVHL truncation, although from a different genomic variant, was reported to be affected by bilateral PH [286]. No follow up clinical information was available on our p.Tyr175X patient who unfortunately died at age 17 due to complications of a brainstem HB. A unique case is represented by the p.Pro59ArgfsX8 mutation that we found associated with PH (Supplementary Table S.5.1), a tumor that should not be associated with alterations predicted to completely abolish pVHL function. Genetic or epigenetic modifiers would reconcile the apparent genotype-phenotype correlation discrepancy in these two mutations, as recently suggested for the risk of RCC and HB [287].

Predicted effects of novel missense variants

Ten novel missense variants have been detected in eleven unrelated individuals (Table 5.1). Of these, seven subjects were referred for VHL disease (two of them carrying the same variant), one presented with an apparently sporadic CNS hemangioblastoma, and three with seemingly sporadic PH. The segregation of the four variants p.Gly93Phe, p.Pro138Thr, p.Arg167Leu, and p.Leu188Arg with the phenotype in other family members further supported their pathogenicity, while for index case 55 the segregation analysis tended to exclude a correlation between p.Pro103Ala and the patient phenotype. Of great relevance was to establish the pathogenicity of the other newly identified

variants, particularly if represented by amino acid substitutions detected in subjects with isolated tumors and unreported family history. One of these, the *de novo* c.250G>C transition found in a case of sporadic pheochromocytoma, results in the known variant p.Val84Leu previously reported to be associated to type 2C VHL. The interpretative problem remained for four variants: p.Glu12Asp, p.Pro59Ser, p.Val66Gly, and p.Val155Gly.

Prediction of pathogenicity for each of the novel missense variants was based on conservation of the mutated residue across orthologous proteins, and in silico prediction of its putative pathogenic effect. Except for variants mapping on the VHL N-terminal region (p.Glu12Asp, p.Pro59Ser, p.Val66Gly), all novel amino acid substitutions occur at conserved positions. It is generally accepted that the canonical splice donor and splice acceptor sites are well recognized although, alterations occurring at positions close to these highly conserved sites are more difficult to predict a priori [288]. However, I excluded that the novel variants could exert their pathogenic role disrupting the normal splicing pattern (Table 5.2 and Supplementary Table S.5.3).

The employed computational methods for prediction of mutation effects performed reasonably well for our test data on VHL, with Eris showing the greatest overall accuracy (Supplementary Table S.5.2). This is not unexpected, given the complexity of the Eris method, which is associated with a larger computational cost due to the thermodynamic calculations. As a faster alternative, Pmut also performed consistently well, correctly predicting the known benign variation and having the advantage that no protein structure is required (Supplementary Table S.5.2). It should be emphasized that all *in silico* predictions can only give an approximation for pathogenicity and can of course not replace experimental validation. In particular, prediction of the effects related to splice site variants and protein interactions require specialized knowledge for the protein under consideration and cannot be easily automated. With this important caveat, the computational predictors can nevertheless provide a fast screening tool to focus further experimental efforts.

The application of seven different computational methods to predict overall pathogenicity of the 10 novel missense variants allowed us to affirm that 7 of these are potentially true pathogenic mutations (Table 5.2 and Supplementary Table S.5.3). Variant p.Pro103Ala could be considered a new benign variant. The major difficulty

using this strategy was to analyze variants occurring at the N-terminus (p.Glu12Asp, p.Pro59Ser). Prediction methods using structural information were indeed not applicable (Supplementary Table S.5.3) and given the low sequence conservation of the N-terminal region, methods based on evolutionary conservation failed to classify these variants. This was also observed for the known benign variant, p.Pro25Leu, reported in the UMD-VHL database as a probably polymorphic variant, that we found in a healthy control patient (Supplementary Table S.5.2). The p.Glu12Asp substitution nonetheless maintains the negative charge of the N-terminal sequence, which seems to be its most important characteristic. This variant is therefore likely not associated with the clinical phenotype of the carrier, referred as the only affected individual of his family, even if no other family members were available for genetic testing. As for p.Pro59Ser, it was predicted to be neutral by three out of four applicable computational methods (Supplementary Table S.5.3). The subject carrying the variant only presented a monolateral PH and a totally silent family history, rendering the pathogenicity of this variant ambiguous (Table 5.2).

To determine effects on protein stability and function, we mapped the novel missense variants on the VHL crystal structure. Our studies show that while two variants are located on the unstructured N-terminus, three mutations affect residues buried in the protein core, one position at the linker region between α and β-domain and four variants affect surface residues. To predict the functional effects of new missense variants, I have considered the positions of altered residues with respect to the three interfaces used by pVHL to interact with partner proteins (Fig. 5.1, Table 5.2, Fig. 5.4). I recently reported an hypothesis by which the three interfaces have different molecular functions. Interface A has a role in protein processing, which comprises ubiquitin ligase complex formation and proteins stabilization, interface B in substrate recognition and interface C in pVHL subcellular localization [267]. The basic hypothesis for this functional evaluation is that substitutions occurring at regions essential for protein binding may affect this interaction.

From a structural point of view, p.Pro138Thr and p.Leu188Arg mutations were predicted to cause complete unfolding and consequent loss of all pVHL functions. This prediction is supported by loss of interactions in both α and β domains and unfolding, as demonstrated for the p.Leu188Gln substitution [289].

Substitutions at Arg167 could have a dominant-negative effect due to partial unfolding and inability to interact with the VCB complex [289-290]. The functional effect of these mutants was reported to be milder, since they likely retain interactions with proteins that do not undergo ubiquitin mediated degradation [291-293].

Our prediction methods indicated that surface mutations do not cause protein unfolding and allow the formation of a stable VCB complex. These mutations nevertheless have the potential to alter functional interfaces of pVHL, as experimentally proven for numerous substitutions at Tyr98 [249-250, 289, 291, 293-295], the residue forming the main interaction with HIF1α. This is also expected for the p.Gly93Phe mutation which alters the same interface, while p.Val155Gly was previously predicted to be involved in interactions with Elongin C [296].

With the exception of a stop codon, no other variants of the β-domain surface residue 66 have been reported. The structural analysis reported here shows that p.Val66Gly causes a local change involving Ser68 which is the phosphorylation site for Glycogen synthase kinase 3 (Supplementary Table S.5.3) [255]. It is tempting to hypothesize that the functional relevance of a substitution at position 66 might therefore be due to alteration of the kinase binding motif.

I have re-evaluated the possible functional implications of the N-terminal substitution of Pro59 by a polar serine residue, which could not be analyzed with the whole array of prediction methods and was left with an unclear interpretation (Table 5.2 and Supplementary Table S.5.3). This variant alters a residue with low conservation but the substitution introduces different biochemical properties that could interfere with functions hypothesized for the pVHL N-terminus, such as microtubule stabilization [255, 297], proper fibronectin matrix deposition [291, 298-299] and ciliary-maintenance mechanisms [255, 300].

Considerations about genotype-phenotype correlations

Consistent with the severe functional impact attributed to variants with a core location, p.Leu188Arg mutation was associated with a phenotype that did not include PH, making it a likely VHL type 1. The fact that p.Pro138Thr was detected in a subject with PH suggests that this core mutation does not cause complete protein unfolding and may be able to maintain interactions with pVHL partners. The latter has been demonstrated

for other core mutations, reported as typically associated with VHL type 2C, e.g. p.Leu188Val and p.Val84Leu [293-294, 299, 301]. Furthermore, other previously described mutations altering the pVHL interface C between residues 114 and 154, such as p.Phe136Cys and p.Pro154Leu, have reduced ability to bind TBP1, necessary for proteasome binding (Fig. 5.4), and have been found in VHL type 2 [250].
All of the new surface variants are associated with PH, confirming that the risk of PH is indeed higher with missense than with loss of function mutations [227].
Finally, it is interesting to note how all surface variants in this study alter pVHL interactions with proteins promoting apoptosis, e.g. p53 [244] , JADE-1 [302], BCL2L11 [303]. This observation seems particularly relevant in view of recent data showing how VHL type 2C mutant proteins are implicated in decreased apoptosis and indicating this mechanism as possibly responsible for PH [304].

5.6. Conclusions

This molecular study increase the list of known VHL mutations and contributes to a better understanding of the molecular pathology of this tumor suppressor gene. I proposed a *in silico* strategy for the evaluation and interpretation of the pathogenicity of novel sequence variants. The adopted computational approach allowed to predict the impact of aminoacid substitutions on the overall stability of pVHL, interference with specific interfaces and possible allosteric effects which could disturb the demonstrated allosteric correlation between the α- and β-domain binding sites [305]. Although not specifically aimed at evaluating genotype–phenotype correlation, this study allowed to observe how classifying missense substitutions according to their predicted effects on pVHL structure enhances the ability to predict the risk of PH occurrence. By integrating genetic information and predicted impact on the protein structure it has been possible to reliably classify as disease-causing 15 of the 18 newly detected VHL variants. An unambiguous interpretation of mutations has an important clinical impact, both in terms of genetic counseling and clinical surveillance and follow up.

6. A computational model of the LGI1 protein suggests a common binding site for ADAM proteins

This chapter has been published in Leonardi E, Andreazza N, Vanin S., Busolin G., Nobile C. and Tosatto S.C.E. A computational model of the LGI1 protein suggests a common binding site for ADAM proteins. PLos ONE 6(3): 2011 March 29,;6(3):e18142

and

Striano P, Busolin G, Santulli L, Leonardi E, Coppola A, Vitiello L, Rigon L, Michelucci R, Tosatto SC, Striano S, Nobile C. Familial temporal lobe epilepsy with psychic auras associated with a novel LGI1 mutation. Neurology. 2011 Mar 29;76(13):1173-6.

6.1. Summary

Mutations of human leucine-rich glioma inactivated (*LGI1*) gene encoding the epitempin protein cause autosomal dominant temporal lateral epilepsy (ADTLE), a rare familial partial epileptic syndrome. The *LGI1* gene seems to have a role on the transmission of neuronal messages but the exact molecular mechanism remains unclear. In contrast to other genes involved in epileptic disorders, epitempin shows no homology with known ion channel genes but contains two domains, composed of repeated structural units, known to mediate protein-protein interactions.

A three dimensional *in silico* model of the two epitempin domains was built to predict the structure-function relationship and propose a functional model integrating previous experimental findings. Conserved and electrostatic charged regions of the model surface suggest a possible arrangement between the two domains and identifies a possible ADAM protein binding site in the β-propeller domain and another protein binding site in the leucine-rich repeat domain. The functional model indicates that epitempin could

mediate the interaction between proteins localized to different synaptic sides in a static way, by forming a dimer, or in a dynamic way, by binding proteins at different times.
The model was also used to predict effects of known disease-causing missense mutations. Most of the variants are predicted to alter protein folding while several others map to functional surface regions. In agreement with experimental evidence, this suggests that non-secreted LGI1 mutants could be retained within the cell by quality control mechanisms or by altering interactions required for the secretion process.
The Arg407Cys is the first mutation with no effect on LGI1 protein secretion. Substitution of Arg407 with a cysteine is predicted to have no effect on propeller domain folding but, under the strongly oxidative conditions present in the extracellular environment, likely forms abnormal disulfide bridges with other molecules, ultimately hampering interaction of LGI1 with its partner protein(s). The uncommon isolated psychic symptoms associated with this mutation suggests that ADLTE encompasses a wider range of auras of temporal origin than hitherto reported.

6.2. Introduction

The human leucine rich, glioma inactivated 1 (*LGI1*; GeneID 9211; MIM# 604619) gene has been linked to two different clinical phenotypes: malignant progression of glioma and autosomal dominant lateral temporal epilepsy (ADLTE; MIM# 600512), a rare familial partial epilepsy syndrome. This gene has been shown to be frequently downregulated in malignant gliomas and to regulate invasiveness of some glioma cell lines [306] by driving the expression of matrix metalloproteinases through the ERK 1/2 pathway. These findings suggest that *LGI1* may serve as a tumor metastasis suppressor gene [307].
ADTLE is an inherited epileptic syndrome characterized by focal seizures with predominant auditory symptoms likely originating from the lateral temporal lobe cortex [308-309]. Mutations causing ADLTE were identified in the *LGI1* gene by positional cloning [310-311]. To date, over 25 mutations have been reported, resulting in either protein truncation or single amino acid substitutions [312], but about half of the ADLTE families have no *LGI1* mutations [308]. *LGI1* is mainly expressed in neurons [311,

313] and shows no similarity to known ion channels. The predicted structure of the LGI1 protein comprises, starting from the N-terminal end, a signal peptide, four leucine-rich repeats (LRR) flanked on both sides by conserved cysteine clusters [96], and seven copies of a repeat of about 45 residues, named EPTP [314] or EAR [315], probably forming a beta-propeller structural domain [316]. Both LRR and beta-propeller domains mediate protein-protein interactions, each motif defining a distinct family of proteins [316-317].

Several different functions and molecular partners have been attributed to LGI1. A recent study provided evidence that LGI1 is associated with a post-synaptic complex containing PSD95 and ADAM22, a receptor associated with the post-synaptic membrane [318]. Through specific binding to ADAM22, LGI1 was shown to participate in the control of synaptic strength at excitatory synapses, whose malfunction may result in epilepsy [318]. Mouse models developed more recently have implicated LGI1 in neuronal maturation processes. In one study, it was shown that LGI1 affects postnatal maturation of glutamatergic synapses, a process involving ADAM22, and mediates dendrite pruning so that LGI1 mutations would result in persistence of immature, untrimmed, dendritic arbor [319]. On the other hand, another study showed that LGI1 preferentially interacts with ADAM23 and through this receptor, which is not located at postsynaptic density, stimulates neurite outgrowth *in vitro* and dendritic arborisation *in vivo* [320]. Finally, analysis of *LGI1* knock-out and transgenic mice suggested that LGI1 may act as a trans-synaptic protein connecting the pre-synaptic ADAM23 with the post-synaptic ADAM22 receptors [321].

To help understand the three dimensional (3D) conformation of LGI1, its binding properties, and ultimately its function(s), we developed an *in silico* model of the protein structure and analysed the amino acid sequence of the LRR and beta-propeller LGI1 domains as well as their phylogenetic relationship. The models were used to assess the significance of known missense mutations. Analysis of possible interaction mechanisms with other proteins suggests a conserved common binding site for members of the ADAM protein family.

6.3. Materials and Methods

Sequence feature analysis

We employed an integrative bioinformatics approach combining sequence and domain database searches with the consensus from predictions of protein structural features. The LGI1 sequence (accession code: O95970) was downloaded from the SwissProt/TrEMBL database [322]. Homologous sequences were retrieved and selected with BLAST [37] from the SwissProt database using standard parameters and visualized using Jalview [43] and ESPript [84]. The secondary structure of LGI1 was predicted using the *consensus* method [99]. Prediction of intrinsic disorder was performed using Spritz [111] and the presence of signal peptides assessed with SignalP [323]. Repetita [97] was used to predict repeat periodicities.

Phylogenetic analysis

In order to reconstruct the phylogeny of the LGIs, 105 vertebrate and one branchiostomid epitempin sequences have been automatically extracted from the available databases using BLAST [37] searches. Full-length amino acid sequences have been recovered from the corresponding nucleotide mRNA or genomic sequences. Multiple alignment was constructed with CLUSTALW [41]. The final alignment has been manually refined at the variable N-terminus and used in the subsequent analysis.

A preliminary quartet puzzling analysis has been performed with the Treepuzzle program [324-325] to test whether a phylogenic approach could be applied to the original data set. Phylogenic studies have been performed according to the maximum likelihood (ML) with the PHYML 2.4 program [44]. The JTT substitution matrix [326] was used during reconstruction, whereas site heterogeneity was modeled with a four-category Γ distribution. Nonparametric bootstrap resampling (BT) [327] was performed with 1,000 replicas to test the robustness of the tree topology. The phylogenetic tree was visualized with the Fig Tree 1.1.1 program (http://tree.bio.ed.ac.uk/software/figtree/).

Alignment construction

Structural templates for the two LGI1 domains were found using MANIFOLD [239]

and MetaServer [328]. Initial alignments were generated through systematic parameter variation from an ensemble of similar alternatives [329]. Given the problematic nature of repeated sequences, the best initial alignment was used as a starting point only. Manual refinement consisted in a method similar to ABRA [330] and Kajava's method [95], with knowledge about the approximate location and number of repeats serving to identify the true repeat boundaries. Knowledge of key residues and secondary structure was used to anchor the aligned repeats.

Molecular modeling

Models for the two LGI1 domains were constructed using the HOMER server (*URL: http://protein.bio.unipd.it/homer/*). The server uses the conserved parts of the structure to generate a raw model, which is then completed by modeling the divergent regions with LOBO, a fast divide and conquer method [88]. Side chains are placed with SCWRL3 [331] and the energy evaluated with FRST [89]. The final models were subjected to a short steepest descent energy minimization with GROMACS [90] to remove energy hotspots before calculating the electrostatic surface with APBS [78]. Evaluation of model quality was performed with QMEAN [100-101]. The structure is visualized using PyMOL (DeLano Scientific, URL: http://pymol.sourceforge.net/). Position-specific conservation scores for each amino acid were calculated with ConSurf [242].

Mutation analysis

Amino acid substitutions have been mapped on the LRR and EPTP domain models and their position evaluated by manual inspection. Four computational methods were used to predict the stability change of the structure caused by these mutations. While I-Mutant 2.0 [190] and MuPro [187] both utilize support vector machines or neural networks to predict the effect of the substitution on protein stability, Eris [188] and PoPMuSiC v2.0 [189] calculate mutational free energy changes of the protein based on its 3D structure.

Cell transfection assay

To ascertain the functional consequences of the Arg407Cys mutation, we transfected

the wild type and 1219C>T LGI1-Flag cDNAs into HEK293 cells, which do not express endogenous Lgi1, and analysed the proteins produced by these cells by immunoblot. Both cell lysates and concentrated (about 20x) serum-free media were analysed using anti-Lgi1 and anti-Flag antibodies. Although some signal was retained in the cell lysate, the Arg407Cys mutated protein was mostly secreted into the medium as was the wild type protein, whereas a mutant protein carrying the ADLTE-causing mutation p.Ala110Asp (c.329C>A), assayed as control, was detected only in the cell lysate (Fig 6.1C). Thus, the Arg407Cys is the first mutation identified in LGI1 that does not affect the secretion process of the protein in culture cells.

6.4. Results and Discussion

Given the fragmented knowledge present in the literature, we performed a full analysis of the LGI protein family starting from the protein sequence. In the following, we will address each step from phylogeny to sequence and structural analysis all the way to new functional hypotheses.

Phylogenetic analysis

The phylogenetic reconstruction was performed using 105 Vertebrate (Chordata; Chraniata) sequences. An additional sequence of *Branchiostoma floridae* (Chordata; Cephalochordata) has been included in the analysis. The obtained reconstruction reported in Figure 6.1 highlights the presence of 4 groups, named 1, 2, 3 and 4. The distribution pattern of LGI family transcripts in the adult mouse brain [332] highlights the tissue specificity of group 1 (see Figure 6.1). Group 1, 2 and 3 present the fish sequences (blue squares) in a basal position, followed in group 1 and 3 by amphibian and bird sequences (red and green arrows). The mammalian sequences present an apical position in all the groups. The *Ornithorhynchus anatinus* protein shares a common node with chicken in group 1 and both are basal to the other mammalians. The phylogeny of LGI1 reveals an early duplication of the gene followed by two other independent duplications as already reported by Gu et al [333], but, in contrast to these authors, the

phylogeny obtained with a larger dataset indicates a closer relationship between the LGI3 and LGI4 sequences as opposed to LGI1 and LGI4.

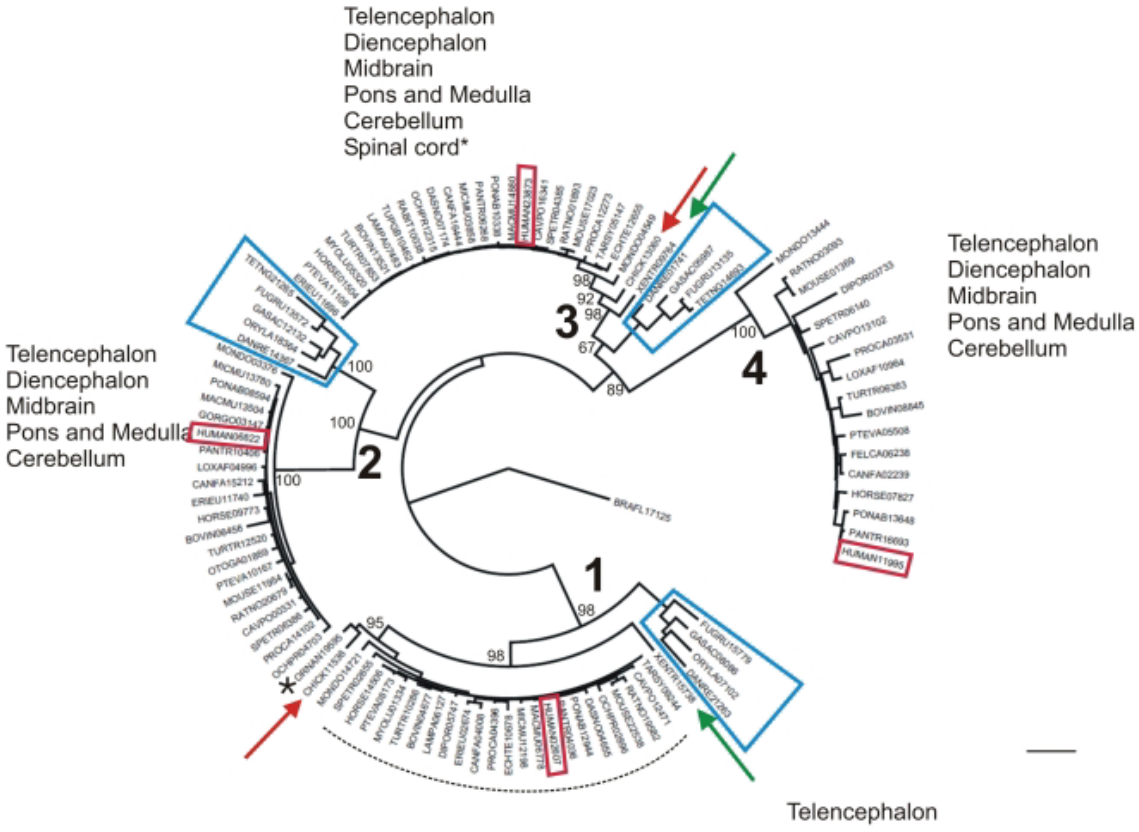


Figure 0.1. Evolutionary relationship among the LGI vertebrate sequences.
The figure shows the best likelihood tree (-lnL= -21148.01332) obtained using the PHYML program. The length of the branches represents the number of reconstructed change of state over all sites (bar represents 0.2 substitutions per site), bootstrap values are reported at the nodes. Blue squares indicate the fish sequences whereas the green and red arrows respectively the amphibian and bird sequences. An asterisk indicates the *Ornithorhynchus anatinus* protein.

Sequence domain organization

We define boundaries of each domain in the LGI1 sequence (Fig. 6.2). The first 35 N-terminal residues contain the signal peptide responsible for its secretion. A cleavage site is also predicted by SignalP in this region. The N-terminal part of the protein from residues 41 to 243 has about 30% sequence identity with LRR domain family proteins, while the C-terminal region between residues 245-552 contains the EPTP repeats. The two domains are also present in all human LGI proteins (LGI1, LGI2, LGI3, LGI4) and conserved across orthologs (Fig. 6.2). Since a structure of LGI1 is not available, a structural analysis was conducted separately for the two domains as they have different characteristics.

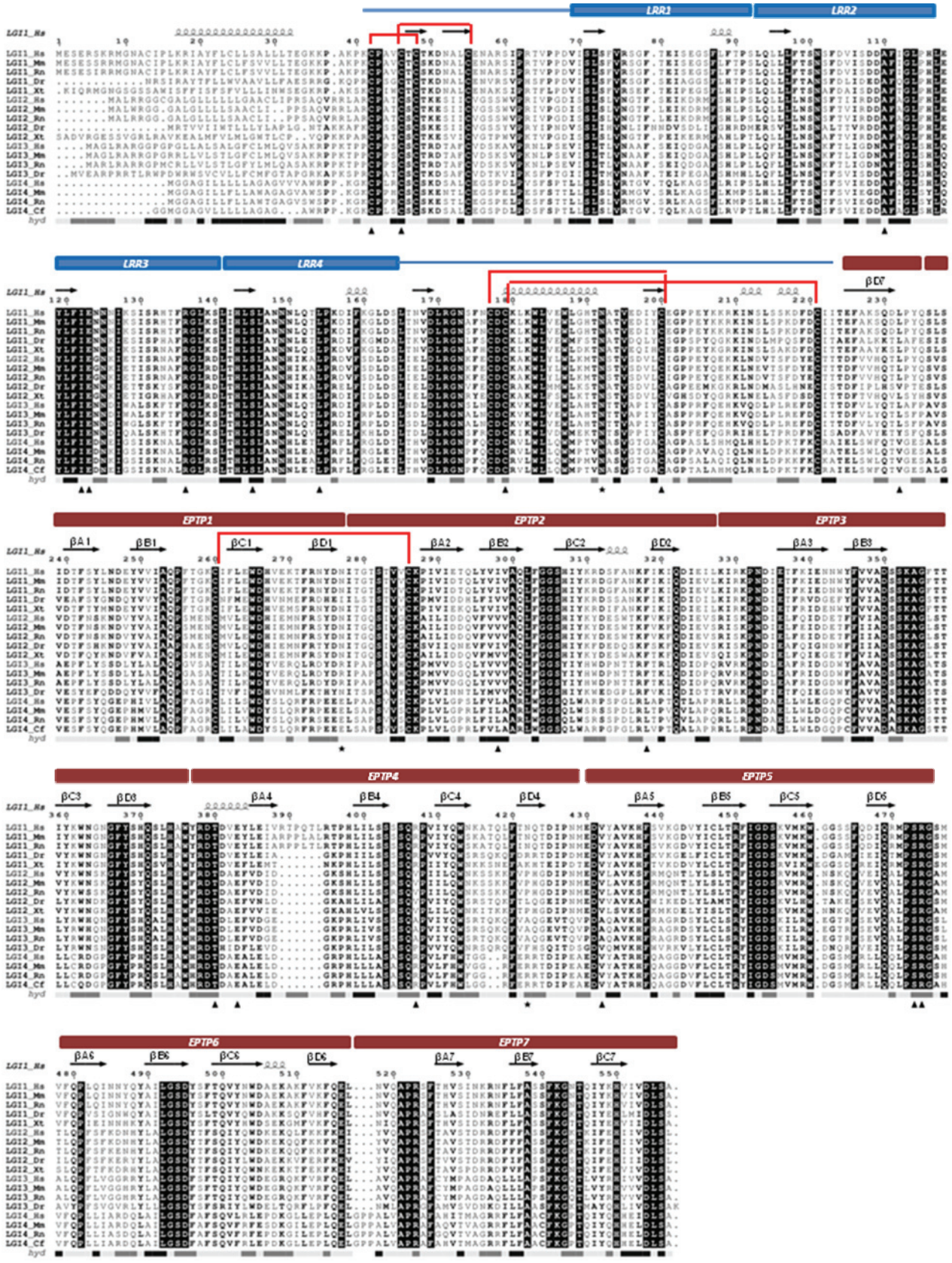

LRR1
LRR2
LRR3
LRR4
βD7
EPTP1
EPTP2
EPTP3
EPTP4
EPTP5
EPTP6
EPTP7

Figure 0.2. Alignment of LGI family members and domain organization.
Multiple alignment of representative homologs in the LGI family. Species are abbreviated as follows: Hs = *Homo sapiens*; Mm = *Mus musculus*; Rn = *Rattus norvegicus*; Dr = *Danio rerio*; Xt = *Xenopus tropicalis*; Cf = *Canis familiaris*. The LGI1 domains and secondary structure are shown on the top part. Missense mutations analyzed in this paper (triangles) and putative glycosylation sites (stars) are indicated on the bottom of the alignment. Red lines are used to connect cysteine residues that form disulphide bridges in the structural model. *acc*: accessibility level from DSSP (black=high and white=low).

Homology modeling of LRR domain and sequence to structure mapping

The LRR domain was predicted using MANIFOLD. It presents two terminal variable regions, LRR-NT and LRR-CT, reported to have high similarity to those in Nogo-66 receptor (NgR) [334] and four repeats between them. Recently, we presented a preliminary model of the LRR domain based on the NgR structure [7]. Modeling was conducted in two separated steps on the N- and C-termini, which were combined successively. Since the NgR protein has a longer LRR-CT and 8 repeats, the analysis of repeat periodicities with Repetita was performed to identify the correct number of LRR repeats in *LGI1*. The program predicts 4 motifs of 24 amino acids length and the template search selected the structure of the third LRR domain of *Drosophila melanogaster* SLIT (PDB code:1W8AA) [335] as the best template with a 32% sequence identity and the same number of repeats. In this way, the curvature of the LRR domain is more accurately modeled and the residues did not change in relative position as the new model is still based on the alignment from our previous work (Fig. 6.3) [312]. Comparison of conserved residues and secondary structures of hLGI1 and dSLIT revealed many correspondences in the alignment. The alignment was used to build the model, with only two gaps located in the LRR-NT and in the first LRR repeat which were modeled with LOBO. LGI family members and their orthologs differ exactly at these positions. This variability may indicate the presence of a specialized region for the specific LRR domain. Evaluation of model quality by QMEAN indicates that the regions of poor quality are located at the N- and C-terminal portion of the structure (Fig. 6.7). However, the N- and C-terminal caps of the LRR domain present two disulfide bonds (C42-C48 and C46-C55) at LRR-NT and two disulfide bonds (C177-C200 and C179-C221) at LRR-CT which confer stability to the structure. Furthermore, the whole model has good quality as indicated by a QMEAN score reflecting predicted model reliability of 0.6 (range 0,…,1; where 0 is worst and 1 best). As expected, the repeated

model core presents all hydrophobic residues forming the consensus sequence in the LRR domain internally buried and polar residues exposed to the solvent (Fig. 6.3).

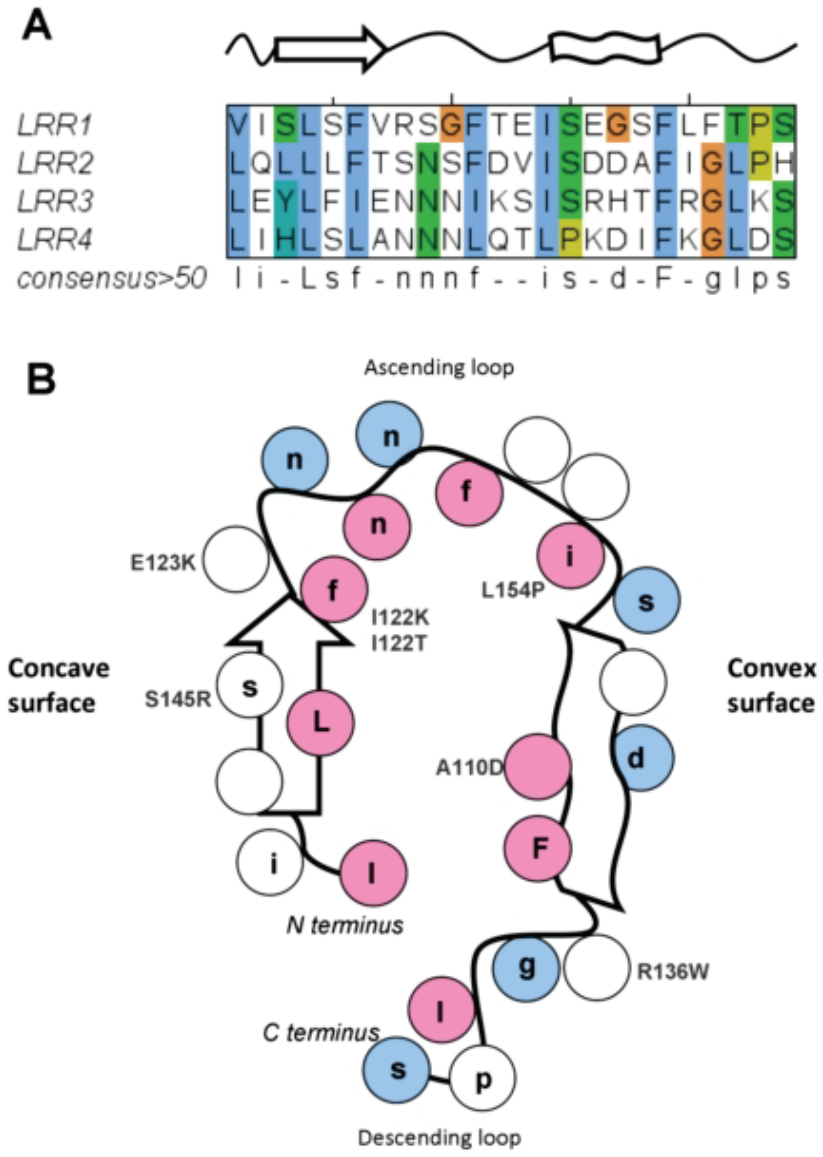


Figure 0.3. LRR repeat overview.
A. Consensus sequence repeat pattern of the LRR domain. Secondary structure is drawn on the top part of the alignment: an arrow represents the β-strand and a ribbon the α-helix connected by curved lines (loops). B. Schematic diagram of repetitive structural units in LGI1 protein. Conserved positions of the consensus pattern are reported on the diagram. Coloured pink spheres for buried residues and blue spheres for exposed residues.

The repeats stack in a parallel arc, allowing to partition the surface into four parts. The concave face, consisting of parallel β-strands, comprises a strong conserved region, while the convex face formed by a tandem arrangement of polyproline II plus β-turns

has only localized regions of conservation. We can also distinguish two other surfaces formed by two arrays of loops: the C-terminal side, which contains the loops linking the C-terminal end of the β-strands to the N-termini of the helices, and the C-terminal side, which forms a negative electrostatic surface (Fig. 6.3 and Fig. 6.4). Conserved negatively charged residues in LRR domains have been found involved in specific hydrogen bonds with NH groups of the backbone and considered important for structural integrity [95]. Other solvent exposed aspartic acid residues have been found to contribute to the twist of the overall LRR structure [336] as in the *Yersinia pestis* cytotoxin YopM [337]. In the LRR domain of LGI1 the negatively charged residues contributing to the negative electrostatic surface are all solvent exposed suggesting that they may be important for protein function.

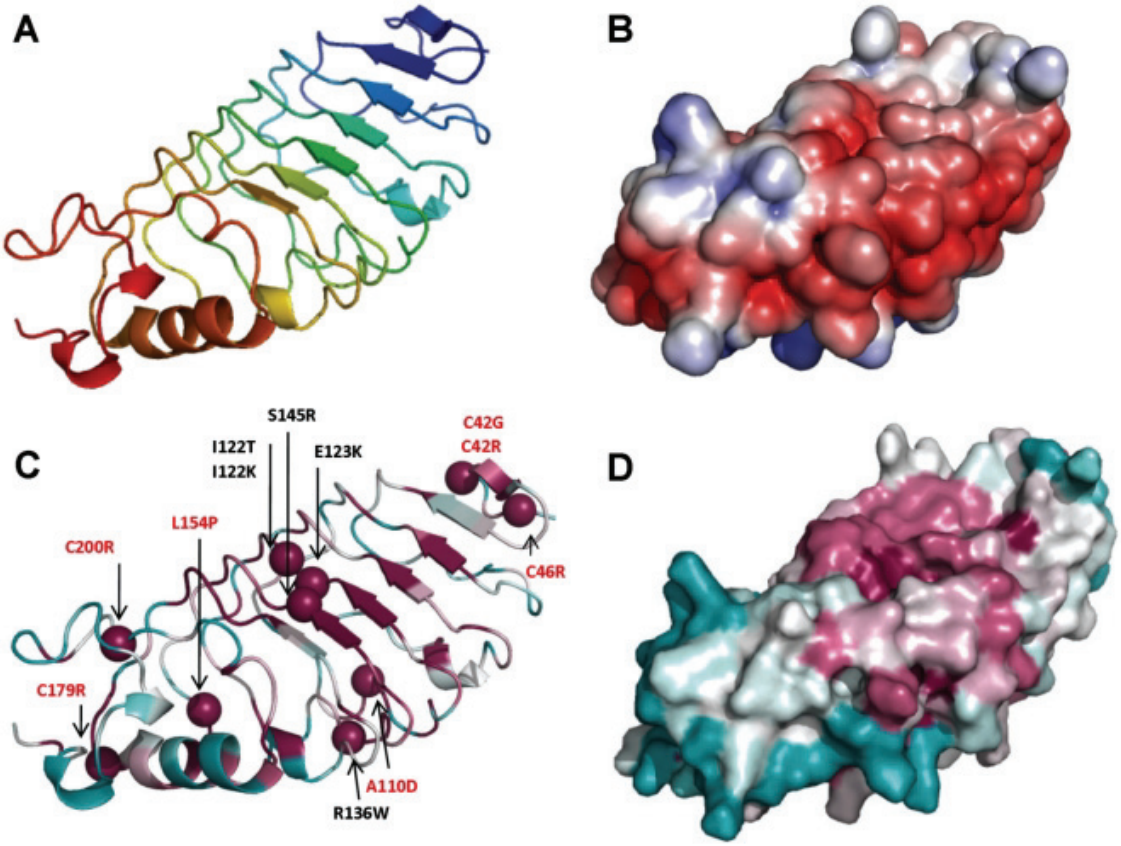


Figure 0.4. LRR model, structural analysis.
A. Cartoon of the LRR model coloured from N-terminal (blue) to C-terminal (red); **B**. Electrostatic surface (negative charge in red and positive charge in blue); **C**. Position of missense mutations, mutated residues are shown as spheres with structural mutations indicated in red; **D**. Conserved surface with ConSurf colour code from unconserved (cyan) to strictly conserved (magenta).

Homology modeling of EPTP domain and sequence to structure mapping

Staub and co-workers [314] proposed that the EPTP repeats could constitute a new class of β-sheet repeats, which fold into a β-propeller structure. The LGI1 β-propeller domain consists of 7 repeats, named EPTP1-7, each comprising a small four-stranded antiparallel β-sheet, whose strands are labeled A to D from N- to C-terminus. Repetita [97] was used to define the boundaries of repeats in the EPTP domain. We built a multiple alignment at the level of single repeats to define the EPTP repeat consensus sequence (Fig. 6.5). In order to classify LGI1 into a specific protein domain family, we searched for the presence of sequence motifs characteristic for different families of β-propellers [338]. The WD motif located at the end of β-strand C is conserved in repeats 1 and 6. In particular, the WD motif at the first repeat is conserved among all LGI proteins. In other blades, tryptophan and aspartatic acid are replaced by amino acids with similar biochemical properties (Fig. 6.2). We applied the Metaserver fold recognition method and selected the structure of human WD repeat protein 5 (WDR5) WD domain (PDB code: 2GNQA) as template, which presents a "velcro" closure and ca. 11% sequence identity. In many β-propellers each sequence repeat contains the first three strands of one blade and the last strand of the next. This is apparently also the case for LGI1. We manually curated the alignment between template and LGI1, keeping in consideration the secondary structure prediction. The gaps were closed with LOBO and fell almost all in loops that are longer in LGI1 than WDR5. Evaluation of the model quality, yielding a QMEAN score of 0.4, reveals that the most high quality regions comprise the core of the propeller formed by circular β-sheets, while the loops forming the bottom and top surface show poorer quality (Fig. 6.7). These regions differ more from the template due to the presence of several insertions/deletions. However, we can suppose that the overall model corresponds to the real structure of LGI1, since the protein core is stabilized by hydrophobic interactions. The modeled structure also presents a likely disulfide bridge between Cys260, in the first blade, and Cys286, in the second blade, which would confer further stability to the overall fold.

The LGI1 structural model has been evaluated for both conserved regions and electrostatic surface (Fig. 6.6). Using the alignment of different sequence families retrieved by BLAST, ConSurf does not reveal any particular conserved region. A conserved feature in all modular sheets from different propeller domains is a set of

positions with non-polar side chains, generally non solvent accessible, located in the central part of the strands. Since the major determinant for propeller assembly is the packing of these residues, amino acids in these positions are free to be replaced by other amino acids with similar biochemical properties [316]. Interestingly, using only sequences of different LGI family members to build the alignment, ConSurf identifies a highly conserved circular region in the top face of the β-propeller. On the bottom face of the protein there are also some conserved sites that correspond to the WD motif and electrostatic surface analysis identifies an extended positively charged region (Fig. 6.6). The top surface is formed by loops connecting strand D of one blade and strand A of the next (DA loops) and loops connecting strand B with strand C in the same blade (BC loops). The bottom surface is formed by loops connecting strand C and D of a blade (CD loops) and loops connecting strand A and B (AB loops) (Fig. 6.5). The alignment of WD repeat sequences allowed the identification of regions of variable length. In some proteins, one or more of these variable regions can be long enough to form an independently folded domain while other insertions form a reverse turn or loop that protrudes from the bottom of the propeller [339]. The LGI1 propeller has an insertion in the AB loop of the fourth repeat, not presents in paralogous LGI members, that protrudes from the bottom surface (Fig. 6.2 and 6.8). This loop may contain a functional motif that contributes to the functional specificity of LGI1.

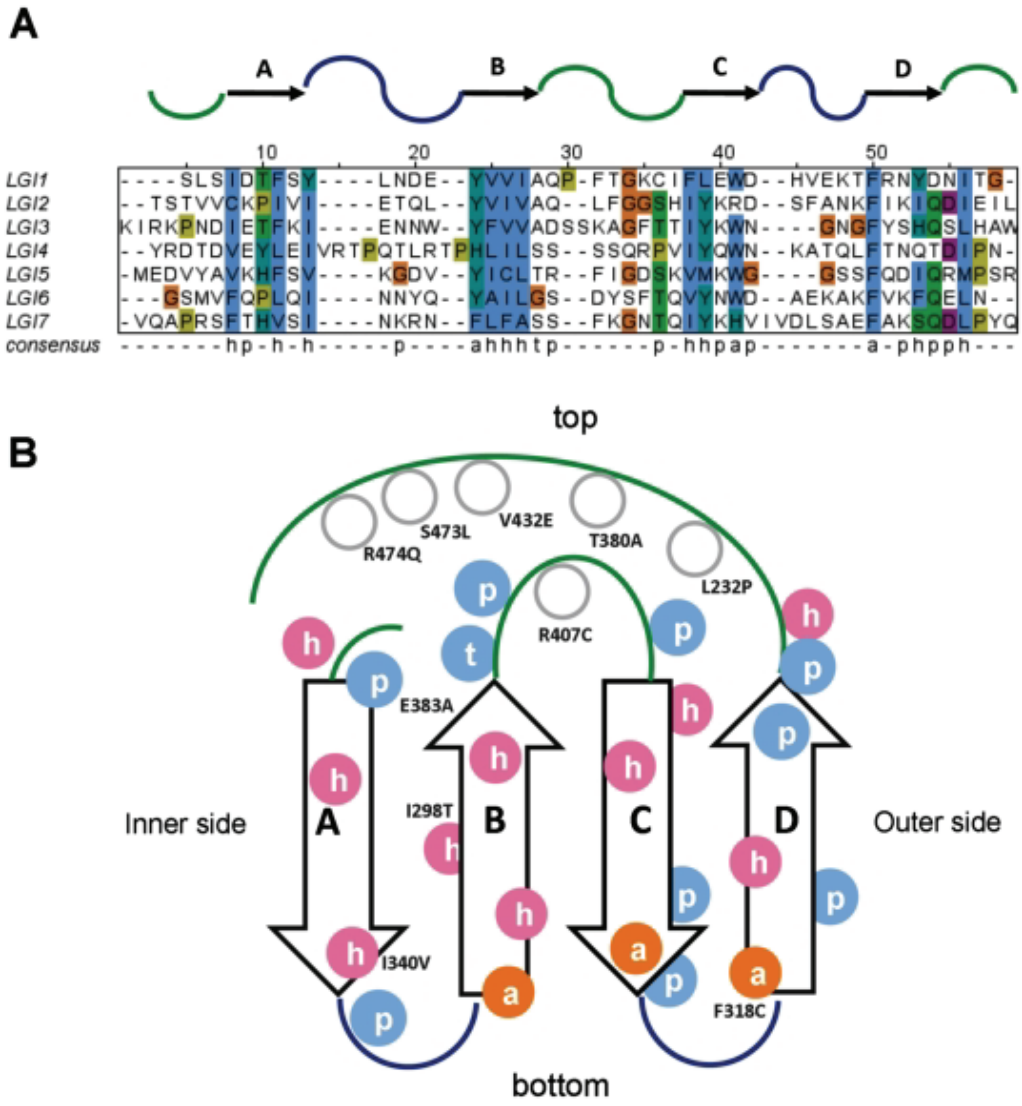


Figure 0.5. EPTP repeat overview.
A. Consensus sequence repeat pattern of EPTP domain. h = hydrophobic residue; p = polar; a = aromatic residue; t = tiny residue. Secondary structure is drawn on the top part of the alignment. Arrows represent β-strands connected by curved lines (loops). Loops forming the top surface are coloured in green, while those forming the bottom surface are coloured in blue. B. Schematic diagram of repetitive structural units in the LGI1 protein. Conserved positions of the consensus pattern are reported on the diagram. Pink and blue spheres indicate buried and exposed residues respectively.

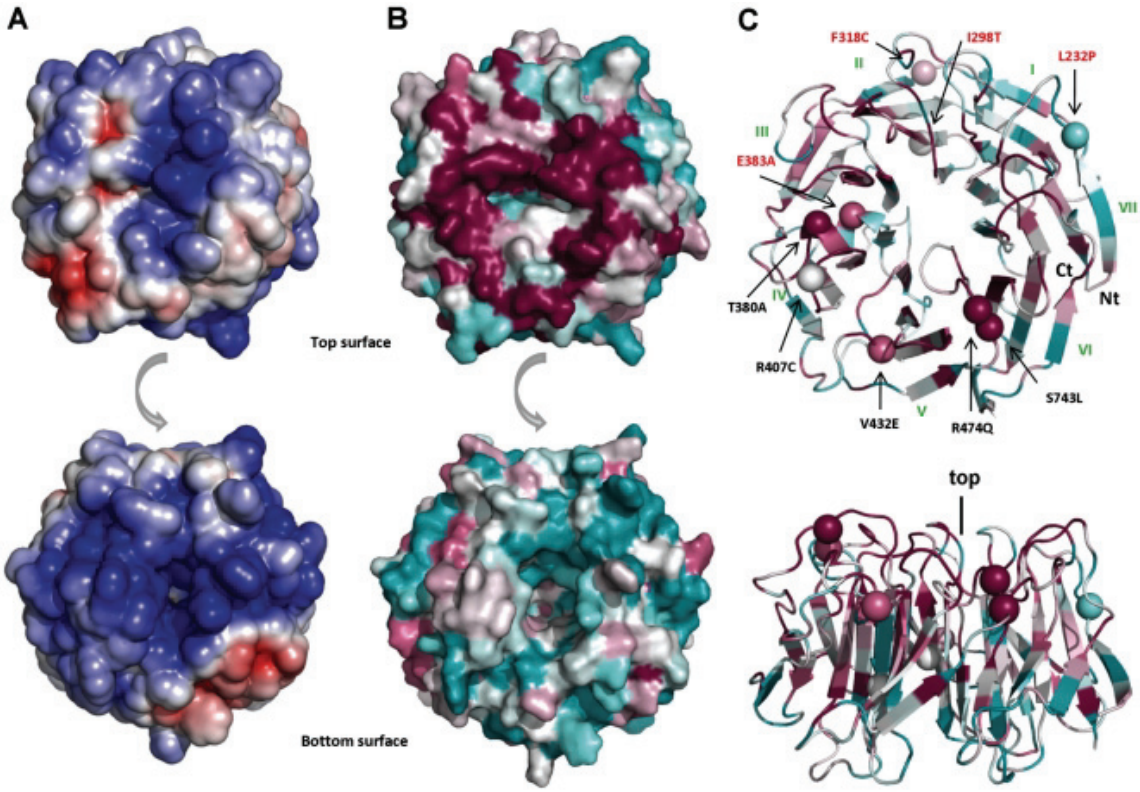


Figure 0.6. EPTP model, structural analysis.
A. Top (up) and bottom (down) view of electrostatic surface of EPTP model (negative charge in red and positive charge in blue); **B**. Top (up) and bottom (down) view of the conserved surface of EPTP model with ConSurf colouring from unconserved (cyan) to strictly conserved (magenta). **C**. Cartoon of the EPTP model in top and lateral view with ConSurf colouring. Spheres indicate residues found mutated in ADTLE patients with structural mutations indicated in red.

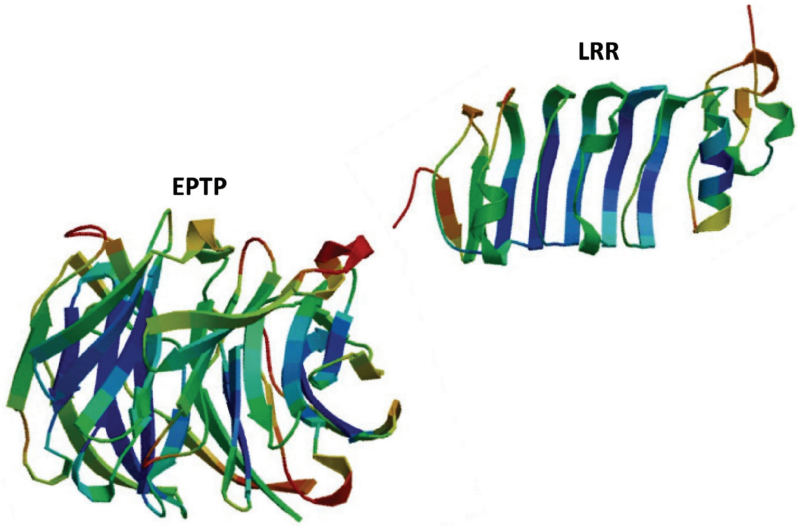


Figure 0.7. QMEAN model quality evaluation.
The estimated residue error is visualised using a colour gradient from blue (most reliable regions) to red (potentially unreliable regions, estimated error above 3.5 Å).

Interactions

LGI1 presents two domains that are known to form multi-protein complexes [316, 340]. It is reasonable to suppose that LGI1 mediates interactions between different proteins using different surfaces in the two domains. The first step is to understand how the two domains are arranged together. As they present two surfaces of opposite charge, it can be expected that an attraction between them exists. However, they are not positioned face to face due to the constraint imposed by the short loop connecting them. Instead, if we position the EPTP domain with the top face resting on a plane, the LRR moves laterally above the plane of the bottom surface exposing the conserved β-sheet (concave surface) (Figure 10A). Even if some LRR proteins use alternative surfaces for ligand binding, it is generally thought that the concave surface of the LRR structure contains the ligand-binding site [341]. LGI1 could interact with one protein through the concave LRR interface and with another protein through the top surface of the EPTP domain. It has been previously observed, that the β-propeller structure creates a stable platform that can form complexes reversibly with several proteins, using three potential interaction interfaces: top, bottom and circumference [339, 342].

The top surface appears to be a specialized region for LGI members because it is particularly conserved across them. The superimposition of LGI1 and the complex of WDR5 with its ligand (PDB code: 3EMH) allowed us to map the putative binding site of a ligand on the top surface of the EPTP domain (Fig. 6.8). LGI1 has been shown to bind through the β-propeller domain to both ADAM22, ADAM23 and ADAM11, although with different affinities [343]. On the other hand, *LGI4* is known to interact with ADAM22 [344]. Since the four members of the LGI family have a common phylogenetic origin (Fig. 6.1), it is reasonable to expect that interactions between various components of the LGI and ADAM protein families likely occur through the same, structurally conserved LGI binding site on the top EPTP surface (Fig. 10A).

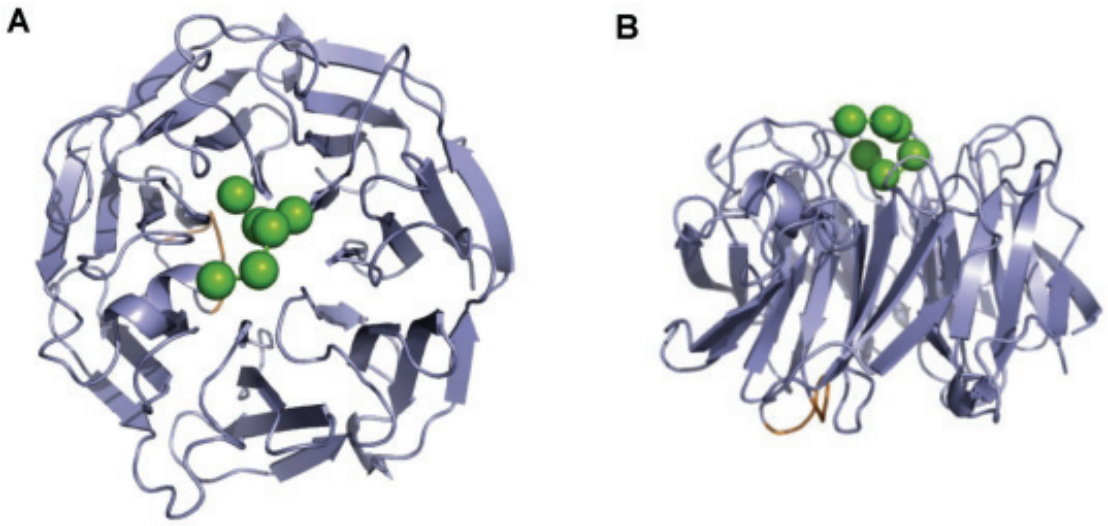


Figure 0.8. EPTP ligand bindind sites.
Top (**A**) and lateral (**B**) view of the hypothetical peptide binding site on the EPTP model. The position of a hypothetical peptide (green spheres) was obtained by superimposition of the EPTP model with the WDR5 structure (PDB code 3EMH). Note that the insertion specific for LGI1 (in yellow) maps on the bottom face of the domain.

Role of LGI1 N-Glycosylation

It is well known that the LRR and EPTP domains in LGI1 are N-glycosylated due to their extracellular localization and Sirerol-Piquer et al. [345] demonstrated that N192Q (LRR-CT, conserved across all LGI members), N277Q (conserved across some LGI1 and LGI2 orthologs) and N422Q (only conserved across mammalians) are sites of N-linked glycosylation in LGI1 (Fig. 6.2). Glycosylation could be essential for proper function of the protein since it can dramatically alter surface properties and thereby affect ligand binding. The effect of the potential N-glycosylation sites have been evaluated on the secretion of LGI1 [345]. Compared to a normal protein, the triple mutant was not secreted and secretion of the N192Q mutant was severely attenuated.

To understand the potential role of LGI1 glycosylation we analyzed their distribution over the domain surfaces. In our model, N192 on the LRR domain and N277 and N422 on the EPTP domain are all solvent exposed, confirming the overall correctness of the model. In the LRR domain, the glycosylation site maps to the N terminal side of the LRR-CT portion, while in the EPTP domain, the glycosylation sites map to the β-strand D of the first and fourth blades on the circumference surface. These findings indicate that, while glycosylation modulates the surface properties of LGI1, the putative ligand binding sites are located in non-glycosylated regions.

However, the glycosylation of N192 is supposed to have a mechanistic role. The presence of an oligosaccharide in this position indeed likely interferes with attraction of the charged surfaces present in the two domains, possibly preventing a too close interaction between them. From this point of view, N-linked glycosylation also appears important for correct protein folding.

In silico analysis of missense mutations

Recently, we have reviewed a total of 25 LGI1 mutations reported in the literature and analyzed their effects on secretion and on the structure using a preliminary model of the LRR domain [312]. Here we present the analysis of all 21 missense mutations found as to date in the LGI1 gene from subjects with familial or sporadic ADLTE, including the recently published p.R407C mutation [346], the two p.I122T and p.C179R mutations (submitted) and the unpublished p.T380A mutation. Twelve variants affect amino acid residues located in the LRR domain while nine are in the EPTP domain (Fig. 6.4 and 6.6). The analysis of structural and/or functional effects of these two variant groups has been conducted separately using our models of the LRR and EPTP domains (Table 6.1). Note that truncating mutations were excluded from our analysis, as no prediction is possible from the structure beyond noting probable protein misfolding.

LRR mutations

Among the twelve variants occurring in the LRR domain, one involves residues on the second LRR repeat, four on the third LRR repeat, two on the fourth LRR repeat and five involve residues at the N- and C-terminus. Some of the considered substitutions mapped at the terminal parts of the LRR domain are of particular interest since they modify conserved cysteine residues flanking the LRR repeats forming disulfide bonds (Fig. 6.2). Substitution of these residues inevitably causes a structural destabilization of the LRR domain. Even if using only protein sequence information, I-Mutant predicts Cys42 and Cys46 as stabilizing, but computational methods are not efficient in predicting protein stability changes due to loss of a disulfide bridge. All LRR variants are predicted to be destabilizing by at least three methods, meaning that all variants could have a negative structural change (Supplementary Table S.6.1). During initial analysis of LRR variants, we observed that it was possible to distinguish two groups of variants on the basis of

their effect on structure or function. The group of structural mutations includes critical mutations of the conserved cysteine residues (p.C42R, p.C42G, p.C46R, C179R and p.C200R), and four mutations of hydrophobic core residues to polar/charged residues (p.A110D, p.I122K, p.I122T, p.L154P). These mutations occur at conserved positions in the LRR repeat alignment having a structural role in folding the LRR domain (Fig. 6.3 and 6.4). The second group (p.E123K, p.R136W, p.S145R) alter residues located at the protein surface which have a potential to maintain the local structure, the details of which may be crucial for interactions with protein partners. Since all of these mutants lost the ability to be secreted, we hypothesize that a change on the surface, if not causing misfolding, should interfere with the secretion process, e.g. hampering attachment of the protein to the membrane. Evaluation of the electrostatic surface of these three mutants revealed that p.E123K and p.S145R affect the conserved concave surface formed by parallel β-strands of the LRR domain (Fig. 6.9). Variant p.R136W has subtle effects on the electrostatic potential of the convex surface, suggesting this could be another protein binding site.

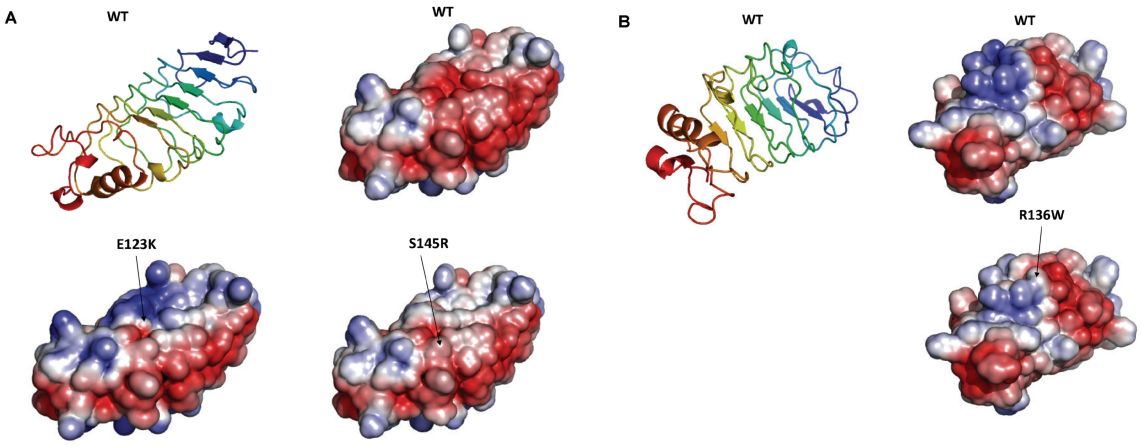


Figure 0.9. Electrostatic potential changes on the LRR surface.
A. Electrostatic potential changes induced by the E123K, S145R; B. Electrostatic potential changes induced by and R136W mutations. Note the different orientation of LRR domain.

Mutations		dbSNP	Position	Structural/functional effects	Secretion
p.C42R	(8)		LRR-NT	Precludes disulfide bridge formation with C48.	NT
p.C42G	(8)		LRR-NT	Precludes disulfide bridge formation with C48.	NT
p.C46R	(8)	rs104894166	LRR-NT	Precludes disulfide bridge formation with C55.	Negative
p.A110D	(8)		LRR2 Core	The mutation leads to three neighboring Asp with possible electrostatic repulsion.	Negative
p.I122K	(8)	rs119488100	LRR3 Core	Insertion of an charged aminoacid (Lys) alters the protein fold.	Negative
p.I122T	(8)		LRR3 Core	Polar residue inside the hydrophobic core. Possible alteration of the LRR domain fold.	NT
p.E123K	(8)		LRR3 Concave surface	The mutation alters the electrostatic surface of a potential peptide binding site on LRR domain.	NT
p.R136W	(5)	rs119488099	LRR4 Convex surface	Arg136 forms a salt bridge with Asp109. The substitution cause the loss of important interactions with neighboring amino acids, leaving tryptophan to protrude from the molecule.	Negative
p.S145R	(9)		LRR4 Concave surface	The mutation alters the electrostatic surface of a potential peptide binding site on LRR domain.	Negative
p.L154P	(6)		LRR4 Core	Having two neighboring proline poses a highly destructive condition.	NT
p.C179R	(9)		LRR-CT	Prevent the disulfide bridge with C241 causing a misfolding of LRR-CT domain	NT
p.C200R	(9)		LRR-CT	Prevent the disulfide bridge with C177 causing a misfolding of LRR-CT domain.	Negative
p.L232P	(2)	rs104894167	EPTP7 Loop D7-A1 ("Velcro")	Failure of "velcro" closure. Possible alteration of the protein fold.	Negative
p.I298T	(5)		EPTP2 βB2	Polar residue inside the hydrophobic core. Possible alteration of the propeller fold.	NT
p.F318C	(7)	rs28939075	EPTP2 βD2 Circumference surface	Position conserved across repeats. Possible alteration of the propeller fold.	Negative
p.T380A	(9)		EPTP4 Loop D3-A4 Top surface	Possible alteration of the functional interactions on the top surface of the propeller.	NT
p.E383A	(8)	rs28937874	EPTP4 βA4	Loss of contacts with neighboring sheets alter the correct fold of the domain.	Negative
p.R407C	(5)		EPTP4 Loop B4-C4 Top surface	Possible alteration of the functional interactions on the top surface of the propeller.	Secreted
p.V432E	(8)		EPTP5 Loop D4-A5 Top surface	The substitution lead to three negatively charged aminoacids. Possible alteration of the local structural integrity.	NT
p.S473L	(9)		EPTP5 Loop D5-A6 Top surface	Possible alteration of the functional interactions on the top surface of the propeller.	NT
p.R474Q	(9)		EPTP5 Loop D5-A6 Top surface	Possible alteration of the functional interactions on the top surface of the propeller.	NT

Table 0.1. Missense mutations overview for the LGI1 protein.
The table summarizes conservation degrees from ConSurf (in parenthesis, range 1-9), positions on the protein and predicted structural and functional effects of mutations found in ADTLE patients. For some of these mutants, the effect on protein secretion was previously investigated. For a recent review see [312].

EPTP mutations

Nine variants affect the EPTP domain and appear distributed through all repeats without any prevalence for a particular one. All mutations except one (p.S473L) were predicted to be destabilizing by at least two of the computational methods used (Supplementary Table S.6.1). We also distinguish between structural and functional mutations for the EPTP domain. Three mutations are classified as structural variants (p.I298T, p.F318C, p.E383A), as they affect conserved positions in the repeat alignment and map into the space between the two β-sheets of repeats 2 and 3 (Fig. 6.4 and 6.5). Indeed, residues forming the consensus sequence of propeller repeats are responsible for the hydrophobic contacts at the inter-sheet cores. It is the packing of these residues that is a major determinant for the assembly of the propeller fold [316]. The variant p.L232P located in the loop between repeats 1 and 7 also has a structural role as it forms part of the Velcro closure conferring stability to the propeller (Fig. 6.5).

Interestingly, other variants (p.T380A, p.R407C, p.V432E, p.S473L, p.R474Q) occur at residues located in the DA and BC loops that form the top surface of the β-propeller (Fig. 6.5 and 6.6). Mutations at the top surface have a potential to interfere with interactions occurring between the β-propeller and molecules such as the known LGI interacting ADAM proteins.

The mutation p.R407C has been found in three affected family members, two of whom had temporal epilepsy with psychic symptoms (déjà-vu, fear) but no auditory or aphasic phenomena, and the third had complex partial seizures without any aura [346] (Fig. 6.10). The pathogenicity of the mutation is supported by a) its cosegregation with epilepsy in the family, b) the evolutionary conservation of the Arg407 residue, c) a high Polyphen score (2.031), and d) its absence in healthy controls. Three of the stability prediction methods classify the mutants as destabilizing, while the Eris method was not able to calculate the energy change for this mutation.

In vitro studies have shown that the Lgi1 protein is secreted [313] and that all *LGI1* mutations tested so far inhibit protein secretion, [312] supporting a loss-of-function effect of mutations. The pR407C is the first mutation that does not prevent secretion of the mutant LGI1 protein (Fig. 6.10).

The structural impact of this mutation has been evaluated on the three dimensional model and indicated that substitution of Arg 407 with a cysteine could have no effect on EPTP domain folding (Fig. 6.11). Because the correct protein folding which is probably

necessary for secretion is preserved, the mutant protein can be secreted. On the other hand, the Cys407 residue is exposed on the top surface of the EPTP domain and, under the strongly oxidative conditions present in the extracellular environment, likely forms abnormal disulfide bridges with other molecules, ultimately hampering interaction of Lgi1 with its partner protein(s). The lack of effect of this mutation on protein secretion has been hypothesized to account for the atypical clinical features observed in this family but further confirmation are required.

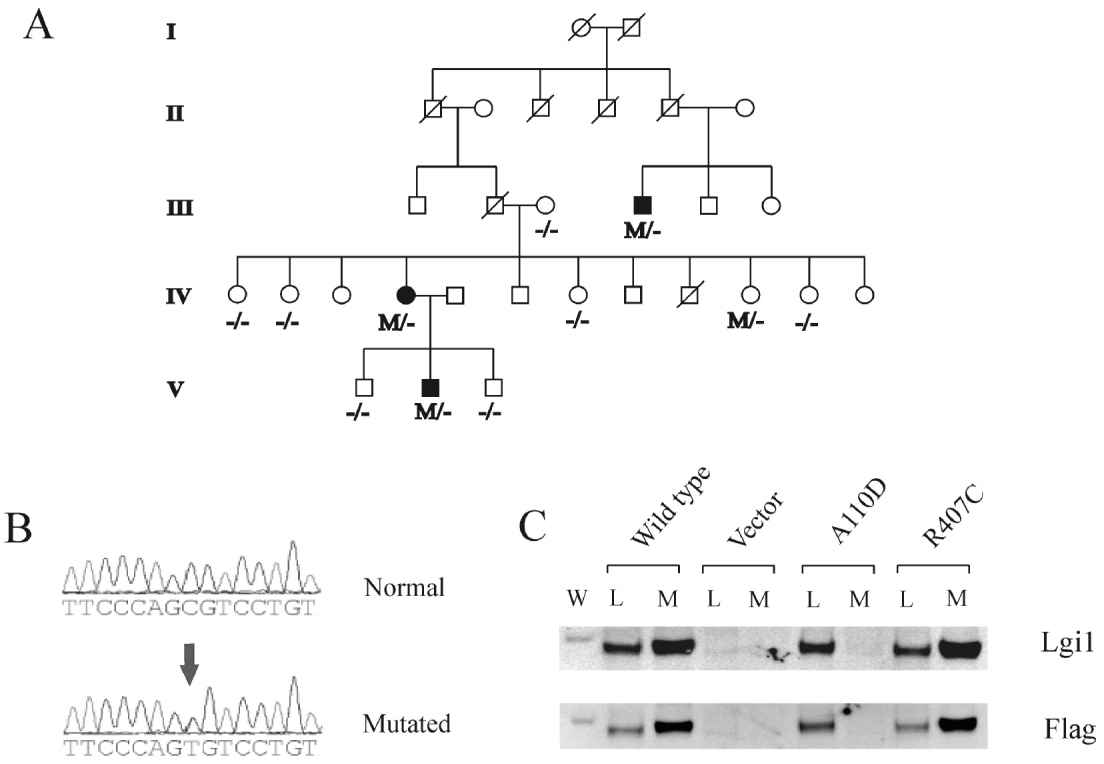


Figure 0.10. Family pedigree and mutation.
A. Pedigree of the family. Circles denote females; squares denote males; blackened symbols denote affected subjects. Individuals carrying one mutant and one normal allele are denoted by M/-, whereas those with no mutations by -/-. B. Original sequence tracings used to detect the disease allele (variant allele denoted by an arrow). C. Immunoblot analysis of transfected HEK293 cells. Cell lysates (L) and concentrated media (M) of HEK293 cells transfected with wild type or mutant [c.1219C>T (Arg407Cys) or c.329C>A (Ala110Asp)] LGI1 expression constructs containing a 3' Flag peptide sequence, or with empty expression vector (vector), were analyzed by western blot using either an anti-Lgi1 or an anti-Flag antibody. W, molecular mass marker.

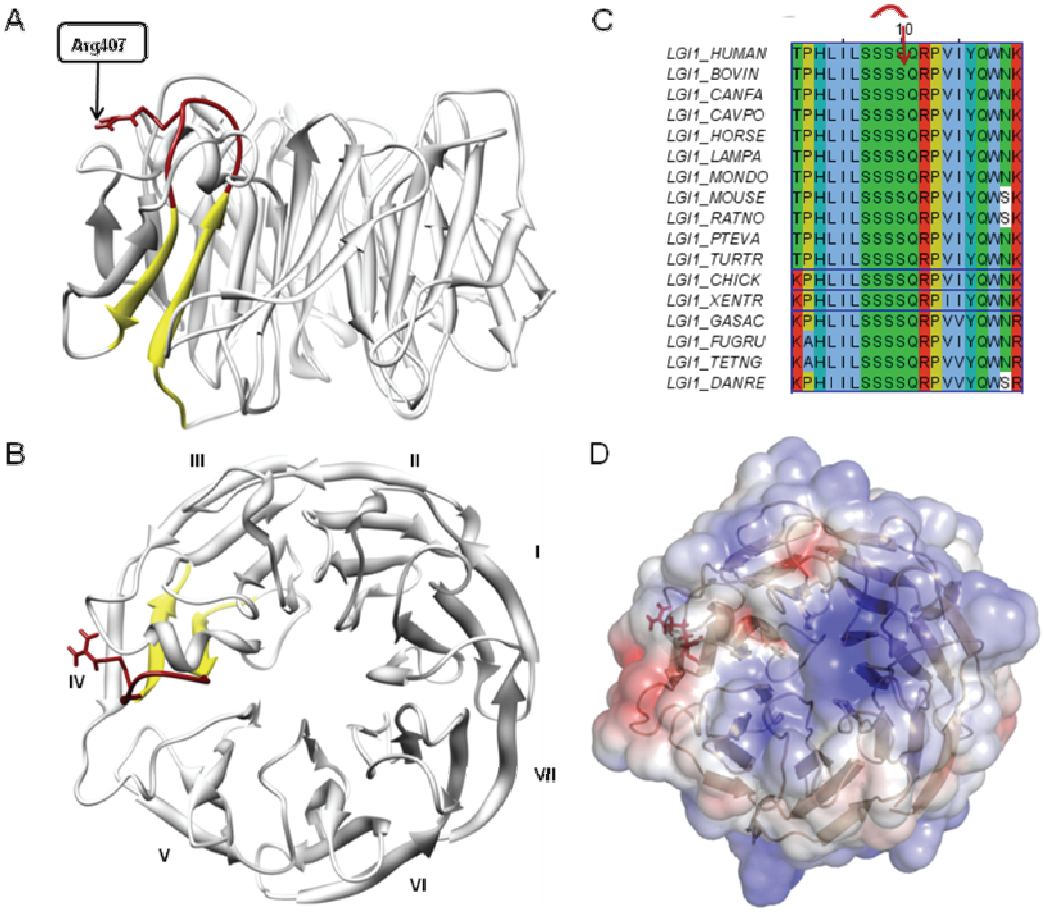


Figure 0.11. Three-dimensional model of the Lgi1 EPTP domain.
Top (A) and lateral (B) view of beta-propeller domain of the Lgi1 protein. The beta-strands and loop carrying Arg407 are coloured in yellow and red respectively. The Arginine maps on the top surface of the beta-propeller. (C) Multiple sequence alignment of the LGI1 orthologous sequences. The organisms are indicated using the OMA nomenclature. Different classes of organisms are grouped together. (D) Top view of the electrostatic surface of the Lgi1 beta-propeller domain. Arg407 is in a local negatively charged region.

Functional model

Although a single transmembrane domain was initially predicted in its central part [347], the LGI1 protein does not contain any transmembrane domains and is presumably secreted into the synaptic space [313]. Fukata et al. [321] have recently proposed a model that assigns LGI1 a role of trans-synaptic adaptor connecting the post-synaptic ADAM22 and the pre-synaptic membrane receptor ADAM23. However, since binding of LGI1 with the ADAM proteins is mediated by the EPTP domain and this interaction likely occurs only through the conserved EPTP bottom surface [167] (and see above), it is unlikely that LGI1 is capable of interactions with two ADAM proteins simultaneously. Thus, rather than forming a stable link between two ADAM receptors

across the synaptic cleft, LGI1 may represent a dynamic link which transports a signal from the pre- to the post-synaptic membrane. In this scenario, binding of a partner protein with the LRR domain removes the EPTP domain from its stable interaction with one ADAM protein and allows the movement of LGI1 to the opposite side of the synapse (Fig. 6.12).

However, it has also been suggested that LGI1 is secreted as an oligomer [318]. Therefore another possible scenario is that LGI1 could form a dimer, in which the LRR domains of two subunits interact by their concave surfaces connecting two ADAM proteins at opposite sides of the synapse (Fig. 6.12). This supports the experimental findings that demonstrated LGI1 connecting the pre- and postsynaptic machinery through ADAM22 and ADAM23 [321].

The hypothesis concerning LGI1 can also be reasonably extended to other LGI family members. As supported by our phylogenetic analysis and conserved surface residues, binding of ADAM family proteins by LGI is probably a conserved feature. The main difference between LGI1 and other family members appears to be the precise arrangement between the LRR and EPTP domains, as suggested by the presence of a unique insertion on the bottom surface of EPTP in the LGI1 sequences. The effect of this insertion may be a reduced binding affinity for the LRR domain and thus an increased propensity for interaction with other proteins and/or LGI homodimerization in LGI1. This adaptation could contribute to explain the unique tissue distribution of LGI1 compared to other family members [332].

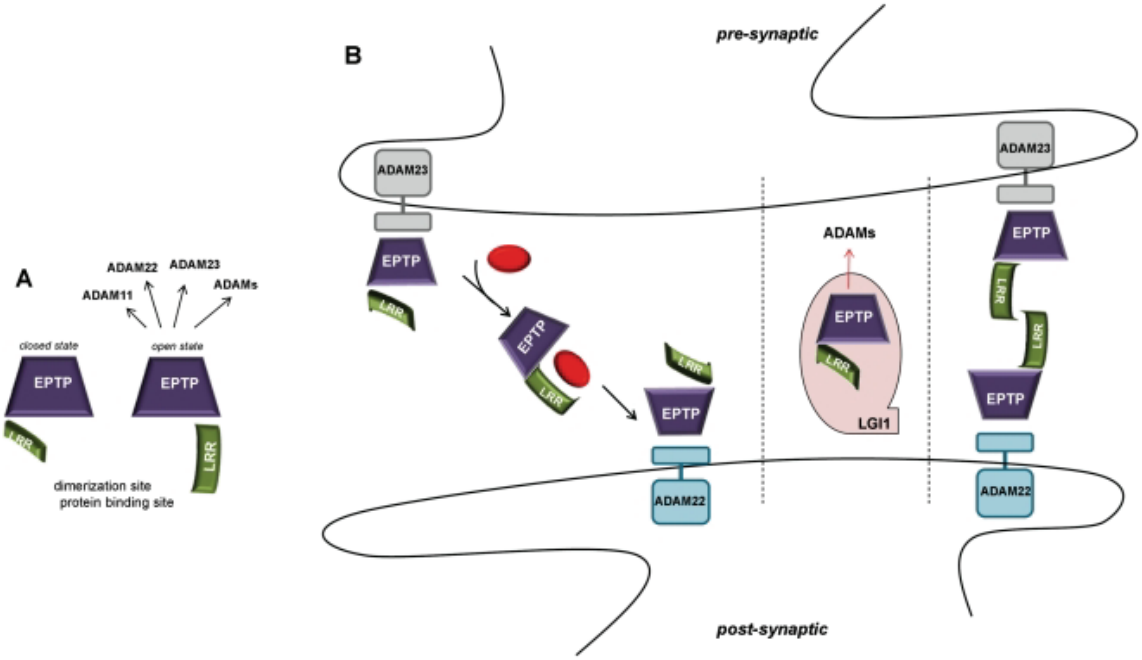


Figure 0.12. Hypothetical structural assembly and interactions.
A. LGI1 is represented as the association of LRR (green arc) and EPTP (violet trapezoid) domains. LGI1 interactions with ADAM proteins likely occur on the top surface of the EPTP domain. B. The two hypothetical ways by which LGI1 could mediate the trans-synaptic interaction between presynaptic ADAM23 and postsynaptic ADAM22.

6.5. Conclusions

An important task of this study was to uncover the relationship between amino acid sequence, 3D structure and putative functions of the LGI1 protein. Evolutionary sequence analysis revealed the presence of peculiar sequence stretches for each LGI protein, e.g. LGI1 contains a unique insertion on the fourth blade facing the bottom surface of the propeller. Using a structure-based sequence profile we identified a pattern among the structural units and obtained the models which validated several underlying assumptions, including the inward orientation of conserved non-polar residues and solvent exposure of N-glycosylated residues.

The three-dimensional model of LGI1 domains showed how the N- and C-terminal regions are intimately related, revealing a possible mechanism by which LGI1 mediates the trans-synaptic interactions between ADAM proteins. The LGI1 protein contains two conserved binding sites at the concave face of the LRR domain and a circular region on the top surface of the propeller domain.

We also evaluated the effect of missense mutations found in ADTLE patients on LGI1 protein and we are able to distinguish between structural and functional mutations, the former potentially causing protein unfolding, while the latter interfere with partner protein interactions. Previously published experiments demonstrated that all but one (p.R407C) tested mutants have a defect on secretion [7] (Striano *et al*., in press). Thus, we could hypothesize that the secretion-defective mutant proteins are either incorrectly folded or have altered electrostatic surfaces, which could affects LGI1 export. This explains why many LGI1 variants could not be secreted and opens a question about the mechanisms involved in the molecular pathogenesis of the disease. On the other hand, the p.R407C mutation is compatible with secretion, but rather may exert its pathogenic effect by disrupting interactions with ADAM proteins. Other functional mutations may have the same extracellular effect.

Experimental knowledge suggests interactions between LGI1 and ADAM proteins to be mediated by the EPTP domain. We showed that these interactions likely occur through the EPTP top surface. Furthermore, based on the assumption that two protein families usually interact in a similar way, with the same binding site, we predict all four LGI family members to use this interface to interact with different ADAM proteins, albeit with different affinity, in a time and space dependent manner. Finally, we suggest two alternative molecular mechanisms by which LGI1 connects ADAM receptors across the synaptic cleft.

6.6. Outlook

Among mutations of the EPTP domain, four (p.T380A, p.V432E, p.S473L, p.R474Q) mapped on the top surface of the β-propeller, as seen for the p.R407C mutation. Only two of these are predicted to be stabilizing by at least one stability prediction method. However, I hypothesized that all of these could have structural effects similar to the p.R407C mutation. Interestingly, the *in vitro* studies of their impact on protein secretion carried out by our cooperation partner confirm that all these mutant proteins are secreted (unpublished data). Furthermore, all individuals carrying these secreted mutants presented a clinical phenotype overlapping with those carrying the p.R407C mutation,

characterized by temporal epilepsy with psychic symptoms. These observations expand the phenotype of *LGI1*-related epilepsy and suggest that *LGI1* mutations should be also searched for in familial temporal epilepsies without auditory symptoms. In order to better understand the molecular mechanism underlying this pathological condition these mutants are being investigated for their ability to reach the cellular membrane and bind ADAM proteins at the extracellular site.

7. Dilated cardiomyopathy in patients with *POMT1*-related congenital and limb-girdle muscular dystrophy

This chapter has been submitted for publication at the European Journal of Human Genetics and it is currently under review: Bello L., Melacini P., Pezzani R., D'Amico A., Piva L., Leonardi E., Soraru' G., Palmieri A., Smaniotto G., Gavassini B., Vianello A., Bertini E., Angelini C., Tosatto S., Torella A., Nigro V. Cardiomyopathy in patients with POMT1-related congenital and limb-girdle muscular dystrophy. European Journal of Human Genetics submitted.

7.1. Summary

Protein-O-mannosyl transferase 1 (*POMT1*) is a glycosyltransferase involved in α-dystroglycan (α-DG) glycosylation. Clinical phenotype in patients harbouring *POMT1* gene mutations ranges from congenital muscular dystrophy (CMD) with structural brain abnormalities, to limb girdle muscular dystrophy (LGMD) with microcephaly and mental retardation and to mild LGMD.

We report three patients who harboured compound heterozygous *POMT1* mutations and developed dilated cardiomyopathy. Two patients had an LGMD phenotype with a normal or close-to-normal cognitive profile, while one had CMD with mental retardation; all patients had normal brain MRI.

Bioinformatics methods were used to study the potential effect of detected aminoacidic substitutions, 2 of which are caused by novel missense mutations. All of the detected mutations are predicted *in silico* to interfere with protein folding and/or catalytic function.

These patients widen the clinical spectrum associated with *POMT1* gene mutations, emphasizing the relevance of a careful follow-up of cardiac function in patients with α-DG glycosylation defects, regardless of the severity of neuromuscular involvement.

7.2. Introduction

POMT1 (protein-O-mannosyl-transferase 1), together with its homologue POMT2, is part of a heteromeric complex involved in the initiation of O-mannosyl glycan synthesis in the endoplasmic reticulum [348-349]. The complex catalyses the first step in the attachment of O-mannose-linked glycan moieties to α-dystroglycan (α-DG) [350]. α-dystroglycan has a relevant, structural role in muscle fiber integrity, connecting the dystrophin-glycoprotein complex to the extracellular matrix [351]. Mutations in *POMT1* result in a reduction of α-dystroglycan glycosylation in skeletal muscle of affected patients [352-354]. The clinical phenotype of *POMT1* (*Protein-O-mannosyl transferase 1*) mutations ranges from severe Walker-Warburg Syndrome (WWS) [352](5), to milder forms of congenital muscular dystrophy (CMD) with microcephaly and mental retardation without eye abnormalities (CMD-MR) [353], and to limb-girdle muscular dystrophy with normal brain structure and different degrees of mental retardation (LGMD2K)[354-355].

Five other genes involved in α-DG glycosylation are known. Mutations in these genes: *protein O-mannosyl transferase 2* (*POMT2*), *protein O-mannose β-1, 2-N-acetylglucosaminyltransferase* (*POMGnT1*), *fukutin (FKTN)*, *fukutin-related protein* (*FKRP*), and *like-glycosyltransferase (LARGE)* lead to heterogeneous phenotypes resulting from the combination of muscular dystrophy, brain and eye involvement [356]. Notably, cardiac involvement has so far been reported only in patients with *FKRP* [357-359] and *FKTN* [360] gene mutations.

We report three patients from three unrelated families, with different neuromuscular phenotypes, who presented dilated cardiomyopathy in association with compound heterozygous *POMT1* mutations.

7.3. Materials and Methods

Patients

We screened muscle biopsies of 247 patients affected by LGMD, CMD, muscle weakness or CK elevation of unknown cause for α-DG glycosylation defect by immunohistochemistry. Dystrophin, α-sarcoglycan, calpain and dysferlin were normal

by immunohistochemistry and/or immunoblotting in patients' biopsy. A mild to complete reduction of immunolabelling was found in 107 patients, who were subsequently screened for mutations in glycosyltransferase genes. POMT1 mutations were found in 9 patients, distributed by phenotype as follows: 3 LGMD, 4 CMD with mental retardation and normal brain MRI, 2 WWS (Table 7.1). All these patients routinely undergo a periodic screening for cardiological abnormalities by EKG end echocardiography; three patients showing signs of cardiomyopathy were selected for the present study.

Patient	Disease onset	Current Age	Phenotype	Mentation/Brain MRI	Cardio-myopathy onset	*POMT1* mutations	
						Nucleotide change	Amino acid change
#1	Birth	14	CMD-MR	MR/normal, microcephaly	14 yrs	c.2005G>a c.1241+1G>A	Ala699Thr p.His384_Thr414del
#2	3 yrs	20	LGMD-MR	Slight MR/, normal	12 yrs	c.430A>G c.1241C>T	Asn144Asp Thr414Met
#3	33 yrs	34	LGMD-NOMR	Normal/normal	34 yrs	c.1864C>T ?	Arg622Stop ?

Table 7.1. Clinical and molecular features of patients.
POMT1: protein-O-mannosyltransferase 1; LGMD-MR: limb-girdle muscular dystrophy with mental retardation; LGMD-NOMR: limb-girdle muscular dystrophy with no mental retardation; CMD-MR: congenital muscular dystrophy with mental retardation; MR: mental retardation; MRI: magnetic resonance imaging

Patient # 1 is a 17-year-old boy described in a previous report [361]. Hypotonic at birth, the patient acquired stable head control at 8 months and the ability to sit unsupported at 15 months, but never learned to walk. He had severe mental retardation and autistic features. A brain MRI carried out at 6 years of age was normal. On neurological examination, diffuse muscle wasting, muscle weakness, mild calf hypertrophy, severe scoliosis with rigid spine and microcephaly were present. Tendon reflexes were normal. Serum creatine kinase (CK) was 6000 U/L. A muscle biopsy, taken when the patient was 12 years old, revealed dystrophic features and reduced immunolabelling of α-DG and dystrophin. Dystrophin gene analysis did not identify any mutations and the observed slight decreased of dystrophin immunostaining was probably secondary to nonspecific proteolysis.

At the age of 16, the patient was admitted to the hospital for respiratory distress. Nocturnal non-invasive ventilation was begun and a gastrostomy was carried out because of severe swallowing disturbances. An electrocardiogram (EKG) and an echocardiography carried out at that time were normal. One year later the patient developed acute respiratory distress, prompting a complete cardiac evaluation. An echocardiography showed a moderate left ventricular dysfunction (left ventricular end diastolic volume index [LVEDVi] 50 ml/m2: n.v. < 70ml/m2 ; left ventricular ejection fraction [LVEF] 40%: n.v. ≥ 50%) but a poor acoustic window due to scoliosis did not permit assessment of right ventricular (RV) function. Diuretic therapy was begun and cardiac ultrasound performed six months later demonstrated stable parameters (LVEDV index 56 ml/ m2, LVEF 44%).

Patient #2 is a 20 year-old man who showed normal psychomotor development, who had come to medical attention at the age of 3 because of the occasional finding of elevated CK levels (<10,000 U/L). At the age of 5 years a muscle biopsy showed mild myopathic alterations and perimysial fibrosis. Immunohistochemical analysis of dystrophin, α-, β-, and γ-sarcoglycan and β-dystroglycan was normal. Dystrophin gene analysis did not identify any mutations. At the age of 12 years the patient, until then asymptomatic, underwent a routine echocardiography which documented a diffuse left ventricular (LV) wall hypokinesia with normal LVEDVi (69 ml/m2) and LVEF (50%). He presented at the age of 17 years with shortness of breath, cough, easy fatigability and abdominal pain. An electrocardiogram (ECG) showed LV hypertrophy (voltage criteria Sokolow-Lyon index = 38mm, n.v. ≤ 35mm) and an echocardiography showed a moderate LV dilation (LVEDVi 81 ml/m2) with oderate-severe systolic dysfunction (LVEF 36%) as well as moderate RV dilation (RVEDVi 88 ml/m2, n.v. ≤ 60 ml/m2). The patient responded to β-blockers and angiotensin receptor 1 blocker (sartanics) therapy. A cardiac echo carried out when the patient was 20 years of age showed a LVEDVi of 92 ml/m2 with an EF of 47% and mild hypokinesia of LV walls. The RV was moderately dilated (RVEDi 98 ml/m2) and kinesis was normal (RVEF 70%). Conventional spirometry showed mild obstruction and a normal forced vital capacity (FVC).

Currently, the patient has no difficult rising from the floor or climbing stairs. A neurological examination showed calf and thigh hypertrophy, relative wasting of the

scapulohumeral girdle, and a mild symmetrical weakness of proximal muscles. A brain MRI was normal, but neuropsychological evaluation showed executive dysfunctions (categorization ability, set-shifting, and planning) and significant visuo-spatial learning impairment. The patient's IQ was in the normal range (82).

Patient # 3 is a 34 year-old man who was well until the age of 33 years, when he began to complain muscle weakness in the lower limbs and myalgias in the shoulder girdle. Serum CK was 981 U/L and a muscle biopsy was consistent with a severe myopathy with type I fiber predominance (90%) and central nuclei and cores in the majority of fibers. A neurological examination revealed calf hypertrophy and moderate weakness of bilateral triceps brachii. He had no difficulty rising from the floor, walking long distances or climbing stairs.

A diagnosis of an initial biventricular dilatation was made when the patient was 34 old on the basis of a cardiac echo which showed a LVEDVi of 78 ml/m2, and a LVEF of 67%; the RV was moderately dilated , the ejection fraction was normal (RVEDi 74 ml/m2, RVEF 59%) as were the kinesis indexes. A conventional spirometry was normal.

α-dystroglycan glycosylation and laminin α2 studies

α-dystroglycan glycosylation was studied on 8 μm thick cryosections of frozen muscle tissue, using an antibody directed against an *O*-glycosylated epitope of α-dystroglycan (IIH-6; Upstate Biotechnology, Lake Placid, NY); laminin α2 was studied using an antibody directed against the carboxyl-terminus of the protein (mAb 1922, 80 kDa, Chemicon, Temecula, CA) (1:1,000).

Gene mutation studies

DNA was extracted from peripheral blood. The complete coding regions, including intron/exon boundaries of *FKRP*, *POMT1*, *POMT2*, *POMGnT1*, *FKTN*, and *LARGE* were screened for mutations either by PCR/SSCP (Single Stranded Conformation Polymorphism)/sequencing or direct sequencing (primers available upon request). Restriction fragment length polymorphism analysis was used to confirm gene variants, to verify segregation in the family and to assess frequency on 110 control chromosomes.

Bioinformatics

An integrative bioinformatics approach was used with the aim of elucidating the sequence-structure-function relationship of POMT1. The human *POMT1* sequence was downloaded from UniProt [173] with accession number Q9Y6A1. PSI-BLAST [37] was used with standard parameters for a single iteration on the UniProt sequence database to search for homologous sequences. InterPro [362] and ELM [52] were used to search for known domains and interacting motifs respectively. The secondary structure was analyzed with the consensus method [99], while disordered regions were searched with SPRITZ [111] and transmembrane helices predicted with TOPCONS [69]. The structure of the MIR domains (found in Mannosyltransferases, Inositol triphosphate receptors and Ryanodin receptors) was modeled with HOMER (URL: http://protein.bio.unipd.it/homer/) from the template structure with PDB code 1T9F previously identified with PSI-BLAST, with loops positioned using a fast divide and conquer approach [88] and the final model being evaluated with FRST [89]. The structure was visualized using PyMol (DeLano Scientific, URL:http://www.pymol.org/). The I-Mutant [190], Mupro [187] and SNPs3D [183] servers were used to estimate effects of the mutations in terms of protein stability. Other two predictor, SNAP [276] and PhD-SNP [181], were used to classify variants as disease-related or as neutral polymorphisms.

7.4. Results

α-dystroglycan glycosylation and laminin α2 studies

Immunofluorescence analysis of muscle biopsies revealed severe reduction of α-DG glycosylation in patient #1 and 2, and a moderate reduction in patient #3 with respect to control (Fig. 7.1). Few α-dystroglycan negative fibers were observed in patients' #1 and #3. Laminin α2 expression was slightly reduced in the patient muscle biopsies compared to control.

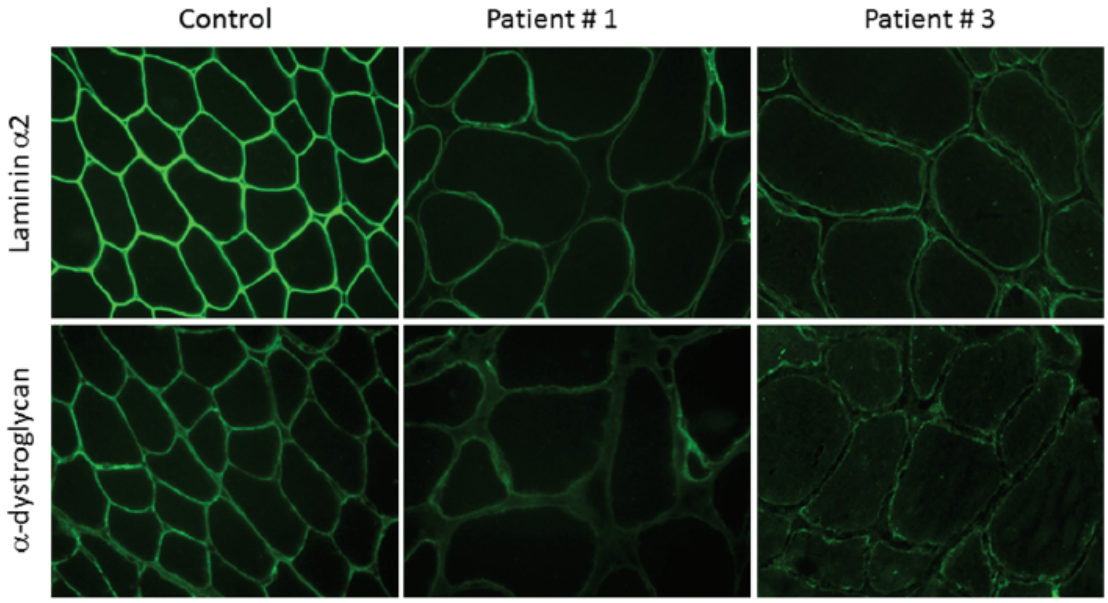


Figure 7.1. Reduced α-dystroglycan glycosylation in POMT1 mutated patients.
α-dystroglycan immunostaining using an antibody directed against a glycosylated epitope shows a normal labeling at the periphery of each fiber in the control's muscle, in comparison with patient #1 and #3 where the majority of myofibers show a faint immunoreaction and variability of the intensity of the labeling. Laminin α2 immunostaining using an antibody directed against the 80 kDa carboxyl-terminus shows a subtle reduction of the labeling in the patients.

Gene mutation studies

Patient #1 was compound heterozygous for two *POMT1* mutations: a missense mutation, c.2005G>A, p.Ala669Thr and a donor splice site mutation in intron 12, POMT1 c.1241+1G>A [361]. cDNA analysis showed that the c.1241+1G>A14 results in the in-frame skipping of exon 12 (p.His384_Thr414del) (data not shown). Two novel *POMT1* missense mutations were identified in patient #2: c.430A>G, p.Asn144Asp and c.1241C>T, p.Thr414Met.

Patient #3 was found to be heterozygous for the nonsense mutation c.1864C>T, predicting a premature stop codon (p.Arg622X). In addition, sequencing of patient's cDNA identified a splice defect that incorporates 5 bases at the junction exon10-exon 11 r.1052_1053insGTAAG. Full sequencing of genomic DNA identified a number of variations in intron 10-exon 11 c.1052+49 g>a (Hom), c.1052+184 g>a (Hom), c.1052+246 g>a (Hom), c.1052+276 t>c (Hom), c.1053-172 c>t (Hom), c.1053-113 c>g (Hom), c.1053-102 g>a (Het), and c.1113 T>C D371D (Hom). All these have unknown significance and none predicted a cryptic splice site compatible with the aberrant

transcript observed. The likely scenario is a leaking splicing defect leading to two different transcripts: one alternative transcript resulting in an out-of-frame insertion of 5 base pairs and a normally spliced transcript consistent with the production of a normal, but reduced protein product, and thereby consistent with partial α-dystroglycan glycosylation defect.

No *FKRP, POMT2*, *POMGnT1*, *FKTN*, and *LARGE* mutations were detected in any of the patients. Identified mutations were not detected in 110 control chromosomes.

In silico *prediction of mutation effects*

The sequence of human *POMT1* (NG_008896) was analyzed with a number of bioinformatics methods in order to characterize the mutation sites. As expected, several transmembrane helices were predicted and the known MIR domains detected. A consensus approach was used to delimit the single transmembrane helices, as different methods provided slightly different predictions, especially for the second and last helices. The structure of the MIR domain was predicted by homology modeling from a template structure with 31.4% sequence identity. Secondary structure and disorder predictions were used in combination with ELM to identify locations of possible functional motifs. Figure 7.2 summarizes the analysis of the POMT1 sequence and the positions of disease-associated mutations.

The mutations were analyzed with several prediction methods to determine possible pathogenicity and compared to known mutations with experimentally measured enzyme activity [349, 363-364]. All substitutions occur at conserved positions (ConSeq score of 7-9), except for G76R which presents a medium score of 5. However, at this position charged residues like Arginine are never present in homologous sequences. Furthermore, all amino acid substitutions are predicted to be destabilizing or pathogenic by most of the used prediction methods (Table 7.2). Two of the identified missense mutations (N144D and A669T) are located in transmembrane helices and have similar predicted effects as previously identified mutations. The two substitutions introduce respectively a negative charged and a polar residue that seem to have a destabilizing effect on protein folding (Table 7.2). The T414M mutation is part of the modeled MIR domain (Fig. 7.2). As can be seen, it is in close proximity to the V428D mutation causing WWS [351], and is likely to destabilize the protein with a similar mechanism.

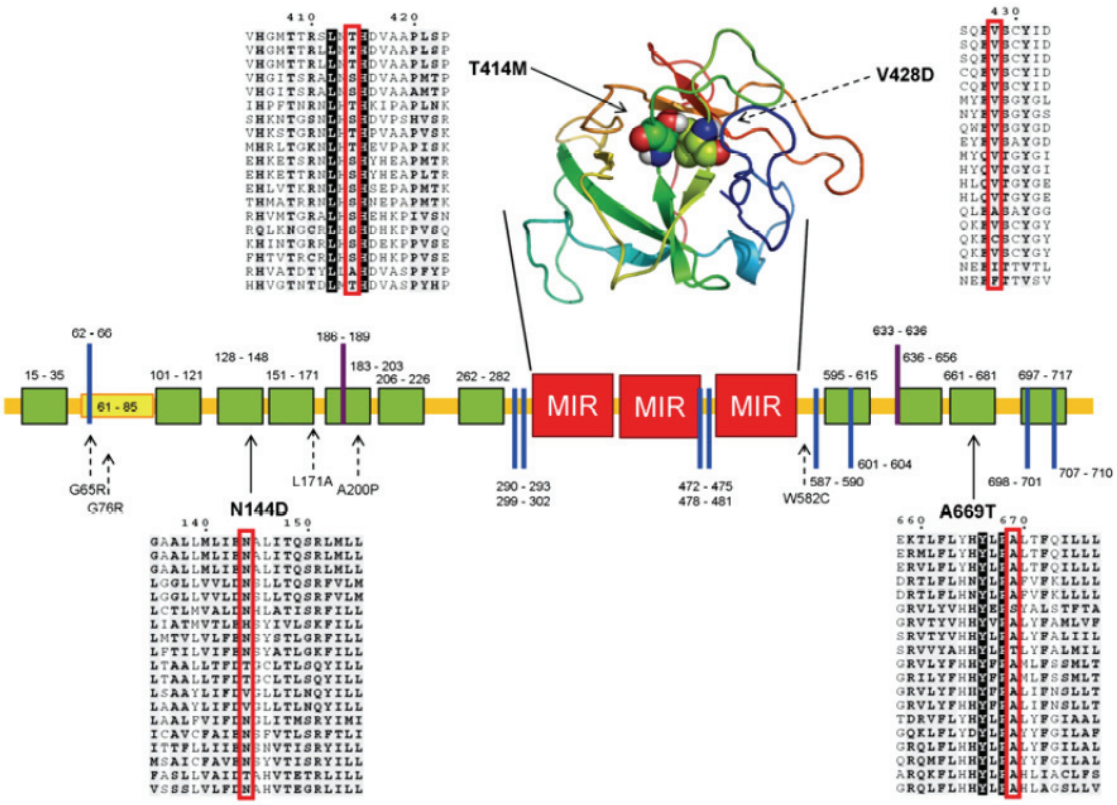


Figure 7.2. Schematic overview of POMT1.
The sequence of human POMT1 is shown as a horizontal line, with transmembrane helices (green), disordered regions (yellow) and MIR domains (red) shown as boxes. The modeled structure of the three MIR domains is shown above the sequence, coloured from N- to C-terminus in blue to red. C-Mannosylation and Glycosaminoglycan attachment sites predicted by ELM are shown as purple and blue bars respectively. Disease-associated mutations are shown with arrows pointing to the relevant position in the sequence together with the local sequence context from a multiple sequence alignment. Previously known mutations are shown with dotted lines. Both mutations falling into the MIR domain are shown with their residues as spheres in the structure.

Mutation	Phenotype	Enzymatic activity	I-Mutant (ΔΔG)	SNPs3D (SVM score)	MUpro (C score)	SNAP (RI, EA)	PhD-SNP (RI)	ConSeq (Conservation level)	AAs
N144D*	LGMD-MR	?	D (-0.41)	D (-1.73)	N (0.3)	D (3, 78%)	D (5)	9	N
T414M*	LGMD-MR	?	D (-0.33)	D (-2.17)	D (-0.66)	D (4, 82%)	N (1)	8	A, S, T
A669T	CMD-MR	0.004 pmol/h/mg proteins [28]	D (-1.37)	D (-2.39)	D (-0.92)	D (1, 63%)	D (5)	9	A, S
G65R	LGMD-MR	0.002 pmol/h/mg protein [28 #]	D (-1.31)	D (-2.66)	D (-0.69)	D (1, 63%)	D (7)	9	H, G
G76R	WWS	None [2] 10% [29]	D (-0.57)	D (-0.90)	N (+0.69)	D (1, 63%)	D (5)	5	A, T, M, I, G, V
L171A	LGMD-MR	40% [29]	D (-2.22)	N (+0.51)	D (-1)	N (0, 53%)	D (8)	8	I,L,V
A200P	LGMD-MR	None [28]	D (-1.05)	D (-0.89)	D (-0.1)	D (3, 78%)	D (7)	7	A, T, D, G, V
V428D	WWS	None [2]	D (-1.11)	D (-3.26)	D (-0.87)	D (6, 93%)	D (9)	8	F, I, V
W582C	LGMD-MR	0.002 pmol/h/mg protein [28 #]	D (-1.53)	N (+0.94)	D (-1)	D (3, 78%)	D (6)	7	F, W

Table 7.2. Summary of missense POMT1 mutation effects.
The mutations are listed with their associated phenotype, enzymatic activity and several *in silico* predictions. I-Mutant, MUpro, and SNPs3D predict changes in protein stability in terms of ΔΔG, whereas SNAP and PhD-SNP predict a variant as disease-related or as neutral polymorphism. In the table we report reliability parameters for each prediction in parentheses. **I-Mutant** calculates the free energy change value (ΔΔG), where a $\Delta\Delta G < 0$ indicates decrease of stability. **SNPs3D** uses a support vector machine (SVM) to find the separation pattern between a set of disease and non-deleterious SNPs. A positive score indicates variants classified as non-deleterious. **MUpro** predictions were reported with the confidence score (C score). A negative score indicates the mutation decreases protein stability, where lower scores imply higher confidence. In **SNAP**, variations are listed as "neutral" or "non-neutral" with reliability indices (RI; range 0–9) and Expected Accuracy (EA; range 1-100%) indicative of confidence in prediction. Higher RI correlates strongly with higher prediction accuracy. Expected accuracy is a number of correctly predicted neutral or non-neutral samples (at a given reliability index) in the SNAP testing set. SNAP only reports predictions that are made with at least 50% accuracy. **PhD-SNP** classifies a mutation as disease-related or as neutral polymorphism. As with SNAP, the reliability index (RI) indicates the confidence of predictions. **ConSeq** scores the sequence conservation from 0 to 9, with 9 being highly conserved and 0 being highly unconserved (i.e. variable). The last column shows the residue types present in that position of the multiple sequence alignment. **Abbreviations**: C score : Confidence score; RI : reliability Index; EA : Expected Accuracy; AAs : amino acids; LGMD-MR : limb-girdle muscular dystrophy with mental retardation; CMD-MR : congenital muscular dystrophy with mental retardation; WWS : Walker-Warburg Syndrome. "*" : novel mutation; "#" : measured in lymphoblasts from a compound heterozygous carrier of p.G65R and p.W582C.

7.5. Discussion

While the clinical spectrum of dystroglycanopathies is broad, cardiac involvement has been reported only in patients with *FKRP [357-360]* and *FKTN [365]* mutations. Similar to other known glycosyltransferases, POMT1 is expressed ubiquitously in all human tissues. Skeletal and cardiac muscles, in particular, show above-average levels of expression3. It is thus quite surprising that no signs of cardiomyopathy have been described in the approximately 40 previously reported patients carrying *POMT1* mutations (Leiden muscular dystrophies pages at http://www.dmd.nl/).

The patients in our series all presented with dilation and/or decreased left ventricular contractility, variable right ventricle involvement, and all had a good response to pharmacological therapy.

Muscle and CNS involvement in the patients was variable ranging from the mild to the severe ends of the POMT1 clinical spectrum. Patient #1, whose clinical immunohistochemical and genetic features were documented prior to the development of cardiomyopathy [361], had CMD with severe mental retardation and a normal brain MRI, a phenotype known to be associated with *POMT1* mutations [363]. Patients #2 and 3, conversely, differ from the classical LGMD2K, which usually includes overt mental retardation [354-355].

It has been hypothesized that mutations which completely disrupt mannosyltransferase activity are associated with more severe phenotypes (WWS), while those allowing residual enzyme activity are linked to milder ones (CMD-MR/LGMD2K)[351, 354]. Recent findings suggest that this correlation is weaker with regards to putative glycosyltransferase genes, such as *FKTN* or *FKRP*, but stronger for genes with a known enzyme product, such as *POMT1* [366]. In fact, studies that measure POMT1 activity in Sf9 cell lines co-expressing mutated POMT1 with wild type POMT22 or in immortalized lymphoblasts from patients carrying *POMT1* mutations [364] has demonstrated a marked reduction in POMT activity in the mutations/patients studied, but were unable to precisely predict phenotype severity. On the other hand, measurement of POMT activity using dermal fibroblasts from *POMT1* mutated patients showed that clinical phenotype severity is inversely correlated with POMT1 activity [367].

A direct correlation between mannosyltransferase activity and clinical severity, however, does not seem to apply to heart involvement which in our patients appears possible with very different degrees of neuromuscular severity, and with both complete and partial glycosylation defects (detected in the skeletal muscle). It remains to be established if the development of cardiomyopathy in our patients can be mutation-dependent and if specific *POMT1* mutations can predispose to cardiac deterioration. Indeed, all the identified mutations in our patients seem to indicate some degree of functional relevance. Some of the identified mutations, such as stop-codon mutations or the in-frame skipping of exon 12, which codes for a portion of the catalytic MIR domain, have an easily predictable deleterious effect on enzyme activity. Novel missense mutations, on the other hand, need further studies in order to better assess pathogenicity. *In silico* predictions of protein structure, summarized in Table 7.2, have localized these mutations into transmembrane helices, probably interfering with protein folding and stability, or into the MIR domain, in close proximity with previously described WWS-associated mutations which completely impair catalytic function, and thus probably alter the protein by means of similar mechanisms. We did not however expect our patients to have a complete defect of POMT1 enzymatic function, especially in those cases in which the phenotype was relatively mild and/or there was residual α-DG immunolabeling with antibodies against glycosylated epitopes. This may explain why some identified mutations have a predicted benign or slightly damaging effect on enzyme structure with some of the bioinformatic models that have been employed, suggesting that they allow for the expression of a partially viable and functioning enzyme.

The mechanism of both cardiomyocyte and muscle fiber damage in dystroglycanopathies is probably loss of dystroglycan function due to insufficient glycosylation and subsequent accumulation of membrane damage in response to exercise-induced stress, as suggested by animal models [368]. In our patients, the myocardium may have been particularly stressed by specific conditions, such as respiratory failure (patient #1) or several years of relatively strenuous exercise in adults with a globally preserved motor function (patients #2 and 3).

We speculate that all or most of patients with severe WWS phenotypes would probably develop cardiomyopathy if their lifespan were longer, while in patients with more

residual enzymatic activity and milder phenotypes, different mutations may determine a different pattern and timing of multisystemic involvement, explaining the absence of cardiomyopathy in most CMD and LGMD patients.

Bioinformatic prediction of the effect of missense mutations upon protein folding and function appears to yield results which are consistent with *in vitro* enzymatic assay findings, and may prove useful especially in those laboratories in which these assays are unavailable. Further studies on α-DG glycosylation in myocardial tissue will help in clarifying genotype-phenotype correlations and the mechanisms by which *POMT1* mutations and dystroglycanopathies in general selectively involve the myocardium.

Our report expands the phenotypical spectrum of *POMT1* mutations, adding cardiomyopathy to LGMD forms with slight cognitive impairment and to CMD-MR. In the light of these findings, we would recommend that clinicians monitor patients with *POMT1*mutations closely, regardless of their neuromuscular phenotype, to detect precocious signs of cardiac dysfunction. Gadolinium-enhanced cardiac MRI studies in these patients may help to detect subclinical heart involvement, making timely therapeutic interventions possible.

8. Deletions and Mutations in the Acidic Lipid-binding Region of the Plasma Membrane Ca^{2+} Pump

This chapter has been published in Brini M, Di Leva F, Ortega CK, Domi T, Ottolini D, Leonardi E, Tosatto SCE, Carafoli E. Deletions and mutations in the acidic lipid-binding region of the plasma membrane Ca2+ pump: a study on different splicing variants of isoform 2. J Biol Chem. 2010 Oct 1;285(40):30779-91.

8.1. Summary

Mutations of PMCA2 have been causally linked to human deafness and ataxia. The plasma membrane calcium ATPase (PMCA) uses energy to pump calcium (Ca2+) ions out of the cytosol into the extracellular milieu, to maintain a relatively low intracellular net Ca2+ load. The transcript of this gene is alternatively spliced at sites A and C to generate several variant proteins. The physiological meaning of the existence of so many isoforms is not clear, but evidently it must be related to the cell-specific demands of Ca(2+) homoeostasis. Tissue-restricted isoforms are indeed more active in exporting Ca2+ than the ubiquitous isoforms, probably due to their higher affinity for the activator calmodulin. The affinity of PMCA2 for Ca2+ is also modulated by the interaction of acidic phospholipids (PL) with two phospholipid-binding sites: the C-terminal PL binding domain and the AL region next to site A of alternative splicing. The calmodulin regulation of the pump has been extensively investigated and is now well understood but that mediated by PL is still unclear.

I built the homology-derived three-dimensional (3D) model of PMCA2 based on the PDB structure of the sarco/endoplasmic Ca2+ - ATPase (SERCA1a). To improve the correctness of target-template alignment, I considered also the results derived from several transmembrane region prediction methods. The electrostatic surface analysis of

the PMCA2 model indicated that the four conserved lysines in the AL PL-binding region, and stretching toward insertion sites A, form a positive charged bend that could easily accommodate a negative charged PL head of the membrane. The activity and the PL sensitivity of different PMCA2 splicing variants have been experimentally investigated by the group of E. Carafoli (Department of Biological Chemistry, Padua). They tested PMCa2 mutants confirming the importance of these lysines.
In order to explain the different sensitivity to CaM of the two variants differing in the insertion site C, I built a model of the PMCA2 CaM-binding region in complex with CaM protein. In the CaM sensitive isoform, this region forms a distinctive pattern of charged and hydrophobic residues crucial for CaM-PMCA2 interaction. This pattern is partially altered in the truncated isoform w/a which share a poor sensitivity to CaM.
On the basis of modelling studies, I proposed a structural interpretation of the interplay of the pump with PL, and the mechanism of their activation.

8.2. Abstract

Acidic phospholipids increase the affinity of the plasma membrane Ca^{2+}-ATPase pump for Ca^{2+}. They interact with the C-terminal region of the pump and with a domain in the loop connecting transmembrane domains 2 and 3 (A_L region) next to site A of alternative splicing. The contribution of the two phospholipid-binding sites and the possible interference of splicing inserts at site A with the regulation of the ATPase activity of isoform 2 of the pump by phospholipids have been analyzed. The activity of the full-length *z/b* variant (no insert at site A), the *w/b* (with insert at site A), and the *w/a* variant, containing both the 45-amino acid A-site insert and a C-site insert that truncates the pump in the calmodulin binding domain, has been analyzed in microsomal membranes of overexpressing CHO cells. The A-site insertion did not modify the phospholipid sensitivity of the pump, but the doubly inserted *w/a* variant became insensitive to acidic phospholipids, even if containing the intact A_L phospholipid binding domain. Pump mutants in which 12 amino acids had been deleted, or single lysine mutations introduced, in the A_L region were studied by monitoring agonist-induced Ca^{2+} transients in overexpressing CHO cells. The 12-residue deletion

completely abolished the ATPase activity of the *w/a* variant but only reduced that of the *z/b* variant, which was also affected by the single lysine substitutions in the same domain. A structural interpretation of the interplay of the pump with phospholipids, and of the mechanism of their activation, is proposed on the basis of molecular modeling studies.

8.3. Introduction

The plasma membrane Ca^{2+}-ATPases (PMCAs) extrude Ca^{2+} from cells, maintaining the resting level of intracellular Ca^{2+} and controlling the Ca^{2+} transients induced by agonists. Four basic PMCA isoforms are encoded by four independent genes. *PMCA1* and *-4* are ubiquitously expressed, whereas *PMCA2* and *-3* are restricted to brain, muscles, and few other tissues; the tissue-restricted isoforms are more active in exporting Ca^{2+} than the ubiquitous isoforms [369], probably due to their higher affinity for the activator calmodulin. The transcript of each gene is subjected to alternative splicing at sites A and C. About 30 splice variants have so far been detected at the RNA or protein levels [370].

The architecture of the PMCAs predicts 10 transmembrane domains, two large intracellular loops, and N- and C-terminal cytoplasmic tails. The 90-residue N-terminal portion appears not to have specific functions even if it contains a consensus binding site for the 14-3-3 protein, which inhibits three of the four pump isoforms [371-372]. The cytosolic loop between transmembrane domains 2 and 3 contains a site that binds activatory acidic phospholipids and site A of alternative splicing upstream of it. Pump variants containing the A-splice site insert are targeted to the apical plasma membrane [373], and the insert has recently been suggested to have a role in the interactions of the pump with lipids in the plasma membrane [374]. The C-terminal tail contains other regulatory sites of the pump, among them the positively charged calmodulin binding domain, which also binds acidic phospholipids [375], the consensus sites for protein kinases A (PKA, isoform-specific) and C (PKC), and high affinity allosteric Ca^{2+}-binding sites.

Under nonactivated conditions, the C-terminal tail of the pump is proposed to fold over to interact with two sites in the first and second cytosolic loops of the enzyme, compromising the access to the active center. Calmodulin then interacts with its binding domain, removing it from its docking sites next to the active center and freeing the pump from autoinhibition.

The calmodulin regulation of the pump has been extensively investigated and is now well understood but that mediated by acidic phospholipids is still unclear. Acidic phospholipids enhance the Ca^{2+} sensitivity of the PMCA to a greater extent than calmodulin [376-379]. The order of stimulatory potency (phosphatidylinositol 4,5-biphosphate > phosphatidylinositol 4-phosphate > phosphatidylinositol ~ phosphatidylserine (PS) ~ phosphatidic acid) is proportional to the number of negative charges on the lipids [380]. The stimulation is appreciably reduced by complexing the negative charges with polyamines or neomycin [381]. Recently, diacylglycerol has also been shown to be a stimulator of the PMCA. Interestingly, the activation induced by diacylglycerol is additional to that produced by calmodulin and PKC, suggesting that diacylglycerol interacts with the PMCA through a specific mechanism [382].

The acidic phospholipid-binding region next to splice A was recently deleted in a variant of PMCA4 containing an inserted exon at splicing site A (variant *xb*) [383-384]. Partial deletions did not alter Ca^{2+} transport activity but made the pump insensitive to acidic phospholipids. However, complete removal of the domain made the pump inactive [383].

The contributions of the two phospholipid-binding sites, and of the alternative splicing at site A next to one of them, to the regulation of the pump have not been analyzed. It was interesting to study these aspects on isoform 2 of the pump, as this isoform has very high activity even in the absence of calmodulin [385-386], but it responds to acidic phospholipids in the same way as PMCA4 [385]. In addition, the splicing mechanisms of PMCA2 generate a larger number of variants than in other isoforms; up to three exons are inserted at site A, generating variant *z* (no exons included), variant *y* (two exons included), variant *x* (one exon included), and variant *w* (all three exons included). Splicing at site C excludes two novel exons (variant *b*, full length) or includes them (variant *a*). The *a* insertion leads to a truncated version of the pump that only contains about half of the original calmodulin binding domain [385-386].

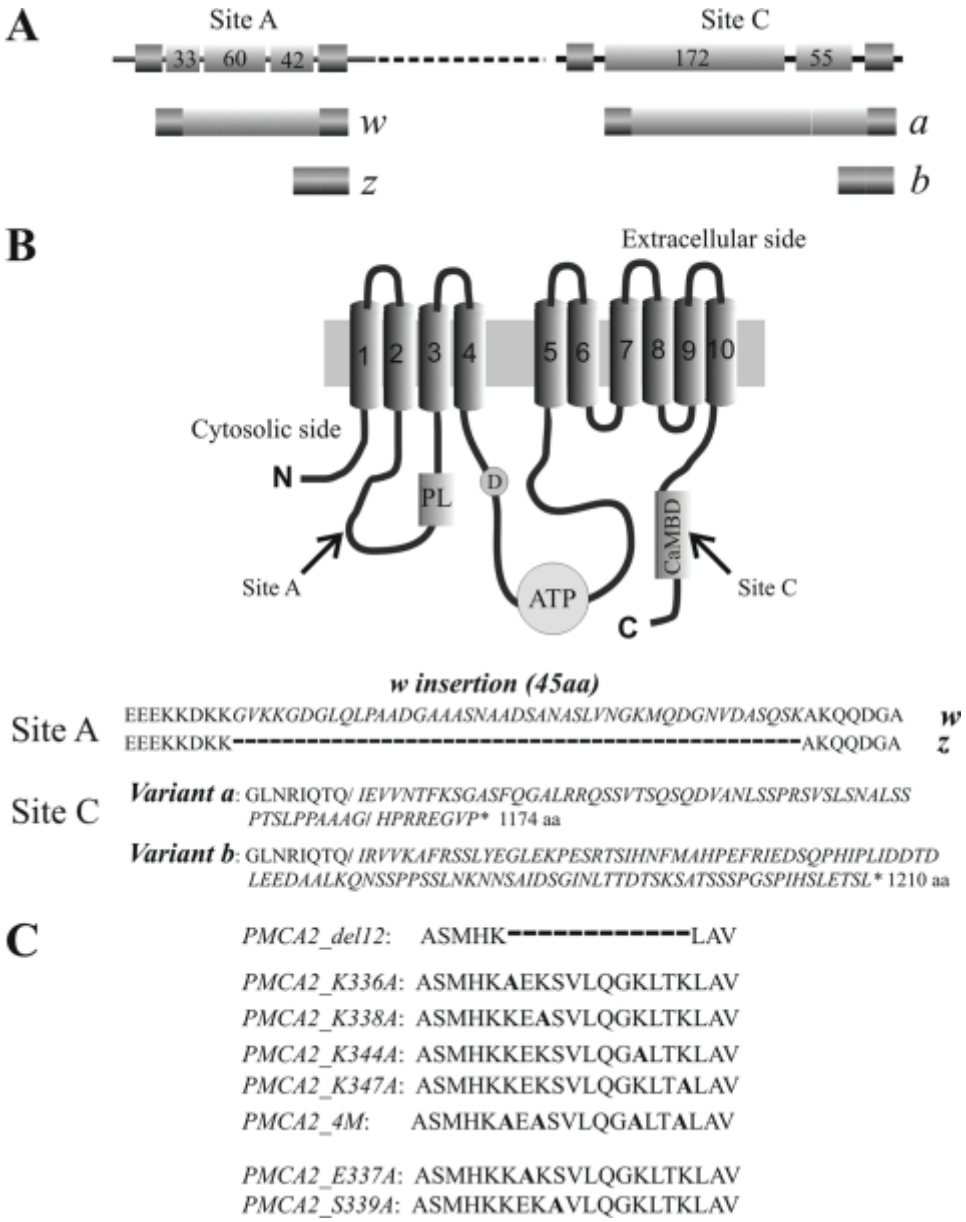


Figure 0.1. Alternative splicing of the PMCA2 transcripts.
A. linear representation of the alternative splicing options at site A and site C of the PMCA2 transcript. Exons are indicated by shadow boxes and introns by the black line. The numbers in the boxes represent the nucleotide number of each exon. B. topography model of the plasma membrane Ca^{2+}-ATPase and sequences of alternative splicing products of isoform 2. The 10 putative transmembrane domains are numbered and indicated by shadow boxes. PL indicates the phospholipid binding domain downstream of site A of alternative splicing; D indicates the catalytic aspartate; ATP and CaMBD indicate the ATP-binding site and the calmodulin binding domain, which contains site C of alternative splicing. C. sequences of the PMCA2 region that have been mutated or deleted in the constructs used in this study. The alanine that replaces the mutated residue in the different constructs is indicated in bold. The dashed line represents the 12 amino acids deletion.

We had previously reported that the *z/a* and *w/b* PMCA2 variants behaved essentially as the full-length, noninserted, (*z/b*) pump (perhaps, they were slightly less efficient) [369,

387-388]. The doubly inserted *w/a* PMCA2 variant had only limited ability to rapidly increase activity when challenged with a Ca^{2+} pulse but had about the same highly nonstimulated (basal) activity of the full-length *z/b* variant [387].

This contribution explores the activation of splicing variants of isoform 2 of the PMCA pump by acidic phospholipids. Because the negative charges on the lipids are likely to be important in the stimulatory effect, the study was performed using a pump variant in which a 12-residue stretch in the A_L acidic phospholipid binding domain, which contains four positively charged residues, was removed. Point mutations that selectively substituted positive residues (Lys), or two other conserved polar residues (Ser and Glu), were also introduced in the stretch. The scheme of Figure 8.1 summarizes graphically the details of the PMCA2 variants and mutants used in this study.

8.4. Experimental Procedures

Cell Cultures and Transfection

CHO cells were cultured in Ham's F-12 nutrient mixture (Invitrogen), supplemented with 10% fetal bovine serum (FBS), 2 mM glutamine, penicillin (60 μg/μl), and streptomycin (120 μg/μl) in 75-cm^2 Falcon flasks at 37 °C. For the microsomes preparation, CHO cells were plated on 150 × 25-mm Petri dishes, allowed to grow to 50% confluence, and transfected according to a calcium-phosphate procedure with 30 μg of total plasmid DNA. For the aequorin and immunocytochemistry experiments, CHO cells were plated onto 13-mm glass coverslips, allowed to grow to 50% confluence, and transfected according to a calcium-phosphate procedure with 3 μg of total plasmid DNA or with 1.5 μg of each plasmid DNA in the case of co-transfection. GFP-tagged PMCA2 *z/b* and *w/b* are of human origin, and GFP-tagged PMCA2 *w/a* variants (WT and del12 mutant) are from rat. Untagged PMCA2 pump variants (*w/a* and *z/b*) of human origin were also used in the Ca^{2+} measurements of experiments in living cells. No differences were observed between the GFP-tagged and -untagged PMCA2 activity.

The average transfection efficiency approached 25%, and the increase of PMCA protein in overexpressing cells, calculated by densitometric analysis of Western blotting

showing the endogenous PMCA (*i.e.* blots developed with the monoclonal antibody 5F10 that recognized all PMCA isoforms) and corrected for the whole cell population, would correspond to about 3-fold the endogenous level (data not shown).

Microsomal Membrane Preparations from CHO Cells

Cells from five 150 × 25-mm dishes were washed once with phosphate buffered saline (PBS) containing 1 mM EDTA and harvested in 10 ml of PBS containing 0.1 mM phenylmethylsulfonyl fluoride (PMSF) and a mixture of EDTA-free protease inhibitors (Roche Applied Science). Cells were collected by centrifugation (2000 × *g*, 10 min) at 4 °C and resuspended in 6 ml of a hypotonic solution of 10 mM Tris-HCl, pH 7.5, 1 mM $MgCl_2$, 0.1 mM PMSF, a mixture of EDTA-free protease inhibitors, and 2 mM dithiothreitol (DTT). The cells were swollen for 15 min on ice and then subjected to three cycles of freeze and thaw. The homogenate was diluted with an equal volume of 0.5 M sucrose, 0.3 M KCl, 2 mM dithiothreitol, 10 mM Tris-HCl, pH 7.5, homogenized again with three cycles of freeze and thaw, and centrifuged at 5000 × *g* for 15 min. KCl was added up to 0.6 M in the supernatant, and to remove calmodulin, an excess of EDTA (1.5 mM) was also added. The suspension was centrifuged at 100,000 × *g* for 40 min to pellet the microsomal fraction. The final pellet was resuspended in a solution containing 0.25 M sucrose, 0.15 M KCl, 10 mM Tris-HCl, pH 7.5, 2 mM DTT, and 20 μM $CaCl_2$, at a protein concentration of 1–3 mg/ml, and stored in liquid N_2.

ATPase Activity Assay

The ATPase activity was measured by the coupled enzyme assay (modified from Ref. [387]) monitoring the absorbance of NADH at 340 nm. The decrease in A_{340} can be converted into ATPase activity where one molecule of NADH oxidized to NAD^+ corresponds to the production of one molecule of ADP by the ATPase. The assay was carried out at 37 °C in a final volume of 1 ml of a mixture containing 20 mM Tris-HCl, pH 7.2, 5 mM $MgCl_2$, 0.5 mM EGTA, 0.1 M KCl, 0.5 mM phosphoenolpyruvate, 0.15 mM NADH, 1.4 units of pyruvate kinase/lactic dehydrogenase (Roche Applied Science), 4 mM ATP, 25 μg of PMCA membranes, and 50 μM $CaCl_2$. The ATPase activity, detected at 340 nm (DU640 Spectrophotometer, Beckman Coulter), was expressed in micromoles of P_i/min/mg of protein (moles of phosphate originated from the ATP

hydrolysis); the maximal activity and the basal activity were calculated by multiplication of the activity curve slope value by a factor considering the NADH molar extinction coefficient (ϵ_{NADH}) and the amount of protein (in micrograms). The real activity was obtained subtracting the basal activity from the maximal activity. The assay was performed in the presence of 5 μg/ml oligomycin and 0.1 μM thapsigargin. To test calmodulin (CaM) or phosphatidylserine (PS) activation of the pump, 200 nM CaM or 25 μM PS was preincubated with the membranes for 5 min at 37 °C before starting the assay.

Generation of PMCA2z/b_del12 and PMCA2z/b Mutant Expression Plasmids

To generate PMCA2*z/b* with the deletion of 12 amino acids in the domain that binds acidic phospholipids, two PCR amplification products that did not contain the portion of 12 amino acids were generated using four different primers bearing restriction sites for EcoRI/HindIII and HindIII/BamHI as follows: 5′-cggGAATTC**atg**ggtgacatgaccaac-3′; 5′-cggTTCGAAgtgcatgctggccttcct-3′; 5′-cggTTCGAAgctgtgcagatcgggaag-3′; cggGGATCCctaaagcgacgtctccag. The PCR products were digested with the respective restriction enzymes and were inserted in a three-part ligation reaction in pcDNA3 vector (Invitrogen) digested with EcoRI and BamHI. The construct was controlled by sequencing.

In vitro site mutagenesis in the PMCA2*z/b* was carried out with QuikChange II site-directed mutagenesis kit (Stratagene) according to the manufacturer's instructions using the following primers:

Lys-336 sense 5′-ccagcatgcacaag**GC**ggagaagtccgtgc-3′ and antisense 5′-gcacggacttctcc**GC**cttgtgcatgctgg-3′;

Lys-338 sense 5′-tgcacaagaaggag**GC**gtccgtgctgcagg-3′ and antisense 5′-cctgcagcacggac**GC**ctccttcttgtgca-3′;

Lys-344 sense 5′-cgtgctgcagggc**GC**gctcaccaagctg-3′ and antisense 5′-cagcttggtgagc**GC**gccctgcagcacg-3′;

Lys-347 sense 5′-gggcaagctcacc**GC**gctggctgtgcagat-3′ and antisense 5′-atctgcacagccagc**GC**ggtgagcttgccc-3′;

Glu-337 sense 5′-gcatgcacaagaagg**C**gaagtccgtgctgcagggc-3′ and antisense 5′-gccctgcagcacggacttc**G**ccttcttgtgcatgc-3′;

Ser-339 sense 5′-gcatgcacaagaaggagaag**G**ccgtgctgcagggc-3′ and antisense 5′-gccctgcagcacgg**C**cttctccttcttgtgcatgc-3′.

The PMCA2_4 M mutant, in which all the four lysines were mutated, was generated by subsequent cycles of PCR amplification using the following primers:

Lys-336_338 sense 5′-ccagcatgcacaag**GC**ggag**GC**gtccgtgctgcagg-3′ and antisense 5′-cctgcagcacggcc**GC**ctcc**GC**cttgtgctgctgg-3′;

Lys-344_347 sense 5′-cgtgctgcagggc**GC**gctcacc**GC**gctggctgtgcaga-3′ and antisense 5′-tctgcacagccagc**GC**ggtgagc**GC**gccctgcagcacg-3′.

Mutated bases are indicated by boldface capital letters.

Immunocytochemistry Analysis

CHO cells were transfected with the different PMCA2 variants and mutants. 36 h after transfection, the cells were washed twice with PBS and fixed with 3.7% formaldehyde for 20 min. The membranes were permeabilized in 0.1% Triton X-100 for 5 min and washed with 1% gelatin (type B, from bovine skin, Sigma) in PBS. The cells were immunostained with primary antibodies against PMCA2 (2N, Sigma) at a 1:100 dilution in PBS and with secondary antibodies Alexa Fluor 594 (Molecular Probes). The images were acquired using a Zeiss Axiovert microscope equipped with a 12-bit digital cooled camera (Micromax-1300Y, Princeton Instruments Inc., Trenton, NJ) using Metamorph software (Universal Imaging Corporation, West Chester, PA).

Preparation of Membranes from CHO Cells, SDS-PAGE, and Western Blotting Analysis

Thirty six hours after transfection, CHO cells were harvested in 10 mM Tris-HCl, pH 8.0, 2 mM EDTA, 2 mM PMSF, 1 mM DTT. They were disrupted by three cycles of freeze and thaw at −80/37 °C, and the insoluble proteins were sedimented at 11,000 × *g* for 30 min (4 °C). The supernatant was discarded, and the pellet was resuspended in 5 mM Tris-HCl, pH 8.0, and 10% sucrose. Proteins were separated by 7.5% SDS-PAGE and transferred to nitrocellulose membranes. 20 μg of membrane proteins were loaded onto each lane. The sheets were probed with a rabbit polyclonal antibody 2N against PMCA2 (Sigma, diluted 1:1000). After incubation with anti-rabbit horseradish peroxidase-conjugated secondary antibodies (Santa Cruz Biotechnology, Santa Cruz, CA), the blots

were developed with ECL reagents (Amersham Biosciences). The quantitative analysis was carried out by densitometric analysis using the Kodak 1D Image Analysis program (Kodak Scientific Imaging System, New Haven, CT). Antibodies against β-tubulin or β-actin were also used to normalize the data obtained from the densitometric analyses.

Cytosolic Ca^{2+} Monitoring with Recombinant Aequorin

CHO cells were plated on 13-mm glass coverslips and transfected according to the calcium-phosphate procedure. 36 h after transfection, the cells were incubated for 3 h with 5 μM of the aequorin prosthetic group coelenterazine WT in Dulbecco's modified Eagle's medium supplemented with 1% FBS at 37 °C in a 5% CO_2 atmosphere.

After incubation with coelenterazine, the coverslips were placed in a perfused thermostated (37 °C) chamber of a luminometer positioned in close proximity to a low noise photomultiplier, with a built-in amplifier discriminator. The experiments were performed in a Krebs-Ringer medium (135 mM NaCl, 5 mM KCl, 0.4 mM KH_2PO_4, 1 mM $MgSO_4$, 20 mM Hepes, pH 7.4, at 37 °C) (KRB) supplemented with 0.1% glucose and 1 mM $CaCl_2$. The cytoplasmic Ca^{2+} concentrations were measured after addition of 100 μM inositol 1,4,5-trisphosphate-generating agonist ATP. The experiments were terminated by lysing the cells with 100 μM digitonin in a hypotonic Ca^{2+}-rich solution (10 mM $CaCl_2$ in H_2O) to discharge the remaining aequorin pool. The light signal from the discriminator was collected by a Thorn-EMI photon counting board and stored in an IBM-compatible computer for further analysis. The aequorin luminescence data were calibrated off line into [Ca^{2+}] values, using a computer algorithm based on the Ca^{2+}-response curve of wild type aequorin [388].

In Silico Analysis

The protein sequence of human PMCA2 was retrieved from the NCBI data base (accession number NP 001674) [235], and amino acid conservation was evaluated with Conseq [132]. Secondary structure and disorder were predicted by a consensus approach [99] and SPRITZ [111], respectively. A consensus of three methods (Prodiv-TMHMM, HMMTOP, and PHOBIUS) was adopted to predict the transmembrane regions. A homology-derived three-dimensional structure model of human PMCA2 was constructed using the Homer-A modeling server based on the PDB structure 2agv

(chain A) of sarco/endoplasmic Ca^{2+}-ATPase (SERCA1a). The loop insertions in human PMCA2 were modeled using a divide and conquer method [88]. The C-terminal PMCA2 CaM-binding region in complex with calmodulin was modeled using the PDB structure 2KNE as template. We used the PyMOL Molecular Graphics System (DeLano Scientific, San Carlo, CA) to map the residue positions in the protein structure and visualize the electrostatic surface calculated by the Adaptive Poisson-Boltzmann Solver tool [78].

Statistical Analysis

Data are reported as means ± S.D. Statistical differences were evaluated by Student's two-tailed *t* test for unpaired samples, with *p* value 0.01 being considered statistically significant.

8.5. Results

Expression of PMCA2 Isoforms in CHO Cells

GFP-tagged PMCA2 splice variants *z/b*, *w/b*, and *w/a* were overexpressed in CHO cells. Crude membranes were prepared, and 20 μg of total proteins were separated by SDS-PAGE and blotted onto nitrocellulose filters. The filter was incubated with polyclonal antibody 2N that recognizes the PMCA2 isoform and with an anti-tubulin antibody. Figure 8.2 shows that all three splice isoforms of the pump were expressed at approximately equivalent levels.

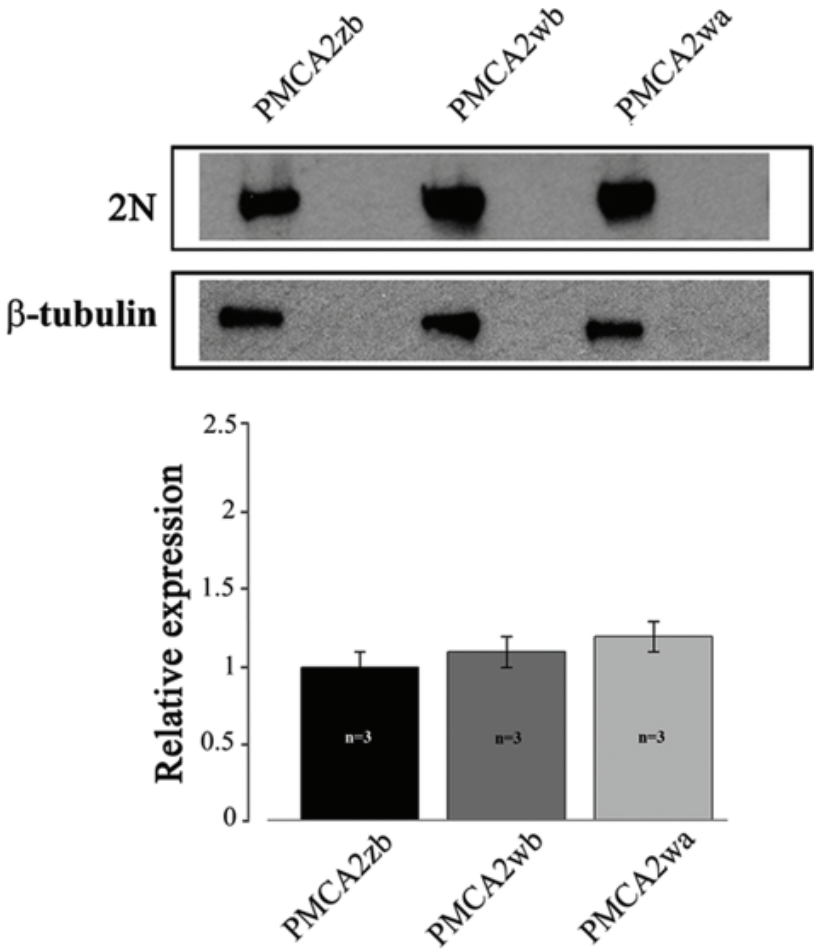


Figure 0.2. Expression of PMCA2 isoforms.
20 μg of crude membrane proteins from transfected CHO cells, prepared by a freeze and thaw method, were separated by SDS-PAGE as described under "Experimental Procedures" and stained with polyclonal antibody 2N, which recognizes isoform 2 of the pump or against tubulin. The lanes correspond to cells transfected with the indicated variants of PMCAs fused to GFP. The data are representative of at least three independent experiments.

Ca^{2+} ATPase Activity in Microsomal Membranes

Microsomal membranes (containing plasma membrane fragments/vesicles) isolated from transfected cells were assayed in the presence of thapsigargin and oligomycin to inhibit the activity of the endogenous sarco/endoplasmic reticulum Ca^{2+}-ATPase pump and the ATP-linked Ca^{2+} uptake by mitochondrial vesicles that could possibly contaminate the microsomal preparation. Figure 8.3, *A* and *D*, shows the PMCA activity in the absence of calmodulin. Both the noninserted full-length PMCA2 *z/b* variant and the *w/b* variant had higher basal activity than the inserted and truncated isoform *w/a*. The calmodulin sensitivity of each isoform was investigated at a fixed Ca^{2+} concentration in the presence of excess (200 nM) calmodulin (Fig. 8.3, *B* and *D*). As

already shown by previous work, the *w/a* variant had reduced stimulation by calmodulin in respect to the full-length *z/b* and *w/b* variants.

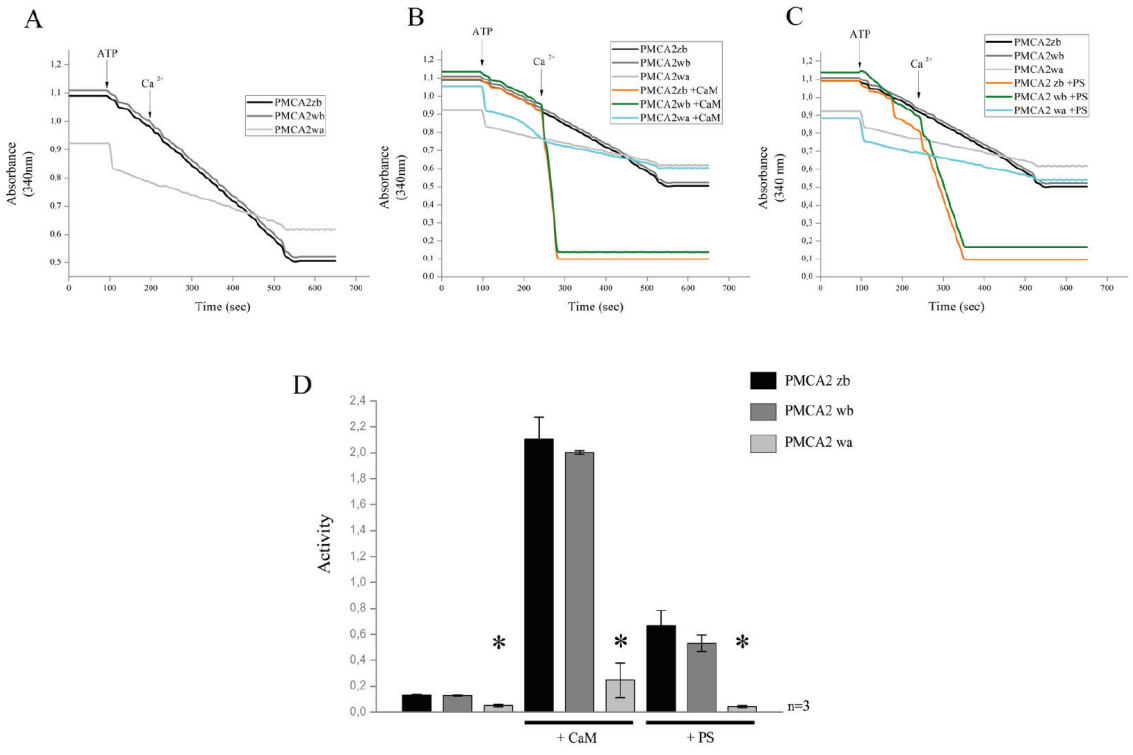


Figure 0.3. Ca^{2+} transport activity of PMCA2.
A, comparison of Ca^{2+} transport activity measured on microsomal membranes isolated from CHO cells overexpressing PMCA2 *z/b*, *w/b*, and *w/a* variants. Membranes vesicles were preincubated at 37 °C, and Ca^{2+} uptake was initiated by the addition of 4 mM ATP (where indicated). 50 µM $CaCl_2$ was added where indicated. ***B***, CaM dependence of Ca^{2+} uptake by microsomal membranes preincubated at 37 °C with 200 nM CaM. ***C***, acidic phospholipid (PS) dependence of Ca^{2+} uptake by microsomal membranes preincubated at 37 °C with 25 µM PS. *A–C*, ATPase activity was indicated as the decrease of the absorbance at 340 nm. ***D***, *histograms* show the means ± S.D. of the activity of the pumps. The activity was expressed as micromoles of P_i/min/µg of protein and calculated as indicated under "Experimental Procedures." The data are representative of at least three experiments with different membranes preparations. *, $p < 0.05$, in respect to the respective controls in the absence of CaM and PS.

The splicing event at site A occurs just upstream of one of the two regions responsible for the binding of acidic phospholipids. The first two spliced exons of PMCA2 encode a relatively hydrophobic stretch of amino acids positioned amid a highly charged region, suggesting possible effects on the overall interaction of the first cytosolic loop of the pump with acidic phospholipids. The response of the three pump variants *z/b*, *w/b*, and *w/a* to phosphatidylserine was thus compared. Figure 8.3, *C* and *D*, shows that isoforms *z/b* and *w/b* had the same response, implying that the A_L acidic phospholipid binding

domain was not affected by the site A insert. Surprisingly, however, the *w/a* variant, which has the phospholipid binding domain contiguous to the site A insert but lacks about half of the C-terminal phospholipid binding domain, was completely insensitive to phosphatidylserine; the response of the full-length variants of the pump (variants *b*) was over 5-fold higher than that of the truncated *w/a* variant. The finding thus suggests a predominant role of the C-terminal phospholipid binding domain in the response to acidic phospholipids.

Mutations in the N-terminal (A_L) Phospholipid Binding Domain

Mutational experiments on the phospholipid binding domains were performed to further explore the molecular mechanism of the activation of the pump by acidic phospholipids. As for the possible mechanism of acidic phospholipid stimulations, in the case of the binding sequence in the C-terminal calmodulin binding domain, it was reasoned that the headgroups of positively charged residues could be neutralized by acidic phospholipids, weakening the autoinhibitory intramolecular interaction of the C-terminal tail of the pump with its receptor sites in the main body of the molecule. The study of the phospholipid binding domain in the C-terminal region was limited to comparison of the *a* variant (in which the splicing truncation removes about half of the CaM binding domain and, presumably, affects the binding of acidic phospholipid binding domain) with the full-length *b* variant. No mutations were introduced in the full-length C-terminal region of the *b* variants.

In the A_L region, structural rearrangements of the transduction (activator) and catalytic domains of the pump could occur following the binding of phospholipids that would facilitate the access of Ca^{2+} to its single high affinity site in the transmembrane sector. The four lysines in the A_L phospholipid-binding region, which are very conserved among the PMCAs isoforms (Fig. 8.4*A*) and in the PMCA across species (Fig. 8.4*B*), could form a charged bend that could easily accommodate a charged phospholipid head. It was thus decided to mutate them. It was also felt that two other well conserved residues in the A_L binding domain (Glu-337 and Ser-339, Fig. 8.4) could also have a role in the interaction (see below). It was thus decided to mutate them as well. It was also decided to study the effect of the deletion of the 12-residue lysine-rich stretch in the A_L domain that had been previously performed by others [389].

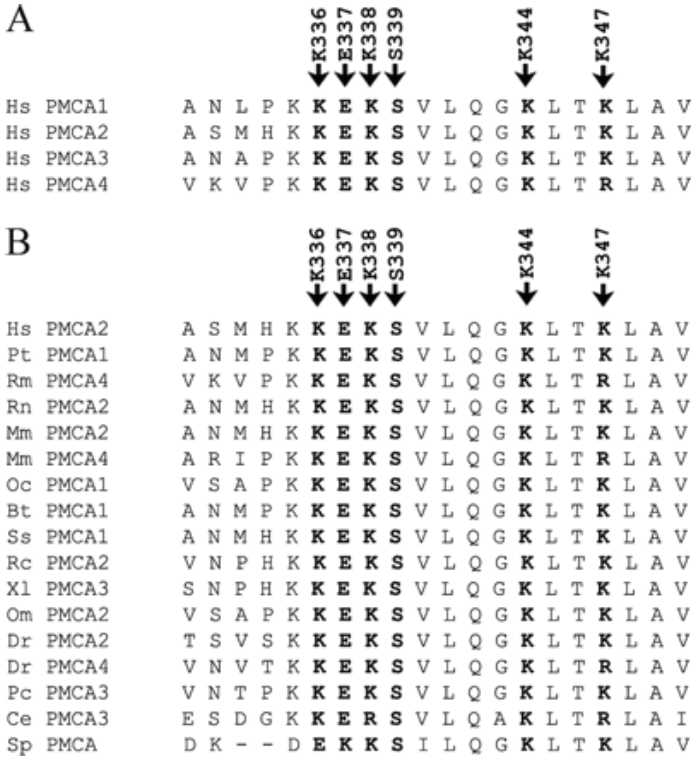


Figure 0.4. **Conservation of the A_L domain.**
The similarity analysis was performed using the ClustalW program. Human PMCA2 sequence (GenBank™ accession number NP_001674) is listed with other human PMCAs isoforms sequences (*A*) and with those of other species (*B*). GenBank™ accession numbers are as follows: NP_001001323 (*Homo sapiens* PMCA1), NP_068768 (*H. sapiens* PMCA3), NP_001675 (*H. sapiens* PMCA4), XP_509257 (*Pan troglodytes* PMCA1), AY928176 (*Rhesus macaque* PMCA4), NP_036640 (*Rattus norvegicus* PMCA2), AAH75643 (*Mus musculus* PMCA2), BC109173 (*M. musculus* PMCA4), Q00804 (*Oryctolagus cuniculus* PMCA1), NP_777121 (*Bos taurus* PMCA1), NP_999517 (*Sus scrofa* PMCA1), AAK11272 (*Rana catesbeiana* PMCA2), BC077905 (*Xenopus laevis* PMCA3), P58165 (*Oreochromis mossambicus* PMCA2), NP_001116710 (*Danio rerio* PMCA2), EU559285 (*D. rerio* PMCA4), AAR28532 (*Procambarus clarkia* PMCA3), AAK68551 (*Caenorhabditis elegans* PMCA3) and AAR13013 (*Stylophora pistillata*).

Generation, Expression, and Activity of PMCA2zb_del12 and PMCA2wa_del12 Mutants

The 12-residue lysine-rich stretch located in the N-terminal portion of the domain was analyzed first, as it had already been shown that the deletion of this stretch failed to affect the plasma membrane targeting of the pump [389]. The effect of the deletion of region 380–391 in the *w/a* variant, which has lost at least half of the C-terminal acidic phospholipid binding domain, and of region 336–347 in the full-length *z/b* variant, which contains it, was studied in over-expressing CHO cells. The activity of the deleted variants of the expressed pump was compared with that of their respective wild type variants. Appropriate controls (Western blotting and immunocytochemistry analysis)

established that the mutant pump variants were expressed at about the same levels with respect to their WT versions and were correctly delivered to the plasma membrane (Fig. 8.5*A*). The *PMCA2wa_del12* was expressed as a GFP fusion chimera; the fusion with GFP did not alter the targeting nor the activity of the pump [389]. As reported previously [390], the *w/a* variant was much less efficient than the *z/b* variant in re-establishing resting cytosolic Ca^{2+} concentrations following the increase induced by the stimulation of the cells with the purinergic agonist ATP (*traces* in Fig. 8.5, *B* and *C*). The Ca^{2+} transient generated by the stimulation reflects the inositol 1,4,5-trisphosphate-mediated Ca^{2+} release from the intracellular stores but also the Ca^{2+} influx from the extracellular medium through channels activated by the depletion of the endoplasmic reticulum stores. The lowering of the Ca^{2+} peak with respect to untransfected cells reflects the ability of the overexpressed pumps to respond with a burst of activation, *i.e.* of Ca^{2+} extrusion, to the arrival of the inositol 1,4,5-trisphosphate-generated Ca^{2+} pulse. The faster clearance of the Ca^{2+} signal is thus due to increased overall pump activity. Fig. 8.5*B* shows the Ca^{2+} response in cells transiently transfected with the *wa_wt* and *wa_del12* variants of PMCA2. Surprisingly, the deletion of the 12 amino acids in the phospholipid binding domain completely abolished the activity of the pump (the heights of the transients were *wa_wt*, 2.77 ± 0.35 μM, $n = 27$; *wa_del12*, 3.58 ± 031 μM, $n = 26$; control (only aequorin), 3.53 ± 0.48 μM, $n = 34$). The half-time of the declining phase was 7.69 ± 1.23 s, $n = 29$, in *wa_wt*, 44.52 ± 4.99 s in *wa_del12*, $n = 27$, and 46.67 ± 7.35 s, $n = 12$ in the control (see Fig. 8.5*B*, *inset*). Fig. 8.5*C* shows pump activity in cells transiently transfected with the *wt_zb* and *zb_del12* PMCA2 variants. The 12-amino acid deletion impaired the activity of the *zb* variant as well (*zb_wt*, 1.31 ± 0.17 μM, $n = 31$; *zb_del12*, 2.72 ± 0.28 μM, $n = 10$), suggesting that the deleted residues are important to pump activity independently of splicing processes. However, at variance with the *wa_del12* variant, the *zb_del12* variant was still partially active; the height of Ca^{2+} transient was reduced with respect to control cells. This finding is also supported by the analysis of the declining phase of the Ca^{2+} traces, in which the half-time of the peak decay was 6.31 ± 0.85 s, $n = 13$, in *zb_wt*, 38.64 ± 5.93 s in *zb_del12*, $n = 14$, *versus* 46.67 ± 7.35 s, $n = 12$, in the control, $p < 0.01$ (see Fig. 8.5*C*, *inset*).

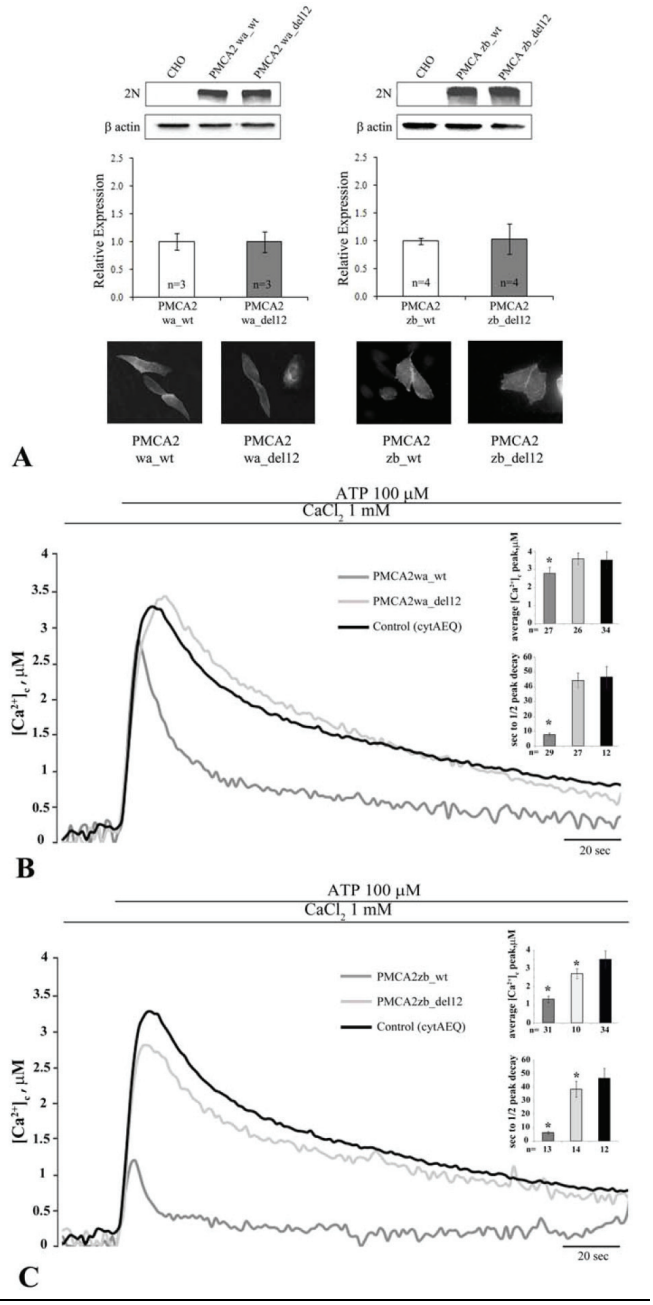


Figure 0.5. Expression and activity of PMCA2 variants.
A, Western blotting and densitometric analysis of the variants of the PMCA2 isoform overexpressed in CHO cells. 20 µg of crude membrane proteins from transfected CHO cells, prepared by a freeze and thaw method, were separated by SDS-PAGE as described under "Experimental Procedures" and stained with polyclonal antibody 2N. The control lane corresponds to nontransfected cells (*CHO*). The other lanes correspond to cells transfected with the WT or mutant variants of the PMCA2 pump. The *panel a*lso shows the immunocytochemistry analysis of the transfected CHO cells. The immunostaining was carried out with the 2N antibody and revealed with the secondary antibody Alexa Fluor 594. *B*, monitoring of cytosolic [Ca^{2+}] in CHO cells transfected with cytAEQ and co-transfected with cytAEQ and the WT *w/a* variant of PMCA2 isoform or deleted PMCA2*wa_del12* mutant. *C*, monitoring of cytosolic [Ca^{2+}] in CHO cells transfected with cytAEQ and co-transfected with cytAEQ and the wt *z/b* variant of PMCA2 isoform or deleted PMCA2*zb_del12* mutant. The *histograms* in *B* and *C* show the means ± S.D. of $[Ca^{2+}]_c$ peaks and of the half-time decays from the peaks. The traces are representative of at least 12 independent experiments. *, $p < 0.01$ calculated with respect to control (CHO cells transfected only with cytAEQ).

Generation, Expression, and Activity of PMCA2 z/b Variants Harboring Single Lys Mutations in the 336–347 Domain

Single amino acids mutants of the PMCA2 *z/b* pump were generated by replacing individual lysines in the 12-residue sequence (336–347 domain). Five mutants were generated as follows: PMCA2*zb_K336A*, PMCA2*zb_K338A*, PMCA2*zb_K344A*, PMCA2*zb_K347A*, and one in which all four lysines were replaced with alanines, PMCA2*zb_4M*. The positions of the mutated lysines in the sequence are shown in Fig. 8.4. The level of expression of all mutants and their correct delivery to the plasma membrane were checked and found to be equivalent (Fig. 8.6*A*). The single mutation of three of the four lysines impaired the activity of the pump (the heights of the peak transients induced by the stimulation were as follows: 1.50 ± 0.15 μM, n = 12 for PMCA2 *zb_K338A*; 1.96 ± 0.09 μM, n = 15 for PMCA2 *zb_K344A*; and 1.56 ± 0.22 μM, n = 15 for PMCA2 *zb_K347 versus* 1.31 ± 0.17 μM, n = 31 for the *zb_wt*, $p < 0.01$) (Fig. 8.6*B*, in which the Ca^{2+} transients were superimposed to that generated in cells overexpressing equivalent levels of PMCA2*zb_wt*). Fig. 8.6*B* shows that instead the mutation of lysine 336 (K336A) had no effect on the Ca^{2+} extruding ability of the pump; the height of the transient was 1.36 ± 0.15 μM, n = 12, as compared with 1.31 ± 0.17 μM, n = 31, in *zb_wt*-expressing cells.

The mutation of all four lysines impaired the Ca^{2+} extrusion activity of the pump. The peak height was 1.78 ± 0.16 μM, n = 15, for PMCA2 *zb_4M versus* 1.31 ± 0.17 μM, n = 31, for *zb_wt*, $p < 0.01$. It also affected the ability of the pump to accelerate the declining phase of the Ca^{2+} transient trace. It did so more significantly than in the case of single lysine mutants, as shown by the *traces* and the *histograms* of Figure 8.6*B*. The half-time of the declining phase was 7.25 ± 1.16 s, n = 8, for PMCA2*zb_K336A*; 7.5 ± 0.83 s, n = 6, for PMCA2*zb_K338A*; 9.37 ± 2.02 s, n = 16, for PMCA2*zb_K344A*; 6.89 ± 0.93 s, n = 13, for PMCA2*zb_K347A*; 13 ± 2 s, n = 16, for PMCA2*zb_4M*, and 6.31 ± 0.85 s, n = 13, for *zb_wt*, $p < 0.01$.

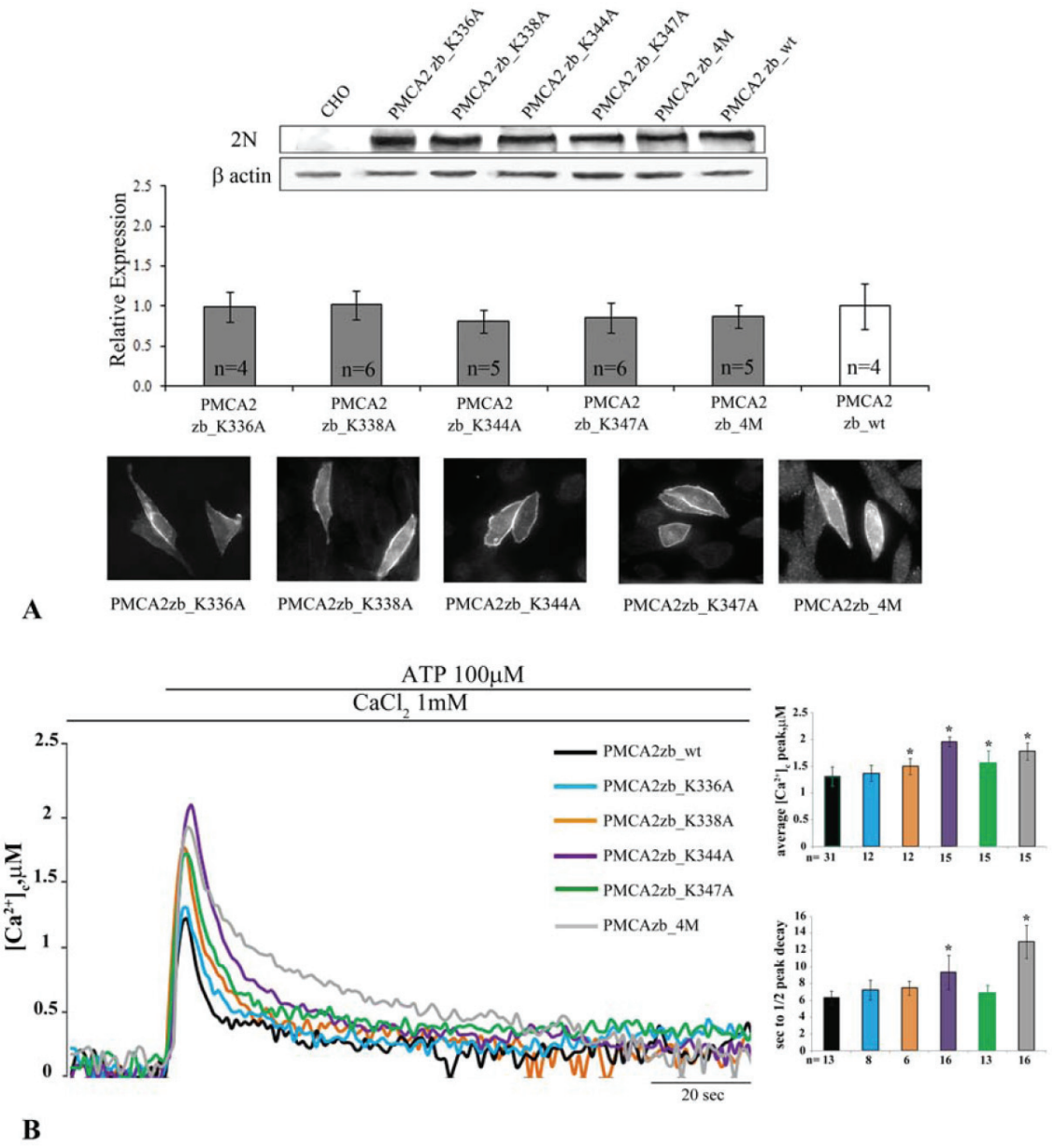


Figure 0.6. Expression and activity of PMCA2zb_del12 and PMCA2wa_del12.
A, Western blotting and densitometric analysis of the Lys mutants of the PMCA2 *z/b* isoform overexpressed in CHO cells. 20 µg of crude membrane proteins from transfected CHO cells, prepared by a freeze and thaw method, were separated by SDS-PAGE as described under "Experimental Procedures" and stained with polyclonal antibody 2N. The control lane corresponds to nontransfected cells (*CHO*). The other lanes correspond to cells transfected with the WT or mutant variants of the PMCA2 pump. The *panel* also shows the immunocytochemistry analysis of the transfected CHO cells. The immunostaining was carried out with the 2N antibody and revealed with the secondary antibody Alexa Fluor 594. ***B***, monitoring of cytosolic $[Ca^{2+}]$ in CHO cells transfected with cytAEQ and co-transfected with cytAEQ and the PMCA2*zb_K336A*, PMCA2*zb_K338A*, PMCA2*zb_K344A*, PMCA2*zb_K347A*, or PMCA2*zb_4M*, alternatively. The *histograms* show the means ± S.D. of $[Ca^{2+}]_c$ peaks and of the half-time decays from the peaks. The traces are representative of at least 12 independent experiments. *, $p < 0.01$ calculated with respect to PMCA2*zb_wt* (CHO cells transfected with wt PMCA2 *z/b* pump).

Generation, Expression, and Activity of E337A or S339A PMCA2 z/b Mutants

The decision to mutate basic residues (lysines) in the 336–347 domain was dictated by the ability of the domain to bind acidic phospholipids. However, the domain also contains a conserved glutamic acid in position 337 (Glu-337) and a serine in position 339 (Ser-339) (see Fig. 8.4). The *in silico* analysis (see below) suggests that these residues could be involved in polar interactions with other portions of the protein. Thus, they were also mutated. Fig. 8.7*A* shows that PMCA2*zb_E337A* and PMCA2 *zb_S339A* were expressed at levels comparable with those of the transfected PMCA*2zb_wt* variant and were correctly delivered to the plasma membrane of the transfected cells.

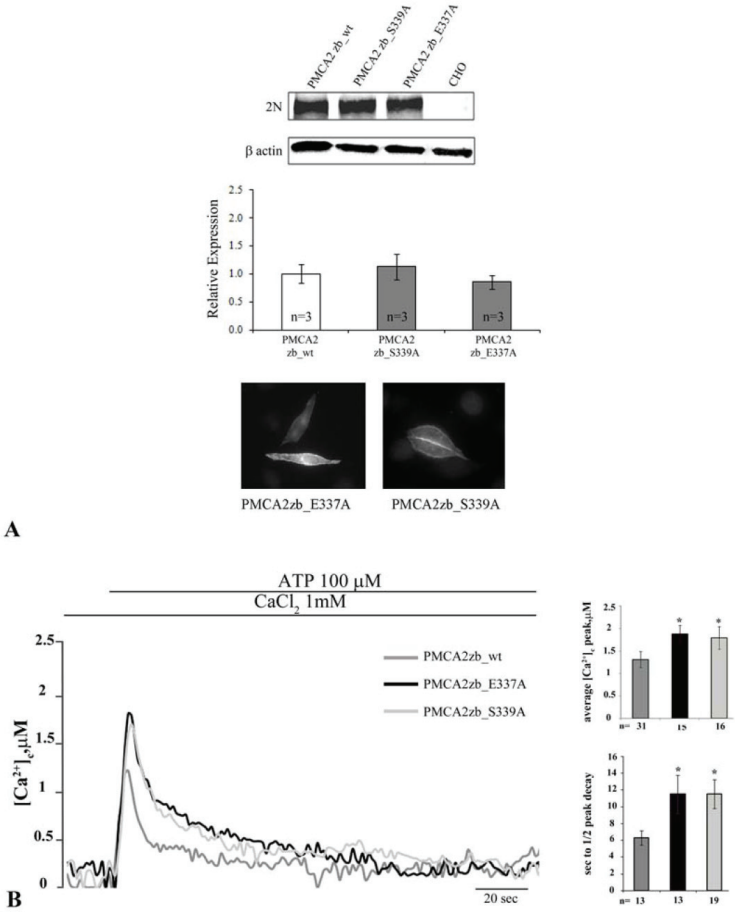


Figure 0.7. Expression, and Activity of E337A or S339A PMCA2 z/b Mutants.
A, Western blotting and densitometric analysis of the Glu and Ser mutants of the PMCA2 *z/b* isoform overexpressed in CHO cells. 20 µg of crude membrane proteins from transfected CHO cells, prepared by a freeze and thaw method, were separated by SDS-PAGE as described under "Experimental Procedures" and stained with polyclonal antibody 2N. The control lane corresponds to nontransfected cells (*CHO*). The other lanes correspond to cells transfected with the WT or mutants variants of the PMCA2 pump. The panel also shows the immunocytochemistry analysis of the transfected CHO cells. The immunostaining was carried out with the 2N antibody and revealed with the secondary antibody Alexa Fluor 594. *B*, monitoring of cytosolic [Ca^{2+}] in CHO cells transfected with cytAEQ and co-transfected with cytAEQ and the PMCA2*zb_E337A* or the PMCA2*zb_S339A*. The *histograms* show the means ± S.D.

of $[Ca^{2+}]_c$ peaks and of the half-time decays from the peaks. The traces are representative of at least 12 independent experiments. *, $p < 0.01$ calculated with respect to PMCA2*zb_wt* (CHO cells transfected with WT PMCA2 *z/b* pump).

The Ca^{2+} measurements showed that the PMCA2*zb_E337A* and the PMCA*zb_S339A* mutants were less efficient than the PMCA*2zb_wt* variant in controlling the peak of the Ca^{2+} transient (1.88 ± 0.19 μM, $n = 15$, for PMCA2*zb_E337A*, 1.79 ± 0.25 μM, $n = 16$, for PMCA2*zb_S339A versus* 1.31 ± 0.17 μM, $n = 31$, for *zb_wt*, $p < 0.01$) (Fig. 8.7*B*). The mutations also severely affected the ability of the pump to restore basal Ca^{2+} levels after cell stimulation, the half-time of the peak decay being 11.52 ± 2.27 s, $n = 13$, in PMCA2*zb_E337A* and 11.26 ± 1.73 s, $n = 19$, in PMCA*zb_S339A*, as compared with 6.31 ± 0.85 s, $n = 13$, in *zb_wt*, $p < 0.01$ (Fig. 8.7*B*).

In Silico Analysis of the Two Phospholipid Binding Domains

Figure 8.8*A* shows a schematic of the PMCA2 in which three of four lysines (Lys-338, Lys-344, and Lys-347) contained in the A_L domain are predicted to be located approximately at the membrane surface, forming a charged bend which, as already mentioned, could easily accommodate a charged phospholipid head. The electrostatic surface potential of PMCA2 (Fig. 8.8*B*) shows that the region surrounding the lysines, and stretching toward insertion site A, is the only positively charged region in contact with the cytoplasmic side of the membrane. Over 30 residues close to insertion site A could not be modeled; thus, the model is only approximate. However, the missing residues are likely to form a mobile flap extruding from the protein structure. Because conformational switches are required for Ca^{2+} transport, it could be reasonably suggested that the three lysines would form a binding pocket for initial phospholipid docking. The model agrees well with the experimental findings on the importance of three of the four lysines, as well as with the effect of the 12-residue deletion. Mutation of all four lysines is likely to slow down phospholipid docking, but the positively charged area surrounding insertion site A could partially compensate for this effect.

The two other mutated residues (Glu-337 and Ser-339), are well conserved in PMCA isoforms and in the PMCA across species (see Fig. 8.4). The model positions glutamic acid between two lysines and exposes it to the protein surface, where it could affect other interactions of the pump. As for the serine, its polar group forms a hydrogen bond

with a glutamine in the α-helix (M3) and with a glutamic acid in the α-helix of domain P (Fig. 8.9). Mutational disruption of hydrogen bonds may have significant structural consequences.

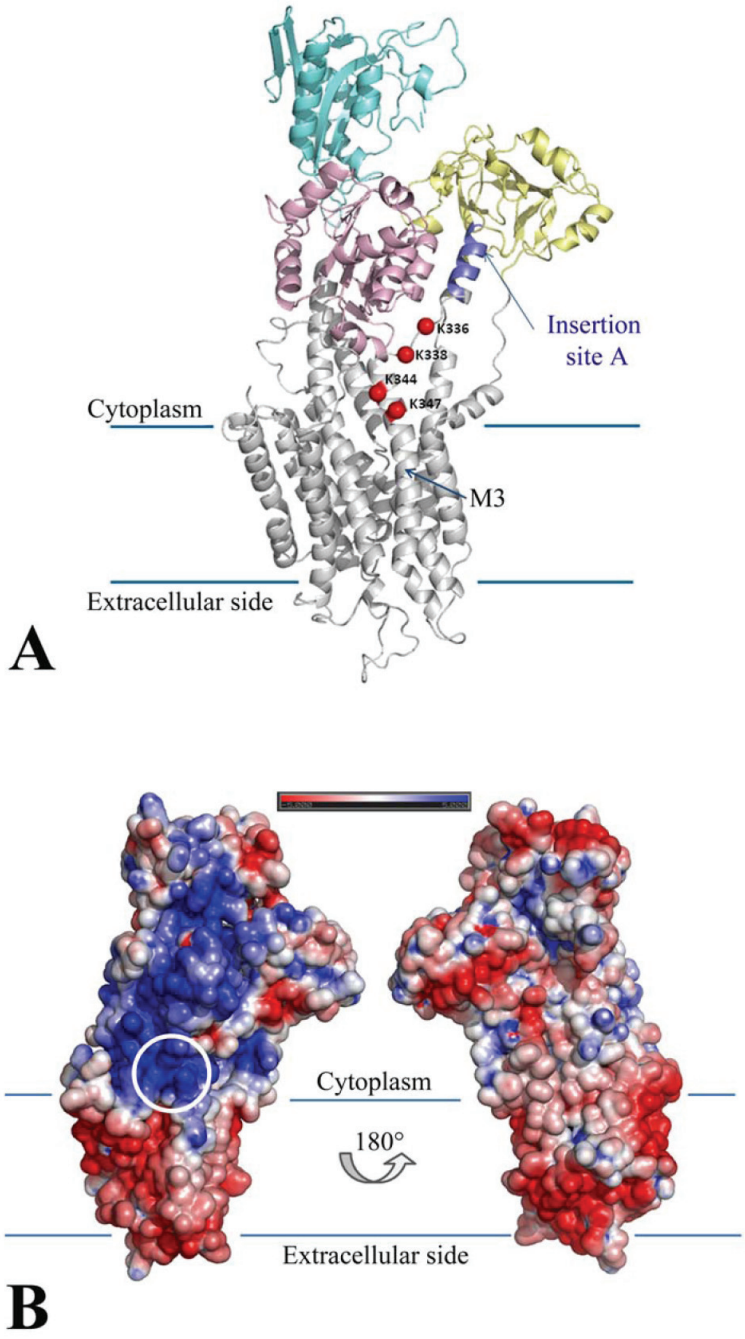


Figure 0.8. PMCA2 model and electrostatic surface.
A, overview of the PMCA2 model, shown in schematics and color-coded for the different canonical domains, with the four mutated lysines highlighted as *red spheres*. The approximate location of the membrane limits are shown with *lines*, and the third transmembrane helix is labeled as *M3*. Note that the C-terminal part of PMCA2 from residue 1088 onward could not be modeled. Insertion site A is *highlighted.* ***B***, electrostatic potential of the PMCA2 accessible surface. The structure is shown in the same orientation as in *A* and rotated around the central axis (*right*). The location of the mutated lysine

residues is *circled*. Note how the area around and between the four lysines and insertion site A is the only PMCA2 region with positive potential in contact with the membrane.

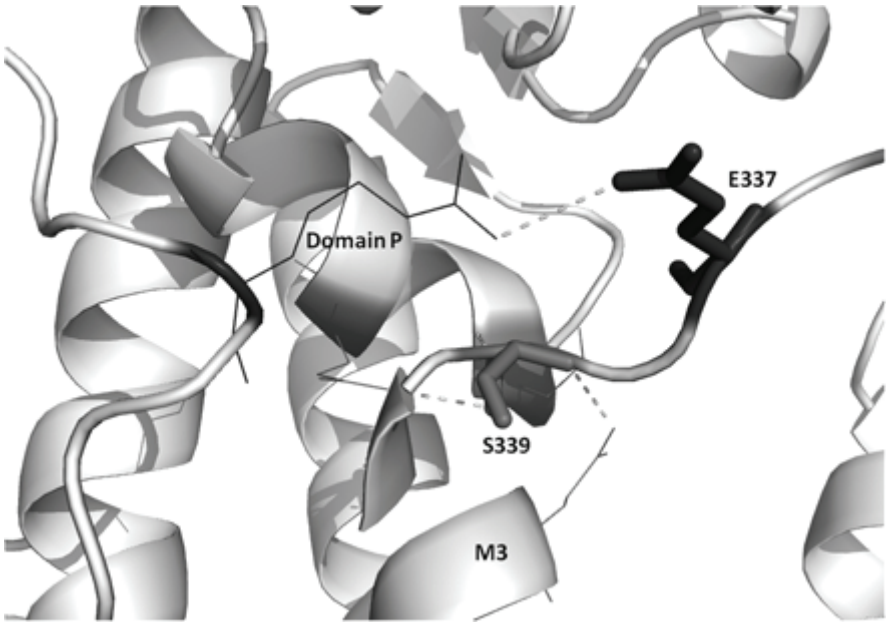


Figure 0.9. Representation of the two residues, Ser-337 and Glu-339.
These are shown as *sticks*, and *dashed lines* indicate interatomic contacts or hydrogen bonds with neighboring residues.

The C-terminal splice variant *w/a* differs from the *w/b* variant by a frameshift mutation affecting the second half of the CaM-binding region. The model generated by the structural analysis (Fig. 8.10*A*) shows that the PMCA2 CaM-binding region (obtained from the recently deposited NMR structure of the PMCA4 CaM-binding region (PDB code 2KNE) could form an amphiphilic α-helix with a distinctive pattern of charged and hydrophobic residues [391]. In the presence of Ca^{2+} ions, CaM folds into a series of α-helices winding around the PMCA2 peptide in a head-to-tail conformation, *i.e.* the N terminus of CaM binds the C terminus of PMCA2. Ca^{2+} could induce a conformational switch through the stabilization of a stretch of negatively charged residues in a turn conformation, yielding the characteristic collapsed structure of CaM. Interestingly, the final conformation has a strongly negative charge and is stabilized through hydrophobic cages between a benzyl ring and a hydrophobic groove at the center of three CaM α-helices (Fig. 8.10*B*). In the model, electrostatic attraction is present, but is not crucial to stabilize the final bound conformation. Given the number of charged residues in CaM,

electrostatic attraction is likely to initiate the folding process of CaM around the PMCA2-binding region. The substitution of two lysine residues in the CaM binding domain of the *w/a* variant (see sequence alignment in Fig. 8.10*A*) could destabilize the CaM interactions necessary to form the hydrophobic cage for proper binding, explaining the poor sensitivity to CaM of the variant.

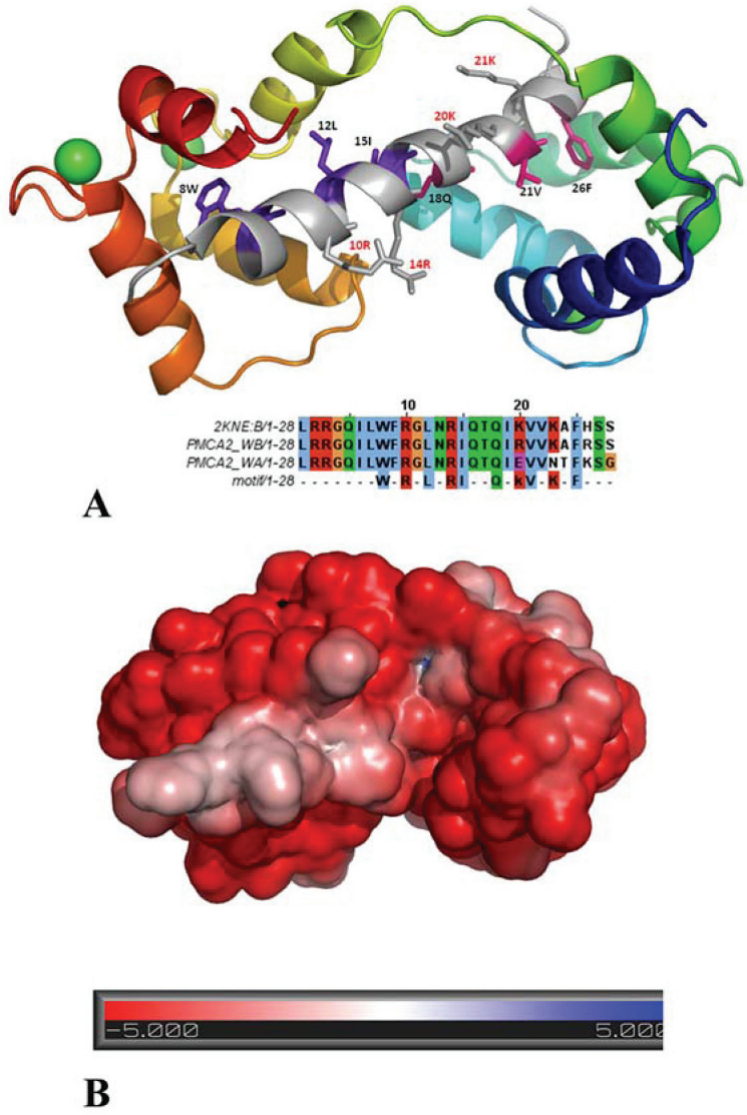


Figure 0.10. Model of CaM-binding region of PMCA2.
A, structural model of the calmodulin-binding region of PMCA2 (*top*) and relative sequence alignment (*bottom*). The amphipathic PMCA2 helix is shown in *gray* at the *center* of the structure, with residues in *purple* and *pink* defining the N- and C-terminal motifs. The calmodulin structure is shown with progressively varying color, from *blue* (N terminus) to *red* (C terminus). Ca^{2+} ions are shown as *green spheres*. The sequence alignment shows the structural template (PMCA4, PDB code 2KNE) together with two PMCA2 variants. The *last line* defines the sequence motif for calmodulin binding. Note how PMCA2 *w/a* lacks two crucial lysine residues for the second motif. ***B***, electrostatic surface of the calmodulin-binding region of PMCA2 with bound CaM in the same orientation as in *A*.

8.6. Discussion

It would be reasonable to expect that the proximity of site A of alternative splicing to the site that binds acidic phospholipids in the A_L domain could influence the sensitivity of PMCAs to acidic phospholipids. The A-site insertion could alter the overall conformation of the second cytosolic loop of the pump. It could thus change the spatial connectivity between the phospholipid binding domain and the sequence further upstream, which is involved in the intramolecular inhibitory interaction with the C-terminal calmodulin binding domain. The finding that the A-site insert is important for the targeting of PMCA pump to the apical membrane [373] underlines its importance in the general properties of the pump. The role of the A_L phospholipid binding domain has always been obscure, particularly in view of the existence of a second phospholipid binding domain in the C-terminal calmodulin binding sequence [375]. One still open question is thus the comparative importance of the two phospholipid binding domains in the regulation of pump activity. Our previous studies on isoform 2 of the PMCA pump had shown differences in the activity of the various A-site splicing variants [390, 392], showing that the *w/a* variant had high basal Ca^{2+} ejection activity but failed to respond rapidly to the sudden arrival of a Ca^{2+} pulse. It had already been reported that both isoforms PMCA2b and -2a have much higher affinity for CaM than the corresponding isoforms of PMCA4, with PMCA2b having the highest affinity. They were both activated at low Ca^{2+}-calmodulin levels and had peculiarly high activity in the absence of activators [386].

The measurements of ATPase activity in microsomal membranes of transfected CHO cells have indicated that the *w/a* variant, as expected, was much less sensitive to CaM than the *z/b* and *w/b* isoforms. However, it was also less sensitive to phosphatidylserine, thus underlining the role of the CaM binding domain in the regulation of pump activity by acidic phospholipids. The finding that the *z/b* and *w/b* isoforms had the same response to phosphatidylserine stimulation had indicated that the splicing insertion upstream of the A_L phospholipid binding domain failed to modify the phospholipid sensitivity of the pump.

The analysis of the A_L 12-amino acid lysine-rich stretch, and the model derived from it, had indicated the importance of the conserved lysines in the stretch in the interaction of

the phospholipid binding domain with the pump microenvironment. The deletion of the 12 amino acids could, for instance, directly affect the structure of M3, which is critical to the sarco/endoplasmic reticulum Ca^{2+}-ATPase pump binding of thapsigargin and could by analogy have special importance to PMCA as well. In the sarco/endoplasmic reticulum Ca^{2+}-ATPase pump, the segment linking M3 to the A domain is essential for the rotation of the latter and for its correct positioning in the active configuration of the catalytic site [393].

By combining the structural information on the four A_L lysines and on the C-terminal CaM-binding region, it could be proposed that the C-terminal domain of the pump that contains the CaM-binding region could anchor Ca^{2+} ions to PMCA; it has indeed been shown that Ca^{2+}-binding sites are present upstream and downstream of the CaM binding domain [394]. Once CaM is bound, the PMCA movements could bring the Ca^{2+} ions closer to the lysine-containing region near insertion site A through electrostatic attraction.

The finding that the deletion of the 12-residue A_L domain completely abolished the activity of the pump in the *w/a* variant, but not in the *z/b* variant in which it only reduced it, indicated that the activity of the PMCA2 *w/a* variant strongly depended on the presence of the A_L(380–391) region and possibly on the acidic phospholipid binding to it. The finding that the *w/a* variant was insensitive to PS in the microsomal membranes assay could mean that its stimulation was already maximal under these conditions, as endogenous acidic phospholipids are present in the membranes and could have saturated the PL binding domain. Further addition of PS could not further stimulate the activity of the *w/a* variant. Evidently, CaM activation is not sufficient to make the *w/a* variant as active as the *z/b* and the *w/b* variant. Thus, the difference between the activities of the *w/a* and *z/b* variants observed in the measurements performed in intact cells could be related to their interaction with acidic phospholipids, as also suggested by the ATPase activity measurements on microsomal membranes. In other words, the *z/b* variant would be more active than the *w/a* variant because of the integrity of its two acidic phospholipid-binding sites. The truncation of the protein induced by the site C splicing drastically affected the ability of the pump to bind activator phospholipids, and the deletion of the 12-residue A_L domain further compromised its activity.

Interestingly, the substitution of all four positively charged residues (lysines) reduced the Ca^{2+} extrusion ability of the pump by about the same extent as the replacement of only Lys-344, suggesting a critical role for Lys-344 in pump activity. However, the mutation of two polar residues (Glu and Ser) in the same region affected the pump activity to about the same extent, suggesting that the disruption of the possible interaction of this region of the pump with the other pump region (or with other proteins) may be as important to pump activity as the impairment of its ability to bind acidic phospholipids.

9. Critical Assessment of Genome Interpretation

The several cases I studied had as major aim to determine the protein structure-function relationship in order to gain insights into genotype-phenotype correlations and to better understand the molecular mechanisms of the related diseases. Recently, science is witnessing a revolution in molecular biology owing to the advances in high-throughput technology. Genome and exome sequencing generate huge amounts of data yielding extensive catalogues of human genetic variations. However, the identification of few causal variants among the extensive background of non-pathogenic polymorphisms remains a major challenge, particularly for rare and common complex diseases [34]. In this context, there is a strong demand to develop efficient and accurate bioinformatics tools for the classification of disease mutations. Currently, several different methods are available for this purpose but the community needs to 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.

The *Critical Assessment of Genome Interpretation* (CAGI, http://www.genomeinterpretation.org/) is a community experiment started in 2011 to assess computational methods predicting the functional impact of genome variations. The organizers provide unpublished genomic data for which they know the associated phenotypes and participating groups have a few months to make predictions. The evaluations, performed by independent assessors, have been made public and discussed at the CAGI meeting in San Francisco on December 2011. In addition to being an opportunity to connect researchers from diverse disciplines, the CAGI experiment aims to identify the critical points in genome interpretation and promising areas of future research.

In this chapter I will describe different applications where computational tools are useful to interpret experimental work or to predict genotype-phenotype correlations. Moreover, I will report the state of the art strategy applied so far for disease-gene

prediction from next-generation sequencing data. These computational approaches have made rapid progress in the last few years as testified by challenges presented in the CAGI competition. I will describe how we addressed them and the results of our predictions are discussed with respect to the experimental evidence and results presented at the CAGI-2011 meeting.

Critical Assessment of Genome Interpretation (CAGI) competitions

The first prototype CAGI experiment was designed in 2010 by Steven Brenner, a computational genomicist at the University of California Berkeley, and John Moult, a computational biologist at the University of Maryland. In 1994 they conceived a similar competition named CASP (Critical assessment of techniques for protein Structure Prediction), which was aimed to improve the ability of researchers to predict the shape of a protein starting from the amino acid sequence. This experiment gave a boost to the development of new approaches for structure prediction, determining tools which are still the best choice in this regard. The goal of the CAGI contest instead is to accelerate the development of software able to predict molecular, cellular, or organismal phenotypic impacts of variations and to process quickly a large amount of genetic data arising from the increased ability of genome sequencing seen in the last decade.

In 2011 the CAGI contest proposed several different challenges which were divided into two main groups, depending on the overall approaches applied, which we will call gene-oriented and phenotype-oriented predictions. Gene-oriented predictions aim to identify the connections between mutations occurring in a specific protein and an observed phenotype. The second group of predictions aimed to identify variants that could be related to a particular human phenotype. In this case we have a large number of single nucleotide variants (SNVs) sparsely located in the genome. The candidate variants to assess are those mapping to genes having a high likelihood of being the cause for the specific phenotype.

Our group, named UniPadova, participated in five competitions aimed at identifying the connections between specific genes and phenotypes. Three of these include the study of mutations on three different proteins: the p53 transcription factor, RAD50 protein, and Nav1.5 channel. In the cases of p53 and Nav1.5 the associated phenotype to predict corresponded with a protein-specific biochemical feature (e.g. tumor suppressor

reactivation, current flow density), while for RAD50 the probability of the variant to occur in individuals with breast cancer had to be predicted.

Two other challenges required a totally different approach. We refer to this predictions as phenotype-oriented because they point to the responsible genes from those indicated by clinical findings or genetic analysis such as genome wide association study (GWAS). In one case, predictors had to distinguish between exomes of Crohn's disease patients and healthy individuals. The Personal Genome Project (PGP) challenge was instead directed at predicting the probability of an individual having a specific human phenotype or trait from a list of forty binary and numerical traits starting from exome sequence data. In the next paragraphs I will describe each CAGI challenge separately, giving a view of how each problem has been addressed and the relative state of the art.

9.1. Single amino-acid changes in the human p53 core domain that can restore activity of inactive p53 found in human cancers

Inactivation of the p53 gene is the most common genetic cause of human cancers [395-396]. In most cases, this inactivation is the direct result of mutations mapping in the DNA core domain of the protein. Restoring p53 activity is possible, as demonstrated in vivo, through intragenic second-site suppressor mutations. The group of Rick Lathrop at University of California Irvine conducted a functional experiment to discover "cancer rescue" mutants. Structural analysis of known rescue mutants identified regions in p53 where perturbations may cause p53 reactivation and highlighted that cancer rescue mutations may also influence protein-DNA interactions or protein stability without necessarily inducing major structural disruptions [397].

For the p53 protein, CAGI participants were called to predict "cancer rescue mutants" on four p53 cancer mutations mapping on the DNA binding domain. The dataset for p53 provided by the CAGI organizers contains 14,668 putative rescue mutants. The experimental assay used to test the p53 function was applied for all possible rescue

mutants of the core domain in the presence of the four cancer mutations: R248Q, R282W, Y220C, and M237I.

The DNA binding domain of p53

The DNA binding domain adopts a defined conformation for which several crystal structures have been reported both in complex with DNA and regulatory proteins, or in a free state. The most recent structures show how p53 tetramers recognize DNA [398-400]. The immunoglobulin-like fold serves as a scaffold for the DNA-binding surface which is formed by two major loops, L2 and L3, and a loop-sheet-helix motif (loop L1, β-strands S2 and S2', C terminal helix H2). A zinc ion stabilizes the position of the two large loops. The DNA molecule makes contact through its major groove at the loop-sheet-helix motif, while the DNA minor groove interacts with residues located at loop L3. This loop, together with the helix H1 in the L2 loop, is also involved in core-domain dimerization. In contrast to previously studied p53-DNA complexes, Petty et colleagues [401] demonstrated that p53 binding to specific DNA sequence causes a conformational switch in loop L1, which alters the kinetic properties of p53 DNA binding. Mutations that facilitate the conformational switch of loop L1 thus have reduced levels of transcriptional activity compared to wild type [401]. To demonstrate this, the authors expressed p53 polypeptides containing both the DNA binding and oligomerization domains, forming stable p53-DNA complexes in solution. The resulting structure seems to be less affected by crystal packing interactions.

Solution structure of the p53 core domain by NMR revealed how loop L1, together with the S7-S8 loop, are the protein regions with high structural flexibility. This provided an hypothesis for the structural basis of the relative instability of p53 [402]. The intrinsic instability of p53 seems to be an evolutionary advantage, since it confers structural plasticity that facilitates the exploitation of several functions involving p53. Biophysical characterization of p53 was quite difficult due to its tendency to melt at temperatures of less than 37°C. Thus the request of more stable p53 proteins guided the development of the stable p53 mutant (T-p53C) containing the point mutations M133L, V203A, N239Y, and N268D. These mutations stabilize the protein core by 2.6 Kcal/mol and provide a more rigid structural framework on which structural effect of cancer mutations can be studied [403].

Effects of common p53 cancer mutants

Bullock and colleagues [404] demonstrated that p53 cancer mutants can be divided in different classes on the basis of their location in the core domain. Mutations affect either stability or DNA binding properties of p53. Effects on stability were found for all tested mutations mapping on the β-sandwich domain (e.g. V143A, F270L, Y220C). Destabilizing effects were observed also for structural mutations altering the DNA binding surface or the zinc binding site. These variants showed a reduced or absent ability to bind DNA. Examples of these mutations are those altering the L3 loop (e.g. R249S, G245S), mutations disrupting the zinc binding region (e.g. R175H), mutations located in the loop-sheet-helix motif (e.g. R282W, H168R), and mutations in loop L1 (e.g. T123A). Other mutants have been classified as DNA-contact mutations since they inactivate p53 replacing residues that form direct contact with DNA. The structures of R273H and R273C mutants maintain the overall topology of the DNA-binding surface, even if they cause loss of DNA contacts [405-406]. Other amino acid substitutions introduce a large hydrophobic side chain that prevent DNA binding by steric clashes (e.g. S241F, R248W, and C277F) [397].

Cancer rescue mutants

The deleterious effect of some cancer mutations can be restored by intragenic second site suppressor mutations. This is of particular interest for p53, because the understanding of molecular mechanisms by which the activity can be restored provides insights for the development of therapeutic anticancer strategies. Recently, Baronio and colleagues [407] used an all codon-scanning strategy to systematically produce all possible single-codon mutations within a defined region of p53, and by using a genetic approach in yeast and mammalian cells, identified diverse second site suppressor mutations. The p53 activity was analyzed using a yeast-based p53 activity assay, where the yeast cells were engineered in a way that they require active p53 for URA3A gene expression. This gene is involved in uracil synthesis, thus cell growth in medium lacking uracil is proportional to p53 activity [408]. This study confirmed that different second site suppressor mutations restore the activity of the protein using different mechanisms, suggesting that different regions in the protein correspond to distinct mechanisms of reactivation [407]. The available structures of p53 mutants provide a

detailed understanding of the structural basis for the role of several mutations in rescuing cancer mutants. It is possible to distinguish between specific and global rescue mutants. A prime example of global rescue mutants are the N239Y and N268D mutations which cause increased stability of the protein without altering its function. Usually, these mutations can reactivate a whole subset of destabilizing mutants [405]. The systematic search of rescue mutants by Danziger and colleagues [408] resulted in the identification of a global suppressor motif involving core domain residues 235, 239 and 240. Other oncogenic mutations are reactivated by specific rescues, as these mutants usually cause a distinct structural change in functional regions of the protein. There are few examples for specific molecular mechanisms restoring the protein activity. One of these is S240R, a specific rescue for the DNA contact mutation R273H. Arg240 indeed compensates the loss of Arg273 creating a novel DNA contact. The other well known example is represented by the H168R/R249S rescue pair, mapping at the DNA binding surface. In this case replacement of Arg249 with Arg168 stabilizes the conformation of loop L3 which is essential for positioning Arg248 in direct contact with DNA [406].

Computational approaches to predict p53 rescue mutants

The *in vitro* testing for p53 functionality of all possible rescue mutants is difficult to do due to time and expense. For each mutant of the DNA binding domain, we should test more than 4,000 putative rescue mutations. Furthermore, predicting the effects of amino acid changes is a difficult problem due to the marginal instability of p53. For many p53 cancer mutants the identification of underlying structural changes affecting folding or protein-DNA contacts remains to be discovered [409]. Recently, Danziger and colleagues [410] applied a method, named Most Informative Positive (MIP) active learning, to discover mutations that reactivate p53 cancer mutants. Active learning using modeled structural features was developed in concert with experiments in order to reduce the number of tests that need to be performed to built an accurate classifier. This method was also used to select gene regions suitable for systematic combinatorial mutagenesis. The computational classifier is trained with a subgroup of examples for which the activity was experimentally determined. The classifier then predicts the set of mutants that should be labeled in order to improve classifier accuracy. These mutants

are then tested by activity assay and added to the set of p53 mutants with known function, and the cycle repeats [410].

9.1.1. Method

In this challenge we used residue interaction networks to infer predictions. This approach has been developed in our laboratory especially for the prediction of cancer rescue mutations. This was not the widely studied situation where we want to predict the pathogenicity of a mutation. Rather, we have to predict the ability of a mutation to restore the alteration caused by another variant. This task is very difficult and at the moment no computational method has been developed for this purpose. We therefore tried to rank the mutations with the aim of reducing the number of putative rescue mutations by an order of magnitude. First of all I performed a structural analysis of the p53 core domain in order to identify relevant residues for p53 function and to predict structural effects of the four cancer mutations in the data. This allows to hypothesize the molecular mechanism which could reactivate the protein function. Information arising from this study was integrated in the prediction process.

Rescue mutants prediction using residue interaction network

To analyze structural effects of amino acid substitutions, our laboratory developed an approach that uses graph theory to represent proteins. The residue interaction network is an interaction graph, where nodes are residues with given properties and edges are the weighted relationships among these nodes. We used RING (see first chapter) [200] to build RINs for wild type p53 and putative rescue mutants of the four p53 cancer rescue mutations (R248Q, R282W, Y220C, and M237I). The mutant models were built using FOLDX [411].

The method considers both local and long range interactions, defining different level of interactions. Nodes and interactions in the networks were annotated with a weight derived from data provided by RING. In this case we only used residue conservation and the type of chemical bond between amino acids (e.g. hydrogen bond, van-der-Waals interaction, ionic bond). It is interesting to note that *a priori* information can be introduced in the network simply by adding or removing node or edge weights (Fig.

9.1). We exploited this feature by adding scores for certain nodes known to represent crucial structural or functional residues. Finally, the measure of relevant nodes was calculated by the page rank algorithm, which relies on a hidden Markov model.

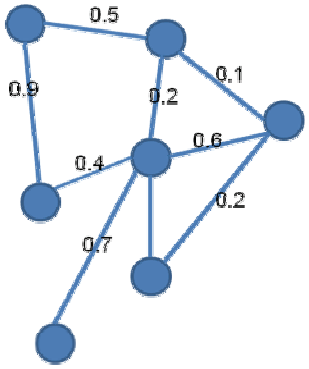


Figure 0.1. Schematic representation of a residue interaction network.
Nodes and edges have a weight.

From page rank we obtain a value for each amino acid, which is used to build a vector describing the impact of each node on the overall graph. Comparison of the wild-type network vector to that of the mutant network is performed by Euclidean distance, providing an index of divergence for the two networks. Clearly, the more similar the networks, the smaller the final difference will be. The submission resulting from this approach was called metric. Another approach was adopted to classify the mutant proteins against a set of known rescue mutants using a k-nearest neighbour algorithm (K-NN) (Fig. 9.2).

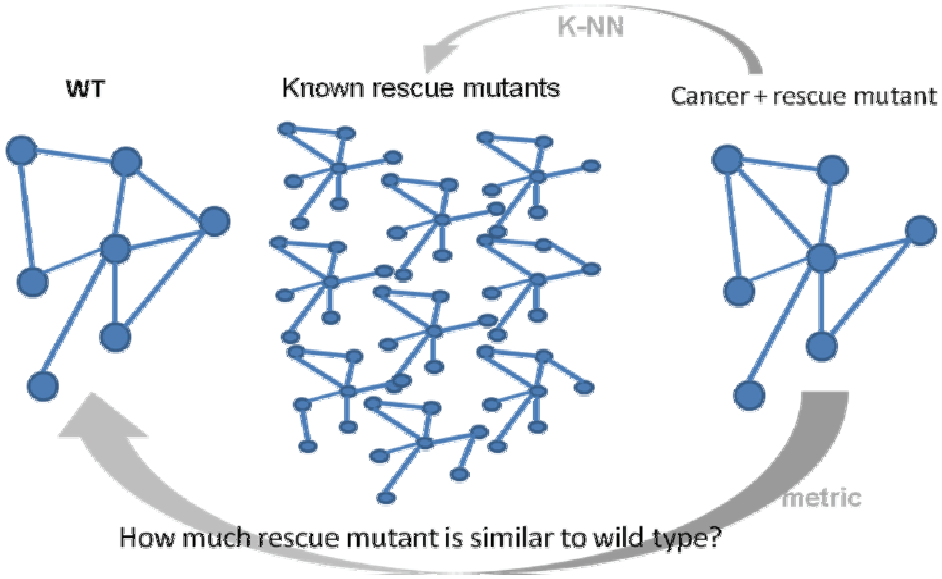


Figure 0.2. Scheme of the two methods used to identify p53 rescue mutants.

Training set

Overall, the training dataset contains 16,772 mutants. For each of these the activity of the p53 protein is known. The experimental assay used a previously described yeast system where cell growth is proportional to p53 activity [408]. The training dataset represents an exhaustive single-point mutagenesis experiment of the entire core domain of p53 for the following p53 cancer mutations: R175H, R273H, and G245S. Additionally, regional saturation mutagenesis of the following p53 cancer mutations are included: H179R, P151S, R280T, P278L, R248L, R273L, R249S, P152L, and R158L. While these mutations comprise most of the dataset, several hundred examples for other p53 cancer mutants are included.

9.1.2. Results and discussion

Structural effects of four tested cancer mutations

The analysis of four cancer mutants was made by visual inspection of the mutant model and by residue interaction network. The aim of this analysis was to identify the possible mechanisms which could restore activity of these mutations. Modeled mutant proteins were built from the wild-type p53 core domain structure (PDB:1TSR). Each mutant model was used as input to build the corresponding residue interaction network. In order to identify changes on residue interactions, the mutant and wild-type RINs were compared. This helped in defining the molecular mechanisms altering p53 function for each mutant (Fig. 9.3). For two of these mutants, R282W and Y220C, a crystal structure has been determined. Here, we could test the accuracy of RIN analysis on the identification of structural changes.

The crystal structure of the **R282W** mutation has been obtained by Joerger and colleagues [405], introducing the mutation into a stabilized variant of p53 core domain (T-p53C). The structure revealed the role of Arg282 in maintaining the conformation of the loop-sheet-helix motif which makes contact with DNA. Comparing the RIN of R282W with that of the wild-type we identified the loss of several strong interactions with residues Thr125, Tyr126, and Ser127 in strand S2, Phe134 in strand S2', and Glu286 in helix H2. All of these interactions are important for the correct packing of the loop-sheet-helix motif.

The same authors determined the crystal structure of **Y220C**, the most common mutation mapping far from the DNA binding surface creating a solvent accessibility cleft in the β-sandwich [405]. Tyr220 maps at the beginning of the loop connecting β-strands S7 and S8 and. Even if it maintains the H-bond with Thr155, its substitution causes the loss of several favorable van-der-Waals interactions with residues Val147 and Thr150 in the loop connecting β-strands S3 and S4, and Thr230 Pro223 in the loop S7-S8. These findings suggest that while the overall topology of the core domain is maintained, this mutation has a destabilizing effect resulting from the loss of hydrophobic interactions. The change in the surface was also visible with the visual inspection of the mutant model surface.

The **R248Q** mutation we had to predict rescue mutations for is classified as a DNA-contact mutant. This class of mutations inactivate p53 replacing residues that form direct contact with DNA without affecting positioning of the neighboring residues. Here, the RIN analysis indeed fails to identify any changes in interactions.

Finally, the other mutant **M237I** maps at the beginning of loop L3 near the DNA binding surface. Even if the RIN analysis did not identify any changes in the residue interaction network, it helped to define the possible effects of this mutations for which a crystal structure is not available. In this case, we observed that the residue Met237, close to Cys238 in L3, forms an interaction with Asn239. It is then possible to hypothesize its role on the correct positioning of loops L3 and L2 maintained by zinc ion coordination. The introduction of Isoleucine at that position, rather than loss of interactions should cause a structural distortion that may directly interfere with zinc binding. The other variant known to have similar effects is R175H. Arg175 protrudes between loops L2 and L3 forming hydrogen bonds with Pro191 and Met237 and a salt bridge with Asp184. Several mutants of this residue position have been experimentally investigated and show different functional effects. It seems that the introduction of bulky residues has a strong impact on p53 function, while R175A, R175C, and R175L mutants have reduced or similar activity to that of the wild-type [397].

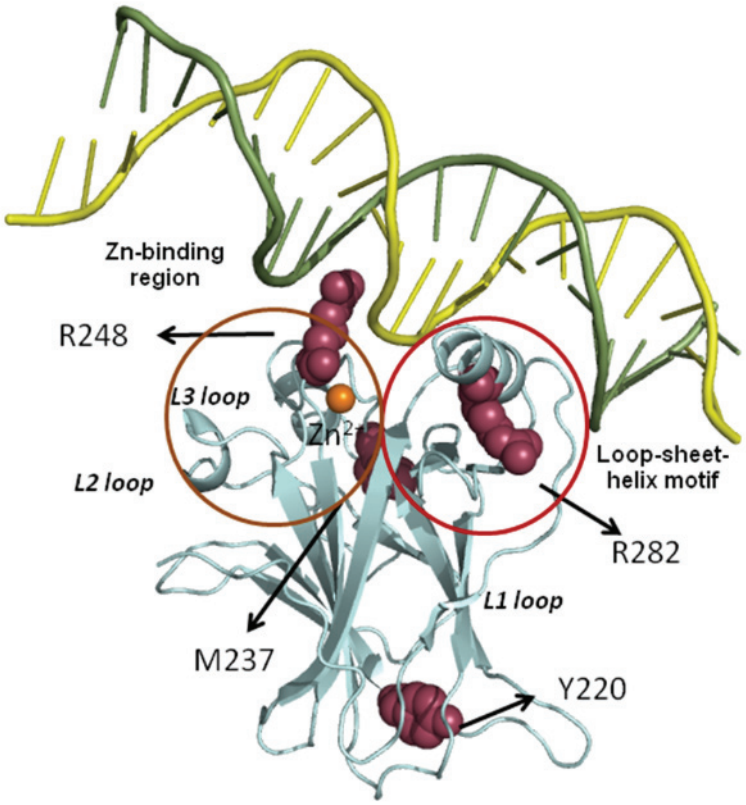


Figure 0.3. The four p53 cancer mutations.
The mutations for which we have to predict rescue mutants are mapped on the crystal structure of the DNA binding domain of p53 protein (PDB code: 1TSR).

Prediction results

For this challenge we submitted two different predictions using different scoring functions, metric and K-NN, to calculate rescue mutant probability. Furthermore, we submitted a random prediction with the 100 side die. For the two predictions, a threshold was chosen in order to obtain 560 and 1287 rescue mutants, which represent ~4% and ~8% of the total putative rescue mutants, respectively. The results for each cancer mutant are summarized in Figure 9.4, which represents only an extrapolation of the data. In these graphs I want to indicate the sum of the probabilities assigned for each mutant at a specific position. It is possible to note that probabilities are not sparsely distributed, instead there is a common tendency of some sequence regions to have high or low probability to contain rescue mutants. This result was also obtained by the MIP active learning approach adopted by Danziger and colleagues, which identified regions of p53 sequence containing the best number of known rescue mutants [410].

Experimental results

After the prediction season was over, the organizers released experimentally determined rescue mutants for the four cancer mutants: M237I is rescued by L137R, R175A, R175P, R175S, R175V, and R175T; R282W is rescued by F212G; Y220C is rescued by L137R. R248Q does not have any rescue mutants. For these positions, our prediction method gave low probabilities to be rescue mutants, with values ranging from 0,3 to 0,39.

Observing the graphs in Figure 9.4, it is possible to note that for the R248Q cancer mutation we obtained generally low probability values to have rescue mutations. In the other cases, the real rescue mutants mapped in regions that seem to have a higher probability to contain rescue mutants for one of the two methods used. For M237I the rescue mutants located in two regions of high restoring probability calculated by the K-NN approach, while for the metric calculation they are positioned in regions with lower values. The same has been also observed for the Y220C mutation. This can be interpreted on the basis of the difference in the two approaches we adopted. The metric calculation finds similarity with wild-type proteins, while K-NN calculates the similarity of the rescue mutants to another which was experimentally determined. These observations suggest that to improve the performance of the method, it is important to evaluate each class of cancer mutants separately with a specific class of known rescue mutations. As I described before, the four cancer mutants for which we have to predict rescue mutants indeed determine different structural impacts on the p53 core domain.

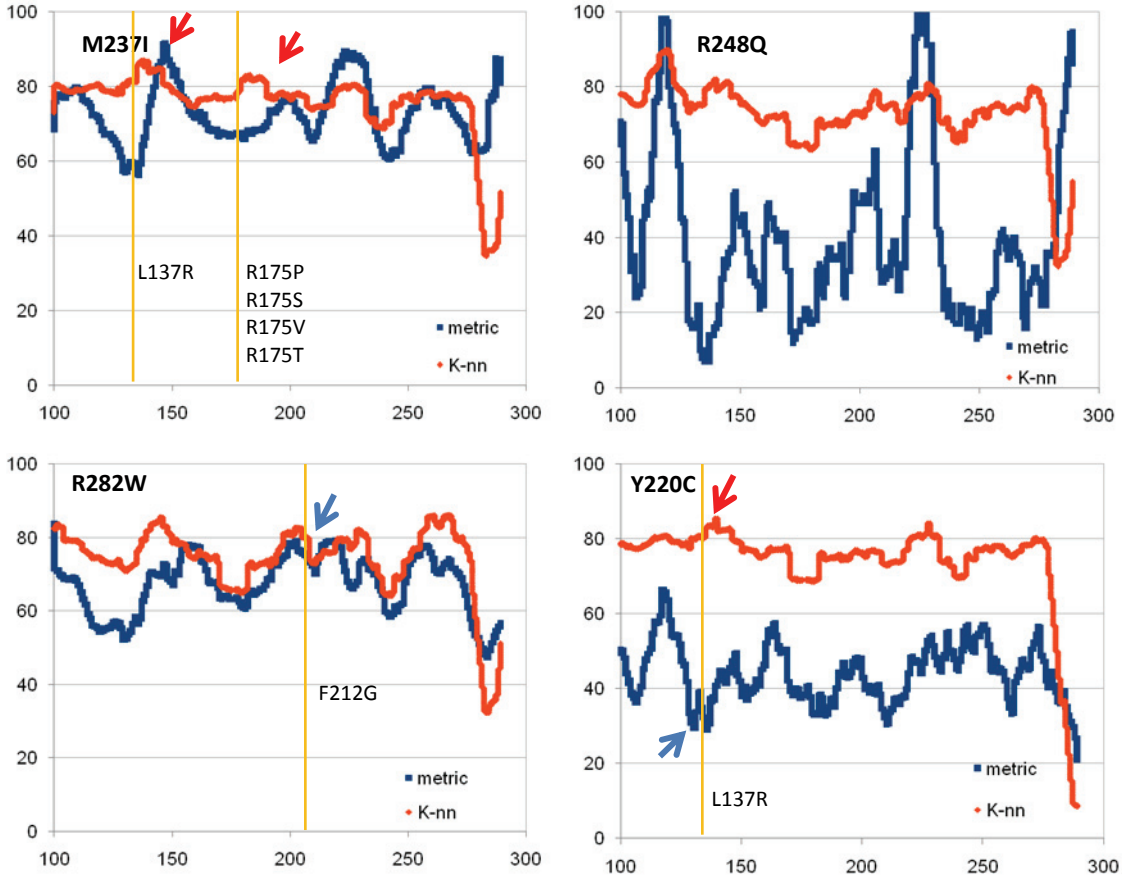


Figure 0.4. Probability to contain rescue mutants for p53 sequence.
The value reported in the y-axis is the sum of probabilities and those in the x-axis the position in the p53 amino acid sequence. For each position of the sequence, several rescue mutants corresponding to all possible amino acid substitution have been obtained and tested. The orange bar indicates the position of experimentally determine rescue mutants.

Comparison with other participating groups

Five groups participated in this competition. Some groups used methods previously developed to evaluate the pathogenicity of missense mutations. Others used information derived from multiple sequence alignment, and the method which had the best prediction used structure features to calculate free energy changes caused by mutations. However, also for this group the identified rescue mutants were distant from the mutant with higher restoring probability, being near the 1,000th position in the list. However, these results confirm that stability is an important feature in the molecular mechanisms involved in p53 mutant reactivation, but also that this challenge is very hard to solve. It is interesting to note that the best group reported to avoid conserved positions from the evaluation, considering the great impact on structure and function due to mutations at

these positions. This was also considered in our method where Arg175 was weighed to have a low probability to restore p53 activity given its role in DNA contacts and stability of the protein.

9.2. RAD50 variants in breast cancer patients and controls

Mutations in the RAD50 gene confer increased risk of breast cancer [412-413]. RAD50 is a DNA repair protein of 1,312 amino acids, which in complex with Mre11 protein forms an ATP dependent molecular clamp in DNA double-strand break repair. This evolutionary conserved complex has a relevant role in several processes emerging from DNA breaks including meiotic recombination, non-homologous end joining, telomere maintenance, and DNA damage checkpoint activation. Defects on these processes may lead to an accumulation of mutations and increased instability of the genome which are conditions predisposing to cancer development [414].

This CAGI challenge was to predict the probability of RAD50 variants to occur in a individual with breast cancer. RAD50 was sequenced in about 1,400 breast cancer cases and 1,200 healthy controls, allowing the identification of 69 variants including 5 truncating mutations, 33 variants leading to amino acid changes, 15 silent mutations, and 14 variants mapping to intronic regions of the gene. While some of these are novel variants, many other variants have been previously identified and are present in the specific database for single nucleotide polymorphisms.

RAD50 proteins

The protein contains two ABC ATPase or P-loop hydrolase domains at the N and C terminal ends, which have 50% sequence identity with yeast Rad50. Structural information for proteins of this family can be derived from partial RAD50 structures in several organisms. The most divergent central domain forms an extended coiled coil structure containing a Zinc-hook motif. RAD50 forms a dimer, where the N- and C-termini assemble into a single ABC domain linked to an antiparallel coiled coil. The coiled coil kinks back in the middle exposing the Zinc-hook motif (Fig. 9.5). In all

organisms studied, RAD50 forms a heterotetramer with Mre11. The architecture of this complex looks like a clamp with bipolar structure with two long tails and a globular head formed by the RAD50 ABC ATPase domains interacting with the Mre11 dimer. The two tails are locked together by the Zinc-hook domains [415]. Crystal structures of the globular head have been determined in several organisms, but the coiled coil structure was revealed only by atomic force microscopy [416]. The relevance of this long coiled coil structure was demonstrated both for the formation of the Mre11 complex and the sister chromatid interactions ([415, 417]. In contrast to its hydrophilic nature, the coiled coil segment contains a conserved hydrophobic surface path corresponding to the binding site of the Mre11 protein [415].

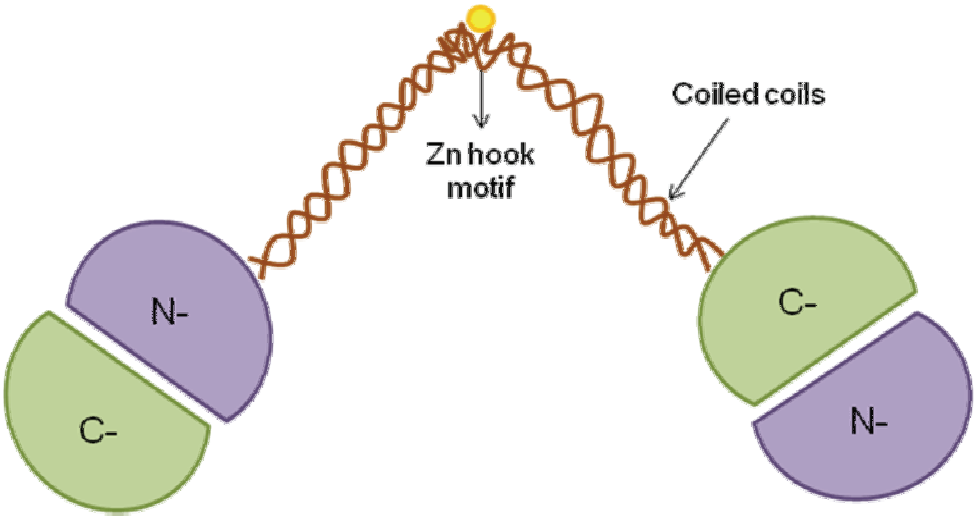


Figure 0.5. Rad50 antiparallel homodimer.
The N- and C-termini assemble into a single ABC domain linked to an antiparallel coiled coil.

Case-control mutation screening

Recently, two studies reported the assessment of variants identified in candidate genes conferring an intermediate-risk of breast cancer, CHEK2 and ATM [35, 418]. The developed methods consist in ranking of the variants, usually missense substitutions, in a scale from least to most likely to be evolutionarily deleterious. The variants were scored using common mutation prediction methods such as SIFT, Polyphen and AlignGVGD. Finally, the frequency distribution of different classes of missense substitutions were compared in the cases versus controls in order to identify specific

trends. They found that truncating variants in ATM and CHECK2 have respectively a modest or moderate probability to confer a risk of cancer. Furthermore, the risk was also associated to rare, evolutionarily unlikely CHEK2 missense substitutions [35]. Strong or moderate statistical evidence was indeed found for rare missense substitutions in the ATM and CHEK2 genes, respectively. In particular, a stronge evidence was observed when the analysis of missense substitutions was focused on ATM key functional domains [418].

9.2.1. Method

We assumed that the probability of variants occurring in a case individual is proportional to its degree of deleteriousness. Thus, we classified the variants on the basis of their functional effects using different approaches on the basis of the mutation type: truncating, missense substitution, silent variant, and intronic variant.

Many computational methods are available that predict the pathogenicity of missense variants, but the performance of these is variable. Thus, we decide to use a combination of these methods to see the agreement between them and exclude ambiguous predictions. Since we had to make several predictions, we chose methods for which the program was available for local installation. This makes the job submissions fast and automatable. We predicted the functional impact of the 35 missense substitutions by a combination of five computational methods using only sequence information: PHDSNP (using only sequence information), PHD-SNP (using MSA information) [181], I-Mutant [190], Mupro [187], and ParePro [419]. After getting predictions from the different methods, we classified variants in two main classes: deleterious and neutral. The probability of a mutation to be associated with breast cancer was calculated using different scoring functions, majority vote or weighed sum. Each method gives a specific value for prediction, which were translated to -1 and +1 if the final prediction was deleterious or neutral, respectively. A simple sum of these indices indicates the overall tendency of a variant to be predicted as deleterious or not by different methods. This scoring function was named *majority vote*. The negative majority vote indicates that a variant is predicted as deleterious. In some cases we weighed the index by a factor depending on the performance of the methods in the data training set. This method was called *weighed sum*. Another prediction was made arbitrarily considering information

derived from structural/functional analysis of the RAD50 protein: the position of the variants in a specific domain, degree of conservation, solvent accessibility value, presence in the dbSNP database, and frequency in the general population.
Silent substitutions and intronic variants were analyzed by two computational methods for the prediction of splice sites: BDGP (http://www.fruitfly.org/seq_tools/splice.html) and NetGene2 [420] (http://www.cbs.dtu.dk/services/NetGene2/). The 15 silent substitutions were also investigated by ESEFinder [421] (http://rulai.cshl.edu/cgi-bin/tools/ESE3/) in order to predict the possible alteration of splicing due to alteration of exonic splicing enhancers (ESEs).

9.2.2. Results and Discussion

Structural analysis of human RAD50

The domain organization of human RAD50 protein was defined using PFAM. An alignment between the sequence of human RAD50 and its orthologs allowed identification of the domain boundaries. The sequence conservation for the two P-loop domains indicates two residues ranges: residues 1-220 and 1109-1300 (Fig. 9.6). From PFAM the Zinc-hook domain mapped to residue 635-734 with the conserved CxxC motif from residue 681 to 684. In order to understand the structural relevance of some sequence regions, the human RAD50 sequence was also aligned with the sequence of *Escherichia coli* Rad50 for which a crystal structure has been determined (PDBcode 3QG5). They share 24% sequence identity with conservation only in the two P-loop domains from residue 1 to 178 and 1065 to 1292.

Prediction results

The sequence alignment of RAD50 and their orthologs was used as input for Consurf analysis, which allowed to classify variants based on their conservation score. Based on the MSA, the server also uses a neural network prediction scheme [422] to annotates residues as exposed (e) or buried (b) (Fig. 9.6). In order to collect frequency data for each variants, the dbSNP database was interrogated. Among the 69 variants, 24 were present in this database and, for 15 of these, a frequency in the general population was available. Among the 35 missense mutations, five occur at the N-terminal P-loop

hydrolase domain, three at the C- terminal P-loop hydrolase domain, three at the Zinc-hook domain, and only one at the C-terminal tail. The other 23 variants map on the coiled coil region. Missense variants mapping on the structured functional domains have been considered to yield a major impact on the protein function and are thus thought likely to be more frequent in case individuals. Seven missense substitutions mapping to the coiled coil region have been predicted as pathogenic for their unexpected high degree of conservation and for the biochemical properties of the substituted residue. The amphiphilic nature of the coiled coil structure allowed to observe a conserved pattern of hydrophobic and hydrophilic residues. If the substitution changes this pattern, the mutation was therefore considered to alter the proper coiled coil structure important for correct complex formation and chromatid interaction.

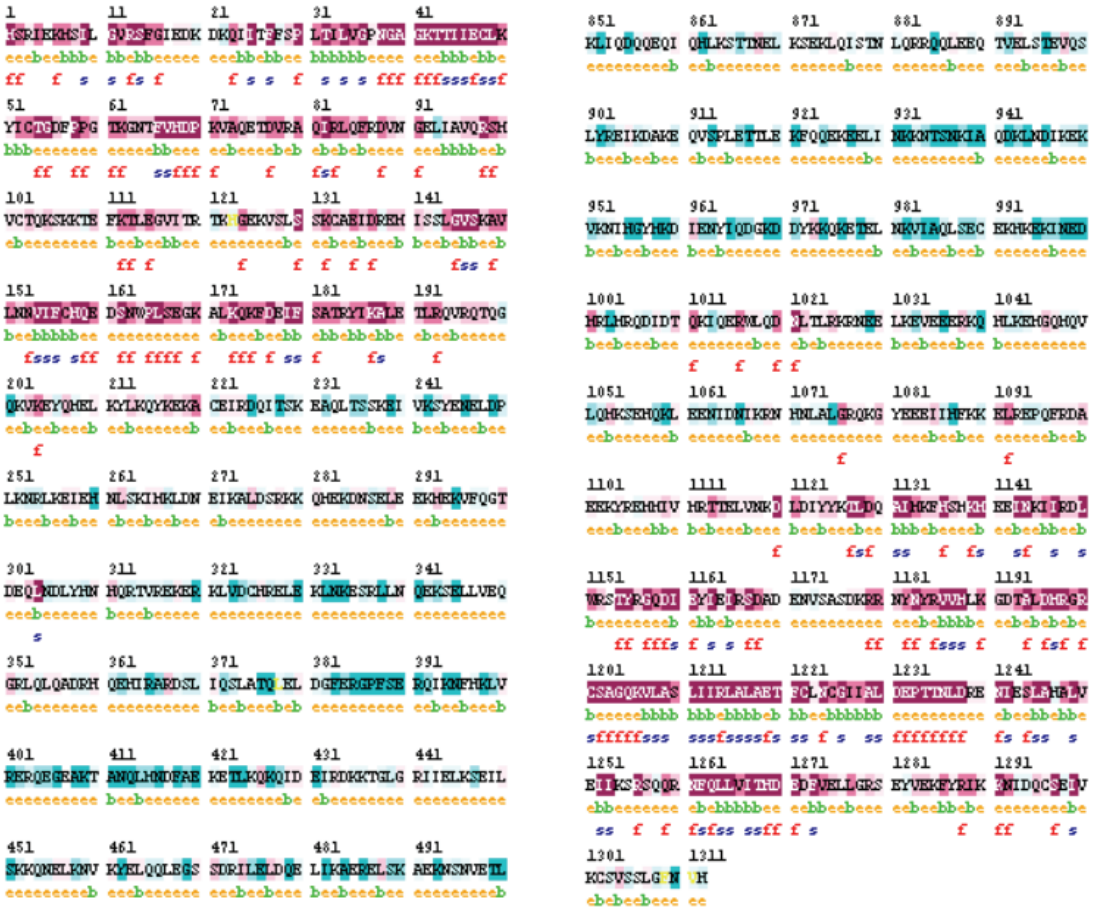


Figure 0.6. Consurf results for the RAD50 sequence.

Among silent variants, three were predicted to alter the pattern of splicing factors binding sites in the coding sequence. One of the three potentially pathogenic variants

was a novel alteration. Only one of the intronic variants was reported in dbSNP, but the frequency in the normal population was not calculated. At least one of the two methods predicted an alteration on splice site recognition for 9 of the 14 intronic variants. Most of the potentially deleterious nucleotide substitutions occur at the conserved positions of the splice site consensus sequence. For three of these the probability of occurrence in case individuals reaches the value of nonsense or frame shift mutations, since splicing alteration may lead the production of truncating proteins. We were not able to predict any protective variant since known protective RAD50 variants were not available for evaluation. For this purpose we used the 100 sided die as the only prediction method.

Comparison with other participant groups

The CAGI organizers provided results for each RAD50 gene variant on the basis of their frequency calculated in a population of 1,400 breast cancer cases and 1,200 healthy controls. However, the frequencies for many variants were very low, often with only one individual carrying the mutation. Figure 9.7 reports the assessment of the participating groups considering three different groups of variants: all rare variants, only missense variants, and only missense substitutions mapping on the protein functional domains. The best prediction has been obtained by the group using a consensus of prediction methods (PON-P). The best prediction from our group (Expert+4 predictors in Fig. 9.7) combines the four prediction methods with the manual prediction based on the structural functional analysis of the RAD50 protein. However, for truncating variants pathogenicity prediction remains difficult. These results highlighted that the most damaging mutations map on the P-loop hydrolase and Zinc-hook domains. About half of the groups predicted well the effect of these mutations.

	Logistic Regression Likelihood ratio test		Receiver Operating Characteristic analysis ¥	
	P-value: Crude	Adjusted †	ROC area	95% CI
Late submission	0.030	0.041	0.54	0.51-0.58
PON-P, del unreliable 0.95	0.146	0.070	0.51	0.43-0.59
PON-P all	0.065	0.115	0.53	0.45-0.61
SNAP	0.179	0.173	0.54	0.47-0.62
PON-P, del unreliable 0.99	0.328	0.228	0.48	0.41-0.56
Naïve Bayes, UCSC multiZ(3)	0.110	0.288	0.56	0.48-0.64
Naïve Bayes, UCSC multiZ(4)	0.111	0.318	0.56	0.47-0.63
SNPs&GO	0.248	0.437	0.55	0.46-0.63
Expert + 4 predictors	0.158	0.555	0.55	0.46-0.63
Predict protective	0.582	0.624	0.47	0.38-0.55
Combine 4 predictors	0.188	0.801	0.55	0.47-0.64
SNPs3D	0.530	0.828	0.52	0.44-0.59
General strategy (?)	0.408	0.940	0.53	0.45-0.61
SIFT + Splice Port	0.586	0.943	0.51	0.42-0.60

	P-value: Crude	Adjusted †	ROC area	95% CI
Late submission	0.030	0.041	0.59	0.51-0.67
PON-P, del unreliable 0.95	0.123	0.045	0.49	0.36-0.63
SNPs&GO	0.051	0.083	0.60	0.48-0.72
SNAP	0.109	0.085	0.57	0.44-0.71
PON-P all	0.058	0.091	0.54	0.40-0.70
Expert + 4 predictors	0.039	0.108	0.58	0.45-0.71
PON-P, del unreliable 0.99	0.286	0.159	0.44	0.31-0.57
Naïve Bayes, UCSC multiZ(3)	0.098	0.161	0.59	0.46-0.72
Combine 4 predictors	0.054	0.163	0.57	0.44-0.70
Naïve Bayes, UCSC multiZ(4)	0.101	0.185	0.58	0.45-0.71
SNPs3D	0.166	0.259	0.54	0.40-0.68
Predict protective	0.129	0.278	0.54	0.40-0.67
General strategy (?)	0.171	0.338	0.54	0.42-0.66
SIFT + Splice Port	0.590	0.686	0.49	0.35-0.62

	P-value: Crude	Adjusted †	ROC area	95% CI
Late submission	0.003	0.005	0.76	0.56-0.96
PON-P, del unreliable 0.95	0.122	0.020	0.67	0.42-0.92
SNPs&GO	0.002	0.020	0.80	0.59-1.00
Expert + 4 predictors	0.004	0.031	0.75	0.42-1.00
SIFT + Splice Port	0.011	0.043	0.71	0.40-1.00
SNAP	0.039	0.047	0.54	0.08-0.99
General strategy (?)	0.009	0.049	0.73	0.34-1.00
Naïve Bayes, UCSC multiZ(3)	0.009	0.057	0.79	0.54-1.00
Naïve Bayes, UCSC multiZ(4)	0.010	0.069	0.77	0.52-1.00
Combine 4 predictors	0.011	0.071	0.72	0.50-0.95
PON-P all	0.024	0.073	0.72	0.37-1.00
PON-P, del unreliable 0.99	0.239	0.152	0.36	0.07-0.64
Predict protective	0.050	0.161	0.39	0.07-0.71
SNPs3D	0.060	0.167	0.46	0.11-0.80

† Adjusted for study center and ethnicity
¥ Only applied to carriers of in-class sequence variants

Figure 0.7. Assessement of predictions submitted for the RAD50 challenge.
Prediction from uniPadova are labeled in blue. The three tables refer to three different groups of variants: all rare variants, only missense variants, and only missense substitutions mapping on the protein functional domains. (Figures provided by CAGI assessor ean V. Tavtigian)

9.3. Novel Nav1.5 channel mutations associated with Brugada Syndrome

Mutations in SCN5A gene coding for Nav1.5 channel appear to be the genetic cause for an estimated 15% to 30% of Brugada Syndrome (BrS) (MIM: 601144) [423]. Mutations in this gene have further been linked with various other pathological conditions including long QT syndrome subtype 3 (LQT3) (MIM: 603830), and cardiac conduction disease (CCD) (MIM: 113900). To date, several other genes encoding channels have been associated to BrS or proposed as risk factors, but the mutations on SCN5A represent the major contribution. The Nav1.5 isoform is the predominant subunit in the heart and is involved in excitability of arterial and ventricular cardiomyocytes and propagation of impulses through a specific conduction system [424]. The syndrome is characterized by syncope and sudden cardiac death resulting from ventricular tachy-arrhythmias and a right pre-cordial ST segment elevation in the ECG. Many patients remain asymptomatic but the disease manifests at young age, with higher prevalence in men. The characteristic physiological alterations originate from an impaired function of the Nav1.5 channel. The functional characterization of mutant channels using the patch-clamp technique revealed how mutations associated to BrS lead to loss of Na+ current through several mechanisms. In contrast, mutations in SCN5A associated to LQT3 syndrome cause an increased persistent sodium current. However, defining a genotype-phenotype relationship is still difficult, since some identical mutations result in non-functional or hyperactive channels, depending on the genetic background of the individual host. Furthermore, after the spread of available genetic tests for BrS, a significant percentage of rare SCN5A variants have been also identified in a control population (2% in healthy Caucasian, 5% in healthy non white subjects) [423]. Melvin Scheinman's group at Department of Medicine, University of California San Francisco, identified three novel mutations in the SCN5A gene in two independent families with BrS and investigated the function of these Nav1.5 mutant channels. The CAGI challenge was to predict the effect of the mutations on Nav1.5 function, in particular in terms of current density.

The Nav1.5 protein channel

The Nav1.5 α-subunit is part of the voltage-gated cardiac sodium channel involved in the initiation and conduction of the action potential (AP). The structure of the Na+ voltage dependent channel is unknown but recently the three-dimensional structure of NavAb from *Arcobacter butzleri,* a probable ancestor of the vertebrate Nav channels, has been determined [425]. The human family of voltage-gated channels includes nine genes, SCN1A-SCN11A, with about 50% of sequence identity which share a common architecture composed of four homologous transmembrane domains DI-DIV linked by intracellular loops (IDLs). Among the six transmembrane segments (S1-S6) composing each domain, the S4 segment has a relevant role in the activation of the channel, while segments S6 and S5 together with their connecting loops (P loops) form the pore channel. The activation state is coupled with a mechanism of inactivation which involves the inactivation gate formed by the DIII-DIV linker (Fig. 9.8) [426].

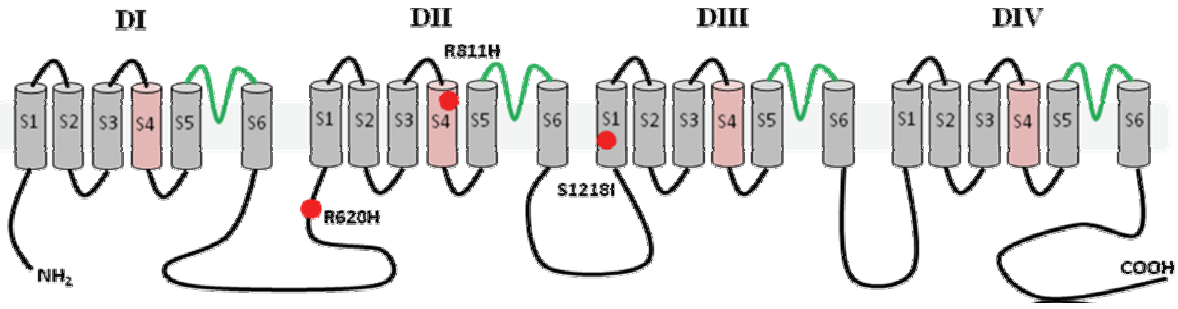


Figure 0.8. Domain organization of SCN5A.
P loops are indicated in green; S4 segments (in pink) are crucial for channel activation; tested mutations are represented as red balls.

Mechanisms of altered Nav1.5 function in BrS

Many mutants channels have been experimentally investigated for their ability to conduct Na+ flux compared to wild type channels. Four distinct loss of function mechanisms have been proposed: i) production of a non-functional channel, creating haploinsufficiency, ii) negative or positive steady state inactivation of the channel, iii) accelerated inactivation, iiii) enhanced intermediate or slow inactivation. Mutations in SCN5A result in channel proteins with reduced or no Na+ current for two possible causes. Either through impaired intracellular trafficking of the ion channel to the plasma

membrane, thereby reducing membrane surface channel expression, or through altered gating properties of the channel. All non-frame or frame-shift mutations are predicted to produce a truncated non-functional channel. Among missense mutations we can distinguish between mutant channels causing protein unfolding, retained in the endoplasmic reticulum for degradation, and others reaching the membrane but exhibiting reduced activity.

Although BS mutations seem to be sparsely distributed across the entire Nav1.5 sequence, their incidence is higher in transmembrane and in pore-forming parts, such as segments S5 and S6 and the interconnecting P-loops. This suggests that non-functional channels could be caused by pore-localizing missense variants [423-424]. Kapplinger and colleagues performed a retrospective analysis of the BrS mutations databases from nine different centers and compared mutation frequency, type, and localization among cases and 1,300 healthy controls from diverse ethnic origins. They found that 50% of the rare unique variants identified in healthy controls localized in two linker domains DI-DII and DII-DIII, suggesting that some of the possible pathogenic rare mutations identified in BrS patients could be false positives.

Mechanisms of altered Nav1.5 function in LQT3

Mutations of the SCN5A gene are also associated to another inherited cardiac disorder named long QT syndrome (LQT3). The disease is characterized by prolonged ventricular repolarization predisposing to ventricular arrhythmias, and sudden cardiac death. Most of the mutations linked to LQT3 syndrome resulted in defects of the inactivation gating of the channels, which show an increased persistent Na+ current during the AP plateau. The conduction of Na+ ions at depolarized membrane potential prolongs the AP duration see as an increased QT-interval on the ECG [427]. In contrast to BrS, LQT3 mutations have not been found in the P loops (S5/S6 linker), while the incidence is higher in S1/S2, S3/S4 and intracellular linkers. Mutations in transmembrane regions are also reported except for segment S1 of each domain and all segments of the domain DII [424].

9.3.1. Method

For this challenge we had to predict the impact of the three mutations, R620H, R811H, and S1218I, on the current density conducted by the resulting channel proteins. First of all, we classified the variants on the basis of their deleteriousness. We predicted the functional impact of the three mutations by a combination of five computational methods using only sequence information: PHD-SNP (version using sequence information), PHD-SNP (version using MSA information), I-Mutant, Mupro, and ParePro. Two different approaches were used to estimate the percentage of current density reduction for each mutant compared to the wild type channel. The first prediction was made using the mutation impact index obtained by the majority voting of the used methods. Another approach attempted to predict the functionality of the channel on the basis of the mutation position on the diverse channel domains. We distinguished between variants associated to BrS which cause reduced or no Na+ current and LQT3 mutations that instead showed unchanged current density. The third prediction was made by the 100 sided die.

9.3.2. Results and Discussion

The domain organization of the Nav1.5 protein channel was defined according to the Pfam classification. A topology diagram of the protein is reported in Figure 9.8. This allowed the mapping of the three tested mutations. It seems that the R620H mutation localizes in the intracellular DI-DII linker, while R811H and S1218I map on transmembrane regions of the protein. The estimate of the current density was calculated considering the position of the variants. In particular, R811H is one of the positively charged residues that move following the depolarization and initiate opening of the pore channel [425]. This mutation is thought to be likely deleterious for the activation of the channel and thus probably impact the current flux. All the three residues are conserved. R620H in the DI-DII linker should have an higher probability to cause no changes in their current flux since many rare variants have been found in this domain in healthy controls. Furthermore, LQT3 mutations showing no changes in current density have high incidence in intracellular linkers. The third mutation S1218I is

located in the segment S1 of the domain DIII (Fig. 9.8). Interesting to note, in segments S1 of each domain no mutations associated to LQT3 have been found. Thus, it has a positive probability to have a reduction in current density such as other BrS mutations. As the range of reduction in current density was from 0 to 100%, we used only three changes: 0%, 50%, and 100%. The established values are reported in Table 9.1 including results from computational methods and from the 100 side die random prediction. The consensus of computational methods predicted as deleterious R811H and S1218I mutations, while R620H has been predicted for the most of the methods having no functional effect on the protein. These predictions agree with those derived from the structural/functional analysis of the protein.

Mutation	Experimental	Consensus	Manual	Random
c.1859G>A R620H	100	10	0	48.98
c.2432G>A R811H	0	100	50	57.69
c.3653G>T S1218I	50	50	50	5

Table 0.1. Experimental and prediction results for SCN5A mutations.

Experimental results

Experimental results of the current density change for each of the three variants revealed that R620H has the largest impact on protein function, completely blocking the ion current through the channel. The S1218 mutant showed a reduced activity. Surprisingly, the R811H mutation, which alters one of the charged residues initiating pore channel opening, seems to maintain the total current density. The predictions for this competition have not been assessed at the CAGI meeting and we thus do not know the explanation for these results.

9.4. Distinguishing exomes of Crohn's disease patients from healthy individuals

Crohn's disease, like ulcerative colitis, is a form of inflammatory bowel disease (IBD, MIM 266600) characterized by a chronic inflammation of the gastrointestinal tract. The disease affects people with an incidence of 1 in 250, but the etiology is still unknown. IBDs are multifactorial diseases where genetic, environmental, and immunological factors all contribute to the establishment of the pathological condition. The hypothesis is that a pathological inflammatory response arouses from an unknown pathogen or from the normal bowel flora within a genetically susceptible individual. The present challenge aimed to predict the probability of an individual to have Crohn's disease or to be healthy starting from the exome sequences.

Molecular mechanisms involved in Crohn's disease

Genome-wide genotyping with high-throughput approaches allowed the identification of associations between about 1,300 loci and 200 diseases or traits [3]. The genotyping is performed by measuring the differences of allele frequencies in case and control individuals throughout the genome. Since the associated alleles contain information about the molecular processes modulating risk to disease, the challenge is therefore to identify the disease-causing pathways that may be targeted for diagnostics and therapeutic drug discovery. Recent genome-wide and candidate gene association studies have identified 71 susceptibility loci for Crohn's disease [428]. It is interesting to note that many of the loci identified were also associated to other complex diseases. Thus, it is emerging that chronic inflammatory disorders and autoimmune diseases (ankylosis spondylitis, rheumatoid arthritis, systemic lupus erythematosus), probably share genetic risk factors which influence common pathways. Usually, the connection between a specific gene and the pathogenic mechanism is implicated by its proximity to a disease-associated locus, or by its appropriate biological function. The functional annotation of genes mapping to Crohn's disease-associated loci suggests that pathogenic mechanisms involve diverse pathways that are in a delicate balance including modulation of T cell and other immune pathways, regulatory functions in self tolerance, and infection

defense functions. Several candidate genes mapping in loci associated both to Crohn's disease and ulcerative colitis are involved in the IL23 pathway (JAK2, STAT3, IL12B, and PTPN2) [429]. The role of Interleukin-23, which is the central protein in this pathway, has been well established in expanding and maintaining T-cells (Th17 cell) involved in antimicrobial immune response which contribute to autoimmunity and tissue inflammation [430]. Another pathway operating in IBD inflammatory response involves TNF signaling which includes NF-kB activation. RANKL is a TNF-related cytokine which activates osteoclast differentiation. Its association discovery was relevant to explain the osteoporosis clinically associated with Crohn's disease [431]. Studying the relationship between genome and transcriptome (eQTL analysis), it seems that regulatory effects are common mechanisms of disease susceptibility. Alterations in DNMT3A activity, a key mediator of epigenetic regulation and regulator of TNF-α, have been associated to Crohn's disease [428]. The connections between genes mapping to the 71 Crohn's disease associated loci have been investigated by the Gene Relationships Across Implicated Loci (GRAIL) approach [432], which identifies correlations between genes based on descriptive features that delineate the underlying pathogenic mechanisms. The candidate genes among the associated loci seem to be non-randomly correlated, but instead an evidence-based connectivity has been predicted [428].

Like Mendelian diseases, there is a growing realization that pathogenic processes could be identified investigating the protein-protein interaction network of the causal genes. Especially for rheumatoid arthritis and Crohn's disease, it has been demonstrated that proteins encoded by disease associated loci are interacting, suggesting that common risk variants may act in a set of proteins involved in the same biological processes [433]. The authors also found proteins known to interact directly linked by common interactors expressed in the same tissue as associated proteins and mapping in genomic regions with significant association. This further suggested that common interactors may carry risk variants [433].

Computational Approaches for rare variant identification in complex disease

Identifying the associated SNPs is relatively easy, but mapping them to the underlying rare causal variants (that is, with a minor allele frequency (MAF) <1%) that functionally

influence the disease risk is the next challenge. Recently, several examples where high-throughput sequencing of genomes allowed identification of rare mutations in genes located near common alleles associated to complex diseases were reported in the literature [9]. In particular, Momozawa and colleagues identified rare variants in the IL23 gene which have a protective role for Crohn's disease [434]. First, it is assumed that the causal variant must be closely correlated and in linkage-disequilibrium with associated variants. It is the rare variant that further explains most of the association evidence. Rare causal variants may include those more likely to be deleterious and, therefore, causing large functional effects that could be easily observed to gain insight on the pathogenic disease mechanism [9]. In addition, the existence of multiple causal variants have to be considered.

So far, studies aimed to discover rare variants associated to complex diseases have been limited to re-sequencing of candidate genes or genomic regions identified by linkage analysis of genome wide association studies. Exome sequencing has been used in conjunction with sampling strategies based on comparing variants found in distantly related individuals from a family or searching for de novo variant in families where only the offspring is affected [435]. Another strategy to identify novel candidate alleles used the sequence of individuals with extreme phenotypes in which the frequency of associated SNPs are more frequent. Furthermore, with reducing costs of exome sequencing and available control exomes, the identification of rare variants by comparing case-control populations is a promising approach. The difficult task however is to detect the association between rare variants and a specific trait. For this end, some methods assessing the role of each variant alone or in concert with others located in a gene or in multiple genes have been developed. A simple test named "burden test" evaluates the distribution of rare variants in a gene of case individuals compared to control individuals (Fig. 9.9) [436]. If a rare variant is enriched in a candidate gene it might play a role in disease pathogenesis. Some methods use multivariate analysis of variants, incorporating a priori information of the functional impact of the variants, gene function, and involved pathway [437]. To enhance the power of association tests, a good approach is to consider only variants more likely to be deleterious, those occurring in conserved positions or predicted to have large functional effects.

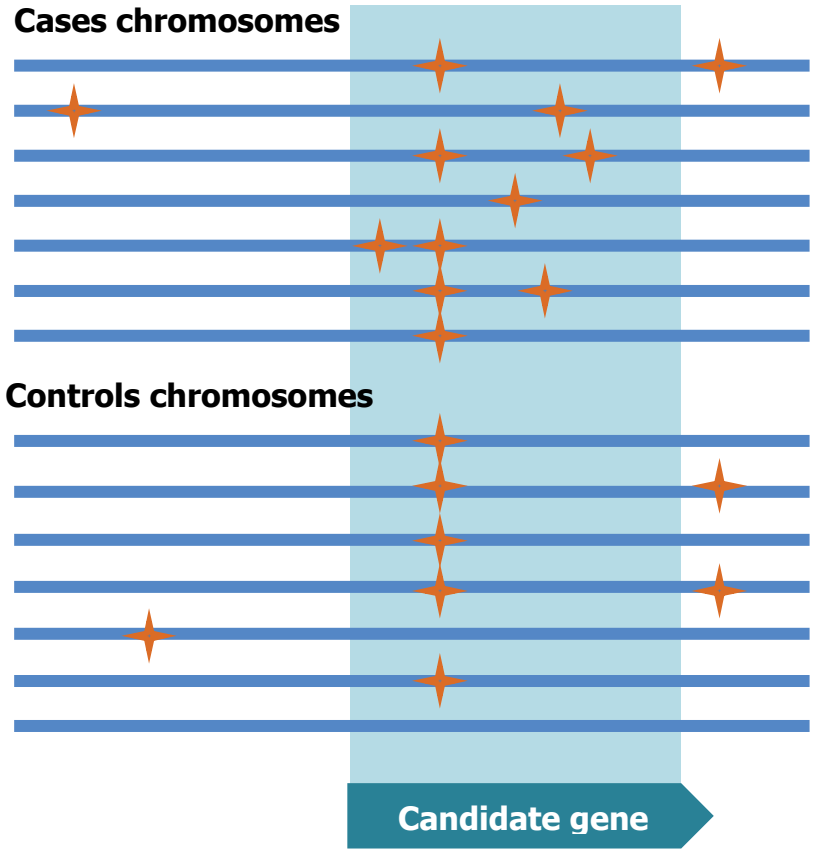


Figure 0.9. **Rare missense variants enriched in candidate genes.**

9.4.1. Method

Our approach consists of three main steps: (1) filtering the genome data, (2) compiling extended lists of putative disease-related genes and (3) deriving probabilities for disease phenotypes (Fig. 9.10). The first step is to apply discrete filtering to reduce the number of candidate genes to a set of high-priority candidates. For this goal we used a recently developed software, ANNOVAR, resulting in a list of about 20 genes per patient containing novel putative causal variants occurring in indispensable genes at conserved positions. ANNOVAR removes polymorphisms found in a reference database (e.g. 1000 genomes project) and performs a first stratification of candidate alleles on the basis of their predicted impact or deleteriousness using SIFT [438].

For each trait we had to predict, we compiled a list of genes known to be associated to the same or similar phenotypes from OMIM. In particular, for complex traits we included PheGenI (http://www.ncbi.nlm.nih.gov/gap/PheGenI), a phenotype-oriented

resource, to collect data from genome-wide association study (GWAS). By comparing the list of candidate genes from ANNOVAR with the list of associated genes, we assessed only rare variants that have been found on candidate genes or in genomic regions identified by GWAS. Candidate alleles were additionally prioritized by existing biological or functional information about a gene. For complex diseases such as rheumatoid arthritis and Crohn's disease, there is evidence that common genetic associations implicate regions encoding physically interacting proteins [433]. The candidate gene list from ANNOVAR was therefore virtually expanded using STRING [161], a protein-protein interaction database, to obtain a list of candidate gene interactors. This second list was compared with the list of disease-associated genes to find a functional or physical relationship of candidate genes to the latter (Fig. 9.10). Finally, the assessment of associations between an individual and a phenotype was performed analyzing the collective effects of rare variants across one or multiple genes. Prior evidence about variants (e.g. known pathogenic mutations) and their severity (missense, frameshift or stop-gain) were incorporated into a probabilistic score taking into account the known genes and their interactors (Fig. 9.10).

ANNOVAR

ANNOVAR is a tool to annotate variants identified by high-throughput sequencing technology [438]. Single nucleotide polymorphisms (SNPs) are identified as silent substitution, missense, stop-gain, stop-loss, or frameshift variant with further annotation based on several databases. ANNOVAR identifies variants in specific genomic regions, such as those recovered from GWAS for complex diseases or loci specific for Mendelian disease. A useful ANNOVAR option is the filter based annotation protocol removing variants less likely to be pathogenic, such as synonymous substitutions or variants with a low conservation score. This protocol also helps the reduction of variants to analyze filtering those present in dbSNP or observed in the 1000 Genome Project, which are less likely to be rare disease variants. The output consist of a list of candidate genes carrying the most likely causal variants. The software can run in a modern PC (3GHz Intel Xeon CPU, 8Gb memory) and it takes only 15 minutes to perform the full variant reduction protocol for an exome.

PheGenI

The Phenotype-Genotype Integrator (PheGenI, http://www.ncbi.nlm.nih.gov/gap/PheGenI) is a web interface provided by NCBI to search results from genome-wide association study (GWAS) results associated to a target phenotype. The tool integrates data from the NHGRI GWAS catalogue with several databases hosting at the NCBI. In addition to phenotype-genotype associations extracted from NHGRI, it contains data submitted to the database of Genotype and Phenotype (dbGaP) at NCBI. PheGenI provides a list of phenotypes or traits reported in the MeSH nomenclature, but the list is incomplete at the moment. The output gives a list of genes identified by GWAS studies with a graphical view of their genomic position and their associated OMIM entries. The results could be download also as list of SNPs complete with genomic location, functional class, and validation status, e.g. those reported by the 1000 Genomes Project or HapMap.

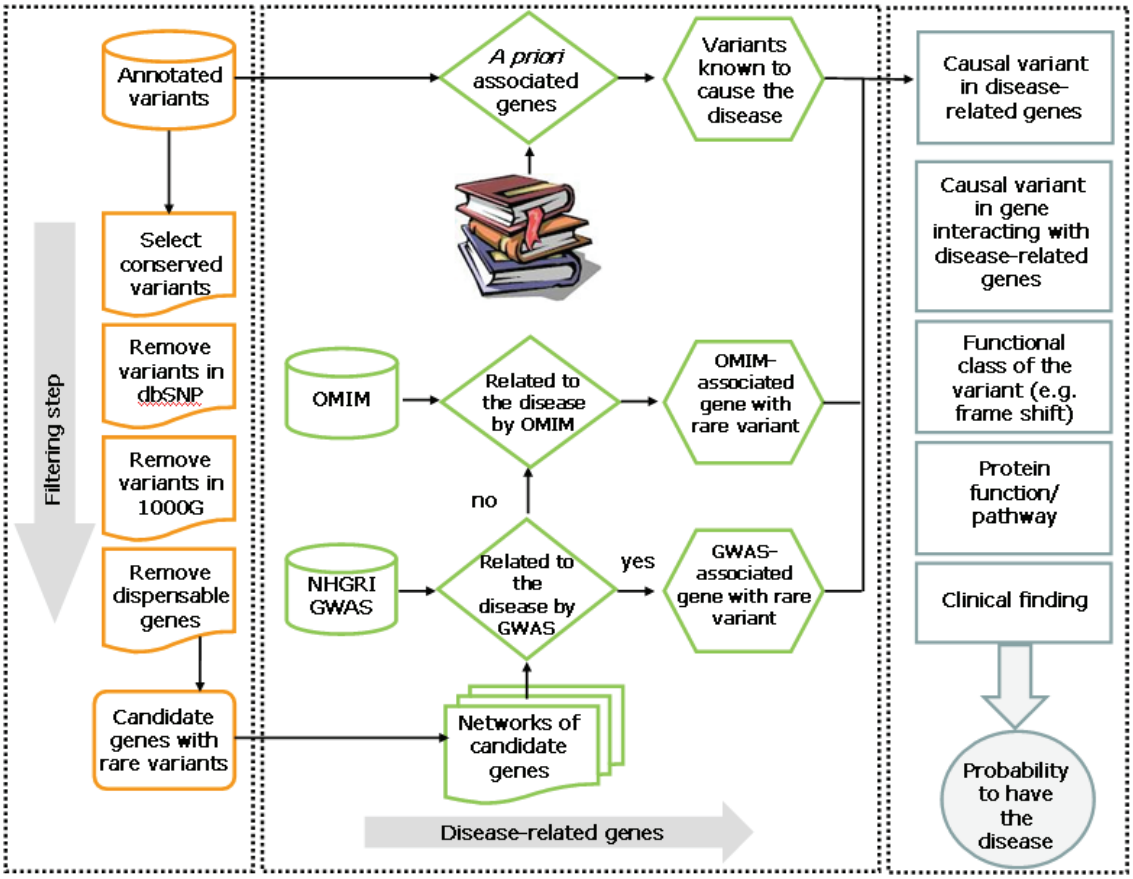


Figure 0.10. Schematic representation of the variant annotation protocol.

9.4.2. Results and Discussion

The data provided by CAGI consisted in exome sequences from 56 individuals with the challenge being to predict the probability of each individual to have Crohn's disease. Variants detected by exome sequencing were annotated using ANNOVAR and each individual presented about 3.5-4.1 million SNVs. The variant reduction protocol of ANNOVAR allowed us to obtain a subset of candidate genes containing potential causal variants for each individual. Comparing the lists of candidate genes obtained from ANNOVAR with the list of candidate genes associated to Crohn's disease from GWAS or OMIM, we observed only 3 individuals (PGP2, PGP23 and PGP53) with potential causal variants (SLC37A4, NOD2, and CNTNAP2) (Table 9.2). These variants were p.A1006P and p.L1007P in NOD2, p.R377X and p.R126Q in SLC37A4, p.T589P and p.H764P in CNTNAP2. We validated the possible involvement of these variants by checking the literature for information on their association with Crohn's disease. Mutations in the LRR domain of the NOD2 gene are implicated in Crohn's disease [439], while mutations in the neuronal apoptosis inhibitor (NACHT) nucleotide binding domain are involved in Blau syndrome [440]. Mutations in SLC37A4, which are not in the GWAS list, are reported in glycogen storage disease 1B (GSD1B). This is a disorder characterized by recurrent infections and neutropenia. In patients with GSD1B, chronic inflammatory bowel disease (IBD, MIM 266600) appears to be a consequence of leucocyte abnormality [441]. The CNTNAP2 gene encodes for Contactin-associated protein-like 2 and was associated to Crohn's disease by GWAS studies. Genetic variation in this gene is the cause of cortical dysplasia-focal epilepsy syndrome (MIM: 610042) and influences susceptibility to autism type 15 (MIM 612100). The others 53 individuals carry variants in genes that are not found to be related with Crohn's disease.

Since we suppose that other genes may be involved in the pathogenesis of the disease, candidate genes identified by ANNOVAR could have some relationship with genes identified by GWAS. Thus, we selected genes from the ANNOVAR list interacting with genes known to be related with Crohn's disease. We found causal variants in about 300 genes interacting with Crohn-related proteins, but only 40 genes were mutated in at least three individuals. Our hypothesis was that patients with Crohn's disease should

display variants in the same gene. The objective thus was to cluster individuals in groups of patients having disease causal variants in the same gene or in a combination of genes. The list of interactors was retrieved from STRING, which scores the relevance of interactions with information deriving from different databases or from the literature. Using as threshold only interactors for which the interaction was experimentally determined with a score of at least 0.5, we found three genes that are physically associated to genes related to Crohn's disease: 42 individuals have mutations in PTPN11, 8 in RUNX2, and 19 in NCOA3 (Table 9.2). Individuals with NCOA3 variants and those with altered RUNX2 have also alterations in PTPN11. The three genes carry a truncating mutation. The PTPN11 variant p.Y197X (dbSNP code: rs76982592) was reported in dbSNP with a frequency in the general population of 0.104, while Q1269X in NCOA3 (dbSNP code: rs75561226) was reported without a frequency. The other NCOA3 missense variant, Q1261H, and Q54X in RUNX2 were never reported before. With this criteria we can make a first stratification of the individuals, retrieving 22 of 56 putative Crohn's disease patients.

Since not much information is not reported in databases for interactions and only a functional relationship could be known, we decided to use a more relaxed selection for interactors, using as threshold the combined STRING score with values of at least 0.5. Doing so we can consider associations found by text mining and have further verify the quality of the interactions. Many individuals were found to carry mutations in other genes likely to be interactors of genes associated to Crohn's disease in this way. These are RBMX (25 individuals), TDG (14 individuals), PRKRA (18 individuals), XFHX3 (9 individuals), CELA1 (5 individuals), and DSPP (3 individuals) (Table 9.2). Even if PTPN11 was found directly to interact with genes associated to Crohn's disease, the variant identified in this gene was frequent in the population. For this reason, we thought the gene to be in some way associated but not sufficient to determine the disease. The TDG gene carries a splicing site variants (c.793-1G>T) and PRKRA presents a splicing variant in addition to several missense substitutions. Another gene presenting several missense variants was RBMX, whose association with Crohn's disease is unclear and different variants were distributed among the individuals. DSSP carries the S960G in all three individuals and its expression seems to be regulated by TGFB1 (associated to Crohn's disease) as well as directly interacting with RUNX2.

Proteins homologous or belonging to the family of ZFHX3 and CELA1 regulate MUC5B expression. A down regulated expression of mucins, including MUC5B, has been found in the ileum and colon of Crohn's disease and ulcerative colitis patients versus controls individuals. In our dataset these genes present a stop gain variant, Q819X, and two missense variants, L210P and G208A. Using information about gene association strength with Crohn's disease, type of variants and frequency in the population, it was possible to draw a ranking of the genes in the following order: RUNX2, NCOA3, TDG, ZFNX3, DSPP, PRKRA, CELA1, RBMX, and PTPN11 (Table 9.2).

Mutant protein	Subjects	GWAS, OMIM	Interaction type	Function	Mutation class
NOD2 (CARD15 or IBD1)	1	GWAS	-	Nucleotide-binding oligomerization domain-containing protein 2, Caspase recruitment domain-containing protein 15, Inflammatory bowel disease protein 1	Missense
SLC37A4	1	OMIM	-	Glucose-6-phosphate translocase, associated to glycogen storage disease type 1B, 1C, 1D	Stop gain
CNTNAP2	1	GWAS	-	Contactin-associated protein-like 2	Missense
RUNX2	8		Physical	Runt-related transcription factor 2, Osteoblast-specific transcription factor 2	Truncating
NCOA3	19		Physical	Nuclear receptor coactivator 3 Lipid metab olic process	Missense
TDG	14		Physical with DNMT3A	G/T mismatch-specific thymine DNA glycosylase	splicing
ZFNX3	9		Functional	Zn finger homeobox protein 3, myogenesis	Stop gain
DSSP	3		Functional		Missense
PRKRA	18		Functional	INF inducible dsRNA dipendent activator	Splicing
CELA1	5		Functional		Missense
RBMX	25		Functional	nuclear ribonucleoprotein G, cellular response to IL1	Missense
PTPN11	43		Physical	Tyrosine-protein phosphatase non-receptor type 11	Truncating

Table 0.2. Priority list of candidate genes used for population clustering.

The priority list was used to score the probability of an individual to have Crohn's disease. One prediction was made manually combining diverse information on protein

function, other four predictions were calculated using four different scoring functions: majority voting, weighted sum, association rules, clustering (Table 9.3). A arbitrary threshold was chose expecting that about 50% of the individuals could had Crohn's disease. A final prediction was performed by the 100 side die to assess the random prediction results.

Phenotypic results

The CAGI organizers eventually released the results revealing that, in contrast with our expectation, 42 of the 56 individuals have Crohn's disease. The manual prediction worked better identifying 31 of the 42 Crohn's disease patients (Table 9.3). However, changing the threshold to discriminate Crohn patients and healthy individuals, we also reached a very good prediction using different scoring functions. In particular, the clustering method identified 40 of the 42 Crohn's individuals with an accuracy of 0.91 (Table 9.4).

	Majority	Weighted sum	Association rules	Clustering	Manual	Random
TP	17	18	22	14	31	24
FP	4	4	4	4	4	8
TN	10	10	10	10	10	6
FN	25	24	20	25	11	18
Sensibility	0.4	0.43	0.52	0.4	0.74	0.57
Specificity	0.71	0.71	0.71	0.71	0.71	0.43
Selectivity	0.81	0.82	0.85	0.81	0.89	0.75
Accuracy	0.48	0.5	0.57	0.48	0.73	0.54

Table 0.3. Predictions of Crohn's disease individuals.
The results obtained with different scoring methods are reported. TP: true positive; FP: false positive; TN: true negatives; FN: false negative; Sens: sensibility; Spec: specificity; Sele: selectivity; Ac: accuracy.

	Majority	Weighted sum	Association rules	Clustering	Manual
TP	26	40	37	40	33
FP	4	4	3	3	3
TN	10	10	11	11	11
FN	16	2	5	2	9
Sensitivity	0.62	0.95	0.88	0.95	0.79
Specificity	0.71	0.71	0.79	0.79	0.79
Selectivity	0.87	0.91	0.93	0.93	0.92
Accuracy	0.64	0.89	0.86	0.91	0.79

Table 0.4. Prediction of Crohn's disease individuals with a modified threshold.
Lowering the threshold used to cluster Crohn-related individuals we had a better performance of the

prediction methods.

During the evaluation of predictions, the organizers decided to take with caution the prediction for height of the control individuals which were sequenced with a different method and present an higher number of variants. The risk is that many of these variants are errors due to sequencing or mutation calling process (Fig. 9.11).

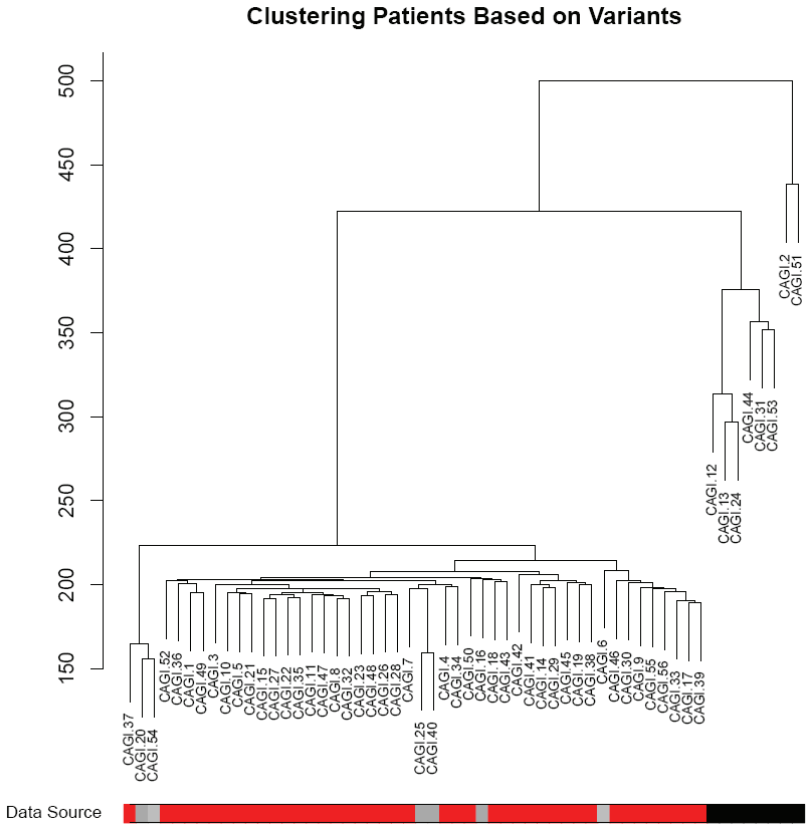


Figure 0.11. Clustering Patient Samples Based on Variants.
A Euclidean distance based hierarchical clustering of the samples based on the reported variants. The samples came from four sources. The 42 samples from German patients with CD are labeled in red in the colored strip under the dendrogram. Eight controls from a population genetics study are indicated in black. Two centenarians are indicated in gray, and a HapMap trio and one German male are indicated by dark gray. Please note that the 8 samples from the population genetics study are separated from the other case and control samples, and they were excluded from some of the further analyses for this reason. (Figure provided by CAGI assessor A. Morgan)

Comparison with other participating groups

The assessor Alexander Morgan noticed that the submissions of most groups cluster very closely to one another, with the interesting exception of one of our submission (sub.95) which clusters separately from the other submissions from our group. This represents the random predictions obtained by the 100 side die (Fig. 9.12 and Fig. 9.13).

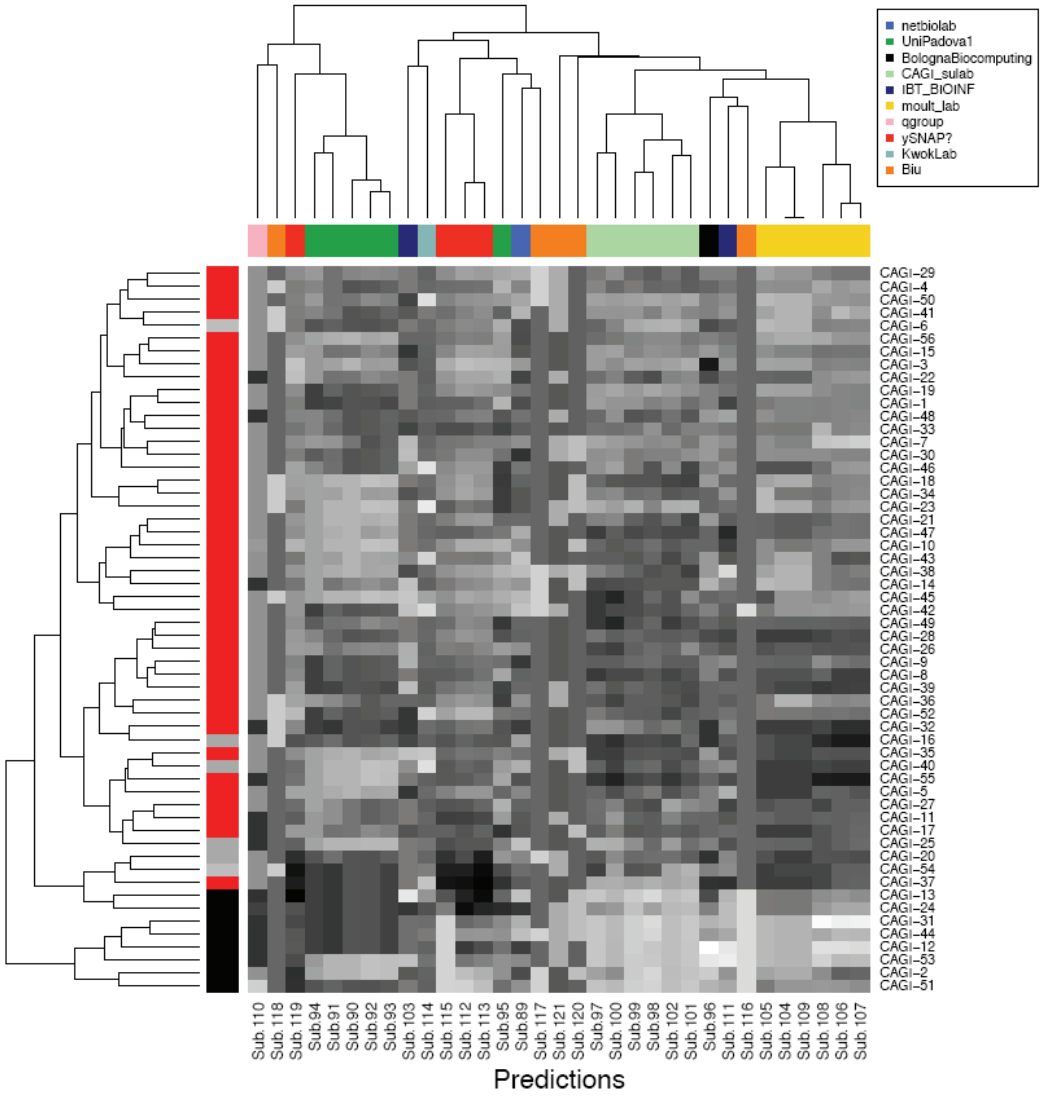


Figure 0.12. **Clustering of Samples Based on Predictions.**
A heat map of the scaled predictions is indicated by each grayscale cell. A darker cell indicates a higher reported prediction of CD likelihood by each submission. Submissions are organized into columns. The color strip above the heat map labels the group making the submission, as indicated by the legend in the upper right corner. The patient samples are in the rows, with the color strip on the left indicating whether the samples were from CD patients (red) or controls (different shades of black or gray as described in Figure 9.11). (Figure provided by CAGI assessor A. Morgan)

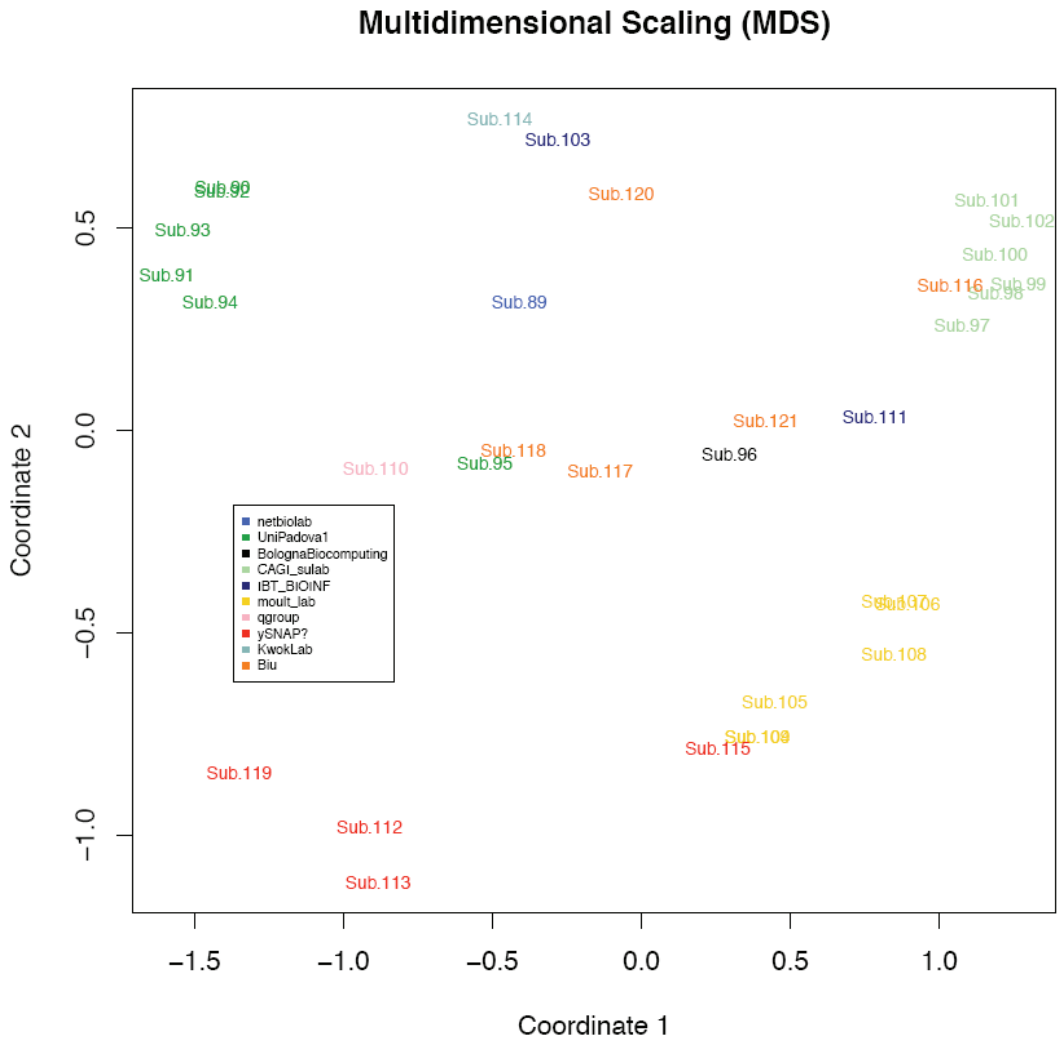


Figure 0.13. Multidimensional Scaling of Submissions.
Similar to the clustering in Figure 9.12, the distances between submissions were mapped onto a two dimensional space. Many of the submissions by the same groups were closely associated to one another through multidimensional scaling. (Figure provided by CAGI assessor A. Morgan)

For each submission, the assessor calculated the ROC curve were the true positive rate (sensitivity) is plotted in function of the false positive rate (specificity) (Fig. 9.14). The area under the ROC curve (AUC) is a measure of how well a participant group can distinguish between Crohn-related and healthy individuals. A perfect prediction has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore, the closer the ROC curve is to the upper left corner, the higher the overall accuracy of the test [442]. One of the participant group reached the best performance identifying 40 of the Crohn-related individuals. However, our predictions worked better than many other groups (Fig. 9.14 and Fig. 9.15). It is interesting to note

that while the best group submitted probabilities in a narrow range around 0.5, our group tried to classify the two populations with a stronger decision (Fig. 19). This leads to a worse assessment of our prediction.

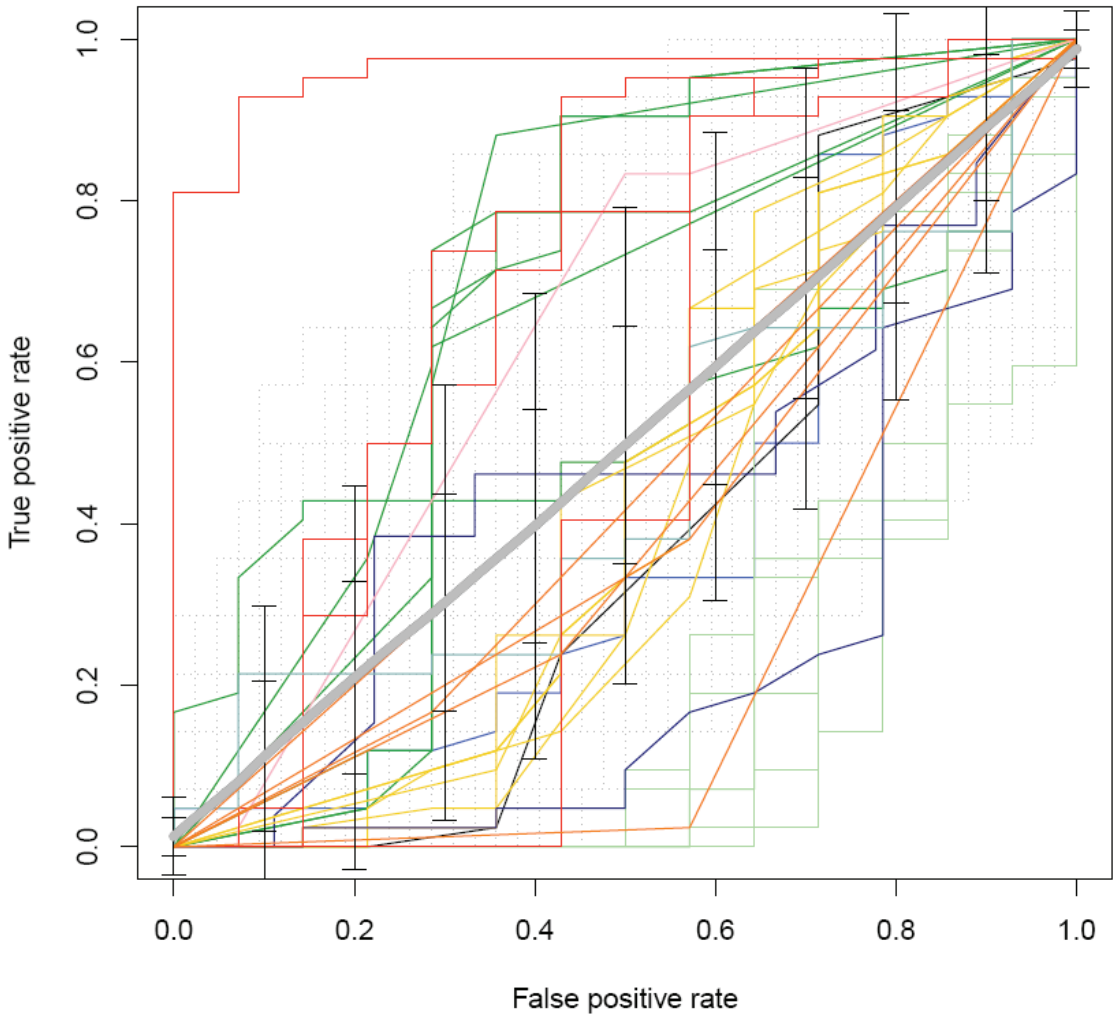


Figure 0.14. **ROC Curve for Predictions.**
The Receiver Operating Characteristic (ROC) Curves for each prediction submission are shown. Also shown are the result of 1,000 random predictions in gray, along with the confidence intervals for 1 and 2 standard deviations from the average of these 1,000 predictions. The curves of the actual 1,000 randomizations are shown with lightly dotted lines. (Figure provided by CAGI assessor A. Morgan)

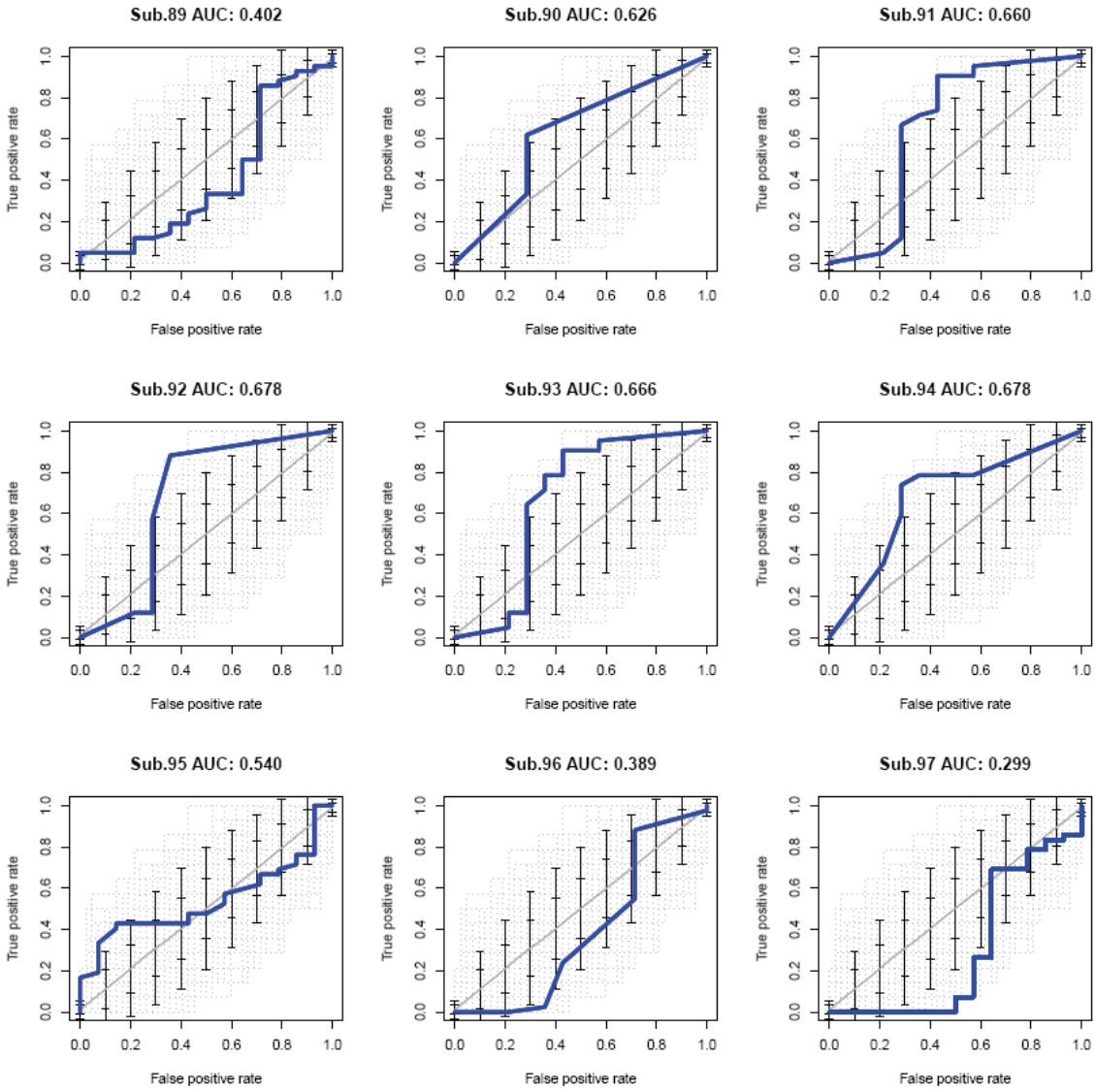
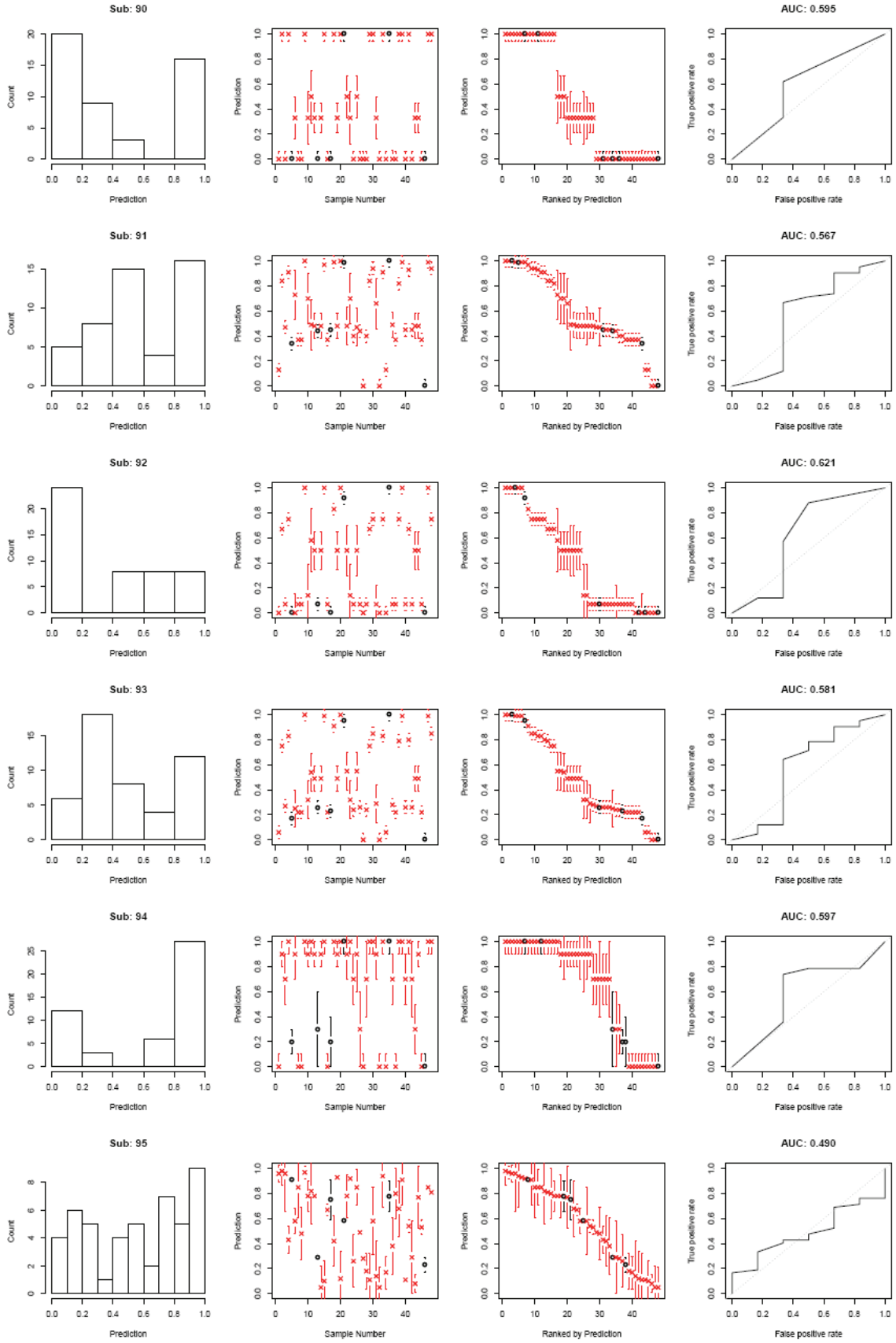


Figure 0.15. ROC Curves for Each Submission.
Similar to Figure 9.14, but instead of overlaying each submission together, all submissions from our group (sub.90-95) and some others are shown separately on an individual ROC curve. The area under the curve (AUC) is indicated at the top. AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. (Figure provided by CAGI assessor A. Morgan)


Sub: 90
AUC: 0.595
Sub: 91
AUC: 0.567
Sub: 92
AUC: 0.621
Sub: 93
AUC: 0.581
Sub: 94
AUC: 0.597
Sub: 95
AUC: 0.490
Count
Prediction
Sample Number
Ranked by Prediction
True positive rate
False positive rate

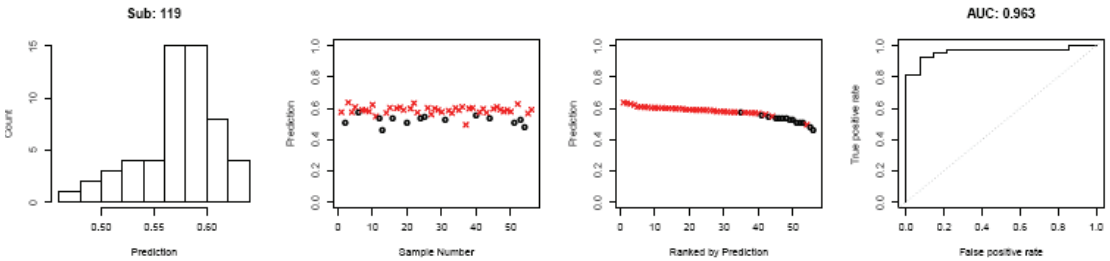


Figure 0.16. **Prediction results evaluation.**
A series of smaller panels for each submission organized by row is shown. Submissions from our group are indicated by numbers 90-95, the last in the bottom Sub:119 is the submission which obtained the best performance. The first column shows a histogram of submitted values. The second column shows a scatter plot of submitted predictions, with all controls indicated by black circles (all control types are colored black) and diseased (CD) samples with red exes. If a standard deviation was reported in the submission, it is indicated by confidence interval bars. Each patient sample is sorted in order along the horizontal axis. In the third column, the CAGI samples are reordered in decreasing order based on prediction value, otherwise it is the same as the second column. The fourth column is the ROC curve for the predictions, with the area under the curve indicated at the top. (Figure provided by CAGI assessor A. Morgan)

The strategy adopted by different groups was to consider the impact of non-synonymous variants found in regions associated to Crohn's disease by GWAS. Some groups evaluated only common variants used in the genotyping process of GWAS, while only our group calculated the probability to have the disease taking into account the unique contribution of rare missense variants found in these individuals. Common diseases are due to multiple deleterious alleles in a mutation-selection balance, sharing a high mutation rate and weak selection. An interesting debate arose from this experiment about the contribution of common and rare variants in complex diseases. Several hypotheses have been formulated. One hypothesis is that common diseases are caused by a multitude of SNPs with small effects and a multitude of rare variants. Another hypothesis is that functional genetic variations have the major contribution and functional information could guide the classification methods. It seems that both common and rare variants have to be considered in the prediction of common disease risk, but the size of their impact has to be re-estimated. All the groups looked for variants that cause functional protein changes using different approaches (e.g. SIFT, PolyPhen, SNAP). The best group, together with our group, used ANNOVAR to select functionally relevant mutations. We used the variant reduction protocol to filter rare variants while the other group used the protocol to find common non-synonymous SNPs. The groups with the best prediction results, including our group, also used non

synonymous variants found in a wider list of genes than those collected by GWAS. In these cases, the interesting genes have been selected using additional information such as expression quantitative trait loci (eQTL) or, in our case, the network of functionally associated proteins provided by the STRING database. A successful prediction strategy should consider common and rare variants prioritized by likely functional effects found in a set of genes including genes functionally related to GWAS loci. Finally, there is a need to create a robust scaling metric weighting severity of modifications, the prevalence of variants, and their impact on protein function.

9.5. Personal Genome Project (PGP) – Predict traits and phenotypes

The ability to sequence the genome in little time and with low costs has increased the possibility to obtain a personal genomic profile. A person may like to know his predisposition to disease or explain an existing trait, or maybe its genetic history. Several companies offer to individual customers the sequence of their entire genome for $400-$2500 in the form of variants identified at specific loci [443]. The declared purpose is to develop a system where genetic information can be used by physicians in diagnosing or application of the appropriate therapeutic strategy. Approaches detecting disease-causing variants are very promising in providing disease diagnosis, including adult onset diseases, detection of carriers, and even prenatal diagnosis. Even if the associations of SNPs with a disease risk give an uncertain prediction of the disease, one of the most desirable outcomes of personal genome profiling is to known the risk to develop common complex diseases. Much progress in this field comes from new sequencing technology which identifies causal variants to discover novel pathogenic mechanisms underlying the disease offering new opportunities for drug development. Pharmacogenetics is another promising field of genomic and personalized medicine. The variability of drug responses among individuals arise from the presence of specific genetic variants. Pharmacogenetics aims to identify genes influencing drug metabolism.

The PGP challenge

The CAGI challenge was to predict the phenotype of ten individuals for whom the exome sequences were available. The clinical profile of each individual including information about age, vital signs (e.g. weight, blood pressure), ethnicity, allergies, medications, medical history, and even facial photographs were also provided. Participants were asked to make prediction for 32 phenotypes (e.g. asthma) and eight numerical traits (e.g. HDL level). An additional challenge was released near the prediction deadline aiming to predict an individual with irritable bowel syndrome (IBS), a color blind man, and a female carrier of familial color blindness. The CAGI challenge focused on the prediction of many common diseases and the task was to give a probability to develop the disease. The sample population provided for CAGI predictions consisted of ten unrelated individuals with probably independent phenotypes. This makes it very difficult to identify the unique causal variants responsible for a specific phenotype. Since, given our experience with Crohn's disease, we assumed that rare variants are more likely to contribute to disease susceptibility, we searched for putative causal variants in disease-related genes and calculated the probability by weighting the contribution of each variant we found.

PGP and Mendelian disease diagnosis

Rare variants have been identified in about 3,000 human genes responsible for Mendelian diseases, and some of these have been implicated in common disease risk. In particular, common alleles associated to blood lipid levels have been found near genes that have been previously known to be involved in lipid metabolism [444]. Some autosomal recessive disorders, such as Cystic Fibrosis, have been extensively studied and a list of mutations is reported in the specific database. There are also disease-causing mutations presenting comparatively high frequency in the population due to founder effects or selection. Examples are the c.35delG mutation in the GJB2 gene and mutations in the HFE gene causing two autosomal recessive disorders, neurosensorial hearing loss and hemochromatosis . However, we can count only a thousand affected individuals for each Mendelian disease and sometimes few mutations have been identified. Even if the majority of individuals affected by a Mendelian disorder carry an already known mutation, a large number of private mutations can be found and the

number of these increase dramatically with the advent of next generation sequencing technology.

Approaches to disease gene discovery adopt different strategies depending on the mode of disease inheritance, the extent of locus heterogeneity, pedigree information, or size and structure of the sample. Recently, exome sequencing has been used as a powerful tool to discover causal variants in genes involved in rare Mendelian diseases, such as Kabubi syndrome, Miller syndrome, and Fowler syndrome (reviewed in [4]). This approach used discrete filtering of all variants observed in the 1000 Genome Project or reported in the publicly available dbSNP database, or even found in a control population. This allowed to reduce the huge number of variants originating from exome sequencing and to focus the investigation on a limited set of candidate genes. Discrete filtering has been more useful for recessive than for dominant diseases, but lowering the MAF cutoff to 0.1% can be helpful in solving dominant disorders [6]. Further stratification of candidate alleles can be obtained by ranking variants on the basis of their predicted functional impact. Nonsense or frameshift mutations resulting in truncating proteins are predicted to be the most important candidates, even if in some cases they can result in a harmless protein loss. For the classification of non-synonymous variants, many of the common computational methods, such as SIFT and Polyphen, use evolutionary annotation [34]. An additional prioritization could be performed by annotating variants for their role in pathways or interactions that could explain the pathogenic mechanism involved in the disease or in similar phenotypes.

Recent studies that successfully identified candidate genes for rare Mendelian diseases focused on the identification of rare or novel variants in the same gene found in unrelated or closely related affected individuals. In unrelated individuals with similar phenotype, we expect to have causal variants in the same candidate gene and for disorders with genetic heterogeneity in a subset of different genes. In case of Kabubi syndrome, investigators applied further genotypic and phenotypic stratification to successful identify variants in the MLL2 gene in a subset of the affected individuals (Ng et al. 2010). Furthermore, with familial information one can further filter variants that do not follow the mode of inheritance expected for the target disease or remove potential causal variants that do not segregate with the disease [432-433].

PGP and common disease risk

Over the last 20 years, genome wide association studies allowed the identification of a large amount of common variants in 800 disease-associated loci for ~150 human disease/traits [445]. However, the identification of these common variants in an individual does not explain his predisposition to develop the disease or the missing heritability of the variant in the family. Complex diseases have a large genetic component and show genetic heterogeneity, but different high-risk variants result in the same phenotype. Sometimes the effect of several moderate-risk variants is aggregative, and the disease seems to have a dominant mode of inheritance. In other families the high-risk variants do not segregate with the disease. Several statistical strategies for association studies involving rare variants have been developed [436]. As described in the previous paragraph, identification of rare high-risk variants in genomic regions through GWAS or in several disease susceptibility genes can be adopted as a successful approach.

Pharmacogenetics

Many people have severe consequences due to adverse drug reactions. Pharmacogenetic is the discipline that study the interaction between drugs and a specific or multiple genes. Since drug responses may be genotype-driven, the discovery of biomarkers that can predict the responsiveness to the drug may have a very real diagnostic value. This approach lead to the field of "personalized medicine" and can be very promise for improve health outcomes, including those related to complex disease

One of the best example of pharmacogenetic association is computing the warfarin dose in treatment and prevention of thromboembolitis. As higher doses of this drug cause bleeding, the therapy is made using the appropriate dose according to several factors including age, gender, weight, diet. Furthermore, two genes have been implicated in the determination of the warfarin dose, CYP2C9 and VKORC1. They are used as predictors of dosing since mutant alleles of these genes are associated with increased bleeding [446].

9.5.1. Method

For this challenge we had to predict the phenotypes of ten individuals submitting the probability of a person having a phenotype among 32 proposed. There were also numerical traits (e.g. HDL and LDL level in mg/dL) for which we had to predict the mean value found in each individual. The list of phenotypes consists of several pathophysiological conditions including Mendelian disorders or traits and complex diseases (Fig. 9.17).

Phenotype list for the CAGI 2011 PGP challenge

Binary traits
1 Asthma *
2 Crohn's disease
3 Ulcerative colitis
4 Irritable bowel syndrome
5 Rheumatoid arthritis
6 Type II Diabetes
7 Coronary artery disease
8 Long QT Syndrome
9 Hypertrophic cardiomyopathy
10 Glaucoma *
11 Color blindness
12 Bipolar disorder
13 Celiac disease
14 Psoriasis
15 Lupus
16 Breast cancer
17 Prostate cancer
18 Migraine *
19 Lactose intolerance
20 Dyslexia
21 Autism
22 Osteoporosis
23 Incontinence
24 Kidney stones
25 Varicose veins
26 Sleep Apnea
27 Tongue rolling (tube)
28 Phenylthiocarbamide tasting
29 Blood type - Has A antigen?
30 Blood type - Has B antigen?
31 Blood type - Is Rh (D) positive?
32 Absolute pitch

Numerical traits
33 Birth weight (in g)
34 HDL level (in mg/dL) *
35 LDL level (in mg/dL) *
36 Triglyceride level (in mg/dL) *
37 Fasting blood glucose level (in mg/dL)
38 Warfarin dose (in mg)
39 Age at Menarche
40 Annual income (in $)

Figure 0.17. **Phenotype list for the PGP challenge.**
The phenotypes to predict include both Mendelian and complex disorders, and numerical traits.

The approach we adopted for the PGP challenge was the same used to predict individuals with Crohn's disease (one of the phenotypes to predict was again Crohn's disease) (Fig. 9.10). In this case, we downloaded a list of candidate genes from GWAS and OMIM for each trait or phenotype and searched for those that presented potentially causal mutations identified by ANNOVAR in each individual (list L0). The variant reduction protocol was again used, putting a threshold of MAF>0.01 for variants

observed in the 1000 Genome Project and reported in dbSNP (build 130). This allowed us to include variants that could be known pathogenic mutations reported in these databases (list L3). We expanded both lists L0 and L3 including genes identified by ANNOVAR which interact with disease associated genes (list L1 and L4 respectively). In cases where a specific gene is known to determine the particular phenotype, we used information retrieved from the literature. For example, blood type (antigen A and B, Rh antigen) was predicted searching for mutations in the ABO and RHD genes respectively, as these are specific for each group [447]. Finally, the assessment of association between an individual and a phenotype was performed analyzing the collective effects of rare variants across one or multiple genes. Prior evidence about variants (e.g. known pathogenic mutations) and their severity (missense, frameshift or stop-gain) were incorporated into a probabilistic score taking into account the known genes and their interactors (Fig. 9.17).

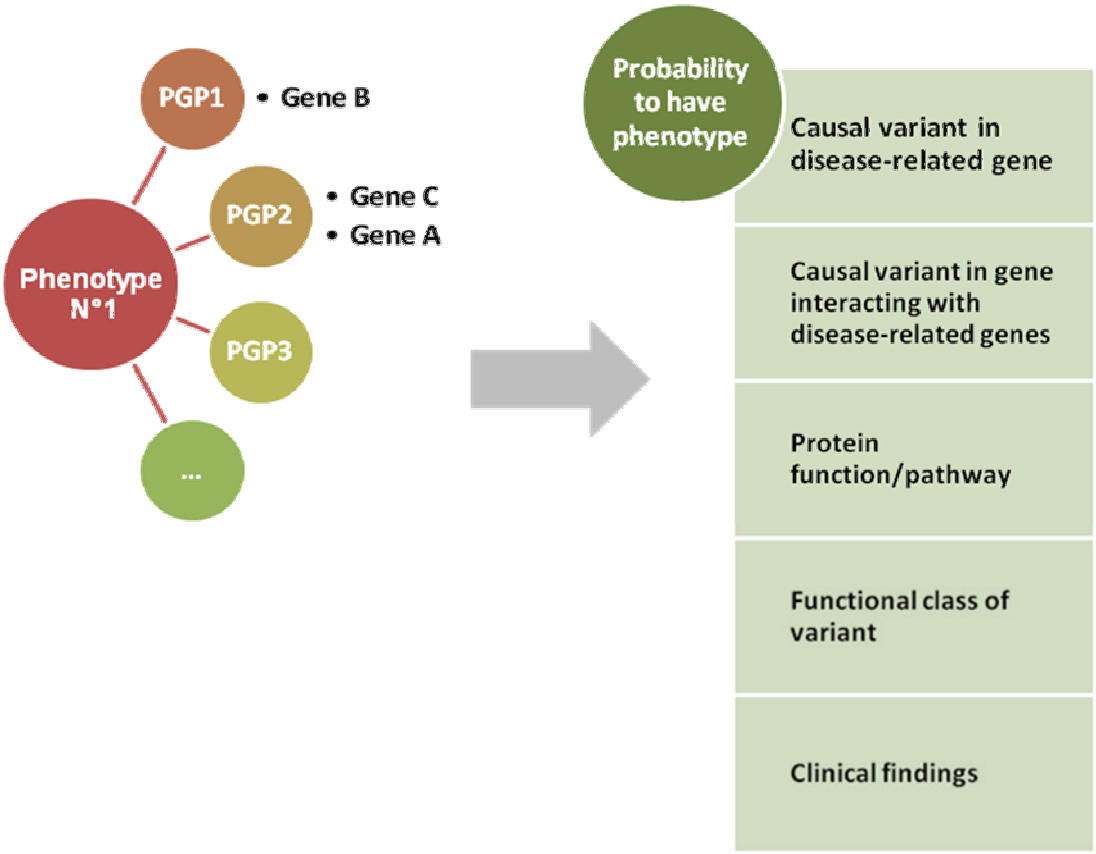


Figure 0.18. Strategy adopted for PGP challenge.

Numerical traits were estimated using both probabilistic scores from causal variants identified in genes known to regulate each trait and the normal range we expected to find in a human individual. We did not find mutations in genes known to regulate Birth weight, Fasting blood glucose level, and Age at Menarche. These values were predicted using a probabilistic score derived from Type II Diabetes. Low birth weight is tightly associated with high risk of Type II Diabetes and also with earlier Menarche [448-449]. Age at Menarche was calculated as the mean age at menarche (12.7 years) minus the genetic index from Diabetes. The birth weight was calculated as the mean weight at birth (3,500 g) minus the product of genetic index and the range of normal weight variance. The normal fasting blood glucose level is 83mg/dL with a normal range of 60-100 mg/dL. The values for each individual were calculated adding to the normal level the product of a genetic index and the glucose level variance. HDL, LDL and triglyceride levels were predicted combining genetic indexes for genes associated to dyslipidemia, for age, and the mean normal level for each numerical trait. Warfarin dose was calculated using the optimal pharmacogenetics algorithm that estimated the daily warfarin dose (mg/day), which was: exp[0.9751 − 0.3238 × VKOR3673G>A + 0.4317 × BSA − 0.4008 × CYP2C9*3− 0.00745 × age − 0.2066 × CYP2C9*2+ 0.2029 × target INR − 0.2538 x amiodarone + 0.0922 × smokes − 0.0901 × African-American race + 0.0664 × DVT/PE], where the SNPs are coded 0 if absent, 1 if heterozygous, and 2 if homozygous, and race is coded as 1 if African American and 0 otherwise [446]. Finally, the annual income was predicted combining a minimum wage with an index for weak (e.g. missense) and strong (e.g. frameshift) variants and an age index.

9.5.2. Results and Discussion

Our group submitted predictions for 38 of the phenotypes/traits present in the list. For tongue rolling and phenylthiocarbamide tasting we did not find information about related genes. We predicted a high probability of having the disease in at least one individual for 19 phenotypes/traits. These results indicated that our prediction may have a high rate of false positives, but we decided to maintain a low threshold in order to avoid loss of information. The organizers provided the phenotypes of each PGP individuals on the basis of their answers for each phenotype or trait (Table 9.5). None of

them answered about the annual income. We were able to predict correctly the individual with IBS, a female with osteoporosis, and the men with colour blindness. For the numerical trait, the challenge was more difficult but we identified the value within a standard deviation for some of the individuals (Table 9.6).

Individual	Binary traits/phenotype
PGP9, PGP10	Asthma
PGP10	Glaucoma
PGP7	Irritable Bowel Syndrome (IBS)
PGP3	Osteoporosis
PGP6	Lactose intolerance
PGP10	Colour blindness
PGP1	Dyslexia
PGP1	Sleep apnea

Table 0.5. Phenotype for the PGP individuals.
Only some disease or trait was observed in the ten individuals. In red are highlighted those that we predicted correctly.

Numerical traits	Correct prediction
Birthday Weight	PGP2, PGP4, PGP5, PGP7
HDL	PGP2
LDL	PGP1, PGP9
TG	PGP2, PGP9
Blood Glucose	PGP2, PGP3, PGP5
Warfarin dose	No one uses the drug
Age at menarche	PGP9 (only 2 females)
Annual income	?

Table 0.6. Correct predictions for numerical traits.

Comparison with other participating groups

Only other two groups participated to this challenge (Fig. 9.18). One of these submitted statistically significant predictions, with a precision of 0.652 and P-value equal to 0. This group (R. Karchin of Johns Hopkins University) was the only group participating in the PGP competition in the pre-pro-CAGI-2010 meeting. They improved their

predictions using a more stringent threshold to assign each phenotype/trait to an individual. Besides using a computational model based on a Bayesian network and GWAS databases, they calculated the probability of having a specific phenotype, using information derived from manual online literature search (e.g. for birth weight [448]). Our group adopted a different computational approach as described above, but for many of the phenotypes/traits to predict, also in our case, the manual online literature search was the only source on which to make predictions. It is interesting to note that the best predictions of true positives were all obtained by the manual method, while the computational model worked well in identifying the true negatives which were very frequent in this set of individuals. This observation suggests that computational approaches are still far from the solution of the personal genome project and that there is a lack of annotations linking specific variants, genes, and pathways to phenotypes. In the words of the assessor, we are still in the "game phase". Some simpler phenotypes are nevertheless already predictable from haplotypes or SNPs.

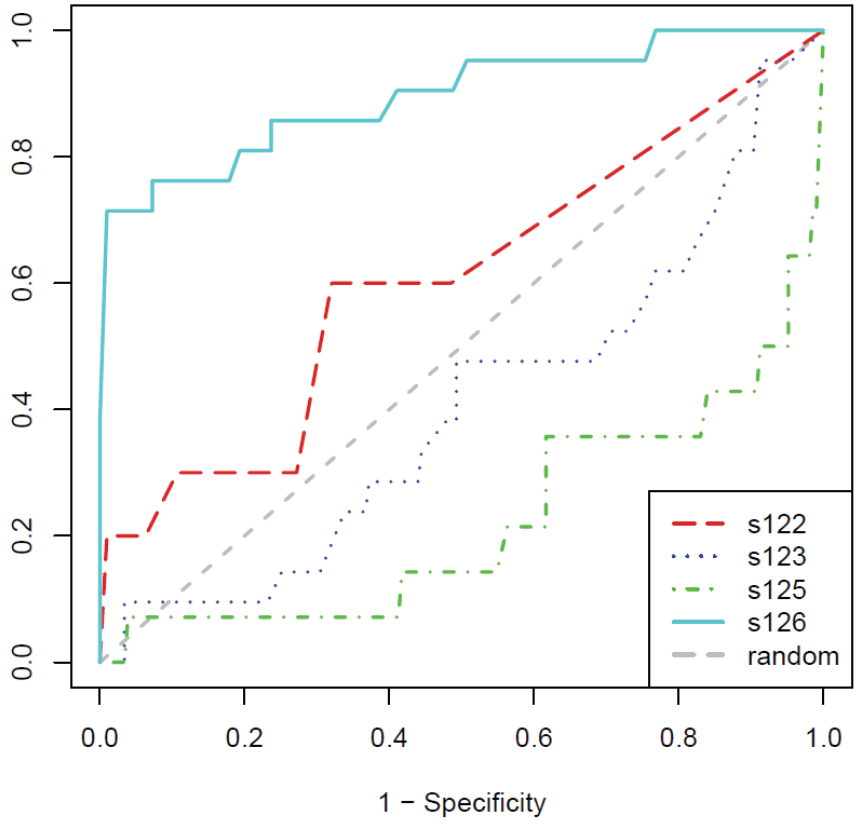


Figure 0.19. ROC curve for binary traits.
The submissions s122 and s123 refer to our group, with s123 obtained with 100 side die. (Figure provided by CAGI assessor S. Mooney)

For numerical traits, the assessment was based on the average distance to the correct answer for each trait and the number of times a group predicted within one standard deviation and z-score results. This task was very hard and we performed unexpectedly well for three of the numerical traits considering that the other groups did not obtain good predictions (Table 9.7).

The birth weight was correctly predicted for 3 individuals, the triglyceride level for one, and the age at menarche for the two women in the group. The question emerging from this challenge is how to use information about other sources of influencing factors, such as those deriving from environment, habits of the individuals, diet, and current treatments. This needs further studies in case-control populations.

Number of PGPs each submission got values within the range (prediction +/- standard deviation):				
	UniPadova	UniPadova	Netbiolab	KarchinLab
Birth weight (in g)	3	3	1	1
HDL level (mg/dL)	0	0	1	0
LDL level (mg/dL)	1	2	2	0
Triglyceride level (mg/dL)	1	1	0	0
Fasting blood glucose level (mg/dL)	3	1	1	3
Warfarin dose (mg)	NA	NA	NA	NA
Age at menarche	2	0	0	0
TOTAL	10	7	5	4

Table 0.7. Number of correct predictions.
In this table the prediction was considered correct when was in the range of the reported Standard Deviation. (Table provided by CAGI assessor S. Mooney)

9.6. Conclusions

Participation in the CAGI experiment allowed us to focus on possible new approaches for mutation effect prediction and to test their performance in a set of variants for which protein function was experimentally tested. Our group developed a new method based on residue interaction methods to predict the impact of missense mutations on protein

structure. This approach is innovative and worked similarly well to the common methods for mutation analysis. In this experiment, it has been possible to highlight the strengths and weakness of the method. From the CAGI experiment one important thing which emerges is that a consensus of prediction methods results in a better prediction of mutations. We used a consensus method for the prediction of intermediate-risk variants of RAD50 and SCN5A mutations, but the results are weaker than PON-P which also uses the consensus of several methods. The difference may be due to the combination of used methods. However, even in the case of PON-P the bottleneck in the prediction is the output format of the different methods and the time of calculation. One of the major remaining challenges is to developed a fast prediction method for the interpretation of huge amounts of genetic variations.

This experiment further highlighted the relevance of a priori knowledge about structure and function of the target protein. In our method, we were able to fix different parameters on the basis of information collected from the literature and structural analysis. My experience on the *in silico* analysis of proteins of different structural and functional classes (e.g. transmembrane proteins, repeat proteins, proteins containing modular domains, kinases, DNA binding proteins) was particularly useful for the participation in the p53, RAD50 and SCN5A challenges.

We also tried to work with data generated from next generation sequencing technologies, which is the current effort in mutation research. Since this is a new field, few research groups have experience in this, especially for complex disease prediction. Here, we introduced new strategies to prioritize candidate genes for several phenotypes. The vision of each biological process as a complex system in which a protein works and exerts its numerous functions together with other protein partners allowed us to introduce a novel approach using functional protein networks for the analysis of genetic variations. Also, in this case the experience in interpretation of genetic rare disease variants has been useful to tackle the prediction of complex disease risk. This highlights how a group with differentiated expertise is especially important for the development of methods for genome interpretation.

10. Conclusions

During my project work I analyzed several proteins of biological and biomedical interest using computational tools. The structural and functional insights obtained through these analyses were used to better understand the molecular mechanisms underlying protein function and to identify a possible genotype-phenotype correlation. Knowledge of the protein structure, either experimental or through modelling, provides insights which can be used to pinpoint finer details, such as the domains or segments of the proteins essential for its biological activity.

The experimental structure of the DNA binding domain of the WT1 transcription factor was analyzed to identify residues involved in binding DNA and to evaluate the effect of different isoforms on the transcription process. The protein structure can also contain more binding sites which mediate interactions with different proteins in order to form the correct protein complex in a specific cellular compartment. In particular, study of the VHL protein aimed to investigate its role in different signaling pathways involved in tumor formation. Since more than 200 interactors have been experimentally identified, I hypothesised that while some proteins interact simultaneously, some other must be mutually exclusive. The structural and functional analysis of 35 VHL interactors with experimentally determined pVHL interaction regions allowed the characterization of three interaction interfaces corresponding to processing, substrate recognition and localization. These interfaces highlight diverse protein interaction types, namely domain-domain (interface A) and domain-peptide (interface B), with interface C being less clear. In particular for interface B it has been possible to better define a hypothetical interaction motifs that can be used to validate the growing number of proteins found to interact with pVHL by high-throughput methods.

When the experimental structure is not available, the model of the protein can be useful to explore the spatial arrangement of known functional residues or to identify possible binding sites with other proteins. For the LGI1 protein I adopted a remote homology modelling approach to model the two repeated LRR and beta-propeller domains. Repeat proteins are difficult to model using a homology modelling approach since they show

poor sequence conservation. In this case, manual refinement of the target-template alignment was crucial. This approach used the ABRA protocol [330] and Kajava's method [95], which suggested to use knowledge of key residues and secondary structure to anchor the aligned repeats. Analysis of the model surface properties suggested a possible arrangement between the two domains and identifies possible protein binding sites in both the β-propeller and leucine-rich repeat domains. The three dimensional *in silico* model of LGI1 allowed the creation of a functional model integrating previous experimental findings and suggesting a possible molecular mechanism involved in the synaptic transmission of neural signals.

Transmembrane proteins are another class of proteins for which structure prediction requires a more complex approach than for soluble proteins. Although progress is hampered by a limited amount of high-resolution experimental 3D structures, the overall prediction of functional and structural features of transmembrane proteins is improving. When few or no experimental data is available, no present method can accurately predict the 3D-structure of any transmembrane protein from sequence alone [450]. However, we can apply computational methods which predict transmembrane segments and their topology through knowledge-based approaches. In the case of the POMT1 protein, modelling of the protein is very difficult as homologous structures are available only for the MIR domain, the catalytic domain protruding from the transmembrane domain on the cytoplasmic side. The transmembrane region of POMT1 was therefore analysed using a consensus of different topology prediction methods. Such an ensemble method consists of taking the output of individual predictors and combining them by majority vote. The ensemble yields a better prediction than each individual method and tends to cancel out the errors, combining the advantages of the different methods. When a high resolution structure of a homologue to the transmembrane protein is available, we can use homology modelling to obtain a prediction of the protein structure. The prediction of transmembrane topology for Plasma Membrane Ca2+ Pump isoform 2 (PMCA2) has been performed taking into account a manually curated alignment with the sarco/endoplasmic Ca2+ - ATPase (SERCA1a). The alignment was subsequently used as template to build a model with standard homology modelling. An analysis of the model surface properties allowed the characterization of a positively charged region which plays a crucial role in the pump

activation mechanism through interaction with phospholipids (PL) on the cellular membrane. This hypothesis has been confirmed experimentally by measuring the PL sensitivity of mutant proteins derived by introducing mutations at four positions normally occupied by conserved lysines responsible for the positive charge observed in this PL binding region.

An in depth analysis of the proteins known to be associated to diseases is useful in particular for the evaluation of the pathogenicity of novel variants. This is essential for guiding medical decisions on treatment and follow up. The use of structural information should improve the prediction since most of the known disease causing variants have been found to destabilize protein structure [16-18]. However, detailed analysis of the known proteins will serve to elucidate the single pieces involved in the regulatory network at the molecular level, in order to formulate hypotheses that may explain the genotype to phenotype correlation of the involved genes. In particular, the effects of single missense mutations will be evaluated, as well as for protein stability changes, on their impact with the interaction partners. In my thesis, I analysed the effects of mutations in different disease associated proteins. Structural analysis was used in combination with genetic information to established the role of previously uncharacterized variants. In particular I analyzed the impact of novel WT1 and POMT1 variants associated to atypical clinical findings. The first study focused on the evaluation of a new WT1 mutation found in three family members with focal segmental glomerulosclerosis but without genital abnormalities or Wilms tumor. *In silico* analysis of the mutant model for the DNA binding domain of WT1 indicated that the functional impact caused by the mutation results in a reduced ability of the KTS-positive isoforms to bind DNA. Novel mutations of the glycosyltransferases POMT1 were found in patients with muscular dystrophy which developed dilated cardiomyopathy. Computational analysis, consistent with *in vitro* enzymatic assays, predicted these novel variants as disease causing mutations and demonstrated how this approach can be especially useful in laboratories where experimental assays are unavailable. Since no particular characteristic has been found to distinguish the novel mutations from other known mutations, cardiac involvement should be considered in the phenotypical spectrum associated to POMT1 mutations. The hypothesis is that patients carrying a mutant protein with residual enzymatic activity may present a different pattern and

timing of multisystemic involvement. Similar conclusions were derived from the *in silico* analysis of 18 variants found for the first time in subjects with VHL syndrome. These variants were classified as structural or functional mutations on the basis of their impact on the protein fold or in interfering with interaction interfaces, respectively. Our approach allowed to improve the ability to predict the risk of pheochromocytoma, which seems to be caused by mutant proteins with residual functions.

In order to analyse the finer details of the perturbation on the protein fold caused by amino acid substitutions, I applied for the first time the residue interaction network (RIN) analysis. Some interactions such as salt bridges or disulfide bonds need a pair of residues with specific physico-chemical characteristics. Thus, if one of these changes, we can predict that this interaction will be lost. These considerations are usually derived by visual inspection of the structure with a molecular visualization tool. The RIN approach simplifies identification of the diverse intra-residue interactions that each residue undergoes in 3D space and in prediction of the local or global structural effects caused by an amino acid substitution.

Missense mutations of the VHL gene were also analysed using a set of seven well known prediction methods, with the final decision about the pathogenicity of the variant taken combining the output of individual predictors. The ensemble method predicted one of the novel variants as neutral. This finding was supported by available genetic information of the family which reported the same variant in the unaffected father. The strategy to use a combination of several prediction methods was not used in previous works, but in the last year appears to have become the best approach to obtain an accurate prediction of mutation pathogenicity.

Finally, the same overall approach was applied for the analysis of LGI1 novel mutations. In this case, the analysis was performed using the structural models of the two repeat domains which allowed classification of the variants in two classes. The structural variants have the potential to destabilize the protein fold, losing its ability to be secreted. This hypothesis is supported by *in vitro* studies of protein secretion of several mutant proteins. Functional variants, while maintaining the overall protein fold, alter residues located at the protein surface, the details of which may be crucial for interactions with protein partners. I predicted as five variants of the β-propeller domain functional and hypothesised that these variants maintain the ability to be secreted even if function

results affected. One of these has been found to segregate in a family with epilepsy and psychic symptoms in absence of the characteristic auditory phenomena associated with LGI1 mutations [346]. *In vitro* studies further revealed that this mutation does not prevent secretion of the mutant LGI1 protein. Recently, the study of protein secretion has been extended to other functional variants I predicted, confirming the hypothesis that these variants act by using a molecular mechanism which differs from loss-of-function mutations. Furthermore, this may explain the atypical clinical features associated to this class of variants (unpublished data).

All these works demonstrate how using a combination of computational tools and resources available on the web it is possible to conduct an in depth analysis of the structure and function of different proteins and to predict the effects of novel variants involved in the pathogenesis of associated diseases. However, a large part of the analysis requires intervention of a bioinformatician in order to decide which prediction method to use and to evaluate the overall results. In particular, while analyzing mutation effects, genetics and clinical information have to be considered. The Critical Assessment of Genome Interpretation (CAGI) experiment aims to assess computational methods for the prediction of phenotypic impacts caused by genomic variations. The interesting point is that participants have to make blind predictions of genetic variants for which the molecular, cellular, or organismal phenotype is already known but unpublished. The predictions were evaluated subsequently by independent assessors in order to understand which method performed better compared to the others. The goals of this experiment are to evaluate the performance of state-of-the art methods and to foster the creation of innovative software in the prediction of mutation effect.

Our participation in the CAGI experiment allowed the development of two new mutation prediction approaches for different applications, a method using residue interaction network and an ensemble prediction approach. The CAGI challenges were particularly difficult, since they represent specific applications differing from the simple prediction of variant pathogenicity. One aimed to identify mutations that can reactivate p53 function in cancer mutants, while another deals with determining the probability of a variant to predispose to cancer in a intermediate-risk breast cancer gene. However, these experiments highlighted that each application could be addressed by a proper software which fits better to the evaluation of the molecular mechanisms involved in the

alterations. For example, in the p53 challenge the stability change of the protein is the feature which improves the prediction of rescue mutants. Our approach evaluated this change by using residue interaction network data based on the idea that the stability of the protein is coupled to its correct folding. The three dimensional fold of the protein is determined by chemical bonds and interactions between amino acid side chains. Thus, alterations causing changes in interaction energy between amino acids may affect the free energy difference of the folded and unfolded states of the protein. The application of this method to the identification of p53 rescue mutants highlighted that the protein may have different stable conformations which can be predicted using different classes of known rescue mutants. We can take advantage from the fact that rescue mutants use specific molecular mechanisms for cancer mutations causing similar structural effects. However, the method works as well or even better than commonly used prediction methods for the identification of pathogenic mutations. Future work will be to improve its accuracy in this field.

Another approach we applied to the CAGI experiment is to use an ensemble prediction method based on the previously described approach I adopted for the analysis of mutations in proteins associated to disease. The limitations of this approach depend on the available structural information of the target protein and, especially in the automation step, on the availability of the different methods. For the RAD50 challenge it was possible to use only methods using sequence information since it was difficult to obtain a model of the entire protein sequence. The first obstacle is to combine different computation times and output formats from the different methods which also require different input data. Furthermore, the final decision process should be benchmarked and improved. For RAD50 the best results were obtained combining the prediction of the ensemble method with those obtained by structural functional analysis of the protein. In this case, adding a priori information derived from experimental data reported in literature or from the functional analysis of the protein also improves accuracy of the prediction.

These approaches can be very difficult to use when we have to predict the phenotype starting from the exome sequences which can contain millions of variants. Recent improvements in large-scale genotyping arrays and of sequence technologies promise to provide DNA tests of genetic markers for a myriad of different diseases/traits and to

gain a large number of personal genomes within the next years. The management of personal genomes has diverse medical, ethical, legal and also technical limitations. Nevertheless, analysis of exome or whole genome data have been successfully used to discover the genetic basis of several diseases that were not identified by traditional genetics for decades (for a review see [4]). The "omics" fields have introduced new approaches of investigation, moving from hypothesis-testing to discovery-based approaches. The discovery-based approach allows the generation and prioritization of high-throughput hypotheses. The CAGI experiment proposed two challenges aimed to predict phenotypes starting from exome sequences. One of these required to distinguish between exomes of healthy individuals from those with Crohn's disease, while the other challenge was to predict the possible phenotypes of ten individual from a list of 40 diseases or traits including rare and common diseases, and several numerical traits such as blood lipid levels. In order to deal with these situations I designed a computational model which can be used in the prediction of diverse diseases of different nature such as Mendelian and complex diseases. The aim was also to use methods and resources available on the web.

The first step of the model was to identify a set of variants predicted to cause a strong functional impact on protein products. For this, I chose the "variant reduction" protocol of ANNOVAR, a tool for functional annotation and filtering of variants detected from genomic sequences. This tool filters millions of variants, removing variations that are either not conserved or previously reported in public SNP databases, such as dbSNP and the 1000 Genome Project. In this way, we obtain a list of about twenty genes containing rare variants likely to have a phenotypic impact. This approach has been previously adopted to analyse genomes of individuals with rare diseases for which a genetic cause had not yet been identified [4, 6-7]. The choice to consider only rare variants for the analysis of complex diseases was based on the assumption that rare causal variants should have a stronger impact on the development of the disease compared to common variants. The good results on the prediction of Crohn's disease patients obtained by this model compared to others that considered only common variants associated to disease confirmed that the hypothesis was successful.

In both Mendelian and complex diseases some genes or genomic regions are known to be associated with the disease, thus the analysis of the exome was focused on these

regions. However, since for complex diseases the association of variants mapping to these regions explains only in part the pathogenesis of the disease, the list of candidate genes was expanded using protein interaction network information. The main idea is that causal variants may act in a set of proteins involved in the same biological process. This approach allowed us to obtain good predictions on the identification of Crohn's disease patients. The accuracy of the predictions was also improved for other approaches that used, besides the GWAS loci, additional information of related genes. With this approach we also expanded the possibility to discover novel molecular mechanisms involved in the pathogenesis of the disease. In particular, for Crohn's disease we identified a gene, Tyrosine-protein phosphatase non-receptor type 11 (PTPN11), which was mutated in 40 of the 42 patients presenting Crohn's disease. Due to the erroneous expectation that a lower number of individuals in our data set should present the illness, we did not consider this data for CAGI. An a posteriori evaluation suggests that this gene may be involved in the disease pathogenesis. Mutations in PTPN11 are associated with several disorders: Leopard syndrome type 1 (MIM: 151100), Noonan syndrome type 1 (MIM: 163950), juvenile myelomonocytic leukemia (JMML) (MIM: 607785), metachondromatosis (MC) (MIM: 156250). However, further experimental and clinical evidence is necessary to confirm this hypothesis.

Knowledge derived from function and structural analysis of the proteins encoded by the candidate genes has also been useful in the interpretation of exome sequences, especially to calculate the contribution of each causal variant identified in the set of candidate genes. This emerges from the results obtained from what we called the manual method. This represents the arbitrary decision on the existing disease association considering all available information such as the type of identified variants, their prevalence in the analyzed set and in the normal population, the protein function and biological processes involved. This suggests that we need to improve the weighting method to get the correct probability of disease association integrating different kinds of parameters. Furthermore, especially in the Personal Genome Project challenge, many of the correctly predicted phenotypes or traits were predicted manually by searching information from the literature and analyzing the known associated genes or SNPs. For example, the individual with colour blindness was identified by looking for variants in CNGB3, one of the six genes that are reported to cause this phenotype. [451]. This

highlights how there is a lack of annotations linking specific mutations, genes and pathways to the associated phenotypes.

At the CAGI meeting some solutions were discussed to improve the ability to predict phenotypes from genome data. It has been highlighted that CAGI is an experiment and not a competition which has a scientifically benefit to predictors. This year about 50 groups from seven different countries participated in diverse competitions. This will be the main challenge for future research on human genome interpretation and scientists are working together towards this aim. Some companies provide services already available on the web (e.g. http://www.decodeme.com/) which offer the complete scan of the genome for a few thousand dollars. They allow to get to know our own genome and learn how we can use it to improve our health. Progress in this field is improving fast, but we still need to understand which is the most appropriate approach to address this problem and how to use and present these results to the interested individuals.

Bibliography

1. Li, C., *Personalized medicine - the promised land: are we there yet?* Clin Genet, 2011. **79**(5): p. 403-12.
2. Tanaka, H., *Omics-based medicine and systems pathology. A new perspective for personalized and predictive medicine.* Methods Inf Med, 2010. **49**(2): p. 173-85.
3. Lander, E.S., *Initial impact of the sequencing of the human genome.* Nature, 2011. **470**(7333): p. 187-97.
4. Ku, C.S., N. Naidoo, and Y. Pawitan, *Revisiting Mendelian disorders through exome sequencing.* Hum Genet, 2011. **129**(4): p. 351-70.
5. Ng, S.B., et al., *Massively parallel sequencing and rare disease.* Hum Mol Genet, 2010. **19**(R2): p. R119-24.
6. Bamshad, M.J., et al., *Exome sequencing as a tool for Mendelian disease gene discovery.* Nat Rev Genet, 2011. **12**(11): p. 745-55.
7. Maxmen, A., *Exome sequencing deciphers rare diseases.* Cell, 2011. **144**(5): p. 635-7.
8. Kuhlenbaumer, G., J. Hullmann, and S. Appenzeller, *Novel genomic techniques open new avenues in the analysis of monogenic disorders.* Hum Mutat, 2011. **32**(2): p. 144-51.
9. Raychaudhuri, S., *Mapping rare and common causal alleles for complex human diseases.* Cell, 2011. **147**(1): p. 57-69.
10. Fernald, G.H., et al., *Bioinformatics challenges for personalized medicine.* Bioinformatics, 2011. **27**(13): p. 1741-8.
11. Chakravarti, A., *It's raining SNPs, hallelujah?* Nat Genet, 1998. **19**(3): p. 216-7.
12. Syvanen, A.C., et al., *First International SNP Meeting at Skokloster, Sweden, August 1998. Enthusiasm mixed with scepticism about single-nucleotide polymorphism markers for dissecting complex disorders.* Eur J Hum Genet, 1999. **7**(1): p. 98-101.
13. Krawczak, M., et al., *Human gene mutation database-a biomedical information and research resource.* Hum Mutat, 2000. **15**(1): p. 45-51.
14. Stenson, P.D., et al., *Human Gene Mutation Database (HGMD): 2003 update.* Hum Mutat, 2003. **21**(6): p. 577-81.
15. Bross, P., et al., *Protein misfolding and degradation in genetic diseases.* Hum Mutat, 1999. **14**(3): p. 186-98.
16. Wang, Z. and J. Moult, *SNPs, protein structure, and disease.* Hum Mutat, 2001. **17**(4): p. 263-70.
17. Yue, P., Z. Li, and J. Moult, *Loss of protein structure stability as a major causative factor in monogenic disease.* J Mol Biol, 2005. **353**(2): p. 459-73.
18. Allali-Hassani, A., et al., *A survey of proteins encoded by non-synonymous single nucleotide polymorphisms reveals a significant fraction with altered stability and activity.* Biochem J, 2009. **424**(1): p. 15-26.

19. Thusberg, J. and M. Vihinen, *Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods.* Hum Mutat, 2009. **30**(5): p. 703-14.
20. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.
21. *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
22. Tosatto, S.C. and S. Toppo, *Large-scale prediction of protein structure and function from sequence.* Curr Pharm Des, 2006. **12**(17): p. 2067-86.
23. Tramontano, A., *The role of molecular modelling in biomedical research.* FEBS Lett, 2006. **580**(12): p. 2928-34.
24. Beltrao, P., C. Kiel, and L. Serrano, *Structures in systems biology.* Curr Opin Struct Biol, 2007. **17**(3): p. 378-84.
25. Ideker, T. and R. Sharan, *Protein networks in disease.* Genome Res, 2008. **18**(4): p. 644-52.
26. Oti, M. and H.G. Brunner, *The modular nature of genetic diseases.* Clin Genet, 2007. **71**(1): p. 1-11.
27. Goh, K.I., et al., *The human disease network.* Proc Natl Acad Sci U S A, 2007. **104**(21): p. 8685-90.
28. Teng, S., E. Michonova-Alexova, and E. Alexov, *Approaches and resources for prediction of the effects of non-synonymous single nucleotide polymorphism on protein function and interactions.* Curr Pharm Biotechnol, 2008. **9**(2): p. 123-33.
29. Ng, P.C. and S. Henikoff, *Predicting the effects of amino acid substitutions on protein function.* Annu Rev Genomics Hum Genet, 2006. **7**: p. 61-80.
30. Jordan, D.M., V.E. Ramensky, and S.R. Sunyaev, *Human allelic variation: perspective from protein function, structure, and evolution.* Curr Opin Struct Biol, 2010. **20**(3): p. 342-50.
31. Lee, W., P. Yue, and Z. Zhang, *Analytical methods for inferring functional effects of single base pair substitutions in human cancers.* Hum Genet, 2009. **126**(4): p. 481-98.
32. Capriotti, E. and R.B. Altman, *Improving the prediction of disease-related variants using protein three-dimensional structure.* BMC Bioinformatics, 2011. **12 Suppl 4**: p. S3.
33. Rivas, M.A., et al., *Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease.* Nat Genet, 2011. **43**(11): p. 1066-73.
34. Cooper, G.M. and J. Shendure, *Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data.* Nat Rev Genet, 2011. **12**(9): p. 628-40.
35. Le Calvez-Kelm, F., et al., *Rare, evolutionarily unlikely missense substitutions in CHEK2 contribute to breast cancer susceptibility: results from a breast cancer family registry case-control mutation-screening study.* Breast Cancer Res, 2011. **13**(1): p. R6.
36. *Reorganizing the protein space at the Universal Protein Resource (UniProt).* Nucleic Acids Res, 2011.

37. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
38. Gilks, W.R., et al., *Modeling the percolation of annotation errors in a database of protein sequences.* Bioinformatics, 2002. **18**(12): p. 1641-9.
39. Schneider, A., C. Dessimoz, and G.H. Gonnet, *OMA Browser--exploring orthologous relations across 352 complete genomes.* Bioinformatics, 2007. **23**(16): p. 2180-2.
40. Schomburg, D. and I. Schomburg, *Enzyme databases.* Methods Mol Biol, 2010. **609**: p. 113-28.
41. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res, 1994. **22**(22): p. 4673-80.
42. Katoh, K., et al., *MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.* Nucleic Acids Res, 2002. **30**(14): p. 3059-66.
43. Clamp, M., et al., *The Jalview Java alignment editor.* Bioinformatics, 2004. **20**(3): p. 426-7.
44. Guindon, S. and O. Gascuel, *A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.* Syst Biol, 2003. **52**(5): p. 696-704.
45. Ashkenazy, H., et al., *ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W529-33.
46. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
47. Almonacid, D.E. and P.C. Babbitt, *Toward mechanistic classification of enzyme functions.* Curr Opin Chem Biol, 2011. **15**(3): p. 435-42.
48. Marchler-Bauer, A., et al., *CDD: a Conserved Domain Database for the functional annotation of proteins.* Nucleic Acids Res, 2011. **39**(Database issue): p. D225-9.
49. Finn, R.D., et al., *The Pfam protein families database.* Nucleic Acids Res, 2010. **38**(Database issue): p. D211-22.
50. Letunic, I., et al., *SMART 5: domains in the context of genomes and networks.* Nucleic Acids Res, 2006. **34**(Database issue): p. D257-60.
51. Tatusov, R.L., et al., *The COG database: an updated version includes eukaryotes.* BMC Bioinformatics, 2003. **4**: p. 41.
52. Hunter, S., et al., *InterPro: the integrative protein signature database.* Nucleic Acids Res, 2009. **37**(Database issue): p. D211-5.
53. Tian, W., A.K. Arakaki, and J. Skolnick, *EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference.* Nucleic Acids Res, 2004. **32**(21): p. 6226-39.
54. Henikoff, J.G., et al., *Blocks-based methods for detecting protein homology.* Electrophoresis, 2000. **21**(9): p. 1700-6.
55. Attwood, T.K., et al., *PRINTS and its automatic supplement, prePRINTS.* Nucleic Acids Res, 2003. **31**(1): p. 400-2.

56. Gnad, F., J. Gunawardena, and M. Mann, *PHOSIDA 2011: the posttranslational modification database.* Nucleic Acids Res, 2011. **39**(Database issue): p. D253-60.
57. Puntervoll, P., et al., *ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins.* Nucleic Acids Res, 2003. **31**(13): p. 3625-30.
58. Nakai, K. and P. Horton, *PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization.* Trends Biochem Sci, 1999. **24**(1): p. 34-6.
59. Nielsen, H., et al., *Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.* Protein Eng, 1997. **10**(1): p. 1-6.
60. McGuffin, L.J., K. Bryson, and D.T. Jones, *The PSIPRED protein structure prediction server.* Bioinformatics, 2000. **16**(4): p. 404-5.
61. Pollastri, G. and A. McLysaght, *Porter: a new, accurate server for protein secondary structure prediction.* Bioinformatics, 2005. **21**(8): p. 1719-20.
62. Krogh, A., et al., *Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.* J Mol Biol, 2001. **305**(3): p. 567-80.
63. Tusnady, G.E. and I. Simon, *The HMMTOP transmembrane topology prediction server.* Bioinformatics, 2001. **17**(9): p. 849-50.
64. Kall, L., A. Krogh, and E.L. Sonnhammer, *A combined transmembrane topology and signal peptide prediction method.* J Mol Biol, 2004. **338**(5): p. 1027-36.
65. Kall, L., A. Krogh, and E.L. Sonnhammer, *An HMM posterior decoder for sequence feature prediction that includes homology information.* Bioinformatics, 2005. **21 Suppl 1**: p. i251-7.
66. Bernsel, A., et al., *TOPCONS: consensus prediction of membrane protein topology.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W465-8.
67. Viklund, H. and A. Elofsson, *OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar.* Bioinformatics, 2008. **24**(15): p. 1662-8.
68. Viklund, H. and A. Elofsson, *Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information.* Protein Sci, 2004. **13**(7): p. 1908-17.
69. Bernsel, A., et al., *Prediction of membrane-protein topology from first principles.* Proc Natl Acad Sci U S A, 2008. **105**(20): p. 7177-81.
70. Hessa, T., et al., *Molecular code for transmembrane-helix recognition by the Sec61 translocon.* Nature, 2007. **450**(7172): p. 1026-30.
71. Berman, H.M., et al., *The Protein Data Bank.* Acta Crystallogr D Biol Crystallogr, 2002. **58**(Pt 6 No 1): p. 899-907.
72. Lo Conte, L., et al., *SCOP: a structural classification of proteins database.* Nucleic Acids Res, 2000. **28**(1): p. 257-9.
73. Orengo, C.A., F.M. Pearl, and J.M. Thornton, *The CATH domain structure database.* Methods Biochem Anal, 2003. **44**: p. 249-71.
74. Knudsen, M. and C. Wiuf, *The CATH database.* Hum Genomics, 2010. **4**(3): p. 207-12.

75. Kabsch, W. and C. Sander, *Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.* Biopolymers, 1983. **22**(12): p. 2577-637.
76. Sayle, R.A. and E.J. Milner-White, *RASMOL: biomolecular graphics for all.* Trends Biochem Sci, 1995. **20**(9): p. 374.
77. Herraez, A., *Biomolecules in the computer: Jmol to the rescue.* Biochem Mol Biol Educ, 2006. **34**(4): p. 255-61.
78. Baker, N.A., et al., *Electrostatics of nanosystems: application to microtubules and the ribosome.* Proc Natl Acad Sci U S A, 2001. **98**(18): p. 10037-41.
79. Pettersen, E.F., et al., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem, 2004. **25**(13): p. 1605-12.
80. Morris, J.H., et al., *structureViz: linking Cytoscape and UCSF Chimera.* Bioinformatics, 2007. **23**(17): p. 2345-7.
81. Shindyalov, I.N. and P.E. Bourne, *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path.* Protein Eng, 1998. **11**(9): p. 739-47.
82. Guda, C., et al., *CE-MC: a multiple protein structure alignment server.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W100-3.
83. Konagurthu, A.S., et al., *MUSTANG: a multiple structural alignment algorithm.* Proteins, 2006. **64**(3): p. 559-74.
84. Gouet, P., et al., *ESPript: analysis of multiple sequence alignments in PostScript.* Bioinformatics, 1999. **15**(4): p. 305-8.
85. Gouet, P., X. Robert, and E. Courcelle, *ESPript/ENDscript: Extracting and rendering sequence and 3D information from atomic structures of proteins.* Nucleic Acids Res, 2003. **31**(13): p. 3320-3.
86. Sali, A., et al., *Evaluation of comparative protein modeling by MODELLER.* Proteins, 1995. **23**(3): p. 318-26.
87. Arnold, K., et al., *The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling.* Bioinformatics, 2006. **22**(2): p. 195-201.
88. Tosatto, S.C., et al., *A divide and conquer approach to fast loop modeling.* Protein Eng, 2002. **15**(4): p. 279-86.
89. Tosatto, S.C., *The Victor/FRST Function for Model Quality Estimation.* J Comput Biol, 2005. **12**(10): p. 1316-27.
90. Van Der Spoel, D., et al., *GROMACS: fast, flexible, and free.* J Comput Chem, 2005. **26**(16): p. 1701-18.
91. Ginalski, K., et al., *3D-Jury: a simple approach to improve protein structure predictions.* Bioinformatics, 2003. **19**(8): p. 1015-8.
92. Holm, L. and C. Sander, *The FSSP database of structurally aligned protein fold families.* Nucleic Acids Res, 1994. **22**(17): p. 3600-9.
93. Tusnady, G.E., Z. Dosztanyi, and I. Simon, *PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank.* Nucleic Acids Res, 2005. **33**(Database issue): p. D275-8.
94. Kelm, S., J. Shi, and C.M. Deane, *MEDELLER: homology-based coordinate generation for membrane proteins.* Bioinformatics, 2010. **26**(22): p. 2833-40.
95. Kajava, A.V. and B. Kobe, *Assessment of the ability to model proteins with leucine-rich repeats in light of the latest structural information.* Protein Sci, 2002. **11**(5): p. 1082-90.

96. Kobe, B. and A.V. Kajava, *The leucine-rich repeat as a protein recognition motif.* Curr Opin Struct Biol, 2001. **11**(6): p. 725-32.
97. Marsella, L., et al., *REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform.* Bioinformatics, 2009. **25**(12): p. i289-95.
98. Atchley, W.R., et al., *Solving the protein sequence metric problem.* Proc Natl Acad Sci U S A, 2005. **102**(18): p. 6395-400.
99. Albrecht, M., et al., *Simple consensus procedures are effective and sufficient in secondary structure prediction.* Protein Eng, 2003. **16**(7): p. 459-62.
100. Benkert, P., S.C. Tosatto, and D. Schomburg, *QMEAN: A comprehensive scoring function for model quality assessment.* Proteins, 2008. **71**(1): p. 261-77.
101. Benkert, P., S.C. Tosatto, and T. Schwede, *Global and local model quality estimation at CASP8 using the scoring functions QMEAN and QMEANclust.* Proteins, 2009. **77 Suppl 9**: p. 173-80.
102. Fuxreiter, M., et al., *Preformed structural elements feature in partner recognition by intrinsically unstructured proteins.* J Mol Biol, 2004. **338**(5): p. 1015-26.
103. Tompa, P., *Intrinsically unstructured proteins.* Trends Biochem Sci, 2002. **27**(10): p. 527-33.
104. Tompa, P., *The interplay between structure and function in intrinsically unstructured proteins.* FEBS Lett, 2005. **579**(15): p. 3346-54.
105. Tompa, P., C. Szasz, and L. Buday, *Structural disorder throws new light on moonlighting.* Trends Biochem Sci, 2005. **30**(9): p. 484-9.
106. Tompa, P., Z. Dosztanyi, and I. Simon, *Prevalent structural disorder in E. coli and S. cerevisiae proteomes.* J Proteome Res, 2006. **5**(8): p. 1996-2000.
107. Iakoucheva, L.M., et al., *Intrinsic disorder in cell-signaling and cancer-associated proteins.* J Mol Biol, 2002. **323**(3): p. 573-84.
108. Ward, J.J., et al., *Prediction and functional analysis of native disorder in proteins from the three kingdoms of life.* J Mol Biol, 2004. **337**(3): p. 635-45.
109. Xie, H., et al., *Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins.* J Proteome Res, 2007. **6**(5): p. 1917-32.
110. Li, X., et al., *Predicting Protein Disorder for N-, C-, and Internal Regions.* Genome Inform Ser Workshop Genome Inform, 1999. **10**: p. 30-40.
111. Vullo, A., et al., *Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W164-8.
112. Walsh, I., et al., *CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W190-6.
113. Dosztanyi, Z., et al., *IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.* Bioinformatics, 2005. **21**(16): p. 3433-4.
114. Dunker, A.K., et al., *Intrinsic disorder and protein function.* Biochemistry, 2002. **41**(21): p. 6573-82.
115. Elbaum, M., *Materials science. Polymers in the pore.* Science, 2006. **314**(5800): p. 766-7.

116. Iakoucheva, L.M., et al., *The importance of intrinsic disorder for protein phosphorylation.* Nucleic Acids Res, 2004. **32**(3): p. 1037-49.
117. Cox, C.J., et al., *The regions of securin and cyclin B proteins recognized by the ubiquitination machinery are natively unfolded.* FEBS Lett, 2002. **527**(1-3): p. 303-8.
118. Khan, A.N. and P.N. Lewis, *Unstructured conformations are a substrate requirement for the Sir2 family of NAD-dependent protein deacetylases.* J Biol Chem, 2005. **280**(43): p. 36073-8.
119. Tompa, P. and P. Csermely, *The role of structural disorder in the function of RNA and protein chaperones.* FASEB J, 2004. **18**(11): p. 1169-75.
120. Mohan, A., et al., *Analysis of molecular recognition features (MoRFs).* J Mol Biol, 2006. **362**(5): p. 1043-59.
121. Vacic, V., et al., *Characterization of molecular recognition features, MoRFs, and their binding partners.* J Proteome Res, 2007. **6**(6): p. 2351-66.
122. Cheng, Y., et al., *Mining alpha-helix-forming molecular recognition features with cross species sequence alignments.* Biochemistry, 2007. **46**(47): p. 13468-77.
123. Oldfield, C.J., et al., *Coupled folding and binding with alpha-helix-forming molecular recognition elements.* Biochemistry, 2005. **44**(37): p. 12454-70.
124. Dosztanyi, Z., B. Meszaros, and I. Simon, *ANCHOR: web server for predicting protein binding regions in disordered proteins.* Bioinformatics, 2009. **25**(20): p. 2745-6.
125. Yap, K.L., et al., *Calmodulin target database.* J Struct Funct Genomics, 2000. **1**(1): p. 8-14.
126. Neduva, V. and R.B. Russell, *DILIMOT: discovery of linear motifs in proteins.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W350-5.
127. Edwards, R.J., N.E. Davey, and D.C. Shields, *SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins.* PLoS One, 2007. **2**(10): p. e967.
128. Davey, N.E., et al., *SLiMFinder: a web server to find novel, significantly over-represented, short protein motifs.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W534-9.
129. Petsalaki, E., et al., *Accurate prediction of peptide binding sites on protein surfaces.* PLoS Comput Biol, 2009. **5**(3): p. e1000335.
130. Dokholyan, N.V., et al., *Discrete molecular dynamics studies of the folding of a protein-like model.* Fold Des, 1998. **3**(6): p. 577-87.
131. Camps, J., et al., *FlexServ: an integrated tool for the analysis of protein flexibility.* Bioinformatics, 2009. **25**(13): p. 1709-10.
132. Berezin, C., et al., *ConSeq: the identification of functionally and structurally important residues in protein sequences.* Bioinformatics, 2004. **20**(8): p. 1322-4.
133. O'Brien, K.P., M. Remm, and E.L. Sonnhammer, *Inparanoid: a comprehensive database of eukaryotic orthologs.* Nucleic Acids Res, 2005. **33**(Database issue): p. D476-80.
134. Nooren, I.M. and J.M. Thornton, *Structural characterisation and functional significance of transient protein-protein interactions.* J Mol Biol, 2003. **325**(5): p. 991-1018.

135. Berchanski, A., B. Shapira, and M. Eisenstein, *Hydrophobic complementarity in protein-protein docking.* Proteins, 2004. **56**(1): p. 130-42.
136. Porter, C.T., G.J. Bartlett, and J.M. Thornton, *The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.* Nucleic Acids Res, 2004. **32**(Database issue): p. D129-33.
137. Lopez, G., A. Valencia, and M. Tress, *FireDB--a database of functionally important residues from proteins of known structure.* Nucleic Acids Res, 2007. **35**(Database issue): p. D219-23.
138. Hendlich, M., F. Rippmann, and G. Barnickel, *LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins.* J Mol Graph Model, 1997. **15**(6): p. 359-63, 389.
139. Brady, G.P., Jr. and P.F. Stouten, *Fast prediction and visualization of protein binding pockets with PASS.* J Comput Aided Mol Des, 2000. **14**(4): p. 383-401.
140. Laskowski, R.A., *SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions.* J Mol Graph, 1995. **13**(5): p. 323-30, 307-8.
141. Laurie, A.T. and R.M. Jackson, *Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites.* Bioinformatics, 2005. **21**(9): p. 1908-16.
142. Huang, B., *MetaPocket: a meta approach to improve protein ligand binding site prediction.* OMICS, 2009. **13**(4): p. 325-30.
143. Lopez, G., A. Valencia, and M.L. Tress, *firestar--prediction of functionally important residues using structural templates and alignment reliability.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W573-7.
144. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
145. Brylinski, M. and J. Skolnick, *FINDSITE: a threading-based approach to ligand homology modeling.* PLoS Comput Biol, 2009. **5**(6): p. e1000405.
146. Laskowski, R.A., *PDBsum new things.* Nucleic Acids Res, 2009. **37**(Database issue): p. D355-9.
147. Bogan, A.A. and K.S. Thorn, *Anatomy of hot spots in protein interfaces.* J Mol Biol, 1998. **280**(1): p. 1-9.
148. Bordner, A.J. and R. Abagyan, *Statistical analysis and prediction of protein-protein interfaces.* Proteins, 2005. **60**(3): p. 353-66.
149. Neuvirth, H., R. Raz, and G. Schreiber, *ProMate: a structure based prediction program to identify the location of protein-protein binding sites.* J Mol Biol, 2004. **338**(1): p. 181-99.
150. Chen, H. and H.X. Zhou, *Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data.* Proteins, 2005. **61**(1): p. 21-35.
151. Liang, S., et al., *Protein binding site prediction using an empirical scoring function.* Nucleic Acids Res, 2006. **34**(13): p. 3698-707.
152. Qin, S. and H.X. Zhou, *meta-PPISP: a meta web server for protein-protein interaction site prediction.* Bioinformatics, 2007. **23**(24): p. 3386-7.
153. Chelliah, V., et al., *Distinguishing structural and functional restraints in evolution in order to identify interaction sites.* J Mol Biol, 2004. **342**(5): p. 1487-504.

154. Bader, G.D., D. Betel, and C.W. Hogue, *BIND: the Biomolecular Interaction Network Database.* Nucleic Acids Res, 2003. **31**(1): p. 248-50.
155. Cesareni, G. and M. Helmer-Citterich, *Searching the MINT database for protein interaction information.* Curr Protoc Bioinformatics, 2003. **Chapter 8**: p. Unit 8 5.
156. Hermjakob, H., et al., *IntAct: an open source molecular interaction database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D452-5.
157. Peri, S., et al., *Development of human protein reference database as an initial platform for approaching systems biology in humans.* Genome Res, 2003. **13**(10): p. 2363-71.
158. Snel, B., et al., *STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene.* Nucleic Acids Res, 2000. **28**(18): p. 3442-4.
159. Kanehisa, M., *The KEGG database.* Novartis Found Symp, 2002. **247**: p. 91-101; discussion 101-3, 119-28, 244-52.
160. Harris, M.A., et al., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Res, 2004. **32**(Database issue): p. D258-61.
161. Szklarczyk, D., et al., *The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.* Nucleic Acids Res, 2011. **39**(Database issue): p. D561-8.
162. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): p. 2498-504.
163. Shoemaker, B.A. and A.R. Panchenko, *Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners.* PLoS Comput Biol, 2007. **3**(4): p. e43.
164. Aloy, P. and R.B. Russell, *Structural systems biology: modelling protein interactions.* Nat Rev Mol Cell Biol, 2006. **7**(3): p. 188-97.
165. Stein, A., R.B. Russell, and P. Aloy, *3did: interacting protein domains of known three-dimensional structure.* Nucleic Acids Res, 2005. **33**(Database issue): p. D413-7.
166. Kiel, C., et al., *Recognizing and defining true Ras binding domains II: in silico prediction based on homology modelling and energy calculations.* J Mol Biol, 2005. **348**(3): p. 759-75.
167. Kiel, C., P. Beltrao, and L. Serrano, *Analyzing protein interaction networks using structural information.* Annu Rev Biochem, 2008. **77**: p. 415-41.
168. Campagna, A., L. Serrano, and C. Kiel, *Shaping dots and lines: adding modularity into protein interaction networks using structural information.* FEBS Lett, 2008. **582**(8): p. 1231-6.
169. Kim, P.M., et al., *Relating three-dimensional structures to protein networks provides evolutionary insights.* Science, 2006. **314**(5807): p. 1938-41.
170. Cai, Z., et al., *Bayesian approach to discovering pathogenic SNPs in conserved protein domains.* Hum Mutat, 2004. **24**(2): p. 178-84.
171. Sherry, S.T., M. Ward, and K. Sirotkin, *dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation.* Genome Res, 1999. **9**(8): p. 677-9.
172. McKusick, V.A., *Mendelian Inheritance in Man and its online version, OMIM.* Am J Hum Genet, 2007. **80**(4): p. 588-604.

173. Apweiler, R., et al., *UniProt: the Universal Protein knowledgebase.* Nucleic Acids Res, 2004. **32**(Database issue): p. D115-9.
174. Fokkema, I.F., J.T. den Dunnen, and P.E. Taschner, *LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach.* Hum Mutat, 2005. **26**(2): p. 63-8.
175. Baker, D. and A. Sali, *Protein structure prediction and structural genomics.* Science, 2001. **294**(5540): p. 93-6.
176. Khan, S. and M. Vihinen, *Spectrum of disease-causing mutations in protein secondary structures.* BMC Struct Biol, 2007. **7**: p. 56.
177. Mathe, E., et al., *Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods.* Nucleic Acids Res, 2006. **34**(5): p. 1317-25.
178. Bromberg, Y., G. Yachdav, and B. Rost, *SNAP predicts effect of mutations on protein function.* Bioinformatics, 2008. **24**(20): p. 2397-8.
179. Kumar, P., S. Henikoff, and P.C. Ng, *Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.* Nat Protoc, 2009. **4**(7): p. 1073-81.
180. Ferrer-Costa, C., et al., *PMUT: a web-based tool for the annotation of pathological mutations on proteins.* Bioinformatics, 2005. **21**(14): p. 3176-8.
181. Capriotti, E., R. Calabrese, and R. Casadio, *Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information.* Bioinformatics, 2006. **22**(22): p. 2729-34.
182. Ramensky, V., P. Bork, and S. Sunyaev, *Human non-synonymous SNPs: server and survey.* Nucleic Acids Res, 2002. **30**(17): p. 3894-900.
183. Yue, P., E. Melamud, and J. Moult, *SNPs3D: candidate gene and SNP selection for association studies.* BMC Bioinformatics, 2006. **7**: p. 166.
184. Bash, P.A., et al., *Free energy calculations by computer simulation.* Science, 1987. **236**(4801): p. 564-8.
185. Funahashi, J., et al., *How can free energy component analysis explain the difference in protein stability caused by amino acid substitutions? Effect of three hydrophobic mutations at the 56th residue on the stability of human lysozyme.* Protein Eng, 2003. **16**(9): p. 665-71.
186. Guerois, R., J.E. Nielsen, and L. Serrano, *Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations.* J Mol Biol, 2002. **320**(2): p. 369-87.
187. Cheng, J., A. Randall, and P. Baldi, *Prediction of protein stability changes for single-site mutations using support vector machines.* Proteins, 2006. **62**(4): p. 1125-32.
188. Yin, S., F. Ding, and N.V. Dokholyan, *Eris: an automated estimator of protein stability.* Nat Methods, 2007. **4**(6): p. 466-7.
189. Gilis, D. and M. Rooman, *PoPMuSiC, an algorithm for predicting protein mutant stability changes: application to prion proteins.* Protein Eng, 2000. **13**(12): p. 849-56.
190. Capriotti, E., P. Fariselli, and R. Casadio, *I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W306-10.

191. Capriotti, E., et al., *A three-state prediction of single point mutations on protein stability changes.* BMC Bioinformatics, 2008. **9 Suppl 2**: p. S6.
192. Masso, M. and Vaisman, II, *AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements.* Protein Eng Des Sel, 2010. **23**(8): p. 683-7.
193. Csermely, P., *Creative elements: network-based predictions of active centres in proteins and cellular and social networks.* Trends Biochem Sci, 2008. **33**(12): p. 569-76.
194. Del Sol, A., et al., *Modular architecture of protein structures and allosteric communications: potential implications for signaling proteins and regulatory linkages.* Genome Biol, 2007. **8**(5): p. R92.
195. Dokholyan, N.V., et al., *Topological determinants of protein folding.* Proc Natl Acad Sci U S A, 2002. **99**(13): p. 8637-41.
196. Soundararajan, V., et al., *Atomic interaction networks in the core of protein domains and their native folds.* PLoS One, 2010. **5**(2): p. e9391.
197. Suel, G.M., et al., *Evolutionarily conserved networks of residues mediate allosteric communication in proteins.* Nat Struct Biol, 2003. **10**(1): p. 59-69.
198. Vendruscolo, M., et al., *Three key residues form a critical contact network in a protein folding transition state.* Nature, 2001. **409**(6820): p. 641-5.
199. Cheng, T.M., et al., *Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms.* PLoS Comput Biol, 2008. **4**(7): p. e1000135.
200. Martin, A.J., et al., *RING: networking interacting residues, evolutionary information and energetics in protein structures.* Bioinformatics, 2011. **27**(14): p. 2003-5.
201. Doncheva, N.T., et al., *Analyzing and visualizing residue networks of protein structures.* Trends Biochem Sci, 2011. **36**(4): p. 179-82.
202. Niaudet, P. and M.C. Gubler, *WT1 and glomerular diseases.* Pediatr Nephrol, 2006. **21**(11): p. 1653-60.
203. Morrison, A.A., et al., *New insights into the function of the Wilms tumor suppressor gene WT1 in podocytes.* Am J Physiol Renal Physiol, 2008. **295**(1): p. F12-7.
204. Pelletier, J., et al., *Germline mutations in the Wilms' tumor suppressor gene are associated with abnormal urogenital development in Denys-Drash syndrome.* Cell, 1991. **67**(2): p. 437-47.
205. Reddy, J.C., et al., *WT1-mediated transcriptional activation is inhibited by dominant negative mutant proteins.* J Biol Chem, 1995. **270**(18): p. 10878-84.
206. Klamt, B., et al., *Frasier syndrome is caused by defective alternative splicing of WT1 leading to an altered ratio of WT1 +/-KTS splice isoforms.* Hum Mol Genet, 1998. **7**(4): p. 709-14.
207. Jeanpierre, C., et al., *Identification of constitutional WT1 mutations, in patients with isolated diffuse mesangial sclerosis, and analysis of genotype/phenotype correlations by use of a computerized mutation database.* Am J Hum Genet, 1998. **62**(4): p. 824-33.
208. Denamur, E., et al., *Mother-to-child transmitted WT1 splice-site mutation is responsible for distinct glomerular diseases.* J Am Soc Nephrol, 1999. **10**(10): p. 2219-23.

209. McTaggart, S.J., et al., *Clinical spectrum of Denys-Drash and Frasier syndrome.* Pediatr Nephrol, 2001. **16**(4): p. 335-9.
210. Kohsaka, T., et al., *Exon 9 mutations in the WT1 gene, without influencing KTS splice isoforms, are also responsible for Frasier syndrome.* Hum Mutat, 1999. **14**(6): p. 466-70.
211. Kaltenis, P., et al., *Slow progressive FSGS associated with an F392L WT1 mutation.* Pediatr Nephrol, 2004. **19**(3): p. 353-6.
212. Demmer, L., et al., *Frasier syndrome: a cause of focal segmental glomerulosclerosis in a 46,XX female.* J Am Soc Nephrol, 1999. **10**(10): p. 2215-8.
213. Hu, M., et al., *A novel mutation of WT1 exon 9 in a patient with Denys-Drash syndrome and pyloric stenosis.* Pediatr Nephrol, 2004. **19**(10): p. 1160-3.
214. Stoll, R., et al., *Structure of the Wilms tumor suppressor protein zinc finger domain bound to DNA.* J Mol Biol, 2007. **372**(5): p. 1227-45.
215. Weiss, T.C. and P.J. Romaniuk, *Contribution of individual amino acids to the RNA binding activity of the Wilms' tumor suppressor protein WT1.* Biochemistry, 2009. **48**(1): p. 148-55.
216. Kikuchi, H., et al., *Do intronic mutations affecting splicing of WT1 exon 9 cause Frasier syndrome?* J Med Genet, 1998. **35**(1): p. 45-8.
217. Tsuda, M., et al., *WT1 nephropathy in a girl with normal karyotype (46,XX).* Clin Nephrol, 1999. **51**(1): p. 62-3.
218. Hahn, H., et al., *Two cases of isolated diffuse mesangial sclerosis with WT1 mutations.* J Korean Med Sci, 2006. **21**(1): p. 160-4.
219. Yang, Y., et al., *WT1 and PAX-2 podocyte expression in Denys-Drash syndrome and isolated diffuse mesangial sclerosis.* Am J Pathol, 1999. **154**(1): p. 181-92.
220. Mucha, B., et al., *Mutations in the Wilms' tumor 1 gene cause isolated steroid resistant nephrotic syndrome and occur in exons 8 and 9.* Pediatr Res, 2006. **59**(2): p. 325-31.
221. Zirn, B., S. Wittmann, and M. Gessler, *Novel familial WT1 read-through mutation associated with Wilms tumor and slow progressive nephropathy.* Am J Kidney Dis, 2005. **45**(6): p. 1100-4.
222. Regev, M., et al., *Vertical transmission of a mutation in exon 1 of the WT1 gene: lessons for genetic counseling.* Am J Med Genet A, 2008. **146A**(18): p. 2332-6.
223. Viney, R.L., et al., *A proteomic investigation of glomerular podocytes from a Denys-Drash syndrome patient with a mutation in the Wilms tumour suppressor gene WT1.* Proteomics, 2007. **7**(5): p. 804-15.
224. Wagner, N., et al., *Intermediate filament protein nestin is expressed in developing kidney and heart and might be regulated by the Wilms' tumor suppressor Wt1.* Am J Physiol Regul Integr Comp Physiol, 2006. **291**(3): p. R779-87.
225. Su, W., et al., *Expression of nestin in the podocytes of normal and diseased human kidneys.* Am J Physiol Regul Integr Comp Physiol, 2007. **292**(5): p. R1761-7.
226. Kaelin, W.G., *Von Hippel-Lindau disease.* Annu Rev Pathol, 2007. **2**: p. 145-73.
227. Ong, K.R., et al., *Genotype-phenotype correlations in von Hippel-Lindau disease.* Hum Mutat, 2007. **28**(2): p. 143-9.

228. Stebbins, C.E., W.G. Kaelin, Jr., and N.P. Pavletich, *Structure of the VHL-ElonginC-ElonginB complex: implications for VHL tumor suppressor function.* Science, 1999. **284**(5413): p. 455-61.
229. Hon, W.C., et al., *Structural basis for the recognition of hydroxyproline in HIF-1 alpha by pVHL.* Nature, 2002. **417**(6892): p. 975-8.
230. Min, J.H., et al., *Structure of an HIF-1alpha -pVHL complex: hydroxyproline recognition in signaling.* Science, 2002. **296**(5574): p. 1886-9.
231. Kaelin, W.G., Jr., *The von Hippel-Lindau tumour suppressor protein: O2 sensing and cancer.* Nat Rev Cancer, 2008. **8**(11): p. 865-73.
232. Frew, I.J. and W. Krek, *Multitasking by pVHL in tumour suppression.* Curr Opin Cell Biol, 2007. **19**(6): p. 685-90.
233. Frew, I.J. and W. Krek, *pVHL: a multipurpose adaptor protein.* Sci Signal, 2008. **1**(24): p. pe30.
234. Devos, D. and R.B. Russell, *A more complete, complexed and structured interactome.* Curr Opin Struct Biol, 2007. **17**(3): p. 370-7.
235. Bairoch, A., et al., *The Universal Protein Resource (UniProt).* Nucleic Acids Res, 2005. **33**(Database issue): p. D154-9.
236. Bateman, A., et al., *The Pfam protein families database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D138-41.
237. Bourne, P.E., et al., *The distribution and query systems of the RCSB Protein Data Bank.* Nucleic Acids Res, 2004. **32 Database issue**: p. D223-5.
238. Camps, J., et al., *FlexServ: An integrated tool for the analysis of protein flexibility.* Bioinformatics, 2009.
239. Bindewald, E., et al., *MANIFOLD: protein fold recognition based on secondary structure, sequence similarity and enzyme classification.* Protein Eng, 2003. **16**(11): p. 785-9.
240. Pearl, F.M., et al., *The CATH database: an extended protein family resource for structural and functional genomics.* Nucleic Acids Res, 2003. **31**(1): p. 452-5.
241. *The Gene Ontology (GO) project in 2006.* Nucleic Acids Res, 2006. **34**(Database issue): p. D322-6.
242. Landau, M., et al., *ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W299-302.
243. Waterhouse, A.M., et al., *Jalview Version 2--a multiple sequence alignment editor and analysis workbench.* Bioinformatics, 2009. **25**(9): p. 1189-91.
244. Roe, J.S., et al., *p53 stabilization and transactivation by a von Hippel-Lindau protein.* Mol Cell, 2006. **22**(3): p. 395-405.
245. Kibel, A., et al., *Binding of the von Hippel-Lindau tumor suppressor protein to Elongin B and C.* Science, 1995. **269**(5229): p. 1444-6.
246. Kamura, T., et al., *VHL-box and SOCS-box domains determine binding specificity for Cul2-Rbx1 and Cul5-Rbx2 modules of ubiquitin ligases.* Genes Dev, 2004. **18**(24): p. 3055-65.
247. Russell, R.C. and M. Ohh, *NEDD8 acts as a 'molecular switch' defining the functional selectivity of VHL.* EMBO Rep, 2008. **9**(5): p. 486-91.
248. Cohen, H.T., et al., *An important von Hippel-Lindau tumor suppressor domain mediates Sp1-binding and self-association.* Biochem Biophys Res Commun, 1999. **266**(1): p. 43-50.

249. Kuznetsova, A.V., et al., *von Hippel-Lindau protein binds hyperphosphorylated large subunit of RNA polymerase II through a proline hydroxylation motif and targets it for ubiquitination.* Proc Natl Acad Sci U S A, 2003. **100**(5): p. 2706-11.
250. Corn, P.G., et al., *Tat-binding protein-1, a component of the 26S proteasome, contributes to the E3 ubiquitin ligase function of the von Hippel-Lindau protein.* Nat Genet, 2003. **35**(3): p. 229-37.
251. Khacho, M., et al., *eEF1A is a novel component of the mammalian nuclear protein export machinery.* Mol Biol Cell, 2008. **19**(12): p. 5296-308.
252. Khacho, M., et al., *Cancer-causing mutations in a novel transcription-dependent nuclear export motif of VHL abrogate oxygen-dependent degradation of hypoxia-inducible factor.* Mol Cell Biol, 2008. **28**(1): p. 302-14.
253. Schoenfeld, A.R., E.J. Davidowitz, and R.D. Burk, *Endoplasmic reticulum/cytosolic localization of von Hippel-Lindau gene products is mediated by a 64-amino acid region.* Int J Cancer, 2001. **91**(4): p. 457-67.
254. Lolkema, M.P., et al., *Tumor suppression by the von Hippel-Lindau protein requires phosphorylation of the acidic domain.* J Biol Chem, 2005. **280**(23): p. 22205-11.
255. Hergovich, A., et al., *Priming-dependent phosphorylation and regulation of the tumor suppressor pVHL by glycogen synthase kinase 3.* Mol Cell Biol, 2006. **26**(15): p. 5784-96.
256. Kajava, A.V., *Review: proteins with repeated sequence--structural prediction and modeling.* J Struct Biol, 2001. **134**(2-3): p. 132-44.
257. Lise, S. and D.T. Jones, *Sequence patterns associated with disordered regions in proteins.* Proteins, 2005. **58**(1): p. 144-50.
258. Kaelin, W.G., Jr. and E.R. Maher, *The VHL tumour-suppressor gene paradigm.* Trends Genet, 1998. **14**(10): p. 423-6.
259. Lonser, R.R., et al., *von Hippel-Lindau disease.* Lancet, 2003. **361**(9374): p. 2059-67.
260. Crossey, P.A., et al., *Identification of intragenic mutations in the von Hippel-Lindau disease tumour suppressor gene and correlation with disease phenotype.* Hum Mol Genet, 1994. **3**(8): p. 1303-8.
261. Nordstrom-O'Brien, M., et al., *Genetic analysis of von Hippel-Lindau disease.* Hum Mutat, 2010. **31**(5): p. 521-37.
262. Franke, G., et al., *Alu-Alu recombination underlies the vast majority of large VHL germline deletions: Molecular characterization and genotype-phenotype correlations in VHL patients.* Hum Mutat, 2009. **30**(5): p. 776-86.
263. Hes, F.J., et al., *Frequency of Von Hippel-Lindau germline mutations in classic and non-classic Von Hippel-Lindau disease identified by DNA sequencing, Southern blot analysis and multiplex ligation-dependent probe amplification.* Clin Genet, 2007. **72**(2): p. 122-9.
264. Ritter, M.M., et al., *Isolated familial pheochromocytoma as a variant of von Hippel-Lindau disease.* J Clin Endocrinol Metab, 1996. **81**(3): p. 1035-7.
265. Crossey, P.A., et al., *Molecular genetic diagnosis of von Hippel-Lindau disease in familial phaeochromocytoma.* J Med Genet, 1995. **32**(11): p. 885-6.
266. Lisztwan, J., et al., *The von Hippel-Lindau tumor suppressor protein is a component of an E3 ubiquitin-protein ligase activity.* Genes Dev, 1999. **13**(14): p. 1822-33.

267. Leonardi, E., A. Murgia, and S.C. Tosatto, *Adding structural information to the von Hippel-Lindau (VHL) tumor suppressor interaction network.* FEBS Lett, 2009.
268. Nordstrom-O'Brien, M., et al., *Genetic analysis of von Hippel-Lindau disease.* Hum Mutat. **31**(5): p. 521-37.
269. Martella, M., et al., *Molecular analysis of two uncharacterized sequence variants of the VHL gene.* J Hum Genet, 2006. **51**(11): p. 964-8.
270. Casarin, A., et al., *Molecular characterization of large deletions in the von Hippel-Lindau (VHL) gene by quantitative real-time PCR: the hypothesis of an alu-mediated mechanism underlying VHL gene rearrangements.* Mol Diagn Ther, 2006. **10**(4): p. 243-9.
271. Kuzmin, I., et al., *Identification of the promoter of the human von Hippel-Lindau disease tumor suppressor gene.* Oncogene, 1995. **10**(11): p. 2185-94.
272. Beroud, C., et al., *Software and database for the analysis of mutations in the VHL gene.* Nucleic Acids Res, 1998. **26**(1): p. 256-8.
273. Reese, M.G., et al., *Improved splice site detection in Genie.* J Comput Biol, 1997. **4**(3): p. 311-23.
274. Hebsgaard, S.M., et al., *Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information.* Nucleic Acids Res, 1996. **24**(17): p. 3439-52.
275. Masso, M. and Vaisman, II, *Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis.* Bioinformatics, 2008. **24**(18): p. 2002-9.
276. Bromberg, Y. and B. Rost, *SNAP: predict effect of non-synonymous polymorphisms on function.* Nucleic Acids Res, 2007. **35**(11): p. 3823-35.
277. Sommer, I., et al., *Improving the quality of protein structure models by selecting from alignment alternatives.* BMC Bioinformatics, 2006. **7**: p. 364.
278. Sutovsky, H. and E. Gazit, *The von Hippel-Lindau tumor suppressor protein is a molten globule under native conditions: implications for its physiological activities.* J Biol Chem, 2004. **279**(17): p. 17190-6.
279. Liu, J. and R. Nussinov, *The mechanism of ubiquitination in the cullin-RING E3 ligase machinery: conformational control of substrate orientation.* PLoS Comput Biol, 2009. **5**(10): p. e1000527.
280. Abbott, M.A., et al., *The von Hippel-Lindau (VHL) germline mutation V84L manifests as early-onset bilateral pheochromocytoma.* Am J Med Genet A, 2006. **140**(7): p. 685-90.
281. Khan, S. and M. Vihinen, *Performance of protein stability predictors.* Hum Mutat, 2010. **31**(6): p. 675-84.
282. Wagner, E. and J. Lykke-Andersen, *mRNA surveillance: the perfect persist.* J Cell Sci, 2002. **115**(Pt 15): p. 3033-8.
283. Gallou, C., et al., *Association of GSTT1 non-null and NAT1 slow/rapid genotypes with von Hippel-Lindau tumour suppressor gene transversions in sporadic renal cell carcinoma.* Pharmacogenetics, 2001. **11**(6): p. 521-35.
284. Nagy, E. and L.E. Maquat, *A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance.* Trends Biochem Sci, 1998. **23**(6): p. 198-9.

285. Zhang, J. and L.E. Maquat, *Evidence that translation reinitiation abrogates nonsense-mediated mRNA decay in mammalian cells.* Embo J, 1997. **16**(4): p. 826-33.
286. Chauveau, D., et al., *Diagnosis of pheochromocytoma and laparoscopic adrenalectomy in two anephric patients with von Hippel-Lindau disease.* Am J Kidney Dis, 2002. **39**(2): p. E6.
287. Turturro, F., *Beyond the Knudson's hypothesis in von Hippel-Lindau (VHL) disease-proposing vitronectin as a "gene modifier".* J Mol Med, 2009. **87**(6): p. 591-3.
288. Pettigrew, C.A. and M.A. Brown, *Pre-mRNA splicing aberrations and cancer.* Front Biosci, 2008. **13**: p. 1090-105.
289. Clifford, S.C., et al., *Contrasting effects on HIF-1alpha regulation by disease-causing pVHL mutations correlate with patterns of tumourigenesis in von Hippel-Lindau disease.* Hum Mol Genet, 2001. **10**(10): p. 1029-38.
290. Lee, C.M., et al., *VHL Type 2B gene mutation moderates HIF dosage in vitro and in vivo.* Oncogene, 2009. **28**(14): p. 1694-705.
291. Ohh, M., et al., *The von Hippel-Lindau tumor suppressor protein is required for proper assembly of an extracellular fibronectin matrix.* Mol Cell, 1998. **1**(7): p. 959-68.
292. Iturrioz, X., et al., *The von Hippel-Lindau tumour-suppressor protein interaction with protein kinase Cdelta.* Biochem J, 2006. **397**(1): p. 109-20.
293. Grosfeld, A., et al., *Interaction of hydroxylated collagen IV with the von hippel-lindau tumor suppressor.* J Biol Chem, 2007. **282**(18): p. 13264-9.
294. Knauth, K., et al., *Renal cell carcinoma risk in type 2 von Hippel-Lindau disease correlates with defects in pVHL stability and HIF-1alpha interactions.* Oncogene, 2006. **25**(3): p. 370-7.
295. Roe, J.S. and H.D. Youn, *The positive regulation of p53 by the tumor suppressor VHL.* Cell Cycle, 2006. **5**(18): p. 2054-6.
296. Forman, J.R., et al., *Structural bioinformatics mutation analysis reveals genotype-phenotype correlations in von Hippel-Lindau disease and suggests molecular mechanisms of tumorigenesis.* Proteins, 2009.
297. Lolkema, M.P., et al., *The von Hippel-Lindau tumour suppressor interacts with microtubules through kinesin-2.* FEBS Lett, 2007. **581**(24): p. 4571-6.
298. Tang, N., et al., *pVHL function is essential for endothelial extracellular matrix deposition.* Mol Cell Biol, 2006. **26**(7): p. 2519-30.
299. Hoffman, M.A., et al., *von Hippel-Lindau protein mutants linked to type 2C VHL disease preserve the ability to downregulate HIF.* Hum Mol Genet, 2001. **10**(10): p. 1019-27.
300. Thoma, C.R., I.J. Frew, and W. Krek, *The VHL tumor suppressor: riding tandem with GSK3beta in primary cilium maintenance.* Cell Cycle, 2007. **6**(15): p. 1809-13.
301. Knauth, K., et al., *VHL mutations linked to type 2C von Hippel-Lindau disease cause extensive structural perturbations in pVHL.* J Biol Chem, 2009. **284**(16): p. 10514-22.
302. Zhou, M.I., et al., *The von Hippel-Lindau tumor suppressor stabilizes novel plant homeodomain protein Jade-1.* J Biol Chem, 2002. **277**(42): p. 39887-98.

303. Guo, Y., M.C. Schoell, and R.S. Freeman, *The von Hippel-Lindau protein sensitizes renal carcinoma cells to apoptotic stimuli through stabilization of BIM(EL).* Oncogene, 2009. **28**(16): p. 1864-74.
304. Lee, S., et al., *Neuronal apoptosis linked to EglN3 prolyl hydroxylase and familial pheochromocytoma genes: developmental culling and cancer.* Cancer Cell, 2005. **8**(2): p. 155-67.
305. Liu, J. and R. Nussinov, *Allosteric effects in the marginally stable von Hippel-Lindau tumor suppressor protein and allostery-based rescue mutant design.* Proc Natl Acad Sci U S A, 2008. **105**(3): p. 901-6.
306. Kunapuli, P., K.S. Chitta, and J.K. Cowell, *Suppression of the cell proliferation and invasion phenotypes in glioma cells by the LGI1 gene.* Oncogene, 2003. **22**(26): p. 3985-91.
307. Kunapuli, P., et al., *LGI1, a putative tumor metastasis suppressor gene, controls in vitro invasiveness and expression of matrix metalloproteinases in glioma cells through the ERK1/2 pathway.* J Biol Chem, 2004. **279**(22): p. 23151-7.
308. Michelucci, R., et al., *Autosomal dominant lateral temporal epilepsy: clinical spectrum, new epitempin mutations, and genetic heterogeneity in seven European families.* Epilepsia, 2003. **44**(10): p. 1289-97.
309. Ottman, R., et al., *Localization of a gene for partial epilepsy to chromosome 10q.* Nat Genet, 1995. **10**(1): p. 56-60.
310. Morante-Redolat, J.M., et al., *Mutations in the LGI1/Epitempin gene on 10q24 cause autosomal dominant lateral temporal epilepsy.* Hum Mol Genet, 2002. **11**(9): p. 1119-28.
311. Kalachikov, S., et al., *Mutations in LGI1 cause autosomal-dominant partial epilepsy with auditory features.* Nat Genet, 2002. **30**(3): p. 335-41.
312. Nobile, C., et al., *LGI1 mutations in autosomal dominant and sporadic lateral temporal epilepsy.* Hum Mutat, 2009. **30**(4): p. 530-6.
313. Senechal, K.R., C. Thaller, and J.L. Noebels, *ADPEAF mutations reduce levels of secreted LGI1, a putative tumor suppressor protein linked to epilepsy.* Hum Mol Genet, 2005. **14**(12): p. 1613-20.
314. Staub, E., et al., *The novel EPTP repeat defines a superfamily of proteins implicated in epileptic disorders.* Trends Biochem Sci, 2002. **27**(9): p. 441-4.
315. Scheel, H., S. Tomiuk, and K. Hofmann, *A common protein interaction domain links two recently identified epilepsy genes.* Hum Mol Genet, 2002. **11**(15): p. 1757-62.
316. Paoli, M., *An elusive propeller-like fold.* Nat Struct Biol, 2001. **8**(9): p. 744-5.
317. Buchanan, S.G. and N.J. Gay, *Structural and functional diversity in the leucine-rich repeat family of proteins.* Prog Biophys Mol Biol, 1996. **65**(1-2): p. 1-44.
318. Fukata, Y., et al., *Epilepsy-related ligand/receptor complex LGI1 and ADAM22 regulate synaptic transmission.* Science, 2006. **313**(5794): p. 1792-5.
319. Zhou, Y.D., et al., *Arrested maturation of excitatory synapses in autosomal dominant lateral temporal lobe epilepsy.* Nat Med, 2009. **15**(10): p. 1208-14.
320. Owuor, K., et al., *LGI1-associated epilepsy through altered ADAM23-dependent neuronal morphology.* Mol Cell Neurosci, 2009. **42**(4): p. 448-57.
321. Fukata, Y., et al., *Disruption of LGI1-linked synaptic complex causes abnormal synaptic transmission and epilepsy.* Proc Natl Acad Sci U S A, 2010. **107**(8): p. 3799-804.

322. Boeckmann, B., et al., *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.* Nucleic Acids Res, 2003. **31**(1): p. 365-70.
323. Bendtsen, J.D., et al., *Improved prediction of signal peptides: SignalP 3.0.* J Mol Biol, 2004. **340**(4): p. 783-95.
324. Strimmer, K., N. Goldman, and A. von Haeseler, *Bayesian probabilities and quartet puzzling.* Molecular Biology and Evolution, 1997. **14**(2): p. 210-211.
325. Strimmer, K. and A. von Haeseler, *Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies.* Molecular Biology and Evolution, 1996. **13**(7): p. 964-969.
326. Jones, D.T., W.R. Taylor, and J.M. Thornton, *The rapid generation of mutation data matrices from protein sequences.* Comput Appl Biosci, 1992. **8**(3): p. 275-82.
327. Felsenstein, J., *Phylogenies and the comparative method.* Amer. Naturalist, 1985. **125**: p. 1-15.
328. Bujnicki, J.M., et al., *Structure prediction meta server.* Bioinformatics, 2001. **17**(8): p. 750-751.
329. Sommer, I., et al., *Improving the quality of protein structure models by selecting from alignment alternatives.* BMC Bioinformatics, 2006. **7**(1): p. 364.
330. Perry, J., N. Kleckner, and G.V. Borner, *Bioinformatic analyses implicate the collaborating meiotic crossover/chiasma proteins Zip2, Zip3, and Spo22/Zip4 in ubiquitin labeling.* Proc Natl Acad Sci U S A, 2005. **102**(49): p. 17594-9.
331. Canutescu, A.A., A.A. Shelenkov, and R.L. Dunbrack, *A graph-theory algorithm for rapid protein side-chain prediction.* Protein Science, 2003. **12**: p. 2001-2014.
332. Herranz-Perez, V., et al., *Regional distribution of the leucine-rich glioma inactivated (LGI) gene family transcripts in the adult mouse brain.* Brain Res, 2010. **1307**: p. 177-94.
333. Gu, W., et al., *Using gene-history and expression analyses to assess the involvement of LGI genes in human disorders.* Mol Biol Evol, 2005. **22**(11): p. 2209-16.
334. Matsushima, N., et al., *Structural analysis of leucine-rich-repeat variants in proteins associated with human diseases.* Cell Mol Life Sci, 2005. **62**(23): p. 2771-91.
335. Howitt, J.A., N.J. Clout, and E. Hohenester, *Binding site for Robo receptors revealed by dissection of the leucine-rich repeat region of Slit.* Embo J, 2004. **23**(22): p. 4406-12.
336. Kobe, B. and A.V. Kajava, *When protein folding is simplified to protein coiling: the continuum of solenoid protein structures.* Trends Biochem Sci, 2000. **25**(10): p. 509-15.
337. Evdokimov, A.G., et al., *Unusual molecular architecture of the Yersinia pestis cytotoxin YopM: a leucine-rich repeat protein with the shortest repeating unit.* J Mol Biol, 2001. **312**(4): p. 807-21.
338. Chaudhuri, I., J. Soding, and A.N. Lupas, *Evolution of the beta-propeller fold.* Proteins, 2008. **71**(2): p. 795-803.
339. Smith, T.F., et al., *The WD repeat: a common architecture for diverse functions.* Trends Biochem Sci, 1999. **24**(5): p. 181-5.

340. Bella, J., et al., *The leucine-rich repeat structure.* Cell Mol Life Sci, 2008. **65**(15): p. 2307-33.
341. Kobe, B. and J. Deisenhofer, *A structural basis of the interactions between leucine-rich repeats and protein ligands.* Nature, 1995. **374**(6518): p. 183-6.
342. Stirnimann, C.U., et al., *WD40 proteins propel cellular networks.* Trends Biochem Sci, 2010.
343. Sagane, K., Y. Ishihama, and H. Sugimoto, *LGI1 and LGI4 bind to ADAM22, ADAM23 and ADAM11.* Int J Biol Sci, 2008. **4**(6): p. 387-96.
344. Ozkaynak, E., et al., *Adam22 is a major neuronal receptor for Lgi4-mediated Schwann cell signaling.* J Neurosci, 2010. **30**(10): p. 3857-64.
345. Sirerol-Piquer, M.S., et al., *The epilepsy gene LGI1 encodes a secreted glycoprotein that binds to the cell surface.* Hum Mol Genet, 2006. **15**(23): p. 3436-45.
346. Striano, P., et al., *Familial temporal lobe epilepsy with psychic auras associated with a novel LGI1 mutation.* Neurology, 2011. **76**(13): p. 1173-6.
347. Chernova, O.B., R.P. Somerville, and J.K. Cowell, *A novel gene, LGI1, from 10q24 is rearranged and downregulated in malignant brain tumors.* Oncogene, 1998. **17**(22): p. 2873-81.
348. Manya, H., et al., *Demonstration of mammalian protein O-mannosyltransferase activity: coexpression of POMT1 and POMT2 required for enzymatic activity.* Proc Natl Acad Sci U S A, 2004. **101**(2): p. 500-5.
349. Akasaka-Manya, K., et al., *Physical and functional association of human protein O-mannosyltransferases 1 and 2.* J Biol Chem, 2006. **281**(28): p. 19339-45.
350. Jurado, L.A., A. Coloma, and J. Cruces, *Identification of a human homolog of the Drosophila rotated abdomen gene (POMT1) encoding a putative protein O-mannosyl-transferase, and assignment to human chromosome 9q34.1.* Genomics, 1999. **58**(2): p. 171-80.
351. Barresi, R. and K.P. Campbell, *Dystroglycan: from biosynthesis to pathogenesis of human disease.* J Cell Sci, 2006. **119**(Pt 2): p. 199-207.
352. Beltran-Valero de Bernabe, D., et al., *Mutations in the O-mannosyltransferase gene POMT1 give rise to the severe neuronal migration disorder Walker-Warburg syndrome.* Am J Hum Genet, 2002. **71**(5): p. 1033-43.
353. van Reeuwijk, J., et al., *The expanding phenotype of POMT1 mutations: from Walker-Warburg syndrome to congenital muscular dystrophy, microcephaly, and mental retardation.* Hum Mutat, 2006. **27**(5): p. 453-9.
354. Balci, B., et al., *An autosomal recessive limb girdle muscular dystrophy (LGMD2) with mild mental retardation is allelic to Walker-Warburg syndrome (WWS) caused by a mutation in the POMT1 gene.* Neuromuscul Disord, 2005. **15**(4): p. 271-5.
355. Godfrey, C., et al., *Refining genotype phenotype correlations in muscular dystrophies with defective glycosylation of dystroglycan.* Brain, 2007. **130**(Pt 10): p. 2725-35.
356. Muntoni, F., S. Torelli, and M. Brockington, *Muscular dystrophies due to glycosylation defects.* Neurotherapeutics, 2008. **5**(4): p. 627-32.
357. Poppe, M., et al., *Cardiac and respiratory failure in limb-girdle muscular dystrophy 2I.* Ann Neurol, 2004. **56**(5): p. 738-41.

358. Bourteel, H., et al., *Clinical and mutational spectrum of limb-girdle muscular dystrophy type 2I in 11 French patients.* J Neurol Neurosurg Psychiatry, 2009. **80**(12): p. 1405-8.
359. Boito, C.A., et al., *Clinical and molecular characterization of patients with limb-girdle muscular dystrophy type 2I.* Arch Neurol, 2005. **62**(12): p. 1894-9.
360. Margeta, M., et al., *Cardiac pathology exceeds skeletal muscle pathology in two cases of limb-girdle muscular dystrophy type 2I.* Muscle Nerve, 2009. **40**(5): p. 883-9.
361. D'Amico, A., et al., *Expanding the clinical spectrum of POMT1 phenotype.* Neurology, 2006. **66**(10): p. 1564-7; discussion 1461.
362. Mulder, N.J., et al., *The InterPro Database, 2003 brings increased coverage and new features.* Nucleic Acids Res, 2003. **31**(1): p. 315-8.
363. Jimenez-Mallebrera, C., et al., *A comparative study of alpha-dystroglycan glycosylation in dystroglycanopathies suggests that the hypoglycosylation of alpha-dystroglycan does not consistently correlate with clinical severity.* Brain Pathol, 2009. **19**(4): p. 596-611.
364. Manya, H., et al., *Protein O-mannosyltransferase activities in lymphoblasts from patients with alpha-dystroglycanopathies.* Neuromuscul Disord, 2008. **18**(1): p. 45-51.
365. Murakami, T., et al., *Fukutin gene mutations cause dilated cardiomyopathy with minimal muscle weakness.* Ann Neurol, 2006. **60**(5): p. 597-602.
366. Mercuri, E., et al., *Congenital muscular dystrophies with defective glycosylation of dystroglycan: a population study.* Neurology, 2009. **72**(21): p. 1802-9.
367. Lommel, M., et al., *Correlation of enzyme activity and clinical phenotype in POMT1-associated dystroglycanopathies.* Neurology, 2010. **74**(2): p. 157-64.
368. Michele, D.E., et al., *Dystroglycan matrix receptor function in cardiac myocytes is important for limiting activity-induced myocardial damage.* Circ Res, 2009. **105**(10): p. 984-93.
369. Brini, M., et al., *A comparative functional analysis of plasma membrane Ca2+ pump isoforms in intact cells.* J Biol Chem, 2003. **278**(27): p. 24500-8.
370. Brini, M. and E. Carafoli, *Calcium pumps in health and disease.* Physiol Rev, 2009. **89**(4): p. 1341-78.
371. Rimessi, A., et al., *Inhibitory interaction of the 14-3-3{epsilon} protein with isoform 4 of the plasma membrane Ca(2+)-ATPase pump.* J Biol Chem, 2005. **280**(44): p. 37195-203.
372. Linde, C.I., et al., *Inhibitory interaction of the 14-3-3 proteins with ubiquitous (PMCA1) and tissue-specific (PMCA3) isoforms of the plasma membrane Ca2+ pump.* Cell Calcium, 2008. **43**(6): p. 550-61.
373. Chicka, M.C. and E.E. Strehler, *Alternative splicing of the first intracellular loop of plasma membrane Ca2+-ATPase isoform 2 alters its membrane targeting.* J Biol Chem, 2003. **278**(20): p. 18464-70.
374. Xiong, Y., et al., *Apical localization of PMCA2w/b is lipid raft-dependent.* Biochem Biophys Res Commun, 2009. **384**(1): p. 32-6.
375. Brodin, P., et al., *Identification of two domains which mediate the binding of activating phospholipids to the plasma-membrane Ca2+ pump.* Eur J Biochem, 1992. **204**(2): p. 939-46.

376. Heim, R., et al., *Expression, purification, and properties of the plasma membrane Ca2+ pump and of its N-terminally truncated 105-kDa fragment.* J Biol Chem, 1992. **267**(34): p. 24476-84.
377. Adamo, H.P. and J.T. Penniston, *New Ca2+ pump isoforms generated by alternative splicing of rPMCA2 mRNA.* Biochem J, 1992. **283 (Pt 2)**: p. 355-9.
378. Enyedi, A., et al., *The maximal velocity and the calcium affinity of the red cell calcium pump may be regulated independently.* J Biol Chem, 1987. **262**(13): p. 6425-30.
379. Papp, B., et al., *Functional domains of the in situ red cell membrane calcium pump revealed by proteolysis and monoclonal antibodies. Possible sites for regulation by calpain and acidic lipids.* J Biol Chem, 1989. **264**(8): p. 4577-82.
380. Missiaen, L., et al., *Phospholipid-protein interactions of the plasma-membrane Ca2+-transporting ATPase. Evidence for a tissue-dependent functional difference.* Biochem J, 1989. **263**(3): p. 687-94.
381. Missiaen, L., et al., *Polyamines and neomycin inhibit the purified plasma-membrane Ca2+ pump by interacting with associated polyphosphoinositides.* Biochem J, 1989. **261**(3): p. 1055-8.
382. Perez-Gordones, M.C., et al., *Diacylglycerol regulates the plasma membrane calcium pump from human erythrocytes by direct interaction.* Arch Biochem Biophys, 2009. **489**(1-2): p. 55-61.
383. Pinto Fde, T. and H.P. Adamo, *Deletions in the acidic lipid-binding region of the plasma membrane Ca2+ pump. A mutant with high affinity for Ca2+ resembling the acidic lipid-activated enzyme.* J Biol Chem, 2002. **277**(15): p. 12784-9.
384. de Tezanos Pinto, F. and H.P. Adamo, *Deletions in the A(L) region of the h4xb plasma membrane Ca(2+) pump. High apparent affinity for Ca(2+) of a deletion mutant resembling the alternative spliced form h4zb.* FEBS Lett, 2006. **580**(6): p. 1576-80.
385. Hilfiker, H., D. Guerini, and E. Carafoli, *Cloning and expression of isoform 2 of the human plasma membrane Ca2+ ATPase. Functional properties of the enzyme and its splicing products.* J Biol Chem, 1994. **269**(42): p. 26178-83.
386. Elwess, N.L., et al., *Plasma membrane Ca2+ pump isoforms 2a and 2b are unusually responsive to calmodulin and Ca2+.* J Biol Chem, 1997. **272**(29): p. 17981-6.
387. Huang, T.G. and D.D. Hackney, *Drosophila kinesin minimal motor domain expressed in Escherichia coli. Purification and kinetic characterization.* J Biol Chem, 1994. **269**(23): p. 16493-501.
388. Brini, M., et al., *Transfected aequorin in the measurement of cytosolic Ca2+ concentration ([Ca2+]c). A critical evaluation.* J Biol Chem, 1995. **270**(17): p. 9896-903.
389. Grati, M., et al., *Molecular determinants for differential membrane trafficking of PMCA1 and PMCA2 in mammalian hair cells.* J Cell Sci, 2006. **119**(Pt 14): p. 2995-3007.
390. Ficarella, R., et al., *A functional study of plasma-membrane calcium-pump isoform 2 mutants causing digenic deafness.* Proc Natl Acad Sci U S A, 2007. **104**(5): p. 1516-21.

391. Juranic, N., et al., *Calmodulin wraps around its binding domain in the plasma membrane Ca2+ pump anchored by a novel 18-1 motif.* J Biol Chem, 2010. **285**(6): p. 4015-24.
392. Domi, T., et al., *Functional specificity of PMCA isoforms?* Ann N Y Acad Sci, 2007. **1099**: p. 237-46.
393. Toyoshima, C., *Structural aspects of ion pumping by Ca2+-ATPase of sarcoplasmic reticulum.* Arch Biochem Biophys, 2008. **476**(1): p. 3-11.
394. Hofmann, F., et al., *The C-terminal domain of the plasma membrane Ca2+ pump contains three high affinity Ca2+ binding sites.* J Biol Chem, 1993. **268**(14): p. 10252-9.
395. Beroud, C. and T. Soussi, *The UMD-p53 database: new mutations and analysis tools.* Hum Mutat, 2003. **21**(3): p. 176-81.
396. Olivier, M., et al., *The IARC TP53 database: new online mutation analysis and recommendations to users.* Hum Mutat, 2002. **19**(6): p. 607-14.
397. Joerger, A.C. and A.R. Fersht, *Structure-function-rescue: the diverse nature of common p53 cancer mutants.* Oncogene, 2007. **26**(15): p. 2226-42.
398. Malecka, K.A., W.C. Ho, and R. Marmorstein, *Crystal structure of a p53 core tetramer bound to DNA.* Oncogene, 2009. **28**(3): p. 325-33.
399. Kitayner, M., et al., *Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs.* Nat Struct Mol Biol, 2010. **17**(4): p. 423-9.
400. Chen, Y., R. Dey, and L. Chen, *Crystal structure of the p53 core domain bound to a full consensus site as a self-assembled tetramer.* Structure, 2010. **18**(2): p. 246-56.
401. Petty, T.J., et al., *An induced fit mechanism regulates p53 DNA binding kinetics to confer sequence specificity.* EMBO J, 2011. **30**(11): p. 2167-76.
402. Canadillas, J.M., et al., *Solution structure of p53 core domain: structural basis for its instability.* Proc Natl Acad Sci U S A, 2006. **103**(7): p. 2109-14.
403. Nikolova, P.V., et al., *Semirational design of active tumor suppressor p53 DNA binding domain with enhanced stability.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14675-80.
404. Bullock, A.N., J. Henckel, and A.R. Fersht, *Quantitative analysis of residual folding and DNA binding in mutant p53 core domain: definition of mutant states for rescue in cancer therapy.* Oncogene, 2000. **19**(10): p. 1245-56.
405. Joerger, A.C., H.C. Ang, and A.R. Fersht, *Structural basis for understanding oncogenic p53 mutations and designing rescue drugs.* Proc Natl Acad Sci U S A, 2006. **103**(41): p. 15056-61.
406. Joerger, A.C., et al., *Structures of p53 cancer mutants and mechanism of rescue by second-site suppressor mutations.* J Biol Chem, 2005. **280**(16): p. 16030-7.
407. Baronio, R., et al., *All-codon scanning identifies p53 cancer rescue mutations.* Nucleic Acids Res, 2010. **38**(20): p. 7079-88.
408. Baroni, T.E., et al., *A global suppressor motif for p53 cancer mutants.* Proc Natl Acad Sci U S A, 2004. **101**(14): p. 4930-5.
409. Martin, A.C., et al., *Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein.* Hum Mutat, 2002. **19**(2): p. 149-64.
410. Danziger, S.A., et al., *Predicting positive p53 cancer rescue regions using Most Informative Positive (MIP) active learning.* PLoS Comput Biol, 2009. **5**(9): p. e1000498.

411. Schymkowitz, J., et al., *The FoldX web server: an online force field.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W382-8.
412. Heikkinen, K., et al., *Mutation screening of Mre11 complex genes: indication of RAD50 involvement in breast and ovarian cancer susceptibility.* J Med Genet, 2003. **40**(12): p. e131.
413. Tommiska, J., et al., *Evaluation of RAD50 in familial breast cancer predisposition.* Int J Cancer, 2006. **118**(11): p. 2911-6.
414. Assenmacher, N. and K.P. Hopfner, *MRE11/RAD50/NBS1: complex activities.* Chromosoma, 2004. **113**(4): p. 157-66.
415. Hopfner, K.P., et al., *Structural biochemistry and interaction architecture of the DNA double-strand break repair Mre11 nuclease and Rad50-ATPase.* Cell, 2001. **105**(4): p. 473-85.
416. Moreno-Herrero, F., et al., *Mesoscale conformational changes in the DNA-repair complex Rad50/Mre11/Nbs1 upon binding DNA.* Nature, 2005. **437**(7057): p. 440-3.
417. Hohl, M., et al., *The Rad50 coiled-coil domain is indispensable for Mre11 complex functions.* Nat Struct Mol Biol, 2011. **18**(10): p. 1124-31.
418. Tavtigian, S.V., et al., *Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer.* Am J Hum Genet, 2009. **85**(4): p. 427-46.
419. Tian, J., et al., *Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines.* BMC Bioinformatics, 2007. **8**: p. 450.
420. Brunak, S., J. Engelbrecht, and S. Knudsen, *Prediction of human mRNA donor and acceptor sites from the DNA sequence.* J Mol Biol, 1991. **220**(1): p. 49-65.
421. Cartegni, L., et al., *ESEfinder: A web resource to identify exonic splicing enhancers.* Nucleic Acids Res, 2003. **31**(13): p. 3568-71.
422. Fariselli, P. and R. Casadio, *RCNPRED: prediction of the residue co-ordination numbers in proteins.* Bioinformatics, 2001. **17**(2): p. 202-4.
423. Kapplinger, J.D., et al., *An international compendium of mutations in the SCN5A-encoded cardiac sodium channel in patients referred for Brugada syndrome genetic testing.* Heart Rhythm, 2010. **7**(1): p. 33-46.
424. Zimmer, T. and R. Surber, *SCN5A channelopathies--an update on mutations and mechanisms.* Prog Biophys Mol Biol, 2008. **98**(2-3): p. 120-36.
425. Payandeh, J., et al., *The crystal structure of a voltage-gated sodium channel.* Nature, 2011. **475**(7356): p. 353-8.
426. Rook, M.B., et al., *Biology of cardiac sodium channel Nav1.5 expression.* Cardiovasc Res, 2012. **93**(1): p. 12-23.
427. Wehrens, X.H., et al., *A novel mutation L619F in the cardiac Na+ channel SCN5A associated with long-QT syndrome (LQT3): a role for the I-II linker in inactivation gating.* Hum Mutat, 2003. **21**(5): p. 552.
428. Franke, A., et al., *Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci.* Nat Genet, 2010. **42**(12): p. 1118-25.
429. Anderson, C.A., et al., *Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47.* Nat Genet, 2011. **43**(3): p. 246-52.

430. Bettelli, E., et al., *Induction and effector functions of T(H)17 cells.* Nature, 2008. **453**(7198): p. 1051-7.
431. Moschen, A.R., et al., *The RANKL/OPG system is activated in inflammatory bowel disease and relates to the state of bone loss.* Gut, 2005. **54**(4): p. 479-87.
432. Raychaudhuri, S., et al., *Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions.* PLoS Genet, 2009. **5**(6): p. e1000534.
433. Rossin, E.J., et al., *Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology.* PLoS Genet, 2011. **7**(1): p. e1001273.
434. Momozawa, Y., et al., *Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease.* Nat Genet, 2011. **43**(1): p. 43-7.
435. Vissers, L.E., et al., *A de novo paradigm for mental retardation.* Nat Genet, 2010. **42**(12): p. 1109-12.
436. Bansal, V., et al., *Statistical analysis strategies for association studies involving rare variants.* Nat Rev Genet, 2010. **11**(11): p. 773-85.
437. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet, 2011. **43**(5): p. 491-8.
438. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Res, 2010. **38**(16): p. e164.
439. Hugot, J.P., et al., *Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease.* Nature, 2001. **411**(6837): p. 599-603.
440. Miceli-Richard, C., et al., *CARD15 mutations in Blau syndrome.* Nat Genet, 2001. **29**(1): p. 19-20.
441. Roe, T.F., et al., *Inflammatory bowel disease in glycogen storage disease type Ib.* J Pediatr, 1986. **109**(1): p. 55-9.
442. Zweig, M.H. and G. Campbell, *Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine.* Clin Chem, 1993. **39**(4): p. 561-77.
443. Ginsburg, G.S. and H.F. Willard, *Genomic and personalized medicine: foundations and applications.* Transl Res, 2009. **154**(6): p. 277-87.
444. Teslovich, T.M., et al., *Biological, clinical and population relevance of 95 loci for blood lipids.* Nature, 2010. **466**(7307): p. 707-13.
445. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.
446. Gage, B.F., et al., *Use of pharmacogenetic and clinical factors to predict the therapeutic dose of warfarin.* Clin Pharmacol Ther, 2008. **84**(3): p. 326-31.
447. Anstee, D.J., *Red cell genotyping and the future of pretransfusion testing.* Blood, 2009. **114**(2): p. 248-56.
448. Rich-Edwards, J.W., et al., *Birthweight and the risk for type 2 diabetes mellitus in adult women.* Ann Intern Med, 1999. **130**(4 Pt 1): p. 278-84.
449. Morris, D.H., et al., *Determinants of age at menarche in the UK: analyses from the Breakthrough Generations Study.* Br J Cancer, 2010. **103**(11): p. 1760-4.

450. Punta, M., et al., *Membrane protein prediction methods.* Methods, 2007. **41**(4): p. 460-74.
451. Lam, K., et al., *Identification of variants in CNGA3 as cause for achromatopsia by exome sequencing of a single patient.* Arch Ophthalmol, 2011. **129**(9): p. 1212-7.

Appendix – Supplementary Tables

In the following, the supplementary tables from published articles used throughout the thesis are listed together with the URL where these have been deposited.

Supplementary Table S.4.1. Information about the 35 pVHL interacting proteins.
URL: doi:10.1016/j.febslet.2009.10.070
http://www.sciencedirect.com/science/article/pii/S001457930900862X#

Supplementary Table S.5.1. Summary of VHL variants and related clinical information.
URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1469-1809.2011.00647.x/suppinfo

Supplementary Table S.5.2. Results of prediction methods for known VHL variants.
URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1469-1809.2011.00647.x/suppinfo

Supplementary Table S.5.3. Results of prediction methods for novel VHL variants.
URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1469-1809.2011.00647.x/suppinfo

Supplementary Table S.6.1. Analysis of LGI mutations with stability change prediction methods.
URL:http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3066209/?tool=pubmed#pone.0018142.s003

Acknowledgements

I thank Prof. Silvio Tosatto for his advice and friendship. He encouraged me to be creative and made my research more enjoyable and productive. I would also like to thank him for guiding my research into interesting directions and giving me the opportunity to travel around the world. In the future I hope to meet other supervisors or colleagues with a similar attitude for research and management of a group.

I also want to thank all past and present members of the BioComputing group, each has given me the possibility to discover a different culture and diverse ways to approach and solve scientific and non-scientific problems. In particular, thanks to Manuel who patiently explained to me how a computer scientist thinks. I cannot forget my best friend, Filippo, thanks for your technical and moral support. Also thanks to all for the enjoyable social events, I hope there will be many more.

A special thanks is due to my family. To my parents, brothers, sisters, nephews and nieces thanks for your support and for the chaotic and funny weekly meetings. I will remember that “life is not only work” but I hope to convey my passion for science.