

Genome of the Blueberry Stem Gall Wasp

Submitted by

Joshua Vollmer

Biology

To

The Honors College

Oakland University

In partial fulfillment of the
requirement to graduate from

The Honors College

Mentor: Dr. Taras K. Oleksyk, Professor of Biological Sciences

Department of Biological Sciences

Oakland University

January 3, 2022

Abstract

The blueberry stem gall wasp, *Hemadas nubilipennis*, is a parasitic wasp species that lays its eggs within the shoots of blueberry plants, inducing the formation of galls. These galls provide larvae with nourishment and protect them from environmental threats, as well as reducing the crop yields of affected blueberry bushes. Blueberries are a major agricultural product in Michigan, so parasitism by blueberry stem gall wasps could have a significant economic impact on the region. The assembly of the *H. nubilipennis* genome may help to develop better pest management strategies. In this project we performed sequencing, size estimation and assembly of the blueberry stem gall wasp genome. First, the DNA of a single organism was extracted and sequenced using Illumina paired-end technology. Then, the data was used to estimate genome size using k-mer counting. Finally, the genome was assembled *de novo* using the SPAdes assembler. The genome size of *H. nubilipennis* was found to be approximately 1Gb, similar to other plant gall forming parasitic wasps, but larger than other closely related non-plant gall forming parasitic wasps. The trend of significantly larger genome size associated with the plant gall forming trait was observed in many other Hymenopteran parasitic wasps. Structural and functional annotation of the assembled reference genome produced in this study will allow for the identification of genes associated with the plant gall forming trait. Additionally, identification of cytochrome p450 genes may be a starting point for the study of insecticide resistance in the species.

Introduction

The blueberry stem gall wasp, *Hemadas nubilipennis*, parasitizes blueberry bushes as a means of reproduction. This species of wasp lays its eggs in the shoots of blueberry plants which results in the formation of a swollen tissue mass known as a gall. These galls protect vulnerable wasp larvae against environmental threats and provide them with nourishment important in their early development. Though it has been estimated that as many as 30,000 arthropod species express the gall forming phenotype, the mechanism by which plant systems are targeted to induce gall formation, in most cases, is not well understood. In oak galls, the larvae of the cynipid gall wasp, *Biorhiza pallida*, was found to differentially and highly express sets of genes encoding secretory peptides. Hearn et al. (2019) hypothesize that these secretory peptides are effector proteins that are transported to tissues surrounding larval deposits to induce an embryogenesis-like process in host tissues. Analysis of the proteomes of gall and non-gall host plant tissues parasitized by three oak gall wasp species indicated that the protein expression profile of gall tissues differs from that of non-gall tissue; the differentially regulated proteins were involved in metabolism, processing of genetic information, and cellular processes of the oak host (Pawlowski, Staszak, Karolewski, & Giertych, 2017). Analysis of the ovary and venom gland transcriptomes of two cynipid gall wasps, *Biorhiza pallida* and *Diplolepis rosae*, identified a set of highly expressed transcripts in maternal wasps likely related to the suppression of host plant defense signaling (Cambier et al., 2019). Thus, the parasitism of host plant tissues by gall wasps involves the manipulation of a wide range of host physiological mechanisms.

The outcome of parasitism by the blueberry stem gall wasp is a reduction in blueberry crop yields. Blueberries are a major agricultural product in Michigan; as of 2019, Michigan ranked fourth nationally in blueberry utilized production, outputting almost 85 million pounds of

blueberries (USDA, 2020). Current management of *H. nubilipennis* is achieved through pruning, which includes burning and mowing of blueberry fields (Hayman, Mackenzie & Reekie, 2003), and the use of foliar applied insecticides (Fanning & Isaacs, 2020). Although both of these methods have been shown to successfully reduce the impact of wasp parasitism, *H. nubilipennis* continues to threaten blueberry crop yields in Michigan. For more effective parasite control methods to be developed, a greater understanding of *H. nubilipennis* must be found through genomic analysis.

While the assembly of the *H. nubilipennis* genome had not previously been completed, the genomes of various other members of the Hymenoptera order have already been assembled and analyzed. Lindsey et al. (2018) sequenced the genome of *Trichogramma pretiosum*, a miniature parasitic wasp species, and used that data to place the species in a phylogenetic tree with 8 other hymenopteran species. Similarly, Chen and van Achterberg (2019) constructed a molecular phylogeny for braconid wasps using genetic markers from mitochondrial and nuclear genome sequencing. These genetic markers included 18S rDNA and 28S rDNA, sequences in the nuclear genome of eukaryotes that code for RNA components of cytosolic ribosomes. In addition to elucidating the evolutionary history of a species, genome construction allows for the analysis of gene family profiles within a species or group of species. As examined in a group of parasitic fig wasp species, a gene family involved in the production of chemosensory proteins was compared between species (Xin et al., 2020). The authors found that the gene family involved in chemical receptor production was upregulated in winged fig wasp morphs when compared to wingless fig wasp morphs.

The relatively moderate size and structure of most Hymenopteran genomes has allowed for the genome assembly of many species in the order, but the majority of the estimated 1 million Hymenopteran species' genomes still remain unsequenced (Branstetter et al., 2018). In fact, much of the extremely diverse Hymenopteran order remains unidentified, with only about 153,000 of the possible million members having been described at this point (Peters et al., 2017). The sequencing and assembly of the *H. nubilipennis* genome will indicate the species' relatedness to other insect species and more specifically, to other parasitic wasp species. The assembled reference genome could also facilitate further studies into variation within local populations of *H. nubilipennis*. Ultimately, this endeavor will result in a better understanding of these blueberry parasites and may allow for the development of better pest control practices.

Background

Sequencing Technologies

Since the development of Sanger sequencing in 1977, also known as the chain termination method, DNA sequencing has yielded massive insight into the molecular basis of biological function and activity. Sanger sequencing relies on four separate elongation aparati corresponding to the four types of fluorescently-labeled terminator nucleotides they contain (A,C,T, and G). After elongation, each aparati contains DNA fragments terminated at variable positions in the sequence that correspond to the nucleotide at that position. The fragments can then be electrophoretically separated to interrogate the complete sequence of the target DNA fragment. This produces relatively long reads (roughly 1 kilobase) with high accuracy base calls, and in combination with methods like shotgun sequencing, has enabled the assembly of some genomes (Heather & Chain, 2015). However, the shifting focus of genomics towards whole genome sequencing (WGS) in recent years has resulted in the development of a new set of sequencing techniques better equipped for large volume sequencing, collectively termed Next Generation Sequencing (NGS) technologies.

NGS technologies operate through a variety of methods that allow for high-throughput, massively parallel sequencing which is well equipped for sequencing large amounts of genetic data (Van Dijk et al., 2014). These characteristics, in addition to the cheaper cost per base allow for entire genomes to be sequenced several times in a single run. While NGS technologies tend to produce shorter reads than Sanger sequencing, they operate more rapidly, with less intensive oversight and experimental preparation, allowing for sequencing output adequate for coverage of

entire genomes many times over. One such NGS technology is Solexa sequencing employed by Illumina that relies on nucleotide terminators binding to the target sequence for base interrogation. However, an important distinction is that the terminators used in Solexa sequencing are reversibly bound and differentially dyed such that combinations of all fluorescent nucleotides can be washed over the target fragment simultaneously and step-wise interrogation of the sequence can occur. The Illumina sequencing used in our project uses bridge amplification of paired-end reads. This means that each read is sequenced from either end and the subsequent reads are paired, indicating that they both come from a single DNA fragment with a certain insert between them that remains unsequenced. The additional information provided by paired-end sequencing about the relationship between the two paired reads allows for easier and more successful assembly.

NGS technologies have become the dominant method for performing WGS while Sanger sequencing remains a useful method for examining shorter genomic intervals. NGS technologies do well to generate high coverage and base call accuracy, however, there have been recent advances in sequencing technology that help to solve the problems associated with short reads. These technologies, called third generation sequencing (TGS), rely on methods that can produce long reads, similar to Sanger sequencing, but with higher throughput (Lee et al., 2016). The longer read length allows for easier downstream processing in assembly and alignment, but there are flaws to the budding technologies that have prevented them from replacing NGS. TGS technologies tend to have higher base calling error rates than NGS and Sanger sequencing which has limited their usefulness in *de novo* assembly. Some genome assemblers, like SPAdes, have adapted to allow for the joint input of short reads from NGS and long reads from TGS, which in

combination produce high accuracy assemblies with fewer assembly gaps. The short reads do well to produce high accuracy contigs while the long reads often span regions of the genome that may be highly repetitive and predisposed to overcollapse with short read sequencing alone. Long read sequencing was not used in this project because we had a low quantity of DNA to begin with and it was unlikely that TGS long read sequencing would have produced enough sequencing data to generate a high quality assembly. Instead, we exploited the strength of NGS to generate high volumes of sequencing data from low amounts of genetic material.

Genome Size Estimation

Size is a fundamental characteristic of genomes and is one that exhibits a massive degree of variation. In eukaryotes alone, genome size varies by at least five orders of magnitude (Blommaert, 2020). Notably, it has been widely recognized that genome size does not vary directly with organismal complexity. Despite this, patterns arise with expanding genome size, including greater proportions accounted for by repetitive regions and transposable elements (Elliott & Gregory, 2015). Though differences in genome size are not a good predictor of the phenotypic features of an organism, the rapid expansion and contraction of genomes may induce genetic changes that are responsible for variation and even speciation. Also, the determination of an organism's or species' genome size has a practical use in planning genetic studies. When conducting a WGS study, it is important that enough sequencing data is produced such that adequate coverage of a genome is achieved. The estimation of genome size prior to a study can help to determine exactly how much time, money, and material will be required, and can help to predict its feasibility.

Traditionally, genome size estimation has been accomplished using flow cytometry, but recently, new computational methods that rely on sequencing data alone have been developed (Genome Size Database; <https://www.genomesize.com/>). Specifically, the k-mer counting method takes sequencing data and breaks it up into k-mers, where k is the length of read-derived subsequences (Hozza et al., 2015). If k is sufficiently large, an appearance of a certain k-mer is likely to come from a unique sequence rather than the repetition of the k-mer at different genomic locations due to random chance (Fig. 1). If a certain k-mer appears more than once in a sequencing data set, it can thus be considered the result of multiple coverage of the genome. K-mer counting algorithms allow us to find the number of unique k-mers in sequencing datasets and in doing so, approximates the size of the genome.

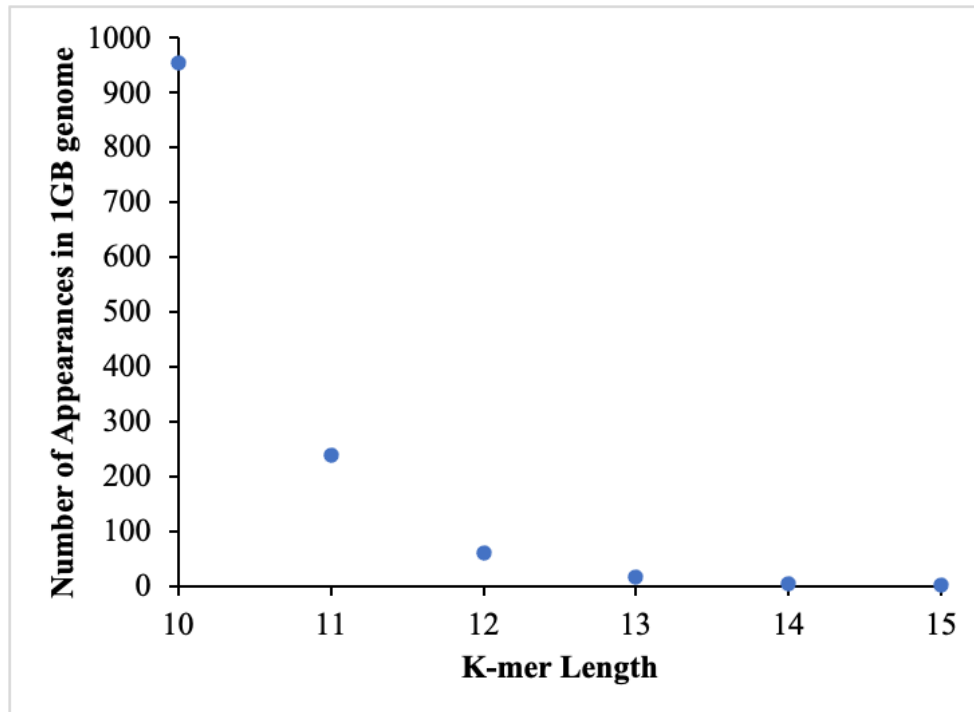


Figure 1. The number of appearances due to random chance in a 1GB genome of k-mers with variable length, k. The genome in question is assumed to have equal base content and sequence content generated by random base selection.

The k-mer counting method was used in this study due to the presence of high quality sequencing data and for its additional ability to generate estimates of coverage. The k-mer counting method for estimating genome size achieves estimates similar to the flow cytometry method. Pflung et al. (2020) suggested that k-mer methods in beetles tended to produce accurate estimates of genome size, but that the estimates were often slightly lower than those produced with flow cytometry. This trend may be the result of genomes that contain large proportions of repetitive sequences and so, for genomes with that characteristic, genome size validation by other means may be useful.

Genome Assembly

At its simplest, the field of computational genomics can be described as seeking an answer to two central problems: assembly and alignment. Assembly is the process of taking DNA sequencing reads, strings of letters corresponding to the base sequence at random locations of an organism's genome, and finding relationships between them. These relationships are the overlaps between the reads that indicate that two reads were taken from locations in the genome directly next to one another. If we imagine a genome as a puzzle, then the reads produced by sequencing are equivalent to the puzzle pieces. Our goal in assembly is to compare each puzzle piece to all others and find which fit next to each other. In an idealized situation, we would have a perfect set of puzzle pieces that would come together to form a complete puzzle, or a complete genome in the case of the analogy. However, in the real data that we use for assemblies, we do not have this idealized dataset. Instead, there are certain puzzle pieces that are missing, equivalent to incomplete coverage in which some genomic intervals are left uninterrogated by sequencing reads. Also, the puzzle pieces will have small differences with the picture that is

supposed to be depicted by the puzzle, equivalent to sequencing errors. These are some of the difficulties that come with genome assembly, and if we depart from the puzzle analogy to the real data and problem, we are confronted with more. Of course, the data associated with assembly is much greater than a simple puzzle and so a genome cannot be assembled in the manual manner in which a puzzle is. This is why we rely on automated computational methods for genome assembly and a number of different algorithmic approaches.

The Overlap-Layout-Consensus (OLC) method is perhaps the most intuitive approach to genome assembly, and as such, was one of the first developed (Li et al., 2011). First, an overlap graph is constructed in which each node is a read and each edge represents an overlap between reads with a particular weight. Once this graph contains all reads and edges, an overall layout of the manner in which the reads fit together is developed by walking through the overlap graph such that each node is visited. Finally, the consensus sequence is found by calling the base at each position in the layout according to the most common base at that position among all overlapped reads. While this method has aided in the assembly of some genomes, it is computationally inefficient. Therefore, it is not a good solution for WGS projects involving the large sequencing output associated with NGS technologies. The overlap graph grows directly with the number of reads considered and so, increased read depth becomes a hindrance in performing the assembly. For this reason, other assembly methods more equipped for NGS output have been found.

The De Bruijn graph approach to genome assembly is one method well-equipped for large sequencing projects (Compeau, 2011). With De Bruijn graph assembly, reads are first broken into k-mers of uniform length. New k-mers become the nodes of the graph and edges are

drawn between k-mers taken from positions immediately left and right of each other in a read. In this approach, the graph grows with the presence of unique k-mers rather than by each additional read, so large amounts of sequencing data remain computationally manageable.

For high coverage short read sequencing by NGS, the De Bruijn graph assembly method is preferable, while OLC may still be useful for low coverage long read sequencing (Li et al., 2011). Recently, as TGS technologies have become more prominent, new assembly methods have been developed that can be used in conjunction with existing methods. The string graph approach is one that may aid in closing assembly gaps that remain even in genomes that have already been thoroughly researched, like the human genome. The method is well suited to use the information contained in long reads to span highly repetitive genomic regions (Huang & Liao, 2016; Grigorev et al., 2018). In summary, different assembly methods perform best on different types of sequencing data and on genomes with different characteristics. In order to generate the highest quality assembly, a genome assembly method should be selected to suit the characteristics of the sequencing data and target genome.

Methods and Materials

DNA Extraction and Sequencing

DNA was extracted from single *H. nubilipennis* organisms using the QIAGEN DNeasy Blood and Tissue Kit (#69504, QIAGEN LLC, Maryland, USA). In order to account for the hard wasp exoskeletons, samples were crushed in microcentrifuge tubes until homogenization with Buffer ATL/Proteinase K mixture occurred. The extraction protocol was further modified by allowing lysis of tissue to occur for 2 hours, incubating lysis mixture in a water bath for at least 8 hours (overnight) at 56 °C, and diluting the DNA in a 80 µL solution of elution buffer. After extraction, the quality and quantity of DNA was evaluated using Qubit Fluorometer 3.0 (Thermo Fisher Scientific, Massachusetts, USA), IMPLN NanoPhotometer (IMPLEN, Munich, Germany), and gel electrophoresis on 1% agarose gel for high molecular weight genomic DNA. Measurements of quantity and quality were used to evaluate sample extractions and the best candidate was chosen for sequencing. The selected sequencing candidate was then sent to Psomagen Inc. (Geithersburg, MD) for Illumina whole genome sequencing with the HiSeq 2500 sequencer. The data was returned in fastq files containing paired-end reads and associated phred-encoded quality scores for each base call.

Sequencing Data Quality Control

The paired-end sequencing data obtained from Illumina NGS was first evaluated for quality using *FastQC* (Andrews, 2012). The output from *FastQC* indicated basic information about the number and length of reads obtained from sequencing, as well as calculated measurements of the sequencing data quality. The two fields that were of particular interest were average quality of base call per position of read and average base quality of whole reads. The

fastq files were subsequently trimmed using *Sickle* (Joshi & Fass, 2011) with a quality threshold of $Q > 25$ (equivalent to 99.7% base call accuracy). The trimmed reads were then once again evaluated with *FastQC* to determine the number of reads that passed quality control thresholds.

Genome Size Variation Analysis

Eight species of parasitic wasps with accessible assemblies were selected for comparative analysis of genome size and categorized according to their family and ability to form plant galls. Four species were selected from the *Pteromalidae* family – the family of *H. nubilipennis* – and four from the *Cynipidae* family, a Hymenopteran family containing gall forming and non-gall forming parasitic wasps. *PhyloT* was used to construct a cladogram of the selected species and the genus of *H. nubilipennis*, Hemadas, which was uploaded to *iTOL* for visualization (Letunic & Bork, 2021). Assemblies for these eight parasitic wasp species were accessed from NCBI and the genome sizes were recorded and compared for variation.

Genome Size Estimation

The trimmed reads were processed with *Jellyfish* (Marçais & Kingsford, 2011) using the count command to create a file containing the number of appearances for all k-mers of size 21. This file was then further processed by histo command within *Jellyfish* to generate a histogram file containing the number of k-mers associated with each frequency of appearance. This histogram file was then passed to *GenomeScope* (Vurture et al., 2017) to obtain an estimation of genome size.

Assembly

The genome assembly was conducted within the *Geneious Prime* software for data analysis (Geneious Prime 2022.0.1; <https://www.geneious.com>). The trimmed paired-end fastq files were uploaded into the geneious environment and merged into a single paired-end fastq file. The file was then digitally normalized to reduce file size and increase the efficiency of assembly. The assembly was then performed using the *SPAdes* assembler (Nurk et al., 2013) within the *Geneious Prime* software environment.

Results

DNA Extraction and Sequencing Output

The sample chosen for sequencing had an A260/280 ratio of 2.201, A260/230 ratio of 0.322 and concentration of 15.3 ng/μL on IMPLEN NanoPhotometer. The DNA concentration on the Qubit Fluorometer was found to be 7.9 ng/μL. The differences in DNA concentrations between testing methods was unexplained, but found in many of the extracted samples. The 1% gel electrophoresis found that extracted DNA was undegraded high molecular weight genomic DNA. The Illumina sequencing produced 392,596,630 paired-end reads of 151 base pair length. The forward and reverse reads had high average accuracy by base position (Fig. 2a-b) and by read (Fig. 3a-b). The percentage of GC content was 35% for both forward and reverse reads. After read trimming with *Sickle* to $Q > 25$, 96.3% (378,175,108/392,596,630) of reads passed the quality threshold.

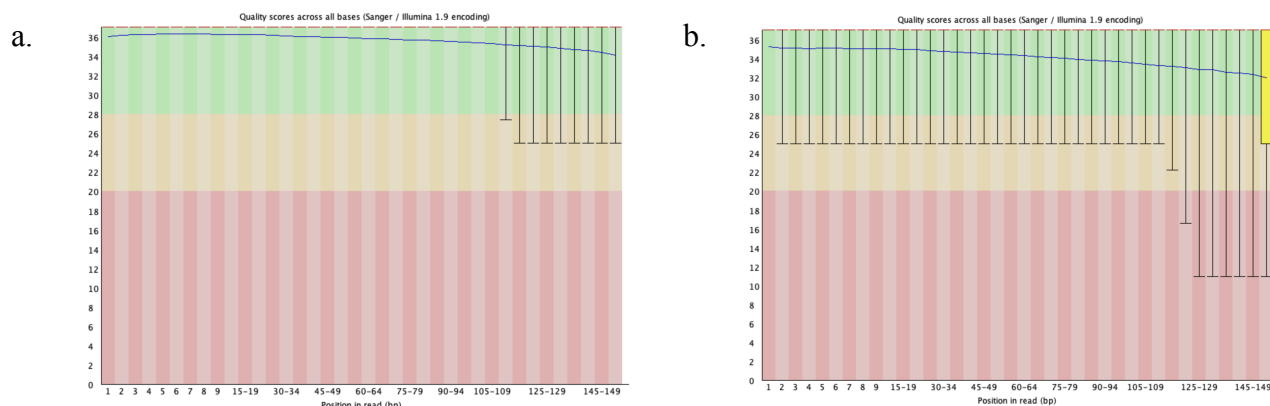


Figure 2. Per base sequence quality of forward (a) and reverse (b) reads determined by FastQC before trimming.

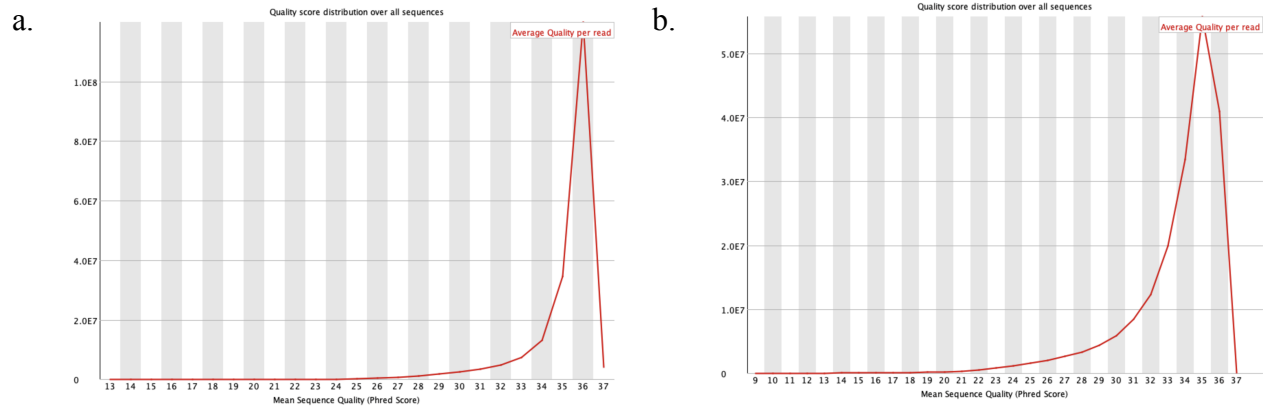


Figure 3. Mean sequence quality for forward (a) and reverse (b) reads determined by FastQC before trimming.

Genome Size

Genome size varied widely even among closely related parasitic wasp species. In this study, eight wasp species in the *Pteromalidae* and *Cynipidae* families were examined and found to have genome sizes ranging from ~200 megabases to 2 gigabases (Table 1). This variability highlighted the necessity to estimate the genome size of *H. nubilipennis* before assembly.

Table 1. Genome size of eight species of parasitic wasp in the *Pteromalidae* and *Cynipidae* families.

Species	Family	Genome Size (Mb)	GenBank Accession
<i>Synergus japonicus</i>	<i>Cynipidae</i>	226.414463	GCA_900474275.1
<i>Belonocnema treatae</i>	<i>Cynipidae</i>	1538.68626	GCA_010883055.1
<i>Andricus grossulariae</i>	<i>Cynipidae</i>	1412.75489	GCA_011634705.1
<i>Pseudoneuroterus saliens</i>	<i>Cynipidae</i>	2060.85117	GCA_900490055.1
<i>Apocrypta bakeri</i>	<i>Pteromalidae</i>	198.585306	GCA_018906985.1
<i>Nasonia vitripennis</i>	<i>Pteromalidae</i>	297.309692	GCA_009193385.2
<i>Trichomalopsis sarcophagae</i>	<i>Pteromalidae</i>	236.484274	GCA_002249905.1
<i>Pteromalus puparum</i>	<i>Pteromalidae</i>	338.170873	GCA_012977825.2

The relatedness of species was not a good predictor of genome size or the presence of the plant gall forming trait. The gall forming trait is present in a number of families in the

Hymenoptera order and many families contain species that both do and do not possess the trait (Fig. 4a). Among the species we examined from the *Cynipidae* family, three species, *B. treatae*, *A. grossulariae*, and *P. saliens*, were gall formers while *S. japonicus* was not. In the *Pteromalidae* family, all the species whose genome assembly was retrieved, *A. bakeri*, *N. vitripennis*, *T. sarcophagae*, and *P. puparum*, were non-gall formers, while *H. nubilipennis* was. Though genome size was not well predicted by relatedness between species, a trend was observed that may predict genome size. Plant gall formers tended to have much larger genomes than non-plant gall formers (Fig. 4b). This raised the question of whether the genome size of *H. nubilipennis* was on the order of a few hundred million base pairs like other members of the *Pteromalidae* family, or if it was similar to the size of other plant gall formers.

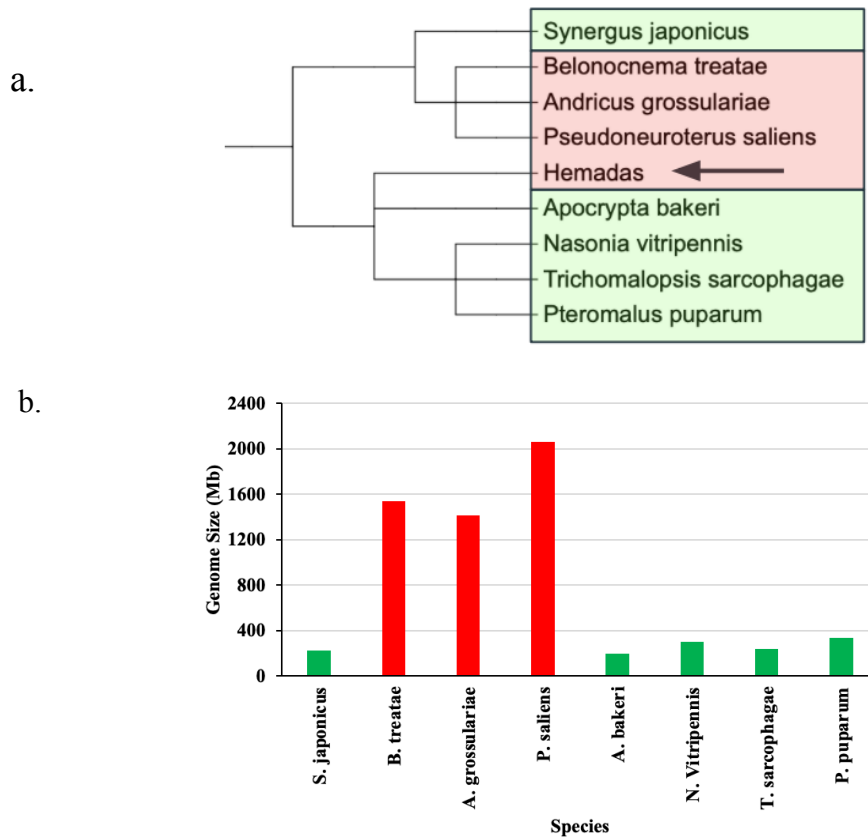


Figure 4. Cladogram of the eight parasitic wasp species and Hemadas, the genus of *H. nubilipennis*, indicated with arrow (a) and corresponding genome sizes (b) with plant gall formers highlighted in red and non-plant gall formers highlighted in green.

The genome size of *H. nubilipennis* was estimated to be 1,014,018,919 base pairs via the k-mer counting method (Fig. 5). The method also provided estimates of nearly 100% homozygosity which was expected due to the sequencing of haploid males, and sequencing coverage of roughly 40x. This 1GB genome size is significantly larger than those of other closely related *Pteromalidae* non-plant gall forming species (Fig. 4b). However, the size is approximately consistent with the genome size of other gall forming species in the *Cynipidae* family which ranged from 1.4 Gb to 2.1 Gb. This may indicate that the plant gall forming trait is associated with the large genomes seen in *Pteromalidae* and *Cynipidae*.

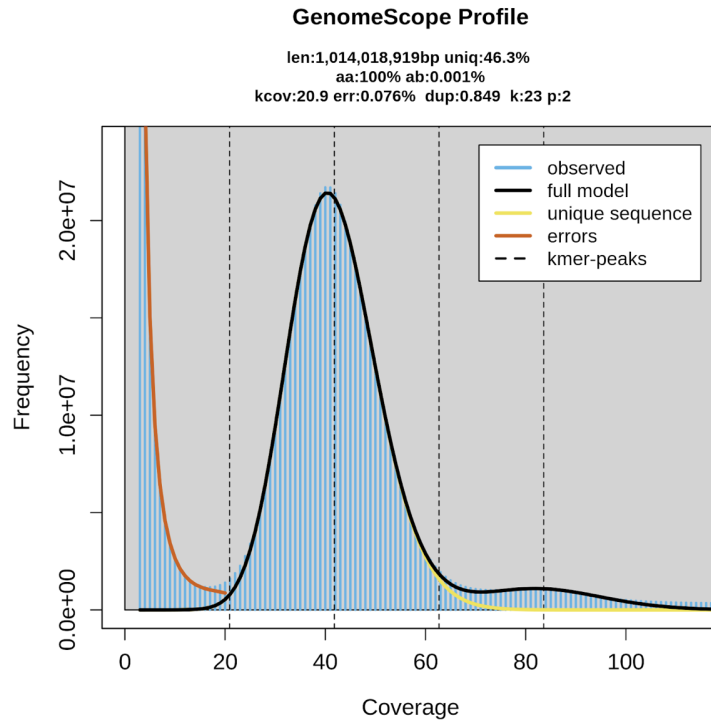


Figure 5. K-mer counting method output from *GenomeScope* (Vurture et al., 2017) with estimates of genome size, homozygosity, and coverage.

Genome Assembly

As of the submission of this thesis, the genome assembly of *H. nubilipennis* remains in-progress. The assembly of the large sequencing dataset requires extensive computational resources and additional data which requires more funds. Thus, the full results of the genome assembly could not be included in the results of this thesis. Upon the completion of the assembly, it will be evaluated for quality, annotated, and then further analyzed. However, with the limited sequencing data available, we could estimate the genome size of the *H. nubilipennis* and evaluate a specific hypothesis about the genome evolution of gall forming wasps.

Discussion

We found that considerable variation in genome size exists among Hymenopteran species, with even closely related species exhibiting major differences in genome size. In the *Cynipidae* family alone, genome size ranged from 226 megabases to 1,539 megabases among the four species examined in this study. *H. nubilipennis* and other plant gall formers had genomes that were significantly larger than related non-plant gall formers. The genome size of non-plant gall forming wasps in this study ranged from 199 megabases to 338 megabases and the genome size of plant gall forming wasps ranged from 1,413 megabases to 2,061 megabases. The genome size of *H. nubilipennis* was estimated to be 1,014 megabases which was less than but similar to that of other plant gall formers.

The variation in genome size within and between species is usually associated with differences in the content of repetitive sequences, including transposable elements, satellite DNA, and duplications (Biémont, 2008). The genome expansion associated with the plant gall forming trait is therefore likely associated with greater genome proportions of repetitive sequences. If high content of repetitive sequences is found in the genome of *H. nubilipennis*, analysis of the enriched regions may help to identify what genes are associated with the plant gall forming trait. Al-Qurainy (2021) found that differences in genome size estimation by flow cytometry and k-mer depth methods may be due to high repeat content and particularly long repetitive regions. The likely high content of repetitive sequences in the *H. nubilipennis* genome may necessitate verification of genome size by a non-computational method like flow cytometry. Our finding that *H. nubilipennis*' genome size was roughly 1 Gb, which was less than the range

found in other plant gall forming wasps (1.4 Gb to 2.1 Gb) may be the result of an underestimation by the k-mer depth method.

In this project, the DNA of *H. nubilipennis* was successfully extracted and was of high enough quality and quantity for comprehensive WGS sequencing and downstream analysis. Though this was true for the sample chosen for sequencing, the extraction of many other samples produced DNA quantities near or below the lower threshold for reliable WGS. The quantity of DNA used in sequencing has a significant effect on downstream analyses like alignment and assembly, with low DNA concentrations resulting in poorer analyses (Modi et al., 2021). Future studies may require extraction protocols that are optimized for small organisms with low initial DNA content. The sequencing output produced by Illumina NGS was of high quality and produced adequate coverage for *de novo* assembly. This demonstrates that Illumina sequencing is a reliable method for WGS of samples with relatively low DNA yield.

The immediate next step in the genomic analysis of *H. nubilipennis* is to structurally annotate the genome. Since RNA-seq data was not collected, this should be performed using an *ab initio* method that finds sequences in the assembled genome that are genes, promoters, and regulatory elements. *BRAKER* is an annotation software that can perform the structural annotation required (Hoff et al., 2019). Once these genomic elements have been identified by structural annotation, their function should be determined through functional annotation. *EggNOG-mapper* is a functional annotation software that does this by processing and comparing novel sequences with a database of orthologous sequences (Cantalapiedra, 2021). This method will allow the function of genic sequences in the assembled genome to be identified and for genetic analysis to proceed. Specifically, genotype-phenotype relationships in *H. nubilipennis* populations can be investigated.

Cytochrome p450 genes have been shown to play an integral role in the development of insecticide resistance in Hymenopteran species (Scott, 1999). We suggest that identification of cytochrome p450 genes in the functionally annotated *H. nubilipennis* genome could yield important information about the diversity of the gene family in the species. Additionally, sequencing of *H. nubilipennis* populations in Michigan could help to better understand the development of blueberry stem gall wasp insecticide resistance in the region. The results could then be used to inform effective insecticide resistant pest management strategies.

Conclusion

In this study, we identified that the genomes of plant gall forming parasitic wasps were larger than that of non-plant gall forming parasitic wasps. This trend predicted genome size more accurately than the relatedness of parasitic wasp species. Plant gall forming parasitic wasp genome size ranged from 1,413 megabases to 2,061 megabases and non-plant gall forming parasitic wasp genome size ranged from 199 megabases to 338 megabases. Through whole genome sequencing and k-mer-based genome size estimation, we found that the genome of *H. nubilipennis* was 1,014 megabases which was similar to other plant gall forming parasitic wasps. We then performed a *de novo* assembly of the *H. nubilipennis* using short read whole genome sequencing data and the SPAdes assembler. This assembly may be further annotated to identify functional genetic elements within the *H. nubilipennis* genome. Doing this may allow for future studies into genetic variation in local populations of *H. nubilipennis* and yield insight into the evolutionary history of parasitic wasp species. This may inform the future development of better pest control strategies for parasitic blueberry stem gall wasps.

References

- Al-Qurainy, F., Gaafar, A.-R. Z., Khan, S., Nadeem, M., Alshameri, A. M., Tarroum, M., Alansi, S., Almarri, N. B., & Alfarraj, N. S. (2021). Estimation of Genome Size in the Endemic Species *Reseda pentagyna* and the Locally Rare Species *Reseda lutea* Using comparative Analyses of Flow Cytometry and K-Mer Approaches. In *Plants* (Vol. 10, Issue 7, p. 1362). MDPI AG. <https://doi.org/10.3390/plants10071362>
- Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
- Biémont, C. (2008). Genome size evolution: Within-species variation in genome size. In *Heredity* (Vol. 101, Issue 4, pp. 297–298). Springer Science and Business Media LLC. <https://doi.org/10.1038/hdy.2008.80>
- Blommaert, J. (2020). Genome size evolution: towards new model systems for old questions. In *Proceedings of the Royal Society B: Biological Sciences* (Vol. 287, Issue 1933, p. 20201441). The Royal Society. <https://doi.org/10.1098/rspb.2020.1441>
- Branstetter, M. G., Childers, A. K., Cox-Foster, D., Hopper, K. R., Kapheim, K. M., Toth, A. L., & Worley, K. C. (2018). Genomes of the Hymenoptera. In *Current Opinion in Insect Science* (Vol. 25, pp. 65–75). Elsevier BV. <https://doi.org/10.1016/j.cois.2017.11.008>
- Cambier, S., Ginis, O., Moreau, S. J. M., Gayral, P., Hearn, J., Stone, G. N., Giron, D., Huguet, E., & Drezen, J.-M. (2019). Gall Wasp Transcriptomes Unravel Potential Effectors Involved in Molecular Dialogues With Oak and Rose. In *Frontiers in Physiology* (Vol. 10). Frontiers Media SA. <https://doi.org/10.3389/fphys.2019.00926>

- Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. In K. Tamura (Ed.), *Molecular Biology and Evolution*. Oxford University Press (OUP). <https://doi.org/10.1093/molbev/msab293>
- Chen, X., & van Achterberg, C. (2019). Systematics, Phylogeny, and Evolution of Braconid Wasps: 30 Years of Progress. In *Annual Review of Entomology* (Vol. 64, Issue 1, pp. 335–358). Annual Reviews. <https://doi.org/10.1146/annurev-ento-011118-111856>
- Compeau, P. E. C., Pevzner, P. A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. In *Nature Biotechnology* (Vol. 29, Issue 11, pp. 987–991). Springer Science and Business Media LLC. <https://doi.org/10.1038/nbt.2023>
- Elliott, T. A., & Gregory, T. R. (2015). What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 370, Issue 1678, p. 20140331). The Royal Society. <https://doi.org/10.1098/rstb.2014.0331>
- Fanning, P. D., & Isaacs, R. (2020). Efficacy of Foliar Applied Insecticides for Control of Blueberry Stem Gall Wasp, 2019. In J. Wise (Ed.), *Arthropod Management Tests* (Vol. 45, Issue 1). Oxford University Press (OUP). <https://doi.org/10.1093/amt/tsaa050>

- Grigorev, K., Kliver, S., Dobrynin, P., Komissarov, A., Wolfsberger, W., Krasheninnikova, K., Afanador-Hernández, Y. M., Brandt, A. L., Paulino, L. A., Carreras, R., Rodríguez, L. E., Núñez, A., Brandt, J. R., Silva, F., Hernández-Martich, J. D., Majeske, A. J., Antunes, A., Roca, A. L., O'Brien, S. J., ... Oleksyk, T. K. (2018). Innovative assembly strategy contributes to understanding the evolution and conservation genetics of the endangered *Solenodon paradoxus* from the island of Hispaniola. In *GigaScience* (Vol. 7, Issue 6). Oxford University Press (OUP). <https://doi.org/10.1093/gigascience/giy025>
- Hayman, D. I., MacKenzie, K. E., & Reekie, E. G. (2003). The Influence of Pruning on Wasp Inhabitants of Galls Induced by *Hemadas nubilipennis* Ashmead (Hymenoptera: Pteromalidae) on Lowbush Blueberry. In *Journal of Economic Entomology* (Vol. 96, Issue 4, pp. 1245–1253). Oxford University Press (OUP). <https://doi.org/10.1093/jee/96.4.1245>
- Hearn, J., Blaxter, M., Schönrogge, K., Nieves-Aldrey, J.-L., Pujade-Villar, J., Huguet, E., Drezen, J.-M., Shorthouse, J. D., & Stone, G. N. (2019). Genomic dissection of an extended phenotype: Oak galling by a cynipid gall wasp. In G. P. Copenhaver (Ed.), *PLOS Genetics* (Vol. 15, Issue 11, p. e1008398). Public Library of Science (PLOS). <https://doi.org/10.1371/journal.pgen.1008398>
- Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. In *Genomics* (Vol. 107, Issue 1, pp. 1–8). Elsevier BV. <https://doi.org/10.1016/j.ygeno.2015.11.003>

- Hoff, K. J., Lomsadze, A., Borodovsky, M., & Stanke, M. (2019). Whole-Genome Annotation with BRAKER. In *Methods in Molecular Biology* (pp. 65–95). Springer New York. https://doi.org/10.1007/978-1-4939-9173-0_5
- Hozza, M., Vinař, T., & Brejová, B. (2015). How Big is that Genome? Estimating Genome Size and Coverage from k-mer Abundance Spectra. In *String Processing and Information Retrieval* (pp. 199–209). Springer International Publishing. https://doi.org/10.1007/978-3-319-23826-5_20
- Huang, Y.-T., & Liao, C.-F. (2016). Integration of string and de Bruijn graphs for genome assembly. In *Bioinformatics* (Vol. 32, Issue 9, pp. 1301–1307). Oxford University Press (OUP). <https://doi.org/10.1093/bioinformatics/btw011>
- Joshi NA, Fass JN. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at <https://github.com/najoshi/sickle>.
- Lee, H., Gurtowski, J., Yoo, S., Nattestad, M., Marcus, S., Goodwin, S., Richard McCombie, W., & Schatz, M. C. (2016). Third-generation sequencing and the future of genomics. Cold Spring Harbor Laboratory. <https://doi.org/10.1101/048603>
- Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. In *Nucleic Acids Research* (Vol. 49, Issue W1, pp. W293–W296). Oxford University Press (OUP). <https://doi.org/10.1093/nar/gkab301>

- Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B., & Fan, W. (2011). Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. In *Briefings in Functional Genomics* (Vol. 11, Issue 1, pp. 25–37). Oxford University Press (OUP). <https://doi.org/10.1093/bfgp/elr035>
- Lindsey, A., Kelkar, Y., Wu, X., Sun, D., Martinson, E., & Yan, Z. et al. (2018). Comparative genomics of the miniature wasp and pest control agent *Trichogramma pretiosum*. *BMC Biology*, 16(1). <https://doi.org/10.1186/s12915-018-0520-9>
- Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. In *Bioinformatics* (Vol. 27, Issue 6, pp. 764–770). Oxford University Press (OUP). <https://doi.org/10.1093/bioinformatics/btr011>
- Modi, A., Vergata, C., Zilli, C., Vischioni, C., Vai, S., Tagliazucchi, G. M., Lari, M., Caramelli, D., & Taccioli, C. (2021). Successful extraction of insect DNA from recent copal inclusions: limits and perspectives. In *Scientific Reports* (Vol. 11, Issue 1). Springer Science and Business Media LLC. <https://doi.org/10.1038/s41598-021-86058-9>
- Nurk, S., Bankevich, A., Antipov, D., Gurevich, A., Korobeynikov, A., Lapidus, A., Prjibelsky, A., Pyshkin, A., Sirotkin, A., Sirotkin, Y., Stepanauskas, R., McLean, J., Lasken, R., Clingenpeel, S. R., Woyke, T., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2013). Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads. In *Lecture Notes in Computer Science* (pp. 158–170). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37195-0_13

- Pawłowski, T. A., Staszak, A. M., Karolewski, P., & Giertych, M. J. (2017). Plant development reprogramming by cynipid gall wasp: proteomic analysis. In *Acta Physiologiae Plantarum* (Vol. 39, Issue 5). <https://doi.org/10.1007/s11738-017-2414-9>
- Peters, R. S., Krogmann, L., Mayer, C., Donath, A., Gunkel, S., Meusemann, K., Kozlov, A., Podsiadlowski, L., Petersen, M., Lanfear, R., Diez, P. A., Heraty, J., Kjer, K. M., Klopstein, S., Meier, R., Polidori, C., Schmitt, T., Liu, S., Zhou, X., Niehuis, O. (2017). Evolutionary History of the Hymenoptera. In *Current Biology* (Vol. 27, Issue 7, pp. 1013–1018). <https://doi.org/10.1016/j.cub.2017.01.027>
- Pflug, J. M., Holmes, V. R., Burrus, C., Johnston, J. S., & Maddison, D. R. (2020). Measuring Genome Sizes Using Read-Depth, k-mers, and Flow Cytometry: Methodological Comparisons in Beetles (Coleoptera). In *G3 & Genes|Genomes|Genetics* (Vol. 10, Issue 9, pp. 3047–3060). Oxford University Press (OUP). <https://doi.org/10.1534/g3.120.401028>
- Scott, J. G. (1999). Cytochromes P450 and insecticide resistance. In *Insect Biochemistry and Molecular Biology* (Vol. 29, Issue 9, pp. 757–777). Elsevier BV. [https://doi.org/10.1016/s0965-1748\(99\)00038-7](https://doi.org/10.1016/s0965-1748(99)00038-7)
- United States of America. United States Department of Agriculture. (2020). *2019 blueberry statistics*. National Agricultural Statistics Service.
- van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. In *Trends in Genetics* (Vol. 30, Issue 9, pp. 418–426). Elsevier BV. <https://doi.org/10.1016/j.tig.2014.07.001>

- Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. In B. Berger (Ed.), *Bioinformatics* (Vol. 33, Issue 14, pp. 2202–2204). Oxford University Press (OUP). <https://doi.org/10.1093/bioinformatics/btx153>
- Xin, Z., Huang, D., Zhao, D., Li, J., Wei, X., & Xiao, J. (2020). Genome-Wide Analysis of Chemosensory Protein Genes (CSPs) Family in Fig Wasps (Hymenoptera, Chalcidoidea). In *Genes* (Vol. 11, Issue 10, p. 1149). MDPI AG. <https://doi.org/10.3390/genes11101149>