**INTRODUCTION paragraph outline notes:**

1. **AA level resolution important, little is known about the role of specific amino acids**
2. **(combine 2 and 3) consequences of substitutions are poorly understood. Increase in available information online has helped identify important AA when there is lack of experimental validation (variant databases and sequence annotation). Score performance depends on what inputs go into prediction. Many types of inputs like conservation (main one), codon constraint, structural, chemical properties... Scores leverage annotations to predict missense pathogenicity, but there is no perfect score for all cases of AA in proteins.**
3. **(paragraph 4) Chemoproteomics provides residue-level insight and a handle to address question of AA importance mentioned in paragraph 1 and potential annotation source for future scores**
4. **(combine 5 and 6 ) Our goal is combining scores with chemoproteomics data to characterize cys and lys proteome-wide, challenges with integrating these things are ...., here is what we do in study, here is what we find from doing it**

Understanding how proteins work is the bedrock of functional biology and drug development. The identification of amino acid residues that 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 understanding their role in cellular networks and potential implication in disease states. Delineation of amino acid functionality is typically accomplished through mutagenesis studies, such as site-directed mutagenesis or saturation mutagenesis. While such studies have proven to be powerful validation tools, they are typically limited in scope, require assays based on prior knowledge of protein functionality, largely restricted to proteins easily expressed *in vitro*, and have yet to be extended genome-wide1.

Challenges with identifying amino acids critical to protein functional properties parallels a central challenge in human genetics with identifying pathogenic variants relevant to disease. The most abundant type of protein-altering variation in an individual’s genome is non-synonymous (missense) single nucleotide variants (nsSNVs), or single amino acid substitutions. Missense variants are historically the most difficult to interpret due to the wide range of possible outcomes, with extremes of no observable impact and causing severe disease, to milder consequences such as inhibiting, activating, stabilizing, or destabilizing proteins2. Because a majority of protein-coding variation is evolutionarily recent, deleterious alleles are enriched in the human population at very low allele frequency3, further complicating their interpretation in a clinical setting.

Diverse *in silico* algorithms are available to aid in the interpretation of sequence variants. Commonly used methods can be subdivided by type: conservation methods (GERP++4, phyloP5, phastCons6, and SiPhy7), function-prediction methods (FATHMM8, LRT9, MutationAssessor10, PolyPhen211, PROVEAN12, and SIFT13), or ensemble methods that use other methods and/or diverse genomic annotations as input (CADD14, REVEL15, M-CAP16, DANN17, and FATHMM-MKL18). Despite the availability of variant interpreters, a majority of missense are annotated as variant of unknown significance (VUS) in databases and mis-identified by scores, which typically perform best at distinguishing lower from higher effect consequences and fail to identify beneficial or moderate effect missense19. To narrow the gap between identifying population genetic variants and accurately annotation of mutations impact on human health, new methods are needed to probe the landscape of amino acid functionality proteome wide.

A frequently overlooked parameter associated with functional ‘hot-spots’ in the proteome is cysteine and lysine reactivity, which can fluctuate depending on the residue’s local and 3-dimensional protein microenvironment. Mass spectrometry-based chemoproteomics methods assay the intrinsic reactivity of thousands of amino acid side chains in native biological systems21–23. Using these methods, previous studies, including our own, revealed that "hyper-reactive" or pKa perturbed cysteine and lysine residues are enriched in known functional pockets21–23. These chemoproteomic methods can even be extended to measure the targetability or "druggability" of amino acid side chains, which has revealed a surprising number of cysteine and lysine side chains that can also be irreversibly labeled by small drug-like molecules. For the vast majority of these chemoproteomic-detected amino acids (CpDAA), the functional impact of chemical labeling remains unknown.

The integration of CpDAA datasets with variant pathogenicity predictors represents a new approach to stratify amino acid functionality and identify novel druggable and disease-associated sites proteome wide. Such multi-omic studies require mapping protein sequence positions to codon coordinates in the human reference genome. *In silico* reverse translation is typically achieved through searching proteomics data against genome-derived protein sequences. Improvements in sequencing technology and experimental data availability promote re-annotation efforts of reference sequences and ultimately the dynamic nature of biological sequence databases. This evolving nature of annotations confounds accurate multi-omics data integration, particularly for residue-level analyses. In the context of CpDAAs, residues identified in legacy datasets can be miss-identified or lost when these datasets are mapped to newer databases. These errors can stem from updates in the canonical or reference sequences, changes to sequences associated with stable identifiers, or discontinuation of legacy protein identifiers in later database releases. Complicating matters further, many missense predictors are based on annotations from human reference GRCh37, which was frozen in 2009. Mapping of UniProtKB proteins, which our lab uses for MS analysis, to genomic coordinates {McGarvey, 2018} is based on human reference GRCh38, which is continually updated. To circumvent the hassles of mapping data between frequently updating gene, transcript, and protein sequences, several tools have been developed to translate equivalent information between commonly used sequence databases such as Ensembl and UniProt43–45. However, the performance of tools is limited by the database mapping files they employ. For most tools, only one specific database version is used for mapping. Therefore, using such methods to accurately characterize CpDAA with annotations built from several databases and versions of sequences can be problematic.

Here we quantitatively assess the impact of mapping methods on residue-level data integration. By illuminating challenges associated with data compatibility and variability between several database release dates, we identify key sources of inaccurate mapping and provide guidelines for multi-omic data integration. Additionally, our study reveals that highly reactive cysteines are enriched for genetic variants that have high predicted pathogenicity (high deleteriousness), which supports general utility of using predictive scores to further power chemoproteomics to help decipher genetic variants. As many databases and sequencing pipelines update to GRCh38, our study should provide a roadmap for more precise mapping of annotations, which should have wide ranging applications for both the proteomics and genetics communities.

1. Cooper, G. M. & Shendure, J. Needles in stacks of needles : finding disease-causal variants in a wealth of genomic data. *Nat. Publ. Gr.* **12**, 628–640 (2011).

2. Samocha, K. E. *et al.* Regional missense constraint improves variant deleteriousness prediction. *bioRxiv* 148353 (2017). doi:10.1101/148353

3. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science (80-. ).* **336**, 64–69 (2012).

4. Davydov, E. V, Goode, D. L., Sirota, M., Cooper, G. M. & Sidow, A. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLoS Comput Biol* **6**, 1001025 (2010).

5. 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).

6. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).

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

8. Shihab, H. A. *et al.* Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. *Hum. Mutat.* **34**, 57–65 (2013).

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

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

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

12. 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**, (2012).

13. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1082 (2009).

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

15. 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).

16. 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).

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

18. Shihab, H. A. *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* **31**, 1536–1543 (2015).

19. Reeb, J., Wirth, T. & Rost, & B. Variant effect predictions capture some aspects of deep mutational scanning experiments. doi:10.1101/859603

20. Miosge, L. A. *et al.* Comparison of predicted and actual consequences of missense mutations. *Proc. Natl. Acad. Sci.* **112**, E5189–E5198 (2015).

21. Weerapana, E. *et al.* Quantitative reactivity profiling predicts functional cysteines in proteomes. *Nature* **468**, 790–797 (2010).

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

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