De novo assembly of a long-read Amblyomma americanum tick genome

We generated a whole-genome assembly for the lone star tick to serve as a reference for downstream efforts where whole-genome maps are required. We created our assembly using pooled DNA from salivary glands of 50 adult female ticks that we sequenced using PacBio HiFi reads.

Published Mar 29, 2023

Arcadia Science

DOI: 10.57844/arcadia-9b6j-q683

Purpose

This is our first public draft assembly of an *Amblyomma americanum* genome. We're continuing to refine this assembly and are only just beginning to explore it. We've already started to find intriguing things (like sequences from the endosymbiont *Coxiella*!) but before we get too deep into analysis, we're sharing our initial work in the hope that it will serve as a valuable reference data set for the tick research community.

- 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.
- **Raw genomic data** is available via NCBI.
- Our **pseudo-haploid genome** is on Zenodo and pending approval at GenBank.
- Our **full assembly** is on Zenodo.
- Our **code** is available in this GitHub repository.

Background and goals

We want to understand how ticks manipulate humans, so we've been studying the lone star tick, *Amblyomma americanum*. Given that there are over 900 species of ticks identified so far, it's easy to overlook *A. americanum* in particular. However, to many in

the eastern and southern United States, *A. americanum* are pests with a rapidly growing presence and predilection for humans as hosts [1][2].

One of our long-time pain points with respect to the study of *A. americanum* has been operating without a reference genome. Despite their relevance to humans, we suspect the deficit in publicly available genomic references for *A. americanum* might stem from their surprising genetic complexity. Based on flow cytometry estimates, the tick's haploid genome is expected to be approximately 3 gigabases (Gb) [3]. For comparison, *Escherichia coli* (bacterium) have 5 megabase (Mb) genomes, *Drosophila melanogaster* (fruit fly) have 140 Mb genomes, *Mus musculus* (mouse) have 2.7 Gb genomes, and *Homo sapiens* (human) have 2.9 Gb genomes [4][5]. Based on these figures, one might not expect that tiny ticks could hold so much genetic information, but believe it or not, the tree of life is peppered with examples more extreme than this. Our ability to accurately and comprehensively map these edge cases with reasonably ordinary resources is a recent phenomenon.

Previously, we and others generated transcriptome assemblies using tissue extracted from *A. americanum* ticks [6][7]. These data sets provide snapshots of the genes being transcribed in the cells we collected. This information was instrumental for generating the protein databases we needed in order to do mass spectrometry-based proteomics on the same tick species. While these transcriptome data sets have proven useful, they can fall short when mass spectrometry detects peptide features that don't correspond to reference transcripts. In these cases, the mass spectrometry features will likely go unidentified in analyses until more transcriptomic data becomes available.

Rather than iteratively sequence tick transcriptomes under various conditions (to capture broader transcript landscapes), we decided that a whole-genome assembly could give us more bang for our buck by capturing all genes encoded by *A. americanum*.

> **SHOW ME THE DATA**: Access our raw genomic data on NCBI, our assembled pseudo-haploid genome on Zenodo, and our full assembly on Zenodo.

The approach

We felt that the time might be perfect to try to generate an *A. americanum* genome because long-read DNA sequencing technologies have become more accessible in recent years, meeting the challenge of mapping large and complex genomes. These long-reads should, in theory, make the assembly of large genomes less computationally expensive (compared to short-reads) while reducing assembly errors. Many groups have rallied around long-reads as a key technology for tick genome assembly in particular, with a notably high-quality *Ixodes scapularis* (deer tick) example unveiled very recently [8][9]. To the best of our knowledge, no such assembly exists for *A. americanum*.

Using long-read HiFi DNA sequencing from Pacific Biosciences (PacBio), we assembled an unphased diploid whole-genome map for female *A. americanum* ticks using the pooled DNA of approximately 50 individuals. We hope that our efforts will provide the research community with a useful resource for advancing work in this important tick species.

Detailed methods

Tick salivary gland extraction

We dissected 50 female ticks for DNA extraction. In an effort to reduce potential contamination from microbes that might inhabit the tick gut, we chose to isolate salivary glands, which incidentally comprise a major mass fraction of the internal tick organ system. We pooled these salivary glands in distilled water, chilled on ice, and extracted DNA immediately after dissection.

DNA extraction

We attempted high-molecular-weight DNA extraction using several different commercially available kits and found that in our hands, the Circulomics high molecular weight DNA tissue kit was most consistent for isolating well-distributed, ≥30 kilobase (kb) genomic DNA fragments (as judged by femto-pulse analysis).

Sequencing

We submitted raw genomic DNA to UC Berkeley's QB3 genomics core for fragment analysis, shearing, and 12–17 kb size selection. Subsequently, we prepared HiFi libraries using PacBio's SMRTbell prep kit 3.0. We sequenced these libraries using two

SMRT Cells (8M) and a Sequel II instrument. UC Berkeley's Vincent J. Coates genomics sequencing lab processed raw sequencing data into circular consensus (CCS) HiFi reads and sent us data in HiFi FASTQ format.

Data analysis

Our long-read assembly and assessment workflow is summarized in Figure 1. We tried assembling the genome with Shasta, Flye, and Hifiasm with default settings on a 10-core 3.7 GHz Xeon workstation containing 224 GB of RAM, and ultimately moved forward with Flye.

The following code blocks depict the command line scripts we used for the assembly and assessment processes:

Concatenate FASTQ files (optional):

```
cat *.fastq > concatenated_file_name.fastq
```

Run Flye assembler (v2.8.3; default settings):

```
flye –pacbio-hifi concatenated_file_name.fastq –g 3g –o output_directory –t 19 –min-ovlp 5000
```

BUSCO assessment (v5.4.4):

```
busco –i assembly.fasta –l arachnida_odb10 –o output_directory –m genome
```

Purge_dups (v1.2.6):

Instructions for execution are available here.

> All other **code**, including the **Jupyter notebook** we used for clean-up, is available at this GitHub repository (DOI: 10.5281/zenodo.7787240).

Data deposition

We deposited raw HiFi reads (FASTQ files) into NCBI (bioproject PRJNA932813). We deposited our pseudo-haploid genome assembly (FASTA file) into NCBI/GenBank but we are still awaiting approval, so we uploaded it to Zenodo (DOI: 10.5281/zenodo.7783368) for now. and We also uploaded our full, unphased assembly (FASTA file) to Zenodo (DOI: 10.5281/zenodo.7747102).

The data

SHOW ME THE DATA: Access our raw genomic data on NCBI, our assembled pseudo-haploid genome on Zenodo, and our full assembly on Zenodo.

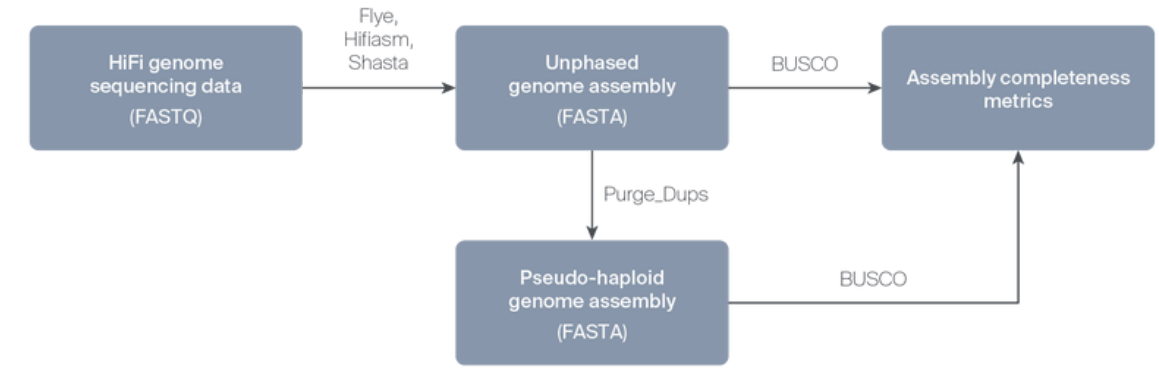

Figure 1. **Bioinformatics workflow**.

We received a sufficient amount of data from each SMRT Cell 8M. Our combined yield totaled 3.3 million HiFi CCS reads, composed of approximately 44.5 billion HiFi CCS bases, for an average HiFi insert length of 13.6 kb. We expected this amount of data to provide approximately 15-fold coverage of the *A. americanum* genome. We subjected the data to a simple long-read assembly and assessment workflow (Figure 1) starting with some cursory test assemblies using Shasta, Flye, and Hifiasm with default settings on a 10-core 3.7 GHz Xeon workstation containing 224 GB of RAM [10][11][12]. We found

that Flye and Hifiasm provided the most BUSCO complete assemblies using data from just one SMRT Cell 8M (Figure 2).

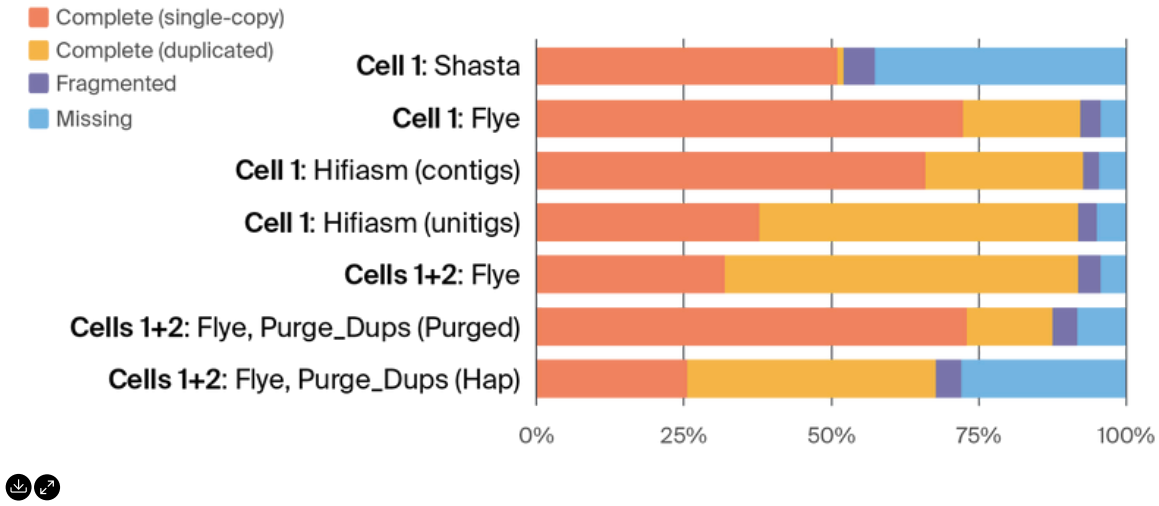

Figure 2. **BUSCO results for each assembly**.

Note that we deposited the entry labeled “Cell 1+2:Flye, Purge_Dups (Hap)” in Zenodo and deposited the entry labeled “Cell 1+2:Flye, Purge_Dups (Purged)” in Zenodo and at NCBI/GenBank, which awaits release.

In our experience, Flye was the fastest assembler that produced reasonably (>75%) complete assemblies. However, Hifiasm produced several assemblies and the largest (unitig) assembly contained the most duplicated BUSCO genes. Flye consumed approximately 1–2 days of processing time for one SMRT Cell worth of data and 3–4 days of processing time using the combined data from both SMRT Cells. Hifiasm consumed 1–2 days for one SMRT Cell and did not complete processing after three weeks for two SMRT Cells. We suspect that Hifiasm might have had trouble with our data set because the genomic DNA we sequenced came from 50 ticks rather than one individual, which would have been the ideal scenario.

For our initial draft assembly, we chose to move forward with Flye due to its speed, convenience, and simplicity of output. We are still examining the outputs and merits of each assembler and would like to note that the unitig assembly that Hifiasm produced is a bit larger and potentially more information-rich than the contig Hifiasm assembly and the default Flye assembly. This could have implications for transcriptome mapping and protein database assembly for proteomics.

During genome deposition at NCBI/GenBank, the raw assembly that Flye produced triggered a few automated error messages, indicating that our assembly needed some light clean-up. Specifically, we had several contigs of less than 200 nucleotides and several duplicate contigs that we needed to remove. We also had a contig containing an adaptor sequence requiring adaptor excision. We generated a Python-based Jupyter notebook to take care of these issues.

> All **code**, including the **Jupyter notebook** we used for clean-up, is available at this GitHub repository (DOI: 10.5281/zenodo.7787240).

The final issue, which a simple Python script could not resolve, was the fact that our assembly was too large compared to NCBI/GenBank estimates. To solve this issue, we used Purge_Dups to split our unphased assembly [13]. This generated a pseudo-haploid assembly which we then cleaned up using our aforementioned Python-based Jupyter notebook. We deposited the resultant assembly at NCBI/GenBank and are awaiting final approval, so we uploaded it to Zenodo to make it available now. We also deposited our unphased diploid genome into Zenodo for anyone interested in accessing our full data set.

Key takeaways

We've assembled an 88% BUSCO-complete long-read pseudo-haploid *Amblyomma americanum* genome of approximately 3 gigabases from 50 individual female ticks. It is available for download and use at Zenodo. The unphased diploid genome is also available in a Zenodo repository.

Next steps

We will continue to make refinements to the *Amblyomma americanum* assembly. Of note, we've preliminarily detected *Coxiella*-like endosymbiont sequences in our assembly along with some potential contaminants that we plan to extract. With a more refined assembly in hand, we are planning to perform gene-finding operations, *in silico*

functional annotations, and finally, we'll construct a more complete protein database for proteomics.

Acknowledgements

Thank you to the QB3 Genomics Facility (RRID:SCR_022170) at UC Berkeley for raw DNA quality control, library preparation, and sequencing.

Contributors (A–Z)

- **Seemay Chou**: Conceptualization, Supervision
- **Tori Doran**: Resources
- **Behnom Farboud**: Critical Feedback
- **Megan L. Hochstrasser**: Editing, Visualization
- **Kira E. Poskanzer**: Supervision
- **Taylor Reiter**: Critical Feedback, Data Curation, Validation
- **MaryClare Rollins**: Project Administration, Resources, Supervision
- **Peter S. Thuy-Boun**: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing

References

1. McClung KL, Little SE. (2023). Amblyomma americanum (Lone star tick). https://doi.org/10.1016/j.pt.2022.10.005
2. Geraci NS, Spencer Johnston J, Paul Robinson J, Wikel SK, Hill CA. (2007). Variation in genome size of argasid and ixodid ticks. https://doi.org/10.1016/j.ibmb.2006.12.007
3. Hotaling S, Kelley JL, Frandsen PB. (2021). Toward a genome sequence for every animal: Where are we now?. https://doi.org/10.1073/pnas.2109019118
4. Archer CT, Kim JF, Jeong H, Park JH, Vickers CE, Lee SY, Nielsen LK. (2011). The genome sequence of E. coli W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of E. coli. https://doi.org/10.1186/1471-2164-12-9
5. 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

6. Chou S, Hochstrasser ML, Poskanzer KE, Thuy-Boun PS. (2022). Robust long-read saliva transcriptome and proteome from the lone star tick, Amblyomma americanum. https://doi.org/10.57844/arcadia-3hyh-3h83
7. De S, Kingan SB, Kitsou C, Portik DM, Foor SD, Frederick JC, Rana VS, Paulat NS, Ray DA, Wang Y, Glenn TC, Pal U. (2023). A high-quality Ixodes scapularis genome advances tick science. https://doi.org/10.1038/s41588-022-01275-w
8. Gulia-Nuss M, Nuss AB, Meyer JM, Sonenshine DE, Roe RM, Waterhouse RM, Sattelle DB, de la Fuente J, Ribeiro JM, Megy K, Thimmapuram J, Miller JR, Walenz BP, Koren S, Hostetler JB, Thiagarajan M, Joardar VS, Hannick LI, Bidwell S, Hammond MP, Young S, Zeng Q, Abrudan JL, Almeida FC, Ayllón N, Bhide K, Bissinger BW, Bonzon-Kulichenko E, Buckingham SD, Caffrey DR, Caimano MJ, Croset V, Driscoll T, Gilbert D, Gillespie JJ, Giraldo-Calderón GI, Grabowski JM, Jiang D, Khalil SMS, Kim D, Kocan KM, Koči J, Kuhn RJ, Kurtti TJ, Lees K, Lang EG, Kennedy RC, Kwon H, Perera R, Qi Y, Radolf JD, Sakamoto JM, Sánchez-Gracia A, Severo MS, Silverman N, Šimo L, Tojo M, Tornador C, Van Zee JP, Vázquez J, Vieira FG, Villar M, Wespiser AR, Yang Y, Zhu J, Arensburger P, Pietrantonio PV, Barker SC, Shao R, Zdobnov EM, Hauser F, Grimmelikhuijzen CJP, Park Y, Rozas J, Benton R, Pedra JHF, Nelson DR, Unger MF, Tubio JMC, Tu Z, Robertson HM, Shumway M, Sutton G, Wortman JR, Lawson D, Wikel SK, Nene VM, Fraser CM, Collins FH, Birren B, Nelson KE, Caler E, Hill CA. (2016). Genomic insights into the Ixodes scapularis tick vector of Lyme disease. https://doi.org/10.1038/ncomms10507
9. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, Sedlazeck FJ, Marschall T, Mayes S, Costa V, Zook JM, Liu KJ, Kilburn D, Sorensen M, Munson KM, Vollger MR, Monlong J, Garrison E, Eichler EE, Salama S, Haussler D, Green RE, Akeson M, Phillippy A, Miga KH, Carnevali P, Jain M, Paten B. (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. https://doi.org/10.1038/s41587-020-0503-6
10. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. (2019). Assembly of long, error-prone reads using repeat graphs. https://doi.org/10.1038/s41587-019-0072-8
11. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. https://doi.org/10.1038/s41592-020-01056-5
12. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. https://doi.org/10.1093/bioinformatics/btaa025