Genomics 109 (2017) 113–122

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Widespread pre-translational regulation of the inclusion of signal peptides in human proteins

Philippe Balthazar [1], Daniel C. Tucunduva [1], Mikael-Jonathan Luce, Michelle S Scott *

Department of Biochemistry and RNA Group, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, Québec J1E 4K8, Canada

ARTICLE INFO

Article history:
Received 8 November 2016
Received in revised form 16 January 2017
Accepted 18 January 2017
Available online 22 January 2017

Keywords:
Signal peptide
Protein subcellular localization
Pre-translational regulation
Alternative splicing
RNA-seq
Tissue-specific regulation

ABSTRACT

Signal peptides (SP) are cleavable N-terminal protein motifs used co-translationally for entry of nascent polypeptides into the secretory pathway. Their co-translational cleavage prevents their extensive post-translational regulation and flexibility in their usage is made possible by the control of their inclusion at a pre-translational level. To characterize this regulation on a transcriptome scale, we analyzed the level and mechanisms of inclusion of the 3298 most likely human SP-encoding genes, 47% of which alternatively express their SP. Analysis of RNA-seq data across different normal human tissues indicates that pre-translational regulation of the SP differs depending on tissue-coverage of the gene, with alternative SP genes more likely to be widely expressed than constitutive SP genes. SP inclusion represents a new metric to measure functional gene expression and its deregulation in disease. Our analysis supports the extensive use of pre-translational regulation of SP inclusion, with functional consequences and implications for biomarker discovery.

1. Introduction

Signal peptides (SP) are short transient N-terminal protein sequences serving as the main targeting signal for entry into the secretory pathway. First described over four decades ago, SPs are widely used in all eukaryotes [1,2]. Most SPs are recognized co-translationally and are bound by the signal recognition particle (SRP) as soon as they emerge from the ribosome [3,4]. The SRP directs the nascent polypeptide and its translating ribosome to the endoplasmic reticulum (ER) membrane where translation resumes, and the polypeptide is translocated across the membrane through a protein-conducting channel [2,5]. Once inserted through the ER membrane, the SP is cleaved by the signal peptidase, a resident ER enzyme. As a consequence, SPs can only be used once and are not subject to post-translational regulation as are protein targeting motifs such as the nuclear localization signals, which can be controlled by limiting their accessibility by conformational change or by post-translation modification [6,7].

Both soluble and membrane spanning proteins can contain SPs. In the case of SP-containing membrane proteins, they have extra hydrophobic transmembrane segments that are laterally inserted into the lipid bilayer when they pass through the translocation channel, while the hydrophilic segments of the protein either cross the ER membrane through the channel or remain in the cytosol [5]. Luminal portions of proteins inserted into the ER can be N-glycosylated and then the proteins are routed to their target destination in a cytoplasmic organelle (including the ER, Golgi apparatus, endosomes, lysosomes, or in some cases the peroxisomes), or to the plasma membrane or outside the cell [8,9].

Although always found at the N-terminus of proteins, SPs are of variable length and consist of three structural segments termed the n, h and c-regions [10]. While the n-region is often positively charged, the h-region is enriched in hydrophobic residues and forms an alpha-helix that spans the ER membrane [10]. The c-region is neutral but dominated by polar residues, except for the residues in positions −3 and −1 (with respect to the cleavage site), which are small and neutral, defining the cleavage site [10–12]. This well-described structure of SPs and the large number of SPs that have been reported have enabled the creation of many predictors that identify SPs with high accuracy [13].

The correct prediction and annotation of SPs is of great importance because these motifs will determine the localization and ultimately the environment, the interaction partners and consequently, the function of the protein. The presence/absence of a SP in a protein is likely to significantly change the fate and cellular role of a protein. Cells can control the inclusion of the signal sequence coding region (SSCR, i.e. the region encoding a SP in a transcript) at pre-translational levels, through the use of alternative promoters, alternative splicing, or through the differential choice of start codons, resulting in protein isoforms with different localization [14,15]. Several examples of pre-translational regulation resulting in the alternative inclusion of an SSCR have been reported. For example, the VEGF-A gene regulates the localization of its protein products with two different promoters resulting in the use of different start codons. While one encodes a protein with a SP at its N-

* Corresponding author.
E-mail address: michelle.scott@usherbrooke.ca (M.S. Scott).

[1] Joint first authors.

http://dx.doi.org/10.1016/j.ygeno.2017.01.001
0888-7543/

terminus, the other encodes a protein without an N-terminal SP, which is cleaved and functions in the nucleus [16,17]. By varying the relative abundance of each form, cells can modulate the predominant localization and function of the encoded products.

The alternative inclusion of SSCRs has been investigated previously for a small subgroup of 145 SSCR-encoding mouse genes [18] as well as for a larger group of 1475 SSCR-encoding mouse genes [15]. In both cases, based on transcript sequence annotations and SP predictions, approximately 50% of these genes were predicted to encode both transcripts with an SSCR and transcripts without. These results suggested the widespread usage of pre-translation regulation mechanisms for the inclusion of SSCRs. However, levels and patterns of the inclusion were not investigated, nor whether the transcripts not encoding an SSCR are even expressed in comparable levels to those containing an SSCR.

To further characterize the pre-translational regulation of the inclusion of SPs, we consider 3298 human genes predicted to encode an SSCR and characterize the prevalence and mechanisms used to alternatively include the motif. We then investigate the inclusion of the SSCR quantitatively in healthy and diseased human tissues, supporting the concept of pre-translation mechanisms being important regulators of the presence of SPs in proteins, which can be controlled in a tissue-specific and condition-specific way.

2. Results

2.1. Prediction of signal peptides in Ensembl transcripts

To characterize the set of genes coding for a SP, we considered the SignalP and Phobius predictors, both of which show high accuracy in the prediction of SPs in eukaryotes [13,19,20]. In addition, these two predictors also show the best performances at discriminating between SPs and transmembrane regions, which is important in the context of a proteome-wide prediction [19]. Of the 19,627 human coding genes defined in Ensembl version 83 [21], SignalP and Phobius predict respectively 3858 and 4881 genes to encode proteins containing a SP (Fig. 1A). In total, 16.8% of human protein-coding genes (3298) are predicted to code for a SP by both predictors (Fig. 1) and this high confidence predicted set (HCP) was used for all further analyses. The HCP was further characterized and validated using genes encoding experimentally confirmed SPs as positive controls and genes annotated as encoding nuclear proteins as negative controls (see Methods). As shown in Fig. S1A, of the 659 genes encoding experimentally validated SPs considered, 94.5% (623 genes) are part of the HCP indicating a high true positive rate for our high confidence dataset. In contrast, of the 6574 genes annotated as encoding nuclear proteins, only 4.3% (281 genes) are part of the HCP indicating a low false positive rate. We note that although most nuclear proteins are not expected to have SPs, a small subset such as FGF1 [22] have been shown to code for proteins with both a SP and a nuclear localization signal. And thus the false positive rate of the HCP might be overestimated.

Considering all Ensembl protein coding transcript annotations, we investigated whether all or only a subset of transcripts per HCP gene encode an SSCR. For 1739 (53%) of the 3298 HCP genes, the SSCR is present in all encoded protein-coding transcripts and these motifs are thus referred to as constitutive SSCRs. In contrast, the remaining 1559 genes (47%) encode both transcripts with and without an SSCR, which are thus referred to as alternative motifs (Fig. 1B).

2.2. Exonic position of SSCRs in transcripts

SSCRs in the HCP are on average 71 ± 18 nucleotides long (SSCR length distribution shown in Fig. S2A). While they are typically short enough to be entirely encoded in one exon (the coding region of the first coding exon in human being on average 188 ± 449 nucleotides long with a median of 95 nucleotides, Fig. S2B), some SSCRs stretch across as many as 3 exons. 70% of constitutive HCP genes encode the SSCR entirely in one exon, which is the expected proportion based on SSCR length distribution and coding region distribution of first coding exon in the human transcriptome (compare HCP constitutive to random, Fig. 2A). Interestingly, this proportion goes down to 62% for alternative HCP genes (p-value $< 10^{-6}$) suggesting that amongst the types of regulation used to alternatively include the SSCR, splicing is one such mechanism. However, even for the SPs encoded in more than one exon, the near totality of the SSCR is contained in the first coding exon (Fig. 2B). Indeed, in contrast to randomly chosen transcripts from the human transcriptome, a strong selective pressure seems to exist to ensure that as much of the SSCR as possible is encoded in the first coding exon (Fig. 2B). As a consequence, half (51%) of HCP SSCRs encoded in more than one exon encode at least 80% of their SSCR in their first coding exon, as compared to 22% expected for randomly chosen regions. As the c-region is typically only 5 to 6 amino acids long and does not increase in length when the SP is longer [10], there is thus strong selective pressure for SPs to encode their entire n- and h-regions in the first coding exon of their transcript (>80% of SPs follow this trend).

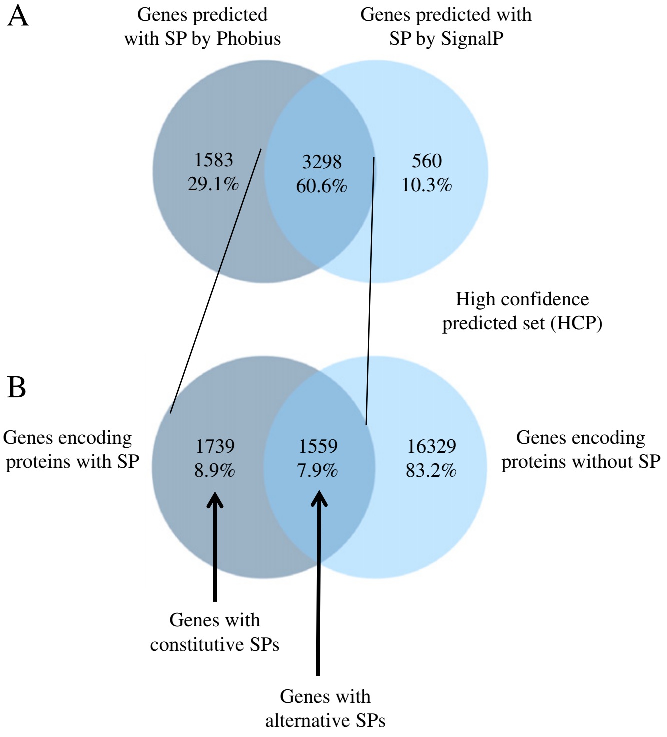
A
Genes predicted with SP by Phobius
Genes predicted with SP by SignalP
1583
29.1%
3298
60.6%
560
10.3%
High confidence predicted set (HCP)
B
Genes encoding proteins with SP
1739
8.9%
1559
7.9%
16329
83.2%
Genes encoding proteins without SP
Genes with constitutive SPs
Genes with alternative SPs

Fig. 1. High confidence predicted SP dataset. The human proteome was scanned using Phobius and SignalP for proteins predicted to encode SPs. 4881 human genes were predicted by Phobius to encode at least one protein with a SP while SignalP predicts 3858 such genes, with an intersection of 3298 SP genes between the two predictors (A). Amongst this high confidence predicted SP set of 3298 genes, the SP is predicted to be present in all proteins encoded by 1739 of these genes (53%) while for the remainder, the SP is predicted to be encoded in only a subset of their transcripts (B).

2.3. Regulation mechanisms for the inclusion of SPs

To further investigate the pre-translational regulation of the SSCR, we considered mechanisms leading to alternatively present SPs. Amongst the HCP alternative SSCRs, we identified three main mechanisms regulating their inclusion, which were classified according to the decision diagram shown in Fig. S3. The most prevalent mechanism, used by 79.8% of alternative HCP genes, is an alternative transcription start site which results in the differential inclusion of the SSCR (Fig. 3A), as was previously described in mouse for a subset of the transcriptome [15] and previously reported by the ENCODE project [23]. Examples of genes employing this mechanism include the parathyroid