Title

Proteogenomic Annotation of Chinese Hamsters Reveals Extensive Novel Translation Events and Endogenous Retroviral Elements.

Journal

Journal of Proteome Research, 18(6)

Authors

Li, Shangzhong

Cha, Seong

Heffner, Kelly

et al.

Publication Date

2019-06-07

DOI

10.1021/acs.jproteome.8b00935

Peer reviewed

HHS Public Access

Author manuscript

J Proteome Res. Author manuscript; available in PMC 2020 June 07.

Published in final edited form as:

J Proteome Res. 2019 June 07; 18(6): 2433–2445. doi:10.1021/acs.jproteome.8b00935.

Proteogenomic annotation of the Chinese hamster reveals extensive novel translation events and endogenous retroviral elements

Shangzhong Li[1,2], **Seong Won Cha**[3], **Kelly Heffner**[4], **Deniz Baycin Hizal**[5], **Michael A. Bowen**[5], **Raghothama Chaerkady**[5], **Robert N. Cole**[6], **Vijay Tejwani**[7], **Prashant Kaushik**[8], **Michael Henry**[8], **Paula Meleady**[8], **Susan T. Sharfstein**[7], **Michael J. Betenbaugh**[4], **Vineet Bafna**[9], and **Nathan E. Lewis**[*,1,2,10]

[1.]Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA [2.]Novo Nordisk Foundation Center for Biosustainability, University of California, San Diego, La Jolla, CA, USA [3.]Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA, USA [4.]Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, Maryland, USA [5.]Antibody Discovery and Protein Engineering, AstraZeneca, Gaithersburg, Maryland USA [6.]The Mass Spectrometry Core, Johns Hopkins School of Medicine, Baltimore, MD, 21205 [7.]Colleges of Nanoscale Science and Engineering, SUNY Polytechnic Institute, Albany, NY, USA [8.]National Institute for Cellular Biotechnology, Dublin City University, Dublin 9, Ireland [9.]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA [10.]Department of Pediatrics, University of California, San Diego, La Jolla, CA USA

Abstract

A high-quality genome annotation greatly facilitates successful cell line engineering. Standard draft genome annotation pipelines are based largely on *de novo* gene prediction, homology, and RNA-Seq data. However, draft annotations can suffer from incorrect predictions of translated sequence, inaccurate splice isoforms and missing genes. Here we generated a draft annotation for the newly assembled Chinese hamster genome and used RNA-Seq, proteomics, and Ribo-Seq to experimentally annotate the genome. We identified 3,529 new proteins compared to the hamster RefSeq protein annotation and 2,256 novel translational events (e.g., alternative splices, mutations, novel splices). Finally, we used this pipeline to identify the source of translated retroviruses contaminating recombinant products from Chinese hamster ovary (CHO) cell lines, including 119 type-C retroviruses, thus enabling future efforts to eliminate retroviruses by reducing the costs incurred with retroviral particle clearance. In summary, the improved annotation provides a more

[*]Corresponding Author nlewisres@ucsd.edu.

Supporting information

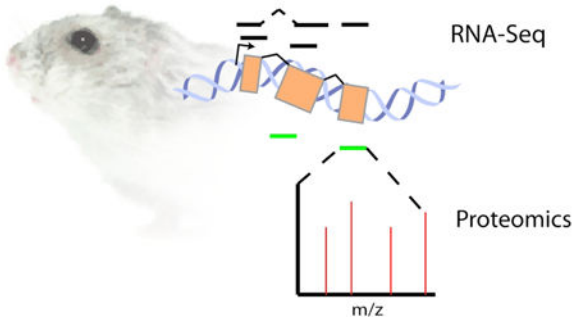
The following supporting information is available free of charge at ACS website http://pubs.acs.org

ACCESSION

RNA-Seq raw data: PRJNA504034; Proteomics raw data in MassIVE: doi:10.25345/C5M597, Identified peptides in Synapse: doi: 10.7303/syn17037373. Annotation will be available at CHOgenome.org. Code is available in https://github.com/LewisLabUCSD/Proteogenomics.

accurate platform for guiding CHO cell line engineering, including facilitating the interpretation of omics data, defining of cellular pathways, and engineering of complex phenotypes.

Graphical Abstract



Keywords

Chinese hamster; genome annotation; proteogenomics; endogenous retrovirus

Introduction

Chinese hamster ovary (CHO) cells are the primary workhorse for therapeutic protein production[1] thanks to its ability to efficiently produce biologically active recombinant proteins[2]. Sequencing and assembly of the CHO and Chinese hamster genomes[3–5] have enabled improvement in protein production using genetic engineering and in cell line process optimization using omics technologies[67]. A recent effort greatly improved the reference Chinese hamster genome assembly by combining Pacific Biosciences Single Molecule Real Time (SMRT) and short-read Illumina sequencing data, thus reducing the number of scaffolds by 28-fold and filling 95% of the sequence gaps[8]. Despite these great improvements in the assembly, the current genome annotation was based primarily on *ab initio* prediction, protein homology, ESTs, and limited publicly available transcriptomic data. However, these pipelines have difficulties in translation confirmation, splice form detection, and complete novel gene identification[9]. To improve cell line engineering success, an accurate genome annotation is necessary.

Proteogenomics provides a way to address such challenges by integrating mass spectrometry-based proteomics, RNA-Seq, and genomic data. For example, peptides can be identified by mapping tandem mass spectra to protein databases derived from RNA-Seq and genome annotation. The peptides are then used to update the annotation with novel coding regions and splice sites. Proteogenomics was first applied to *Mycoplasma pneumoniae*[10] to identify new and extended open reading frames (ORFs) and remove low quality gene models. It has also been applied to many eukaryotes including plants[11], yeast[12], and human[13]. In addition to improving annotations, the proteomic data can also identify proteomic variation (e.g., in cancer[14] and post translational modifications[15,16]).

In addition to proteomics data, Ribo-Seq (data from sequencing ribosome-protected coding reactions at single nucleotide resolution[17]), provides a global view of actively translated mRNAs, and thus has been used to predict ORFs and translation frames for proteins[18] and to identify additional predicted proteins for proteogenomics[19]. When analyzed together, transcriptomic, Ribo-Seq, and proteomic data can be invaluable for refining annotation about proteins.

To obtain a data-supported refinement of the Chinese hamster genome annotation, here we integrated proteomics, RNA-Seq, and Ribo-Seq to verify coding regions, update gene models, identify novel translated genes, and verify protein-coding variants in different CHO cell lines (Figure 1). To further demonstrate the increased value of this resource, we investigated the challenge associated with the Food and Drug Administration (FDA) requirement to ensure that viral particles (particularly endogenous retroviral particles) are eliminated from the therapeutic protein product, which contributes to the high costs in bioprocessing[20]. Specifically, we identified all translated retrovirus particles in CHO cells, including previously unannotated translated loci, thus providing potential knockout targets to increase drug purity and reduce demands on viral clearance. This proteogenomic resource will be invaluable for future efforts to study and engineer CHO cells for bioprocessing.

Methods

Proteomic sample preparation

Proteomic data were acquired at two different locations using different protocols, different biological samples, and different treatments. This increased the diversity in spectra used for annotation. These are referred to as batch 1 and batch 2, as follow.

Tissue Sample Collection for batch 1

Chinese hamsters were generously provided by Dr. George Yerganian (Cytogen Research, Roxbury, MA). Hamsters were euthanized by CO_2 and verified by puncture. Harvested liver and ovary tissues were flash frozen on dry ice and stored at −80°C until analysis.

Cell Culture Sample Collection for batch 1

Suspension CHO cell lines (including CHO-S and CHO DG44) were grown in shake-flask batch culture. CHO-S cells were cultured in CD-CHO medium supplemented with 8mM glutamine (Thermo Fisher Scientific, Waltham, MA), and CHO DG44 cells were culture in DG44 medium supplemented with 2mM glutamine (Thermo Fisher Scientific, Waltham, MA). Samples were collected on day 2 for exponential phase and day 4/5 for stationary phase. Cells were incubated at 37°C, 8% CO_2, and 120RPM. For sample collection, approximately 3 million cells were spun down, washed with PBS on ice, frozen rapidly on dry ice, and stored at −80°C until analysis.

Cell lysate and Tissue Sample preparation for batch 1

Cell culture lysates and tissue samples were thawed on ice and suspended in 2% sodium dodecyl sulfate (SDS) supplemented with 0.1mM phenylmethane sulfonyl fluoride (PMSF) and 1mM ethylenediaminetetraacetic acid (EDTA), pH 7–8. Samples were lysed by

sonicating for 60 seconds at 20% amplitude followed by a 90 second pause (for three cycles). Protein concentration was measured with a bicinchoninic acid (BCA) protein assay after briefly spinning to remove cell debris. Three hundred micrograms of each sample were reduced in 10mM tris(2-carboxyethyl)phosphine (TCEP), pH 7–8, at 60°C for 1hr on a shaking platform. After bringing each sample to room temperature, iodoacetamide was added to alkylate the sample to 17mM final concentration for 30 minutes. Next, samples were cleaned using 10kDa filters to reduce the SDS concentration as suggested by the filter aided sample preparation (FASP) protocol[21]. The samples were finally digested using a trypsin/LysC enzyme mix at an enzyme to substrate ratio of 1:10 (Promega V507A, Madison, WI), overnight at 37°C on a shaking platform.

Identification of Proteins by Mass Spectrometry for batch 1

Digested peptides (100μg from each protein digest) were fractionated on a basic reversed phase column (XBridge C18 Guard Column, Waters, Milford, MA). Fractions were concatenated into 48 prior to second dimension LC and MS analysis. The use of fractionation with equal peptides in each was designed to mimic biological replicates for each sample. Tandem MS/MS analysis of the peptides was carried out on the LTQ Orbitrap Velos MS (Thermo Fisher Scientific, Waltham, MA) interfaced to the Eksigent nanoflow liquid chromatography system (Eksigent, Dublin, CA) with the Agilent 1100 auto sampler (Agilent Technologies, Santa Clara, CA). Peptides were enriched on a 2cm trap column (YMC, Kyoto, Japan), fractionated on Magic C18 AQ, 5μm, 100Å, 75μm × 15cm column (Bruker, Billerica, MA), and electrosprayed through a 15μm emitter (SIS, Ringoes, NY). The reversed phase solvent gradient consisted of solvent A (0.1% formic acid) with increasing levels of solvent B (0.1% formic acid, 90% acetonitrile) over a period of 90 minutes. LTQ Orbitrap Velos parameters included 2.0kV spray voltage, full MS survey scan range of 350–1800m/z, data dependent HCD MS/MS analysis of the top 10 precursors with a minimum signal of 2000, isolation width of 1.9, 30s dynamic exclusion limit and normalized collision energy of 35. Precursor and fragment ions were analyzed at 60000 and 7500 resolutions, respectively.

Cell Culture Sample Collection for batch 2

Chinese hamster ovary cell clones that produce a recombinant monoclonal humanized IgG with different specific productivities (qP) were a generous gift from an industrial collaborator. These cell lines were developed by cotransfecting two plasmids, one containing IgG heavy chain (HC) and dihydrofolate reductase (DHFR) genes and the other containing IgG light chain (LC) and neomycin phosphotransferase (Neo) genes. Transfected cell lines were initially selected in medium containing 400 ug/mL neomycin (G418). After selection, the neomycin was removed, and all subsequent cultures were performed in the absence of neomycin. Subsequently, gene amplification was performed by stepwise selection with increasing methotrexate concentrations. Cell clones A1 and A1 have been previously described[22,23] (Jiang et al. 2006; Jiang and Sharfstein 2009). Cells were cultured in a serum-free modification of DME-F12 and alpha MEM as described by Dahodwala et al. with Glutamax (ThermoFisher Scientific) used instead of glutamine and 5 mg/l of insulin used. Methotrexate was not added during the culture period.

Cell clones were seeded at~ 0.1 × 106 cells/mL into six flasks per cell line and grown at 37°C, 5% CO2 in 125 ml shake-flasks (Thomson Scientific) shaken at 125 rpm. Three flasks per cell line were harvested in exponential phase (74 hours) and the remaining three flasks in stationary phase (96 hours). Cells were harvested by centrifugation and prepared for proteomic studies as described below.

In-solution digestion of whole cell lysate for proteomics batch 2

A second batch of samples were prepared and analyzed using a different approach. For these, 1mg of protein sample was transferred to a centrifuge tube, and all samples were equalized to the same volume using the same lysis buffer. A fresh stock of 0.5M reducing agent dithiothreitol (DTT) was prepared, and an appropriate volume of DTT was added to achieve a final concentration of 5mM. Samples were incubated for 25 minutes at 56°C. Before alkylation samples were cooled to room temperature and an appropriate volume of freshly prepared 0.5M iodoacetamide was added to a final concentration of 14mM and incubated for 30 minutes at room temperature. Untreated iodoacetamide was quenched by a second addition of 0.5M DTT to make total concentration of DTT equal to 10mM and incubated for 15 min at room temperature. The protein mixture was diluted 1:5 in 25 mM Tris-HCl, pH 8.2, to reduce the concentration of urea to 1.6 M. A double digestion by trypsin was performed by adding trypsin to 1/50 enzyme: substrate ratio and incubated at 37°C. After 4 hours of primary incubation the trypsin was topped up (enzyme: substrate ratio 1/100), and the protein mixture was left to digest overnight at 37°C. After the overnight digestion, unused trypsin was quenched by adding TFA to a final concentration of 0.4%.

Peptide sample clean-up for proteomics batch 2

Digested peptides were desalted and cleaned up using Sep-Pak c18 Vac cartridge, 200mg sorbent per cartridge, 55–500 ^m Particle size (WAT054945) using negative pressure. The C18 cartridge was washed and conditioned by using 9ml of ACN followed by 3ml of 50%ACN and 0.5% acetic acid. C18 resin was then equilibrated with 9ml of 0.1% TFA and samples were loaded in 0.4%TFA. Loaded samples were desalted with 9ml 0.1%TFA. TFA was removed with 1ml 0.5% acetic acid. Desalted peptides were eluted with 3ml of 50%ACN 0.5% acetic acid. The eluted fraction was applied twice and collected in a 15ml conical tube. The eluate was snap frozen in liquid nitrogen and lyophilized overnight or until the white (sometimes yellow) fluffy powder was observed. Dried peptides were stored at −20°C or otherwise dissolved in the appropriate buffer for phosphopeptide enrichment.

Draft genome annotation generation

68 RNA-Seq samples from multiple CHO cell lines and hamster tissues were trimmed and aligned to the newly assembled hamster reference genome using Trimmomatic 0.36[24] and STAR 2.5.2[25], respectively. The aligned reads were assembled into transcripts using stringtie v1.3.1c[26] for each sample and then the transcripts were consolidated into a union transcript set using stringtie-merge[26]. To improve the transcript coverage, we also mapped the hamster RefSeq (GCF_000419365.1, annotation updated in December 2017) transcript sequences to the newly assembled hamster genome using gmap 2018–07-04[27], which were then integrated with transcripts generated from stringtie-merge using the PASA pipeline[28]. Potential proteins encoded in the transcripts were predicted using transdecoder[29]. Finally the

functions of predicted proteins were determined by mapping them to the hamster RefSeq proteins and UniProt Swiss-Prot proteins[30] using BLASTP 2.7.1+[31]. Proteins whose mapping lengths were greater than 80% were considered and percentage identity of 60% when mapping to hamster RefSeq and of 50% when mapping to UniProt were used as threshold to further classify proteins into 4 categories with 1 to 4 indicating decreasing confidence scores as follows: 1. Percentage of identity (pident) and percentage of length (plen) are larger than the threshold and have the same gene name between hamster RefSeq and UniProt. 2. Pident and plen are larger than the threshold and have different names between hamster RefSeq and UniProt. 3. Pident and plen are less than the threshold and have the same gene name between hamster RefSeq and UniProt. 4. Pident and plen are less than the threshold and have different gene names between hamster RefSeq and UniProt. LncRNAs were predicted by aligned transcripts to hamster lncRNAs using BLASTN 2.7.1+ with pident larger than 60% and plen larger than 80%.

Proteogenomics database construction

We prepared 4 protein databases for mass spectrum matching: a known protein database (KnownDB), SNP database (SnpDB), splice database (SpliceDB) and a six-frame translation of the genome database (SixframeDB). The KnownDB includes protein sequences extracted from our draft annotation, while the rest serve as novel protein databases. SnpDB was constructed by translating RNA-Seq reads that have mutations[32]. Mutations were called using GATK3.7[33] and annotated using Annovar[34]. The SpliceDB was constructed by translating all RNA-Seq reads that span splice junctions[35]. The SixframeDB was derived from peptides fragments between stop codons in all frames of the reference genome assembly.

Peptide identification The original MS/MS spectra were converted from RAW format to Mascot Generic Format (MGF) using msconvert[36] and searched against each database independently using MSGF+[37]. Since different databases have different false discovery rates, it is recommended to perform multistage FDR correction with 1% cut off for the databases[14], which means the spectra failed to pass FDR correction were fed to the next database to correct again. We corrected FDR for databases in PSM level in the following order: KnownDB, SnpDB, SpliceDB, SixframeDB. Then for the final known and novel PSM results we corrected FDR in the peptide and protein levels.

New translational event prediction

Significant peptide-spectrum matches (PSM) against the novel databases were used to discover new translation events using Enosi pipeline[35]. Briefly, identified novel peptides were mapped to the novel databases to get the loci relative to their mapped proteins. Protein headers in the novel databases have loci relative to the reference genome assembly. A custom python script was used to deduce peptide loci relative to the reference genome assembly from those two loci. Then the loci were compared with the draft annotation to decide event type. Peptides in the same translation frame that are less than 300 nucleotides (100 amino acids) away are grouped to represent the same event. For SNPs and short INDELs, we filtered the false positives by variant calling using Illumina reads from sequencing hamster genomic DNA against the reference genome assembly. To determine the

confidence of the new translational events, we used the following equation to calculate the probability of an event being true:

$$P\_\text{event} = 1 - \prod_{i=1}^{N}\left(1 - \frac{1}{h_i}\right)$$

h_i represents the number of loci a peptide hits in the database. If the probability is greater than 0.5, we consider the event is true.

Protein prediction using Ribo-Seq data

We used previously published Ribo-Seq data of the CHO CS CS13–1.0 cell line[38]. All Ribo-Seq and RNA-seq data of each biological sample were trimmed and aligned to the reference genome assembly using Trimmomatic 0.36[24] and HISAT 2.2.1[39] respectively. The aligned bam files were sorted and merged into one Ribo-Seq and one RNA-Seq bam files using Samtools 1.6[40]. The potential translated regions were predicted using RiboTaper[41], which takes advantage of coverage in coding regions and triplet periodicity of ribosomal footprints. A custom python script was used to compare the translation prediction and draft annotation.

We treated the predicted proteins from Ribo-Seq as a novel protein database and combined it with the KnownDB. Then we mapped all the mass spectra to these two databases using the proteogenomics pipeline. In this case, the identified novel peptides would be the unique peptides from the Ribo-Seq database.

Retrovirus in the draft annotation

All viral proteins in the draft annotation were identified by mapping to UniProt and hamster RefSeq proteins using BLASTP. Then the retroviral proteins were identified based on the full gene names. Peptides supporting annotated retroviral proteins were identified by mapping the known peptides to the KnownDB. LTRs were predicted using LTRharvest[42]. Retroviral proteins located between LTR were identified by overlapping LTR regions with retroviral annotations using Bedtools 2.27[43].

New retrovirus discovery

Since viral proteins lack introns, we filtered novel peptides by removing those with splice sites. Retroviral proteins and filtered novel peptides were aligned to the reference genome assembly using tblastn[44] and mappings with more than 60 pident and 55 plen were considered for downstream analysis. Virus mapping and peptide mapping were overlapped using Bedtools[43] to get the virus sites with peptide support, which were then further overlapped with LTR regions to get virus sites with both LTR and peptides support.

We decided the thresholds for virus tblastn mapping as follows (Supplementary Figure S6). First, we started with low thresholds (30% for each). Then we assessed the overlap between filtered mapping and the draft annotation to identify those associated with known viral genes. If the mappings overlap with many non-viral genes, thresholds were incrementally increased and overlapped with draft annotation again. This was repeated until the mappings overlap with few non-viral genes and the number ceased to decrease.

Discovery of unique retroviruses in the CHO-S cell line

In-house CHO-S SMRT sequence data were used to check if CHO-S has unique retroviral elements, compared to hamster. Firstly, we subsampled hamster SMRT reads to the same depth as CHO-S SMRT data, and both datasets were corrected using Illumina paired-end reads[3] through LoRDEC[45]. Secondly, retroviral proteins and filtered novel peptides from our proteogenomics pipeline were mapped to corrected CHO-S and hamster SMRT reads using the same threshold as the previous tblastn mapping. Thirdly, viral and peptide mappings were overlapped to get virus sites with peptide support for CHO-S and hamster separately. Fourthly, we filtered CHO-S virus sites with hamster SMRT virus sites and mapped the unique CHO-S sites to the reference genome assembly. The mapped sites are the new retroviral sites and the unmapped sites are unique retroviral elements in CHO-S.

Type-C retrovirus detection in CHO cell lines

The functions of all the identified retroviral proteins were determined by mapping the protein sequences to UniProt using BLASTP. The full function descriptions of the proteins have the organism resource. Therefore, the types of all the retroviruses were determined by manually matching their full virus names to the types defined previously (see appendix “Retroviral Taxonomy, Protein Structures, Sequences, and Genetic Maps” of the following book: [46]). Peptide coverage of a protein is defined as the number of amino acids covered by peptides divided by the protein length. The RNA-Seq coverage along the protein body was calculated using pysam[47].

Results

Draft annotation for the genome

The CHO-K1[3] and Chinese hamster genome sequences[4,5] were originally assembled using short read (99bp or 150bp) technology, and therefore resulted in fragmented contigs and scaffolds. Thus, efforts to annotate the genomes resulted in errors in protein and gene models. The RefSeq pipeline has corrected some such errors; however, the complete reassembly of the Chinese hamster genome[8] provides an opportunity to obtain a much improved annotation of coding regions in the genes and their corresponding protein sequences. Therefore, we generated a new draft annotation here (Supplementary Figure S1), including predicted protein sequences.

68 RNA-Seq samples were prepared from multiple CHO cell lines and hamster tissues (**see** Methods). Transcripts were assembled for each sample separately using stringtie[26] and merged using stringtie-merge[26], yielding 26,530 genes with 68,082 transcripts. Then 38,654 hamster RefSeq transcripts were mapped to the newly assembled hamster reference genome using GMAP[27]. Finally, the RefSeq alignments and RNA-Seq assembled transcripts were merged to 86,790 transcripts and grouped into 38,511 genes based on genomic locations using Program to Assemble Spliced Alignments (PASA).

We then applied TransDecoder[29] to the 86,790 transcripts and predicted that 63,331 of them encode proteins. These proteins have 47,829 unique protein sequences. To annotate the function of the proteins, we aligned the protein sequences to the hamster RefSeq and

UniProt Swiss-Prot protein databases using BLASTP[31]. (Supplementary Table S1). We assigned UniProt gene names to the proteins in our draft annotation except for those that only map to hamster RefSeq proteins. Furthermore, we identified 4,640 non-coding transcripts by aligning the transcripts to the hamster RefSeq non-coding transcripts using BLASTN.

Proteogenomics identifies novel proteins in the draft annotation

To quantify novel proteins predicted in the draft annotation compared to the hamster RefSeq proteins, we mapped 47,829 unique draft proteins to the RefSeq proteins using BLASTP. We classified the mappings into 5 main categories: (1) 15,787 perfectly mapped proteins; (2) 7,483 proteins mapping perfectly on only one end between the draft and RefSeq sequences; (3) 11,780 high quality mapped proteins (over 90% percentage of identity (pident) and over 80% percentage of length (plen) on both sides of homology protein mapping pair between draft and RefSeq); (4) 6,688 high quality mapped proteins (over 90% pident and over 80% plen on either side), but only mapping well on one end; (5) 5,820 low quality or non-mapping proteins. 289 proteins failed to map. We defined proteins that were not in category (1) as novel proteins. Among the one-sided perfect mapping proteins, 3,336 proteins are shorter in the draft, compared to RefSeq, while 4,147 proteins are longer than RefSeq. Interestingly, isoforms of some of the former proteins map perfectly to RefSeq. This indicates the draft annotation pipeline is sensitive to splice sites, which resulted in more isoforms being assembled than seen in RefSeq.

Next, we sought peptide support for the novel proteins in the draft annotation. We acquired and prepared12,870,725 mass spectra from multiple CHO cell lines and hamster tissues. We merged the draft and RefSeq proteins and extracted the unique protein sets as a reference protein database. Then we used MS-GF+[37] to search the peptide-spectrum matches (PSMs) with 1% FDR correction and identified 205,294 peptides with 1% FDR correction in peptide level. Here we only consider proteins with at least two uniquely-mapping peptides of at least 9 amino acids in length[48]. For each pair of homologous draft and RefSeq proteins, the draft protein was considered as novel if it has extra peptide support compared to the corresponding RefSeq protein. As a result, we identified 3,529 draft novel proteins, 3,389 of which have additional unique peptide support not seen in the RefSeq sequence (Figure 2A). The remaining 140 novel proteins have extra peptides mapping to multiple locations. (Figure 2B). The high-quality protein mappings category (>90% pident and >80% plen) has the most novel proteins. 5,608 draft proteins have the same peptide support as similar RefSeq proteins, which may require additional data to verify their novel features. The numbers of novel proteins in each mapping category are depicted in Figure 2.

Proteogenomics and Ribosome profiling identify additional translational events

In addition to verifying novel protein sequences, proteomics can also help identify other translational events, e.g., novel splice sites, gene fusions, etc. (Figure 3A). Thus, to obtain a more comprehensive view of protein sequence verification and identification of other translation events, we created 4 putative protein databases: (1) KnownDB-predicted from draft annotation, (2) SnpDB-translated from RNA-Seq reads that have non-synonymous mutations and short INDELs, (3) SpliceDB-translated from spliced RNA-Seq reads, and (4)

SixframeDB-peptides between stop codons in all 6 frames of the genome (Figure 1). As previously recommended[14], we performed multi-stage 1% FDR correction for the databases sequentially (Supplementary Figure S2), and 1% FDR correction in the peptide level. This pipeline (Supplementary Figure S3) identified 3,656,801 (28%) significant PSMs resulting in 194,470 significant unique peptides mapping to the KnownDB and 8,003 peptides mapping to the remaining databases (Figure 3B). Among all the peptides identified from KnownDB, 168,862 (87%) map to unique genomic locations.

We required each validated protein to have at least two uniquely mapping peptides with at least 9 amino acids (i.e., to only one locus) as recommended in Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1[48]. Using this strategy, we verified 35,112 proteins, which represent 73.4% of the sequences in the KnownDB.

After known protein sequence validation, we explored additional novel peptide events. To guarantee high confidence of the events, we filtered out those with fewer than 3 RNA-Seq supporting reads and required each event to have at least one uniquely mapped peptide. Most proteins are longer than 100 amino acids. Thus, if novel peptides are close and in the same translational frame, they are likely to support the same protein. Therefore, we clustered novel identified peptides that were in the same translational frame and fewer than 100 amino acids away from each other to represent the same event. In total, we discovered 2,256 new translational events, 86.5% of which are novel splice and nonsynonymous single nucleotide polymorphisms (SNPs) (Figure 3A). Novel splice sites represent 44% of the total events, covering 857 genes. We also identified 54 alternative splice events, 20 reverse strand translation events, 6 novel ORFs (not predicted by the draft annotation) and 4 gene fusion events.

Ribo-Seq offers orthogonal evidence to further support protein sequences and can help discover new proteins as well. Ribo-Seq and corresponding RNA-Seq data sets for an IgG-producing CHO cell line were acquired at both exponential and stationary phase[38]. We used RiboTaper[41] to predict the translating ORFs under the guidance of the draft annotation. 28,700 transcripts were predicted to encode proteins, with 24,709 (86%) having a single ORF (Figure 3C). Among these, 13,666 transcripts have the same protein sequences as the draft annotation. In addition, 1,318 “non-coding transcripts” in the draft annotation are predicted to encode proteins. The remaining Ribo-Seq predicted sequences were classified into two groups: (1) in the same frame (8775) and (2) in different frames (950) with the draft annotation (Figure 3D, Supplementary Table S2).

Proteins predicted by Ribo-Seq can expand the databases for proteomics so that more proteins can be verified. Here we used Ribo-Seq predicted proteins as a novel database and ran the proteogenomics pipeline together with KnownDB. After filtering all peptides identified in the previous proteogenomics pipeline, the Ribo-Seq data facilitated the identification of 2,286 new peptides. Here we require at least two unique peptides with length of at least 9 amino acids in an ORF to verify translation. Among 1,628 non-protein-coding transcripts in the draft annotation, 218 were verified to encode proteins by Ribo-Seq and proteomics. While Ribo-Seq enabled the successful identification of many new peptides, including those in transcripts previously thought to be non-coding, it was less successful at

identifying the translation start sites. Supporting peptides were found for only 8 out of the 305 ORFs that were predicted to be longer than the draft annotation. In addition, Ribo-Seq helped identify translation events in 5'UTR and 3'UTR regions in 213 genes, consistent with previous reports in human[49]. For more details, see Supplementary Table S2.

Proteomic-based validation of SNPs and INDELs in CHO cell lines and hamster tissues

The peptides obtained from proteomic studies enabled the validation of genetic variants in the various CHO cell lines and hamster tissues. To discover these mutations, we used our SnpDB (Figure 1) as the novel database in the proteogenomics pipeline, which includes peptides translated from all the RNA-Seq reads supporting SNPs and small INDELs. Proteomics identified fewer mutations than RNA-Seq (Supplementary Table S3), mainly because of its lower depth of coverage compared to RNA-Seq. Furthermore, mutated proteins can be degraded, and therefore not detected. In total we identified 959 nonsynonymous SNPs, located in 722 genes. Most genes have one SNP while there are 6 genes with more than 5 SNPs: GAPDH, GOLGB1, AHNAK2, PKM, MYH9 and EEF1A1. Surprisingly, only one protein lost a stop codon (ribosome protein gene RPS23) while others change amino acids. 75% of the SNPs are homozygous, which indicates CHO cell lines may have developed and retained those mutations after long periods of evolution. Furthermore, the distributions of the 6 SNP types showed that transitions occurred more frequently than transversions, and the proteomic data captured a similar distribution of mutations as RNA-Seq data (Figure 4A, Figure 4B). We identified 43 insertions and 6 deletions, located in 42 and 6 genes respectively. There are more homogeneous frameshift INDELs than other types. The 8 genes harboring both SNPs and insertions include AHNAK2, CALR, HNRNPUL1, HSP90B1, PLEKHG5, PTMA, RIF1, and VAT1, while only SIK3 had both SNPs and deletions.

Next, we looked at the mutation distribution across the protein body. Since there are far fewer INDELs than SNPs, we focused on the distribution of SNPs. Figure 4C shows that SNPs are distributed relatively evenly across the protein body.

Many different CHO cell lines (e.g., CHO-K1, CHO-S and DG44), have been used to develop different recombinant protein-producing cells. Each cell line has a lengthy history of mutation and selection during cell line development[4], and therefore can have unique genomic variants[4,50]. To check the variations between these CHO cell lines, we compared their peptide-supported SNPs in the coding regions. Most of the SNPs are shared among the cell lines, which means these variants have been conserved during the long period of cell line development, either due to early mutations obtained in CHO cells when derived in 1957 or genetic drift of the Chinese hamster colony since then (Figure 4D).

Proteomics elucidate translated retroviral elements in the genome

For decades, it has been known that CHO cells shed retroviral particles[51]; while these were shown to be non infectious[52,53], the safety concern has required companies to filter out all such viral particles and conduct extensive testing to verify non-infectivity. This adds a substantial cost to production. Viral particles have been isolated, but the few mRNAs that have been sequenced from these particles were all non-coding, in that they contained many

early stop codons. Thus, it remains unclear which loci encode the translated viral particles. Here, we analyzed the transcriptomic and proteomic data to identify the loci of expressed and translated retroviral particles, to enable further efforts to eliminate these particles and reduce drug purification costs.

To identify translated endogenous retroviruses in CHO cells, we first extracted peptides that support retroviral proteins from our draft annotation (Figure 5A). Since retroviruses can have multiple similar copies across the whole genome, we consider peptides that can map to multiple locations. We found 457 retroviral genes covered by 723 transcripts, 151 of these genes (corresponding to 209 transcripts) have peptide support and 104 transcripts map well to RefSeq (>60% pident, >80% plen) and UniProt (>50% pident, >80% plen). Infectious retroviral DNA is flanked by two identical non-coding repeats called long terminal repeats (LTR)[54], which aid in retroviral mobility and integration into the host genome, along with regulation of retroviral gene expression. In the hamster genome, we identified 3,324 LTR pairs in the reference genome using LTRharvest[42] and found only 40 retroviral transcripts locate between those LTRs, all of which have peptide support. If one LTR is disrupted, the other side can still effectively induce transcription. Thus, genes not flanked by LTR pairs can still produce retroviral particles.

We next aimed to identify unannotated translated retroviral sites in the genome (Figure 5A). For this, we aligned all retroviral proteins from NCBI to the reference genome using tblastn[44]. Then we overlapped the mapping sites with the 4,265 novel peptides identified against novel protein databases from our proteogenomics pipeline and obtained 41 novel retroviral sites, 1 of which localized between an LTR pair and was covered by 5 peptides. The site showed homology to the gag proteins from Gibbon ape leukemia virus and Spleen focus-forming virus. Finally, since CHO cells were originally derived in 1957, we further checked if new infections may have emerged in CHO (Supplementary Figure S4). For this analysis, we aligned all known retroviral proteins and novel peptides from our proteogenomics pipeline, using tblastn, to in-house Illumina-corrected single molecule real time (SMRT) sequence data from CHO-S and the Chinese hamster[8]. After filtering out the putative viral sites identified in hamster, we found no evidence that wild type CHO-S cell lines have acquired any new retroviral sites.

Mammalian retroviruses have been classified into different types based on the genomic compositions. We mapped retroviral protein sequences identified from previous steps to UniProt and used protein full names to discover 3 main retroviral types in hamster: type-A, type-B, and type-C[55]. We found type-C retroviruses to be the most highly transcribed and translated (Supplementary Figure S5). In addition, type-C viral particles have been identified in CHO cell lines before and regulatory agencies now require the verification that products are non-infectious type-C particles[53,56]. The vast majority of type-C proteins had little or no peptide support, suggesting these are silenced or noncoding. Of the 119 type-C retroviral proteins with peptide support that we identified, most were gag and envelope proteins (Figure 5B). Only 4 proteins had more than 20 detected peptides, and most proteins had low coverage of supporting peptides (Figure 5C). Figure 5D shows an example of highly translated envelope protein with 33 peptides, covering 30% of the coding sequence. Although it has many secondary reads, it also has many uniquely mapping reads, which

indicates this locus is truly expressed. The proteins with high coverage and more supported peptides should be prioritized in efforts to eliminate viral particle production. The genomic locations, RNA-Seq and peptide coverage of all endogenous retroviral genes are provided in Supplementary Table S4.

Discussion and Conclusion

Here we presented the first proteogenomic reannotation of the Chinese hamster genome, in which we utilized RNA-Seq, Ribo-Seq, and proteomics to improve the annotation. To identify as many peptide-supported proteins as possible, we mapped spectra to a known protein database from a draft annotation and several novel protein databases derived from different data types. We identified 3,529 novel proteins in the draft annotation compared to the hamster RefSeq protein database and 2,256 novel translational events and mutations in hamster and CHO cell lines. Furthermore, we identified the potential sources of retroviral particles shed from CHO cells, including 119 type-C retrovirus genes, 4 of which are supported by more than twenty peptides.

Usually an annotation is required before running proteogenomics pipeline. The typical first step of genome annotation is masking the repeats to avoid getting millions of seeds during BLAST[57]. However, masking the genome can hide important annotation information, including common domains and retroviral elements. To avoid this loss of information, we aligned assembled transcripts to the unmasked genome using gmap. Doing so resulted in only 0.7% transcript mappings to be ambiguous (i.e., with >2 mappings). This enabled the annotation and analysis of endogenous retroviral genes, which usually have multiple similar copies. Masking repeats often removes such information[58]. Thus, future annotation efforts would benefit from the acquisition and alignment of high-quality transcripts or protein sequences to the genome of interest.

Different pipelines have been designed for genome annotation in higher eukaryotes[57], and most include *ab initio* prediction, mapping of homologous protein sequences, and transcript assembly and alignment. As the throughput and resolution of mass spectrometry techniques is increasing, more studies are integrating these data types to refine annotations[59]. Here we discovered thousands of genes and novel translational events using proteomics data. Despite their value, proteomics data can be sparse since the chemical properties of some peptides are less compatible with the experimental setup (e.g., due to hydrophobicity) or size of the peptides post-digestion[60]. Furthermore, many peptides have diverse post-translational modifications, thus making it difficult to align peptides to predicted peptide sequences from genomes. Thus, Ribo-Seq provides a complementary method to further discover new genes or correct annotations[61]. The Ribo-Seq datasets we used here are from the DG44 CHO cell line, and CHO cells may only express half to two thirds of their genes[62,63]. Thus, further annotation efforts would benefit from the acquisition of Ribo-Seq from many other hamster tissues and developmental stages. In summary, as more tools and proteomic and Ribo-Seq data accumulates, these data types will become increasingly integrated into standard pipelines for genome annotation.

Finally, the reannotation here provides an invaluable resource for the development of improved CHO cell lines. While CHO cells provide several advantages as an expression host for recombinant protein production, HCPs are continuously secreted[64,65], thus impacting recombinant protein quality and safety. Thus, expensive chromatographic columns, filtration systems, and assays for infection and HCPs are required during downstream processing. This adds considerable cost to biopharmaceuticals. One particular regulatory concern has been the endogenous retroviruses that are shed by CHO cells[56]. For these, assays were developed to quantify retroviral particles and ensure infectious retroviral particles are not found in the drug product after extensive filtration[66]. Here, we identified, among hundreds of retroviral genes in the hamster genome, which ones are expressed and translated in several CHO cell lines. This information will enable future efforts to remove these from CHO cells, thereby reducing burdens to downstream processing. To further facilitate such efforts, many endogenous retroviral genes show high levels of homology. In our work, this was manifested in the identification of many retroviral RNA-Seq reads and tryptic peptides that map to multiple genomic loci. As many of these also share DNA sequence, multiple viruses can be knocked out simultaneously by targeting the conserved regions, as accomplished in the pig[67]. Doing this in CHO cells can reduce the costs by simplifying expensive purification steps where product can also be lost, and also simplify viral testing steps for the final product.

In conclusion, our work provides a refined and more extensive annotation of the Chinese hamster genome, which will enable more accurate CHO cell line engineering[6,68–70], as systems are targed such as glycosylation[71–74] and apoptosis pathways[75,76]. The improved annotation will also facilitate improved processing of omics data[77] as the gene models improve. Finally, a more complete list of all genes will enable efforts to map out molecular pathways in CHO cells to enable systems approaches to cell line development[6,78].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This work was supported by generous funding from the Novo Nordisk Foundation provided to the Center for Biosustainability at the Technical University of Denmark (grant no. NNF16CC0021858), and SL was supported with funding from the Frontiers of Innovation Scholars Program at UCSD. VB was supported in part by grants from the NIH 1R01GM114362, and P-41-RR24851. M. Betenbaugh also acknowledges support from AstraZeneca.

References

(1). Golabgir A; Gutierrez JM; Hefzi H; Li S; Palsson BO; Herwig C; Lewis NE Quantitative Feature Extraction from the Chinese Hamster Ovary Bioprocess Bibliome Using a Novel Meta-Analysis Workflow. Biotechnol. Adv. 2016, 34 (5), 621–633. 10.1016/j.biotechadv.2016.02.011. [PubMed: 26948029]

(2). Lin FK; Suggs S; Lin CH; Browne JK; Smalling R; Egrie JC; Chen KK; Fox GM; Martin F; Stabinsky Z Cloning and Expression of the Human Erythropoietin Gene. Proc. Natl. Acad. Sci. U. S. A. 1985, 82 (22), 7580–7584. [PubMed: 3865178]

(3). Xu X; Nagarajan H; Lewis NE; Pan S; Cai Z; Liu X; Chen W; Xie M; Wang W; Hammond S; et al. The Genomic Sequence of the Chinese Hamster Ovary (CHO)-K1 Cell Line. Nat. Biotechnol. 2011, 29 (8), 735–741. 10.1038/nbt.1932. [PubMed: 21804562]

(4). Lewis NE; Liu X; Li Y; Nagarajan H; Yerganian G; O'Brien E; Bordbar A; Roth AM; Rosenbloom J; Bian C; et al. Genomic Landscapes of Chinese Hamster Ovary Cell Lines as Revealed by the Cricetulus Griseus Draft Genome. Nat. Biotechnol. 2013, 31 (8), 759–765. 10.1038/nbt.2624. [PubMed: 23873082]

(5). Brinkrolf K; Rupp O; Laux H; Kollin F; Ernst W; Linke B; Kofler R; Romand S; Hesse F; Budach WE; et al. Chinese Hamster Genome Sequenced from Sorted Chromosomes. Nat. Biotechnol. 2013, 31, 694. [PubMed: 23929341]

(6). Kuo C-C; Chiang AW; Shamie I; Samoudi M; Gutierrez JM; Lewis NE The Emerging Role of Systems Biology for Engineering Protein Production in CHO Cells. Curr. Opin. Biotechnol. 2018, 51, 64–69. https://doi.org/10.1016Zj.copbio.2017.11.015. [PubMed: 29223005]

(7). Stolfa G; Smonskey MT; Boniface R; Hachmann A-B; Gulde P; Joshi AD; Pierce AP; Jacobia SJ; Campbell A CHO-Omics Review: The Impact of Current and Emerging Technologies on Chinese Hamster Ovary Based Bioproduction. Biotechnol. J. 2017, 1700227 10.1002/biot.201700227.

(8). Rupp O; MacDonald ML; Li S; Dhiman H; Polson S; Griep S; Heffner K; Hernandez I; Brinkrolf K; Jadhav V; et al. A Reference Genome of the Chinese Hamster Based on a Hybrid Assembly Strategy. Biotechnol. Bioeng. 2018, 115 (8), 2087–2100. 10.1002/bit.26722. [PubMed: 29704459]

(9). Castellana N; Bafna V Proteogenomics to Discover the Full Coding Content of Genomes: A Computational Perspective. J. Proteomics 2010, 73 (11), 2124–2135. 10.1016/jjprot.2010.06.007. [PubMed: 20620248]

(10). Jaffe JD; Berg HC; Church GM Proteogenomic Mapping as a Complementary Method to Perform Genome Annotation. Proteomics 2004, 4 (1), 59–77. 10.1002/pmic.200300511. [PubMed: 14730672]

(11). Castellana NE; Payne SH; Shen Z; Stanke M; Bafna V; Briggs SP Discovery and Revision of Arabidopsis Genes by Proteogenomics. Proc. Natl. Acad. Sci. 2008, 105 (52), 21034–21038. [PubMed: 19098097]

(12). Yagoub D; Tay AP; Chen Z; Hamey JJ; Cai C; Chia SZ; Hart-Smith G; Wilkins MR Proteogenomic Discovery of a Small, Novel Protein in Yeast Reveals a Strategy for the Detection of Unannotated Short Open Reading Frames. J. Proteome Res. 2015, 14 (12), 5038–5047. 10.1021/acs.jproteome.5b00734. [PubMed: 26554900]

(13). Kim M-S; Pinto SM; Getnet D; Nirujogi RS; Manda SS; Chaerkady R; Madugundu AK; Kelkar DS; Isserlin R; Jain S; et al. A Draft Map of the Human Proteome. Nature 2014, 509 (7502), 575–581. 10.1038/nature13302. [PubMed: 24870542]

(14). Woo S; Cha SW; Bonissone S; Na S; Tabb DL; Pevzner PA; Bafna V Advanced Proteogenomic Analysis Reveals Multiple Peptide Mutations and Complex Immunoglobulin Peptides in Colon Cancer. J. Proteome Res. 2015, 14 (9), 3555–3567. 10.1021/acs.jproteome.5b00264. [PubMed: 26139413]

(15). Cesnik AJ; Shortreed MR; Sheynkman GM; Frey BL; Smith LM Human Proteomic Variation Revealed by Combining RNA-Seq Proteogenomics and Global Post-Translational Modification (G-PTM) Search Strategy. J. Proteome Res. 2016, 15 (3), 800–808. 10.1021/acs.jproteome. 5b00817. [PubMed: 26704769]

(16). Kaushik P; Henry M; Clynes M; Meleady P The Expression Pattern of the Phosphoproteome Is Significantly Changed During the Growth Phases of Recombinant CHO Cell Culture. Biotechnol. J. 2018, 13 (10), e1700221 10.1002/biot.201700221. [PubMed: 30076757]

(17). Ingolia NT; Ghaemmaghami S; Newman JRS; Weissman JS Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science 2009, 324 (5924), 218–223. 10.1126/science.1168978. [PubMed: 19213877]

(18). Calviello L; Ohler U Beyond Read-Counts: Ribo-Seq Data Analysis to Understand the Functions of the Transcriptome. Trends Genet. TIG 2017, 33 (10), 728–744. https://doi.org/10.1016Zj.tig. 2017.08.003. [PubMed: 28887026]

(19). Crappé J; Ndah E; Koch A; Steyaert S; Gawron D; De Keulenaer S; De Meester E; De Meyer T; Van Criekinge W; Van Damme P; et al. PROTEOFORMER: Deep Proteome Coverage through Ribosome Profiling and MS Integration. Nucleic Acids Res. 2015, 43 (5), e29 10.1093/nar/gku1283. [PubMed: 25510491]

(20). Strauss DM; Lute S; Brorson K; Blank GS; Chen Q; Yang B Removal of Endogenous Retrovirus-like Particles from CHO-Cell Derived Products Using Q Sepharose Fast Flow Chromatography. Biotechnol. Prog. 2009, 25 (4), 1194–1197. 10.1002/btpr.249. [PubMed: 19452543]

(21). Wiśniewski JR; Zougman A; Nagaraj N; Mann M Universal Sample Preparation Method for Proteome Analysis. Nat. Methods 2009, 6 (5), 359–362. 10.1038/nmeth.1322. [PubMed: 19377485]

(22). Jiang Z; Huang Y; Sharfstein ST Regulation of Recombinant Monoclonal Antibody Production in Chinese Hamster Ovary Cells: A Comparative Study of Gene Copy Number, MRNA Level, and Protein Expression. Biotechnol. Prog. 2006, 22 (1), 313–318. 10.1021/bp0501524. [PubMed: 16454525]

(23). Dahodwala H; Nowey M; Mitina T; Sharfstein ST Effects of Clonal Variation on Growth, Metabolism, and Productivity in Response to Trophic Factor Stimulation: A Study of Chinese Hamster Ovary Cells Producing a Recombinant Monoclonal Antibody. Cytotechnology 2012, 64 (1), 27–41. 10.1007/s10616-011-9388-z. [PubMed: 21822681]

(24). Bolger AM; Lohse M; Usadel B Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. Bioinformatics 2014, 30 (15), 2114–2120. 10.1093/bioinformatics/btu170. [PubMed: 24695404]

(25). Dobin A; Davis CA; Schlesinger F; Drenkow J; Zaleski C; Jha S; Batut P; Chaisson M; Gingeras TR STAR: Ultrafast Universal RNA-Seq Aligner. Bioinforma. Oxf. Engl. 2013, 29 (1), 15–21. 10.1093/bioinformatics/bts635.

(26). Pertea M; Pertea GM; Antonescu CM; Chang T-C; Mendell JT; Salzberg SL StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads. Nat. Biotechnol. 2015, 33 (3), 290–295. 10.1038/nbt.3122. [PubMed: 25690850]

(27). Wu TD; Watanabe CK GMAP: A Genomic Mapping and Alignment Program for MRNA and EST Sequences. Bioinforma. Oxf. Engl. 2005, 21 (9), 1859–1875. 10.1093/bioinformatics/bti310.

(28). Haas BJ; Delcher AL; Mount SM; Wortman JR; Smith RK; Hannick LI; Maiti R; Ronning CM; Rusch DB; Town CD; et al. Improving the Arabidopsis Genome Annotation Using Maximal Transcript Alignment Assemblies. Nucleic Acids Res. 2003, 31 (19), 5654–5666. 10.1093/nar/gkg770. [PubMed: 14500829]

(29). Haas BJ; Papanicolaou A; Yassour M; Grabherr M; Blood PD; Bowden J; Couger MB; Eccles D; Li B; Lieber M; et al. De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis. Nat. Protoc. 2013, 8 (8), 1494–1512. 10.1038/nprot.2013.084. [PubMed: 23845962]

(30). Apweiler R; Bairoch A; Wu CH; Barker WC; Boeckmann B; Ferro S; Gasteiger E; Huang H; Lopez R; Magrane M; et al. UniProt: The Universal Protein Knowledgebase. Nucleic Acids Res. 2004, 32 (Database issue), D115–119. 10.1093/nar/gkh131. [PubMed: 14681372]

(31). Gish W; States DJ Identification of Protein Coding Regions by Database Similarity Search. Nat. Genet. 1993, 3 (3), 266–272. 10.1038/ng0393-266. [PubMed: 8485583]

(32). Woo S; Cha SW; Na S; Guest C; Liu T; Smith RD; Rodland KD; Payne S; Bafna V Proteogenomic Strategies for Identification of Aberrant Cancer Peptides Using Large-Scale next-Generation Sequencing Data. Proteomics 2014, 14 (23–24), 2719–2730. 10.1002/pmic.201400206. [PubMed: 25263569]

(33). Auwera G. A. V. der; Carneiro MO; Hartl C; Poplin R; Angel G.del; Levy-Moonshine A; Jordan T; Shakir K; Roazen D; Thibault J; et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr. Protoc. Bioinforma. 2013, 43 (1), 11.10.1–11.10.33. 10.1002/0471250953.bi1110s43.

(34). Wang K; Li M; Hakonarson H ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data. Nucleic Acids Res. 2010, 38 (16), e164 10.1093/nar/gkq603. [PubMed: 20601685]

(35). Woo S; Cha SW; Merrihew G; He Y; Castellana N; Guest C; MacCoss M; Bafna V Proteogenomic Database Construction Driven from Large Scale RNA-Seq Data. J. Proteome Res. 2014, 13 (1), 21–28. 10.1021/pr400294c. [PubMed: 23802565]

(36). Kessner D; Chambers M; Burke R; Agus D; Mallick P ProteoWizard: Open Source Software for Rapid Proteomics Tools Development. Bioinforma. Oxf. Engl. 2008, 24 (21), 2534–2536. 10.1093/bioinformatics/btn323.

(37). Kim S; Pevzner PA MS-GF+ Makes Progress towards a Universal Database Search Tool for Proteomics. Nat. Commun. 2014, 5, 5277 10.1038/ncomms6277. [PubMed: 25358478]

(38). Kallehauge TB; Li S; Pedersen LE; Ha TK; Ley D; Andersen MR; Kildegaard HF; Lee GM; Lewis NE Ribosome Profiling-Guided Depletion of an MRNA Increases Cell Growth Rate and Protein Secretion. Sci. Rep. 2017, 7, 40388. [PubMed: 28091612]

(39). Kim D; Langmead B; Salzberg SL HISAT: A Fast Spliced Aligner with Low Memory Requirements. Nat. Methods 2015, 12 (4), 357–360. 10.1038/nmeth.3317. [PubMed: 25751142]

(40). Li H; Handsaker B; Wysoker A; Fennell T; Ruan J; Homer N; Marth G; Abecasis G; Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map Format and SAMtools. Bioinforma. Oxf. Engl. 2009, 25 (16), 2078–2079. 10.1093/bioinformatics/btp352.

(41). Calviello L; Mukherjee N; Wyler E; Zauber H; Hirsekorn A; Selbach M; Landthaler M; Obermayer B; Ohler U Detecting Actively Translated Open Reading Frames in Ribosome Profiling Data. Nat. Methods 2016, 13 (2), 165–170. 10.1038/nmeth.3688. [PubMed: 26657557]

(42). Ellinghaus D; Kurtz S; Willhoeft U LTRharvest, an Efficient and Flexible Software for de Novo Detection of LTR Retrotransposons. BMC Bioinformatics 2008, 9 (1), 18 10.1186/1471-2105-9-18. [PubMed: 18194517]

(43). Quinlan AR; Hall IM BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. Bioinforma. Oxf. Engl. 2010, 26 (6), 841–842. 10.1093/bioinformatics/btq033.

(44). Altschul SF; Gish W; Miller W; Myers EW; Lipman DJ Basic Local Alignment Search Tool. J. Mol. Biol. 1990, 215 (3), 403–410. 10.1016/S0022-2836(05)80360-2. [PubMed: 2231712]

(45). Salmela L; Rivals E LoRDEC: Accurate and Efficient Long Read Error Correction. Bioinformatics 2014, 30 (24), 3506–3514. 10.1093/bioinformatics/btu538. [PubMed: 25165095]

(46). Retroviruses; Coffin JM, Hughes SH, Varmus HE, Eds.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor (NY), 1997.

(47). pysam: htslib interface for python — pysam 0.15.0 documentation https://pysam.readthedocs.io/en/latest/ (accessed Dec 26, 2018).

(48). Deutsch EW; Overall CM; Van Eyk JE; Baker MS; Paik Y-K; Weintraub ST; Lane L; Martens L; Vandenbrouck Y; Kusebauch U; et al. Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1. J. Proteome Res. 2016, 15 (11), 3961–3970. 10.1021/acs.jproteome.6b00392. [PubMed: 27490519]

(49). Ingolia NT; Brar GA; Stern-Ginossar N; Harris MS; Talhouarne GJS; Jackson SE; Wills MR; Weissman JS Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. Cell Rep. 2014, 8 (5), 1365–1379. https://doi.org/10.1016Zj.celrep.2014.07.045. [PubMed: 25159147]

(50). van Wijk XM; Dohrmann S; Hallström BM; Li S; Voldborg BG; Meng BX; McKee KK; van Kuppevelt TH; Yurchenco PD; Palsson BO; et al. Whole-Genome Sequencing of Invasion-Resistant Cells Identifies Laminin A2 as a Host Factor for Bacterial Invasion. mBio 2017, 8 (1). 10.1128/mBio.02128-16.

(51). Lieber MM; Benveniste RE; Livingston DM; Todaro GJ Mammalian Cells in Culture Frequently Release Type C Viruses. Science 1973, 182 (4107), 56–59. 10.1126/science.182.4107.56. [PubMed: 4125845]

(52). Anderson KP; Low MA; Lie YS; Keller GA; Dinowitz M Endogenous Origin of Defective Retroviruslike Particles from a Recombinant Chinese Hamster Ovary Cell Line. Virology 1991, 181 (1), 305–311. [PubMed: 1704658]

(53). Dinowitz M; Lie YS; Low MA; Lazar R; Fautz C; Potts B; Sernatinger J; Anderson K Recent Studies on Retrovirus-like Particles in Chinese Hamster Ovary Cells. Dev. Biol. Stand. 1992, 76, 201–207. [PubMed: 1282476]

(54). Temin HM Function of the Retrovirus Long Terminal Repeat. Cell 1982, 28 (1), 3–5. 10.1016/0092-8674(82)90367-1. [PubMed: 7066985]

(55). Weiss RA Retrovirus Classification and Cell Interactions. J. Antimicrob. Chemother. 1996, 37 Suppl B, 1–11.

(56). Lie YS; Penuel EM; Low MA; Nguyen TP; Mangahas JO; Anderson KP; Petropoulos CJ Chinese Hamster Ovary Cells Contain Transcriptionally Active Full-Length Type C Proviruses. J. Virol. 1994, 68 (12), 7840–7849. [PubMed: 7966574]

(57). Yandell M; Ence D A Beginner's Guide to Eukaryotic Genome Annotation. Nat. Rev. Genet. 2012, 13 (5), 329–342. 10.1038/nrg3174. [PubMed: 22510764]

(58). Slotkin RK The Case for Not Masking Away Repetitive DNA. Mob. DNA 2018, 9 (1), 15 10.1186/s13100-018-0120-9.

(59). Nesvizhskii AI Proteogenomics: Concepts, Applications and Computational Strategies. Nat. Methods 2014, 11 (11), 1114–1125. 10.1038/nmeth.3144. [PubMed: 25357241]

(60). Chandramouli K; Qian P-Y Proteomics: Challenges, Techniques and Possibilities to Overcome Biological Sample Complexity. Hum. Genomics Proteomics HGP 2009, 2009 https://doi.org/10.406½009/239204.

(61). Ingolia NT; Brar GA; Rouskin S; McGeachy AM; Weissman JS Genome-Wide Annotation and Quantitation of Translation by Ribosome Profiling. Curr. Protoc. Mol. Biol. 2013, Chapter 4, Unit 4–18. 10.1002/0471142727.mb0418s103.

(62). Rupp O; Becker J; Brinkrolf K; Timmermann C; Borth N; Pühler A; Noll T; Goesmann A Construction of a Public CHO Cell Line Transcript Database Using Versatile Bioinformatics Analysis Pipelines. PloS One 2014, 9 (1), e85568 10.1371/journal.pone.0085568. [PubMed: 24427317]

(63). Singh A; Kildegaard HF; Andersen MR An Online Compendium of CHO RNA-Seq Data Allows Identification of CHO Cell Line-Specific Transcriptomic Signatures. Biotechnol. J. 2018, 13 (10), e1800070 10.1002/biot.201800070. [PubMed: 29762913]

(64). Kumar A; Baycin-Hizal D; Wolozny D; Pedersen LE; Lewis NE; Heffner K; Chaerkady R; Cole RN; Shiloach J; Zhang H; et al. Elucidation of the CHO Super-Ome (CHO-SO) by Proteoinformatics. J. Proteome Res. 2015, 14 (11), 4687–4703. 10.1021/acs.jproteome.5b00588. [PubMed: 26418914]

(65). Hogwood CE; Bracewell DG; Smales CM Measurement and Control of Host Cell Proteins (HCPs) in CHO Cell Bioprocesses. Curr. Opin. Biotechnol. 2014, 30, 153–160. https://doi.org/10.1016Zj.copbio.2014.06.017. [PubMed: 25032907]

(66). de Wit C; Fautz C; Xu Y Real-Time Quantitative PCR for Retrovirus-like Particle Quantification in CHO Cell Culture. Biologicals 2000, 28 (3), 137–148. 10.1006/biol.2000.0250. [PubMed: 10964440]

(67). Yang L; Güell M; Niu D; George H; Lesha E; Grishin D; Aach J; Shrock E; Xu W; Poci J; et al. Genome-Wide Inactivation of Porcine Endogenous Retroviruses (PERVs). Science 2015, 350 (6264), 1101–1104. 10.1126/science.aad1191. [PubMed: 26456528]

(68). Richelle A; Lewis NE Improvements in Protein Production in Mammalian Cells from Targeted Metabolic Engineering. Curr. Opin. Syst. Biol. 2017, 6, 1–6. 10.1016/j.coisb.2017.05.019. [PubMed: 29104947]

(69). Lee JS; Grav LM; Lewis NE; Faustrup Kildegaard H CRISPR/Cas9-Mediated Genome Engineering of CHO Cell Factories: Application and Perspectives. Biotechnol. J. 2015, 10 (7), 979–994. 10.1002/biot.201500082. [PubMed: 26058577]

(70). Fischer S; Handrick R; Otte K The Art of CHO Cell Engineering: A Comprehensive Retrospect and Future Perspectives. Biotechnol. Adv. 2015, 33 (8), 1878–1896. 10.1016/j.biotechadv.2015.10.015. [PubMed: 26523782]

(71). Yang Z; Wang S; Halim A; Schulz MA; Frodin M; Rahman SH; Vester-Christensen MB; Behrens C; Kristensen C; Vakhrushev SY; et al. Engineered CHO Cells for Production of Diverse, Homogeneous Glycoproteins. Nat. Biotechnol. 2015, 33 (8), 842–844. 10.1038/nbt.3280. [PubMed: 26192319]

(72). Amann T; Hansen AH; Kol S; Hansen HG; Arnsdorf J; Nallapareddy S; Voldborg B; Lee GM; Andersen MR; Kildegaard HF Glyco-Engineered CHO Cell Lines Producing Alpha-1-

Antitrypsin and C1 Esterase Inhibitor with Fully Humanized N-Glycosylation Profiles. Metab. Eng. 2019, 52, 143–152. 10.1016/j.ymben.2018.11.014. [PubMed: 30513349]

(73). Spahn PN; Hansen AH; Kol S; Voldborg BG; Lewis NE Predictive Glycoengineering of Biosimilars Using a Markov Chain Glycosylation Model. Biotechnol. J. 2017, 12 (2). 10.1002/biot.201600489.

(74). Wang Q; Yin B; Chung C-Y; Betenbaugh MJ Glycoengineering of CHO Cells to Improve Product Quality. Methods Mol. Biol. Clifton NJ 2017, 1603, 25–44. 10.1007/978-1-4939-6972-2_2.

(75). Baek E; Noh SM; Lee GM Anti-Apoptosis Engineering for Improved Protein Production from CHO Cells. Methods Mol. Biol. Clifton NJ 2017, 1603, 71–85. 10.1007/978-1-4939-6972-2_5.

(76). Xiong K; Marquart KF; Karottki K. J. la C.; Li S; Shamie I; Lee JS; Gerling S; Yeo NC; Chavez A; Lee GM; et al. Reduced Apoptosis in Chinese Hamster Ovary Cells via Optimized CRISPR Interference. Biotechnol. Bioeng. 0 (ja). 10.1002/bit.26969.

(77). Chen C; Le H; Goudar CT Evaluation of Two Public Genome References for Chinese Hamster Ovary Cells in the Context of Rna-Seq Based Gene Expression Analysis. Biotechnol. Bioeng. 2017, 114 (7), 1603–1613. 10.1002/bit.26290. [PubMed: 28295162]

(78). Hefzi H From Random Mutagenesis to Systems Biology in Metabolic Engineering of Mammalian Cells. Pharm. Bioprocess. 2014, 2 (5), 355–358. 10.4155/pbp.14.36.

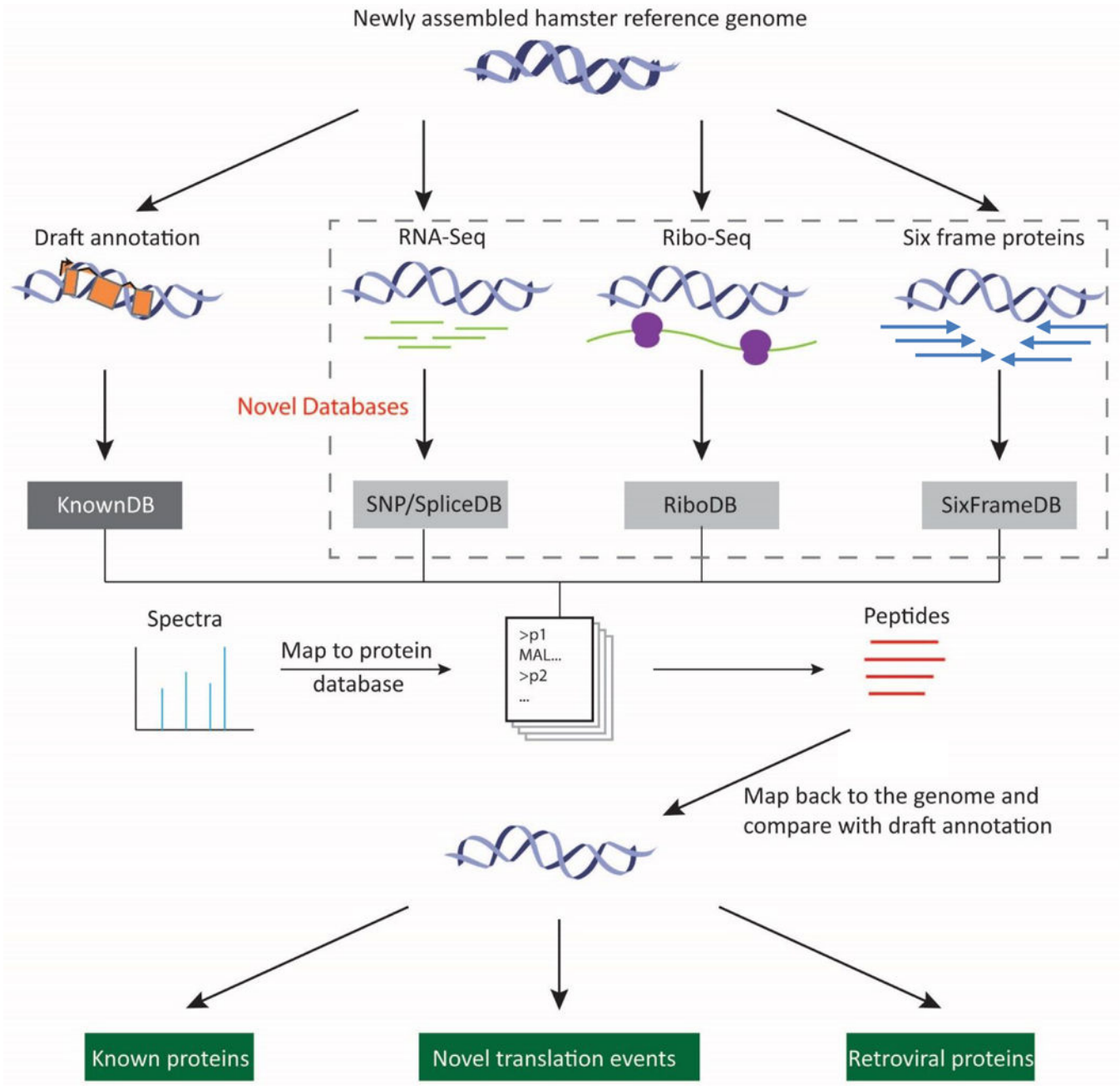


Figure 1: Overview of the proteogenomic pipeline.

Multiple databases of putative protein sequences were generated based on the newly assembled hamster genome[7] and additional data. The KnownDB contains protein sequences from our draft annotation generated here. The SNP/SpliceDB was derived from RNA-Seq samples, and contains candidate mutated or novel spliced proteins compared to the draft annotation. The RiboDB was derived from predicted translated ORFs from Ribo-Seq and RNA-Seq. The SixFrameDB is derived from the reference genome[7]. After database construction, mass spectra were mapped against the protein databases using MSGF+ to identify the peptides. The peptides were then mapped back to the genome and compared

with the draft annotation to verify translated known proteins, enumerate novel translation events, and the identity of retroviral proteins.

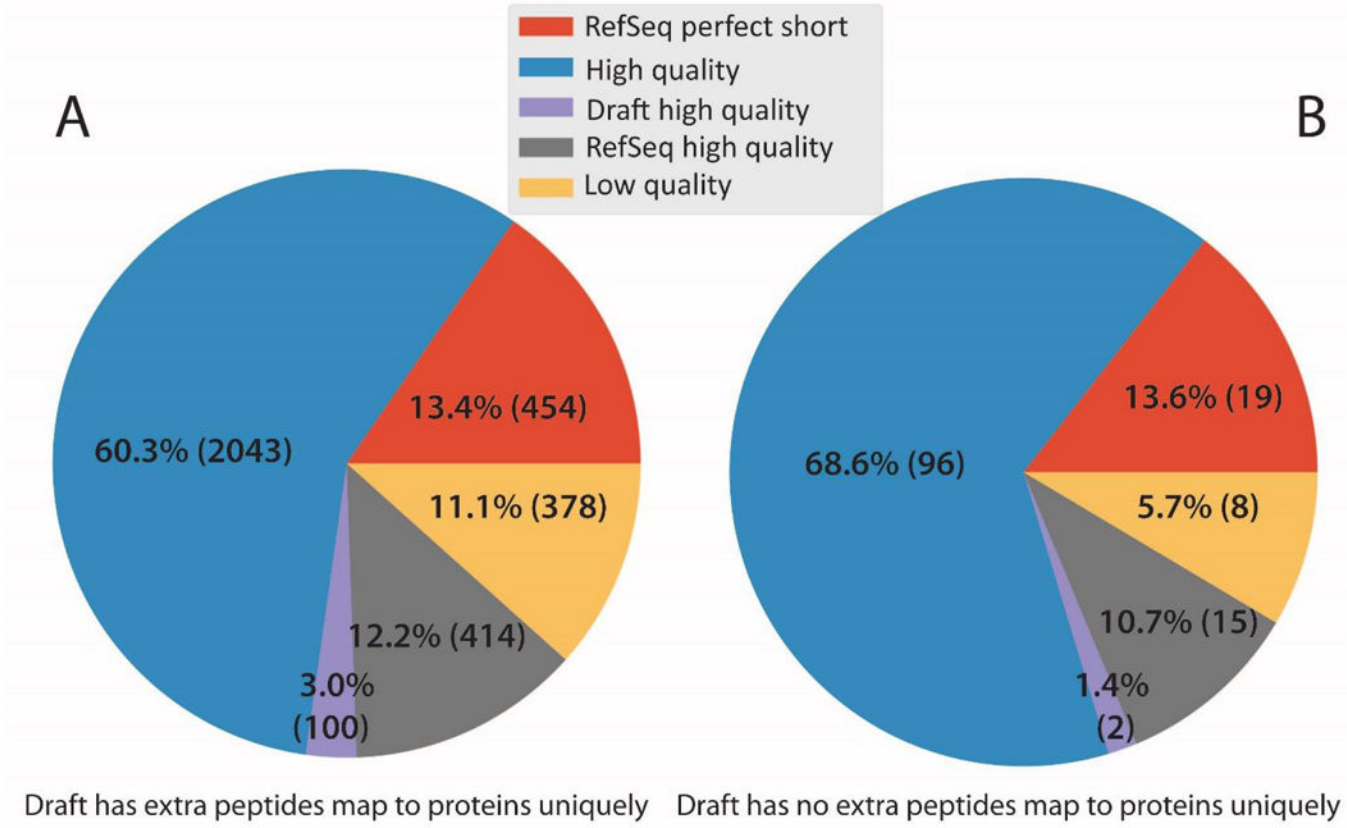
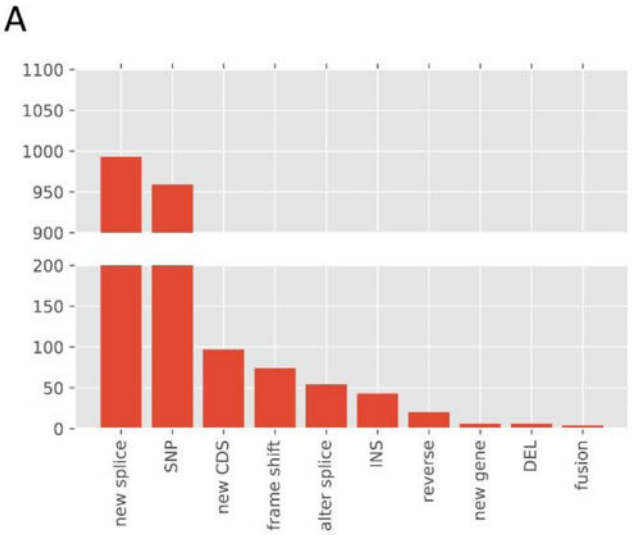


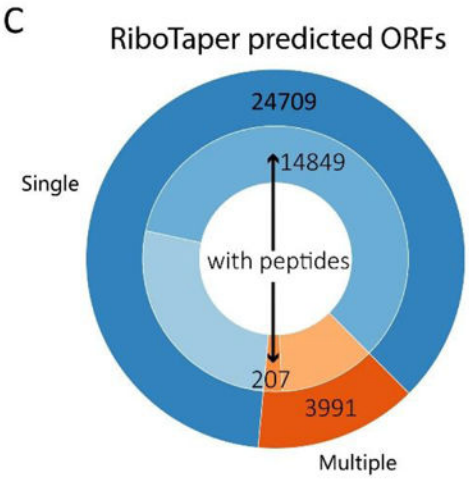
Figure 2: Number of novel draft proteins verified by draft-only peptides in different categories.

The draft annotation predicted thousands of novel protein sequences. **(A)** Of these, 3,389 had peptides mapping to proteins uniquely supporting the novel protein sequences. **(B)** Only 140 did not have extra peptide support from peptides that map to proteins uniquely, and thousands provided peptide support. **RefSeq perfect short**: RefSeq proteins map perfectly but are shorter than draft proteins; **High quality**: high quality mapping proteins between draft and RefSeq; **Draft high quality**: draft proteins map to RefSeq with high quality, but the reverse doesn't hold; **RefSeq high quality**: RefSeq proteins map to draft with high quality, but the reverse doesn't hold; **Low quality**: low quality mapping between draft and RefSeq.



B

Total spectrum	12,870,725
Identified protein isoforms	35,112
Known peptides	194,470
Novel peptides	8,003



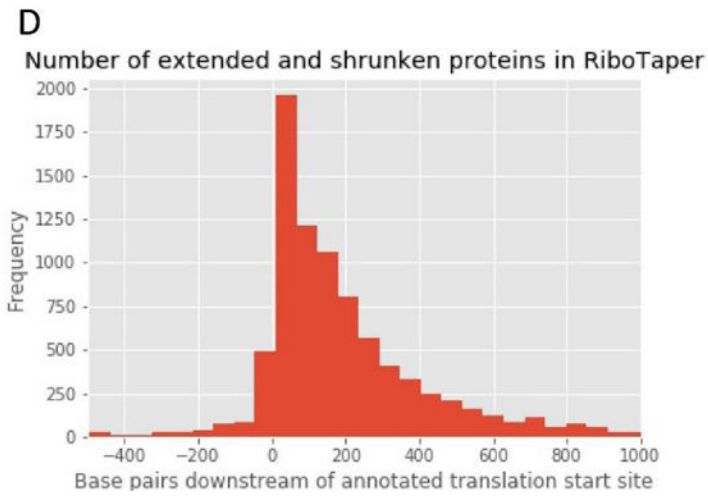


Figure 3: Proteogenomics and RiboTaper verified predicted protein sequences and identified novel translation events.

(A) Numerous novel translational events were identified, including novel splice sites that are not in the draft annotation file (new splice), non-synonymous mutations (SNP), peptides that map to UTR regions or to transcripts with no CDS (new CDS), alternative splice sites (alter splice), peptide mapping to reverse strand of reference CDS (reverse), insertions (INS), peptide mapping to intergenic regions (new gene), deletion (DEL), and gene fusions connecting two genes (fusion). **(B)** Statistics for the number of spectra, peptides and protein isoforms identified in proteogenomics. **(C)** Number of ORFs identified using RiboTaper. **Outer circle**: Number of transcripts predicted with single ORF (blue) or multiple ORFs (orange). Inner circle: Number of transcripts with (darker blue and orange) or without (light blue and orange) peptide support. **(D)** Number of proteins that are shorter/longer than the draft annotation. Positive x axis means the RiboTaper proteins are shorter (i.e., start later) than the draft annotation.

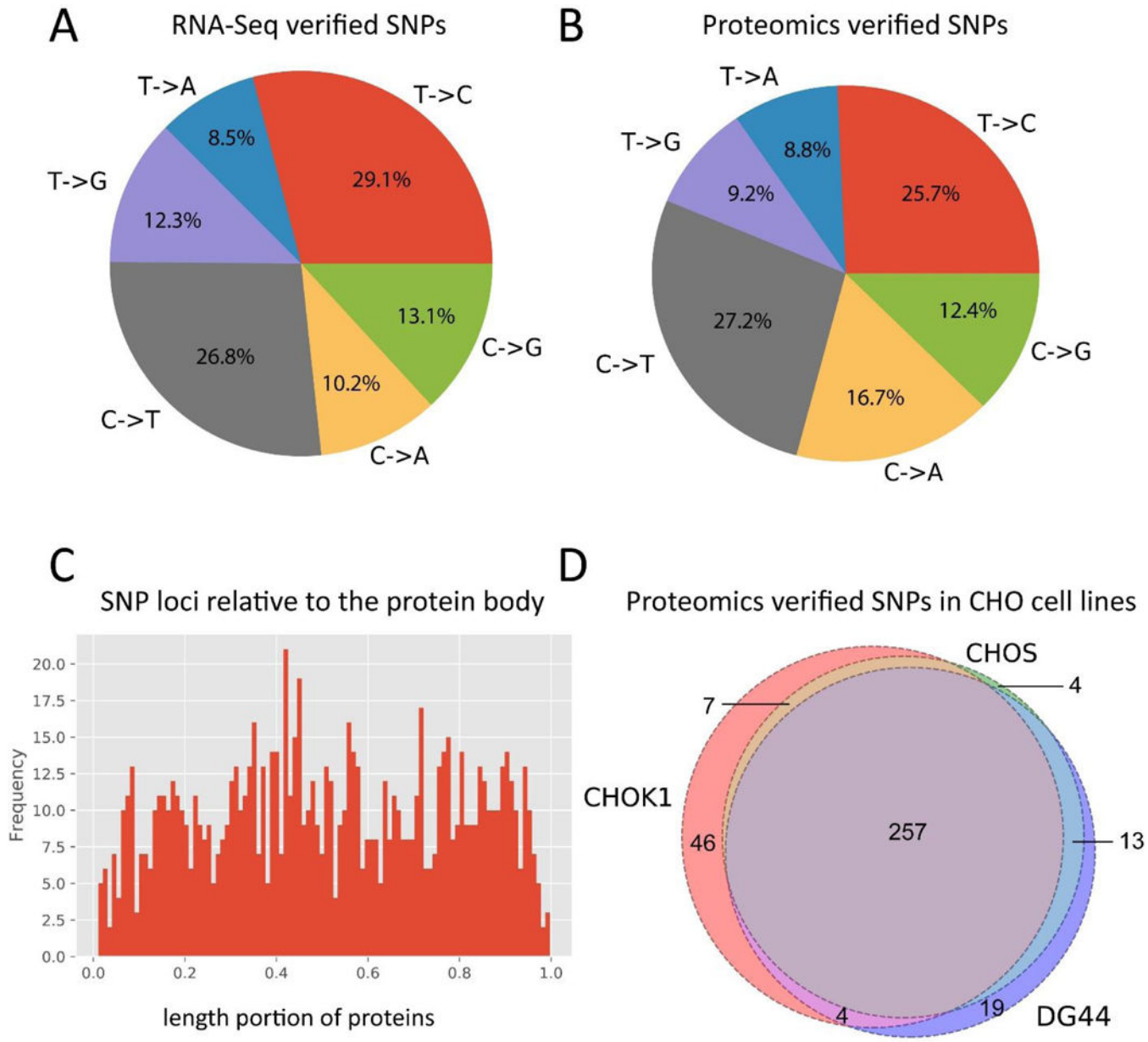


Figure 4: Hundreds of SNPs in hamster and different CHO cell lineages are validated.

A comparison of the **(A)** distribution of SNP types identified from RNA-Seq and **(B)** SNP types verified by proteomics validates the overall distribution of SNPs. **(C)** Peptide-validated non-synonymous SNPs are located throughout the protein bodies. The length of each protein is scaled to 1, and 0 represents the start codon. SNPs that locate below 0 or above 1 represent peptide-supported SNPs in 5'-UTR and 3'-UTR regions, respectively. **(D)** Venn diagram of 353 peptide-supported SNPs from CHO-K1, CHO-S and DG44 cell lines shows that most SNPs are shared across cell lines.

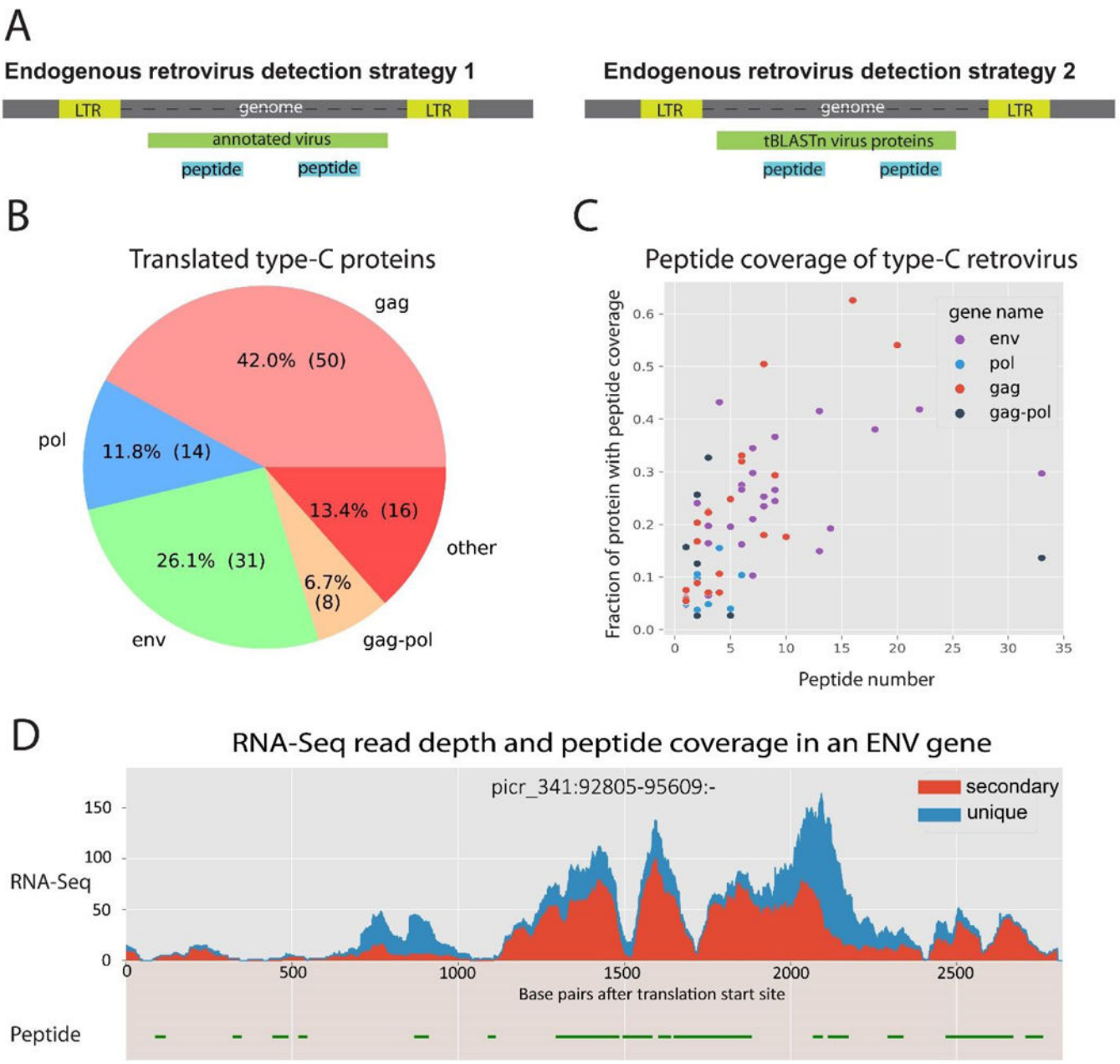


Figure 5: A proteogenomic identification of the source of translated endogenous retroviral particles shed from CHO cells.

(A) Two strategies were taken to identify translated retroviral loci. In strategy 1, peptides were mapped to the annotated retroviral proteins. For strategy 2, the sequences from the NCBI retroviral protein database were aligned to the genome using BLASTP. Then we evaluated the overlap of these aligned peptides with the novel peptides identified from the novel databases in our proteogenomics pipeline. **(B)** The strategies recovered 119 type-C peptide-supported retroviral proteins in CHO cell lines (the “other” category represents non-typical retroviral proteins, such as the p12 protein). **(C)** Peptide-supported type-C virus proteins were analyzed to assess the portion of protein sequence covered by peptides against

peptide number. **(D)** Coverage of an envelope protein in reverse strand. **uni**: reads map uniquely to the locus, **sec**: reads are secondary reads and map to multiple loci.