Prohormone Predictor - Tissue-based human prohormone prediction for the identification of novel small, secreted peptides

Laetitia Coassolo[1,2], Quennie Nguyen[1], Niels Danneskiold-Banhos Samsoee[1], David Toomer[1], Katrin J. Svensson[1,2*]

[1]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA.

[2]Stanford Diabetes Research Center, Stanford University School of Medicine, Stanford, CA, USA

*Corresponding author: katrinjs@stanford.edu

SUMMARY

Peptide hormones represent a class of small peptides that regulate a wide range of physiological functions. Systematic efforts to identify and characterize secreted bioactive polypeptides have traditionally been hampered by their low abundance, small size, and the difficulty predicting their functions. Using evolution as a tool, we developed a method to map peptides based on their conserved cleavage sequences as a predictor of release of small peptide hormones. By annotating all secreted protein sequences with one of the most common conserved dibasic cleavage sites within the entire human secretome, this tissue-wide peptide secretome of 2,683 peptides provides new insights to extent of which posttranslational processing contribute to diversifying our secretome.

INTRODUCTION

Peptide hormones are a class of small (<100 amino acids), low abundant, and bioactive peptides involved in regulating physiological processes such as food intake and body weight regulation, making them attractive targets for modulation of energy metabolism (*1*–*3*). Recently, modified GLP-1 peptide analogs such as liraglutide and semaglutide have had transformative efficacy in reducing body weight in humans (*4*, *5*). Traditionally, novel bioactive peptide hormones have been identified by biochemical purification from endocrine organs, including insulin, glucagon, and oxyntomodulin from pancreatic or gut extracts (*6*, *7*), and neuropeptide Y (NPY) and gonadotropin-releasing hormone (GnRH) from brain extracts (*8*, *9*). More recently, extensive proteomics and peptidomics efforts have demonstrated detection and quantitation of peptide hormones and neuropeptides from complex biological tissues (*10*–*12*) or blood (*13*, *14*), but there are still significant challenges in detection owing to their low abundance. Furthermore, because many peptide hormones are synthesized as part of larger precursors further processed into active fragments by posttranslational endoproteolytic cleavage, their dynamic regulation are undetected by RNA sequencing analyses or conventional proteomic analyses. Thus, we know very little about other endogenous peptides that control feeding behavior and obesity development, and the extent to which these uncharacterized peptides contribute to modulation of energy balance remains unclear.

RESULTS

Sequence pattern recognition predicts small, secreted human peptide hormones and their expression across human tissues

The functional identification and characterization of small, proteolytically cleaved peptides have traditionally been challenging because of their low abundance and the ability to distinguish bioactive fragments from inactive fragments or degradation products (*15*, *16*). We therefore made use of the fact that peptide hormones are often synthesized as part of larger prohormones that are processed into peptides by posttranslational endoproteolytic cleavage, similar to GLP-1 (*17*, *18*). This process occurs at specific dibasic amino acid residues, KR/RR/RK/KK, followed by a non-basic, non-aliphatic amino acid (KRH being an exception)(*19*) (**Fig 1a**). The proteolytic cleavage is mediated by enzymes present in the secretory pathway, including the subtilisin-like proprotein convertases furin, prohormone convertase 2 (PC2), and prohormone convertase 1/3 (PC1/PC3)(*17*, *18*). Using these conserved sites as a criterion, we generated a code for sequence pattern recognition using Regular Expression (RegEx) for matching text patterns of protein sequences. Using this method, we are able to annotate all amino acid sequences with a signal peptide (2,082 proteins) to their tissue of origin (*20*), that also have more than 4 of any combination of the KR/RR/RK/KK cleavage sites (**Fig 1b**). The space between cleavage sites was set to 3 to generate peptides > 4 aa in length to exclude tripeptides (**Fig 1c**). The minimum number of cleavage sites per protein was set to 4, based on the number of cleavage sites to generate at least five peptides after cleavage (**Fig 1d**). In addition, the next criterion was that the protein needs to be smaller than 2,000 amino acids to exclude extremely long peptide sequences, to enrich for prohormones with high cleavage site density, such as pre-proglucagon

(**Fig 1e**). In total, out of 2,082 secreted proteins, we identify 373 prohormones predicted to generate 2,683 novel peptides, the majority of which are entirely unknown (**Fig 1b**). To analyze the tissue distribution of the identified prohormones, we categorized the prohormones according to distinct or shared tissue expression (*21*) (**Fig 1f and Fig S1**). The largest group of peptides identified, 601 peptides, have wide expression (**Fig S1b**), while the second largest number of peptides, 366, belong to the brain (**Fig 1f**). Interestingly, the brain prohormone prediction accurately predicts 9 already known prohormones and their cleaved neuropeptides, and an additional 50 brain-enriched proteins with unknown functions (**Fig 1g**). For example, we identify peptides derived from proenkephalin-A, pituitary adenylate cyclase activating polypeptide (PACAP), secretogranin-II, and thyrotropin-releasing hormone (TRH) (**Fig 1g**). Furthermore, we identify many known peptides, including vasoactive intestinal peptide in the intestine (**Fig 1h**), neuropeptide precursors in the pituitary gland, including secretogranin-I, pro-opiomelanocortin (POMC) (**Fig 1i**) and proenkephalin-A in the adrenal gland (**Fig 1j**). Lastly, we find that the pancreas is enriched for proglucagon, which validates that the computational approach can predict true peptide hormones (**Fig 1k**). In conclusion, the brain (**Fig 1g**) and liver (**Fig S1c**) are major contributors to the release of small, secreted peptides, most of which have no previously annotated function.

Bioinformatic analyses for Prohormone Predictor

To generate the program for sequence pattern recognition using Regular Expression (RegEx) for matching text patterns of protein sequences, we used the FASTA files for all reviewed, secreted, human genes (as in secreted.fasta) retrieved from UniProtKB API. The code for the Prohormone Predictor can be found at https://github.com/Svensson-Lab/pro-hormone-predictor.

This program predicts whether a secreted gene has prohormone activity based on the number of cleavage sites it contains. For tissue distribution, prohormones were categorized according to distinct or shared tissue expression based on tissue expression data from Human Protein Atlas (*21*).The following criteria were used to annotate all predicted prohormones and their subsequent peptides: > 4 of any combination of the KR/RR/RK/KK cleavage sites, > 3 cleavage sites per protein, > 4 cleavage sites per protein, prohormone < 2,000 amino acids in size.

Discussion

The peptide drug discovery field has revolutionized medicine with the introduction of over 60 peptide drugs approved in the US. To date, a number of peptides have been identified as modulators of food intake and obesity, including leptin, ghrelin, and glucagon-derived peptides. While peptide hormones traditionally have been identified by biochemical purification, our computational analysis has unlocked exciting possibilities to target selective aspects of metabolism via distinct mechanisms from current drugs.

Posttranslational cleavage enables rapid processing and release of mature, active peptides into circulation in response to nutritional or dietary changes. This posttranslational mechanism is especially important in the context of fasting and feeding —a process where instant responses (within minutes) are needed, which excludes traditional regulation such as transcriptional activation and protein stability alterations. Hormonal signaling through protein phosphorylation is one of the most important post-translational modifications allowing for rapid changes in cellular metabolic states, including feeding control.

Author contributions

Conceptualization: L.C., K.J.S.; methodology: L.C., Q.N., N.D.B., A.W., D.T., M.Z., E.B., V.L.; investigation: L.C., Q.N., N.D.B., A.W., D.T., M.Z., E.B., V.L., L.W., D.H.C., K.J.S.; supervision and funding acquisition : K.J.S.; Writing – original draft: L.C., K.J.S.

Competing Interests

The authors do not declare any conflict of interests.

Data and materials availability

All data and peptide sequences generated or analyzed during this study are included in the manuscript and supporting files. The code for the Prohormone Predictor can be found at https://github.com/Svensson-Lab/pro-hormone-predictor. Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Dr. Katrin J. Svensson (katrinjs@stanford.edu).

FIGURES AND FIGURE LEGENDS

Figure 1

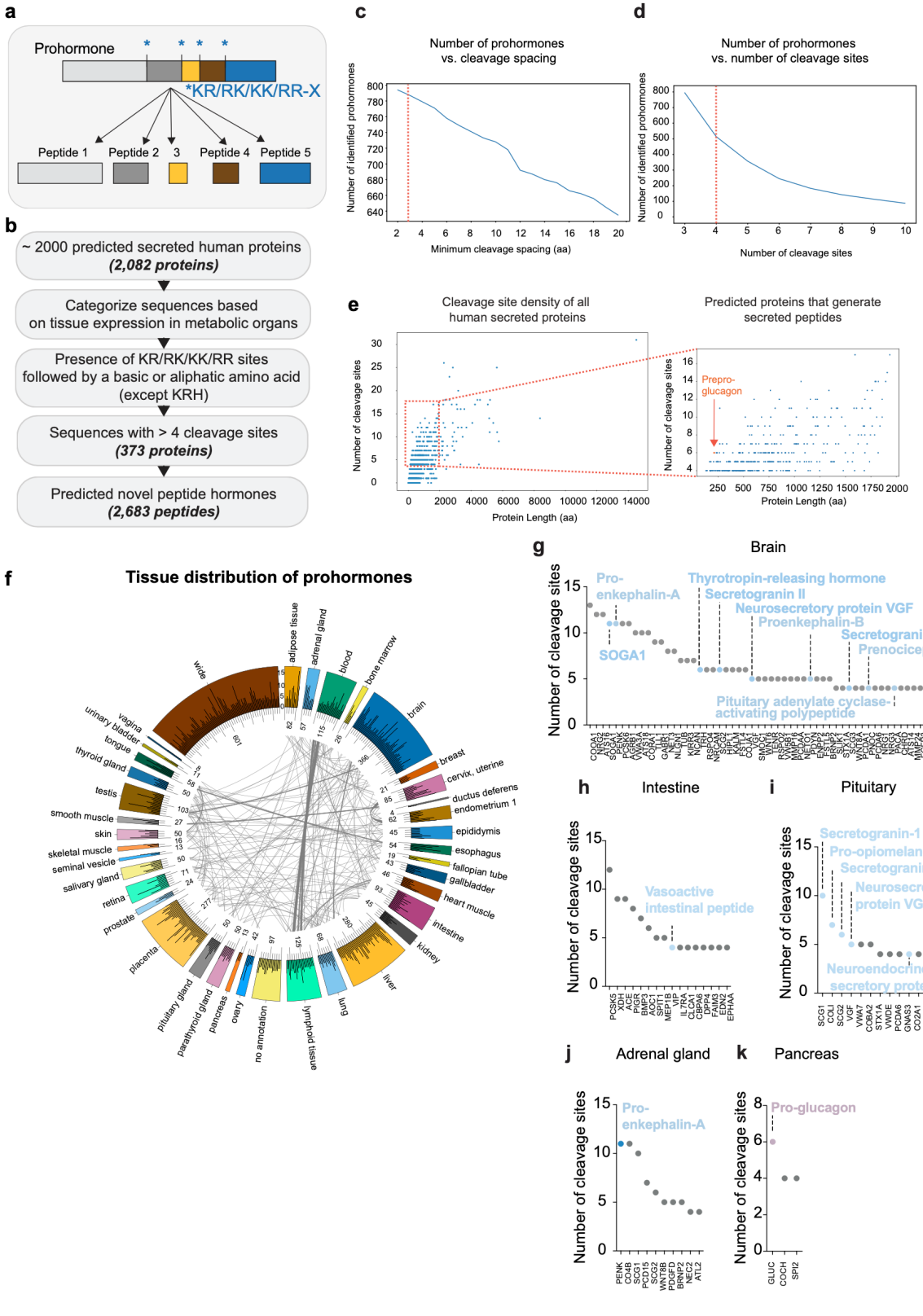

Figure 1. Sequence pattern recognition predicts small, secreted human polypeptide hormones expressed across human tissues.

a. Evolutionary conserved dibasic cleavage sites KR/RK/KK/RR-X in protein sequences will predict cleaved peptides.

b. Criteria for annotation of prohormones based on the presence of >4 dibasic cleavage sites, size longer than 4 aa, and a high cleavage density. This prediction generated 2,683 novel human peptides.

c. Number of prohormones with a minimum cleavage spacing of 3 amino acids between cleavage sites.

d. Number of prohormones with 4 or more cleavage sites per protein.

e. Cleavage site density as determined by the number of cleavage sites relative to the protein length. Boxed area depicts all proteins with a size < 2000 amino acids with > 4 cleavage sites. Preproglucagon is marked as a positive internal control.

f. Distribution diagram of shared and unique peptides from each organ. Inner grey links show peptides detected in two or more tissues. The ticks indicate the number of peptides. Large ticks indicate 50 peptides, and small ticks indicate 5 peptides. The inner ring displays the total number of peptides from a single tissue.

g-k. Detected prohormones and the number of cleavage sites per prohormone in brain (g), intestine (h), pituitary gland (i), adrenal gland (j), and pancreas (k). Blue: known prohormones. Grey: unannotated functions.

Figure S1

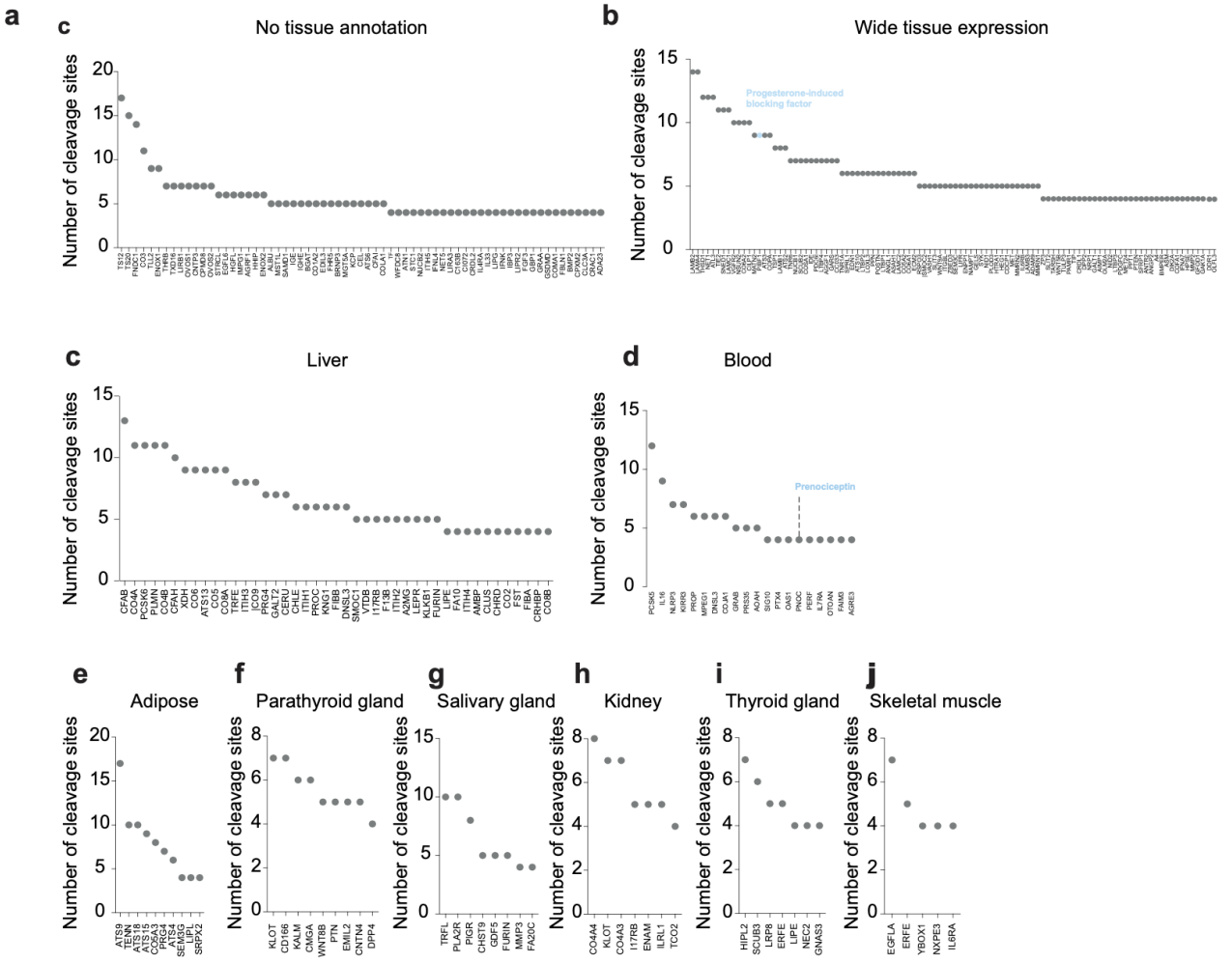

Figure S1. Sequence pattern recognition predicts small, secreted human polypeptide hormones expressed across human tissues.

a-j. Detected prohormones and the number of cleavage sites per prohormone with no tissue annotation (a), wide tissue expression (b), in liver (c), blood (d), adipose tissue (e), parathyroid gland (f), salivary gland (g), kidney (h), thyroid gland (i) and skeletal muscle (j) Blue: known prohormones. Grey: unannotated functions.

REFERENCES

1. C. Sobrino Crespo, A. Perianes Cachero, L. Puebla Jiménez, V. Barrios, E. Arilla Ferreiro, Peptides and food intake. *Frontiers in endocrinology*. **5**, 58 (2014).
2. J. E. Campbell, D. J. Drucker, Pharmacology, physiology, and mechanisms of incretin hormone action. *Cell metabolism*. **17**, 819–837 (2013).
3. M. Muttenthaler, G. F. King, D. J. Adams, P. F. Alewood, Trends in peptide drug discovery. *Nature Reviews Drug Discovery*. **20**, 309–325 (2021).
4. C. H. Lin, L. Shao, Y. M. Zhang, Y. J. Tu, Y. Zhang, B. Tomlinson, P. Chan, Z. Liu, An evaluation of liraglutide including its efficacy and safety for the treatment of obesity. *Expert opinion on pharmacotherapy*. **21**, 275–285 (2020).
5. J. P. H. Wilding, R. L. Batterham, S. Calanna, M. Davies, L. F. Van Gaal, I. Lingvay, B. M. McGowan, J. Rosenstock, M. T. D. Tran, T. A. Wadden, S. Wharton, K. Yokote, N. Zeuthen, R. F. Kushner, Once-Weekly Semaglutide in Adults with Overweight or Obesity. *The New England journal of medicine*. **384**, 989–1002 (2021).
6. H. G. Pollock, J. W. Hamilton, J. B. Rouse, K. E. Ebner, A. B. Rawitch, Isolation of peptide hormones from the pancreas of the bullfrog (Rana catesbeiana). Amino acid sequences of pancreatic polypeptide, oxyntomodulin, and two glucagon-like peptides. *The Journal of biological chemistry*. **263**, 9746–9751 (1988).
7. I. Vecchio, C. Tornali, N. L. Bragazzi, M. Martini, The discovery of insulin: An important milestone in the history of medicine. *Frontiers in Endocrinology*. **9** (2018), p. 613.
8. K. Tatemoto, Neuropeptide Y: complete amino acid sequence of the brain peptide. *Proceedings of the National Academy of Sciences of the United States of America*. **79**,

5485–5489 (1982).

9. D. A. Lovejoy, W. H. Fischer, S. Ngamvongchon, A. G. Craig, C. S. Nahorniak, R. E. Peter, J. E. Rivier, N. M. Sherwood, Distinct sequence of gonadotropin-releasing hormone (GnRH) in dogfish brain provides insight into GnRH evolution. *Proceedings of the National Academy of Sciences of the United States of America*. **89**, 6373–6377 (1992).

10. K. L. Lee, M. J. Middleditch, G. M. Williams, M. A. Brimble, G. J. S. Cooper, Using Mass Spectrometry to Detect, Differentiate, and Semiquantitate Closely Related Peptide Hormones in Complex Milieu: Measurement of IGF-II and Vesiculin. *Endocrinology*. **156**, 1194–1199 (2015).

11. J. E. Lee, Neuropeptidomics: Mass Spectrometry-Based Identification and Quantitation of Neuropeptides. *Genomics & informatics*. **14**, 12–19 (2016).

12. L. D. Fricker, J. Lim, H. Pan, F.-Y. Che, Peptidomics: Identification and quantification of endogenous peptides in neuroendocrine tissues. *Mass Spectrometry Reviews*. **25**, 327–344 (2006).

13. B. Muthusamy, G. Hanumanthu, S. Suresh, B. Rekha, D. Srinivas, L. Karthick, B. M. Vrushabendra, S. Sharma, G. Mishra, P. Chatterjee, K. S. Mangala, H. N. Shivashankar, K. N. Chandrika, N. Deshpande, M. Suresh, N. Kannabiran, V. Niranjan, A. Nalli, T. S. K. Prasad, K. S. Arun, R. Reddy, S. Chandran, T. Jadhav, D. Julie, M. Mahesh, S. L. John, K. Palvankar, D. Sudhir, P. Bala, N. S. Rashmi, G. Vishnupriya, K. Dhar, S. Reshma, R. Chaerkady, T. K. B. Gandhi, H. C. Harsha, S. S. Mohan, K. S. Deshpande, M. Sarker, A. Pandey, Plasma Proteome Database as a resource for proteomics research. *Proteomics*. **5**, 3531–3536 (2005).

14. J. M. Schwenk, G. S. Omenn, Z. Sun, D. S. Campbell, M. S. Baker, C. M. Overall, R. Aebersold, R. L. Moritz, E. W. Deutsch, The Human Plasma Proteome Draft of 2017: Building on the Human Plasma PeptideAtlas from Mass Spectrometry and Complementary Assays. *Journal of proteome research*. **16**, 4299–4310 (2017).

15. B. R. Southey, A. Amare, T. A. Zimmerman, S. L. Rodriguez-Zas, J. V Sweedler, NeuroPred: a tool to predict cleavage sites in neuropeptide precursors and provide the masses of the resulting peptides. *Nucleic acids research*. **34**, W267-72 (2006).

16. B. R. Southey, J. V Sweedler, S. L. Rodriguez-Zas, A python analytical pipeline to identify prohormone precursors and predict prohormone cleavage sites. *Frontiers in neuroinformatics*. **2**, 7 (2008).

17. D. F. Steiner, The proprotein convertases. *Current opinion in chemical biology*. **2**, 31–39 (1998).

18. M. Zheng, R. D. Streck, R. E. Scott, N. G. Seidah, J. E. Pintar, The developmental expression in rat of proteases furin, PC1, PC2, and carboxypeptidase E: implications for early maturation of proteolytic processing capacity. *The Journal of neuroscience : the official journal of the Society for Neuroscience*. **14**, 4656–4673 (1994).

19. J. R. Peinado, H. Li, K. Johanning, I. Lindberg, Cleavage of recombinant proenkephalin and blockade mutants by prohormone convertases 1 and 2: an in vitro specificity study. *Journal of neurochemistry*. **87**, 868–878 (2003).

20. M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. M. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A. K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P. H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk,

M. Hamsten, K. Von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. Von Heijne, J. Nielsen, F. Pontén, Tissue-based map of the human proteome. *Science*. **347**, 1260419 (2015).

21. R. Petryszak, M. Keays, Y. A. Tang, N. A. Fonseca, E. Barrera, T. Burdett, A. Füllgrabe, A. M. P. Fuentes, S. Jupp, S. Koskinen, O. Mannion, L. Huerta, K. Megy, C. Snow, E. Williams, M. Barzine, E. Hastings, H. Weisser, J. Wright, P. Jaiswal, W. Huber, J. Choudhary, H. E. Parkinson, A. Brazma, Expression Atlas update - An integrated database of gene and protein expression in humans, animals and plants. *Nucleic Acids Research*. **44**, D746–D752 (2016).