data

MDPI

Data Descriptor

First Draft Genome Assembly of the Malaysian Stingless Bee, *Heterotrigona itama* (Apidae, Meliponinae)

Chien-Yeong Wee [1,*], Amin-Asyraf Tamizi [1], Nazrul-Hisham Nazaruddin [1], Siuk-Mun Ng [2], Jia-Shiun Khoo [2] and Rosliza Jajuli [3]

1 Biotechnology and Nanotechnology Research Centre, Malaysian Agricultural Research and Development Institute (MARDI), Serdang 43400, Selangor, Malaysia; aminasyraf@mardi.gov.my (A.-A.T.); nazrulhisham@mardi.gov.my (N.-H.N.)
2 Codon Genomics S/B, Jalan Dutamas 7, Taman Dutamas Balakong, Seri Kembangan 43200, Selangor, Malaysia; siukmun@codongenomics.com (S.-M.N.); jiaxun@codongenomics.com (J.-S.K.)
3 Agrobiodiversity and Environmental Research Centre, Malaysian Agricultural Research and Development Institute (MARDI), Serdang 43400, Selangor, Malaysia; rosliza@mardi.gov.my
* Correspondence: cywee@mardi.gov.my

Received: 29 September 2020; Accepted: 21 November 2020; Published: 30 November 2020

Abstract: The Malaysian stingless bee industry is hugely dependent on wild colonies. Nevertheless, the availability of new queens to establish new colonies is insufficient to meet the growing demand for hives in the industry. *Heterotrigona itama* is primarily utilized for honey production in the region and the major source of stingless bee colonies comes from the wild. To propagate new colonies domestically, a fundamental understanding of the biology of queen development, especially from the genomics aspect, is necessary. The whole genome was sequenced using a paired-end 150 strategy on the Illumina HiSeq X platform. The shotgun sequencing generated approximately 89 million raw pair-end reads with a total output of 13.37 Gb and a GC content of 37.31%. The genome size of the species was estimated to be approximately 272 Mb. Phylogenetic analysis showed *H. itama* are much more closely related to the bumble bee (*Bombus* spp.) than they are to the modern honey bee (*Apis* spp.). The genome data provided here are expected to contribute to a better understanding of the genetic aspect of queen differentiation as well as of important molecular pathways which are crucial for stingless bee biology, management and conservation.

Dataset: This genome sequencing project has been deposited at the European Nucleotide Archive (ENA) under the accession BioProject ID PRJEB34838. The data can be accessed at ENA (https://www.ebi.ac.uk/ena/browser/view/PRJEB34838).

Dataset License: The European Nucleotide Archive (ENA) policies on data release follow the International Nucleotide Sequence Database Collaboration (INSDC) Policy, remaining permanently accessible as part of the scientific record. The archive is an open, supported platform for the management, sharing, integration, archiving and dissemination of sequence data.

Keywords: eusocial insect; genomics; Illumina sequencing; pollinator; queen differentiation

1. Summary

Malaysia is home to about 50 different species of Meliponines (stingless bees) previously belonging to 13 genera [1]. However, they recently have undergone taxonomic revisions and the 13 genera are now reduced to only seven—with some of the genera having been revised into the subgenus level [2,3].

While some taxa have limited distributions, *Heterotrigona itama* populations are scattered in Malesia (the Malay Archipelago) spanning over Peninsular Malaysia, Borneo's East Malaysia, Southern Thailand, Singapore and Indonesia (Kalimantan, Java and Sumatra) [1,4]. Based on our recent species diversity survey, Meliponines found in Malaysia produce edible, fragrant kinds of honey, the taste of which is generally sweet-tangy and whose color varies depending on the season. Among these, *H. itama* has become the most common species reared or kept for its highly prized medicinal honey; this species has also been regarded as an economically-important insect in agriculture due to its high adaptability to a wide range of habitats and floristic attributes [5–9].

H. itama is a member of the Apidae family within the order of Hymenoptera, which displays eusociality behavior and a caste polyphenism in females. Female larvae develop into two interdependent adult castes, the queen and the worker, during caste differentiation. Chen et al. [10] stated that caste differentiation depends on the differential expression of entire sets of genes involved in the larval fate of queens and workers. The substantial physiological and morphological differences are due to the differential expression of these genes [10–13]. To date, the genetic aspect of queen differentiation in Malaysian stingless bees has not been fully addressed, yet this area is crucial not only to enhance our understanding of caste differentiation but also to complement the ongoing research on producing queens using an in vitro platform. Tamizi et al. [14] conducted a related study focusing on the transcriptomics of a queen larva which discovered that *H. itama* most highly follows a conserved caste differentiation pathway based on the detection of a few sets of annotated genes related to queen differentiation. Other studies related to *H. itama* queens were on oviposition behavior and the presence of virgin queen eggs after colony splitting [6,8]; however, these two did not address the caste differentiation during the developmental stage. Here, we report the genome sequencing, de novo assembly and annotation data of the Malaysian stingless bee, *H. itama*, which can help researchers to better access genes and pathways related to queen differentiation from a much larger nucleotide reference—the genome. In addition, the data can be useful as a genomic reference to facilitate future studies on sustainable cultivation and conservation of stingless bees. The sequenced *H. itama* genome is the first report for the Indo-Malayan/Australasian stingless bee group and the fourth in the world after two Neotropical stingless bees, *Melipona quadrifasciata* and *Frieseomelitta varia* [15,16], and an Asian species from Taiwan, *Lepidotrigona ventralis hoosana* (GenBank accession: PRJNA387986), which had been revised as *Lepidotrigona hoozana* (Strand) Rasmussen [1].

2. Data Description

The whole genome sequencing of a queen of *H. itama* (Figure 1A) generated approximately 89 million raw pair-end reads with a total output of 13.37 Gb sequencing data. A total of 90.83% of the sequencing data were retained after pre-processing; more than 82 million high-quality reads with approximate 12.15 Gb total bases were generated (Table 1A).

Table 1. Summary statistics of Malaysian stingless bee, *H. itama* draft genome.

	(**A**) ***Sequencing Reads***	
	Total number	***Total bases (bp)***
Raw data	89,154,444	13,373,166,600
Pre-processed data	82,979,630 (93.07%)	12,146,499,923 (90.83%)
	(**B**) ***Assembly Data***	
No. contigs (≥1000 bp)		13,733
Total length (≥1000 bp)		262,450,989
N50		43,263
N75		14,434
L50		1649
L75		4534
GC content		37.31%

Table 1. *Cont.*

(**C**) ***Structural Annotation***		
Number of predicted protein-coding genes		12,496
Total length of CDS (bp)		5,951,300
Number of predicted protein-coding genes (≥99 bp)		12,482
Total length of CDS (bp) (≥99 bp)		5,951,047
Number of tRNA		398
Number of rRNA		14
(**D**) ***Functional Annotation***		
Predicted protein-coding genes (≥99bp)		12,496 (100%)
Protein with RefSeq blast hits		10,388 (83.22%)
Protein with Swiss-Prot blast hits		8,214 (65.81%)
Protein with GO assignments		7,999 (64.08%)
Total GO annotation		175,576
Total EC annotation		613
Total KEGG pathway annotation		142
(**E**) ***Annotated Gene Ontology Related to Caste Differentiation and Insect Hormone***		
(Note: Number in the parentheses indicates the number of the genes of interest)		
Caste differentiation	GO:0048650	Caste determination, influenced by environmental factors (1)
Insect hormone (Top 5)	GO:0009725	Response to hormone (77)
	GO:0045433	Male courtship behavior (30)
	GO:0048749	Compound eye development (25)
	GO:0045169	Fusome (21)
	GO:0060562	Epithelial tube morphogenesis (35)

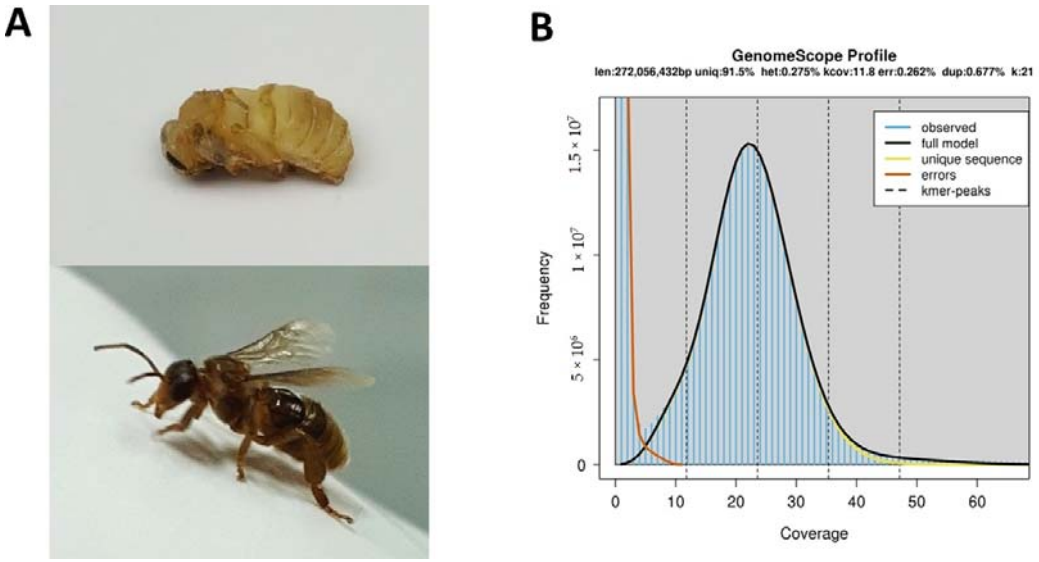

Figure 1. Genome size estimation of Malaysian stingless bee, *H. itama*. (**A**) Photograph of a queen pupa (unpigmented) and newly emerged queen of *H. itama* (slightly pigmented). (**B**) 21-mer GenomeScope profile showed the predicted genome size with other matrices for genome profiling.

2.1. Genome Size Estimation

Analysis of *k*-mer distribution used the high-quality reads to estimate the genome size, heterozygosity and repeat content of the stingless bee genome. Based on the *k*-mer spectrum (Figure 1B), a simple Poisson profile denotes the low heterozygosity of the organism (0.275%) [17]. In addition, the genome was found to be slightly repetitive, and the estimated genome size of the Malaysian stingless bee (*H. itama*) was discovered to be approximately 272 Mb. This is similar to the size of previously reported genomes of the honey bee (*Apis mellifera*—236 Mb) [18], bumble bee (*Bombus terrestris*—249 Mb) [19] and two Neotropical stingless bee species (*M. quadrifasciata*—256 Mb,

F. varia—275 Mb) [15,16]. Based on the draft genome size estimated, subsequent de novo assembly and genome annotation were performed with the sequencing depth of approximately 49X coverage.

2.2. Genome Assembly and Structural Annotation

The draft genome with a total assembly size of 262.45 Mb has 13,733 contigs (≥1000 bp) with GC content of 37.31%, the longest contig of 438,094 bp and contig N50 sizes of 43,263 bp (Table 1B). The BUSCO integrity assessment detected a 94.4% completeness score of complete single-copy orthologs (53-mer) in the genome assembly (Figure 2A), indicating a high level of completeness of the draft genome. In addition, the BUSCO assessment reported a completeness score of 94.6% on the genome assembly based on hymenoptera_odb9 profiles, indicating a comparable level of completeness to that of other bee genomes (Figure 2B). Subsequently, a total of 12,496 protein-coding genes (>99 bp), 398 tRNA and 14 rRNA genes were predicted from structural annotation (Table 1C). The mean exon per gene annotated is 6.57. The predicted genes achieved 84.6% completeness based on hymenoptera_odb9 BUSCO profiles.

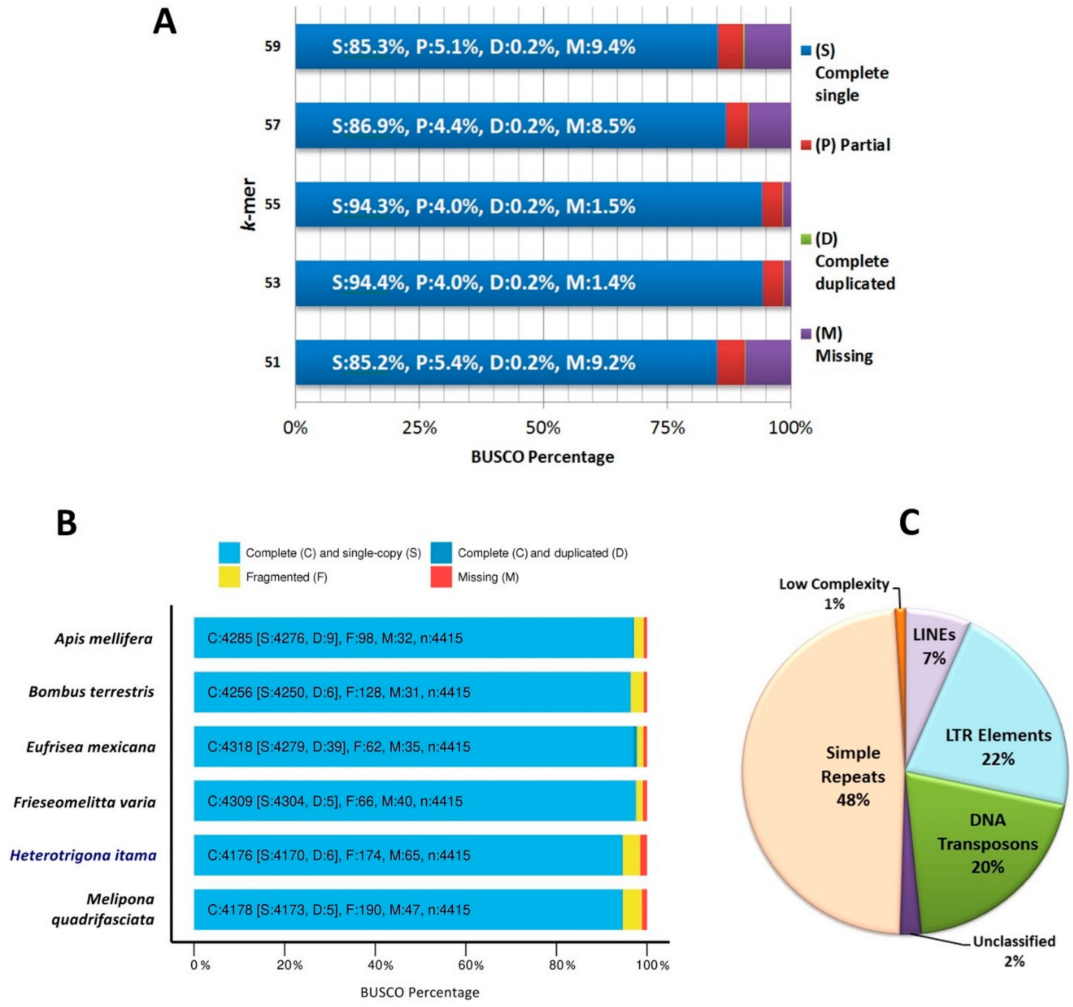

Figure 2. Genome assembly and structural annotation. (**A**) Benchmarking Universal Single Copy Orthologs (BUSCO) assessment based on different *k*-mer (51- to 59-mer). (**B**) Benchmarking Universal Single Copy Orthologs (BUSCO) assessment for six different bee species including the *H. itama* assembled in this study. (**C**) Repeat elements of the draft genome.

A total of 2.29 Mb, equivalent to 0.87% repeats belonging to different classes of interspersed repeats were masked for the stingless bee draft genome. The repetitive elements identified in the draft

genome comprising 0.19% of long terminal repeats (LTRs), 0.18% of DNA transposons, and 0.06% of long interspersed nuclear elements (LINEs) (Figure 2C). The low amount of repetitive sequences detected was due to the sequencing technology used and the relatively low sequencing coverage. Short reads generated could have been collapsed into contiguous contig or resulted in a fragmented assembly. In addition, the relatively low sequencing coverage of ~49× might not be enough to sequence every genomic region and thus some repetitive regions might be missed out from the sequencing.

2.3. Functional Annotation

From the predicted protein-coding genes (>99 bp) obtained, the peptide sequences (≥33 amino acid) were found to be 83.22% and 65.81% functionally annotated against RefSeq [20] and Swiss-Prot [21] databases, respectively (Table 1D). The RefSeq top-hit species annotation revealed that most of the stingless bees' proteins hit to those of the bumble bee, *B. terrestris* dataset. Further, protein structure characterization successfully discovered a total of 7999 genes from Gene Ontology (GO) [22] and categorized to biological function, cellular component and molecular function. A total of 142 KEGG pathways were mapped with the most enzymes identified in the biosynthesis of the antibiotics pathway, the purine metabolism pathway and the cysteine and methionine metabolism pathway (Figure 3). Subsequently, genes of interest related to caste differentiation and insect hormone were data-mined from the annotated GO terms (Table 1E). Several key genes related to insect hormone, which were reported to play an important role in caste development [11,23], were mapped to the insect hormone biosynthesis pathway (KEGG pathway: map 00981).

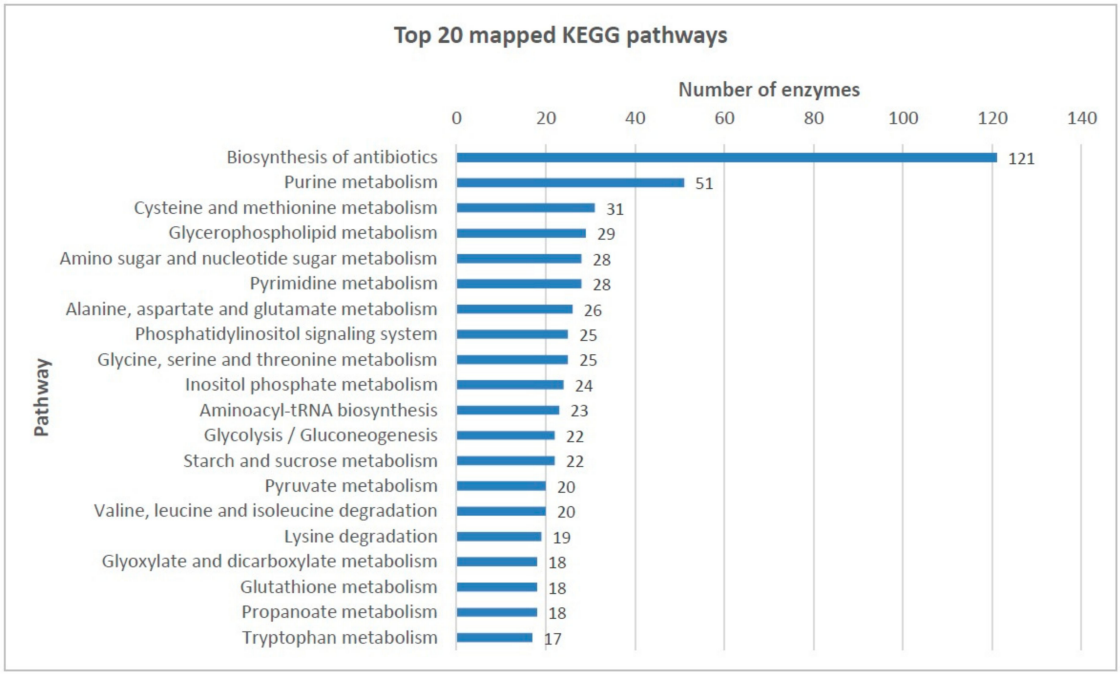

Figure 3. Distribution of the top twenty Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways mapped with the most number of enzymes through KEGG pathway analysis.

2.4. Orthologous and Phylogenetic Analysis

Ortholog analysis showed that 10,067 orthologous clusters were formed based on the protein data set from the four Hymenoptera bee species. The numbers in the Venn diagram represent the number of orthologous clusters that *H. itama* shares with the three other species. In total, 7573 orthologous clusters were found to be common in all four bee species, suggesting their conservation in the lineage after speciation. In addition, the diagram shows that there are 70 clusters exclusive to the Malaysian

stingless bee, *H. itama* (Figure 4A). These 70 clusters contain unannotated genes whose function have not been studied at this stage (Table S1).

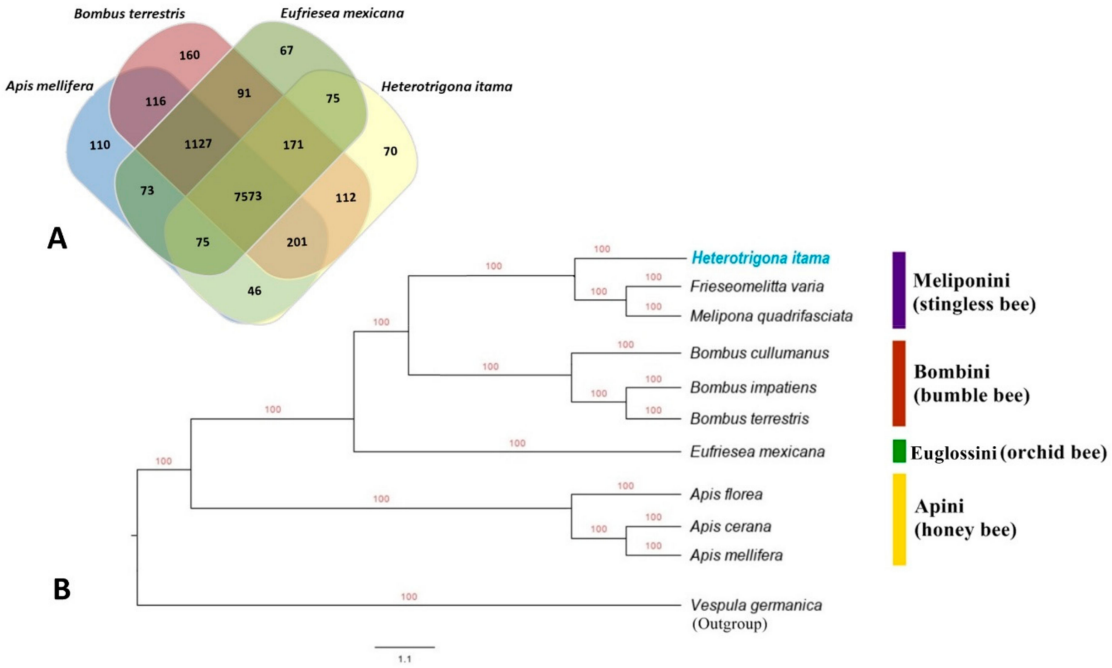

Figure 4. Orthologous and phylogenetic analysis. (**A**) Venn diagram of orthologous gene clusters between four extant representative members (Apini, Bombini, Euglossini and Meliponini) of the bee clade Corbiculata including *H. itama*. (**B**) Bayesian inference (BI) tree obtained for ten bee species based on 39 conserved single-copy orthologs rooted on *Vespula germanica* (branches transformed proportionally; posterior probability (%) labeled; species names labeled at tips).

As for the phylogenetic analysis, a total of 39 single-copy orthologs that were found to be conserved across the ten bees and one wasp species were adopted and analyzed. The eleven species that were analyzed were made up of different tribes, including the stingless bee, honey bee, bumble bee and euglossine bee, with the wasp as the outgroup. Phylogeny of these concatenated genes showed that the *H. itama* evolved closely with two other stingless bees, *F. varia* and the *M. quadrifasciata*, which share a common ancestor with the bumble bees (Figure 4B). The phylogenetic inference also indicated that stingless bees are much closely related to bumble bees than they are to the modern honey bee group (*Apis* spp.). This is further supported by a study done by Tamizi et al. [14], which found that the *H. itama* transcripts shared higher similarity with *B. impatiens* and *B. terrestris* (30%–39%) than with *A. mellifera* and *A. florea* (<8.1%).

In summary, we report the first draft genome of *H. itama*, the fourth Meliponine genome to be sequenced. Since there are limited genomic sequence resources for this intriguing group of social insects, our study hopes to provide more insights into and understanding of the biological processes which could contribute to the conservation and sustainable management of stingless bees in Malaysia.

3. Methods

3.1. Sample Preparation and Sequencing

A single queen of *H. itama* (pupal stage) was collected from an active colony placed at MARDI, Serdang, Selangor. The main reason a single individual (particularly the queen) was used is the size advantage of the queen compared to a drone (male). A single queen pupa can provide just enough DNA material needed for genome sequencing while three-four workers or drones are needed to produce an equal amount of DNA. The biology and behavior of *H. itama* are slightly different than those of

Apis spp. While drones from a single healthy Apini colony are most likely produced by a single queen, drones from Meliponini could come from different lineages as some species may house multiple queens at a time; even the workers could actively participate in laying haploid eggs that turn into drones. Genomic DNA was extracted from the whole pupa following a modified cetyl trimethyl ammonium bromide (CTAB) protocol [24]. Next generation sequencing library preparations were constructed following the manufacturer's protocol (VAHTS Universal DNA Library Prep Kit for Illumina). Prior to the library construction, the quality assessment of the genomic DNA sample was performed using Qubit 2.0 fluorometer (Life Technologies Corporation) for concentration determination and agarose gel electrophoresis for integrity testing. The prepared library was then loaded onto an Illumina HiSeq X Ten instrument according to the manufacturer's instructions (Illumina, San Diego, CA, USA). Sequencing was carried out using a 2 × 150 paired-end (PE) configuration; image analysis and base calling were conducted by the HiSeq Control Software (HCS) + OLB + GAPipeline-1.6 (Illumina) on the HiSeq instrument. The data quality of the raw sequencing reads was assessed using FastQC version 0.11.3 [25]. Pre-processing was then carried out using Trimmomatic version 0.38 [26] with the following parameters: ILLUMINACLIP: TruSeq3-PE-2. fa: 2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:8:20 MINLEN:50. Reads that were at least QV20, read length of ≥50 bp and adapters-free paired sequences were collected as clean reads.

3.2. Genome Size Estimation and Genome Assembly

The high-quality reads were used for *k*-mer distribution analysis to estimate the genome size, heterozygosity and repeat content of the stingless bee genome. Jellyfish v2.1.1 (University of Maryland, College Park, US) [27] was used to calculate the *k*-mer occurrences in DNA with a *k*-mer size of 21. Following that, GenomeScope v1.0.0 (Cold Spring Harbor Laboratory, Laurel Hollow, US) [28], a *k*-mer analysis software that uses the *k*-mer frequencies output from jellyfish, was employed to predict the genome size along with several other metrics for genome profiling.

The draft genome was de novo assembled using Velvet v1.2.10 (KBase, US) [29]. The assembly was carried out using different *k*-mer lengths (ranging from 51 to 59) with the parameters as follows: -shortPaired -separate -fastq for velveth; exp_cov = auto, read_trkg = yes for velvetg. Among the assemblies built with different *k*-mer sizes, the best assembly was chosen based on the number of contigs, assembly size, N50 and BUSCO completeness score. Therefore, QUAST v3.1 (St. Petersburg Academic University, St. Petersburg, Russia) [30] was used to access the quality of the draft genome in terms of contiguity through three different contig thresholds that were set at 300 bp, 500 bp and 1000 bp respectively. Benchmarking Universal Single-Copy Orthologs (BUSCO) v2.0 (Université de Genève, Geneva, Switzerland) [31] was used to assess the completeness of the draft genome. The hymenoptera_odb9 profile was chosen as the reference profile for this sample. After both assessments, filtering was applied to remove contigs shorter than 1000 bp in order to improve the overall contiguity without losing much genetic information. In other words, all contigs longer than 1000 bp were retained for downstream analyses.

3.3. Gene Structural Annotation and Functional Annotation

The draft genome was annotated using the MAKER v3.0 (University of Utah, Salt Lake City, US) [32] genome annotation pipeline which combines repeat masking, different prediction tools with evidence-based quality control and gene-model editing. The custom repeat library was used by RepeatMasker within the MAKER pipeline to mask repetitive elements. Transcript assembly from the same species [14] was adopted as EST evidence. In addition, four different sets of protein sequences from *Apis* spp., *Bombus* spp., *Melipona* spp. and *Scaptotrigona* spp. were downloaded from a public database and used as evidences to aid gene predictions. As MAKER was run iteratively for three times, repeat masking and evidence alignment were first performed with the following parameters: est = Trinity.fasta, protein = apis.fasta,bombus.fasta,melipona.fasta,scaptotrigona.fasta, model_org = hymenoptera, est2genome = 1 and protein2genome = 1. The resulting general feature

format (GFF3) file was used as input for all subsequent MAKER runs. Gene predictions were performed within MAKER using SNAP. For the second and third rounds of MAKER, the following parameters were used: est2genome = 0 and protein2genome = 0 in order to specify ab initio gene prediction. tRNA was predicted using tRNAscan-SE v1.3.1 [33] with default parameters. This was followed by rRNA prediction using RNammer v1.2 [34] by including these parameters: -S euk and -multi. After gene prediction, the full repertoire of peptide sequences (≥33 amino acid) was assessed for completeness using BUSCO v2.0 [31]. The hymenoptera_odb9 profile was chosen as the reference profile. Transposable elements and repetitive regions in the genome were identified using RepeatMasker v4.0.5 (Institute for Systems Biology, Seattle, US) [35] with the following parameters: Hymenoptera as the source species of query DNA, NCBI search engine, without masking the small RNA (pseudo) genes or low complexity DNA or simple repeats but masking only the interspersed repeats using a sensitive slow-search mode.

The Protein homology BLAST search was performed against NCBI Reference Sequence protein (RefSeq) [20] and Swiss-Prot [21] protein databases. Diamond v0.9.22 (The University of Tübingen, Tübingen, Germany) [36] was used to blast the peptide sequences to the RefSeq database, while ncbi-blast v2.7.1+ was used to blast the same peptide set against Swiss-Prot. The cut-off for both BLAST searches was set at a maximum expect value (E-value) of 1e-5. Subsequently, the BLAST outputs from both databases were used for Gene Ontology (GO) [22] and Kyoto Encyclopedia of Genes and Genomes (KEGG) [37] analysis using standalone Blast2GO v2.5 (BioBam, Valencia, Spain) [38]. In addition, protein structure characterization was carried out with local InterProScan lookup service v5.4-47.0 (EMBL-EBI, Cambridgeshire, UK) [39].

3.4. Orthologous and Phylogenetic Analysis

Ortholog analysis was carried out using OrthoMCL v2.0.9 (University of Pennsylvania, Philadelphia, US) [40] with default parameters. Three protein data sets from *A. mellifera*, *B. terresteris* and *E. mexicana* together with the protein data of the Malaysian stingless bee were used for orthologous group clustering.

Genome sequences from a total of eleven species (namely the *H. itama* sequenced and assembled in this study, *A. cerana* (RefSeq accession: GCF_001442555.1), *A. florea* (RefSeq accession: GCF_000184785.3), *A. mellifera* (GenBank accession: GCA_000002195.1), *B. cullumanus* (GenBank accession: GCA_014737535.1), *B. impatiens* (RefSeq accession: GCF_000188095.3), *B. terrestris* (RefSeq accession: GCF_000214255.1), *E. mexicana* (GenBank accession: GCA_001483705.1), *F. varia* (GenBank accession: GCA_011392965.1), *M. quadrifasciata* (GenBank accession: GCA_001276565.1) and *V. germanica* (GenBank accession: GCA_014466195.1)) were assessed for their single-copy orthologs using BUSCO v2.0 [31], utilizing metazoa_odb9 (N = 978) BUSCO profiles. A subset of single-copy orthologs (Table 2) that were annotated as complete and present in all eleven species were aligned using MAFFT v7.471 (AIST, Tsukuba, JP) [41]. Manual inspection and trimming were carried out for single-copy orthologs with gaps or poor alignments in ≥50% of the sequences. A phylogeny based on the refined concatenated multiple sequence alignments of 39 single-copy orthologs was generated using MrBayes v3.2.7 [42,43] utilizing the Bayesian Markov Chain Monte Carlo (MCMC) algorithm. The analysis was conducted by sampling a mixture of models: fixed rate matrices as well as gamma-distributed variable- and invariable matrices [44]. Input alignments were carried out with sampling frequency for every 500 generations. A burn-in of 25% from the beginning of the cold chain was discarded. An average standard deviation of split frequencies <0.01 was achieved. The plot of generation versus log probability of the data did not show a noticeable trend, and potential scale reduction factor (PSRF) close to 1.0 was set for all parameters. The Bayesian posterior probability of tree was based on a total of 50,000 generations. The phylogenetic tree was visualized using FigTree v1.4.4 (The University of Edinburgh, Edinburg, UK) [45] and rooted using *V. germanica*.

Table 2. List of single-copy orthologs that were annotated as complete and present in all eleven species used for phylogenetic analysis.

BUSCO ID	Description
EOG091G00GQ	thyroid hormone receptor interactor 12
EOG091G00L0	integrator complex subunit 1
EOG091G01TH	CSE1 chromosome segregation 1-like (yeast)
EOG091G01VH	Aminoacyl-tRNA synthetase, class Ia
EOG091G01WY	SCY1-like, kinase-like 2
EOG091G01XI	vacuolar protein sorting 18 homolog (S. cerevisiae)
EOG091G02C8	component of oligomeric golgi complex 1
EOG091G02LY	cleavage stimulation factor, 3′ pre-RNA, subunit 3, 77kDa
EOG091G02O6	Vacuole morphology and inheritance protein 14
EOG091G02P2	component of oligomeric golgi complex 4
EOG091G02V5	eukaryotic translation initiation factor 3, subunit B
EOG091G03AN	UFM1-specific ligase 1
EOG091G03I1	SAC1 suppressor of actin mutations 1-like (yeast)
EOG091G03JX	mediator complex subunit 17
EOG091G03M4	nucleolar protein 10
EOG091G03P0	component of oligomeric golgi complex 6
EOG091G03PD	Protein arginine N-methyltransferase
EOG091G03QC	mitochondrial translational initiation factor 2
EOG091G03RW	exocyst complex component 5
EOG091G03XO	tyrosyl-tRNA synthetase
EOG091G03Z1	asunder spermatogenesis regulator
EOG091G03ZS	Radical SAM
EOG091G04FV	SMG8 nonsense mediated mRNA decay factor
EOG091G04GK	PDZ domain containing 8
EOG091G04JK	neurochondrin
EOG091G04WH	negative elongation factor complex member B
EOG091G017G	Ribosomal protein S5 domain 2-type fold
EOG091G017T	N-acetyltransferase 10 (GCN5-related)
EOG091G024J	vacuolar protein sorting 53 homolog (S. cerevisiae)
EOG091G025T	Integrator complex subunit 2
EOG091G040D	methyltransferase like 13
EOG091G046H	USO1 vesicle transport factor
EOG091G049W	lin-9 DREAM MuvB core complex component
EOG091G0262	excision repair cross-complementation group 2
EOG091G0321	vacuolar protein sorting 51 homolog (S. cerevisiae)
EOG091G0349	exocyst complex component 8
EOG091G0495	zinc finger, RAN-binding domain containing 1
EOG091G0525	NOP9 nucleolar protein
EOG091G03Z1	asunder spermatogenesis regulator

Supplementary Materials: The following are available online at http://www.mdpi.com/2306-5729/5/4/112/s1, Table S1: List of 70 orthologous gene clusters exclusive to the Malaysian stingless bee, H. itama.

Author Contributions: Conceptualization, C.-Y.W., A.-A.T. and N.-H.N.; methodology, C.-Y.W., A.-A.T. and N.-H.N.; software, S.-M.N., J.-S.K. and N.-H.N.; validation, C.-Y.W., A.-A.T. and N.-H.N.; investigation, C.-Y.W., A.-A.T. and N.-H.N.; resources, A.-A.T. and R.J.; data curation, N.-H.N.; writing—original draft preparation, C.-Y.W., A.-A.T. and N.-H.N.; writing—review and editing, C.-Y.W., A.-A.T., N.-H.N., S.-M.N. and J.-S.K.; visualization, S.-M.N., J.-S.K., N.-H.N. and A.-A.T.; project administration, C.-Y.W. and R.J.; funding acquisition, C.-Y.W. and R.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by MARDI Development Fund, Ministry of Agriculture and Food Industry of Malaysia, grant number 21003004050001-7.2 and National Conservation Trust Fund (NCTF), Ministry of Water, Land and Natural Resources (KATS) of Malaysia, grant number KAT(S)600-2/1/48/2J.

Acknowledgments: We thank Muhammad Faris Mohd Radzi for assisting in stingless bee rearing and sampling as well as Ivan Hoh, Aeris Chow, Su-Boon Chua and Lisa Tho for technical support and helpful suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rasmussen, C. Catalog of the Indo-Malayan/Australasian stingless bees (Hymenoptera: Apidae: Meliponini). *Zootaxa* **2008**, *1935*, 1–80. [CrossRef]
2. Engel, M.S.; Rasmussen, C. A new subgenus of *Heterotrigona* from New Guinea (Hymenoptera: Apidae). *J. Melittology* **2017**, *73*, 1–16. [CrossRef]
3. Rasmussen, C.; Thomas, J.C.; Engel, M.S. A new genus of Eastern Hemisphere stingless bees (Hymenoptera, Apidae), with a key to the supraspecific groups of Indomalayan and Australasian Meliponini. *Am. Mus. Novit.* **2017**, *3888*, 1–33. [CrossRef]
4. Brummitt, R.K. *World Geographical Scheme for Recording Plant Distributions*, 2nd ed.; International Working Group on Taxonomic Databases for Plant Sciences (TDWG); Hunt Institute for Botanical Documentation, Carnegie Mellon University: Pittsburgh, UK, 2001; pp. 43–115.
5. Ghazalli, M.N.; Tamizi, A.A.; Talip, N. Impak Ekonomi Famili Sapindaceae. In *Debunga Rambutan Hutan (Sapindaceae) dan Kepentingan Taksonomi*; Jaafar, S., Ed.; UKM Press: Bangi, Malaysia; Universiti Kebangsaan Malaysia: Bangi, Malaysia, 2020; pp. 112–114.
6. Ahmad-Jailani, N.M.; Mustafa, S.; Mustafa, M.Z.; Mariatulqabtiah, A.R. Nest characteristics of stingless bee *Heterotrigona itama* (Hymenoptera: Apidae) upon colony transfer and splitting. *Pertanika J. Trop. Agric. Sci.* **2019**, *42*, 861–869.
7. Wong, P.; Hii, S.L.; Koh, C.C.; Moh, T.S.Y.; Anak Gindi, S.R. Chemical analysis on the honey of *Heterotrigona itama* and *Tetrigona binghami* from Sarawak, Malaysia. *Sains Malays.* **2019**, *48*, 1635–1642. [CrossRef]
8. Fahimee, J.; Nursyazwani, N.; Fairuz, K.; Rosliza, J.; Mispan, M.R.; Idris, A.B. Variation the oviposition behavior by the stingless bee, *Heterotrigona itama* (Hymenoptera, Apidae, Meliponini). *J. Asia-Pac. Entomol.* **2018**, *21*, 322–328. [CrossRef]
9. Md-Zaki, N.N.; Abd-Razak, S.B. Pollen profile by stingless bee (*Heterotrigona itama*) reared in rubber smallholding environment at Tepoh, Terengganu. *Malays. J. Microsc.* **2018**, *14*, 38–54.
10. Chen, X.; Hu, Y.; Zheng, H.; Cao, L.; Niu, D.; Yu, D.; Sun, Y.; Hu, S.; Hu, F. Transcriptome comparison between honey bee queen- and worker-destined larvae. *Insect Biochem. Mol. Biol.* **2012**, *42*, 665–673. [CrossRef]
11. Barchuk, A.R.; Cristino, A.S.; Kucharski, R.; Costa, L.F.; Simoes, Z.L.P.; Maleszka, R. Molecular determinants of caste differentiation in the highly eusocial honeybee *Apis mellifera*. *BMC Dev. Biol.* **2007**, *7*, 70. [CrossRef]
12. Evans, J.D.; Wheeler, D.E. Differential gene expression between developing queens and workers in the honey bee, *Apis mellifera*. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 5575–5580. [CrossRef]
13. Severson, D.W.; Williamson, J.L.; Aiken, J.M. Caste-specific transcription in the female honey bee. *Insect Biochem.* **1989**, *19*, 215–220. [CrossRef]
14. Tamizi, A.A.; Nazaruddin, N.H.; Yeong, W.C.; Mohd-Radzi, M.F.; Jaafar, M.A.; Sekeli, R. The first dataset of de novo transcriptome assembly of *Heterotrigona itama* (Apidae, Meliponinae) queen larva. *Data Brief* **2020**, *29*, 105235. [CrossRef] [PubMed]
15. de Paula-Freitas, F.C.; Lourenco, A.P.; Nunes, F.M.F.; Paschoal, A.R.; Abreu, F.C.P.; Barbin, F.O.; Bataglia, L.; Cardoso-Junior, C.A.M.; Cervoni, M.S.; Silva, S.R.; et al. The nuclear and mitochondrial genomes of *Frieseomelitta varia*-a highly eusocial stingless bee (Meliponini) with a permanently sterile worker caste. *BMC Genom.* **2020**, *21*, 386. [CrossRef] [PubMed]
16. Kapheim, K.M.; Pan, H.; Li, C.; Salzberg, S.L.; Puiu, D.; Magoc, T.; Robertson, H.M.; Hudson, M.E.; Venkat, A.; Fischman, B.J.; et al. Genomic signatures of evolutionary transitions from solitary to group living. *Science* **2015**, *348*, 1139–1143. [CrossRef] [PubMed]
17. Nowak, R.M.; Jastrzebski, J.P.; Kusmirek, W.; Salamatin, R.; Rydzanicz, M.; Sobczyk-Kopciol, A.; Sulima-Celinska, A.; Paukszto, L.; Makowczenko, K.G.; Płoski, R.; et al. Hybrid de novo whole-genome assembly and annotation of the model tapeworm *Hymenolepis diminuta*. *Sci. Data* **2019**, *6*, 302. [CrossRef] [PubMed]
18. Weinstock, G.M.; Robinson, G.E.; Gibbs, R.A.; Worley, K.C.; Evans, J.D.; Maleszka, R.; Robertson, H.M.; Weaver, D.B.; Beye, M.; Bork, P.; et al. Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* **2006**, *443*, 931–949. [CrossRef]
19. Sadd, B.M.; Barribeau, S.M.; Bloch, G.; de Graaf, D.C.; Dearden, P.; Elsik, C.G.; Gadau, J.; Grimmelikhuijzen, C.J.P.; Hasselmann, M.; Lozier, J.D.; et al. The genomes of two key bumblebee species with primitive eusocial organization. *Genome Biol.* **2015**, *16*, 76. [CrossRef]

20. O'Leary, N.A.; Wright, M.W.; Brister, J.R.; Ciufo, S.; Haddad, D.; McVeigh, R.; Rajput, B.; Robbertse, B.; Smith-White, B.; Ako-Adjei, D.; et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **2016**, *44*, D733–D745. [CrossRef]
21. Bairoch, A.; Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **2000**, *28*, 45–48. [CrossRef]
22. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
23. Bomtorin, A.D.; Mackert, A.; Rosa, G.C.C.; Moda, L.M.; Martins, J.R.; Bitondi, M.M.G.; Hartfelder, K.; Simoes, Z.L.P. Juvenile hormone biosynthesis gene expression in the *corpora allata* of Honey Bee (*Apis mellifera*) Female Castes. *PLoS ONE* **2014**, *9*, e86923. [CrossRef] [PubMed]
24. Murray, M.G.; Thompson, W.F. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Res.* **1980**, *8*, 4321–4325. [CrossRef] [PubMed]
25. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data. 2010. Available online: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 13 September 2019).
26. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]
27. Marcais, G.; Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **2011**, *27*, 764–770. [CrossRef]
28. Vurture, G.W.; Sedlazeck, F.J.; Nattestad, M.; Underwood, C.J.; Fang, H.; Gurtowski, J.; Schatz, M.C. GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics* **2017**, *33*, 2202–2204. [CrossRef]
29. Zerbino, D.R.; Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **2008**, *18*, 821–829. [CrossRef]
30. Gurevich, A.; Saveliev, V.; Vyahhi, N.; Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **2013**, *29*, 1072–1075. [CrossRef]
31. Simao, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [CrossRef]
32. Cantarel, B.L.; Korf, I.; Robb, S.M.; Parra, G.; Ross, E.; Moore, B.; Holt, C.; Sánchez-Alvarado, A.; Yandell, M. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **2008**, *18*, 188–196. [CrossRef]
33. Lowe, T.M.; Eddy, S.R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **1997**, *25*, 955–964. [CrossRef]
34. Lagesen, K.; Hallin, P.; Rodland, E.A.; Staefeldt, H.-H.; Rognes, T.; Ussery, D.W. RNAmmer: Consistent and rapid annotation of Ribosomal RNA genes. *Nucleic Acids Res.* **2007**, *35*, 3100–3108. [CrossRef] [PubMed]
35. Smit, A.F.A.; Hubley, R.; Green, P. RepeatMasker Open-4.0. 2013–2015. Available online: http://www.repeatmasker.org/ (accessed on 8 November 2019).
36. Buchfink, B.; Xie, C.; Huson, D.H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **2015**, *12*, 59–60. [CrossRef] [PubMed]
37. Kanehisa, M.; Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [CrossRef] [PubMed]
38. Conesa, A.; Götz, S.; García-Gómez, J.M.; Terol, J.; Talón, M.; Robles, M. Blast2GO: A universal tool for annotation, visualisation and analysis in functional genomics research. *Bioinformatics* **2005**, *21*, 3674–3676. [CrossRef] [PubMed]
39. Jones, P.; Binns, D.; Chang, H.-Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; et al. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **2014**, *30*, 1236–1240. [CrossRef] [PubMed]
40. Li, L.; Stoeckert, C.J.; Roos, D.S. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.* **2003**, *12*, 2178–2189. [CrossRef]
41. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef]

42. Huelsenbeck, J.P.; Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **2001**, *17*, 754–755. [CrossRef]
43. Ronquist, F.; Teslenko, M.; van der Mark, P.; Ayres, D.L.; Darling, A.; Höhna, S.; Larget, B.; Liu, L.; Suchard, M.A.; Huelsenbeck, J.P. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **2012**, *61*, 539–542. [CrossRef]
44. Ronquist, F.; Huelsenbeck, J.P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **2003**, *19*, 1572–1574. [CrossRef]
45. Rambaut, A. FigTree v1.4.4: A Graphical Viewer of Phylogenetic Trees. 2014. Available online: http://tree.bio.ed.ac.uk/software/figtree/ (accessed on 10 November 2020).