NAR Genomics and Bioinformatics, 2023, **5**, 1–17
https://doi.org/10.1093/nargab/lqad093
Advance access publication date: 18 October 2023
Standard Article

Pathogenic signal peptide variants in the human genome

Sneider Alexander Gutierrez Guarnizo[1,†], Morgana K. Kellogg[1,†], Sarah C. Miller[1], Elena B. Tikhonova[1], Zemfira N. Karamysheva[2] and Andrey L. Karamyshev[1,*]

[1]Department of Cell Biology and Biochemistry, Texas Tech University Health Sciences Center, Lubbock, TX 79430, USA
[2]Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA

[*]To whom correspondence should be addressed. Tel: +1 806 743 4102; Email: andrey.karamyshev@ttuhsc.edu
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Abstract

Secreted and membrane proteins represent a third of all cellular proteins and contain N-terminal signal peptides that are required for protein targeting to endoplasmic reticulum (ER). Mutations in signal peptides affect protein targeting, translocation, processing, and stability, and are associated with human diseases. However, only a few of them have been identified or characterized. In this report, we identified pathogenic signal peptide variants across the human genome using bioinformatic analyses and predicted the molecular mechanisms of their pathology. We recovered more than 65 thousand signal peptide mutations, over 11 thousand we classified as pathogenic, and proposed framework for distinction of their molecular mechanisms. The pathogenic mutations affect over 3.3 thousand genes coding for secreted and membrane proteins. Most pathogenic mutations alter the signal peptide hydrophobic core, a critical recognition region for the signal recognition particle, potentially activating the Regulation of Aberrant Protein Production (RAPP) quality control and specific mRNA degradation. The remaining pathogenic variants (about 25%) alter either the N-terminal region or signal peptidase processing site that can result in translocation deficiencies at the ER membrane or inhibit protein processing. This work provides a conceptual framework for the identification of mutations across the genome and their connection with human disease.

Introduction

Eukaryotic cells have multiple intracellular compartments that require coordinated protein sorting. Ribosomes synthesize thousands of proteins that must be transported to different organelles, integrated into membranes, or secreted outside of the cell (1). Different intrinsic signals in the amino acid sequence act like postal codes to deliver proteins to specific cellular locations (2,3). The most numerous targeting signals include signal peptides, N-terminal amino acid sequences that direct the targeting and translocation of many secreted and membrane proteins to the endoplasmic reticulum (ER) (4–7). Secreted and membrane proteins represent over 30% of the human proteome and participate in essential biological processes such as cell signaling, transport, and cell recognition (8–10). Defects in the trafficking of secreted or membrane proteins contribute to the pathogenesis of many human diseases (7,11–15).

Co-translational protein targeting involves nascent peptide recognition by a ribonucleoprotein complex known as the signal recognition particle (SRP). SRP co-translationally binds the signal peptides or transmembrane domains of secreted or membrane proteins once they are exposed from the ribosome's exit tunnel, forming SRP-ribosome-nascent-chain complexes (SRP-RNC). When SRP binds to the RNC, it temporarily arrests translation to allow targeting of the complex to the SRP receptor in the ER membrane. SRP then hands the RNC to the SEC61 translocon for co-translational translocation into the ER lumen. Signal peptidase cleaves signal peptides after translocation, releasing the protein into the ER lumen. These events are summarized in Figure 1A and have been described in detail in multiple reviews (7,16–21).

Signal peptides do not have a consensus sequence, but contain three distinct regions with conserved physicochemical features: a positively charged N-terminal region (N-region); a hydrophobic region (H-region), which consists of primarily aliphatic amino acid residues; and a C-terminal region (C-region), which includes the cleavage site recognized by signal peptidase (4,5,22) (Figure 2A). The H-region is the most critical region for SRP recognition, which is directly involved in hydrophobic interactions between the signal peptide and the methionine-lined binding pocket of the SRP54 subunit. Some mutations in the H-region have been shown to lead to severe defects in SRP recognition of the signal peptide (15,23–25) (Figure 1B). In contrast, some signal peptide mutations in the C-region have been shown to disrupt signal peptide cleavage, leading to the accumulation of precursor protein in the ER membrane (26,27) (Figure 1C). Mutations in the N-region can interfere with protein translocation efficiency, induce protein mistargeting and aggregation, or can lead to the accumulation of precursor proteins at the ER membrane because the signal peptide is not oriented correctly in the translocon (14,23,28,29) (Figure 1D).

Misfolded and mistargeted proteins are often cytotoxic. Cells have developed several quality control mechanisms to prevent mistargeting or the accumulation of misfolded secretory and membrane proteins (30–32). These quality control mechanisms occur in the cytosol during protein targeting or at the ER during translocation. One of them is the preemptive quality control pathway, the Regulation of Aberrant Protein Production (RAPP), which senses the failure of SRP to recognize signal peptides, and specifically targets secretory and membrane protein mRNAs for degradation

Received: July 13, 2023. **Revised:** September 5, 2023. **Editorial Decision:** September 25, 2023. **Accepted:** September 29, 2023

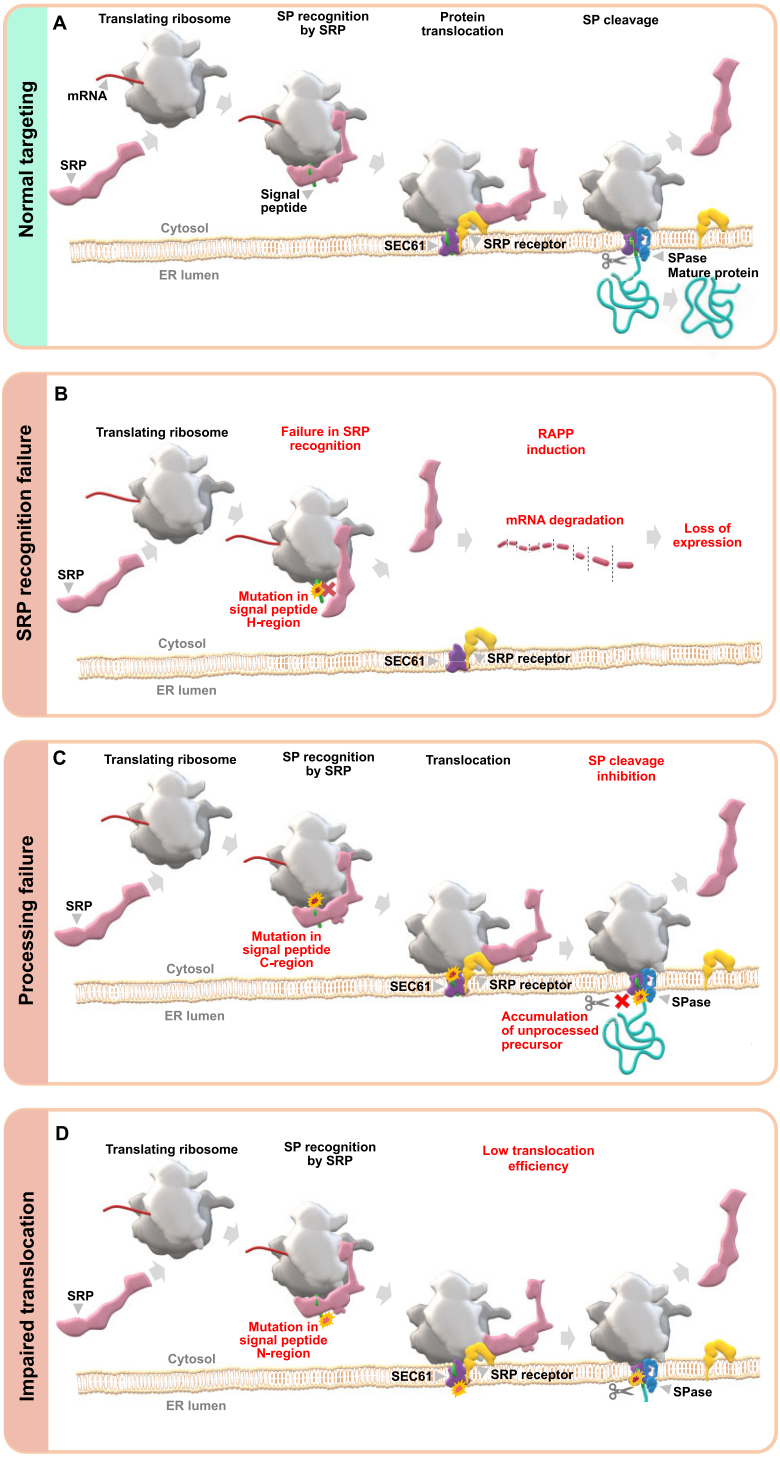


Figure 1. Mutations in signal peptides affect different molecular mechanisms. (**A**) Illustration of a normal targeting of secreted/membrane proteins to ER. When a nascent polypeptide containing signal peptide is exposed from the ribosome tunnel, it is co-translationally recognized by SRP. Then, the SRP-ribosome nascent chain complex is targeted to SRP receptor in the ER membrane. Finally, the nascent polypeptide is co-translationally translocated into the ER lumen through protein-conducting channel in SEC61 translocon, the signal peptide is cleaved off by signal peptidase (SPase), and the mature protein is released from the ER membrane and transported further to extracellular space, or integrated into the plasma membrane, or remained in the ER lumen. (**B**) Mutations affecting hydrophobicity of signal peptide H-region lead to SRP recognition failure, activation of the RAPP pathway and degradation of the protein's mRNA. (**C**) Mutations in signal peptide cleavage site may affect protein processing leading to accumulation of unprocessed protein in ER. (**D**) Mutations in signal peptide N-terminus may affect its interaction with SEC61 translocon decreasing protein translocation efficiency.

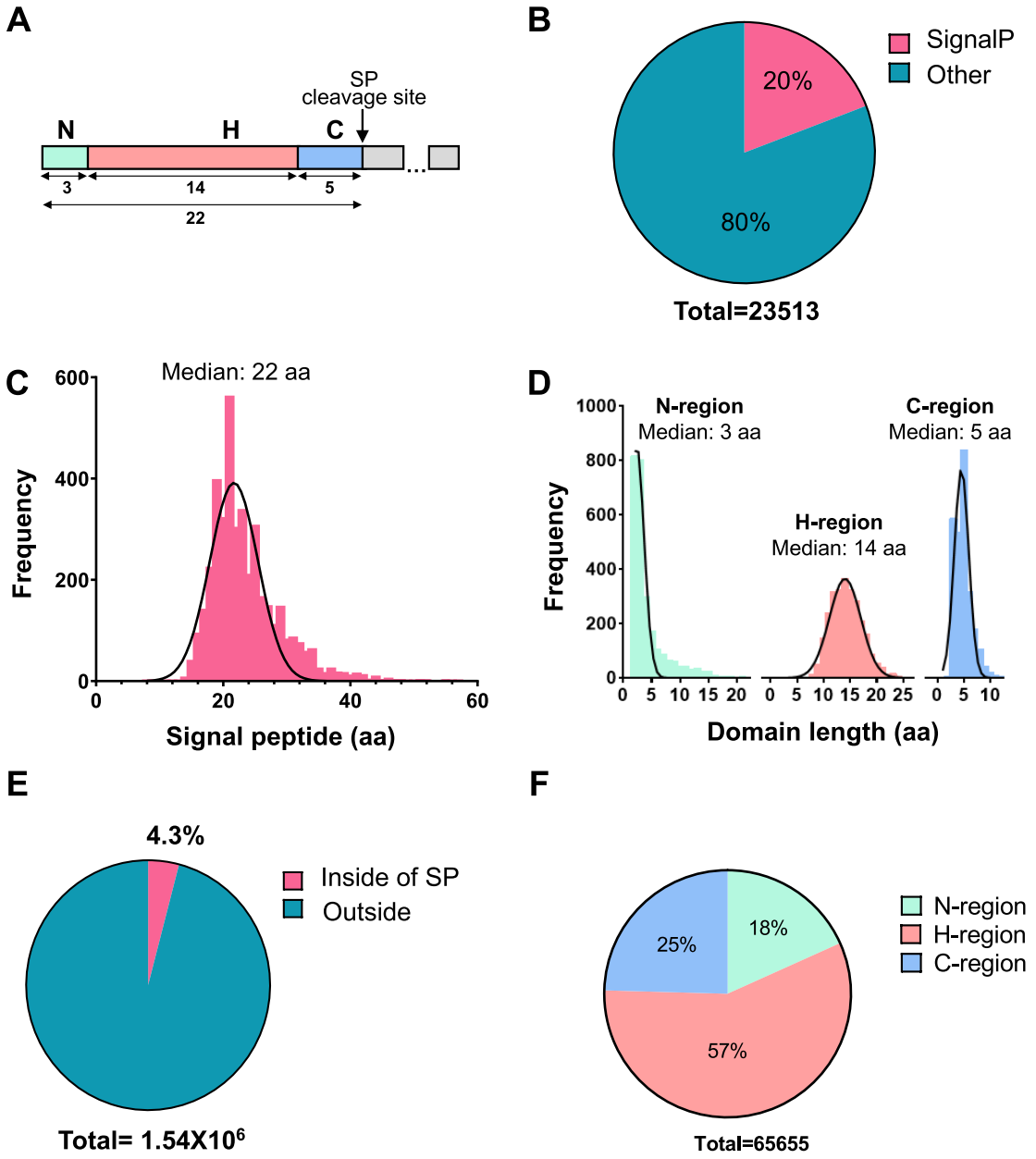


Figure 2. Detection of signal peptides and signal peptide missense variants at the whole human genome. (**A**) Graphical representation of signal peptide regions: N-terminal region (N), hydrophobic region (H), C-terminal region with cleavage site for signal peptidase (C). The numbers are median lengths of the human signal peptides and their regions as determined in this work as shown in the figure panels C and D. (**B**) Relative number of proteins containing cleavable signal peptides among other proteins in human proteome. Genes coding proteins containing signal peptide sequences were selected from the UniProtKB/Swiss-Prot and verified by SignalP 6.0. The number of proteins were determined by the exclusive stable Ensembl peptide identifiers. (**C**) Distribution of signal peptides relatively to their length in amino residues among all signal peptides. (**D**) Frequency distribution of signal peptide regions per amino acid sequence length in all signal peptides. (**E**) Percentage of missense variants detected inside and outside of the signal peptide coding sequences. (**F**) Distribution of the detected signal peptide missense variants between different signal peptide regions. The signal peptide variants include those located in the position +1 because it may affect signal peptide cleavage (protein processing).

(15,24,25,33,34) (Figure 1B). Other quality control mechanisms in the cytosol involve the recognition of aberrant proteins that have already been synthesized and released from the ribosome. These are known to be targeted for degradation by the ubiquitin/proteasome system in the cytoplasm (35). At the ER, misfolded or accumulated proteins activate the Unfolded Protein Response (UPR), or ER-associated degradation (ERAD) (32,36–38). Unprocessed precursors due to mutations at the signal peptidase cleavage site or in the N-region are modified with ubiquitin and subjected to degradation by the proteasome, reviewed in (39).

The association between signal peptide variants and human diseases is likely underestimated. While there are several reports of missense variants in signal peptides that are associated with human diseases, researchers have experimentally validated only a few variants (7,15,24,40–55). Since it is technically impossible to study the contribution of all signal peptide variants to the pathogenesis of human diseases,

strategies of classification and prediction are needed to identify pathogenic signal peptide variants. Here, we report single nucleotide polymorphisms (SNPs) in signal peptide coding sequences across the human genome and predicted their pathogenicity. Using a rational stepwise classification strategy, we identified more than eleven thousand predicted pathogenic variants (PPVs) and their possible molecular mechanisms in a number of human diseases. We validated some PPVs by computational modeling of their interactions with SRP compared to the wildtype sequence. Our analysis provides a conceptual framework for how PPVs may alter the targeting, translocation, and processing of more than three thousand membrane and secreted proteins. Identifying PPVs in signal peptides will contribute to our understanding of the secretory pathway and the genetic basis for human diseases.

Materials and methods

Identification of human proteins with signal peptides

Gene-coding proteins with signal peptides were identified by using the UniProtKB/Swiss-Prot database. The search comprised the following parameters: 'annotation:(type: signal) AND reviewed: yes AND organism: 'Homo sapiens (Human) [9606]' (56). These proteins were annotated with stable ensemble protein IDs (https://m.ensembl.org/info/genome/stable_ids/index.html) and the amino acid sequences were recovered as .fasta format by using the function 'getSequence' of the BiomaRt package (57). The SignalP 6.0 slow mode algorithm through Python Programming Language 3.10.0 was used to predict the presence of signal peptides and signal peptide regions (https://www.python.org/) (58). See Supplementary File S1.

Identification of human signal peptide variants

To identify genetic variants associated with signal peptides, the single nucleotide variation database (dbSNP) from NCBI was used (59). The SnpEff algorithm predicted each variant's effect at the protein level, using the GRCh38.p13 human genome for reference (60), Ensembl's Variant Effect Predictor algorithm selected missense variants that affected canonical proteins (61).

The generated datasets containing (i) signal peptide proteins and (ii) human missense variants were matched based on the Ensembl protein IDs. The variants that fell within the signal peptide protein sequence were selected. The last amino acid residue in the signal peptide just before the cleavage site was denoted as '–1', and the first amino acid residue in the mature part as '+1' (the location of the signal peptide cleavage site is between the –1 and + 1 amino acid positions). The +1 amino acid residues were included in the analysis of wild type signal peptides. The mutant versions were built by replacing the reference (wildtype) amino acid with the respective amino acid incorporated by the missense variant in Microsoft Excel.

Identification of signal peptide pathogenic variants

The identification of a predicted pathogenic variant (PPV) included a stepwise strategy of classification performed partially in the R language and with Microsoft Excel (Supplementary Files S3, S9). Three parameters were defined to detect missense variants that interfere with signal peptide recognition in the hydrophobic core:

- missense variants that modify the hydrophobic core (H-region);
- variants that decrease the hydrophobicity;
- variants that reduce the potential binding between the signal peptide and SRP.

First, missense variants outside of the H-region were filtered out. Next, the change in hydrophobicity was estimated by subtracting the hydrophobicity of the amino acid of the variant from the amino acid of the wild type. The hydrophobicity estimation was based on the Kyte & Doolittle scale (62). Signal peptides with significant hydrophobicity decreases were considered for further analysis of the potential change in SRP interaction via the Boman Index. The potential protein interaction index, or the Boman Index, was used to measure the variant's overall impact on the signal peptide interaction with SRP. This parameter was estimated by subtracting the Boman Index of the wild type amino acid from its variant, followed by multiplication by 100 for scaling (63). Based on previous experiments (15), the change in the Boman Index $\leq$–20 was used as the cutoff for possible SRP recognition failure, RAPP activation, and induced mRNA decay. The variants that fulfilled these three parameters were classified as PPVs in the H-region.

Detection of variants leading to signal peptidase failure was based on the selection of missense variants modifying the signal peptide C-region or the position + 1. The amino acid position '+1', '–1' and '–3' were identified based on the cleavage site. The predicted variants incorporating less frequent amino acids and potentially affecting signal peptidase recognition were classified as PPVs.

To identify pathogenic variants in the N-region, the missense variants that incorporate negatively charged amino acids aspartate (D) and glutamate (E) instead of polar, nonpolar, or positively charged amino acids were also classified as PPVs.

Computational modelling of the signal recognition particle targeting a signal peptide

DeepMind's AlphaFold (64,65) was used to predict how SRP54 and the signal peptide dock *de novo*. AlphaFold uses machine learning and artificial intelligence to predict secondary, tertiary, and quaternary structures from the primary amino acid sequence. AlphaFold minimizes the root mean square deviation (RMSD) between atoms. RMSD is ultimately a measure of accuracy, the square root of the differences between predicted and observed values. RMSD values closer to or below 0 indicate better models than those above 0. Thus, minimizing RMSD leads to more optimal models. Rosetta Online Server that Includes Everyone (ROSIE) uses AlphaFold's model as a starting point before predicting how SRP54 and the signal peptide interact by further minimization of the RMSD and its own internal Rosetta Total Score generated by 1000 model iterations (66,67).

Models are submitted to ModelArchive (modelarchive.org). The individual links for the models are:

SRP54 with ALK WT signal peptide: modelarchive.org/doi/10.5452/ma-owxf7

SRP54 with ALK W8R signal peptide: modelarchive.org/doi/10.5452/ma-w701z

SRP54 with ALK S15Y signal peptide: modelarchive.org/doi/10.5452/ma-cm6gn

The models were further validated according to CAPRI quality assessment criteria with correct models defined as follows:

Fnat (frequency of native contacts) > 0.1 -OR-

Ligand rms (root mean squared) < 10.0 -AND- interface rms (irms) < 4.0

Supplementary File S11 provides validation of SRP54 Rosetta models with different ALK signal peptides. SRP54 with ALK WT and SRP54 with ALK S15Y are valid as models, but SRP54 with ALK W8R is not valid as a model that demonstrates SRP54 and ALKW8R do not interact, and, thus, in agreement with our prediction.

Molecular images

The molecular images were created in PyMol (68). The SRP subunits were aligned on 7SL RNA using PDB coordinates 1RY1, 5WRW, 5WRV, 4P3F and 1MFQ (69–72). The image of the human signal peptidase complex was prepared by using coordinates 7P2P (73).

Association of signal peptide variants and human diseases

Genes with a PPV were surveyed for their association with human disease. The analysis was based on the Genetic Association Database (GAD) using the DAVID Bioinformatics Resources 6.8, NIAID/NIH. We filtered significantly enriched diseases and disease categories by a false discovery rate lower than 0.05. We estimated the association between a decrease in protein expression and disease or cell phenotype using the function 'BioProfiler' in IPA software (Qiagen, Version: 763620684).

Association of signal peptide genes and biological processes

The set of genes with PPVs was analyzed based on Protein ANalysis THrough Evolutionary Relationships, a bioinformatical analysis tool that groups genes based on evolutionary relationships (74) (Supplementary File S10). We used a cutoff of 50 genes to define a group of genes that clustered for a biological process.

Results

Signal peptides in human proteome

There were 3607 genes encoding proteins with annotated signal peptides in the UniProtKB/Swiss-Prot database (56) (Supplementary File S1 UniProt). We assigned 4504 different proteins stable Ensembl peptide identifiers to avoid redundancy (see Methods) and to accurately denote different isoforms (Supplementary File S1 Ensembl). To validate the presence of signal peptides in the recovered amino acid sequences and to determine the signal peptide regions, we analyzed these by the predictive algorithm SignalP6.0 that was recently published using the slow mode (58). As a result, we confirmed the presence of 4142 different proteins with signal peptides (Supplementary File S1 SignalP6.0Slow). We chose the most abundant protein isoforms as the reference for canonical transcripts. Proteins with cleavable signal peptides contribute to 20% of the total human proteome (Figure 2B, Supplementary File S1). Signal peptides are variable in length with a median of 22 amino acids (Figure 2A, C). As previously mentioned, signal peptides have conserved functional domains. To annotate these domains for each signal peptide, we used the slow mode algorithm of SignalP6.0, which defines the borders of each of the three domains for each signal peptide (58). These domains vary in length with median values of 3 residues for the N-region, 14 residues for the H-region and 5 residues in the C-region (Figure 2A, D, Supplementary File S1). The C-region is the most conserved in length. Furthermore, signal peptide regions are characterized by specific features. The N-region contains 16.5% of positively charged amino acids, with almost three times more arginine than lysine. Other amino acids such as alanine, glycine, and proline are also frequently observed in the N-terminal region with a frequency of 9.78%, 8.24% and 9.22%, respectively. The least frequent amino acid in the N-region is tyrosine (0.28%). The central H-domain predominantly contains hydrophobic amino acids such as leucine (36.93%), valine (8.8%), and alanine (9.68%), while aspartate (0.21%), asparagine (0.39%), and lysine (0.41%) are rare. Finally, the C-region contains mostly alanine (20.14%), glycine (16.16%), and serine (11.06%), while the presence of methionine (0.9%), phenylalanine (0.99%), and asparagine (1%) are rare (Supplementary File S2). Overall, these results highlight that although human SPs are variable in amino acid sequences, signal peptide regions preferentially contain specific groups of amino acids that support conserved functional features.

Missense variants modifying signal peptides

To detect signal peptide variants, we used the dbSNP-NCBI repository to recover all reported human single nucleotide polymorphisms (SNPs) (75). We further categorized these SNPs by effect: synonymous, missense, upstream, downstream, in-frame and out-of-frame shifts, and deletions using SnpEff 5.0 software (60). We selected missense variants and annotated these SNPs with the Variant Effect Predictor algorithm (61). Out of 1 540 002 missense variants in genes coding for proteins with signal peptides, 65 655 or over 4% of missense variants were actually located within the signal peptides of 3506 unique proteins (Figure 2E, Supplementary File S3). This means that 82% of signal peptide-containing proteins have at least one missense mutation in their signal peptide (Supplementary File S3). The distribution of missense variants showed that 12 003 change the N-region; 37 526 change the H-region; and 16 126 change the C-region. (Figure 2F, Supplementary File S3). Most variants (57%) were found in the H-region, the lengthiest domain in signal peptides.

As stated above, we used two approaches for detection of signal peptides in the proteins, UniProt and SignalP6.0 (slow mode). Comparing signal peptides detected by these methods, we observed that while the majority (83%) of the signal peptides are the same regardless of the method used, 17% of signal peptides have alternative lengths (see Supplementary Text and Supplementary Figure S1). Moreover, we observed that UniProt and SignalP6.0 distributed residues differently between the H- and C-regions. Although it is impossible to evaluate whether UniProt or SignalP6.0 is more precise without reliable experimental data, our results suggest that the research groups working with secretory proteins need to pay careful attention to the algorithms used to predict signal peptides. Although we completed analysis of all alternative variants (see Supplementary Text, Supplementary Files and Supplementary Figures S1–S3), we chose to utilize SignalP6.0

exclusively as it is the most up-to-date software for signal peptide prediction.

Numerous variants modifying the signal peptide H-region are predicted to disrupt SRP recognition and activate RAPP

The central hydrophobic signal peptide H-region contains most of the missense variants (*n*: 37 526, or 57%) (Figure 2F). This region is crucial for signal peptide interaction with SRP, the first step for targeting secreted proteins to the ER (23) (Figure 1A). Previously, we have shown that mutations that reduce the hydrophobicity of the H-region impair SRP recognition and lead to mutant protein mRNA degradation through the RAPP pathway (15,23–25) (Figure 1B). To predict mutations that activate RAPP, we selected missense variants in the H-region protein coding sequence (Figure 3A) and classified these variants by their effect on H-region hydrophobicity. SRP recognition of the H-region is impaired when changes between wildtype and mutant signal peptides result in a notable hydrophobicity decrease (Figure 3B). In contrast, SRP recognition still occurs when hydrophobicity increased or slightly changed. We previously demonstrated with a range of mutations that the more drastic a hydrophobicity change in the H-region causes a loss of SRP client mRNA (15). We applied the Boman Index (63), a parameter to estimate protein binding, to predict SRP interaction with each signal peptide H-region variant. The Boman Index is the sum of the solubility values for all amino acid residues in a protein sequence divided by the number of residues. In our analysis, Boman Index provides an overall estimate of peptide binding to SRP. When the Boman Index value is high, the protein has a high binding potential. Based on our previous experimental analyses of disease-causing mutations in the signal peptides (15), we found that changes in the Boman index of these secretory and membrane proteins correlate well with the observed changes in the mRNA expression (Figure 3C). Therefore, changes in the Boman index score induced by signal peptide variants can be used to predict decay of the mutant protein mRNAs. As a result, we identified 8539 missense variants, or 23% of total H-region missense variants which substantially decreased both the hydrophobicity and the signal peptide-SRP interaction and classified them as PPVs due to their propensity to activate RAPP (Figure 3D, Supplementary File S3 PPV Class).

Variants at or near the signal peptidase cleavage site leading to preprotein processing failure

We identified 16 126 missense variants in the C-terminal region and +1 position upstream of the cleavage site (Figure 4, Supplementary File S3 PPV Class). While variants in the signal peptide H-region can impair recognition by SRP, variants in the signal peptide C-region can affect the cleavage of the signal peptide that is required to generate mature proteins (Figure 1C). The signal peptide C-region is characterized by the (–3, –1) rule which specifies restrictions for the amino acids at –3 and –1 positions near the cleavage site (–1 is the amino acid on the N-terminal side of the cleavage site while +1 is the amino acid on the C-terminal side of the cleavage site) (22). In our analysis, positions –3 and –1 are notably more conserved (e.g. alanine consists of >25% of residues in –1 and –3) than position +1 (Figure 4B), which agrees with the '–3, –1' rule that small, neutral amino acids are predominant in these positions (Figure 4C). We used residue abundance at each position in the C-region in the wild type sequence to predict missense variants that lead to impaired signal peptide cleavage (Figure 4D). For example, tyrosine is a rare (<0.5%) amino acid in the –3, –1 and +1 positions (Figure 4C); thus, a missense mutation to tyrosine would be considered a PPV in any of those positions. As a result, 2267 missense variants (14% of all missense variants detected in the C-region) were retrieved and classified as PPVs. We show their distribution among amino acid positions in the C-region in Figure 4E.

To validate PPVs in the C-region, we compared our prediction with the data available in the ClinVar repository. ClinVar archives the relationships between human genetic variation and phenotypic expression with references and automatically archives any variant reported in other databases. Using this tool, we observed that 16 out of the 19 C-region variants (84%) in the ClinVar repository affect the –3 and –1 positions, and we identified 12 of them (75%) as PPVs by our strategy (Supplementary file S4).

Signal peptide N-region variants that likely affect protein translocation through the SEC61 translocon

While the roles of the signal peptide H- and C-regions in SRP recognition and signal peptide cleavage are relatively well established, the function of the N-region is not well understood. Residues in the N-region affect the secretion efficiency of bacterial proteins (76), and the presence of positively charged amino acids (lysine and arginine) in this region is important for correct orientation of preproteins in the SEC61 translocon in eukaryotes (77). The positive charge of the N-region is also particularly important for the efficient translocation of small secretory proteins in humans (28). Thus, missense variants affecting the positive charge of the N-region likely result in decreased translocation efficiency and contribute to disease (Figures 1D, 5A). The association of N-region variants with clinical disease is less evident than for the H-region or the C-region, and most of the negatively charged N-region variants in the ClinVar repository annotated 'uncertain significance' due to limited data available. We identified in wild-type N-region sequences that acidic residues are not common (Figure 5B). Further, our analysis revealed 12 003 missense variants in the N-region, and 705 (6%) of these introduced negatively charged amino acids and were identified as PPVs (Figure 5C). Although the clinical data for signal peptide N-domain variants are still minimal, our evaluation provides a concept for contribution of mutations in this region to human diseases.

Computational modeling of SRP and signal peptide interactions

To test how mutations in signal peptides potentially affect interaction with SRP and verify our prediction of the pathogenic variants and their possible molecular mechanisms, we created *in-silico* models to show the interaction of wild-type and mutated signal peptides with SRP54, a subunit of SRP. We selected ALK protein (ALK receptor tyrosine kinase) for this analysis. ALK is a representative membrane protein, and as we found, it is one of the proteins containing multiple mutations in the signal peptide (Supplementary Files S3 and S5). Our model used amino acids 1–20 of the ALK signal peptide; this region covers the entire N- and H-regions and a part of the C-region as determined by SignalP6.0 (Figure 6A). We chose two different mutations: W8R, which is predicted to affect hydrophobicity dramatically and is designated a PPV

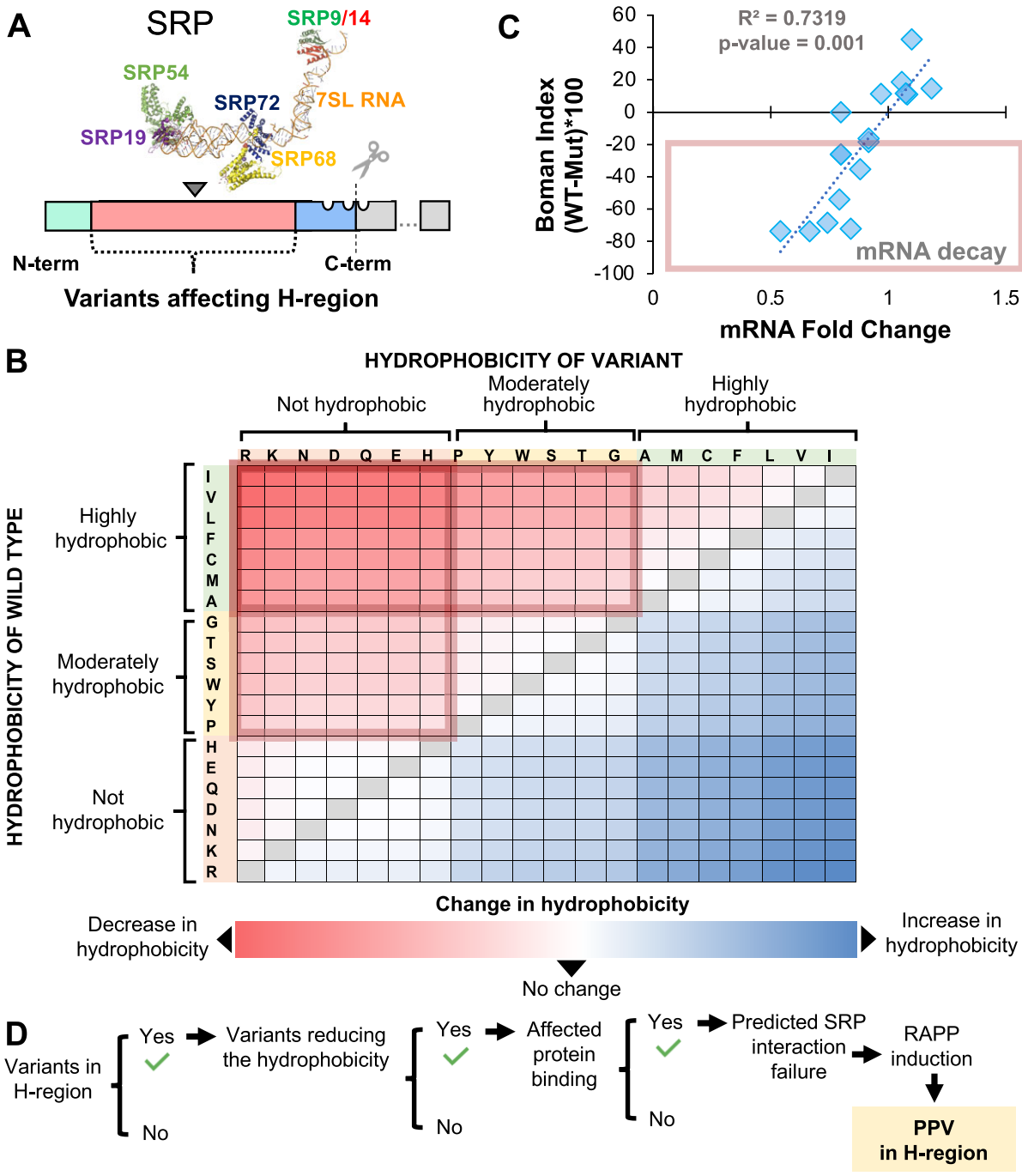
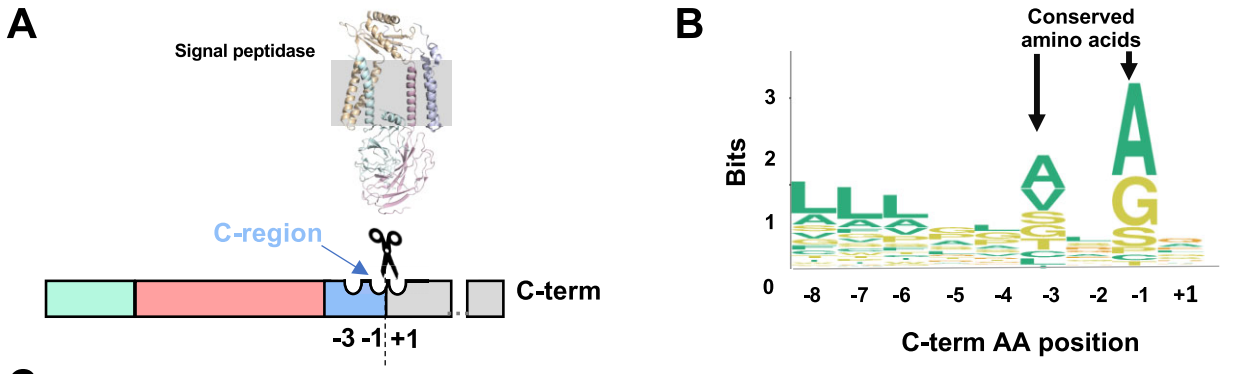
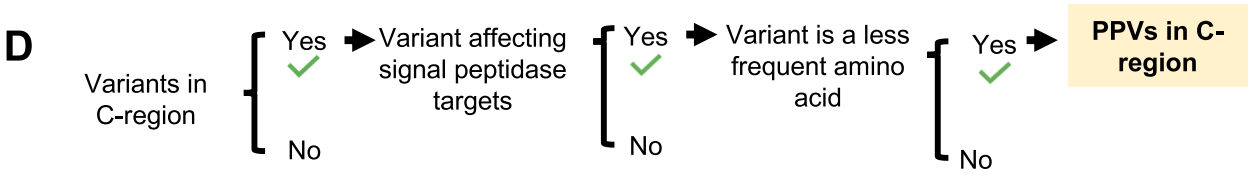


Figure 3. Gene variants affecting the signal peptide H-region can be pathogenic by inducing SRP recognition failure. (**A**) Schematic representation of positions of missense variants modifying the signal peptide H-region sequence. These variants can affect interaction with SRP. SRP image was created in PyMol by aligning the SRP subunits on 7SL RNA. The SRP subunits are marked. (**B**) Hydrophobicity scale of amino acids substitutions resulted from missense variants relatively to wildtype amino acids in signal peptides. Color gradient represents the effect on the hydrophobicity after replacing wildtype amino acid with a mutant variant—blue is high and red is low hydrophobicity. The scores of hydrophobicity per amino acid were estimated by Kyte-Doolittle scale. (**C**) Missense variants potentially reducing the SRP interaction and activating protein's mRNA degradation via RAPP pathway. The changes in the potential protein interaction across signal peptide sequences (Boman index, Y-axis) positively correlate with the changes in protein's mRNA levels experimentally detected (X-axis). Correlation analysis was completed by Pearson test. *R*-squared: 0.732. *F* distribution value: 40.95. Freedom degrees of numerator: 1. Freedom degrees of denominator: 15. *P* value: <0.0001. (**D**) Summary of the strategy used for the identification of PPVs affecting signal peptide H-region.



C

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Unknown
-3	29.91%	5.11%	0.17%	0.15%	0.24%	9.61%	0.22%	3.82%	0.12%	3.94%	1.45%	0.29%	0.70%	0.19%	0.22%	11.86%	9.70%	22.02%	0.15%	0.12%	0.00%
-1	44.67%	4.60%	0.10%	0.24%	0.14%	25.51%	0.17%	0.00%	0.02%	0.39%	0.51%	0.07%	6.99%	1.08%	0.24%	11.04%	3.71%	0.46%	0.05%	0.00%	0.00%
+1	10.32%	2.68%	5.01%	9.31%	2.58%	9.84%	6.41%	1.59%	3.30%	6.58%	0.92%	1.81%	0.53%	18.97%	4.99%	6.39%	3.18%	2.99%	0.84%	1.76%	0.00%



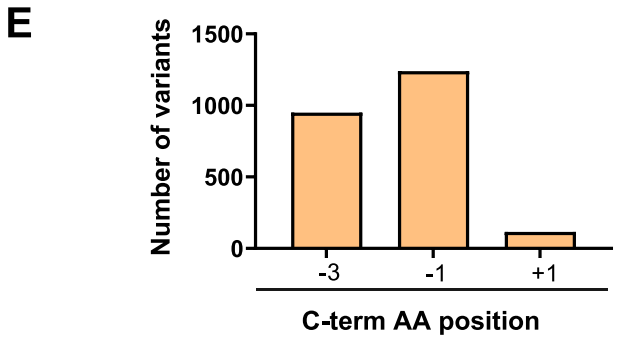


Figure 4. The incorporation of less frequent amino acids in signal peptidase cleavage region potentially leads to signal peptide processing failure and disease. (**A**) Schematic representation of the signal peptidase targeting the signal peptide C-region. The structure of signal peptidase complex is visualized within the ER lipid bilayer. (**B**) Protein sequence logo plot. Amino acid postions respectively to the signal peptide cleavage site are shown on the X-axis. The height of amino acid symbols within the stack indicates the relative frequency of each amino per position measure in bits (Y-axis). (**C**) Table of the relative frequency of each amino acid per signal peptide position. Less to more frequent amino acids are indicated with blue and orange, respectively. Middle values are indicated with white. (**D**) Summary of the strategy used for identifying PPVs that potentially affect the signal peptide cleavage by signal peptidase. (**E**) Distribution of PPVs per signal peptide amino acid position.

according to our algorithm, and S15Y, which does not decrease hydrophobicity and is not a PPV (Figure 6). We used ROSIE (Rosetta Online Server that Includes Everyone) (66,67,78) to determine the best protein folding model and find whether the H-domain of ALK signal peptides will dock into the SRP54 M-domain hydrophobic pocket. We graphed the distribution of the top 100, and 1000 models in Rosetta Total Score and root-mean-square-deviation (RMSD) coordinates in Figure 6B–D, central panels. Lower RMSD values indicate more similar structures to the reference, and we indicated the most minimized, and therefore more reliable, models. Comparing models of SRP54 with WT and S15Y signal peptides, we observed that both signal peptides (WT and S15Y) are predicted to be in the hydrophobic pocket of SRP54 (bottom left dots in the graphs, Figure 6B, C, central panels). In contrast, the ALK W8R mutant does not show models in the bottom left of the graph and therefore does not have a structure that matches the reference inside the SRP54 hydrophobic pocket (Figure 6D, central panel). We then used PyMol (68) to visualize the indicated top Rosetta models to determine

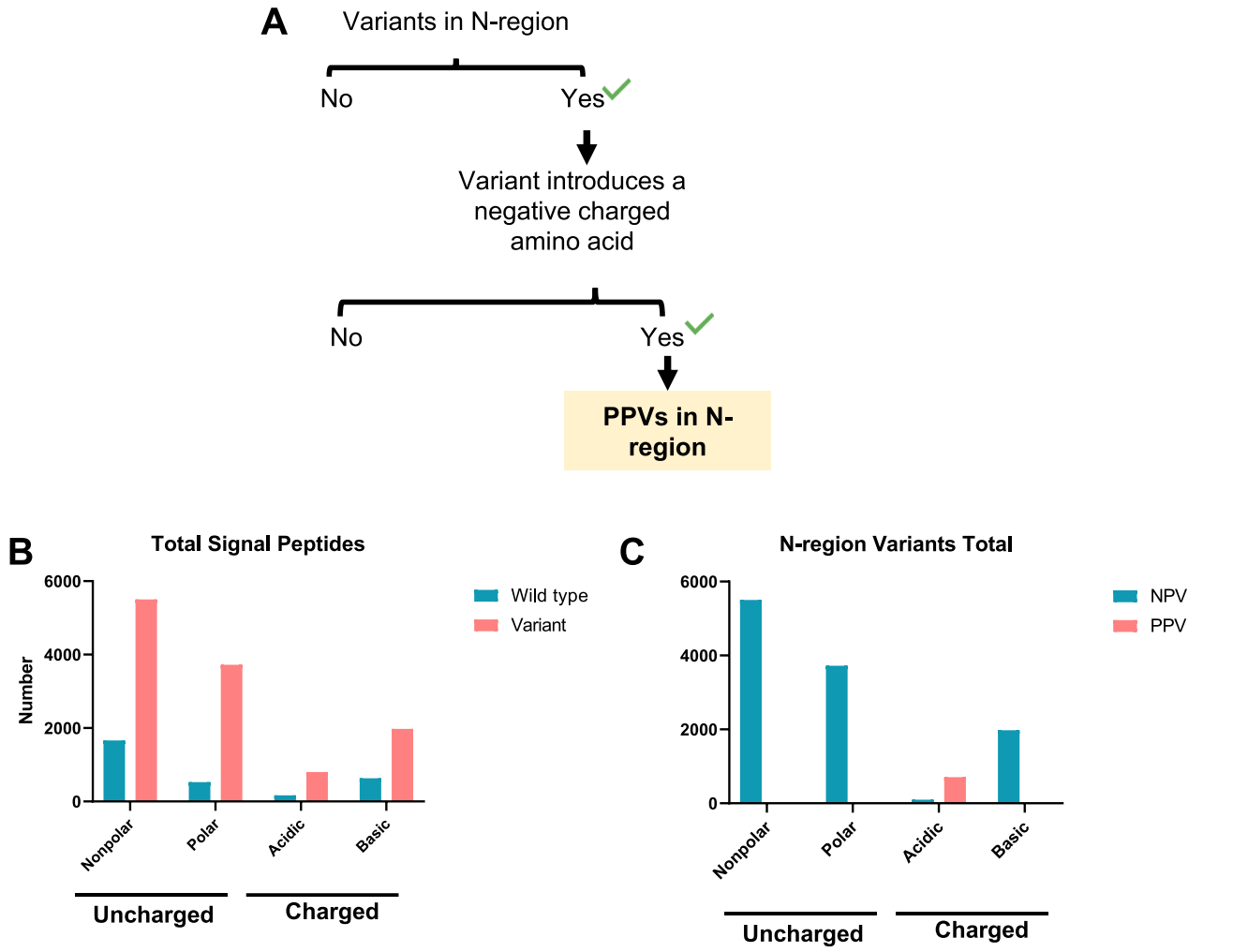


Figure 5. Variants introducing negatively charged amino acids in the signal peptide N-region are potentially pathogenic. The presence of positively charged amino acids is a main trait of signal peptide N-region. The N-terminal positive charge is required to orientate the nascent polypeptide across SEC61 translocon. The incorporation of negatively charge amino acids potentially impair signal peptide orientation. (**A**) Strategy for selecting PPVs in the N-region. (**B**) Bar chart summarizing the non-polar, polar, negative and positive amino acids detected in all wildtype and mutated signal peptides. (**C**) Bar chart summarizing the total variant amino acid distribution in non-pathogenic and pathogenic variants.

where the signal peptides are docking with SRP54 (Figure 6B–D, right panels). Supplementary File S11 provides validation of SRP54 Rosetta models with different ALK signal peptides. SRP54 with ALK WT and SRP54 with ALK S15Y are valid as models, but SRP54 with ALK W8R is not valid as a model demonstrating that SRP54 and ALKW8R do not interact, and, thus, in agreement with our prediction.

Our analyses demonstrate that while wild-type and S15Y variant signal peptides successfully dock in the SRP54 hydrophobic pocket crucial for signal peptide recognition, W8R does not (Figure 6B–D, right panels). Earlier, we showed that the loss of SRP interaction with signal peptide triggers RAPP protein quality control leading to mRNA degradation. Thus, we propose that activation of the RAPP pathway and loss of the ALK expression is a molecular mechanism of pathologies associated with W8R mutation. Thus, using the *in-silico* molecular modelling, we are able to distinguish between predicted pathogenic and non-pathogenic variants. This approach may be helpful for a detailed evaluation of mutation outcomes for other proteins in wide applications.

Validation of PPVs by experimental data extracted from the literature

To verify that our strategy can identify variants that lead to RAPP and induce mRNA decay, we searched the literature for studies that report mRNA expression for proteins with signal peptide variants. Though few studies quantitatively analyzed signal peptide mutant mRNA and protein expression, our predictions match the experimentally evaluated mRNA levels expressed in different mutant cell lines (Table 1).

Once we classified variants as PPV, we compared our list with the information available in ClinVar. Based on clinical data, this database classifies human variants in terms of their pathogenic effect. As a result, of the 43 verified pathogenic variants that affect the signal peptide H-region, 30 (~70%) were correctly identified as PPV by our algorithm (Supplementary file S4). Thus, we could predict most of the clinical variants modifying the signal peptide H-region.

Together, our data demonstrate that the proposed bioinformatic strategy can be used to identify human signal peptide variants which impair SRP recognition and activate protein quality control pathways leading to different human diseases.

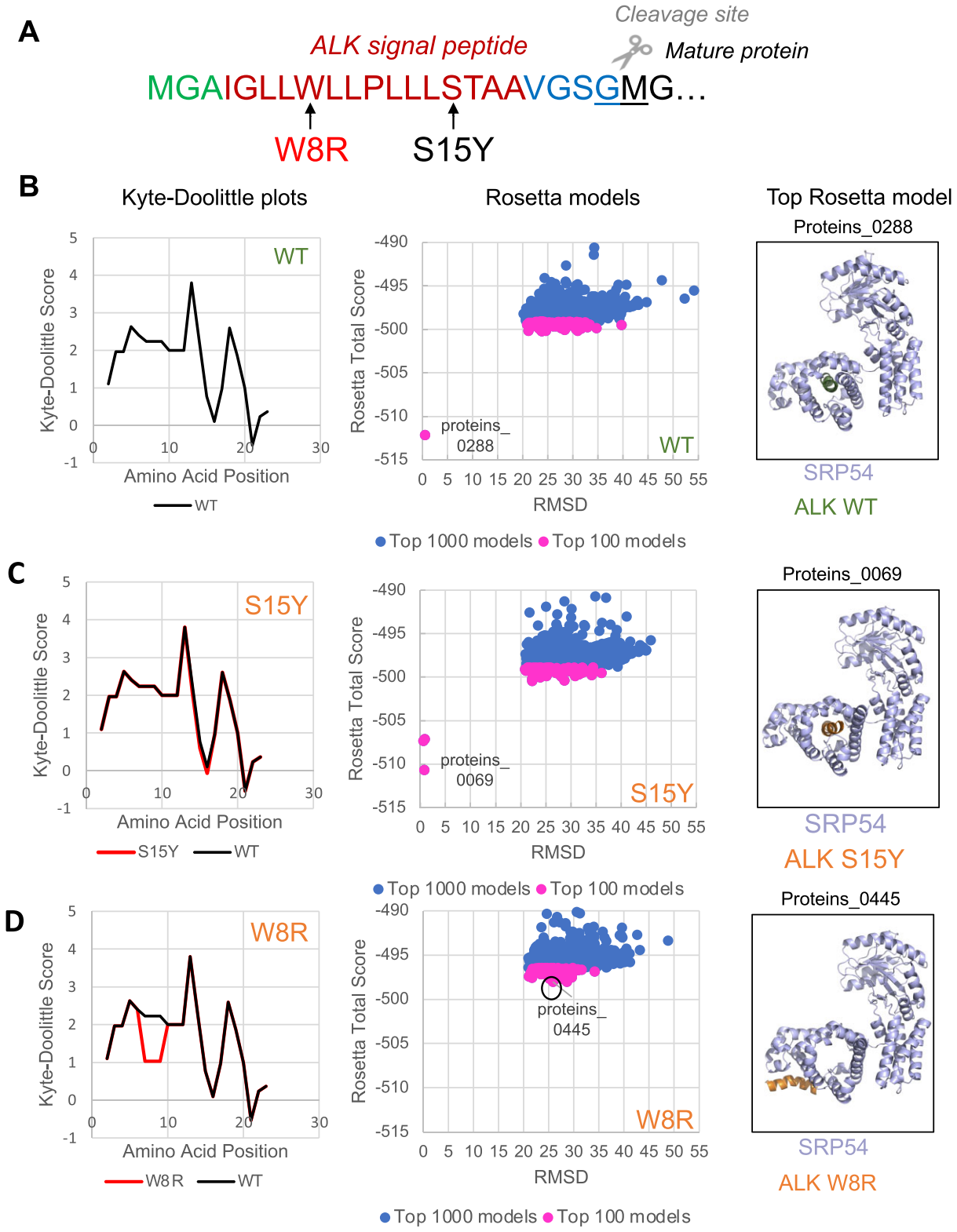


Figure 6. *in silico* molecular modeling of SRP54 interactions with signal peptides. (**A**) Positions of mutations W8R (PPV, marked in red) and S15Y (NPV, marked in black) in the signal peptide of ALK receptor tyrosine kinase. Signal peptide was predicted by SignalP6.0. (B–D) Modeling of the wild-type ALK signal peptide (**B**) and the mutants, S15Y (**C**), W8R (**D**). The hydrophobicity of each signal peptide was determined using the Kyte-Doolittle scale and shown in the left panels. Signal peptide mutants (red line) are compared to the WT ALK signal peptide (black line) to determine the predicted change in hydrophobicity. The top 100 (pink) and top 1000 (blue) models of the corresponding signal peptides and SRP are created using ROSIE (Rosetta Online Serve that Includes Everyone) and their distribution is shown in the central panels. The top models, which are most minimized, are labeled. These top models were selected and visualized by the use of PyMol and shown in the right panels. Wild-type signal peptide (WT) is shown in green, mutated signal peptides are in orange, and M-domain of SRP54, a subunit of SRP, is shown in light blue.

Table 1. Predicted and observed mRNA expression levels for signal peptide variants

Gene symbol	Protein length	Variant	Amino acid modification	Wild type signal peptide	Mutant signal peptide	Prediction	Experimental evidence	Cell line	Reference
AGA	346	rs386833429	15(L/R)	MARKSNLPVLLVPF**L**LCQALVRC|S	MARKSNLPVLLVPF**R**LCQALVRC|S	mRNA decay	Decreased mRNA	HeLa	(15)
CTSK	329	rs1057517252	7(L/P)	MWGLKVLL**L**PVVSFA|L	MWGLKV**P**LLPVVSFA|L	mRNA decay	Decreased mRNA	HeLa	(15)
CTSK	329	rs1057517252	9(L/P)	MWGLKVLL**L**PVVSFA|L	MWGLKVLL**P**PVVSFA|L	mRNA decay	Decreased mRNA	HeLa	(15)
GRN	593	rs63751243	9(A/D)	MWTLVSWV**A**LTAGLVAG|T	MWTLVSWV**D**LTAGLVAG|T	mRNA decay	Decreased mRNA	HeLa	(24)
INS	110	rs121908259	6(R/H)	MALWM**R**LLPLLALLALWG PDPAAA|F	MALWM**H**LLPLLALLALWG PDPAAA|F	No mRNA decay	No change in mRNA	HEK293	(51)
INS	110	rs121908278	6(R/C)	MALWM**R**LLPLLALLALWG PDPAAA|F	MALWM**C**LLPLLALLALWG PDPAAA|F	No mRNA decay	No change in mRNA	HEK293	(51)
LHCGR	699	rs4539842	16(L/Q)	MKQRFSALQLLKLLLLLQ PPLPRA|L	MKQRFSALQLLKLLL**Q**LQ PPLPRA|L	mRNA decay	Decreased mRNA	HEK293T	(52)
LHCGR	699	rs917607255	10(L/P)	MKQRFSALQ**L**LKLLLLLQ PPLPRA|L	MKQRFSALQ**P**LKLLLLLQ PPLPRA|L	No mRNA decay	No change in mRNA	COS-7	(53)
NDP	133	rs104894879	13(L/R)	MRKHVLAASFSM**L**SLLVI MGDTDS|K	MRKHVLAASFSM**R**SLLVI MGDTDS|K	mRNA decay	Decreased mRNA	HeLa	(15)
POMC	267	rs779629993	15(A/G)	MPRSCCSRSGALLL**A**LLLQ ASMEVRG|W	MPRSCCSRSGALLL**G**LLLQ ASMEVRG|W	No mRNA decay	No change in mRNA	β-TC3	(54)
PTH	115	rs104894271	18(C/R)	MIPAKDMAKVMIVMLAI**C**F LTKSDG|K	MIPAKDMAKVMIVMLAI**R**F LTKSDG|K	mRNA decay	Decreased mRNA	HeLa	(15)
SERPINE1	402	rs6092	15(A/T)	MQMSPALTCLVLGL**A**LV FGEGSA|V	MQMSPALTCLVLGL**T**LV FGEGSA|V	No mRNA decay	No change in mRNA	HeLa	(15)
TGFB1	390	rs1800470	10(P/L)	MPPSGLRLL**P**LLLPLLWL LVLTPGRPAAG|L	MPPSGLRLL**L**LLLPLLWL LVLTPGRPAAG|L	No mRNA decay	No change in mRNA	HeLa	(15)
UGT1A1	533	rs111033541	15(L/R)	MAVESQGGRPLVLG**L**LLC VLGPVVSHA|G	MAVESQGGRPLVLG**R**LLC VLGPVVSHA|G	mRNA decay	Decreased mRNA	HeLa	(15)
LIPA	399	rs1051338	16(T/P)	MKMRFLGLVVCLVLW**T**LH SEGSG|G	MKMRFLGLVVCLVLW**P**LH SEGSG|G	No mRNA decay	Increased mRNA	Human monocytes (CD14+)	(55)

Variants affecting the H-region were classified by their predicted effect on the mRNA level based on the potential RAPP induction. The predicted variant effects were compared with published experimental data. All variants matched the predicted and observed results. Aspartylglucosaminidase (AGA), cathepsin K (CTSK), granulin precursor (GRN), insulin (INS), luteinizing hormone/choriogonadotropin receptor (LHCGR), lipase A (LIPA), norrin cystine knot growth factor (NDP), proopiomelanocortin (POMC), parathyroid hormone (PTH), serpin family E member 1 (SERPINE1), transforming growth factor beta 1 (TGFB1), UDP glucuronosyltransferase family 1 member A (UGT1A1). Sequences presented are signal peptides plus one amino acid, the cleavage site is marked by a red line. Changed amino acids in mutant signal peptides are in red bold font, corresponding amino acids in the wild-type signal peptides are in black bold font.

Predicted pathogenic variants in human diseases

Using our strategy and taking into account all signal peptide regions, we identified 65 655 variants and 11 622 of them as PPVs (Supplementary file S3). These variants affect 3506 genes, and we found PPVs in different locations of the corresponding signal peptides in 3344 of these genes (162 genes of 3506 do not have any classified PPVs) (Figure 7A, Supplementary File S3). Based on our analysis and potential molecular mechanisms (Figure 1), we propose that PPVs associated with the signal peptide N-region affect signal peptide orientation and efficiency of protein translocation through the SEC61 translocon; PPVs associated with H-region induce the RAPP pathway with mRNA degradation of the secreted/membrane protein; and PPVs associated with C-region inhibit signal peptide cleavage affecting protein processing (Figure 7B). To evaluate the potential association of PPVs with particular diseases, we investigated the linkage of these genes with human disease using the Genetic Association Disease Database (https://geneticassociationdb.nih.gov/). Most genes with detected PPV are connected with metabolic and cardiovascular diseases and cancer. Reproductive and vision disease variants had a higher fold of enrichment compared to genes associated with a particular disease (Figure 7C). Remarkably, several PPVs contribute to the most severe human diseases defined in mortality (79). The identified PPVs are associated with the development of coronary heart disease (ACE, THBD, LDLR, etc.), ischemic stroke (F5, GP1BA), Alzheimer's (APOE, CD33, TREM2, SORL1), neonatal conditions(SFTPA1, SFTPA2, SFTPB, SFTPD), lung cancer (SEMA3B, WNT5B, RECK), respiratory infection (ADAM33, CCL1, CXCL1, MUC1), diabetes mellitus (INS, EGFL8, KIR3DL1), diarrheal diseases (LTF, UGT1A7, UGT1A8, UGT1A9, UGT2B7), and kidney disease (ACE, AGT, PXDN, COLEC11). The Supplementary files S6-8 summarize the distribution of genes with associated PPV per illness class and disease.

Furthermore, since PPVs may reduce protein levels and lead to loss of protein function, we also evaluated the association between decreased protein expression and associated diseases and cellular function by using Ingenuity Pathway Analysis (QIAGEN). Based on experimental data, the loss of protein function is mainly associated with increased cell death and cancer (Supplementary file S8). Together, our results demonstrate that the bioinformatics approaches applied in this study allow us to associate PPVs with particular human diseases and connect them with a molecular basis of the disorders based on the severity of the mutation and its position in the signal peptide regions.

Discussion

The association between gene variants of protein targeting signals and human diseases remains poorly explored. In this work, we developed a bioinformatic approach to identify and classify PPVs affecting signal peptide coding sequences. We identified 65 655 missense variants in signal peptides across the human genome and predicted 11 622 of these variants as pathogenic. Our data indicates that previously reported pathogenic variants affecting signal peptides are only a tip of the iceberg – our findings significantly widen the scope of the studies on diseases associated with secretory/membrane proteins. We have highlighted that signal peptides can be impaired differently depending on their affected regions, resulting in distinct molecular mechanisms (Figure 1B, C, Figures 3–5). PPV mechanisms are summarized in Figure 7B. PPVs likely induce loss of protein expression through mRNA degradation, decreased protein translocation efficiency, protein mistargeting, or cleavage inhibition (7,24,50). The predicted pathogenic variants can be used for further analysis with future medical applications, including targeting mutations' effects by drug development.

SRP-signal peptide interaction is an essential step in most secreted and membrane protein biogenesis. Mutations in the SRP subunits are associated with many human diseases (13). The signal peptide H-region plays the most crucial role in the process of signal peptide recognition by SRP (23). The current work identified 37 526 mutations in this region, and classified 8614 as PPVs. We also used computational modeling to simulate SRP interactions with wild-type and mutant ALK signal peptides to demonstrate how PPVs may interact with SRP. These models clearly show that the ALK signal peptide containing a charged amino acid in the H-domain cannot interact with the SRP54 subunit of SRP (Figure 6). We have previously demonstrated that mutations in the signal peptide H-region trigger the RAPP protein quality control leading to the degradation of the mutant protein mRNAs (15,24,25). In the current work, we predict that variants reducing the signal peptide hydrophobicity, and decreasing the potential signal peptide binding to SRP (Figure 3), trigger mRNA decay as a characteristic for the RAPP activation (see comparison of the prediction with published experimental data in Table 1). Thus, pathological activation of the RAPP pathway is the most likely molecular mechanism of these signal peptide variants.

However, even if a mutated secretory protein is still targeted to the ER membrane, it may be translocated inefficiently. Signal peptide hydrophobicity also determines whether signal peptides can autonomously facilitate the opening of the SEC61 translocon ('strongly' gating signal peptides) or if their translocation requires additional components, such as the translocon-associated protein (TRAP), SEC62 or SEC63 (80–83). Compared to strong hydrophobic signal peptides, proteins with weak hydrophobic signal peptides translocate less efficiently. It was shown recently that some proteins with weak hydrophobic transmembrane spans and signal peptides retained at the SEC61 translocon, and they need engagement of SEC63 and BiP for their release from SEC61, their translocation and folding in ER (83). Thus, some mutations that decrease the hydrophobicity of the signal peptide may result in less efficient translocation and may engage other components of the protein transport machinery to compensate the translocation defects or result in degradation of defective proteins.

In contrast, mutations that increase the hydrophobicity of the H-region are predicted to increase the interaction between SRP and the signal peptide. It was demonstrated previously that an increase in signal peptide hydrophobicity also leads an increase in pulling force, suggesting faster translocation through the SEC61 translocon (84). In fact, there appears to be both lower and upper hydrophobicity bounds for signal peptide insertion in the ER membrane. There is a hydrophobicity barrier that has to be overcome for the signal peptide or transmembrane span to be pulled through SEC61 (84–86). It was also shown that increasing the signal peptide hydrophobicity of pseudorabies virus glycoprotein gC impaired proper ER translocation (87). It was shown that highly hydrophobic signal peptides do not engage BiP during translocation, thus,

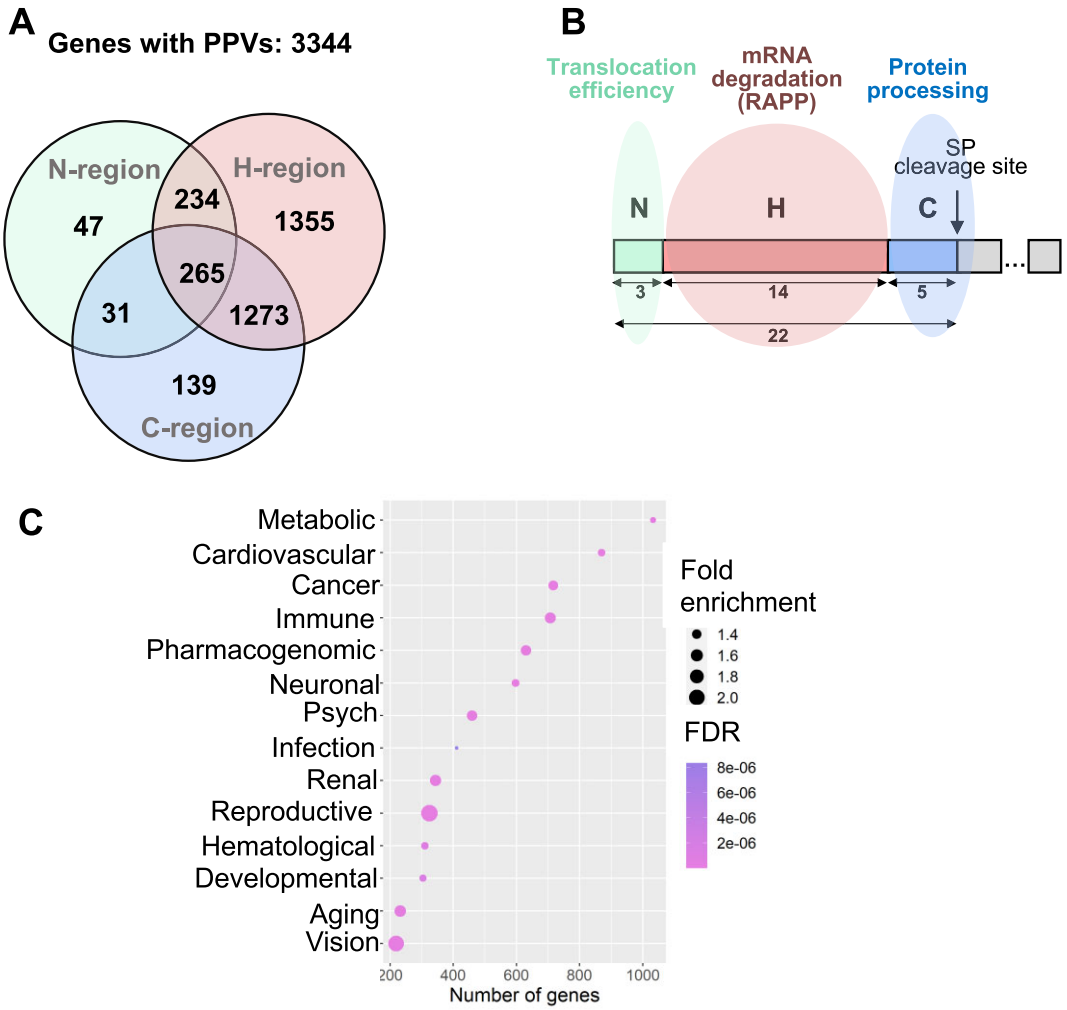


Figure 7. Potentially pathogenic signal peptide variants are connected with disease-associated genes and cause multiple molecular mechanisms. (**A**) Venn diagram summarizing the distribution of genes with PPVs per affected signal peptide region. (**B**) Possible molecular mechanisms of PPVs. Scheme of the typical signal peptide with marked regions is shown. Numbers represent average lengths of the signal peptide or its respected regions in amino acid residues as determined in this work. PPVs, localyzed in the signal peptide N-region, affect protein translocation efficiency through the translocon; PPVs, localyzed in the H-region, trigger mRNA degradation of the secreted/membrane protein through the RAPP pathway; and PPVs localyzed in the C-region inhibit protein processing. (**C**) Dot plot showing the number of genes per disease class. False Discovery Rate (FDR) as raw *P*-value correction. Fold enrichment obtained through comparing the background frequency of total genes annotated per disease class to the sample frequency representing the number of genes inputted that fall under the same disease class.

these proteins may indicate a propensity to misfolding and aggregation in stress conditions (83). Only 11% of H-region variants we analyzed moderately or significantly increased hydrophobicity (Supplementary File S3), and only seven variants were found in ClinVar. Thus, the pathology of mutations increasing signal peptide hydrophobicity is still questionable. However, if some of these variants are associated with a disease, the molecular mechanism is likely due to defects in their translocation and folding, but not related to inefficient recognition by SRP inducing RAPP.

The variants introducing changes to the signal peptide in the C-region, which modify the specific amino acids recognized by signal peptidase (position –3, –1 and +1 with respect to the cleavage site), are also predicted as pathogenic by decreasing the likelihood of correct signal peptide cleavage. We show in the Figure 4B, C, that there are clear preferences for specific amino acids in the positions –3 and –1, while the appearance of the others was scarce. The +1 position was less conservative than the –1 and –3 positions, with only methionine, proline, and tryptophan being relatively rare. Variants incorporating less conserved amino acids are potentially more likely to confer disease risk and were considered PPVs (88–90) (Figure 4C, D). Most PPVs detected in the C-region affect the –1 position (Figure 4E). The –1 position corresponds to the more conserved position in the signal peptide C-region of wild-type sequences (Figure 4B). Most pathogenic variants affecting the C-region in the ClinVar repository are at the –1 position (Supplementary file S4). More information regarding the distribution of amino acids in the C-region, particularly between algorithms and the distribution of PPVs, is presented in Supplementary Files S2 and S3. These results support the idea that human missense variants affecting the amino acids targeted by signal peptidase and particularly in the –1 position promote the retention of membrane/secreted pre-proteins at the ER membrane and, induce protein loss of function and increase the risk of human diseases (26,27). Thus, the molecular mechanism of the pathogenic signal peptide variants near the cleavage site is likely associated with defects in the protein

processing and, consequently, protein degradation and the loss of function.

The signal peptide N-region is less conserved than the C-region. In bacteria, this region determines the efficiency of protein secretion and possible interaction with membrane lipids. Mutations in this region do not dramatically interfere with SRP recognition, although they slightly modulate that process in mammals. Few experimental findings suggest that the correct orientation of signal peptide across the SEC61 translocon requires positively charged amino acids distributed in the N-terminal region. Small proteins are susceptible to losing the positive charge in the signal peptide N-region. As a result, the newly synthesized pre-proteins are accumulated intracellularly because of impaired translocation (28). We propose that variants incorporating negatively charged amino acids in the signal peptide N-region have a higher chance of inducing impaired translocation by affecting the required signal peptide N-terminal positive charge (Figure 5). Indeed, a group of 38 small proteins, including HCRT (narcolepsy, rs1327645071), CCL2 (neural tube defects and HIV, rs898151976), CCL7 (nephrogenic systemic fibrosis and toxic black mold infections, rs1439804640), CCL8 (tenosynovitis and T-Cell non-Hodgkin Lymphoma), INSL5 (Cryptorchidism, rs751653318), IGF1 (Pituitary Gland Disease, rs3730195), GYPA (malaria susceptibility, rs371519566) and others showed gene variants affecting the N-region charge. With our classification system, it will be much easier to identify and validate experimentally whether the human missense variants disturbing the N-region charge affect the translocation of these proteins.

Our findings provide a conceptual map for establishing the molecular basis for many human diseases associated with signal peptide mutations. It can serve as a source of PPVs and their association with the molecular mechanism of diseases for a broad group of academic and clinical researchers. Moreover, our analyses can be used to identify new and currently unknown PPVs using genome sequencing data obtained after this publication. Our classification system provides an important starting point for further models that will need to include other factors such as dominant, recessive, or dominant-negative mutations, variant frequency, gene copy number, mature protein features (size and domains), presence of homologous proteins, the effect of loss of function mutations, SRP dependency, gender, and estimated impact on fundamental biological processes.

Data availability

The analyses are presented in the Supplementary Data to the manuscript and available in the online version.

Models are submitted to ModelArchive (modelarchive.org). The individual links for the models are:

SRP54 with ALK WT signal peptide: modelarchive.org/doi/10.5452/ma-owxf7

SRP54 with ALK W8R signal peptide: modelarchive.org/doi/10.5452/ma-w701z

SRP54 with ALK S15Y signal peptide: modelarchive.org/doi/10.5452/ma-cm6gn

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

The authors would like to acknowledge data repositories and bioinformatic tools used in this study: BiomaRt, Ensembl, Variant Effect Predictor, SignalP 6.0, DAVID Bioinformatics, Galaxy Project, PANTHER, PyMol and R project. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Funding

National Institute of General Medical Sciences of the National Institutes of Health [R01GM135167].

Conflict of interest statement

None declared.

References

1. Thul,P.J., Åkesson,L., Wiking,M., Mahdessian,D., Geladaki,A., Ait Blal,H., Alm,T., Asplund,A., Björk,L., Breckels,L.M., *et al.* (2017) A subcellular map of the human proteome. *Science*, **356**, eaal3321.
2. Sommer,M.S. and Schleiff,E. (2014) Protein targeting and transport as a necessary consequence of increased cellular complexity. *Cold Spring Harb. Perspect. Biol.*, **6**, a016055.
3. Kunze,M. and Berger,J. (2015) The similarity between N-terminal targeting signals for protein import into different organelles and its evolutionary relevance. *Front. Physiol.*, **6**, 259.
4. von Heijne,G. (1985) Signal sequences. The limits of variation. *J. Mol. Biol.*, **184**, 99–105.
5. von Heijne,G. (1990) The signal peptide. *J. Membr. Biol.*, **115**, 195–201.
6. Nielsen,H., Tsirigos,K.D., Brunak,S. and von Heijne,G. (2019) A brief history of protein sorting prediction. *Protein J.*, **38**, 200–216.
7. Karamyshev,A.L., Tikhonova,E.B. and Karamysheva,Z.N. (2020) Translational control of secretory proteins in health and disease. *Int. J. Mol. Sci.*, **21**, 2538.
8. Uhlén,M., Karlsson,M.J., Hober,A., Svensson,A.S., Scheffel,J., Kotol,D., Zhong,W., Tebani,A., Strandberg,L., Edfors,F., *et al.* (2019) The human secretome. *Sci. Signal*, **12**, eaaz0274.
9. Lang,S. and Zimmermann,R. (2022) Mechanisms of ER Protein Import.. *Int. J. Mol. Sci.*, **23**, 5315.
10. Uhlen,M., Fagerberg,L., Hallstrom,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,A., Kampf,C., Sjostedt,E., Asplund,A., *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.
11. Zimmermann,R., Muller,L. and Wullich,B. (2006) Protein transport into the endoplasmic reticulum: mechanisms and pathologies. *Trends Mol. Med.*, **12**, 567–573.
12. Hebert,D.N. and Molinari,M. (2007) In and out of the ER: protein folding, quality control, degradation, and related human diseases. *Physiol. Rev.*, **87**, 1377–1408.
13. Kellogg,M.K., Tikhonova,E.B. and Karamyshev,A.L. (2022) Signal recognition particle in human diseases. *Front. Genet.*, **13**, 898083.
14. Rane,N.S., Chakrabarti,O., Feigenbaum,L. and Hegde,R.S. (2010) Signal sequence insufficiency contributes to neurodegeneration caused by transmembrane prion protein. *J. Cell Biol.*, **188**, 515–526.
15. Tikhonova,E.B., Karamysheva,Z.N., von Heijne,G. and Karamyshev,A.L. (2019) Silencing of aberrant secretory protein expression by disease-associated mutations. *J. Mol. Biol.*, **431**, 2567–2580.
16. Kellogg,M.K., Miller,S.C., Tikhonova,E.B. and Karamyshev,A.L. (2021) SRPassing co-translational targeting: the role of the signal recognition particle in protein targeting and mRNA protection. *Int. J. Mol. Sci.*, **22**, 6284.

17. Hsieh,H.H. and Shan,S.O. (2021) Fidelity of cotranslational protein targeting to the endoplasmic reticulum. *Int. J. Mol. Sci.*, **23**, 281.
18. Akopian,D., Shen,K., Zhang,X. and Shan,S.O. (2013) Signal recognition particle: an essential protein-targeting machine. *Annu. Rev. Biochem.*, **82**, 693–721.
19. Aviram,N. and Schuldiner,M. (2017) Targeting and translocation of proteins to the endoplasmic reticulum at a glance. *J. Cell Sci.*, **130**, 4079–4085.
20. Hegde,R.S. and Keenan,R.J. (2022) The mechanisms of integral membrane protein biogenesis. *Nat. Rev. Mol. Cell Biol.*, **23**, 107–124.
21. Elvekrog,M.M. and Walter,P. (2015) Dynamics of co-translational protein targeting. *Curr. Opin. Chem. Biol.*, **29**, 79–86.
22. von Heijne,G. (1983) Patterns of amino acids near signal-sequence cleavage sites. *Eur. J. Biochem.*, **133**, 17–21.
23. Nilsson,I., Lara,P., Hessa,T., Johnson,A.E., von Heijne,G. and Karamyshev,A.L. (2015) The code for directing proteins for translocation across ER membrane: SRP cotranslationally recognizes specific features of a signal sequence. *J. Mol. Biol.*, **427**, 1191–1201.
24. Pinarbasi,E.S., Karamyshev,A.L., Tikhonova,E.B., Wu,I.H., Hudson,H. and Thomas,P.J. (2018) Pathogenic signal sequence mutations in progranulin disrupt SRP interactions required for mRNA stability. *Cell Rep.*, **23**, 2844–2851.
25. Karamyshev,A.L., Patrick,A.E., Karamysheva,Z.N., Griesemer,D.S., Hudson,H., Tjon-Kon-Sang,S., Nilsson,I., Otto,H., Liu,Q., Rospert,S., *et al.* (2014) Inefficient SRP interaction with a nascent chain triggers a mRNA quality control pathway. *Cell*, **156**, 146–157.
26. Park,S.Y., Ye,H., Steiner,D.F. and Bell,G.I. (2010) Mutant proinsulin proteins associated with neonatal diabetes are retained in the endoplasmic reticulum and not efficiently secreted. *Biochem. Biophys. Res. Commun.*, **391**, 1449–1454.
27. Cui,J., Chen,W., Sun,J., Guo,H., Madley,R., Xiong,Y., Pan,X., Wang,H., Tai,A.W., Weiss,M.A., *et al.* (2015) Competitive inhibition of the endoplasmic reticulum signal peptidase by non-cleavable mutant preprotein cargos. *J. Biol. Chem.*, **290**, 28131–28140.
28. Guo,H., Sun,J., Li,X., Xiong,Y., Wang,H., Shu,H., Zhu,R., Liu,Q., Huang,Y., Madley,R., *et al.* (2018) Positive charge in the n-region of the signal peptide contributes to efficient post-translational translocation of small secretory preproteins. *J. Biol. Chem.*, **293**, 1899–1907.
29. Guo,H., Xiong,Y., Witkowski,P., Cui,J., Wang,L.J., Sun,J., Lara-Lemus,R., Haataja,L., Hutchison,K., Shan,S.O., *et al.* (2014) Inefficient translocation of preproinsulin contributes to pancreatic beta cell failure and late-onset diabetes. *J. Biol. Chem.*, **289**, 16290–16302.
30. Karamyshev,A.L. and Karamysheva,Z.N. (2018) Lost in translation: ribosome-associated mRNA and protein quality controls. *Front. Genet.*, **9**, 431.
31. Wang,L. and Ye,Y. (2020) Clearing traffic jams during protein translocation across membranes. *Front. Cell Dev. Biol.*, **8**, 610689.
32. Sun,Z. and Brodsky,J.L. (2019) Protein quality control in the secretory pathway. *J. Cell Biol.*, **218**, 3171–3187.
33. Tikhonova,E.B., Gutierrez Guarnizo,S.A., Kellogg,M.K., Karamyshev,A., Dozmorov,I.M., Karamysheva,Z.N. and Karamyshev,A.L. (2022) Defective human SRP induces protein quality control and triggers stress response. *J. Mol. Biol.*, **434**, 167832.
34. Karamysheva,Z.N. and Karamyshev,A.L. (2023) Aberrant protein targeting activates quality control on the ribosome. *Front. Cell Dev. Biol.*, **11**, 1198184.
35. Buchberger,A., Bukau,B. and Sommer,T. (2010) Protein quality control in the cytosol and the endoplasmic reticulum: brothers in arms. *Mol. Cell*, **40**, 238–252.
36. Volpi,V.G., Touvier,T. and D'Antonio,M. (2017) Endoplasmic reticulum protein quality control failure in myelin disorders. **9**, 162.
37. Tsai,Y.C. and Weissman,A.M. (2010) The unfolded protein response, degradation from endoplasmic reticulum and cancer. *Genes Cancer*, **1**, 764–778.
38. Phillips,B.P., Gomez-Navarro,N. and Miller,E.A. (2020) Protein quality control in the endoplasmic reticulum. *Curr. Opin. Cell Biol.*, **65**, 96–102.
39. Vembar,S.S. and Brodsky,J.L. (2008) One step at a time: endoplasmic reticulum-associated degradation. *Nat. Rev. Mol. Cell Biol.*, **9**, 944–957.
40. Abu-Safieh,L., Alrashed,M., Anazi,S., Alkuraya,H., Khan,A.O., Al-Owain,M., Al-Zahrani,J., Al-Abdi,L., Hashem,M., Al-Tarimi,S., *et al.* (2013) Autozygome-guided exome sequencing in retinal dystrophy patients reveals pathogenetic mutations and novel candidate disease genes. *Genome Res.*, **23**, 236–247.
41. Sunthornthepvarakul,T., Churesigaew,S. and Ngowngarmratana,S. (1999) A novel mutation of the signal peptide of the preproparathyroid hormone gene associated with autosomal recessive familial isolated hypoparathyroidism. *J. Clin. Endocrinol. Metab.*, **84**, 3792–3796.
42. Baker,M., Mackenzie,I.R., Pickering-Brown,S.M., Gass,J., Rademakers,R., Lindholm,C., Snowden,J., Adamson,J., Sadovnick,A.D., Rollinson,S., *et al.* (2006) Mutations in progranulin cause tau-negative frontotemporal dementia linked to chromosome 17. *Nature*, **442**, 916–919.
43. Baumer,A., Belli,S., Trüeb,R.M. and Schinzel,A. (2000) An autosomal dominant form of hereditary hypotrichosis simplex maps to 18p11.32-p11.23 in an Italian family. *Eur. J. Hum. Genet.*, **8**, 443–448.
44. Machado,R.D., Southgate,L., Eichstaedt,C.A., Aldred,M.A., Austin,E.D., Best,D.H., Chung,W.K., Benjamin,N., Elliott,C.G., Eyries,M., *et al.* (2015) Pulmonary arterial hypertension: a current perspective on established and emerging molecular genetic defects. *Hum. Mutat.*, **36**, 1113–1127.
45. Chou,Y.H., Pollak,M.R., Brandi,M.L., Toss,G., Arnqvist,H., Atkinson,A.B., Papapoulos,S.E., Marx,S., Brown,E.M., Seidman,J.G., *et al.* (1995) Mutations in the human Ca(2+)-sensing-receptor gene that cause familial hypocalciuric hypercalcemia. *Am. J. Hum. Genet.*, **56**, 1075–1079.
46. Malmgren,B., Lindskog,S., Elgadi,A. and Norgren,S. (2004) Clinical, histopathologic, and genetic investigation in two large families with dentinogenesis imperfecta type II. *Hum. Genet.*, **114**, 491–498.
47. Fischer,J., Bouadjar,B., Heilig,R., Huber,M., Lefèvre,C., Jobard,F., Macari,F., Bakija-Konsuo,A., Ait-Belkacem,F., Weissenbach,J., *et al.* (2001) Mutations in the gene encoding SLURP-1 in Mal de Meleda. *Hum. Mol. Genet.*, **10**, 875–880.
48. Dickinson,J.L., Sale,M.M., Passmore,A., FitzGerald,L.M., Wheatley,C.M., Burdon,K.P., Craig,J.E., Tengtrisorn,S., Carden,S.M., Maclean,H., *et al.* (2006) Mutations in the NDP gene: contribution to Norrie disease, familial exudative vitreoretinopathy and retinopathy of prematurity. *Clin. Experiment. Ophthalmol.*, **34**, 682–688.
49. Karamysheva,Z.N., Tikhonova,E.B. and Karamyshev,A.L. (2019) Granulin in frontotemporal lobar degeneration: molecular mechanisms of the disease. *Front. Neurosci.*, **13**, 395.
50. Jarjanazi,H., Savas,S., Pabalan,N., Dennis,J.W. and Ozcelik,H. (2008) Biological implications of SNPs in signal peptide domains of human proteins. *Proteins*, **70**, 394–403.
51. Meur,G., Simon,A., Harun,N., Virally,M., Dechaume,A., Bonnefond,A., Fetita,S., Tarasov,A.I., Guillausseau,P.J., Boesgaard,T.W., *et al.* (2010) Insulin gene mutations resulting in early-onset diabetes: marked differences in clinical presentation, metabolic status, and pathogenic effect through endoplasmic reticulum retention. *Diabetes*, **59**, 653–661.
52. Potorac,I., Trehan,A., Szymanska,K., Fudvoye,J., Thiry,A., Huhtaniemi,I., Daly,A.F., Beckers,A., Parent,A.S. and

Rivero-Muller,A. (2019) Compound heterozygous mutations in the luteinizing hormone receptor signal peptide causing 46,XY disorder of sex development. *Eur. J. Endocrinol.*, **181**, K11–K20.
53. Vezzoli,V., Duminuco,P., Vottero,A., Kleinau,G., Schulein,R., Minari,R., Bassi,I., Bernasconi,S., Persani,L. and Bonomi,M. (2015) A new variant in signal peptide of the human luteinizing hormone receptor (LHCGR) affects receptor biogenesis causing leydig cell hypoplasia. *Hum. Mol. Genet.*, **24**, 6003–6012.
54. Mencarelli,M., Zulian,A., Cancello,R., Alberti,L., Gilardini,L., Di Blasio,A.M. and Invitti,C. (2012) A novel missense mutation in the signal peptide of the human POMC gene: a possible additional link between early-onset type 2 diabetes and obesity. *Eur. J. Hum. Genet.*, **20**, 1290–1294.
55. Evans,T.D., Zhang,X., Clark,R.E., Alisio,A., Song,E., Zhang,H., Reilly,M.P., Stitziel,N.O. and Razani,B. (2019) Functional characterization of LIPA (lysosomal acid Lipase) variants associated with coronary artery disease. *Arterioscler. Thromb. Vasc. Biol.*, **39**, 2480–2491.
56. UniProt Consortium (2020) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
57. Durinck,S., Spellman,P.T., Birney,E. and Huber,W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, **4**, 1184–1191.
58. Teufel,F., Almagro Armenteros,J.J., Johansen,A.R., Gíslason,M.H., Pihl,S.I., Tsirigos,K.D., Winther,O., Brunak,S., von Heijne,G. and Nielsen,H. (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.*, **40**, 1023–1025.
59. Smigielski,E.M., Sirotkin,K., Ward,M. and Sherry,S.T. (2000) dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res.*, **28**, 352–355.
60. Cingolani,P., Platts,A., Wang le,L., Coon,M., Nguyen,T., Wang,L., Land,S.J., Lu,X. and Ruden,D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: sNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly.*, **6**, 80–92.
61. McLaren,W., Gil,L., Hunt,S.E., Riat,H.S., Ritchie,G.R.S., Thormann,A., Flicek,P. and Cunningham,F. (2016) The ensembl variant effect predictor. *Genome Biol.*, **17**, 122.
62. Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
63. Boman,H.G. (2003) Antibacterial peptides: basic facts and emerging concepts. *J. Intern. Med.*, **254**, 197–215.
64. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
65. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
66. Lyskov,S. and Gray,J.J. (2008) The RosettaDock server for local protein–protein docking. *Nucleic Acids Res.*, **36**, W233–W238.
67. Chaudhury,S., Berrondo,M., Weitzner,B.D., Muthu,P., Bergman,H. and Gray,J.J. (2011) Benchmarking and analysis of protein docking performance in Rosetta v3.2. *PLoS One*, **6**, e22477.
68. *The PyMOL Molecular Graphics System, Version 2.5*, Schrödinger, LLC.
69. Halic,M., Becker,T., Pool,M.R., Spahn,C.M., Grassucci,R.A., Frank,J. and Beckmann,R. (2004) Structure of the signal recognition particle interacting with the elongation-arrested ribosome. *Nature*, **427**, 808–814.
70. Gao,Y., Zhang,Q., Lang,Y., Liu,Y., Dong,X., Chen,Z., Tian,W., Tang,J., Wu,W., Tong,Y., *et al.* (2017) Human apo-SRP72 and SRP68/72 complex structures reveal the molecular basis of protein translocation. *J. Mol. Cell Biol.*, **9**, 220–230.
71. Grotwinkel,J.T., Wild,K., Segnitz,B. and Sinning,I. (2014) SRP RNA remodeling by SRP68 explains its role in protein translocation. *Science*, **344**, 101–104.
72. Kuglstatter,A., Oubridge,C. and Nagai,K. (2002) Induced structural changes of 7SL RNA during the assembly of human signal recognition particle. *Nat. Struct. Biol.*, **9**, 740–744.
73. Liaci,A.M., Steigenberger,B., Telles de Souza,P.C., Tamara,S., Grollers-Mulderij,M., Ogrissek,P., Marrink,S.J., Scheltema,R.A. and Forster,F. (2021) Structure of the human signal peptidase complex reveals the determinants for signal peptide cleavage. *Mol. Cell*, **81**, 3934–3948.
74. Thomas,P.D., Ebert,D., Muruganujan,A., Mushayahama,T., Albou,L.P. and Mi,H. (2022) PANTHER: making genome-scale phylogenetics accessible to all. *Protein Sci.*, **31**, 8–22.
75. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
76. Nesmeyanova,M.A., Karamyshev,A.L., Karamysheva,Z.N., Kalinin,A.E., Ksenzenko,V.N. and Kajava,A.V. (1997) Positively charged lysine at the N-terminus of the signal peptide of the Escherichia coli alkaline phosphatase provides the secretion efficiency and is involved in the interaction with anionic phospholipids. *FEBS Lett.*, **403**, 203–207.
77. Goder,V., Junne,T. and Spiess,M. (2003) Sec61p Contributes to Signal Sequence Orientation According to the Positive-Inside Rule. *Mol. Biol. Cell*, **15**, 1470–1478.
78. Lyskov,S., Chou,F.C., Conchúir,S., Der,B.S., Drew,K., Kuroda,D., Xu,J., Weitzner,B.D., Renfrew,P.D., Sripakdeevong,P., *et al.* (2013) Serverification of molecular modeling applications: the Rosetta Online Server that Includes Everyone (ROSIE). *PLoS One*, **8**, e63906.
79. World Health Organization (2020) The Top 10 Causes of Death. World Health Organization.
80. Kriegler,T., Kiburg,G. and Hessa,T. (2020) Translocon-Associated Protein Complex (TRAP) is crucial for co-translational translocation of pre-proinsulin. *J. Mol. Biol.*, **432**, 166694.
81. Jadhav,B., McKenna,M., Johnson,N., High,S., Sinning,I. and Pool,M.R. (2015) Mammalian SRP receptor switches the Sec61 translocase from Sec62 to SRP-dependent translocation. *Nat. Commun.*, **6**, 10133.
82. Hassdenteufel,S., Johnson,N., Paton,A.W., Paton,J.C., High,S. and Zimmermann,R. (2018) Chaperone-mediated Sec61 channel gating during ER import of small precursor proteins overcomes Sec61 inhibitor-reinforced energy barrier. *Cell Rep.*, **23**, 1373–1386.
83. Sun,S., Li,X. and Mariappan,M. (2023) Signal sequences encode information for protein folding in the endoplasmic reticulum. *J. Cell Biol.*, **222**, e202203070.
84. Kriegler,T., Magoulopoulou,A., Amate Marchal,R. and Hessa,T. (2018) Measuring endoplasmic reticulum signal sequences translocation efficiency using the Xbp1 arrest peptide. *Cell Chem Biol*, **25**, 880–890.
85. Hessa,T., Kim,H., Bihlmaier,K., Lundin,C., Boekel,J., Andersson,H., Nilsson,I., White,S.H. and von Heijne,G. (2005) Recognition of transmembrane helices by the endoplasmic reticulum translocon. *Nature*, **433**, 377–381.
86. Hessa,T., Meindl-Beinker,N.M., Bernsel,A., Kim,H., Sato,Y., Lerch-Bader,M., Nilsson,I., White,S.H. and von Heijne,G. (2007) Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature*, **450**, 1026–1030.
87. Tomilo,M., Wilkinson,K.S. and Ryan,P. (1994) Can a signal sequence become too hydrophobic?*J. Biol. Chem.*, **269**, 32016–32021.
88. Pérez-Palma,E., May,P., Iqbal,S., Niestroj,L.M., Du,J., Heyne,H.O., Castrillon,J.A., O'Donnell-Luria,A., Nürnberg,P., Palotie,A., *et al.* (2020) Identification of pathogenic variant enriched regions across genes and gene families. *Genome Res.*, **30**, 62–71.

89. Miller,M.P. and Kumar,S. (2001) Understanding human disease mutations through the use of interspecific genetic variation. *Hum. Mol. Genet.*, **10**, 2319–2328.

90. Vockley,J.G., Goodman,B.K., Tabor,D.E., Kern,R.M., Jenkinson,C.P., Grody,W.W. and Cederbaum,S.D. (1996) Loss of function mutations in conserved regions of the human arginase I gene. *Biochem. Mol. Med.*, **59**, 44–51.

Received: July 13, 2023. **Revised:** September 5, 2023. **Editorial Decision:** September 25, 2023. **Accepted:** September 29, 2023