RESEARCH ARTICLE

A draft genome sequence for the *Ixodes scapularis* cell line, ISE6 [version 1; referees: 2 approved]

Jason R. Miller[1,2], Sergey Koren[3], Kari A. Dilley[1], Derek M. Harkins[1], Timothy B. Stockwell[1,4], Reed S. Shabman[1,5], Granger G. Sutton[1]

[1]J. Craig Venter Institute, Rockville, MD, 20850, USA
[2]Shepherd University, Shepherdstown, WV, 25443, USA
[3]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, MD, 20892, USA
[4]NBACC, Fort Detrick, MD, 21702, USA
[5]ATCC, Gaithersburg, MD, 20877, USA

v1 **First published:** 08 Mar 2018, **7**:297 (doi: 10.12688/f1000research.13635.1)
Latest published: 08 Mar 2018, **7**:297 (doi: 10.12688/f1000research.13635.1)

Abstract

Background: The tick cell line ISE6, derived from *Ixodes scapularis*, is commonly used for amplification and detection of arboviruses in environmental or clinical samples.
Methods: To assist with sequence-based assays, we sequenced the ISE6 genome with single-molecule, long-read technology.
Results: The draft assembly appears near complete based on gene content analysis, though it appears to lack some instances of repeats in this highly repetitive genome. The assembly appears to have separated the haplotypes at many loci. DNA short read pairs, used for validation only, mapped to the cell line assembly at a higher rate than they mapped to the *Ixodes scapularis* reference genome sequence.
Conclusions: The assembly could be useful for filtering host genome sequence from sequence data obtained from cells infected with pathogens.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
version 1 published 08 Mar 2018	✓ report	✓ report

1 **Scott Emrich**, University of Tennessee, USA

2 **Chan Heu**, Pennsylvania State University, USA

Discuss this article

Comments (0)

Corresponding author: Jason R. Miller (jmill02@shepherd.edu)

Author roles: Miller JR: Formal Analysis, Writing – Original Draft Preparation, Writing – Review & Editing; **Koren S**: Formal Analysis, Writing – Original Draft Preparation; **Dilley KA**: Investigation; **Harkins DM**: Investigation; **Stockwell TB**: Conceptualization, Funding Acquisition; **Shabman RS**: Conceptualization, Funding Acquisition, Supervision; **Sutton GG**: Conceptualization, Funding Acquisition, Supervision

Competing interests: SK has received travel support to speak at Oxford Nanopore conferences.

How to cite this article: Miller JR, Koren S, Dilley KA *et al.* **A draft genome sequence for the *Ixodes scapularis* cell line, ISE6 [version 1; referees: 2 approved]** *F1000Research* 2018, **7**:297 (doi: 10.12688/f1000research.13635.1)

Grant information: JCVI staff was supported by DHS contract HSHQDC-15-C-B0059. SK was supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 08 Mar 2018, **7**:297 (doi: 10.12688/f1000research.13635.1)

Introduction

The *Ixodes scapularis* embryonic 6 (ISE6) cell line is a widely used resource that is permissive to pathogens including human pathogens transmitted by ticks. Two decades ago, a collection of *I. scapularis* cell lines were derived from embryonated eggs including IDE lines derived from northern ticks and ISE lines derived from southern ticks (Munderloh *et al.*, 1994). Recent proteomics analysis suggested that the ISE6 line is derived from neuronal cells (Oliver *et al.*, 2015). ISE6 cells have been used to isolate and analyze bacterial pathogens including: the causative agent of human granulocytic ehrlichiosis (HGE) (Munderloh *et al.*, 1999); *Borrelia burgdorferi*, the causative agent of Lyme disease (Obonyo *et al.*, 1999); the causative agent of southern tick-associated rash illness (STARI) (Varela *et al.*, 2004); and *Rickettsia felis*, the causative agent of spotted fever (Pornwiroon *et al.*, 2006). ISE6 cells have been used to study viral pathogens including: Semliki Forest virus (SFV) and Hazara virus (arbovirus, family *Bunyaviridae*, genus *Nairovirus*) (Garcia *et al.*, 2005); and Langat virus (LGTV), a Flavivirus (Grabowski *et al.*, 2016). The cells have also been used to study RNAi and genome engineering in ticks; reviewed in (Oliver *et al.*, 2015).

The *Ixodes scapularis* (black-legged tick) genome had been estimated to harbor 70% repeat content (Ullmann *et al.*, 2005) when our lab participated in a community effort to sequence and assemble a tick genome. A reference sequence built from 3.8X Sanger sequencing and the Celera Assembler (Istrail *et al.*, 2004; Myers *et al.*, 2000) software was fractured into 570,640 contigs (369,495 scaffolds) with a contig N50 of only 2,942 bp. The total contig span was 1.388 Gbp though the genome size was estimated at 2.1 Gbp. The assembly supported the annotation of 20,486 protein-coding genes and an extensive analysis of tick biology (Gulia-Nuss *et al.*, 2016). The assembly is maintained at VectorBase (Giraldo-Calderón *et al.*, 2015) under the name IscaW1.

A genome assembly for the ISE6 genome would assist investigations of ISE6 as a biological system. It would also provide a host subtraction tool for ISE6-based sequencing assays. Host subtraction is the bioinformatics process of filtering reads whose origin is host DNA and RNA (Daly *et al.*, 2015). Host subtraction enriches the non-host component of sequence datasets and is especially attractive for assays involving high-throughput sequencing technologies that generate short reads in high volume where data reduction can realize cost savings. Following host subtraction, remaining reads can be mapped to references and counted, or used as queries to sequence databases, or assembled to reconstruct novel transcript or genome sequences. With an expectation that the IES6 genome would be as challenging as the tick genome, we sequenced IES6 with high coverage and long reads that, taken together, might generate a high quality reference genome assembly.

Methods

Cell growth

ISE6 cells were obtained from the American Type Culture Collection (ATCC), cell line CRL-11974, lot number 100005, patent 5,869,335. This cell line had been isolated from ticks collected in Georgia, USA. As generally described by ATCC technical bulletins and personal communications, cells were grown in Leibovitz L-15B media pH 7.0 (Leibovitz, 1963) (ThermoFisher Scientific) supplemented with 80 mM glucose, 10% tryptose phosphate broth, 0.1% bovine lipoprotein cholesterol concentrate and 2% heat inactivated fetal bovine serum (Munderloh & Kurtti, 1989). An addition of 0.7% non-essential amino acids (NEAA) concentrate (ThermoFisher Scientific) was also added prior to incubation. The cells were incubated in a gently shaking flask at 31°C with no $C0_2$.

Sequencing

For short-read sequencing, genomic DNA was isolated from the cell line using a Qiagen genomic DNA isolation kit. Bioanalyzer analysis confirmed high molecular weight DNA was recovered. The library was size selected using Pippin Prep and prepared using the NextGen paired end barcoded genomic library construction protocol. Library quantification and normalization was performed by qPCR. The library was sequenced on the Illumina NextSeq 500 platform to generate 2x150 paired reads. Reads were demultiplexed which removed barcodes and sequencing adapters, and further treated with CutAdapt 1.8.1 to remove any remaining adapter.

For long-read sequencing, cells were grown until they attached to the flask. Genomic DNA was extracted from ISE6 cells using a Qiagen Genomic DNA isolation kit, stopping prior to the G2 isolation step. Frozen pellets and frozen cells were shipped to the Icahn School of Medicine at Mount Sinai for library construction and sequencing using standard SMRTbell template preparation kits (Pacific Biosciences). A total of 52 SMRTcells were run on the PacBio Sequel platform using standard PacBio protocols.

Assembly and analysis

The long reads were corrected and assembled with the Canu assembler (Koren *et al.*, 2017) version 1.6. Canu was run with the SGE grid engine and Java 1.8 using default parameters except: minOverlapLength = 1000 bp, corMhapSensitivity = "low", and genomeSize = 1 Gbp. The contig consensus sequences were polished using SMRT Link version 5.0.1.9585 which includes Arrow version 2.2.1, blasr 5.3, and pbalign 0.3.1 (Pacific Biosciences).

Short reads were mapped with bowtie2 (Langmead & Salzberg, 2012) version 2.2.5 using either end-to-end or local-alignment mode as indicated in the text. Using default settings, the mapper reported at most one mapping per read and reported read maps individually though it used reads as pairs to select alignments. Mappings were analyzed with samtools (Li *et al.*, 2009) version 1.2.1 and bedtools (Quinlan, 2014; Quinlan & Hall, 2010) version 2.26. K-mers were counted using Jellyfish (Marçais & Kingsford, 2011) version 2.2.6 for several values of K: 11, 15, 21, 41, 51, and 61. Each computed K-mer histogram was analyzed with GenomeScope (Vurture *et al.*, 2017). Single-copy gene analysis used BUSCO (Simão *et al.*, 2015) version 3.0.2 with Arthropoda OrthoDB version 9. Alignments between the

tick and cell line assemblies were computed on contigs from each assembly using nucmer, the local aligner in the MUMmer package (Kurtz *et al.*, 2004) version 3.1. The BLAST analysis used TBLASTN in NCBI BLAST+ (Camacho *et al.*, 2009) version 2.2.31. The Ixodes protein predictions were downloaded from UniProt in Nov 2017.

Results

Assembly

PacBio sequencing yielded 192.5 Gbp in 27.3 million unpaired reads, providing approximately 92X coverage of the estimated 2.1 Gbp tick genome. This dataset included 190.7 Gbp in reads >= 1 Kbp and 115.9 Gbp in reads >= 10 Kbp. To overcome high base call error observed in single-molecule long-read data, the long reads were subjected to the Canu correction process which filtered, trimmed, and polished reads based on alignment with other long reads (Koren *et al.*, 2017). Correction yielded 36.7 Gbp in 2.1 million reads with a 17,680 bp N50. Read counts are shown in Table S1. Read length distributions, before and after correction, are shown in Figure S1.

The corrected long reads were assembled in isolation with the Canu long-read assembler (Koren *et al.*, 2017). The initial assembly, named Ise6_asm0, contained 18,717 contigs. The uncorrected long reads were used to polish the contig consensus sequences using the Arrow process. This was run in two iterations to produce assemblies Ise6_asm1 and Ise6_asm2 respectively. The released assembly, named Ise6_asm2, contained 2,691,078,110 bases in 18,717 contigs with a 269,660 bp contig N50. Statistics for each assembly are shown in Table S2.

Assembly quality assessment

Since the tick contigs are generally smaller than the cell line contigs, it seemed likely that some tick contigs would be wholly contained by cell line contigs. The IscaW1 and Ise6_asm0 assemblies were compared with the MUMmer local alignment software; see Table S3. A high-confidence alignment subset (filtered with delta-filter -1) was used. The average identity of aligned bases was 96.58%. Both assemblies were almost fully covered by alignments. Using alignments of 1 Kbp or more, IscaW1 contigs with at least one alignment made up 78.6% of IscaW1 bases. The set of IscaW1 contigs with at least 90% coverage in any one alignment made up 35.5% of IscaW1 bases.

To provide an orthogonal dataset for assembly assessment, the genome was also sequenced on a short-read platform. Illumina sequencing yielded 25 Gbp in 171 million 2x150 bp paired reads, providing approximately 24X short-read coverage of the 2.1 Gbp estimated genome size of the tick. Comparative map statistics are shown in Table S4. When paired reads were mapped to Ise6_asm0 using a local alignment algorithm, 97.83% of reads mapped as a concordant pair. For comparison, when the reads were mapped to the IscaW1 tick reference assembly, 91.68% of reads mapped as a concordant pair. This result indicates that the ISE6 assembly is more representative of ISE6 genome structure than the *Ixodes* reference.

For consensus quality assessment, the paired reads were mapped to Ise6_asm2 and IscaW1 contigs using a more stringent, global (end-to-end) alignment algorithm; see Table S4. Among these alignments, the read sequence disagreement with the contig consensus was 1.79% for Ise6_asm2 and 5.03% for IscaW1. This demonstrates that the ISE6 consensus is more representative of ISE6 genome sequence than the *Ixodes* reference.

The rates of concordant pair mapping to zero, one, or multiple sites were 23%, 29%, and 48% respectively for Ise6_asm2 and 44%, 30%, and 25% for IscaW1; see Table S4. Thus, by paired-read mappability, both assemblies contain 29%–30% unique sequence while the ISE6 assembly captures an additional 23% of reads and these align to repeat sequences in the assembly.

The global alignment 23% unmapped rate in Ise6_asm2 is an order of magnitude larger than the unmapped rate among the local alignments. It is possible that the long and short read sequencing captured genuine differences at unstable regions of the cell line genome. It seems more likely that the genome harbors repeat instances that are similar-but-not-identical to those in the assembly.

Using the global alignments and accepting all mapped reads (whether mapped as a pair or not), the Ise6_asm2 assembly mapped 81% of reads while IscaW1 mapped 65%. Thus, the Ise6_asm2 assembly outperformed the IscaW1 assembly as a host subtraction tool using pairwise local, pairwise global, and read-wise global alignments.

The assembly was assessed for completeness using gene content analysis. The latest UniProt protein predictions on the IscaW1 tick genome assembly were used as TBLASTN query sequences against the cell line assembly. Out of 20,473 predicted proteins: 20,290 (99.1%) had at least one hit in Ise6_asm0 while 183 predictions had no hit. The Ise6_asm2 assembly was analyzed for gene content using the BUSCO collection of genes thought to be single-copy in arthropod genomes; Table S5. Of 1066 genes searched, 1.4% were fragmented, 3.6% were missing, and 95% were complete. These results indicate that the assembly is fairly complete for single-copy genes.

Genome size analysis

The Ise6_asm2 contig span is 2.8 Gbp which exceeds the 1.4 Gbp contig span of the IscaW1 tick reference assembly as well as the 2.1 Gbp estimated genome size for tick. The discrepancy could be due to several factors. It is possible that the cell line genome is larger than the tick genome, or that the assembly contains dual representations of heterozygous loci that assembled separately, or that the IscaW1 reference assembly underrepresents repeats present in the tick and ISE6 genomes. These possibilities were explored with several analyses.

K-mer analysis (Vurture *et al.*, 2017) provides an assembly-free genome size estimate extrapolated from the frequency distributions of short, contiguous sequences extracted from the

sequencing reads and counted using an exact-match algorithm (Marçais & Kingsford, 2011). In K-mer analysis of our short-read data, the observed distribution could not be fit to a model. Analysis of the corrected long read K-mers was similarly inconclusive. As illustrated by the representative plot in Figure S2, the distribution does not include a strong peak other than the mode at 1X. These results are possibly due to the low coverage in both datasets, which would be especially low if heterozygous haplotypes were represented separately in contigs.

Next, genome size estimation was attempted using coverage analysis of paired reads mapped to contigs. Pairs mapped end-to-end provided 39.9 Gbp of mapped bases. In the distribution of read coverage per contig base, there is a smooth peak with a 9X mode and a tail at higher coverage; see Figure S3. Using the mapped base count divided by 9X to represent the average coverage of unique sequence in the genome, the extrapolated genome size is 4.43 Gbp, which is about twice as large as the tick estimated genome size. This suggests that the assembly process separated the haplotypes of a 2.22 Gbp diploid genome.

Long-read coverage was analyzed next. The uncorrected long reads had been mapped to Ise6_asm1 contigs for the final iteration of Arrow consensus polish. Based on those results, the per-base coverage peaks at 34X with a shoulder at almost twice that level; see Figure 1. This suggests that 34X represents the coverage mode for haplotype-separated sequence. With 175.48 Gbp in mapped reads, 158.47 Gbp of read sequence aligned to 151.93 Gbp of contig sequence. These mapping and coverage results combine to indicate 151.93 / 34 = 4.47 Gbp size for the combined haplotypes (diploid) and 2.24 Gbp for the haploid genome. Thus, the completeness of the 2.8 Gbp assembly is uncertain and the assembly appears to harbor double-representation of loci that are haplotype-separated and under-representation of genomic repeats, as indicated in Figure 1.

Local alignments between assemblies were examined for support of the haplotype separation hypothesis. The cell line Ise6_asm0 contigs were aligned to the tick reference IscaW1 contigs using the nucmer local alignment software. As shown in Figure S4, the largest IscaW1 contig aligns full-length to two contigs of Ise6_asm0. The full set of alignments was filtered to retain one best alignment per Ise6_asm0 position (with delta-filter -q) and to retain only IscaW1 contigs with at least 50% coverage in one such alignment. Using these alignments to distinguish IscaW1:ISE6 multiplicities, 43% of IscaW1 bases are in contigs with 1:1 relation to ISE6, while 17% of bases are in contigs with 1:2 relations, and 3% of bases are in contigs with

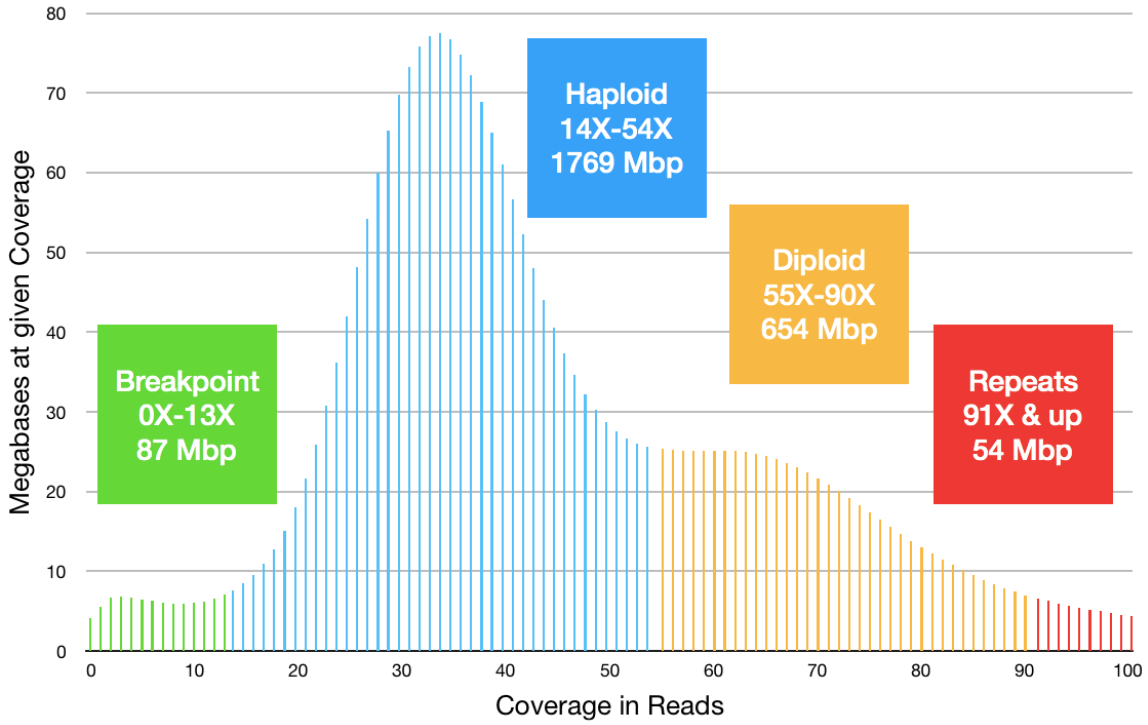

Figure 1. Long read coverage of the cell line assembly. Read coverage per base was computed by mapping uncorrected long reads to assembled contigs. Colors were added to highlight the following interpretation. Green: the minor peak of low-coverage bases suggests 87 Mbp of contig bases lies near breakpoints such as contig ends and false joins. Blue: the dominant peak, with mode at 34X, suggests that 1769 Mbp of the assembly is haplotype-separated sequence possibly representing 885 Mbp of the diploid genome. Yellow: the shoulder near 64X suggests that about 654 Mbp of the diploid genome is captured as diploid-consensus sequence. Red: high coverage (including bases not shown with coverage over 100X) suggests that under-represented genomic repeats occupy 54 Mbp of the assembly but more of the genome.

relations of 1:3 or higher; see Table S3. Not all contigs could be categorized. Further work would be required to fully partition and phase all duplicated contigs with high confidence.

Of 1066 BUSCO genes searched against Ise6_asm2, 67% were complete and single-copy, and 28% were complete and duplicated. This indicates that almost a third of single-copy genes are duplicated in the assembly, presumably due to haplotype separation. If this rate of sequence duplication applies to the assembly overall, the 2.67 Gbp assembly indicates a 2.29 Gbp genome size.

Non-Ixodes sequence

During the assembly submission to GenBank, NCBI flagged one contig for sequence similarity to *Rickettsia*, a genus of endosymbionts common to *Ixodes scapularis* (Gillespie *et al.*, 2012; Labruna *et al.*, 2007; Steiner *et al.*, 2008; Zeringóta *et al.*, 2017). The long read coverage in the flagged region differs from the flanking regions; see Figure S5. The contig was split and re-submitted without the flagged region.

Discussion

The ISE6 cell line, derived from the Lyme disease tick *Ixodes scapularis*, is widely used but has so far lacked a genome reference sequence. With this report, the genome has been sequenced, assembled, and released. The genome sequence was generated from high coverage in PacBio Sequel long reads, assembled with the Canu assembler, and polished with Arrow. Gene content analysis indicated that the assembly is largely complete though read mapping analysis indicated that some genomic repeats are under-represented in the assembly. Read mapping and single-gene analysis indicated that portions of the genome are represented as a diploid consensus while other portions are represented in haplotype-separated copies. The new assembly provides a more accurate representation of the cell line genome compared to the previous closest reference, which was an assembly of the *I. scapularis* tick genome. In our mapping of cell line gDNA short read pairs that were not used for the assembly, the cell line assembly was more effective for identifying host reads compared to the tick reference. Thus, the new assembly provides a resource for analysis of the cell line and for host subtraction to assist the detection of pathogens present in the cells.

Comparable genome size estimates were obtained by three methods. Short-read coverage analysis indicated 2.22 Gbp. Long-read coverage indicated 2.24 Gbp. Single-copy gene analysis indicated 2.29 Gbp. The tick genome was previously estimated to be 2.1 Gbp so the cell line may harbor some ISE6-specific sequence. Identification of such sequences is left for future work. Our local alignments of the tick and ISE6 assemblies covered nearly all of both assemblies, so any cell-line-specific sequence is likely to involve genomic repeats.

We hope to enhance the assembly resource in several ways. We hope to generate a second version of the sequence using different assembly parameters. A larger genome size estimate, in particular, should lead to higher generated coverage in corrected reads. Additional corrected reads would necessarily be shorter than the ones used here, but they could provide additional depth for repeat detection and repeat resolution. We hope to provide gene annotation, repeat content analysis, and more specific haplotype separation analysis of the second assembly.

Data availability

The gDNA Illumina and PacBio reads are available at NCBI SRA under BioSample SAMN06329993. This entire assembly has been deposited at GenBank under the accession GCA_002892825.1. Whole Genome Shotgun sequencing project is available under accession PKSA00000000.

Competing interests

SK has received travel support to speak at Oxford Nanopore conferences.

Grant information

JCVI staff was supported by DHS contract HSHQDC-15-C-B0059. SK was supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

This study utilized the computational resources of the Biowulf system at the National Institutes of Health, Bethesda, MD (https://biowulf.nih.gov). Vinita Puri assisted the sample preparation at JCVI. PacBio sequencing was performed at the Icahn School of Medicine at Mount Sinai by Robert P. Sebra, Melissa L. Smith, and Ying-Chih Wang.

Supplementary material

Figure S1. Length distributions for PacBio long reads before (blue) and after (green) correction by the Canu assembler. The comparison indicates that the correction step removed most short reads while reducing the number and lengths of long reads. Each blue or green vertical bar represents the number of reads in a 10 bp window. Reads longer than 30 Kbp (not shown) provided 1,623,767,067 bp in 47,967 reads before correction and 333,923,002 bp in 10,211 reads after correction.
Click here to access the data.

Figure S2. K-mer frequency distribution plotted by GenomeScope from K-mers found in PacBio long reads by Jellyfish with K=41. Similar plots were generated by GenomeScope for selected values of K ranging from 11 to 61 from the corrected long reads and also the (uncorrected) short reads. The distribution is dominated by K-mers that occur once (coverage = 1X), which can be attributed to random base call error.

Click here to access the data.

Figure S3. Histogram of short-read coverage of large contigs. The mode of the distribution is at 9X. The minor peak at 0X indicates uncovered contig bases, which are frequent at contig ends. The pseudo-peak at 51X represents all bases covered at 51X or greater. Reads were mapped to all contigs but only large contigs (length >= 10Kbp) were plotted here. The mapping used end-to-end alignments and reported at most one map per read. The mapping used reads as pairs but reported all mapped reads. A map report produced with 'samtools stats' reported 277,543,724 reads mapped including 261,690,992 properly paired in the mapping, and 39,913,302,735 based mapped and no clipping. Assuming 9X is the average read coverage of unique sequence in the genome, the genome size is estimated as 278 Gbp / 9X = 4.4 Gbp, or roughly twice the tick estimated genome size. This suggests that the assembly captures many diploid loci twice and that 9X is the average coverage of a haplotype-separated sequence in the assembly.

Click here to access the data.

Figure S4. Alignment of two ISE6 contigs to one IscaW1 contig. Contig ABJB010274751.1, the longest IscaW1 contig, has length 117,687 bp and is represented full-length on the X-axis. This contig has nearly-complete, full-length coverage in alignments to two ISE6_ASM1 contigs, each represented full-length on the Y-axis. Alignments were computed with nucmer and plotted with mummerplot using minimum length 1000 bp. ISE6_ASM1 contig tig00013410 (top) has length 309,049 bp and 19 local alignments including the 13 large alignments shown. ISE6_ASM1 contig contig tig00009157 has length 719,239 bp (bottom) and 30 local alignments including the 20 large alignments shown.

Click here to access the data.

Figure S5. Sequence identified as *Rickettsia* has unusual read coverage distribution. As part of the Ise6_asm2 assembly submission to GenBank, NCBI flagged one contig for having sequence similarity to *Rickettsia*, A search using NCBI megablast and the nt database confirmed that tig00009859, of length 618,170 bp, had a portion of sequence that was most similar to *Rickettsia* (various species). Uncorrected PacBio reads had been mapped to the Ise6_asm1 version of the assembly as part of the Arrow polish process that generated Ise6_asm2. Read coverage at every base was extracted from that mapping using samtools. Read coverage was plotted for the three coordinate ranges 1–250,000 (orange, upstream), 250,000–350,000 (blue, contig middle), and 350,000–600,000 (downstream, green). The flagged region falls within the contig middle coordinates. Because the contig middle appeared to have a different coverage profile from both flanking sequencings, the flagged region (268 Kbp to 330 Kbp) of the contig was removed and the upstream and downstream contigs were re-submitted to GenBank.

Click here to access the data.

Table S1. ISE6 sequencing results.

Click here to access the data.

Table S2. ISE6 contig size statistics for several assemblies.

Click here to access the data.

Table S3. Nucleotide local alignments between the contigs of IscaW1 and Ise6_asm0. Alignments were generated and analyzed with the MUMmer package including the delta-filter alignment filter.

Click here to access the data.

Table S4. Alignment rates for paired reads mapped to contigs. Short reads were mapped with bowtie2 in local mode or global (end-to-end) mode.

Click here to access the data.

Table S5. Single-gene content analysis with BUSCO. Analysis of Ise6_asm2 using the Arthropoda OrthoDB version 9 single-copy genes.

Click here to access the data.

References

Camacho C, Coulouris G, Avagyan V, *et al.*: **BLAST+: architecture and applications.** *BMC Bioinformatics.* 2009; **10**: 421.
PubMed Abstract | Publisher Full Text | Free Full Text

Daly GM, Leggett RM, Rowe W, *et al.*: **Host Subtraction, Filtering and Assembly Validations for Novel Viral Discovery Using Next Generation Sequencing Data.** *PLoS One.* 2015; **10**(6): e0129059.
PubMed Abstract | Publisher Full Text | Free Full Text

Garcia S, Billecocq A, Crance JM, *et al.*: **Nairovirus RNA sequences expressed by a Semliki Forest virus replicon induce RNA interference in tick cells.** *J Virol.* 2005; **79**(14): 8942–8947.
PubMed Abstract | Publisher Full Text | Free Full Text

Gillespie JJ, Joardar V, Williams KP, *et al.*: **A Rickettsia genome overrun by mobile genetic elements provides insight into the acquisition of genes characteristic of an obligate intracellular lifestyle.** *J Bacteriol.* 2012; **194**(2): 376–394.
PubMed Abstract | Publisher Full Text | Free Full Text

Giraldo-Calderón GI, Emrich SJ, MacCallum RM, *et al.*: **VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases.** *Nucleic Acids Res.* 2015; **43**(Database issue): D707–13.
PubMed Abstract | Publisher Full Text | Free Full Text

Grabowski JM, Perera R, Roumani AM, *et al.*: **Changes in the Proteome of Langat-Infected *Ixodes scapularis* ISE6 Cells: Metabolic Pathways Associated with Flavivirus Infection.** *PLoS Negl Trop Dis.* 2016; **10**(2): e0004180.
PubMed Abstract | Publisher Full Text | Free Full Text

Gulia-Nuss M, Nuss AB, Meyer JM, *et al.*: **Genomic insights into the *Ixodes scapularis* tick vector of Lyme disease.** *Nat Commun.* 2016; **7**: 10507.
PubMed Abstract | Publisher Full Text | Free Full Text

Istrail S, Sutton GG, Florea L, *et al.*: **Whole-genome shotgun assembly and comparison of human genome assemblies.** *Proc Natl Acad Sci U S A.* 2004; **101**(7): 1916–1921.
PubMed Abstract | Publisher Full Text | Free Full Text

Koren S, Walenz BP, Berlin K, *et al.*: **Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation.** *Genome Res.* 2017; **27**(5): 722–736.
PubMed Abstract | Publisher Full Text | Free Full Text

Kurtz S, Phillippy A, Delcher AL, *et al.*: **Versatile and open software for comparing large genomes.** *Genome Biol.* 2004; **5**(2): R12.
PubMed Abstract | Publisher Full Text | Free Full Text

Labruna MB, Pacheco RC, Richtzenhain LJ, *et al.*: **Isolation of *Rickettsia rhipicephali* and *Rickettsia bellii* from *Haemaphysalis juxtakochi* ticks in the state of São Paulo, Brazil.** *Appl Environ Microbiol.* 2007; **73**(3): 869–873.
PubMed Abstract | Publisher Full Text | Free Full Text

Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods.* 2012; **9**(4): 357–359.
PubMed Abstract | Publisher Full Text | Free Full Text

Leibovitz A: **The growth and maintenance of tissue-cell cultures in free gas exchange with the atmosphere.** *Am J Hyg.* 1963; **78**(2): 173–180.
PubMed Abstract | Publisher Full Text

Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–2079.
PubMed Abstract | Publisher Full Text | Free Full Text

Marçais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers.** *Bioinformatics.* 2011; **27**(6): 764–770.
PubMed Abstract | Publisher Full Text | Free Full Text

Munderloh UG, Jauron SD, Fingerle V, *et al.*: **Invasion and intracellular development of the human granulocytic ehrlichiosis agent in tick cell culture.** *J Clin Microbiol.* 1999; **37**(8): 2518–2524.
PubMed Abstract | Free Full Text

Munderloh UG, Kurtti TJ: **Formulation of medium for tick cell culture.** *Exp Appl Acarol.* 1989; **7**(3): 219–229.
PubMed Abstract | Publisher Full Text

Munderloh UG, Liu Y, Wang M, *et al.*: **Establishment, maintenance and description of cell lines from the tick Ixodes scapularis.** *J Parasitol.* 1994; **80**(4): 533–543.
PubMed Abstract | Publisher Full Text

Myers EW, Sutton GG, Delcher AL, *et al.*: **A whole-genome assembly of *Drosophila*.** *Science.* 2000; **287**(5461): 2196–2204.
PubMed Abstract | Publisher Full Text

Obonyo M, Munderloh UG, Fingerle V, *et al.*: ***Borrelia burgdorferi* in tick cell culture modulates expression of outer surface proteins A and C in response to temperature.** *J Clin Microbiol.* 1999; **37**(7): 2137–2141.
PubMed Abstract | Free Full Text

Oliver JD, Chávez AS, Felsheim RF, *et al.*: **An *Ixodes scapularis* cell line with a predominantly neuron-like phenotype.** *Exp Appl Acarol.* 2015; **66**(3): 427–442.
PubMed Abstract | Publisher Full Text | Free Full Text

Pornwiroon W, Pourciau SS, Foil LD, *et al.*: ***Rickettsia felis* from cat fleas: isolation and culture in a tick-derived cell line.** *Appl Environ Microbiol.* 2006; **72**(8): 5589–5595.
PubMed Abstract | Publisher Full Text | Free Full Text

Quinlan AR: **BEDTools: The Swiss-Army Tool for Genome Feature Analysis.** *Curr Protoc Bioinformatics.* 2014; **47**: 11.12.1–34.
PubMed Abstract | Publisher Full Text | Free Full Text

Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics.* 2010; **26**(6): 841–842.
PubMed Abstract | Publisher Full Text | Free Full Text

Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics.* 2015; **31**(19): 3210–3212.
PubMed Abstract | Publisher Full Text

Steiner FE, Pinger RR, Vann CN, *et al.*: **Infection and co-infection rates of *Anaplasma phagocytophilum* variants, *Babesia* spp., *Borrelia burgdorferi*, and the rickettsial endosymbiont in *Ixodes scapularis* (Acari: Ixodidae) from sites in Indiana, Maine, Pennsylvania, and Wisconsin.** *J Med Entomol.* 2008; **45**(2): 289–297.
PubMed Abstract | Publisher Full Text

Ullmann AJ, Lima CM, Guerrero FD, *et al.*: **Genome size and organization in the blacklegged tick, *Ixodes scapularis* and the Southern cattle tick, *Boophilus microplus*.** *Insect Mol Biol.* 2005; **14**(2): 217–222.
PubMed Abstract | Publisher Full Text

Varela AS, Luttrell MP, Howerth EW, *et al.*: **First culture isolation of *Borrelia lonestari*, putative agent of southern tick-associated rash illness.** *J Clin Microbiol.* 2004; **42**(3): 1163–1169.
PubMed Abstract | Publisher Full Text | Free Full Text

Vurture GW, Sedlazeck FJ, Nattestad M, *et al.*: **GenomeScope: fast reference-free genome profiling from short reads.** *Bioinformatics.* 2017; **33**(14): 2202–2204.
PubMed Abstract | Publisher Full Text

Zeringóta V, Maturano R, Luz HR, *et al.*: **Molecular detection of *Rickettsia rhipicephali* and other spotted fever group *Rickettsia* species in *Amblyomma* ticks infesting wild birds in the state of Minas Gerais, Brazil.** *Ticks Tick Borne Dis.* 2017; **8**(1): 81–89.
PubMed Abstract | Publisher Full Text

Open Peer Review

Current Referee Status: ✔ ✔

Version 1

Referee Report 03 April 2018

doi:10.5256/f1000research.14813.r31976

✔ **Chan Heu**
Pennsylvania State University, State College, PA, USA

The authors assembled a draft genome of ISE6 using long-read sequencing. In addition, they used short-read sequencing to validate and distinguish the ISE6 assembly through differences in genome size and mapping coverage from the already sequenced, albeit incomplete, *Ixodes scapularis* genome. The primary aim is to generate the genome of ISE6 for host genomic for potential application such as the subtraction from genomic analysis of pathogens and/or symbionts cultured in ISE6. The authors demonstrated very well that the draft genome of ISE6 showed more coverage than the tick genome. It was also a great approach to use three different methods to estimate the genome size.

There were a few areas that could have been improved upon. This article would have benefited from demonstrating if there is any difference in using the ISE6 genome assembly vs. the tick genome for subtracting host genomic background, or in some other application in which using the assembled ISE6 genome would be more beneficial than the tick genome. In addition, the authors could discuss further the origin of the contig that was similar to *Rickettsia*. For instance, the authors could hypothesize whether this contig came from a lateral gene transfer event, rickettsial contamination of the starting material, or technical error in the assembly process. Nevertheless, having a draft genome of ISE6 will be useful for in vitro research involving tick-associated bacteria, protozoans, and viruses.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Referee Expertise: Rickettsiology, tick-associated bacteria, molecular biology, genetics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 22 March 2018

doi:10.5256/f1000research.14813.r31696

Scott Emrich
University of Tennessee, Knoxville, TN, USA

Miller et al. present a long-read based assembly of the ISE6 tick cell line. Although mostly in the background/discussion sections of this paper, the Ixodes genome has been hard to assemble given a large number of repeats and heterogeneity between individuals. Similar to other recent long read efforts of mosquitoes, continuity / completeness has been improved by at the cost of residual haplotypes; only ~66% of the genome is single copy.

On the one hand this does not impact the main use case of this effort, namely an improved reference for host extraction. As a nice twist the authors included an Illumina experiment, for example, and showed more host (tick) DNA aligned to their reference. Given this cell line is most used for pathogen-related work this is a valuable contribution even though the assembly itself contains haplotypes.

On the other hand, this paper in of itself does not really provide any new solutions or even discussion on how to improve assembly or a reference tick genome, which makes it more of a data note (and fine for F1000Research). For example, getting DNA from a single individual is sometimes hard so other projects have resorted to using full sibs. Given that the cell line was derived from ticks (plural) is it even feasible to dilute the cell line to get what amounts to clonal flasks and try and sequence that? That would not help the repeat issue but may reduce the amount of heterogeneity.

I have a few remaining specific questions/comments listed below:

1. On page 3 you mention because 6.7% more pairs are concurrent you state "This result indicates that the ISE6 assembly is more representative of ISE6 genome structure than the Ixodes reference." This is not wrong, but I have two followup questions. First, in the next section you observe that there is over 5% divergence between your line and the Ixodes reference (which makes sense). Could this affect your short-read alignment parameters and cause less pairs to map? Second, what if your contigs are 2X bigger that map these additional pairs but they are repeated in your assembly but not the genome (residual heterogyosity). I guess my issue is using "structure" vs. a more generic "contents" given the issues reported in the build.

2. Similarly, you state on page 3 that the 25% of reads that map to multiple sites are repeats, but couldn't they map to un merged alleles given the later results in the manuscript? I realize repeats is ambiguous and the haplotype issue is discussed later but maybe generalize to "repeated sequences present in the assembly" (or similar) to allow for assembly artifacts and "real" repetitive sequences present in tick

3. I was curious as to the BUSCO results of the old assembly ... the Ixodes reference was done long enough ago I am not sure they would have used it and a reader shouldn't have to look it up. In short, if we ignore the allele duplication issue are genes assembled better? The answer was yes for Aedes and I assume so here but it is unclear.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Not applicable

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: A resource on which I am am a principal investigator (VectorBase) is the NIH NIAID-supported repository for all vector genomes. In this role I received a brief email from the lead author about potential new genomes (including this one) for inclusion in our resource. I also collaborate indirectly with Sergei Koren on long read assembly either by providing our mosquito data to Adam Phillippy (NIH) to hopefully improve Canu for difficult arthropod genomes, or as a senior vector bioinformatician on independent genome publications (An. gambiae Y, less so for Ae. aegypti). I personally have never worked on tick and therefore have no relationship to this manuscript or the results presented here.

Referee Expertise: Genome informatics, vector genomics, sequence-based bioinformatics

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.