An Information System Enabling the Organization and Automatic Annotation of Proteomic Experiments with User-specific Proteins

Devin Lee Drew[1], Daniel Lopez-Ferrer[1], Emanuela Milani[2], Jacob Poder[3], Martin Damsbo[3], Bernd Wollscheid[2] and Andreas FR Hühmer[1]

[1]Thermo Fisher Scientific; San Jose, United States; [2]Institute of Molecular Systems Biology; ETH Zurich, Switzerland [3]Thermo Fisher Scientific; Odense, Denmark

ABSTRACT

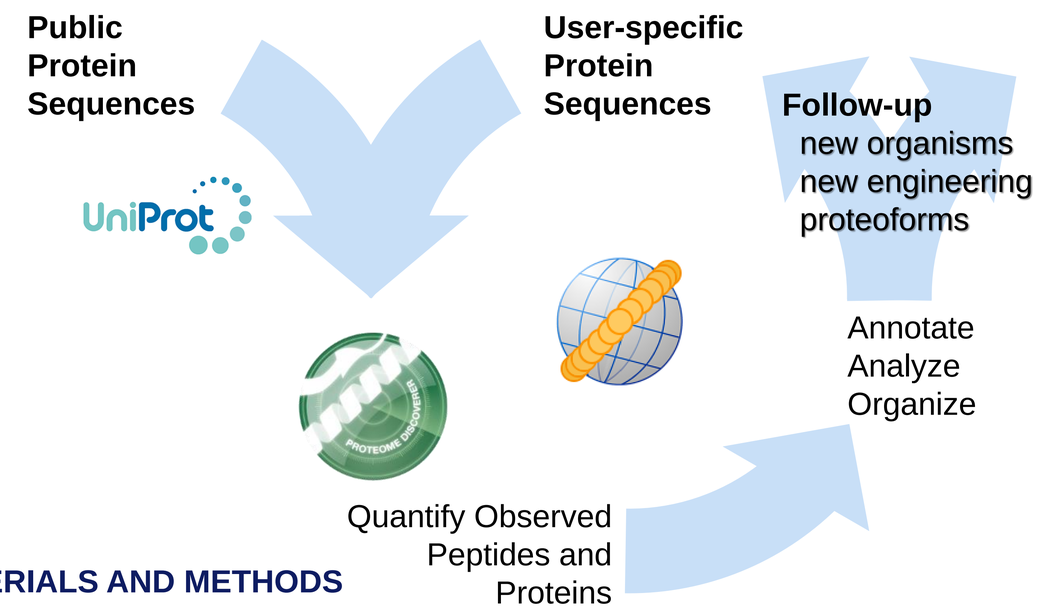
Purpose: Researchers studying engineered proteins or other proteins not defined in public sequence repositories wish to quantitatively analyze and annotate their results in the same manner as with publicly known sequences. We extend the capabilities of ProteinCenter to meet this need. Using a label-free differential analysis approach we analyze host proteins that interact with a wild type virus protein, and a mutant of the viral sequence that is postulated to disrupt host protein interactions.

Methods: Demonstrate pre-release ProteinCenter and Proteome Discoverer workflow capabilities to meet the user's engineered protein analysis needs. Apply the workflow to a functional protein interaction analysis where collaborators generated datasets for 4 culture replicates of a viral bait fusion protein and 4 replicates of a non-functional point mutant to test their hypotheses on viral protein interaction. Apply differential label free LC/MS/MS protein interactome analysis and analyze significant pathways in the functional and point mutant context.

Results: Our workflow successfully processed two non-public engineered sequence proteins to achieve label-free interactor quantities, and annotated significant pathways that are likely relevant to the function of the viral protein.

INTRODUCTION

Thermo Scientific™ ProteinCenter™ software provides an annotation framework allowing the interpretation of proteomics results in a biological context using established protein databases and annotated sequence variants, including proteoforms. However, researchers often utilize engineered proteins such as fusion proteins and introduce amino acid sequence mutations to test their hypotheses. To extend the benefits of ProteinCenter beyond currently published protein sequences we have enabled users to store their protein sequences of interest within the system thereby allowing quantitative comparison in proteomes where non-canonical protein sequences are involved. To demonstrate the new functionalities of ProteinCenter we show a differential analysis from an interactomics study using a proximity-dependent strategy where a point-mutation is expected to interfere with biologically relevant interactions.



MATERIALS AND METHODS

Sample Preparation and Data Acquisition

Cells were processed in biological quadruplicates and analyzed via a mass spectrometry based workflow using data-dependent acquisition. Stable transfection of HEK cells was performed with each of two fusion proteins encoding a Flag tag, BirA per methods described for BioID interaction studies[3], and either a wild type viral bait protein or a point mutant viral bait protein. Briefly: BirA promiscuously catalyzes biotinylation of proteins that interact with the polypeptide that BirA is fused to. Samples are enriched for biotinylated proteins and subjected to tryptic digestion and LC/MS/MS; a Thermo Scientific™ Q Exactive™ Plus mass spectrometer was used to collect DDA data. Sequences, virus, and protein names are not enumerated here due to the pre-publication nature of the dataset.

Data Analysis

Pre-release Thermo Scientific™ Proteome Discoverer™ 2.2 software was used to process and quantify identified peptide features in a label-free fashion, as well as initial two-groups analysis and absent-present search. Summed protein intensities were log2 transformed and median shifted to zero for subsequent analyses with R / InfeRno and ProteinCenter. A development release of ProteinCenter was used for profiling, and differential analysis interpretation in the human proteome background biological context.

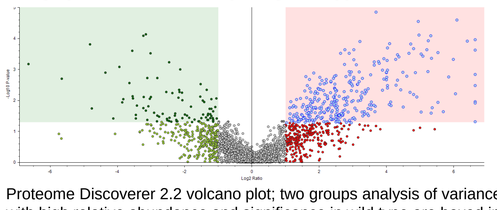
Figure 1. Engineered proteins

Table 1. Polypeptides that were disregarded by ProteinCenter software before these software changes.

Polypeptide	Length
DYKDDDDK	8
BirA	321
Translated cloning remnants	~14 per junction

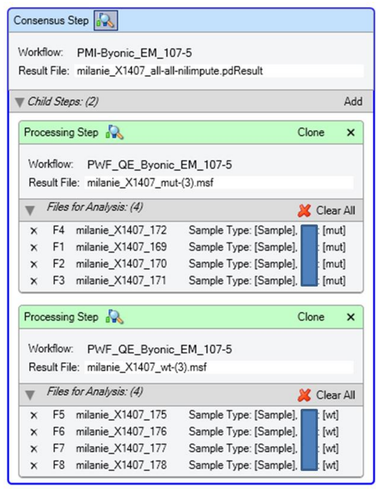
ProteinCenter previously ignored sequences that were not derived from the public repository records, such as the flag sequence, BirA fusion, and amino acid coding remnants introduced by DNA cloning sites[1,2].The pre-release changes employed here address the full engineered polypeptide sequence. Additionally, a mechanism to consolidate public and custom sequence records is underway such that the public record is linked with ProteinCenter datasets after publication and repository updates for example UniProt.

Figure 2. Preliminary volcano plots of significant high fold-change proteins when comparing wild type and mutant bait



Proteome Discoverer 2.2 volcano plot; two groups analysis of variance. Proteins with high relative abundance and significance in wild type are boxed in red. High relative abundance and significance from the mutant interactions are boxed in green. The point mutation was designed to disrupt known interaction pathways.

Figure 3. Proteome Discoverer study setup with dual Byonic processing workflows using search files specific for each engineered protein sequence.



Study Design:

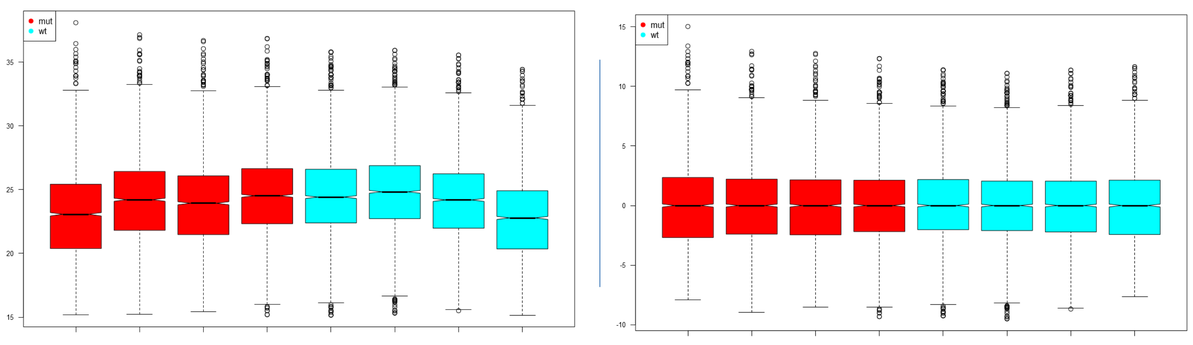
4 cell culture **replicates** of **wild type** virus bait fusion protein

4 cell culture **replicates** of **point mutant** virus bait fusion protein

Search databases:
release9606+wt-viral-fusion.fasta & release9606+mutant-viral-fusion.fasta

Static modification: Carbamidomethyl @ C
Dynamic modifications: Oxidation @ M; GG, LRGG or Biotin @ K

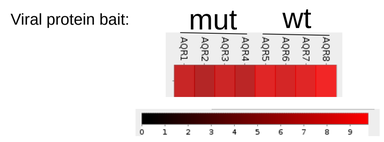
Figure 4. Data normalization



R/InfeRno Log2 transform of raw intensities (left), normalized with median central tendency and median subtraction (right). This minimally manipulative normalization retains the original data's consistent spread, visible in the consistent inter quartile ranges.

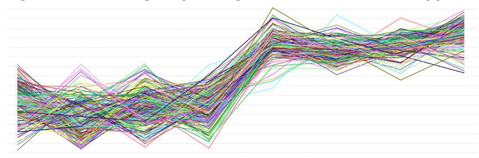
RESULTS

Figure 5. Stable transfected engineered protein quantities are consistent across samples



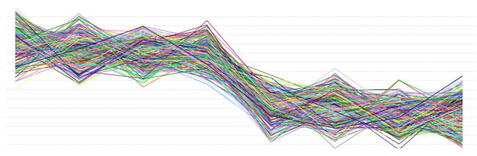
Fusion protein viral protein bait is highly expressed with slightly higher overall expression in its wild type form. These values can be used as an optional normalization step in the interactions analysis. Scale shown is log2 of raw intensities followed by median subtraction.

Figure 6. Profile group 1, high abundance in wild type



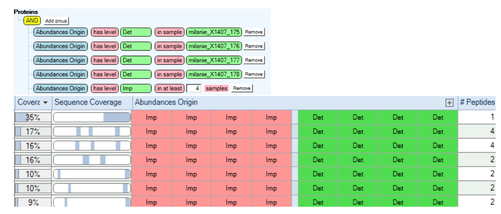
Profiles that are high abundance in wild type from proteins with 1.5 fold or greater absolute log2 fold change, and ANOVA p-value < .01

Figure 7. Profile group 2, high abundance in mutant



Profiles that are high abundance in mutant, from proteins with 1.5 fold or greater absolute log2 fold change, and ANOVA p-value < .01

Figure 8. Absence-presence analysis



Among proteins with confident identities and quantities determined in all four wild type samples, 68 proteins were uniformly undetermined among all four mutant samples. Upper panel shows filter implementing the analysis in Proteome Discoverer.

Figure 9. Venn comparison of highly abundant wild type interactors and the dataset analysis organizational layout.

Following a label-free MS1 quantitation workflow in Proteome Discoverer 2.2 (prerelease), R/InfeRno and ProteinCenter software were used to assess the quantitative differences of interacting proteins between the engineered fusion wild type and mutant baits. The sequences' biological effects were assessed using profiling, and pathway analysis in ProteinCenter. Non-public peptide & protein analysis, from user defined sequences in ProteinCenter, is made possible by the new software workflow, and is demonstrated here.

Table 2. Over-represented pathways from profile group 1 and absence presence analysis: high abundance in wild type interactors

Pathway	Count	Ref. Count	Raw p-value	FDR p-value	Source
NOD pathway (WP1433)	12	281	7.9E-13	4.6E-11	WikiPathways
Focal Adhesion (WP306)	12	2646	8.5E-03	3.5E-02	WikiPathways
Regulation of Actin Cytoskeleton (WP51)	10	1920	6.4E-03	2.8E-02	WikiPathways
Endocytosis (hsa04144)	10	324	1.9E-03	4.6E-02	KEGG Pathways
RNA transport (hsa03013)	9	190	1.5E-04	1.1E-02	KEGG Pathways
Glycogen Metabolism (WP500)	7	373	1.3E-05	2.5E-04	WikiPathways
Insulin signaling pathway (hsa04910)	7	168	1.8E-03	4.6E-02	KEGG Pathways
NLR Proteins (WP288)	6	138	4.6E-07	1.3E-05	WikiPathways
mRNA Processing (WP411)	6	1297	5.2E-02	1.5E-01	WikiPathways
Androgen receptor signaling pathway (WP138)	6	1342	6.0E-02	1.5E-01	WikiPathways
Protein modification; protein ubiquitination.	6	417	4.1E-02	4.5E-02	Uniprot Pathways
Ubiquitin mediated proteolysis (hsa04120)	6	160	6.4E-03	1.0E-01	KEGG Pathways
SREBP signalling (WP1982)	5	410	1.6E-03	1.4E-02	WikiPathways
Translation Factors (WP107)	5	469	2.9E-03	1.9E-02	WikiPathways
AMPK Signaling (WP1403)	5	539	5.3E-03	2.5E-02	WikiPathways
EBV LMP1 signaling (WP262)	4	145	2.4E-04	3.5E-03	WikiPathways
TOR Signaling (WP1471)	4	190	6.7E-04	7.8E-03	WikiPathways
IL1 and megakaryotyces in obesity (WP2865)	4	248	1.8E-03	1.4E-02	WikiPathways
Structural Pathway of Interleukin 1 (IL-1) (WP2637)	4	297	3.4E-03	1.9E-02	WikiPathways
Rac1/Pak1/p38/MMP-2 pathway (WP3303)	4	465	1.6E-02	5.4E-02	WikiPathways
IL-1 signaling pathway (WP195)	4	511	2.1E-02	7.0E-02	WikiPathways
Aminoacyl-tRNA biosynthesis (hsa00970)	4	54	2.3E-03	4.6E-02	KEGG Pathways
Glucuronidation (WP698)	3	236	1.3E-02	5.0E-02	WikiPathways
Signaling of Hepatocyte Growth Factor Receptor (WP313)	3	427	5.9E-02	1.5E-01	WikiPathways
Proteasome (hsa03050)	3	47	1.2E-02	1.6E-01	KEGG Pathways

Table 3. Over-represented pathways from profile group 2: high abundance in mutant interactors

KEGG Pathway	Count	Ref. Count	Raw p-value	FDR p-value
Ribosome (hsa03010)	34	138	4.0E-28	4.26E-26
Spliceosome (hsa03040)	18	157	2.4E-09	1.30E-07
RNA transport (hsa03013)	15	190	7.0E-06	2.34E-04
Biosynthesis of amino acids (hsa01230)	10	86	8.9E-06	2.34E-04
Carbon metabolism (hsa01200)	12	137	2.1E-05	4.34E-04
Protein processing in endoplasmic reticulum (hsa04141)	15	211	2.4E-05	4.34E-04
Citrate cycle (TCA cycle) (hsa00020)	6	35	6.5E-05	9.86E-04
Glycolysis / Gluconeogenesis (hsa00010)	8	83	2.7E-04	3.25E-03
Pyruvate metabolism (hsa00620)	6	45	2.7E-04	3.25E-03

The workflow makes possible novel data normalization approaches including scaling to fusion protein abundance across distinct primary sequences. The interpreted context of the study shows relevant pathway over-representation for the wild type samples and typical abundant and irrelevant representation for the mutant. The bias eliminating study design shows the power of point mutation and molecular cloning and the BioID approach as applied to interaction proteomics analysis.

CONCLUSIONS

We demonstrate a software workflow system for analyzing, annotating, and organizing user-defined protein sequences in their proteome-wide biological context. Application to a functional hypothesis evaluation of a viral protein's interactors gave positive results.

REFERENCES

1. DNA cloning using in vitro site-specific recombination Hartley, J. L. and Temple, G. F. and Brasch, M. A. Genome Res 2000 Nov 1788-1795 Vol 10
2. https://www.thermofisher.com/us/en/home/life-science/cloning/gateway-cloning.html
3. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. Roux, Kyle J. and Kim, Dae In and Raida, Manfred and Burke, Brian J Cell Biol 2012 Mar Number 6 p801-810 Vol 196

TRADEMARKS/LICENSING