Asian Giant Hornet (*Vespa mandarinia*) Genome Project Methods Description

Sample collection:

A specimen of a large hornet of unknown taxonomy was turned in to local authorities on Vancouver Island, in British Columbia, in May of 2019. It was subsequently identified, by Canadian and Japanese entomologists, as well as DNA fingerprinting, to be *V. mandarinia*. Local beekeepers in Nanaimo, BC, managed to track down the nest, located at 49.152809 N, 123.943125 W. The beekeepers visited the nest later that evening and subdued all the insects with fire retardant (CO2), to both asphyxiate them and stun them with lower temperatures. The nest was then immediately dug out of the ground and all insects were transferred quickly into 70% ethanol. Once dead, the insects were removed from the alcohol and stored at -20˚C until they could be transferred to the University of British Columbia, where they were subsequently stored at -80˚C. The specimen used for sequencing was stored at -20˚C for approximately 2 weeks.

DNA extraction & Library preparation:

High molecular weight (HMW) DNA extraction was performed using the Qiagen MagAttract HMW DNA Kit. The tissue sample of the specimen was shipped on dry ice and stored at -80° C prior to extraction. To process the sample, it was removed from the freezer and maintained frozen in liquid nitrogen. The specimen was divided into the major body segments (head, thorax and abdomen). Half of the hornet’s thorax was split into two replicates (reaction A and reaction B) and placed in pre-frozen 2.0 ml reinforced tubes (SPEX SamplePrep catalog item 2310) each containing a sterilized 5/32inch 440 Stainless Steel Ball Bearing G25 (BC PERCISION) and submerged in liquid nitrogen. The frozen sample tubes were then placed in a pre- frozen aluminum tube block and loaded into the Geno/Grinder2010 (SPEX Sample prep). Samples were ground for 30 seconds at 1750 rpm until pulverized into a fine powder. Immediately after grinding, 100 μl 1x PBS and 10 μl Proteinase K was added to each sample and inverted 5 times to mix. A mixture containing 2 μl of RNase A and 75 μl Buffer AL were added to each sample, mixed by pulse-vortexing, and then spun down. The samples were allowed to incubate at 25° C for 2 hours. Following incubation, the lysate of each replicate was carefully transferred to new 1.5 ml microcentrifuge tubes using wide-bore pipette tips. MagAttract Suspension G was resuspended and 10 μl was added to each sample followed by 140 μl Buffer MB. The samples were placed in a thermomixer and incubated for 3 minutes at 25° C and 1400 rpm. The samples were centrifuged briefly, placed on a magnetic rack, and incubated on the benchtop until the solutions cleared. The supernatant was discarded, and the samples were removed from the magnetic rack. Buffer MW1 (350 μl) was added directly to each bead pellet and the samples were returned to the thermomixer and incubated at 25° C and 1400 rpm for 1 min. The above procedure was repeated for a total of 2 washes with Buffer MW1 and 2 washes with 350 μl Buffer PE. Following the second wash with Buffer PE the supernatant was discarded. While the samples remained on the magnetic rack, 350 μl nuclease-free water was carefully pipetted down the side of each tube opposite the bead pellet. The samples were allowed to incubate for exactly 1 min and the supernatant was quickly and carefully discarded. This process was repeated for a total of 2 washes with nuclease-free water. HMW genomic DNA was eluted by adding 100 μl Buffer AE directly to each bead pellet and incubated in a thermomixer at 37° C and 1400 rpm for 5 min. The samples were briefly centrifuged, placed on a magnetic rack, and allowed to incubate on the benchtop until the solutions cleared. The eluate containing purified gDNA was then transferred into a new 1.5 ml microcentrifuge tube using wide-bore pipette tips.

The extracted DNA was quantified using dsDNA BR Qubit assay and read using the Fluorometer feature on a Denovix DS-11 Series. Additionally, the purity of the DNA was assessed using the UV-Vis spectrometer feature on a Denovix DS-11 Series.

To improve the purity of the extracted DNA prior to shearing, a 1.75x bead clean-up was performed by adding 175 μl of room temperature polyethylene glycol containing solid-phase reversible immobilization beads solution [1] to each sample. The samples were mixed thoroughly by slow pipetting with wide-bore pipette tips and incubated at room temperature for 15 min. The samples were briefly centrifuged, placed on a magnetic rack, and allowed to incubate on the benchtop until the solutions cleared. The supernatant was carefully removed and discarded. While samples remained on the magnet, 500 µl of freshly prepared 80% ethanol was added to each tube. After a minute, the ethanol was removed and replaced with another 500 µl of 80% ethanol. After a minute the ethanol was removed, and the samples were spun down and placed back on to the magnet. Any residual ethanol was removed, and the samples were allowed to air dry for 5 min. The samples were removed from the magnet and resuspended in 26 µl elution buffer (5 mM Tris-HCl, pH 8.5) and incubated at 37° C for 15 min. The samples were briefly centrifuged, placed on a magnetic rack, and allowed to incubate on the benchtop until the solutions cleared. The eluate from each sample was carefully transferred using wide-bore pipette tips to new 1.5 ml tubes. A second elution was performed by resuspending each bead pellet in an additional 26 µl of elution buffer (5 mM Tris-HCl, pH 8.5). The samples were again incubated at 37° C for 15 min, briefly centrifuged and placed on a magnetic rack. The samples were allowed to incubate on the benchtop until the solution cleared and the eluates were carefully added to their respective 1.5 ml tube containing the first elution using wide-bore pipette tips.

The cleaned DNA was quantified using dsDNA BR Qubit assay and read using the Fluorometer feature on a Denovix DS-11 Series. Additionally, the purity of the DNA was assessed using the UV-Vis spectrometer feature on a Denovix DS-11 Series. The DNA size was qualified on the Agilent 4200 Tapestation System with genomic Screen Tape (Agilent Technologies).

DNA shearing was performed using the Diagenode Megaruptor 2. The Megaruptor was prepared by initializing the system and performing a full wash cycle followed by a 10 mM Tris-HCl pH 7.5 rinse. The machine was fitted with a long Hydropore (Diagenode catalog item E07010002) and pre-loaded. For the construction of a Pacific Biosciences HiFi SMRTbell library, HMW gDNA from each replicate was transferred into new 0.5 ml Hydro Tubes (Diagenode catalog item C30010018); 1.2 μgs of reaction A and 0.615 μgs of reaction B. The volume of each tube was brought to a total 300 μl with 5 mM Tris-HCl, pH 8.5). Each replicate was sheared using the specifications of 20 kb and volume of 300 μl.

After shearing, a 2x bead clean-up was performed by adding 690 μl of room temperature polyethylene glycol containing solid-phase reversible immobilization beads solution [1] to each sample. The samples were mixed thoroughly by slow pipetting with wide-bore pipette tips and incubated at room temperature for 15 min. The samples were briefly centrifuged, placed on a magnetic rack, and allowed to incubate on the benchtop until the solutions cleared. The supernatant was carefully removed and discarded. While samples remained on the magnet, 500 µl of freshly prepared 80% ethanol was added. After a minute, the ethanol was removed and replaced with another 500 µl of 80% ethanol. After a minute the ethanol was removed, and the samples were spun down and placed back on to the magnet. Any residual ethanol was removed, and the samples were allowed to air dry for 5 min. The samples were removed from the magnet and resuspended in 25 µl elution buffer and incubated at 37° C for 15 min. The samples were briefly centrifuged, placed on a magnetic rack, and allowed to incubate on the benchtop until the solutions cleared. The eluates were carefully transferred using wide-bore pipette tips to new 1.5 ml tubes. A second elution was performed by resuspending the beads in an additional 25 µl of elution buffer. The samples were again incubated at 37° C for 15 min, briefly centrifuged and placed on a magnetic rack. The samples were allowed to incubate on the benchtop until the solutions cleared and the eluates were carefully added to their respective 1.5 ml tubes containing the first elution using wide-bore pipette tips.

The sheared gDNA was quantified using a dsDNA BR Qubit assay and read using the fluorometer feature on a Denovix DS-11 Series. Additionally, an aliquot of sheared DNA was taken and diluted to approximately 1ng/μl and then sized on the Fragment Analyzer using a HS Large Fragment 50kb Kit (Agilent Technologies).

To prepare the SMRTbell library, working on ice a master mix of Diluted DNA Prep Additive was prepared in a new 0.2 ml PCR tube by combining 0.2 μl DNA Prep Additive and 0.8 μl Enzyme Dilution Buffer for each sample. For each replicate, up to 45.4 µl of sheared DNA was placed in a new 0.2 mL PCR tube and then the following were added to each reaction: 7.0 µl DNA Prep Buffer, 0.6 µl NAD, 1.0 µl Diluted DNA Prep Additive, and 1.0 µl DNA Prep Enzyme. The samples were mixed thoroughly by slow pipetting using wide-bore pipette tips and the contents were spun down. The reactions were incubated in a thermocycler at 37° C for 15 min to remove single-stranded overhangs and then returned to 4° C. Following the incubation, 2.0 μl DNA Damage Repair Mix v2 was added to each sample, mixed well, and spun down briefly. The reactions were allowed to incubate for an additional 30 min at 37° C to repair DNA damage. The reactions were allowed to cool to 4° C before 3.0 μl End Prep Mix was added to each sample and mixed. The reactions were incubated in a thermocycler with the following settings to allow for end repair and the addition of an A-tail: 20° C for 10 min, 65° C for 30min and hold at 4° C. Barcoded Overhang Adapter (5.0 μl of bc1011), 30.0 μl Ligation Mix, 1.0 μl Ligation Additive, 1.0 μl Ligation Enhancer was added to each sample, mixed and allowed to ligate at 20° C for 3 hours. Following the ligation, the reactions were incubated for an additional 10 min at 65° C to inactivate the ligase and returned to 4° C. A nuclease treatment to remove damaged SMRTbell library was performed by adding 2.0 μl Enzyme A, 0.5 μl Enzyme B, 0.5 μl Enzyme C, and 1.0 μl Enzyme D to each sample. The reactions were mixed well, spun down briefly, and incubated at 37° C for 1 hr.

A 0.8x Ampure PB clean-up was performed by adding 81 μl of room temperature Ampure PB beads to each sample. The samples were mixed thoroughly by slow pipetting with wide-bore pipette tips and incubated at room temperature for 15 min. The samples were briefly centrifuged, placed on a magnetic rack, and allowed to incubate on the benchtop until the solutions cleared. The supernatant was carefully removed and discarded. While samples remained on the magnet, 500 µl of freshly prepared 80% ethanol was added. After a minute, the ethanol was removed and replaced with another 500 µl of 80% ethanol. After a minute the ethanol was removed, and the samples were spun down and placed back on to the magnet. Residual ethanol was removed, and the samples were allowed to air dry for 5 min. The samples were removed from the magnet and resuspended in 75 µl elution buffer and incubated at 37° C for 15 min. The samples were briefly centrifuged, placed on a magnetic rack, and allowed to incubate on the benchtop until the solutions cleared. The eluates were carefully transferred using wide-bore pipette tips to new 1.5 ml tubes. A second elution was performed by resuspending each bead pellet in an additional 25 µl of elution buffer. The samples were again incubated at 37° C for 15 min, briefly centrifuged and placed on a magnetic rack. The samples were allowed to incubate on the benchtop until the solutions cleared and the eluates were carefully added to their respective 1.5 ml tubes containing the first elution using wide-bore pipette tips.

A 3kb size selection was performed using 2.2x volume of 40% diluted Ampure PB beads. 220 μl of room temperature 40% Ampure PB beads was added to each sample and mixed thoroughly by slow pipetting with wide-bore pipette tips. The samples were incubated at room temperature for 15 min, briefly centrifuged, and placed on a magnetic rack. The samples were allowed to incubate on the benchtop until the solutions cleared. The supernatant was carefully removed and discarded. While samples remained on the magnet, 500 µl of freshly prepared 80% ethanol was added. After a minute, the ethanol was removed and replaced with another 500 µl of 80% ethanol. After a minute the ethanol was removed, and the samples were spun down and placed back on to the magnet. Residual ethanol was removed, and the samples were allowed to air dry for 5 min. The samples were removed from the magnet and resuspended in 10 µl elution buffer and incubated at 37° C for 15 min. The samples were briefly centrifuged, placed on a magnetic rack, and allowed to incubate on the benchtop until the solutions cleared. The eluates were carefully transferred using wide-bore pipette tips to new 1.5 ml tubes. A second elution was performed by resuspending each bead pellet in an additional 7 µl of elution buffer. The samples were again incubated at 37° C for 15 min, briefly centrifuged and placed on a magnetic rack. The samples were allowed to incubate on the benchtop until the solutions cleared and the eluates were carefully added to their respective 1.5 ml tubes containing the first elution using wide-bore pipette tips.

The SMRTbell libraries were quantified using a dsDNA BR Qubit assay and read using the fluorometer feature on a Denovix DS-11 Series. Additionally, an aliquot of sheared DNA was taken and diluted to approximately 1ng/μl and then sized on the Fragment Analyzer (Agilent Technologies) using a HS Large Fragment 50kb Kit. Size-fractionation of SMRTbell Library was performed on a SageELF (Sage Science, INC) using a 0.75% agarose gel cassette (1kb-40kb) and marker 75. By following the manufacturer's instructions the instrument was calibrated and the gel cassettes were prepared. The target value was set to 3450 and all wells were selected ON to collect. Each of the remaining prepared SMRTbell Libraries (644ng of reaction A and 695ng of reaction B) were brought up to a total volume of 30 µl with the elution buffer and gently mixed with 10 µL of Sage Science’s Marker 75. Each mixture was loaded into individual gel cassettes and allowed to run. For each cassette, eluate from the 13 collection wells were transferred using wide-bore pipette tips to new 0.2 mL tubes. Each well was rinsed twice with 30 µl of elution buffer and combined with its respective fraction. All size-fractionation were quantified using dsDNA BR Qubit assay using the Fluorometer feature on a Denovix DS-11 Series and then sized on the Fragment Analyzer using a HS Large Fragment 50kb Kit (Agilent Technologies).

From reaction A, well 5(avg. size: 17.9kb), well 6(avg. size: 14.3 kb), and well 7(avg. size: 12 kb) were pooled together with well 7(avg. size: 10.8 kb) from reaction B. The pooled fractions were cleaned and concentrated by adding 720 μl (2x volume) of room temperature polyethylene glycol containing solid-phase reversible immobilization beads solution [1]. The sample was mixed thoroughly by slow pipetting with wide-bore pipette tips and incubated at room temperature for 15 min. The sample was briefly centrifuged, placed on a magnetic rack, and allowed to incubate on the benchtop until the solution cleared. The supernatant was carefully removed and discarded. While the sample remained on the magnet, 500 µl of freshly prepared 80% ethanol was added. After a minute, the ethanol was removed and replaced with another 500 µl of 80% ethanol. After a minute the ethanol was removed, and the sample was spun down and placed back on to the magnet. Residual ethanol was removed, and the sample was allowed to air dry for 5 min. The sample was removed from the magnet and resuspended in 10 µl elution buffer and incubated at 37° C for 15 min. The sample was briefly centrifuged, placed on a magnetic rack, and allowed to incubate on the benchtop until the solution cleared. The eluate was carefully transferred using wide-bore pipette tips to new 1.5 ml tubes. A second elution was performed by resuspending the bead pellet in an additional 7 µl of elution buffer. The sample was again incubated at 37° C for 15 min, briefly centrifuged and placed on a magnetic rack. The sample was allowed to incubate on the benchtop until the solution cleared and the eluate was carefully added to the 1.5 ml tubes containing the first elution using a wide-bore pipette tip.

The completed HiFI SMRTbell Library was quantified using dsDNA BR Qubit assay and read using the Fluorometer feature on a Denovix DS-11 Series. An aliquot was diluted to 1ng/µl and sized on the Fragment Analyzer using a HS Large Fragment 50kb Kit (Agilent Technologies).

Sequencing:

The library SGAP075 was assigned the local identifier LIB107643. Sequencing was performed on a Sequel II System using Bind Kit 2.0, Sequencing kit v2.0, and SMRT Cell 8M. To target HiFi reads, the library was loaded at 100pM concentration using 30 hour movie times on two cells. Raw subreads were converted to HiFi data by processing with CCS to call a single high quality consensus sequence for each molecule, using a 99.5% consensus accuracy cutoff.

HiFi read stats:

|  |  |  |
| --- | --- | --- |
|  | SMRT Cell 1 | SMRT Cell 2 |
| # of HiFi reads | 980,541 | 1,191,311 |
| Yield | 15.5 Gb | 18.9 Gb |
| Mean insert length | 15,822 bp | 15,883 bp |
| Median read quality | Q28 | Q29 |

Reads were adapter barcode trimmed using [lima](https://github.com/PacificBiosciences/barcoding). Kmer analysis of the HiFi reads using [GenomeScope](http://qb.cshl.edu/genomescope/genomescope2.0/) [2] yielded 66x average haploid coverage, corresponding to a genome size estimate of 260 Mb.

Assembly:

The nuclear genome assembly was carried out using IPA (Sović, Kronenberg *et al.* [3,4], using the [bioconda release](https://github.com/PacificBiosciences/pbbioconda/wiki/Improved-Phased-Assembler)) with default settings. Primary contigs were subjected to [purge\_dups](https://github.com/dfguan/purge_dups) [5] to generate the primary assembly file. The purged contigs were combined with the original haplotigs file, followed by an additional round of purge\_dups.

|  |  |  |
| --- | --- | --- |
|  | Primary | Haplotigs |
| Assembly size | 248 Mb | 200 Mb |
| # of contigs | 262 | 2,523 |
| Contig N50 | 3.14 Mb | 0.16 Mb |
| BUSCO (insecta)1 completeness (n=1,658) | C:99.3% [S:98.9%,D:0.4%] F:0.4%,M:0.3% | C:59.9% [S:59.5%,D:0.4%] F:2.1%,M:38.0% |
| Conserved gene frameshift error analysis (BUSCO insecta)2 | 20 frame shift inducing errors in 1,308 kb gene alignment space (Q48.2) | 41 frame shift inducing errors in 776 kb gene alignment space (Q42.8) |

1https://busco.ezlab.org/; 2Korlach *et al.* (2017) Gigascience 6: 1-16

The second round of purge\_dups purged 56 Mb from the haplotigs file, consisting of repeat sequence
(BUSCO C:0.1%[S:0.1%,D:0.0%],F:0.0%,M:99.9%)

During screening it was determined that three adapters remained in the primary assembly and two in the haplotig assembly in addition to two runs of Ns in the primary assembly which were introduced by purge\_dups. These sequences were removed by splitting the contigs, yielding a final count of 267 contigs in the primary assembly and 2,525 in the hapolotigs.

The mitochondrial genome was assembled following a general strategy described previously [6]. Mitochondrial reads were selected by alignment of the full HiFi read dataset against the *Vespa mandarinia* mtgenome reference (NCBI accession [NC\_027172](https://www.ncbi.nlm.nih.gov/nuccore/827045924)) using [minimap2](https://github.com/lh3/minimap2) [7]. IPA was run on the mitochondrial read set, followed by manual end overlap trimming and start-point rotation (to the NCBI reference start point) of the resulting mitochondrial genome contigs. Multiple contigs were obtained due to the presence of an extended variable number tandem repeat (VNTR) region corresponding to the control region (this mtgenome portion is absent in the NCBI reference), with different copy numbers (ranging from 5 to 9) of an 823 bp repeat unit. Remapping of the mitochondrial read set provided estimates of the relative abundance of the mtgenome VNTR variants in the hornet individual:

|  |  |  |  |
| --- | --- | --- | --- |
| Mtgenome VNTR variant | Mtgenome length | Coverage | Relative abundance |
| 5 | 20,009 bp | 160x | 5% |
| 6 | 20,831 bp | 1800x | 54% |
| 7 | 21,654 bp | 850x | 25% |
| 8 | 22,474 bp | 500x | 15% |
| 9 | 23,300 bp | 30x | 1% |

We designated the most abundant mtgenome variant (6 repeat copies) as the mtgenome sequence, and also provide the sequences of all five mtgenome VNTR variants. Remapping revealed additional heterogeneity in the VNTR region, with 15 heteroplasmy SNP locations and a 92 bp insertion/deletion structural heteroplasmy.

Data availability & File designations:

SRA: SRR12366675
ihVesMand1\_HiFi\_reads\_Cell1.ccs.bam: HiFi reads from SMRT Cell 1
ihVesMand1\_HiFi\_reads\_Cell2.ccs.bam: HiFi reads from SMRT Cell 2

BioProject: PRJNA649644, BioSample: SAMN15675875, GenBank: JACHAV000000000

ihVesMand1\_IPA\_purge\_primary.fasta: IPA Primary contig assembly (after purge\_dups)

ihVesMand1\_mtgenome.fasta: mitochondrial genome (VNTR6)

BioProject: PRJNA649643, BioSample: SAMN15675875, GenBank: JACHAW000000000
ihVesMand1\_IPA\_purge\_htig.fasta: IPA Haplotigs contig assembly (after purge\_dups)

Ag Data Commons
ihVesMand1\_IPA\_purged\_from\_htig.fasta: IPA purged from haplotigs contig set
ihVesMand1\_mt\_reads.fasta: mitochondrial HiFi read set
ihVesMand1\_mtgenome\_all\_VNTR\_variants.fasta: all mtgenome VNTR variants (5-9 repeat copies)

References:

1. DeAngelis *et al.* (1995) “Solid-phase reversible immobilization for the isolation of PCR products”, Nucleic acids research 23(22): 4742–4743.
2. Ranallo-Benavidez *et al.* (2020) “GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes”, Nature Communications 11: 1432.
3. Kronenberg (2020) “Direct Phased Genome Assembly Using Nighthawk on HiFi Reads”, [PacBio blog post](https://www.pacb.com/blog/direct-phased-genome-assembly-using-nighthawk-on-hifi-reads/).
4. Sović (2020), “Improved & Phased Assembly using HiFi Data”, [SMRT Leiden conference presentation](https://www.gotostage.com/channel/824b0f358f7b437eb3bb2c37ae5d063b/recording/a65c3e6f1f0441aeae07f686fb4e6b90/watch?source=CHANNEL).
5. Guan *et al.* (2020) “Identifying and removing haplotypic duplication in primary genome assemblies”, Bioinformatics 36: 2896.
6. Soorni *et al.* (2017) “Organelle\_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data”, BMC Genomics 18: 49.
7. Li (2018) “Minimap2: pairwise alignment for nucleotide sequences”, Bioinformatics 34: 3094.