Non-coding after all: Biases in proteomics data do not explain observed absence of lncRNA translation products

Kenneth Verheggen, Pieter-Jan Volders, Pieter Mestdagh, Gerben Menschaert, Petra Van Damme, Kris Gevaert, Lennart Martens, and Jo Vandesompele

J. Proteome Res., Just Accepted Manuscript • Publication Date (Web): 23 May 2017
Downloaded from http://pubs.acs.org on May 24, 2017

"Just Accepted" manuscripts have been peer-reviewed and accepted for publication. They are posted online prior to technical editing, formatting for publication and author proofing. The American Chemical Society provides "Just Accepted" as a free service to the research community to expedite the dissemination of scientific material as soon as possible after acceptance. "Just Accepted" manuscripts appear in full in PDF format accompanied by an HTML abstract. "Just Accepted" manuscripts have been fully peer reviewed, but should not be considered the official version of record. They are accessible to all readers and citable by the Digital Object Identifier (DOI®). "Just Accepted" is an optional service offered to authors. Therefore, the "Just Accepted" Web site may not include all articles that will be published in the journal. After a manuscript is technically edited and formatted, it will be removed from the "Just Accepted" Web site and published as an ASAP article. Note that technical editing may introduce minor changes to the manuscript text and/or graphics which could affect content, and all legal disclaimers and ethical guidelines that apply to the journal pertain. ACS cannot be held responsible for errors or consequences arising from the use of information contained in these "Just Accepted" manuscripts.
Non-coding after all: Biases in proteomics data do not explain observed absence of IncRNA translation products

Kenneth Verheggen\textsuperscript{‡,1,2,6}, Pieter-Jan Volders\textsuperscript{‡,1,2,3,5,6}, Pieter Mestdagh\textsuperscript{3,5,6}, Gerben Menschaert\textsuperscript{4}, Petra Van Damme\textsuperscript{1,2}, Kris Gevaert\textsuperscript{1,2}, Lennart Martens\textsuperscript{1,2,6,*}, Jo Vandesompele\textsuperscript{3,5,6}

\textsuperscript{1} Medical Biotechnology Center, VIB, Ghent 9000, Belgium
\textsuperscript{2} Department of Biochemistry, Ghent University, Ghent 9000 Belgium
\textsuperscript{3} Center for Medical Genetics, Ghent University, Ghent 9000, Belgium
\textsuperscript{4} Department of Mathematical Modeling, Statistics and Bioinformatics, Ghent University, Ghent 9000, Belgium
\textsuperscript{5} Cancer Research Institute Ghent (CRIG), Ghent University, Ghent 9000, Belgium
\textsuperscript{6} Bioinformatics Institute Ghent (BIG N2N), Ghent University, Ghent 9000, Belgium

\‡ equally contributing authors

* Corresponding author: Prof. Dr. Lennart Martens, A. Baertsoenkaai 3, B-9000 Gent, Belgium, lennart.martens@vib-ugent.be, tel: +32 9 264 93 58, fax: +32 9 264 94 84

Keywords: public data, IncRNA, mass spectrometry, msms, micropeptide, translation
ABSTRACT

Over the past decade, long non-coding RNAs (lncRNAs) have emerged as novel functional entities of the eukaryotic genome. However, the scientific community remains divided over the amount of true non-coding transcripts among the large number of unannotated transcripts identified by recent large scale and deep RNA-sequencing efforts. Here, we systematically exclude possible technical reasons underlying the absence of lncRNA-encoded proteins in mass spectrometry datasets, strongly suggesting that the large majority of lncRNAs is indeed not translated.
INTRODUCTION

Advances in sequencing technologies have uncovered pervasive transcription of the eukaryotic genome outside of annotated protein-coding loci. Most of these novel transcripts are long (> 200 nucleotides), lack large open reading frames (ORFs) and homology to annotated protein-coding genes\(^1\). Termed long non-coding RNAs (lncRNAs), these transcripts comprise a vast, diverse and largely unexplored class of RNA, outnumbering any other class of genetic entities in the human genome\(^2\). Those that have been studied in detail play important roles in a wide range of cellular processes during normal development and in homeostasis and disease, including cancer\(^3\).

Similar to lncRNAs, short open reading frame (sORF)-encoded polypeptides (SEPs) or micropeptides have gained increased attention over the past few years. While classical bioactive peptides are enzymatically cleaved from longer protein precursors, micropeptides are small peptides (< 100 amino acids) directly translated from single sORFs. So far, only a limited number of these micropeptides have been discovered and functionally characterized\(^4\).

The coding potential of newly discovered RNA transcripts is typically assessed by means of prediction algorithms\(^5-7\). While each algorithm has its own strengths and weaknesses, they are all biased to current annotations and may thus be unsuitable for the detection of small or non-conserved proteins including micropeptides.

Although the advent of ribosome profiling\(^8\) (sequencing of ribosome protected RNA fragments) promised to provide evidence for (the lack of) translation of expressed ORFs, much is still open to interpretation. Numerous studies report substantial ribosome occupancy of lncRNA transcripts\(^9-12\). The striking similarities in the pattern and size of ribosome protected fragments covering protein-coding
transcripts and lncRNAs have led some researchers to conclude that up to 90% of the lncRNA transcriptome bears coding ORFs. Other researchers report much more conservative numbers. For instance, if the relative abundance of ribosomes before and after stop codons (termed ribosome release) is used to discriminate between protein-coding and non-coding transcripts, only a few novel coding ORFs are found. When taking into account the phased movement of ribosomes across translated ORFs, only a small number of novel peptides arising from transcripts annotated as lncRNAs are identified. Different research groups have thus developed different metrics and methodologies to detect coding ORFs in ribosome profiling data. Without a consensus, the true coding potential of lncRNA transcripts remains open to speculation.

Mass spectrometry is often considered as the gold standard in detection and characterization of proteins or peptides. So far, few studies have turned to mass spectrometry to study micropeptides and lncRNA-encoded proteins. In our previous work, we have reprocessed large quantities of tandem mass spectrometry data obtained from the PRoteomics IDEntifications (PRIDE) database. In brief, we reanalyzed raw data from 2,493 PRIDE experiments, containing 39,463,035 fragmentation mass spectra covering 68 human tissues using a combinatorial database consisting of Uniprot protein sequences and six reading frame translated LNCipedia lncRNAs. In these searches, less than 1% of the lncRNA genes in LNCipedia were covered by at least two unique peptide to spectrum matches (PSMs), compared to approximately 87% of Uniprot proteins (Volders et al., 2015; Supplemental Tables S1 and S2). The results of these searches are publicly available through the LNCipedia portal.
Other groups have reported numbers similar numbers, ranging from less than 100 up to 1,600 putative lncRNA-encoded proteins in human\textsuperscript{16-18}. Compared to the more than 60,000 reported lncRNA genes\textsuperscript{2,19}, these numbers are fairly low and definitely much lower than those reported by various ribosome profiling studies.

This discrepancy in the reported amounts of potentially coding lncRNAs is the source of spirited discussion in the field. Indeed, a resolution of this conflict has direct relevance for further investigations into the biological roles of lncRNAs.

The most direct observation of coding lncRNAs is the actual detection by mass spectrometry based proteomics of the encoded proteins. As such, the absence of large amounts of detected lncRNA-derived proteins strongly hints at a limited coding potential for lncRNAs. The main criticism of this approach however, is that mass spectrometry-based proteomics is somehow biased against the detection of lncRNA products.

Here, we therefore examine the possible biases of mass spectrometry to detect and characterize lncRNA-encoded proteins based on a detailed yet exhaustive reprocessing of very large amounts of public proteomics data. Our findings clearly show that there are no obvious technical reasons why mass spectrometry would have largely missed (micro)peptides originating from non-coding RNA transcripts, thus eliminating the possibility that mass spectrometry would be biased against the detection of putative lncRNA-encoded proteins.

THE INFLUENCE OF PROTEIN COMPOSITION ON DETECTABILITY BY MASS SPECTROMETRY

Mass spectrometry enables high-throughput protein identification in complex samples. However, there is some controversy regarding the limitations of this
technique in terms of detectability of peptides and thus, by extension, proteins. Several potential causes have been proposed, including biases due to the size of the protein sequence, the amino acid composition, the abundance, and the half-life of proteins\textsuperscript{20-22}. Here, we investigate these presumed issues and identify potential reasons as to why certain predicted ORF products evade detection. The applied strategy revolves around the reprocessing of publicly available data in PRIDE\textsuperscript{23}, one of the world’s leading mass spectrometry repositories\textsuperscript{24}. Sequence database searches were performed using an automated reprocessing pipeline, consisting of pride-asap\textsuperscript{25} for the detection of data set specific parameters, SearchGUI\textsuperscript{26} to match the fragmentation mass spectra against peptides derived from protein sequence databases, and PeptideShaker\textsuperscript{27} to integrate the identifications and control these at a 1\% false discovery rate at the peptide-to-spectrum match level.

Because the combination of known canonical human protein sequences and hypothetical lncRNA derived sequences can hamper protein inference, the overlap between both datasets must be investigated. This was achieved by matching the full set of tryptic peptides originating from the six reading frame translated LNCipedia database (version 3.1) against the full set of human canonical proteins in UniProtKB/SwissProt\textsuperscript{28}. Out of 8,645,916 hypothetical tryptic peptide sequences, only 277,412 had one or more identical matches in the protein sequence dataset. The overlap is thus minimal (approximately 3.21\%), which is about equal to the between-protein tryptic peptide overlap for the human complement of UniProtKB/Swiss-Prot, which does not consider any splice isoforms. This indicates that uniquely identifying lncRNA polypeptides should be no more difficult than identifying unique human proteins.
A first potential factor that may contribute to a detection bias is the size of a protein. In order to analyse this, publicly available submissions of human projects to PRIDE were searched against the human complement of the UniProtKB/SwissProt protein sequence database using our reprocessing pipeline. The resulting set of proteins was ranked according to sequence length. A simple spectral count over all PRIDE assays in which a protein was identified, was used to indicate the number of times the protein was observed. Q8WZ42, the megadalton protein titin, represented by its canonical isoform of 34,350 residues, was identified 298 times in 183 assays. This indicates that large proteins are picked up despite their length, as is to be expected due to the relatively higher number of potential MS/MS-identifiable peptides following enzymatic cleavage of larger proteins. At the same time, short proteins are also frequently identified across a broad range of assays (Table 1). It is noteworthy that out of 20,207 human entries in UniProtKB/SwissProt, only 36 –(mainly) tissue or cell specific– proteins (0.18%) are smaller than the shortest reported protein sequences in Table 1. These numbers provide a strong indication that protein length is not likely a major determining factor in protein detectability by mass spectrometry using standard sampling protocols.

<table>
<thead>
<tr>
<th>protein</th>
<th>gene name</th>
<th>length (AA)</th>
<th>average MW (da)</th>
<th>spectral count</th>
<th>assay count</th>
</tr>
</thead>
<tbody>
<tr>
<td>P62328</td>
<td>TMSB4X</td>
<td>44</td>
<td>4921.46</td>
<td>787</td>
<td>287</td>
</tr>
<tr>
<td>P63313</td>
<td>TMSB10</td>
<td>44</td>
<td>4894.48</td>
<td>366</td>
<td>229</td>
</tr>
<tr>
<td>Q8N4H5</td>
<td>TOMM5</td>
<td>51</td>
<td>6035.31</td>
<td>88</td>
<td>70</td>
</tr>
<tr>
<td>P62891</td>
<td>RPL39</td>
<td>51</td>
<td>6275.49</td>
<td>109</td>
<td>52</td>
</tr>
<tr>
<td>Q59GN2</td>
<td>RPL39P5</td>
<td>51</td>
<td>6322.59</td>
<td>107</td>
<td>51</td>
</tr>
</tbody>
</table>
Table 1. The ten shortest human proteins identified by reprocessing of the reprocessed PRIDE data.

<table>
<thead>
<tr>
<th>Accession</th>
<th>Protein</th>
<th>MW</th>
<th>p</th>
<th>Isoelectric point</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q5VTU8</td>
<td>ATP5E2</td>
<td>51</td>
<td>53</td>
<td>43</td>
</tr>
<tr>
<td>P56381</td>
<td>ATP5E</td>
<td>51</td>
<td>53</td>
<td>43</td>
</tr>
<tr>
<td>Q96IX5</td>
<td>USMG5</td>
<td>58</td>
<td>112</td>
<td>86</td>
</tr>
<tr>
<td>P62861</td>
<td>FAU</td>
<td>59</td>
<td>248</td>
<td>141</td>
</tr>
<tr>
<td>P13640</td>
<td>MT1G</td>
<td>62</td>
<td>71</td>
<td>47</td>
</tr>
</tbody>
</table>

A second feature that could impose a bias on protein detection using mass spectrometry is the amino acid sequence composition. The existence of such a potential bias was investigated by comparing the composition of peptides that have been identified at high confidence with the composition of in silico generated peptide sequences. A theoretical digest of the human UniProtKB/SwissProt database was therefore created using dbtoolkit with tryptic cleavage rules, allowing for two missed cleavages. Both empirical peptides from the reprocessing of the human data in PRIDE and in silico obtained peptide sequences from the in silico digest of UniProtKB/SwissProt were filtered to sizes between 5 and 30 amino acids, which is the common range of observed peptide lengths in practice. The amino acid composition of both theoretical and observed peptides was then calculated by counting the occurrence rate of an amino acid per position in the sequence (Figure 1). There is a high positive correlation between both datasets (Spearman $\rho = 0.952$, $p < 0.01$), hinting that there is no reason to assume that the composition of proteins identified by the reprocessing of PRIDE and those generated by in silico digestion is very different. The higher occurrence rates for R and K in the experimental data are most likely related to the fact that these are the residues that are targeted by the most
common sample preparation procedure, which involves protein digestion by trypsin. This is indeed the confirmed case for the majority of PRIDE projects. In addition, these residues are strong bases and therefore strongly promote ionization. The explanation for the slightly lower occurrence rate of S in the experimental data can be related to the fact that S can be phosphorylated in vivo, and the somewhat lower efficiency in the detection of phosphorylated residues.

Figure 1. Comparison between theoretical ( UniProtKB/SwissProt) and observed (reprocessed PRIDE data) peptide sequence amino acid composition for human data from PRIDE and UniProtKB/SwissProt.

Another important property that can affect detection by mass spectrometry is protein (and thus peptide) abundance in the sample. Although there are examples of successful enrichment protocols, the detection of products of rare translation events is not straightforward. In order to investigate the influence of the abundance on the detectability of proteins by mass spectrometry, we first make use of the study by Anderson and Hunter that reports empirically obtained protein quantification values.
in human blood plasma. Reprocessing of the subset of PRIDE data sets derived from human blood was carried out, and their estimated abundances were mapped to the values reported by Anderson and Hunter (Figure 2). While it is clear that the lowest abundant proteins are not detected, the abundance range of human plasma is quite extreme at eleven orders of magnitude, of which at least eight are covered reliably in the PRIDE data. This analysis thus shows that mass spectrometry based proteomics is only biased against the very least abundant proteins.

**Figure 2.** Reprocessing results for PRIDE data sets derived from human blood plasma mapped onto the abundance values reported by Anderson and Hunter. The size of a bubble corresponds to the number of PRIDE assays in which that protein was identified.
Another possibility for detection bias is provided by the half-life of a protein as rapidly degraded proteins may escape detection as well. In order to assess a possible bias based on protein half-life, we make use of the study by Schwanhäusser et al., where half-life values for murine proteins are reported. Because PRIDE also contains murine data, extensive reprocessing of these murine data sets against the mouse complement of the UniProtKB/SwissProt database was performed and the reprocessed identifications were mapped to the originally reported half-life data (Figure 3). This analysis reveals that the PRIDE data cover the entire half-life range, indicating no influence of protein half-life values on detectability.
Figure 3. Reprocessing results for all PRIDE murine data mapped onto the half-life values reported by Schwanhäusser et al. The size of the bubble corresponds to the number of PRIDE assays in which the protein was identified.

In addition, we calculated the N-terminal instability index of human proteins as described by Guruprasad et al.\textsuperscript{34}. This metric is based on the dipeptide composition of a protein and provides a crude estimation of protein half-life when large-scale experimental data are lacking, as is the case for human proteins. The underlying assumption is that a protein’s half-life correlates negatively to its relative instability. We therefore compared the calculated instability indices for all proteins in the human complement of UniProtKB/SwissProt with those calculated for the identified proteins from the human data sets in PRIDE. Only a minor deviation is revealed between the instability index distributions of observed and theoretical proteins (Figure 4), providing additional proof that the degradation rate of a protein is of little, if any, influence on its detectability.
**Figure 4.** The instability index distributions of human UniProtKB/SwissProt proteins, and of identified proteins from reprocessed human data sets in PRIDE.

**LncRNA Expression and Composition Show No Indication of Coding Potential**

The expression profile of lncRNAs differs extensively from that of protein coding mRNAs (Figure 5a). LncRNAs are generally expressed at a lower level and are more abundant in the nucleus. While mRNAs are transported to the cytoplasm for ribosomal translation, several lncRNAs have a documented function in the nucleus\(^{35}\). As such, the nuclear enrichment of lncRNAs suggests a non-coding role for the majority of the lncRNA transcripts.

We have observed that very low protein abundance can hamper the detection by mass spectrometry (Figure 2) and lncRNAs are expressed at lower levels compared to mRNAs. Because expression level is a good predictor for protein concentration\(^{36}\), one might speculate that lncRNAs give rise to proteins at concentrations below the mass spectrometry detection limit. To examine this issue, we first compared lncRNA and mRNA expression levels in the GENCODE v7 dataset\(^1\) (see Supplementary Material for details). While the average expression level of lncRNAs is below that of protein coding genes, the expression range is very similar (Figure 5b). In addition, a substantial number of lncRNAs are expressed at levels similar to typical mRNA transcripts. To evaluate the protein detectability as a function of its mRNA expression, we compared mRNA expression levels to the normalized spectral abundance factor (NSAF+)\(^{37}\) of the corresponding protein. The expression level is defined as the maximally observed RPKM (reads per kilobase per million mapped reads) for a particular mRNA across 11 cell lines in the GENCODE dataset. The
maximally observed NSAF+ for each protein from the 4,413 assays in PRIDE that originate from these cell lines is reported. The NSAF+ and RPKM show a low but significant correlation (Spearman $\rho = 0.32$, p-value < 0.01), which is particularly apparent in the higher expression ranges (Figure 5c). Importantly, even though low abundant proteins are more difficult to detect, detected proteins cover the entire expression range. Thus, should IncRNAs give rise to proteins, their concentrations should be detectable by mass spectrometry.
Figure 5. LncRNA and mRNA expression profile and detectability. a) Two-dimensional kernel density plot of LncRNA and mRNA expression levels and subcellular localization. The enrichment of nuclear over cytosolic expression versus the expression in the whole-cell extract is shown. Selected LncRNA and protein coding genes are depicted. Especially low abundant LncRNAs show nuclear
enrichment compared to mRNAs (adapted from Djebali et al.) b) Whole-cell expression distribution for IncRNAs and mRNAs. Although IncRNAs are generally expressed at lower levels, a substantial overlap is observed. c) Normalized spectral abundance factor (NSAF) of the detected protein as a function of its RNA expression level. While mRNA expression and NSAF are moderately correlated, the entire range of expression is clearly covered and thus detectable with mass-spectrometry.

The fact remains that most (if not all) IncRNAs contain canonical ORFs. While predictions classify these as non-coding (hence the annotation as IncRNA), it is conceivable that these ORFs represent recent evolutionary adaptations and are thus difficult to detect by in silico analyses. To evaluate if IncRNA ORFs are evolutionary retained or products of random nucleotide progression, we examined the relative size of these ORFs. By using the reverse complement of the sequence as a control, it is obvious that mRNA ORFs are much larger than random ORFs in the reverse complement sequence (see Supplementary Material for details). In contrast, IncRNA ORFs do not differ in size from randomly occurring ORFs (Figure 6), suggesting that they are indeed the product of random nucleotide progression. In addition, it was previously shown that IncRNA ORFs do not show the within-species substitution patterns expected of recently evolved proteins.
Figure 6. Relative size of the largest canonical ORF in mRNA and lncRNA transcripts. Using the reverse complement sequence as a control, it is apparent that lncRNA (as opposed to mRNA) ORFs are not larger than what would be expected from random nucleotide progression.
CONCLUSIONS

Investigations into the proportion of coding IncRNAs have resulted in very different estimates. RNA-based analyses, including ribosome profiling, has led to very high estimates, while the more direct measurement of IncRNA-derived proteins via mass spectrometry has turned up only a small percentage of putatively coding IncRNAs. In order to help resolve this discrepancy, we here performed a detailed yet thorough analysis across the very large amounts of publicly data available for the human and murine proteomes to eliminate possible biases of mass spectrometry based proteomics in detecting IncRNA-derived proteins. Our analyses reveal that the detection of proteins by mass spectrometry displays only limited bias, relating to proteins with very low abundance and/or very short sequence lengths (shorter than 44 amino acids). Nevertheless, it should be noted that specialized methods can circumvent the observed protein detection biases. Targeted sampling of less studied tissues may still reveal the existence of IncRNA-encoded, tissue specific\(^1\) translation products. Short translation products can be picked up using peptidomics approaches\(^3^8\), and enrichment protocols\(^3^1\) can boost yet unseen (micro-)peptides above the mass spectrometry detection threshold. Our analyses thus also delineate useful methods and protocols for comprehensive analysis strategies that are tailored towards finding yet unfound putative protein products from IncRNAs.

Even though mass spectrometry has its limitations in the detection of very low abundant or very small proteins, we firmly demonstrate here that these limitations alone cannot explain the discrepancy between the observed number of IncRNA-encoded proteins and the predicted number by various ribosome profiling studies. In addition, we show that the putative protein products of IncRNA ORFs do not differ in protein sequence length or composition from currently well-detectable proteins. It is
thus unlikely that the majority of the current lncRNA annotation consists of miss-
classified protein coding genes. These findings confirm that ribosome association
alone is insufficient to define novel coding ORFs, as was already suggested by some
ribosome profiling studies.

ASSOCIATED CONTENT

Supporting Information

The following files are available free of charge at ACS website http://pubs.acs.org:

Supporting Information.docx. Supplementary methods, supplementary tables S-1 –
S-4.

AUTHOR INFORMATION

Corresponding Author

* Corresponding author: Prof. Dr. Lennart Martens, A. Baertsoenkaai 3, B-9000 Gent,
Belgium, lennart.martens@vib-ugent.be, tel: +32 9 264 93 58, fax: +32 9 264 94 84

Author Contributions

The manuscript was written through contributions of all authors. All authors have
given approval to the final version of the manuscript. ‡These authors contributed
equally.

Funding Sources

This work was supported by the Multidisciplinary Research Partnership
‘Bioinformatics: From Nucleotides to Networks’ Project of Ghent University
[01MR0310W to P.V]; Fund for Scientific Research Flanders [FWO; to P.M. P.D.
and G.M.]; The European Union 7th Framework Program ‘PRIME-XS’ [262067 to L.M.]; Ghent University [to K.V. and J.V.].
REFERENCES


(27) Vaudel, M.; Burkhart, J. M.; Zahedi, R. P.; Oveland, E.; Berven, F. S.; Sickmann, A.; Martens, L.; Barsnes, H. PeptideShaker enables reanalysis of


for TOC only