Università degli Studi di Padova

Sede Amministrativa: Università degli Studi di Padova
Dipartimento di Scienze Biomediche

CORSO DI DOTTORATO DI RICERCA IN SCIENZE BIOMEDICHE SPERIMENTALI
CICLO: XXX

FROM HIGH-THROUGHPUT ANALYSIS OF GENETIC VARIANTS TO THE EXPERIMENTAL VALIDATION OF PUTATIVE PROTEIN FUNCTION

Coordinatore: Ch.mo Prof. Paolo Bernardi
Supervisore: Ch.mo Prof. Silvio C. E. Tosatto
Co-Supervisore: Dott.ssa Emanuela Leonardi

Dottoranda: Alessandra Gasparini

Abstract

The state-of-the-art approach for the genetic molecular cause research relies on massively parallel gene sequencing, which represents a challenge both in data handling and variant prioritization. The univocal assignment of disease pathogenicity to the sequence variants is often difficult, and requires the integration of different lines of evidence for a comprehensive interpretation. During my thesis, I contributed to the development of novel approaches to evaluate rare variant contribution to the clinical phenotype. These methods were presented and evaluated at the Critical Assessment of Genome Interpretation, ranking among top programs considering either performance or the number of correct assigned disease predictions. A similar strategy was employed for identification of disease genes linked to neurodevelopmental disorders (NDDs) comorbidity. In this case, computational methods were applied to select the most promising candidate genes for the design of diagnostic panel, which is currently used for patient screening at Pediatrics Clinic of the University of Padova. The variants found within the panel genes have been selected according frequency, pathogenicity prediction and variant segregation analysis within the family. Furthermore, I took advantage of different computational tools to investigate the mutated gene function, and used this information to demonstrate the impact of likely pathogenic variant on clinical phenotype. In several cases, likely pathogenic mutations mapped to intrinsically disordered regions (IDRs), which lack a fixed three-dimensional structure. Coherently, several studies demonstrate that mutations in IDRs are often associated with the pathogenesis of various human diseases. Thus, IDRs classification could represent a critical step for understanding the impact of possibly disease-causative variants mapping in these regions. Due to the influence of intrinsically disordered proteins (IDPs) in diseases, I participated to the manual curation and update of entries in the DisProt database, the primary repository of disorder-related data on sequence. Interestingly, increasing evidence from literature highlights the IDPs involvement in neuronal signal transduction. Among the proteins encoded by diagnostic panel genes, TANC2 especially emerged as intrinsically disordered protein with a possible role in synaptic signal transduction. As TANC2 and its protein family function was poorly characterized, I performed an *in silico* analysis to characterize the TANC protein activity, and the implicated biological processes. The functional hypothesis emerged from the

bioinformatics analysis was used to drive further experimental investigations. *In vitro* validation of predicted TANC2-CDKL5 interaction highlighted the relevance of the IDRs in regulating degradation of CDKL5, whose mutations are associated with a heterogeneous set of NDD phenotypes. Furthermore, I demonstrated that TANC2 contributes to downregulate CDKL5 expression levels. For this reason, TANC2 protein could represent a novel therapeutic target to design new drugs for the treatment of CDKL5 over-expression associated diseases.

Riassunto

La strategia di elezione per l'identificazione di varianti causative di malattie genetiche consiste nell'utilizzo di piattaforme di *Next Generation Sequencing.* Questo tipo di approccio rappresenta una sfida, sia per quanto riguarda la gestione della mole di dati da sequenziamento, che per l'interpretazione clinica dei risultati. L'identificazione di varianti chiaramente implicate nella determinazione della patologia è un processo complesso, che richiede l'integrazione di diversi tipi di informazione. Durante il mio dottorato, ho contributo all'implementazione di metodi computazionali per predire la probabilità che un determinato genotipo sia associato al fenotipo clinico di interesse. Questi metodi sono stati presentati, e valutati, in occasione del *Critical Assessment of Genome Interpretation* (CAGI), dove si sono posizionati tra i migliori classificati sia per prestazioni che numero di predizioni corrette. Una strategia analoga è stata applicata all'identificazione di geni implicati nella comorbidità tra disordini del neurosviluppo. Anche in questo caso, l'utilizzo di tecniche bioinformatiche si è reso fondamentale per la selezione di geni candidati, che sono stati poi utilizzati nella progettazione di un pannello genico diagnostico attualmente in uso presso la Clinica Pediatrica dell'Università di Padova. Data la gran quantità di dati prodotti per esperimento, le varianti trovate nei geni inclusi nel pannello sono state filtrate in base alla frequenza, alla predizione di patogenicità e all'analisi di segregazione all'interno della famiglia. In alcuni casi, un ulteriore contributo a supporto dell'effettiva patogenicità della variante è stato dato dall'analisi bioinformatica della proteina mutata. Frequentemente, la variante candidata provoca alterazioni a livello di regioni intrinsecamente disordinate (IDR), caratterizzate dall'assenza di una conformazione tridimensionale stabile. Questo dato è coerente con la più recente letteratura: diversi studi, infatti, dimostrano l'implicazione di mutazioni nelle IDR in diverse patologie umane. La classificazione delle IDR, quindi, può rappresentare un primo passo per comprendere l'impatto di eventuali varianti causative all'interno di queste regioni. Data la rilevanza delle IDR a livello biologico e clinico, ho partecipato alla curazione manuale e all'aggiornamento delle voci presenti nel database DisProt, la principale banca dati relativa al disordine nelle proteine. È interessante notare che, tra i vari processi biologici in cui le IDR sono coinvolte, queste regioni svolgono un ruolo molto importante nel *signaling* neuronale. Tra le proteine codificate dai geni inclusi nel pannello genico, TANC2 si è distinta per essere una proteina

disordinata, probabilmente implicata alla trasduzione del segnale a livello delle sinapsi neuronali. Dato che la funzione di TANC2 e della rispettiva famiglia proteica risultava ancora poco chiara, ho eseguito un'analisi *in silic*o delle proteine TANC, grazie alla quale è stato possibile caratterizzare le funzioni e i diversi processi cellulari in cui queste sono coinvolte. Le ipotesi funzionali emerse dall'analisi bioinformatica sono state utilizzate per condurre ulteriori indagini sperimentali. In particolare, la validazione *in vitro* dell'interazione TANC2-CDKL5 ha evidenziato l'estrema importanza di regioni intrinsecamente disordinate nella regolazione della degradazione di CDKL5, le cui mutazioni sono associate con manifestazioni cliniche legate a disordini del neurosviluppo. Inoltre, gli esperimenti hanno dimostrato che TANC2 contribuisce alla down-regolazione dei livelli di espressione di CDKL5. Per questo motivo, TANC2 si candida a rappresentare un nuovo target terapeutico per lo sviluppo di nuovi composti per il trattamento di condizioni cliniche associate all'over-espressione di CDKL5.

Contents

Figure Index

Table Index

1 Introduction

Since the first publication of the human genome sequence, human genetics and genomics have been significantly improving[1]. The release of the raw sequence by the Human Genome Project (HGP) prompted multiple secondary studies, aimed at improving our understanding in genome architecture and function. According to the initial annotation, there are at least 20,000-25,000 protein-coding genes in human genome[1,2]. Understanding the function of genes and related encoded proteins is the unavoidable starting point to design effective diagnostic tools and therapies for genetic pathologies[2]. Next-generation sequencing (NGS) has revolutionized medical research and clinical diagnostics in the last decades[3]. Indeed, NGS employs powerful massively parallel sequencing, allowing to screen from subsets of few genes to the full human DNA sequence at once[3,4]. Targeted exon capture before genomic sequencing, i.e. the whole exome sequencing (WES), covers the analysis of most of the coding regions, which is less than 2% of the genome. Conversely, whole genome sequencing (WGS) can inspect also non-coding and regulative regions, allowing the identification of splicing site variants, enhancers, as well as promoter regions[2,4]. In both cases, hundreds to millions of genetic variants are detected from a single individual[5]. Thus, the most demanding challenge in human genetics currently consists in isolating a (small) subset of genomic variations that can be proven to be causative for a disease phenotype[2,5,6]. There are plentiful online prioritization tools and interpretation sources to address the variant filtering and interpretation. However, the first common step required is annotation, by which information about the variant location and effects are added. The variant effect describes how the variation influences the reference sequence characteristics, and it is defined by the Sequence Ontology. A common practice is to annotate variants based on the transcripts with the most severe effects, allowing to identify all potentially causative variants[5]. Furthermore, additional information can be added by the annotation tools, such as data from disease associated mutation databases (e.g. ClinVar, and COSMIC), population allele frequency, and large scale genome and exome sequencing, such as the 1000 Genome Project and ExAC consortium[5]. Despite being a mainly automated and preliminary process, this step is a powerful resource for setting the *a priori* expectation for mutation impact on the pathology, allowing the variant prioritization[5]. Generally, an initial selection is made considering Sequence Ontology variant classification. The assumption is

that variants affecting protein-coding regions, e.g. non-synonymous or frameshift single nucleotide variants (SNVs), are more likely to be damaging in respect to synonymous or intronic variations[5]. Conventional approaches use conservation and protein structure to predict the consequence on the protein function from missense changes, and integrate allele frequency and gene conservation into a prioritization workflow (see Chapter 2 for more details). Even though generally correct, these methods convey some limitations. Stop-codon and frameshift SNVs are either ignored at all, or systematically assigned with the highest damaging score by prediction tools[7]. Moreover, some classes of genes, e.g. large genes and paralog gene family members, are more likely to bear "damaging" variants by chance, either for the amount of nucleotides comprised in the gene, or due to probe mapping errors[7]. Of note, not all damaging variants are causal mutations, as they do not necessarily have an effect on the clinical phenotype[6,7]. For these reasons, variant genotype frequency in population should be taken into consideration, excluding those variants recurrently found in healthy individuals, and with a frequency higher than the disease prevalence[6,7]. Further complications are represented by non-coding and synonymous SNVs, which do not directly affect encoded protein function. The difficulty in predicting the non-coding variant effects relies in the insufficient knowledge of regulatory elements in non-coding DNA[7]. Conversely, emerging observations are consistent with the concept that synonymous variants can influence a broad range of molecular mechanisms, such as splicing and miRNA regulation, though only early achievements have been made in predicting their impact[8]. In addition to the SNVs and small insertion/deletions (*indels*), variants implicated in human diseases also include structural variants (SVs), such as copy number variations (CNVs), large deletions and duplication, and balance rearrangements[2]. Due to their width, their phenotypic effects are potentially large, but hard to assess, as they can affect gene dosage, disrupting several genes, and regulatory elements at once[5]. Moreover, structural variants are difficult to observe, and WGS is the preferred approach for SV identification, despite the cost and complexity of results[5]. A body of literature[9–11], however, proves that structural variant identification to improve our knowledge and comprehension of mechanisms underlying both Mendelian and complex diseases, and provides useful information needed to establish a proficient diagnosis[5]. Exome sequencing is now the most commonly used tool for genetic disease gene discovery, i.e. the identification of novel disease associated genes[4,6,12]. Considered for diagnostic purpose, WES is extremely useful and cost effective when applied to cases remaining undiagnosed from previous multiple single-gene tests[6,12]. Indeed, targeted exon sequencing does not require knowledge *a priori* about patients'

disease[6,12], as it does not focus on a precise subset of genes, making it a better diagnostic tool for disorders with nonspecific symptoms[12]. Despite their contribution to human genetic research[4,6,12], WES/WGS routinely application to medical practice is generally complicated, both by the cost, the unprecedented scale of data to handle, and challenges in variant interpretation[4,12]. Moreover, these approaches strike a balance between coverage depth and the large quantity of targeted regions, resulting in a reduced clinical sensitivity for lower covered regions[12]. In cases of diseases with a limited genetic heterogeneity, a more appropriate tool for efficient molecular diagnosis consists in the targeted re-sequencing of a selected subset of well-established disease genes, i.e. a gene panel[12]. The gene inclusion criteria consider manifold aspects, ranging from the previously single-tested genes with a strong disease association, to the novel gene gathering from literature and specific case reports[2,12]. Targeted gene re-sequencing generally involves coding regions of the panel genes, reaching a significantly higher coverage depth in all sequenced regions compared to WES or WGS[12]. In clinical practice, gene panel tests are systematically complemented with Sanger sequencing, either for in low-coverage and missed exons coverage, or to confirm variants[12]. As in other NGS applications, the interpretation outcome largely depends on the curation of clinical phenotype information, and the strength of evidence supporting gene association to the disease[6,7]. Indeed, after the preliminary automated prioritization, the variants interpretation mainly relies on experts review and human curation according standardized guidelines[5]. Despite validated by direct sequencing, genetic data itself does not provide a direct diagnosis, but has a primarily role. Generally, several types of clinical and genetic tests are combined to maximize the diagnostic yield[4,12]. A major challenge is shifting genetic information, i.e. variant prioritization score, into the main diagnostic tool to drive clinical decisions and patients' long-term treatment upon integration with clinical observations [4,12]. In this context, methodologies investigating the mutated gene function can be useful in demonstrating the impact of candidate mutations on clinical phenotype, as well as in identifying the disease related molecular mechanisms [4,6]. Lots of possibly disease variants map to newly discovered genes, whose role in determining a clinical phenotype is still under validation[12]. A first line of evidence can be obtained by an extensive analysis *in silico*, aimed at collecting as much data as possible about protein structure, function, and interactions[13] (see Chapter 2 for details). Structural characterization is a key step in understanding the role of proteins in cellular processes[13]. Indeed, either experimentally determined or *in silico* modeled structures allow the identification of residues and regions, which are fundamental for protein functioning, e.g. globular

domains[13]. Moreover, functional motifs are also located in unstructured regions, to include motifs mediating either direct binding or post-translational modifications, and signals for cellular localization[14,15]. Proteins do not work as isolated system. They exert their activity through interactions with other polypeptides, forming the so called protein-protein interaction networks[16,17]. As causative mutations generally yields perturbation of key cellular pathways[17,18], the modeling of protein-protein interactions can help to shed light on molecular mechanisms involved in a specific pathology[16]. Secondly, functional hypothesis emerging from bioinformatics predictions can drive further experimental validation. Several techniques *in vivo* can be employed for this purpose, including animal model systems and, evidence derived from patient's tissues, though a valuable support can be provided by cell culture tests or *in-vitro* assays for protein-protein interaction assessment as well[2,4,6].

1.1 Contribution of the thesis

The leading thread of my thesis consists in deciphering the molecular causes of genetic diseases. The project involves both the application of bioinformatics tools for variant interpretation and protein analysis, and *in vitro* experiments for hypothesis validation (see Chapter 2). In the former section of this thesis (Chapters 3-5), I present the application of computational methods for NGS data analysis. This section includes the publications related to the Critical Assessment of Genome Interpretation (CAGI) international experiment and the work aimed to the design of a gene panel for the neurodevelopmental disorder (NDD) diagnosis. Chapter 6 consists in the publication related to the last DisProt database release. Chapter 7 summarizes the bioinformatics strategies employed for functional and structural characterization of TANC protein family. The hypothesis emerged by the *in silico* analysis was used to drive further experimental investigations, reported in Chapter 8. This section is followed by a conclusive part, where I discuss the results of my PhD research. Hereafter, a summary of each contribution to the different tasks addressed in my thesis project is reported.

1.1.1 Challenges in genotype-phenotype association in genetic diseases

Despite the increasing employment of NGS technologies in medical practice, the computational sequence data analysis remains challenging and critical for successful

interpretation of results[3]. In this context, Critical Assessment of Genome Interpretation (CAGI) international experiment aimed to assess the accuracy of different bioinformatics tools in predicting the functional impact of genetic variants and their relevance in determining a clinical phenotype[19], promoting their application in the medical practice. Chapter 3 is based on the published article: *Daneshjou, R., et al. Working towards precision medicine: predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. Human Mutation 38, no. 9 (September 1, 2017): 1182-1192. doi:10.1002/humu.23280.* This Chapter summarizes the results for exome-sequencing based challenges of the fourth CAGI edition (Crohn's disease, bipolar disorder, and warfarin dosing). Our group participated to two challenges: the bipolar disorder, and Crohn's disease challenges. Both bipolar disorder and Crohn's diseases are complex pathologies, for which the interplay among genetic and environmental factors is still debated. Bipolar Disorder (BD) is a common mood disorder characterized by episodes of manias and depression with a high component of heritability[19]. However, two decades of research with linkage and association studies failed to identify any susceptibility gene[19]. Crohn's disease (CD) is a chronic inflammatory bowel disease, caused by the complex interplay between autoimmune response and environmental factors in genetically susceptible individuals[19]. For both the diseases, the challenge was the (blind) identification of affected patients from a cohort of hundreds of exomes, in which both affected patients and healthy controls were present. My contribution to the prediction method definition consisted in integrating genetic information used for variant filtering and prioritization. I performed an extensive literature review to identify, list and select all genes and genetic variants associated with either BD or CD traits. For the BD challenge, the variant sets were selected favoring those mapping to genes involved in nervous system development or pathways impaired by the pathology, and used to train the neural network implemented for variant effect prediction. In the case of Crohn's disease, manually curated lists allowed to filter patient variants mainly considering known variants from WES studies. A particular attention was placed on very early onset cases and rare variants (MAF $\leq$ 5%) probably affecting the activity and/or expression of disease associated genes or direct interactors of Crohn's genes identified by STRING[20]. In both cases, our methods were statistically more efficient than random algorithm in assigning the correct phenotype, better discriminating between genetic and environmental factors.

Chapter 4 is based on *Chandonia, J.-M, et al. Lessons from the CAGI-4 Hopkins clinical panel challenge. Human Mutation 38, no. 9 (September 1, 2017): 1155–68. doi:10.1002/humu.23225.* This Chapter recapitulates the results of Hopkins clinical panel challenge from the fourth edition of CAGI. The dataset was provided by the Johns Hopkins' DNA Diagnostic Laboratory, and it included data from 106 patients, resulting from the targeted re-sequencing of 83 disease genes. The participants were asked to group the panel genes into 14 diverse disease classes (e.g. lung disorder, and craniofacial disorders), and to predict the putative disease-causing genes and variants for each patient basing only on sequencing data. In this challenge, I curated the Hopkins panel gene assignment to the clinical phenotypes based on literature. I also contributed to the candidate variant selection, considering only exonic variants and filtering out common (MAF > 5%) and synonymous variants. Finally, I participated in the development of a disease-probability scoring scheme, helping with the variant effects ranking according to the disease inheritance mode and variant state (e.g. heterozygosity/homozygosity). Our method ranked third according to the number of correct predicted phenotypes, resulting in one of the state-of-the-art algorithms for predicting clinical phenotype-genotype association.

In different genetic conditions, disease-associated genes are numerous and, thus, single-gene testing generally fails to provide an accurate diagnosis[21]. This is the case of Neurodevelopmental disorders, which represent the main thread of my PhD research. NDDs are common conditions including clinically heterogeneous diseases. Due to the wide genetic heterogeneity and recurrent overlapping clinical features, single-gene testing for diagnosis of NDD is especially challenging[21,22]. As consequence, high-throughput methods, such as NGS targeted gene re-sequencing are increasingly employed for NDD genetic testing[21,22]. Chapter 5 collects the computational techniques used for identification of a subset of genes involved in autism and intellectual disability co-morbidity for the development of a diagnostic gene-panel. In this project, I worked on the candidate gene list generation, gathering data from publicly available sources such as disease-specific databases, exome sequencing studies and meta-analysis publications. During the gene selection process, I curated the annotation and enrichment analysis of the candidate genes used for the final panel gene list identification. The so-selected gene panel is currently employed in clinical screening of individuals affected by intellectual disability (ID) and/or autism spectrum disorder (ASD) and referred to the Molecular Genetics of Neurodevelopment Laboratory (Paediatric Department, University of Padova) for genetic testing. Given the huge amount of NGS data per patient, the following analysis step

consisted in variant selection. Specifically, I contributed to the filtering and interpretation of patient genetic variants, integrating predictions with literature, clinical findings and case specific research data. For the most promising missense variants, I performed an *in silico* evaluation of the effects on protein function/structure. Our analyses allow us to assign a molecular diagnosis to twenty-four of the screened patients, with a diagnostic yield of 16,4%. For twenty-three probands, at least one likely pathogenic (LP) variant has been detected. The causative role of LP variants is under assessment and will be established through segregation analysis.

1.1.2 Annotation and classification of intrinsically disordered proteins

Intrinsically disordered proteins (IDPs) are proteins lacking a fixed or ordered three-dimensional structure, which folding state ranges from fully unstructured to partially structured. Interestingly, many proteins encoded by genes in ASD/ID panel contain intrinsically disordered regions (IDRs) that were found mutated in 45,6% (26/57) of patients bearing at least a possibly causative variant. This is consistent with the evidence in literature proving the association of mutations in IDRs with numerous human diseases[23]. Indeed, despite the lack of stable conformation, IDPs play an important role in regulatory and signaling processes of the cell, frequently acting as hubs in protein-protein interaction network[24]. Due to the relevance of IDPs in cellular processes and clinical phenotypes, characterization and classification of IDRs and IDPs should be viewed as a crucial step for understanding the impact of possibly disease-causative variants mapping within these regions. For these reasons, our group decided to update and manual curate the entries forming the DisProt database. DisProt is the primary database of disorder-related data on sequence- and functional annotations, being primarily focused on IDPs or IDRs with experimental validation. Chapter 6 is based on the published article: *Piovesan, D., et al. DisProt 7.0: a major update of the database of disordered proteins. Nucleic Acids Res. 45, D1123–D1124*. In this work, I contribute to the manual curation of 32 DisProt entries, totaling 101 annotated intrinsically disorder regions. The annotation process involved both newly added IDPs and entries from the previous release of the database. In both cases, the first step of analysis consisted of identifying the position of each IDR, either using MobiDB[25] predictions or retrieving experimental data reported in literature. Then, each entry was associated to the PubMed ID related to the experimental evidence, e.g. circular dichroism or nuclear magnetic resonance, confirming the disordered state and the function

of the annotated region. The new release, DisProt 7.0, contains more than 800 intrinsically disordered proteins, entries of the previous one, resulting in the most valuable resource for a better understanding of the structural disorder.

1.1.3 Molecular mechanisms involved in neurodevelopmental disorders: focus on TANC2

Due to their structural and functional plasticity, IDPs are highly represented in regulation and recognition processes, including signal transduction in neurons[26]. Indeed, disordered proteins mediate interactions with their targets with relatively high specificity and low affinity, which could trigger signaling events and favoring rapid disassociation when signaling is completed[26]. In neurons, these functions are exerted by the scaffold proteins. Scaffold proteins typically contain IDRs and domains for protein binding, by which selectively gather specific proteins within signaling pathways, influencing the specificity and kinetics of signaling interactions[27]. The ASD/ID panel comprises known and most promising candidate disease-associated genes, including post-synaptic scaffold proteins. In some cases, the scaffold protein function is not fully understood, and *in silico* analyses aimed at characterizing of protein structure/function could address this issue. Two examples are the TANC2 protein, and its paralog TANC1, which I extensively analyzed taking into advantage of different bioinformatics tools. Chapter 7 is based on *Gasparini A., Dynamic scaffolds for neuronal signaling: in silico analysis of the TANC protein family Sci Rep. 2017 Jul 28;7(1):6829.* This work had two main purposes: the functional and structural characterization of TANC protein family and the identification of the TANC interaction network, aimed to elucidate TANCs influence on neuronal pathways. Integrating structural and functional elements, this work provides the basis for interpretation of genetic variants found in TANC encoding genes. Furthermore, our findings shed light on possibly molecular mechanisms, which would be worth of experimental validation, such as the assessment of the interactions with key proteins in regulating synapse function. Chapter 8 describes the experimental techniques employed for the validation of TANC2 interaction with PP1 and CDKL5, and the functional significance of the TANC2-CDKL5-PP1 interplay (paper in preparation, presenting author). To validate our hypothesis, I assessed the interaction among endogenous full length proteins, both by co-localization in different cell lines, and immune-precipitation from rat synaptosomes. Furthermore, I outlined TANC2, PP1 and CDKL5 minimal interacting regions through

yeast two-hybrid system. To assess the functional relationship among regulative TANC2 activity and CDKL5 protein levels, I evaluated the effects of TANC2 silencing in SHSY5Y cells. The experiments suggest that TANC2 forms complexes both with CDKL5 and PP1, and links the phosphatase PP1 to its substrate CDKL5, allowing its dephosphorylation and subsequent degradation. Given its role in downregulation of CDKL5 expression levels, we propose TANC2 as a new therapeutic target for the treatment of clinical phenotypes associated to CDKL5 duplication, though further investigations are required to confirm our hypothesis.

2 NGS data analysis and interpretation

Next-generation sequencing (NGS) has revolutionized medical research and clinical diagnostics in the last decades, allowing to screen from subsets of few genes to the whole genome at once[3]. Millions of single nucleotide variants (SNVs) are identified per genome, resulting in the most common type of genetic differences within population[28]. Therefore, distinguishing diseases causing genetic variants from the thousands of potential candidates represents the prevailing challenge in human genetics[2,3,28]. In this Chapters, I discuss the workflow aimed to identification and interpretation of disease causative variants from sequencing data. Data interpretation is a multidimensional task, which relies on the analysis of sequence data *per se*, patient phenotype and literature review, and can be integrated with functional information inferred by *in silico* analysis. The first part of this Chapter deals about the strategies commonly employed for variant/gene prioritization, and publicly available sources of genetic data used for variant interpretation, whereas the description of computational methods to predict protein function is provided in the second section. As disease genes are highly interconnected[29], prediction and validation of protein-protein interactions can shed light on molecular mechanism involved in disease pathogenesis. The approaches for protein-protein interaction (PPI) assessment are discussed in the last part of the Chapter.

2.1 Variant filtering

After obtaining the data from a NGS experiment, the variant filtering aims at detecting potentially causing disease variants. The process relies on a multistage workflow; whose first step consists in focusing on those variants mapping to known disease associated genes or discovered by gene prioritization. Then, different parameters are taken into account for variant selection, e.g. allele frequency, predicted pathogenicity, *a priori* knowledge about the variant and/or the pathology, which will be discussed as following.

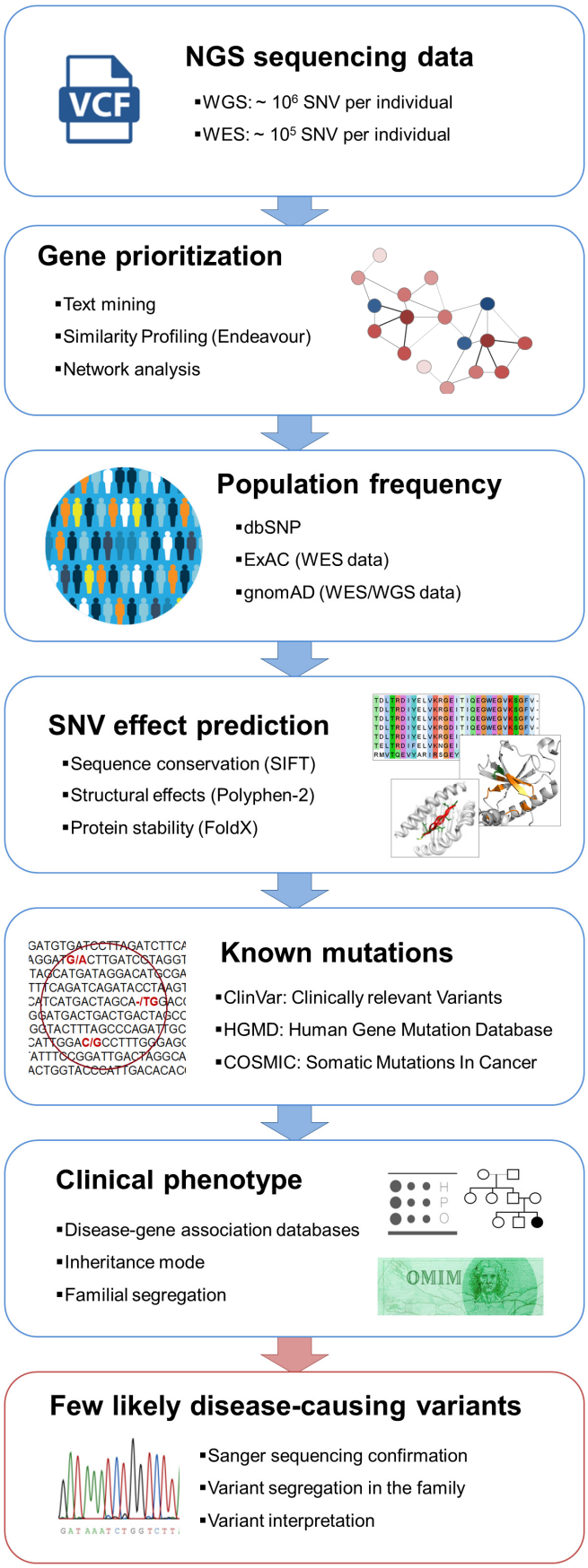


Figure 2.1: Schematic representation of a general variant investigation workflow

2.1.1 Gene prioritization

The main goal of gene prioritization is the identification of the most promising disease related genes (and variants). Gene prioritization largely relies on prior knowledge about gene association to the disease, and can be applied in different scenarios, either for the design of targeted re-sequencing panel, or to direct exome/genome sequencing data analysis[6,12]. However, the need of prioritization methods is particularly pressing in case of highly impacting clinical conditions, for which underlying molecular causes remain mostly uncharacterized[30]. Moreover, besides the consequences on patients' diagnosis, identification of novel disease genes represents the first step in the understanding of the protein and molecular pathways involved in genetic disorder, and, thus in developing targeted therapeutic strategies[4,6]. *In silico* techniques for prioritization are numerous, and require two main inputs, i.e. a list of candidate genes to prioritize, and the criteria considered for gene ranking, typically a list of keywords, or a set of training genes[31], respectively. The choice of training keywords/ "seed" genes is crucial for prioritization outcome and selected elements generally have a well-established connection with the clinical phenotype of interest. It is necessary to assess the relevance of each seed gene to the disease across different sources, e.g. from a locus based repository or specific research data. Moreover, the number of seed genes greatly affect the outcome of the gene prioritization process. A dataset with less than five genes is uninformative, whereas large gene lists are likely to be too heterogeneous and could produce unreliable results[31]. Another element greatly influencing the analysis quality is the identification of candidate genes for prioritization, which is extrapolated from different kind of sources, such as disease specific databases (e.g. AutismKB[32], SFARI database[33]). Prioritized candidate genes can be either filtered, selecting only a small subset of the most promising genes, or ranked according to the training list. In the latter case, gene ranking strategies can be further divided in three categories: text mining, similarity profiling and network analysis. The first step of text mining workflow includes gathering of keywords/data relevant for the clinical phenotype from literature. Then, gene terms present in the texts are gathered, and statistically assessed for strength of extracted information. Data mining from literature represent the most conservative ranking approach, as it generally identifies only genes with the strongest evidence of disease association[31]. Conversely, similarity-based profiling methods compare candidate genes to the training set, favoring genes which are similar to known seed genes according to a particular feature, e.g. the cellular process. These methods are based on the

assumption that at least one among prioritized genes is enriched for the feature of interest[34]. One example is Endeavour [34], which integrates seventy-five different sources (e.g. genomic data, expression levels, and clinical phenotype) into a comprehensive ranking score, allowing prediction of novel disease genes[34]. Endeavour prioritization can be started either by selecting the species of interest or defining the seed gene list. It allows to choose all data sources to be used in the prioritization of the candidate loci. The seed gene list is used to build a feature of interest model according to each source (e.g. gene ontology) by which the candidate genes are scored and ranked. The individual weights are then integrated in a global score correlating with the gene association to the phenotype of interest, e.g. a specific disease trait. Despite their effectiveness, similarity profiling algorithms have been recently integrated or replaced by network analysis strategies. Network-based methods rely on the assumption of "guilt by association", by which genes share molecular and phenotypic features with their direct interactors[31]. Indeed, genes and respective encoded proteins exert their function by interacting with other molecules, generally meaning that mutations in disease associated genes yield to perturbations in key cellular pathways, directly affecting the protein interaction network (PIN)[18]. A network is defined by means of nodes (proteins) and edges (e.g. protein-protein interactions, PPI), where the number of connections linking nodes reflects its centrality, thus the importance of the gene/protein within the network and for the clinical condition[18]. Genes involved in a same pathology display more interactions among themselves than what would be expected for a random set of genes, allowing to identity disease specific clusters[17]. The nodes can be retrieved either from experimentally determined interaction repositories (such as InAct[35], Biogrid[36], and STRING[20]) or from large scale proteomics studies. As for other prioritization methods, the network based algorithms aim at identifying genes (i.e. nodes) relevant to the clinical phenotype or biological process, and take advantage of prior information, such as gene expression or presence of known associated causative mutations, to select the best candidates[17,31]. Different approaches can be considered in analyzing networks. The so-called linkage methods consider the direct interactors of disease genes in the network as disease genes themselves. Conversely, disease module based strategies focus on identifying modules enriched for disease-associated genes in a given larger network (e.g. tissue specific interactomes). Other algorithms, such as the network propagation methods, based on random walker algorithms, that are allowed to diffuse from a node along all the connections of the network, moving to any neighboring gene with equal probability. In this way, the most often "visited" nodes are assigned with a high probabilistic weight based on the

number of connections with known disease genes, which is in turn used for gene ranking[18]. Both linkage and disease modules based techniques are quite effective strategies for gene discovery, and they are generally validated by demonstrating that selected genes belong to related pathways, or are expressed in the same cell type[18]. This kind of information can be obtained automatically by gene set enrichment analysis. Enrichment algorithms use statistical approaches to assess if the selected gene set displays over-represented features (e.g. gene ontology terms, or clinical phenotype) in respect to random gene-feature associations, thus helping in discriminating disease genes from background[37].

2.1.2 Genomic data repositories

Even focusing only on a small subset of candidate genes, the amount of data to handle is almost huge, with the need of other parameters, by which refining the candidate list, is evident. An absolutely essential resource for variant interpretation consists of catalogues of genetic variants among human populations, where allele frequencies for each observed variant are reported[5]. Variants can be prioritized by allele frequency, discriminating from potential causative alterations, assumed to be rare (minor allele frequency < 1%), and common neutral SNVs. The Database for Short Genetic Variations (dbSNP) is the first public domain archive of genetic variations. dbSNP contains both known disease associated variants and non-pathogenic single nucleotide polymorphisms (SNPs), and aimed to facilitate large-scale association studies[38]. However, with the improvement of sequencing technologies, several databases have been developed by assembling genomes and/or exomes data, with the main goal to provide a deep catalogue of protein-coding variations for both population studies, and for the clinical interpretation of variants[5]. First attempts in this direction are the 1000 Genomes Project (1000GP), which contains variants resulting from WGS of 2,504 individuals[39], and the NHLBI Exome Sequencing Project that comprises exome data of approximately 6,500 patients with heart, lung and blood disorders, providing a catalogue of extremely rare protein-coding genomic alterations[40]. More recent efforts resulted in the Exome Aggregation Consortium (ExAC), comprising of 60,706 exomes from 6 broad populations and 14 disease cohorts, and the genome Aggregation Database (gnomAD), which catalogues genetic variants observed in 12,3136 WES and 15,496 WGS of unrelated individuals[41,42]. Another important source in candidate prioritization is represented by variant–phenotype databases, such as ClinVar[43], COSMIC[44] and the Human Gene Mutation Database (HGMD)[45]. These databases collect variant

known to be linked with different pathologies, ranging from rare Mendelian disorders, common diseases and cancer. The different lines of evidence for disease association are retrieved from multiple scientific publications, locus specific databases and clinical case reports[43,44,46]. In particular, ClinVar keeps a record of support for each variant interpretation submission, by which conflicts about the variant clinical significance are kept, allowing a better understanding of genetic variations in disease pathogenesis[43]. Moreover, disease inheritance pattern and patient's phenotype should be taken into account for variant interpretation, providing further evidence of variant causality[5]. The Online Mendelian Inheritance in Man (OMIM) repository collects information on all known Mendelian disorders and over 15,000 genes, focusing on relationships between genes and diseases[47]. Analogously, Human Phenotype Ontology (HPO) represents a standardized and hierarchical catalogue of phenotype descriptions (symptoms), for which the association to known disease genes is provided[48]. Both HPO and OMIM clinical definitions can be employed in variant/gene prioritization pipelines to link genomic alterations to patient's phenotype, or to discover new gene-disease associations[5].

2.1.3 Variant effect prediction

A common approach to cope with variant filtering consists in focusing only on rare non-synonymous variants (NSVs) to restrict the candidate list, allowing the identification of few potentially disease causative variants[5]. Indeed, variants in coding sequences have a clear potential for altering their phenotype, though only a small fraction of NSVs have a clear damaging functional effect on protein function[5]. Numerous computational strategies have been developed to prioritize variants on the basis of their functional/structural effects on proteins[5]. The theoretical basis for interpretation of coding SNVs rely on two main concepts: i) changes in protein sequence modify its functioning, and ii) different residues contribute to protein function to different levels[28]. Key residues with prominent functional role can be either inferred from evolutionary conservation resulting from the comparison with orthologous sequences, or identified directly inspecting the protein structure[49]. Evolutionary conservation among homologous proteins reflects the effects of negative selection against harming variants, with residue substitutions being tolerated only in regions free from functional constraints[50]. Thus, the first step of sequence conservation methods consists in building a multiple sequence alignment (MSA) of homologous sequences, allowing the identification of regions affected by amino acid replacement

constraints. The quality of MSA has wide effects on predictions, and restricting the analysis to orthologous proteins have been demonstrated to increase the method accuracy[28]. Replacement probabilities are empirically derived from MSA, measuring positional conservation according to probabilistic scoring functions. The multiple sequence alignments can be derived from PSI-BLAST, as in the case of SIFT[51], or constructed from specific protein sequence repositories, such as the non-redundant protein sequence database UniRef100 (Polyphen-2[52]) or SwissProt (Mutation Assessor[53])[54]. However, the scoring schemes of the presented tools rely on sequence homology and physic-chemical amino acid similarities, assuming that variations in orthologous alignment are functionally neutral[52,55]. A SNV could map to specific functional motifs, including active sites, and regions for cofactor binding, or elements required for interaction with other proteins[56]. In most of the cases this information can be retrieved from protein databases, such as Uniprot[54], and integrated in a machine learning method to improve missense prediction[28], as proposed for SNAP2[57]. In other cases, experimentally determined protein structure is available, allowing the assessment of effect variant on protein stability[28]. Despite causing free energy fluctuation of few kcal/mol, mutations can substantially affect structure stability, e.g. substituting hydrophobic residues with charged ones in the protein core[58]. Stability methods rely on free energy differences among folded and unfolded protein states ($\Delta\Delta G$). An amino acid substitution can shift protein conformational equilibrium either in favor of folded or unfolded states, and $\Delta\Delta G$ differences among wild-type and mutated protein reflects the variant impact on protein stability[58]. Most stability prediction tools rely on physic-base potential algorithms, which require the tridimensional structure to calculate the protein force field from the full atomic description (e.g. FoldX)[28,59]. The information can be integrated with available experimental and structural knowledge, atomic modeling and fast side chain packing, to improve method predictive power[60,61]. As a consequence, these methods can be very compute-intensive and allow the analysis of few mutations at once. Other approaches for stability prediction, such as I-Mutant2.0[20] and MUpro[21], consists in machine learning, trained on known protein substitutions with experimentally determined free energy variations, allowing to run predictions directly on protein sequence[62,63]. Despite similar performances, variant effect predictions obtained from distinct tools significantly differ one from each other[5,28]. In the last years, several meta-predictions method, based on machine learning approaches, have been proposed, integrating different data sources such as scores schemes from multiple tools for variant impact assessment[28]. The first implemented meta-predictor was PON-P, which combines conservation with structural

stability and machine learning methods to statistically derive the final prediction[64]. Numerous other tools were created following this integrative approach due to the high performance on benchmark sets, including computational methods for non-coding variant interpretation[5,28]. These prediction programs range from tools of DNA conservation assessment, e.g. GERP++[65], to unified methods combining experimental data and biochemical features for variant prioritization, including chromatin methylation state, such as FitCons[66], and presence of transcription factor target sites, e.g. CADD[67]. Another example is SnpEff[68], a multi-platform program that allows variant prioritization, including splice site variants. Moreover, SnpEff uses different types of annotations, such as ENSEMBL[69] functional site prediction, DNA methylation, histone modifications, chromatin accessibility and small RNA transcripts mapping from NIH Roadmap[70], and proteomic annotation from Nextpro database[71], to categorize candidate SNVs according their effect impact, favoring the identification of most significant variants[68].

2.2 Variant effects interpretation

Albeit being extremely informative, SNV filtering represents only the first step in proving the causality between variants and clinical phenotypes. For some variants, case-control studies, segregation, and prior literature based knowledge strongly support the direct association with the disease. More often, the SNVs map to novel disease associated genes identified by single case reports. These studies mainly rely on genetic and clinical tests, and rarely provide information about encoded protein. Bioinformatics analysis can address this issue, allowing to predict gene product function, thus, structural or functional effects of variants mapping to these genetic regions[13].

2.2.1 Inferring protein function from primary sequence

Intuitively, the first step for *in silico* protein characterization consists in retrieving the protein reference sequence of interest. The central protein sequence repository is the UniProt Knowledgebase (UniProtKB), which combines manually curated datasets (UniProtKB/Swiss-Prot) and proteins automatically derived from high throughput DNA sequencing experiments (UniProtKB/TrEMBL)[54]. UniProtKB provides different annotations, from protein function, to clinical phenotypes, and many links to other sources, facilitating the data mining process. Generally, protein functional regions are under

evolutionary selective constraint and are conserved among homologous proteins across species (orthologous). Consequently, the multiple alignment (MSA) of an unknown protein with its orthologous sequences is a powerful strategy to identify function-discriminating residues.

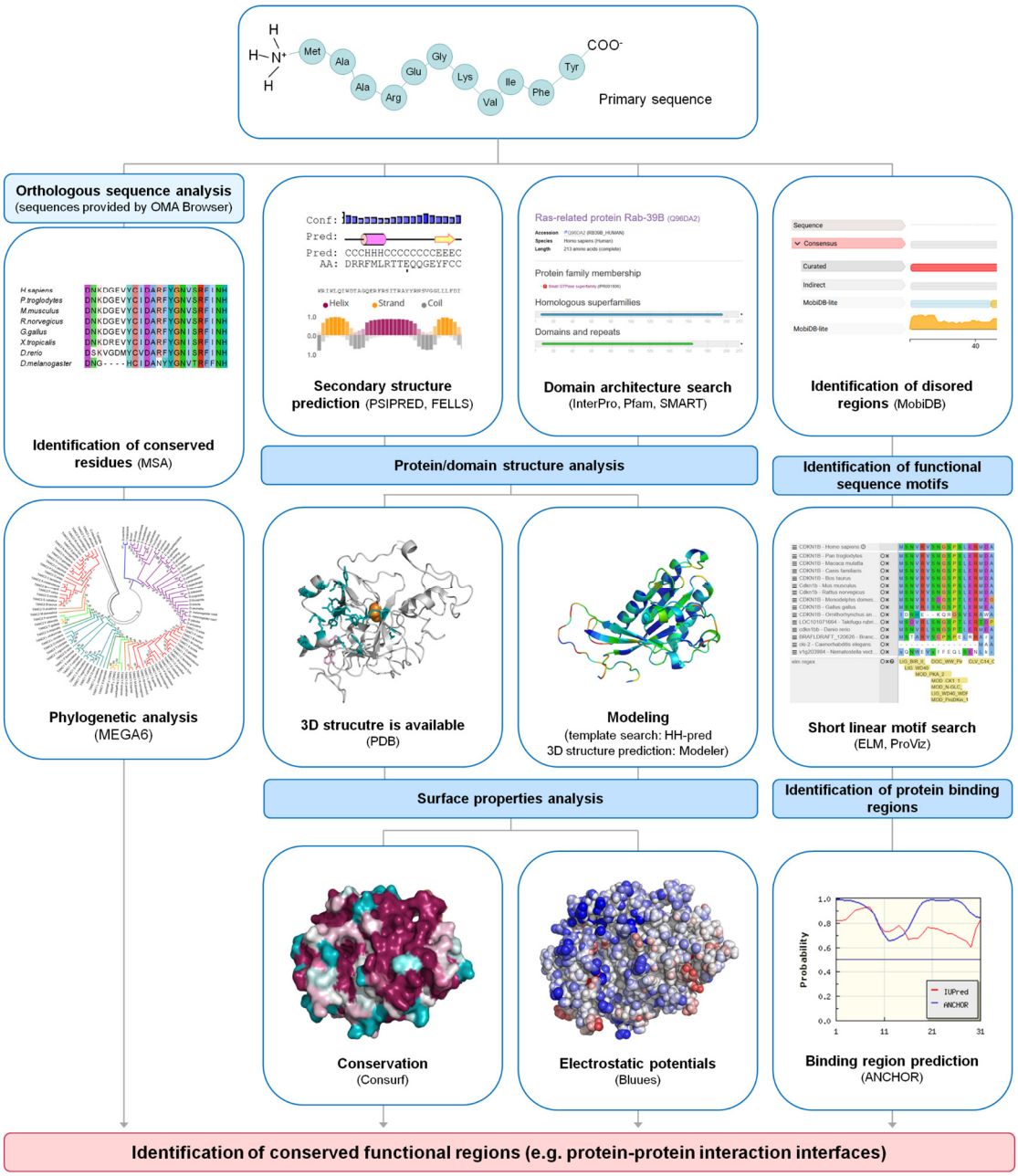


Figure 2.2: In silico protein analysis pipeline. Protein functional regions can be predicted by sequence and structure analysis.

Orthologous sequences can be obtained from specific orthologous protein database, such as the Orthologous MAtrix (OMA) browser[72], and used for the MSA. Although the MSA constructions can be performed by different algorithms depending on the considered

dataset[73], MSA manual inspection is always recommended. Indeed, alignment can be edited and refined according prior structural information, avoiding gaps in conserved structural elements. Manually curated MSA can be used to assess the evolutionary relationships within a protein family by phylogenic analysis. In spite of the variety of algorithms, sequence evolution process in small divergent group of sequences is usually evaluated by maximum likelihood approach, whereas the robustness of the phylogenetic tree reconstruction is generally assessed with bootstrap resampling[74]. Another useful application of phylogenic analysis is the inter-species protein annotation transfer, according to the concept that sequence similarity suggests structural and functional similarities[75]. However, protein structure, as well as the overall domain architecture, can be directly inferred from the single protein sequence. Several web servers, e.g. InterPro[76], provide functional analysis of proteins, integrating classification into protein families, structural domain annotation and functional sites predictions from different sources. Assessment of secondary structure elements represents the first step in characterizing of predicted protein domains. Among others, PSIPRED[77] is a machine learning based method, which uses PSI-BLAST profiles to calculate secondary structure propensity, and to classify residues according their alpha-helix (H), extended (E) or beta-strand, and coil (C) propensity. Besides structural domains, a large portion of both eukaryotic and prokaryotic proteins contains intrinsically disordered regions. IDRs are characterized by the lack of a fixed tridimensional structure, low hydrophobicity and a net composition bias for charged residues[78]. Intrinsically disorder is encoded in protein sequence and can be predicted from amino acid composition[79]. Among the available prediction methods, in-house MobiDB database[25] has been recognized as the primary repository of ID annotations, due to integration of the DisProt database manually curated information, IDR automatically derived from the Protein Data Bank (PDB) and the consensus computed from several different ID prediction tools[80]. Another in-house tool is FELL[78], a program for latent local structure prediction. It estimates secondary structure, intrinsic disorder or amphipathicity propensity from protein primary sequence. Despite unstructured, IDRs play an important role in cellular processes, mainly thank to the presence of functional interaction modules, known as short linear motifs (SLiMs)[81]. Indeed, SLiMs-mediated interactions have been proven to be involved in many molecular pathways, such as cell cycle regulation, proteosomal degradation, and protein complex structure stabilization[81]. Moreover, these short modules are target of post translation modifications (PTMs), allowing the context-dependent, integrative modulation of protein activity[81]. The ELM repository is the most

important database for the annotation and classification of linear motifs. The database also present a prediction tool module for identification of motif instances[81]. As short and highly degenerated, detected linear motifs are more likely false positive predictions[82]. Discriminating true matches among all the predicted short linear motifs is a complex task that should take in consideration: i) conservation in multiple sequence alignment, ii) presence of structural elements, iii) experimentally determined interactors[82]. All these data can be gathered and annotated on protein sequence by ProViz (Protein Vizualisation)[83]tool web-based visualization program. ProViz main advantage consists in providing a comprehensive graphical annotation of functional elements, e.g. domain structure, PTM, short linear motif, known sequence variant, on MSA related to query protein. Integrating both functional and evolutionary information with ANCHOR[67]protein binding site prediction, this tool can be used to predict most likely interactions regions in the query protein[83].

2.2.2 Identification of structural features relevant for protein function

Structural similarities correlate with function, thus characterizing the three-dimensional folding of a protein is a key step in understanding its role in cellular processes. The Protein Data Bank (PDB) represents the hub for collecting experimentally determined structure. PDB is divided in three main organizations, PDBe (UK), PDBj (Japan), and RCSB (USA), whose primary goal consists in maintaining a global and uniform repository of biological macromolecules. PDB files are lists of structure atomic coordinates, which are determined by X-ray crystallography, nuclear magnetic resonance (NMR), or cryo-electron microscopy (cryoEM). The file can contain either single proteins, or proteins forming complexes with other molecules, such as ligands, nucleic acids or other proteins. Unfortunately, PDB structures are not always available for the protein of interest, and protein fold should be determined by structure modeling. Homology modeling methods use protein sequence of interest for the identification of homologous protein with experimentally determined structures, which can be used as templates for three-dimensional model construction[84]. The prediction mainly relies on the principle that structures diverge with a lower rate compared to protein sequences, and homologous proteins can display similar folding even with divergent sequences[84,85]. Despite the higher conservation of protein tertiary structure, templates for homology modeling should share at least 30% sequence identity with query sequence, as most of protein pairs with lower sequence identity are unrelated[84]. The

template search can be performed using HHpred[86], which compares Hidden Markov Model profiles (HMMs) from alignment databases, such as Pfam[87] or SMART[88]. It returns a list of pairwise alignments between the query sequence and detected distant homologous protein structures, allowing to select the best one accordingly to sequence identity, or secondary structure prediction. Then, selected HHpred alignment can be forwarded to Modeller[89] for comparative modeling of the protein/domain of interest[86]. Model reliability is determinant to infer function, and thus, should be assessed. One of the most used quality model assessment program is the QMEAN web server[90]. QMEAN is a composite scoring function that calculates absolute quality estimates both for the entire model (global score) and for each residue in the structure (local score)[90]. Moreover, model quality can be directly explored by structure visualization. Pymol is an open-source molecular visualization software available for different operating systems[91]. Its widespread diffusion is due to the user-friendly interface, and the possibility to easily inspect structures in details, or to integrate results from other tools, such as Consurf[92] and Bluues[93]. ConSurf server identifies conserved amino acids from multiple sequence alignment and maps them on protein structures[92]. Residues on structure are assigned with different colors according to the evolutionary conservation rate, ranging from magenta, i.e. most conserved residues, to cyan for variable amino acids[92]. Generally buried residues are important for protein folding and, thus, are more conserved than the exposed ones. However, conserved residues on the surface may suggest the presence of functional sites, e.g. binding site[94]. Together with conservation, highly charged surfaces can indicate nucleic acid binding sites, or catalytic regions in proteins[94]. Hence, electrostatic potential analysis can be used to infer functional sites. The Bluues web server computes generalized born radii and surface potential either from PDB structure, or from user defined model. The software uses this information to build a PQR file, which can be visualized with Pymol[93]. The electrostatic potentials are represented on the solvent accessible surface, with blue indicating negative amino acids and red positive charged residues[93].

2.3 Variant effects on protein interactions

Protein-protein interactions (PPIs) are essential for the proper functioning of the molecular mechanisms underlying cellular homeostasis, and events affecting PPIs play a major causal role in disease pathogenesis[18,95]. In medical research, the disease PPI network analysis represents a useful strategy for the identification of new therapeutic targets and treatment-

responsive biomarkers for diagnostic and prognostic purposes[95]. Besides the clinical applicability of PPI studies, PPI information are used to predict the function of a uncharacterized protein on the evidence of its interactions with other proteins, whose function is already established[96]. Due to recent progresses in proteomics, experimentally determined PPIs have been impressively increased, mainly thanks to the application of high-throughput techniques, e.g. yeast two-hybrid (Y2H) screening and mass spectrometry[96]. To organize and process the extensive production of PPI data, several databases have been developed. BioGRID[36] (biological general repository for interaction datasets) contains protein-protein, genetic and chemical interactions, and post-translational modifications from thirteen different species. The entries are retrieved by an extensive curation of data from the literature and large-scale experiments. Another PPI database is IntAct[35], which provides both textual and graphical representations of interactions. Other sources, such as STRING[20], rely on integrated networks of PPIs, which assign probabilistic scores to each interaction based on a variety of sources, such as text mining and experimental evidence, to identify all possible functional associations. Indeed, PPIs identified by high-throughput methods generally suffer from high false positive rate, e.g. unspecific interactions, and they should be considered as a starting point for further experimental validations[95]. Moreover, the specific interaction interfaces are usually not detected, and require further analysis to be determined. Thus, computational methods are often employed both to reduce large lists of potential binding partners and for the identification of the putative binding regions[96]. Interactions occur both among different classes of domains, and between domains and short linear motifs [97,98]. In both cases, the PPI specificity is linked to the complementarities between interacting regions, and typically depends on few highly conserved residues. These residues encode the affinity and specificity of the binding, and more variable surfaces that selectively contribute to the interaction[97,98]. This means that a domain is expected to interact with a specific counterpart module, either a folded domain type or a peptide motif class. Thus, information about interactions can be transferred from a functionally defined protein to the target one based on similarity[96], as interfaces are generally conserved among homologous proteins, but also considering interactions complementarities. E.g., a protein contains a conserved SRC Homology domain 3 (SH3), which specifically recognizes the poly-proline motif PxxP[98]. Hence, possible interactors should display an exposed conserved PxxP motif, mapping in an accessible region of the protein, such as an IDR. These considerations can be used to filter putative interactors retrieved from large scale experiment data or predicted by

computational tools (e.g. ELM), allowing to focusing only on the most promising binding partners. After these preliminary filters, it is clearly important to experimentally validate selected PPIs by additional low-scale experimental techniques to exclude artifacts[99]. One strategy consists in evaluating the cellular colocalization of the proteins involved in the interaction. Colocalization analysis is based on the hypothesis that interacting molecules should be placed in the same physical location, and should display an overlapped intracellular distribution[99]. The colocalization experiments can be performed either with transfected cells, where target macromolecules are over-expressed, or on endogenous proteins. In both cases, target proteins are labeled with antibodies conjugated with fluorophores. Fluorophores are chosen so that they emit at different wavelengths, allowing to monitor their distribution within the cells. When the proteins colocalize, fluorophore signals are overlapped. Images are recorded by confocal microscopy, and statistically analyzed to assess correlation among protein distributions[100]. Despite supporting the interaction, colocalization does not discriminate between direct binding and PPIs mediated by common interaction partner. Thus, a common practice consists in combining colocalization analysis with other experimental tests for direct-binding assessment, such as the co-immunoprecipitation (co-IP). In co-IP experiments, the antibodies are used to capture proteins in native form, bound to other interacting macromolecules. The main advantage of this approach is that physiological endogenous protein complexes are studied. However, co-IP cannot recognize the specific regions involved in protein binding[99]. Yeast two hybrid system can be used to address this issue. Y2H is one of the most used methods to screen and confirm PPIs, as it allows the direct identification of PPI between protein pairs. The Y2H system takes advantage of the modularity of Gal4 transcription factor, which is formed by two modules: i) DNA binding domain (DBD), and ii) an activation domain (AD). Both domains are required for the transcription of a reporter gene, e.g. *HIS3* gene, which enables the yeast to grow on a selective medium[96,99]. Moreover, Y2H can be used to examine sub-regions of proteins, by which uncover interactions not revealed by full-length proteins and validate interaction minimal regions; e.g. validate linear motif mediated interactions[101].

3 Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges

This Chapter has been published in "R. Daneshjou, Y. Wang, Y. Bromberg, S. Bovo, P. Martelli, G. Babbi, P. Di Lena, R. Casadio, M. D. Edwards, D. K. Gifford, D. T. Jones, L. Sundaram, R. Bhat, X. Li, L. R. Pal, K. Kundu, Y. Yin, J. Moult, Y. Jiang, V. Pejaver, K. A. Pagel, B. Li, S. Mooney, P. Radivojac, S. Shah, M. Carraro, A. Gasparini, E. Leonardi, M. Giollo, C. Ferrari, S.C.E. Tosatto, E. Bachar, J. Azaria, Y. Ofran, R. Unger, A. Niroula, M. Vihinen, B. Chang, M. H. Wang, A. Franke, B. Petersen, M. Pirooznia, P. Zandi, R. McCombie, J. B. Potash, R. B. Altman, T. Klein, R. Hoskins, S. Repo, S. E. Brenner, A. Morgan. Hum Mutat. 2017 Jun 21." For Supplementary Materials, check the online version of the paper.

3.1 Summary

Precision medicine aims to predict a patient's disease risk and best therapeutic options by using that individual's genetic sequencing data. The Critical Assessment of Genome Interpretation (CAGI) is a community experiment consisting of genotype-phenotype prediction challenges; participants build models, undergo assessment, and share key findings. For CAGI 4, three challenges involved using exome sequencing data: bipolar disorder, Crohn's disease, and warfarin dosing. Previous CAGI challenges included prior versions of the Crohn's disease challenge. Here, we discuss the range of techniques used for phenotype prediction and discuss the methods used for assessing predictive models. Additionally, we outline some of the difficulties associated with making predictions and evaluating them. The lessons learned from the exome challenges can be applied to both research and clinical efforts to improve phenotype prediction from genotype. In addition, these challenges serve as a vehicle for sharing clinical and research exome data in a secure manner with scientists who have a broad range of expertise, contributing to a collaborative effort to advance our understanding of genotype-phenotype relationships.

3.2 Introduction

Precision medicine aims to use a patient's genomic and clinical data to make predictions about medically relevant phenotypes such as disease risk or drug efficacy [102,103]. The Critical Assessment of Genome Interpretation (CAGI) is a community experiment, which aims to advance methods for phenotype prediction from genotypes through a series of "challenges" with real data (CAGI, 2011). Exome sequencing data, which captures exons and nearby flanking regulatory regions, is already being used clinically to solve medical mysteries with well-defined symptoms [104]. However, in order to advance precision medicine, clinicians and scientists will need to be able to make inferences about disease risk or drug efficacy from genetic data. Interpretation of genetic data is one of the major difficulties in the implementation of precision medicine [105].

CAGI is an example of the Common Task Framework, a phrase coined by Mark Liberman to describe the approach of using shared training and testing datasets and evaluation metrics to advance machine learning [106,107]. The Common Task Framework has been called the 'secret sauce' behind the recent successes in machine learning [107]. Starting with common task challenges in the 1980's for machine translation, this approach has led to significant gains in speech recognition and dialog systems, protein structure prediction, biomedical natural language processing, autonomous vehicles, and collaborative filtering for consumer preferences [108–112]. Through this same approach, CAGI aims to push forward the field of precision medicine.

At CAGI 4 held in 2016, three challenges involved making predictions using exome sequence data: A Crohn's disease challenge, a bipolar disorder, and a warfarin dosing challenge. These challenges represent the spectrum of phenotypes seen in clinical practice. Bipolar disorder and Crohn's disease are discrete phenotypes, with the former being a clinical diagnosis (based on meeting clinical criteria) and the latter a pathological diagnosis (based on biopsies). Therapeutic warfarin dose, on the other hand, is a continuous phenotype. The Crohn's disease challenge has been a part of previous CAGI iterations, while the warfarin dosing and bipolar disorder challenges debuted during CAGI 4. We will describe the nature of each challenge in greater detail. The number of groups participating in each challenge can be found in Table 3.1.

Table 3.1 The number of predictors and predictions for each CAGI challenge.

Challenge	Number of predictors	Number of Predictions
Crohn's Disease Exomes Challenge	CAGI 2 – 10 groups CAGI 3 – 14 groups CAGI 4 – 14 groups	CAGI 2 – 33 predictions CAGI 3 – 58 (+3 late) predictions CAGI 4 – 46 predictions
Bipolar Exomes Challenge	CAGI 4 – 9 groups	CAGI 4 – 29 predictions
Warfarin Exomes Challenge	CAGI 4 – 3 groups	CAGI 4– 9 predictions

3.2.1 Crohn's Disease Challenge

Crohn's disease is a chronic inflammatory bowel disease marked by transmural inflammation of the gastrointestinal tract that can occur anywhere from the mouth to the rectum [113]. Symptoms include pain and debilitating diarrhea, which can lead to malnutrition [113].

Monozygotic twin studies have shown a concordance of 40-50%, and genome wide association studies have identified genetic risk loci [113,114]. Age of onset is typically between 20-40 years old, but early age of onset, such as in early childhood is associated with more severe disease features [115].

The 2011 (CAGI 2) dataset has 56 exomes (42 cases, 14 controls), all of German ancestry [116]. The 2013 (CAGI 3) dataset has 66 exomes (51 cases, 15 controls). Though these samples were also of German ancestry; cases were selected from pedigrees of German families with multiple occurrences of Crohn's disease. As such, some of these cases were related. For the most part, the samples sequenced as controls were unrelated healthy individuals; the exceptions to this were the unaffected parents of three cases and the unaffected twin of one case. The most recent challenge, CAGI 4 in 2016, was to identify cases from controls in 111 unrelated German ancestry exomes (64 cases, 47 controls). For CAGI 4, submitting groups were allowed to use the data from the Crohn's disease CAGI challenges of 2011 and 2013. In all iterations of the challenge, groups were asked to report a probability of Crohn's disease (between 0 to 1) for each individual and a standard deviation representing their confidence in that prediction. For the most recent Crohn's disease evaluation, teams were also asked to predict if age of onset was greater or less than age 10; an age cutoff selected by CAGI based on the literature [115]. Additional details of the CAGI 4 challenge can be found under Supplementary Exhibit 3.1.

3.2.2 Bipolar Disorder Challenge

Bipolar disorder is a mood disorder marked by elevated mood (mania or hypomania) and depressed mood that disrupts an individual's ability to function[117]. In the general population, the lifetime risk of bipolar disorder is 0.5-1% [118]. However, bipolar disorder has a high component of heritability, with studies demonstrating a 40-70% monozygotic twin concordance [118]. In this CAGI 4 challenge, 1000 exomes of unrelated bipolar disorder cases and age/ancestry-matched controls of Northern European ancestry were provided. 500 exomes were used as the training set and 500 exomes were for the prediction set [119]. Groups were asked to report a probability of bipolar disorder (between 0 to 1) for each individual and a standard deviation representing their confidence in that prediction. Additional information on the challenge can be found under Supplementary Exhibit 3.2.

3.2.3 Warfarin Dosing Challenge

Warfarin is an anticoagulant with over 30 million prescriptions written in 2011 (IMS, 2012). Warfarin remains a clinical staple despite the introduction of novel oral anticoagulants because of multiple factors – warfarin's lower cost, longer half-life, and clinical indications for which novel oral anticoagulants have not yet been approved [120]. However, warfarin is responsible for one third of hospitalizations due to adverse drug events because of its narrow therapeutic index and high inter-individual dose variability [121]. Both clinical and genetic factors affect the therapeutic dose of warfarin [122]. For this challenge, participants were provided with exomes of African Americans on tail ends of the warfarin dose distribution (≤ 35 mg or ≥ 49 mg) [123]. Clinical covariates were provided for all exomes. The training set consisted of 50 exomes, and participants submitted dose predictions with standard deviations on 53 test set exomes. Additional details of the challenge can be found under Supplementary Exhibit 3.3.

3.3 Methods

3.3.1 Data Distribution

Data was distributed to the participants who consented to the CAGI data use agreement. Data providers worked with their home institution to ensure adherence with local privacy regulations and predicting groups agreed not to share the anonymized data. Data was provided as described above, with genetic variant data shared in the VCF file format.

3.3.2 Predicting Phenotypes

Predicting groups were required to return a simple text file with appropriate predicted values (such as disease status and confidence in prediction) for each sample. They were also provided with a validation script to check their output formatting. Submitting groups were asked to submit a methods description for each submission. The prediction results from selected groups that submitted predictions and methods descriptions were presented at the CAGI meeting. Additionally, the ground truth data and scoring scripts used to perform the evaluation were shared with participants.

3.3.3 Data Quality

For the Crohn's disease and bipolar disorder exome challenges, biases in the data were assessed using principal component analysis and clustering after pruning for linkage disequilibrium using plink [124]. For the warfarin challenge, data had previously undergone QC using ancestry informative markers to confirm self-reported ancestry and identity by State (IBS) analysis in order to ensure that samples were not related, as previously described [123].

3.3.4 Assessing Discrete Phenotypes (Crohn's Disease and Bipolar Disorder)

A simple accuracy of prediction per sample score, such as derivable from setting a threshold for prediction (such as 0.5), although tantalizing in its simplicity neither supports the goals of CAGI nor is it representative of a likely clinically relevant scenario for prediction. Because the genetic datasets from CAGI are drawn from case-control studies, as well as pedigree studies in families with a strong burden of disease, it does not represent a random sampling of the population.

Requiring a fixed threshold for evaluation and reporting a basic accuracy score of prediction in such a dataset would obscure interpretation. Also, using this as a Figure of merit for ranking encourages participants to optimize their system predictions for the anticipated case/control distribution instead of focusing on features that selectively prioritize and rank disease likelihood in the absence of that calibration. The use of Receiver Operator Characteristics (ROC) curves for genomic test evaluation has been previously investigated by Wray et, al [125]. The ROC offers many advantages for evaluating a test, and is often used to characterize clinical tests. The shape of a ROC curve can help differentiate between highly sensitive tests, which could rule in a possible diagnosis, and highly specific

tests that could rule out a diagnosis. The prediction of Crohn's disease status from sequencing data might be used in either of those situations depending on clinical presentation, risk factors, or stage of patient evaluation. Additionally, ROC curves allow easy selection of a classification threshold (based on selecting a position on the curve). Based on the selected threshold, a positive or negative likelihood ratio can be derived and applied in standard evidence based techniques of patient diagnosis, which rely on a Bayesian framework that takes into account the pre-test probabilities and the characteristics of a given test depending on the threshold chosen for prediction [126]. Additionally, we evaluated the robustness of the prediction accuracy when making predictions on different subsamples of exomes and assessed the confidence intervals reported by the participants. To capture confidence intervals on the predictions, multiple samples with replacement were drawn. Each prediction was then modified by adding a random amount drawn from a normal distribution with a mean of zero and a standard deviation equivalent to the standard deviation reported for the original prediction. If no confidence interval was reported for the original prediction, the standard deviation was taken to be zero. If a prediction for a particular exome was missing, the prediction score for that sample was set to the mean reported prediction value in that submission. In order to compare submissions by a single Figure of merit, the average area under the ROC curves from the bootstrap sampling was used, accompanied by the bootstrapped confidence interval around that area under the curve, to estimate the robustness of differences between prediction performances. The evaluation scripts were provided to all participants. A cross-validated logistic regression based meta-classifier using lasso regularization was also trained on the submissions as features for CAGI 4 Crohn's disease and CAGI 4 bipolar disorder. This step allowed us to assess whether combining the features selected across the different groups would improve prediction over a single method. The meta-classifier could perform better than any single method if the different methods use significantly different predictive features.

3.3.5 Assessing Continuous Phenotypes (Therapeutic Warfarin Dose)

For the warfarin exomes challenge, several metrics of assessment were used. Each participant provided a predicted therapeutic dose of warfarin for each individual as well as a standard deviation for that prediction. To look at the amount of variation in dose explained by the predicted doses, we used linear regression with the linear model function (lm) in the R statistical package (v 2.15.3). We evaluated each method using the R2 and the sum of

squared errors. Additionally, we compared each prediction against one of the best performing warfarin predictive algorithms, the International Warfarin Pharmacogenetic Consortium (IWPC) algorithm [122]. To assess, on average, how many participant-provided standard deviations the predicted dose was from the actual dose, we used a mean of the absolute value of the z-score for each prediction, as seen in equation 1. Here, dose_actual is the known therapeutic dose of warfarin for each individual *i*, while dose_predicted is the therapeutic dose predicted by that group for that individual. SD_predicted is the standard deviation for each individual's predicted dose, as provided by the participant's prediction method. The number of individuals is n.

Equation 1:

$$\frac{\sum_{i=1}^{n} \left| \frac{dose\_actual_i - dose\_predicted_i}{SD\_predicted_i} \right|}{n}$$

To assess the range of each prediction's standard deviation compared to the predicted dose, we calculated the mean of the coefficient of variation, which was the mean of the standard deviation for each prediction divided by the predicted dose, as seen in equation 2. Equation 2:

$$\frac{\sum_{i=1}^{n} \frac{SD\_predicted_i}{dose\_predicted_i}}{n}$$

We also evaluated the mean absolute value of the z-score multiplied by the mean coefficient of variation for each method. This value allowed us to assess the mean z-scores with a penalization for mean z-scores whose values were closer to 0 because of larger standard deviations. Additionally, we calculated rho and p-values using the spearman rank correlation between 1) each group's predicted warfarin doses and the actual therapeutic doses across individuals and 2) each group's predicted warfarin doses and the IWPC predicted doses across individuals. These calculations were made with the spearmanr command from the stat package in scipy (python v 2.7.5).

3.4 Results

With each year, CAGI has expanded the number of challenges and participants. Table 1 displays the number of participants and predictions for each CAGI challenge.

3.4.1 Crohn's Disease Exomes Challenge (CAGI 2-4)

For the 2011 Crohn's disease (CAGI 2) challenge, during the assessment phase, a substantial batch effect was discovered in the data as a side effect of sample preparation and sequencing (Figure 3.1).

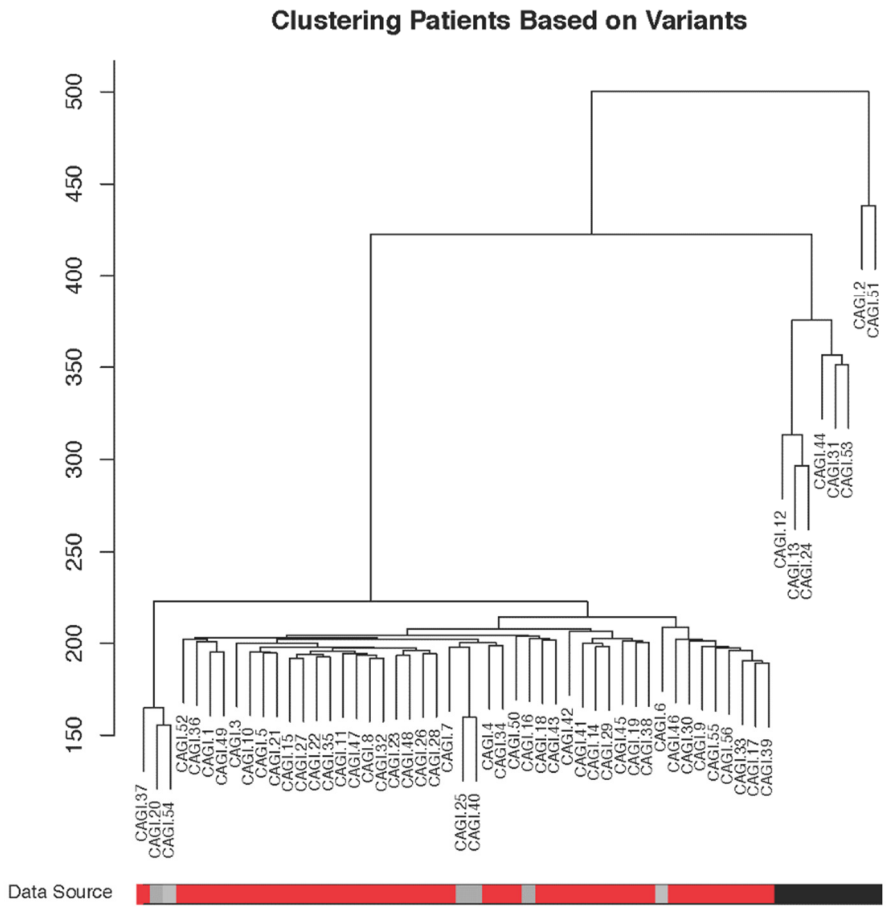


Figure 3.1: Clustering of patients from the CAGI 2 Crohn's Disease Challenge. The black and gray bars at the bottom represent the controls; the red represents the cases. Many of the controls cluster together, likely due to batch effects. For instance, the controls represented in black were sequenced separately from the gray controls and the cases.

Overall, the control samples that clustered separately due to this batch effect had overall fewer variants reported that did not match the reference genome. The participants were not

aware of this batch effect; their methods were not designed to exploit it. However, this raises the possibility that techniques that used a very large list of genes were more likely to correctly identify case samples as coming from individuals with Crohn's disease. Indeed, many different methods did better than random based on AUC, with a maximum AUC of 0.94, and in general approaches that favored a large list of potentially Crohn's disease related genes and gave more weight to rarer variants did the best. A full description of all methods used by the participants can be found in the supplement under Exhibit 3.1: CAGI 2. Supplemental File 3.1 shows comparative results of the CAGI 2 Crohn's disease challenge predictive methods. It is certainly biologically plausible that increased burden of variation in a large number of Crohn's disease related genes leads to increased likelihood of disease; however, it is also possible that there was systematic over reporting of variation as a batch effect. Therefore, it was important to re-evaluate with more data. In the 2013 CAGI 3, a much greater effort was made to carefully collect and prepare samples in a completely consistent way. In this case, case samples were collected from German families with a particularly high burden of Crohn's disease (two or more effected family members), including a pair of twins discordant for disease, and another pair of twins concordant with disease. Additional healthy controls were drawn from the unaffected German general population. During the 2013 CAGI 3, there was once again a substantial difference in clustering between cases and controls, but in this dataset there was substantially more homogeneity in the cases. Individuals from different case families clustered much more closely with other high Crohn's burden family individuals (Figure 3.2). This prompted two possible hypotheses. The first is that there might be a hidden founder effect and that these families with a high burden of disease may all actually be closely related. The second is that reduced heterogeneity and perhaps increased ancestor consanguinity may contribute to increased risk of Crohn's disease in these families with a high burden. Either one alone or a mixture of both possibilities is biologically plausible. In this instantiation of CAGI, groups that simply did some version of partitioning the test datasets based on hierarchical clustering did quite well, and the top performing methods had an AUC of 0.87. Once again, all of these methods were implemented without awareness of the bias in the data. A full description of all methods used by the participants can be found in the supplement under Exhibit 3.1: CAGI 3. Supplemental File 3.2 shows comparative results of the CAGI 3 Crohn's disease challenge.

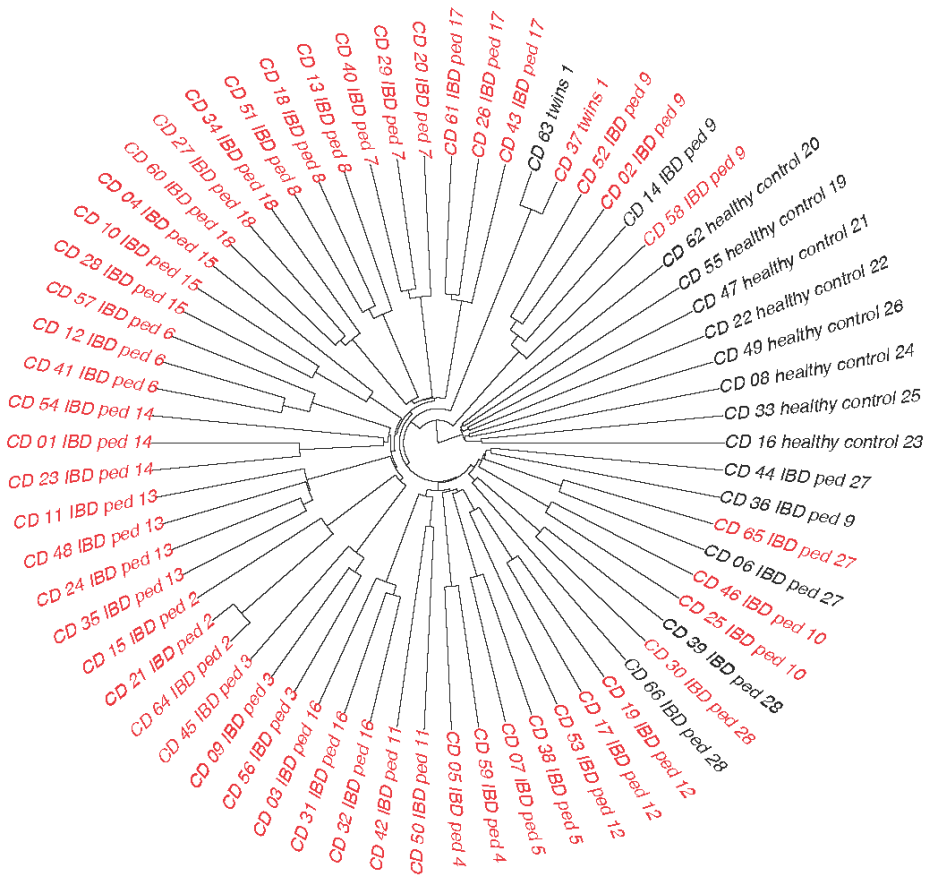


Figure 3.2: Clustering of samples for CAGI 3 Crohn's Disease challenge.
Black represents controls, while red represents cases. This dataset included healthy family members of cases as well as random controls. Samples with a "ped" designation in the sample name came from a pedigree; samples that share the same "ped" number came from the same pedigree.

In CAGI 4, the 111 exomes were derived from a mix of 64 Crohn's disease patients, with a skew toward early onset of disease, and 47 healthy controls, all taken from individuals of German descent. With this data, the simple separation of cases and controls based on genetic variants was not present (Figure 3.3), suggesting the problems with batch effects and sampling bias were no longer present; the only noticeable structure indicated the possibility of a few related samples, as seen in the PCA and IBD plots shown in Supplementary Figure S1 and Supplementary Figure S2. Correspondingly, the peak performance dropped from previous CAGI iterations down to an AUC of 0.72. However, given the elimination of biases in the data, this incarnation of the Crohn's disease challenge is likely the best reflection of how the prediction methods perform. A meta-classifier created by the assessment team using all submitted methods for this challenge, as shown in Supplementary Figure S3.3, had an AUC of 0.78, a small improvement over the top

method. The distribution of AUCs across methods is shown in Figure 3.4. A full description of all methods used by the participants can be found in the supplement Exhibit 3.1: CAGI 4. Supplemental File 3.3 shows comparative results of the CAGI 4 Crohn's disease challenge.

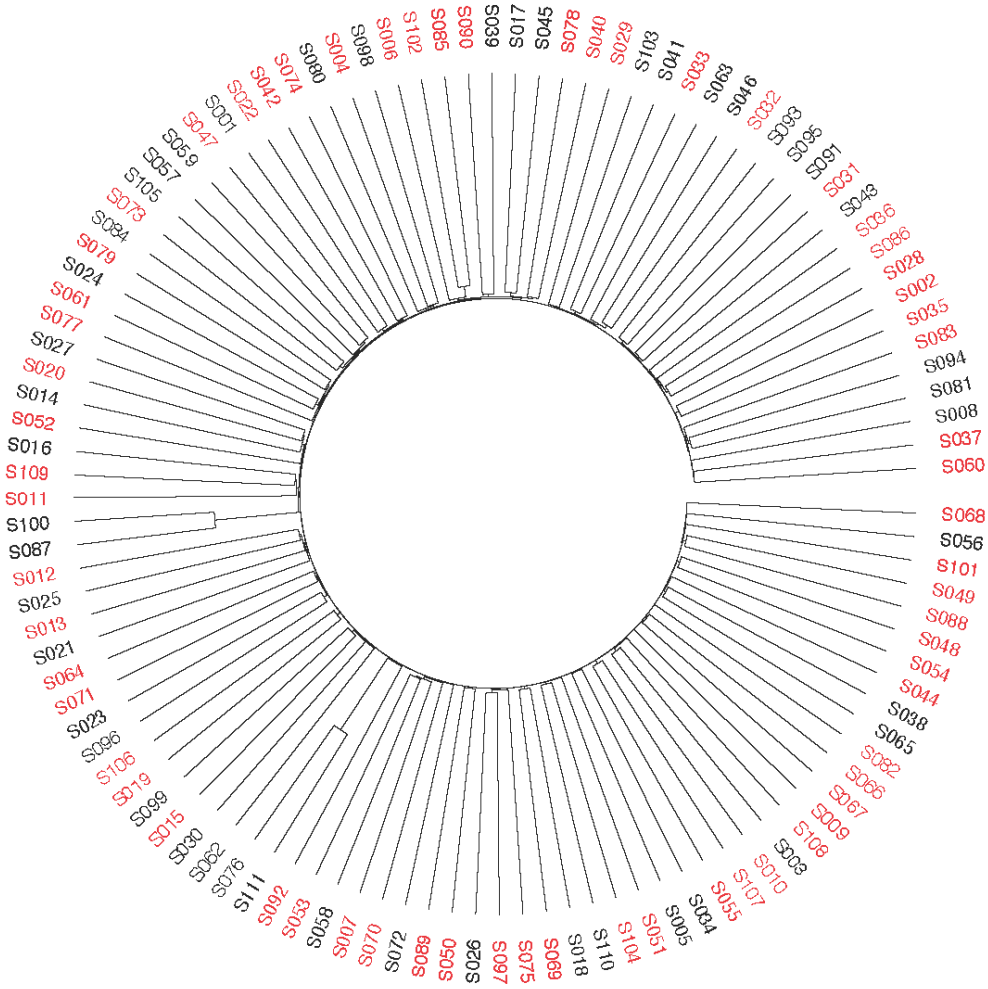


Figure 3.3: Clustering of samples for CAGI 3 Crohn's Disease challenge

The top approach in CAGI 4 used a compiled list of genes and genomic regions associated with Crohn's disease from prior studies, used imputation to evaluate risk contribution from known regions associated with Crohn's disease but not covered by exome sequencing, and used the Welcome Trust Case Control Consortium (WTCCC) Crohn's disease genotyping array data to train a disease classifier to score relative risk for each sample. Across participants, numerous methods were used for selecting the covariates, highlighting the many different approaches to building a Crohn's disease classifier. Similar to the top

approach, many groups used variants previously found to be associated in genome-wide association studies; the NHGRI catalog was a popular choice to identify these associated variants (Welter, et al., 2014). Other approaches relied on gene lists of associated and "predicted" Crohn's disease genes to select variants of interest. To create the "predicted" list of Crohn's disease genes, groups used a variety of methods. Examples include using (1) existing tools such as Phenolyzer, which associates disease terms with genes based on prior research, expands the gene list by using gene-gene relationships, and then creates a ranked list of candidate genes and (2) creating gene lists based on GO pathways enriched with Crohn disease associated variants (3) using natural language processing to identify genes of interest from Pubmed abstracts[127,128]. From a gene level, different groups would then devise different strategies to select variants of interest. For some approaches, population level frequency data was used to help distinguish variants more likely to be pathogenic. Other methods relied on pathogenicity prediction tools such as SNAP, PON-P2, SNPs&GO, and Variant Effect Predictor to inform variant selection and weighting[129–132]. A range of machine learning approaches were used to actually build the classifiers-naïve Bayes, logistic regression, neural nets, and random forests. Additionally, some groups improved on prior iterations by creating meta-classifiers based on combinations of prior methods.

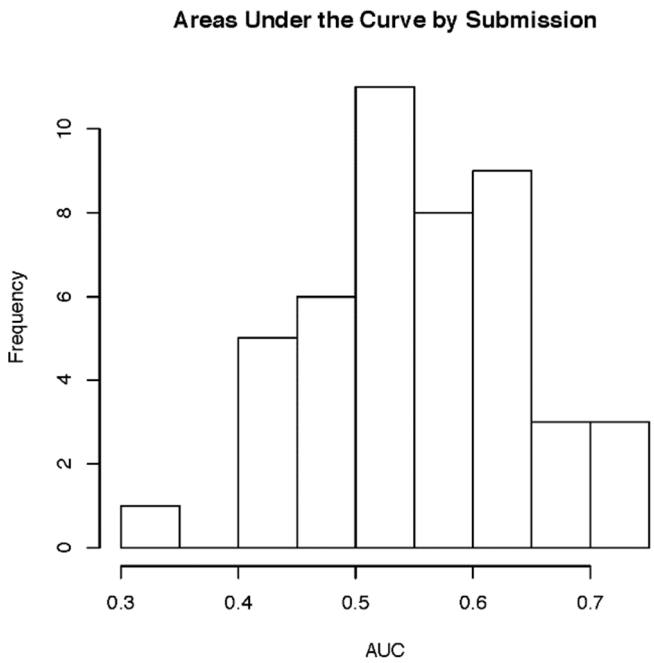


Figure 3.4: CAGI 4 Crohn's disease challenge distribution of AUCs across all methods.

3.4.2 Bipolar Disorder Exomes Challenge (CAGI 4)

As noted, a substantial difference between the Crohn's disease phenotypic prediction challenge and the bipolar disorder challenge, was that a substantial amount of training data was provided for the bipolar disorder challenge, with 500 of the 1000 exomes randomly selected and provided as training data for the challenge. These samples were unrelated, and analysis steps assessing the relationships between samples can be found in Supplementary Figures S3.4, S3.5, and S3.6. The top performing group had a method with an AUC of 0.64. The distribution of AUCs across methods is shown in Figure 3.5. Although many groups used approaches similar to those used for the Crohn's disease challenge, the top performing group (which did not apply this method to Crohn's disease data), treated the genotype data as linear features and trained a neural network with 3 hidden layers, with the middle layers looking at local features in the linear space of the ordered SNPs of the VCF file, tuning for performance using cross validation on the test data. Importantly, this approach used essentially no prior knowledge of genetics or the results of prior studies on disease-gene relationships. Supplemental File 3.4 shows comparative results of the CAGI 4 bipolar disorder challenge. Overall descriptions of prediction methods are available under Exhibit 3.2: CAGI 4. A meta-classifier created by the assessment team using all submitted methods for this challenge, as shown in Supplementary Figure S3.7, had an AUC of 0.64, which was not significantly different from the top method.

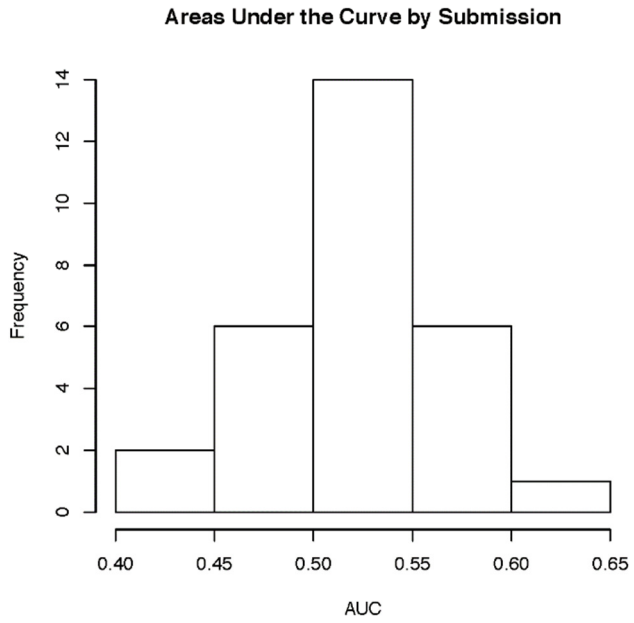


Figure 3.5: CAGI 4 bipolar disorder challenge distribution of AUCs across all methods

3.4.3 Warfarin Exomes Challenge (CAGI 4)

With the warfarin exomes challenge, similar to the Crohn's disease challenge, many groups used a priori data to create a list of covariates used. This included known pharmacokinetic and pharmacodynamics warfarin genes, genes mentioned in the literature, and also using tools to find functional neighbors of the known gene set. One prediction method (Group 50, Prediction 1) was ahead of the others when looking across multiple performance metrics described in the methods section - R2, mean absolute value of z-score, and mean absolute value of z score multiplied by the coefficient of variation (Figures 3.6A-D, Supplementary Table S3.1). The R2 of the top prediction method was 0.25, compared to 0.35 for the IWPC prediction method, one of the best performing predictive algorithms. A visualization of the predictions compared to the actual dose can be seen in Supplementary Figures S3.8 and S3.9. Details of all methods can be found under Supplementary Exhibit 3.3: CAGI 4. The methods submitted for this challenge had several similar features. Every method submitted took advantage of the fact that the range of doses were published in the paper from which the data came. Thus, these methods either fit rankings to the dose range or set doses above or below the known range to the lower or upper limits. Additionally, most methods used prior information from the literature to help set the initial clinical and genetic covariates to consider in their models.

3.5 Discussion

The CAGI exome challenges revealed lessons specific to each particular challenge as well as generalizable principles for future genotype-phenotype prediction challenges.

3.5.1 Crohn's Disease

Overall, there were substantial challenges with bias and population stratification in the datasets that make evaluation and comparison of techniques for identifying Crohn's Disease status from exome data difficult. In the latest crop of prediction systems, it may be that techniques such as using imputation to infer variants in regions not covered by the exome sequencing and large external microarray SNP chip datasets are key factors in superior performance. The top AUC varied across the three evaluations, demonstrating the substantial differences in the data sets. Groups who created meta-classifiers based on combining previous methods from previous CAGI challenges demonstrated the value of

applying the Common Task Framework to genetic problems – through iteratively improving their methods based on prior learning. Importantly, across the three CAGI evaluations, the average system performance performed better than random, including in the most recent, CAGI 4, implying that there is some level of useful information in predicting likelihood of Crohn's disease from exome data in the population, something previously not demonstrated.

3.5.2 Bipolar Disorder

Surprisingly, the group that created the best performing prediction in the Bipolar disorder challenge acknowledged having little background in biomedicine or genetics. This group approached the problem as purely a data classification challenge. On the one hand this may be hailed another example of the unreasonable effectiveness of data and the success of machine learning over human expertise; the quotation "Every time I fire a linguist, the performance of our speech recognition system goes up," has been attributed to Fred Jelinek in the 1980's, and something similar may be afoot in genomics, promising an exciting future as datasets expand and machine learning techniques improve. However, one of the major challenges is that prediction accuracy with case-control data does not really reflect most applications we can envision for a phenotypic prediction system. Moreover, while not detected by any of our quality control methods, it is still possible that the top performing method picked up on hidden population stratification/biases in the data. Although we were unable to find evidence of this, a sophisticated machine learning system may be identifying features which partition the cases and controls but which are not related to biological drivers of disease risk. Unfortunately, the tools to dissect the deep neural net architecture in the context of genomic features are currently too primitive to help us deepen our biological understanding using these results. There has been recent work into advanced techniques to understand the decisions made by previous black box systems in areas like image processing and natural language processing; however, similar tools for understanding genomic prediction systems are less developed [133].

3.5.3 Warfarin

Predicting warfarin dose using clinical information and genetics is a difficult problem; one of the best performing algorithms (IWPC) has an R2 of 0.35 on this data set. Existing algorithms have poorer performance on diverse populations since most algorithms are

trained on European descent [122,123]. For this challenge, the winning method had an R2 of 0.25.

The warfarin exomes challenge had several limitations. The sample size was limited, with only 50 samples for training and 53 for testing. This data was generated at a time when exome sequencing was more expensive; falling costs may allow an expansion of available exome data. Additionally, all groups used the known dose range of the cohort when assigning their predicted doses. Because of the use of this known range, some of these methods may be tailored particularly to this challenge and not be generalizable to the wider population.

3.5.4 Overall lessons from CAGI exomes challenge

An advantage of the common task structure is the ability to iterate quickly and learn from the setbacks of the groups analyzing the data. The exome challenges allowed us to glean several important lessons that will inform future iterations of CAGI.

The importance of population stratification, batch effects, and hidden biases became evident early on with CAGI 2 Crohn's disease challenge (Figure 3.1). In that particular instance, either population stratification or batch effects created a discernable difference between cases and controls that was unlikely related to actual disease status. Based on that finding in CAGI 2, every subsequent CAGI challenge included a pre-analysis of the whole exome data trying to identify if there were samples that clustered together inappropriately based on case-control status. Population stratification has long been an issue in genetic studies. The most obvious issue arises when cases and controls come from distinctly different ancestral populations – such as comparing Northern European cases against Chinese controls. However, less obvious stratification can also be an issue – such as differences in admixture/population substructure or cryptic relatedness [134]. Batch effects can occur at many different steps in the pipeline, for example if samples from the cases and controls have differences in sample preparation, DNA quality, sequencing coverage, or genotype calling. Any of the above can result in prediction methods that perform well due to systemic biases between cases and controls rather than true features that define case-control status. How these challenge datasets emulate the real world was another important consideration and was a topic of discussion among the CAGI 4 community.

A majority of the challenges used samples of Northern European ancestry –only the warfarin dose prediction challenge used samples of African ancestry. In order for the

methods to be generalizable to real world populations, representation of human diversity is necessary, particularly since disease risk and pharmacogenetic variants can be population-specific [135]. Moreover, the CAGI exome datasets all came from research studies, which are often designed to maximize the possibility of picking up a significant signal. One way to achieve this is through selecting for extreme phenotypes – a strategy employed by both the Crohn's disease exome dataset (which selected a subset of cases who had early-onset Crohn's disease) and the warfarin prediction exome dataset (selected from individuals requiring "low" and "high" doses to achieve the therapeutic index)[136]. However, while this strategy works well for increasing signal strength in research, using such data for building a classifier may lead to a biased predictor that has difficulty differentiating between the subtler variations seen in the real world. Having larger datasets and using data generated for clinical use may help remedy some of these issues in the future. And finally, one of the most promising lessons from CAGI was on the effectiveness of data. As mentioned before, for complex tasks, the common task framework has provided a way to have many people work on a problem and iterate quickly. After a challenge has ended, sharing the evaluation scripts and the challenge answers allows participants to analyze when their prediction methods succeed or fail in order to improve further. Additionally, large datasets, even if imperfect, have also been shown to be a critical part of developing algorithms to tackle a complicated task [137]. Critical to accumulating large enough datasets is data sharing, and the open data movement aims to encourage increased biomedical data sharing [138]. However, one of the difficulties with genetic data that includes protected health information is sharing data in a secure manner. CAGI, which includes data encryption and verifies the groups participating can provide a platform to facilitate sharing such data. As a result of the data accumulated thus far, CAGI has demonstrated how data can, in certain cases, surmount prior biological knowledge. For CAGI 4, the Bipolar Disease challenge was the best example; individuals with no biological background, but a strong background in data science had the best performance. In particular, this should inspire a more multi-disciplinary approach to genotype-phenotype prediction and a greater effort to engage those whose backgrounds are more data-driven rather than biologically-driven.

Overall, the CAGI exome challenges provided an opportunity to begin building the classifiers required to implement precision medicine. While there is still a long road ahead for genotype-phenotype prediction, the accumulation of larger datasets and the participation of more groups with every subsequent CAGI holds promise for continued improvement.

4 Lessons from the CAGI-4 Hopkins clinical panel challenge

This Chapter has been published in "Chandonia J.M, Adhikari A., Carraro M., Chhibber A., Cutting G.R., Fu Y., Gasparini A., Jones D.T., Kramer A., Kundu K., Lam H.Y.K., Leonardi E., Moult J., Pal L.R., Searls D.B., Shah S., Sunyaev S., Tosatto S.C.E., Yin Y., Buckley B.A. Lessons from the CAGI-4 Hopkins clinical panel challenge. Hum Mutat. 2017 Apr 11.". For Supplementary Materials, check the online version of the paper.

4.1 Summary

The CAGI-4 Hopkins clinical panel challenge was an attempt to assess state of the art methods for clinical phenotype prediction from DNA sequence. Participants were provided with exonic sequences of 83 genes for 106 patients from the Johns Hopkins DNA Diagnostic Laboratory. Five groups participated in the challenge, predicting both the probability that each patient had each of fourteen possible classes of disease, as well as one or more causal variants. In cases where the Hopkins laboratory reported a variant, at least one predictor correctly identified the disease class in 36 of 43 patients (84%). Even in cases where the Hopkins laboratory did not find a variant, at least one predictor correctly identified the class in 39 of 63 patients (62%). Each prediction group correctly diagnosed at least one patient that was not successfully diagnosed by any other groups. We discuss the causal variant predictions by the different groups and their implications for further development of methods to assess variants of unknown significance. Our results suggest that clinically relevant variants may be missed when physicians order small panels targeted on a specific phenotype. We also quantify the false positive rate of DNA-guided analysis in the absence of prior phenotypic indication.

4.2 Introduction

DNA sequencing tests are increasingly used in medical practice to confirm or assign clinical diagnoses [4]. However, the interpretation and classification of novel sequence variants identified in a patient remains difficult, even for well-studied disorders like cystic fibrosis [139]. Improved computational methods may aid in the interpretation of sequence

variants and, when used in conjunction with clinical data, could increase the confidence of a diagnosis [140]. Until recently, genetic testing was limited to genes associated with a specific clinical phenotype. However, recent technological advances have made it feasible to sequence large gene panels, exomes, and genomes [141–143]. As the number of genes sequenced per patient increases, the number of novel, rare, and unclassified variants also increases. Clinical molecular geneticists must determine which variants, if any, are likely to contribute to the patient's clinical presentation. The current gold standards for assessing a variant's pathogenicity are segregation of the variant with the clinical phenotype in multiple pedigrees, and functional assays demonstrating a detrimental effect of that specific nucleotide change. In most instances, when a novel genetic variant is identified there is no rapid and reliable method to assess its pathogenicity. Predictive software tools are interrogated, but none are considered strong evidence to assert a novel variant's pathogenicity [144]. The shift towards analyzing large datasets has led to a need for high-throughput methods to aid in variant classification and also for computation tools to help better interrogate the increasing number of variants of uncertain clinical significance.

Crowd sourced data analysis challenges such as the 4th Critical Assessment of Genome Interpretation (CAGI-4) have emerged as a framework to compare predictive methods and assess the overall state of particular analysis areas [145]. In the CAGI-4 Hopkins Clinical Panel challenge, participants were asked to develop or use existing computational methods to analyze data from a next generation sequencing (NGS) panel in order to match a patient's genotype to their clinical phenotype in the absence of additional clinical information. The Johns Hopkins DNA Diagnostic Laboratory (henceforth, Hopkins), a CLIA and CAP certified lab that specializes in clinical molecular testing for rare, inherited disorders, provided data for this challenge. The Hopkins lab offers testing for approximately 50 phenotypes and disorders totaling 3,500 tests annually. They offer NGS-based tests targeted for ~20 specific phenotypes. The same NGS capture probe set is used for all panels and only the requested genes are analyzed in each patient. Hopkins provided CAGI-4 organizers with the VCF files for the entire NGS panel for 106 patients with a range of clinical presentations. The genetic disorders associated with variants in the 83 genes on the panel were grouped into 14 'disease classes' which include lung disorders, peroxisomal disorders, aneurysm disorders and craniofacial disorders (Table 4.1, Supp. Table S-4.4). The goal of the challenge was for the participants to match each patient to a disease class based on

informatics analysis of the sequence data. A further part of the challenge was to predict the specific gene and variant(s) that is/are the underlying cause of disease.

4.3 Materials and methods

4.3.1 Sequencing, variant calling, and analysis by the Hopkins lab

Gene sequences were captured using one of two custom probe sets (Agilent SureSelectXT Target Enrichment Kit) and sequenced by a NGS platform (Illumina MiSeq, 2x100 nt reads). The NGS panels used to test assessed exons and exon-adjacent sequences for 64 or 83 loci (Supp. Table S-4.4, Supp. Table S-4.5). Sequences were aligned to the human reference genome (GRCh37/hg19) using the Burrows-Wheeler Aligner (bwa). Sequence variants were called individually for each patient to produce two Variant Call Format (VCF) files, one for single nucleotide variants (SNVs; GATK UnifiedGenotyper, v2.7-4) and one for insertion-deletion variants (InDels; GATK HaplotypeCaller, v2.7-4). De-identified VCF files were provided to the CAGI-4 organizers. Note that the CAGI-4 organizers combined individual VCF files for each patient into a single VCF, resulting in potentially misleading data in the INFO and FILTER fields of the file. The panel of 83 genes was sequenced in 96 of the 106 patients; for the other 10 patients, a partially overlapping list of 64 genes were sequenced (Supp. Table S-4.5). Although the whole NGS panel was sequenced in all patients, only the genes selected on the patient's test requisition form were analyzed by the lab (n=1-24 genes/patient).

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

The Hopkins lab included variants in the genes they analyzed that were classified as Variants of Uncertain Significance (VUS), Likely Pathogenic, and Pathogenic as an answer key. The disease class of each patient was also provided in the answer key and reflects the test selected by the patient's physician on the test requisition form. The ~20 phenotypes that Hopkins tests for were narrowed down to 14 disease classes in order to simplify the challenge (Supp. Table S-4.1). Some disease classes were not represented by any patients and were included as red herrings (Supp. Figure S-4.9).

4.3.2 Challenge format

Participants in the Hopkins clinical panel challenge were provided with the two VCF files above, a detailed description of the 14 disease classes given in Table 4.1, a submission template, a submission validation script, and the gene capture regions used in sequencing the patients (in Browser Extensible Data, or BED format). Participants were also instructed that every patient matched exactly one disease class.

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

Table 4.1: A summary of the 14 disease classes in the CAGI-4 Hopkins clinical panel challenge

Disease class	Description
Cystic fibrosis and CF-related disorders	Classic cystic fibrosis consists of progressive lung disease, exocrine pancreatic insufficiency, and male infertility.
Diffuse lung disease	Diffuse lung disease is an umbrella term encompassing multiple lung disease phenotypes.
Primary ciliary dyskinesia	Primary ciliary dyskinesia is a genetically heterogeneous group of disorders resulting from dysfunction in different parts of the cilia.
Peroxisomal beta-oxidation defects	The majority of patients with peroxisomal beta-oxidation defects have liver disease, brain malformations, developmental retardation, sensory deficits, and dysmorphic craniofacial features.
Rhizomelic chondrodysplasia punctata	Symptoms of rhizomelic chondroplasia punctata include proximal shortening of the limbs, cataracts, severe intellectual disability, seizures, and calcific stippling of cartilage.
Zellweger spectrum disorders	Zellweger spectrum disorders consist of Zellweger syndrome (cerebro-hepato-renal syndrome; most severe phenotype), neonatal adrenoleukodystrophy (intermediate phenotype), and infantile Refsum disease (mildest phenotype).
Loeys-Dietz syndrome	Loeys-Dietz syndrome is a connective tissue disorder that predisposes individuals to aortic aneurysms.
Marfan syndrome	Marfan syndrome is an inherited connective tissue disorder that affects the skeletal, ocular, and cardiovascular systems.

Continues

Table 4.1 (Continued)

Disease class	Description
Thoracic aortic aneurysm and dissection	Thoracic aortic aneurysm and dissection is a cardiovascular disease characterized by dilation of the aorta, which leads to aortic aneurysms (most commonly in the ascending aorta) and aortic dissection.
Ataxia telangiectasia	Ataxia-telangiectasia is a disorder of childhood onset progressive cerebellar ataxia and occulocutaneous telangiectasias.
Liddle syndrome	Liddle syndrome is a rare genetic disorder characterized by early onset high blood pressure (hypertension) and low blood potassium (hypokalemia).
Pseudohypoaldosteronism type 1	Pseudohypoaldosteronism type 1 is a salt-wasting disease with onset during infancy.
Telomere shortening disorders	Telomere shortening disorders represent a spectrum of phenotypes that result from mutations in genes involved in telomere maintenance protein complexes.
Treacher Collins and related syndromes	Treacher Collins syndrome is a rare disorder affecting craniofacial development.

4.3.3 Assessment

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

In other cases, assessors calculated one metric (nCorrect) in which the number of correct predictions was counted, giving ties full credit; another metric ($nCorrect_{tie}$) was calculated in which N-way ties were given 1/N credit.Finally, assessors calculated a third metric ($nCorrect_{var}$) in which they counted the number of predictions for which the disease class

was correct (giving ties full credit) AND for which at least one of the variants submitted in the corresponding column for that disease class matched one of the variants noted by Hopkins.

Assessors also calculated the following metrics for each submission:

avgPCorrect – the average probability assigned by the predictor to the correct disease class. This statistic provides an assessment of predictions that is not dependent on whether the submitter's highest probability prediction was correct.

$avgPCorrect_{norm}$– the average probability assigned by the predictor to the correct disease class, after normalizing all probabilities predicted in each submission for each patient to sum to 1.0. (Exception: if all probabilities for a patient were zero, they were not normalized).

avgRank – the average rank assigned by the predictor to the correct disease class. Ties were assigned the average rank of each set of tied predictions; e.g., if the two highest probability disease classes had equal rank, both were assigned a rank of 1.5; a 3-way tie for 2nd highest probability would be assigned a rank of 3. Note that because there were 14 disease classes, an all-zero prediction would have an avgRank score of 7.5 (i.e., was scored as a 14-way tie).

avgError – the average error in predictions, where the error was measured as the absolute difference between the probability assigned each disease class and zero (if not the correct disease class) or one (if the correct disease class). Like avgPcorrect, avgError assesses predictions independent of their rank, but also includes correct negative predictions.

4.3.4 Prediction Methodology

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

Group 57 (Jones)**:** The Jones-UCL group made use of one-class Support Vector Machine (SVM) classifiers to automatically assign disease classes according to the supplied exome data. In a normal machine learning experiment, sufficient positive and negative cases are needed to define a hypersurface which separates the two classes. Standard SVMs attempt to define this hypersurface such that the chance of misclassifying new cases is minimized. In some applications, however, only positive or negative cases are readily available, but not both. One-class SVMs [146] have been proposed for problems where either negative or positive case data is unavailable. In this situation, the SVM attempts to identify outliers from a distribution modeled on the available single class of data, and it is assumed that the

outliers belong to the alternative class. In this CAGI challenge, of course, neither negative nor positive training data was readily available. However, the assumption was made that the 1000 Genomes data set [147] could be used as a proxy for negative case data. This is a reasonable assumption if we assume that the diseases in question are relatively rare. To start with, gene variants relating to each disease class were collated using ClinVar [148]. Feature sets were generated for each disease class by encoding variant 0/0, 0/1 and 1/1 calls as 0, 1 and 2 respectively, and for each disease-specific feature set, a one class ν-SVM (using a RBF kernel) was trained. The single parameter ν, which controls both the number of support vectors and the misclassification cost, was optimized for each disease class so as to minimize the number of outliers detected in the 1000 Genome training data. Once trained, the SVM was then applied to the test sample data, and the distance to decision boundary was used as a proxy for classification confidence. The most important variant was identified in each case by systematically removing each variant from the feature set and recalculating the confidence scores.

Group 58 (Tosatto): The analysis started with a manually curated association between the genes of the panel and the 14 clinical phenotypes of interest based on literature review. Sequencing data was annotated with ANNOVAR [149], considering for each variant the corresponding affected gene, frequency estimated from the 1000 Genomes Project [150] and predicted pathogenicity score from SIFT [151] and PolyPhen2 [52]. The method to define association between genetic data and phenotypes was based mainly on two phases. For each individual, variations that are less probable to be disease causing were filtered out and a probability to be affected based on the analysis of variants defined. Only coding and splice-site variants which can affect protein function were considered according to the Common Disease-Rare Variant Hypothesis (CDRVH)[152]. Common (MAF > 5%) and/or synonymous single nucleotide variations (SNVs) were filtered out. Insertion and deletions were excluded as their impact on protein function is difficult to predict compared to SNVs. Only insertions and deletions (indels) affecting the coding part of a gene and predicted to be "damaging" or known to be pathogenic were considered. Heterozygous indels in genes with autosomal recessive inheritance, occurring in GC-rich or repeated regions were filtered out from the disease candidate mutation pool. An empirically derived scoring scheme was implemented to define association between patients and phenotypes, considering both disease inheritance and predicted SNV pathogenicity (Supp. Table S-4.2). Different weights were assigned to different mutation types, i.e. a high score for known variants associated with a specific disease (mainly by literature review) and a lower score for

mutations not affecting protein function according to predictor output (i.e. tolerated, benign and unknown). For autosomal dominant (AD) pathologies, only heterozygous variants plus few manually curated homozygous mutations were considered (i.e. the one with the highest probability score). The disease cutoffs were set at different values between submissions, allowing the stringency of the analysis to vary. Both homozygous and compound heterozygous variants were considered for autosomal recessive (AR) conditions. When more than one match per patient occurred, only the most likely was considered (e.g. the one with higher probability score). Different submissions correspond to different sets of weights. In particular, in the first submission, a slightly lower weight was assigned to variants whose effect is more difficult to assess (i.e. compound heterozygous, homozygous variants with uncertain significance, variants affecting different genes coding for subunits of the same complex) with respect to submission 4.

Group 59 (Qiagen Bioinformatics): All 106 samples were uploaded to Ingenuity Variant Analysis (QIAGEN- Hereditary Disease Solution) and set up an analysis with all samples to filter low quality (call quality < 20) and common variants (>0.5% MAF in 1000 Genomes [147], NHLBI-EVS (http://evs.gs.washington.edu/EVS/), ExAC [153], and Allele Frequency Community (www.allelefrequencycommunity.org), using the Confidence and Common Variants filters, respectively. The Allele Frequency Community is a QIAGEN hosted allele frequency database, founded by QIAGEN and participating members in 2014. It is a freely accessible "opt-in" community resource designed to facilitate sharing of anonymized, pooled allele frequency statistics among community members. The Predicted Deleterious filter was used to keep only those variants that are previously published and classified Pathogenic or Likely Pathogenic, using ACMG guidelines, DM variants (pathological mutations reported to be disease causing in the original literature report) present in HGMD, along with other loss of function (frameshift, start/stop loss or gain, splice site) and missense variants. Finally, the biological context filter was applied to find variants linked to each one of the 14 categories and patient disease category was predicted based variant-disease connection, using path-to-phenotype evidence.

Group 60 (RSS): Gene phenotype associations were mined from the Hopkins diagnostic panels, OMIM [154], and GeneReviews [155]. Inheritance mode and penetrance information were extracted from online resources for each gene-phenotype pair. Variants with low quality or high population allele frequencies were filtered out and the functional impact was annotated with Variant Effect Predictor[156]. To estimate the probability that a variant is damaging to protein function, we integrated multiple prediction methods to score all types

of variants, e.g. missense, nonsense, indels and intronic variants. The damaging scores were scaled and normalized to reflect the relative deleteriousness, e.g. frame-shift / nonsense variants would have higher scores than missense variants. We then used the damaging scores to estimate the probability that each individual has a particular phenotype with a probabilistic model, i.e. calculated as the probability that at least one associated gene in the individual causes the phenotype. For a particular gene, the probability the gene causes the phenotype was calculated as the probability that the gene is disrupted (taking into account inheritance mode) multiplied by its penetrance score. The confidence level of the prediction was calculated from the distribution of the estimated probabilities across phenotypes and across individuals. Considering the 14 phenotypes are Mendelian like diseases, if one individual has high prediction scores across phenotypes, it is more likely to be false positive. Thus high confidence was assigned to individuals with high variability across phenotypes. A more detailed description of this group's prediction methods is included in the Supplementary Information.

Group 61 (Moult): The method (implemented in Python) has four modules – Variant annotation, QC (quality check), Variant Prioritization, and Probability scoring for the disease. The modules were executed sequentially. Inputs were the two gVCF files and a gene configuration file containing the genes associated with each disease class and their inheritance pattern. The Varant tool (http://compbio.berkeley.edu/proj/varant) was used to annotate variants with: region of occurrence in the genome, allele frequency from ExAC [153], predicted pathogenicity based on four methods [52,157–159] (for missense), and previously reported disease associations in databases [148,160]. Three QC analyses were run: (1) Variant counts (common vs. rare vs. novel & homozygous vs. heterozygous) per sample, (2) Read depth for each gene in each sample was obtained by averaging DP values over all bases in a gene recorded in the gVCF file, and (3) Exons with relatively low or no coverage compared to other exons in a gene. The QC qualified variants per sample were prioritized by first assigning them to one of three classes, ranked by the likelihood that the variant is causative and further grouping the variants in each class by frequency based on its ExAC MAF (group 1 – novel, 2 - very rare (MAF <= 0.005), or 3 – rare (MAF <= 0.01). Class-1 identified variants previously reported in disease databases as pathogenic, Class-2 identified loss of function, splice and missense variants predicted damaging by in-silico prediction tools, and Class-3 identified missense variants (not predicted damaging), UTR, and intronic variants. Variants were further filtered for inheritance model. For each sample, once putative causative variants were found, the process was terminated (e.g. if a suitable

variant or variants were found using Class-1, Class-2 and Class-3 were not executed). Finally, a probability score for a sample to have a particular disease was computed based on the type of prioritized variant(s) and inheritance pattern. For the missense variants, the probability model was based on the extent of consensus among the four prediction methods, using a previous HGMD derived calibration. For other variant types, subjective probability rules were used.

4.4 Results

4.4.1 Summary of submissions

Five groups submitted predictions (with 4, 2, 2, 2, and 1 distinct prediction per group). An overview of the challenge and results is shown in Figure 4.1. The 106 patients in the challenge can be roughly grouped into two difficulty classes: 1) patients for whom Hopkins noted a potentially causal variant in the answer key (43 patients) and 2) patients for whom Hopkins did not note any variants (63 patients) (Figure 4.1A). At least one CAGI-4 predicting group correctly predicted the disease class for 36 of the 43 patients who had a reported variant (Figure 4.1B). Fewer groups correctly predicted both the disease class and at least one of the variant(s) that Hopkins reported (Figure 4.1C). CAGI-4 predictors were not as accurate at predicting disease classes for the remaining 63 patients for whom Hopkins did not note a variant, although at least one group correctly predicted the disease class for the majority of these patients (Figure 4.1D). The lower prediction accuracy is perhaps unsurprising given the negative test results for these 63 patients.

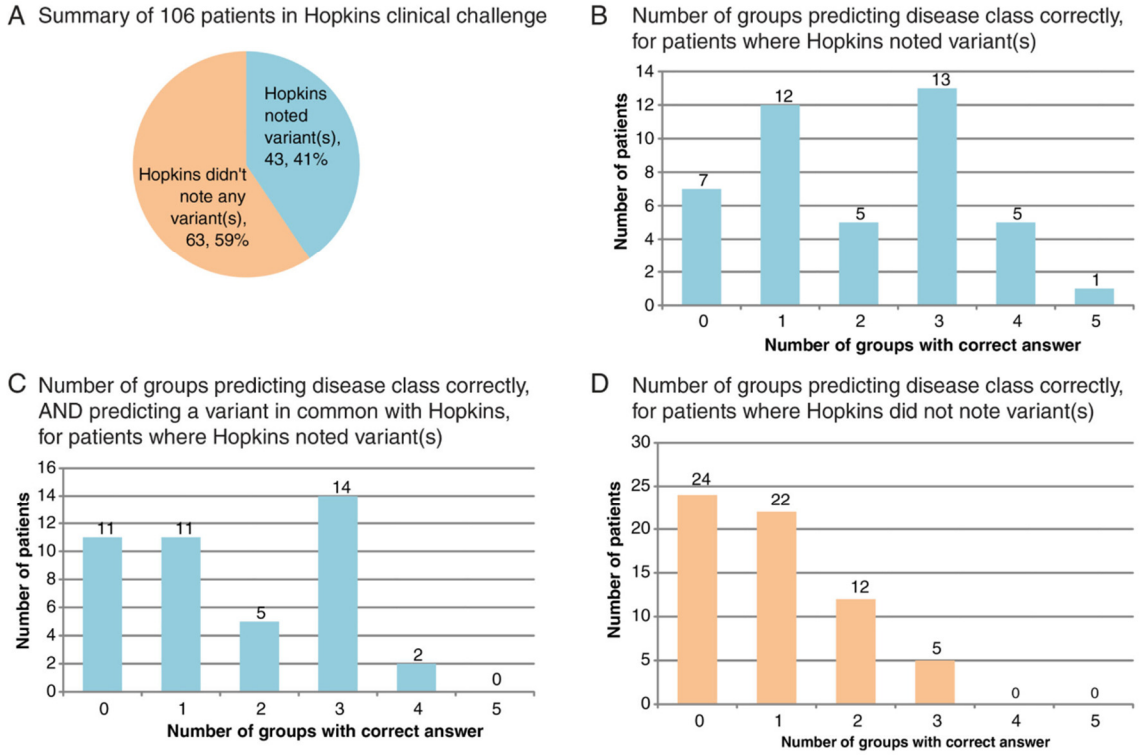


Figure 4.1: Summary of CAGI-4 Hopkins clinical panel challenge and results.

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

4.4.2 Numeric assessment summary

Table 4.2 summarizes our numeric assessment metrics for each non-redundant, submitted prediction, for all patients. Table 4.3 shows the same statistics for only the 43 patients for which Hopkins noted at least one potentially causal variant. The best values for each metric in each Table are indicated in bold. Each group's overall performance is briefly discussed below. Table 4.4 shows a summary of the performance of all predicting groups on each patient. An expanded version of Table 4.4 with additional columns is provided as Supplementary Information (Supp. Table S-4.6). Tables 4.5 and 4.6 summarize the most frequent combinations of groups that predicted the correct disease class for patients (Table

4.5 ignores causal variant predictions, while Table 4.6 requires each group to predict one of the variants noted by Hopkins).

Group 57 (Jones): Group 57's primary submission (57.1) scored much higher than their other submissions by our metrics. Their method was less accurate than other groups in cases where Hopkins reported a potential causal variant, but it was more accurate at predicting the correct disease class in cases where Hopkins didn't report a variant. Group 57's primary submission was also the most accurate among all submissions at rank-ordering the disease classes. As seen in Table 4.5, Group 57 predicted disease classes correctly for 18 patients that no other group predicted correctly, with seven of these cases in their primary submission. This method was unique in that it did not attempt to mimic a traditional clinical genetics approach. No attempt was made to independently predict the pathogenicity of the ClinVar variants used as features or to correct for linkage disequilibrium, which may explain why the method was able to make correct inferences where no causal variants were reported and why correct inference can arise without reporting the correct variants. A possibility is that some or even a majority of the variants relied on by the classifiers were non-causal variants which simply happen to be in linkage disequilibrium with one or more true causal variants. Thus the occurrence of these variants were sufficient to identify the sample as a genetic outlier, though not indicating true causation. It is possible that by addressing these issues, the method might be further enhanced to make more accurate predictions relating to true causal variants. It would be interesting to test this method on a larger dataset to rule out the possibility that there is some underlying structure in this dataset that the algorithm is detecting.

Group 58 (Tosatto): As seen in Table 4.5, most cases that Group 58 predicted correctly were also predicted by at least one other group. However, Group 58 predicted the disease class for one patient (P81) that no other groups predicted; they also assigned 100% probability of the correct disease to that patient, and predicted exactly the same causal variants as noted by Hopkins. Many of the diseases in this challenge result from loss of function variants in a given gene, thus by excluding frameshift variants (out of frame deletions and/or insertions within an exon) Group 58 missed these cases. The genes and molecular mechanisms associated with each of the 14 disease classes were not provided as part of the dataset, which increased the difficulty of the matching exercise (Supp. Table S-4.2).

Group 59 (Qiagen): Group 59 had the highest average P values for the correct disease classes, after normalization; they also had some of the best scores in the avgError metric.

Group 59 correctly predicted the disease class for five patients that no other groups predicted. Among all the groups, they were the only group for which both P values and SD values were independent and positively correlated with the values they were expected to correlate with (see discussion of P and SD, below). This challenge was well-suited for the Qiagen group, as they specialize in large scale variant interpretation [161].

Table 4.2: Summary of assessment metrics for each non-redundant, submitted prediction, for all patients

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

Note: Predictions are numbered according to the group's (formerly anonymized) group number (57, Jones; 58, Tosatto; 59, Qiagen Bioinformatics; 60, RSS; 61, Moult) and the group's submission number (1, most confident prediction; other non-redundant predictions are as numbered by the submitters, up to six per group).

Table 4.3: Summary of assessment metrics for each non-redundant, submitted prediction, for the 43 patients for which Hopkins noted at least one potentially causal variant

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

Note: Predictions are numbered as in Table 4.2.

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

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

4.4.3 Accuracy of P and SD values

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

Table 4.4: Summary of the performance of all predicting groups on each patient

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

(Continues)

Table 4.4 (Continued)

Patient	nC	nCV	Correct groups	Correct groups, with variant	Correct predictions	Correct predictions, with variant
P34	3	3	58, 59, 61	58, 59, 61	59.2, 58.4, 61.1, 58.1, 59.1	59.2, 58.4, 61.1, 58.1, 59.1
P35	0	N/A	None	N/A	None	N/A
P36	0	0	None	None	None	None
P37	0	0	None	N/A	None	N/A
P38	3	3	58, 60, 61	58, 60, 61	58.4, 60.1, 60.2, 61.1, 58.1	58.4, 60.1, 60.2, 61.1, 58.1
P39	1	0	59	None	59.2, 59.1	None
P40	0	N/A	None	N/A	None	N/A
P41	0	N/A	None	N/A	None	N/A
P42	2	1	59, 61	61	59.2, 61.1, 59.1	61.1
P43	2	N/A	57, 61	N/A	61.1, 57.1	N/A
P44	1	N/A	59	N/A	59.2, 59.1	N/A
P45	1	N/A	57	N/A	57.1	N/A
P46	0	N/A	None	N/A	None	N/A
P47	1	1	61	61	61.1	61.1
P48	0	0	None	None	None	None
P49	1	N/A	57	N/A	57.1	N/A
P50	0	N/A	None	N/A	None	N/A
P51	1	N/A	61	N/A	61.1	N/A
P52	3	3	58, 59, 61	58, 59, 61	59.2, 58.4, 61.1, 59.1	59.2, 58.4, 61.1, 59.1
P53	2	N/A	58, 59	N/A	59.2, 58.4, 58.1, 59.1	N/A
P54	1	N/A	57	N/A	57.3	N/A
P55	0	0	None	None	None	None
P56	2	1	58, 59	59	59.2, 58.1, 59.1	59.2, 59.1
P57	1	1	61	61	61.1	61.1
P58	0	N/A	None	N/A	None	N/A
P59	0	0	None	None	None	None
P60	2	2	59, 61	59, 61	59.2, 61.1, 59.1	59.2, 61.1, 59.1
P61	1	N/A	57	N/A	57.4, 57.3	N/A
P62	2	N/A	57, 60	N/A	60.1, 60.2, 57.1	N/A
P63	3	N/A	57, 59, 61	N/A	59.2, 61.1, 57.1, 59.1	N/A
P64	1	1	60	60	60.1, 60.2	60.1, 60.2
P65	3	N/A	58, 60, 61	N/A	58.4, 60.1, 60.2, 61.1	N/A
P66	0	N/A	None	N/A	None	N/A
P67	0	0	None	None	None	None
P68	1	N/A	59	N/A	59.2, 59.1	N/A
P69	0	0	None	None	None	None
P70	0	N/A	None	N/A	None	N/A
P71	0	N/A	None	N/A	None	N/A

(Continues)

Table 4.4 (Continued)

Patient	nC	nCV	Correct groups	Correct groups, with variant	Correct predictions	Correct predictions, with variant
P72	3	2	57, 59, 61	59, 61	59.2, 61.1, 57.2, 59.1	59.2, 61.1, 59.1
P73	4	3	57, 58, 59, 61	58, 59, 61	59.2, 58.4, 61.1, 57.1, 58.1, 59.1	59.2, 58.4, 61.1, 58.1, 59.1
P74	0	N/A	None	N/A	None	N/A
P75	0	N/A	None	N/A	None	N/A
P76	0	N/A	None	N/A	None	N/A
P77	0	N/A	None	N/A	None	N/A
P78	1	N/A	61	N/A	61.1	N/A
P79	0	N/A	None	N/A	None	N/A
P80	3	3	57, 59, 61	57, 59, 61	59.2, 61.1, 57.1, 57.2, 59.1	59.2, 61.1, 57.1, 57.2, 59.1
P81	1	1	58	58	58.4, 58.1	58.4, 58.1
P82	1	N/A	57	N/A	57.2	N/A
P83	1	N/A	57	N/A	57.4, 57.3	N/A
P84	4	4	57, 58, 59,61	57, 58, 59, 61	59.2, 58.4, 61.1, 57.2, 58.1, 59.1	59.2, 58.4, 61.1, 57.2, 58.1,59.1
P85	2	N/A	57, 61	N/A	57.4, 61.1	N/A
P86	2	N/A	58, 59	N/A	59.2, 58.4, 58.1, 59.1	N/A
P87	3	N/A	57, 58, 61	N/A	57.4, 58.4, 61.1, 57.1, 58.1, 57.3	N/A
P88	1	N/A	57	N/A	57.1	N/A
P89	1	N/A	57	N/A	57.4, 57.1	N/A
P90	2	N/A	58, 61	N/A	58.4, 61.1, 58.1	N/A
P91	1	N/A	57	N/A	57.2	N/A
P92	3	3	58, 59, 61	58, 59, 61	59.2, 58.4, 61.1, 59.1	59.2, 58.4, 61.1, 59.1
P93	1	1	60	60	60.1, 60.2	60.1, 60.2
P94	2	2	59, 61	59, 61	59.2, 61.1, 59.1	59.2, 61.1, 59.1
P95	3	3	58, 59, 61	58, 59, 61	59.2, 58.4, 61.1, 58.1, 59.1	59.2, 58.4, 61.1, 58.1, 59.1
P96	1	0	57	None	57.3	None
P97	0	N/A	None	N/A	None	N/A
P98	2	N/A	57, 61	N/A	61.1, 57.1	N/A
P99	1	N/A	59	N/A	59.2, 59.1	N/A
P100	0	N/A	None	N/A	None	N/A
P101	2	N/A	57, 61	N/A	61.1, 57.1	N/A
P102	1	N/A	57	N/A	57.3	N/A
P103	0	N/A	None	N/A	None	N/A
P104	3	3	58, 59, 61	58, 59, 61	59.2, 58.4, 61.1, 58.1, 59.1	59.2, 58.4, 61.1, 58.1, 59.1
P105	3	3	58, 59, 61	58, 59, 61	59.2, 58.4, 61.1, 58.1, 59.1	59.2, 58.4, 61.1, 58.1, 59.1
P106	1	N/A	61	N/A	61.1	N/A

Note: An expanded version of Table 4.4 with additional columns has been provided in Supp. Information (Suppl. Table S4.6). nC, number of groups predicting the disease class correctly, among all submissions from each group (counting ties, except in cases where all 14 disease classes were assigned equal probability; nCV, number of groups predicting both the correct disease class and at least one variant noted by Hopkins; correct groups, a list of groups in which the disease class was predicted correctly in at least one submission (counting ties, except in cases where all 14 disease classes were assigned equal probability).Groups are numbered as in Table 4.2; correct groups, with variant, a list of groups with at least one prediction of the correct disease class, and also at least one variant noted by Hopkins (N/A in this field indicates that Hopkins did not note any variants). Predictions are numbered as in Table 4.2; correct predictions, with variant, same as above, but indicating individual submission numbers that were correct.

4.4.4 Commentary on novel variant predictions

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

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

Table 4.5: Frequency with which each combination of groups correctly diagnosed patients.

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

Notes: This Table summarizes the number of times each combination of groups correctly diagnosed patients, as shown in the "correct groups" column of Table 4.4.

Table 4.6: Frequency with which each combination of groups correctly diagnosed patients, and also noted a Hopkins variant.

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

Notes: This Table summarizes the number of times each combination of groups correctly diagnosed patients and predicted at least one variant noted by Hopkins, as shown in the "correct groups with variant" column of Table 4.4

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

Patient P7 – Groups 57 (submission 4), 58, 59, and 61 all predicted Telomere Shortening Disorders, and the latter 3 groups consistently noted a missense variant in *TERT*. The

patient's diagnosis was Cystic Fibrosis and CF-Related disorders, and Hopkins did not note any reportable variants and did not analyze the *TERT* gene. The *TERT* variant is described in the literature; it leads to telomere shortening and is involved in bone marrow failure. Telomere shortening due to mutations in *TERT* is known to be involved in pulmonary fibrosis. Clinical presentation of pulmonary fibrosis is very different from cystic fibrosis. This *TERT* variant is annotated in ClinVar as involved in pulmonary fibrosis, but literature support for this phenotype is unclear. The variant is found in 120 ExAC participants including 2 homozygotes.

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

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

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

the background frequencies of these two variants in further detail, in order to decide whether they are likely to be causative.

4.5 Discussion

Overall, we found that current state of the art computational prediction methods does a reasonable job of predicting clinical phenotype from genotype, even when blinded to clinical diagnoses. At the same time, current genotype-driven prediction methodologies generate false positives and false negatives at a rate unacceptable for clinical use. In cases where the Hopkins lab reported a variant, predictors did relatively well, with at least one group correctly identifying the disease class in 36 of 43 patients (84%), and at least one group identifying the correct disease class and variant in 33 of 43 cases (77%). In cases where the Hopkins lab did not find a reportable variant in the genes they analyzed, at least one group correctly matching the disease class in 39 of 63 patients (62%). In the latter cases, methods based on machine learning (SVM) technology appeared to be most effective at correctly identifying the disease. Interestingly, despite the ability to correctly match genotype to phenotype, the SVM-based method could not correctly identify the pathogenic variant. It is unclear what is happening in cases where groups correctly identify the disease class, but not the causal variant. In retrospect, it would have been prudent to include a list of gene-disease associations as well as modes of inheritance to the predictors to aid in the matching process. Different groups performed better depending on which metric was used; there was no clear "winner" that dominated performance across all metrics. Indeed, every group predicted at least one patient's disease class correctly that no other group predicted correctly. This result suggests that a "meta-predictor" or a human clinical expert with access to all groups' results might improve on the performance of each individual group. Currently, clinical genetic testing is almost entirely phenotype-driven: given a clinical diagnosis, laboratories analyze variants in genes known to be relevant to the diagnosed disease. This is partially due to the historic technical limitations on genetic testing, e.g., sequencing costs limited the number of genes for which data could be obtained. The standards for reporting variants to the patient are also currently conservative, in part because common, benign polymorphic variants have caused many false positives in past genetic analyses [162,163]. However, as whole-exome and whole-genome sequencing become more economical, the phenotype-driven paradigm may be replaced by a genomics-driven approach, in which all rare, putatively functional variants in a patient's genome are first

identified, then evaluated based on the plausibility that they may be pathogenic. The genomics-driven approach has the potential for higher sensitivity, due to more genes being analyzed, and also has the potential to diagnose diseases not identified by the referring physician. However, the main tradeoff compared to phenotype-driven approaches is a potentially higher false positive rate. Multiple CAGI-4 groups in the Hopkins challenge were in consensus in identifying several possible causative variants that were not identified by the current panel testing paradigm. They also identified several other variants that were likely to be false positives. Distinguishing these two possibilities, and identifying which variants to report to the patient, is a topic that requires further research. The American College of Medical Genetics and Genomics has published guidelines for the interpretation of sequence variants in order to help codify variant assessment [144]. However, even when adhering to these guidelines there are still elements of variant interpretation that are subjective and vary between labs [164,165]. Given large databases of "control" exomes (i.e., without a known phenotype), researchers could develop statistical models to predict whether particular variants are in fact causative [153]. Such models could inform the development of new statistically justified reporting standards based on, for example, particular thresholds on the probability that the prediction of a causal variant is a false positive. This challenge was designed to reflect the range of cases seen in the Hopkins diagnostic lab (Figure 4.1A). This includes a high percentage of cases for which no likely pathogenic variant was identified, despite the patient presenting with a clinical phenotype. Even for clinical exome sequencing, nearly 75% of cases are negative [141,142]. Negative cases proved especially challenging to participants, as 'phenotype not discernable' was not listed as a matching option. Despite the fact that no pathogenic variants were identified by the Hopkins lab, most groups were able to make a disease prediction and to identify putative pathogenic alleles in these negative cases. Indeed, the reason data from all 83 genes was included in the challenge was to highlight the difficulty in interpreting a large data set of rare variants that are unrelated to the patient's phenotype. The presence of negative cases in the data set reflects clinical practice and cautions on the over-interpretation of rare variants. Unlike prior prediction challenges, where the activity of an enzyme had been quantitatively measured in the laboratory, there was no definitive answer key for this challenge. The predictors were asked to match sequencing data to a phenotype, and many groups did so by first identifying a causative variant. Only in a minority of cases (~23% in this dataset) could it be said with high confidence that a variant was likely contributing to disease in a patient. When a clinical laboratory reports a variant as Pathogenic, this is often

because the variant has previously been reported in patients with the same phenotype or the nucleotide change introduces a premature termination codon in a gene where loss-of-function variants cause disease [144]. Thus, with a foundation in clinical genetics and access to online resources one could identify a large proportion of the 'Pathogenic' variants in this dataset. However, many of the variants detected in the clinical laboratory are rare missense or synonymous variants that have not previously been reported in the literature; these are almost always classified as variants of uncertain clinical significance. It is for these variants of uncertain significance, that are difficult to interpret and for which there is no answer key, that better assessment tools are needed. A CAGI challenge focused on the interpretation of variants of uncertain clinical significance would be more relevant to current clinical genetics practice. A clinical lab may upgrade a variant's classification from 'Uncertain' to 'Pathogenic' based on new clinical information, segregation of a variant within a family, or identification of the variant in multiple unrelated individuals. Many molecular diagnostic labs maintain internal variant databases; such databases could be mined to curate a challenge set of 'Uncertain' variants for which there is unpublished data to support pathogenicity. In this proposed challenge, participants would have to correctly identify these 'Pathogenic' variants from a set of 'Uncertain' variants (for which there was unpublished data that they were NOT likely to contribute to disease). This would more directly test the challengers' ability to predict pathogenicity without relying on allele frequency or online databases and without requiring knowledge of gene-disease associations. Assessment of the challenge would benefit from having fully vetted data and a clear answer key. This type of challenge, while still lacking a phenotype component, would more accurately mirror the clinical challenge of interpreting rare variants. Obtaining this data set would also invite communication between clinical testing labs (both academic and commercial) and the research community. In this vein, the development of a clinically useful variant assessment tool will require collaboration between clinical geneticists and data scientists. Discussions resulting from the Hopkins Clinical challenge demonstrated that although most participants incorporated genetic principles into their pipelines, they approached variant interpretation in a very different manner than a clinical laboratory. In future challenges, it would be interesting to pair an informatics group with a clinical group as a challenge team, particularly for whole exome sequencing challenges. Ideally, the back-and-forth between clinical and informatics groups would produce a method that could outperform that of either group alone. Diverse collaborations at CAGI could help bridge the communication gap between fields and pave the way for development of better tools.

5 Design of a diagnostic gene-panel for the diagnosis of neurodevelopmental disorders

5.1 Summary

Neurodevelopmental disorders (NDDs) are common genetic conditions including clinically heterogeneous phenotypes, such as intellectual disability (ID) and autism spectrum disorder (ASD)[166]. Due to a wide genetic heterogeneity and recurrent overlapping clinical features, single-gene testing for diagnosis of NDDs is especially challenging[21,22]. As consequence, high-throughput methods, as NGS targeted gene re-sequencing, are increasingly employed for NDD genetic testing[21,22]. This part of my project deals with the identification of a subset of genes involved in the ID/ASD co-morbidity to develop a new, cost effective, comprehensive gene panel for NDD diagnosis. Candidate ID and ASD gene lists were generated by gathering data from public databases, exome sequencing and meta-analysis studies. These lists were filtered, selecting only known causative genes from literature, top ranked by gene prioritization, and meeting network parameters. The final panel set resulted in a manually curated gene list (74 genes), used to design the diagnostic gene-panel, which is currently used for clinical screening of affected patients at the Molecular Genetics of Neurodevelopment Laboratory (Paediatric Department, University of Padova).

5.2 Introduction

Neurodevelopmental disorders are common highly heritable diseases, including autism spectrum disorder and intellectual disability [166,167]. ASD is the most severe manifestation among NDDs, with a high prevalence in population (1%, 3-4% combined with ID)[167]. Common features shared among ASD patients are the restrictive and stereotypic behaviors, difficulties in reciprocal social interactions and communication. These clinical conditions are mostly associated with psychomotor development delay, intellectual disability, and seizures[9,168,169]. ID is defined by a below average intellectual function (IQ <70), often associated with limitation in adaptive skills[167]. ID can arise alone or in paired to other clinical phenotypes, such as craniofacial dimorphisms, neurologic impairment, seizures and behavioral issues[167]. As consequence, both ASD and ID show a high level of comorbidity,

with the 70% of ASD patients presenting mild to severe ID, and at least 10% of ID affected individuals showing also autism features[167]. This phenomenon is linked to the impairment of closely related molecular mechanisms in both pathologies[9,169]. Indeed, ASD-ID associated genes converge on common pathways involved in synaptic development, plasticity and signaling in neurons of central nervous system[170]. Moreover, alterations in genes belonging to shared pathways (e.g. *SHANK2*, *NRXN1*, and *CNTNAP2)* are associated with ASD as well as ID[167], while more than 1,000 causative loci were identified so far[21]. Pathogenic mutations in the known ASD-ID genes are both *de novo* and germline rare variants, which comprise chromosome abnormalities, copy-number variation (CNV) and single-nucleotide variation (SNV)[9]. In addition to the high genetic heterogeneity, the causative variants present a very low prevalence in their respective patient populations, posing even more challenges to diagnostic use of DNA testing[167]. As consequence, high-throughput methods, as targeted gene re-sequencing, are increasingly employed for NDD genetic testing[21,22]. In this context, next generation sequencing (NGS) analysis of selected gene panels holds many advantages. Firstly, this approach allows of screening dozens of genes with high coverage in a single experiment pairing with a considerably moderate sequencing costs. Secondly, the restricted targeting decreases the incidental finding occurrence, allowing to identify rare variants with large effect sizes in an enriched target space[3,171]. For these reasons, we worked toward the development of a novel comprehensive, and cost effective, gene panel for ASD-ID diagnosis. The gene panel was applied to identify rare *de novo* SNVs that may be causative for sporadic and non-syndromic ASD-ID cases.

5.3 Methods

5.3.1 Patient cohort

The patients were referred from the clinical geneticists of seventeen Italian public hospitals with a diagnosis of non-specific neurodevelopmental disorder. The enrolled patients have negative high-resolution karyotype or array-CGH, Fragile-X test and metabolic screening. Clinical data were collected with a standardized record describing the clinical and family history, the disease phenotype (i.e. auxological parameters, neurological development, physical features, behavioral profile) and the presence of associated disorders. Data from the neurophysiological profile (i.e. ictal and interictal video-EEG-polygraphy, during sleep and wakefulness, evoked potentials when appropriate) and MRI brain, were also collected.

Table S5.1 summarizes the clinical data of the patients. A written informed consent was obtained from all the patient's parents or legal representatives. This study was approved by the Local Ethic Committee of the University-Hospital of Padova

5.3.2 Gene selection

For the construction of an efficient and low-cost GP, we selected the most promising ID and/or ASD genes gathering data from public databases (AutismKB http://autismkb.cbi.pku.edu.cn [32], and SFARI https://sfari.org/resources/sfari-gene [33]), OMIM, and PubMed. In particular, candidate genes were extracted from the recent exome sequencing and meta-analysis studies (Table S5.2). We collected a list of 972 genes scored according to the recurrence in different sources, and annotated for the clinical phenotype, gene function, subcellular localization and interaction with other known causative genes. Separated lists were generated considering ASD or ID association. Using data from STRING 9.0 (https://string-db.org)[20], a disease protein-protein interaction (PPI) network was built starting from 66 high confidence genes (intersection list), shared both by ASD and ID gene lists. The emerging features of the network were assessed by an enrichment analysis with Enrichr webserver[37]. The intersection list was used as training set for the Endeavour gene prioritization (https://endeavour.esat.kuleuven.be/) [34]. Hub direct interactors (STRING score above 0.45) belonging to the top ranking prioritized list, but not included in the intersection, were also included in the most promising candidate genes list. The inclusion conditions are: i) either the gene introduction in the network allows to connect a one or more unconnected nodes (i.e. genes) to the core gene set, ii) or the gene is one of the connected nodes after linker introduction. To this list, we added also the top ranked ID or ASD associated genes only (genes with at least 5 evidences for ID or ASD). The final panel set resulted in a manually curated gene list (74 genes), comprising the selected known causative genes, top ranked by gene prioritization and meeting PPI network parameters.

5.3.3 Gene Panel Sequencing

The nucleic acids were extracted from patients' blood samples using Wizard genomic DNA Promega Kit (Promega Corporation). Multiplex, PCR-based primer panel were designed with Ion AmpliSeq™ Designer (Thermo Fisher Scientific) to amplify all the exons and the flanking regions (10 bp) of the 74 selected genes. The amplicon libraries were prepared

with Ion AmpliSeq Library kit v2.0 (Thermo Fisher Scientific) with a barcoding protocol. The template preparation and enrichment were performed with Ion One Touch 2 and Ion One Touch ES System (Thermo Fisher Scientific). The sequencing was performed on the Ion PGM System using the Ion PGM Sequencing Hi-Q Kit, and Ion 316 v2 and Ion 318 v2 BC chips (Thermo Fisher Scientific).

5.3.4 Variant ranking

Reads alignment to the human genome reference (hg19/GRCh37) and variant calling were performed with the Ion Torrent Suite Software v5.02 (Thermo Fisher Scientific), whereas SNVs were annotated using wANNOVAR[172]. Detected variants were ranked for their allelic frequency (AF) in the 146 patients and in control cohorts of 1000G [150], ExaC [173], and ESP6500 [174] databases. We excluded those SNVs found in more than two patients of our cohort and with AF higher than expected for the disorder[7]. SNVs with low frequency were filtered considering the consensus among eight computational methods (SIFT[55], Polyphen-2[52] HDIV and HVAR version, Mutation Taster [175], Mutation Assessor [53], FATHMM[176], MetaSVM [177], MetaLR [177], and conservation scores (CADD[67], GERP++ [65], PhyloP [178], SiPhy[179]). The selected SNVs were also evaluated *in silico* for their impact on splicing and on protein structure and function. The Integrated Genome Viewer (IGV) platform was used to exclude errors in the alignment process around selected SNVs.

5.3.5 *In silico* analysis of candidate variants

The canonical protein sequences were retrieved from UniProt[180], and the protein domain predicted by InterProScan[181]. The orthologous sequences were downloaded from OMA Browser[72] and aligned with MAFFT[73] to evaluate evolutionary conservation. When available, the crystal structures were retrieved from PDB[182]. Otherwise, the domain structures templates were searched using HHpred[86] and the structure folding predicted with MODELLER [86] (automatic best template selection). The disorder content and the presence of short linear motifs for protein interactions were assessed combining MobiDB 3.0[25], ELM[183] and using the interactive exploration tool ProViz[83].

5.3.6 Sanger sequencing validation

The candidate variants were validated by Sanger sequencing. When the DNA samples were available, patient relatives were sequenced to assess familiar segregation. The PCR

products were directly sequenced using the BigDye® Terminator version 3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA, USA). The reaction was run in an ABI Prism 3100XL automated sequencer (Applied Biosystems, Foster City, CA, USA) and the results were analyzed with Chromas 2.6.4 software (Technelysium Pty Ltd, Australia).

5.4 Results and discussion

5.4.1 ASD/ID shared genes define a core network, enriched for regulation of membrane excitability and synaptic trafficking

The filtering pipeline allowed us to select 66 genes shared by both the ASD and ID gene lists, and confidently linked to the comorbidity between the two pathologies. As mutations in disease associated genes generally yield perturbations in key cellular pathways[18], we decided to reconstruct the intersection gene network according to the STRING 9.0 protein interaction annotations (Figure 5.1). A network is defined by means of nodes (proteins encoded by selected genes) and edges (interactions), where the number of connection linking one node to another reflects the centrality, and thus the importance, of the gene/protein within the considered network[18]. STRING based analysis revealed that genes the intersection list forms a highly interconnected core set of 30 nodes, a small group of *DMD*-related genes (referred hereafter as "quartet"), and two pairs, i.e. *TANC2-KIRREL2* and *CC2D1A-DEAF1* (Figure 5.1). The core gene set comprises multiple genes involved in neuronal membrane excitability regulation (e.g. *SCN2A*[184], *GRI2B*[185], and *SLC6A1*[186]), which is a crucial factor in synapse formation during neurodevelopment[187]. Conversely, the "quartet" is characterized by genes related to the muscle functionality, such as *DMD* and *DAG1*. The *DMD* gene encodes for the dystrophin, a protein that interacts with many glycoproteins, e.g. the Dystrophin-Associated Glycoprotein 1 (DAG1). The dystrophin-glycoprotein complex links cytoskeletal components to extracellular matrix, regulating cellular stability[188]. This mechanisms is prevalently studied in muscle fibers, as mutations in *DMD* gene are generally associated with the Duchenne muscular dystrophy onset[188]. However, increasing lines of evidence highlight how mutations in this gene can be related to developmental cognitive and behavioral abnormalities, either in presence or in absence of neuromuscular symptoms[189–192]. As regard the pairs, the role in ASD-ID comorbity of *CC2D1A* and *DEAF1* mutations is well established[193,194], whereas *KIRREL3* is mainly

associated with intellectual disability[195], and the *TANC2* involvement in NDDs is mainly inferred from its functions in synapse modulation[196,197].

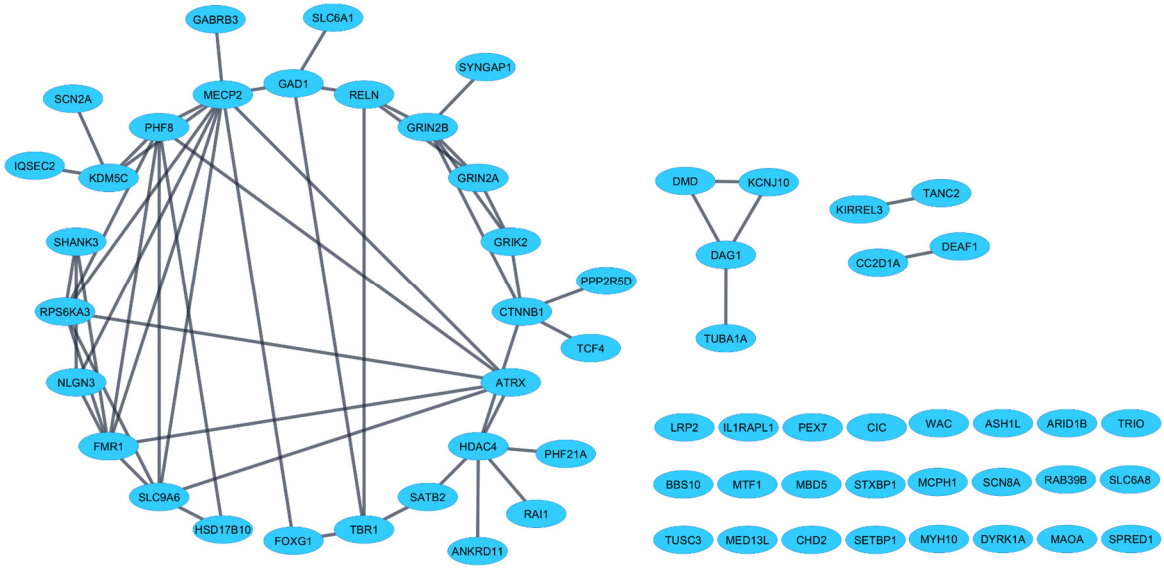


Figure 5.1: Network analysis results for intersection gene list.
Graphical representation of connectivity among high confidence intersection list. Interactions are represented with grey lines between nodes/genes (light blue).

In the latter scenario, the integration of the gene ontology terms and phenotypic features could be useful to provide a more general context to the functional consequences of gene/protein alterations[16]. Thus, the intersection list was used for the enrichment analysis, aimed to better characterize the affected molecular pathway, and the phenotypes resulting by mutations in our genes. Unsurprising, our gene set is enriched in terms related to regulation of synaptic plasticity and neuronal transmission, as well as to post-synaptic localization (Table S5.3). More interestingly, HPO term enrichment shows that our genes are both likely linked to autism typical features, e.g. stereotypic, aggressive and abnormal social behavior, and to intellectual disability[167]. Moreover, the clinical features, like the tented upper lip vermilion and the deep set eyes, are characteristic of pathologies where autism and intellectual disability are comorbid (e.g. the Pitt-Hopkins syndrome[198]), further supporting the effectiveness of this gene selection method.

5.4.2 Network expansion

Once the clinical impact of our high confidence gene list was established, we decided to expand the protein network, taking into consideration the previously excluded top ranked ASD and ID genes. These genes were prioritized with Endeavour[34]. The software integrates many other sources (e.g. gene expression, functional annotation and regulatory information), allowing to rank the genes by a comprehensive score based on the training set, i.e. the intersection list[34]. A perturbation-response inspired method was employed to determine the inclusion of prioritized genes in the candidate gene panel (GP) list. Briefly, we considered the protein network built on the intersection list and how the introduction of the prioritized gene affects network connectivity. Genes were introduced in the network following the Endeavour ranking order. If the number of connections among intersection genes increased after introduction, both the prioritized gene (linker) and connected gene are added to the final GP list (Figure 5.2).

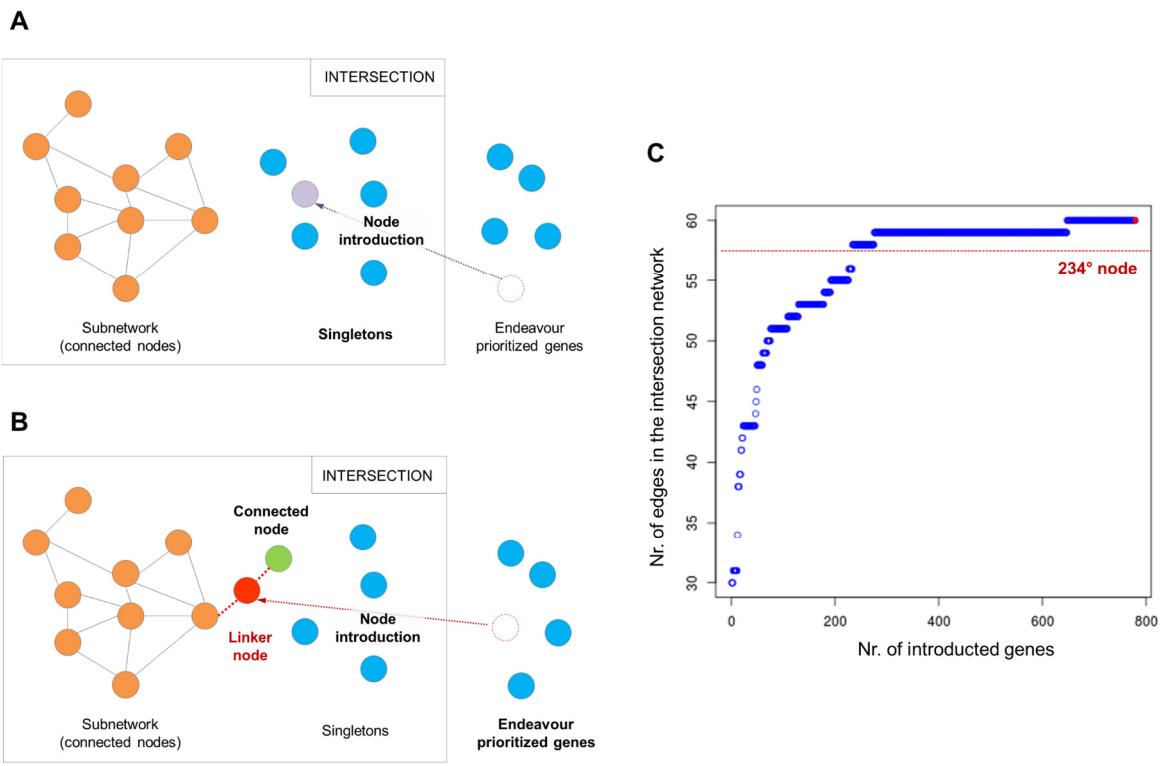


Figure 5.2: Gene introduction approach for intersection network expansion.
Genes we introducted according Endeavour prioritization ranking order. The introduction of a new node (gene/protein) does not affect protein connectivity **(A)**, or can increase the number of edges among intersection genes/proteins **(B)**. **C)** Effects of the introduction of prioritized genes on network connectivity. Notably, after introducing the 234th gene, the number of interactions reaches plateu.

Interestingly, we prioritized 906 genes, but after the introduction of the 234th gene, the number of interactions reaches the plateu, indicating that the other genes are not relevant for our disease protein network. This analysis allowed us to add seven genes, two linkers (*CREBBP*, and *CASK*), and four connected nodes (*MED13L, KIRREL3, ASH1L* and *MIB1*) to the candidate list, which were also mantained in the final GP. Except for the *MIB1* encoded protein (an E3 ubiquitin-protein ligase), the linker/connected genes are either scaffold proteins or chromatin modifiers, highlighting the important role of these two protein classes in synapse functioning (Table S5.4).

5.4.3 The ASD/ID gene panel screening results

After the network expansion, the candidate encoding regions were manually curated, and the resulting list was used for the diagnostic panel design (Table 5.1). The final ASD/ID gene-panel (GP) contains 29 Fragile X Mental Retardation protein targets, 21 genes encoding postsynaptic proteins (including TANC2 gene) and 16 chromatin modification factors (Tables S5.4)

Gene	Chr. band	Associated syndrome (Phenotype MIM)	Inheritance pattern
ADNP	20q13.13	Helsmoortel-Van Der Aa Syndrome	AD
ANKRD11	16q24.3	Kbg Syndrome	AD
AP1S2	Xp22.2	_	XL
ARFGEF2	20q13.13	Periventricular Heterotopia With Microcephaly, Autosomal Recessive	AR
ARID1B	6q25.3	Coffin-Siris Syndrome, Mental Retardation, Autosomal Dominant 12	AD/AR
ARX	Xp21.3	Corpus Callosum, Agenesis Of, With Abnormal Genitalia, Lissencephaly, X-Linked, 2, Mental Retardation, X-Linked, With Or Without Seizures, Arx-Related, Epileptic Encephalopathy, Early Infantile 1, Partington X-Linked Mental Retardation Syndrome	XLR
ASH1L	1q22		AD
ATRX	Xq21.1	Alpha-Thalassemia Myelodysplasia Syndrome, Alpha-Thalassemia/Mental Retardation Syndrome, Mental Retardation-Hypotonic Facies Syndrome	XLR
CASK	Xp11.4	Fg Syndrome, Mental Retardation And Microcephaly With Pontine And Cerebellar Hypoplasia	XLD
CC2D1A	19p13.12	Mental Retardation, Autosomal Recessive 3	AR

(Continues)

Table 5.1 (Continued)

Gene	Chr. band	Associated syndrome (Phenotype MIM)	Inheritance pattern
CDKL5	Xp22.13	Epileptic Encephalopathy, Early Infantile, 2; Eiee2	XLD
CHD8	14q11.2		AD
CNTNAP2	7q35-q36	Pitt-Hopkins-Like Syndrome 1	AD
CREBBP	16p13.3	Rubinstein-Taybi Syndrome 1	AD
CTNNB1	3p22.1	Mental Retardation, Autosomal Dominant 19	AD
DEAF1	11p15.5	Mental Retardation, Autosomal Dominant 24	AD
DYRK1A	21q22.13	Mental Retardation, Autosomal Dominant 7	AD
EHMT1	9q34.3	Kleefstra Syndrome	AD
FMR1	Xq27.3	Fragile X Tremor/Ataxia Syndrome, Fragile X Syndrome , Premature Ovarian Failure 1	XLD
FOXG1	14q12	Rett Syndrome	AD
FOXP1	3p13	Mental Retardation With Language Impairment And With Or Without Autistic Features	AD
GABRB3	15q12	_	AD
GAD1	2q31.1	Cerebral Palsy, Spastic Quadriplegic	AR
GATAD2B	1q21.3	Mental Retardation, Autosomal Dominant 18	AD
GRIA3	Xq25	Mental Retardation, X-Linked	XLR
GRIK2	6q16.3	Mental Retardation, Autosomal Recessive 6	AR
GRIN2A	16p13.2	Epilepsy, Focal, With Speech Disorder And With Or Without Mental Retardation; Fesd	AD
GRIN2B	12p13.1	Mental Retardation, Autosomal Dominant 6, With Or Without Seizures , Epileptic Encephalopathy, Early Infantile	AD
HDAC4	2q37.3	_	AD
IL1RAPL1	Xp21.2-Xp21.3	Mental Retardation, X-Linked 21	XLR
IQSEC2	Xp11.22	Mental Retardation, X-Linked 1	XLD
KATNAL2	18q21.1	_	AD
KDM5C	Xp11.22	Mental Retardation, X-Linked, Syndromic, Claes-Jensen Type	XLR
KIRREL3	11q24.2	Mental Retardation, Autosomal Dominant 4	AD
MBD5	2q23.1	Mental Retardation, Autosomal Dominant 1	AD
MCPH1	8p23	Microcephaly 1, Primary, Autosomal Recessive	AR
MECP2	Xq28	Mental Retardation, X-Linked, Syndromic 13, Lubs X-Linked Mental Retardation Syndrome, Encephalopathy, Neonatal Severe, Rett Syndrome	XLR
MED12	11q24.2	Ohdo Syndrome, X-Linked , Opitz-Kaveggia Syndrome, Lujan-Fryns Syndrome	AD
MED13L	12q24.21	Mental Retardation And Distinctive Facial Features With Or Without Cardiac Defects	AD

(Continues)

Table 5.1 (Continued)

Gene	Chr. band	Associated syndrome (Phenotype MIM)	Inheritance pattern
MEF2C	5q14.3	Mental Retardation, Autosomal Dominant 20	AD
MIB1	18q11.2	Left Ventricular Noncompaction 7	AD
MTF1	1p33	_	AD
MYH10	17p13.1	_	XL
NLGN3	Xq13.1	_	XL
NLGN4X	Xp22.32-p22.31	Mental Retardation, Usceptibility To Autism, X-Linked	XL
NRXN1	2p16.3	Pitt-Hopkins-Like Syndrome 2	AR
NTNG1	1p13.3	_	AD
OPHN1	Xq12	Mental Retardation, X-Linked, With Cerebellar Hypoplasia And Distinctive Facial Appearance	XLR
PHF21A	11p11.2	_	AD
PHF8	Xp11.22	Siderius X-Linked Mental Retardation Syndrome	XLR
PPP2R5D	6p21.1	Mental Retardation, Autosomal Dominant 35	AD
PQBP1	Xp11.23	Renpenning Syndrome 1	XLR
PTCHD1	Xp22.11	_	XLR
PTEN	10q23.31	Bannayan-Riley-Ruvalcaba Syndrome, Cowden Syndrome 1, Macrocephaly/Autism Syndrome	AD
PTPN4	2q14.2	_	AD
RAB39B	Xq28	_	XLR
RAI1	17p11.2	Smith-Magenis Syndrome	AD
RELN	7q22	Lissencephaly 2, Epilepsy, Familial Temporal Lobe, 1	AD/AR
RPS6KA3	Xp22.12	Mental Retardation, X-Linked 19, Coffin-Lowry Syndrome	XLD
SATB2	2q33.1	Glass Syndrome	AD
SCN2A	2q24.3	Seizures, Benign Familial Infantile, 3, Epileptic Encephalopathy, Early Infantile, 11	AD
SETBP1	18q12.3	Schinzel-Giedion Midface Retraction Syndrome , Mental Retardation, Autosomal Dominant 29	AD
SHANK2	11q13.3-q13.4	_	AD
SHANK3	22q13.33	Phelan-Mcdermid Syndrome	AD
SLC6A1	3p25.3	Myoclonic-Atonic Epilepsy	AD
SLC9A6	Xq26.3	Mental Retardation, X-Linked, Syndromic, Christianson Type	XLD
SYNGAP1	6p21.32	Mental Retardation, Autosomal Dominant 5	AD
TANC2	17q23.2	_	AD
TBR1	2q24	_	AD
TCF4	18q21.2	Pitt-Hopkins Syndrome	AD
TRIO	5p15.2	_	AD

(Continues)

Table 5.1 (Continued)

Gene	Chr. band	Associated syndrome (Phenotype MIM)	Inheritance pattern
TUSC3	8p22	Mental Retardation, Autosomal Recessive 7	AR
UBE3A	15q11.2	Angelman Syndrome	AD
WAC	10p12.1-p11.2	Desanto-Shinawi Syndrome	AD

Table 5.1: ASD/ID gene panel list.
For each gene the chromosome location, the related OMIM syndrome, and inheritance pattern (AD = autosomal dominant, AR = autosomal recessive; XLD = X-linked dominant; XLR = X-linked recessive) are reported.

The panel was employed in clinical screening of 146 individuals referred to the Molecular Genetics of Neurodevelopment Laboratory. An average of 94,8% of the target regions achieved a read depth of 20X and a mean depth of coverage of 263X for each individual. Overall, we detected in forty-seven of the subjects analyzed (47/146) fifty-seven rare single nucleotide variants, not yet identified or with low frequency in public databases. Sixteen of these rare SNVs are likely disrupting variants, since they are predicted to code for truncated proteins (14/57) or to affect transcript splicing (2/57). The other rare SNVs are missense (thirty-nine). Eleven missense variants are classified as possibly causative, since they are predicted to be deleterious by several computational tools, located in conserved positions critical for the protein function and supported by segregation analysis. Among the identified candidate disease-causing variants (Table 5.2), four were X-linked mutations (in *CASK, MECP2* and *RAB39B*), nineteen involve autosomal dominant genes (*KATNAL2, ANKRD11, ARID1B, DYRK1A, EHMT1, FOXP1, GRIN2B, SATB2, SETBP1, SHANK3, SLC6A1, SYNGAP1*, and *TRIO*), and five were detected in autosomal recessive genes (*CC2D1A, GAD1, GRIK2* and *TUSC3*). However, we identified the second possible causative mutation, a splice site variant, only in the case of *GAD1* gene. In four cases, the pathogenic variants were already described in literature in patients with a phenotype consistent with the clinical conditions of our probands (*GRIN2A* c.G2087A[199], *MECP2* c.C502T[200], *MECP2* c.2194_2198del[201] and *SATB2* c.C715T[202]). Familial segregation analysis has been performed in the 47,9 % (70/146) of patients carrying rare SNVs, and in three cases allowed to support the pathogenicity of the variant. The remaining twenty-eight missense SNVs were classified likely pathogenic, as segregation analysis is still ongoing, and thus, cannot support their role in disease pathogenesis (see Table S5.5).

Patient	Sex	Gene	I.P.	Variant	cDNA position	AA change	Variant segregation	dbSNP	D/T
602_01	F	CASK	XLD	chrX:41401980:G:A	c.C2119T	p.Q707X	*de novo*	_	stop codon
1974_01	M	RAB39B	XLR	chrX:154490151:A:C	c.T579G	p.F193L	maternal, affected	rs782042596	2/9
1975_01	M	KATNAL2	AD	chr18:44595922:C:T	c.C743T	p.A248V	maternal	rs140833601	9/9
984_01	M	TRIO	AD	chr5:14390392:C:T	c.C4111T	p.H1371Y	*de novo*	_	5/9
1985_01	**M**	**SATB2**	**AD**	**chr2:200213882:G:A**	**c.C715T**	**p.R239X**	***de novo***	**rs137853127**	**stop codon**
2033_01	M	ANKRD11	AD	chr16:89345974:CCTTCGGGG:C	c.6968_6975del	p.A2323fs	*de novo*	_	stop codon
1970_01	F	SHANK3	AD	chr22:51159830:A:TTC	c.3548insTTC	p.D1183Vfs5X	*de novo*	_	stop codon
2127_01	F	ARID1B	AD	chr6:157528165:G:T	c.G5851T	p.E1951X	*de novo*	_	stop codon
2145_01	**F**	**MECP2**	**XLD**	**chrX: 153296399:C:T**	**c.880C>T**	**Arg294X**	***de novo***	**rs61751362**	**stop codon**
2165_01	M	TRIO	AD	chr5:14394159:C:T	c.C4231T	p.R1411X	maternal, affected	_	Stop codon
2166_01	M	EHMT1	AD	chr9:140728837:G:C	c.G3577C	p.G1193R	*de novo*	_	8/9
2019_01	**M**	**GRIN2B**	**AD**	**chr12:13724822:C:T**	**c.G2087A**	**p.R696H**	***de novo***	**_**	**6/9**
2222_01	M	DYRK1A	AD	chr21:38858777:G:C	c.G525C	p.K175N	*de novo*	_	5/9
2230_01	F	SHANK3	AD	chr22:51153476:G:A	c.2244+1G>A	_	*de novo*	_	splicing
2271_01	M	SHANK3	AD	chr22:51159718:C:T	c.C3436T	p.R1146X	*de novo*	_	stop codon
2243_01	M	EHMT1	AD	chr9:140657209:GA:G	c.1585delA	p.S529fs	_	_	stop codon
1749_01	M	SHANK3	AD	chr22:51160432:GA:G	c.4150delA	p.E1384Dfs35X	_	_	stop codon
2274_01	M	SETBP1	AD	chr18:42531499:AGAGC:-	c.2194_2198del	p.E734Afs18X	*de novo*	_	stop codon
2233_01	M	SYNGAP1	AD	chr6:33411228:C:T	c.C2899T	p.R967X	_	_	stop codon
2347_01	F	ANKRD11	AD	chr16:89346137:A:G	c.6812_6813del	p.P2271fs	_	_	stop codon
2368_01	M	SLC6A1	AD	chr3:11078549:G:A	c.G1697A	p.R566H	_	rs767142926	8/13
414_01	**F**	**MECP2**	**XLD**	**chrX:153296777:G:A**	**c.502C>T**	**p.R168X**	**_**	**rs61748421**	**stop codon**
1769_01	M	FOXP1	AD	chr3:71026867:A:C	c.T1355G	p.I452S	paternal	_	4/9
1769_01	M	CNTNAP2	AD	chr7:146829502:G:T	c.G1249T	p.D417Y	maternal	rs147815978	6/9
2053_01	F	GAD1	AR	chr2:171678594:T:C	c.83-3T>C		maternal	rs769951300	splicing
2053_01	F	GAD1	AR	chr2:171702114:C:T	c.C850T	p.L284F	paternal	rs780519382	6/9

Table 5.2: Likely causative variants detected in our cohort.

Variant annotation scheme: Patient identification code, patient sex, affected gene, inheritance pattern (I.P.), chromosomal position, cDNA and amino acid change, variant segregation analysis results, dbSNP code (if available), and the ratio among damaging prediction over the total (D/T). Variants already reported in literature are in bold.

5.4.4 In silico analysis of possible causative variants

Further evidence supporting the causality between missense SNVs and clinical phenotype was provided by the evaluation of the variant effects on protein function/structure. In the following section, two examples are reported, i.e. DYRK1A p. K175Q and EHMT1 p. G1193R.

DYRK1A p. K175Q

DYRK1A (dual specificity tyrosine-phosphorylation-regulated kinase 1A) is a protein kinase that plays a key role in neurogenesis, neuronal differentiation and proliferation and synaptic plasticity[203]. The encoding gene maps to the Down syndrome critical region of chromosome 21, and extra copy of *DYRK1A* accounts for the majority of Down syndrome clinical phenotype. However, heterozygous single-nucleotide variants resulting in *DYRK1A* gene haploinsufficiency have been demonstrated to be associated with ID and mental retardation[203]. DYRK1A p. K175Q substitution was detected in the patient 2222_01, who presented severe ID, microcephaly, absent speech, behavioral and movement abnormalities, and corpus callosum agenesis.

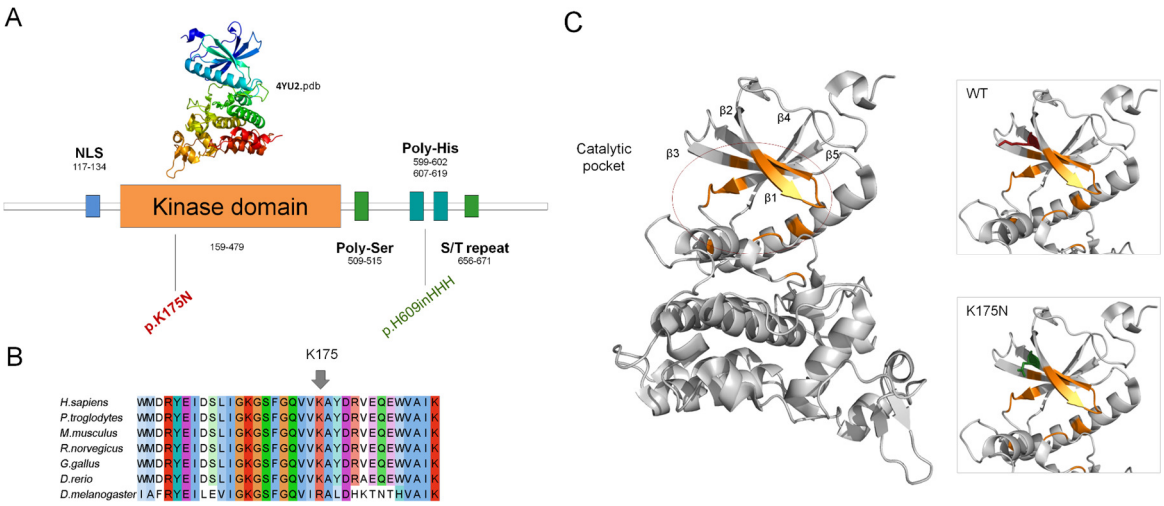


Figure 5.3: DYRK1A missense mutation affect catalytic pocket of kinase domain
A) Variants identified in our patient cohort are mapped to the DYRK1A domain architecture (red: possibly causative, green: likely benign). Protein sequence presents regions biased toward polar (serine and threonine) and aromatic (histidine residues). NLS = nuclear localization signal. **B)** DYRK1A p. Lys175 and neighboring residues are conserved among orthologous sequences. Amino acids are colored by conservation, according to ClustalX color code. **C)** DYRK1A p. K175N variant and wild type residues are mapped to kinase domain structure (4yu2.pdb, chain A). Residues involved in nucleotide binding are represented in orange, wild-type lysine in red and asparagine in green.

DYRK1A p.K175Q met all our filtering criteria: occurred *de novo*, was not listed in the SNV databases, e.g. dbSNP [204] and gnomAD[173], and was predicted being damaging by five out of nine variant prediction tools (see Table 5.2). Moreover, it maps to the catalytic region of kinase domain (β-2 strand), where the conserved Lys175 faces the nucleotide binding region in catalytic pocket[205] (Figure 5.3). The variant causes the substitution of a positively charged residue with an amidic bulkier asparagine. This could affect both the charge distribution and spatial constraints required for nucleotide binding, resulting in the alteration of catalytic pocket, and consequently, in the disruption of DYRK1A kinase activity, which could be connected with disease pathogenesis in our patient.

EHMT1 p. G1193R

Euchromatin histone methyltransferase 1 (EHMT1) catalyzes the mono- and dimethylation of histone H3 N-terminal lysine 9 (H3K9me2), which is associated with gene transcription repression, and with learning and memory processes[206,207]. Heterozygous *EHMT1* variants abrogating methyltransferase activity cause Kleefstra syndrome (KS), characterized by the comorbity of autistic-like traits and ID. Moreover, the KS clinical phenotype includes mild to severe developmental delay, hypotonia causing feeding difficulties, speech and motor delay [206]. In addition to being classified as damaging from the majority (8/9) of prediction tools (Table 5.2), EHMT1 p. G1193R substitution was detected *de novo* in a patient (2166_01) with psychomotor delay, mild ID, hypotonia and absent speech. The variant maps to a highly conserved glycine residue in the SET domain (Figure 5.4). The SET domain mediates the transfer of a methyl group from S-adenosyl-L-methionine (SAM) co-factor to the H3K9[208]. Variants mapping to the SET binding pocket would lead to an alteration of its activity, resulting in changes of H3K9me levels and gene transcription repression. The selected missense SNV leads to the substitution of a small glycine with a large, positively charged arginine in the SET domain (Figure 5.4). Besides the introduction of a bulkier residue, EHMT1 p.G1193R adds a positive charge within the H3K9 binding pocket, whose key interacting residues are prevalently non-polar or negatively-charged, with only two arginine accepted at positions 1214 and 1180[208]. Thus, the variants may affect the charge distribution and histone binding affinity within the interaction region, triggering variation in methyltransferase activity, and may explain the clinical phenotype in patient 2166_01.

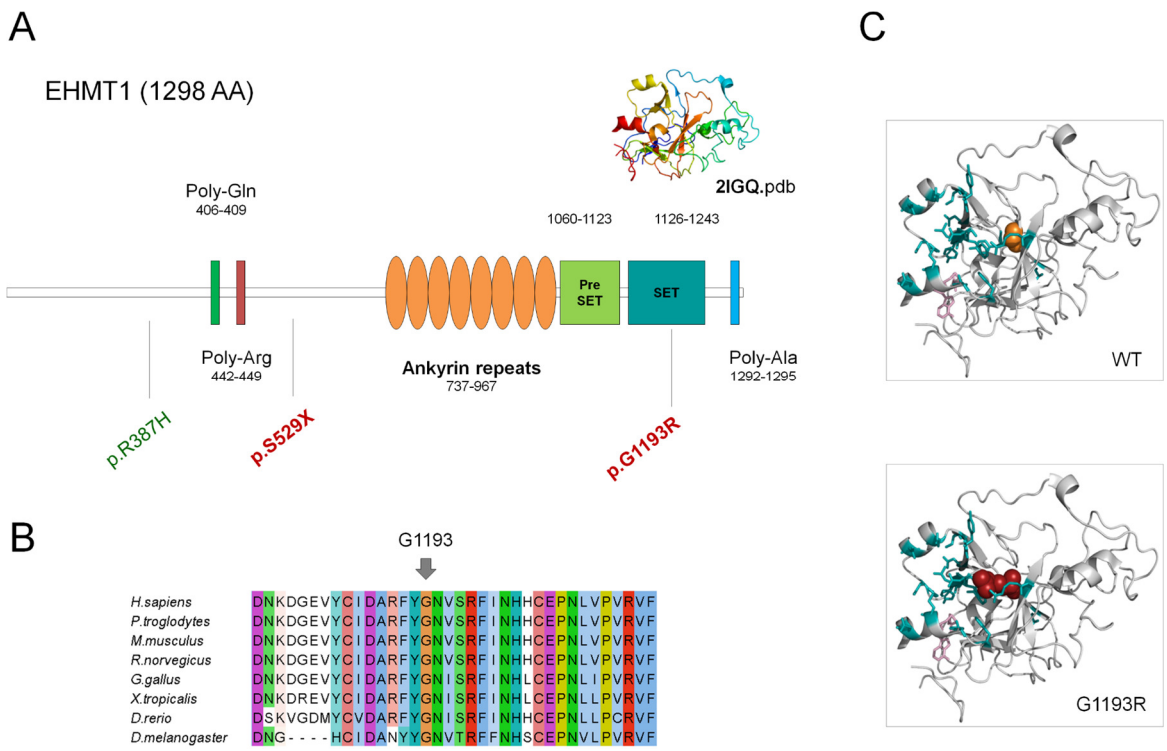


Figure 5.4: EHMT1 missense mutation affect SET domain.
Variants identified in our patient cohort are mapped to the EHMT1 domain architecture (red: possibly causative, green: likely benign). Protein sequence presents regions biased toward polar (glutamine and arginine) and a poly-alanine motif. The ankyrin domain (orange) is involved with the histone H3K9me binding. The Pre-SET domain (green) contributes to SET domain stabilization **B)** EHMT1 p. Lys1193 and neighboring residues are conserved among orthologous sequences. Amino acids are colored by conservation, according ClustalX color code. **C)** EHMT1 p. G1193R variant (red) and wild type glycine (orange) are mapped to SET domain structure (2igq.pdb, chain A) Residues involved in H3K9 binding and the S-adenosyl-L-methionine molecule are represented in sticks.

5.5 Conclusions

In this work, we develop a new and comprehensive gene panel for ASD/ID diagnosis comorbidity. It consists of 74 genes, including the previously single-tested genes, known ASD/ID associated genes and the best candidate genes from network and prioritization analysis. The linkage based network analysis, combined with the gene prioritization, allow us to filter the most promising candidate genes associated with ASD/ID comorbidity, which were added to the final list. Moreover, HPO and G.O. ontologies enrichment analyses provided additional evidence supporting the selected loci involvement in cellular processes and clinical conditions consistent with the ASD/ID comorbidity. The resulting panel was

employed in clinical screening of 146 individuals referred to the Molecular Genetics of Neurodevelopment Laboratory (Pediatrics' Department, University of Padova) for the genetic testing. The GP screening allow us to assign a molecular diagnosis to twenty-four of the screened patients, with a diagnostic yield of 16,4% (24/146). The selected variants met all the filtering criteria, i.e. they are *de novo* or low frequency SNVs, are classified as damaging by most of the prediction tools and follow the disease phenotype segregation in the family. In some cases, evidence supporting the variant pathogenicity was provided by bioinformatics analysis, by which the effect on related functional region was assessed. In the examples presented here, mutations affect structural domains that are critical for protein activity, e.g. the kinase catalytic domain in DYRK1A and the histone 3 binding pocket in EHMT1 protein, explaining the variant impact on patients' pathology. Besides diagnosed cases, at least one likely pathogenic variant was detected for additional twenty-three patients (15,7%), and ongoing investigation will assess their role in disease onset. These results confirm the diagnostic value of this targeted gene panel for investigating children affected by ID and ASD, both presenting co-occurring phenotype, as well as for their differential diagnosis.

6 DisProt 7.0: a major update of the database of disordered proteins

This Chapter has been published in "Piovesan D., Tabaro F., Mičetić I., Necci M., Quaglia F., Oldfield C.J., Aspromonte M.C., Davey N.E., Davidović R., Dosztányi Z., Elofsson A., Gasparini A., Hatos A., Kajava A.V., Kalmar L., Leonardi E., Lazar T., Macedo-Ribeiro S., Macossay-Castillo M., Meszaros A., Minervini G., Murvai N., Pujols J., Roche D.B., Salladini E., Schad E., Schramm A., Szabo B., Tantos A., Tonello F., Tsirigos K.D., Veljković N., Ventura S., Vranken W., Warholm P., Uversky V.N., Dunker A.K., Longhi S., Tompa P., Tosatto S.C.E. Nucleic Acids Res. 2017 Jan 4;45(D1):D219-D227."

6.1 Summary

The Database of Protein Disorder (DisProt, URL: www.disprot.org) has been significantly updated and upgraded since its last major renewal in 2007. The current release holds information on more than 800 entries of IDPs/IDRs, i.e. intrinsically disordered proteins or regions that exist and function without a well-defined three-dimensional structure. We have re-curated previous entries to purge DisProt from conflicting cases, and also upgraded the functional classification scheme to reflect continuous advance in the field in the past 10 years or so. We define IDPs as proteins that are disordered along their entire sequence, i.e. entirely lack structural elements, and IDRs as regions that are at least five consecutive residues without well-defined structure. We base our assessment of disorder strictly on experimental evidence, such as X-ray crystallography and nu-clear magnetic resonance (primary techniques) and a broad range of other experimental approaches (secondary techniques). Confident and ambiguous annotations are highlighted separately. DisProt 7.0 presents classified knowledge regarding the experimental characterization and functional annotations of IDPs/IDRs, and is intended to provide an invaluable resource for the research community for a better understanding structural disorder and for developing better computational tools for studying disordered proteins.

6.2 Introduction

Our traditional view of protein structure and function is deeply rooted in the structure–function paradigm which stated that the polypeptide chain of proteins needs to fold into a stable three-dimensional (3D) structure, which is a prerequisite of the functioning of the protein. The extreme explanatory power and success of this model is at-tested by more than hundred thousand high-resolution structures in the Protein Data Bank (PDB) [209] and many Nobel Prizes awarded for describing structures central to understanding important cell-biological phenomena. It has been suggested almost 20 years ago, however, that many proteins or regions of proteins in various proteomes lack such stable 3D structure, and are rather intrinsically disordered under native, physiological-like conditions (thus named IDPs/IDRs, respectively) [210–212]. The recognition of this structural phenomenon brought a radical change in the structure–function paradigm, and critically extended the general appreciation of the role of dynamics in protein function. It has been recognized that structural disorder, which is prevalent in all organisms, plays roles primarily in cellular signaling and regulation [213]. Because of that, IDPs/IDRs are often implicated in diseases [214] and represent important drug targets [215].

The structural and functional characterization of disordered proteins represents a special challenge, because they exist as an ensemble of rapidly interconverting conformations. Although they cannot be crystallized and thus cannot be directly characterized by X-ray crystallography, there are a variety of techniques that can report on their highly dynamic structural state at low- or even high spatial and temporal resolution [211]. The current best structural description of IDPs/IDRs is by structural ensembles, which can be solved by a combination of experimental and computational approaches and are collected into a dedicated structural database, PED [216].

Studies of the structure–function relationship of disordered proteins have shown that in certain cases their function arises directly from the disordered state (entropic chains), whereas in many other cases their function emanates from molecular recognition accompanied by induced folding to specific binding partners, such as another protein, RNA or DNA molecule [217,218]. In these functions, the sensitivity to regulated remodeling of the disordered structural ensemble is an excellent substrate for protein regulation [219], as exemplified by frequent post-translational modifications and special modes of allosteric regulation [220] involving IDPs/IDRs.

Due to the prevalence and importance of structural disorder, several dedicated databases covering various aspects of IDPs/IDRs have appeared in the past decade. DisProt is the primary repository of disorder-related data on sequence-and functional annotations, focusing on disordered proteins or regions with experimental verification [221,222]. Several other databases are based on predictions of disorder, such as D2P 2, which contains disorder protein predictions by a variety of predictors on 1765 complete proteomes [223], MobiDB, which features three levels of annotations, manually curated, indirect and predicted for all UniProt sequences (over 80 million) [224], and IDEAL, which contains manual annotations of interaction regions undergoing induced folding, sites of post-translational modifications and assignments of structural domains [225]. In addition, as already mentioned, PED is the database that gathers structural in-formation on IDPs/IDRs, in the form of structural ensembles [216]. The interaction of IDPs/IDRs with their target(s) is most often mediated by short continuous stretches of amino acids such as Molecular Recognition Elements/Features (MoREs/MoRFs) [226] and short/eukaryotic linear motifs (SLiMs/ELMs), which have been collected in the ELM database [227]. Less frequently, partner interactions of IDPs/IDRs may also be mediated by intrinsically dis-ordered domains (IDDs), i.e. longer regions that conform to the definition of domains as functional, evolutionary and structural units [228]. Although probably still underappreciated, some of these IDDs may be found in the Pfam database of protein families which includes their annotations and underlying multiple sequence alignments [87].

DisProt is central to all IDP-related research efforts, because it collects and presents in a structured way the core experimental evidence reported for structural disorder in proteins. To give a new impetus to the field, we have significantly updated and upgraded it with new features. This new release—DisProt 7.0—contains more than 800 entries of IDPs/IDRs. We have also re-defined and extended functional categories laying the basis for a functional ontology of IDPs, now encompassing 7 major classes and 35 sub-classes, all based on published experimental data.

6.2.1 Detection and characterization of IDPs

Technical advances in the field of biophysical and structural biology in the last 50 years have provided the scientific community with an arsenal of techniques to tackle the challenging characterization of IDPs/IDRs [212,229]. The various methods differ in their extent of sophistication, and hence in their technical demand, as well as in the nature of the

information they provide. Nuclear magnetic resonance (NMR) and X-ray crystallography provide site-specific information, whereas other methods provide more qualitative and global information (e.g. far-UV circular dichroism, size-exclusion chromatography; SEC). The rise of the field of protein disorder has greatly benefited from structural biology, because structures deposited in the PDB [209] have been instrumental for the development of disorder predictors, often trained on regions of missing electron density. Developments of multidimensional heteronuclear NMR also enabled the structural characterization of disordered proteins of increasing size [230,231]. In particular, heteronuclear single quantum coherence (HSQC) experiments are most commonly used to define protein disorder irrespective of whether residue-specific chemical shifts are available or not, as crowded HSQC spectra, characterized by a poor spread of resonances, are typical of IDPs/IDRs. The same feature of low spread of proton resonances is also apparent in one-dimensional proton-based NMR spectra, which offers the obvious advantage of not requiring isotopic labeling. Following assignment of the spectrum, quantitative estimations of disorder can be obtained through various NMR observables, such as chemical shifts, relaxation rates, residual dipolar couplings and resonance intensities in paramagnetic relaxation enhancement experiments. These data enable probing sequence-specific structural information in IDPs/IDRs. A particular strength of NMR is that it can be increasingly applied under truly in vivo conditions, in live cells [232]. Therefore, these two experimental approaches, X-ray crystallography and multidimensional NMR, are considered as the 'primary techniques' providing evidence for structural disorder on a per residue basis in DisProt.

It should not miss our attention, though, that due to the expenses of isotopic labeling in NMR and the high rate of failure in protein crystallization, it would be unreason-able to only rely on these two approaches to document protein disorder. Therefore, beyond X-ray crystallography and NMR, a plethora of alternative biochemical and bio-physical approaches (termed 'secondary techniques') pro-vide orthogonal information on protein disorder in DisProt [212,229]. The various approaches are of course not equivalent in terms of reliability, resolution and accuracy and suffer from specific drawbacks and limitations. Structural disorder is often based on far-UV CD spectroscopy, which is overall quite reliable, but does not enable discrimination between ordered and molten globular forms. Near-UV CD, beyond being able to unveil the lack of ordered structure, has the advantage of distinguishing between globular and molten globule forms. Another hallmark of disorder is anomalous sodium dodecyl sulphate-polyacrylamide gel electrophoresis migration,

where IDPs have a high apparent molecular mass. IDPs/IDRs also behave anomalously in SEC, light scattering (DLS, MALS), and in small-angle X-ray scattering in that they display hydrodynamic radii (RH) and radii of gyration (Rg) higher than expected, reflecting an extended conformation.

Fluorescence spectroscopy is another common method to assess disorder. Intrinsic fluorescence probing the chemical environment of tryptophan residues provides information about their solvent-accessibility, whereas thermal differential scanning fluorimetry—similar to differential scanning calorimetry—can highlight the lack of a cooperative thermal transition and hence absence of ordered structure. Fluorescence resonance energy transfer between external fluorophores can even generate information on distance distributions and help solve the structural ensemble of the IDP [233]. Hyper-sensitivity to proteolysis is also commonly used to map out disordered regions of proteins. Recently, native mass spectrometry exploiting nanoelectrospray ionization [234,235] and high-speed atomic force microscopy operating at the single-molecule level [236] have emerged as attractive alternatives to address structural disorder.

As a last statement, it is noteworthy that the higher the number of independent experimental lines supporting dis-order, the higher the reliability of the annotation. Further-more, multi-dimensional information may help realize that structural disorder is not a single homogeneous structural state along an order-disorder binary classification coordinate, it rather represents a continuum of states from the fully ordered to the fully disordered. Similarly, many examples of biological relevant disorder in fragments that are missing from the full length protein have been reported. Further-more, numerous functional examples of 'conditional disorder', i.e. instances where a disordered region functions by transitions to or from a folded state [237], or when disorder is only observed in a fraction of similar structures [238], lead to ambiguity and clearly points to the need for carrying out complementary experiments. In addition, an extreme case leading to conflicting results is represented by instances where a protein region, predicted to be ordered, is not defined in the electron density in one crystal structure while being ordered in another one (for an example see [239] and DisProt entry DP00133). Do these ambiguous regions represent a new class of disorder that escape detection using the currently available disorder predictors (thus setting the scene for their improvement), or *a contrario* are they the result of static disorder that arises from experimental conditions or domain wobbling? Combining information from a variety of sources may help clarify these cases and also improve meaningful

descriptions of IDPs as conformational ensembles [240,241], which may lead to future descriptions of the structure–function relationship of IDPs.

6.2.2 Database structure and implementation

Database records. The technology of DisProt has been up-dated and is now based on a document-oriented MongoDB database. Stored documents are of two types, 'protein' including general information about the protein and 'disordered region (DR)' including evidence of disorder from literature. Protein information is retrieved from UniProt and includes cleavage sites and chain/peptide boundaries for polyproteins and processed proteins. DisProt is sequence-centric and different isoforms correspond to different entries as in the previous version. Cleaved proteins are merged into a single entry as they are products of the same native sequence. DisProt accession numbers now follow a single format and all previous entries with a 'xxx' suffix were re-moved. DR records are evidence-centric, i.e. different documents are stored for different experiments even when related to the same region. Forcing a one-to-one paradigm allows to track annotation evidence type and the corresponding literature source unambiguously. DR records also include experimental evidence quality tags for ambiguous annotations. Sometimes experiments are carried out on engineered sequences or fragments which may prove ambiguous to generalize for the entire sequence (AMBSEQ). Moreover, disorder boundaries are occasionally not clear from the literature (AMBLIT) or experiments are performed under extremely non-physiological conditions (AMBEXP). The major improvement from previous versions is the manually cu-rated functional annotation of the regions. Whenever possible, curator-associated functions based on literature evidence are indicated by selecting terms from a new ontology built for describing disorder-related functional modes. If none of the current terms in the new ontology give a proper description of the functional mode, the curator may propose a new term to be added to the ontology. Acceptance of the new term will require approval by the IDP/IDR ontology committee.

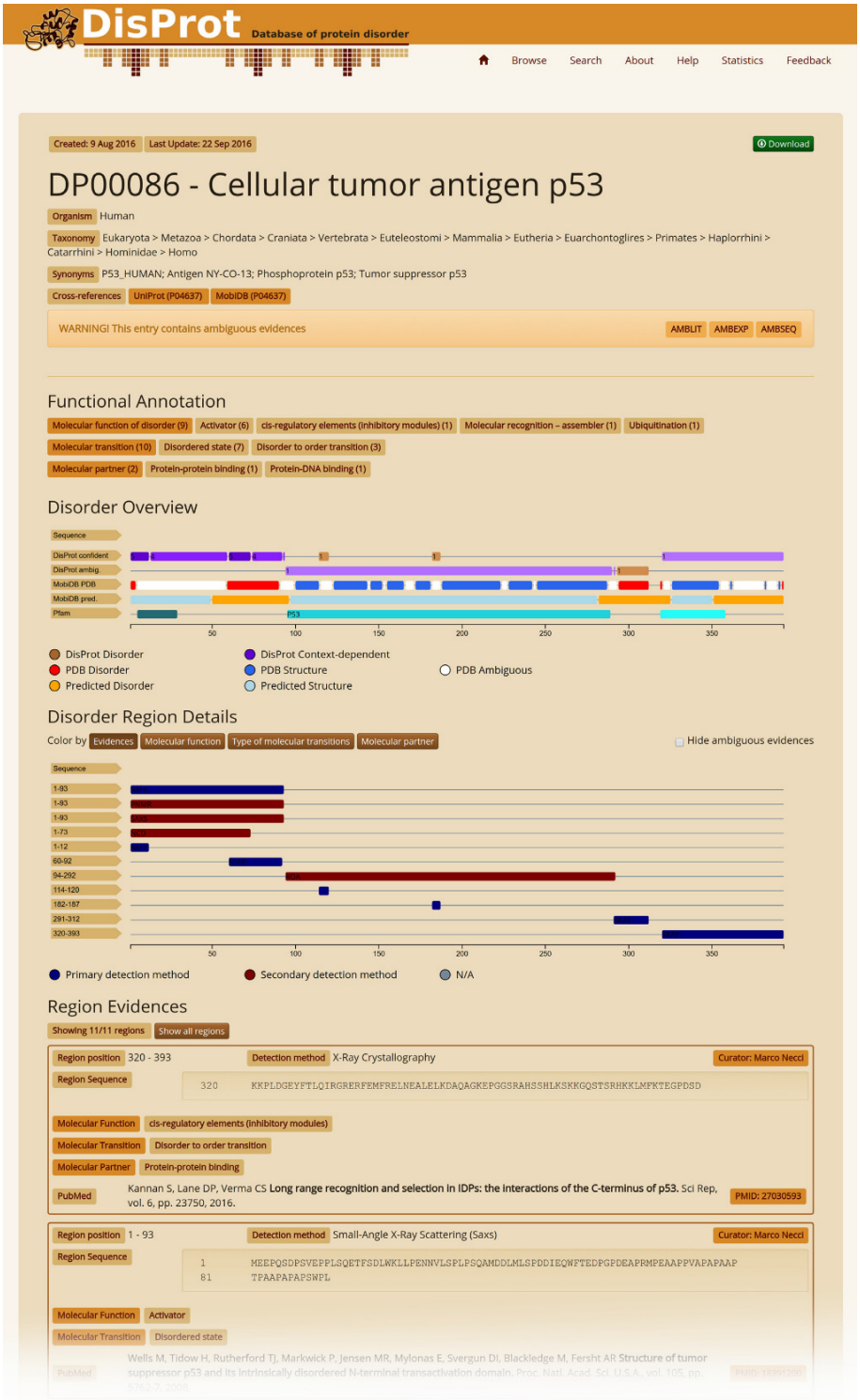


Figure 6.1: DisProt sample entry, human p53 protein (DP00086).

Several experiments have been carried out to characterize the human p53 protein. DisProt reports literature evidence for IDRs. In particular, 11 different IDR evidences (Region Evidences) have been collected from nine different papers by two different curators. Most of these are related to the N-terminus and come from different types of experiments (Disorder Region Details). Disorder regions and the number of DisProt evidences, separated into confident and ambiguous annotations, can be compared with structural information from the Pfam and MobiDB databases in the Disorder Overview. DisProt also provides function annotation of IDRs by reporting molecular function, transition and partner terms (Functional Annotation). A literature reference is provided for each annotated IDR, linked to the relevant PubMed entry.

Annotation pipeline. The new DisProt data have been generated by a community effort through a web server interface accessible upon registration. The same infrastructure can be used both to create and update entries. Curators provide an annotation through a submission form where all fields are validated on the client-side and a sequence viewer allows the comparison of assigned regions with structure information (Pfam domains, MobiDB disorder). Of note, the name of the curator is clearly visible in the entry to allow proper attribution of credit. The pipeline is fully automatic and can be potentially applied to the entire UniProt database. The DisProt public database is a snapshot of the community annotations.

Entry page. The entry page features four different sections (Figure 6.1). A protein information Table gives the protein name, gene, synonyms, identifiers, taxonomy and 'homologous' entries inferred from sequence similarity. An interactive feature viewer reports DisProt disorder regions separated into confident and ambiguous annotations, colored brown for intrinsically disordered regions and purple for context-dependent regions. Pfam domains along with PDB and predicted disorder derived from MobiDB are also shown. Below, a detailed feature viewer provides different visualization layers to highlight different functional aspects (ontology terms) and the strength of available disorder evidence. Each position in the sequence is colored according to the number and type of evidence. Last but not least, the full curator-generated list of region evidences is reported on the bottom of the page and can be filtered by selecting an element (region) in the feature viewer. Figure 6.1 shows the current DisProt annotation for the human p53 protein. The combination of DisProt and PDB annotation clearly shows how p53 contains several segments undergoing disorder to order transitions. Evidence for disorder from the literature in the central p53 DNA binding domain, for which many crystal structures are available in the PDB, is ambiguous and highlighted with AMBLIT. Similar conflicts can probably be found in scores of DisProt entries and demonstrate the importance of flagging ambiguous data.

Browsing and searching data. Both browsing and searching functionalities are provided in a single solution from the 'Browse' page. A sortable, customizable and filterable Table lists all entries by protein. Alternatively, another Table listing all regions is available and accessible through the 'regions' button. Complex queries can be simulated applying different filters to different columns. Specific entries can be selected manually and customized views can be generated by adding or removing columns. Filtered and/or selected data can be downloaded both in text and JSON formats. Alternatively, the 'Search'

page allows the user to search for specific words in a free-text form or to search for DisProt entries similar to a query sequence. Output for either search is a provided in a simplified form.

Feedback page. DisProt users are highly encouraged to suggest additional disorder annotations or changes to existing annotations using the 'Feedback' page. This contains a drop-down menu guiding the choice of feedback provided (e.g. website experience, novel annotations) and a message field. For feedback related to data entries, the user is asked to provide either the UniProt or DisProt ID and (where possible) a PubMed reference. All messages are reviewed by the curators and integrated in the database as time permits.

Web technology. The DisProt server is implemented in Node.js (https://nodejs.org) using the REST (Representational State Transfer) architecture. The data can be accessed through the web interface or programmatically exploiting the RESTful functionality. Please refer to the 'Help' section of the website for details on using the DisProt web services. The web interface is built using Angular.js (https://angularjs.org) and Bootstrap (http://getbootstrap.com) frameworks. The feature viewer is implemented on top of the Bio.js library.

6.2.3 Database content: upgrades and updates

Entries in DisProt 7.0 came from three major sources: (i) from the previous version of DisProt (where conflicting cases have been re-annotated), (ii) novel cases identified as PDB entries with long regions of missing electron density and (iii) proteins identified by text-mining in PubMed abstracts for keywords 'intrinsically disordered', 'intrinsically unstructured' and 'structural disorder'. New proteins selected based on disorder content (estimated based on MobiDB data) were prioritized (if appropriate information was available in SwissProt) to concentrate on well-studied and most interesting cases. New proteins were also selected by curators themselves to exploit their specific previous knowledge. All entries from previous versions were re-annotated to remove inconsistencies. One hundred and ninety-eight previous entries were completely removed and 469 modified. Recurring problems being fixed were wrong organism or isoform assignments, wrong IDR positioning, untracked disorder evidence (e.g. missing explicit literature reference) and weak evidence (e.g. based on very short fragments, please note that the minimal length of an IDR in DisProt 7.0 is 5 residues). Moreover, disorder annotations based on not traceable author/curator statements were discarded. Where

necessary, a curator comment now highlights criticisms relative to a given evidence/experiment, e.g. if the experiment has been carried out on an engineered protein. Regions annotated as structured in previous DisProt releases were removed (33 regions). Information related to experiments has been simplified by skipping technical details regarding experimental conditions. However, weak experimental evidence is filtered out by the curator during annotation and tagged with one of three ambiguous labels. Overall, DisProt 7.0 includes 804 entries and 2167 disordered regions, with a total of 92 432 amino acids with clear experimental and functional annotations (Table 6.1), and the length distribution of disordered regions has significantly changed from the last release of DisProt (Figure 6.2).

Table 6.1: DisProt annotation content

Method/function	Proteins	Regions	Residues
Nuclear magnetic resonance (NMR)	333	592	32 926
X-ray crystallography	326	683	20 742
Circular dichroism (CD) spectroscopy, far-UV	261	352	53 935
Sensitivity to proteolysis	75	95	13 961
Size exclusion/gel filtration chromatography	62	67	12 206
Proton-based NMR	53	69	7723
SDS-PAGE gel, aberrant mobility on	34	34	6326
Other methods	237	273	41 833
Disorder transition	564	1505	151 498
Molecular function	489	1199	106 670
Molecular partner	444	1108	119 665

Distribution of DisProt annotation based on experimental evidence (method) and disorder function (function). As each annotated disorder region corresponds to one piece of experimental evidence, multiple regions can map to the same sequence segment. If a protein is annotated multiple times with the same type of experiment it is counted once. The number of residues is the sum of region lengths.

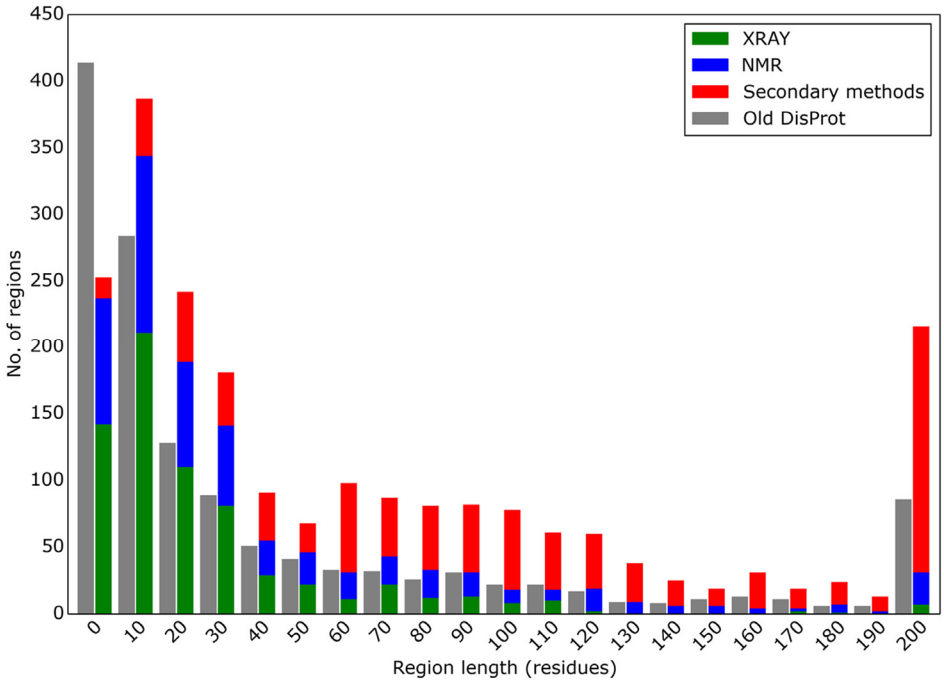


Figure 6.2: Distribution of disorder segment lengths.

Segment lengths are binned in groups of 10 residues, e.g. the column 10 showing lengths between 10 and 19 residues. The current DisProt release is distinguished by experimental technique (X-ray in green, NMR in blue and other methods in red). The previous DisProt release is shown in a single gray bar as it did not have the experimental technique in a machine-readable format.

6.2.4 New feature: functional classification

IDPs/IDRs carry out important functions in the cell. The field has settled on the notion that structural disorder rep-resents a continuum of states from fully folded to fully un-folded (random coil-like), and function may come from any of the states and transitions between them. That is, their function may come directly from the disordered state or from molecular recognition and binding to partner molecule(s). We derive our classification from the logic of the gene ontology classification scheme [242], which is based on three structured ontologies ascribing functional terms to gene products (proteins) in terms of their associated biological processes (BP), cellular components (CC) and molecular functions (MF). Apparently, the CC and BP ontologies do not depend on the disordered status of the protein, they simply reflect the intracellular location of the protein and the BP it participates in, which can be kept without reference for the disordered status [242]. The situation is entirely different with MF, which describes the elemental activities of a protein at the molecular level. In this regard, IDPs basically differ from folded proteins, such as enzymes or ligand-binding receptors, because their mode of action and type of function are usually completely different from those of folded proteins. Therefore, we have developed a novel classification

scheme that merges and expands previous schemes that suggested thirty [243] and six [217] different categories, to provide classified descriptors for their MFs. Because previous categories [217,243] lacked coherence (for example, they treated structural transitions and interaction partners at the same level), we created a rational scheme that distinguishes these different types of ontologies (cf. Table 6.2 and ref. [211]).The three sub-ontologies are as follows: (i) molecular function of disorder (MFUN): describes the type of functional readout of function (such as molecular chaperone); (ii) molecular transition (TRAN) necessary for function (such as disorder-to-order transition); and (iii) molecular partner (PART) that is recognized by the disordered protein (such as protein/RNA/DNA/small molecule). The MFUN ontology is described in detail in Table 1. The TRAN ontology can be further simplified to two IDR states (disorder and transition) to highlight different types of behavior, e.g. in the feature viewer of each DisProt entry.

Table 6.2: Major functional categories of the MFUN ontology of DisProt

MFUN code	Generic functional category	Functional category
MFUN_01	Entropic chain	Flexible linker/spacer
		Entropic bristle
		Entropic clock
		Entropic spring
		Structural mortar
		Self-transport through channel
MFUN_02	Molecular recognition: assembler	Assembler
		Localization (targeting)
		Localization (tethering)
		Prion (self-assembly, polymerization)
		Liquid-liquid phase separation demixing (self-assembly)
MFUN_03	Molecular recognition: scavenger	Neutralization of toxic molecules
		Metal binding/metal sponge
		Water storage
MFUN_04	Molecular recognition: effector	Inhibitor
		Disassembler
		Activator
		cis-regulatory elements (inhibitory modules)
		DNA bending
		DNA unwinding
MFUN_05	Molecular recognition: display site	Phosphorylation
		Acetylation
		Methylation
		Glycosylation
		Ubiquitination
		Fatty acylation (myristolation and palmitoylation)
		Limited proteolysis

(Continues)

Table 6.2 (Continued)

MFUN code	Generic functional category	Functional category
MFUN_06	Molecular recognition: chaperone	Protein detergent/solvate layer
		Space filling
		Entropic exclusion
		Entropy transfer

The functional schemes are an open hierarchy. One goal of sharing information with the community through DisProt is to refine our views of the functional modes of IDPs

6.3 Conclusions and future work

We have presented an updated and completely re-worked version of the DisProt database. It now features state-of-the-art database and web technology, enabling programmatic access of interested parties. The content was expanded by defining a standardized set of experimental techniques and a novel functional ontology of disordered segments. Both allow for a richer description of disorder which may be used for further analyses. The other main improvement in DisProt is a complete re-annotation of existing entries to remove inconsistencies and an expansion of ca. 50% over the previous release, which also resulted in a significant shift in the length coverage of disordered regions in the database. This advance was made possible by a distributed annotation effort coordinated by the COST Action NGP-net (URL: ngp-net.bio.unipd.it) involving a dozen different groups and close to 40 annotators. The longer term maintenance of DisProt is provided by the Italian node of the European bioinformatics infrastructure Elixir. In the future we hope that DisProt can be able to provide disorder annotations for UniProt.

Finally, we hope that the upgrade of DisProt will encourage the scientific community to deposit experimental evidence for disorder within this unique repository, and that this renewed momentum will lead to an increased awareness of the importance of intrinsic disorder in proteins

7 Dynamic scaffolds for neuronal signaling: in silico analysis of the TANC protein family

This Chapter has been published in "Gasparini A., Tosatto S.C.E, Murgia A., Leonardi E. Sci Rep. 2017 Jul 28;7(1):6829." For Supplementary Material, check the online version of the paper.

7.1 Summary

The emergence of genes implicated across multiple comorbid neurologic disorders allows to identify shared underlying molecular pathways. Recently, investigation of patients with diverse neurologic disorders found TANC1 and TANC2 as possible candidate disease genes. While the TANC proteins have been reported as postsynaptic scaffolds influencing synaptic spines and excitatory synapse strength, their molecular functions remain unknown. Here, we conducted a comprehensive *in silico* analysis of the TANC protein family to characterize their molecular role and understand possible neurobiological consequences of their disruption. The known Ankyrin and tetratricopeptide repeat (TPR) domains have been modeled. The newly predicted N-terminal ATPase domain may function as a regulated molecular switch for downstream signaling. Several putative conserved protein binding motifs allowed to extend the TANC interaction network. Interestingly, we highlighted connections with different signaling pathways converging to modulate neuronal activity. Beyond a known role for TANC family members in the glutamate receptor pathway, they seem linked to planar cell polarity signaling, Hippo pathway, and cilium assembly. This suggests an important role in neuron projection, extension and differentiation.

7.2 Introduction

Neurodevelopmental disorders (NDDs) are common conditions including clinically and genetically heterogeneous diseases, such as intellectual disability (ID), autism spectrum disorder (ASD), and epilepsy[244]. Advances in next generation sequencing have identified a large number of newly arising disease mutations which disrupt convergent molecular pathways involved in neuronal plasticity and synaptic strength[9,22,245–247]. In particular,

scaffold proteins seem to play a critical role in glutamatergic neurotransmission, organizing different components of glutamate receptor complexes at post-synaptic densities (PSDs) and determining synaptic strength and plasticity[248]. Among these, the TANC1 and TANC2 genes, encoding for the recently described scaffold proteins, are emerging as candidate genes for NDD[249,250]. TANC2 gene mutations were found in patients with different clinical conditions, ranging from ID and ASD to schizophrenia[245,251,252]. A case of *de novo* inversion encompassing TANC1, causing psychomotor-retardation was also recently reported in the literature[253]. TANC1 interacts with PSD95, one of the most important and well characterized scaffold proteins, as well as additional postsynaptic proteins including glutamate receptors[249,250]. The TANC proteins are expressed in the hippocampus and over-expression of either has been shown to increase dendritic spines and excitatory synapse strength in mice, although in vivo assays suggest differences in expression timing and knock-out phenotype. TANC1 reaches the highest levels in the adult brain and its depletion seems to impair spatial memory in mice. TANC2 is higher expressed during the early embryonic stages and seems to be involved in proper fetal development, with knock-outs causing in utero lethality[249]. Although available experimental evidence suggests an important role for these proteins in neuronal development, little is still known about the pathogenic mechanisms involved[249]. The TANC1 and TANC2 proteins were named on the basis of their domain architecture, predicted to contain tetratricopeptide (TPR) and ankyrin (ANK) repeats as well as a coiled-coil domain[250]. Furthermore, a P-loop ATPase domain was first observed at the N-terminus of the rolling pebbles orthologous of TANC proteins by Leipe and colleagues using sequence profile analysis and sequence based structure prediction to define the novel class of STAND (Signal Transducing ATPase with Numerous Domains) NTPase [254]. STAND proteins, unlike other NTPases, present a C-terminal helix bundle fused to the NTPase domain thought to transmit conformational changes due to NTP hydrolysis to downstream effector domains[254]. As an example, the closely related APAF1 protein is activated by the release of cytochrome *c*, which together with nucleotide binding, induces a conformational change in the P-loop ATPase driving apoptosome assembly[255]. Even though the nucleotide binding activity of the TANC P-loop domain and its functional role have to be demonstrated, this particular multi-domain architecture suggests at least a mechanistic similarity in molecular functions for TANC protein, combining a regulatory molecular switch with scaffold properties to assembly highly dynamic protein complexes. In this work, we employed a combined bioinformatics strategy, integrating sequence and phylogenetic analysis with *in silico* modeling of

structural domains to better characterize the structure-function relationship of the two TANC proteins. Furthermore, we conducted an in depth computational analysis to identify compositionally biased regions and candidate short linear motifs (SLiMs) in intrinsically disordered regions (IDR) of the proteins, which may provide further interaction surfaces mediating dynamic protein complex assembly. Experimental evidence for protein-protein interactions (PPI) in the literature or from protein-protein interaction (PPI) databases and predicted functional elements have been used to infer novel putative interactors for the two TANC members. Predicted and collected data highlight TANC involvement in orchestrating different neuronal signaling pathways, which may be implicated in the pathogenesis of diverse NDDs. This analysis suggests structural and functional elements that will help the interpretation of newly discovered TANC mutations. It would be worthwhile to follow up experimentally to support the hypothesis of a functional mechanism for TANC as a dynamically regulated scaffold.

7.3 Methods

7.3.1 Sequence feature analysis

TANC1 and TANC2 (UniProt accession codes: Q9C0D5 and Q9HCD6, respectively) were downloaded from UniProt[180], aligned using the MAFFT multiple sequence alignment software[256] and visualized with Jalview[257]. Secondary structure was predicted using PSIPRED[77], whereas domains, repeats and other features were predicted with InterproScan[76] COILS[258], MARCOIL[259] and CCHMM-PROF[260] were used to assess previously predicted coiled-coil regions, and TPR modules were predicted with TPRpred [261]. *A* repeat consensus was manually curated with Jalview from the MAFFT alignment. Further periodicities were searched with TRUST[262], RADAR[263] and Repetita[264]. Regions outside predicted domains, as well as N- and C-terminal protein sequences, were assessed for intrinsic disorder, presence of compositionally biased regions (i.e., repeating amino acids) and short linear motifs (SLiMs) using MobiDB[224] and ELM[81]. Since SLiMs have a high chance of random occurrence and their prediction often has low specificity, we selected for consideration only those mapping to disordered regions conserved among orthologous. Accessibility and localization in alternatively spliced regions are further evidences supporting the validation of the predicted SLiM[183].

7.3.2 Known TANC interactors analysis

A list of experimentally determined TANC interactors was compiled and manually annotated from the literature and the publicly available databases BioGrid[265], IntAct[266], and STRING[267] (see Table 7.1). Three significant interactions (false positive rate < 0.1) identified in the Cilium were also included[268]. Each interactor was annotated with its protein domain architecture and biological processes in which it is involved, retrieved from the InterPro[76], UniProt[180] and KEGG[269] databases. Furthermore, PubMed was searched for papers describing the involvement of TANC in neuronal development using selected keywords. Interaction details (i.e. residues, sequence motif and domain) were manually curated from the literature.

7.3.3 TANC interaction prediction

TANC interaction predictions were made either for the binding site of known interactions or to infer novel interactors. For each interactor we searched for the putative domain or linear motifs predicted to mediate TANC interaction. We assume that if the known interactor is a class of protein or presents the domain known to bind a predicted TANC linear motif this may be the putative interactor binding site. Collected PPI data and predicted binding sites/domains were used to infer novel putative interactors for the two TANC members. Proteins belonging to the same family usually interact in a similar way with a specific protein domain[270]. We assumed that when a protein has been found to interact with only one of the two TANC paralogs it is possible to infer it could interact with both paralogs as long as they share a common conserved SLiM predicted to mediate this interaction.

7.3.4 Mutation analysis

Pathogenicity of NDD associated variants in TANC proteins was assess using twelve different prediction tools: Align-GVGD[271], I-Mutant2.0[62], MUpro[63], MutationAssesor[53], MutationTaster[175], PhD-SNP[272], Polyphen2 [273], PROVEAN[274], SIFT [55], SNAP2 [275] and UMD-Predictor[276].

7.3.5 Phylogenetic analysis

The TANC orthologous were downloaded from OMA Browser [277] to reconstruct the phylogeny of the protein family. Eighty-one vertebrate sequences, representative of each

infrasubphyla, were retrieved. Taking into account teleost lineage-specific genome duplication[278], only one copy of each TANC protein was considered. The analysis comprised also earlier species in which duplication of TANC gene did not occurred: 2 arthropoda sequences (*Strigamia maritima* and *Ixodes scapularis*), 32 insect sequences, *Trichinella spiralis* (Nematoda), *Ciona intestinalis* (Tunicata). Multiple alignments were computed with ClustalO [279] and manually curated using Jalview. Phylogenetic analysis and visualization were performed with MEGA6[280], using Maximum Likelihood based on the JTT model + G (Gamma distributed Sites) with 500 bootstrap replicates.

7.3.6 Homology modeling

The predicted domains were modeled separately in order to build more reliable models. Sequences for TANC1 and TANC2 domains were submitted to the homology detection method HHpred[281]. Multiple sequence alignment-based template detection was performed with HHblits (local alignment) against pdb70, taking into account also target-template secondary structure similarity (for details see Supplementary Table S7.4). The resulting target-template alignments were manually curated using the repeat consensus map and consensus secondary structure prediction[282] in analogy to our previous work[94]. Two models for each domain were built by homology with Modeller [89] and their model quality was estimated with QMEAN[90]. The electrostatic surface of each model was calculated with Bluues[93] and Consurf[283] was used to map conservation for each residue based on OMA orthologous alignment. The structures were finally visualized using Pymol (DeLano Scientific LLC).

7.4 Results

Despite the emerging role of the TANC protein family in neuronal and embryonic development, little is known about their specific functions and molecular mechanisms[249]. A computational analysis of the TANC proteins starting from primary structure to explore the function of these twin proteins was thus performed. TANC1 and TANC2 are large proteins, of 1,861 and 1,990 residues respectively, sharing 51.9% overall amino acid identity, with similar multi-domain architecture (Table S7.1) resulting from an early duplication event (Supplementary Figure S7.1). InterproScan identifies two domains in both TANC protein sequences, an ankyrin (ANK) and tetratricopeptide repeat (TPR)

domain (Figure 7.1). An N-terminal P-loop containing nucleoside triphosphate hydrolase (NTPase) domain is predicted only in TANC2 and a likely a false negative for TANC1. The predicted domains are highly conserved among TANC paralogs. The N- and C-terminal disordered regions are quite variable. Crystal structures are not available for the TANC proteins, nor for any closely related proteins with similar domain architecture. To characterize the protein structure, each domain was modeled separately. The N- and C-terminal disordered regions were analyzed for the presence of a stretch rich in particular amino acid residues or conserved sequences containing predicted linear motifs likely to mediate protein interactions. Known interactors for both TANC proteins were downloaded from BioGrid[265], IntAct[266], and STRING[267]. Additional interactors were manually curated from the relevant literature (see Table 7.1). These findings were used to expand and curate a TANC protein interaction network (Table 7.2). While many interactors are in common between both TANC proteins, there are a two sets of proteins with experimental evidence for binding only one protein. In the following, we will describe each TANC region separately in more detail before using the predicted functional and structural elements to infer the possible impact of reported TANC2 missense mutations.

7.4.1 N-terminus

The N- and C- termini are intrinsically disordered and share rather low identity between TANC paralogs, suggesting functional divergence (Figure 7.1 and 7.2). A heterogeneous group of TANC2 sequences, comprising mammals, a bird and fish (*O. aries, M. putorius furo, M. domestica, F. albicollis, L. oculatus, O. niloticus*) defines the largest group sharing a 54 residue segment with TANC1 but no other TANC2 orthologous. A CK1 phosphorylation site and two SH3 binding motifs map to this sequence. This suggests that the N-terminal sequence was present in TANC1 first, duplicated in TANC2 orthologous and lost in other organisms, possibly to fine-tune the TANC2 interaction network. An alignment of TANC sequences highlights the presence of short conserved sequences, shared across all members, containing putative linear motifs (Figure 7.2). The N-terminus for instance presents two highly conserved motifs both involved in initiation of ubiquitin-dependent degradation and two protein phosphatase 1 (PP1) docking motifs (RVxF), almost identical in all considered sequences can be recognized (Figure 1 and Supplementary Figure S7.2). Shared linear motifs also comprise several post-translational modification sites recognized by different kinases, such as GSK3, MAPK, and NEK2.

These “hot spots” map in highly conserved serine-rich regions (SRR) (170-224). The TANC1 isoform Q9C0D5-2 is missing residues 122-227, which contains the conserved PP1 docking motif and SRR, suggesting a regulatory role for these regions (Figure 7.1).

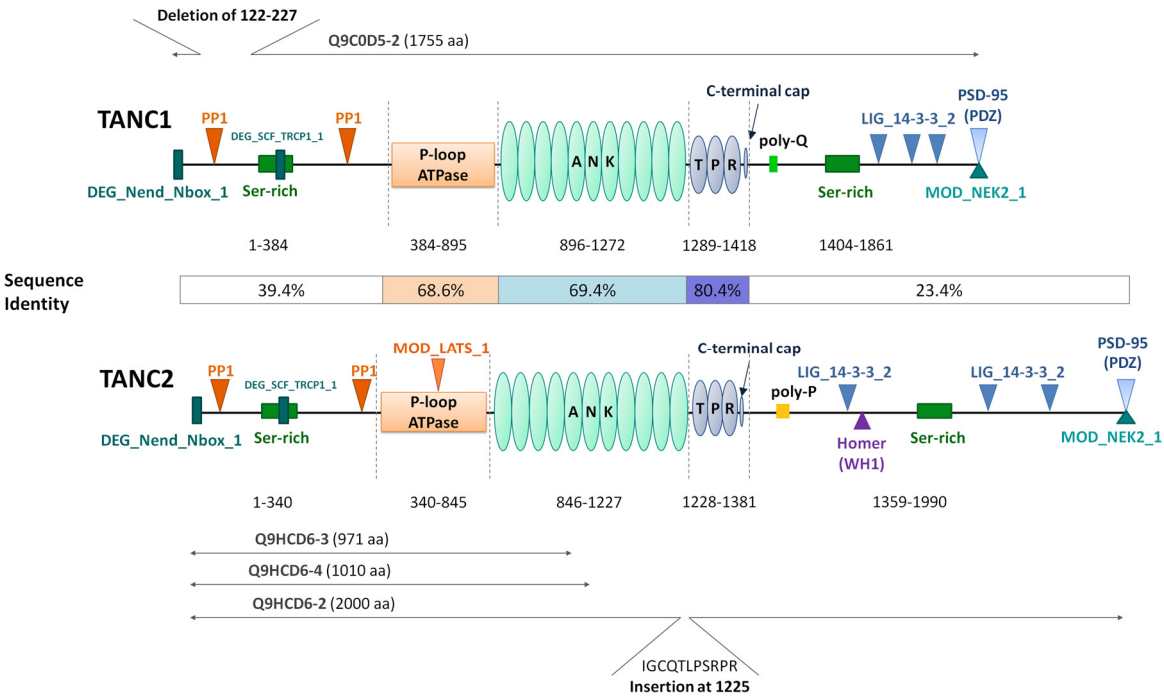


Figure 7.1: Sequence analysis of TANC proteins

An overview of TANC family domain architecture is here reported. Both TANC proteins are characterized by a putative P-loop NTPase domain (orange), an Ankyrin repeat containing domain (light teal) and a tetratricopeptide repeat region (blue). For each domain, the sequence boundaries and sequence identity between the two proteins are indicated. Conserved linear motifs are represented as follow: PDZ binding sequences (light blue triangles); PP1 docking motif (RVxF) (orange triangles); degrons (DEG_Nend_Nbox_1 and DEG_SCF_TRCP1_1) (deep teal rectangles); 14_3_3 binding sites (LIG_14-3-3_2) (blue triangles); TANC2 Homer binding motif (LIG_EVH1_1) (Purple triangle); LATS1 kinase (light orange triangle); NEK2 phosphorylation motif (MOD_NEK2_1) (Teal triangle). Serine-rich regions are represented with green rectangles (TANC1 residues 170-243 and 1659–1689; TANC2 residues 125-189 and 1775-1865). The TANC1 glutamine-rich region (Poly-Q) region and TANC2 glutamine/proline rich region (polyP) are in light green and yellow respectively. Alternative TANC protein isoforms are reported in grey. The TANC1 isoform Q9C0D5-2 (1755 residues) is missing the region 122-227. The TANC2 isoform Q9HCD6-2 is longer (2000 residues) due to an insertion at position 1225 (I > IGCQTLPSRPR). Q9HCD6-3 (971 residues) is truncated at residue 97 with different substitution in the region from position 944 to 971 (VDHLDKNGQCALVHAALRGHLEVVKFLI > VLAAQLCCFSSLFLYFRCILFLISSVTS). Q9HCD6-4 (1,010 residues) is truncated at residue 1011 with different substitution in the region from position 1006 to 1010 (IVSYL > VRSRQ).

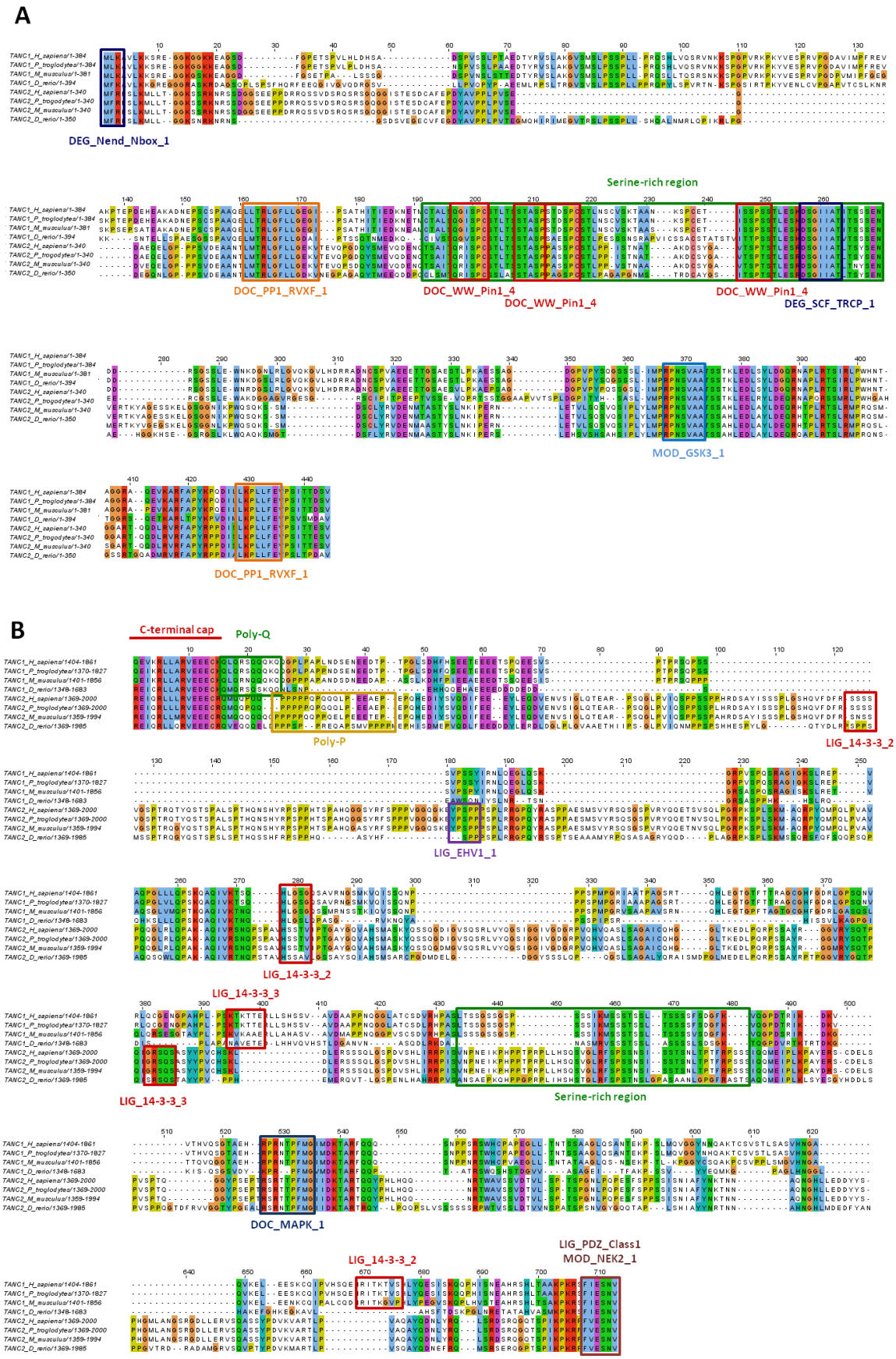


Figure 7.2: Multiple alignment of TANC N- and C- termini.

Color code based on ClustalX scheme. Linear motifs identified by ELM analysis are reported: DEG_Nend_Nbox_1: N-terminal motif that initiates protein degradation by binding to the N-box of N-

recognins; DOC_PP1_RVXF_1: Protein phosphatase 1 catalytic subunit (PP1c) interacting motif; DOC_WW_Pin1_4: IV WW domain interaction motif; MOD_GSK3_1: GSK3 phosphorylation recognition site; MOD_NEK2_1: NEK2 phosphorylation motif; DEG_SCF_TRCP1_1: DSGxxS phospho-dependent degron recognized by F box protein of the SCF-betaTrCP1 complex; LIG_14-3-3_2: 14-3-3-binding motif ; DOC_MAPK_1: MAPK docking motifs; LIG_PDZ_Class_1: PDZ-binding motif; LIG_14-3-3_3: 14-3-3-binding motif ; LIG_EVH1_1: Proline-rich motif binding to signal transduction class I EVH1 domains. A). N-terminus, B). C-terminus.

7.4.2 P-loop containing nucleoside triphosphate hydrolase (NTPase) domain

The P-loop NTPase domain contains two sub-domains, the conserved NTPase α/β fold and a regulative region, known as helical third domain of STAND (HETHS). The NTPase domain of both TANC proteins was modeled considering both regions together. A HHpred search selected human apoptotic-protease activating factor 1 (APAF1; PDB code: 1z6t) as the best template, with 12.1% sequence identity for TANC1 and 12.5% for TANC2 (see Supplementary Table S7.1). Despite the presence of insertions between the TANC and template sequences, the conserved secondary structure elements superimpose well, especially in functional motifs on the catalytic core (Walker A, Walker B, and ASCE motifs). The Walker A and Walker B motifs define P-loop NTPase domains and are involved in nucleotide and Mg^{2+} cations binding respectively. The ASCE ("additional strand, catalytic E") motif, typically situated between both Walker motifs, determines ATP as preferred substrate (Figure 7.3). Moreover, residues placed in the catalytic pocket form a positively charged surface and are highly conserved in TANC orthologous (Figure 3, Supplementary Figure 7.2). While NTPase signature elements are rather conserved, the HETHS domain is quite variable among STAND family members and seems involved in family-specific regulative functions[254]. Indeed, since TANCs and APAF1 belong to different STAND NTPase families, their HETHS domains are more divergent in sequence and secondary structure. The 3D model quality evaluation of TANC2 and TANC1 is typical of more remotely homologous structures, with QMEAN scores of 0.421 and 0.391 respectively. However, lower quality regions are located in insertions corresponding to long disordered loops in TANC, while elements defining the catalytic core have low positional variability and higher reliability.

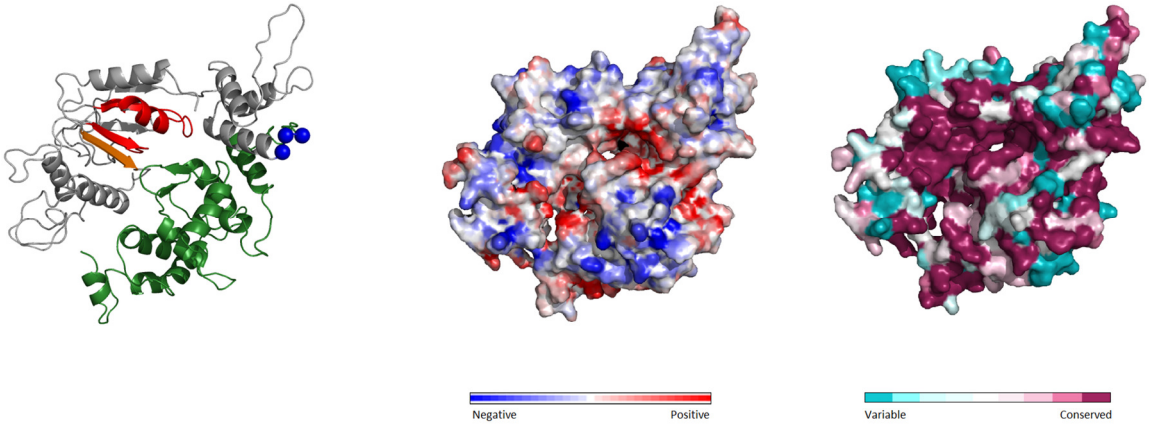


Figure 7.3: Structural analysis of ATPase domain in TANC1.

Cartoon of TANC1 ATPase domain model (front part) is colored as following: Walker motifs is in red, ASCE in orange, HETHS domain in green, GxP motif in blue spheres. Electrostatic properties of front surfaces are shown: negative charges in blue and red charges in red. ConSurf analysis of front surfaces, color code from unconserved (cyan) to conserved (purple) residues.

7.4.3 Ankyrin (ANK) repeat domain

Ankyrin repeats are a relatively conserved motifs of ca. 33 residues with a consistent pattern of key residues essential for structural integrity (Figure 7.4)[284]. The structural unit consists of a β-turn followed by two antiparallel α-helices and a loop connecting to the next repeat[284]. In both TANC proteins, eleven ankyrin (ANK) repeats are predicted by InterProScan. The alignment of ANK repeats reveals that the key conserved positions are overall maintained in both TANC1 and TANC2 (Figure 7.4, Supplementary Figure 7.3). Despite high sequence identity, the TANC1 repeat pattern is more regular, supporting divergent evolution. TANC2 presents longer loops and a peculiar negatively charged loop between the fifth and sixth repeat. Given its length, this loop separates the ANK domain into two regions and could be involved in TANC2 specific functions, as it is highly conserved among other species but not in TANC1 (Figure 7.4).

For both TANC1 and TANC2, HHpred selected the human ankyrin-R (PDB code: 1n11_A) crystal structure with 12 ANK repeats as template. Using the same template should allow a more accurate identification of structural differences between both proteins. The HHpred alignments were manually refined using the previously defined ANK repeats to maintain the structural integrity of each repeat. Both models have good quality, with QMEAN scores of 0.787 for TANC1 and 0.737 for TANC2. Each ANK domain is composed of eleven tandem repeats stacked together to form a linear solenoid structure. The linker loops of

neighboring repeats are connected in a tail to head order to form a hairpin-like β-sheet usually involved in protein-protein interactions in most ANK proteins [285]. Conservation and electrostatic surface analysis highlighted specific features for each TANC protein (Figure 7.4). TANC2 presents higher overall conservation than TANC1, with a negative charge in the concave region compensated by the prevalently positive convex surface (Supplementary Figure S7.3C). TANC1 presents a more significant separation between conserved residues belonging to the convex surface and unconserved positions in the concave region. The electrostatic surface follows the same pattern of TANC2, though more pronounced (Supplementary Figure 7.3). This region could be involved in electrostatic interactions with TANC binding partners.

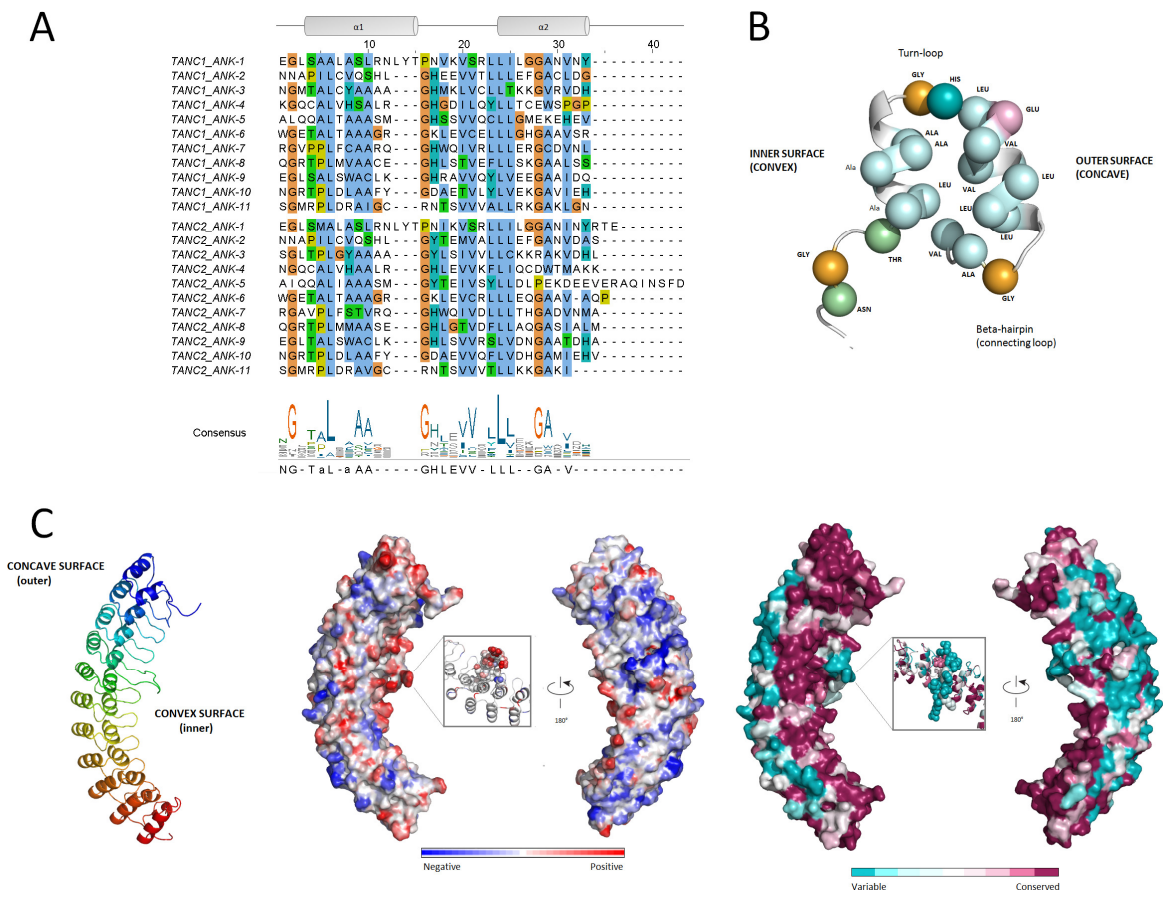


Figure 7.4: Ankyrin repeat overview and TANC1 Ankyrin domain model.

A) Consensus sequence of TANC ankyrin modules and related sequence logo. Residues that match the published consensus [284] are reported in upper case. Secondary structure is shown above the alignment: the inner alpha helix (α1) and the outer alpha helix (α2) are connected by a turn-loop (black line). **B)** Graphic representation of ANK repeats structure in TANC proteins. Conserved positions of ankyrin consensus pattern are reported in the diagram as spheres. Color code refers to consensus logo: hydrophobic amino acids (A, L and V) are in light blue, glycine in orange, threonine and asparagine in green, histidine in teal, glutamate in violet and proline in yellow. Residues matching the published consensus [284] are reported in bold. **C)** Cartoon

of TANC1 AR domain model is colored from N-terminus (blue) to C-terminus (red). Electrostatic properties of turn-loop surfaces and connecting-loop surfaces are shown: negative charges in blue and red charges in red. ConSurf analysis of turn-loop surface and connecting-loop surface, color code from unconserved (cyan) to conserved (purple) residues.

7.4.4 Tetratrico-peptide (TPR)-like repeat domain

Both TANC proteins are predicted to contain three TPR repeats which are extremely conserved among orthologous sequences. TPRs consist of 34 residues, whose consensus is defined by a pattern of small and large amino acids (Figure 7.5). Each module is formed by two antiparallel α helices, forming a superhelical helix-turn-helix fold. TPRs are typically involved in protein-protein interactions and assembly of protein complexes[286,287]. Despite the high sequence identity of human TANC TPR domains (80,4%), the template search selected different structures for homology modeling: human FK506-binding protein 52 (FKBP52, PDB code: 1P5Q) and *B. taurus* cyclophilin 40 (CYPD; PDB code: 1ihg) for TANC1 and TANC2 respectively. The HHpred alignments were again manually refined using the previously defined TPR repeat alignment. QMEAN shows a rather high reliability for both models, with scores of 0.749 for TANC1 and 0.732 for TANC2. The TPR models were evaluated for both conservation and electrostatic properties (Figures 7.5, Supplementary Figure 7.4). ConSurf revealed the presence of highly conserved regions in the TPR domains corresponding to the convex surfaces, with prevailing positively charged surfaces in both TANC proteins. On the other hand, the concave part seems to be less conserved, with the exception of negatively charged residues at the C-terminus. A coiled-coil region was previously thought to map downstream from the TPR domains [250]. Unlike TPRs, most domain predictors did not recognize any significant coiled coil region in TANC proteins (Supplementary Table 7.2). The presence of coiled-coil structures was assessed using three different tools. In both TANC sequences, all coiled coil predictors recognize a region downstream of the TPR-region with a low reliability score (Supplementary Table 7.2). However, secondary structure and a further manual evaluation of coiled coil motifs do not support this prediction [288]. Only one helix could be recognized downstream from the TPR domain for both TANC proteins (Supplementary Table 7.3). To exclude the presence of degenerate repeats in this position, TPR prediction was performed using TPR-pred. The analysis highlighted a low confidence TPR module (P-value> e-03) in TANC1, but not in TANC2. We conclude that the helix is neither a coiled coil nor a TPR repeat, but may represent a C-terminal cap for the TPR domain. Similar C-terminal capping structures

consisting of a 22 residue helix stabilizing the TPR fold [286,289] are present in both TANC1 and TANC2.

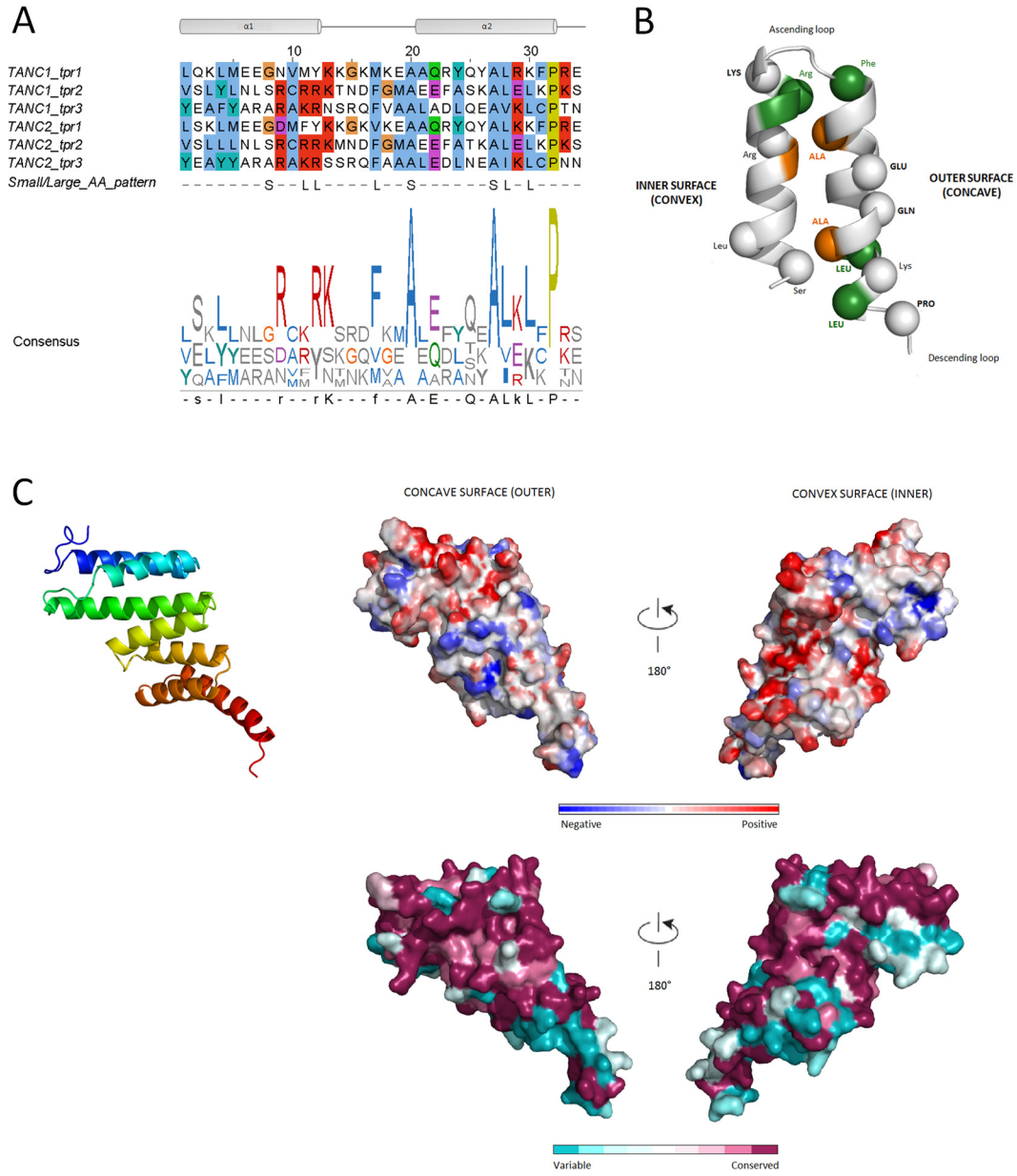


Figure 7.5: TPR repeat overview and TANC2 TPR model.

A) Consensus sequence repeat pattern of the TANC TPR domain and related sequence logo. Secondary structure is shown above the alignment: two alpha helices (grey shapes) connected by a loop (black line). Below the alignment, pattern of conserved small/large residues typical of TPR modules is reported: S indicates small residues, L for large residues. Residues that match the consensus are reported in upper case. **B)** Graphic representation of repeats structure in TANC proteins. Conserved positions of TPR consensus pattern are reported in the diagram (spheres). Residues that match the consensus are reported in bold. Conserved small-large residue pattern is also represented: dark green for large residues and orange for small residues. **C)** Cartoon of TANC2 TPR domain model is colored from N-terminus (blue) to C-terminus (red). Electrostatic properties of concave and convex surfaces are shown: negative charges in blue and red charges in red. ConSurf analysis of turn-loop surfaces and connecting-loop surfaces, color code from unconserved (cyan) to conserved (purple) residues.

7.4.5 C-terminus

In each TANC protein, the C-terminal region is preceded by the TPR C-capping helix and ends with the final PDZ binding motif. The Q9HCD6-3 and Q9HCD6-4 TANC isoforms are missing most of the ANK domain and the C-terminus, with the third - fourth ankyrin repeats and the fifth ankyrin repeat modified, respectively. Only few sequence stretches in the C-terminus have a significant similarity between TANC proteins and their orthologous (Figures 1 and Supplementary Figure 7.2). As expected, ELM recognized the highly conserved PDZ-binding motif in both paralogs, which has been demonstrated to mediate TANC interaction with PSD95 and SCRIB[249]. A poly-glutamine region followed by a proline stretch and a serine-rich region (SRR) are present in both TANC C-termini. Furthermore, MAPK and WW binding sites are predicted in the C-terminus of all TANC homologs sequences (Figure 7.2). These sites are partially overlapping and located in a region predicted to be phosphorylated by different kinases. Several 14-3-3 binding motifs are also predicted on different positions in the TANC C-termini.

The TANC2 C-terminus, but not its paralog, presents an unusual number of 27 conserved tyrosine residues showing a periodicity of ca. 12 residues. The presence of possible repetitive modules was therefore assessed. As expected, no repeat pattern was identified for TANC1, whereas both TRUST and RADAR recognized four repetitive regions in the sequence preceding the SRR. Further manual curation of TANC2 repeats suggests the presence of shorter modules of approximately twelve residues, in which the tyrosine residue represents the main signature. Taken together, these findings confirm the presence of a regular pattern that could organize the C-terminus and have a regulative role in protein function.

7.4.6 TANC network

We manually curated 24 TANC1 and 20 TANC2 interacting proteins. Thirteen TANC1 interactors and five TANC2 interactor were retrieved directly from publications. The remaining interactors have been determined by High-Throughput Screening (HTS) methods and deposited in publicly available PPI databases. The TANC interacting regions have been experimentally determined for only six TANC1 and two TANC2 interactors (see Table 7.1). Three PDZ domain proteins interacting with the C-terminal PDZ binding motif in TANC are considered mutually exclusive. For one known TANC1 and five TANC2 interactors we predicted a putative interacting site. These proteins present a domain or

belong to a class of proteins, which may recognize a conserved linear motif mapping in a disordered TANC regions. Exportin-1 and LATS-2 have a predicted binding motif on the structured TANC2 ATPase domain. The motifs are located in a loop that may be exposed upon conformational changes of the domain.

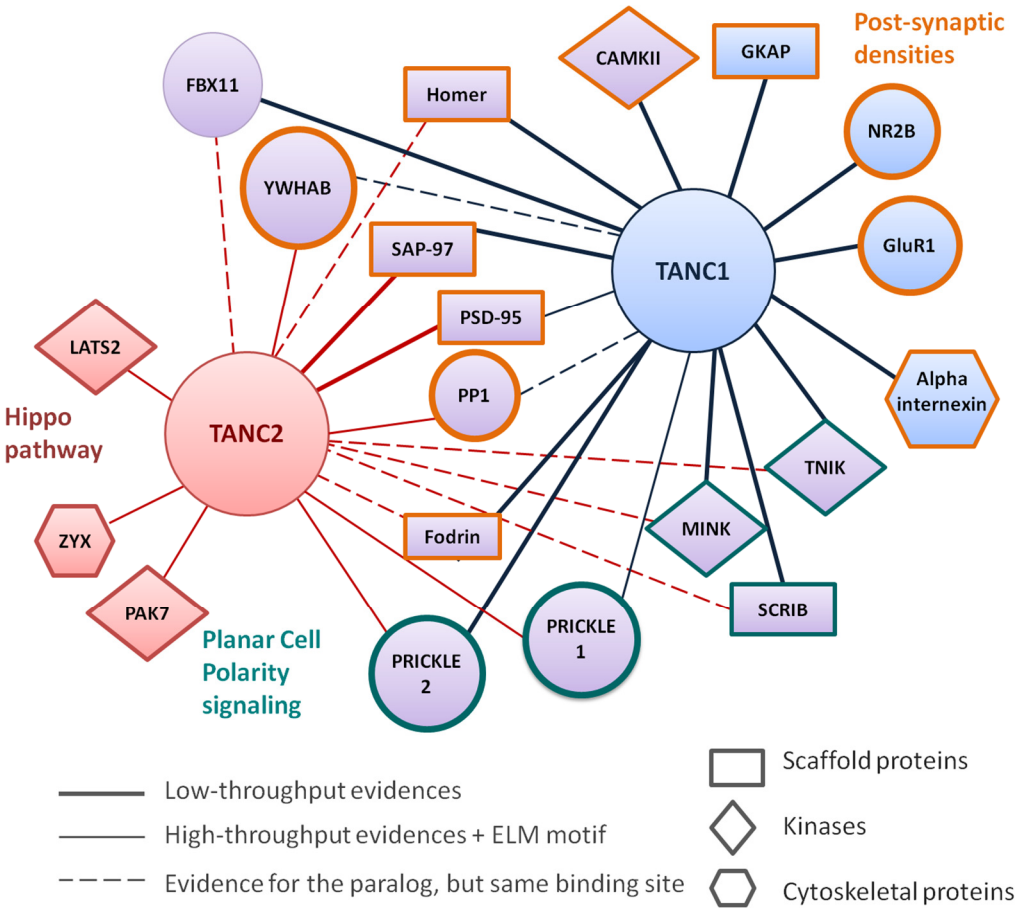


Figure 7.6: TANC protein interaction network.

TANC interaction partners identified by low throughput data (solid lines), PPI database evidence (thinner lines or linear motifs prediction. Interactions that are proved only in one paralog, but mediated by binding sites (linear motif or structural domain) that are identical in both proteins are reported as dotted edges. TANC1 interactors only are colored in light blue; TANC2 interactors only in red; while TANC interactors both are in violet. Interactors are represented with different shapes based on specific molecular function: scaffold proteins (rectangles), protein kinases (rhombus); cytoskeleton proteins (hexagons). TANC proteins are connected with different neuronal regulative proteins, belonging to Planar Cell Polarity signaling (teal outline), Hippo pathway (dark red outline) and glutamate signaling (orange outline).

We inferred novel interactors for each TANC member based on known interactors and shared conserved linear motifs of the paralog (see Table 7.2). Three TANC1 interactors may also bind TANC2 through shared linear motifs. The three proteins found to interact with the globular domains of TANC1 (Fodrin, MINK, and TNIK) may also bind TANC2,

although surface analysis of the ANK and TPR domains did not highlight a common conserved region. Finally, the N- and C-termini of both TANC proteins contain shared conserved binding motifs for different kinases and WW domain proteins. We hypothesize that these proteins may mediate post-translational TANC modifications.

TANC	Interacting protein			Experimental evidence	Ref.	TANC region
	Name	Domain architecture	Pathway			
1	α-internexin	Intermediate filament head, DNA-binding	Cytoskeleton organization	Co-IP	250	_
1	CAMKIIα	Kinase	Glutamate Receptor signaling	Co-IP	250	_
1	CASK	Casein Kappa	Glutamate Receptor signaling	Pull-down assay	250	_
2	CBY1	Chibby_fam	Wnt/Wingless signaling, Cilium assembly	SF-TAP/MS	268	_
2	CDC5L	Myb- HTH DNA binding type 1 and 2, Myb/Cef1 domain	Spliceosome assembly	HTS AC-MS	290	_
2	CENPQ	CENP-Q domain, Coiled coil	Nucleosome assembly at the centromere	HTS AC-MS	291	_
2	CEP120*	2 C2, Coiled coil	Centrosome organization, Cilium assembly	HTS AC-MS	291	_
1	CEP128	Coiled coil	Centrosome organization, Cilium assembly	HTS AC-MS	292	_
1	CNTRL	4 LRR, 4 Coiled coil	Centrosome organization, Cilium assembly	HTS AC-MS	292	_
1	FBXW11	F-box, 7 WD repeats	Ubiquitin-mediated degradation	Co-IP	293	_
2	FMRP	Agenet-like, KH, FXMRP1_C_core, FXMR_C2	Regulation of translation	CLIP	294	_
1	Fodrin	23 Spectrin repeats, SH3,3 EF-hand	Cytoskeleton organization	Pull-down assay	250	ANK and TPR
1	GKAP	3 Coiled coil	Glutamate Receptor signaling	Co-IP	250	_
1	GluR1	TM	Glutamate Receptor signaling	Co-IP	250	_
1	GRIP	7 PDZ	Glutamate Receptor signaling	Pull-down assay	250	_

(Continues)

Table 7.1 (Continued)

TANC	Interacting protein			Experimental evidence	Ref.	TANC region
	Name	Domain architecture	Pathway			
1	Homer	WH1/EVH1, Coiled coil	Glutamate Receptor signaling	Pull-down assay	250	_
2	INPP5E	13 repeats of P-X-X-P, Phosphatase	Cilium trafficking	SF-TAP/MS	268	_
2	LATS2	UB associated Kinase, AGC-kinaseC-terminal	Hippo pathway	HTS PL-MS	295	_
2	MAPRE1	CH, EB1_C	Microtubule cytoskeleton regulation	SF-TAP/MS	268	_
1	MOV10	P-loop ATPase domain	RNA-mediated gene silencing	HTS AC-RNA	296	_
1	MINK	Kinase, CNH	Rap2-mediated signaling	Immunoblotting	297	TPR
1	NINL	4 EF-hand, 4 Coiled coil, KEN-box, D-box	Centrosome organization, cilium assembly	HTS AC-MS	292	_
1	NR2B	Transmembrane receptor	Glutamate Receptor signaling	Co-IP	297	_
2	NR2C2	ZF- C4, NHR ligand binding	Nuclear receptor signaling pathways	HTS AC-MS	291	_
1	NXF1	RRM, 4 LLR repeats, NTF2, TAP-C	mRNA export from nucleus	HTS AC-RNA	296	_
2	PAK7	CRIB, kinase	Planar Cell Polarity pathway	HTS AC-MS	298	_
1	PCM1	Coiled coil, GTPase, molybdopterin domain	Centrosome organization, Cilium assembly	HTS AC-MS	292	_
2	PPP1CA	Ser/Thr phosphatase	Glutamate Receptor signaling, Hippo, Wnt signaling	HTS AC-MS	298	_
2	PPP1CC	Ser/Thr phosphatase	GluR, Hippo signaling	HTS AC-MS	298	_
1 & 2	PRICKLE1	PET, 3 LIMs	Planar Cell Polarity pathway	LC-MS/MS	299	_
1 & 2	PRICKLE2	PET, 3 LIMs	Planar Cell Polarity pathway	LC-MS/MS	299	_
1 & 2	PSD-95	3 PDZ, SH3, GK	Glutamate Receptor signaling	Y2H, Pull-down assay	196	LIG_PDZ_Class_1
1 & 2	SAP97	L27, 3 PDZ, SH3, GK	Glutamate Receptor signaling	Y2H, Pull-down assay	196	LIG_PDZ_Class_1

(Continues)

Table 5.1 (Continued)

TANC	Interacting protein			Experimental evidence	Ref.	TANC region
	Name	Domain architecture	Pathway			
1	SCRIB	16 LRR repeats, 4 PDZ	Planar Cell Polarity pathway	SPR	[300]	LIG_PDZ_Class_1
1	SHANK1	6 ANK, SH3, PDZ, SAM	Glutamate Receptor signaling	Pull-down assay	[250]	_
2	SPIRE2	KIND, 3 WH2, ZF	Vesicles transport	HTS AC-MS	[291]	_
1	TNIK	Kinase, CNH	Rap2-mediated and Wnt signaling	Immunoblotting	[297]	TPR
2	XPO1	Importin_N-term, 10 ARM/HEAT repeat like	Nuclear export	Pull down	[301]	_
2	YWHAB	14-3-3	Glutamate Receptor signaling, Hippo signaling	HTS AC-MS	[295]	_
2	ZYX	3 LIM, Zn binding	Hippo pathway	HTS AC-MS	[298]	_

Table 7.1: List of TANC interactors.

For each interactor, the interacting TANC protein, the detection method and the binding region (experimentally validated) are here listed. Y2H: Yeast two hybrid; Co-IP: Co-immunoprecipitation; SPR: Surface plasmon resonance HTS: High-Throughput System; AC: Affinity Capture; PL: Proximity Label; MS: Mass spectrometry; CLIP: Cross-Linking Immuno-Precipitation SF-TAP/MS: systematic tandem affinity purifications coupled to mass spectrometry. SLiMs are named according to the ELM nomenclature.

TANC	Predicted Interactor		Method	Predicted TANC interacting region
	Name	Domain architecture		
1	FBXW11	F-box, 7 WD repeats	TANC1 interactor, Conserved SLiM in IDR	DEG_SCF_TRCP1 DEG_Nend_Nbox_1
1	YWHAB	14-3-3	TANC2 interactor, same SLiM in TANC1	LIG_14-3-3_2
1	PPP1CA	Ser/Thr phosphatase		DOC_PP1_RVXF_1
1	XPO1	Importin_N-term, 10 ARM/HEAT repeat like		TRG_NES_CRM1_1
2	YWHAB	14-3-3	TANC2 interactor, Conserved SLiM in IDR	LIG_14-3-3_2 LIG_14-3-3_3
2	PPP1CA	Ser/Thr phosphatase		DOC_PP1_RVXF_1
2	LATS2	UB associated Kinase, AGC-kinase C-terminal		MOD_LATS_1
2	NR2C2	ZF- C4, NHR ligand binding		LIG_NRBOX (score 0,3)
2	XPO1	Importin_N-term, 10 ARM/HEAT repeat like	TANC2 interactor, Conserved SLiM in P-loop domain	TRG_NES_CRM1_1
2	Homer	WH1/EVH1, Coiled coil	TANC1 interactor	LIG_EVH1_1

(Continues)

Table 7.2 (Continued)

TANC	Predicted Interactor		Method	Predicted TANC interacting region
	Name	Domain architecture		
2	SAP97	L27, 3 PDZ, SH3, GK	TANC1 interactor, same SLiM in TANC2	LIG_PDZ_Class1
2	SCRIB	16 LRR repeats, 4 PDZ		LIG_PDZ_Class1
2	FBXW11	F-box, 7 WD repeats		DEG_SCF_TRCP1 DEG_Nend_Nbox_1
2	Fodrin	23 Spectrin repeats, SH3, 3 EF-hand	TANC1 interactor, same domain in TANC2	ANK and TPR
2	MINK	Kinase, CNH		TPR
2	TNIK	Kinase, CNH		TPR
1	CDK	Kinase	Conserved SLiM in IDR	MOD_CDK_1
1&2	G- Actin?	Actin domain		LIG_Actin_WH2_2
1&2	Cyclins	Cyclin, N-terminal		DOC_CYCLIN_1
1&2	MAPK	Kinase		DOC_MAPK_gen_1
1&2	WW domain-containing proteins	WW domain		DOC_WW_Pin1_4
1&2	Atg8 protein family	Autophagy		LIG_LIR_Gen_1
1&2	CK1	Kinase		MOD_CK1_1
1&2	GSK3	Kinase		MOD_GSK3_1
1&2	NEK2	Kinase		MOD_NEK2_1

Table 7.2.: Predictions for TANC interactors. SLiM: Short Linear motif; IDR: Intrinsically disordered region.

7.4.7 Missense mutation analysis

Three TANC2 missense mutations have been reported in three unrelated patients with different neuropsychiatric phenotypes[245,252,302]. The two variants p.Arg760Cys and p.Ala794Val map on the ATPase regulative domain. The former has been found de novo in a pediatric patient presenting intellectual disability[302]. The p. Arg760Cys variant maps on a buried loop facing the ASCE strand within the ATPase regulative domain in catalytic pocket. The substitution of a charged arginine residue with a cysteine may have some effect on the catalytic pocket, where charged residues coordinate Mg^{2+} ions and binding of ATP molecules. The p.Ala794Val was inherited from the father in a patient with schizophrenia[245]. It affects a buried residue in the ninth helix of the regulative region that could affecting folding due to steric clashes. Both mutations are predicted as pathogenic by most prediction tools (11/12 for R760C and 12/12 for A794A, details in Supplementary Table 7.4) and likely affect regulative domain stability and ATPase activity. A third

inherited mutation mapping to the C-terminal tail (p.His1689Arg) was found in a patient with autism spectrum disorder[252]. Although it maps within a conserved region, the Histidine to Arginine substitution is only predicted to be damaging by six of twelve tested methods (Supplementary Table S7.7).

7.5 Discussion

Recently, evidence from mouse models and human patients suggested the TANC proteins as candidates for NDD. Despite different expression profiles in the brain, TANC1 and TANC2 have both been shown to positively regulate dendritic spines and excitatory synapses[249] . The TANC family has been described as PSD95 partners found to localize and interact with several postsynaptic proteins[250]. Here, we report an in depth *in silico* analysis of the TANC family structure and function to gain insights on their molecular function as well as to elucidate the role of these proteins in NDDs. The P-loop domain model suggests that the TANC proteins may have an ATPase activity since all functional elements are conserved, although the regulative domain differs from other proteins of this class and its role has to be demonstrated. Modeling the repeat domains allowed identifying conserved PPI interfaces for both ANK and TPR domains, with different electrostatic charges possibly involved in protein binding. Despite previously reported predictions, sequence and structural analysis of the TPR domain allowed to exclude the presence of coiled-coil region in TANC, as the mispredicted region corresponds to a stabilizing C-capping element of the TPR domain. Along the N and C- terminal disordered regions of TANC we predicted several conserved SLiMs supporting interactions from high-throughput experiments known to have false positives (Table 7.2, Figure 7.6). The prediction of putative interacting regions, besides inferring novel interactors, allowed to define some proteins as mutually exclusive interactors. PSD95 and SCRIB interact with the TANC PDZ linear motif anchoring TANC proteins to the glutamate receptor or in PCP signaling[249,250,300,303,304]. Although most short motif patterns have a high chance of random occurrence and their prediction may have low specificity, we used stringent criteria [82] to select putative protein binding sites. To be considered, a binding site has to be conserved among orthologous, or shared among paralogs, and mapping to a disordered region. Alternatively, spliced regions are also favorable factor of being a true binding site. Moreover, the putative motif is supported if its binding proteins known TANCs interactors or involved in the same biological processes[82]. We expanded the functional network of

TANC proteins, integrating prediction and high throughput data, and inferring protein partners based on information of one of the two TANC family member (see Figure 7.6). We found that the TANC N-termini present several conserved linear motifs, which may be involved in a broader range of cell regulation, including phosphodegrons and phosphatase docking motifs. These motifs could be the target of two TANC interactors identified by high-throughput screening, protein phosphatase 1 (PP1) and FBXW11. The latter is a component of the SCF E3 ubiquitin-protein ligase complex implicated in recognition of phosphorylated proteins targeted for degradation[305]. PP1 is one of the three phosphatases expressed in neurons regulating NMDAR-dependent Long Term Depression (LTD) during development[306].n Another mechanism involved in functional TANC regulation is suggested by the findings that RNAs of both TANCs are targets of the MOV10 RNA helicase and the Fragile X mental retardation protein (FMRP)[307,308]. Recently, MOV10 was found to be a functional partner of FMRP[308]. MOV10 promotes miRNA-mediated translational suppression of its target RNAs, while FMRP regulates synaptic strength at glutamatergic synapses by controlling translation of specific RNAs. TANC regulation may also occur through post-translational modifications (PTM) sites we have predicted. Different kinases, such as CAMKII, MINK TNIK, PAK7, LATS2, have been identified as TANC interactors and PTM sites have been frequently shown to conditionally switch motif-mediated interactions[81] triggering different signaling pathways. Predicted and collected PPI data allowed us to position TANC proteins in several biological processes, other than post-synaptic density proteins, such as the planar cell polarity pathway[299,309], Wnt signaling and Cilium assembly. We also found for TANC1 and TANC2 specific connections with Rap2-mediated and Hippo signaling[295], respectively, that may explain different roles of TANC1 and TANC2 in brain function. However, all of these pathways contribute in different ways to correct neuronal development and maintenance [297,299,300,304]. The TANC proteins thus appear to be regulated at several levels from synthesis to degradation, while being involved in pathways controlling neural development and maintenance. It is likely that alterations of these proteins may affect different processes, thus explaining the broader range of disease phenotypes associated with TANC variants. The performed analysis allowed us to discover structural and functional elements that will help the interpretation of newly discovered TANC gene variants. It would be worth following them up experimentally to support a mechanistic model for TANC function as a dynamically regulated scaffold.

8 Unraveling TANC2-CDKL5-PP1 interactions: intrinsically disordered regions mediating novel pathways in neurodevelopment

8.1 Summary

This Chapter summarizes the experimental approaches used in validating TANC2 interactions with PP1 and CDKL5, and the functional significance of the TANC2-CDKL5-PP1 interplay. The interaction among endogenous full length proteins was assessed both by co-localization analysis in primary hippocampal cultures from E18 rats and neuroblastoma cell line, and immunoprecipitation from rat synaptosomes. Binary interactions between TANC2, PP1 and CDKL5 sub-clones were validated by yeast two-hybrid assays. TANC2 silencing was performed in SHSY5Y cells to assess the functional relationship among regulative TANC2 activity and CDKL5 protein levels. Our findings suggest that TANC2 not only is able to form a complex with either CDKL5 or PP1, but it also functions as a scaffold linking the phosphatase PP1 to its substrate CDKL5, allowing the dephosphorylation and subsequent kinase degradation.

8.2 Introduction

Neurodevelopmental disorders (NDDs) are common conditions including clinically heterogeneous diseases, such as intellectual disability (ID) and autism spectrum disorders (ASD). In most cases, NDDs are genetically determined and, up to now, several hundreds of causative loci have been already identified[166]. CDKL5 disorder is caused by mutations on the CDKL5 gene, resulting in early onset seizures and severe neurodevelopmental impairment[310]. Although biological functions of CDKL5 remain largely unknown, its role in synapse development and neuronal plasticity has been established demonstrating that it regulates neuronal morphogenesis via Rac1 signaling [310]. The CDKL5 ensures excitatory synapse stability by reinforcing the NLG1-PSD95 association at the postsynaptic compartment[311] and its synaptic localization is regulated through its direct interaction with a palmitoylated form of PSD-95 [312]. It was demonstrated that upon NMDAR stimulation,

CDKL5 is dephosphorylated by PP1, causing its proteasome-dependent degradation in mature neurons[313]. This step appeared to be relevant for an activity-dependent signaling cascade, regulating synapse composition, shape and strength. However, the mechanism targeting PP1 activity towards CDKL5 is still debated [313]. We speculated TANC2 could mediate CDKL5 degradation, and different lines of evidence contributing to the formulation of this hypothesis. TANC2 is recently emerging as a candidate gene for neurodevelopmental disorders (NDDs), as TANC2 mutations have been associated with several forms of NDDs, including autism and intellectual disability[197], both characteristics shared with CDKL5 disorder[310]. TANC2 and CDKL5 are post-synaptic proteins known to interact with PSD-95, which is required for their proper localization at post-synaptic densities (PSDs) and for their association with additional postsynaptic proteins, including the glutamate receptors[196,311]. Although TANC2 function in brain cells remains unclear, this scaffold protein is proposed to play a critical role in organizing different components of glutamate receptor complexes at PSD and determining synaptic strength and plasticity[196]. CDKL5 presents two isoforms, i.e. CDKL5-1 (1,033 AA) and CDKL5-2 (960 AA). CDKL5-2 is the most abundant isoform both in fetal and adult brain, where it is constitutively expressed[314]. Contrarily to the isoform 2, CDKL5-1 displays differences in expression profiles over time, with higher levels in fetal brain than in adult one[314]. Like CDKL5-1, TANC2 is highly expressed during the early embryonic stages, where it is thought to be involved in proper fetal development, with knock-out causing lethality *in utero*[196]. In an our previous investigation[197], we found TANC2 to harbor several conserved linear motifs, including phosphatase docking motifs. These motifs could be the target of interactors identified by high-throughput screening, such as the protein phosphatase 1 (PP1). PP1 dephosphorylates hundreds of key biological targets and its activity is controlled by hundreds of regulative proteins directing its specificity[315]. Thus, we hypothesize that the TANC2 scaffold protein could mediate and regulate PP1 function on CDKL5. Therefore, this work aims to fully validate PP1/TANC2/CDKL5 interactions taking advantage of different *in vitro* techniques, and to better understand the role of TANC2 in PP1-mediated CDKL5 degradation.

8.3 Materials and methods

8.3.1 SHSY5Y and primary hippocampal neurons cultures

Human neuroblastoma SH-SY5Y cells were plated on laminin-coated (1 μg/cm^2, Sigma) coverslips in 24-well plate at a density of 5×10^4 cells/mm^2. Cells were maintained in Dulbecco Modified Eagle Medium (DMEM) supplemented with 10% heat-inactivated fetal bovine serum (FBS), 2mM of glutamine and antibiotics (penicillin, 100 U/ml; streptomycin, 100 μg/ml), in a humidified incubator with an atmosphere of 95% air and 5% CO_2, at 37°C for six days. Primary cultures of hippocampal neurons were prepared[316] from the isolated hippocampi of E18 rat embryos purchased from BrainBits®. Briefly, hippocampi were dissociated with trypsin 0,06% at 37 °C for 15minutes, washed in Ca^{2+}- and Mg^{2+}-free Hank's balanced salt solution (HBSS; Gibco®) containing 10% FBS to block trypsin enzymatic activity, and homogenated in Neurobasal medium supplemented with 1% B27 supplement, 25 μM glutamate, 0.5 mM glutamine, and 50 μg/ml gentamycin. Dissociated cells were filtered with a 100 μm cell strainer to obtain a uniform single cell suspension. Isolated cells were plated on poly-d-lysine-coated coverslips (0.1 mg/ml, Sigma) in 24-well plate at a density of 1.52×10^5 cells/mm^2 and maintained in normal growth conditions (atmosphere of 95% air and 5% CO_2, at 37°C) for 15 days. After seven days, half of the culture medium was replaced with fresh medium without glutamate.

8.3.2 Immunofluorescence protocol

Immunofluorescence of SH-SY5Y cells was performed after an incubation time of 72 hours with differentiation medium (IGF-1 50 nM, 2 mM of glutamine and antibiotics). The cells were fixed with 3.7% formaldehyde (v/v) for 20 minutes and permeabilized for 5 minutes with 0.1% Triton X-100 in PBS (v/v). Washes were performed using 1% gelatin/PBS 1X (w/v) to block nonspecific antibody-binding sites. Immunostaining was performed incubating the coverslips with the primary antibodies in PBS, 90 minutes at 37°C. After three washes with 1% gelatin/PBS and other three with PBS 1X, coverslips were incubated with the secondary antibodies, for 30 minutes at room temperature. Hippocampal neurons (15 DIV) were fixed with 4% paraformaldehyde/sucrose in PBS and permeabilized with 0.3 % Triton X-100 in PBS for 5 min. The cells were then incubated in PBS/10 % BSA for one hour at room temperature. Immunostaining was performed incubating the coverslips with the primary antibodies in PBS/3 % BSA, two hours at room temperature. Each sample

was washed six times with PBS and was incubated with the secondary antibodies, for one hour at room temperature. In both cases, the coverslips were mounted with Dako fluorescence mounting medium. Primary antibodies: goat anti-TANC2 S16 (1:20, Santa Cruz Biotechnology), rabbit anti-CDKL5 A304-172 (1:20, Bethyl) and rabbit anti-PP1 A300-904 (1:20, Bethyl). Secondary antibodies: donkey anti-goat IgG conjugated with Alexa Flour 555 (1:50, Invitrogen) and donkey anti-rabbit IgG conjugated with Alexa Flour 647 (1:50, Invitrogen). All antibody dilutions for immunostaining experiments were done in PBS 1X.

8.3.3 Colocalization analysis

Imaging of immunostained samples was performed with Leica SP5 confocal microscope, using a 63× oil objective (1,024x1,024 resolutions, 200Hz, z-stack step ≥ 0.5 µm, PA 1AU). Confocal image z-stacks were analyzed using the ImageJ2/Fiji colocalization plugin Coloc2 (https://imagej.net/Coloc_2). Pixel intensity correlation was calculated by the automated Manders' method[317] with a point spread function of 1 and 10 shuffling iterations for the Costes' significance test[100].

8.3.4 Rat cortex synaptosome preparation

Purified synaptosomes were prepared according to the Nicholls method[318]. In brief, brains from 3 months old male rats were homogenized in lysis buffer (4mM HEPES-Na, 0.32 M Sucrose, pH 7.3) with 12 strokes in a glass homogenizer on ice. Homogenate was centrifuged using AvantiJ-centrifuge at 5,000 r.p.m for 3 minutes with JA25.50 rotors at 4°C. Then, supernatant was centrifuged at 1,1000 r.p.m for 12 minutes at 4°C. The pellet was resuspended in 3 ml of lysis buffer taking care to not perturb the mitochondrial button, and diluted to 1:2 factor with a HEPES-buffered solution (140 mM NaCl, 5 mM KCl, 20 mM HEPES-Na pH 7.4, 5mM $NaHCO_3$, 1 mM $MgCl_2$, 1.2m M Na2HPO4 and 10 mM Glucose). The resuspension was divided in 2 ml tubes, centrifuged at 4°C for 10 seconds, and resuspended in HEPES-buffered solution with a volume up to 1.5 ml. The obtained synaptosomes were lysated in the lysis buffer (40mM Tris-HCl pH 8, 150 mM NaCl, 50mM Na-Citrate, 5mM CHAPS, 1X PIC) and used for co-immunoprecipitation experiments.

8.3.5 TANC2-CDKL5 co-immunoprecipitation experiments

Endogenous TANC2/CDKL5 complexes were immunoprecipitated from rat synaptosomes using rabbit anti-TANC2 A303-023A antibody (Bethyl). The antibody (2µg in 200 µl TBS/0.05% Tween-20) was incubated 1 h at 4°C in agitation (500 rpm) with 5 µl of Protein a Magnetic Beads (Pierce ThermoFisher). Then, after three TBS-T 0.05% washes, lysates were incubated with conjugated resin for 1 h at low temperature (4°- 6°C), washed three times with 200 µl of the lysis buffer (see above), and eluted with NuPAGE LDS Sample Buffer 4X (ThermoFisher). The eluted samples were added of 100 mM DTT and incubated for 10' at 70°C in agitation (500 rpm). Afterwards, samples were analyzed by immunoblotting with rabbit anti-CDKL5 A304-172A antibody (Bethyl). IrDye 800 anti-rabbit (LI-COR) was used as secondary antibody at 1:10000 dilutions in BSA 5%/TBST (TBS-Tween20) 0.1%. To confirm the results, the experiments were replicated using 2 µg rabbit anti-CDKL5 antibody for immunoprecipitation and anti-TANC2 A303-023A antibody for immunoblotting.

8.3.6 Plasmids, oligonucleotides, site directed mutagenesis

Assessment of TANC2-PP1 and TANC2-CDKL5 interactions was performed using the yeast two-hybrid system, as described previously in Minervini et al. 2017[319]. In brief, pcDNA3.1-derived plasmids (GenScript) carrying the full-length synthetic cDNA of TANC2 (NM_025185), CDKL5 (NM_001037343) and PP1 (GenEZ™ORF clone OHu17166) were used as templates to transfer different coding regions in yeast vectors pGADT7 and pGBKT7, provided by Clontech. The DNA fragments were amplified by PCR (primer sequences in Supplementary Tables 8.1 and 8.2), and cloned in the final plasmids, near di EcoRI restriction site, using the In-Fusion® HD Cloning Kit (Clontech) standard protocol. When possible, shorter overlapping fragments of both CDKL5 and TANC2 C-terminal region (1,228-1,990) were obtained using QuikChange II XL Site-Directed Mutagenesis Kit (Agilent Technologies), following the supplier's instructions. All recombinant plasmids express TANC2, CDKL5 or PP1 fragments fused with either the DNA binding domain (DBD, pGBKT7-derived plasmids), or the activation domain (AD, pGADT7-derived plasmids) of the Gal4 transcription factor.

8.3.7 Yeast-two-hybrid (Y2H) experiments

Minimal interaction sub-regions between TANC2, CDKL5, PP1 were investigated through yeast two-hybrid assay, by the Matchmaker® Gold Two Hybrid system (ClonTech).The co-transformation Y190 reporter yeast strain was performed following the one-step transformation protocol [320]. Generally, multiple transformations (2-3) were performed, from which two or more independent colonies were tested, using 10-fold dilutions on solid medium. Positive interactions were identified by growth on His− Leu− Trp− 30 mM and 60 mM 3-Amino-1,2,4-Triazol (3AT) plates. To assess the interaction strength of tested combinations, a positive control strain, containing both Gal4 AD-SV40 large T-antigen and Gal4 DBD-murine p53 (fragment 72–390), and a negative control, corresponding to Y190 co-transformed with empty pGADT7 and pGBKT7, were employed. Yeast stain growth was monitored for 3 and 6 days at 30°C.

8.3.8 *S*equence feature analysis and known CDKL5 interactors analysis

CDKL5 isoforms (UniProt accession code: O76039-1, O76039-2) were downloaded from UniProt[180], and aligned with orthologous sequences retrieved from OMA Browser[277] (forty five sequences, accession date: 15/08/17). The C-terminal protein sequences were assessed for intrinsic disorder, presence of compositionally biased regions (i.e., repeating amino acids) and short linear motifs (SLiMs) using FELLS[78] and ELM[81], whereas secondary structure was predicted using PSIPRED[77]. Since SLiMs can be highly degenerated, and many putative SLiMs could be false positives, we selected only those mapping to ANCHOR[321] predicted binding sites in conserved disordered regions (perfect match in at least twenty three out of forty five sequences used for the multiple alignment). A list of experimentally determined CDKL5 interactors was gathered from the literature and the publicly available databases BioGrid[36], and IntAct[35] (see Table 8.2). For each binding partner, the interaction details from the literature (i.e. involved sequence motifs, domains, and regions) and the involved biological processes from the InterPro[76], UniProt[180] and KEGG[269] databases were annotated

8.3.9 TANC2 silencing

TANC2 silencing was performed in SHSY5Y cells, using the FlexiTube siRNA premix system (QIAGEN). In brief, shortly before transfection, cells were seeded in a 24-well plate at a concentration of 0.8 x 10^5/well, in 0.6 ml of in supplemented DMEM (see above), and

briefly incubated in normal growth condition (5% CO2, at 37°C). Two different FlexiTube siRNA sequences were tested, Hs_TANC2_1 (SI03131639) and Hs_DKFZP564D166_3 (SI00367003), both targeting TANC2 NM_025185 transcripts. The All Stars Negative Control siRNA (QIAGEN) was used as negative transfection control. 25 nM of each FlexiTube premix, which contains both transfection reagents and the siRNAs, was added drop-wise to the cells. Cells were harvested 72h after transfection, using a lysis buffer (50 mM Tris-HCl pH7.5, 150mM NaCl, 10% glycerol 1mM EDTA 0,5% NP-40). TANC2 and CDKL5 expression levels were assessed on cell lysates by immunoblotting against the two proteins. The samples were diluted with Leamli sample buffer 5X (60 mM Tris–HCl pH 6.8, 10% glycerol, 2% SDS, 0.001 % bromophenol blue, and 5% β-mercaptoethanol) and 100 mM DTT. Protein samples were separated by SDS-PAGE in 10 % polyacrylamide gels and transferred to nitrocellulose membranes. Membranes were blocked in 5 % milk in TBS-T 0.1%, and incubated with primary antibodies overnight at 4°C. Finally, the membranes were washed and exposed to alkaline phosphatase-conjugated secondary antibodies. To quantify sample band intensities, β-actin was used as internal normalization standard. Significance of difference in protein levels using Microcal Origin 8.0 (Malvern Instruments, Worcestershire, United Kingdom) by Welch's *t* test, with *p-value* < 0.05 considered statistically significant. Data were graphed using Microsoft Excel and expressed as mean ± S.E.M.

Primary antibodies: rabbit anti-CDKL5 A304-172A antibody (1:3,000, Bethyl), rabbit anti-TANC2 A303-023A antibody (1: 3,000 Bethyl), mouse anti-β-actin antibody A5441 (1: 1,000, Sigma) Secondary antibodies: goat anti-rabbit IgG (whole molecule) – peroxidase A6154 antibody (1: 10,000 Sigma), rabbit anti-mouse IgG (whole molecule) – peroxidase A9044 antibody (1: 10,000 Sigma). All antibody dilutions for western blot experiments were done in BSA 1%/TBS-T 0,1%.

8.4 Results

8.4.1 TANC2 colocalizes with CDKL5 and PP1, and CDKL5-TANC2 interaction is confirmed by co-IP in rat synaptosome

TANC2-CDKL5 and TANC2-PP1 interactions between endogenous proteins was tested by double-labeled immunofluorescence in different cell types. Colocalization analysis revealed that there is a high degree of spatial correlation between TANC2 and PP1, both in primary hippocampal neurons (Pearson's correlation coefficient (PCC) = 0.44 ± 0.03; n > 10 cells from 4 different coverslips), and SHSY5Y neuroblastoma cells (PCC = 0.49 ± 0.04; n > 20 from 6 different coverslips; Figure 8.1), confirming the predicted interaction between the two proteins. The same was observed for TANC2-CDKL5 immunostained samples, with PCC = 0.59 ± 0.04 in hippocampal neurons (n > 10 in 6 different coverslips) and PCC = 0.47 ± 0.03 in SHSY5Y cells (n > 20 from 5 different coverslips), suggesting that CDKL5 and TANC2 could be interaction partners. In addition, to evaluate colocalization, we performed with Coloc2 plugin, which returns an intensity correlation quotient (ICQ), ranging from -0.5 (perfect anti-correlation) and 0.5 (perfect colocalization)[322]. In our samples, the ICQ values confirm the colocalization of TANC2 with PP1 and with CDKL5 (TANC2-PP1: ICQ = 0.33 ± 0.01 and 0.35 ± 0.02 in primary and neuroblastoma cells respectively; TANC2-CDKL5: ICQ = 0.36 ± 0.02 in hippocampal neurons and 0.32 ± 0.03 in SHSY5Y cells). However, we noticed that TANC2 localization slightly differs from CDKL5 and PP1 distributions within cells. Indeed, PP1 and CDKL5 signals are more homogeneous in the cytoplasm, whereas TANC2 is more concentrated in dendrites and at membrane level, in accordance with TANC2 being a scaffold protein (Figure 8.1). To confirm CDKL5-TANC2 interaction in conditions near physiological environment, co-immunoprecipitation of the two proteins was performed using rat hippocampal synaptosomes. The experiments confirm that CDKL5-TANC2 complex can be detected in *ex vivo* preparations. Indeed, TANC2 can be detected in the sample immunoprecipitated using an antibody against kinase domain of CDKL5, as well as CDKL5 can be detected in the sample immunoprecipitated using the αTANC2-antibody (Figure 8.1).

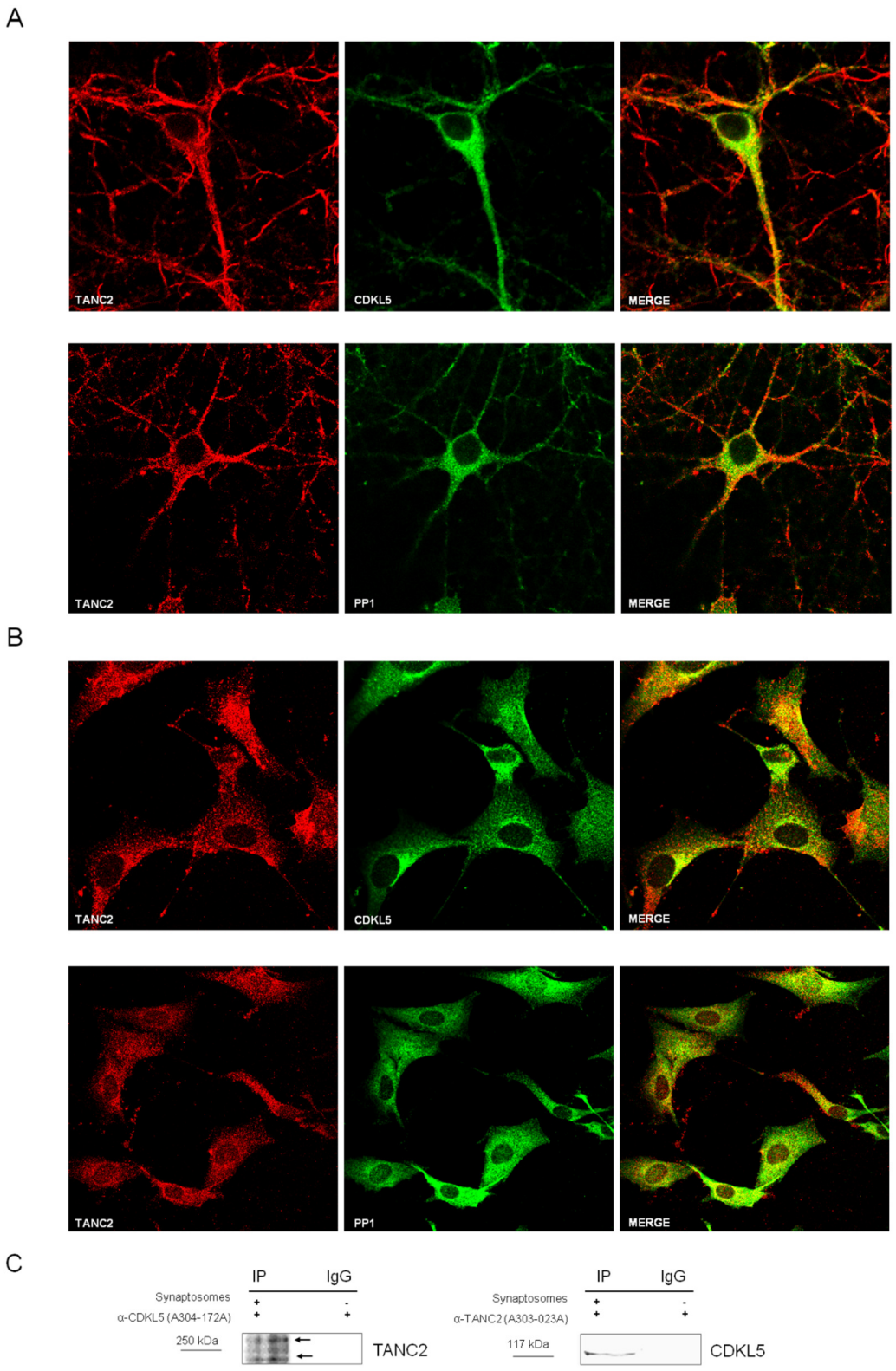


Figure 8.1: Validation of PP1-TANC2 and CDKL5-TANC2 interactions. **A)** Representative images of endogenous TANC2, CDKL5 and PP1 immunostaining in primary hippocampal neurons. **B)** Representative images of endogenous TANC2, CDKL5 and PP1 immunostaining in SHSY5Y cells. **C)** Validation of TANC2-CDKL5 interaction assessing immunoprecipitation of endogenous complexes from rat synaptosome extracts.

8.4.2 In yeast, TANC2 interacts with PP1 through the its N-terminus, whereas CDKL5 interaction is mediated by TANC2 C-terminus

The characterization of TANC2-PP1 and TANC2-CDKL5 interacting sub-regions was performed using the yeast two hybrid system. TANC2 sequence was divided in different sub-clones (Table 8.1), which were tested separately against PP1 and CDKL5-1 full length proteins.

TANC2		
Region	Position	Domains
TANC2.I	1-845	N-terminus
TANC2.II	1-845	N-terminus + ATPase domain
TANC2.III	1-1,227	N-terminus + ATPase domain + ANK domain
TANC2.IV	341-1,227	ATPase domain + ANK domain
TANC2.V	846-1,227	ANK domain
TANC2.VI	1228-1,990	TPR domain + C-terminus
TANC2.VII	1,228-1,358	TPR domain (C-terminal cap not included)
TANC2.VIII	1,244-1,542	full TPR domain + PolyP region
TANC2.IX	1,538-1,628	C-terminus
TANC2.X	1,623-1,990	C-terminus
CDKL5		
Region	Position	Domains
CDKL5.I	1-1,030	full-length protein
CDKL5.II	1-300	kinase domain only
CDKL5.III	301-615	C-terminal tail (first part)
CDKL5.IV	587-1,030	C-terminal tail (second part)
CDKL5.IVA	587-740	_
CDKL5.IVB	731-802	Poly-Lys region
CDKL5.IVC	796-1,030	_
PP1		
Region	Position	Domains
PP1	1-330	full-length protein

Table 8.1: CDKL5, PP1 and TANC2 regions investigated through Y2H assay.
First column: name of TANC2/CDKL5 clones; second column: clone mapping on protein sequence. A brief description of the relevant structural features for each clone is reported in third column.

At first, our attention was focused on PP1-TANC2 interaction. TANC2 was found to interact with PP1 by high-throughput mass spectrometry[298], although the interacting peptide was not indicated. Two highly conserved RVxF docking motifs map to TANC2 N-terminus, making this region a good candidate for interaction with the phosphatase. As expected, only the clones containing the N-terminal tail (TANC2.I = 1-340, TANC2.II =

1-845, TANC2.III = 1-1,227) were able to directly interact with PP1, resulting in yeast growth (Figure 8.2). No other TANC2 domain is necessary to PP1-TANC2 interaction (Figure S8.1).

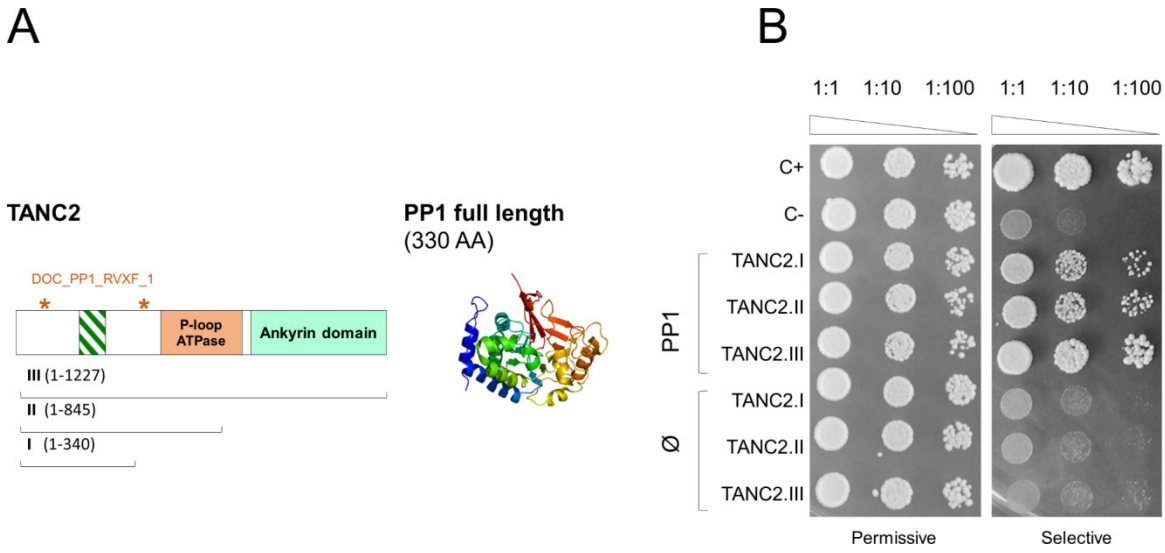


Figure 8.2: Yeast two-hybrid dissection for interaction between TANC2 and PP1.
Co-transformation of Y190 reporter yeast strain was performed, allowing the validation of PP1/TANC2 interactions. **A)**: graphic representation of tested TANC2 subclones (orange triangles = PP1 docking motif) and PP1 structure (pdb code 3EGG); **B)**: Serial dilutions of yeast cells were spotted on both permissive (left) and selective (right) media, and incubated for six days at 30 °C. C+ and C− are positive and negative controls. The image is representative of two independent experiments, each with three different clones analyzed. Clone autoactivation was evaluated co-transforming the tested subclones with an p. GBKT7 empty vector (indicated with ø).

As regard CDKL5-TANC2 interaction, it was detected between TANC2.VI clone (1228-1990), containing both TPR domain and C-terminus, and the full-length kinase sequence (CDKL5.I) (Figure 8.3, Figure S8.2). To further outline the sub-regions involved in TANC2-CDKL5 interaction, TANC2.VI and CDKL5 plasmids were mutagenized, obtaining shorter overlapping fragments (Table 8.1). CDKL5 binds the C-terminus of TANC2 (TANC2.X = 1623-1990) via its C-terminal tail (CDKL5.IV = 587-1030). Indeed, when only kinase domain was tested (CDKL5.II =1-300), the growth assay results were negative (Figures S8.3, S8.5). Interestingly, the TANC2-CDKL5 interaction seems to be mediated by a bipartite region, involving CDKL5. IV. A (587-740) and CDKL5.IV.C (796-1030), but not CDKL5.IV.B (731-802) (Figures 8.3, Figures S8.3, S8.4 and S8.5). Of note, the region excluded from the interaction is CDKL5.IV.B (731-802), containing the Poly-lysine segment, predicted as nuclear import signal (see below)[323].

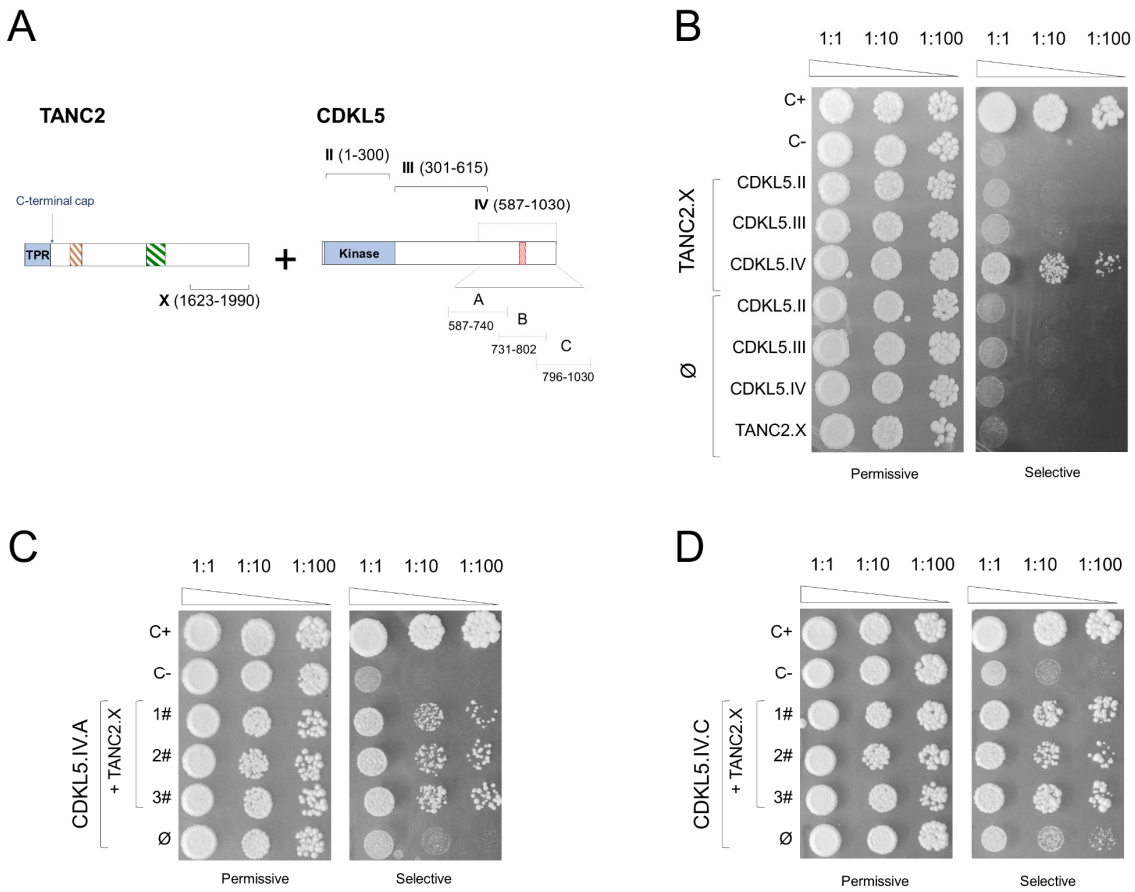


Figure 8.3: Fine mapping of TANC2-CDKL5 interaction with Y2H system.
A): graphic representation of tested TANC2 and CDKL5 subclones; **B)**: Drop-test results for the interaction of TANC2.X and different CDKL5 subclone (CDKL5.II, CDKL5.III, and CDKL5.IV) on permissive and selective media. Results of TANC2.X interaction with CDKL5.IV.A and CDKL5.IV.C are reported in **C)** and **D)** respectively. C+ and C− are positive and negative controls. The image is representative of two independent experiments, each with three different clones analyzed. Clone autoactivation was evaluated co-transforming the tested subclones with an appropriate empty vector (indicated with ø).

8.4.3 CDKL5 sequence analysis: C-terminus mediated protein interactions

CDKL5 presents an N-terminal serine/threonine kinase domain (1-297), homologous to the MAP kinases and cyclin-dependent kinases [324,325]. Aside from the kinase domain, CDKL5 is characterized by a long C-terminus, known to negatively regulate its catalytic activity and to mediate proper sub-cellular localization [324,326]. Despite this important role in protein function regulation, the CDKL5 C-terminal tail is still largely uncharacterized[324,325]. Thus, we performed a comprehensive sequence analysis of CDKL5 C-terminus aimed to the identification of conserved short linear motifs (SLiMs) likely to mediate protein interactions. Furthermore, we searched for literature and protein-protein interaction databases to collect information about known or candidate CDKL5 interactors. We

manually curated twenty-seven CDKL5 interacting proteins, eight interactors from publications, and nineteen identified by high-throughput screening in publicly available protein-protein interactions (PPI) databases. Low-throughput experimental evidence for CDKL5 interactors revealed the region 550-850 binds different interaction partners, including MECP2 [327] and DNMT1B [328], and PSD-95 [312]. This region is characterized by the presence of a Poly-lysine stretch (784 – 789), which ELM predicted to be enriched in nuclear localization signals (NLS). The assessment of candidate disordered binding regions with ANCHOR [321] highlighted the presence of several putative protein interaction sites (see Supplementary Table S8.3) that were employed during selection of candidate protein binding motifs. One example is the conserved APC/C-binding destruction motif (DEG_APCC_KENBOX2, 697-701), which map to an ANCHOR predicted binding site and may support the interaction with APC/C and likely regulating CDKL5 protein stability. Interesting, we predicted many other conserved motifs mapping to the C-terminus predicted interaction sites, which may regulate CDKL5 expression levels, either favoring its ubiquitin-dependent proteosomal degradation (DEG_Nend_UBRbox_1 and DEG_SPOP_SBC_1) or USP7-mediated de-ubiquitination (DOC_USP7_MATH_2 and DOC_USP7_UBL2_3), highlighting the relevance of CDKL5 availability regulation. For three of the interactors retrieved from databases, we predicted the putative interacting sites (see Figure 8.4, and Table 8.2). CDKL5 was found to interact with three SH3-domain containing proteins, ABL1 and FYN kinases, and GRB2 protein (see Figure 8.4, and Table 8.2). These interactions may recognize the SH3-domain binding motif mapping to the most distal part of the CDKL5 tail, which is specific for CDKL5 isoform1 (Figure 8.4, and Table 8.2) and is present only in *Hominidea* family (Figure 8.4, Figure S8.6). CDKL5-1 is expressed in the brain prevalently during foetal stages [314], and SH3-binding motif mediated interactions link the kinase to different pathways regulating cell proliferation, axon guidance and neuron projections[329–331], which are essential for synaptic plasticity and brain development[332].

Interactor	Biological process	Experimental evidence	Ref	CDKL5 region	CDKL5-interactor region
ABL1	Axon guidance, neuron projection extension	Peptide array	333	*LIG_SH3_2	*SH3 domain
CDH1	Cell adhesion, synapse assembly	PL-MS	334	_	_
DHX16	mRNA processing, mRNA splicing	HTS AC-MS	335	_	_
DHX38	mRNA processing, mRNA splicing	HTS AC-MS	335	_	_
DNMT1	Chromatin regulator, nervous system development	GST pull-down	328	650–850	1-290
DYRK1A	Synaptic plasticity	IF/coIP	326	#Ser308	1-497 (kinase domain)
FYN	Neuron death regulation, Wnt /PCP pathway regulation	Peptide array	333	*LIG_SH3_2	*SH3 domain
GPALPP1	Chaperon	HTS AC-MS	298	_	_
GRB2	Cell proliferation, axon guidance	Peptide array	333	*LIG_SH3_2	*SH3 domain
GTF2F1	Transcription regulation	HTS AC-MS	298	_	_
HDAC4	Chromatin regulator, nervous system development	CoIP	336	299-1030	_
HDGFRP2	Cell proliferation	HTS AC-MS	298	_	_
HDGFRP3	Cell proliferation, neuron projection extension, MT polymerization	HTS AC-MS	335	_	_
IQGAP1	Neuron projection extension	Y2H, IF/coIP	337	299-1,030	CHD domain, 160-260
KLHL20	Protein transport, Ubl conjugation pathway	HTS AC-MS	298	_	_
MACF1	Wnt signaling pathway	HTS AC-MS	298	_	_
MECP2	Transcription regulation, neuron maturation	GST pull-down	327	450–550, 551–650	202–486
MGMT	DNA damage repair	HTS AC-MS	338	_	_
NGL-1	Neuron projection extension	IF/coIP	311	_	550-556
PRKG2	Glutamate Receptor signaling, synaptic plasticity	HTS AC-MS	339	_	_
PSD-95	Glutamate Receptor signaling	Pull-down assay, IF	312	671–834	1–64
RABL6	Cell proliferation	HTS AC-MS	298	_	_
SAR1B	Protein transport	HTS AC-MS	298	_	_
SERPINB8	Protease inhibitor	HTS AC-MS	298	_	_
Shootin1	Neuron projection morphogenesis and extension	Y2H, IF/coIP	325	299-1,030	53-290
SKIV2L	mRNA processing	Y2H screening	340	_	_
UBAP2	positive regulation of gene expression	HTS AC-MS	298	_	_

Table 8.2: CDKL5 known and predicted interactors. For each interactor, the biological process, the detection method and the binding regions, both experimentally validated (in bold) and predicted from SLiMs analysis (labeled with *), are here listed. Linear motifs are named according to the ELM database nomenclature. # = CDKL5 Ser308, phosphorylated by DYRK1A.Y2H: Yeast two hybrid; Co-IP: Coimmunoprecipitation; HTS: High-Throughput System; AC: Affinity Capture; PL: Proximity Label; MS: Mass spectrometry.

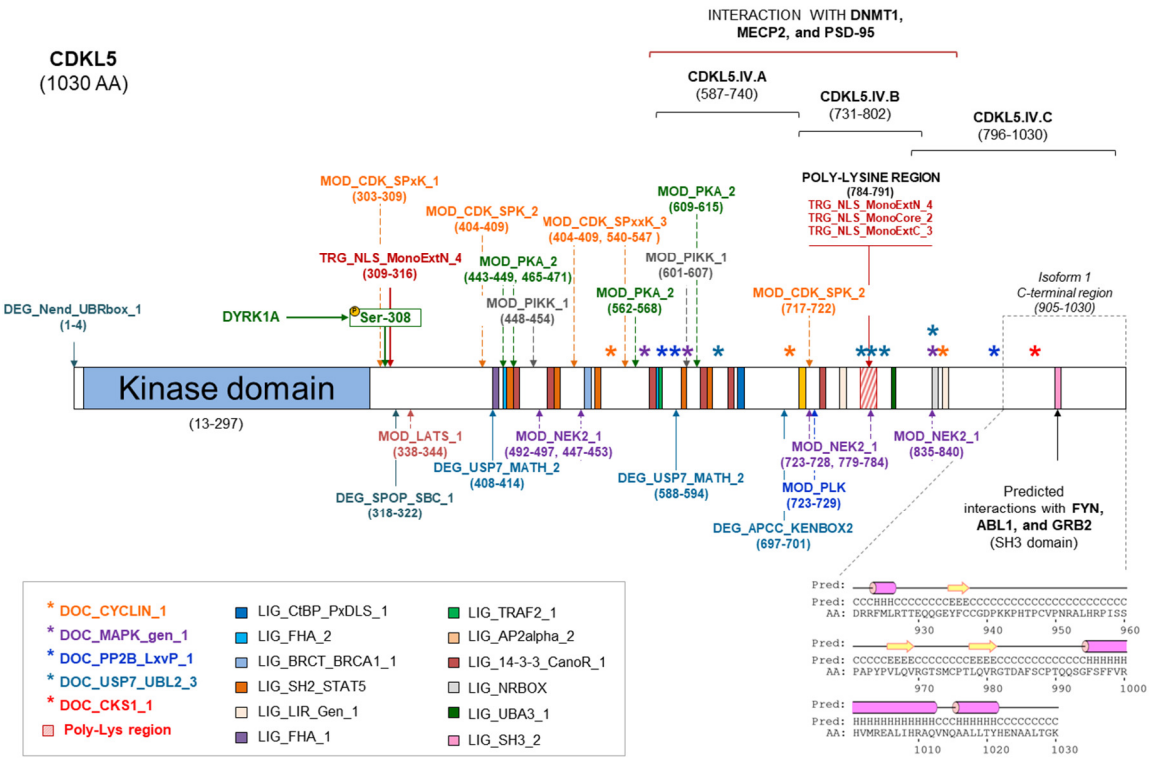


Figure 8.4: CDKL5 sequence analysis. Linear motifs are named according to the ELM database nomenclature.

8.4.4 TANC2 downregulation leads to an increase of CDKL5 levels

Our hypothesis relies on the assumption that TANC2 could modulate CDKL5 degradation, binding and targeting PP1 activity to the kinase. Accordingly, TANC2 down-regulation should correspond to an increasing in CDKL5 protein levels. After 72h incubation with siRNAs, TANC2 and CDKL5 were quantified by Western Blot. Welch's *t* test analysis revealed that only Hs_TANC2_1 siRNA significantly decreases the expression of the scaffold-protein (p-value = 0.03997, n = 7, statistical significance threshold: p-value < 0.05), though the halving of TANC2 level mean can be observed also in samples incubated with Hs_DKFZP564D166_3 (Figure 8.5B). On the contrary, in both tested conditions, a significant increase in CDKL5 availability (p-value = 0.01071 and p-value = 0.03329 for Hs_TANC2_1 and Hs_DKFZP564D166_3, respectively) was observed (Figure 8.5C), suggesting a role of TANC2 in CDKL5 degradation, which is reported to be mediated by PP1[313].

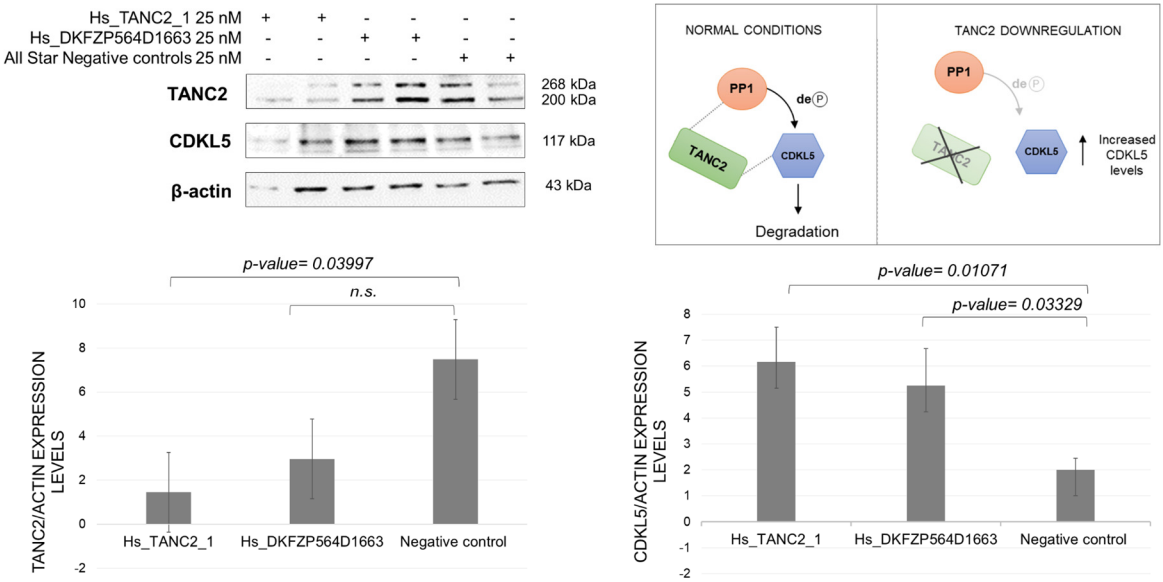


Figure 8.5: TANC2 silencing affect CDKL5 protein levels.
A) Representative western blot of cell lysates previously treated with: Hs_TANC2_1 (first three lanes), Hs_DKFZP564D1663 (from the forth to the sixth lanes), and negative control siRNA (last three lanes). **B)** Statistical analysis (Welch's t test) of TANC2 downregulation (statistical significance threshold: p-value < 0.05, n.s. = not significant). **C)** Statistical analysis (Welch's t test) of CDKL5 upregulation (statistical significance threshold: p-value < 0.05). **D)** TANC2-CDKL5-PP1 working model: when TANC2 is downregulated, a significant fraction of CDKL5 escapes PP1-dephosphorylation and subsequent degradation, resulting in an increasing of kinase availability.

8.5 Discussion

In this study, we demonstrated that TANC2 is able to interact with PP1 and CDKL5, possibly contributing to the kinase degradation. Both interactions with CDKL5 and PP1 involve intrinsically disordered regions (IDRs), characterized by the presence of different SLIMs and PTM sites[26]. IDRs are generally implicated in transient and potentially promiscuous interactions, and different mechanisms aim to avoid non-specific protein interactions, including on-site synthesis[26]. Proteins translated *in situ* are generally present in low concentration, their abundance rapidly increases after stimuli (e.g. extrasynaptic glutamate signaling) and display distinct phosphorylation dynamics[26]. In mature neurons, CDKL5 is upregulated in response to NMDAR-mediated membrane depolarization due to the localized activation of protein synthesis in the dendritic spines[313]. After a brief protein expression increasing, the kinase returns to basal levels as result of PP1-mediated dephosphorylation and consequent proteasome-dependent degradation[313].This suggests that both CDKL5 protein availability and phosphorylation levels need to be tightly regulated for the proper signaling of the synaptic plasticity[313]. The performed experiments

allowed us to identify part of the mechanism by which PP1 activity is targeted toward CDKL5, yielding its proteosomal degradation, though a more complex regulation mechanism is suggested by the presence of degrons and deubiquitinase target sites. We demonstrated that TANC2 scaffold downregulation causes an increasing in CDKL5 protein levels, supporting TANC2 involvement in influencing glutamatergic neuron signaling, by the modulation of synaptic plasticity[196]. We found that the interaction among the two post-synaptic proteins involves TANC2 C-terminus (1623-1990) and a bipartite motif of the kinase carboxyl-tail, which includes the known binding site of MECP2, DNMT1B and PSD95 and the specific region of the longer CDKL5-1 isoform (a.a. 904-1,030), which could increase CDKL5 binding affinity toward TANC2. Indeed, this latter region was previously suggested to contribute in regulating CDKL5 protein stability, as the longer CDKL5 isoform, detected in fetal brain[313], seems to be more actively degraded via proteasome pathway compared to CDKL5-2[341]. Splicing isoforms generally involve IDRs, resulting in tissue specific sets of short linear motifs and post-translational sites, allowing the recruitment of a same biological activity to different molecular contexts[26]. The fine tuning of CDKL5-1 function in fetal brains may affect synaptic plasticity and could require specific interactions, e.g. with TANC2 protein. Interestingly, Like CDKL5-1, the highest levels of TANC2 protein are present in embryonic stages, with a maximum in E18 in mouse[196]. Given its potential role in the regulation of CDKL5 protein levels, alteration of TANC2 activity would result in the over-expression of the post-synaptic kinase. The consequence of increased amount of CDKL5 on clinical phenotype is still poorly understood, and only few cases of CDKL5 duplications are available in literature[342]. Szafranski and colleagues [342] reported familial cases with short duplications (<1Mb) in *CDKL5*. Unlike patients with CDKL5 deletions, patients with CDKL5 duplications presented autistic traits, hyperactivity, development and language impairment, but not epilepsy. Authors suggested that the clinical phenotype could be linked to perturbation of synaptic plasticity, learning and neuron excitability[342,343]. Indeed, CDKL5 controls postsynaptic localization of GluN2B-containing NMDAR receptors, influencing their activity[343]. CDKL5 knock-out is characterized by an over-accumulation of the receptor subunits, resulting in enhancement of glutamatergic synaptic transmission and causing an increased seizure susceptibility[343]. Thus, it is conceivable that CDKL5 over-expression may down-regulate glutamate receptor subunit targeting to the PSD and, consequently, NMDAR activity. Interestingly, it has been demonstrated that GluN1 mutations, another subunit of NMDAR, reduce the channel trafficking to the synaptic membrane, and are

associated with non-syndromic intellectual disability without seizures[344]. This may suggest that deficits linked to these mutations could be more likely linked to deficit in NMDAR targeting rather than the alteration of receptor activity[344], which could be consistent with CDKL5 up-regulation, but also with clinical phenotypes associated with TANC2 malfunctioning[345]. Here, we found that the TANC2 scaffold protein could contribute to the regulation of postsynaptic CDKL5 availability, although further investigations are required for a better understanding of TANC2-PP1 complex activity. These include the identification of other synaptic substrates, which might contribute to CDKL5 expression levels. Additional validation of the proposed pathway could suggest TANC2 as a putative target of pharmacological interventions in CDKL5 duplication disorders.

9 Conclusions

In the last years, the next generation sequencing has been established as the state-of-the-art strategy for causative mutation discovery. NGS data handling and interpretation are still challenging for the vast majority of genetic studies[7]. The disease causality assignment of sequence variants often consists in a demanding process, requiring the integration of different sources for a comprehensive analysis[2]. An extensive part of my thesis deals about development of novel instruments for causative variants detection. On parallel, the identification of the molecular basis explaining the insurgence of different genetic diseases is also presented. The methods that our group developed during the first part of my Ph.D. have been assessed during my participation to the Critical Assessment of Genome Interpretation (CAGI). Our group was requested to implement different methods aimed at defining disease probability starting either from exome- or gene-targeted re-sequencing-derived data. These methods were employed for data interpretation in a board range of clinical conditions, from complex diseases, such as bipolar disorder and Crohn's Disease, to Mendelian disorders. Predictions were blinded to the clinical diagnosis, meaning that the analysis could not be driven by prior information about patients [19,346]. Our methods allowed to mostly assign a correct phenotype for each sequenced individual, correctly predicting the clinical phenotype-genotype associations. The prioritization of candidate disease-associated genes has played a critical role in variant selection, due to the huge number of genes and variants potentially involved in these pathologies. This specific strategy has been successfully used to identify the genes involved in the autism disease (ASD) and intellectual disability (ID) comorbidity, and to develop a diagnostic gene-panel aimed to assess their co-morbidity. Indeed, both ASD and ID are characterized by a highly genetic heterogeneity as well as overlapping clinical features, making single-gene testing insufficient for an accurate disease diagnosis[21]. The relevance of ASD/ID clinical diagnosis is almost clear, as it allows to shed light on molecular mechanism underlying the clinical phenotype, and has important implications for disease management and treatment. We considered several parameters for the candidate variant selection, including the pathogenicity prediction, frequency in publicly available database and segregation analysis. Through the ASD/ID panel screening, we assigned a clear molecular diagnosis to twenty-four of the tested patients (16,4%), with at least one likely pathogenic variant being detected

for additional twenty-three probands, highlighting the diagnostic value of a such developed AD/ID panel. Further evidence supporting causality between the likely pathogenic variations and the clinical phenotypes was provided by means of evaluation of variants effects on protein function/structure. Indeed, I performed a comprehensive analysis of selected ASD/ID -associated proteins to decipher variant pathogenicity, e.g. DYRK1A and EHMT1 proteins. The *in silico* analysis of the mutated proteins reveals that the 45,6% of the likely pathogenic variants to map in intrinsically disordered regions (IDRs). Consistently, an increasing number of studies shows that mutations in IDRs are associated with different pathologies, as they are often involved in protein function regulation[24,26]. Considered the important role played by intrinsically disordered protein in human diseases insurgence, I participated the update of DisProt, the most relevant database of protein disorder. I personally contributed to the manual curation of several proteins and dozens of intrinsically disorder regions. The annotation process involved the identification of IDR position and the association to the Pubmed ID related to supporting publications. The new release, DisProt 7.0, contains more hundreds of IDRs, eight-fold data respect to the earlier version, resulting in the most valuable resource for a better understanding of the structural disorder. Besides being extremely widespread, the intrinsically disorder proteins are essential components of regulative molecular pathways, including neuronal signaling[26]. In neurons, cellular signaling is tightly regulated by scaffold proteins, which play an important role in coordinating alternate signal paths, e.g. targeting glutamate receptors and their interactors to synaptic memenbrane[170]. Scaffolds are characterized by different modules, such as structural domain for protein-protein binding as well as short linear motif in IDRs[27]. Among the post-synaptic proteins encoded by the ASD/ID panel genes, TANC2 emerged for being a really promising scaffold protein with extended IDRs. Like other neuronal scaffold proteins, TANC2 seems to regulate neurotransmission and synaptic plasticity, although its function is still poorly understood. For all these reasons, I conducted an extensive *in silico* investigation of TANC2. The analysis was focused on its structural and functional characterization, including its close human paralog TANC1. Furthermore, the TANC family was also investigated at the pathway level, performing an interaction network characterization to elucidate TANCs collective role on neuronal pathways. This work provides the basis for interpretation of genetic variants found in TANC encoding genes[197]. Functional hypotheses emerged from bioinformatics study were experimentally validated. As molecular context is essential to understand macromolecules activity, I employed different experimental techniques to validate TANC2 predicted interactions with PP1 and

CDKL5 and the functional significance of the TANC2-CDKL5-PP1 interplay, as well. The results suggest that the expression of splice isoforms may play a significant role in modulating CDKL5-1 activity. Splicing isoforms generally involve IDRs, resulting in tissue specific sets of short linear motifs and post-translational sites, mediating isoform specific PPIs that occurs in a determined time window[26], such as interaction with TANC2. Indeed, we proved that TANC2 forms complexes with both CDKL5 isoform 1 and PP1 by means of its IDRs, and TANC2 downregulation corresponds to an increase of CDKL5 protein levels. Thus, TANC2 possibly links the phosphatase PP1 to its substrate CDKL5, allowing its dephosphorylation and subsequent degradation. It was demonstrated that this step is particularly critical for proper neuronal development[313], as increased levels of CDKL5 are associated with neurobehavioral and neurodevelopment features, e.g. ID and ASD traits[342]. These findings support the hypothesis that imbalance in CDKL5 levels affects neuronal function[342], suggesting TANC2 as a new putative pharmacological target for clinical conditions related to CDKL5 overexpression. Collectively, this study adds important elements to extend our understanding of the CDKL5 disorders. In particular, the roles that we propose for the IDRs contained in TANC2 and CDKL5 look promising. On the other side, this study clearly shows that the correct interpretation of next generation sequencing derived data is almost ready to be a routinely task in basic research, as well as a powerful tool for disease-variant association study. Indeed, determining the association between genetic variants and diseases is an interdisciplinary task, which cannot be limited to the identification of causative variants, but it should also provide the causality between variant, protein function alteration and disease.

10 Bibliography

1. Naidoo, N., Pawitan, Y., Soong, R., Cooper, D. N. & Ku, C.-S. Human genetics and genomics a decade after the release of the draft sequence of the human genome. *Hum. Genomics* **5,** 577–622 (2011).
2. Antonarakis, S. E. & Beckmann, J. S. Mendelian disorders deserve more attention. *Nat Rev Genet* **7,** 277–282 (2006).
3. Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. & Mardis, E. R. The Next-Generation Sequencing Revolution and Its Impact on Genomics. *Cell* **155,** 27–38 (2013).
4. Katsanis, S. H. & Katsanis, N. Molecular genetic testing and the future of clinical genomics. *Nat. Rev. Genet.* **14,** 415–426 (2013).
5. Eilbeck, K., Quinlan, A. & Yandell, M. Settling the score: variant prioritization and Mendelian disease. *Nat Rev Genet* **18,** 599–612 (2017).
6. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508,** 469–476 (2014).
7. Eilbeck, K., Quinlan, A. & Yandell, M. Settling the score: variant prioritization and Mendelian disease. *Nat Rev Genet* **18,** 599–612 (2017).
8. Hunt, R. C., Simhadri, V. L., Iandoli, M., Sauna, Z. E. & Kimchi-Sarfaty, C. Exposing synonymous mutations. *Trends Genet.* **30,** 308–321 (2014).
9. Pinto, D. *et al.* Convergence of genes and cellular pathways dysregulated in autism spectrum disorders. *Am. J. Hum. Genet.* **94,** 677–694 (2014).
10. Stefansson, H. *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* **505,** 361–366 (2014).

11. Pehlivan, D. *et al.* The role of combined SNV and CNV burden in patients with distal symmetric polyneuropathy. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **18,** 443–451 (2016).

12. Xue, Y., Ankala, A., Wilcox, W. R. & Hegde, M. R. Solving the molecular diagnostic testing conundrum for Mendelian disorders in the era of next-generation sequencing: single-gene, gene panel, or exome/genome sequencing. *Genet Med* **17,** 444–451 (2015).

13. Im, W. *et al.* Challenges in structural approaches to cell modeling. *J. Mol. Biol.* **428,** 2943–2964 (2016).

14. Bah, A. & Forman-Kay, J. D. Modulation of Intrinsically Disordered Protein Function by Post-translational Modifications. *J. Biol. Chem.* **291,** 6696–6705 (2016).

15. Davey, N. E. *et al.* Attributes of short linear motifs. *Mol. Biosyst.* **8,** 268–281 (2012).

16. Yi, S. *et al.* Functional variomics and network perturbation: connecting genotype to phenotype in cancer. *Nat Rev Genet* **18,** 395–410 (2017).

17. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nat Rev Genet* **18,** 551–562 (2017).

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

19. Daneshjou, R. *et al.* Working towards precision medicine: predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. *Hum. Mutat.* (2017). doi:10.1002/humu.23280

20. Szklarczyk D., von M. C. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible.

21. Soden, S. E. *et al.* Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. *Sci. Transl. Med.* **6,** 265ra168-265ra168 (2014).

22. De Rubeis, S. *et al.* Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature* **515,** 209–215 (2014).

23. Uversky, V. N. Wrecked regulation of intrinsically disordered proteins in diseases: pathogenicity of deregulated regulators. *Front. Mol. Biosci.* **1,** 6 (2014).

24. Tompa, P., Schad, E., Tantos, A. & Kalmar, L. Intrinsically disordered proteins: emerging interaction specialists. *Curr. Opin. Struct. Biol.* **35,** 49–59 (2015).

25. Potenza, E., Domenico, T. D., Walsh, I. & Tosatto, S. C. E. MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res.* **43,** D315–D320 (2015).

26. Babu, M. M. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem. Soc. Trans.* **44,** 1185–1200 (2016).

27. Cortese, M. S., Uversky, V. N. & Dunker, A. K. Intrinsic disorder in scaffold proteins: Getting more from less. *Prog. Biophys. Mol. Biol.* **98,** 85–106 (2008).

28. Tang, H. & Thomas, P. D. Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. *Genetics* **203,** 635 (2016).

29. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network Medicine: A Network-based Approach to Human Disease. *Nat. Rev. Genet.* **12,** 56–68 (2011).

30. Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Disease gene identification strategies for exome sequencing. *Eur J Hum Genet* **20,** 490–497 (2012).

31. Moreau, Y. & Tranchevent, L.-C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet* **13,** 523–536 (2012).

32. Xu, L.-M. *et al.* AutismKB: an evidence-based knowledgebase of autism genetics. *Nucleic Acids Res.* **40,** D1016–D1022 (2012).

33. Abrahams, B. S. *et al.* SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Mol. Autism* **4,** 36–36 (2013).

34. Tranchevent, L.-C. *et al.* Candidate gene prioritization with Endeavour. *Nucleic Acids Res.* **44,** W117–W121 (2016).

35. Orchard, S. *et al.* The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42,** D358–D363 (2014).

36. Chatr-aryamontri, A. *et al.* The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw1102

37. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44,** W90–W97 (2016).

38. Schaafsma, G. C. P. & Vihinen, M. VariSNP, A Benchmark Database for Variations From dbSNP. *Hum. Mutat.* **36,** 161–166 (2015).

39. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526,** 75–81 (2015).

40. Auer, P. L. *et al.* Guidelines for Large-Scale Sequence-Based Complex Trait Association Studies: Lessons Learned from the NHLBI Exome Sequencing Project. *Am. J. Hum. Genet.* **99,** 791–801 (2016).

41. Walsh, R. *et al.* Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet Med* **19,** 192–203 (2017).

42. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

43. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44,** D862–D868 (2016).

44. Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45,** D777–D783 (2017).

45. Stenson, P. D. *et al.* The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.* **136,** 665–677 (2017).

46. Cooper, D. N., Stenson, P. D. & Chuzhanova, N. A. The human gene mutation database (HGMD) and its exploitation in the study of mutational mechanisms. *Curr Protoc Bioinforma.* **Chapter 1,** (2006).

47. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* gku1205 (2014). doi:10.1093/nar/gku1205

48. Köhler, S. *et al.* The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res.* **42,** D966–D974 (2014).

49. Panchenko, A. R., Kondrashov, F. & Bryant, S. Prediction of functional sites by analysis of sequence and structure conservation. *Protein Sci.* **13,** 884–892 (2004).

50. Camps, M., Herman, A., Loh, E. & Loeb, L. A. Genetic Constraints on Protein Evolution. *Crit. Rev. Biochem. Mol. Biol.* **42,** 10.1080/10409230701597642 (2007).

51. Velankar, S. *et al.* SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res.* **41,** D483-489 (2013).

52. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. *Curr. Protoc. Hum. Genet. Editor. Board Jonathan Haines Al* **0 7,** Unit7.20-Unit7.20 (2013).

53. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39,** e118–e118 (2011).

54. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45,** D158–D169 (2017).

55. Sim, N.-L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40,** W452-457 (2012).

56. David, A. & Sternberg, M. J. E. The Contribution of Missense Mutations in Core and Rim Residues of Protein–Protein Interfaces to Human Disease. *J. Mol. Biol.* **427,** 2886–2898 (2015).

57. Hecht, M., Bromberg, Y. & Rost, B. Better prediction of functional effects for sequence variants. *BMC Genomics* **16,** S1–S1 (2015).

58. Giollo, M., Martin, A. J., Walsh, I., Ferrari, C. & Tosatto, S. C. NeEMO: a method using residue interaction networks to improve prediction of protein stability upon mutation. *BMC Genomics* **15,** S7–S7 (2014).

59. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res.* **33,** W382–W388 (2005).

60. Dehouck, Y. *et al.* Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinforma. Oxf. Engl.* **25,** 2537–2543 (2009).

61. Yin, S., Ding, F. & Dokholyan, N. V. Eris: an automated estimator of protein stability. *Nat Meth* **4,** 466–467 (2007).

62. Capriotti, E., Fariselli, P. & Casadio, R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* **33,** W306–W310 (2005).

63. Cheng, J., Randall, A. & Baldi, P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins Struct. Funct. Bioinforma.* **62,** 1125–1132 (2006).

64. Olatubosun, A., Väliaho, J., Härkönen, J., Thusberg, J. & Vihinen, M. PON-P: Integrated predictor for pathogenicity of missense variants. *Hum. Mutat.* **33,** 1166–1174 (2012).

65. Davydov, E. V. *et al.* Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput. Biol.* **6,** e1001025 (2010).

66. Gulko, B., Gronau, I., Hubisz, M. J. & Siepel, A. Probabilities of Fitness Consequences for Point Mutations Across the Human Genome. *bioRxiv* (2014). doi:10.1101/006825

67. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).

68. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w(1118); iso-2; iso-3. *Fly (Austin)* **6,** 80–92 (2012).

69. Aken, B. L. *et al.* Ensembl 2017. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw1104

70. Chadwick, L. H. The NIH Roadmap Epigenomics Program data resource. *Epigenomics* **4,** 317–324 (2012).

71. Gaudet, P. *et al.* The neXtProt knowledgebase on human proteins: 2017 update. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw1062

72. Schneider, A., Dessimoz, C. & Gonnet, G. H. OMA Browser--exploring orthologous relations across 352 complete genomes. *Bioinformatics* **23,** 2180–2182 (2007).

73. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).

74. Yang, Z. & Rannala, B. Molecular phylogenetics: principles and practice. *Nat Rev Genet* **13,** 303–314 (2012).

75. Yu, H. *et al.* Annotation Transfer Between Genomes: Protein–Protein Interologs and Protein–DNA Regulogs. *Genome Res.* **14,** 1107–1118 (2004).

76. Finn, R. D. *et al.* InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Res.* **45,** D190–D199 (2017).

77. Buchan, D. W. A., Minneci, F., Nugent, T. C. O., Bryson, K. & Jones, D. T. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Res.* **41,** W349–W357 (2013).

78. Piovesan, D., Walsh, I., Minervini, G. & Tosatto, S. C. E. FELLS: fast estimator of latent local structure. *Bioinformatics* **33,** 1889–1891 (2017).

79. Best, R. B. Computational and theoretical advances in studies of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **42,** 147–154 (2017).

80. Necci, M., Piovesan, D., Dosztányi, Z. & Tosatto, S. C. E. MobiDB-lite: Fast and highly specific consensus prediction of intrinsic disorder in proteins. *Bioinformatics* btx015 (2017). doi:10.1093/bioinformatics/btx015

81. Dinkel, H. *et al.* The eukaryotic linear motif resource ELM: 10 years and counting. *Nucleic Acids Res.* **42,** D259–D266 (2014).

82. Gibson, T. J., Dinkel, H., Roey, K. V. & Diella, F. Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad. *Cell Commun. Signal.* **13,** 42 (2015).

83. Jehl, P., Manguy, J., Shields, D. C., Higgins, D. G. & Davey, N. E. ProViz—a web-based visualization tool to investigate the functional and evolutionary features of protein sequences. *Nucleic Acids Res.* **44,** W11–W15 (2016).

84. Martí-Renom, M. A. *et al.* Comparative Protein Structure Modeling of Genes and Genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 291–325 (2000).

85. Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5,** 823–826 (1986).

86. Alva, V., Nam, S.-Z., Söding, J. & Lupas, A. N. The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucleic Acids Res.* **44,** W410–W415 (2016).

87. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44,** D279–D285 (2016).

88. Letunic, I., Doerks, T. & Bork, P. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res.* **43,** D257–D260 (2015).

89. Eswar, N. *et al.* Comparative Protein Structure Modeling Using Modeller. in *Current Protocols in Bioinformatics* (eds. Bateman, A., Pearson, W. R., Stein, L. D., Stormo, G. D. & Yates, J. R.) 5.6.1-5.6.30 (John Wiley & Sons, Inc., 2006).

90. Benkert, P., Tosatto, S. C. E. & Schomburg, D. QMEAN: A comprehensive scoring function for model quality assessment. *Proteins Struct. Funct. Bioinforma.* **71,** 261–277 (2008).

91. Rigsby, R. E. & Parker, A. B. Using the PyMOL application to reinforce visual understanding of protein structure. *Biochem. Mol. Biol. Educ.* **44,** 433–437 (2016).

92. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **38,** W529–W533 (2010).

93. Walsh, I. *et al.* Bluues server: electrostatic properties of wild-type and mutated protein structures. *Bioinformatics* **28,** 2189–2190 (2012).

94. Leonardi, E. *et al.* A computational model of the LGI1 protein suggests a common binding site for ADAM proteins. *PloS One* **6,** e18142 (2011).

95. Kuzmanov, U. & Emili, A. Protein-protein interaction networks: probing disease mechanisms using model systems. *Genome Med.* **5,** 37 (2013).

96. Rao, V. S., Srinivas, K., Sujini, G. N. & Kumar, G. N. S. Protein-Protein Interaction Detection: Methods and Analysis. *Int. J. Proteomics* **2014,** 147648 (2014).

97. Keskin, O., Gursoy, A., Ma, B. & Nussinov, R. Principles of Protein−Protein Interactions: What are the Preferred Ways For Proteins To Interact? *Chem. Rev.* **108,** 1225–1244 (2008).

98. Davey, N. E., Cyert, M. S. & Moses, A. M. Short linear motifs – ex nihilo evolution of protein regulation. *Cell Commun. Signal. CCS* **13,** 43 (2015).

99. Berggård, T., Linse, S. & James, P. Methods for the detection and analysis of protein–protein interactions. *PROTEOMICS* **7,** 2833–2842 (2007).

100. Costes, S. V. *et al.* Automatic and Quantitative Measurement of Protein-Protein Colocalization in Live Cells. *Biophys. J.* **86,** 3993–4003

101. Galletta, B. J. & Rusan, N. M. A Yeast Two-Hybrid approach for probing protein-protein interactions at the centrosome. *Methods Cell Biol.* **129,** 251–277 (2015).

102. Ashley, E. A. The precision medicine initiative: a new national effort. *Jama* **313,** 2119–2120 (2015).

103. Ashley, E. A. *et al.* Clinical evaluation incorporating a personal genome. *Lancet* **375,** 1525–1535 (2010).

104. Brown, T. L. & Meloche, T. M. Exome sequencing a review of new strategies for rare genomic disease research. *Genomics* **108,** 109–114 (2016).

105. Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K. J. & Altman, R. B. Bioinformatics challenges for personalized medicine. *Bioinformatics* **27,** 1741–1748 (2011).

106. National Academies of Sciences, E., and Medicine. *Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop*. (The National Academies Press, 2016). doi:10.17226/21915

107. Donoho, D. 50 years of Data Science. in (2015).

108. Bell, R. M. & Koren, Y. Lessons from the Netflix prize challenge. *Acm Sigkdd Explor. Newsl.* **9,** 75–79 (2007).

109. Morgan, A. A. *et al.* Overview of BioCreative II gene normalization. *Genome Biol.* **9,** S3–S3 (2008).

110. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)--round x. *Proteins* **82 Suppl 2,** 1–6 (2014).

111. Thrun, S. *et al.* Stanley: The robot that won the DARPA Grand Challenge. *J. Field Robot.* **23,** 661–692 (2006).

112. Walker, M. A., Passonneau, R. & Boland, J. E. Quantitative and qualitative evaluation of DARPA Communicator spoken dialogue systems. in 515–522 (Association for Computational Linguistics, 2001).

113. Cho, J. H. The genetics and immunopathogenesis of inflammatory bowel disease. *Nat Rev Immunol* **8,** 458–466 (2008).

114. Halfvarson, J., Bodin, L., Tysk, C., Lindberg, E. & Järnerot, G. Inflammatory bowel disease in a Swedish twin cohort: a long-term follow-up of concordance and clinical characteristics. *Gastroenterology* **124,** 1767–1773 (2003).

115. Uhlig, H. H. *et al.* The diagnostic approach to monogenic very early onset inflammatory bowel disease. *Gastroenterology* **147,** 990–1007 (2014).

116. Ellinghaus, D. *et al.* Association Between Variants of PRDM1 and NDP52 and Crohn's Disease, Based on Exome Sequencing and Functional Studies. *Gastroenterology* (2013). doi:10.1053/j.gastro.2013.04.040

117. Craddock, N. & Sklar, P. Genetics of bipolar disorder. *The Lancet* **381,** 1654–1662 (2013).

118. Craddock, N. & Jones, I. Genetics of bipolar disorder. *J. Med. Genet.* **36,** 585–594 (1999).

119. Monson, E. T. *et al.* Assessment of Whole-Exome Sequence Data in Attempted Suicide within a Bipolar Disorder Cohort. *Mol. Neuropsychiatry* **3,** 1–11 (2017).

120. BAUER, K. A. Recent progress in anticoagulant therapy: oral direct inhibitors of thrombin and factor Xa. *J. Thromb. Haemost.* **9,** 12–19 (2011).

121. Budnitz, D. S., Lovegrove, M. C., Shehab, N. & Richards, C. L. Emergency Hospitalizations for Adverse Drug Events in Older Americans. *N. Engl. J. Med.* **365,** 2002–2012 (2011).

122. The International Warfarin Pharmacogenetics Consortium. Estimation of the Warfarin Dose with Clinical and Pharmacogenetic Data. *N. Engl. J. Med.* **360,** 753–764 (2009).

123. Daneshjou, R. *et al.* Genetic variant in folate homeostasis is associated with lower warfarin dose in African Americans. *Blood* **124,** 2298–2305 (2014).

124. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

125. Wray, N. R., Yang, J., Goddard, M. E. & Visscher, P. M. The genetic interpretation of area under the ROC curve in genomic profiling. *PLoS Genet.* **6,** e1000864 (2010).

126. Nomogram for Bayes's Theorem. *N. Engl. J. Med.* **293,** 257–257 (1975).

127. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29 (2000).

128. Yang, H., Robinson, P. N. & Wang, K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods* **12,** 841–843 (2015).

129. Bromberg, Y. & Rost, B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* **35,** 3823–3835 (2007).

130. Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L. & Casadio, R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.* **30,** 1237–1244 (2009).

131. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).

132. Niroula, A., Urolagin, S. & Vihinen, M. PON-P2: Prediction Method for Fast and Reliable Identification of Harmful Variants. *PLoS ONE* **10,** e0117380 (2015).

133. Ribeiro, M. T., Singh, S. & Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. in 1135–1144 (ACM, 2016).

134. Altshuler, D. M. *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).

135. Rosenberg, N. A. *et al.* Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11,** 356–366 (2010).

136. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).

137. Halevy, A., Norvig, P. & Pereira, F. The unreasonable effectiveness of data. *IEEE Intell. Syst.* **24,** 8–12 (2009).

138. McNutt, M. #IAmAResearchParasite. *Science* **351,** 1005 (2016).

139. Sosnay, P. R. *et al.* Applying Cystic Fibrosis Transmembrane Conductance Regulator Genetics and CFTR2 Data to Facilitate Diagnoses. *Cyst. Fibros. Found. Consens. Guidel. Diagn. Cyst. Fibros.* **181, Supplement,** S27–S32.e1 (2017).

140. Schulz, W. L., Tormey, C. A. & Torres, R. Computational Approach to Annotating Variants of Unknown Significance in Clinical Next Generation Sequencing. *Lab. Med.* **46,** 285–289 (2015).

141. Lee, H. *et al.* Clinical Exome Sequencing for Genetic Identification of Rare Mendelian Disorders. *JAMA* **312,** 1880–1887 (2014).

142. Posey, J. E. *et al.* Molecular Diagnostic Experience of Whole-Exome Sequencing in Adult Patients. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **18,** 678–685 (2016).

143. Vassy, J. L. *et al.* The MedSeq Project: a randomized trial of integrating whole genome sequencing into clinical medicine. *Trials* **15,** 85 (2014).

144. Richards, S. *et al.* Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **17,** 405–424 (2015).

145. Hill, S. M. *et al.* Inferring causal molecular networks: empirical assessment through a community-based effort. *Nat. Methods* **advance online publication,** (2016).

146. Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J. & Williamson, R. C. Estimating the Support of a High-Dimensional Distribution. *Neural Comput.* **13,** 1443–1471 (2001).

147. The 1000 Genomes Project Consortium. A map of human genome variation from population scale sequencing. *Nature* **467,** 1061–1073 (2010).

148. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44,** D862–D868 (2016).

149. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38,** e164 (2010).

150. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).

151. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–3814 (2003).

152. El-Fishawy, P. Common Disease-Rare Variant Hypothesis. in *Encyclopedia of Autism Spectrum Disorders* (ed. Volkmar, F. R.) 720–722 (Springer New York, 2013). doi:10.1007/978-1-4419-1698-3_1997

153. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

154. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33,** D514–D517 (2005).

155. Kirmani, S. & Young, W. F. Hereditary Paraganglioma-Pheochromocytoma Syndromes. in *GeneReviews* (eds. Pagon, R. A. et al.) (University of Washington, Seattle, 1993).

156. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17,** 122 (2016).

157. Yue, P., Melamud, E. & Moult, J. SNPs3D: Candidate gene and SNP selection for association studies. *BMC Bioinformatics* **7,** 166–166 (2006).

158. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc.* **4,** 1073–1081 (2009).

159. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).

160. Stenson, P. D. *et al.* Human Gene Mutation Database (HGMD®): 2003 update. *Hum. Mutat.* **21,** 577–581 (2003).

161. Tricarico, R. *et al.* Assessment of the InSiGHT Interpretation Criteria for the Clinical Classification of 24 MLH1 and MSH2 Gene Variants. *Hum. Mutat.* **38,** 64–77 (2017).

162. Manrai, A. K. *et al.* Genetic Misdiagnoses and the Potential for Health Disparities. *N. Engl. J. Med.* **375,** 655–665 (2016).

163. Walsh, R. *et al.* Reassessment of Mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genet. Med.* **19,** 192–203 (2017).

164. Amendola, L. M. *et al.* Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. *Am. J. Hum. Genet.* **98,** 1067–1076 (2016).

165. Garber, K. B. *et al.* Reassessment of Genomic Sequence Variation to Harmonize Interpretation for Personalized Medicine. *Am. J. Hum. Genet.* **99,** 1140–1149 (2016).

166. Stein, J. L., Parikshak, N. N. & Geschwind, D. H. Rare Inherited Variation in Autism: Beginning to See the Forest and a Few Trees. *Neuron* **77,** 209–211 (2013).

167. Srivastava, A. K. & Schwartz, C. E. Intellectual disability and autism spectrum disorders: Causal genes and molecular mechanisms. *Common Mech. Intellect. Disabil. Chall. Transl. Outlooks* **46,** 161–174 (2014).

168. An, J. Y. *et al.* Towards a molecular characterization of autism spectrum disorders: an exome sequencing and systems approach. *Transl. Psychiatry* **4,** e394 (2014).

169. Krumm, N., O'Roak, B. J., Shendure, J. & Eichler, E. E. A de novo convergence of autism genetics and molecular neuroscience. *Trends Neurosci.* **37,** 95–105 (2014).

170. Gao, C., Tronson, N. C. & Radulovic, J. Modulation of behavior by scaffolding proteins of the post-synaptic density. *Neurobiol. Learn. Mem.* **105,** 3–12 (2013).

171. Sun, Y. *et al.* Next-Generation Diagnostics: Gene Panel, Exome, or Whole Genome? *Hum. Mutat.* **36,** 648–655 (2015).

172. Yang, H. & Wang, K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat. Protoc.* **10,** 1556–1566 (2015).

173. Kobayashi, Y. *et al.* Pathogenic variant burden in the ExAC database: an empirical approach to evaluating population data for clinical variant interpretation. *Genome Med.* **9,** 13 (2017).

174. Rosenthal, E. A. *et al.* Association between absolute neutrophil count and variation at TCIRG1: the NHLBI Exome Sequencing Project. *Genet. Epidemiol.* **40,** 470–474 (2016).

175. Schwarz, J. M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Meth* **7,** 575–576 (2010).

176. Shihab, H. A. *et al.* Ranking non-synonymous single nucleotide polymorphisms based on disease concepts. *Hum. Genomics* **8,** 11–11 (2014).

177. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum. Mol. Genet.* **24,** 2125–2137 (2015).

178. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20,** 110–121 (2010).

179. Garber, M. *et al.* Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics* **25,** i54–i62 (2009).

180. The UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43,** D204–D212 (2015).

181. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30,** 1236–1240 (2014).

182. Rose, P. W. *et al.* The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.* **45,** D271–D281 (2017).

183. Gibson, T. J., Dinkel, H., Roey, K. V. & Diella, F. Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad. *Cell Commun. Signal.* **13,** 42 (2015).

184. Anderson, L. L. *et al.* Antiepileptic Activity of Preferential Inhibitors of Persistent Sodium Current. *Epilepsia* **55,** 1274–1283 (2014).

185. Tu, Y.-C. & Kuo, C.-C. The differential contribution of GluN1 and GluN2 to the gating operation of the NMDA receptor channel. *Pflüg. Arch. - Eur. J. Physiol.* **467,** 1899–1917 (2015).

186. Carvill, G. L. *et al.* Mutations in the GABA Transporter SLC6A1 Cause Epilepsy with Myoclonic-Atonic Seizures. *Am. J. Hum. Genet.* **96,** 808–815 (2015).

187. Gatto, C. L. & Broadie, K. Genetic Controls Balancing Excitatory and Inhibitory Synaptogenesis in Neurodevelopmental Disorder Models. *Front. Synaptic Neurosci.* **2,** 4 (2010).

188. Gao, Q. & McNally, E. M. The Dystrophin Complex: structure, function and implications for therapy. *Compr. Physiol.* **5,** 1223–1239 (2015).

189. Mehler, M. F. Brain dystrophin, neurogenetics and mental retardation. *Brain Res. Rev.* **32,** 277–307 (2000).

190. Srour, M. *et al.* An Instructive Case of an 8-Year-Old Boy With Intellectual Disability. *Case Stud. Pediatr. Neurol. Number 3* **15,** 154–155 (2008).

191. Banihani, R. *et al.* A Novel Mutation in DMD (c.10797+5G>A) Causes Becker Muscular Dystrophy Associated with Intellectual Disability. *J. Dev. Behav. Pediatr.* **37,** (2016).

192. de Brouwer, A. P. *et al.* A 3-base pair deletion, c.9711_9713del, in DMD results in intellectual disability without muscular dystrophy. *Eur. J. Hum. Genet.* **22,** 480–485 (2014).

193. Manzini, M. C. *et al.* CC2D1A Regulates Human Intellectual and Social Function as well as NF-κB Signaling Homeostasis. *Cell Rep.* **8,** 647–655 (2014).

194. Vulto-van Silfhout, A. T. *et al.* Mutations Affecting the SAND Domain of DEAF1 Cause Intellectual Disability with Severe Speech Impairment and Behavioral Problems. *Am. J. Hum. Genet.* **94,** 649–661 (2014).

195. Martin, E. A. *et al.* The intellectual disability gene Kirrel3 regulates target-specific mossy fiber synapse development in the hippocampus. *eLife* **4,** e09395 (2015).

196. Han, S. *et al.* Regulation of Dendritic Spines, Spatial Memory, and Embryonic Development by the TANC Family of PSD-95-Interacting Proteins. *J. Neurosci.* **30,** 15102–15112 (2010).

197. Gasparini, A., Tosatto, S. C. E., Murgia, A. & Leonardi, E. Dynamic scaffolds for neuronal signaling: in silico analysis of the TANC protein family. *Sci. Rep.* **7,** 6829 (2017).

198. Marangi, G. & Zollino, M. Pitt–Hopkins Syndrome and Differential Diagnosis: A Molecular and Clinical Challenge. *J. Pediatr. Genet.* **4,** 168–176 (2015).

199. Swanger, S. A. *et al.* Mechanistic Insight into NMDA Receptor Dysregulation by Rare Variants in the GluN2A and GluN2B Agonist Binding Domains. *Am. J. Hum. Genet.* **99,** 1261–1280 (2016).

200. Inui, K. *et al.* Mutational analysis of MECP2 in Japanese patients with atypical Rett syndrome. *Brain Dev.* **23,** 212–215 (2001).

201. Lundvall, M., Samuelsson, L. & Kyllerman, M. Male Rett Phenotypes in T158M and R294X MeCP2-mutations. *Neuropediatrics* **37,** 296–301 (2007).

202. Leoyklang, P. *et al.* Heterozygous nonsense mutation SATB2 associated with cleft palate, osteoporosis, and cognitive defects. *Hum. Mutat.* **28,** 732–738 (2007).

203. Ji, J. *et al.* DYRK1A haploinsufficiency causes a new recognizable syndrome with microcephaly, intellectual disability, speech impairment, and distinct facies. *Eur J Hum Genet* **23,** 1473–1481 (2015).

204. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).

205. Soundararajan, M. *et al.* Structures of Down Syndrome Kinases, DYRKs, Reveal Mechanisms of Kinase Activation and Substrate Recognition. *Structure* **21,** 986–996 (2013).

206. Balemans, M. C. M. *et al.* Reduced Euchromatin histone methyltransferase 1 causes developmental delay, hypotonia, and cranial abnormalities associated with increased bone gene expression in Kleefstra syndrome mice. *Dev. Biol.* **386,** 395–407 (2014).

207. Bart Martens, M. *et al.* Euchromatin histone methyltransferase 1 regulates cortical neuronal network development. *Sci. Rep.* **6,** 35756 (2016).

208. Wu, H. *et al.* Structural Biology of Human H3K9 Methyltransferases. *PLOS ONE* **5,** e8570 (2010).

209. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

210. Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu. Rev. Biochem.* **83,** 553–584 (2014).

211. van der Lee, R. *et al.* Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* **114,** 6589–6631 (2014).

212. Habchi, J., Tompa, P., Longhi, S. & Uversky, V. N. Introducing protein intrinsic disorder. *Chem. Rev.* **114,** 6561–6588 (2014).

213. Xie, H. *et al.* Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J. Proteome Res.* **6,** 1882–1898 (2007).

214. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu. Rev. Biophys.* **37,** 215–246 (2008).

215. Metallo, S. J. Intrinsically disordered proteins are potential drug targets. *Curr. Opin. Chem. Biol.* **14,** 481–488 (2010).

216. Varadi, M. *et al.* pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. *Nucleic Acids Res.* **42,** D326–D335 (2014).

217. Tompa, P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.* **579,** 3346–3354 (2005).

218. Wright, P. E. & Dyson, H. J. Linking folding and binding. *Curr. Opin. Struct. Biol.* **19,** 31–38 (2009).

219. Iakoucheva, L. M. *et al.* The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* **32,** 1037–1049 (2004).

220. Tompa, P. Multisteric Regulation by Structural Disorder in Modular Signaling Proteins: An Extension of the Concept of Allostery. *Chem. Rev.* **114,** 6715–6732 (2014).

221. Vucetic, S. *et al.* DisProt: a database of protein disorder. *Bioinformatics* **21,** 137–140 (2005).

222. Sickmeier, M. *et al.* DisProt: the Database of Disordered Proteins. *Nucleic Acids Res.* **35,** D786-793 (2007).

223. Oates, M. E. *et al.* D(2)P(2): database of disordered protein predictions. *Nucleic Acids Res.* **41,** D508–D516 (2013).

224. Potenza, E., Di Domenico, T., Walsh, I. & Tosatto, S. C. E. MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res.* **43,** D315-320 (2015).

225. Fukuchi, S. *et al.* IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature. *Nucleic Acids Res.* **40,** D507-511 (2012).

226. Mohan, A. *et al.* Analysis of Molecular Recognition Features (MoRFs). *J. Mol. Biol.* **362,** 1043–1059 (2006).

227. Dinkel, H. *et al.* ELM 2016--data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Res.* **44,** D294-300 (2016).

228. Tompa, P. *et al.* Close encounters of the third kind: disordered domains and the interactions of proteins. *BioEssays News Rev. Mol. Cell. Dev. Biol.* **31,** 328–335 (2009).

229. Receveur-Bréchot, V., Bourhis, J.-M., Uversky, V. N., Canard, B. & Longhi, S. Assessing protein disorder and induced folding. *Proteins* **62,** 24–45 (2006).

230. Kosol, S., Contreras-Martos, S., Cedeño, C. & Tompa, P. Structural Characterization of Intrinsically Disordered Proteins by NMR Spectroscopy. *Molecules* **18,** 10802–10828 (2013).

231. Felli, I. C. & Pierattelli, R. Recent progress in NMR spectroscopy: Toward the study of intrinsically disordered proteins of increasing size and complexity. *IUBMB Life* **64,** 473–481 (2012).

232. Theillet, F.-X. *et al.* Structural disorder of monomeric α-synuclein persists in mammalian cells. *Nature* **530,** 45–50 (2016).

233. Schuler, B., Soranno, A., Hofmann, H. & Nettels, D. Single-Molecule FRET Spectroscopy and the Polymer Physics of Unfolded and Intrinsically Disordered Proteins. *Annu. Rev. Biophys.* **45,** 207–231 (2016).

234. Kaltashov, I. A., Bobst, C. E. & Abzalimov, R. R. Mass spectrometry-based methods to study protein architecture and dynamics. *Protein Sci. Publ. Protein Soc.* **22,** 530–544 (2013).

235. Borysik, A. J., Kovacs, D., Guharoy, M. & Tompa, P. Ensemble Methods Enable a New Definition for the Solution to Gas-Phase Transfer of Intrinsically Disordered Proteins. *J. Am. Chem. Soc.* **137,** 13807–13817 (2015).

236. Miyagi, A. *et al.* Visualization of Intrinsically Disordered Regions of Proteins by High-Speed Atomic Force Microscopy. *ChemPhysChem* **9,** 1859–1866 (2008).

237. Jakob, U., Kriwacki, R. & Uversky, V. N. Conditionally and Transiently Disordered Proteins: Awakening Cryptic Disorder To Regulate Protein Function. *Chem. Rev.* (2014). doi:10.1021/cr400459c

238. DeForte, S. & Uversky, V. N. Resolving the ambiguity: Making sense of intrinsic disorder when PDB structures disagree. *Protein Sci. Publ. Protein Soc.* **25,** 676–688 (2016).

239. Blocquel, D. *et al.* Coiled-coil deformations in crystal structures: the measles virus phosphoprotein multimerization domain as an illustrative example. *Acta Crystallogr. Sect. D* **70,** 1589–1603 (2014).

240. Sterckx, Y. G. J. *et al.* Small-Angle X-Ray Scattering- and Nuclear Magnetic Resonance-Derived Conformational Ensemble of the Highly Flexible Antitoxin PaaA2. *Structure* **22,** 854–865 (2014).

241. Aznauryan, M. *et al.* Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E5389–E5398 (2016).

242. The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43,** D1049–D1056 (2015).

243. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradović, Z. Intrinsic Disorder and Protein Function. *Biochemistry (Mosc.)* **41,** 6573–6582 (2002).

244. Stein, J. L., Parikshak, N. N. & Geschwind, D. H. Rare inherited variation in autism: beginning to see the forest and a few trees. *Neuron* **77,** 209–211 (2013).

245. Fromer, M. *et al.* De novo mutations in schizophrenia implicate synaptic networks. *Nature* **506,** 179–184 (2014).

246. Gilman, S. R. *et al.* Rare De Novo Variants Associated with Autism Implicate a Large Functional Network of Genes Involved in Formation and Function of Synapses. *Neuron* **70,** 898–907 (2011).

247. Krumm, N., O'Roak, B. J., Shendure, J. & Eichler, E. E. A de novo convergence of autism genetics and molecular neuroscience. *Trends Neurosci.* **37,** 95–105 (2014).

248. Ting, J. T., Peça, J. & Feng, G. Functional consequences of mutations in postsynaptic scaffolding proteins and relevance to psychiatric disorders. *Annu. Rev. Neurosci.* **35,** 49–71 (2012).

249. Han, S. *et al.* Regulation of dendritic spines, spatial memory, and embryonic development by the TANC family of PSD-95-interacting proteins. *J. Neurosci. Off. J. Soc. Neurosci.* **30,** 15102–15112 (2010).

250. Suzuki, T. *et al.* A novel scaffold protein, TANC, possibly a rat homolog of *Drosophila* rolling pebbles (rols), forms a multiprotein complex with various postsynaptic density proteins. *Eur. J. Neurosci.* **21,** 339–350 (2005).

251. de Ligt, J., Veltman, J. A. & Vissers, L. E. Point mutations as a source of de novo genetic disease. *Curr. Opin. Genet. Dev.* doi:10.1016/j.gde.2013.01.007

252. Iossifov, I. *et al.* De Novo Gene Disruptions in Children on the Autistic Spectrum. *Neuron* **74,** 285–299 (2012).

253. Granot-Hershkovitz, E. *et al.* Complex chromosomal rearrangement in a girl with psychomotor-retardation and a de novo inversion: inv(2)(p15;q24.2). *Am. J. Med. Genet. A.* **155,** 1825–1832 (2011).

254. Leipe, D. D., Koonin, E. V. & Aravind, L. STAND, a Class of P-Loop NTPases Including Animal and Plant Regulators of Programmed Cell Death: Multiple, Complex Domain Architectures, Unusual Phyletic Patterns, and Evolution by Horizontal Gene Transfer. *J. Mol. Biol.* **343,** 1–28 (2004).

255. Yuan, S., Topf, M., Reubold, T. F., Eschenburg, S. & Akey, C. W. Changes in Apaf-1 conformation that drive apoptosome assembly. *Biochemistry (Mosc.)* **52,** 2319–2327 (2013).

256. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).

257. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25,** 1189–1191 (2009).

258. Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252,** 1162–1164 (1991).

259. Delorenzi, M. & Speed, T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinforma. Oxf. Engl.* **18,** 617–625 (2002).

260. Bartoli, L., Fariselli, P., Krogh, A. & Casadio, R. CCHMM_PROF: a HMM-based coiled-coil predictor with evolutionary information. *Bioinformatics* **25,** 2757–2763 (2009).

261. Karpenahalli, M. R., Lupas, A. N. & Söding, J. TPRpred: a tool for prediction of TPR-, PPR-and SEL1-like repeats from protein sequences. *BMC Bioinformatics* **8,** 2 (2007).

262. Szklarczyk, R. & Heringa, J. Tracking repeats using significance and transitivity. *Bioinformatics* **20,** i311–i317 (2004).

263. Heger, A. & Holm, L. Rapid automatic detection and alignment of repeats in protein sequences. *Proteins Struct. Funct. Bioinforma.* **41,** 224–237 (2000).

264. Marsella, L., Sirocco, F., Trovato, A., Seno, F. & Tosatto, S. C. E. REPETITA: detection and discrimination of the periodicity of protein solenoid repeats by discrete Fourier transform. *Bioinformatics* **25,** i289–i295 (2009).

265. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2015 update. *Nucleic Acids Res.* **43,** D470-478 (2015).

266. Orchard, S. *et al.* The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42,** D358-363 (2014).

267. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45,** D362–D368 (2017).

268. Boldt, K. *et al.* An organelle-specific protein landscape identifies novel diseases and molecular mechanisms. *Nat. Commun.* **7,** 11491 (2016).

269. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw1092

270. Shoemaker, B. A., Panchenko, A. R. & Bryant, S. H. Finding biologically relevant protein domain interactions: Conserved binding mode analysis. *Protein Sci. Publ. Protein Soc.* **15,** 352–361 (2006).

271. Tavtigian, S. V. *et al.* Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J. Med. Genet.* **43,** 295–305 (2006).

272. Capriotti, E., Calabrese, R. & Casadio, R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* **22,** 2729–2734 (2006).

273. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).

274. Choi, Y. A fast computation of pairwise sequence alignment scores between a protein and a set of single-locus variants of another protein. in *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine* 414–417 (ACM, 2012).

275. Hecht, M., Bromberg, Y. & Rost, B. Better prediction of functional effects for sequence variants. *BMC Genomics* **16 Suppl 8,** S1 (2015).

276. Salgado, D. *et al.* UMD-Predictor: A High-Throughput Sequencing Compliant System for Pathogenicity Prediction of any Human cDNA Substitution. *Hum. Mutat.* **37,** 439–446 (2016).

277. Altenhoff, A. M. *et al.* The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res.* **43,** D240–D249 (2015).

278. Lu, J., Peatman, E., Tang, H., Lewis, J. & Liu, Z. Profiling of gene duplication patterns of sequenced teleost genomes: evidence for rapid lineage-specific genome expansion mediated by recent tandem duplications. *BMC Genomics* **13,** 246 (2012).

279. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539–539 (2014).

280. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol. Biol. Evol.* **30,** 2725–2729 (2013).

281. Soding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33,** W244–W248 (2005).

282. Albrecht, M., Tosatto, S. C. E., Lengauer, T. & Valle, G. Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Eng.* **16,** 459–462 (2003).

283. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* **38,** W529-533 (2010).

284. Mosavi, L. K., Minor, D. L. & Peng, Z.-Y. Consensus-derived structural determinants of the ankyrin repeat motif. *Proc. Natl. Acad. Sci. U. S. A.* **99,** 16029–16034 (2002).

285. Groves, M. R. & Barford, D. Topological characteristics of helical repeat proteins. *Curr. Opin. Struct. Biol.* **9,** 383–389 (1999).

286. D'Andrea, L. D. & Regan, L. TPR proteins: the versatile helix. *Trends Biochem. Sci.* **28,** 655–662 (2003).

287. Main, E. R. G., Xiong, Y., Cocco, M. J., D'Andrea, L. & Regan, L. Design of Stable α-Helical Arrays from an Idealized TPR Motif. *Structure* **11,** 497–508 (2003).

288. Li, C. *et al.* Critical evaluation of *in silico* methods for prediction of coiled-coil domains in proteins. *Brief. Bioinform.* bbv047 (2015). doi:10.1093/bib/bbv047

289. Petters, E., Krowarsch, D. & Otlewski, J. Design, expression and characterization of a highly stable tetratricopeptide-based protein scaffold for phage display application. *Acta Biochim. Pol.* **60,** 585–590 (2013).

290. Llères, D., Denegri, M., Biggiogera, M., Ajuh, P. & Lamond, A. I. Direct interaction between hnRNP-M and CDC5L/PLRG1 proteins affects alternative splice site choice. *EMBO Rep.* **11,** 445–451 (2010).

291. Hein, M. Y. *et al.* A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell* **163,** 712–723 (2015).

292. Gupta, G. D. *et al.* A Dynamic Protein Interaction Landscape of the Human Centrosome-Cilium Interface. *Cell* **163,** 1484–1499 (2015).

293. Kim, T. Y. *et al.* Substrate Trapping Proteomics Reveals Targets of the βTrCP2/FBXW11 Ubiquitin Ligase. *Mol. Cell. Biol.* **35,** 167–181 (2015).

294. Darnell, J. C. *et al.* FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146,** 247–261 (2011).

295. Couzens, A. L. *et al.* Protein interaction network of the mammalian Hippo pathway reveals mechanisms of kinase-phosphatase interactions. *Sci. Signal.* **6,** rs15–rs15 (2013).

296. Castello, A. *et al.* Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell* **149,** 1393–1406 (2012).

297. Nonaka, H. *et al.* MINK is a Rap2 effector for phosphorylation of the postsynaptic scaffold protein TANC1. *Biochem. Biophys. Res. Commun.* **377,** 573–578 (2008).

298. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162,** 425–440 (2015).

299. Daulat, A. M. *et al.* Mink1 Regulates -Catenin-Independent Wnt Signaling via Prickle Phosphorylation. *Mol. Cell. Biol.* **32,** 173–185 (2012).

300. Luck, K. *et al.* Putting into Practice Domain-Linear Motif Interaction Predictions for Exploration of Protein Networks. *PLoS ONE* **6,** e25376 (2011).

301. Kırlı, K. *et al.* A deep proteomics perspective on CRM1-mediated nuclear export and nucleocytoplasmic partitioning. *eLife* **4,** e11466 (2015).

302. de Ligt, J. *et al.* Diagnostic exome sequencing in persons with severe intellectual disability. *N. Engl. J. Med.* **367,** 1921–1929 (2012).

303. Emoto, K. The growing role of the Hippo--NDR kinase signalling in neuronal development and disease. *J. Biochem. (Tokyo)* **150,** 133–141 (2011).

304. Moreau, M. M. *et al.* The Planar Polarity Protein Scribble1 Is Essential for Neuronal Plasticity and Brain Function. *J. Neurosci.* **30,** 9738–9752 (2010).

305. Kim, T. Y. *et al.* Substrate trapping proteomics reveals targets of the βTrCP2/FBXW11 ubiquitin ligase. *Mol. Cell. Biol.* **35,** 167–181 (2015).

306. Lohmann, C. & Kessels, H. W. The developmental stages of synaptic plasticity. *J. Physiol.* **592,** 13–31 (2014).

307. Suhl, J. A., Chopra, P., Anderson, B. R., Bassell, G. J. & Warren, S. T. Analysis of FMRP mRNA target datasets reveals highly associated mRNAs mediated by G-quadruplex structures formed via clustered WGGA sequences. *Hum. Mol. Genet.* **23,** 5479–5491 (2014).

308. Kenny, P. J. *et al.* MOV10 and FMRP regulate AGO2 association with microRNA recognition elements. *Cell Rep.* **9,** 1729–1741 (2014).

309. Paemka, L. *et al.* Seizures Are Regulated by Ubiquitin-specific Peptidase 9 X-linked (USP9X), a De-Ubiquitinase. *PLOS Genet.* **11,** e1005022 (2015).

310. Chen, Q. *et al.* CDKL5, a Protein Associated with Rett Syndrome, Regulates Neuronal Morphogenesis via Rac1 Signaling. *J. Neurosci.* **30,** 12777–12786 (2010).

311. Ricciardi, S. *et al.* CDKL5 ensures excitatory synapse stability by reinforcing NGL-1–PSD95 interaction in the postsynaptic compartment and is impaired in patient iPSC-derived neurons. *Nat Cell Biol* **14,** 911–923 (2012).

312. Zhu, Y.-C. *et al.* Palmitoylation-dependent CDKL5-PSD-95 interaction regulates synaptic targeting of CDKL5 and dendritic spine development. *Proc. Natl. Acad. Sci.* **110,** 9118–9123 (2013).

313. La Montanara, P. *et al.* Synaptic Synthesis, Dephosphorylation, and Degradation: A NOVEL PARADIGM FOR AN ACTIVITY-DEPENDENT NEURONAL CONTROL OF CDKL5. *J. Biol. Chem.* **290,** 4512–4527 (2015).

314. Hector, R. D. *et al.* Characterisation of CDKL5 Transcript Isoforms in Human and Mouse. *PLoS ONE* **11,** e0157758 (2016).

315. Peti, W., Nairn, A. C. & Page, R. Structural basis for protein phosphatase 1 regulation and specificity: Protein phosphatase 1 regulation and specificity. *FEBS J.* **280,** 596–611 (2013).

316. Mele, M., Aspromonte, M. C. & Duarte, C. B. Downregulation of GABAA Receptor Recycling Mediated by HAP1 Contributes to Neuronal Death in In Vitro Brain Ischemia. *Mol. Neurobiol.* 1–13 (2016). doi:10.1007/s12035-015-9661-9

317. MANDERS, E. M. M., VERBEEK, F. J. & ATEN, J. A. Measurement of co-localization of objects in dual-colour confocal images. *J. Microsc.* **169,** 375–382 (1993).

318. McMahon, H. T. *et al.* Tetanus toxin and botulinum toxins type A and B inhibit glutamate, gamma-aminobutyric acid, aspartate, and met-enkephalin release from synaptosomes. Clues to the locus of action. *J. Biol. Chem.* **267,** 21338–21343 (1992).

319. Minervini, G. *et al.* Novel interactions of the von Hippel-Lindau (pVHL) tumor suppressor with the CDKN1 family of cell cycle inhibitors. **7,** 46562 (2017).

320. Chen, D.-C., Yang, B.-C. & Kuo, T.-T. One-step transformation of yeast in stationary phase. *Curr. Genet.* **21,** 83–84 (1992).

321. Dosztányi, Z., Mészáros, B. & Simon, I. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics* **25,** 2745–2746 (2009).

322. Li, Q. *et al.* A Syntaxin 1, $G\alpha_o$, and N-Type Calcium Channel Complex at a Presynaptic Nerve Terminal: Analysis by Quantitative Immunocolocalization. *J. Neurosci.* **24,** 4070 (2004).

323. Kilstrup-Nielsen, C. *et al.* What We Know and Would Like to Know about CDKL5 and Its Involvement in Epileptic Encephalopathy. *Neural Plast.* **2012,** 728267 (2012).

324. Bertani, I. *et al.* Functional Consequences of Mutations in CDKL5, an X-linked Gene Involved in Infantile Spasms and Mental Retardation. *J. Biol. Chem.* **281,** 32048–32056 (2006).

325. Nawaz, M. S. *et al.* CDKL5 and Shootin1 Interact and Concur in Regulating Neuronal Polarization. *PLoS ONE* **11,** e0148634 (2016).

326. Oi, A. *et al.* Subcellular distribution of cyclin-dependent kinase-like 5 (CDKL5) is regulated through phosphorylation by dual specificity tyrosine-phosphorylation-regulated kinase 1A (DYRK1A). *Biochem. Biophys. Res. Commun.* **482,** 239–245 (2017).

327. Mari, F. CDKL5 belongs to the same molecular pathway of MeCP2 and it is responsible for the early-onset seizure variant of Rett syndrome. *Hum. Mol. Genet.* **14,** 1935–1946 (2005).

328. Kameshita, I. *et al.* Cyclin-dependent kinase-like 5 binds and phosphorylates DNA methyltransferase 1. *Biochem. Biophys. Res. Commun.* **377,** 1162–1167 (2008).

329. Grossman, E. N., Giurumescu, C. A. & Chisholm, A. D. Mechanisms of Ephrin Receptor Protein Kinase-Independent Signaling in Amphid Axon Guidance in Caenorhabditis elegans. *Genetics* **195,** 899–913 (2013).

330. Zhou, L., Talebian, A. & Meakin, S. O. The Signaling Adapter, FRS2, Facilitates Neuronal Branching in Primary Cortical Neurons via Both Grb2- and Shp2-Dependent Mechanisms. *J. Mol. Neurosci.* **55,** 663–677 (2015).

331. Lu, X. *et al.* The SH2 domain is crucial for function of Fyn in neuronal migration and cortical lamination. *BMB Rep.* **48,** 97–102 (2015).

332. Chaudhury, S., Sharma, V., Kumar, V., Nag, T. C. & Wadhwa, S. Activity-dependent synaptic plasticity modulates the critical phase of brain development. *Brain Dev.* **38,** 355–363 (2016).

333. Wu, C. *et al.* Systematic identification of SH3 domain-mediated human protein–protein interactions by peptide array target screening. *PROTEOMICS* **7,** 1775–1785 (2007).

334. Guo, Z. *et al.* E-cadherin interactome complexity and robustness resolved by quantitative proteomics. *Sci. Signal.* **7,** rs7 (2014).

335. Huttlin, E. L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545,** 505–509 (2017).

336. Trazzi, S. *et al.* HDAC4: a key factor underlying brain developmental alterations in CDKL5 disorder. *Hum. Mol. Genet.* **25,** 3887–3907 (2016).

337. Barbiero, I. *et al.* The neurosteroid pregnenolone reverts microtubule derangement induced by the loss of a functional CDKL5-IQGAP1 complex. *Hum. Mol. Genet.* (2017). doi:10.1093/hmg/ddx237

338. Niture, S. K., Doneanu, C. E., Velu, C. S., Bailey, N. I. & Srivenugopal, K. S. Proteomic analysis of human O6-methylguanine-DNA methyltransferase by affinity

chromatography and tandem mass spectrometry. *Biochem. Biophys. Res. Commun.* **337,** 1176–1184 (2005).

339. Varjosalo, M. *et al.* The Protein Interaction Landscape of the Human CMGC Kinase Group. *Cell Rep.* **3,** 1306–1320 (2013).
340. Wang, J. *et al.* Toward an understanding of the protein interaction network of the human liver. *Mol. Syst. Biol.* **7,** 536–536 (2011).
341. Williamson, S. L. *et al.* A novel transcript of cyclin-dependent kinase-like 5 (CDKL5) has an alternative C-terminus and is the predominant transcript in brain. *Hum. Genet.* **131,** 187–200 (2012).
342. Szafranski, P. *et al.* Neurodevelopmental and neurobehavioral characteristics in males and females with CDKL5 duplications. *Eur J Hum Genet* **23,** 915–921 (2015).
343. Okuda, K. *et al.* CDKL5 controls postsynaptic localization of GluN2B-containing NMDA receptors in the hippocampus and regulates seizure susceptibility. *Neurobiol. Dis.* **106,** 158–170 (2017).
344. Chen, W. *et al.* GRIN1 mutation associated with intellectual disability alters NMDA receptor trafficking and function. *J Hum Genet* **62,** 589–597 (2017).
345. de Ligt, J. *et al.* Diagnostic Exome Sequencing in Persons with Severe Intellectual Disability. *N. Engl. J. Med.* **367,** 1921–1929 (2012).
346. Chandonia, J.-M. *et al.* Lessons from the CAGI-4 Hopkins clinical panel challenge. *Hum. Mutat.* **38,** 1155–1168 (2017).
347. Cukier, H. N. *et al.* Exome sequencing of extended families with autism reveals genes shared across neurodevelopmental and neuropsychiatric disorders. *Mol. Autism* **5,** 1–1 (2014).

348. Betancur, C. Etiological heterogeneity in autism spectrum disorders: More than 100 genetic and genomic disorders and still counting. *Emerg. Neurosci. Autism Spectr. Disord.* **1380,** 42–77 (2011).

349. Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485,** 242–245 (2012).

350. O'Roak, B. J. *et al.* Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.* **43,** 585–589 (2011).

351. Sanders, S. J. *et al.* De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature* **485,** 237–241 (2012).

352. Jiang, Y. *et al.* Detection of Clinically Relevant Genetic Variants in Autism Spectrum Disorder by Whole-Genome Sequencing. *Am. J. Hum. Genet.* **93,** 249–263 (2013).

353. Toma, C. *et al.* Exome sequencing in multiplex autism families suggests a major role for heterozygous truncating mutations. *Mol Psychiatry* **19,** 784–790 (2014).

354. Rauch, A. *et al.* Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. *The Lancet* **380,** 1674–1682 (2012).

355. Vissers, L. E. L. M. *et al.* A de novo paradigm for mental retardation. *Nat. Genet.* **42,** 1109–1112 (2010).

356. Hamdan, F. F. *et al.* De Novo Mutations in Moderate or Severe Intellectual Disability. *PLoS Genet.* **10,** e1004772 (2014).

357. Redin, C. *et al.* Efficient strategy for the molecular diagnosis of intellectual disability using targeted high-throughput sequencing. *J. Med. Genet.* **51,** 724–736 (2014).

358. Agha, Z. *et al.* Exome Sequencing Identifies Three Novel Candidate Genes Implicated in Intellectual Disability. *PLoS ONE* **9,** e112687 (2014).

359. Iossifov, I. *et al.* The contribution of de novo coding mutations to autism spectrum disorder. *Nature* **515,** 216–221 (2014).

360. Majumder, P., Chu, J.-F., Chatterjee, B., Swamy, K. B. S. & Shen, C.-K. J. Co-regulation of mRNA translation by TDP-43 and Fragile X Syndrome protein FMRP. *Acta Neuropathol. (Berl.)* **132,** 721–738 (2016).

361. Ascano, M. *et al.* FMR1 targets distinct mRNA sequence elements to regulate protein expression. *Nature* **492,** 382–386 (2012).

11 Appendix –Supplementary Materials

The supplementary Figures and Tables of unpublished data are listed in the following section.

	Patients (n=146)
Gender	
Female	59
Male	87
Age (at diagnosis, years)	Mean 10y
Familial History	
Sporadic	117
Familial (sib-pair)	29
ID severity	
Mild/Borderline	33
Moderate	35
Severe	32
Psychomotor delay	115
Comorbidity	
ASD	71
Behavioral abnormality	90
Epilepsy	53
Hypotonia	28
Ataxia	10
Microcephaly	19
Macrocephaly	8
Hyperactivity	28
Stereotypy	30
Learning disability	2
Language	
Delay	72
Absent	33
Specific Language Impairment	8
Dimorphisms	
Palate anomaly	7
Dentition anomaly	4
Malar Hypoplasia	2
Skeletal abnormalities	
Digital anomaly	10
Clubbed toes	3
Abnormality of other organs	
Suspected RETT	41
EEG anomaly	55
MRI anomaly	36

Supplementary Table S5.1: Description of the cohort of 146 individuals enrolled for the study of ID/ASD comorbidity.

Reference for known ASD/ID associated genes		
Phenotype	**Source**	
ASD	Cukier et al. 2014[347]	Table S3: ASD candidate genes (only those with at least 2 references and Betancur[348] list)
ASD	Neale et al. 2012[349]	Table S5: ASD genes
ASD	Pinto et al. 2014[9]	Table S6A: ASD genes
ID	Neale et al. 2012[349]	Table S6: ID genes
ID	Pinto et al. 2014[9]	Table S6C: ID genes
Reference for candidate ASD/ID genes from whole exome/panel sequencing		
Phenotype	**Source**	
ASD	O'Roak et al. 2011[350]	Table 2: Summary of confirmed de novo events
ASD	O'Roak et al. 2011[350]	Table 1: Top de novo ASD risk contributing mutations
ASD	Sanders et al. 2012[351]	Table 2: loss of function in probands
ASD	Iossifov et al. 2012[252]	Table 3: Likely disrupting mutations in affected children
ASD	Neale et al. 2012[349]	Table S1: All validated de novo events and annotations (no synonymous variants)
ASD	Jiang et al. 2013[352]	Table 1: Summary ASD genes found with mutations in this study
ASD	Cukier et al. 2014[347]	Supporting Table 5: Genes with Damaging, Validated Variants in More Than One Family
ASD	An et al. 2014[168]	Supplementary Table 3: The list of de novo variants identified in 48 ASD cases (only frameshift, nonsynonymous, stop gain)
ASD	Toma et al. 2014[353]	Table 2: Gene disrupting rare variants shared by affected sibs in each multiplex family
ID	Rauch et al. 2014[354]	Table 3: Missense, nonsense, frameshift, and splice site de-novo variants in genes associated with intellectual disability in each patient–parent trio
ID	Rauch et al. 2014[354]	Table 4: Probable disease-causing de-novo variants in each patient–parent trio
ID	de Ligt et al. 2012[345]	Table 2: Genes Affected by De Novo Mutations Associated with Intellectual Disability
ID	Vissers et al. 2010[355]	Table 2: Overview of all de novo variants identified by exome sequencing in ten individuals with unexplained mental retardation
ID	Hamdman et al. 2014[356]	Table 2: Top risk DNMs identified in this study
ID	Redin et al. 2014[357]	Table 2: List of all causative/possibly causative mutations identified in our cohort
ID	Agha et al 2014[358]	Tables 2-4: **Homozygous and compound heterozygous variants validation using Sanger sequencing and *in silico* prediction**

Supplementary Table S5.2: list of the references considered for initial ASD/ID gene gathering

BIOLOGICAL PROCESS	P-value
Regulation of synaptic transmission (GO:0050804)	0,0000063
Regulation of membrane potential (GO:0042391)	0,0000143
Regulation of excitatory postsynaptic membrane potential (GO:0060079)	0,00000403
Behavior (GO:0007610)	0,000107
Learning (GO:0007612)	0,0000362
Regulation of synaptic plasticity (GO:0048167)	0,0000494
Cognition (GO:0050890)	0,000472
CELLULAR COMPONENT	**P-value**
Postsynaptic membrane (GO:0045211)	0,00000147
Synaptic membrane (GO:0097060)	0,000004518
Synapse (GO:0045202)	0,000007111
Dystrophin-associated glycoprotein complex (GO:0016010)*	0,00006884
Dendrite (GO:0030425)	0,00005336
Neuronal postsynaptic density (GO:0097481)*	0,00008008
Synapse part (GO:0044456)	0,0002044
Neuron spine (GO:0044309)*	0,0002779
MOLECULAR FUNCTION	**P-value**
Ionotropic glutamate receptor activity (GO:0004970)*	0,0001031
Extracellular-glutamate-gated ion channel activity (GO:0005234)*	0,00008931
Metal ion transmembrane transporter activity (GO:0046873)	0,000006874
Glutamate receptor activity (GO:0008066)*	0,0002654
Chromatin binding (GO:0003682)	0,00006805
Substrate-specific channel activity (GO:0022838)	0,0003147
Passive transmembrane transporter activity (GO:0022803)	0,0004382
HUMAN PHENOTYPE ONTOLOGY	**P-value**
Autism (HP:0000717)	0,000000259
Tented upper lip vermilion (HP:0010804)	0,000537
Stereotypic behavior (HP:0000733)	0,000738
Aggressive behavior (HP:0000718)	0,000366
Intellectual disability, severe (HP:0010864)	0,000744
Abnormal social behavior (HP:0012433)	0,000004266
Impaired social interactions (HP:0000735)	0,000004266
Deeply set eye (HP:0000490)	0,0000082

Supplementary Table S5.3: Functional enrichment of the intersection network.
The enrichment analysis was performed with Enricher webserver[37] and it is based on Gene Ontology (GO) classes, and Human Phenotype Ontology (HPO). The p-value indicates the probability that the GO term, or HPO phenotype is assigned by chance to the selected subset of genes.

Gene	Chr. band	I.P.	Associated traits	FMR Targets	PSD proteins	Chromatin Modifiers
ADNP	20q13.13	AD	OCD, ASD, Stereotypy, ADHD, ID, Seizures, Muscular hypotonia, DD , Language impairment	yes	_	_
ANKRD11	16q24.3	AD	Microcephaly, ID, DD	yes	_	_
AP1S2	Xp22.2	XL	Macrocephaly, Aplasia/Hypoplasia of the cerebellum, Poor speech, Muscular hypotonia, ID, Aggressive behavior, DD, ASD, High palate.	_	_	_
ARFGEF2	20q13.13	AR	Microcephaly, Progressive microcephaly, ASD, ID, Seizures, Muscular hypotonia, DD, Hypoplasia of the corpus callosum	_	yes	_
ARID1B	6q25.3	AD/AR	Microcephaly, Hearing impairment, Aggressive behavior, ASD, ID, Seizures, Partial agenesis /Hypoplasia of the corpus callosum, Severe expressive language delay, DD, Muscular hypotonia.	yes	_	yes
ARX	Xp21.3	XLR	Microcephaly, Seizures, DD, Agenesis of corpus callosum, Neonatal hypotonia, ID, Seizures, Muscular hypotonia, Spasticity, Agenesis of corpus callosum, Learning disability, Lissencephaly, DD, Delayed speech and language development, EEG abnormality.	_	_	_
ASH1L	1q22	AD	_	yes	_	yes
ATRX	Xq21.1	XLR	Microcephaly, Sensorineural hearing impairment, ID, Seizures, Spasticity, DD, ADHD, Seizures, Muscular hypotonia	_	_	yes
CASK	Xp11.4	XLD	Sensorineural hearing impairment, ID, Seizures, DD, Muscular hypotonia, Microcephaly, Progressive microcephaly, Spasticity, Absent speech	_	yes	_
CC2D1A	19p13.12	AR	ADHD, ADHD, DD, Incomprehensible speech, Juvenile onset, ID.	_	yes	_
CDKL5	Xp22.13	XLD	Progressive microcephaly, Stereotypy, ASD, Muscular hypotonia, DD, Myoclonus, Generalized myoclonic seizures, ID, profound, DD, Infantile spasms, Epileptic encephalopathy.	yes	yes	_
CHD8	14q11.2	AD	_	yes	_	yes
CNTNAP2	7q35-q36	AD	Impaired social interactions, ADHD, ID, Seizures, DD, Progressive language deterioration,	_	yes	_
CREBBP	16p13.3	AD	Microcephaly, Hearing impairment, ASD, Stereotypy, Delayed speech and language development, ADHD, ID, Seizures, Muscular hypotonia, Agenesis of corpus callosum, EEG abnormality	yes	_	yes
CTNNB1	3p22.1	AD	Microcephaly, ID, Muscular hypotonia, DD, Hypoplasia of the corpus callosum, Neurological speech impairment	yes	yes	_
DEAF1	11p15.5	AD	Aggressive behavior, Mood swings, ASD, Sacral dimple, ID, DD	_	_	_
DYRK1A	21q22.13	AD	Microcephaly, ASD, ID, Cerebral cortical atrophy, Febrile seizures, Severe DD	_	_	_
EHMT1	9q34.3	AD	Brachycephalic, Microcephaly, Hearing impairment, , Aggressive behavior, OCD, Stereotypy, Delayed speech and language development, ID, Seizures, Muscular hypotonia	yes	_	yes

(Continues)

Table S5.4 (Continued)

Gene	Chr. band	I.P.	Associated traits	FMR Targets	PSD proteins	Chromatin Modifiers
FMR1	Xq27.3	XLD	Depression, Dementia, Disinhibition, Anxiety, Dysarthria, Memory impairment, Diffuse cerebral atrophy, OCD, Macrocephaly, ASD, ADHD,ASD, Seizures, ID.	yes	_	yes
FOXG1	14q12	AD	Progressive microcephaly, ASD, Seizures, Spasticity, Neonatal hypotonia, Hypoplasia of the corpus callosum, Apraxia, EEG abnormality, ID.	_	_	_
FOXP1	3p13	AD	Macrocephaly, Aggressive behavior, Stereotypy, Delayed speech and language development, ADHD, ID, DD	_	_	_
GABRB3	15q12	AD	ID, Generalized tonic seizures, Atonic seizures, Atypical absence seizures, Personality disorder, Aggressive behavior, ADHD, Encephalopathy, Mental deterioration, Generalized tonic-clonic seizures, Myoclonus, ASD, EEG with focal sharp slow waves, Falls.	yes	_	_
GAD1	2q31.1	AR	Microcephaly, ID, Seizures, DD, Infantile onset, Cerebral palsy	_	_	_
GATAD2B	1q21.3	AD	Inappropriate laughter, DD, Neonatal hypotonia, Language impairment, ID	_	_	yes
GRIA3	Xq25	XLR	Brachycephaly, Macrocephaly, ASD, Aggressive behavior, ID, Seizures, Myoclonus, Distal muscle weakness, ID.	_	yes	_
GRIK2	6q16.3	AR	ID, Seizures, DD, Dystonia, Myoclonus, Infantile onset	_	_	_
GRIN2A	16p13.2	AD	Delayed speech and language development, ID, Seizures, DD, Dysphasia, Aphasia, Attention deficit ADHD disorder, Agnosia, Speech apraxia, EEG with centrotemporal focal spike waves	yes	yes	_
GRIN2B	12p13.1	AD	Behavioral abnormality, ID, EEG abnormality, Seizures, Muscular hypotonia, DD, Absent speech, Epileptic encephalopathy	yes	yes	_
HDAC4	2q37.3	AD	ID, Seizures, Muscular hypotonia, DD, Microcephaly.	yes	_	yes
IL1RAPL1	Xp21.2--p21.3	XLR	ASD, ADHD, Seizures, ID.	_	yes	_
IQSEC2	Xp11.22	XLD	ID	yes	yes	_
KATNAL2	18q21.1	AD	_	_	_	_
KDM5C	Xp11.22	XLR	High palate, Microcephaly, Macrocephaly, Facial hypotonia, Aggressive behavior, Low frustration tolerance, Seizures, ID	yes	_	yes
KIRREL3	11q24.2	AD	ID	_	yes	_
MBD5	2q23.1	AD	Microcephaly, Aggressive behavior, ADHD, ID, Ataxia, ,Febrile seizures, Language impairment, DD	yes	_	_
MCPH1	8p23	AR	Microcephaly, ID, Seizures, Small cerebral cortex, Increased rate of premature chromosome condensation, Short stature	yes	_	_

(Continues)

Table S5.4 (Continued)

Gene	Chr. band	I.P.	Associated traits	FMR Targets	PSD proteins	Chromatin Modifiers
MECP2	Xq28	XLR	Microcephaly, Facial hypotonia, Delayed speech and language development, Seizures, Ataxia, ID, DD, EEG abnormality, Brachycephaly, Microcephaly, Macrocephaly, ASD, Muscular hypotonia, Progressive microcephaly, Encephalopathy, Myoclonus, EEG abnormality, Congenital onset, ID,ASD.	yes	_	yes
MED12	11q24.2	AD	Hearing impairment, ID, DD, Seizures, Motor delay, Neonatal hypotonia, Partial agenesis of the corpus callosum, Macrocephaly, ADHD ASD, Aggressive behavior, OCD, Impaired social interactions, Low frustration tolerance, ADHD,	yes	_	_
MED13L	12q24.21	AD	ASD, Ataxia, Dysarthria, DD, Motor delay, ID	yes	_	_
MEF2C	5q14.3	AD	ASD, Seizures, Muscular hypotonia, Motor delay, ID, Epileptic encephalopathy	yes	_	_
MIB1	18q11.2	AD	_	yes	_	_
MTF1	1p33	AD	_	yes	_	_
MYH10	17p13.1	XL	_	yes	yes	_
NLGN3	Xq13.1	XL	_	yes	_	_
NLGN4X	Xp22.32-p22.31	XL	ASD, Stereotypy, Delayed speech and language development, , ID, Seizures, Lack of peer relationships, EEG abnormality	_	yes	_
NRXN1	2p16.3	AR	Muscular hypotonia, ID, Epileptic encephalopathy	yes	yes	_
NTNG1	1p13.3	AD		_	_	_
OPHN1	Xq12	XLR	Macrocephaly, Delayed speech and language development, ADHD, ID, Seizures, Muscular hypotonia, Spasticity, DD, Cerebellar hypoplasia,	yes	_	_
PHF21A	11p11.2	AD	Seizures, DD, Brachycephaly	_	_	yes
PHF8	Xp11.22	XLR	ID	yes	_	yes
PPP2R5D	6p21.1	AD	Macrocephaly, ID, Seizures, DD, Generalized hypotonia.	_	yes	_
PQBP1	Xp11.23	XLR	Brachycephaly, Microcephaly, Hearing impairment, Anxiety, ID, Seizures, Spasticity.	_	_	_
PTCHD1	Xp22.11	XLR	ID	_	_	_
PTEN	10q23.31	AD	Macrocephaly, Delayed speech and language development, ID, Seizures, Muscular hypotonia, hick corpus callosum, Hypoplasia of the maxilla, Hearing impairment, Progressive macrocephaly, ASD, DD.	yes	_	_
PTPN4	2q14.2	AD	_	_	_	_
RAB39B	Xq28	XLR	Macrocephaly, ID, Seizures, Neurological speech impairment.	yes	_	_
RAI1	17p11.2	AD	Brachycephaly, Hearing impairment, Stereotypy, Self-mutilation, Delayed speech and language development, ADHD, ID, Seizures, DD, Generalized hypotonia, EEG abnormality, Sleep disturbance, Impaired pain sensation.	yes	_	yes

(Continues)

Table S5.4 (Continued)

Gene	Chr band	I.P.	Associated traits	FMR Targets	PSD proteins	Chromatin Modifiers
RELN	7q22	AD/AR	Microcephaly, Type I Lissencephaly, Thick cerebral cortex, Generalized tonic-clonic seizures with focal onset, Auditory auras	yes	_	_
RPS6KA3	Xp22.12	XLD	Delayed speech and language development, ID, Muscular hypotonia, Motor delay, Sensorineural hearing impairment	_	_	_
SATB2	2q33.1	AD	Microcephaly, Malar Aggressive behavior, Delayed speech and language development, ADHD, Arachnodactyly, ID, Seizures, DD, Happy demeanor.	_	_	yes
SCN2A	2q24.3	AD	Generalized tonic-clonic seizures with focal onset, Focal seizures, Dialeptic seizures, Focal seizures, afebrile, DD, Epileptic encephalopathy	yes	_	_
SETBP1	18q12.3	AD	ID, Seizures, Hypoplasia of the corpus callosum, DD, Brachycephaly, Dental crowding, Motor delay, Absent speech, ADHD.	_	_	_
SHANK2	11q13.3-q13.4	AD	_	yes	yes	_
SHANK3	22q13.33	AD	Microcephaly, Macrocephaly, Dolichocephaly, Hearing impairment, ,ASD, Aggressive behavior, Delayed speech and language development, ASD, Seizures, DD, Motor delay, Hypotonia, ID.	yes	yes	_
SLC6A1	3p25.3	AD	Ataxia, Tremor, Absence seizures, Scoliosis	yes	_	_
SLC9A6	Xq26.3	XLD	Muscular hypotonia, DD, Absent speech, ID, Photosensitive tonic-clonic seizures, happy demeanor.	yes	_	_
SYNGAP1	6p21.32	AD	Microcephaly, ASD, ID, Seizures, Muscular hypotonia, DD, Motor delay, EEG abnormality, Developmental regression, Language impairment, Epileptic encephalopathy	yes	yes	_
TANC2	17q23.2	AD	_	yes	yes	_
TBR1	2q24	AD	DD, Small face, ASD, ID, Seizures, Hand clenching, Growth delay, Neonatal hypotonia,	_	_	_
TCF4	18q21.2	AD	Microcephaly, Aggressive behavior, Seizures, Muscular hypotonia, Motor delay, Encephalopathy, Absent speech, Hypoplasia of the corpus callosum, ID	yes	_	_
TRIO	5p15.2	AD	_	yes	yes	_
TUSC3	8p22	AR	ID	_	_	_
UBE3A	15q11.2	AD	ADHD, Seizures, Muscular hypotonia, DD, Motor delay, Absent speech, EEG abnormality, Postnatal microcephaly, ID.	yes	_	_
WAC	10p12.1-p11.2	AD	Hearing impairment, Seizures, DD.	_	_	yes

Table S5.4: ASD/ID gene panel list. For each gene the chromosome location, the inheritance pattern (I.P.), and the related HPO phenotype are reported. ID = intellectual disability, ASD = autism spectrum disorders, DD= developmental delay, OCD = obsessive-compulsive disorder, ADHD = Attention-Deficit/Hyperactivity Disorder. Genes are classified as FMR (Fragile X mental retardation protein 1) targets, post-synaptic proteins, or chromatin modifiers according to different studies[359–361].

Patient	Sex	Gene	I.P.	Variant	cDNA position	AA change	dbSNP	D/T
2344_01	M	ASH1L	AD	chr1:155450025:T:C	c.A2636G	p.Q879R	_	3/9
2272_01	M	PTEN	AD	chr10:89690828:G:A	c.G235A	p.A79T	rs202004587	5/9
2007_01	M	SHANK2	AD	chr11:70644598:G:A	c.C1727T	p.P576Q	_	8/9
0243_01	F	GRIN2B	AD	chr12:13720096:C:G	c.G2461C	p.V821L	_	5/9
2278_01	M	GRIN2B	AD	chr12:13761626:T:G	c.A1921C	p.I641L	_	6/9
2141_01	M	CHD8	AD	chr14:21876977:G:A	c.C2372T	p.P791L	rs372717272	9/9
1749_01	M	CHD8	AD	chr14:21897467:G:A	c.C871T	p.L291F	rs192989929	8/9
2039_01	M	CREBBP	AD	chr16:3788561:C:T	c.G4393A	p.G1465R	_	9/9
1635_01	M	CREBBP	AD	chr16:3900813:C:T	c.G283A	p.V95M	rs756802946	7/9
1730_01	F	ANKRD11	AD	chr16:89348493:C:G	c.G4457C	p.R1486P	_	5/12
2264_01	M	ANKRD11	AD	chr16:89349967:T:C	c.A2983G	p.K995E	_	4/9
2374_01	F	MYH10	AD	chr17:8393828:C:T	c.G4621A	p.V1541M	rs200997387	6/13
2360_01	F	MBD5	AD	chr2:149226891:C:T	c.C1379T	p.S460L	_	7/12
2360_01	F	MBD5	AD	chr2:149226984:G:C	c.G1472C	p.R491T	_	6/12
2340_01	M	SCN2A	AD	chr2:166165900:C:T	c.C644T	p.A215V	rs149024364	6/9
0270_01	F	HDAC4	AD	chr2:239990233:C:T	c.G2806A	p.V936M	rs757439665	6/9
0270_01	F	ARID1B	AD	chr6:157454311:G:A	c.G2482A	p.G828S	rs150249745	5/9
2387_01	M	ARID1B	AD	chr6:157522326:C:T	c.C4559T	p.T1520I	_	8/12
2241_01	M	RELN	AR, AD	chr7:103130201:C:T	c.G9751A	p.E3251K	rs376520049	
2251_01	F	RELN	AR, AD	chr7:103214555:C:G	c.G4495C	p.D1499H	rs200428576	6/9
2374_01	F	CNTNAP2	AD	chr7:148080864:C:T	c.C3599T	p.S1200L	rs778312206	9/13
1543_01	M	CNTNAP2	AD	chr7:148112649:A:C	c.A3937C	p.N1313H	_	7/12
2140_01	M	GRIA3	XLR	chrX:122460015:G:A	c.G647A	p.R216Q	rs753214982	
196_01	M	CASK	XLD	chrX:41448842:A:G	c.T1159C	p.Y387H	_	4/9
2344_01	M	PHF8	XLR	chrX:53964467:A:G	c.T2794C	p.C932R	rs782094119	1/9
1730_01	F	OPHN1	XLR	chrX:67273488:C:T	c.G2323A	p.V775M	_	7/13
2276_01	M	ATRX	XLR	chrX:76909661:T:C	c.A4244G	p.N1415S	_	7/9
2022_01	M	ATRX	XLR	chrX:76939451:T:C	c.A1297G	p.K433E	_	5/9

Supplementary Table S5.5: Likely pathogenic (LP) variants detected in our cohort. Variant annotation scheme: SNV class (LC/LP), Patient identification code, patients' sex, affected gene, inheritance pattern (I.P), cDNA and amino acid change of the variant, dbSNP code (if available), and the ratio among damaging prediction over the total (D/T).

	Region	Cloning in pGBKT7 vector	Cloning in pGADT7 vector
TANC2	I (1-340)	TANCA-KT7-F: 5’-catggaggccgaattcATGTTTCGGAATAGTCTCAAG TANC-NT-KT7-R: 5’-ggatccccgggaattgTTACACTGACTCTGTAGTGATCG	TANCA-GAD-F:5’-ggaggccagtgaattcATGTTTCGGAATAGTCTCAAG TANC-NT-GAD-R:5'-cacccgggtggaattgTTACACTGACTCTGTAGTGATCG
	II (1-845)	TANCA-KT7-F: 5’-catggaggccgaattcATGTTTCGGAATAGTCTCAAG TANCA-KT7-R:5’-ggatccccgggaattgTTAAGTCTGCTGTCGGTTTAG	_* _*
	III (1-1227)	TANCA-KT7-F: 5’-catggaggccgaattcATGTTTCGGAATAGTCTCAAG TANCB-KT7-R:5-ggatccccgggaattgTTATGGACCTATCTTGGCTCC	TANCA-GAD-F:5’-ggaggccagtgaattcATGTTTCGGAATAGTCTCAAG TANCB-GAD-R:5’-cacccgggtggaattgTTATGGACCTATCTTGGCTCC
	IV (341-1227)	TANCC-KT7-F:5’-catggaggccgaattcTTTGTTGGCCGAGATTGGG TANCB-KT7-R:5-ggatccccgggaattgTTATGGACCTATCTTGGCTCC	TANCC-GAD-F:5’-ggaggccagtgaattcTTTGTTGGCCGAGATTGGG TANCB-GAD-R:5’-cacccgggtggaattgTTATGGACCTATCTTGGCTCC
	V (846-1227)	TANCD-KT7-F: 5’-catggaggccgaattcGAAGGTCTTTCCATGGCAC TANCB-KT7-R:5-ggatccccgggaattgTTATGGACCTATCTTGGCTCC	_* _*
	VI (1228-1990)	TANC-TPR-KT7-F: 5’-catggaggccgaattcTTGAGCAAGCTGATGGAAGAG TANC-CT-KT7-R: 5’-ggatccccgggaattgTTAAACATTAGACTCCACGAAC	TANC-TPR-GAD-F: 5’-ggaggccagtgaattcTTGAGCAAGCTGATGGAAGAG TANC-CT-GAD-R: 5’-cacccgggtggaattgTTAAACATTAGACTCCACGAAC
CDKL5	I (full length)	CDKL5-KT7-F: 5’-catggaggccgaattcATGAAGATTCCTAACATTGG CDKL5-FULL-KT7-R: 5’-ggatccccgggaattgTCACTTGCCCGTCAGTGCCGC	CDKL5-GAD-F: 5’-ggaggccagtgaattcATGAAGATTCCTAACATTGG CDKL5-FULL-GAD-R: 5’-cacccgggtggaattgTCACTTGCCCGTCAGTGCCGC
PP1	full length	_* _*	PP1-GAD-F: 5’-ggaggccagtgaattcATGTCCGACAGCGAGAAGC PP1-GAD-R: 5’-cacccgggtggaattgCTATTTCTTGGCTTTGGCGG

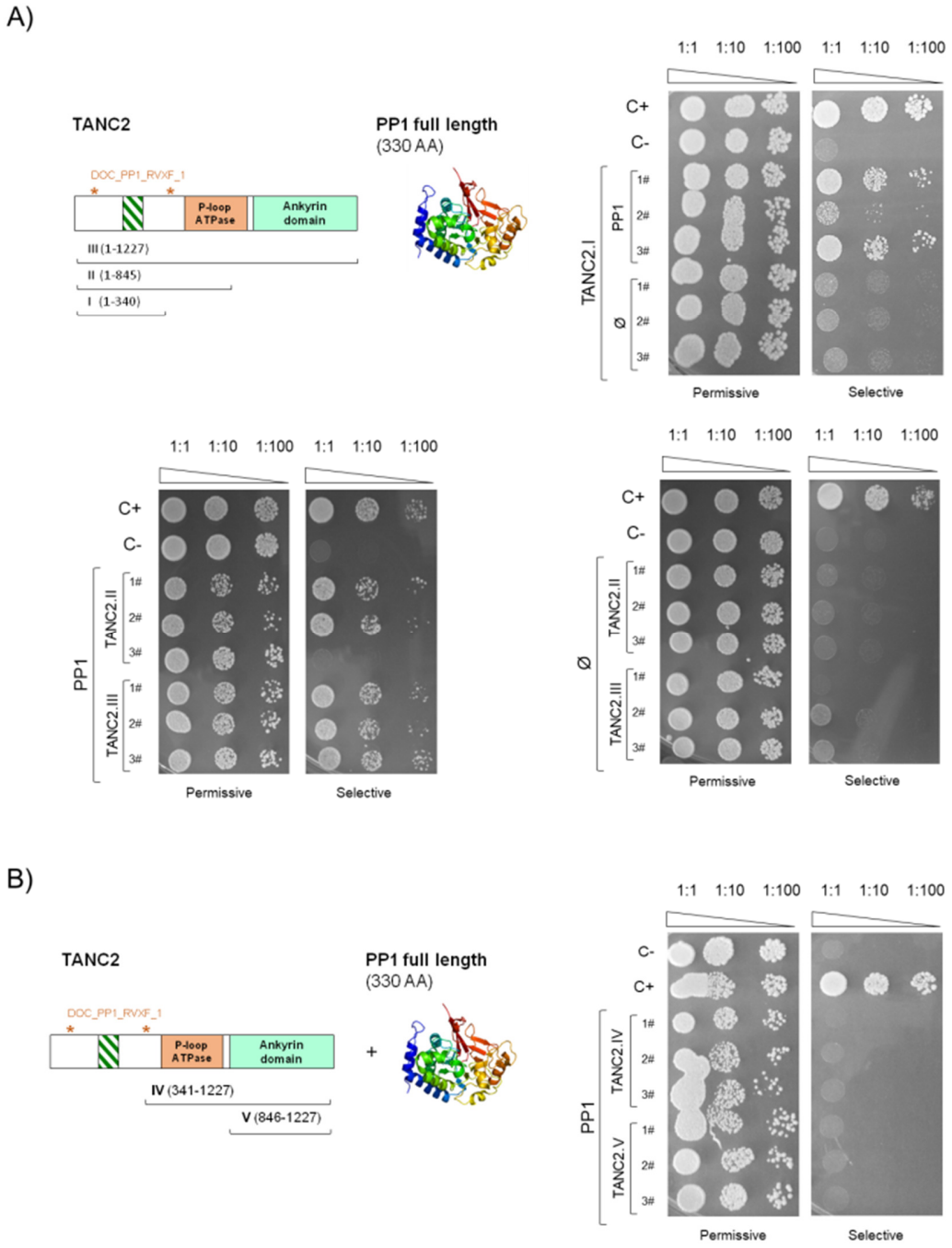
Supplementary Table S8.1: primer sequences used for cloning PP1 and TANC2 regions: In first place, TANC2-PP1 interaction was assessed with TANC2 subclones in p.GBKT7 vector, and PP1 in p.GADT7. However, in this conformation, the experiments presented a high rate of autoactivation, that was eliminated shifting TANC2 in p.GADT7 and PP1 in p.GBKT7. * plasmid switch performed with standard subcloning protocol with EcoRI-HF/BamHI-HF

	Region	Cloning in pGBKT7 vector
TANC2	VII (1228-1358)	TANC-TPR-KT7-F: 5’-catggaggccgaattcTTGAGCAAGCTGATGGAAGAG TANC-TPR-KT7-R: 5’-ggatccccgggaattgTTAGTTGTTGGGACACAGCTTGATG
	VIII (1244-1542)	del301-747: 5'-ggccctcagtatcggtaacaattcccgggg del301-747-antisense: 5'-ccccgggaattgttaccgatactgagggcc
	IX (1538-1628)	TANC-polyQP-KT7-F: 5’-catggaggccgaattcAGAGGCCCTCAGTATCGGG TANC-polyQP-KT7-R: 5’-ggatccccgggaattgTTAGATGACGGTGCTTGAATGG
	X (1623-1990)	del20-399: 5'-catggaggccgaattctcaagcaccgtcatcc del20-399-antisense: 5'-ggatgacggtgcttgagaattcggcctccatg
		Cloning in pGADT7 vector
CDKL5	II (1-300)	CDKL5-GAD-F: 5’-ggaggccagtgaattcATGAAGATTCCTAACATTGG CDKL5-GAD-R: 5’-cacccgggtggaattgTTACTGGGTTTGAAATGTAGGG
	III (301-615)	GAD-CDKL5-central_F: 5’-ggaggccagtgaattcAGACTTCTGGATCGTTCTCC GAD-CDKL5-central_R: 5’-cacccgggtggaattgTTACACATACATAGAATGCCTATG
	IV (587-1030)	GAD-CDKL5- Δ2-F: 5’- ggaggccagtgaattcCATTCCCATTCACTGTCTGC GAD-CDKL5- Δ2-R: 5’- cacccgggtggaattgTCACTTGCCCGTCAGTGCCGC
	IV.A (587-740)	GAD-CDKL5-central_F: 5’-ggaggccagtgaattcAGACTTCTGGATCGTTCTCC GAD-CDKL5-C1_R: 5’-cacccgggtggaattgTCACTCTGATGGTAGAGAAGAAAC
	IV.B (731-802)	GAD-CDKL5-C2_F:5’- ggaggccagtgaattcAATGTGTCAACTAGAGTTTC GAD-CDKL5-C2_R: 5’-cacccgggtggaattgTCAGGATTTCTGCAACGTCAGAAG
	IV.C (796-1030) IV.A (587-740)	GAD-CDKL5-C3_F:5’-ggaggccagtgaattcTCCGACAGCCCTGATCTTC CDKL5-FULL-GAD-R: 5’-cacccgggtggaattgTCACTTGCCCGTCAGTGCCGC GAD-CDKL5-central_F: 5’-ggaggccagtgaattcAGACTTCTGGATCGTTCTCC GAD-CDKL5-C1_R: 5’-cacccgggtggaattgTCACTCTGATGGTAGAGAAGAAAC

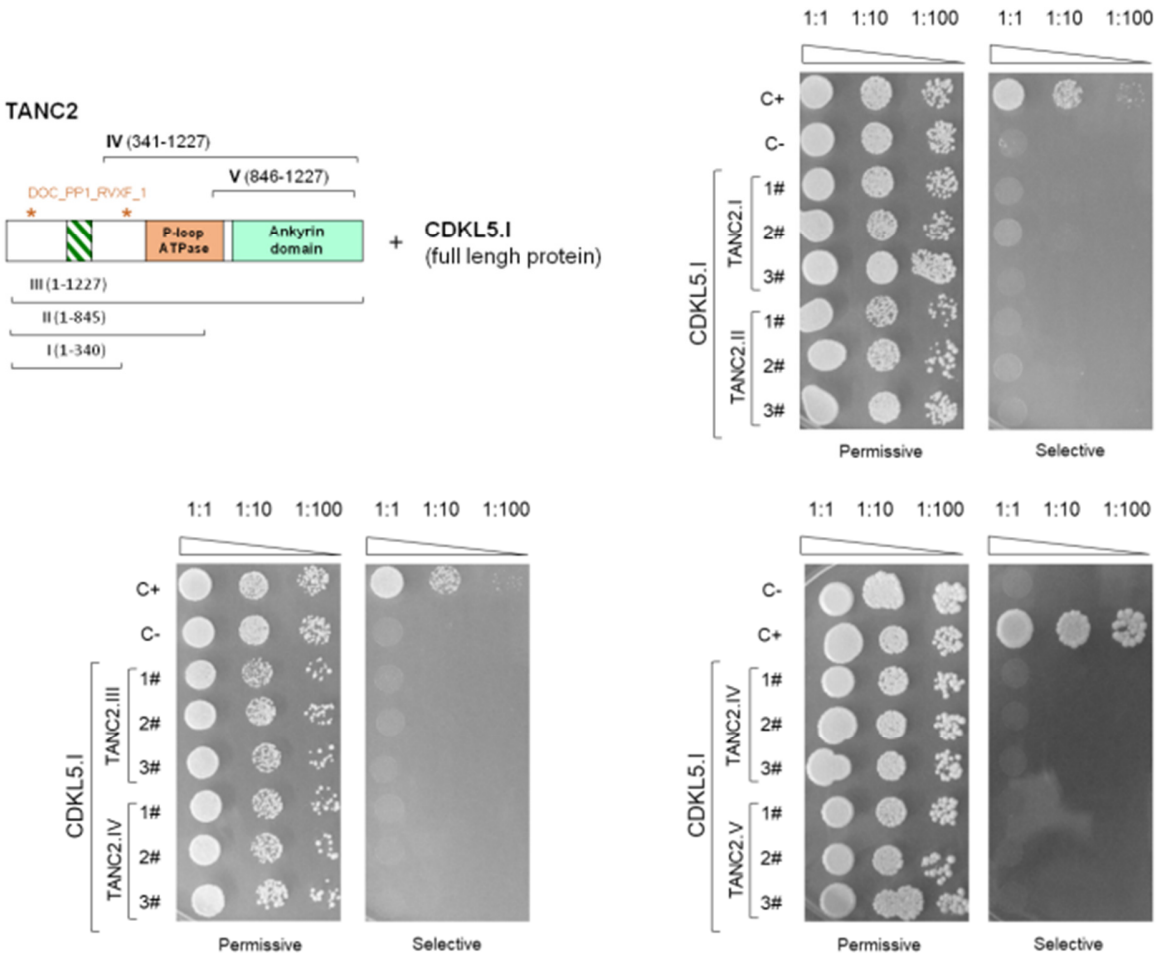
Supplementary Table S8.2: primer sequences used for mutagenesis/fine mapping of regions interacting in CDKL5/ TANC2 complex. In bold, primers used with QuikChange II XL Site-Directed Mutagenesis Kit (Agilent). They allow to produce deleted mutants independently from the vector in which the fragment of interest is. Other primers (not in bold) were used for In-fusion cloning.

Predicted Interacting region	From	To	Length	CDKL5 subclone
1	331	337	7	CDKL5.III
2	344	386	43	CDKL5.III
3	397	423	27	CDKL5.III
4	431	458	28	CDKL5.III
5	464	478	15	CDKL5.III
6	480	526	47	CDKL5.III
7	539	546	8	CDKL5.III
8	*566*	*604*	*39*	*CDKL5.IV.A*
9	*609*	*649*	*41*	*CDKL5.IV.A*
10	*656*	*666*	*11*	*CDKL5.IV.A*
11	*670*	*690*	*21*	*CDKL5.IV.A*
12	*699*	*721*	*23*	*CDKL5.IV.A*
13	*727*	*741*	*15*	*CDKL5.IV.B*
14	*753*	*760*	*8*	*CDKL5.IV.B*
15	*771*	*788*	*18*	*CDKL5.IV.B*
16	*798*	*811*	*14*	*CDKL5.IV.C*
17	*826*	*851*	*26*	*CDKL5.IV.C*
18	856	889	34	CDKL5.IV.C
19	896	909	14	CDKL5.IV.C
20	921	939	19	CDKL5.IV.C
21	995	1003	9	CDKL5.IV.C

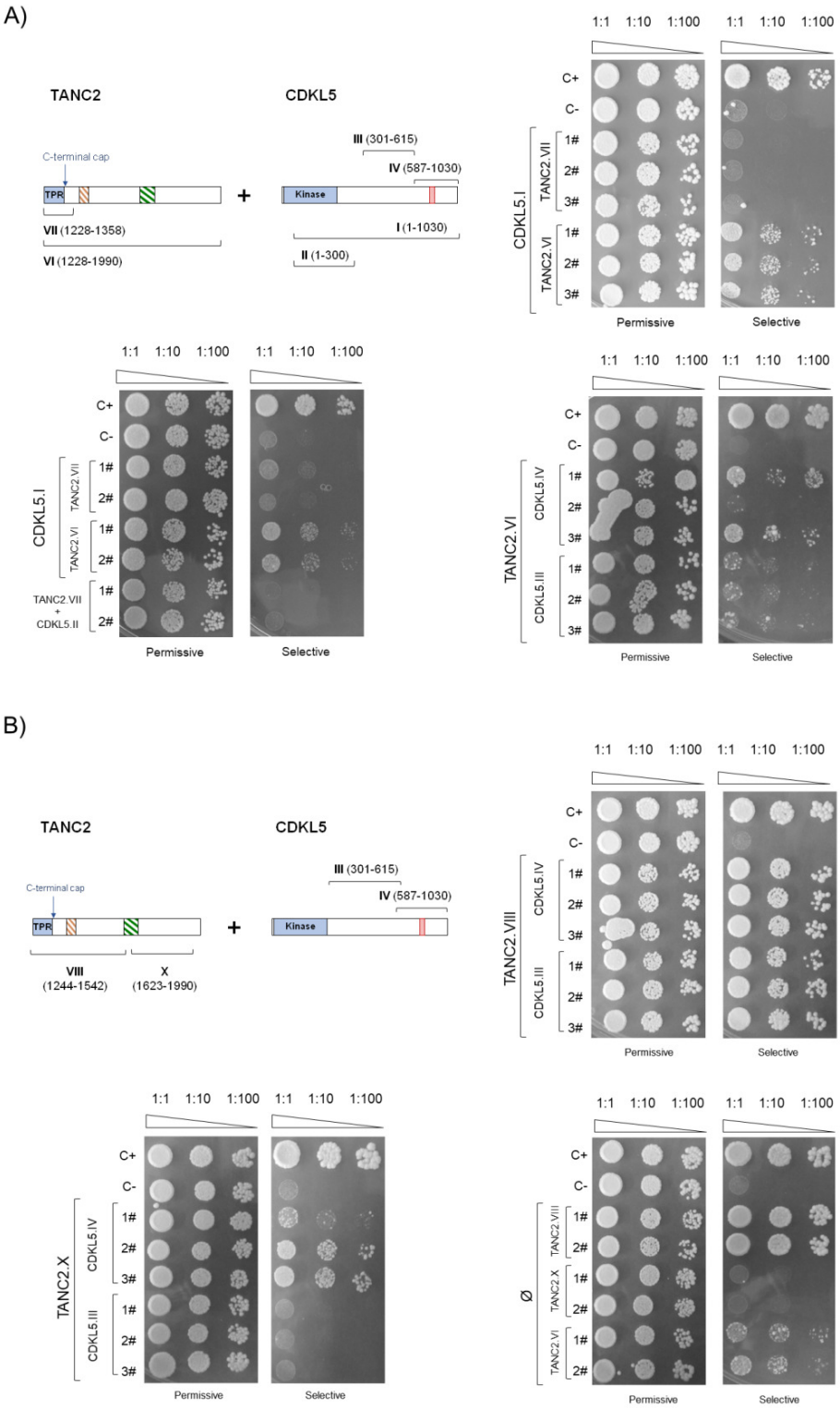
Supplementary Table S8.3: predicted disordered binding regions by ANCHOR. In italic: predicted binding regions mapping to experimentally validated interaction sites.



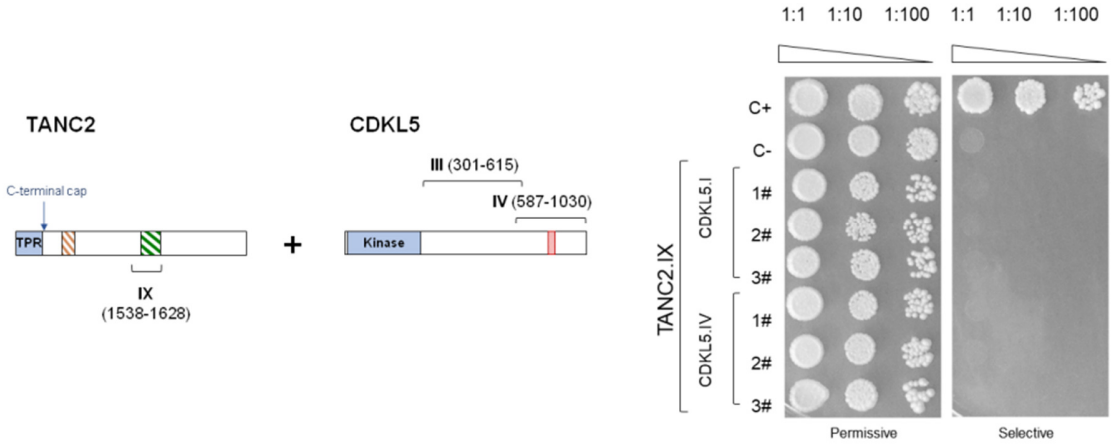
Supplementary Figure S8.1: TANC2/PP1 combinations assessed with Y2H system. A) replicates of interactions reported in the main text (Figure 1). B) Drop test of TANC2.III and TANC2.IV with PP1 (negative results).



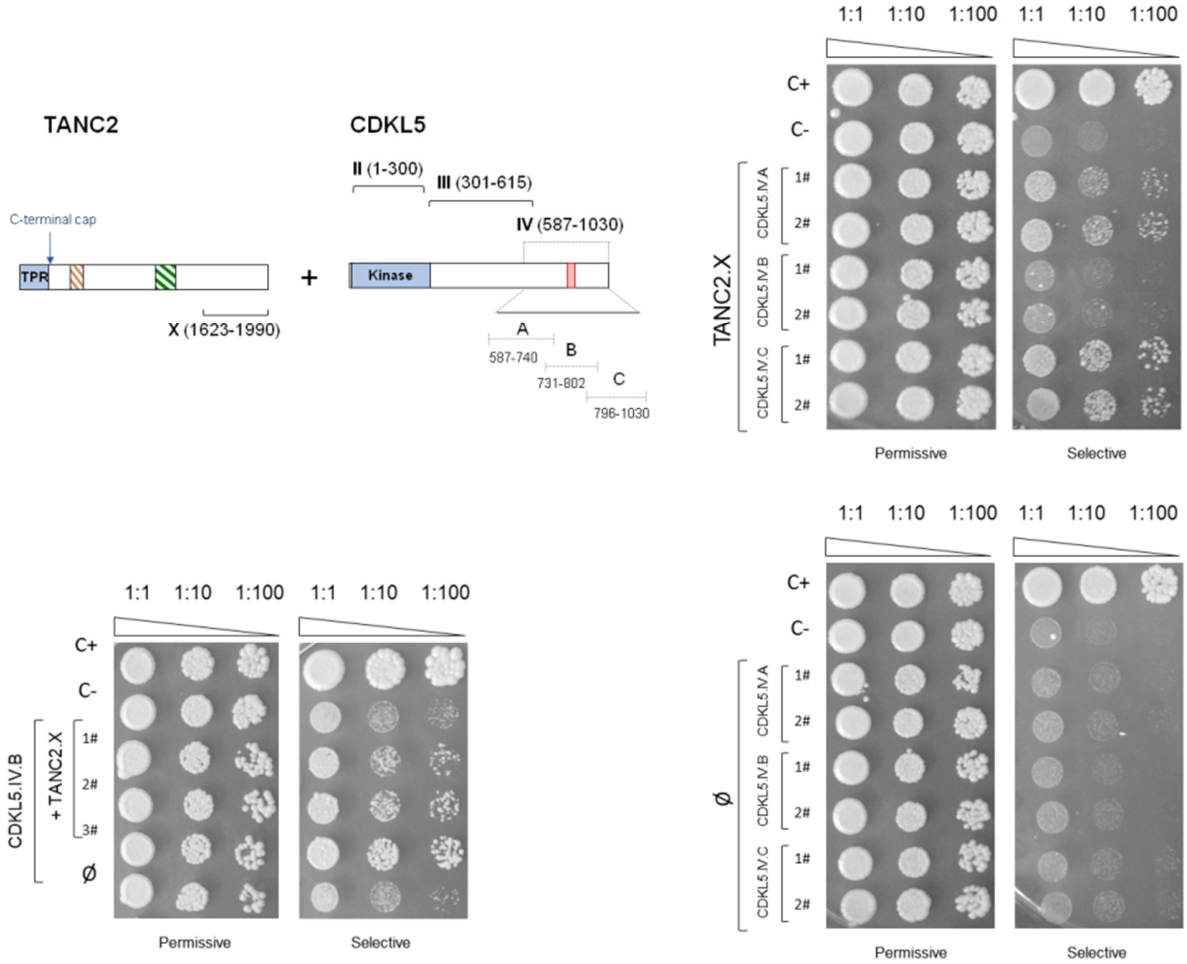
Supplementary Figure S8.2: TANC2/CDKL5 combinations assessed with Y2H system. Interactions between CDKL5.I and TANC2.I, TANC2.II, TANC2.III, and TANC2.IV (negative results).



Supplementary Figure S8.3: TANC2/CDKL5 combinations assessed with Y2H system. A) replicates of interactions between TANC2 fragments and CDKL5.I presented in the main text. B) Drop test between CDKL5.III and CDKL5.III and TANC2 C-terminus subclones (replicate, and autoactivation tests on the left of the panel).

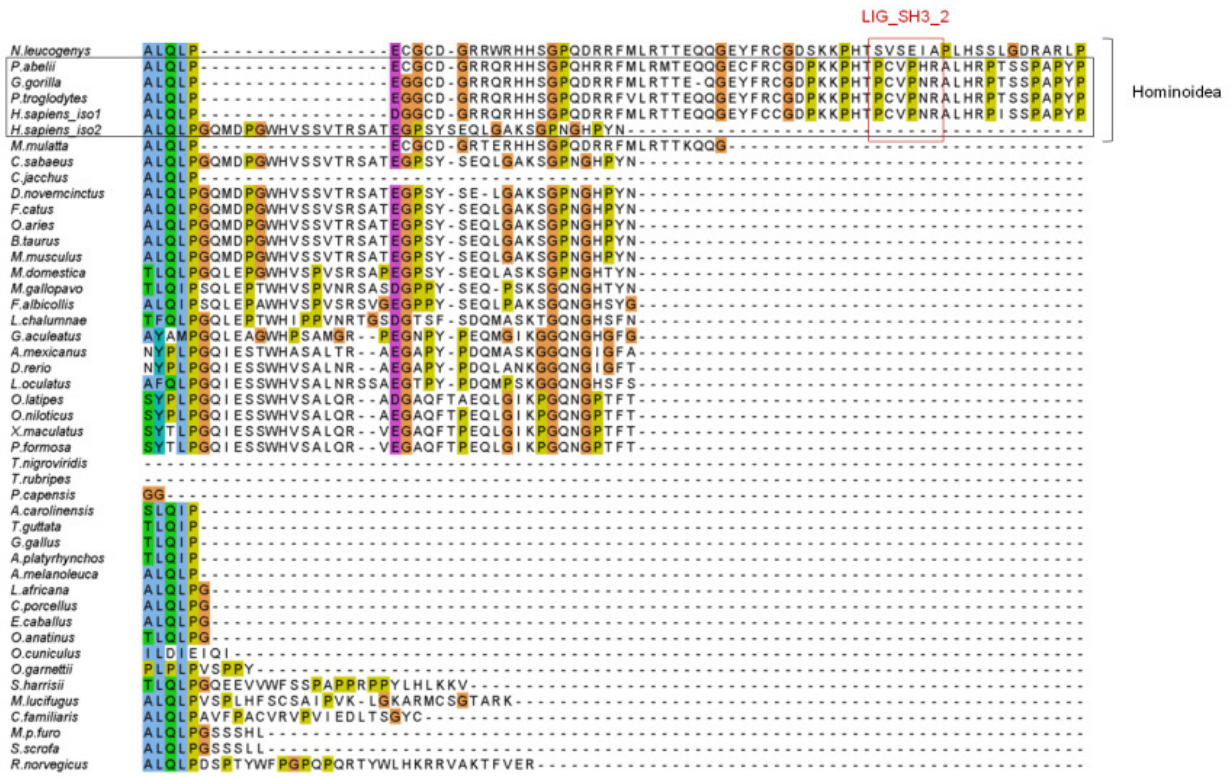


Supplementary Figure S8.4: TANC2.IX does not interact with CDKL5 in Y2H system. Drop test between CDKL5.I and CDKL5.IV and TANC2.IX (negative results).

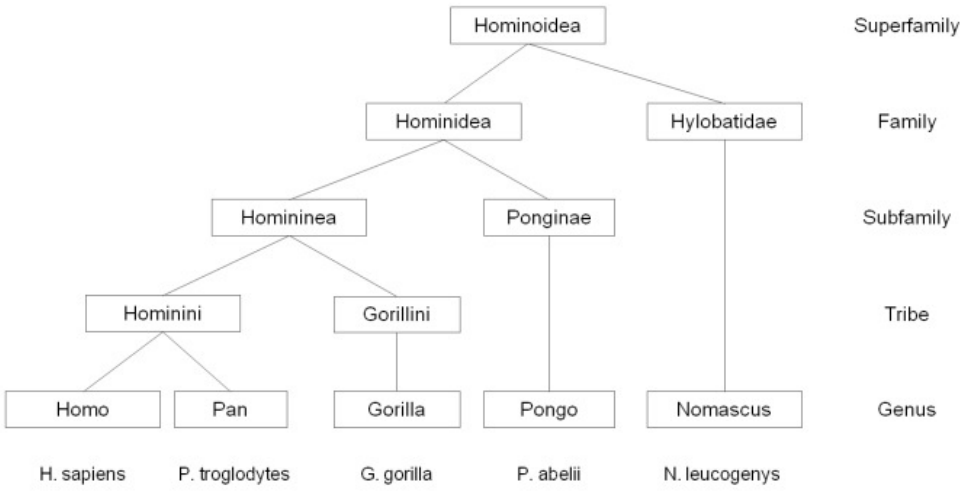


Supplementary Figure S8.5: TANC2.X interacts with CDKL5.IV.A and CDKL5.IV.C. Replicate of the drop test in the main text. On the left of the panel, autoactivation test for CDKL5.IV.A and CDKL5.IV.C subclones.

A)



B)



Supplementary Figure S8.6: CDKL5 isoform 1 C-terminus similarity in *Hominidae* subfamily.
A) Multiple sequence alignment of CDKL5 orthologous C-terminal sequences (from residue 904 of CDKl5 isoform 1). LIG_SH3_2= SH3 domain binding site **B)** *Hominoidea* superfamily tree.