G3, 2024, 14(10), jkae160
https://doi.org/10.1093/g3journal/jkae160
Advance Access Publication Date: 19 July 2024
Genome Report

Chromosome-scale genome assembly of the hunt bumble bee, *Bombus huntii* Greene, 1860, a species of agricultural interest

Jonathan Berenguer Uhuad Koch [1,*], Sheina B. Sim [2], Brian Scheffler,[3] Jeffrey D. Lozier,[4] Scott M. Geib[2]

[1]U.S. Department of Agriculture, Agricultural Research Service, Pollinating Insects Biology, Management, Systematics Research Unit, Logan, UT 84341, USA
[2]U.S. Department of Agriculture, Agricultural Research Service, U.S. Pacific Basin Agricultural Research Center, Tropical Pest Genetics and Molecular Biology Research Unit, Hilo, HI 96720, USA
[3]U.S. Department of Agriculture, Agricultural Research Service, Jamie Whitten Delta States Research Center, Genomics and Bioinformatics Research Unit, Stoneville, MS 38776, USA
[4]Department of Biological Sciences, The University of Alabama, Tuscaloosa, AL 35487, USA

*Corresponding author: U.S. Department of Agriculture, Agricultural Research Service, Pollinating Insects Biology, Management, Systematics Research Unit, Logan, UT 84341, USA. Email: jonathan.koch@usda.gov

The Hunt bumble bee, *Bombus huntii*, is a widely distributed pollinator in western North America. The species produces large colony sizes in captive rearing conditions, experiences low parasite and pathogen loads, and has been demonstrated to be an effective pollinator of tomatoes grown in controlled environment agriculture systems. These desirable traits have galvanized producer efforts to develop commercial *Bombus huntii* colonies for growers to deliver pollination services to crops. To better understand *Bombus huntii* biology and support population genetic studies and breeding decisions, we sequenced and assembled the *Bombus huntii* genome from a single haploid male. High-fidelity sequencing of the entire genome using PacBio, along with HiC sequencing, led to a comprehensive contig assembly of high continuity. This assembly was further organized into a chromosomal arrangement, successfully identifying 18 chromosomes spread across the 317.4 Mb assembly with a BUSCO score indicating 97.6% completeness. Synteny analysis demonstrates shared chromosome number ($n = 18$) with *Bombus terrestris*, a species belonging to a different subgenus, matching the expectation that presence of 18 haploid chromosomes is an ancestral trait at least between the subgenera *Pyrobombus* and *Bombus* sensu stricto. In conclusion, the assembly outcome, alongside the minimal tissue sampled destructively, showcases efficient techniques for producing a comprehensive, highly contiguous genome.

Keywords: *Bombus huntii*; pollinator; biodiversity; conservation; bombiculture; bees; genome assembly

Introduction

Bumble bees (Hymenoptera: Apidae, *Bombus* Latreille, 1802) are significant pollinators of flowering plants, resulting in at least seven species being subjected to domestication by commercial enterprises for crop pollination since the mid 1980s (Velthuis and van Doorn 2006). Bumble bees are most heavily used to pollinate tomatoes grown in controlled environment agriculture (Strange 2015). The most widely used bumble bee pollinator is arguably *Bombus terrestris* (Linnaeus, 1758), as their commercial use has expanded beyond its native range of Europe and Asia. To date, *B. terrestris* has been used to pollinate crops throughout Mesoamerica, South America, Japan, New Zealand, and Australia (Velthuis and van Doorn 2006). However, other countries, including the United States of America (USA) and Canada, have specific policies in place limiting the movement of non-native bumble bees across borders. As of 2022, three bumble bee species native to North America, *B. impatiens*, *B. huntii*, and *B. vosnesenskii*, have been made available by major bumble bee producers across specific regions of Canada, Mexico, and the USA.

The Hunt bumble bee, *B. huntii*, is native to western North America, with its geographic range spanning the countries of Canada, Mexico, and the USA (Koch *et al.* 2018) (Fig. 1). Population genetic analyses identified two to five genetic populations that correspond to past climate change and geographic variation (Koch *et al.* 2018). Relative to other North American bumble bees, *B. huntii* populations in the USA are an excellent candidate for commercial colony production due to high abundance in the wild, captive rearing success, low disease prevalence, and pollination effectiveness of tomatoes grown in controlled environment agriculture (Koch *et al.* 2015; Strange 2015; Baur *et al.* 2019; Mullins *et al.* 2020; Strange *et al.* 2023). However, unlike *B. impatiens* and *B. vosnesenskii*, there are no genomic resources for *B. huntii* (Sadd *et al.* 2015; Heraghty *et al.* 2020). Given the demonstration of genetic variation across populations, a high-quality genome resource will be important for linking genotypes to phenotypes and ecotypes that will not only be useful for understanding the evolution and ecology of this species but also may be used to improve its utility as a domesticated pollinator. For example, pollination effectiveness, the production of gynes and males, immunity to pathogens and parasites, gyne overwintering survival, and

Received on 14 May 2024; accepted on 02 July 2024
Published by Oxford University Press on behalf of The Genetics Society of America 2024.

captive-rearing success are key traits that can greatly influence the market availability, profitability, and sustainability of bumble bees used in agriculture (Thorp 2003; Velthuis and Van Doorn 2006). Thus, the study of biological traits in bumble bees could be enhanced by understanding their underlying genetic basis.

In this article, we present a near-chromosome-level assembly for *B. huntii* (iyBomHunt1.1), marking one of the initial genomes assembled and annotated under the Beenome100 initiative (http://beenome100.org). The primary goal of this consortium of scientists is to produce high-quality reference genomes representing more than 100 native bee species distributed in the USA. Employing PacBio HiFi generated data combined with a HiC library, we present an annotated genome assembly of *B. huntii* that is structured into 18 scaffolds mirroring 18 bumble bee chromosomes. The genome's quality stands out favorably compared to previously sequenced bumble bee genomes and promises to be an asset for genomic investigations concerning this bee species, which holds ecological and agricultural significance.

Methods and materials

Organism/strain origin and derivation

A male *B. huntii* specimen was used to develop the reference genome assembly. The specimen was collected by hand from a colony reared in captivity following bombiculture techniques described in Rowe *et al.* (2023). The foundress queen of the colony (mother of the male specimen, F_0) was collected in North Logan in Cache County, Utah, USA (Coordinates: 41°45′54″N 111°48′48″W, 1400 m). The male (F_1) specimen was flash frozen in liquid nitrogen and maintained at −80°C until it was shipped to the United States Department of Agriculture—Agricultural Research Service (USDA-ARS)—Pacific Basin Agricultural Research Center (PBARC) in Hilo, Hawaii, USA.

Sequencing methods and preparation details

The *B. huntii* male specimen was sent to the USDA-ARS PBARC to undergo DNA extraction and PacBio and HiC library preparation. Genomic DNA was extracted from a slice of abdominal tissue from the *B. huntii* male (ToLID iyBomHunt1). The fresh or frozen tissue protocol of the Qiagen MagAttract HMW DNA Kit (Qiagen, Hilden Germany) was followed to obtain DNA that was sufficiently of high-molecular weight for PacBio sequencing. Isolated genomic DNA was purified using 2:1 polyethylene glycol with solid-phase reversible immobilization beads (DeAngelis *et al.* 1995). The resulting DNA was quantified using a dsDNA broad range (BR) Qubit assay (Thermo Fisher Scientific, Waltham, Massachusetts, USA) and the fluorometry function of a DS-11 Spectrophotometer and Fluorometer (DeNovix Inc, Wilmington, Delaware, USA). Purity was determined using OD 260/230 and 260/280 ratios from the UV–vis spectrometer feature of the DS-11. The high-molecular-weight DNA sample was then sheared to a mean size of 20 kb with a Diagenode Megaruptor 2 (Denville, New Jersey, USA) and the subsequent size distribution was assessed with an Agilent Fragment Analyzer (Agilent Technologies, Santa Clara, California, USA) using a high sensitivity (HS) large fragment kit. The PacBio SMRTBell library was prepared using the SMRTBell Express Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, California, USA). The prepared library was bound and sequenced on a Pacific Biosciences 8 M SMRT cell on a sequel IIe system (Pacific Biosciences, Menlo Park, California, USA) at the USDA-ARS Genomics and Bioinformatics Research Unit in Stoneville, Mississippi, USA. The run was performed with a 2-h pre-extension followed by a 30-h movie collection time. After sequencing, consensus sequences from the PacBio Sequel IIe subreads were obtained using the SMRTLink v8.0 software.

A HiC library was also prepared from a slice of abdominal tissue from the same male *B. huntii* used for HiFi sequencing using the Arima HiC kit (Arima Genomics, San Diego, California, USA) from crosslinked tissue prepared following the Arima HiC low input protocol. Following proximity ligation, DNA was sheared using a Bioruptor (Diagenode, Dennville, New Jersey, USA) and DNA fragments in the range of 200–600 bp were selected as the input for Illumina library preparation using the Swift Accel NGS 2S Plus kit (Integrated DNA Technologies, Coralville, Iowa, USA). Illumina sequencing (150 bp paired-end) was performed on a NovaSeq 6000 at the Hudson Alpha Genome Sequencing Center (Huntsville, Alabama, USA), and adapter trimming after sequence collection was performed using BaseSpace software (Illumina, San Diego, California, USA).

Data analysis methods

Genome assembly methods largely follow Koch *et al.* (2023) but are briefly described below. First, HiFi reads containing artifact adapter sequences were removed using the program HiFiAdapterFilt v2.0 (Sim *et al.* 2022). The filtered HiFi reads were assembled into a contig assembly using HiFiASM v0.15.1-r329 (Cheng *et al.* 2021) using the default parameters and the output was converted to .fasta format using any2fasta (Seeman, 2018, https://github.com/tseemann/any2fasta). Scaffolding with HiC data were performed following the Arima Genomics mapping pipeline (Ghurye *et al.* 2019, https://github.com/ArimaGenomics/mapping_pipeline) and using the YaHS scaffolding software (Zhou *et al.* 2023; https://github.com/c-zhou/yahs). The Arima Genomics mapping pipeline uses BWA mem (Li 2013) to align the paired Illumina reads separately to the HiFiASM contig assembly and uses the mapping pipeline script "filter_five_end.pl" to retain reads mapped in the 5′ orientation. The individual read alignments were then processed with the "two_read_bam_combiner.pl" script to produce a sorted and quality-filtered paired-end bam file. Picard Tools "MarkDuplicates" (Picard Tools, 2019, https://broadinstitute.github.io/picard/) was used to remove PCR duplicates. The resulting .bam file and the HiFiASM contig assembly were input into YaHS for scaffolding using the "no contig error correcting" option and converted to Juicebox (Durand *et al.* 2016) compatible files using the "juicer_pre" function. Manual curation was then performed in Juicebox (v2.15) and edits were applied to the scaffold assembly using "juicebox_assembly_converter.py" from Phase Genomics (https://github.com/phasegenomics/juicebox_scripts).

To separate *Bombus* sequences from those deriving from symbionts or pathogens, we aligned contigs to taxa identified in nucleotide (NT) and protein databases using the rule "bestsumorder" of blobtoolkit v.2.6.1 (Challis *et al.* 2020). This tool assigns contigs to a taxon first based on alignments to the National Center for Biotechnology Information (NCBI) NT database and then followed by alignments to the protein database if there were no hits to the NT database. Taxonomic assignment of assembled scaffolds was tertiarily conducted using NCBI Foreign Contamination Screen (FCS, https://github.com/ncbi/fcs/wiki) tool suite using the fcs-gx function, which uses the genome cross-species aligner (GX) to identify contaminants of which there were none in the final assembly. Coverage per scaffold and contig record was calculated using minimap2 v2.2-r1101 (Li 2021).

The HiC scaffold assembly was assessed for completeness using a Benchmark of Universal Single Copy Orthologs (BUSCOs), with all relevant taxonomic databases for the genome (Eukaryota, Metazoa, Arthropoda, Insecta, and Endopterygota)

and only the most derived database, Endopterygota for the protein set. Ab initio annotations on the scaffold assembly were performed using Metaeuk v.4.a0f584d (Levy Karin *et al.* 2020) for the Eukaryota, Arthropoda, Insecta, and Endopterygota odb10 databases, and Augustus v3.4.0 (Stanke *et al.* 2008) was used to detect the Metazoa odb10 orthologs. Designation of genes as complete single copy, duplicated, fragmented, or missing were determined using BUSCO v5.2.2 (Manni *et al.* 2021) in "genome" mode for the genome assembly and "protein" for the annotated protein set. Identification for off-target (non-*B. huntii*) contigs in the assembly was performed by aligning all contigs to the NCBI NT database (accessed 2022-02-14) using the "blastn" function of BLAST + v2.5.9+ (Camacho *et al.* 2009). The contigs were secondarily aligned to the UniProt protein database (accessed 2020-03) using Diamond (Buchfink *et al.* 2021).

Coverage, taxonomic assignment, and BUSCO results were aggregated using blobtoolkit and subsequently summarized using blobblurb v2.0.1 (Sim 2022). Expected genome size was estimated using GenomeScope v2.0 (Ranallo-Benavidez *et al.* 2020), which uses k-mer frequency analysis of k-mer counts performed by KMC v3.2.1 (Kokot *et al.* 2017). The level of duplicate artifacts in the assembly was assessed using BUSCO results for both the genome and the protein set and using k-mer abundance in the raw HiFi reads relative to their representation in the final assembly as determined by k-mer Analysis Toolkit v2.4.2 (KAT; Mapleson *et al.* 2017).

The *B. huntii* genome was submitted to the NCBI RefSeq (Rajput *et al.* 2019) for annotation using the NCBI Eukaryotic Genomic Annotation Pipeline v10.0. This method was selected to provide consistency and standardization in annotation methods with other bumble bee species (Sadd *et al.* 2015; Heraghty *et al.* 2020; Koch *et al.* 2023) and with the ongoing USDA-ARS sequencing initiatives AgPest100 and Beenome100. The annotation pipeline utilized > 6 billion RNA sequencing reads from thoracic muscle and brain tissue of other closely related North American species in the subgenus *Pyrobombus* (i.e. *B. vosnesenskii*, *B. vancouverensis*, *B. bifarius*, and *B. impatiens*) available on GenBank. A description of the RNA sequencing reads is available on the NCBI *Bombus huntii* annotation release 100 report (https://www.ncbi.nlm.nih.gov/refseq/annotation_euk/Bombus_huntii/100/).

We assigned the annotated genes from *B. huntii* to orthologous groups along with publicly available gene annotations from *Apis mellifera* (NCBI GenBank accession: GCF_003254395.2), *B.* (*Bombus*) *terrestris* (NCBI RefSeq accession: GCF_910591885.1 and ToLID: iyBomTerr1), and *B.* (*Pyrobombus*) *hypnorum* Linnaeus, 1758 (NCBI GenBank and Ensembl accession GCA_911387925.2 and ToLID: iyBomHypn1 (Crowley and Sivell 2023)] using OrthoFinder (Emms and Kelly 2019) with default parameters. We visualized the results from OrthoFinder using a riparian plot produced by GENESPACE (Lovell *et al.* 2022) to depict chromosomal distribution patterns, synteny, and identify potential regions of the genome that are associated with genomic gaps, inversions, and translocations. Finally, we depicted shared sets of genes between taxa in an upset plot using UpSetR (Lex *et al.* 2014; Conway *et al.* 2017).

Results and discussion

Genome assembly

PacBio Sequel II HiFi and HiC sequencing of *B. huntii* produced sufficient data for a highly contiguous contig assembly with chromosomal resolution. In the preliminary contig assembly, 55.2% (sum of reads = 183 M) of the contigs did not match any taxonomic group (Fig. 2a), potentially signaling a significant number of contaminants, which is not unexpected for abdominal tissue. The other contigs had BLAST hits to Arthropoda (37.2%), Streptophyta (7%), Nematoda (1%), and Ciliophora (<1%; Fig. 2a,b). Additional rounds of filtering with blobtools and alignment to HiC sequences, including the removal of "No Hit" sequences resulted in the removal of most non-Arthropod contaminants in the remaining scaffolds (Fig. 2c,d). The 29 of the remaining 31 contigs taxonomically assigned as "non-Arthropod" collectively have 559 annotated genes and indicates spurious assignment as non-Arthropod.

Fig. 1. Lateral view of *B. huntii* gyne. Image by Joseph S. Wilson, with permission.

The initial *B. huntii* assembly resulted in 225 scaffolds with a scaffold N50 of 17.5 Mb (Table 1; Fig. 3a). The assembly exceeds the minimum reference standard of 6.C.Q40 (>1.0 Mb contig and 10.0 Mb scaffold N50) identified by the Earth BioGenome Project (Lewin *et al.* 2018; Lawniczak *et al.* 2022). HiC contact mapping was able to recover 18 chromosome-length scaffolds (Fig. 3b). The size of the scaffolded *B. huntii* assembly is 317.4 Mb (L50 = 8). Total haploid assembly size is estimated to be 288.73 Mb based on k-mer analysis with GenomeScope v2 (k-mer = 21, k-cov = 83; Fig. 3c). The assembly is hypothesized to have minimal error (=0.223%), likely due to low sequencing error and low repetitive content of sequences (=1.04%). The assembly is larger and more intact than the most recent *B. impatiens* Cresson, 1863 genome assembly (version BIMP_2.2) and other published genomes of closely related North American species in the subg. *Pyrobombus* (*B. bifarius* Cresson, 1878 = 266.8 Mb, *B. vancouverensis* Cresson, 1878 = 282.1 Mb, and *B. vosnesenskii* Radoszkowski, 1862 = 275.6 Mb; Fig. 3b), demonstrating the quality of the assembly compared to other *Bombus* genomes assembled without the advantages of PacBio and HiC data (Table 1). GC content of the *B. huntii* assembly is comparable to the species assemblies included in our study at 37% (average = 37.8% +/− 0.20%; Fig. 3b; Table 1).

BUSCO scores also indicated a highly complete genome assembly (Fig. 3b; Table 1). The *B. huntii* assembly performed similarly to other published assemblies in the subg. *Pyrobombus*, with 97.6% of the 5,991-benchmarking universal single-copy orthologs represented in OrthoDB v1.10 Hymenoptera lineage dataset (hymenoptera_odb10). Most of the genes were single copy (97.3%) with 0.27% duplicated, 0.45% fragmented, or 2.42% missing (Fig. 3b; Table 1). Results of the BUSCO analysis of the genome annotation are presented in Table 1.

Genome annotation and synteny

In total, the NCBI Eukaryotic Genomic Annotation Pipeline predicted 15,072 genes and pseudogenes, of which 14,643 genes

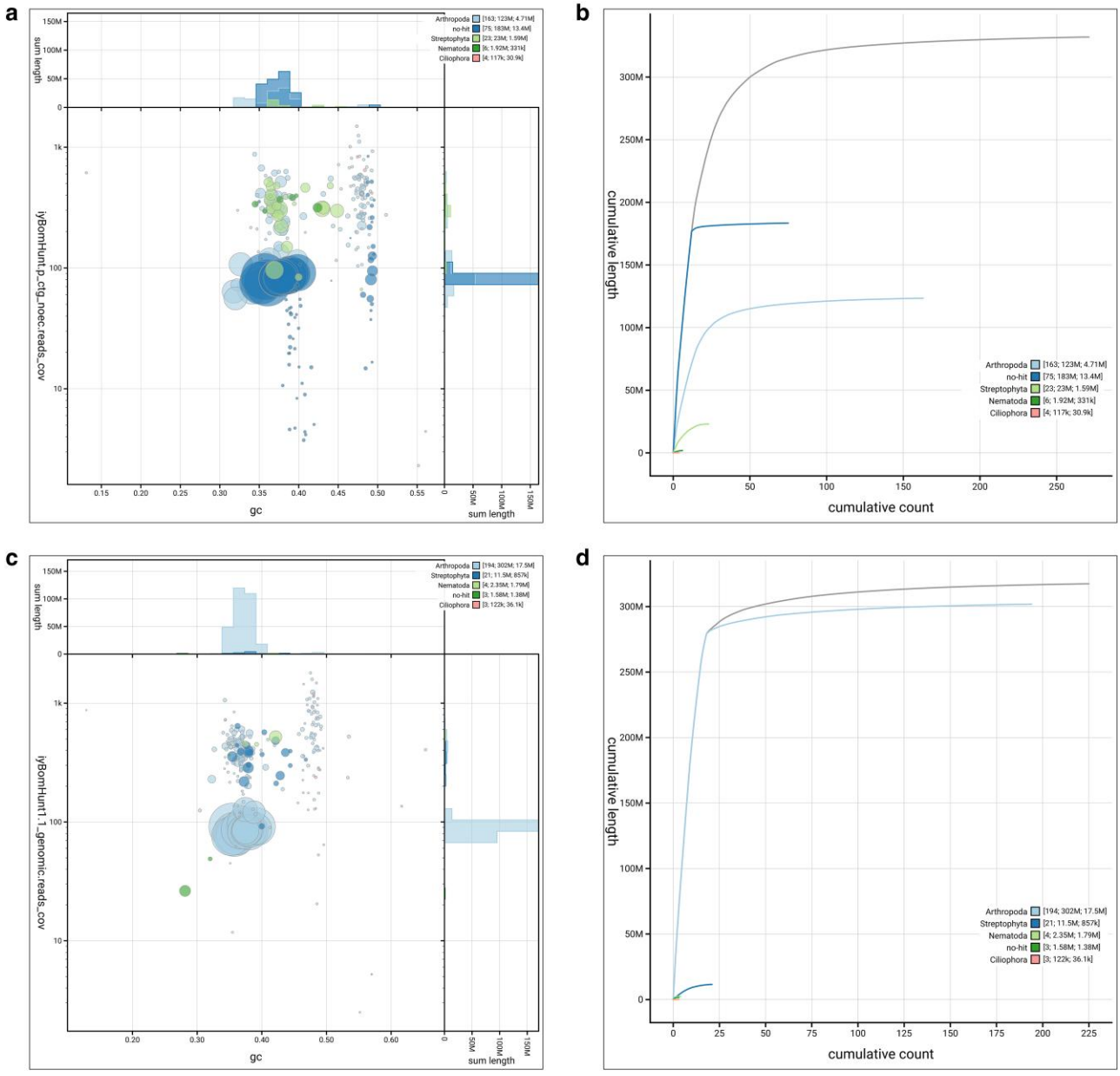

Fig. 2. Blob plots from PacBio data showing read depth of coverage, GC content, and size of a) contigs, b) cumulative count and length of contigs per category, c) scaffolds, and d) cumulative count and length of scaffolds per category.

Table 1. Assembly statistics and BUSCO analysis for *B. huntii* genome assembly in comparison to other *Bombus* genome assemblies.

Species	Assembly version	Len. (Mb)	N50 (Mb)	L50	No. scaffolds	GC %	Complete [single, duplicated]	Frag.	Miss.
B. huntii[a]	iyBomHunt1.1	317.4	17.5	8	225	37	97.6% [97.3%, 0.3%]	0.45%	2.42%
B. bifarius	Bbif_JDL3187	266.8	2.2	30	1,249	37.96	98.1% [97.7%, 0.4%]	0.60%	1.30%
B. vancouverensis	Bvanc_JDL1245	282.1	3.06	24	1,162	38.02	98.4% [97.9%, 0.5%]	0.60%	1.00%
B. vosnesenskii	Bvos_JDL3184-5_v1.1	275.6	2.83	24	1,429	37.93	98.2% [98.0%, 0.2%]	0.60%	1.20%
B. impatiens	BIMP_2.2	242	1.41	54	5,460	38	99.0% [98.8%, 0.2%]	0.30%	0.70%
B. hypnorum	iyBomHypn1.2	297.3	24.3	6	52	37.5	97.6% [97.4%, 0.2%]	0.5%	1.9%
B. terrestris	iyBomTerr1.2	393	14.6	10	249	38.5	98.9% [98.5%, 0.4%]	0.10%	1.00%

[a] *B. huntii* BUSCO analysis of gene annotation: Complete = 98.7% [single = 98.4%, duplicated = 0.3%]; Fragmented = 0.2%; Missing = 1.2%; $n = 5{,}991$. For NCBI *Bombus huntii* annotation release 100 report: https://www.ncbi.nlm.nih.gov/refseq/annotation_euk/Bombus_huntii/100/.

gave rise to 34,820 transcripts. The number of predicted genes in the *B. huntii* genome is greater than the number of genes predicted in *B. impatiens* and the *B. huntii* sister species *B. vosnesenskii* by 14.5 to 11.43%, although the number of predicted protein coding genes (11,088) is similar to other species (Table 2). Much of the difference in gene numbers could be attributed to the annotation of 3,554

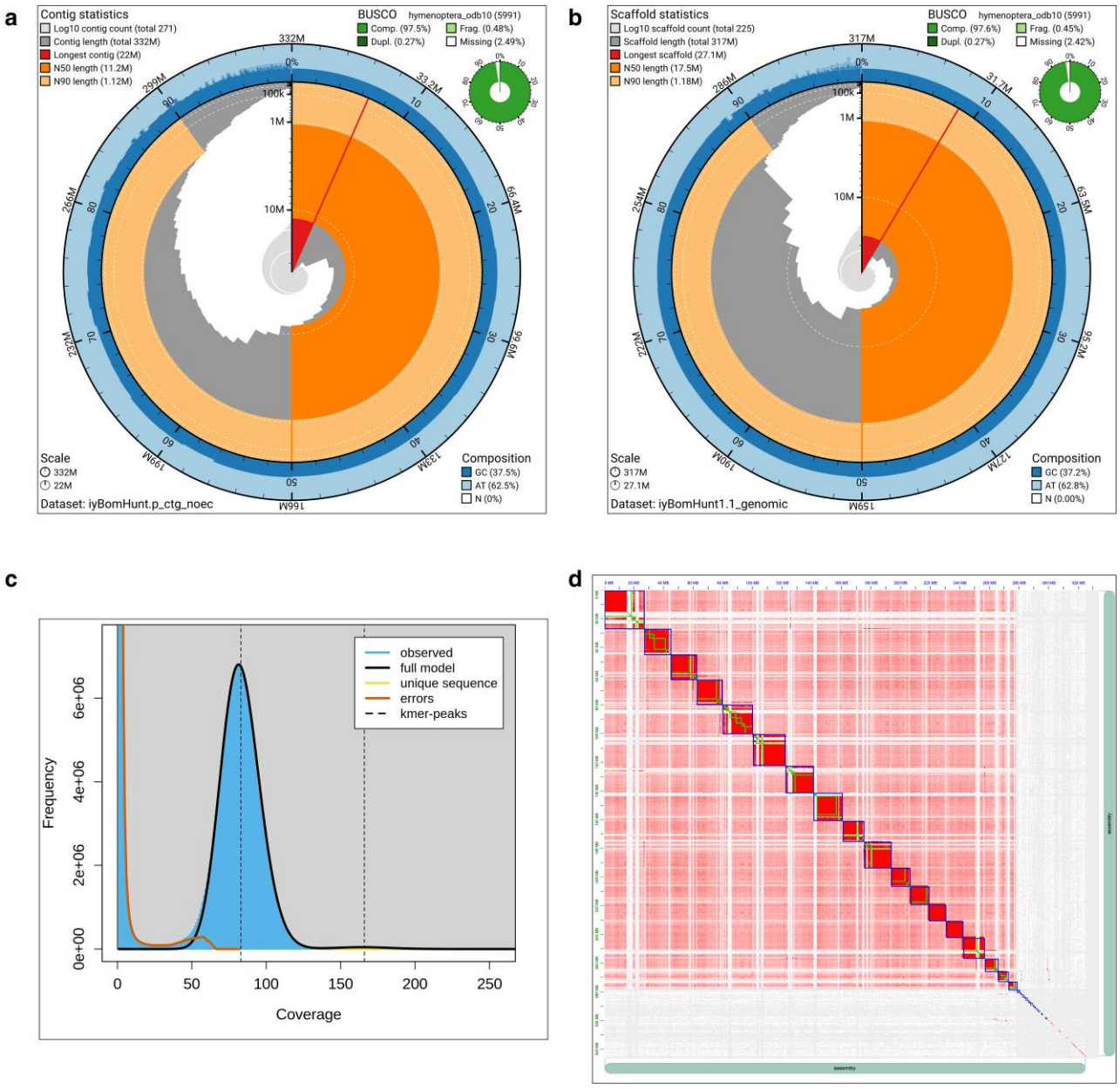

Fig. 3. Snail plot visualization and genome assembly statistics for *B. huntii* with a) contig and b) scaffold resolution, c) haploid assembly size estimated with GenomeScope v., and d) HiC contact map of *B. huntii* genome assembly. The HiC contact map demonstrates chromosomes ordered by size from left to right and top to bottom. The snail plot (contig and scaffold) is divided into one thousand size-ordered bins around the circumference. Each bin represents 0.1% of the assembly (contig = 332 Mb; scaffold = 317 Mb). The distribution of the scaffold lengths are shown in gray. The plot radius is scaled to the longest scaffold present in the assembly (contig = 22 Mb; scaffold = 27.1 Mb, both shown in red). The orange and pale-orange arcs show the N50 and N90 lengths (contig N50 = 11.2 Mb, contig N90 = 1.12 Mb; scaffold N50 = 17.5 Mb, scaffold N90 = 1.18 Mb). The blue and pale-blue ring around the outside of the plot is the distribution of GC, AT, and N percentages in the same bins as the inner plot. A summary of the BUSCO genes in the hymenoptera_odb10 dataset is shown in the top right of both contig and scaffold assemblies.

noncoding genes in *B. huntii*, whereas an average of 1,884 noncoding genes were detected in the genome assemblies of *B. bifarius*, *B. vancouverensis*, *B. vosnesenskii*, and *B. impatiens*. The elevated number of noncoding genes detected in the assembly may be due to the quality of the PacBio HiFi sequencing technology. Increased detection of noncoding genes in a bumble bee genome was also observed in the *B. affinis* genome assembly which used the same sequencing platform and bioinformatics methods described in this study (Koch *et al.* 2023). The verification, importance, and function of these noncoding genes in *Bombus* would benefit from further research.

We analyzed the synteny of *B. huntii* (*B. Pyrobombus*) with *B. terrestris* (*B. Bombus*), and *B. hypnorum* (*B. Pyrobombus*) as these species are represented by chromosomal resolution assemblies on NCBI. We also included *A. mellifera* in the synteny analysis as an outgroup for comparative purposes. The subgenus *Pyrobombus* (i.e. *B. huntii* and *B. hypnorum*) diverged from the subgenera *Bombus* (i.e. *B. terrestris*) and *Alpinobombus* around 19 mya during the Miocene (Hines 2008). Furthermore, *B. hypnorum* diverged from the that clade includes *B. huntii* around 15 mya, also during the Miocene (Hines 2008). Synteny analysis demonstrates that *B. huntii* and *B. terrestris* share the same number of linkage groups,

Table 2. Annotation statistics from the NCBI eukaryotic genome annotation pipeline for *B. huntii* and other *Bombus* genome assemblies available on NCBI.

	B. huntii	*B. bifarius*	*B. vancouverensis*	*B. vosnesenskii*	*Impatiens*	*B. hypnorum*	*B. terrestris*
Genes and pseudogenes	15,072	13,325	13,687	13,527	13,161	13,622	13,398
Protein-coding	11,088	11,148	11,338	11,184	10,632		10,310
Noncoding	3,554	1,653	1,802	1,789	2,293		2,939
Nontranscrib. pseudogenes	429	524	547	554	236		148
Genes with variants	6,278	5,083	5,162	5,166	5,453		6,016
Other	1	0	0	0	0		1
mRNAs	27,776	23,896	24,285	24,067	24,471		25,755
Fully supported	26,590	22,904	23,144	23,079	23,904		25,165
With > 5% ab initio	621	674	779	643	299		277
Partial	67	195	208	141	218		46
With filled gap(s)	0	0	0	0	29		1
Known RefSeq (NM_)	0	0	0	0	4		25
Model RefSeq (XM_)	27,776	23,896	24,285	24,067	24,467		25,730
Noncoding RNAs	7,044	2,731	2,964	2,974	3,542		5,477
Fully supported	6,124	2,401	2,584	2,647	3,261		5,117
With > 5% ab initio	0	0	0	0	0		0
Partial	2	0	1	0	0		2
Model RefSeq (XR_)	6,840	2,525	2,760	2,774	3,339		5,271
CDSs	27,776	23,896	24,285	24,067	24,471		25,755
Fully supported	26,590	22,904	23,144	23,079	23,904		25,165
With > 5% ab initio	723	727	833	685	340		311
Partial	67	195	208	141	214		43
With major correction(s)	68	122	118	173	188		77
Known RefSeq (NP_)	0	0	0	0	4		25
Model RefSeq (XP_)	27,776	23,896	24,285	24,067	24,467		25,730

The *B. hypnorum* assembly was annotated with ensemble rapid annotation pipeline (Crowley and Sivell 2023). Thus, the NCBI annotation statistics are not available for a comparative assessment with other focal *Bombus*.

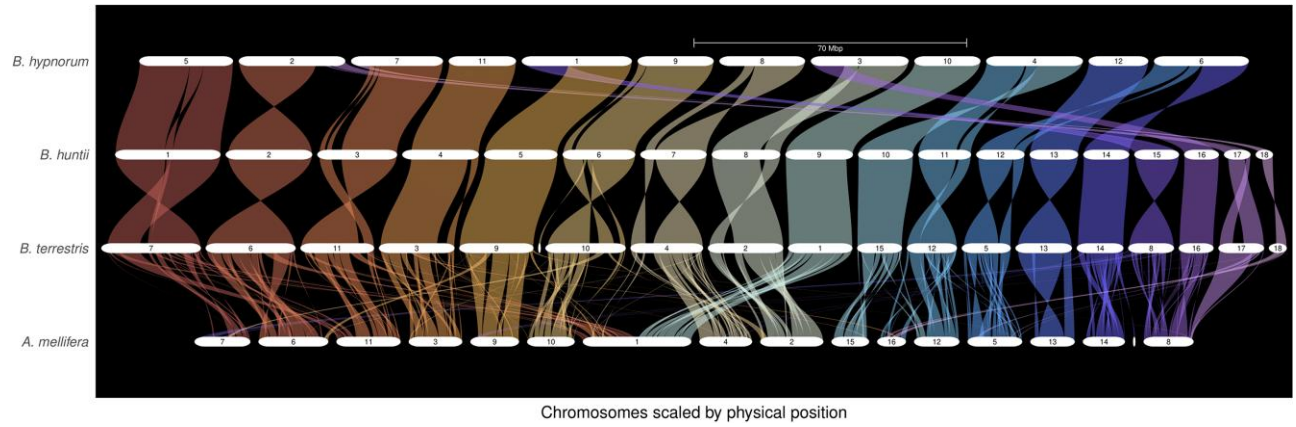

Fig. 4. A GENESPACE-generated synteny map (bottom to top) of *A. mellifera* (outgroup), *B. terrestris*, *B. huntii* (study taxon), and *B. hypnoroum*. The white horizontal segments represent chromosomes. The colors (red to purple) represent the orthologous *B. huntii* chromosomes (1–18) and braids represent the syntenic blocks between bee genomes. X-axis positions are scaled by physical position.

hereafter identified as "chromosomes" ($n = 18$; Fig. 4). However, *B. hypnorum*, while in the same subgenus as *B. huntii*, has a reduced number of chromosomes ($n = 12$; Fig. 4). A chromosome number of 12 have also been reported in *B. perplexus* Cresson, 1,863 through karyotyping (Owen *et al.* 1995). *B. perplexus* belongs to the same clade as *B. hypnorum* and is a sister taxon to *B. hypnorum* (Cameron *et al.* 2007; Hines 2008). Consistent with the hypothesis for *Bombus* as a genus (Owen *et al.* 1995), a haploid number of 18 chromosomes appear to represent the ancestral state shared between at least the subgenera *Bombus* and *Pyrobombus*, with 12 chromosomes to be a derived trait found in clade that includes *B. hypnorum* and *B. perplexus*. However, further sampling of taxa across *Pyrobombus* and *Bombus* will be required to confirm this hypothesis and investigate the potential timing and importance of this reduction in chromosome number in some *Pyrobombus*.

Visualization of chromosome arrangements across *B. huntii*, *B. terrestris*, and *B. hypnorum* demonstrate several rearrangements including inversions, gaps, and repeats (Fig. 4). Of course, the structural rearrangements could indicate assembly artifacts or genuine changes that occurred within the past 20 million years since the divergence from the common ancestors of *B. terrestris* with *B. huntii* and *B. hypnorum*, as well as between *B. huntii* and *B. hypnorum* (~15 mya). Interestingly, *B. hypnorum* does not appear to have lost major chromosomal regions with the reduction in chromosome number, with most *B. huntii* regions represented, indicating chromosome fusion events as the mechanism of this reduction (Owen *et al.* 1995). Finally, we compared the number of orthogroups shared across *B. huntii*, *B. hypnorum*, *B. terrestris*, and *A. mellifera*. We determined that 86% ($n = 9,810$) of the orthogroups identified through NCBI annotation were shared across the four

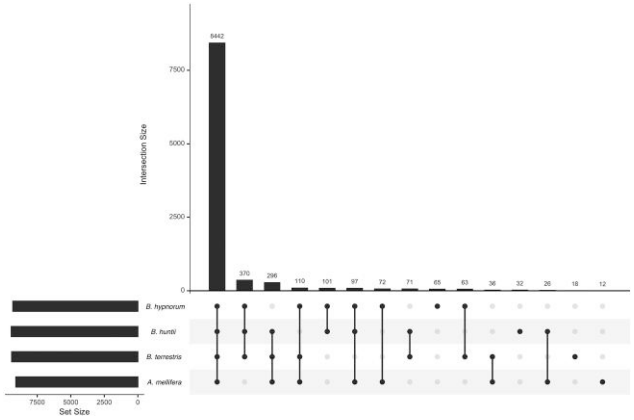

Fig. 5. Orthogroup barplot of annotated genes for *B. huntii*, *B. hypnorum*, *B. terrestris*, and *A. mellifera*. Size of bar represents the number of orthogroups identified for each taxon, shared between taxa, or unique to a taxon (i.e. species-specific orthogroup).

focal species with 7,933 orthogroups found in single-copy and the remaining 509 orthogroups found in multiple copy in one or more of the four species. Furthermore, we identified 32 genes unique to *B. huntii* and 101 orthogroups shared between *B. huntii* and *B. hypnorum* (Fig. 5).

As the market for mass produced bumble bees grows in tandem with market needs, their domestication can be improved by genetic data and genomic resources. This is especially true for regulatory agencies and policy makers who make decisions on the importation of bumble bees across state, provincial, and national borders. Many of these decisions are driven in part by concerns on the movement of pathogens across bee species, hybridization, and increased interactions amongst native and invasive species. Even within species, there is concern about the use and escape of regional genotypes that could contribute to outbreeding depression among native locally adapted *Bombus* populations (Lozier *et al.* 2015). Furthermore, while *B. huntii* populations in the USA appear to be abundant and genetically diverse (Koch *et al.* 2015, 2018), populations in Mesoamerica are projected to be vulnerable to climate variation that is expected to occur over the next 30 years and beyond (Martínez-López *et al.* 2021). Identifying genetic loci under selection in the context of environmental variability for *B. huntii* will be useful in managing the commercial populations for pollination in diverse agricultural systems (Jackson *et al.* 2020).

Data availability

The *B. huntii* genome assembly has NCBI Bioproject ID #PRJNA858880 (ToLID: iyBomHunt1) in the Beenome100 umbrella BioProject #PRJNA923301. GenBank accession is GCA_024542735.1 and RefSeq accession is GCF_024542735.1. PacBio HiFi reads are available on the NCBI Sequence Read Archive (SRA) at accession SRX16249823. Illumina sequences used for HiC scaffolding are available on SRA at SRX16249961. Scripts to run the pipeline are on Ag Data Commons (https://doi.org/10.15482/USDA.ADC/25762431.v1; Sim 2024).

Acknowledgments

The genome assembly was generated as part of the USDA-ARS Beenome100 Initiative (https://www.beenome100.org/). This research used resources provided by the SCINet project of the USDA-ARS project number 0500-00093-001-00-D, Tropical Pest Genetics and Molecular Biology project number 2040-22430-028-000-D, and Pollinating Insect Research Unit project number 2080-21000-019-000-D. The authors thank the members of the USDA-ARS Beenome100 and Ag100Pest Team for sequencing and analysis support. We also thank Tien Lindsay for her support in bombiculture and specimen curation. All opinions expressed in this article are the authors and do not necessarily reflect the policies and views of USDA. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA. The USDA is an equal opportunity provider and employer.

Funding

This work was supported by grants from the National Science Foundation to JDL (DEB-2126418) and JBUK (2126417).

Conflicts of interests

The authors declare no conflicts of interest.

Literature cited

Baur A, Strange JP, Koch JB. 2019. Foraging economics of the hunt bumble bee, a viable pollinator for commercial agriculture. Environ Entomol. 48(4):799–806. doi:10.1093/ee/nvz075.

Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 18(4): 366–368. doi:10.1038/s41592-021-01101-x.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10(1):421. doi:10.1186/1471-2105-10-421.

Cameron SA, Hines HM, Williams PH. 2007. A comprehensive phylogeny of the bumble bees (*Bombus*). Biol J Linn Soc Lond. 91(1): 161–188. doi:10.1111/j.1095-8312.2007.00784.x.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. 2020. BlobToolKit—interactive quality assessment of genome assemblies. G3 (Bethesda). 10(4):1361–1374. doi:10.1534/g3.119.400908.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 18(2):170–175. doi:10.1038/s41592-020-01056-5.

Conway JR, Lex A, Gehlenborg N. 2017. Upsetr: an R package for the visualization of intersecting sets and their properties. Bioinformatics. 33(18):2938–2940. doi:10.1093/bioinformatics/btx364.

Crowley L, Sivell O. 2023. The genome sequence of the Tree Bumblebee, Bombus hypnorum (Linnaeus, 1758). Wellcome Open Research. 8(21):21.

DeAngelis MM, Wang DG, Hawkins TL. 1995. Solid-phase reversible immobilization for the isolation of PCR products. Nucleic Acids Res. 23(22):4742–4743. doi:10.1093/nar/23.22.4742.

Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. 2016. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 3(1):99–101. doi:10.1016/j.cels.2015.07.012.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20(1):238. doi:10.1186/s13059-019-1832-y.

Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. 2019. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLOS Comput Biol. 15(8): e1007273. doi:10.1371/journal.pcbi.1007273.

Heraghty SD, Sutton JM, Pimsler ML, Fierst JL, Strange JP, Lozier JD. 2020. De Novo genome assemblies for three North American bumble bee species: *bombus bifarius, Bombus vancouverensis*, and *Bombus vosnesenskii*. G3 (Bethesda). 10(8):2585–2592. doi:10.1534/g3.120.401437.

Hines HM. 2008. Historical biogeography, divergence times, and diversification patterns of bumble bees (Hymenoptera: Apidae: *Bombus*). Syst Biol. 57(1):58–75. doi:10.1080/10635150801898912.

https://broadinstitute.github.io/picard/2019 Picard Toolkit. Broad Institute.

Jackson JM, Pimsler ML, Oyen KJ, Strange JP, Dillon ME, Lozier JD. 2020. Local adaptation across a complex bioclimatic landscape in two montane bumble bee species. Mol Ecol. 29(5):920–939. doi:10.1111/mec.15376.

Koch JB, Lozier J, Strange JP, Ikerd H, Griswold T, Cordes N, Solter L, Stewart I, Cameron SA. 2015. USBombus, a database of contemporary survey data for North American bumble bees (Hymenoptera, Apidae, *Bombus*) distributed in the United States. Biodivers Data J. 30(3):e6833. doi:10.3897/BDJ.3.e6833.

Koch JBU, Sim SB, Scheffler B, Geib SM, Smith TA. 2023. Chromosome-scale genome assembly of the rusty patched bumble bee, *Bombus affinis* (cresson) (Hymenoptera: Apidae), an endangered North American pollinator. G3 (Bethesda). 13(8):jkad119. doi:10.1093/g3journal/jkad119.

Koch JB, Vandame R, Mérida-Rivas J, Sagot P, Strange J. 2018. Quaternary climate instability is correlated with patterns of population genetic variability in *Bombus huntii*. Ecol Evol. 8(16):7849–7864. doi:10.1002/ece3.4294.

Kokot M, Dlugosz M, Deorowicz S. 2017. KMC 3: counting and manipulating k-mer statistics. Bioinformatics. 33(17):2759–2761. doi:10.1093/bioinformatics/btx304.

Lawniczak MKN, Durbin R, Flicek P, Lindblad-Toh K, Wei X, Archibald JM, Baker WJ, Belov K, Blaxter ML, Marques Bonet T, *et al.* 2022. Standards recommendations for the earth BioGenome project. Proc Natl Acad Sci U S A. 119(4):e2115639118. doi:10.1073/pnas.2115639118.

Levy Karin E, Mirdita M, Söding J. 2020. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. Microbiome. 8(1):48. doi:10.1186/s40168-020-00808-x.

Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, Durbin R, Edwards SV, Forest F, Gilbert MTP, *et al.* 2018. Earth BioGenome project: sequencing life for the future of life. Proc Natl Acad Sci U S A. 115(17):4325–4333. doi:10.1073/pnas.1720115115.

Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. 2014. Upset: visualization of intersecting sets. IEEE Trans Vis Comp Graph. 20(12):1983–1992. doi:10.1109/TVCG.2014.2346248.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN]. https://doi.org/10.48550/arXiv.1303.3997, preprint: not peer reviewed.

Li H. 2021. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 37(23):4572–4574. doi:10.1093/bioinformatics/btab705.

Lovell JT, Sreedasyam A, Schranz ME, Wilson M, Carlson JW, Harkess A, Emms D, Goodstein DM, Schmutz J. 2022. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. elife. 11:e78526. doi:10.7554/eLife.78526.

Lozier JD, Cameron SA, Duennes MA, Strange JP, Williams PH, Goulson D, Brown MJF, Morales C, Jepsen S. 2015. Relocation risky for bumblebee colonies. Science. 350(6258):286–287. doi:10.1126/science.350.6258.286-b.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol 38(10):4647–4654. doi:10.1093/molbev/msab199.

Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. 2017. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics. 33(4):574–576. doi:10.1093/bioinformatics/btw663.

Martínez-López O, Koch JB, Martínez-Morales MA, Navarrete-Gutiérrez D, Enríquez E, Vandame R. 2021. Reduction in the potential distribution of bumble bees (Apidae: *Bombus*) in mesoamerica under different climate change scenarios: conservation implications. Glob Chang Biol. 27(9):1772–1787. doi:10.1111/gcb.15559.

Mullins JL, Strange JP, Tripodi AD. 2020. Why are queens broodless? Failed nest initiation not linked to parasites, mating status, or ovary development in two bumble bee species of *Pyrobombus* (Hymenoptera: Apidae: *Bombus*). J Econ Entomol. 113(2):575–581. doi:10.1093/jee/toz330.

Owen RE, Richards KW, Wilkes A. 1995. Chromosome numbers and karyotypic variation in bumble bees (Hymenoptera: Apidae. Bombini). J Kans Entomol. Soc. 68(3):290–302.

Rajput B, Pruitt KD, Murphy TD. 2019. RefSeq curation and annotation of stop codon recoding in vertebrates. Nucleic Acids Res. 47(2):594–606. doi:10.1093/nar/gky1234.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 11(1):1432. doi:10.1038/s41467-020-14998-3.

Rowe G, Hagadorn MA, Lindsay T-TT, Malfi R, Williams NM, Strange JP. 2023. Chapter 20—production of bumblebees (Hymenoptera: Apidae) for pollination and research. In: Morales-Ramos JA, Rojas MG, Shapiro-Ilan DI, editors. Mass Production of Beneficial Organisms. 2nd ed. London: Academic Press. p. 559–579.

Sadd BM, Barribeau SM, Bloch G, de Graaf DC, Dearden P, Elsik CG, Gadau J, Grimmelikhuijzen CJP, Hasselmann M, Lozier JD, *et al.* 2015. The genomes of two key bumblebee species with primitive eusocial organization. Genome Biol. 16(1):76. doi:10.1186/s13059-015-0623-3.

Seeman T. any2fasta. 2018. [accessed 2021 May 1]. https://github.com/tseemann/any2fasta

Sim SB. blobblurb. 2022. [accessed 2024 Feb 2]. https://github.com/sheinasim/blobblurb02/01/2024

Sim SB. Bombus huntii genome assembly scripts. Ag Data Commons. 2024. [accessed 2024 Feb 2]. doi:10.15482/USDA.ADC/25762431.v1

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(1):157. doi:10.1186/s12864-022-08375-1.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 24(5):637–644. doi:10.1093/bioinformatics/btn013.

Strange JP. 2015. *Bombus huntii, Bombus impatiens*, and *Bombus vosnesenskii* (Hymenoptera: Apidae) pollinate greenhouse-grown tomatoes in western North America. J Econ Entomol. 108(3):873–879. doi:10.1093/jee/tov078.

Strange JP, Tripodi AD, Lindsay T-TT, Herndon JD, Knoblett J, Christman ME, Barkan NP, Koch JBU. 2023. Variation in North American bumble bee nest success and colony sizes under captive rearing conditions. J. Insect Sci. 23(3):10. doi:10.1093/jisesa/iead032.

Thorp RW. 2003. Bumble bees (Hymenoptera: Apidae): commercial use and environmental concerns. In: Strickler K, Cane J, editors. For Nonnative Crops, Whence Pollinators of the Future?. Lanham [MD]: Thomas Say Publications in Entomology. p. 23.

Velthuis HHW, Van Doorn A. 2006. A century of advances in bumblebee domestication and the economic and environmental aspects of its commercialization for pollination. Apidologie (Celle). 37(4): 421–451. doi:10.1051/apido.

Zhou C, McCarthy SA, Durbin R. 2023. YaHS: yet another Hi-C scaffolding tool. Bioinformatics. 39(1):btac808

Editor: K. Vogel