2

May 24, 2022

A parallel transcriptomics and proteomics workflow for organisms with minimal reference protein databases V.2

Peter Thuy-Boun[1]

[1]Arcadia Science

1

protocol .

Peter Thuy-Boun

This is a method that lets researchers perform mass spectrometry-based proteomics experiments and build a new reference protein database to help interpret that data in parallel work streams. In our case, we used it to study the lone star tick, *Amblyomma americanum*, but the approach is applicable to any understudied organism.

Peter Thuy-Boun 2022. A parallel transcriptomics and proteomics workflow for organisms with minimal reference protein databases. **protocols.io** https://protocols.io/view/a-parallel-transcriptomics-and-proteomics-workflow-b9dar22e
Peter Thuy-Boun

protocol ,

May 14, 2022

May 24, 2022

62594

Proteomics Sample Preparation 6h 5m

1 Dissect salivary gland tissue from 10 adult female *Amblyomma americanum* ticks into approximately 100 µL of sterile water and store at -20°C. Note that tissue/cells from other organisms can be substituted for *Amblyomma americanum* salivary gland tissue. These other tissue/cell types may require additional processing steps. **01:00:00** 1h

2 Thaw tissue on ice, dilute to 1 mL with water, then transfer into 2 mL bead beating tubes containing 1 mm zirconium beads. Homogenize salivary glands using a bead bug bead beater 15m

(Benchmark Scientific) operating at 300 rpm for 3 pulses lasting 30 seconds each with 30 second ice-bath submersions between each pulse. Upon completion, chill tubes on ice and allow beads to settle. **00:15:00**

3 Transfer 600 µL of lysate to a clean microcentrifuge tube. Centrifuge the tube at 12,000 g for 5 minutes and separate supernatant from pellet. 5m **00:05:00**

4 Fractionate supernatant from the previous step using a 3 KDa MWCO centrifugal concentrator unit (PES, 0.5 mL capacity) and water over 3 centrigation cycles at 10,000 g for 30 minute intervals. 2h **02:00:00**

4.1 The first round of centrifugation should yield a 100 µL retentate and 400 µL filtrate.

4.2 During the second round of centrifugation, add 100 µL of water to the 100 µL retentate (200 µL total). This should yield 50 µL of retentate and 150 µL of filtrate after centrifugation.

4.3 During the third round of centrifugation, add 50 µL of water to the 50 µL retentate fraction (100 µL total). This should yield 50 µL of filtrate and 50 µL of retentate after centrifugation.

4.4 Pool the filtrate fractions in a clean microcentrifuge tube. Transfer retentate into a clean microcentrifuge tube. Resuspend residue from the concentrator unit in 50 µL of water and combine with the retentate fraction.

5 Measure protein concentration of the retentate fraction (dissolved in 100 µL of water). This can be done using a Pierce/Thermo BCA assay kit. 45m **00:45:00**

6 Based on BCA results, transfer 10 µg of retentate into a clean microcentrifuge tube. Dry sample using a vacuum centrifuge (25ºC) over 30 minutes. This should yield a white residue. 35m **00:35:00**

7 Denature, reduce, and alkylate proteome to unfold proteins and cap cysteines. This enables subsequent proteolysis to proceed more effectively. 1h **01:00:00**

7.1 Resuspend residue in 10 µL of denaturation buffer (5M Urea + 100 mM bicine, pH 8; sterile filtered), then treat with 1 µL of TCEP·HCl solution (14 mg/mL in water). Mix by pipetting, then incubate at ambient temperature.

urea: Fisher [57-13-6]
bicine: Fisher [150-25-4]
TCEP·HCl: ThermoFisher [51805-45-9]

7.2 After 15 minutes, add 0.7 µL of 2-chloroacetamide solution (47 mg/mL in water). Mix tube contents by pipetting, then incubate at ambient temperature for 45 minutes.

2-chloroacetamide: TCI [79-07-2]

7.3 Expand contents of the tube with the addition of 40 µL of modified trypsin buffer (50 mM + 1 mM calcium chloride, pH 7.6; sterile filtered) and mix thoroughly by pipette.

triethylammonium bicarbonate: Aldrich [15715-58-9]
calcium chloride (anhydrous): Aldrich [10043-52-4]

8 Reconstitute a 20 µg vial of lyophilized sequencing grade trypsin with 100 µL of modified trypsin buffer. Then add 1 µL of reconstituted trypsin solution to the tube from the previous step. Mix contents thoroughly, then incubate at 37°C overnight. **Overnight**

sequencing grade trypsin: Promega (V5111)

9 After overnight incubation, add 2.5 µL of LC-MS grade formic acid. Mix well then centrifuge at 12,000 g for 5 minutes to pellet insolubles. Transfer supernatant into clean microcentrifuge tubes in 2x25 µL portions (containing 5 µg peptides in each tube). 10m **00:10:00**

formic acid: ThermoFisher [64-18-6], LC-MS grade

10 Desalt peptides from previous step into clean microcentrifuge tubes using C18 ZipTips (one tip per 5 µg peptides) following manufacturer instructions. Perform standard elution using a 1:1 water/acetonitrile mixture. 15m **00:15:00**

C18 ZipTips: Millipore, 10 µL format, 0.6 µL beds
acetonitrile: Aldrich [75-05-8], LC-MS grade

11 Dry desalted peptides in microcentrifuge tubes by vacuum centrifugation over 30 minutes (45ºC) then store at -20°C until shipment to the mass spectrometry core for analysis. Dried peptides can be sent by overnight delivery at ambient temperature. 30m **00:30:00**

Mass Spectrometry Data Collection 3h 35m

12 This section was performed by the University of Florida Scripps Biomedical Research Proteomics Core. Reconstitute dried desalted peptides in 50 µL of LC-MS grade water + 0.1% formic acid. **00:05:00** 5m

13 Adjust these parameters as necessary for your experimental setup and desired outcome. The following describes our specific approach. 2 µg of peptides dissolved in 10 µL of water were injected onto an EASY PepMap RSLC C18 column (2 µm, 100 Å, 75 µm ID x 50 cm, Thermo Scientific). Separation was performed using a binary solvent gradient (A: water + 0.1% formic acid; B: 80:20 acetonitrile/water + 0.1% formic acid) at a flowrate of 250 nL/min delivered using a nEasy-LC1000 nano liquid chromatography system (Thermo Scientific). The following gradient was used: 5-25% B over 90 min, followed by 25-44% B over 30 min, 44-80% B over 0.1 min, a 10 min hold at 80% B, a return to 5% B over 3 min, and finally a 3 min hold at 5% B. An extended cleaning gradient was then applied: 98% B over 3 min, hold at 98% B for 10 min, return to 5% B over 3 min, hold at 5% B for 3 min, increase to 98% B over 3 min, hold at 98% B for 10 min, return to 5% B over 3 min, hold at 5% B for 3 min, increase to 98% B over 3 min, hold at 98% B for 10 min. Ions were generated at 2.4 kV using an EASY Spray ion source (Thermo Scientific) held at 50ºC. Data was acquired using a Thermo Orbitrap Fusion Tribrid mass spectrometer. Data dependent scanning was performed by the Xcalibur software package (v 4.0.27.10) using survey scans at 120,000 resolution in the Orbitrap mass analyzer scanning between m/z 380-2000 with an automatic gain control (AGC) target of 1e5 and maximum injection time of 50 ms. This was followed by higher-energy collisional dissociation (HCD) fragmentation at normalized collision energy (NCE) 30% for the topN peaks with an AGC setting of 1e4. Precursors were selected by monoisotopic precursor selection (MIPS) setting to peptide and MS/MS was performed in the Orbirap on ions with charges +2 to +8 at 30,000 resolution. Dynamic exclusion was set to exclude ions after 2 times within a 30 sec window for 20 sec. **03:30:00** 3h 30m

14 In our case, data were obtained in Thermo .raw format. Data were further processed in the Metamorpheus software package (see Proteomics Data Analysis section).

RNA Extraction, Library Preparation, and Transcriptome Sequencing

15 Dissect salivary gland tissue from 10 adult female *Amblyomma americanum* ticks into approximately 300 µL of ultrapure RNAse-free water. Extract tissue ≤1 hour prior to RNA extraction. Note that tissue/cells from other organisms can be substituted for *Amblyomma americanum* salivary gland tissue. These other tissue/cell types may require additional processing steps. **01:00:00** 1h

16 Perform RNA extraction using the RNeasy mini kit (other kits will work, this is just the kit we chose). Add 10 µL of beta-mercaptoethanol to 1 mL of buffer RLT (included in kit). Add 600 µL 5m

of buffer RLT (+beta-mercaptoethanol) to the 300 µL suspension of tick salivary gland tissue held on ice. **00:05:00**

RNeasy mini kit: Qiagen (2019)
beta-mercaptoethanol: Aldrich [60-24-2]

17 Transfer tissue suspension into a 2 mL bead beating tube containing 1 mm zirconium beads. 5m
Seal tube, vortex briefly, then homogenize using a bead bug bead beater (Benchmark Scientific) operating at 300 rpm for 30 seconds. Immediately chill homogenized tick tissue on ice. **00:05:00**

18 Perform the next steps in RNA extraction and purification as outlined in the Qiagen RNeasy 2h
mini handbook, "purification of total RNA from animal tissues" section (October 2019 edition, pages 45-52). **02:00:00**

19 Elute washed and bound RNA with 50 µL of ultrapure RNase-free water. Measure crude quality 5m
and yield using a nanodrop spectrophotometer (ThermoFisher). **00:05:00**

20 All following steps in this section were performed by the Vincent J. Coates Genomics 1d
Sequencing Laboratory (UC Berkeley QB3) guided by recommendations optimized for the Iso-seq workflow (Pacific Biosciences; Iso-seq express library preparation using SMRTbell express template prep kit 2.0). Adjust as necessary as new workflows become available. 8 µg of total RNA were submitted to the sequencing core for femtopulse analysis (Agilent) yielding an electropherogram with a single apparent 18S rRNA peak with no substantial signal in the 28S rRNA region. A RNA integrity score (RIN) was not calculable. Others have noted that the 28S rRNA subunits tend to be unstable collapsing down to apparent 18S-sized fragments during the electrophoresis sample preparation and analysis steps. **24:00:00**

21 Separate mRNA from rRNA using polyA-enrichment strategy. Use PolyA-enriched mRNA for cDNA synthesis and amplification. Size-select cDNA targeting sequences >3 kb. In our case this procedure yielded a majority of transcripts in the 2-10 kb range centered around 5 kb. Use purified cDNA sequences to construct a SMRTbell library (Pacific Biosciences).

22 Load SMRTbell library onto one SMRT Cell 8M and perform sequencing on a Sequel II (Pacific 4w
Biosciences) system. Adjust these steps as new sequencing instruments and modalities supplant current ones. **672:00:00**

23 In our case data were obtained as a .bam file which was transferred to the Computation Genomics Resource Laboratory (UC Berkeley QB3) for further processing (see Transcriptome

Data Analysis section).

Transcriptome Data Analysis 5d 4h 30m

24 In our case, raw sequencing data from the Sequel II system were processed at the Computational Genomics Resource Laboratory using the Iso-Seq data analysis workflow (excluding optional genome alignment step as no public genome assemblies yet exist for *Amblyomma americanum*). Resultant polished full length non-concatemer circular consensus reads outputted as *.flnc.polished.hq.hq.fastq.gz and *.flnc.polished.hq.hq.fasta.gz files were uncompressed and used for the next steps including coding sequence prediction. **04:00:00** 4h

25 Coding sequence prediction using TransDecoder: Operate TransDecoder (v.5.5.0) with the following linux command line input: "TransDecoder.LongOrfs -G universal -m X -t *.fasta", where *.fasta is the uncompressed Iso-seq output file referenced above and X is the minimum protein length recognized. In this work X was set to 25, 50, and 100 during three independent runs. Pool output files in *.pep format and deduplicated using CD-HIT (v4.8.1) with a similarity cut-off of 100 (c=1.0) using the following command line input: "cd-hit -i input -o output -c 1 -g 1 -d 0". Input is the input protein fasta file and output is the prefix for resultant summarized .fasta and .clstr files. Note that the -g 1 and -d 0 flags will continue to be used whenever CD-HIT is referenced unless otherwise stated. **06:00:00** 6h

26 Coding sequence prediction using the Coding Potential Assessment Tool (CPAT): Operate CPAT (v.3.0.3) with the following linux command line input: "cpat.py --antisense --hex=fly_Hexamer.tsv --logitModel=fly_logitModel.RData --top-orf=5 --min-orf=75 --gene=*.fasta --outfile=output" where *.fasta is the input fasta file referenced above and output is the chosen output filename. Translate output nucleotide sequences were to protein sequences using one of many utilities including the EMBOSS transeq utility (e.g. "transeq -sequence *.fa -outseq outputproteinfile" where *.fa is the input nucleotide fasta file and outputproteinfile is the output file name). Pool output files containing protein sequences in fasta format and deduplicated using CD-HIT with a similarity cut-off of 100 (c=1.0). **06:00:00** 6h

27 Coding sequence prediction using ANGEL (v3.0; re-implementation of ANGLE protein coding region prediction tool): note that this tool is depricated and that resources and guides for its use are available at the Pacific Biosciences github page. ANGEL-based prediction is a three step process and our workflow deviated from that outlined on github during the "dumb ORF prediction" step where "--min_aa_length" was set to 100. The output *.ANGEL.pep files contain predicted protein sequences which need to be stripped of end-of-sequence '*' characters before CD-HIT deduplication at a similarity cut-off of 100 (c=1.0). **20:00:00** 20h

28 Concatenate output from TransDecoder, CPAT, and ANGEL then deduplicate using CD-HIT 30m

with a similarity threshold set to 100 (c=1.0). This deduplicated assembly of protein sequences in .fasta format will be used for proteomics database searching. **00:30:00**

29 3d
Cluster the deduplicated assembly of protein sequences at a 95% similarity cut-off using CD-HIT (c=0.95). Assemble representative sequences for each 95% output cluster into a new protein fasta file. Submit this file for Interproscan (v5.51-85.0) analysis using the following command line input "interproscan -dp -cpu X -goterms -i input.fasta -o outputfile" where input.fasta is the 95% cut-off clustered protein fasta file and outputfile is the desired prefix for the output .tsv file. In our case, output .tsv files were parsed for annotations using custom python scripts. Protein sequences without annotation did not contain entries in the output .tsv files. **72:00:00**

30 16h
To assess transcriptome and proteome completeness, run BUSCO (v5.2.2) on desired datasets. In our case, the following four *Amblyomma americanum* datasets were used: (1) Mulenga transcriptome-based proteome, (2) NCBI proteome- downloaded early 2021, (3) our Iso-seq transcriptome fasta generated in step 1, and (4) our CD-HIT deduplicated (c=1.0) TransDecoder/CPAT/ANGEL-based proteome. Apply the following command line input "busco -i inputfile -l arachnida_odb10 -o output -m mode" where inputfile is the fasta file containing the collection of sequences under examination, output is the desired output directory, and mode represents the sequence type under examination (genome/transcriptome/proteome). In our case, outputs for all four datasets under examination were combined in a bar plot.
16:00:00

Proteomics Data Analysis 3h 20m

31 5m
For our analysis, Metamorpheus (v.0.0.320) was used for proteomics database searching, but there are a variety of other proteomics search tools available as well. Load thermo .raw file obtained above into the search software. For our search database, our Iso-seq-based proteome (assembled using coding sequence prediction tools TransDecoder, CPAT, and ANGEL; here on out referred to as the Trove proteome) was concatenated with the Mulenga short-read transcriptome-based proteome (retrieved from the PRIDE repository) as well as the collection of *Amblyomma americanum* protein sequences downloaded from the NCBI in early 2021 into a single .fasta file. Load a .fasta file up as the primary proteomics search database. Load up the additional set of common contaminants included in the Metamorpheus software package and add this to the search space. **00:05:00**

32 15m
For our analysis, a basic classic search was performed with the Metamorpheus package employing a target-decoy strategy to control for false discovery. The following settings were applied: 5 ppm precursor mass tolerance, 20 ppm fragment mass tolerance, protease set to trypsin with up to 4 missed cleavages allowed, a maximum of 2 modifications per peptide, minimum peptide length of 7 residues, HCD fragmentation, variable initiator methionine, no quantification, no protein parsimony, static modification:carbamidomethylation of cysteine

and selenocysteine residues, and variable modification: oxidation of methionine. Please customize search parameters as necessary. **00:15:00**

33 Use "AllPSMs.psmtsv" output file to map peptide spectrum matches (PSMs), peptides, and protein groups. In our case, we generated figures from this output using custom python scripts. Only PSMs with a QValue < 0.01 were considered for further analysis. We assembled protein groups by applying CD-HIT with a similarity cut-off of 65% (c=0.65) to concatenated Trove, Mulenga, and NCBI proteomes used for proteomics database searching. 3h **03:00:00**