Supporting Information

CysDB: A Human Cysteine Database based on Experimental Quantitative Chemoproteomics.

Lisa M. Boatner¹,², Maria F. Palafox³, Devin K. Schweppe⁴ and Keriann M. Backus¹,²,⁵,⁶,⁷,⁸*

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 Human Genetics, David Geffen School of Medicine, UCLA, Los Angeles, CA, 90095, USA.

4. Department of Genome Sciences, University of Washington, Seattle, WA, 98185, 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
Contents

(A) Supplementary Figures 3
(B) Methods 29
(C) References 31
(A) Supplementary Figures

Figure S1. General chemoproteomics workflow for measuring intrinsic cysteine-reactivity towards electrophilic probes such as iodoacetamide alkyne (IAA). For these intrinsic reactivity studies, Isotopic Tandem Orthogonal Proteolysis‒ABPP (isoTOP‒ABPP), relative cysteine labeling by a high (10x) and low (1x) concentration of IAA or other probes are compared using isotopically labeled enrichment handles and MS1-based quantification. Hyper-reactive residues are those that show R_{10:1} ratios close to 1, indicating saturation of labeling at the lower reagent concentration.

Figure S2. General chemoproteomics workflow for measuring cysteine ligandability using competitive isoTOP-ABPP. Proteomes are treated with electrophilic fragments or vehicle (DMSO), labeled with an iodoacetamide (IA)-alkyne probe, and conjugated to isotopically-differentiated, TEV-cleavable tags by click chemistry. Treated and control samples are combined, processed, and analyzed by LC-MS/MS, where the isotopic label is used to distinguish between peptides from control and fragment-treated samples, with elevated R_{H:L} ratios indicative of a liganded cysteine.
Figure S3. Total number of compounds for each warhead category in CysDB Lig: acrylamide (AA), bromoacetamide (BA), chloroacetamide (CA), dimethyl fumarate (DMF) and other (OTHER).
Figure S4. Total number proteins for each category in CysDB & in UniProtKB/Swiss-Prot.
Figure S5. Number of shared cysteines identified between different cell lines.
Figure S6. Entity-relationship diagram of all nine tables in CysDB and relationships with external data sources, such as UniProt, COSMIC, ClinVar and the Human Protein Atlas (HPA).
**Figure S7.** Total number of proteins in the human proteome from UniProtKB/Swiss-Prot and targeted by FDA approved drugs (left). Total number of CysDB ligandable proteins, CysDB hyperreactive proteins and targeted by FDA approved drugs (left).
Figure S8. Total number of cysteines liganded by each warhead per dataset.
Figure S9. Total number of proteins (left) and cysteines (right) liganded by the following electrophiles: chloroacetamides (CA), acrylamides (AA), other (OTHER), dimethyl fumarate (DMF) and bromoacetamides (BA).
Figure S10. Total number of proteins liganded by both acrylamides and chloroacetamides, exclusively acrylamides and exclusively chloroacetamides.
Figure S11. Distribution of amino acids annotated as binding sites in UniProt (left) and in UniProt proteins with an associated PDB structure (right).
Figure S12. Distribution of amino acids annotated as active sites in UniProt (left) and in UniProt proteins with an associated PDB structure (right).
**Figure S13.** Distribution of UniProtKB proteins (grey) and CysDB ID proteins (blue) as annotated binding sites (left), active sites (center) and have an associated PDB structure (right).

**Figure S14.** Number of CysDB ID cysteines that are annotated binding sites and cysteines that are not binding sites but near a binding site. Primary sequences were searched +/- 10 amino acids from the location of a detected cysteine. If another binding site was within this +/- 10 amino acid window, the cysteine was considered ‘near’ a binding site.
Figure S15. Number of CysDB ID cysteines that are annotated active sites and cysteines that are not active sites but near an active site. Primary sequences were searched +/- 10 amino acids from the location of a detected cysteine. If another active site was within this +/- 10 amino acid window, the cysteine was considered 'near' a active site.
Figure S16. Number of UniProt proteins in the human proteome, with an associated PDB structure, residue mapped SIFTS file and having a cysteine resolved in the corresponding associated PDB (right).
Figure S17. Number of UniProtKB proteins with an annotated binding site, associated PDB structure, annotated as an binding site resolved in an associated PDB structure and with cysteines near an annotated binding site. The distance from the sulfur atom of each cysteine to an annotated binding site residue was calculated. Cysteines within 10 Angstroms of the annotated binding site residue were considered as cysteines ‘near’ binding sites.
**Figure S18.** Number of UniProt proteins with an annotated active site, associated PDB structure, annotated as an active site resolved in an associated PDB structure and with cysteines near an annotated active site. The distance from the sulfur atom of each cysteine to an annotated active site residue was calculated. Cysteines within 10 Angstroms of the annotated active site residue were considered as cysteines ‘near’ active sites.
Figure S19. Number of CysDB ID cysteines identified, resolved in an associated PDB and CysDB ID cysteines that are not annotated binding sites but are near an annotated binding site in 3D space. Proteins with an annotated binding site, annotated as a binding site resolved in an associated PDB structure and with cysteines near an annotated binding site. The distance from the sulfur atom of each cysteine to an annotated binding site residue was calculated. Cysteines within 10 Angstroms of the annotated binding site residue were considered as cysteines 'near' binding sites.
Figure S20. Number of CysDB ID cysteines identified, resolved in an associated PDB and CysDB ID cysteines that are not annotated active sites but are near an annotated active site in 3D space. Proteins with an annotated active site, annotated as an active site resolved in an associated PDB structure and with cysteines near an annotated binding site. The distance from the sulfur atom of each cysteine to an annotated active site residue was calculated. Cysteines within 10 Angstroms of the annotated active site residue were considered as cysteines 'near' active sites.
Figure S21. Top-10 enriched protein domains from Pfam-term enrichment analysis of liganded proteins with gene counts.
Figure S22. Top-10 enriched protein domains from Pfam-term enrichment analysis of hyper-reactive proteins with gene counts.
**Figure S23.** Top-10 enriched pathways from Panther-term enrichment analysis of liganded proteins with gene counts.
Figure S24. Top-10 enriched pathways from Panther-term enrichment enrichment analysis of hyperreactive proteins with gene counts.

Figure S25. Top-10 enriched pathways from OMIM-term enrichment enrichment analysis of ligandable proteins (left) and hyperreactive proteins (right).
**Figure S26.** Top-10 enriched pathways from OMIM-term enrichment enrichment analysis of ligandable proteins with gene counts.
Figure S27. Top-10 enriched pathways from OMIM-term enrichment enrichment analysis of hyperreactive proteins with gene counts.

Figure S28. Overlap between the number of CysDB identified proteins and UniprotKB proteins associated with Cancer Gene Census (CGC) genes (756) (left). Overlap between
the number of CysDB identified proteins and proteins associated with ClinVar variants (right).

**Figure S29.** Overlap between the number of FDA targeted proteins, Uniprot proteins associated with Cancer Gene Census (CGC) genes (756) and ClinVar variants.
Figure S30. Overlap between the number of benign, variants of unknown significance (VUS) and pathogenic ClinVar missense variants for CysDB ID proteins.
(B) Methods

**ChEMBL, DrugBank and FDA Annotations of Detected Cysteines**

Custom Python scripts classified protein functions based on annotations in the UniProtKB/Swiss-Prot database (2209-release). UniProt accessions were collected from proteins with available ChEMBL and DrugBank UniProt annotations. Data from the Human Protein Atlas (HPA) version 21.1 was downloaded and parsed to obtain genes targeted by FDA approved drugs. HGNC gene symbols were mapped to UniProt accessions to collect proteins targeted by FDA approved drugs.

**Functional Annotations of Ligandable Cysteines**

Custom Python scripts classified protein functions based on annotations in the UniProtKB/Swiss-Prot database (2209-release), HPA version 21.1 and the ScaPD database. UniProt keywords were parsed to classify proteins into five broad functional categories: chaperones/transporter/channel/receptor, enzyme, nucleic acid and small molecule binding, scaffolding/modulator/adaptor, transcription factor/regulator and uncategorized. Transcription factors, channels and transporters were also found using protein class descriptions from the HPA. In addition, examples of experimentally validated scaffolding proteins were collected from the ScaPD database. For proteins in more than one category, annotations were prioritized based on the following: enzyme > chaperones/transporter/channel/receptor > scaffolding/modulator/adaptor > transcription factor/regulator > nucleic acid and small molecule binding.

**Active Site, Binding Site, Catalytic Activity, Disulfide Bond and Redox Potential Annotations of Detected Cysteines**

Counts of how many CysDB proteins had UniProtKB annotations for active sites, binding sites, catalytic activity, disulfide bonds and redox potentials were calculated based on matches between the position of the identified residue and UniProtKB functional annotation. Further parsing of UniProt active and binding site annotations were extracted to obtain specific residues and amino acid numbers. Positions of binding and active sites that were not cysteine residues were discarded. Exact amino acid positions of UniProt cysteine active and binding sites were cross-referenced with CysDB cysteine identifiers.

**Cysteines Near Active or Binding Sites using Primary Sequences**

CysDB cysteines ‘near’ UniProt annotated active or binding sites were assessed using primary protein sequences. Positions of identified cysteines were found via their amino acid numbering. Annotated active or binding sites within +/-10 amino acids from the identified cysteine was considered as a cysteine ‘near’ an active or binding site.

**Cysteines Resolved in Crystal Structures**
Protein Data Bank (PDB) identifiers were found from UniProt annotations. Proteins without PDB structures were filtered out. PDB structures for proteins with PDB annotations were downloaded and parsed for amino acid numbering and residue names. A list of cysteines resolved in each PDB was stored for further processing. SIFTS files, providing residue level mapping between PDB sequences and protein sequences, were downloaded for each PDB. Cysteines resolved in each PDB were mapped to their appropriate UniProt protein sequence and identifiers for PDB to UniProt pairs were created: PDB_C#_UniProtKBID_C#. From these paired identifiers, the number of unique UniProtKBID_C# records were counted to determine the number of UniProt cysteines resolved in PDBs.

Cysteines Near Active or Binding Sites using 3D Crystal Structures
CysDB cysteines 'near' UniProt annotated active or binding sites were assessed using 3D PDB structures. From the workflow described below (determining cysteines in PDB structures), PDB structures were parsed to find all neighboring residues within a 10 Angstrom distance of a cysteine residue. PDB_UniProt identifiers were created for each cysteine and corresponding list of neighboring residues. If the UniProt 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.

Enrichment Analysis of Significant Terms from Publicly Accessible Libraries
Enrichment of Panther 2016, Pfam Domains 2019 and OMIM Disease gene set library terms were performed using the Enrichr package. UniProt protein identifiers were mapped to Entrez gene symbols as input for Enrichr. P-values were computed from Fisher’s exact test to determine the significance of each enriched term. The negative log of these p-values were calculated using R.

(C) References

structure-based annotations for proteins. *Nucleic acids research*, 47(D1), D482-D489.


