Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics

of the Ruperto Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

presented by

Juan Carlos González Sánchez

born in Murcia, Spain

Oral examination:
06.12.2022

Assessing functional impact of amino acid alterations in proteins

Referees:

Prof. Dr. Robert B. Russell

Prof. Dr. Britta Brügger

Prof. Dr. Frauke Gräter

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Rob for giving me the opportunity of doing research in his group, for his mentorship and guidance through all these years, and specially, for never giving up on me. I am very grateful for all I have learned and achieved.

I would like to thank also the members of my advisory committee, Prof. Dr. Britta Brügger and Prof. Dr. Frauke Gräter, for their useful feedback and assistance; as well as to Prof. Dr. Julio Sáez Rodríguez for kindly accepting to be part of the thesis committee.

A big thanks to all my colleagues, past and present. I have met only excellent people since I came to the group, and you have all shaped me one way or another. I want to express my special gratitude to Yvonne who not only is always eager to help but she does so with a smile.

Of course, it would not have been possible to do this without the constant support of my longest partners in crime. Gurdeep, with whom I've shared an office since the days the shutters still worked, I thank you for pushing me… and then pulling me, I finally get it. Torsten, who has kept me a float more times than I can count, I thank you for your infinite patience and your unconditional friendship in and out of the battlefield.

I would like to thank my parents, who always believed in me and supported me from the distance (often in the form of packages of Spanish goodies), and, in special, my new family, the one I have made along the way: Isabel and Boo. With me in every step, encouraging me and sharing the moments that kept me going.

Abstract

Genome sequencing efforts, coupled with technological advances and cost reductions, have led to the discovery of an increasing number of disease-related genetic variants. For the vast majority of these variants there is no known molecular mechanism for how they are related to the disease. This problem is particularly evident for diseases with complex genotype-phenotype relationships, such as cancer. Fortunately, the parallel growth of data on protein families, structures, interactions, modifications, and other aspects of function, in addition to the development of new computational methods provide the means to predict or identify disease variant mechanism.

In this thesis, I first present a systematic analysis of a large dataset of pan-cancer missense mutations to investigate whether positive selection of certain types of amino acid substitutions can reveal interaction-disrupting cancer driver mutations. Hundreds of mechanistically interesting variants were identified in both potentially novel cancer-associated proteins and well-established cancer driver genes. I discuss new insights and for some instances, attempt functional interpretations by integrating information on protein structure and interactions that suggest putative novel mechanisms that question the classical oncogene/tumour suppressor paradigm.

There is a wealth of publicly available resources that already provide valuable information on all aspects that define gene and protein function. This information has been collected from thousands of experiments or publications and has usually been manually verified or predicted using new approaches. This means that interpreting variants can be a tedious process of manually consulting and integrating the different functional data from multiple databases. Mechnetor was developed to aid this process: a freely available web tool that helps users understand the mechanism of protein variants. With a simple input from the user, Mechnetor automatically collects and integrates various relevant functional data and presents them in an interactive network that allows easy visualisation and interpretation of the results.

Many databases are created from the individual efforts of hundreds of labs conducting similar experiments, combining their results to build and increase the confidence of biological knowledge. I had the opportunity to collaborate with the group of Prof. Dr. Felix

Wieland (Heidelberg University Biochemistry Center) in analysing and interpreting the results of one such experiment: a proteome-wide study of *S*-palmitoylation in *Drosophila melanogaster*. *S*-palmitoylation is an important reversible post-translational modification that controls protein membrane location and trafficking and is thus linked to many cellular processes. In contrast to humans, palmitoylation target proteins and responsible enzymes are largely unknown in invertebrates. Here, we identified and characterised the most complete set of *S*-palmitoylated proteins in *Drosophila* to date, as well as the putative substrate profiles of 10 *Drosophila* palmitoyl acyl transferases. Our results provide new insights and reveal many functional similarities of palmitoylation between *Drosophila* and humans.

.

Zusammenfassung

Genomsequenzierungsbemühungen, gepaart mit technologischen Fortschritten und Kostensenkungen, haben zur Entdeckung einer zunehmenden Zahl von krankheitsbedingten genetischen Varianten geführt. Für die überwiegende Mehrheit dieser Varianten gibt es keinen bekannten molekularen Mechanismus dafür, wie sie mit der Krankheit zusammenhängen. Dieses Problem ist besonders offensichtlich bei Krankheiten mit komplexen Genotyp-Phänotyp-Beziehungen wie Krebs. Glücklicherweise bietet das parallele Wachstum von Daten zu Proteinfamilien, Strukturen, Wechselwirkungen, Modifikationen und anderen Funktionsaspekten neben der Entwicklung neuer Rechenmethoden die Möglichkeit, Krankheitsvariantenmechanismen vorherzusagen oder zu identifizieren.

In dieser Dissertation stelle ich zunächst eine systematische Analyse eines großen Datensatzes von „Missense"-Mutationen bei Krebserkrankungen vor, um zu untersuchen, ob eine positive Selektion bestimmter Arten von Aminosäuresubstitutionen Interaktions-unterbrechende Krebstreiber-Mutationen aufdecken kann. Hunderte von mechanistisch interessanten Varianten wurden sowohl in potenziell neuen krebsassoziierten Proteinen als auch in gut etablierten Onkogenen und Tumorsuppressor-genen identifiziert. Ich diskutiere neue Erkenntnisse und versuche in einigen Fällen funktionelle Interpretationen, indem ich Informationen über Proteinstruktur und -wechselwirkungen integriere, die auf mutmaßlich neue Mechanismen hindeuten, die das klassische Onkogen/Tumorsuppressor-Paradigma in Frage stellen.

Es gibt eine Fülle öffentlich zugänglicher Ressourcen, die bereits wertvolle Informationen zu allen Aspekten liefern, die die Funktion von Genen und Proteinen definieren. Diese Informationen wurden aus Tausenden von Experimenten oder Veröffentlichungen gesammelt und normalerweise manuell verifiziert oder mit neuen Ansätzen vorhergesagt. Das bedeutet, dass das Interpretieren von Varianten ein mühsamer Prozess sein kann, bei dem die verschiedenen Funktionsdaten aus mehreren Datenbanken manuell konsultiert und integriert werden müssen. Mechnetor wurde entwickelt um diesen Prozess zu unterstützen: ein frei verfügbares Webtool, welches Benutzern hilft den Mechanismus von Proteinvarianten zu verstehen. Mit einer einfachen Eingabe des Benutzers sammelt und integriert Mechnetor automatisch verschiedene relevante Funktionsdaten und präsentiert sie in einem interaktiven Netzwerk, das eine einfache Visualisierung und Interpretation der Ergebnisse ermöglicht.

Viele Datenbanken werden aus den individuellen Bemühungen von Hunderten von Labors erstellt, die ähnliche Experimente durchführen und ihre Ergebnisse kombinieren, um das Vertrauen in biologisches Wissen aufzubauen und zu stärken. Ich hatte die Gelegenheit, mit der Wieland-Gruppe (Biochemiezentrum der Universität Heidelberg) zusammenzuarbeiten, um die Ergebnisse eines solchen Experiments zu analysieren und zu interpretieren: einer proteomweiten Studie zur *S*-Palmitoylierung in Drosophila melanogaster. Die *S*-Palmitoylierung ist eine wichtige reversible posttranslationale Modifikation, die die Lage und den Transport von Proteinmembranen kontrolliert und somit mit vielen zellulären Prozessen verbunden ist. Im Gegensatz zum Menschen sind die Zielproteine der Palmitoylierung und die verantwortlichen Enzyme bei Wirbellosen weitgehend unbekannt. Hier identifizierten und charakterisierten wir den bisher vollständigsten Satz von S-palmitoylierten Proteinen in Drosophila sowie die mutmaßlichen Substratprofile von 10 Drosophila-Palmitoylacyltransferasen. Unsere Ergebnisse liefern neue Einblicke und offenbaren viele funktionelle Ähnlichkeiten der Palmitoylierung zwischen Drosophila und dem Menschen.

Table of Contents

Publications

Most of the content of this thesis has been published in independent publications where I am one of the lead authors. For those, I detail my contributions. In addition, I also participated in other collaborative projects during my time as a PhD student that are not described in detail here.

First/Co-first authorships

1. González-Sánchez, J. C.†, Raimondi, F.†, & Russell, R. B. (2018). "Cancer Genetics Meets Biomolecular Mechanism—Bridging an Age-Old Gulf." *FEBS Letters*, 592(4): 463–74. https://doi.org/10.1002/1873-3468.12988

   I contributed equally to the writing of this review paper. Chapter 1 is partially adapted from this publication.

2. González-Sánchez, J. C., Ibrahim, M. F. R., Leist, I. C., Weise, K. R., & Russell, R. B. (2021). "Mechnetor: A Web Server for Exploring Protein Mechanism and the Functional Context of Genetic Variants." *Nucleic Acids Research*, 49 (W1): W366-74. https://doi.org/10.1093/nar/gkab399

   I participated in the conceptualization of the project, data analysis, development of both front-end and back-end of the tool, and preparation of the manuscript. Chapter 3 is based on this publication but significantly extended from it.

3. Diwan, G. D.†, González-Sánchez, J. C.†, Apic, G., & Russell, R. B. (2021). "Next Generation Protein Structure Predictions and Genetic Variant Interpretation." *Journal of Molecular Biology*, 433(20): 167180. https://doi.org/10.1016/j.jmb.2021.167180

   I participated in the data analysis and writing of the manuscript.

4. Porcellato, E.†, González-Sánchez, J. C.†, Ahlmann-Eltze, C., Elsakka, M. A., Shapira, I., Fritsch, J., Navarro, J. A., Anders, S., Russell, R. B., Wieland, F. T. & Metzendorf, C. (2022). "The S-palmitoylome and DHHC-PAT interactome of Drosophila melanogaster S2R+ cells indicate a high degree of conservation to mammalian palmitoylomes." *PLOS ONE*, *17*(8), e0261543. https://doi.org/10.1371/journal.pone.0261543

   I participated in the computational analysis and integration of the data, interpretation of results and preparation of the manuscript.

Contributions

5. Bordin, N., González-Sánchez, J. C., & Devos D. P. (2018). “PVCbase: An Integrated Web Resource for the PVC Bacterial Proteomes.” *Database : The Journal of Biological Databases and Curation* 2018 (2018). https://doi.org/10.1093/database/bay042
6. Raimondi, F., Inoue, A., Kadji, F., Shuai, N., González-Sánchez, J. C., Singh, G., de la Vega, A. A., Sotillo, R., Fischer, B., Aoki, J., Gutkind, J. S., & Russell, R. B. (2019). “Rare, functional, somatic variants in gene families linked to cancer genes: GPCR signaling as a paradigm.” *Oncogene*, *38*(38), 6491–6506. https://doi.org/10.1038/s41388-019-0895-2
7. López, C., Schleussner, N., Bernhart, S. H., Kleinheinz, K., Sungalee, S., Sczabiel, H. L., … González-Sánchez, J. C., … & Siebert R. (2022). “Focal structural variants revealed by whole genome sequencing disrupt the histone demethylase *KDM4C* in B cell lymphomas.” *Haematologica.* https://doi.org/10.3324/haematol.2021.280005
8. Muñoz-Prieto, A., Rubić, I., González-Sánchez, J. C., Kuleš, J., Martínez-Subiela, S., Cerón, J. J., … Tvarijonaviciute, A. (2022). “Saliva changes in composition associated to COVID-19: a preliminary study.” *Scientific Reports 2022 12:1*, *12*(1), 1–14. https://doi.org/10.1038/s41598-022-14830-6

List of Figures and Tables

Figures obtained or adapted from publications I was involved in were, unless specified otherwise, originally created by me.

Abbreviations

1kG	1000 Genomes
ABE	Acyl-Biotin Exchange
Acyl-RAC	Acyl-Resin Assisted Capture
BioID	Proximity-dependent biotin identification
CGC	Cancer Gene Census
COSMIC	Catalogue Of Somatic Mutations In Cancer
DDI	Domain-Domain Interaction
DHHC	Asp-His-His-Cys motif
DMI	Domain-Motif Interaction
ELM	Eukaryotic Linear Motif
GPCR	G protein-coupled receptor
ICGC	International Cancer Genome Consortium
IDR	Intrinsically Disordered Regions
NGS	Next-Generation Sequencing
PAT	Palmitoyl Acyl-Tranferase
PPI	Protein-Protein Interaction
PSP	PhosphoSitePlus
PTM	Post-Translational Modification
SLiM	Short Linear Motif
SNARE	soluble N-ethylmaleimide-sensitive factor attachment receptor
SNV	Single Nucleotide Variant
TCGA	The Cancer Genome Atlas
TSG	Tumour Suppressor Gene

Chapter 1

Introduction

1.1 The genetic variant explosion: a challenge for molecular biologists

The Next-Generation Sequencing (NGS) revolution had an enormous impact on clinical research as it completely changed the paradigm of genomics, shifting its scope from the study of single genes linked to disease to the analysis of whole genomes (Koboldt et al., 2013; Soon et al., 2013). Fast and inexpensive sequencing has been widely applied to the genomes of both healthy and diseased individuals, enabling the identification of an ever-growing number of genetic variants of every kind: common human genetic variation (1000 Genomes Project Consortium et al., 2015; Sherry et al., 2001), rare Mendelian disease-causing variants (Amberger et al., 2019), and somatic mutations underpinning most cancers (Tate et al., 2019). As a result, large volumes of genomic data have accumulated and these are often publicly available. The challenge now lies in the identification of variants related to disease and, in particular, in their functional interpretation.

Unfortunately, the genetic basis of most human traits, and especially human diseases, is often very complex. Historically, precise correlations between genotype and phenotype have only been established for a few genetic diseases caused by single, highly penetrant alleles, where the functional interpretation of variants is usually simpler (Amberger et al., 2019; Boycott et al., 2013). Cystic fibrosis, for example, is caused by mutations in CFTR (in 70% of cases by the deletion of a single residue, p.Phe508) (O'Sullivan & Freedman, 2009), and Rett syndrome is caused by loss-of-function mutations in MECP2 (Amir et al., 1999). For most diseases, however, the genotype-phenotype relationship is less clear. Either positively identified causative variants have low penetrance, leaving many others to be found, or the particular mechanisms by which they cause disease pathology are not well understood. The result is that the vast majority of variants uncovered by sequencing are still classified as 'variants of uncertain significance' (Federici & Soddu, 2020; Richards et al., 2015).

The need to understand the molecular basis of disease has increased in recent years owing to the advent of precision medicine (Ashley, 2016). With NGS and other advanced technologies applied to individuals, it is increasingly possible to determine the specific genetic variants responsible for disease for each individual patient (Perkins et al., 2018). This now means that many diagnostic and treatment decisions depend on the interpretation of variants specific to a patient (Suwinski et al., 2019), which is particularly important in the field of precision oncology (Malone et al., 2020). The need for better tools to interrogate such variants has never been more acute.

1.2 Cancer is a genetically complex and diverse disease

Cancer is probably the most representative case of complex genotype-phenotype relationship, as it actually refers to a large group of diseases that show great genetic and phenotypic diversity. Cancers occur in almost every organ and tissue, and are the leading cause of death worldwide. Their unifying characteristic is they are all primarily caused by an accumulation of genetic abnormalities that lead to uncontrolled cell growth and division. Ultimately, these cells form a tumour that invades normal tissues and organs and can spread throughout the body (Stratton et al., 2009). Currently, the medical community has recognised around 200 cancer types according to the histology and subtype of tumour[1].

Cell transformation from normal to malignant is a multistep process driven by somatic mutations that are acquired progressively and then positively selected (Hanahan & Weinberg, 2000) (Figure **1.1**), while negative selection was surprisingly found to be an almost absent force during cancer development (Martincorena et al., 2017). Inherited genetic variation however can contribute to an increased susceptibility to certain cancer types, and at least 5-10% of cancers are considered to arise due to highly penetrant germline mutations (Nagy et al., 2004).

Cancer causative mutations, known as drivers, inhibit or alter the function of certain genes (cancer driver genes), disrupting the processes that regulate normal cell growth and homeostasis, and therefore, promoting tumorigenesis (Stratton et al., 2009). The analysis of the molecular processes driving cancer is usually aimed at the identification of twos of driver

[1] https://www.cancer.gov/types

genes: oncogenes and tumour suppressors (TSGs). Proto-oncogenes are genes that promote cell growth and division, which upon gain-of-function mutations, become oncogenes with an increased and uncontrolled activity. TSGs, in contrast, are those that suffer loss-of-function mutations which diminish or completely inhibit their ability to restrict proliferation and stimulate DNA repair (Imbeaud et al., 2010; Lee & Muller, 2010; Weinberg, 1994).

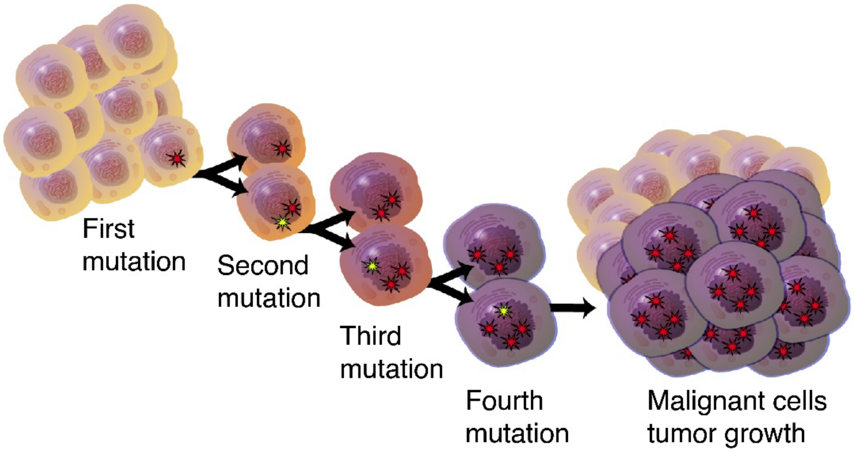


Figure 1.1: Standard model of genetic mutation-driven tumour evolution. Credit: Darryl Leja, National Human Genome Research Institute.

The collection of active cancer driver genes and mutations varies greatly between different cancer types, but also between different tumours of the same type, and even between cancer cells of the same tumour (Gerlinger et al., 2012; Park et al., 2010). These differences have important clinical significance, and are ultimately reflected in diagnosis, treatment and prognosis (Malone et al., 2020; Oser et al., 2015). In addition to wildly variable penetrance, driver mutations are often hidden between many other genetic lesions that are phenotypically neutral but equally passed along the lineage, known as passenger mutations (Lawrence et al., 2014; Stratton et al., 2009). For these reasons, understanding the particular molecular mechanisms behind every cancer is as valuable as it is challenging.

1.2.1 Sequencing in cancer genomics

NGS has been especially helpful for cancer research. First, because obtaining and comparing the genomes of tumour and normal cells from the same individual allows the identification of somatic mutations (Dou et al., 2018; Watson et al., 2013). Second, because the most obvious sign that a somatic mutation is positively selected for driving cancer is their statistically significant recurrence across different tumour samples (Martincorena et al., 2017). Thus, the abundance of sequencing data from different patients has provided the required statistical power to identify driver genes and mutations from the vast majority of effectively neutral passenger mutations, regardless of any previous knowledge of the context of the affected gene products (Campbell et al., 2020). Thereby, proteins, protein regions and protein positions in which mutations occur with a higher frequency than expected by random chance, immediately become putative cancer-driving candidates that warrant further investigation (Kim & Jeong, 2019; Yang et al., 2015).

Coordinated efforts to sequence thousands of cancer genomes, such us The Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas Research et al., 2013) or the International Cancer Genome Consortium (ICGC) (The International Cancer Genome Consortium et al., 2010), have provided a massive collection of genetic changes associated to all possible cancer types (Campbell et al., 2020). Although mutations in non-coding regions have gained more attention in the last years and some driver mutations have been identified—such as those in the promoter of the telomerase reverse transcriptase gene in melanomas (Horn et al., 2013)—, they are still relatively unusual (Elliott & Larsson, 2021). Thus, while cancer sequencing projects have increasingly moved from whole-exome (Kandoth et al., 2013) to whole-genome sequencing (Campbell et al., 2020), the focus has traditionally been in the identification of driver mutations affecting protein-coding genes. In consequence, the catalogue of genes causally implicated in cancer, and the variants in those that lead to the acquisition of their oncogenic properties, have been greatly expanded in the last decade (Sondka et al., 2018). These efforts, together with advances in molecular oncology, have led to determining the hallmarks of cancer, a series of biological capabilities that are successively acquired by cells in their evolution to tumours, providing a framework to understand the neoplastic process (Hanahan & Weinberg, 2011) (Figure **1.2**).

Comparative genomics has also revealed patterns of somatic mutations across cancer samples that are distinctive for both types of drivers. Tumour suppressors are characterized by inactivating, usually truncating, mutations spread across multiple sites in their sequences, while in oncogenes, activating and mostly missense mutations are usually located in well-defined hotspots (Martincorena et al., 2017; Vogelstein et al., 2013) (e.g. Figure **1.3A**). All currently confirmed human cancer driver genes are catalogued according to their role in the disease in the Cancer Gene Census (CGC), an ongoing project from the Catalogue Of Somatic Mutations In Cancer (COSMIC) (Sondka et al., 2018; Tate et al., 2019), that provides expert-curated functional descriptions of these genes. In COSMIC v95, the CGC

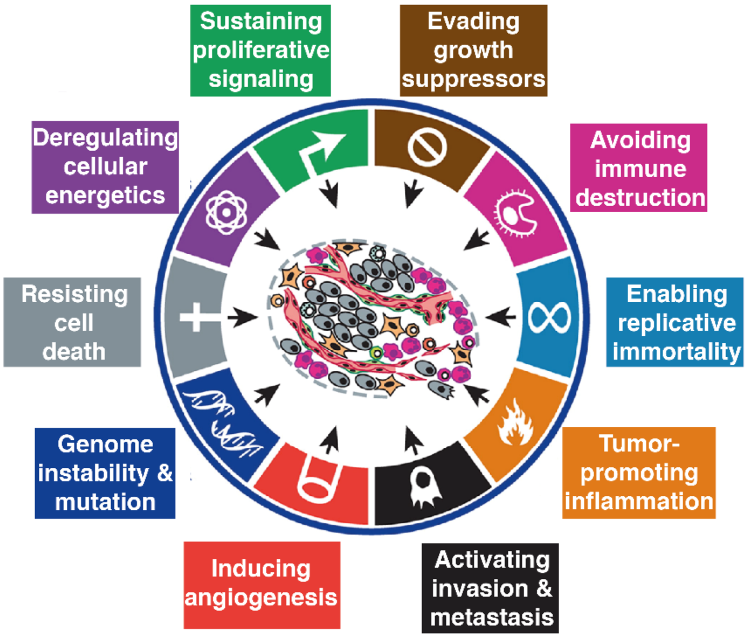

Sustaining proliferative signaling
Evading growth suppressors
Avoiding immune destruction
Enabling replicative immortality
Tumor-promoting inflammation
Activating invasion & metastasis
Inducing angiogenesis
Genome instability & mutation
Resisting cell death
Deregulating cellular energetics

Figure 1.2: The hallmarks of cancer. The ten acquired capabilities that are required for neoplastic transformation. Adapted from its original publication (Hanahan & Weinberg 2011).

comprises 563 genes exactly divided into 50% oncogenes and 50% TSGs if excluding fusion genes, those that due to chromosomic rearrangements are involved in fusions with other oncogenes or TSGs (Mertens et al., 2015) (Figure **1.3B**). A set of 73 genes are classified as both oncogenes and TSGs, suggesting that they are supposed to act through opposite mechanisms. There is indeed increasing evidence supporting the notion that many proteins can have oncogenic or tumour-suppressing functions depending on the cellular context. For instance, this is the case of many transcription factors (e.g. TP53, RB1, or FOXO) (Shen et al., 2018) , but also of other proteins, such as Rho small GTPases (e.g. RHOA, RAC1), which are signal transducers involved in many biological processes (Zandvakili et al., 2017). These further highlight the importance of biological context in fully understanding the molecular mechanisms of gene variants underlying neoplastic transformation, or any other disease.

1.3 Identifying disease variants

The first step of variant interpretation is to determine which are causal of the disease. Many computational methods have been developed with the purpose of prioritizing the most likely pathogenic variants, and applied both to germline and somatic. In cancer, high allele frequency has been widely used to identify candidate driver genes and variants (Figure **1.4**).

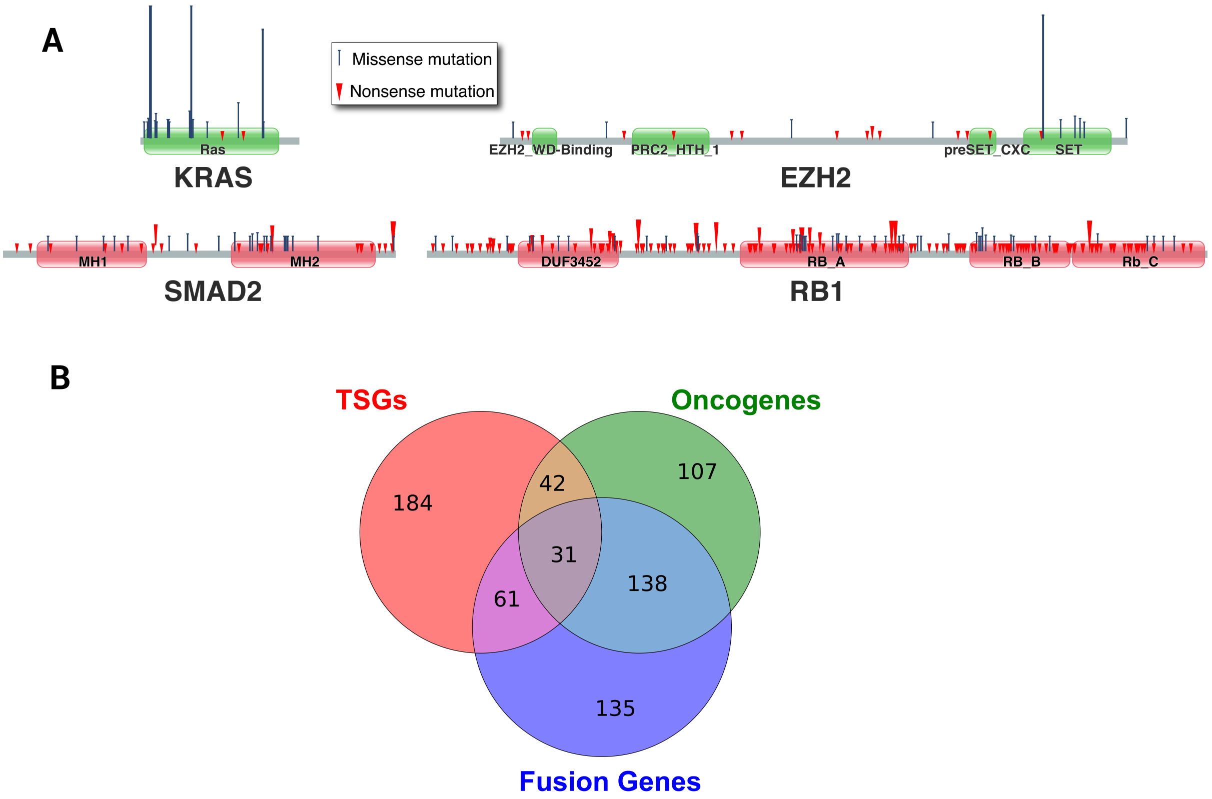


Figure 1.3: **(A)** Distribution of somatic missense mutations and nonsense mutations sites across the sequence of oncogenes KRAS and EZH2 and tumour suppressors SMAD2 and RB1. Lengths of mutation flags are proportional to sample count. Data were obtained from COSMIC. All mutations come from whole-genome sequencing, and are both confirmed somatic and predicted as pathogenic. **(B)** Classification of the Cancer Gene Census genes according to their role in cancer. A significant number of genes belong to more than one category.

However, this can often be too simplistic. The steep growth of genomic data increased also the risk of obtaining false positives, which in turn propelled the development of more accurate background models to account for mutational heterogeneity when determining mutation rates (i.e. gene-specific models) (Lawrence et al., 2013). Moreover, others have identified groups of otherwise low-frequency mutations that are functionally equivalent, by looking for patterns of mutual exclusivity across different samples (Canisius et al., 2016). A study from our own group identified such patterns between several genes, like for instance inactivating mutations in G_i/G_o-protein coupled receptors and oncogenic mutations in $G_s\alpha$, which analogously enhance cAMP signalling (Raimondi et al., 2019).

In general, most conventional variant prioritization tools have heavily relied on phylogenetic conservation for assessing putative impact, on the basis that mutations on highly conserved residues are more likely to be deleterious (e.g. SIFT (P. Kumar et al., 2009)

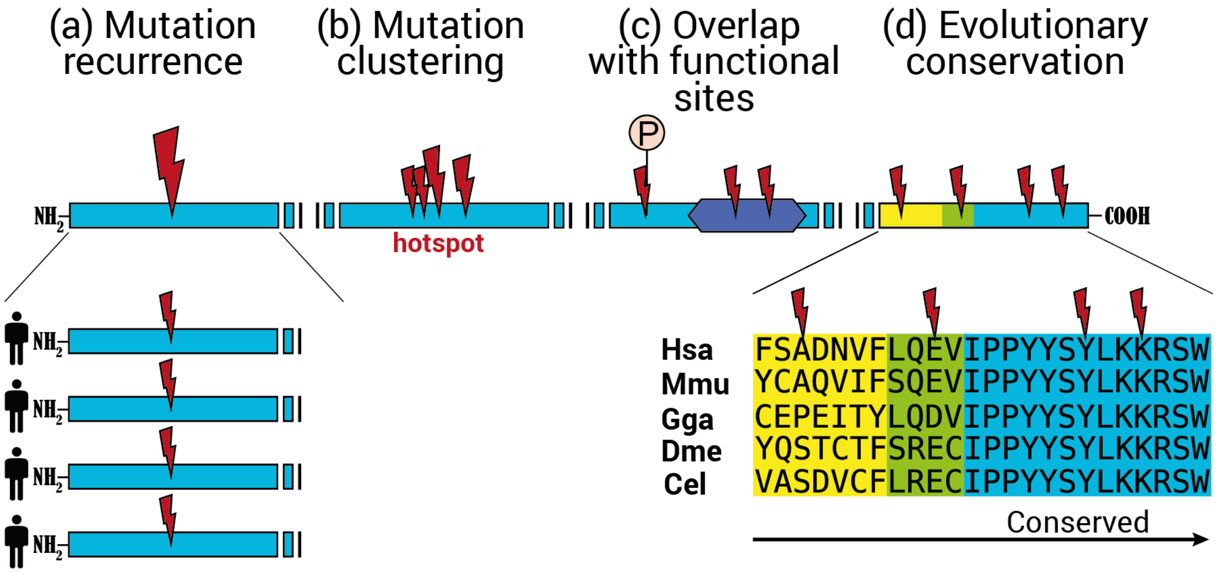

(a) Mutation recurrence
(b) Mutation clustering
(c) Overlap with functional sites
(d) Evolutionary conservation
P
NH2
COOH
hotspot
Hsa FSADNVFLQEVIPPYYSYLKKRSW
Mmu YCAQVIFSQEVIPPYYSYLKKRSW
Gga CEPEITYLQDVIPPYYSYLKKRSW
Dme YQSTCTFSRECIPPYYSYLKKRSW
Cel VASDVCFLRECIPPYYSYLKKRSW
Conserved

Figure 1.4: Main features traditionally exploited by variant prioritization methods: (a) recurrence across different samples/individuals (or allele frequency); (b) non-uniformal distribution; (c) overlap with functional sites, such as domains or phosphosites; (d) phylogenetic conservation of affected residues.

and MutationAssessor (Reva et al., 2011)) (Figure **1.4**). Some algorithms have also been trained using features based on protein sequence and known or predicted structure (e.g. PolyPhen-2 (Adzhubei et al., 2010) and CHASM (Carter et al., 2009, 2010)), as well as knowledge from already established disease-causing polymorphisms (e.g. MutationTaster2 (Schwarz et al., 2014)). Others search for mutations clustered along the protein sequence (e.g. MuSiC (Dees et al., 2012) and OncodriveCLUST (Tamborero et al., 2013)) or for mutations overlapping with diverse protein functional sites (e.g. ActiveDriver (Reimand & Bader, 2013)) (Figure **1.4**). In addition, some tools combine the scores from several of these algorithms in order to obtain a more consensual prediction (e.g. CADD (Kircher et al., 2014) and OncodriveFM (Gonzalez-Perez & Lopez-Bigas, 2012)). The advantage of all these approaches is that they can be quickly applied to the typically large number of variants identified in any sequencing experiment to obtain a reduced list of variants that can be ranked by a predicted pathogenicity score. However, they do not provide information on how the effect of these variants is achieved, typically only identifying whether variants are "pathogenic" or "damaging". Moreover, since they focus on individual protein features, they usually ignore all functional or mechanistic context.

1.4 A gulf between genetics and molecular biology

The large volume of sequencing data and the fast pace of new disease variant discovery cannot be matched by the experiments required to accurately characterize their functional impact. Understanding what variants might do to function is moreover often limited by a poor understanding of the affected genes or pathways. As a result, there is a growing number of confidently classified disease-causing variants that lack any reasonable explanation for the underlying mechanisms that translate them into the disease phenotype; there is thus a gulf between the fields of genetics and molecular biology (González-Sánchez et al., 2018).

Fortunately, thanks to a parallel development of high-throughput proteomics techniques, structural biology and computational power in the last years, mechanistic data are also on the rise in the form of biomolecular structures, interactions, or post-translational modifications. Accordingly, different new methods and systematics studies have integrated mechanistic with sequencing data to provide new context-specific interpretations for disease genetic variation.

1.5 Understanding molecular mechanism

1.5.1 Pathways and networks

Pathway analysis is one way to study genetic variants in a functional context. First, this higher level of organization allows to find rare, and a priori independent, genetic alterations that are functionally related because they affect common pathways (Figure **1.5**). Second, identified variants are already associated to familiar biological processes and are easier to interpret. Under this principle, several pathway- and network-based methods have been applied to cancer data sets to identify new driver genes and their mechanisms, and expand the repertoire of disturbed cellular functions (Akavia et al., 2010; Creixell, Reimand, et al., 2015). In order of complexity, these approaches include: gene set enrichment analysis (GSEA) (Subramanian et al., 2005), which consists in the identification of enrichment of fixed gene sets associated to certain biological categories (e.g. g:Profiler (Raudvere et al., 2019); de-novo construction of interaction networks with mutated genes, which can help discover non-mutated genes likely to be involved in the disease (e.g. GENEMANIA (Franz et al., 2018), STRING (Szklarczyk et al., 2019)); and pathway modelling, which tries to predict how the activity of known networks or pathways is altered by mutations in terms of changes in different qualitative and quantitative parameters (e.g. PARADIGM-SHIFT (Ng et al., 2012)).

For example, using PARADIGM-SHIFT, our group found particular pathway alterations preferences for different cancer types, like nephrin/Neph1 signalling in kidney tumours and ephrin A reverse signalling in thyroid cancer (González-Sánchez et al., 2018).

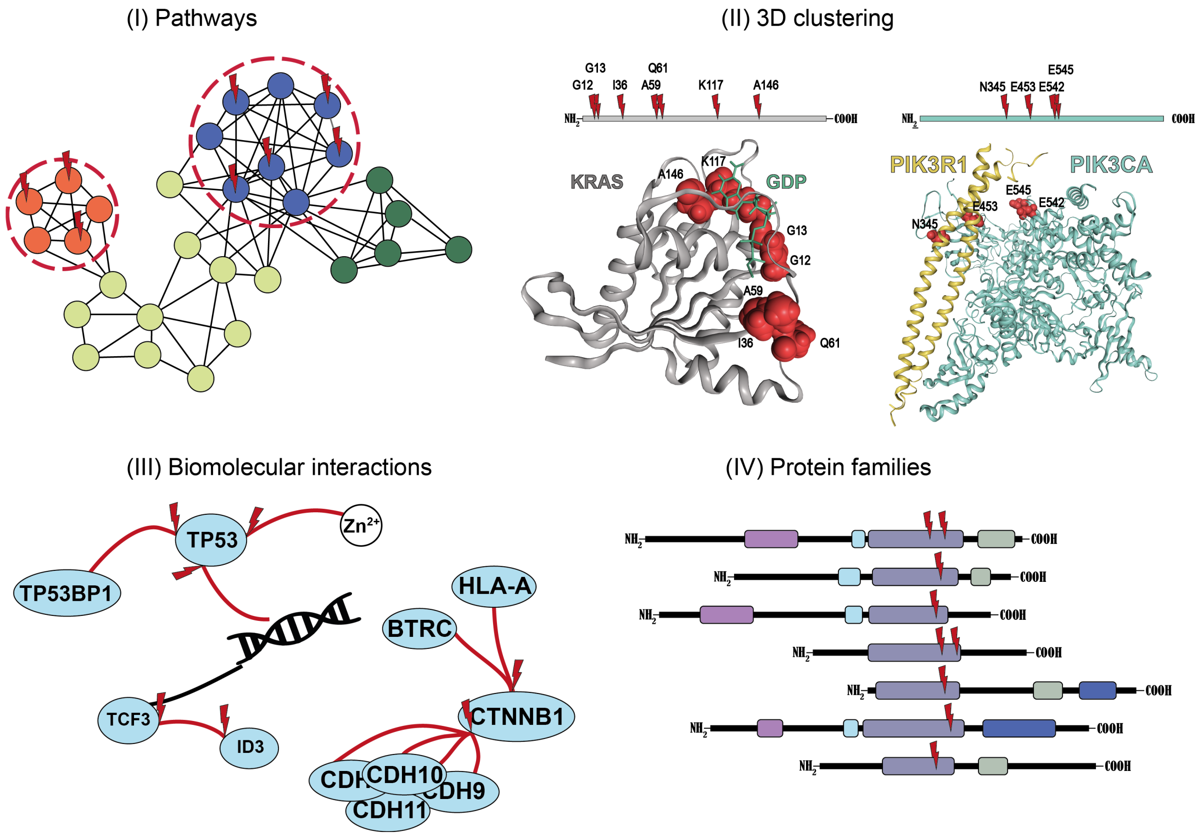


Figure 1.5: Approaches at multiscale resolution used by variant interpretation tools providing deeper mechanistic insights: (a) identification of significantly mutated sets of genes that are all part of the same pathway; (b) identification of mutations that are not proximal on the protein sequence but that cluster on the 3D structure, commonly affecting binding sites, for instance; (c) analysis of *edgetic* effects, different variants in the same protein can affect different interfaces; (d) detection of otherwise scarce mutations in different proteins that are located in equivalent positions within a shared functional domain, thus resulting in a similar effect.

1.5.2 Structures and interactions

Biological processes are mediated by intricate networks of protein-protein interactions (PPI), comprising fast and transient contacts or stable macromolecular complexes, but also interactions with other (i.e. non-protein) biomolecules (Robinson et al., 2007). Large-scale and targeted interaction discovery experiments have illuminated the complex landscape of protein interactions in both human (Havugimana et al., 2012; Huttlin et al., 2015; Rolland et al., 2014), and other model organisms (Li et al., 2004; Rajagopala et al., 2014; Yu et al., 2008). Although many interactions are relatively well understood, for example, those linking important cellular machineries like the ribosome (Ben-Shem et al., 2011) or nuclear pore complex (Alber et al., 2007), others remain more elusive. The total number of known, experimentally validated PPI thus far is estimated to represent only a fraction of the complete interactome for any organism, including humans (Luck et al., 2020). Some resources have tried to fill this gap by integrating computationally predicted interactions, based on genomic context, co-expression, transfer between organisms, or automated text mining (Franz et al., 2018; Szklarczyk et al., 2019).

In parallel, efforts from structural biologists have resulted in an increasing number of high-resolution three-dimensional (3D) structures of single and interacting proteins—typically deposited in the Protein Data Bank (PDB) (Berman et al., 2000)—that in turn has allowed the understanding in full molecular detail of how proteins interact. Although structures are far from being available for all known protein interactions, homologous pairs of interacting proteins were shown to bind through similar interfaces (Aloy et al., 2003), even to the point that it was suggested early on that there is a limited number of interaction types in nature (Aloy & Russell, 2004). Consequently, computational approaches based on homology modelling (e.g. Interactome3D (Mosca et al., 2013)) have expanded the structural coverage on both known single PPI and large protein complexes, as well as predicted completely new interactions (Aloy et al., 2004; Aloy & Russell, 2003).

For variant interpretation, resources like Mechismo (Betts et al., 2015) or dSysMap (Mosca et al., 2015) have systematically integrated protein interactions and structures with variant data to predict their functional impact and identify putative mechanisms (e.g. (Rohde et al., 2014), Figure **1.6**). Other methods look for mutational clusters in 3D space—groups of otherwise rare missense mutations that are in close proximity in structure— based on the assumption that they might elicit similar functional consequences and phenotypes (Figure **1.5**). Such 3D clusters have been found in both known and potentially new cancer genes (Fujimoto et al., 2016; Gao et al., 2017), many located within binding interfaces with proteins, nucleic acids and other small molecules, thus already hinting at a putative mechanism of action (Kamburov et al., 2015; Porta-Pardo et al., 2015). Analyses following these methodologies have highlighted several instances where different mutations in the same driver gene perturb different interaction interfaces. These interaction-specific—or *edgetic*—effects (Zhong et al., 2009) can lead to distinct phenotypes that might correlate with cancer severity in some cases (Raimondi et al., 2016).

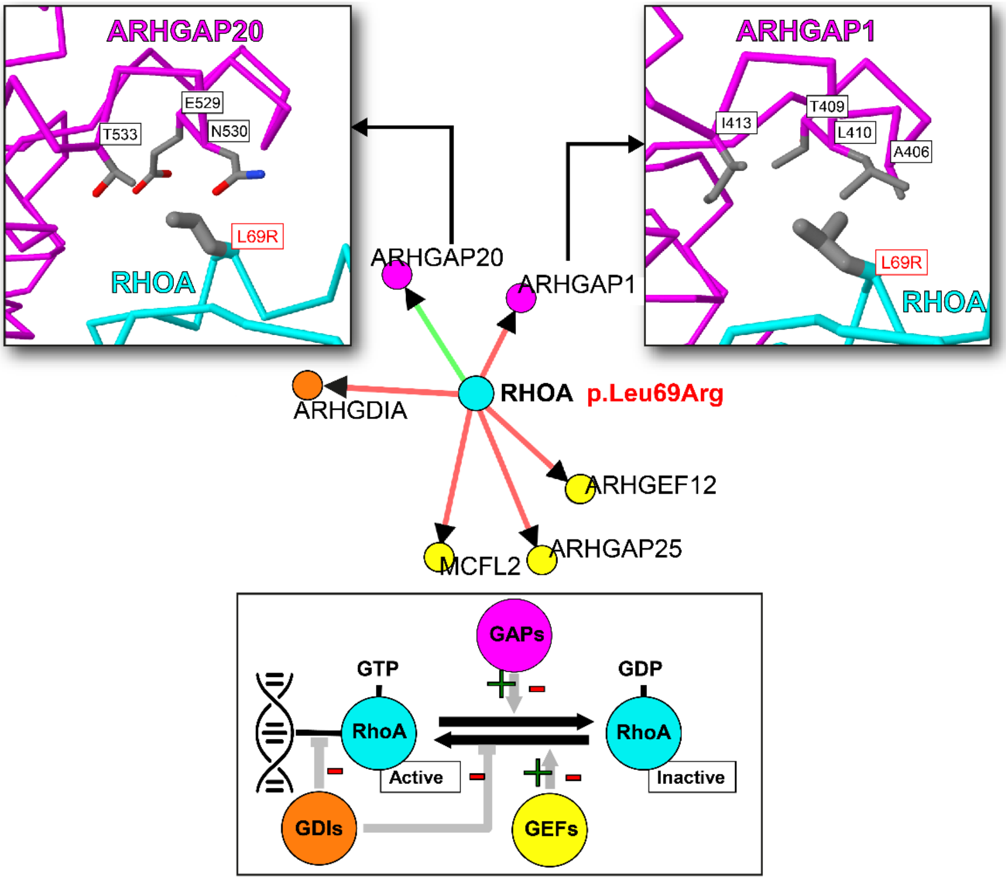


Figure 1.6: Predicted effects of the small GTPase RHOA variant p.Leu69Arg (found in pediatric Burkitt lymphoma) on different interactions, using Mechismo (Betts et al., 2015). The same variant is predicted to have a negative impact (red arrows) on the interaction with several Guanine exchange factors (GEFs), a Guanine dissociation inhibitor (GDI), and the GTPase-activating protein (GAP) ARHGAP1, while also having an enhancing effect (green arrow) on the interaction with the GAP ARHGAP20. Top panels show the structural context of the Leu69, and how it lies in a polar/negatively charged pocket in ARHGAP20 (top left), but in a hydrophobic pocket with ARHGAP1 (top right), which are respectively favourable and unfavourable for a positively charged Arg. The bottom panel shows a schematic of how the balance between RHOA active and inactive forms is maintained by the action of GAPs, GEFs and GDIs. Rather than clearly oncogenic or tumour suppressive, the suggested mechanism for this variant was a subtler interaction tinkering that results in the shift of RHOA towards particular pathways. Figure adapted from (Rohde et al., 2014).

Structure-based approaches are bottlenecked by the availability of template structures. Fortunately, the recently published AlphaFold2 (Jumper et al., 2021), which widely outperforms every other method, has produced such highly accurate predictions that the community of structural biology has called the protein folding problem to be solved. While this is certainly bound to have a tremendous impact in the advancement of molecular biology —in barely a year, the structural coverage of the entire human proteome has expanded from 17% to 58% of residues and to 98.5% of proteins (Tunyasuvunakool et al., 2021)—, my colleagues and I have argued that its impact on protein variant interpretation might be less significant than anticipated, owing for instance to the increasing predominance of disease variants in disordered or tandem-repeat protein regions and in protein interfaces, which are more challenging or impossible to model) (Diwan et al., 2021).

1.5.3 Domains and protein families

Proteins are generally composed of domains: conserved protein regions that fold into stable 3D structures and behave as separate units that can function and evolve independently (Ponting & Russell, 2002). Proteins belonging to the same family typically share the same function and domain composition. This has been exploited to identify rare somatic mutations that affect conserved, functionally equivalent residues within shared protein domains, and that otherwise would escape detection (Miller et al., 2015; Peterson et al., 2017) (Figure **1.5**). Domain-centric approaches have detected mutation hotspots at the domain-level both in oncogenes and TSGs in many different cancer types, hinting not only at new potential drivers but also at the mechanistic consequences of these mutations, revealing similarities in mechanism between mutated proteins (Yang et al., 2015).

Moreover, domains also have an important role mediating interactions. Through the analysis of interfaces in the 3D structures of protein complexes, thousands of domain-domain interactions (DDIs) have been identified and classified (Finn et al., 2014; Mosca et al., 2014). In addition, several computational methods have expedited the discovery of more DDIs through the detection of correlated domain signatures in protein-protein interactions (Deng et al., 2002; Sprinzak & Margalit, 2001), based on the notion that recurrently finding a pair of domains in interacting protein pairs might be indicative that these domains mediate the interaction. Because domains behave as independent interacting elements in the protein, DDIs can be used to infer putative binding mechanisms for other known interacting protein pairs or even to predict novel interactions. DDIs thus extend the landscape of mechanistic possibilities that can help interpret disease variants (Yang et al., 2015).

1.5.4 Linear motifs

Domains can also interact with smaller interfaces, known as short linear motifs (SLiMs), that play critical functional roles in the cell. In contrast to domains, SLiMs only comprise between 3 to 15 amino acids and are generally located in disordered—flexible and easily accessible—protein regions (Davey et al., 2012). Binding motifs typically contain only a few key residues that are actually involved in the interaction, and thus they mediate weak, transient and, consequently, reversible PPIs, which are essential for the dynamic networks that regulate many cellular processes (Davey et al., 2012; Perkins et al., 2010). Furthermore, motifs also act as post-translational modifications sites which are directly and specifically recognized by regulatory enzymes. Motifs thus have a key role in functions like cell signalling, protein trafficking, modification and degradation (Dinkel et al., 2012; Neduva & Russell, 2005; Van Roey et al., 2014). Due to their low-affinity, domain-motif interactions are difficult to identify experimentally; but thanks to curation efforts, there is an increasing catalogue of functional motif instances (notably the Eukaryotic Linear Motif (ELM) resource[2] (Kumar et al., 2019), and thanks to computational methods, there are means of discovering new protein-motif pairs (Neduva & Russell, 2006).

It has been shown that a significant fraction (~22%) of disease mutations occur in intrinsically disordered regions (Vacic et al., 2012). Moreover, in these regions mutations are enriched in SLiMs, and tend to occur at functionally important residues within them (Uyar et al., 2014). Mutations affecting SLiMs can deregulate many processes and have disastrous effects for the cell. A missense mutation of a single key residue of a SLiM is often enough to ablate function, but even mutations in flanking residues can have consequences (Van Roey et al., 2014). For instance, mutations associated with Noonan syndrome were found in a 14-3-3 binding phosphopeptide motif in RAF1 (ELM motif: LIG_14-3-3_CanoR_1)[3], resulting in the inhibition of the interaction and an overactive RAF1 mutant (Pandit et al., 2007).

1.5.5 Post-translational modifications

Post-translational modifications (PTMs) alter the physicochemical properties of a single residue, which might also affect protein properties such us stability, folding, binding affinity with other molecules, and thus modulate protein function. Thanks to high-throughput proteomics coupled with curation efforts, data regarding all types of modifications has increased substantially in recent years, particularly for the best studied modifications: phosphorylation, acetylation, ubiquitination, methylation, glyosylation and diverse types of lipidation; such data is accessible in different repositories like PhosphositePlus (PSP) (Hornbeck et al., 2015) or UniProt (Bateman et al., 2021). Knowledge of experimentally confirmed PTMs can help determining real motif instances, as due to the short size and often

[2] http://elm.eu.org
[3] http://elm.eu.org/elms/LIG_14-3-3_CanoR_1.html

simple motif patterns, prediction of new SLiM instances by simple sequence matching usually identifies a large number of false positives (not functional instances).

While natural human genetic variation generally avoids PTM sites (Reimand et al., 2015), dysregulation of PTMs is targeted by genetic alterations in several diseases, including most cancers. Indeed, PSP also provides with a sub-dataset of more than 25,000 PTMs that intersected with variants from thousands of genetic diseases and all types of cancers (Hornbeck et al., 2015). In particular, phosphorylation and cancer have been the subject of intense study, both at the level of regulatory enzymes (kinases and phosphatases) and phosphosites, highlighting the role of protein signalling rewiring as a prominent cancer driving mechanism (Creixell, Schoof, et al., 2015; Reimand & Bader, 2013). Other types of PTMs have more recently emerged with similar roles in disease, including acetylation and ubiquitination (Narayan et al., 2016), lysine methylation (Carlson & Gozani, 2016) and many more (Krassowski et al., 2021).

1.6 Thesis outline

The general aim of this thesis is to enhance the mechanistic understanding of protein variants through the prism of Systems Biology, i.e., considering their effect on the multiple components of the integrated molecular system controlling cellular activity. This work has been possible through the use of a wide range of computational tools and large datasets, all of which are properly credited. Each of the next three chapters includes their own introduction to general concepts, methodology, results and discussion, and conclusions.

In Chapter 2, I present a systematic study of a large genome-wide pan cancer mutation dataset to identify instances where positive selection of particular amino acid substitutions could be hinting at interaction-specific, possibly interaction switching events. These instances are then explored more in detail in order to make further mechanistic hypotheses, and discussed in the context of challenging the classical oncogene/TSG paradigm.

In Chapter 3, I deal with another tangential problem in the task of achieving a mechanistic understanding of proteins and their variants: the need of integrating data from numerous and diverse databases and resources, coupled with the difficulty of simultaneously visualizing those insights. As a solution, I present Mechnetor, an online tool that automatically performs these tasks, and provides interactive visualizations that facilitate investigating protein mechanism (González-Sánchez et al., 2021). Moreover, I discuss case studies that showcase the advantage of examining protein sequence features in combination with various interaction information, as it can provide a detailed mechanistic understanding of disease-related variants.

Finally, in Chapter 4, I focus on a proteome-wide study of a particular type of PTM, *S*-palmitoylation, in *Drosophila melanogaster*, a model organism where this modification has barely been studied before. In this proteomics project, a close collaboration with Dr. Elena Porcellato, Dr. Christoph Metzendorf and others from the group of Prof. Dr. Felix T. Wieland (at Heidelberg University Biochemistry Center), we first identified and characterized the most complete palmitoylome in *Drosophila* to date, and then coupled it with a high-throughput interaction study in order to identify the potential spectrum of client protein-enzyme (palmitoyl-acid transferases) (Porcellato et al., 2022).

Although independent, these chapters are, overall, thematically linked and some concepts and terminology might be shared. Thus, reading them in order is advisable.

Chapter 2

Investigating positive selection of interaction-perturbing mutations in cancer

2.1 Introduction

Many studies have already highlighted the contribution of amino acid changes at protein interfaces to human disease, evidenced by their enrichment at protein interaction interfaces (David et al., 2012). These mutations can result in structural and physicochemical changes that affect the stability and conformation dynamics of PPI interfaces leading to the loss of the interaction (Schuster-Böckler & Bateman, 2008), although their effects can also be more nuanced, resulting in a weaker or even a stronger interaction (Kucukkal et al., 2015). Moreover, disease mutations can have interaction-specific or 'edgetic' effects, meaning that the same variant in a particular interface does not affect all interactions equally, it might only impede certain interactions while at the same time others are stimulated or not affected at all (Sahni et al., 2015; Zhong et al., 2009). Potentially, two different variants at the same interface might have complete opposite effects that lead to a different phenotype. Edgetic perturbations and their subtler effects are often neglected in classic disease variant interpretations, particularly in cancer where everything is scrutinized under the oncogene/TSG paradigm, but can provide a mechanistic model to understand more complex relationships between genotype and phenotype.

From an edgetic viewpoint, complete protein activation or inactivation (gain and loss of function) can be achieved solely by changes in interactions. In some other cases though, the effect of these changes could be rather interpreted as a switch of function, in the sense that a switch from one set of interactions to another results in an altered protein function (Reva et al., 2011), like for instance, the single Burkitt lymphoma missense RHOA variant p.Leu69Arg, showcased in the previous chapter (Figure **1.6)** (Rohde et al., 2014). Interaction fine-tuning as a driving cancer mechanism can only be the result of subtle missense mutations, implying that the amino acid substitutions of the residues located at a targeted protein interface are probably very specific in order to simultaneously favour and/or disfavour particular interactions, without having more drastic effects.

In this chapter, I extend the work on this subject. The availability of many thousands of genomes for cancer patients provides interesting possibilities to study selection at work within cancers—whether collectively or as individual diseases—and to see if this selection can be illuminating or predictive about biological function and disease progression. Here, I focus in particular on the relationship between amino acid changes preferences in cancer and interaction switching events. To determine these preferences, it is necessary to establish a reference background to which cancer-related changes can be compared with. Many studies have previously used random-generated background models, even accounting for cancer type-, gene- and patient-specific mutation rates (Lawrence et al., 2014; Youn & Simon, 2011). However, non-disease related mutations or natural variants, which are also freely available in large repositories such as the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2015), represent a more suitable experimental benchmark dataset of changes that are in theory phenotypically neutral (de Beer et al., 2013). Using this, first I did a general characterization of amino acid substitutions preferences and the patterns underlying them in human cancer. Next, I identified hundreds of positions within cancer datasets that appear to prefer changes likely to have drastic effects on interactions. I review a few interesting cases where cancer variants cause specific edgetic perturbations in interaction networks, as predicted through tertiary structure, that might lead to distinct phenotypic consequences. For some examples, this finding is borne out by oncogenic mechanisms already known, though for others these could imply potentially novel mechanisms that warrant further investigation.

2.2 Results and discussion

2.2.1 Cancer and natural variants datasets

As source of cancer variants, I used the COSMIC database (v94) (Tate et al., 2019). COSMIC, or The Catalogue Of Somatic Mutations In Cancer, is the largest repository of human cancer somatic mutations, curated by experts from thousands of peer-reviewed publications of both gene-targeted and genome-wide screens, including also data from the TCGA (The Cancer

Genome Atlas Research et al., 2013) and the ICGC (The International Cancer Genome Consortium et al., 2010). Comprising nearly 1.5 million samples, it provides an ample coverage of the cancer somatic mutation landscape and the mechanisms that promote the disease. In order to avoid biases due to targeted sequencing, I only used data coming from whole-genome sequencing (WGS). In total, I collected more than 190 thousand missense somatic variants, belonging to more than a hundred cancer types (full numbers in Figure **2.1**; Methods **2.3.1**).

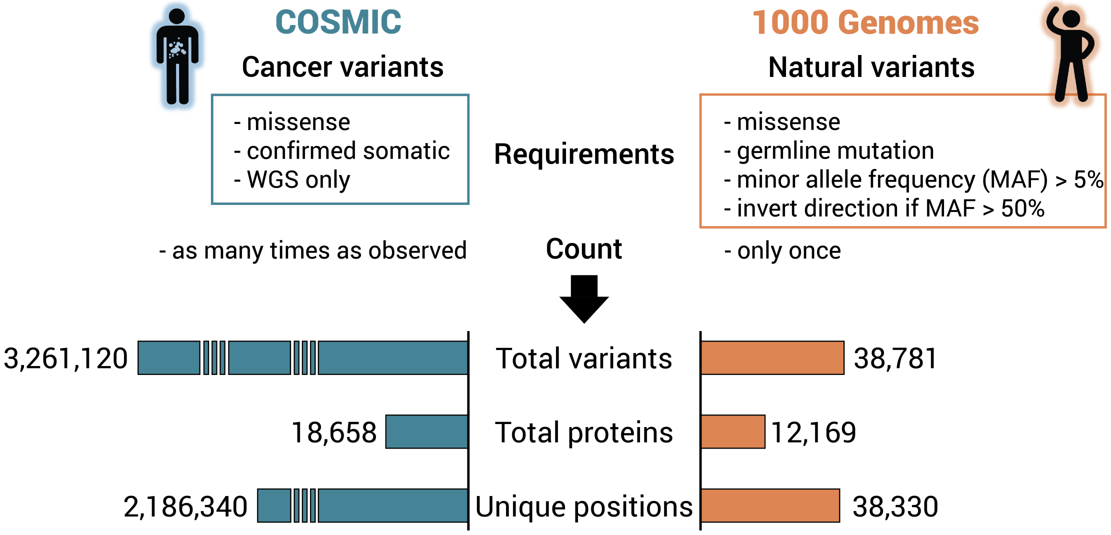


Figure 2.1: Cancer and natural mutations were extracted from COSMIC and 1kG Project datasets and processes, yielding significantly different numbers.

To properly characterize and assess the significance of cancer variants, I also gathered natural variants from the 1000 Genomes Project dataset (1kG) (1000 Genomes Project Consortium et al., 2015). The 1000 Genomes dataset was (now archived) a catalogue of common human genetic variation, created from sequencing 2,504 healthy individuals belonging to five mayor populations in the globe, thus it can be used as a natural background of germline amino acid changes in humans to determine differences of disease-associated variants. Despite this, to avoid rare variants that could potentially be related to disease, I only extracted those with a minor allele frequency >5%, to obtain more than 38 thousand non-synonymous variants (Figure **2.1**).

The huge disparity between the total numbers of somatic and natural variants is due to two factors. First, the COSMIC database comprise data from a much larger number of individuals and it continues to grow, while 1000 Genomes Project was a finite effort of capturing human genetic variation completed in 2015. Second, natural variants are only counted once, independently of the number of individuals that have them. This is because the same minor allele occurring in several individuals is most likely due to inheritance and not to different mutational events. Cancer variants in contrast represent somatic mutations that occurred in different samples independently, and thus are counted as many times they are observed.

2.2.2 General evaluation of amino acid exchange preferences in cancer

2.2.2.1 Cancer and natural variants are differentially distributed

Using recently predicted 3D structures for almost all human proteins (Tunyasuvunakool et al., 2021; Varadi et al., 2022), I calculated the relative solvent accessibility of all residues affected by either natural or cancer somatic variants to found that somatic variants are located in slightly less solvent-accessible (or buried) positions than natural polymorphisms (29% and 17% respectively; Figure **2.2**); results that are in line with previous studies of genetic disease-associated mutations (de Beer et al., 2013; Savojardo et al., 2021). The general notion is that amino acid substitutions in buried positions are more likely to affect protein stability (Martelli et al., 2016; Sunyaev et al., 2001). Similarly, I assessed the distribution of variants within or outside functional protein regions, which were defined by the presence of Pfam domains (Mistry et al., 2021). Also as expected, somatic variants were found to be more enriched in residues located in functional domains (56% versus 42%; Figure **2.2**). Lastly, in line with other studies (Yates & Sternberg, 2013), somatic variants were also found to be significantly over-represented in interaction interfaces predicted through tertiary structure (see Methods **2.3.2**) compared to natural variants (17% versus 4%), supporting the edgetic notion of interaction perturbations as a disease-leading mechanism.

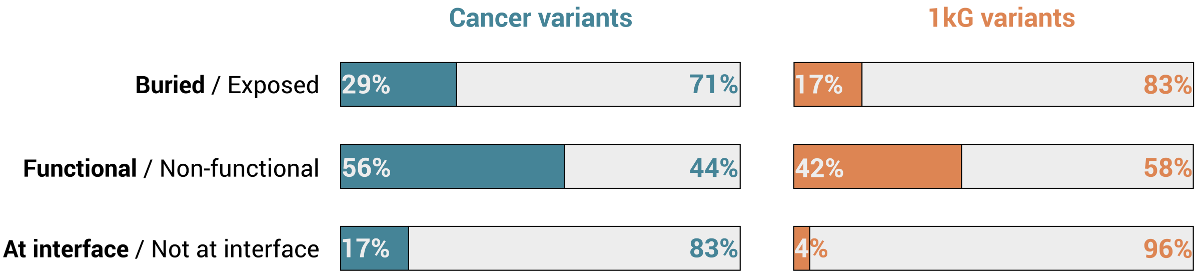


Figure 2.2: Distribution differences of residues affected by cancer or 1kG variants. Cancer variants are more likely to affect buried, functional and/or interface residues.

2.2.2.2 Cancer and natural variants have similar amino acid mutabilities

The probabilities of the 20 amino acids to mutate, or mutabilities, are very different (Creixell et al., 2012). de Beer et al. established that the different amino acid mutabilities observed in natural variants of the 1000 Genomes Project mostly reflect underlying genetic properties, such as the degeneration of the genetic code, the codon usage biases, and the diverse mutation rates of codons due to their CpG content; while in comparison, the impact of protein structure restrictions on mutabilities was found to be relatively small (de Beer et al., 2013). For example, arginine (Arg) has by far the highest mutability among all amino acids (Cooper & Youssoufian, 1988). This arises from the fact that four of the six codons that code for Arg contain the CpG dinucleotide, which is known to have a mutation rate 10-50 times higher than other dinucleotides (Coulondre et al., 1978; Walser & Furano, 2010). de Beer et al also compared the mutabilities of amino acids between the 1000 Genome variants and a dataset of disease-associated variants obtained from OMIM and found no correlation whatsoever (de Beer et al., 2013). In a stark contrast, the comparison of amino acid mutabilities between the 1kG variants and the cancer dataset showed instead a moderate positive correlation (Figure **2.3**). This difference is likely due to the fact that OMIM variants come from many

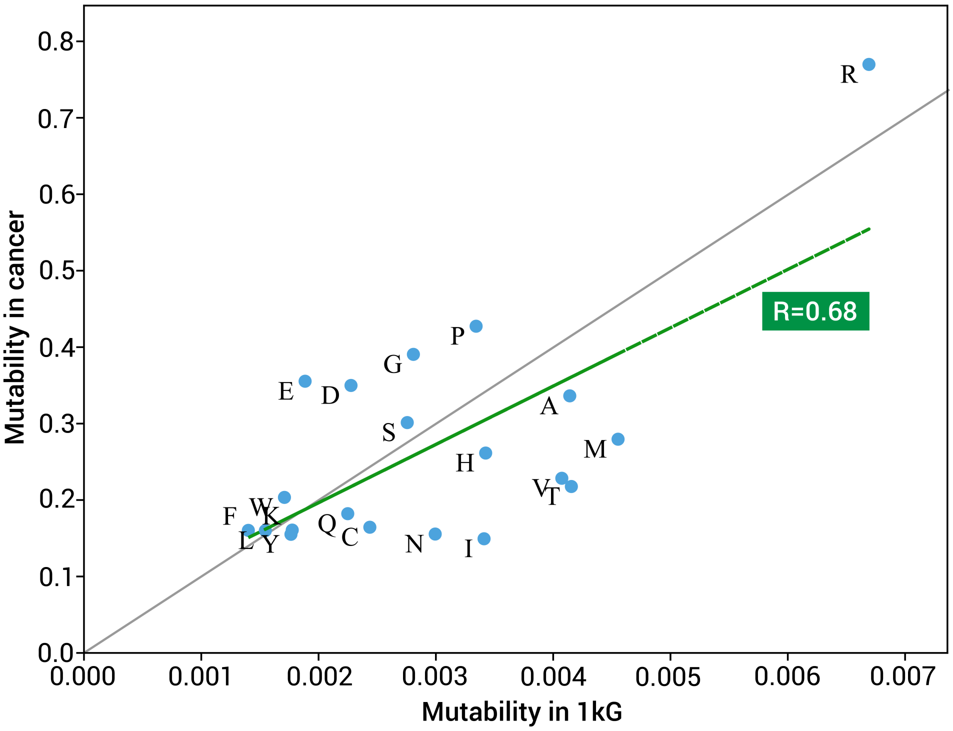


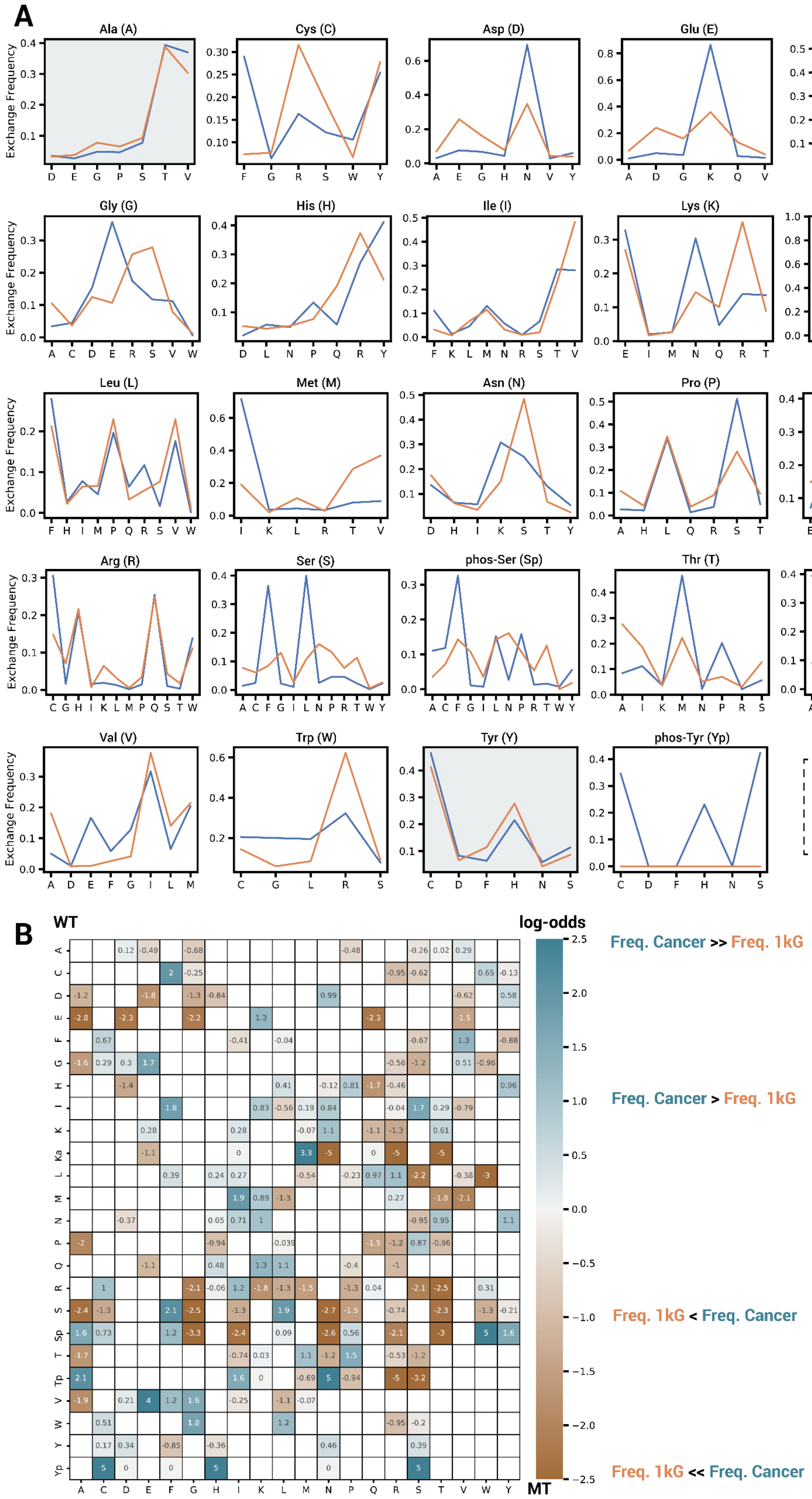
Figure 2.3: Overall, mutability values in cancer variants are much higher than in natural variants, due to a much higher number of mutations. However, proportionally, amino acids show a moderate positive correlation between both data sets.

different experiments targeted to particular proteins. In contrast, the cancer variant dataset used here contains variants from whole-genome sequencing only and thus, it is generally not biased towards any protein or variants. In conclusion, this suggests that, overall, cancer variants are affected by the same constraints at the level of DNA than natural ones, although differences in some amino acids (Arg and Glu mutabilities are higher in cancer, for instance) may be due to positive selection for pathogenicity.

2.2.2.3 Cancer differ from natural variants in residue exchange preferences

The genetic code bias and the diverse codon mutabilities also determine the different probabilities of every amino acid substitution that is accessible through single-nucleotide variants (SNVs), which ultimately are reflected in amino acid exchange rates. For instance, Ala residues are most frequently mutated to either Thr or Val because not only they are obtainable from each of alanine's four codons but also those changes imply a base transition (G→A for Thr, or C→T for Val) rather than a base transversion, which are generally less likely to occur (Collins & Jukes, 1994). In order to account for these general biases and be able to detect trends that are exclusively due to specific selection events, amino acid exchange rates from somatic cancer variants were compared to those from the 1kG natural variants dataset, used as a control. Considering all 150 amino acid substitutions that are accessible through non-synonymous SNVs, a statistical test confirmed that there is a significant difference in their frequencies between both variant types (χ^2 *P*-value << 0.01). Individual comparison of amino acid mutation profiles calculated from both datasets showed that, with the exception of Ala and Tyr, there are considerable differences in the exchange preferences of all other amino acids (Figure **2.4A**). For example, Lys shows a strong preference to mutate to Arg in natural variants, while in cancer this substitution is diminished and Asn is considerably more frequent. For a more accurate assessment, I then calculated the frequency ratios to identify which particular amino acid substitutions were significantly enriched or depleted in cancer (cancer frequency > 1kG frequency *or vice versa*) (Figure **2.4B**).

Figure 2.4: (A) Linear plots showing the mutation profiles for every amino acid calculated from cancer somatic variants (blue) and from 1kG natural variants (orange). For clarity, on the X axis only amino acids accessible by SNVs are displayed. Grey background indicates that exchange frequencies for that amino acid were not found statistically different between both datasets. Mutation profiles for phosphorylated serine (Sp), threonine (Tp) and tyrosine (Yp), as well as acetylated lysine (Ka) are shown separately from their unmodified counterparts. Notice that no natural variants were found on known Yp sites. **(B)** Heatmap that shows the enrichment (blue) or depletion (orange) of the same amino acid substitutions in cancer, calculated as log-odds between the observed frequency (frequency in cancer) and expected frequency (frequency in 1kG). The amino acid rows correspond to the wild type while the columns indicate the mutant amino acids.


A
Exchange Frequency
Ala (A)
Cys (C)
Asp (D)
Glu (E)
Phe (F)
Gly (G)
His (H)
Ile (I)
Lys (K)
ac-Lys (Ka)
Leu (L)
Met (M)
Asn (N)
Pro (P)
Gln (Q)
Arg (R)
Ser (S)
phos-Ser (Sp)
Thr (T)
phos-Thr (Tp)
Val (V)
Trp (W)
Tyr (Y)
phos-Tyr (Yp)
Dataset
Cancer
Natural
B
WT
MT
log-odds
Freq. Cancer >> Freq. 1kG
Freq. Cancer > Freq. 1kG
Freq. 1kG < Freq. Cancer
Freq. 1kG << Freq. Cancer
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
A C D E F G H I K L M N P Q R S T V W Y
A C D E F G H I K Ka L M N P Q R S Sp T Tp V W Y Yp

In particular, there are significant differences in the preferences of Lys, Ser and Thr, and their modified versions. For instance, while for Lys, substitution for Met is rather infrequent and not enriched in cancer, for ac-Lys this is almost the only observed change; which is single-handedly due to the well-known K27M mutation (and known ac-Lys) in Histone 3 (Khuong-Quang et al., 2012). Phos-Ser and phos-Thr both show a significant preference for Ala mutations that contrast with the depletion of this substitution in unmodified Ser and Thr.

2.2.2.4 Cancer variants show a preference for drastic amino acid exchanges

Amino acid physicochemical properties

A glance at these changes while in consideration of amino acid physicochemical properties (Betts & Russell, 2007; W. R. Taylor, 1986) (Figure **2.5**) hints at a general trend where, within the cancer somatic variants, drastic changes are enriched while more conservative ones are avoided, in particular the second part (Figure **2.4B**). For instance, Gly is the smallest amino acid but avoids substitutions for Ala or Ser, which are also tiny residues and thus could be considered fairly neutral changes. Instead, Gly shows a strong preference to mutate to glutamic acid (Glu), a negatively-charged and polar amino acid, and thus a more unfavourable change. Furthermore, Gly, which due to its neutrality and small size can be tolerated in almost any protein site, is one of the most avoided substitutions for other amino acids. The same can be said about Ser, small and polar, but quite avoided even by other polar residues, in particular Thr, which differs from Ser only in the presence of an extra methyl group. Moreover, the opposite change, Ser→Thr, is equally depleted. Instead, Ser shows an extreme preference for exchanges with Phe and Leu, both hydrophobic amino acids, although the opposite is not seen. Conversely, Ile, which is not only hydrophobic and aliphatic but also C-beta-branched (its C-beta carbon is attached to two non-hydrogen substituents whereas most other amino acids only have one) does not show a significant preference for the rather similar Leu or the other C-beta branched residues (Val and Thr), but to polar residues like Ser and Lys instead.

But perhaps the most remarkable cases involve charged amino acids, which usually prefer to substitute for other similarly charged amino acids, but in cancer variants these rather neutral substitutions seem to be generally avoided as, for instance, Asp for Glu and *vice versa*. Asp instead shows a strong preference for Asn, which although it is also polar and fairly similar (its only difference is an amino group in place of an oxygen), it lacks the negative charge, meaning it could only be tolerated in contexts where the negative charge of Asp is not involved in an interaction with a cation or a positively charged protein residue (salt-

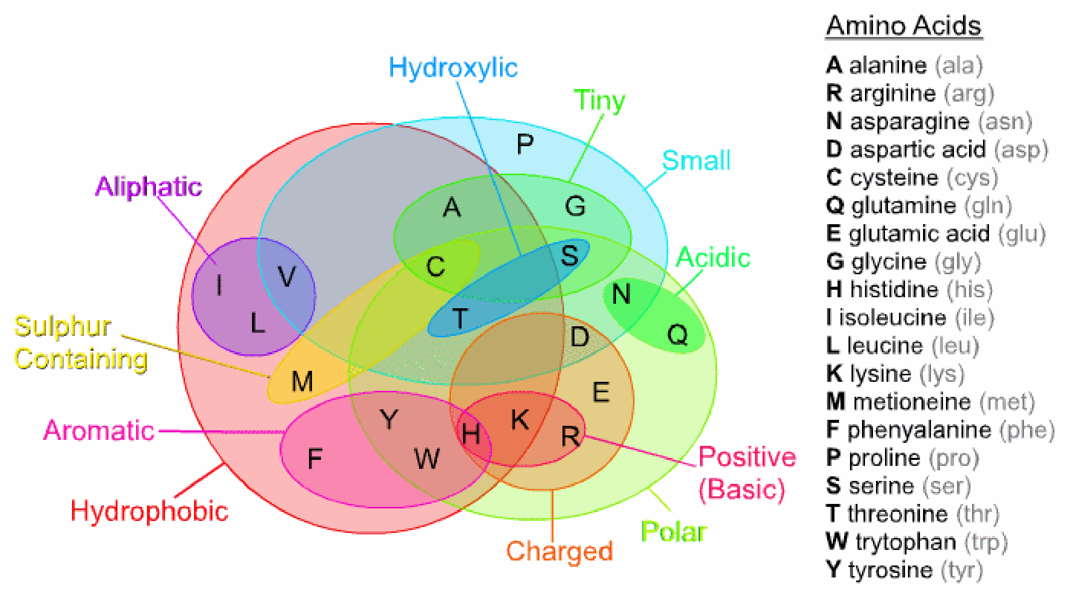
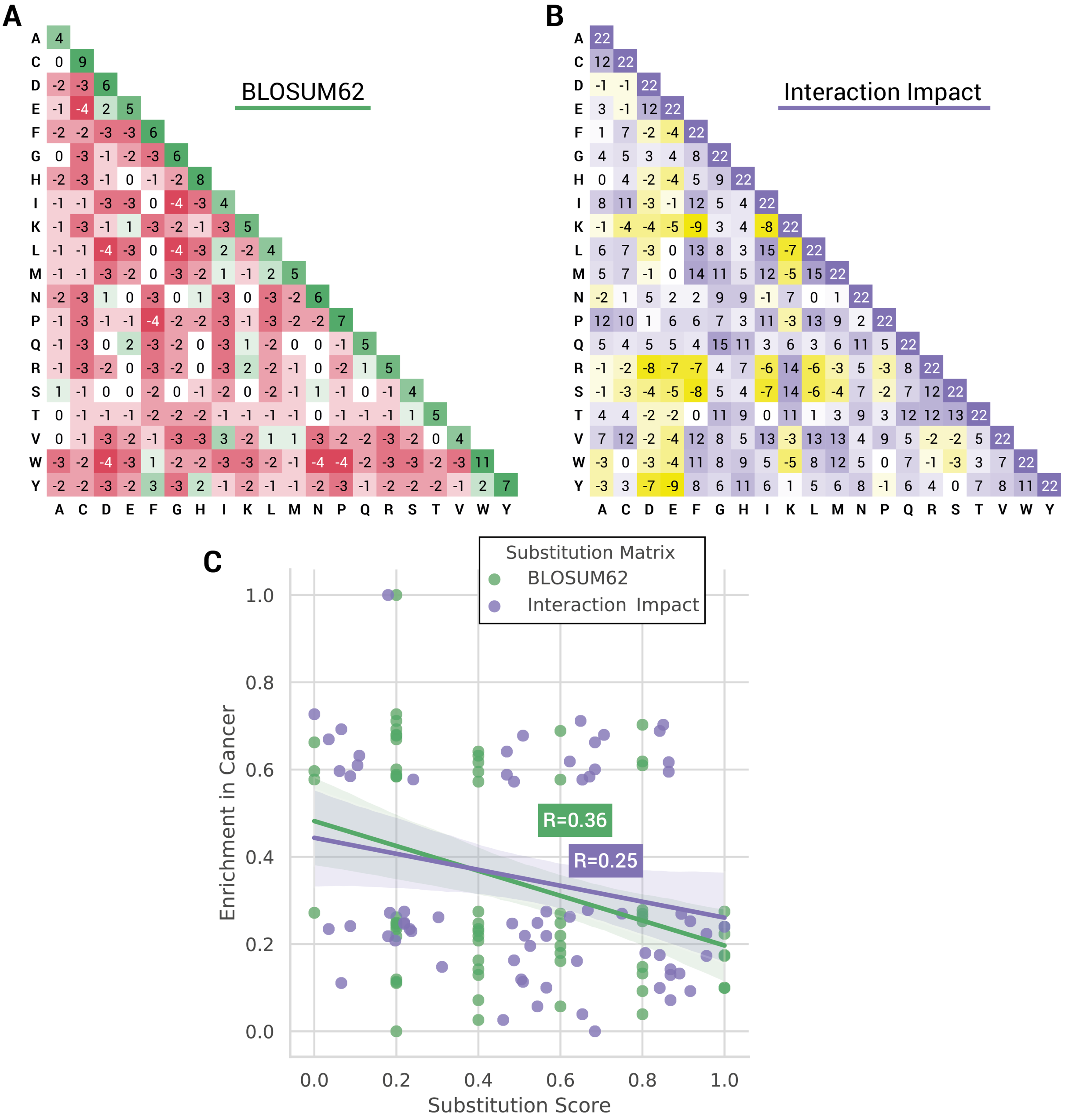


Figure 2.5: Venn diagram showing amino acid physicochemical properties classification originally proposed by William Ramsay Taylor (W. R. Taylor, 1986). Figure adapted from (Betts & Russell, 2007).

bridges). Glutamine, on the other hand, shows a strong preference for Lys, a substitution that implies a side-chain charge reversal, what can potentially be very disruptive. Lastly, for Lys, Arg, the other residue with positive charge (Lys cannot mutate to His by a single nucleotide change) is its most significantly avoided substituent.

Evolutionary conservation substitution matrix (BLOSUM62)

A more accurate way of establishing the significance of the substitution of an amino acid by another is to use a substitution matrix. Such matrices indicate the relative probabilities of all amino acid substitutions as observed in alignments of a large number of proteins sequences. Perhaps the most commonly used one is the BLOSUM62 matrix (Figure **2.6A**) (Henikoff & Henikoff, 1992)—it is by default the matrix used in protein sequence database search tools like BLAST—, which is derived from the substitution frequencies observed in ungapped local alignments of protein regions sharing an identity of 62% or less. Positive and negative scores indicate substitutions regularly seen or not in evolution, and thus likely to either conserve or have a negative effect on protein function. In line with the previous observation of physicochemical properties, a comparison of the enrichment values of amino acid substitutions in cancer with their scores from the BLOSUM62 substitution matrix shows a moderate negative correlation between both: enriched substitutions tend to have lower scores that those that are diminished (Figure **2.6C**).



	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4																			
C	0	9																		
D	-2	-3	6																	
E	-1	-4	2	5																
F	-2	-2	-3	-3	6															
G	0	-3	-1	-2	-3	6														
H	-2	-3	-1	0	-1	-2	8													
I	-1	-1	-3	-3	0	-4	-3	4												
K	-1	-3	-1	1	-3	-2	-1	-3	5											
L	-1	-1	-4	-3	0	-4	-3	2	-2	4										
M	-1	-1	-3	-2	0	-3	-2	1	-1	2	5									
N	-2	-3	1	0	-3	0	1	-3	0	-3	-2	6								
P	-1	-3	-1	-1	-4	-2	-2	-3	-1	-3	-2	-2	7							
Q	-1	-3	0	2	-3	-2	0	-3	1	-2	0	0	-1	5						
R	-1	-3	-2	0	-3	-2	0	-3	2	-2	-1	0	-2	1	5					
S	1	-1	0	0	-2	0	-1	-2	0	-2	-1	1	-1	0	-1	4				
T	0	-1	-1	-1	-2	-2	-2	-1	-1	-1	-1	0	-1	-1	-1	1	5			
V	0	-1	-3	-2	-1	-3	-3	3	-2	1	1	-3	-2	-2	-3	-2	0	4		
W	-3	-2	-4	-3	1	-2	-2	-3	-3	-2	-1	-4	-4	-2	-3	-3	-2	-3	11	
Y	-2	-2	-3	-2	3	-3	2	-1	-2	-1	-1	-2	-3	-1	-2	-2	-2	-1	2	7

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	22																			
C	12	22																		
D	-1	-1	22																	
E	3	-1	12	22																
F	1	7	-2	-4	22															
G	4	5	3	4	8	22														
H	0	4	-2	-4	5	9	22													
I	8	11	-3	-1	12	5	4	22												
K	-1	-4	-4	-5	-9	3	4	-8	22											
L	6	7	-3	0	13	8	3	15	-7	22										
M	5	7	-1	0	14	11	5	12	-5	15	22									
N	-2	1	5	2	2	9	9	-1	7	0	1	22								
P	12	10	1	6	6	7	3	11	-3	13	9	2	22							
Q	5	4	5	5	4	15	11	3	6	3	6	11	5	22						
R	-1	-2	-8	-7	-7	4	7	-6	14	-6	-3	5	-3	8	22					
S	-1	-3	-4	-5	-8	5	4	-7	14	-6	-4	7	-2	7	12	22				
T	4	4	-2	-2	0	11	9	0	11	1	3	9	3	12	12	13	22			
V	7	12	-2	-4	12	8	5	13	-3	13	13	4	9	5	-2	-2	5	22		
W	-3	0	-3	-4	11	8	9	5	-5	8	12	5	0	7	-1	-3	3	7	22	
Y	-3	3	-7	-9	8	6	11	6	1	5	6	8	-1	6	4	0	7	8	11	22

Figure 2.6: (A) BLOSUM62 matrix. **(B)** Interaction impact substitution matrix. **(C)** Lineplot showing the relationship between significant enrichment/depletion scores of amino acid substitutions in cancer variants (log-odds>1, or -1>log-odds, respectively) and their score in each of the substitution matrices. All scores have been normalized between 0 and 1.

Interaction impact substitution matrix

Not surprisingly, the different physiochemical properties of amino acid side chains mean that they have different affinities for each other, therefore residue substitutions can have potentially significant effects on protein-protein interactions. If the interaction affinity between the substitute amino acid and its contact is lower or higher than with the original amino acid, the interaction can be reduced or enhanced, respectively. Interactions affinities for each pair of residues were previously calculated as pair-potentials derived from protein interfaces in a non-redundant set of structures, and indicate whether contacts between the residues are more frequent than expected by chance given the abundance of the amino acids at those interfaces. In other words, a positive value indicates a tendency to interact and a negative one a tendency to avoid each other (Aloy & Russell, 2002). I used an updated version of these pair-potentials, first used in the Mechismo system (Betts et al., 2015) and now calculated from a recent PDB release with a much larger number of protein structures, to construct a substitution matrix that scores the potential impact of amino acid substitutions on protein interactions (Figure **2.6B**). The scores are based on the overall difference of interaction affinities between the original residue and its substitute (see Methods **2.3.4**). Here, high positive scores indicate similar interaction affinities between the two residues and thus a more favourable substitution, while negative scores indicate the opposite. However, these scores show an even weaker negative correlation with cancer variant preferences than BLOSUM62 (Figure **2.6C**), therefore, it is not possible to say with certainty that cancer variants are generally enriched in amino acid substitutions more likely to impact on protein function or interactions.

It is important to note, however, that the scores from the two matrices are not always correlated and their interpretation must take into account the functional context. For example, according to the interaction impact matrix, Pro is not a bad substitution for most other residues, but it is well-known that its conformational rigidity means that it cannot substitute well for many amino acids, and this is reflected in the generally negative scores of the BLOSUM62 matrix. Conversely, the previously highlighted charge-reversing Glu→Lys substitution is positively scored (+1) by BLOSUM62, which means that it is not an uncommon substitution in evolution, and therefore, not particularly deleterious. The biological reason for this could be that both residues are polar and thus prone to be in contact with water when they are on the surface (~70% of residues in the human proteome). In contrast, this substitution is unfavourable in terms of affinity shift, most likely due to the fact that it results in a charge reversal of the side chain, which if occurring within a salt-bridge or a hydrogen-bonded Glu-Lys or Glu-Arg interaction pair, would certainly have a disruptive effect.

2.2.3 Site-specific analysis of amino acid exchange selection in cancer variants

2.2.3.1 Site-specific exchange preferences may hint at underlying specific mechanisms

General preferences in amino acid substitution can only give an idea of how a cancer is driven, but the effects of mutations depend entirely on the functional context: the same amino acid substitution may be harmless in one context but catastrophic in another. Residues can have opposite physicochemical preferences depending on whether they are buried or on the protein surface (i.e. in different microenvironments), which in turn depends on the cellular location of the protein (e.g. cytoplasm, membranes or extracellular space). In addition, residues may be located in functional regions, such as short active sites or broader interaction interfaces (Betts & Russell, 2007).

Recurrence of mutations across independent cancer tumours is usually considered the first sign of positive selection. This is usually sufficient to establish that the affected protein position is functionally important and that changes at this site are likely to affect protein function. However, depending on the functional context of this residue, the nature of the mutant amino acid may be of critical importance. In tumour suppressors, highly mutated positions usually show no preference for a particular variant, meaning that the substitute amino acid makes no difference and all have the same, usually disruptive, effect (Vogelstein et al., 2013). On the contrary, if a particular variant is significantly enriched over the rest of the possible but rarely seen variants at the same position, this could indicate that the resulting amino acid is selected for a very specific functional reason (typically in gain-of-function mutations). Particularly in the case of mutations that affect protein interfaces, there is the possibility that the same variant can, in principle, affect different interactions differently (Sahni et al., 2015). I hypothesise that this mechanistic subtlety means that such variants are highly specific and that this specificity must be reflected in their mutational profiles if they are subject to positive selection.

Consequently, I applied an approach similar to the one used to determine general amino acid preferences, but now at the level of individual positions that are recurrently affected by mutations, to identify protein positions that show a statistically significant enrichment for a particular amino acid substitution (henceforth called selected variants) (see Methods **2.3.6** for details).

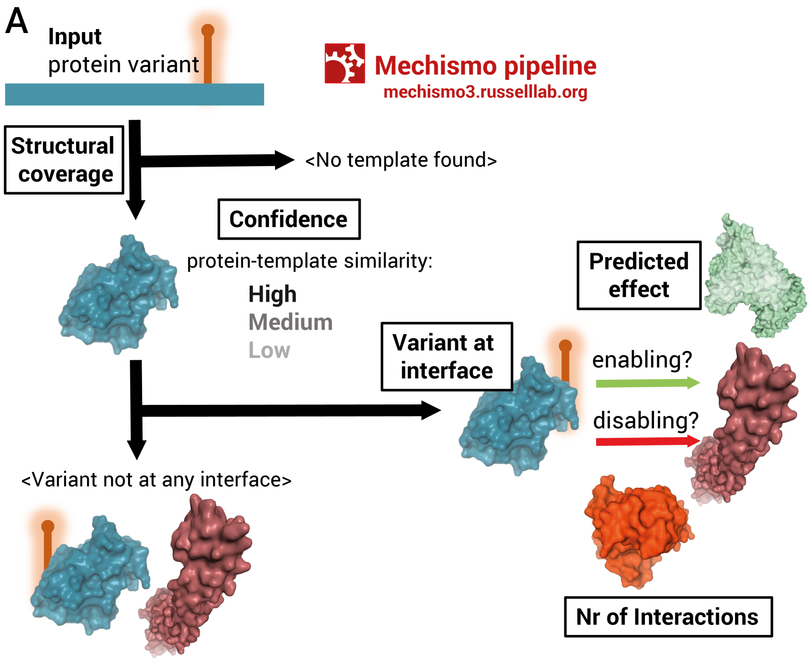
2.2.3.2 Selected variants are more likely to have effects on protein interactions

To test the hypothesis that these variants are more likely to affect protein interactions, I used Mechismo, a tool for mapping variants to 3D structures and predicting their impact on protein interactions (Betts et al., 2015) (Figure **2.7A;** Methods **2.3.7**). I made predictions for all the missense mutations, both natural and cancer ones, then calculated various parameters for different subsets of variants. The subsets consist of: (i) all natural variants; (ii) all cancer variants; (iii) all cancer variants in the genes of the COSMIC cancer gene census (CGC); (iv) all cancer variants that are recurrent (present in at least 5 samples); and (v) the set of selected variants.

The comparison of these parameters between the sets (Figure **2.7B**) reveals that, as established before, cancer variants in general are more prone to have an impact on protein interactions than natural variants. Cancer-associated proteins, and in particular those in the CGC, have also a wider structural coverage, owing to their greater biomedical interest and larger number of targeted studies towards these proteins. Selected variants have a lower structural coverage in comparison exactly because of the opposite: many of the proteins they affect have not been causally implicated in cancer and not deeply studied yet. Within cancer, recurrent variants already have an increased preference for protein interfaces, but selected variants show an even higher enrichment at interfaces (also when considering only high-confidence predictions), and are also predicted to affect a larger number of interactions in those cases (Figure **2.7B**).

2.2.3.3 Selected variants suggest novel driver genes and new mechanisms for established ones

In total, this approach identified 5,209 unique selected variants (45,770 by total sample count) located in 5,183 positions of 3,463 proteins. This means that proteins with this type of mutations have on average less than two highly selected mutations (1.49 to be exact), which in principle is in line with the typical oncogene mutational pattern where gain-of-function mutations stand in single hotspots. Moreover, only ~21% of sites recurrently affected by cancer mutations (mutations observed in more than 5 samples; 24,432 sites in total) exhibit a significant preference for a particular amino acid substitution. The protein with the largest number of selected variants is, unsurprisingly, TP53, which is not only the most frequently mutated protein in cancer but also the most studied one, and it is considered a TSG that can act as an oncogene in certain contexts (Rivlin et al., 2011). It is followed by MUC16, proposed as an oncogene (Aithal et al., 2018); HLA-A, involved in oncogenic fusion



B

	All natural variants	All cancer variants	CGC gene variants	Recurrent variants	Selected variants
Unique variants	33,854	247,964	164,404	25,071	5,209
Structural coverage (proteins)	50.9%	75.6%	90.7%	62.8%	54.8%
Structural coverage (variants)	40.0%	51.9%	53.3%	56.6%	52.7%
Variants at interface	3.9%	7.2%	10.1%	10.4%	15.6%
Variants at interface (HC)	2.5%	4.9%	7.7%	7.6%	10.3%
Avg. number of affected interactions	4.6	5.5	6.0	6.5	7.4

Figure 2.7: **(A)** Workflow of the Mechismo prediction tool. Variants are first mapped onto PDB structures, with different confidence levels depending on sequence similarity between protein and template. If possible, variants are then checked for location at interfaces with other proteins. Variants can be at interfaces affecting interactions with a number of proteins, with either an enabling or disabling effect. **(B)** Comparison of different parameters calculated from all Mechismo predictions for different subsets of variants.

with ROS1 (Uguen & De Braekeleer, 2016); and CDC27, for which roles as either TSG or oncogene has been suggested in different neoplasms (Kazemi-Sefat et al., 2021). However, the majority of selected variants (92%) are in proteins that have not been (yet) causally implicated in cancer, or at least not yet included in COSMIC's Cancer Gene Census (CGC) (Figure **2.8A**), and thus they are potentially interesting targets for future investigation. For instance, p.Val384Asp in the DNA mismatch repair protein Mlh1 (MLH1) and p.Met293Lys in the trans-acting T-cell-specific transcription factor GATA-3 (GATA3) are the single-most enriched variant in each protein (Figure **2.8A,B**). From the variants found in known cancer driver genes, more than half are located at oncogenes or proteins with dual TSG/oncogene dual functionality, but a significant proportion affect TSGs as well (Figure **2.8C**).

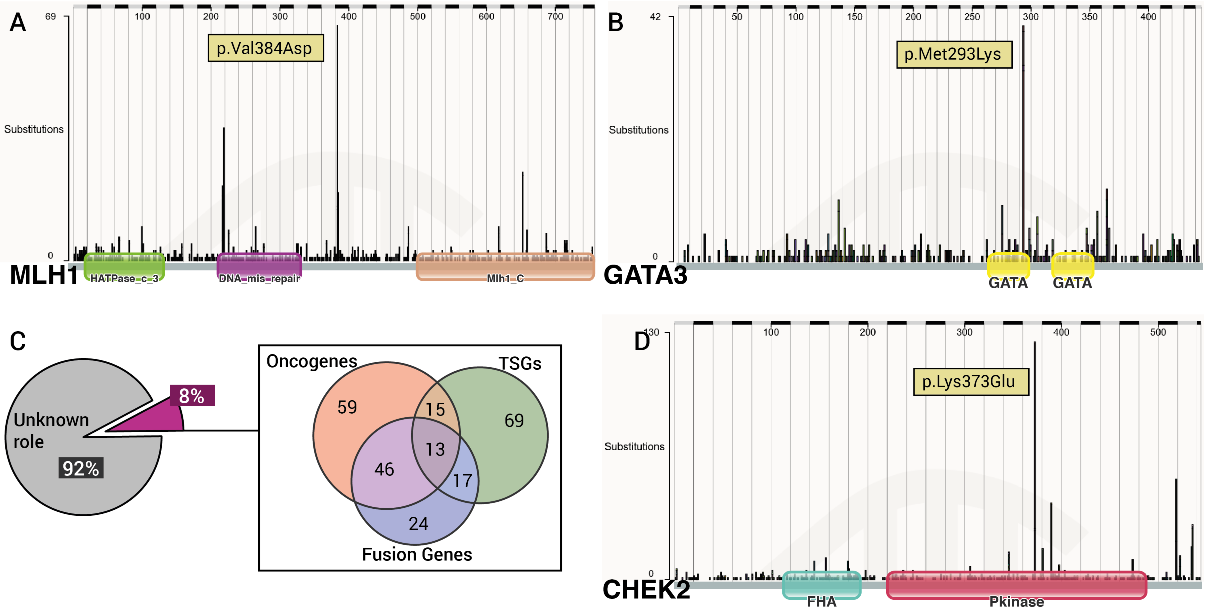


Figure 2.8: (**A,B,D**) Mutation profiles of MLH1, GATA3 and CHEK2, where a highly enriched variants represent the largest peaks. **(C)** Role in cancer of genes affected by selected variants.

The presence of selected variants in TSGs contrasts with their typical characterization of being altered by truncating mutations throughout their sequences, and could suggest two things. The first is that loss of function could be achieved through a subtler mechanism that does not result in complete protein abrogation. Often, tumour suppressors are multi-functional which means that cells presumably need other functions to still work. For example, variants in TP53 are clinically different. The most common ones affect the DNA binding function specifically; only rarer variants destroy the zing binding site and the protein, but these evolve later in cancers (Raimondi et al., 2017). The second is that these genes could perhaps act as oncogenes in particular contexts. For instance, I identified the highly enriched p.Lys373Glu mutation in the serine/threonine-protein kinase Chk2 (CHEK2) (Figure **2.8D**). CHEK2 has a well-established role as a tumour suppressor, regulating cell cycle arrest, DNA repair and apoptosis upon DNA damage through the phosphorylation of numerous substrates (Bartek & Lukas, 2003; Cai et al., 2009). Generally, CHEK2 variants are considered to be inactivating but the presence of the highly recurrent Lys373Glu results in a mutational spectrum typical of oncogenes. Although it has been proposed that Lys373Glu also impairs CHEK2 function (Higashiguchi et al., 2016), there are several interactions potentially affected by this change according to Mechismo, but the most pronounced observation is that a similar variant in an equivalent kinase position is known to lead to constitutive activation, which at least warrants further investigation.

2.2.4 Case study: structural subunit A of the serine/threonine-protein phosphatase 2A

The *PPP2R1A* gene encodes the 65kDa structural subunit A (UniProtKB accession: P30153) of the serine/threonine-protein phosphatase 2A (PP2A), an enzyme that has a major role in the negative regulation of cell growth and division, being implicated in the regulation of cell cycle initiation and most of its checkpoints, by dephosphorylating more than 300 substrates, and thus, a known tumour suppressor (Eichhorn et al., 2009; Wlodarchak & Xing, 2016). This broad range of activity is due to its particular structure. PP2A is a heterotrimeric enzyme formed by a dimeric core composed of subunit A and the 36KDa catalytic subunit C, and a regulatory subunit B. Subunit A is composed of 15 HEAT repeats arranged in a horseshoe-like structure (or alpha solenoid) and acts as the scaffold that coordinates the assembly of the complete complex (Groves et al., 1999; Xu et al., 2006) (Figure **2.9A**). It comes in two flavours: the alpha and beta isoforms (PP2A-Aα and PP2A-A β), encoded by genes *PPP2R1A* and *PPP2R1B* respectively, and although they share 87% sequence similarity, they have different binding affinities to the other PP2A subunits (Hemmings et al., 1990). However, PP2A-Aα, the one discussed here, is the one contained by the large majority of PP2A holoenzymes in adult tissues, as the β isoform is underexpressed in comparison (Zhou et al., 2003).

The dimeric core can associate with a wide variety of mutually exclusive regulatory B subunits, classified into four major families that share no sequence similarity, resulting in diverse PP2A holoenzymes with different specificities (Figure **2.9A**). The regulatory subunit B is thus the one that mediates the substrate specificity of the whole holoenzyme. Although the different subunits B are structurally different, subunit A is able to suffer big conformational changes to adapt to all of them, while the interaction between B and C subunits remains very limited (Xu et al., 2008).

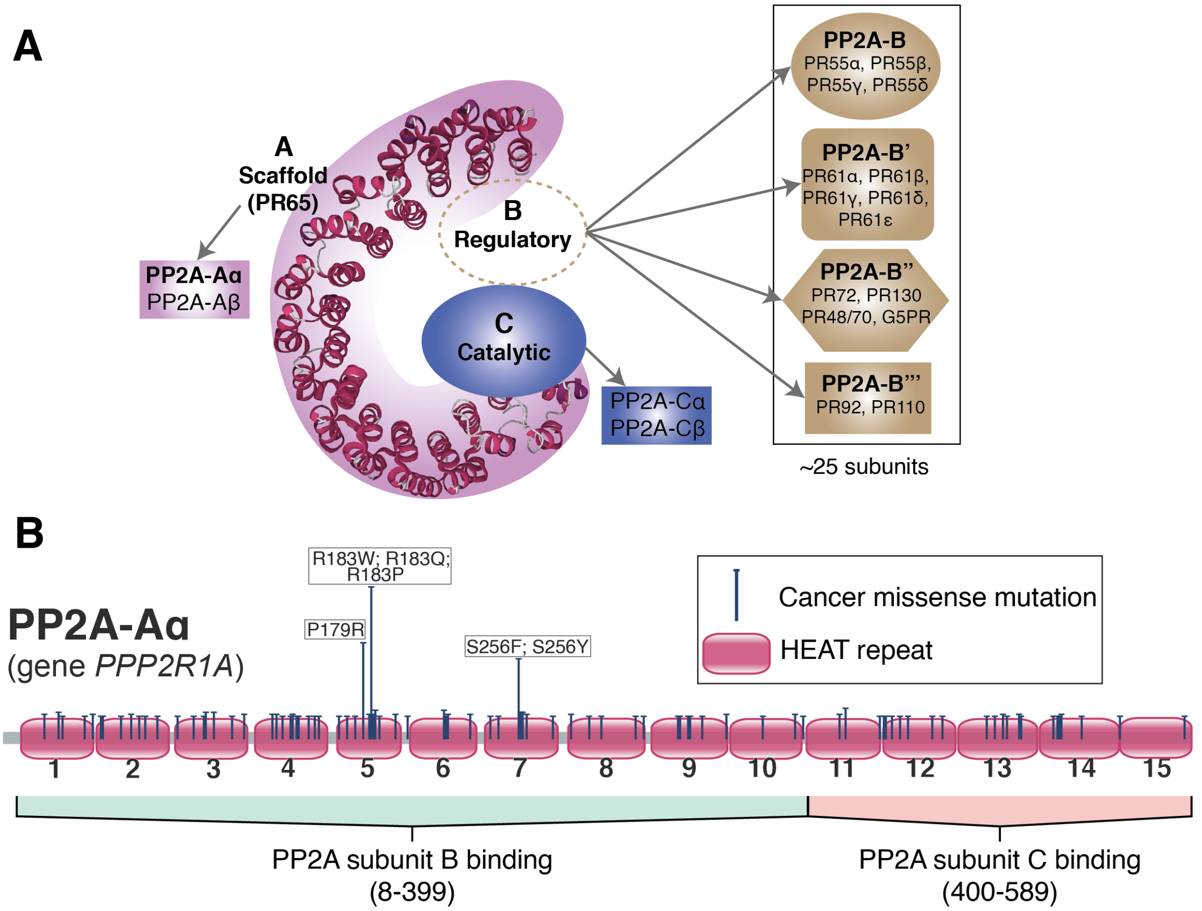


Figure 2.9: **(A)** The serine/threonine-protein phosphatase 2A (PP2A) holoenzyme is a heterotrimeric complex formed by a scaffolding subunit A (two isoforms), a catalytic subunit C (two isoforms) and a regulatory subunit B, for which multiple options exist and are classified into four different families. **(B)** Linear representation of PP2A-Aα, which is an α-solenoid composed of 15 HEAT repeats. Binding regions of the other two PP2A subunits are indicated below. All cancer missense mutations in this protein are shown with flags proportional to the number of tumour samples they have been observed. Amino acid changes are specified at the three major hotspots.

In general, mutations in PP2A-Aα are almost exclusive of endometrial and ovarian carcinomas. Missense mutations in this protein display a pattern that is more typical of gain-of-function mutations and oncogenes than tumour suppressors: they cluster in two very clear hotspots in the subunit B-binding portion of the protein, with residues Pro179, Arg183 (both located at HEAT repeat 5) and Ser256 (at HEAT repeat 7) as the most significantly affected (Figure **2.9B**). In particular, variants Pro179Arg, Ser256Phe and Ser256Tyr were found to be highly selected, and moreover, to show different predicted effects on interactions with different regulatory subunits B of the holoenzyme (Figure **2.10**). Variants in the two sites have enabling effects on interactions with the several isoforms of regulatory subunit B' family (PR61α–ε), while Pro179Arg in addition has a disabling effect on two regulatory subunits B'' (PR72/PR130 and PR48/PR70). Additionally, these predictions were compared to those made for other possible variants at the same residues that are either much less frequent or totally absent in cancer, and revealed that the particular affinity changes were specific to these observed variants (Figure **2.10**). The exact B' enabling and B'' disabling effects of Pro179Arg are not predicted for any of the other possible (and mostly absent) protein changes. In the case, of Ser256, both selected variants have the same enabling effect on B' subunits, which is not predicted for other absent variants with the exception of Ser256Pro, although with weaker enabling scores (one could argue that a Pro might nevertheless not sit too well in the structure).

A similar deduction can be made for position Arg183, where although variants at this position do not show a statistically significant deviation from expected frequencies according to my analysis, the two major variants seen here—Arg183Trp and Arg183Glu—also show exclusive predicted effects on the interactions with the different subunits B. In line with selected Pro179Arg, they are predicted to favour interactions with subunits of the B' family while disfavouring those with the ones from B'' family (including now G5PR among them), with the addition of a disabling effect also on the interactions with B family regulatory subunits (PR55α-δ) (Figure **2.10**).

Mutations in P179 have already been linked to these cancers (Nagendra et al., 2012; Shih & Wang, 2011), and in fact, their distinct effect on interactions with regulatory subunits has been reported (Houge et al., 2015). Pro179Arg has been also suggested to impair holoenzyme formation and its enzymatic activity specifically by increasing the rigidity in the alpha solenoid due to newly form internal interactions, ultimately hindering the binding of catalytic subunit C (Taylor et al., 2019); but they did not discard however more contextual effects, in line with the notion from Houge et al. Given the central role of regulatory subunits in determining the holoenzyme substrate specificity, Mechismo predictions suggest an affinity-shifting mechanism that results in a more nuanced dysregulation of PP2A activity rather than complete inhibition. This is further supported by the fact that both isoforms of the catalytic subunit of PP2A (*PPP2AC*), perhaps a more effective target for obliterating the enzyme, show a very low point mutation burden in cancer. Instead, both upregulation and downregulation of PP2Ac expression have been linked to different types of cancers, at least suggesting multiple mechanisms of action (Gong et al., 2016; Yang et al., 2021; Yong et al., 2018). Strikingly, the predicted effect of favouring interactions with B' subunits might hint at an oncogenic mechanism, as it has been shown that oncogene MDM2 can be dephosphorylated by PP2A holoenzymes specifically containing B' subunits, when recruited

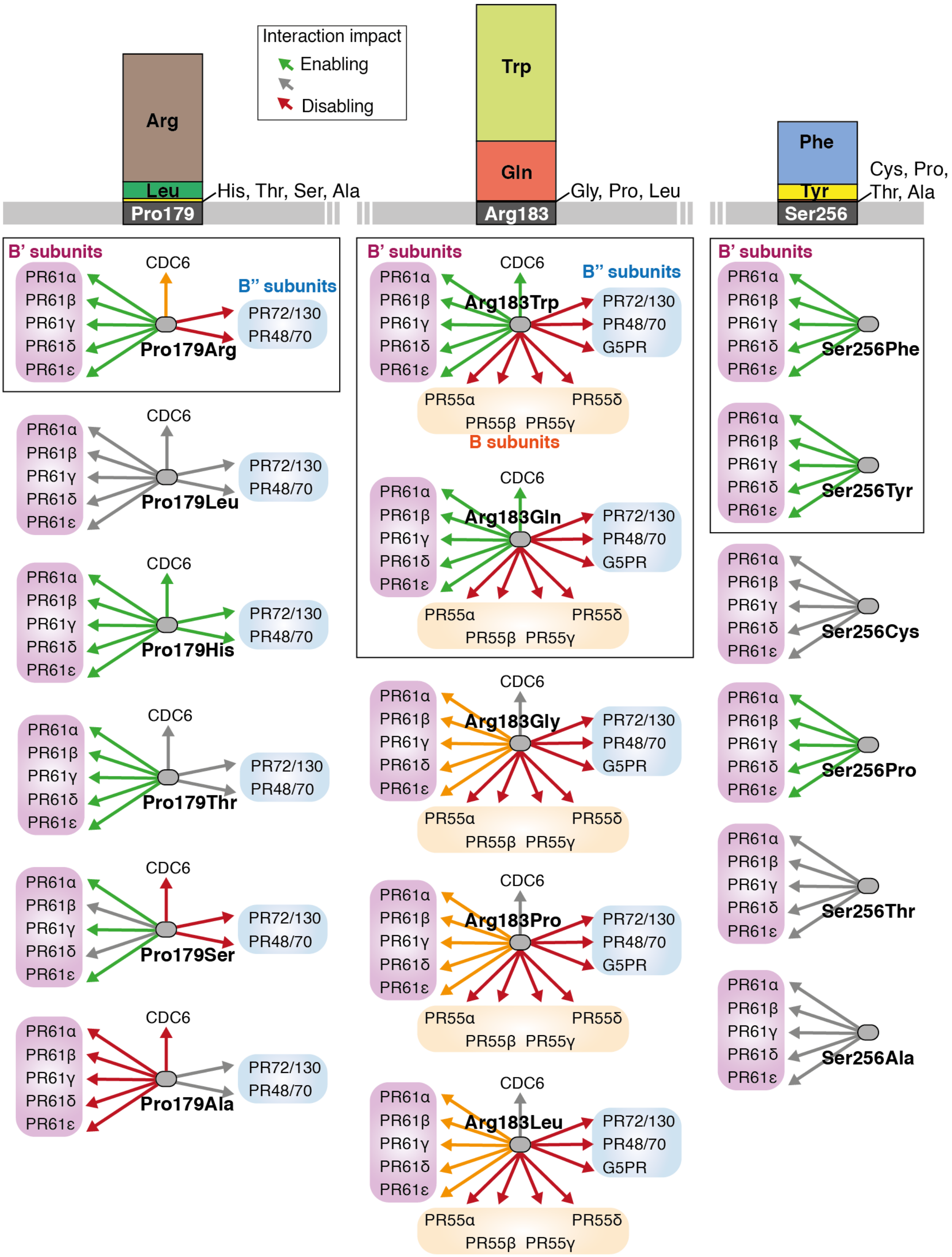


Figure 2.10: Barplots on top show the proportion of observed variants at each position, and list those that are possible by SNVs but are rarely observed or not at all. Diagrams below show predicted effects for each variant on the interactions with different types of B subunits, which can be: enhancing (green), disabling (red), none (grey; when there is a contact but the effect is too weak), and mixed (yellow; when the same amino acid is favouring and disfavouring the interaction with two residues at the same interface).

by cyclin G (Okamoto et al., 2002). Dephosphorylation of MDM2 leads to its activation, which can then mediate TP53 ubiquitination and promote its degradation.

2.3 Materials and methods

2.3.1 Collecting cancer somatic and natural variants

Non-synonymous, ‘confirmed somatic’ protein variants were extracted from the COSMIC v94 (Tate et al., 2019) whole genome screen-only dataset, present in at least 5 independent samples. This dataset thus only contains variants identified in studies that surveyed all genes and should not contain any bias. ‘Confirmed somatic’ indicates that the variant allele from the tumour was confirmed to be different from the germline allele of the same individual. The minimum sample cut-off was used to remove the most likely passenger variants which are neutral to cancer development. A total of 3,261,120 somatic variants, affecting 2,186,340 unique protein positions within 18,658 human proteins, from 103 different cancer types were collected. From the 1000 Genomes Project (1kG) (1000 Genomes Project Consortium et al., 2015) dataset of common genetic variation, I extracted a total of 38,781 missense germline variants with a minor allele frequency >5% (to avoid rare, potentially disease-related variants), located in 38,330 unique residues of 12,169 human proteins.

All variants were mapped to the same canonical UniProtKB (The UniProt Consortium, 2019) protein sequences of the human reference proteome (20,328 proteins, May 2018) *via* alignments between Ensembl transcripts and the UniprotKB sequences. Datasets were crosschecked to additionally exclude likely natural variants from cancer mutations. A summary of the data collected and its processing can be found in Figure **2.1**.

2.3.2 Determining solvent accessibility, domains and interfaces.

Residues affected by cancer or 1kG variants were counted according to the following classifications (total fractions are shown in Figure **2.2**.).

The accessible surface area (ASA) of every protein residue was calculated with the DSSP program (Kabsch & Sander, 1983) from 3D structures predicted by Alphafold (Varadi et al., 2022). The relative solvent accessibility (RSA) of a protein residue is calculated by normalizing the solvent accessible surface area observed in the crystal structure by the maximum possible ASA for that residue, for which the recommended reference values provided by Tien et al. (Tien et al., 2013) were used. RSA is a metric to describe residues as either buried or exposed. I adopted the generally accepted convention that residues can be classified as buried if their RSA is <20%, and exposed otherwise. According to this, 71% of residues in the human proteome are exposed.

Protein residues were classified as either functional or non-functional depending on whether they are inside or outside globular domains, which were extracted from Pfam (El-Gebali et al., 2019). For the complete human proteome, 48% of residues are defined as functional.

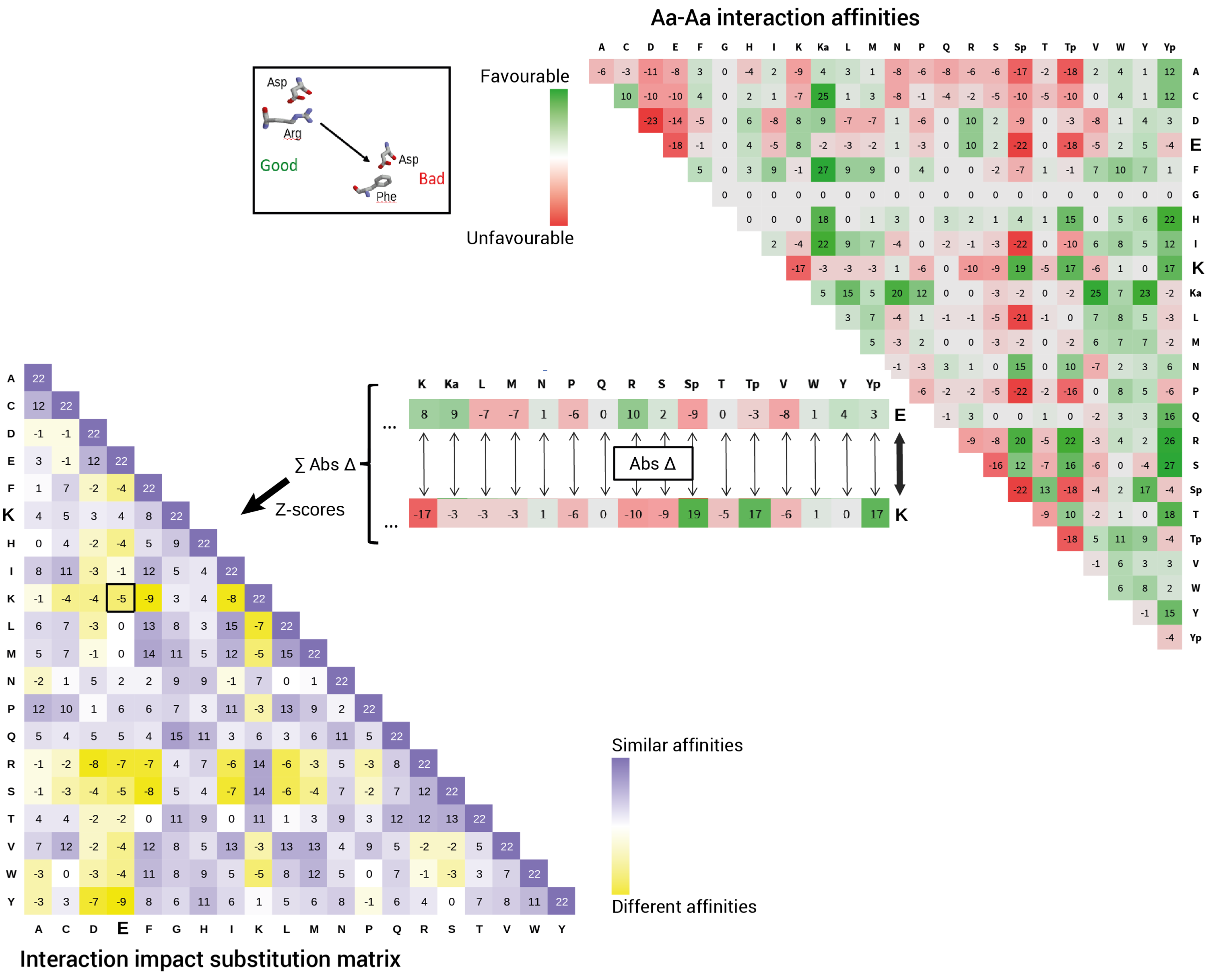
Residues location within interfaces was determined using the predictions from Mechismo (Betts et al., 2015), as part of the step described in section **2.3.7**.

2.3.3 Collecting post-translational modification data

To study PTM-related variants, I retrieved 39,826 phosphorylation and 4,143 acetylation sites in 19,013 human proteins by non-redundantly combining annotations from UniProt and the PhosphositePlus database (Hornbeck et al., 2015). For the subsequent analysis, the amino acids presenting these modifications were considered separately from their non-modified equivalents, but equally analysed, as four additional residues: acetyllysine (Ka), phosphoserine (Sp), phosphothreonine (Tp) and phosphotyrosine (Yp).

2.3.4 Calculating amino acid mutabilities and exchange frequencies

Amino acid exchanges from the cancer and the 1kG variant datasets were independently counted and classified. The mutability of each amino acid was calculated as the total number of mutations for that amino acid in the dataset divided by its frequency of occurrence in the human UniProtKB reference proteome (Figure **2.3**). The normalized frequencies of occurrence of each amino acid exchange were calculated by dividing the count by the total number of mutations observed for that amino acid. For each amino acid, the frequencies in its two mutation profiles (as determined from cancer or 1kG variants) were compared through a chi-square (χ^2) test to determine if there was a significant difference (Figure **2.4A**). A similar test was done using all exchange frequencies to determine, in this case, if there was an overall significant difference between both datasets. The enrichment or depletion of every amino acid substitution was computed as the log of the odds ratio between the observed frequency in cancer and the expected background frequency (as observed in the 1KG variant dataset), then plotted into a heatmap (Figure **2.4B**). Maximum and minimum values of 5 and -5 were set for those cases where the score cannot be calculated due to missing observed or expected frequency of a particular amino acid subsitution. This happens with some modified residues, for example, acetyllysine (Ka).



Aa-Aa interaction affinities

A	C	D	E	F	G	H	I	K	Ka	L	M	N	P	Q	R	S	Sp	T	Tp	V	W	Y	Yp	
-6	-3	-11	-8	3	0	-4	2	-9	4	3	1	-8	-6	-8	-6	-6	-17	-2	-18	2	4	1	12	A
	10	-10	-10	4	0	2	1	-7	25	1	3	-8	-1	-4	-2	-5	-10	-5	-10	0	4	1	12	C
		-23	-14	-5	0	6	-8	8	9	-7	-7	1	-6	0	10	2	-9	0	-3	-8	1	4	3	D
			-18	-1	0	4	-5	8	-2	-3	-2	1	-3	0	10	2	-22	0	-18	-5	2	5	-4	E
				5	0	3	9	-1	27	9	9	0	4	0	0	-2	-7	1	-1	7	10	7	1	F
					0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	G
						0	0	0	18	0	1	3	0	3	2	1	4	1	15	0	5	6	22	H
							2	-4	22	9	7	-4	0	-2	-1	-3	-22	0	-10	6	8	5	12	I
								-17	-3	-3	-3	1	-6	0	-10	-9	19	-5	17	-6	1	0	17	K
									5	15	5	20	12	0	0	-3	-2	0	-2	25	7	23	-2	Ka
										3	7	-4	1	-1	-1	-5	-21	-1	0	7	8	5	-3	L
											5	-3	2	0	0	-3	-2	0	-2	6	7	7	-2	M
												-1	-3	3	1	0	15	0	10	-7	2	3	6	N
													-6	-2	-2	-5	-22	-2	-16	0	8	5	-6	P
														-1	3	0	0	1	0	-2	3	3	16	Q
															-9	-8	20	-5	22	-3	4	2	26	R
																-16	12	-7	16	-6	0	-4	27	S
																	-22	13	-18	-4	2	17	-4	Sp
																		-9	10	-2	1	0	18	T
																			-18	5	11	9	-4	Tp
																				-1	6	3	3	V
																					6	8	2	W
																						-1	15	Y
																							-4	Yp

K	Ka	L	M	N	P	Q	R	S	Sp	T	Tp	V	W	Y	Yp	
8	9	-7	-7	1	-6	0	10	2	-9	0	-3	-8	1	4	3	E
-17	-3	-3	-3	1	-6	0	-10	-9	19	-5	17	-6	1	0	17	K

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	22																			
C	12	22																		
D	-1	-1	22																	
E	3	-1	12	22																
F	1	7	-2	-4	22															
K	4	5	3	4	8	22														
H	0	4	-2	-4	5	9	22													
I	8	11	-3	-1	12	5	4	22												
K	-1	-4	-4	-5	-9	3	4	-8	22											
L	6	7	-3	0	13	8	3	15	-7	22										
M	5	7	-1	0	14	11	5	12	-5	15	22									
N	-2	1	5	2	2	9	9	-1	7	0	1	22								
P	12	10	1	6	6	7	3	11	-3	13	9	2	22							
Q	5	4	5	5	4	15	11	3	6	3	6	11	5	22						
R	-1	-2	-8	-7	-7	4	7	-6	14	-6	-3	5	-3	8	22					
S	-1	-3	-4	-5	-8	5	4	-7	14	-6	-4	7	-2	7	12	22				
T	4	4	-2	-2	0	11	9	0	11	1	3	9	3	12	12	13	22			
V	7	12	-2	-4	12	8	5	13	-3	13	13	4	9	5	-2	-2	5	22		
W	-3	0	-3	-4	11	8	9	5	-5	8	12	5	0	7	-1	-3	3	7	22	
Y	-3	3	-7	-9	8	6	11	6	1	5	6	8	-1	6	4	0	7	8	11	22

Interaction impact substitution matrix

Figure 2.11: Calculation of the interaction impact substitution matrix (purple-yellow matrix) from the interaction affinities of amino acid (green-red matrix). For clarity, values from both matrices are multiplied by 10 and stripped of decimals. Affinities are log-odds that measure how often a pair of residues is seen in contact at an interface compared to the expected frequency given the abundance of those residues generally at interfaces. The interaction impact scores measure how similar are the affinities of two residues with all others.

2.3.5 Calculating the interaction impact substitution matrix

As scores of interaction affinities between amino acids, I used pair potentials first developed by (Aloy & Russell, 2002) using a non-redundant set of structures, later updated and extended to consider phosphorylated and acetylated residues to be used by Mechismo (Betts et al., 2015). I used an even more updated version of these pair-potentials—calculated from the

PDB 2018 version, which includes a much higher number of protein interfaces structures— to construct a substitution matrix that scores the overall impact of an amino acid substitution if happening at a protein-protein interface. The scores are calculated as the sum of the absolute differences in interaction affinities between every two amino acids and all other amino acids, then normalized by calculating the standard deviation from the mean (Z-score). Finally, the scores were inverted so that large positive numbers indicate a low impact change (similar interaction affinities), and large negative ones, changes that have a high impact (amino acids with different interaction affinities) (Figure **2.11**).

2.3.6 Defining selected sites and variants

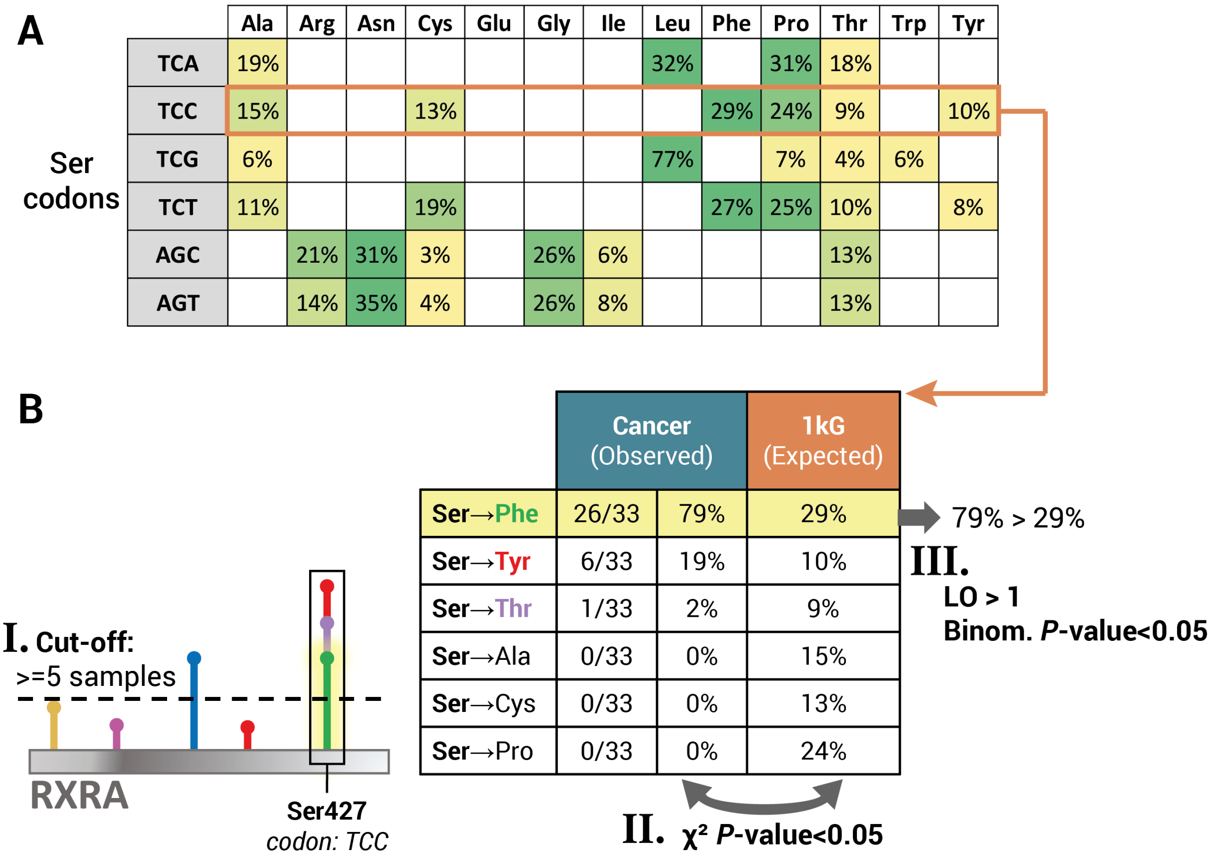
To identify positions and variants likely under positive selection in cancer, I first calculated expected background substitution rates for each amino acid-coding codon based on the 1kG variant dataset, resulting in 61 codon-amino acid mutation profiles. I took this approach to account for the fact that substitution rates for amino acids encoded by more than one codon generally differ between these codons, and there are substitutions that are possible (through SNVs) only for certain codons (Figure **2.12A**). Then, I collected cancer variants that fulfilled the following conditions: (i) the variant is present in at least 5 samples; (ii) there is a statistically significant difference between the observed and expected mutation frequency for the residue/codon (χ^2 *P*-value < 0.05); (iii) the variant is significantly enriched relative to its expected frequency (log-odds > 1 and binomial test *P*-value < 0.05) (Figure **2.12B**). *P*-values were adjusted with the Bonferroni correction.

2.3.7 Assessing functional impact of variants on protein interactions

To predict the effects of protein variants on protein interactions, I used Mechismo (Betts et al., 2015). Mechismo[4] is an online tool that attempts to map protein variants or modifications to available 3D structures of interacting proteins (using sequence homology to extend its coverage) and then assess their impact on interactions if located at interfaces. This impact is given in the form of a positive or negative score indicating whether the variant/modification enhances or hinders the interaction. The coverage of Mechismo is limited by the availability of known or homologous structures in the Protein Databank (Burley et al., 2019). The degree of sequence similarity between protein and structural template defines three confidence levels (low, medium or high).

[4] http://mechismo3.russelllab.org/

I ran Mechismo with all variants included in this project (both cancer and natural variants). I then computed various parameters for performance comparisons between different subsets of variants, such as the proportion of variants that mapped to a structure, the proportion of variants located at a protein interface, and the number of interactions affected by each variant; all these, under different confidence levels.



	Ala	Arg	Asn	Cys	Glu	Gly	Ile	Leu	Phe	Pro	Thr	Trp	Tyr
TCA	19%							32%		31%	18%		
TCC	15%			13%					29%	24%	9%		10%
TCG	6%							77%		7%	4%	6%	
TCT	11%			19%					27%	25%	10%		8%
AGC		21%	31%	3%		26%	6%				13%		
AGT		14%	35%	4%		26%	8%				13%		

	Cancer (Observed)		1kG (Expected)
Ser→Phe	26/33	79%	29%
Ser→Tyr	6/33	19%	10%
Ser→Thr	1/33	2%	9%
Ser→Ala	0/33	0%	15%
Ser→Cys	0/33	0%	13%
Ser→Pro	0/33	0%	24%

Figure 2.12: (A) Codon-specific amino acid substitution frequencies for serine calculated from the 1kG dataset of natural variants. **(B)** The approach for determining selected variants is illustrated using mutations seen at Ser427 of the retinoic acid receptor RXRA. The observed amino acid exchange frequencies are compared to the expected codon-specific frequencies to determine enrichment and asses the significance.

2.4 Conclusions

Conventional methods for identifying cancer drivers focus mainly on the frequency of mutations, as sites recurrently affected by point mutations are usually enough sign of positive selection and to warrant interest. However, considering the nature of the observed amino acid substitutions can be illuminating as it allows us to make hypothesis using prior mechanistic knowledge.

Here, I first showed that cancer has preferences for particular amino acid substitutions, as well as avoids other ones, generally favouring changes that are a priori drastic. Of course, the effects of amino acid substitutions are highly dependent of the functional context of the protein in general and the protein site in particular. Consequently, I then sought to explore whether the presence of significantly enriched amino acid substitutions in cancer can be linked to oncogenic mechanisms, in particular, to subtler phenotypes that imply changes in interactions affinities with one or more different partners. I identified a large number of genes and positively selected mutations within those that may contain several potential novel driver genes as well as suggest new mechanisms for established ones. I reviewed a few cases and highlighted predicted effects on protein interactions that point to interesting (subtler) mechanisms.

Characterising protein variants and the interactions they influence is key to interpreting functional effects at the protein level. In the context of cancer genomics, given the increasing amount of genetic information and the wide diversity of cancers, this type of molecular interpretation can help to better understand the relationship between genotype and phenotype, promote the discovery of potential drug targets, and help guide medical diagnosis and treatment.

Chapter 3

Interactive Visualization of Protein Mechanism with Mechnetor

3.1 Introduction

Efforts to catalogue functional information have resulted in a large number of resources that show or predict insights into protein function and mechanism at different levels, many of which were already mentioned in section **1.5**. These include deposited protein sequences, together with their subcellular locations and functional descriptions based on literature review, protein families and domains identified through sequence alignments, but also a growing number of post-translational modifications, protein structures, interactions and pathways. All these multiple data types offer valuable insights on their own, but together they can give a more complete picture. By such a synthesis there is a great potential to perform systematic mechanistic analyses of all genetic variants.

However, the volume and diversity of available protein -omics data makes the process of gathering, integrating and interpreting data to deduce mechanism very challenging (Gomez-Cabrero et al., 2014; Subramanian et al., 2020). Assembling heterogeneous information to create a unified view requires coping with the lack of format standardization, multiple and asynchronous data updates, and many other issues. This process needs a systematic approach because manual assembly of data in different formats is cumbersome and prone to errors. Moreover, visualization is crucially important as it is often key to seeing the critical functional details that can explain the mechanism of particular genetic variants. The pace of data generation makes visualization essential for the interpretation of increasingly complex biological data—especially for data-driven research—and is equally important for communicating hypothesis and discoveries (O'Donoghue et al., 2010, 2018). Over the last two decades, many visualization tools have been published, and almost every previously existing resource has added or improved its data visualization options.

Proteins, due to the particularities of their organization, function and interactions, pose particular visualization challenges. Because they are linear amino acid sequences, they are usually represented in 2D diagrams where different functional features, such as domains, PTMs, variants, or any other sequence annotations can be displayed while maintaining their corresponding sequence positions and dimension. Such diagrams are for example used by Pfam (Mistry et al., 2021) and its popular domain diagram creation tool[5] to represent domain architecture (Figure **3.1**). Furthermore, simultaneous visualization of different features can enhance the ability to detect patterns; for instance, ProtVista (Watkins et al., 2017) is the feature viewer tool used by UniprotKB, and it allows integrative visualization of the many curated sequence features in their database (Figure **3.1**). However convenient these one-dimensional depictions are, proteins, of course, adopt specific three-dimensional structures that ultimately determine their function. There are dozens of visualization programs for protein structures, such as PyMOL (Schrödinger & DeLano, 2020), RasMol (Sayle & Milner-White, 1995), Jmol[6], Chimera (Pettersen et al., 2004), VMD (Humphrey et al., 1996), including several (e.g. JSmol) that can be readily embedded into web applications, for instance, to permit interactive relations with alignments or domain diagrams.

Moreover, proteins function through coordinated interactions with other proteins and molecules. Databases such as IntAct (Orchard et al., 2014), BioGRID (Oughtred et al., 2021) or STRING (Szklarczyk et al., 2019) collect thousands of protein interactions and allow to visualize them in canonical protein networks that simply represent proteins as spherical nodes, and their binary interactions as edges between them (Figure **3.1**). This kind of network however is only useful to indicate if an interaction happens or not rather than explaining how. Protein interactions occur through specific regions (interfaces), which are often functionally and structurally conserved sub-sequences known as protein signatures (domains, motifs, active sites). Moreover, proteins can interact with different partners through distinct interfaces. In consequence, single lines linking two nodes may not be really representative of an interaction if the particular interfaces that are involved are known. There are some tools that provide more detailed visualizations such as iELM (Weatheritt et al., 2012), which allows to generate PPI networks composed of the involved interaction-mediating linear motifs; ComplexViewer (Combe et al., 2017), which represents the specific binding regions along the

[5] https://pfam.xfam.org/generate_graphic
[6] Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/

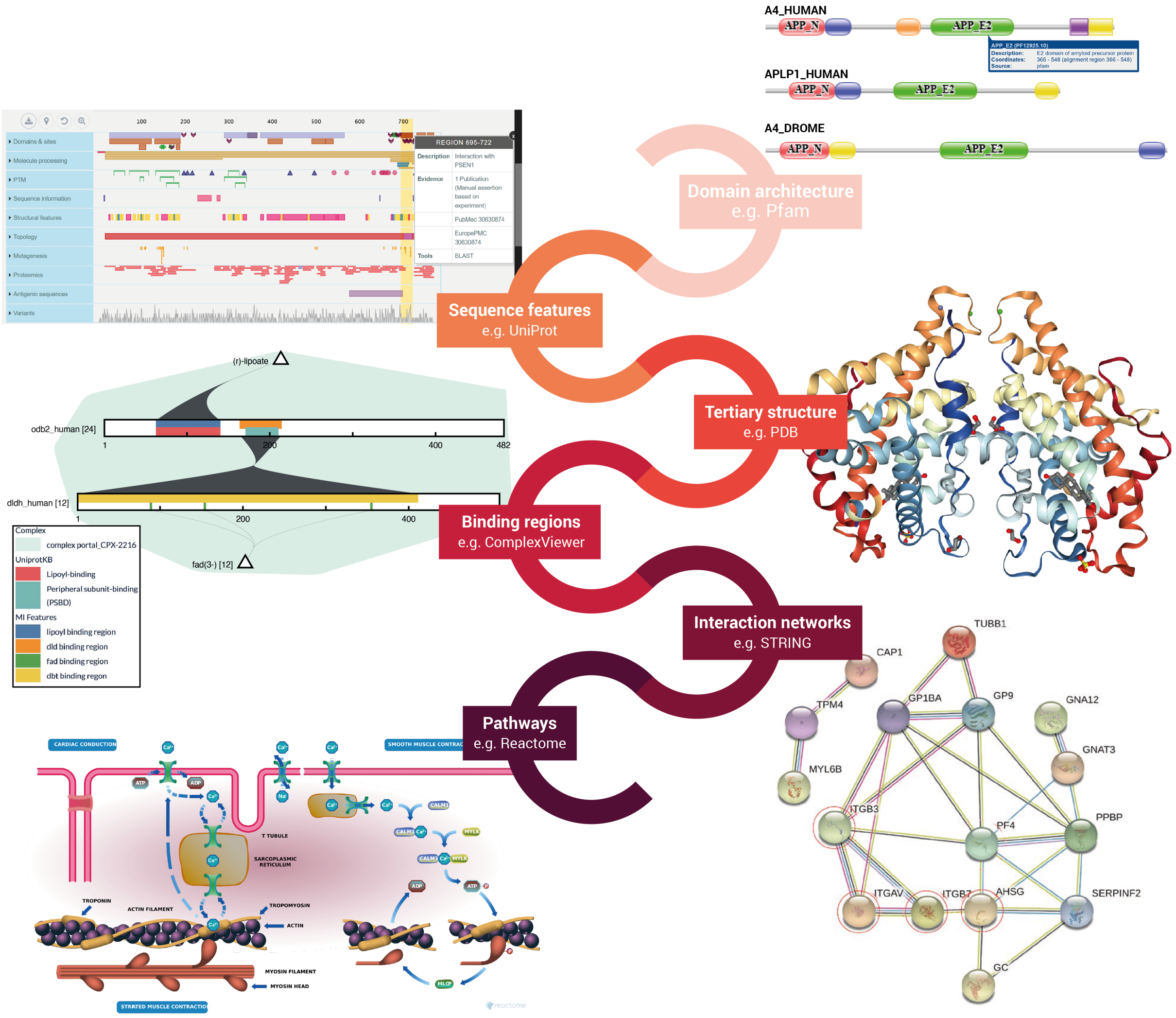
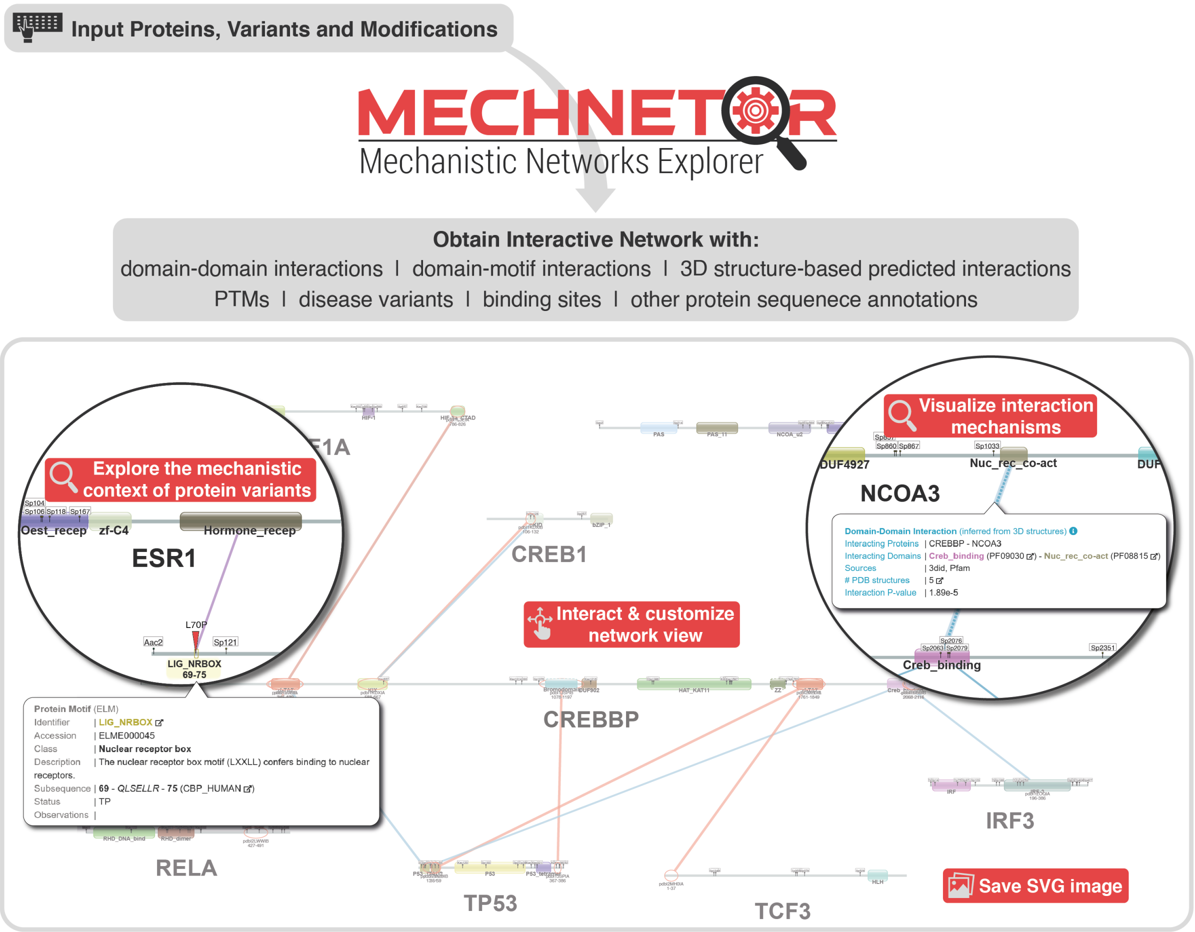


Figure 3.1: Examples of different kinds of protein features visualization.

protein sequences and maintains the topology and stoichiometry of macromolecular complexes (Figure **3.1**); or Interactome3D (Mosca et al., 2013), which annotates PPI networks with available structures of the interacting proteins. However, these tools still lack plenty of other molecular details, for example, they do not allow the examination of positional differences, such as mutations or PTMs. There is a gap between the way protein sequence features are typically visualized, in linear diagrams, one protein at a time, and protein interaction network depictions in the form of graphs, where proteins are represented by simple nodes that do not allow to display any features. Since there is no way to readily visualize both at the same time, studying positional information such as changes, modifications or other annotations, in the context of protein interactions and the implicated protein regions is still challenging.

For this purpose, I, with the help from others, developed Mechnetor (Mechanistic Networks Explorer), a novel web resource that helps understanding protein mechanism in groups of interacting proteins, and studying protein changes in the right mechanistic context (González-Sánchez et al., 2021). Mechnetor helps in two main ways. First, it automatically gathers and integrates diverse interaction data—binary interactions with experimental evidence, known and predicted domain-domain (DDIs) and domain-motifs interactions (DMIs), and 3D structure-based interactions—coupled with information regarding function, post-translational modifications, variants and other annotations. Second, it represents the results into a fully interactive network that is visually appealing and easy to interpret, and which enables users to examine complex interaction mechanisms simultaneously. Figure **3.2** outlines the Mechnetor web-server.


Input Proteins, Variants and Modifications
MECHNETOR
Mechanistic Networks Explorer
Obtain Interactive Network with:
domain-domain interactions | domain-motif interactions | 3D structure-based predicted interactions
PTMs | disease variants | binding sites | other protein sequenece annotations
Explore the mechanistic context of protein variants
Oest_recep
zf-C4
Hormone_recep
ESR1
L70P
Aac2
Sp121
LIG_NRBOX 69-75
Interact & customize network view
CREB1
CREBBP
RELA
TP53
TCF3
IRF3
Visualize interaction mechanisms
DUF4927
Nuc_rec_co-act
NCOA3
Creb_binding
Save SVG image

Protein Motif (ELM)
Identifier | LIG_NRBOX
Accession | ELME000045
Class | Nuclear receptor box
Description | The nuclear receptor box motif (LXXLL) confers binding to nuclear receptors.
Subsequence | 69 - *QLSELLR* - 75 (CBP_HUMAN)
Status | TP
Observations |

Domain-Domain Interaction (inferred from 3D structures)
Interacting Proteins | CREBBP - NCOA3
Interacting Domains | Creb_binding (PF09030) - Nuc_rec_co-act (PF08815)
Sources | 3did, Pfam
PDB structures | 5
Interaction P-value | 1.89e-5

Figure 3.2: Graphical overview of the Mechnetor web-server. Users can input interacting proteins, protein variants and/or protein modifications. Mechnetor will automatically annotate them with diverse mechanistic data (functional, structural, interaction data) and present them to the user in an interactive network (mechanistic network). This network can be explored and customized by using several interactivity options, as well as exported as a vector image

The next sections will describe the development of this tool, from obtaining and integrating the data, to a description of Mechnetor's features and all mechanistic information it provides. At last, I present a statistical analysis of the data to evaluate its coverage and usefulness, as well as discuss particular examples that illustrate Mechnetor's features and use.

3.2 Mechnetor's database

3.2.1 Data sources

Mechnetor builds upon multiple data freely available in databases or resources that themselves were constructed from the experimental results, and thus the efforts, of many researchers. In order to exploit these data without propagating errors, it is essential to understand the nature of the information they contain and how it is structured.

UniProt: The Universal Protein Resource

UniProt (Bateman et al., 2021) is a freely accessible, comprehensive repository of protein sequence and functional information. It is an initiative of the UniProt Consortium, a collaboration between the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). The central component is the UniProt Knowledgebase which contains highly curated data for more than 120 million proteins across all living organisms (more than 80 thousand species). UniProtKB is itself divided in two datasets. The largest of the two, the UniProtKB/TrEMBL, contains protein records that were automatically annotated by a computational pipeline. It was initially created to keep pace with the high sequence data generation volume, and to this day all deposited protein sequences are placed here. Records are then selected for full manual annotation by an expert curation team and integration into the second database, the UniProtKB/Swiss-Prot, which contains only curated (termed "reviewed" by UniProt), non-redundant protein records with high-quality information. Thus, data from multiple sources are integrated, interpreted and standardized with the goal of creating the most comprehensive functional description of a protein in a single record (Boutet et al., 2016). This information ranges from location, gene/functional ontologies, domains, interactions, orthology, post-translational modifications, polymorphisms, alternative splicing and much more. In later releases, UniProt placed a special emphasis on the annotation of the functional impact and clinical significance of thousands of protein variants, many of which could be associated with particular Mendelian diseases (Amberger et al., 2019; Sherry et al., 2001).

For all these reasons, UniProtKB is an excellent starting point to gather information about any protein of interest. It provides a central hub of biological knowledge that is linked to data sources or other databases, and can also be accessed programmatically (via API) or downloaded (via FTP) in community-recognised formats suitable for systematic approaches. UniProt identifiers for protein sequences and protein sequence features are also unique, stable and traceable. This is essential for data integration as diverse data repositories can reference the exact same proteins allowing for data cross-examination.

Pfam: The database of Protein Families

The Pfam database (Mistry et al., 2021) is a large collection of protein families and domains, comprising more than 19,000 entries in its release 34.0. Each entry is defined by a seed alignment of representative sequences and a profile hidden Markov model (HMM) built from this seed alignment, which can be used to find new members in other sequence databases. Thus, Pfam entries are evolutionary conserved modules at the sequence level, and are classified into one of six types: *Family*, *Domain*, *Motif*, *Repeat*, *Coiled-Coil* or *Disordered*, although the vast majority belong to the first two. A Pfam *family* is the most generic class and only indicates that the proteins are related, but a Pfam *domain* corresponds to a single, compact, globular structure, and thus, an autonomous unit that can be found in different protein contexts. Many of these are the product of the sequence-based, domain-hunting activities of the 1990s, which captured all the common domains (e.g. kinase catalytic domains, SH3, PH, Ras, etc.). Pfam *motifs* are short units found outside globular domains (e.g. AT-hook, IQ calmodulin-binding motif), while *repeats* are small units that form a stable structure only when two or more adjacent copies are present (e.g. WD40, TPR, HEAT, etc.). Pfam entries also include manually annotated functional information, as well as their phylogenetic distribution, structural models and interactions between domains observed in structures from PDB, if any (Finn et al., 2014). There are still numerous entries (around 25%) for domains of unknown function and uncharacterized protein families, but these are being annotated over time. Pfam uses UniProtKB sequences as reference, and its sequence and residue coverage is of ∼77% and ∼53%, respectively.

ELM: The Eukaryotic Linear Motif Resource

The eukaryotic linear motif (ELM) resource (Kumar et al., 2019) is the most comprehensive repository of short linear motifs, or short protein interfaces (defined in detail in section 1.5.4), in eukaryotic organisms. ELM stores experimentally determined motif instances that are curated manually from the literature and classified into carefully annotated motif classes, according to the interaction they mediate. A motif class is mainly defined by a regular expression, derived from the observed sequence patterns of the motif instances it comprises, which specifies the residues that confer affinity and specificity to the interaction. Annotation of motif classes also includes detailed descriptions about their function, their binding domains and their taxonomic range. ELM classes are classified into several types according to their broad function as ligand (e.g. *LIG_SH3_1*, a SH3 domain-binding motif), in subcellular targeting (e.g. *TRG_ER_KDEL_1*, a Golgi-to-ER targeting signal), in proteolytic cleavage (e.g. *CLV_C14_Caspase3-7*, caspase-3 and -7 cleavage motif), in docking (e.g. *DOC_PP1_RVXF_1*, docking motif for PP1c), in degradation (e.g. *DEG_APCC_DBOX_1*, destruction motif recognize by the anaphase-promoting ubiquitin ligase complex APC/C) or in post-translational modification sites (e.g. *MOD_Plk_1*, phosphorylation site of Polo-like kinases). As of March 2021, ELM contains 291 motif classes, created from 3,542 experimentally validated instances, which interact with 147 globular domains. The provided patterns can be used to identify new potential motif instances in almost every protein sequence although, for many motif classes with simple patterns, the vast majority of identified instances will be false positives. Any predicted instance needs to be scrutinized carefully or cross-referenced with other information to assess their validity.

BioGRID: Biological General Repository for Interaction Datasets

BioGRID (Oughtred et al., 2021) is a database of protein-protein, genetic and protein-chemical interactions manually curated from experimental evidence in the biomedical literature. It was initially created in 2006 as a comprehensive compendium of all biological interactions in the budding yeast. Today, BioGRID includes data for more than 70 species, over 1.93 million protein and genetic interactions (670,000 interactions for human), extracted from more than 63,000 publications, although human and yeast data make up for almost two thirds of the total. Interactions in the database are described according to a controlled vocabulary, and interaction evidence is classified according to the experiment system. This allows for interactions between the same proteins to be quantified according to the number of experimental sources, both high and low throughput.

As opposed to other resources such as STRING (Szklarczyk et al., 2019) or GeneMANIA (Franz et al., 2018), BioGRID does not include predicted interactions. BioGRID is a high-confidence interaction repository and is often used as a gold standard in many studies. Data can be queried online, accessed programmatically, or downloaded in different formats. BioGRID data is not UniProtKB-centric meaning it uses their own protein identifiers as well as gene symbols.

3did: database of three-dimensional interacting domains

3did (Mosca et al., 2014) is a database of protein interactions mediated by globular domains binding other domains (domain-domain interactions) or binding short linear peptides (domain-motif interactions), for which high resolution 3D structural templates are known. These data are generated by searching for domains, as defined by Pfam, in the protein sequences of PDB structures, and then estimating the number of intrachain and interchain contacts between any pair of domains to determine whether they interact or not. In turn, domain-motif interactions are detected by a machine learning method (described in (Stein & Aloy, 2010)) that exploits the particular structural features of these peptides.

3did website allows queries of particular domains or motifs of interest to obtain their interacting domains and motifs, and the 3D templates from the PDB that support those interactions. Moreover, it is possible to download the complete list of interacting domain-domain and domain-motif pairs along with all the instances in 3D structures where the interaction is observed. This information allows users to infer mechanism and structurally characterize new protein-protein interactions.

PhosphositePlus

Reliable knowledge on PTMs can improve understanding of the fundamental mechanisms of cellular signalling, and the particular role of cellular regulation in health and disease. PhosphositePlus (PSP) (Hornbeck et al., 2015) is an online resource for the study of experimentally observed PTMs, including phosphorylation, acetylation, methylation, ubiquitination, and O-glycosylation. PSP data comes from manual curation of both low- and high-throughput data sources in the literature, as well as from their own mass spectrometry experiments carried out at Cell Signaling Technology Inc. PTM sites are already annotated on UniProtKB protein sequences, what greatly facilitates cross-referencing with other protein data.

COSMIC: the Catalogue of Somatic Mutations in Cancer

COSMIC (Tate et al., 2019), the most comprehensive database of somatic mutations in human cancers, was already introduced in the previous chapter (Section **2.2.1**). COSMIC data are annotated with information regarding source, sample, cancer type, change in protein sequence, and cross-references with other datasets, and are regularly updated. Its web portal allows an in-depth exploration of the functional effects of cancer mutations, through detailed tables and interactive visualizations.

Protein-coding variants in COSMIC are unfortunately not mapped to UniProtKB protein sequences. This means that there are instances where the provided amino acid change does not match the corresponding UniProtKB sequence. Thus additional checking and remapping steps might be required if one wants to use COSMIC data in an UniProtKB-centric fashion.

3.2.2 Data integration and database creation

Mechnetor uses an internal PostgreSQL[7] database, tailor-made to support its needs. Data from the above sources is pre-assembled and integrated, dealing at this point with data quality control and identifier matching issues. This step is crucial as it ensures that all relevant information for any user query can be retrieved fast and efficiently.

UniProtKB is the starting point for basic protein data: protein sequences and their identifiers, genes and descriptions. For every target organism, all UniProtKB proteins were obtained and this defines the complete set of proteins that are available in Mechnetor: in total, close to 317,000 protein entries, corresponding to approximately 152,000 genes in 8 model organisms (plus SARS-CoV2) (Figure **3.3**); all other data must be matched to these proteins/genes. From UniProtKB, multiple protein sequence annotations were also extracted, including chemical ligand and metal ion binding sites, mutagenesis-altered sites, as well as PTMs and disease-linked protein variants—the vast majority of them being human germline changes involved in Mendelian diseases described in the OMIM database (Amberger et al., 2019). The complete list of UniProtKB sequence features is shown in Table **3.1**.

PTMs from UniProt were integrated with those from PhosphositePlus database into a single, more exhaustive dataset. In addition, for human proteins, cancer-related missense variants from the COSMIC database are included (the same dataset used in the previous chapter; see Methods **2.3.1**). For each proteome, Pfam domains matches were obtained

[7] https://www.postgresql.org

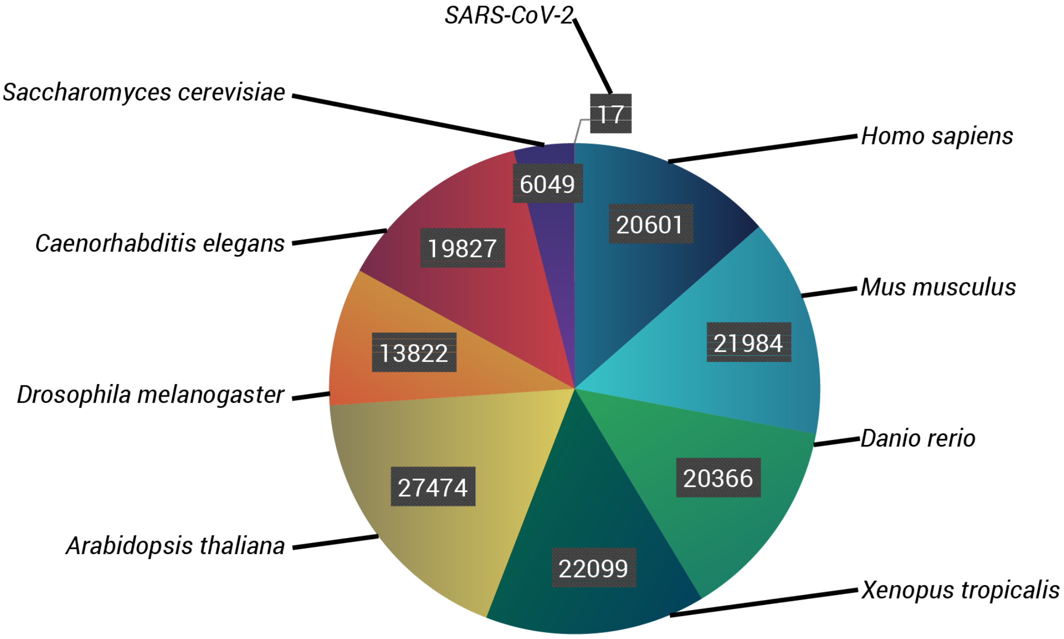


Figure 3.3: Number of genes of every species included in the Mechnetor database.

directly from the Pfam database (release 34.0)[8]. In addition, the PfamScan tool (Madeira et al., 2019) was used to search for domain matches in those protein sequences that were not yet present in the Pfam database. In total, 61% of all proteins in Mechnetor's database have at least one match to a Pfam entry, with a total residue coverage of 29.9%. On average, proteins with domains have 1.8 domain matches per sequence.

More than 1.5 million protein-protein interactions were extracted from the BioGRID database (version 4.2.191), including their corresponding experiment throughput and publication source, and matched to UniProtKB accessions. Around 16,000 domain-domain interactions were non-redundantly obtained from Pfam and 3did, which are already provided as pairs of Pfam domain identifiers together with the PDB codes of the 3D structures that support them. Unfortunately, Pfam does not provide information about the structure where their domain interactions originate anymore; however, these DDI were kept as I noticed that a significant number were only present in their dataset (Figure **3.4**).

Thousands of known short linear motif (SLiM) instances and their binding-domains (domain-motifs interactions) were gathered from two sources: motifs curated from the literature, obtained from ELM, and motifs observed in 3D structures, obtained from 3did. The regular expressions of all SLiM classes were then used to perform a proteome-wide search of motif matches in every organism and identify all potential motif instances. The total

[8] *http://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/proteomes/*

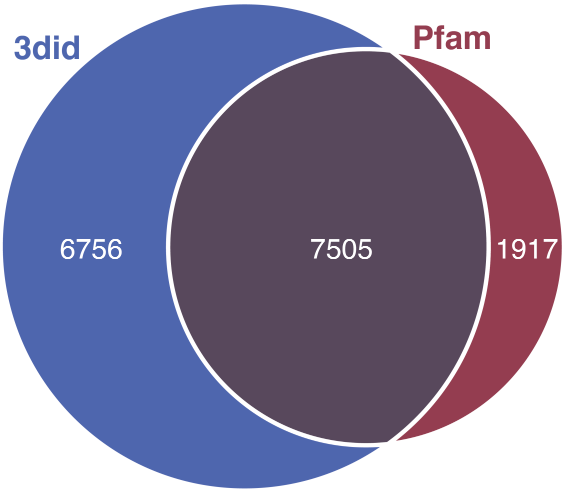


Figure 3.4: Number of unique and common domain-domain interactions extracted from the 3did and Pfam databases.

number of matches for any motif class is inversely proportional to the complexity of its pattern. For example, in human, the complex pattern of the WH1 domain-binding motif *LIG_WH1* (`ES[RK][FY].F[HR][PST][IVLM][DES][DE]`) is only found in the sequences of three WASP proteins which correspond to the three experimentally verified instances of this class. Opposite, the pattern of a simple motif such as the Casein kinase 1 phosphorylation site *MOD_CK1_1* (`S..([ST])...`) is present in more than 90% of human proteins. Thus, most identified motifs are simply sequence pattern matches and likely to be false positives, only a tiny fraction of those correspond to experimentally confirmed motifs (<1% in human, for instance).

3.2.3 Manual reviewing of domain-motif interactions from ELM

The domain-motif interaction (DMI) data from the ELM resource simply contains a list of motif class and domain identifiers pairs[9]. The basic assumption is that DMIs can be extrapolated to any protein pair where the motif and its interaction domain are present, but the reality is that although motifs may interact with the same domain type, not all of them can interact with the same proteins. For instance, more than 30 ELM motif classes are known to interact with the protein kinase domain (*Pkinase* or *PF00069* in Pfam), which is present in hundreds of proteins only in the human proteome. However, most protein kinases only recognize particular motif classes. For instance, the *DOC_MAPK_gen_1* motif is the exclusive docking site of members of the MAP kinase family (MAPKs), while *MOD_NEK2_1* is the

[9] http://elm.eu.org/interactiondomains

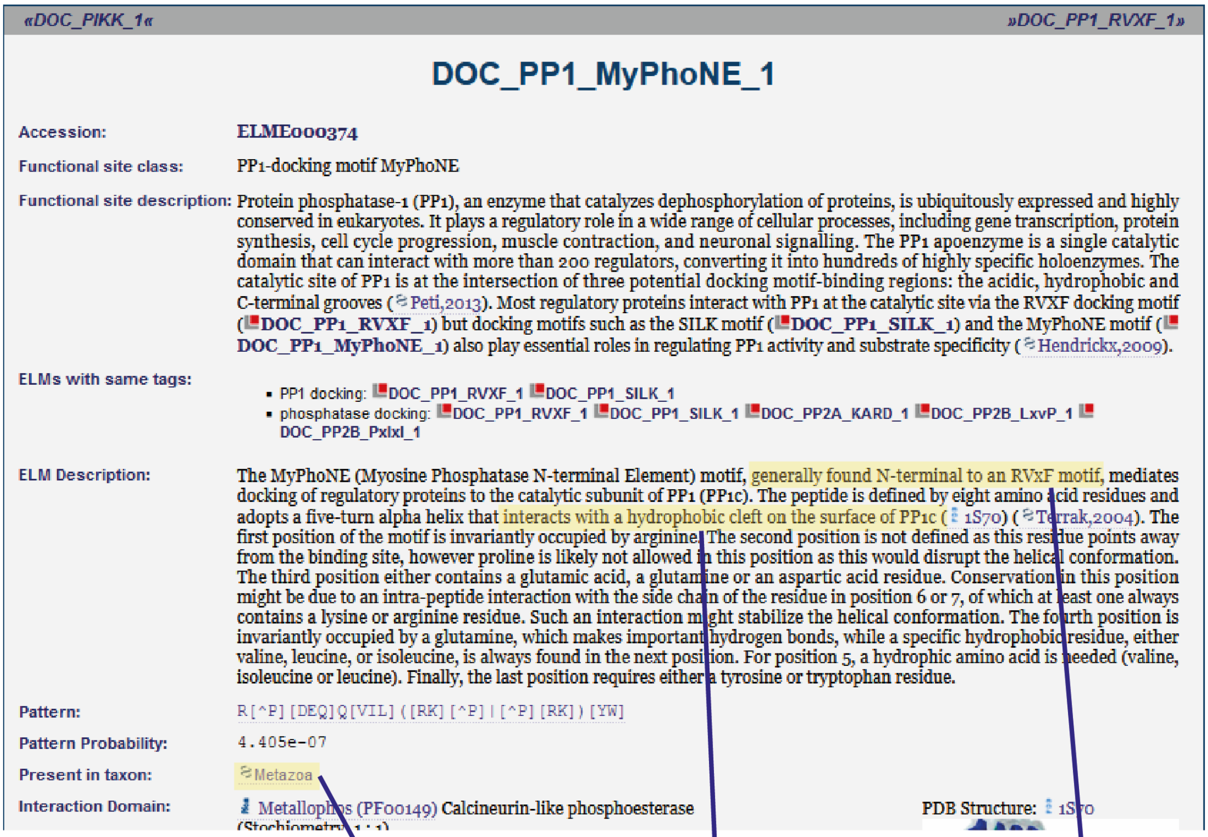
specific phosphorylation site of the Serine/Threonine-protein kinase NEK2. Moreover, some motifs are exclusively located in certain proteins, like the Pex14-binding site in Pex5 (*LIG_PEX14_1*); while others are functional only in certain taxons, such as *LIG_PAM2_2*, which is a metazoan-specific variant of the PABP-interacting motif. Thus instances of motifs out of their specific functional contexts can be completely disregarded.

All these annotations can be used to significantly narrow down DMIs to the most likely cases, but unfortunately they can only be found in the manually reviewed entries of each motif class at the ELM website, and not in a systematic format that can be automatically used. Consequently, I reviewed all entries in the ELM database and manually created a series of parameters and restrictions for each domain-motif pair that allows to automate the process of DMI filtering. These include restricting the DMI to certain taxons, restricting the interaction domain and/or the motif to only certain proteins, requiring the presence of other linear motifs in the same protein, or requiring the presence of an experimentally determined phospho-Ser/Thr/Tyr within the motif (Figure **3.5**). In order to present only the most biologically relevant information, Mechnetor is by default configured to only show DMI that fulfil these requirements.

3.2.4 Scoring and inferring DDI and DMI interactions

In addition to the interactions obtained from Pfam and 3did, I also used a method to infer domain-domain interactions based on domain co-occurrence and first proposed by (Sprinzak & Margalit, 2001). Also known as the association method, it tries to identify pairs of sequence-signatures (such as domains) that, given their frequencies in the proteome, co-occur in interacting proteins more often than expected by chance and thus are likely to mediate those interactions.

For each organism independently, a subset of non-redundant PPI reported by two or more experiments were extracted from the full PPI set defined above. Next, for every possible combination of domain pairs, the number of interacting proteins pairs containing said combination were counted. For example, for domain A and domain B, this is the number of interacting protein pairs where one protein contains domain A and its partner contains domain B, then it was divided by the total number of interacting proteins pairs to obtain its observed frequency ($Obs_{A,B}$). Domains were counted only once per protein, even if the same domain appeared two or more times. Correspondingly, the frequencies expected by chance were calculated as the combined probability of finding a protein with domain A and finding



«DOC_PIKK_1« »DOC_PP1_RVXF_1»

DOC_PP1_MyPhoNE_1

Accession: ELME000374

Functional site class: PP1-docking motif MyPhoNE

Functional site description: Protein phosphatase-1 (PP1), an enzyme that catalyzes dephosphorylation of proteins, is ubiquitously expressed and highly conserved in eukaryotes. It plays a regulatory role in a wide range of cellular processes, including gene transcription, protein synthesis, cell cycle progression, muscle contraction, and neuronal signalling. The PP1 apoenzyme is a single catalytic domain that can interact with more than 200 regulators, converting it into hundreds of highly specific holoenzymes. The catalytic site of PP1 is at the intersection of three potential docking motif-binding regions: the acidic, hydrophobic and C-terminal grooves (Peti,2013). Most regulatory proteins interact with PP1 at the catalytic site via the RVXF docking motif (DOC_PP1_RVXF_1) but docking motifs such as the SILK motif (DOC_PP1_SILK_1) and the MyPhoNE motif (DOC_PP1_MyPhoNE_1) also play essential roles in regulating PP1 activity and substrate specificity (Hendrickx,2009).

ELMs with same tags:

- PP1 docking: DOC_PP1_RVXF_1 DOC_PP1_SILK_1
- phosphatase docking: DOC_PP1_RVXF_1 DOC_PP1_SILK_1 DOC_PP2A_KARD_1 DOC_PP2B_LxvP_1 DOC_PP2B_PxIxI_1

ELM Description: The MyPhoNE (Myosine Phosphatase N-terminal Element) motif, generally found N-terminal to an RVxF motif, mediates docking of regulatory proteins to the catalytic subunit of PP1 (PP1c). The peptide is defined by eight amino acid residues and adopts a five-turn alpha helix that interacts with a hydrophobic cleft on the surface of PP1c (1S70) (Terrak,2004). The first position of the motif is invariantly occupied by arginine. The second position is not defined as this residue points away from the binding site, however proline is likely not allowed in this position as this would disrupt the helical conformation. The third position either contains a glutamic acid, a glutamine or an aspartic acid residue. Conservation in this position might be due to an intra-peptide interaction with the side chain of the residue in position 6 or 7, of which at least one always contains a lysine or arginine residue. Such an interaction might stabilize the helical conformation. The fourth position is invariantly occupied by a glutamine, which makes important hydrogen bonds, while a specific hydrophobic residue, either valine, leucine, or isoleucine, is always found in the next position. For position 5, a hydrophic amino acid is needed (valine, isoleucine or leucine). Finally, the last position requires either a tyrosine or tryptophan residue.

Pattern: `R[^P][DEQ]Q[VIL]([RK][^P]|[^P][RK])[YW]`

Pattern Probability: 4.405e-07

Present in taxon: Metazoa

Interaction Domain: Metallophos (PF00149) Calcineurin-like phosphoesterase (Stochiometry: 1 : 1)

PDB Structure: 1S70

ELM Identifier	Interaction Domain Identifier	In taxon	Not in taxon	ELM-containing protein	Domain-containing protein (Human)	Domain-containing protein	Required Phospho-sites	Other Motifs required
DOC_PP1_MyPhoNE_1	PF00149	Metazoa			PPP1C*	PPP1C* Pp1*[Dme] Gsp*[Cel]		DOC_PP1_RVXF_1
TRG_LysEnd_GGAAcLL_2	PF00790	Vertebrata		GGA1, GGA3	GGA1, GGA3	GGA1, GGA3	1	
LIG_Pex14_3	PF04695	Eukaryota	Fungi	PEX5	PEX14	PEX14		

Figure 3.5: Examples of additional annotations for domain-motif interactions which are the result of manually reviewing ELM entries: the metazoan-specific PP1-dockin motif (*DOC_PP1_MyPhoNE_1*); the auto-inhibitory interaction between motif *TRG_LysEnd_GGAAcLL_2* and VHS domain (*PF00790*) in GGA1/3 which is regulated by Ser phosphorylation (first residue in the motif); and the Pex14 ligand motif (*LIG_Pex14_3*) in the peroxisomal import receptor Pex5, which varies in Fungi (Fungi instead contain the *LIG_Pex14_4*). Original ELM table only lists the interacting motif and domain (first two columns; although interaction domain name and descriptions are also included, they have been omitted here), the rest (yellow) are new. For specifying the proteins that have to contain the motif instance and/or the domains for the interaction to be accepted, regular expressions were used to catch different members of the same family

a protein with domain B in the whole dataset, which is the product of their individual frequencies, F_A and F_B:

$$Exp_{A,B} = F_A\, F_B = \left(\frac{N_A}{N}\right)\left(\frac{N_B}{N}\right)$$

where N_A and N_B are the number of proteins in the set that respectively contain domains A and B at least once, and N is the total number of proteins.

Finally, to determine the degree of correlation, a comparison between observed and expected frequencies was calculated as the logarithm of the odds ratio, or log-odds. Thus the association score is:

$$A_{A,B} = \log_2\left(\frac{Obs_{A,B}}{Exp_{A,B}}\right)$$

A sequence-signature pair is enriched if its association score is greater than or equal to 2, but only if its observed count is also greater than or equal to 5 and the individual counts of proteins containing each of the signatures are greater than or equal to 4. These latter restrictions on the occurrence numbers are set to avoid pair of signatures that, despite having a high association score, are very rare.

Additionally, to estimate the significance of these domain associations, a *P*-value was calculated through a binomial test. This was also done for all previously extracted DMIs. The *P*-value is reported along with all other interaction details and can be used as a threshold to exclude interactions/associations that are statistically irrelevant.

3.2.5 Predicting interactions through tertiary structure

Since the number of interaction types is limited (Aloy & Russell, 2004), it is to a degree possible to infer that proteins homologous to a known interacting protein pair will interact in a similar way (Aloy et al., 2003). This idea has been explored to determine the level of sequence similarity that is required for a pair of proteins to interact in an analogous way to one of known 3D structure. The web tool InterPreTS, when given a pair of protein sequences, will try to find homologues of known structure that are suitable for modelling the interaction between them (Aloy & Russell, 2002, 2003). More specifically, the method looks for regions in the query proteins that are homologous to those participating in an interface in the 3D structure of interacting proteins or protein complex. Then, the required contacts for the interaction are tested in these regions. If the template is suitable, the result is not only the identification of a potential interaction between the query proteins but also the identification of the protein segments that might mediate said interaction.

For Mechnetor, I adapted and integrated an internal version of InterPreTS (this version can run in batch, without graphic interface) to search for templates within the PDB (version 2019) that can model the interaction between every protein pair included in the query. To save users time, InterPreTS predictions for the majority of known PPI of the currently supported organisms were pre-computed, but not for all protein pairs as that would take a very long time. As a solution, the database was designed to be populated with new predictions as new protein pairs are queried for the first time, saving time in future queries.

3.3 The Mechnetor web-server

Mechnetor is a web-based resource where users can directly query their own proteins or protein modifications of interest. The general workflow is represented in Figure **3.6**. Upon input submission, Mechnetor will systematically gather and integrate relevant functional and interaction information from its internal database, as well as, compute some additional data. As a result, the user will be presented with an interactive protein network where they can explore all the integrated data in a comprehensive way, with the aid of tools and options that facilitate visualization and interpretation. In addition, the results page also contains a fully searchable table that lists all interaction evidence contained in the network and that can be downloaded in different formats for local analysis. In its version 1.0, Mechnetor supports eight of the most common model organisms: human (*Homo sapiens*), mouse (*Mus musculus*), zebra fish (*Dario rerio*), frog (*Xenopus laevis*), fruit fly (*Drosophila melanogaster*), worm (*Caenorhabditis elegans*), mouse-ear cress (*Arabidopsis thaliana*) and yeast (*Saccharomyces cerevisiae*). The proteome of *SARS-CoV-2* is also included and can be queried in combination with human proteins (Figure **3.3**).

For use by non-experts, Mechnetor was built with a simple and intuitive interface, and provided it with optional functionalities to extend the repertoire of tasks it can be applied for. Each query is given a unique URL which means that the result page can be bookmarked for later access or shared with other people. Mechnetor is free and open to all users without login or registration requirements[10].

[10] *mechnetor.russelllab.org*

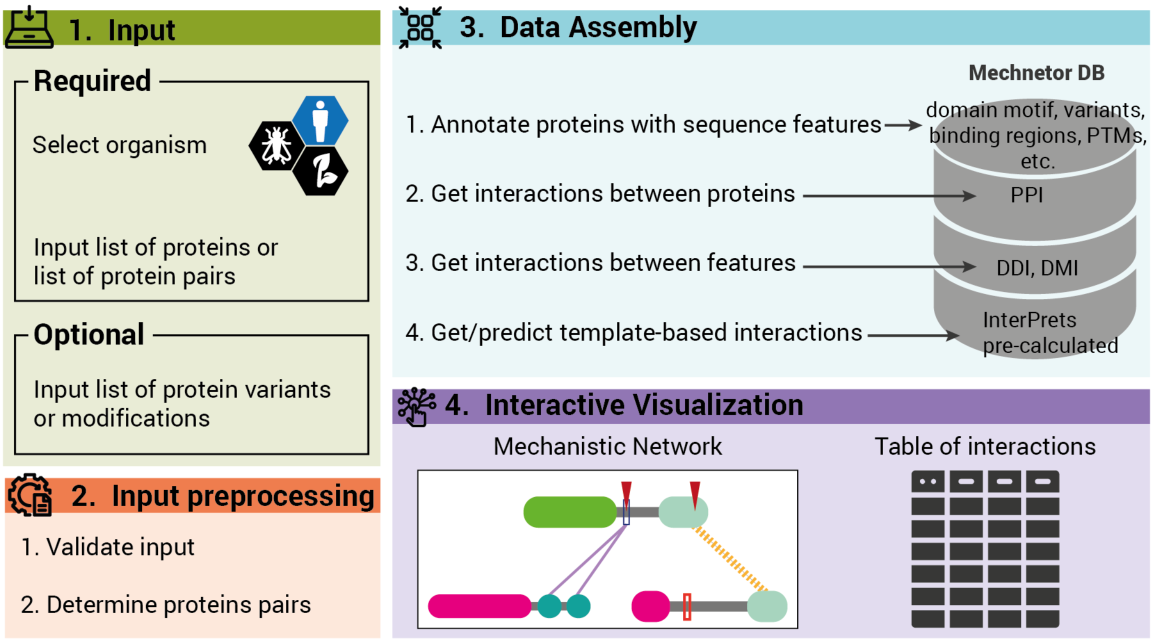

1. Input
Required
Select organism
Input list of proteins or list of protein pairs
Optional
Input list of protein variants or modifications
2. Input preprocessing
1. Validate input
2. Determine proteins pairs
3. Data Assembly
Mechnetor DB
1. Annotate proteins with sequence features
domain motif, variants, binding regions, PTMs, etc.
2. Get interactions between proteins
PPI
3. Get interactions between features
DDI, DMI
4. Get/predict template-based interactions
InterPrets pre-calculated
4. Interactive Visualization
Mechanistic Network
Table of interactions

Figure 3 6: The Mechnetor web-server pipeline.

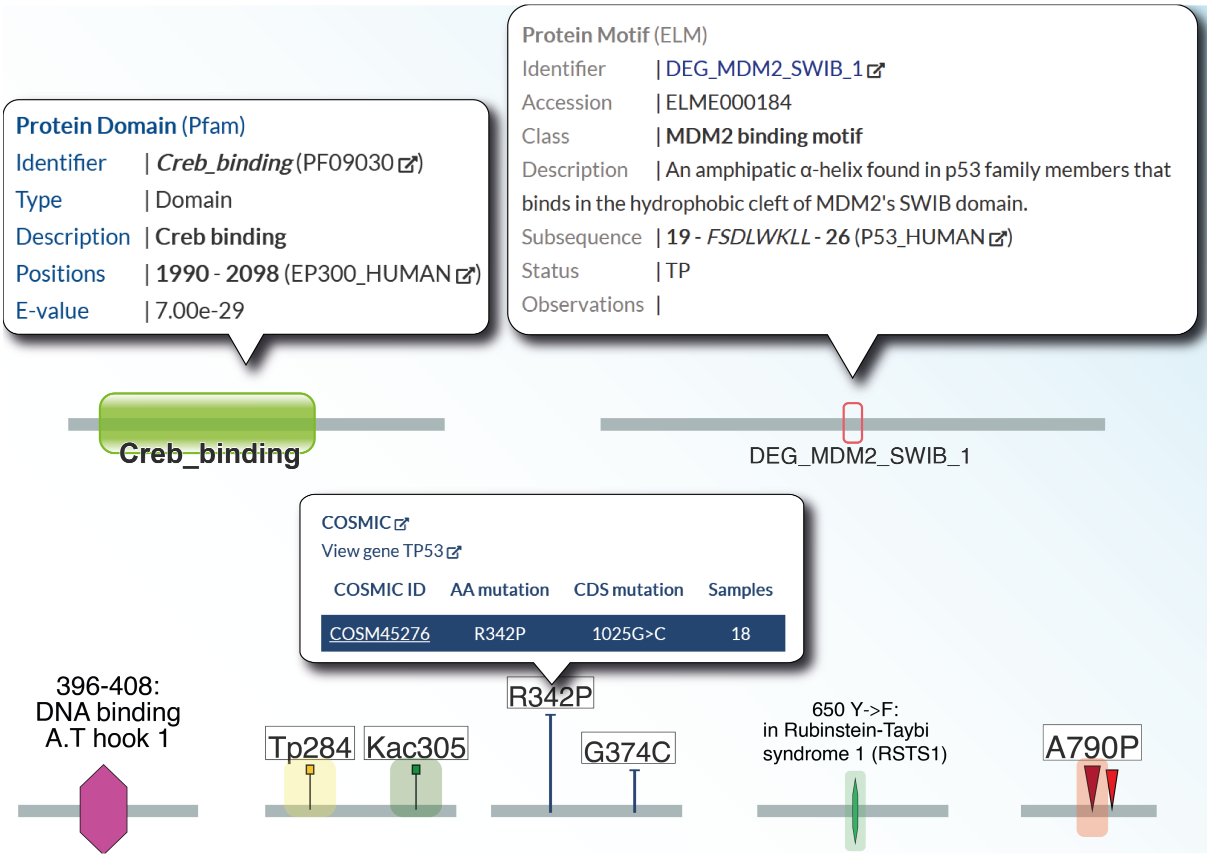
3.3.1 The mechanistic network

The main component of Mechnetor is the network viewer, which allows to manually explore all mechanistic data gathered and integrated for the query proteins through an interactive visualization. As many other network representations, it is comprised of nodes representing proteins and edges representing interactions between them. However, instead of simple-shaped nodes, here proteins are depicted as linear diagrams, proportional to their sequence lengths, with different functional elements (domain, linear motifs, PTMs, etc.) displayed in their corresponding protein positions/regions in a visually distinct fashion. This allows to show different types of interactions with edges that link entire proteins or specifically link the protein elements involved in the interaction, and are coloured according to the type of interaction they represent. Clicking on almost every element in the network will display a tooltip that shows specific information. In addition, some of the edges are weighted according to particular parameters, which is reflected in edge thickness to visually indicate the extent of interaction evidence; and/or are associated to a *P*-value (see section **3.2.4**) that indicates the strength of the association and can be used to set a maximum threshold. Users can utilize several interactivity options to explore and customize the network which are explained in section **3.3.4**. The specific details of all sequence features and interaction types are discussed in the next two sections.

3.3.2 Display of protein sequence features

Domains

Pfam domain architecture is shown roughly as described previously (Finn et al., 2016). Domains are depicted as rounded rectangles of different colours (randomly assigned each time the program is run) with the same domain in different proteins having the same colour, and with sizes that are proportional to domain lengths (Figure **3.7**). The domain tooltip shows the full domain name, start and end positions, type (*Family*, *Domain*, *Motif* or *Repeat*), E-value (of the particular protein region aligned to the domain's HMM) and identifier, which links to its Pfam entry. Mechnetor does not allow for overlapping domains.



Protein Domain (Pfam)

Identifier	*Creb_binding* (PF09030)
Type	Domain
Description	**Creb binding**
Positions	**1990 - 2098** (EP300_HUMAN)
E-value	7.00e-29

Protein Motif (ELM)

Identifier	DEG_MDM2_SWIB_1
Accession	ELME000184
Class	**MDM2 binding motif**
Description	An amphipatic α-helix found in p53 family members that binds in the hydrophobic cleft of MDM2's SWIB domain.
Subsequence	**19** - *FSDLWKLL* - **26** (P53_HUMAN)
Status	TP
Observations	

COSMIC
View gene TP53

COSMIC ID	AA mutation	CDS mutation	Samples
COSM45276	R342P	1025G>C	18

Figure 3.7: Mechnetor representations of different types of protein sequence features. From left to right and from top to bottom: domain, linear motif, DNA binding region (as an example of UniProtKB sequence annotation), PTMs (phosphorylation and acetylation), cancer missense mutations, genetic disease variant, and user-input variant.

Short linear motifs

SLiMs are represented by empty rectangles of randomly assigned colours (Figure **3.7**), which is shared among all instances of the same motif class in all proteins. For clarity, unlike domains, motif overlap is allowed but not all SLiMs identified in a protein are shown, only those for which a binding domain is present in another protein in the network, and therefore are potentially involved in an interaction. There is no E-value associated to SLiM instances either, as they are identified by sequence pattern matching (either they match or they do not), thus the number of motif instances identified in a protein is usually very high, but the majority are not relevant for the particular functional context defined by the other proteins in the network and they will not be displayed.

The motif tooltip specifies whether the particular instance is real (status: 'true positive') or a predicted one (status: 'unknown), its coordinates on the protein, and the corresponding sequence pattern. In addition, only for instances of ELM motifs (3did ones lack any annotation), tooltips include functional descriptions and links to the corresponding ELM entry. For SLiMs that require to overlap with phosphosites (e.g. a phosphorylated residue is required for recognition by a domain), their presence or not within the motif sequence is indicated.

Post-translational modifications

Three types of PTMs can be displayed independently on protein sequences: acetylations, phosphorylations and glycosylations. PTM sites are indicated by small flags, that are coloured differently by type (similar to 'lollipop' plots), and the corresponding text label (Figure **3.7**).

Sequence annotations

These include a variety of regions or positions of interest in the protein—binding sites, secondary structure, diverse functional regions—, which were obtained from the curated entries in UniProtKB (Table **3.1**). All these elements are coloured differently and can be toggled on and off independently, while relevant annotation is shown in their labels (Figure **3.7**). UniProtKB sequence features often provide functional information not found in the other data sources. For example, deletion mutagenesis experiments that lead to loss of particular interaction partners or inhibition of enzymatic activities are often captured in UniProtKB, as often are the mechanistic effects of particular PTMs or disease variant positions.

Table 3.1: Diverse protein features extracted from UniProtKB entries.

UniProtKB feature	Description
Binding site	Single-residue binding site for a chemical ligand (e.g. ATP)
Metal binding	Single-residue binding site for a metal ion (e.g. Iron)
DNA binding	DNA-binding domain (e.g. Homeobox)
Transmembrane	Membrane-spanning segments
Disulfide bond	Pair of cysteines residues forming disulphide bonds
Mutagenesis	Site altered by experimental mutation and its effect (e.g. W124A – No catalytic activity)
Region	Other functional regions of interest (e.g. regions mediating protein-protein interactions or regions involved in localization)

Human cancer variants

It is also possible to toggle on protein missense mutations observed in human cancers (pre-filtered from genome-wide sequencing dataset in COSMIC). They are represented by dark blue T-shaped arrows, which are length-proportional to the number of samples where the mutation has been observed (Figure **3.7**). Different mutations in the same protein residue are merged into a single node. Clicking on this node will display a table listing the different amino acid mutations, their corresponding nucleotide change in the coding DNA sequence, and the number of samples, as well as, links to the COSMIC entries of each mutation and the whole gene overview. By default, only cancer variants observed in more than two samples are displayed, but this limit can be changed via a slider.

Mendelian disease variants

There is also available a second pre-loaded set of variants, in this case related germline changes involved in Mendelian diseases described in the OMIM database and obtained from UniProt. Thin nodes indicate the position of the variant, while the amino acid change and disease are shown in node labels. Node colour is distinct for each disease to help identify variants related with different pathologies in the same protein (Figure **3.7**).

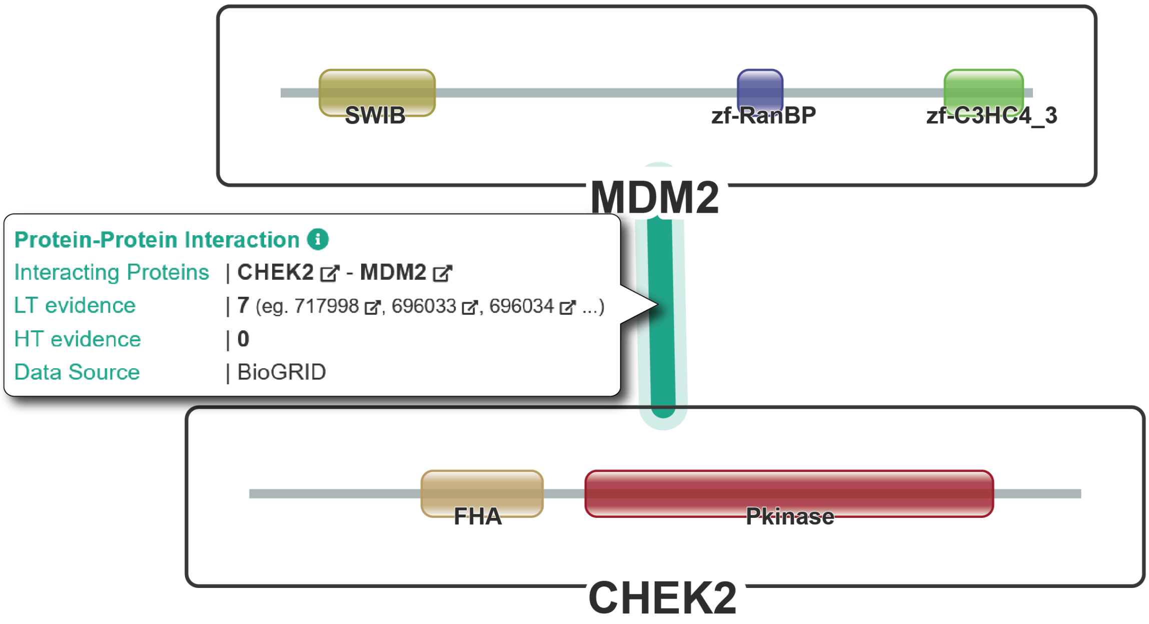
User-input variants

Protein variants, PTMs or other modifications submitted by the user will be added as an independent set of nodes, drawn as red inverted triangles at the specified protein positions. Labels indicate the variant/modification, as well as any custom annotation the user might have included (Figure **3.7**).

3.3.3 Display of interactions

Protein-protein interactions

Grey-coloured edges (green on mouse-over) represent the current experimental evidence supporting the interaction between pairs of proteins (binary protein-protein interactions). These edges simply link entire proteins without specifying the protein segments that are potentially involved in the interaction. They are weighted according to the total number of experiments—both high and low throughput—that determined the interaction between the two proteins. Both of these numbers are displayed in the edge tooltip, which also contains links to the BioGRID website entries of the two proteins (Figure **3.8**).



Protein-Protein Interaction ℹ

Interacting Proteins | **CHEK2** ↗ - **MDM2** ↗

LT evidence | **7** (eg. 717998 ↗, 696033 ↗, 696034 ↗ ...)

HT evidence | **0**

Data Source | BioGRID

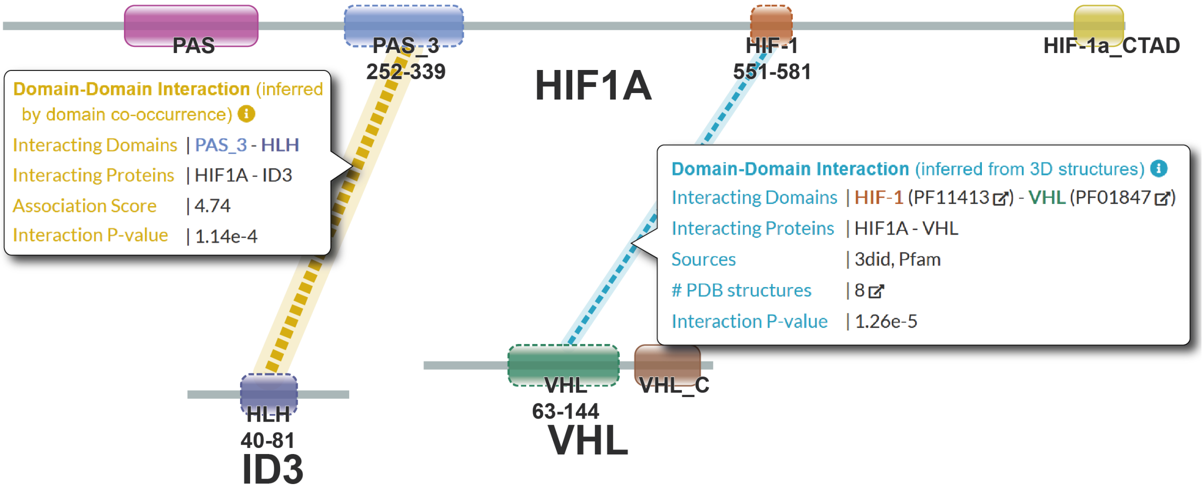
Figure 3.8: Example of PPI representation in Mechnetor network showing the interaction between MDM2 and CHEK2, which is supported by seven low-throughput experiments according to the BioGRID database.

Domain-domain interactions

Based on the fact that domains are evolutionarily conserved modules that interact independently, interactions between two proteins can be inferred when these have a pair of known interacting domains. Accordingly, for any pair of domains present in the network (in two different proteins), an edge is drawn between the two if there is evidence for the two domain classes to interact. The same edge is drawn for all domain instances of the same class, even if the domain is repeated in the same protein. Based on the type of evidence, there are two types of domain-domain interactions (DDIs):

- DDIs derived from contacts observed in 3D structures (described in section **3.2.2**), represented by cyan edges whose thickness is proportional to the number of 3D templates in the PDB where the domain interaction is present, thus indicating higher or lower confidence. The edge tooltip also includes the number of PDB structures, in addition to the pair of proteins and domains that are being linked, the source of the DDI (3did, Pfam, or both) and the interaction *P*-value (Figure **3.9)**.

- DDIs inferred by domain co-occurrence (following the method described in section **3.2.4**) which are represented by yellow edges that are weighted according to their *association score*. The higher the score, the most enriched is the domain pair association. A cut-off value for this score that limits which interactions of this type are shown in the network, can be easily modified with a slider. The exact *association score* and interaction *P*-values can be consulted in the tooltips (Figure **3.9)**.



Domain-Domain Interaction (inferred by domain co-occurrence) ℹ
Interacting Domains | PAS_3 - HLH
Interacting Proteins | HIF1A - ID3
Association Score | 4.74
Interaction P-value | 1.14e-4

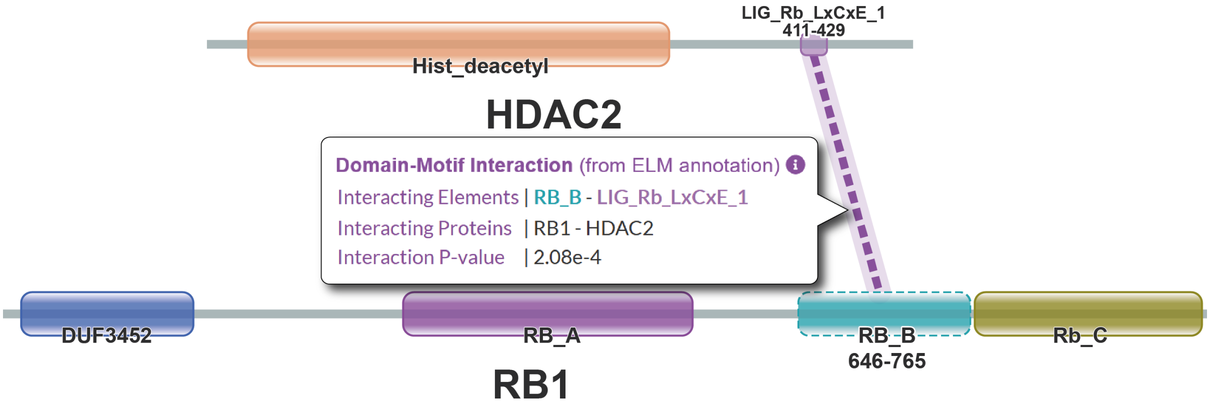
Domain-Domain Interaction (inferred from 3D structures) ℹ
Interacting Domains | HIF-1 (PF11413) - VHL (PF01847)
Interacting Proteins | HIF1A - VHL
Sources | 3did, Pfam
PDB structures | 8
Interaction P-value | 1.26e-5

Figure 3.9: Examples of the two DDI types supported by Mechnetor. DDIs that were inferred from 3D structures (cyan edges), or predicted by domain co-occurrence (method by (Sprinzak & Margalit, 2001); yellow edges).

Domain-motif interactions

Under a similar principle than DDIs, interactions edges are drawn between any linear motif and its known binding domain in any two proteins in the network, regardless the motif instance has been confirmed or just detected by sequence matching. Two types of domain-motif interactions (DMIs) were defined based on the data source, both unweighted but providing different levels of confidence:

- DMIs obtained from the ELM database—displayed as purple edges (Figure **3.10**)—are the most confident ones as they come from manual literature curation efforts. Based on the rich annotations of ELM motif classes, Mechnetor imposes a series of additional restrictions to reduce the number of false positive motif and DMI instances (process described in detail in section **3.2.3**), instead of drawing every possible edge. Moreover, it is possible to manually activate an option to show only those experimentally confirmed instances.
- DMIs inferred from 3D structures—displayed as pink edges—involve the motif classes obtained from the 3did database. Since these were identified from protein structures through an automatic pipeline (Stein & Aloy, 2010), they are generally less reliable. Because these motifs are not annotated in any way, there is not any possibility of narrowing them down. For this reason, it is recommended to toggle on only confirmed instances, which in this case correspond to those actually observed in 3D structures.

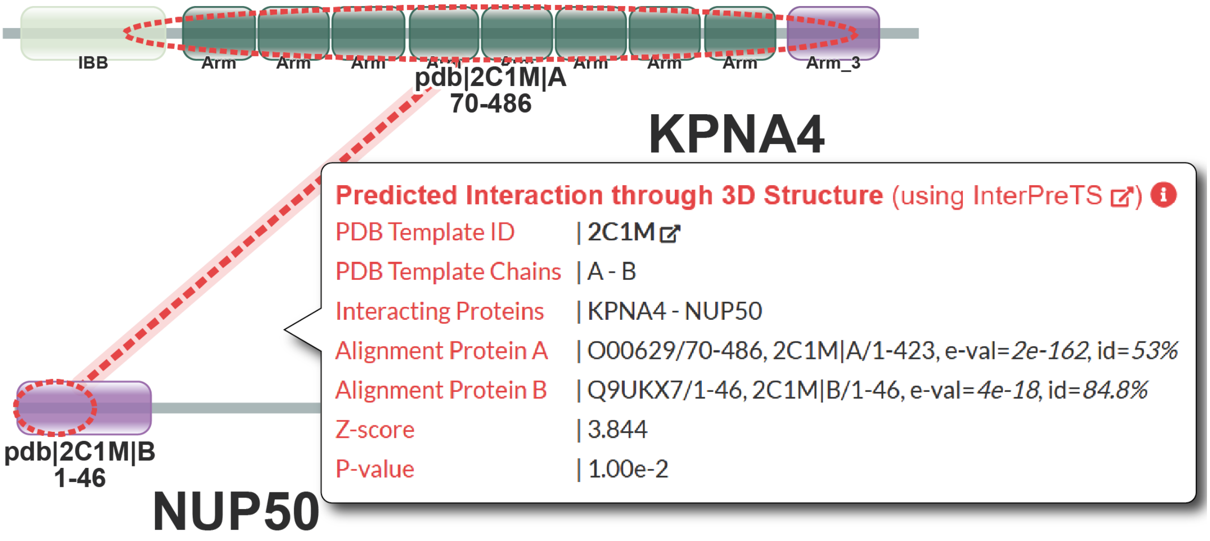


Domain-Motif Interaction (from ELM annotation)
Interacting Elements | RB_B - LIG_Rb_LxCxE_1
Interacting Proteins | RB1 - HDAC2
Interaction P-value | 2.08e-4

Figure 3.10: DMI between the LxCxE motif in Histone deacetylase 2 and the B pocket of the retinoblastoma-associated protein.

Tertiary structure-based predicted interactions

Protein interactions predicted de-novo through tertiary structure homology modelling with InterPreTS (described in **3.2.5**) can be also visualized in the network as their own interaction type. These are illustrated as red edges linking two red ellipses that indicate the corresponding predicted interfaces in each protein, and they are labelled with the PDB structure used as template (Figure **3.11**). It should be noted that these interacting protein regions do not necessarily correspond to other known protein modules (such as Pfam domains or ELM motifs), they simply indicate those regions that significantly aligned with the sequences of the interacting template chains and that have the required contacts for an interaction to take place. Consequently, InterPreTS regions cannot be toggled on independently as on their own they do not mean anything. Tooltips displayed when clicking on the region nodes or interaction edge show the alignment scores (E-value, % identity), coordinates of the protein region–3D structure chain alignment, and scores that indicate the strength of the prediction (*P*-value and Z-score).



Predicted Interaction through 3D Structure (using InterPreTS)

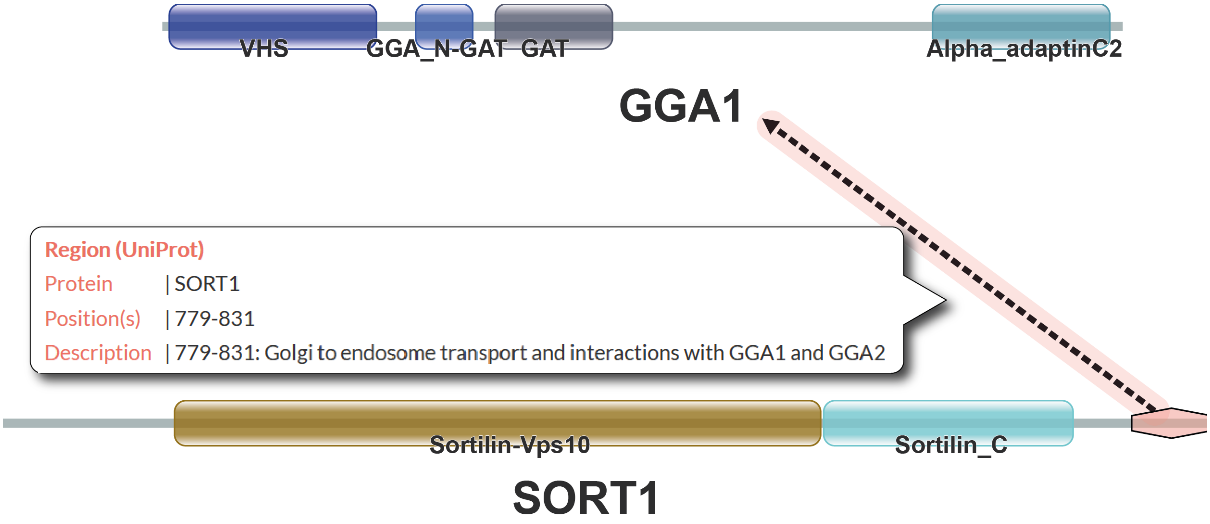
PDB Template ID | 2C1M
PDB Template Chains | A - B
Interacting Proteins | KPNA4 - NUP50
Alignment Protein A | O00629/70-486, 2C1M|A/1-423, e-val=*2e-162*, id=*53%*
Alignment Protein B | Q9UKX7/1-46, 2C1M|B/1-46, e-val=*4e-18*, id=*84.8%*
Z-score | 3.844
P-value | 1.00e-2

Figure 3.11: Mechnetor visualization of the interaction and interfaces between KPNA4 and NUP50 as predicted by InterPreTS (Aloy & Russell, 2003), using PDB structure:2C1M as template.

Interactions inferred from sequence annotations.

UniProtKB does not systematically provide information on interfaces between proteins, in the sense that there is not a single field in their protein entries following a particular format. However, this type of information can be sometimes found in certain sequence features. For example, many UniProtKB regions of interest are annotated as mediating interactions with other proteins, and sites affected by mutagenesis experiments are often annotated to impair certain protein interactions. Mechnetor automatically extracts these connections and draws edges linking the particular protein region or site to the referred protein. The functional nature of this connection is specified in both the region label and the interaction tooltip

(Figure **3.12**). As opposed to all other types of edges, these appear automatically when the corresponding sequence feature is toggled on.



Region (UniProt)

Protein	SORT1
Position(s)	779-831
Description	779-831: Golgi to endosome transport and interactions with GGA1 and GGA2

Figure 3.12: A UniProt region of interest in protein SORT1 is annotated as being involved in the interactions with GGA1 and 2 (apart from Golgi to endosome transport). Since protein GGA1 is also present in the current network, Mechnetor automatically draws an arrow linking both.

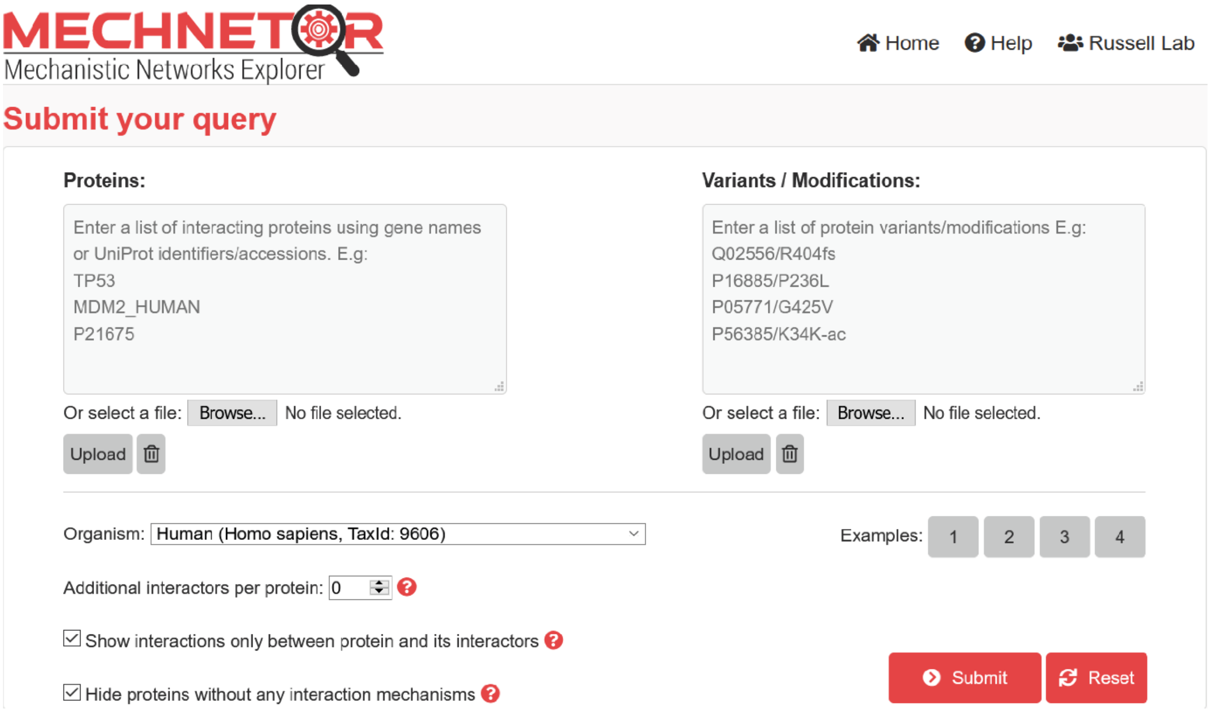
3.3.4 Usage

3.3.4.1 Input

Two types of input data that can be submitted to Mechnetor: proteins and protein variants/modifications. Proteins will define the elements of the resulting network while variants/modifications will be mapped into the corresponding positions within those proteins. Both can be typed directly into input boxes or uploaded as text files, following the same formats. The user also has to select the corresponding organism (Figure **3.13**).

- **Protein input box.** Proteins can be specified by their UniProtKB identifier (e.g. M3K20_HUMAN), UniProtKB accession (e.g. Q9NYL2) or gene symbol (e.g. MAP3K20). Users can input a list of proteins (one protein per line) or a list of protein pairs (two proteins separated by whitespace per line). In the first case, Mechnetor will search for interactions between all input proteins, while in the second, it will only search for interactions between the specified pairs. It is also possible to mix both formats, in which case, proteins from specified pairs will be kept separately but individual proteins will be searched against all the others.

- **Variant/Modification input box.** Users can optionally submit protein variants or PTMs. They have to be introduced one per line following a particular format: the protein, followed by a forward slash, followed by the residue modification, which itself has to contain the original residue, its position and the modified residue, in that order (e.g. the protein variant Q9NYL2/F368C, or the phosphorylation site MAP3K20/S599S-p). Proteins with modifications are automatically added to the network even if they were not included in the protein input box, meaning you do not need to enter them twice.



Proteins:

Enter a list of interacting proteins using gene names or UniProt identifiers/accessions. E.g:
TP53
MDM2_HUMAN
P21675

Or select a file: No file selected.

Variants / Modifications:

Enter a list of protein variants/modifications E.g:
Q02556/R404fs
P16885/P236L
P05771/G425V
P56385/K34K-ac

Or select a file: No file selected.

Organism: Human (Homo sapiens, TaxId: 9606)

Additional interactors per protein: 0

☑ Show interactions only between protein and its interactors

☑ Hide proteins without any interaction mechanisms

Figure 3.13: Mechnetor web-server index page. Users can query lists of proteins, protein pairs and/or protein variants and modifications, for any of 8 model organisms, including human. Additional options allow to configure the final network as well as automatically import known interactors for the user's proteins.

A few additional options allow to further customize user queries:

- **Additional interactors**. This allows to specify a number of known interactors for each protein in the user input that will be automatically imported to the query (none by default). These interactors are extracted from the PPI database, from highest to lowest number of experiments supporting them. Thanks to this feature, users can input proteins without specifying a set of interactions—or even input just a single protein—and study them in the context of their best known interactome.

- **Only interactions between input proteins and known interactors**. When the previous feature is used, automatically imported interactors will by default be checked for interactions against every other protein in the network (on by default). This accessory option can be activated to avoid that, forcing the final network to contain only interactions between the known interacting protein pairs. This is useful if one does not care how these additional interactors interact with each other but only with their input protein of interest, and it can save computational time and result in a simpler network.

- **Hide unconnected proteins**. This option (on default) makes proteins in the resulting network to be hidden if no mechanistic connection (any interaction evidence that is not a general PPI) to any other protein could be found. Hidden proteins can be later toggled on manually.

- **Examples**. To illustrate the input formats as well as the different interaction types Mechnetor supports, four examples are available and can be loaded by simply clicking in the corresponding buttons.

There is no restriction on how many proteins the user can submit, but very large numbers can result in longer job computation times. Besides, Mechnetor is not a tool suitable to visualize large protein networks because those would be too convoluted and one would not be able to see all details clearly. For this reason, there is a limit of 20 proteins to be contained in a network at most. If the user input contains a larger number of proteins, Mechnetor will give priority to those for which mechanistic information is found when building the network, however, the table below the network will still contain everything.

3.3.4.2 Interactivity options in the network

Upon job completion, results will be presented in a network where users can utilize several interactivity options to explore the data and customize the view (Figure **3.14**). These include:

- **Mouseover effects**. To enhance user-experience, all network nodes and edges display some kind of visual change (in size, colour or style) when the mouse is over them. Protein region labels will change to show the corresponding sequence coordinates, while labels of single position features (PTMs, variants) are shown on mouseover. In the case of edges, hovering over them will also highlight the two elements involved in the interaction.

- **Tooltips**. Clicking on almost any element will display a popup box containing more information as well as direct links to original data sources. Some extra customization options are available for some nodes. For example, the main protein tooltip allows you to switch the displayed protein name to either the gene symbol, the UniProtKB accession or the

UniProtKB identifier. For interactions edges, these tooltips specify the proteins and protein elements involved and the evidence or scores supporting the interactions.

- **Toggles**. The sidebar checkboxes allow the user to toggle on or off all different types of protein features and interactions individually. In addition, some offer options to filter the displayed elements. For example, it is possible to toggle on only experimentally verified instances of SLiMs, or to use sliders to set a minimum required score for predicted domain-domain interactions, as well as to adjust the *P*-value cut-off for all interactions.

3.3.4.3 Table of interactions

The table located below the network will always contain all types of interactions found for the input proteins independently of what the network is showing (Figure **3.14**). Each row specifies the pair of interacting proteins, the interaction type, the interacting protein regions (if applies), as well as those user-input variants that are located within those regions. The table can be searched for any term and/or sorted by the values of any column.

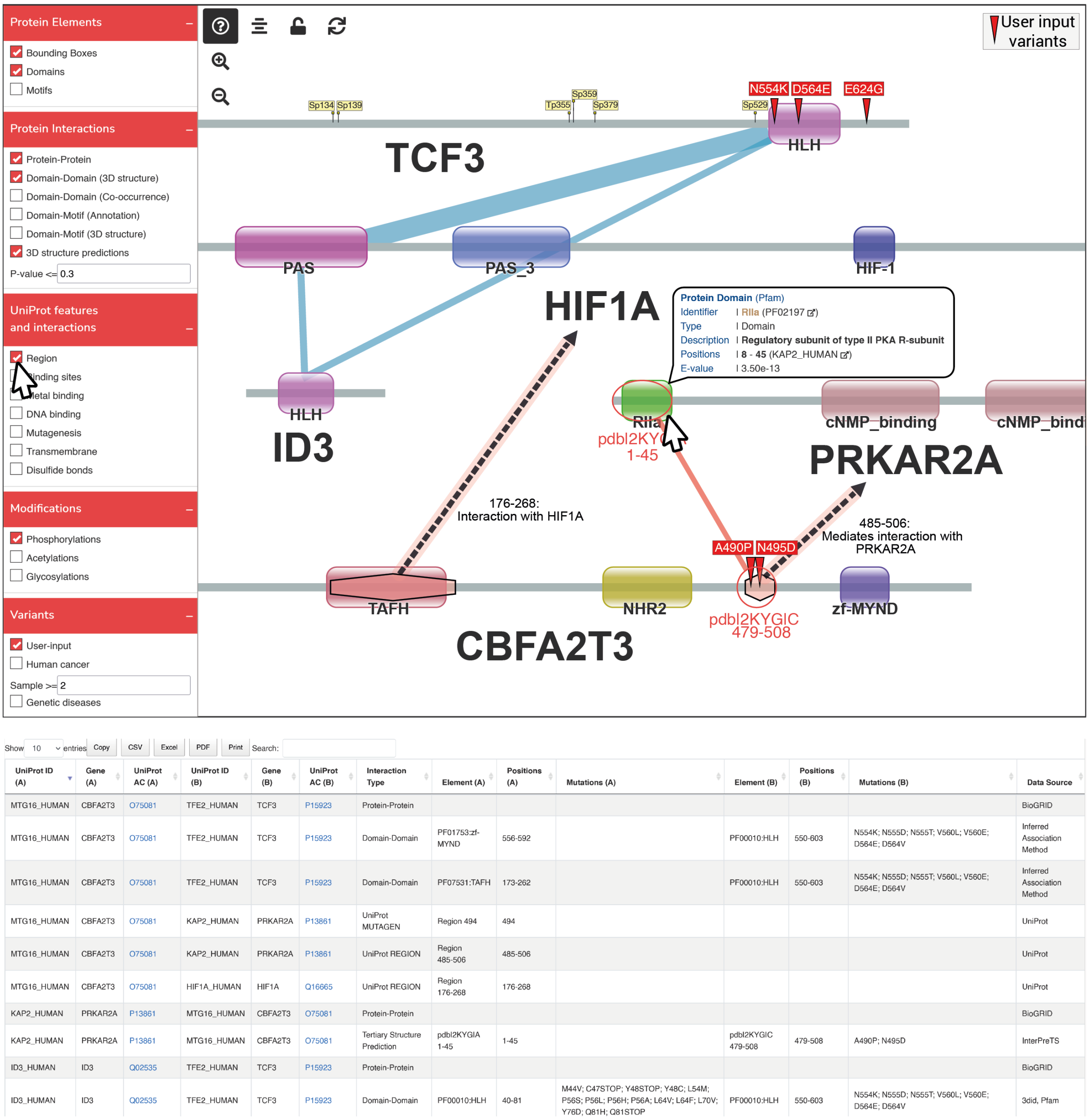
3.3.4.4 Export files

Mechnetor can export a snapshot of the network view at any time as an image file (JPG or PNG), or export the full network as a vector graphic (SVG) which can be edited with any vector image processing software to prepare publication-quality figures. The table can also be exported in several formats (CSV, Excel or PDF).

3.3.5 Technical specifications

Mechnetor is implemented as a web server using Python (Python 3.6.8) and the Flask micro web framework[11]. The InterPreTS tool was fully rewritten in Python so that it could be integrated into the pipeline. The interactive graph component is rendered using cytoscape.js[12], an open-source JavaScript-based graph library that offers a wide range of visual and performance features for creating highly customizable and interactive networks that can be easily integrated into web interfaces (Franz et al., 2015). The biggest challenge in cytoscape.js implementation was finding a way to create networks where proteins, traditionally represented by single nodes of simple shape, could instead be represented as linear arrangements of their diverse functional elements (domains, motifs, PTMs and other

[11] https://flask.palletsprojects.com
[12] https://js.cytoscape.org



Protein Elements –
☑ Bounding Boxes
☑ Domains
☐ Motifs

Protein Interactions –
☑ Protein-Protein
☑ Domain-Domain (3D structure)
☐ Domain-Domain (Co-occurrence)
☐ Domain-Motif (Annotation)
☐ Domain-Motif (3D structure)
☑ 3D structure predictions
P-value <= 0.3

UniProt features and interactions –
☑ Region
☐ Binding sites
☐ Metal binding
☐ DNA binding
☐ Mutagenesis
☐ Transmembrane
☐ Disulfide bonds

Modifications –
☑ Phosphorylations
☐ Acetylations
☐ Glycosylations

Variants –
☑ User-input
☐ Human cancer
Sample >= 2
☐ Genetic diseases

Protein Domain (Pfam)
Identifier | RIIa (PF02197)
Type | Domain
Description | Regulatory subunit of type II PKA R-subunit
Positions | 8 - 45 (KAP2_HUMAN)
E-value | 3.50e-13

Show 10 entries Copy CSV Excel PDF Print Search:

UniProt ID (A)	Gene (A)	UniProt AC (A)	UniProt ID (B)	Gene (B)	UniProt AC (B)	Interaction Type	Element (A)	Positions (A)	Mutations (A)	Element (B)	Positions (B)	Mutations (B)	Data Source
MTG16_HUMAN	CBFA2T3	O75081	TFE2_HUMAN	TCF3	P15923	Protein-Protein							BioGRID
MTG16_HUMAN	CBFA2T3	O75081	TFE2_HUMAN	TCF3	P15923	Domain-Domain	PF01753:zf-MYND	556-592		PF00010:HLH	550-603	N554K; N555D; N555T; V560L; V560E; D564E; D564V	Inferred Association Method
MTG16_HUMAN	CBFA2T3	O75081	TFE2_HUMAN	TCF3	P15923	Domain-Domain	PF07531:TAFH	173-262		PF00010:HLH	550-603	N554K; N555D; N555T; V560L; V560E; D564E; D564V	Inferred Association Method
MTG16_HUMAN	CBFA2T3	O75081	KAP2_HUMAN	PRKAR2A	P13861	UniProt MUTAGEN	Region 494	494					UniProt
MTG16_HUMAN	CBFA2T3	O75081	KAP2_HUMAN	PRKAR2A	P13861	UniProt REGION	Region 485-506	485-506					UniProt
MTG16_HUMAN	CBFA2T3	O75081	HIF1A_HUMAN	HIF1A	Q16665	UniProt REGION	Region 176-268	176-268					UniProt
KAP2_HUMAN	PRKAR2A	P13861	MTG16_HUMAN	CBFA2T3	O75081	Protein-Protein							BioGRID
KAP2_HUMAN	PRKAR2A	P13861	MTG16_HUMAN	CBFA2T3	O75081	Tertiary Structure Prediction	pdbl2KYGIA 1-45	1-45		pdbl2KYGIC 479-508	479-508	A490P; N495D	InterPreTS
ID3_HUMAN	ID3	Q02535	TFE2_HUMAN	TCF3	P15923	Protein-Protein							BioGRID
ID3_HUMAN	ID3	Q02535	TFE2_HUMAN	TCF3	P15923	Domain-Domain	PF00010:HLH	40-81	M44V; C47STOP; Y48STOP; Y48C; L54M; P56S; P56L; P56H; P56A; L64V; L64F; L70V; Y76D; Q81H; Q81STOP	PF00010:HLH	550-603	N554K; N555D; N555T; V560L; V560E; D564E; D564V	3did, Pfam

Figure 3.14: Mechnetor results are primarily presented in a mechanistic network where users can explore the data using different interactivity options. In addition, a table below lists all types of interactions found in the network, including all user-input variants located in the relevant interacting elements.

sequence features) allowing for the possibility of drawing interactions between those elements independently; which is not natively supported by cytoscape.js. The solution was to make proteins a composition of a thin rectangular node proportional to protein length representing its sequence, plus independent, visually distinct nodes for each protein element that can be assigned to it. The latter are superimposed over the protein sequence node by giving them the same y axis coordinates, but placed in their corresponding positions by varying their x axis values. All nodes belonging to the same protein are attached to a common invisible parent node that makes possible that the protein can be dragged as whole without disrupting its internal node layout.

Mechnetor's source code and the code required to build its database is freely available[13].

3.4 Case studies

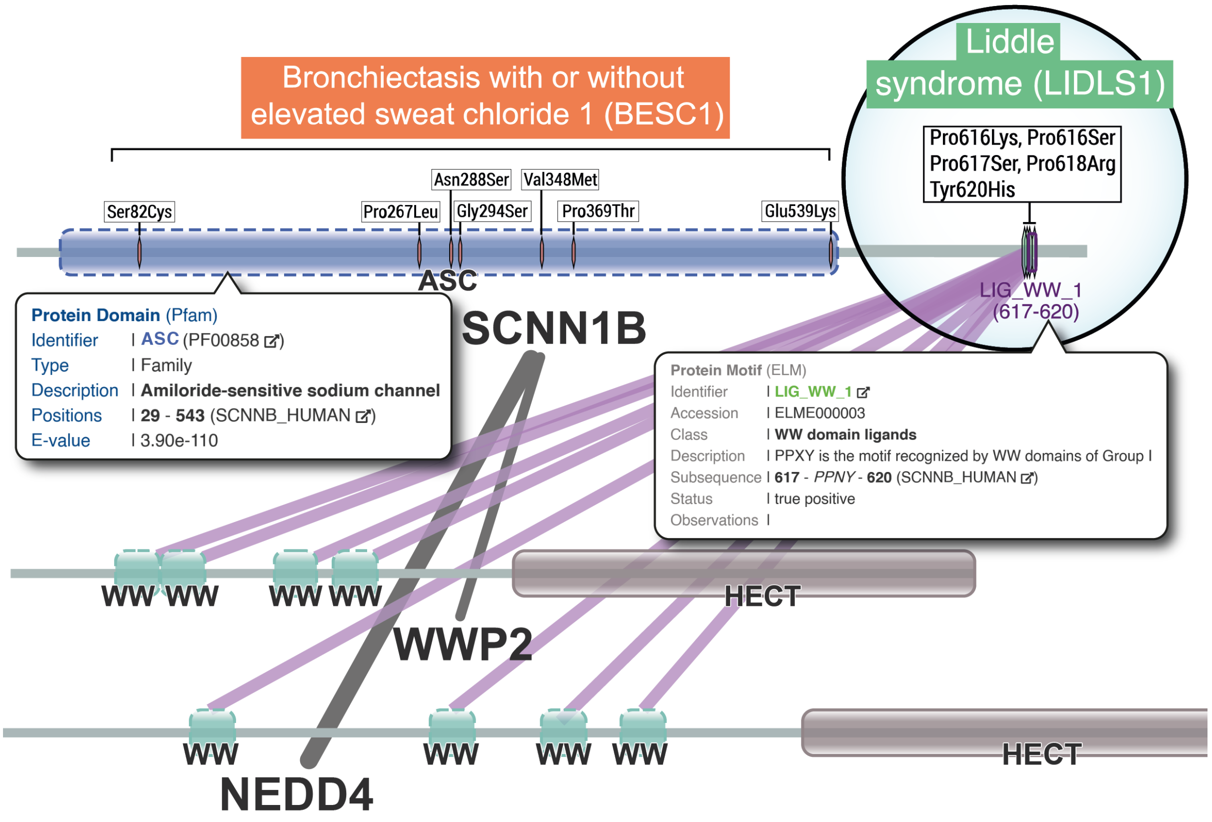
To find cases where I could showcase usability, I looked for instances where mechanistic differences highlighted by Mechnetor correspond to different pathologies. Next, I describe in detail two particular cases together with the view that can be obtained from Mechnetor.

3.4.1 β subunit of the heterotrimeric epithelial sodium channel (SCNN1B)

The epithelial sodium channel (ENaC, or also amiloride-sensitive sodium channel) mediates the first step of active sodium reabsorption across the apical membranes of tight or high resistance epithelial cells, in particular in the distal nephron of kidneys. This channel plays a key role in maintaining electrolyte and water homeostasis, and regulating extracellular volume and blood pressure. ENaC is an heterotrimer composed of homologous subunits α (or δ), β and γ (Garty, 1994; Hanukoglu & Hanukoglu, 2016). β subunit (or SCNN1B) is annotated (UniProtKB: P51168) with disease-causing variants related to two different genetic diseases: bronchiectasis with or without elevated sweat chloride 1 (BESC1), characterized by an abnormal persistent dilatation of the bronchi, excess mucus build-up and other symptoms; and Liddle syndrome 1 (LIDLS1), an autosomal dominant disorder that causes severe hypertension (Shimkets et al., 1994). With Mechnetor it is possible to infer the mechanisms through which these variants affect protein function (Figure **3.15**). We can observe that BESC1 variants are distributed along the large conserved protein region that corresponds to the actual amiloride-sensitive sodium channel (ASC; Pfam accession: PF00858), which suggests that these variants are more likely to be deleterious and result in decreased channel activity (Fajac et al., 2008). In contrast, LIDLS1-causing variants are located in a 4-residue long region at the C-terminus (residues 616-620) where they clearly disrupt a WW domain binding motif (ELM identifier*: LIG_WW_1*). This motif is recognized by

[13] https://github.com/JCGonzS/mechnetor

the WW domains of E3 ubiquitin ligases such as WWP2 and NEDD4, which are experimentally verified interactors of SCNN1B. It seems reasonable to deduct that LIDLS1 variants impair these interactions and thus inhibit ubiquitination and subsequent degradation of the ENaC. A constitutively active channel ultimately leads to an increase of blood volume and pressure (Abriel et al., 1999; Furuhashi et al., 2005).



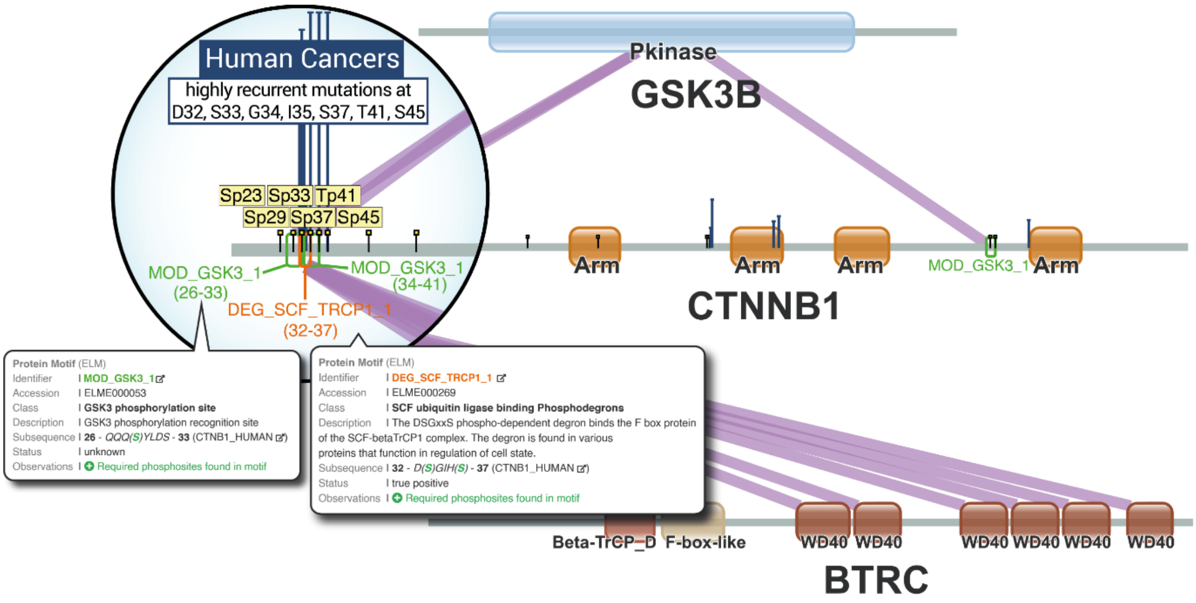
Protein Domain (Pfam)
Identifier | ASC (PF00858)
Type | Family
Description | Amiloride-sensitive sodium channel
Positions | 29 - 543 (SCNNB_HUMAN)
E-value | 3.90e-110

Protein Motif (ELM)
Identifier | LIG_WW_1
Accession | ELME000003
Class | WW domain ligands
Description | PPXY is the motif recognized by WW domains of Group I
Subsequence | 617 - *PPNY* - 620 (SCNNB_HUMAN)
Status | true positive
Observations |

Figure 3.15: Mechnetor view of the epithelial sodium channel β subunit (SCNN1B) and E3 ubiquitin ligases WWP2 and NEDD4. Experimentally verified interactions between two proteins are represented by grey lines, while domain-linear motif interactions are shown as purple lines. Positions corresponding to disease variants are indicated along SCNN1B (Bronchiectasis variants in orange; Liddle Syndrome variants in green). The C-terminal region of SCNN1B has been amplified to enhance visualization of the overlap of Liddle syndrome variants and the *LIG_WW_1* motif.

3.4.2 Catenin beta-1 (CTNNB1)

Mechnetor can also help to understand mechanism of somatic cancer variants, as demonstrated in Figure **3.16** with catenin beta-1 (CTNNB1), a well-established oncogene with roles in cell adhesion regulation, and gene transcription as part of the Wnt signalling pathway (UniProtKB: P35222) (MacDonald et al., 2009). Here, by loading cancer missense variants with a sample count of at least 5, a hotspot of highly recurrent mutations becomes clearly apparent at the N-terminal region of CTNNB1, targeting several GSK3B (UniProtKB: P49841) phosphorylation sites (S33, S37, T41 and S45), as indicated by the overlap with the kinase GSK3B recognition motifs (ELM identifier: *MOD_GSK3_1*). Due to the simplicity of *MOD_GSK3* motif pattern (...[ST]...[ST]), several matches are found on the CTNNB1 sequence, but the overlap with experimentally-proved phosphosites can be used as an indicative of functional motif instances. Furthermore, another binding motif overlaps with this region: a diphospho-dependent degron (ELM identifier: *DEG_SCF_TRCP1_1*), which is required for CTNNB1 recognition by the WD40 β-propeller of the E3 ubiquitin-protein ligase complex component BTRC (UniProtKB: Q9Y297). Since BTRC binding to CTNNB1 is dependent on these phosphosites, by inhibiting N-terminal phosphorylation, these somatic mutations ultimately prevent ubiquitination and degradation of CTNNB1 by the proteasome. The accumulation of CTNNB1 is the same effect as the activation of the Wnt signalling pathway, and results in unrestricted transcription of its target genes (Shang et al., 2017).



Protein Motif (ELM)
Identifier | **MOD_GSK3_1**
Accession | ELME000053
Class | **GSK3 phosphorylation site**
Description | GSK3 phosphorylation recognition site
Subsequence | **26** - *QQQ(S)YLDS* - **33** (CTNB1_HUMAN)
Status | unknown
Observations | ⊕ Required phosphosites found in motif

Protein Motif (ELM)
Identifier | **DEG_SCF_TRCP1_1**
Accession | ELME000269
Class | **SCF ubiquitin ligase binding Phosphodegrons**
Description | The DSGxxS phospho-dependent degron binds the F box protein of the SCF-betaTrCP1 complex. The degron is found in various proteins that function in regulation of cell state.
Subsequence | **32** - *D(S)GIH(S)* - **37** (CTNB1_HUMAN)
Status | true positive
Observations | ⊕ Required phosphosites found in motif

Figure 3.16: Mechnetor view of catenin beta-1 (CTNNB1), the kinase GSK3B and E3 ubiquitin-protein ligase complex component BTRC, showing their domain composition. Domain-linear motif interactions are shown as purple lines. Displayed along CTNNB1 sequence: phosphosites (small yellow flags), cancer missense variants (present in at least 5 samples; blue T-shaped flags of height proportional to number of samples), the GSK3 recognition motifs (*MOD_GSK3*, green empty box) and the BTRC-binding phospho-dependent degron (*DEG_SCF_TRCP1_1*; orange empty box). Popup boxes show more detailed annotations and let us know that the required phosphosites are found within these motifs. Zoomed CTNNB1 N-terminal region better shows the overlap between these element

3.5 Conclusions

It is a challenge to make sense of the wealth of variant data that is increasingly available. There is thus a growing need for tools that facilitate the integration of different datasets and the extraction of meaningful information in order to answer particular biological questions. This is becoming more relevant as we are moving into the era of precision medicine where, to be able to tailor patient-specific medical treatments, understanding individual variability at the molecular level will be crucial. Mechnetor was created with this purpose in mind.

Mechnetor is a tool that considers multiple existing protein and mechanistic data and presents them where they might be applicable. Although formally it does not contain any trained prediction algorithm, the information gathered by this tool can often predict interactions between proteins not known to interact and/or unveil novel mechanistic details. In particular, domain-domain and domain-motif interaction data are extracted from relatively small sets of verified instances. Mechnetor is by default configured to infer mechanism where it makes sense even though the protein pair being studied is not present in any of the original datasets. For this reason, it is very important to provide results together with tools so that users can explore and understand the information provided and ultimately arrive at their own conclusions.

Consequently, one of the priorities was to make Mechnetor very user-friendly so, in essence, it can be used as simply as inputting a pair of proteins, clicking the submit button and, in just a few seconds, visualizing different interaction evidence between them. Further possibilities include studying larger datasets of interacting protein pairs by directly downloading the integrated data for local analysis, mapping custom protein variants into proteins to investigate them in a mechanistic context, or representing and visualizing custom protein interactions and interfaces.

In addition, there is the possibility of adapting Mechnetor into an individual component that can then be easily implemented into other tools and projects. For example, the Mechnetor framework was used to interrogate MS-based cross-linking information[14], allowing us to make mechanistic suggestions about what particular cross-links could be doing and thus aiding in deciphering structures of large macromolecular assemblies.

[14] xlinterpreter.russelllab.org

Chapter 4

Analysis of *S*-palmitoylation in *Drosophila melanogaster*

4.1 Introduction

S-acylation, commonly referred to as *S*-palmitoylation, is a post-translational modification of proteins that consists in the attachment of palmitic acid—a 16-carbon saturated fatty acid—to a cysteine via thioester linkage and, unlike all other protein lipid modifications, it is fully reversible (Chamberlain & Shipston, 2015; Schmidt & Schlesinger, 1979) (Figure **4.1**). Although other fatty acids can be attached via *S*-acylation (stearate or oleate), palmitate is the most commonly found in endogenous *S*-acylated proteins, and thus both terms are generally used synonymously (Muszbek et al., 1999).

Similar to other lipid modifications, the main effect of *S*-palmitoylation is to increase protein hydrophobicity and thus defining roles in membrane targeting and trafficking: the attached palmitic acid acts as a membrane anchor of cytosolic proteins that normally lack transmembrane domains, and promotes the distribution of membrane proteins to different subcellular compartments (Chamberlain & Shipston, 2015). Since *S*-palmitoylation is a reversible process, cycles of palmitoylation/depalmitoylation provide a dynamic regulation of localization and function to proteins in a wide range of cell types and tissues. It is thus implicated in the control of many cellular processes including GPCR signalling (Jia et al., 2014), trafficking of membrane proteins from early secretory pathways to the plasma membrane (Smotrys & Linder, 2004) or synaptic plasticity (Fukata and Fukata 2010). Palmitoylation has been also shown to regulate the stability of integral membrane proteins by impeding their ubiquitination and subsequent degradation (Valdez-Taubas & Pelham, 2005). Particular examples of dynamically regulated proteins and processes include trafficking of mammalian H-Ras and N-Ras between Golgi and plasma membrane to modulate Ras signalling (Goodwin et al., 2005); and lateral distribution of proteins on the plasma membrane to lipid rafts, such as with the PKA anchoring protein AKAP79 (Delint-Ramirez et al., 2011).

Although some proteins have been shown to autopalmitoylate spontaneously (Chan et al., 2016), for the majority, the palmitoylation cycle is mediated by the action of two kinds of enzymes (Figure **4.1**). Protein *S*-palmitoylation is catalysed by palmitoyl acyl-transferases (PATs), which are integral membrane proteins harbouring a 50 residue-long cysteine-rich domain that itself contains a conserved Asp-His-His-Cys (DHHC) motif, thus giving them the common name DHHC PATs (Mitchell et al., 2006). The opposite process—enzymatic removal of S-acyl modifications—is catalysed by thioesterases (Hunt & Alexson, 2002).

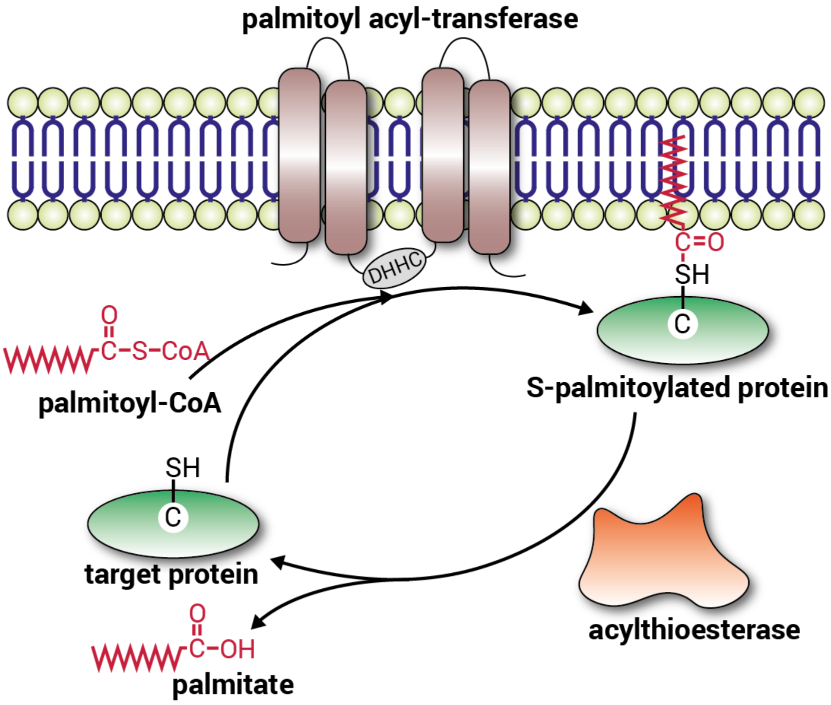


Figure 4.1: Palmitoylation & depalmitoylation cycle mediated by palmitoyl acyltransferases & acylthioesterases.

In mammalian genomes, more than 20 different PATs, which distinctly reside in different cell membranes, have been identified or predicted (Chamberlain & Shipston, 2015; Fukata et al., 2004; Gottlieb & Linder, 2017). In contrast, the number of thioesterases is much lower: mammals typically contain two lysosomal palmitoyl-protein thioesterases (PPT1 and PPT2), in charge of removing palmitate during lysosomal degradation (Camp & Hofmann, 1993; Soyombo & Hofmann, 1997); and two cytosolic serine hydrolases, the acyl-protein thioesterases (APT1 and APT2) (Duncan & Gilman, 1998; Tomatis et al., 2010), which are responsible for depalmitoylation of a wider range of substrates directly on membrane surfaces. More recently, the members of the α/β-hydrolase domain-containing protein 17 family (ABDH17A, ABDH17B and ABDH17C) were established as novel depalmitoylation enzymes, acting on substrates like NRAS (Lin & Conibear, 2015). This imbalance in the number of both type of enzymes is also seen in *Drosophila* where 22 DHHC PATs (many of which show several isoforms), and only three thioesterases (Ppt1, Ppt2 and Apt1) have been identified to date (Bannan et al., 2008).

The importance of this modification has grown in the last years especially because alterations in the palmitoylation cycle have been linked to several diseases, including nervous system disorders (e.g. Huntington's disease) (Sanders et al., 2015) and cancer (Ko & Dixon, 2018). Consequently, this process is well characterized in mammalian cell lines and tissues with both target proteins and associated enzymes largely identified (Blanc et al., 2015; Sanders et al., 2015). In contrast, little is known about *S*-palmitoylation in invertebrates. The first systematic identification of *S*-palmitoylated proteins (i.e. the first palmitoylome) for an invertebrate was performed in *Caenorhabditis elegans* (Edmonds & Morgan, 2014). In *Drosophila melanogaster*, the first palmitoylome was published recently (Strassburger et al., 2019). Whereas the SwissPalm database—the largest compendium of *S*-palmitoylated proteins—comprises around 3700 genes, mainly from the aggregation of human, mouse and rat experimental studies, there are fewer than 200 palmitoylated proteins from invertebrates (Blanc et al., 2015). Moreover, just a handful of DHHC PATs–substrate interactions in *Drosophila* are known. These include the *Drosophila* Huntingtin-interacting protein 14 (dHip14), ortholog of the mammalian DHHC PAT ZDHHC17/HIP14, which interacts with SNAP25, cysteine string protein (CSP) and the short gastrulation (Sog) protein (Kang & Bier, 2010; Ohyama et al., 2007; Stowers & Isacoff, 2007). The DHHC family member approximated (app), ortholog of human ZDHHC14, was shown to have both DHHC PAT-dependent and independent functions. app binds to and localizes Dachs to the apical junctional region of imaginal discs and palmitoylates the large protocatherine Fat, resulting in repression of Fat function and promoting tissue growth (Matakatsu & Blair, 2008). In more recent work, Strassburguer et al. published the first palmitoylome, and identified 13 of those proteins as potential targets of dZDHHC8. Among them, Ras64B stability was found to be strongly dependent on palmitoylation by this enzyme (Strassburger et al., 2019).

Besides this handful of proteins, knowledge of *S*-palmitoylation targets in *Drosophila* is acutely lacking. Here we set out to provide a new and more comprehensive catalogue of *S*-palmitoylated proteins in this model organism. For this, our collaborators experimentally purified and identified *S*-palmitoyl-proteins from the transmembrane fraction of S2R+ cells (*Drosophila* embryonic cell line). They also developed a novel BioID-based method allowing the identification of potential and specific interaction partners of 10 selected *Drosophila* DHHC PATs. I was responsible for the subsequent systematic analysis to validate and integrate the results, provide functional annotations and assess the level of conservation with mammalian palmitoylomes. This work resulted in the second and most complete palmitoylome in *Drosophila melanogaster*, which when coupled to the experimentally-determined DHHC PAT interaction profiles allows the identification of their potential client proteins. We provide novel insights into the scope and mechanisms of this important post-translational modification. This work expands the understanding of *S*-palmitoylation in invertebrates and, in some instances, provides insights into mammalian orthologs.

4.2 Results and Discussion

4.2.1 The *Drosophila melanogaster* palmitoylome

4.2.1.1 Identification of *S*-palmitoylated proteins

Our collaborators used the acyl-resin assisted capture (acyl-RAC) assay (Forrester et al., 2011) to purify *S*-palmitoylated proteins from the membrane fraction of S2R+ (S2 receptor plus) *Drosophila melanogaster* embryonic cells (Yanagawa et al., 1998), coupled with LC-MS/MS for their subsequent identification. Acyl-RAC is an alternative to the widely used acyl-biotin exchange (ABE) assay, a biochemical technique for capturing and identifying *S*-acylated proteins (Drisdel & Green, 2004), with the advantage that it replaces the biotinylation step of ABE—detection of biotinylated proteins requires complex and expensive methods—with direct conjugation of free cysteines by a thiol-reactive resin (Figure **4.2**). Thus, acyl-RAC is fast, has fewer steps, and can be also applied to a wider range of samples.

From the results of this experiment, 1188 proteins were initially identified which were then filtered to keep those with a fold change equal to or above 2 (FC>=2), a false discovery rate below 0.1 (FDR<0.1), and that did not lack cysteine residues. In addition, six proteins were excluded for being likely false positives as they were enzymes with thioester bonds known to not play a role in protein lipidation. The remaining 198 proteins were deemed *S*-palmitoylated. Of these, 51 were further classified as high confidence (HC) due to a high enrichment (FC>=20), with the others being referred to as normal confidence (Figure **4.3A**).

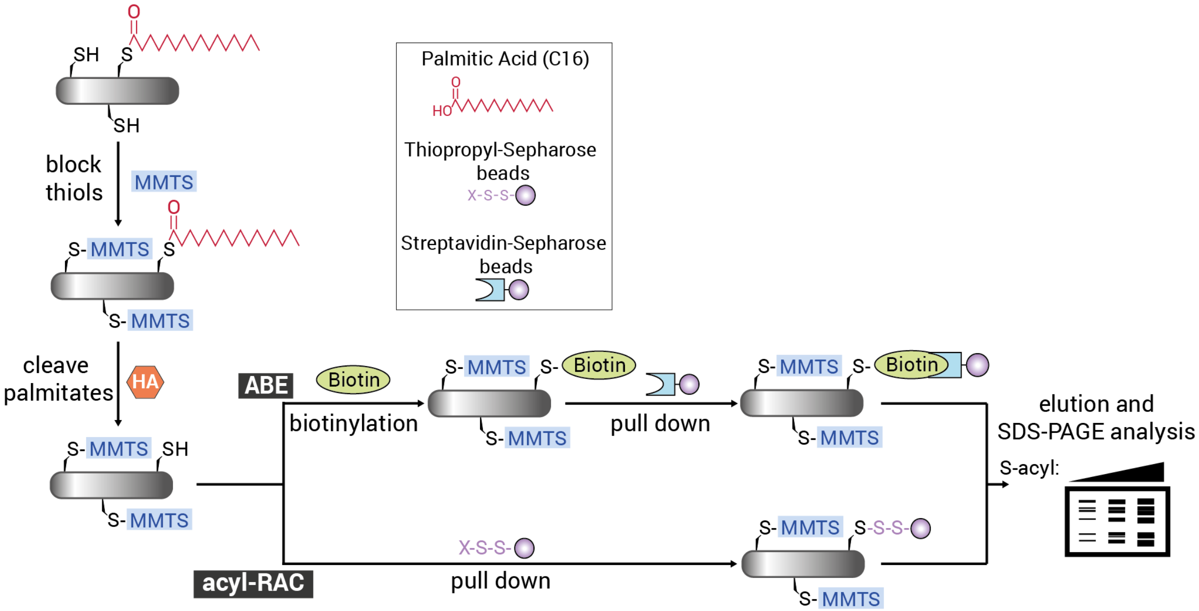


Figure 4.2: Overview of the acyl-resin assisted capture acyl-RAC and acyl-biotin exchange (ABE) assays. First, free thiols groups are blocked with methyl methanethiosulfate (MMTS). Next, thioester-linked S-palmitic fatty acids are cleaved using neutral hydroxylamine (NH_2OH). In ABE, newly freed thiols are first biotinylated and then captured with streptavidin-sepharose beads. In acyl-RAC, these are directly captured with thiopropyl-sepharose beads. After pull-down assay, captured proteins are eluted with reductant and subsequently analysed by SDS-PAGE with either protein staining or immunoblotting.

A cursory glance already finds proteins among these that are well-known to be palmitoylated in mammals: Snap24, an Snap25 homologue; the cysteine-string protein (Csp), a chaperone of the DnaJ family that functions in regulated exocytosis in synaptic vesicles (Greaves et al., 2008); and Flotillin-1/-2, membrane-associated scaffolding proteins involved in endocytosis, cell signalling and protein trafficking (Morrow et al., 2002). For all these, palmitoylation is known to play key roles in regulating function via increasing membrane association. The following sections go over different approaches I used in order to try to validate the whole putative *Drosophila* S2R+ palmitoylome.

4.2.1.2 The *Drosophila* palmitoylome differs between whole larvae and S2R+ cells

Until publishing these results, the only list of *S*-palmitoylated proteins in *Drosophila* had been obtained from instar L2 larvae and comprised 159 proteins (Strassburger et al., 2019). This is roughly 25% smaller than our embryonic S2R+ cell palmitoylome, which is surprising as larvae are more complex, and thus expected to have a wider range of proteins subjected to palmitoylation. Most likely, this difference is due to the fact that Strassburger *et al.* used the

ABE assay instead of acyl-RAC, where more proteins can be lost in the required extra steps. Embryonic and larval palmitoylomes were found to have 61 proteins in common (Figure **4.3B**), which represents 30% and 38% of each total respectively, and thus are the set of *S*-palmitoylated proteins in *Drosophila* with the highest support. We can only speculate whether the larger non-overlapping palmitoylome fractions are indeed due to experimental

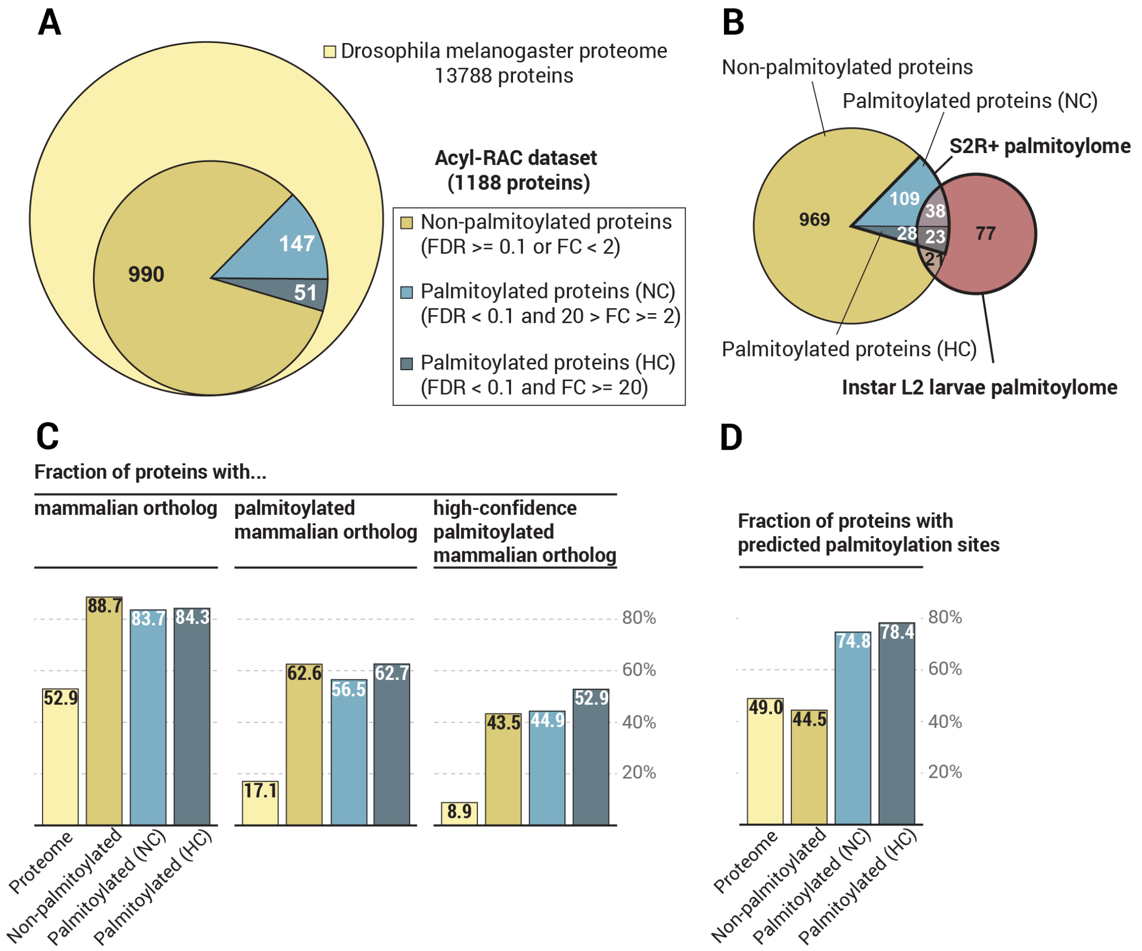


Figure 4.3: Identification and validation of *Drosophila* S2R+ cell palmitoylome. **(A)** The acyl-RAC assay identified 1188 proteins from the membrane fraction of S2R+ cells. Proteins fulfilling FDR and FC cut-offs are defined as putative palmitoylated proteins with either normal (NC) or high confidence (HC). **(B)** Overlap between S2R+ and larval palmitoylome from Strassburger et al. 2019. **(C)** Barplots indicate the fraction (%) of proteins that have: a mammalian ortholog (left), a mammalian ortholog that is palmitoylated (centre), and a mammalian ortholog that is palmitoylated with “high” confidence (right); in each of the protein sets defined above. The palmitoylation status of the mammalian orthologs was obtained from the SwissPalm database. **(D)** Barplot that shows the fraction (%) of proteins predicted to be palmitoylated by CSS-Palm in each of the protein sets.

differences, or if they really reflect a distinct palmitoylation status in different stages of development (or probably, a mix of both). The fact that only 21 proteins that were identified in the acyl-RAC experiment but deemed as not significantly *S*-palmitoylated (25% of the total 990) are actually included in the larval palmitoylome rather points to the former. Since the total number of *S*-palmitoylated proteins is unknown, it cannot also be ruled out that these two experiments are only small samples of a much larger set. This highlights the importance of obtaining more data in *Drosophila*, as it will progressively allow to confidently establish which proteins are really subject to palmitoylation and in which tissues.

4.2.1.3 *S*-palmitoylated proteins are conserved between *Drosophila* and mammals

To further validate our results and assess the degree of conservation, the *Drosophila* S2R+ palmitoylome was compared to the much better-studied and complete mammalian palmitoylome (combining those from human, mouse and rat; see Methods). As for palmitoylome size, although initial estimations suggested that at least 10% of the human proteome is susceptible to palmitoylation (Sanders et al., 2015), based on more recent data from SwissPalm, this number is close to 17.5% (see Methods). In *Drosophila*, considering that S2R+ cells were previously reported to express 5885 genes on average (Cherbas et al., 2011), their 198-protein palmitoylome only represents 3.4% of the total. However, a more realistic estimation of this number at the whole-body level can be calculated as the fraction of proteins from the full *Drosophila* proteome (data from UniProt) that can be confidently matched to a *S*-palmitoylated mammalian ortholog, which ranges between 8.9% and 17.1%, depending on the confidence of the palmitoylation status of the mammalian ortholog (Figure **4.3C**; see Methods). Although this does not necessarily imply that orthologs in *Drosophila* are also *S*-palmitoylated, the similarity of this fraction to the mammalian one seems to indicate a high degree of conservation between palmitoylomes. It also supports the idea that our experiments only identified a small subset from a larger pool of proteins that can be targeted by palmitoylation. Considering that *Drosophila* contains approximately 14,000 protein-coding genes, a conservative estimation would point at a size of roughly 1000 *S*-palmitoylated proteins.

In addition, overlap with mammalian palmitoylome could be used to further corroborate that *Drosophila* S2R+ proteins were correctly determined to be *S*-palmitoylated. Mammalian ortholog searches were done for the three subsets of proteins defined from the acyl-RAC experiment: non-palmitoylated, palmitoylated with normal confidence and palmitoylated with high confidence (Figure **4-3C**). As expected, the percentage of proteins with *S*-

palmitoylated mammalian orthologs is much higher within the S2R+ normal and high confidence palmitoylomes (56.5% and 62.7%) than in the complete proteome (17.1%). However, a significant fraction of non-palmitoylated proteins from the acyl-RAC dataset also have mammalian *S*-palmitoylated orthologs (62.6%). Further restricting the criteria to define a mammalian ortholog as *S*-palmitoylated—thus having higher confidence—these fractions are all reduced, but the difference between Drosophila palmitoylated (44.9%-52.9%) and non-palmitoylated proteins (43.5%) becomes more apparent (Figure **4-3C**).

Overall, these results support the validity of the acyl-RAC assay to recover real *S*-palmitoylated proteins. However, it also suggests that although many proteins were not found palmitoylated in S2R+ cells, they are likely real palmitoylation targets that either were missed due to the experimental conditions or they are only subjected to palmitoylation in other cell types or tissues.

4.2.1.4 Palmitoylation predictions support the *Drosophila* S2R+ palmitoylome but yield a high number of false positives

Machine learning algorithms have been developed to predict palmitoylation sites in proteins, and their accuracies have steadily improved thanks to the growing set of experimentally-determined palmitoylation sites that can be used for training them (Kumari et al., 2014; Ren et al., 2008; Zhou et al., 2006). However, since these data come almost exclusively from mammals (and despite the relatively high degree of conservation between *Drosophila* and mammalian palmitoylomes), how well these prediction methods perform in *Drosophila* is unknown. Nevertheless, using the most recent version of one such method—CSS-Palm 4.0 (Ren et al., 2008)—I predicted the number of palmitoylated proteins for the complete *Drosophila* proteome and the three protein subsets derived from the acyl-RAC assay (Figure **4-3D**, see Methods).

From the results, two conclusions are immediately apparent. The first is that, as expected, there is a very significant increase (>25%) in the fraction of predicted sites in the two subsets of experimentally determined *S*-palmitoylated proteins (with normal or high confidence) in comparison with either the complete proteome or the non-palmitoylated set. However, the second is that there is a considerably high number of proteins wrongly predicted as palmitoylated, as 49% for the whole proteome is, by all means, a strong overestimation (9%-17% was just determined to be a plausible range). To check that this high number was not due to *Drosophila*-specific artefacts, I similarly used CSS-Palm to predict the fraction of palmitoylated proteins of the complete human proteome, and this was also around 50%, which is significantly much higher than the currently estimate of 17.5%. Hence, although CSS-Palm does seem to be able to predict palmitoylated proteins, it does so with a high false positive rate regardless of the query organism. This can be explained perhaps by the fact that this algorithm is intended for predicting palmitoylation sites on known/suspected palmitoylated proteins, rather than predicting new palmitoylation target proteins. It is likely that the lack of context in these prediction methods could explain the over-prediction, which might, in the future, be solved by considering protein networks or similar biological contexts as has proved successful for phosphorylation sites (Linding et al., 2007).

4.2.1.5 Functional characterization of *Drosophila* S2R+ *S*-palmitoylated proteins

To corroborate if the putative *Drosophila* S2R+ palmitoylome is in broad functional agreement with other sets of *S*-palmitoylated proteins, I did a Gene Ontology (GO) term enrichment analysis and obtained functional terms describing their location in the cell, molecular activities and broad physiological roles (the three main GO ontologies), which were over-represented in the set of putative *S*-palmitoylated proteins (Figure **4.4A**).

In terms of cellular location, the huge enrichment in membrane proteins is itself not meaningful as only the membrane fraction was used in the acyl-RAC assay. However, there is also enrichment of particular endomembrane systems components, including the plasma membrane, the Golgi apparatus, the endoplasmic reticulum (ER), cytoplasmic vesicles, or organelle membranes, but excluding the nuclear membrane. This is in agreement with the known common compartments of *S*-palmitoylated proteins: membrane compartments in general and the ER/Golgi to cytoplasmic membrane system in particular. Although palmitoylation in the nucleus and/or nuclear membrane proteins has been reported (Fontana et al., 2019), it seems to be limited to very few targets and would likely not lead to significant enrichment.

Enriched terms from biological process and molecular function ontologies also reveal that the palmitoylome of *Drosophila* S2R+ cells mostly comprises proteins with localization and transporter activities (Figure **4.4A**). For example, these include several different SNARE proteins which mediate vesicle fusion to the target membrane in vesicular transport from ER to Golgi (e.g. Bet1), retrograde transport from Golgi to ER (e.g. Sec20), or from Golgi to plasma membrane (e.g. Snap24), which has an important role regulating neurotransmitter release via synaptic vesicle fusion to neuronal membrane. Palmitoylation of SNARE proteins has been also observed in yeast (Roth et al., 2006) and mammals (Valdez-Taubas & Pelham, 2005). Other transporter proteins directly enable the movement of ions and molecules across

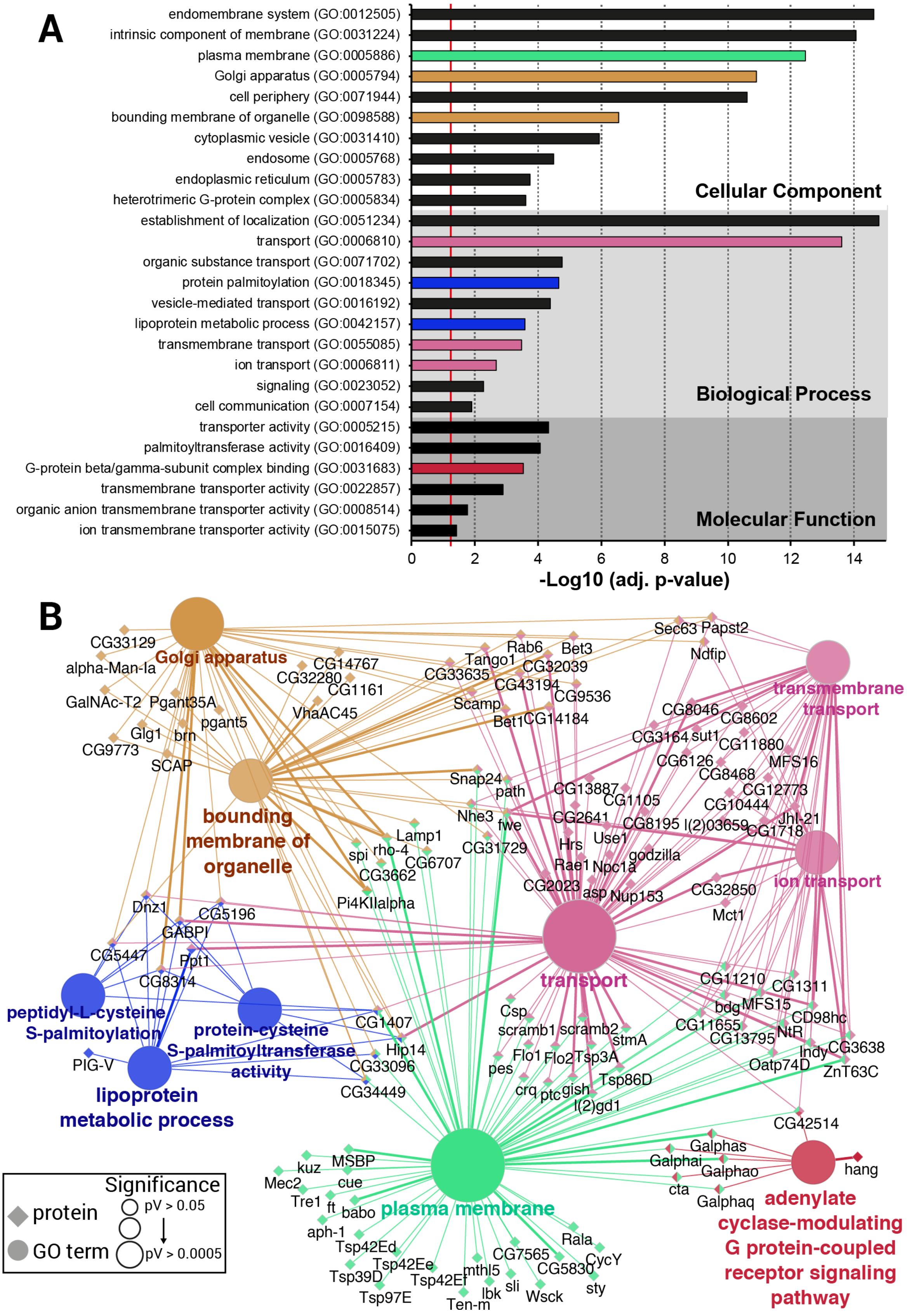

A
endomembrane system (GO:0012505)
intrinsic component of membrane (GO:0031224)
plasma membrane (GO:0005886)
Golgi apparatus (GO:0005794)
cell periphery (GO:0071944)
bounding membrane of organelle (GO:0098588)
cytoplasmic vesicle (GO:0031410)
endosome (GO:0005768)
endoplasmic reticulum (GO:0005783)
heterotrimeric G-protein complex (GO:0005834)
establishment of localization (GO:0051234)
transport (GO:0006810)
organic substance transport (GO:0071702)
protein palmitoylation (GO:0018345)
vesicle-mediated transport (GO:0016192)
lipoprotein metabolic process (GO:0042157)
transmembrane transport (GO:0055085)
ion transport (GO:0006811)
signaling (GO:0023052)
cell communication (GO:0007154)
transporter activity (GO:0005215)
palmitoyltransferase activity (GO:0016409)
G-protein beta/gamma-subunit complex binding (GO:0031683)
transmembrane transporter activity (GO:0022857)
organic anion transmembrane transporter activity (GO:0008514)
ion transmembrane transporter activity (GO:0015075)
Cellular Component
Biological Process
Molecular Function
0
2
4
6
8
10
12
14
-Log10 (adj. p-value)
B
Golgi apparatus
bounding membrane of organelle
transmembrane transport
ion transport
transport
peptidyl-L-cysteine S-palmitoylation
protein-cysteine S-palmitoyltransferase activity
lipoprotein metabolic process
plasma membrane
adenylate cyclase-modulating G protein-coupled receptor signaling pathway
CG33129
alpha-Man-Ia
GalNAc-T2
Pgant35A
pgant5
Glg1
brn
CG9773
SCAP
CG14767
CG32280
CG1161
VhaAC45
Rab6
Bet3
Tango1
CG32039
CG33635
CG43194
CG9536
Scamp
Bet1
CG14184
Sec63
Papst2
Ndfip
CG8046
CG8602
CG3164
sut1
CG11880
CG6126
MFS16
CG8468
CG12773
CG10444
Jhl-21
l(2)03659
CG1718
Snap24
path
CG13887
CG1105
Nhe3
fwe
CG2641
CG8195
Use1
Lamp1
spi
rho-4
CG6707
CG3662
CG31729
Hrs
godzilla
Rae1
Npc1a
CG2023
asp
Nup153
CG32850
Mct1
Pi4KIIalpha
Dnz1
CG5196
GABPI
CG5447
Ppt1
CG8314
PIG-V
CG1407
Hip14
CG33096
CG34449
Csp
scramb1
scramb2
stmA
pes
Flo1
Flo2
Tsp3A
crq
ptc
gish
Tsp86D
l(2)gd1
CG11210
CG1311
bdg
MFS15
CD98hc
CG11655
CG13795
NtR
Indy
CG3638
Oatp74D
ZnT63C
CG42514
Galphas
Galphai
Galphao
cta
Galphaq
hang
kuz
MSBP
Mec2
cue
Tre1
ft
babo
aph-1
Tsp42Ed
Tsp42Ee
Tsp39D
Tsp42Ef
mthl5
Tsp97E
Ten-m
lbk
sli
Wsck
CG7565
CG5830
Rala
CycY
sty
Significance
protein
GO term
pV > 0.05
pV > 0.0005

Figure 4.4: GO enrichment analysis of *S*-palmitoylated proteins in S2R+ cells. **(A)** Barplot of the most significantly enriched functional terms in the three GO ontologies. Red vertical line indicates the adjusted p-value cut-off (0.05). Some bars are coloured differently to match the figure below. **(B)** Alternative network-like visualization of selected, non-redundant, enriched GO terms and proteins associated to those, created with the ClueGo+CluePedia plugin for Cytoscape (Bindea et al., 2009). Circles represent GO terms and their size is proportional to their significance (although all p-values < 0.05); small diamonds represent proteins. Colours represent functional groups of closely related GO terms. For clarity, terms associated with 'intrinsic component of membrane', which are virtually associated to every protein, were excluded.

the plasma membrane, like: Indy ("I'm not dead yet"), which transports Krebs cycle intermediates through the gut epithelium; the putative sodium-dependent transporter bdg (bedraggled) involved in R3/R4 photoreceptor cell fate commitment; and NtR, a predicted neurotransmitter-gated ion channel required for synaptic transmission (Figure **4.4B**). In addition, there is significant enrichment in the G protein-coupled receptor (GPCR) signalling pathway, due to the presence of several different heterotrimeric G protein alpha (Gα) subunits (Galphai, Galphao, Galphas, Galphaq and cta, in Figure **4.4B**). In mammals, palmitoylation of Gα targets the fully assembled G protein to the cytoplasmic face of the plasma membrane, where it binds to the GPCR and remains inactive until the arrival of a signal (Wedegaertner et al., 1993). Indeed, palmitoylation modifies almost every component of G protein signalling, including GPCRs and regulator proteins, and thus cycles of palmitoylation/depalmitoylation have a key regulatory role in the whole pathway (Smotrys & Linder, 2004). GPCR signalling mediates the response to numerous extracellular stimuli and is thus involved in a wide variety of physiological processes, such as sensory transmission, immune system activity regulation, or cell growth and metastasis. This observation suggests that palmitoylation-dependent regulation of this key signalling pathway is conserved in *Drosophila*.

Another important subset of proteins are the DHHC PATs themselves, which result in enrichment in protein-cysteine *S*-palmitoyltransferase activity. Of the 22 DHHC PATs known so far in *Drosophila*, nine were identified in our acyl-RAC experiment (suggesting that at least those nine are expressed in S2R+ cells), of which seven were palmitoylated. Most of them are known (or inferred by homology) to be located in the Golgi apparatus (Dnz1, GABPI, CG5196, CG8314), and while for the others location in plasma membrane is also predicted (CG34449, Hip14 and CG1407), it has only been experimentally determined for CG1407 (Figure **4.4B**). That these enzymes were positively identified in the palmitoylome is not surprising as it has been shown that many DHHC PATs modify their substrates via a two-step ping-pong mechanism where they undergo auto-acylation (Stix et al., 2020) (Figure

4.5A). Regarding the opposite process—depalmitoylation—, our palmitoylated protein set also includes several thioesterases. In mammals, *S*-palmitoylation of APT1/2 and the ABHD-family thioesterases is also required for proper tethering to the plasma or endosomal membranes, while they can detach from them via auto-depalmitoylation (Figure **4.5B**). Thus, thioesterase activity is also regulated by cycles of palmitoylation/depalmitoylation (Kong et al., 2013; Lin & Conibear, 2015). The *Drosophila* ortholog of the thioesterase APT1 (Apt1 or CG1885) was identified but not found to be significantly enriched by acyl-RAC, and thus we

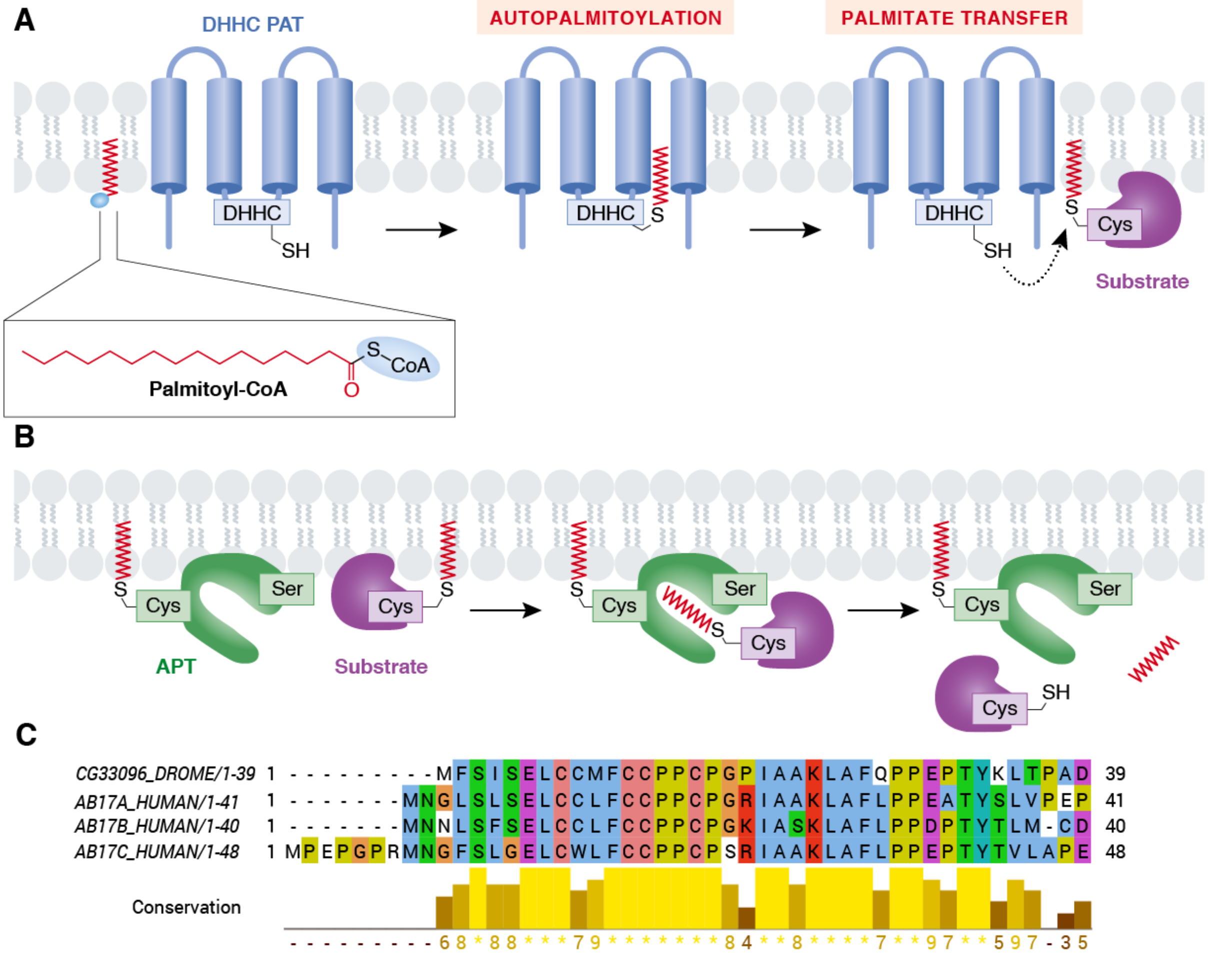

A
DHHC PAT
AUTOPALMITOYLATION
PALMITATE TRANSFER
DHHC
SH
DHHC
S
DHHC
SH
S
Cys
Substrate
S
CoA
O
Palmitoyl-CoA
B
S
Cys
Ser
APT
Cys
S
Substrate
S
Cys
Ser
S
Cys
S
Cys
Ser
Cys
SH
C
CG33096_DROME/1-39 1 ---------MFSISELCCMFCCPPCPGPIAAKLAFQPPEPTYKLTPAD 39
AB17A_HUMAN/1-41 1 -------MNGLSLSELCCLFCCPPCPGRIAAKLAFLPPEATYSLVPEP 41
AB17B_HUMAN/1-40 1 -------MNNLSFSELCCLFCCPPCPGKIASKLAFLPPDPTYTLM-CD 40
AB17C_HUMAN/1-48 1 MPEPGPRMNGFSLGELCWLFCCPPCPSRIAAKLAFLPPEPTYTVLAPE 48
Conservation
---------68*88***79*******84**8****7**97**597-35

Figure 4.5: (A) DHHC PATs are integral membrane proteins which are first autopalmitoylated on the cysteine residue of the DHHC motif, which is located at the cytoplasmic face. Then, the palmitic acid is transferred to an acceptor cysteine of the substrate protein. **(B)** Mammalian acylprotein thiosterases (ATP1/2) undergo *S*-palmitoylation to localize to the membrane where they can carry out depalmitoylation of their substrates. In this process, their hydrophobic pocket accepts the palmitate and positions the substrate's palmitoylated cysteine close to the serine residue of their active site. Finally, it can depalmitoylate itself to detach from the membrane. **(C)** Cysteine-rich N-terminal region of mammalian thioesterases ABHD17A-C where *S*-palmitoylation takes place is conserved in their Drosophila ortholog CG33096 (UniProtKB accession Q9VBXB). Figures A-B were obtained from (Ko & Dixon, 2018).

cannot confirm that is palmitoylated. However, I did find CG33096, a palmitoylated ortholog of the ABHD17 family, that shows conservation of the cysteine-rich N-terminal region where palmitoylation takes place (Figure **4.5C**). This is the first evidence that members of this group of thioesterases are conserved and active in *Drosophila*. Thirdly, I also found palmitoylated Ppt1 (Figure **4.4**), ortholog of the lysosomal thioesterase, which in mammals is also known to be subjected to palmitoylation, although it has been proposed that this does not affect its location but rather results in decreased enzyme activity (Segal-Salto et al., 2016).

Overall, location and function of the proteins in the *Drosophila* S2R+ palmitoylome agree with known roles for *S*-palmitoylated proteins and further confirm the reliability of our dataset. Moreover, despite the fact that GPCR repertoire of invertebrates and vertebrates varies considerably, these results suggest that dynamic *S*-palmitoylation regulation of GPCR signalling could have an origin prior to the formation of the major GPCR superfamilies (Nordström et al., 2011).

4.2.2 The *Drosophila melanogaster* DHHC PAT interactome

4.2.2.1 Identification of DHHC PAT interactors

To identify potential substrates of DHHC PAT enzymes, our collaborators used BioID or proximity biotinylation. This is a technique that allows the *in vivo* identification of protein-protein interactions through the expression of a protein of interest fused to the bacterial biotin ligase mutant BirA (R118G), leading to covalent biotinylation of nearby proteins (and likely interaction partners). Standard affinity purification followed my mass spectrometry can then be used to identify biotin-tagged proteins (Roux et al., 2018) (Figure **4.6A**). The main advantage of BioID is that it can capture weak or transient interactions, such as those between enzymes and their substrates, with the possible disadvantage of also retrieving non-interacting proteins near to the target (Liu et al., 2018). BioID had been successfully applied to many mammalian systems, but so far not in *Drosophila*. Our collaborators optimized conditions and developed a protocol that enabled the use of BioID in the S2R+ embryonic hemocyte-like cell line of this organism. The viability of this protocol for identifying specific interactions between DHHC PAT and their putative client proteins was first verified using the *Drosophila* Huntingtin-interacting protein 14 (dHip14) and Snap25, which is an interaction known and shared with their mammalian orthologs (Figure **4.6B**).

After establishing that BioID works in this organism, it was extended overexpressing BioID-fusion constructs for ten different *Drosophila* DHHC PATs (dHip14, CG8314, CG5196, CG5880, Patsas, app, GabPI, CG1407, CG4676 and Dnz1), and performed two independent experiments (with three replicates per DHHC PAT) to identify their interactomes and putative client spectra. A total of 2162 proteins were identified between both experiments, of which 487 proteins were enriched in at least one DHHC PAT-BioID sample compared with the negative control (S2R+ cells transfected with empty vector), and thus considered potential interactors for one or more of the ten target enzymes.

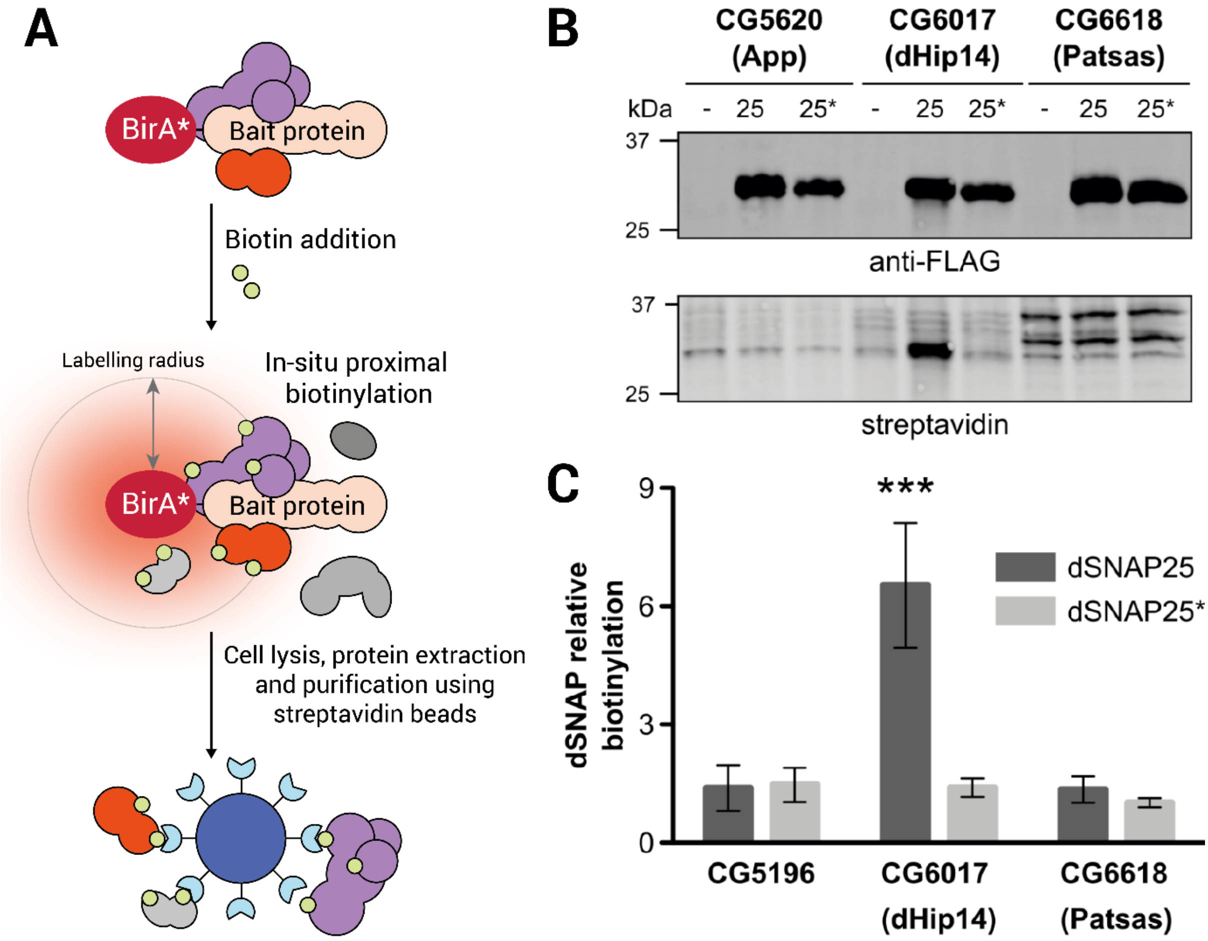


Figure 4.6: (A) Schematic illustration of the basis behind BioID method. Target protein (bait) is fused with the biotin ligase BirA which, after adding biotin, is able to biotinylate *in situ* all proteins located near the bait (preys) and thus candidate interactors. **(B-C)** Western blot and quantification corresponding to BioID test assay with the co-overexpressed DHHC-PATs dHip14, Patsas and app as BioID-fusion target proteins, together with FLAG-dSnap25 wild type (25) and FLAG-dSnap25 proline mutant (25*). dSnap25* contains a proline-to-alanine mutation at a key residue in the known dHip14 binding motif which is known to inhibit the interaction (Lemonidis et al., 2015). As expected, only the interaction between dHip14 and the wild type version of dSnap25 is recovered, confirming that the assay is specific enough.

Next, I checked the overlap with the previously determined palmitoylome and found that only 25% of *S*-palmitoylated proteins could be assigned to any DHHC PAT (Figure **4.7A**). The fact that no interactions were found for the large majority could be due to several factors. First, the experiments only covered 10 out of the 22 total DHHC-PATs (known so far) in *Drosophila*, thus many of these *S*-palmitoylated proteins might be substrates of the enzymes that were not tested. Second, for the ten chosen enzymes, we did not consider existing alternative splicing isoforms. Most DHHC PATs have several isoforms (e.g. CG1407 has at least 6) that differ mainly in their C-terminal cytoplasmic tails, which are believed to be important for protein recognition, and therefore result in different substrate spectra (Howie et al., 2014). Lastly, despite the high sensitivity of the BioID technique, this new protocol for the *Drosophila* S2R+ cell line was just established by our collaborators and it might still not be as efficient as in other systems. It is likely that many enzyme-substrate transient interactions are missed.

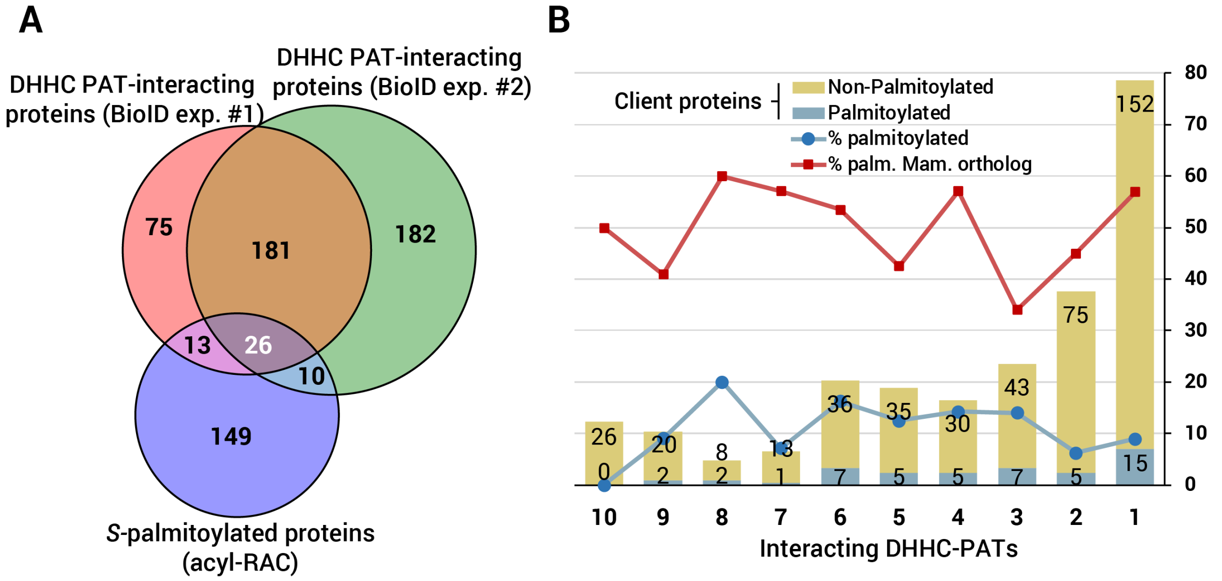


Figure 4.7: (A) Venn diagram showing the overlap between the proteins determined to have at least one putative DHHC-PAT interaction in the BioID experiment #1 (pink bubble), BioID experiment #2 (green bubble) and the proteins determined to be *S*-palmitoylated in the acyl-RAC assay (purple bubble). **(B)** Barplot showing the distribution in the number of DHHC-PAT interactions (up to 10) for those proteins that showed at least one interaction in any or both BioID experiments. Bars are divided into proteins that were deemed palmitoylated and those that were not. In addition, line plots in the right axis indicate the percentage of palmitoylated proteins and the percentage of proteins with a palmitoylated mammalian ortholog in each bar.

4.2.2.2 Target proteins interact with several DHHC PATs

Substrate specificity of the different DHHC PATs still remains obscure even for better-researched organisms. For many important palmitoylation targets, the responsible DHHC PATs have not been yet clearly discerned. Although *S*-palmitoylated proteins can be modified by more than one DHHC PAT (Hou et al., 2009), the degree of promiscuity among PATs is still unknown. According to yeast and mammalian DHHC PAT–substrate analyses, most client proteins are unlikely to be palmitoylated by more than 4-6 different enzymes (Greaves & Chamberlain, 2011). However, in yeast most of the seven DHHC PATs can be knocked out without adverse effects, indicating that some proteins might be palmitoylated by any available enzyme (Roth et al., 2006) or that they could suffer spontaneous palmitoylation in the presence of palmitoyl-CoA, as suggested previously (Corvi et al., 2001). Conversely, BioID has a chance of recovering proteins that are not direct substrates or binders of the target DHHC PAT, but simply neighbouring, accessory proteins within the 10 nm labelling-distance, which are not functionally relevant (Figure **4.6A**). Thus the question that remains open is how high the number of DHHC PAT interactions per client protein can be without being simply the result of co-localization.

In *Drosophila* S2R+ cells, I found that identified putative client proteins interact with 3.5 DHHC PATs on average. I also looked at the distribution in the number of DHHC PAT interactions per protein, under the assumption that enrichment of interacting *S*-palmitoylated proteins could be indicative of real enzyme client proteins (Figure **4.7B**). I was, however, not able to infer a range or maximum limit to the number of DHHC PAT interactions per client protein, as the fraction of *S*-palmitoylated does not show a clear preference (blue line in Figure **4.7B**). However, the range between 4-6 DHHC PATs seems to be the most enriched, which agrees to the previous estimations from mammals. In addition, I used the percentage of proteins with palmitoylated mammalian orthologs as another indicative (red line in Figure **4.7B**), but found these numbers to be too variable to indicate any enrichment. I did notice that there were zero *S*-palmitoylated proteins within those that interact with the 10 DHHC PAT. Despite the high degree of overlap in their client spectra, it seems unlikely that real target proteins can be palmitoylated by all 10 enzymes even if only for their different cellular locations, thus it can be concluded that many of these 10-interaction proteins are suspected of being general accessory or palmitoylation-irrelevant proteins.

4.2.2.3 DHHC PATs can be divided in two groups according to preference for transmembrane substrates or substrates with the Hip14-binding motif

Next, I studied and compared the results for the different DHHC PATs (Figure **4.8A**). While the number of potential interactors for each enzyme is quite diverse—ranging from little more than 100 in app and Hip14 to more than 200 in CG8314 and CG4676—the fraction of these that are S-palmitoylated (according to the *Drosophila* S2R+ palmitoylome) remains relatively constant, at ~10%, indicating that the different sizes in substrate spectra might be real.

The current paradigm is that most acyltransferases show strong overlap in their substrate spectra (Hou et al., 2009). However, studies in yeast and mammals have shown that PATs can be grouped according to the relative position of the palmitate-accepting cysteine residue in relation to transmembrane domains in their substrates (Roth et al., 2006), or according to their preference for soluble proteins and for integral membrane proteins (Ohno et al., 2012). To check whether substrate preferences could be inferred from our data, I obtained transmembrane domains (TMD) annotations for all target proteins (either from UniProtKB directly or predicted; see Methods), and found a certain bias of a group of four DHHC PATs (group 1: CG8314, CG5196, CG5880 and Patsas) towards substrates with TMDs compared to the other six enzymes (group 2: app, Hip14, GabPI, CG1407, CG4676 and Dnz1) (Figure **4.8B**). On average 39% of interactors for enzymes in group 1 have at least one TMD, as opposed to 22.3% for enzymes in group 2 (*t* test *p*-value < 0.01). If this comparison is restricted to *S*-palmitoylated interactors, which are more likely to be real DHHC PATs substrates, this difference becomes even more apparent: 84% v. 52.4% (*t* test *p*-value << 0.01). Further classifying interactors into single TMD- or multiple TMD-containing proteins revealed that there were no significant differences between the group 1 and 2 of DHHC PATs in the fraction of multi-pass membrane proteins, rather the big gap was due to a sharp difference between the percentage of single-pass membrane proteins (60% v 24%, *t* test *p*-value << 0.01; not shown in Figure). This is consistent with previous findings that already pointed to DHHC PATs preference differences in regard to the number of membrane-spanning domains of their substrate proteins (Ohno et al., 2012).

I did then a similar analysis calculating now the fraction of interactors that have the known Hip14 (or zDHHC17)-recognition motif. When focusing on *S*-palmitoylated target proteins, I observed an interesting correlation between the same two groups of DHHC PATs: those enzymes that show a preference for transmembrane proteins seem to disfavour interactions with proteins containing the Hip14-binding motif, and *vice versa* (Figure **4.8C**).

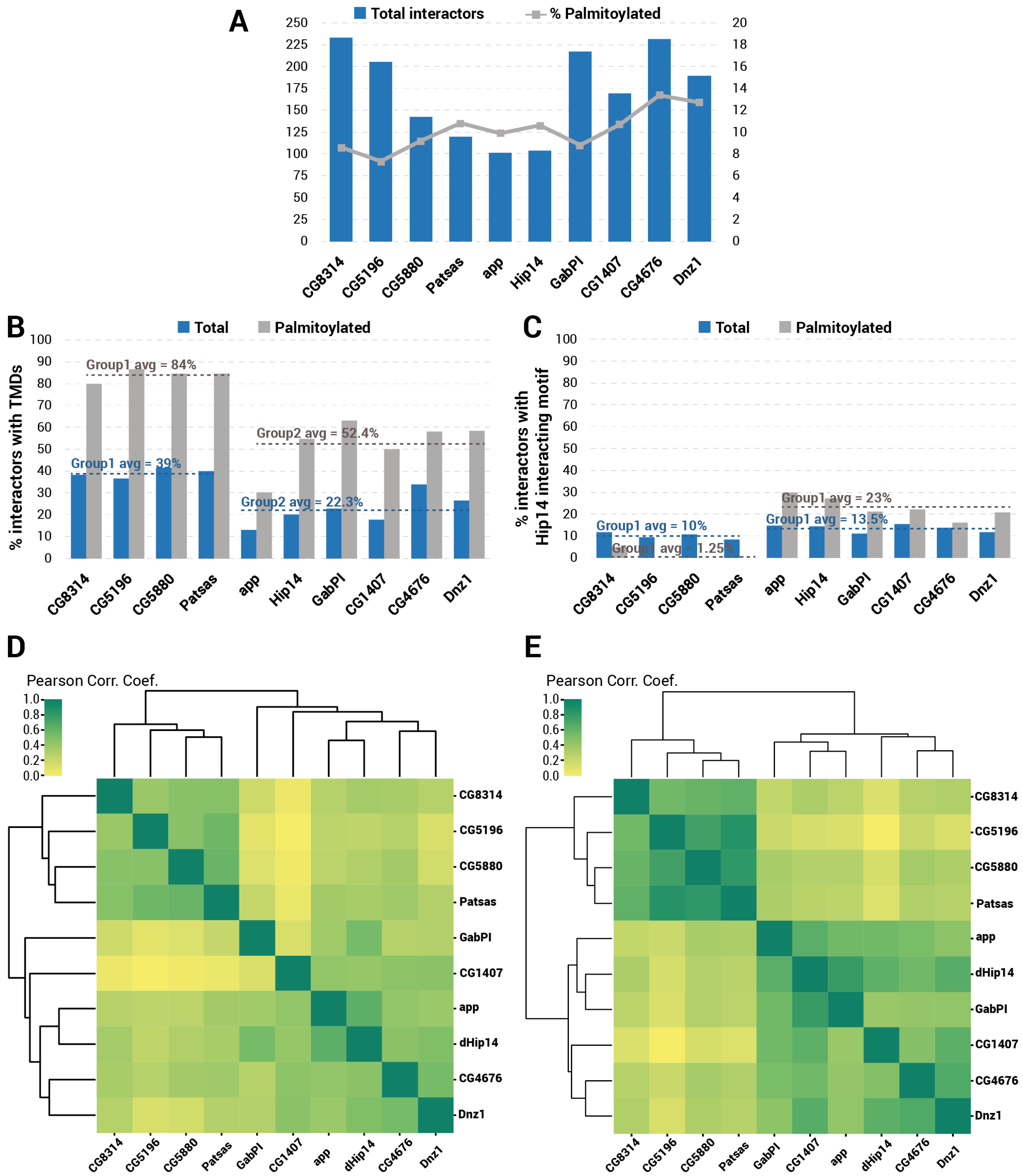
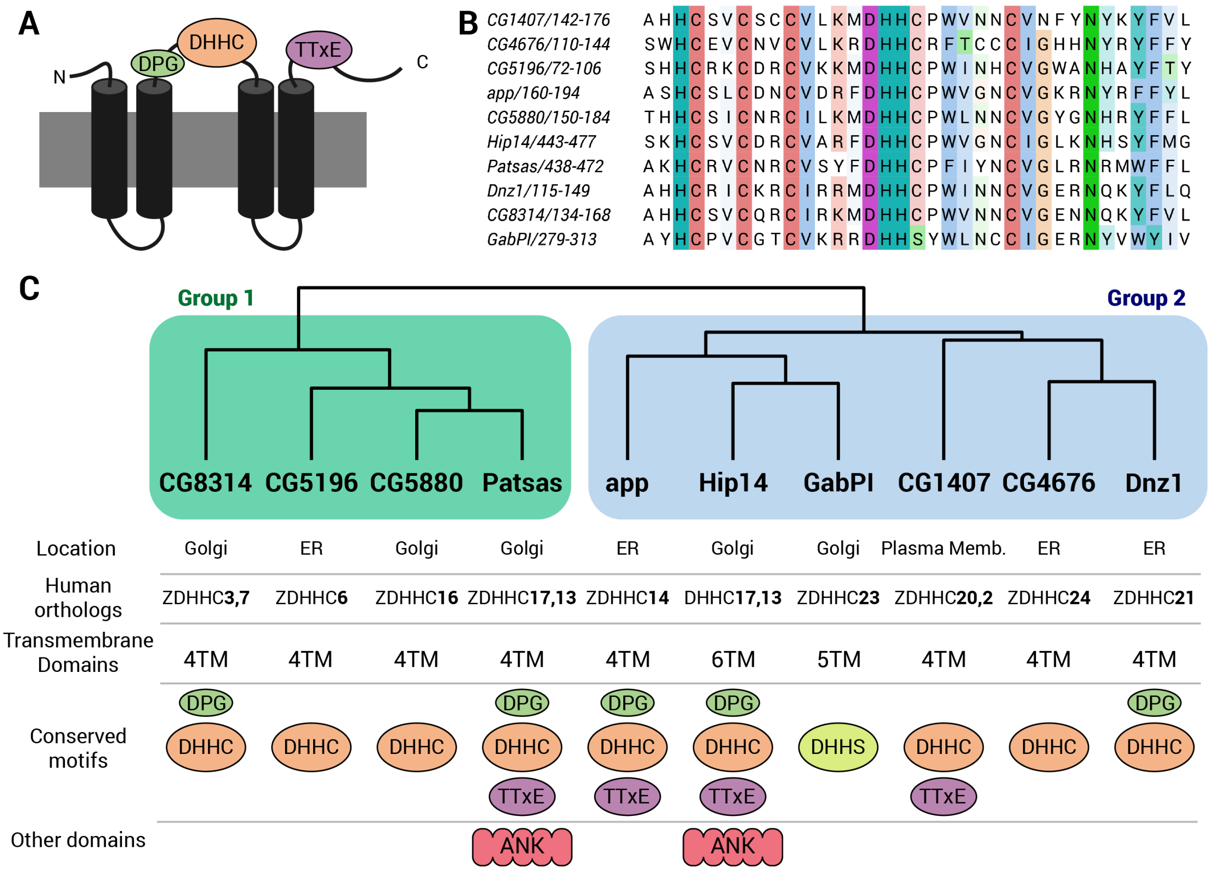


Figure 4. 8: **(A)** Number of identified interactors in the BioID experiments (bar plot) and the fraction of those that were found palmitoylated (line plot), per DHHC PAT. **(B)** Fraction (%) of interactors with at least one transmembrane domain with respect to the total number of interactors (blue bars) or the number of *S*-palmitoylated interactors (grey bars). **(C)** Same as B but showing the fraction of interactors that contain the Hip14-binding motif. **(D)** Cluster analysis of DHHC PATs according to the Pearson correlation coefficients calculated by pairwise comparison of their interaction profiles with respect to all interactors or **(E)** only *S*-palmitoylated interactors. In both cases, the DHHC PATs cluster into two clearly defined groups of four and six members, respectively.

Group 1 enzymes, with the exception of CG8314, have no *S*-palmitoylated interactors that contain the Hip14 motif; in contrast, 23% of *S*-palmitoylated interactors for group 2 enzymes have the motif (1.25% v. 23%, *t* test p-value <<< 0.1). Cluster analyses of the DHHC PATs interaction profiles using all identified interactors (Figure **4.8D**) —and more confidently, when using only those that are S-palmitoylated (Figure **4.8E**) — confirmed the broad subdivision of the ten enzymes into two groups according to their substrate preferences.

More specific substrate similarities between DHHC PATs are difficult to infer and even more difficult to relate to particular structural features of the enzymes. According to the cluster analysis, the protein sharing the most similar interaction profiles are CG4676 and Dnz1, both of which are located at the ER and contain 4 TMDs; as well as Patsas and CG5880, also having 4 TMDs but located at the Golgi (Figure **4.9C**). dHip14 and Patsas are both orthologs to the zDHHC17/13 mammalian *S*-acyltransferases (mammalian HIP14), and thus the only enzymes that have Ankyrin repeat regions which mediate substrate recognition via the Hip14-binding motif (Lemonidis et al., 2015). Strikingly, they not only cluster in different groups but none of Patsas *S*-palmitoylated interactors actually contain this motif (Figure **4.8C**). This agrees with the previous co-expression experiments with Snap25 and mutant Snap25 that showed that only dHip14 interacts with this known substrate via the Hip14-binding motif (Figure **4.6B-C**). This is also in agreement with the notion that dHip14 and Patsas are more distantly related than their mammalian orthologs (ZDHHC17/HIP14 and ZDHHC13/HIP14L) (Bannan et al., 2008).

In terms of cellular location, *Drosophila* DHHC PATs are mainly located at the ER or the Golgi with the exception of CG1407 that, like its human counterpart ZDHHC20, localizes on the plasma membrane (Bannan et al., 2008; Ohno et al., 2006) (Figure **4.9C**). Enrichment analysis of the target proteins for the two DHHC PAT groups revealed that, while location at the Golgi is shared equally, there is specific enrichment of endoplasmic reticulum location in target proteins of group 1 enzymes (CG8314, CG5196, CG5880 and Patsas), and of plasma membrane location in the targets of the group 2 enzymes, which is likely due to the client proteins of CG1407 exclusively. Considering that, out of the four group 1 enzymes, CG5196 is reportedly the only one primarily located at the ER (the others localize in the Golgi), enrichment of their target proteins in this compartment is surprising. This might be another hint that the same client protein might be palmitoylated at different endomembrane system components.


A
N
DPG
DHHC
TTxE
C
B
CG1407/142-176 AHHCSVCSCCVLKMDHHCPWVNNCVNFYNYKYFVL
CG4676/110-144 SWHCEVCNVCVLKRDHHCRFTCCCIGHHNYRYFFY
CG5196/72-106 SHHCRKCDRCVKKMDHHCPWINHCVGWANHAYFTY
app/160-194 ASHCSLCDNCVDRFDHHCPWVGNCVGKRNYRFFYL
CG5880/150-184 THHCSICNRCILKMDHHCPWLNNCVGYGNHRYFFL
Hip14/443-477 SKHCSVCDRCVARFDHHCPWVGNCIGLKNHSYFMG
Patsas/438-472 AKHCRVCNRCVSYFDHHCPFIYNCVGLRNRMWFFL
Dnz1/115-149 AHHCRICKRCIRRMDHHCPWINNCVGERNQKYFLQ
CG8314/134-168 AHHCSVCQRCIRKMDHHCPWVNNCVGENNQKYFVL
GabPI/279-313 AYHCPVCGTCVKRRDHHSYWLNCCIGERNYVWYIV
C
Group 1
Group 2

	CG8314	CG5196	CG5880	Patsas	app	Hip14	GabPI	CG1407	CG4676	Dnz1
Location	Golgi	ER	Golgi	Golgi	ER	Golgi	Golgi	Plasma Memb.	ER	ER
Human orthologs	ZDHHC**3,7**	ZDHHC**6**	ZDHHC**16**	ZDHHC**17,13**	ZDHHC**14**	DHHC**17,13**	ZDHHC**23**	ZDHHC**20,2**	ZDHHC**24**	ZDHHC**21**
Transmembrane Domains	4TM	4TM	4TM	4TM	4TM	6TM	5TM	4TM	4TM	4TM
Conserved motifs	DPG DHHC	DHHC	DHHC	DPG DHHC TTxE	DPG DHHC TTxE	DPG DHHC TTxE	DHHS	DHHC TTxE	DHHC	DPG DHHC
Other domains				ANK		ANK				

Figure 4.9: (A) Canonical structure of DHHC PAT with conserved motifs. **(B)** The DHHC motif within the cysteine-rich region is conserved among the 10 PATs studied here, with the exception of GabPI where Cys is substituted by Ser. **(C)** Schematic showing different structural features and corresponding mammalian orthologs for the 10 DHHC PATs, which are divided in group 1 (green) and group 2 (blue) defined by substrates preferences and clustering analysis (see Fig 4.8).

Examples of proteins that seem to be preferential clients of the group 1 DHHC PATs are the SNARE proteins Bet1 and Use1, both single-span transmembrane proteins involved in vesicle-mediated transport. As reported before (Valdez-Taubas & Pelham, 2005), DHHC PATs might have here a protective role in cellular quality control of integral membrane proteins, since palmitoylation of Cys residues on the cytoplasmic end of their TMDs protects proteins from premature ubiquitination. Alternatively, exclusive targets of group 2 enzymes include several plasma membrane proteins such as the proton-coupled amino acid transporter-like protein pathetic (path), the phospholipid scramblase (scramb2) or the protein ben (be), all with roles in the synaptic function. In addition, CG33096, the *Drosophila* ortholog of the mammalian thioesterase family ABHD17A-C was found to be a substrate of the DHHC PAT Dnz1, providing further evidence that it is a thioesterase (as discussed in section **4.2.1.5**, thioesterases require to be palmitoylated). The DHHC PAT Dnz1 is also interesting because even though it has been reported to locate at the ER (Bannan et al., 2008), it is most likely an ortholog of human ZDHHC21, which is primarily localized at the plasma membrane (Ohno et al., 2006). I found that, together with CG1407, Dnz1 is the DHHC PAT with the highest fraction of plasma membrane client proteins, and thus hypothesize that this enzyme, similarly to ZDHHC21, could have more than one location in the cell.

Finally, in contrast to the other DHHC family members, GabPI does not actually contain the conserved DHHC motif but a DHHS one, replacing the functionally essential Cys residue for a Ser (Figure **4.9B**). In fact, GabPI has been reported to lack palmitoylation activity and instead to have a role in the proper localization of galactosyltransferases in the Golgi through tight interactions (Johswich et al., 2009). The high number of interactors that were identified for GabPI, if not simply the experimental result of promiscuous binding, could suggest a similar activity towards a wider range of proteins.

4.3 Materials and methods

4.3.1 Experimental procedures

Extensive information about experimental methods can be found in our publication (Porcellato et al., 2022). All experimental work was exclusively done by Dr. Elena Porcellato, Dr. Christoph Metzendorf and others at the group of Prof. Dr. Felix Wieland, thus it will not be covered here.

4.3.2 Data analysis & statistics

The MaxQuant software (Tyanova et al., 2016) was used for processing the raw mass-spectrometry data. The resulting protein intensities were normalized by the median difference between each sample and the negative controls. Subsequent analysis, calculation of fold-changes and respective significance was done using a novel method, proDA (version 0.1.), developed by two collaborators, Constantin Ahlmann-Eltze and Dr. Simon Anders, and which specifically deals with the common problem of label-free proteomics that is the large number of non-random missing values. Instead of simple value imputation, proDA uses the overall dropout-probability for each intensity and empirical Bayesian priors to calculate a principled statistical test, recovering more true positives while controlling the false discovery rate (Ahlmann-Eltze & Anders, 2019).

4.3.3 Bioinformatics analysis

Protein identifiers, sequences, descriptions, and gene names were obtained from UniProt (Bateman et al., 2021). As UniProt contains many redundant and poorly annotated entries for *Drosophila melanogaster*, all proteins were mapped to unique entries of Flybase (Larkin et al., 2021). Flybase is a dedicated and centralized resource with highly curated data of *Drosophila melanogaster* genes.

To determine mammalian orthologs for *Drosophila* proteins, I used the DRSC Integrative Ortholog Prediction Tool (DIOPT) (Hu et al., 2011) to search against human, mouse and rat proteomes, selecting only those orthologs with the highest confidence. Proteins were considered to have mammalian orthologs if there was a hit in any of these three organisms, although human alone was sufficient for 95% of proteins, as *Drosophila* orthologs found in rodents and not humans are rare. The palmitoylation status of these mammalian orthologs was then determined by searching in SwissPalm (Release 3 2019-09-08) (Blanc et al., 2015). If present in this database, mammalian proteins were considered palmitoylated; in addition, if they were reported by at least one targeted study, or by at least two different experimental techniques, they were additionally classified as high confidence. The human palmitoylome fraction (17.5%) was calculated as the current number of human proteins in SwissPalm (3593) divided by the human proteome size in UniProt (20577 in release 2022_01).

Palmitoylation sites were predicted using the program CSS-Palm 4.0 (Ren et al., 2008) with high threshold. CSS-Palm was the first palmitoylation site predictor (F. Zhou et al., 2006) and has been updated and improved since, partly thanks to the progressive publication of more experimentally determined palmitoylation sites that can be used as training data. Since the algorithm predicts palmitoylation sites, any protein with predicted sites was considered palmitoylated (Figure **4.3D**).

Data about transmembrane domains were obtained from UniProt if available, otherwise predicted using TMHMM-2.0 (Krogh et al., 2001).

Functional enrichment analyses of Gene Ontology (GO) terms were performed using Fisher's Exact test and False Discovery Rate (FDR) for multiple-testing correction, through the PANTHER overrepresentation tool (Mi et al., 2019). For putative *S*-palmitoylation proteins resulting from the acyl-RAC experiment, the full list of proteins identified by the experiment was used as the *appropriate* background gene list.

All computer analyses were performed using Python.

4.3.4 Data availability

Fully annotated tables containing the S2R+ cell palmitoylome and DHHC-PAT interactome are available online at russelllab.org/jcgonzalez

4.4 Conclusion

Due to its broad range of substrates, palmitoylation is involved in multiple cellular processes and has thus gained attention in recent years, particularly since it has been linked to several neurological diseases and also to cancer. Consequently, knowledge of the mammalian palmitoylome and involved enzymes has also increased substantially. In contrast, there is almost no information of palmitoylation in *Drosophila*. Here, we provided the second and most comprehensive list of putative *S*-palmitoylated proteins in this model organism, the first interactome—including potential substrate-client spectra—for 10 DHHC PATs, and many new insights that highlight the functional similarities and high degree of conservation between palmitoylation-regulated proteins and processes in *Drosophila* and mammals.

These new insights and available data provide a useful resource that the community can build upon to further characterize this important protein modification. Future experiments could iteratively help to expand the catalogue of *S*-palmitoylation targets as well as to validate those previously established and ultimately elucidate the level of substrate specificity and redundancy for the DHHC protein family.

Bibliography

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Abriel, H., Loffing, J., Rebhun, J. F., Pratt, J. H., Schild, L., Horisberger, J. D., Rotin, D., & Staub, O. (1999). Defective regulation of the epithelial Na+ channel by Nedd4 in Liddle's syndrome. *Journal of Clinical Investigation*. https://doi.org/10.1172/JCI5713

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*(4), 248–249. https://doi.org/10.1038/nmeth0410-248

Ahlmann-Eltze, C., & Anders, S. (2019). proDA: Probabilistic Dropout Analysis for Identifying Differentially Abundant Proteins in Label-Free Mass Spectrometry. *BioRχiv*. https://doi.org/https://doi.org/10.1101/661496

Aithal, A., Rauth, S., Kshirsagar, P., Shah, A., Lakshmanan, I., Junker, W. M., Jain, M., Ponnusamy, M. P., & Batra, S. K. (2018). MUC16 as a Novel Target for Cancer Therapy. *Expert Opinion on Therapeutic Targets*, *22*(8), 675. https://doi.org/10.1080/14728222.2018.1498845

Akavia, U. D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H. C., Pochanard, P., Mozes, E., Garraway, L. A., & Pe'Er, D. (2010). An integrated approach to uncover drivers of cancer. *Cell*, *143*(6), 1005–1017. https://doi.org/10.1016/J.CELL.2010.11.013

Alber, F., Dokudovskaya, S., Veenhoff, L. M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Sali, A., & Rout, M. P. (2007). The molecular architecture of the nuclear pore complex. *Nature 2007 450:7170*, *450*(7170), 695–701. https://doi.org/10.1038/nature06405

Aloy, P., Böttcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A. C., Bork, P., Superti-Furga, G., Serrano, L., & Russell, R. B. (2004). Structure-Based Assembly of Protein Complexes in Yeast. *Science*. https://doi.org/10.1126/science.1092645

Aloy, P., Ceulemans, H., Stark, A., & Russell, R. B. (2003). The relationship between sequence and interaction divergence in proteins. *Journal of Molecular Biology*. https://doi.org/10.1016/j.jmb.2003.07.006

Aloy, P., & Russell, R. B. (2002). Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.092147999

Aloy, P., & Russell, R. B. (2003). InterPreTS: Protein Interaction Prediction through Tertiary Structure. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/19.1.161

Aloy, P., & Russell, R. B. (2004). Ten thousand interactions for the molecular biologist. In *Nature Biotechnology* (Vol. 22, Issue 10, pp. 1317–1321). Nature Publishing Group. https://doi.org/10.1038/nbt1018

Amberger, J. S., Bocchini, C. A., Scott, A. F., & Hamosh, A. (2019). OMIM.org: Leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Research*, *47*(D1), D1038–D1043. https://doi.org/10.1093/nar/gky1151

Amir, R. E., Van Den Veyver, I. B., Wan, M., Tran, C. Q., Francke, U., & Zoghbi, H. Y. (1999). Rett syndrome is caused by mutations in X-linked MECP2, encoding methyl-CpG-binding protein 2. *Nature Genetics*, *23*(2), 185–188. https://doi.org/10.1038/13810

Ashley, E. A. (2016). Towards precision medicine. *Nature Reviews Genetics 2016 17:9*, *17*(9), 507–522. https://doi.org/10.1038/nrg.2016.86

Bannan, B. A., Van Etten, J., Kohler, J. A., Tsoi, Y., Hansen, N. M., Sigmon, S., Fowler, E., Buff, H., Williams, T. S., Ault, J. G., Glaser, R. L., & Korey, C. A. (2008). The Drosophila protein palmitoylome: characterizing palmitoyl-thioesterases and DHHC palmitoyl-transferases. *Fly*, *2*(4), 198–214. https://doi.org/10.4161/FLY.6621

Bartek, J., & Lukas, J. (2003). Chk1 and Chk2 kinases in checkpoint control and cancer. *Cancer Cell*, *3*(5), 421–429. https://doi.org/10.1016/S1535-6108(03)00110-7

Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Silva, A. Da, Denny, P., Dogan, T., Ebenezer, T. G., Fan, J., Castro, L. G., … Zhang, J. (2021). UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Research*, *49*(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100

Ben-Shem, A., De Loubresse, N. G., Melnikov, S., Jenner, L., Yusupova, G., & Yusupov, M. (2011). The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science*, *334*(6062), 1524–1529. https://doi.org/10.1126/SCIENCE.1212642/SUPPL_FILE/BEN-SHEM.SOM.PDF

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., & Shindyalov, I. N. (2000). The Protein Data Bank (www.rcsb.org). *Nucleic Acids Research*. https://doi.org/10.1093/nar/28.1.235

Betts, M. J., Lu, Q., Jiang, Y., Drusko, A., Wichmann, O., Utz, M., Valtierra-Gutierrez, I. A., Schlesner, M., Jaeger, N., Jones, D. T., Pfister, S., Lichter, P., Eils, R., Siebert, R., Bork, P., Apic, G., Gavin, A. C., & Russell, R. B. (2015). Mechismo: Predicting the mechanistic impact of mutations and modifications on molecular interactions. *Nucleic Acids Research*, *43*(2), e10. https://doi.org/10.1093/nar/gku1094

Betts, M. J., & Russell, R. B. (2007). Amino-Acid Properties and Consequences of Substitutions. In *Bioinformatics for Geneticists: A Bioinformatics Primer for the Analysis of Genetic Data: Second Edition*. https://doi.org/10.1002/9780470059180.ch13

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W. H., Pagès, F., Trajanoski, Z., & Galon, J. (2009). ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, *25*(8), 1091. https://doi.org/10.1093/BIOINFORMATICS/BTP101

Blanc, M., David, F., Abrami, L., Migliozzi, D., Armand, F., Bürgi, J., & van der Goot, F. G. (2015). SwissPalm: Protein Palmitoylation database. *F1000Research*. https://doi.org/10.12688/f1000research.6464.1

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bansal, P., Bridge, A. J., Poux, S., Bougueleret, L., & Xenarios, I. (2016). Uniprotkb/swiss-prot, the manually annotated section of the uniprot knowledgebase: How to use the entry view. In *Methods in Molecular Biology*. https://doi.org/10.1007/978-1-4939-3167-5_2

Boycott, K. M., Vanstone, M. R., Bulman, D. E., & MacKenzie, A. E. (2013). Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics*, *14*(10), 681–691. https://doi.org/10.1038/nrg3555

Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., Di Costanzo, L., Christie, C., Dalenberg, K., Duarte, J. M., Dutta, S., Feng, Z., Ghosh, S., Goodsell, D. S., Green, R. K., Guranović, V., Guzenko, D., Hudson, B. P., Kalro, T., Liang, Y., … Zardecki, C. (2019). RCSB Protein Data Bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky1004

Cai, Z., Chehab, N. H., & Pavletich, N. P. (2009). Structure and activation mechanism of the CHK2 DNA damage checkpoint kinase. *Molecular Cell*, *35*(6), 818–829. https://doi.org/10.1016/J.MOLCEL.2009.09.007

Camp, L. A., & Hofmann, S. L. (1993). Purification and properties of a palmitoyl-protein thioesterase that cleaves palmitate from H-Ras. *Journal of Biological Chemistry*, *268*(30), 22566–22574. https://doi.org/10.1016/S0021-9258(18)41567-0

Campbell, P. J., Getz, G., Korbel, J. O., Stuart, J. M., Jennings, J. L., Stein, L. D., Perry, M. D., Nahal-Bose, H. K., Ouellette, B. F. F., Li, C. H., Rheinbay, E., Nielsen, G. P., Sgroi, D. C., Wu, C. L., Faquin, W. C., Deshpande, V., Boutros, P. C., Lazar, A. J., Hoadley, K. A., … Zhang, J. (2020). Pan-cancer analysis of whole genomes. *Nature*, *578*(7793), 82. https://doi.org/10.1038/S41586-020-1969-6

Canisius, S., Martens, J. W. M., & Wessels, L. F. A. (2016). A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome Biology*, *17*(1), 261. https://doi.org/10.1186/s13059-016-1114-x

Carlson, S. M., & Gozani, O. (2016). Nonhistone Lysine Methylation in the Regulation of Cancer Pathways. *Cold Spring Harbor Perspectives in Medicine*, *6*(11). https://doi.org/10.1101/CSHPERSPECT.A026435

Carter, H., Chen, S., Isik, L., Tyekucheva, S., Velculescu, V. E., Kinzler, K. W., Vogelstein, B., & Karchin, R. (2009). Cancer-specific high-throughput annotation of somatic mutations: Computational prediction of driver missense mutations. *Cancer Research*, *69*(16), 6660–6667. https://doi.org/10.1158/0008-5472.CAN-09-1133

Carter, H., Samayoa, J., Hruban, R. H., & Karchin, R. (2010). Prioritization of driver mutations in pancreatic cancer using cancer-specific high-throughput annotation of somatic mutations (CHASM). *Cancer Biology & Therapy*, *10*(6), 582–587. https://doi.org/10.4161/CBT.10.6.12537

Chamberlain, L. H., & Shipston, M. J. (2015). The Physiology of Protein S-acylation. *Physiological Reviews*, *95*(2), 341. https://doi.org/10.1152/PHYSREV.00032.2014

Chan, P., Han, X., Zheng, B., Deran, M., Yu, J., Jarugumilli, G. K., Deng, H., Pan, D., Luo, X., & Wu, X. (2016). Autopalmitoylation of TEAD proteins regulates transcriptional output of the Hippo pathway. *Nature Chemical Biology*, *12*(4), 282–289. https://doi.org/10.1038/NCHEMBIO.2036

Cherbas, L., Willingham, A., Zhang, D., Yang, L., Zou, Y., Eads, B. D., Carlson, J. W., Landolin, J. M., Kapranov, P., Dumais, J., Samsonova, A., Choi, J. H., Roberts, J., Davis, C. A., Tang, H., Van Baren, M. J., Ghosh, S., Dobin, A., Bell, K., … Cherbas, P. (2011). The transcriptional diversity of 25 Drosophila cell lines. *Genome Research*, *21*(2), 301–314. https://doi.org/10.1101/GR.112961.110

Collins, D. W., & Jukes, T. H. (1994). Rates of Transition and Transversion in Coding Sequences since the Human-Rodent Divergence. *Genomics*, *20*(3), 386–396. https://doi.org/10.1006/GENO.1994.1192

Combe, C. W., Sivade, M., Hermjakob, H., Heimbach, J., Meldal, B. H. M., Micklem, G., Orchard, S., & Rappsilber, J. (2017). ComplexViewer: Visualization of curated macromolecular complexes. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btx497

Cooper, D. N., & Youssoufian, H. (1988). The CpG dinucleotide and human genetic disease. *Human Genetics*, *78*(2), 151–155. https://doi.org/10.1007/BF00278187

Corvi, M. M., Soltys, C. L. M., & Berthiaume, L. G. (2001). Regulation of mitochondrial carbamoyl-phosphate synthetase 1 activity by active site fatty acylation. *The Journal of Biological Chemistry*, *276*(49), 45704–45712. https://doi.org/10.1074/JBC.M102766200

Coulondre, C., Miller, J. H., Farabaugh, P. J., & Gilbert, W. (1978). Molecular basis of base substitution hotspots in Escherichia coli. *Nature*, *274*(5673), 775–780. https://doi.org/10.1038/274775A0

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J., Sander, C., Raphael, B. J., Marks, D. S., Francis Ouellette, B. F., Valencia, A., Bader, G. D., Boutros, P. C., Stuart, Joshua M, Linding, R., Lopez-Bigas, N., & Stein, L. D. (2015). Pathway and network analysis of cancer genomes. *Nature Methods*, *2*(3), 1–6. https://doi.org/10.1038/NMETH

Creixell, P., Schoof, E. M., Simpson, C. D., Longden, J., Miller, C. J., Lou, H. J., Perryman, L., Cox, T. R., Zivanovic, N., Palmeri, A., Wesolowska-Andersen, A., Helmer-Citterich, M., Ferkinghoff-Borg, J., Itamochi, H., Bodenmiller, B., Erler, J. T., Turk, B. E., & Linding, R. (2015). Kinome-wide Decoding of Network-Attacking Mutations Rewiring Cancer Signaling. *Cell*, *163*(1), 202–217. https://doi.org/10.1016/j.cell.2015.08.056

Creixell, P., Schoof, E. M., Tan, C. S. H., & Linding, R. (2012). Mutational properties of amino acid residues: implications for evolvability of phosphorylatable residues. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1602), 2584. https://doi.org/10.1098/RSTB.2012.0076

Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., & Gibson, T. J. (2012). Attributes of short linear motifs. *Molecular BioSystems*. https://doi.org/10.1039/c1mb05231d

David, A., Razali, R., Wass, M. N., & Sternberg, M. J. E. (2012). Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Human Mutation*, *33*(2), 359–363. https://doi.org/10.1002/HUMU.21656

de Beer, T. a P., Laskowski, R. a., Parks, S. L., Sipos, B., Goldman, N., & Thornton, J. M. (2013). Amino Acid Changes in Disease-Associated Variants Differ Radically from Variants Observed in the 1000 Genomes Project Dataset. *PLoS Computational Biology*, *9*(12). https://doi.org/10.1371/journal.pcbi.1003382

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., Mooney, T. B., Callaway, M. B., Dooling, D., Mardis, E. R., Wilson, R. K., & Ding, L. (2012). MuSiC: Identifying mutational significance in cancer genomes. *Genome Research*, *22*(8), 1589–1598. https://doi.org/10.1101/gr.134635.111

Delint-Ramirez, I., Willoughby, D., Hammond, G. V. R., Ayling, L. J., & Cooper, D. M. F. (2011). Palmitoylation Targets AKAP79 Protein to Lipid Rafts and Promotes Its Regulation of Calcium-sensitive Adenylyl Cyclase Type 8. *The Journal of Biological Chemistry*, *286*(38), 32962. https://doi.org/10.1074/JBC.M111.243899

Deng, M., Mehta, S., Sun, F., & Chen, T. (2002). Inferring domain-domain interactions from protein-protein interactions. *Genome Research*. https://doi.org/10.1101/gr.153002

Dinkel, H., Michael, S., Weatheritt, R. J., Davey, N. E., Van Roey, K., Altenberg, B., Toedt, G., Uyar, B., Seiler, M., Budd, A., Jödicke, L., Dammert, M. A., Schroeter, C., Hammer, M., Schmidt, T., Jehl, P., McGuigan, C., Dymecka, M., Chica, C., … Gibson, T. J. (2012). ELM--the database of eukaryotic linear motifs. *Nucleic Acids Research*, *40*(Database issue), D242-51. https://doi.org/10.1093/nar/gkr1064

Diwan, G. D., Gonzalez-Sanchez, J. C., Apic, G., & Russell, R. B. (2021). Next Generation Protein Structure Predictions and Genetic Variant Interpretation. *Journal of Molecular Biology*, *433*(20), 167180. https://doi.org/10.1016/J.JMB.2021.167180

Dou, Y., Gold, H. D., Luquette, L. J., & Park, P. J. (2018). Detecting somatic mutations in normal cells. *Trends in Genetics : TIG*, *34*(7), 545. https://doi.org/10.1016/J.TIG.2018.04.003

Drisdel, R. C., & Green, W. N. (2004). Labeling and quantifying sites of protein palmitoylation. *BioTechniques*, *36*(2), 276–285. https://doi.org/10.2144/04362RR02

Duncan, J. A., & Gilman, A. G. (1998). A cytoplasmic acyl-protein thioesterase that removes palmitate from G protein alpha subunits and p21(RAS). *The Journal of Biological Chemistry*, *273*(25), 15830–15837. https://doi.org/10.1074/JBC.273.25.15830

Edmonds, M. J., & Morgan, A. (2014). A systematic analysis of protein palmitoylation in Caenorhabditis elegans. *BMC Genomics*, *15*(1), 1–16. https://doi.org/10.1186/1471-2164-15-841/FIGURES/5

Eichhorn, P. J. A., Creyghton, M. P., & Bernards, R. (2009). Protein phosphatase 2A regulatory subunits and cancer. *Biochimica et Biophysica Acta*, *1795*(1), 1–15. https://doi.org/10.1016/J.BBCAN.2008.05.005

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky995

Elliott, K., & Larsson, E. (2021). Non-coding driver mutations in human cancer. *Nature Reviews. Cancer*, *21*(8), 500–509. https://doi.org/10.1038/S41568-021-00371-Z

Fajac, I., Viel, M., Sublemontier, S., Hubert, D., & Bienvenu, T. (2008). Could a defective epithelial sodium channel lead to bronchiectasis. *Respiratory Research*. https://doi.org/10.1186/1465-9921-9-46

Federici, G., & Soddu, S. (2020). Variants of uncertain significance in the era of high-throughput genome sequencing: A lesson from breast and ovary cancers. *Journal of Experimental and Clinical Cancer Research*, *39*(1), 1–12. https://doi.org/10.1186/S13046-020-01554-6/FIGURES/1

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., & Bateman, A. (2016). The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, *44*(D1), D279–D285. https://doi.org/10.1093/nar/gkv1344

Finn, R. D., Miller, B. L., Clements, J., & Bateman, A. (2014). IPfam: A database of protein family and domain interactions found in the Protein Data Bank. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkt1210

Fontana, G. A., Hess, D., Reinert, J. K., Mattarocci, S., Falquet, B., Klein, D., Shore, D., Thomä, N. H., & Rass, U. (2019). Rif1 S-acylation mediates DNA double-strand break repair at the inner nuclear membrane. *Nature Communications*, *10*(1). https://doi.org/10.1038/S41467-019-10349-Z

Forrester, M. T., Hess, D. T., Thompson, J. W., Hultman, R., Moseley, M. A., Stamler, J. S., & Casey, P. J. (2011). Site-specific analysis of protein S-acylation by resin-assisted capture. *Journal of Lipid Research*. https://doi.org/10.1194/jlr.D011106

Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O., & Bader, G. D. (2015). Cytoscape.js: A graph theory library for visualisation and analysis. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btv557

Franz, M., Rodriguez, H., Lopes, C., Zuberi, K., Montojo, J., Bader, G. D., & Morris, Q. (2018). GeneMANIA update 2018. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky311

Fujimoto, A., Okada, Y., Boroevich, K. A., Tsunoda, T., Taniguchi, H., & Nakagawa, H. (2016). Systematic analysis of mutation distribution in three dimensional protein structures identifies cancer driver genes. *Scientific Reports*, *6*(1), 26483. https://doi.org/10.1038/srep26483

Fukata, M., Fukata, Y., Adesnik, H., Nicoll, R. A., & Bredt, D. S. (2004). Identification of PSD-95 palmitoylating enzymes. *Neuron*, *44*(6), 987–996. https://doi.org/10.1016/J.NEURON.2004.12.005

Fukata, Y., & Fukata, M. (2010). Protein palmitoylation in neuronal development and synaptic plasticity. *Nature Reviews. Neuroscience*, *11*(3), 161–175. https://doi.org/10.1038/NRN2788

Furuhashi, M., Kitamura, K., Adachi, M., Miyoshi, T., Wakida, N., Ura, N., Shikano, Y., Shinshi, Y., Sakamoto, K. I., Hayashi, M., Satoh, N., Nishitani, T., Tomita, K., & Shimamoto, K. (2005). Liddle's syndrome caused by a novel mutation in the proline-rich PY motif of the epithelial sodium channel β-subunit. *Journal of Clinical Endocrinology and Metabolism*. https://doi.org/10.1210/jc.2004-1027

Gao, J., Chang, M. T., Johnsen, H. C., Gao, S. P., Sylvester, B. E., Sumer, S. O., Zhang, H., Solit, D. B., Taylor, B. S., Schultz, N., & Sander, C. (2017). 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Medicine*, *9*(1), 4. https://doi.org/10.1186/s13073-016-0393-x

Garty, H. (1994). Molecular properties of epithelial, amiloride-blockable Na+ channels. *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*, *8*(8), 522–528. https://doi.org/10.1096/FASEBJ.8.8.8181670

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N. Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C. R., … Swanton, C. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *The New England Journal of Medicine*, *366*(10), 883–892. https://doi.org/10.1056/NEJMOA1113205

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., & Tegnér, J. (2014). Data integration in the era of omics: current and future challenges. *BMC Systems Biology*, *8 Suppl 2*(Suppl 2), I1. https://doi.org/10.1186/1752-0509-8-S2-I1

Gong, S. J., Feng, X. J., Song, W. H., Chen, J. M., Wang, S. M., Xing, D. J., Zhu, M. H., Zhang, S. H., & Xu, A. M. (2016). Upregulation of PP2Ac predicts poor prognosis and contributes to aggressiveness in hepatocellular carcinoma. *Cancer Biology & Therapy*, *17*(2), 151. https://doi.org/10.1080/15384047.2015.1121345

Gonzalez-Perez, A., & Lopez-Bigas, N. (2012). Functional impact bias reveals cancer drivers. *Nucleic Acids Research*, *40*(21). https://doi.org/10.1093/nar/gks743

González-Sánchez, J. C., Ibrahim, M. F. R., Leist, I. C., Weise, K. R., & Russell, R. B. (2021). Mechnetor: a web server for exploring protein mechanism and the functional context of genetic variants. *Nucleic Acids Research*, *49*(W1), W366–W374. https://doi.org/10.1093/NAR/GKAB399

González-Sánchez, J. C., Raimondi, F., & Russell, R. B. (2018). Cancer genetics meets biomolecular mechanism—bridging an age-old gulf. In *FEBS Letters* (Vol. 592, Issue 4, pp. 463–474). https://doi.org/10.1002/1873-3468.12988

Goodwin, J. S., Drake, K. R., Rogers, C., Wright, L., Lippincott-Schwartz, J., Philips, M. R., & Kenworthy, A. K. (2005).

Depalmitoylated Ras traffics to and from the Golgi complex via a nonvesicular pathway. *The Journal of Cell Biology*, *170*(2), 261. https://doi.org/10.1083/JCB.200502063

Gottlieb, C. D., & Linder, M. E. (2017). Structure and function of DHHC protein S-acyltransferases. *Biochemical Society Transactions*, *45*(4), 923–938. https://doi.org/10.1042/BST20160304

Greaves, J., & Chamberlain, L. H. (2011). DHHC palmitoyl transferases: Substrate interactions and (patho)physiology. In *Trends in Biochemical Sciences*. https://doi.org/10.1016/j.tibs.2011.01.003

Greaves, J., Salaun, C., Fukata, Y., Fukata, M., & Chamberlain, L. H. (2008). Palmitoylation and membrane interactions of the neuroprotective chaperone cysteine-string protein. *The Journal of Biological Chemistry*, *283*(36), 25014–25026. https://doi.org/10.1074/JBC.M802140200

Groves, M. R., Hanlon, N., Turowski, P., Hemmings, B. A., & Barford, D. (1999). The Structure of the Protein Phosphatase 2A PR65/A Subunit Reveals the Conformation of Its 15 Tandemly Repeated HEAT Motifs. *Cell*, *96*(1), 99–110. https://doi.org/10.1016/S0092-8674(00)80963-0

Hanahan, D., & Weinberg, R. A. (2000). The Hallmarks of Cancer. *Cell*, *100*(1), 57–70. https://doi.org/10.1016/S0092-8674(00)81683-9

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hanukoglu, I., & Hanukoglu, A. (2016). Epithelial sodium channel (ENaC) family: Phylogeny, structure-function, tissue distribution, and associated inherited diseases. In *Gene*. https://doi.org/10.1016/j.gene.2015.12.061

Havugimana, P. C., Hart, G. T., Nepusz, T., Yang, H., Turinsky, A. L., Li, Z., Wang, P. I., Boutz, D. R., Fong, V., Phanse, S., Babu, M., Craig, S. A., Hu, P., Wan, C., Vlasblom, J., Dar, V. U. N., Bezginov, A., Clark, G. W., Wu, G. C., … Emili, A. (2012). A census of human soluble protein complexes. *Cell*. https://doi.org/10.1016/j.cell.2012.08.011

Hemmings, B. A., Adams-Pearson, C., Maurer, F., Müller, P., Goris, J., Merlevede, W., Hofsteenge, J., & Stone, S. R. (1990). α- and βForms of the 65-kDa Subunit of Protein Phosphatase 2A Have a Similar 39 Amino Acid Repeating Structure. *Biochemistry*, *29*(13), 3166–3173. https://doi.org/10.1021/bi00465a002

Henikoff, S., & Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(22), 10915–10919. https://doi.org/10.1073/PNAS.89.22.10915

Higashiguchi, M., Nagatomo, I., Kijima, T., Morimura, O., Miyake, K., Minami, T., Koyama, S., Hirata, H., Iwahori, K., Takimoto, T., Takeda, Y., Kida, H., & Kumanogoh, A. (2016). Clarifying the biological significance of the CHK2 K373E somatic mutation discovered in The Cancer Genome Atlas database. *FEBS Letters*, *590*(23), 4275–4286. https://doi.org/10.1002/1873-3468.12449

Horn, S., Figl, A., Rachakonda, P. S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., Schadendorf, D., & Kumar, R. (2013). TERT promoter mutations in familial and sporadic melanoma. *Science (New York, N.Y.)*, *339*(6122), 959–961. https://doi.org/10.1126/SCIENCE.1230062

Hornbeck, P. V., Zhang, B., Murray, B., Kornhauser, J. M., Latham, V., & Skrzypek, E. (2015). PhosphoSitePlus, 2014: Mutations, PTMs and recalibrations. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gku1267

Hou, H., John Peter, A. T., Meiringer, C., Subramanian, K., & Ungermann, C. (2009). Analysis of DHHC acyltransferases implies overlapping substrate specificity and a two-step reaction mechanism. *Traffic (Copenhagen, Denmark)*, *10*(8), 1061–1073. https://doi.org/10.1111/J.1600-0854.2009.00925.X

Houge, G., Haesen, D., Vissers, L. E. L. M., Mehta, S., Parker, M. J., Wright, M., Vogt, J., McKee, S., Tolmie, J. L., Cordeiro, N., Kleefstra, T., Willemsen, M. H., Reijnders, M. R. F., Berland, S., Hayman, E., Lahat, E., Brilstra, E. H., Van Gassen, K. L. I., Zonneveld-Huijssoon, E., … Janssens, V. (2015). B56δ-related protein phosphatase 2A dysfunction identified in patients with intellectual disability. *The Journal of Clinical Investigation*, *125*(8), 3051. https://doi.org/10.1172/JCI79860

Howie, J., Reilly, L., Fraser, N. J., Walker, J. M. V., Wypijewski, K. J., Ashford, M. L. J., Calaghan, S. C., McClafferty, H., Tian, L., Shipston, M. J., Boguslavskyi, A., Shattock, M. J., & Fuller, W. (2014). Substrate recognition by the cell surface palmitoyl transferase DHHC5. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(49), 17534–17539. https://doi.org/10.1073/PNAS.1413627111/SUPPL_FILE/PNAS.1413627111.SFIG03.PDF

Hu, Y., Flockhart, I., Vinayagam, A., Bergwitz, C., Berger, B., Perrimon, N., & Mohr, S. E. (2011). An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-12-357

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, *14*(1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5

Hunt, M. C., & Alexson, S. E. H. (2002). The role Acyl-CoA thioesterases play in mediating intracellular lipid metabolism. *Progress in Lipid Research*, *41*(2), 99–130. https://doi.org/10.1016/S0163-7827(01)00017-0

Huttlin, E. L., Ting, L., Bruckner, R. J., Gebreab, F., Gygi, M. P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., Dong, R., Guarani, V., Vaites, L. P., Ordureau, A., Rad, R., Erickson, B. K., Wühr, M., Chick, J., Zhai, B., … Gygi, S. P. (2015). The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*. https://doi.org/10.1016/j.cell.2015.06.043

Imbeaud, S., Ladeiro, Y., & Zucman-Rossi, J. (2010). Identification of novel oncogenes and tumor suppressors in hepatocellular carcinoma. *Seminars in Liver Disease*, *30*(1), 75–86. https://doi.org/10.1055/S-0030-1247134/ID/44

Jia, L., Chisari, M., Maktabi, M. H., Sobieski, C., Zhou, H., Konopko, A. M., Martin, B. R., Mennerick, S. J., & Blumer, K. J. (2014). A mechanism regulating G protein-coupled receptor signaling that requires cycles of protein palmitoylation and depalmitoylation. *The Journal of Biological Chemistry*, *289*(9), 6249–6257. https://doi.org/10.1074/JBC.M113.531475

Johswich, A., Kraft, B., Wuhrer, M., Berger, M., Deelder, A. M., Hokke, C. H., Gerardy-Schahn, R., & Bakker, H. (2009). Golgi targeting of Drosophila melanogaster β4GalNAcTB requires a DHHC protein family–related protein as a pilot. *The Journal of Cell Biology*, *184*(1), 173. https://doi.org/10.1083/JCB.200801071

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577–2637. https://doi.org/10.1002/BIP.360221211

Kamburov, A., Lawrence, M. S., Polak, P., Leshchiner, I., Lage, K., Golub, T. R., Lander, E. S., & Getz, G. (2015). Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(40), E5486-95. https://doi.org/10.1073/pnas.1516373112

Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., Leiserson, M. D. M., Miller, C. A., Welch, J. S., Walter, M. J., Wendl, M. C., Ley, T. J., Wilson, R. K., Raphael, B. J., & Ding, L. (2013). Mutational landscape and significance across 12 major cancer types. *Nature 2013 502:7471*, *502*(7471), 333–339. https://doi.org/10.1038/nature12634

Kang, K. H., & Bier, E. (2010). dHIP14-dependent palmitoylation promotes secretion of the BMP antagonist Sog. *Developmental Biology*, *346*(1), 1–10. https://doi.org/10.1016/J.YDBIO.2010.06.024

Kazemi-Sefat, G. E., Keramatipour, M., Talebi, S., Kavousi, K., Sajed, R., Kazemi-Sefat, N. A., & Mousavizadeh, K. (2021). The importance of CDC27 in cancer: molecular pathology and clinical aspects. *Cancer Cell International*, *21*(1), 1–11. https://doi.org/10.1186/S12935-021-01860-9/FIGURES/3

Khuong-Quang, D. A., Buczkowicz, P., Rakopoulos, P., Liu, X. Y., Fontebasso, A. M., Bouffet, E., Bartels, U., Albrecht, S., Schwartzentruber, J., Letourneau, L., Bourgey, M., Bourque, G., Montpetit, A., Bourret, G., Lepage, P., Fleming, A., Lichter, P., Kool, M., Von Deimling, A., … Hawkins, C. (2012). K27M mutation in histone H3.3 defines clinically and biologically distinct subgroups of pediatric diffuse intrinsic pontine gliomas. *Acta Neuropathologica*, *124*(3), 439. https://doi.org/10.1007/S00401-012-0998-0

Kim, S., & Jeong, S. (2019). Mutation Hotspots in the β-Catenin Gene: Lessons from the Human Cancer Genome Databases. *Molecules and Cells*, *42*(1), 8–16. https://doi.org/10.14348/MOLCELLS.2018.0436

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Ko, P.-J., & Dixon, S. J. (2018). Protein palmitoylation and cancer. *EMBO Reports*, *19*(10), e46666. https://doi.org/10.15252/EMBR.201846666

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, *155*(1), 27. https://doi.org/10.1016/J.CELL.2013.09.006

Kong, E., Peng, S., Chandra, G., Sarkar, C., Zhang, Z., Bagh, M. B., & Mukherjee, A. B. (2013). Dynamic palmitoylation links cytosol-membrane shuttling of acyl-protein thioesterase-1 and acyl-protein thioesterase-2 with that of proto-oncogene H-ras product and growth-associated protein-43. *The Journal of Biological Chemistry*, *288*(13), 9112–9125. https://doi.org/10.1074/JBC.M112.421073

Krassowski, M., Pellegrina, D., Mee, M. W., Fradet-Turcotte, A., Bhat, M., & Reimand, J. (2021). ActiveDriverDB: Interpreting Genetic Variation in Human and Cancer Genomes Using Post-translational Modification Sites and Signaling Networks (2021 Update). *Frontiers in Cell and Developmental Biology*, *9*, 626821. https://doi.org/10.3389/FCELL.2021.626821

Krogh, A., Larsson, B., Von Heijne, G., & Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology*, *305*(3), 567–580. https://doi.org/10.1006/jmbi.2000.4315

Kucukkal, T. G., Petukh, M., Li, L., & Alexov, E. (2015). Structural and Physico-Chemical Effects of Disease and Non-Disease nsSNPs on Proteins. *Current Opinion in Structural Biology*, *32*, 18. https://doi.org/10.1016/J.SBI.2015.01.003

Kumar, M., Gouw, M., Michael, S., Sámano-Sánchez, H., Pancsa, R., Glavina, J., Diakogianni, A., Valverde, J. A., Bukirova, D., Čalyševa, J., Palopoli, N., Davey, N. E., Chemes, L. B., & Gibson, T. J. (2019). ELM—the eukaryotic linear motif resource in 2020. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz1030

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols*, *4*(7), 1073–1082. https://doi.org/10.1038/nprot.2009.86

Kumari, B., Kumar, R., & Kumar, M. (2014). PalmPred: an SVM based palmitoylation prediction method using sequence profile information. *PloS One*, *9*(2). https://doi.org/10.1371/JOURNAL.PONE.0089246

Larkin, A., Marygold, S. J., Antonazzo, G., Attrill, H., dos Santos, G., Garapati, P. V, Goodman, J. L., Gramates, L. S., Millburn, G., Strelets, V. B., Tabone, C. J., Thurmond, J., Consortium, F., Perrimon, N., Gelbart, S. R., Agapite, J., Broll, K., Crosby, M., dos Santos, G., … Lovato, T. (2021). FlyBase: updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Research*, *49*(D1), D899–D907. https://doi.org/10.1093/NAR/GKAA1026

Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S., & Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, *505*(7484), 495–501. https://doi.org/10.1038/nature12912

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V, Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., … Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, *499*(7457), 214–218. https://doi.org/10.1038/nature12213

Lee, E. Y. H. P., & Muller, W. J. (2010). Oncogenes and Tumor Suppressor Genes. *Cold Spring Harbor Perspectives in Biology*, *2*(10). https://doi.org/10.1101/CSHPERSPECT.A003236

Lemonidis, K., Sanchez-Perez, M. C., & Chamberlain, L. H. (2015). Identification of a novel sequence motif recognized by the ankyrin repeat domain of zDHHC17/13 S-acyltransferases. *Journal of Biological Chemistry*, *290*(36), 21939–21950. https://doi.org/10.1074/jbc.M115.657668

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D. J., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J. F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., … Vidal, M. (2004). A Map of the Interactome Network of the Metazoan C. elegans. *Science*. https://doi.org/10.1126/science.1091403

Lin, D. T. S., & Conibear, E. (2015). ABHD17 proteins are novel protein depalmitoylases that regulate N-Ras palmitate turnover and subcellular localization. *ELife*, *4*(DECEMBER2015). https://doi.org/10.7554/ELIFE.11306

Linding, R., Jensen, L. J., Ostheimer, G. J., van Vugt, M. A. T. M., Jørgensen, C., Miron, I. M., Diella, F., Colwill, K., Taylor, L., Elder, K., Metalnikov, P., Nguyen, V., Pasculescu, A., Jin, J., Park, J. G., Samson, L. D., Woodgett, J. R., Russell, R. B. B., Bork, P., … Pawson, T. (2007). Systematic discovery of in vivo phosphorylation networks. *Cell*, *129*(7), 1415–1426. https://doi.org/10.1016/J.CELL.2007.05.052

Liu, X., Salokas, K., Tamene, F., Jiu, Y., Weldatsadik, R. G., Öhman, T., & Varjosalo, M. (2018). An AP-MS- and BioID-compatible MAC-tag enables comprehensive mapping of protein interactions and subcellular localizations. *Nature Communications*, *9*(1). https://doi.org/10.1038/S41467-018-03523-2

Luck, K., Kim, D. K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F. J., Charloteaux, B., Choi, D., Coté, A. G., Daley, M., Deimling, S., Desbuleux, A., Dricot, A., Gebbia, M., Hardy, M. F., Kishore, N., … Calderwood, M. A. (2020). A reference map of the human binary protein interactome. *Nature 2020 580:7803*, *580*(7803), 402–408. https://doi.org/10.1038/s41586-020-2188-x

MacDonald, B. T., Tamai, K., & He, X. (2009). Wnt/β-catenin signaling: components, mechanisms, and diseases. *Developmental*

Cell, *17*(1), 9. https://doi.org/10.1016/J.DEVCEL.2009.06.016

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D., & Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkz268

Malone, E. R., Oliva, M., Sabatini, P. J. B., Stockley, T. L., & Siu, L. L. (2020). Molecular profiling for precision cancer therapies. *Genome Medicine 2020 12:1*, *12*(1), 1–19. https://doi.org/10.1186/S13073-019-0703-1

Martelli, P. L., Fariselli, P., Savojardo, C., Babbi, G., Aggazio, F., & Casadio, R. (2016). Large scale analysis of protein stability in OMIM disease related human protein variants. *BMC Genomics*, *17*(Suppl 2). https://doi.org/10.1186/S12864-016-2726-Y

Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., Davies, H., Stratton, M. R., & Campbell, P. J. (2017). Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, *171*(5), 1029. https://doi.org/10.1016/J.CELL.2017.09.042

Matakatsu, H., & Blair, S. S. (2008). The DHHC palmitoyltransferase approximated regulates Fat signaling and Dachs localization and activity. *Current Biology : CB*, *18*(18), 1390–1395. https://doi.org/10.1016/J.CUB.2008.07.067

Mertens, F., Johansson, B., Fioretos, T., & Mitelman, F. (2015). The emerging complexity of gene fusions in cancer. *Nature Reviews. Cancer*, *15*(6), 371–381. https://doi.org/10.1038/NRC3947

Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. D. (2019). PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky1038

Miller, M. L., Reznik, E., Gauthier, N. P., Ciriello, G., Schultz, N., Miller, M. L., Reznik, E., Gauthier, N. P., Aksoy, A., Korkut, A., & Gao, J. (2015). *Pan-Cancer Analysis of Mutation Hotspots in Protein Article Pan-Cancer Analysis of Mutation Hotspots in Protein Domains*. 197–209. https://doi.org/10.1016/j.cels.2015.08.014

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Mitchell, D. A., Vasudevan, A., Linder, M. E., & Deschenes, R. J. (2006). Protein palmitoylation by a family of DHHC protein S-acyltransferases. *Journal of Lipid Research*, *47*(6), 1118–1127. https://doi.org/10.1194/JLR.R600007-JLR200

Morrow, I. C., Rea, S., Martin, S., Prior, I. A., Prohaska, R., Hancock, J. F., James, D. E., & Parton, R. G. (2002). Flotillin-1/reggie-2 traffics to surface raft domains via a novel golgi-independent pathway. Identification of a novel membrane targeting domain and a role for palmitoylation. *The Journal of Biological Chemistry*, *277*(50), 48834–48841. https://doi.org/10.1074/JBC.M209082200

Mosca, R., Céol, A., & Aloy, P. (2013). Interactome3D: Adding structural details to protein networks. *Nature Methods*. https://doi.org/10.1038/nmeth.2289

Mosca, R., Céol, A., Stein, A., Olivella, R., & Aloy, P. (2014). 3did: A catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkt887

Mosca, R., Tenorio-Laranga, J., Olivella, R., Alcalde, V., Céol, A., Soler-López, M., & Aloy, P. (2015). dSysMap: Exploring the edgetic role of disease mutations. In *Nature Methods*. https://doi.org/10.1038/nmeth.3289

Muszbek, L., Haramura, G., Cluette-Brown, J. E., Van Cott, E. M., & Laposata, M. (1999). The pool of fatty acids covalently bound to platelet proteins by thioester linkages can be altered by exogenously supplied fatty acids. *Lipids*, *34 Suppl*(1). https://doi.org/10.1007/BF02562334

Nagendra, D. C., Burke, J., Maxwell, G. L., & Risinger, J. I. (2012). PPP2R1A mutations are common in the serous type of endometrial cancer. *Molecular Carcinogenesis*, *51*(10), 826–831. https://doi.org/10.1002/MC.20850

Nagy, R., Sweet, K., & Eng, C. (2004). Highly penetrant hereditary cancer syndromes. *Oncogene 2004 23:38*, *23*(38), 6445–6470. https://doi.org/10.1038/sj.onc.1207714

Narayan, S., Bader, G. D., & Reimand, J. (2016). Frequent mutations in acetylation and ubiquitination sites suggest novel driver mechanisms of cancer. *Genome Medicine*, *8*(1). https://doi.org/10.1186/s13073-016-0311-2

Neduva, V., & Russell, R. B. (2005). Linear motifs: Evolutionary interaction switches. In *FEBS Letters*. https://doi.org/10.1016/j.febslet.2005.04.005

Neduva, V., & Russell, R. B. (2006). DILIMOT: Discovery of linear motifs in proteins. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkl159

Ng, S., Collisson, E. A., Sokolov, A., Goldstein, T., onzalez-Perez, A., Lopez-Bigas, N., Benz, C., Haussler, D., & Stuart, J. M. (2012). PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics*, *28*(18). https://doi.org/10.1093/bioinformatics/bts402

Nordström, K. J. V., Sällman Almén, M., Edstam, M. M., Fredriksson, R., & Schiöth, H. B. (2011). Independent HHsearch, Needleman--Wunsch-based, and motif analyses reveal the overall hierarchy for most of the G protein-coupled receptor families. *Molecular Biology and Evolution*, *28*(9), 2471–2480. https://doi.org/10.1093/MOLBEV/MSR061

O'Donoghue, S. I., Baldi, B. F., Clark, S. J., Darling, A. E., Hogan, J. M., Kaur, S., Maier-Hein, L., McCarthy, D. J., Moore, W. J., Stenau, E., Swedlow, J. R., Vuong, J., & Procter, J. B. (2018). Visualization of Biomedical Data. *Annual Review of Biomedical Data Science*. https://doi.org/10.1146/annurev-biodatasci-080917-013424

O'Donoghue, S. I., Gavin, A. C., Gehlenborg, N., Goodsell, D. S., Hériché, J. K., Nielsen, C. B., North, C., Olson, A. J., Procter, J. B., Shattuck, D. W., Walter, T., & Wong, B. (2010). Visualizing biological data—now and in the future. In *Nature Methods*. https://doi.org/10.1038/nmeth.f.301

O'Sullivan, B. P., & Freedman, S. D. (2009). Cystic fibrosis. *The Lancet*, *373*(9678), 1891–1904. https://doi.org/10.1016/S0140-6736(09)60327-5

Ohno, Y., Kashio, A., Ogata, R., Ishitomi, A., Yamazaki, Y., & Kihara, A. (2012). Analysis of substrate specificity of human DHHC protein acyltransferases using a yeast expression system. *Molecular Biology of the Cell*, *23*(23), 4543. https://doi.org/10.1091/MBC.E12-05-0336

Ohno, Y., Kihara, A., Sano, T., & Igarashi, Y. (2006). Intracellular localization and tissue-specific distribution of human and yeast DHHC cysteine-rich domain-containing proteins. *Biochimica et Biophysica Acta*, *1761*(4), 474–483. https://doi.org/10.1016/J.BBALIP.2006.03.010

Ohyama, T., Verstreken, P., Ly, C. V., Rosenmund, T., Rajan, A., Tien, A. C., Haueter, C., Schulze, K. L., & Bellen, H. J. (2007). Huntingtin-interacting protein 14, a palmitoyl transferase required for exocytosis and targeting of CSP to synaptic vesicles. *The Journal of Cell Biology*, *179*(7), 1481–1496. https://doi.org/10.1083/JCB.200710061

Okamoto, K., Li, H., Jensen, M. R., Zhang, T., Taya, Y., Thorgeirsson, S. S., & Prives, C. (2002). Cyclin G recruits PP2A to dephosphorylate Mdm2. *Molecular Cell*, *9*(4), 761–771. https://doi.org/10.1016/S1097-2765(02)00504-X

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., … Hermjakob, H. (2014). The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, *42*(Database issue), D358-63. https://doi.org/10.1093/nar/gkt1115

Oser, M. G., Niederst, M. J., Sequist, L. V., & Engelman, J. A. (2015). Transformation from non-small-cell lung cancer to small-cell lung cancer: molecular drivers and cells of origin. *The Lancet Oncology*, *16*(4), e165–e172. https://doi.org/10.1016/S1470-2045(14)71180-5

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B. J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., Dolma, S., Coulombe-Huntington, J., Chatr-aryamontri, A., Dolinski, K., & Tyers, M. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, *30*(1), 187–200. https://doi.org/10.1002/pro.3978

Pandit, B., Sarkozy, A., Pennacchio, L. A., Carta, C., Oishi, K., Martinelli, S., Pogna, E. A., Schackwitz, W., Ustaszewska, A., Landstrom, A., Bos, J. M., Ommen, S. R., Esposito, G., Lepri, F., Faul, C., Mundel, P., López Siguero, J. P., Tenconi, R., Selicorni, A., … Gelb, B. D. (2007). Gain-of-function RAF1 mutations cause Noonan and LEOPARD syndromes with hypertrophic cardiomyopathy. *Nature Genetics*, *39*(8), 1007–1012. https://doi.org/10.1038/NG2073

Park, S. Y., Gönen, M., Kim, H. J., Michor, F., & Polyak, K. (2010). Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *The Journal of Clinical Investigation*, *120*(2), 636. https://doi.org/10.1172/JCI40724

Perkins, B. A., Caskey, C. T., Brar, P., Dec, E., Karow, D. S., Kahn, A. M., Hou, Y. C. C., Shah, N., Boeldt, D., Coughlin, E., Hands, G., Lavrenko, V., Yu, J., Procko, A., Appis, J., Dale, A. M., Guo, L., Jönsson, T. J., Wittmann, B. M., … Venter, J. C. (2018). Precision medicine screening using whole-genome sequencing and advanced imaging to identify disease risk in adults. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(14), 3686–3691. https://doi.org/10.1073/PNAS.1706096114/SUPPL_FILE/PNAS.201706096SI.PDF

Perkins, J. R., Diboun, I., Dessailly, B. H., Lees, J. G., & Orengo, C. (2010). Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. In *Structure*. https://doi.org/10.1016/j.str.2010.08.007

Peterson, T. A., Gauran, I. I. M., Park, J., Park, D. H., & Kann, M. G. (2017). Oncodomains: A protein domain-centric framework for analyzing rare variants in tumor samples. *PLoS Computational Biology*, *13*(4). https://doi.org/10.1371/journal.pcbi.1005428

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605–1612. https://doi.org/10.1002/JCC.20084

Ponting, C. P., & Russell, R. R. (2002). The natural history of protein domains. In *Annual Review of Biophysics and Biomolecular Structure*. https://doi.org/10.1146/annurev.biophys.31.082901.134314

Porcellato, E., González-Sánchez, J. C., Ahlmann-Eltze, C., Elsakka, M. A., Shapira, I., Fritsch, J., Navarro, J. A., Anders, S., Russell, R. B., Wieland, F. T., & Metzendorf, C. (2022). The S-palmitoylome and DHHC-PAT interactome of Drosophila melanogaster S2R+ cells indicate a high degree of conservation to mammalian palmitoylomes. *PLOS ONE*, *17*(8), e0261543. https://doi.org/10.1371/JOURNAL.PONE.0261543

Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J., & Godzik, A. (2015). A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. *PLoS Computational Biology*, *11*(10). https://doi.org/10.1371/journal.pcbi.1004518

Raimondi, F., Betts, M. J., Lu, Q., Inoue, A., Gutkind, J. S., & Russell, R. B. (2017). Genetic variants affecting equivalent protein family positions reflect human diversity. *Scientific Reports*, *7*(1), 12771. https://doi.org/10.1038/s41598-017-12971-7

Raimondi, F., Inoue, A., Kadji, F. M. N., Shuai, N., Gonzalez, J. C., Singh, G., de la Vega, A. A., Sotillo, R., Fischer, B., Aoki, J., Gutkind, J. S., & Russell, R. B. (2019). Rare, functional, somatic variants in gene families linked to cancer genes: GPCR signaling as a paradigm. *Oncogene*. https://doi.org/10.1038/s41388-019-0895-2

Raimondi, F., Singh, G., Betts, M. J., Apic, G., Vukotic, R., Andreone, P., Stein, L., & Russell, R. B. (2016). Insights into cancer severity from biomolecular interaction mechanisms. *Scientific Reports*, *6*, 34490. https://doi.org/10.1038/srep34490

Rajagopala, S. V., Sikorski, P., Kumar, A., Mosca, R., Vlasblom, J., Arnold, R., Franca-Koh, J., Pakala, S. B., Phanse, S., Ceol, A., Häuser, R., Siszler, G., Wuchty, S., Emili, A., Babu, M., Aloy, P., Pieper, R., & Uetz, P. (2014). The binary protein-protein interaction landscape of escherichia coli. *Nature Biotechnology*. https://doi.org/10.1038/nbt.2831

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., & Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, *47*(W1), W191–W198. https://doi.org/10.1093/NAR/GKZ369

Reimand, J., & Bader, G. D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Molecular Systems Biology*, *9*(1), 637. https://doi.org/10.1038/msb.2012.68

Reimand, J., Wagih, O., & Bader, G. D. (2015). Evolutionary constraint and disease associations of post-translational modification sites in human genomes. *PLoS Genetics*, *11*(1). https://doi.org/10.1371/JOURNAL.PGEN.1004919

Ren, J., Wen, L., Gao, X., Jin, C., Xue, Y., & Yao, X. (2008). CSS-Palm 2.0: An updated software for palmitoylation sites prediction. *Protein Engineering, Design and Selection*. https://doi.org/10.1093/protein/gzn039

Reva, B., Antipin, Y., & Sander, C. (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. *Nucleic Acids Research*, *39*(17). https://doi.org/10.1093/nar/gkr407

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., & Rehm, H. L. (2015). Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *17*(5), 405. https://doi.org/10.1038/GIM.2015.30

Rivlin, N., Brosh, R., Oren, M., & Rotter, V. (2011). Mutations in the p53 tumor suppressor gene: Important milestones at the various steps of tumorigenesis. In *Genes and Cancer*. https://doi.org/10.1177/1947601911408889

Robinson, C. V., Sali, A., & Baumeister, W. (2007). The molecular sociology of the cell. *Nature*, *450*(7172), 973–982. https://doi.org/10.1038/NATURE06523

Rohde, M., Richter, J., Schlesner, M., Betts, M. J., Claviez, A., Bonn, B. R., Zimmermann, M., Damm-Welk, C., Russell, R. B., Borkhardt, A., Eils, R., Hoell, J. I., Szczepanowski, M., Oschlies, I., Klapper, W., Burkhardt, B., & Siebert, R. (2014).

Recurrent RHOA mutations in pediatric Burkitt lymphoma treated according to the NHL-BFM protocols. *Genes Chromosomes and Cancer*, *53*(11), 911–916. https://doi.org/10.1002/gcc.22202

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S. D., Yang, X., Ghamsari, L., Balcha, D., Begg, B. E., Braun, P., Brehme, M., Broly, M. P., … Vidal, M. (2014). A proteome-scale map of the human interactome network. *Cell*. https://doi.org/10.1016/j.cell.2014.10.050

Roth, A. F., Wan, J., Bailey, A. O., Sun, B., Kuchar, J. A., Green, W. N., Phinney, B. S., Yates, J. R., & Davis, N. G. (2006). Global Analysis of Protein Palmitoylation in Yeast. *Cell*, *125*(5), 1003. https://doi.org/10.1016/J.CELL.2006.03.042

Roux, K. J., Kim, D. I., Burke, B., & May, D. G. (2018). BioID: A Screen for Protein-Protein Interactions. *Current Protocols in Protein Science*, *91*(1), 19.23.1-19.23.15. https://doi.org/10.1002/CPPS.51

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J. I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G. I., Wang, Y., Kovács, I. A., Kamburov, A., Krykbaeva, I., Lam, M. H., Tucker, G., Khurana, V., Sharma, A., Liu, Y. Y., Yachie, N., … Vidal, M. (2015). Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders. *Cell*, *161*(3), 647. https://doi.org/10.1016/J.CELL.2015.04.013

Sanders, S. S., Martin, D. D. O., Butland, S. L., Lavallée-Adam, M., Calzolari, D., Kay, C., Yates, J. R., & Hayden, M. R. (2015). Curation of the Mammalian Palmitoylome Indicates a Pivotal Role for Palmitoylation in Diseases and Disorders of the Nervous System and Cancers. *PLOS Computational Biology*, *11*(8), e1004405. https://doi.org/10.1371/journal.pcbi.1004405

Savojardo, C., Manfredi, M., Martelli, P. L., & Casadio, R. (2021). Solvent Accessibility of Residues Undergoing Pathogenic Variations in Humans: From Protein Structures to Protein Sequences. *Frontiers in Molecular Biosciences*, *7*, 460. https://doi.org/10.3389/FMOLB.2020.626363/BIBTEX

Sayle, R. A., & Milner-White, E. J. (1995). RASMOL: biomolecular graphics for all. *Trends in Biochemical Sciences*, *20*(9), 374–376. https://doi.org/10.1016/S0968-0004(00)89080-5

Schmidt, M. F. G., & Schlesinger, M. J. (1979). Fatty acid binding to vesicular stomatitis virus glycoprotein: a new type of post-translational modification of the viral glycoprotein. *Cell*, *17*(4), 813–819. https://doi.org/10.1016/0092-8674(79)90321-0

Schrödinger, L. L. C., & DeLano, W. (2020). *PyMOL*. http://www.pymol.org/pymol

Schuster-Böckler, B., & Bateman, A. (2008). Protein interactions in human genetic diseases. *Genome Biology*, *9*(1). https://doi.org/10.1186/GB-2008-9-1-R9

Schwarz, J. M., Cooper, D. N., Schuelke, M., & Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. *Nature Methods*, *11*(4), 361–362. https://doi.org/10.1038/nmeth.2890

Segal-Salto, M., Sapir, T., & Reiner, O. (2016). Reversible Cysteine Acylation Regulates the Activity of Human Palmitoyl-Protein Thioesterase 1 (PPT1). *PLoS ONE*, *11*(1). https://doi.org/10.1371/JOURNAL.PONE.0146466

Shang, S., Hua, F., & Hu, Z. W. (2017). The regulation of β-catenin activity and function in cancer: Therapeutic opportunities. In *Oncotarget*. https://doi.org/10.18632/oncotarget.15687

Shen, L., Shi, Q., & Wang, W. (2018). Double agents: genes with both oncogenic and tumor-suppressor functions. *Oncogenesis*, *7*(3). https://doi.org/10.1038/S41389-018-0034-X

Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M., & Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, *29*(1), 308–311. https://doi.org/10.1093/NAR/29.1.308

Shih, I. M., & Wang, T. L. (2011). Mutation of PPP2R1A: a new clue in unveiling the pathogenesis of uterine serous carcinoma. *The Journal of Pathology*, *224*(1), 1–4. https://doi.org/10.1002/PATH.2884

Shimkets, R. A., Warnock, D. G., Bositis, C. M., Nelson-Williams, C., Hansson, J. H., Schambelan, M., Gill, J. R., Ulick, S., Milora, R. V., Findling, J. W., Canessa, C. M., Rossier, B. C., & Lifton, R. P. (1994). Liddle's syndrome: Heritable human hypertension caused by mutations in the β subunit of the epithelial sodium channel. *Cell*. https://doi.org/10.1016/0092-8674(94)90250-X

Smotrys, J. E., & Linder, M. E. (2004). Palmitoylation of intracellular signaling proteins: regulation and function. *Annual Review of Biochemistry*, *73*, 559–587. https://doi.org/10.1146/ANNUREV.BIOCHEM.73.011303.073954

Sondka, Z., Bamford, S., Cole, C. G., Ward, S. A., Dunham, I., & Forbes, S. A. (2018). The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. In *Nature Reviews Cancer*. https://doi.org/10.1038/s41568-018-0060-1

Soon, W. W., Hariharan, M., & Snyder, M. P. (2013). High-throughput sequencing for biology and medicine. *Molecular Systems Biology*, *9*. https://doi.org/10.1038/MSB.2012.61

Soyombo, A. A., & Hofmann, S. L. (1997). Molecular cloning and expression of palmitoyl-protein thioesterase 2 (PPT2), a homolog of lysosomal palmitoyl-protein thioesterase with a distinct substrate specificity. *The Journal of Biological Chemistry*, *272*(43), 27456–27463. https://doi.org/10.1074/JBC.272.43.27456

Sprinzak, E., & Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology*. https://doi.org/10.1006/jmbi.2001.4920

Stein, A., & Aloy, P. (2010). Novel peptide-mediated interactions derived from high- resolution 3-dimensional structures. *PLoS Computational Biology*, *6*(5), 1–16. https://doi.org/10.1371/journal.pcbi.1000789

Stix, R., Lee, C. J., Faraldo-Gómez, J. D., & Banerjee, A. (2020). Structure and Mechanism of DHHC Protein Acyltransferases. *Journal of Molecular Biology*, *432*(18), 4983–4998. https://doi.org/10.1016/J.JMB.2020.05.023

Stowers, R. S., & Isacoff, E. Y. (2007). Drosophila huntingtin-interacting protein 14 is a presynaptic protein required for photoreceptor synaptic transmission and expression of the palmitoylated proteins synaptosome-associated protein 25 and cysteine string protein. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *27*(47), 12874–12883. https://doi.org/10.1523/JNEUROSCI.2464-07.2007

Strassburger, K., Kang, E., & Teleman, A. A. (2019). Drosophila ZDHHC8 palmitoylates scribble and Ras64B and controls growth and viability. *PLoS ONE*, *14*(2). https://doi.org/10.1371/journal.pone.0198149

Stratton, M. R., Campbell, P. J., & Futreal, P. A. (2009). The cancer genome. *Nature*, *458*(7239), 719. https://doi.org/10.1038/NATURE07943

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102

Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights*, *14*. https://doi.org/10.1177/1177932219899051

Sunyaev, S., Ramensky, V., Koch, I., Lathe, W., Kondrashov, A. S., & Bork, P. (2001). Prediction of deleterious human alleles. *Human Molecular Genetics*, *10*(6), 591–597. https://doi.org/10.1093/HMG/10.6.591

Suwinski, P., Ong, C. K., Ling, M. H. T., Poh, Y. M., Khan, A. M., & Ong, H. S. (2019). Advancing personalized medicine through the application of whole exome sequencing and big data analytics. *Frontiers in Genetics*, *10*(FEB), 49. https://doi.org/10.3389/FGENE.2019.00049/BIBTEX

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., & Von Mering, C. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky1131

Tamborero, D., Gonzalez-Perez, A., & Lopez-Bigas, N. (2013). OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, *29*(18), 2238–2244. https://doi.org/10.1093/bioinformatics/btt395

Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., … Forbes, S. A. (2019). COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky1015

Taylor, S. E., O'Connor, C. M., Wang, Z., Shen, G., Song, H., Leonard, D., Sangodkar, J., LaVasseur, C., Avril, S., Waggoner, S., Zanotti, K., Armstrong, A. J., Nagel, C., Resnick, K., Singh, S., Jackson, M. W., Xu, W., Haider, S., DiFeo, A., & Narla, G. (2019). The highly recurrent PP2A Aα-subunit mutation P179R alters protein structure and impairs PP2A enzyme function to promote endometrial tumorigenesis. *Cancer Research*, *79*(16), 4242. https://doi.org/10.1158/0008-5472.CAN-19-0218

Taylor, W. R. (1986). The classification of amino acid conservation. *Journal of Theoretical Biology*, *119*(2), 205–218. https://doi.org/10.1016/S0022-5193(86)80075-3

The Cancer Genome Atlas Research, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet*, *45*(10), 1113–1120. https://doi.org/10.1038/ng.2764

The International Cancer Genome Consortium, Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., Guttmacher, A., Guyer, M., Hemsley, F. M., Jennings, J. L., Kerr, D., Klatt, P., Kolar, P., Kusada, J., … Wainwright, B. J. (2010). International network of cancer genome projects. *Nature*, *464*(7291), 993–998. https://doi.org/10.1038/nature08987

The UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky1049

Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., & Wilke, C. O. (2013). Maximum Allowed Solvent Accessibilites of Residues in Proteins. *PLOS ONE*, *8*(11), e80635. https://doi.org/10.1371/JOURNAL.PONE.0080635

Tomatis, V. M., Trenchi, A., Gomez, G. A., & Daniotti, J. L. (2010). Acyl-protein thioesterase 2 catalyzes the deacylation of peripheral membrane-associated GAP-43. *PloS One*, *5*(11). https://doi.org/10.1371/JOURNAL.PONE.0015045

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., … Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature 2021 596:7873*, *596*(7873), 590–596. https://doi.org/10.1038/s41586-021-03828-1

Tyanova, S., Temu, T., & Cox, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols*, *11*(12), 2301–2319. https://doi.org/10.1038/NPROT.2016.136

Uguen, A., & De Braekeleer, M. (2016). ROS1 fusions in cancer: a review. *Future Oncology (London, England)*, *12*(16), 1911–1928. https://doi.org/10.2217/FON-2016-0050

Uyar, B., Weatheritt, R. J., Dinkel, H., Davey, N. E., & Gibson, T. J. (2014). Proteome-wide analysis of human disease mutations in short linear motifs: Neglected players in cancer? *Molecular BioSystems*. https://doi.org/10.1039/c4mb00290c

Vacic, V., Markwick, P. R. L., Oldfield, C. J., Zhao, X., Haynes, C., Uversky, V. N., & Iakoucheva, L. M. (2012). Disease-Associated Mutations Disrupt Functionally Important Regions of Intrinsic Protein Disorder. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1002709

Valdez-Taubas, J., & Pelham, H. (2005). Swf1-dependent palmitoylation of the SNARE Tlg1 prevents its ubiquitination and degradation. *The EMBO Journal*, *24*(14), 2524–2532. https://doi.org/10.1038/SJ.EMBOJ.7600724

Van Roey, K., Uyar, B., Weatheritt, R. J., Dinkel, H., Seiler, M., Budd, A., Gibson, T. J., & Davey, N. E. (2014). Short linear motifs: Ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chemical Reviews*, *114*(13), 6733–6778. https://doi.org/10.1021/CR400585Q/ASSET/IMAGES/CR400585Q.SOCIAL.JPEG_V03

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Zídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., … Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, *50*(D1), D439–D444. https://doi.org/10.1093/NAR/GKAB1061

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz Jr., L. A., & Kinzler, K. W. (2013). Cancer Genome Landscapes. *Science*, *339*(6127), 1546–1558. https://doi.org/10.1126/science.1235122

Walser, J.-C., & Furano, A. V. (2010). The mutational spectrum of non-CpG DNA varies with CpG content. *Genome Research*, *20*(7), 875. https://doi.org/10.1101/GR.103283.109

Watkins, X., Garcia, L. J., Pundir, S., & Martin, M. J. (2017). ProtVista: Visualization of protein sequence annotations. *Bioinformatics*, *33*(13), 2040–2041. https://doi.org/10.1093/bioinformatics/btx120

Watson, I. R., Takahashi, K., Andrew Futreal, P., & Chin, L. (2013). Emerging patterns of somatic mutations in cancer. *Nature Publishing Group*, *14*. https://doi.org/10.1038/nrg3539

Weatheritt, R. J., Jehl, P., Dinkel, H., & Gibson, T. J. (2012). iELM-a web server to explore short linear motif-mediated interactions. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gks444

Wedegaertner, P. B., Chu, D. H., Wilson, P. T., Levis, M. J., & Bourne, H. R. (1993). Palmitoylation is required for signaling functions and membrane attachment of Gq alpha and Gs alpha. *Journal of Biological Chemistry*, *268*(33), 25001–25008. https://doi.org/10.1016/S0021-9258(19)74563-3

Weinberg, R. A. (1994). Oncogenes and tumor suppressor genes. *CA: A Cancer Journal for Clinicians*, *44*(3), 160–170.

https://doi.org/10.3322/CANJCLIN.44.3.160

Wlodarchak, N., & Xing, Y. (2016). PP2A as a master regulator of the cell cycle. *Critical Reviews in Biochemistry and Molecular Biology*, *51*(3), 162. https://doi.org/10.3109/10409238.2016.1143913

Xu, Y., Chen, Y., Zhang, P., Jeffrey, P. D., & Shi, Y. (2008). Structure of a Protein Phosphatase 2A Holoenzyme: Insights into B55-Mediated Tau Dephosphorylation. *Molecular Cell*, *31*(6), 873–885. https://doi.org/10.1016/J.MOLCEL.2008.08.006

Xu, Y., Xing, Y., Chen, Y., Chao, Y., Lin, Z., Fan, E., Yu, J. W., Strack, S., Jeffrey, P. D., & Shi, Y. (2006). Structure of the Protein Phosphatase 2A Holoenzyme. *Cell*, *127*(6), 1239–1251. https://doi.org/10.1016/J.CELL.2006.11.033

Yanagawa, S. I., Lee, J. S., & Ishimoto, A. (1998). Identification and characterization of a novel line of Drosophila Schneider S2 cells that respond to wingless signaling. *The Journal of Biological Chemistry*, *273*(48), 32353–32359. https://doi.org/10.1074/JBC.273.48.32353

Yang, C. L., Qiu, X., Lin, J. Y., Chen, X. Y., Zhang, Y. M., Hu, X. Y., Zhong, J. H., Tang, S., Li, X. Y., Xiang, B. De, & Zhang, Z. M. (2021). Potential Role and Clinical Value of PPP2CA in Hepatocellular Carcinoma. *Journal of Clinical and Translational Hepatology*, *9*(5), 661. https://doi.org/10.14218/JCTH.2020.00168

Yang, F., Petsalaki, E., Rolland, T., Hill, D. E., Vidal, M., & Roth, F. P. (2015). Protein Domain-Level Landscape of Cancer-Type-Specific Somatic Mutations. *PLoS Computational Biology*, *11*(3). https://doi.org/10.1371/journal.pcbi.1004147

Yates, C. M., & Sternberg, M. J. E. (2013). The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein-protein interactions. *Journal of Molecular Biology*, *425*(21), 3949–3963. https://doi.org/10.1016/J.JMB.2013.07.012

Yong, L., YuFeng, Z., & Guang, B. (2018). Association between PPP2CA expression and colorectal cancer prognosis tumor marker prognostic study. *International Journal of Surgery*, *59*, 80–89. https://doi.org/10.1016/J.IJSU.2018.09.020

Youn, A., & Simon, R. (2011). Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics*, *27*(2), 175–181. https://doi.org/10.1093/BIOINFORMATICS/BTQ630

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J. F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., … Vidal, M. (2008). High-quality binary protein interaction map of the yeast interactome network. *Science*. https://doi.org/10.1126/science.1158684

Zandvakili, I., Lin, Y., Morris, J. C., & Zheng, Y. (2017). Rho GTPases: Anti- or pro-neoplastic targets? *Oncogene*, *36*(23), 3213–3222. https://doi.org/10.1038/onc.2016.473

Zhong, Q., Simonis, N., Li, Q.-R., Charloteaux, B., Heuze, F., Klitgord, N., Tam, S., Yu, H., Venkatesan, K., Mou, D., Swearingen, V., Yildirim, M. A., Yan, H., Dricot, A., Szeto, D., Lin, C., Hao, T., Fan, C., Milstein, S., … Vidal, M. (2009). Edgetic perturbation models of human inherited disorders. *Molecular Systems Biology*, *5*, 321. https://doi.org/10.1038/MSB.2009.80

Zhou, F., Xue, Y., Yao, X., & Xu, Y. (2006). CSS-Palm: Palmitoylation site prediction with a clustering and scoring strategy (CSS). *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btl013

Zhou, J., Pham, H. T., Ruediger, R., & Walter, G. (2003). Characterization of the Aα and Aβ subunit isoforms of protein phosphatase 2A: Differences in expression, subunit interaction, and evolution. *Biochemical Journal*, *369*(2), 387–398. https://doi.org/10.1042/BJ20021244