

UCLA

UCLA Electronic Theses and Dissertations

Title

Multi-omic stratification of the missense variant and redox-sensitive cysteinome

Permalink

<https://escholarship.org/uc/item/81w6f7hd>

Author

Desai, Heta

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Multi-omic stratification of the missense variant
and redox-sensitive cysteinome

A dissertation submitted in partial satisfaction of the
requirements for the degree

Doctor of Philosophy in Molecular Biology

by

Heta Sunil Desai

2023

© Copyright by

Heta Sunil Desai

2023

ABSTRACT OF THE DISSERTATION

Multi-omic stratification of the missense variant and redox-sensitive cysteinome

by

Heta Sunil Desai

Doctor of Philosophy in Molecular Biology

University of California, Los Angeles, 2023

Professor Keriann M. Backus, Chair

Cysteine-directed chemoproteomic profiling methods yield high-throughput inventories of redox-sensitive and ligandable cysteine residues. They are enabling techniques for functional biology. Due to their nucleophilicity and sensitivity to alkylation, cysteines have emerged as attractive sites to target with chemical probes. Cysteine-reactive covalent compounds can access small and poorly defined binding sites and efficiently block high-affinity interactions or compete with high concentrations of endogenous biomolecules. Furthermore, cysteine is the most frequently acquired amino acid due to missense variants in cancer databases. Acquired cysteines are both driver mutations and sites targeted by precision therapies; however, despite their ubiquity, nearly all acquired cysteines remain uncharacterized. Regardless of improvements in sample preparation workflows, cysteine chemoproteomic experiments still only sample a small fraction of the human cysteinome due to biological factors such as protein abundance, restricted protein

expression profiles, and technical factors such as unoptimized data analysis workflows not tailored to chemoproteomics, including database searches that do not sample the mutation-induced variant proteome. The cumbersome nature of these sample preparation workflows along with reagent costs hinder most chemoproteomics studies. In this work, we develop two new chemoproteomics platforms to enable high-throughput identification of redox sensitive and ligandable cysteines, including gain-of-cysteines. First, we tailor our single-pot, solid-phase-enhanced sample preparation (SP3) method to specifically probe the redox proteome, which showcases the utility of the SP3 platform in multistep sample-preparation workflows. Application of the SP3-Rox method to cellular proteomes identified cysteines sensitive to the oxidative stressor GSNO and cysteine oxidation state changes that occur during T cell activation. By implementing a customized workflow in the FragPipe computational pipeline, we achieve accurate MS1-based quantification, including for peptides containing multiple cysteine residues. We also present “chemoproteogenomics”, combining proteogenomics with established chemoproteomics methods to study human missense variation resulting in neo cysteine residues or mutations nearby cysteine residues. For both cancer and healthy genomes, we find that cysteine acquisition is a ubiquitous consequence of genetic variation that is further elevated in the context of decreased DNA repair. Our chemoproteogenomics platform integrates chemoproteomic, whole exome, and RNA-seq data, with a customized 2-stage false discovery rate (FDR) error controlled proteomic search enhanced with a user-friendly FragPipe interface to improve coverage of acquired cysteine variants and proximal variants using a panel of 11 cancer cell lines. These two established pipelines allow us to extend activity-based profiling methods, including small molecule screening

and redox-profiling, to gain-of-cysteine variants and cysteines proximal to variants. We expect widespread utility in guiding proteoform-specific biology and therapeutic discovery.

The dissertation of Heta Sunil Desai is approved.

Hilary A. Coller

Andrew Goldstein

James A. Wohlschlegel

Matteo Pellegrini

Keriann M. Backus, Committee Chair

University of California, Los Angeles

2023

TABLE OF CONTENTS

LIST OF FIGURES.....	viii
ACKNOWLEDGEMENTS	x
CIRRICULUM VITAE	xiii
Chapter 1: Introduction.....	1
References	9
Chapter 2: SP3-Enabled Rapid and High Coverage Chemoproteomic Identification of Cell-State–Dependent Redox-Sensitive Cysteines	16
Abstract	17
Introduction.....	16
Experimental Procedures	21
Results	28
Discussion	45
Supporting Information	53
References	104
Chapter 3: Multi-omic stratification of the missense variant cysteinome	118
Abstract	119
Introduction.....	120
Results	124
Discussion	158

Supporting Information	164
References	240
Chapter 4: Conclusion.....	256

LIST OF FIGURES

Chapter 2

Figure 1. Synthesis and evaluation of isotopic isopropyl iodoacetamide alkyne probes.	29
Figure 2. Quantification of IPIAA probes in various heavy to light ratios using IonQuant.	32
Figure 3. Application of SP3-Rox with FragPipe-IonQuant to identify oxidized cysteines.	37
Figure 4. Application of SP3-Rox to identify redox sensitive cysteines.	40
Figure 5. Application of SP3-Rox to identify redox changes during T cell maturation. ..	44

Chapter 3

Figure 1. Acquired cysteines are prevalent across cancer genomes, particularly for high missense burden cell lines	129
Figure 2. dMMR cell lines are enriched for rare, predicted deleterious gain-of-cysteine mutations.....	134
Figure 3. Variant peptide identification implementing an MSFragger-search pipeline.	137
Figure 4. Variant peptide identification on tumor cell lines	143
Figure 5. Comparison of variants identified from cysteine enrichment and bulk proteomics.....	148

Figure 6. Assessing ligandability of variant proximal cysteines and gain-of-cysteines..	
.....	152
Figure 7. Expanding HLA cysteine peptide coverage and gel-based ABPP of HLA covalent labeling	155
Figure 8. 2-stage search implemented into FragPipe GUI with Percolator rescoring ..	157

ACKNOWLEDGMENTS

This work would be impossible without the support of the Backus lab members. I want to thank my advisor, Dr. Keri Backus, for her enthusiasm and dedication to science. She goes above and beyond in making sure we have everything we need to carry out our work. I admire her broad scope of knowledge and hope I can leave the lab with even just a fraction of it. In the last 5 years, Keri has put together an excellent team of people: Jose Castellón, Dr. Maria Palafox, Dr. Sunny Yan, Lisa Boatner, Nik Burton, Ashley Julio, Andrew Becker, Flow Shikwana, Miranda Villanueva, Alex Turmon, Eli Bilech, Dr. Sho Takechi, Dr. Sam Ofori, Dr. Jian Cao. They are the kindest and friendliest bunch on campus, and I will deeply miss their company, whether its eating meals together, watching movies, talking science, or dropping whatever I'm doing to spend time with them.

I'm grateful for everyone that I was able to work with during graduate school. I want to thank the Nesvizhskii lab members that have been instrumental to this work. My deepest gratitude goes to Dr. Fengchao Yu and Dr. Alexey Nesvizhskii for helping with our proteomics search pipelines. I want to thank Dr. Valle Ojeda from the Merchant lab, who is such a patient and kind collaborator. I would also like to thank my committee members Dr. Hilary Coller, Dr. Andrew Goldstein, Dr. James A. Wohlschlegel, and Dr. Matteo Pellegrini for providing helpful feedback through my last three years and always being responsive to any of my questions. I want to acknowledge my home area director Dr. Feng Guo and the MBI and Biological Chemistry administration.

During grad school, I've come to recognize how vital dance is to me and any of my achievements. First, I need to thank Samahang Modern for providing a home-away-from-home during my first year living across the country and for filling in my dance gap during

quarantine. I'm super thankful for the memories, culture, and friendship we shared during this short but very intense time together. Second, I need to thank *the* dance. I'm immensely grateful to have trained with Joseph Wiggan during the last 4 years. Every session, I've him watched him tap dance and it always blows my mind. I hope to carry everything I've learned from him in this lifelong passion.

I have a few more acknowledgments. Giving up a car in LA was certainly a choice. Thanks to people giving me car rides (mostly Nik, also Jose, and sometimes Lisa) and bus drivers for getting me to campus to carry out this work. Also, a very special shout out to the Burnside crew (Billy, Lisa, Nik, Michael) for all the good times we had.

The biggest of all thanks goes to my family (Sushma Desai, Sunil Desai, Dr. Jigar Desai, Baa, Dada, and Covlie) for their support since the beginning. I want to thank Covlie for being the cutest Slack photo ever, and I want to thank my brother, Jigar, for being my science inspiration.

Chapter 1 is a version of Desai, H. S.; Yan, T.; Backus, K. M. SP3-FAIMS-Enabled High-Throughput Quantitative Profiling of the Cysteinome. *Current Protocols* **2022**, 2 (7), e492. <https://doi.org/10.1002/CPZ1.492> Reprinted with permission. © 2022 Wiley Periodicals LLC. TY and HD co-authored this protocol paper with K. M. B. supervising.

Chapter 2 is a version of Desai, H. S.; Yan, T.; Yu, F.; Sun, A. W.; Villanueva, M.; Nesvizhskii, A. I.; Backus, K. M. SP3-Enabled Rapid and High Coverage Chemoproteomic Identification of Cell-State–Dependent Redox-Sensitive Cysteines. *Molecular & Cellular Proteomics* **2022**, 21 (4), 100218.

<https://doi.org/10.1016/J.MCPRO.2022.100218> © 2022 The Authors. Published by Elsevier Inc on behalf of American Society for Biochemistry and Molecular Biology. This version is made available under the CC-BY-NC-ND 4.0 license. H. S. D. and K. M. B. conceptualization; H. S. D., T. Y., and M. V. investigation; H. S. D. and T. Y. methodology; H. S. D. formal analysis; H. S. D. validation; H. S. D. and F. Y. data curation; H. S. D. visualization; H. S. D. and K. M. B writing–original draft; H. S. D., A. W. S., A. I. N., and K. M. B. writing–review and editing; A. W. S. resources; A. I. N. and K. M. B. supervision; A. I. N. and K. M. B. funding acquisition.

Chapter 3 is a version of Desai, H. S.; Ofori, S.; Boatner, L. M.; Yu, F.; Villaneuva, M.; Ung, N.; Nesvizhskii, A.; Backus, K. M. Multi-Omic Stratification of the Missense Variant Cysteine. *bioRxiv* **2023**. <https://doi.org/10.1101/2023.08.12.553095>. H. S. D., K. M. B., and A.I.N. conceptualization; H. S. D. formal analysis; H. S. D. visualization; H. S. D. validation; H. S. D., L.M.B., F. Y., K.M.B data curation; H. S. D., S. O., and M. V. investigation; H. S. D., F. Y., and N.U. methodology; H. S. D. and K. M. B writing–original draft; H. S. D., S.O., L.M.B., F.Y., A. I. N., and K. M. B. writing–review and editing; A. I. N. and K. M. B. supervision; A. I. N. and K. M. B. funding acquisition.

CIRRICULUM VITAE

EDUCATION

- 2018—present PhD | University of California Los Angeles (UCLA)
Molecular Biology Interdepartmental Doctoral Program
- 2014—2018 BS | University of North Carolina at Chapel Hill (UNC-Chapel Hill)
Chemistry

PUBLICATIONS

Desai, H.S., Ofori, S, Boatner, L.M., Yu, F. Villaneuva, M., Ung, N., Nesvizhskii, A., Backus K.M. (2023). Multi-Omic Stratification of the Missense Variant Cysteinome. *bioRxiv*, <https://doi.org/10.1101/2023.08.12.553095>

Yan, T., Julio, A. R., Villanueva, M., Jones, A. E., Ball, A. B., Boatner, L. M., Turmon, A. C., Nguyễn, K. B., Yen, S. L., **Desai, H. S.**, Divakaruni, A. S., Backus, K. M. (2023). Proximity-Labeling Chemoproteomics Defines the Subcellular Cysteinome and Inflammation-Responsive Mitochondrial Redoxome. *Cell Chemical Biology* 30 (7), 811-827.e7

Desai, H.S., Yan, T., & Backus, K. M. (2022). SP3-FAIMS-Enabled High-Throughput Quantitative Profiling of the Cysteinome. *Current Protocols*, 2(7), e492

Desai, H.S., Yan, T., Yu F., Sun A., Villanueva M., Nesvizhskii A., & Backus, K.M. (2022). SP3-enabled Rapid and High Coverage Chemoproteomic Identification of Cell-State Dependent Redox-Sensitive Cysteines. *Molecular & Cellular Proteomics*, 21(4), 100218

Cao, J., Boatner, L. M., **Desai, H. S.**, Burton, N. R., Armenta, E., Chan, N. J., Castellón, J. O., & Backus, K. M. (2021). Multiplexed CuAAC Suzuki–Miyaura Labeling for Tandem Activity-Based Chemoproteomic Profiling. *Analytical Chemistry*, 93(4), 2610–2618

Yan, T., **Desai, H. S.**, Boatner, L. M., Yen, S. L., Cao, J., Palafox, M. F., Jami-Alahmadi, Y., & Backus, K. M. (2021). SP3-FAIMS Chemoproteomics for High-Coverage Profiling of the Human Cysteinome. *ChemBioChem*, cbic.202000870

Palafox, M. F., **Desai, H. S.**, Arboleda, V. A., & Backus, K. M. (2021). From chemoproteomic-detected amino acids to genomic coordinates: insights into precise multi-omic data integration. *Molecular Systems Biology*, 17(2), e9840

TEACHING EXPERIENCE

Winter 2021 TA

Biomedical Research: Essential Skills and Concepts (BMD RES 5HB)

Led one 2 hr discussion and one office hour per week; graded weekly problem sets; reviewed the scientific method--formulating questions hypotheses, and specific aims--introduced writing research proposals and reviewed biochemistry techniques

Winter 2020 TA

Intro to Cell Biology (MCDB 165A)

Led three 1hr discussions and two office hours per week; graded exams and weekly quizzes, reviewed cell biology journal articles and experimental techniques

AWARDS AND FELLOWSHIPS

- Summer Mentored Research Fellowship (SMRF), UCLA **2021**
- Sigman Symposium Poster Award, UCLA **2021**
- Research Experience for Undergraduates (NSF-REU), Syracuse University **2017**
- Carolina Research Scholar Award, UNC **2017**
- Summer Undergraduate Fellowship (SURF), UNC **2016**

OUTREACH

2020-2023 AWiSE (Advancing Women in Science and Engineering) Member

- Judge for Annual Portola Middle School Science Fair, evaluated middle school science fair projects
- Judge for Annual ENVISION international research competition for female and nonbinary high school students, evaluated scientific project proposals

2019-2022 UCLA Molecular Biology Department

- Panel member for McNair Scholars Research Conference graduate student panels
- MBIDP Peer Mentorship Program, maintained contact with incoming graduate student
- Panel member for MBIDP Transitioning into Graduate School panels

2016-2018 UNC-Learning Center

- Peer tutor for organic chemistry courses
- Summer Undergraduate Research Fellowship Writing Advisor; helped students put together research proposals for SURF fellowships

Chapter 1: Introduction

Distinguished by their nucleophilicity, sensitivity to oxidative stress, propensity to coordinate metals, numerous post-translational modifications, and ability to form disulfides, cysteine residues play key roles in the structure and function of most human proteins. Nearly all proteins contain at least one cysteine, and in aggregate, the human genome encodes ~262,000 cysteines (Yan et al., 2021). Many bioactive compounds and drugs have been developed that react with specific cysteine residues. For example, the blockbuster covalent kinase inhibitor therapeutics afatinib and ibrutinib, together with second-generation molecules, function by targeting non-catalytic cysteines in the active sites of the kinases EGFR and BTK, respectively (Liu et al., 2013; Singh, Petter, Baillie, & Whitty, 2011; W. Zhou et al., 2009). More recently, the Gly12Cys-mutant form of KRAS, an oncoprotein long thought to be “undruggable,” has been successfully targeted by a number of cysteine-reactive compounds, including the US Food and Drug Administration–approved drug sotorasib (Canon et al., 2019; Lim et al., 2014; Ostrem, Peters, Sos, Wells, & Shokat, 2013). Consequently, the identification and functional characterization of the “cysteinome” is a central challenge for both functional biology and drug development efforts.

Mass spectrometry (MS)-based chemoproteomics is an enabling technology capable of rapidly identifying functional and “druggable” residues proteome-wide (Backus et al., 2016; Hacker et al., 2017; Kuljanin et al., 2021; Vinogradova et al., 2020; Weerapana et al., 2010). Among all amino acids that can be captured by chemical probes, cysteine is unique given its nucleophilicity, sensitivity to oxidative modifications, and

therapeutic relevance, as showcased by the number of drugs and clinical candidates that function through cysteine covalent modification. Although recent work, including our own studies, has showcased a range of technical innovations that together have substantially increased the throughput and coverage of chemoproteomics workflows, several important opportunities and challenges remain for cysteine-directed chemoproteomics studies.

First, as our work has shown, such experiments still sample only a small fraction of the total cysteinome (~10-15%; Yan et al., 2021). Achieving more complete coverage is an essential step to fully harness the therapeutic and mechanistic relevance of the cysteinome. Second, cysteine-directed chemoproteomic workflows remain laborious, are low throughput, and often require relatively large amounts of input material (e.g., 2 mg proteome/per sample used in our prior studies (Backus et al., 2016; Palafox, Desai, Arboleda, & Backus, 2021), which can preclude the analysis of rare cell populations and clinical samples. Lastly, the data analysis pipelines for quantifying labeling are not standardized and can be challenged by false-positive values, prohibitive cost, and inaccessibility to those beyond the pipeline's developers. Standard cysteine-directed chemoproteomics platforms (Backus, 2019) all rely on the same general workflows with minor modifications. First, cells or lysates are treated with cysteine modifier of interest (e.g., induction of oxidative stressor or cysteine-reactive compound or vehicle) and then labeled with a pan-cysteine-reactive probe (e.g., iodoacetamide, maleimide, or other, more tailored electrophiles that are coupled to an enrichment handle, such as biotin or an alkyne/azide moiety that can be biotinylated by "click" chemistry). Biotinylated samples are then subjected to enrichment on avidin resin and sequence specific proteolysis. Labeled peptides are eluted from the resin and analyzed by liquid chromatography -

tandem mass spectrometry (LC-MS/MS). In their most simplistic form, these studies will report spectral counts for the peptides containing the specific cysteines modified by a probe (Weerapana, Speers, & Cravatt, 2007). Such semiquantitative approaches, although useful for identifying modification sites, have, however, been largely supplanted by precursor-ion mass spectrometry (MS1)- and tandem mass spectrometry (MS2)-based quantification of fractional occupancy afforded by a capping step with a pan-cysteine-reactive probe (Kuljanin et al., 2021; Weerapana et al., 2010).

For example, in the widely adopted isoTOP-ABPP method, after labeling with iodoacetamide alkyne (IAA), treated and control samples are conjugated via copper-catalyzed azide-alkyne cycloadditions (CuAAC), or “click” chemistry, to isotopically labeled (“heavy” and “light”) tobacco etch virus (TEV)-protease-cleavable biotinylated peptide enrichment reagents (TEV tags). The heavy and light labeled samples are pooled and subjected to enrichment on streptavidin resin, and then undergo sequential proteolysis with trypsin and TEV proteases. LC-MS/MS analysis of the TEV-protease-eluted fraction reports the relative labeling of treated and control samples by IAA on the basis of changes to the relative MS1 chromatographic peak areas. Various modified versions of this workflow have subsequently been developed that incorporate modified enrichment handles. Examples include the desthiobiotin-based tags (Zanon, Lewald, & Hacker, 2020), as well as chemical and photocleavable tags (Fu et al., 2020; Qian & Weerapana, 2017; Rabalski, Bogdan, & Baranczak, 2019). Integration of isobaric tagging strategies that rely upon commercially available tandem mass tags (TMT) have recently extended these chemoproteomic workflows from MS1- to MS2-based quantification (TMT-based methods; Kuljanin et al., 2021; Mnatsakanyan et al., 2019; Vinogradova et

al., 2020). A key advantage of MS2-based quantification is increased sample throughput afforded by up to 18-plex sample multiplexing (Li et al., 2021). For both MS1- and MS2-based quantification, a key, and still largely unmet, limitation is the interpretation of so-called singleton ions, meaning ions in which only either the light or the heavy precursor can be identified at the MS1 level. Although these ions can correspond to peptides with highly elevated ratios consistent with a large fold change between treated and control samples (e.g., a high-occupancy labeling event), they should be treated with an additional level of statistical rigor because of the increased propensity for irreproducible ratios observed for these singleton species.

Our research revealed several additional key challenges and limitations that are ubiquitous in chemoproteomics studies. One key problem is achieving efficient biotinylation, which is essential for obtaining high-coverage samples. We have found that widely used click chemistry conditions afford only ~30-40% yields of labeled peptides, leaving a substantial fraction of alkyne-modified peptides unconjugated (Yan et al., 2021). Our research also revealed that using a modest 2-fold excess of the biotin-azide reagent relative to the alkyne probe, together with a 10-fold increase in the concentrations of both reagents (e.g., from 200 μ M IAA and 400 μ M biotin-azide to 2 mM IAA and 4 mM biotin-azide), afforded >80% yields for biotinylation.

A second key challenge for these protocols is the relatively poor solubility of most biotin reagents in both aqueous and organic solvent systems. This can limit the absolute concentrations of reagents used for labeling; for example, the aforementioned example of increased biotin-azide affording increased biotinylation is not possible with some of the more elaborated reagents (e.g., TEV tags). More problematic still is efficient

decontamination of samples to remove excess biotin reagents. Our recent work has shown that the choice of desalting method (e.g., protein precipitation or solid-phase extraction) can dramatically affect the yield of peptides after affinity enrichment. Our comparison of standard chloroform/methanol precipitation to the recently developed single-pot solid-phase sample preparation (SP3) method revealed a marked increase in both PSMs and unique peptides for samples processed when using SP3 (Yan et al., 2021). This finding is particularly striking for samples that require a larger excess of biotin reagent to achieve efficient labeling—for example, our Suzuki-Miyaura chemoproteomic method (Cao et al., 2021), which requires a large excess of biotin–boronic acid coupling partner. Contamination of samples with excess biotin reagents can result in poor recovery of labeled peptides and low-coverage datasets.

Data analysis is another area that remains challenging and non-standardized for the cysteine chemoproteomics field. A number of different quantification pipelines have been developed or applied to cysteine chemoproteomics. For search, multiple algorithms including SEQUEST (Eng, McCormack, & Yates, 1994), ProLuCiD (Xu et al., 2015), MaxQuant (Tyanova, Temu, & Cox, 2016), Proteome Discoverer (Mnatsakanyan et al., 2019), and MSFragger (Kong, Lprevost, Avtonomov, Mellacheruvu, & Nesvizhskii, 2017; Yu, Haynes, et al., 2020; Yu, Teo, et al., 2020) have been found to be compatible with cysteine chemoproteomics data. MS1-based quantification has been achieved using CENSUS (Park, Venable, Xu, & Yates, 2008), CIMAGE (Gao et al., 2021), Perseus (Tyanova et al., 2016), IonQuant (Yu et al., 2020; Yu, Haynes, & Nesvizhskii, 2021), and Skyline (MacLean et al., 2010). MS2- and MS3-based quantification has been reported using ProLuCiD together with the Integrated Proteomics Pipeline (IP2; Vinogradova et

al., 2020). Real-time, search-enabled, MS3-based mass spectrometry (RTS-SPS-MS3) analysis of TMT-labeled cysteine chemoproteomic samples was achieved using Orbiter together with Comet (Schweppe et al., 2020; <https://uwpr.github.io/Comet/>). Differences between the ways that these packages calculate ratios, for example by area under the curve, linear regression, ion intensity, or some combination thereof, can substantially affect the ratios calculated. The prohibitive cost of some of these softwares is another challenge that limits their adoption. Additionally, lack of flexibility in how peptides containing multiple different modifications are treated, for example, with two cysteines where one is carbamidomethylated and one contains a heavy or light label, can hinder accurate analysis for multi-cysteine-containing peptides. Both MaxQuant and Skyline suffer from this limitation. Our recent work yielded a custom plugin-for Skyline that enables such quantification (<https://proteome.gs.washington.edu/~nicksh/kbackus/AddLabelType/setup.exe>).

It is also worth highlighting several areas in which recent advances in instrumentation and data acquisition have improved chemoproteomics workflows. First, the advent of new instruments with improved acquisition speeds has dramatically increased the coverage of modified peptides achieved in single-shot experiments. Second, the advent of MS3-based quantification, particularly when paired with RTS, has improved the detection of peptides labeled with isobaric reagents, and has substantially decreased the ratio compression observed with these samples at the MS2 level (Schweppe et al., 2020). Incorporation of a field asymmetric ion mobility (FAIMS) device has substantially increased coverage of cysteine-containing peptides—of note, FAIMS does afford mild ratio compression, which should be accounted for during data

processing. Lastly, incorporation of labile-ion search tools in MSFragger has enabled the identification of diagnostic or signature ions for peptides enriched in chemoproteomics experiments, innovations that enabled further improved coverage of chemoproteomics experiments (Yan et al., 2022).

Chemoproteomics has emerged as an enabling technology capable of assaying the reactivity, redox sensitivity, and ligandability of cysteines proteome-wide. Small-molecule chemoproteomic screens using libraries of electrophilic compounds have mapped hundreds of novel cysteine-ligand interactions, which can serve as starting points for the development of new selective probes and even. Cysteine chemoproteomic methods began with the ICAT method pioneered by Gygi and coworkers (Gygi et al., 1999). The OxICAT method, developed by Leichert et al. (2008), applied the isotopically differentiated ICAT probes in a sequential labeling workflow tailored to quantify cysteine oxidation (García-Santamarina et al., 2014). A number of more recent methods that assay cysteine oxidation state have been developed, including Cys-Boost (Mnatsakanyan et al., 2019), QTRP (Fu et al., 2020), and methods that rely on isotopically differentiated iodoacetamide alkyne probes, including one from Abo and co-workers (Abo et al., 2018). Alongside this plethora of redox proteomics methods, a comparable number of chemoproteomic platforms capable of assaying cysteine ligandability have emerged. The use of alternative enrichment reagents, such as chemically cleavable and photocleavable enrichment tags, have improved the compatibility of these methods with alternative sequence specific proteases. Incorporation of isobaric tags into these workflows, including for example in the streamlined cysteine activity-based protein

profiling (SLC-ABPP) method, have enabled increased sample throughput via multiplexed MS2- or MS3-level quantification (Kuljanin et al., 2021).

Despite the continuous improvements in instrumentation and sample preparation workflows, cysteine chemoproteomic experiments still only sample a small fraction of the human cysteinome. Reasons for this gap include biological factors such as protein abundance, restricted protein expression profiles or structural unavailability, e.g., due to disulfides; and also technical factors such as inefficient biotinylation, loss of sample during cleanup, unoptimized acquisition workflows, or data analysis workflows not tailored to chemoproteomics, including non-sample specific database searches that do not sample the mutation-induced variant proteome.

As we have achieved substantial expansions of cysteine coverage, my work extends and adapts our improved methodology to new redox proteomics applications, described in Chapter 2, through optimization of experimental sample preparation and data analysis workflows. Detailed in Chapter 3, I establish a chemoproteogenomics platform that combines genomics data with our chemoproteomics data to identify the *hidden* landscape of cysteine acquisition.

References

- Abo, M., Li, C., & Weerapana, E. (2018). Isotopically-labeled iodoacetamide-alkyne probes for quantitative cysteine-reactivity profiling. *Molecular Pharmaceutics*, 15(3), 743-749. doi: 10.1021/acs.molpharmaceut.7b00832
- Backus, K.M. (2019). Applications of reactive cysteine profiling. *Current Topics in Microbiology and Immunology*, 420, 375–417. doi: 10.1007/82_2018_120
- Backus, K. M., Correia, B. E., Lum, K. M., Forli, S., Horning, B. D., González-Páez, G. E., ... Cravatt, B. F. (2016). Proteome-wide covalent ligand discovery in native biological systems. *Nature*, 534(7608), 570–574. doi: 10.1038/nature18002
- Canon, J., Rex, K., Saiki, A. Y., Mohr, C., Cooke, K., Bagal, D., ... Lipford, J. R. (2019). The clinical KRAS(G12C) inhibitor AMG 510 drives anti-tumour immunity. *Nature*, 575(7781), 217–223. doi: 10.1038/S41586-019-1694-1
- Cao, J., Boatner, L. M., Desai, H. S., Burton, N. R., Armenta, E., Chan, N. J., ... Backus, K. M. (2021). Multiplexed CuAAC Suzuki–Miyaura labeling for tandem activity-based chemoproteomic profiling. *Analytical Chemistry*, 93(4), 2610–2618. doi: 10.1021/acs.analchem.0c04726
- Eng, J. K., McCormack, A. L., & Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, 5(11), 976–989. doi: 10.1016/1044-0305(94)80016-2
- Fu, L., Li, Z., Liu, K., Tian, C., He, J., He, J., ... Yang, J. (2020). A quantitative thiol reactivity profiling platform to analyze redox and electrophile reactive cysteine

proteomes. *Nature Protocols*, 15(9), 2891–2919. doi: 10.1038/s41596-020-0352-2

Gao, J., Liu, Y., Yang, F., Chen, X., Cravatt, B.F., & Wang, C. (2021). CIMAGE2.0: An expanded tool for quantitative analysis of activity-based protein profiling (ABPP) data. *Journal of Proteome Research*, 20(10), 4893–4900. doi: 10.1021/ACS.JPROTEOME.1C00455

García-Santamarina, S., Boronat, S., Domènech, A., Ayté, J., Molina, H., & Hidalgo, E. (2014). Monitoring in vivo reversible cysteine oxidation in proteins using ICAT and mass spectrometry. *Nature Protocols*, 9(5), 1131–1145. doi: 10.1038/nprot.2014.065

Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., & Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, 17(10), 994–999. doi: 10.1038/13690

Hacker, S. M., Backus, K. M., Lazear, M. R., Forli, S., Correia, B. E., & Cravatt, B. F. (2017). Global profiling of lysine reactivity and ligandability in the human proteome. *Nature Chemistry*, 9(12), 1181–1190. doi: 10.1038/nchem.2826

Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., & Nesvizhskii, A. I. (2017). MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nature Methods*, 14(5), 513–520. doi: 10.1038/nmeth.4256

Kuljanin, M., Mitchell, D. C., Schweppe, D. K., Gikandi, A. S., Nusinow, D. P., Bulloch, N.J., ... Gygi, S. P. (2021). Reimagining high-throughput profiling of reactive

- cysteines for cell-based screening of large electrophile libraries. *Nature Biotechnology*, 39(5), 630–641. doi: 10.1038/s41587-020-00778-3
- Leichert, L. I., Gehrke, F., Gudiseva, H. V., Blackwell, T., Ilbert, M., Walker, A. K., ... Jakob, U. (2008). Quantifying changes in the thiol redox proteome upon oxidative stress in vivo. *Proceedings of the National Academy of Sciences of the United States of America*, 105(24), 8197–8202. doi: 10.1073/pnas.0707723105
- Li, J., Cai, Z., Bomgarden, R. D., Pike, I., Kuhn, K., Rogers, J. C., ... Paulo, J. A. (2021). TMTpro-18plex: The expanded and complete set of TMT-pro reagents for sample multiplexing. *Journal of Proteome Research*, 20(5), 2964–2972. doi: 10.1021/ACS.JPROTEOME.1C00168
- Lim, S. M., Westover, K. D., Ficarro, S. B., Harrison, R. A., Choi, H. G., Pacold, M. E., ... Gray, N. S. (2014). Therapeutic targeting of oncogenic K-ras by a covalent catalytic site inhibitor. *Angewandte Chemie International Edition*, 53(1), 199–204. doi: 10.1002/ANIE.201307387
- Liu, Q., Sabnis, Y., Zhao, Z., Zhang, T., Buhrlage, S. J., Jones, L. H., & Gray, N. S. (2013). Developing irreversible inhibitors of the protein kinase cysteinome. *Chemistry & Biology*, 20(2), 146–159. doi: 10.1016/J.CHEMBIOL.2012.12.006
- MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., ... MacCoss, M. J. (2010). Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics*, 26(7), 966–968. doi: 10.1093/BIOINFORMATICS/BTQ054
- Mnatsakanyan, R., Markoutsas, S., Walbrunn, K., Roos, A., Verhelst, S. H. L., & Zahedi, R. P. (2019). Proteome-wide detection of S-nitrosylation targets and motifs using

- biorthogonal cleavable-linker-based enrichment and switch technique. *Nature Communications*, 10(1), 1–12. doi: 10.1038/s41467-019-10182-4
- Ostrem, J.M., Peters, U., Sos, M. L., Wells, J. A., & Shokat, K. M. (2013). K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nature*, 503(7477), 548–551. doi: 10.1038/nature12796
- Palafox, M. F., Desai, H. S., Arboleda, V. A., & Backus, K. M. (2021). From chemoproteomic-detected amino acids to genomic coordinates: Insights into precise multi-omic data integration. *Molecular Systems Biology*, 17, e9040. doi: 10.15252/msb.20209840
- Park, S. K., Venable, J. D., Xu, T., & Yates, J. R. (2008). A quantitative analysis software tool for mass spectrometry–based proteomics. *Nature Methods*, 5(4), 319–322. doi: 10.1038/nmeth.1195
- Qian, Y., & Weerapana, E. (2017). A quantitative mass-spectrometry platform to monitor changes in cysteine reactivity. *Methods in Molecular Biology*, 1491, 11–22. doi: 10.1007/978-1-4939-6439-0_2
- Rabalski, A. J., Bogdan, A. R., & Baranczak, A. (2019). Evaluation of chemically-cleavable linkers for quantitative mapping of small molecule-cysteinome reactivity. *ACS Chemical Biology*, 14(9), 1940–1950. doi: 10.1021/ACSCHEMBIO.9B00424
- Schweppe, D. K., Eng, J. K., Yu, Q., Bailey, D., Rad, R., Navarrete-Perea, J., ... Gygi, S. P. (2020). Full-featured, real-time database searching platform enables fast and accurate multiplexed quantitative proteomics. *Journal of Proteome Research*, 19(5), 2026–2034. doi: 10.1021/ACS.JPROTEOME.9B00860

- Singh, J., Petter, R. C., Baillie, T. A., & Whitty, A. (2011). The resurgence of covalent drugs. *Nature Reviews Drug Discovery*, 10(4), 307–317. doi: 10.1038/nrd3410
- Tyanova, S., Temu, T., & Cox, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols*, 11(12), 2301–2319. doi: 10.1038/nprot.2016.136
- Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., ... Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods*, 13(9), 731–740. doi: 10.1038/nmeth.3901
- Vinogradova, E. V., Zhang, X., Remillard, D., Lazar, D. C., Suciu, R. M., Wang, Y., ... Cravatt, B. F. (2020). An activity-guided map of electrophile-cysteine interactions in primary human T cells. *Cell*, 182(4), 1009–1026.e29. doi: 10.1016/J.CELL.2020.07.001
- Weerapana, E., Speers, A. E., & Cravatt, B. F. (2007). Tandem orthogonal proteolysis-activity-based protein profiling (TOP-ABPP)—a general method for mapping sites of probe modification in proteomes. *Nature Protocols*, 2(6), 1414–1425. doi: 10.1038/nprot.2007.194
- Weerapana, E., Wang, C., Simon, G. M., Richter, F., Khare, S., Dillon, M. B. D., ... Cravatt, B.F. (2010). Quantitative reactivity profiling predicts functional cysteines in proteomes. *Nature*, 468(7325), 790–797. doi: 10.1038/nature09472
- Xu, T., Park, S. K., Venable, J. D., Wohlschlegel, J. A., Diedrich, J. K., Cociorva, D., ... Yates, J. R. (2015). ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity. *Journal of Proteomics*, 129, 16–24. doi: 10.1016/J.JPROT.2015.07.001

- Yan, T., Desai, H. S., Boatner, L. M., Yen, S.L., Cao, J., Palafox, M. F., ... Backus, K. M. (2021). SP3-FAIMS chemoproteomics for high coverage profiling of the human cysteinome. *ChemBioChem*, 22, 1841. doi: 10.1002/cbic.202000870
- Yan, T., Palmer, A. B., Geiszler, D. J., Polasky, D.A., Boatner, L. M., Burton, N. R., ... Backus, K. M. (2022). Enhancing cysteine chemoproteomic coverage through systematic assessment of click chemistry product fragmentation. *Analytical Chemistry*, 94(9), 3800–3810. doi: 10.1021/acs.analchem.1c04402
- Yu, F., Haynes, S. E., & Nesvizhskii, A. I. (2021). IonQuant enables accurate and sensitive label-free quantification with FDR-controlled match between-runs. *Molecular and Cellular Proteomics*, 20, 100077. doi: 10.1016/J.MCPRO.2021.100077
- Yu, F., Haynes, S. E., Teo, G. C., Avtonomov, D. M., Polasky, D. A., & Nesvizhskii, A. I. (2020). Fast quantitative analysis of timsTOF PASEF data with MSFragger and IonQuant. *Molecular and Cellular Proteomics*, 19(9), 1575–1585. doi: 10.1074/MCP.TIR120.002048
- Yu, F., Teo, G. C., Kong, A. T., Haynes, S. E., Avtonomov, D. M., Geiszler, D. J., & Nesvizhskii, A. I. (2020). Identification of modified peptides using localization-aware open search. *Nature Communications*, 11(1), 1–9. doi: 10.1038/s41467-020-17921-y
- Zanon, P. R. A., Lewald, L., & Hacker, S. M. (2020). Isotopically labeled desthiobiotin azide (isoDTB) tags enable global profiling of the bacterial cysteinome. *Angewandte Chemie*, 132(7), 2851–2858. doi: 10.1002/ANGE.201912075

Zhou, W., Ercan, D., Chen, L., Yun, C. H., Li, D., Capelletti, M., ... Jänne, P. A. (2009).
Novel mutant-selective EGFR kinase inhibitors against EGFR T790M. *Nature*,
462(7276), 1070–1074. doi: 10.1038/nature08622

Chapter 2: SP3-Enabled Rapid and High Coverage Chemoproteomic Identification of Cell-State–Dependent Redox-Sensitive Cysteines

Chapter 2 is a version of Desai, H. S.; Yan, T.; Yu, F.; Sun, A. W.; Villanueva, M.; Nesvizhskii, A. I.; Backus, K. M. SP3-Enabled Rapid and High Coverage Chemoproteomic Identification of Cell-State–Dependent Redox-Sensitive Cysteines. *Molecular & Cellular Proteomics* **2022**, 21 (4), 100218.

<https://doi.org/10.1016/J.MCPRO.2022.100218> © 2022 The Authors. Published by Elsevier Inc on behalf of American Society for Biochemistry and Molecular Biology. This version is made available under the CC-BY-NC-ND 4.0 license.

Heta S. Desai^{1,2}, Tianyang Yan^{1,3}, Fengchao Yu⁴, Alexander W. Sun¹, Miranda Villanueva^{1,2}, Alexey I. Nesvizhskii^{5,4}, and Keriann M. Backus^{1,3,2,6,7,8,*}

¹Biological Chemistry Department, David Geffen School of Medicine, ²Molecular Biology Institute, and ³Department of Chemistry and Biochemistry, UCLA, Los Angeles, California, USA; ⁴Department of Pathology, and ⁵Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA; ⁶DOE Institute for Genomics and Proteomics, ⁷Jonsson Comprehensive Cancer Center, and ⁸Eli and Edy the Broad Center of Regenerative Medicine and Stem Cell Research, UCLA, Los Angeles, California, USA *For correspondence: Keriann M. Backus, kbackus@mednet.ucla.edu.

Abstract

Proteinaceous cysteine residues act as privileged sensors of oxidative stress. As reactive oxygen and nitrogen species have been implicated in numerous pathophysiological processes, deciphering which cysteines are sensitive to oxidative modification and the specific nature of these modifications is essential to understanding protein and cellular function in health and disease. While established mass spectrometry-based proteomic platforms have improved our understanding of the redox proteome, the widespread adoption of these methods is often hindered by complex sample preparation workflows, prohibitive cost of isotopic labeling reagents, and requirements for custom data analysis workflows. Here, we present the SP3-Rox redox proteomics method that combines tailored low cost isotopically labeled capture reagents with SP3 sample cleanup to achieve high throughput and high coverage proteome-wide identification of redox-sensitive cysteines. By implementing a customized workflow in the free FragPipe computational pipeline, we achieve accurate MS1-based quantitation, including for peptides containing multiple cysteine residues. Application of the SP3-Rox method to cellular proteomes identified cysteines sensitive to the oxidative stressor GSNO and cysteine oxidation state changes that occur during T cell activation.

Introduction

Oxidative stress plays an essential role in human health and disease. Both reactive oxygen species (ROS) and reactive nitrogen species (RNS) have been implicated in many pathophysiological processes, including cancers, neurodegenerative disorders, and atherosclerosis (1, 2, 3). Oxidative stress also plays an essential role in the

modulation of innate and adaptive immune responses, with abnormal cellular activation occurring at both hypo- and hyper-levels of ROS and RNS (4, 5, 6). Consequently, the identification of mechanisms by which cells respond to oxidative stressors is an essential step to improving the treatment and prevention of a wide range of human disorders.

Inherently nucleophilic and sensitive to oxidative stress, proteinaceous cysteine residues function as key sensors of ROS and RNS through oxidative modifications, including disulfide formation, S-nitrosation, and sulfenylation (3, 7). Therefore, the proteome-wide identification of redox-sensitive cysteine residues has emerged as a useful strategy to gain insight into cellular response to oxidative stress (8). Numerous proteomic methods, including OxiCat, Biotin Switch, and modifications to these methodologies (8, 9, 10, 11, 12, 13, 14) have enabled the high throughput identification of cysteines sensitive to oxidative modifications. Exemplifying these technical innovations, recent application of isobaric mass tagging coupled with immobilized metal affinity chromatography to the organism-wide identification of redox-sensitive cysteines identified ~34,000 unique cysteines across 10 murine tissues (15).

Most of these platforms rely upon the same general workflow: first, all reduced cysteines are capped using a cysteine reactive probe such as iodoacetamide (IA). After removal of all excess IA, oxidized cysteines are then reduced by the application of reducing agents tailored to the oxidized species of interest (e.g., gentler reductant such as sodium ascorbate for selective identification of nitrosylated cysteines or stronger reducing agents such as tris(2-carboxyethyl)phosphine (TCEP) or DTT for identification of all oxidized cysteines, including disulfides and sulfenylated cysteines) (11, 16, 17, 18). Newly liberated thiol side chains are then capped using a second cysteine-reactive

electrophile, typically one that features an enrichment handle (e.g., iodoacetamide alkyne (IAA) or iodoacetamide-desthiobiotin). After enrichment and proteolytic digestion, oxidized cysteines are then identified using standard LC-MS/MS analysis workflows. Incorporation of isotopic labeling strategies, including stable isotope labeling by amino acids in cell culture and other isotopic labeling reagents, such as isotope-coded affinity tag reagents, tandem mass tag labeling reagents, isobaric tags for relative and absolute quantification, and heavy- and light-IAA, enables inter- and intra-sample quantitation of relative and absolute cysteine oxidation (9, 15, 19, 20, 21, 22, 23).

While already widely adopted, redox proteomic methods suffer from several shared limitations. Reliance on costly isotopically labeled reagents has made large scale redox studies cost-prohibitive for many groups. Efficient removal of excess cysteine-labeling reagent, which is essential to achieving high fidelity identification of redox-sensitive cysteines, requires laborious sample decontamination steps, such as protein precipitation or buffer exchange. Such sample manipulation can easily result in material loss or spurious results due to inefficient decontamination. The large amount of sample input required for most redox proteomic methods has hindered application to samples with limited available material, such as primary cells and biopsy samples.

Several hurdles also exist for quantitative analysis of redox proteomics datasets. Accurate quantitation (frequently at the MS1 level) is an essential component of most chemoproteomic data analysis workflows, including methods aimed at measuring thiol oxidation state. Prior studies, including our own, have relied upon custom software to report accurate measures of relative MS1 chromatographic peak areas for the extracted ion chromatograms of heavy- and light-reagent-labeled peptides. While some

commercial software packages, such as ProLuCID/Census and Byonic have been successfully employed for residue-level quantification of chemoproteomics experiments, the cost of these tools has precluded widespread adoption (24). Powerful free software packages for both data search and quantitation, including Skyline and FragPipe, have been widely employed by the mass spectrometry community (25, 26, 27, 28, 29, 30, 31, 32, 33, 34). However, the adoption of these tools for chemoproteomics remains limited to a handful of studies, due in part to incompatibility with several specific applications, including the quantification of peptides containing multiple modifications (e.g., peptides with two or more cysteine residues). As multi-cysteine-containing peptides are ubiquitous in redox motifs (e.g., CXXC), this limitation is particularly problematic for measures of thiol-oxidation state (35, 36, 37). Furthermore, the relative performance of these tools for chemoproteomic applications remains unexplored.

An optimal redox proteomic method would achieve near complete removal of excess reagents with minimal sample loss, be compatible with all cell and tissue types, use a minimal amount of input material, report oxidative modifications with high sensitivity and specificity, and be cost-effective. The single-pot, solid-phase-enhanced sample-preparation (SP3) method is poised to enable such a method. The single-pot, solid-phase-enhanced sample-preparation method employs carboxyl-coated magnetic beads to achieve efficient sample decontamination, even for small sample sizes (38, 39). Our recent findings revealed that not only is SP3 compatible with cysteine chemoproteomics, it even affords improved coverage of labeled peptides when benchmarked against protein-precipitation-based cleanup methods (40). SP3 also facilitates multiple rounds of

sample cleanup, using a simple magnetic capture and elution system. Whether SP3 can enable identification of redox sensitive cysteines remains unexplored.

To enable rapid, cost-effective, and high throughput redox proteomics, we first synthesized novel and low-cost isotopically labeled IAA probes. We then combined these reagents with a new multi-step SP3 redox sample preparation workflow (SP3-Rox) compatible with FragPipe-IonQuant based quantification to achieve rapid and high coverage identification of redox-sensitive cysteines. We validate the accuracy of IonQuant quantification by comparison to widely adopted Skyline-based MS1 level quantification. Application of the SP3-Rox method identified redox sensitive cysteines sensitive to S-nitrosoglutathione (GSNO) in proteomes derived from an immortalized T lymphocyte cell line and cysteines in primary human T cells that showed cell-state dependent oxidation states. The technical innovations and expanded portrait of the redox proteome enabled by SP3-Rox provides a roadmap for a proteome-wide understanding of the cellular mechanisms underlying response to oxidative stress.

Experimental Procedures

Experimental Design and Statistical Rationale

We used a total of 58 datasets in this work. In all datasets, we estimated the identification false-discovery rate using the target-decoy approach (41). For MSFragger, peptide-spectrum matches (PSMs), peptides, and proteins were filtered at 1% PSM and 1% protein identification FDR. For Skyline analyses, default settings were used unless otherwise noted. All experiments were performed in duplicate or triplicate for quantification accuracy assessment and method validation. S-nitrosoglutathione and T

cell experiments utilized experimental duplicates with two additional technical replicates per condition (n = 4 for \pm GSNO and resting/activated T cells). Aggregated mean $\log_2(\text{heavy}:\text{light ratios})$ were used to assess peptide quantification accuracy in comparison to ground-truth ratios. Comparison of the number of unique quantified cysteine-containing peptides was used to evaluate the performance of quantitation pipelines. Means of reported \log_2 ratio values for each condition (\pm GSNO or unstimulated/stimulated T cells) were calculated for all replicates per condition, and the difference of the \log_2 mean values were reported. Variances were calculated for each sample-condition pairing and a corresponding two-sample t test was performed on the raw \log_2 ratios to generate p-values (n = 4 for \pm GSNO and unstimulated/stimulated T cells); p-values were adjusted for multiple comparisons using Benjamini–Hochberg procedure. Quantified peptides identified in all replicate samples per condition for \pm GSNO and unstimulated/stimulated T cell experiments were used for statistical analyses. Difference values above 2 and 1.5 were used in subsequent gene ontology and expression analyses for GSNO and T cell experiments respectively.

Cell Culture and Preparation of Cell Lysates

All cell lines were obtained from ATCC and maintained at a low passage number (<20 passages) and tested regularly for mycoplasma. The cells were cultured in DMEM/RPMI-1640 supplemented with 10 % fetal bovine serum and penicillin-streptomycin. Media was filtered (0.22 μm) prior to use. Cells were maintained in a humidified incubator at 37 °C with 5 % CO₂. Cells were harvested by centrifugation, washed twice with cold DPBS, resuspended in DPBS, and sonicated. Blood from deidentified healthy donor was

obtained from UCLA/CFAR Virology Core (5P30 AI028697) after informed consent. After Trima filter isolation, peripheral blood mononuclear cells were purified over Ficoll–Hypaque gradient (Sigma-Aldrich), and T cells were isolated via negative selection with magnetic beads (EasySep Human T Cell Iso Kit, 17951, STEMCELL). The isolated T cells were washed with sterile PBS. Unstimulated cells were harvested by centrifugation. The remaining cells were then resuspended in RPMI-1640 supplemented with FBS, penicillin, streptomycin, and glutamine (2 million cells per ml), and 200,000 cells per well were seeded on nontreated tissue culture, 96-well transparent plates that had been coated with anti-CD3 (1:200, BioXcell) and anti-CD28 (1:500, Biolegend) in PBS (100 μ l per well). After 72 h, the cells were then harvested, washed with PBS, and the cell pellets lysed by sonication in PBS. Protein concentrations were determined using a Bio-Rad DC protein assay kit from Bio-Rad Life Science, and the lysate diluted to the working concentrations as indicated.

Gel-Based Proteome Labeling

HEK293T proteome (25 μ l of 2 mg/ml) was labeled with 30 μ M isopropyl iodoacetamide alkyne (IPIAA)-H (5), IPIAA-L (4), or IAA (6) (0.75 μ l of 1 mM stocks) for 1 h. Copper-catalyzed azide–alkyne cycloaddition (CuAAC) was performed with rho-azide (#) (1 μ l of 1.25 mM stock in DMSO, final concentration = 50 μ M), TCEP (0.5 μ l of fresh 50 mM stock in water, final concentration = 1 mM), tris((1-benzyl-4-triazolyl)methyl)amine (1.5 μ l of 1.7 mM stock in DMSO/t-butanol 1:4, final concentration = 100 μ M), and CuSO₄ (0.5 μ l of 50 mM stock in water, final concentration = 1 mM). Samples were allowed to react for 1 h at ambient temperature. All samples were denatured (5 min, 95 °C) and

analyzed by SDS-PAGE using Criterion TGX Stain-free gels obtained from Bio-Rad. Loading control images were obtained after Coomassie staining.

SP3-Rox Proteomic Sample Preparation

Details for each experiment are provided in Supporting Information. Lysate samples (200 μ l of 2 mg/ml) were incubated with vehicle or 1 mM GSNO for 30 min at room temperature (RT) followed by labeling with 2 mM IPIAA-L (4; 2 μ l of 200 mM stock solution in DMSO, final concentration = 2 mM) for 1 h at 37 °C. The samples were incubated with 0.5 μ l benzonase (Fisher Scientific, 70–664–3) for 30 min at 37 °C. SP3 bead slurries were then transferred to the proteome samples, incubated for 10 min at RT with shaking (1000 rpm). Absolute ethanol (400 μ l) was added to each sample, and the samples were incubated for 5 min at RT with shaking (1000 rpm). Using the magnetic rack as described above, supernatants were then removed and discarded and the beads were further washed two times with 80% ethanol in water (400 μ l). Beads were then resuspended in 200 μ l PBS containing 2 M urea. TCEP (10 μ l of 100 mM stock in water, final concentration = 5 mM) was added into each sample, and the sample was incubated at 56 °C for 30 min, shaking (500 rpm). Beads were washed with absolute ethanol for 5 min followed by 80% ethanol twice as described and resuspended in 200 μ l PBS containing 2 M urea. The samples were then labeled with IPIAA-H (5; 2 μ l of 200 mM stock solution in DMSO, final concentration = 2 mM) for 1 h at 37 °C with shaking (500 rpm). Absolute ethanol (400 μ l) was added to each sample, and the samples were incubated for 5 min at RT with shaking (1000 rpm). Samples were then placed on a magnetic rack, and the beads are allowed to settle. Supernatants were then removed and discarded. Beads were then resuspended

in 200 μ l 0.5% SDS in PBS. Copper-catalyzed azide–alkyne cycloaddition was performed with biotin-azide 7 (4 μ l of 200 mM stock in DMSO, final concentration = 4 mM), TCEP (4 μ l of fresh 50 mM stock in water, final concentration = 1 mM), tris((1-benzyl-4-triazolyl)methyl)amine (12 μ l of 1.7 mM stock in DMSO/t-butanol 1:4, final concentration = 100 μ M), and CuSO₄ (4 μ l of 50 mM stock in water, final concentration = 1 mM). Samples were allowed to react for 1 h at ambient temperature with shaking (500 rpm). Samples were washed and subjected to trypsin digestion as described below.

SP3 Digest and Elution

Absolute ethanol (400 μ l) was added to each sample, and the samples were incubated for 5 min at RT with shaking (1000 rpm). Beads were washed with 80% ethanol as described above. The beads were then resuspended in 200 μ l 0.5% SDS in PBS containing 2 M urea. Dithiothreitol (DTT; 10 μ l of 200 mM stock in water, final concentration = 10 mM) was added into each sample, and the sample was incubated at 65 °C for 15 min. To this, iodoacetamide (10 μ l of 400 mM stock in water, final concentration = 20 mM) was added and the solution was incubated for 30 min at 37 °C with shaking. After that, beads were washed with ethanol as described. Next, the beads were resuspended in 150 μ l PBS containing 2 M urea followed by an addition of 3 μ l trypsin 1 mg/ml solution. Digest was allowed to proceed overnight at 37 °C with shaking. After digestion, ~4 ml acetonitrile was added to each sample and the mixtures were incubated for 10 min at RT with shaking (1000 rpm). Supernatants were then removed and discarded using the magnetic rack, and the beads were washed (3 \times 1 ml acetonitrile). Peptides were then eluted from SP3 beads with 100 μ l of 2% DMSO in

molecular biology (MB) grade water for 30 min at 37 °C with shaking two times for a total of 200 µl eluent.

NeutrAvidin Enrichment of Labeled Peptides

For each sample, 50 µl of NeutrAvidin Agarose resin slurry (Pierce) was washed three times in 10 ml immunoaffinity purification buffer (50 mM MOPS–NaOH (pH 7.2), 10 mM Na₂HPO₄, 50 mM NaCl) and then resuspended in 500 µl immunoaffinity purification buffer. Peptide solutions eluted from SP3 beads were then transferred, and the samples were then rotated for 2 h at RT. After incubation, the beads were pelleted by centrifugation and washed. Bound peptides were eluted with 60 µl of 80 % acetonitrile in MB water containing 0.1% formic acid (10 min at RT & 10 min at 72 °C). The eluants were dried (SpeedVac). The samples were then reconstituted with 5% acetonitrile and 1% formic acid in MB grade water and analyzed by LC-MS/MS.

LC-MS/MS Analysis

Details are provided in Supporting Information document. The samples were analyzed by liquid chromatography tandem mass spectrometry using Orbitrap Eclipse Tribrid Mass Spectrometer (Thermo Scientific) coupled to an Easy-nLC 1200 system.

FragPipe Peptide Identification and Quantitation

Details are provided in Supporting Information document and workflow file in PRIDE (PXD029500). Raw data collected by LC–MS/MS were converted to mzML format (for PTMProphet datasets) or left as raw files and searched using FragPipe GUI v16.0 with

MSFragger (version 3.3) (28, 32), Philosopher (version 4.0.0) (34) and IonQuant (version 1.7.5) (29, 30) were enabled. Precursor and fragment mass tolerance was set as 20 ppm. Missed cleavages were allowed up to 2. Peptide length was set 7 to 50, and peptide mass range was set 500 to 5000. Cysteine residues were searched with variable modifications at cysteine residues for carboxyamidomethylation (+57.02146), IPIAA-L (+463.2366), and IPIAA-H (+467.2529) labeling allowing for 3 max occurrences and all mods used in first search checked. Peptide and protein level FDR were set to 1%. Permissive IonQuant parameters allowed minimum scan/isotope numbers set to 1. PTMProphet information was obtained from psm.tsv using 'heavy' and 'light' localizations scores. A FASTA database from UniProtKB homo sapien FASTA file containing canonical, nonredundant sequences (08/2018) used for all searches and is provided in the PRIDE (PXD029500 and PXD031647) repository. Proteomic workflow of FragPipe and its collection of tools are outlined in Supporting Information.

Skyline Quantitation

Details are provided in Supporting Information document. Interact.pep.xml files from FragPipe searches were imported into Skyline v21.1.0.146 (42) with a probability threshold corresponding to the 1% peptide-ion level FDR in the dataset. Following the standard DDA analysis workflow for isotopically labeled dataset using the following modifications: carboxyamidomethylation (+57.02146) as a variable modification, light structural modification (+463.2366), and heavy isotope modification (+4.01634). As Skyline automatically places a heavy isotopic label on all modified cysteines, including carbamidomethylated residues and those modified by the IPIAA-H reagent, quantification

fails for peptides containing two or more modified cysteines (e.g., one carbamidomethyl residue and one IPIAA-residue). A custom plugin was generated to remove the heavy mass from all carbamidomethylated residues.

Data Analysis and Processing

Custom R scripts were implemented to compile labeled peptide datasets from peptide_label_quant.tsv FragPipe outputs. Unique cysteines were quantified for each dataset using unique identifiers consisting of UniprotID_Cysteine Number. Details of data processing are in Supplemental Information.

Synthesis of Reagents

Compound 6 and 7 were prepared as has been reported (43). Detailed syntheses of compounds 4 and 5 are provided in the electronic Supplementary Information.

Results

Synthesis and Benchmarking of Low-Cost Isotopically Labeled IAA Probes

Relatively cost-effective isotopically labeled IAA probes have been achieved, using ¹³C-benzaldehyde (\$115/mmol) as a low-cost isotope source (9). Here, we envisioned the synthesis of even lower cost second generation IAA probes, by employing isotopically labeled ¹³C-acetone (\$60/mmol) for isotopic barcoding. The desired light and heavy isopropyl iodoacetamide probes (IPIAA-L and IPIAA-H) were obtained in three steps, and the key isotopic labels were incorporated by reductive alkylation with propargyl amine and acetone, using either ¹²C- or ¹³C-acetone and NaBH₃CN or NaBD₃CN for the light and

heavy probes, respectively, as shown in Figure 1A. Analysis of the reagents by LC-MS confirmed that the desired isotopologues had been obtained at equal concentrations (Fig. 1B).

With probes validated, we next established labeling conditions to achieve high occupancy cysteine capping required for redox proteomics. Benchmarking of the IPIAA probes against a standard unsubstituted IAA probe using a gel-based assay revealed comparable labeling and banding pattern for all three probes (Fig. 1C). Using a competitive gel-based assay under denaturing conditions (2M Urea), we found that at 2 mM concentration, all three probes efficiently blocked lysate labeling by IA-rhodamine. This finding concurs with our prior study (40), which indicated that low mM concentrations of iodoacetamide reagents are sufficient to cap most cysteines with near completion (supplemental Fig. S1). Somewhat unexpectedly, a comparison of cysteine labeling at pH 7 and pH 8.4 revealed near complete competition of IA-rhodamine labeling at both neutral and more basic pH (supplemental Fig. S2). Next, we carried out a mass spectrometry-based competitive assay by subjecting Jurkat lysates to labeling with 2 mM IPIAA-H and IPIAA-L probes followed by 20 mM IA to completely cap all unlabeled cysteines. After tryptic digest and LC-MS/MS analysis, search using MSFragger revealed that for both the light and heavy probes, >80% of all cysteines were labeled with the IPIAA (supplemental Fig. S3). The observed slight decrease in labeling efficiency compared to our prior study using unsubstituted IAA (40) can likely be ascribed to the increased steric bulk afforded by the isopropyl modification.

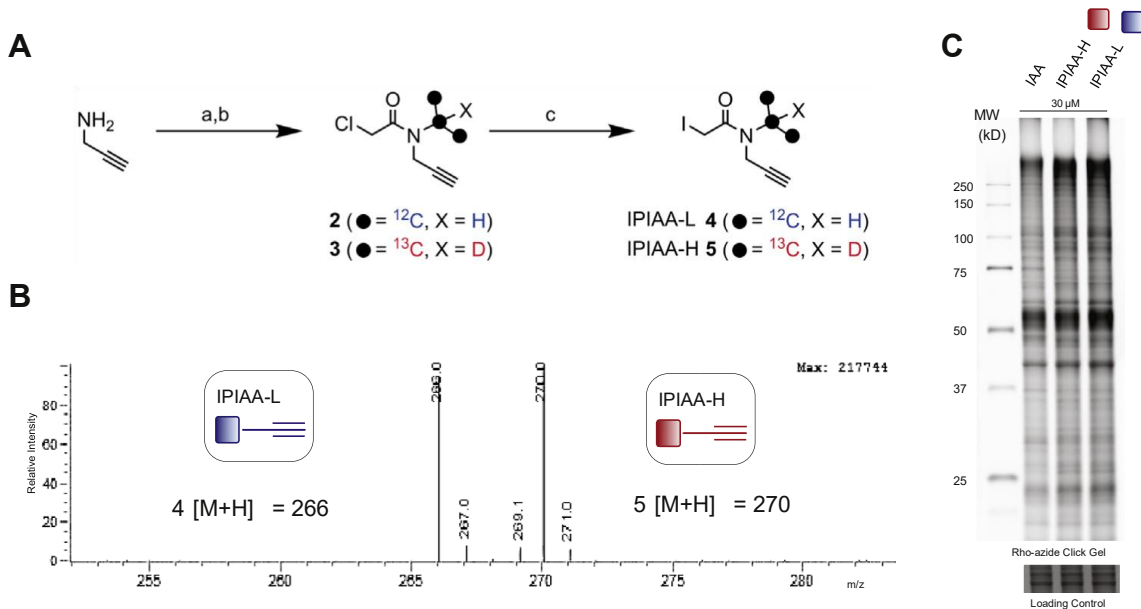


FIG. 1. Synthesis and evaluation of isotopic isopropyl iodoacetamide alkyne probes. A, synthesis of heavy and light iodoacetamide alkyne probes. a, acetone (^{13}C or light), NaBH_3CN or NaBD_3CN , $\text{CH}_2\text{Cl}_2/\text{MeOH}$ (2:1); rt b, chloroacetyl chloride, NEt_3 ; c, NaI , acetone, CH_2Cl_2 , 6.2% yield for 4 over 3 steps, 11% yield for 5 over 3 steps, and 11% yield for 5 over 3 steps. B, LC-MS of IPIAA probes in equimolar mixture. C, gel-based visualization of 30uM IPIAA labeling followed by CuAAC to rhodamine-azide in HEK293T lysates. Loading control was generated using InstantBlue Coomassie protein stain. CuAAC, copper-catalyzed azide–alkyne cycloaddition; IAA, iodoacetamide alkyne; IPIAA, isopropyl iodoacetamide alkyne.

FragPipe-IonQuant and Skyline-Based Quantitation for Chemoproteomic Analysis

With probes in hand, we sought to establish a robust platform for chemoproteomic dataset quantitation. We subjected HEK293T cell lysates to labeling with different ratios of the IPIAA-H and IPIAA-L probes (1:1 and 4:1) (Fig. 2A). The heavy- and light-labeled samples were then subjected to CuAAC to biotin-azide, SP3 sample cleanup, on-resin digest, enrichment on Neutravidin, and LC-MS/MS analysis. Guided by the precedent of recent reports, including our prior study (40) that showcased the compatibility of MSFragger - ultrafast, easy-to-use search algorithm - with chemoproteomics, we opted to use FragPipe - the MSFragger-powered computational pipeline for proteomics data analysis.

In a conventional 'closed' MSFragger search with FragPipe v16.0, we identified 8,350 cysteine-containing peptides, 7,126 light-labeled, and 4,017 heavy-labeled peptides in one 1:1 (H:L) replicate.

Next, we sought to identify a quantitation algorithm compatible with our SP3-Rox method, including for multi-cysteine containing peptides. While there are many available quantitation algorithms, we opted to compare FragPipe's built-in IonQuant tool (29, 30) to the quantitation achieved by Skyline v21.1.0.146 (42). These packages were selected based on widespread adoption by the proteomics community and availability of options for compatibility with multi-cys containing peptides and with Field Asymmetric Ion Mobility Spectrometry (FAIMS) data (44). To quantify data in FragPipe, we customized a workflow (see Supplemental Information) that substantially increased the fraction of peptides quantified when compared to the default settings (supplemental Fig. S5). One key modification that was implemented was selection of the "use all mods in first search" that retains all modifications from the user's setting in the initial search, which enables improved mass calibration for peptides. Without this option, MSFragger only keeps common modifications in the first search; for samples with rare modifications, the number of high quality PSMs would be insufficient for mass calibration.

We quantified 5748, 5736, and 8269 cysteines in the 1:1, 4:1, and 1:10 (H:L) datasets respectively with IonQuant (Figs. 2 and S4), where at least 93% of all peptides identified fall within ratio windows [$1 > \log_2(H/L) > -1$, $4 > \log_2(H/L) > 1$, and $1 > \log_2(H/L) > -4$], supporting accurate quantification of labeled cysteines. More variability in ratios was observed for the 1:10 datasets consistent with decreased accuracy for larger fold change MS1 chromatographic peak area ratios. We also observed a marked increase in the

fraction of peptides detected with only one isotopic label (e.g., only heavy-label identified) for higher reagent ratio combinations (Fig. 2C). We found for 1:10 samples, ~1,200 unique cysteines and for 4:1 samples ~100 unique cysteines with single labels (Figs. 2C and S4). For these peptides, we set the ratio value to a maximal ratio of 20 or minimal ratio of 1/20 (\log_2 values of ± 4.32).

To further benchmark and optimize the IonQuant quantification, we next established a pipeline for Skyline quantification of these model datasets. From FragPipe searches, interact.pep.xml files were imported into Skyline following the standard DDA analysis workflow for isotopically labeled dataset with the following modifications. As Skyline automatically places a heavy isotopic label on all modified cysteines, including carbamidomethylated residues and those modified by the IPIAA-H reagent, quantification fails for peptides containing two or more modified cysteines (e.g., one carbamidomethyl residue and one IPIAA-residue). Therefore, a custom plugin was generated to remove the heavy mass from all carbamidomethylated residues. Accurate quantification of FAIMS data was achieved using the Ion Mobility settings (See Supplementary Information). While Skyline initially outperformed IonQuant in terms of absolute numbers of peptides quantified (supplemental Fig. S5), we found that this difference in performance could be eliminated by adjusting the IonQuant parameters by setting minimum scan and minimum isotopes required for feature detection to 1 (supplemental Fig. S5). With these changes, we observed similar quantification with Skyline and IonQuant using our 1:1 and 4:1 datasets (Fig. 2B).

As FAIMS-based separation of peptides has only recently been implemented for chemoproteomics experiments (40) the impact of FAIMS on quantitation of detected ions

in chemoproteomics experiments has not, to our knowledge, been explored. Therefore, our next step was to compare the relative intensities of peptide ratios of samples analyzed with and without FAIMS. After observing the increased ratio variability in the 1:10 datasets, we opted to compare quantitation for the 1:10 samples analyzed with and without FAIMS. This comparison revealed FAIMS-associated ratio compression consistent with that of reported Stable Isotope Labeling by Amino acids in Cell culture experiments, with more pronounced compression evident for the higher ratio 1:10 datasets in comparison to the 4:1 datasets (supplemental Fig. S4) (45). Interestingly, FAIMS nearly eliminates the large fraction of single-isotopic labeled quantified peptides in both our FragPipe and Skyline analysis (peptides with ratio L:H ratio values >3,000 in Skyline datasets) (Figs. 2C and S4). We also found that FAIMS data analyzed using FragPipe (mean ratio = -2.57) showed reduced ratio compression relative to Skyline quantitation with Ion Mobility settings (mean ratio= -1.79) (Fig. 2C), supporting the use of a FAIMS device coupled with IonQuant for quantitative chemoproteomic applications.

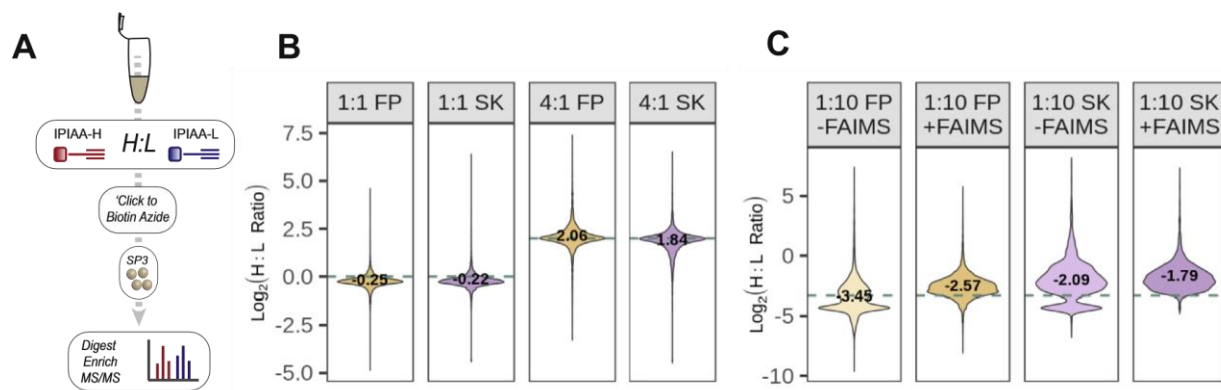


FIG. 2. Quantification of IPIAA probes in various heavy to light ratios using IonQuant. A, workflow to assess the quantification of IPIAA probes. B, heavy to light intensity ratios from 1:1 and 4:1 IPIAA-H: IPIAA-L mixtures aggregated from triplicate datasets (+FAIMS) using FragPipe coupled with IonQuant (FP) or ratios of integrated peak areas from ion chromatograms Skyline (SK), mean ratio values reported. C, FragPipe IonQuant ratio reproducibility with and without FAIMS using 1:10 IPIAA-H: IPIAA-L mixtures, mean ratio values reported. Dashed lines indicate ground-truth \log_2 ratio. Experiments were performed in triplicate for \pm FAIMS comparisons and 4:1 comparisons and duplicate for 1:1 comparisons. Data available

in supplemental Table S1. FAIMS, field asymmetric waveform ion mobility spectrometry; IPIAA, isopropyl iodoacetamide alkyne.

Single-Pot, Solid Phase-Enhanced Sample-Preparation for Identification of Relative Cysteine Oxidation and PTMProphet-Enabled Site-of-Labeling Localization With heavy- and light-IPIAA probes in hand and quantitation workflow established, we next developed our chemoproteomic redox platform. Motivated by the improved coverage that we had achieved previously with SP3 solid phase sample cleanup, we tested whether the SP3 method could be extended to measure cysteine oxidation state with the goal of harnessing SP3-enabled repeated binding and elution on resin to conduct sequential capping, reduction, and capping steps with the heavy and light IPIAA probes.

We adapted the optimized workflow from our previous study to produce the SP3-Rox workflow, shown in Figure 3A (additional methodological details are shown in supplemental Fig. S6). Initial samples subjected to labeling in nondenaturing conditions indicated only partial or incomplete labeling of reduced cysteines in comparison to denaturing conditions with apparent ratio values nonrepresentative of a reduced thiol proteome (supplemental Fig. S7) (46, 47). To ensure optimal labeling, the samples were subjected to labeling with IPIAA-L in 2M urea first to cap all reduced cysteines; labeled proteins were then bound to SP3 resin and washed to remove excess IPIAA-L. Following on-SP3 resin thiol reduction and a second round of SP3 desalting to remove excess reductant, all formally oxidized cysteines were capped with IPIAA-H. The samples were then conjugated by CuAAC to biotin-azide, and excess reagents were removed by subsequent wash steps using SP3 beads. After sequence-specific digest (trypsin) and enrichment on neutravidin resin, labeled peptides were analyzed by LC-MS/MS. Notably,

minimal lysis-induced oxidation was observed for both sonication and urea-based lysis (supplemental Fig. S8).

SP3-Rox analysis of Jurkat cell lysates in biological triplicate yielded 7,805 quantified cysteines in two or more replicates after filtering the data (Figs. 3, A and B and S9 and S10). SP3-Rox afforded robust (90–95%) enrichment of cysteine-containing peptides using this workflow, consistent with our previous study (40) (supplemental Fig. S11) and adequate peptide recovery from 400 μ g input material (supplemental Fig. S12). Gratifyingly, and consistent with the aforementioned vetting of both Skyline and FragPipe quantitation, we observed a strong correlation between ratios reported for both methods, both with and without FAIMS (Figs. 3B and S10A). We also find that the median IonQuant H/L ratio is -3 , ($\sim 10\%$ oxidation) (supplemental Fig. S9), which is consistent with previously reported values for the median oxidation state of proteinaceous cysteines (46, 47). Analysis of annotated disulfides (UniProtKB) revealed a strong enrichment for higher heavy/light ratios anticipated for heavily oxidized thiols with a 3-fold increase in disulfide annotated cysteines for IonQuant $\log_2(\text{H/L})$ ratios > -3 (more than $\sim 10\%$ oxidized) in comparison to $\log_2(\text{H/L})$ ratios < -3 . Notable disulfides identified include IGF2R Cys161, TF Cys137, CD1C Cys120, and HLA-A Cys188. Many disulfides with lower heavy/light ratios were found to be primarily in redox-active motifs (e.g., thioredoxin, peroxiredoxin), which rationalizes the observed partial labeling by the IPIAA-L probe (Fig. 3B). We also observe modest FAIMS-induced ratio compression in comparison to samples run without FAIMS (supplemental Fig. S10B). To maintain consistency with previous studies, we provide both FragPipe reported ratios as well as a % oxidation metric in the proteomic datasets reported in supplemental Tables S2, S4, and S5.

Many redox-active cysteines are located in tryptic peptides that contain multiple cysteines. The accurate identification of the labeling site and quantification of the area ratio for these multi-cysteine containing peptides is confounded by several key factors. First, the default assumption of many software packages (e.g., MaxQuant and Skyline) is that all cysteines within a given peptide are labeled by the same variable modification. Therefore, for methods that use multiple labels such as the SP3-Rox method, the identification and quantification of multi-cysteine peptides frequently fail using these software packages. To fix this issue, we implemented a customized plugin for Skyline analysis to compare multi-cysteine peptide ratios with those quantified by IonQuant in FragPipe. Second, as is the case with all posttranslational modifications, and described in detail for phosphoproteomics, identification of the labeling site for peptides containing multiple possible amino acid labeling sites can be confounded by the uncertainty caused by incomplete coverage of b and y ions (48). The lack of confidence metrics for posttranslational modification site assignments in most search engines makes differentiating these sites of labeling more difficult.

In contrast with other search algorithms, MSFragger has the flexibility to accommodate multiple unique modifications on cysteines. Six percentage of all unique peptides (over 450) quantified by the SP3-Rox method contain multiple cysteines (Fig. 3C). Of these, ~300 were found to contain light labeling together with carbamidomethyl modifications and only ~15 with both light and heavy labels. Comparison of the ratios generated for multi-cysteine containing peptides quantified by IonQuant and Skyline revealed a strong correlation, consistent with similar performance of both packages (Figs. 3D and S13). Analysis of individual Skyline MS1 peak areas for multi-cys peptides

revealed high concordance with FragPipe identified ratios. Notably, partially labeled peptides containing one carbamidomethyl modification were observed to have substantially reduced intensities compared to fully labeled peptides, consistent with near complete IPIAA labeling afforded by the SP3-Rox method (supplemental Fig. S13). Validated against Skyline, we proceeded to use IonQuant for all subsequent analyses. Using the PTMProphet tool (49) which is built into FragPipe and has recently been tailored to localize any assigned modifications, we assessed the probability score for localization and found that for most peptides, the localization probability is greater than 0.8 (Fig. 3E). Showcasing the accuracy of PTMProphet scoring, the known redox-active cysteine Cys152 in GAPDH was identified with a light labeling and with carbamidomethyl labeling at Cys156 with localization scores ranging from 0.86 to 0.99 for each site (50). Similarly, for the Cys169, Cys171 multi-cys peptide from LYPLA1 thioesterase, Cys171 showed preferential light labeling with a localization score of 0.93. In contrast, a multi-cys peptide from the IKZF1 DNA-binding protein was identified with more ambiguous localization scores for the nonactive site, zinc-finger cysteines Cys147 and Cys150 positions (supplemental Table S2) (51).

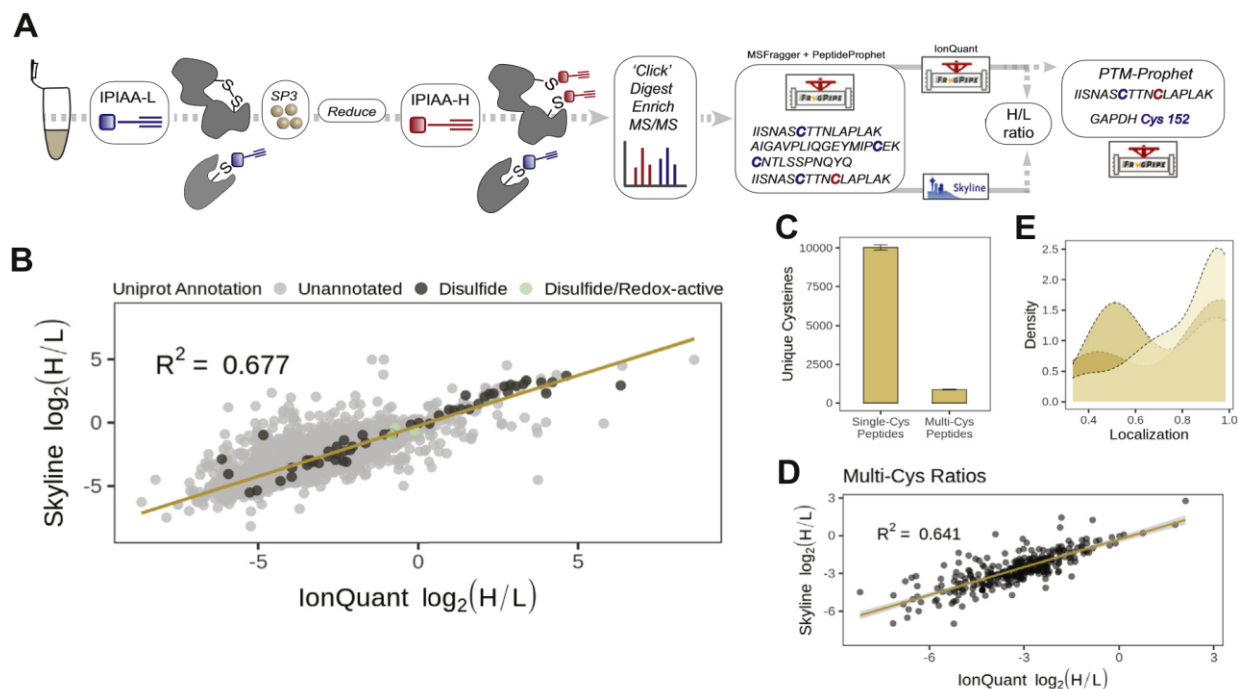


FIG. 3. Application of SP3-Rox with FragPipe-IonQuant to identify oxidized cysteines. A, workflow to assess the quantification of oxidation state ratios using SP3-Rox workflow. B, pairwise comparison of ratios quantified using FragPipe-IonQuant or Skyline aggregated from triplicate datasets; dark color indicates annotated disulfides, green color indicates annotated redox-active disulfides. C, comparison of single and multi-cysteine peptides aggregated from triplicate datasets. D, pairwise comparison of multi-cysteine peptide ratios using FragPipe-IonQuant or Skyline aggregated from triplicate datasets. E, localization scores on multi-cysteine peptides for three replicate experiments. Experiments were performed in triplicate. Data available in supplemental Table S2. IPIAA, isopropyl iodoacetamide alkyne; SP3, single-pot, solid-phase-enhanced sample-preparation.

SP3-Rox Identification of Cysteines Sensitive to Reversible Oxidation Cysteine S-

nitrosylation has been implicated in numerous biological processes, spanning apoptosis, proliferation, and angiogenesis and contributes to the pathology of multiple diseases (e.g., Alzheimer's and Parkinson's disease) as well as host immune responses (7). Motivated by these wide ranging and significant biological impacts, we opted to apply our workflow to identify reversibly oxidized cysteines prone to S-nitrosation by GSNO (Fig. 4A). Immortalized T lymphocyte cell (Jurkat) proteome was treated with 1 mM GSNO or vehicle followed by SP3-Rox sample preparation to establish our method using samples with robust labeling. We identified 15,226 unique sites and 4,479 unique proteins using

our workflow. 4,060 unique sites were identified in all four replicates per condition. The ratio difference between treated and control samples was then calculated for these 4,060 high confidence cysteines (Fig. 4B). Two thousand seventy two cysteines showed significant ratio changes (Difference >2, p-value < 0.5) sensitivity to GSNO labeling (Fig. 4C). Of these residues, a number with high ratio changes have been previously reported to be involved in the oxidative stress response including superoxide dismutase (SOD1 Cys147, ~6.3), parkin 7 (PARK7 Cys106, ~4.7), and thioredoxin (TXN Cys73, ~3.7) (Fig. 4C) (52, 53, 54). In particular, parkin 7 Cys106 oxidation is necessary for mitochondrial relocation and protection against neuronal death (3, 55). The majority of Swiss-Prot reviewed and annotated S-nitrosocysteines identified in our dataset present significant and high ratio changes (Fig. 4C) (56).

Looking beyond proteins implicated in cellular response to oxidative stress, we next assessed whether other subsets of proteins were enriched for harboring redox sensitive cysteines. Gene ontology enrichment analysis (57, 58) of proteins exhibiting significant high ratio-change cysteines (residues highlighted in green, Fig. 4B) did not reveal enrichments in redox pathways, possibly due to the excess of GSNO (1 mM) used in our labeling study (supplemental Fig. S14). Immune-relevant proteins have been found to harbor numerous reactive and ligandable cysteines (59). Consistent with these findings, we uncovered a number of immune-relevant proteins containing GSNO-sensitive cysteines (supplemental Table S3).

Given that treatment with somewhat superphysiological concentrations of GSNO could easily afford nonspecific cysteine oxidation, we next asked whether SP3-Rox could stratify the relative GSNO sensitivity of individual cysteines. Across a panel of 50 proteins

analyzed that all contain ≥ 6 quantified cysteines with significantly different ratio changes, we find that nearly all show preferential labeling of one or only a handful of cysteines across all residues detected (green, Fig. 4D). One notable exception is CNBP in which nearly all identified cysteines exhibit elevated ratios, consistent with nitrosylation. As CNBP contains multiple zinc finger motifs, we speculate that the observed GSNO promiscuous labeling is likely indicative of cysteine-modification-induced protein unfolding.

As redox-sensitive cysteines are frequently found in close proximity, both in sequence and 3D space, to additional cysteines, we hypothesized that the GSNO-sensitive cysteine subset would be enriched for peptides that contain multiple cysteines. We find that, of the sites we annotated as GSNO sensitive (Difference >2 , p-value < 0.05), 6% belong to general multi-cys peptides (supplemental Fig. S15A) in comparison to 5% of sites we established as insensitive (Difference <2 or p-value > 0.05) (supplemental Fig. S15B). We then extended this analysis to the subset of peptides that contain a putative redox motif (either 'CXXC' or 'CXXXC'). In the GSNO-sensitive subset (green), nearly 40% of the multi-cys peptides contain a redox motif (supplemental Fig. S15C) compared to 28% in the insensitive subset (gray) (supplemental Fig. S15D). While we acknowledge that the relatively modest number of putative redox motif-containing peptides may preclude generalization of our findings, we believe that this observed enrichment highlights the importance of considering multi-cys motifs in quantification pipelines, particularly those aimed at profiling redox-sensitive cysteines.

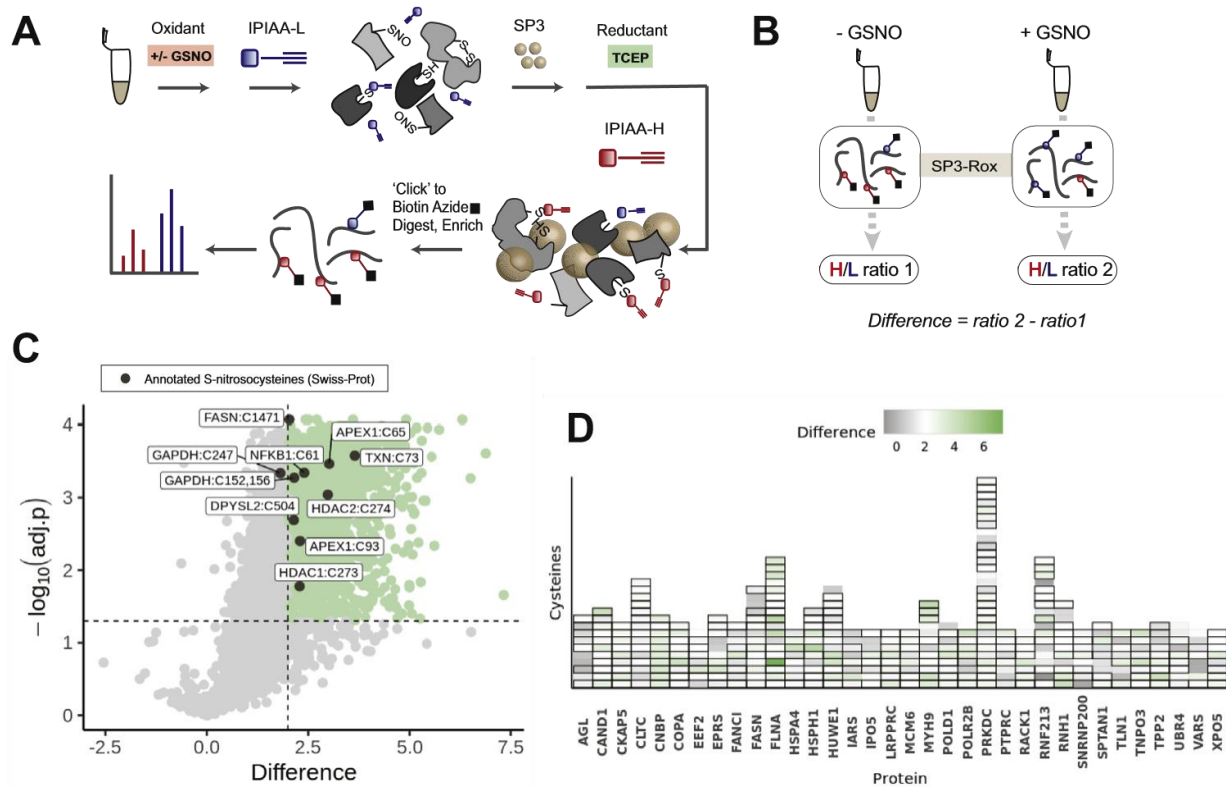


FIG. 4. Application of SP3-Rox to identify redox-sensitive cysteines. A, workflow to assess cysteines sensitive to nitrosation by GSNO. B, quantification of differential oxidation states. C, difference in mean log₂ ratio upon GSNO treatment prior to labeling (Two-sample t test; green are significant Difference >2, p-value < 0.05 and dark gray are Swiss-Prot annotated S-nitrosocysteines; p-values adjusted using Benjamini–Hochberg procedure). D, heatmap showing log₂ ratios of cysteines in multi-cys detected proteins, significant values bolded in gray. Experimental duplicates with two technical replicates (n = 4) were used for each treated and control condition. Data available in supplemental Table S4. GSNO, S-nitrosoglutathione; IPIAA, isopropyl iodoacetamide alkyne; SP3, single-pot, solid-phase-enhanced sample-preparation.

SP3-Rox for Identification of Cell-State-Dependent Redox-Sensitive Cysteines

Reactive oxygen species have been extensively implicated in the activation of peripheral T cells. T cell activation requires ROS production; however, inappropriately high levels of ROS are proinflammatory and have been linked to DNA damage and cell death (60, 61, 62, 63, 64). Several oxidation-sensitive cysteines have been implicated in the regulation of appropriate T cell function (e.g., residues in NRF2 and NFKB) (65, 66).

As T cells produce a burst of ROS during cellular activation, we postulated that a comparison of resting (unstimulated) versus activated (anti-CD3 and anti-CD28 stimulated) T cell redoxomes would reveal changes in the oxidation state of cysteines involved in cellular activation. To test this model, T cells isolated from healthy donor blood were subjected to the SP3-Rox workflow either before or after stimulation with immobilized anti-CD3 and anti-CD28 antibodies (Fig. 5A).

Using the IonQuant workflow illustrated in Figure 3A for stimulated and resting T cells, we quantified 13,411 total cysteines including 4,061 high confidence cysteines identified in all four replicates per condition. We also find 243 multi-cysteine containing peptides (~5%) with varying and quantifiable ratios (supplemental Fig. S16). We found 55 cysteines with significant changes in cysteine redox state during activation (Difference < -1.5 or > 1.5) (Fig. 5, A and B). As expected, we primarily observe increases in mean ratios indicating an oxidizing environment relative to unstimulated T cells. As in the GSNO analysis, we detect many proteins with multiple quantified cysteines, including a subset that exhibit high difference values (Fig. 5C). Comparison to the prior chemoproteomic inventory of ligandable cysteines in T cells revealed 33 ligandable and redox-sensitive cysteines (Difference > 1), with several high ratio-change cysteines including RRP1B Cys197 and RPS6KA5 Cys475, which provides additional clues for prioritization of residues for future small molecule probe development campaigns (Fig. 5B) (59).

Gratifyingly, we identified a number of cysteines in proteins with known involvement in adaptive immune response and T cell activation. Consistent with the important contributions of protein kinases, particularly serine-threonine kinases in T cell function (67), we detect cell-state-dependent cysteine-oxidation changes in multiple

kinase cysteines demonstrating changes in oxidative state with differences in the 1.5 to 2.0 range. For example in PRKCQ, which is implicated in T cell receptor signaling, we identify Cys193 (Difference = 1.77) (68). In PRKDC, the DNA-dependent kinase involved in DNA damage repair and telomere stability, Cys1312 shows increased oxidation during T cell activation (Difference = 1.61) (Fig. 5B) (69). Interestingly, a prior study of cysteine dimethyl-fumarate sensitivity in native and activated T cells identified PRKDC (Cys4045) and PRKCQ (Cys14/17 and Cys322) (70). We also identify serine protein kinase ATM Cys2021 (Difference = 1.73) and bifunctional polynucleotide phosphatase/kinase PNKP Cys445 (Difference = 2.34) implicated in DNA damage response critical to T cells (71, 72).

Among the cysteines demonstrating very high ratio changes (>2.5) is GATOR complex protein MIOS Cys276 (Difference = 3.00), which indirectly positively regulates the mTORC pathway as well as GCN1 Cys1482 (Difference = 3.32), a positive regulator of the EIF2AK4 kinase pathway, and E2F1 activator RRP1B Cys197 (Difference = 2.65) (73, 74, 75). We observe many nonkinase proteins also linked to T cell regulation and viral and tumor immune response such as SHMT2 Cys412, ITPKB Cys693, ZC3HAV1 Cys581, CTPS1 Cys216, and SERPINB9 Cys98 (76, 77, 78, 79, 80). Several proteins demonstrating high ratio changes do not have direct links to T cell activation and may play roles not yet elucidated. We also see a small number of cysteines with more reduced redox state upon stimulation including GADD45GIP1 Cys124, a negative regulator of cell-cycle progression (81), which is more reduced in activated T cells and upregulated during activation (82) as well as IRF2BP2 Cys521, an important transcriptional repressor typically down-regulated during T cell activation (83).

Altered protein abundance can complicate the interpretation of putative oxidative modification events identified by workflows requiring inter-sample ratio comparisons such as IsoTOP-ABPP and adaptations to the OxICAT method. A key advantage of the SP3-Rox method is its general insensitivity to abundance changes, even for proteins with only single cysteines identified. To further assess the ability of the SP3-Rox method to delineate oxidative modifications in the presence of expression changes, we analyzed unstimulated and activated T cell transcriptome profiling datasets from Zhao et al. to determine if significantly high ratio-change cysteine residues in Figure 5B reside in genes that show differential expression during activation. Comparing 0 h to 72 h differential expression datasets, we find that the subset of genes with high ratio-change (Difference >1.5) cysteines (green) show insignificant difference in the distribution of $\log_2(R)$ values from all other genes (Fig. 5D).

Gene ontology-analysis of proteins that harbor cysteines with significant difference >1.5 revealed an enrichment for proteins involved in peptidyl-serine phosphorylation and telomere maintenance, consistent with well-established literature on loss of telomere length during T cell activation (84, 85) (Fig. 5E). Many identified cysteines reside in or near ATP-binding sites or enzyme active sites such as SQOR, which catalyzes the oxidation of hydrogen sulfide with active site Cys201 (Difference = 1.76), functioning as part of the catalytic disulfide bond (86) with Cys379. We also identify CAMK2D Cys290 (Difference = 2.06) near the calmodulin-binding site 291 to 301, which is intriguing given CAMK2D's roles in T cell proliferation (Fig. 5, F and G) (87).

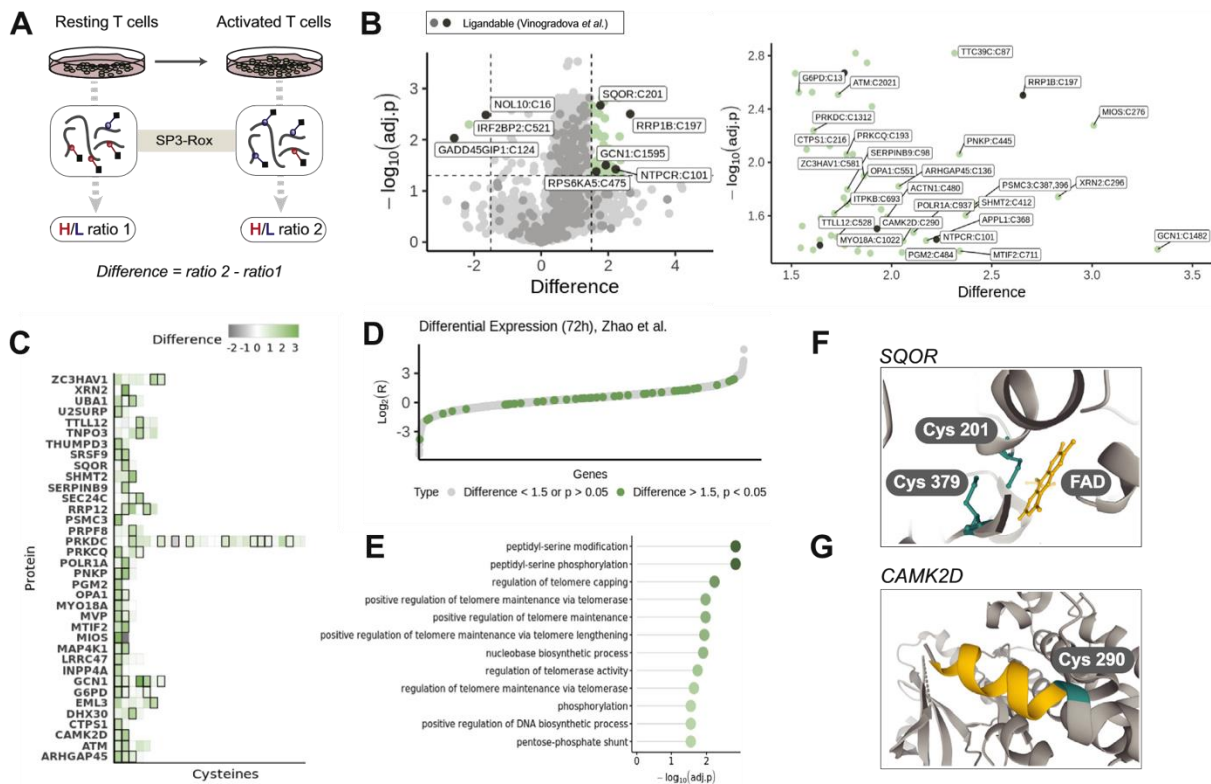


FIG. 5. Application of SP3-Rox to identify redox changes during T cell maturation. A, workflow to assess redox state in resting and activated T cells. B, difference in mean log₂ ratio upon activation (Two-sample t test, green are significant Difference >1.5, p-value < 0.05; darker gray points are ligandable cysteines from Vinogradova et al.) (left), including immune-relevant cysteines indicated (right). C, heatmap showing log₂ ratios of cysteines for proteins with multiple cysteines with at least one residue with significant Difference >1.5, significant values bolded in gray. D, comparison of differential expression for cysteines with Difference >1.5, p-value < 0.05 and all other cysteines using transcriptome datasets from Zhao et al (Mann-Whitney U test, p-value = 0.52). E, gene ontology-analysis of cysteines with Difference >1.5, p-value < 0.05 from panel B. F, crystal structure of SQOR indicating FAD-binding site and detected Cys201 and Cys379 (PDB 6OIB). G, crystal structure of CAMK2D indicating detected Cys290 and calmodulin-binding site (yellow) (PDB 5VLO). Experimental duplicates with two technical replicates (n = 4) were used for each treated and control condition. Data available in supplemental Table S5. SP3, single-pot, solid-phase-enhanced sample-preparation.

Discussion

Cysteine oxidative modifications are known to regulate most biological processes. Consequently, methods to determine the sites and fractional occupancy of thiol side chain modifications are widely applicable. Many such techniques are available including OxICAT, Cys-BOOST, and adaptations of IsoTOP-ABPP such as QTRP (10, 12, 87).

Near ubiquitous-shared features of these methods are the requirement for multiple rounds of sample decontamination (typically accomplished by precipitation or size exclusion) and for relatively large amounts of input material and for costly and complex isotopic reagents. Addressing these challenges, here, we combined new cost-effective isotopically labeled cysteine-reactive probes with SP3 sample cleanup to develop the SP3-Rox method, which is capable of quantifying cysteine oxidation state for small sample sizes, including for primary T cells derived from healthy donors.

To enable this method, we built upon our SP3 chemoproteomics sample cleanup workflow to incorporate two rounds of cysteine labeling, first capping all reduced cysteines and then all oxidized cysteines, with light and heavy isotopically differentiated iodoacetamide probes, respectively (40). While our method derives substantial inspiration from the pioneering OxICAT platform and workflows developed by Weerpana et al., the SP3-Rox method offers several clear advantages over these prior techniques. First, our reduced cost isotopically labeled IAA-reagents make subjecting samples to higher concentrations of reagents required for near complete capping of cysteines less cost prohibitive. Second, our use of SP3 sample cleanup streamlines standard redox preparation workflows by eliminating requirements for sample precipitation or other more laborious decontamination steps that frequently cause substantial material loss. As SP3 allows for repeated rounds of high yielding binding and elution to carboxyl-coated magnetic beads, samples can readily be capped, reduced, capped, and subjected to click chemistry, all in the same pot with minimal manipulation.

Critical to the success of the SP3-Rox method was our application of the FragPipe computational pipeline for highly sensitive peptide identification, as well as rigorous

vetting of the FragPipe's built-in quantification module IonQuant by comparison with Skyline quantification. While proteomics tools such as FragPipe and Skyline have been widely adopted by the proteomics community, their use in chemoproteomics and redox proteomics has been limited to a handful of examples (40, 88, 89). Building upon our prior findings that MSFragger is compatible with search of chemoproteomics data generated using a FAIMS device for online fractionation, we first confirmed that search of samples labeled with the heavy and light IAA probes at a 1:1 stoichiometry afforded comparable PSMs for light- and heavy-labeled peptides. Next, we evaluated the performance of IonQuant and Skyline for quantification of samples labeled with different ratios to IAA probes (1:1, 4:1, 1:10), which revealed comparable performance for both quantification methods. Comparison of quantification for samples analyzed with and without FAIMS revealed improved quantification with FAIMS, albeit with modestly increased ratio compression. Of note, we find that FragPipe-processing affords less substantial FAIMS-induced compression when compared to Skyline. As chemoproteomics workflows have only recently incorporated online fractionation methods such as FAIMS, our work highlights the utility of FAIMS for simultaneously improving coverage and quantification of peptides detected in chemoproteomics studies. To enable widespread adoption of these free and versatile search and quantification tools by the chemoproteomics community, we provide a workflow file that automates implementation of our analysis pipeline (supporting file Sp3-Rox.workflow), and we include detailed step-by-step methods for our data analysis workflows.

We then combined our isotopically labeled IAA reagents, data analysis pipeline, and SP3 cleanup to generate the SP3-Rox method. Application of this method to Jurkat

cell lysates identified 9,687 cysteines, including 219 known disulfides. In lysates subjected to 1 mM of the oxidant GSNO, we identified 2,072 oxidant-sensitive cysteines; follow-up studies should consider a dose-range of GSNO concentrations to mimic native cellular environments. One potential limitation of the SP3-Rox method is its reliance on a ratio-to-ratio comparison between oxidant and vehicle-treated samples, which may reduce the coverage of cysteines found in tough-to-detect peptides, including low abundance sequences and those that ionize poorly. This limitation is in part mitigated by the high coverage of the SP3-Rox method, which is achieved by our combination of Orbitrap Eclipse mass spectrometer (Thermo Scientific), FAIMS online fractionation, and enhanced coverage afforded by SP3 sample cleanup. One potential advantage of using relative ratio changes is the ability to control for any artificial oxidation introduced in sample preparation steps. The small and comparable amounts of methionine oxidation observed here for samples generated using two different lysis methods (sonication versus freeze/thaw lysis in urea) supports the generally minimal impact of artificial oxidation on our analyses (supplemental Fig. S8). In follow-up studies measuring general oxidation, steps should be taken to avoid artificial oxidation such as using mild and inactivating lysis conditions when possible.

Use of alternative methods, such as isoTOP-ABPP, which rely on competitive decrease in IAA labeling upon cysteine oxidation, provide a complementary alternative to the SP3-Rox method, with the possible advantage of increased coverage of tough-to-detect residues. However, such methods can be confounded by changes in protein abundance. In contrast, the SP3-Rox method and other related platforms should prove largely insensitive to expression-level changes, with the exception of complete activation

or inactivation of a gene, and thus are well suited to applications where substantial proteome remodeling is expected. Furthermore, we expect that integration of the SP3-Rox method with probes tailored to capture specific oxidative modifications (e.g., recently repurposed Wittig reagents (90)) likely will capture modification sites missed by in vitro labeling studies, which should prove particularly relevant for more labile or transient modifications.

Multiple redox active cysteines are often found in close proximity, both in peptide sequence and 3D space. Many search and quantification algorithms are not well equipped to handle peptides containing multiple modifications. Addressing this issue, here, we show that FragPipe is capable of identifying multi-cysteine containing peptides with two or more different modifications and that IonQuant and Skyline both accurately quantify isotopic ratios of labeled peptides. We find that PTMProphet can be applied to score the confidence of labeling sites. Collectively, these findings provide a toolbox for the chemoproteomics community to enable rapid and accurate search and quantitation of chemoproteomics datasets, including for multi-cysteine containing peptides. While these multi-cysteine peptides represent a smaller fraction of all data, given their potential biological relevance, we anticipate that this localization data should prove useful for future studies aimed at functional annotation of the redox-sensitive cysteinome. Future efforts will include the incorporation of PTMProphet localizations in IonQuant quantification results.

Showcasing the utility of the SP3-Rox method, we apply our platform to identify cell-state dependent changes in cysteine oxidation. We find that during T cell activation, a number of cysteines show increased oxidation during cellular activation. As prior studies

have demonstrated the contributions of low-level ROS to T cell activation and proliferation, we expect that a subset of the residues identified here are likely functional regulators of cellular activation. Although our workflow controls for protein abundance changes upon activation, we further used unmatched RNA-seq data to rule out confounding expression changes of genes harboring identified residues; future studies would benefit from matched transcriptomic data. The clinical relevance of cysteine-reactive molecules as immunomodulators points to future opportunities in the design and synthesis of tailored compounds aimed at targeting such functional residues (91). Future efforts to functionally validate the impact of these oxidative modifications should help to stratify functional and bystander oxidative modifications.

Data Availability

The MS data have been deposited to the ProteomeXchange Consortium (<http://proteomecentral.proteomexchange.org/>) via the PRIDE (92, 93) partner repository with the dataset identifiers PXD029500 and PXD031647. File and peptide details are listed in supplemental Table S6.

Supplemental data

This article contains supplemental data (28, 29, 30, 34, 38, 42, 43, 93, 94).

Conflicts of interest

The authors declare no competing interests.

Acknowledgments

We thank all members of the Backus lab for helpful suggestions and the CFAR virology core 5P30 AI028697 for providing donor T cells as well as the UCLA Proteome Research Center for assistance with mass spectrometry-based proteomic data collection. We thank Nicholas Shulman for providing the custom Skyline plugin.

Funding and additional information

This study was supported by a Beckman Young Investigator Award (K. M. B.), V Scholar Award V2019-017 (K. M. B.), UCLA Jonsson Comprehensive Cancer Center Seed Grant (K. M. B.), TRDRP T31DT1800 (T. Y.), and the National Institutes of Health grant R01-GM-094231 (A. I. N.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

H. S. D. and K. M. B. conceptualization; H. S. D., T. Y., and M. V. investigation; H. S. D. and T. Y. methodology; H. S. D. formal analysis; H. S. D. validation; H. S. D. and F. Y. data curation; H. S. D. visualization; H. S. D. and K. M. B. writing—original draft; H. S. D., A. W. S., A. I. N., and K. M. B. writing—review and editing; A. W. S. resources; A. I. N. and K. M. B. supervision; A. I. N. and K. M. B. funding acquisition.

Abbreviations

The abbreviations used are: CuAAC, copper-catalyzed azide–alkyne cycloaddition; FAIMS, field asymmetric wave form ion mobility spectrometry; GSNO, S-nitrosoglutathione; IA, iodoacetamide; IAA, iodoacetamide alkyne; IPIAA, isopropyl iodoacetamide alkyne; PSMs, peptide-spectrum matches; ROS, reactive oxygen species; RNS, reactive nitrogen species; SP3, single-pot, solid-phase enhanced sample preparation; TCEP, tris(2-carboxyethyl) phosphine.

Received November 23, 2021, and in revised form, February 17, 2022 Published,
MCPR0 Papers in Press, February 25, 2022, [https://
doi.org/10.1016/j.mcpro.2022.100218](https://doi.org/10.1016/j.mcpro.2022.100218)

Chapter 2 Supporting Information:

SP3-enabled rapid and high coverage chemoproteomic identification of cell-state dependent redox-sensitive cysteines

Supplementary Figures

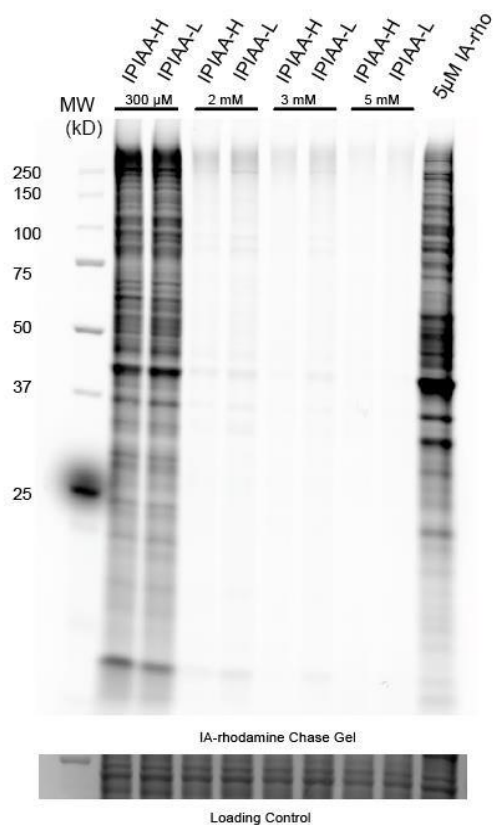


Figure S1. Evaluation of isotopic iodoacetamide alkyne probes by in-gel fluorescence. Gel-based visualization of gradient IPIAA lysate labeling followed by chase with 5 μ M IA-rhodamine in PBS buffer. Loading control was generated using InstantBlue Coomassie Protein Stain.

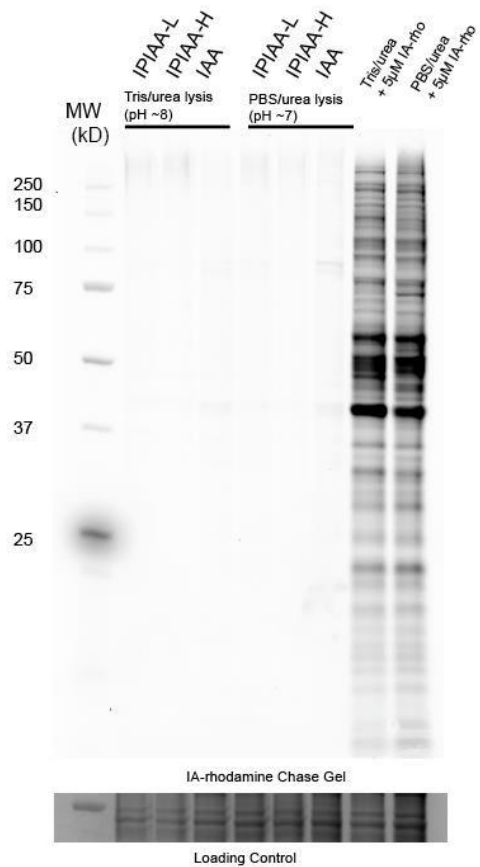


Figure S2. Evaluation of isotopic iodoacetamide alkyne probes by in-gel fluorescence. Gel-based visualization of 2mM IPIAA lysate labeling followed by chase with 5 μ M IA-rhodamine in \sim pH7 and \sim pH8 buffers. Loading control was generated using InstantBlue Coomassie Protein Stain.

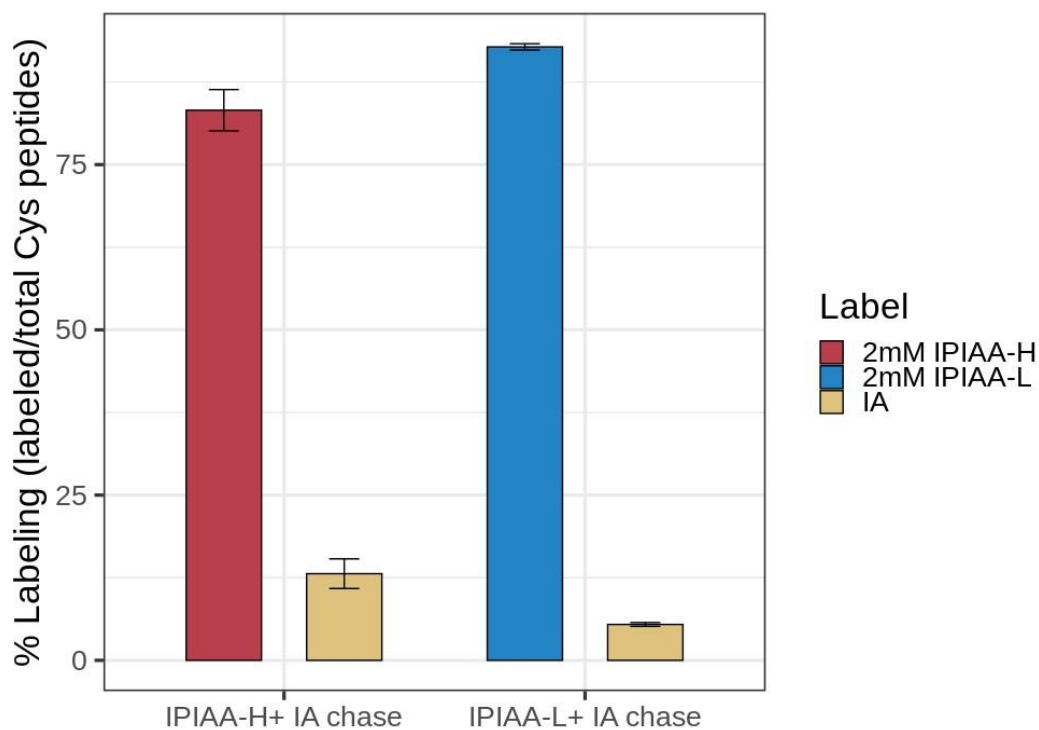


Figure S3. Evaluation of isotopic iodoacetamide alkyne probes by MS. Cysteine labeling with 2mM IPIAA probe followed by chase with 20mM IA by MS analysis. Peptides identified with indicated label as a fraction of the total cysteine containing peptides identified. Experiments were performed in triplicate.

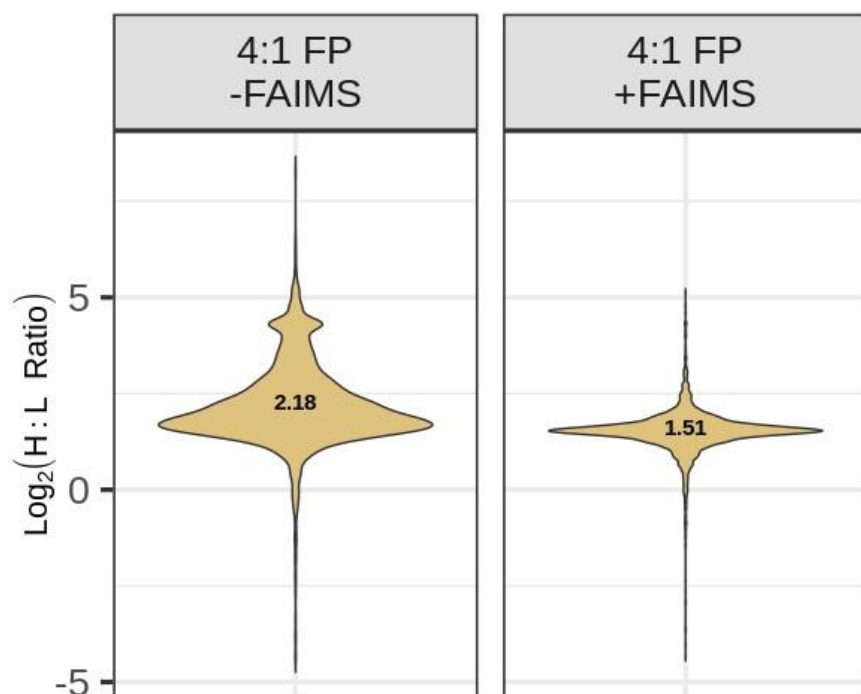


Figure S4. Comparison of quantification with and without FAIMS for 4:1, H:L samples using FragPipe-Ion Quant (FP). Experiments were performed in triplicate (Figure 2). Data provided in **Table S1**.

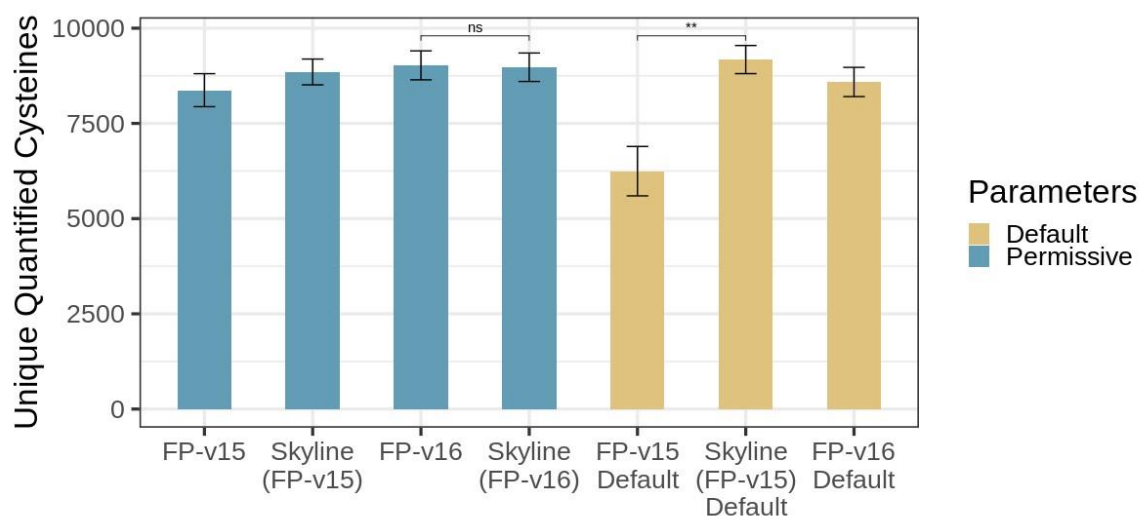


Figure S5. Comparison of IonQuant quantification using previous FragPipe iterations. Details of parameters outlined in Supplementary Information. Statistical significance was calculated using unpaired Student's t-test, ** $p < 0.01$. Experiments were performed in triplicate (Figure 3). Data provided in **Table S2**.

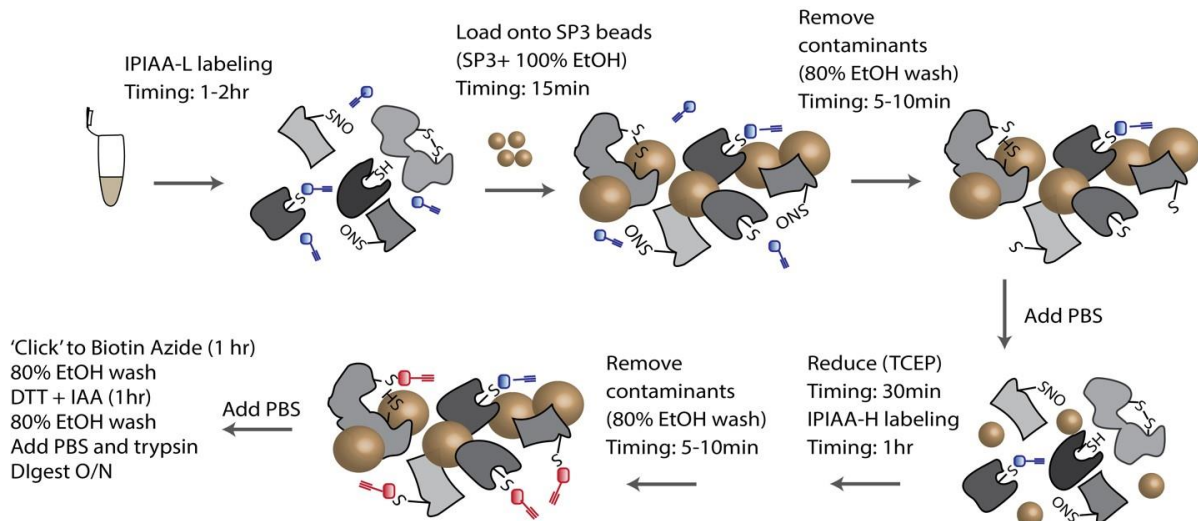


Figure S6. SP3 workflow detailing steps and timing.

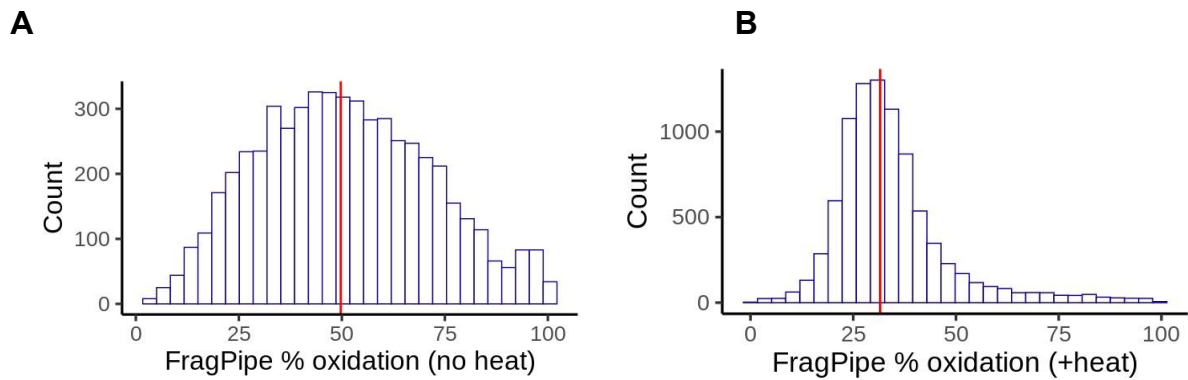


Figure S7. Histogram of FragPipe ratios converted to % oxidation generated using Figure 3A workflow +/- heat (65C) prior to light labeling. Red lines indicate median values: A) 49.7%, no heat and B) 31.5%, with heat denaturation. Data provided in **Table S2**.

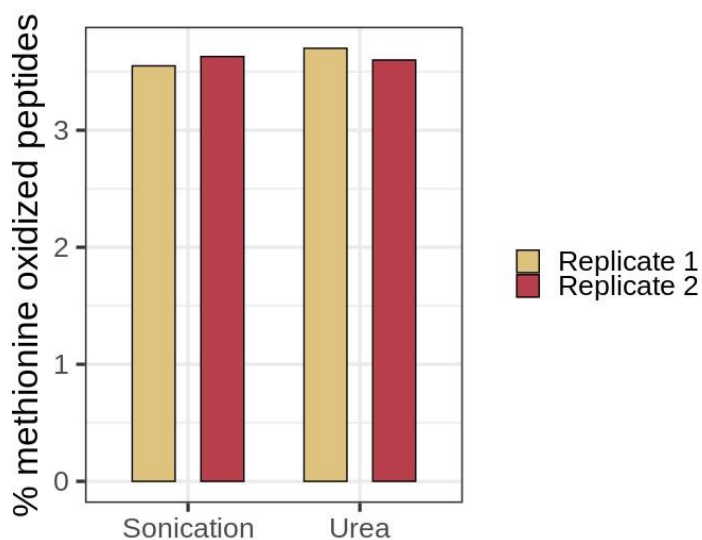


Figure S8. Comparison of methionine oxidation from urea and sonication lysis in HEK293T cells. Data provided in **Table S6**.

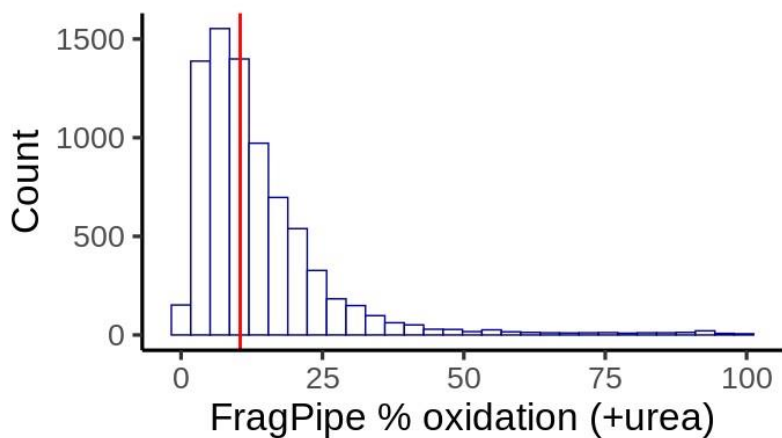


Figure S9. Histogram of FragPipe ratios converted to % oxidation generated from Figure 3B data using denaturing conditions. Median value indicated in red: 10.5%. Data provided in **Table S2**.

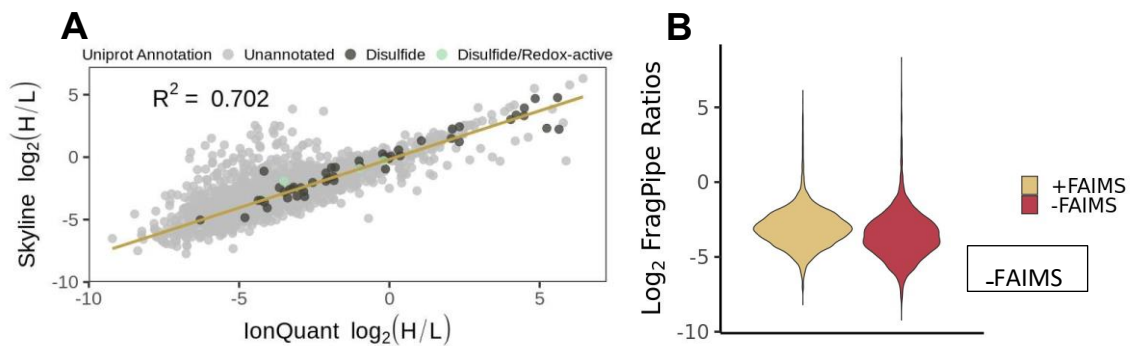


Figure S10. Comparison of FragPipe ratios Figure 3A workflow +/- FAIMS. (A) Pairwise comparison of Skyline and Fragpipe ratios corresponding to samples showing in Figure 3B analyzed without FAIMS (-FAIMS) (B) Distribution of ratios +/- FAIMS. Experiments were performed in triplicate. Data provided in **Table S2**.

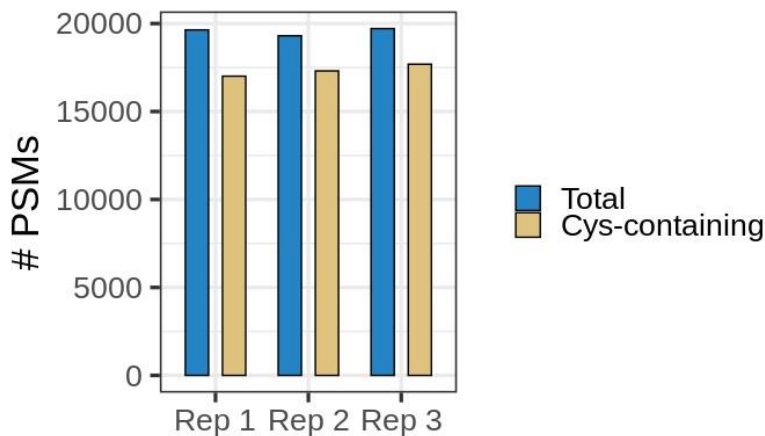


Figure S11. Enrichment efficiency of cysteine-containing peptides from 3 replicate experiments using SP3-Rox workflow (Figure 3B). Data provided in **Table S6**.

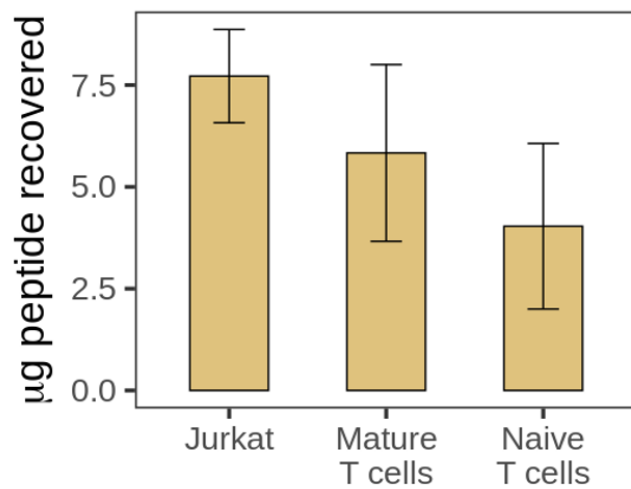


Figure S12. Recovered amount of peptide following SP3-Rox workflow using input of 400 µg protein in 3 replicate samples per cell line.

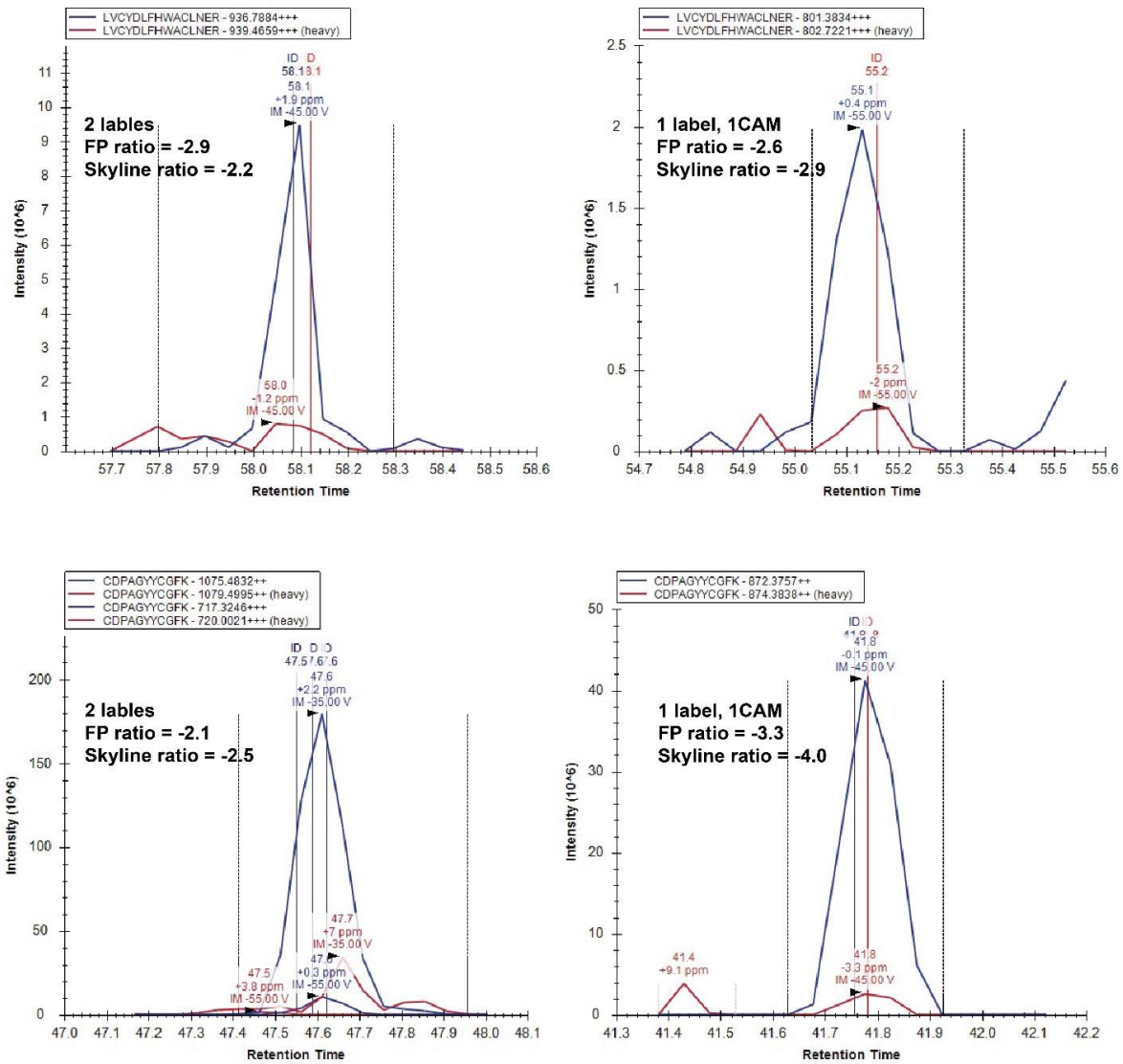


Figure S13. MS1 peak areas of representative multi-cysteine containing peptides in ZFPL (top) and PSA6 (bottom). Indicated peptides with detection of only 1 heavy/light label and 1 carbamidomethylation (CAM) or 2 heavy/light labels per pair of precursors. FragPipe ratio is reported $\log_2(H/L)$ ratio.

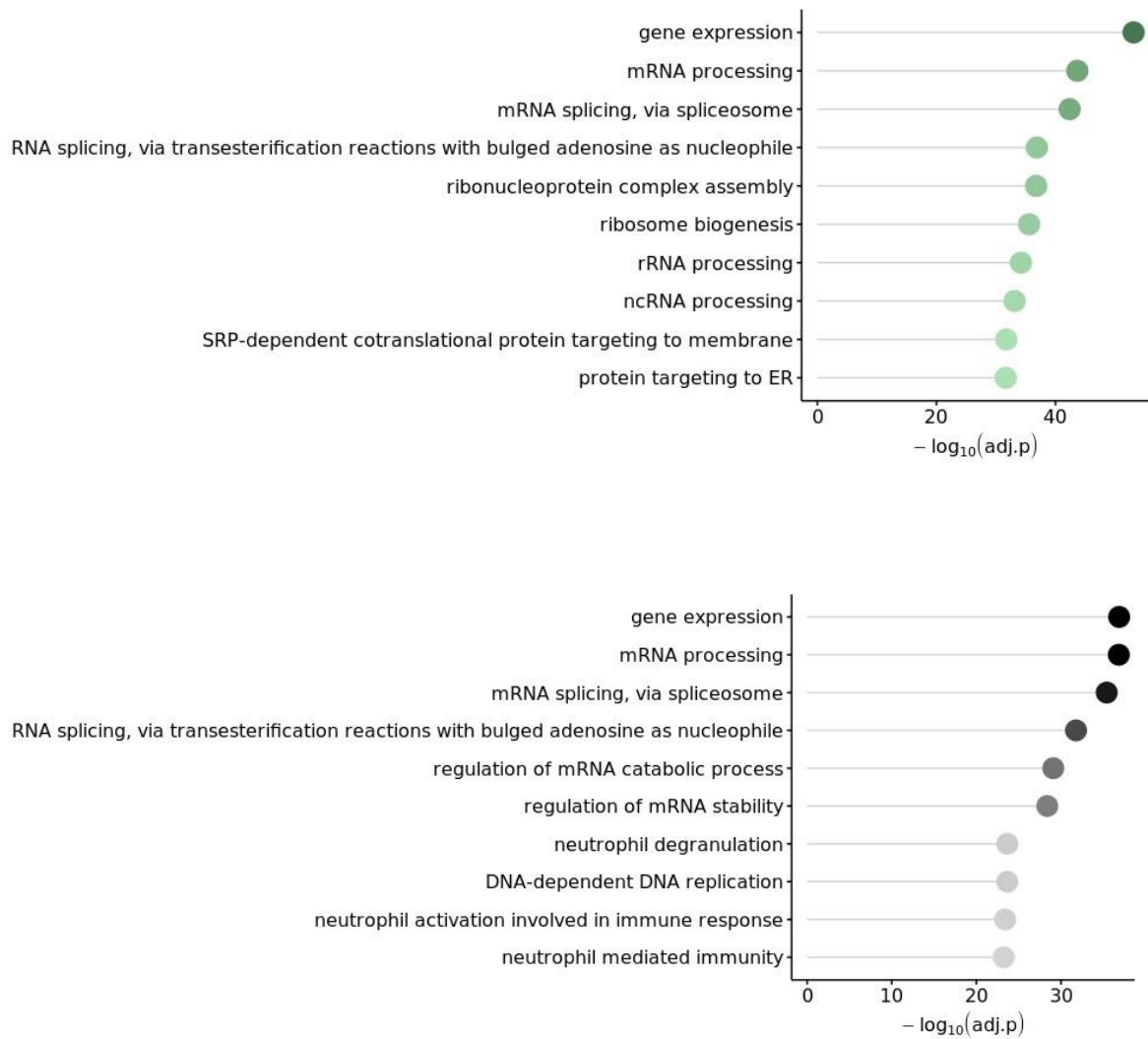


Figure S14. GO-analysis of GSNO-sensitivity datasets. Enriched GO terms of significant and high ratio-change (Difference > 2, p-value < 0.05) sites (green) and GO terms for all other sites (grey). Data aggregated from triplicate experiments (Figure 4B)

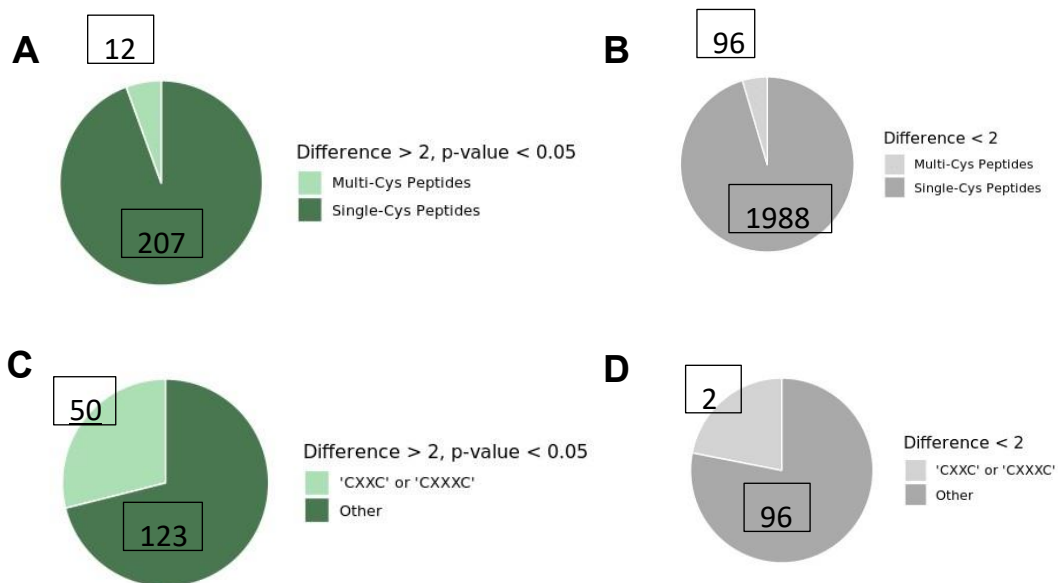


Figure S15. Multi-cysteine peptide analysis from Figure 4 data (A) Fraction of quantified peptides containing multiple cysteines of those with a significant difference in \log_2 ratios > 2 upon GSNO treatment (green) (B) Fraction of quantified peptides containing multiple cysteines for all other sites (grey) (C) Fraction of multi-cys peptides containing redox motifs of those with a significant difference in \log_2 ratios > 2 upon GSNO treatment (green) (D) Fraction of multicys peptides containing redox motifs for all other sites (grey). Data aggregated from triplicate experiments (Figure 4B). Data aggregated from **Table S4**.

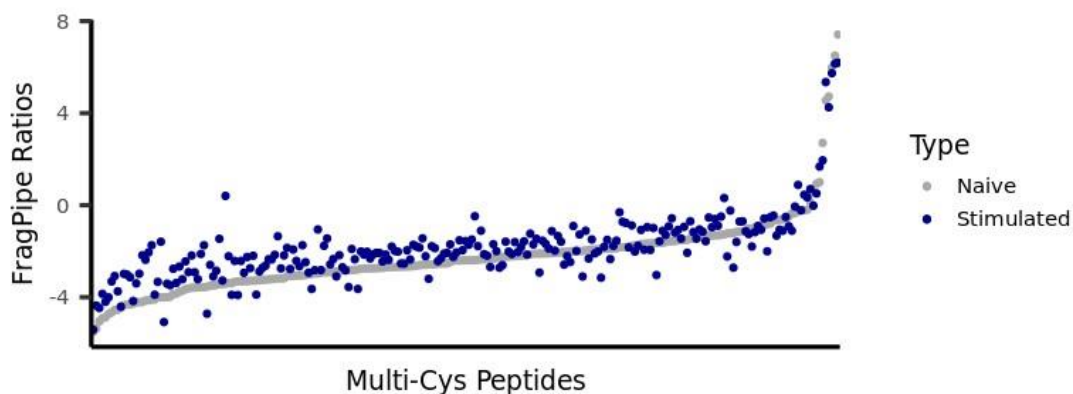
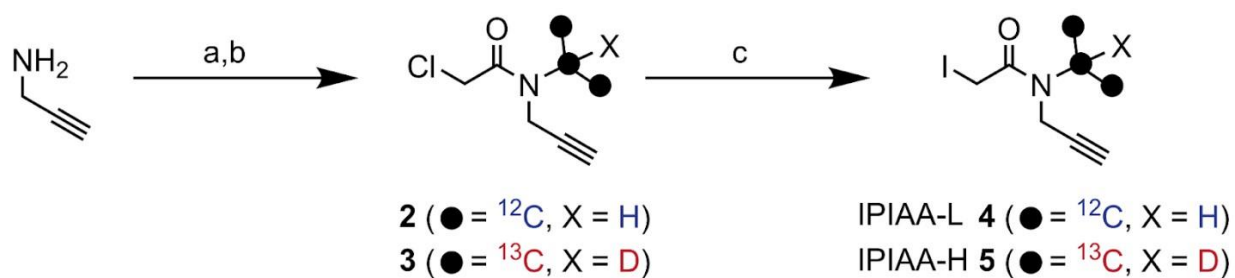


Figure S16. Distribution of multi-cys peptide ratios from peptides quantified in all T cell experiments. Ratios shown are $\text{Log}_2(\text{H/L})$. Data provided in **Table S5**.

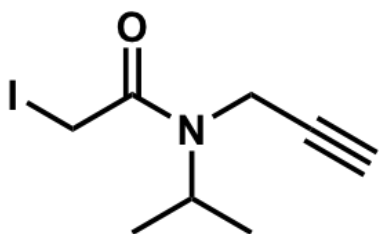
Methods

General Synthetic Methods. General Methods. All reactions were performed in oven dried glassware under an inert atmosphere unless stated otherwise. Silica gel P60 (SiliCycle) was used for column chromatography and SiliCycle 60 F254 silica gel (precoated sheets, 0.25 mm thick) was used for analytical thin layer chromatography. Plates were visualized by fluorescence quenching under UV light or by staining with iodine. Other reagents were purchased from SigmaAldrich (St. Louis, MO), Alfa Aesar (Ward Hill, MA), EMD Millipore (Billerica, MA), Fisher Scientific (Hampton, NH), Oakwood Chemical (West Columbia, SC), Combi-blocks (San Diego, CA), Click Chemistry Tools (Scottsdale, AZ) and Cayman Chemical (Ann Arbor, MI) and used without further

purification. ^1H NMR and ^{13}C NMR spectra for characterization of new compounds and monitoring reactions were collected in CDCl_3 or C_6D_6 (Cambridge Isotope Laboratories, Cambridge, MA) on a Bruker AV 400 or 500 MHz spectrometer in the Department of Chemistry & Biochemistry at University of California, Los Angeles. All chemical shifts are reported in the standard notation of parts per million using the peak of residual proton signals of the deuterated solvent as an internal reference. Coupling constant units are in Hertz (Hz). Splitting patterns are indicated as follows: br, broad; s, singlet; d, doublet; t, triplet; q, quartet; m, multiplet; dd, doublet of doublets; dt, doublet of triplets.



Scheme S1: A. Synthesis of heavy and light iodoacetamide alkyne probes. a. Acetone (^{13}C or ^{12}C), NaBH_3CN or NaBD_3CN , $\text{CH}_2\text{Cl}_2/\text{MeOH}$ (2:1), rt b. Chloroacetyl chloride, NEt_3 , c. NaI , acetone, CH_2Cl_2 , 6.2% yield for **4** over 3 steps, 11% yield for **5** over 3 steps.

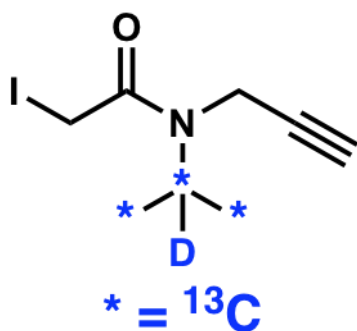


2-Iodo-N-isopropyl-N-(prop-2-yn-1-yl)acetamide (4).

To a solution of propargyl amine (1.16 mL, 18.2 mmol, 1.1 equiv) in dichloromethane (30 mL, 0.55 M) was added acetone (1.22 mL, 16.5 mmol, 1 equiv) at room temperature. The solution was stirred at room temperature for 30 min and then sodium triacetoxyborohydride (4.62 g, 19.8 mmol, 1.2 equiv) was added in several portions. The solution was stirred for 16 h at room temperature and was then quenched with 10 mL saturated NaHCO₃ (aq). The organic solvent was removed and the aqueous phase was then extracted with dichloromethane (2 x 10 mL) followed by extraction with EtOAc (2 x 10 mL). The combined organic layers were dried over Na₂SO₄ and then concentrated to provide the crude reductive amination product as a yellow oil. This crude mixture was carried on directly to the acylation step:

The crude amine from the reductive amination step was dissolved in dichloromethane (30 mL) and cooled to 0 °C. NEt₃ (3.8 g, 24.8 mmol, 1.5 equiv) was added followed by chloroacetyl chloride (2.17 g, 24.8 mmol, 1.5 equiv) dropwise. A precipitate formed within seconds of addition of chloroacetyl chloride. The reaction was allowed to warm to room temperature and then stirred for 30 min. It was then quenched with 10 mL saturated NaHCO₃ (aq) and diluted with dichloromethane (10 mL). The aqueous phase was extracted with dichloromethane (3 x 15 mL). The combined organic phases were dried over anhydrous Na₂SO₄ and concentrated under reduced pressure. The crude product was purified through a silica plug (10 → 20 % EtOAc/Hexanes) to yield 430 mg of chloroacetamide as a yellow oil, which was impure by ¹H NMR. We elected to carry this impure product through the final iodination step:

To a solution of 360 mg of light chloroacetamide **2** dissolved in acetone (10 mL, 2M) was added NaI (1.5 g, 10.4 mmol, 5 equiv) at room temperature. A brown precipitate formed within seconds of addition of NaI. The reaction was stirred for 30 min at room temperature and was then directly diluted with 5 mL saturated NaCl(aq). The aqueous phase was extracted with EtOAc (3 x 5 mL). The combined organic phases were dried over anhydrous Na₂SO₄ and concentrated under reduced pressure. The crude iodoacetamide was then purified by silica gel flash chromatography (10 → 20 % EtOAc/hexanes) to afford the light iodoacetamide **4** as a yellow oil (270 mg, 6.2% yield over 3 steps): **¹H NMR (400 MHz, Benzene-*d*₆)** δ 4.71 (hept, *J* = 6.7 Hz, 1H), 3.71 (s, 1H), 3.53 (s, 1H), 3.40 (s, 2H), 3.22 (s, 1H), 1.73 (m, 1H), 0.80 (d, *J* = 6.8 Hz, 6H). **¹³C NMR (101 MHz, Benzene-*d*₆)** δ (a mixture of two rotamers) 167.0, 166.3, 128.1, 81.1, 80.2, 72.3, 70.6, 50.3, 46.2, 32.4, 29.7, 20.5, 19.4, -2.00, -3.24. **HRMS (ESI):** *m/z* calc'd for C₈H₁₃INO [M+H]⁺: 266.0042, found 266.0065.



2-iodo-*N*-(prop-2-yn-1-yl)-*N*-(propan-2-yl-1,2,3-¹³C₃-2-*d*)acetamide (5). To a solution of propargyl amine (384 μL, 6.0 mmol, 1.1 equiv) in dichloromethane (10 mL, 0.55 M) was added **Acetone-¹³C₃** (333 mg, 5.5 mmol, 1 equiv) at room temperature. The solution was stirred at room temperature for 30 min and then sodium cyanoborodeuteride (431 mg,

6.5 mmol, 1.2 equiv) was added in several portions. The solution was stirred for 16 h at room temperature and was then quenched with 5 mL saturated NaHCO₃ (aq). The organic solvent was removed and the aqueous phase was then extracted with dichloromethane (2 x 10 ml) followed by extraction with EtOAc (2 x 10 mL). The combined organic layers were dried over Na₂SO₄ and then concentrated to provide the crude reductive amination product as a yellow oil. This crude mixture was carried on directly to the acylation step:

The crude heavy amine from the reductive amination step was dissolved in dichloromethane (10 mL) and cooled to 0 °C. NEt₃ (1.14 mL, 8.2 mmol, 1.5 equiv) was added followed by chloroacetyl chloride (644 g, 8.2 mmol, 1.5 equiv) dropwise. A precipitate formed within seconds of addition of chloroacetyl chloride. The reaction was allowed to warm to room temperature and then stirred for 30 min. It was then quenched with 5 mL saturated NaHCO₃ (aq) and diluted with dichloromethane (5 mL). The aqueous phase was extracted with dichloromethane (3 x 10 ml). The combined organic phases were dried over anhydrous Na₂SO₄ and concentrated under reduced pressure. The crude product was purified through a silica plug (10 → 20 % EtOAc/Hexanes) to yield 330 mg of heavy chloroacetamide as a yellow oil, which was impure by ¹H NMR. We elected to carry this impure product through the final iodination step:

To a solution of 140 mg of heavy chloroacetamide **3** dissolved in acetone (5 mL, 0.16 M) was added NaI (591 mg, 3.9 mmol, 5 equiv) at room temperature. A brown precipitate formed within seconds of addition of NaI. The reaction was stirred for 30 min at room

temperature and was then directly diluted with 5 mL saturated NaCl(aq). The aqueous phase was extracted with EtOAc (3 x 5 mL). The combined organic phases were dried over anhydrous Na₂SO₄ and concentrated under reduced pressure. The crude iodoacetamide was then purified by silica gel flash chromatography (10 → 20 % EtOAc/hexanes) to afford the heavy iodoacetamide **5** as a yellow oil (74 mg, 11% yield over 3 steps): **¹H NMR (400 MHz, Benzene-*d*₆)** δ (a mixture of rotamers) 3.78 – 3.66 (m, 1H), 3.52 (s, 1H), 3.44 – 3.32 (m, 1H), 3.21 (s, 1H), 1.81 – 1.66 (m, 1H), 1.05 – 0.88 (m, 3H), 0.68 – 0.57 (m, 3H). **¹³C NMR (101 MHz, C₆D₆)** δ (13C spectrum exhibits complex splitting due to the presence of deuterium in the molecule) 174.6, 128.3, 128.1, 127.8, 50.7, 50.5, 50.4, 50.2, 50.0, 49.8, 49.6, 47.2, 47.0, 46.9, 46.6, 46.5, 46.4, 46.3, 46.1, 46.1, 45.9, 45.7, 20.81, 20.45, 20.4, 20.3, 20.0, 19.9, 19.7, 19.4, 19.3, 19.0, 18.9, -6.7. **HRMS (ESI):** *m/z* calc'd for C₅¹³C₃H₁₂DINO [M+H]⁺: 270.0205, found 270.0197.

Compounds 6 (IAA) and 7 (biotin-azide) were prepared as previously reported (43). Rhodamineazide was purchased from Click Chemistry Tools (AZ109-5).

Cell culture and preparation of cell lysates. Cell culture reagents including Dulbecco's phosphate-buffered saline (DPBS), Dulbecco's modified Eagle's medium (DMEM)/high glucose media, Roswell Park Memorial Institute (RPMI) media, trypsin-EDTA and penicillin/streptomycin (Pen/Strep) were purchased from Fisher Scientific. Fetal Bovine Serum (FBS) were purchased from Avantor Seradigm (lot # 214B17). All cell lines were obtained from ATCC and were maintained at a low passage number (< 20 passages). HEK293T (ATCC: CRL-3216) cells were cultured in DMEM supplemented with 10% FBS

and 1% antibiotics (Penn/Strep, 100 U/mL). Jurkat cells were cultured in RPMI---Media was filtered (0.22 μ m) prior to use. Cells were maintained in a humidified incubator at 37 °C with 5% CO₂. Cells were harvested by centrifugation (4500g, 5 min, 4 °C) and washed twice with cold DPBS. Cell pellets were then lysed with sonication (amp=10, 10 x1 sec pulses) or lysed in 6M urea with one freeze/thaw cycle and clarified by briefly centrifuging (4500g, 1 min). Blood from a healthy donor was obtained from UCLA/CFAR Virology Core (5P30 AI028697) after informed consent. After Trima filter isolation, peripheral blood mononuclear cells (PBMCs) were purified over Ficoll–Hypaque gradients (Sigma-Aldrich) and T cells were purified via negative selection with magnetic beads (EasySep Human T Cell Iso Kit, 17951, STEMCELL). The purified T cells were washed with sterile PBS. Unstimulated cells were harvested by centrifugation. The remaining cells were then resuspended in RPMI-1640 supplemented with FBS, penicillin, streptomycin and glutamine (2 million cells per ml) and 200,000 cells per well were seeded on non-treated tissue culture, 96-well transparent plates that had been coated with anti-CD3 (1:200, BioXcell) and anti-CD28 (1:500, Biolegend) in PBS (100 μ l per well). After 72h, the cells were then harvested, washed with PBS and the cell pellets lysed by sonication in PBS and proteomics samples prepared as described below. The lysates were then transferred to a new microcentrifuge tube. Protein concentrations were determined using a BioRad DC protein assay kit from Bio-Rad Life Science (Hercules, CA) and the lysate diluted to the working concentrations indicated below.

Gel-based proteome labeling with IPIAA probes. HEK293T proteome (25 μ L of 2 mg/mL) was labeled with 30 μ M IPIAA-L (4), IPIAA-H (5), or IAA (6) (0.75 μ L of 1 mM

stocks) for 1 hour. CuAAC was performed with rho-azide (8) (1 μ L of 1.25 mM stock in DMSO, final concentration = 50 μ M), TCEP (0.5 μ L of fresh 50 mM stock in water, final concentration = 1 mM), TBTA (1.5 μ L of 1.7 mM stock in DMSO/*t*-butanol 1:4, final concentration = 100 μ M), and CuSO₄ (0.5 μ L of 50 mM stock in water, final concentration = 1 mM). Samples were allowed to react for 1h at RT. All samples were denatured (5 min, 95 °C) and analyzed by SDS-PAGE using Bio-Rad Criterion TGX Stain-Free. Loading control images were obtained after Coomassie staining.

IA-rhodamine chase gels. Jurkat proteome (25 μ L of 2 mg/mL) lysed in PBS/2M urea or Tris/2M urea were labeled with 5 mM IPIAA-L (4), IPIAA-H (5), IAA (6), or vehicle (2.5 μ L of 50 mM stocks) for 1 hour at 37°C and chased with 5 μ M IA-rhodamine for 1 hour at 37°C (2.5 μ L of 50 μ M stock). In separate tubes, Jurkat proteome (25 μ L of 2 mg/mL) from PBS/urea lysate was labeled with 300 μ M, 2 mM, 3 mM, and 5 mM IPIAA-H (5) and IPIAA-L (4) (2.5, 1.5 and 1 μ L of 50 mM stocks or 1.25 μ L of 10 mM stock) or vehicle for 1 hour at 37°C and chased with 5 μ M IA-rhodamine for 1 hour at 37°C (2.5 μ L of 50 μ M stock). All samples were denatured (5 min, 95 °C) and analyzed by SDS-PAGE using Criterion TGX Stain-Free gels obtained from Bio-Rad. Loading control images were obtained after Coomassie staining.

IPIAA labeling MS experiments. Jurkat proteome (25 μ L of 2 mg/mL) from PBS/urea lysate was labeled with 2 mM IPIAA-H (5) or IPIAA-L (4) (1 μ L of 50 mM stocks) for 1 hour at RT and chased with 20 mM iodoacetamide (1.2 μ L of 400 mM stock) for 1 hour at

RT. 5 μ L of a 1:1 bead mixture was washed using a magnetic rack as described below in *SP3 sample loading*. 5 μ L of the bead mixture was added to each sample and incubated for 10 min at 1000rpm at RT. Absolute ethanol (50 μ L) was added to each sample, and the samples were incubated for 5 min at RT with shaking (1000 rpm). Samples were then placed on a magnetic rack, and beads allowed to settle. Supernatants were then removed and discarded. Using the magnetic rack, the beads were further washed two times with 80% ethanol in water (50 μ L). Samples were resuspended in 2M urea/PBS (50 μ L 2M urea/PBS) and subjected to digest with 1 μ L trypsin overnight. Digest was allowed to proceed overnight at 37 °C with shaking. After digestion, Acetonitrile (495 μ L, > 95% of the final volume) was added to each sample and the mixtures were incubated for 10 min at RT with shaking (1000 rpm). Supernatants were then removed and discarded using the magnetic rack, and the beads were washed (3 \times 125 μ L acetonitrile). Peptides were then eluted from SP3 beads with 94 μ L of 2% DMSO in MB water for 30 min at 37°C with shaking (1000 rpm). To each elution, 5 μ L of acetonitrile and 1 μ L formic acid was added. Samples were analyzed by LC-MS/MS. Experiment was done in triplicate for each IPIAA label. For buffer comparison, PBS/urea and Tris/urea lysates (25 μ L) from gel-chase experiments were labeled with 2 mM IPIAA-H, IPIAA-L, or IAA (1 μ L from 50 mM stocks). Samples were subjected to SP3 sample loading, wash, and digest as described above and analyzed by LC-MS/MS.

IPIAA quantification MS/FAIMS experiments. IPIAA probes were mixed in ratios of 1:1, 4:1 using initial concentrations of 500mM. HEK293T proteome (200 μ L of 1 mg/mL) in 1mM DTT was labeled with 2mM probe mixture (0.8 μ L of 500 mM stocks) for 1 hour at

RT. CuAAC was performed with biotin-azide (**7**) (4 μ L of 200 mM stock in DMSO, final concentration = 4 mM), TCEP (4 μ L of fresh 50 mM stock in water, final concentration = 1 mM), TBTA (12 μ L of 1.7 mM stock in DMSO/*t*-butanol 1:4, final concentration = 100 μ M), and CuSO₄ (4 μ L of 50 mM stock in water, final concentration = 1 mM). Samples were allowed to react for 1h at RT. After CuAAC, 10 μ L of 20% SDS was added to each sample. Samples were incubated with 0.5 μ L benzonase (Fisher Scientific, 70-664-3) for 30 min at 37°C. Samples were loaded onto beads as described in *SP3 sample loading*. After 10 min incubation at 1000rpm, samples were then subjected to *SP3 digest and elution* described below followed by *Neutravidin enrichment* described below. Experiment was conducted in duplicate for 1:1 samples and triplicate for 4:1 samples. Samples were identically prepared in triplicate for FAIMS analysis using 1:10 and 4:1 IPIAAH:IPIAA-L ratios for labeling in Jurkat 2mg/mL proteome samples.

Proteomic sample preparation for urea and sonication lysis comparison. HEK293T cells were harvested and lysed with sonication (amp=10, 10 x1 sec pulses) or lysed in 6M urea with one freeze/thaw cycle. Samples were adjusted to 2 mg/mL, 2M urea/PBS. Samples were then labeled with 2 mM IAA (2 μ L of 200 mM stock solution in DMSO, final concentration = 2 mM) for 1h at RT (700 rpm). Samples were incubated with 0.5 μ L benzonase (Fisher Scientific, 70-6643) for 30 min at 37°C. The samples were then subjected to *SP3 sample loading*, *SP3 digest and elution*, *NeutrAvidin enrichment* and *LC-MS/MS analysis*, as described below. Experiments were conducted in duplicate for each condition.

Proteomic sample preparation for disulfide analysis. Jurkat proteome samples (200 μ L of 2 mg/mL, prepared as described in *preparation of cell lysates*) were incubated for 5 min in 2M urea. Samples were then labeled with 2mM IPIAA-L (2 μ L of 200 mM stock solution in DMSO, final concentration = 2 mM) for 1h at RT (700rpm). Samples were incubated with 0.5 μ L benzonase (Fisher Scientific, 70-664-3) for 30 min at 37°C. The samples were then subjected to *SP3 sample loading, SP3-Rox reduction and CuAAC, SP3 digest and elution, NeutrAvidin enrichment and LCMS/MS analysis*, as described below. Experiments were conducted in duplicate for each condition with an additional technical replicate per injection, totaling 4 samples per condition.

Proteomic sample preparation for disulfide analysis and GSNO sensitivity experiments. Jurkat proteome samples (200 μ L of 2 mg/mL, prepared as described in *preparation of cell lysates*) were incubated with vehicle or 1mM GSNO for 30 min. Samples were then labeled with 2mM IPIAA-L (2 μ L of 200 mM stock solution in DMSO, final concentration = 2 mM) for 1h at RT (700 rpm). Samples were incubated with 0.5 μ L benzonase (Fisher Scientific, 70-664-3) for 30 min at 37°C. The samples were then subjected to *SP3 sample loading, SP3-Rox reduction and CuAAC, SP3 digest and elution, NeutrAvidin enrichment and LC-MS/MS analysis*, as described below. Experiments were conducted in duplicate for each condition with an additional technical replicate per injection, totaling 4 samples per condition.

Proteomic sample preparation for T-cell experiments. Proteome samples (200 μ L of 2 mg/mL, prepared as described in *preparation of cell lysates*) were labeled with 2mM

IPIAA-L (4) (2 μ L of 200 mM stock solution in DMSO, final concentration = 2 mM) for 1 h at RT (700rpm). Samples were incubated with 0.5 μ L benzoylase (Fisher Scientific, 70-664-3) for 30 min at 37°C. The samples were then subjected to *SP3 sample loading, SP3-Rox reduction and CuAAC, SP3 digest and elution, NeutrAvidin enrichment and LC-MS/MS analysis*, as described below. Experiments were conducted in duplicate for each condition with an additional technical replicate per injection, totaling 4 samples per condition.

SP3 sample loading. SP3 sample cleanup was performed generally at a bead/protein ratio of 10:1 (wt/wt) (38). For each 200 μ L sample (2 mg/mL or 1 mg/mL protein concentration), 20 μ L Sera-Mag SpeedBeads Carboxyl Magnetic Beads, hydrophobic (GE Healthcare, 65152105050250, 50 μ g/ μ L, total 1 mg) and 20 μ L Sera-Mag SpeedBeads Carboxyl Magnetic Beads, hydrophilic (GE Healthcare, 45152105050250, 50 μ g/ μ L, total 1 mg) were aliquoted into a single microcentrifuge tube and gently mixed. Tubes were then placed on a magnetic rack until the beads settled to the tube wall, and the supernatants were removed. The beads were removed from the magnetic rack, reconstituted in 1 mL of MB water, and gently mixed. Tubes were then returned to the magnetic rack, beads allowed to settle, and the supernatants removed. Washes were repeated for two more cycles, and then the beads were reconstituted in 40 μ L MB water. The bead slurries were then transferred to the proteome samples, incubated for 10 min at RT with shaking (1000 rpm). Samples were then subjected to *SP3-Rox reduction and CuAAC*.

SP3-Rox reduction and CuAAC. Samples are prepared as described in *Proteomic sample preparation* and *SP3 sample loading*. Absolute ethanol (400 μ L) was added to each sample, and the samples were incubated for 5 min at RT with shaking (1000 rpm). Samples were then placed on a magnetic rack, and beads allowed to settle. Supernatants were then removed and discarded. Using the magnetic rack as described above, the beads were further washed two times with 80% ethanol in water (400 μ L). Beads were then resuspended in 200 μ L PBS containing 2 M urea. TCEP (10 μ L of 100 mM stock in water, final concentration = 5 mM) was added into each sample and the sample was incubated at 56 °C for 30 min, shaking (500rpm). Absolute ethanol (400 μ L) was added to each sample, and the samples were incubated for 5 min at RT with shaking (1000 rpm). Beads were washed twice with 80% ethanol as described above and resuspended in 200 μ L PBS containing 2 M urea. Samples were then labeled with IPIAA-H (**5**) (2 μ L of 200 mM stock solution in DMSO, final concentration = 2 mM) for 1h at 37 °C with shaking (500 rpm). Absolute ethanol (400 μ L) was added to each sample, and the samples were incubated for 5 min at RT with shaking (1000 rpm). Beads were washed twice with 80% ethanol. Samples were then placed on a magnetic rack, and beads allowed to settle. Supernatants were then removed and discarded. Beads were then resuspended in 200 μ L 0.5% SDS in PBS. CuAAC was performed with biotinazide (**7**) (4 μ L of 200 mM stock in DMSO, final concentration = 4 mM), TCEP (4 μ L of fresh 50 mM stock in water, final concentration = 1 mM), TBTA (12 μ L of 1.7 mM stock in DMSO/*t*-butanol 1:4, final concentration = 100 μ M), and CuSO₄ (4 μ L of 50 mM stock in water, final concentration = 1 mM). Samples were allowed to react for 1h at RT temperature with

shaking (500rpm). Samples were subjected to *SP3 digest and elution* followed by *NeutrAvidin enrichment* as described below.

SP3 digest and elution. Absolute ethanol (400 μ L) was added to each sample, and the samples were incubated for 5 min at RT with shaking (1000 rpm). Beads were washed twice with 80% ethanol as described above. Beads were then resuspended in 200 μ L 0.5% SDS in PBS containing 2 M urea. DTT (10 μ L of 200 mM stock in water, final concentration = 10 mM) was added into each sample and the sample was incubated at 65 °C for 15 min. To this iodoacetamide (10 μ L of 400 mM stock in water, final concentration = 20 mM) was added and the solution was incubated for 30 min at 37 °C with shaking. After that, absolute ethanol (400 μ L) was added to each sample, and the samples were incubated for 5 min at RT with shaking (1000 rpm). Beads were then again washed three times with 80% ethanol in water (400 μ L). Next, beads were resuspended in 150 μ L PBS containing 2 M urea followed by addition of 3 μ L trypsin solution (Worthington Biochemical, LS003740, 1 mg/mL in 666 μ L of 50 mM acetic acid and 334 μ L of 100 mM CaCl₂, final weight = 2 ng). Digest was allowed to proceed overnight at 37 °C with shaking. After digestion, ~ 4 mL acetonitrile (> 95% of the final volume) was added to each sample and the mixtures were incubated for 10 min at RT with shaking (1000 rpm). Supernatants were then removed and discarded using the magnetic rack, and the beads were washed (3 \times 1 mL acetonitrile). Peptides were then eluted from SP3 beads with 100 μ L of 2% DMSO in MB water for 1 hour at 37 °C with shaking (1000 rpm). The elution was repeated again with 100 μ L of 2% DMSO in MB water. Peptide concentration

assay (Pierce, 23275) was performed to test the concentration of the peptide. The elution can be used for NeutrAvidin enrichment or analyzed by LC-MS/MS.

NeutrAvidin enrichment of labeled peptides. For each sample, 50 μ L of NeutrAvidin® Agarose resin slurry (Pierce, 29200) was washed three times in 10 mL IAP (immunoaffinity purification) buffer (50 mM MOPS–NaOH (pH 7.2), 10 mM Na₂HPO₄, 50 mM NaCl) and then resuspended in 500 μ L IAP buffer. Peptide solutions eluted from SP3 beads were then transferred to the NeutrAvidin® Agarose resin suspension, and the samples were then rotated for 2h at RT. After incubation, the beads were pelleted by centrifugation (21,000 g, 1 min) and washed by centrifugation (3 \times 1 mL PBS, 6 \times 1 mL water). Bound peptides were eluted with 60 μ L of 80% acetonitrile in MB water containing 0.1% FA (10 min at RT). The samples were then collected by centrifugation (21,000 g, 1 min) and residual beads separated from supernatants using Micro BioSpin columns (Bio-Rad). The remaining peptides were then eluted from pelleted beads with 60 μ L of 80% acetonitrile in water containing 0.1% FA (10 min, 72 °C). Beads were then separated from the eluants using the same Bio-Spin column. Eluants were collected by centrifugation (21,000 g, 1 min) and dried (SpeedVac). The samples were then reconstituted with 5% acetonitrile and 1% FA in MB water and analyzed by LC-MS/MS.

Liquid-chromatography tandem mass-spectrometry (LC-MS/MS) analysis. The samples were analyzed by liquid chromatography tandem mass spectrometry using a Thermo Scientific™ Orbitrap Eclipse™ Tribrid™ mass spectrometer coupled with a High Field Asymmetric Waveform Ion Mobility Spectrometry (FAIMS) Interface. Peptides were

resuspended in 5% formic acid and fractionated online using a 18cm long, 100 µM inner diameter (ID) fused silica capillary packed in-house with bulk C18 reversed phase resin (particle size, 1.9 µm; pore size, 100 Å; Dr. Maisch GmbH). The 70-minute water acetonitrile gradient was delivered using a Thermo Scientific™ EASY-nLC™ 1200 system at different flow rates (Buffer A: water with 3% DMSO and 0.1% formic acid and Buffer B: 80% acetonitrile with 3% DMSO and 0.1% formic acid). The detailed gradient includes 0 – 5 min from 3 % to 10 % at 300 nL/min, 5 – 64 min from 10 % to 50 % at 220 nL/min, and 64 – 70 min from 50 % to 95 % at 250 nL/min buffer B in buffer A. Data was collected with charge exclusion (1, 8,>8). Data was acquired using a Data-Dependent Acquisition (DDA) method comprising a full MS1 scan (Resolution = 120,000) followed by sequential MS2 scans (Resolution = 15,000) to utilize the remainder of the 1 second cycle time. Time between master scans was set 1 s. HCD collision energy of MS2 fragmentation was 30 %.

FragPipe v16.0 peptide identification and quantitation. Details of parameters are in the supplemental parameter workflow file and below. Raw data collected by LC-MS/MS and converted to mzML format with peakPicking for MS levels 1 and 2 using MSConvert (ProteoWizard release 3.0.20287) (94) were searched using FragPipe GUI v16.0 with MSFragger (version 3.3), Philosopher (version 4.0.0) and IonQuant (version 1.7.5) enabled (28,29,30,34). Precursor and fragment mass tolerance was set as 20 ppm. Missed cleavages were allowed up to 2. Peptide length was set 7 - 50 and peptide mass range was set 500 - 5000. Cysteine residues were searched with variable modifications at cysteine residues for carboxyamidomethylation (+57.02146), IPIAA-L (+463.2366), and

IPIAA-H (+467.2529) labeling allowing for 3 max occurrences and all mods used in first search checked. Peptide and protein level FDR were set to 1%. Permissive IonQuant parameters allowed minimum scan/isotope numbers set to 1. PTMProphet information was obtained from psm.tsv using 'heavy' and 'light' localizations scores.

FragPipe v15.0. peptide identification and quantitation Raw data collected by LC-MS/MS and converted to mzML format with peakPicking for MS levels 1 and 2 using MSConvert (ProteoWizard release 3.0.20287) were searched using FragPipe GUI v15.0 with MSFragger (version 3.2) Philosopher (version 3.4.13) and IonQuant (version 1.5.5) enabled. Precursor and fragment mass tolerance was set as 20 ppm. Missed cleavages were allowed up to 2. Peptide length was set 7 - 50 and peptide mass range was set 500 - 5000. Cysteine residues were searched with variable modifications at cysteine residues for carboxyamidomethylation (+57.02146), IPIAA-L (+463.2366), and IPIAA-H (+467.2529) labeling allowing for 2 max occurrences.

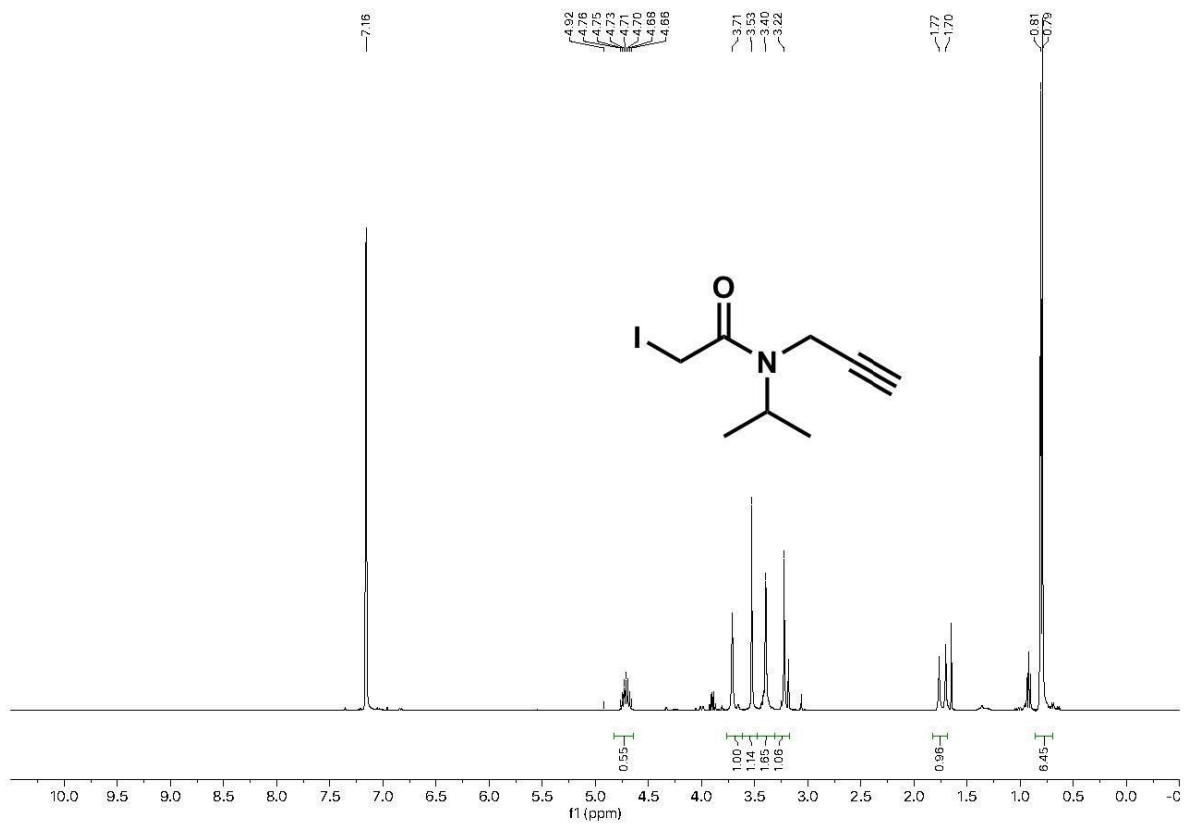
Skyline quantitation. Details below (44). Data was imported into Skyline with a probability threshold corresponding to the 1% peptide-ion level FDR in the dataset. following the standard DDA analysis workflow for isotopically labeled dataset with the following modifications. As Skyline automatically places a heavy isotopic label on all modified cysteines, including carbamidomethylated residues and those modified by the IPIAA-H reagent, quantification fails for peptides containing two or more modified cysteines (e.g. one carbamidomethyl residue and one IPIAA-residue). A custom plugin was generated to remove the heavy mass from all carbamidomethylated residues.

The MS search results have been deposited to the ProteomeXchange Consortium (<http://proteomecentral.proteomexchange.org>) via the PRIDE partner repository (93) with the dataset identifiers PXD029500 and PXD031647.

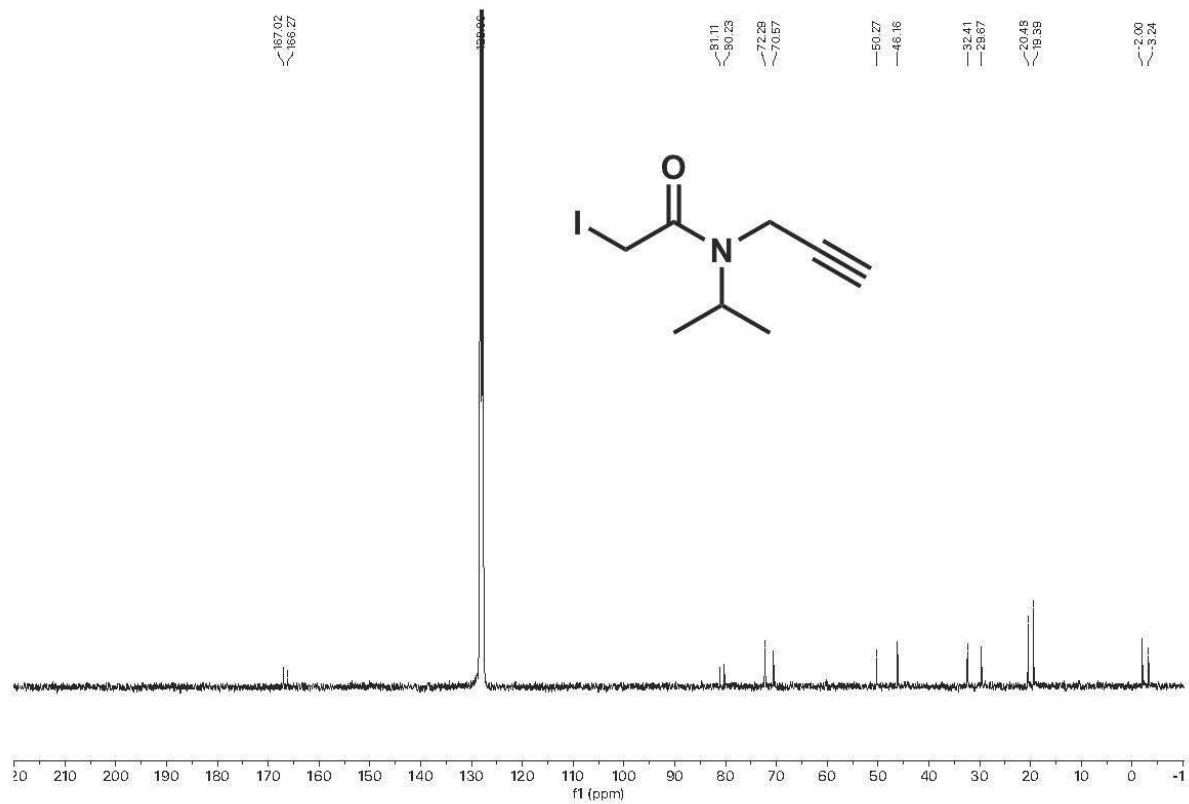
Data Analysis and Statistics. Custom R scripts were implemented to compile peptide_label_quant.tsv outputs from FragPipe to count unique quantified cysteines. Unique cysteines and unique peptide-spectrum matches (PSMs) were quantified for each dataset using unique identifiers consisting of a UniProt protein ID and the amino acid number of the modified cysteine and an additional parameter specifying single or double isotopic labeling (heavy and/or light). Unique proteins were established based on UniProt protein IDs. Residue numbers were found by aligning the peptide sequence to the corresponding UniProt ID protein sequence specified by FragPipe outputs. **Figure S3** data is aggregated from peptide.tsv by counting unique modified and unmodified cysteines containing peptides. All violin plots, **Figure 3B**, and **Figure 3D** data of FragPipe and Skyline quantitative outputs were generated by taking the median H:L ratio among all tryptic peptides for unique cysteines in replicate datasets (peptide_label_quant.tsv); mean ratio values were calculated across replicate datasets; quantified cysteines appearing in at least two replicates with ratio SD < 1 were kept. Skyline ratio values were converted to log₂(H/L) format to match FragPipe output prior to SD-filtering in **Figure 3B** and **Figure 3D**, and duplicated ratios assigned to multiple proteins were aggregated in cysteine counts in **Figure S5**. For **Figure 3C**, single and multi-cys analysis counts were aggregated from each replicate and reported as mean +

SD. **Figure 3E** data is generated from heavy localization scores in psm.tsv for all multi-cysteine containing peptides. The following formula was used to convert unlogged, average H:L ratios (R) to % oxidation: $(R/(1+R))*100$. **Figure 4C** and **Figure 5B** data is aggregated as described by calculating the median of output log₂ ratio values within replicates. Means of reported logged ratio values for each condition (+/-GSNO or unstimulated/stimulated T cells) were calculated for all replicates per condition, and the difference of the log₂ mean values were reported. Variances were calculated for each sample-condition pairing and a corresponding two-sample T-test was performed on the raw ratios to generate p-values for scatter plots (n=4 for +/- GSNO and resting/stimulated T cells); p-values were adjusted for multiple comparisons using Benjamini-Hochberg procedure.

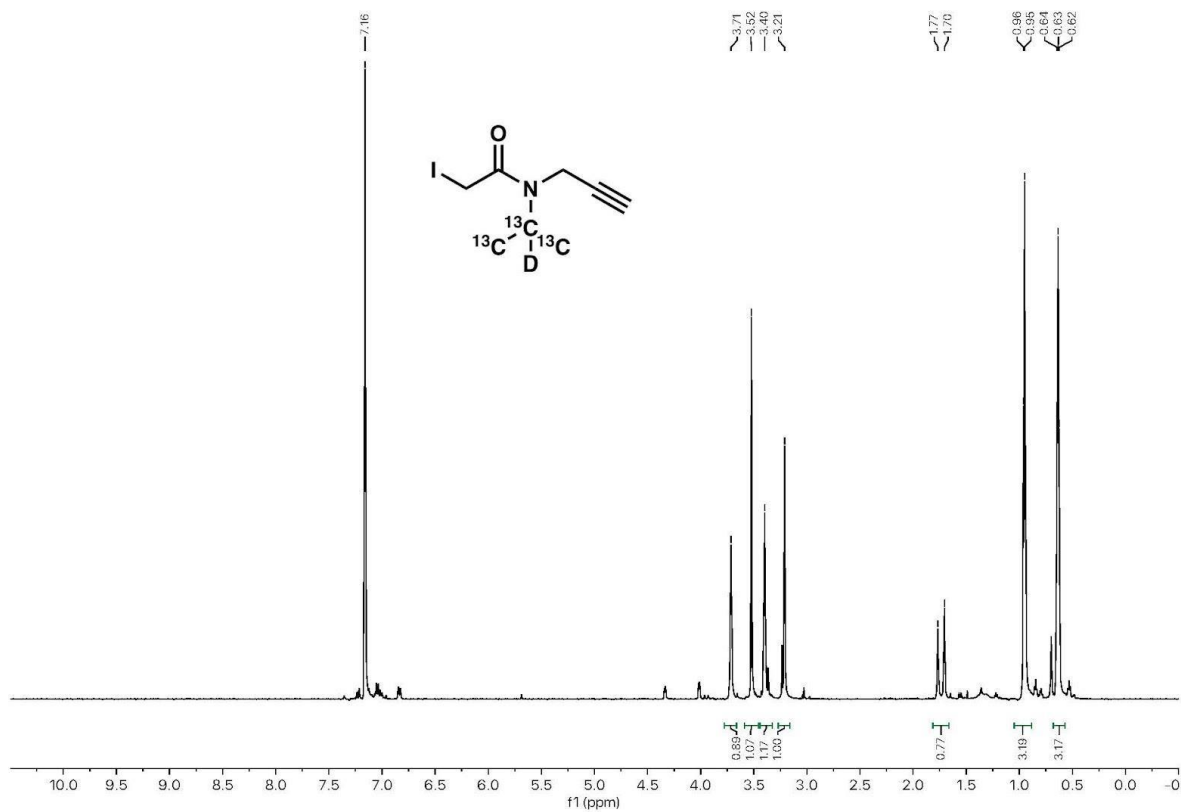
1H-NMR (4)



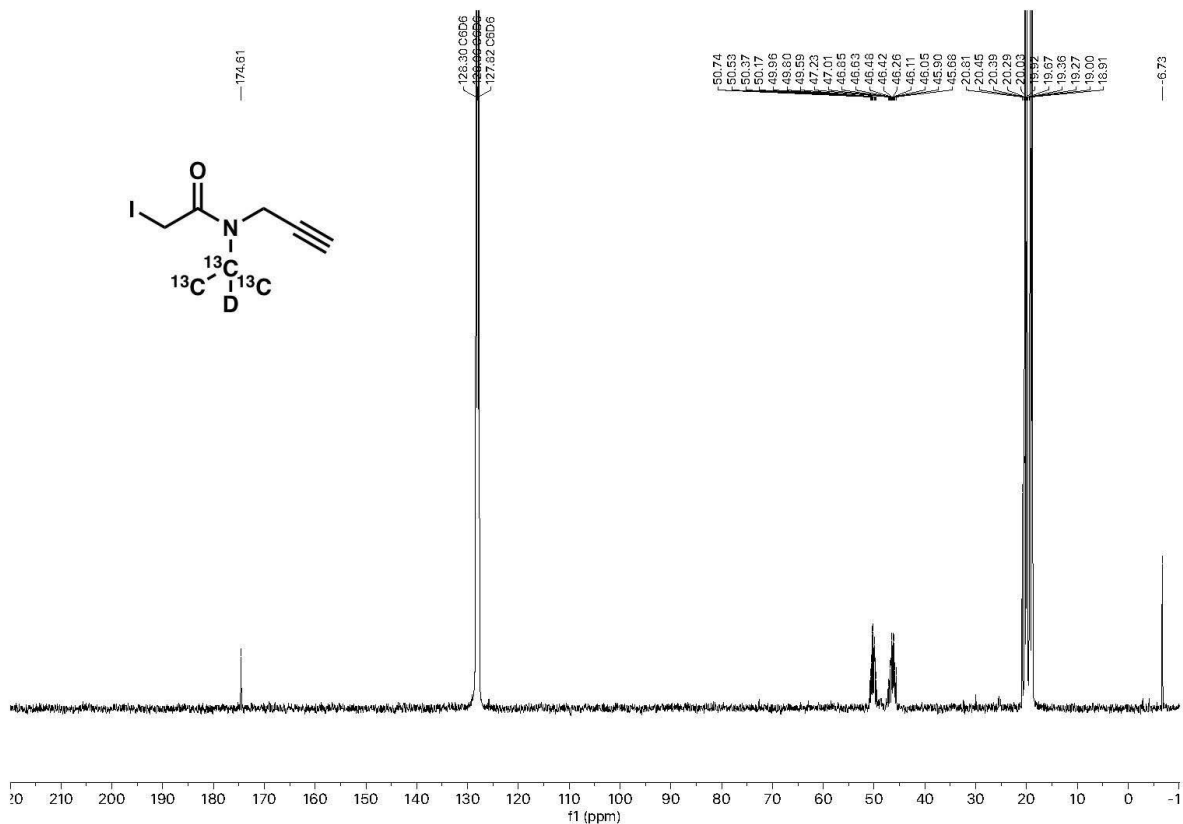
13C-NMR (4)



1H-NMR (5)



13C-NMR (5)



MSFragger search parameters

FragPipe (v16.0)

Config Workflow Umbridge Database MSFragger Validation Quant (MS1) Quant (Isobaric) PTMs Spec Lib Run

Run MSFragger

Save Config Load Custom MSFragger parameter file from disk

Common Options (Advanced Options are at the end of the page)

Peak Matching

Precursor mass tolerance PPM -20 20 Fragment mass tolerance PPM 20

Calibration and Optimization Mass calibration, parameter optimization Isotope error 0/1/2/3 Data type DDA

Protein Digestion

Load rules stricttrypsin Enzyme name stricttrypsin Cut after KR But not before

Cleavage ENZYMATC Missed cleavages 2 Clip N-term M

Peptide length 7 50 Peptide mass range 500 5,000 Split database 1

Modifications

Variable Modifications

Modifications

Variable modifications

Max variable mods on a peptide 3 Max combinations 5,000 Use all mods in first search

Enabled	Site (editable)	Mass Delta (e...)	Max occurren...
<input type="checkbox"/>	C	406.2151	3
<input type="checkbox"/>	K	8.014199	2
<input type="checkbox"/>	R	10.008269	2
<input checked="" type="checkbox"/>	C	463.2366	3
<input checked="" type="checkbox"/>	C	467.2529	3
<input checked="" type="checkbox"/>	C	57.02146	3
<input type="checkbox"/>	C	0	3
<input type="checkbox"/>	site 16	0	1

Fixed modifications

Enabled	Site	Mass Delta (editable)
<input checked="" type="checkbox"/>	N-Term Protein	0
<input checked="" type="checkbox"/>	G (glycine)	0
<input checked="" type="checkbox"/>	A (alanine)	0
<input checked="" type="checkbox"/>	S (serine)	0
<input checked="" type="checkbox"/>	P (proline)	0
<input checked="" type="checkbox"/>	V (valine)	0
<input checked="" type="checkbox"/>	T (threonine)	0
<input checked="" type="checkbox"/>	C (cysteine)	0
<input checked="" type="checkbox"/>	L (leucine)	0
<input checked="" type="checkbox"/>	I (isoleucine)	0
<input checked="" type="checkbox"/>	M (methionine)	0

IonQuant parameters

MS1 Quantification

Run MS1 quant [Load Quant defaults](#)

IonQuant Match between runs (MBR) Protein quant **Top-N** Min ions **1** Normalize

Feature detection

m/z tolerance (ppm) **10** RT tolerance (minutes) **0.4** IM tolerance (1/k0) **0.05**

Match between runs (MBR)

MBR RT tolerance (minutes) **1** MBR IM tolerance (1/k0) **0.05**
MBR min correlation **0** MBR top runs **10**
MBR ion FDR **0.01** MBR peptide FDR **1** MBR protein FDR **1**

Top-N options

Top N ions **3** Min freq **0.5** Min expts **1**

Labeling-based quant

Labels: Light **C463.2366** Medium Heavy **C467.2529** Re-quantify

Advanced options

Excluded mods
Min scans **1** Min isotopes **1** Keep index on disk

FreeQuant (alternative tool) RT Window (minutes) **0.4** m/z Window (ppm) **10**

PTM-prophet parameters

Crystal-C

Run Crystal-C Crystal-C performs additional search results cleanup. Recommended for Open Searches only.

PSM Validation

Run PSM Validation

Run PeptideProphet Defaults for: **Closed Search** [Load](#) Single **combined** pepxml file per experiment / group

Cmd line opts: `--decoyprobs --ppm --accmass --nonparam --expectscore`

Run Percolator keep TSV files

Cmd line opts: `--only-psms --no-terminate --post-processing-tdc`

PTM Site Localization

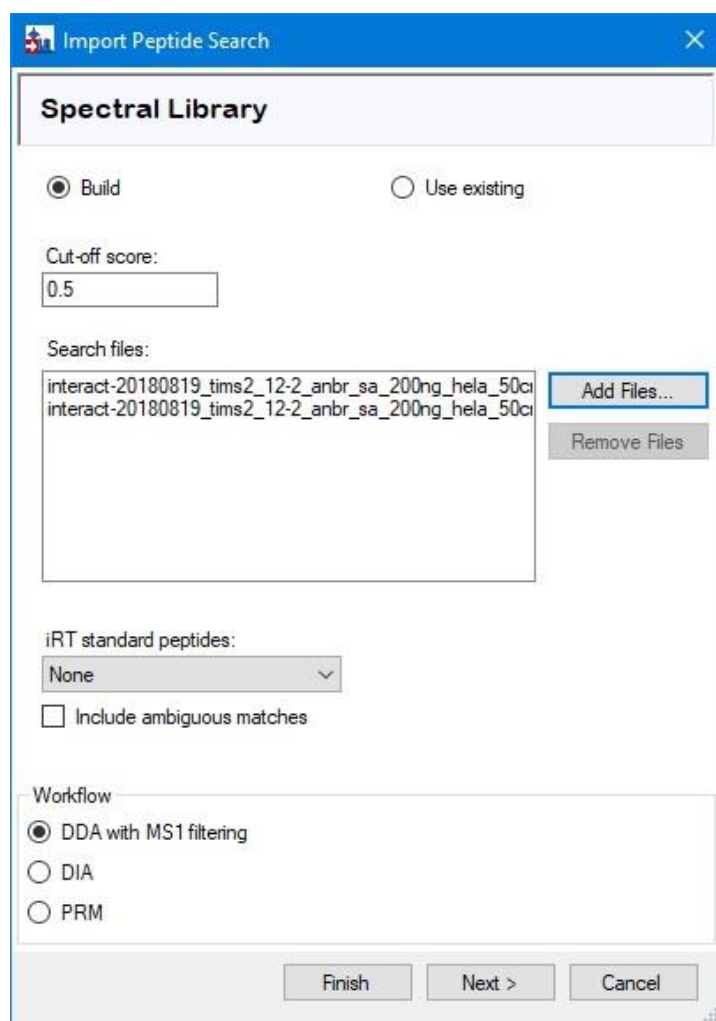
Run PTMProphet [Load defaults](#) Not for open searches. Mods format example: STY:79.966331,M:15.9949

Cmd line opts `--keepold --static --em 1 --nions b --mods C:463.2366,C:467.2529,C:57.02416,M:15.9949 --minprob 0.5`

Skyline-based Quantification of MSFragger searches

1. Search with mzML or raw files using FragPipe
Write calibrated mgf and set deisotoping to 0 or use uncalibrated.mgf output.
2. For non FAIMS data, use either (A) mzML files, interact.pepxml files and protein.fas files or (B) raw files, mgf files, interact.pepxml files, and protein.fas files. FAIMS data are processed with B only.
3. Make sure all above files are in the same root folder.
4. Download the latest version of Skyline:
https://skyline.ms/wiki/home/software/Skyline/page.view?name=SkylineInstall_64_21-1
5. Download the latest version of AddLabelType:
<https://proteome.gs.washington.edu/~nicksh/kbackus/AddLabelType/setup.exe>
6. Open Skyline and select '**Import DDA peptide search**'. Save the resulting file in the designated folder.
7. Once the new document is saved, go to '**Add Files**' and navigate to the 'interactpep.xml' file for the search results you want to import.
8. Specify the peptide probability threshold that corresponds to 1% ion FDR as the '**Cut-off score**'. (This can be found in the FragPipe log or .log file, a cut-off of 0.9024 would be used in this example: *INFO[15:56:25] Converged to 1.00 % FDR with 57143 ions decoy=576 threshold=0.9024 total=57719*)

9. Once a cut-off score has been set, press **'Next'**, prompting Skyline to build the spectral library from these peptide search results.



The screenshot shows the 'Import Peptide Search' dialog box with the 'Spectral Library' section active. The 'Build' radio button is selected, and the 'Cut-off score' is set to 0.5. The 'Search files' list contains two entries: 'interact-20180819_tims2_12-2_anbr_sa_200ng_hela_50ci'. The 'iRT standard peptides' dropdown is set to 'None', and the 'Include ambiguous matches' checkbox is unchecked. The 'Workflow' section has 'DDA with MS1 filtering' selected. The 'Next >' button is highlighted.

Import Peptide Search

Spectral Library

Build Use existing

Cut-off score:
0.5

Search files:
interact-20180819_tims2_12-2_anbr_sa_200ng_hela_50ci
interact-20180819_tims2_12-2_anbr_sa_200ng_hela_50ci

Add Files...
Remove Files

iRT standard peptides:
None

Include ambiguous matches

Workflow
 DDA with MS1 filtering
 DIA
 PRM

Finish Next > Cancel

10. Check all the modifications found by the program and add the unknown modifications:

Edit Structural Modification ×

Name:

Amino acid: Terminus: Variable

Crosslinker

Chemical fomula:

Monoisotopic mass: Average mass:

Add heavy as 4.0163 (mass difference between heavy and light).

Dialog box titled "Edit Isotope Modification" with a close button (X) in the top right corner.

Name:

Amino acid: Terminus:

Chemical formula

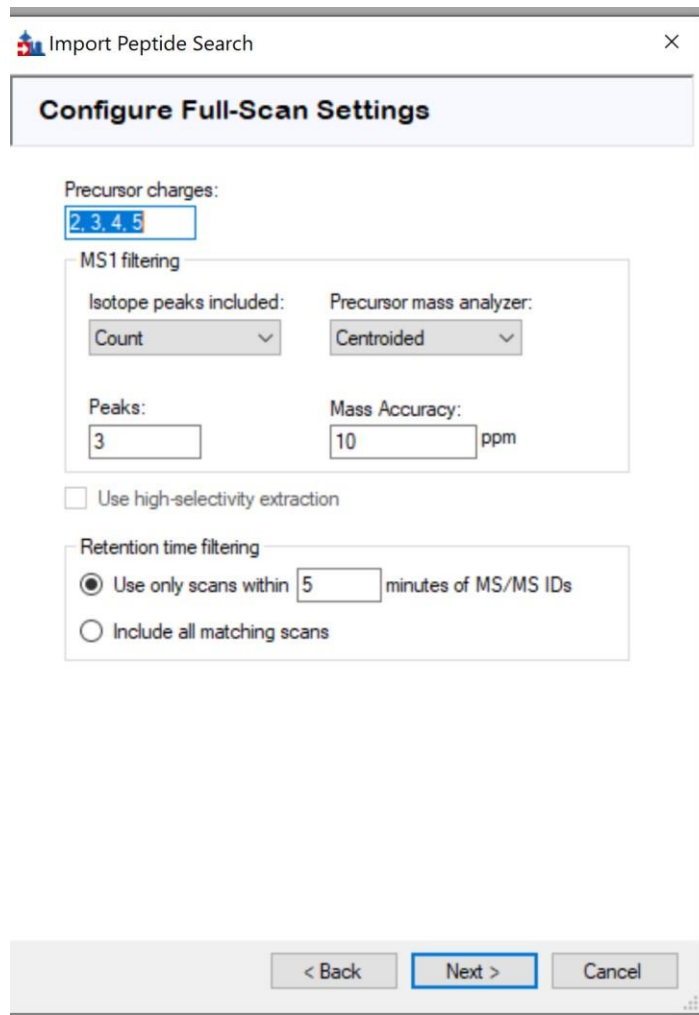
Chemical formula:

Monoisotopic mass: Average mass:

Relative retention time:

Buttons: OK, Cancel

11. Set precursor to 2,3,4,5 and 0.4 retention time filtering (as used by IonQuant)



12. Navigate to the protein.fas file for the search and include in Optional 'Add Fasta' tab with 2 miscleavages and continue with 'Keep All'.
13. Under Settings>Peptide Settings in the Modifications tab set the IA modification to 'variable' by selecting the modification and clicking 'edit'.

Edit Structural Modification

Name: Carbamidomethyl (C)

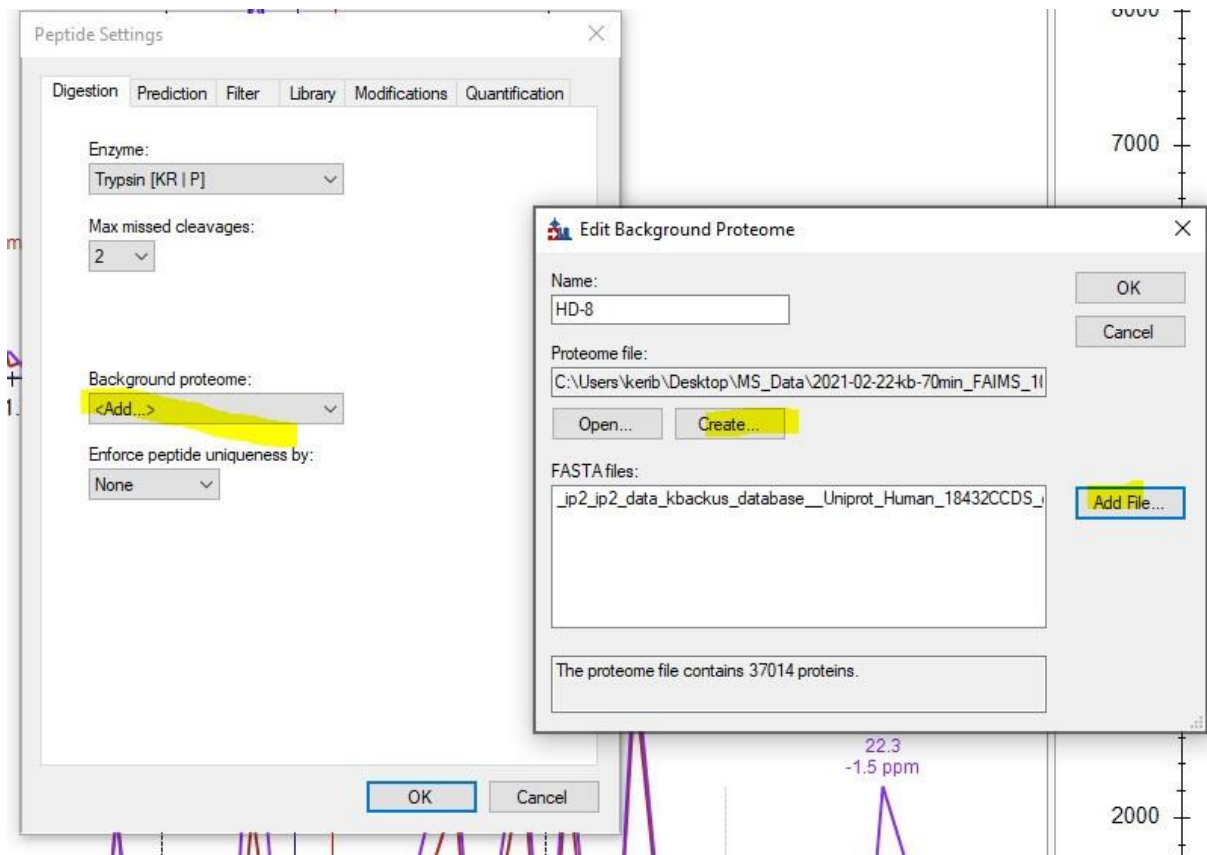
Amino acid: C Terminus: Variable Crosslinker

Chemical formula: H3C2NO

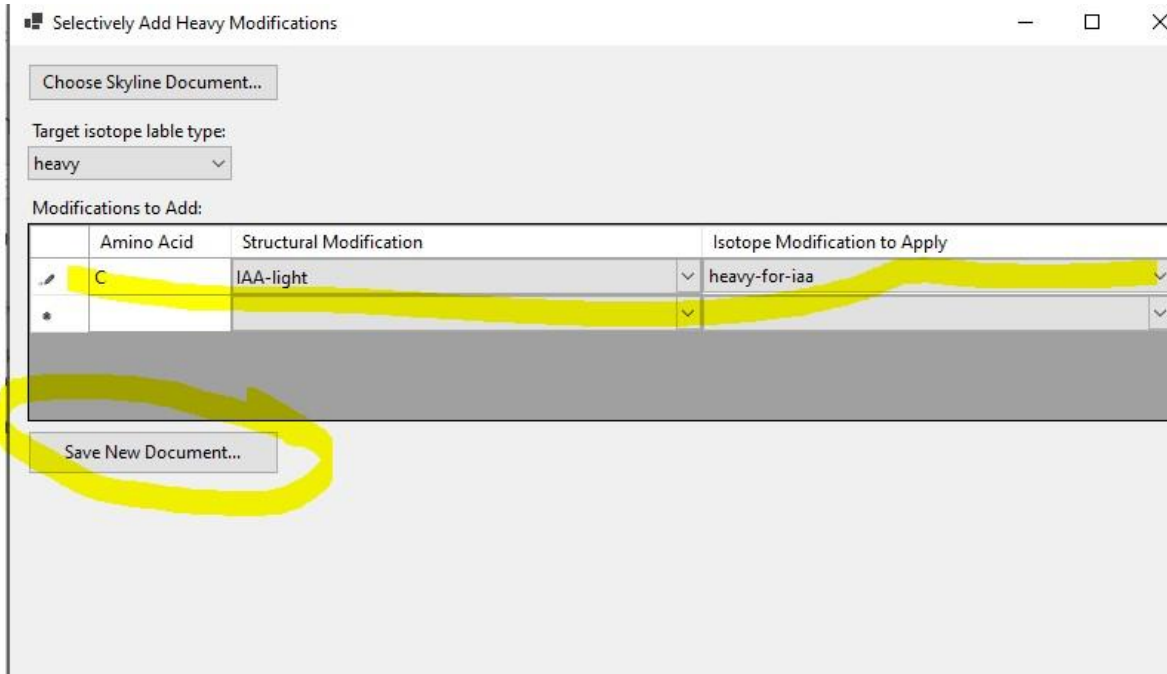
Monoisotopic mass: 57.021464 Average mass: 57.05162

OK Cancel Loss >>

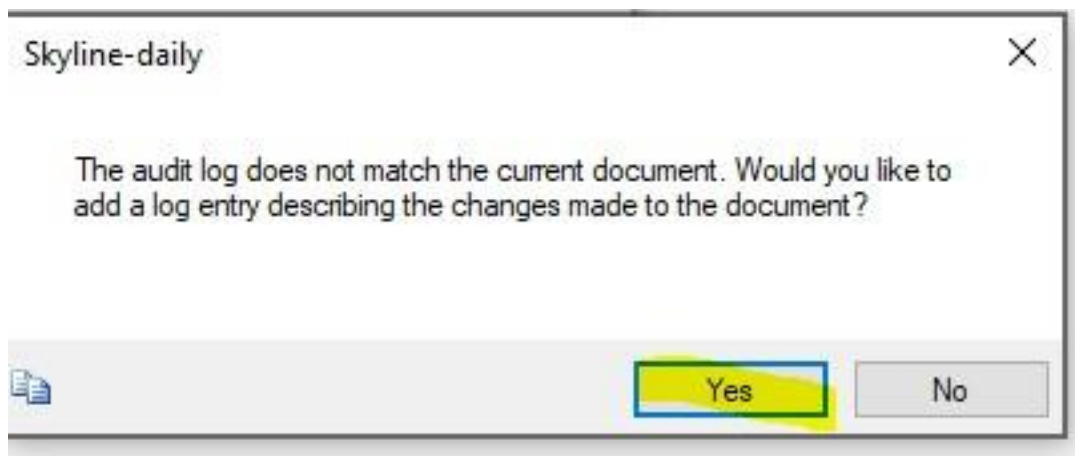
14. Set the internal standard to 'none' in the Peptide Settings>Modifications tab at the bottom instead of 'heavy'
15. Under Digestion tab settings, select 'add background proteome' and navigate to the reference database without reverse sequences. Set the max missed cleavages to 2



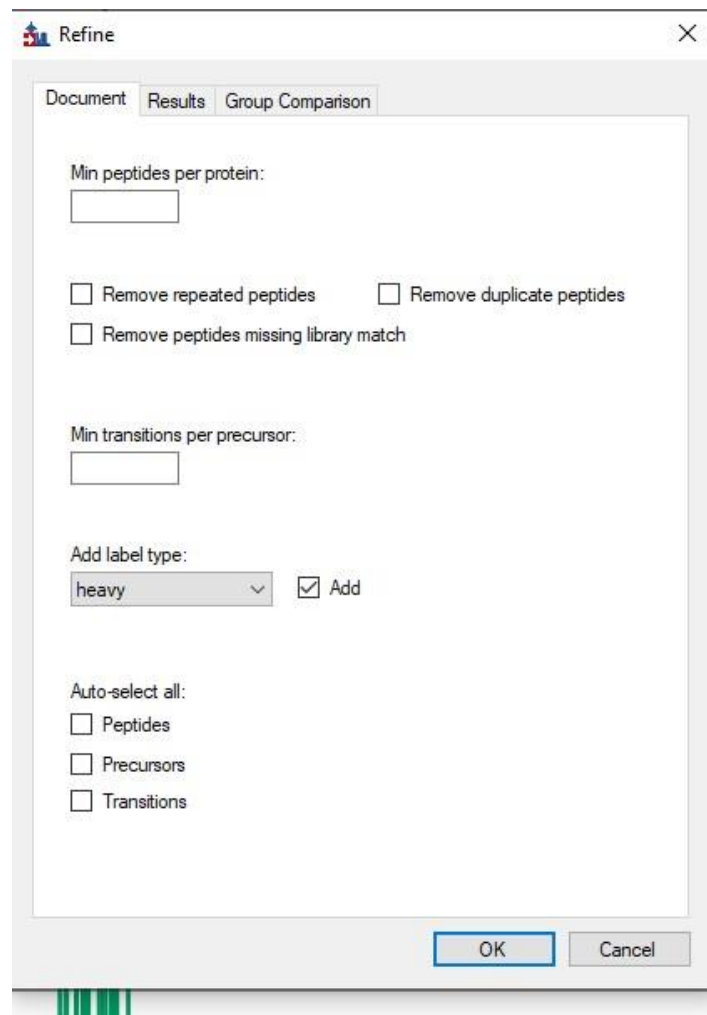
16. In Settings>Transition settings dropdown, in the first tab, set ion charges to 1,2,3.
17. In the 'View' dropdown, select Spectral Libraries and select associate all proteins, OK, 'add all' and wait for proteins to load.
18. Save the Skyline document and exit.
19. Open 'Add Label Type Program'.



20. Save the document, overwriting the previous version.



21. Reopen the file and select Refine > Advanced and click the "Add" checkbox next to "Remove Label Type" and tell Skyline to add the Heavy label type.



22. **Reimport your file.**

Press Ctrl+R.

Reimport and select the file.

Click 'OK'

23. **If non-FAIMS data, STOP HERE: Under 'File' dropdown, select Export> Report.**

Create a report and select desired columns: **Protein Name, Protein Gene, Peptide Sequence, Ratio LightToHeavy, Quantification, Peptide Modified Sequence**. You may have to go several layers in the tabs

The screenshots illustrate the process of selecting columns for a report in a software interface. The columns are organized in a tree structure under 'Proteins' and 'Peptides'.

Top Left Screenshot: Shows the 'Proteins' tree. The 'Protein Name' column is selected (checked) and highlighted in blue. Other columns include Protein Description, Protein Accession, Protein Preferred Name, Protein Gene, Protein Species, Protein Sequence, Auto Select Peptides, Protein Sequence Coverage, Protein Note, Protein Locator, and Replicates.

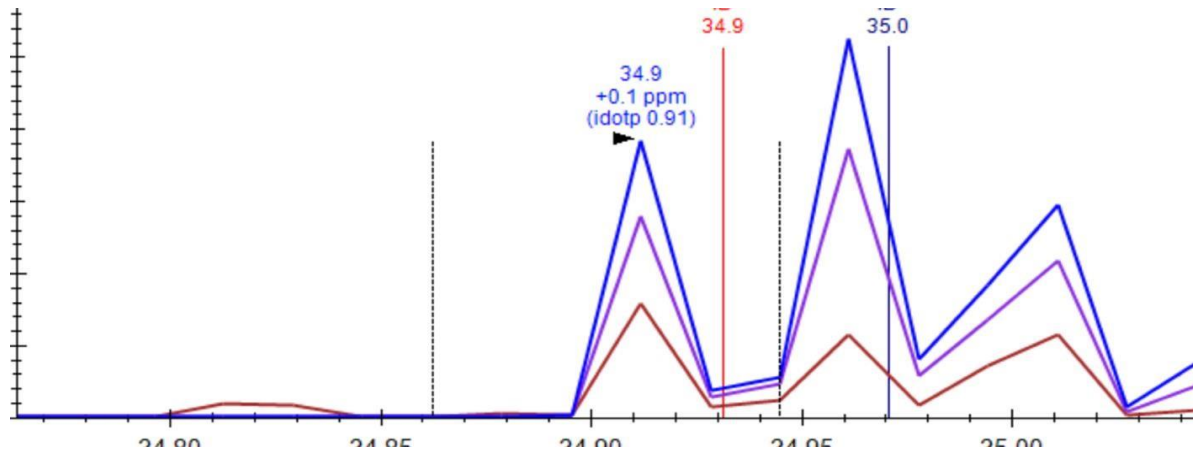
Top Right Screenshot: Shows the 'Peptides' tree. The 'Quantification' column is selected (checked) and highlighted in blue. Other columns include Peptide Peak Found Ratio, Peptide Retention Time, Predicted Result Retention Time, Ratio To Standard, Best Replicate, Modified Area Proportion, Attribute Area Proportion, Exclude From Calibration, Replicate Calibration Curve, Batch Figures Of Merit, Explicit Analyte Concentration, Peptide Result Locator, RatioLightToHeavy, DotProductLightToHeavy, Peptide Sequence, Peptide Sequence Length, and Peptide Modified Sequence.

Bottom Left Screenshot: Shows the 'Peptides' tree. The 'Peptide Sequence' column is selected (checked) and highlighted in blue. Other columns include Peptide Peak Found Ratio, Peptide Retention Time, Predicted Result Retention Time, Ratio To Standard, Best Replicate, Modified Area Proportion, Attribute Area Proportion, Exclude From Calibration, Quantification, Replicate Calibration Curve, Batch Figures Of Merit, Explicit Analyte Concentration, Peptide Result Locator, RatioLightToHeavy, DotProductLightToHeavy, Peptide Sequence Length, and Peptide Modified Sequence.

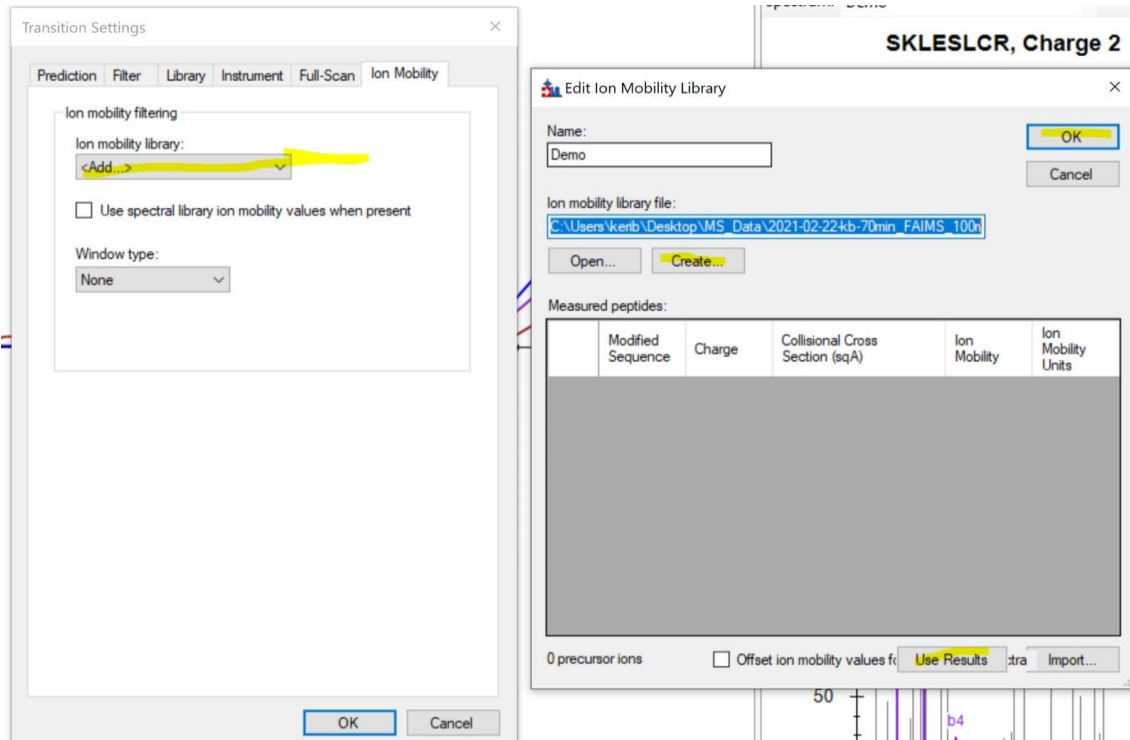
Bottom Right Screenshot: Shows the 'Peptides' tree. The 'RatioLightToHeavy' column is selected (checked) and highlighted in blue. Other columns include Peptide Peak Found Ratio, Peptide Retention Time, Predicted Result Retention Time, Ratio To Standard, Best Replicate, Modified Area Proportion, Attribute Area Proportion, Exclude From Calibration, Quantification, Replicate Calibration Curve, Batch Figures Of Merit, Explicit Analyte Concentration, Peptide Result Locator, RatioLightToHeavy, DotProductLightToHeavy, Peptide Sequence, and Peptide Modified Sequence.

24. **If FAIMS data:** Turn on Ion mobility and do the following:

Current settings display peaks as shown below:



25. Navigate to Settings > Transition Settings > Prediction. Hit 'Add' and 'OK'. *This will only display the peak for a peptide from the 'best CV'*



26. Peptides should be listed as shown.

Edit Ion Mobility Library

Name: OK Cancel

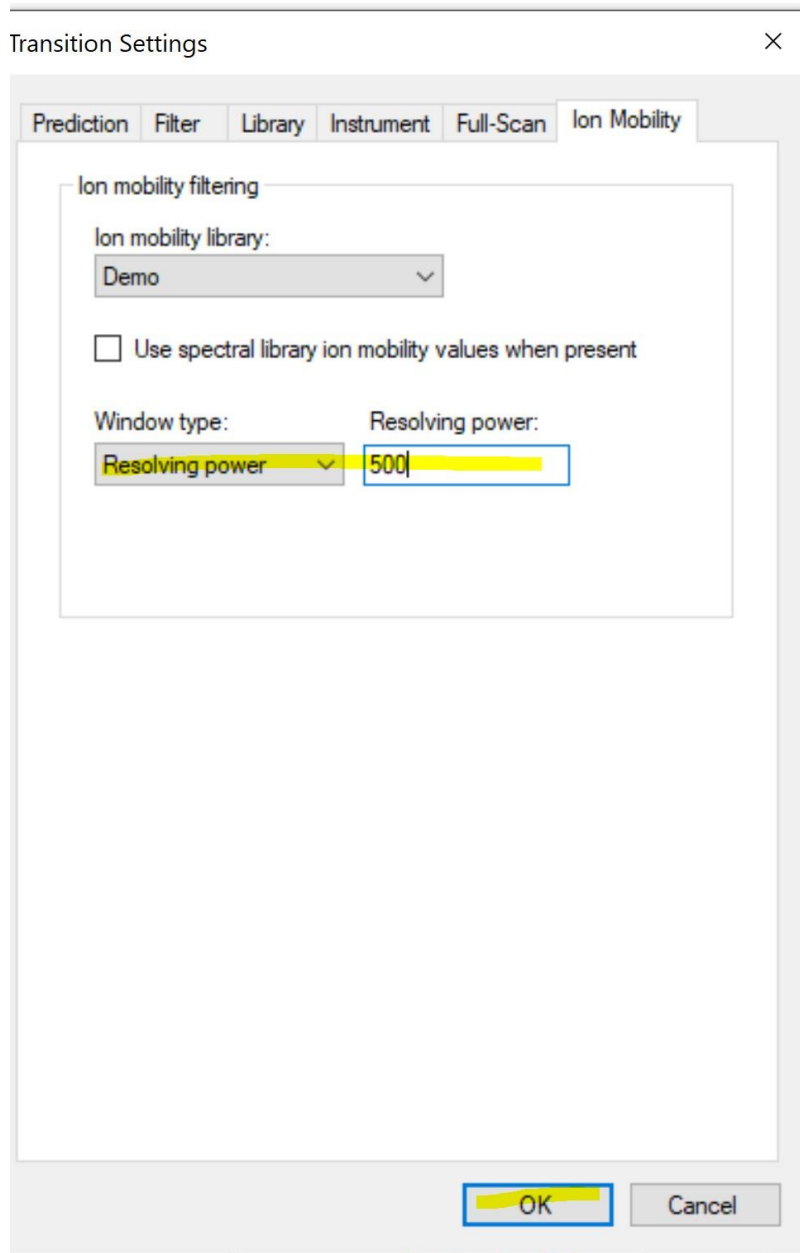
Ion mobility library file: Open... Create...

Measured peptides:

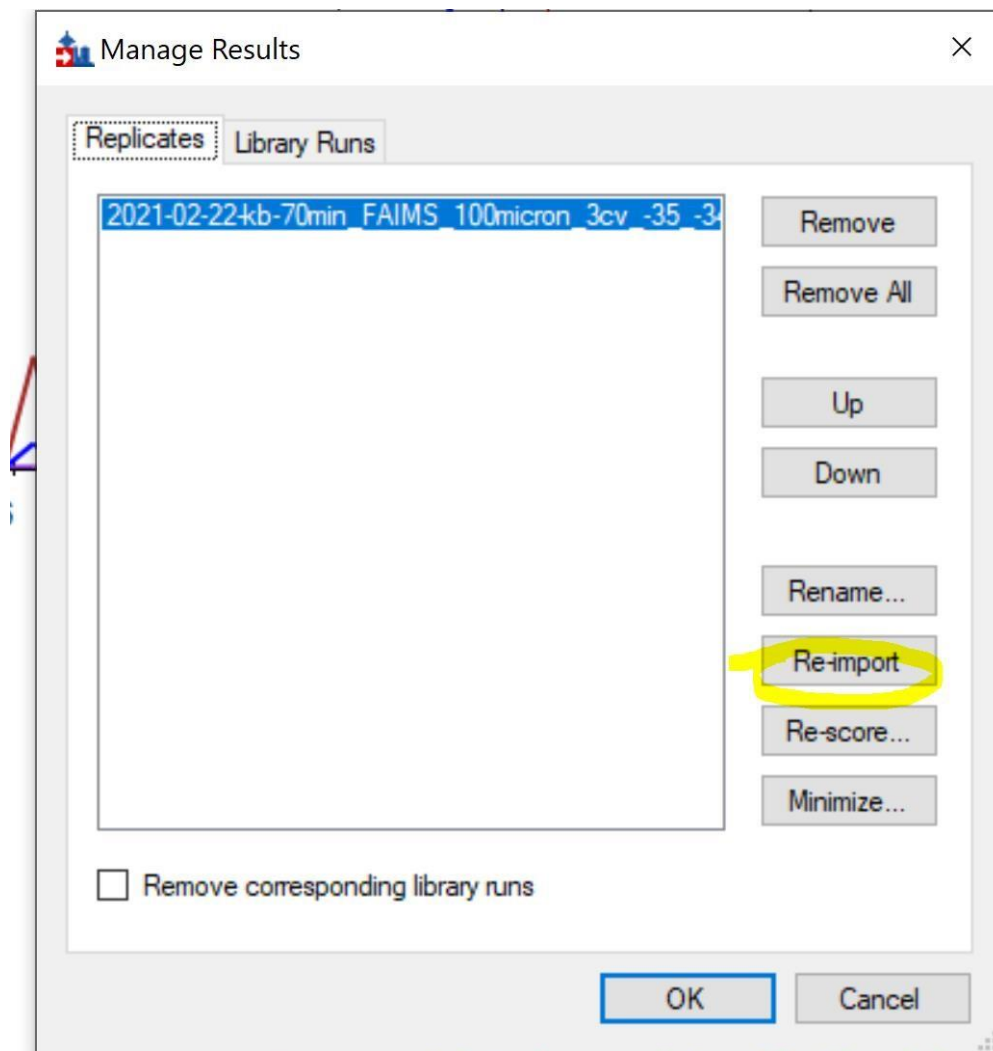
	Modified Sequence	Charge	Collisional Cross Section (sqÅ)	Ion Mobility	Ion Mobility Units
▶	EAHHEATC[+463.2366]R	2		-45.00000	Compensation Voltage (V) ▼
	EAHHEATC[+467.25294]R	2		-45.00000	Compensation Voltage (V) ▼
	SIGSLPC[+463.2366]LEHVS...	3		-35.00000	Compensation Voltage (V) ▼
	SIGSLPC[+467.25294]LEHVS...	3		-35.00000	Compensation Voltage (V) ▼
	QWEALC[+463.2366]GR	2		-45.00000	Compensation Voltage (V) ▼
	QWEALC[+467.25294]GR	2		-55.00000	Compensation Voltage (V) ▼
	FC[+463.2366]EDWMQAFLN...	2		-35.00000	Compensation Voltage (V) ▼
	FC[+467.25294]EDWMQAFLN...	2		-35.00000	Compensation Voltage (V) ▼
	EC[+463.2366]GLWLSLGGF...	3		-45.00000	Compensation Voltage (V) ▼
	EC[+467.25294]GLWLSLGGF...	3		-45.00000	Compensation Voltage (V) ▼

13972 Precursor Ions Offset ion mobility values for high energy spectra Use Results Import...

27. Set resolving power to 500.



28. Reimport the data again (Ctrl+R)



29. Export data as described above.

References

1. Abo, M., and Weerapana, E. (2019) Chemical probes for redox signaling and oxidative stress. *Antioxid. Redox Signal.* 30, 1369–1386.
2. Backus, K. M. (2019) Applications of reactive cysteine profiling. In: Cravatt, B., Hsu, K. L., Weerapana, E., eds., Vol. 420., *Activity-Based Protein Profiling. Current Topics in Microbiology and Immunology*, Springer, Cham: 375–417.
3. Fra, A., Yoboue, E. D., and Sitia, R. (2017) Cysteines as redox molecular switches and targets of disease. *Front. Mol. Neurosci.* 10, 167.
4. Cremers, C. M., and Jakob, U. (2013) Oxidant sensing by reversible disulfide bond formation. *J. Biol. Chem.* 288, 26489–26496.
5. D'Autr'eaux, B., and Toledano, M. B. (2007) ROS as signaling molecules: Mechanisms that generate specificity in ROS homeostasis. *Nat. Rev. Mol. Cell Biol.* 8, 813–824.
6. Sun, L., Wang, X., Saredy, J., Yuan, Z., Yang, X., and Wang, H. (2020) Innate-adaptive immunity interplay and redox regulation in immune response. *Redox Biol.* 37, 101759.
7. Paulsen, C. E., and Carroll, K. S. (2013) Cysteine-mediated redox signaling: Chemistry, biology, and tools for discovery. *Chem. Rev.* 113, 4633–4679.
8. Chouchani, E. T., James, A. M., Fearnley, I. M., Lilley, K. S., and Murphy, M. P. (2011) Proteomic approaches to the characterization of protein thiol modification. *Curr. Opin. Chem. Biol.* 15, 120–128.
9. Abo, M., Li, C., and Weerapana, E. (2018) Isotopically-labeled iodoacetamide-alkyne probes for quantitative cysteine-reactivity profiling. *Mol. Pharm.* 15, 743–749.

10. Fu, L., Li, Z., Liu, K., Tian, C., He, J., He, J., He, F., Xu, P., and Yang, J. (2020) A quantitative thiol reactivity profiling platform to analyze redox and electrophile reactive cysteine proteomes. *Nat. Protoc.* 15, 2891–2919
11. García-Santamarina, S., Boronat, S., Domènech, A., Ayté, J., Molina, H., and Hidalgo, E. (2014) Monitoring in vivo reversible cysteine oxidation in proteins using ICAT and mass spectrometry. *Nat. Protoc.* 9, 1131–1145
12. Leichert, L.I., Gehrke, F., Gudiseva, H.V., Blackwell, T., Ilbert, M., Walker, A.K., Strahler, J.R., Andrews, P.C., and Jakob, U. (2008) Quantifying changes in the thiol redox proteome upon oxidative stress in vivo. *Proc. Natl. Acad. Sci. U.S.A.* 105, 8197–8202
13. Zhou, Y., Wynia-Smith, S.L., Couvertier, S.M., Kalous, K.S., Marletta, M.A., Smith, B.C., and Weerapana, E. (2016) Chemoproteomic strategy to quantitatively monitor transnitrosation uncovers functionally relevant S nitrosation sites on Cathepsin D and HADH2. *Cell Chem. Biol.* 23, 727–737
14. Guo, J., Gaffrey, M. J., Su, D., Liu, T., Camp, D. G., Smith, R. D., and Qian, W. J. (2013) Resin-assisted enrichment of thiols as a general strategy for proteomic profiling of cysteine-based reversible modifications. *Nat. Protoc.* 9(1), 64–75.
15. Xiao, H., Jedrychowski, M. P., Schweppe, D. K., Huttlin, E. L., Yu, Q., Heppner, D. E., Li, J., Long, J., Mills, E. L., Szpyt, J., He, Z., Du, G., Garrity, R., Reddy, A., Vaites, L. P., et al. (2020) A quantitative tissue-specific landscape of protein redox regulation during aging. *Cell* 180, 968–983.e24.
16. Jaffrey, S. R., and Snyder, S. H. (2001) The biotin switch method for the detection of S-nitrosylated proteins. *Sci. Signal.* 2001. <https://doi.org/10.1126/stke.2001.86.pl1>.

17. Lind, C., Gerdes, R., Hamnell, Y., Schuppe-Koistinen, I., Von Löwenhielm, H. B., Holmgren, A., and Cotgreave, I. A. (2002) Identification of S-glutathionylated cellular proteins during oxidative stress and constitutive metabolism by affinity purification and proteomic analysis. *Arch. Biochem. Biophys.* 406, 229–240.
18. Saurin, A. T., Neubert, H., Brennan, J. P., and Eaton, P. (2004) Widespread sulfenic acid formation in tissues in response to hydrogen peroxide. *Proc. Natl. Acad. Sci. U.S.A.* 101, 17982–17987.
19. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* 17, 994–999.
20. Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., and Mann, M. (2002) Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell Proteomics* 1, 376–386.
21. Vajrychova, M., Salovska, B., Pimkova, K., Fabrik, I., Tambor, V., Kondelova, A., Bartek, J., and Hodny, Z. (2019) Quantification of cellular protein and redox imbalance using SILAC-iodoTMT methodology. *Redox Biol.* 24, 101227.
22. Weerapana, E., Wang, C., Simon, G. M., Richter, F., Khare, S., Dillon, M. B. D., Bachovchin, D. A., Mowen, K., Baker, D., and Cravatt, B. F. (2010) Quantitative reactivity profiling predicts functional cysteines in proteomes. *Nature* 468, 790–797.
23. Shakir, S., Vinh, J., and Chiappetta, G. (2017) Quantitative analysis of the cysteine redoxome by iodoacetyl tandem mass tags. *Anal. Bioanal. Chem.* 409, 3821–3830.

24. Park, S. K., Venable, J. D., Xu, T., and Yates, J. R. (2008) A quantitative analysis software tool for mass spectrometry-based proteomics. *Nat. Methods* 5, 319–322.
25. Bae, J. W., Kim, S., Kim, V. N., and Kim, J. S. (2021) Photoactivatable ribonucleosides mark base-specific RNA-binding sites. *Nat. Commun.* 12, 1–10.
26. [preprint] Demichev, V., Yu, F., Teo, G. C., Szyrwił, L., Rosenberger, G. A., Decker, J., Kaspar-Schoenefeld, S., Lilley, K. S., Mülleder, M., Nesvizhskii, A. I., and Ralser, M. (2021) High sensitivity dia-PASEF proteomics with DIA-NN and FragPipe. *bioRxiv*. <https://doi.org/10.1101/2021.03.08.434385>.
27. Fisher, J., Mohanty, T., Karlsson, C. A. Q., Khademi, S. M. H., Malmström, E., Frigyesi, A., Nordenfelt, P., Malmstrom, J., and Linder, A. (2021) Proteome profiling of recombinant DNase therapy in reducing NETs and aiding recovery in COVID-19 patients. *Mol. Cell. Proteomics* 20, 100113.
28. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., and Nesvizhskii, A. I. (2017) MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* 14, 513–520.
29. Yu, F., Haynes, S. E., and Nesvizhskii, A. I. (2021) IonQuant enables accurate and sensitive label-free quantification with FDR-controlled match between runs. *Mol. Cell. Proteomics* 20, 100077.
30. Yu, F., Haynes, S. E., Teo, G. C., Avtonomov, D. M., Polasky, D. A., and Nesvizhskii, A. I. (2020) Fast quantitative analysis of timsTOF PASEF data with MSFragger and IonQuant. *Mol. Cell. Proteomics* 19, 1575–1585.

31. Yu, F., Teo, G. C., Kong, A. T., Haynes, S. E., Avtonomov, D. M., Geiszler, D. J., and Nesvizhskii, A. I. (2020) Identification of modified peptides using localization-aware open search. *Nat. Commun.* 11, 1–9.
32. Teo, G. C., Polasky, D. A., Yu, F., and Nesvizhskii, A. I. (2020) Fast deisotoping algorithm and its implementation in the MSFragger search engine. *J. Proteome Res.* 20, 498–505.
33. Van Gelder, K., Virta, L. K. A., Easlick, J., Prudhomme, N., McAlister, J. A., Geddes-McAlister, J., and Akhtar, T. A. (2021) A central role for polyprenol reductase in plant dolichol biosynthesis. *Plant Sci.* 303, 110773.
34. da Veiga Leprevost, F., Haynes, S. E., Avtonomov, D. M., Chang, H. Y., Shanmugam, A. K., Mellacheruvu, D., Kong, A. T., and Nesvizhskii, A. I. (2020) Philosopher: A versatile toolkit for shotgun proteomics data analysis. *Nat. Methods* 17, 869–870.
35. Day, N. J., Gaffrey, M. J., and Qian, W. J. (2021) Stoichiometric thiol redox proteomics for quantifying cellular responses to perturbations. *Antioxidants* 10, 499.
36. PT, C., KE, P., and RT, R. (1997) The CXXC motif: A rheostat in the active site. *Biochemistry* 36, 4061–4066.
37. Quan, S., Schneider, I., Pan, J., Von Hacht, A., and Bardwell, J. C. A. (2007) The CXXC motif is more than a redox rheostat. *J. Biol. Chem.* 282, 28823–28833.
38. Hughes, C. S., Moggridge, S., Müller, T., Sorensen, P. H., Morin, G. B., and Krijgsveld, J. (2019) Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat. Protoc.* 14, 68–85.

39. Müller, T., Kalxdorf, M., Longuespée, R., Kazdal, D. N., Stenzinger, A., and Krijgsveld, J. (2020) Automated sample preparation with SP3 for low-input clinical proteomics. *Mol. Syst. Biol.* 16, e9111.
40. Yan, T., Desai, H. S., Boatner, L. M., Yen, S. L., Cao, J., Palafox, M. F., Jami-Alahmadi, Y., and Backus, K. M. (2021) SP3-FAIMS chemoproteomics for high-coverage profiling of the human Cysteineome*. *Chembiochem* 22, 1841–1851.
41. Nesvizhskii, A. I. (2010) A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* 73, 2092–2123.
42. MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., Kern, R., Tabb, D. L., Liebler, D. C., and MacCoss, M. J. (2010) Skyline: An open-source document editor for creating and analyzing targeted proteomic experiments. *Bioinformatics* 26, 966–968.
43. Backus, K. M., Correia, B. E., Lum, K. M., Forli, S., Horning, B. D., González-Páez, G. E., Chatterjee, S., Lanning, B. R., Teijaro, J. R., Olson, A. J., Wolan, D. W., and Cravatt, B. F. (2016) Proteome-wide covalent ligand discovery in native biological systems. *Nature* 534, 570–574.
44. Swearingen, K. E., and Moritz, R. L. (2012) High-field asymmetric waveform ion mobility spectrometry for mass spectrometry-based proteomics. *Expert Rev. Proteomics* 9, 505–517.
45. Pfammatter, S., Bonneil, E., McManus, F. P., and Thibault, P. (2019) Accurate quantitative proteomic analyses using metabolic labeling and high field asymmetric waveform ion mobility spectrometry (FAIMS). *J. Proteome Res.* 18, 2129–2138.

46. Hansen, R. E., Roth, D., and Winther, J. R. (2009) Quantifying the global cellular thiol–disulfide status. *Proc. Natl. Acad. Sci. U.S.A.* 106, 422–427.
47. Brandes, N., Reichmann, D., Tienson, H., Leichert, L. I., and Jakob, U. (2011) Using quantitative redox proteomics to dissect the yeast redoxome. *J. Biol. Chem.* 286, 41893–41903.
48. Kim, M. S., Zhong, J., and Pandey, A. (2016) Common errors in mass spectrometry-based analysis of post-translational modifications. *Proteomics* 16, 700.
49. Shteynberg, D. D., Deutsch, E. W., Campbell, D. S., Hoopmann, M. R., Kusebauch, U., Lee, D., Mendoza, L., Midha, M. K., Sun, Z., Whetton, A. D., and Moritz, R. L. (2019) PTMProphet: Fast and accurate mass modification localization for the trans-proteomic pipeline. *J. Proteome Res.* 18, 4262–4272.
50. Samson, A. L., Knaupp, A. S., Kass, I., Kleifeld, O., Marijanovic, E. M., Hughes, V. A., Lupton, C. J., Buckle, A. M., Bottomley, S. P., and Medcalf, R. L. (2014) Oxidation of an exposed methionine instigates the aggregation of glyceraldehyde-3-phosphate dehydrogenase. *J. Biol. Chem.* 289, 26922–26936.
51. Hirano, T., Kishi, M., Sugimoto, H., Taguchi, R., Obinata, H., Ohshima, N., Tatei, K., and Izumi, T. (2009) Thioesterase activity and subcellular localization of acylprotein thioesterase 1/lysophospholipase 1. *Biochim. Biophys. Acta* 1791, 797–805.
52. Weichsel, Andrzej, Brailey, Jacqueline L., and Montfort, W. R. (2007) Buried S-nitrosocysteine revealed in crystal structures of human thioredoxin. *Biochemistry* 46, 1219–1227.
53. Arnesano, F., Banci, L., Bertini, I., Martinelli, M., Furukawa, Y., and O’Halloran, T. V. (2004) The unusually stable quaternary structure of human Cu, Zn-superoxide

dismutase 1 is controlled by both metal occupancy and disulfide status. *J. Biol. Chem.* 279, 47998–48003.

54. Richarme, G., Mihoub, M., Dairou, J., Chi Bui, L., Leger, T., and Lamouri, A. (2015) Parkinsonism-associated protein DJ-1/park7 is a major protein deglycase that repairs methylglyoxal- and glyoxal-glycated cysteine, arginine, and lysine residues. *J. Biol. Chem.* 290, 1885–1897.
55. Canet-Avil'es, R. M., Wilson, M. A., Miller, D. W., Ahmad, R., McLendon, C., Bandyopadhyay, S., Baptista, M. J., Ringe, D., Petsko, G. A., and Cookson, M. R. (2004) The Parkinson's disease protein DJ-1 is neuroprotective due to cysteine-sulfinic acid-driven mitochondrial localization. *Proc. Natl. Acad. Sci. U.S.A.* 101, 9103–9108.
56. Consortium, T. U., Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Da Silva, A., et al. (2021) UniProt: The universal protein knowledge base in 2021. *Nucleic Acids Res.* 49, D480–D489.
57. Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., Clark, N. R., and Ma'ayan, A. (2013) Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 14, 128.
58. Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., and Ma'ayan, A. (2016) Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* 44, W90–W97.

59. Vinogradova, E. V., Zhang, X., Remillard, D., Lazar, D. C., Suci, R. M., Wang, Y., Bianco, G., Yamashita, Y., Crowley, V. M., Schafroth, M. A., Yokoyama, M., Konrad, D. B., Lum, K. M., Simon, G. M., Kemper, E. K., et al. (2020) An activity-guided map of electrophile-cysteine interactions in primary human T cells. *Cell* 182, 1009–1026.e29.
60. Angelini, G., Gardella, S., Ardy, M., Ciriolo, M. R., Filomeni, G., Trapani, G. D., Clarke, F., Sitia, R., and Rubartelli, A. (2002) Antigen-presenting dendritic cells provide the reducing extracellular microenvironment required for T lymphocyte activation. *Proc. Natl. Acad. Sci. U.S.A.* 99, 1491–1496.
61. Belikov, A. V., Schraven, B., and Simeoni, L. (2015) T cells and reactive oxygen species. *J. Biomed. Sci.* 22, 1–11.
62. Chen, X., Song, M., Zhang, B., and Zhang, Y. (2016) Reactive oxygen species regulate T cell immune response in the tumor microenvironment. *Oxid. Med. Cell Longev.* 2016, 1580967.
63. Sena, L. A., Li, S., Jairaman, A., Prakriya, M., Ezponda, T., Hildeman, D. A., Wang, C. R., Schumacker, P. T., Licht, J. D., Perlman, H., Bryce, P. J., and Chandel, N. S. (2013) Mitochondria are required for antigen-specific T cell activation through reactive oxygen species signaling. *Immunity* 38, 225–236.
64. Hildeman, D. A., Mitchell, T., Kappler, J., and Marrack, P. (2003) T cell apoptosis and reactive oxygen species. *J. Clin. Invest.* 111, 575–581.
65. Kesarwani, P., Murali, A. K., Al-Khami, A. A., and Mehrotra, S. (2013) Redox regulation of T-cell function: From molecular mechanisms to significance in human health and disease. *Antioxid. Redox Signal.* 18, 1497–1534.

66. Michalek, R. D., Nelson, K. J., Holbrook, B. C., Yi, J. S., Stridiron, D., Daniel, L. W., Fetrow, J. S., King, S. B., Poole, L. B., and Grayson, J. M. (2007) The requirement of reversible cysteine sulfenic acid formation for T cell activation and function. *J. Immunol.* 179, 6456–6467.
67. Navarro, M. N., and Cantrell, D. A. (2014) Serine-threonine kinases in TCR signaling. *Nat. Immunol.* 15, 808–814.
68. Isakov, N., and Altman, A. (2003) Protein kinase C θ in T cell activation. *Annu. Rev. Immunol.* 20, 761–794.
69. Mohiuddin, I. S., and Kang, M. H. (2019) DNA-PK as an emerging therapeutic target in cancer. *Front. Oncol.* 9, 635.
70. Blewett, M. M., Xie, J., Zaro, B. W., Backus, K. M., Altman, A., Teijaro, J. R., and Cravatt, B. F. (2016) Chemical proteomic map of dimethyl fumarate-sensitive cysteines in primary human T cells. *Sci. Signal.* 9, rs10.
71. Jilani, A., Ramotar, D., Slack, C., Ong, C., Yang, X. M., Scherer, S. W., and Lasko, D. D. (1999) Molecular cloning of the human gene, PNKP, encoding a polynucleotide kinase 3'-phosphatase and evidence for its role in repair of DNA strand breaks caused by oxidative damage. *J. Biol. Chem.* 274, 24176–24186.
72. Bhoumik, A., Takahashi, S., Breitweiser, W., Shiloh, Y., Jones, N., and Ronai, Z. (2005) ATM-dependent phosphorylation of ATF2 is required for the DNA damage response. *Mol. Cell* 18, 577–587.
73. Bar-Peled, L., Chantranupong, L., Cherniack, A. D., Chen, W. W., Ottina, K. A., Grabiner, B. C., Spear, E. D., Carter, S. L., Meyerson, M., and Sabatini, D. M. (2013)

A tumor suppressor complex with GAP activity for the rag GTPases that signal amino acid sufficiency to mTORC1. *Science* 340, 1100–1106.

74. Cambiaghi, T. D., Pereira, C. M., Shanmugam, R., Bolech, M., Wek, R. C., Sattlegger, E., and Castilho, B. A. (2014) Evolutionarily conserved IMPACT impairs various stress responses that require GCN1 for activating the eIF2 kinase GCN2. *Biochem. Biophys. Res. Commun.* 443, 592–597.
75. Paik, J. C., Wang, B., Liu, K., Lue, J. K., and Lin, W. C. (2010) Regulation of E2F1-induced apoptosis by the nucleolar protein RRP1B. *J. Biol. Chem.* 285, 6348–6363.
76. Hayakawa, S., Shiratori, S., Yamato, H., Kameyama, T., Kitatsuji, C., Kashigi, F., Goto, S., Kameoka, S., Fujikura, D., Yamada, T., Mizutani, T., Kazumata, M., Sato, M., Tanaka, J., Asaka, M., et al. (2010) ZAPS is a potent stimulator of signaling mediated by the RNA helicase RIG-I during antiviral responses. *Nat. Immunol.* 12, 37–44.
77. Jiang, L., Wang, Y. J., Zhao, J., Uehara, M., Hou, Q., Kasinath, V., Ichimura, T., Banouni, N., Dai, L., Li, X., Greiner, D. L., Shultz, L. D., Zhang, X., Sun, Z. Y. J., Curtin, I., et al. (2020) Direct tumor killing and immunotherapy through anti-SerpinB9 therapy. *Cell* 183, 1219–1233.e18.
78. Ma, E. H., Bantug, G., Griss, T., Condotta, S., Johnson, R. M., Samborska, B., Mainolfi, N., Suri, V., Guak, H., Balmer, M. L., Verway, M. J., Raissi, T. C., Tsui, H., Boukhaled, G., Henriques da Costa, S., et al. (2017) Serine is an essential metabolite for effector T cell expansion. *Cell Metab.* 25, 345–357.
79. Martin, E., Palmic, N., Sanquer, S., Lenoir, C., Hauck, F., Mongellaz, C., Fabrega, S., Nitschké, P., Esposti, M. D., Schwartzentruber, J., Taylor, N., Majewski, J., Jabado,

- N., Wynn, R. F., Picard, C., et al. (2014) CTP synthase 1 deficiency in humans reveals its central role in lymphocyte proliferation. *Nature* 510, 288–292.
80. Pouillon, V., Hascakova-Bartova, R., Pajak, B., Adam, E., Bex, F., Dewaste, V., VanLint, C., Leo, O., Erneaux, C., and Schurmans, S. (2003) Inositol 1, 3, 4, 5-tetrakisphosphate is essential for T lymphocyte development. *Nat. Immunol.* 4, 1136–1143.
81. Chung, H. K., Yi, Y. W., Jung, N. C., Kim, D., Suh, J. M., Kim, H., Park, K. C., Song, J. H., Kim, D. W., Hwang, E. S., Yoon, S. H., Bae, Y. S., Kim, J. M., Bae, I., and Shong, M. (2003) CR6-interacting factor 1 interacts with Gadd45 family proteins and modulates the cell cycle. *J. Biol. Chem.* 278, 28079–28088.
82. Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K., and Liu, X. (2014) Comparison of RNA-seq and microarray in transcriptome profiling of activated T cells. *PLoS One* 9, e78644.
83. S'ecca, C., Faget, D. V., Hanschke, S. C., Carneiro, M. S., Bonamino, M. H., de Araujo-Souza, P. S., and Viola, J. P. B. (2016) IRF2BP2 transcriptional repressor restrains naive CD4 T cell activation and clonal expansion induced by TCR triggering. *J. Leukoc. Biol.* 100, 1081–1091.
84. Patrick, M. S., Cheng, N. L., Kim, J., An, J., Dong, F., Yang, Q., Zou, I., and Weng, N. P. (2019) Human T cell differentiation negatively regulates telomerase expression resulting in reduced activation-induced proliferation and survival. *Front. Immunol.* 10, 1993.

85. Röth, A., Yssel, H., P`ene, J., Chavez, E. A., Schertzer, M., Lansdorp, P. M., Spits, H., and Luiten, R. M. (2003) Telomerase levels control the lifespan of human T lymphocytes. *Blood* 102, 849–857.
86. Jackson, M. R., Loll, P. J., and Jorns, M. S. (2019) X-ray structure of human sulfide:quinone oxidoreductase: Insights into the mechanism of mitochondrial hydrogen sulfide oxidation. *Structure* 27, 794–805.e4.
87. Mnatsakanyan, R., Markoutsas, S., Walbrunn, K., Roos, A., Verhelst, S. H. L., and Zahedi, R. P. (2019) Proteome-wide detection of S-nitrosylation targets and motifs using bioorthogonal cleavable-linker-based enrichment and switch technique. *Nat. Commun.* 10, 1–12.
88. Yan, T., Palmer, A. B., Geiszler, D. J., Polasky, D. A., Armenta, E., Nesvizhskii, A. I., and Backus, K. M. (2021) Enhancing Cysteine Chemoproteomic Coverage Through Systematic Assessment of Click Chemistry Product Fragmentation.
89. Zanon, P. R. A., Yu, F., Musacchio, P., Lewald, L., Zollo, M., Krauskopf, K., Mrdovi`c, D., Raunft, P., Maher, T. E., Cigler, M., Chang, C., Lang, K., Toste, F. D., Nesvizhskii, A. I., and Hacker, S. M. (2021) Profiling the proteome-wide selectivity of diverse electrophiles. *ChemRxiv*. <https://doi.org/10.26434/chemrxiv.14186561.v1>.
90. Shi, Y., Fu, L., Yang, J., and Carroll, K. S. (2021) Wittig reagents for chemoselective sulfenic acid ligation enables global site stoichiometry analysis and redox-controlled mitochondrial targeting. *Nat. Chem.* 2021, 1–11.
91. Backus, K. M., Cao, J., and Maddox, S. M. (2019) Opportunities and challenges for the development of covalent chemical immunomodulators. *Bioorg. Med. Chem.* 27, 3421–3439.

92. Deutsch, E. W., Csordas, A., Sun, Z., Jarnuczak, A., Perez-Riverol, Y., Ternent, T., Campbell, D. S., Bernal-Llinares, M., Okuda, S., Kawano, S., Moritz, R. L., Carver, J. J., Wang, M., Ishihama, Y., Bandeira, N., et al. (2017) The ProteomeXchange consortium in 2017: Supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.* 45, D1100–D1106.
93. Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., Perez, E., Uszkoreit, J., Pfeuffer, J., Sachsenberg, T., Yilmaz, S., et al. (2019) The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Res.* 47, D442–D450.
94. Chambers, M. C., Maclean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., Hoff, K., Kessner, D., Tasman, N., Shulman, N., Frewen, B., et al. (2012) Across platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* 30, 918–920.

Chapter 3: Multi-omic stratification of the missense variant cysteinome

Chapter 3 is a version of Desai, H. S.; Ofori, S.; Boatner, L. M.; Yu, F.; Villaneuva, M.; Ung, N.; Nesvizhskii, A.; Backus, K. M. Multi-Omic Stratification of the Missense Variant Cysteinome. bioRxiv **2023**. <https://doi.org/10.1101/2023.08.12.553095>

Heta Desai^{1,5}, Samuel Ofori¹, Lisa Boatner^{1,2}, Fengchao Yu⁴, Miranda Villanueva^{1,5}, Nicholas Ung, Alexey I. Nesvizhskii^{3,4}, Keriann Backus^{1,2,5,6,7,8,*}

1. Biological Chemistry Department, David Geffen School of Medicine, UCLA, Los Angeles, CA, 90095, USA.
2. Department of Chemistry and Biochemistry, UCLA, Los Angeles, CA, 90095, USA.
3. Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, 48109, USA.
4. Department of Pathology, University of Michigan, Ann Arbor, MI, 48109, USA.
5. Molecular Biology Institute, UCLA, Los Angeles, CA, 90095, USA.
6. DOE Institute for Genomics and Proteomics, UCLA, Los Angeles, CA, 90095, USA.
7. Jonsson Comprehensive Cancer Center, UCLA, Los Angeles, CA, 90095, USA.
8. Eli and Edythe Broad Center of Regenerative Medicine and Stem Cell Research, UCLA, Los Angeles, CA, 90095, USA.

*Corresponding Author: kbackus@mednet.ucla.edu

Abstract

Cancer genomes are rife with genetic variants; one key outcome of this variation is gain-of-cysteine, which is the most frequently acquired amino acid due to missense variants in COSMIC. Acquired cysteines are both driver mutations and sites targeted by precision therapies. However, despite their ubiquity, nearly all acquired cysteines remain uncharacterized. Here, we pair cysteine chemoproteomics—a technique that enables proteome-wide pinpointing of functional, redox sensitive, and potentially druggable residues—with genomics to reveal the hidden landscape of cysteine acquisition. For both cancer and healthy genomes, we find that cysteine acquisition is a ubiquitous consequence of genetic variation that is further elevated in the context of decreased DNA repair. Our chemoproteogenomics platform integrates chemoproteomic, whole exome, and RNA-seq data, with a customized 2-stage false discovery rate (FDR) error controlled proteomic search, further enhanced with a user-friendly FragPipe interface. Integration of CADD predictions of deleteriousness revealed marked enrichment for likely damaging variants that result in acquisition of cysteine. By deploying chemoproteogenomics across eleven cell lines, we identify 116 gain-of-cysteines, of which 10 were liganded by electrophilic druglike molecules. Reference cysteines proximal to missense variants were also found to be pervasive, 791 in total, supporting heretofore untapped opportunities for proteoform-specific chemical probe development campaigns. As chemoproteogenomics is further distinguished by sample-matched combinatorial variant databases and compatible with redox proteomics and small molecule screening, we expect widespread utility in guiding proteoform-specific biology and therapeutic discovery.

Introduction

The average human genome is rife with sequence variation and differs from the reference at roughly 3.5 million sites¹. This profound genetic variation gives rise to human diversity and disease. While the fraction of single nucleotide variants (SNVs) that occur in protein-coding make up a small fraction of all known variants, most known disease-causing mutations are found in protein coding sequences. Nearly all (>98%) of nonsynonymous protein-coding SNVs are missense variants that result in the substitution of single amino acids². There are over 2 million coding mutations that have been identified in human cancers (Catalogue of Somatic Mutations [COSMIC] database), of which >90% are missense variants^{3,4}. However, only a tiny fraction of these genetic variants (~3,400) have been identified as putative missense driver mutations⁵ that confer selective growth advantages to cancer cells with the remaining mutations acting as “passengers.”

Quite surprisingly given the relative rarity of cysteine (2.3% of all residues in a human reference proteome)⁶, cysteine is the most commonly acquired amino acid due to somatic mutations in human cancers⁷. Given the unique chemistry of the cysteine thiol, including its nucleophilicity and sensitivity to oxidative stress, a subset of these residues almost unquestionably have a substantial impact on protein function. Exemplifying this paradigm, a number of driver mutations are gained cysteines, including Gly12Cys KRAS Tyr279Cys SHP2, Ser249Cys FGFR, and Arg132Cys IDH1⁸⁻¹². A likely reason for the ubiquity of cysteine acquisition is the comparative instability of CpG motifs; C-T transitions are nearly ten times more common than other missense mutations in cancer¹³, and these transitions should favor gain-of-cysteine codons.

Due to its nucleophilicity and sensitivity to alkylation, cysteine residues have emerged as attractive sites to target with chemical probes. Covalent compounds can access small and poorly defined binding sites and can efficiently block high-affinity interactions (e.g. protein-protein interactions) or compete with high concentrations of endogenous biomolecules (e.g. ATP). There are numerous examples of cysteine-reactive clinical candidates and drugs, including the blockbuster covalent kinase inhibitors (e.g. Afatinib and Ibrutinib¹⁴⁻¹⁶) and covalent compound that react with the Gly12Cys mutated oncogenic form of the GTPase KRAS (e.g. ARS-1620 and sotorasib^{9,17-19}), a protein previously thought to be undruggable.

Mass spectrometry-based chemical proteomic methods, including those developed by our lab, have begun to unlock the therapeutic potential of the cysteinome. By capturing and enriching cysteines using highly reactive chemical probes, such as iodoacetamide alkyne (IAA) and iodoacetamide desthiobiotin, the studies have assayed the ligandability of upwards of 25% of all cysteines in the human proteome²⁰⁻²⁹. Cysteine chemoproteomics has even enabled the discovery of new lead molecules that target specific cysteines, including JAK³⁰, SARM1³¹, PPP2R1A³², XRCC5³³, NRB01³⁴, and pro-CASP8²⁹. Several new strategies have made substantial inroads into stratifying cysteine functionality to achieve function-first readouts of the likelihood of a covalent modification altering the labeled protein, including quantifying intrinsic cysteine nucleophilicity²⁵, by pairing of chemoproteomics with CRISPR-base editing³⁵, by performing proteomic stratification of covalent-modification induced altered protein complexes³⁶, and our own work combining computational predictions of genetic pathogenicity with cysteine chemoproteomics²⁷.

Single amino acid variants (SAAVs) encoded by missense mutations, including those that result in acquisition of cysteine, are almost universally missed by chemoproteomic studies. A key reason for this gap is that most genetic variants are not found in reference protein sequence databases used to identify peptides from acquired tandem mass spectrometry (MS/MS) data^{20–29}. Understanding whether a genetic variant is translated into protein is a critical step for characterizing the functional impact and therapeutic relevance of genomic variation. Proteogenomic studies that implement custom variant-containing sequence databases for search have enabled proteome-wide detection of protein coding variants, including SAAVs and splice variants^{37–43}. When compared to variant calling at the genomic level, the coverage of these studies remains comparatively small, spanning tens to hundreds of peptides, with the exception of recent studies employing ultra deep fractionation^{44,45} resulting in thousands of identified variants. These studies all share general data processing pipelines. Variant calling is performed on next-gen sequencing (NGS) data, then customized databases featuring both canonical protein sequences and sequences encoding SAAV-, insertion/deletions (indels)-, or splice variant-proteins are generated, using customized tools, such as Spritz⁴⁶, CustomProDB⁴⁷, Galaxy-P⁴⁸, and sapFinder⁴⁹. While targeted proteomics methods, such as parallel reaction monitoring (PRM) have enabled focused monitoring of high value variant-containing peptides⁵⁰, including encoding driver mutations, the broader landscape of translated SAAVs remains to be fully explored.

There are two central complexities to these pipelines that have only recently begun to be addressed. The first challenge is that, by relying on exome-only sequencing and short read sequencing, the relative proximity of two or more variants in the same gene

(whether they are on the same or opposite chromosomes) is not typically apparent. A notable exception is the recent integration of long read sequencing for de-novo database construction with sample-specific proteomics to characterize novel protein isoforms⁵¹. Consequently, multi-variant peptides are typically not detected by most proteogenomics workflows that rely on databases featuring either single-each or all-in-one SAAV-containing proteins. Such search strategies also introduce higher chances of false positive identification⁵². All possible cancer-derived aberrant peptide sequences, reflecting increased genetic complexity of tumor genomes, increases the size of the custom databases and thus search spaces. One solution to the false discovery rate (FDR) challenge is to calculate a class-specific FDR (separating the FDR calculations for the variant-containing peptides and reference peptides)⁵². An alternative strategy to ensure class-specific FDR control is to perform a 2-stage database search⁵³. In this strategy, the first search of acquired MS/MS spectra is performed against a reference database of canonical protein sequences. Subsequently, peptide to spectrum (PSM) matches identified with a certain high level of confidence (e.g. passing 1% FDR) are removed, and the remaining spectra are then searched against a variant-containing database. While implementation of such strategies in prior proteogenomic studies highlights the importance of rigorous statistical validation of identified variant-containing peptides^{53–55}, the requirement for customized pipelines has so far limited widespread adoption.

Here we develop and deploy chemoproteogenomics as an integrated platform tailored to capture the missense variant cysteinome. Chemoproteogenomics unites a missense-variant focused proteogenomic pipeline with mass spectrometry-based cysteine chemoproteomics. By mining publically available datasets, including COSMIC,

dbSNP, and ClinVar, we reveal that gain-of-cysteine variants are a ubiquitous consequence of genetic variation. We further reveal that DNA repair deficient cell lines are particularly enriched for acquired cysteines, together with a general high burden of rare and predicted deleterious variants. Guided by these discoveries, we generate combinatorial cell-specific custom databases built from whole exome and RNA-Seq data for eleven cell lines. Chemoproteogenomic analysis with a user-friendly FragPipe computational platform, extended to support 2-stage database search and FDR estimation, identified >1,400 total unique variants, including 629 chemoproteomic enriched variant-proximal cysteines and 103 gain-of-cysteines. Chemoproteogenomics also robustly identifies ligandable SAAVs that alter cysteine oxidation state and outperforms bulk proteogenomic analysis for capture of SAAVs with lower variant allele frequency. The utility of chemoproteogenomics is further showcased through our identification of iodoacetamide-labeled Cys67 (Cys91) in the highly variable peptide binding-groove of HLA-B. In sum, chemoproteogenomics sets the stage for enhanced global understanding of the functional and therapeutic relevance of the missense variant proteome.

Results

High missense burden cancer cell lines are rich in acquired cysteines, including in census genes. Our first step to realize variant-directed chemoproteomics was to mine existing publicly available missense repositories to assess the scope of acquired cysteines present in cancer genomes (COSMIC) and healthy genomes (dbSNP) (**Figure 1A**). By doing so, we sought to achieve three goals: (1) validate prior reports of high

cysteine acquisition in cancer^{7,56,57} (2) determine whether cysteine acquisition is a privileged feature of cancer genomes, and (3) establish a panel of variant rich cell lines. We analyzed publicly available sequencing data of 1,020 cell lines, found in the Catalogue of Somatic Mutations in Cancer Cell Lines Project database^{58,59} (COSMIC-CLP, release v96), to establish a panel of high mutational burden tumor cell lines; our hypothesis was high missense burden cell lines would be enriched for acquired cysteine SAAVs, including those found in Census genes⁶⁰ and residues that are driver mutations. The top 15 cell lines with the highest mutational burden (**Figure 1B, S1A, Table S1**) encode 77,693 total unique missense variants, which represents ~18% of all unique missense variants in COSMIC-CLP.

We next evaluated whether these identified missense-rich cell line genomes were similarly enriched for gained cysteine SAAVs. We calculated the net gain amino acid changes (total gained minus total lost) encoded by all coding missense variants in this cell line panel (**Figure S2**), which revealed a marked enrichment for acquired histidines and cysteines together with loss of arginine, both for the aggregate cell line panel and for individually analyzed cell line datasets (**Figure S3**). As calculations of net gain can fail to distinguish high versus low missense burden cell lines, we also further stratified these cell lines based on total gained and total lost amino acids (**Figure S1B, S4, S5**), which further substantiated the enrichment for gain-of-cysteine across all of the top 15 missense variant burden cell lines analyzed (**Figure 1C, S1B**). This marked cysteine enrichment in cancer cell line genomes is consistent with previously reported aggregate analysis, not stratified by cell line, of all available COSMIC missense data^{7,56,57}. Our own analysis of all COSMIC-CLP mutations shows cysteine as the second most gained residue (**Figure 1D**).

The genomes of the top 15 missense cell lines encoded 4,725 total gained cysteines, found in 3,688 genes. Showcasing the potential therapeutic relevance of this set, <10% of these identified genes have been targeted by FDA approved drugs^{20,61} (**Figure 1E**). Notably, 219/738 Census genes (v98) were found to harbor one or more gained cysteines, including NRAS (G12C), which is found in the Molt-4 cell line; TP53 (R273C) found in the KARPAS-45 cell line; GNAS (R218C) in the CW-2, SNU-175, and HT-115 cell lines; FBXW7 (R505C) in Jurkat and KARPAS-45 cell line; ASXL1 (W796C) found in HCT-15 cell line, and KEAP1 (Y33C) found in the Hec-1 cell line (**Table S1**).

dMMR cell lines are enriched for SAAVs, including acquired cysteines. Cancer genomes display characteristic patterns of mutations, or signatures, that have developed from biological processes specific to the course of the cancer^{62,63}. Endogenous and exogenous sources of DNA damage, left uncorrected due to faulty repair pathways, often lead to high tumor mutational burdens. Microsatellite instability (MSI) is a hypermutable phenotype caused by deficiency in mismatch repair (dMMR). High MSI tumors have higher mutational burdens; the converse is not true as high mutational burden tumors do not always display MSI⁶⁴. Eight out of fifteen of the top missense burden cell lines reported in COSMIC were observed to be derived from colorectal carcinoma (CRC) (**Figure 1B, S3**). As ~15% of CRCs are reported to have elevated MSI⁶⁵⁻⁶⁷, this high CRC missense burden is to be expected^{64,68}. While Jurkat, Molt-4 and Hec-1B cells are not CRC, both have previously been reported as dMMR with mutations in mismatch repair machinery^{69,70}. Unexpectedly, MeWo cells, which are derived from metastasized melanoma and reported to be microsatellite stable (MSS)⁷¹, also exhibited a high burden

of missense mutations. The majority of missense rich cell lines, including the dMMR lines were observed to encode between 200 and 500 acquired cysteine SAAVs (**Figure S1B**). However, a significant depletion of gained cysteines relative to total variant burden was observed for MeWo and SW684 (**Figure 1C**).

Acquired cysteines are ubiquitous in both healthy and diseased genomes. We next asked whether this marked enrichment for gained cysteines was specific to cancer genomes or a more universal consequence of human genetic variation, with the overarching goal of facilitating efforts to pinpointing acquired cysteines with therapeutic relevance. Complicating matters, gain-of-cysteine missense variants are also expected to be ubiquitous in healthy genomes, due to the comparative instability of CpG—a key consequence of this instability is the frequent loss of arginine codons (4/6 CG dinucleotides)⁷². We aggregated and quantified the amino acid changes resulting from common missense variants reported by dbSNP⁷³, a repository of single nucleotide polymorphisms and ClinVar⁷⁴, a repository of variants with reported pathogenicity. We find that cysteine acquisition is the third most common consequence of missense variants identified in dbSNP (**Figure 1D, Table S1**) for common variants—common variants are defined by NCBI as of germline origin and/or with a minor allele frequency (MAF) of ≥ 0.01 in at least one major population, with at least two unrelated individuals having the minor allele. Analogous stratification of variants reported by ClinVar also revealed a preponderance of gained cysteines compared with lost cysteines, albeit to a more modest degree than that observed for cancer genomes (**Figure S6 and Table S1**). For the

pathogenic variant subset of ClinVar, both gain- and loss-of-cysteine and gain-of-proline were frequently observed (**Figure S6**).

An expanded cell line panel incorporates high value acquired cysteines. Across the >2 million missense variants reported in COSMIC, 52 acquired cysteines are reported as putative driver mutations (dN/dS values)⁷⁵ in the Cancer Mutation Census (**Table S1**). Consequently, nearly all acquired cysteine SAAVs are of uncertain functional significance for tumor cell growth and survival. Given that one of our key objectives is to enable rapid proteomic identification and subsequent electrophilic compound screening of functional variants, we next stratified the top missense variant cell lines based on known driver mutations and damaging variants. We find that top missense cell lines that are readily available for purchase encode NRAS G12C, KRAS G12D, PIK3CA E545K, and TP53 R248Q variants among other known driver mutations (**Table S1**). Given the considerable interest in targeting G12C KRAS, we opted to add several KRAS mutated cell lines to our panel (MIA-PACA-2, H2122, and H358) in order to favor detection of the G12C peptide. Notably, the smoking-associated mutational signature is C→A/G→T⁷⁶, which should favor gain-of-cysteines. Therefore, we additionally sought to test whether smoking associated NSCLC-derived H2122 and H1437 adenocarcinoma cell lines would be enriched for acquired cysteines when compared to other proficient mismatch repair (pMMR) cell lines, including lung cancer cell lines (H358 NSCLC and H661 metastatic large cell undifferentiated carcinoma (LCUC) lung cancer cell lines). Lastly, we opted to include CACO-2 cells, an MSS CRC cell line, to test the feasibility of capturing driver mutations located proximal to chemoproteomics detectable cysteines—Caco-2 cells

express mutant SMAD4 (D351H), a variant implicated in blocking SMAD homo- and hetero-oligomerization⁷⁷ and located proximal to two previously chemoproteomics detected cysteines (C345 and C363)^{21,22}. Our prioritized cell line panel features 11 cell lines in total (2 female and 9 male) spanning 6 tumor types and encoding 22,559 somatic variants and 1,296 somatic acquired cysteines, as annotated by COSMIC-CLP (**Figure 1F, Table S1-S2**), with aggregate enrichment for gained cysteines observed for the entire panel (**Figure S7, S8**). Of the proteins that harbor gained cysteines, 486 are Census genes and 5% are targeted by FDA approved drugs (**Table S1**).

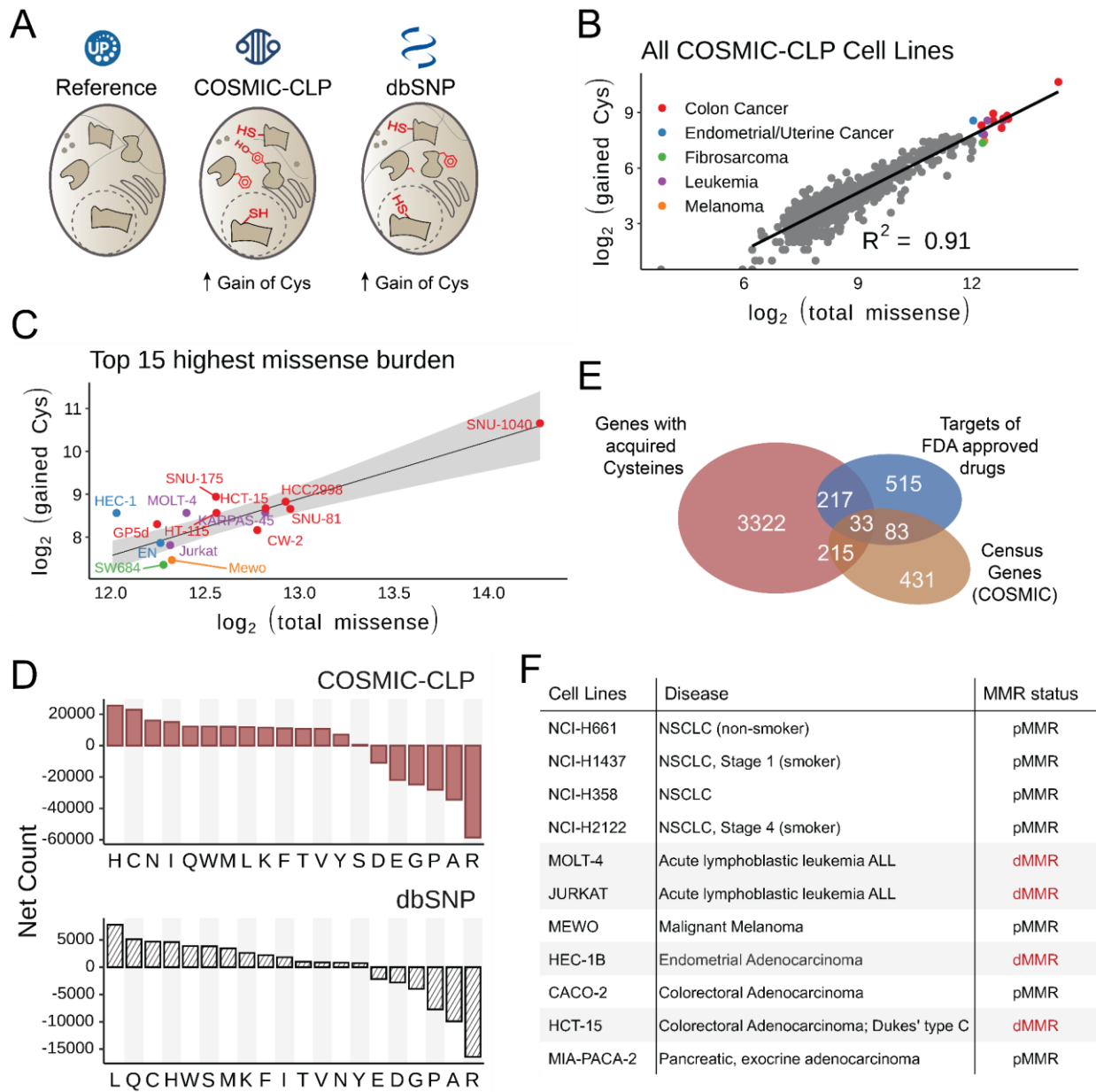


Figure 1. Acquired cysteines are prevalent across cancer genomes, particularly for high missense burden cell lines. A) The full scope of acquired cysteines in the COSMIC Cell Lines Project (COSMIC-CLP, cancer.sanger.ac.uk/cell_lines) (v96)^{58,59} and dbSNP (4-23-18)⁷³ were analyzed. B) 1,020 cell lines stratified by number of gained cysteines and total missense mutations; color indicates cancer type for top 15 highest missense count cell lines. C) Top 15 cell lines with highest missense burden from panel B; linear regression and 95% confidence interval shaded in gray. D) Net missense mutations (gained-lost) from COSMIC-CLP (v96) and common SNPs (dbSNP 4-23-18). E) Overlap of genes with acquired cysteines in top 15 subset from panel B with Census genes and targets of FDA approved drugs. F) Panel of cell lines used in this study with MMR status (dMMR= deficient mismatch repair, pMMR=proficient mismatch repair). Data found in **Table S1**.

dMMR cell lines are enriched for rare predicted missense changes, including acquired cysteines. Given the preponderance of acquired cysteine SAAVs observed across COSMIC, ClinVar, and dbSNP, we postulated that cancer genomes would be enriched for both rare and common gain-of-cysteine mutations. To both test this hypothesis and enable the building of sequence databases for proteogenomics search, we sequenced exomes and RNA of our cell lines and subjected NGS reads to variant-calling (**Figure 2A, Figure S9**). For all 11 cell lines sequenced, we identified on average 82% of the variants reported in COSMIC-CLP and 70% of missense mutations reported by Cancer Cell Line Encyclopedia (CCLE)⁷¹ databases (**Table S2**). Driver mutations (CMC significant, dN/dS q-values) identified include KRAS G12C for MIA-PACA-2, H358, and H2122 cell lines, PIK3CA E545K in HCT-15, and FBXW7 R505C in Jurkat cells (**Table S2**). 9,190 total rare variants were identified that had been not previously reported in COSMIC-CLP, including 435 variants encoding acquired cysteines (**Table S2**).

As with our analysis COSMIC-CLP (**Figure 1B**), we detected a high missense burden for the dMMR cell lines compared to the pMMR cell lines. MeWo cells were an exception, with a missense burden comparable to that of the dMMR cell lines (**Figure 2B**). Analysis of DNA damage repair-associated genes revealed specific mutations (**Table S2**), including DDB2 R313* in MeWo cells, which provide an explanation for the previously unreported high missense burden—inactivating mutations in DDB2 are implicated in deficient nucleotide excision repair⁷⁸.

We next subsetted the data into rare and common variant categories, using dbSNP common variants (04-23-2018 00-common_all.vcf.gz) (**Table S2**)⁷³. The dMMR cell lines, together with the MeWo cells, have proportionally more rare variants compared to

common variants (**Figure 2B**), irrespective of sequencing coverage (**Figure S10**). Further SAAV analysis revealed net gain of histidine, isoleucine, and cysteine as the most frequent amino acids gained across the common and rare subsets (**Figure 2C**). We find that cysteine acquisition is a more frequent consequence of common variants detected in pMMR cell lines (**Figure 2D**).

In contrast with the common variants, the net gained SAAV signatures encoded by rare variants differed markedly between dMMR and pMMR cell lines (**Figure 2D, S11-13**). No significant difference between the number of gained cysteines was observed for the smoking-associated lung cancer cell lines (**Figure S14**). By contrast, in the dMMR cell lines, we detected a sizable increase, when compared to the pMMR cell lines, of acquired rare SNVs encoding Cys, along with His, Ile, Asn, Tyr, and Tryp (**Figure 2D, Figure S11**). Beyond cysteine acquisition, the SAAV signature for MeWo cells was observed to be distinct, with pronounced gain-of rare Phe and Lys detected (**Figure S11-13**), consistent with UV radiation induced pyrimidine dimers, which result in gain-of F and K (**Figure S15, Table S2**). These findings together with our analysis of the top missense cell lines in COSMIC-CLP indicate that previously reported widespread cysteine acquisition in cancer genomes is predominated by mismatch repair deficient cell lines.

Rare gained cysteines in dMMR cell lines are enriched for high CADD scores. With the overarching goal of facilitating identification of likely functional variants, we next stratified the predicted deleteriousness of the identified missense variants (**Figure 2E, Table S2**). We focused on the Combined Annotation Dependent Depletion (CADD) score, due to its high reported specificity and sensitivity⁷⁹ and our prior findings that showed

strong association between cysteine functionality and high CADD score²⁷. Unsurprisingly, our analysis revealed higher CADD scores for rare variants compared to common variants, across the cell line panel (**Figure 2E, Table S2**). More unexpectedly, we observed a more marked increase in the predicted pathogenicity of the rare variants detected in dMMR cell lines compared with pMMR cell lines (the top 1% most predicted deleterious mutations have CADD phred-scaled scores > 20) (**Figure 2F-G, S16-17**). This enrichment for high CADD score rare variants held true for the MeWo cells. Further stratification by specific gained or lost amino acids (**Figure 2H, Figure S18-21**), revealed that gained cysteine missense are the most significantly enriched for high predicted deleterious scores across all pMMR and dMMR cell lines (**Figure S19, Table S2**)—a notable exception are the MeWo cell line variants for which gain-of Phe, Lys, and Leu codons are the most high CADD scoring variants (**Figure S22**).

As only a small fraction of the acquired cysteines are known driver mutations, we next restricted our analysis to include only the 388 total variants localized to hotspot mutations, as annotated by CCLE and The Cancer Genome Atlas (TCGA). We find that gain of cysteine within TCGA hotspot mutations is markedly enriched for high CADD score variants (**Figure S21**). Notable high CADD score hotspot acquired cysteines include the tumor suppressor FBXW7 R505C in Jurkat cells, the metalloprotease ADAMTS1 R604C in Molt-4 cells, and extrin-associated protein SCYL3 R61C in MeWo cells. 98% (50/51) of these cysteines are gained due to loss of arginine, which aligns with the observed parallel enrichment for high CADD scores at loss of arginine hotspot variants (**Figure S23**).

dMMR rare variants are enriched for proximity to known functional sites. To further broaden our understanding of the functional landscape of cysteine acquisition, we also analyzed proximity to known functional sites and sites of post translational modification (**Table S2**). We find that the dMMR rare variant set is enriched for known proximal active site/binding site residues (**Figure 2I**). Intriguingly, analysis of known PTM modified sites reported by Phosphosite⁸⁰ revealed a significant association between arginine methylation sites and rare variants in dMMR cell lines (**Figure 2I**). These findings are consistent with loss of arginine as a frequent consequence of exonic CpG mutability^{72,81} together with roles of MMR in protecting against CpG associated deamination⁸². As 60% of the gained cysteines in our data resulted from loss of arginine (**Figure S24**), we expected that many of these variants will result in altered PTM status (**Figure 2J**).

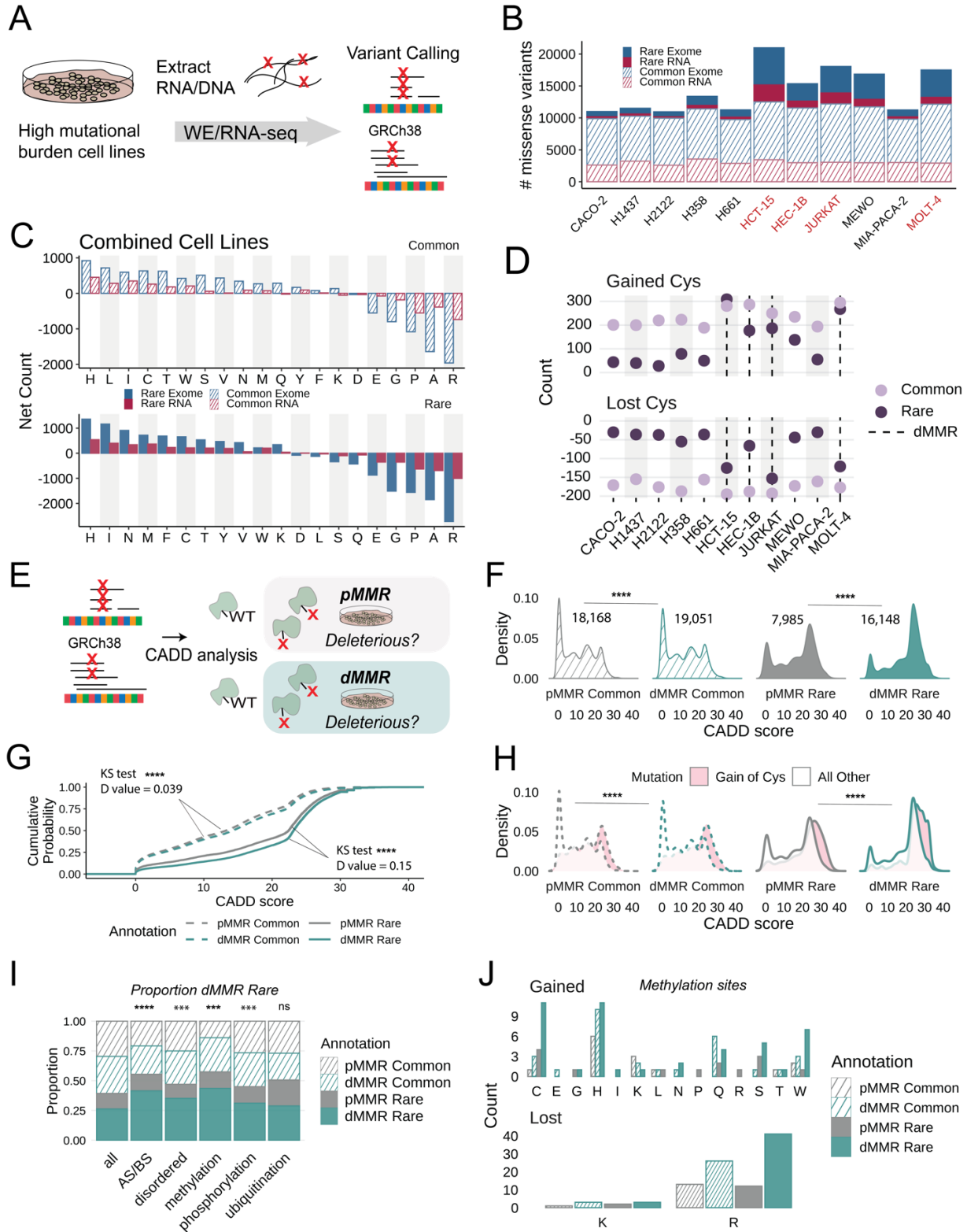


Figure 2. dMMR cell lines are enriched for rare, predicted deleterious gain-of-cysteine mutations. A) Sequencing portion of the ‘chemoproteogenomic’ workflow to identify chemoproteomic detected variants—extracted genomic DNA or RNA from cell lines undergo sequencing followed by variant calling using

Platypus (v0.8.1)⁸³ and GATK-Haplotype Caller (v4.1.8.1)⁸⁴ for RNA and exomes respectively and predicted missense changes were computed. B) Total numbers of missense mutations identified from either RNA-seq or WE-seq; stripe vs solid denotes common and rare variants, red text indicate dMMR cell lines. C) Net amino acid changes for all cell lines combined. D) Totals of gained and lost cysteine in each cell line separated by rare and common variants, dashed line indicates dMMR cell lines. E) Scheme of CADD score analysis for two dMMR and non-dMMR cell lines. F) Distribution of CADD scores for indicated variant grouping; statistical significance was calculated using Mann-Whitney U test, **** $p < 0.0001$. G) Empirical cumulative distributions (ECDF) were computed for CADD scores with indicated grouping; statistical significance was calculated using two-sample Kolmogorov-Smirnov test, **** $p < 0.0001$. H) CADD score distributions for cysteine gained amino acid indicated separated by grouping; statistical significance between gained Cys values was calculated using Mann-Whitney U test, **** $p < 0.0001$. I) Proportion of variants belonging to the indicated sites; AS/BS = in or near active site/binding site as annotated by UniProtKB or Phosphosite; statistical significant calculated using two-sample test of proportions, *** $p < 0.001$, **** $p < 0.0001$, ns $p > 0.05$. J) Amino acid changes at protein methylation sites as identified by Phosphosite. Data found in **Table S2**.

Variant peptide identification enabled by MSFragger 2-stage database search and false discovery rate (FDR) estimation. To enable chemoproteomic detection of acquired cysteine SAAV-containing peptides and SAAVs found in peptides with canonical cysteines, we next established a customized proteogenomics pipeline (**Figure 3A, B**). Motivated by the prior report³⁸ that demonstrated proteogenomic sample searches performed with sample-specific databases both improved coverage (~45% more variants) and decreased rates of SAAV peptide false discovery, we generated cell line-specific variant peptide databases from HEK293T RNA-seq data (**Figure 3A, Table S3**). Next, to afford a reduction to the likelihood that a variant peptide will be mismatched to wild-type spectra⁵³, we established a 2-stage database search and FDR control scheme (**Figure 3B**), using MSFragger (v3.5)/Philosopher^{85,86} command line pipeline within FragPipe computational platform (detailed in Methods).

We then subjected our chemoproteogenomics pipeline to benchmarking by generating a set of high coverage cysteine chemoproteomics datasets (**Figure 3B**) in which cell lysates labeled with iodoacetamide alkyne (IAA)²⁵ and conjugated isotopically labeled 'light' (¹H₆) or 'heavy' (²H₆) biotin-azide reagents⁸⁷ (+6 Da mass difference

between the reagents) were combined pairwise in biological triplicate at different H/L ratios (1:1, 10:10, 1:4, 4:1, 1:10, and 10:1). By searching these datasets using our 2-stage search, we sought to validate the accuracy of variant identification. Peptide quantification using IonQuant^{88,89}, following the workflow shown in **Figure 3A**, revealed MS1 intensity ratios for both canonical and variant peptide sequences that matched closely with the expected values (**Figure 3C, Table S3**). We also compared the retention times of the heavy- and light-peptides and observed an ~2-3 sec shift for the deuterated heavy sequences for both the variant and canonical peptide sequences (**Figure 3D, Table S3**). These retention time shifts are consistent with our previous study⁸⁷ and with prior reports^{90,91}. Analogous to studies that utilize isotopically enriched synthetic peptide standards to validate novel peptide sequences⁹²⁻⁹⁴, the observed co-elution of both heavy and light variant peptides provides further evidence to support the low FDR of our data processing pipeline. Lastly, the high concordance between observed and expected MS1 ratios provides compelling support for the use of the heavy and light biotin azide reagents in competitive cysteine-reactive compound screens, in which elevated MS1 intensity ratios are indicative of a compound modified cysteine.

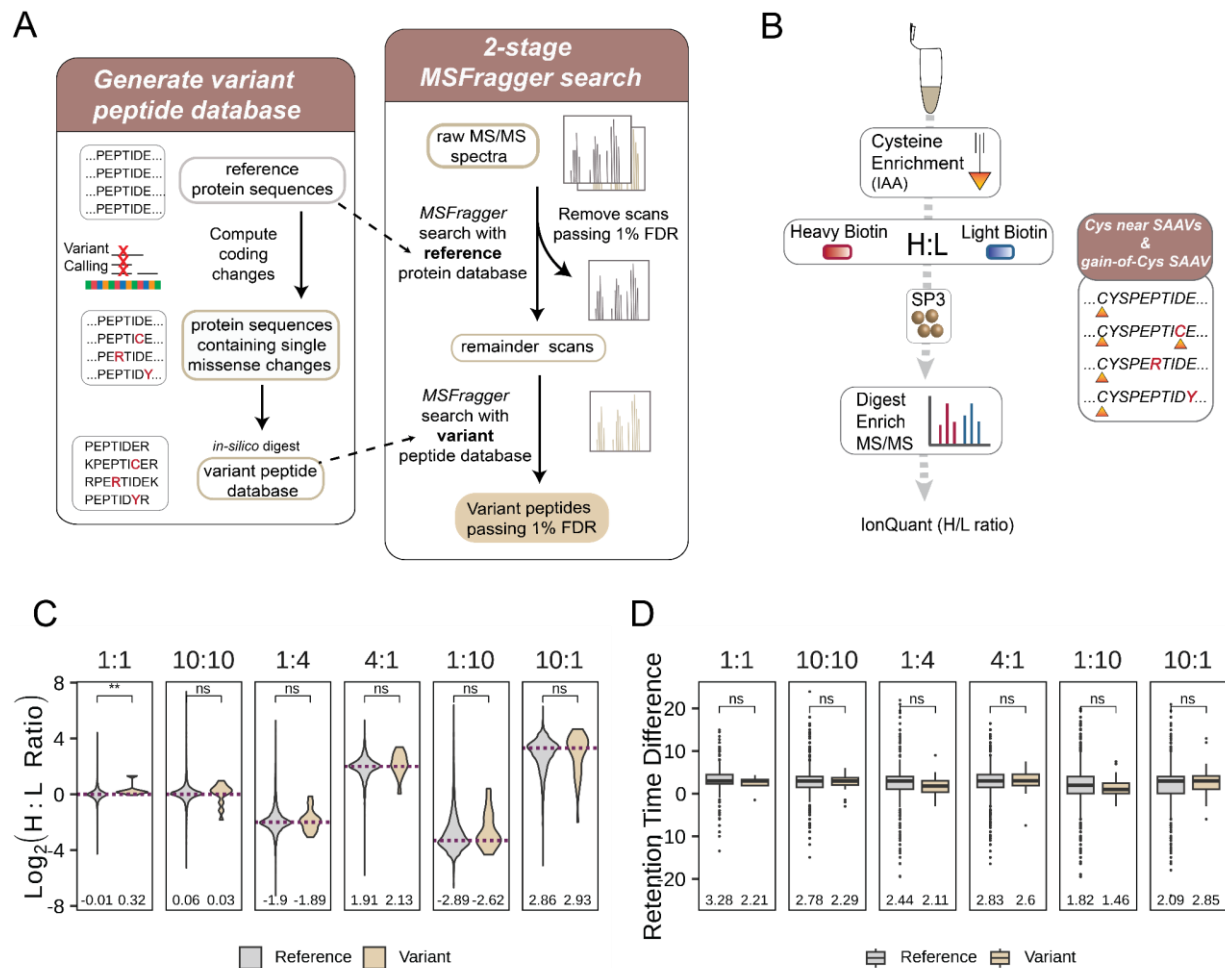


Figure 3. Variant peptide identification implementing an MSFragger-search pipeline A) 2-stage MSFragger-enabled variant searches—variant databases are generated from non-redundant reference protein sequences that are *in-silico* mutated to incorporate sequencing-derived missense variants followed by 2-stage MSFragger/PeptideProphet search to identify confident variant-containing peptides. First, raw spectra are searched against a normal reference protein database, confidently matched spectra (passing 1% FDR) are removed and remainder spectra are searched with a variant tryptic database. B) Chemoproteomics workflow to validate heavy and light biotin⁸⁷. HEK293T cell lysates were labeled with pan-reactive iodoacetamide alkyne (IAA) followed by ‘click’ conjugation onto heavy or light biotin azide enrichment handles in known ratios. Following neutravidin enrichment, samples are digested and subjected to MS/MS analysis. C) Heavy to light ratios (H:L) from triplicate datasets comparing identifications from reference and variant searches; mean ratio value indicated, *dashed lines* indicate ground-truth log₂ ratio, statistical significance was calculated using Mann-Whitney U test, ** p < 0.01, ns p > 0.05. D) Retention time difference for heavy and light identified peptides for reference and variant-searches; mean value indicated, statistical significance was calculated using Mann-Whitney U test, ns p > 0.05. Data found in **Table S3**.

Chemoproteomics with combinatorial databases improves coverage of acquired cysteines and proximal variants. We next set out to apply our validated search scheme

for chemoproteogenomic variant detection (**Figure 4A**). Inspired by the recent report⁹⁵ of combinatorial databases to improve detection of proximal SAAVs—we expect such variants to be prevalent in heterogeneous cell populations, such as a mismatch repair deficient tumor cell line—we established an algorithm (**Figure S27**) to generate all combinations of SAAVs derived from both RNA/WE-seq data within 30 amino acids flanking the variant site. These combinations were then converted into a peptide FASTA database containing two tryptic sites flanking each variant site (**Figure 4B**). On average, >4,500 total multi-variant peptide sequences were generated per cell line. Our approach differs from most prior custom database generators, which offer ‘Single-Each’^{47,92,96,97} or ‘All-in-One’ outputs^{98,99} for the former, all protein sequences harbor one SAAV each; for the latter, each protein harbors all SAAV detected. While establishing our combinatorial databases, we observed that a small number of highly polymorphic genes (**Table S4**) markedly increased database size—exemplifying this increased complexity, upwards of 1 billion combinations ($2^n - 1$) are possible for protein sequences with 30 or more SAAVs. To determine the practical limit for the number of SAAVs/protein, we performed test searches where we limited the numbers of variants to combine (**Table S4**). We find that nearly all variants are retained with databases that include combinations for proteins with up to 25 variants (**Table S4**). For the small set of highly polymorphic protein sequences (e.g. HLA, MUC, and OBSCN, (**Table S4**), Single-Each sequences were searched (**Figure S27**).

Next, for all 11 sequenced cell lines (**Table S2**), we prepared and acquired a set of high coverage cysteine chemoproteomics datasets (**Figure 4A**). In aggregate, 32,638 total canonical cysteines were identified on 7,233 total proteins, with 9,349 cysteines

unique to individual cell lines and 25,223 shared across the entire dataset (**Figure S25, Table S4**). 2,318 cysteines on 1,406 total proteins had not previously been reported in the CysDB database²⁰ (**Figure S26**). 2-stage MSFragger search using our sample specific combinatorial databases identified a total of 59 gained cysteines and 302 SAAVs located proximal to 343 reference cysteines (**Figure 4C, Table S4**). 74 canonical sequence cysteines located proximal to variants and 60 acquired cysteines had not been previously reported in CysDB (**Figure 4D**)²⁰. Notable examples of acquired cysteine variants not reported in CysDB include acquired cysteines KRAS G12C and PRKDC R2899C. Consistent with the aforementioned genomic data findings, we observe arginine as the most frequently lost out of detected Cys-proximal SAAVs (**Figure 4E**). We detect 15 total cysteines in peptides that harbor gain/loss of arginine that were previously too long or too short to be identified (**Figure 4F, Table S4**). For the cysteine protease cathepsin B (CTSB), we identify Cys207 in HCT-15 cells which was not identified in CysDB—a K209E mutation that creates a longer tryptic peptide sequence compared to reference sequence ('CSK' to 'CSEICEPGYSPTYKQDK'). In the well-studied Jurkat proteome, we detect stromal cell derived factor 2 SDF2, Cys88, which is also not reported in CysDB, is found in a peptide harboring a proximal R93Q mutation that creates a longer, detectable peptide sequence ('CGQPIR' to 'CGQPIQLTHVNTGR'). Showcasing the utility of the combinatorial exome and RNA-seq SAAV databases, we identify six multi variant-containing peptides (**Table S4**). One noteworthy example is the peptide L86P/F92C peptide from the mitochondrial enzyme HADH, which catalyzes beta-oxidation of fatty acyl-CoAs—two variants, one from RNA-seq and one from exome-seq were detected in this peptide. For the I105V, A114V peptide from enzyme GSTP1, the I105V variants were

flagged as bad quality reads from RNA-seq data but passed filters from the exome-seq data (**Table S4**). Of these combination variants, two are exome-seq only derived variants that span exon boundaries.

Chemoproteomic identified variants are in diverse functional sites across protein families. We next asked whether the chemoproteogenomic-identified SAAVs might be of functional significance. By stratifying the the CADD scores of identified SAAVs, we find that the enrichment of high CADD score missense variants in the dMMR rare variant subset was maintained for SAAVs identified by chemoproteogenomics, including for gain-of-cysteine SAAVs (**Figure S28, S29**).

As CADD scores only provide a prediction of deleteriousness, we also asked whether any of the identified variants are located in Census genes or have been reported in Clinvar. We identify 77 variants previously reported in ClinVar (**Table S4**), with nearly all annotated as benign. A total of 16 mutations and 7 putative driver mutations (dN/dS p-values) were identified in Census genes. One prevalent driver was KRAS G12C, which was identified in several of the cell lines known to harbor this variant as a driver mutation (MIA-PACA-2 and H358 but not H2122). As KRAS expression is known to vary across cell lines⁷¹, this data suggests both H358 and MIA-PACA-2 cell lines are suitable for chemoproteogenomic target engagement analysis of G12C-directed compounds. However, as a cautionary example in mapping peptides, we identify several SAAV-peptides that match to multiple protein sequences, including sequences in human leukocyte antigens (HLA) and POTE ankyrin domain family proteins (**Figure 4G**). Most notably, the RHOT2 R425C mitochondrial GTPase peptides in H358 cells have exact

sequence similarity to KRAS G12C peptides; these half-tryptic peptides are also identified in H1437 cells that do not harbor the KRAS G12C variant.

Chemoproteogenomics failed to capture several key Census gene SAAVs that we detected on the genomic level (e.g. SMAD4 (D351H) in CaCo-2, FBXWY (R505C) in Jurkat and CDK6 (R220C) in Molt-4 cells). Several Census gene SAAVs did, however, stand out due both to their high CADD scores and proximity to known pathogenic mutation sites. These variants of interest include MLH1 R385C, RAD17 L557R (proximal Cys551/556), MSN R180C, HIF1A S790N (proximal Cys800) and CTCF R320C, a likely pathogenic position in this protein (CADD score = 29.4) (**Figure 4H, Table S4**).

Exemplifying the utility of the chemoproteogenomics to uncover new variants, we find that 20 of the identified SAAVs have not been previously reported in COSMIC, CCLE or ClinVar (**Table S4**). One variant of unknown significance, not reported in ClinVar, is HMGB1 R110C labeled in the Molt-4 cell line (**Figure 4I**) (CADD score = 24.1). Adjacent Cys106 is a cysteine under highly controlled redox state that is responsible for inactivating the immunostimulatory state of HMGB1¹⁰⁰. We also identify SARS R302H (proximal Cys300; CADD = 32), a mutation in the ATP binding site of serine-tRNA ligase, which is a tRNA ligase involved in negative regulation of VEGFA expression¹⁰¹.

Given the comparatively limited set of variants at or proximal to known damaging sites, we next broadened our analysis to include SAAVs at or proximal to UniProtKB annotated active sites (AS) and binding sites (BS) (**Figure 4J**). We find that 27 SAAVs are located within the permissive range of 10 amino acids of a known functional residue, including 4 active sites and 24 binding sites. Specific examples of high value SAAVs include tRNA synthetase EPRS R1152 (proximal Cys1148; CADD = 33), a mutation

known to cause complete loss of tRNA glutamate-proline ligase activity¹⁰². Interestingly, EPRS has mTORC-mediated roles in regulating fat metabolism¹⁰³. We also capture a variant proximal to the active site of BLM hydrolase I75T (proximal Cys73,78; CADD = 27.6), a cysteine protease responsible for BLM anti-tumor drug resistance¹⁰⁴. More broadly, analysis of SAAV location by protein domains, reveals no marked bias for variants located in specific domain types, with the ubiquitous P-loop NTPase domain as the most SAAV-rich domain (**Figure S30, Table S4**).

As cysteines play critical roles in protein structure via disulfide bond formation together with additional cysteine oxidative modifications¹⁰⁵, we asked whether identified loss of cysteine variants (10 in total) were annotated as involved in disulfides. Likely due to the comparatively small number of loss-of-cys variants, none were observed with disulfide annotations. To further pinpoint whether any variants are sensitive to oxidative modification, we subjected our previously reported Jurkat cell redox chemoproteomics datasets to reanalysis¹⁰⁶. In total, our reanalysis quantified 7 acquired cysteines and 54 variants proximal to acquired cysteines. For nearly all of the cysteines quantified both in our reference database searches and now also identified with proximal variants, we observed a high concordance between variant- and reference sequence oxidation ($R^2=0.77$). One notable exception was the Mitochondrial-processing peptidase enzyme (PMPCA) Cys225, for which markedly different cysteine oxidation states were measured for the reference peptide Cys (~3% oxidation) and variant peptide Cys (~88% oxidation) (**Figure 4K**). These data provide evidence that the proximal P226S mutation profoundly impacts Cys225 sensitivity to oxidative modifiers.

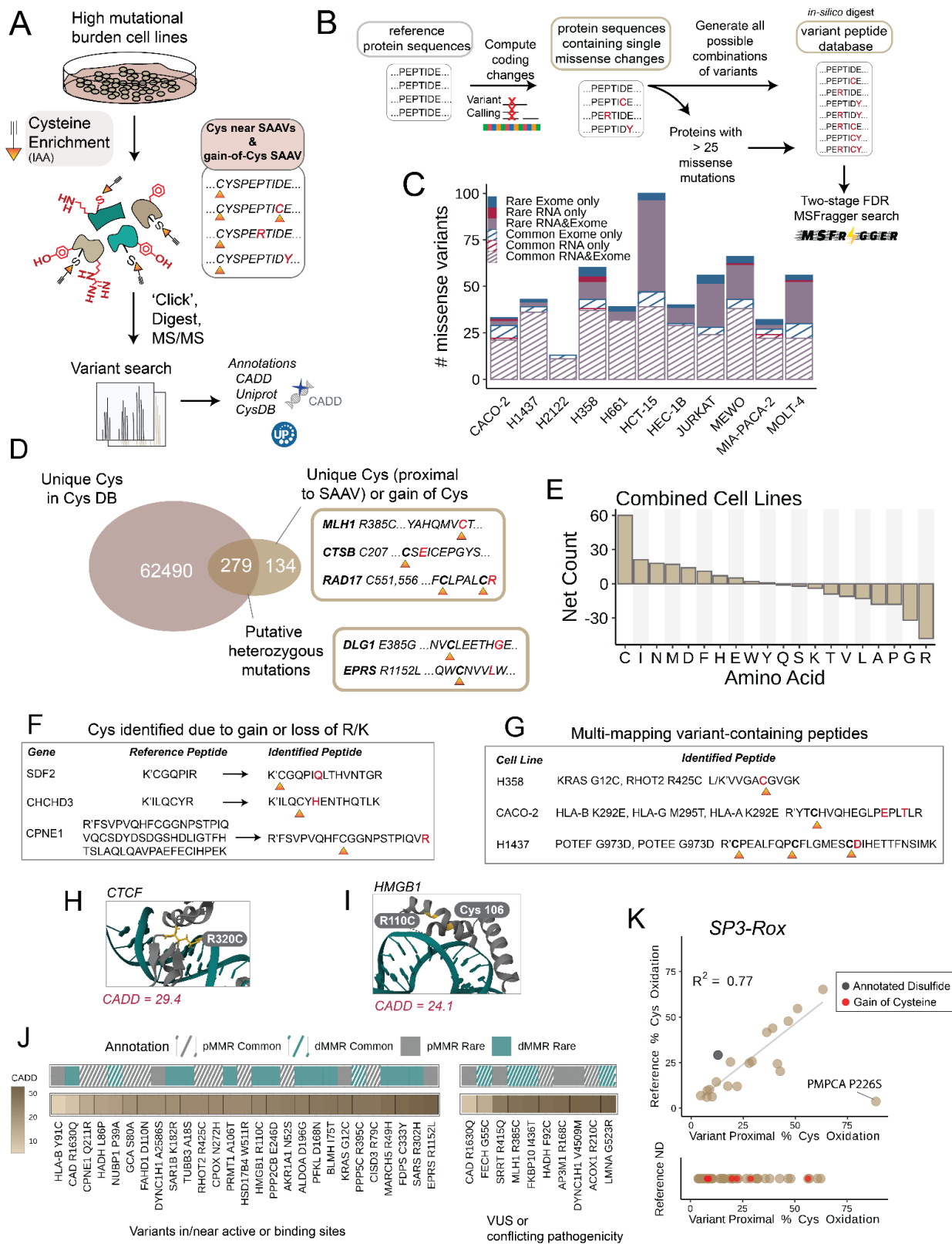


Figure 4. Variant peptide identification on tumor cell lines A) Cell lysates were labeled with pan-reactive iodoacetamide alkyne (IAA) followed by 'click' conjugation onto biotin azide enrichment. Samples were

prepared and acquired using our SP3-FAIMS chemoproteomic platform^{22,23,107} using single pot solid phase sample preparation (SP3)¹⁰⁸ sample cleanup, neutravidin enrichment, sequence specific proteolysis, and LC-MS/MS analysis with field asymmetric ion mobility (FAIMS) device¹⁰⁹. Experimental spectra are searched using the custom fasta for variant identification. Sample set includes both reanalysis of previously reported datasets from Yan et al. (Molt-4, Jurkat, Hec-1B, HCT-15, H661, and H2122 cell line) with newly acquired datasets (H1437, H358, Caco-2, Mia-PaCa-2 and MeWo cell lines). B) Non-synonymous changes are incorporated into reference protein sequences and combinations of variants are generated for proteins with less than 25 variant sites to make customized fasta databases. Details in methods. C) Total numbers of unique missense variants identified from either RNA-seq or WE-seq or both after using 2-stage MSFragger search and philosopher validation from duplicate datasets; stripe vs solid denotes common and rare variants, red text indicate dMMR cell lines Indicated is sequencing source and type of variant. D) Overlap of identified cysteines from variant searches with cysteines in CysDB database²⁰. E) Net amino acid changes for all cell lines combined F) Example of cysteines identified from loss of R/K peptides G) Examples of multi-mapping variant sites. H) Crystal structure of CTCF indicating detected Cys320 (*yellow*) and DNA-binding site (PDB: 5T0U). I) Crystal structure of HMGB1 indicating detected Cys110 and nearby Cys106 (*yellow*) (PDB: 6CIL). J) Variants identified in or near active and binding sites with CADD score, common/rare, cell line dMMR/pMMR annotations. K) Re-analysis of SP3-Rox¹⁰⁶ oxidation state data in Jurkat cells. Data found in **Table S4**.

Assessing how differential expression impacts chemoproteogenomic detection.

Our comparatively modest coverage of SAAVS achieved by chemoproteogenomics (particularly when compared to our genomics datasets) is on par with the coverage reported by most prior proteogenomics studies^{41,43,53}. A notable exception is the recent study by Coon and colleagues that implemented ultra-deep fractionation to achieve more global coverage of variants⁴⁴. Inspired by this work, we next sought to ask whether chemoproteogenomics, with its built in enrichment step, would enable sampling of variants not detectable by fractionation methods (**Figure 5A**). We subjected lysates from HCT-15 and Molt-4 cells, which were chosen based on high rare missense burden, to tryptic digest, off-line high pH fractionation, and LC-MS/MS analysis. In aggregate across both cell lines, we identified 8,435 proteins and 149,006 peptides, including 1,069 unique SAAVs found in 1,352 total peptides using our 2-stage MSFragger search (**Figure 5B,S31,Table S5**). 26 peptides were identified that contained multiple variants, including peptides that would only be detected by our combinatorial databases (**Figure 4B**) as well

as those readily detected by combined ‘Single-Each’ and ‘All-in-One’ database searches (**Table S5**).

Comparison of this unenriched dataset to the chemoproteogenomic dataset for the matched HCT-15 and Molt-4 proteomes (145 total SAAVs identified by chemoproteogenomics for these two cell lines) revealed 70 SAAVs, including eight acquired cysteines, uniquely identified with chemoproteogenomics (**Table S4-S5**), (**Figure 5C**). Despite the lower numbers of total SAAVs in the chemoproteogenomics datasets, we find that chemoproteomic enrichment afforded a ~5-fold boost in the relative fraction of acquired cysteines captured (**Figure 5D**). Further stratification of the net detected amino acid changes (**Figure S32-S33**) revealed that, again, cysteine was a top gainer and arginine was the most lost amino acid for both enriched and unenriched datasets.

We next asked whether protein or RNA abundance might rationalize the differences in SAAV coverage for each method. Comparison of normalized transcript counts for SAAV-matched genes identified either by chemoproteogenomics or in our bulk proteomic dataset, for HCT-15 cells analysis revealed no significant difference between measured transcript abundance between the sets (**Figure 5E, Table S5**). A notable subset of SAAVs (3,262 total, including PIK3CA E545K, TP53 S241F, SMARCA4 R885C TCGA hotspot mutations) with low abundance transcripts (less than 4000 normalized counts) were not detected in either the chemoproteogenomics or bulk proteogenomics. Providing further evidence that lower transcript abundance decreases the likelihood of detection, we find that an even more sizable fraction of cysteines found in reference

protein sequences matched with low abundance genes are not detected, both for high-pH fractionated samples and chemoproteomics enriched samples (**Figure S34**).

Given the likely disconnect between transcript abundance and protein abundance^{110–112} for some SAAVs analyzed, we also extended these analyses to measures of protein abundance. Using label-free quantification (LFQ) analysis, we find that for proteins with proteomic-detectable SAAV peptides, the quantified protein intensities were significantly higher when compared to proteins for which the corresponding variants were only detected via genomic analysis. No difference was observed between the bulk fractionated samples and the chemoproteogenomic samples (**Figure 5F, Table S5**).

As both the transcript and protein abundance analyses do not delineate reference from variant-specific transcript/protein sequences, we also compared the variant allele frequencies (VAF) for SAAVs detected by each method. We find that high-pH variant allele frequencies (VAF) were significantly higher than the chemoproteogenomic detected SAAVs, which were comparable to the aggregate bulk RNA-seq VAFs (**Figure 5G, Table S5**). This enrichment for lower VAF for the chemoproteogenomic detected SAAVs extended to the acquired cysteine subset (**Figure S34**).

Given that cysteine chemoproteomics requires peptide derivatization, with a comparatively large (463 Da) biotin modification, we postulated that some differences in coverage might be ascribed to behavior of peptides during sample acquisition. Comparing the properties of the SAAV peptides detected by chemoproteogenomics versus proteogenomics we observed a more restricted charge state distribution for cysteine-enriched samples and no appreciable differences in the amino acid content beyond

enrichment for cysteine (**Figure S35**). While we did not observe differences in the peptide lengths in our comparison of between the chemoproteomic-enriched and high pH detected SAAV peptides, a marked significant increase in SAAV peptide length (average 5AA longer) was observed compared to reference peptides in both datasets (**Figure 5H**). This increased peptide length is consistent with the ubiquity of loss-of-arginine SAAVs in both datasets, which are favored in the longer length peptides (**Figure S36**).

Protein families analysis revealed slight differences between the two datasets with enzymes making up a larger fraction of cys-enriched detected variant proteins. Significantly higher CADD scores were also observed for enrichment data (**Figure S37**). Notable high-CADD score variants identified only from enrichment include lysine demethylase KDM3B D1444Y, RNA polymerase POLRMT R805C, glycoprotein transporter LMAN2 R218C and Serine/threonine-protein phosphatase PP1-alpha catalytic subunit PPP1CA D203N (**Figure 5C**). Addition of the bulk proteomic analysis yielded coverage of 85 notable variants belonging to Census genes, including BRD4 E451G and KRAS G13D, and 26 rare and common variants of uncertain significance in ClinVar, including rare gain-of-cysteines ubiquitin hydrolase USP8 Y1040C and LMNA R298C (**Figure 5I, Table S5**).

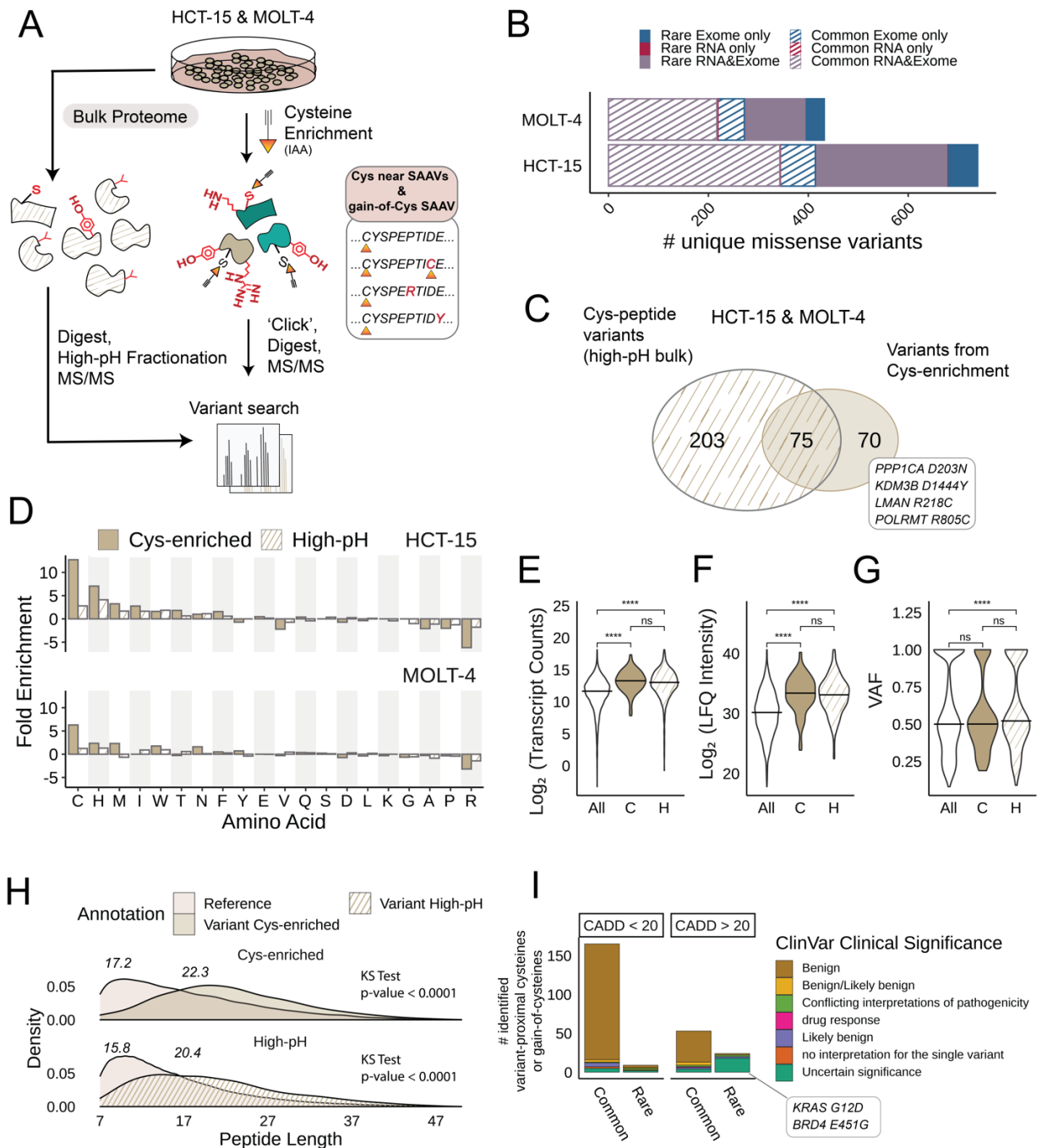


Figure 5. Comparison of variants identified from cysteine enrichment and bulk proteomics A) Workflow for high-pH fractionation of lysates. Cell lysates are treated with DTT and iodoacetamide followed by digestion, high-pH fractionation, and LC-MS/MS analysis. Triplicate high-pH sets for HCT-15 and Molt-4 cells were used. B) Total numbers of unique missense variants identified from either RNA-seq or WE-seq or both after using 2-stage MSFragger search of high-pH datasets. C) Overlap of cysteine-containing peptide variants identified from bulk fractionation and cysteine enrichment datasets. D) Fold enrichment of amino acids as a ratio of the net amino acid frequency (gain minus loss) to the amino acid frequency in all

missense-containing proteins detected in high-pH and cys-enriched datasets. E) DE-seq normalized transcript counts for all RNA variants 'All', variants detected from cys-enrichment 'C', and variants detected from high-pH fractionation 'H' in HCT-15 cells; bar indicates mean value. F) Label free quantitation (LFQ) intensities for proteins matched to all RNA variants 'All', variants detected from cys-enrichment 'C', and variants detected from high-pH fractionation 'H' in HCT-15 cells; bar indicates mean value. G) Variant allele frequencies (VAF) (total reads/total coverage per site) for RNA-seq variants called in HCT-15 and Molt-4 cells. E-G; bar indicates median, statistical significance was calculated using Kolmogorov-Smirnov test, **** $p < 0.0001$, ns $p > 0.05$. H) Peptide lengths of reference and variant peptides identified in dataset types. I) High-pH detected variants stratified by CADD score and ClinVar clinical significance. Data found in **Table S5**.

Chemoproteogenomics enables ligandability screening. As demonstrated by our previous studies, cysteine chemoproteomics platforms are capable of pinpointing small-molecule targetable cysteine residues^{21,22,26,29}. Therefore, we next paired our 2-stage search method with cysteine-reactive small molecule ligandability analysis to establish a chemoproteogenomic small molecule screening platform (**Figure 6A**). We first opted to use the widely employed scout fragment **KB02**²⁹ (**Figure 6B**) to compare the ligandable variant proteomes for three high variant burden dMMR cell lines (HCT-15, Jurkat, and Molt-4). For **KB02** treated samples, we identified 210 total variants. The high concordance for ratios detected for variant peptides with multiple alleles provides evidence of the robustness of our platform and hints that most cysteine proximal variants do not substantially alter cysteine ligandability (**Figure 6C**).

We next subjected the HCT-15 proteome to more in depth analysis using a small panel of custom electrophilic fragments (**Figure 6B**). We observed 27 total liganded variant peptides in 27 proteins in the HCT-15 proteome, which are labeled by one or more compounds (**Figure 6C**). As with the **KB02** cell line comparison, nearly all multi-allelic peptides showed comparable ratios (**Figure 6E**). Nucleotide analogue **SO-105** was observed to be more promiscuously reactive (**Figure 6F**) when compared to the less elaborate fragments.

In aggregate across all ligandability datasets, we identified 259 total variants found in 232 total proteins (**Figure 6D**). Of these variants, 57 were acquired cysteines, in 55 proteins; 22 were ligandable ($\text{Log}_2(\text{HL})$ ratio > 2), variant-proximal cysteines and 10 were ligandable gain-of-cysteines (**Figure 6D**). Notable liganded sites we identify include Cullin-associated NEDD8-dissociated protein 1 (CAND1) G1069C—a site which mutated in the Arabidopsis ortholog reduces auxin response¹¹³ and Tubulin beta 6 (TUBB6) G71C (**Figure 6G**). Some sites with differing reference and variant ratios include EPRS P1482T—the mutated proline nearby Cys 1480 may be requisite for labeling by electrophilic fragments. We also identify 3 ligandable variants of uncertain significance or conflicting pathogenicity that we show may be modulated for study with small molecules and could act as potential starting points for biological analyses (**Figure 6C**). As multi-allelic acquired cysteine sites cannot be captured sans cysteine, no analogous ratio comparison could be performed for the 6 total quantified acquired cysteines (**Figure 6G**).

To understand functionality of the ligandable variant sites in 3D protein space, we analyzed active site and binding sites within 10 angstrom distance of the ligandable Cys residues and Cys-proximal variant sites (**Table S6**). We find three ligandable cysteines near or in active/binding sites including previously identified HMGB1 Cys106 (R110C) (**Figure 4I**), as well as Aldolase A ALDOA Cys178 (G196G) and HLA-B/C Cys125 (V127L/S123Y). Intriguingly HLA-B/C Cys125 (C101 post signal peptide cleavage), near peptide binding region sites Y183 is liganded by **KB02** in HCT-15 cells which harbor HLA-B*08:01 and HLA-B*35:01 (**Figure 6F**). This conserved cysteine plays important roles in HLA structure¹¹⁴. Ligandability of this site is unexpected as this site is known to be disulfided with C188 in cell surface HLA¹¹⁵; however, we find in our sequencing that HCT-

15 cells harbor truncated beta-2-microglobulin (β 2m) protein (B2M Y30*) (**Table S2**). β 2m is known to stabilize this specific disulfide^{115,116}, facilitating protein folding and translocation to the cell surface^{117–119}. In HLA-B27 allelic variants, Cys125 is known to be exposed without β 2m¹²⁰.

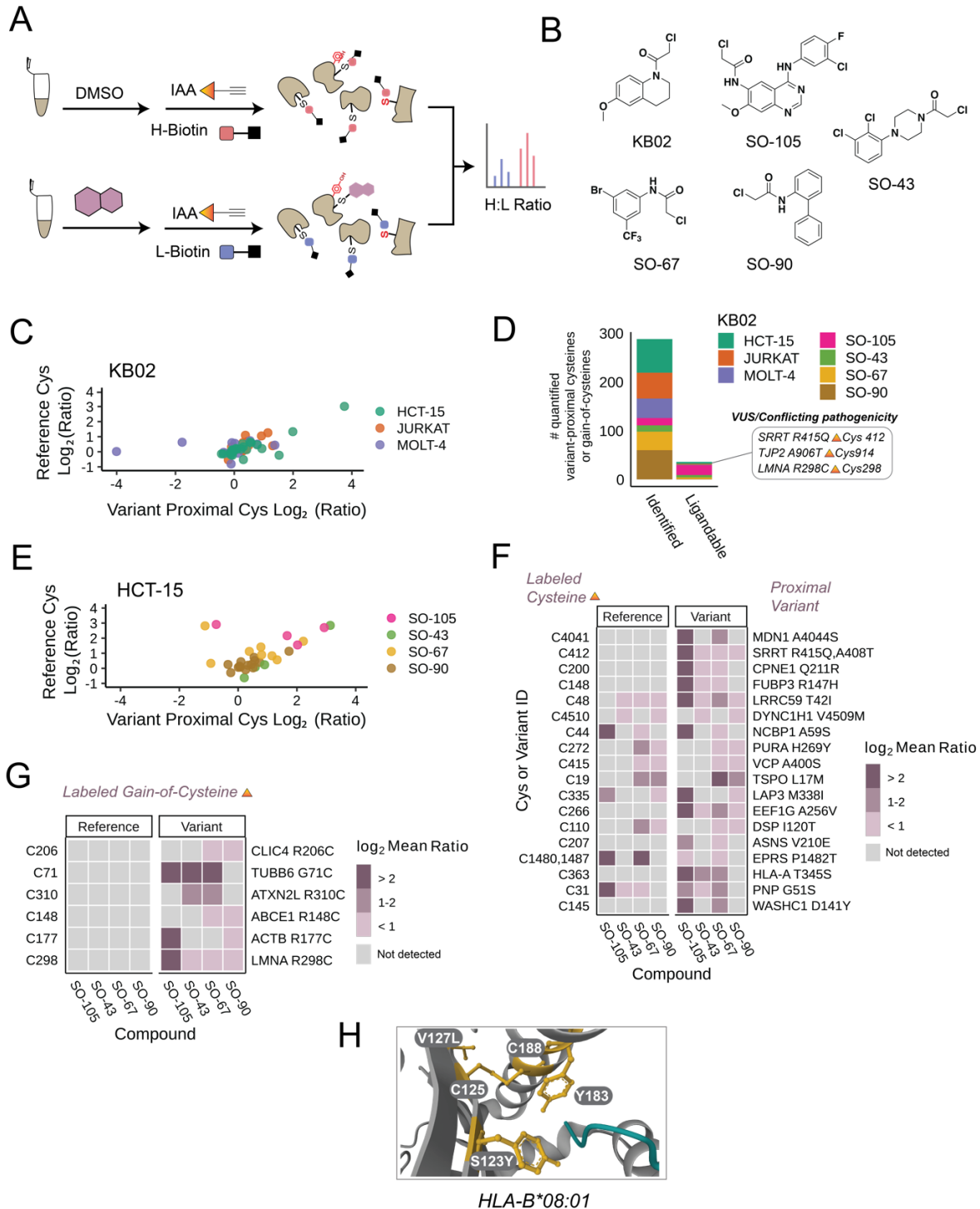


Figure 6. Assessing ligandability of variant proximal cysteines and gain-of-cysteines. A) Schematic of activity-based screening of Cys reactive compounds; cell lysates are labeled with compound or DMSO followed by chase with IAA and 'click' conjugation to heavy or light biotin click conjugation to our isotopically differentiated heavy and light biotin-azide reagents, tryptic digest, LC-MS/MS acquisition, and MSFragger analysis. B) Chloroacetamide compound library. C) Total quantified variants and total ligandable variants

(Log₂ Ratio > 2) identified stratified by cell line (KB02 data) or compound (HCT-15 cell line). D) Correlation of high-confidence variant containing and reference cysteine ratio values from KB02 data. E) Correlation of high-confidence variant containing and reference cysteine ratio values from SO compound data. F) Log₂ heavy to light ratio values for variant containing and reference cysteine peptides. G) Subset of gain of cysteine peptide variant log₂ ratios. H) Crystal structure of HLA-B*08:01 protein liganded Cys125, disulfide Cys188, and binding site residue Y183 as well as variant sites V127 and S123 (PDB: 3X13). Data provided in **Table S6**.

Expanding HLA cysteine peptide coverage and gel-based ABPP of HLA covalent labeling. Major Histocompatibility Complex (MHC) Class I molecules (known as HLA molecules in humans) present intracellularly derived protein fragments, either self-derived or from pathogens in the context of cross-presentation, on the cell surface for recognition by T cells and subsequent immune response; noncovalent assembly of a polymorphic heavy chain with a light chain (β 2m) and peptide occurs in the endoplasmic reticulum (ER) followed by translocation via the Golgi to the cell surface¹²¹. Recent reports of allele-specific HLA-binding compounds, most notably abacivir HIV drug¹²², together with efforts to develop covalent modulators of MHC Class I and II complexes^{123–125} prompted us to assess the impact of chemoproteogenomics on achieving improved coverage of highly polymorphic genes (**Figure 7A**). 15,000 HLA alleles have been reported in the human population¹²⁶. Exemplifying this impact on proteomic sequence coverage, our panel of cell lines alone harbor >25 HLA-A, B and C alleles (**Table S2**), while most protein reference databases only contain one copy of each MHC Class I and Class II molecule.

Through search of sample-specific databases of both chemoproteomics and high pH fractionated samples, we achieved ~50% more coverage of HLA-A sequence in comparison to reference searches (**Figure 7B and Figure S39**). A key finding of our analysis was detection of HLA-B Y91C (C67 post signal peptide cleavage), which lies in the extracellular peptide binding pocket of HLA-B and was identified as IAA-labeled in

MeWo cells (**Figure 4J**). The MeWo cell line HLA alleles (HLA-B*14:02 and HLA-B*38:01) both harbor this comparatively rare Cys (**Figure 7C**). Notably this cysteine is also a key feature of the pathogenic ankylosing spondylitis associated allele HLA-B*27^{127,128}. To test whether this cysteine was amenable to gel-based ABPP analysis and to determine whether this IAA labeling extends to HLA-B*27:05, we co-expressed c-terminal FLAG tagged HLA-B*38:01, HLA-B*27:05, HLA-B*38:01 C91S, and HLA-B*27:05 C91S with beta-2-microglobulin (β 2m) and subjected cells to in situ IAA labeling followed by lysis, FLAG immunoprecipitation to enhance the detectability of the HLA cysteine, and click conjugation to rhodamine azide (**Figure 7D**). Gratifyingly, we observed a Cys67-specific rhodamine signal (**Figure 7E**), showcasing the utility of gel-based ABPP in visualizing HLA small molecule interactions. Notably IAA labeling was also observed for HLA-B27:05, although the presence of a strong co-migrating band in the HLA-B27:05 C67S immunoprecipitated sample complicates interpretation of the specificity of this labeling to Cys67. We were unable to observe comparable signal in lysate-based labeling studies, supporting enhanced accessibility of this cysteine to cell-based labeling (**Figure S40**).

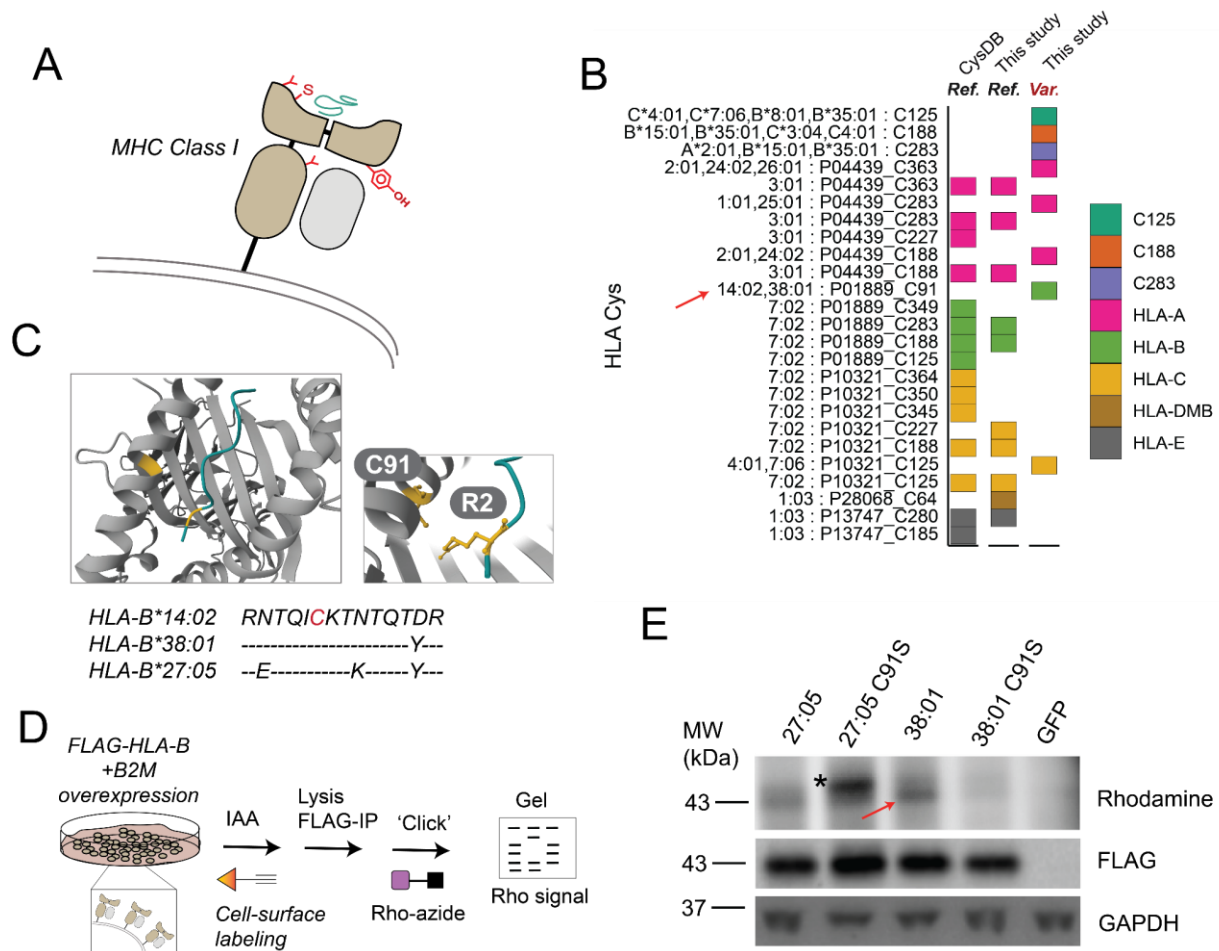


Figure 7. Expanding HLA cysteine peptide coverage and gel-based ABPP of HLA covalent labeling.

A) Schematic of highly variable HLA binding pocket containing cysteine with bound peptide. B) Coverage of HLA cysteines from this study and in CysDB; color indicates HLA type or multi-mapped cysteines. C) Crystal structure of HLA-B 14:02 (PDB: 3BXN) with highlighted Cys67 and Arg P2 position of bound peptide; alignments of Cys91 regions of three HLA-B alleles. D) Workflow to visualize HLA cysteine labeling; first cells were harvested and treated with IAA followed by lysis, FLAG immunoprecipitation, and click onto rhodamine-azide. E) Cys-dependent cell surface labeling of HLA-B alleles with IAA, band indicated with red arrow and non-specific band represented with asterisk (representative of 2 two biological replicates). Data provided in **Table S7**.

FragPipe graphical user interface with improved 2-stage MSFragger search and FDR estimation. Motivated by the multi-faceted uses of the 2-stage search pipeline,

including those reported here and future envisioned applications, we also sought to facilitate the utilization of the 2-stage search strategy by the scientific community. Therefore, we enhanced FragPipe by establishing semi-automated execution of these searches while also providing an option to run MSBooster and Percolator (instead of PeptideProphet) to further improve the sensitivity of identification of variant peptides **(Figure 8A)**.

In the first stage pass, with the "write sub mzML" option enabled, FragPipe utilizes MSFragger^{85,129} for mass calibration, search parameter optimization, and database searching. Following this, FragPipe applies MSBooster¹³⁰ to compute the deep-learning scores¹³⁰, Percolator¹³¹ for PSM rescoring, ProteinProphet¹³² for protein inference, and Philosopher for FDR filtering. Subsequently, FragPipe generates new mzML files, which include the scans that did not pass the FDR filtering (default is 1%) and those with a probability higher than a predefined threshold (default is 0).

In the second search, as the mass spectral files have already been calibrated and only scans that remained unidentified in the first search have been retained, the mass calibration should be disabled. Moreover, Percolator modeling might fail in the second pass due to a lack of sufficient number of high-scoring PSMs. Therefore, FragPipe lets Percolator reuse the model from the initial pass. FragPipe then generates a new workflow file containing optimized parameters, and a new manifest file with the new (subset) mzML files specified for the second-pass search. The user is merely required to load these two files without needing any further adjustments.

Using the new GUI features, we observe comparable coverage for both the command-line and automated GUI implementations of the 2-stage search with a slight

increase in numbers of identifications observed for datasets processed with MSBooster and Percolator (Figure S41, Table S8). The ratio differences between variant and reference Cys peptide are comparable (Figure 8B).

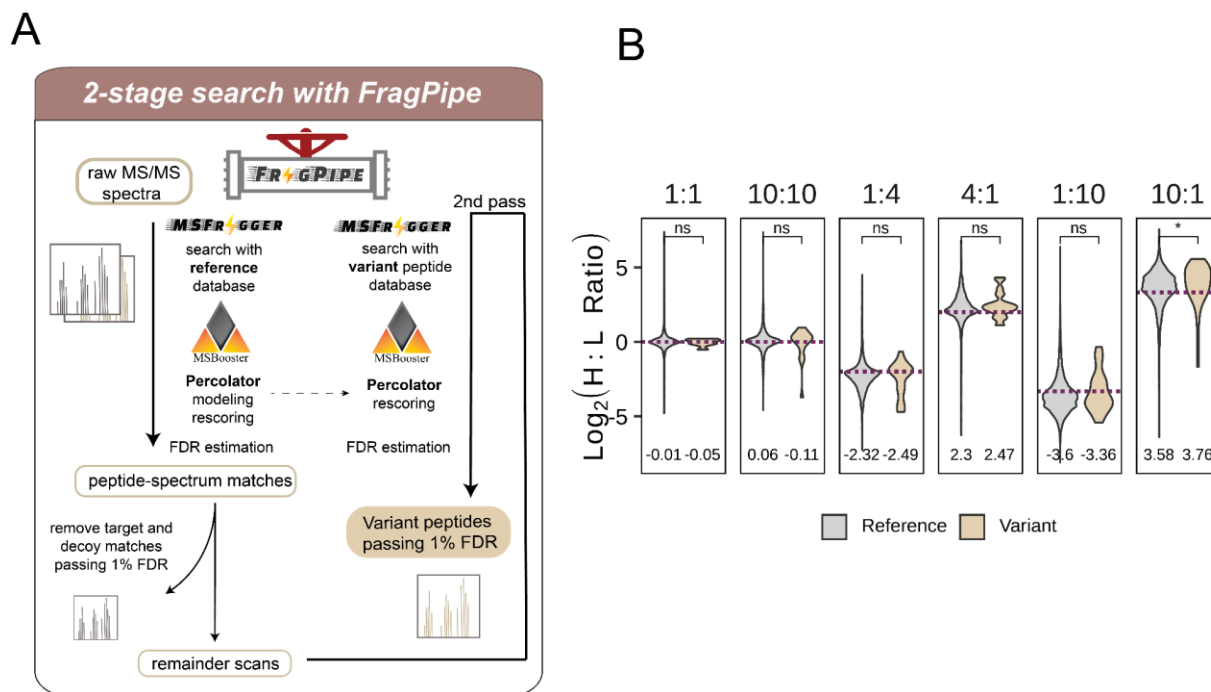


Figure 8. 2-stage search implemented into FragPipe GUI with Percolator rescoring A) 2-stage search incorporation into FragPipe GUI workflow. B) Heavy to light ratios (H:L) from triplicate datasets comparing identifications from reference and variant searches; mean ratio value indicated, *dashed lines* indicate ground-truth log₂ ratio, statistical significance was calculated using Mann-Whitney U test, * $p < 0.05$, ** $p < 0.01$, ns $p > 0.05$. Data provided in **Table S8**.

Discussion

SAAVs are a ubiquitous feature of human proteins, which remain under sampled in established proteomics pipelines. Here, we merged genomics with mass spectrometry-based chemoproteomics to establish chemoproteogenomics as an integrated platform tailored to capture and functionally assess the missense variant cysteinome. Our chemoproteogenomics study is distinguished by a number of features including: (1)

genomic stratification of the predicted pathogenicity of acquired cysteine residues, (2) cell-line paired custom combinatorial search databases, (3) FragPipe enabled 2-stage database search platform ensuring class-specific FDR estimation, and (4) capacity to pinpoint both redox-sensitive and ligandable genetic variants proteome-wide. To facilitate widespread adoption of our approach, including for applications beyond the study of the variant cysteinome, the user-friendly GUI-based FragPipe platform now features a robust semi-automated version of our 2-stage search (**Figure 8**).

To build chemoproteogenomics, we started by analyzing publically available datasets in Clinvar, COSMIC, and dbSNP, which revealed that cysteine acquisition is a ubiquitous feature of human genetic variation, which predominates in the context of DNA damage repair responses. The instability of CpG motifs is a key driver of bulk cysteine acquisition, which occurs largely hand-in-hand with bulk arginine depletion, across both cancer genomes and healthy genomes and rare and common variants. Many colon cancer cell lines and other MSI high cell lines are particularly enriched for cysteine acquisition—however, nearly all of the acquired residues in these lines are not driver mutations, which complicates their use as models for assessing the potentially druggability of variants with established clinical connections and highlights the value of future efforts to analyze additional missense variant rich cell lines and perform CRISPR-Cas9 base editing to engineer variants of interest into endogenous loci^{35,133–136}.

Armed with a set of variant rich cell lines, we next generated combinatorial SAAV-peptide databases for cell-line specific SAAVs as identified in cell-line matched whole exome and transcriptome datasets. In total, across 11 cell lines sequenced, we identified 1,453 missense variants, of which 116 led to gain-of-cysteine. Looking towards future

iterations of chemoproteogenomics, we expect that the use of tumor-normal paired variant calling with tools such as MuTect2¹³⁷ will further decrease the likelihood of false discovery introduced by factors such as cell heterogeneity and low read quality—for cell lines that lack matched normal controls, we expect that the pairing of publically available datasets (e.g. DepMap, <https://depmap.org/>) with custom sequencing data, will prove another useful strategy to further bolster the quality and accessibility of variant-containing databases. Such multi-pronged approaches will likely prove most useful when paired with combinatorial custom databases, such as the peptide-based databases reported here, which were designed to minimize increased search space complexity while also more fully accounting for cell heterogeneity.

By building upon prior reports describing 2-stage database searches for class-specific FDR control⁵³⁻⁵⁵ as a rigorous search strategy that reduces the likelihood of a false positive variant peptide detection, here we deployed a 2-stage search approach in FragPipe, first as a custom command-line workflow and subsequently as a user-friendly semi-automated workflow in the FragPipe GUI. Enabled by our previously reported isotopically enriched heavy- and light-biotin-azide capture reagents⁸⁷, we provide compelling evidence to support the low rates of false discovery of variant peptides using the 2-stage search—spurious false discovery of variant peptides would easily be detected from MS1 precursor ion ratios that deviate from the expected spike-in values (**Figure 3,8**). Our isotopic labeling strategy also enabled the assessment of the ligandability and redox sensitivity of variant peptides. Our discovery of a cysteine in PMPCA that exhibits variant-dependent changes in oxidation provides an intriguing anecdotal example that supports the future utility of chemoproteogenomics in more broadly characterizing the missense

variant redox proteome. Given the critical role that disulfides play in protein structure and folding and the causal roles for cysteine mutations in human disease, for example the NOTCH mutations that cause the neurodegenerative disorder CADASIL¹³⁸, we expect a subset of these lost cysteines could be implicated in altered protein abundance or activity. Through cysteine chemoproteomic capture, we identified ligandable variant-proximal cysteines in Census genes such as RAD17, including one gain-of-cysteine of uncertain significance in LMNA (R298C). Other liganded cysteines proximal to variants of uncertain significance include TJP2 (A906R) and SRRT (R415Q). Demonstrating the utility of our approach, we identified a Cys91 (Cys67) as labeled by IAA both by proteomics and gel-based ABPP. As this cysteine is shared with the pathogenic HLA-B27, it is exciting to speculate about the impact of covalent modification on HLA peptide presentation. Our application of chemoproteogenomics to screening of a focused library of electrophilic compounds, identified 32 ligandable variant-proximal Cys which demonstrates that cysteine ligandability can be assessed proteome-wide in a proteoform-specific manner. Looking beyond our current study, we anticipate multiple high value applications for chemoproteogenomics. Application to immunopeptidomics should uncover additional covalent neoantigen sites, analogous to the recent reports for Gly12Cys KRAS^{124,139}. Pairing of chemoproteogenomics with ultra-deep offline fractionation should further increase coverage and allow delineation of variants that alter protein stability, including the numerous high CADD score acquired cysteines, which we find were underrepresented in our proteomics analysis when compared to genomic identification. Inclusion of genetic variants beyond SAAVs will allow for capture of additional therapeutically relevant targets that result from indels, alternative splicing^{39,140},

translocations, transversions, or even undiscovered open reading frames such as microproteins^{141,142}. Thus chemoproteogenomics is poised to guide discovery of proteoform-directed therapeutics.

Acknowledgments

We thank all members of the Backus lab for helpful suggestions. We thank the UCLA Technology Center for Genomics and Bioinformatics (TCGB). Additionally, we thank Jigar Desai for guidance on NGS data processing, Angela Wei for guidance on Kallisto data processing, and Ian Ford for providing a CuAAC-compatible IP protocol. The results here are in part based upon data generated by the COSMIC-CLP: https://cancer.sanger.ac.uk/cell_lines and TCGA Research Network: <https://www.cancer.gov/tcga>. This study was supported by a Beckman Young Investigator Award (K. M. B.), V Scholar Award V2019-017 (K. M. B.), UCLA Jonsson Comprehensive Cancer Center Seed Grant (K. M. B.), and the National Institutes of Health grants R01-GM094231 and U24-CA271037 (A. I. N.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author Contributions

H. S. D., K. M. B., and A.I.N. conceptualization; H. S. D. formal analysis; H. S. D. visualization; H. S. D. validation; H. S. D., L.M.B., F. Y., K.M.B data curation; H. S. D., S. O., and M. V. investigation; H. S. D., F. Y., and N.U. methodology; H. S. D. and K. M. B writing—original draft; H. S. D., S.O., L.M.B., F.Y., A. I. N., and K. M. B. writing—review and editing; A. I. N. and K. M. B. supervision; A. I. N. and K. M. B. funding acquisition.

Conflicts of Interest

The authors declare no financial or commercial conflict of interest.

Methods

Experimental details and Tables S1-S9 can be found in the Supporting Information.

Chapter 3 Supporting Information:

Multi-omic stratification of the missense variant cysteinome

Supplementary Figures

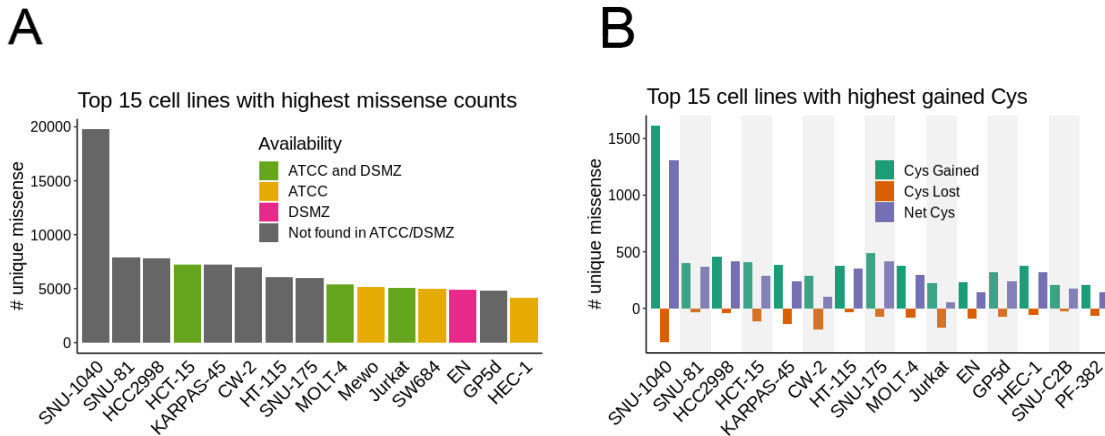


Figure S1. Missense counts in subsets of COSMIC Cell Lines A) Top 15 cell lines with highest missense burden in COSMIC-CLP (COSMIC Cell Lines Project release v96) color indicates availability of high mutational burden cell lines in American Type Culture Collection (ATCC) and German Collection of Microorganisms and Cell Cultures (Deutsche Sammlung von Mikroorganismen und Zellkulturen, DSMZ). B) Top 15 cell lines with highest gained cysteines showing net gain, total gained and total lost.

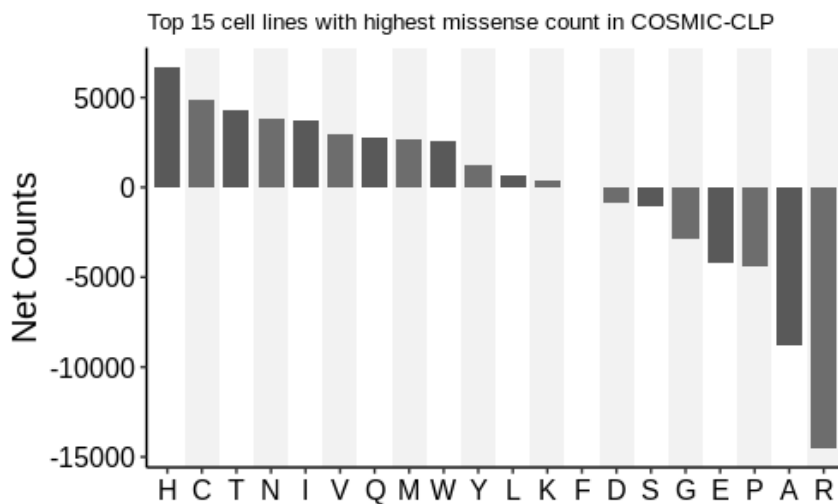


Figure S2. Aggregated net amino acid changes (gained counts - lost counts) in combined top 15 cell lines with highest missense burden in COSMIC-CLP (COSMIC Cell Lines Project release v96).

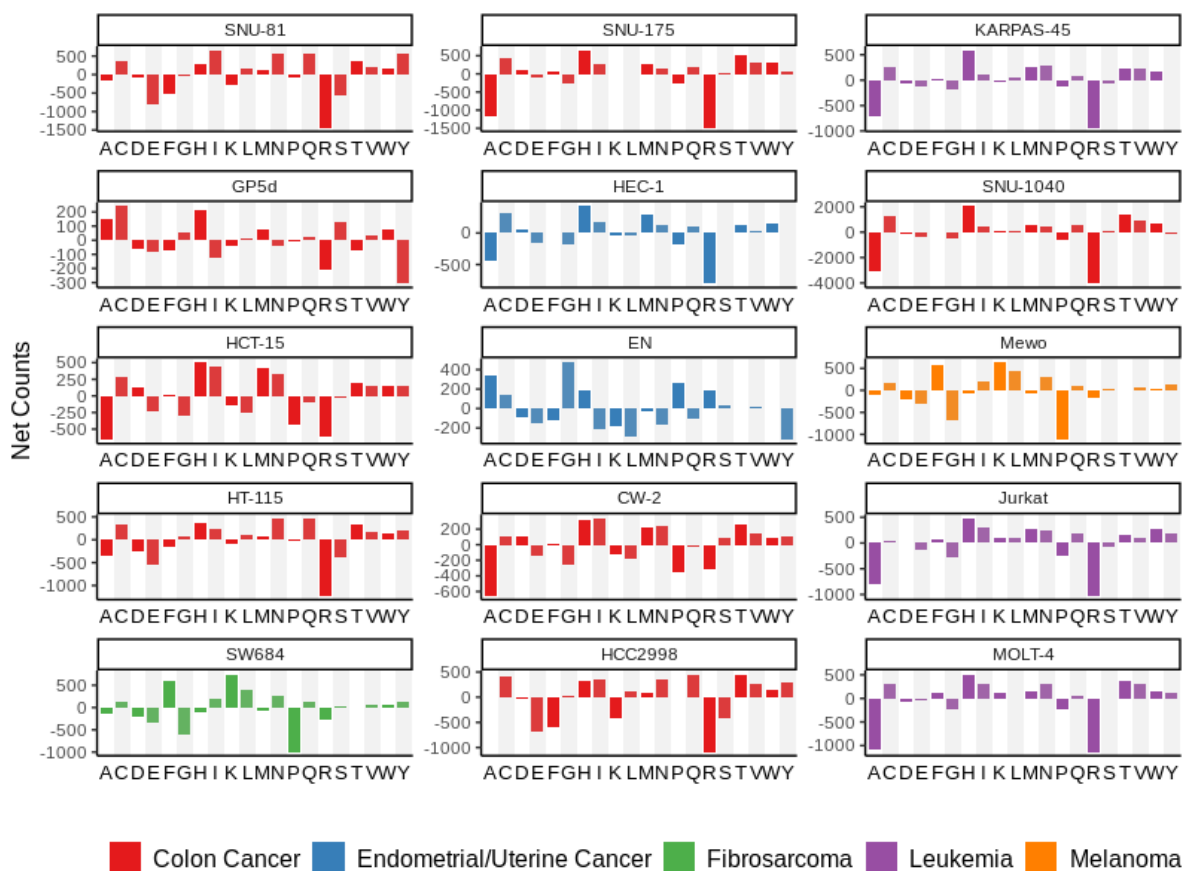


Figure S3. Net single amino acid mutation counts (gained counts - lost counts) in top 15 cell lines with highest missense burden (COSMIC Cell Lines Project release v96). Several cell lines show marked depletion of alanine, for example SNU-175, KARPAS-45, Jurkat (A3 subclone) and MOLT-4 cell lines, which represent two colon cancer cell lines, and two T lymphoblast cell lines, respectively. In contrast, several cell lines showed marked net gain of alanine, including EN, Gp5d, and HCC2998 cells, which are an endometrial/uterine cancer lines and two colon cancer cell lines, respectively.

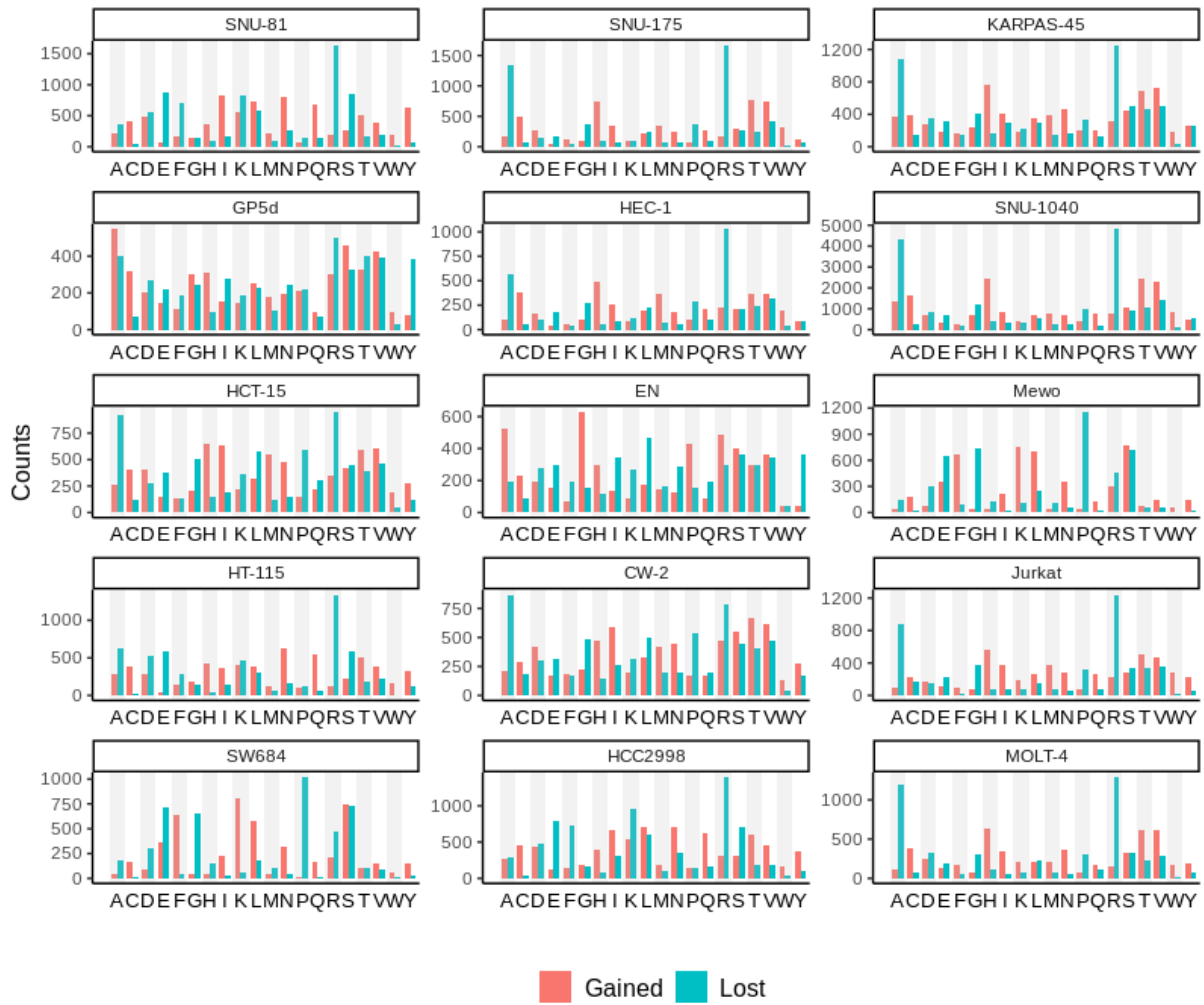


Figure S4. Total gain and loss missense mutation counts in the top 15 cell lines with highest missense burden (COSMIC CLP v96).

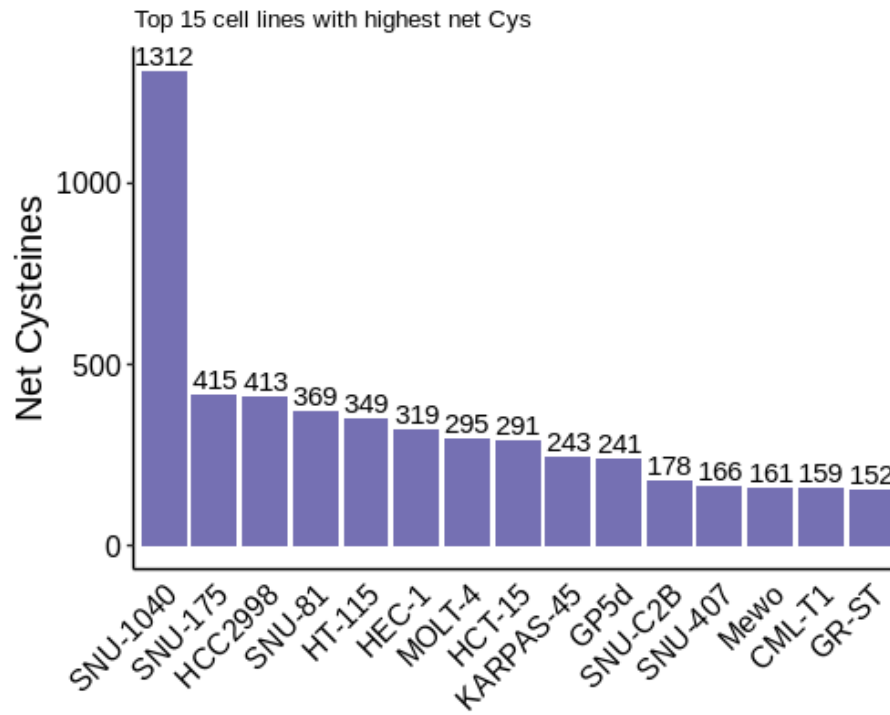


Figure S5. Cell lines with highest net gained cysteines (COSMIC CLP v96).

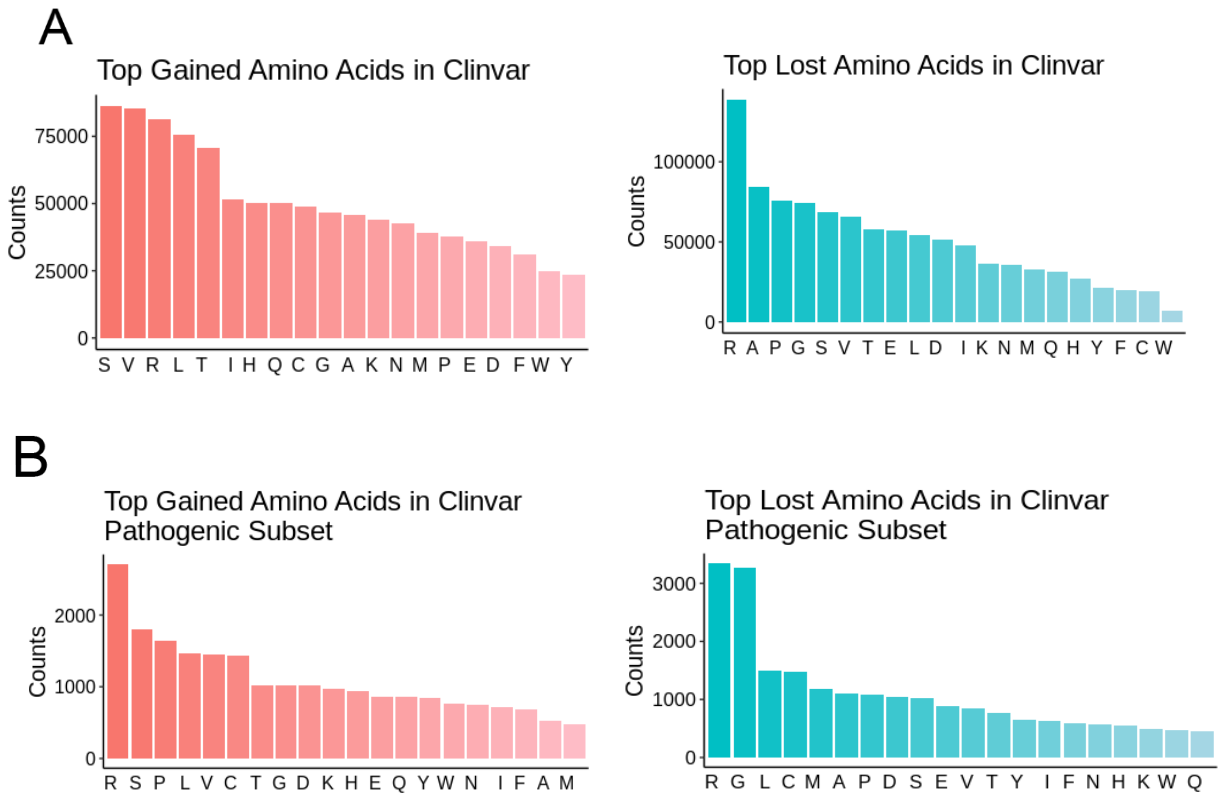


Figure S6. ClinVar amino acid changes A) Total gained and lost amino acids reported in all ClinVar data for unique gene name, protein position and amino acid change B) Total gained and lost amino acids for pathogenic missense variants in ClinVar. Red indicates gain amino acids and blue indicates lost amino acids

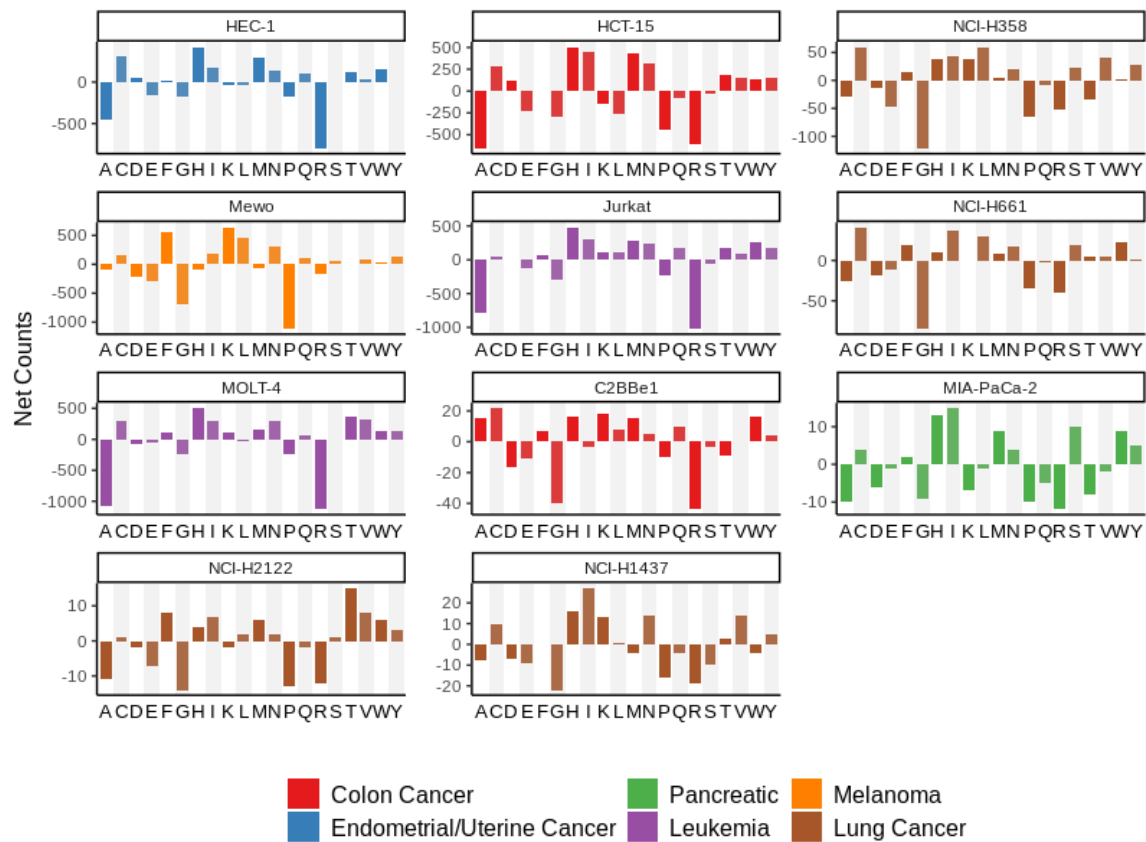


Figure S7. Net single amino acid mutation counts (gained counts - lost counts) in the panel of cell lines in this study (COSMIC CLP v96). C2BBE1 is a CaCo-2 cell line clone in COSMIC-CLP.

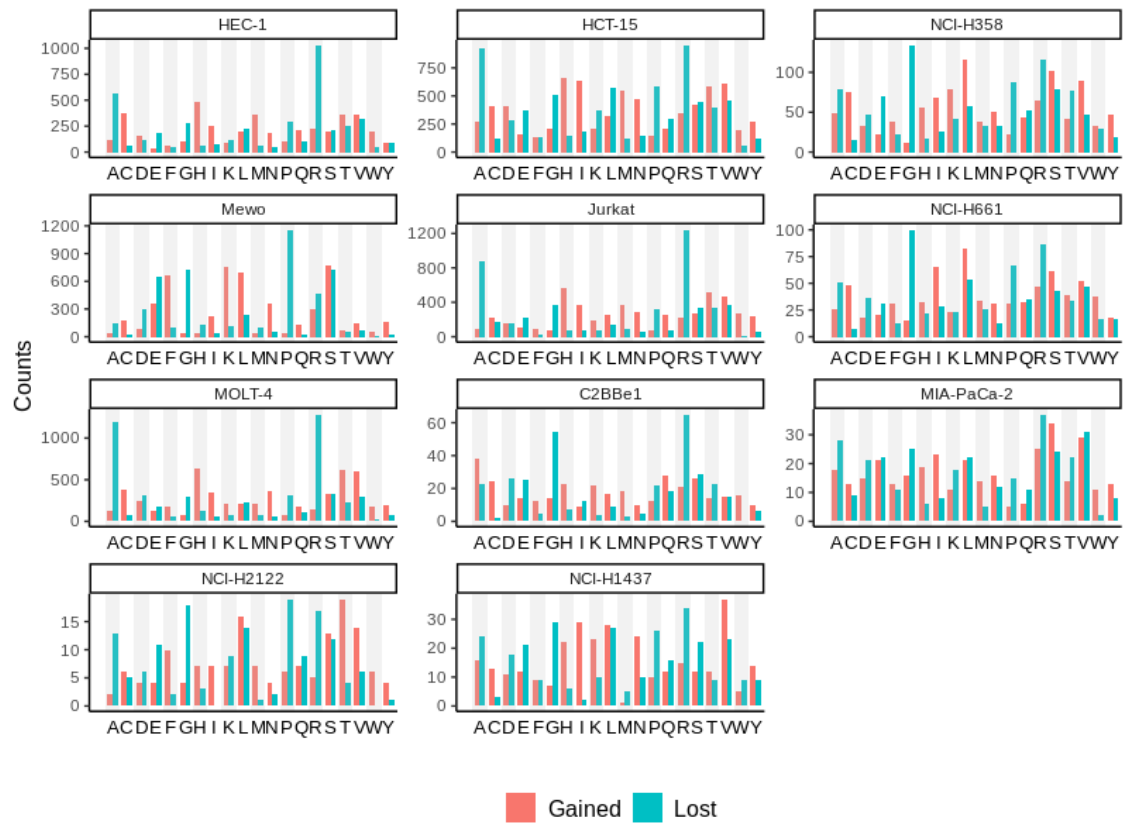


Figure S8. Total gain and loss missense mutation counts in a panel of cell lines in this study (COSMIC CLP v96). C2BBE1 is a CaCo-2 cell line clone in COSMIC-CLP.

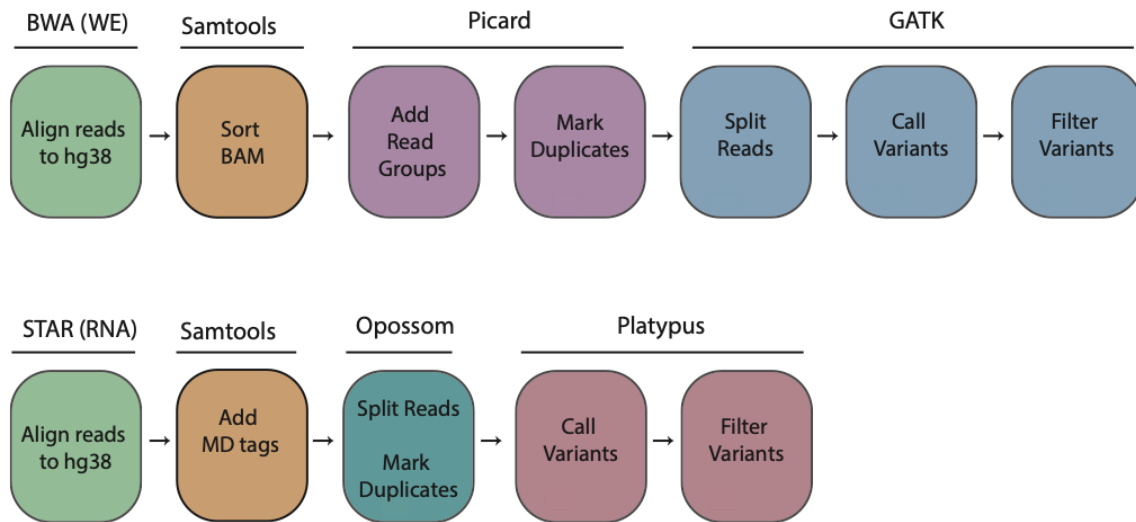


Figure S9. Variant calling pipelines for RNA and whole-exome (WE) datasets. Details in methods. Raw reads submitted to Sequence Read Archive (SRA) as BioProject PRJNA997729. Color indicates commands under toolkit listed.

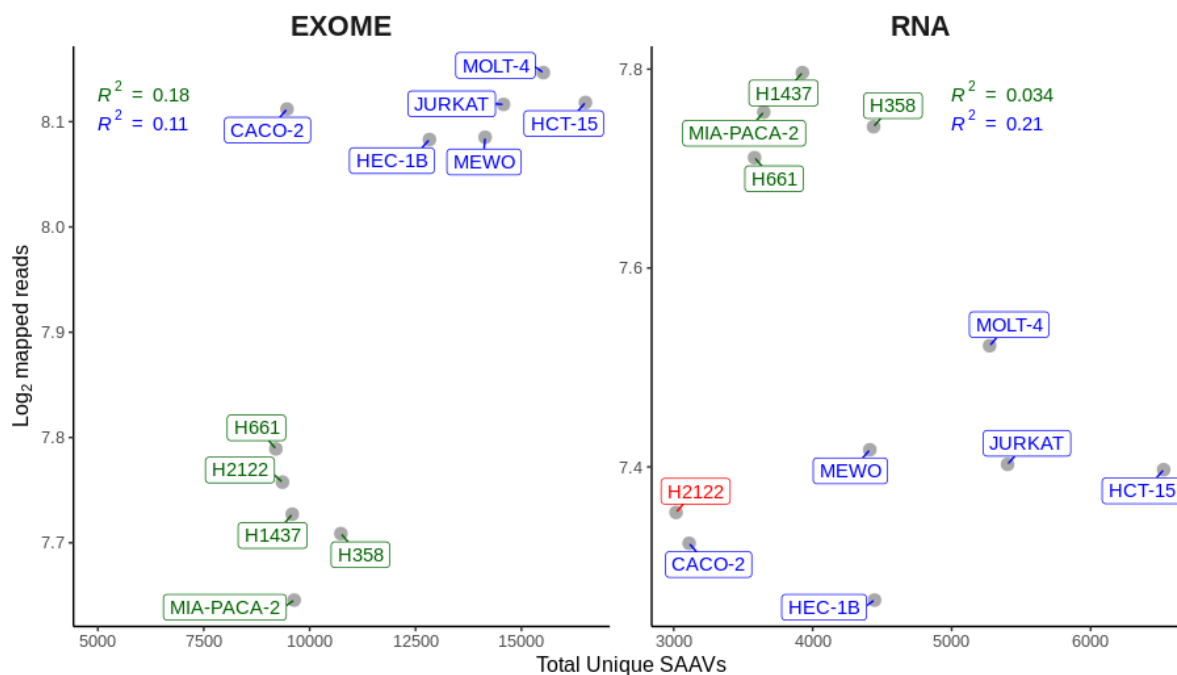


Figure S10. Number of mapped reads per cell line after BWA or STAR mapping and total unique SAAVs (single amino acid variants) identified per cell line. Color indicates sequencing batch

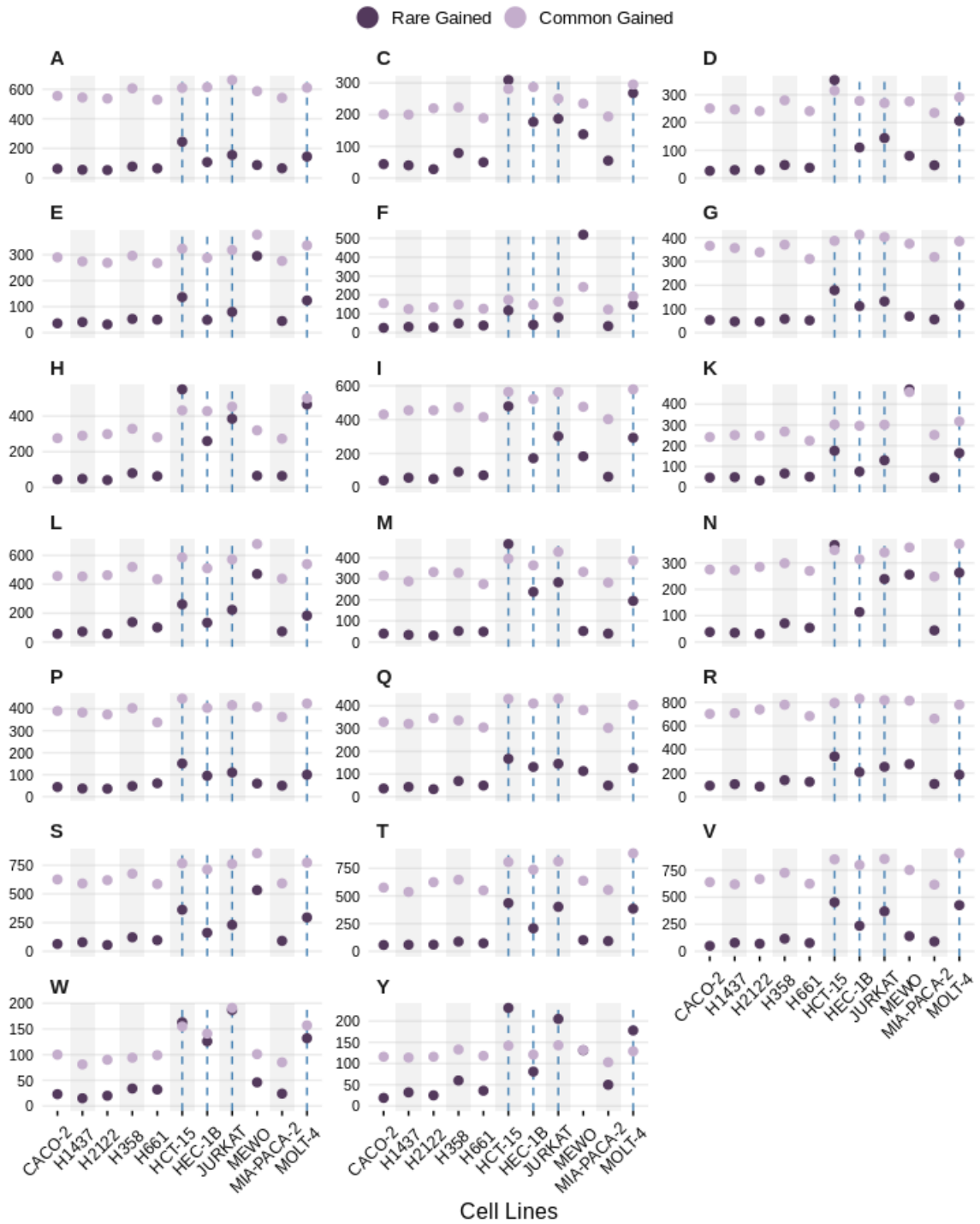


Figure S11. Gained amino acid counts identified in individual cell lines separated by common and rare. Dotted lines indicate dMMR cell lines

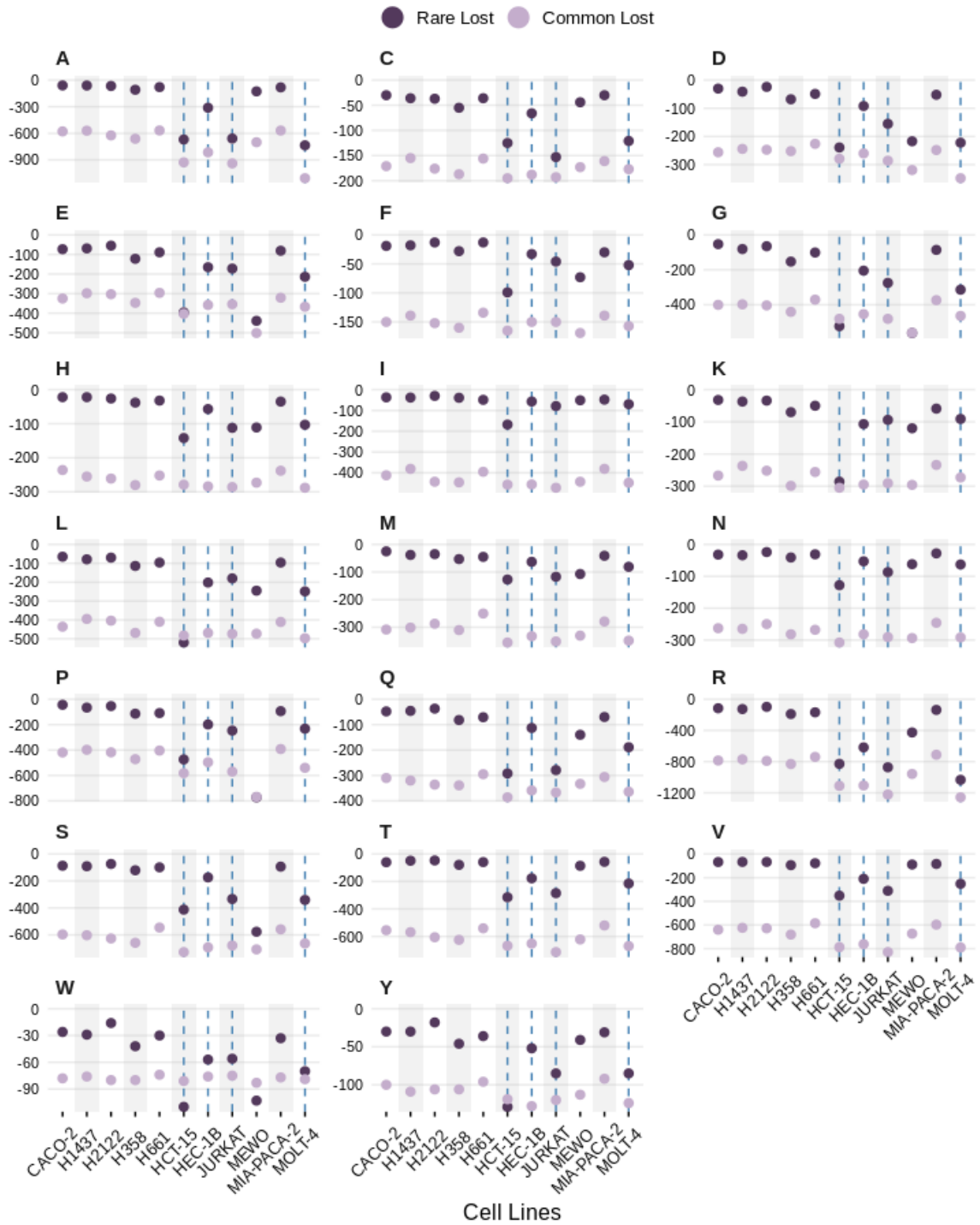


Figure S12. Lost amino acid counts identified in individual cell lines separated by common and rare. Dotted lines indicate dMMR cell lines.

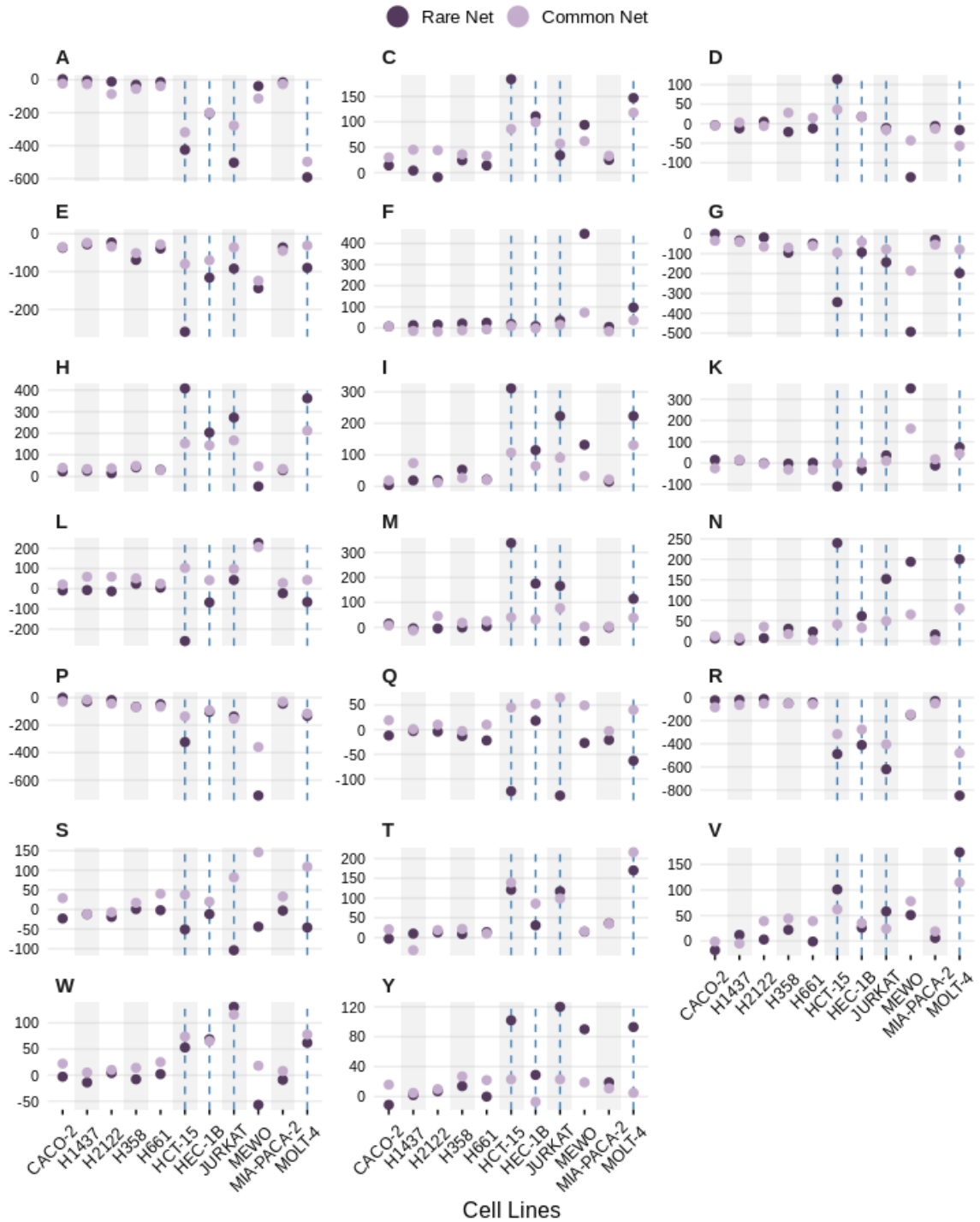


Figure S13. Net amino acid counts in identified individual cell lines separated by common and rare. Dotted lines indicate dMMR cell lines.

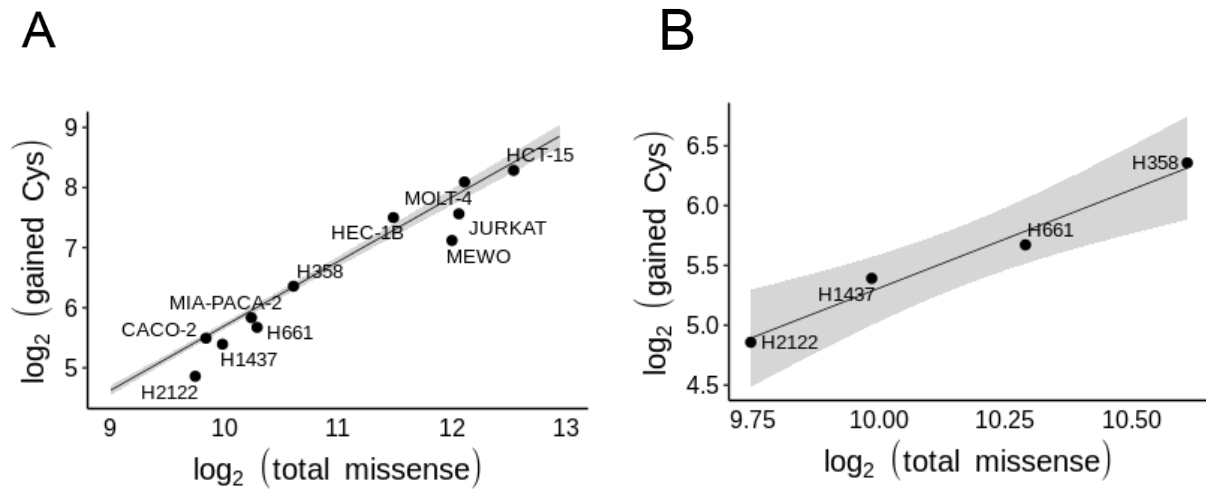
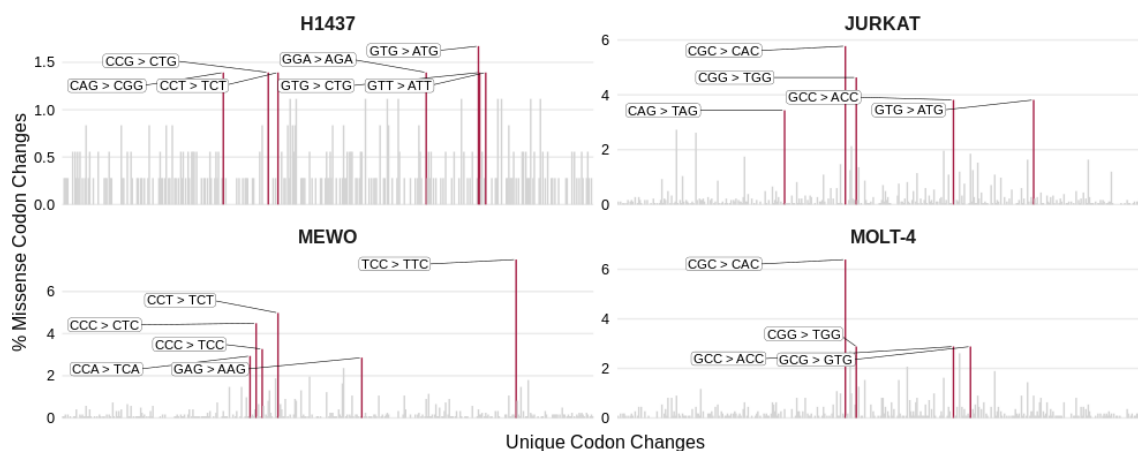


Figure S14. A) Gained cysteine missense mutations relative to total missense mutations from sequencing data fit to all COSMIC-CLP linear model from Figure 1B and 1C. B) Lung cancer lines subset with linear regression and 95% confidence interval shaded in gray—suggests the comparatively modest impact of smoking on cysteine acquisition for the cell lines evaluated.

A



B

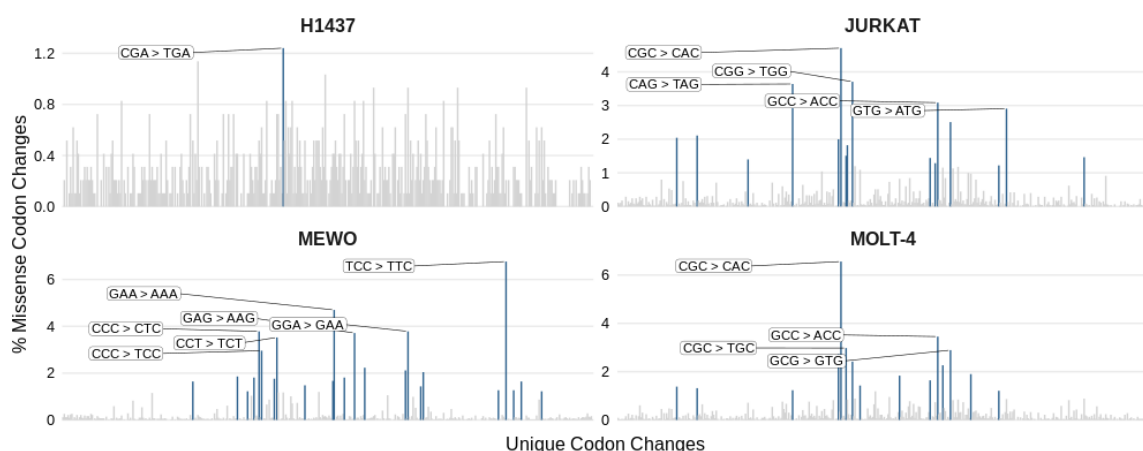


Figure S15. Unique codon changes in Jurkat, NCI-H1437, MOLT-4 and MeWo cell lines; RNA changes in red and exome changes in blue; analysis of rare variants. Tobacco-smoke carcinogens, such as benzo[*a*]pyrene (B[*a*]P) diol epoxides, uniquely confer G→T (C→A) transversions that we see enriched only in our smoking LC line. dMMR and UV mutations are predominantly C→T (G→A) transitions, the flanking nucleotides differ as UV radiation induces pyrimidine dimers which largely result in CC→TT transitions (and counterpart GG→AA) which cause gain of F/K and loss of S/E.

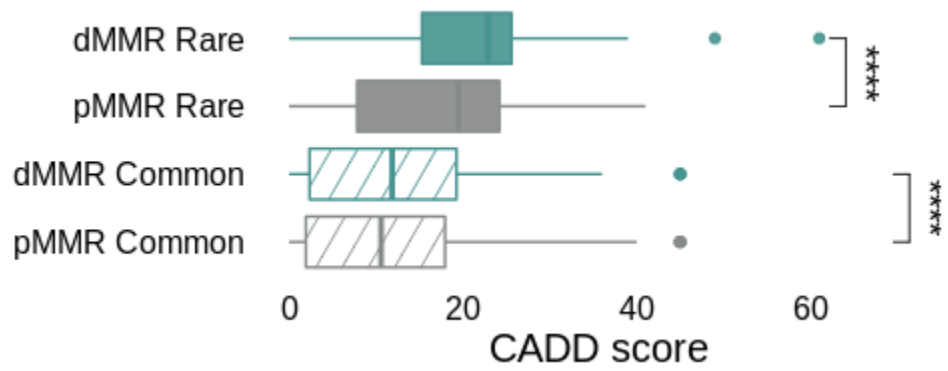


Figure S16. Distribution of CADD (Combined Annotation Dependent Depletion) scores for indicated variant grouping from Figure 2E data, Statistical significance was calculated using Mann-Whitney U test, **** $p < 0.0001$.

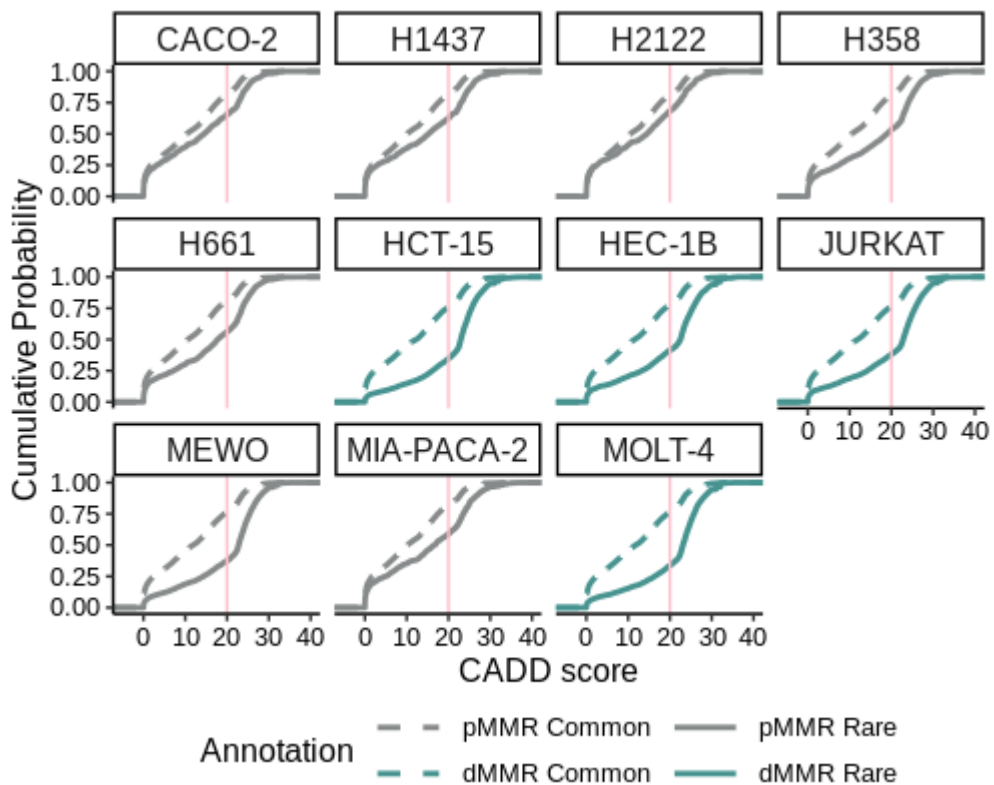


Figure S17. CADD-phred cumulative probabilities stratified by cell line. Pink lines indicate CADD-phred = 20

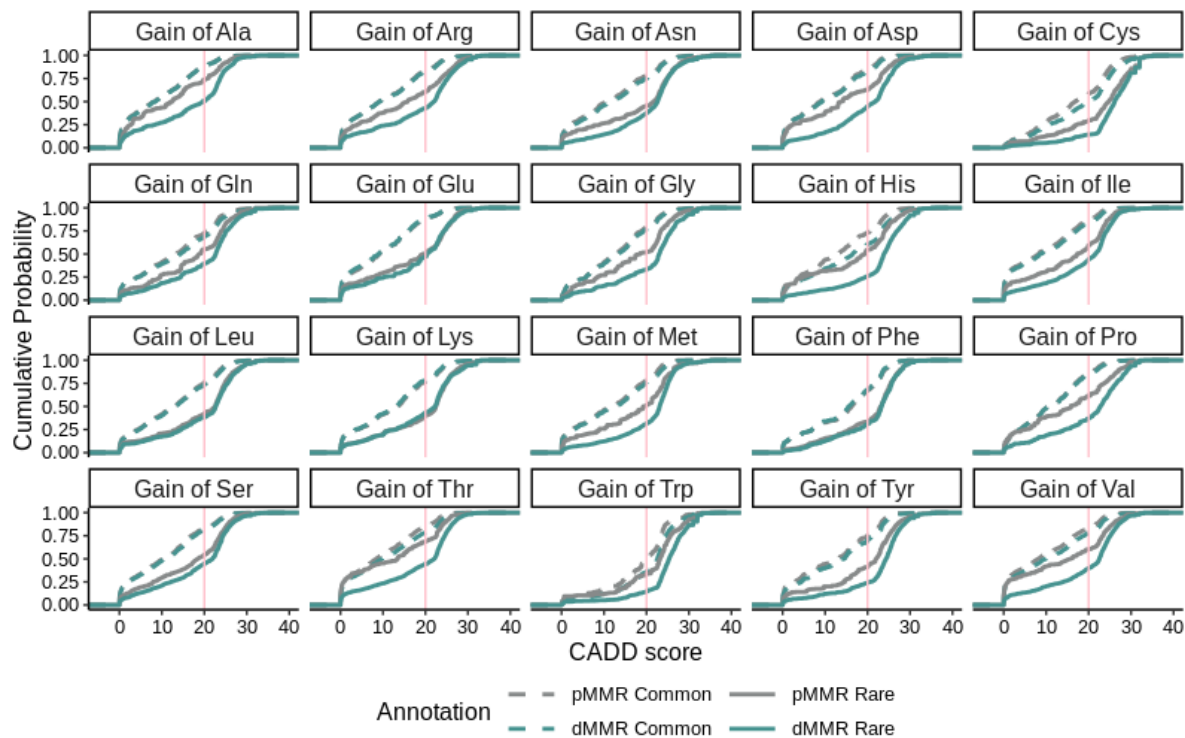
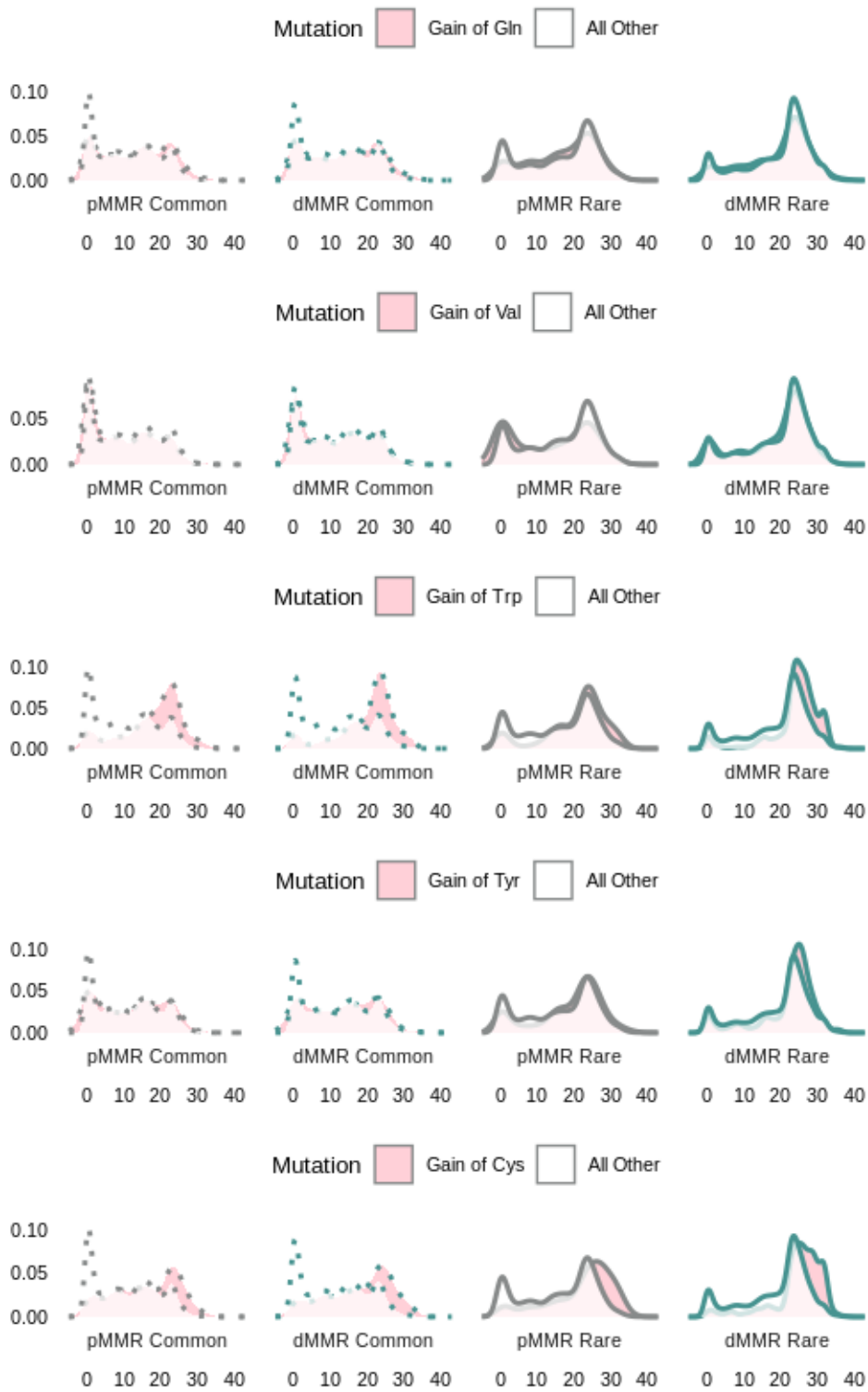
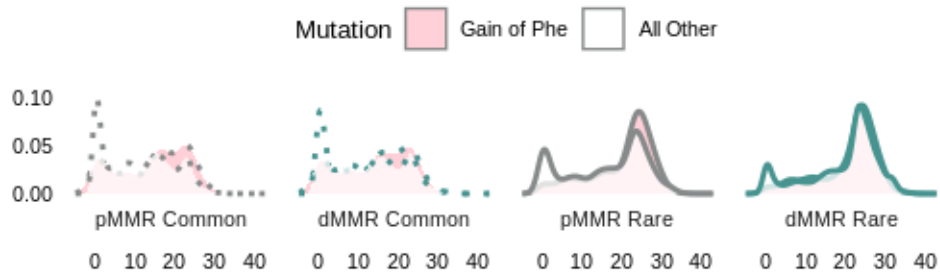
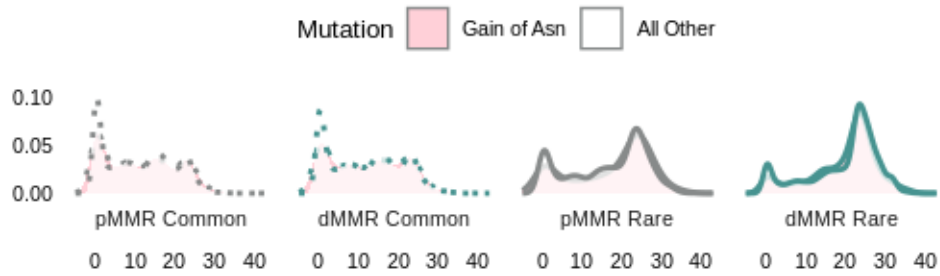
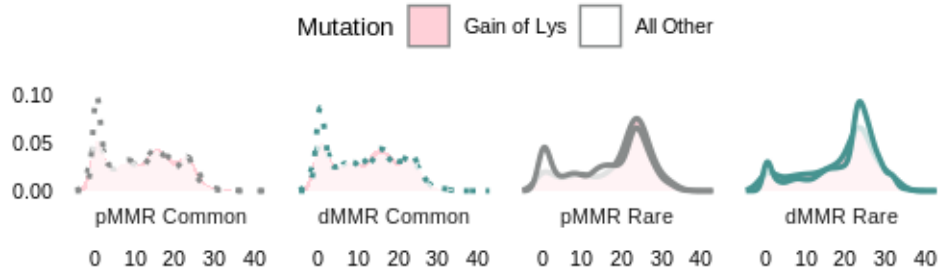
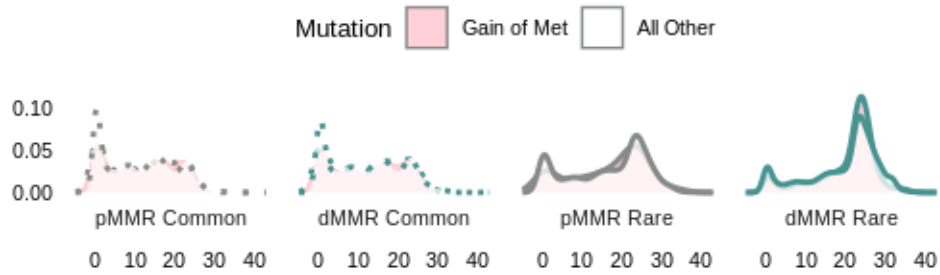
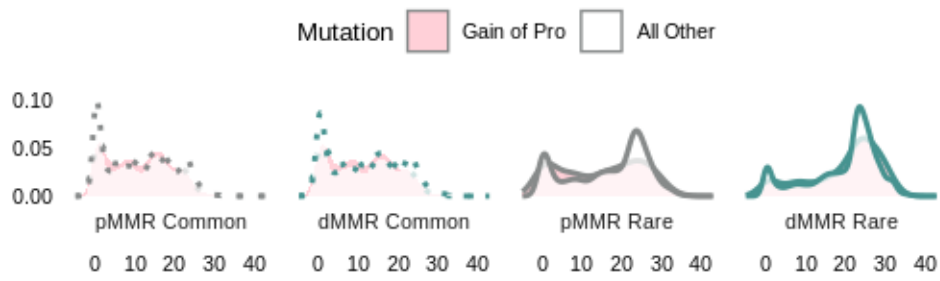
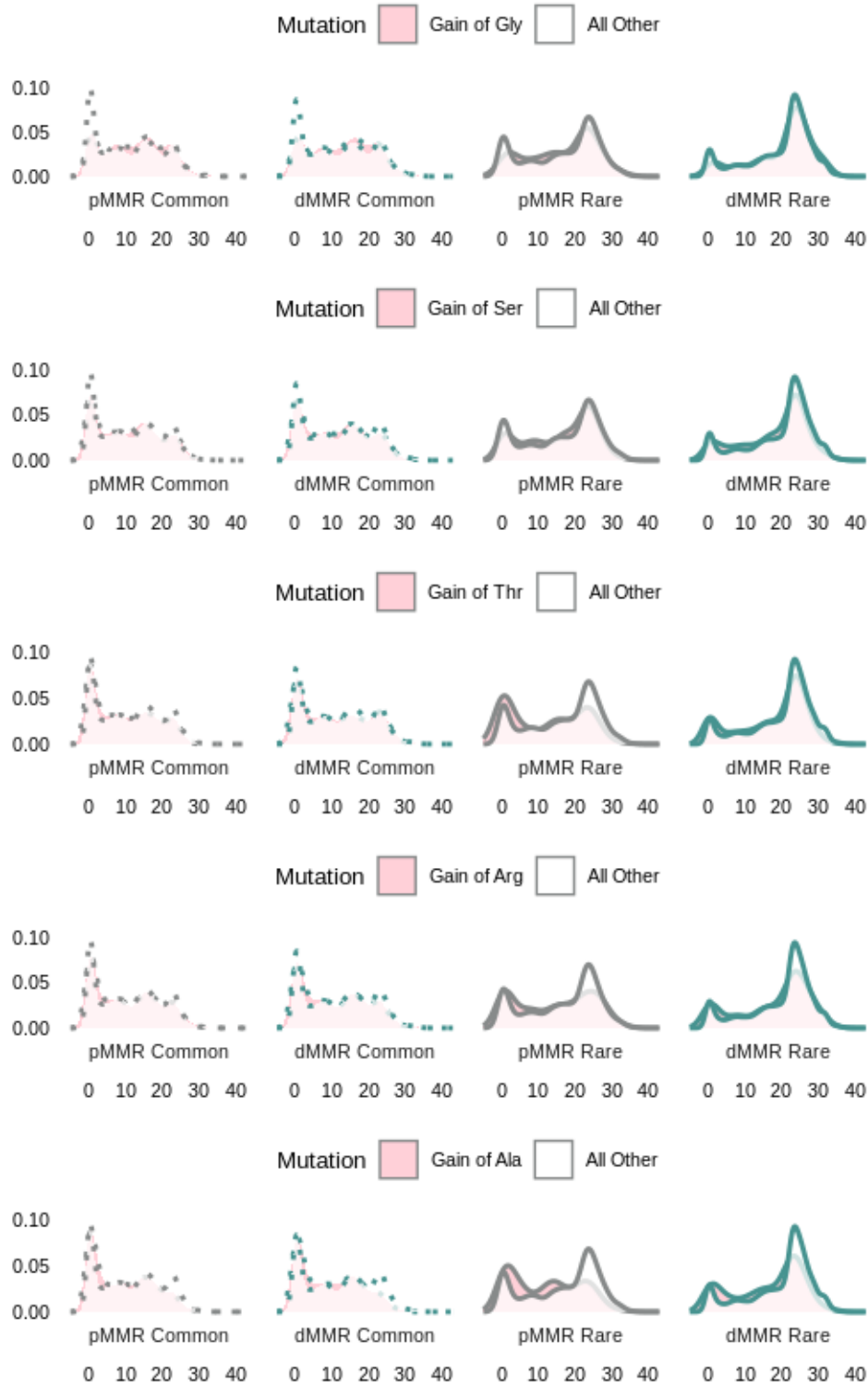


Figure S18: CADD-phred cumulative probabilities stratified by gained amino acid. Pink lines indicate CADD-phred = 20. One sided Mann Whitney U test p-values are in **Table S2** (tab S16).







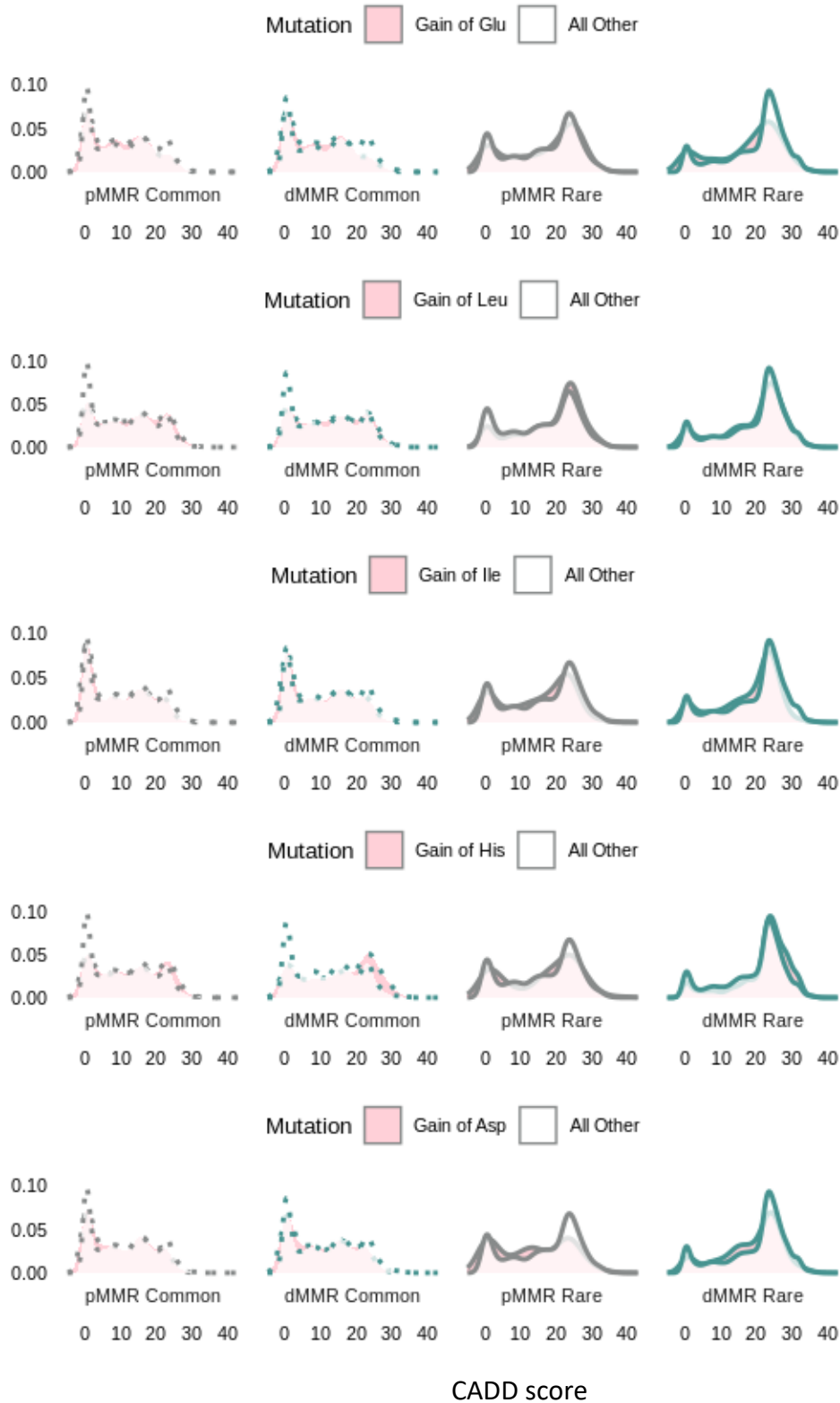


Figure S19. Distribution of CADD-phred scores for indicated amino-acid gained grouping. One sided Mann Whitney U test p-values are in **Table S2** (tab S16).

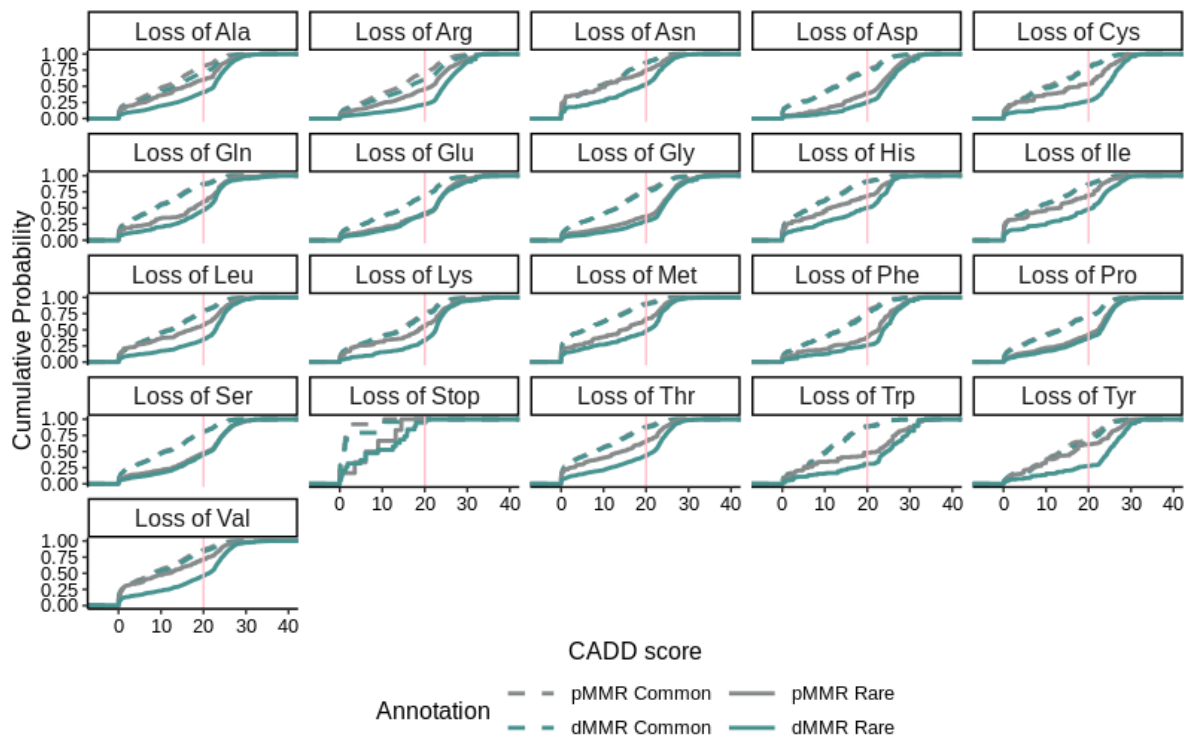
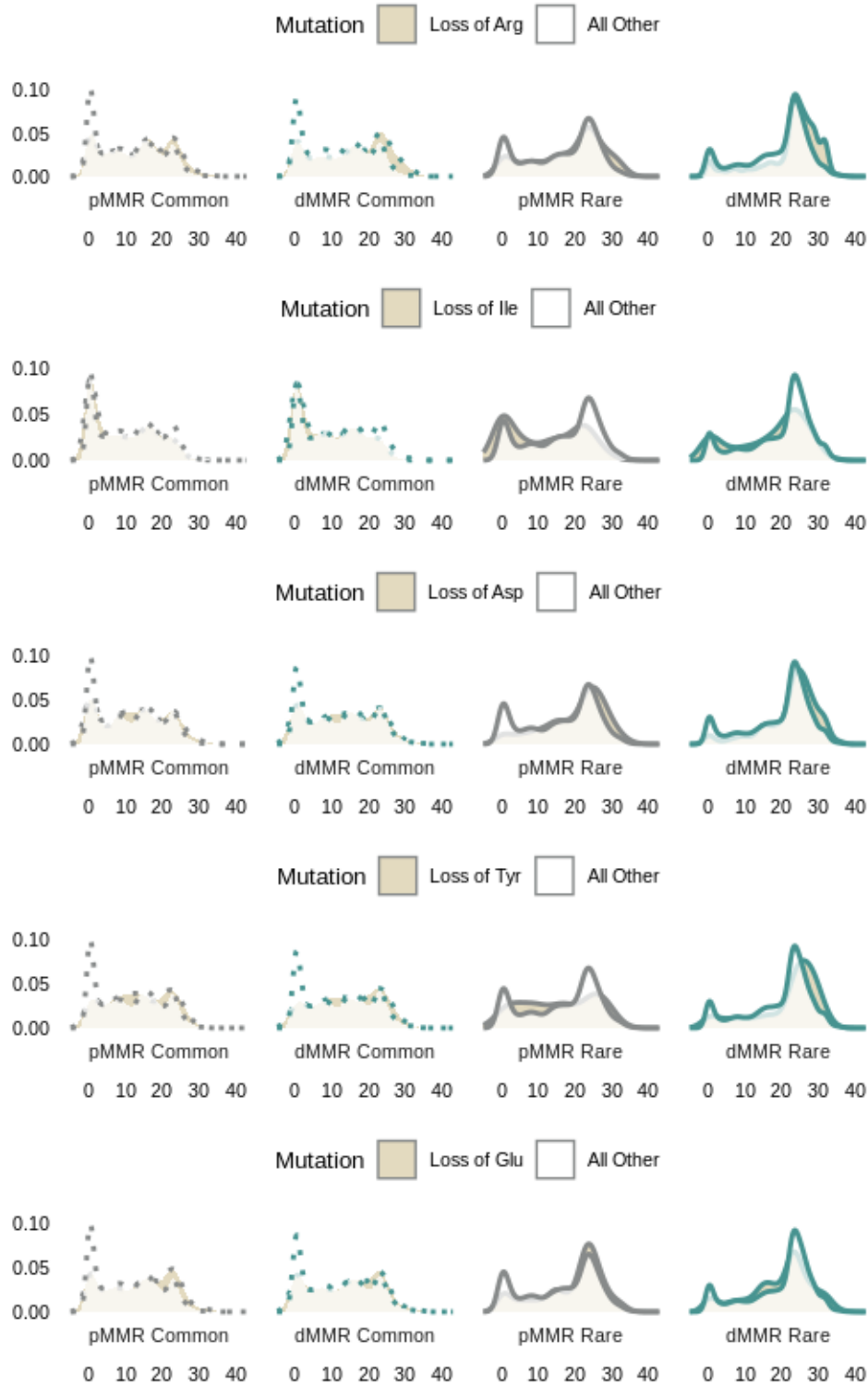
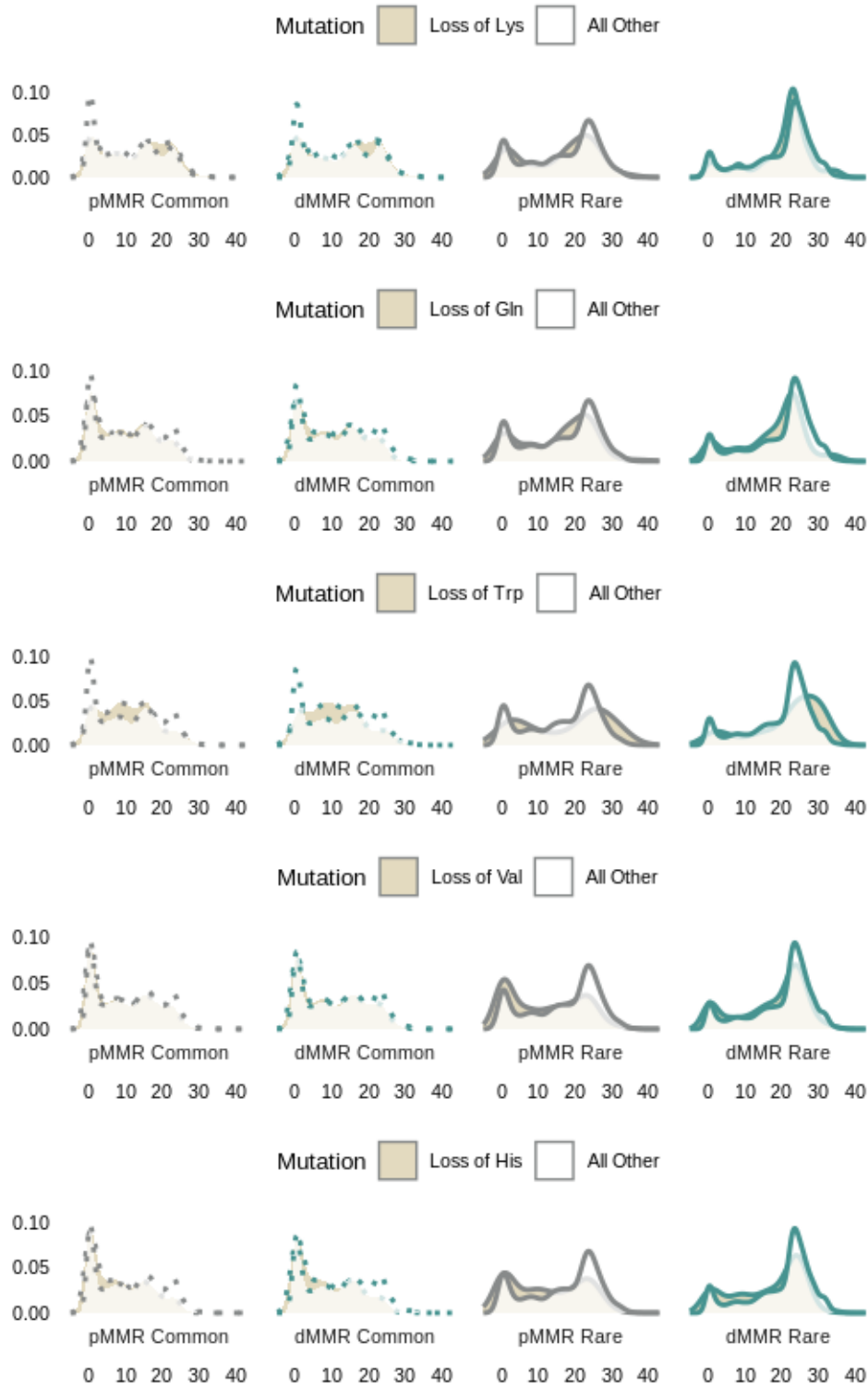
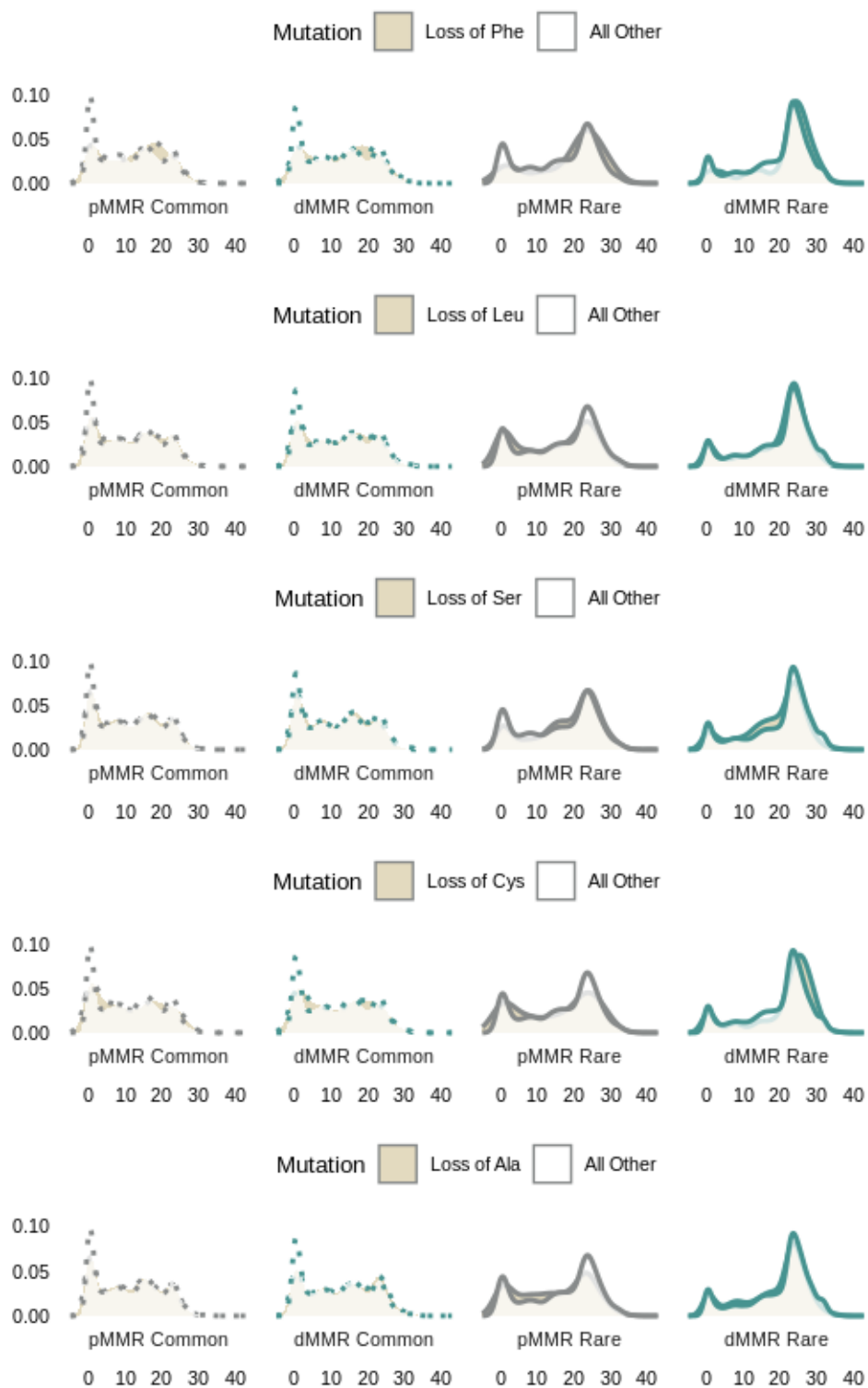


Figure S20. CADD-phred cumulative probabilities stratified by lost amino acid. Pink lines indicate CADD-phred = 20. One sided Mann Whitney U test p-values are in **Table S2** (tab S16).







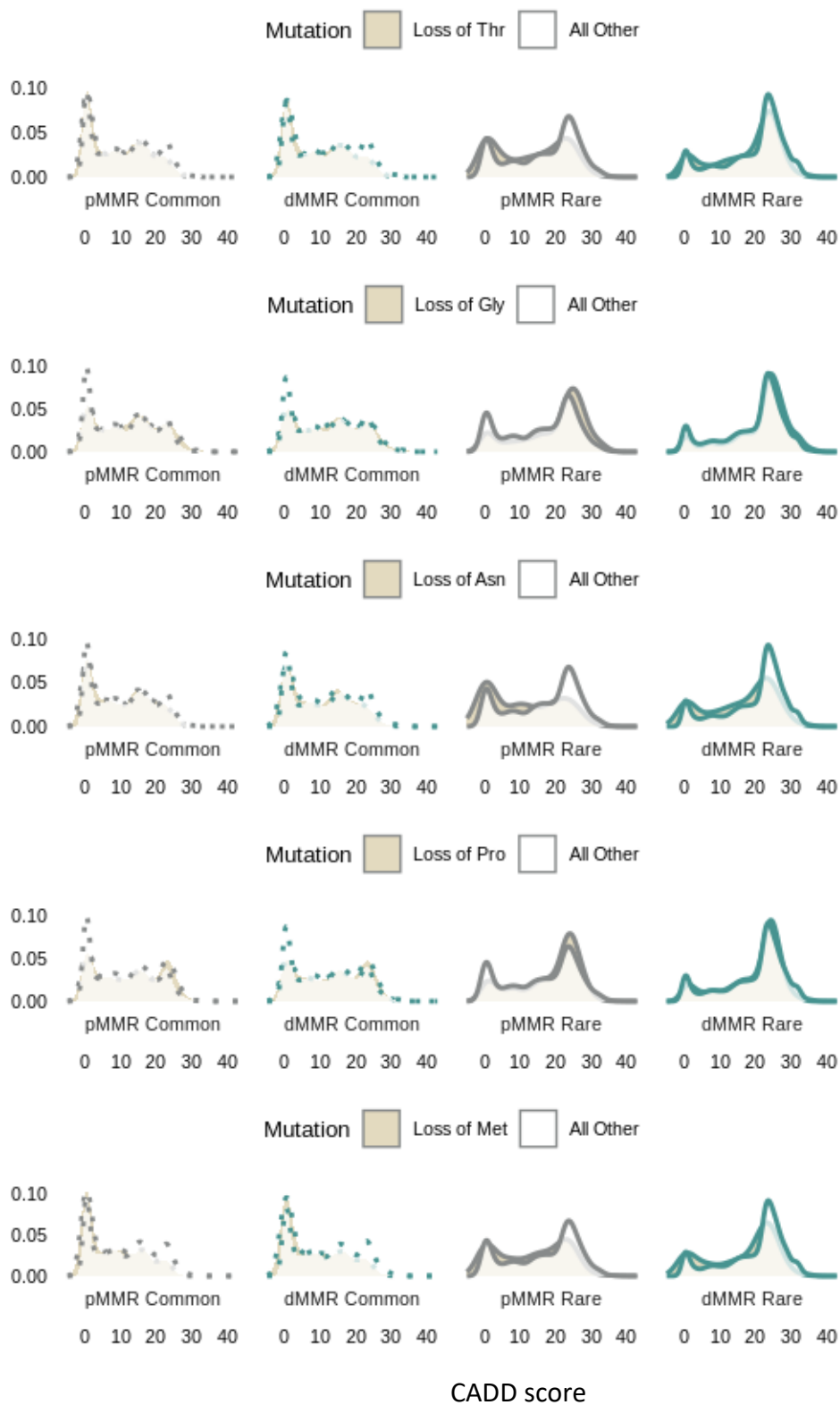


Figure S21. Distribution of CADD-phred scores for indicated amino-acid loss grouping. One sided Mann Whitney U test p-values are in **Table S2** (tab S16).

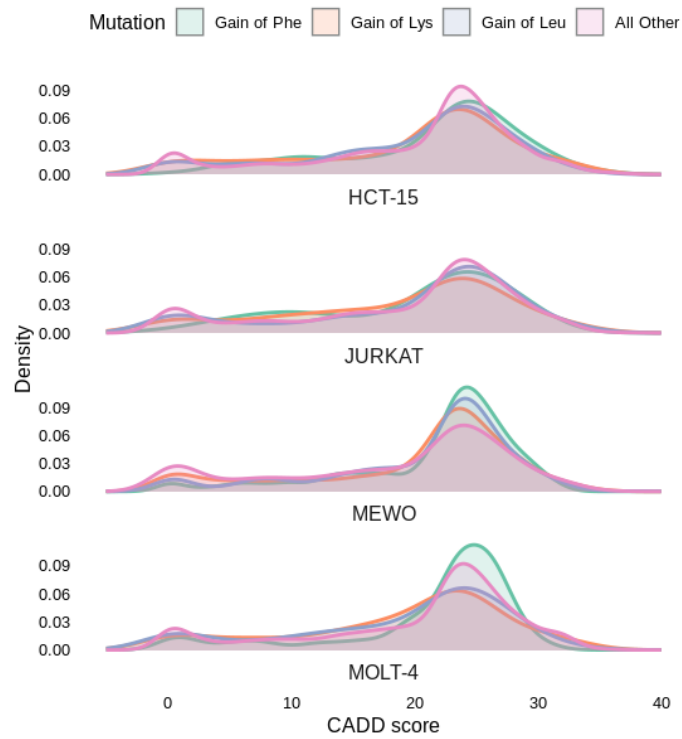


Figure S22. Distribution of CADD-phred scores for selected amino acid rare variants in MeWo, HCT-15, MOLT-4, and Jurkat cell lines

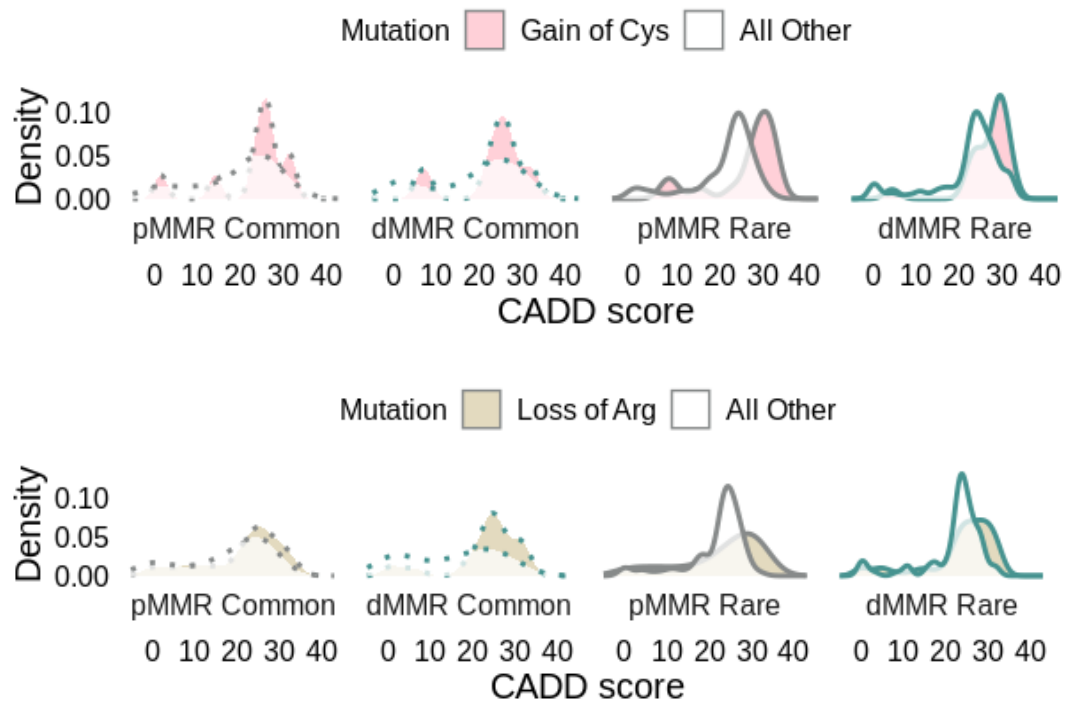


Figure S23. The Cancer Genome Atlas (TCGA) hotspot mutation distribution of CADD scores for indicated grouping. Highly deleterious gain of cysteines that are also TCGA hotspot mutations include FBXW7 R505C in Jurkat and ADAMTS1 R604C in MOLT-4, and SCYL3 R61C in MeWo. n=10, 6,16,21 for Cys ; n=25, 34, 31, 90 for Arg across from left.

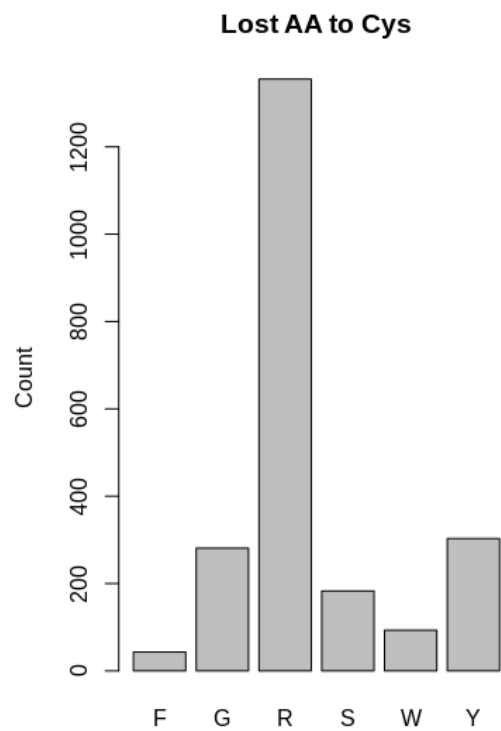


Figure S24. Distribution of reference/lost amino acids of gained cysteines in **Figure 2** data (60% arginine).

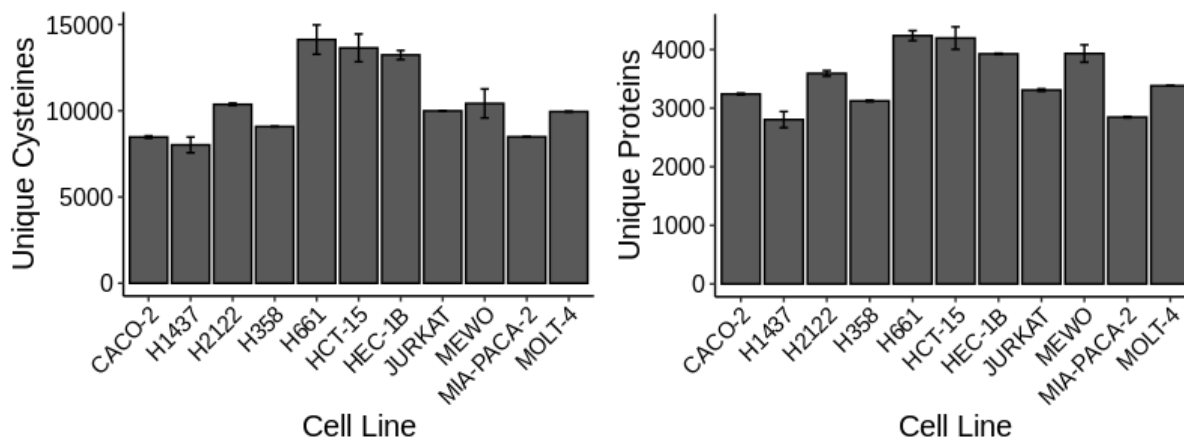


Figure S25. Reference cysteines and proteins identified per cell line in **Figure 4** for duplicate datasets.

Unique Cys
in Cys DB

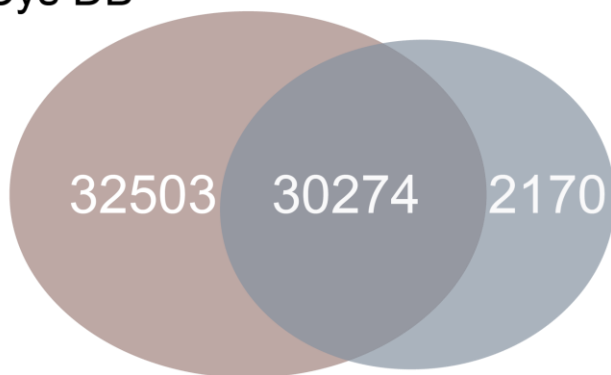


Figure S26. **Figure 4** reference cysteines overlap with CysDB cysteines. Red indicates Cys DB dataset, blue indicates **Figure 4** dataset.

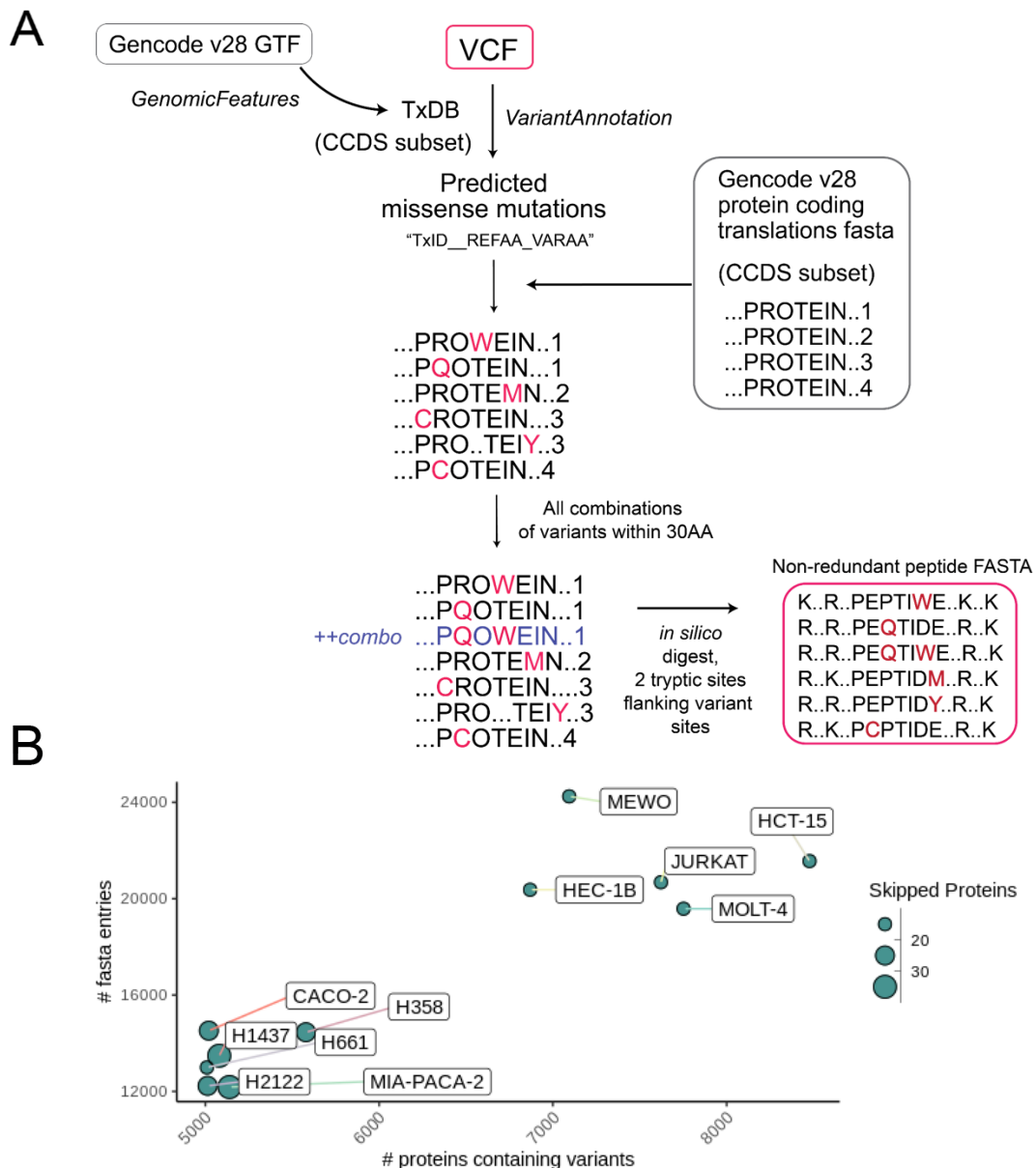


Figure S27: Details of database generation. A) Pipeline of packages and inputs to obtain variant peptide databases—*VariantAnnotation*¹ package is used to obtain predicted changes that replace reference sequence residues from Gencode by matching internal transcript IDs (TxID). Details in methods. Combinations are omitted from proteins with > 25 or 15 variants. B) Variant database sizes calculated as the number of FASTA entries. The skipped proteins indicate the number of proteins that were omitted from combinations in **Table S4** (tab S24).

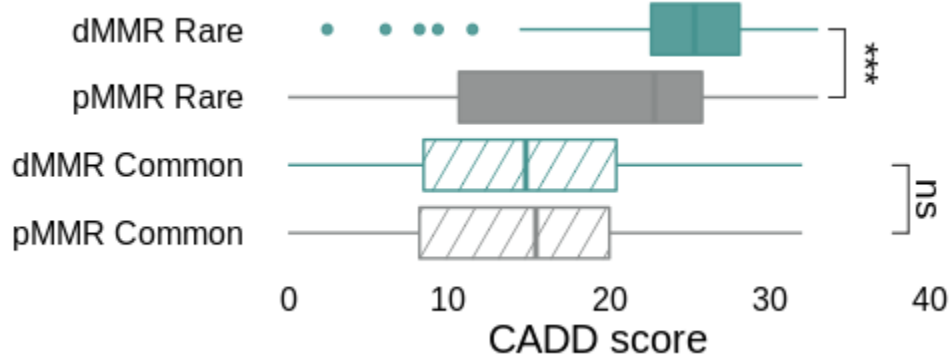


Figure S28: Distribution of CADD-phred scores for indicated variant grouping from **Figure 4** data. Statistical significance was calculated using Mann-Whitney U test, *** $p < 0.001$.

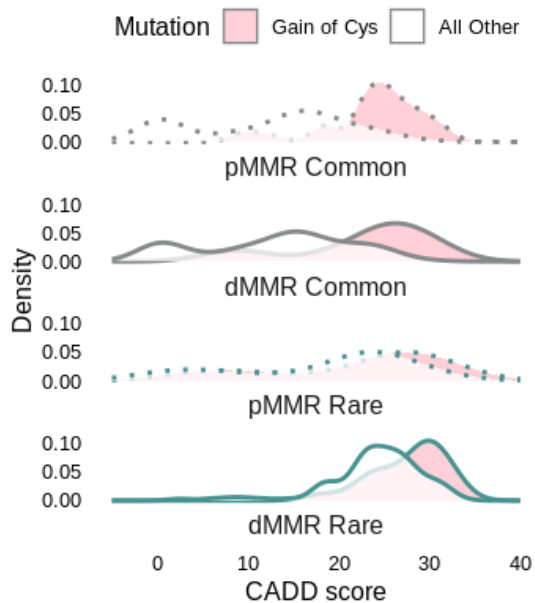


Figure S29: Distribution of CADD-phred scores for indicated gained cysteines vs all other from **Figure 4** data.

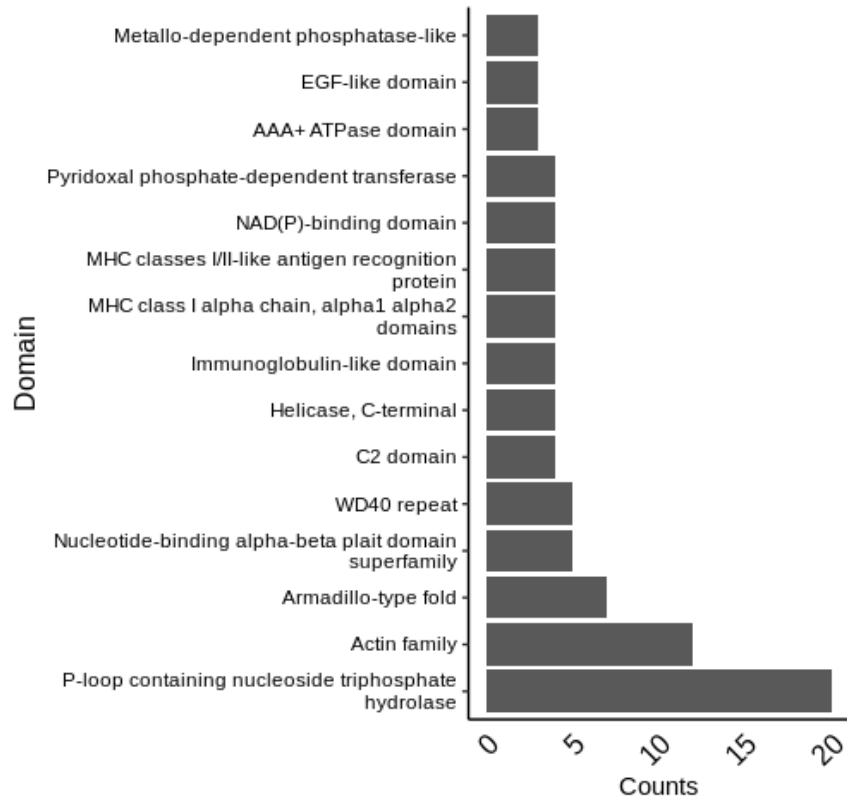


Figure S30: Identified variants' domain residence counts from Figure 4 data.

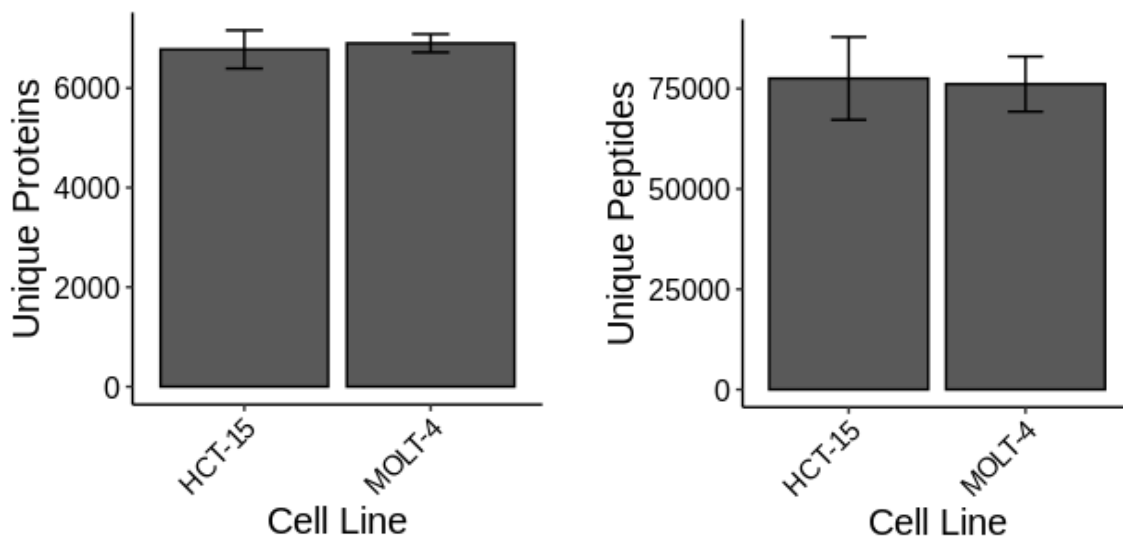


Figure S31. Reference peptides and proteins identified per cell line in Figure 5 for triplicate sets of high pH fractionated samples per cell line.

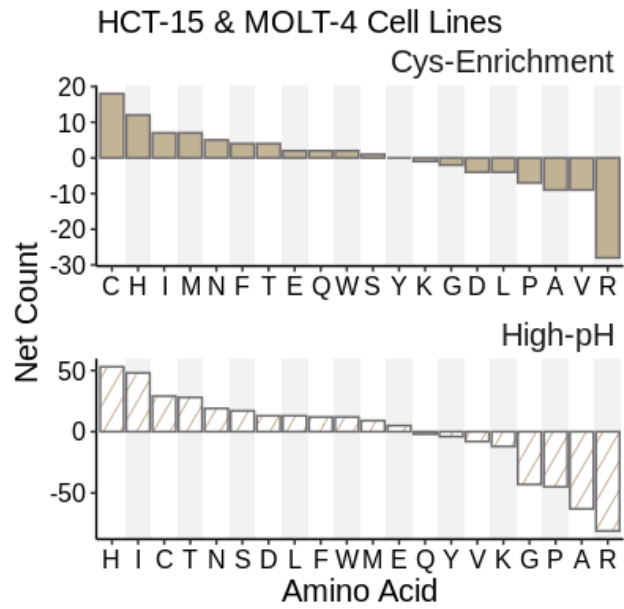


Figure S32. Net counts of SAAVs identified in Figure 5.

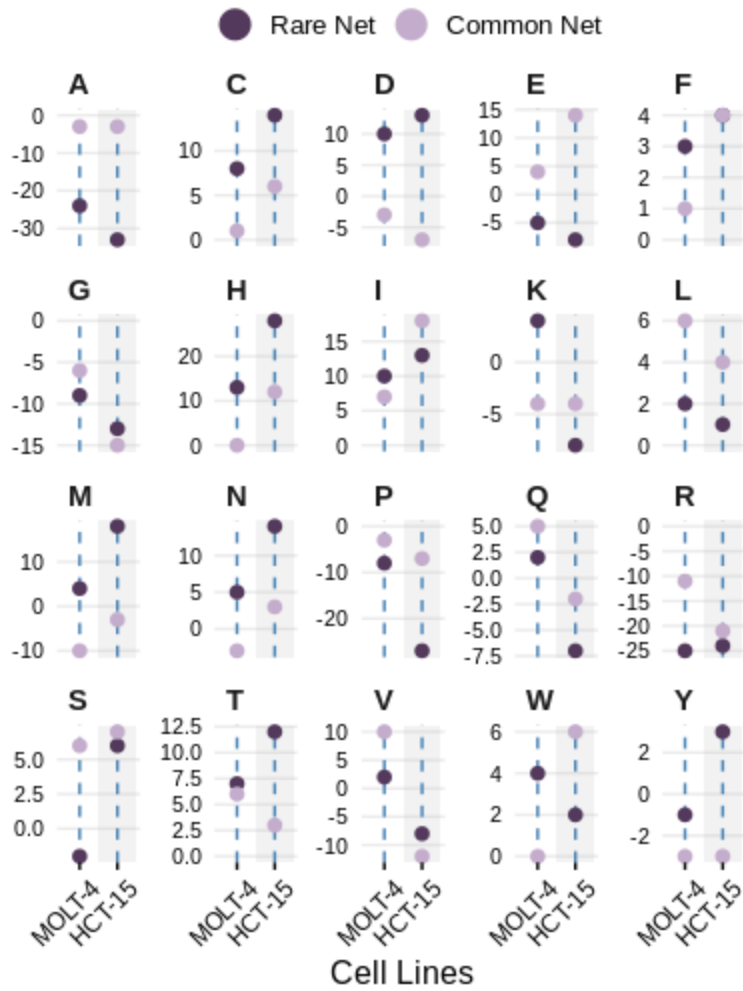


Figure S33. Net counts of SAAVs stratified by amino acid and common vs rare identified in Figure 5.

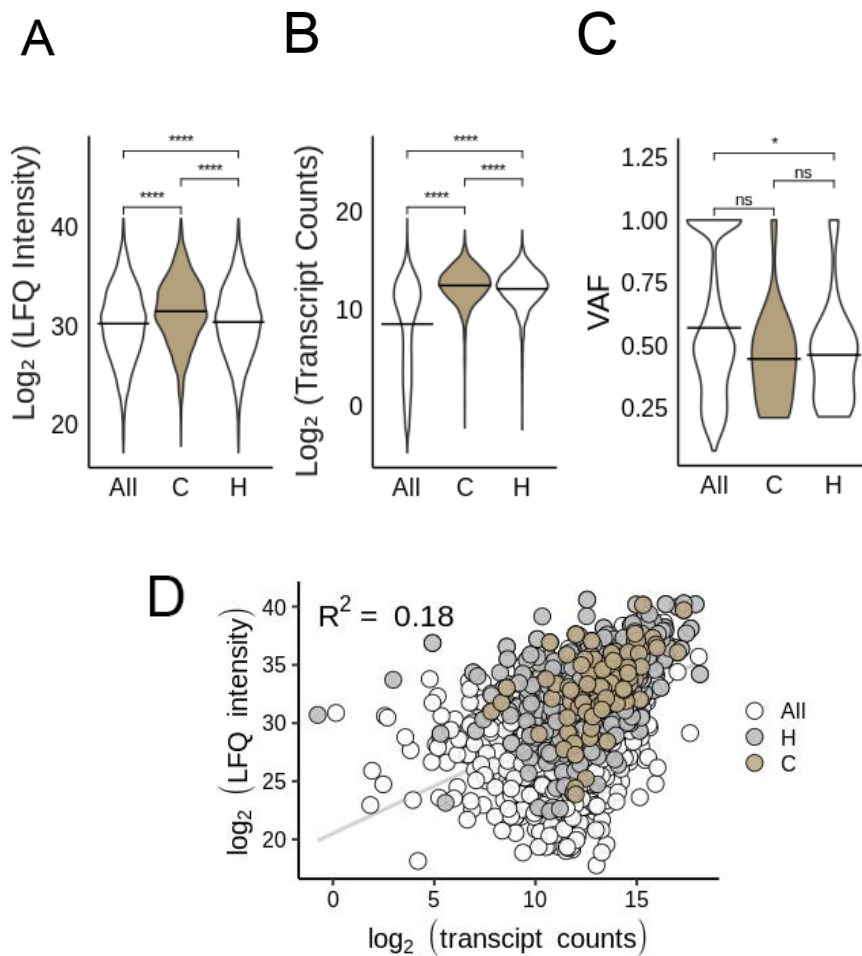


Figure S34: A) Label free quantitation (LFQ) intensities comparisons and B) DE-seq normalized transcript comparisons from HCT-15 cell line reference database searches. C) Variant allele frequencies for gain-of-cysteines subset in HCT-15 and Molt-4 searches. D) Matched LFQ intensities and normalized transcript count correlation of variant containing proteins/transcripts; All=all proteins in LFQ search or all transcripts, C=proteins from cys-enrichment search, H=proteins from high-pH fractionation search.

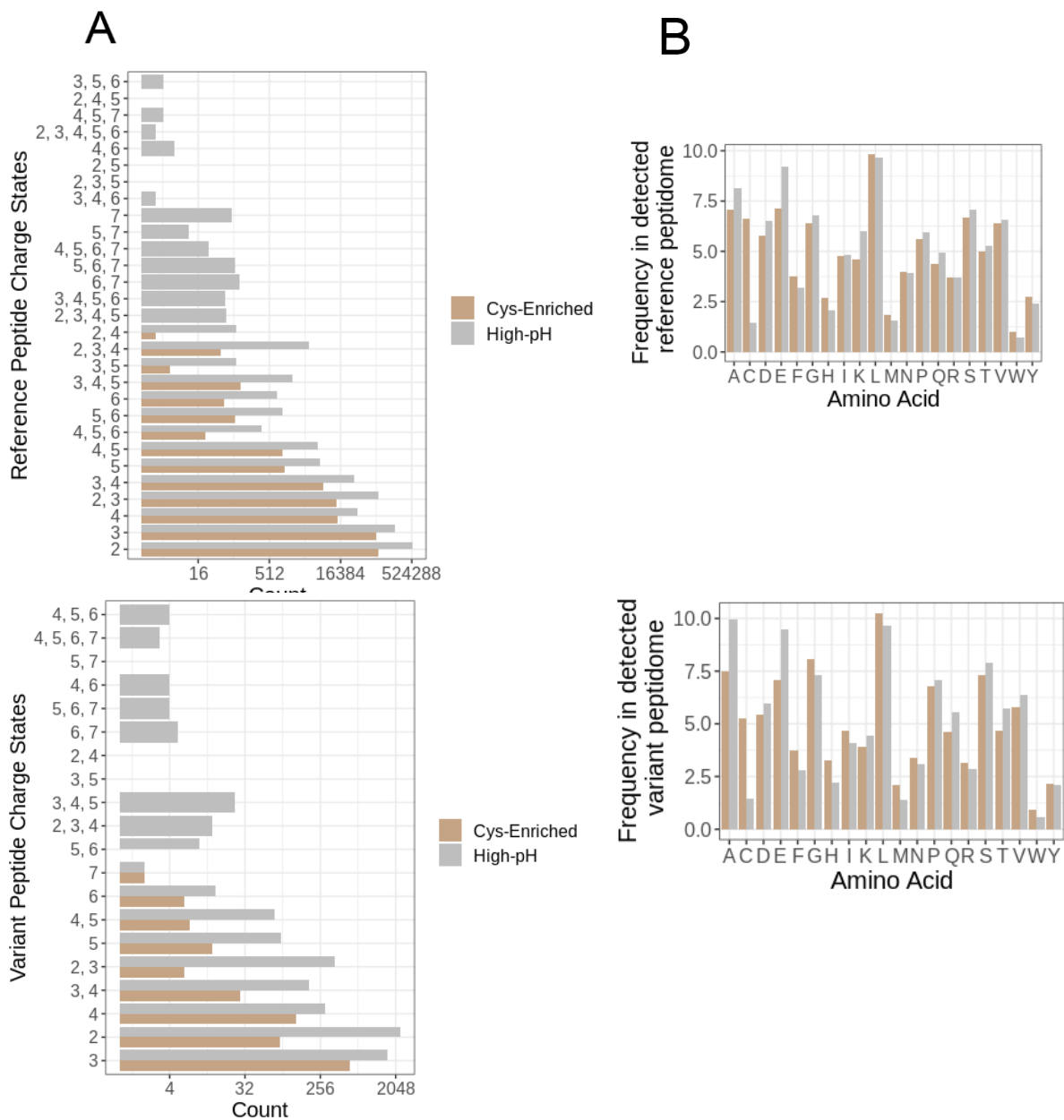


Figure S35: Peptide properties of detected reference peptides from cys-enrichment and high-pH fractionation A) Charge states of detected reference peptides and variant peptides. B) Abundance of amino acids in detected reference and variant peptides.

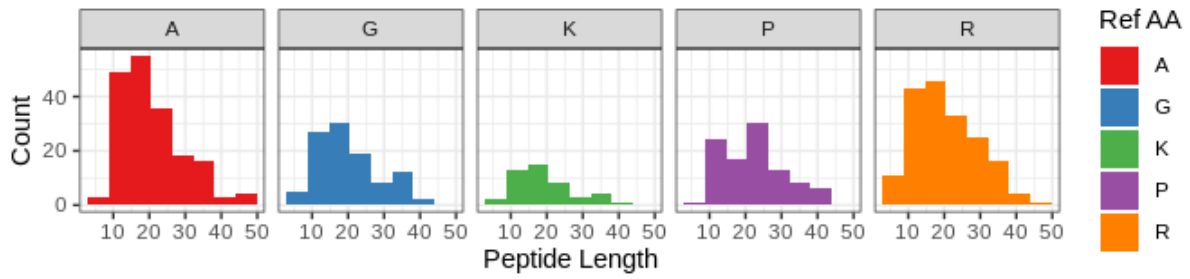


Figure S36: High-pH identified variant peptide length and SAAV reference amino acid for top lost amino acids.

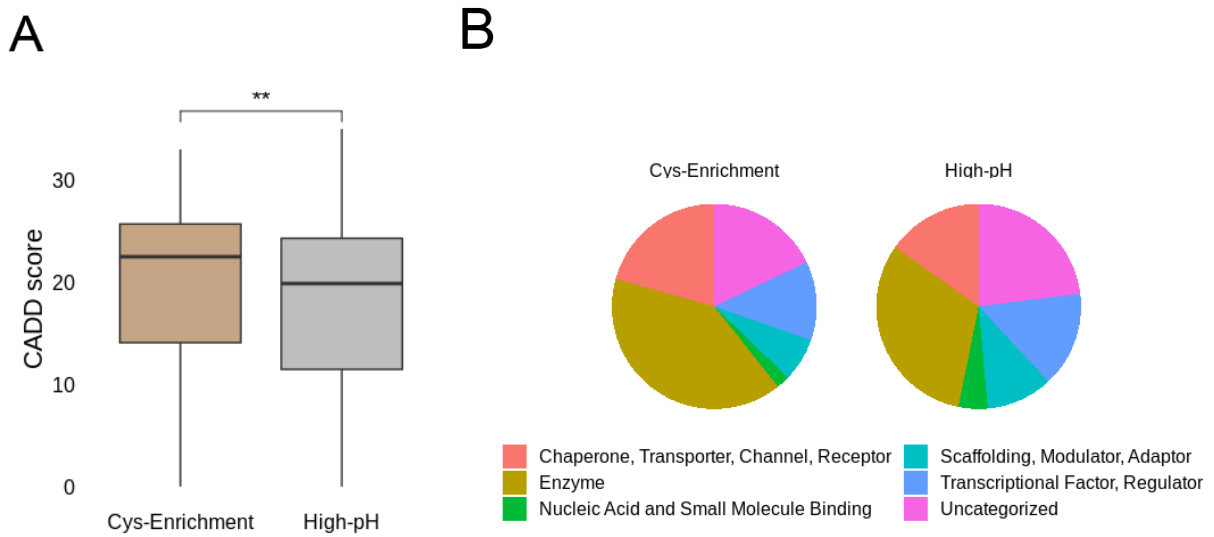


Figure S37: A) CADD score comparison between cys-enriched and high-pH detected variants B) Protein classes of cys-enriched and high-pH detected variants.

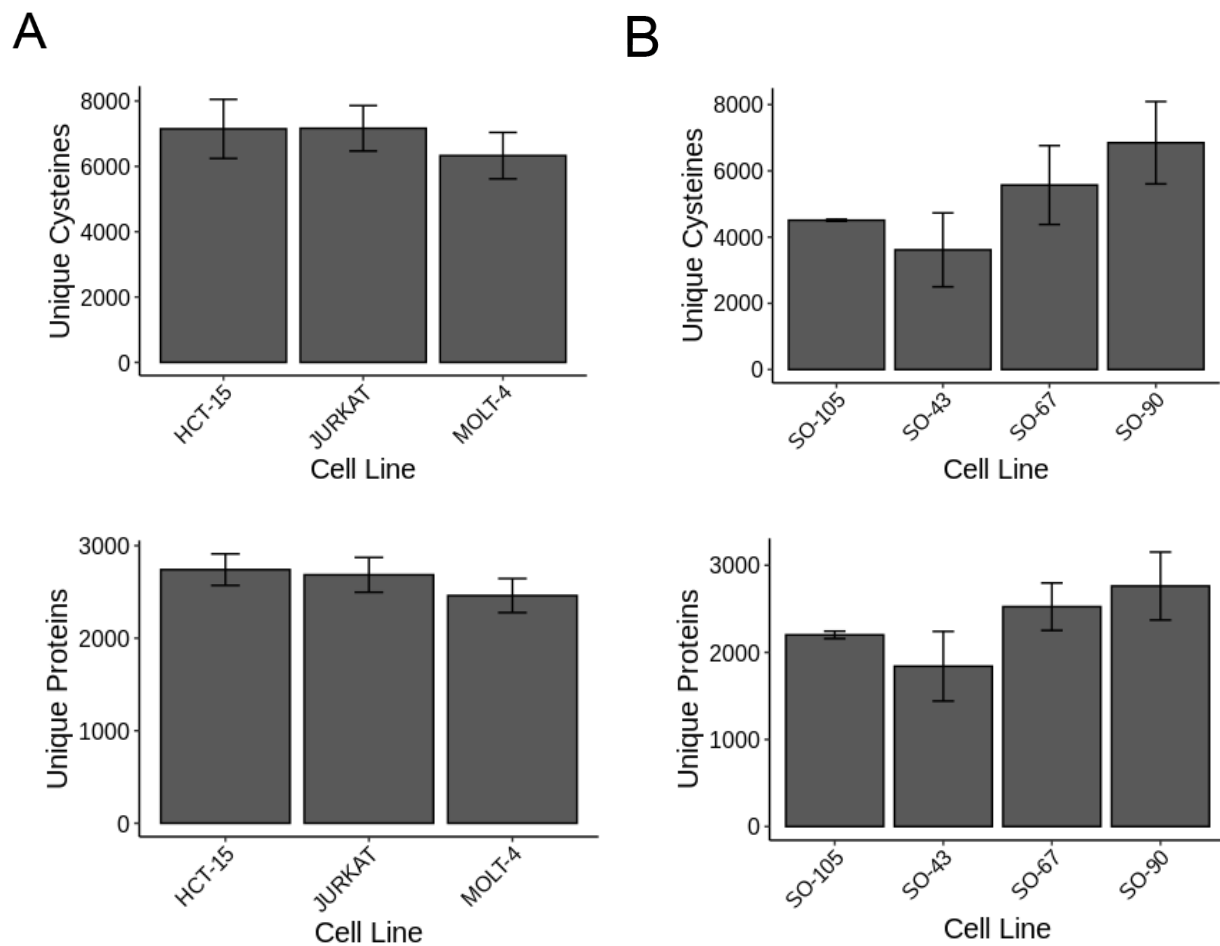


Figure S38. Reference identifications in Figure 6 for for A) KB02 datasets or B) HCT-15 datasets.

Variant search
 Reference search
 Enriched Cys

A HLA-A

```

1  MAVMAPRTL L L L L L S G A L A L T Q T W A G S H S M R Y F F T S V S R P G R G E P R F I A V G
51 Y V D D T Q F V R F D S D A A S Q R M E P R A P W I E Q E G P E Y W D Q E T R N V K A Q S Q T D
99 R V D L G T L R G Y Y N Q S E A G S H T I Q I M Y G C D V G S D G R F L R G Y R Q D A Y D G K D Y I A
150 L N E D L R S W T A A D M A A Q I T K R K W E A A H E A E Q L R A Y L D G T C V E W L R R Y L E N G K
201 E T L Q R T D P P K T H M T H H P I S D H E A T L R C W A L G F Y P A E I T L T W Q R D G E D Q T Q D T
253 E L V E T R P A G D G T F Q K W A A V V V P S G E E Q R Y T C H V Q H E G L P K P L T L R W E L S S Q
304 P T I P I V G I I A G L V L L G A V I T G A V V A A V M W R R K S S D R K G G S Y T Q A A S S D S A Q G S D
358 V S L T A C K V
  
```

B HLA-B

```

1  M L V M A P R T V L L L L S A A L A L T E T W A G S H S M R Y F Y T S V S R P G R G E P R F I S V G
51 Y V D D T Q F V R F D S D A A S P R E E P R A P W I E Q E G P E Y W D R N T Q I Y K A Q A Q T D
99 R E S L R N L R G Y Y N Q S E A G S H T L Q S M Y G C D V G P D G R L L R G H D Q Y A Y D G K D Y I A
150 L N E D L R S W T A A D T A A Q I T Q R K W E A A R E A E Q R R A Y L E E C V E W L R R Y L E N G K
201 D K L E R A D P P K T H V T H H P I S D H E A T L R C W A L G F Y P A E I T L T W Q R D G E D Q T Q D T
253 E L V E T R P A G D R T F Q K W A A V V V P S G E E Q R Y T C H V Q H E G L P K P L T L R W E P S S Q
304 S T V P I V G I V A G L A V L A V V I G A V V A A V M C R R K S S G G K G G S Y S Q A A C S D S A Q G S D
358 V S L T A
  
```

Figure S39. Sequence coverage of Uniprot references A) HLA-A*3:01 and B) HLA-B*7:02 peptides from HCT-15 reference and variant searches in Figure 4-6 datasets; yellow indicates enriched cysteines and red are variant sites.

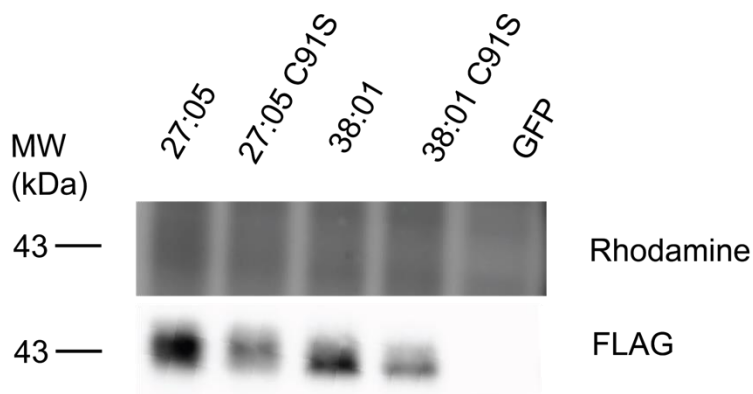


Figure S40. Cysteine lysate labeling of overexpressed HLA-B alleles with iodoacetamide alkyne (IAA) (representative of 2 two biological replicates) conjugation by click chemistry to a rhodamine-azide tag described in methods.

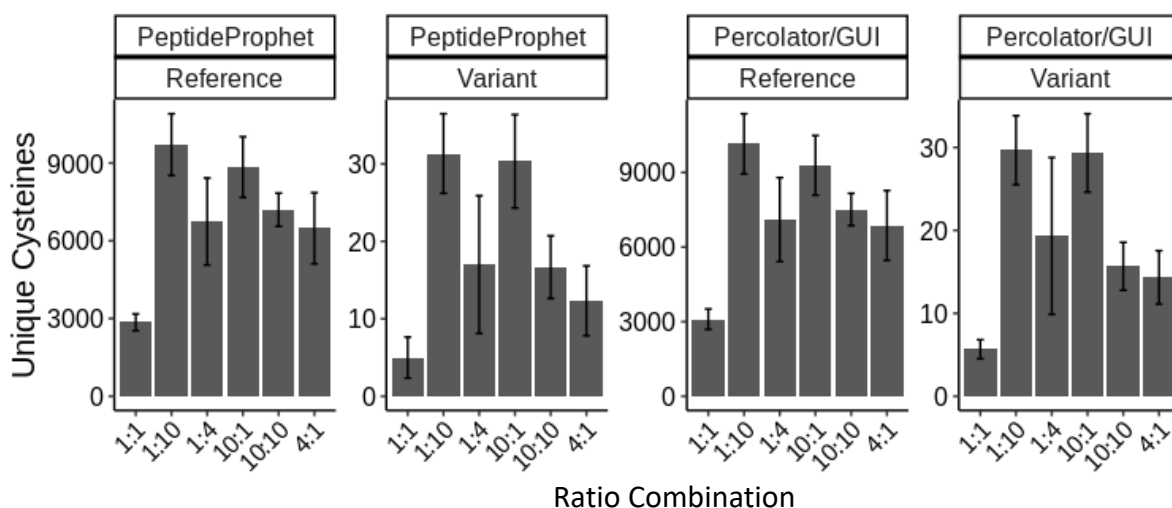


Figure S41. Coverage of FragPipe GUI implementing MSBooster with Percolator rescoring in comparison to PeptideProphet command line quantified cysteine identifications from validation datasets.

Methods

Biology:

Cell culture and preparation of cell lysates. Cell culture reagents including Dulbecco's phosphate-buffered saline (DPBS), Dulbecco's modified Eagle's medium (DMEM)/high glucose media, Eagle's Minimum Essential Medium (EMEM), Roswell Park Memorial Institute (RPMI) media, trypsin-EDTA and penicillin/streptomycin (Pen/Strep), and Horse Serum, heat inactivated (26-050-070) was purchased from Fisher Scientific. Fetal Bovine Serum (FBS) was purchased from Avantor Seradigm (lot # 214B17). All cell lines were obtained from ATCC and were maintained at a low passage number (< 20 passages). HEK293T (ATCC: CRL-3216) cells were cultured in DMEM supplemented with 10% FBS and 1% antibiotics (Penn/Strep, 100 U/mL). MIA-PaCa-2 (ATCC: CRL-1420) cells were

cultured in DMEM supplemented with 10% FBS, 1% antibiotics (Penn/Strep, 100 U/mL), and 2.5% horse serum. H661 (ATCC: HTB-183), H1437 (ATCC: CRL-5872), H358 (ATCC: CRL-5807), HCT-15 (ATCC: CCL-225), Jurkat (ATCC: TIB-152), MOLT-4 (ATCC: CRL-1582) and H2122 (ATCC: CRL-5985) cells were cultured in RPMI-1640 supplemented with 10% FBS and 1% antibiotics (Penn/Strep, 100 U/mL). HEC-1-B (ATCC: HTB-113), MeWo (ATCC: HTB-65), CaCo-2 (ATCC: HTB-37) cells were cultured in EMEM supplemented with 10% FBS and 1% antibiotics (Penn/Strep, 100 U/mL). Cells were maintained in a humidified incubator at 37 °C with 5% CO₂. Cells were harvested by centrifugation (4500g, 5 min, 4 °C) and washed twice with cold DPBS. Cell pellets were then lysed with sonication (amp=10, 10 x1 sec pulses). The lysates were then transferred to a new microcentrifuge tube. Protein concentrations were determined using a BioRad DC protein assay kit from Bio-Rad Life Science (5000113, 5000114) and the lysate diluted to the working concentrations indicated below.

RNA-seq variant calling. Total RNA was extracted from cells using the Invitrogen Purelink RNeasy Plus Mini Kit (Qiagen, 166043750) or PureLink RNA mini kit (ThermoFisher, 12183018A) or Library preparation and RNA sequencing was carried out by the UCLA Technology Center for Genomics and Bioinformatics (TCGB). Libraries were prepared using the KAPA stranded mRNA kit. Paired-end sequencing (2x150) was performed to a depth of 50-60x with an Illumina HiSeq3000 system. RNAFastq paired-end reads for each cell line were aligned to Gencode reference genome hg38 (GRCh38.p13) using STAR-2 PASS alignment (v2.7.3a)². We ran STAR with default settings for paired-reads and the following additional parameters:–outSAMtype BAM

SortedByCoordinate, --outSAMunmapped Within, and --sjdbFileChrStartEnd in second alignment. Samtools (v1.7)^{3,4} calmd was used to add MD tags to sorted BAM files. Opposum (v0.2, 02-23-2017)⁵ was used to split reads and mark duplicates using default parameters and an additional parameter --SoftClipsExist True. Platypus (v0.8.1)⁶ was used to call variants and generate VCF files using default parameters. Samtools flagstat was used to obtain BAM file mapped read counts. Raw reads submitted to Sequence Read Archive (SRA) as BioProject PRJNA997729.

WE-seq variant calling. Genomic DNA was extracted from cells using the Zymo Quick DNA Miniprep Plus Kit kit (Fisher Sci, 50-444-149). Library preparation and exome-sequencing was carried out by the UCLA Technology Center for Genomics and Bioinformatics (TCGB). Libraries were prepared using the Nimblegen Capturing Kit. Paired-end sequencing (2x150) was performed to a depth of 50-60x with an Illumina HiSeq3000 system. Fastq paired-end reads for each cell line were aligned to Gencode reference genome hg38 (GRCh38.p13) using BWA-MEM alignment and default parameters for paired-reads. Output SAM files were converted to BAM files using Samtools (v1.7) and Picard (v2.21.4) (<https://broadinstitute.github.io/picard/>) was used to generate coordinate-sorted BAM files with read groups added. Samtools was used to index the files and duplicates were marked with Picard. GATK-HaplotypeCaller (v4.1.8.1)⁷ was used to split reads. Since we do not have matched normal samples, we opted to use the germline caller GATK-HaplotypeCaller for exome data. Variants were called using default parameters with the exception of the ploidy option which was set to the value outlined in **Table S2** (tab S11). GATK was used to index the VCF file and filter the variants

using the following parameters: `-window 35 -cluster 3 --filter-name FS --filter-expression "FS > 30.0" --filter-name QD --filter-expression "QD < 2.0"`. Samtools flagstat was used to obtain BAM file mapped read counts. Raw reads submitted to Sequence Read Archive (SRA) as BioProject PRJNA997729.

HCT-15 expression analysis. 5 biological replicates RNA extracts were sequenced as described and aligned to hg38 as described in *RNA-seq variant calling*. Kallisto (v0.46.1)⁸ was used to estimate transcript counts with indexed Gencode v28 transcriptome (gencode.v28.transcripts.fa) and `-b` (bootstrap) set to 100. Abundance transcript files were normalized with DE-seq2 (v1.28.1)⁹. Counts table was subsetted to a curated set of nonredundant CCDS transcript ID's (24,950) in **Table S2** (tab S9) and mean counts were calculated for downstream analysis. Raw reads submitted to Sequence Read Archive (SRA) as PRJNA997729.

Predicting amino acid changes: Table S2 (S9) (nonredundant CCDS transcript ID's) was used to remove redundant proteins. VCFs from variant calling pipelines for both RNA and WES were processed using R package 'Variant Annotation'¹. First, a TxDB object was made using the Gencode v28 annotation GTF file. The 'predictCoding' function using genome hg38 (GRCh38.p13) was used to obtain protein level changes from the VCFs, and 'nonsynonymous' and 'nonsense' changes were extracted; the resulting table includes a set of internal transcript IDs labeled 'TXID'. A database of common SNPs from NCBI (04-23-2018 00-common_all.vcf.gz) was used to annotate SNPs from rare mutations. The output missense table (**Table S2** (tab S10)) lists reference/variant codons

and amino acids. This table was filtered to contain matches to the CCDS set of 24,950 Ensembl transcript IDs only and those that resulted in single amino acid variants (SAAVs), ignoring small indels and multi-nucleotide variants (48,552 variants). Variants *passing* variant-calling filters were used in Figures 1-2 (48,301 variants) and non-PASS variants are included in the proteomics analyses following.

Generation of sample-specific custom databases with all combinations of variants.

Several R packages were used in generating custom databases: *VariantAnnotation*¹, *GenomicFeatures*¹⁰, *biomaRt*¹¹ and *BSgenome.Hsapiens.UCSC.hg38* (10.18129/B9.bioc.BSgenome.Hsapiens.UCSC.hg38). A curated set of CCDS transcript ID's (24,950) (**Table S2**) was used to subset the Gencode v28 protein coding translations FASTA file by Ensembl transcript IDs. These sequences consist of a non-redundant UniProtKB¹² subset of cross-referenced CCDS proteins. Using the previously generated TxDB object from '*Predicting amino acid changes*', and the *biomaRt* select function, corresponding TXID headers for the protein FASTA file were obtained by selecting 'TXID' with Ensembl transcript ID keys ('TXNAME'). Matching TXIDs from all SAAVs with new protein sequence TXIDs, positions in the corresponding wild-type protein sequences were replaced with the corresponding variant amino acid to generate a list of protein sequences containing only one variant per sequence. Protein sequences containing variants shared between RNA and exome-derived variants were grouped as one sequence. For proteins containing multiple variants, all possible combinations were generated for variants within 30 amino acids windows for proteins with 25 (or 15, see SI tables) or fewer total variants. Output sequences were written to a FASTA file with headers containing corresponding

Uniprot-ID, Gene ID, Ensembl transcript ID, and missense changes as well as cell-line, and sequencing origin (RNA or WE). To limit the increased search space, the database variant protein sequences were *in-silico* digested. A custom python script was used to *in-silico* digest the FASTA to generate tryptic peptides containing 2 misscleavages (two tryptic sites flanking amino acids surrounding the individual variant). Any duplicated peptide sequences were removed to leave unique sequences. For compatibility with MSFragger-based searches, simplified FASTA headers were used containing only the Uniprot ID. Result peptides are mapped back to detailed FASTA files for variant information. Scripts are available at <https://github.com/hdesai17/chemoproteogenomics.git>.

Proteomic sample preparation for unenriched sample analysis. HCT-15 and MOLT-4 lysates were incubated in 2 M urea/PBS at RT (final concentration = 2 mg/mL). DTT (10 μ L of 200 mM stock in water, final concentration = 10 mM) was added into each sample and the sample was incubated at 65 °C for 15 min. To this, iodoacetamide (10 μ L of 400 mM stock in water, final concentration = 20 mM) was added and the solutions were incubated for 30 min at 37 °C. Following addition of 3 μ L trypsin solution (Worthington Biochemical, LS003740, 1 mg/mL in 666 μ L of 50 mM acetic acid and 334 μ L of 100 mM CaCl₂, final weight = 2 ng), digest was allowed to proceed overnight at 37 °C with shaking. The next day, 90 μ L from each digest was combined with 210 μ L water and 0.3 μ L TFA (final concentration ~0.1% TFA and ~180 μ g peptides). Samples were fractionated into low-bind eppendorf tubes using a high-pH reversed phase fractionation kit (Pierce, 84868). Fractions were dried (Speed Vac) then reconstituted with 15 μ L 5% acetonitrile

and 1% FA in MB water and analyzed by LC-MS/MS. Samples were fractionated in triplicates for a total of 48 samples.

Proteomic sample preparation for cysteine-enrichment sample analysis. Proteome samples (200 μ L of 1 mg/mL, prepared as described in preparation of cell lysates) Samples were then labeled with 2 mM IAA (2 μ L of 200 mM stock solution in DMSO, final concentration = 2 mM) for 1h at RT (700rpm). CuAAC was performed with biotin azide (2) (4 μ L of 200 mM stock in DMSO, final concentration = 4 mM), TCEP (4 μ L of fresh 50 mM stock in water, final concentration = 1 mM), TBTA (12 μ L of 1.7 mM stock in DMSO/tbutanol 1:4, final concentration = 100 μ M), and CuSO₄ (4 μ L of 50 mM stock in water, final concentration = 1 mM) for 1h at ambient temperature. After CuAAC, 10 μ L of 20% SDS was added to each sample. Samples were incubated with 0.5 μ L benzonase (Fisher Scientific, 70-664-3) for 30 min at 37°C. The samples were then subjected to SP3 sample loading, SP3 digest and elution, NeutrAvidin enrichment and LC MS/MS analysis, as described below. Experiments were conducted in duplicate for each cell line.

Proteomic sample preparation for ligandability screening. HCT-15, MOLT-4, and MeWo proteome samples (200 μ L of 2 mg/mL, prepared as described in preparation of cell lysates). Compound (500 μ M) or DMSO vehicle was added to lysates for 1 hr (2 μ L 50 mM stocks or 2 μ L DMSO). Samples were chased with 2 mM IAA (2 μ L of 200 mM stock solution in DMSO, final concentration = 2 mM) for 1hr. CuAAC was performed with heavy biotin azide (DMSO samples) or light biotin azide (compound labeled samples) (4 μ L of 200 mM stock in DMSO, final concentration = 4 mM), TCEP (4 μ L of fresh 50 mM

stock in water, final concentration = 1 mM), TBTA (12 μ L of 1.7 mM stock in DMSO/tbutanol 1:4, final concentration = 100 μ M), and CuSO₄ (4 μ L of 50 mM stock in water, final concentration = 1 mM) for 1h at ambient temperature. After CuAAC, 10 μ L of 20% SDS was added to each sample. Samples were incubated with 0.5 μ L benzonase (Fisher Scientific, 70-664-3) for 30 min at 37 °C. The samples were then subjected to *SP3 sample loading* using 80 μ L total bead volumes, *SP3 digest and elution*, *NeutrAvidin enrichment* and *LC MS/MS analysis*, as described below. Experiments were conducted in triplicate for each compound per cell line.

SP3 sample loading. SP3 sample cleanup was performed generally at a bead/protein ratio of 10:1 (wt/wt) (38). For each 200 μ L sample, 20 μ L (or 40 μ L) Sera-Mag SpeedBeads Carboxyl Magnetic Beads, hydrophobic (GE Healthcare, 65152105050250, 50 μ g/ μ L, total 1 mg) and 20 μ L (or 40 μ L) Sera-Mag SpeedBeads Carboxyl Magnetic Beads, hydrophilic (GE Healthcare, 45152105050250, 50 μ g/ μ L, total 1 mg) were aliquoted into a single microcentrifuge tube and gently mixed. Tubes were then placed on a magnetic rack until the beads settled to the tube wall, and the supernatants were removed. The beads were removed from the magnetic rack, reconstituted in 1 mL of MB water, and gently mixed. Tubes were then returned to the magnetic rack, beads allowed to settle, and the supernatants removed. Washes 20 were repeated for two more cycles, and then the beads were reconstituted in 40 μ L MB water. The bead slurries were then transferred to the proteome samples, incubated for 10 min at RT with shaking (1000 rpm).

SP3 digest and elution. Absolute ethanol (400 μ L) was added to each sample, and the samples were incubated for 5 min at RT with shaking (1000 rpm). Beads were washed twice with 80% ethanol as described above. Beads were then resuspended in 200 μ L 0.5% SDS in PBS containing 2 M urea. DTT (10 μ L of 200 mM stock in water, final concentration = 10 mM) was added into each sample and the sample was incubated at 65 $^{\circ}$ C for 15 min. To this iodoacetamide (10 μ L of 400 mM stock in water, final concentration = 20 mM) was added and the solution was incubated for 30 min at 37 $^{\circ}$ C with shaking. After that, absolute ethanol (400 μ L) was added to each sample, and the samples were incubated for 5 min at RT with shaking (1000 rpm). Beads were then again washed three times with 80% ethanol in water (400 μ L). Next, beads were resuspended in 150 μ L PBS containing 2 M urea followed by addition of 3 μ L trypsin solution (Worthington Biochemical, LS003740, 1 mg/mL in 666 μ L of 50 mM acetic acid and 334 μ L of 100 mM CaCl_2 , final weight = 2 ng). Digest was allowed to proceed overnight at 37 $^{\circ}$ C with shaking. After digestion, \sim 4 mL acetonitrile (> 95% of the final volume) was added to each sample and the mixtures were incubated for 10 min at RT with shaking (1000 rpm). Supernatants were then removed and discarded using the magnetic rack, and the beads were washed (3 \times 1 mL acetonitrile). Peptides were then eluted from SP3 beads with 100 μ L of 2% DMSO in MB water for 1 hour at 37 $^{\circ}$ C with shaking (1000 rpm). The elution was repeated again with 100 μ L of 2% DMSO in MB water. Peptide concentration assay (Pierce, 23275) was performed to test the concentration of the peptide. The elution can be used for NeutrAvidin enrichment or analyzed by LC-MS/MS.

NeutrAvidin enrichment of labeled peptides. For each sample, 50 μ L of NeutrAvidin® Agarose resin slurry (Pierce, 29200) was washed three times in 10 mL IAP (immunoaffinity purification) buffer (50 mM MOPS–NaOH (pH 7.2), 10 mM Na₂HPO₄, 50 mM NaCl) and then resuspended in 500 μ L IAP buffer. Peptide solutions eluted from SP3 beads were then transferred to the NeutrAvidin® Agarose resin suspension, and the samples were then rotated for 2h at RT. After incubation, the beads were pelleted by centrifugation (21,000 g, 1 min) and washed by centrifugation (3 \times 1 mL PBS, 6 \times 1 mL water). Bound peptides were eluted with 60 μ L of 80% acetonitrile in MB water containing 0.1% FA (10 min at RT). The samples were then collected by centrifugation (21,000 g, 1 min) and residual beads separated from supernatants using Micro BioSpin columns (Bio-Rad). The remaining peptides were then eluted from pelleted beads with 60 μ L of 80% acetonitrile in water containing 0.1% FA (10 min, 72 °C). Beads were then separated from the eluants using the same Bio-Spin column. Eluents were collected by centrifugation (21,000 g, 1 min) and dried (SpeedVac). The samples were then reconstituted with 5% acetonitrile and 1% FA in MB water and analyzed by LC-MS/MS.

Liquid-chromatography tandem mass-spectrometry (LC-MS/MS) analysis. The samples were analyzed by liquid chromatography tandem mass spectrometry using a Thermo Scientific™ Orbitrap Eclipse™ Tribrid™ mass spectrometer coupled with a High Field Asymmetric Waveform Ion Mobility Spectrometry (FAIMS) Interface. Peptides were resuspended in 5% formic acid and fractionated online using a 18cm long, 100 μ m inner diameter (ID) fused silica capillary packed in-house with bulk C18 reversed phase resin (particle size, 1.9 μ m; pore size, 100 Å; Dr. Maisch GmbH). The 70-minute water

acetonitrile gradient was delivered using a Thermo Scientific™ EASY-nLC™ 1200 system at different flow rates (Buffer A: water with 3% DMSO and 0.1% formic acid and Buffer B: 80% acetonitrile with 3% DMSO and 0.1% formic acid). The detailed gradient includes 0 – 5 min from 3 % to 10 % at 300 nL/min, 5 – 64 min from 10 % to 50 % at 220 nL/min, and 64 – 70 min from 50 % to 95 % at 250 nL/min buffer B in buffer A. For bulk fractionation data, the detailed 80 min gradient includes 0 – 3 min from 1 % to 10 % at 300 nL/min, 3 – 63 min from 10 % to 40 % at 220 nL/min, 63 – 73 min from 40 % to 50 % at 220 nL/min, and 73 – 80 min from 50 % to 95 % at 250 nL/min buffer B in buffer A. Data was collected with charge exclusion (1, 8,>8). Data was acquired using a Data-Dependent Acquisition (DDA) method comprising a full MS1 scan (Resolution = 120,000) followed by sequential MS2 scans (Resolution = 15,000) to utilize the remainder of the 1 second cycle time. Time between master scans was set 1 s and 3s for compound labeling datasets, validation datasets, and fractionation datasets. HCD collision energy of MS2 fragmentation was 30 %. Raw file names used for figures are in **Table S9**.

Command-line MSFragger-based variant peptide identification and quantitation.

Raw data collected by LC-MS/MS were searched using a 2-stage search scheme implemented using custom bash scripts: MSFragger (version 3.5), Philosopher (version 4.2.2) and IonQuant (version 1.8.0) enabled^{13–16}. Precursor and fragment mass tolerance was set as 20 ppm. Missed cleavages were allowed up to 2. Peptide length was set 7 - 50 and peptide mass range was set 500 - 5000. Cysteine residues were searched with variable modifications at cysteine residues for carboxyamidomethylation (+57.02146), biotin-azide (+463.2366), and heavy biotin-azide (+469.2742) added for quant searches

in Figure 5 datasets. Labeling was set allowing for 3 max occurrences and 'all mods used in first search' checked. Peptide and protein level FDR were set to 1%. For ligandability screening, permissive IonQuant parameters allowed minimum scan/isotope numbers set to 1. First, raw spectra are searched with normal reference protein sequences (CCDS set) and peptide to spectrum matches (PSM) are filtered to 1% FDR. Custom bash scripts were used to extract 1% FDR filtered PSM scan numbers from this first search. Prior to a second search using the same parameters and custom database, a text file of these scan numbers is generated with leading zeros removed and included as option 'excluded_scan_list_file', allowing remaining scans to be searched with a cell line-specific custom database containing Uniprot identifiers and tryptic peptide sequences as described in *Generation of sample-specific custom databases*. PeptideProphet¹⁷ was used for rescoring for both searches. Bash scripts are available at <https://github.com/hdesai17/chemoproteogenomics.git>.

FragPipe label-free quantitation of HCT-15 fractionation data. Raw data collected by LC-MS/MS were searched with the default LFQ-MBR workflow provided by FragPipe. With each experimental group corresponding to fractionation set for a total of three intensity values per protein in combined.protein.tsv output. Mean LFQ intensities were calculated and the non-redundant set of Uniprot IDs (**Table S2** (tab S9)) was used in downstream analyses.

The MS search results and fasta files have been deposited to the ProteomeXchange Consortium (<http://proteomecentral.proteomexchange.org>) via the PRIDE partner

repository¹⁸ with the dataset identifiers PXD043879 for newly generated data, and PXD023059 and PXD029500 for re-analyzed data.

FragPipe GUI with improved 2-stage search. FragPipe generates a "fragpipe-second-pass.workflow" and a "fragpipe-files-second-pass.fp-manifest" after the first search. The manifest file points to the calibrated mzML files generated from the first pass. The workflow file has mass calibration and optimization turned off. Thus, using those two files, the second-pass search skips the calibration and searches the calibrated data. See document on 2-stage searches for more details.

Transient expression of HLA-B alleles. Expression plasmids (pTwist CMV) containing HLA-B*38:01, and HLA-B*27:05, HLA-B*38:01 C91S ,and HLA-B*27:05 C91S inserts with C-terminal FLAG-tags were obtained from Twist Bioscience. pDONR223_B2M_WT was a gift from Jesse Boehm & William Hahn & David Root (Addgene plasmid # 81810; <http://n2t.net/addgene:81810>; RRID:Addgene_81810) and subcloned using GateWay cloning into C-terminal FLAG destination vector generated from a pRK5 backbone vector, which was a kind gift from T Wucherpfennig. Plasmids were co-transfected into 60% confluent 6cm plated 293T cells using 14µL PEI, 140 µL serum-free DMEM, and 1 µg co-transfections or 2 µg eGFP expression plasmid. Cells were harvested after 24 hour transfections

FLAG-IP and Gel based-ABPP of HLA-B alleles: cell surface and lysate labeling. Cells were washed once with PBS and resuspended in 100µL serum-free DMEM. One-

half of cells were rotated at RT in 200 μ M IAA (1 μ L of 10 mM IAA stock) for 1 hr for cell-surface labeling. After spinning down at 1800 xg, the supernatant was removed. Cells were lysed in 30 μ L 2% CHAPS/PBS for 30 min on ice. Remainder cells were lysed in 30 μ L 2% CHAPS/PBS. Dilute all samples to 300 μ L with PBS and spin 1800 xg for 1 min. Samples were adjusted to 2 mg/mL using a Bio-Rad DC protein assay kit from Bio-Rad Life Science (Hercules, CA). 200 μ L of unlabeled lysates were incubated with 200 μ M IAA (2 μ L of 20 mM IAA stock) for 1 hr RT. 50 μ L EZred FLAG bead suspension per sample (Sigma, F2426) were washed with tris-buffered saline (TBS) buffer according to manufacturer recommendations. 50 μ L washed beads were added to each sample and rotated for 2 hours at 4C. Beads were washed 3x with 500 μ L TBS pelleted at 8200 xg and resuspended in 50 μ L PBS with 250 μ g/mL 3x FLAG peptide (Sigma, F4799) and 0.2% NP-40 alternative (Millipore Sigma, 492016) and rotated for 30 min at 4C. Beads were pelleted at 8200 xg to capture eluted proteins. Eluant was clicked on to rhodamine-azide (Click Chemistry Tools, AZ109-5) (25 μ M rhodamine-azide (1.25 mM stock), 1 mM Tris(2carboxyethyl)phosphine (TCEP) (SigmaAldrich) (50 mM stock), 100 μ M Tris[(1-benzyl-1H-1,2,3-triazol-4-yl)methyl]amine (TBTA) (Sigma-Aldrich) (1.7 mM stock), 1 mM CuSO₄ (50 mM stock)) for 1 hour at RT. All samples were denatured (5 min, 95 °C) and loaded onto 4-12% Criterion™ XT Bis-Tris gels with XT MOPS running buffer from Bio-Rad followed by semi-dry transfer to nitrocellulose membrane. 1:2000 dilution of anti-FLAG rabbit antibody (14793, Cell Signaling) and followed by 1:5000 IRDye® 800CW Goat anti-Rabbit IgG (102673-330, VWR) as well as 1:3000 GAPDH rabbit antibody (2118S, Cell Signaling) followed by 1:5000 IRDye® 680RD Goat anti-Rabbit IgG was used for visualization of loading and rho signal for IAA labeling.

Data processing for cys-enrichment and quantitative ratio peptide analysis. Custom R scripts were implemented to compile modified_peptide_label_quant.tsv (quant) outputs from command line MSFragger pipeline or FragPipe to count unique quantified cysteines. Unique cysteines were quantified for each dataset using unique identifiers consisting of a UniProt protein ID and the amino acid number of the modified cysteine and an additional parameter specifying single or double isotopic labeling (heavy and/or light). Unique proteins were established based on UniProt protein IDs. Residue numbers were found by aligning the peptide sequence to the corresponding UniProt ID protein sequence specified by FragPipe outputs. Variant residue sites were obtained by mapping peptide sequences to respective custom FASTA files containing variant info in the headers. For enriched samples, nonspecific non-Cys containing peptides were omitted from analysis. For ratios data, methionine oxidized peptides were omitted, unpaired heavy or light-identified peptides were kept by setting ratios to $\log_2(20)$ or $\log_2(1/20)$. Outputs were generated by taking the median H:L ratio among all tryptic peptides for unique cysteines in replicate datasets (modified_peptide_label_quant.tsv); mean ratio values were calculated across replicate datasets; quantified cysteines appearing in at least two replicates with ratio SD ≤ 1 were kept (Figure 6 and Figure 4 data) and no SD filter was applied for Figure 3 data to interpret ratio skew. For comparisons to CysDB cystines, unique UniprotID_CysPosition identifiers were used and FragPipe assigned Protein ID was used; for ClinVar variant matching, chromosome position and nucleotide changes of associated variants were used. The MS search results and FASTA files have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset

identifiers PXD043879 for newly generated data and FASTA files. R and Python scripts are available at <https://github.com/hdesai17/chemoproteogenomics.git>.

Data processing for high-pH data analysis. Custom R scripts were implemented to compile peptide.tsv outputs from command line MSFragger pipelines. Unique proteins were established based on UniProt protein IDs. Variant residue sites were obtained by mapping peptide sequences to respective custom FASTA files containing variant info in the headers. Residue numbers were found by aligning the peptide sequence to the corresponding UniProt ID protein sequence specified by FragPipe outputs. For ClinVar variant matching, chromosome position and nucleotide changes of associated variants were used.

Linear sequence and spatial site-analysis. For linear sequence analysis in Figure 2-5 datasets, residues 'in or near' UniProtKB annotated active or binding sites, DisProt annotated sites¹⁹⁻²¹, or PhosphoSite annotated sites²² were within +/-10 amino acids in linear sequence. For Figure 6 datasets, residues 'in or near' UniProtKB annotated active or binding sites, disprot annotated sites, or phosphosite annotated sites were assessed using 3D Protein Data Bank (PDB) structures. PDB structures were parsed to find all neighboring residues within a 10 Angstrom distance of the liganded Cys (alpha carbon atom). PDB_UniProtKB identifiers were created for each cysteine and corresponding list of neighboring residues. If the UniProtKB annotated active or binding sites were resolved in an associated crystal structure and found within the 10 Angstroms net, it was classified as a cysteine proximal to a known active or binding site.

Data Analysis and Statistics:

Fig 1B, C, S1-S5: COSMIC Cell Lines Project Complete mutation data 'CosmicCLP_MutantExport.tsv.gz' was downloaded (release v96, 31st May 2022). For each cell line, the 'Substitution-Missense' amino acid changes were totaled up; the gained counts (example Cys, X→C) were subtracted from the lost counts (example Cys, C→X) to obtain net counts for each amino acid. The analysis was limited to accession numbers matching a curated set of non-redundant Ensembl transcript IDs (24,950) to limit any over/under counting due to mutations in multi-transcript proteins. **Fig S6:** ClinVar (3-28-23) data filtered for unique gene name, protein position, and amino acid change to calculate total gains and losses. **Fig 1D:** common SNPs from NCBI (04-23-2018 00-common_all.vcf.gz) were filtered for missense mutations (645,395) and further filtered to include only variants in the non-redundant Ensembl transcript set. Net changes were calculated from unique 'GeneID_amino acid change identifiers'.

Fig 2B,C,D,S24: Missense table in **Table S2** (tab S10) was filtered for unique 'CellLine_Protein_AAchange' identifiers, removing transcripts IDs that result in the same amino acid changes per protein. **Fig 2F,G,H,S11-S23:** CADD-PHRED score distributions were obtained by filtering the missense table for only single nucleotide base substitutions and uploading them as a vcf ('#CHROM', 'POS', 'ID', 'REF', 'ALT') to the CADD query tool <https://cadd.gs.washington.edu/score> using GRCh38-v1.6. Distribution of CADD scores shown by indicated grouping, unique 'CellLine_Protein-AAchange-CADD-score'

identifiers were used to remove variants identified in both RNA and exomes. **Fig 2I,2J:** Active site/binding site analysis: We consider single variant residues in or within +/- 10 amino acids in primary sequence from an annotated active/binding site range as annotated by Uniprot.

Fig 3C (and Fig 8B): Quantitative output obtained as described in *Data processing for cys-enrichment and quantitative ratio peptide analysis* for triplicate datasets. **Fig 3D:** From combined_ion_label_quant.tsv outputs for reference and variant searches, apex light - apex heavy retention times per ion were calculated for each replicate set. Mean values were obtained for ions appearing in at least two of three replicates.

Fig 4: Data was processed as described in methods for *Data processing for cys-enrichment* for duplicate datasets. **Figure S26:** Interpro ascension numbers and interpro family/domain names were pulled for all identified variant genes using EnsDb.Hsapiens.v86 from the *ensemldb* package²³ **Fig 4F:** The provided reference FASTA was *in-silico* digested to check for cysteines theoretically detectable using the FragPipe peptide length limits of 7-50 amino acids. Cysteine peptides with loss or gain of Arg/Lys were manually checked against theoretically detectable cysteines. **Fig 4J:** Site level analyses from Fig2. **Fig 4K:** SP3-Rox dataset were re-searched with 2-stage search and custom database; quantitative output obtained as described in data processing for triplicate datasets.

Fig 5: Data was processed as described in methods for *Data processing for high-pH data*

analysis. **Fig 5D:** For each cell line, protein sequences from the subsetted Gencode FASTA provided in supplemental information (24,950 proteins) and as described in *custom database generation* were used to calculate individual amino acid frequencies (# amino acid/total # amino acids) for every protein that contained a missense mutation. The total gained AA (identified in proteomic data) was used to calculate the gained frequencies (# gained AA/total missense changes identified). The fold enrichment was calculated by taking the ratio of the gained frequency to the AA frequency. **Fig 5E:** Transcript counts and LFQ data for 'high-ph' and 'all' categories were restricted to those containing at least one cysteine residue.

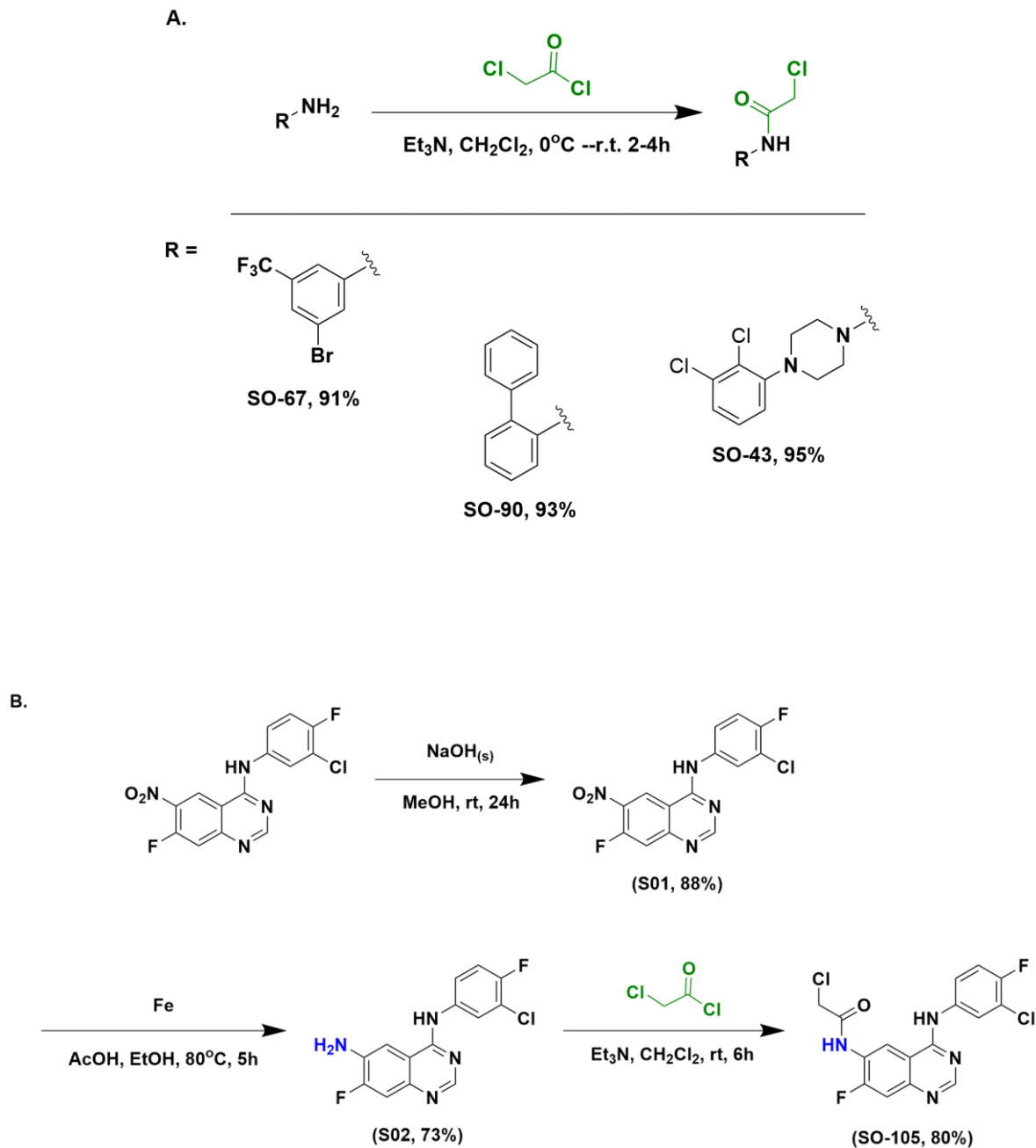
Fig 6: Quantitative output obtained as described in *Data processing for cys-enrichment and quantitative ratio peptide analysis*. **Fig 6D:** Ligandable cys residues were restricted to those appearing in more than one replicate, with Log2 ratios > 2, and with standard deviations of ratios across replicates ≤ 1 . **Fig 6C-G:** Values are restricted to those that appear in more than one replicate and with standard deviations ≤ 1 for both reference and variant Cys peptides.

Chemistry:

General Methods. All solution-phase reactions were performed in dried glassware under an atmosphere of dry N₂ except where water was used as a solvent. Silica gel P60 (SiliCycle) was used for column chromatography. Plates were visualized by fluorescence quenching under UV light or by staining with iodine. Other reagents were purchased from

Sigma-Aldrich (St. Louis, MO), Alfa Aesar (Ward Hill, MA), EMD Millipore (Billerica, MA), Fisher Scientific (Hampton, NH), Oakwood Chemical (West Columbia, SC), Combi-blocks (San Diego, CA) and Cayman Chemical (Ann Arbor, MI) and used without further purification. ^1H NMR and ^{13}C NMR spectra for characterization of new compounds and monitoring reactions were collected in CDCl_3 , CD_3OD , or DMSO-d_6 (Cambridge Isotope Laboratories, Cambridge, MA) on a Bruker AV 400 MHz spectrometer in the Department of Chemistry & Biochemistry at The University of California, Los Angeles. All chemical shifts are reported in the standard notation of parts per million using the peak of residual proton signals of the deuterated solvent as an internal reference. Coupling constant units are in Hertz (Hz). Splitting patterns are indicated as follows: br, broad; s, singlet; d, doublet; t, triplet; q, quartet; m, multiplet; dd, doublet of doublets; dt, doublet of triplets. Low-resolution mass spectroscopy was performed on an Agilent Technologies InfinityLab LC/MSD single quadrupole LC/MS (ESI source).

Scheme:



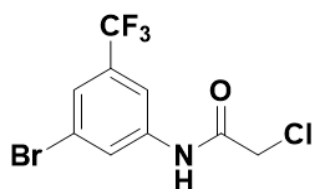
Scheme S1. Synthetic routes to obtain (A) electrophilic fragments SO-67, SO-90, and SO-43, and (B) prototype kinase inhibitor SO-105.

Synthesis

General Procedure 1:

To a solution of the amine (1 equiv.) and triethylamine (1.5 equiv.) in dichloromethane (5 mL), 2-chloroacetyl chloride (1.1 equiv.) was added dropwise and the reaction mixture was stirred at 0°C for 0.5 h then left to warm to room temperature (2 - 6h). The reaction progress was monitored via TLC. On reaction completion, the reaction mixture was poured slowly into distilled water (10ml). After separation of the phases, the organic layer was dried with anhydrous sodium sulfate, filtered and concentrated to obtain the crude product. This crude product was then coated on silica gel, then purified via flash chromatography (in 15-30% Ethyl acetate-hexanes) to yield the analytically pure product.

Synthesis of N-(3-bromo-5-(trifluoromethyl)phenyl)-2-chloroacetamide (SO-67)

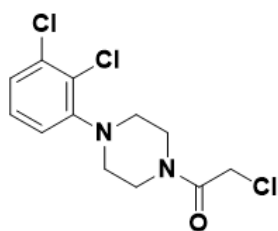


Prepared according to general procedure 1 using 3-bromo-5-(trifluoromethyl)aniline (200mg, 0.833 mmol) as the amine source. Product: white solid . **Yield:** 240 mg (91%)

¹H NMR (400 MHz, CDCl₃) δ 8.34 (s, 1H), 8.03 (s, 1H), 7.76 (s, 1H), 7.57 (s, 1H), 4.21 (s, 2H). **¹³C NMR** (100 MHz, CDCl₃) δ 164.09, 138.38, 133.17, 132.84, 132.51, 132.20, 132.13, 125.93, 124.93, 123.23, 115.40, 42.73. **¹⁹F NMR** (400 MHz, CDCl₃) δ -62.94.

HRMS (ESI-TOF) [M+H]⁺ = C₉H₇BrClF₃NO⁺ : calculated for 315.9305 ; Found 315.9352

Synthesis of 2-chloro-1-(4-(2,3-dichlorophenyl)piperazin-1-yl)ethan-1-one (SO-43)

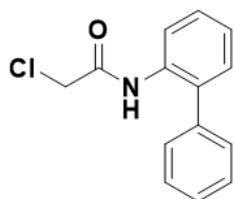


Prepared according to general procedure 1 using 1-(2,3-Dichlorophenyl)piperazine.HCl (300 mg, 1.12 mmol) as the amine source. Product: off-white solid. **Yield:** 95% (326mg).

¹H NMR (400 MHz, CDCl₃) δ 7.24 – 7.14 (m, 2H), 6.94 (dd, J = 7.8, 1.7 Hz, 1H), 4.12 (s, 2H), 3.77 (dt, J = 43.0, 4.9 Hz, 4H), 3.16 – 3.00 (m, 4H). **¹³C NMR** (101 MHz, CDCl₃) δ 169.96, 165.55, 150.36, 134.26, 127.63, 125.43, 118.86, 51.47, 50.92, 46.67, 42.47, 40.76.

HRMS (ESI-TOF) [M+H]⁺ = C₁₂H₁₄Cl₃N₂O⁺ : calculated for 307.0172 ; Found 307.0040

Synthesis of N-([1,1'-biphenyl]-2-yl)-2-chloroacetamide (SO-90)

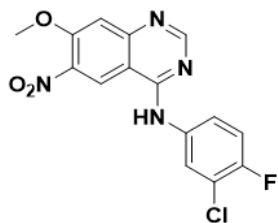


Prepared according to general procedure 1 using [1,1'-biphenyl]-2-amine (300 mg, 1.12 mmol) as the amine source. Product: off-white solid. **Yield:** 93% (410mg).

¹H NMR (400 MHz, CDCl₃) δ 8.46 (s, 1H), 8.36 (d, J = 8.2 Hz, 1H), 7.53 – 7.47 (m, 2H), 7.45 – 7.37 (m, 4H), 7.30 (dd, J = 7.6, 1.7 Hz, 1H), 7.24 (dd, J = 7.4, 1.2 Hz, 1H), 4.08 (s, 2H). **¹³C NMR** (101 MHz, CDCl₃) δ 163.64 , 137.43 , 133.92 , 132.65 , 130.14 , 129.15 , 128.22 , 124.98 , 120.64 , 43.03.

HRMS (ESI-TOF) [M+H]⁺ = C₁₄H₁₂ClNO⁺ : calculated for 246.0607 ; Found 246.0623

Synthesis of N-(3-chloro-4-fluorophenyl)-7-methoxy-6-nitroquinazolin-4-amine (S01)



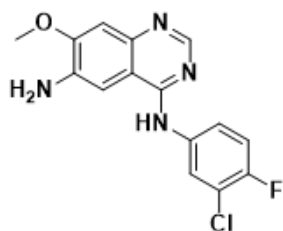
To the N-(3-chloro-4-fluorophenyl)-7-fluoro-6-nitroquinazolin-4-amine (1 g, 2.97 mmol, 1 eqv.) in methanol (23.7 mL, 551 mmol, 200 eqv.) was added sodium hydroxide pellets (1.2 g, 29.8 mmol, 10 eqv.). The reaction was refluxed overnight and monitored by TLC. Next, the reaction mixture was taken off the heat source and allowed to cool to room temperature. Then, the mixture was poured into a saturated solution of sodium bicarbonate, and then filtered under vacuum. The resultant residue was washed with water (1x), then methanol (200 mL). Then the solid residue was dried under high vacuum to yield the target compound as yellowish solid.. **Yield:** 88% (0.91g).

¹H NMR (400 MHz, DMSO-d₆) δ 9.09 (s, 1H), 8.51 (s, 1H), 8.05 (s, 1H), 7.71 – 7.64 (m, 1H), 7.43 – 7.28 (m, 2H), 4.02 (s, 3H). **¹³C NMR** (100 MHz, DMSO-d₆) δ 158.12, 154.93, 154.15, 152.39, 138.63, 124.23, 123.18, 122.64, 119.22, 116.72, 109.59, 57.41.

LC-MS (ESI) [M+H]⁺ = C₁₅H₁₁ClFN₄O₃⁺ : calculated for 348.04 ; Found 348.0

Synthesis of N⁴-(3-chloro-4-fluorophenyl)-7-methoxyquinazoline-4,6-diamine (S02)

Procedure: To a stirred mixture of N-(3-chloro-4-fluorophenyl)-7-methoxy-6-nitroquinazolin-4-amine (800 mg, 0.435 mol) in ethanol (100 mL) and AcOH (5 mL) was added Fe (641 mg, 11.5 mmol) and the reaction was heated to reflux. Further, more ethanol (50 mL) and AcOH (5 mL) were added, and reflux continued for 5 h, with reaction progress monitored by LC-MS. The reaction was cooled to room temperature, when deemed complete. Next, the crude solution was filtered by passing it through celite. The filtrate was then concentrated to about one-third its original volume. The resultant precipitate was isolated and dried under high vacuum to give the target compound as a

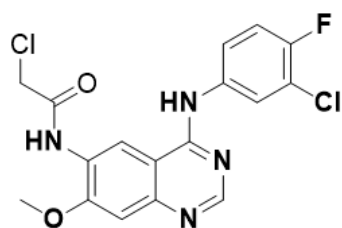


dark-brownish solid, (530mg, 73%) – confirmed by LC/MS – and this was used, as is, without further purification. **¹H NMR** (400 MHz, DMSO-*d*₆) δ 9.38 (s, 1H), 8.28 (d, J = 76.7 Hz, 2H), 7.81 (d, J = 8.9 Hz, 1H), 7.24 (d, J = 112.2 Hz, 3H), 5.37 (s, 2H), 3.96 (s, 3H). **¹³C**

NMR (101 MHz, DMSO-*d*₆) δ 153.20 , 150.73 , 137.99 , 122.92 , 106.36 , 101.30 , 56.25.

LC-MS (ESI) [M+H]⁺ = C₁₅H₁₃ClFN₄O⁺ : calculated for 318.01; Found 319.01

Synthesis of 2-chloro-N-(4-((3-chloro-4-fluorophenyl)amino)-7-methoxyquinazolin-6-yl) acetamide (SO-105)



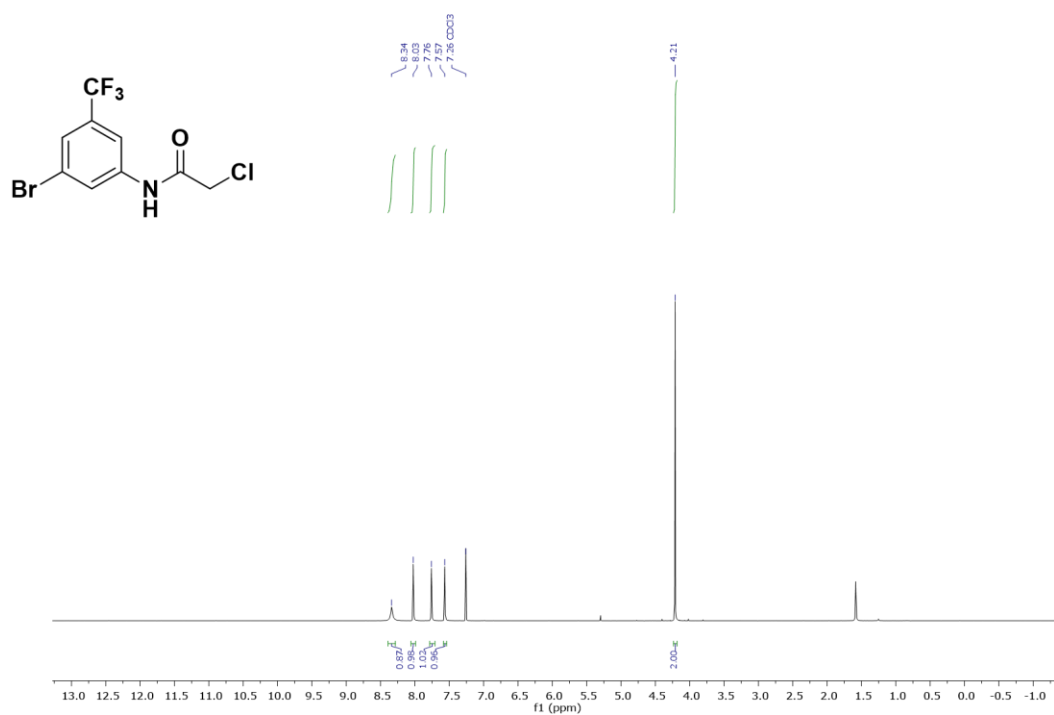
Prepared according to general procedure 1 using N⁴-(3-chloro-4-fluorophenyl)-7-methoxyquinazoline-4,6-diamine (500 mg, 1.57 mmol) as the amine source. Product: pale-brownish solid.

Yield: 80% (500mg).

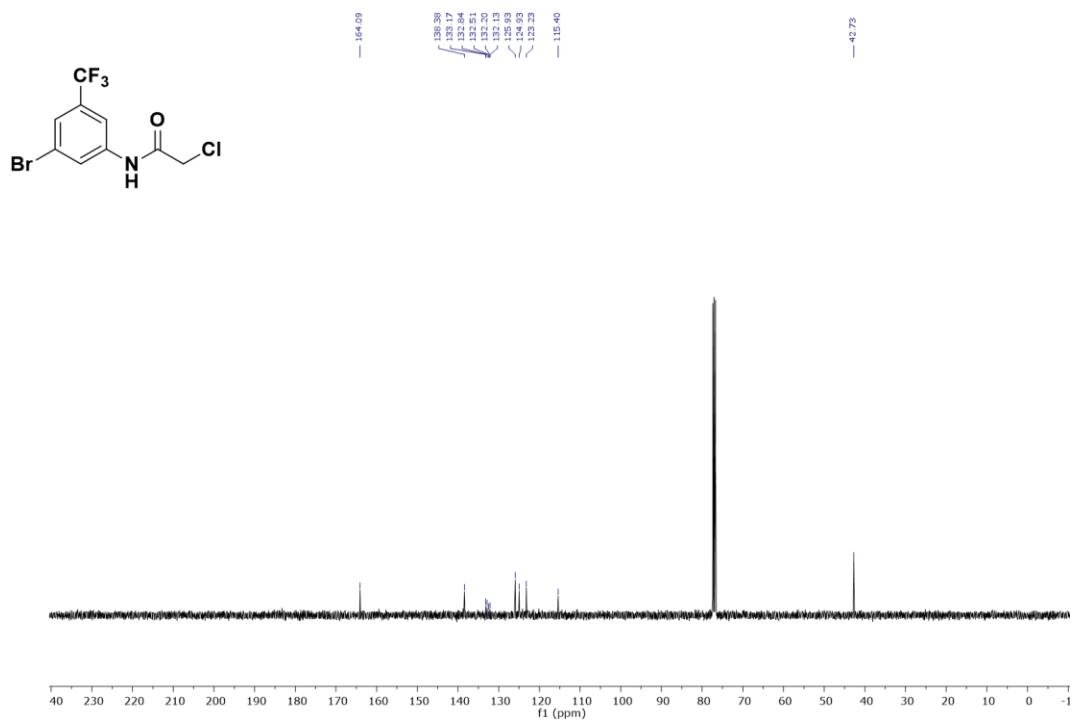
¹H NMR (400 MHz, DMSO-*d*₆) δ 9.99 (s, 1H), 8.94 (s, 1H), 8.63 (s, 1H), 8.06 (dd, J = 6.8, 2.6 Hz, 1H), 7.75 (dd, J = 6.7, 2.3 Hz, 1H), 7.45 (s, 1H), 7.34 (s, 1H), 4.49 (s, 2H), 4.27 (s, 1H), 4.04 (s, 3H). **¹³C NMR** (101 MHz, DMSO-*d*₆) δ 169.04, 165.68, 157.83, 155.99, 153.41, 152.99, 136.53, 127.56, 125.01, 123.77, 119.36, 116.89, 116.19, 108.83, 105.43, 56.97, 43.78.

HRMS (ESI-TOF) [M+H]⁺ = C₁₇H₁₄Cl₂FN₄O₂⁺ : calculated for 395.0478 ; Found 395.0389

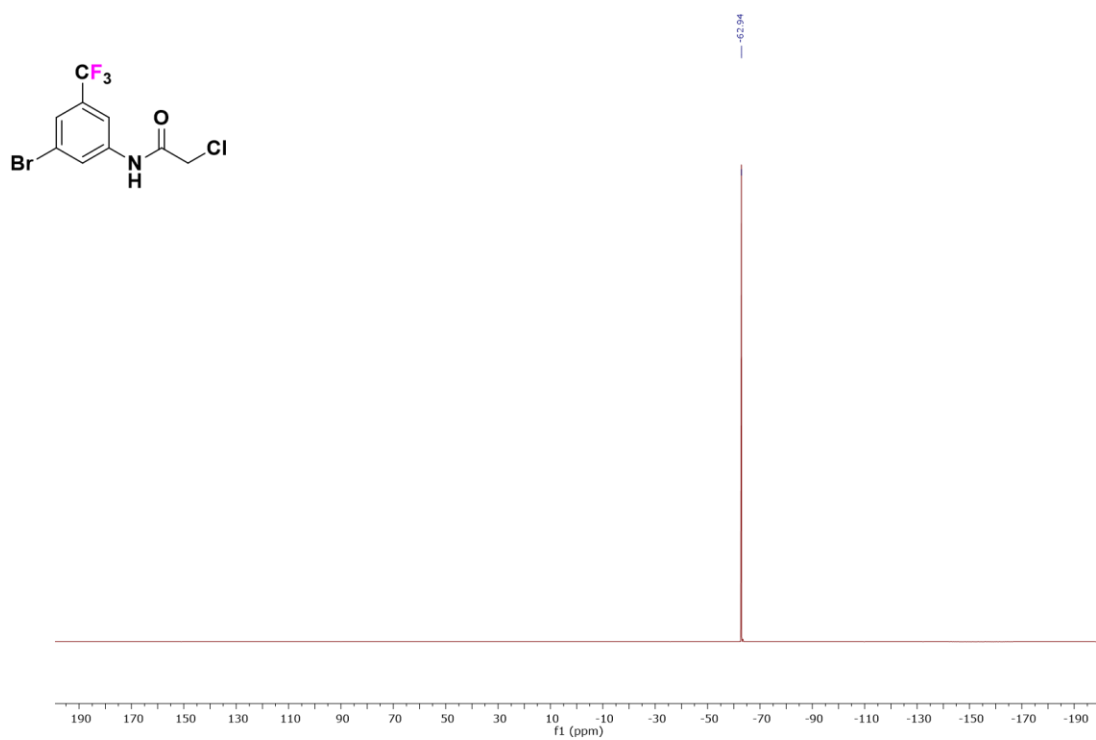
^1H NMR of N-(3-bromo-5-(trifluoromethyl)phenyl)-2-chloroacetamide, **SO-67** in CDCl_3



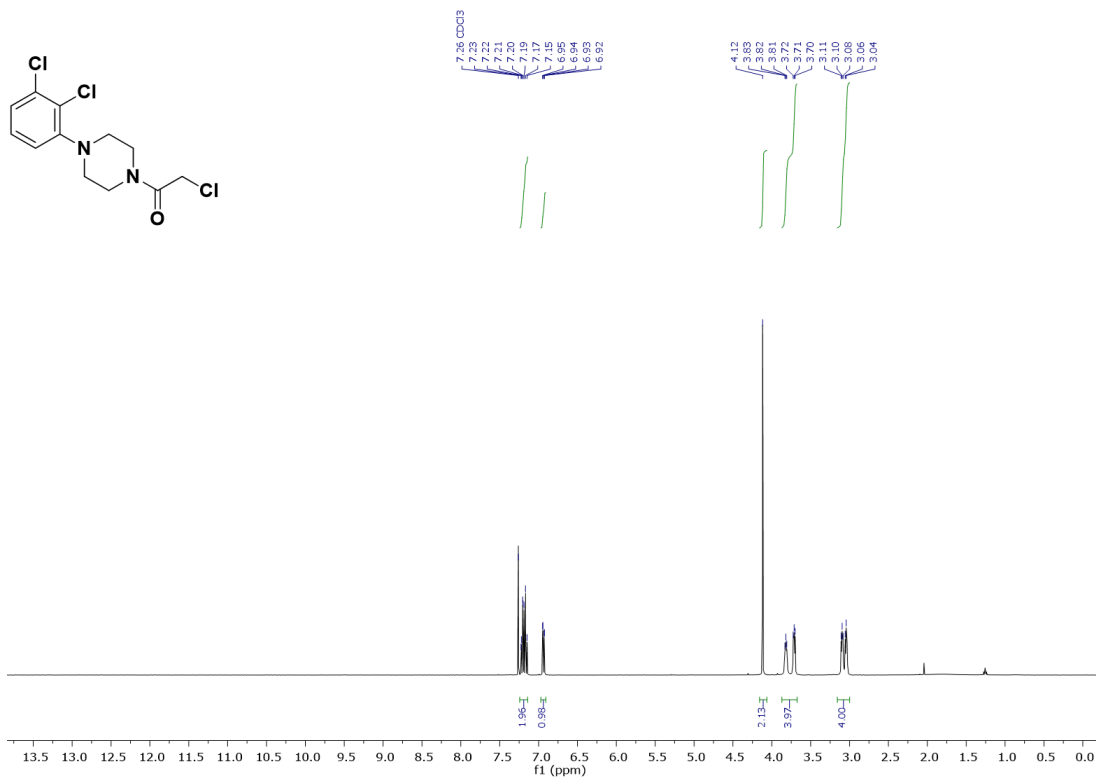
^{13}C NMR of N-(3-bromo-5-(trifluoromethyl)phenyl)-2-chloroacetamide, **SO-67** in CDCl_3



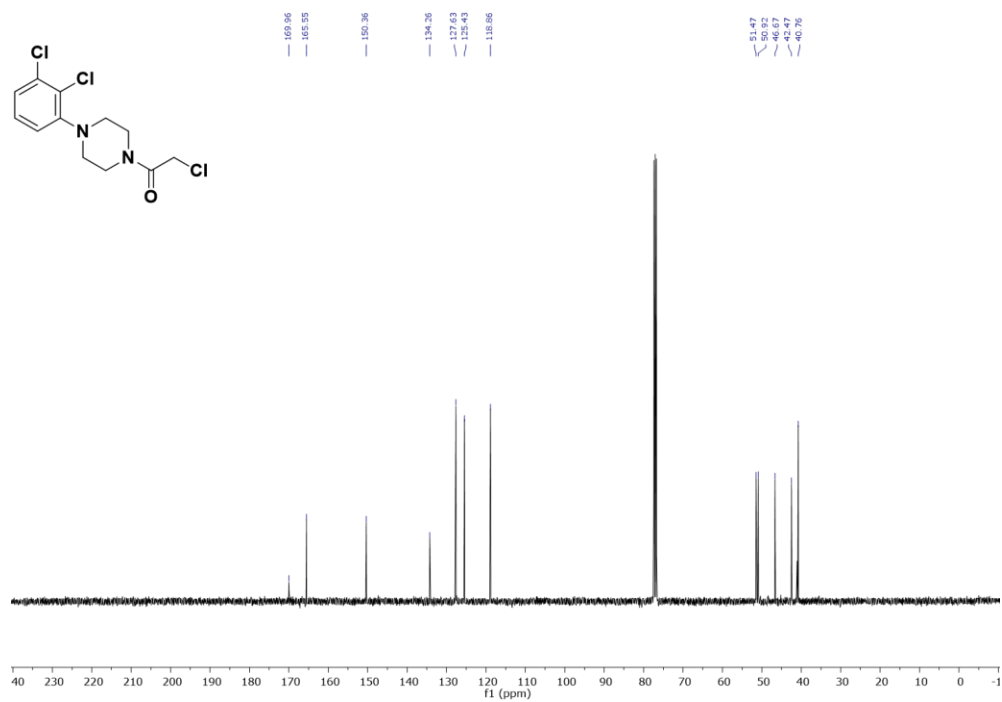
^{19}F NMR of N-(3-bromo-5-(trifluoromethyl)phenyl)-2-chloroacetamide, **SO-67** in CDCl_3



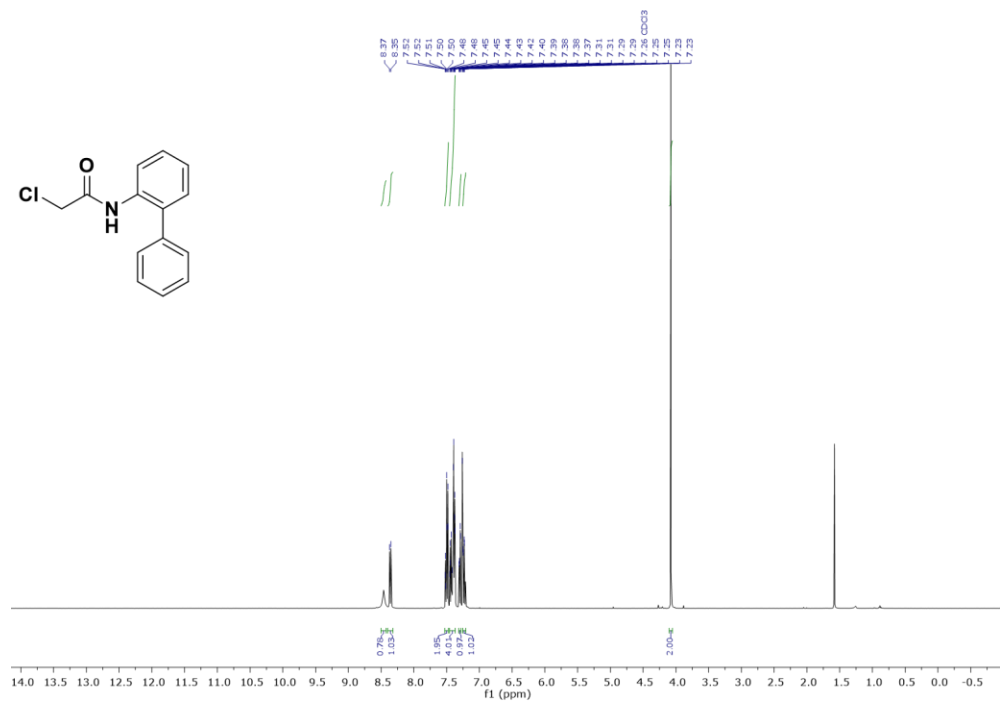
^1H NMR of 2-chloro-1-(4-(2,3-dichlorophenyl)piperazin-1-yl)ethan-1-one, **SO-43** in CDCl_3



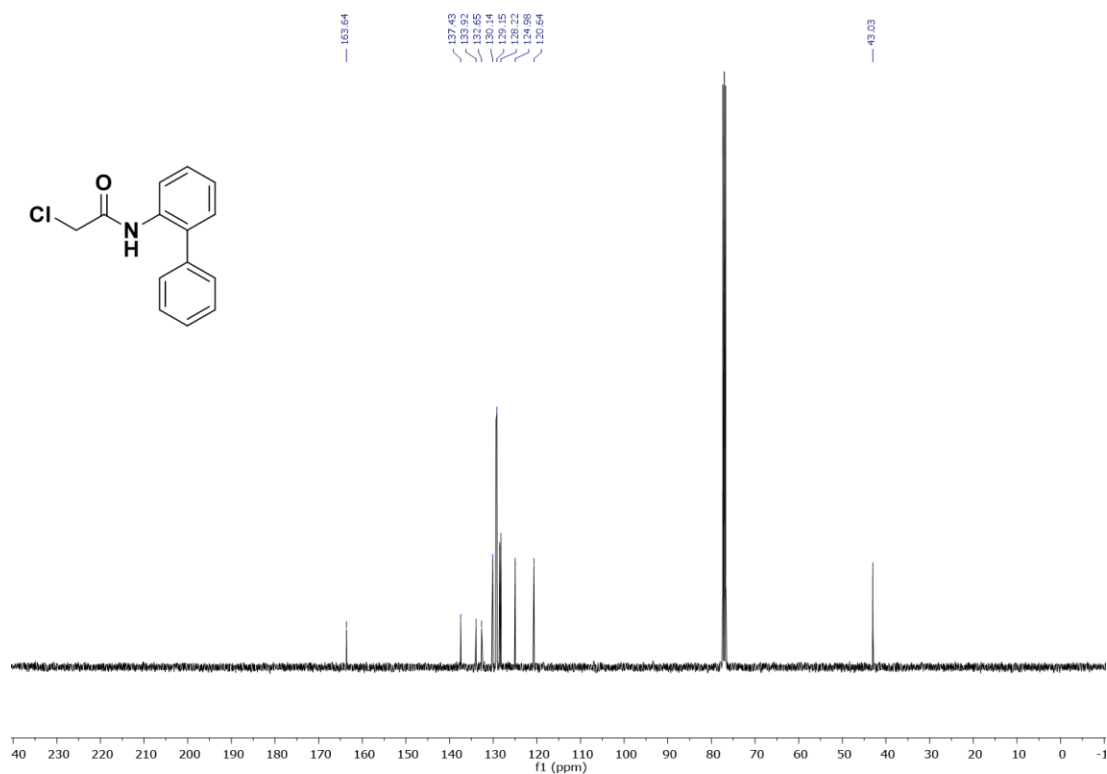
^{13}C NMR of 2-chloro-1-(4-(2,3-dichlorophenyl)piperazin-1-yl)ethan-1-one, **SO-43** in CDCl_3



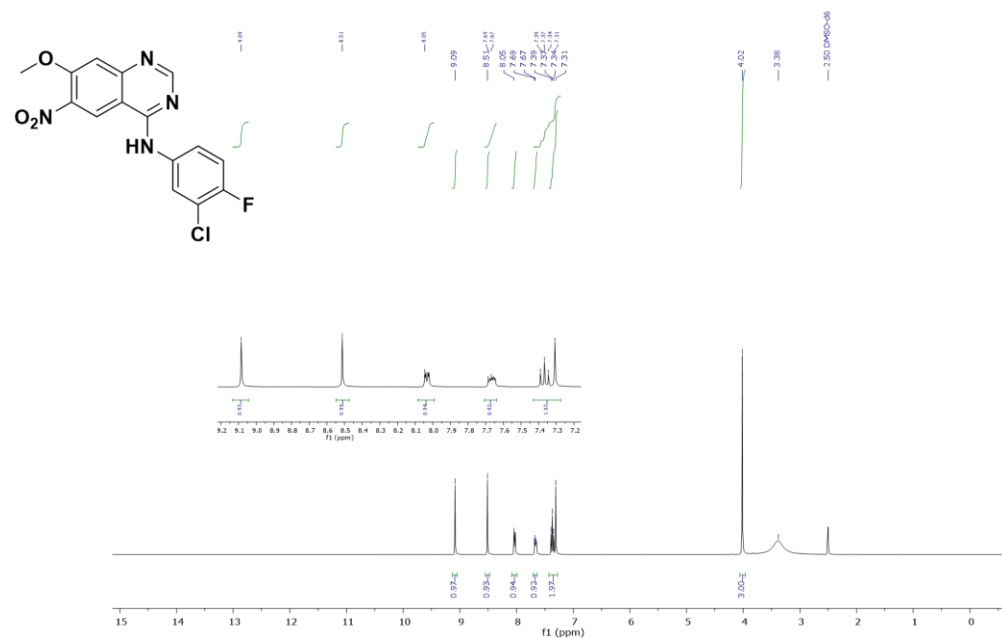
^1H NMR of N-([1,1'-biphenyl]-2-yl)-2-chloroacetamide, **SO-90** in CDCl_3



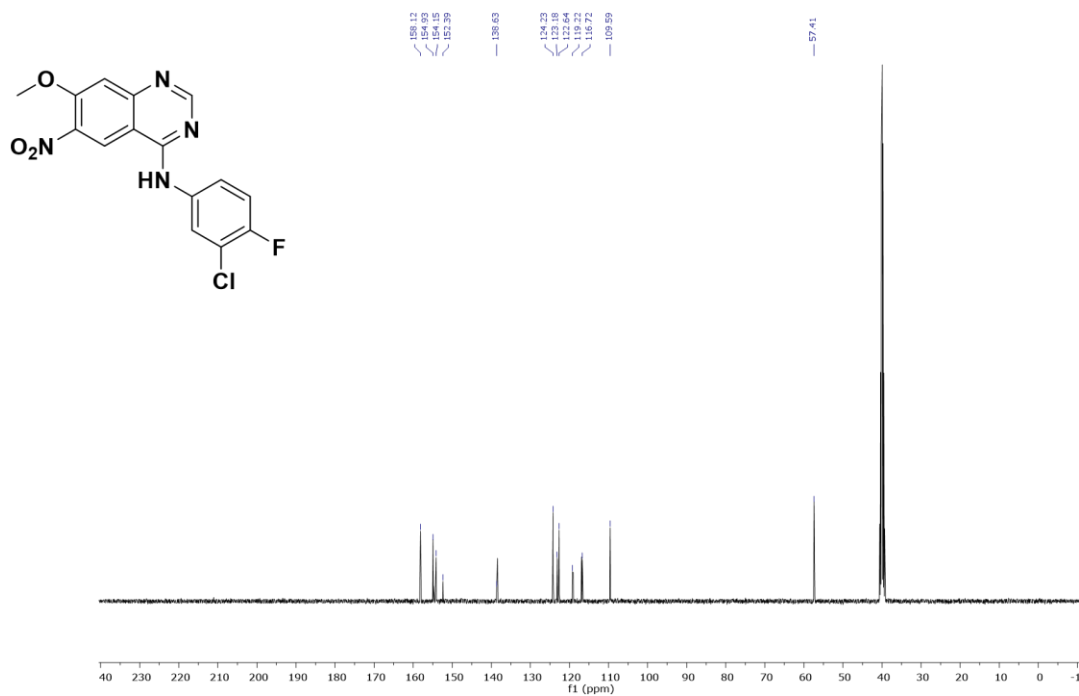
^{13}C NMR of N-([1,1'-biphenyl]-2-yl)-2-chloroacetamide, **SO-90** in CDCl_3



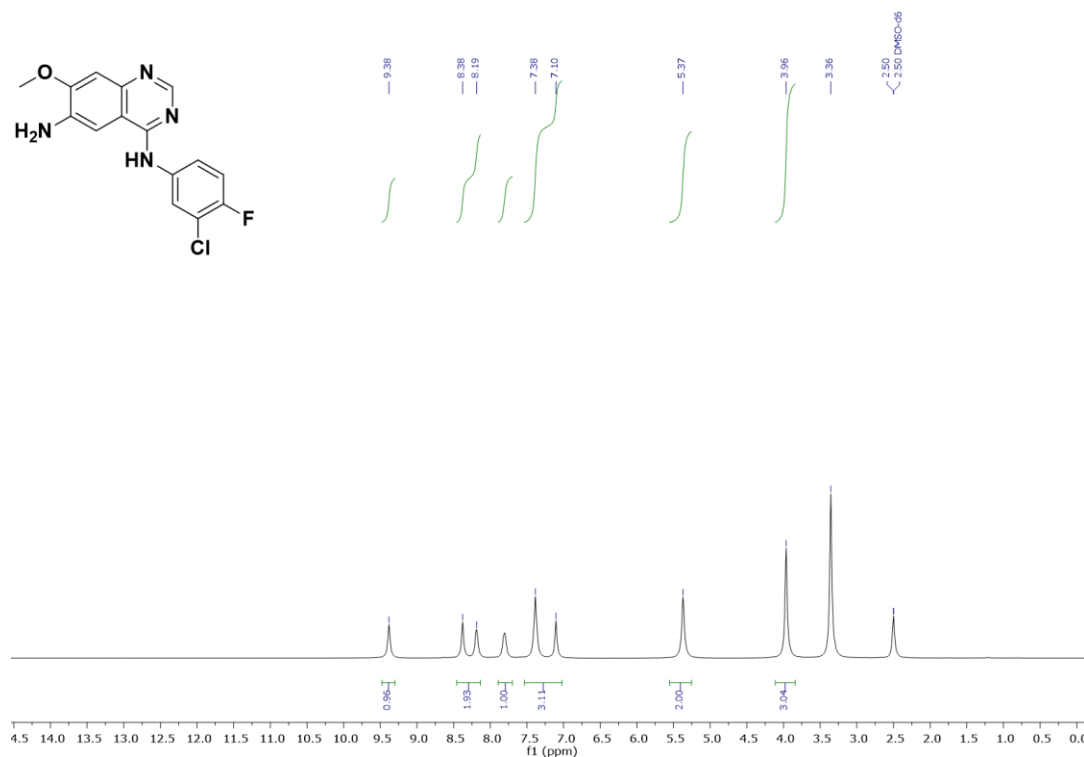
^1H NMR of N-(3-chloro-4-fluorophenyl)-7-methoxy-6-nitroquinazolin-4-amine (**S-01**) in CDCl_3



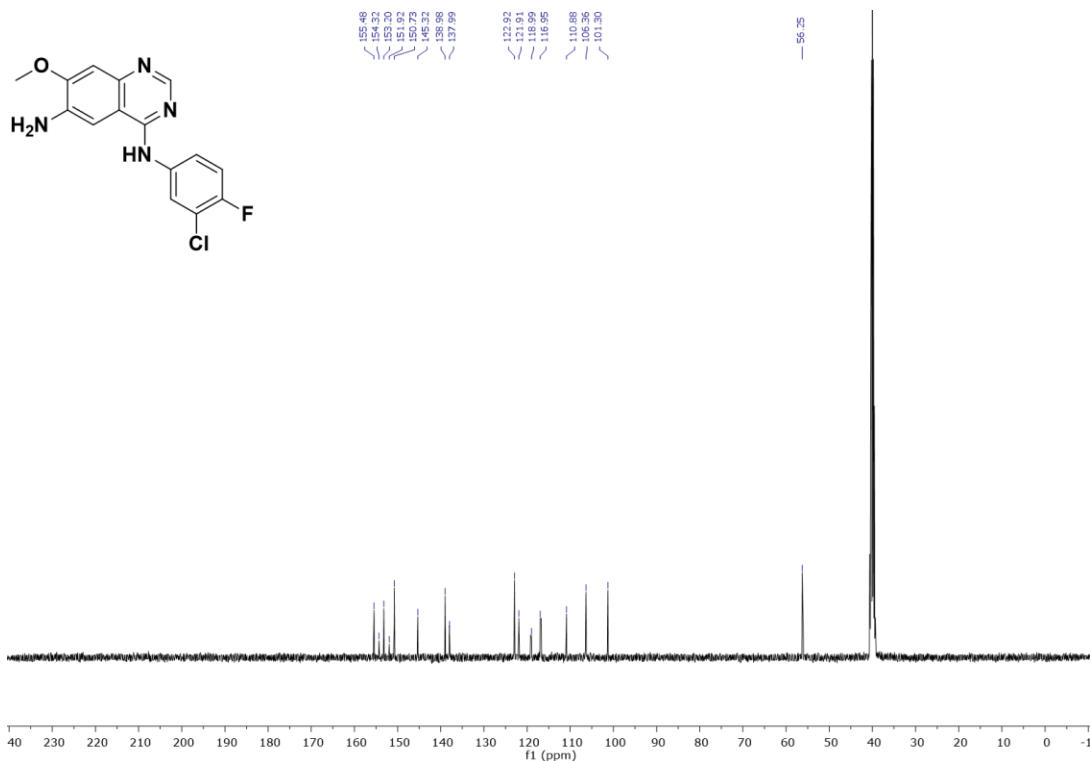
^{13}C NMR of N-(3-chloro-4-fluorophenyl)-7-methoxy-6-nitroquinazolin-4-amine (**S-01**) in CDCl_3



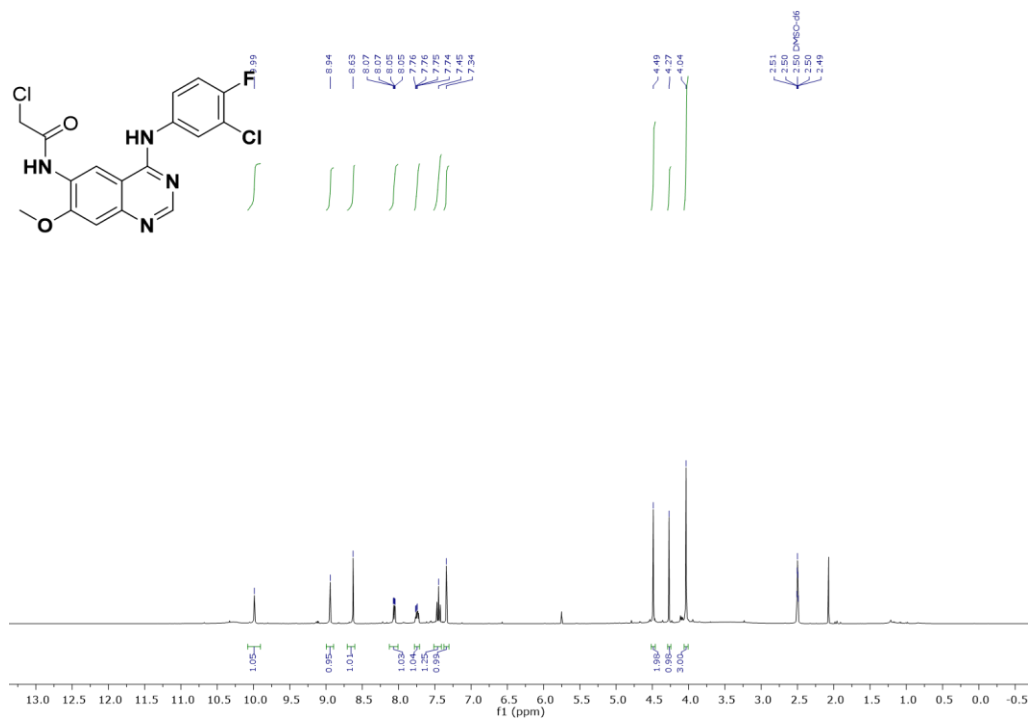
^1H NMR of N⁴-(3-chloro-4-fluorophenyl)-7-methoxyquinazoline-4,6-diamine (**S-02**) , in $\text{DMSO-}d_6$



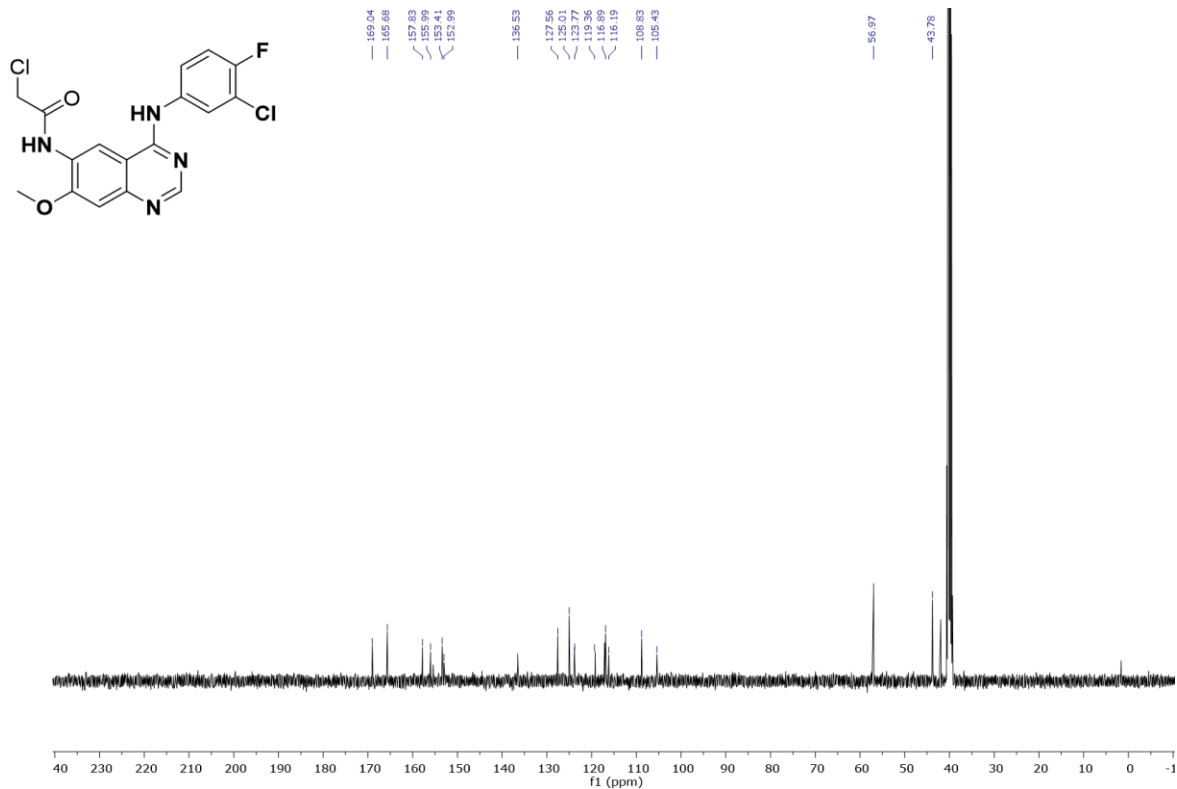
^{13}C NMR of N^4 -(3-chloro-4-fluorophenyl)-7-methoxyquinazoline-4,6-diamine(**S-02**), in $\text{DMSO-}d_6$



^1H NMR of 2-chloro-N-(4-((3-chloro-4-fluorophenyl)amino)-7-methoxyquinazolin-6-yl)acetamide, **SO-105**, in $\text{DMSO-}d_6$



¹³C NMR of 2-chloro-N-(4-((3-chloro-4-fluorophenyl)amino)-7-methoxyquinazolin-6-yl)acetamide, **SO-105**, in DMSO

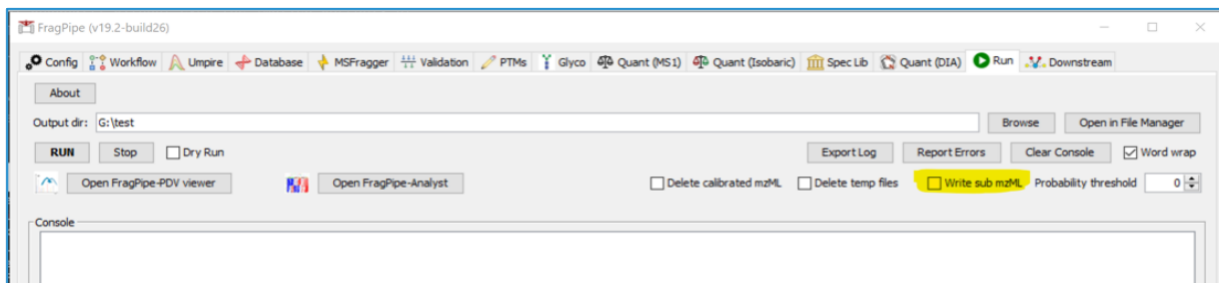


Running 2-stage search with FragPipe v20+

Make sure the tools are up-to-date

FragPipe v20.0+, MSFragger version 3.8+, Philosopher version 5.0.0+

1. Set up a search as normally done with the normal reference database and with experiment specific parameters (Percolator rescoring recommended).
2. Select 'Print decoys' in the Validation tab
3. In the last "Run" tab, there is a checkbox called "write sub mzML":



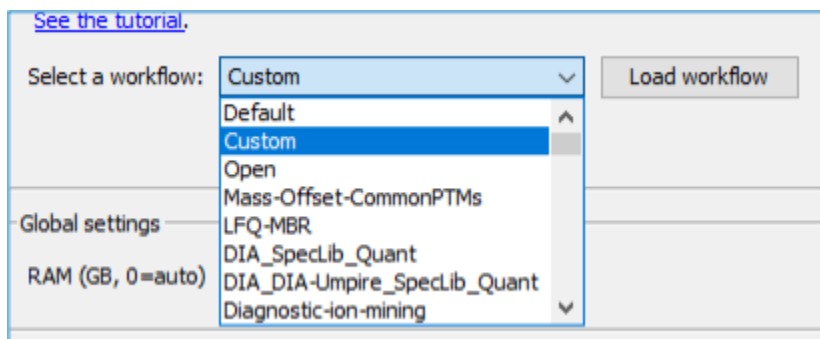
Check this box and leave the threshold set to 0.

When it is enabled, FragPipe runs the search, and writes new mzML files with "_sub.mzML" as the file name suffix, "fragpipe-second-pass.workflow", and "fragpipe-files-second-pass.fp-manifest" files to the specified result folder. In the first-pass search, FragPipe will write decoy PSMs to psm.tsv if they pass 1% FDR.

Both "fragpipe-second-pass.workflow" and "fragpipe-files-second-pass.fp-manifest" have adjusted parameters and mzML file paths that can be directly used in the second search.

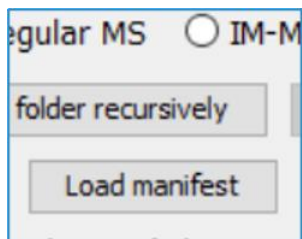
4. Name the result folder to distinguish it from the reference search ("first_pass", "reference", "1") and run the first search.

5. After the first search is finished, in your open FragPipe window, navigate to the workflow tab and load the "fragpipe-second-pass.workflow" located in your first results folder file by selecting the "custom" and clicking "load workflow". Select the fragpipe-second-pass.workflow which should then be automatically loaded.



Note: If using Percolator: Percolator weights saving and loading is automatic. FragPipe lets Percolator save the weight files during the first-pass search. And then, in the second-pass, FragPipe gives Percolator the weight files if it can find it (the weight files are in the sub mzML files' folder)

6. Clear all the .raw/mzML files you have loaded. Load the “fragpipe-files-second-pass.fp-manifest” by selecting load manifest and navigating to it in your first results folder. This should re-add your files with the “_sub.mzML” suffix.



Note: You should not remove/add any new modifications in the second search. The MSFragger parameters should be the same, just a new database and using _sub.mzML files

7. Go to the Database tab and navigate to the variant-containing database for your cell line.

8. Specify the reverse header (ex. “REV”).

Note: For databases generated with <https://github.com/hdesai17/chemoproteogenomics.git> tool, header information containing variant info will appear as Protein/Protein-ID/Protein.Description; other database formats will depend on FragPipe compatibility. For FASTA databases lacking

header information such as simple Uniprot IDs (>PXXXX), additional processing/mapping is required to obtain variant IDs.

9. Make sure the MSFragger parameters are the same as the first search.
10. Navigate to the “Run” tab, uncheck the write sub mzML box, and specify a new results folder to distinguish variant search (“second_pass” or “variant”, “2”) and run the second search.

Supplementary Information References

1. Obenchain, V. *et al.* VariantAnnotation : a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics* **30**, 2076–2078 (2014).
2. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
3. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
4. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
5. Oikkonen, L. & Lise, S. Making the most of RNA-seq: Pre-processing sequencing data with Opossum for reliable SNP variant detection. *Wellcome Open Res.* **2**, 6 (2017).
6. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
7. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

8. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
9. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
10. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013).
11. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
12. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699 (2018).
13. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520 (2017).
14. Yu, F., Haynes, S. E. & Nesvizhskii, A. I. IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs. *Mol. Cell. Proteomics* **20**, 100077 (2021).
15. Yu, F. *et al.* Fast Quantitative Analysis of timsTOF PASEF Data with MSFragger and IonQuant. *Mol. Cell. Proteomics* **19**, 1575–1585 (2020).
16. Leprevost, F. da V. *et al.* Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat. Methods* **2020 179** **17**, 869–870 (2020).
17. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Anal. Chem.* **74**, 5383–5392 (2002).

18. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D450 (2019).
19. Hatos, A. *et al.* DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* **48**, D269–D276 (2020).
20. Quaglia, F. *et al.* DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Res.* **50**, D480–D487 (2022).
21. Piovesan, D. *et al.* DisProt 7.0: a major update of the database of disordered proteins. *Nucleic Acids Res.* **45**, D219–D227 (2017).
22. Hornbeck, P. V., Chabra, I., Kornhauser, J. M., Skrzypek, E. & Zhang, B. PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics* **4**, 1551–1561 (2004).
23. Rainer, J., Gatto, L. & Weichenberger, C. X. ensemblDb: an R package to create and use Ensembl-based annotation resources. *Bioinformatics* **35**, 3151–3153 (2019).

References

1. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
2. Shen, H. *et al.* Comprehensive Characterization of Human Genome Variation by High Coverage Whole-Genome Sequencing of Forty Four Caucasians. *PLOS ONE* **8**, e59494 (2013).
3. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* **339**, 1546–1558 (2013).
4. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* **43**, D805–D811 (2015).
5. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e18 (2018).
6. Miseta, A. & Csutora, P. Relationship Between the Occurrence of Cysteine in Proteins and the Complexity of Organisms. *Mol. Biol. Evol.* **17**, 1232–1239 (2000).
7. Tsuber, V., Kadamov, Y., Brautigam, L., Berglund, U. W. & Helleday, T. Mutations in Cancer Cause Gain of Cysteine, Histidine, and Tryptophan at the Expense of a Net Loss of Arginine on the Proteome Level. *Biomol.* 2017 Vol 7 Page 49 **7**, 49 (2017).
8. Kim, J. Y., Plaman, B. A. & Bishop, A. C. Targeting a Pathogenic Cysteine Mutation: Discovery of a Specific Inhibitor of Y279C SHP2. *Biochemistry* **59**, 3498–3507 (2020).
9. Ostrem, J. M., Peters, U., Sos, M. L., Wells, J. A. & Shokat, K. M. K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. *Nat.* 2013 5037477 **503**, 548–551 (2013).
10. Slebos, R. J. C. *et al.* K-ras Oncogene Activation as a Prognostic Marker in

- Adenocarcinoma of the Lung. *N. Engl. J. Med.* **323**, 561–565 (1990).
11. Tomlinson, D. C., Hurst, C. D. & Knowles, M. A. Knockdown by shRNA identifies S249C mutant FGFR3 as a potential therapeutic target in bladder cancer. *Oncogene* **26**, 5889–5899 (2007).
 12. Dang, L. *et al.* Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature* **462**, 739–744 (2009).
 13. Sved, J. & Bird, A. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc. Natl. Acad. Sci.* **87**, 4692–4696 (1990).
 14. Li, D. *et al.* BIBW2992, an irreversible EGFR/HER2 inhibitor highly effective in preclinical lung cancer models. *Oncogene* **27**, 4702–4711 (2008).
 15. Honigberg, L. A. *et al.* The Bruton tyrosine kinase inhibitor PCI-32765 blocks B-cell activation and is efficacious in models of autoimmune disease and B-cell malignancy. *Proc. Natl. Acad. Sci.* **107**, 13075–13080 (2010).
 16. Pan, Z. *et al.* Discovery of Selective Irreversible Inhibitors for Bruton's Tyrosine Kinase. *ChemMedChem* **2**, 58–61 (2007).
 17. Janes, M. R. *et al.* Targeting KRAS Mutant Cancers with a Covalent G12C-Specific Inhibitor. *Cell* **172**, 578-589.e17 (2018).
 18. Lanman, B. A. *et al.* Discovery of a Covalent Inhibitor of KRASG12C (AMG 510) for the Treatment of Solid Tumors. *J. Med. Chem.* **63**, 52–65 (2020).
 19. Canon, J. *et al.* The clinical KRAS(G12C) inhibitor AMG 510 drives anti-tumour immunity. *Nature* **575**, 217–223 (2019).
 20. Boatner, L. M., Palafox, M. F., Schweppe, D. K. & Backus, K. M. CysDB: a human cysteine database based on experimental quantitative chemoproteomics. *Cell*

- Chem. Biol.* (2023) doi:10.1016/j.chembiol.2023.04.004.
21. Kuljanin, M. *et al.* Reimagining high-throughput profiling of reactive cysteines for cell-based screening of large electrophile libraries. *Nat. Biotechnol.* 2021 395 **39**, 630–641 (2021).
 22. Yan, T. *et al.* SP3-FAIMS Chemoproteomics for High-Coverage Profiling of the Human Cysteinome**. *ChemBioChem* cbic.202000870 (2021) doi:10.1002/cbic.202000870.
 23. Cao, J. *et al.* Multiplexed CuAAC Suzuki–Miyaura Labeling for Tandem Activity-Based Chemoproteomic Profiling. *Anal. Chem.* **93**, 2610–2618 (2021).
 24. Li, Z., Liu, K., Xu, P. & Yang, J. Benchmarking Cleavable Biotin Tags for Peptide-Centric Chemoproteomics. *J. Proteome Res.* **21**, 1349–1358 (2022).
 25. Weerapana, E. *et al.* Quantitative reactivity profiling predicts functional cysteines in proteomes. (2010) doi:10.1038/nature09472.
 26. Vinogradova, E. V. *et al.* An Activity-Guided Map of Electrophile-Cysteine Interactions in Primary Human T Cells. *Cell* **182**, 1009-1026.e29 (2020).
 27. 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**, e9840 (2021).
 28. Yang, F., Jia, G., Guo, J., Liu, Y. & Wang, C. Quantitative Chemoproteomic Profiling with Data-Independent Acquisition-Based Mass Spectrometry. *J. Am. Chem. Soc.* jacs.1c11053 (2022) doi:10.1021/JACS.1C11053.
 29. Backus, K. M. *et al.* Proteome-wide covalent ligand discovery in native biological systems. (2016) doi:10.1038/nature18002.

30. Eberl, H. C. *et al.* Chemical proteomics reveals target selectivity of clinical Jak inhibitors in human primary cells. *Sci. Rep.* **9**, 14159 (2019).
31. Feldman, H. C. *et al.* Selective inhibitors of SARM1 targeting an allosteric cysteine in the autoregulatory ARM domain. *Proc. Natl. Acad. Sci.* **119**, e2208457119 (2022).
32. Grossman, E. A. *et al.* Covalent Ligand Discovery against Druggable Hotspots Targeted by Anti-cancer Natural Products. *Cell Chem. Biol.* **24**, 1368-1376.e4 (2017).
33. Abegg, D. *et al.* Chemoproteomic Profiling by Cysteine Fluoroalkylation Reveals Myrocin G as an Inhibitor of the Nonhomologous End Joining DNA Repair Pathway. *J. Am. Chem. Soc.* **143**, 20332–20342 (2021).
34. Bar-Peled, L. *et al.* Chemical Proteomics Identifies Druggable Vulnerabilities in a Genetically Defined Cancer. *Cell* **171**, 696-709.e23 (2017).
35. Li, H. *et al.* Assigning functionality to cysteines by base editing of cancer dependency genes. 2022.11.17.516964 Preprint at <https://doi.org/10.1101/2022.11.17.516964> (2022).
36. Lazear, M. R. *et al.* Proteomic discovery of chemical probes that perturb protein complexes in human cells. *Mol. Cell* **83**, 1725-1742.e12 (2023).
37. Rivero-Hinojosa, S. *et al.* Proteogenomic discovery of neoantigens facilitates personalized multi-antigen targeted T cell immunotherapy for brain tumors. *Nat. Commun.* **12**, 6689 (2021).
38. Sheynkman, G. M., Shortreed, M. R., Frey, B. L., Scalf, M. & Smith, L. M. Large-scale mass spectrometric detection of variant peptides resulting from nonsynonymous nucleotide differences. *J. Proteome Res.* **13**, 228–240 (2014).

39. Lau, E. *et al.* Splice-Junction-Based Mapping of Alternative Isoforms in the Human Proteome. *Cell Rep.* **29**, 3751-3765.e5 (2019).
40. Chen, Y. J. *et al.* Proteogenomics of Non-smoking Lung Cancer in East Asia Delineates Molecular Signatures of Pathogenesis and Progression. *Cell* **182**, 226-244.e17 (2020).
41. Vasaiakar, S. *et al.* Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. *Cell* **177**, 1035-1049.e19 (2019).
42. Wang, X. *et al.* Protein Identification Using Customized Protein Sequence Databases Derived from RNA-Seq Data. *J. Proteome Res.* **11**, 1009–1017 (2012).
43. Sheynkman, G. M., Shortreed, M. R., Cesnik, A. J. & Smith, L. M. Proteogenomics: Integrating Next-Generation Sequencing and Mass Spectrometry to Characterize Human Proteomic Variation. *Annu. Rev. Anal. Chem.* **9**, 521–545 (2016).
44. Sinitcyn, P. *et al.* Global detection of human variants and isoforms by deep proteome sequencing. *Nat. Biotechnol.* 1–11 (2023) doi:10.1038/s41587-023-01714-x.
45. Sinitcyn, P., Gerwien, M. & Cox, J. MaxQuant Module for the Identification of Genomic Variants Propagated into Peptides. in *Proteomics in Systems Biology: Methods and Protocols* (ed. Geddes-McAlister, J.) 339–347 (Springer US, 2022). doi:10.1007/978-1-0716-2124-0_23.
46. Cesnik, A. J. *et al.* Spritz: A Proteogenomic Database Engine. *J. Proteome Res.* (2020) doi:10.1021/acs.jproteome.0c00407.
47. Wang, X. & Zhang, B. customProDB: an R package to generate customized protein databases from RNA-Seq data for proteomics search. *Bioinformatics* **29**, 3235–3237 (2013).

48. Sheynkman, G. M. *et al.* Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. *BMC Genomics* **15**, 703 (2014).
49. Wen, B. *et al.* sapFinder: an R/Bioconductor package for detection of variant peptides in shotgun proteomics experiments. *Bioinformatics* **30**, 3136–3138 (2014).
50. Kennedy, J. J. *et al.* Internal Standard Triggered-Parallel Reaction Monitoring Mass Spectrometry Enables Multiplexed Quantification of Candidate Biomarkers in Plasma. *Anal. Chem.* **94**, 9540–9547 (2022).
51. Miller, R. M. *et al.* Enhanced protein isoform characterization through long-read proteogenomics. *Genome Biol.* **23**, 69 (2022).
52. Nesvizhskii, A. I. Proteogenomics: Concepts, applications and computational strategies. *Nat. Methods* **11**, 1114–1125 (2014).
53. Woo, S. *et al.* Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. *PROTEOMICS* **14**, 2719–2730 (2014).
54. Woo, S. *et al.* Advanced Proteogenomic Analysis Reveals Multiple Peptide Mutations and Complex Immunoglobulin Peptides in Colon Cancer. *J. Proteome Res.* **14**, 3555–3567 (2015).
55. Wen, B., Li, K., Zhang, Y. & Zhang, B. Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis. *Nat. Commun.* **2020 111** **11**, 1–14 (2020).
56. Szpiech, Z. A. *et al.* Prominent features of the amino acid mutation landscape in cancer. *PLOS ONE* **12**, e0183273 (2017).
57. Anoosha, P., Sakthivel, R. & Michael Gromiha, M. Exploring preferred amino acid

- mutations in cancer genes: Applications to identify potential drug targets. *Biochim. Biophys. Acta BBA - Mol. Basis Dis.* **1862**, 155–165 (2016).
58. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
59. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740–754 (2016).
60. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).
61. Wishart, D. S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**, D668–672 (2006).
62. Helleday, T., Eshtad, S. & Nik-Zainal, S. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* **15**, 585–598 (2014).
63. Petljak, M. *et al.* Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* **176**, 1282 (2019).
64. Chalmers, Z. R. *et al.* Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* **9**, 1–14 (2017).
65. Vilar, E. & Gruber, S. B. Microsatellite instability in colorectal cancer—the stable evidence. *Nat. Rev. Clin. Oncol.* **7**, 153 (2010).
66. Aaltonen, L. A. *et al.* Clues to the Pathogenesis of Familial Colorectal Cancer. *Science* **260**, 812–816 (1993).
67. Ionov, Y., Peinado, M. A., Malkhosyan, S., Shibata, D. & Perucho, M. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature* **363**, 558–561 (1993).

68. Stadler, Z. K. *et al.* Reliable Detection of Mismatch Repair Deficiency in Colorectal Cancers Using Mutational Load in Next-Generation Sequencing Panels. *J. Clin. Oncol.* **34**, 2141–2147 (2016).
69. Glaab, W. E. *et al.* Characterization of Distinct Human Endometrial Carcinoma Cell Lines Deficient in Mismatch Repair That Originated from a Single Tumor. *J. Biol. Chem.* **273**, 26662–26669 (1998).
70. Matheson, E. C. & Hall, A. G. Assessment of mismatch repair function in leukaemic cell lines and blasts from children with acute lymphoblastic leukaemia. *Carcinogenesis* **24**, 31–38 (2003).
71. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).
72. Schulze, K. V., Hanchard, N. A. & Wangler, M. F. Biases in arginine codon usage correlate with genetic disease risk. *Genet. Med.* 1–6 (2020) doi:10.1038/s41436-020-0813-6.
73. Sherry, S. T., Ward, M. & Sirotkin, K. dbSNP—Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. *Genome Res.* **9**, 677–679 (1999).
74. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
75. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).
76. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618–622 (2016).

77. Shi, Y., Hata, A., Lo, R. S., Massagué, J. & Pavletich, N. P. A structural basis for mutational inactivation of the tumour suppressor Smad4. *Nature* **388**, 87–93 (1997).
78. Van Houten, B. & Kong, M. Eukaryotic Nucleotide Excision Repair. *Encycl. Cell Biol.* **1**, 435–441 (2016).
79. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
80. Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43**, D512–520 (2015).
81. Ying, H. & Huttley, G. Exploiting CpG Hypermutable to Identify Phenotypically Significant Variation Within Human Protein-Coding Genes. *Genome Biol. Evol.* **3**, 938–949 (2011).
82. Fang, H. *et al.* Deficiency of replication-independent DNA mismatch repair drives a 5-methylcytosine deamination mutational signature in cancer. *Sci. Adv.* **7**, eabg4398 (2021).
83. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
84. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
85. Kong, A. T., Lerevost, F. V., Avtonomov, D. M., Mellacheruvu, D. & Nesvizhskii, A. I. MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods* **14**, 513–520 (2017).

86. Leprevost, F. da V. *et al.* Philosopher: a versatile toolkit for shotgun proteomics data analysis. *Nat. Methods* 2020 179 **17**, 869–870 (2020).
87. Yan, T. *et al.* Enhancing Cysteine Chemoproteomic Coverage through Systematic Assessment of Click Chemistry Product Fragmentation. *Anal. Chem.* **94**, 3800–3810 (2022).
88. Yu, F., Haynes, S. E. & Nesvizhskii, A. I. IonQuant enables accurate and sensitive label-free quantification with FDR-controlled match-between-runs. *Mol. Cell. Proteomics* **20**, 100077 (2021).
89. Yu, F. *et al.* Fast Quantitative Analysis of timsTOF PASEF Data with MSFragger and IonQuant. *Mol. Cell. Proteomics* **19**, 1575–1585 (2020).
90. Boutilier, J. M., Warden, H., Doucette, A. A. & Wentzell, P. D. Chromatographic behaviour of peptides following dimethylation with H₂/D₂-formaldehyde: Implications for comparative proteomics. *J. Chromatogr. B* **908**, 59–66 (2012).
91. Zhang, R., Sioma, C. S., Thompson, R. A., Xiong, L. & Regnier, F. E. Controlling Deuterium Isotope Effects in Comparative Proteomics. *Anal. Chem.* **74**, 3662–3669 (2002).
92. Zhu, Y. *et al.* Discovery of coding regions in the human genome by integrated proteogenomics analysis workflow. *Nat. Commun.* **9**, (2018).
93. Yeom, J. *et al.* A proteogenomic approach for protein-level evidence of genomic variants in cancer cells. *Sci. Rep.* **6**, 35305 (2016).
94. Wang, D. *et al.* A deep proteome and transcriptome abundance atlas of 29 healthy human tissues. *Mol. Syst. Biol.* **15**, e8503 (2019).
95. Choong, W. K., Wang, J. H. & Sung, T. Y. MinProtMaxVP: Generating a minimized

- number of protein variant sequences containing all possible variant peptides for proteogenomic analysis. *J. Proteomics* **223**, 103819 (2020).
96. Alfaro, J. A. *et al.* Detecting protein variants by mass spectrometry: A comprehensive study in cancer cell-lines. *Genome Med.* **9**, (2017).
 97. Zhang, M. *et al.* CanProVar 2.0: An Updated Database of Human Cancer Proteome Variation. *J. Proteome Res.* **16**, 421–432 (2017).
 98. Robin, T., Bairoch, A., Müller, M., Lisacek, F. & Lane, L. Large-Scale Reanalysis of Publicly Available HeLa Cell Proteomics Data in the Context of the Human Proteome Project. *J. Proteome Res.* **17**, 4160–4170 (2018).
 99. Krug, K., Popic, S., Carpy, A., Taumer, C. & Macek, B. Construction and assessment of individualized proteogenomic databases for large-scale analysis of nonsynonymous single nucleotide variants. *PROTEOMICS* **14**, 2699–2708 (2014).
 100. Venereau, E. *et al.* Mutually exclusive redox forms of HMGB1 promote cell recruitment or proinflammatory cytokine release. *J. Exp. Med.* **209**, 1519–1528 (2012).
 101. Xu, X. *et al.* Unique domain appended to vertebrate tRNA synthetase is essential for vascular development. *Nat. Commun.* **3**, 681 (2012).
 102. Son, J. *et al.* Conformational changes in human prolyl-tRNA synthetase upon binding of the substrates proline and ATP and the inhibitor halofuginone. *Acta Crystallogr. D Biol. Crystallogr.* **69**, 2136–2145 (2013).
 103. Arif, A. *et al.* EPRS is a critical mTORC1-S6K1 effector that influences adiposity in mice. *Nature* **542**, 357–361 (2017).
 104. Sebti, S. M., Jani, J. P., Mistry, J. S., Gorelik, E. & Lazo, J. S. Metabolic inactivation:

- a mechanism of human tumor resistance to bleomycin. *Cancer Res.* **51**, 227–232 (1991).
105. Finkel, T. Signal transduction by reactive oxygen species. *J. Cell Biol.* **194**, 7–15 (2011).
106. Desai, H. S. *et al.* SP3-Enabled Rapid and High Coverage Chemoproteomic Identification of Cell-State–Dependent Redox-Sensitive Cysteines. *Mol. Cell. Proteomics* **21**, 100218 (2022).
107. Desai, H. S., Yan, T. & Backus, K. M. SP3-FAIMS-Enabled High-Throughput Quantitative Profiling of the Cysteinome. *Curr. Protoc.* **2**, e492 (2022).
108. Hughes, C. S. *et al.* Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat. Protoc.* **14**, 68–85 (2019).
109. Hebert, A. S. *et al.* Comprehensive Single-Shot Proteomics with FAIMS on a Hybrid Orbitrap Mass Spectrometer. *Anal. Chem.* **90**, 9529–9537 (2018).
110. Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol.* **19**, 1720–1730 (1999).
111. Fortelny, N., Overall, C. M., Pavlidis, P. & Freue, G. V. C. Can we predict protein from mRNA levels? *Nature* **547**, E19–E20 (2017).
112. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–550 (2016).
113. Chuang, H., Zhang, W. & Gray, W. M. Arabidopsis ETA2, an apparent ortholog of the human cullin-interacting protein CAND1, is required for auxin responses mediated by the SCF(TIR1) ubiquitin ligase. *Plant Cell* **16**, 1883–1897 (2004).
114. Bjorkman, P. J. *et al.* Structure of the human class I histocompatibility antigen, HLA-

- A2. *Nature* **329**, 506–512 (1987).
115. Peaper, D. R. & Cresswell, P. Regulation of MHC Class I Assembly and Peptide Binding. <https://doi.org/10.1146/annurev.cellbio.24.110707.175347>
<https://www.annualreviews.org/doi/abs/10.1146/annurev.cellbio.24.110707.175347>
(2008) doi:10.1146/annurev.cellbio.24.110707.175347.
116. Warburton, R. J. *et al.* Mutation of the $\alpha 2$ domain disulfide bridge of the class I molecule HLA-A*0201 Effect on maturation and peptide presentation. *Hum. Immunol.* **39**, 261–271 (1994).
117. Gattoni-Celli, S., Kirsch, K., Timpane, R. & Isselbacher, K. J. $\beta 2$ -Microglobulin Gene Is Mutated in a Human Colon Cancer Cell Line (HCT) Deficient in the Expression of HLA Class I Antigens on the Cell Surface¹. *Cancer Res.* **52**, 1201–1204 (1992).
118. Martayan, A. *et al.* Conformation and surface expression of free HLA-CW1 heavy chains in the absence of $\beta 2$ -microglobulin. *Hum. Immunol.* **53**, 23–33 (1997).
119. Hughes, E. A., Hammond, C. & Cresswell, P. Misfolded major histocompatibility complex class I heavy chains are translocated into the cytoplasm and degraded by the proteasome. *Proc. Natl. Acad. Sci.* **94**, 1896–1901 (1997).
120. Lenart, I. *et al.* The MHC Class I Heavy Chain Structurally Conserved Cysteines 101 and 164 Participate in HLA-B27 Dimer Formation. *Antioxid. Redox Signal.* **16**, 33–43 (2012).
121. Neefjes, J., Jongma, M. L. M., Paul, P. & Bakke, O. Towards a systems understanding of MHC class I and MHC class II antigen presentation. *Nat. Rev. Immunol.* **11**, 823–836 (2011).
122. Mallal, S. *et al.* HLA-B*5701 Screening for Hypersensitivity to Abacavir. *N. Engl. J.*

- Med.* **358**, 568–579 (2008).
123. Weiss, G. A. *et al.* Covalent HLA-B27/peptide complex induced by specific recognition of an aziridine mimic of arginine. *Proc. Natl. Acad. Sci.* **93**, 10945–10948 (1996).
124. Zhang, Z. *et al.* A covalent inhibitor of K-Ras(G12C) induces MHC class I presentation of haptenated peptide neoepitopes targetable by immunotherapy. *Cancer Cell* **40**, 1060-1069.e7 (2022).
125. Grob, N. M. *et al.* Electrophile Scanning Reveals Reactivity Hotspots for the Design of Covalent Peptide Binders. Preprint at <https://doi.org/10.26434/chemrxiv-2023-hvq1k> (2023).
126. Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. HLA variation and disease. *Nat. Rev. Immunol.* **18**, 325–339 (2018).
127. Brewerton, D. A. *et al.* ANKYLOSING SPONDYLITIS AND HL-A 27. *The Lancet* **301**, 904–907 (1973).
128. Alvarez, I. *et al.* The Cys-67 Residue of HLA-B27 Influences Cell Surface Stability, Peptide Specificity, and T-cell Antigen Presentation *. *J. Biol. Chem.* **276**, 48740–48747 (2001).
129. Teo, G. C., Polasky, D. A., Yu, F. & Nesvizhskii, A. I. Fast Deisotoping Algorithm and Its Implementation in the MSFragger Search Engine. *J. Proteome Res.* **20**, 498–505 (2020).
130. Yang, K. L. *et al.* MSBooster: Improving Peptide Identification Rates using Deep Learning-Based Features. 2022.10.19.512904 Preprint at <https://doi.org/10.1101/2022.10.19.512904> (2022).

131. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).
132. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry. *Anal. Chem.* **75**, 4646–4658 (2003).
133. Jinek, M. *et al.* A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**, 816–821 (2012).
134. Bennis, H. J. *et al.* CRISPR-based oligo recombineering prioritizes apicomplexan cysteines for drug discovery. *Nat. Microbiol.* **7**, 1891–1905 (2022).
135. Gaudelli, N. M. *et al.* Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).
136. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).
137. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
138. Joutel, A. *et al.* Notch3 mutations in CADASIL, a hereditary adult-onset condition causing stroke and dementia. *Nature* **383**, 707–710 (1996).
139. Hattori, T. *et al.* Creating MHC-Restricted Neoantigens with Covalent Inhibitors That Can Be Targeted by Immune Therapy. *Cancer Discov.* **13**, 132–145 (2023).
140. Desai, J., Francis, C., Longo, K. & Hoss, A. Predicting exon criticality from protein sequence. *Nucleic Acids Res.* **50**, 3128–3141 (2022).

141. Cao, X. *et al.* Comparative Proteomic Profiling of Unannotated Microproteins and Alternative Proteins in Human Cell Lines. *J. Proteome Res.* **19**, 3418–3426 (2020).
142. Chen, Y., Cao, X., Loh, K. H. & Slavoff, S. A. Chemical labeling and proteomics for characterization of unannotated small and alternative open reading frame-encoded polypeptides. *Biochem. Soc. Trans.* **51**, 1071–1082 (2023).

Chapter 4: Conclusion

Collectively, our work has generated two novel chemoproteomic platforms that together combine SP3 chemoproteomics workflows with new sample preparation and data processing methods to improve proteome-wide identification of cysteines.

In Chapter 2, we establish SP3-Rox to improve our understanding of how cells are impacted by and respond to oxidative stressors. We first created heavy and light isotopically differentiated iodoacetamide probes to enable quantitation in our two-round labeling workflow. These reduced cost reagents streamline traditional redox workflows that require several decontamination steps. We validated our probes using an improved version of FragPipe that accurately quantifies cysteine labeling with and without the use of a FAIMS device comparably to Skyline. We also demonstrated PTM-prophet enabled site-of-labeling localization for multi-Cys containing peptides. We then generated SP3-Rox for high-coverage Cys oxidation state quantitation of thousands of cysteines, identifying > 9,000 cysteines and > 2,000 oxidant sensing cysteines from lysates subjected to 1mM GSNO. We use SP3-Rox to identify cell-state dependent changes in cysteine oxidation. We found that a number of cysteines showed increased oxidation during cellular activation. We expect that a subset of the residues identified are likely functional regulators of cellular activation.

In Chapter 3, we fused cysteine chemoproteomics with genomics to unveil *hidden* cysteines in the proteome. Chemoproteomics which is capable of proteome-wide assessment of functional, redox-sensitive or druggable residues, has historically been limited to reference proteome detection. In this work, we found that, for both cancer and

healthy genomes, cysteine acquisition is a ubiquitous result of spontaneous genetic variation that is increased in the context of decreased DNA mismatch repair. We find that CADD predictions of deleteriousness show an enrichment for likely damaging variants that result in acquisition of cysteine in cell lines with decreased mismatch repair. In this work, we create combinatorial databases from sample-matched NGS data—these databases include all combinations of variants. To address elevated false discovery rates from sample-specific database searches, we also established an MSFragger-based two-stage search implemented into the FragPipe GUI. Chemoproteogenomics across 11 cancer cell lines identified 116 gain-of-cysteines of which 10 were liganded by druglike molecules. In total we identified > 1,400 unique variants through chemoproteomics and bulk fractionation. We found 791 reference cysteines proximal to variants highlighting opportunities to develop proteoform-specific probes. We also found that chemoproteogenomics is compatible with redox and ligandability profiling, as evident by our re-analysis of SP3-Rox data, and small-molecule screening.

We anticipate that chemoproteogenomics is capable of proteoform-specific probe development or neoantigen discovery that is further enhanced when supplemented with information beyond missense mutations (e.g. alternative splicing, indels, translocations/transversions, new open reading frames). The SP3-Rox and chemoproteogenomics platforms together will provide relational insight, in addition to variant effects, on cysteine ligandability and redox-sensitivity.