Journal of proteome research

Article

pubs.acs.org/jpr

Noncoding after All: Biases in Proteomics Data Do Not Explain Observed Absence of lncRNA Translation Products

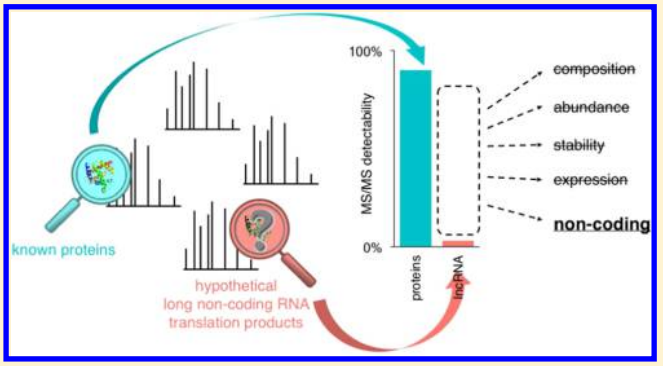
Kenneth Verheggen,[∇,†,‡,#] Pieter-Jan Volders,[∇,†,‡,§,⊥,#] Pieter Mestdagh,[§,⊥,#] Gerben Menschaert,[∥,⊥,#] Petra Van Damme,[†,‡] Kris Gevaert,[†,‡] Lennart Martens,[*,†,‡,#] and Jo Vandesompele[§,⊥,#]

[†]VIB-UGent Center for Medical Biotechnology, Ghent 9000, Belgium

[‡]Department of Biochemistry; [§]Center for Medical Genetics; [∥]Department of Mathematical Modeling, Statistics and Bioinformatics; [⊥]Cancer Research Institute Ghent (CRIG); and [#]Bioinformatics Institute Ghent (BIG N2N), Ghent University, Ghent 9000 Belgium

Supporting Information

ABSTRACT: Over the past decade, long noncoding RNAs (lncRNAs) have emerged as novel functional entities of the eukaryotic genome. However, the scientific community remains divided over the amount of true noncoding 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 data sets, strongly suggesting that the large majority of lncRNAs is indeed not translated.



KEYWORDS: *public data, lncRNA, mass spectrometry, msms, micropeptide, translation*

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) and lack large open reading frames (ORFs) and homology to annotated protein-coding genes.[1] Termed long noncoding 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 nonconserved 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.[10] Other researchers report much more conservative numbers.[11−14] For instance, if the relative abundance of ribosomes before and after stop codons (termed ribosome release) is used to discriminate between protein-coding and noncoding transcripts, only a few novel coding ORFs are found.[11] 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[13] 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,[15] we have reprocessed large quantities of tandem mass spectrometry data obtained from the PRoteomics IDEntifica-

Received: February 14, 2017
Published: May 23, 2017

DOI: 10.1021/acs.jproteome.7b00085
J. Proteome Res. 2017, 16, 2508−2515

tions (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; Tables S-1 and S-2). The results of these searches are publicly available through the LNCipedia portal.

Other groups have reported similar numbers, ranging from less than 100 up to 1,600 putative lncRNA-encoded proteins in human.[16−18] Compared to the more than 60,000 reported lncRNA genes,[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 noncoding RNA transcripts, thus eliminating the possibility that mass spectrometry would be biased against the detection of putative lncRNA-encoded proteins.

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.[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,[23] one of the world's leading mass spectrometry repositories.[24] Sequence database searches were performed using an automated reprocessing pipeline, consisting of pride-asap[25] for the detection of data set specific parameters, SearchGUI[26] to match the fragmentation mass spectra against peptides derived from protein sequence databases, and PeptideShaker[27] to integrate the identifications and control these at a 1% false discovery rate at the peptide-to-spectrum match level.

Table 1. Ten Shortest Human Proteins Identified by Reprocessing of the Reprocessed PRIDE Data

protein	gene name	length (AA)	average MW (Da)	spectral count	assay count
P62328	TMSB4X	44	4921.46	787	287
P63313	TMSB10	44	4894.48	366	229
Q8N4H5	TOMM5	51	6035.31	88	70
P62891	RPL39	51	6275.49	109	52
Q59GN2	RPL39P5	51	6322.59	107	51
Q5VTU8	ATP5EP2	51	5806.87	53	43
P56381	ATP5E	51	5648.57	53	43
Q96IX5	USMG5	58	6326.38	112	86
P62861	FAU	59	6647.86	248	141
P13640	MT1G	62	6647.86	71	47

Because the combination of known canonical human protein sequences and hypothetical lncRNA-derived sequences can hamper protein inference, the overlap between both data sets 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.[28] Out of 8,645,916 hypothetical tryprtic peptide sequences, only 277,412 had one or more identical matches in the protein sequence data set. 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 analyze this, publicly available submissions of human projects to PRIDE were searched against the human complement of the UniProtKB/SwissProt[28] 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.

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[29] with tryptic cleavage rules, allowing for two missed cleavages. Both empirical peptides from the reprocessing of the human data in PRIDE and *in silico* obtaining 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.[30]

DOI: 10.1021/acs.jproteome.7b00085
J. Proteome Res. 2017, 16, 2508−2515

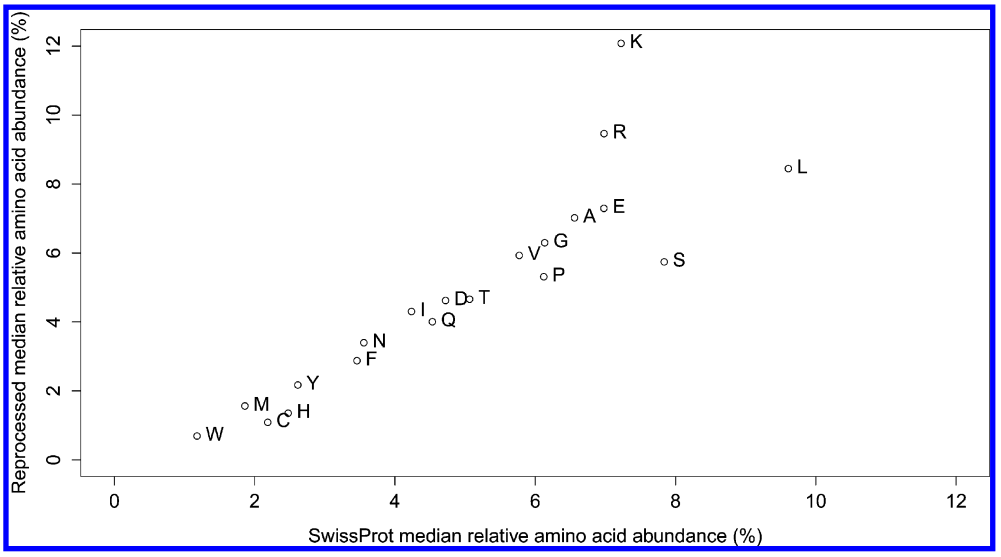


Figure 1. Comparison between theoretical (UniProtKB/SwissProt) and observed (reprocessed PRIDE data) peptide sequence amino acid composition for human data from PRIDE and UniProtKB/SwissProt.

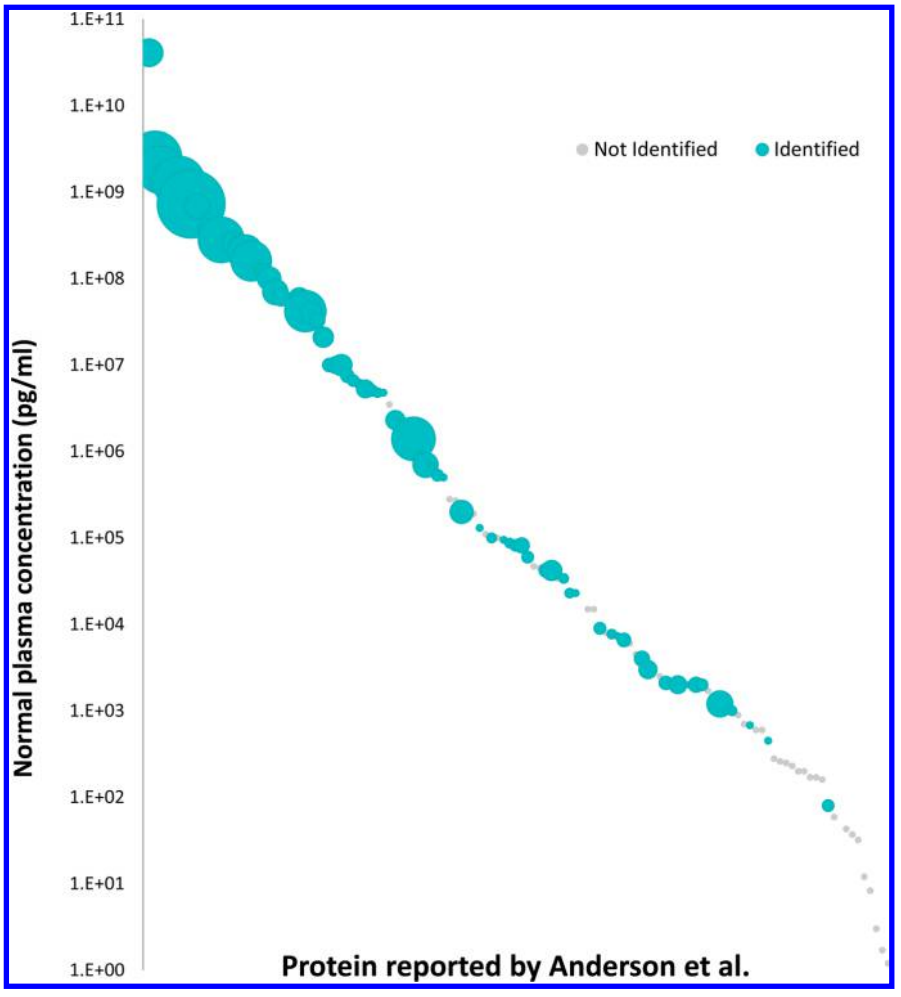


Figure 2. Reprocessing results for PRIDE data sets derived from human blood plasma mapped onto the abundance values reported by Anderson and Hunter.[32] The size of a bubble corresponds to the number of PRIDE assays in which that protein was identified.

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 data sets (Spearman ρ = 0.952, $p < 0.01$), hinting that there is no reason to assume that the compositions of proteins identified by the reprocessing of PRIDE and those generated by *in silico* digestion are very different. The higher occurrence rates for R and K in the

DOI: 10.1021/acs.jproteome.7b00085
J. Proteome Res. 2017, 16, 2508−2515

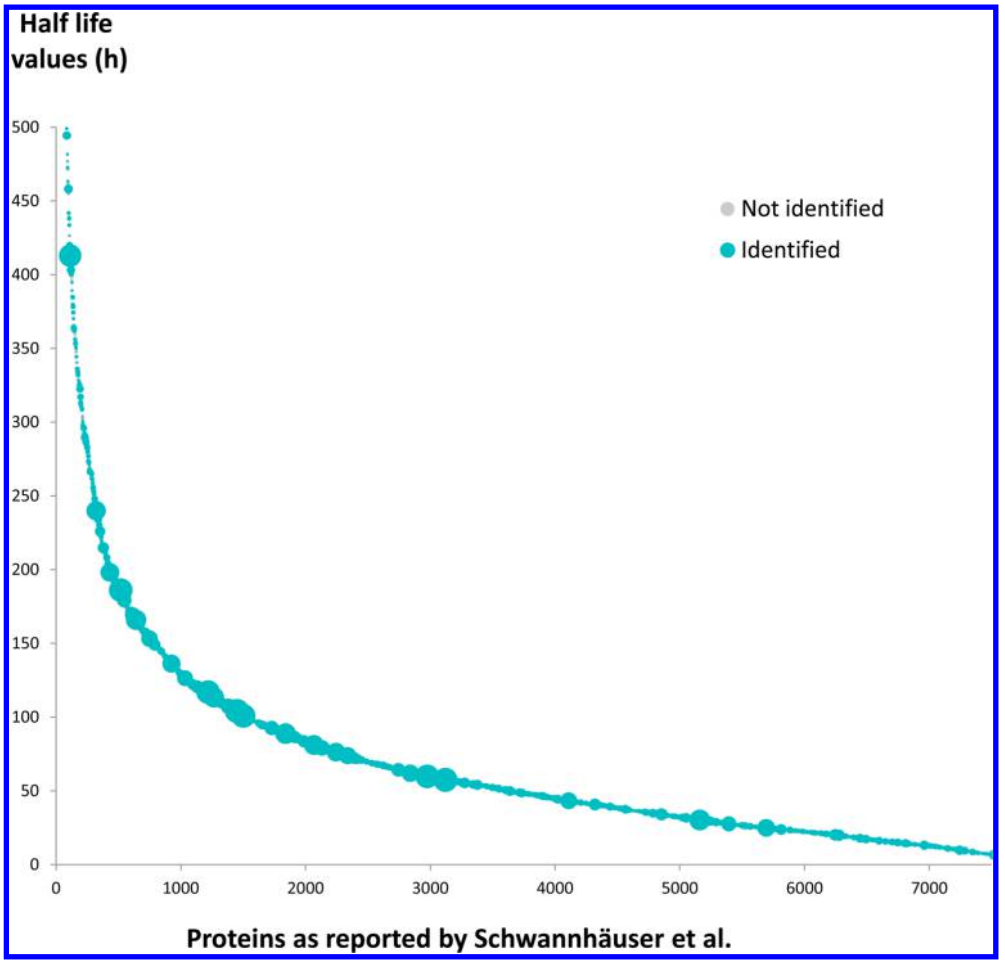


Figure 3. Reprocessing results for all PRIDE murine data mapped onto the half-life values reported by Schwanhäusser et al.[33] The size of the bubble corresponds to the number of PRIDE assays in which the protein was identified.

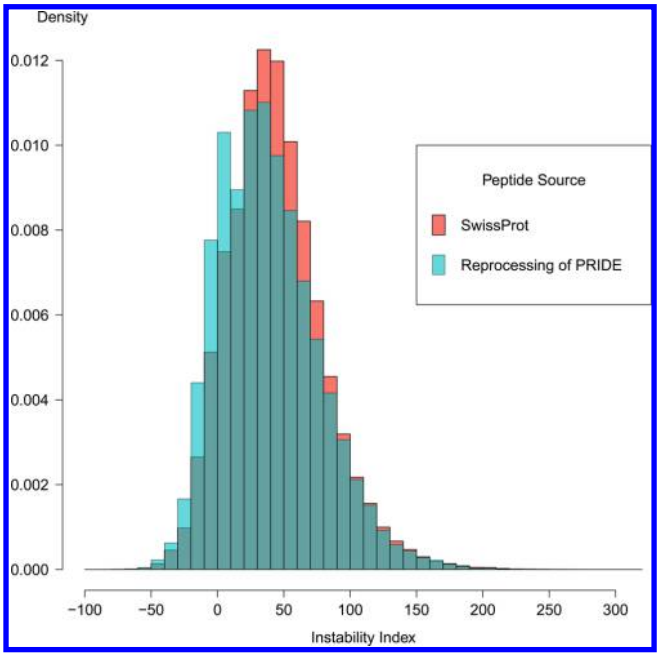


Figure 4. Instability index distributions of human UniProtKB/SwissProt proteins, and of identified proteins from reprocessed human data sets in PRIDE.

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 to the somewhat lower efficiency in the detection of phosphorylated residues.

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,[31] 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[32] 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 11 orders of magnitude, of which at least eight are covered reliably in the PRIDE data. This analysis thus shows that mass

DOI: 10.1021/acs.jproteome.7b00085
J. Proteome Res. 2017, 16, 2508−2515

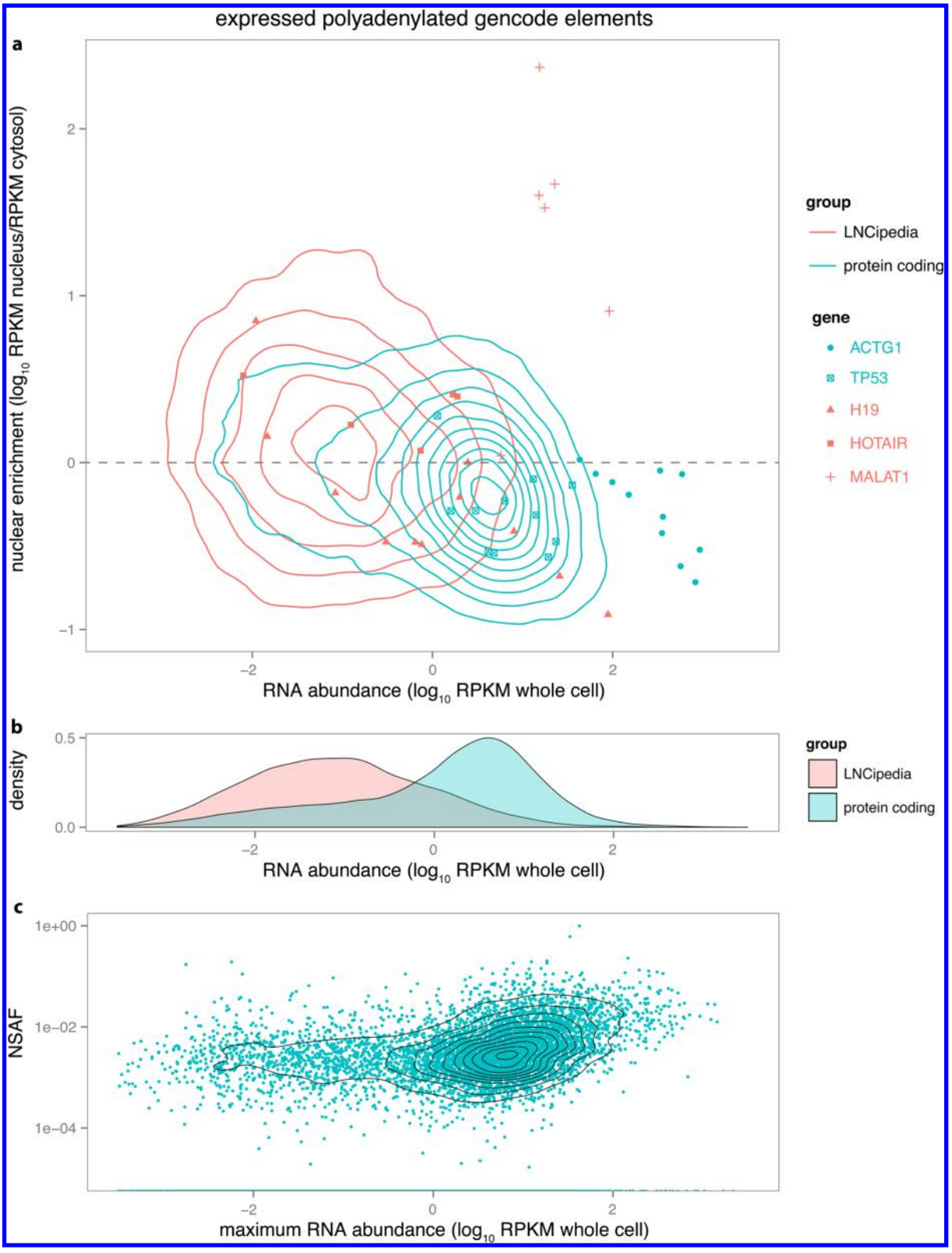


Figure 5. LncRNA and mRNA expression profile and detectability. (a) Two-dimensional kernel density plot of lncRNA and mRNA expression levels and subcellular localization. The enrichment of nuclear over cytosolic expression versus the expression in the whole-cell extract is shown. Selected lncRNA and protein-coding genes are depicted. Especially low abundant lncRNAs show nuclear enrichment compared to mRNAs (adapted from Djebali et al.[1]). (b) Whole-cell expression distribution for lncRNAs and mRNAs. Although lncRNAs 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.

DOI: 10.1021/acs.jproteome.7b00085
J. Proteome Res. 2017, 16, 2508−2515

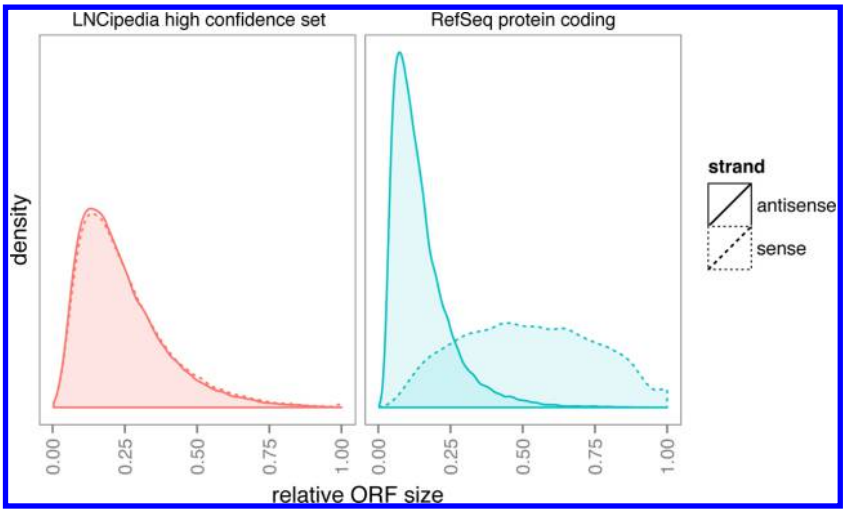


Figure 6. Relative size of the largest canonical ORF in mRNA and lncRNA transcripts. Using the reverse complement sequence as a control, it is apparent that lncRNA (as opposed to mRNA) ORFs are not larger than what would be expected from random nucleotide progression.

spectrometry-based proteomics is only biased against the very least abundant proteins.

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.[33] 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.

In addition, we calculated the N-terminal instability index of human proteins as described by Guruprasad et al.[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 has little, if any, influence on its detectability.

LNCRNA EXPRESSION AND COMPOSITION SHOW NO INDICATION OF CODING POTENTIAL

The expression profiles of lncRNAs differ extensively from those 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 noncoding 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 data set[1] (see Supporting Information 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 data set. 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 ρ = 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 lncRNAs give rise to proteins, their concentrations should be detectable by mass spectrometry.

The fact remains that most (if not all) lncRNAs contain canonical ORFs. While predictions classify these as noncoding (hence the annotation as lncRNA), it is conceivable that these ORFs represent recent evolutionary adaptations and are thus difficult to detect by *in silico* analyses. To evaluate if lncRNA 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 Supporting Information for details). In contrast, lncRNA 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 lncRNA ORFs do not show the within-species substitution patterns expected of recently evolved proteins.[11]

DOI: 10.1021/acs.jproteome.7b00085
J. Proteome Res. 2017, 16, 2508−2515

CONCLUSIONS

Investigations into the proportion of coding lncRNAs have resulted in very different estimates. RNA-based analyses, including ribosome profiling, have led to very high estimates, while the more direct measurement of lncRNA-derived proteins via mass spectrometry has turned up only a small percentage of putatively coding lncRNAs. 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 lncRNA-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 lncRNA-encoded, tissue specific[1] translation products. Short translation products can be picked up using peptidomics approaches,[38] and enrichment protocols[31] 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 toward finding yet unfound putative protein products from lncRNAs.

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 lncRNA-encoded proteins and the predicted number by various ribosome profiling studies. In addition, we show that the putative protein products of lncRNA 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 Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.7b00085.

Supplementary methods,; supplementary tables S-1 to S-4, showing overview of the obtained coverage, processed RNA-sequencing datasets, and overview of the PRIDE projects (PDF)

AUTHOR INFORMATION

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.

ORCID

Pieter-Jan Volders: 0000-0002-2685-2637
Kris Gevaert: 0000-0002-4237-0283
Lennart Martens: 0000-0003-4277-658X

Author Contributions

∇K.V. and P.-J.V. contributed equally.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

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.]; SBO grant "InSPECtor" of Flanders Innovation & Entrepreneurship (VLAIO) [120025 to L.M.]; and Ghent University [to K.V. and J.V.].

REFERENCES

(1) Djebali, S.; Davis, C. A.; Merkel, A.; Dobin, A.; Lassmann, T.; Mortazavi, A. M.; Tanzer, A.; Lagarde, J.; Lin, W.; Schlesinger, F.; et al. Landscape of transcription in human cells. *Nature* **2012**, *489* (7414), 101−108.

(2) Iyer, M. K.; Niknafs, Y. S.; Malik, R.; Singhal, U.; Sahu, A.; Hosono, Y.; Barrette, T. R.; Prensner, J. R.; Evans, J. R.; Zhao, S.; et al. The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **2015**, *47* (3), 199−208.

(3) Mercer, T. R.; Dinger, M. E.; Mattick, J. S. Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.* **2009**, *10* (3), 155−159.

(4) Crappé, J.; Van Criekinge, W.; Menschaert, G. Little things make big things happen: A summary of micropeptide encoding genes. *EuPa Open Proteomics* **2014**, *3*, 128−137.

(5) Lin, M. F.; Jungreis, I.; Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **2011**, *27* (13), I275−I282.

(6) Wang, L.; Park, H. J.; Dasari, S.; Wang, S.; Kocher, J.-P.; Li, W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **2013**, *41* (6), e74−e74.

(7) Kong, L.; Zhang, Y.; Ye, Z. Q.; Liu, X. Q.; Zhao, S. Q.; Wei, L.; Gao, G. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* **2007**, *35*, W345−W349.

(8) Ingolia, N. T. Genome-Wide Translational Profiling by Ribosome Footprinting. In *Guide to Yeast Genetics: Functional Genomics, Proteomics, and Other Systems Analysis*; Methods in Enzymology; Elsevier, 2010; Vol. *470*, pp 119−142.

(9) Ingolia, N. T.; Lareau, L. F.; Weissman, J. S. Ribosome Profiling of Mouse Embryonic Stem Cells Reveals the Complexity and Dynamics of Mammalian Proteomes. *Cell* **2011**, *147* (4), 789−802.

(10) Ingolia, N. T.; Brar, G. A.; Stern-Ginossar, N.; Harris, M. S.; Talhouarne, G. J. S.; Jackson, S. E.; Wills, M. R.; Weissman, J. S. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell Rep.* **2014**, *8* (5), 1365−1379.

(11) Guttman, M.; Russell, P.; Ingolia, N. T.; Weissman, J. S.; Lander, E. S. Ribosome Profiling Provides Evidence that Large Noncoding RNAs Do Not Encode Proteins. *Cell* **2013**, *154* (1), 240−251.

(12) Chew, G.-L.; Pauli, A.; Rinn, J. L.; Regev, A.; Schier, A. F.; Valen, E. Ribosome profiling reveals resemblance between long non-coding RNAs and 5′ leaders of coding RNAs. *Development* **2013**, *140* (13), 2828−2834.

(13) Bazzini, A. A.; Johnstone, T. G.; Christiano, R.; Mackowiak, S. D.; Obermayer, B.; Fleming, E. S.; Vejnar, C. E.; Lee, M. T.; Rajewsky, N.; Walther, T. C.; et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* **2014**, *33* (9), 981−993.

(14) Lee, S.; Liu, B.; Lee, S.; Huang, S.-X.; Shen, B.; Qian, S.-B. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U. S. A.* **2012**, *109* (37), E2424−E2432.

(15) Volders, P.-J.; Verheggen, K.; Menschaert, G.; Vandepoele, K.; Martens, L.; Vandesompele, J.; Mestdagh, P. An update on

DOI: 10.1021/acs.jproteome.7b00085
J. Proteome Res. 2017, 16, 2508−2515

LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res.* **2015**, *43* (Database issue), D174−D180.
(16) Menschaert, G.; Van Criekinge, W.; Notelaers, T.; Koch, A.; Crappé, J.; Gevaert, K.; Van Damme, P. Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Mol. Cell. Proteomics* **2013**, *12* (7), 1780−1790.
(17) Crappé, J.; Van Criekinge, W.; Trooskens, G.; Hayakawa, E.; Luyten, W.; Baggerman, G.; Menschaert, G. Combining in silico prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. *BMC Genomics* **2013**, *14* (1), 648.
(18) Slavoff, S. A.; Mitchell, A. J.; Schwaid, A. G.; Cabili, M. N.; Ma, J.; Levin, J. Z.; Karger, A. D.; Budnik, B. A.; Rinn, J. L.; Saghatelian, A. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **2012**, *9* (1), 59−64.
(19) Volders, P.-J.; Verheggen, K.; Menschaert, G.; Vandepoele, K.; Martens, L.; Vandesompele, J.; Mestdagh, P. An update on LNCipedia: a database for annotated human lncRNA sequences. *Nucleic Acids Res.* **2015**, *43* (Database issue), D174−D180.
(20) Brewis, I. A.; Brennan, P. Proteomics technologies for the global identification and quantification of proteins. *Adv. Protein Chem. Struct. Biol.* **2010**, *80*, 1−44.
(21) Klie, S.; Martens, L.; Vizcaíno, J. A.; Côté, R.; Jones, P.; Apweiler, R.; Hinneburg, A.; Hermjakob, H. Analyzing large-scale proteomics projects with latent semantic indexing. *J. Proteome Res.* **2008**, *7* (1), 182−191.
(22) Leary, D. H.; Hervey, W. J.; Deschamps, J. R.; Kusterbeck, A. W.; Vora, G. J. Which metaproteome? The impact of protein extraction bias on metaproteomic analyses. *Mol. Cell. Probes* **2013**, *27* (5−6), 193−199.
(23) Vizcaíno, J. A.; Côté, R. G.; Csordas, A.; Dianes, J. A.; Fabregat, A.; Foster, J. M.; Griss, J.; Alpi, E.; Birim, M.; Contell, J.; et al. The PRoteomics IDEntifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.* **2013**, *41* (Database issue), D1063−D1069.
(24) Vizcaíno, J. A.; Deutsch, E. W.; Wang, R.; Csordas, A.; Reisinger, F.; Ríos, D.; Dianes, J. A.; Sun, Z.; Farrah, T.; Bandeira, N.; et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **2014**, *32* (3), 223−226.
(25) Hulstaert, N.; Reisinger, F.; Rameseder, J.; Barsnes, H.; Vizcaíno, J. A.; Martens, L. Pride-asap: automatic fragment ion annotation of identified PRIDE spectra. *J. Proteomics* **2013**, *95*, 89−92.
(26) Vaudel, M.; Barsnes, H.; Berven, F. S.; Sickmann, A.; Martens, L. SearchGUI: An open-source graphical user interface for simultaneous OMSSA and X!Tandem searches. *Proteomics* **2011**, *11* (5), 996−999.
(27) Vaudel, M.; Burkhart, J. M.; Zahedi, R. P.; Oveland, E.; Berven, F. S.; Sickmann, A.; Martens, L.; Barsnes, H. PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat. Biotechnol.* **2015**, *33* (1), 22−24.
(28) UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2014**, *42* (Databaseissue), D191−D198.
(29) Martens, L.; Vandekerckhove, J.; Gevaert, K. DBToolkit: processing protein databases for peptide-centric proteomics. *Bioinformatics* **2005**, *21* (17), 3584−3585.
(30) Vandermarliere, E.; Mueller, M.; Martens, L. Getting intimate with trypsin, the leading protease in proteomics. *Mass Spectrom. Rev.* **2013**, *32* (6), 453−465.
(31) Mustafa, G. M.; Larry, D.; Petersen, J. R.; Elferink, C. J. Targeted proteomics for biomarker discovery and validation of hepatocellular carcinoma in hepatitis C infected patients. *World J. Hepatol* **2015**, *7* (10), 1312−1324.
(32) Anderson, L.; Hunter, C. L. Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. *Mol. Cell. Proteomics* **2005**, *5* (4), 573−588.
(33) Schwanhäusser, B.; Busse, D.; Li, N.; Dittmar, G.; Schuchhardt, J.; Wolf, J.; Chen, W.; Selbach, M. Corrigendum: Global quantification of mammalian gene expression control. *Nature* **2013**, *495* (7439), 126−127.
(34) Guruprasad, K.; Reddy, B. V.; Pandit, M. W. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng., Des. Sel.* **1990**, *4* (2), 155−161.
(35) Rinn, J. L.; Chang, H. Y. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **2012**, *81*, 145−166.
(36) Wilhelm, M.; Schlegl, J.; Hahne, H.; Moghaddas Gholami, A.; Lieberenz, M.; Savitski, M. M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; et al. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, *509* (7502), 582−587.
(37) Zybailov, B.; Mosley, A. L.; Sardiu, M. E.; Coleman, M. K.; Florens, L.; Washburn, M. P. Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J. Proteome Res.* **2006**, *5* (9), 2339−2347.
(38) Schulz-Knappe, P.; Schrader, M.; Zucht, H.-D. The peptidomics concept. *Comb. Chem. High Throughput Screening* **2005**, *8* (8), 697−704.

DOI: 10.1021/acs.jproteome.7b00085
J. Proteome Res. 2017, 16, 2508−2515