Title
An Integrated Chemoproteomics- and Genetics-based Approach to Identify Functional Amino Acids

Author
Palafox, Maria

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

An Integrated Chemoproteomics- and Genetics-based Approach to Identify Functional Amino Acids

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Human Genetics

by

Maria Francis Palafox

2023

ABSTRACT OF THE DISSERTATION

An Integrated Chemoproteomics- and Genetics-based Approach to Identify Functional Amino Acids

by

Maria Francis Palafox

Doctor of Philosophy in Human Genetics

University of California, Los Angeles, 2023

Professor Valerie A Arboleda, Co-Chair

Professor Keriann Marie Backus, Co-Chair

Deciphering the functional and therapeutic relevance of missense variants—mutations that change a single amino acid to an alternative residue—is a central challenge in modern genetics. In this work, we address this gap using an innovative approach that integrates genetic variants, *in silico* predictions of pathogenicity, and proteomic measures of amino acid functionality. First, we found that chemoproteomic methods that use a mass spectrometry-based approach to quantify amino acid sidechain reactivity proteome-wide can identify amino acid positions enriched for disease-associated missense variants. Second, by globally characterizing the positional and contextual relationships between reactive residues and genetic variation, we prioritized several

likely functional amino acids proximal to rare variants of uncertain significance in monogenic disorder genes. While many advanced methods exist to discern the pathogenicity of genetic variants, this work uniquely focuses on pathogenicity of missense and nucleophilic (reactive) residues in human proteins. In summary, this work has important implications in variant prioritization and in therapeutics development, where it can support drug discovery efforts through prioritization of attractive and function amino acids for small molecule targeting.

The dissertation of Maria Francis Palafox is approved.

Jason Ernst

Heather R Christofk

Ellen May Sletten

Keriann Marie Backus, Committee Co-Chair

Valerie A Arboleda, Committee Co-Chair

University of California, Los Angeles

2023

DEDICATION

This dissertation is dedicated to my Nana y mi hermano de laboratorio José Omar Castellón.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I'd like to thank all those who helped me complete this research. Thanks first to Keri Backus and Valerie Arboleda, whose faith in me allowed me to pursue research passionately, and guidance helped me develop and define this work. Your expertise and advice have been invaluable. The other members of my committee Heather Christofk, Jason Ernst, and Ellen Sletten, have been generous with their time, insightful feedback, and words of support that have fueled me over the years. I have enjoyed working with them and look forward to continuing our relationship. To my mentor Mike Carey, for constant support and honest discussions related to my career path. To all members of the cysteine chapel and Arboleda lab, past and present, for camaraderie and motivation, in particular: Brianna Hill-Payne, José Omar Castellón, Emily Yang, Albert Chan, Alex Winnett, Kathryn Elliot, Ernest Armenta, Neil Chan, Alex Sun, Sunny Yan, Heta Desai, Meghna Singh, Lisa Boatner, and Aileen Nava. To other UCLA friends and beyond that that I have learned a tremendous amount from: Laurent Vergnes, Adriana Arneson, Robert Brown, Colin Farrell, Arun Kumar Durvasula, Juan De La Hoz, Annique Claringbould, and Abby Gibbs. To my personal support team: my mom Katherine; my brother William; my roommate Kyle; and my cohort family Katie Leap, Shirley Nieves-Rodriguez, Gregory Rosenberg, and Eric Heinrichs. I would also like to thank my dad Andrew for sharing relevant tweets, Esteban Dell'Angelica for all the emails, and Sam Niles-Jensen and Katie Leap for help with the final step. Lastly, I want to thank Bun for encouraging me to do computational research and Pudding Pie for always pooping in the box, allowing me more time for research.

Chapter 2 is a version of Palafox MF, Desai HS, Arboleda VA, Backus KM (2021), From Chemoproteomic-Detected Amino Acids to Genomic Coordinates: Insights into Precise Multi-omic Data Integration. *Mol. Syst. Biol*. 17. MFP is supported by the Chemistry Biology Interface Training Program T32GM008496. VAA is supported by DP5OD024579. KMB is supported by a Beckman Young Investigator Award and the V Scholar Award V2019-017. We thank Dennis W. Wolan and Gonzalo E. González-Páez for the recombinant caspase-8 proteins. We gratefully acknowledge all members of the Backus Lab and Arboleda Lab for their helpful suggestions.

Chapter 3 is a version of Palafox MF, Boatner LM, Wilde BR, Christofk HR, Backus KM, Arboleda VA (2023), Prioritizing disease-associated missense variants with chemoproteomic-detected amino acids. *Manuscript in preparation.*

VITA

EDUCATION

2016 Bachelor of Science, Advanced Nutritional Sciences Honors
University of Texas, Austin

PEER-REVIEWED PUBLICATIONS

Zore, T., Palafox, M. & Reue, K. Sex differences in obesity, lipid metabolism, and inflammation—A role for the sex chromosomes? *Mol. Metab.* 1–10 (2018). doi:10.1016/j.molmet.2018.04.003

Yan, T. *et al.* SP3-FAIMS Chemoproteomics for High-Coverage Profiling of the Human Cysteinome**. *ChemBioChem* **22**, 1841–1851 (2021).

Yabumoto, M. *et al.* Novel variants in KAT6B spectrum of disorders expand our knowledge of clinical manifestations and molecular mechanisms. *Mol. Genet. Genomic Med.* **9**, 1–21 (2021).

Palafox, M. F., Desai, H. S., Arboleda, V. A. & Backus, K. M. From chemoproteomic-detected amino acids to genomic coordinates: insights into precise multi-omic data integration. *Mol. Syst. Biol.* **17**, (2021).

Boatner, L. M., Palafox, M. F., Schweppe, D. K. & Backus, K. M. CysDB: A Human Cysteine Database based on Experimental Quantitative Chemoproteomics. *ChemRxiv* 1–46 (2023). (to be published in *CellChemBio*)

MANUSCRIPTS IN PREPARATION

Palafox, M.F., Boatner, L.M., Wilde, B.R., Christofk, C.R., Backus, K.M., Arboleda, V.A., Prioritizing disease-associated missense variants with chemoproteomic-detected amino acids.(2023).

Chapter 1. Introduction

The average person's genome contains over 15,000 missense variants (Lek et al 2016), or single nucleotide variants (SNVs) that change an amino acid to an alternative amino acid in proteins. A major goal of human genetics is to distinguish which DNA variants confer disease risk from those that have no noticeable impact or are otherwise considered neutral. Experimental and computational approaches have been developed to better estimate the probability that a given variant is associated with deleterious consequences for an individual's health. However, experimental approaches are not yet widely used for interpreting the large number of candidate genetic variants in a patient's genome primarily because they are challenging to implement and costly to scale (Cooper et al 2011). Computational approaches, despite their algorithmic diversity and availability, are not accurate enough yet on their own to identify disease-causing variants in a clinical setting. Despite recent advances in both experimental and computational methods, identifying which missense variants are associated with disease remains a difficult task.

A parameter associated with protein function is amino acid sidechain reactivity, which fluctuates depending on a residue's chemical microenvironment. Mass spectrometry-based chemoproteomics methods have been developed to assay the intrinsic reactivity and targetability of residues in native biological systems (Weerapana et al 2010; Backus et al 2016; Hacker et al 2017). These methods have successfully identified drug vulnerabilities in cancer (Bar-Peled et al 2017) and the targets of FDA-approved drugs (Blewett et al 2016). While these methods have emerged as powerful

tools for uncovering residues that are critical to protein function, a limitation of this approach is that biochemical measurements from chemoproteomics experiments alone fail to distinguish residue sites essential to proper protein functionality. A major goal following chemoproteomics experiments is to identify which detected residues, out of thousands of residues detected in a single experiment, play pivotal roles in protein functionality and ultimately, are sites amenable to therapeutic drug targeting.

In this dissertation, I integrate chemoproteomics with human genetics to inform whether and to what extent biochemical measurements such as reactivity and computational methods for predicting missense variant pathogenicity can be used to guide the identification of functional residues and disease-associated missense variants in the human genome. In the following dissertation chapters, I explore the utility of my approach towards accomplishing major goals in both areas of science and highlight the strengths of using a multi-disciplinary approach.

In **Chapter 2** I combine chemoproteomic, genomic, and genetic-variant annotation data to understand how genetics can support prioritization of functional Chemoproteomic-Detected Amino Acids (CpDAAs). Accurate and precise inter-database mapping of amino acids is an essential component of such multi-omic studies. While prior proteogenomics studies, which aim to identify unknown proteins by providing protein-level evidence of gene- and isoform-specific expression, showcase the utility of multi-omic approaches, such studies have not extended to chemoproteomics. In this chapter, I evaluate two mapping approaches to match CpDAAs to their genetic coordinates. My analysis sheds new light on the challenges associated with accurate

residue-to-codon mapping and reveals how databases update cycles and a reliance on stable identifiers can lead to pervasive mis-mapping and misidentification of CpDAAs in many functionally important proteins such as PRMT1, G6PD, and TP53. More broadly, this work provides a roadmap for more precise inter-database comparisons with wide ranging applications for both communities of proteomics and genetics researchers.

In **Chapter 3** I apply the mapping insights gained from Chapter 2 and explore what CpDAAs can do for prioritizing protein-altering variants with pathogenic potential. We find that detected proteins are relevant to studying monogenic disorders. Reactive residues on the surface of proteins are associated with missense constraint and can provide evidence of missense intolerance when genetic based scores may be less accurate for a particular context. A global analysis of monogenic disorder genes showed an overall lower abundance of cysteine and higher abundance of glycine and I discuss the significance of amino acid evolutionary history for insights into tolerance of gains and losses by missense variation. I then stratified chemoproteomic detected amino acids based on their 1D and 3D environment relations to missense alleles and predicted pathogenicity scores and confirmed the utility of my approach using functional data for fumarate hydratase protein.

Chapter 2. From chemoproteomic-detected amino acids to genomic coordinates: insights into precise multi-omic data integration

2.1 Introduction

Understanding how proteins work is the bedrock of functional biology and drug development. The identification of amino acids that directly regulate a protein's activity (e.g., catalytic residues, residues that drive interactions, or residues important for folding or stability) is an essential step to functionally characterize a protein. Delineation of amino acid-specific functions is typically accomplished using site-directed mutagenesis (Hemsley et al 1989; Starita et al 2015). While such studies can identify functional hotspots in human proteins, they are typically limited in scope and largely restricted to proteins easily expressed in vitro. With the advent of next-generation sequencing and CRISPR-based mutagenesis, deep mutational analysis can now be scaled to individual genes (e.g., TP53 and BRCA1) (Starita et al 2015; Boettcher et al 2019), but such studies have not been extended genome-wide.

This problem of identifying the functional properties of a specific amino acid parallels one of the central challenges of modern genetics: interpreting the pathogenicity of the millions of genetic variants found in an individual's genome. Many computational methods, such as M-CAP (Jagadeesh et al 2016), Combined Annotation Dependent Depletion (CADD) (Kircher et al 2014), PolyPhen (Adzhubei et al 2010), and SIFT (Vaser et al 2016) integrate data such as sequence conservation, metrics of sequence constraint, and other functional annotations to provide a quantitative assessment of variant deleteriousness. In the absence of experimental data, these scores provide a

metric to rank genetic variants for their effect on a phenotype, something particularly important in the era of genome-wide association and sequencing studies.

Beyond genetic variation, a frequently overlooked parameter that defines functional hotspots in the proteome is amino acid side chain reactivity, which can fluctuate depending on the residue's local and 3-dimensional protein microenvironment. Mass spectrometry-based chemoproteomics methods have been developed that can assay the intrinsic reactivity of thousands of amino acid side chains in native biological systems (Weerapana et al 2010; Backus et al 2016; Hacker et al 2017). Using these methods, previous studies, including our own, revealed that "hyper-reactive" or pKa-perturbed cysteine and lysine residues are enriched in functional pockets. These chemoproteomics methods can even be extended to measure the targetability or "druggability" of amino acid side chains, which has revealed that a surprising number of cysteine and lysine side chains can also be irreversibly labeled by small drug-like molecules (Weerapana et al 2010; Backus et al 2016; Hacker et al 2017). Complicating matters, for the vast majority of these chemoproteomic-detected amino acids (CpDAA), the functional impact of a missense mutation or chemical labeling remains unknown. Integrating chemoproteomics data with genomic-based annotations represents an attractive approach to stratify CpDAA functionality and to identify therapeutically relevant disease-associated pockets in human proteins.

Such multi-omic studies require mapping a protein's sequence back to genomic coordinates, through the transcript isoforms, in essence reverse engineering the central dogma of molecular biology. Accurate mapping between amino acid positions and genomic coordinates remains particularly challenging, due in part to the diversity of cell

type-specific transcript and protein isoforms and the non-linear relationship between gene, transcript, and protein sequences. One approach to address these challenges is through proteogenomics (Ruggles et al 2017), where custom FASTA files are generated from whole exome or RNA-sequencing data. However, such approaches are not scalable or cost-effective. Furthermore, many proteomic datasets, particularly previously acquired and public datasets, lack matched genomic data, precluding proteogenomic analysis.

Many computational tools have been developed for inter-database mapping, including using unique identifiers (Durinck et al 2009; Smith et al 2019; Agrawal & Prabakaran, 2020), methods to map genomic coordinates to protein sequences and structures (David & Yip, 2008; Sehnal et al 2017; Sivley et al 2018; Stephenson et al 2019), and tools for codon-centric-based annotation of genetic variants (Gong et al 2014; Schwartz et al 2019). One key application of these tools is the improved prediction of variant pathogenicity (Guo et al 2017). However, while many predictive genetic scores are built on the GRCh37 genome assembly (frozen in 2014), the UniProt Knowledge Base (UniProtKB) (McGarvey et al 2019) proteomic reference is based on genome assembly GRCh38. Further complicating data integration, the unsynchronized and frequent updates to widely used databases, such as UniProtKB and Ensembl, result in a constantly evolving landscape of genome-, transcriptome- and proteome-level sequences and annotations, which further confounds multi-omic data integration, particularly for residue-level analyses.

Focusing initially on previously identified CpDAAs (Weerapana et al 2010; Backus et al 2016; Hacker et al 2017), we first assess how choice of databases,

including release dates, and the use of isoform-specific, versioned, or stable identifiers impact residue-coordinate mapping and the fidelity of data integration. We then apply an optimized mapping strategy to annotate CpDAA positions with predictions of genetic variant pathogenicity, for both previously published and newly generated chemoproteomic analyses of amino acid reactivity. Our study uncovers key sources of inaccurate mapping and provides fundamental guidelines for multi-omic data integration. We also reveal that highly reactive cysteines, including those identified previously (Weerapana et al 2010) and newly identified CpDAAs, are enriched for genetic variants that have high predicted pathogenicity (high deleteriousness), which supports both the utility of predictive scores to further power proteomics datasets and the use of chemoproteomics to add another layer of interpretation to missense genetic variants. As many databases move to GRCh38, we anticipate that our findings will provide a roadmap for more precise inter-database comparisons, which will have wide-ranging applications for both the proteomics and genetics communities.

2.2 Results

Characterizing the dynamic mapping landscape relevant to CpDAA data integration

Our first step to achieve high-fidelity multi-omic data integration was to establish a comprehensive set of test data. For this, we aggregated publicly available cysteine and lysine chemoproteomics datasets (Weerapana et al 2010; Backus et al 2016; Hacker et al 2017), resulting in a total of 6,510 CpD cysteines and 9,327 CpD lysines detected in

4,119 unique proteins. These 15,837 CpDAAs are further sub-categorized by the residues labeled by cysteine- or lysine-reactive probes (iodoacetamide alkyne [IAA] or pentynoic acid sulfotetrafluorophenyl ester [STP], respectively) and those residues with additional measures of intrinsic reactivity (categorized as high-, medium-, and low-reactive residues).

As our overarching objective was to characterize CpDAAs using functional annotations based on different versions of protein, transcript, and DNA sequences (**Figure 2-1A**), our next step was to develop a high-fidelity data analysis pipeline for intra- and inter-database mapping. To guide our analyses, we first referenced established methods for such data mapping, including ID mapping (Huang et al 2008; Meyer, Geske, & Yu, 2016; Xin et al 2016), residue–residue mapping (Martin, 2005; David & Yip, 2008; Dana et al 2019), and residue–codon mapping (Zhou et al 2015; Li et al 2016) (See **Table 2-1** for detailed descriptions of each type of mapping).

We suspected that the frequent and unsynchronized update cycles of independent databases (**Figure 2-1B**) might complicate accurate residue-level mapping. Supporting this hypothesis, quantification of the average update cycle for each database across this time period revealed that UniProtKB has the shortest mean update cycle (~ 6 weeks; **Figure 2-1C**). In contrast, NCBI is only updated yearly. These different update cycles can create a lag between versions of databases used to create identifier cross-reference (a.k.a. External Reference [xref]) files. For example, ID mapping files provided by Ensembl for UniProtKB proteins may not share identical sequences if not used within the short 4-week window between UniProtKB updates.

To enable further characterization of how database update cycles and mapping strategy impact the fidelity of data integration, we collected a test set of Ensembl releases (**Figure 2-2**). Specific releases were prioritized that (i) represented reference releases based on the GRCh37 or GRCh38 reference genome, (ii) were compatible with the latest Consensus Coding Sequence (CCDS) update for the human genome (release 22), (iii) were used in database for nonsynonymous functional predictions (dbNSFP) v4.0a and CADDv1.4, two resources that integrate functional annotations for all possible nonsynonymous single nucleotide variants (SNV) (Kircher et al 2014; Liu et al 2016; Rentzsch et al 2019), and (iv) were associated with a commonly used version of the Ensembl Variant Effect Predictor (VEP) (McLaren et al 2016).

With these prioritized datasets in hand, we next tracked the loss of CpDAA-containing protein IDs during intra-database mapping of UniProtKB releases and inter-database ID mapping to different Ensembl releases. Gratifyingly, only a handful of the original 4,119 protein IDs were lost due to database updates, both for Ensembl (e.g., 37 IDs for v97 release of Ensembl) and for UniProtKB (e.g., 26 IDs for 2012 UniProtKB; **Figure 2-2**). The greatest identifier loss was observed from mapping UniProtKB-based legacy data to the 2018 UniProtKB-SwissProt CCDS cross-referenced curation of the human proteome, with 119 IDs not found in the 2018 dataset. We ascribe this identifier loss to both UniProtKB updates and to the higher level of curation for proteins in the 2018 dataset, which includes only Swiss-Prot canonical protein sequences with a cross-referenced (“xref”) entry term in the CCDS database. Of note, CCDS gene IDs are manually reviewed and linked to UniProtKB-SwissProt. The TREMBL database is comprised of automatically generated protein IDs, which, as a result, comprises a

substantially larger set of UniProtKB IDs, when compared to the manually curated SwissProt CCDS subset (**Figure 2-3)**. From these analyses, we concluded that using the CCDS UniProtKB release was optimal for integrating functional annotations with chemoproteomics datasets.

Updates to canonical sequences assigned to UniProtKB stable identifiers can lead to intra-database mismapping of CpDAAs

Proteomics datasets, including published CpDAA datasets, are routinely searched against FASTA files containing only canonical UniProtKB proteins (**Table 2-1**), for two main reasons. First, canonical proteins reduce the redundancy and complexity of proteome search databases. Second, these sequences are identified by stable identifiers (also known as the UniProtKB primary accessions) and offer the seeming advantage of remaining constant through database update cycles. However, one particularly confusing aspect of the stable identifier is that the word “stable” in this context does not mean permanent or immutable. Specifically, the associated sequence linked to a stable identifier can change over database releases.

Therefore, we next assessed whether and to what extent updates to the canonical sequences assigned to UniProtKB stable identifiers resulted in mismapping. To confirm the integrity of our CpDAA dataset, we started this process by validating that over 99% of the CpDAA protein IDs and residue positions matched with those found in a 2012 UniProt FASTA file, corresponding to the reference proteome originally used to process the datasets (see Materials and Methods). The small fraction of data lost was due to missing stable identifiers and mis-matched CpDAA positions, which likely stems

from slight inconsistencies between the original processing pipeline and our current workflow. We then mapped the 6,404 CpD cysteines and 9,213 CpD lysines from 4,084 canonical proteins identified in the 2012 dataset to the 2018 UniProtKB CCDS canonical sequence subset of the human proteome. Mapping to CCDS sequences enabled us to take advantage of the extensive array of tools that facilitate forward and reverse annotation between gene, transcript, and protein sequences and would allow for residue-specific mapping to genomic functional annotations (Zhou et al 2015; Meyer, Geske, & Yu, 2016; McGarvey et al 2019). Updating to the 2018 release was a requisite step for using these tools, as they overwhelmingly require recent cross-reference files using the newest reference genome GRCh38. For all CpDAA positions, we performed residue–residue mapping—defined as a one-to-one correspondence between amino acids in proteins from different databases or release dates—to match the 2012 canonical UniProtKB sequences with their 2018 counterparts. This dataset mapping resulted in the loss of 121 protein IDs, with 108 simply not found in the 2018 reference file and the remaining 13 found to have different canonical sequences, resulting in mismapping or loss of the originally identified CpDAA residues.

The high concordance between these two UniProtKB releases, separated by 6 years, indicates that for the vast majority of UniProtKB updates, differences in release date should not complicate re-mapping legacy proteomics data to more recently released gene, transcript, and protein sequences. However, we were surprised to find that several widely studied proteins, including protein arginine N-methyltransferase 1 (PRMT1 or ANM1, Q99873), serine/threonine protein kinase, (SIK3; Q9Y2K2) (Walkinshaw et al 2013), and tropomyosin alpha-3 chain (TPM3, P06753), had

canonical protein sequence differences resulting in all or nearly all CpDAA positions to be missed using the 2018 position index. We observed two main reasons for these losses: (i) changes to the canonical sequence associated with the UniProtKB stable ID and (ii) changes to which isoform is assigned as the canonical sequence. While both 2012 and 2018 sequences of PRMT1 are associated with UniProtKB stable ID Q99873, the 2018 sequence contains an additional short N-terminal sequence, not present in the 2012 sequence (**Figure 2-4A**). As a result, all 13 PRMT1 CpDAAs failed to map to the 2018 UniProtKB release. In the 2012 release of UniProtKB, the canonical sequence of the peptidyl-prolyl cis-trans isomerase FKBP7 is associated with the versioned (isoform) ID Q9Y680-1, whereas in the 2018 release, the canonical sequence is associated with the versioned (isoform) ID Q9Y680-2, which lacks a short sequence (AAΔ125:162) in the middle of the protein. For FKBP7, this update fortuitously does not result in loss of CpD Lys83, as it is located N-terminal to the deletion. These updates to the protein sequence are, in essence, masked by the stable IDs, which do not flag sequence updates or changes to which isoform sequence is assigned as the canonical. Exemplifying this problem, we identified 45 stable identifiers with non-identical canonical protein sequences in the 2012 and 2018 UniProtKB releases.

To further understand how the presence or absence of protein isoforms impacts the fidelity of data mapping during intra-database (UniProtKB) mapping, we identified all isoforms associated with CpDAA stable protein IDs. Analysis of this dataset revealed that 58% of protein stable IDs have between 2–5 associated isoform sequences (**Figure 2-4B**). Catenin delta-1 protein (CTNND1, O60716) had 32 isoforms, which was the greatest number of isoforms in our dataset. Protein isoforms are identified by the “-X”

after the UniProtKB ID, where X represents the isoform name. A common assumption of most mapping tools and proteomics databases is that the “-1” sequence is the canonical sequence. However, a key finding from our isoform analysis is that the canonical sequence does not always correspond to the “-1” isoform ID provided by UniProtKB. In fact, for 288 proteins in the UniProtKB 2018 release, the non-“-1” entry corresponds to the canonical isoforms, and for 55 CpDAA-containing proteins in our dataset (~ 2%), the canonical sequence is not the “-1” isoform (**Figure 2-4C**). Strikingly, the canonical sequence can even be the “-10” isoform, as is the case for the Ras-associated and pleckstrin homology domains-containing protein (RAPH1, Q70E73). In the context of database mapping, all of these non-“-1” canonical proteins will likely result in mismapping using established tools.

Accurate residue-level inter-database mapping between UniProtKB and Ensembl is dependent on database update cycles

To investigate how sequence versions impact inter-database mapping, we next turned to ID cross-reference files that are released by Ensembl and UniProtKB. Cross-reference files can be used to convert between UniProtKB and Ensembl ID types. Three major challenges arise with ID cross-referencing: (i) when cross-reference stable IDs match, but corresponding sequences are not identical, (ii) multi-mapping, where a UniProtKB ID maps to many Ensembl protein (ENSP), transcript, and gene IDs, and (iii) when the origin, both the time of the releases and the specific database provided cross-reference files used, determines the mapping accuracy of datasets.

Glucose-6-phosphate dehydrogenase (G6PD, P11413) exemplifies how sequence updates associated with a stable ID can lead to mismapping of gene-, transcript-, and protein-level annotations for CpDAAs (**Figure 2-4D**). For G6PD, the same UniProtKB ID maps to four unique ENSP IDs with identical sequences (see first row in “Identical”) as well as four different ENSP IDs with non-identical sequences (see second row in “Non-identical”). For G6PD, this significant redundancy is also observed at the gene and transcript level, both for stable and versioned IDs (**Figure 2-5A**). Overall, genes undergo the highest frequency of sequence re-annotation due to continual refinement of the reference genome. In contrast, protein IDs remain largely fixed across releases (**Figure 2-5B**).

To assess how pervasive multi-mapping is across the entire CpDAA dataset, we quantified the mean number of Ensembl IDs per UniProtKB ID. We counted both versioned and stable Ensembl IDs types (gene, transcript, and protein IDs), for all CpD UniProtKB proteins grouped by single (**Figure 2-5C**) or multi-isoform (**Figure 2-5D**) associated stable IDs. We suspected that database updates for all data types (gene, transcript, and protein) and the presence of UniProtKB isoforms would contribute to the observed multi-mapping of CpD protein IDs in our dataset. Of note, Ensembl versioned IDs indicate changes to the associated sequence rather than the presence of isoforms. For example, for protein tropomyosin alpha-4 chain (TPM4, P67936), during the update from v96 to v97, the stable protein identifier showed version change from “.3” to “.4” (ENSP00000300933.3 to ENSP00000300933.4), which corresponds to a difference of 165 amino acids in the primary sequence caused by the update. Not surprisingly, we found that UniProtKB stable identifiers with multiple associated protein isoforms have a

higher average of cross-referenced Ensembl ID types per UniProtKB stable identifier, when compared to UniProtKB stable IDs associated with only one protein isoform. In addition, single isoform UniProtKB stable IDs are more likely to cross-reference identical ENSPs, when compared to multi-isoform UniProtKB stable IDs (**Figure 2-6A** and **Figure 2-6B**).

One last challenge we identified is that the origin of the cross-reference file (whether it was created by UniProtKB or by Ensembl) affected the outcome of our mapping procedures. Across the five Ensembl releases, only 56.9% of all Ensembl-UniProtKB cross-referenced IDs had identical protein sequences (**Figure 2-6**). We then used a cross-reference file from UniProtKB that, unlike the Ensembl mapping files, contains mappings with canonical isoform protein identifiers for UniProtKB proteins to Ensembl stable protein IDs, to test whether inclusion of isoform name details improves the accuracy of inter-database ID mapping. This approach allowed for > 99% identical protein sequence cross-references for UniProtKB-ENSP IDs and substantially reduced the burden of identifier multi-mapping (**Figure 2-7A** and **Figure 2-7B**). Our study demonstrates that high-fidelity ID cross-referencing requires attention to details regarding database updates, multi-mapping, and identifier types used in cross-reference file sources. We also observed that sequences associated with mapped UniProtKB and Ensembl stable IDs varied significantly in alignment distance depending on the Ensembl version (**Figure 2-4E**; **Figure 2-10**), with temporally close releases showing generally greater sequence similarity.

Assessment of pathogenicity predictions for CpD cysteine and lysine codons, using residue–codon mapping

Our next objective was to apply residue–codon mapping to the prioritization of functional CpDAAs. Cysteines and lysines are both highly conserved, with 97% (Miseta & Csutora, 2000) and 80% (Hacker et al 2017) median conservation, respectively. Consequently, sequence motif conservation cannot distinguish between functional and non-functional residues within chemoproteomics datasets. To identify cysteine- and lysine-centric genetic features suitable for pathogenicity prioritization, we tailored our pipeline to reverse-translate CpD cysteine and lysine positions in canonical UniProtKB proteins to genomic coordinates from both major genome assemblies (GRCh37 and GRCh38) and genomic-based functional annotations. For all proteins within our CpDAA dataset, referred to as detected proteins, we also processed undetected equivalent residue types in CpD Cys- and/or CpD Lys-containing proteins (**Figure 2-8**). Cysteines and lysines were required to have valid coordinates in GRCh37 and GRCh38 reference genome assemblies, as some functional genetic variant annotations are only available in one genome assembly. Probe-labeled cysteines and lysines represent ~ 15% of all cysteines (6,057 CpD Cys out of 40,107 total Cys) and ~ 6% of all lysines (8,868 CpD Lys out of 149,520 total Lys) found in chemoproteomic-identified proteins (n = 3,840 UniProtKB IDs successfully mapped; **Figure 2-9A** and **Figure 2-9B**).

Next, genomic coordinates of cysteine and lysine codons from 3,840 detected proteins were annotated by a panel of functional scores (Quang, Chen, & Xie, 2015; Shihab et al 2015; Ioannidis et al 2016; Jagadeesh et al 2016; preprint: Samocha et al 2017; Sundaram et al 2018; Rentzsch et al 2019). With the goal of assessing the

correlation between individual scores and chemoproteomics identification labels, we selected complementary pan-genome and missense deleteriousness prediction scores based on either GRCh37 or GRCh38 reference genome assemblies for our analysis. For the CADD score, which is available for both assemblies, we observed a trend of slightly higher scores with CADD38 compared to CADD37 (**Figure 2-12**). We calculated the Spearman's correlation of scores for all possible nonsynonymous SNVs overlapping cysteine and lysine codons and saw a higher correlation between the deleteriousness predictions for CpD cysteine substitutions (**Figure 2-9C**) than for CpD lysine substitutions (**Figure 2-9D**). For the subset of scores that provide deleteriousness scores for all possible nonsynonymous variants, we did not observe substantial differences between the correlation of scores for chemoproteomic- detected and undetected lysines or cysteines (**Figure 2-13**).

Pathogenicity thresholds, which are provided by a subset of the scores investigated (e.g., CADD, functional analysis through hidden markov models [fathmm-MKL], and Deleterious Annotation of genetic variants using Neural Networks [DANN]), provide a useful cut-off for assessing whether substitutions at specific amino acids are likely to be deleterious to protein function. Therefore, we next assessed whether substitutions at detected vs undetected cysteines or lysines were more likely to be predicted damaging. We first assessed the amino acid substitutions for cysteine and lysine resulting in the greatest chemical property change, or highest Grantham score (Grantham, 1974), Cys > Trp and Lys > Ile. For CADD38 (Kircher et al 2014), fathmm-MKL coding (Shihab et al 2014), and DANN (Quang, Chen, & Xie, 2015), substitutions of detected cysteines were less likely to be predicted damaging compared to

substitutions of undetected cysteines (**Figure 2-9E**, red). In contrast, substitutions of detected lysines were more likely to be predicted damaging compared to substitutions of undetected lysines (**Figure 2-9E**, blue). This trend for cysteine and lysine predicted deleterious score enrichment extended to all missense types (**Figure 2-16A**).

We next tested if these trends would extend to clinically validated “pathogenic” and “likely pathogenic” missense mutations, as identified by the ClinVar database (Landrum et al 2018). ClinVar is the gold standard repository of genomic variants associated with monogenic disorders. In total, the filtered ClinVar dataset contained 2,225 disease-associated missense variants that change from a cysteine (1,653 variants) or lysine (572 variants). We found no significant enrichment of disease-associated variants in detected over undetected cysteines (**Figure 2-9F**, red). In contrast, detected lysines showed a significant enrichment for disease-associated variants relative to undetected lysines (**Figure 2-9F**, light blue). Combining cysteine and lysine data revealed detected residues in general as more likely to harbor disease-associated mutations relative to equivalent undetected residues in 3,840 detected proteins (**Figure 2-9F**, dark blue). Given the challenges associated with accurately diagnosing missense variants, we expect that chemoproteomic detection, particularly for lysine residues, could be used as an additional metric to improve pathogenicity predictions for genetic variants.

Chemoproteomics data combined with pathogenicity scores can help prioritize functional residues

We next assessed correlations between genetic-based pathogenicity score and amino acid reactivity, as assessed by chemoproteomics. We chose CADD as the optimal score to evaluate, as it integrates other nucleotide variant predictors into its model and is available for both reference genome assemblies, GRCh37 and GRCh38. Chemoproteomic reactivity measurements were binned into low, medium, and high reactivity categories, defined as low (R10:1 > 5), medium (2 < R10:1 < 5), high (R10:1 < 2) isoTOP-ABPP ratios, respectively (Weerapana et al 2010; Hacker et al 2017). These ratios quantify the relative labeling of a residue at different probe concentrations (e.g., 1× vs 10×). A ratio closer to one indicates that labeling is saturated at low probe concentration, which corresponds to a cysteine or lysine with higher intrinsic reactivity.

To adapt CADD scores from the nucleotide level to the amino acid level for CpDAAs, the mean and max CADD score for all possible nonsynonymous SNVs per codon (see Methods) were calculated. For both max (**Figure 2-11A**) and mean (**Figure 2-16B**) CADD codon scores, we found that highly reactive cysteines show significantly higher predicted deleteriousness. In contrast, lysine reactivity did not correlate with predicted pathogenicity (**Figure 2-11B** and **Figure 2-16C**).

As the legacy cysteine reactivity dataset was relatively small (94 high reactivity cysteines in total), we next sought to verify these striking correlations, using a larger dataset. For this, we subjected lysates from the immortalized human T lymphocyte Jurkat cell line to isoTOP-ABPP reactivity profiling, comparing cysteine labeling with 10 or 100 μM iodoacetamide alkyne probe, as has been described previously (Weerapana et al 2010). In aggregate, we identified 4,291 cysteines across five replicate

experiments (~ 4-fold more cysteines than were assayed by (Weerapana et al 2010)), including 322 high, 1,448 medium, and 2,247 low reactivity residues. A strong correlation (Pearson correlation coefficient = 0.5) was observed between values reported in our new dataset and those reported previously (**Figure 2-15**). This rich dataset allowed us to further verify our finding that the codons of highly reactive CpDAAs are enriched for high pathogenicity scores. Gratifyingly, our initial finding was reproduced with this new and larger dataset (**Figure 2-16B** and **C**), supporting both the validity of our approach and the robustness of our findings.

As a first case study to explore the utility of integrating genetic-based pathogenicity predictions with CpDAA reactivity measures, we turned to the well-characterized essential enzyme glucose-6-phosphate dehydrogenase (G6PD). Associated with over 160 different genetic mutations, G6PD deficiency is one of the most common genetic enzymopathies (Hwang et al 2018). As G6PD deficiency is associated with both acute and chronic hemolytic anemia (Porter et al 1964; Miwa & Fujii, 1996) (OMIM #300908), and with malaria resistance (Luzzatto, Usanga, & Reddy, 1969) (OMIM #611162), identifying functionally important residues in G6PD should inform the diagnosis and treatment of G6PD-associated genetic disorders. To visualize CADD pathogenicity scores along protein sequence length, we plotted the first 300 amino acids in G6PD with lines tracking max CADD GRCh38 scores, including the positions of all 15 residues identified in prior chemoproteomics studies (**Figure 2-17A**). Of particular interest to us were K171 and K205, which are both located proximal to the enzyme active site (**Figure 2-17B**). While K171 and K205 had very different intrinsic reactivities (R10:1 = 1.3 and R10:1 = 9.2, respectively), both showed high max CADD

scores (28.8 and 32, respectively; **Figure 2-17A**). Consistent with the observed high CADD scores, chemical modification at K205 (e.g., by aspirin) has been found to block G6PD activity (Jeffery, Hobbs, & Jörnvall, 1985; Ai et al 2016) and mutations at K171 have been implicated in anemia (Hirono et al 1989; Au et al 2000). These prior data, when combined with our analysis of CADD and reactivity measurements support our finding that the propensity of lysines to react with electrophilic probes, but not measured differences in their intrinsic probe reactivity, correlate with predicted pathogenicity (**Figure 2-9E** and **Figure 2-9F**, and **Figure 2-16A** and **Figure 2-16C**).

We next sought to determine whether the utility of integrating genetic-based pathogenicity predictions with CpDAA reactivity measures could extend to the de novo discovery of functional residues. We turned to the well-characterized enzyme caspase-8, a member of the cysteine-aspartic acid protease (caspase) family and a key initiator of extrinsic apoptosis. Pathogenic mutations in caspase-8 result in autoimmune lymphoproliferative syndrome (ALPS, OMIM# 607271) (Chun et al 2002; Kanderova et al 2019) and are associated with certain types of cancer. Our chemoproteomic reactivity dataset (**Figure 2-18A**) revealed that caspase-8 harbors two iodoacetamide alkyne-reactive cysteines: the catalytic cysteine (Cys360, R10:1 = 3.8) and a second non-catalytic cysteine (Cys409, R10:1 = 2.9). Consistent with its function as the catalytic nucleophile, the codon of Cys360 has a high mean CADD score (29.3), whereas the codon of Cys409 has a lower CADD score (21.4), indicative that mutations that alter Cys409 should be less damaging to caspase-8 (**Figure 2-11C**). Cys409 is located on a flexible loop ~ 11.8 Å from the active site, as revealed by our projection of the max CADD codon scores onto the CASP8 X-ray structure (**Figure 2-11D**). As, to our

knowledge, the functional impact of Cys409 mutations has not been assessed, we tested whether mutations at Cys409 would impact protein function, as indicated by the elevated measured reactivity, but not the moderate CADD score. Activity assays revealed that mutations at Cys409 do indeed impact protein function, completely blocking proteolytic activity (**Figure 2-11E**). Taken together, these analyses highlight the utility of integration of chemoproteomic measures with pathogenicity predictions to improve stratification of functional and pathogenic residues.

2.3 Discussion

We conducted an in-depth assessment of multiple mapping strategies to facilitate multi-omic analysis of chemoproteomics datasets. We then applied our optimal mapping strategy to analyze the relationship between missense pathogenicity scores and chemoproteomic measures of the intrinsic reactivity of cysteine and lysine residues. Our study revealed a number of challenges that limit the precision of multi-omic data analyses when using publicly available chemoproteomics datasets. To increase awareness of identifier mapping problems and to highlight important considerations for those analyzing similar datasets, we have summarized a list of best practices for accurate curation of functional annotations for CpDAA below.

Recommended best practices for inter- and intra-database integration for chemoproteomics datasets

Entry Recommendation

A Support integration of quantitative chemoproteomics studies by (i) providing reference UniProtKB FASTA files alongside raw proteomic data files and (ii) including genomic coordinates for the codons of identified amino acids in the reference files

B Perform proteomics database searches against reference database sequences that map to known transcript and gene coordinates (e.g., CCDS)

C Perform sequence identity checks, which will identify and minimize mismapping caused by canonical sequence updates between UniProtKB releases

D Map data to the appropriate genome assembly for downstream applications. Genome assembly updates can introduce or refine genome resolution and in doing so alter the genomic coordinates of codons. Not all downstream pathogenicity predictors are compatible with both GRCh37 and GRCh38.

The availability of raw proteomics data in public repositories (e.g., PRIDE (Côté et al 2012), PeptideAtlas, and Panorama (Sharma et al 2014)) might suggest an obvious solution to address the challenges associated with reprocessing published data: to re-search raw data using a new UniProtKB reference. However, reprocessing raw proteomic datasets can be both computationally expensive and time-limiting. An important alternative is to re-map the processed residue-level data to a release of UniProtKB that serves as the reference proteome for all functional annotations of interest, facilitating comparisons between annotated datasets. Complicating matters, providing the reference search databases (typically a custom UniProtKB FASTA file) alongside the raw proteomics files is not routine, and, although UniProtKB is updated

monthly, only annual releases are maintained long-term in the database archives. Simply put, the original reference search sequences used in a chemoproteomics study may no longer be accessible for subsequent follow-up studies. Use of non-matched reference files can result in data loss and annotation errors, which may confound interpretation. For example, when we remapped legacy protein identifiers to multiple UniProtKB releases, we lost 28–199 of proteins, which ranges from 0.6 to 4.8% of the original total CpDAA proteins (**Figure 2-2**). While this may, at first glance, seem to be a paltry fraction of all data, these losses can still prove problematic when key proteins of interest are lost due to database release differences.

There are several interconnected causes for our observed data loss at the protein level. The absence of protein isoform-specific identifiers in most proteomics search databases, particularly when combined with database updates to canonical sequences, can lead to mismapping, as shown for PRMT1 and FKBP7 (**Figure 2-4A**). The small number of UniProtKB sequences for which the canonical sequence is not the UniProtKB “-1” entry can also lead to further mismapping, especially when using mapping software that relies on this assumption (**Figure 2-4C**). Making reference FASTA files publicly available alongside raw data files is a relatively simple solution to facilitate data integration.

Reversing the central dogma to map protein identifiers back to transcript and gene identifiers and CpDAA positions back to transcript and genomic coordinates adds several additional layers of mapping complexity. Ensembl stable identifiers (gene, transcript, and protein), which are linked to UniProtKB stable identifiers are useful for facilitating this process. However, the number of redundant sequences maintained by

Ensembl and the dynamic landscape of Ensembl entries across releases complicates the use of Ensembl stable IDs for inter-database mapping. For example, for the protein G6PD, across the five Ensembl releases investigated, we identified seven stable protein IDs, of which only one was consistently identical to the UniProtKB canonical sequence for G6PD (**Figure 2-4D**). The unsynchronized and frequent database update cycles are a cause of mismapping, which is particularly problematic for large-scale residue-level annotation projects (**Figure 2-4E** and **Figure 2-2**). Practically speaking, what this means is that a CpDAA from an available proteomics dataset could easily be mapped to the incorrect amino acid in an ENSP, followed by the incorrect transcript position, incorrect genomic coordinates, and incorrect pathogenicity score. Although there are a number of tools (e.g., TransVar and BISQUE (Zhou et al 2015; Meyer, Geske, & Yu, 2016) that facilitate inter-database cross-referencing, their performance can be limited by all the challenges outlined above. An important and easily implementable solution to these problems is to search proteomics data against a highly curated reference file, such as the UniProtKB subset of cross-referenced CCDS proteins. Additionally, where possible, sequence identity checks should be performed to verify the mapping of identified residues.

Choice of reference genome further complicates data mapping. While many studies have now transitioned to GRCh38, many useful annotations, including variant-, sub-gene-, and gene-level metrics (e.g., MPC, PrimateAI, M-CAP, CCR, LOEUF), were built using GRCh37 genome assembly and are generally incompatible with the more recent GRCh38 genome assembly. As GRCh37 was frozen in 2014, mismapping can occur from invalid coordinates of proteomics datasets generated using newer reference

proteomes based on GRCh38 coordinates. For many annotations, the solution to different genome assemblies is to “lift-over” annotations to the other genome assembly. However, not all functional annotations are compatible with liftover. Local sequence alignment tools can be used to address problems when transitioning between GRCh37 and GRCh38 but can be challenging to scale genome-wide. The transition of all relevant annotations to the GRCh38 reference genome is ongoing and will address many of the aforementioned issues. However, this move is a substantial undertaking that requires rerunning of large-scale datasets and extensive quality control measures. To make full use of these scores, we recommend mapping proteomics data to genomic coordinates for both assemblies.

Together our analysis of inter-database mapping enabled us to compile a rigorously curated dataset of CpDAAs that mapped to both GRCh37- and GRCh38-based scores (data can be visualized in our CpDAA R Shiny app https://mfpalafox.shinyapps.io/CpDAA/). Using this dataset, we were then able to ask a number of novel questions, including how different scores compare across all identified cysteine and lysine residues and whether the codons of specific residues are enriched for predicted pathogenic mutations. For all nucleotide substitutions that result in a CpD cysteine or lysine amino acid change, we observed generally high concordance between scores (**Figure 2-9C** and **Figure 2-9D**). While mutations at detected cysteine codons were, in general, predicted to be less deleterious than those at undetected cysteine codons, the subset of CpD cysteines with heightened reactivity were predicted to be more damaging than cysteines of lower reactivity (**Figure 2-9E**, **Figure 2-14** and **Figure 2-16A**). No such trend was observed for highly reactive lysines (**Figure 2-11B**;

Figure 2-14). These intriguing findings suggest that cysteine hyper-reactivity is a privileged feature that could be used to inform the functions of genetic variants. As a demonstration of the utility of cysteine reactivity measures for identification of functional residues, we found that mutation of the non-catalytic Cys409 in caspase-8, which had an elevated reactivity ratio but relatively modest CADD score, completely ablated proteolytic activity, which supports that reactivity measurements likely can help to functionally stratify amino acids when CADD scores are less than conclusive.

We can foresee a multitude of applications for chemoproteomic and genomic data integration. While prior studies that revealed hyper-reactive cysteine residues are enriched in redox-active sites and enzyme active sites (Weerapana et al 2010; Backus et al 2016) and that hyper-reactive lysines were depleted in post-translational modification sites (Hacker et al 2017), most CpDAAs still lack functional annotation. Predictive tools, such as those highlighted here, will undoubtedly aid in stratification of residues identified by chemoproteomics studies, pinpointing potentially druggable and disease-linked protein regions. To further aid in integration of CpDAA functional data, we have developed the CpDAA database to house all datasets used in this study together with their associated annotations.

Another area that we expect will benefit from such multi-omic approaches is interpretation of the impact of rare missense variants identified in patients with monogenic disorders. Protein-level functional data can aid in the interpretation of variants of uncertain significance (VUS), including those identified in clinical genetic testing, and can guide follow-up research studies. We anticipate that chemoproteomic methods should prove enabling for VUS interpretation, providing a high-throughput

means to stratify amino acid functionality that is complementary to established genetic approaches, including site-directed mutagenesis. Application of chemoproteomics data to clinical studies will require careful data integration and sequence level mapping, particularly given that the reference sequences and choice of identifiers employed by clinical vs research studies are typically non-identical.

Addition of protein structural data to such pipelines will likely further improve their utility and predictive power. As a starting point to such structure-based data integration, we mapped CADD predictive scores directly to the structures of CASP8 and G6PD (**Figure 2-11D** and **Figure 2-17C**). This 3-dimensional data integration highlighted key residues that form a common function in 3D space but are not easily identified using predictions associated with conservation in the linear-space of DNA. Looking to the future, we anticipate that such multi-omic studies will likely prove most enabling when combined with rigorous functional validation, for example by combining CRISPR-Cas9 mutagenesis with phenotypic assays. The use of CRISPR-Cas9 base editors (Kim et al 2019; Grünewald et al 2019, 2020) should facilitate such studies, particularly when combined with protein-centric guide RNA design packages (e.g., CRISPR-TAPE) (Anderson et al 2020). In sum, we anticipate that such studies represent the next frontier for both the genetics and chemoproteomics communities.

2.4 Methods

Data sources

All data sources are listed in Reagents and Tools table. CpDAA datasets were obtained from the following studies (Weerapana et al 2010; Backus et al 2016; Hacker et al

2017). UniProtKB-SwissProt human proteome filtered by canonical isoform and cross-reference in CCDS database was downloaded August 06, 2018 (2018_06; see Reagents and Tools table). Two cross-reference file sources were used to map UniProtKB protein IDs to Ensembl IDs: (i) UniProtKB ID mapping (idmapping.dat) (McGarvey et al 2019) or (ii) Ensembl release-specific mapping files (xref files) (Aken et al 2016). ENSPs and identifiers were extracted from five release-specific FASTA files (Ensembl database version v85, v92, v94, v96, and v97) downloaded November 19, 2019. CADDv1.4 (Kircher et al 2014) scores were downloaded on July 03, 2019. DANN (Quang, Chen, & Xie, 2015), fathmm-MKL (Shihab et al 2014), M-CAP v1.3 (Jagadeesh et al 2016), MPC release 1 (preprint: Samocha et al 2017), REVEL (Ioannidis et al 2016), and PrimateAI (Sundaram et al 2018) scores were extracted from dbNSFPv4.0a (Liu et al 2016) downloaded on June 11, 2019. “Pathogenic” and “likely pathogenic” labeled variants were extracted from the July 24, 2019, release of ClinVar (Landrum et al 2018).

Database update cycles

Average time between Ensembl, GENCODE, CCDS, and NCBI updates was quantified using all releases between August 2013 and July 2019 (5 years and 11 months window of time). Dates counted refer to the public release date posted on each databases' ftp site. For the UniProtKB update cycle length, values provided by the UniProtKB website on typical time between Knowledgebase releases from 2019 (4 weeks) and 2020 (8 weeks) were averaged. UniProtKB, Ensembl, GENCODE, CCDS, and NCBI releases

were selected based on proximity to the release dates of the five Ensembl database versions analyzed in the current study.

Mapping CpDAA data to more recent UniProtKB releases

CpDAA datasets had been previously searched against a non-redundant reverse concatenated UniProtKB reference FASTA file (Weerapana et al 2010; Backus et al 2016; Hacker et al 2017) from the November 2012 (2012_11) release and amino acids in labeled peptides were annotated with the corresponding UniProtKB stable ID, amino acid letter, and position (e.g., P11413_C205). The author-provided UniProtKB 2012_11 FASTA file was referenced to check the UniProtKB IDs and CpDAA positions. Legacy chemoproteomic-detected cysteine and lysine positions that did not match positions in the canonical sequences from the 2012_11 release were dropped from further analysis. The UniProtKB 2012 canonical protein-based CpDAA residue numbers were then checked against UniProtKB canonical proteins from the 2018_06 release of CCDS cross-referenced human proteome dataset (See GitHub for python script). Chemoproteomic-detected proteins were excluded from further analysis if (i) UniProtKB canonical sequence from 2018 release was missing chemoproteomic-detected positions (e.g., natural variant overlaps detected cysteine position), (ii) UniProtKB ID flagged with “caution” on UniProt's website (e.g., https://www.uniprot.org/uniprot/Q8WUH1), and (iii) UniProtKB IDs not cross-referenced in all five Ensembl release-specific mapping files.

Assessment of isoforms per stable UniProtKB ID

The UniProtKB homo sapien FASTA file containing canonical and isoform sequences was downloaded August 06, 2018. Isoform IDs per UniProt entry (referred to as stable ID in this study) were counted in the FASTA file. Canonical isoform IDs marked by lack of isoform name details (e.g., P11413) were excluded.

Identification of UniProtKB canonical isoform ID numbers

UniProtKB canonical isoform ID numbers (e.g., P11413-X, “X” representing the isoform name) were identified for multi-isoform associated UniProtKB entries by comparing the 2018 UniProtKB FASTA file (used to count total isoforms per UniProtKB entry) and the UniProtKB ID mapping (idmapping.dat) file from August 01, 2018, release downloaded August 06, 2018. The FASTA file displays the canonical protein isoform ID with no isoform name details, but the idmapping.dat file displays the canonical isoform protein ID with these details.

Inter-database identifier mapping (ID mapping) of CpDAA residues between UniProtKB and ENSPs

Two methods were used to cross-reference stable or versioned protein IDs between UniProtKB and five Ensembl releases:

Method A

Ensembl mapping: Ensembl mapping (“xref”) files from the five releases studied (v85, v92, v94, v96, and v97) were used for inter-database identifier mapping. Ensembl gene

(ENSG), transcript, and associated protein IDs cross-referencing the curated set of 3,953 CpD UniProtKB stable IDs were extracted and grouped by single or multi-isoform status of the cross-referenced UniProtKB entry. Ensembl IDs cross-referencing UniProt CpD protein IDs were then used to filter the five Ensembl release-specific peptide FASTA files for associated protein sequences.

Method B

UniProtKB isoform-specific mapping: UniProtKB ID mapping (idmapping.dat) file from August 01, 2018, release was used for inter-database identifier mapping. Ensembl IDs cross-referenced by the UniProtKB canonical protein isoform IDs for multi-isoform entries and stable IDs for single isoform entries were pooled and used to filter release-specific Ensembl peptide FASTA files for associated protein sequences.

Assessing identifier multi-mapping between UniProtKB and Ensembl

From Method A ID mapping, the total number of unique Ensembl IDs (versioned and stable) from five releases that cross-reference CpD UniProt proteins was calculated for each UniProtKB ID. The mean number of unique multi-mapping Ensembl IDs per CpD UniProtKB protein ID was calculated for single and multi-isoform entries. Sequence identity was checked for all cross-referenced Ensembl and UniProtKB proteins and marked by an additional Boolean column (“False” for non-identical and “True” for identical Ensembl-UniProt canonical proteins; see GitHub for python script). From Method B ID mapping, as with analysis for Method A, identifier multi-mapping was calculated for single and multi-isoform UniProtKB entries and sequence identity of

cross-reference proteins was marked by an additional Boolean column. Student's unpaired t-test was used to assess all ID multi-mapping differences between versioned and stable ENSG, transcript, and protein IDs cross-referencing our curated set of 3,953 CpD UniProt protein IDs found in all Ensembl release-specific mapping files.

Identification of frequently updated Ensembl sequence types and non-identical cross-referenced UniProtKB-ENSPs

CpD UniProtKB canonical protein IDs were used to filter five Ensembl peptide FASTA files (Method A). A total of 8,861 unique Ensembl stable protein IDs were shared across all five Ensembl releases, cross-referencing a total of 3,887 CpD UniProtKB canonical proteins IDs. The 8,861 ENSP IDs with their associated stable gene and transcript IDs in each Ensembl release file were combined into a stable key ID (formatted as “ENSG_ENST_ENSP”, for gene, transcript, and protein Ensembl stable IDs). Ensembl versioned IDs were additionally extracted from the release-specific FASTA files. To identify differences between ENSG, transcript, and protein sequence re-annotation rates, ID version number increments (signifying sequence re-annotation updates) relative to the v85 versioned IDs were summed for each ID biotype (gene, transcript, and protein ID extension numbers “.X”). To identify “dated” ID mappings, in which the cross-referenced ENSPs are no longer identical to canonical proteins from the 2018 UniProtKB release (current study's reference proteome for CpDAA positions and functional annotations), sequence distance (IDs from Method B) was scored using the Hamming normalized distance metric (Frederick, Sedlmeyer, & White, 1993) and the Levenshtein normalized distance metric (Yujian & Bo, 2007). Normalized scale is 0 to 1,

with 0 indicating identical Ensembl-UniProtKB proteins and 1 indicating significant differences between the two sequences.

Residue mapping to pathogenicity scores

CpDAA-containing UniProt protein IDs and residue positions were mapped to dbNSFPv4.0a for annotations of missense deleteriousness scores. Additionally, undetected cysteine and lysine positions in CpD proteins were also pulled from dbNSFPv4.0. Genomic coordinate keys (formatted as “chr_pos_ref_alt”) were made from the dbNSFP columns for GRCh37 and GRCh38 genome assemblies. Coordinate keys from dbNSFP were then used to map CADDv1.4 model annotation files. Missense overlapping cysteine and lysine codons in CpD proteins were required to have valid coordinates in both genome assemblies and annotations for all possible nonsynonymous SNVs (stop-gained missense consequences were filtered out from our analysis). The deleteriousness scores with no missing annotations for loss-of-cysteine and loss-of-lysine missense (CADD, fathmm-MKL, and DANN) were summarized by taking the max or mean of all nonsynonymous variants per cysteine and lysine codon in successfully annotated CpD proteins (see GitHub for python scripts).

Correlation of deleteriousness scores for CpD cysteine and CpD lysine missense variants

Relationship between missense deleteriousness prediction scores and chemoproteomic detection was assessed by Spearman's rank-order correlation using the SciPy stats module both for CpDAAs and for non-detected cysteine and lysine residues. A

nonparametric correlation test was chosen based on non-normal distributions of missense scores for cysteines and lysines in CpD proteins. All correlations are based on a subset of cysteine and lysine missense variants with no missing score annotations. CADD raw scores were used instead of the PHRED scores, with “CADD37” denoting raw score from the CADD GRCh37 model and “CADD38” denoting raw score from the CADD GRCh38 model.

Enrichment analysis of predicted and known pathogenic missense variants for cysteine and lysine residues in detected proteins

For the analysis with predicted deleteriousness scores, cysteine and lysine residues from 3,840 successfully annotated CpD proteins were filtered for Cys > Trp and Lys > Ile specific substitutions. Deleterious missense thresholds were set as follows: CADD PHRED scores from the GRCh38 model (CADD38) greater than or equal to 25, fathmm-MKL scores greater than or equal to 0.95, and DANN scores greater than or equal to 0.98. For each group, an odds ratio (OR) along with the 95% confidence interval (CI) was calculated using Fisher's exact test on a 2 × 2 contingency matrix. Evidence for statistical significance of association was determined using the Bonferroni-adjusted P-value cut-off of 0.004. For the analysis with ClinVar “pathogenic” and “likely pathogenic” variants, the downloaded ClinVar variant data were filtered for loss-of-cysteine and loss-of-lysine missense consequences (n = 2,225 pathogenic variants) by parsing the Human Genome Variation Society Sequence Variant Nomenclature column (HGVS, e.g., p.Cys36Trp). In total, 389 pathogenic variants overlapped the genomic coordinates of cysteines and lysines in 3,840 CpD proteins. For each group, an

estimate of fold enrichment or odds ratio (OR), along with the 95% confidence interval (CI) was obtained using Fisher's exact test on a 2 × 2 contingency matrix. Evidence for statistical significance of association was determined based on the Bonferroni-adjusted P-value cut-off of 0.0167.

Bootstrap analysis of CADD38 PHRED max codon scores

The bootstrapping procedure for calculating the 95% confidence interval of median CADD38 PHRED max codon scores and further characterizing the differences between low, medium, and highly reactive residues was performed as follows: original CADD38 max scores for each sub-group were resampled 20,000 times with replacement, with the median of each bootstrapped sample calculated. This process produced 20,000 samples with 895 low, 412 medium, and 94 high observations for CpD Cys, and 3,401 low, 660 medium, and 302 high observations for CpD Lys.

Mapping deleteriousness scores to protein structures

For the UniProtKB canonical proteins G6PD (P11413) and CASP8 (Q14790), CADD GRCh38 model PHRED scores for missense overlapping all amino acid positions were extracted and summarized by taking the max or mean of all missense scores per residue. Scores for all residue positions were extracted from the dbNSFPv4.0a file (see Reagents and Tools table). After checking the canonical protein positions against the cross-referenced Protein Data Bank (PDB) ID pulled from the UniProtKB website (PDB ID 3KJN for CASP8 and 2BH9 for G6PD), residue max CADD PHRED scores were

mapped to protein structure through assignment of scores as beta factor values of protein structure alpha carbons (GitHub for python script; Dataset EV21).

IsoTOP-ABPP sample preparation and analysis

IsoTop-ABPP samples were prepared as described previously (Weerapana et al 2010; Backus et al 2016). Briefly, cells were harvested and lysed by sonication in PBS. Proteomes were adjusted to 1 mg/ml. Samples were labeled for 1 h at ambient temperature with either 10 or 100 µM iodoacetamide alkyne (IA-alkyne, 5 µl of 1 or 10 mM stock in DMSO). Samples were conjugated by CuAAC to either the light (fragment treated) or heavy (DMSO treated) TEV tags (10 µl of 5 mM stocks in DMSO, final concentration = 100 µM), with TCEP (10 µl of fresh 50 mM stock in water, final concentration = 1 mM), TBTA (30 µl of 1.7 mM stock in DMSO/t-butanol 1:4, final concentration = 100 µM), and CuSO4 (10 µl of 50 mM stock in water, final concentration = 1 mM). After 1h, the samples were pelleted and the pellets sonicated in ice-cold methanol (500 µl) and combined pairwise. The pellets were solubilized in PBS containing 1.2% SDS (1 ml) with sonication and heating (5 min, 95°C) and any insoluble material was removed by an additional centrifugation step at ambient temperature (14,000 g, 1 min). Samples were then enriched on streptavidin resin (100 µl slurry) in PBS (10 ml) with rotating for 90 min. Beads were then washed (2× PBS and 2× water), resuspended in 6 M urea reduced (20 mM DTT), and alkylated (40 mM iodoacetamide). Samples were then diluted to 2 M urea and 6 µl (2 µg) reconstituted MS grade trypsin (Promega V5111) was added and the samples were allowed to digest overnight. The beads were then pelleted, washed (3× PBS and 3× water), and then resuspended in

75 µl TEV buffer (50 mM Tris, pH 8, 0.5 mM EDTA, 1 mM DTT). 5 µl TEV protease (80 µM) was added and the reactions were rotated for 7 h at 29°C. The samples were then cleaned using Micro Bio-Spin columns, desalted using Pierce C18 100 µl bed zip-tips, concentrated by speedvac and reconstituted in 20 µl 5% ACN and 1% formic acid.

Liquid chromatography tandem mass spectrometry (LC-MS/MS) analysis

The samples were analyzed by liquid chromatography tandem mass spectrometry using a Q Exactive™ mass spectrometer (Thermo Scientific) coupled to an Easy-nLC™ 1000 pump. Peptides were resolved on a C18 reversed phase column (3 µM, 100 Å pores), packed in-house, with 100 µm internal diameter and 18 cm of packed resin. The peptides were eluted using a 140-min gradient of buffer B in buffer A (buffer A: water with 3% DMSO and 0.1% FA; buffer B: acetonitrile with 3% DMSO and 0.1% FA) and a flow rate of 220 nl/min with electrospray ionization of 2.2 kV. The regular gradient includes 0–5 min from 1 to 5%, 15–130 min from 5 to 27%, 15–137 min from 27 to 35%, and 137–138 min from 35 to 80% buffer B in buffer A. Data were collected in data-dependent acquisition mode with dynamic exclusion (15 s), and charge exclusion (1, 7, 8, > 8) was enabled. Data acquisition consisted of cycles of one full MS scan (400–1,800 m/z at a resolution of 70,000) followed by 12 MS2 scans of the nth most abundant ions at resolution of 17,500.

Peptide and protein identification

The MS2 spectra data were extracted from a raw file using RAW Xtractor (version 1.1.0.22; available at http://fields.scripps.edu/rawconv/). MS2 spectra data were

searched using the ProLuCID algorithm (publicly available at http://fields.scripps.edu/yates/wp/?page_id=17 using a reverse concatenated, non-redundant variant of the Human UniProtKB database (release-2020_01). Cysteine residues were searched with a static modification for carboxyamidomethylation (+57.02146) and isoTOP differential modification at cysteine residues (+464.28595 for light and +470.29976 for heavy). Peptides were required to have at least one tryptic terminus, allowed one missed cleavage event and to contain the isoTOP modification. ProLuCID data were filtered through DTASelect (version 2.0) to achieve a peptide false-positive rate below 1%.

Proteomic data processing

Custom python and R scripts were implemented to filter and compile labeled peptide datasets. Peptides with one tryptic terminus were filtered out before further analysis. Unique proteins, unique residues (cysteines or lysines), and unique peptide-spectrum matches (PSMs) were quantified for each dataset, using unique identifiers. Unique proteins were established based on UniProtKB protein ID. Unique residues were classified by an identifier consisting of a UniProtKB protein ID and the residue number of the modified cysteine/lysine; residue numbers were found by aligning the peptide sequence to the corresponding UniProtKB protein sequence. Unique peptides were found based on sequences containing modified residue location. If a peptide was labeled at multiple residues, an identifier was generated for each protein ID and modified residue location. IsoTOP-ABPP ratios from each experiment were averaged and reported with ± SD.

Recombinant caspase-8 expression and purification

Recombinant caspase-8 (residues 217–479) without the CARD domain subcloned into pET23b (Novagen) with C-terminal His6-affinity tags was expressed as has been described (Backus et al 2016) previously. Site-directed mutagenesis (Liu & Naismith, 2008) was conducted as has been described previously, using the primers shown in the Reagents and Resources Table.

Caspase-8 activity assay

Caspase-8 assay was conducted with CASP8 activity assay kit (BioVision; K112-100), following the manufacturer's instructions. Briefly, recombinant protein was diluted to 500 nM into assay buffer (50 µl/well in a 96-well plate) following which IETD-AFC substrate (4 mM stock in DMSO of IETD-AFC) was added to each well (5 µl stock diluted into 50 µl assay buffer for a final concentration of 200 µM substrate) and the samples were incubated at ambient temperature for 1 h. Caspase activity was measured from the increase in fluorescence (excitation 380 nm emission 460 nm). Experiments were performed in triplicate. Background was calculated from samples lacking the recombinant caspase.

Data availability

Code for the mapping analysis and figures is available on the GitHub site https://github.com/mfpfox/MAPPING. All chemoproteomics datasets along with functional annotations are made available to download through the CpDAA database

https://mfpalafox.shinyapps.io/CpDAA/ an R Shiny-based web interface. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier: PXD022151 and https://doi.org/10.6019/PXD022151

2.5 Figures

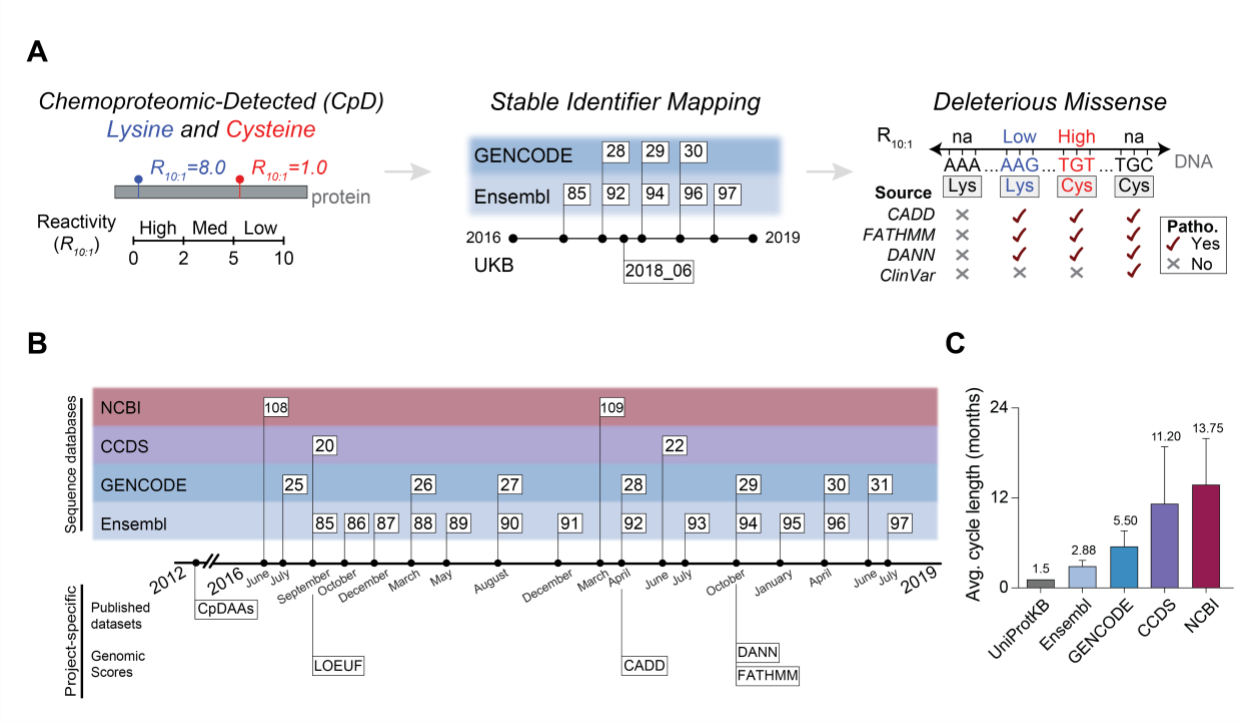

Figure 2-1. Landscape of sequence annotation information updates

A. Schematic representation of mapping chemoproteomic detected amino acids (CpDAAs) to pathogenicity scores.

B. Timeline of gene annotation database release dates and project-specific datasets, including Ensembl releases tested for compatibility to CpDAA coordinates based on

canonical UniProtKB protein sequences and the database reference corresponding to the genomic pathogenicity scores.

C. Average database release cycle length for releases between August 2013 - July 2019. All values are mean ± s.d. Total of 25 Ensembl, 13 GENCODE, 6 CCDS (homo sapien only), and 5 NCBI releases were counted. UniProtKB average cycle length was calculated using data provided by the UniProtKB website.

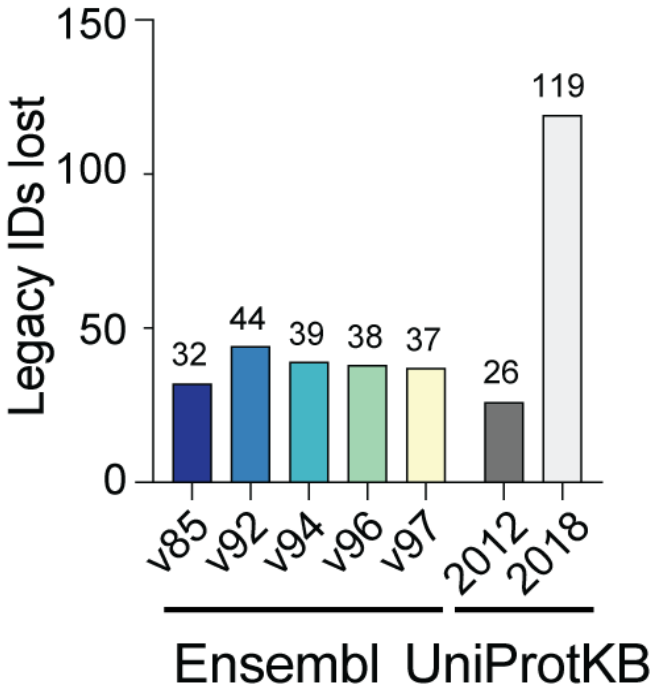

Figure 2-2. Data losses that result from re-mapping chemoproteomic datasets to new releases of Ensembl and UniProtKB

Shows the number of stable UniProtKB protein IDs from cysteine and lysine chemoproteomics studies in original legacy chemoproteomics dataset (4,119 Uniprot stable IDs in aggregate) (Hacker *et al*, 2017; Backus *et al*, 2016; Weerapana *et al*, 2010) that fail to map to IDs in more recent releases of Ensembl and UniProtKB. While

all Ensembl datasets showed similar losses, Ensembl v85 modestly outperformed more recent versions, consistent with the v85 release date being closest in time to the UniProtKB release on which legacy data was based.

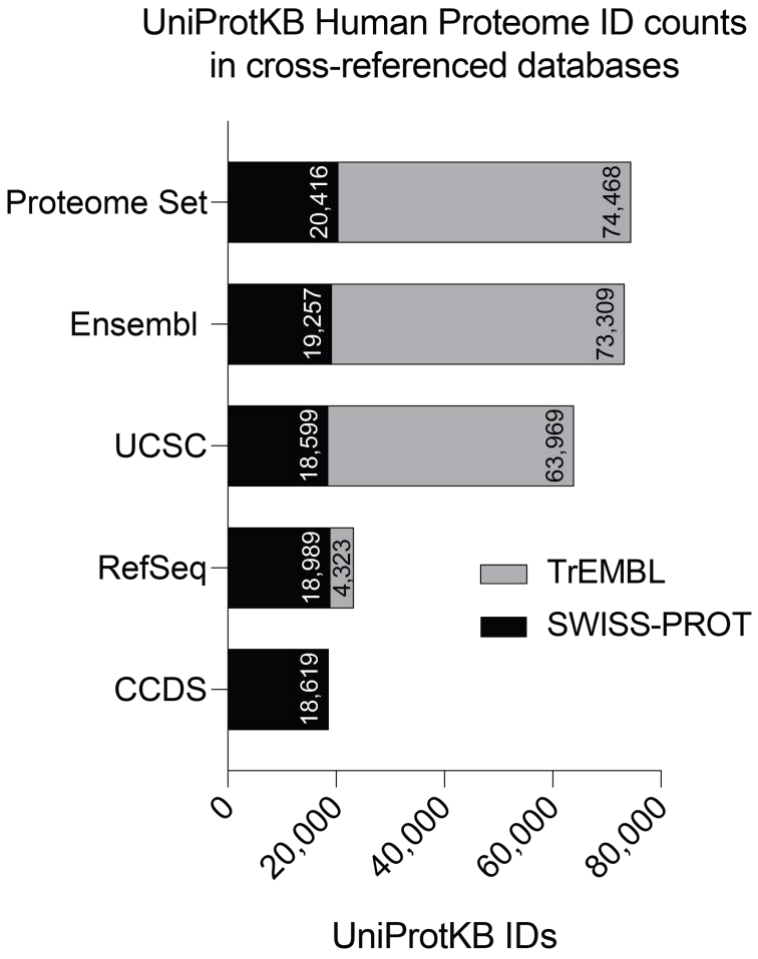

Figure 2-3. UniprotKB Human Proteome ID counts in cross-referenced databases

The UniProtKB/TrEMBL subset (automated translations of coding sequences) are shown in grey and the UniProtKB/Swiss-Prot subset (manually curated sequences) are shown in black. Ensembl, USCS, and RefSeq contain both automated (TrEMBL) and

manually curated (Swiss-Prot) entries. Sequences derived from the consensus coding sequence (CCDS) project are associated with the UniProtKB/Swiss-Prot subset.

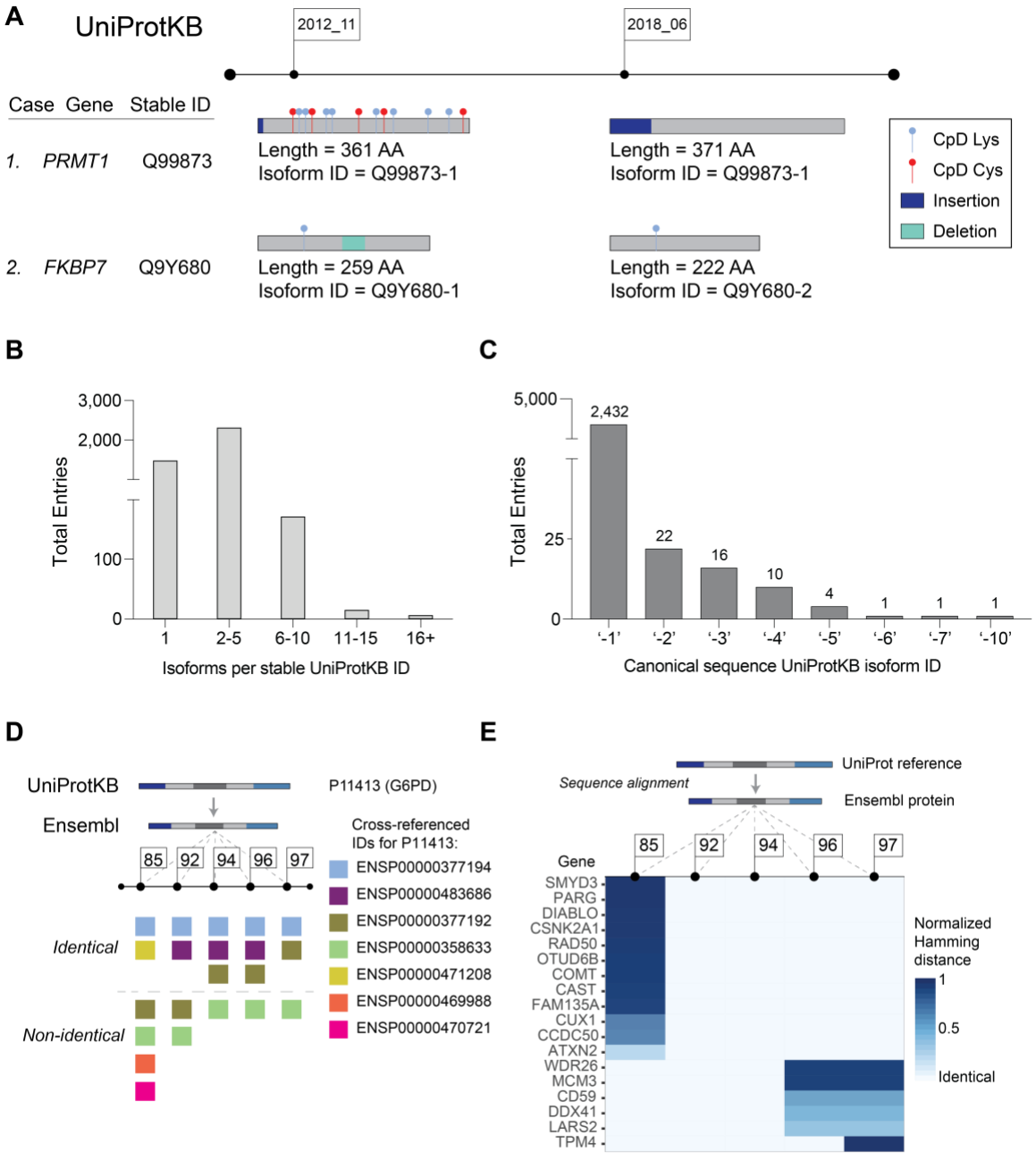

Figure 2-4. Challenges with residue-level mapping and UniProtKB canonical protein sequences

A. Schematic depiction of mapping scenarios from updating chemoproteomic-detected protein sequences using stable or versioned identifiers.

B. Distribution of number of isoforms per stable UniprotKB ID for 3,953 detected proteins.

C. Frequency of (not displayed) specific isoform name for 2,487 multi-isoform UniProtKB canonical proteins.

D. Schematic depiction of glucose-6-phosphate dehydrogenase (G6PD, UniProtKB ID P11413) cross-referencing both identical and non-identical sequences of Ensembl Stable IDs from five releases.

E. Heatmap of protein sequence distance scores for detected UniProtKB and cross-referenced Ensembl proteins from five releases. Each gene name corresponds to one unique stable Ensembl protein ID.

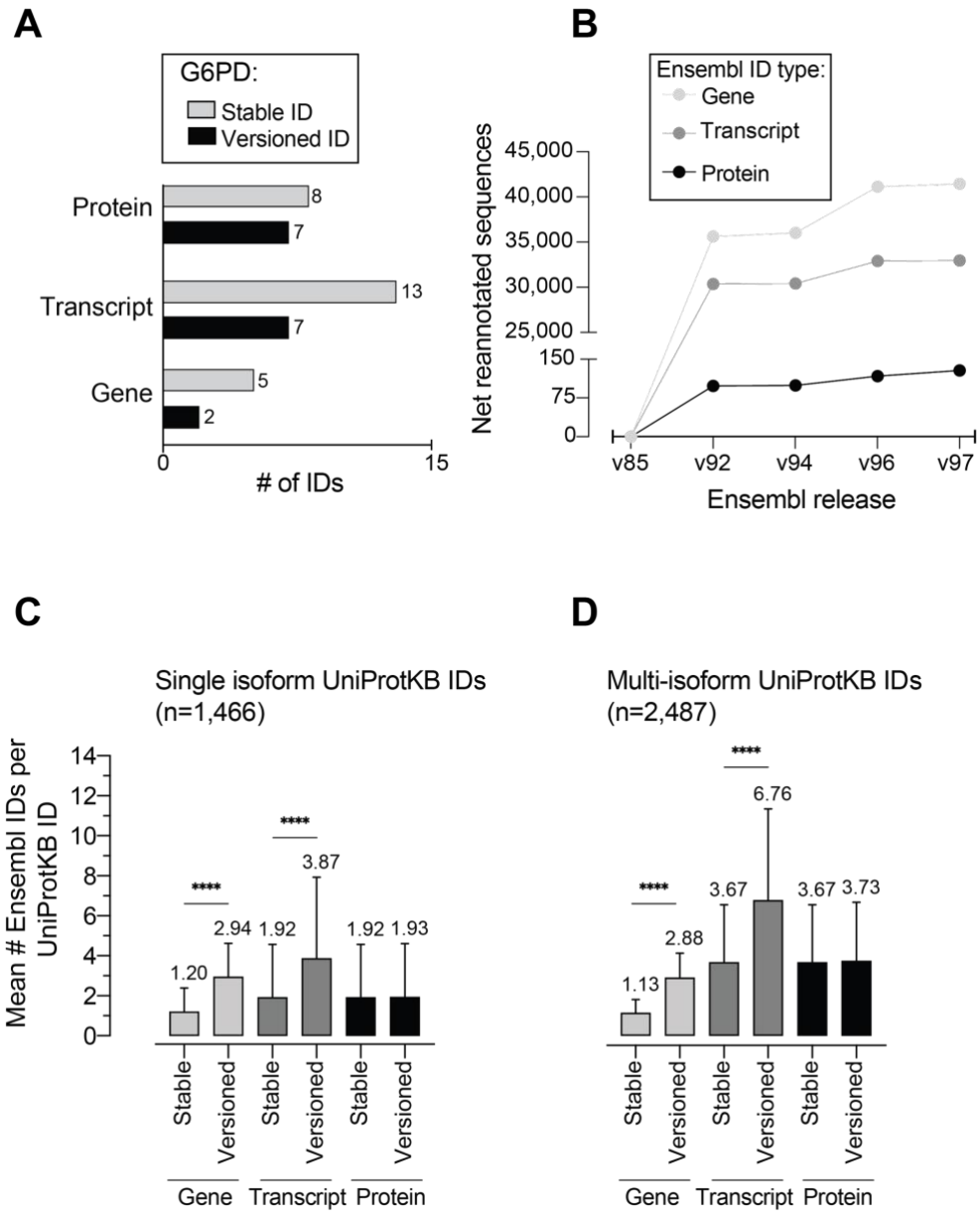

Figure 2-5. Mapping of Ensembl IDs to UniprotKB shows heterogeneity at gene, transcript and protein levels

A. Number of stable and versioned Ensembl gene, transcript and protein IDs for G6PD across all five Ensembl releases.

B. Cumulative sequence re-annotations for Ensembl gene, transcript, and protein IDs since the v85 release.

C-D. Average number of Ensembl gene, transcript, and protein IDs for (C) single isoform (n=1,466) and (D) multi-isoform (n=2,487) CpDAA UniProt entries. Bar plots represent mean values ± s.d. for the number of Ensembl IDs per stable UniProtKB ID. Statistical significance was calculated using an unpaired Student's T-test, **** p-value <0.0001.

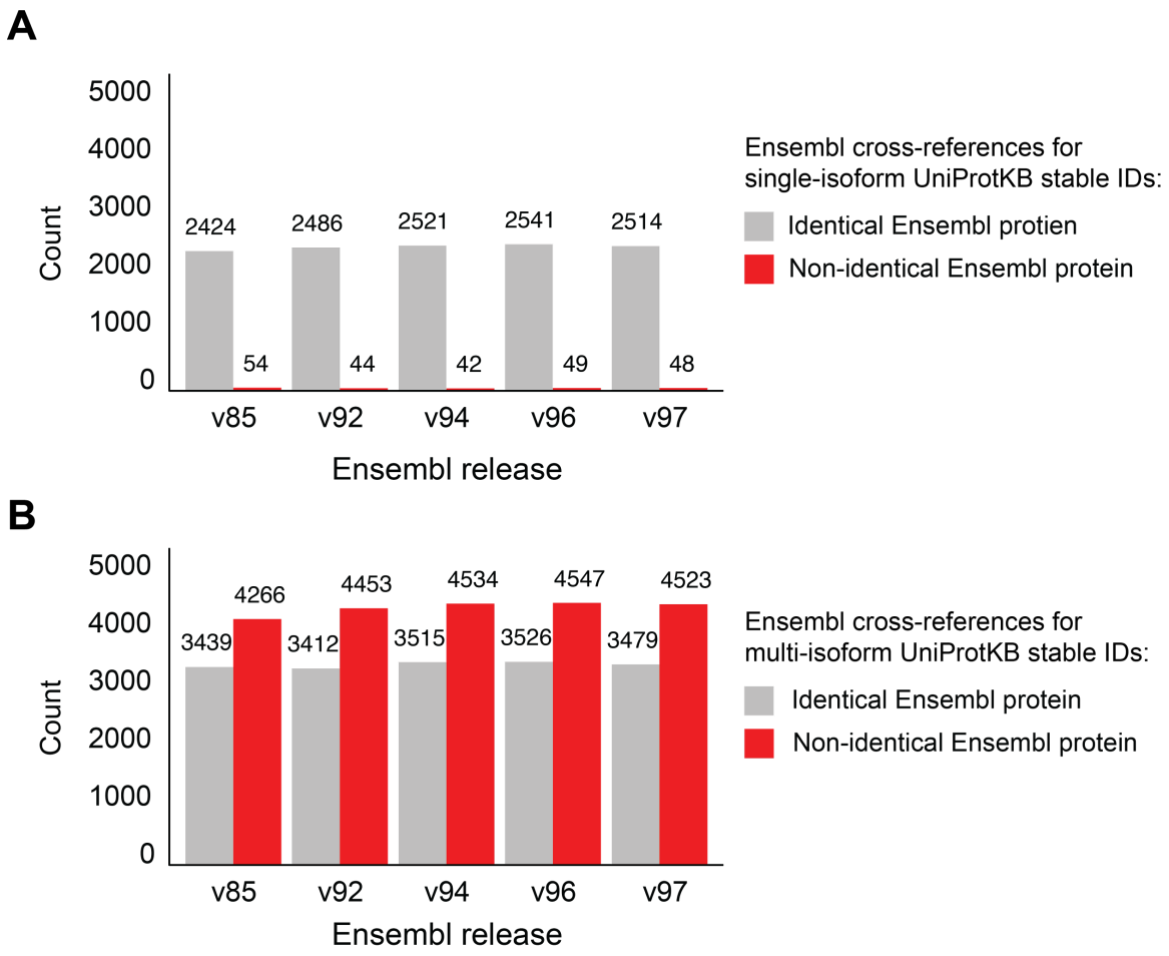

Figure 2-6. Comparison of single and multi-isoform UniProtKB protein cross-references to Ensembl proteins, using the Ensembl xref files

Using five Ensembl xref files (Materials and Methods, Method A) containing only stable ID cross-references to UniProtKB IDs, protein sequences were compared for A) 1,466 single isoform UniProKB IDs and B) 2,487 multi-isoform UniProKB IDs contained in our CpDAA-containing protein dataset.

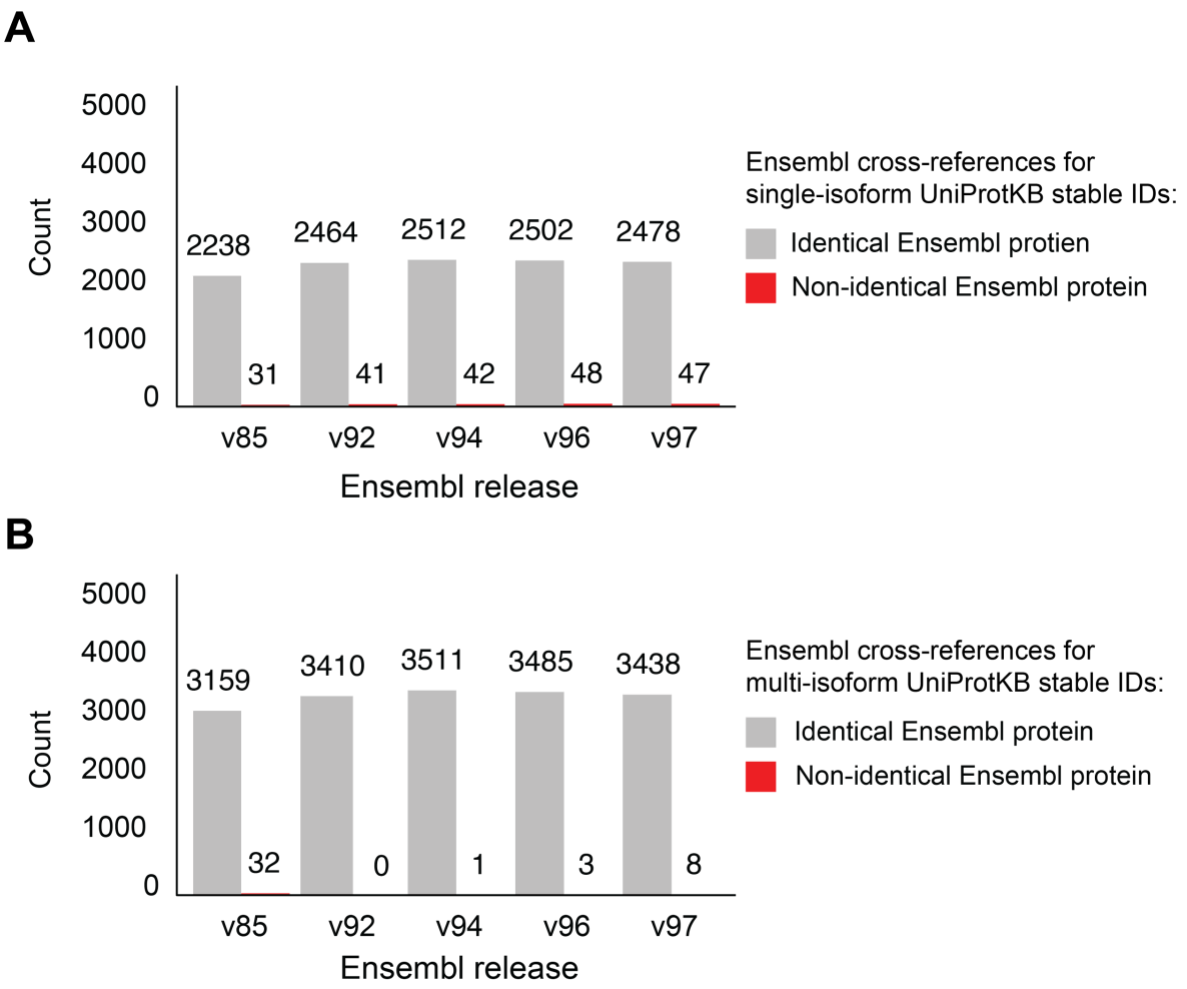

Figure 2-7. Comparison of single and multi-isoform UniProtKB protein cross-references to Ensembl proteins, using the UniProtKB mapping file

Using UniProtKB mapping file (Materials and Methods, Method B) provided canonical protein isoform ID cross-references to Ensembl stable protein IDs. Comparisons between UniProtKB canonical proteins from 2018_06 release were made to Ensembl proteins from five releases. Results of sequence identity comparison was performed for A) 1,466 single isoform UniProtKB IDs and B) 2,487 multi-isoform UniProKB IDs contained in our CpDAA-containing protein dataset.

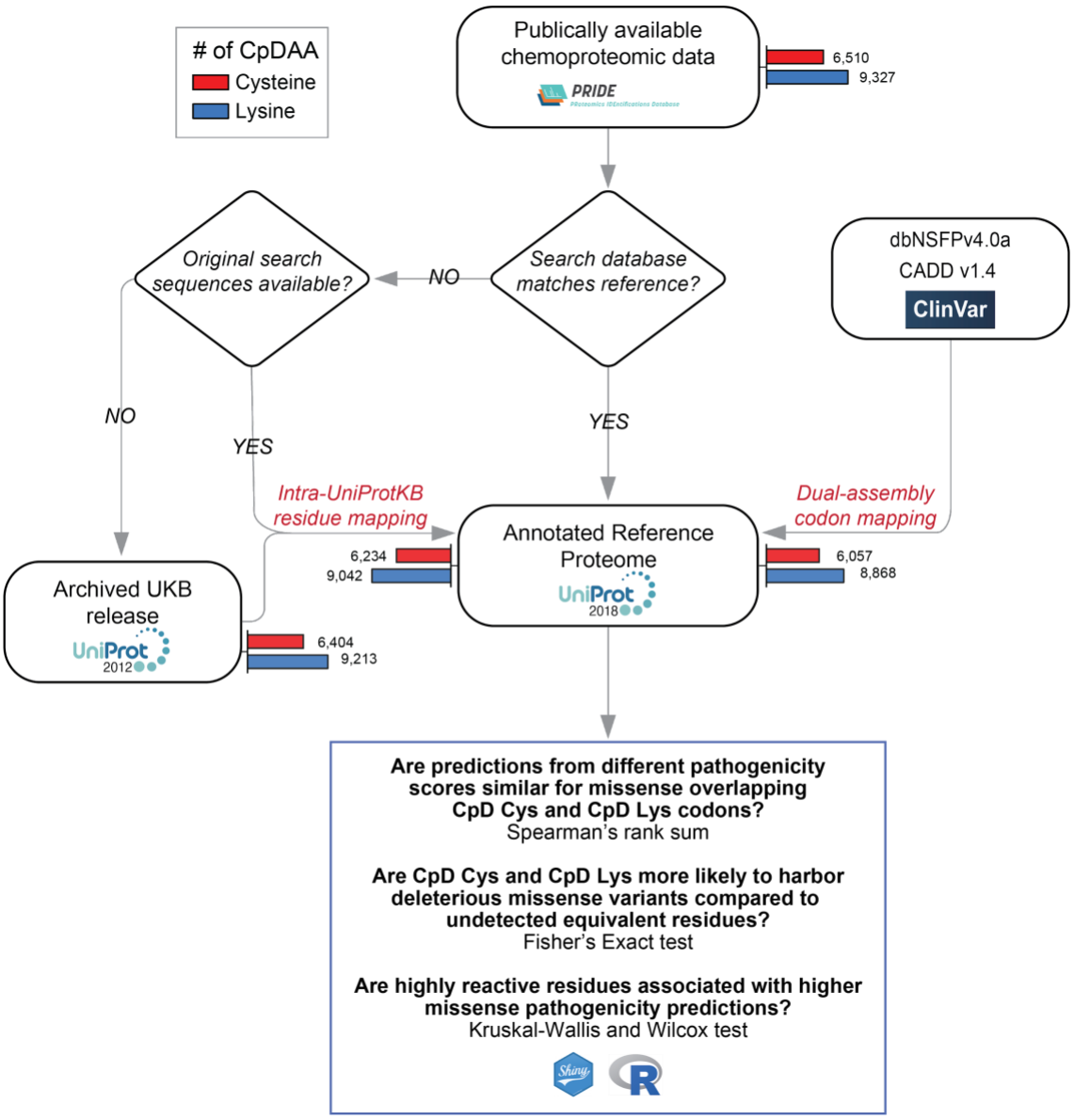

Figure 2-8. Flowchart of the mapping strategy and data analysis

CpD cysteines and lysines from three publicly available datasets were processed and filtered according to our optimized mapping pipeline. Number of CpD cysteines (red) and CpD lysines (blue) retained following each step shown as barplots.

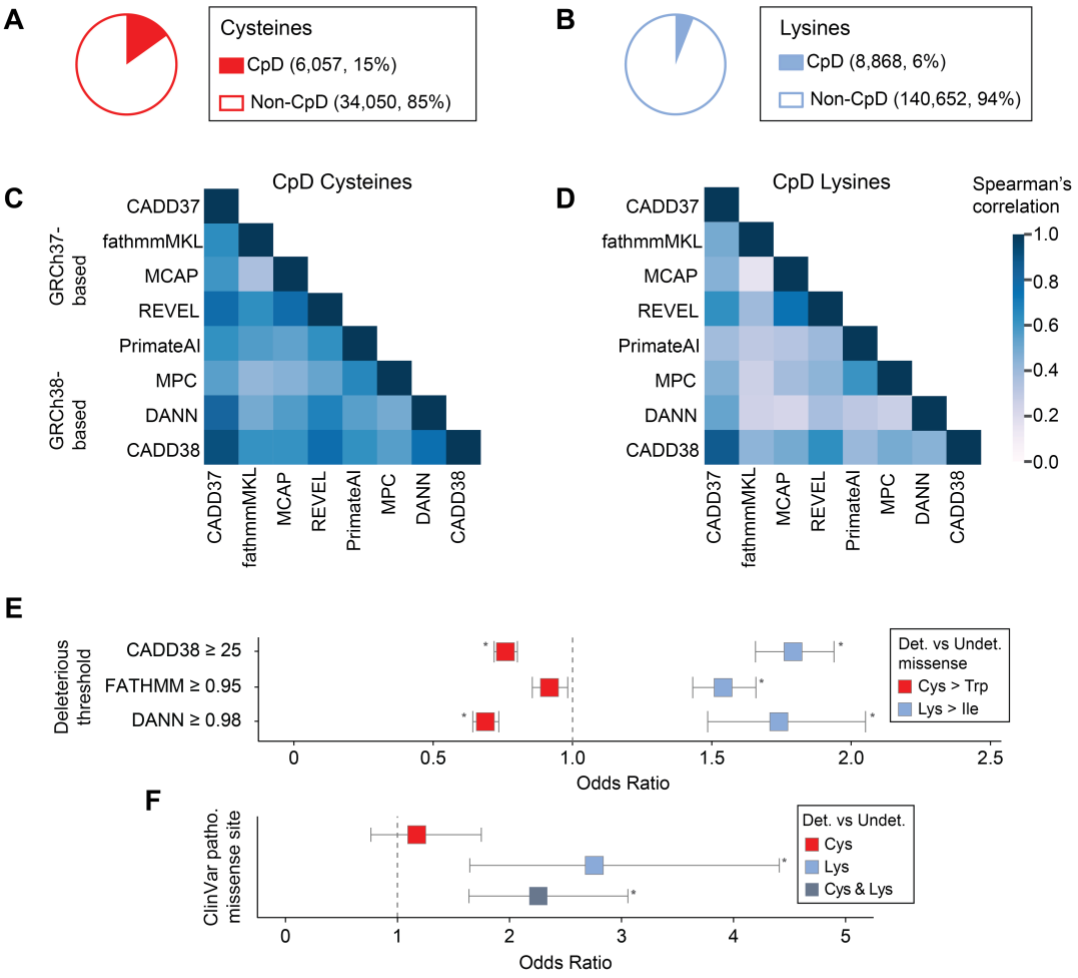

Figure 2-9. Analysis of pathogenic missense at Detected versus Undetected Cysteines and Lysines

A-B. Aggregate number of detected and undetected cysteines (A) and lysines (B) in 3,840 CpDAA-containing proteins.

C. Heatmap of missense score correlations for all possible non-synonymous SNVs at CpD Cysteine (29,541 missense) for eight pathogenicity scores. Overall, Spearman's rank r were between 0.36 and 0.91.

D. CpD Lysine (41,850 missense) heatmap for missense score correlations for all possible non-synonymous SNVs. Spearman's rank r between 0.16 and 0.81. Color

intensity represents two-tailed Spearman's rank-order correlation coefficients between 0 and 1.

E. Odds of predicted deleterious Cys>Trp (red) missense at detected (n=6,057) versus undetected (n=34,049) residues in 3840 detected proteins. Deleterious missense defined by CADD38, FATHMM, and DANN score thresholds (y axis). CADD38 OR = 0.76, p = 3.40e-22; FATHMM OR = 0.92 , p = 0.02; DANN OR = 0.690 , p = 6.69e-26. Odds of predicted deleterious Lys>Ile (blue) missense at detected (n=3,581) versus undetected (n=63,385) residues in 3840 detected proteins. CADD38 OR = 1.80, p = 1.03e-53; FATHMM OR = 1.55, p = 3.47e-33; DANN OR = 1.75, p = 9.21e-14. *p < 0.0042 Bonferroni adjusted (two-tailed Fisher's Exact test)

F. Odds of ClinVar pathogenic variant overlapping detected (6,057 Cys; 8,868 Lys) versus undetected (34,050 Cys; 140,652 Lys) residues in 3,840 detected proteins. Cys detected in ClinVar pathogenic site (red, OR = 1.17, p = 0.457) and Lys detected at ClinVar Pathogenic site (light blue, OR = 2.76, p = 1.03e-04). Combined Cys and Lys (dark blue, OR = 2.26, p = 9.99e-07) *p < 0.0167 Bonferroni adjusted (two-tailed Fisher's Exact test).

Data information: In (E-F), 95% confidence intervals (line segments) and odds ratios (squares).

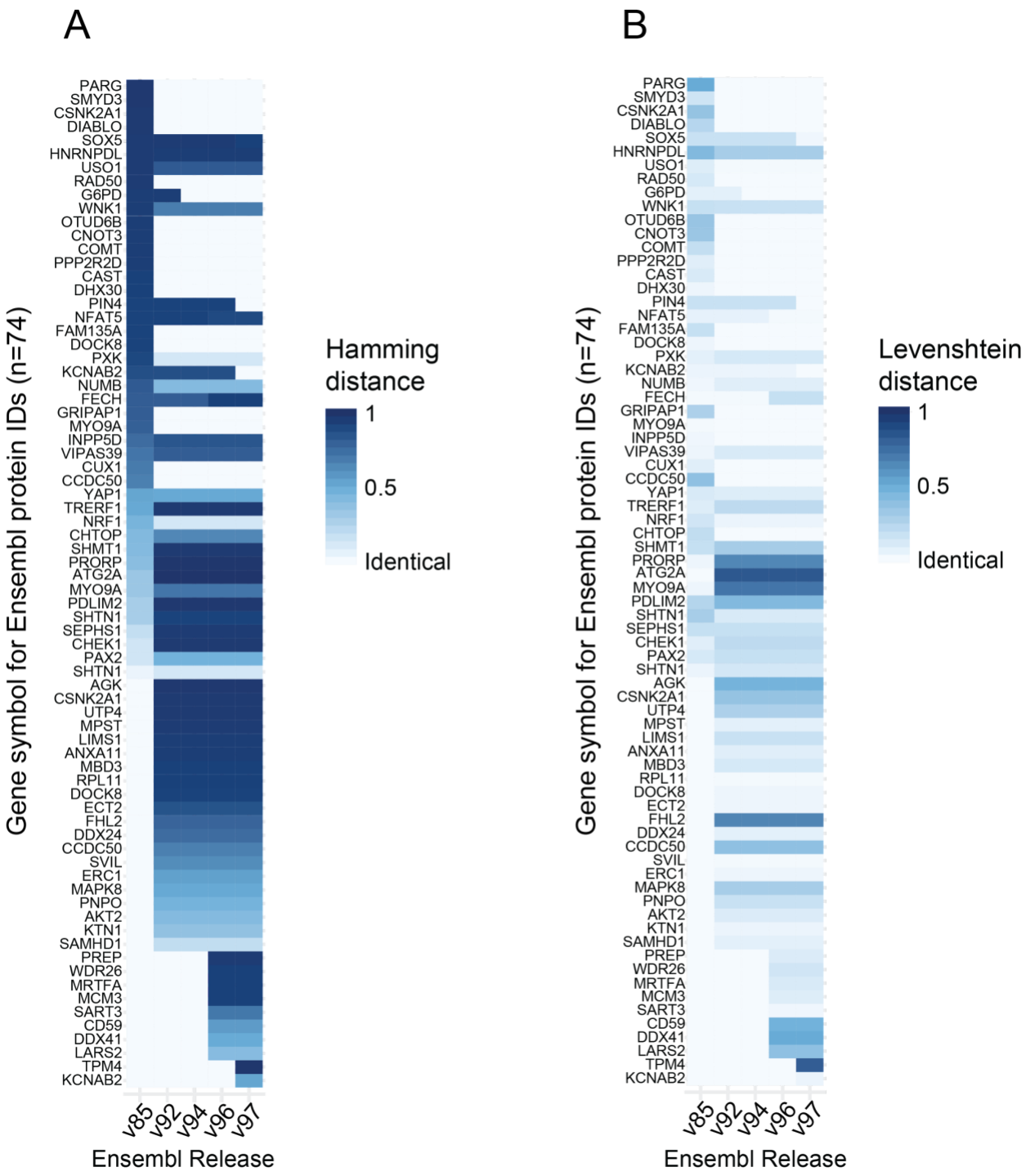

Figure 2-10. Sequence similarity between UniProtKB protein sequences and protein sequences associated with Ensembl stable IDs across releases

Heatmaps show A) normalized Hamming distance and B) normalized Levenshtein distance for sequence alignments of the protein sequences associated with the top 74

stable Ensembl gene, transcript, and protein IDs with an identical cross-referenced Ensembl protein sequence in one release, but non-identical sequences in additional releases. Scores range from 0 to 1, with 0 indicating identical to the canonical sequence in the 2018 UniProtKB CCDS release.

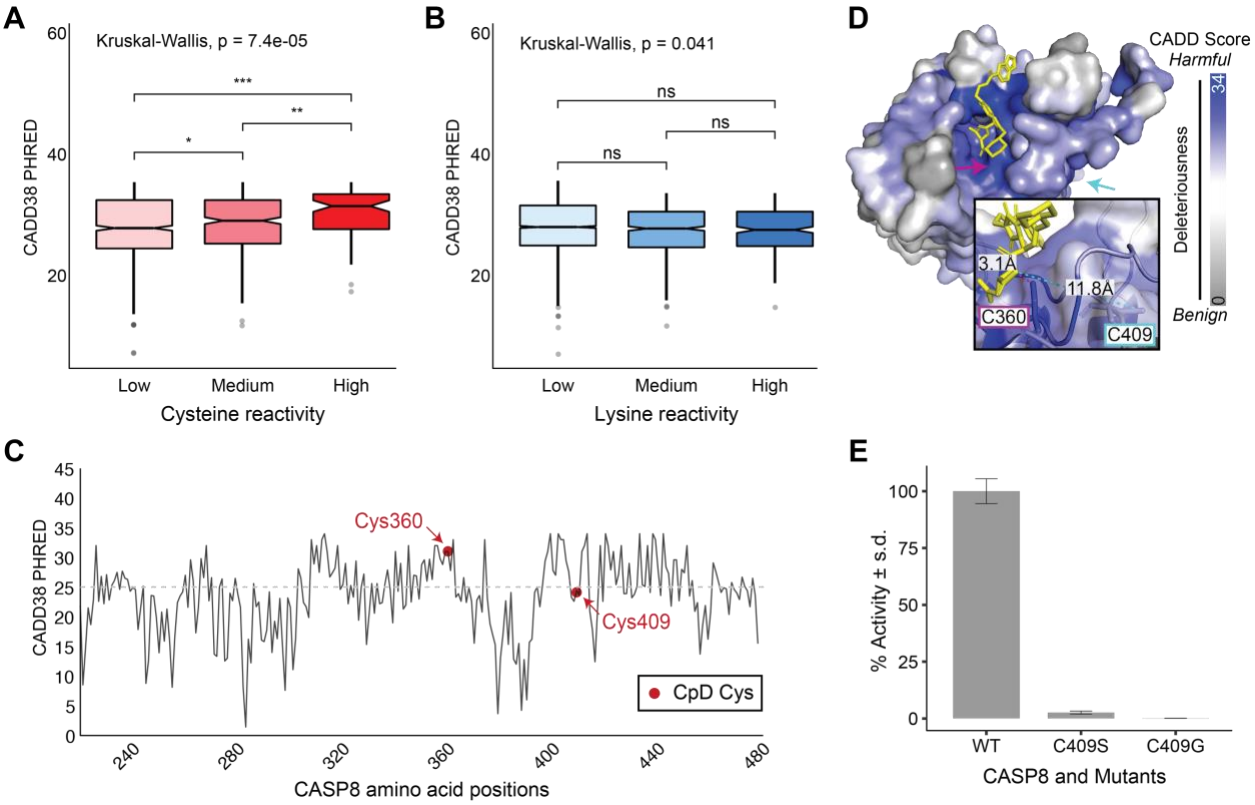

Figure 2-11. Association between amino acid reactivity and CADD score

A-B. Distribution of the max CADD38 PHRED (model for GRCh38) scores for (A) cysteine (n=1,401) and (B) lysine (n=4,363) CpDAAs of low, medium, and high intrinsic reactivities, defined by isoTOP-ABPP ratios, low ($R_{10:1}$>5), medium (2<$R_{10:1}$<5), high (R<2) (Weerapana et al. 2010; Hacker et al. 2017). Kruskal-Wallis nonparametric test to examine reactivity group difference, Wilcox test used for pairwise comparisons (BH-adjusted p-values, *$p.\ adj$ = 0.04, **$p.\ adj$ = 0.0037, ***$p.\ adj$ = 0.00013). Median of the

CADD38 max codon scores with bootstrapped 95% confidence intervals for reactive groups are: low CpD Cys 27.3 [26.9, 28.0], medium CpD Cys 28.55 [27.80, 29.05], high CpD Cys 31 [28.8, 32.0], low CpD Lys 29.5 [29.3, 29.6], medium CpD Lys 29.25 [28.85, 29.50], high CpD Lys 29.05 [28.50, 29.55].

C. Shows CADD38 max codon scores for nonsynonymous SNVs at residues 220-479 of CASP8 (UniProt ID Q14790).

D. Crystal structure of CASP8 (PDB ID: 3KJN) highlighting C360 and C409. Bound covalent inhibitor B93 in yellow, with distance between detected cysteines and inhibitor measured in Angstroms. Protein surface color represents CADD38 max codon scores. Image generated in PyMOL (R. H. B. Smith, Dar, and Schlessinger, n.d.; DeLano and Others 2002).

E. Activity of recombinant caspase-8 protein assayed using fluorogenic IETD-AFC substrate. Percentage activity shown relative to wild-type (WT) protein for three replicate experiments.

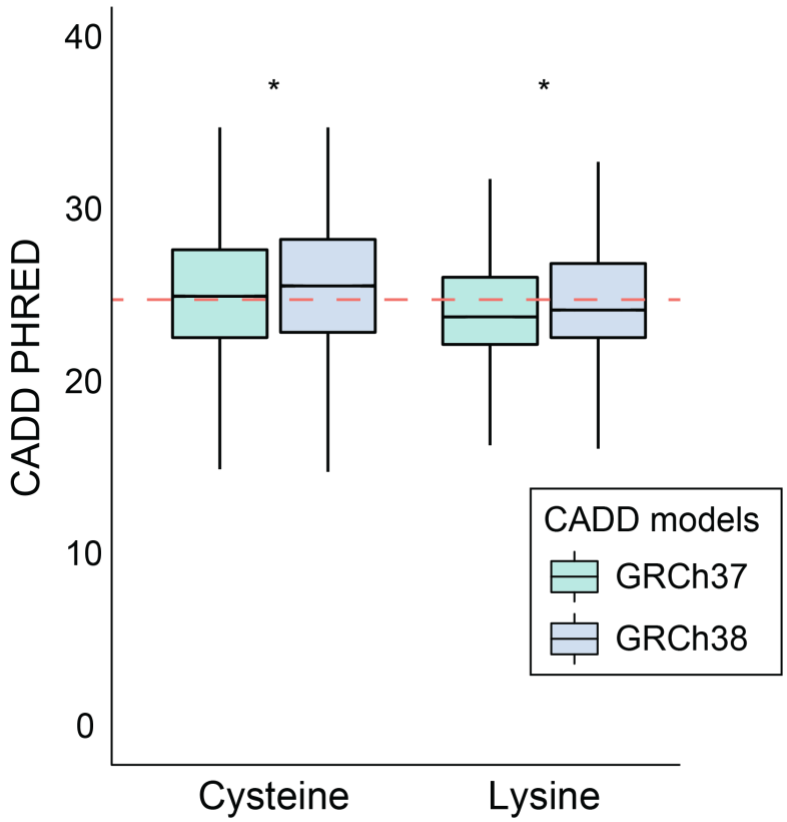

Figure 2-12. Comparison of GRCh37 and GrCh38 CADD models for loss of cysteine and loss of lysine

Loss of cysteine (n= 280,748) and loss of lysine (n= 1,046,638) missense overlapping coordinates of residues in 3,840 detected proteins. Deleterious missense threshold for CADD PHRED score of 25 marked by red dashed line. Cysteine missense score average is 24.34 +- 5.88 s.d. for CADD GRCh37 model (green) and 25.51 +- 5.10 s.d. for CADD GRCh38 model. Lysine missense score average is 23.55 +- 4.97 s.d. for CADD GRCh37 model and 24.70 +- 4.14 s.d. for CADD GRCh38 model. Wilcox test used for pairwise comparisons, * = p-value <2e-16 (CADD38-CADD37 PHRED score mean difference of 1.16 +- 2.56 s.d. for all cysteine and lysine residues in 3,840 detected proteins).

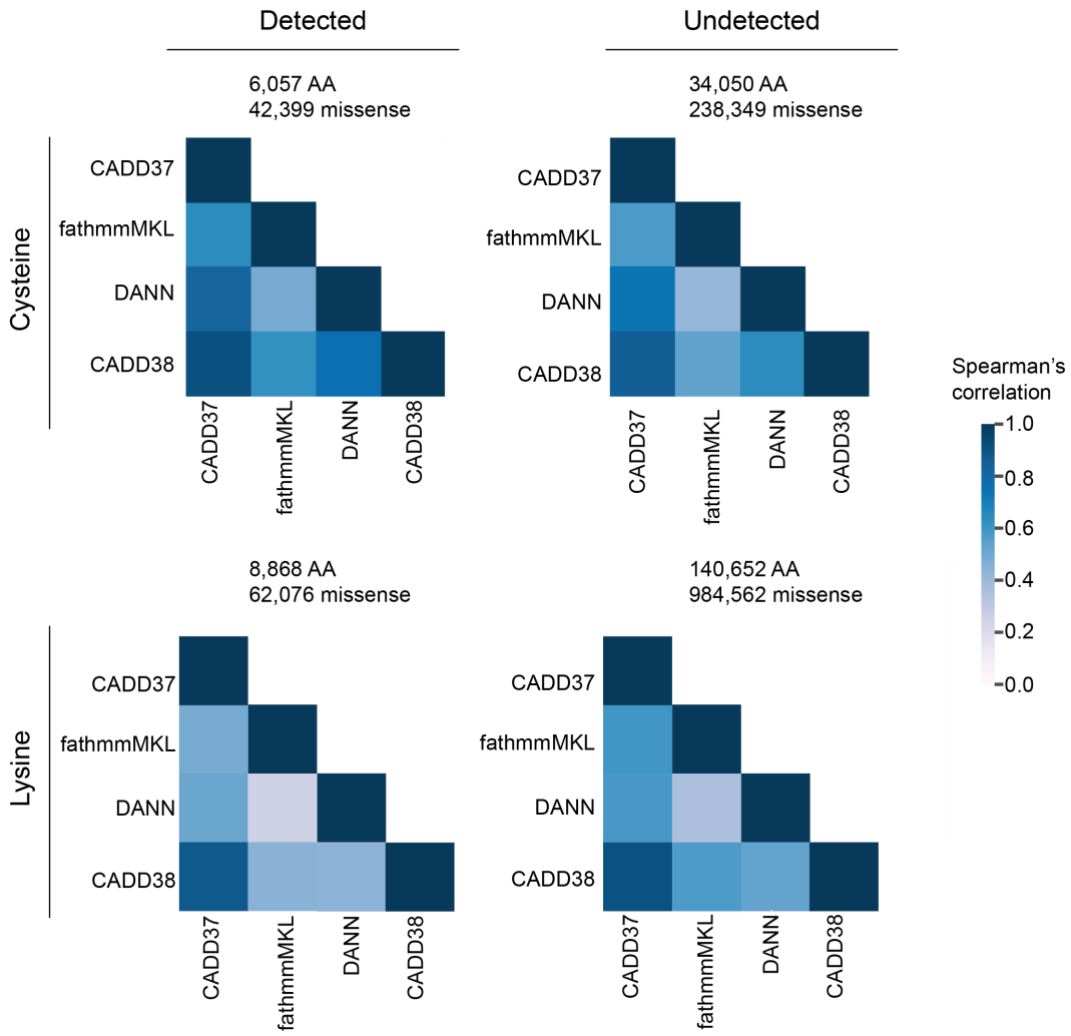

Figure 2-13. Correlation of pathogenicity scores for all possible non-synonymous SNVs at codons of detected or undetected cysteine and lysine residues

Heatmap represents two-tailed Spearman's rank-order correlation coefficients for all possible non-synonymous SNVs at codons of detected or undetected cysteine and lysine residues in 3,840 detected proteins. Only the subset of scores that provide pathogenicity annotations for all possible non-synonymous variants were included in this analysis.

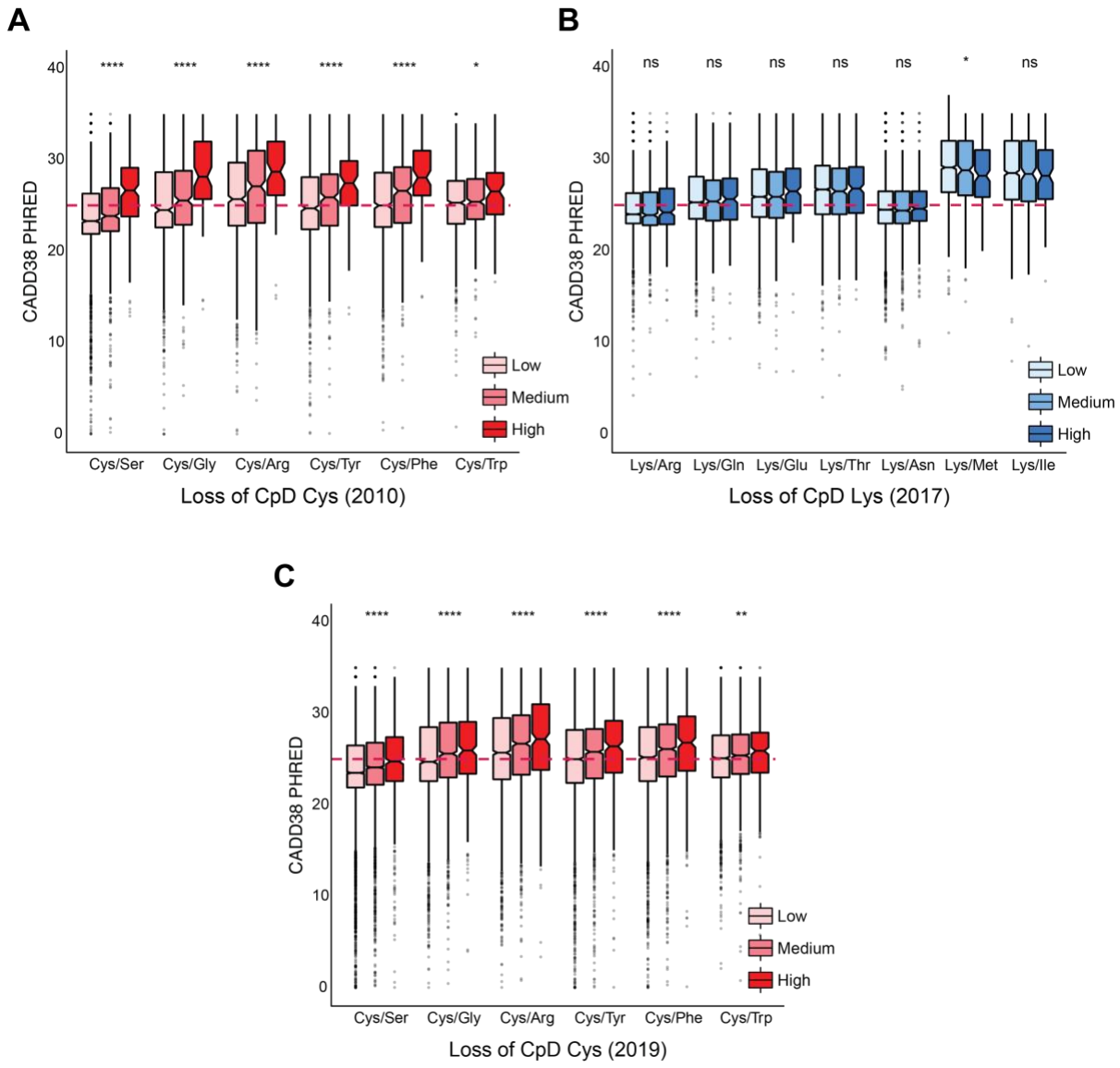

Figure 2-14. CADD38 PHRED scores for all possible missense variants at CpD cysteine and lysine codons, ordered by Grantham score

Distribution of CADD38 (model for GRCh38) PHRED scores for cysteine and lysine CpDAAs of Low, Medium, and High intrinsic reactivities, defined by isoTOP-ABPP ratios, Low ($R_{10:1}$>5), Medium (2<$R_{10:1}$<5), High ($R_{10:1}$<2)(Weerapana et al, 2010; Hacker et al, 2017). A) Enrichment of predicted deleterious missense variants for highly reactive cysteine residues identified by (Weerapana *et al*, 2010). B) No enrichment of predicted deleterious missense variants for highly reactive lysine residues identified by

(Hacker *et al*, 2017). C) Enrichment of predicted deleterious missense variants for highly reactive cysteine residues identified in the current study. Kruskal-Wallis nonparametric test to examine reactivity group difference, **p value* = 0.01, ***p value* = 0.003, *****p value* = <1.1e-06.

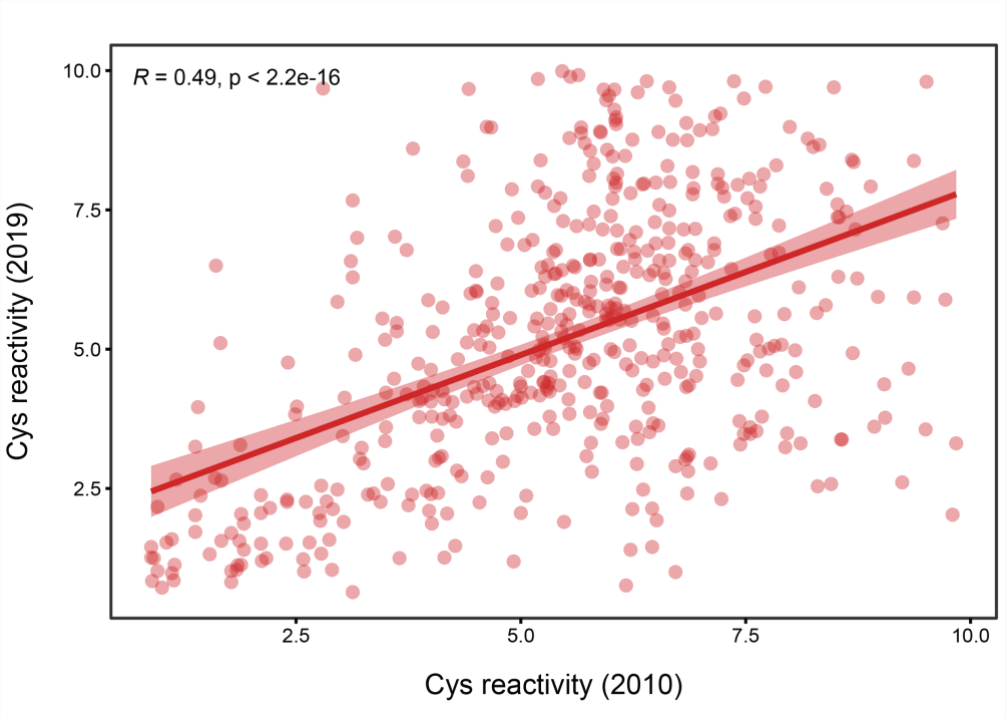

Figure 2-15. Correlation of cysteine reactivity between different chemoproteomic datasets

Total of 502 CpDAA are shared between the 2010 CpD Cys (Weerapana *et al*, 2010) and 2019 CpD Cys reactivity dataset. Pearson's correlation coefficient (R) = 0.49, p value < 2.2e-16, and 95% confidence interval of coefficient [0.425, 0.558].

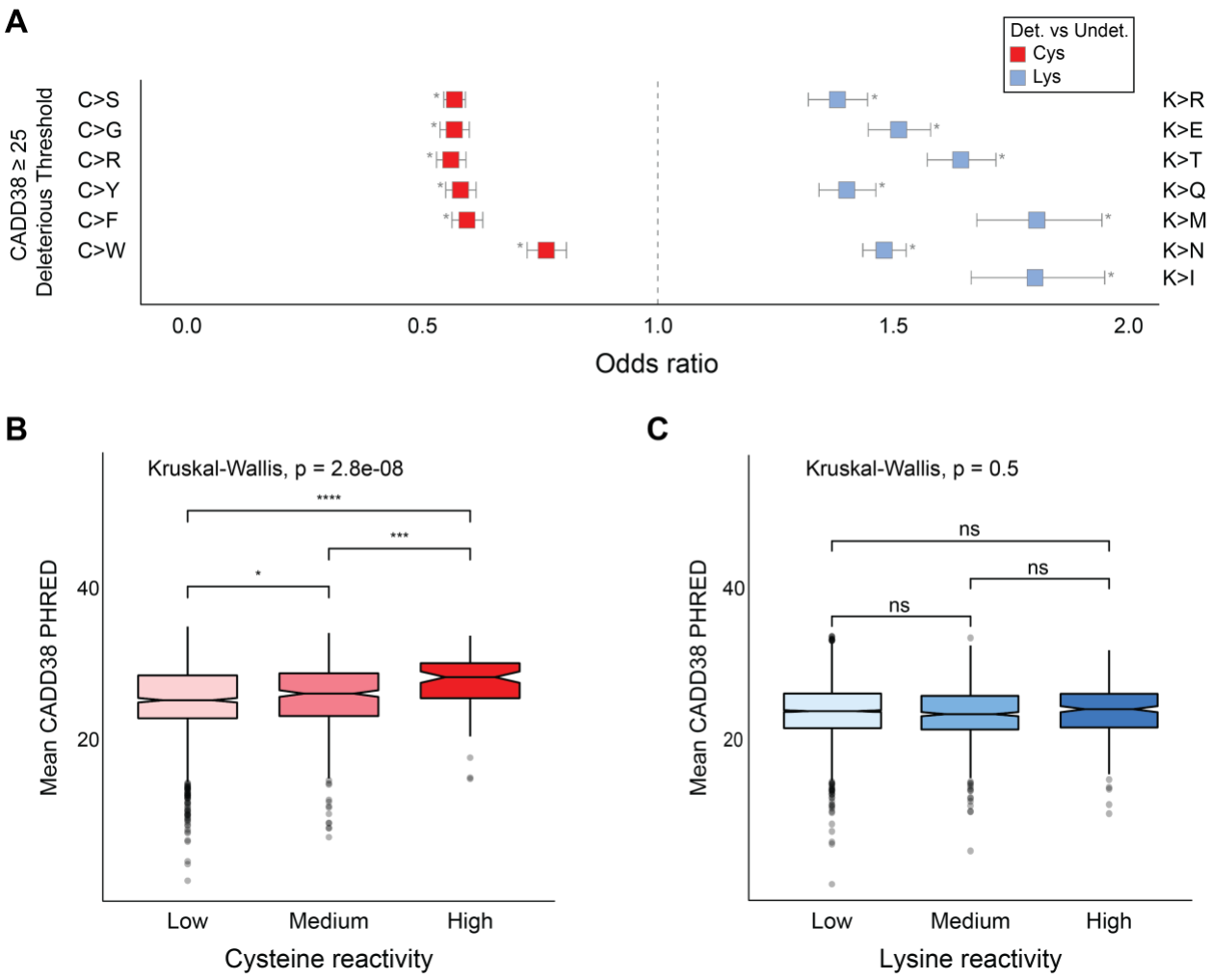

Figure 2-16. Assessment of missense pathogenicity between detected-undetected and reactivity groups for CPD cysteine and lysine residues

A. Enrichment of predicted deleterious missense variants for detected versus undetected cysteine (red) and lysine (blue) missense variants in 3,840 proteins. Missense types in order of increasing Grantham score. Odds Ratio (OR) for all possible nonsynonymous SNVs at cysteine codons between 0.56-0.76, at lysine codons fall between 1.38-1.80. 95% confidence intervals (line segments) and odds ratios (squares), two-tailed Fisher's Exact test, *p < pcut-off, 0.0019 Bonferroni corrected (0.05/26).

B-C. Distribution of mean CADD38 (model for GRCh38) PHRED scores for (B) cysteine (n=1,401) and (C) lysine (n=4,363) CpDAAs of low, medium, and high intrinsic

reactivities, defined by isoTOP-ABPP ratios, Low ($R_{10:1}$>5), Medium (2<$R_{10:1}$<5), High ($R_{10:1}$<2)(Weerapana *et al*, 2010; Hacker *et al*, 2017). Kruskal-Wallis nonparametric test to examine reactivity group difference, Wilcox test used for pairwise comparisons (BH-adjusted p values, **p. adj* = 0.013, ****p. adj* = 2.80e-05, *****p. adj* = 5.30e-08).

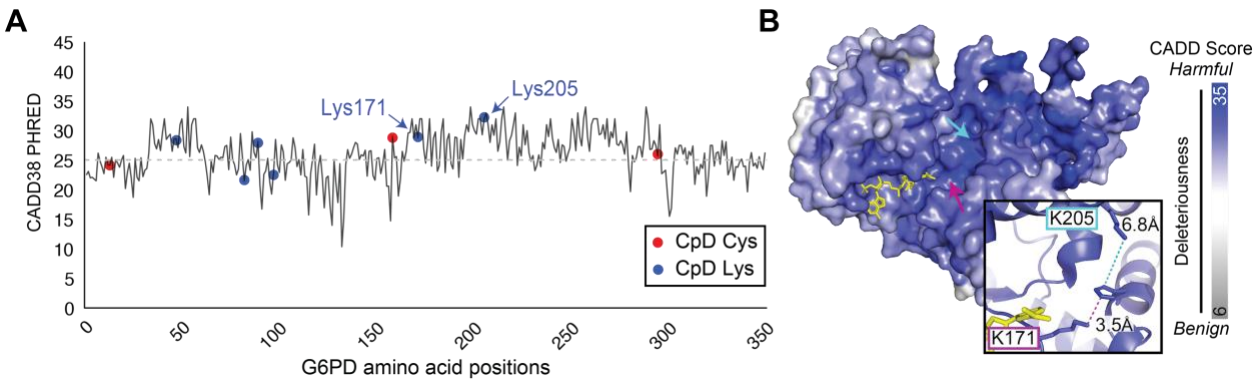

Figure 2-17. Functional Validation of reactive lysine in G6PD

A. Shows CADD38 max codon missense scores for residues 1-350 of G6PD (UniProt ID P11413). CpD K205 has the highest score out of all positions in protein. CpDAA positions above CADD38 deleterious threshold (grey dash line) include K47, K89, C158, K171, K205, and C294

B. Crystal structure of G6PD (PDB ID: 2BH9) shows K205 and K171 located within the enzyme active site. $NADP^+$ cofactor shown in yellow. Surface colored by CADD38 max codon missense scores. Image generated in PyMOL(Smith *et al*; DeLano & Others, 2002).

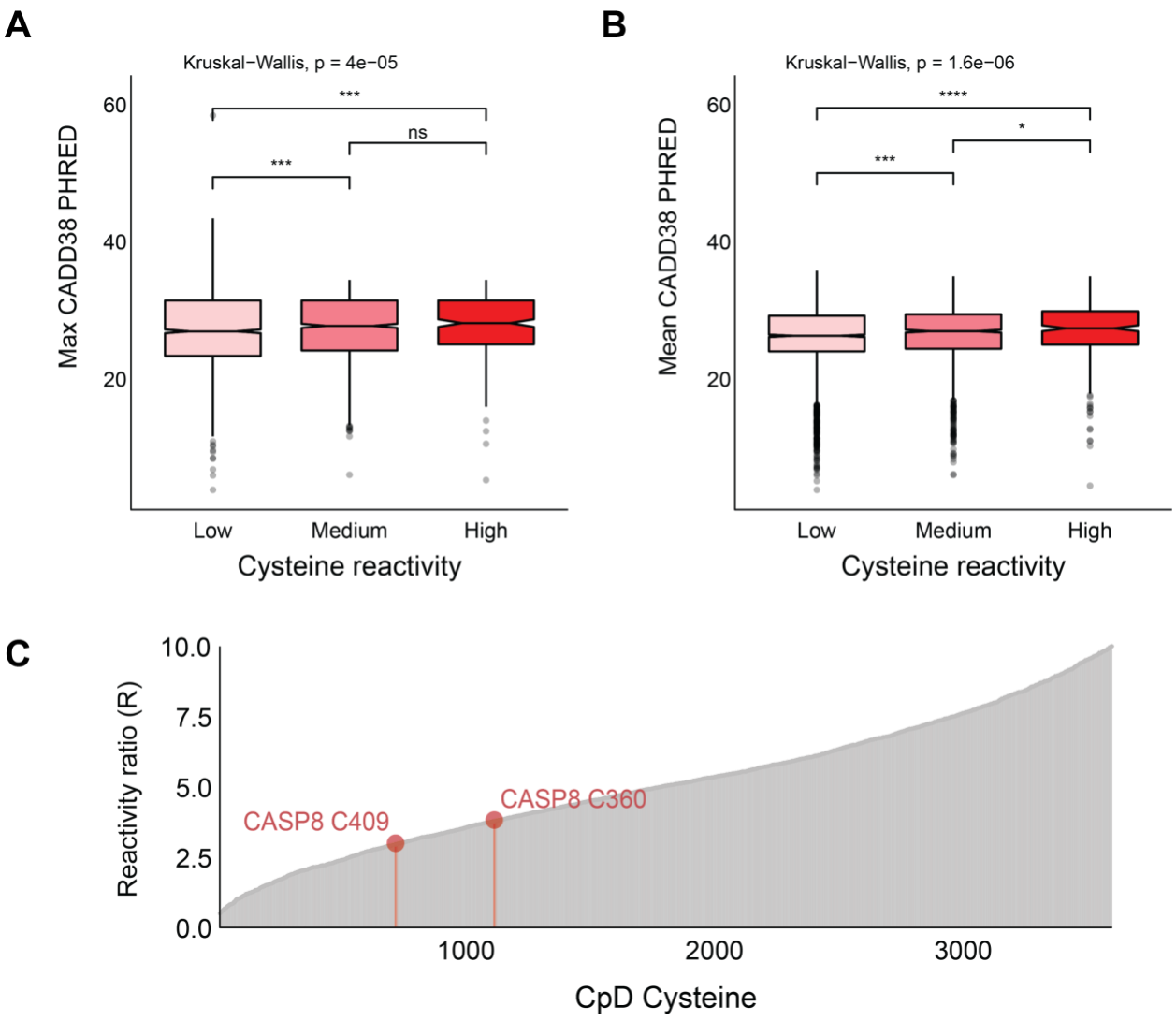

Figure 2-18. 2019 Cysteine Chemoproteomics Data identify caspase-8 residues that are pathogenic

Association between cysteine reactivity levels and CADD score for the 2019 Chemoproteomics Test Data Set. CADD38 scores were taken as either the max score for a missense change at that codon (A) or the mean score for all missense changes at the codon (B). The results show similar results, with increasing CADD predicted pathogenicity as reactivity increases.

A. Distribution of the max CADD38 (model for GRCh38) PHRED scores for cysteine of Low (n=2,247), Medium (n=1,448), and High (n=322) intrinsic reactivities, defined by isoTOP-ABPP ratios, Low ($R_{10:1}$>5), Medium (2<$R_{10:1}$<5), High ($R_{10:1}$<2)(Weerapana *et al*, 2010; Hacker *et al*, 2017). Kruskal-Wallis nonparametric test to examine reactivity group difference, Wilcox test used for pairwise comparisons (BH-adjusted p values, Low vs Med ****p. adj* = 0.00099, Low vs High ****p. adj* = 0.00086).

B. Association between 2019 cysteine reactivity data and mean CADD score. Distribution of the mean CADD38 (model for GRCh38) PHRED scores for cysteine of Low (n=2,247), Medium (n=1,448), and High (n=322) intrinsic reactivities, defined by isoTOP-ABPP ratios, Low ($R_{10:1}$>5), Medium (2<$R_{10:1}$<5), High ($R_{10:1}$<2) (Weerapana *et al*, 2010; Hacker *et al*, 2017). Kruskal-Wallis nonparametric test to examine reactivity group difference, Wilcox test used for pairwise comparisons (BH-adjusted p values, Low vs Med ****p. adj* = 4.0e-04, Med vs High **p. adj* = 0.023, Low vs High *****p. adj* = 3.90e-05.

C. Plot of cysteine reactivity ratios for 3590 out of 4017 total profiled residues in 2019 isoTOP-ABPP study. Represented are 322 High, 1448 Medium, and 1820 Low threshold cysteines.

2.6 Tables

Table 2-1. Definitions of key terms

	Term	Definition	References
1	Database update	Updated compilation of database resources, typically driven by gene or transcript re-annotation projects.	(Breuza *et al*, 2016; Potter *et al*, 2004)
2	Cross-reference	Referred to as 'xref' by UniProt and Ensembl, these files contain ID translations to equivalent sequences in other databases. These translations are what many mapping tools reference in order to translate user input.	(McGarvey *et al*, 2019; Ruffier *et al*, 2017)
3	Stable ID	The main citable identifier type from Ensembl and UniProtKB (primary accession). Ensembl IDs lack version number extensions (".#") and UniProtKB IDs lack specific isoform names ("-#").	https://uswest.ensembl.org/info/genome/stable_ids/index.html
4	Canonical protein isoform ID	For UniProtKB canonical proteins, the stable ID refers to both the canonical sequence and all known protein isoforms of a given gene. Canonical protein isoform IDs display the specific isoform name of the canonical	https://www.uniprot.org/help/canonical_and_isoforms

		protein with a "-#" extension.	
5	Versioned ID	Ensembl IDs with a '.#' extension, with increments to protein IDs indicating that the associated sequence has changed.	https://uswest.ensembl.org/info/genome/stable_ids/index.html
6	Mapping methods	1. ID mapping, translating IDs between different databases.	(Meyer *et al*, 2016; Huang *et al*, 2008)
		2. Residue-residue mapping, a one-to-one correspondence between amino acids in proteins from different databases.	(David & Yip, 2008; Martin, 2005; Dana *et al*, 2019)
		3. Residue-codon mapping, a one-to-three correspondence between an amino acid and nucleotide coordinates (codon) in a reference genome	(Li *et al*, 2016; Zhou *et al*, 2015; Stephenson *et al*, 2019)

Chapter 3. Prioritizing disease-associated missense variants with chemoproteomic-detected amino acids

3.1 Introduction

Efforts to understand the genetic basis of monogenic disorders and associated molecular mechanisms underlying disease pathology are ongoing and warrant further research. Rare diseases affect fewer than 1 person every 2,000 by US standards, and although non-genetic origins of rare disease are known, an overwhelming fraction of rare diseases are known as Mendelian disorders, follow monogenic patterns of inheritance, and originate from consequences of essential gene disruption by genetic variation (Vaisitti et al 2021; Wright et al 2018).

About 38% of Mendelian disorders are caused by missense variation, or single nucleotide changes in DNA codon that cause single amino acid substitutions in protein polypeptide chains (Landrum et al 2015). These single amino acid changes to alternative amino acids can affect protein structure and function, with effect sizes largely depending on factors like the physical and chemical differences between the substituted residues and context of mutant residue's structural location. The other protein-level effects from single nucleotide variants besides missense are called nonsense and silent mutations. Nonsense mutations cause premature truncation of the protein and are typically associated with protein loss-of-function (LoF) outcomes, whereas silent, or synonymous, mutations are mostly considered to have neutral protein consequences given they do not change the sequence of amino acids in the polypeptide chain.

Compared to nonsense and silent variants, determining whether missense variants have positive, deleterious, or neutral consequences is uniquely challenging; most missense variants of the millions reported in clinically relevant genetic databases currently are classified as VUS for variant of uncertain significance. Significant progress has been made in connecting human disease phenotypes to genotypes (Lee et al 2019; Yang et al 2013), but the missense VUS classification burden continues to limit individuals with rare disorders from receiving definitive diagnoses, prognoses, and therapies treating the underlying causes of genetic disorders.

Experimental approaches such as mutagenesis methods have proven successful in delineating deleterious missense variants from less functional mutations in cellular contexts, but these methods are currently costly to scale and not widely applied yet to poorly characterized and understudied genes and proteins (Cooper et al 2011). In contrast to mutagenesis approaches, computational methods can be used for *in silico* mutagenesis to predict the functional consequence of missense variation, but many of these tools are overly dependent on degree of sequence conservation (Sun et al 2019) and have lower performance in predicting gain-of-function or more moderate effect missense mutations (Reeb et al 2020). More recent computational models that incorporate information about the non-random distribution of genetic variation across gene exons and protein functional regions and domains (Melamed et al 2022; Monroe et al 2022) show performance gains over methods that do not consider local mutation distribution information (Waring et al 2020; Hicks et al 2019; Perez-Palma et al 2020; Quinodoz et al 2022). Such computational models can help in identifying pathogenic mutational hotspots of genes and corresponding protein products.

Besides local mutational distributions, a largely overlooked feature associated with protein functional hotspots is amino acid side-chain reactivity. Nucleophilic amino acids can react with many biological entities and can play important roles in protein stability, interactions, and regulation in cellular pathways. Amino acid side chain reactivity can also fluctuate depending on the residue's protein microenvironment and protein's subcellular environment. Detected by chemoproteomic methods, reactive amino acid residues have exposed important drug vulnerabilities in cancer (Bar-Peled et al 2017) and been the targets of blockbuster FDA-approved drugs (Blewett et al 2016). In this study, chemoproteomic-detected amino acids (CpDAAs) are used to prioritize rare missense variants with associated pathogenic potential. We hypothesized some CpDAA residues to play pivotal roles in disease-associated proteins, and that these sites could be leveraged to prioritize likely functional rare missense variants.

We analyzed chemoproteomic-detected cysteine, lysine, and tyrosine residues in combination with the missense variants of monogenic disorder genes and demonstrated the utility of our approach from identification of rare missense variants likely implicated in HLRCC. Our results highlight one third of monogenic disorder associated proteins had cysteine, lysine, and/or tyrosine reactive sites previously reported by results from four chemoproteomic profiling studies. From the missense analysis focused on monogenic disorder genes, gain of proline and loss of lysine were strongly associated with pathogenic variation relative to population variation. Cysteine codons were also significantly depleted in monogenic disorder genes relative to all other protein-coding genes, adding to the attractiveness of these proteins for selective small molecule targeting.

3.2 Results

Cysteine lysine and tyrosine profiling studies

To study chemoproteomic annotations in monogenic disorder, publicly available datasets of residue profiling experiments specific to human cell lines were curated. The results of six profiling experiments from five independent studies (Weerapana et al 2010; Backus et al 2016; Palafox et al 2021, Hacker et al 2017, Hahm et al 2020), were used as sources of cysteine, lysine, and tyrosine positions ligandability and/or reactivity annotations (**Table 3-1**). The 18,827 detected residue positions contain 4,535 total CpD proteins, with an average of 1.63 cysteine, 2 lysine, 0.52 tyrosine sites per protein (**Table 3-2**). All CpD data sets were combined into a non-redundant inventory to use in the remainder of the analysis. Significant overlaps between the detected protein groups for profiled cysteine, lysine, and tyrosine residues are shown in **Figure 3-1**. Of note, proteins belonging to all three detected residue groups (CpD-CKY proteins) represented 14% of the total CpD proteins.

Chemoproteomic-detected proteins are enriched for genes causing monogenic disorders

Probes from activity-based protein profiling-based (ABPP) chemoproteomic strategies allow for pharmacological interrogation of previously difficult targets (Backus et al 2016) and facilitate drug discovery efforts when coupled with irreversible small-molecule modulators of protein targets (Roberts et al 2017). However, the advantages of coupling

data from ABPP platforms with rare variant prioritization efforts are unknown and under explored.

To explore this application of chemoproteomic detected proteins and residues, I began by describing the genes set of CpD proteins in terms of rare disease and current drug targets. FDA approved drug targets were sourced from the Human Proteome Atlas (Uhlén et al 2015; Wishart DS et al 2006) on May 14, 2021. Monogenic disorders and their associated genes were sourced from the Online Mendelian Inheritance in Man, or OMIM (McKusick et al 2007), June 23, 2021. The 8,322 disease phenotypes from OMIM were filtered for known molecular basis, single gene disorder association, and identifier compatibility with our universal cross-reference file. Our final OMIM dataset contained 5,622 unique diseases and 3,990 unique genes (**Table 3-3**). For brevity, 'OMIM' is used to refer to Mendelian disease phenotypes, genes, or proteins.

Gene set overlap analysis showed FDA drug targets cover ~10% of the OMIM gene set and CpD proteins cover ~31% of the OMIM gene set (**Figure 3-2a**). Minimal overlap between CpD proteins and FDA approved drugs indicate the potential for advances as ~94% of OMIM-CpD genes (*n* = 1168) are not FDA approved drug targets and ~79% of OMIM-FDA genes (*n* = 275) are not CpD proteins. To determine the significance of CpD proteins association to OMIM genes, a Fisher's exact test was used with homozygous LoF tolerant genes (Lek et al 2016) serving as the negative control for disease association, and FDA approved drug targets serving as a positive control for disease association. All CpD protein groups based on different detected residue types were significantly associated with Mendelian disease genes relative to all other human genes (**Figure 3-3**).

CpD proteins are less bias than FDA targets for OMIM phenotypes

Genetic disease phenotype-genotype connections are discovered at a rate of ~260 per year and this shows no sign of slowing (Posey et al 2019; Wenger et al 2016). This means there are many diseases and corresponding pathogenic variants that remain undescribed. It is also not uncommon for a single gene to have multiple connections to human disease. For example, *CFTR* gene is primarily associated with cystic fibrosis, but has connections to other diseases such as chronic pancreatitis and congenital bilateral absence of the vas deferens (CBAVD). Different mutations in the same gene can cause similar or different disease phenotypes.

The FDA approved drug targets are expected to associate with genes linked to multiple disease phenotypes (King et al 2019, Minikel et al 2020, and Nelson et al 2015). To establish the connectedness of CpD proteins to OMIM disease phenotypes with respect to FDA targets, the unique phenotype counts per gene were used for group comparisons. The number of disease phenotypes per gene were described by four levels: '1' for single phenotype, '2' for two phenotypes, '3' for three phenotypes, and '4+' for four or more phenotypes. FDA approved drug target genes had significantly higher mean number of disease phenotypes per genes than genes of CpD proteins (mean of 1.69 vs 1.31 respectively, Wilcox test *p*.adj = 3.7e-11) and were associated with multiple OMIM phenotypes. In contrast, genes of CpD proteins represented both ends of the OMIM phenotype count spectrum, with most only associated to a single OMIM phenotype (**Figure 3-2b**) and potentially more phenotypes not described yet for human disease traits.

CpD protein annotations as orthogonal support for missense constraint

Towards our goal of using chemoproteomics datasets to prioritize rare missense variants with greater pathogenic potential, gene intolerance to protein altering mutations based on gnomAD metrics (Karczewski et al 2020) were next assessed for CpD proteins. Specifically, a gene's intolerance to LoF, missense, and silent variation was described by the observed / expected upper bound fraction of the observed / expected confidence interval metrics (MOEUF, LOEUF, and SOEUF for missense, LoF, and silent constraint, respectively). Lower scores indicate stronger gene intolerance to a specific type of variation while higher values indicate greater tolerance to variation. For example, a MOEUF score of 0.35 indicates that ~35% of missense variation expected by chance was observed in a large population, indicating that the gene is likely under strong selective pressure against this variant type.

To establish the significance of CpD gene intolerance to variation their enrichment in constrained genes was compared to all OMIM genes. Both a strict constraint cutoff of 0.35 recommended by the gnomAD authors and a less strict constraint cutoff based on the bottom decile of scores was used in our analysis. Our results support that CpD proteins are more intolerant to missense variation and LoF variation compared to OMIM genes. To further understand missense intolerance associated with the CpD proteins, specific subsets of CpD proteins based on their detected amino acid type annotations were analyzed along with homozygous LoF tolerant genes (*n* = 285; Lek et al 2016) serving as a negative control. The results showed CpD-CKY proteins had the highest odds for missense gene intolerance (OR =

6.794, p = 1.49e-89) with respect to all other human genes and tested gene sets (**Figure 3-2d**). Notably, proteins detected by cysteine, lysine, and tyrosine reactive probes showed significant association with intolerance to missense variation.

Residue intrinsic reactivity is an important feature for interactivity of proteins and chemical moieties, and the high constraint associated with CpD proteins were suspected to be confounded by protein-protein interaction partners, which is a protein level feature previously described to associate with gene essentiality (Khurana et al 2013; Pei et al 2020). The comparison of CpD proteins to all other proteins interaction partner counts showed CpD were significantly more connected in biological networks (**Figure 3-4**) compared to the average non CpD protein. Closer investigation of the specific CpD groups showed CpD-CKY proteins as more connected in networks compared to other CpD groups and all human proteins (**Figure 3-5**). These CpD subsets tracked perfectly with the odds ratio and MOEUF results described by **Figure 3-2c**.

In contrast to the total of 2,797 genes considered constrained for LoF variation by LOEUF cutoff <0.35, only 114 genes (<1%) of all human genes scored by these metrics are constrained for missense variation. Surprisingly, the CpD proteins included in this study from cysteine, lysine, and tyrosine profiling represented 83% of these genes (n = 95) (**Figure 3-6**). Only 41% (39 out of 95) of these missense constrained genes had associated monogenic disorder phenotypes (**Figure 3-7**), and many may represent important opportunities to connect phenotype to genotype through exploration of genes prioritized by chemoproteomic based annotations (**Table 3-4**).

OMIM genes and hypermutable arginine codons

Missense variant trends are largely influenced by the composition of codons and amino acids in each gene and protein product (Vitkup et al 2003; Gao et al 2015; Khan et al 2007; David et al 2015). The composition of OMIM gene protein sequences were investigated in order to apply insights gained towards: (1) the identification of novel genes underlying Mendelian disorders based on resemblance to OMIM genes and protein products, and (2) the stratification of missense VUS in OMIM genes based on potential disruption of functional CpDAA sites. To begin, codon compositions of OMIM genes were compared to all other human genes.

Twenty proteinogenic amino acids are coded by 61 nucleotide triplets, or codons, making the genetic code degenerate (Crick et al 1961). Apart from methionine and tryptophan, each amino acid corresponds to either 2, 3, 4, or 6 synonymous codons. Biased utilization of synonymous codons is evident across species and even across subsets of genes in a single genome (Grantham et al 1980; Nakamura et al 2000). Significant differences in the average codon abundances for all OMIM vs non-OMIM genes were established using a two-sided, two-sample Welch's t test and 1000 permutations without replacement test. The Relative Codon Synonymous Usage (RCSU) (Sharp et al 1987) metric was also used to quantify usage biases between gene sets of interest.

Shown in **Figure 3.8**, glycine GGC codon ('G.GGC' label in plot) had the largest positive difference between OMIM genes (n = 3744) relative to all other genes (n = 13543), with more frequent usage in OMIM genes (p = 1.61e-08). Glycine codon GGT was also significantly more abundant in OMIM genes (p = 1.39e-08) while the other

synonymous glycine codons had positive differences but were not significantly different between OMIM and all other genes from Welch's t test and across 1000 permutations for which a randomly sampled set of non-OMIM genes more comparable in size to the OMIM gene set was tested (n = 4000; p < 5.0e-05 in 1000/1000 instances; **Table 3-5**). Additionally, the synonymous codons for the following amino acids were more frequently observed in OMIM gene compositions: aspartate (D.GAC; D.GAT), valine (V.GTC), isoleucine (I.ATT), tyrosine (Y.TAC), asparagine (N.AAC), and arginine (R.CGT). Notably, this arginine codon has a CpG dinucleotide at position one, and methylation of the cytosine make these sites 10-50 times more mutable compared to all other possible dinucleotides in the genome because of methylation-deamination (Walser JC et al 2010; Kong A et al 2012). The other CpG arginine codons, referred to as CGN codons, were also more frequently observed in OMIM genes relative to all other genes but the differences were not significant based on p < 5.0e-05. The largest negative difference between OMIM genes relative to all other genes was found for TGT and TGC, which are the synonymous codons for cysteine (p = 3.04e-38 and p = 4.29e-18, respectively; **S1.7**). Arginine's non-CGN codons, AGA and AGG, were also significantly less abundant on average in OMIM sequences, as well as synonymous codons for lysine (K.AAA), serine (S.TCT; S.TCA; S.TCC; S.AGT), and glutamine (Q.CAA) compared to all other genes.

The codon usage bias analysis based on the RSCU metric showed a preference amongst all human genes for arginine AGA codon usage, which differed from the subset of OMIM genes, which preferred usage of arginine CGG codon (**Table 3-6**). Our findings agree with earlier studies showing human genes are depleted of CpG

dinucleotides compared to non-coding regions, but the depletion signal is weaker for human genes involved in essential developmental processes and transcription factor genes (Branciamore et al 2010; Antonarakis et al 2005; Cooper et al 1988). More recently, Schulze et al 2020 reports genes more frequently using arginine CGN codons as more likely to underlie single gene disorders, particularly for dominant phenotypes. The authors also show these codons are hotspots for pathogenic mutations, with most tending to involve C>T transitions. The C>T transitions at CGN codon sites can cause non-synonymous gains of cysteine, glycine, serine, or tryptophan residues which are discussed in detail in a later section.

Higher glycine and lower cysteine composition correlate with monogenic disorders

The average composition of proteins, or frequency of amino acid occurrence, is well-conserved amongst related species (**Figure 3-9**), and positively correlates with synonymous codon number. Composition differences between proteins of a single organism are driven by differences pertaining to function but also selection pressures to minimize the deleterious impact of missense mutations. To understand the context of OMIM proteins and pressures of these essential genes to minimize deleterious structural impacts of mutations, the composition of OMIM proteins were described by their normalized abundances for 20 amino acids. The mean of all OMIM proteins composition of residues was compared to the average of all non-OMIM proteins. Significant differences between the mean abundance of amino acids between the two

groups was tested using the same resampling approach used for codon abundance group differences.

Glycine (p = 1.25e-09) and cysteine (p = 1.22e-36) residues showed the greatest positive and negative composition differences, respectively, for OMIM proteins relative to all other proteins (**Figure 3-10**; **Table 3-7**). This finding was also reflected at the codon-level from the previous section. Interestingly, arginine residues were less frequent in OMIM proteins with a modest p value of 0.00052066 relative to all other proteins.

Ancient proteins are suggested to have been rich in amino acids first incorporated into the genetic code and poor in amino acids lastly incorporated into the genetic code (Trifonov et al 2004). Given the highly conserved nature of OMIM genes and importance of their functional roles in cells, frequency differences between OMIM and all other proteins were expected. To test, the amino acids consensus order of recruitment into genetic code values (Trifonov et al 2004) were summed for the top and bottom four residues that had composition differences between OMIM and all other proteins (**Figure 3-10**). The summed order of recruitment for residues with higher frequency of occurrence in OMIM proteins equaled 10 (summed consensus orders: 1st Glycine; 3rd Aspartate; 4th Valine; 2nd Alanine). The summed order of recruitment for residues with lower frequency of occurrence in OMIM proteins equaled 47 (summed consensus orders: 16th Cysteine; 6th Serine; 15th Lysine; 10th Arginine). This difference of 37 supports the idea of OMIM proteins and associated genes having longer evolutionary histories compared to other human genes, making them richer in

firstly incorporated amino acids like glycine and poorer in lastly incorporated amino acids like cysteine.

Further, our analysis of OMIM amino acid compositions suggested that glycine and cysteine frequencies might also be associated with greater likelihood of genetic constraint and could potentially support identification of genes where variants in heterozygous individuals may impact phenotype. To this end, gnomAD constraint metrics were tested with residue abundances. There was a modest but significant negative correlation between glycine frequency and observed-over-expected (o/e) ratios of LoF (r = -0.035, p = 6.5e-06, Pearson correlation) and missense (r = -0.029, p = 0.00018, Pearson correlation) constraint scores (**Figure 3-11**). The opposite trend was true for cysteine frequency, which had a stronger positive correlation with o/e ratios of LoF (r = 0.13, p = 6.8859e-63, Pearson correlation) and missense (r = 0.12, p = 1.5263e-52, Pearson correlation) constraint scores (**Figure 3-12**).

Mutability of amino acids gained and lost in missense substitutions

With sequence composition insights in hand, the spectrum of missense mutations in OMIM genes we next described. From gnomAD (Karczewski et al 2020) and ClinVar (Landrum et al 2016), a high confidence mutually exclusive set of pathogenic, likely neutral, and VUS missense we created. The missense from gnomAD were assigned to one of three categories based on allele frequency: common, rare, and rarest. Either all ClinVar benign and gnomAD population variants, referred to as Background in figures, were used for comparisons with pathogenic missense or only common and benign

considering set considering that some rare gnomAD alleles may be pathogenic in a homozygous state.

Our analysis of variants in OMIM genes focused first on the frequency of certain types of missense. First, gains and losses of specific amino acids by population and pathogenic alleles were shown, with visual annotations from protein evolution theory first described by Zuckerkandl et al 1971 and later supported by Jordan et al 2005, that suggests the order in which the genetic code was assembled over 3.5 billion years ago continues to influence the evolution of proteins today. Jordan et al 2005 reports cysteine (C), methionine (M), histidine (H), serine (S), and phenylalanine (F) as accruing, while proline (P), alanine (A), glutamate (E), and glycine (G) are declining in frequency based on nucleotide polymorphisms data for human genomes and ortholog sequence substitutions for genomes from 15 taxa representing all three domains of life (Bacteria, Archaea and Eukaryota). The background variant frequency of occurrence plotted in **Figure 3-13** supports the trend of five strong gainers and four strong losers in our curated set of OMIM genes, with strong losers shown to fall below the equilibrium line (slope = 1) and strong gainer residue points shown above this line. This same analysis but using the pathogenic alleles is shown in **Figure 3-14** revealed Methionine, a strong gainer, as falling below the equilibrium line, while proline, a strong loser, was above the equilibrium line and more often gained by pathogenic alleles in OMIM genes

Proportions of pathogenic missense involving loss of cysteine and glycine clearly distinguished the pathogenic variant set, making up ~20% of the pathogenic allele category compared to only ~7-8% of the neutral variant categories (**Figure 3-15a**). Clear distinctions between pathogenic and likely neutral alleles were also observed for

cysteine and proline residue gains, which accounted for ~15% of the pathogenic category and only ~7-8% of the neutral variant categories (**Figure 3-15b**). Although the proportion of missense involving loss or gain or arginine residues was high across the pathogenic and likely neutral variant categories, loss of arginine and loss of glycine together accounted for ~30% of pathogenic alleles. This result is consistent with Vitkup et al 2003, finding arginine and glycine residues together responsible for 30% of genetic diseases.

The relative mutability (Rm) of an amino acid in each category was calculated based on Khan et al 2007 (**Table 3-8**). In short, a residue's gain or loss mutability in a specific variant category is relative to the least mutable residue, defined as the amino acid with the lowest ratio of observed / expected mutations for a given missense category in OMIM genes (denoted by Rm values equal to 1). Arginine had the highest Rm for residue losses by substitution in both the pathogenic and background missense categories. The hypermutable nature of arginine CGN codons is likely related to this finding. Relative mutability results that differentiated pathogenic from background alleles involved losses of cysteine (C), glycine (G), and tryptophan (W) residues, which were about 7x, 6x, and 4x more mutable than lysine for pathogenic missense, respectively (**Figure 3-16**). Notably, tryptophan for the background lost-by-missense variants was the least mutable amino acid (**Figure 3-16**). For amino acid gains by missense substitutions, lysine (K) and proline (P) were about 3.5x and 3.3x more mutable than alanine for pathogenic missense, respectively (**Figure 3-17**), and differentiated the pathogenic category from the background Rm results. Notably, proline for the background gain-by-missense variants was the least mutable residue (**Figure 3-17**).

Enrichment of residues gained and lost by missense alleles in OMIM

To distinguish pathogenic from background missense variation with respect to the unique observed composition traits of OMIM sequences, the fold enrichment (FE) of specific substitution types for pathogenic and common/benign alleles were calculated. The calculations were based on a given residue's frequency of occurrence in OMIM proteins for loss-of-residue mutation types (Visscher et al 2016) and based on mutated codon frequency of occurrence and all possible non-synonymous changes for gain-of-residue mutation types (**Figure 3-18a** and **b**). FE values for missense outcomes were log transformed to clearly distinguish large enrichments from depletions of specific missense types in the different variant categories. Significant enrichments for residue losses and gains were confirmed using a Fisher's exact test and Bonferroni correction for multiple testing, with the magnitude of pathogenic enrichment calculated with respect to the common/benign category for a controlled set of 2,873 OMIM genes impacted by both pathogenic (n = 33,872) and common/benign (n = 29,778) missense categories.

For loss-by-missense substitutions, enrichment of cysteine (p = 4.17e-267), glycine (p = 2.44e-247), tryptophan (p = 1.99e-95), and tyrosine (p = 9.66e-60) distinguished pathogenic alleles from likely neutral variants in OMIM genes (**Figure 3-18a**). For gain-by-missense substitutions, enrichment of proline (p = 1.83e-213), arginine (p = 1.72e-109), tyrosine (p = 1.97e-60), and lysine (p = 8.28e-04) distinguished pathogenic alleles from likely neutral variants in OMIM genes (**Figure 3-18b**). Interestingly, some residues had opposite fold enrichments for lost-by-missense and gained-by-missense substitutions. For example, loss-of-cysteine was depleted in

the common/benign category but gain-of-cysteine was enriched in the common/benign category. For cases of opposite trends involving pathogenic alleles, loss-of-lysine and loss-of-proline were depleted in this category, but gains of both these residues were enriched in OMIM genes. The reverse was true for alleles involving glycine, with losses depleted and gains enriched in the pathogenic category of missense variants.

Spectrum of missense substitutions in OMIM genes

For a higher-resolution view of missense trends in monogenic disorder genes, the specific types of substitutions frequencies of occurrence in the pathogenic and common/benign categories were compared using the controlled subset of OMIM genes impacted by variants from both missense categories. Differences in frequency of occurrence for pathogenic vs common/benign substitutions replicated findings from previous studies (Khan et al 2007; Vitkup et al 2003), with conservative exchanges more frequently observed in the common/benign alleles and less conservative exchanges more frequently observed in the pathogenic alleles (**Figure 3-18c**). For example, valine to isoleucine (V>I), and the reciprocal exchange of these two residues, had higher abundance in common/benign alleles and are considered conservative substitutions of two hydrophobic residues. In contrast to these population variations, in the case of pathogenic disease associated alleles, the two most abundant changes were glycine to arginine (G>R) and leucine to proline (L>P) substitutions.

A well-known observation of the genetic code is that similar amino acids are more likely to substitute each other, thereby minimizing the structural impact of mutation and mis-readings (Freeland et al 1998). Evidence of this observation is clear from the

arrangement of the genetic code table, where mutations at the first and third codon positions are generally tolerated by proteins better compared to mutations at the second position, which result in exchanges of more dissimilar residues in terms of their chemical properties. non-synonymous changes to four synonymous leucine codons can lead to proline gain, and all involve mutating the second codon position. In addition to simple evidence of deleteriousness presented by the genetic code, at the protein-level, leucine is commonly found in protein α-helices and in this context, proline can cause major disturbances to protein stability by disrupting the hydrogen bridge system of the α-helices and exposing a hydrogen bridge acceptor. The second most common pathogenic route for proline gain appears to stem from loss of arginine, which also involves nucleotide substitutions at the second position of codons (**Figure 3-18c**).

Unlike the other 19 amino acid types that have clear association with one or multiple groups of residues based on shared chemical properties, glycine is a “borderline” member of the group of amino acids with uncharged polar groups (Lehninger et al 1975). The uniqueness of this residue is also evident considering that it is the smallest amino acid, taking up 3Å^3, and the next smallest amino acid (alanine) is 10 times its size. If all non-synonymous mutations of glycine are considered with equal probability, the average substitution of this residue will result in gain of an amino acid about 26 times larger (Graur et al 1985). The abundance of G>R mutations in pathogenic alleles makes sense considering these simple features of glycine, and from **Figure 3-18c**, it is evident that only G>A exchanges were nearly equal in abundance between pathogenic and common/benign alleles, with all other glycine exchanges skewed towards bias abundance in the pathogenic category. I also provide a higher resolution view of the

missense exchanges and their normalized abundance differences between pathogenic and common/benign variants in a 61x61 codon-level heatmap (**Figure 3-19**).

The magnitude of enrichments for specific amino acid exchanges in the pathogenic category of variants relative to the background category of variants was lastly calculated using a Fisher's exact test for all possible missense derivatives for 20 amino acids. I highlight findings specific to cysteine loss and gain exchanges in **Figure 3-18d**. Cysteine losses (**Figure 3-18d;** left panel), no matter which residue type was gained by the exchange, had strong enrichment in pathogenic alleles relative to background alleles, with the highest enrichment observed for the cysteine to phenylalanine (C>F) and TGC>TTC codon exchanges (OR = 4.875, p = 4.57e-60). In contrast to cysteine loss, exchanges involving cysteine gains (**Figure 3-18d**; right panel) were not all enriched in the pathogenic category, with two out of the four possible serine to cysteine (S>C) exchanges shown as enriched for background alleles (depleted in pathogenic alleles) (TCC>TGC OR = 0.52, p = 1.39e-06; TCT>TGT OR = 0.33, p = 1.53e-14). The strongest enrichment for cysteine gain substitutions were observed for the glycine to cysteine (G>C) and GGT>TGT codon exchanges (OR = 2.95, p = 2.66e-26) and the TGG>TGT codon W>C residue exchange (OR = 2.90, p = 5.31e-20). The lysine and tyrosine specific results are provided along with cysteine for a complete picture of chemoproteomic-detected amino acids and their pathogenic substitution propensities (**Figure 3-20** and **Figure 3-21**). With the insights gained from the analysis of OMIM sequence compositions and mutational spectrum results, the analysis advanced to defining the mutational and contextual landscapes of our CpDAA residue sites in the curated set of OMIM genes and proteins.

1D relationships of CKY positions to missense in OMIM proteins

To define the context of detected amino acids in terms of proximity to disease missense and population missense positions in OMIM proteins, the CKY residue landscape in 1D sequence space were analyzed (**Figure 3-22a**). To evaluate the significance of CpD-CKY positions, they were compared to undetected CKY positions in regards to missense allele positions in OMIM proteins (**Figure 3-22b**). CpD cysteine positions accounted for ~14% of all cysteine positions, which was significantly higher than the percent of positions that CpD lysine or CpD tyrosine positions accounted for out of all equivalent residues in OMIM&CpD proteins (**Figure 3-22b**). Cysteine residues higher percentage than CpD lysine and CpD tyrosine positions is most likely related to global rareness of cysteine in diverse species of life (**Figure 3-23; Figure 3-24**) as well as the results described from previous sections such as lower abundance in OMIM proteins compared to all other human proteins.

Preference of pathogenic missense effecting detected over undetected CKY sites

Distances of zero, or direct overlaps of CKY positions and positions of missense alleles, were first summarized. About 1.9% of CpDAA positions overlapped pathogenic missense alleles compared to 0.9% percent of undetected CKY (**Figure 3-25a**). The percentage of positions overlapping positions of missense alleles from the benign category or other non-pathogenic categories of missense differed by less than one percentage point for the detected and undetected CKY positions (**Figure 3-25b**). Detected positions were significantly enriched over undetected positions for pathogenic

missense allele overlaps (OR = 1.54; *p* = 1.75e-04) and significantly depleted over undetected positions for background missense allele overlaps (OR = 0.89, *p* = 1.56e-04) in analysis of 3,907 OMIM proteins (**Figure 3-22c**). CpD-lysine positions relative to undetected lysine positions showed the greatest odds for pathogenic missense overlap (OR = 2.56, *p* = 8.08e-05), followed by CpD-tyrosine positions (OR = 2.56, *p* = 8.08e-05) relative to undetected tyrosine positions (**Figure 3-26**). CpD-cysteine positions showed positive but unsignificant associations with pathogenic allele overlap (OR = 1.341, *p* = 0.0754) relative to all other cysteine codons (**Figure 3-26**). Decreasing odds of missense variant overlap at detected versus undetected positions associated with increasing residue relative mutability in the pathogenic allele category (**Figure 3-26**).

Residue-level annotations of CpD lysine show greater potential for identifying positions in proteins more likely to overlap pathogenic missense alleles compared to positions of the other residue types studied and highlights a potential strength of this annotation type for variant prioritization given that lysine is the least mutable amino acid in the pathogenic allele category from our relative mutability results discussed in the previous section.

Detected residues are closer to pathogenic missense than population missense in 1D sequence space

To determine if detected residues are closer to pathogenic versus common/benign missense alleles in OMIM proteins, the distributions of 1D distances for CpDAA residue types to the nearest positions of common/benign and pathogenic missense disease were visualized (**Figure 3-22c**). Only OMIM&CpD proteins with pathogenic and

common/benign allele positions were used to minimize bias stemming from protein length differences. The pathogenic allele to detected residue distances were smaller than distances between common/benign alleles and detected residue positions, with pathogenic-to-CpDAA distances shown as more skewed towards zero (**Figure 3-22c**).

Because the allele-to-CpDAA distances are also confounded by differences between protein lengths and ratio of pathogenic to common/benign missense (**Figure 3-27**), additional tests of missense 1D distances to detected residues were calculated for three sets of OMIM&CpD proteins: all sequences, short sequences, and long sequences. All missense allele distances (based on unique DNA change) to detected residue positions were included in this analysis and short and long proteins were defined by lengths less than or greater than the median protein length calculated for all OMIM&CpD sequences, respectively (**Figure 3-28**). Since the 1D distances are not normally distributed, the median was used for bootstrapped 95% confidence interval estimates in comparisons between the missense categories and detected residue 1D distances. Pathogenic variants were closer to detected residues for all OMIM&CpD proteins, short OMIM&CpD proteins, and long OMIM&CpD proteins tested based on the median 95% confidence intervals describing pathogenic-to-CpDAA distances no overlapping estimates for the other four missense categories tested and all protein subsets for short and long sequences tested (**Figure 3-29**). From our analysis of 1D distances between positions of ClinVar and gnomAD variants and detected residues, it was concluded the detected cysteine, lysine, and tyrosine positions are closer to positions of pathogenic alleles in 1D sequence space.

1D distances comparing detected versus undetected CKY positions show pathogenic missense closer to detected positions

Are missense variants closer to detected residue positions relative to undetected equivalent residue positions? To answer this question, we considered nearest 1D distances for pathogenic and common/benign alleles to all CKY residue positions in 926 OMIM&CpD controlled for having at least one pathogenic and one common/benign missense position. Since Figure 3-22c showed an enrichment of pathogenic alleles at detected residue codons relative to undetected residue codons, 1D distances were also filtered to exclude distances of zero to prevent overlaps of CKY residues and missense variants from biasing comparisons between detected and undetected residues. The Wilcoxon test was used to compare the mean 1D distances between pathogenic and common/benign missense positions to detected versus undetected residue positions. Detected residues were closer to pathogenic missense positions compared to undetected equivalent residue positions (p < 2e-16; **Figure 3-30**) in OMIM&CpD proteins based on our analysis. In addition, distances of common/benign allele positions were significantly smaller for undetected residues compared to detected residue positions (p = 3.6e-10; **Figure 3-30**).

For further validation of detected versus undetected 1D distance findings, the analysis was repeated separately for the three residue types that are the focus of our study: cysteine, lysine, and tyrosine. The residue type specific results replicated initial findings related to pathogenic alleles and detected residue positions, with pathogenic missense positions found to be closer on average to detected residue positions than to undetected equivalent residue positions in OMIM&CpD proteins (p < 2e-16; **Figure 3-**

31). For CKY 1D distances to common/benign alleles, undetected lysine positions were significantly closer (*p* = 2.8e-11; **Figure 3-31**) than detected lysine positions, but distances for cysteine and tyrosine positions were not significantly different between detected versus undetected residues (**Figure 3-31**). Based on these results, it was concluded that detected positions are closer to pathogenic missense alleles and farther from common/benign missense alleles than undetected positions in OMIM&CpD proteins, supporting CpDAA annotations potential in missense variant prioritization pipelines.

Burden of variants in 1D windows of CpDAA positions

Missense variant occurrences at positions flanking CpDAA residues were next calculated to provide more contextual information about CpDAA positions nearby missense alleles for OMIM proteins. The 1D window sizes used in our analysis included ±3 amino acids, ±6 amino acids, or ±15 amino acids flanking reference CKY residue positions. Missense variants overlapping reference CKY positions were excluded from total missense occurrence calculations for 1D windows and only sequences with pathogenic and common/benign missense variants were used for 1D window analysis. About 15% - 20% of CpDAA 1D windows included positions of pathogenic missense alleles compared to only ~8% - 9% of 1D windows for common/benign missense positions (**Figure 3-32**). Surprisingly, local missense VUS was true for nearly half of all CpDAA window cases for size of ±6 amino acids (**Figure 3-32**).

Fisher's exact test was used to estimate the likelihood of pathogenic and common/benign alleles occurrence in 1D windows of detected versus undetected

cysteine, lysine, and tyrosine reference positions. Occurrence of pathogenic missense variants was significantly enriched for ±6 amino acid 1D windows of CpDAA positions (OR = 1.95, *p* = 1.87e-49) relative to 1D windows of undetected CKY positions in 926 OMIM&CpD proteins (**Figure 3-33**). In contrast, occurrence of common/benign missense variants was significantly enriched for ±6 amino acid 1D windows of undetected CKY positions (OR = 0.74, *p* = 1.01e-08) relative to 1D windows of detected CKY residues (**Figure 3-33**). All results were replicated in the alternative window sizes used for 1D window analyses (**Figure 3-34**). Significant occurrence of pathogenic missense in detected residue 1D also replicated in separate analysis of specific residue types (**Figure 3-35**). Pathogenic missense occurrence showed the highest odds with detected tyrosine 1D windows (OR = 2.221, *p* = 6.49e-12), followed by detected lysine 1D windows (OR = 2.066, *p* = 3.40e-26), and detected cysteine 1D windows (OR = 1.663, *p* = 3.46e-12) relative to undetected equivalent residue 1D windows in 926 OMIM&CpD proteins (**Figure 3-22e**). Significant depletion of common/benign missense occurrence was also found for detected lysine 1D windows relative to all other lysine residue 1D windows in OMIM&CpD proteins (**Figure 3-22**e). Our results support chemoproteomic annotation's association to regions of OMIM proteins impacted by deleterious missense alleles.

3D distances between missense variants and cysteine, lysine, and tyrosine CpDAA residues

To build upon characterization of detected CKY residues based on 1D distances to missense alleles, 3D protein environments of detected CKY residues were investigated

for the presence of missense variant alleles. Our PDB mapping pipeline provided distances from the terminal atoms of cysteine, lysine, and tyrosine residues (SG, NZ, and OH atoms, respectively) to all other atoms of neighboring residues within a max environment boundary set for 10 $Å^3$ from the reference terminal atom coordinate. To calculate 3D environment missense burden, environment boundaries of $6Å^3$, $8Å^3$, and $10Å^3$ and all unique missense alleles mapping to environment based on terminal atom coordinates for cysteine, lysine, and tyrosine residues were considered (**Figure 3-36a**).

Out of 926 OMIM&CpD proteins used for 1D distance analysis, resolved structures were available for 419 proteins in the Proteins Data Bank, accounting for ~ 45% of total CpDAA positions from 1D distance characterization available for 3D environment characterization (**Table 3-9**). The dramatic decrease in available CKY positions for 3D mapping hindered comparison between detected and undetected residues, so the focus of our analysis was on detected residues and their 3D environments.

Prioritizing CpDAAs based on 3D environment VUS count and predicted deleteriousness for substitutions at CpDAA sites

Analysis of missense variants in 3D environments revealed 38.2% of CpD-cysteine were within 8Å of a pathogenic missense variant, followed by 36.2% of CpD-tyrosine, and 23.2% of CpD-lysine, compared to ~10% of CpDAA within 8Å of a common/benign missense variant position (**Figure 3-36b**). Notably, more than half of all CpDAA environments contained missense VUS within 8Å. The greatest association is shown for

CpD-cysteine environments, with 57.7% harboring at least one local missense VUS (**Figure 3-36b**).

To understand unique features of detected cysteine, lysine, and tyrosine, associations of missense variant categories in the environments of specific residue types versus all other CpDAA 3D environments were assessed. For comparison to pathogenic variants, the background set of variants was used instead of the common/benign variant set due to the limited counts of available common/benign missense that might bias interpretation of the results. Both detected cysteine and detected tyrosine environments were enriched for proximal pathogenic missense relative to all other CpDAA types (**Figure 3-36c**). In contrast, detected lysine environments were significantly depleted in proximal pathogenic variants with respect to cysteine and tyrosine CpDAA positions (**Figure 3-36c**). Detected cysteine environments were also uniquely enriched for local background missense and VUS missense alleles (**Figure 3-36d**). CpD-tyrosine environments were also less abundant in local background missense variants compared to CpD-cysteine and CpD-lysine environments (**Figure 3-36c**). Notably, CpD-lysine environments were depleted of pathogenic alleles despite finding significant overlap of pathogenic alleles in the previous section.

Missense pathogenicity predictors are commonly used by genetic variant annotation pipelines to provide orthogonal evidence of a variant's potential deleteriousness or neutral molecular consequences. To support stratification of CpDAA with higher functional potential, the Combined Annotation Dependent Depletion, or CADD, metric was used based on previous work demonstration of its performance

across diverse classes of genes (Anderson et al, 2018; Ghosh et al, 2017). CADD scores also have the added advantage of minimal missense values for missense predictions of all possible non-synonymous changes in the human genome. Deleterious missense predictions for possible CpDAA substitutions were also tested for significant associations to environment features such as local pathogenic variant. A mean phred score for all possible non-synonymous exchanges above 25 defined deleterious CADD scores.

CpDAA environments with local pathogenic variants were significantly associated with deleterious loss-of-CpDAA predictions (**Figure 3-37**). In contrast to local pathogenic, environments with common/benign or background alleles were significantly depleted of CADD scores above 25 threshold (**Figure 3-37**). Separate analysis of each residue type showed local pathogenic allele and CpDAA environments significant for CpD-tyrosine relative to all other equivalent residue types in OMIM&CpD proteins (**Figure 3-38**). Furthermore, CpD-lysine environments with common/benign variant allele positions associated more with CADD missense scores below the deleterious threshold relative to all other lysine positions in OMIM&CpD proteins (**Figure 3-38**).

A major goal of this study was to use information about CpDAAs to prioritize missense VUS as pathogenic. CpDAAs with the highest counts of proximal VUS missense along with mean CADD scores were visualized (**Figure 3-36d**). Sorting the detected positions by highest counts of proximal missense VUS unintentionally prioritized residues located at multimeric protein interfaces because CpDAA residues at interfaces have more interactions with neighboring amino acids on the same chain and

different protein chain(s) of a multimeric protein's quaternary structure compared to residues at the surface of monomeric proteins.

CpDAA environments heavily burdened by missense VUS included lysine at position 1296 of DNA mismatch repair protein MSH6 (*n* = 59 environment VUS; **Figure 3-36d**) and lysine at position 65 of DNA mismatch repair protein MSH2. Notably, MSH6 and MSH2 form a heterodimer complex as part of their role in the post-replicative DNA mismatch repair (MMR) system (PubMed:26300262). CpDAA environments of MSH2 and MSH6 lacked local pathogenic alleles based on our ClinVar dataset, whereas environment of CpDAA residues in MLH1, VHL, and FH proteins had VUS and pathogenic allele containing environments (**Figure 3-36d**; red colored position ID labels). For further investigation, CpDAA residue environments with high missense VUS count and deleterious substitution scores based on mean CADD phred scores above the deleterious threshold of 25 were prioritized (**Figure 3-36d**).

Functional validation of our approach in FH protein

For genes currently associated with a disease phenotype, establishing the causal role of individual variants within the gene remains problematic, and many patients with suspected rare genetic diseases are left without a definitive diagnosis (MacArthur et al 2014). To demonstrate the utility of our approach in stratifying CpDAA residues based on potential genetic disease phenotype connection, we selected an OMIM and CpD protein from a list of ideal candidate proteins for experimental validation of missense molecular mechanism studies. Criteria for ideal candidate protein: gene has mapped missense position from each variant category, available PDB structure for protein

structure analysis, less than 1000 amino acid protein length, ClinVar and gnomAD allele position interactions with CpDAA in 3D environment, and in CKY detected protein group. After applying these criteria to OMIM and CpD proteins, 28 candidate proteins remained for consideration. Fumarate hydratase protein FH was selected from the 28 candidates list because of its important role in metabolism fumarate to L-malate interconversions, lack of FDA-approved drugs targeting protein, and many structurally resolved CpDAA sites (*n* = 16 CpDAA sites; 2 cysteine, 13 lysine, and 1 tyrosine) (**Figure 3-39a**). Mutations in *FH* are linked to hereditary leiomyomatosis and renal cell cancer (HLRCC), and fumarase deficiency OMIM disease phenotypes . Based on our missense variant inventory, *FH* had 410 total non-synonymous single nucleotide variants (52 pathogenic, 3 common/benign, 12 rare, 85 rarest, and 258 VUS), that together impacted 341 unique residue positions of FH protein. All CpDAA positions for the FH protein were clustered based on local missense allele counts from 1D and 3D analyses to identify interesting residue sites for experimental follow-up (**Figure 3-39b**). Clustering based on proximal missense variant counts resulted in two main groups of FH CpDAA sites. The residue group associated with higher local VUS counts included two CpD-cysteine residues (C333 and C434) and one CpD-lysine residue (K311). All three CpDAAs in the top cluster were also associated with deleterious CADD scores for possible residue substitutions by missense variants (**Figure 3-39b**). The cysteine at position 333 in FH protein was within 8Å of four pathogenic missense alleles, six gnomAD alleles in the rarest allele frequency category, and twenty-four missense VUS alleles combined from ClinVar and a validation screen by Wilde et al 2022 (**Figure 3-40**). The prioritized cysteine at position 333 (**Figure 3-39c**) was validated as a loss-of-

function consequence associated with disruption of FH protein multimerization (**Figure 3-39d**). This supports the utility of our approach to stratify functional CpDAA sites based on spatial relations to missense alleles.

3.3 Discussion

Missense are known major causes of human diseases, but knowledge of their impact on protein function and regulation is largely unknown. Scalable and creative sources of information are needed to annotate residues so that estimates of deleteriousness and functional consequences can be assessed. In this study, protein- and residue-level annotations from chemoproteomics studies of cysteine (C), lysine (K), and tyrosine (Y) profiling were investigated based on their functionality and ability to provide orthogonal support for missense variant interpretation. Multi-detected proteins that have CpD-cysteine, lysine, and tyrosine probe labeled positions were shown to be centrally located in protein-protein interaction networks and intolerant of missense variation (**Table 3-10**). This may be a particularly useful insight for assessing the missense intolerance of smaller genes (containing fewer codons) or highly paralogous genes for which constraint metrics based on observed over expected ratios of rare missense variation are less reliable. Lysine detection annotations may be important for rare variant interpretation given that lysine residues overall had the lowest relative mutability in the pathogenic allele category, but detection was significantly associated with pathogenic allele overlap relative to all other lysine positions in monogenic disorder associated proteins. Cysteine detection annotation may serve as important markers of pathogenic allele burden region so genes and proteins based on 1D and 3D spatial analysis of

missense alleles. Incorporation of unsupervised machine learning based annotations such as the CADD score appear to support stratification of CpDAA and complement analysis of proximal missense burden analysis of CKY residue positions.

A caveat to this study is that it pertains to only a subset of human genes implicated in monogenic disorders and therefore conclusions do not apply possibly to larger set of reactive and probe detected sites in human proteins. However, a multilevel understanding of OMIM genes and proteins, the spectrum of disease and population variation in these context, and CpDAA were most relevant to our overall goal of stratifying missense VUS, most of which map to OMIM genes.

Future work should continue investigating the context of reactive amino acids and their relation to human phenotypic variation. I imagine many of these solvent exposed residues prioritized by multi-level approaches will lead to the unraveling of compensatory targeting opportunities of gain-of-function mutants.

3.4 Methods

Data source	URL	Version
UniProtKB	https://www.uniprot.org/downloads	August 2021
dbNSSFP	https://sites.google.com/site/jpopgen/dbNSFP	4.2a
ClinVar	https://www.ncbi.nlm.nih.gov/clinvar/	June 10,2021
gnomAD constraint		2.1.1
OMIM	https://www.omim.org/downloads	June 24, 2021
Human Protein Atlas	http://www.proteinatlas.org	20.1
HGNC	https://www.genenames.org/download/custom/	September 2020

Software	URL	Version
Python	https://www.python.org/	3.7.4
R	https://www.r-project.org/	3.6.2
Tidyverse	https://doi.org/10.21105/joss.01686	1.3.0

Pandas	https://pandas.pydata.org/	0.25.1
Numpy	https://numpy.org/	1.17.2
SciPy	https://www.scipy.org/	1.3.1
Adobe Illustrator	Adobe, Inc	

Curation and standardization of chemoproteomics datasets

To curate and standardize available residue-level results of reactivity-based protein profiling, we processed all result files through the same quality control pipeline. This process included filtering out peptides with multiple amino acids marked as modified (e.g. MAC*ALRC*Y) and removing peptides that do not meet the minimum number of detections in replicate samples. For datasets from reactivity profiling experiments, reactivity ratios ($R_{10:1}$) assigned to each residue were averaged across peptide replicates for final assignment of one $R_{10:1}$ value. Reactivity labels for Low, Medium, and High were re-assigned based on the following bins: Low $R_{10:1}>5$, Medium $2<R_{10:1}<5$, High $R_{10:1}<2$. All residue identities and positions were checked against the reference set of protein sequences from UniProtKB to prevent analysis errors caused by residue position mis-mapping. To compare cysteine, lysine, and tyrosine detected proteins, we combined experimental datasets for cysteine specific reactivity profiling after confirming the reactivity ratios of residues detected in both studies were significantly correlated (Pearson's R=0.49).

Assigning proteins to subcellular location information

We used the COMPARTMENTS (Binder et al 2014) database to assign each protein to their main subcellular location(s) based on the database provided highest location score values. This resource integrates evidence on protein subcellular localization from

manually curated literature, high-throughput screens, automatic text mining, and sequence-based prediction methods. If a protein had multiple subcellular locations with the highest score, more than one main subcellular location was assigned to the protein.

Calculating amino acid and codon mean abundances

The codon composition of 17,287 human genes with Ensembl CDS sequences was calculated by taking the frequency of occurrence of 61 codons and normalizing by the total number of counted codons per gene.

3D distance calculations

Proteins with CpDAAs were cross-referenced with the Protein Data Bank (PDB) downloaded June 23, 2022. All biological assembly files of entries were processed. For each CpD protein associated with a PDB, the SIFTS database (2019 release) was used to map protein sequence residue positions to PDB structure residue positions. The author determined biological unit annotations were extracted from each PDB, as well as the exact 3D coordinates of a CpDAA. Specifically, distances were calculated with respect to locations of the SG atom of cysteine residues, NZ atom of lysine residues, and OH atom of tyrosine residues to all other atoms of neighboring amino acids within 10 Angstroms. The smallest distance between terminal cysteine, lysine, or tyrosine atoms and atoms of neighboring missense variant positions were stored for statistical analyses. Multiple PDB structures assigned to a given uniprot identifier used for missense environment counting. At the amino acid level, all proximal amino acids to a detected residue were assigned a distance pair identifier, composed of the protein

identifier and protein position of a given residue. The distance pairs for a single environment (in the consistent direction from missense protein position to CpDAA protein position) were all unique.

3.5 Figures

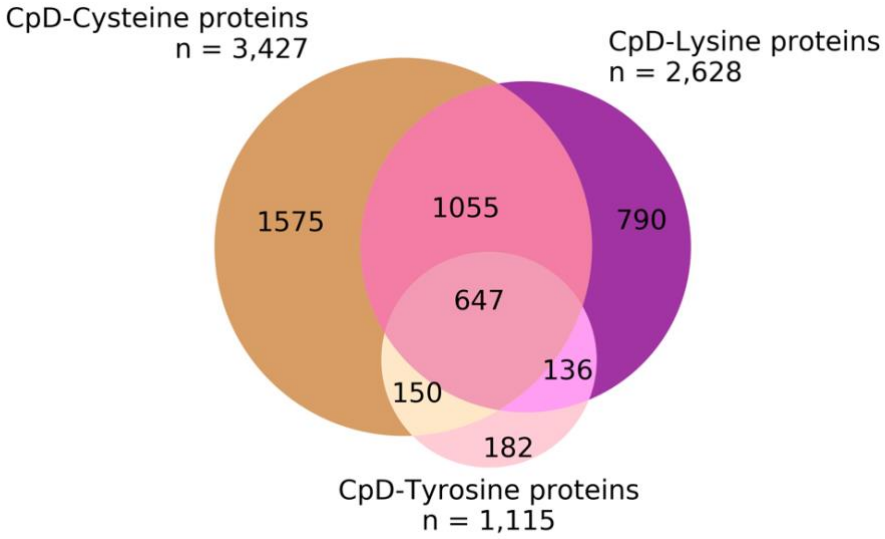

Figure 3-1. Overlaps of chemoproteomic detected protein subsets.

The set of all 4,535 CpD proteins in our study are associated to 4,572 unique gene symbols. Numbers annotating the figure are protein counts based on unique UniProt protein.

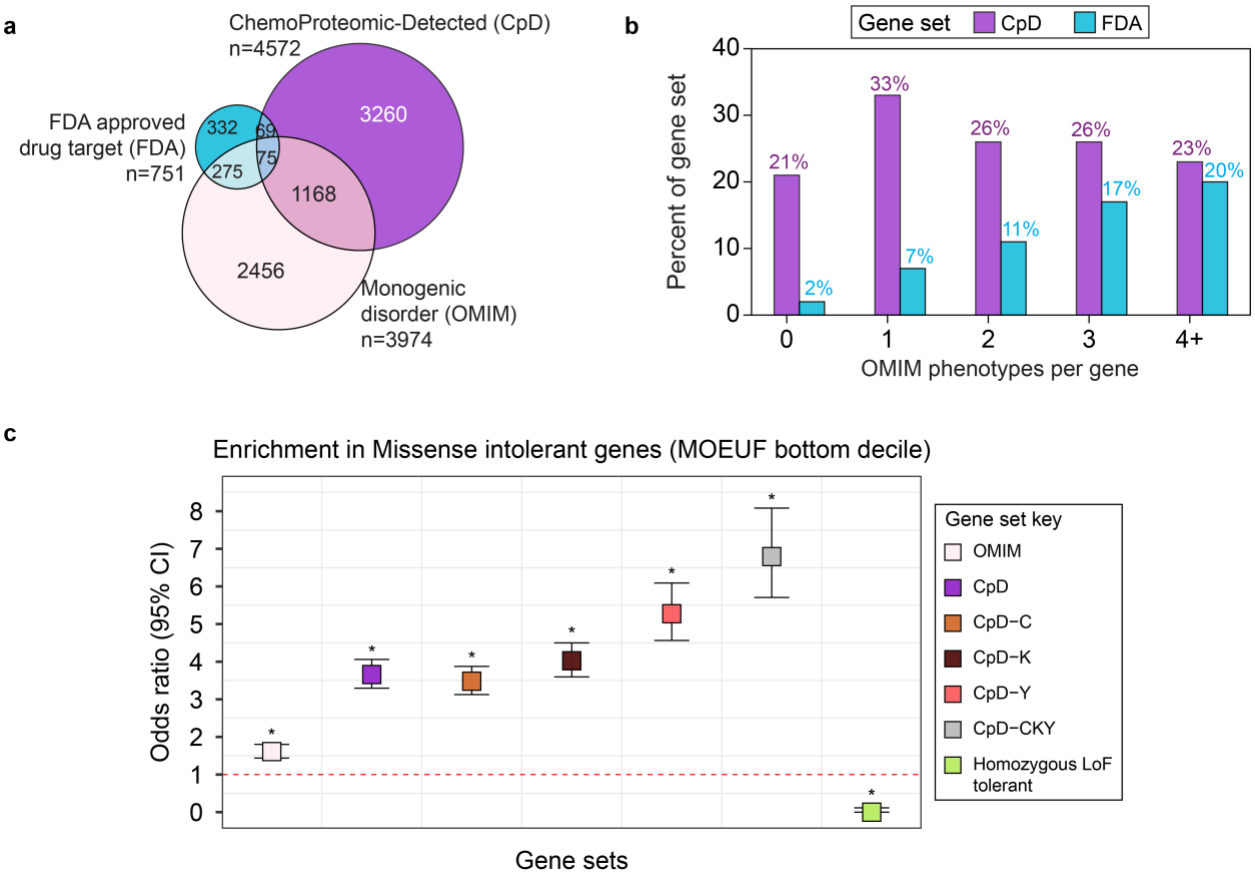

Figure 3-2. CpD-proteins are associated with monogenic disorders and gene-level intolerance to missense variation

a) Venn diagram of the overlaps between the FDA approved drug target gene set, gene set of ChemoProteomic Detected (CpD) proteins, and gene set associated with monogenic disorders (OMIM).

b) Distribution of genes from the CpD protein (purple) and FDA approved drug target (cyan) gene sets across five levels based on the number of OMIM monogenic disorder phenotypes per gene (x axis). Data is based on a reference set of 20,210 total human protein coding genes. Bar height represents the fraction of each gene set associated with a level of phenotypes counted per gene. Levels: “0” connected phenotypes (n=16,236 total genes in universe, CpD=21% , FDA=2%); “1” connected phenotype

(n=2,985 total genes in universe, CpD=33%, FDA=7%); “2” connected phenotypes (n=644 total genes in universe, CpD=26%, FDA=11%); “3” connected phenotypes (n=202 total genes in universe, CpD=26%, FDA=17%); “4+” connected phenotypes (n=143 total genes in universe, CpD=23%, FDA=20%).

c) Enrichment for missense intolerant genes in gene sets related to CpD-proteins, monogenic disorders (OMIM), and homozygous LoF tolerant genes. The threshold for missense (MOEUF) constraint is based on the bottom decile of all 16812 sorted and scored genes included in this analysis. Fisher’s exact test was used to compare gene counts. Point estimates (odds ratio) > 1 indicate a gene set’s enrichment for missense intolerant genes based on the described threshold. Significant associations are marked by *. Bonferroni correction for multiple testing. Horizontal bars show the 95% confidence interval of the odds ratio point estimates.

Abbreviations: FDA, FDA approved drug target genes. CpD, ChemoProteomic-Detected protein genes. OMIM, Monogenic disorder-associated genes.

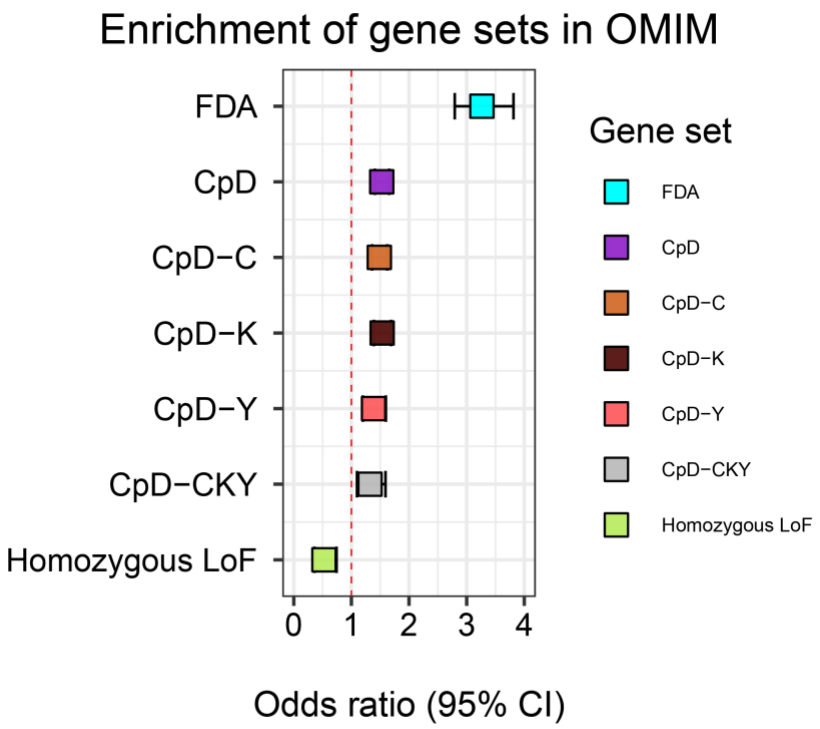

Figure 3-3. Enrichment of protein groups in OMIM genes.

Gene set sizes: total = 16812, OMIM = 3703, CpD-C = 3239, CpD-K = 2488, CpD-Y = 1026, CpD-CKY = 615, FDA = 707, Homozygous LoF tolerant (Lek et al 2016) = 285. Based on universe reference file with no missing values for all gene sets. *p threshold < 0.007142857, all points were significant.

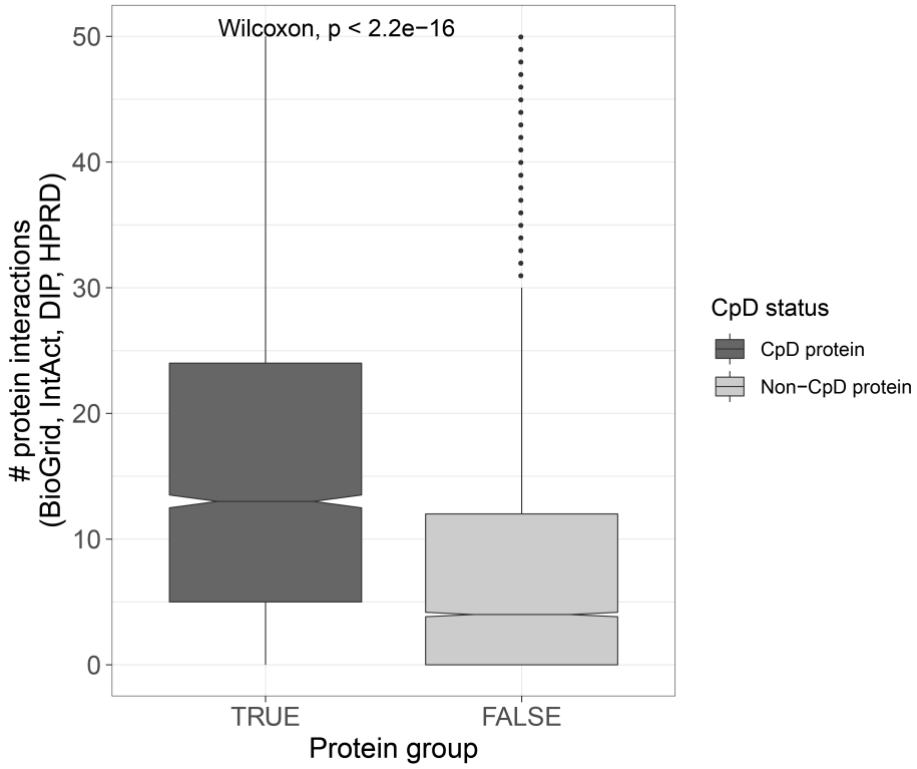

Figure 3-4. Significant association between CpD as an annotation related to higher interactivity of proteins.

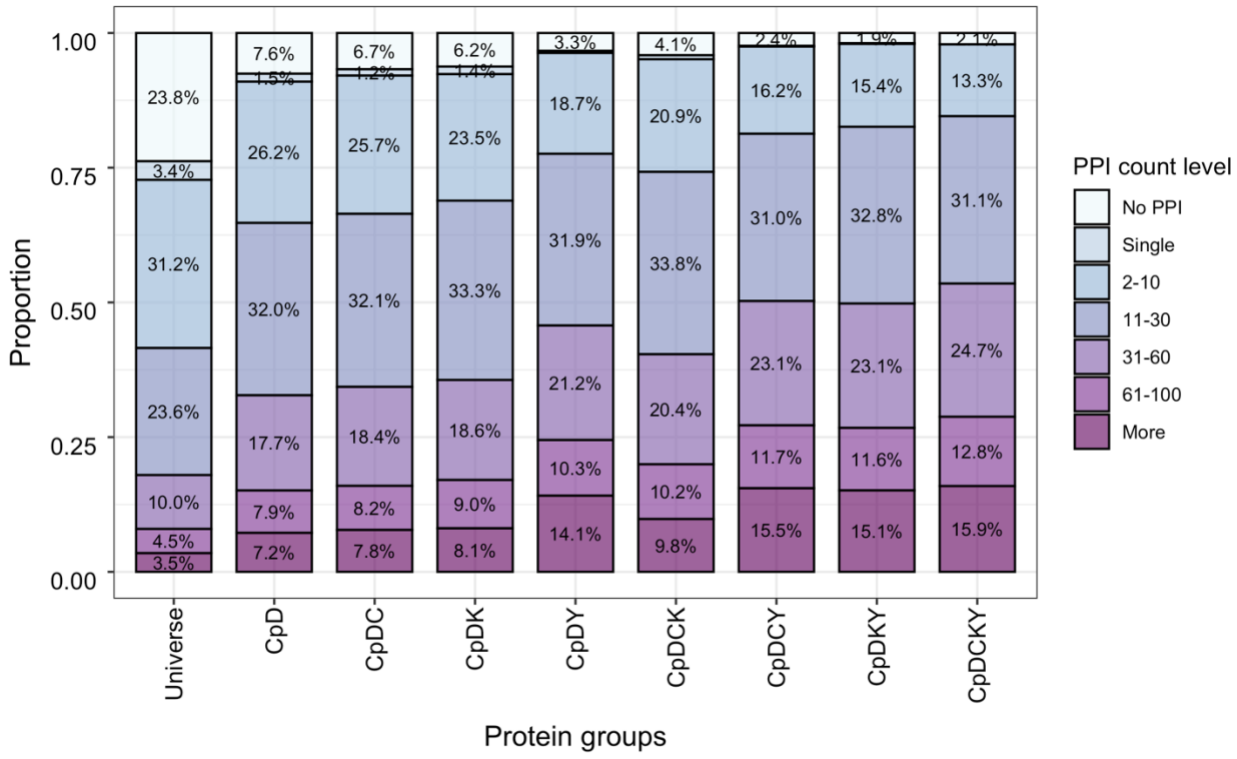

Figure 3-5. Distribution of protein interaction levels amongst all proteins, CpD proteins, and subsets of CpD proteins based on profiled residues types detected for each protein.

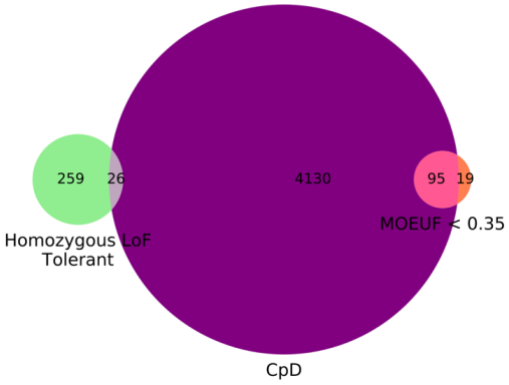

Figure 3-6. Venn diagram shows overlaps of CpD proteins with missense constrained genes (based on constraint cut-off < 0.35) and homozygous LoF tolerant genes.

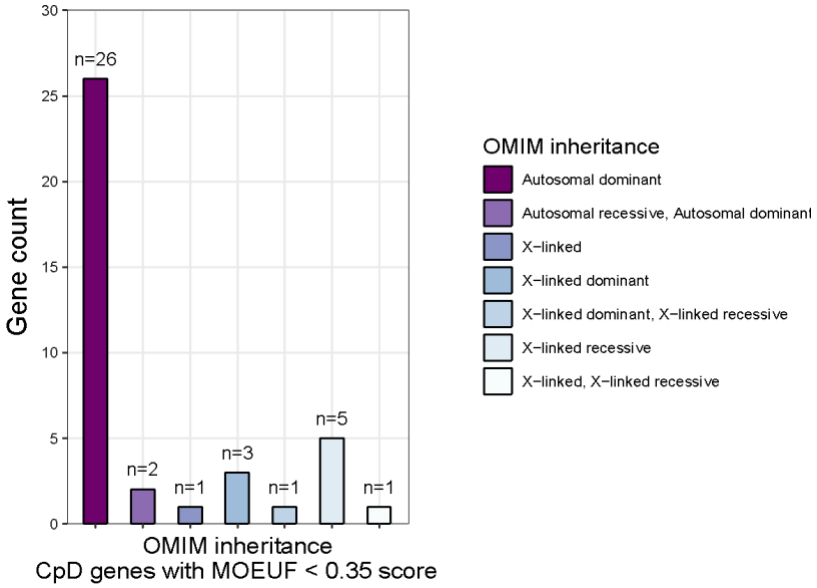

Figure 3-7. OMIM inheritance of single gene disorder phenotypes and CpD genes highly constrained to missense mutations (gnomAD MOEUF < 0.35).

Genes with missing values for phenotype inheritance are not shown, CpD *n*=39 genes out of 95 total CpD missense constrained genes in the universe of 16,812 genes.

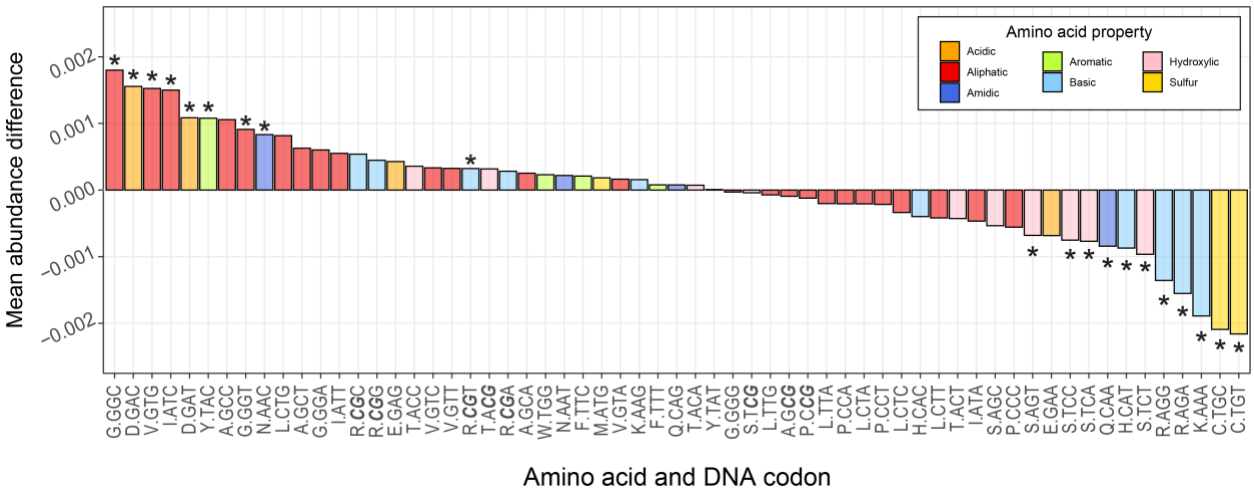

Figure 3-8. Differences in mean abundance of 61 codons in OMIM genes (n=3744) versus all other genes (n=13543).

Total of 17287 human genes included in the analysis. The codon frequency of each gene was normalized by total codons counted per gene and averaged for all genes in the OMIM and non-OMIM gene sets. The codon labels shown on the x axis are formatted as the single amino acid letter abbreviation followed by the synonymous DNA codon (i.e. 'C.TGT' for Cysteine's TGT codon). Eight codons on the x axis contain CpG dinucleotides marked by bold italic font. The y axis shows mean normalized codon abundance differences. Bar colors are based on physiochemical properties of the encoded amino acid residues. Significant abundance differences were determined using a two-sided, two-sample Welch's t test and permutation without replacement test. *p values < 5.0e-05 in Welch's t test and 1000/1000 permutation instances.

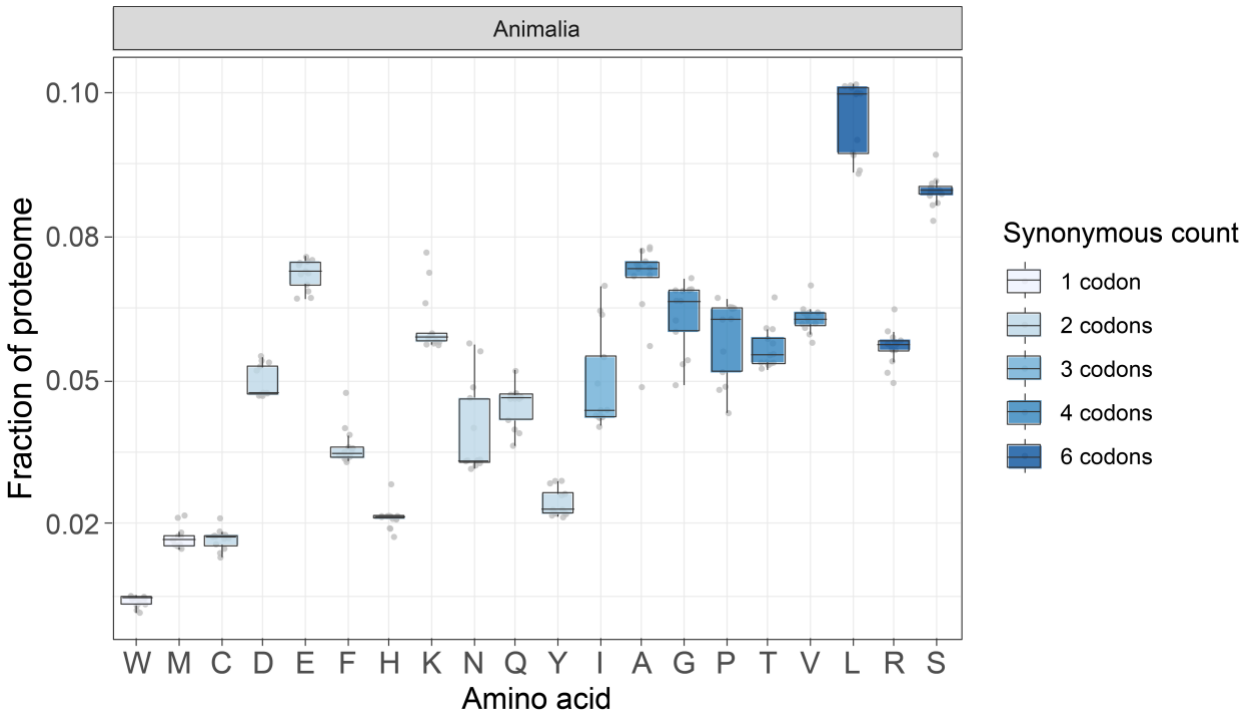

Figure 3-9. Amino acid average occurrence frequencies in animals.

The boxplot shows the distribution of average residue occurrence frequency across 13 diverse organisms from the Animal kingdom. Amino acid normalized abundances per protein were first calculated and the average frequency for all proteins in each complete proteome set was assigned to each organism, represented by the grey points on the plot. The interquartile ranges for most residues are tight, reflecting the well-conserved nature of amino acid frequencies across species from the same kingdom. Positive correlation of residue abundance and synonymous codon count (color of boxplots) is also shown by the figure.

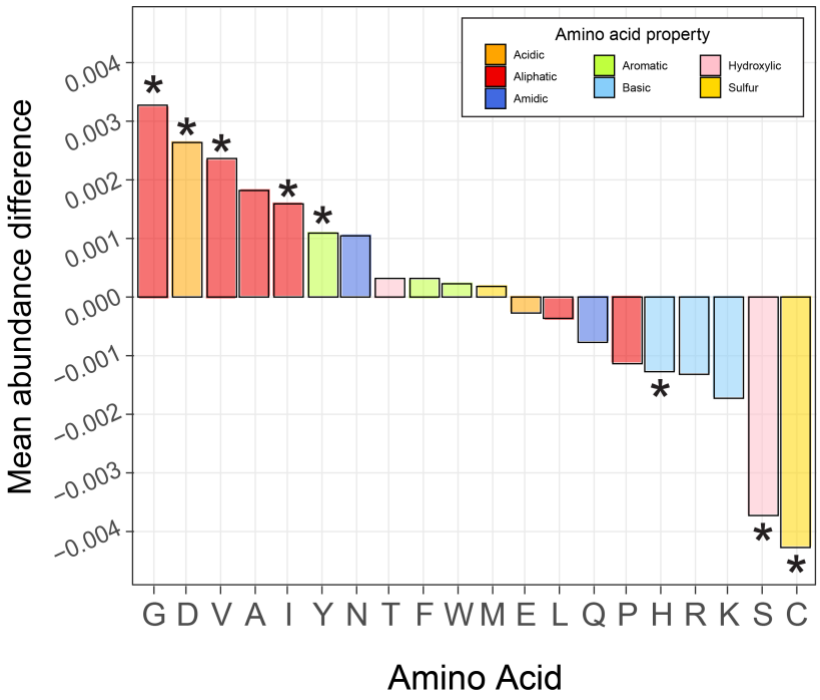

Figure 3-10. Differences in mean abundance for 20 amino acids in OMIM proteins (*n* = 3744) versus all other proteins (*n* = 13543).

Amino acid frequency per protein was normalized by protein sequence length and averaged for all proteins in the OMIM and non-OMIM gene sets. The mean abundance of each letter from the OMIM subset of proteins was subtracted from the mean abundance of non-OMIM proteins, with values shown on the y axis. Bar colors are based on physiochemical properties of amino acid residues. Significant abundance differences were determined using a two-sided, two-sample Welch's t test and permutation without replacement test. **p* values < 5.0e-05 in Welch's t test and 1000/1000 permutation instances.

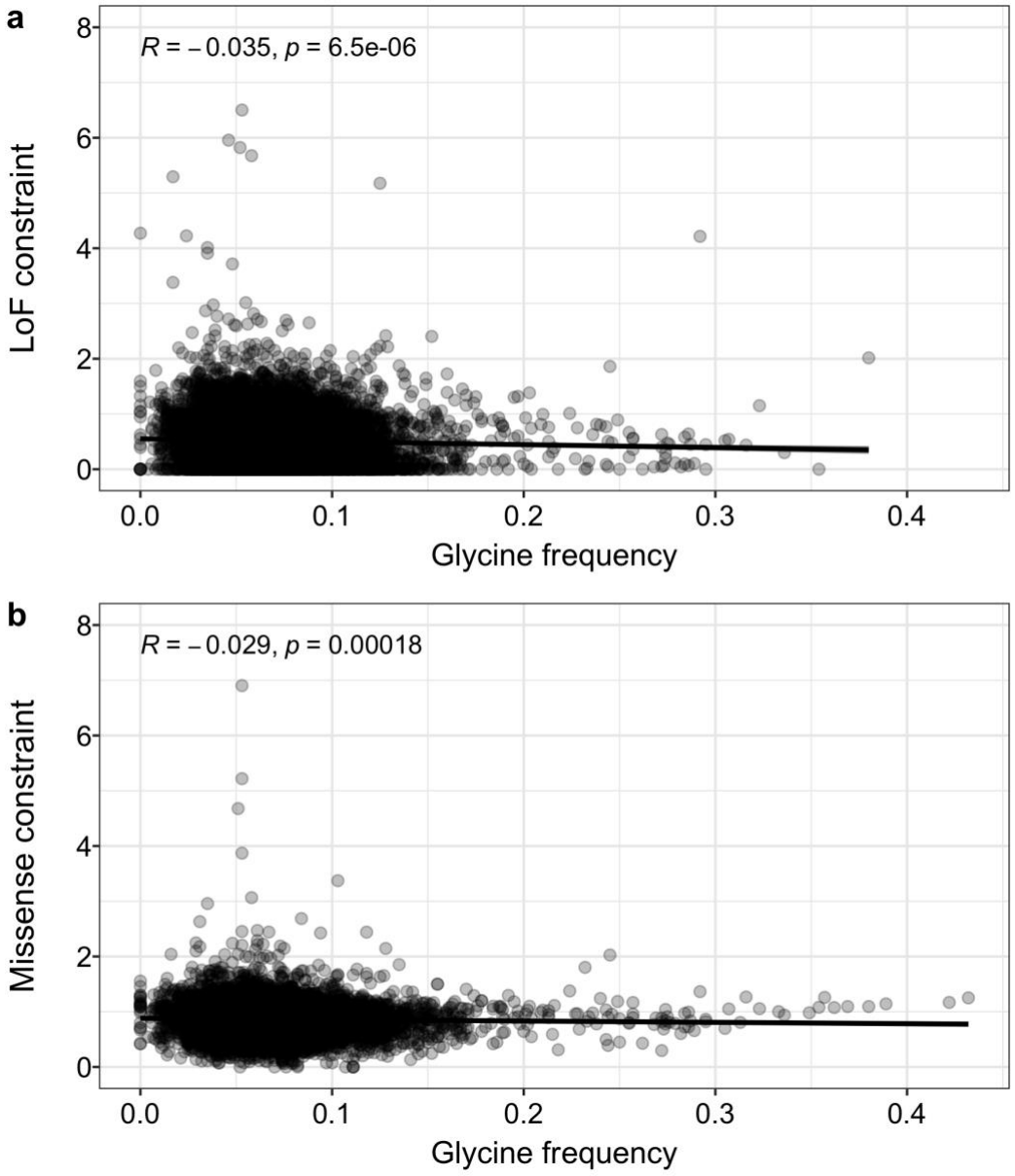

Figure 3-11. Glycine frequency and gnomAD constraint.

The observed-over-expected (o/e) loss-of-function (LoF) constraint metric (**a**) and the missense o/e constraint metric (**b**) are shown with the Pearson correlation test results annotated on each plot. The points in the scatter plots represent unique proteins. Total of 16,639 human proteins represented in each plot.

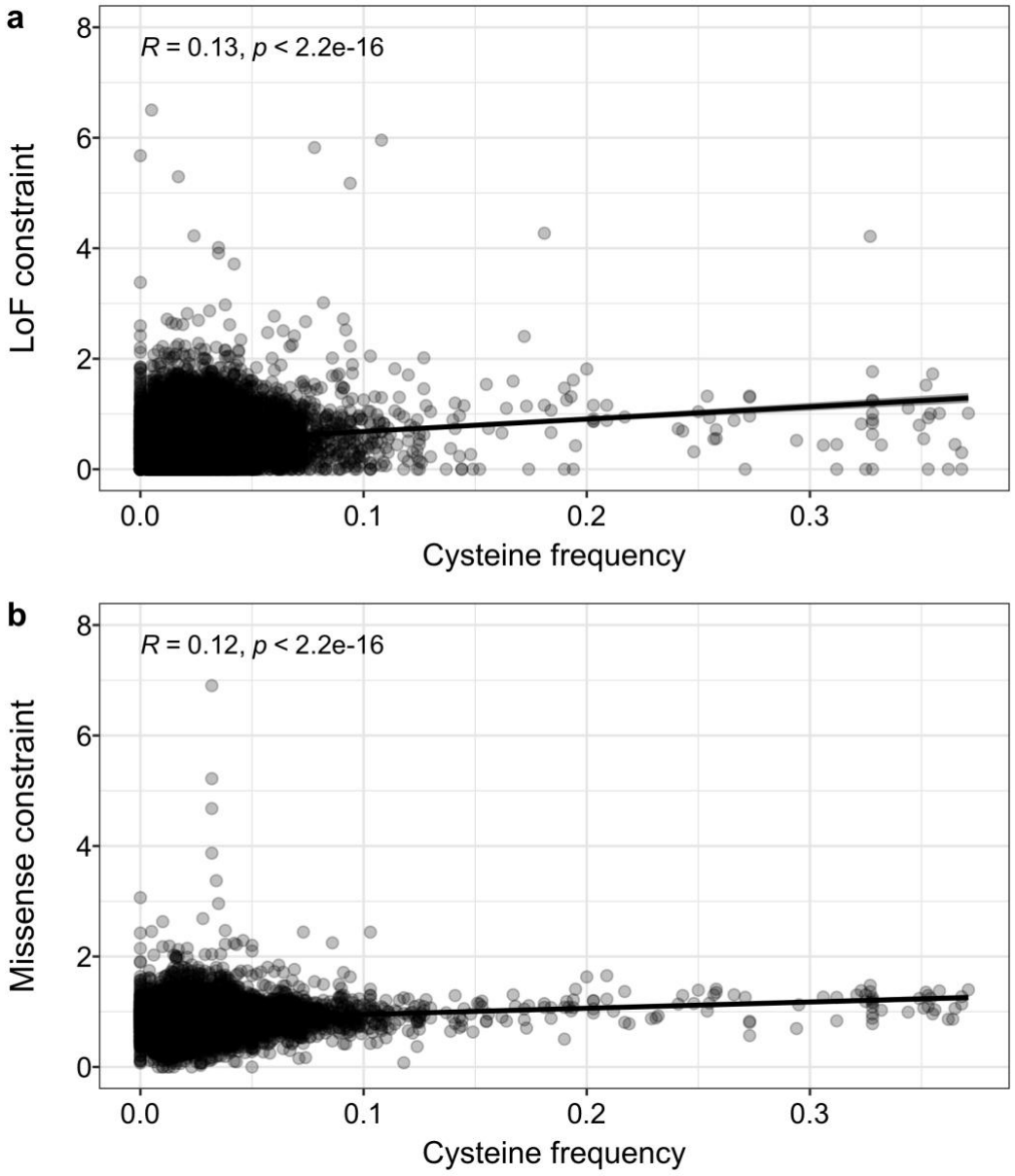

Figure 3-12. Cysteine frequency and gnomAD constraint.

The observed-over-expected (o/e) loss-of-function (LoF) constraint metric (**a**) and the missense o/e constraint metric (**b**) are shown with the Pearson correlation test results annotated on each plot. The points in the scatter plots represent unique proteins. Total of 16,639 human proteins represented in each plot.

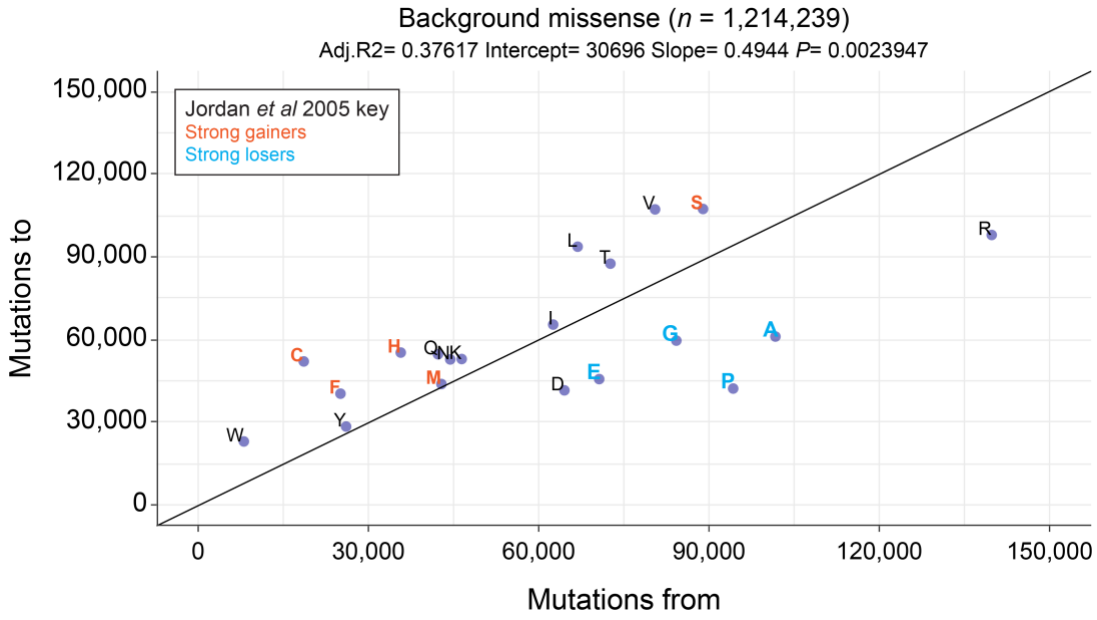

Figure 3-13.Comparisons between the number of background missense involving loss and gain of specific amino acid types.

The scatter plot shows the asymmetry between how often a residue is lost and gained by missense substitutions specifically for OMIM gene and all gnomAD population missense that mapped to this gene set. The x and y axis reflect raw missense counts and the black line shows a slope of one, representing perfect equilibrium between forward and reverse mutations. Each amino acid point is colored according to the missense category (dark blue for background), with letter colors showing amino acids reported as strong gainers (orange) or strong loser (light blue) by Jordan et al 2005. Results of fitting a linear regression model are also included under the plot title.

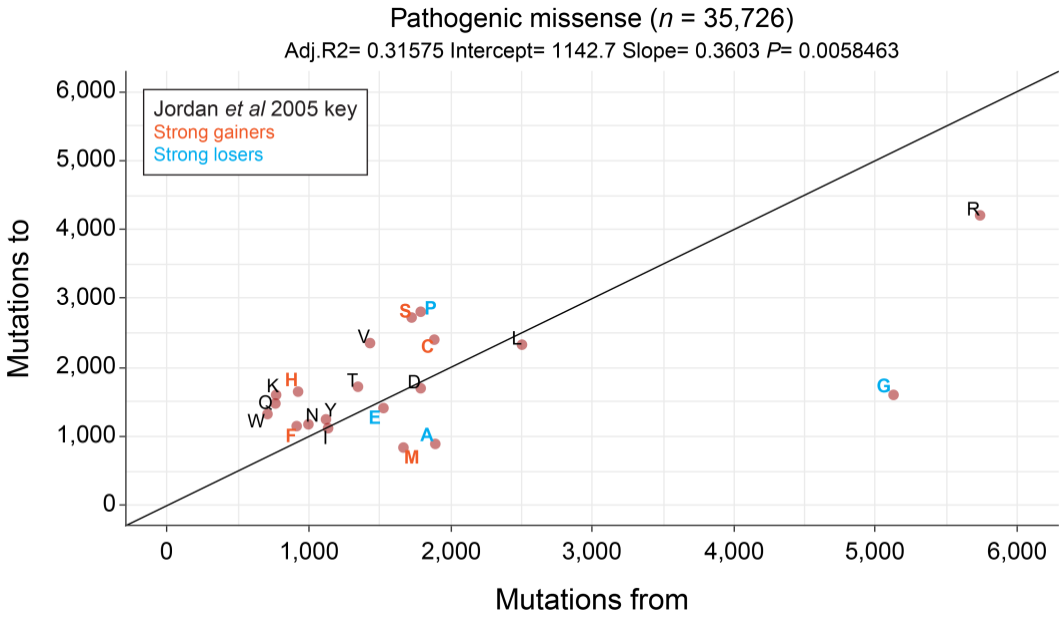

Figure 3-14. Comparisons between the number of pathogenic missense involving loss and gain of specific amino acid types.

The scatter plot shows the asymmetry between how often a residue is lost and gained by missense substitutions specifically for OMIM gene and all ClinVar pathogenic missense that mapped to this gene set. The x and y axis reflect raw missense counts and the black line shows a slope of one, representing perfect equilibrium between forward and reverse mutations. Each amino acid point is colored according to the missense category (red for pathogenic), with letter colors showing amino acids reported as strong gainers (orange) or strong loser (light blue) by Jordan et al 2005. Results of fitting a linear regression model are also included under the plot title.

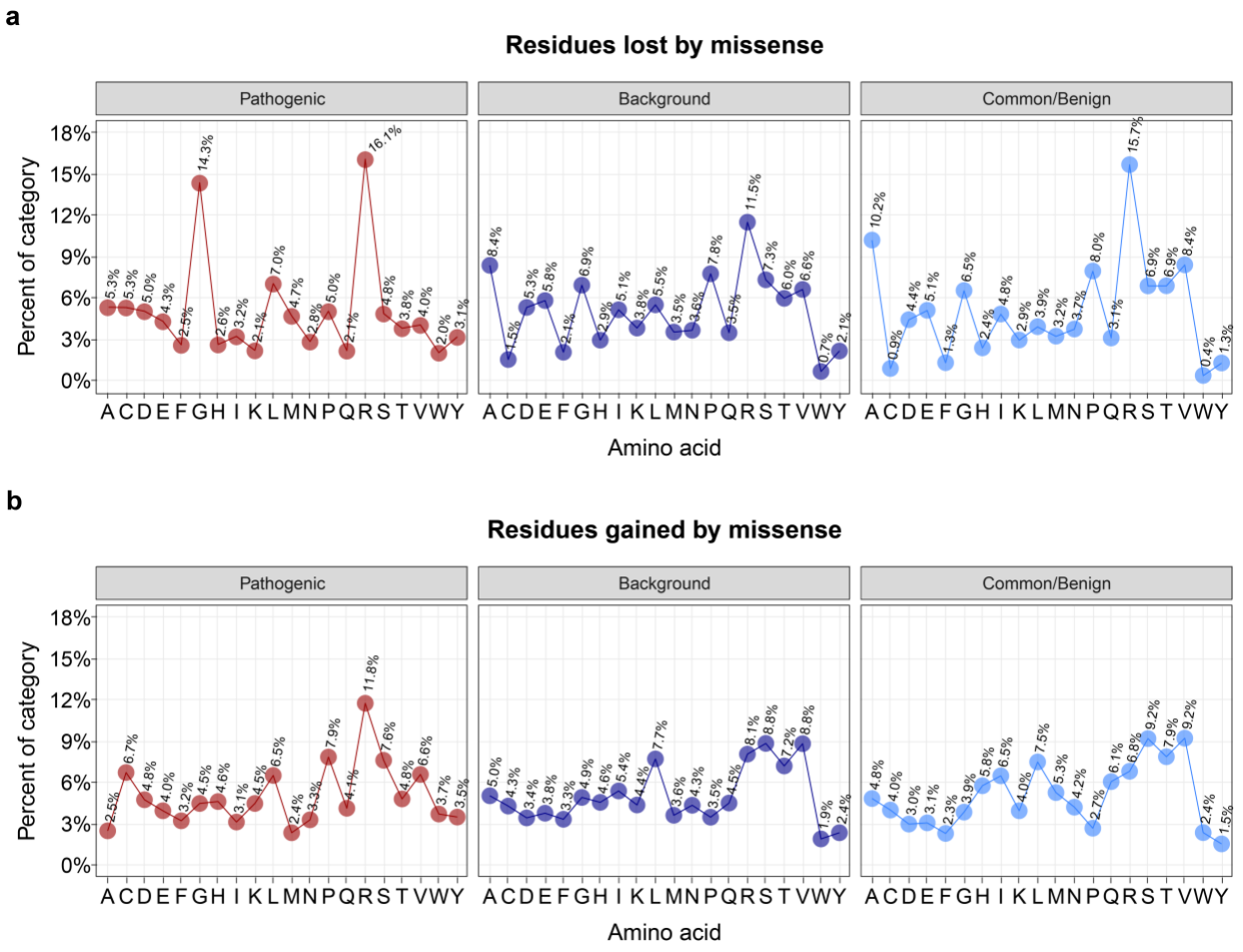

Figure 3-15. Proportion of pathogenic and likely neutral missense substitutions that involve a specific amino acid in OMIM genes.

The missense substitutions are divided into lost amino acids (a) and gained amino acids (b). The y axis shows the percent of missense that involved the gain or loss of a specific residue shown on the x axis. The three variant categories separated by panels are colored as red for pathogenic, dark blue for background, and light blue for common/benign.

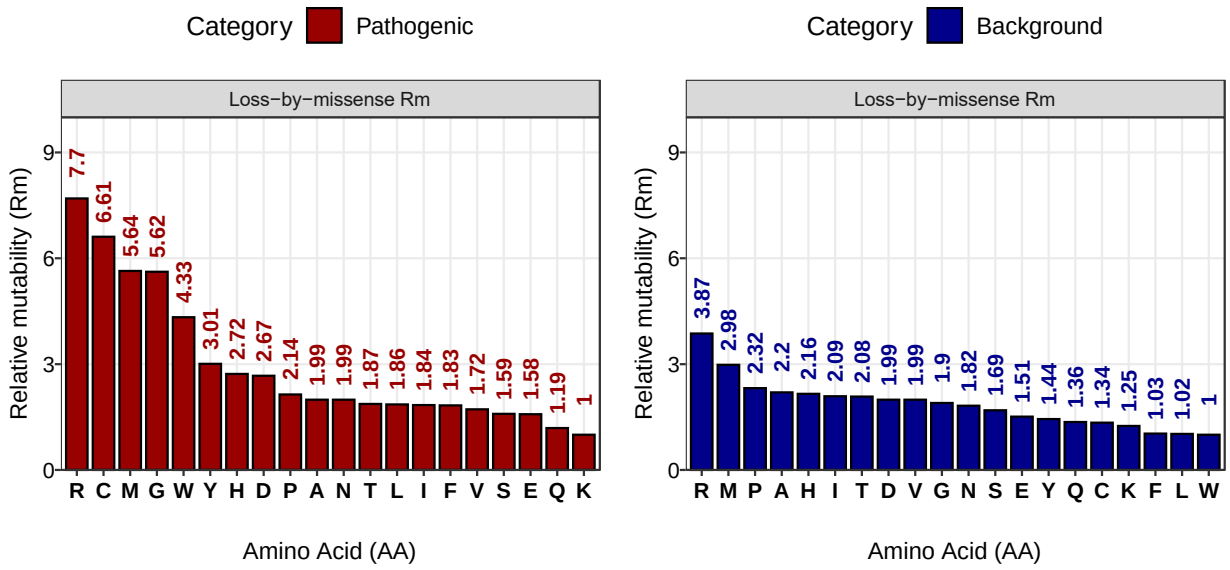

Figure 3-16. Relative mutability of amino acids lost-by missense substitutions in the pathogenic and background categories for OMIM genes.

The x axis shows the single letter abbreviations of the twenty amino acids analyzed. The y axis represents residue relative mutability (Rm) calculated using an equation from Khan et al 2007. The bar heights reflect residue Rm values and are ordered in each plot from left to right by decreasing Rm values.

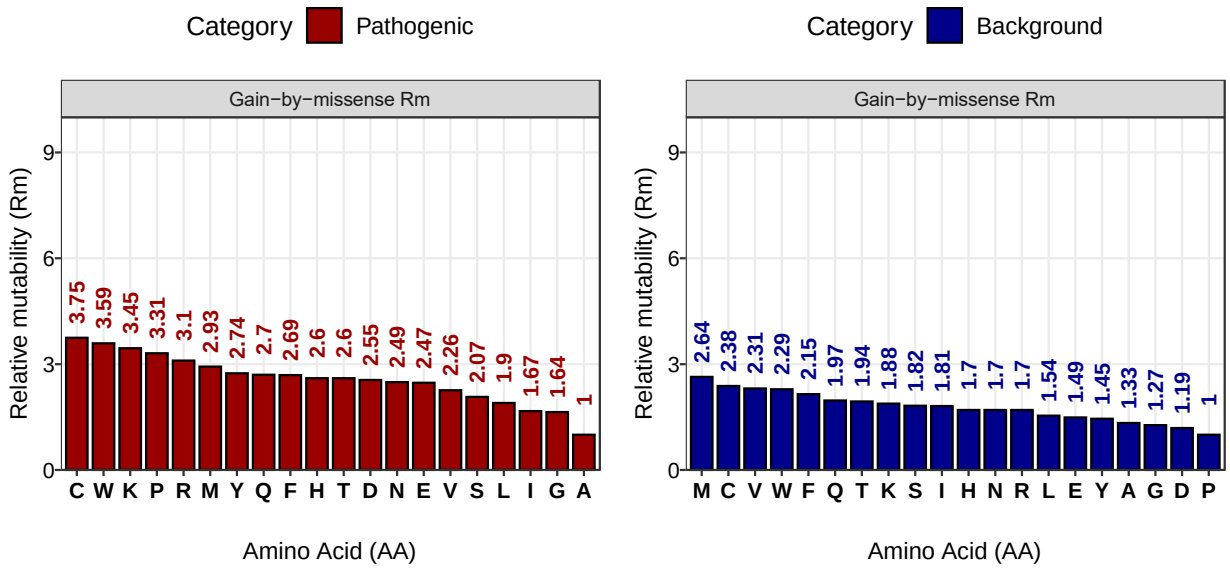

Figure 3-17. Relative mutability of amino acids gained-by missense substitutions in the pathogenic and background categories for OMIM genes.

The x axis shows the single letter abbreviations of the twenty amino acids analyzed. The y axis represents residue relative mutability (Rm) calculated using an equation from Khan et al 2007. The bar heights reflect residue Rm values and are ordered in each plot from left to right by decreasing Rm values.

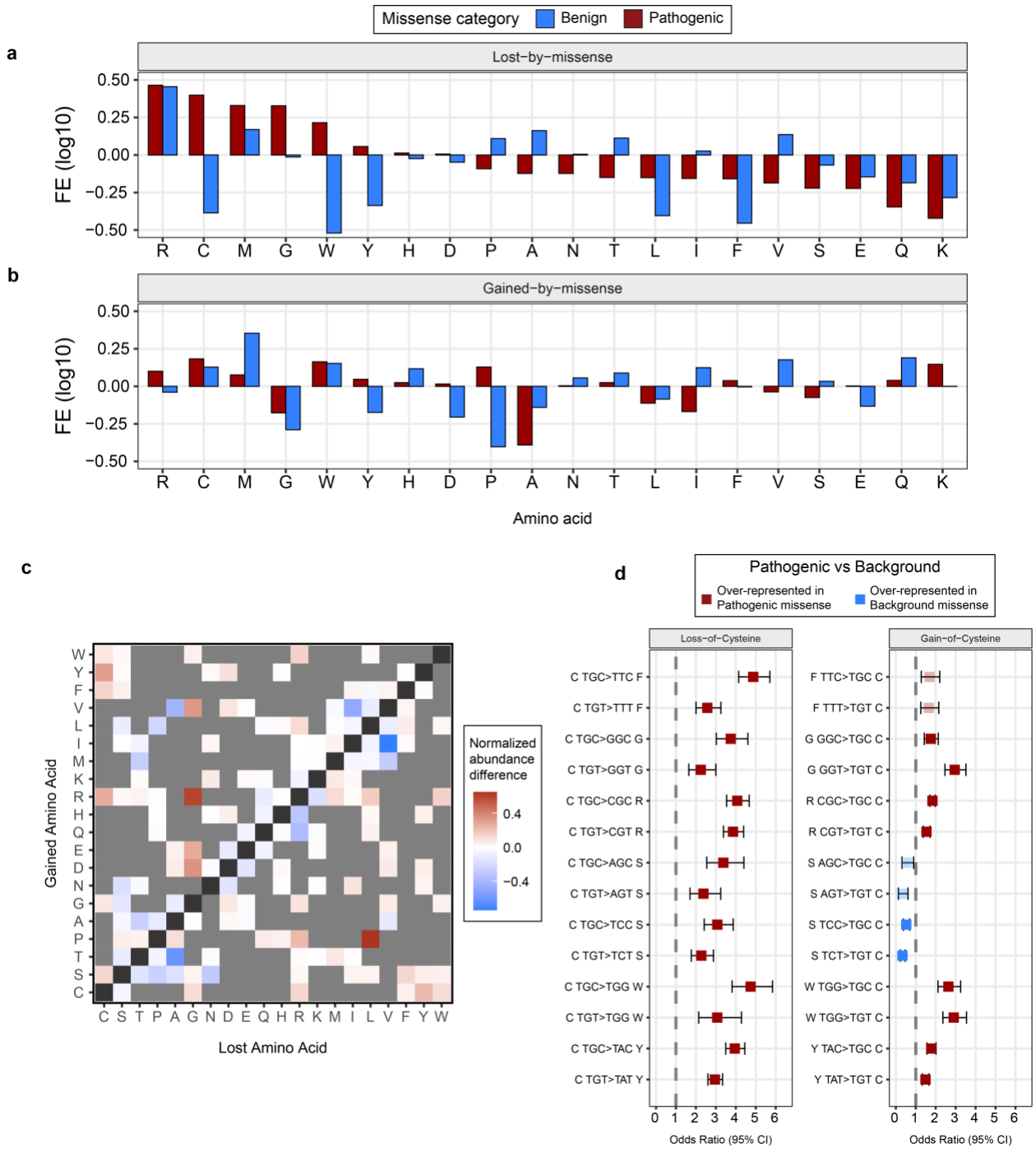

Figure 3-18. Asymmetry of residue gains and losses by missense substitutions in OMIM genes.

a) Lost-by-missense amino acid fold enrichment (FE) relative to residue abundance for pathogenic (red) and common/benign (blue) variants.

b) Gained-by-missense amino acid fold enrichment (FE) relative to possible missense based on mutated codon abundance for pathogenic (red) and common/benign (blue) variants.

a.b. Fold enrichment was calculated as the abundance of a specific type of missense substitution in a category (i.e. % of Cysteine-loss substitutions in pathogenic missense) divided by the amino acid's abundance in monogenic disorder-associated proteins (i.e. total cysteine / total amino acids in protein subset). FE values are positive and were log10 transformed to clearly reflect enriched substitution types as positive bars and un-enriched substitution types as negative bars (values between 0-1 FE before transformation).

c) Heatmap of the difference between Pathogenic and Common/Benign missense normalized abundance. Heatmap colors reflect the difference between Pathogenic minus Common/Benign category for a specific missense substitution, with red squares indicating greater abundance in the Pathogenic category, blue squares indicating greater abundance in the Common/Benign category, white squares indicating no difference between the two categories, gray squares indicating substitutions not possible by single nucleotide change, and black squares are silent (synonymous) variants not included in the study. The analysis was restricted to a subset of 2,873 proteins that have annotations for both categories of missense variant, resulting in a final set of 33,872 Pathogenic and 29,778 Common/Benign variants reelected in the heatmap. The x axis shows the lost (or mutated) amino acid and the y axis shows the

gained (or mutant) amino acid. Amino acid single letter labels are ordered based on the side-chain chemistry {Mount DW: Bioinformatics Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 2001.} Substitutions closer to the diagonal line of black squares are considered more conserved, and substitutions farther from the diagonal line are considered less conserved. Pathogenic and Common/Benign missense abundance matrices were normalized so that the sum over all mutation frequencies equals 1.

(C) sulfhydryl; (STPAG) small hydrophilic; (NDEQ) acid, acid amide and hydrophilic; (HRK) basic; (MILV) hydrophobic; (FYW) large hydrophobic/aromatic.

d) Magnitude of enrichment for missense involving cysteine in the Pathogenic versus Background missense categories. 95% confidence intervals (line segments) and odds ratios (squares). All possible substitutions by single nucleotide variants were counted, resulting in 3489 Pathogenic and 65505 Background mutations involving a gain or loss of cysteine. Red squares are substitutions enriched in the Pathogenic category and blue squares are substitutions enriched in the Background category. Non-significant odds are shown as transparent squares. Significant Bonferroni-adjusted $p < 6.38e-05$ (two-tailed Fisher's exact test).

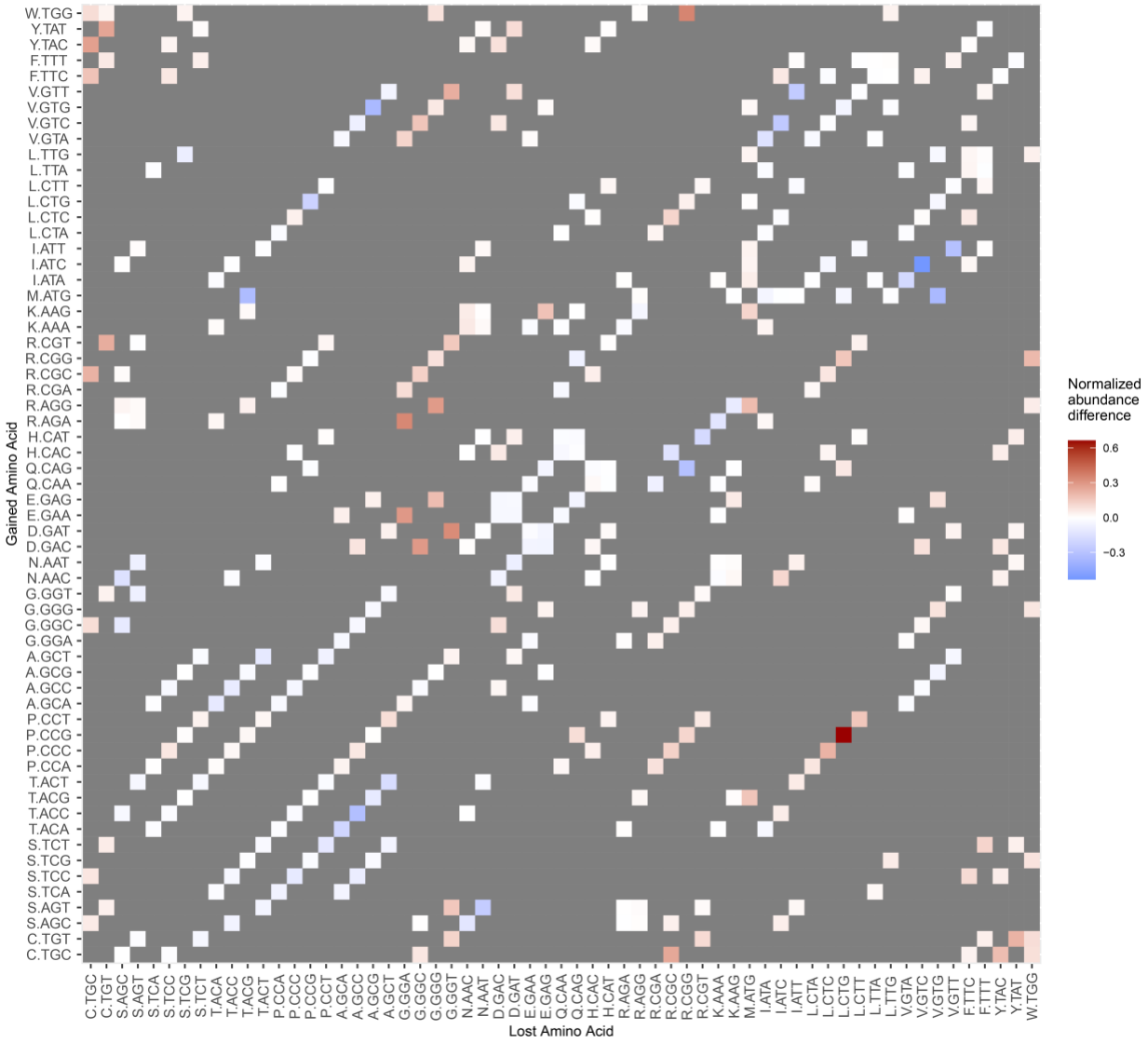

Figure 3-19. Difference in codon exchanges for pathogenic vs common/benign missense in OMIM genes.

Heatmap colors reflect the difference between the abundance of a specific missense substitution in the Pathogenic category minus the Common/Benign category, with red indicating greater abundance in the Pathogenic category, and blue indicating greater abundance in the Common/Benign category. White squares indicate no difference

between the two categories and gray indicates either a substitution not possible by single nucleotide change or a silent (synonymous) substitution (the heatmap identity line). The analysis was restricted to the subset of 2,873 OMIM proteins that have annotations for both categories of missense variants, resulting in a final set of 33,872 Pathogenic and 29,778 Common/Benign variants included in the analysis. Axis labels are formatted as the single letter for an amino acid and its corresponding DNA codon (i.e. cysteine as ‘C.TGT’ and ‘C.TGC’). The x axis represents the lost (or mutated) codon, and the y axis represents the gained (or mutant) codon. The three stop codons (TAA, TAG, TGA) were excluded from our analysis, resulting in a 61x61 matrix. The codons are ordered based on the same amino acid order shown in Figure 2c heatmap, with more conservative substitutions of residues tending closer to the heatmap identity line. The missense category matrices used as input to calculate the difference matrix were normalized first so that the sum over all mutation frequencies in a category equals 1.

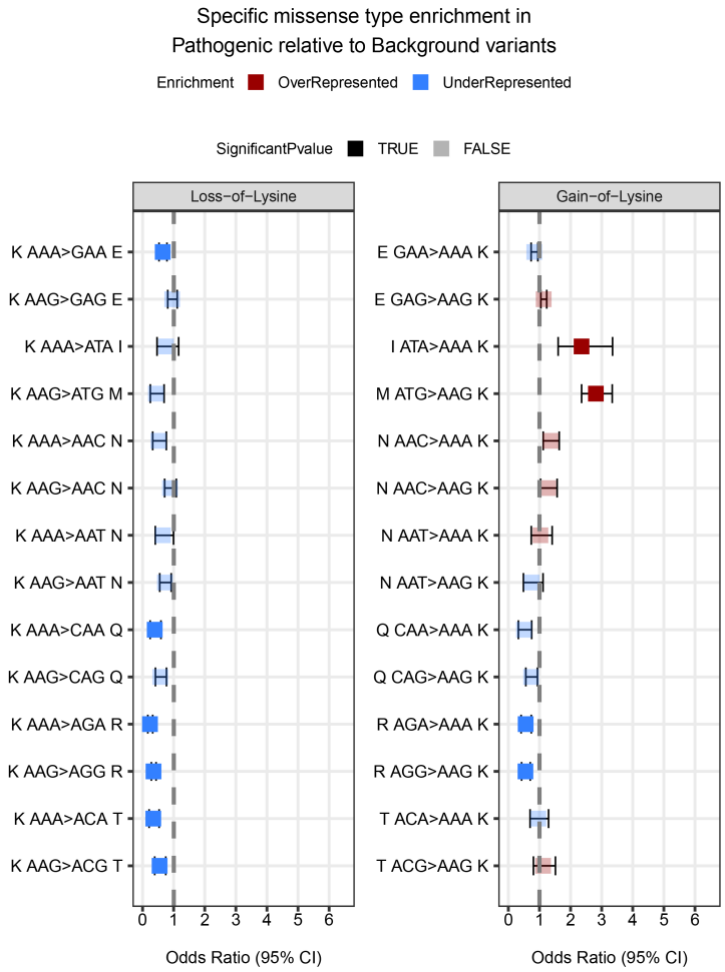

Figure 3-20. Magnitude of enrichment for missense involving lysine in the pathogenic versus background missense categories.

All possible substitutions by single nucleotide variants were counted; 95% confidence intervals (line segments) and odds ratios (squares). Red squares are substitutions enriched in the Pathogenic category and blue squares are substitutions enriched in the Background category. Non-significant odds are shown as transparent squares.

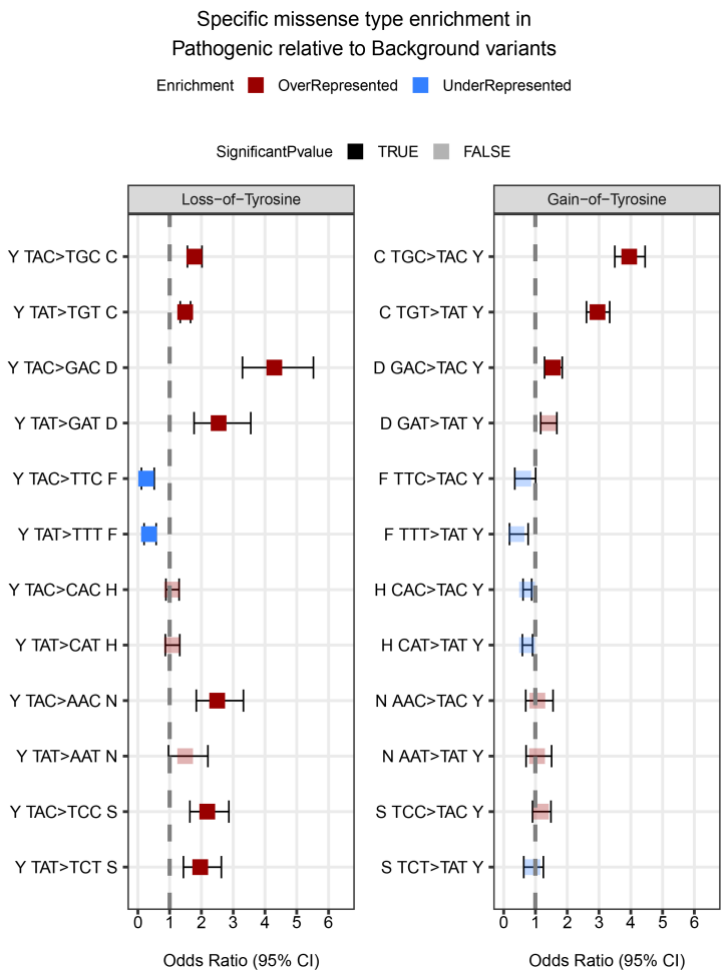

Figure 3-21. Magnitude of enrichment for missense involving tyrosine in the pathogenic versus background missense categories.

All possible substitutions by single nucleotide variants were counted; 95% confidence intervals (line segments) and odds ratios (squares). Red squares are substitutions enriched in the Pathogenic category and blue squares are substitutions enriched in the Background category. Non-significant odds are shown as transparent squares.

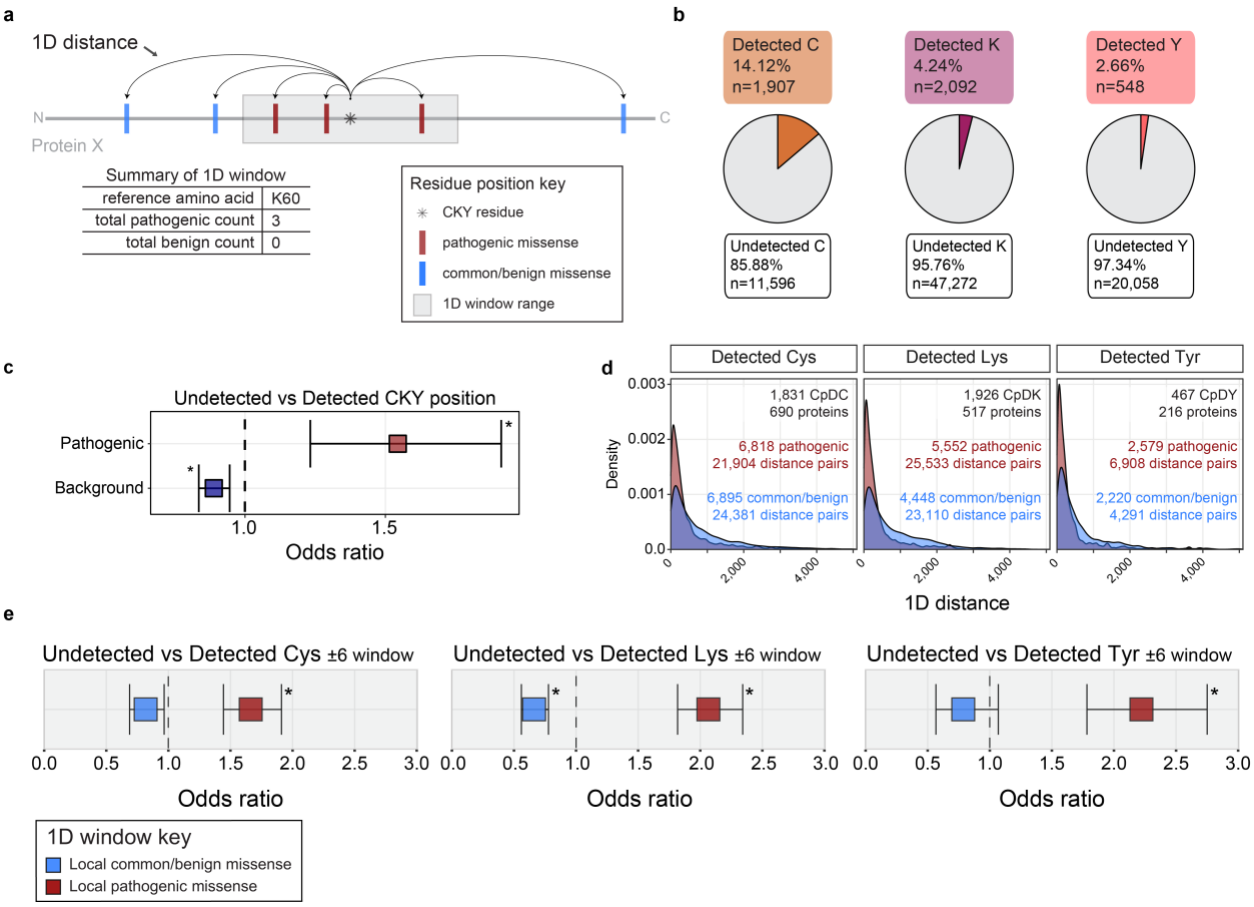

Figure 3-22. Chemoproteomic-detected amino acids are more associated to pathogenic missense than undetected residues in 1D space.

a) Schematic of 1D distance calculations between CKY residue positions to positions of missense variants. The absolute number of amino acids away a variant position was from a reference CKY position was assigned to an ID representing the distance pair (missense identifier plus protein position identifier of CKY residue). A sequence window centered on a reference CKY position is shown as a grey box in the cartoon. A summary table for the reference CKY position in the cartoon shows the total number of unique missense variants within a sequence range of lysine position 60 (K60) in protein X. The window summed variant counts excluded missense variants overlapping codons of reference CKY positions in our 1D window analysis.

b) Proportion of detected (*n* = 4,547) and undetected (*n* = 78,926) cysteine, lysine, and tyrosine residue positions in 926 OMIM&CpD proteins.

c) Odds greater than 1 indicate enrichment of specific missense category in detected residue windows. Odds less than 1 indicate enrichment of specific missense category in undetected residue windows, and depletion for detected residue windows.

Odds of pathogenic and background missense variant overlapping a detected (*n* = 5,854) versus undetected (*n* = 304,889) CKY residue position in OMIM proteins (*n* = 3,907). Bonferroni-corrected two-sided *p* value < 0.05 calculated by Fisher's exact test, **p* < 0.0042.

d) Detected residue distances to pathogenic versus common/benign missense positions in OMIM proteins. Counts differ from Figure 3b because proteins were controlled to have a specific CpDAA type and at least one pathogenic and one common/benign variant position for this analysis. Distance distributions are distance pairs are based on unique missense alternative amino acid change positions to unique positions of detected residue types. The distributions represent distances between nearest category missense positions to each CpDAA positions in the same protein and includes 1D distances of 0 for CpDAA-missense position overlaps.

e) Odds of pathogenic and common/benign missense variants in 1D windows of detected versus undetected CKY residues for 926 OMIM&CpD proteins. Error bars of points reflect 95% CI for missense within ±6 amino acid windows of reference CKY positions. Odds greater than 1 indicate enrichment of specific missense category in detected residue windows. Odds less than 1 indicate enrichment of specific missense category in undetected residue windows, and depletion for detected residue windows.

Overlaps between missense variant positions and window reference CKY positions were excluded from the analysis. Bonferroni-corrected two-sided p value < 0.05 calculated by Fisher's exact test, *p < 0.0083.

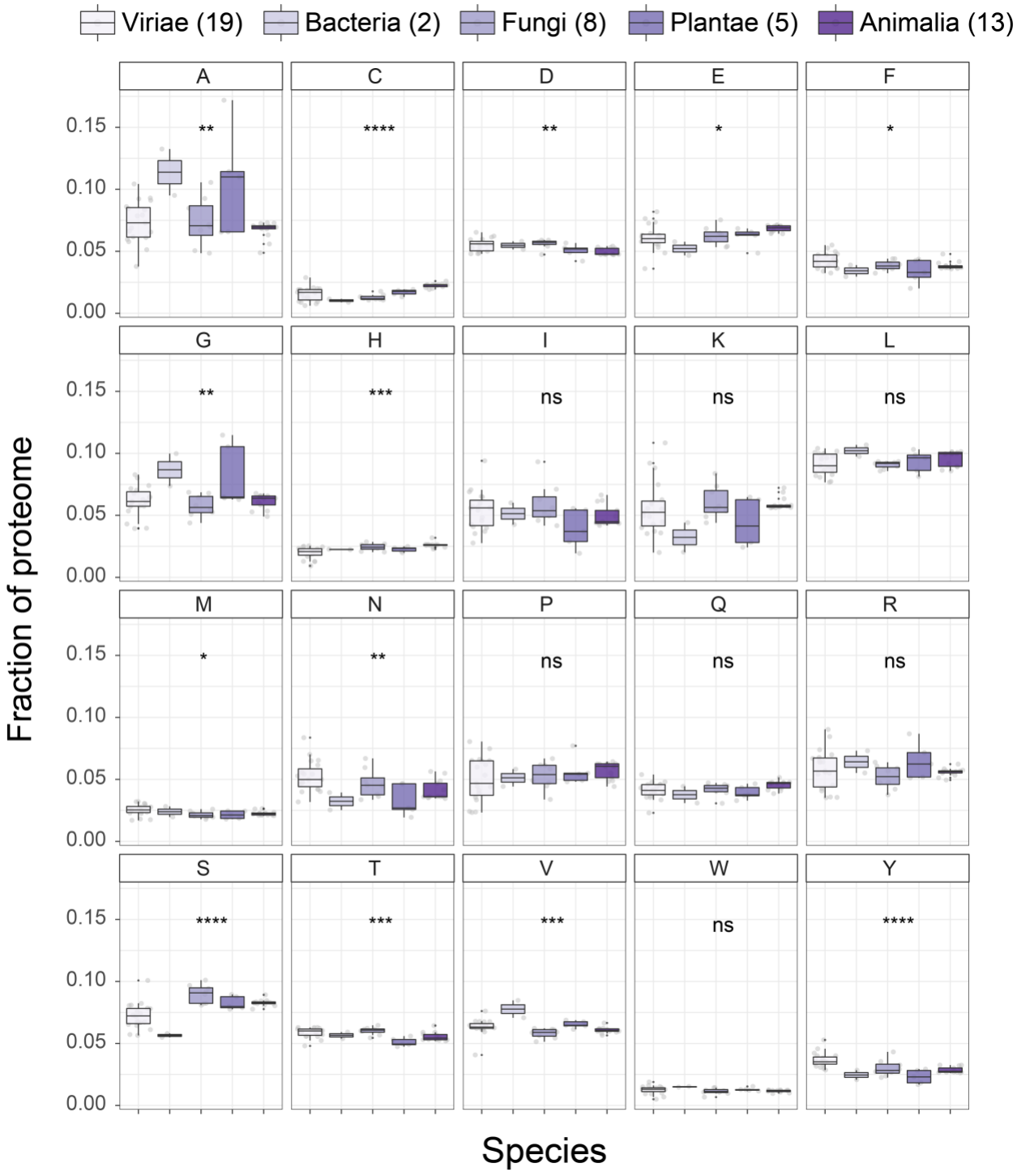

Figure 3-23. Distribution of amino acid frequencies in the total proteome sets of 47 species representing four kingdoms of life and viruses.

The species were grouped by kingdoms and virus, resulting in five species groups for comparisons. ANOVA was used to test significant mean differences for all groups. Total

numbers for unique species in each of the five groups is noted in parentheses next to the group label in the figure key. Specific species used in the analysis are shown in the ancestry circle plot in **Figure 3-24.**

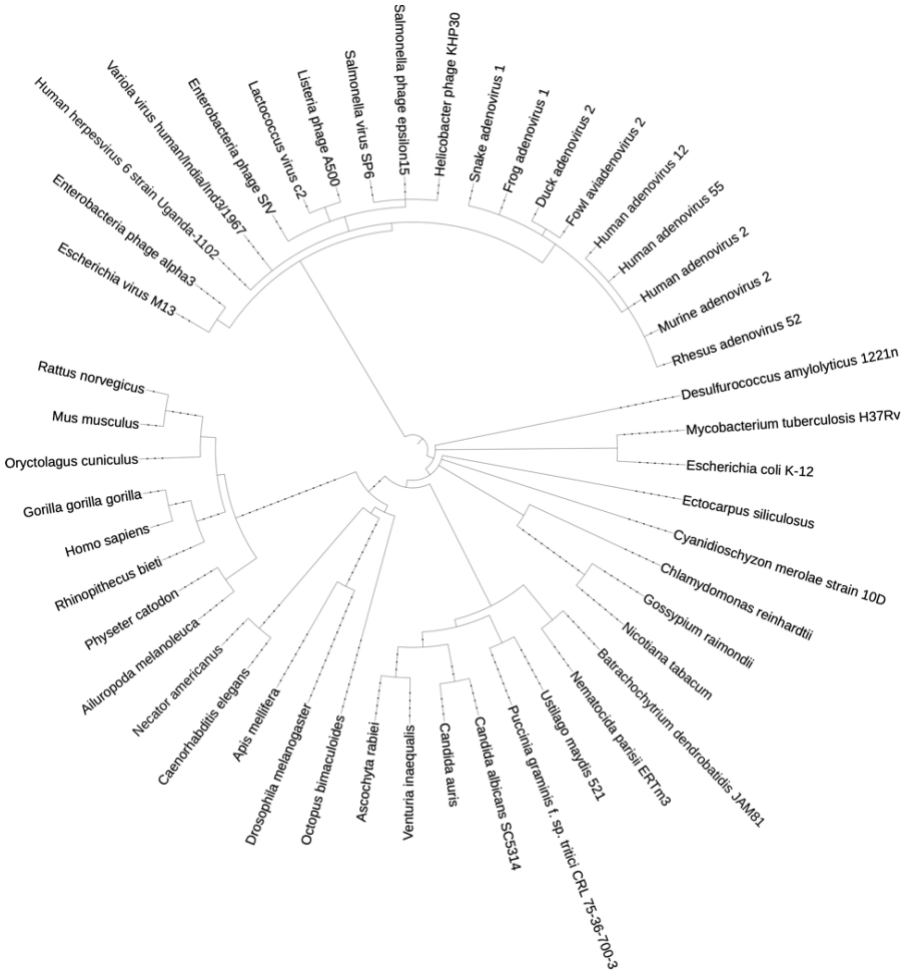

Figure 3-24. NCBI circle ancestry plot of species from the following major branches: Animalia, Plantae, Fungi, Bacteria, and Viriae.

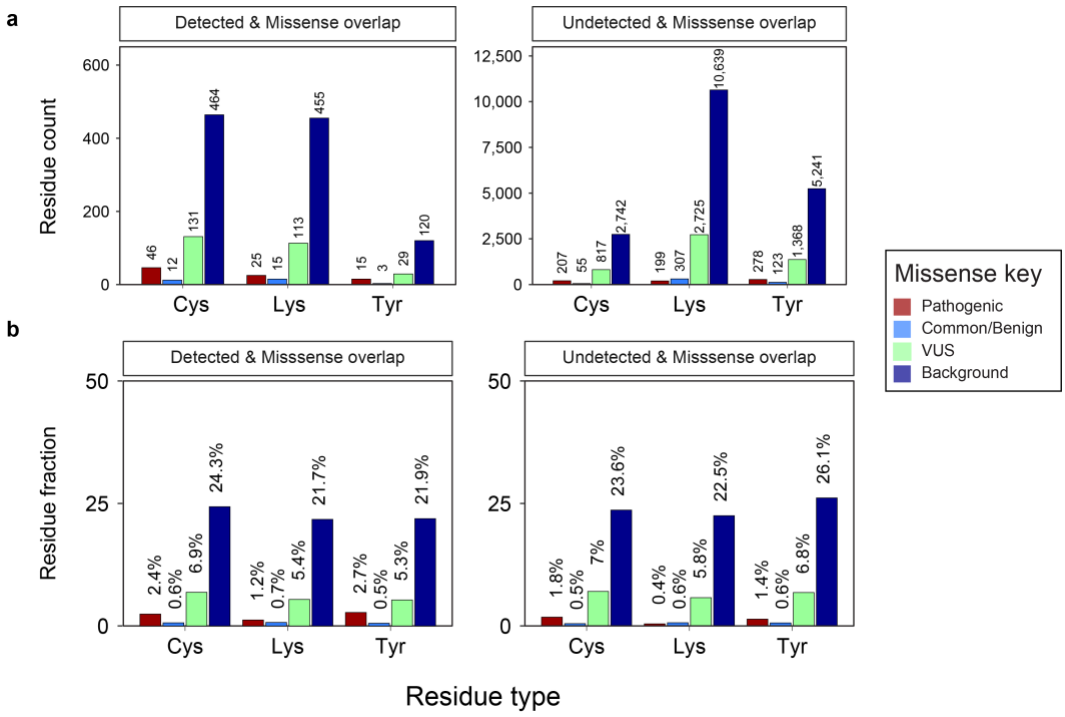

Figure 3-25. Detected versus undetected CKY positions overlapping missense variants in OMIM&CpD proteins.

The counts of CKY positions overlapping missense alleles (a) and the proportion of CKY positions overlapping missense alleles are based on 926 OMIM&CpD proteins.

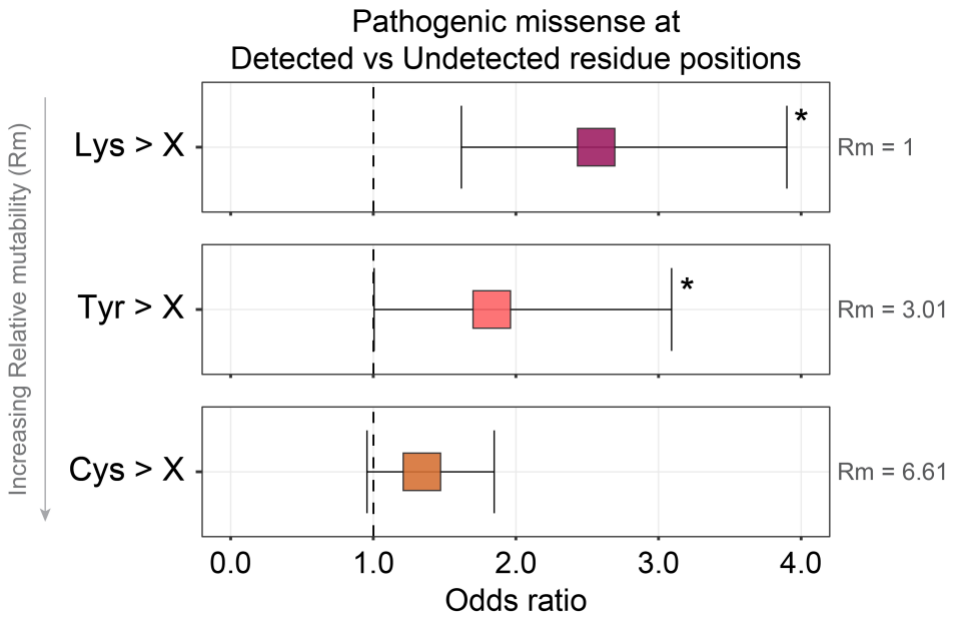

Figure 3-26. Odds of missense categories overlapping detected versus undetected cysteine, lysine, and tyrosine residues.

Bonferroni-corrected two-sided *p* value < 0.05. The x axis corresponds to the odds ratio for overlapping pathogenic missense variants. Values greater than 1 indicate pathogenic missense enrichment at codons of detected residue types. The figure panels are ordered by increasing amino acid relative mutability (Rm) for the missense variant pathogenic category. Error bars represent 95% CI. Based on 1,234 OMIM&CpD proteins

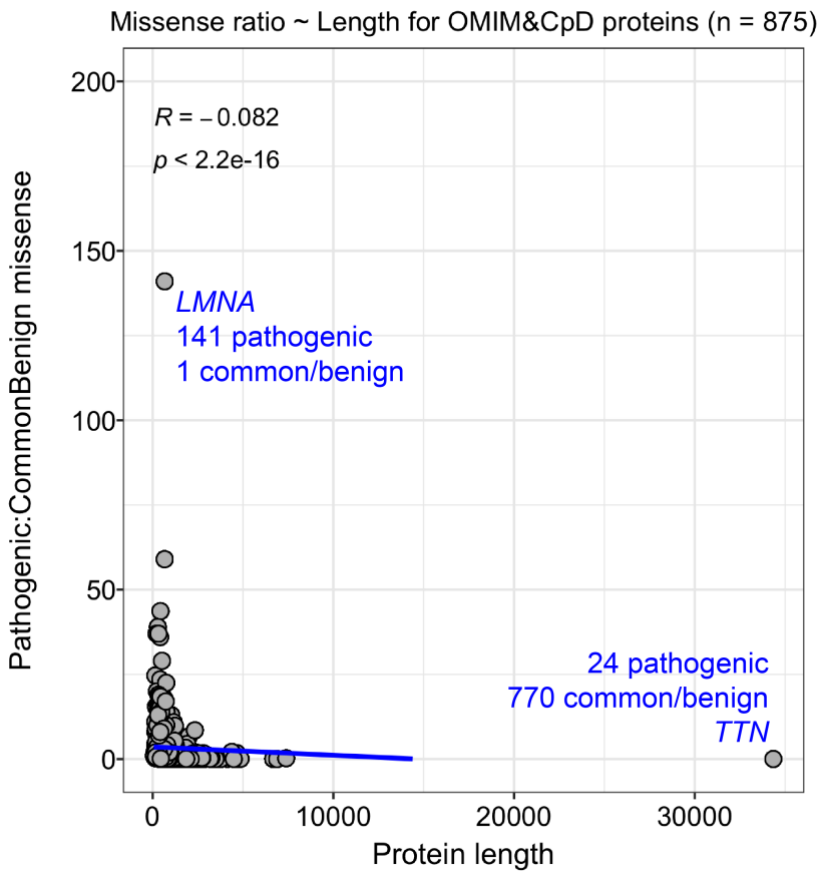

Figure 3-27. Ratio of pathogenic:common/benign missense and protein length. Outlier gene examples are annotated on plot along with the counts of unique pathogenic and common/benign missense per gene.

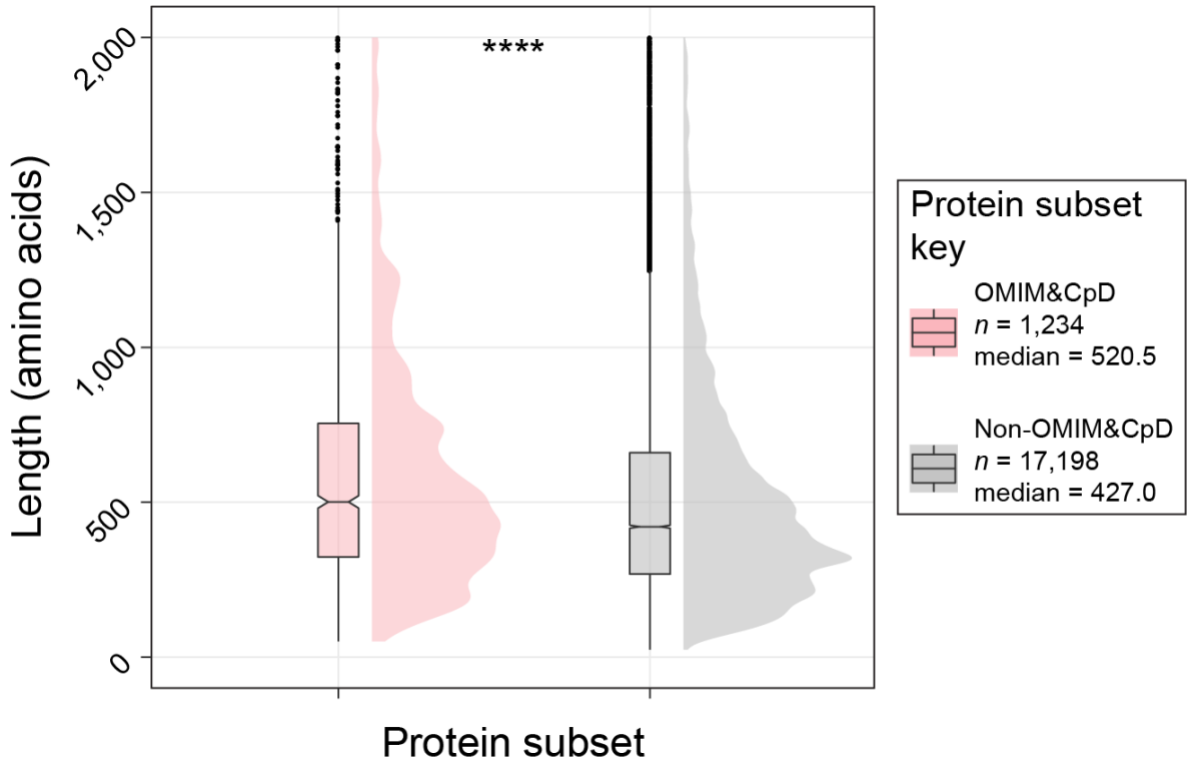

Figure 3-28. Protein length of OMIM&CpD versus all other proteins.

Protein counts and group median values shown in figure key. Wilcoxon test used for group mean comparison with FDR adjustment of *p* values. ****p < 2e-16. The median length for all human proteins was 434 amino acids. Plot based on 18,432 canonical UniProtKB human proteins.

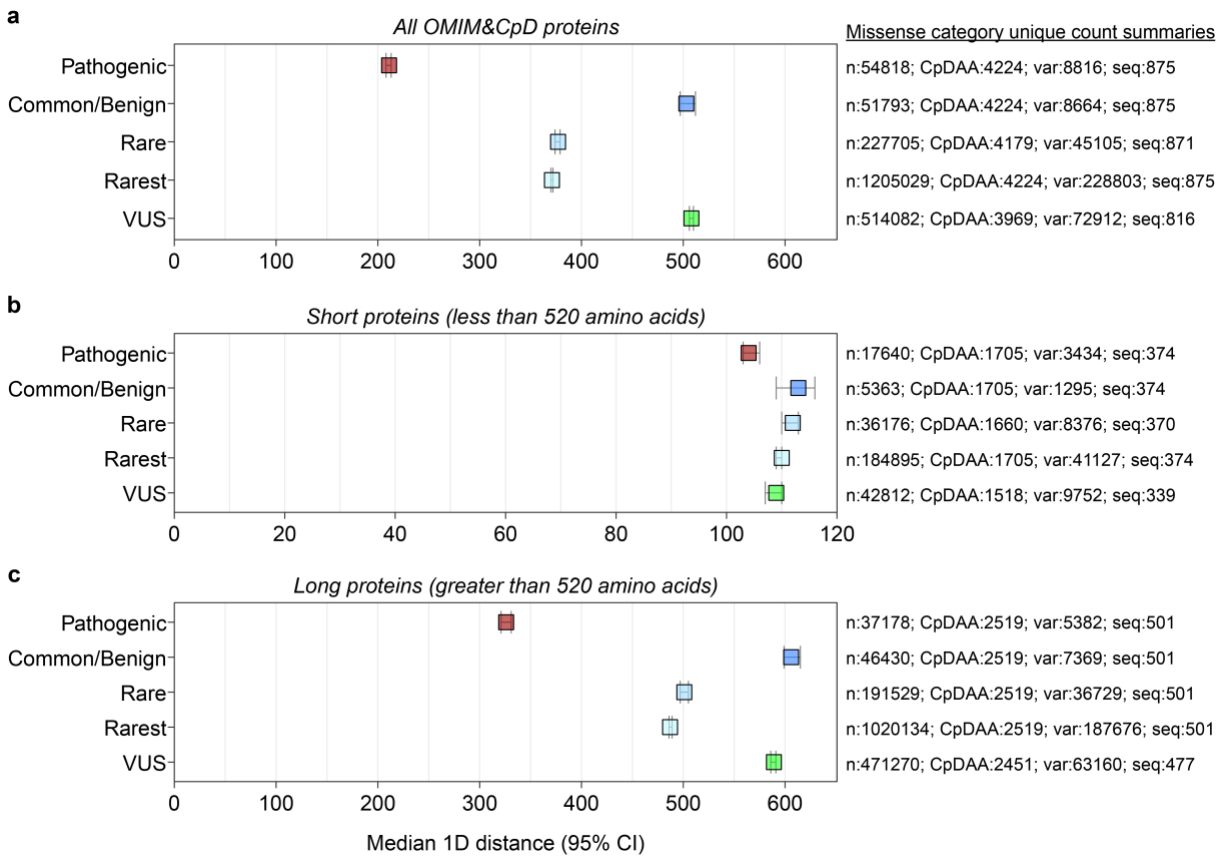

Figure 3-29. Missense categories to detected residue 1D distances.

Medians with bootstrapped 95% CI shown for OMIM&CpD proteins. The three subplots show (**a**) all OMIM&CpD proteins, (**b**) short OMIM&CpD proteins, and (**c**) long OMIM&CpD proteins with 1D distance unique count summaries of total distance pairs (n), detected CKY positions (CpDAA), missense alleles (var), and proteins (seq) for the missense categories (y axis) shown on the right of each subplot.

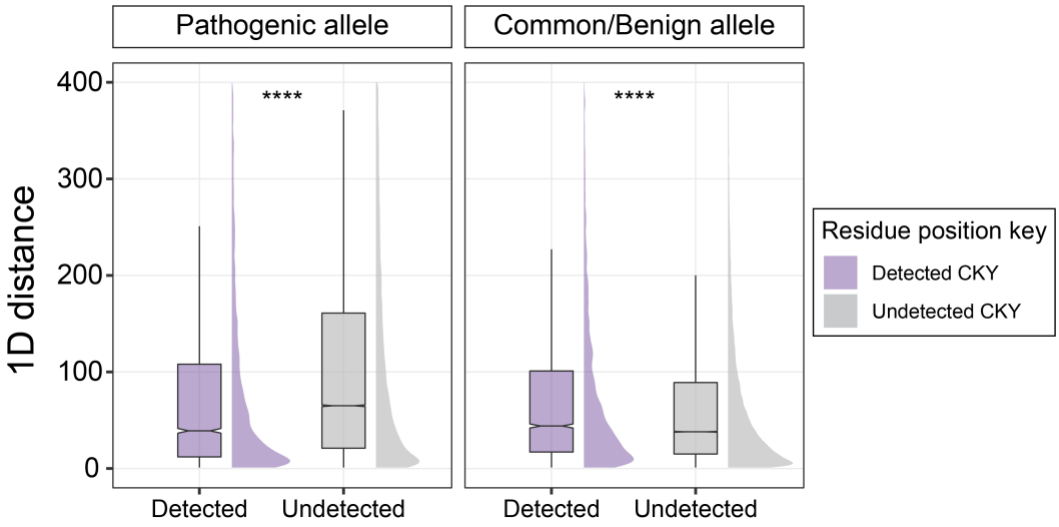

Figure 3-30. Distance to pathogenic and common/benign missense for detected versus undetected positions.

Nearest distances for a given category to a unique CKY reference position counted for 926 OMIM&CpD proteins. Distances of zero were excluded from the analysis and proteins were controlled to contain at least on pathogenic and one common/benign missense positions. Wilcoxon test for mean comparison with FDR adjustment of *p* values. ****$p < 2e\text{-}16$.

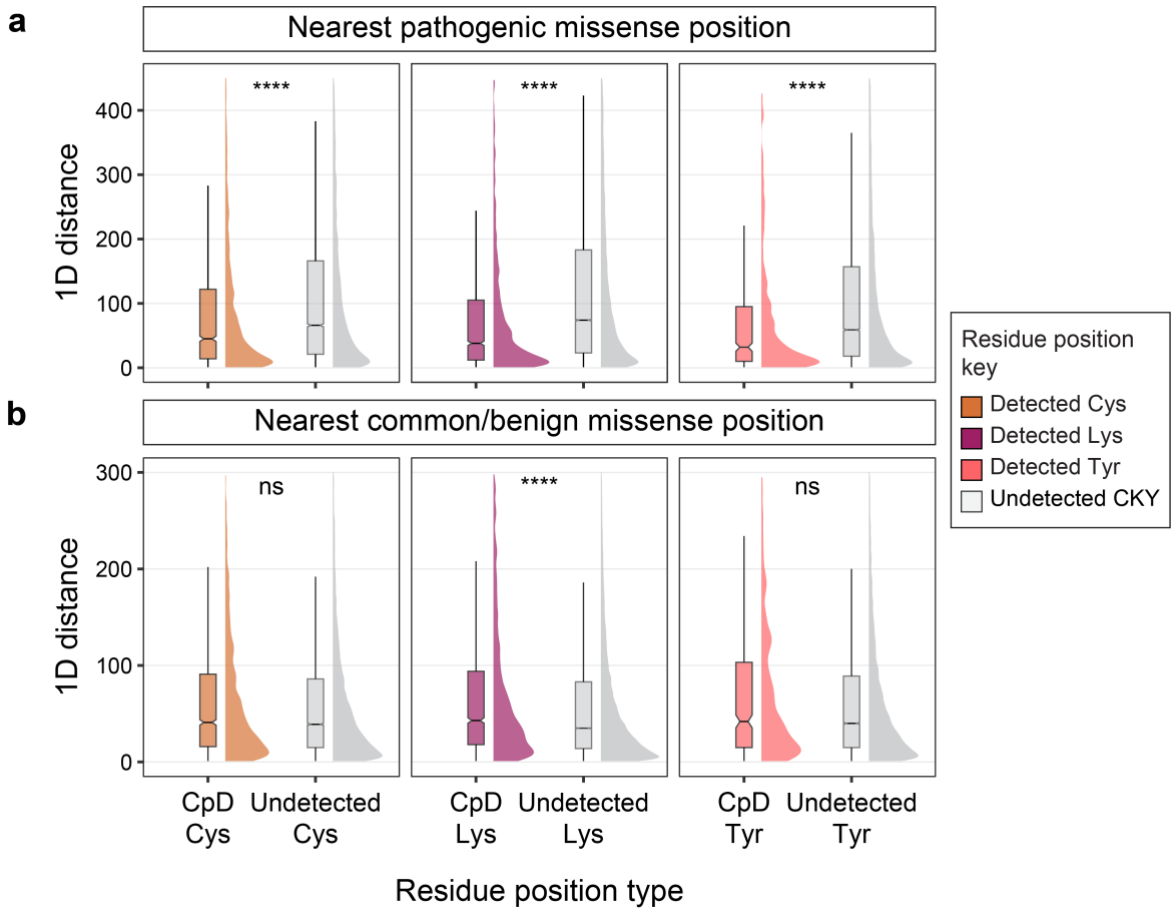

Figure 3-31. Distance to pathogenic and common/benign missense for detected versus undetected CKY specific positions.

Nearest distances for a given category to a unique CKY reference position counted for 926 OMIM&CpD proteins. Distances of zero were excluded from the analysis and proteins were controlled to contain at least on pathogenic and one common/benign missense positions. (**a**) 1D distances between pathogenic missense positions and detected vs undetected cysteine (left panel), lysine (middle panel), and tyrosine (right panel) positions. Wilcoxon test for mean comparison with FDR adjustment of *p* values. ****p < 2e-16. (**b**) 1D distances between common/benign missense positions and detected vs undetected cysteine (left panel), lysine (middle panel), and tyrosine (right

panel) positions. Wilcoxon test for mean comparison with FDR adjustment of *p* values. ****p = 2.8e-11; ns = not significant; cysteine p = 0.084; tyrosine p = 0.150.

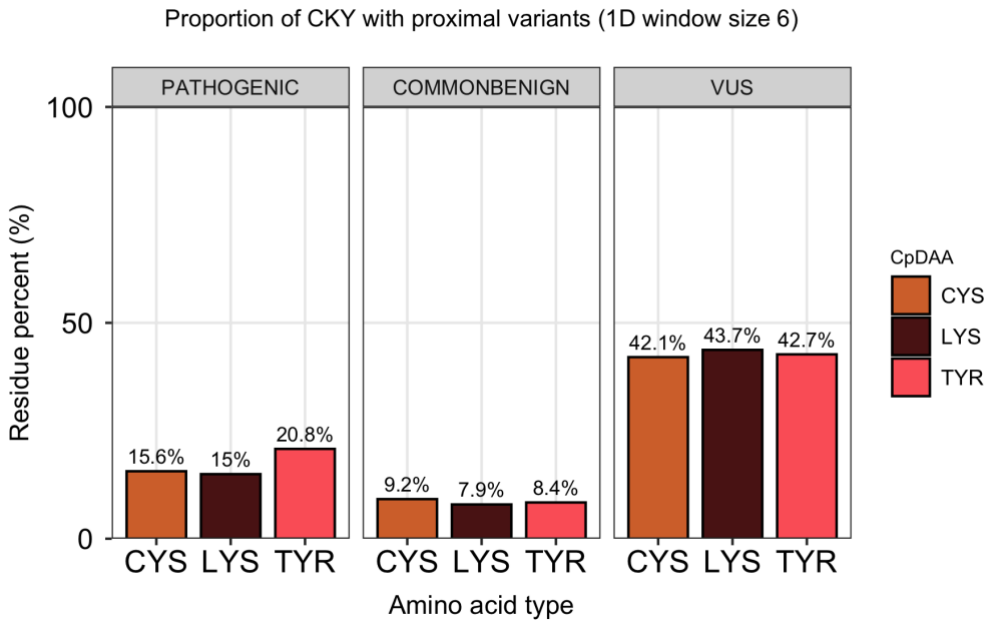

Figure 3-32. Proportion of detected residue 1D windows with pathogenic, common/benign, and VUS missense alleles.

The 1D windows are based on ±6 amino acids from their position in 1D-sequence space. Analysis includes 926 OMIM&CpD proteins that have been filtered to contain at least one pathogenic and one common/benign missense variant.

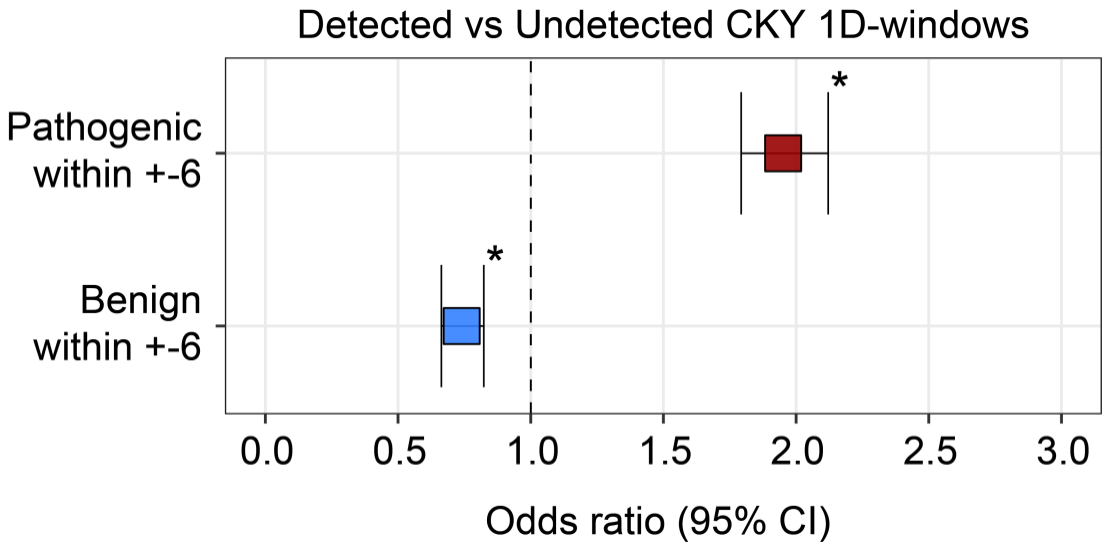

Figure 3-33. Odds of missense in 1D window of detected versus undetected residues.

Significant associations calculated by Fisher's exact test with Bonferroni-corrected two-sided *p* value < 0.05. The x axis corresponds to the odds ratio for ±6 amino acid sized 1D windows. Values greater than 1 indicate a missense category's enrichment for 1D windows of detected residue positions (red for pathogenic enrichment) and values less than 1 indicate a missense category's enrichment for 1D-windows of undetected residue positions (blue for common/benign enrichment). Error bars represent 95% CI. Adjusted significance threshold set as $p < 0.0083$.

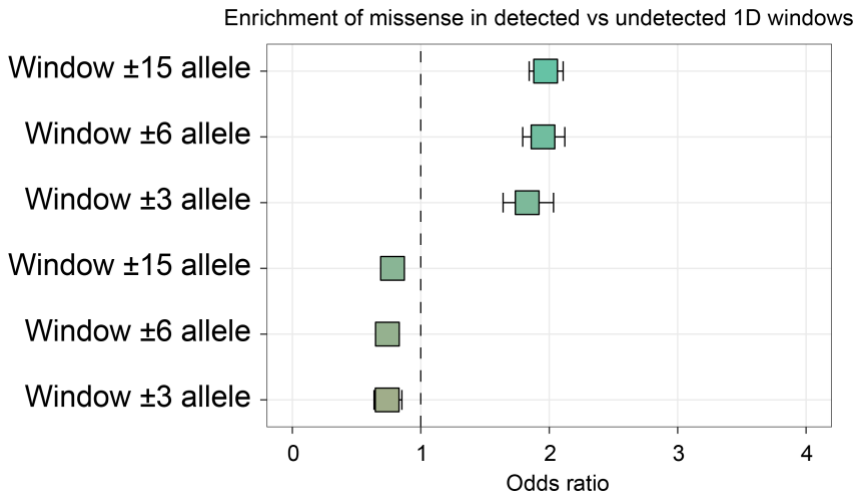

Figure 3-34. Odds of missense in alternative sized 1D windows of detected vs undetected residues.

Significant associations calculated by Fisher's exact test with Bonferroni-corrected two-sided *p* value < 0.05. Detected residues: 1,907 cysteine, 2,092 lysine, 548 tyrosine positions. Undetected residues: 11,596 cysteine, 47,272 lysine, 20,058 tyrosine positions. Windows with pathogenic or common/benign missense variants were analyzed for 926 OMIM&CpD proteins. The x axis corresponds to the odds ratio and the y axis has missense category and1D window size group labels sorted by highest to lowest odds. ±3 amino acid windows, ±6 amino acid windows, and ±15 amino acid window sizes included. Odds greater than 1 indicate a missense category's enrichment in 1D-windows of detected residues while odds of less than 1 indicate a missense category's enrichment for 1D-windows of undetected residue positions.

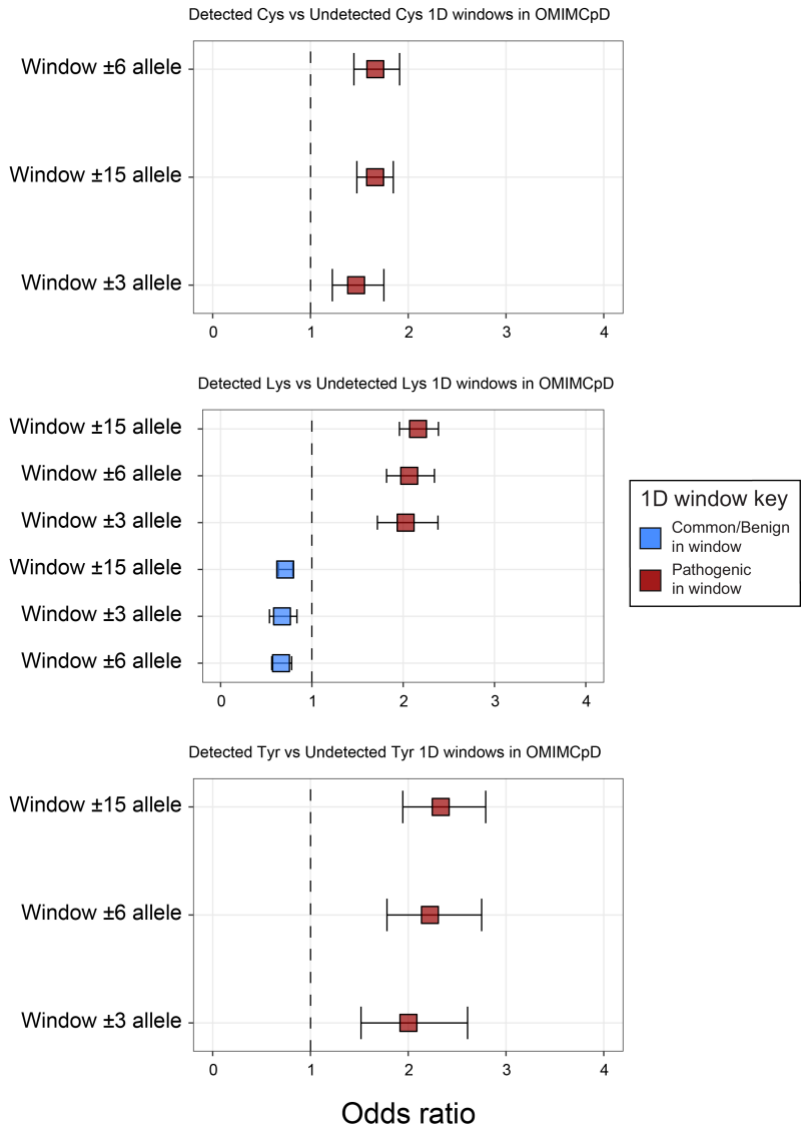

Figure 3-35. Odds of missense in 1D window of CKY specific detected versus undetected residues.

Significant associations calculated by Fisher's exact test. The x axis corresponds to the odds ratio and the y axis has missense category and 1D window size group labels sorted by highest to lowest odds. ±3 amino acid windows, ±6 amino acid windows, and ±15 amino acid window sizes included. Odds greater than 1 indicate a missense category's enrichment in 1D windows of detected residues while odds of less than 1 indicate a missense category's enrichment for 1D-windows of undetected residue

positions. Error bars represent 95% CI. Adjusted significance threshold set as $p < 0.0083$, only significant odds shown. Missense overlapping reference CKY window positions were not included in 1D window enrichment analysis.

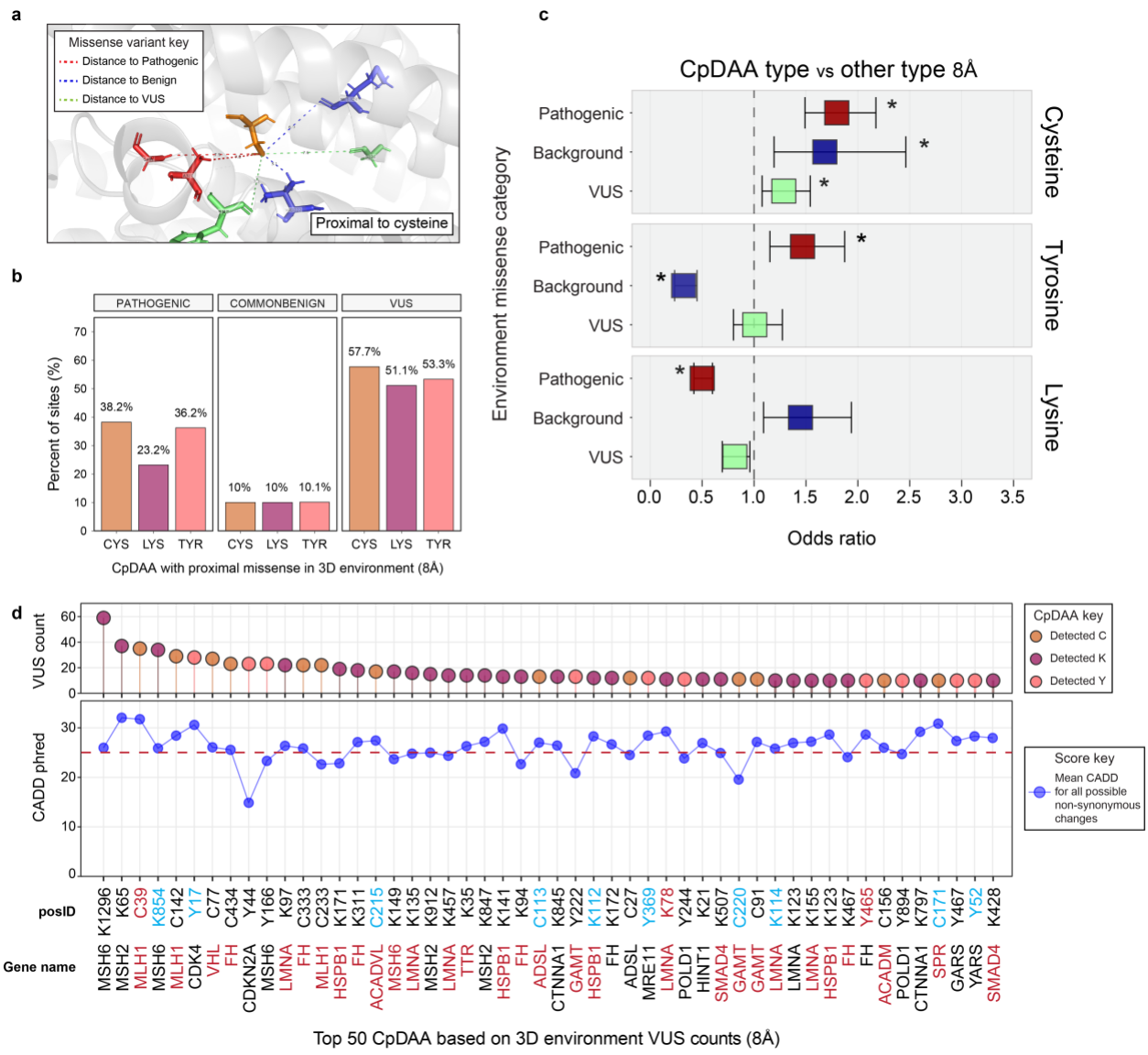

Figure 3-36. 3D environments of CpDAA residues are burdened by VUS missense alleles

a) Cartoon of a CpD-cysteine's 3D protein environment. Measured distances between missense variant impacted positions and CpDAA were based on the terminal atom of CpDAA residues to nearest atom of neighboring residues for environment size 8Å^3. The distances are shown in the cartoon by the dashed lines, with color corresponding to missense categories: pathogenic (red), background (blue), and VUS (green). Residue

positions in the protein structure that overlapped missense alleles are also colored by missense categories.

b) Proportion of CpDAA environments with at least one local pathogenic (left) or VUS (right) missense variant position for $8Å^3$ sized environments. A total of 419 OMIM&CpD proteins are represented by data used to create the figure.

c) Significant associations (Bonferroni-corrected two-sided p value < 0.05) calculated by Fisher's exact test for missense categories in 3D-environments of specific CpDAA types versus all other CpDAA types included in this study for 419 OMIM&CpD proteins. The x axis corresponds to the odds ratio for a particular CpDAA type having a proximal missense variant within 8Å of their terminal atoms. Values greater than 1 indicate enrichment of proximal pathogenic (red), background (blue), or VUS (green) missense for detected cysteine (top panel, n=570), lysine (middle panel, n=1209), or tyrosine (bottom panel, n=287). Error bars represent 95% CI.

d) Top 50 CpDAA positions based on local missense VUS counts within $8Å^3$ environment. The top panel lollipop plot is arranged by decreasing VUS counts within CpDAA environments. The lower panel plot shows mean CADD phred scores for all possible non-synonymous substitutions for each CpDAA codon. The deleterious threshold of 25 is marked by the horizontal dashed red line in the plot. CpDAA residues with scores above this threshold were considered important for the purpose of stratification. CpDAA position IDs for both panels are shown on the x axis with red colored gene names indicating a pathogenic alleles within the CpDAA environment, and red colored amino acid letter and position for pathogenic aliele overlap and blue and common/benign allele overlap.

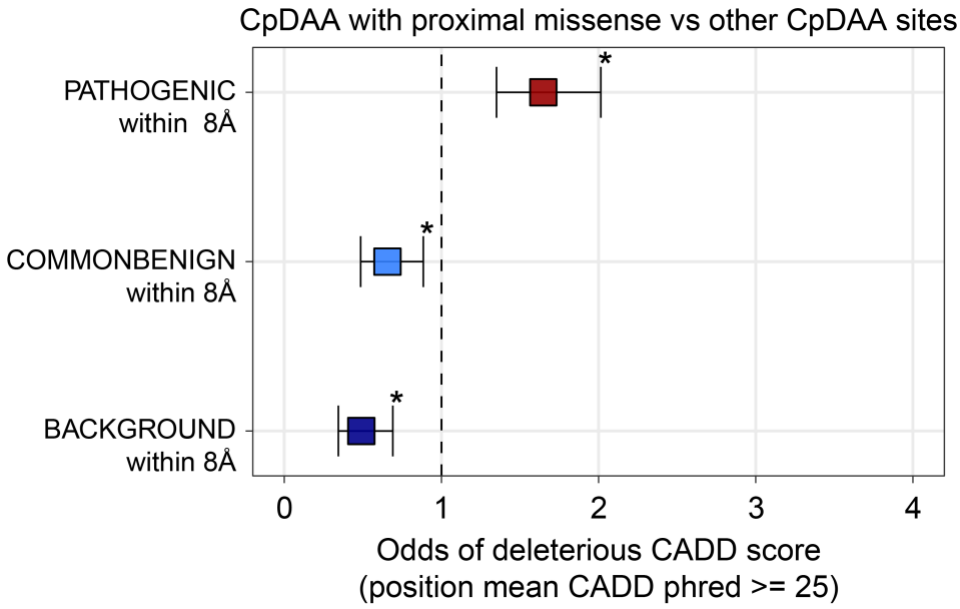

Figure 3-37. Odds of deleterious CADD score based on missense environment of CpDAA residues.

Significant associations calculated by Fisher's exact test for local missense in 3D environment and CADD deleterious scores of all possible substitutions of CpDAA codons. Bonferroni-corrected two-sided p value < 0.05p; x axis corresponds to the odds ratio for $8Å^3$ environment; error bars represent 95% CI.

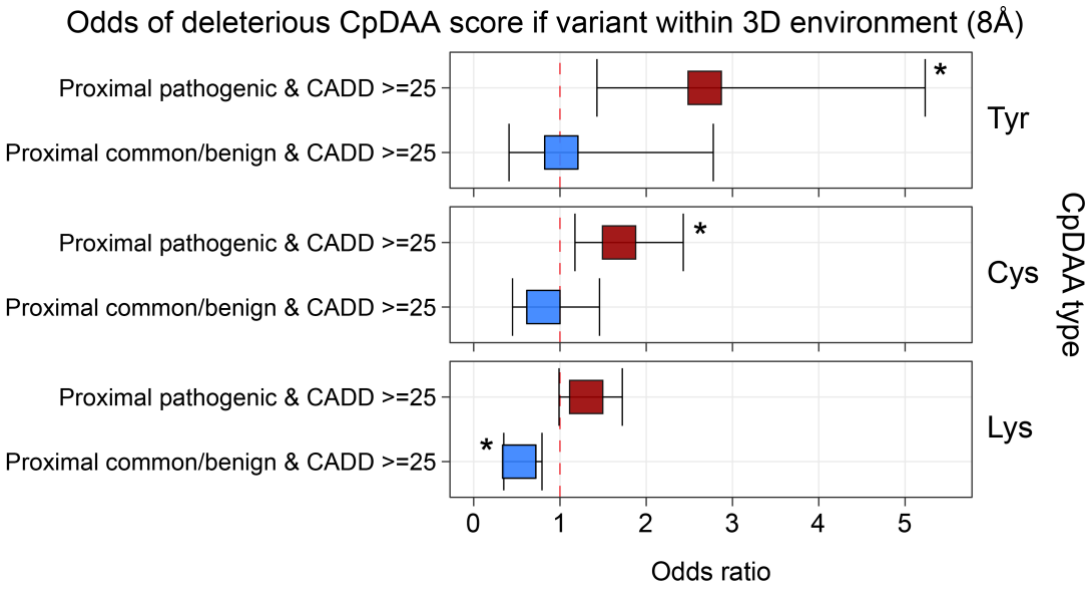

Figure 3-38. Odds of deleterious CADD score based on missense environment of specific CKY detected residues.

CpDAA with missense alleles within 8Å^3 environment were compared to CpDAA with no local missense alleles in 3D environment. Analysis based on 419 OMIM&CpD proteins.

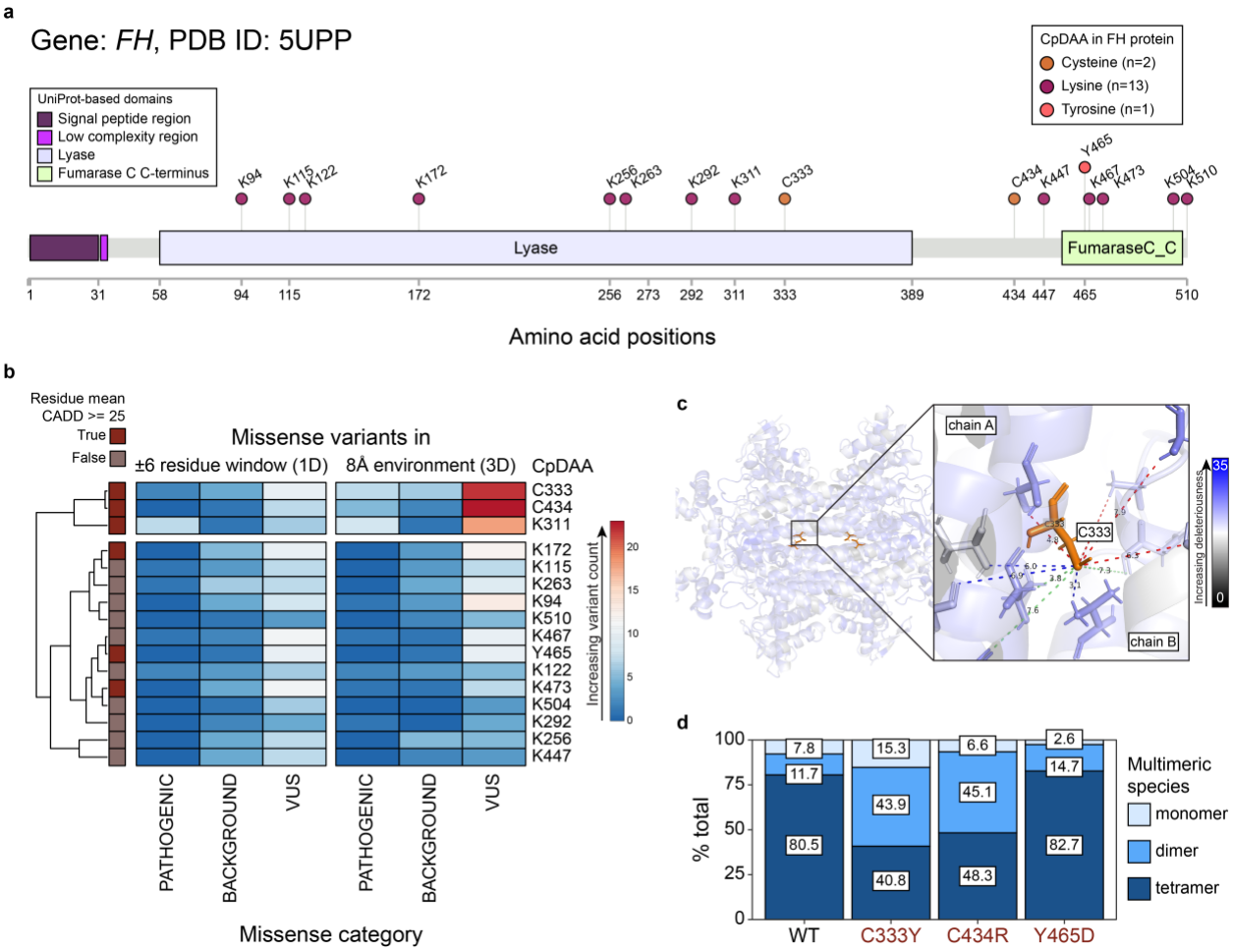

Figure 3-39. Tetramerization of FH protein is disrupted by loss of detected cysteine residues.

a) Schematic of the fumarate hydratase, FH, protein with lollipops marking CpDAA positions. Total count of cysteine, lysine, and tyrosine residues are shown in the figure key in parenthesis. Protein domains shown as colored rectangles and sourced from UniProt for the canonical FH protein sequence (length of protein = 510).

b) Heatmap of proximal missense variant counts in 1D-sequence space (left three columns) and 3D-structural space (right three columns) for detected residue sites (CpDAA) in the FH protein. Count of variants from the pathogenic, background, and VUS missense categories (x axis) were used to compare CpDAA (n=16 residues).

CpDAA rows were clustered based on correlation and average distance. Data preprocessing included row centering and no unit variant scaling. Two groups resulted from row clustering and are shown by vertical space separating the top 3 rows from the other rows. Row annotations for CpDAAs with mean CADD scores greater than or equal to the deleterious score threshold are shown in the outermost left column.

c) Crystal structure of FH (PDB ID: 5UPP) highlighting C333. Distances between positions impacted by missense variants and Cys333 were measured in Angstroms, with dashed line color denoting the missense category (red for pathogenic, blue for background, green for VUS). The VUS positions are specific to the tested inactivating VUS proximal to C333. Protein cartoon color represents CADD phred mean codon scores. Image generated in PyMOL (DeLano et al. 2002).

d) Densitometry was used to quantify the percentage of each multimerization species for *FH* variants in a previous study (Wilde et al. 2022). Missense variants that overlap codons of CpDAAs are shown in red text on the x axis compared to the wildtype (WT) protein.

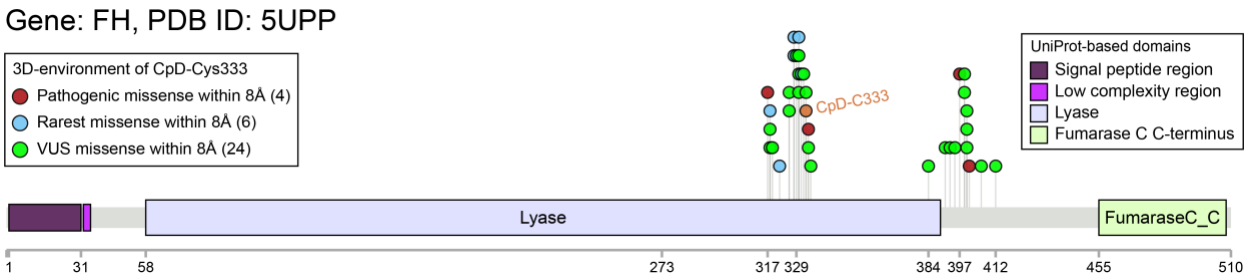

Figure 3-40. Missense in 8Å environment of FH cysteine 333 shown in 1D sequence space.

METHODS

Data availability

Data source	URL	Version
UniProtKB	https://www.uniprot.org/downloads	August 2021
dbNSSFP	https://sites.google.com/site/jpopgen/dbNSFP	4.2a
ClinVar	https://www.ncbi.nlm.nih.gov/clinvar/	June 10,2021
gnomAD constraint		2.1.1
OMIM	https://www.omim.org/downloads	June 24, 2021
Human Protein Atlas	http://www.proteinatlas.org	20.1
HGNC	https://www.genenames.org/download/custom/	September 2020

Software	URL	Version
Python	https://www.python.org/	3.7.4
R	https://www.r-project.org/	3.6.2
Tidyverse	https://doi.org/10.21105/joss.01686	1.3.0
Pandas	https://pandas.pydata.org/	0.25.1
Numpy	https://numpy.org/	1.17.2
SciPy	https://www.scipy.org/	1.3.1
Adobe Illustrator	Adobe, Inc	

Curation and standardization of chemoproteomics datasets

To curate and standardize available residue-level results of reactivity-based protein profiling, we processed all result files through the same quality control pipeline. This process included filtering out peptides with multiple amino acids marked as modified (e.g. MAC*ALRC*Y) and removing peptides that do not meet the minimum number of detections in replicate samples. For datasets from reactivity profiling experiments, reactivity ratios ($R_{10:1}$) assigned to each residue were averaged across peptide replicates for final assignment of one $R_{10:1}$ value. Reactivity labels for Low, Medium, and High were re-assigned based on the following bins: Low $R_{10:1}>5$, Medium $2<R_{10:1}<5$, High $R_{10:1}<2$. All residue identities and positions were checked against the reference set of protein sequences from UniProtKB to prevent analysis errors caused by residue position mis-mapping. To compare cysteine, lysine, and tyrosine detected proteins, we combined experimental datasets for cysteine specific reactivity profiling after confirming the reactivity ratios of residues detected in both studies were significantly correlated (Pearson's R=0.49).

Assigning proteins to subcellular location information

We used the COMPARTMENTS (Binder et al 2014) database to assign each protein to their main subcellular location(s) based on the database provided highest location score values. This resource integrates evidence on protein subcellular localization from manually curated literature, high-throughput screens, automatic text mining, and sequence-based prediction methods. If a protein had multiple subcellular locations with the highest score, more than one main subcellular location was assigned to the protein.

Calculating odds ratios

For each group, an estimate of fold enrichment or odds ratio (OR), along with the 95% confidence interval (CI) was obtained using Fisher's exact test on a 2 × 2 contingency matrix. Evidence for statistical significance of association was determined based on the Bonferroni-adjusted *P*-value cut-off of 0.0167.

Bootstrap analysis of CADD38 PHRED max codon scores

The bootstrapping procedure for calculating the 95% confidence interval of median CADD38 PHRED max codon scores and further characterizing the differences between low, medium, and highly reactive residues was performed as follows: original CADD38 max scores for each sub-group were resampled 20,000 times with replacement, with the median of each bootstrapped sample calculated. This process produced 20,000 samples with 895 low, 412 medium, and 94 high observations for CpD Cys, and 3,401 low, 660 medium, and 302 high observations for CpD Lys.

Calculating amino acid and codon mean abundances

The codon composition of 17,287 human genes with Ensembl CDS sequences was calculated by taking the frequency of occurrence of 61 codons and normalizing by the total number of counted codons per gene.

Curation of disease and population missense variants

Unique nucleotide substitutions were counted once, thereby excluding missense recurrence information from this study. Below are exclusive missense filtering rules-

1. for pathogenic, did not drop anything

2. for common/benign, dropped pathogenic and vus. includes rare and rarest gnomad

3. for rare, dropped all other categories including benign

4. for rarest, dropped all other categories including benign

5. for vus, dropped pathogenic, benign, and common.

Fold enrichment calculations for residue loss-by-missense and gain-by-missense mutations

The fold enrichment of residue gain outcomes is relative to all possible amino acid gains that could result from all possible nonsynonymous single nucleotide alterations at every occurrence of the lost codon type in OMIM genes.

3D distance calculations

Proteins with CpDAAs were cross-referenced with the Protein Data Bank (PDB) downloaded June 23, 2022. All biological assembly files of entries were processed. For each CpD protein associated with a PDB, the SIFTS database (2019 release) was used to map protein sequence residue positions to PDB structure residue positions. The author determined biological unit annotations were extracted from each PDB, as well as the exact 3D coordinates of a CpDAA. Specifically, distances were calculated with respect to locations of the SG atom of cysteine residues, NZ atom of lysine residues, and OH atom of tyrosine residues to all other atoms of neighboring amino acids within 10 Angstroms. The smallest distance between terminal cysteine, lysine, or tyrosine

atoms and atoms of neighboring missense variant positions were stored for statistical analyses.

Multiple PDB structures assigned to a given uniprot identifier used for missene environment counting. At the amino acid level, all proximal amino acids to a detected residue were assigned a distance pair identifier, composed of the protein identifier and protein position of a given residue. The distance pairs for a single environment (in the consistent direction from missense protein position to CpDAA protein position) were all unique.

3.6 Tables

Table 3-1. Curation of chemoproteomics studies

Residue	Chemical probe	Experimental design	Citation	Screen	Detected Proteins	Detected Residues
Cysteine	IA-alkyne	In vitro, profiling in MDA-MB-231, JURKAT, and MCF7 cell lines, soluble fraction only	Weerapana et al. 2010	Reactivity	985	1,432
Cysteine	Compound library		Backus et al. 2016	Ligandability	2,791	5,970
Cysteine	IA-alkyne	In vitro, profiling in JURKAT cell line, grouped soluble and membrane fractions	Desai 2020 PXD022151	Reactivity	1,717	2,610

Lysine	STP-alkyne	In vitro, profiling in MDA-MB-231 breast cancer, Ramos, and JURKAT cell lines, soluble fraction only	Hacker et al. 2017	Reactivity	1,548	4,431
Lysine	Compound library		Hacker et al. 2017	Ligandability	2,374	8,126
Tyrosine	HHS-465	In situ, profiling in HEK293T cell line, SILAC heavy and light soluble proteomes treated with 250 or 25 uM (10:1)	Hahm et al. 2020	Reactivity	1,154	2,445

Table 3-2. Summary of detected residue counts per CpD protein.

	Cysteine (C)	Lysine (K)	Tyrosine (Y)
total	7406	9058	2363
mean residue count	1.63	2.00	0.52
std	1.94	3.78	1.35
min	0	0	0
25%	1	0	0
50%	1	1	0
75%	2	2	0
max residue count	25	105	18

Table 3-3. OMIM phenotypes and gene stats summary for June 23, 2022 release.

	Total phenotypes	Total genes*not exclusive groups

Total single gene disorders and traits(june 23)	5855	4093
molecular (MB) basis known	6114	Note: includes non-single gene phenotypes
unknown MB	1526	
suspected mendelian basis	1756	
known+unknown	9396	
susceptibility to complex	691	497
nondiseases	152	119
somatic cell genetic disease	231	130
susceptibility+nondisease+somatic	1074	
Total penetrant disease phenotypes	8322	Note: took total number of phenotypes with molecular basis known and unknown, then subtracted the number of molecular basis known phenotypes that describe susceptibility to complex disease or infection, non-diseases, and somatic cell genetic disease. The 8322 phenotypes include unknown molecular basis
subtracted unknown basis phenotypes	6796	Note: All penetrant phenotypes for which molecular basis known, but not all phenotypes are associated with severe single gene disorders
Total single gene disorder phenotypes	5855	4093
Post mapping to universal xref file, total single gene disorder phenotypes	5622	3990

Table 3-4. CpD proteins with MOEUF < 0.35 gene constraint scores and no associated monogenic disorder phenotypes as per OMIM June 23, 2021 release.

Gene	MOEUF	LOEUF	Length	UniProt protein name
ARF3	0.342	0.716	181	ADP-ribosylation factor 3
ARF6	0.32	0.722	175	ADP-ribosylation factor 6
ARIH1	0.303	0.166	557	E3 ubiquitin-protein ligase ARIH1 (EC 2.3.2.31) (H7-

				AP2) (HHARI) (Monocyte protein 6) (MOP-6) (Protein ariadne-1 homolog) (ARI-1) (UbcH7-binding protein) (UbcM4-interacting protein) (Ubiquitin-conjugating enzyme E2-binding protein 1)
CFL1	0.303	0.716	166	Cofilin-1 (18 kDa phosphoprotein) (p18) (Cofilin, non-muscle isoform)
COPS2	0.277	0.101	443	COP9 signalosome complex subunit 2 (SGN2) (Signalosome subunit 2) (Alien homolog) (JAB1-containing signalosome subunit 2) (Thyroid receptor-interacting protein 15) (TR-interacting protein 15) (TRIP-15)
CSTF3	0.342	0.303	717	Cleavage stimulation factor subunit 3 (CF-1 77 kDa subunit) (Cleavage stimulation factor 77 kDa subunit) (CSTF 77 kDa subunit) (CstF-77)
CUL1	0.261	0.144	776	Cullin-1 (CUL-1)
DCAF7	0.258	0.348	342	DDB1- and CUL4-associated factor 7 (WD repeat-containing protein 68) (WD repeat-containing protein An11 homolog)
DDX39B	0.261	0.218	428	Spliceosome RNA helicase DDX39B (EC 3.6.4.13) (56 kDa U2AF65-associated protein) (ATP-dependent RNA helicase p47) (DEAD box protein UAP56) (HLA-B-associated transcript 1 protein)
DHX15	0.297	0.105	795	Pre-mRNA-splicing factor ATP-dependent RNA helicase DHX15 (EC 3.6.4.13) (ATP-dependent RNA helicase #46) (DEAH box protein 15)
DYNLL1	0.324	0.759	89	Dynein light chain 1, cytoplasmic (8 kDa dynein light chain) (DLC8) (Dynein light chain LC8-type 1) (Protein inhibitor of neuronal nitric oxide synthase) (PIN)
EIF1AX	0.173	0.386	144	Eukaryotic translation initiation factor 1A, X-chromosomal (eIF-1A X isoform) (Eukaryotic translation initiation factor 4C) (eIF-4C)
EIF4A2	0.332	0.205	407	Eukaryotic initiation factor 4A-II (eIF-4A-II) (eIF4A-II) (EC 3.6.4.13) (ATP-dependent RNA helicase eIF4A-2)
ELAVL1	0.346	0.222	326	ELAV-like protein 1 (Hu-antigen R) (HuR)
ERH	0.249	0.421	104	Enhancer of rudimentary homolog
ETF1	0.246	0.251	437	Eukaryotic peptide chain release factor subunit 1 (Eukaryotic release factor 1) (eRF1) (Protein Cl1) (TB3-1)
GNAQ	0.346	0.302	359	Guanine nucleotide-binding protein G(q) subunit alpha (Guanine nucleotide-binding protein alpha-q)
HNRNPH1	0.346	0.109	449	Heterogeneous nuclear ribonucleoprotein H (hnRNP H) [Cleaved into: Heterogeneous nuclear ribonucleoprotein H, N-terminally processed]
KPNB1	0.271	0.092	876	Importin subunit beta-1 (Importin-90) (Karyopherin subunit beta-1) (Nuclear factor p97) (Pore targeting complex 97 kDa subunit) (PTAC97)

MAGOH	0.291	0.492	146	Protein mago nashi homolog
NEDD8	0.271	0.481	81	NEDD8 (Neddylin) (Neural precursor cell expressed developmentally down-regulated protein 8) (NEDD-8) (Ubiquitin-like protein Nedd8)
NRF1	0.327	0.189	503	Nuclear respiratory factor 1 (NRF-1) (Alpha palindromic-binding protein) (Alpha-pal)
NUDT21	0.187	0.232	227	Cleavage and polyadenylation specificity factor subunit 5 (Cleavage and polyadenylation specificity factor 25 kDa subunit) (CPSF 25 kDa subunit) (Cleavage factor Im complex 25 kDa subunit) (CFIm25) (Nucleoside diphosphate-linked moiety X motif 21) (Nudix motif 21) (Nudix hydrolase 21) (Pre-mRNA cleavage factor Im 68 kDa subunit)
PCBP2	0.282	0.228	365	Poly(rC)-binding protein 2 (Alpha-CP2) (Heterogeneous nuclear ribonucleoprotein E2) (hnRNP E2)
PHF5A	0.161	0.455	110	PHD finger-like domain-containing protein 5A (PHD finger-like domain protein 5A) (Splicing factor 3B-associated 14 kDa protein) (SF3b14b)
POLR2B	0.312	0.344	1174	DNA-directed RNA polymerase II subunit RPB2 (EC 2.7.7.6) (DNA-directed RNA polymerase II 140 kDa polypeptide) (DNA-directed RNA polymerase II subunit B) (RNA polymerase II subunit 2) (RNA polymerase II subunit B2)
PSMC1	0.342	0.138	440	26S proteasome regulatory subunit 4 (P26s4) (26S proteasome AAA-ATPase subunit RPT2) (Proteasome 26S subunit ATPase 1)
PSMC5	0.305	0.355	406	26S proteasome regulatory subunit 8 (26S proteasome AAA-ATPase subunit RPT6) (Proteasome 26S subunit ATPase 5) (Proteasome subunit p45) (Thyroid hormone receptor-interacting protein 1) (TRIP1) (p45/SUG)
PSMD14	0.258	0.298	310	26S proteasome non-ATPase regulatory subunit 14 (EC 3.4.19.-) (26S proteasome regulatory subunit RPN11) (26S proteasome-associated PAD1 homolog 1)
RAB14	0.349	0.214	215	Ras-related protein Rab-14
RAB2A	0.297	0.22	212	Ras-related protein Rab-2A
RACK1	0.334	0.16	317	Receptor of activated protein C kinase 1 (Cell proliferation-inducing gene 21 protein) (Guanine nucleotide-binding protein subunit beta-2-like 1) (Guanine nucleotide-binding protein subunit beta-like protein 12.3) (Human lung cancer oncogene 7 protein) (HLC-7) (Receptor for activated C kinase) (Small ribosomal subunit protein RACK1) [Cleaved into: Receptor of activated protein C kinase 1, N-terminally processed (Guanine nucleotide-binding protein subunit beta-2-like 1, N-terminally processed)]
RAN	0.176	0.262	216	GTP-binding nuclear protein Ran (Androgen receptor-associated protein 24) (GTPase Ran) (Ras-like protein

				TC4) (Ras-related nuclear protein)
RBBP4	0.198	0.253	425	Histone-binding protein RBBP4 (Chromatin assembly factor 1 subunit C) (CAF-1 subunit C) (Chromatin assembly factor I p48 subunit) (CAF-I 48 kDa subunit) (CAF-I p48) (Nucleosome-remodeling factor subunit RBAP48) (Retinoblastoma-binding protein 4) (RBBP-4) (Retinoblastoma-binding protein p48)
RBX1	0.28	0.323	108	E3 ubiquitin-protein ligase RBX1 (EC 2.3.2.27) (EC 2.3.2.32) (E3 ubiquitin-protein transferase RBX1) (Protein ZYP) (RING finger protein 75) (RING-box protein 1) (Rbx1) (Regulator of cullins 1) (ROC1) [Cleaved into: E3 ubiquitin-protein ligase RBX1, N-terminally processed (E3 ubiquitin-protein transferase RBX1, N-terminally processed)]
RHOA	0.182	0.664	193	Transforming protein RhoA (EC 3.6.5.2) (Rho cDNA clone 12) (h12)
RPS18	0.293	0.261	152	40S ribosomal protein S18 (Ke-3) (Ke3) (Small ribosomal subunit protein uS13)
SF3B1	0.224	0.066	1304	Splicing factor 3B subunit 1 (Pre-mRNA-splicing factor SF3b 155 kDa subunit) (SF3b155) (Spliceosome-associated protein 155) (SAP 155)
SKP1	0.286	0.698	163	S-phase kinase-associated protein 1 (Cyclin-A/CDK2-associated protein p19) (p19A) (Organ of Corti protein 2) (OCP-2) (Organ of Corti protein II) (OCP-II) (RNA polymerase II elongation factor-like protein) (SIII) (Transcription elongation factor B polypeptide 1-like) (p19skp1)
SMU1	0.33	0.108	513	WD40 repeat-containing protein SMU1 (Smu-1 suppressor of mec-8 and unc-52 protein homolog) [Cleaved into: WD40 repeat-containing protein SMU1, N-terminally processed]
SNRPD1	0.338	0.564	119	Small nuclear ribonucleoprotein Sm D1 (Sm-D1) (Sm-D autoantigen) (snRNP core protein D1)
SRSF1	0.19	0.242	248	Serine/arginine-rich splicing factor 1 (Alternative-splicing factor 1) (ASF-1) (Splicing factor, arginine/serine-rich 1) (pre-mRNA-splicing factor SF2, P33 subunit)
SRSF3	0.248	0.291	164	Serine/arginine-rich splicing factor 3 (Pre-mRNA-splicing factor SRP20) (Splicing factor, arginine/serine-rich 3)
TNPO2	0.325	0.129	897	Transportin-2 (Karyopherin beta-2b)
TUBA1B	0.092	0.315	451	Tubulin alpha-1B chain (Alpha-tubulin ubiquitous) (Tubulin K-alpha-1) (Tubulin alpha-ubiquitous chain) [Cleaved into: Detyrosinated tubulin alpha-1B chain]
U2AF1	0.253	0.216	240	Splicing factor U2AF 35 kDa subunit (U2 auxiliary factor 35 kDa subunit) (U2 small nuclear RNA auxiliary factor 1) (U2 snRNP auxiliary factor small subunit)

U2AF2	0.311	0.133	475	Splicing factor U2AF 65 kDa subunit (U2 auxiliary factor 65 kDa subunit) (hU2AF(65)) (hU2AF65) (U2 snRNP auxiliary factor large subunit)
UBE2D2	0.146	0.286	147	Ubiquitin-conjugating enzyme E2 D2 (EC 2.3.2.23) ((E3-independent) E2 ubiquitin-conjugating enzyme D2) (EC 2.3.2.24) (E2 ubiquitin-conjugating enzyme D2) (Ubiquitin carrier protein D2) (Ubiquitin-conjugating enzyme E2(17)KB 2) (Ubiquitin-conjugating enzyme E2-17 kDa 2) (Ubiquitin-protein ligase D2) (p53-regulated ubiquitin-conjugating enzyme 1)
UBE2D3	0.143	0.665	147	Ubiquitin-conjugating enzyme E2 D3 (EC 2.3.2.23) ((E3-independent) E2 ubiquitin-conjugating enzyme D3) (EC 2.3.2.24) (E2 ubiquitin-conjugating enzyme D3) (Ubiquitin carrier protein D3) (Ubiquitin-conjugating enzyme E2(17)KB 3) (Ubiquitin-conjugating enzyme E2-17 kDa 3) (Ubiquitin-protein ligase D3)
UBE2H	0.161	0.355	183	Ubiquitin-conjugating enzyme E2 H (EC 2.3.2.23) ((E3-independent) E2 ubiquitin-conjugating enzyme H) (EC 2.3.2.24) (E2 ubiquitin-conjugating enzyme H) (UbcH2) (Ubiquitin carrier protein H) (Ubiquitin-conjugating enzyme E2-20K) (Ubiquitin-protein ligase H)
UBE2I	0.169	0.287	158	SUMO-conjugating enzyme UBC9 (EC 2.3.2.-) (RING-type E3 SUMO transferase UBC9) (SUMO-protein ligase) (Ubiquitin carrier protein 9) (Ubiquitin carrier protein I) (Ubiquitin-conjugating enzyme E2 I) (Ubiquitin-protein ligase I) (p18)
UBE2K	0.185	0.231	200	Ubiquitin-conjugating enzyme E2 K (EC 2.3.2.23) (E2 ubiquitin-conjugating enzyme K) (Huntingtin-interacting protein 2) (HIP-2) (Ubiquitin carrier protein) (Ubiquitin-conjugating enzyme E2-25 kDa) (Ubiquitin-conjugating enzyme E2(25K)) (Ubiquitin-conjugating enzyme E2-25K) (Ubiquitin-protein ligase)
UBE2L3	0.19	0.499	154	Ubiquitin-conjugating enzyme E2 L3 (EC 2.3.2.23) (E2 ubiquitin-conjugating enzyme L3) (L-UBC) (UbcH7) (Ubiquitin carrier protein L3) (Ubiquitin-conjugating enzyme E2-F1) (Ubiquitin-protein ligase L3)
UBE2N	0.349	0.431	152	Ubiquitin-conjugating enzyme E2 N (EC 2.3.2.23) (Bendless-like ubiquitin-conjugating enzyme) (E2 ubiquitin-conjugating enzyme N) (Ubc13) (UbcH13) (Ubiquitin carrier protein N) (Ubiquitin-protein ligase N)
XPO1	0.317	0.051	1071	Exportin-1 (Exp1) (Chromosome region maintenance 1 protein homolog)
YWHAZ	0.316	0.357	245	14-3-3 protein zeta/delta (Protein kinase C inhibitor protein 1) (KCIP-1)

Table 3-5. Codon abundance observed mean difference for OMIM versus all other genes.

Codon (with amino acid prefix)	GroupMember	NongroupMember	Tobs	welch.ttest.Pvalue
A.GCA	0.015818078	0.015565689	0.000252389	0.116105188
A.GCC	0.030233278	0.029177103	0.001056175	0.001030102
A.GCG	0.00888759	0.008980299	-9.27E-05	0.648308956
A.GCT	0.018561677	0.017933141	0.000628535	0.000438607
C.TGC	0.012326069	0.01441835	-0.002092281	4.29E-18
C.TGT	0.009217765	0.011378128	-0.002160364	3.04E-38
D.GAC	0.026577848	0.025022589	0.001555259	1.12E-11
D.GAT	0.021712788	0.020628338	0.00108445	5.53E-06
E.GAA	0.027927978	0.028611736	-0.000683758	0.06168943
E.GAG	0.040444785	0.0400177	0.000427084	0.276093191
F.TTC	0.021471272	0.021261253	0.000210019	0.353578083
F.TTT	0.017247207	0.017169572	7.76E-05	0.706498873
G.GGA	0.016689508	0.016087812	0.000601697	0.00577626
G.GGC	0.025123319	0.023324671	0.001798649	1.61E-08
G.GGG	0.016480335	0.016509348	-2.9E-05	0.874174155
G.GGT	0.011034367	0.010124617	0.00090975	1.39E-08
H.CAC	0.014950297	0.015347929	-0.000397632	0.023138822
H.CAT	0.009833071	0.010704104	-0.000871033	8.09E-10
I.ATA	0.006929773	0.007395606	-0.000465833	0.000273677
I.ATC	0.022322743	0.020823251	0.001499492	2.32E-10
I.ATT	0.016015502	0.015464589	0.000550913	0.008299954
K.AAA	0.023099882	0.024990438	-0.001890556	8.42E-09
K.AAG	0.033369014	0.033212323	0.000156691	0.608469958
L.CTA	0.006804516	0.007011695	-0.000207179	0.034986198
L.CTC	0.019864737	0.020201891	-0.000337153	0.100990032
L.CTG	0.041962677	0.041145826	0.000816851	0.054355342
L.CTT	0.012456008	0.012874922	-0.000418913	0.011231601
L.TTA	0.007022954	0.007226134	-0.00020318	0.154366789
L.TTG	0.012632627	0.012705077	-7.25E-05	0.613123505
M.ATG	0.023039714	0.022855541	0.000184173	0.309969419
N.AAC	0.019616075	0.018784548	0.000831527	2.11E-06
N.AAT	0.016229514	0.01601092	0.000218594	0.300665461
P.CCA	0.015985612	0.016191817	-0.000206205	0.311561332
P.CCC	0.020072384	0.020629505	-0.000557121	0.030578142
P.CCG	0.007862578	0.007983874	-0.000121296	0.48971093
P.CCT	0.016371432	0.016586644	-0.000215212	0.278741947

Q.CAA	0.011163613	0.012006024	-0.000842411	5.97E-07
Q.CAG	0.033892757	0.03381556	7.72E-05	0.785550128
R.AGA	0.010625236	0.012177413	-0.001552177	7.45E-20
R.AGG	0.010843039	0.012199651	-0.001356612	3.13E-25
R.CGA	0.006405115	0.00612311	0.000282006	0.00169777
R.CGC	0.011969832	0.011431021	0.000538811	0.010766364
R.CGG	0.012358729	0.011911709	0.00044702	0.010747574
R.CGT	0.004699804	0.00437708	0.000322724	4.53E-05
S.AGC	0.019264267	0.019799831	-0.000535564	0.010934016
S.AGT	0.010844059	0.011523978	-0.000679919	2.65E-06
S.TCA	0.010678358	0.011446729	-0.000768371	8.47E-08
S.TCC	0.017203916	0.017955078	-0.000751162	3.12E-05
S.TCG	0.004780501	0.004821665	-4.12E-05	0.665152736
S.TCT	0.013344461	0.014308389	-0.000963928	2.21E-09
T.ACA	0.01409138	0.014018441	7.29E-05	0.644447729
T.ACC	0.01887371	0.018516992	0.000356718	0.054240211
T.ACG	0.006413223	0.006095558	0.000317666	0.001311989
T.ACT	0.012151812	0.012579801	-0.000427989	0.00378447
V.GTA	0.006960422	0.006796957	0.000163465	0.144524312
V.GTC	0.014797151	0.014463733	0.000333417	0.025342254
V.GTG	0.029695472	0.028170858	0.001524614	2.64E-09
V.GTT	0.010755507	0.010429566	0.000325941	0.040016336
W.TGG	0.012891861	0.012662244	0.000229617	0.141953355
Y.TAC	0.016715644	0.015637192	0.001078452	1.8E-09
Y.TAT	0.012354819	0.012346381	8.44E-06	0.959514637

Table 3-6. Comparison of Relative Synonymous Codon Usage (RSCU) between gene sets representing the human protein-coding genes and OMIM.

Table only shows amino acid with RSCU differences between the compared gene sets. The preferred codon used to encode an amino acid in each gene set is marked by * following the RSCU value. DNA codons with a G or C in the 3rd position, referred to as GC3 codons, were marked by underlining the 3rd position nucleotide letter in the codon column. CpG dinucleotide-containing codons are marked by bolded font of CG in the codon column.

Amino acid	Codon	Exome RSCU	OMIM RSCU
ARG (R)	AGA	1.284*	1.233
ARG (R)	AGG	1.241	1.172
ARG (R)	**CG**G	1.227	1.247*

ARG (R)	**CG**T	0.476	0.511
ARG (R)	**CG**C	1.114	1.146
ARG (R)	**CG**A	0.658	0.691

Table 3-7. Observed mean abundance differences for 61 codons between OMIM and all other genes.

Amino	GroupMember	NongroupMember	Tobs	welch.ttest.Pvalue
A	0.0735016025641026	0.0716615963966625	0.00184000616744008	9.71643835599188E-05
C	0.021522702991453	0.0258004873366315	-0.00427778434517848	1.22032010430963E-36
D	0.0483162393162393	0.0456695709960865	0.00264666832015278	8.34208756936076E-20
E	0.0683896901709402	0.0686539171527727	-0.000264226981832483	0.585544973691577
F	0.0387406517094017	0.0384315144355017	0.00030913727389998	0.291296370919043
G	0.0693274572649573	0.0660640183120431	0.00326343895291414	1.24899547748334E-09
H	0.0247865918803419	0.0260569297792217	-0.00127033789887986	4.06357543060537E-09
I	0.04528125	0.0436844864505649	0.00159676354943514	3.90912040916302E-06
K	0.0564791666666667	0.0582150188289153	-0.00173585216224864	0.000235719045353498
L	0.100774572649573	0.101156243077605	-0.000381670428032013	0.462592331251908
M	0.023045405982906	0.0228509931329838	0.000194412849922155	0.2841431883955
N	0.035860844017094	0.0348010042088164	0.00105983980827765	5.25115657721252E-05
P	0.0602532051282051	0.0613825592557041	-0.00112935412749892	0.0373623403465337
Q	0.0450625	0.0458252972015063	-0.000762797201506314	0.0173387154641126
R	0.0568832799145299	0.058198774274533	-0.00131549436000306	0.000520664633646661
S	0.0760854700854701	0.0798317211843757	-0.0037462510989056	9.51290964574544E-17
T	0.0515082799145299	0.0511987004356494	0.000309579478880502	0.265336721152646
V	0.0622235576923077	0.0598656132319279	0.00235794446037976	1.25713906346819E-12
W	0.0128675213675214	0.0126414383814517	0.000226082986069695	0.146599565032661
Y	0.0290819978632479	0.0279920254005759	0.00108997246267192	6.15861499146455E-06

Table 3-8. Relative residue mutability for five missense categories.

Rm values are sorted by ascending with the lowest for each Type and Category marked with bolded text.

Type	Category	Amino acid	*Rm*	Observed	Expected	OE
Loss	**Background**	**W**	**1**	**7899**	**14605.5952354**	**0.540820135892508**
Loss	Background	L	1.02458705429823	66771	120499.75483949	0.554117309939272

Loss	Background	F	1.03242343751397	24941	44668.68364621	0.558355383774918
Loss	Background	K	1.24562493953374	46321	68760.30250609	0.673659049069736
Loss	Background	C	1.33560768601226	18465	25563.33723983	0.722323530248228
Loss	Background	Q	1.35556443194315	42143	57484.72130102	0.733116540294546
Loss	Background	Y	1.43668458527564	25936	33380.18293818	0.776987952643441
Loss	Background	E	1.50807541457727	70569	86524.29123156	0.815597550647832
Loss	Background	S	1.68579065431948	88881	97488.28656684	0.911709530755383
Loss	Background	N	1.8235448943692	44300	44919.44828449	0.986209797578839
Loss	Background	G	1.90274773455311	84178	81802.11576056	1.02904428837017
Loss	Background	D	1.98811190449929	64436	59928.70513305	1.07521095036082
Loss	Background	V	1.99206479651248	80426	74651.7779917	1.07734875395656
Loss	Background	T	2.07701897671018	72548	64585.06885025	1.12329368523572
Loss	Background	I	2.09013092948986	62455	55251.09223335	1.13038489331984
Loss	Background	H	2.16308492400494	35572	30407.5801575	1.16983988254739
Loss	Background	A	2.20238650831213	101653	85344.16020507	1.19109497071319
Loss	Background	P	2.32028554231577	94199	75067.50914052	1.25485714230464
Loss	Background	M	2.98411319171508	42747	26487.2881221	1.61386850186197
Loss	Background	R	3.86857053992533	139799	66819.11075918	2.09220084511217
Loss	**Common/Benign**	**W**	**1**	**127**	**420.760428**	**0.301834468140621**
Loss	Common/Benign	F	1.16372622231968	452	1286.8229022	0.351252685375155
Loss	Common/Benign	L	1.3065704288454	1369	3471.3770718	0.394367990478815
Loss	Common/Benign	C	1.36314049302839	303	736.4328906	0.411442785714166
Loss	Common/Benign	Y	1.52626713843703	443	961.6218876	0.460680029970649
Loss	Common/Benign	K	1.72272115133888	1030	1980.8582838	0.519976622468968
Loss	Common/Benign	Q	2.16266276877702	1081	1656.0294564	0.652766166581335
Loss	Common/Benign	E	2.36989351558272	1783	2492.6062392	0.715315548825816
Loss	Common/Benign	S	2.83712312235402	2405	2808.4588488	0.856341548685184
Loss	Common/Benign	D	2.96489371587576	1545	1726.436151	0.894907117824828
Loss	Common/Benign	H	3.13536575223695	829	875.98665	0.946361454252756
Loss	Common/Benign	G	3.21386203514255	2286	2356.5690192	0.970054338054586
Loss	Common/Benign	N	3.35135761662512	1309	1294.0469718	1.01155524376306
Loss	Common/Benign	I	3.51980237267656	1691	1591.682697	1.06239767711692
Loss	Common/Benign	P	4.26360281948155	2783	2162.5573464	1.28690228938106
Loss	Common/Benign	T	4.29319508417217	2411	1860.577455	1.29583425485503
Loss	Common/Benign	V	4.53075322853882	2941	2150.580894	1.36753749101242
Loss	Common/Benign	A	4.814761142192	3573	2458.6088274	1.45326086857765
Loss	Common/Benign	M	4.897643209724	1128	763.050222	1.47827753334957
Loss	Common/Benign	R	9.45075036997899	5491	1924.9361076	2.85256221145238
Loss	**Pathogenic**	**K**	**1**	**767**	**2023.10300306**	**0.379120587948261**
Loss	Pathogenic	Q	1.18835160182245	762	1691.34672268	0.450528557972184
Loss	Pathogenic	E	1.58006228891757	1525	2545.76473704	0.599034143969305
Loss	Pathogenic	S	1.58535759964393	1724	2868.35336856	0.601041705285252
Loss	Pathogenic	V	1.71726886636444	1430	2196.4451978	0.651051982281331
Loss	Pathogenic	F	1.8303498110023	912	1314.26629514	0.693923296498181
Loss	Pathogenic	I	1.83998635115383	1134	1625.6276739	0.697576707266216

Loss	Pathogenic	L	1.86216046059583	2503	3545.40929866	0.705983368675097
Loss	Pathogenic	T	1.86694950643269	1345	1900.2570085	0.707798994548479
Loss	Pathogenic	N	1.98577976979027	995	1321.64442866	0.75284999385865
Loss	Pathogenic	A	1.98636993977138	1891	2511.04228038	0.753073739448876
Loss	Pathogenic	P	2.1352951993064	1788	2208.67706568	0.80953437140414
Loss	Pathogenic	D	2.67469973755936	1788	1763.2549437	1.01403373708857
Loss	Pathogenic	H	2.71826301129019	922	894.668355	1.03054947103835
Loss	Pathogenic	Y	3.00527202557198	1119	982.12989012	1.13936049727931
Loss	Pathogenic	W	4.33339077935527	706	429.7337636	1.64287766007874
Loss	Pathogenic	G	5.6187692319722	5127	2406.82632304	2.1301910947709
Loss	Pathogenic	M	5.638712035416	1666	779.3233914	2.13775182213785
Loss	Pathogenic	C	6.60702216486968	1884	752.13840622	2.50485812773259
Loss	Pathogenic	R	7.69843180393801	5738	1965.98820412	2.91863399178857
Loss	**Rare**	**W**	**1**	**936**	**2493.1198076**	**0.375433221117857**
Loss	Rare	F	1.08013923844022	3092	7624.77517574	0.405520153543401
Loss	Rare	L	1.15679015220359	8933	20568.85192006	0.434297452999212
Loss	Rare	C	1.39663166025717	2288	4363.56488002	0.524341922925531
Loss	Rare	K	1.44717832889924	6377	11737.12330046	0.543318821550599
Loss	Rare	Q	1.61513172476652	5950	9812.42427988	0.606374105958733
Loss	Rare	Y	1.67120892305974	3575	5697.87084492	0.627427349145222
Loss	Rare	E	1.99750699476625	11076	14769.36891864	0.749930485250544
Loss	Rare	S	2.19766510283061	13730	16640.88141096	0.825076488494003
Loss	Rare	N	2.44106279867774	7027	7667.57975006	0.916456069458556
Loss	Rare	D	2.57360122599779	9884	10229.6030667	0.966215398149218
Loss	Rare	H	2.59202341489553	5051	5190.458805	0.973131699867137
Loss	Rare	I	2.71523024192045	9614	9431.1522549	1.01938763580081
Loss	Rare	G	2.72399987309091	14280	13963.31144464	1.02268004667915
Loss	Rare	V	3.14105501253539	15027	12742.7758598	1.17925640106455
Loss	Rare	T	3.16723908914557	13109	11024.4267235	1.18908677328831
Loss	Rare	P	3.44045312646278	16551	12813.73959288	1.29166039937292
Loss	Rare	A	3.55659049731835	19452	14567.92502058	1.33526222660539
Loss	Rare	M	3.73268279540886	6336	4521.2797974	1.40137312529155
Loss	Rare	R	8.16841311384886	34978	11405.76921892	3.06669364675362
Loss	**Rarest**	**W**	**1**	**6836**	**11691.7149998**	**0.584687533019488**
Loss	Rarest	L	1.00124679856287	56469	96459.52584763	0.585416520595383
Loss	Rarest	F	1.02345072783912	21397	35757.08556827	0.598398881227255
Loss	Rarest	K	1.20916426204027	38914	55042.32092183	0.706983269387657
Loss	Rarest	Q	1.30502954679678	35112	46016.26756474	0.763034506234152
Loss	Rarest	C	1.32674064294844	15874	20463.33946921	0.77572871348221
Loss	Rarest	Y	1.40290872740299	21918	26720.69020566	0.820263242876762
Loss	Rarest	E	1.4250504207381	57710	69262.31607372	0.833209214929744
Loss	Rarest	S	1.59431410264499	72746	78038.94630708	0.93217557953368
Loss	Rarest	N	1.71060911699072	35964	35957.82156263	1.00017182457395
Loss	Rarest	G	1.7659420707371	67612	65482.23529672	1.0325243127946
Loss	Rarest	V	1.78757850729879	62458	59758.4212379	1.04517486751119

Loss	Rarest	T	1.8865712893364	57028	51700.06467175	1.10305471302749
Loss	Rarest	D	1.88979866649938	53007	47972.66591535	1.10494172021904
Loss	Rarest	A	1.96843295181527	78628	68317.62635709	1.15091820651114
Loss	Rarest	I	1.97798036926169	51150	44228.25728145	1.15650046246459
Loss	Rarest	H	2.08629058628418	29692	24341.1347025	1.21982809605628
Loss	Rarest	P	2.13080658489013	74865	60091.21220124	1.24585604546109
Loss	Rarest	M	2.84606762978639	35283	21202.9581027	1.66406026126642
Loss	Rarest	R	3.17612035156821	99330	53488.40543266	1.8570379729314
Loss	**VUS**	**F**	**1**	**4656**	**8549.16871166**	**0.544614354568743**
Loss	VUS	L	1.00380699908798	12608	23062.52725054	0.546687700919889
Loss	VUS	W	1.03652055920503	1578	2795.3744684	0.564503975348679
Loss	VUS	K	1.19684647742834	8578	13160.07947414	0.651819771822507
Loss	VUS	Q	1.23083482950172	7375	11002.03857892	0.670330316249805
Loss	VUS	Y	1.4758489754416	5135	6388.65514428	0.803768537201066
Loss	VUS	C	1.48691829303114	3962	4892.58391018	0.809797046455609
Loss	VUS	S	1.52338158568921	15480	18658.34721864	0.829655479052036
Loss	VUS	E	1.52370915320995	13742	16559.94094776	0.829833877025922
Loss	VUS	N	1.77184004385691	8296	8597.16272054	0.964969521884182
Loss	VUS	G	1.86429012949928	15896	15656.16068176	1.01531916560613
Loss	VUS	H	1.94068053287075	6151	5819.726745	1.05692247583353
Loss	VUS	D	1.96874640136177	12298	11469.7942503	1.07220755068717
Loss	VUS	T	1.97594567043889	13302	12360.9787615	1.07612837596898
Loss	VUS	V	2.00186066792535	15577	14287.6528382	1.09024205559872
Loss	VUS	I	2.02498752442283	11662	10574.5428441	1.1028372736233
Loss	VUS	P	2.13007851833391	16667	14367.21989592	1.16007133744317
Loss	VUS	A	2.15439432799684	19165	16334.07489522	1.17331407642856
Loss	VUS	M	2.82808521676388	7808	5069.4194766	1.54021580499326
Loss	VUS	R	4.66027180248486	32458	12788.55351028	2.5380509198252
Gain	**Background**	**P**	**1**	**42309**	**71975.9524886368**	**0.587821328334342**
Gain	Background	D	1.1949725133019	41660	59308.3729663761	0.702430330092152
Gain	Background	G	1.26819797689863	59692	80072.563849831	0.745473819371477
Gain	Background	A	1.32550198312072	61160	78494.9568535677	0.779158336427829
Gain	Background	Y	1.45003022335691	28531	33472.997069381	0.852358692018599
Gain	Background	E	1.4893035851592	45732	52238.6109131446	0.875444411721381
Gain	Background	L	1.53936758063824	93745	103600.162729644	0.904873096045591
Gain	Background	R	1.696479995585	97996	98268.4862782774	0.997227124497412
Gain	Background	N	1.69679288141033	52865	53002.2203390363	0.997411045458878
Gain	Background	H	1.69918466443412	55353	55418.5609038705	0.99881698653301
Gain	Background	I	1.80549595693097	65503	61719.0639513747	1.06130903170545
Gain	Background	S	1.81769131768749	107417	100532.74626159	1.06847772486486
Gain	Background	K	1.8784046788458	53011	48009.976751848	1.10416633346858
Gain	Background	T	1.93909695266829	87595	76848.3333685016	1.13984254648655
Gain	Background	Q	1.96755225337904	54744	47333.0960110006	1.1565691791485
Gain	Background	F	2.14764740214396	40452	32042.8899142317	1.26243294872206
Gain	Background	W	2.29091500831114	23124	17171.5161820341	1.34664870328654

Gain	Background	V	2.30824664183704	107266	79055.9448621112	1.35683660712793
Gain	Background	C	2.3765896134352	52135	37318.9867153237	1.39701006347508
Gain	Background	M	2.63691443717587	43949	28353.5615902176	1.55003454716472
Gain	**Common/Benign**	**P**	**1**	**936**	**2367.50901180857**	**0.395352243785116**
Gain	Common/Benign	G	1.29971409404549	1345	2617.52144188937	0.513844883360022
Gain	Common/Benign	D	1.57825710647916	1042	1669.95880316079	0.62396748831634
Gain	Common/Benign	Y	1.69638750834508	540	805.164254639083	0.670670607753267
Gain	Common/Benign	A	1.8309783802281	1693	2338.78087543283	0.723881410945214
Gain	Common/Benign	E	1.86396595208071	1075	1458.76817899316	0.736923121494166
Gain	Common/Benign	L	2.07804643070937	2623	3192.70531829885	0.821560319070598
Gain	Common/Benign	R	2.31086674018385	2385	2610.533516825	0.913606350820081
Gain	Common/Benign	F	2.50831104317921	806	812.773328598064	0.991666399031884
Gain	Common/Benign	K	2.52647824244992	1380	1381.59042883778	0.998848842026853
Gain	Common/Benign	S	2.73315151101766	3217	2977.16665186895	1.08055758248551
Gain	Common/Benign	N	2.87673044664419	1474	1296.02716860516	1.13732183684574
Gain	Common/Benign	T	3.10245696982575	2758	2248.55899849063	1.22656332426738
Gain	Common/Benign	H	3.31215669716525	2019	1541.84684364734	1.30946858199218
Gain	Common/Benign	I	3.36803206353875	2271	1705.51957737725	1.33155903346026
Gain	Common/Benign	C	3.39869690011482	1398	1040.42439847287	1.34368244540591
Gain	Common/Benign	W	3.59451488933187	826	581.240078131936	1.42109952681636
Gain	Common/Benign	V	3.80264365705892	3228	2147.15644144544	1.50338370213348
Gain	Common/Benign	Q	3.91705750052304	2122	1370.25446151114	1.5486174718671
Gain	Common/Benign	M	5.70622787254585	1842	816.500221965728	2.25596999296017
Gain	**Pathogenic**	**A**	**1**	**891**	**2191.94354069475**	**0.406488572108747**
Gain	Pathogenic	G	1.63748308155568	1602	2406.78529682107	0.6656181596738
Gain	Pathogenic	I	1.6696853920642	1119	1648.72072978269	0.678708030891011
Gain	Pathogenic	L	1.90054033564896	2330	3015.99411213818	0.772547927273026
Gain	Pathogenic	S	2.07208151247617	2724	3234.08869946757	0.842277455299372
Gain	Pathogenic	V	2.25859485299834	2353	2562.92119456825	0.918092996767459
Gain	Pathogenic	E	2.47457161425803	1411	1402.7447321413	1.00588508206058
Gain	Pathogenic	N	2.48522199847278	1175	1163.11950018209	1.01021434153245
Gain	Pathogenic	D	2.54822819520921	1698	1639.2720296214	1.03582564047784
Gain	Pathogenic	H	2.60282979533626	1649	1558.57083644057	1.05802056694834
Gain	Pathogenic	T	2.60412835055352	1720	1624.86663429354	1.05854841480441
Gain	Pathogenic	F	2.68559175269994	1149	1052.52323927784	1.09166235682203
Gain	Pathogenic	Q	2.6956955597498	1477	1347.91129186105	1.09576943892259
Gain	Pathogenic	Y	2.74042439989585	1247	1119.43862402663	1.11395120128563
Gain	Pathogenic	M	2.93090636230914	838	703.386023482076	1.19137994219948
Gain	Pathogenic	R	3.09992776622319	4206	3337.86950451778	1.26008521133232
Gain	Pathogenic	P	3.3078987471511	2808	2088.31763236849	1.34462303840976
Gain	Pathogenic	K	3.4489917474871	1602	1142.67313261996	1.40197573065088
Gain	Pathogenic	W	3.58594356479759	1324	908.314389275246	1.45764507931712
Gain	Pathogenic	C	3.7497365083335	2403	1576.53885641941	1.52422503905652
Gain	**Rare**	**P**	**1**	**5462**	**14054.9448363412**	**0.388617676099103**
Gain	Rare	G	1.35615927625121	8167	15496.3460565981	0.527027466356988

Gain	Rare	D	1.53315052221198	5854	9825.28988006008	0.595809393052147
Gain	Rare	A	1.7408504248511	9221	13629.9421893499	0.676525246541769
Gain	Rare	E	1.93697833298716	6494	8627.10276148417	0.752744018419783
Gain	Rare	Y	1.97394842252092	3793	4944.52402624783	0.76711124869957
Gain	Rare	L	2.20878275219798	16179	18848.470849333	0.858372020166959
Gain	Rare	R	2.4358782168767	14729	15559.4821980831	0.946625331903049
Gain	Rare	S	2.62326740093865	17962	17619.3376909652	1.01944808113931
Gain	Rare	N	2.79079945704973	8504	7841.0111476748	1.0845539994573
Gain	Rare	K	2.8637733861843	8685	7803.84479837699	1.1129129582134
Gain	Rare	F	3.04161088642488	5719	4838.31224143441	1.18202375428017
Gain	Rare	I	3.0878308660899	11311	9425.94600978066	1.19998565536694
Gain	Rare	T	3.12885134967952	15558	12795.1766536567	1.215926940372
Gain	Rare	H	3.26425431766141	12019	9474.6199316644	1.26854692712604
Gain	Rare	V	3.63203681212051	18145	12855.3581481264	1.41147370543267
Gain	Rare	Q	3.73790796834287	12348	8500.5194630389	1.45261710812972
Gain	Rare	C	4.14501854003516	10729	6660.55191119074	1.61082747241616
Gain	Rare	W	4.31706539881316	6173	3679.46858048339	1.67768792285462
Gain	Rare	M	5.49190780939486	10214	4785.75062610947	2.13425245023754
Gain	**Rarest**	**P**	**1**	**35911**	**55599.4532683072**	**0.645887646173492**
Gain	Rarest	D	1.12588480879088	34764	47805.6033728324	0.727195089012435
Gain	Rarest	A	1.24404135431529	50246	62533.0625490389	0.803510942081186
Gain	Rarest	G	1.25296112082836	50180	62006.3380871049	0.809272109078727
Gain	Rarest	Y	1.3523867951682	24198	27702.6664410959	0.8734899238473
Gain	Rarest	E	1.40196398845443	38163	42145.253570646	0.905511220522834
Gain	Rarest	L	1.42213097196742	74943	81589.5431471668	0.918536826034457
Gain	Rarest	H	1.44063953266915	41315	44401.2760063063	0.930491276740156
Gain	Rarest	N	1.51483549477031	42887	43833.2040541426	0.978413532057254
Gain	Rarest	R	1.5640258085687	80882	80066.526586094	1.01018494805103
Gain	Rarest	I	1.58989781796977	51921	50561.1399735477	1.02689535930487
Gain	Rarest	Q	1.66403250639903	40274	37471.923084863	1.07477803871424
Gain	Rarest	S	1.67006232396462	86238	79948.2606030104	1.07867262338854
Gain	Rarest	K	1.71300304292206	42946	38815.7165173296	1.10640750328096
Gain	Rarest	T	1.73567994956745	69279	61798.0805076389	1.12105423713664
Gain	Rarest	W	1.93091273674921	16125	12929.4514025396	1.24715268250536
Gain	Rarest	F	1.99155007614751	33927	26375.2903944558	1.28631759091955
Gain	Rarest	V	2.07670326127262	85893	64036.3174419872	1.34131698122419
Gain	Rarest	C	2.09097590859025	40008	29623.8046084617	1.35053550780483
Gain	Rarest	M	2.1704775929579	31893	22750.0883834304	1.40188466358789
Gain	**VUS**	**P**	**1**	**8501**	**14309.0558198518**	**0.594099296768838**
Gain	VUS	G	1.12971596132598	10908	16252.3746893261	0.671163458172296
Gain	VUS	D	1.1423115033056	7677	11312.2228485444	0.678646460804813
Gain	VUS	A	1.18512280502666	10371	14729.847166648	0.704080625051053
Gain	VUS	Y	1.41352749892368	5250	6251.66939612474	0.839775693073973
Gain	VUS	E	1.42313438349405	8387	9919.77206701104	0.845483136441372
Gain	VUS	L	1.43200401539277	17230	20252.6568066067	0.850752578514999

Gain	VUS	R	1.67507598085845	17986	18073.4490653517	0.995161462262378
Gain	VUS	S	1.69533713259566	19501	19361.6234510914	1.00719859826118
Gain	VUS	N	1.74671042617249	10118	9750.22694041869	1.03771943584787
Gain	VUS	I	1.76835538993662	11906	11332.8016954316	1.05057869359873
Gain	VUS	H	1.85846809167297	11827	10711.7505251965	1.10411458633023
Gain	VUS	T	1.95803819916775	16400	14098.1994260009	1.16326911717208
Gain	VUS	K	1.96979528276316	10590	9049.31755838289	1.17025399226817
Gain	VUS	F	2.03066139634918	7322	6069.22409685688	1.20641450754667
Gain	VUS	Q	2.09496196261016	11615	9332.19991633861	1.24461542874416
Gain	VUS	V	2.23957181868767	19991	15024.8618298581	1.33052804254565
Gain	VUS	W	2.60952130559155	5671	3657.96682092697	1.55031477255524
Gain	VUS	C	2.66442751949471	11797	7452.61404282061	1.58293451562335
Gain	VUS	M	2.88534454893287	9346	5452.16583721186	1.71418116745682

Table 3-9. Unique counts of OMIM proteins and residues available for analysis following PDB structure mapping.

Top row of table shows protein and residue counts for 501 unique OMIM&CpD proteins. Middle row of table shows counts based on a filtered subset of 419 unique OMIM&CpD proteins containing at least one pathogenic and one common/benign missense variant position. Bottom row shows the percent of data available to use for 3D analysis relative to data used for 1D window analysis.
Abbreviations: CpDC: Chemoproteomic-Detected Cysteine; CpDK: Chemoproteomic-Detected Lysine; CpDY: Chemoproteomic-Detected Tyrosine.

	CpDC protein	**CpD-C**	**Other-C**	**CpK protein**	**CpD-K**	**Other-K**	**CpDY protein**	**CpD-Y**	**Other-Y**
OMIM&CpD	270	696	703	426	1,551	3,988	174	351	1,472
Control for missense categories	211	570	634	330	1,209	3,654	142	287	1,406
% of 1D available for 3D	29.1%	29.9%	5.5%	59.5%	57.8%	7.7%	58.7%	52.4%	7.0%

Table 3-10. Multi-detected OMIM&CpD proteins with cysteine, lysine, and tyrosine CpDAA positions.

Gene	MOEUF constrained (<0.35)	PPIs	VUS	Phenotype count	OMIM phenotype(s)	OMIM inheritance

ACTB	**TRUE**	More	25	**2**	['?Dystonia, juvenile-onset', 'Baraitser-Winter syndrome 1']	Autosomal dominant
ACTN1	FALSE	More	18	1	['Bleeding disorder, platelet-type, 15']	Autosomal dominant
ACTN4	FALSE	61-100	24	1	['Glomerulosclerosis, focal segmental, 1']	Autosomal dominant
ADSL	FALSE	2-10	138	1	['Adenylosuccinase deficiency']	Autosomal recessive
AHCY	FALSE	11-30	22	1	['Hypermethioninemia with deficiency of S-adenosylhomocysteine hydrolase']	Autosomal recessive
AIMP2	FALSE	61-100	3	1	['Leukodystrophy, hypomyelinating, 17']	Autosomal recessive
AK1	FALSE	2-10	2	1	['Hemolytic anemia due to adenylate kinase deficiency']	Autosomal recessive
AKT2	FALSE	61-100	12	2	['Hypoinsulinemic hypoglycemia with hemihypertrophy', 'Diabetes mellitus, type II']	Autosomal dominant
AKT3	FALSE	11-30	9	1	['Megalencephaly-polymicrogyria-polydactyly-hydrocephalus syndrome 2']	Autosomal dominant
ALDOA	FALSE	31-60	7	1	['Glycogen storage disease XII']	Autosomal recessive
ANXA11	FALSE	11-30	0	1	['Amyotrophic lateral sclerosis 23']	Autosomal dominant
AP3D1	FALSE	11-30	16	1	['?Hermansky-Pudlak syndrome 10']	Autosomal recessive
APRT	FALSE	11-30	6	1	['Adenine phosphoribosyltransferase deficiency']	Autosomal recessive
ARCN1	FALSE	11-30	4	1	['Short stature, rhizomelic, with microcephaly, micrognathia, and developmental delay']	Autosomal dominant
ARF1	**TRUE**	31-60	3	**1**	['Periventricular nodular heterotopia 8']	Autosomal dominant
ARHGDIA	FALSE	61-100	0	1	['Nephrotic syndrome, type 8']	Autosomal recessive
ARPC1B	FALSE	11-30	7	1	['Immunodeficiency 71 with inflammatory disease and congenital thrombocytopenia']	Autosomal recessive
ASNS	FALSE	2-10	14	1	['Asparagine synthetase deficiency']	Autosomal recessive
ATIC	FALSE	11-30	4	1	['AICA-ribosiduria due to ATIC deficiency']	Autosomal recessive
ATP6V1A	FALSE	11-30	8	2	['Cutis laxa, autosomal recessive, type IID', 'Developmental and epileptic encephalopathy 93']	Autosomal recessive, Autosomal dominant
BAG3	FALSE	61-100	223	2	['Cardiomyopathy, dilated, 1HH', 'Myopathy, myofibrillar, 6']	Autosomal dominant
BLVRA	FALSE	11-30	0	1	['Hyperbiliverdinemia']	Autosomal

						dominant, Autosomal recessive
CAD	FALSE	31-60	41	1	['Developmental and epileptic encephalopathy 50']	Autosomal recessive
CCT5	FALSE	31-60	42	1	['Neuropathy, hereditary sensory, with spastic paraplegia']	Autosomal recessive
CDC73	FALSE	61-100	184	3	['Parathyroid adenoma with cystic changes', 'Hyperparathyroidism-jaw tumor syndrome', 'Hyperparathyroidism, familial primary']	Autosomal dominant
CFL2	FALSE	11-30	19	1	['Nemaline myopathy 7, autosomal recessive']	Autosomal recessive
CHD4	FALSE	61-100	23	1	['Sifrim-Hitz-Weiss syndrome']	Autosomal dominant
CLTC	**TRUE**	More	10	**1**	['Mental retardation, autosomal dominant 56']	Autosomal dominant
CNBP	FALSE	11-30	0	1	['Myotonic dystrophy 2']	Autosomal dominant
COASY	FALSE	2-10	23	2	['Neurodegeneration with brain iron accumulation 6', 'Pontocerebellar hypoplasia, type 12']	Autosomal recessive
COG1	FALSE	2-10	55	1	['Congenital disorder of glycosylation, type IIg']	Autosomal recessive
COPA	FALSE	11-30	98	1	['{Autoimmune interstitial lung, joint, and kidney disease}']	Autosomal dominant
COPB1	FALSE	61-100	0	1	['Baralle-Macken syndrome']	Autosomal recessive
COPB2	FALSE	31-60	0	1	['?Microcephaly 19, primary, autosomal recessive']	Autosomal recessive
CTPS1	FALSE	11-30	35	1	['Immunodeficiency 24']	Autosomal recessive
CUL3	FALSE	More	5	2	['Neurodevelopmental disorder with or without autism or seizures', 'Pseudohypoaldosteronism, type IIE']	Autosomal dominant
CYB5R3	FALSE	31-60	8	2	['Methemoglobinemia, type I', 'Methemoglobinemia, type II']	Autosomal recessive
DCPS	FALSE	11-30	9	1	['Al-Raqad syndrome']	Autosomal recessive
DCTN1	FALSE	61-100	229	2	['Perry syndrome', 'Neuronopathy, distal hereditary motor, type VIIB']	Autosomal dominant
DCXR	FALSE	2-10	0	1	['[Pentosuria]']	Autosomal recessive
DDX3X	**TRUE**	31-60	27	**1**	['Intellectual developmental disorder, X-linked, syndrome, Snijders Blok type']	X-linked dominant, X-linked

						recessive
DDX6	FALSE	61-100	0	1	['Intellectual developmental disorder with impaired language and dysmorphic facies']	Autosomal dominant
DNAJB6	FALSE	31-60	55	1	['Muscular dystrophy, limb-girdle, autosomal dominant 1']	Autosomal dominant
DNM2	FALSE	61-100	201	4	['Lethal congenital contracture syndrome 5', 'Charcot-Marie-Tooth disease, axonal type 2M', 'Centronuclear myopathy 1', 'Charcot-Marie-Tooth disease, dominant intermediate B']	Autosomal recessive, Autosomal dominant
DNMT1	FALSE	61-100	223	2	['Cerebellar ataxia, deafness, and narcolepsy, autosomal dominant', 'Neuropathy, hereditary sensory, type IE']	Autosomal dominant
DYNC1H1	FALSE	31-60	524	3	['Mental retardation, autosomal dominant 13', 'Spinal muscular atrophy, lower extremity-predominant 1, AD', 'Charcot-Marie-Tooth disease, axonal, type 20']	Autosomal dominant
ECHS1	FALSE	31-60	11	1	['Mitochondrial short-chain enoyl-CoA hydratase 1 deficiency']	Autosomal recessive
EDC3	FALSE	31-60	0	1	['?Mental retardation, autosomal recessive 50']	Autosomal recessive
EEF1A2	**TRUE**	31-60	40	**2**	['Developmental and epileptic encephalopathy 33', 'Mental retardation, autosomal dominant 38']	Autosomal dominant
EEF2	FALSE	31-60	13	1	['?Spinocerebellar ataxia 26']	Autosomal dominant
EIF4A3	FALSE	31-60	1	1	['Robin sequence with cleft mandible and limb anomalies']	Autosomal recessive
ELP1	FALSE	31-60	261	1	['Dysautonomia, familial']	Autosomal recessive
ENO3	FALSE	11-30	30	1	['?Glycogen storage disease XIII']	Autosomal recessive
FARSA	FALSE	11-30	1	1	['?Rajab interstitial lung disease with brain calcifications 2']	Autosomal recessive
FH	FALSE	11-30	369	2	['Fumarase deficiency', 'Leiomyomatosis and renal cell cancer']	Autosomal recessive, Autosomal dominant
FHL1	FALSE	31-60	75	6	['Reducing body myopathy, X-linked 1a, severe, infantile or early childhood onset', 'Scapuloperoneal myopathy, X-linked dominant', 'Reducing body myopathy, X-linked 1b, with late childhood or adult onset', '?Uruguay	X-linked dominant, X-linked recessive, X-linked

					faciocardiomusculoskeletal syndrome', 'Emery-Dreifuss muscular dystrophy 6, X-linked', 'Myopathy, X-linked, with postural muscle atrophy']	
FLNA	FALSE	More	490	10	['Otopalatodigital syndrome, type I', 'Congenital short bowel syndrome', 'Otopalatodigital syndrome, type II', 'Intestinal pseudoobstruction, neuronal', 'Melnick-Needles syndrome', 'Cardiac valvular dysplasia, X-linked', '?FG syndrome 2', 'Heterotopia, periventricular, 1', 'Terminal osseous dysplasia', 'Frontometaphyseal dysplasia 1']	X-linked dominant, X-linked, X-linked recessive
FLNB	FALSE	31-60	145	5	['Larsen syndrome', 'Atelosteogenesis, type I', 'Boomerang dysplasia', 'Spondylocarpotarsal synostosis syndrome', 'Atelosteogenesis, type III']	Autosomal recessive, Autosomal dominant
FLNC	FALSE	31-60	1009	4	['Cardiomyopathy, familial hypertrophic, 26', 'Myopathy, myofibrillar, 5', 'Cardiomyopathy, familial restrictive 5', 'Myopathy, distal, 4']	Autosomal dominant
FTO	FALSE	2-10	17	1	['Growth retardation, developmental delay, facial dysmorphism']	Autosomal recessive
FUS	FALSE	61-100	37	1	['Essential tremor, hereditary, 4']	Autosomal dominant
G6PD	FALSE	2-10	29	1	['Hemolytic anemia, G6PD deficient (favism)']	X-linked dominant
GANAB	FALSE	11-30	11	1	['Polycystic kidney disease 3']	Autosomal dominant
GDI1	FALSE	11-30	9	1	['Mental retardation, X-linked 41']	X-linked dominant
GFPT1	FALSE	2-10	55	1	['Myasthenia, congenital, 12, with tubular aggregates']	Autosomal recessive
GLUL	FALSE	31-60	11	1	['Glutamine deficiency, congenital']	Autosomal recessive
GPI	FALSE	11-30	6	1	['Hemolytic anemia, nonspherocytic, due to glucose phosphate isomerase deficiency']	Autosomal recessive
HADH	FALSE	2-10	21	2	['3-hydroxyacyl-CoA dehydrogenase deficiency', 'Hyperinsulinemic hypoglycemia, familial, 4']	Autosomal recessive
HCFC1	FALSE	61-100	75	1	['Mental retardation, X-linked 3 (methylmalonic acidemia and homocysteinemia, cblX type)']	X-linked recessive

HINT1	FALSE	11-30	26	1	['Neuromyotonia and axonal neuropathy, autosomal recessive']	Autosomal recessive
HK1	FALSE	11-30	62	4	['Hemolytic anemia due to hexokinase deficiency', 'Neuropathy, hereditary motor and sensory, Russe type', 'Neurodevelopmental disorder with visual defects and brain anomalies', 'Retinitis pigmentosa 79']	Autosomal recessive, Autosomal dominant
HMGB3	FALSE	2-10	0	1	['?Microphthalmia, syndromic 13']	X-linked
HNRNPA1	FALSE	61-100	4	2	['Amyotrophic lateral sclerosis 20', '?Inclusion body myopathy with early-onset Paget disease without frontotemporal dementia 3']	Autosomal dominant
HNRNPH2	TRUE	31-60	2	1	['Mental retardation, X-linked, syndromic, Bain type']	X-linked dominant
HNRNPK	FALSE	More	4	1	['Au-Kline syndrome']	Autosomal dominant
HNRNPU	FALSE	More	92	1	['Developmental and epileptic encephalopathy 54']	Autosomal dominant
HPRT1	FALSE	11-30	10	2	['Hyperuricemia, HRPT-related', 'Lesch-Nyhan syndrome']	X-linked recessive
HSD17B10	FALSE	31-60	11	1	['HSD10 mitochondrial disease']	X-linked dominant
HSPA9	FALSE	61-100	2	2	['Even-plus syndrome', 'Anemia, sideroblastic, 4']	Autosomal recessive, Autosomal dominant
HSPD1	FALSE	More	25	2	['Spastic paraplegia 13, autosomal dominant', 'Leukodystrophy, hypomyelinating, 4']	Autosomal recessive, Autosomal dominant
HUWE1	FALSE	61-100	87	1	['Mental retardation, X-linked syndromic, Turner type']	X-linked
HYOU1	FALSE	11-30	13	1	['?Immunodeficiency 59 and hypoglycemia']	Autosomal recessive
LDHA	FALSE	31-60	10	1	['Glycogen storage disease XI']	Autosomal recessive
LMNA	FALSE	More	370	11	['Muscular dystrophy, congenital', 'Lipodystrophy, familial partial, type 2', 'Charcot-Marie-Tooth disease, type 2B1', 'Cardiomyopathy, dilated, 1A', 'Heart-hand syndrome, Slovenian type', 'Hutchinson-Gilford progeria', 'Restrictive dermopathy, lethal', 'Mandibuloacral dysplasia', 'Emery-Dreifuss muscular	Autosomal recessive, Autosomal dominant

					dystrophy 2, autosomal dominant', 'Emery-Dreifuss muscular dystrophy 3, autosomal recessive', 'Malouf syndrome']	
MAGED2	FALSE	11-30	0	1	['Bartter syndrome, type 5, antenatal, transient']	X-linked recessive
MAPK1	**TRUE**	More	0	**1**	['Noonan syndrome 13']	Autosomal dominant
MCM4	FALSE	61-100	44	1	['Immunodeficiency 54']	Autosomal recessive
MCM5	FALSE	31-60	1	1	['?Meier-Gorlin syndrome 8']	Autosomal recessive
MDH1	FALSE	11-30	0	1	['?Developmental and epileptic encephalopathy 88']	Autosomal recessive
MESD	FALSE	61-100	0	1	['Osteogenesis imperfecta, type XX']	Autosomal recessive
MRE11	FALSE	61-100	561	1	['Ataxia-telangiectasia-like disorder 1']	Autosomal recessive
MSH6	FALSE	31-60	2576	2	['Mismatch repair cancer syndrome 3', 'Colorectal cancer, hereditary nonpolyposis, type 5']	Autosomal recessive, Autosomal dominant
MSN	FALSE	31-60	3	1	['Immunodeficiency 50']	X-linked recessive
MVD	FALSE	2-10	3	1	['Porokeratosis 7, multiple types']	Autosomal dominant
MYH9	FALSE	61-100	106	2	['Deafness, autosomal dominant 17', 'Macrothrombocytopenia and granulocyte inclusions with or without nephritis or sensorineural hearing loss']	Autosomal dominant
NAA10	FALSE	61-100	15	2	['Ogden syndrome', 'Microphthalmia, syndromic 1']	X-linked, X-linked dominant, X-linked recessive
NAA15	FALSE	11-30	9	1	['Mental retardation, autosomal dominant 50']	Autosomal dominant
NAXE	FALSE	2-10	3	1	['Encephalopathy, progressive, early-onset, with brain edema and/or leukoencephalopathy']	Autosomal recessive
NCAPD2	FALSE	11-30	5	1	['?Microcephaly 21, primary, autosomal recessive']	Autosomal recessive
NCAPH	FALSE	11-30	3	1	['?Microcephaly 23, primary, autosomal recessive']	Autosomal recessive
NDRG1	FALSE	11-30	110	1	['Charcot-Marie-Tooth disease, type 4D']	Autosomal recessive
NHLRC2	FALSE	2-10	1	1	['FINCA syndrome']	Autosomal recessive
NSUN2	FALSE	2-10	39	1	['Mental retardation, autosomal recessive 5']	Autosomal recessive
P4HB	FALSE	61-100	4	1	['Cole-Carpenter syndrome 1']	Autosomal

						dominant
PABPN1	FALSE	11-30	1	1	['Oculopharyngeal muscular dystrophy']	Autosomal dominant
PCNA	FALSE	More	1	1	['?Ataxia-telangiectasia-like disorder 2']	Autosomal recessive
PDXK	FALSE	2-10	0	1	['Neuropathy, hereditary motor and sensory, type VIC, with optic atrophy']	Autosomal recessive
PFKM	FALSE	11-30	30	1	['Glycogen storage disease VII']	Autosomal recessive
PGK 1	FALSE	11-30	14	1	['Phosphoglycerate kinase 1 deficiency']	X-linked recessive
PGM3	FALSE	2-10	47	1	['Immunodeficiency 23']	Autosomal recessive
PLAA	FALSE	2-10	4	1	['Neurodevelopmental disorder with progressive microcephaly, spasticity, and brain anomalies']	Autosomal recessive
PLPBP	FALSE	No PPI	7	1	['Epilepsy, early-onset, vitamin B6-dependent']	Autosomal recessive
PLS3	FALSE	11-30	0	1	['Bone mineral density QTL18, osteoporosis']	X-linked dominant
PNP	FALSE	11-30	43	1	['Immunodeficiency due to purine nucleoside phosphorylase deficiency']	Autosomal recessive
PNPO	FALSE	2-10	67	1	["Pyridoxamine 5'-phosphate oxidase deficiency"]	Autosomal recessive
POLD1	FALSE	31-60	1154	1	['Mandibular hypoplasia, deafness, progeroid features, and lipodystrophy syndrome']	Autosomal dominant
PPA2	FALSE	2-10	0	2	['?Sudden cardiac failure, alcohol-induced', 'Sudden cardiac failure, infantile']	Autosomal recessive
PPP1CB	**TRUE**	61-100	5	**1**	['Noonan syndrome-like disorder with loose anagen hair 2']	Autosomal dominant
PPP2CA	**TRUE**	More	1	**1**	['Neurodevelopmental disorder and language delay with or without structural brain abnormalities']	Autosomal dominant
PPP2R1A	FALSE	More	9	1	['Mental retardation, autosomal dominant 36']	Autosomal dominant
PQBP1	FALSE	11-30	22	1	['Renpenning syndrome']	X-linked recessive
PRDX1	FALSE	31-60	0	1	['Methylmalonic aciduria and homocystinuria, cblC type, digenic']	Autosomal recessive
PRKDC	FALSE	More	425	1	['Immunodeficiency 26, with or without neurologic abnormalities']	Autosomal recessive
PRPF8	FALSE	61-100	110	1	['Retinitis pigmentosa 13']	Autosomal dominant
PRPS1	**TRUE**	11-30	11	**5**	['Charcot-Marie-Tooth disease, X-linked recessive, 5',	X-linked, X-linked

					'Phosphoribosylpyrophosphate synthetase superactivity', 'Deafness, X-linked 1', 'Arts syndrome', 'Gout, PRPS-related']	recessive
PSAT1	FALSE	2-10	29	2	['Neu-Laxova syndrome 2', '?Phosphoserine aminotransferase deficiency']	Autosomal recessive
PSMG2	FALSE	2-10	2	1	['?Proteasome-associated autoinflammatory syndrome 4']	Autosomal recessive
PTPN11	FALSE	More	99	3	['LEOPARD syndrome 1', 'Metachondromatosis', 'Noonan syndrome 1']	Autosomal dominant
PYGL	FALSE	11-30	49	1	['Glycogen storage disease VI']	Autosomal recessive
RAB7A	FALSE	61-100	21	1	['Charcot-Marie-Tooth disease, type 2B']	Autosomal dominant
RAC1	**TRUE**	More	4	**1**	['Mental retardation, autosomal dominant 48']	Autosomal dominant
RBPJ	FALSE	31-60	6	1	['Adams-Oliver syndrome 3']	Autosomal dominant
RDX	FALSE	11-30	15	1	['Deafness, autosomal recessive 24']	Autosomal recessive
RNASEH2A	FALSE	No PPI	67	1	['Aicardi-Goutieres syndrome 4']	Autosomal recessive
RNASEH2B	FALSE	No PPI	37	1	['Aicardi-Goutieres syndrome 2']	Autosomal recessive
RPL5	FALSE	31-60	23	1	['Diamond-Blackfan anemia 6']	Autosomal dominant
SAMHD1	FALSE	11-30	68	2	['?Chilblain lupus 2', 'Aicardi-Goutieres syndrome 5']	Autosomal recessive, Autosomal dominant
SEC23B	FALSE	11-30	26	2	['?Cowden syndrome 7', 'Dyserythropoietic anemia, congenital, type II']	Autosomal recessive, Autosomal dominant
SHMT2	FALSE	11-30	1	1	['Neurodevelopmental disorder with cardiomyopathy, spasticity, and brain abnormalities']	Autosomal recessive
SMC1A	**TRUE**	61-100	63	**2**	['Developmental and epileptic encephalopathy 85, with or without midline brain defects', 'Cornelia de Lange syndrome 2']	X-linked dominant
SNRNP200	FALSE	31-60	135	1	['Retinitis pigmentosa 33']	Autosomal dominant
SORD	FALSE	2-10	0	1	['Sorbitol dehydrogenase deficiency with peripheral neuropathy']	Autosomal recessive
SPART	FALSE	11-30	43	1	['Troyer syndrome']	Autosomal recessive
STAT1	**TRUE**	More	47	**3**	['Immunodeficiency 31C, chronic mucocutaneous	Autosomal recessive,

					candidiasis, autosomal dominant', 'Immunodeficiency 31A, mycobacteriosis, autosomal dominant', 'Immunodeficiency 31B, mycobacterial and viral infections, autosomal recessive']	Autosomal dominant
TBC1D23	FALSE	2-10	7	1	['Pontocerebellar hypoplasia, type 11']	Autosomal recessive
TKT	FALSE	11-30	2	1	['Short stature, developmental delay, and congenital heart defects']	Autosomal recessive
TUBA1A	**TRUE**	More	16	**1**	['Lissencephaly 3']	Autosomal dominant
TUBA4A	FALSE	61-100	0	1	['Amyotrophic lateral sclerosis 22 with or without frontotemporal dementia']	Autosomal dominant
TUBB	**TRUE**	More	8	**2**	['Symmetric circumferential skin creases, congenital, 1', 'Cortical dysplasia, complex, with other brain malformations 6']	Autosomal dominant
TUBB2B	**TRUE**	31-60	22	**1**	['Cortical dysplasia, complex, with other brain malformations 7']	Autosomal dominant
TUBB3	**TRUE**	31-60	21	**2**	['Fibrosis of extraocular muscles, congenital, 3A', 'Cortical dysplasia, complex, with other brain malformations 1']	Autosomal dominant
TUBB6	FALSE	11-30	2	1	['?Facial palsy, congenitla, with ptosis and velopharyngeal dysfunction']	Autosomal dominant
UBA1	FALSE	More	82	1	['Spinal muscular atrophy, X-linked 2, infantile']	X-linked recessive
UBA5	FALSE	11-30	3	2	['?Spinocerebellar ataxia, autosomal recessive 24', 'Developmental and epileptic encephalopathy 44']	Autosomal recessive
UBE2T	FALSE	11-30	0	1	['Fanconi anemia, complementation group T']	Autosomal recessive
UMPS	FALSE	11-30	19	1	['Orotic aciduria']	Autosomal recessive
UROD	FALSE	11-30	10	2	['Porphyria, hepatoerythropoietic', 'Porphyria cutanea tarda']	Autosomal dominant, Autosomal recessive
USP7	FALSE	More	6	1	['Hao-Fountain syndrome']	Autosomal dominant
USP9X	FALSE	61-100	52	2	['Mental retardation, X-linked 99', 'Mental retardation, X-linked 99, syndromic, female-restricted']	X-linked dominant, X-linked recessive
VCP	**TRUE**	More	62	**2**	['Inclusion body myopathy with early-onset Paget disease and frontotemporal dementia 1',	Autosomal dominant

					'Charcot-Marie-Tooth disease, type 2Y']	
VIM	FALSE	More	8	1	['Cataract 30, pulverulent']	Autosomal dominant
VPS35	FALSE	11-30	27	1	['{Parkinson disease 17}']	Autosomal dominant
VPS4A	FALSE	11-30	1	1	['CIMDAG syndrome']	Autosomal dominant
WBP2	FALSE	31-60	1	1	['Deafness, autosomal recessive 107']	Autosomal recessive
WDR1	FALSE	11-30	9	1	['Periodic fever, immunodeficiency, and thrombocytopenia syndrome']	Autosomal recessive
YWHAG	FALSE	More	3	1	['Developmental and epileptic encephalopathy 56']	Autosomal dominant

REFERENCES

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

Ainscough, B. J. *et al.* DoCM: A database of curated mutations in cancer. *Nat. Methods* **13**, 806–807 (2016).

Alirezaie, N., Kernohan, K. D., Hartley, T., Majewski, J. & Hocking, T. D. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. *Am. J. Hum. Genet.* **103**, 474–483 (2018).

Althari, S. *et al.* Unsupervised Clustering of Missense Variants in HNF1A Using Multidimensional Functional Data Aids Clinical Interpretation. *Am. J. Hum. Genet.* **107**, 670–682 (2020).

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–D798 (2015).

Armstrong, C. T., Mason, P. E., Anderson, J. L. R. & Dempsey, C. E. Arginine side chain interactions and the role of arginine as a gating charge carrier in voltage sensitive ion channels. *Sci. Rep.* **6**, 1–10 (2016).

Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

Backus, K. M. *et al.* Proteome-wide covalent ligand discovery in native biological systems. *Nature* **534**, 570–574 (2016).

Bandyopadhyay, U. *et al.* Leucine retention in lysosomes is regulated by starvation. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2114912119 (2022).

Bartas, M., Červeň, J., Guziurová, S., Slychko, K. & Pečinka, P. Amino acid composition in various types of nucleic acid-binding proteins. *Int. J. Mol. Sci.* **22**, 1–12 (2021).

Bennet, S. A., Cohen, M. A. & Gonnet, G. H. Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng. Des. Sel.* **7**, 1323–1332 (1994).

Branciamore, S., Chen, Z. X., Riggs, A. D. & Rodin, S. N. CpG island clusters and pro-epigenetic selection for CpGs in protein-coding exons of HOX and other transcription factors. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 15485–15490 (2010).

Brooks, D. J. & Fresco, J. R. Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor. *Mol. Cell. Proteomics* **1**, 125–131 (2002).

Brooks, D. J., Fresco, J. R., Lesk, A. M. & Singh, M. Evolution of amino acid frequencies in proteins over deep time: Inferred order of introduction of amino acids into the genetic code. *Mol. Biol. Evol.* **19**, 1645–1655 (2002).

Buljan, M., Blattmann, P., Aebersold, R. & Boutros, M. Systematic characterization of pan-cancer mutation clusters. *Mol. Syst. Biol.* **14**, e7974 (2018).

Butterfield, R. J. *et al.* Position of glycine substitutions in the triple helix of COL6A1, COL6A2, and COL6A3 is correlated with severity and mode of inheritance in collagen vi myopathies. *Hum. Mutat.* **34**, 1558–1567 (2013).

Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14 Suppl 3**, S3 (2013).

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS One* **7**, e46688 (2012).

Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res.* **19**, 1553–1561 (2009).

Clark, M. M. *et al.* Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *npj Genomic Med.* **3**, 16 (2018).

Coban-Akdemir, Z. *et al.* Identifying Genes Whose Mutant Transcripts Cause Dominant Disease Traits by Potential Gain-of-Function Alleles. *Am. J. Hum. Genet.* **103**, 171–187 (2018).

Cummings, B. B. *et al.* Transcript expression-aware annotation improves rare variant interpretation. *Nature* **581**, 452–458 (2020).

Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6**, e1001025 (2010).

de Beer, T. A. P. *et al.* Amino Acid Changes in Disease-Associated Variants Differ Radically from Variants Observed in the 1000 Genomes Project Dataset. *PLoS Comput. Biol.* **9**, (2013).

Desai, D. *et al.* Intragenic codon bias in a set of mouse and human genes. *J. Theor. Biol.* **230**, 215–225 (2004).

Dhindsa, R. S., Copeland, B. R., Mustoe, A. M. & Goldstein, D. B. Natural Selection Shapes Codon Usage in the Human Genome. *Am. J. Hum. Genet.* **107**, 83–95 (2020).

Dietz, H. C., Saraiva, J. M., Pyeritz, R. E., Cutting, G. R. & Francomano, C. A. Clustering of fibrillin (FBN1) missense mutations in Marfan syndrome patients at cysteine residues in EGF-like domains. *Hum. Mutat.* **1**, 366–374 (1992).

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

Favalli, V. *et al.* Machine learning-based reclassification of germline variants of unknown significance: The RENOVO algorithm. *Am. J. Hum. Genet.* **108**, 682–695 (2021).

Gao, M., Zhou, H. & Skolnick, J. Insights into Disease-Associated Mutations in the Human Proteome through Protein Structural Analysis. *Structure* **23**, 1362–1369 (2015).

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

Geisheker, M. R. *et al.* Hotspots of missense mutation identify neurodevelopmental disorder genes and functional domains. *Nat. Neurosci.* **20**, 1043–1051 (2017).

González-Pérez, A. & López-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.* **88**, 440–449 (2011).

Grantham, R. *et al.* Volume 8 Number 9 1980 Nucleic A c i d s Research. **8**, (1980).

Graur, D. Amino acid composition and the evolutionary rates of protein-coding genes. *J. Mol. Evol.* **22**, 53–62 (1985).

Grimm, D. G. *et al.* The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.* **36**, 513–523 (2015).

Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* **47**, 276–283 (2015).

Hacker, S. M. *et al.* Global profiling of lysine reactivity and ligandability in the human proteome. *Nat. Chem.* **9**, 1181–1190 (2017).

Hahm, H. S. *et al.* Global targeting of functional tyrosines using sulfur-triazole exchange chemistry. *Nat. Chem. Biol.* **16**, 150–159 (2020).

Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat. Genet.* **51**, 88–95 (2019).

Hayeck, T. J. *et al.* Improved Pathogenic Variant Localization via a Hierarchical Model of Sub-regional Intolerance. *Am. J. Hum. Genet.* **104**, 299–309 (2019).

Hormoz, S. Amino acid composition of proteins reduces deleterious impact of mutations. *Sci. Rep.* **3**, 1–10 (2013).

Huang, Y. F. Y. F. Y. F. Unified inference of missense variant effects and gene constraints in the human genome. *PLoS Genet.* **16**, 1–24 (2020).

Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.* **99**, 877–885 (2016).

Ionita-Laza, I., Mccallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat. Genet.* **48**, 214–220 (2016).

Iqbal, S. *et al.* Genomic analysis of AlphaFold2-predicted structures identifies maps of 3D essential sites in 243 neurodevelopmental disorder-associated proteins. *Biophys. J.* **121**, 165a-166a (2022).

Iqbal, S. *et al.* Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 28201–28211 (2020).

Jagadeesh, K. A. *et al.* M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat. Genet.* **48**, 1581–1586 (2016).

Jian, X. & Liu, X. In silico prediction of deleteriousness for nonsynonymous and splice-altering single nucleotide variants in the human genome. in *Methods in Molecular Biology* **1498**, 191–197 (2017).

Jordan, I. K. *et al.* A universal trend of amino acid gain and loss in protein evolution. *Nature* **433**, 633–638 (2005).

Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

Kamburov, A. *et al.* Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E5486–E5495 (2015).

Karczewski, K. J. K. J. K. J. K. J. K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

Khan, S. & Vihinen, M. Spectrum of disease-causing mutations in protein secondary structures. *BMC Struct. Biol.* **7**, 1–18 (2007).

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

Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: Implications for complex disease and association studies. *Am. J. Hum. Genet.* **80**, 727–739 (2007).

Landrum, M. J. *et al.* ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).

Lee, M. S. *et al.* Comprehensive analysis of missense variations in the BRCT domain of BRCA1 by structural and functional assays. *Cancer Res.* **70**, 4880–4890 (2010).

Lelieveld, S. H. *et al.* Spatial Clustering of de Novo Missense Mutations Identifies Candidate Neurodevelopmental Disorder-Associated Genes. *Am. J. Hum. Genet.* **101**, 478–484 (2017).

Li, Y., Zhang, Y., Li, X., Yi, S. & Xu, J. Gain-of-Function Mutations: An Emerging Advantage for Cancer Biology. *Trends in Biochemical Sciences* **44**, 659–674 (2019).

Lin, S. *et al.* Redox-based reagents for chemoselective methionine bioconjugation. *Science (80-.).* **355**, 597–602 (2017).

Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med.* **12**, 1–8 (2020).

Liu, X. *et al.* A Tyrosine Phosphoproteome Analysis Approach Enabled by Selective Dephosphorylation with Protein Tyrosine Phosphatase. *Anal. Chem.* **94**, 4155–4164 (2022).

Lu, Q. *et al.* A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci. Rep.* **5**, 10576 (2015).

Medina-Carmona, E. *et al.* Insight into the specificity and severity of pathogenic mechanisms associated with missense mutations through experimental and structural perturbation analyses. *Hum. Mol. Genet.* **28**, 1–15 (2019).

Molnár, J., Szakács, G. & Tusnády, G. E. Characterization of disease-associated mutations in human transmembrane proteins. *PLoS One* **11**, e0151760 (2016).

Nakamura, Y., Gojobori, T. & Ikemura, T. Codon usage tabulated from international DNA sequence databases: Status for the year 2000. *Nucleic Acids Res.* **28**, 292 (2000).

Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).

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

O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).

Parkin, J. Des *et al.* Mapping structural landmarks, ligand binding sites, and missense mutations to the collagen IV heterotrimers predicts major functional domains, novel

interactions, and variation in phenotypes in inherited diseases affecting basement membranes. *Human Mutation* **32**, 127–143 (2011).

Pejaver, V., Mooney, S. D. & Radivojac, P. Missense variant pathogenicity predictors generalize well across a range of function-specific prediction challenges. *Hum. Mutat.* **38**, 1092–1108 (2017).

Pérez-Palma, E. *et al.* Identification of pathogenic variant enriched regions across genes and gene families. *Genome Res.* **30**, 62–71 (2020).

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

Ponzoni, L., Peñaherrera, D. A., Oltvai, Z. N. & Bahar, I. Rhapsody: Predicting the pathogenicity of human missense variants. *Bioinformatics* **36**, 3084–3092 (2020).

Pottinger, T. D. *et al.* Pathogenic and Uncertain Genetic Variants Have Clinical Cardiac Correlates in Diverse Biobank Participants. *J. Am. Heart Assoc.* **9**, e013808 (2020).

Qi, H. *et al.* MVP predicts the pathogenicity of missense variants by deep learning. *Nat. Commun.* **12**, 510 (2021).

Qiu, Y. *et al.* Collagen Gly missense mutations: Effect of residue identity on collagen structure and integrin binding. *J. Struct. Biol.* **203**, 255–262 (2018).

Quang, D., Chen, Y. & Xie, X. DANN: A deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–763 (2015).

Quinodoz, M. *et al.* Analysis of missense variants in the human genome reveals widespread gene-specific clustering and improves prediction of pathogenicity. *Am. J. Hum. Genet.* **109**, 457–470 (2022).

Raimondi, D. *et al.* DEOGEN2: Prediction and interactive visualization of single amino acid variant deleteriousness in human proteins. *Nucleic Acids Res.* **45**, W201–W206 (2017).

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).

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

Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).

Rogers, M. F. *et al.* FATHMM-XF: Accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* **34**, 511–513 (2018).

Samocha, K. E., Kosmicki, J. A., Karczewski, K. J., O'Donnell-Luria, A. H., Pierce-Hoffman, E., MacArthur, D. G., Neale, B. M., & Daly, M. J. (2017). Regional missense constraint improves variant deleteriousness prediction. *BioRxiv*, 148353. https://doi.org/10.1101/148353

Schulze, K. v., Hanchard, N. A., & Wangler, M. F. (2020). Biases in arginine codon usage correlate with genetic disease risk. *Genetics in Medicine*, *22*(8), 1407–1412. https://doi.org/10.1038/s41436-020-0813-6

Schwarz, J. M., Rödelsperger, C., Schuelke, M., & Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. In *Nature Methods* (Vol. 7, Issue 8, pp. 575–576). Nature Publishing Group. https://doi.org/10.1038/nmeth0810-575

Sharp, P. M., & Li, W. H. (1987). The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, *15*(3), 1281–1295. https://doi.org/10.1093/nar/15.3.1281

Stenson, P. D., Mort, M., Ball, E. v., Chapman, M., Evans, K., Azevedo, L., Hayden, M., Heywood, S., Millar, D. S., Phillips, A. D., & Cooper, D. N. (2020). The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. *Human Genetics*, *139*(10), 1197–1207. https://doi.org/10.1007/s00439-020-02199-3

Sundaram, L., Gao, H., Padigepati, S. R. S. R., McRae, J. F. J. F., Li, Y., Kosmicki, J. A. J. A., Fritzilas, N., Hakenberg, J., Dutta, A., Shon, J., Xu, J., Batzloglou, S., Li, X., Farh, K. K.-H. H., Batzoglou, S., Li, X., Farh, K. K.-H. H., Batzloglou, S., Li, X., … al., et. (2018). Predicting the clinical impact of human mutation with deep neural networks. *Nature Genetics*, *50*(8), 1161–1170. https://doi.org/10.1038/s41588-018-0167-z

Tavtigian, S. v., Greenblatt, M. S., Harrison, S. M., Nussbaum, R. L., Prabhu, S. A., Boucher, K. M., & Biesecker, L. G. (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genetics in Medicine*, *20*(9), 1054–1060. https://doi.org/10.1038/gim.2017.210

Vaser, R., Adusumalli, S., Leng, S. N., Sikic, M., & Ng, P. C. (2016). SIFT missense predictions for genomes. *Nature Protocols*, *11*(1), 1–9. https://doi.org/10.1038/nprot.2015.123

Veitia, R. A., Caburet, S., & Birchler, J. A. (2018). Mechanisms of Mendelian dominance. In *Clinical Genetics* (Vol. 93, Issue 3, pp. 419–428). https://doi.org/10.1111/cge.13107

Waring, A. A. J. (2020). *Exploration of rare-missense variant clustering in Mendelian disease-genes*. University of Oxford.

Weerapana, E., Wang, C., Simon, G. M., Richter, F., Khare, S., Dillon, M. B. D., Bachovchin, D. A., Mowen, K., Baker, D., & Cravatt, B. F. (2010). Quantitative reactivity profiling predicts functional cysteines in proteomes. *Nature*, *468*(7325), 790–797. https://doi.org/10.1038/nature09472

Weerapana, E., Simon, G. M., & Cravatt, B. F. (2008). Disparate proteome reactivity profiles of carbon electrophiles. *Nature Chemical Biology*, *4*(7), 405–407. https://doi.org/10.1038/nchembio.91