Robust long-read saliva transcriptome and proteome from the lone star tick, *Amblyomma americanum*

The way you generate a reference database has a real impact on the completeness and results of proteomics experiments.

Version 2, published Feb 22, 2023. Originally published May 30, 2022.

Arcadia Science DOI: 10.57844/arcadia-3hyh-3h83

Purpose

At Arcadia, we're studying diverse organisms and sharing both our discoveries and the tools we develop along the way. In one of our first efforts, we're trying to understand how ticks manipulate their hosts. In this pub, we describe how we established a proteome reference for *A. americanum* ticks by collecting a long-read transcriptome from their salivary glands. We also show you where you can access all the data. We hope it will be useful for other tick researchers or anyone interested in doing omics in the absence of a complete genome.

- 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.
- **Data** from this pub is accessible in the SRA (transcriptome) and in the PRIDE repository (proteome).
- The **method** we used to generate this data is more fully described in this pub.

Background and goals

These experiments are part of our effort to build omics tools to study ticks. Ticks feed on us and other animals for prolonged periods, which suggests that ticks have powerful means for suppressing host surveillance systems. We are identifying the components of the tick molecular toolkit and developing them into new therapies for patients living with otherwise intractable skin conditions.

To begin, we decided to take a peek at all the proteins we could find in the saliva of *Amblyomma americanum* (a.k.a. the lone star tick).

> SHOW ME THE DATA: Access our transcriptomics data and proteomics data.

The approach

To begin unraveling the intricacies of tick saliva, we've chosen to examine the salivary proteome using tandem mass spectrometry-based proteomics as a key technology. For now, we're taking the more straightforward bottom-up approach. Proteomics experiments come in many flavors, but they can be categorized into one of three bins according to the size of the peptide analytes being examined. 1) Top-down proteomics is generally concerned with the analysis of intact proteins and/or their complexes. 2) Bottom-up proteomics is generally concerned with the analysis of peptides generated by chemical or enzymatic digestion of parent proteins. 3) Middle-down proteomics takes an intermediate approach wherein parent proteins are minimally digested, creating peptides larger than those considered for bottom-up work but still smaller than intact proteins. Tools and techniques for bottom-up proteomics have been in development for much longer than the other two styles are thus more reliable and accessible.

We hope that mass spectrometry will be advantageous in this context because it will enable the analysis of cell-free secretions. Importantly, it is suited for the detection of non-encoded molecules/modifications, including protein post-

translational modifications (e.g. phosphorylation, sulfation, lipidation, glycosylation, etc.), non-ribosomal peptides, and small molecules (metabolomics).

In order to interpret proteomic data from tandem mass spectrometry, we need a reference proteome, which can be inferred from genome and/or transcriptome sequencing efforts. Unfortunately, ticks aren't model organisms (yet) and apart from *Ixodes scapularis* (a.k.a. the deer tick), there are few previously deposited data sets for the other ~900 known tick species, including the tick species we're studying, *A. americanum* (Figure 1). One notable exception is the combined short-read transcriptome and matched time-resolved salivary proteome deposited in the PRIDE repository by the Mulenga lab [1]. This rich data set serves as a great scientific resource.

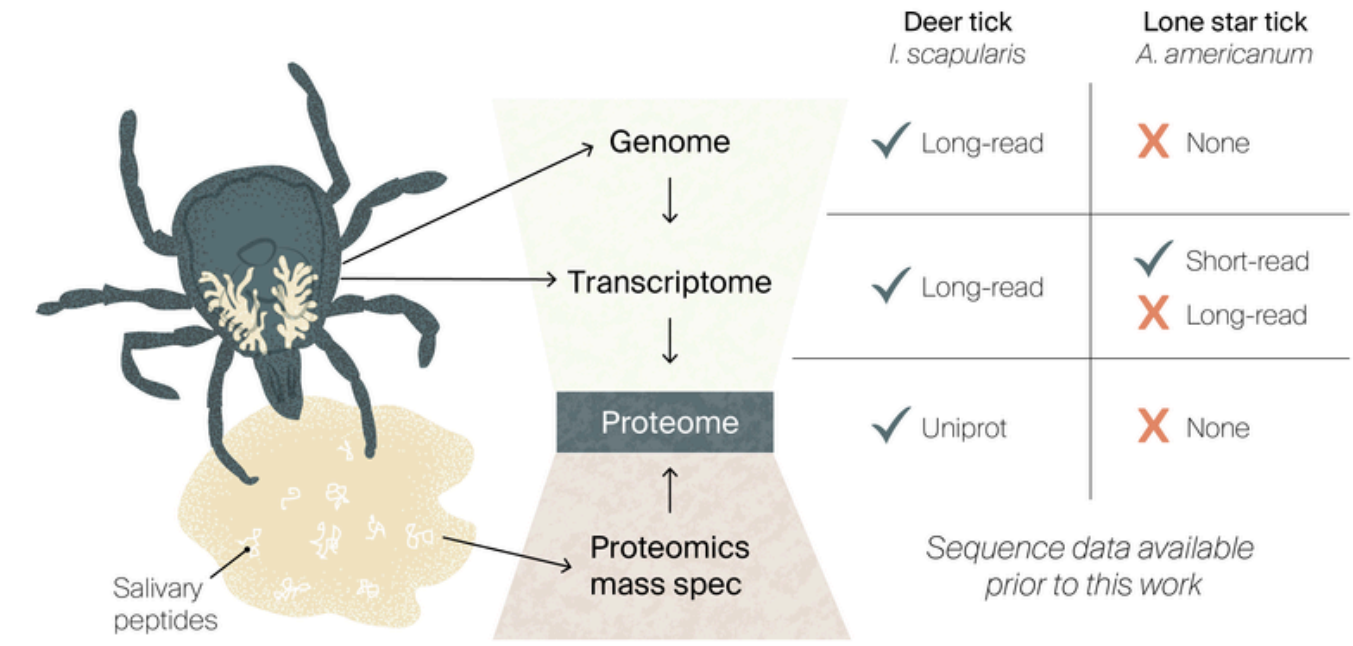

Figure 1. **In contrast to *Ixodes scapularis*, *A. americanum* reference data sets are incompletely represented in public repositories**.

A. americanum genome and transcriptome assembly will enable the creation of a comprehensive proteome database for LC-MS/MS-based proteomics analysis. In this work, we focused on assembling a new transcriptome to inform our proteomic analysis.

Before we performed our mass spectrometry experiments, we decided to develop our own proteome database, adding to the Mulenga lab's work and enriching the reference data available to the tick research community. We considered

sequencing the *A. americanum* genome, but it would require more time, money, and expertise than RNA sequencing. We therefore decided to do long-read RNA sequencing (specifically PacBio's HiFi Iso-seq methodology) because it can provide insights into full transcript structures. We figured it would provide a great complement to the Mulenga lab's short-read data set collected on the same tick species.

Our overall method is summarized in the text below and in Figure 2. For more information on why we took this approach, see our companion method piece. For a detailed, step-by-step protocol, see our protocols.io entry.

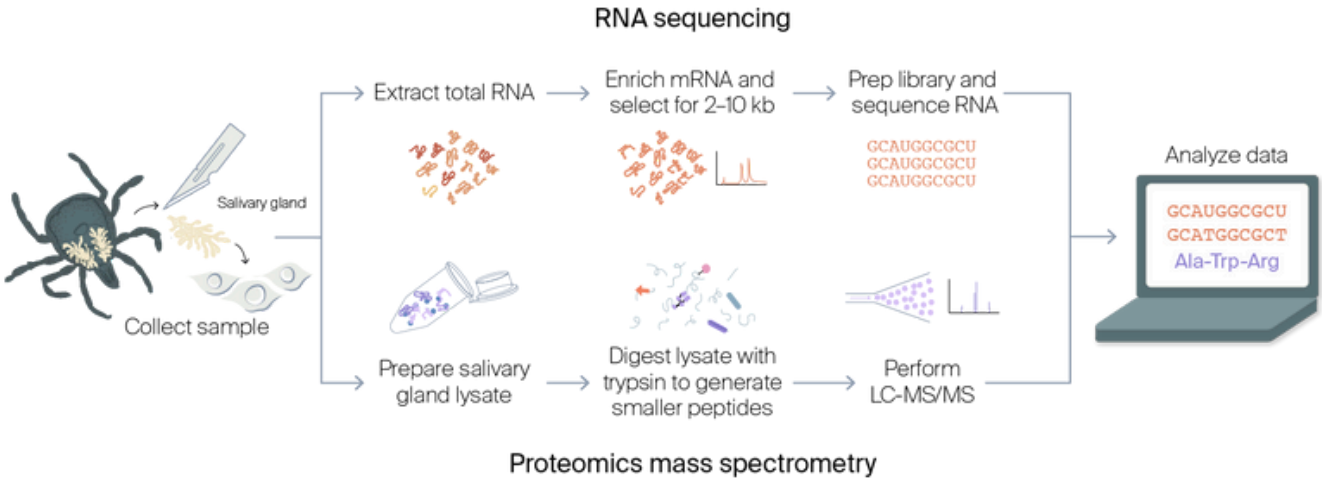

Figure 2. **Overview of the parallel transcriptomic (top) and proteomic (bottom) work streams**.

Sample collection and RNA preparation

We collected our tissue of interest by excising salivary glands [2] (which comprise a major mass fraction of the tick anatomy) from unfed female *A. americanum* ticks.

RNA extraction, processing, and sequencing

We pooled salivary gland tissue from about 10 ticks, homogenized by bead beating, 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. We enriched mRNA via oligo-(dT) primers, which target mRNA containing poly-A tails.

Finally, we submitted our RNA 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.

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.

Mass spectrometry

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

Transcriptome and proteome processing and analysis

Our analysis process is summarized in Figure 3. We identified coding sequences in our transcriptome data using TransDecoder [5], CPAT [6], and ANGEL [7]. We collapsed sequences down by CD-HIT clustering [8] for subsequent proteomics mapping. Clusters were submitted for Interproscan analysis [9] and BUSCO analysis [10] to identify protein families and assess completeness of our

transcriptome data set, respectively. We assigned fragmentation spectra with a basic proteomic search.

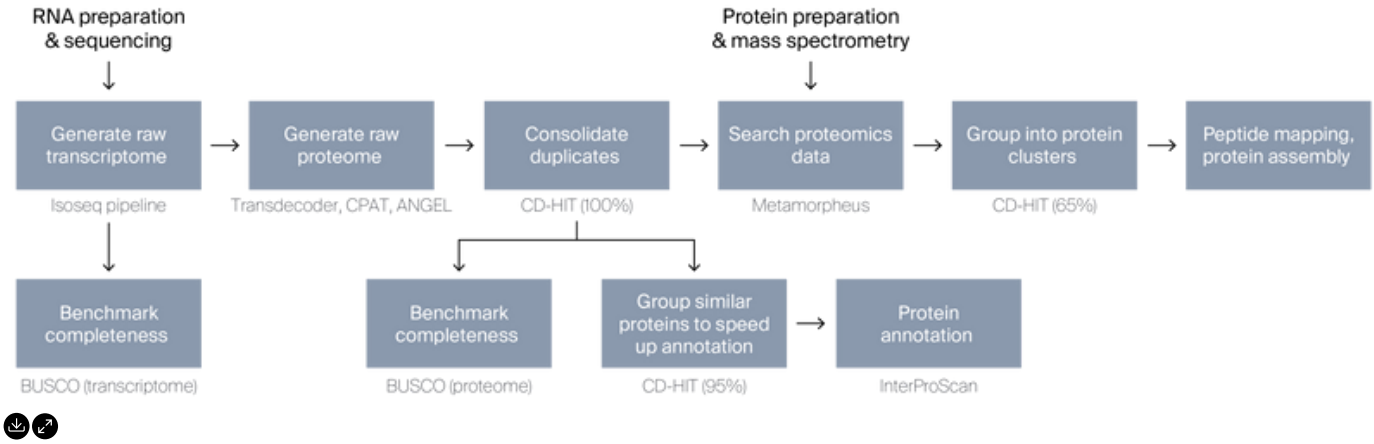

Figure 3. **Overview of data analysis workflow and tools**.

The results

> SHOW ME THE DATA: Access our transcriptomics data and proteomics data.

Transcriptomic data

Once our transcriptome data arrived, we identified protein-coding sequences using TransDecoder, CPAT, and ANGEL. We combined our resultant protein output and collapsed sequences down by CD-HIT clustering with a similarity setting of 100% (c=1.0) to group redundant sequences, yielding 222,632 predicted proteins (down from a total of 307,541). We used these CD-HIT-collapsed non-redundant protein sequences for subsequent proteomics mapping. For functional analysis, we reasoned that proteins with closely related sequences would likely have the same function. Thus in order to reduce compute time, we grouped closely related protein sequences using CD-HIT, except this time with a similarity setting of 95% (c=0.95). This yielded 121,223 protein clusters (Figure 4, A). Each cluster contained one or more members and one representative sequence; for single-member clusters, one sequence is both a member and a representative. Representative

sequences for each of these 95% cut-off clusters were submitted for Interproscan analysis to classify proteins into families and identify domains, resulting in annotation for 68,705 clusters (57%) but no annotation for 52,518 clusters (43%) (Figure 4, B).

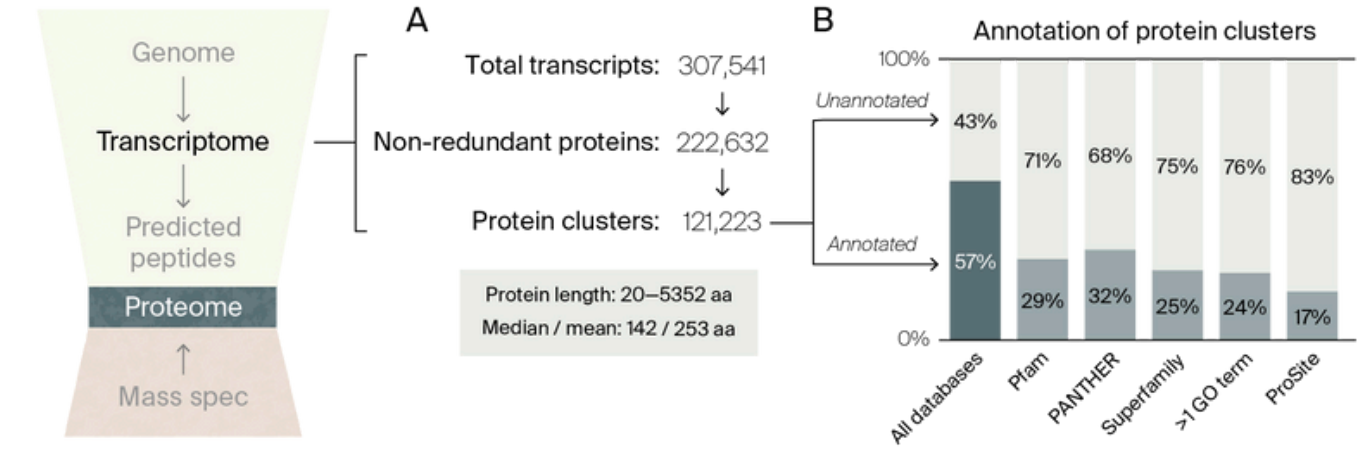

Figure 4. **Processing and annotating protein clusters within our *A. americanum* transcriptome**.
(A) Overview of *A. americanum* long-read transcriptome data. Protein-coding sequences were predicted from poly-A-enriched and 5-kb-size-selected transcripts.
(B) Protein-coding sequences were clustered using CD-HIT and functional annotation by Interproscan reveals a large subset (43%) of unannotated protein clusters.

In addition, BUSCO analysis, which assesses completeness of a transcriptome, revealed a slight gain in completeness compared to the previous short-read transcriptome (Figure 5, A). Finally, we compared our new long-read transcriptome database with the short-read Mulenga database and a database forged from NCBI sequences (Figure 5, B). It's striking how divergent our data set and the Mulenga data set appear to be, but the real test of usefulness for our data set will be determined by proteomics mapping results.

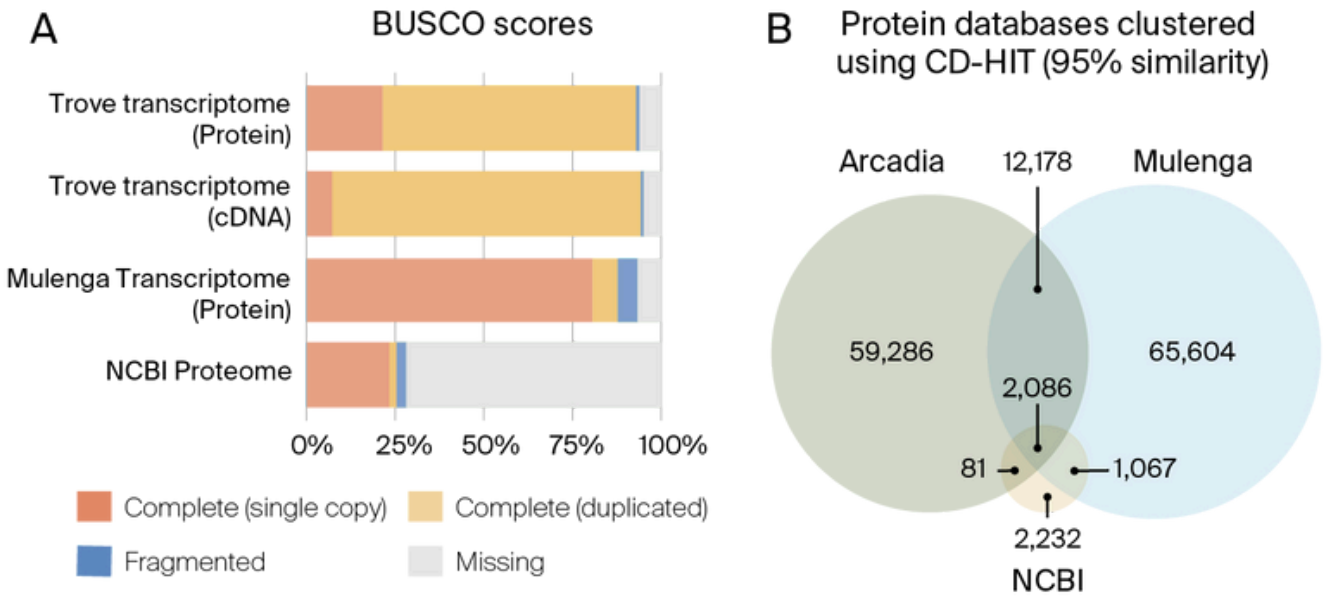

Figure 5. **Comparing the transcriptome generated through this method to previous resources for the same organism, *A. americanum*.**

(A) BUSCO analysis reveals our long-read transcriptome ("Arcadia") is slightly more complete than the short-read transcriptome from the Mulenga lab.

(B) CD-HIT clustering reveals only small overlap between protein cluster membership between Arcadia, Mulenga, and NCBI proteomes.

Proteomic data from mass spectrometry

With a basic proteomics database search, we were able to assign approximately 40% of all collected fragmentation spectra between all databases. 37% were assigned by our new database and 36% by the Mulenga database, with a fairly large overlap. We observe approximately 8% more peptide-spectrum matches (PSMs) and 9% more peptides than are represented in the Mulenga transcriptome and NCBI databases alone (Figure 6).

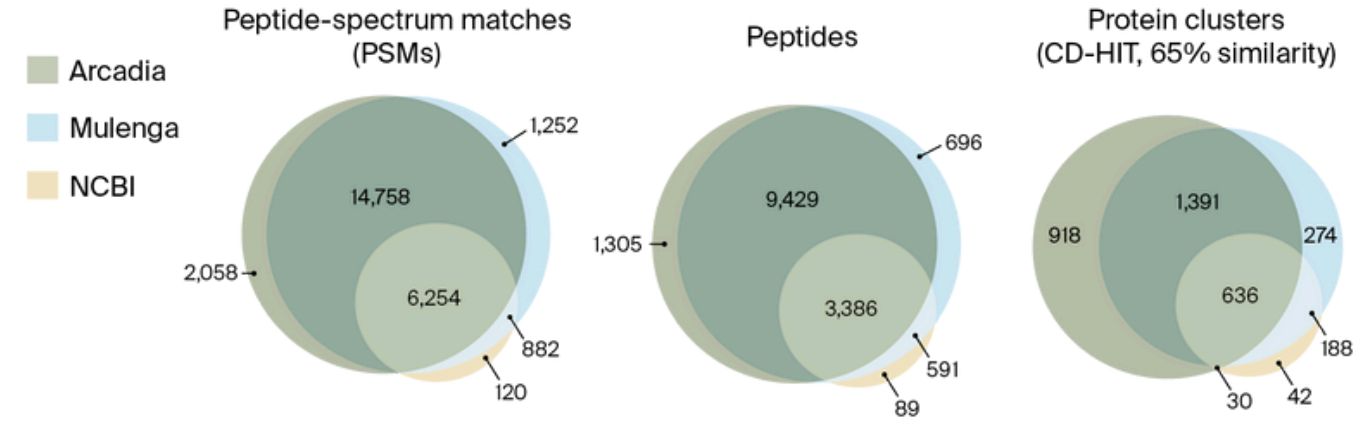

Figure 6. **Comparison of LC-MS/MS-based proteomics mapping results when we used the Arcadia, Mulenga, and NCBI transcriptome-based proteomes as mapping databases**.
Venn diagrams depicting overlap at the peptide-spectrum match (PSM)-, peptide-, and protein cluster-level.

Background on peptide-spectrum matches vs. peptides:

During a proteomics database search, theoretical mass spectra are generated from peptides in our search database. We compare these theoretical mass spectra to experimental mass spectra, and when there is reasonable agreement during a comparison, we assign an experimental mass spectrum with the matching peptide identity. This assignment is called a peptide-spectrum match (PSM). During a tandem mass spectrometry run, many mass spectra are collected and often, several of the same spectra are collected, especially if a peptide is abundantly represented in a mixture. Thus, it is possible for a single distinct peptide to be represented by many spectra.

At the protein cluster level (CD-HIT clustering at 65% similarity cut-off; c=0.65), we observe a 38% increase in cluster detection. Interestingly, when we compare all database protein sequences against all protein sequences detected by proteomics, an unexpected distribution emerges revealing that proteins detected by our database tend to skew toward longer sequences (Figure 7). We hope that

for further studies, having longer protein sequences will enable a more complete understanding of function.

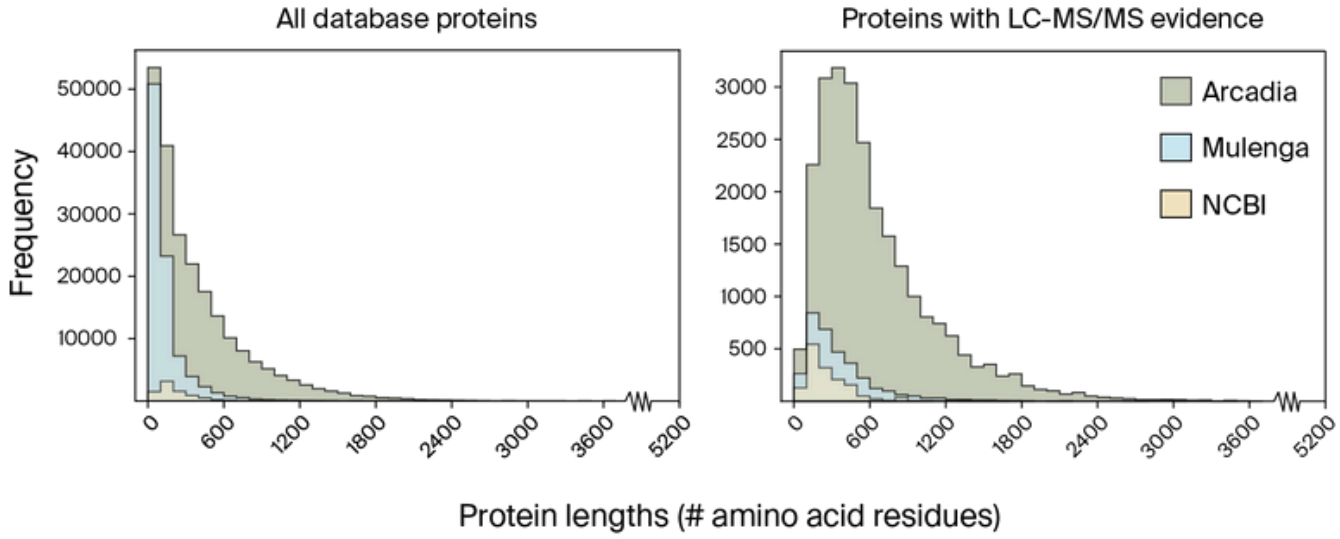

Figure 7. **Histograms of protein sequence length distribution for all proteins (left) and only proteins with LC-MS/MS evidence (right)**.
Note that y-axes are different.

Key takeaways

To sum this up, it looks like our long-read transcriptome-based proteome database compares reasonably well with the Mulenga lab's short-read transcriptome-based proteome database.

While our database enables the detection of approximately 8% more PSMs and 9% more peptides than the previous Mulenga database, it is in no way a replacement, as 5% of all PSMs are only detectable thanks to the Mulenga database. The short-protein skew of the Mulenga database appears to be complementary to the long-protein skew of our own database.

Finally, the assignment of 40% of all fragmentation spectra is reasonable but there are likely many more assignable spectra awaiting deconvolution. >80% assignment is highly unlikely based on many factors (and personal experience), but leaping to a value between 40% and 80% may be achievable.

What's next?

Building a more complete protein database will allow us to assign a greater percentage of fragmentation spectra. For this, we'd need a fully assembled *A. americanum* genome, which, to the best of our knowledge, is not yet available in a public repository. As such, assembling an *A. americanum* genome will probably be the next item on our checklist. We're also still analyzing this data and specifically exploring post-translational modifications. Ultimately, we hope to identify active salivary molecules.

Acknowledgements

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

Contributors (A–Z)

- **Seemay Chou**: Conceptualization, Editing, Supervision
- **Tori Doran**: Resources
- **Behnom Farboud**: Critical Feedback
- **Juliana Gil**: Critical Feedback
- **William Hatleberg**: Visualization
- **Megan L. Hochstrasser**: Editing, Visualization, Writing
- **Greg Huber**: Conceptualization, Critical Feedback
- **Kira E. Poskanzer**: Conceptualization, Supervision
- **MaryClare Rollins**: Project Administration, Resources
- **Peter S. Thuy-Boun**: Editing, Formal Analysis, Investigation, Methodology, Writing
- **Elizabeth Tseng**: Conceptualization, Critical Feedback
- **Joan Wong**: Critical Feedback

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. 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. https://github.com/pacificbiosciences/angel
8. 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
9. 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
10. 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