Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome

Benoit Vanderperre1, Jean-François Lucier2, Cynthia Bissonnette1, Julie Motard1, Guillaume Tremblay1, Solène Vanderperre1, Maxence Wisztoriski3, Michel Salzet3, François-Michel Boisvert4, Xavier Roucou1*

1 Département de biochimie, Faculté de Médecine et des Sciences de la Santé, Université de Sherbrooke, Québec, Canada, 2 Département de microbiologie, Faculté de Médecine et des Sciences de la Santé, Université de Sherbrooke, Québec, Canada, 3 PRISM, Laboratoire de Protéomique, Réponse Inflammatoire, Spectrométrie de Masse, EA 4550, SN3, Université Lille 1, Villeneuve d’Ascq, France, 4 Département d’anatomie et de biologie cellulaire, Faculté de Médecine et des Sciences de la Santé, Université de Sherbrooke, Québec, Canada

Abstract

A fully mature mRNA is usually associated to a reference open reading frame encoding a single protein. Yet, mature mRNAs contain unconventional alternative open reading frames (AltORFs) located in untranslated regions (UTRs) or overlapping the reference ORFs (RefORFs) in non-canonical +2 and +3 reading frames. Although recent ribosome profiling and footprinting approaches have suggested the significant use of unconventional translation initiation sites in mammals, direct evidence of large-scale alternative protein expression at the proteome level is still lacking. To determine the contribution of alternative proteins to the human proteome, we generated a database of predicted human AltORFs revealing a new proteome mainly composed of small proteins with a median length of 57 amino acids, compared to 344 amino acids for the reference proteome. We experimentally detected a total of 1,259 alternative proteins by mass spectrometry analyses of human cell lines, tissues and fluids. In plasma and serum, alternative proteins represent up to 55% of the proteome and may be a potential unsuspected new source for biomarkers. We observed constitutive co-expression of RefORFs and AltORFs from endogenous genes and from transfected cDNAs, including tumor suppressor p53, and provide evidence that out-of-frame clones representing AltORFs are mistakenly rejected as false positive in cDNAs screening assays. Functional importance of alternative proteins is strongly supported by significant evolutionary conservation in vertebrates, invertebrates, and yeast. Our results imply that coding of multiple proteins in a single gene by the use of AltORFs may be a common feature in eukaryotes, and confirm that translation of unconventional ORFs generates an as yet unexplored proteome.


Editor: Eric Jan, University of British Columbia, Canada

Received April 16, 2013; Accepted June 20, 2013; Published August 12, 2013

Copyright: © 2013 Vanderperre et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: Most of this work was supported by an operating grant from the Canadian Institutes for Health Research MOP-89881 to XR. BV was supported in part by a scholarship from the Fonds de Recherche Nature et Technologies du Québec and CB by a scholarship from the National Sciences and Engineering Research Council of Canada. Work in François-Michel Boisvert was supported by an operating grant from the Canadian Institutes for Health Research (MOP-123469). Work in Michel Salzet’s laboratory was supported by the Ministère de L’Education Nationale de l’Enseignement Supérieur et de la Recherche (to MS), Agence Nationale de la Recherche, and a collaboration between PRISM Laboratory and Thermo Fisher. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: There was no financial support from Thermo Fisher, only material support (to Dr Salzet’s laboratory). Importantly, this collaboration does not alter the authors’ adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: xavier.roucou@usherbrooke.ca

Introduction

The proteome impacts all aspects of health and disease and deciphering the human proteome represents an important challenge in the post-genomic era. A typical fully processed mRNA includes one RefORF and is associated with a reference protein (Fig. 1A). Reference proteins populate current protein databases used to support research in life sciences. For example, protein databases are central to the success of mass spectrometry-based protein identification for the discovery of expression and interaction proteomics, and of biomarkers [1].

Two cellular mechanisms have evolved to promote proteomic diversity by encoding more than one protein per gene, increasing the diversity of the transcriptome or producing more than one protein from a single transcript. Transcriptome diversity [2] is achieved by utilization of alternative promoters [3], reiterative transcription [4], or post-transcriptional processing, including alternative splicing [5], alternative polyadenylation [6] and RNA editing [7]. On the other hand, N-terminal extension [8], ribosomal frameshifting [9,10] and utilization of multiple coding ORFs in one transcript [11,12] can generate functional or disease-related proteins.

Out-of-frame alternative translation initiation in the same transcript is used to encode proteins of totally different amino acid composition and is mainly observed in viruses and bacteriophages [13]. This mechanism provides such small organisms with increased coding capacity. Until recently, very few examples were documented in human [11,12,14–18], and it was assumed such mechanisms were anecdotal in eukaryotes considering the flexibility and coding capacity provided by their large genome. Two recent studies described the discovery of
several protein products resulting from translation of non-canonical, alternative open reading frames (AltORFs) [19,20], present in 5’ UTRs, overlapping the RefORF, or in 3’ UTRs. These studies demonstrate that the human proteome is more complex than previously appreciated and suggest that there are many more alternative proteins that remain undiscovered.

Paradoxically, several databases predicting AltORFs in eukaryotes are available [21]. Yet, most of these databases did not actually predict the rare examples of AltORFs translation products documented in humans [11,12,14–18]. Most importantly these databases did not include AltORFs present within UTRs. To address this issue, we generated a database of predicted AltORFs present in mature human mRNAs.

Here, we provide evidence that AltORFs located within UTRs or overlapping the RefORF in a different reading frame are translated and significantly contribute to the human proteome. We demonstrate that basic transfection of cDNAs containing a RefORF and an AltORF results in coexpression of the reference and alternative proteins. In addition, we provide evidence that alternative proteins are conserved with a high sequence identity in vertebrates and invertebrates, suggesting an important function for this unexplored proteome.

Methods

Ethics protocol (human tissues)
Ovarian, fallopian tube and endometrial formalin fixed, paraffin-embedded tissues were obtained from the CHRU de Lille pathology department (institutional review approval from the Ethical Research Committee CPP Nord Ouest IV 12/10). The ethical committee considered contacting the patients, often many years after surgery, to be unnecessary and waived the need for written informed consent from the participants.

AltORFs database
Few databases predicting alternative ORFs (AltORFs) in eukaryotes are available. Yet, most of these databases did not actually predict the very rare examples of out-of-frame alternative translation products documented in humans. Most importantly these databases did not include AltORFs present within UTRs.
Other criteria included in previous predictions—such as conservation among species, presence of an optimal Kozak context around the initiator AUG codon, location within the reference coding sequence—were not taken into account in this study to reduce biases in prediction and generate a comprehensive database.

The AltORF database for each species was built as previously described [21] with minor modifications to the pipeline of Perl scripts that populate a MySQL database, to retain AltORFs regardless of the Kozak context or the location on the mRNA. GenBank entries (http://www.ncbi.nlm.nih.gov/genbank/) were used to generate the AltORF databases for human (release 37), chimpanzee (release 102), mouse (release 103), cow (version 1) and frog (version 1), whereas Ensembl entries (http://www.ensembl.org/) were used to generate the databases for fly (version 70.546), nematode (version 70.230) and yeast (version 70.4). The AltORF databases are provided as excel files (Databases S1–S8).

Antibodies

Primary antibodies used for western blot were monoclonal anti-HA (Covance), monoclonal anti-GFP (SantaCruz), polyclonal anti-NPTII (Millipore), monoclonal anti-BRCA1 (Bethyl). Secondary antibodies used for western blots were horseradish peroxidase (HRP)-conjugated sheep anti-mouse IgG (NA931V) and HRP-conjugated donkey anti-rabbit IgG (NA934V) (GE Healthcare). Primary antibody used for immunofluorescence was monoclonal anti-BRCA1 (Novus Biologicals). Secondary antibodies used for immunofluorescence was Alexa Fluor 568-conjugated goat anti-mouse IgG (Invitrogen).

Clones

These clones were generated to verify translation initiation at predicted AUG codons, to confirm the coexpression of reference and alternative proteins from the same mRNA, and to visualize the expression of alternative proteins using a GFP tag.

**Predicted AltORFs with initiation and stop codons in 5’UTRs.** SLCO3A1 and ZNF83 (GenBank mRNA entries NM_080670 and NM_001105592, respectively): gBlocks (IDT) gene fragments containing the 5’UTR sequence up to the last coding nucleotide of their respective AltORF were designed and inserted into the BamHI restriction site of pEGFP-N1 using the Gibson assembly kit (New England Biolabs) according to the manufacturer’s protocol.

**Predicted AltORFs overlapping the 5’UTR and the reference CDS.** IDH3B (NM_174856). This construct was synthesized as gBlocks gene fragments (IDT). The construct was assembled and inserted into the BamHI site of pcDNAHA-EGFP+3 using the Gibson assembly kit. pcDNAHA-EGFP+3 was designed to express HA-tagged reference proteins and GFP-tagged alternative proteins with the corresponding AltORF in the +3 reading frame.

**Predicted AltORFs entirely included in the reference CDS.** NIPA1 (NM_001142275), BDH2 (NM_020139), SCARB2 (NM_005506), LGALS3BP (NM_005567), CDC42 (NM_144681), VEGFC (NM_005429), BDKRB2 (NM_000623), TP53 (NM_000546), and SRSF1 (NM_006924). For NIPA1, BDH2, SCARB2, LGALS3BP, CDC42 and VEGFC, constructs were synthesized as gBlocks gene fragments (IDT), then assembled and inserted into the BamHI site of pcDNAHA-EGFP+2 using the Gibson assembly kit. For TP53 and SRSF1, inserts were PCR amplified from cDNA containing plasmids (Addgene). The resulting PCR fragments were inserted into the BamHI site of pcDNAHA-EGFP+2. BDRKB2 was PCR amplified from a cDNA containing plasmid (clone SC119794, OriGene), and the resulting PCR fragments were inserted into the BamHI site of pcDNAHA-EGFP+3.

**Figure 2. Endogenous expression of alternative proteins in cultured cells.** (A) Alternative proteins expression was analyzed by LC-MS/MS in HeLa cells, and a schematic distribution of AltORFs in absolute numbers is shown. There were a total of 129 identified alternative proteins (indicated in the center). (B) Average peptide intensity plot of both the reference and alternative proteins that were co-expressed from 42 genes. (C) Same as (A) with colon cell lines. (D) Venn diagram showing the number of alternative proteins identified in HeLa cells and colon cell lines. The overlap identifies a common list of 14 alternative proteins.

doi:10.1371/journal.pone.0070698.g002
Mutation of the predicted alternative initiation ATG codons to AAG was performed with complementary oligonucleotides containing the mutation, using the QuikChange kit (Stratagene). Each construct contained the endogenous Kozak sequence (positions 2 to +4) around both the RefORF and AltORF putative initiation codons. 5' UTR regions were present in the constructs only when the corresponding AltORF was at least partially contained within this region.

The AltMRVI1EGFP fusion was obtained by inserting a gBlock gene fragment containing the AltMRVI1 coding sequence (GenBank mRNA entry NM_001100167, nucleotides 5403 to 5688) into the BamHI digested pEGFP-N1 plasmid. Primers and sequence verified gBlocks gene fragments were purchased from IDT. All constructs were sequenced in both orientations.

Cell culture and transfection
Human epithelial kidney (HEK293), human cervical cancer HeLa and human colon CCL227, CCL228, CCL233, CRL1459, and HCT116 cells were grown in Dulbecco’s Modified Eagle’s Medium supplemented with 10% Fetal Bovine Serum and penicillin/streptomycin. Human colon cell lines CCL227, CCL228, CCL233, CRL1459, and HCT116 were purchased from ATCC. Transfections were carried out using GeneCellin according to the manufacturer’s protocol (BioCellChallenge).

Protein sample preparation, immunoprecipitation and western blot
For validation of alternative proteins expression by HA- and GFP-tagging, HeLa cells were grown in 12-well plates for 24 h and were then transfected as described above. Cells were rinsed with PBS and lysed in SDS-PAGE sample buffer (0.5% SDS (w/v), 1.25% 2-β-mercaptoethanol (v/v), 4% glycerol (v/v), 0.01% bromophenol blue (w/v), 15 mM Tris-HCl, pH 6.8). After electrophoresis, proteins were transferred to PVDF membranes and detected by western blot using anti-HA (Covance, 1/1000), anti-GFP (Santa Cruz, 1/1000), and anti-NPTII (Millipore, 1/1000) antibodies.

For immunoprecipitation of AltMRVI1EGFP, HEK 293 cells were grown in 100 mm plates for 24 h before transfection. After 24 h, cells were rinsed twice with ice-cold PBS and lysed in 1 mL NETN buffer (50 mM Tris-HCl, pH 8.0, 0.15 M NaCl, 1 mM EDTA, 0.5% NP-40, with protease inhibitors (Roche) and protein phosphatase inhibitors (Thermo Scientific)) for 15 min at 4°C. Nuclei were broken by successive passing in 18G, 20G, 21G and 25G needles, and the lysate was centrifuged at 15,000 g for 15 min at 4°C. Protein concentration was quantified using BCA protein assay reagent (Pierce). Ten µL GFP-Trap agarose beads (ChromoTek) were used for GFP immunoprecipitation of 0.75 mg sample (1 mg/mL in NETN buffer) during 1 h at 4°C. The beads were then washed three times with 1 mL of NETN buffer, and the bound proteins eluted by incubating for 5 min at 95°C in SDS-PAGE sample buffer. Proteins were detected by western blot using anti-GFP (Santa Cruz, 1/1000), and anti-BRCA1 (Bethyl, 1/1000) antibodies.

Protein samples preparation and in-gel digestion of proteins from normal colon tissue and colon cell lines prior to LC-MS/MS analysis were performed as previously described [22]. For LC-MS/MS analysis of proteins contained between the 4.6 and 10 kDa markers of an 1D SDS-PAGE, HeLa cells were lysed in a buffer containing 4% SDS/100 mM DTT/100 mM Tris-HCl pH 8.0 and 5 μg of total proteins was loaded per lane. For the analysis of proteins contained between the 10 kDa and 25 kDa markers of an 1D SDS-PAGE, detergent samples were directly analyzed by LC-MS/MS.

Table 1. Summary of LC-MS/MS analyses of human samples.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Alternative proteins (peptides)</th>
<th>Reference proteins (peptides)</th>
<th>Alternative proteins (% of total identified proteins)</th>
<th>% sequence coverage (alternative/reference)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fractionated HeLa cells</td>
<td>129 (280)</td>
<td>5,558 (68,035)</td>
<td>2.27</td>
<td>28.8/32.3</td>
</tr>
<tr>
<td>Colon cell lines</td>
<td>45 (63)</td>
<td>3,512 (39,285)</td>
<td>1.27</td>
<td>17.1/28.0</td>
</tr>
<tr>
<td>Colon tissue</td>
<td>13 (15)</td>
<td>1,985 (16,068)</td>
<td>0.65</td>
<td>19.8/23.7</td>
</tr>
<tr>
<td>Normal ovary</td>
<td>4 (4)</td>
<td>1,389 (6,224)</td>
<td>0.29</td>
<td>19.5/12.7</td>
</tr>
<tr>
<td>Serous ovary</td>
<td>6 (7)</td>
<td>1,935 (8,372)</td>
<td>0.31</td>
<td>17.7/11.8</td>
</tr>
<tr>
<td>Serous fallopian tube</td>
<td>3 (3)</td>
<td>1,379 (5,090)</td>
<td>0.22</td>
<td>8.8/11.0</td>
</tr>
<tr>
<td>Endometrioid ovary</td>
<td>8 (8)</td>
<td>1,451 (5,762)</td>
<td>0.55</td>
<td>13.0/11.7</td>
</tr>
<tr>
<td>Normal endometrium</td>
<td>3 (3)</td>
<td>1,240 (4,649)</td>
<td>0.24</td>
<td>9.3/10.1</td>
</tr>
<tr>
<td>&lt;10kDa HeLa</td>
<td>14 (18)</td>
<td>44 (109)</td>
<td>24.14</td>
<td>25.4/35.3</td>
</tr>
<tr>
<td>PetideAtlas accession numbers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colony tissue</td>
<td>40 (60)</td>
<td>2,373 (16,987)</td>
<td>1.66</td>
<td>19.3/18.4</td>
</tr>
<tr>
<td>Cerebrospinal fluid</td>
<td>16 (22)</td>
<td>266 (1,963)</td>
<td>5.67</td>
<td>18.4/19.0</td>
</tr>
<tr>
<td>Urine</td>
<td>47 (50)</td>
<td>754 (2,898)</td>
<td>5.87</td>
<td>24.9/15.5</td>
</tr>
<tr>
<td>Plasma</td>
<td>90 (96)</td>
<td>70 (92)</td>
<td>56.25</td>
<td>34.7/4.9</td>
</tr>
<tr>
<td>Serum A</td>
<td>311 (326)</td>
<td>192 (351)</td>
<td>61.83</td>
<td>23.4/3.5</td>
</tr>
<tr>
<td>Serum B</td>
<td>269 (293)</td>
<td>339 (847)</td>
<td>44.24</td>
<td>28.4/9.7</td>
</tr>
<tr>
<td>Serum C</td>
<td>158 (160)</td>
<td>279 (977)</td>
<td>36.16</td>
<td>43.1/8.0</td>
</tr>
<tr>
<td>Serum D</td>
<td>230 (248)</td>
<td>190 (365)</td>
<td>54.76</td>
<td>N/A</td>
</tr>
<tr>
<td>Serum total</td>
<td>928 (1,002)</td>
<td>754 (2,066)</td>
<td>55.17</td>
<td>N/A</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1,259 (1,525)</td>
<td>7,341 (85,311)</td>
<td>14.64</td>
<td>N/A</td>
</tr>
</tbody>
</table>

*Total identified proteins = alternative + reference proteins.

**PeptideAtlas accession numbers associated with lung, cerebrospinal fluid, urine, plasma and serum samples are indicated in the Methods section. The Serum total line indicates the sum of distinct proteins detected in Serum A, B, C and D.

doi:10.1371/journal.pone.0070698.t001
Immunofluorescence and confocal microscopy

Immunofluorescence and confocal microscopy analyses were carried out as previously described [24,25].

LC-MS/MS analyses of HeLa cell lysates, colon cell lines and colon tissues

LC-MS/MS analyses of HeLa cell lysates, colon cell lines (CCL227, CCL228, CCL233, CRL1459 and HCT116 cells) and colon tissues were performed as previously described [22].

LC-MS/MS analyses of cancerous ovarian, normal ovarian, cancerous fallopian tube and normal endometrial tissues

Formalin fixed, paraffin-embedded tissues were obtained from the CHRU de Lille pathology department (institutional review approval CPP Nord Ouest IV 12/10). For the histological imaging prior to LC-MS/MS analysis, 4-μm-thick tissue sections were cut from the formalin-fixed, paraffin-embedded (FFPE) whole-mount ovarian tissue blocks. The sections were placed on ITO-coated slides and heated for 60 min at 58°C [26]. The tissue was counterstained with hematoxylin, cosin and safran (HES), dehydrated using graded ethanol solutions, and air-dried for histological examination by our staff pathologist. The tissues appeared heterogeneous and contained cancerous, hyperplastic, and normal regions, with stromal tissue in each region [28]. The International Federation of Gynecology and Obstetrics (FIGO) stages were determined.

For tissue de-waxing, sections (10 μm) were generated using a microtome and were applied to conductive glass slides that were coated with ITO (indium tin oxide) on one side. The paraffin was removed by submersion in toluene twice for 5 min, followed by a light rehydration in ethanol baths (100%, 96%; 70% and 30%) before the slides were dried in a desiccator at room temperature [26,27].

Citric acid antigen retrieval was performed by immersing the slides in 10 mM of citric acid for 20 min at 90°C and then drying them in a desiccator for 10 min. Prior to the enzymatic digestion, the slides were incubated in 10 mM NH4HCO3 twice to remove the remaining antigen retrieval solution and to condition the tissue for effective enzyme activity.

Ten milliliters of a solution of 40 mM trypsin in 50 mM ammonium bicarbonate was dropped onto each tissue region of interest. The slides were then incubated for 4 hours at 37°C in a customized humidity chamber (a 10 cm × 15 cm box filled with water to one quarter of the box height and placed in a 37°C incubator). After trypsin digestion, 10 μL of a 10 mg/mL HCCA solution in aqueous TFA 0.1%/ACN (3:7) was dropped onto each section [28,29].

Tryptsin-digested peptides were manually extracted from specific tissue regions. Using a micropipette, specific regions were subjected to 20 successive washes with 100 μL of a solution of 80% ACN in water. The extract solution was then submitted to freeze-dried with a SpeedVac (Savant). The dried peptides were then re-dissolved in 10 μL of 0.1% TFA. Salts were removed from the solutions, and peptides were concentrated using a solid-phase extraction procedure with the Millipore ZipTip device with a final 10 μL 80% ACN elution solution. The solution was then dried again using the SpeedVac. Dried samples were solubilized in water/5% acetonitrile/0.1% formic acid. Samples were separated by online reversed-phase chromatography using a Thermo Scientific Proxeon Easy-nLC system equipped with a Proxcen trap column (100 μm ID × 2 cm, Thermo Scientific) and a C18 packed-tip column (100 μm ID ×15 cm, NikkyoTechnos Co. Ltd). Peptides were separated using a 5%-40% gradient of acetonitrile over 110 minutes at a flow rate of 300 nL/min. The
LC-MS/MS analyses of lung tissue, cerebrospinal fluid, urine, plasma and serum

Raw data were obtained from the PeptideAtlas online repository [30] under the following accession numbers: lung (PAe000171), cerebrospinal fluid (PAe001777), urine (PAe000761 and PAe000763), plasma (PAe000846), serum (A: PAe000135; B: PAe000347; C: PAe000281; D: PAe000331, PAe000332, PAe000333, PAe000334, PAe000335, PAe000337, PAe000338).

Bioinformatics analysis

Quantitation was performed using the program MaxQuant version 1.2.2.5 [31,32]. The derived peak list generated by Quant.exe (the first part of MaxQuant) was searched against Andromeda as the database search engine for peptide identifications using the human GenBank protein entries (release 37) containing 37,390 proteins, to which the 83,886 predicted alternative proteins and 175 commonly observed contaminants and all the reversed sequences had been added. The first search mass tolerance was set to 20 p.p.m. and main search mass tolerance was 6 p.p.m. Enzyme was set to trypsin/p with 2 missed cleavages. Carbamidomethylation of cysteine was searched as a fixed modification, whereas N-acetylation of methionine were searched as variable modifications. For the serum sample PAe000347, asparagine deamidation was added as a fixed modification, and glutamine deamidation as a variable modification. For analysis of fractionated HeLa cells, colon cell lines, colon tissue, <10 kDa HeLa cells proteins, and paraffin-embedded human tissues, identification was set to a false discovery rate of 5%, determined by the use of a reverse database (similar results were obtained by using a randomly generated database). To achieve reliable identifications, only the proteins associated with a PEP value of less than 0.05, and identified with at least one unique peptide were retained. Indeed, the detection of more than one tryptic fragment for small proteins is unlikely [19], though we detected at least two tryptic fragments for 148 alternative proteins (Table S4). For analysis of lung tissue and fluids, identification was set to a false discovery rate of 1%, with at least one unique peptide. Protein isoforms and proteins that cannot be distinguished based on the peptides identified are grouped and displayed on a single line, but with a single GenBank accession number (Tables S1–S3).

Accession numbers for detected alternative proteins are provided in Table S4 and are publicly available on the nucleotide database of the European Bioinformatics Institute website (http://www.ebi.ac.uk/).

The conservation analyses of alternative and reference proteins were carried out with BLASTP [33] (version 2.2.27+). The AltORFs databases and reference databases from different species were searched against the corresponding human database with an expectation (E) value cutoff of \( \leq 10^{-5} \). Only the best matching hit for each predicted protein was considered. The reference protein databases match the GenBank and Ensembl releases used to generate the AltORFs databases.

Results

To generate our database, we defined AltORFs as ORFs located in a non-canonical reading frame of the RefORF, in the 5’ and 3’UTR regions of an mRNA, or partially overlapping with both the RefORF and an UTR region (Fig. 1A). We based our prediction algorithm on characteristics of known AltORFs (AUG as TIS, alternative stop codon different from the RefORF stop codon), and added a size cut-off of 40 codons to keep the database.
Alternative Proteins Expand the Human Proteome

Figure 5. The alternative protein AltMRVI1 interacts with BRCA1. (A) Sequence alignment between the third BRCA1 interacting protein 3 (BRCA1 IP3) identified by Liu et al by Y2H [40] and AltMRVI1. (B) Localization of BRCA1 and AltMRVI1EGFP in the nucleus. HeLa cells transfected with AltMRVI1EGFP (green channel) were immunostained with anti-BRCA1 (red channel) antibodies. Nuclei were stained with Hoechst (blue channel). Scale bar, 10 μm.

doi:10.1371/journal.pone.0070698.g005

Alternative Proteins Expand the Human Proteome

A

Table: The expression of alternative proteins in different tissues

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Alternative Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plasma</td>
<td>16</td>
</tr>
<tr>
<td>Serum</td>
<td>47</td>
</tr>
<tr>
<td>Fetal fallopian tube</td>
<td>90</td>
</tr>
<tr>
<td>Normal ovary</td>
<td>928</td>
</tr>
<tr>
<td>Endometrioid ovary</td>
<td>1699</td>
</tr>
<tr>
<td>Serous ovary</td>
<td>192</td>
</tr>
<tr>
<td>Normal lung</td>
<td>13</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>40</td>
</tr>
</tbody>
</table>

B

Figure 5A: Alignment of BRCA1 IP3 and AltMRVI1. The alignment reveals conservation in the region of the protein that interacts with BRCA1.

C

Figure 5B: Localization of BRCA1 and AltMRVI1 in the nucleus. HeLa cells expressing AltMRVI1EGFP (green channel) were immunostained with anti-BRCA1 (red channel) antibodies. Nuclei were stained with Hoechst (blue channel).

D

Figure 5C: Detection of AltMRVI1 in lysates from HeLa cells expressing AltMRVI1EGFP. Molecular weight markers in kDa are indicated on the right.

Figure 5D: Detection of BRCA1 and AltMRVI1 in the nucleus. HeLa cells transfected with AltMRVI1EGFP (green channel) were immunostained with anti-BRCA1 (red channel) antibodies. Nuclei were stained with Hoechst (blue channel).

Figure 5E: Detection of AltMRVI1 in lysates from HeLa cells expressing AltMRVI1EGFP. Molecular weight markers in kDa are indicated on the right.
alternative proteins represent approximately 55% of the proteins identified in plasma and serum (Table 1). Overall, we detected a total of 1,259 alternative proteins (Table 1), and 47 were expressed in different cell lines and/or tissues (Table S4).

In accordance with the scanning model of translation initiation, we used the first AUG rule in order to predict the TIS of AltORFs present in our database. Since other non-AUG codons can be used as TIS [34], we tested the reliability of our TIS prediction for the alternative proteins previously detected by two independent methods. First, the detection of N-acetylated peptides, a modification specific to protein N-termini [38], in 889 out of the 1,259 total alternative proteins detected throughout our different LC-MS/MS experiments allowed us to determine that in most cases (866/889), the alternative TIS predicted in our database was correct (Table S4). Second, we randomly selected and tested the co-expression of 6 alternative proteins and their corresponding reference proteins from the 129 alternative proteins detected by LC-MS/MS in the fractionated HeLa cells lysate. A strategy based on the transfection of constructs with two tags, an HA tag in frame with the reference protein and a GFP tag in frame with the alternative protein, was used to report the co-expression of both proteins in transfected cells (Fig. 3). The corresponding alternative proteins were all detected by both western blot and GFP fluorescence. Importantly, inactivating mutations (AUG to AAG) on the transfection of constructs with two tags, an HA tag in frame with the reference protein and a GFP tag in frame with the alternative protein, was used to report the co-expression of both proteins in transfected cells (Fig. 3). The corresponding alternative proteins were all detected by both western blot and GFP fluorescence. Importantly, inactivating mutations (AUG to AAG) on the transfection of constructs with two tags, an HA tag in frame with the reference protein and a GFP tag in frame with the alternative protein, was used to report the co-expression of both proteins in transfected cells (Fig. 3). The corresponding alternative proteins were all detected by both western blot and GFP fluorescence. Importantly, inactivating mutations (AUG to AAG) on the transfection of constructs with two tags, an HA tag in frame with the reference protein and a GFP tag in frame with the alternative protein, was used to report the co-expression of both proteins in transfected cells (Fig. 3). The corresponding alternative proteins were all detected by both western blot and fluorescence (Fig. 4C). Diverse subcellular distributions could be observed among the tested constructs (Fig. 4C, see also Fig. 3), suggesting a variety of possible functions associated with alternative proteins.

Many cDNA clones identified in large scale screening assays, including yeast two-hybrid (Y2H) studies do not match any known protein of the conventional proteome because they represent out-of-frame clones [39,40]. In Y2H, these unknown interacting proteins are usually rejected as false positive hits; yet, we reasoned that a proportion of such clones could represent alternative proteins with real affinity for the bait. We found in the literature the partial sequence of five out-of-frame clones from a Y2H experiment performed with the tandem BRCT domain of breast cancer susceptibility protein 1 (BRCA1) [40]. One sequence was 100% identical to an alternative protein from our database whose AltORF is located in the 3'UTR of the mRNA produced from the MRVI1 gene (Fig. 5A). AltMRVI1EGFP was cloned and transfected into HeLa cells. Similar to BRCA1, AltMRVI1EGFP localized to the nucleus (Fig. 5B). We confirmed the interaction between BRCA1 and AltMRVI1EGFP by co-immunoprecipitation (Fig. 5C). Thus, AltMRVI1 is possibly a novel BRCA1 interacting protein that was already identified by Y2H, but mistakenly rejected as a false positive hit.

To assess the evolutionary conservation of predicted alternative proteins, we generated a database of AltORF's present in mature mRNAs from different eukaryote species (Databases S2–S8), and predicted 5019 distinct alternative proteins in Saccharomyces cerevisiae, 35,532 in Caenorhabditis elegans, 38,248 in Drosophila melanogaster, 52,454 in Xenopus tropicalis, 82,305 in Mus musculus, 57,492 in Bos taurus, and 79,874 in Pan troglodytes (Fig. S2; Table S5). The median AltORF size of these predicted proteins ranges from 50 to 57 codons, showing that as observed for humans, the alternative proteome is composed of small proteins compared to the reference proteome (Table S5). Although the homology between reference proteins is greater than that of alternative proteins, we also identified thousands of human alternative proteins conserved with predicted AltORFs in vertebrates, hundreds in invertebrates, and were even surprised to find...
that 13 alternative proteins are conserved between human and yeast with a median sequence identity of 47.8% (Fig. 6).

Discussion

The potential of eukaryotic genomes for encoding alternative proteins from non-canonical open reading frames is well known and was recently featured in ribosome profiling studies [34,41,42]. Yet, a large scale approach allowing the detection of the corresponding alternative proteins was lacking. In this study, we have generated a database of predicted alternative protein-coding ORFs with a minimum length of 40 codons present in human mRNAs. The data presented here indicate an average of 3.8 AltORFs per human mRNA with a median length of 57 amino acids. Using this database, 1259 human alternative proteins were detected by mass spectrometry in the present study, 3 of which (accession numbers HF548059, HF547970, HF548029) were previously detected [19,43,44]. This result strongly supports the hypothesis that the complexity of the proteome has been underestimated and alternative translation initiation already well characterized in viruses cannot be ignored in humans. Importantly, evolutionary conservation of alternative proteins between vertebrates and invertebrates implies that these proteins have significant biological functions.

It is very likely that similar to proteins translated from canonical ORFs, alternative proteins display a wide variety of biological functions. This is suggested by the great diversity of subcellular localizations that we observed in our fluorescence microscopy experiments (Fig. 3, 4C, 5B), and by the fact that a growing number of important functions are attributed to small proteins and peptides [45–48]. The polycistronic nature of AltORF encoding mRNAs can potentially lead to intriguing functional interplays between the reference and alternative proteins, such as direct interaction between the reference and alternative proteins [11,49]. There is also evidence that an upstream ORF not only regulates the expression of a downstream RefORF by interfering with canonical AUG recognition by scanning ribosomes, but also reduces the translational efficiency of the RefORF in trans [45,50]. AltORFs translation products could also be of particular importance during the thymal selection of T lymphocytes, serving as “cryptic T-cell epitopes”. In some cases, this has been shown to lead to the selection of lymphocytes with antiviral or antitumor activities [51].

Our databases of AltORFs will be useful to identify genes containing multiple protein coding ORFs and to unravel their functions. This is particularly important in experimental settings where gene expression studies (cDNA transfection, knock-down, transgenes, and gene therapy) could result in the expression or down regulation of a reference protein and an unnoticed alternative protein, leading to confounding results [12]. Another striking example is the co-expression of therapeutic transgenes and their associated alternative proteins, which elicit a cytotoxic T lymphocyte response [52]. Thus, transgene sequences should be carefully examined for possible AltORFs to decrease potential adverse immune responses during therapeutic gene transfer.

We also propose that proteins recalcitrant to mass spectrometry identification or proteins with no sequence homology with the conventional proteome identified in large-scale cDNAs screens should be revisited with the AltORFs databases. Additionally, large fundamental and clinical proteomic studies using organs and tissues would likely benefit from the AltORFs database to achieve complete catalogs of proteins in different tissues [33,54].

The presence of a large fraction of alternative proteins in plasma and serum is particularly interesting as there is a constant need for biomarkers to identify a variety of disorders at an early stage [55–57]. The reason why so many alternative proteins are secreted is currently unknown. We did not find any enrichment for classical export signal peptides (not shown), and their secretion mechanism remains to be investigated.

As for any databases, the AltORFs database has some limitations. Although AUG remains the main translation initiation site, recent ribosome profiling studies clearly indicate the use of non-AUG start sites [34]. Yet, we did not take into account non-AUG initiation sites as an accurate prediction method for such functional translation start sites is not yet available [50]. Although there is strong evidence that short peptides are also translated [19], we introduced a cut-off of 40 amino acids to discard from our databases polypeptides shorter than 40 residues, which are less readily detected by conventional mass spectrometry approaches, and to keep the database to a reasonable size. Finally, we used the NCBI reference sequence database (RefSeq) as a source for RNA transcripts. This non-redundant and well-annotated database is fairly conservative, and thus is a quality source for identifying candidate AltORFs, but it would be interesting to compare with other databases (e.g. Gencode) to verify if different mRNA isoforms could also serve as template for the expression of corresponding alternative proteins. Nevertheless, the location of AltORFs is comparable with the distribution obtained in a peptidomic study of small ORFs encoded polypeptides, with the exception of AltORFs located in 5’UTRs [19]. For these AltORFs, the apparent discrepancy probably results from our prediction of AltORFs initiating at AUG sites only [35].

In conclusion, we have provided compelling evidence that alternative proteins significantly contribute to the human proteome by identifying 1259 new proteins and many more will likely be detected in further MS experiments. A comprehensive knowledge of the proteome is of crucial interest to unravel the cellular mechanisms underlying health and disease. We believe that proteomics approaches supported by ribosome profiling will further benefit the establishment of an exhaustive catalog of proteins to fulfill this goal in the future.

Supporting Information

Figure S1 Hematoxylin, eosin and saffron-stained sections of normal and cancerous tissues in two patients. Patient 1, sections of normal and serous ovarian tissues and normal and serous cancerous fallopian tissue. Patient 2, sections of normal, borderline and endometrioid ovarian tissues and normal and hyperplastic endometrial tissues. Annotations of the tissues were performed by a pathologist (Dr. O. Kerdraon, Centre Oscar Lambret, Lille, France).

(TIF)

Figure S2 AltORFs distribution among eukaryote species. Distribution in % across the different mRNA regions. The number of distinct proteins predicted for each species is displayed in the insert.

(TIF)

Table S1 Alternative and reference proteins list in diverse cell lines. For a given alternative protein, N-terminal N-acetylated peptides and sequence coverage (%) are indicated in additional columns. When co-expression of the reference and alternative proteins is observed for a particular gene, the lane is highlighted in gray. The sequence covered by all detected peptides is underlined in the alternative protein amino acid sequence column. For the HeLa cell line whole-protein analysis, a total of 129 alternative proteins identified by 280 peptides, and 5,558
reference proteins identified by 68,035 peptides were detected. For the HeLa cell line analysis of proteins between the 4.6 and 10 kDa marker of a 1-D SDS-PAGE, a total of 14 alternative proteins identified by 18 peptides, and 44 reference proteins identified by 109 peptides were detected. We excluded from the identified proteins those with expected molecular weights above 10 kDa since fragments below 10 kDa likely represent breakdown products. For the colon cell lines, a total of 45 alternative proteins identified by 63 peptides, and 3,512 reference proteins identified by 39,265 peptides were detected.

Table S2 Alternative and reference proteins list in human tissues. For a given alternative protein, N-terminal N-acetylated peptides and sequence coverage (%) are indicated in additional columns. The sequence covered by all detected peptides is underlined in the alternative protein amino acid sequence column. For the colon tissue, a total of 13 alternative proteins identified by 17 peptides, and 1,985 reference proteins identified by 16,068 peptides were detected. For the lung tissue, a total of 40 alternative proteins and 2,373 reference proteins were detected. For the normal endometrium, endometrioid ovary, serous ovary, normal ovary, and serous fallopian tube, a total of 19 alternative proteins and 2,748 reference proteins were detected.

Table S3 Alternative and reference proteins list in human fluid. For a given alternative protein, N-terminal N-acetylated peptides and sequence coverage (%) are indicated in additional columns. The sequence covered by all detected peptides is underlined in the alternative protein amino acid sequence column. In cerebrospinal fluid, a total of 16 alternative proteins and 266 reference proteins were detected. In urine, a total of 47 alternative proteins and 754 reference proteins were detected. In plasma, a total of 90 alternative proteins and 70 reference proteins were detected. In serum, a total of 928 alternative proteins and 754 reference proteins were detected.

Table S4 Combined alternative proteins list. All alternative proteins identified across all analysed samples are shown. For a given alternative protein, if N-terminal N-acetylated peptides were detected, they are indicated in an additional column and the entire lane is highlighted in gray. The sequence covered by all detected peptides is underlined in the alternative protein amino acid sequence column.

Table S5 Summary of AltORFs and alternative proteins characteristics in different eukaryote species.

Database S1 A database of alternative ORFs in Homo sapiens. For each AltORF, the gene name and accession number of the mRNA in which it is encoded are provided. Other information can also be found for both the reference ORF and the alternative ORF, including the reading frame and the coordinates of the start and stop codon (with respect to the first nucleotide of the mRNA). The predicted amino acid sequence of the alternative protein is also indicated.

Database S2 A database of alternative ORFs in Pan troglodytes. For each AltORF, the gene name and accession number of the mRNA in which it is encoded is provided in the first tab. Other information can also be found for both the reference ORF and the alternative ORF, including the reading frame and the coordinates of the start and stop codon (with respect to the first nucleotide of the mRNA). The predicted amino acid sequence of the alternative protein is also indicated.

Database S3 A database of alternative ORFs in Mus musculus. For each AltORF, the gene name and accession number of the mRNA in which it is encoded is provided in the first tab. Other information can also be found for both the reference ORF and the alternative ORF, including the reading frame and the coordinates of the start and stop codon (with respect to the first nucleotide of the mRNA). The predicted amino acid sequence of the alternative protein is also indicated. The second tab displays the results of alternative protein conservation analysis against human. The gene name and accession number for chimpanzee and human homolog proteins are provided, as well as standard BLASTP output, and similarity percentage.

Database S4 A database of alternative ORFs in Bos taurus. For each AltORF, the gene name and accession number of the mRNA in which it is encoded is provided in the first tab. Other information can also be found for both the reference ORF and the alternative ORF, including the reading frame and the coordinates of the start and stop codon (with respect to the first nucleotide of the mRNA). The predicted amino acid sequence of the alternative protein is also indicated. The second tab displays the results of alternative protein conservation analysis against human. The gene name and accession number for cow and human homolog proteins are provided, as well as standard BLASTP output, and similarity percentage.

Database S5 A database of alternative ORFs in Xenopus tropicalis. For each AltORF, the gene name and accession number of the mRNA in which it is encoded is provided in the first tab. Other information can also be found for both the reference ORF and the alternative ORF, including the reading frame and the coordinates of the start and stop codon (with respect to the first nucleotide of the mRNA). The predicted amino acid sequence of the alternative protein is also indicated. The second tab displays the results of alternative protein conservation analysis against human. The gene name and accession number for frog and human homolog proteins are provided, as well as standard BLASTP output, and similarity percentage.

Database S6 A database of alternative ORFs in Drosophila melanogaster. For each AltORF, the gene name and accession number of the mRNA in which it is encoded is provided in the first tab. Other information can also be found for both the reference ORF and the alternative ORF, including the reading frame and the coordinates of the start and stop codon (with respect to the first nucleotide of the mRNA). The predicted amino acid sequence of the alternative protein is also indicated. The second tab displays the results of alternative protein conservation analysis against human. The gene name and accession number for fly and human homolog proteins are provided, as well as standard BLASTP output, and similarity percentage.

Database S7 A database of alternative ORFs in Caenorhabditis elegans. For each AltORF, the gene name and
accession number of the mRNA in which it is encoded is provided in the first tab. Other information can also be found for both the reference ORF and the alternative ORF, including the reading frame and the coordinates of the start and stop codon (with respect to the first nucleotide of the mRNA). The predicted amino acid sequence of the protein is also indicated. The second tab displays the results of alternative protein conservation analysis against human. The gene name and accession number for nematode and human homolog proteins are provided, as well as standard BLASTP output, and similarity percentage. (XLSX)

**Database S8** A database of alternative ORFs in *Saccharomyces cerevisiae*. For each AltORF, the gene name and accession number of the mRNA in which it is encoded is provided in the first tab. Other information can also be found for both the reference ORF and the alternative ORF, including the reading frame and the coordinates of the start and stop codon (with respect to the first nucleotide of the mRNA). The predicted amino acid sequence of the alternative protein is also indicated. The second tab displays the results of alternative protein conservation analysis against human. The gene name and accession number for yeast and human homolog proteins are provided, as well as standard BLASTP output, and similarity percentage. (XLSX)

**Acknowledgments**

We thank Dr Michelle Scott (Université de Sherbrooke), Francis Gaudreault (Université de Sherbrooke), Dr Paul Harrison (McGill University) and members of the Roucou lab for helpful comments. The plasmid containing the SR5SF1 coding sequence was a kind gift from Dr Benoit Chabot (Université de Sherbrooke).

**Author Contributions**

Conceived and designed the experiments: XR BV. Performed the experiments: BV JM GT SV MW. Analyzed the data: XR BV CB FMB MS. Contributed reagents/materials/analysis tools: BV JFL CB FMB MS MW SV JM GT. Wrote the paper: XR BV CB MS. Responsible for the project: XR. Contributed to computational analyses: JFL BV CB. Generated the databases: JFL. Performed the conservation analyses: CB.

**References**

15. Ronsin C, Chung-Scott V, Poullion I, Aknouche N, Gaudin C, et al. (1999) A novel human gene encoding a 150 kDa transmembrane protein is also indicated. The second tab displays the results of alternative protein conservation analysis against human. The gene name and accession number for yeast and human homolog proteins are provided, as well as standard BLASTP output, and similarity percentage. (XLSX)