Published on May 31, 2022 by Arcadia Science **DOI**: 10.57844/arcadia-3w15-6f78

Performing mass spectrometry-based proteomics in organisms with minimal reference protein databases

If you're interested in generating proteomics data but your organism of interest doesn't have a sequenced genome to use as a reference database, it is straightforward and useful to collect a transcriptome instead.

Contributors (A-Z)

Seemay Chou, Tori Doran, Behnom Farboud, Juliana Gil, William Hatleberg, Megan L. Hochstrasser, Greg Huber, Kira E. Poskanzer, MaryClare Rollins, Peter S. Thuy-Boun, Elizabeth Tseng, Joan Wong

Version 6 · Mar 31, 2025

Purpose

When we first started planning our project to find useful biomolecules in tick saliva, we struggled with the lack of sequenced genomes and other omics datasets. We were most interested in proteomics, but our tick species of interest lacked a reference database, so we decided to simultaneously develop a transcriptome and a mass spec-based proteome.

We're sharing our method and detailed protocol to make it easier for researchers studying other non-model organisms to apply this approach, and hope it will be especially helpful for those without a background in sequencing or proteomics.

- This pub is part of the **project**, "Ticks as treasure troves: Molecular discovery in new organisms." Visit the project narrative for more background and context.
- We used this method to generate a **dataset** from tick salivary glands, described here.
- This method features a detailed **protocol**, which you can view here.

The problem

Don't need background? Jump to "The method."

Bottom-up, tandem mass spectrometry-based proteomics is a key technology for detecting both protein sequences and post-translational modifications like phosphorylation, sulfation, lipidation, or glycosylation. However, using this technique requires a database containing all protein sequences expected to exist in a biological sample set (Figure 1).

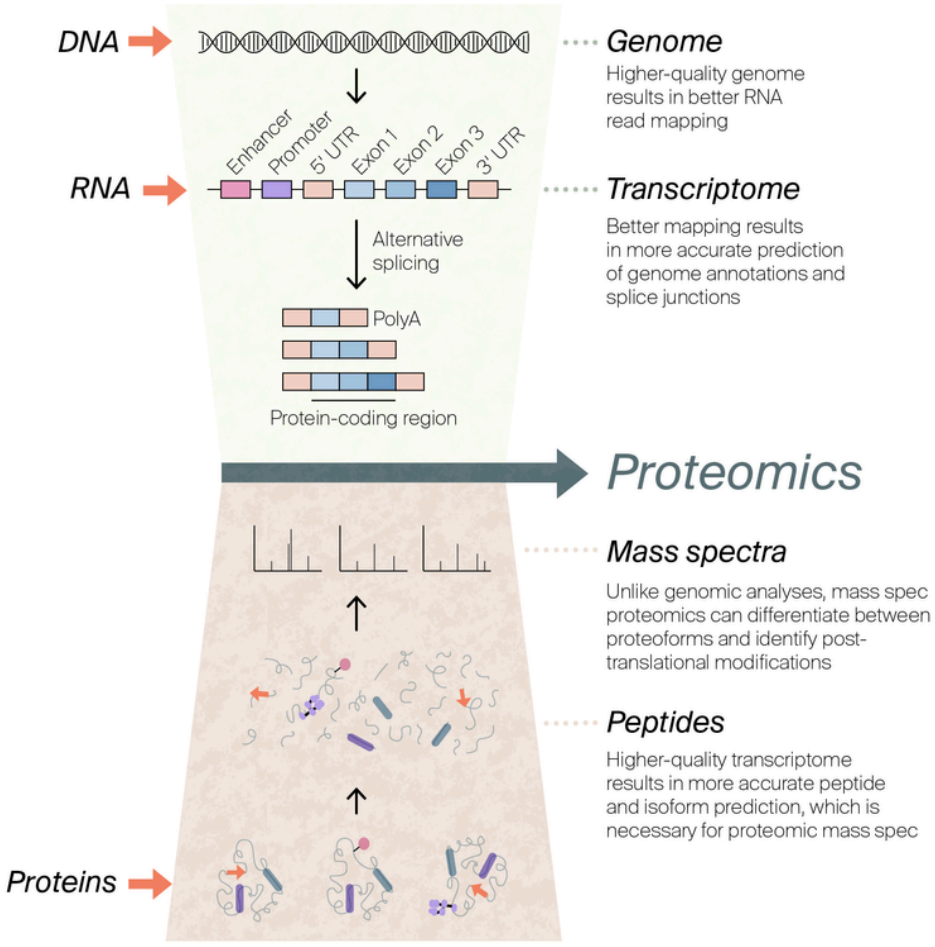

Figure 1

How proteomic information can be generated or inferred, and which types of information depend on each other to be useful.

Experimental mass spectrometry data (lower trapezoid) are decoded using hints generated by genomics and transcriptomics data (upper trapezoid). These two data types converge during the proteomics data analysis process, wherein experimental fragmentation mass spectra are compared to theoretical mass spectra generated from genomic and transcriptomic sequencing experiments.

Why do we need a protein database?

Modern mass spectrometry-based protein identification techniques involve shattering peptides to generate patterns called fragmentation spectra. For the most part, each spectrum (generated from the fragmentation of a given molecule) is unique, like a fingerprint. Fingerprints are useful when we have something to which we can compare them. In that sense, a protein database is like a fingerprint database—it lets us 1) match each experimental spectrum (fingerprint at a crime scene) to a known/predicted peptide (fingerprint in a database), and 2) it tells us what larger protein that peptide came from (whose finger left the print). Additionally, a fingerprint database is very useful for matching imperfect mass spectra, which tend to be the majority of spectra collected. Sometimes peptides fail to fully fragment, yielding small ambiguous segments within a larger sequence. Knowing that these imperfect fragmentation spectra can map to only a limited number of possible peptides gives us confidence in an otherwise ambiguous assignment.

Before we do mass spec, we treat our experimental proteome sample (protein mixture) with a protease that chews all the proteins into smaller fragments, or peptides. Next, the peptides are run through the mass spectrometer, generating a pattern of unique fragmentation spectra. How do we interpret these spectra? When we have a protein database for the organism we're studying, we can computationally predict all the peptide sequences that will result from digesting all possible proteins in the organism), and generate what their fragmentation spectra would look like. By comparing these theoretical spectra to those from our experimental sample, we can decode the signal and deduce which of the reference peptides are actually in our sample. This process can be high-throughput, letting us identify many proteins very quickly.

Many organisms lack reference databases

Well-studied "model organisms" are highly represented in public sequencing repositories and a quick trip to the NCBI or UniProt will likely yield good-quality reference proteomes assembled by other researchers. But for non-model organisms, reference databases are scarce. **The method describes parallel work streams in which we 1) use transcriptomics to build a reference protein database and 2) perform mass spectrometry-based proteomics experiments. The work streams**

converge during data analysis, when we use the new reference protein database to help interpret the mass spec data.

We want to find interesting components of tick saliva, especially those that interact with the human body. We used this new method to generate a dataset from the salivary glands of lone star ticks, but we hope this approach will be broadly useful in enabling proteomics in any organism for which there is a paucity of reference genomic, transcriptomic, or proteomic data.

Why is this useful?

Using mass spectrometry for proteomic analysis is straightforward for organisms with pre-existing reference databases, but most non-model organisms lack such information. **The approach described here lets scientists simultaneously gather new proteomic data from mass spectrometry while doing RNA sequencing to create a protein database to compare with the proteomic data**.

Notably, while many transcriptomics studies rely on short-read RNA sequencing, our method uses long-read sequencing. This can be advantageous for resolving long repetitive genomic regions, speeding up genome assemblies, yielding more complete contigs, and in this case, providing insights into the full structures of transcripts without assembly.

Ultimately, this method generated a robust, long-read, transcriptome-based proteome database that compares reasonably well to pre-existing data. Our approach enabled detection of approximately 9% more peptide spectrum matches (PSMs, the number of experimental spectra that we can match to a theoretical spectrum) and peptides (the number of peptides identified; a given peptide may have many mass spectra) than were represented in the prior database, and favored detection of longer protein sequences, which may enable a more complete understanding of function.

It may be helpful to check out our full description of the tick salivary gland dataset that we generated through this approach.

The strategy

We set out to create a comprehensive method for learning about the proteome in tissues from non-model organisms. We decided to use mass spectrometry to detect proteins in our sample of interest. Because specific protein sequences in mass spec data can generally only be identified by comparing to a reference, we knew we'd also need a reference protein database. There is a paucity of genomic, transcriptomic, and proteomic data for many non-model organisms, so we decided to split our method into two parallel work streams (Figure 2) after initial sample collection: one includes RNA sequencing to develop a reference protein database; the other includes performing proteomic mass spectrometry. The two work streams come together for the final step, data analysis, as the mass spec data is best interpreted using a transcriptome-based protein database.

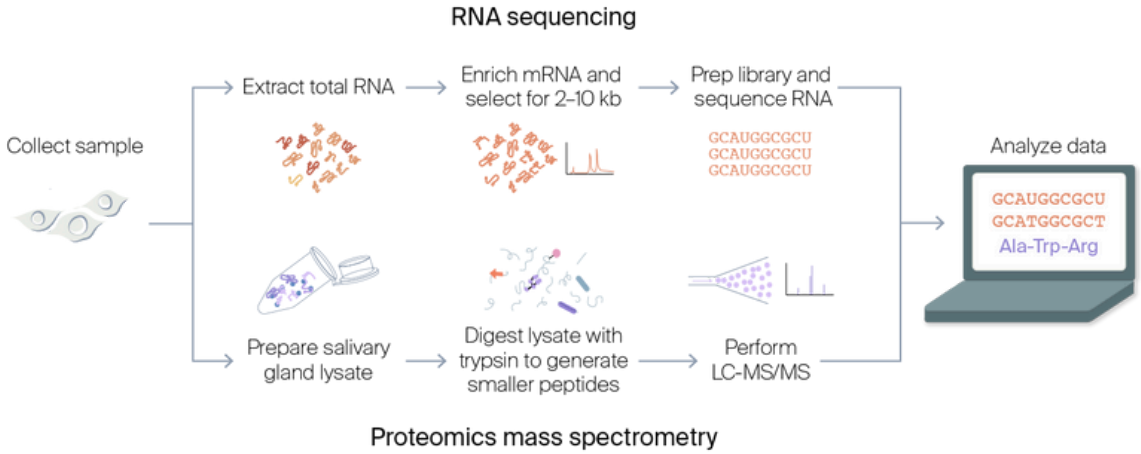

Figure 2

Overview of the parallel transcriptomic (top) and proteomic (bottom) work streams.

We encountered a few key decision points in designing our approach, which are described in depth below (or you can skip to the step-by-step description of the overall method). Let us know if you try this and tweak any of these procedural options—we'd be curious to hear how it may influence the quality or nature of the resulting data.

mRNA enrichment — Poly-A enrichment vs. rRNA depletion

Ribosomal RNA (rRNA) tends to dominate in the total RNA mixture extracted from samples (~80% of total RNA composition) and occludes the protein-coding messenger RNA (mRNA) transcripts that we're interested in profiling. Thus, we needed a way to enrich mRNA. One approach involves the negative enrichment of rRNA, using capture techniques hinged on complementary nucleotides specifically designed for each species's rRNA sequences. The other, more common approach is the positive enrichment of mRNA via oligo-(dT) primers that target mRNA containing poly-A tails. rRNA negative enrichment advantageously enables the detection of non-coding RNA and mRNA without poly-A tails, but comes with the added burden of troubleshooting rRNA probe design for non-model organisms. Since this was our first shot at transcriptome profiling, we took the path of least resistance and performed mRNA poly-A based enrichment using oligo-(dT) probes instead.

RNA sequencing — Long-read vs. short-read

Sequencing technology selection was our most crucial decision point. Illumina powers the dominant platform and enables the assembly of genomes and transcriptomes via highly accurate nucleotide fragments hundreds of base pairs in length (short-read sequencing). In contrast, the dominant long-read sequencing platforms supported by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) are able to assay contiguous nucleotide fragments in the multi-kilobase and megabase range, respectively. PacBio and ONT have lagged behind Illumina over the last decade due to lower-accuracy basecalls and lack of sequencing depth, but recent technological improvements have brought their platforms' sequencing accuracy within competitive range of Illumina.

Long-read sequencing data can provide insights into the full structures of transcripts without assembly. Our interest in full transcript structures brought us to PacBio's relatively mature HiFi Iso-seq methodology as a first choice. In addition, we figured it would provide a great complement to the Mulenga lab's short-read dataset collected on the same tick species **[1]**.

Protein identification — Mass spectrometry vs. immunoprecipitation or Edman degradation

We hope that mass spectrometry will be advantageous in this context because it lets us analyze cell-free secretions. Importantly, it is suited for the detection of non-encoded molecules/modifications, which can include protein post-translational modifications (e.g. phosphorylation, sulfation, lipidation, glycosylation, etc.), non-ribosomal peptides, and small molecules (metabolomics). Other protein identification tools like immunoprecipitation and Edman degradation are also available options, but these methods can be low-throughput and require non-trivial amounts of purified protein (which can be difficult to obtain in some settings).

The method

The following is a high-level overview of our approach, also visually summarized in Figure 2*. You can view a detailed, step-by-step protocol on protocols.io.*

Sample collection

Our efforts began with the excision of salivary glands from unfed female *Amblyomma americanum* ticks **[2]**. While our interest lies in ticks, this method should work with tissue from any organism.

RNA extraction and quality control

We pooled about 10 ticks worth of salivary gland tissue and obtained total RNA using a standard extraction kit.

We collected electropherograms to calculate RNA integrity number (RIN), which is a ratio of the 28S:18S ribosomal RNA (rRNA) subunit peak areas and a proxy for RNA quality.

A note on electropherograms from arthropod RNA:

We were surprised to find only one peak corresponding to the 18S subunit where we would normally see two peaks: one corresponding to the 18S subunit and one to the 28S subunit.

Some quick literature searches suggested that this is a commonly observed phenomenon with arthropod RNA. It's thought that arthropods' 28S subunit can fragment (due to structural instability) during sample preparation, yielding two peaks that overlap with the 18S subunit's peak **[3][4]**.

We took a chance and proceeded with transcriptomic library preparation without a RIN readout for RNA quality. To ensure that future extraction are adequate before library preparation, we'd like to identify fast and easy alternative assays for RNA quality. Suggestions are highly appreciated.

mRNA enrichment

Next, we needed to enrich mRNA from the total RNA mixture, as rRNA tends to dominate. We used positive enrichment of mRNA via oligo-(dT) primers, which target mRNA containing poly-A tails.

RNA sequencing

We submitted our samples to the UC Berkeley QB3 genomics core for size-selection (> 3 kb), PacBio's library preparation, Sequel II HiFi sequencing, and Iso-seq analysis.

Tandem mass spectrometry-based proteomics

In parallel to the RNA processing and sequencing steps, we prepared tryptic peptides from *A. americanum* salivary gland lysate and analyzed them by data-dependent LC-MS/MS using a high-resolution strategy on an Orbitrap mass spectrometer.

Transcriptomic and proteomic data analysis

Our overall computational pipeline is summarized in Figure 3. We identified coding sequences in our transcriptome data using TransDecoder **[5]**, CPAT **[6]**, and ANGEL **[7]**. We combined our resultant output and submitted all sequences for BUSCO analysis **[8]**. Next, we collapsed sequences down by CD-HIT clustering with a similarity setting of 100% (c = 1.0) **[9][10]** to deduplicate, and then we used these CD-HIT-collapsed sequences for subsequent proteomics mapping. For functional analysis, we further clustered these sequences down using CD-HIT with a similarity setting of 95% (c = 0.95) in order to group closely related sequences. Representative sequences for each of these 95% cutoff clusters were submitted for Interproscan analysis **[11]**.

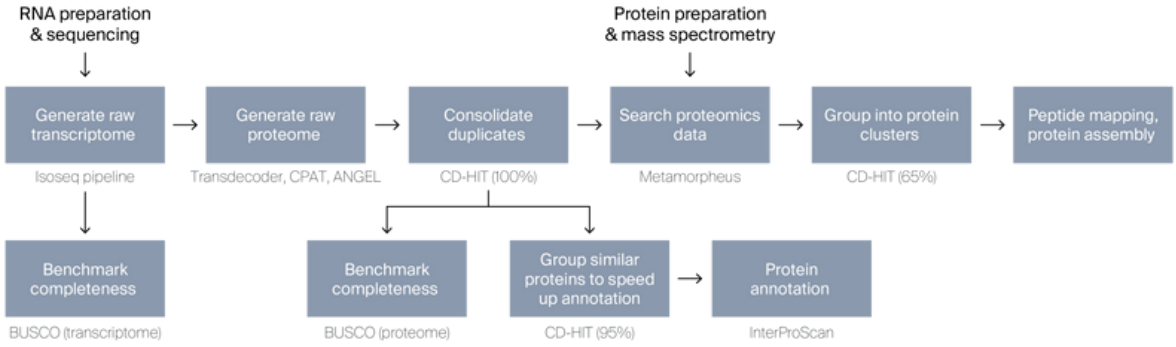

Figure 3

Overview of data analysis workflow and tools.

We assigned fragmentation spectra with a basic proteomic search. We compared the overlap between PSMs and peptides identified using various databases. In order to compare protein-level results, we further clustered sequences using CD-HIT at a 65% similarity cutoff (c = 0.65). Our intent was to group moderately related sequences and gain an orthogonal view of the number of protein clusters identified by each database. Since we developed this method to study lone star ticks and don't have a genome to which we can map transcripts, it can be difficult to group transcripts accurately due to alternative splicing events. One of the easiest operations we can do until we have a genome is to cluster the sequences we do have by a similarity metric.

To see a representative output from this method, check out our tick salivary gland dataset.

What's next?

We developed this method to gain insight into the tick salivary gland proteome, and are now analyzing that dataset.

If you decide to try this or a similar method in your own research, we'd love to hear how it goes. Let us know if you have any questions!

Acknowledgements Thank you to the QB3 Genomics Facility at UC Berkeley (RRID:SCR_022170) for RNA library prep and sequencing.

References

1. Kim TK, Tirloni L, Pinto AFM, Diedrich JK, Moresco JJ, Yates JR, da Silva Vaz I, Mulenga A. (2020). Time-resolved proteomic profile of Amblyomma americanum tick saliva during feeding. https://doi.org/10.1371/journal.pntd.0007758
2. Patton TG, Dietrich G, Brandt K, Dolan MC, Piesman J, Gilmore Jr. RD. (2012). Saliva, Salivary Gland, and Hemolymph Collection from Ixodes Scapularis Ticks. https://doi.org/10.3791/3894
3. McCarthy SD, Dugon MM, Power AM. (2015). 'Degraded' RNA profiles in Arthropoda and beyond. https://doi.org/10.7717/peerj.1436
4. DeLeo DM, Pérez-Moreno JL, Vázquez-Miranda H, Bracken-Grissom HD. (2018). RNA profile diversity across arthropoda: guidelines, methodological artifacts, and expected outcomes. https://doi.org/10.1093/biomethods/bpy012
5. TransDecoder tool for finding coding regions within transcripts (GitHub). https://github.com/TransDecoder/TransDecoder

6 Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. https://doi.org/10.1093/nar/gkt006

7 ANGEL tool for robust open reading frame prediction (GitHub). https://github.com/PacificBiosciences/ANGEL

8 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. https://doi.org/10.1093/bioinformatics/btv351

9 Li W, Godzik A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. https://doi.org/10.1093/bioinformatics/btl158

10 Fu L, Niu B, Zhu Z, Wu S, Li W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. https://doi.org/10.1093/bioinformatics/bts565

11 Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong S-Y, Lopez R, Hunter S. (2014). InterProScan 5: genome-scale protein function classification. https://doi.org/10.1093/bioinformatics/btu031