Predicting bioactive peptides from transcriptome assemblies with the peptigate workflow

Peptigate predicts bioactive peptides from transcriptomes. It integrates existing tools to predict sORF-encoded peptides, cleavage peptides, and RiPPs, then annotates them for bioactivity and other properties. We welcome feedback on expanding its capabilities.

Published Aug 7, 2024

Arcadia Science

DOI: 10.57844/arcadia-6500-9be8

Purpose

Peptides are small protein sequences (less than 100 amino acids in length) with significant therapeutic and biotechnological potential due to their small size and the wide variety of biological pathways they participate in. Despite these appealing traits, experimental discovery of peptides remains challenging, and computational tools suffer from false positives.

In this pub, we introduce peptigate (peptide + investigate), a workflow that predicts and annotates bioactive peptides from transcriptomes. Peptigate unites functions previously distributed across different tools. It predicts small open reading frame (sORF)-encoded proteins, cleavage peptides, and ribosomally synthesized and post-translationally modified peptides (RiPPs) from transcriptomes. Peptigate then annotates them for bioactivity, chemical properties, similarity to known sequences, and signal peptide presence.

We used peptigate to predict peptides in the human transcriptome, resulting in 2,949 distinct peptides. Comparing these predictions against experimental datasets, we validated an average of 23% of peptides (49% general cleavage, 20%

RiPPs, 22% sORF-encoded peptides). A major challenge during this project was the lack of gold-standard data for validation, as peptide annotations are incomplete even for humans. We used noisy and incomplete proxies like mass spectrometry peptidomics databases and ribosomal profiling. With only a quarter of our predictions confirmed, it's unclear whether mismatches arise from gaps in these data sources or incorrect predictions. We welcome suggestions for more reliable ground truth data to improve our pipeline's assessment.

We anticipate that peptigate may be a jumping-off point for new peptide discovery. For example, if a researcher is interested in identifying peptides in a tumor microenvironment, they might interact with peptigate as follows. First, the researcher would identify a transcriptome or group of transcriptomes from their tumor and non-tumor samples. Next, they'd run peptigate on the transcriptomes. Using the peptigate output, they'd filter to peptides that are only present in the tumor samples or perform differential expression analysis and retain transcripts that encode peptides that are differentially expressed in the tumor microenvironment.

What else could they do with this information? The researcher could use the metadata reported by peptigate to form a hypothesis about the cellular role of these peptides. For example, if the peptides contain a secretory pathway-targeting signal peptide, they're likely secreted and interact with other cells. Using these predictions, the researcher could design wet-lab experiments to follow up on their research interests.

What do you think?

If you think this example resonates with work you're doing, we'd love to hear about it and possibly help. We are also open to learning about other peptigate use cases that others come up with.

- This pub is part of the **project**, "Software: Implementing useful and innovative computing." Visit the project narrative for more background and context.

- The **peptigate pipeline** is available in this GitHub repository.
- The **code** and **associated data** we used to evaluate the peptigate pipeline are available in this GitHub repository, including the results and evaluation of running peptigate on the human RefSeq transcriptome.

The context

Peptides are a diverse class of biological molecules present in all three domains of life. They participate in activities like cellular signaling [1], chemical messaging [2], and defense/immunity [3][4]. Peptide synthesis occurs via many pathways, including ribosomal synthesis of small open reading frames (sORFs) [5], cleavage from precursor proteins, and synthesis by non-ribosomal enzymes [6].

Due to their high specificity and potency, peptides are increasingly recognized for their therapeutic and biotechnological potential. When compared to small molecules, peptides offer the advantages of lower toxicity and relative ease of synthesis. However, they often face challenges such as a short half-life and the requirement for non-oral delivery methods to bypass digestive degradation and effectively reach target tissues [7]. In contrast to other biologics like monoclonal antibodies, peptides theoretically benefit from simpler synthesis processes, shorter research and development phases, and faster regulatory approvals. Despite these advantages, peptides generally exhibit lower stability during storage and handling, similar to the stability issues observed with biologics, necessitating advanced formulation strategies to ensure efficacy.

Our working definition of “peptide”

While the definition of a peptide varies, for this pub, we’ll define peptides as small polypeptides comprised of 2–100 amino acids with standalone biological activity. We refer to these peptides as “bioactive” to denote their distinct physiological functions, unlike peptide fragments from protein degradation or those that don't function independently, such as intermediary or cleaved signal peptides [8][9].

The problem

Endogenous peptide discovery is difficult, especially when predicting many peptides from many species. Before the advent of DNA sequencing technologies, peptide discovery was primarily an experimental endeavor. Early discoveries in the first half of the 20th century focused on single peptides implicated in specific biological actions [10][11][12][13]. Advances in chromatography and mass spectrometry ushered in a high-throughput discovery era via peptidomics [14][15]. Discoveries facilitated by these technologies as well as genome sequencing highlighted the underappreciation of peptides as a biological class [16].

In the intervening decades, further refinement of these technologies and appreciation of different ways peptides are synthesized endogenously have led to more discoveries of peptides [17][18]. Even still, blind spots persist. Some peptides are only present under hyper-specific conditions [19], while peptidomics and ribosomal profiling require expensive infrastructure and expertise and may require sample-specific preparation techniques that limit usability for new sample types [20][21][22][23].

Computational tools address this experimental bottleneck by predicting peptides from genomes and transcriptomes [17]. Sequencing data in particular is amenable to peptide discovery because it can be analyzed in many different ways, which fits with the natural diversity of peptides themselves; multiple tools can detect different types of peptides. However, detecting peptides from sequencing data is still fairly challenging. Apart from the many different kinds of peptides, the short nature of peptide sequences makes them difficult to detect and makes detection susceptible to false positives [5].

Our solution

We introduce peptigate, a workflow that applies previously developed best-in-class tools to predict and annotate diverse bioactive peptides from transcriptomes (Figure 1). “Peptigate” is a portmanteau of *peptide* and *investigate*. Peptigate currently predicts sORF-encoded proteins, cleavage peptides, and ribosomally synthesized and post-translationally modified peptides (RiPPs). These peptides are then annotated to predict bioactivity, chemical properties, similarity to known peptide sequences, and the presence of a signal peptide. These functions were previously scattered in disparate tools; peptigate unites them to make diverse peptide prediction simpler.

For multiple reasons, we chose to use transcriptomes as the input. RNA-seq data, and thus transcriptome assemblies, are comparatively more available than genomes, especially for less developed research organisms. Transcriptomes are also smaller and have a higher ratio of gene content than genomes, which reduces false positives in peptide discovery. However, to make peptigate more flexible, we also provide a reduced pipeline that takes predicted protein sequences as input.

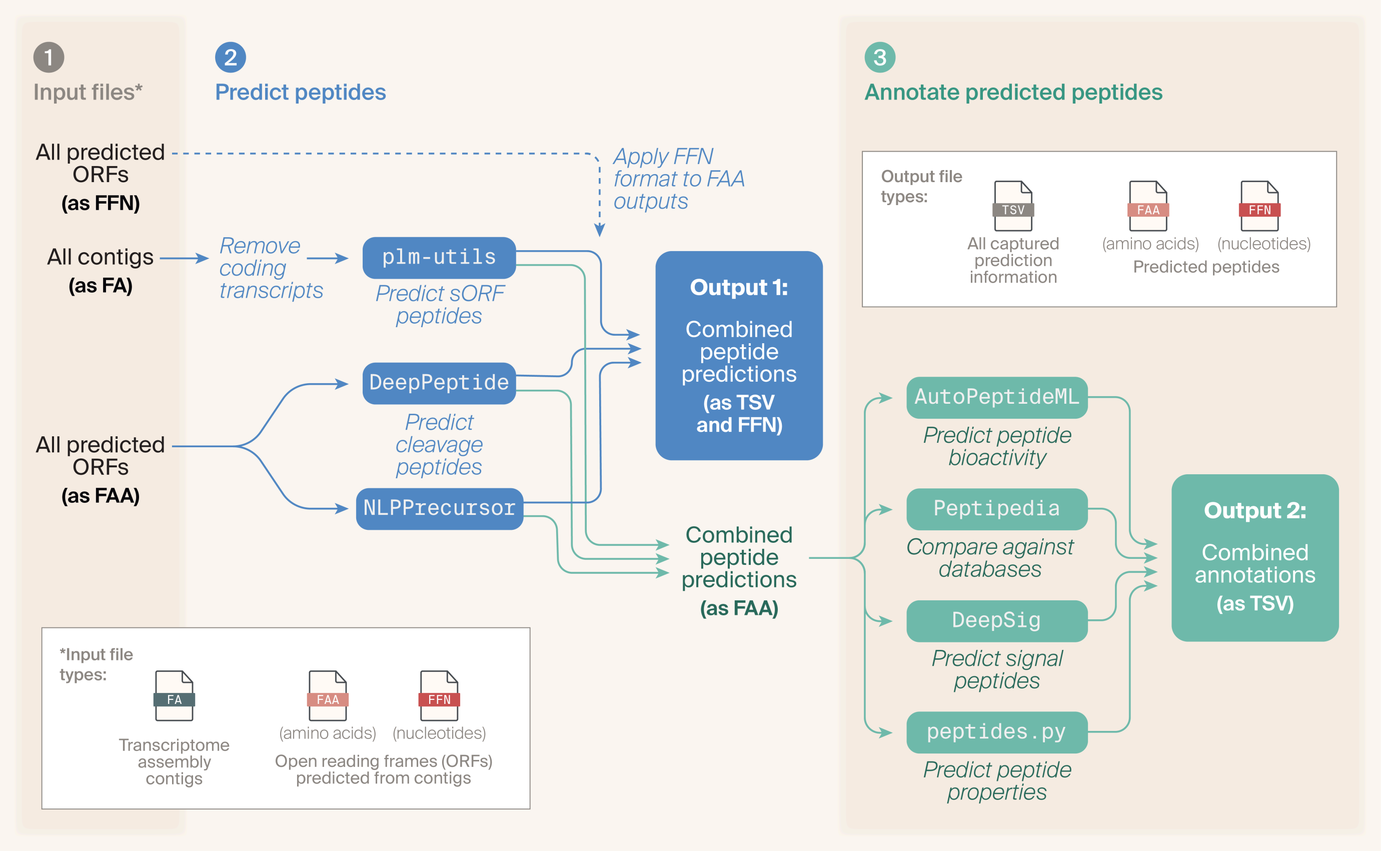
1
Input files*
2
Predict peptides
3
Annotate predicted peptides
All predicted ORFs (as FFN)
Apply FFN format to FAA outputs
All contigs (as FA)
Remove coding transcripts
plm-utils
Predict sORF peptides
DeepPeptide
Predict cleavage peptides
All predicted ORFs (as FAA)
NLPPrecursor
Output 1: Combined peptide predictions (as TSV and FFN)
Combined peptide predictions (as FAA)
Output file types:
TSV
All captured prediction information
FAA
(amino acids)
FFN
(nucleotides)
Predicted peptides
AutoPeptideML
Predict peptide bioactivity
Peptipedia
Compare against databases
DeepSig
Predict signal peptides
peptides.py
Predict peptide properties
Output 2: Combined annotations (as TSV)
*Input file types:
FA
Transcriptome assembly contigs
FAA
(amino acids)
FFN
(nucleotides)
Open reading frames (ORFs) predicted from contigs

Figure 1. **An overview of the peptigate workflow for predicting bioactive peptides**. Peptigate takes a transcriptome assembly and open reading frames (ORFs) predicted from the transcriptome contigs as input. It uses these files to predict sORF peptides with plm-utils, cleavage peptides with DeepPeptide, and RiPP peptides with NLPPrecursor. Predicted peptides are then annotated for bioactivity using AutoPeptideML, compared to known peptides in the metadatabase Peptipedia, annotated for signal peptides with DeepSig, and chemical properties calculated with the Python package peptides.py. The peptide prediction and annotation outputs are reported in a pair of TSV files. The predicted peptide sequences are also provided in nucleotide (FFN) and amino acid (FAA) format for convenience. We've omitted many intermediate steps in the workflow to focus on the parts of the workflow that perform predictive tasks.

The resource

The **peptigate pipeline** is available in this GitHub repository (DOI: 10.5281/zenodo.12775316).

Peptigate is a Snakemake pipeline that combines existing tools to predict bioactive peptides from transcriptomes. Below, we highlight how each part of peptide prediction works, covering sORF, cleavage, and RiPP peptide prediction and annotation.

Predicting small open reading frames

Background on sORFs

Small open reading frames (sORFs) encode peptides that are short upon synthesis [24] (rather than cleaved later). They're also known as "short" open reading frames [25]. The functional peptide products are referred to as sORF-encoded polypeptides (SEPs or sPEPs), microproteins, or micropeptides [26]. DNA transcription and ribosomal translation from open reading frames 300 nucleotides or shorter produces these peptides. Most sORFs use non-traditional start codons like UUG, CUG, GUG, and ACG, each of which differs by one nucleotide from the start codon AUG [27].

While most genomes contain many sORFs, only a few are actively translated and transcribed. Most transcribed sORFs are within a transcript's 5′ or 3′ UTR of the primary coding domain sequence (uORF and dORF, respectively). They often play regulatory roles by influencing the translation of the mRNA [5]. However, some sORFs encode peptides that are translated into functional small proteins. The majority of these sORFs have been identified in what were presumed to be long non-coding RNAs [5][28].

How peptigate predicts sORFs

The peptigate pipeline targets sORFs on contigs without longer ORFs to identify sORFs that encode functional peptides (as opposed to translation regulators). The pipeline begins sORF prediction by removing contigs in the transcriptome assembly that have predicted open reading frames (supplied by the user). Next, the pipeline tries to remove fragmented contigs that likely contain longer open reading frames by comparing each remaining transcript against the UniRef50 database using DIAMOND `blastp` [29]. If a contig has a match to a protein in UniRef50 that's longer than 300 nucleotides, we remove these transcripts. Peptigate then scans the remaining contigs for open reading frames using common sORF start codons (AUG, UUG, CUT, GUG, and ACG [27]) and retains all predicted ORFs 300 nucleotides or shorter. It then predicts whether each sequence is coding or non-coding using the Python package plm-utils [30]. Plm-utils uses latent information in the large protein language model Evolutionary Scale Modeling (ESM) [31][32] to determine whether a short open reading frame is coding or non-coding [30].

Predicting cleavage peptides

Background on cleavage peptides

Cleavage peptides are generated by enzymatic cleavage (proteolysis) of precursor proteins. These peptides are initially ribosomally translated while embedded in the precursor protein and then cleaved to become biologically active. Peptides can be proteolytically released from proteins by specific [33] or general proteases [34] or receive additional modifications after cleavage [35]. Cleavage peptides participate in a variety of biological tasks, including stress response (corticotropin-releasing hormone), blood sugar regulation (insulin and glucagon), blood clotting (thrombin), and inflammation (C3a), and phagocytosis (C3b).

Cleavage peptides are different from propeptides and proteolytic degradation products. Propeptides are parts of proteins that are cleaved during protein maturation and don’t have a biological function once cleaved. Similarly, proteolytic degradation products are generated by the ubiquitin or lysosomal pathways and mostly don't generate functional products, although individual amino acids are often recycled for new protein synthesis [36][37].

How peptigate predicts cleavage peptides

The peptigate pipeline predicts two classes of cleavage peptides: cleavage peptides with protease cut sites as well as ribosomally synthesized and post-translationally modified peptides (RiPPs). Peptigate uses the DeepPeptide tool to identify cleavage peptides with protease cut sites [38]. DeepPeptide is built atop the ESM2 large protein language model [31][32] and predicts peptides and propeptides from protein sequences. The peptides range in length from 5–50 amino acids.

Peptigate uses NLPPrecursor for RiPP prediction [39]. NLPPrecursor was trained using only bacterial RiPP sequences and thus may work best when run on bacterial protein sequences [39]. However, many cyclic eukaryotic peptides are

RiPPs [40]. When run against eukaryotic protein sequences, we think it's possible that the RiPP peptides detected were once horizontally transferred from bacteria to eukaryotes; however, we haven't followed up on this hypothesis.

Annotating predicted peptide sequences

Functional annotation of peptide sequences is difficult for many reasons. Most protein functional annotation tools use sequence similarity or orthology to compare new protein sequences to proteins of known function. These methods often generate statistically unreliable results when applied to short sequences; for short sequences, sequence similarity comparisons typically only work to find matches that are very similar to sequences that have already been discovered [41]. In some species, peptides encoded by sORFs are under lower purifying selection [42], or they’re evolutionarily young so they’re not present in other closely related species [25], decreasing sequence similarity.

Moreover, peptides can exhibit varied functions in different biological contexts due to their ability to adopt multiple conformations [43], which complicates functional annotation based solely on sequence similarity. Because some peptide functions, such as antimicrobial activity, are easier to assay, these functions may improperly propagate even though they don't reflect *in vivo* functions [44].

Peptigate attempts to overcome these challenges by annotating predicted peptide sequences using multiple approaches. First, peptigate compares against known peptide sequences by BLASTing each predicted peptide sequence against the Peptipedia database using DIAMOND `blastp` [29][45][46]. Peptipedia is a metadatabase with peptide sequences from 76 databases encompassing 213 bioactivities (as of March 23, 2024). Peptigate reports the top match for each peptide.

Next, peptigate annotates signal peptides in predicted peptide sequences using DeepSig [47]. Signal peptides are short peptide sequences (16–30 amino acids long) that mark proteins for secretion [48]. Signal peptides can provide clues as to the function of a protein depending on the presence and the class of the signal peptide [49].

Peptigate predicts the function of predicted peptide sequences using AutoPeptideML [50]. AutoPeptideML is a tool that allows users to build and use models for peptide bioactivity prediction through machine learning best practices. It uses the ESM large protein language model (ESM2-8M) internally to improve prediction accuracy [31][32]. Currently, peptigate uses 16 models built in the AutoPeptideML preprint (antibiotic, anticancer, ACE inhibitor, antifungal, anti-MRSA, antimalarial, antimicrobial, antioxidant, antiparasitic, antiviral, blood-brain barrier crossing, neuropeptide, quorum sensing, toxic, and tumor t-cell antigen) [50]. However, the Matthews correlation coefficient of these models ranges from approximately 0.02 to 0.73, indicating a wide performance range and a general inability to predict peptide bioactivity. Nevertheless, this approach is state-of-the-art, so we’ve included it in the peptigate pipeline.

Peptigate also calculates peptide chemical characteristics using the Python package peptides.py. Peptigate calculates metrics like molecular weight, charge, and hydrophobicity. These attributes can be used to compare peptides or to assess whether a given peptide is suitable for a downstream task (e.g., removing hydrophobic peptides because they’ll be difficult to synthesize).

Last, peptigate determines the nucleotide sequences that encode the predicted peptide protein sequences. The nucleotide sequences are three times as long as the amino acid sequences, which can improve sequence searches against large databases and other comparisons. Peptigate doesn't use these sequences directly for annotation, but they're provided to the user as an output so they can be further analyzed (e.g., via sequence similarity clustering with MMseqs2 [51]).

We think there's still room for improvement in our approach to peptide annotation, especially for bioactivity prediction. We welcome feedback or suggestions on how to improve our approach.

Limitations of the peptigate pipeline

While we tried to generate a comprehensive tool, peptigate is still limited. Below, we outline specific tasks that peptigate doesn't yet perform and highlight why including these approaches is difficult.

Peptigate doesn't predict non-bioactive peptides. It's focused on predicting bioactive peptides, so it doesn't predict degradation products from the ubiquitin or lysosomal degradation pathways or digestion (e.g., tryptic cleavage). It also doesn't predict sORFs that occur in the 5′ or 3′ UTR of longer ORFs, as most of these sORFs regulate the translation of the transcripts they occur in and don't have bioactivity beyond this niche role [5].

There are also some classes of bioactive peptides that peptigate doesn't yet predict. In particular, peptigate doesn't predict nonribosomal peptides synthesized by nonribosomal peptide synthetase(s) (NRPSs). NRPSs synthesize peptides independent of messenger RNA and ribosomes. Each enzyme typically contains multiple catalytic domains that help accomplish a specific peptide synthesis step. Multiple NRPS enzymes are usually required to synthesize a peptide, and these enzymes are usually co-located together in the genome (and co-expressed on polycistronic transcripts in the case of bacteria). We didn't include NRPS prediction in peptigate because we were unsure how to identify which NRPS enzymes belong to a single NRPS peptide. We were also unsure if we'd be able to predict the peptide sequence generated through this mechanism.

There are also several annotation tasks that peptigate doesn't currently perform. In general, we omitted tools that are only accessible through a browser, don't have commercial-compatible licenses, or aren’t easily installable through a package manager or a container. We considered including the tools DeepLoc to predict the sub-cellular localization of a peptide [52], PeptideRanker to assess the likelihood that a peptide is bioactive [53], and PepScore to assess whether a peptide is stable in humans [54], but ultimately didn’t include them. We're also interested in predicting the immunogenicity of peptide predictions but didn't find an accurate tool for this.

Peptigate pipeline inputs and outputs

The peptigate pipeline takes three user-provided input files: a transcriptome assembly and annotated ORFs from that assembly in both amino acid and nucleotide format. These files are then used to predict sORF and cleavage peptides.

Peptigate also relies on databases and models. These are either packaged in the peptigate repository or the pipeline downloads them. The sORF prediction tool plm-utils, the cleavage peptide prediction tool NLPPrecursor, and the bioactivity annotation tool AutoPeptideML all require model weights. The plm-utils model is packaged in the peptigate GitHub repository, while the pipeline downloads the AutoPeptideML and NLPPrecursor models. Peptigate also downloads the two databases on which it depends, UniRef50 and Peptipedia. Once downloaded and prepared, the peptigate pipeline will use these same files repeatedly unless they're moved or changed.

Peptigate outputs six files, two FASTA files, and four TSV files. The two main outputs are a pair of TSV files, “peptide_predictions.tsv” and “peptide_annotations.tsv.” The predictions file provides the peptide identifiers, sequences, and the tools that predicted each peptide. The second annotation file provides information from each annotation approach discussed above. The FASTA files and the partner TSV files provide the predicted peptides’ amino acid and nucleotide sequences.

We also adapted peptigate to run when the user only has protein sequences as input. In this scenario, peptigate predicts sORF proteins by length-filtering to proteins less than 100 amino acids. Cleavage peptide prediction and annotation proceed as in the main pipeline, although without nucleotide reporting.

Evaluating the peptigate pipeline

> The **code** and **associated data** we used to evaluate the peptigate pipeline are available in this GitHub repository (DOI: 10.5281/zenodo.13239486), including the results and evaluation of running peptigate on the human RefSeq transcriptome.

We used peptigate to predict peptides in the human transcriptome to understand the tool’s accuracy. Starting from the human RefSeq transcriptome (click here to download the transcriptome), we predicted open reading frames using TransDecoder. We recognize that this approach doesn't fully take advantage of existing annotations for the human transcriptome, but it matches our recommended preprocessing for peptigate. Peptigate predicted 4,235 distinct peptides in the human transcriptome (Table 1). After removing DeepPeptide-predicted propeptides — a part of a protein cleaved during activation or maturation that lacks independent function — 2,949 peptide sequences remained.

We next wanted to evaluate the accuracy of these predictions. Because not all human peptides have been annotated, we lacked a ground truth against which to compare our peptide predictions. We decided to compare the predicted peptide sequences against orthogonal data sources such as databases of previously observed peptides, public annotations, and ribosomal profiling data. We reasoned that if we observed matches between these data sources and our predictions, this would provide evidence that the peptide is likely real. However, this approach is flawed because any disagreement could mean that databases are incomplete, our predictions are at least partially wrong, or some combination of the two. Even still, we moved forward with this approach because we were unable to identify a better gold standard dataset for evaluation.

Prediction tool within peptigate pipeline	Number of predicted peptides	Peptipedia	NCBI metadata	RibORF	Total (distinct)
DeepPeptide (predicts cleavage peptides)	263	130	NA	NA	130 (49%)
NLPPrecursor (predicts RiPPs)	431	87	NA	NA	87 (20%)
plm-utils (predicts sORFs)	2,255	291	287	288	486 (22%)
Total	2,949	508	287	288	703 (24%)

Table 1. **Summary of peptides predicted by peptigate and orthogonal validation information**.

“NA” indicates that orthogonal information wasn't available. “Distinct” refers to distinct amino acid sequences; each sequence is counted once even if it’s validated by multiple datasets. It represents the fraction of predicted peptides validated by orthogonal datasets.

We started by comparing peptigate's predictions to peptides in the Peptipedia database [46]. Peptipedia is a metadatabase comprised of peptides from 76 databases, including human peptide-containing databases like Peptide Atlas [55]. Using the annotation results generated by the peptigate pipeline, we checked whether the predicted peptides had a hit against any peptide in Peptipedia. More cleavage peptides had hits to peptides in Peptipedia than sORF peptides: 49% of peptides predicted by DeepPeptide, 20% of peptides predicted by NLPPrecursor, and 13% of peptides predicted by plm-utils had hits against at least one peptide in the database (Table 1). Our findings suggest that at least one-quarter of peptigate-predicted peptides are likely real.

View the analysis code we used to investigate peptide matches against the Peptipedia peptide database.

For cleavage peptides, we expected to predict more peptides than are present in databases because the DeepPeptide paper predicted 1.3× the known number of peptides in humans (352 in UniProt, 458 predicted) [38]. To determine whether predicted cleavage peptides that didn't have matches in the Peptipedia database might still be real, we looked for signals associated with cleavage peptides. For example, most (but not all) annotated cleavage peptides are cleaved from precursor proteins that contain an N-terminal signal peptide [56]. Signal peptides target a protein to the secretory pathway and allow cleaved peptides to reach their final destination [57]. Many cleavage peptides function as hormones or other signaling molecules, making export from the cell a key step in their biogenesis [57]. Of the 133 predicted peptides with no BLAST hit, 28 are predicted from precursor proteins with signal peptides. We also investigated whether the precursor proteins contained propeptides, as many precursor proteins contain these constructs that help with protein folding, stability, or targeting [58]. A further eight precursor proteins contained propeptides. These results suggest that some cleavage peptide predictions that didn't match known peptides are biologically plausible.

View the analysis code we used to identify signal peptides and propeptides in the precursor proteins of cleavage peptides.

We anticipated that sORF-encoded peptides would have a lower hit rate than cleavage peptides when compared against peptides in databases. While Peptipedia contains 76 databases, it doesn’t include dedicated sORF catalogs like sORFs.org [59]. Further, cleavage peptides were discovered many decades before sORF-encoded peptides [60][61][62][63], and so we expect more cleavage peptides to be annotated than sORFs. In addition, many sORFs are thought to be evolutionarily young [25], meaning we wouldn’t expect homology to peptides from other species. Even still, because so few sORF-encoded peptides had matches against the Peptipedia database, we next focused on validating this class of peptide predictions.

We first looked at the annotations for each transcript. Since we started with a RefSeq transcriptome, all transcripts are labeled as curated coding, curated non-coding, predicted coding, or predicted non-coding by their accession number. Of the 2,255 predicted sORF-encoded peptides, 13% are labeled as curated coding (Table 1). We anticipate that many more transcripts are actually coding; recent research has shown that many transcripts labeled as non-coding actually contain sORFs that encode peptides [64][65][66][67][68][69][70][71][72]. However, the observed overlap validates a subset of our sORF predictions and demonstrates that the Peptipedia database is partially incomplete with regard to sORF-encoded peptides with known coding potential.

Given that Peptipedia is incomplete with regards to sORFs, we tested how many predicted sORF-encoded peptides are supported by ribosomal profiling data. Ribosome profiling data is generated by sequencing fragments of mRNA that are protected by ribosomes, offering a snapshot of translation in action [73] — if one of our predicted sORF-encoded peptides appears in a ribosome profiling dataset, it would lend credence to the idea that this is a real, translated peptide. A recent set of papers developed a tool called RibORF that predicts open reading frames from ribosomal profiling data and uses this tool to re-analyze over 600 ribosomal profiling datasets from humans [54][74]. 13% of sORF-predicted peptides overlapped with RibORF predictions (Table 1), 265 (189 canonical, 61 non-coding, nine extension, and six truncation). This overlap supports the idea that these sORFs are translated into proteins.

View the analysis code we used to compare sORF-encoded peptides against ribosomal profiling data.

The fraction of sORF-predicted peptides that appeared in ribosomal profiling data underwhelmed us, so we tried to validate these sequences using other orthogonal datasets. First, we checked whether peptigate predicted true non-coding RNAs as coding. Of the three we tested (XIST, HOTAIR, NEAT1), peptigate predicted none to be coding. These findings confirm that peptigate effectively discriminates between coding and non-coding RNAs.

View the analysis code we used to search for non-coding RNAs in sORF-encoded peptides.

We next wanted to measure the relative translation potential of the predicted sORFs. If an sORF is able to recruit a ribosome for translation, it's potentially more likely to be translated into a protein. To estimate translation potential, we measured the Kozak sequence similarity score for each predicted sORF and compared the distribution against ORFs > 300 nucleotides in the human transcriptome. The Kozak consensus sequence functions as a translation initiation start site and enhances translation efficiency by directing ribosomes to the correct start codon [75]. Six base pairs occur upstream and one base pair downstream of the start codon in a transcript [75]. The exact sequence varies, so each Kozak sequence can be scored in comparison to the most common sequence motif [76]. We scored each Kozak sequence as performed in [76]: using the sequence motif GccA/Gcc**AUG**G, we designated upper-case letters as highly conserved (scored +3) and lower-case letters as common (scored +1). We didn't score the start codon (bolded letters). The maximum score is 13. On average and across transcript types (inherited from RefSeq labels), sORFs have lower Kozak sequence scores than other transcripts (Welch’s two-sample t-test, estimate = 1.4, $p < 0.001$, 95% CI [0.8, 1.07]). However, the sORF Kozak sequence scores occurred within the same range as those of other transcripts, with both coding and non-coding sequences achieving the maximum Kozak sequence score of 13. Given the range of Kozak scores observed, these results suggest that some predicted sORFs are likely to recruit ribosomes and be translated into proteins.

> View the analysis code for calculating and comparing Kozak sequence scores in sORF-encoded peptides versus normal open reading frames.

Overall, we struggled to identify a gold-standard, ground-truth dataset to use when evaluating peptigate. It's unclear to us what a "good" expected hit rate is against different orthogonal datasets. We expect some hits, as we'd expect some fraction of our predicted peptides to have been previously discovered. However, it's unclear how many bioactive peptides exist or how many have been discovered. A peptidomics mass spectrometry and machine learning paper published in 2022 suggested that, to date, only 300 peptides in humans have confirmed bioactivity [56], so our predictions aren't many orders of magnitude away from what we might expect, and there may be room for new human peptide discovery. We welcome suggestions for different validation datasets that can be used to validate computational peptide predictions.

Additional methods

We used ChatGPT to help refactor some Python scripts executed by the Snakefile, write first drafts of doc strings, and clean up character lines to reduce them to under 100 characters. We also used ChatGPT and Notion AI to suggest wording ideas, and then we chose which small phrases or sentence structure ideas to use.

Key takeaways

1. Peptigate is a workflow for predicting and annotating bioactive peptides from transcriptomes. It combines existing state-of-the-art tools to predict peptides encoded by small open reading frames and cleavage peptides. It

annotates predicted peptides to provide insights into their potential function.

2. Peptigate is designed to better inform researchers as they make decisions about follow-up functional studies. This may require multiple peptigate prediction runs across diverse transcriptomes or additional prediction tasks. For example, if a researcher is interested in a specific bioactivity that isn't tested in peptigate, it may be useful to build additional bioactivity prediction models with AutoPeptideML.
3. Only about a quarter of peptigate predictions match peptides predicted in orthogonal datasets, highlighting a need for more comprehensive and reliable validation methods and datasets.

Next steps

1. **Identifying ground-truth data**. One of the things we struggled with during this project was a lack of gold-standard data for prediction. Given that peptide annotations are incomplete, even for the human genome and proteome, it wasn't clear what to use as ground truth, true positive, and true negative data. We used orthogonal datasets like mass spectrometry peptidomics databases and ribosomal profiling as proxies, but these datasets are noisy and incomplete. We'd love new ideas for ground truth data we can use to assess our pipeline.
2. **Improving bioactivity annotations**. Bioactive peptides participate in almost all aspects of metabolism, making them interesting for both basic and translational research. Even if we can produce confident peptide sequence predictions, it’s difficult to computationally predict the bioactivity of those sequences because of their short length. We're interested in identifying new tools or orthogonal tests that we can incorporate into peptigate to improve bioactivity annotations.
3. **Including more tools for peptide prediction and annotation**. As described in the “Limitations…” section above, peptigate doesn't predict all types of peptides or incorporate all possible annotation tools. We'd like to expand the types of peptides and annotations included if we can overcome the challenges outlined in the limitations section.
4. **Making the pipeline easier to use**. We wrote peptigate as an experimental pipeline. While we tried to assemble a reasonable pipeline, we identified many areas where we could improve the quality of our software engineering.

If peptigate proves useful, we plan to improve the quality of the software by adding things like installation from a package manager and automated tests.

Contributors (A–Z)

- **Audrey Bell**: Visualization
- **Adair L. Borges**: Supervision
- **Feridun Mert Celebi**: Supervision
- **Keith Cheveralls**: Methodology, Software
- **Seemay Chou**: Conceptualization, Supervision
- **Megan L. Hochstrasser**: Editing
- **Taylor Reiter**: Formal Analysis, Investigation, Methodology, Software, Visualization, Writing
- **Emily C.P. Weiss**: Conceptualization, Critical Feedback

References

1. Nässel DR, Larhammar D. (2013). Neuropeptides and Peptide Hormones. https://doi.org/10.1007/978-3-642-10769-6_11
2. Verbeke F, De Craemer S, Debunne N, Janssens Y, Wynendaele E, Van de Wiele C, De Spiegeleer B. (2017). Peptides as Quorum Sensing Molecules: Measurement Techniques and Obtained Levels In vitro and In vivo. https://doi.org/10.3389/fnins.2017.00183
3. Robinson S, Norton R. (2014). Conotoxin Gene Superfamilies. https://doi.org/10.3390/md12126058
4. Zhang L-j, Gallo RL. (2016). Antimicrobial peptides. https://doi.org/10.1016/j.cub.2015.11.017
5. Andrews SJ, Rothnagel JA. (2014). Emerging evidence for functional peptides encoded by short open reading frames. https://doi.org/10.1038/nrg3520
6. Dell M, Dunbar KL, Hertweck C. (2022). Ribosome-independent peptide biosynthesis: the challenge of a unifying nomenclature. https://doi.org/10.1039/d1np00019e
7. Rossino G, Marchese E, Galli G, Verde F, Finizio M, Serra M, Linciano P, Collina S. (2023). Peptides as Therapeutic Agents: Challenges and Opportunities in the Green Transition Era. https://doi.org/10.3390/molecules28207165

8. Martini S, Tagliazucchi D. (2023). Bioactive Peptides in Human Health and Disease. https://doi.org/10.3390/ijms24065837
9. Wang L, Wang N, Zhang W, Cheng X, Yan Z, Shao G, Wang X, Wang R, Fu C. (2022). Therapeutic peptides: current applications and future directions. https://doi.org/10.1038/s41392-022-00904-4
10. Bayliss WM, Starling EH. (1902). The mechanism of pancreatic secretion. https://doi.org/10.1113/jphysiol.1902.sp000920
11. (1922). On a remarkable bacteriolytic element found in tissues and secretions. https://doi.org/10.1098/rspb.1922.0023
12. Dubos RJ. (1939). STUDIES ON A BACTERICIDAL AGENT EXTRACTED FROM A SOIL BACILLUS. https://doi.org/10.1084/jem.70.1.11
13. Sarges R, Witkop B. (1964). Gramicidin A. IV. Primary Sequence of Valine and Isoleucine Gramicidin A. https://doi.org/10.1021/ja01063a049
14. Schrader M, Schulz-Knappe P, Fricker LD. (2014). Historical perspective of peptidomics. https://doi.org/10.1016/j.euprot.2014.02.014
15. Rodríguez AA, Otero-González A, Ghattas M, Ständker L. (2021). Discovery, Optimization, and Clinical Application of Natural Antimicrobial Peptides. https://doi.org/10.3390/biomedicines9101381
16. Hellinger R, Sigurdsson A, Wu W, Romanova EV, Li L, Sweedler JV, Süssmuth RD, Gruber CW. (2023). Peptidomics. https://doi.org/10.1038/s43586-023-00205-2
17. Tharakan R, Kreimer S, Ubaida-Mohien C, Lavoie J, Olexiouk V, Menschaert G, Ingolia NT, Cole RN, Ishizuka K, Sawa A, Nucifora LG. (2020). A methodology for discovering novel brain-relevant peptides: Combination of ribosome profiling and peptidomics. https://doi.org/10.1016/j.neures.2019.02.006
18. Lai ZW, Petrera A, Schilling O. (2014). The emerging role of the peptidome in biomarker discovery and degradome profiling. https://doi.org/10.1515/hsz-2014-0207
19. Li L, Wu J, Lyon CJ, Jiang L, Hu TY. (2023). Clinical Peptidomics: Advances in Instrumentation, Analyses, and Applications. https://doi.org/10.34133/bmef.0019
20. Meo AD, Pasic MD, Yousef GM. (2016). Proteomics and peptidomics: moving toward precision medicine in urological malignancies. https://doi.org/10.18632/oncotarget.8931
21. Fleites LA, Johnson R, Kruse AR, Nachman RJ, Hall DG, MacCoss M, Heck ML. (2020). Peptidomics Approaches for the Identification of Bioactive

Molecules from Diaphorina citri. https://doi.org/10.1021/acs.jproteome.9b00509

22. Yu Q, OuYang C, Liang Z, Li L. (2014). Mass spectrometric characterization of the crustacean neuropeptidome. https://doi.org/10.1016/j.euprot.2014.02.015
23. Schlesinger D, Elsässer SJ. (2021). Revisiting sORFs: overcoming challenges to identify and characterize functional microproteins. https://doi.org/10.1111/febs.15769
24. Sandmann C-L, Schulz JF, Ruiz-Orera J, Kirchner M, Ziehm M, Adami E, Marczenke M, Christ A, Liebe N, Greiner J, Schoenenberger A, Muecke MB, Liang N, Moritz RL, Sun Z, Deutsch EW, Gotthardt M, Mudge JM, Prensner JR, Willnow TE, Mertins P, van Heesch S, Hubner N. (2023). Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. https://doi.org/10.1016/j.molcel.2023.01.023
25. Wright BW, Yi Z, Weissman JS, Chen J. (2022). The dark proteome: translation from noncanonical open reading frames. https://doi.org/10.1016/j.tcb.2021.10.010
26. Cao X, Slavoff SA. (2020). Non-AUG start codons: Expanding and regulating the small and alternative ORFeome. https://doi.org/10.1016/j.yexcr.2020.111973
27. Yeasmin F, Yada T, Akimitsu N. (2018). Micropeptides Encoded in Transcripts Previously Identified as Long Noncoding RNAs: A New Chapter in Transcriptomics and Proteomics. https://doi.org/10.3389/fgene.2018.00144
28. Buchfink B, Reuter K, Drost H-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. https://doi.org/10.1038/s41592-021-01101-x
29. Borges AL, Celebi FM, Cheveralls K, Reiter T. (2024). Using protein language models to predict coding and non-coding transcripts with plm-utils. https://doi.org/10.57844/arcadia-fa56-ee23
30. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, Fergus R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. https://doi.org/10.1073/pnas.2016239118
31. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. https://doi.org/10.1126/science.ade2574

32. Kim Y-G, Lone AM, Nolte WM, Saghatelian A. (2012). Peptidomics approach to elucidate the proteolytic regulation of bioactive peptides. https://doi.org/10.1073/pnas.1203195109

33. Fitzgerald K. (2020). Furin Protease: From SARS CoV-2 to Anthrax, Diabetes, and Hypertension. https://doi.org/10.7812/tpp/20.187

34. Arnison PG, Bibb MJ, Bierbaum G, Bowers AA, Bugni TS, Bulaj G, Camarero JA, Campopiano DJ, Challis GL, Clardy J, Cotter PD, Craik DJ, Dawson M, Dittmann E, Donadio S, Dorrestein PC, Entian K-D, Fischbach MA, Garavelli JS, Göransson U, Gruber CW, Haft DH, Hemscheidt TK, Hertweck C, Hill C, Horswill AR, Jaspars M, Kelly WL, Klinman JP, Kuipers OP, Link AJ, Liu W, Marahiel MA, Mitchell DA, Moll GN, Moore BS, Müller R, Nair SK, Nes IF, Norris GE, Olivera BM, Onaka H, Patchett ML, Piel J, Reaney MJT, Rebuffat S, Ross RP, Sahl H-G, Schmidt EW, Selsted ME, Severinov K, Shen B, Sivonen K, Smith L, Stein T, Süssmuth RD, Tagg JR, Tang G-L, Truman AW, Vederas JC, Walsh CT, Walton JD, Wenzel SC, Willey JM, van der Donk WA. (2013). Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. https://doi.org/10.1039/c2np20085f

35. Varshavsky A. (2005). Regulated protein degradation. https://doi.org/10.1016/j.tibs.2005.04.005

36. https://www.ncbi.nlm.nih.gov/books/nbk9957/

37. Teufel F, Refsgaard JC, Madsen CT, Stahlhut C, Grønborg M, Winther O, Madsen D. (2023). DeepPeptide predicts cleaved peptides in proteins using conditional random fields. https://doi.org/10.1093/bioinformatics/btad616

38. Merwin NJ, Mousa WK, Dejong CA, Skinnider MA, Cannon MJ, Li H, Dial K, Gunabalasingam M, Johnston C, Magarvey NA. (2019). DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products. https://doi.org/10.1073/pnas.1901493116

39. Luo S, Dong S-H. (2019). Recent Advances in the Discovery and Biosynthetic Study of Eukaryotic RiPP Natural Products. https://doi.org/10.3390/molecules24081541

40. Coelho LP, Santos Júnior CD, de la Fuente Nunez C. (2024). Challenges in computational discovery of bioactive peptides in ’omics data. https://doi.org/10.1002/pmic.202300105

41. Slavoff SA, Mitchell AJ, Schwaid AG, Cabili MN, Ma J, Levin JZ, Karger AD, Budnik BA, Rinn JL, Saghatelian A. (2012). Peptidomic discovery of short

open reading frame–encoded peptides in human cells. https://doi.org/10.1038/nchembio.1120

42. Ghosh G, Barman R, Mukherjee A, Ghosh U, Ghosh S, Fernández G. (2021). Control over Multiple Nano and Secondary Structures in Peptide Self Assembly. https://doi.org/10.1002/anie.202113403
43. Ruiz-Blanco YB, Agüero-Chapin G, Romero-Molina S, Antunes A, Olari L-R, Spellerberg B, Münch J, Sanchez-Garcia E. (2022). ABP-Finder: A Tool to Identify Antibacterial Peptides and the Gram-Staining Type of Targeted Bacteria. https://doi.org/10.3390/antibiotics11121708
44. Quiroz C, Saavedra YB, Armijo-Galdames B, Amado-Hinojosa J, Olivera-Nappa Á, Sanchez-Daza A, Medina-Ortiz D. (2021). Peptipedia: a user-friendly web application and a comprehensive database for peptide research supported by Machine Learning approach. https://doi.org/10.1093/database/baab055
45. Savojardo C, Martelli PL, Fariselli P, Casadio R. (2017). DeepSig: deep learning improves signal peptide detection in proteins. https://doi.org/10.1093/bioinformatics/btx818
46. Owji H, Nezafat N, Negahdaripour M, Hajiebrahimi A, Ghasemi Y. (2018). A comprehensive review of signal peptides: Structure, roles, and applications. https://doi.org/10.1016/j.ejcb.2018.06.003
47. Teufel F, Almagro Armenteros JJ, Johansen AR, Gíslason MH, Pihl SI, Tsirigos KD, Winther O, Brunak S, von Heijne G, Nielsen H. (2022). SignalP 6.0 predicts all five types of signal peptides using protein language models. https://doi.org/10.1038/s41587-021-01156-3
48. Fernandez-Diaz R, Cossio-Pérez R, Agoni C, Lam HT, Lopez V, Shields DC. (2023). AutoPeptideML: A study on how to build more trustworthy peptide bioactivity predictors. https://doi.org/10.1101/2023.11.13.566825
49. Steinegger M, Söding J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. https://doi.org/10.1038/nbt.3988
50. Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. https://doi.org/10.1093/bioinformatics/btx431
51. Mooney C, Haslam NJ, Pollastri G, Shields DC. (2012). Towards the Improved Discovery and Design of Functional Peptides: Common Features of Diverse Classes Permit Generalized Prediction of Bioactivity. https://doi.org/10.1371/journal.pone.0045012
52. Yang H, Li Q, Stroup EK, Wang S, Ji Z. (2024). Widespread stable noncanonical peptides identified by integrated analyses of ribosome

profiling and ORF features. https://doi.org/10.1038/s41467-024-46240-9

53. Desiere F. (2006). The PeptideAtlas project. https://doi.org/10.1093/nar/gkj040
54. Madsen CT, Refsgaard JC, Teufel FG, Kjærulff SK, Wang Z, Meng G, Jessen C, Heljo P, Jiang Q, Zhao X, Wu B, Zhou X, Tang Y, Jeppesen JF, Kelstrup CD, Buckley ST, Tullin S, Nygaard-Jensen J, Chen X, Zhang F, Olsen JV, Han D, Grønborg M, de Lichtenberg U. (2022). Combining mass spectrometry and machine learning to discover bioactive peptides. https://doi.org/10.1038/s41467-022-34031-z
55. Bean AJ, Zhang X, Hökfelt T. (1994). Peptide secretion: what do we know?. https://doi.org/10.1096/fasebj.8.9.8005390
56. Pei J, Kinch LN, Cong Q. (2024). Computational analysis of propeptide containing proteins and prediction of their post cleavage conformation changes. https://doi.org/10.1002/prot.26702
57. Leong AZ-X, Lee PY, Mohtar MA, Syafruddin SE, Pung Y-F, Low TY. (2022). Short open reading frames (sORFs) and microproteins: an update on their identification and validation measures. https://doi.org/10.1186/s12929-022-00802-5
58. Ward CW, Lawrence MC. (2011). Landmarks in Insulin Research. https://doi.org/10.3389/fendo.2011.00076
59. Steiner DF, Cunningham D, Spigelman L, Aten B. (1967). Insulin Biosynthesis: Evidence for a Precursor. https://doi.org/10.1126/science.157.3789.697
60. Anderson D, Anderson K, Chang C-L, Makarewich C, Nelson B, McAnally J, Kasaragod P, Shelton J, Liou J, Bassel-Duby R, Olson E. (2015). A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. https://doi.org/10.1016/j.cell.2015.01.009
61. Jackson R, Kroehling L, Khitun A, Bailis W, Jarret A, York AG, Khan OM, Brewer JR, Skadow MH, Duizer C, Harman CCD, Chang L, Bielecki P, Solis AG, Steach HR, Slavoff S, Flavell RA. (2018). The translation of non-canonical open reading frames controls mucosal immunity. https://doi.org/10.1038/s41586-018-0794-7
62. Tajbakhsh S. (2017). lncRNA-Encoded Polypeptide SPAR(s) with mTORC1 to Regulate Skeletal Muscle Regeneration. https://doi.org/10.1016/j.stem.2017.03.016
63. Nelson BR, Makarewich CA, Anderson DM, Winders BR, Troupes CD, Wu F, Reese AL, McAnally JR, Chen X, Kavalali ET, Cannon SC, Houser SR, Bassel-Duby R, Olson EN. (2016). A peptide encoded by a transcript annotated as

long noncoding RNA enhances SERCA activity in muscle. https://doi.org/10.1126/science.aad4076

64. Min K-W, Davila S, Zealy RW, Lloyd LT, Lee IY, Lee R, Roh KH, Jung A, Jemielity J, Choi E-J, Chang JH, Yoon J-H. (2017). eIF4E phosphorylation by MST1 reduces translation of a subset of mRNAs, but increases lncRNA translation. https://doi.org/10.1016/j.bbagrm.2017.05.002

65. Huang J-Z, Chen M, Chen D, Gao X-C, Zhu S, Huang H, Hu M, Zhu H, Yan G-R. (2017). A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. https://doi.org/10.1016/j.molcel.2017.09.015

66. Wu S, Zhang L, Deng J, Guo B, Li F, Wang Y, Wu R, Zhang S, Lu J, Zhou Y. (2020). A Novel Micropeptide Encoded by Y-Linked LINC00278 Links Cigarette Smoking and AR Signaling in Male Esophageal Squamous Cell Carcinoma. https://doi.org/10.1158/0008-5472.can-19-3440

67. Guo B, Wu S, Zhu X, Zhang L, Deng J, Li F, Wang Y, Zhang S, Wu R, Lu J, Zhou Y. (2019). Micropeptide CIP 2A   BP encoded by LINC 00665 inhibits triple negative breast cancer progression. https://doi.org/10.15252/embj.2019102190

68. Niu L, Lou F, Sun Y, Sun L, Cai X, Liu Z, Zhou H, Wang H, Wang Z, Bai J, Yin Q, Zhang J, Chen L, Peng D, Xu Z, Gao Y, Tang S, Fan L, Wang H. (2020). A micropeptide encoded by lncRNA MIR155HG suppresses autoimmune inflammation via modulating antigen presentation. https://doi.org/10.1126/sciadv.aaz2059

69. Brar GA, Weissman JS. (2015). Ribosome profiling reveals the what, when, where and how of protein synthesis. https://doi.org/10.1038/nrm4069

70. Ji Z. (2018). RibORF: Identifying Genome Wide Translated Open Reading Frames Using Ribosome Profiling. https://doi.org/10.1002/cpmb.67

71. Kozak M. (1987). An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. https://doi.org/10.1093/nar/15.20.8125

72. Castelo-Szekely V, Arpat AB, Janich P, Gatfield D. (2017). Translational contributions to tissue specificity in rhythmic and constitutive gene expression. https://doi.org/10.1186/s13059-017-1222-2