OXFORD

G3 Genes | Genomes | Genetics

otb: an Automated HiC/HiFi Pipeline Assembles the *Prosapia bicincta* Genome

David C. Molik *,[1], Amanda R. Stahlke [2], Sharu P. Sharma [3], Tyler J. Simmonds [4], Renee L. Corpuz [4], Angela N. Kauwe [4], Jeremy E. Schrader [4], Charles J. Mason [4], Sheina B. Sim [4] and Scott M. Geib [4]

[1] Arthropod-borne Animal Diseases Research Unit, Center for Grain and Animal Health Research, United States Department of Agriculture, Agricultural Research Service, 1515 College Ave, Manhattan, KS 66502 USA

[2] Colorado Mesa University, Department of Biological Sciences, Wubben Hall and Science Center, 1100 North Avenue, Grand Junction, CO 81501-3122 USA

[3] Genome Informatics Facility, Iowa State University, 2200 Osborn Drive, Ames, IA 50011-4009 USA

[4] Tropical Pest Genetics and Molecular Biology Research Unit, Daniel K. Inouye U.S. Pacific Basin Agricultural Research Center, United States Department of Agriculture, Agricultural Research Service, 64 Nowelo St, Hilo HI 96720 USA

*Corresponding author: david.molik@usda.gov

Abstract

The implementation of a new genomic assembly pipeline named only the best (otb) has effectively addressed various challenges associated with data management during the development and storage of genome assemblies. otb, which incorporates a comprehensive pipeline involving a setup layer, quality checks, templating, and the integration of Nextflow and Singularity. The primary objective of otb is to streamline the process of creating a HiFi/HiC genome, aiming to minimize the manual intervention required in the genome assembly process. The Two-lined spittlebug, (*Prosapia bicincta*, Hemiptera: Cercopidae), a true bug insect herbivore, serves as a practical test case for evaluating otb. The two-lined spittlebug is both a crucial agricultural pest and a genomically understudied insect belonging to the order Hemiptera. This insect is a significant threat to grasslands and pastures, leading to plant wilting and phytotoxemia when infested. Its presence in tropical and subtropical regions around the world poses a long-term threat to the composition of plant communities in grassland landscapes, impacting rangelands, and posing a substantial risk to cattle production.

Keywords: genome assembly; non-model organism; haplotype phasing; next-generation sequencing; assembly error correction

Introduction

The USDA-ARS AgPest 100 Initiative (Ag100pest) aims to generate high-quality genome assemblies of existing and/or emerging pest insect species that threaten agricultural production (Childers *et al.* 2021). High-quality genome assemblies can inform both basic and applied research. The time cost in the production of multiple genome assemblies can cause inefficiencies in projects such as Ag100pest; a HiFi/HiC assembly pipeline is required for efficiencies in these projects. HiFi reads refer to high-fidelity sequencing reads generated by the HiFi (High-Fidelity) sequencing technology, providing accurate and long-read DNA sequences for improved genomic analysis. HiC reads are DNA sequencing reads generated using Hi-C technology, which captures spatial proximity information of genomic loci, enabling the study of chromatin interactions and three-dimensional genome structure. To induce the implementation of such a pipeline, such a pipeline must be utilized in a test scenario. This pipeline should also be a complete assembly pipeline, instead of a polishing pipeline such as PolishCLR, although ideally both are utilized Chang *et al.* (2023).

As Ag100Pest often works with true bugs, and true bug genomics is relatively unexplored Jiang *et al.* (2021), an agriculturally significant true bug is the ideal test case for such a pipeline, Two-line spittlebugs (*Prosapia bicincta*) are such test cases. Two-lined spittlebug (Hemiptera: Cercopidae) is an insect herbivore distributed throughout the eastern part of the United States (Potter *et al.* 1991; Braman and Abraham 1995). In 2016, the two-lined spittlebug was first detected on Hawai'i Island (Thorne *et al.* 2017). The immature lifestages of this species are significant pests of turfgrass and pasturelands, where feeding causes wilting and phytotoxemia resulting in plant mortality (Byers and Wells 1966; Fagan and Kuitert 1969; Joseph and Jespersen 2021). Since its establishment in Hawai'i, two-lined spittlebug has had significant cascading effects on plant communities in rangelands, altering the composition of plant communities of grass-dominated landscapes and posing a significant threat to cattle production (Bremer *et al.* 2021). Uncovering the behavioral and metabolic strategies that insects use to exploit plants is an important step in determining their pest status. Like other cercopids, the two-lined spittlebug eats a nutritionally impoverished diet in xylem sap (Mattson Jr 1980). The processing of this diluted diet produces the characteristic spittle masses at the base of the plant. Spittlebugs have some metabolic innovations to contend with these diluted diets. Like other hemipterans that feed on phloem and xylem, spittlebugs harbor endosymbiotic bacteria that reside in specialized structures called bacteriomes. Spittlebugs harbor two symbionts in independent organs. *Sulcia muelleri* is ubiquitous in Auchenorrhyncha, whereas the other symbiont can be *Zinderia insectola* or a *Sodalis*-like microorganism (Koga *et al.* 2013; Koga and Moran 2014). These symbionts help provide complementary sets of essential amino acids through complex and intertwined metabolic pathways (Ankrah *et al.* 2020).

Management of two-lined spittlebugs in grassland ecosystems is inherently challenging. Adults are long-lived and highly fecund (Peck 1998), few commercial grass cultivars exhibit resistance and / or tolerance to this insect (Braman *et al.* 2014; Joseph and Jespersen 2021), and nymphs feed in protected areas at the base of the plant, which facilitates their escape from natural enemies (Nachappa *et al.* 2006). In the case study we use here, a genome assembly of two-lined spittlebug is a critical first step towards understanding the physiology, ecology, and evolution of this herbivorous pest and may yield novel targets to exploit for sustainable pest management.

To create HiC/HiFi genomes of the two-lined spittlebug and other insects, we developed a new HiC/HiFi genomic assembly pipeline called otb, or Only The Best [genome assembly tools]. Our pipeline reduces the time spent organizing data, installing and calibrating bioinformatic tools, and, therefore, performing analysis. otb is possibly the first nextflow HiC/HiFi genomic assembly pipeline, the the time of creation no other nextflow or snakemake pipelines were found. By implementing this pipeline, we reduced the amount of time required to produce a usable genome. The careful implementation of data management and standardization also significantly reduced team effort in genome assembly creation. otb is a software tool that utilizes the nextflow programming language (Tommaso *et al.* 2017) and is accessed using a bash script. To ensure a consistent computing environment between users, otb is implemented within a singularity container management software, which enables users to share containers with other users within the same environment. The use of nextflow provides the benefit of parallel task execution and efficient management of compute resources, while singularity ensures a consistent and reproducible compute environment. This also eliminates the need for software duplication in a high-performance computing (HPC) cluster. The development of otb was primarily for the United States Department of Agriculture, Agricultural Research Services' Ag100pest and Beenome projects, where large numbers of reference genomes needed to be created. However, otb can be used for any project that requires the automation of HiFi genome assembly; additionally, otb aims to automate the genome assembly process to a point where human involvement is necessary: HiC contig rearrangement. otb is https://github.com/molikd/otb and the documentation of otb is available in the otb wiki.

Materials and Methods

Sampling

Male and Female Two-lined spittlebugs were collected from the University of Hawai'i's Kona research station (79-7381 Hawai'i Belt Road, Holualoa, HI) where they were reared to adulthood. The samples were flash-frozen at the Hawai'i's Kona research station site on August 13, 2020. Pooled samples from the same population were used in the genome assembly.

DNA Extraction, PacBio Library Preparation, and Sequencing

High molecular weight DNA (HMW DNA) was extracted for the preparation of the PacBio HiFi library from a single adult male P. bicincta. The sample was cryoground using a Spex GenoGrinder 2010, and DNA was extracted from the ground tissue using the Qiagen MagAttract HMW DNA kit (Cat# 67563) following the kit protocol. The concentration of the extracted HMW DNA was quantified using the Qubit 1x dsDNA HS kit (Q33230), and DNA purity was assessed using UV-vis spectroscopy. The size distribution of the extracted HMW DNA was evaluated using an Agilent Femto Pulse instrument with the Genomic DNA 165kbp kit (Cat # FP-1002-0275).

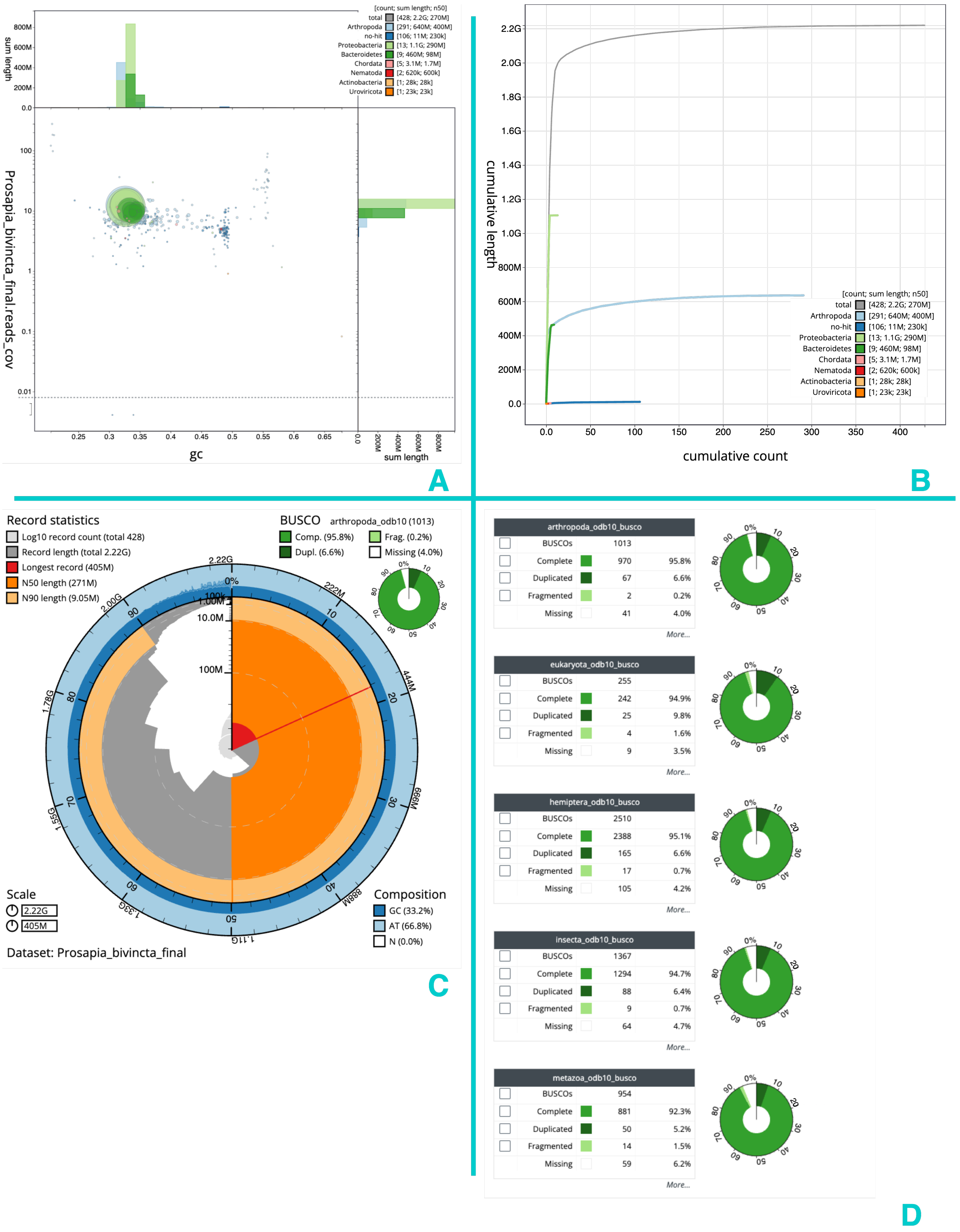
Before preparation of the PacBio HiFi library, the extracted HMW DNA was sheared using the Diagenode Megaruptor 2 with the 20 kbp shearing program, to target a sheared DNA size of approximately 10-15kbp. Sheared DNA was used to prepare PacBio HiFi libraries, using PacBio's Express Template Prep Kit 2.0 (PN: 102-088-900) following the kit protocol, with the optional nuclease digestion step after library preparation. PacBio libraries were size selected with 40% diluted AMPure PB beads (PN: 102-182-500) to remove library molecules shorter than 3kbp following PacBio's protocol. The final libraries selected for size were quantified using the Qubit 1x dsDNA HS kit, and the library size was checked using the Agilent Femto Pulse with the Genomic DNA 165kbp kit. The PacBio libraries were sequenced on a PacBio Sequel IIe using a 30 hour movie time with 2 hours of pre-extension. Sequencing reaction was prepared using the Sequel II Binding Kit 2.2. Three pooled samples were used in the assembly of the genome, sequence quality can be found in the supplement.

HiC Library Preparation

To prepare a HiC library, a single adult male two-lined spittlebug was cryoground and then fixed in freshly prepared TC fixation buffer, following the low-input crosslinking protocol for the Arima HiC 2.0 kit. Proximity ligation was performed on cross-linked samples using the Arima HiC 2.0 kit following the manufacturer's protocol. Prior to preparation of the Illumina library, the proximity-ligated DNA was sheared using a Diagenode Biorupter and then size-selected to enrich the sheared DNA in the 200-600bp range. The size distribution of the selected size sheared DNA was checked using an Agilent TapeStation with the High Sensitivity D5000 ScreenTape (Cat # 5067-5588) before proceeding to library preparation. The Illumina HiC library was prepared using the Swift Accel NGS 2S Plus DNA Library kit following the protocol outlined in the Arima HiC 2.0 kit. Library amplification was performed using the KAPA Library Amplification Kit with Primer Mix (Cat# KK2620), with 8 cycles of PCR. The final libraries were quantified using the Qubit 1x dsDNA HS kit, and the library size distribution was checked using an Agilent TapeStation. The HiC library was sequenced on an Illumina NovaSeq 6000.

otb Genome Assembly

otb was written as an automated genome assembly pipeline and run on PacBio HiFi and Illimina HiC sequences. Starting with a setup, first otb will check its environment and any set environmental variables, as well as called flags and any modes it should be running it, then otb will check if all required containers are available and working, if not, otb will download the required singularity containers and proceed to calling the nextflow run, the main body of the analysis. The sequencing data will be filtered and then assembled with HiFiASM and hicstuff (Cheng *et al.* 2021; Matthey-Doret *et al.* 2020). If the user requests it, Busco can be optionally run at this point. shhquis.jl, an in-house script is run after the initial assembly to cluster and orient the contigs, it runs on top of YaHS results to put contigs in order according to computed HiC contacts. Optionally, otb can undergo some genome assembly polishing. "polishing" is described as utilizing error corrected reads or variants to try span gaps and reduce the numeber of contigs, or in the case of DeepVariant and Merfin: select the mostly likely variant in the case where multiple Single Nucleotide Polymorphisms (SNPs) were found in the raw HiFi reads. otb provides this in three ways: Merfin, DeepVariant (Formenti *et al.* 2022b; Poplin *et al.*



	arthropoda_odb10_busco			
□	BUSCOs		1013	
□	Complete	🟩	970	95.8%
□	Duplicated	🟩	67	6.6%
□	Fragmented	🟩	2	0.2%
	Missing	□	41	4.0%

More...

	eukaryota_odb10_busco			
□	BUSCOs		255	
□	Complete	🟩	242	94.9%
□	Duplicated	🟩	25	9.8%
□	Fragmented	🟩	4	1.6%
	Missing	□	9	3.5%

More...

	hemiptera_odb10_busco			
□	BUSCOs		2510	
□	Complete	🟩	2388	95.1%
□	Duplicated	🟩	165	6.6%
□	Fragmented	🟩	17	0.7%
	Missing	□	105	4.2%

More...

	insecta_odb10_busco			
□	BUSCOs		1367	
□	Complete	🟩	1294	94.7%
□	Duplicated	🟩	88	6.4%
□	Fragmented	🟩	9	0.7%
	Missing	□	64	4.7%

More...

	metazoa_odb10_busco			
□	BUSCOs		954	
□	Complete	🟩	881	92.3%
□	Duplicated	🟩	50	5.2%
□	Fragmented	🟩	14	1.5%
	Missing	□	59	6.2%

More...

Figure 1 Blobtools reports: A: blobplot, showing contig length and GC content B: cumulative plots showing length of total contigs and thier assignment of reads to orders C: snail plot showing record statistics and Busco D: Busco plots showing complete buscos of several taxonomic classifications (while this report shows the genome assembly after contamination removal, likely misattributed contigs remain), ordered alphabetically

2018), or a "Simple". In "Simple" the error corrected reads are used to scaffold, while both Merfin and DeepVariant methods do this as well, they will also undergo variant calling in an effort to select the right variant. Busco can optionally be run at this point. The user can then optionally run Yahs, and Busco can optionally run at that point as well (Kokot *et al.* 2017; Manni *et al.* 2021a).

Genome assembly of two-lined spittlebug raw HiFi and HiC reads was completed with otb. Broadly, otb was set to use the "merfin" variant polishing option; additionally, at the k-mer creation step, KMC3 was used, and all tool-level options were kept to defaults. In its initiation, otb runs genomescope (see Supplement for GenomeScope results). Juicebox was then utilized for HiC correction, and the assembly was modified accordingly. shhquis.jl was utilized post-assembly to rearrange chromosomes for hic mapping, reducing the manual time required in the hic rearrangement step; shhquis.jl is a software tool written in the Julia programing language which clusters contigs on the HiC contig map according the the computed HiC contact map, it does not affect the genome assembly, but makes the manual hic rearrangement step slightly easier, putting contigs which are likely to be combined closer together in the HiC map. Manual hic rearrangement, the initiating map for which was created with YaHS from hicstuff exported HiC data, was used in a Juicebox manual contig rearrangement , and the final genome assembly report, was completed with Blobtools. Small contigs of less than and equal to 1,000 base pairs, and obvious contamination were removed; however, large contigs marked with one or more genes from other organisms, especially bacteria, and especially when the contig in question was beyond the size of a typical bacterial genome were kept (see Fig. 1). YaHs is a Hi-C scaffolding tool which helps create a visual mapping for use in Juicebox Zhou *et al.* (2022). shhquis.jl works on top of YaHS. otb was written for this project in Nextflow, in addition to a functionalized bash script, which pre-downloads and checks software containers of the constituent assembly tools. otb was written so that if errors occur in the pipeline, otb can be rerun from that point. otb also was written so that software versions of all tools are exported into a final reports folder (see Supplement for software versions). otb was also written with configurations for local as well as slurm and sge high performance computing cluters, and is packaged with an optional slurm template script to run the pipeline.

Results and Discussion

The goal of otb is to deliver a genome assembly as close to a polished genome as possible (i.e. reduce manual task time). otb takes several steps to reach this point (see Fig. 2, Table 1). The result is that the maximum number of assembly steps is performed, saving the user from having to perform each step individually. By including Hi-C contact map rearranging in the pipeline, and steps to reduce the number of contigs in the draft assembly, less work is needed in the contig rearrangement step (Molik 2022). Even still otb does not fully automate the contact map rearrangement and in the assembly of the two-lined spittle bug, an estimated two hours was needed to rearrange the assembly. However, since configuration of the pipeline can be used for multiple assemblies on the same compute system, once otb is setup, the amount of reconfiguration for each additional assembly is negligible.

otb was tested with two-lined spittle bug data. The genome assembly had an N50 of 270.86 megabases, a total scaffold length of 2.22 gigabases, a GC content of 33.22%, had an average scaffold length of 5.19 megabases, a total scaffold length of 2.22 gigabases, 33. 22% GC content, and had an N50 of 270.86 megabases (see Table 2 for expanded statistics). Blobtools hemiptera showed 95. 1% Buscos (see Fig. 1. Representing a high-quality contig-level assembly. The genome assembly of the two-lined spittle bug will provide a basis for further work into its interaction with its obligate hosts and, no doubt, into the control of the pest. The genome assembly will also further resources for hemipteran comparative genomics. In 2022 there were 63 species of hemiptera with genome assemblies in NCBI (Pacheco *et al.* 2022). There are a number of Hemipterans with genomes over gigabase in size. Hemipterans are also notable for their transposable elements, of an analysis of 42 arthropod species, hemipterans were found to have the greatest diversity of transposable elements Petersen *et al.* (2019).

otb has the ability to create phased genomes, which determine both chromosomes of a diploid, are a valuable resource for analyzing genetic variation within populations (Snyder *et al.* 2015). This is particularly important in agricultural research, where understanding genetic variation is crucial to understanding the evolution and spread of resistance genes that can impact insect pest outbreaks (Leftwich *et al.* 2015). The creation of a large number of phased arthropod genomes, a number of which could be enabled by more accessible and hands-off bioinformatic pipelines, has several applications (Tewhey *et al.* 2011). Having phased genomes to better understand the genetic variation of two-lined spittlebugs in an invasive population could lead to valuable insights into how the insect adapts to insecticides or adapts to new environments.

While otb was tested on the two-lined spittlebug, it was designed in principle for use in the Ag100Pest project of the USDA ARS, and represents a standard use of a HiFiASM based assembly pipeline using standard insect assembly practices, therefore it should be usable on any arthropod assembly. Hemipterans, a difficult to assemble order due to their genome size, and number transposons Pacheco *et al.* (2022); Petersen *et al.* (2019), and represents something of a worst case scenario for the pipeline. The introduction of otb, a new HiC/HiFi phased genomics assembly pipeline, has solved several data management problems that were prevalent in the creation and storage of genome assemblies. otb is written in nextflow and utilizes singularity to ensure uniformity of the computing environment. Offers parallel task execution and resource management, while also reducing the time spent organizing data, installing tools, and performing analysis. With the implementation of otb, genome creation can be automated, especially with regard to projects such as the Ag100pest.

Data availability

Code used in the creation of this genome is in the public domain per United States 17 U.S.C. § 105. The code is freely available for use and modification:

	github	DOI
otb	molikd/otb	10.5281/zenodo.6689816
Shhquis.jl	molikd/Shhquis.jl	10.5281/zenodo.6315237

Data published for this article are in the public domain per United States 17 U.S.C. § 105. The data are freely available for use and modification:

Table 1 Software Tools Utilized by otb

Software	References	Latest Title
BamTools	(Barnett *et al.* 2011)	BamTools: a C++ API and toolkit for analyzing and managing BAM files
BBTools	(Bushnell *et al.* 2017)	BBMerge – Accurate paired shotgun read merging via overlap
BCFTools	(Li 2011; Danecek *et al.* 2021)	Twelve years of SAMtools and BCFtools
bwa	(Li and Durbin 2009)	Fast and accurate short read alignment with Burrows–Wheeler transform
fcs-adaptor	(CGR 2022a,b)	Foreign Contamination Screen (FCS) tool for GenBank submissions
GFAstats	(Formenti *et al.* 2022a)	Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs
BUSCO	(Manni *et al.* 2021b,a; Seppey *et al.* 2019) (Waterhouse *et al.* 2018, 2017; Simão *et al.* 2015)	BUSCO Update: Novel and streamlined workflows along with a wider and deeper phylogenetic coverage for the scoring of eukaryotic, prokaryotic, and viral genomes
DeepVariant	(Poplin *et al.* 2018)	A universal SNP and small-indel variant caller using deep neural networks
GenomeScope2	(Ranallo-Benavidez *et al.* 2020; Vurture *et al.* 2017)	GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes
hicstuff	(Matthey-Doret *et al.* 2020, 2021)	koszullab/hicstuff: Use miniconda layer for docker and improve P(s) normalization.
HiFiAdapterFilt	(Sim *et al.* 2022)	HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly
hifiasm	(Cheng *et al.* 2021)	Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm
Jellyfish	(Marçais and Kingsford 2011)	A fast, lock-free approach for efficient parallel counting of occurrences of k-mers
KMC 3	(Kokot *et al.* 2017)	KMC 3: counting and manipulating k-mer statistics
Merfin	(Formenti *et al.* 2022b)	Merfin: improved variant filtering, assembly evaluation and polishing via k-mer validation
RagTag	(Alonge *et al.* 2021, 2019)	Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing
SAMTools	(Li 2011; Danecek *et al.* 2021)	Twelve years of SAMtools and BCFtools
Shhquis.jl	(Molik 2022)	molikd/Shhquis.jl: Inital Release
VCFTools	(Danecek *et al.* 2011)	The variant call format and VCFtools
Yahs	(Zhou *et al.* 2022)	YaHS: yet another Hi-C scaffolding tool

Table 2 Vital Statistics of *Prosapia bicincta* Assembly

Genome size	2.2 Gb
Total ungapped length	2.2 Gb
Number of scaffolds	428
Scaffold N50	270.9 Mb
Scaffold L50	4
Number of contigs	833
Contig N50	10.1 Mb
Contig L50	50
GC percent	33
Estimated Genome coverage	23.0x

	Identifier
BioProject	PRJNA987615
BioSample	SAMN35984262
Assembly	GCA_036971475.1

Acknowledgments

Special thanks to Shannon Wilson who reared the two-lined spittlebug. The McDonnell Genome Institute (MGI) provided several docker containers used by otb, especially through the works of contributors to these containers: John Garza, Thomas B. Mooney, and Alex Paul. Special thanks to Paul Stodghill of the USDA ARS in Ithaca for providing another container used by otb.

Funding

David Molik was supported through this project by the USDA Agricultural Research Service (ARS) HQ Research Associate program in Big Data. This work was carried out in part by the Tropical Pest Genetics and Molecular Biology Research Unit, ARS Project number 2040-22430-028-000D. This research used resources provided by the SCINet project of the USDA Agricultural Research Service, ARS project number 0500-00093-001-00-D.

Conflicts of interest

none declared.

Footnotes

The U.S. Department of Agriculture is an equal opportunity lender, provider, and employer.

Mention of trade names or commercial products in this report is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

Literature cited

Alonge M, Lebeigle L, Kirsche M, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S. 2021. Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing. biorxiv. doi: 10.1101/2021.11.18.469135, pre-print: not peer-reviewed.

Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, Lippman ZB, Schatz MC. 2019. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biology. 20. doi: 10.1186/s13059-019-1829-6.

Ankrah NY, Wilkes RA, Zhang FQ, Zhu D, Kaweesi T, Aristilde L, Douglas AE. 2020. Syntrophic splitting of central carbon metabolism in host cells bearing functionally different symbiotic bacteria. The ISME journal. 14:1982–1993.

Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. 2011. Bamtools: a c++ api and toolkit for analyzing and managing bam files. Bioinformatics. 27:1691–1692. doi: 10.1093/bioinformatics/btr174.

Braman S, Hanna W, Schwartz B, Nair S. 2014. Evaluation of chinese centipedegrasses and other turfgrass taxa for potential resistance to twolined spittlebug, prosapia bicincta (say). Journal of entomological science. 49:121–129.

Braman SK, Abraham CM. 1995. Twolined spittlebug. Handbook of turfgrass insect pests. Entomological Society of America, Lanham, MD. pp. 88–90.

Bremer LL, Nathan N, Trauernicht C, Pascua P, Krueger N, Jokiel J, Barton J, Daily GC. 2021. Maintaining the many societal benefits of rangelands: The case of hawai‘i. Land. 10:764.

Bushnell B, Rood J, Singer E. 2017. Bbmerge–accurate paired shotgun read merging via overlap. PloS one. 12:e0185056. doi: 10.1371/journal.pone.0185056.

Byers R, Wells HD. 1966. Phytotoxemia of coastal bermudagrass caused by the two-lined spittlebug, prosapia bicincta (homoptera: Cercopidae). Annals of the Entomological Society of America. 59:1067–1071.

CGR NCGR. 2022a. Fcs. https://github.com/ncbi/fcs.

CGR NCGR. 2022b. Foreign contamination screen (fcs) tool for genbank submissions. Technical report. NCBI Inights. Bethesda, MD. Accessed March 2023. https://ncbiinsights.ncbi.nlm.nih.gov/2022/07/28/fcs-beta-tool/.

Chang J, Stahlke AR, Chudalayandi S, Rosen BD, Childers AK, Severin AJ. 2023. polishCLR: A Nextflow Workflow for Polishing PacBio CLR Genome Assemblies. Genome Biology and Evolution. 15:evad020.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods. 18:170–175. doi: 10.1038/s41592-020-01056-5.

Childers AK, Geib SM, Sim SB, Poelchau MF, Coates BS, Simmonds TJ, Scully ED, Smith TP, Childers CP, Corpuz RL *et al*. 2021. The usda-ars ag100pest initiative: high-quality genome assemblies for agricultural pest arthropod research. Insects. 12:626.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST *et al*. 2011. The variant call format and VCFtools. Bioinformatics. 27:2156–2158. doi: 10.1093/bioinformatics/btr330.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM *et al*. 2021. Twelve years of SAMtools and BCFtools. GigaScience. 10. doi: 10.1093/gigascience/giab008.

Fagan EB, Kuitert L. 1969. Biology of the two-lined spittlebug, prosapia bicincta, on florida pastures (homoptera: Cercopidae). Florida Entomologist. pp. 199–206.

Formenti G, Abueg L, Brajuka A, Brajuka N, Gallardo-Alba C, Giani A, Fedrigo O, Jarvis ED. 2022a. Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs. Bioinformatics. 38:4214–4216.

Formenti G, Rhie A, Walenz BP, Thibaud-Nissen F, Shafin K, Koren S, Myers EW, Jarvis ED, Phillippy AM. 2022b. Merfin: improved variant filtering, assembly evaluation and polishing via k-mer validation. Nature Methods. 19:696–704. doi: 10.1038/s41592-022-01445-y.

Jiang T, Yin Z, Cai R, Yu H, Lu Q, Zhao S, Tian Y, Yan Y, Guo J, Chen X. 2021. Chromosomal-Level Genome Assembly of a True Bug, Aspongopus chinensis Dallas, 1851 (Hemiptera: Dinidoridae). Genome Biology and Evolution. 13:evab232.

Joseph SV, Jespersen D. 2021. Influence of relative humidity on the expression of twolined spittlebug (hemiptera: Cercopidae) feeding injury in turfgrass genotypes. Arthropod-Plant Interactions. 15:197–207.

Koga R, Bennett GM, Cryan JR, Moran NA. 2013. Evolutionary replacement of obligate symbionts in an ancient and diverse insect lineage. Environmental microbiology. 15:2073–2081.

Koga R, Moran NA. 2014. Swapping symbionts in spittlebugs:

evolutionary replacement of a reduced genome symbiont. The ISME journal. 8:1237–1246.

Kokot M, Długosz M, Deorowicz S. 2017. KMC 3: counting and manipulating k-mer statistics. Bioinformatics. 33:2759–2761. doi: 10.1093/bioinformatics/btx304.

Leftwich PT, Bolton M, Chapman T. 2015. Evolutionary biology and genetic techniques for insect control. Evolutionary Applications. 9:212–230. doi: 10.1111/eva.12280.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics. 27:2987–2993. doi: 10.1093/bioinformatics/btr509.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 25:1754–1760.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021a. BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Molecular Biology and Evolution. 38:4647–4654. doi: 10.1093/molbev/msab199.

Manni M, Berkeley MR, Seppey M, Zdobnov EM. 2021b. BUSCO: Assessing genomic data quality and beyond. Current Protocols. 1. doi: 10.1002/cpz1.323.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 27:764–770. doi: 10.1093/bioinformatics/btr011.

Matthey-Doret C, Baudry L, Bignaud A, Cournac A, Remi-Montagne, Guiglielmoni N, Foutel-Rodier T, Scolari VF. 2020. hicstuff: Simple library/pipeline to generate and handle hi-c data. doi: 10.5281/zenodo.4066351.

Matthey-Doret C, Baudry L, Mortaza S, Moreau P, Koszul R, Cournac A. 2021. Normalization of chromosome contact maps: Matrix balancing and visualization, In: , Springer US. pp. 1–15. doi: 10.1007/978-1-0716-1390-0_1.

Mattson Jr WJ. 1980. Herbivory in relation to plant nitrogen content. Annual review of ecology and systematics. 11:119–161.

Molik D. 2022. molikd/shhquis.jl: Inital release. doi: 10.5281/ZENODO.6315238.

Nachappa P, Guillebeau L, Braman S, All J. 2006. Susceptibility of twolined spittlebug (hemiptera: Cercopidae) life stages to entomophagous arthropods in turfgrass. Journal of economic entomology. 99:1711–1716.

Pacheco ID, Walling LL, Atkinson PW. 2022. Gene editing and genetic control of hemipteran pests: Progress, challenges and perspectives. Frontiers in Bioengineering and Biotechnology. 10.

Peck DC. 1998. Natural history of the spittlebug prosapia nr. bicincta (homoptera: Cercopidae) in association with dairy pastures of costa rica. Annals of the Entomological Society of America. 91:435–444.

Petersen M, Armisén D, Gibbs RA, Hering L, Khila A, Mayer G, Richards S, Niehuis O, Misof B. 2019. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. BMC Ecology and Evolution. 19:1–15.

Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT *et al.* 2018. A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology. 36:983–987. doi: 10.1038/nbt.4235.

Potter DA, Braman SK *et al.* 1991. Ecology and management of turfgrass insects. Annual Review of Entomology. 36:383–406.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and smudgeplot for reference-free profiling of polyploid genomes. Nature Communications. 11. doi: 10.1038/s41467-020-14998-3.

Robinson JT, Turner D, Durand NC, Thorvaldsdóttir H, Mesirov JP, Aiden EL. 2018. Juicebox.js provides a cloud-based visualization system for hi-c data. Cell Systems. 6:256–258.e1.

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: Assessing genome assembly and annotation completeness, In: , Springer New York. pp. 227–245. doi: 10.1007/978-1-4939-9173-0_14.

Sim SB, Corpuz RL, Simmonds TJ, Geib SM. 2022. Hifiadapterfilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in pacbio hifi reads and their negative impacts on genome assembly. BMC Genomics. 23:157. doi: 10.1186/s12864-022-08375-.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31:3210–3212. doi: 10.1093/bioinformatics/btv351.

Snyder MW, Adey A, Kitzman JO, Shendure J. 2015. Haplotype-resolved genome sequencing: experimental methods and applications. Nature Reviews Genetics. 16:344–358. doi: 10.1038/nrg3903.

Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ. 2011. The importance of phase information for human genomics. Nature Reviews Genetics. 12:215–223. doi: 10.1038/nrg2950.

Thorne M, Fukumoto G, Curtiss R, Hamasaki R. 2017. New spittlebug on pasture grasses in hawai'i two-lined spittlebug, prosapia bicincta. Research communication. College of Tropical Agriculture, University of Hawai'i at Manoa: Honolulu, HI, USA. Honolulu, HI.

Tommaso PD, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. 2017. Nextflow enables reproducible computational workflows. Nature Biotechnology. 35:316–319. doi: 10.1038/nbt.3820.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC. 2017. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 33:2202–2204. doi: 10.1093/bioinformatics/btx153.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, Zdobnov EM. 2017. BUSCO applications from quality assessments to gene prediction and phylogenomics. Molecular Biology and Evolution. 35:543–548. doi: 10.1093/molbev/msx319.

Waterhouse RM, Seppey M, Simão FA, Zdobnov EM. 2018. Using BUSCO to assess insect genomic resources, In: , Springer New York. pp. 59–74. doi: 10.1007/978-1-4939-8775-7_6.

Zhou C, McCarthy SA, Durbin R. 2022. YaHS: yet another Hi-C scaffolding tool. Bioinformatics. 39:btac808.

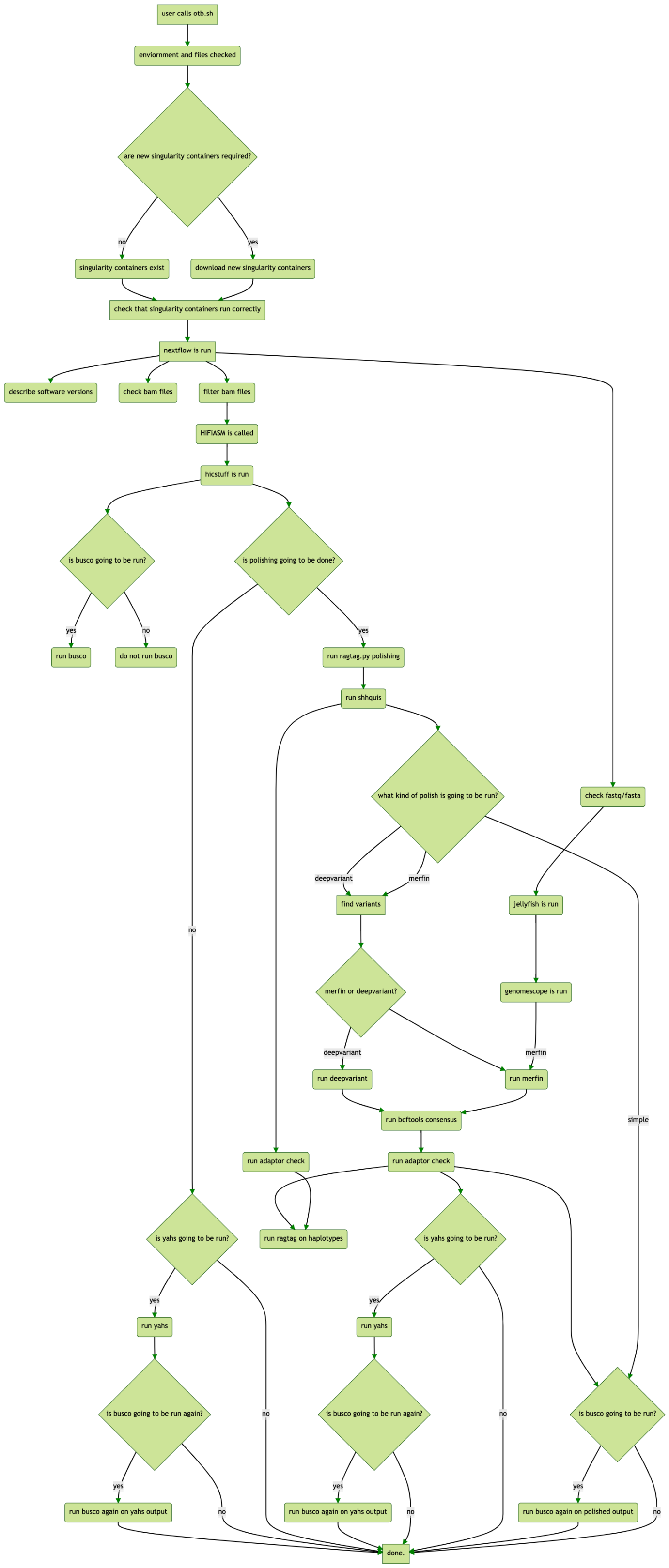


Figure 2 otb flowchart. Workflow diagram of otb showing the provess of otb running, otb.sh the entry point for otb run software and container checks, followed by the assessment of the type of hifiasm assembly to be created, otb allows also allows for multiple types of a sequence based polishing run, including a "simple" or reuse of error correct reads remapped using ragtag.py, a "deep variant" which uses deepvariant, and "merfin". Busco and sequence stats are run at multipe points in the pipeline. Yahs is run to produce HiC maps which can be utilize in JuiceBox Robinson *et al.* (2018).