

Comparative Bioinformatic Analysis of Proteome and Transcriptome Derived from a Single Cell Type

Martin Damsbo, Jacob Poder, Erik Nielsen, Christian Ravnsborg, Alexandre Podtelejnikov
Thermo Fisher Scientific, Odense, Denmark



Overview

Purpose: Demonstration of fast and effective way to perform statistical, comparative and bioinformatic analysis of large scale proteome and transcriptome studies

Methods: Set of bioinformatic data-mining tools

Results: Comprehensive bioinformatic analysis of data derived from large-scale LC-MS/MS proteomics study of HeLa cells and comparison to transcriptome data

Introduction

With the current development of mass spectrometry-based proteomics techniques, it is becoming possible to perform in-depth analyses of the entire proteome. With protein identifications of more than ten thousand proteins, it is essential to be able to perform sophisticated bioinformatic analyses of such datasets and compare to gene expression data. Here we present such a comparison for HeLa cell analysis and a new development in web-based Thermo Scientific ProteinCenter software—an integration with Cytoscape®—that enables visualization of molecular interaction networks.

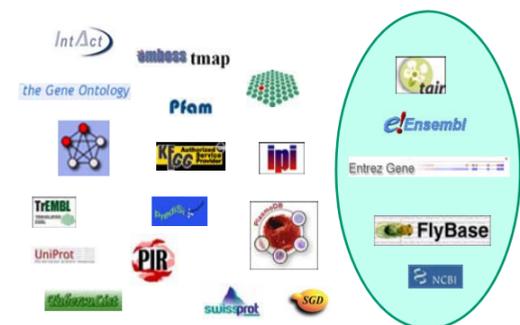
Methods

ProteinCenter™ software is a protein-centric tool that integrates the contents of a large number of public protein sequence databases (Figure 1), experimental protein identifications and a number of bioinformatics tools. It contains a non-redundant database with more than 10 million proteins from the major public protein databases - distilled from 100 million accession codes from past and present versions of public databases.

To test ProteinCenter software, we used data from a recent publication "Deep proteome and transcriptome mapping of a human cancer cell line" [1]. Data was logged to ProteinCenter software and included accession codes, ion intensities, iBAQ [2] and FPKM values, absolute quantification in fmol, and number of copies per cell. Genomic information was restricted only for proteins detected in the study. Statistical analysis of datasets was performed based on Benjamini-Hochberg correction of p-values with false discovery rate (FDR) values of 5%. Several features like profiling, gene ontology and KEGG pathways were applied to assist with the visualization and bioinformatic analysis of data.

The Cytoscape (<http://www.cytoscape.org>) version was 2.8.2. Several plug-ins like GeneMANIA (<http://www.genemania.org>) and Reactome (www.reactome.org) were used to assist protein interaction network visualisation and analysis. Interaction data was imported into Cytoscape [3] as a network SIF file. Quantitative values were imported as 'node attributes' as a separate .csv file.

FIGURE 1. Consolidated database. Genomic databases supported by ProteinCenter software are highlighted by the circle.



Results

Statistical Analysis of HeLa Proteome

Very often proteomics and transcriptomics experiments are performed in different laboratories at different times and using different experiment set-ups. As a result, the degree of correlation between the protein and mRNA levels can be difficult to interpret. While ProteinCenter software removes the obstacles of database entry dynamics, the values observed during these experiments can introduce a level of uncertainty. In the best case scenario, both experiments are performed at one place in the same experimental set-up. That is why we selected the very well controlled study of HeLa cells. A total of 10,255 proteins with 99% confidence were identified; the average identification of the fragmentation spectra was 43%. The average sequence coverage was close to 39%. Currently it is the most exhaustive proteome analysis of human single cell type and we consider to use it as a reference dataset.

Statistical analysis of transcriptome and proteome based on gene ontology, signal peptides, transmembrane regions did not reveal significant differences and was in close proximity to statistical analysis of reference databases, IPI and Swiss-Prot® (Figure 2). This can be explained by high proteome coverage; high mass accuracy mass spectrometry based analysis with a FASP purification method covers nearly the entire functional proteome including membrane proteins. That is why we focused on analysis of pathways in the study.

In total, 242 pathways were detected out of 251 available from KEGG for all human proteins. Among under-represented pathways were asthma, autoimmune thyroid disease, allograft rejection, olfactory transduction and others. All of them are not expected to be functionally relevant in HeLa cells. At the same time, some of the over-represented pathways (Figure 3) like nucleotide excision repair (Figure 4) have more than 92% coverage. All in all, average pathway coverage is around 46% and suggests that the detected proteome covers a very large part of functional pathways.

FIGURE 2. General statistics

General Info	Proteins msb201181-s2	Homo sapiens (SP)
Proteins	10235	35852
Exp. Data	10237	0
Peptides	0	0
Unique peptide sequences	0	0

Taxonomy	Proteins msb201181-s2
Taxonomies	1

Annotation	Proteins msb201181-s2
Genes	9201
Chromosomes	26
Enzymes	841
Keywords	617
Diseases	2561

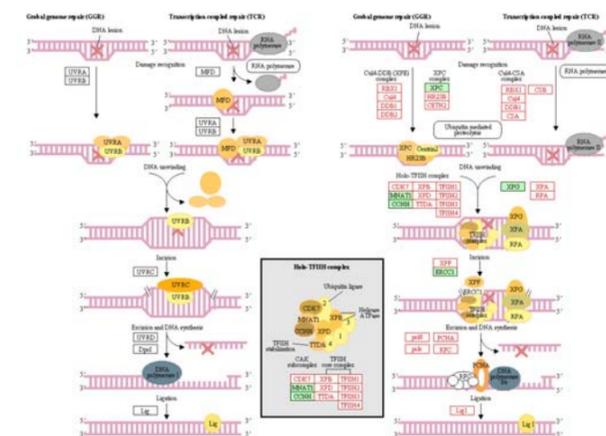
Domains	Proteins msb201181-s2
Pfam domain proteins	9045
InterPro domain proteins	8501
Transmembrane proteins	4459
Signal peptide proteins	1377

Pathways	Proteins msb201181-s2
KEGG pathways	242
UniProt pathways	215

FIGURE 3. Over-representation analysis of KEGG pathways using Swiss-Prot as a reference database

Description	Occurrence	Count	Ref. Count	Raw p-value	FDR p-value
Spliceosome (hsa03040)	26	114	3.993E-21	5.709E-19	
Ribosome (hsa03010)	69	85	6.884E-14	3.936E-12	
RNA transport (hsa03013)	34	127	8.298E-14	3.936E-12	
Metabolic pathways (hsa01100)	310	991	4.385E-12	1.639E-10	
Nucleotide excision repair (hsa03420)	24	37	1.782E-10	5.096E-9	
Cell cycle (hsa04110)	25	105	4.852E-10	1.157E-8	
Aminocycl-RNA biosynthesis (hsa09270)	32	36	4.744E-9	9.651E-8	
Ribosome biogenesis in eukaryotes (hsa03008)	49	64	1.27E-8	2.271E-7	
mRNA surveillance pathway (hsa03015)	51	68	2.155E-8	3.424E-7	
RNA degradation (hsa03018)	46	60	3.33E-8	4.762E-7	
Ubiquitin mediated proteolysis (hsa04120)	80	121	3.865E-8	5.024E-7	

FIGURE 4. Visualisation of nucleotide excision repair KEGG pathway. (Proteins highlighted in red were detected in the proteome analysis.)



Comparative Bioinformatic Analysis of HeLa Proteome and Transcriptome

Historically, mRNA and protein levels correlate poorly partially because of the inability to sequence significant number of proteins. Until recently, the proteomics dataset from a single LC/MS experiment was in the range of several hundreds of proteins identified. With the improvement of proteomics techniques, like sub-ppm mass accuracy and attomole sensitivity of modern hybrid mass spectrometers, it is becoming possible to achieve identification levels up to ten thousand proteins as it was shown in the selected study of HeLa cells.

A single gene can give rise to multiple distinct proteins due to alternative splicing, sequence polymorphisms and post-translational modifications. As shown in Figure 5, the numbers of protein counts in a set of particular KEGG pathways were higher than the numbers of genes. This can be explained by the different protein isoforms detected in the study.

FIGURE 5. Proteome versus transcriptome

Description	Occurrence	Count	Ref. Count
Systemic lupus erythematosus (hsa05322)	26	14	
Uric acid junction (hsa04530)	52	46	
Viral myocarditis (hsa05416)	18	13	
Long-term potentiation (hsa04720)	28	22	
Melanogenesis (hsa04916)	31	25	
Transcriptional misregulation in cancer (hsa05202)	39	33	
Insulin signaling pathway (hsa04910)	61	54	
Thyroid cancer (hsa05216)	13	10	
Long-term depression (hsa04730)	25	21	
Gap junction (hsa04540)	41	36	

Recent developments in labelling technology have provided useful techniques for the quantitative study of protein expression. The quantitative values FPKM (fragments per kilobase of exon per million fragments mapped) and absolute protein quantification by intensity-based absolute quantification (iBAQ) approach were used for comparison of HeLa proteome and transcriptome. iBAQ protein intensities were translated to protein copy number estimates based on the overall protein amount in the analyzed sample. While most publications found that the overall correlation between quantified mRNA and protein abundances is modest or poor [4], the current study showed relatively good correlation. Partially this can be explained as the datasets were generated from the same set of cells under the same experimental conditions.

Cellular protein abundances span 8 orders of magnitude from approximately 0.2 copy per cell for Centriolin up to 3.3E+7 for Cofilin-1. Highly abundant proteins (top 100) are mainly of cytoplasmic origin and involved in cell development, response to chemical stimulus and cell death. On the level of molecular function they are involved in ATP binding, structural molecule activity, catalytic activity.

Bioinformatic Analysis of Low Abundant Proteins

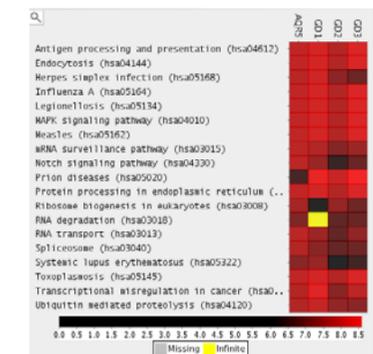
We also extracted the list of proteins with the lowest iBAQ values to create a small dataset of low-abundance proteins. These proteins were expressed on the level up to 50 copies per cell in the range of 0.2 to 40 attomoles. Among these proteins 54 have at least one transmembrane domain, 20 contain signal peptide signature and more than 60% of their sequences contain alternative spliced variants. To characterize that pool of proteins we performed the Gene Ontology analysis on three levels: Molecular Function (GO MF), Cellular Components (GO CC), and Biological Processes (GO BP). GO MF showed serine-type exopeptidase activity, GTPase binding, and transmembrane receptor protein serine/threonine kinase activity. Among GO CC were gap junction, Golgi-associated vesicle and ruffle. GO BP analysis showed that these proteins are heavily involved in regulation of signal transduction (Figure 6).

FIGURE 6. Gene Ontology Biological Processes analysis of low abundant proteins

Description	Occurrence	Count	Ref. Count
regulation of signal transduction	22	3461	
cellular response to transforming growth factor beta stimulus	3	53	
response to transforming growth factor beta stimulus	3	62	
cellular component organization	39	8577	
regulation of response to stimulus	25	4704	
negative regulation of response to extracellular stimulus	2	23	
negative regulation of response to nutrient levels	2	23	
cellular component organization or biogenesis	39	8758	

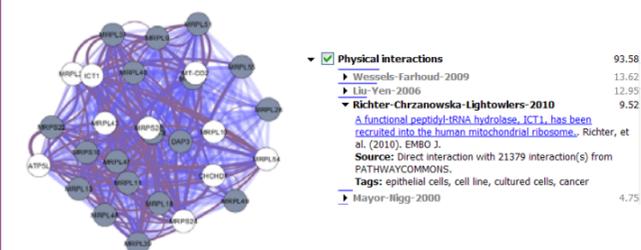
A heat map can assist the visualization of the quantitative data. Figure 7 shows a comparison of FPKM (AQR5) and iBAQ values for trypsin, LysC and GluC digestions (GD1, GD2 and GD3 correspondingly) for a fraction of a 60S large ribosomal subunit that were summarized on the level of pathway analysis.

FIGURE 7. Heat map generated from quantitative values of proteome and transcriptome comparison



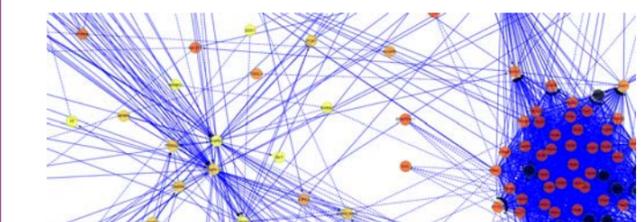
While the knowledge of over-represented GO terms or pathways is helpful, it does not always explain the molecular mechanism relevant to the explored proteins and investigation of protein-protein interaction (PPI) networks can be useful. To complete the bioinformatic analysis of the data set we performed PPI analysis using Cytoscape graph algorithms and the VizMapper visualization tool, where ProteinCenter software data were used to colour-code nodes according to the protein quantitative values. Integration of ProteinCenter software with Cytoscape also allows access to a significant number of plug-ins and provides exhaustive analysis of interactome. For example, Figure 8 shows the interactome of mitochondrial ribosomal proteins on the level of physical interactions, prediction and co-localizations. All of these interactions are well curated and have the links to the corresponding publications. Nodes highlighted in grey were detected in the study.

FIGURE 8. Mitochondrial ribosomal interactome analysis by GeneMANIA



In-depth proteome coverage opens an opportunity to link pathway profiles to disease phenotypes. An excellent tool for such an analysis is Reactome FI plug-in for Cytoscape. Figure 9 shows the fragment of network where the genes highlighted by yellow reflect the association with a cervical carcinoma.

FIGURE 9. Visualization of cancer related genes by Reactome FI



Conclusion

- We demonstrated the comparison of proteome and transcriptome data followed by gene ontology, over-represented analysis, pathways analysis, PPI networks analysis and overall statistics
- The state-of-art mass spectrometric technique overcomes the major limitation of peptide-centric proteomics, which is biased towards more-abundant peptides. The data show that the detected proteome is closed to almost the entire functional transcriptome and covers a very large part of functional pathways.
- Due to the high proteome coverage and high quality of data, the current study can be used as a reference proteome data set similar to HUPO brain, liver and plasma studies.

References

- Nagaraj, N.; Wisniewski, J.R.; Geiger, T.; Cox, J.; Kircher, M.; Kelso, J.; Pääbo, S.; Mann, M. Deep proteome and transcriptome mapping of a human cancer cell line. *Mol Syst Biol.* **2011**, *7*, 337–342.
- Schwanhauser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M. Global quantification of mammalian gene expression control. *Nature* **2011**, *473*, 337–342.
- Cline, M.S., et al. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc.* **2007**, *2*(10), 2366–82.
- de Sousa Abreu, R.; Penalva, L.O.; Marcotte, E.M.; Vogel, C. Global signatures of protein and mRNA expression levels. *Mol Biosyst.* **2009**, *5*(12):1512–26.

Cytoscape is a registered trademark of The Cytoscape Consortium. Swiss-Prot is a registered trademark of Institut Suisse de Bioinformatique (SIB) Foundation Switzerland. All other trademarks are the property of Thermo Fisher Scientific and its subsidiaries. This information is not intended to encourage use of these products in any manner that might infringe the intellectual property rights of others.