Naive Pluripotent and Trophoblastic Stem Cell Lines as a Model for Detecting Missing Proteins in the Context of the Chromosome-Centric Human Proteome Project

Océane Girard, Régis Lavigne, Simon Chevolleau, Constance Onfray, Emmanuelle Com, Pierre-Olivier Schmit, Manuel Chapelle, Thomas Fréour, Lydie Lane, Laurent David, et al.

Naive Pluripotent and Trophoblastic Stem Cell Lines as a Model for Detecting Missing Proteins in the Context of the Chromosome-Centric Human Proteome Project

Océane Girard [1], Régis Lavigne [3,4], Simon Chevolleau [1], Constance Onfray [1], Emmanuelle Com [3,4], Pierre-Olivier Schmit [5], Manuel Chapelle [5], Thomas Fréour [1,6,7], Lydie Lane [8], Laurent David [1,2], and Charles Pineau [3,4,*]

[1] Nantes Université, CHU Nantes, Inserm, CR2TI, UMR 1064, F-44000 Nantes, France

[2] Nantes Université, CHU Nantes, Inserm, CNRS, BioCore, F-44000 Nantes, France

[3] Univ Rennes, Inserm, EHESP, Irset (Institut de Recherche en Santé, Environnement et Travail) - UMR_S 1085, F-35000 Rennes, France

[4] Univ Rennes, CNRS, Inserm, Biosit UAR 3480 US_S 018, Protim Core Facility, F-35000 Rennes, France

[5] Bruker Daltonique SA, 34 rue de l'Industrie, F-67166 Wissembourg cedex, France

[6] CHU Nantes, Service de Biologie de la Reproduction, F-44000 Nantes, France

[7] Department of Obstetrics, Gynecology and Reproductive Medicine, Dexeus University Hospital, Barcelona, Spain

[8] CALIPHO Group, SIB Swiss Institute of Bioinformatics and University of Geneva, Geneva, Switzerland

Corresponding author: charles.pineau@inserm.fr

Keywords: human proteome project, missing proteins, human naive pluripotent stem cells, human trophoblastic stem cells

Project Name: Proteomics of human naive pluripotent and trophoblastic stem cells

Project accession: PXD035768

Project DOI: 10.6019/PXD035768

ABSTRACT

The Chromosome-centric Human Proteome Project (C-HPP) aims at identifying the proteins as gene products encoded by the human genome, characterizing their isoforms and functions. The existence of products has now been confirmed for 93.2% of the genes at the protein level. The remaining mostly correspond to proteins of low abundance or difficult to access. Over the past years, we have significantly contributed to the identification of missing proteins in the human spermatozoa. We pursue our search in the reproductive sphere with a focus on early human embryonic development. Pluripotent cells, developing into the fetus, and trophoblast cells, giving rise to the placenta, emerge during the first weeks. This emergence is a focus of scientists working in the field of reproduction, placentation and regenerative medicine. Most knowledge has been harnessed by transcriptomic analysis. Interestingly, some genes are uniquely expressed in those cells, giving the opportunity to uncover new proteins that might play a crucial role in setting up the molecular events underlying early embryonic development. Here, we analyzed naive pluripotent and trophoblastic stem cells and discovered 4 new missing proteins, thus contributing to the C-HPP. The mass spectrometry proteomics data was deposited on ProteomeXchange under the dataset identifier PXD035768.

INTRODUCTION

The Human Proteome Project (HPP) is the flagship initiative of the global Human Proteome Organization (HUPO). Its main objective is to catalog proteins as gene products encoded by the human genome and credibly identify these essentially, but not entirely, by mass spectrometry[1]. In 2020, the HPP celebrated its 10th anniversary with a major achievement, *i.e.,* a high stringency blueprint of the Human proteome detailing the detection of over 90% of all predicted human proteins[2]. It has since increased this effort to 93.2% in 2022. In this context, the number of experimentally validated proteins (PE1) is annually updated by neXtProt, the reference protein knowledge-base for the HPP (www.nextprot.org) and has now reached 18,407 whereas the actual count of missing proteins (MPs) scored as PE2 or 3 or 4, now stands at 1,343.

There are many reasons, often combined, why missing proteins still lack evidence, either because they are expressed at low-copy numbers, because of their biology (*e.g.,* time- or stress-dependent, restricted to specific cell types or during pathophysiological situations) or because of detection limits by mass spectrometry due to some peculiar physicochemical properties (*e.g.,* hydrophobicity, basicity). In such regard, Lane and collaborators suggested that the proteins that have been systematically missed might be restricted to a few unusual organs or cell types, particularly the testis[3]. Unsurprisingly, the very high number of testis-specific genes supported the hypothesis that the testis was a promising organ in which to search for missing proteins. In a series of works, we succeeded in identifying in spermatozoa over 250 missing proteins whose expression is restricted and/or specific to the post-meiotic germ cell lineage[4–7]. Our mining of publicly available transcriptomic expression datasets indicates that both the male and female reproductive sphere organs are sources for detecting a significant number of missing proteins[7]. However, it can be anticipated that most of these will be particularly difficult to access due to their physicochemical properties (membrane proteins, defensins, etc.) or to narrow windows of expression.

In the present work, we pursue our search for missing proteins in the reproductive sphere with a focus on the early human embryonic development.

During pre-implantation development, the early human embryo successively undergoes two main morphological events, *i.e.,* compaction and cavitation. Alongside these morphological events and embryo growth, two cell fates specification are required at the morula and blastocyst stages. The first cell fate decision in the morula segregates the outside trophectoderm (TE) cells from inner cell mass (ICM). Subsequently, the second cell fate decision at the early blastocyst stage processes ICM cells to form the pluripotent epiblast (EPI) and primitive endoderm (PrE), precursors of the embryo and yolk sac, respectively[8] (**Figure 1**). Proteomic analysis of human embryos has been scarce up to now, not only owing to the technical challenge, but also to ethical and legal issues, resulting in a limited access to human embryos for research.

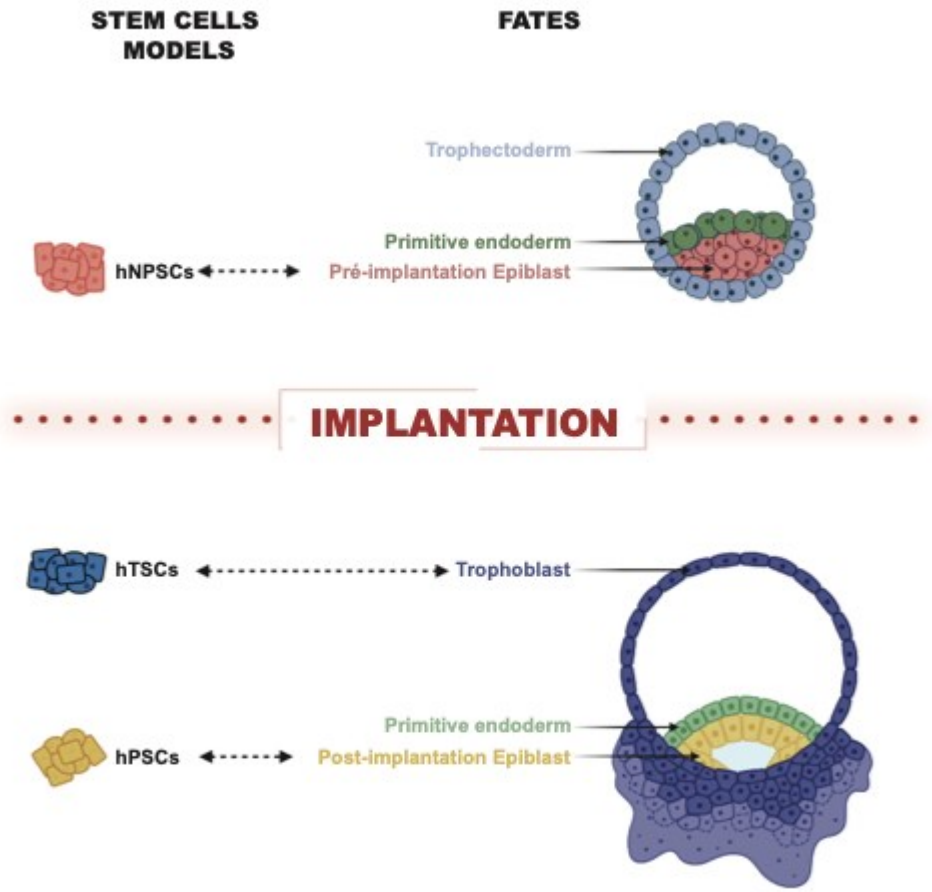

Figure 1: Schematic representation showing the correlation between embryo cell fate and stem cells. On the right, the three cell fates in the embryo before and after implantation with the trophectoderm (blue), the epiblast (pink and yellow) and the primitive endoderm (green). On the left, the stem cell models corresponding to the cell

fate with the hNPSCs (pink) as equivalent to the pre-implantation epiblast, the hTSCs (blue) as equivalent to the trophectoderm and the hPSCs (yellow) as equivalent to the post-implantation epiblast (*Figure created with BioRender.com*).

We thus hypothesized that missing proteins might be accessible in human embryos. To overcome the above-mentioned challenges, stem cell models are commonly used as a proxy for human embryos. In order to study human peri-implantation, 3 main stem cell types are used. The first is naive pluripotent stem cells (hNPSCs) with the characteristics of the preimplantation epiblast[9]. The second is trophoblastic stem cells (hTSCs) with the characteristics of post-implantation cytotrophoblasts[10]. Those stem cells models are often compared to the primed pluripotent stem cells (hPSCs), not studied here, which represent post-implantation epiblast, a later stage of development than the hNPSCs[9] (**Figure 1**).

Here we describe a strategy combining the analysis of RNASeq datasets and MS-based experiments to identify and validate missing proteins in hNPSCs and hTSCs. “We validated 4 missing proteins with at least 2 non nested unique peptides, according to the HPP guidelines[11], whereas single peptides were detected for 6 additional proteins. The use of the PaSER 2022 database search platform (Bruker Daltonik GmbH, Bremen, Germany) was also shown to be relevant and crucial for validating several unique peptides out of the MS proteomics data.

MATERIALS AND METHODS

Cell culture

All cell lines were cultured on a feeder cell layer, under hypoxic conditions (5%O^2, 5%CO^2). Culture medium was replaced daily. 10µM Y27632 (Axon Medchem, Groningen, The Netherlands) was added to the culture medium upon thawing and passaging of the cells.

hTSCs were cultured in hTSCs medium [DMEM/F12 (Gibco, ThermoFisher, Les Ulis, France) supplemented with 0.1mM 2-mercaptoethanol (Gibco), 0.2% FBS, 0.5% penicillin-streptomycin, 0.3% Bovine Serum Albumin (BSA, Sigma-Aldrich, St Louis, United States), 1% Insulin-Transferrin-Selenium-Ethanolamine supplement (ITS-X, Gibco), 1.5 mg/ml L-ascorbic acid (Sigma-Aldrich), 50 ng/ml hEGF (Miltenyi Biotec, Bergisch Gladbach, Germany), 2 µM CHIR99021 (Axon Medchem), 0.5 µM A83-01 (Tocris, Bristol, United Kingdom), 1 µM SB431542 (Tocris), 0.8 mM valproic acid (Sigma-Aldrich) and 5 µM Y27632]. hiTSCs (human induced naive pluripotent stem cells) were passaged with TrypLE (5 min, 37°C, Life Technologies Corporation, Carlsbad, United States) every 4 to 5 days at a cell density between 1.04*104 and 2.08 *104 cells per cm^2.

hNPSCs were cultured in PXGLY medium [47.5% Neurobasal medium (GIBCO) and 47.5% DMEM/F12 (Gibco) supplemented with 1mM N2 (GIBCO), 2mM B27 (Gibco), 1mM glutamax (Gibco), 1mM non-essential amino acids (GIBCO), 0.45% BSA, 1mM sodium pyruvate (Gibco), 0.1% 2-mercaptoethanol (Gibco), 0.5% penicillin-streptomycin, 1µM PD0325901 (Axon Medchem), 2µM XAV939 (Axon medchem), 2µM Gö6983 (Axon medchem), 10ng/mL hLif (PeproTech, Neuilly-sur-Seine, France), 10µM Y27632 (Axon Medchem)]. hNPSCs were passaged every 4 to 5 days with TrypLE (5 min, 37°C, Life Technologies) at a cell density of 2.08 *10^4 cells per cm^2.

Prior to the experiment, cells were submitted to two dissociation and lysis protocols. In the first one, cells were dissociated using TrypLE (5min, 37°C), feeders were removed by incubation on 0.1% gelatin-coated plates layer for 1 hour and cells were lysed with the iST Kit (PreOmics GmbH, Planegg, Germany) following the manufacturer's instructions. In the second one, cells were plated overnight after feeder removal on respectively 0.1% geltrex coated plate for hNPSCs or on 3µg/mL

vitronectin and 1µg/mL laminin coated plate for hTSCs. 24 hour after seeding, cells were processed with the iST Kit.

Protein Extraction, Digestion, and Liquid Chromatography—Tandem Mass Spectrometry (LC-MS/MS) Analyses

Briefly, samples were thawed and lysed (denatured, reduced and alkylated) for 10 min at 95°C then Trypsin/LysC digested for 3hours at 37°C. Purification of peptides was then carried out at room temperature on spin cartridge and peptides were finally eluted in 10 µL of LC-load buffer. Simultaneously, a protein assay has been realized to precisely know the quantity of proteins present in the samples. Once purified, hTSCs and hNPSCs samples were peptide assayed to prepare for mass spectrometry injection at an amount of 10µg in 10µL.

Approximately 200ng each of tryptic peptides samples were separated onto a 75µm x 250mm IonOpticks Aurora 2 column (Ion Opticks Pty Ltd, Australia) packed with a 120 Å pore, 1.6µm particle size C18 beads. A reversed-phase gradient of basic buffers (Buffer A: 0.1% formic acid, 98% H2O MilliQ, 2% acetonitrile ; Buffer B: 0.1% formic acid, 100% acetonitrile) was run on a NanoElute HPLC System (Bruker Daltonik) at a flow rate of 400 nl/min at 50°C. The liquid chromatography (LC) run lasted for 120 min with a starting concentration of 2% of buffer B increasing to 15% over the initial 60 min, a further increase in concentration to 25% over 30 min, then to 37% in ten minutes, and finally to 95% in ten minutes again. This elution gradient was followed by a 95% wash during ten minutes and re-equilibration.

Temperature of the separation column was regulated thanks to the Sonation column oven. The constant column temperature makes measurement significantly more accurate and higher column temperatures could also be applied allowing us to apply the flow of 400 nl/min while maintaining the same pressure. The NanoElute HPLC system was coupled online to a Tims TOF Pro mass spectrometer (Bruker Daltonik) with a CaptiveSpray ion source (Bruker Daltonik). The CaptiveSpray nanoflow ESI source was directly attached to a vacuum inlet capillary via a short capillary extension heated using the instrument's drying gas. High voltage for the electrospray (ESI) process was applied to the vacuum capillary inlet, whereas the sprayer was kept at ground. Temperature of the ion transfer capillary was set at 180°C. The spray type was automatically mechanically aligned on the axis with the capillary inlet without the need for any adjustment. Ions were accumulated for 114 ms, and mobility separation

was achieved by ramping the entrance potential from -160 V to -20 V within 114 ms. The acquisition of the MS mass spectra with the TIMS TOF Pro is done with an average resolution of 60 000 FWHM (mass range 100-1700 m/z). To enable the PASEF method, precursor m/z and mobility information was first derived from full scan TIMS-MS experiments (with a mass range of m/z 100-1700). Resulting quadrupole mass, collision energy and switching times were automatically transferred to the instrument controller as a function of the total cycle time. The quadrupole isolation width was set to 2 and 3 Th and, for fragmentation, the collision energies varied between 31 and 52 eV depending on precursor mass and charge. TIMS, MS operation and PASEF were controlled and synchronized using the control instrument software OtofControl 6.2 (Bruker Daltonik). LC-MS/MS data were acquired using the PASEF method with a total cycle time of 1.28 s, including 1 TIMS MS scan and 10 PASEF MS/MA scans. The 10 PASEF scans (110ms each) contain on average 12 MS/MS scans per PASEF scan. In addition, the most abundant precursors which could have been sequenced in previous scan cycles are dynamically excluded from re-sequencing. The acquisition of the MS/MS mass spectra with the TIMS TOF Pro is also done with an average resolution of 50 000 FWHM (mass range 100-1700 m/z).

Protein identification

Ion mobility resolved mass spectra, nested ion mobility vs m/Z distributions, as well as summed fragment ion intensities were extracted from the raw data file with DataAnalysis 6.0 (Bruker Daltonik). Signal-to-noise (S/N) ratio were increased by summations of individual TIMS scans. Mobility peak positions and peak half-widths were determined based on extracted ion mobilograms (±0.05 Da) using the peak detection algorithm implemented in the DataAnalysis software. Features detection were also performed using DataAnalysis 6.0 software and exported in .mgf format.

Peptides identification were performed with the Mascot search engine (version 2.6.2, Matrix Sciences), applying the previously described search parameters and using its automatic decoy database search to calculate the false discovery rate (FDR)[4,7]. The search database was the complete human proteome homo sapiens UP000005640 from UniProtKB 2022_02 restricted to one protein sequence per gene. Briefly, 1 miscleavage for trypsin was allowed and mass tolerance of peptides and fragments was established at 15 ppm and 0,05 Da. Moreover, mass modifications of peptides

are taken into account. For fixed modifications, carbamidomethylation of cysteines and for variable modifications, oxidations and acetylation of lysines and N-term proteins were considered. After interrogations on Mascot, data processing was performed using the Proline software (version 2.1.0). All the results of the queries performed on Mascot were imported into Proline with a subset threshold of 1. After importation, the results were validated with a peptide pretty rank of 1, an FDR for PSM of 1% on adjusted e-value and an FDR for protein set of 1% with a standard scoring.
Using the Mascot search engine, the total number of expected false positives for hNPSCs classically dissociated from cultures dishes with trypsin, was 0.06% for PSM, 0.09% for peptide set and 0.51% for protein set. For hNPSCs directly dissociated from culture plates with the lysis buffer, the total number of expected false positives was 0.06% for PSM, 0.09% for peptide set and 0.52% for protein set. The merge of the two hNPSCs conditions provided an expected false positive rate of 0.12% for PSM and peptide set and 0.77% for protein set. Similarly, the total number of expected false positives for hTSCs classically dissociated from cultures dishes with trypsin, was 0.06% for PSM, 0.09% for peptide set and 0.53% for protein set. For hTSCs directly dissociated from culture plates with the lysis buffer, the total number of expected false positives was 0.06% for PSM, 0.09% for peptide set, 0.52% for protein set. The merge of the two hTSCs conditions provided an expected false positive rate of 0.13% for PSM and peptide set and 0.85% for protein set.

On top of the standard Mascot search, all datasets were searched using the Graphics Processing Units (GPU)-based PaSER 2022 V3 solution (Bruker). PaSER was configured to use the ProLucid search engine[12] with the complete human proteome homo sapiens UP000005640 from UniProtKB 2022_02 restricted to one protein sequence per gene and with modification definitions as the ones used with Mascot. With PaSER, the acetylation of lysines and N-term proteins was parameterized at the peptide level and not at the protein level like Mascot. The mass tolerance of peptides for PaSER was fixed at 30 ppm. Both protein and peptide FDR thresholds were set to 1%. TIMScore was enabled to allow the use of the peptide Collisional Cross Section (CCS) during the scoring process.
When using the PaSER search engine, the total number of expected false positives for hNPSCs classically dissociated from cultures dishes with trypsin, was 0.22% for PSM, 0.25% for peptide set, 0.99% for protein set, whereas for hNPSCs directly

dissociated from culture plates with the lysis buffer, the total number of expected false positives was 0.30% for PSM, 0.36% for peptide set, 0.98% for protein set. For hTSCs classically dissociated from cultures dishes with trypsin, the total number of expected false positives was 0.31% for PSM, 0.36% for peptide set, 0.99% for protein set, whereas for hTSCs directly dissociated from culture plates with the lysis buffer, the total number of expected false positives was 0.25% for PSM, 0.33% for peptide set, 0.98% for protein set.

An essential component to TIMScore is defining the deviation between experimental and predicted CCS values. Machine learning was used in order to accurately predict CCS values from a peptide's primary amino acid sequence. A training dataset of hun-dreds of thousands of tryptic and phosphorylated peptides was used, where the dataset included peptides of doubly, triply and quadruply charge states. A transformer model of peptide CCS was developed from this training set. The model was tested for accu-racy against an independent dataset it had previously not seen. For doubly, triply and quadruply charged peptides the accuracy in predicting a peptide CCS from the primary amino acid sequence was 95% for tryptic pep-tides. Upon setting up the parameters file, *in silico* peptide candidates are sent to the CCS prediction model to generate a pre-dicted CCS value. The PaSER search algorithm (in our case, ProLucid) is run as normal and the search algorithm compares the predicted and measured CCS values and calculates a correlation score, namely TIMScore for the top 5 peptide candidates for each spectra.

Peptide and protein identifications summaries were generated for each sample with both the Mascot and the PaSER search engines. Datasets were further analyzed in accordance with version 3.0 the HPP data interpretation guidelines[11]. Finally peptide-to-protein mappings were checked using the neXtProt uniqueness checker[13].

3'SRP datasets

3'seq-RNA Profiling (3'SRP)[14] datasets were obtained from Kilens et al (2018) and Castel et al (2020). Concerning the plots made from 3'SRP datasets, the values lower than 0.1 mRNA molecules per million of mRNA molecules have been arbitrarily represented as 0.1 mRNA molecules per million of mRNA molecules for convenience.

Data Availability

The mass spectrometry proteomics data, including raw files and identification files, form a complete submission with the ProteomeXchange Consortium[15]. Data were submitted via the PRIDE partner repository under the dataset identifiers PXD035768 and 10.6019/PXD035768".

RESULTS & DISCUSSION

The aim of the present work was to pursue our ongoing project of possibly detecting numerous missing proteins in human tissues and cells of the reproductive sphere. Here, we assessed whether hNPSCs and hTSCs were relevant biological material for searching for missing proteins by analyzing expression in these cells of 1343 mRNAs corresponding to referenced missing proteins in the latest neXtProt release (2022.02.25).

hNPSCs and hTSCs are recently discovered cell types[16,17]. They are also representative of a unique developmental time, with specific features: hNPSCs are modeling the pre-implantation epiblast, that need to remain pluripotent and proliferate to form the source of the fetus, while hTSCs represent cytotrophoblast, the “placental stem cells”. We surmised that those unique and transient developmental stages might use specific gene sets, potentially restricted to this developmental stage. As a consequence, we started by analyzing 3’SRP datasets[10,18] corresponding to the two stem cell lines. 3’SRP uses unique molecular identifiers (UMI) to correct errors in quantification of mRNA, hence reflecting the likeliness of a gene to be expressed at a level that corresponds to protein expression, and not background. This was for example the case for DPPA5 which is often detected by qPCR in both hNPSCs and hPSCs, but that is only detected by western blot when the expression levels are around 5000 mRNA molecules per million of mRNA molecules (in hNPSCs) and not detected when expression levels are around 10 molecules per million of mRNA molecules[18].

811 genes corresponding to 1343 missing proteins were unambiguously annotated in transcriptomic datasets. Out of these 811 genes, only 14 genes expressed in hNPSCs (**Figure 2A**) and 9 genes expressed in hTSCs (**Figure 2B**) had abundance over the threshold for credible expression, which is 20 mRNA molecules per million of RNAs. We organized proteins based on their relative expression in hNPSCs or hTSCs. We also displayed expression levels in hPSCs to highlight that the identified missing proteins are globally specific of hNPSCs and might therefore correspond to the preimplantation epiblast (**Figure 2C**).

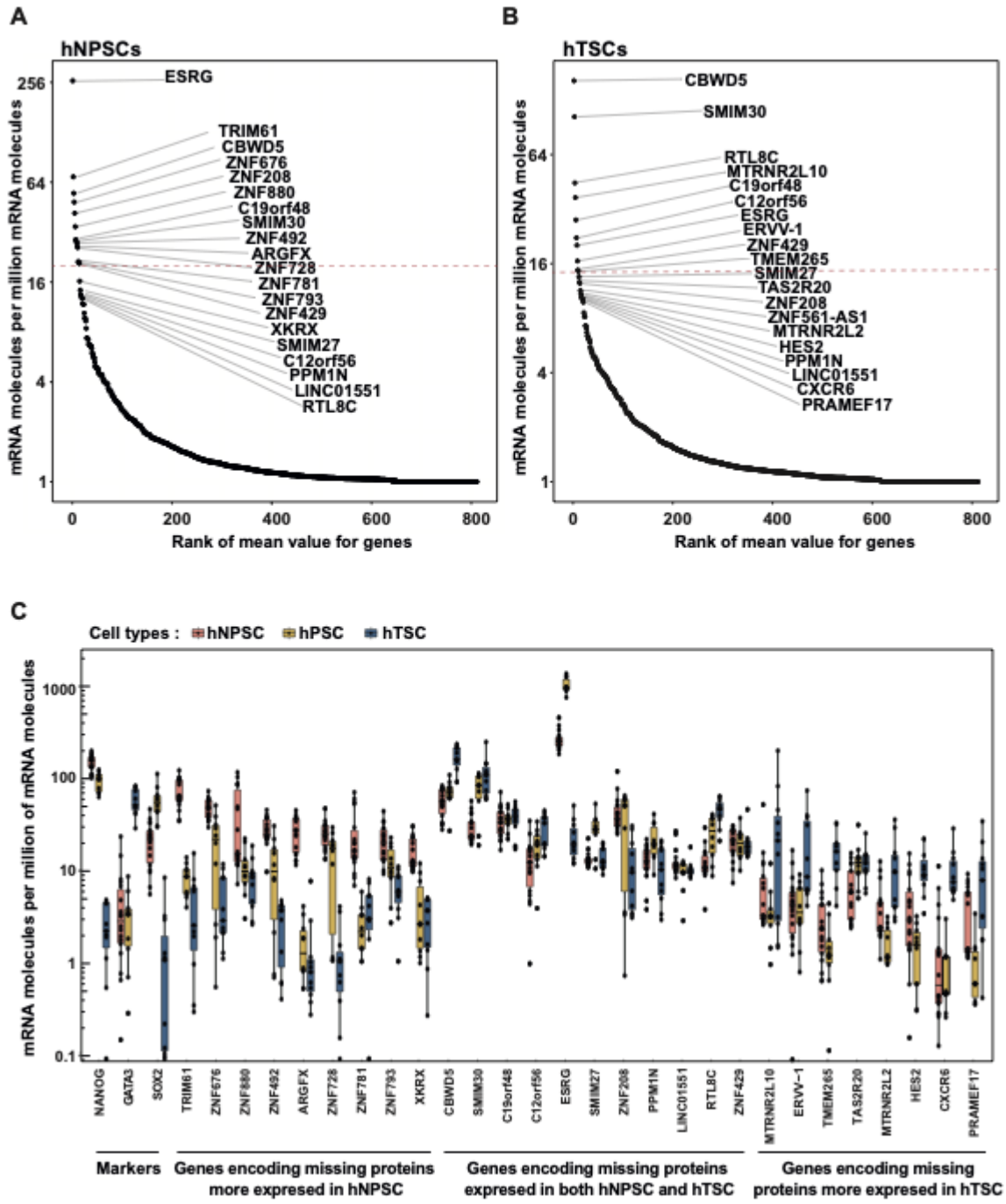

Figure 2: Transcriptomic expression levels of missing proteins in hNPSCs and hTSCs. (**A**, **B**). Distribution of average transcriptomic expression levels (log2(mRNA molecules per million of mRNA molecules+1)) for hNPSCs (**A**) and hTSCs (**B**). We highlighted the top 20 genes; for hNPSCs: ESRG, TRIM61, CBWD5, ZNF676, ZNF208, ZNF880, C19orf48, SMIM30, ZNF492, ARGFX, ZNF728, ZNF781, ZNF793, ZNF429, XKRX, SMIM27, C12orf56, PPM1N, LINC01551 and RTL8C; for hTSCs: CBWD5, SMIM30, RTL8C, MTRNR2L10, C19orf48, c12orf56, ESRG, ERVV-1,

ZNF429, TMEM265, SMIM27, TAS2R20, ZNF208, ZNF561-AS1, MTRNR2L2, HES2, PPM1N, LINC01551, CXCR6 and PRAMEF17.
(**C**) Gene expression levels of the 20 most expressed MPs in hNPSCs (pink), hPSCs (yellow) and hTSCs (blue). NANOG and SOX2 are markers of pluripotency, with higher expression in hNPSCs and hPSCS, respectively. GATA3 is a marker of hTSCs. TRIM61, ZNF676, ZNF880, ZNF492, ARGFX, ZNF728, ZNF781, ZNF793, XKRX are in the 20 most expressed MPs in hNPSCs when RTL8C, MTRNR2L10, ERVV-1, ZNF429, TMEM265, TAS2R20, MTRNR2L2, HES2, CXCR6 , PRAMEF17 are in the 20 most expressed MPs in hTSCs. RTL8C, CBWD5, SMIM30, C19orf48, C12orf56, ESRG, SMIM27, ZNF208, PPM1N, LINC01551 are present in the 2 top 20 MPs most expressed (hTSCs and hNPSCs). For each box, the median, the first and third quartile are displayed.

Interestingly, the 9 genes with highest expression levels in hNPSCS but not in hTSCs are also 3 to 30 times more expressed in hNPSCs than hPSCS, therefore of great potential interest to understand the specificity of preimplantation vs postimplantation EPI (**Figure 2C, Supplementary Table 1**).

Total protein digests from cells dissociated with two different protocols were analyzed by MS/MS using Mascot and Proline search engines. This analysis allowed the identification of 44247 peptides for hNPSCs and 44618 peptides for hTSCs, further mapping to 5150 and 5253 proteins respectively. We analyzed the gene expression levels of genes corresponding to proteins identified by MS/MS. This showed that in both hNPSCs and hTSCs, genes that were expressed over 20 mRNA molecules per million of mRNA molecules had 70% MS identification for hNPSCs and 66% MS identification for hTSCs (**Supplementary Figure 1**). This further confirmed the likelihood of identifying missing proteins in hNPSCs and hTSCs (**Figure 2A, B**).

In a second time the same datasets were processed using PaSER with the objective to catch new peptides that would not have been seen by Mascot. When using the PaSER search engine, 58446 peptides from 5906 proteins were identified in hNPSCs classically dissociated from cultures dishes with trypsin, and 51273 peptides from 5297 proteins were identified in hNPSCs directly dissociated from culture plates with the lysis buffer. As many as 61339 peptides from 6073 proteins were identified in

hTSCs dissociated from culture dishes with trypsin, and 51408 peptides from 5513 proteins in hTSCs directly dissociated from culture plates with the lysis buffer. Detailed information on the proteins identified from our MS datasets is reported in **Supplementary Table 2**.

The list of all identified proteins in our study was searched for missing proteins against neXtProt data (release 2022-02-25). The overall workflow for the detection and validation of missing proteins was used as previously described[6]. Applying this workflow, we produced a list of 16 and 18 “candidate missing proteins” entries in hNPSCs and hTSCs respectively.

In hNPSCs, UQCRHL, CTAGE15, MAP1LC3B2, ZNF98, ZNF732, ZNF728, ZNF208, ZNF804B, ZNF117, ZNF676, ZNF492, RGPD1, CPSF4L, TRIM61, NANOGP8 and WASH2P were considered. In hTSCs, the missing proteins examined correspond to 18 genes: PPIAL4E, PPIAL4C, PPIAL4D, PPIAL4F, TRBV18, RGPD1, CTAGE6, CTAGE15, MAP1LC3B2, DDTL, CGB1, CGB2, CGB7, ZNF732, FBXO47, WASH2P, OR1M1 and OR5M8.

These two subsets were further analyzed in line with version 3.0 of the HPP mass spectrometry data interpretation guidelines[11]. Peptides smaller than 9 amino acids in length were removed and peptide-to-protein mappings were checked using the neXtProt uniqueness checker[13], considering alternative mappings by taking into account the 9.7 million single amino acid variants currently available in neXtProt. As a result, 4 missing proteins could be validated in hNPSCs with at least 2 unique, non-nested peptides of at least 9 amino acids, but none in hTSCs. In addition, single “one-hit wonder” peptides uniquely mapping to 6 other missing proteins could be proposed to the community. Full details (description, number of unique peptides, chromosome location, etc.) of this analysis are reported in **Table 1**.

Table 1: List of Missing Proteins (PE2-PE4) identified inhNPSCs.

Table 1. List of Missing proteins (PE2-PE4) identified in the two datasets

Localization	UniProt ID	Genes ID	Chromosome location	Peptide sequence	Peptide identified by	Spectrum quality[1]	PE level	Protein name	Functional note	Number of validated peptides	SAAV
hNPSCs	P0DKX9	ZNF728	19p12	DFNQSSHLTTHK	PaSER+Mascot /Proline	Medium	PE2	Zinc Finger Protein 728	Involved in regulation of transcription by RNA polymerase II	5	No
				AFSWVSVLNKHK	Mascot/Proline	Low					No
				IGCTNVDECK	PaSER	High					No
				AFIWSSRLSEHK	PaSER	Medium					No
				YANIFHKCSNSK	PaSER	Low					No
	Q5EBN2	TRIM61	4q32.3	FISNPQLGSLTEIAK	PaSER+Mascot /Proline	High	PE2	Putative tripartite motif-containing protein 61	Ubiquitin protein ligase activity	2	No
				LEEYNAPWK	PaSER	Medium					No
	Q8N7Q3	ZNF676	19p12	GFSSVSTLNTHK	PaSER	High	PE2	Zinc Finger Protein 676	Involved in regulation of transcription by RNA polymerase II	2	No
				AFSWSSILTEHKIIHTGEK	PaSER	Low					No
	A4D1E1	ZNF804B	7q21.13	DFSVILKSNHISMTSK	PaSER	Low	PE2	Zinc Finger Protein 804B	Enables metal ion binding	2	No
				CQEQSSNVEISSNSCK	PaSER	Low					No

All peptides but three could only be identified thanks to a search via the PaSER engine (**Table 1**). The capacity of the PaSER search pipeline to detect more proteins can be attributed to the use of the peptides's CCS value to add an extra scoring dimension (TIMScore). The true benefit of TIMScore can be realized during the peptide-validation and FDR estimation steps of the proteomics pipeline. In a non-CCS enabled algorithm, such as Mascot, only two dimen-sions can be utilized to estimate the FDR rate, and

so a discriminate line is fit to a 1% error to distinguish forward and reverse pep-tide candidates. With TIMScore, and the extra CCS dimension, the peptide-candidates can be vectorized in 3-dimensions, allowing a discriminate contoured plane to be applied to achieve the same 1% error. Applying a discriminate plane provides increased accuracy and precision, helping to validate formerly poorly scoring PSMs in the standard two dimensions. Thus, the key effect of TIMScore is derived from the additional dimension of CCS that it provides in assigning true positives from decoy peptide sequences. TIMScore works in a bidirectional fashion, boosting the confidence of borderline peptides under strict FDR thresholds while simultaneously lowering the prob-ability score of a peptide candidate such that it falls below the level of detection. Additionally, the probability score differentiates ambiguous PSMs where the traditional search score cannot distinguish between the first and second (or more) best candidates. Spectra corresponding to peptides exclusively identified using PaSER in this study are provided as **Supplemental material 1**.

Fluorescence immunocytochemical studies on hNPSCs in suspension were undertaken using HPA antibodies to provide orthogonal evidence of the expression of TRIM61 and ZNF728 selected based on our data mining process. However, results were not convincing enough using both antibodies and the approach requires further optimization before drawing any conclusion.

We then carried out a rapid search of the literature and knowledge bases to highlight the potential relevance of the identified missing proteins to the reproduction field. Putative tripartite motif-containing protein 61 (TRIM61) is a RING finger domain protein that is predicted to bind four zinc cations. Many proteins containing a RING finger play a key role in the ubiquitination pathway and TRIM61 is predicted to possess a ubiquitin protein ligase activity[19]. According to the Human Protein Atlas (https://www.proteinatlas.org), the gene is expressed in a large array of tissues and the protein may be localized in the cytoplasm, nucleoli fibrillar center and endoplasmic reticulum. In the mouse, Trim61, also called Rnf35*,* is transcribed temporally in the preimplantation mouse embryo, predominantly at the two-cell embryonic stage. However, the gene is permanently silenced before the blastocyst stage of development. It is thus supposedly implicated in zygotic gene expression[20]. Huang and collaborators have later demonstrated that Rnf35 was actively transcribed from

the newly formed embryonic genome between the late 1-cell and 2-cell stages of early development[21]. Finally, the Trim61 mRNA was shown to bind to Cpeb, a sequence-specific RNA-binding protein that regulates polyadenylation-induced translation, that controls oocyte growth and follicle development in the mouse[22]. The pool of information gathered here on TRIM61 is an example of what should be provided for all PE1 proteins and centralized on a specific repository in the frame of the newly launched HPP “Grand Project”.

The three other missing proteins identified in the present study are classical C2H2 type zinc-finger proteins, *i.e.,* ZNF728, ZNF676 and ZNF804B. They contain one (ZNF804B) or several (ZNF728, ZNF676) zinc-finger domains that consist of short 30 amino acid motifs making tandem contacts with a target molecule. ZNF motifs are stabilized by one or more zinc ions. In numerous C2H2 type zinc-finger proteins, the motif mediates direct interaction with DNA. ZNF728 and ZNF676 are hominoid specific proteins that were predicted to be DNA-binding transcription factors modulating the transcription of specific gene sets transcribed by the RNA polymerase II[19]. Additionally, they both contain a Krüppel associated box (KRAB) domain, typically found in transcription repressors. ZNF676, and possibly ZNF728, have been recently shown to repress the transcriptional activity of a subset of ERV-embedded regulatory sequences active during gametogenesis and early development of the egg[23].

Using our 3’SRP datasets analysis[10,18] corresponding to hNPSCs cells, we observed a concordance between our proteomics and transcriptomics analyses (**Figure 3**).

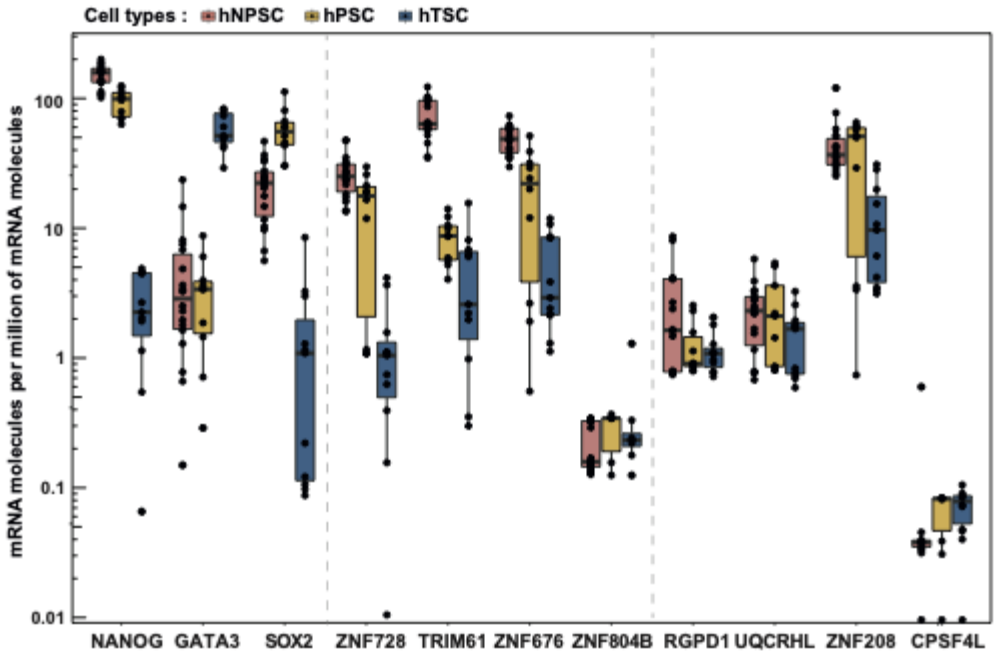

Figure 3. Gene expression levels of missing proteins identified for hNPSCs, hPSCs and hTSCs. Expression levels of the 8 MPs specifically studied in hNPSCs (pink), hPSCs (yellow) and hTSCs (blue). NANOG, GATA3 and SOX2 are included as controls, as in Figure 2. For each box, the median, the first and third quartile are displayed.

Indeed, three MPs identified (TRIM61, ZNF676, ZNF728) at protein level are present in the top 11 MPs most expressed at transcriptomic level. We also observed a difference in expression between hNPSCs and hTSCs/hPSCs with greater transcriptomic expression in hNPSCs than in hTSCs and hPSCs for these three MPs. That is consistent with our proteomic analysis by mass spectrometry. Of note, for our last identified MPs, *i.e.,* ZNF804B, the transcriptomic expression is low in hNPSCs, hTSCs and hPSCs. Additional unique peptides for this protein need to be identified for reinforcing its identification (**Table 1**).

Six other proteins (PE2) (RGPD1, UQCRHL, ZNF208, CPSF4L, TRBV18 and OR5M8) were detected with only one distinct uniquely mapping peptide of length ≥ 9 amino acids (**Table 2**). For these 6 MPs, no additional peptide considered unique but smaller than 9 amino acids could be identified in our datasets. As a consequence, we must be particularly careful with these peptide identifications. Thus, the unambiguous

validation of the corresponding MPs will rely on the identification of other unique peptides.

Table 2: List of “one-hit wonder” proteins to be further validated.

Table 2. List of "one-hit wonder" proteins to be further validated

Localization	UniProt ID	Genes ID	Chromosome location	Peptide sequence	Peptide identified by	Spectrum quality*	PE level	Protein name	Functional note	Number of validated peptides	SAAV
hNPSCs	P0DJD0	RGPD1	2p11.2	MNVMQFNTDR	PaSER	Low	PE2	Ran-binding protein 2-like 6	Contributes to GTPse activator activity	1	No
	A0A096LP55	UQCRHL	1p36.21	ERLELYDEHVSSR	PaSER	Medium	PE2	Cytochrome b-c1 complex subunit 6-like, mitochondrial	Involved in mitochondrial electron transport, ubiquinol to cytochrome c	1	No
	O43345	ZNF208	19p12	WSSTLSYHK	PaSER	Medium	PE2	Zinc Finger Protein 208	Involved in regulation of transcription by RNA polymerase III	1	No
	A6NMK7	CPSF4L	17q25.1	MVVCKHWLR	PaSER	Medium	PE2	Putative cleavage and polyadenylation specificity factor subunit 4-like protein	Enables metal ion and RNA binding	1	No
hTSCs	A0A087X0M5	TRBV18	7q34	FMVYLQKENIIDESGMPK	PaSER	Low	PE2	T cell receptor beats variable 18	Involved in adaptative response	1	No
	Q8NGP6	OR5M8	11q12.1	ELSMKIYFS	PaSER	Low	PE2	Olfactory receptor 5M8	Enables G protein-coupled receptor activity and enables olfactory receptor activity	1	No

For most of these proteins, limited information is available in the literature. The RANBP2-like and GRIP domain-containing protein 1 (RGPD1) is expressed in a large

array of tissues, group enriched in the cervix, placenta and testis, and mainly localized in nuclear membranes. The protein is predicted to contribute to GTPase activator activity[19]. Of note is that this protein is identical to RGPD2 (NX_P0DJD1) except in the first 16 amino acids. The peptide identified in the present work (aa 1-10) is the only one that allows the two proteins to be distinguished. Further validation of this protein in the frame of the HPP will be impossible with the current guidelines.This is clearly a case that justifies a derogation in the current HPP mass spectrometry data identification guidelines or the addition of a dedicated paragraph in its future version.

UQCRHL, the Cytochrome b-c1 complex subunit 6-like, mitochondrial protein is expressed in a large set of tissues and cell-type enriched in cardiomyocytes. The protein is predicted to be localized in the mitochondria inner membrane where it would be a component of the respiratory chain complex III and catalyze the oxidation of ubiquinol by oxidized cytochrome c1. Yet another unique peptide for UQCRHL has been reported in Peptide Atlas. Thanks to our contribution, this protein should be validated in the coming future.

ZNF208 is another Krüppel C2H2-type zinc-finger protein with low tissue specificity and that can be considered as a transcription factor. It is among the highest expressed transcripts in hNPSCs and hTSCs. Of note is that another unique peptide for ZNF208 has been reported in Peptide Atlas. Again, thanks to our contribution, this protein should be soon validated.

CPSF4L, the Putative cleavage and polyadenylation specificity factor subunit 4-like is predicted to be a RNA-binding protein involved in pre-mRNA cleavage required for polyadenylation[19]. CPSF4L mRNA expression level is surprisingly low, questioning whether this protein is present in the hNPSCs or whether its expression is difficult to detect by 3'SRP.

TRBV18 is the variable region 18 of the T cell receptor beta chain located in the plasma membrane. This region is responsible for recognizing fragments of antigen as peptides bound to major histocompatibility complex (MHC) molecules.

Finally, OR5M8 is the olfactory receptor 5M8. It is predicted to be a G protein-coupled receptor, involved in the detection of chemical stimuli and sensory perception of smell. The unique peptide identified in our study corresponds to a cytoplasmic domain of the protein. Considering no olfactory receptor has been so far unambiguously identified by mass spectrometry in the frame of the HPP, additional effort is planned in our

laboratory to further analyze hNPSCs and hTSCs protein extracts. As a matter of fact, other unique peptides corresponding to extracellular domains of OR5M8 could potentially be obtained through a controlled sample digestion by some other enzymes. Interestingly there is another long unique peptide reported by MassIVE: ESVEQGKMVAVFYTTVIPMLNLIIYSLRNKNVKEALIK (mzspec:PXD022531:j7912_PDIA6.mzXML:scan:9162:ESVEQGKMVAVFYTTVIPMLNLIIYSLRNKNVKEALIK/3). Of note is that the mRNA corresponding to TRBV18 and OR5M8 were not or barely detected in the list of 24,849 transcripts generated during our transcriptomic analysis; which eventually makes the identification of these unique peptides questionable and calls for a cross-validation by the trans-Proteomics Pipeline. As far as OR5M8 is concerned, as the identification of the unique peptide was good, the only credible explanation is that its mRNA was at the limit of detection of the 3'seq-RNA profiling approach used.

The present study shows that we have reached lower technical limits for identifying missing proteins in mass spectrometry but also in sample preparation. In the present work, the use of the TIMScore and peptide Collisional Cross Section (CCS), only made possible on a Trapped Ion Mobility Spectrometry instrument (*i.e.,* Tims TOF Pro; Bruker Daltonik), was shown to be a valuable additional feature that strengthens peptide identification by mass spectrometry. Thus, the PaSER search engine will be systematically used in our future studies.
As regards sample preparation, interestingly, MPs evidenced in this work are generally recognized as discrete gene products, *e.g.,* transcription factors. We are thus continuing our efforts to identify missing proteins in hNPSCs and hTSCs extracts thanks to additional enrichment strategies offering better resolution to favor the selective extraction of membranous, membrane-bound and nuclear proteins. The protein digestion tool recently developed by neXtProt (www.nextprot.org/tools/protein-digestion) can also be used wisely to determine enzymes alternative to trypsin and select the experimental conditions that would yield additional unique peptides to confidently identify missing proteins.

Among the top 20 missing proteins expected to be found in hNPSCs and hTSCs cellular extracts based on mRNA expression, several were not identified in the present study. Interestingly, according to the literature and to HUGO Gene Nomenclature

Committee curators, the LINC01551 and ZNF561-AS1 genes are probably not protein coding. Additionally, the expression of several proteins cannot be validated using the current HPP guidelines. A few examples are CBWD5 that has only 1 amino acid difference with CNWD3; the mature form of SMIM30 is 35 amino acids long and MTRNR2L10 and MTRNR2L2 are also short proteins <30 amino acids.
We should then focus on proteins such as ESRG that are highly expressed and should be found even with conventional trypsin digestion. Sample preparation will be key here to access the missing proteins that could not be seen in the first round. Additionally, we plan to use targeted mass assays using a PRM acquisition approach to target unique peptides matching the current HPP data interpretation guidelines.

CONCLUSION

We demonstrate here that early development stages harbor unique missing proteins. Indeed, this is the first time that these 4 proteins have been found in particularly understudied samples. They might be involved in events restricted to the naive stem cells lineage, and have a crucial role in setting up the molecular events that underlie early embryonic development. Identifying the transcription factors involved in the establishment and maintenance of human naive pluripotency is an important focus for the field. Knowing that those transcription factors exist at the protein level will support further biological investigations, and the development of gene-editing and invalidation approaches to create knock-out cell lines (*e.g.,* CRISPR-Cas9). Moreover, identification of those transcription factors will also trigger their further study in human embryos.

In this paper several peptides unique to missing proteins were identified thanks to the PaSER search engine. This new validation strategy, supported by the PaSER search engine, makes use of the correlation between theoretical (predicted) and measured CCS value for each peptide as an extra scoring dimension. This correlation is combined to the more classical fragmentation-based correlation pattern to rescue out of the 1% FDR plane some of the peptides showing a low-quality MS/MS fragmentation pattern but with a good CCS correlation. The feasibility of an accurate CCS value prediction for peptides has been demonstrated recently and this feature is now exploited as an extra filtering value for peptide candidates along the identification

process, and is also used in the FDR calculation process. All presented peptides have passed a 1% peptide FDR threshold.

The HPP mass spectrometry data interpretation guidelines version 3.0 do not take into consideration CCS values. To date, CCS values can only be generated when mass spectrometry analyses are performed on a mass spectrometer that is equipped for ion mobility. Yet major constructors have launched ion mobility instruments. It appears to us that ion mobility is the future of mass spectrometry in proteomics. As a consequence, the use of CCS values to interpret spectra and validate peptide should be taken into consideration and discussed when preparing the next version of the guidelines.

We focused our study on the prominent stem cell types modeling peri-implantation, but given the technical progresses of mass spectrometry for proteomics in recent years, the analysis of primed PSCs could also lead to the identification of another set of missing proteins.

AUTHORS CONTRIBUTIONS

CP and LD co-coordinated the study. CP conceived and designed the HPP mass spectrometry experiments and analyses. SC processed transcriptomics data. CO produced stem cell lines. OG performed the proteomics sample fractionation and preparation. RL performed MS/MS analyses. RL, EC processed and analyzed the MS/MS datasets. MC and POS performed PaSER analyses. OG and EC performed the bioinformatics analyses. OG, CP, LD and LL performed data/ literature mining of the identified proteins and selected candidates. OG, SC and LD prepared the figures, tables, and supporting Information. OG, CP and LD drafted the manuscript.

Corresponding Author *

Charles Pineau: charles.pineau@inserm.fr, Tel: +33 (0)2 2323 5072

ORCID number: 0000-0002-7461-5433

ACKNOWLEDGMENTS

O. Girard holds a BIRTH GRANT 2021 fellowship. This work was also supported by structural grants from Biogenouest, Infrastructures en Biologie Santé et Agronomie (IBiSA), and the Conseil Régional de Bretagne awarded to C.P. We are grateful to Cecilia Lindskog (Human Protein Atlas, Uppsala, Sweden) for the gift of TRIM61 and ZNF728 antibodies.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ABBREVIATIONS

C-HPP, chromosome-centric human proteome project; HPP, Human Proteome Project; HUPO, Human Proteome Organization; PE1, existence based on evidence at protein level; MPs, missing proteins; PE2, existence based on evidence at transcript level; TE, trophectoderm; ICM, inner cell mass; EPI, pluripotent epiblast; PrE, primitive endoderm; hPSCs, human primed stem cells; hNPSCs, human naive pluripotent stem cells; hiTSCs, human induced naive pluripotent stem cells; hTSCs, human trophoblastic stem cells; FDR, false discovery rate; PSM, peptide spectrum match; CCR, collisional cross section; BSA,Bovine Serum Albumin; 3'SRP, 3'seq-RNA Profiling; UMI, unique molecular identifiers; RPL, Ribosomal Protein Large; RPS, ribosomal protein small.

REFERENCES

(1) Legrain, P.; Aebersold, R.; Archakov, A.; Bairoch, A.; Bala, K.; Beretta, L.; Bergeron, J.; Borchers, C. H.; Corthals, G. L.; Costello, C. E.; Deutsch, E. W.; Domon, B.; Hancock, W.; He, F.; Hochstrasser, D.; Marko-Varga, G.; Salekdeh, G. H.; Sechi, S.; Snyder, M.; Srivastava, S.; Uhlén, M.; Wu, C. H.; Yamamoto, T.; Paik, Y.-K.; Omenn, G. S. The Human Proteome Project: Current State and Future Direction. *Mol Cell Proteomics* **2011**, *10* (7), M111.009993. https://doi.org/10.1074/mcp.M111.009993.

(2) Adhikari, S.; Nice, E. C.; Deutsch, E. W.; Lane, L.; Omenn, G. S.; Pennington, S. R.; Paik, Y.-K.; Overall, C. M.; Corrales, F. J.; Cristea, I. M.; Van Eyk, J. E.; Uhlén, M.; Lindskog, C.; Chan, D. W.; Bairoch, A.; Waddington, J. C.; Justice, J. L.; LaBaer, J.; Rodriguez, H.; He, F.; Kostrzewa, M.; Ping, P.; Gundry, R. L.; Stewart, P.; Srivastava, S.; Srivastava, S.; Nogueira, F. C. S.; Domont, G. B.; Vandenbrouck, Y.; Lam, M. P. Y.; Wennersten, S.; Vizcaino, J. A.; Wilkins, M.; Schwenk, J. M.; Lundberg, E.; Bandeira, N.; Marko-Varga, G.; Weintraub, S. T.; Pineau, C.; Kusebauch, U.;

Moritz, R. L.; Ahn, S. B.; Palmblad, M.; Snyder, M. P.; Aebersold, R.; Baker, M. S. A High-Stringency Blueprint of the Human Proteome. *Nat Commun* **2020**, *11* (1), 5301. https://doi.org/10.1038/s41467-020-19045-9.

(3) Lane, L.; Bairoch, A.; Beavis, R. C.; Deutsch, E. W.; Gaudet, P.; Lundberg, E.; Omenn, G. S. Metrics for the Human Proteome Project 2013–2014 and Strategies for Finding Missing Proteins. *J. Proteome Res.* **2014**, *13* (1), 15–20. https://doi.org/10.1021/pr401144x.

(4) Jumeau, F.; Com, E.; Lane, L.; Duek, P.; Lagarrigue, M.; Lavigne, R.; Guillot, L.; Rondel, K.; Gateau, A.; Melaine, N.; Guével, B.; Sergeant, N.; Mitchell, V.; Pineau, C. Human Spermatozoa as a Model for Detecting Missing Proteins in the Context of the Chromosome-Centric Human Proteome Project. *J Proteome Res* **2015**, *14* (9), 3606–3620. https://doi.org/10.1021/acs.jproteome.5b00170.

(5) Carapito, C.; Duek, P.; Macron, C.; Seffals, M.; Rondel, K.; Delalande, F.; Lindskog, C.; Fréour, T.; Vandenbrouck, Y.; Lane, L.; Pineau, C. Validating Missing Proteins in Human Sperm Cells by Targeted Mass-Spectrometry- and Antibody-Based Methods. *J Proteome Res* **2017**, *16* (12), 4340–4351. https://doi.org/10.1021/acs.jproteome.7b00374.

(6) Vandenbrouck, Y.; Lane, L.; Carapito, C.; Duek, P.; Rondel, K.; Bruley, C.; Macron, C.; Gonzalez de Peredo, A.; Couté, Y.; Chaoui, K.; Com, E.; Gateau, A.; Hesse, A.-M.; Marcellin, M.; Méar, L.; Mouton-Barbosa, E.; Robin, T.; Burlet-Schiltz, O.; Cianferani, S.; Ferro, M.; Fréour, T.; Lindskog, C.; Garin, J.; Pineau, C. Looking for Missing Proteins in the Proteome of Human Spermatozoa: An Update. *J Proteome Res* **2016**, *15* (11), 3998–4019. https://doi.org/10.1021/acs.jproteome.6b00400.

(7) Melaine, N.; Com, E.; Bellaud, P.; Guillot, L.; Lagarrigue, M.; Morrice, N. A.; Guével, B.; Lavigne, R.; Velez de la Calle, J.-F.; Dojahn, J.; Pineau, C. Deciphering the Dark Proteome: Use of the Testis and Characterization of Two Dark Proteins. *J Proteome Res* **2018**, *17* (12), 4197–4210. https://doi.org/10.1021/acs.jproteome.8b00387.

(8) Gerri, C.; McCarthy, A.; Alanis-Lobato, G.; Demtschenko, A.; Bruneau, A.; Loubersac, S.; Fogarty, N. M. E.; Hampshire, D.; Elder, K.; Snell, P.; Christie, L.; David, L.; Van de Velde, H.; Fouladi-Nashta, A. A.; Niakan, K. K. Initiation of a Conserved Trophectoderm Program in Human, Cow and Mouse Embryos. *Nature* **2020**, *587* (7834), 443–447. https://doi.org/10.1038/s41586-020-2759-x.

(9) Pera, M. F.; Rossant, J. The Exploration of Pluripotency Space: Charting Cell State Transitions in Peri-Implantation Development. *Cell Stem Cell* **2021**, *28* (11), 1896–1906. https://doi.org/10.1016/j.stem.2021.10.001.

(10) Castel, G.; Meistermann, D.; Bretin, B.; Firmin, J.; Blin, J.; Loubersac, S.; Bruneau, A.; Chevolleau, S.; Kilens, S.; Chariau, C.; Gaignerie, A.; Francheteau, Q.; Kagawa, H.; Charpentier, E.; Flippe, L.; François-Campion, V.; Haider, S.; Dietrich, B.; Knöfler, M.; Arima, T.; Bourdon, J.; Rivron, N.; Masson, D.; Fournier, T.; Okae, H.; Fréour, T.; David, L. Induction of Human Trophoblast Stem Cells from Somatic Cells and Pluripotent Stem Cells. *Cell Rep* **2020**, *33* (8), 108419. https://doi.org/10.1016/j.celrep.2020.108419.

(11) Deutsch, E. W.; Lane, L.; Overall, C. M.; Bandeira, N.; Baker, M. S.; Pineau, C.; Moritz, R. L.; Corrales, F.; Orchard, S.; Van Eyk, J. E.; Paik, Y.-K.; Weintraub, S. T.; Vandenbrouck, Y.; Omenn, G. S. Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 3.0. *J. Proteome Res.* **2019**, *18* (12), 4108–4116. https://doi.org/10.1021/acs.jproteome.9b00542.

(12) Xu, Y.; Zhao, J.; Ren, Y.; Wang, X.; Lyu, Y.; Xie, B.; Sun, Y.; Yuan, X.; Liu, H.; Yang, W.; Fu, Y.; Yu, Y.; Liu, Y.; Mu, R.; Li, C.; Xu, J.; Deng, H. Derivation of Totipotent-like Stem Cells with Blastocyst-like Structure Forming Potential. *Cell Res* **2022**. https://doi.org/10.1038/s41422-022-00668-0.

(13) Schaeffer, M.; Gateau, A.; Teixeira, D.; Michel, P.-A.; Zahn-Zabal, M.; Lane, L. The NeXtProt Peptide Uniqueness Checker: A Tool for the Proteomics Community. *Bioinformatics* **2017**, *33* (21), 3471–3472. https://doi.org/10.1093/bioinformatics/btx318.

(14) Charpentier, E.; Cornec, M.; Dumont, S.; Meistermann, D.; Bordron, P.; David, L.; Redon, R.; Bonnaud, S.; Bihouée, A. *3' RNA Sequencing for Robust and Low-Cost Gene Expression Profiling*; preprint; Protocol Exchange, 2021. https://doi.org/10.21203/rs.3.pex-1336/v1.

(15) Vizcaíno, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Ríos, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; Binz, P.-A.; Xenarios, I.; Eisenacher, M.; Mayer, G.; Gatto, L.; Campos, A.; Chalkley, R. J.; Kraus, H.-J.; Albar, J. P.; Martinez-Bartolomé, S.; Apweiler, R.; Omenn, G. S.; Martens, L.; Jones, A. R.; Hermjakob, H. ProteomeXchange Provides Globally Coordinated Proteomics Data Submission and Dissemination. *Nat Biotechnol* **2014**, *32* (3), 223–226. https://doi.org/10.1038/nbt.2839.

(16) Takashima, Y.; Guo, G.; Loos, R.; Nichols, J.; Ficz, G.; Krueger, F.; Oxley, D.; Santos, F.; Clarke, J.; Mansfield, W.; Reik, W.; Bertone, P.; Smith, A. Resetting Transcription Factor Control Circuitry

toward Ground-State Pluripotency in Human. *Cell* **2014**, *158* (6), 1254–1269. https://doi.org/10.1016/j.cell.2014.08.029.
(17) Okae, H.; Toh, H.; Sato, T.; Hiura, H.; Takahashi, S.; Shirane, K.; Kabayama, Y.; Suyama, M.; Sasaki, H.; Arima, T. Derivation of Human Trophoblast Stem Cells. *Cell Stem Cell* **2018**, *22* (1), 50-63.e6. https://doi.org/10.1016/j.stem.2017.11.004.
(18) Kilens, S.; Meistermann, D.; Moreno, D.; Chariau, C.; Gaignerie, A.; Reignier, A.; Lelièvre, Y.; Casanova, M.; Vallot, C.; Nedellec, S.; Flippe, L.; Firmin, J.; Song, J.; Charpentier, E.; Lammers, J.; Donnart, A.; Marec, N.; Deb, W.; Bihouée, A.; Le Caignec, C.; Pecqueur, C.; Redon, R.; Barrière, P.; Bourdon, J.; Pasque, V.; Soumillon, M.; Mikkelsen, T. S.; Rougeulle, C.; Fréour, T.; David, L.; Milieu Intérieur Consortium. Parallel Derivation of Isogenic Human Primed and Naive Induced Pluripotent Stem Cells. *Nat Commun* **2018**, *9* (1), 360. https://doi.org/10.1038/s41467-017-02107-w.
(19) Gaudet, P.; Livstone, M. S.; Lewis, S. E.; Thomas, P. D. Phylogenetic-Based Propagation of Functional Annotations within the Gene Ontology Consortium. *Brief Bioinform* **2011**, *12* (5), 449–462. https://doi.org/10.1093/bib/bbr042.
(20) Chen, H.-H.; Liu, T. Y.-C.; Li, H.; Choo, K.-B. Use of a Common Promoter by Two Juxtaposed and Intronless Mouse Early Embryonic Genes, Rnf33 and Rnf35: Implications in Zygotic Gene Expression. *Genomics* **2002**, *80* (2), 140–143. https://doi.org/10.1006/geno.2002.6808.
(21) Huang, C.-J.; Wu, S.-C.; Choo, K.-B. Transcriptional Modulation of the Pre-Implantation Embryo-Specific Rnf35 Gene by the Y-Box Protein NF-Y/CBF. *Biochem J* **2005**, *387* (Pt 2), 367–375. https://doi.org/10.1042/BJ20041364.
(22) Racki, W. J.; Richter, J. D. CPEB Controls Oocyte Growth and Follicle Development in the Mouse. *Development* **2006**, *133* (22), 4527–4537. https://doi.org/10.1242/dev.02651.
(23) Iouranova, A.; Grun, D.; Rossy, T.; Duc, J.; Coudray, A.; Imbeault, M.; de Tribolet-Hardy, J.; Turelli, P.; Persat, A.; Trono, D. KRAB Zinc Finger Protein ZNF676 Controls the Transcriptional Influence of LTR12-Related Endogenous Retrovirus Sequences. *Mobile DNA* **2022**, *13* (1), 4. https://doi.org/10.1186/s13100-021-00260-0.

(Figure created with BioRender.com)

FOR TOC ONLY