**NOTES from results read thru:**

* Legacy data mapping to 11\_2012, ID search and residue-level mapping (no alignment), where 2012 file is from with reverse seq and trembl

ID mapping

* ID search against 2018, Ensembl releases, xref file details, 2018 reference details
* Uniprot cross reference subsets, downloaded XXX, counts refer to XXX, file sourceXXX web source
* Timeline and average update methods, files uses, how i calculated it,

Residue-residue mapping

* Intra ukb 2012 to 2018 update, ID search, residue-residue mapping
* Isoform and Canonical ID

Inter database mapping

* UKB Ensembl multimapping- release-specific and pooled multimapping based on all xref for stable IDs from ensembl - downloaded for version specific ID
* Single isoform and multi isoform filtered
* Single and multi isoform ukbs identified from isoform analysis (used idat 01\_2018and ukb fasta from 06 2018 release) using filtered\_by\_col.py, multimapping code and methods pull from markdown

**-------------------------------**

**Database update cycles**

Average time between database updates was quantified for all releases between August 2013 - July 2019, total 5 years and 11 months. Dates counted refer to the public release date posted on each databases’ ftp site. There was a total of 25 Ensembl, 13 GENCODE, 6 CCDS (only homo sapien releases), and 5 NCBI releases used to calculate average time between each update. \*\*NCBI homo sapiens genomes includes Genbank and Refseq. RefSeq updates are daily following new submissions to website. NCBI plans to increase frequency of updates (from 1 per year to every 2 months). CCDS follows coordinated NCBI/Ensembl/WTSI (Havana) updates.\*\* **Timeline** iincludes some, but not all, dates used to calculate average update cycle length. The timeline dates selected focus on releases proximal to or overlapping the five Ensembl releases analyzed for multi-mapping burden.

Links to database ftp site:

**Data sources**

Our chosen reference human proteome for this study is the UniProtKB-SwissProt CCDS cross-referenced set of canonical sequences downloaded August 06, 2018. We selected this subset of UniProtKB because of the reduced peptide redundancy and compatibility with RefSeq and Ensembl transcript translations. The UniProtKB fasta file from November, 2012 releases was provided by (). CpD ID refers to the UniProtKB-SP stable entry or canonical ID for chemoproteomic detected proteins.

UniProtKB-SP 2018 fasta file:

CCDStype\_canonicalONLY\_uniprot-database%3A%28type%3Accds%29-filtered-reviewed%3Ayes+AND+organism%3A\_H--.fasta

From Ensembl, we downloaded the peptide fasta files for v85 (Homo\_sapiens.GRCh37.pep.all.fa), and v92, v94, v96, v97 (Homo\_sapiens.GRCh38.pep.all.fa). For ID mappings to UniProt, we downloaded release-specific xref pipeline files {Potter et al. 2016}. All cross-reference files used in this study are listed in the following table:

|  |  |  |
| --- | --- | --- |
| File name | N unique stable IDs | Description |
| HUMAN\_9606\_idmapping\_2018\_08\_01download.dat | 3,953 | To convert between isoform-specific UniProt IDs to stable Ensembl IDs |
| Homo\_sapiens.GRCh37.85.uniprot.tsv | 10,183 | To convert between Ensembl and UniProt stable IDs |
| Homo\_sapiens.GRCh38.92.uniprot.tsv | 10,395 | “” |
| Homo\_sapiens.GRCh38.94.uniprot.tsv | 10,612 | “” |
| Homo\_sapiens.GRCh38.96.uniprot.tsv | 10,663 | “” |
| Homo\_sapiens.GRCh38.97.uniprot.tsv | 10,564 | “” |

**Cross-reference sources used in this study.** For analysis all files were filtered for chemoproteomic detected stable ID matches. All Ensembl files have been filtered to map to a shared set of 3,953 CpD IDs. Some numbers quoted in the main text reflect only the subset of Ensembl IDs shared between all Ensembl releases (maps to 3,887 unique CpD IDs). Some limitations to using Ensembl and UniProt cross-reference files for ID conversions are:

* Versions of databases used in ID conversions dated or unknown to user
* ID redundancy between protein identifiers (partly due to Ensembl automated pipeline)
* For Ensembl xref files, identity score column does not mention what specific UniProt version ID the Ensembl protein is identical too (source says canonical UniProt ID, but this should not be assumed to be “-1” isoform and can change between UniProt updates).
* Unclear isoform/version related to ID conversion

A close up of a piece of paper

Description automatically generated

**ID mapping between legacy CpD proteins and later database releases**

To compare ID mapping efficiencies between legacy chemoproteomic data and later Ensembl and UniProt releases, we counted the total number of legacy IDs not found in each database release. Chemoproteomic-detected proteins were excluded from further analysis if: (1) matching stable UniProt ID was not found in database cross-reference file, (2) UniProt ID flagged with ‘caution’ on UniProt’s website (example:<https://www.uniprot.org/uniprot/Q8WUH1>), (3) UniProt sequence from 2018 releases compared to 2012 reference is missing detected cysteine and/or lysine positions (example: natural variant overlaps detected cysteine position).

* What files did i search?
* What script did i use?

**Residue-level mapping of CpD cysteine and lysine from original data to multiple UniProtKB releases**

The original chemoproteomic data of cysteine and lysine was provided in the format of UniProt stable ID, amino acid letter, and position (i.e. P11413\_C205). First, ID search and residue-level mapping code was run for the 11\_2012 provided fasta file and original CpD data.

OG 2012 results: I lost total 9 unique IDs. no sequence alignment option. A total of 12 cysteines, 4 iodoactemide labeled, 1 target, 7 reactive ratio. File ‘OG\_to\_11\_2012\_POS\_NOT\_FOUND\_DROP\_THESE\_IDS.csv

**CpD protein isoform count and canonical isoform name annotation**

Files used to calculate isoforms per UniProt stable ID:

UniprotKB\_isoforms\_canonical\_filtered\_homosapiens\_9606\_2018\_06release.fasta

Files used to calculate canonical isoform name:

UniprotKB\_isoforms\_canonical\_filtered\_homosapiens\_9606\_2018\_06release.fasta

HUMAN\_9606\_idmapping\_2018\_08\_01download.dat

Canonical isoform name calculated for UniProt IDs with more than one isoform.

|  |  |
| --- | --- |
| idat\_isoformNumber:  ['1', '2', '3', '4', 0] | fasta\_isoformNumber:  [0, '2', '3', '4'] |

**Multi-mapping analysis of Ensembl IDs to CpD stable IDs**

Statistical significance between different types of IDs (stable or version) within each biotype (gene, transcript, and protein) was calculated for Ensembl IDs cross-referencing UniProt IDs with only one isoform based on the 06\_2018 UniProtKB. We excluded UniProt IDs with multiple isoforms to analysis the redundancy in total Ensembl IDs linked to one protein sequence in UniProtKB**.** To calculate the degree of ID multi-mapping per UniProt stable ID, we pooled Ensembl IDs from all five releases studied and calculated the number of unique IDs cross-referencing a UniProt ID. Our goal was to visualize which biotype level (gene, transcript, or protein) has the greatest burden of multi-mapping IDs to UniProt stable IDs. The y axis represents the mean number of Ensembl IDs (grouped by biotype and stable or version type) cross-referencing a given UniProt stable ID, with error bars as SD. Unpaired t-test used for significance testing between ID type means for each biotype. \*\*\*\* for P value <0.0001. Alternative way for multimapping instead of using single isoform only: Multimapping correction check using UniProt mapping file to ensembl stable IDs. That way it reduced the ambiguous uniprot protein mapping. Ideally you should see 1:1 for protein protein ID conversions. grouped by unique UKB version ID, still have multimapping cases of ensembl stable protein I.D.s.

**Calculating top updating Ensembl gene, transcript, and protein IDs**

**Markdwon i has increment code**

To identify IDs that have undergone more updates over the course of the five Ensembl releases studied, we used unique Ensembl stable IDs shared across all releases (stableID\_keys {ENSG\_ENST\_ENSP} n=8,861 Ensembl stable transcript and protein IDs, n=X Ensembl stable gene IDs). The 8,861 Ensembl protein IDs cross-reference 3,887 unique CpD UniProt stable IDs. To identify update differences between ID biotypes, we used the version IDs in each release associated with the shared set of stable IDs, and summed the ID version number increments since the v85 Ensembl release. From v85 to v97, Ensembl gene IDs had a total of 41,439 sequence version updates, transcript IDs had 32,949 sequence version updates, and protein IDs had 128 sequence version updates.

*RANGE\_ENSGv\_top25\_8861stableKEY\_level.csv*

*RANGE\_ENSTv\_top25\_8861stableKEY\_level.csv*

*RANGE\_ENSPv\_top25\_8861stableKEY\_level.csv*

Distance scoring

**Annotation Mapping**

To annotate all cysteines and lysines in chemo-detected UniProt protein sequences with missense pathogenicity scores, a combination of stable ID and residue-level mapping was used to pull information from dbNSFPv4.0. Reference genome hg19 and hg38 coordinates from dbNSFPv4.0 used to map CADDv1.4 scores and additional annotations.

A screenshot of a cell phone

Description automatically generated

Statistical analysis

**Data and source code availability**

Analyses utilized Python 3.7.4 and R 3.6.2. Data and code sufficient to produce the plots and analyses in this paper are available at https://github.com/mfpfox/MAPPING

**---------------------------------------------------------------------------------------------------------------**

**Notes for genome annotation process available in TEXTBOOKS/Genome\_annotation\_chap2**

**Table of hg19 and hg38 stats:** regional, conservation, mutation density: Potential problem with redundancy since i am using all CADD positions and position changes...maybe I can make AA column and then drop all redundant rows. T test calculated in Prism

Check powerpoint slide draft for first exam for better version of this table

|  |  |  |  |
| --- | --- | --- | --- |
|  | GRCh37 | GRCh38 | BOTH |
| Functional | MCAP, REVEL, MPC, PrimateAI, CADD | CADD | Grantham, SIFT, Polyphen, BLOSUM |
| Conservation | GERP, PhastCons, PhyloP | GERP, PhastCons, PhyloP |  |
| Regional | Mutation density, GC, CpG | Mutation density, GC, CpG |  |

**MISMAP related:**

***Mismap score for multimapping Ensembl IDs to UniProt IDs =***

*(total # of positions not found / total # positions checked) / (total # of Ensembl IDs linked to UniProt ID)*

*\*\*Note: Total # positions checked is the # of chemo-detected residues detected in single UniProt ID times total # Ensembl IDs linked.* Our mismap score ignores the fact that amino acid positions could be recovered on Ensembl protein sequences through the use of pairwise sequence alignment tools (Tcoffee, muscle).  While these tools would most likely eliminate positions being “missed”, we chose to not use these resources in order quantify stable identifier mapping issues from UniProt to different Ensembl releases.]

To quantify the proportion of residues that can be mis-annotated using stable-identifier mapping, we compared the 2018 re-mapped legacy data to multiple versions of Ensembl proteins and calculated the fraction of detected positions not found with respect to the UniProt sequence positions. From calculating the proportion of detected positions not identical to UniProt positions in stable identifier cross-referenced Ensembl proteins, we found that each Ensembl release had similar ‘mis-mapped’ fraction averages. When considering the same set of Ensembl stable IDs in five different database versions, 18%-20% of all detected cysteine and lysine positions are missed (not matching position in UniProt canonical sequence). When only considering proteins with detected cysteines, 19%-22% of cysteines are missed, and in proteins with detected lysines, 16%-19% of lysines are missed.

**Mismapping score between releases expectation**

Since our stable identifier cross-referencing ignores UniProt specific isoform details, we expect that a fraction of Ensembl protein sequence to be identical to isoforms of the UniProt entry rather than identical to the canonical sequence. Because of this, we expect to see missed cysteine and lysine positions for a minority of all cross-referenced Ensembl identifiers.

**Mismap results or supplemental commentary?**

From calculating the proportion of detected positions not identical to UniProt positions in stable identifier cross-referenced Ensembl proteins, we found that each Ensembl release had similar ‘mis-mapped’ fraction averages. When considering the same set of Ensembl stable IDs in five different database versions, 18%-20% of all detected cysteine and lysine positions are missed (not matching position in UniProt canonical sequence). When only considering proteins with detected cysteines, 19%-22% of cysteines are missed, and in proteins with detected lysines, 16%-19% of lysines are missed. From searching legacy chemoproteomic-detected UniProt stable IDs against later Ensembl and UniProt versions, we expect variable ID loss between different releases. The Ensembl v85 release date is closest in time to the UniProt release legacy data is based on.

* LASER PLOTS:
  + use distance from ukb sequence instead of mismap (also could be heatmap)
  + Files with all ENSP + distance + 8861 label + True/False identity label:
  + v85.to\_csv("ENSEMBL\_SEQ\_v85\_3953\_UKBIDs.csv")
  + v92.to\_csv("ENSEMBL\_SEQ\_v92\_3953\_UKBIDs.csv")
  + v94.to\_csv("ENSEMBL\_SEQ\_v94\_3953\_UKBIDs.csv")
  + v96.to\_csv("ENSEMBL\_SEQ\_v96\_3953\_UKBIDs.csv")
  + v97.to\_csv("ENSEMBL\_SEQ\_v97\_3953\_UKBIDs.csv")

D) supplemental: Exploring correlated features with mismap, what associated with mismap:

*multimap and mismap score @*

*UKB\_level\_MISMAP\_3953\_v85.csv*

*UKB\_level\_MISMAP\_3953\_v92.csv*

*UKB\_level\_MISMAP\_3953\_v94.csv*

*UKB\_level\_MISMAP\_3953\_v96.csv*

*UKB\_level\_MISMAP\_3953\_v97.csv*