Revised manuscript accepted for publication in Journal of Mass Spectrometry, 2019, doi: 10.1002/jms.4464.

Widespread presence of bovine proteins in human cell lines

Simon Sugár[1,2], Lilla Turiák[1], Károly Vékey[1], László Drahos[1]

[1]MS Proteomics Research Group, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Magyar tudósok körútja 2, H-1117 Budapest, Hungary

[2]Semmelweis University, Károly Rácz School of PhD Studies, Üllői út 26, H-1085 Budapest, Hungary

Abstract

HPLC-MS/MS analysis of various human cell lines show the presence of a major amount of bovine protein contaminants. These likely originate from fetal bovine serum (FBS), typically used in cell cultures. If evaluated against a human protein database, on average 10% of the identified human proteins will be misleading (bovine proteins, but indicated as if they were human). Bovine contaminants therefore may cause major bias in proteomic studies of cell cultures, if not considered explicitly.

Introduction

Identification and quantitation of proteins in human cell lines are typical objectives of proteomics and are mostly performed by HPLC-MS[1]. In a recent study on the glycosylation of HeLa cellular proteins[2] we have observed that the commercial sample contained a significant amount of bovine serum proteins. It is a reasonable assumption that these impurities come from fetal bovine serum (FBS), a universally used component of cell culture media[3, 4]. This observation prompted us to examine whether this is a common problem or if it was related to a particular sample only.

We have selected 11 frequently studied human cell lines based on a study published by Geiger et al[5]: Cervical cancer (HeLa), hepatoma (HEP2G), lung carcinoma (A549), glioblastoma (GAMG), embryonic kidney cells (HEK293), chronic myeloid leukemia (K562), acute T-cell leukemia (Jurkat), breast cancer (MCF7), colon carcinoma (RKO), prostate carcinoma (LnCap) and osteosarcoma (U2OS). We have analyzed the commercially available HeLa cell line lysate using high-resolution HPLC-MS/MS proteomics. We have also downloaded and analyzed HPLC-MS/MS results of the various cell lines listed above from the Pride database (Table 1), corresponding to research performed by Geiger et al. In this paper we have performed a comparative analysis of these results, with the objective of identifying bovine contaminants in human cell lines and estimating its influence on human proteome studies.

Experimental

Analysis of the HeLa cell lysate:

The HeLa cell lysate analyzed was purchased from Thermo Fischer Scientific (Waltham, MA). The solvents used were all HPLC-MS grade and purchased from Merck (Darmstadt, Germany). Samples were analyzed using a Maxis II QTOF instrument (Bruker Daltonik GmbH, Bremen, Germany) equipped with CaptiveSpray nanoBooster ion source coupled to a Dionex UltiMate 3000 RSLCnano system (Sunnyvale, CA, USA). Peptides were separated on an Acquity M-Class BEH130 C18 analytical column (1.7 µm, 75 µm × 250 mm Waters, Milford, MA) using gradient elution (4–50% linear gradient of eluent B in 90 min) following trapping on an Acclaim PepMap100 C18 (5 µm, 100 µm × 20 mm, Thermo Fisher Scientific, Waltham, MA) trap column. Solvent A consisted of water + 0.1% formic acid, while Solvent B was acetonitrile + 0.1% formic acid. The cycle time was set at 2.5 sec in the 700-2000 m/z range, preferred charges states were set between +2 and +5. MS spectra were acquired at 3 Hz, while MS/MS spectra at 4 or 16 Hz depending on the intensity of the precursor. Following each run raw data were recalibrated using the Compass DataAnalysis software 4.3 (Bruker Daltonics, Bremen, Germany).

Analysis of the other samples:

Results on various other cell cultures were taken from the manuscript published by Geiger et al. For details see Table 1. Experiments were based on LC-MS/MS, analogous to those described above. Experimental data were downloaded from the Pride database[6] (https://www.ebi.ac.uk/pride/archive/).

Data evaluation

Qualitative analysis of peptides and proteins studied by LC-MS/MS as described above were evaluated by Byonic[7] (version 2.15.7, Protein Metrics Inc., Cupertino, CA, USA), Andromeda[8] and Mascot[9] (version 2.5.1, www.matrixscience.com) softwares. In the case of Byonic, program settings and modifications for each sample were determined by Byonic's Preview function for optimal results. The Andromeda search was performed by the MaxQuant software (version 1.5.8.3). Each dataset was searched against the SwissProt Human and SwissProt Bovine database[10] (downloaded from https://www.uniprot.org/) in two separate runs. The commercial HeLa sample was searched against a combined human/bovine, and an all species database as well. In general, only proteins with at least two peptide hits were considered.

Results and Discussion

Commercial HeLa cell digest

The commercial HeLa cell lysate, analyzed by HPLC-MS/MS, was evaluated against human, bovine, a combined human/bovine, and an all species database. Initial studies were performed using Byonic, Mascot and Andromeda search engines. The results were similar,

although Byonic gave the most peptide and proteins hits, therefore in the discussion below we refer to results obtained by Byonic. However, all conclusions drawn in the present paper were confirmed by Mascot and Andromeda searches as well.

Using the all species database most proteins identified were either human or bovine, but there were a few hits corresponding to other animals as well. The latter was most closely related to humans (like orangutan or gorilla) or bovines (like camel) and were based on few peptide hits only. Detailed analysis showed that the peptides supposedly identifying e.g. an orangutan protein by the search engine were identical to those of the corresponding human protein (i.e. due to sequence homologies). As orangutan proteins are unlikely contaminants in a HeLa cell lysate we safely concluded that it was a human protein, misidentified by the search engine. The same conclusion was obtained using Mascot search as well. Subsequently, the sample was treated to contain human and bovine proteins only.

The same HPLC-MS/MS results were evaluated using human, and in a separate search using bovine database. The two searches found over 1800 supposedly human, and over 600 supposedly bovine proteins, respectively. In most cases, the same proteins (in their human or bovine versions) were found in the two searches, due to sequence homologies between human and bovine proteins. We have compared (using an Excel macro) peptide sequences found in the two searches and annotated which sequences are unique in humans, in bovines, and which are common peptides. Among the peptides found we have identified over 8000 human-specific and over 1000 bovine-specific peptides, and approximately 5000 further peptides, which are identical in humans and in bovines (accurate numbers are listed in Table 2). These results clearly show that the HeLa sample contains a large number of bovine impurities. Based on the number of species-specific peptides, approximately 15% of the proteins in the commercial HeLa sample are of bovine origin.

We have analyzed the results at the protein level as well. We have determined the number of human-specific and “common” peptides (i.e. those, which were found both in the human and in the bovine searches) identifying a certain protein. Subsequently, we have identified the bovine protein corresponding to these “common” peptides. Using the bovine search results, we have determined the number of bovine-specific peptides found in the sample. This way we have determined the number of human-specific, bovine-specific and common peptides corresponding to each protein found in the sample.

Based on these comparisons, the sample components can be categorized as unequivocally or predominantly human, or unequivocally or predominantly bovine proteins. We term “unequivocally human” those proteins, which are identified based on only human-specific peptides or both human-specific and “common” peptides, but the bovine search did not find any bovine-specific peptides for the corresponding bovine protein. Proteins were considered predominantly human when there were more human-specific than bovine-specific peptide hits. There were a few proteins where the number of human and bovine-specific peptide hits were identical. As a human cell line was studied, these were also considered as “predominantly human”. Furthermore, there were a few protein identifications where beside “common” peptides neither human- nor bovine-specific peptides were found. These might be either human or bovine, but as HeLa is a human sample, these were also considered predominantly human proteins. Similarly, the list of bovine proteins consists of bovine proteins misidentified as human and proteins only found in the bovine search.

The relative amount of bovine and human proteins analogs is estimated by the number of unique bovine and human peptides, respectively. One such example is Gelsolin (GELS_BOVIN / GELS_HUMAN), in which case 13 common, 11 bovine specific and 5 human specific peptides were found. These are listed in Supplementary Table S2. As an example, MS/MS spectra of a pair of human and bovine peptide homologs are shown in Supplementary Figs S1 and S2, showing that these can be identified with high certainty.

Based on these considerations, approximately 2000 proteins were found in the HeLa sample (see Table 2. for precise numbers). Among these only, approximately 1200 are unequivocally human, while approximately 200 are unequivocally bovine proteins. A little over 700 further proteins were found, which had both human and bovine components. Altogether there were 1605 predominantly human and 331 predominantly bovine proteins in this HeLa sample. We have checked manually, that all abundant bovine proteins were serum proteins. This clearly suggests that the bovine contaminants derive from FBS, typically used as culture media.

Some of the bovine contaminants were identified based on bovine-specific peptides only. These will be "hidden" contaminants (not identified using a human database search). In most studies, these will not have adverse effect on the proteomic studies of human cell lines. However, most bovine proteins will be found using a human protein database, and these will be misidentified as human proteins. In the case of the HeLa sample studied, approximately 1900 human proteins were found using the human database (Table 2). However, among these only 1605 are in reality human proteins, while there are 280 predominantly bovine proteins, which were misidentified as human. This is a major error, which may cause significant bias in proteomics studies.

We have analyzed the same results using a combined human/bovine FASTA database as well. In this respect Byonic and Mascot search engines behaved differently. We have found that identifying predominantly human and predominantly bovine proteins was unreliable using Byonic. Mascot gave the same results as discussed above for identifying whether a certain protein is predominantly of human or of bovine origin. However, Mascot results were unfeasible to evaluate automatically, so we have used our macro-based analysis (described above) for analyzing the other samples.

We have also checked the analysis of an analogous sample, which did not get into contact with bovine proteins. However, in all cases we found in PRIDE either used bovine supplements, or experimental information was lacking in the original paper to decide, whether bovine supplements were used in the cell culture. So, we have selected a human tissue sample[11] and studied it in an analogous way to that described above. Even though this clearly did not contain bovine impurities, proteomic analysis using Byonic search did find a few bovine-specific peptides, 1.3% of all peptides identified. This is clearly an error of the database search. It is slightly higher than the false discovery rate (1%) but far smaller than the proportion of bovine-specific peptides in the commercial HeLa sample. While this indicates limitation of most proteomic analyses, the high proportion of bovine peptides found in HeLa (15% of specific peptide hits) shows unequivocally that the sample is contaminated by a massive amount of bovine proteins.

Presence of bovine impurities in other cell lines

We have established above that the commercial HeLa cell lysate studied by us contained a large amount of bovine impurities. Next, we have downloaded (from the Pride database)

results of a previous LC-MS/MS study on a HeLa cell lysate studied by a leading research group. We have evaluated these results in a manner analogous to that discussed above using human and bovine database searches; analyzing the human and bovine specific peptide fragments. In this case, approximately 3000 human proteins and 500 bovine proteins were found in the sample (Table 2). Using only a human database, 13% of the proteins identified as human proteins were, in fact, bovine proteins.

Next, we have downloaded results from the PRIDE database on 10 further, commonly used cell lines. The comparative proteomics analysis of these cell lines has been previously reported[5]. Data evaluation analogous to that above showed, that these cell lines were also contaminated by a large amount of bovine proteins (Table 3). On average, 10 % of the human proteins found using the human database were, in fact, bovine proteins, misidentified as human proteins. The bovine protein contaminants in the various samples were similar, but not identical to each other. The correlation coefficient of the number of peptide hits of various bovine proteins between two cell lines was, on average, 0.78. For the complete list of bovine and human proteins (listed as total H and total B in Table 3) see the Supplementary Material.

Bovine proteins most commonly found in the various cell lines discussed above are listed in Table 4. These were selected as proteins that have been found with a large number of bovine-specific peptides and were present in most cell lines studied. In our opinion, these proteins are well suited to be markers for identifying the presence of bovine contaminants in human samples originating from cell cultures.

Conclusions

Bovine serum components were found to be common, abundant contaminants of human cell lines. These are the remains of fetal bovine serum (FBS), a generally used cell culture media. These represent over 10% of the protein content of most cell culture samples and therefore should be taken into account. Their presence may or may not cause bias or compromise research results – it depends on the type of analysis, type of data evaluation, and type of information sought. In any case, major effort must be taken, that such contaminants should not compromise research results.

In the case of ‘straightforward’ proteomics, proteins are identified based on a human proteomic or genetic database. In such type of data evaluation, ca. 10% of the protein content of cell cultures will be misidentified: they will appear as human proteins, when in fact they are of bovine origin. Both their presence and (in quantitative studies) their amount will be misleading. This type of error or bias can be avoided if the experimental results are evaluated against a bovine database as well, identifying the predominantly bovine protein contaminants.

In the case of studies on post-translational modifications, the situation is more complicated. A good example is protein glycosylation, which may be analyzed as glycopeptides or (more often) as released glycans. When glycopeptides are analysed[12, 13], it can be ascertained if the relevant peptide sequence is human- or bovine specific. When the amino acid sequence of the glycopeptide is common in human and in bovine, it is still possible to ascertain, if the corresponding protein is predominantly of human or bovine origin. This approach may be

used to categorize the glycopeptide as predominantly human or bovine. Glycosylation profiles analyzed this way will be unaffected by the presence of bovine impurities. A different, simpler and more commonly used experimental strategy is released glycan analysis[14, 15]. In this case information on the origin of the released glycan is lost, and the results will represent some average between the glycosylation profile of human cellular and bovine serum glycosylation. This will be an inherently erroneous result. Furthermore, serum proteins are heavily glycosylated, while only few cellular proteins are glycosylated. This will amplify the errors, which will be even larger, than the amount of bovine contaminants may suggest. In the case of other PTMs analogous considerations need to be used to ascertain the validity of the results.

Acknowledgments

L.T. and K.V. are grateful for funding from the National Research Development and Innovation Office (NKFIH PD-121187, and NKFIH K-119459). L.T. acknowledges support from the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

Table 1: Short description of the cell lines analyzed

Sample identifier	Short description	Cell culture sample preparation	PRIDE project number
HeLa S	Cervical cancer cells, sample measured by the authors	Dulbecco's modified Eagle's medium, 10% FBS and antibiotics	PXD013930
HeLa M	Cervical cancer cells, sample measured by Geiger et al.	Dulbecco's modified Eagle's medium, 10% FBS and antibiotics	PXD002395
HEP2G	Hepatoma cells, sample measured by Geiger et al.	Dulbecco's modified Eagle's medium, 10% FBS and antibiotics	PXD002395
Jurkat	Acute T-cell Leukemia cells, sample measured by Geiger et al.	RPMI, 10% FBS and antibiotics	PXD002395
LnCap	Prostate carcinoma cells, sample measured by Geiger et al.	RPMI, 10% FBS and antibiotics	PXD002395
MCF7	Mammary carcinoma cells, sample measured by Geiger et al.	Dulbecco's modified Eagle's medium, 10% FBS and antibiotics	PXD002395
K562	Chronic myeloid leukemia cells, sample measured by Geiger et al.	Dulbecco's modified Eagle's medium, 10% FBS and antibiotics	PXD002395
A549	Lung carcinoma cells, sample measured by Geiger et al.	Dulbecco's modified Eagle's medium, 10% FBS and antibiotics	PXD002395
RKO	Colon carcinoma cells, sample measured by Geiger et al.	Dulbecco's modified Eagle's medium, 10% FBS and antibiotics	PXD002395
U2OS	Osteosarcoma cells, sample measured by Geiger et al.	Dulbecco's modified Eagle's medium, 10% FBS and antibiotics	PXD002395
GAMG	Glioblastoma cells, sample measured by Geiger et al.	Dulbecco's modified Eagle's medium, 10% FBS and antibiotics	PXD002395
HEK293	Embryonic kidney cells, sample measured by Geiger et al.	Dulbecco's modified Eagle's medium, 10% FBS and antibiotics	PXD002395

Table 2: Human and bovine peptides and proteins in HeLa

Sample	Peptide hits			Protein hits						
	Human-specific	Bovine specific	Common	Human-specific	Pred. Human	Bovine specific	Pred. Bovine	Total Human	Total Bovine	Misid. Human
HeLa S	8552	1396	5152	1218	387	192	139	1605	311	280
HeLa M	16421	2800	5557	2510	711	354	193	3221	547	583

Table 3: Presence of human and bovine proteins in various human cell lines

Sample	Total number of Human proteins	Total number of Bovine proteins	Number of misidentified Human proteins
HeLa S	1605	331	280
HeLa M	3221	659	583
HEP2G	2242	232	216
Jurkat	1889	231	203
LnCap	2654	216	196
MCF7	2618	349	330
K562	1968	240	215
A549	2436	226	200
RKO	1906	242	218
U2OS	2509	247	228
GAMG	2610	266	236
HEK293	2742	395	353

Table 4: Most common bovine proteins in various cell lines

Bovine contaminants	Identified peptides		
	Bovine specific	Human-specific	Common
A2MG	13	2	2
ALBU	21	3	3
FETA	8	0	0
A1AT	7	0	0
TRFE	8	2	1
FETUA	5	0	1
THRB	4	2	1
APOA1	4	0	0

References:

[1] H.J. Issaq. The role of separation science in proteomics research. *ELECTROPHORESIS.* **2001**, *22*, 3629.

[2] L. Turiák, S. Sugár, A. Ács, G. Tóth, Á. Gömöry, A. Telekes, K. Vékey, L. Drahos. Site-specific N-glycosylation of HeLa cell glycoproteins. *Scientific Reports.* **2019**, *9*, 1.

[3] W.J. Bettger , W.L. McKeehan. Mechanisms of cellular nutrition. *Physiological Reviews.* **1986**, *66*, 1.

[4] H. Eagle. Nutrition Needs of Mammalian Cells in Tissue Culture. *Science.* **1955**, *122*, 501.

[5] T. Geiger, A. Wehner, C. Schaab, J. Cox, M. Mann. Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins. *Molecular & Cellular Proteomics.* **2012**, *11*, M111.014050.

[6] J.A. Vizcaíno, A. Csordas, N. del-Toro, J.A. Dianes, J. Griss, I. Lavidas, G. Mayer, Y. Perez-Riverol, F. Reisinger, T. Ternent, Q.-W. Xu, R. Wang, H. Hermjakob. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Research.* **2015**, *44*, D447.

[7] M. Bern, Y.J. Kil, C. Becker. Byonic: Advanced Peptide and Protein Identification Software. *Current Protocols in Bioinformatics.* **2012**, *40*, 13.20.1.

[8] J. Cox, N. Neuhauser, A. Michalski, R.A. Scheltema, J.V. Olsen, M. Mann. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *Journal of Proteome Research.* **2011**, *10*, 1794.

[9] D.N. Perkins, D.J.C. Pappin, D.M. Creasy, J.S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *ELECTROPHORESIS.* **1999**, *20*, 3551.

[10] C. The UniProt. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research.* **2018**, *47*, D506.

[11] L. Turiák, O. Ozohanics, G. Tóth, A. Ács, Á. Révész, K. Vékey, A. Telekes, L. Drahos. High sensitivity proteomics of prostate cancer tissue microarrays to discriminate between healthy and cancerous tissue. *Journal of proteomics.* **2019**, *197*, 82.

[12] M. Wuhrer, M.I. Catalina, A.M. Deelder, C.H. Hokke. Glycoproteomics based on tandem mass spectrometry of glycopeptides. *Journal of Chromatography B.* **2007**, *849*, 115.

[13] L. Cao, Y. Qu, Z. Zhang, Z. Wang, I. Prytkova, S. Wu. Intact glycopeptide characterization using mass spectrometry. *Expert Review of Proteomics.* **2016**, *13*, 513.

[14] A. Varki, R.D. Cummings, J.D. Esko, H.H. Freeze, P. Stanley, C.R. Bertozzi, G.W. Hart, M.E. Etzler, *Essentials of Glycobiology*, Cold Spring Harbor Laboratory Press, **2009**.

[15] S.M. Haslam, S.J. North, A. Dell. Mass spectrometric analysis of N- and O-glycosylation of tissues and cells. *Current Opinion in Structural Biology.* **2006**, *16*, 584.