# Hybrid *de novo* Assembly and Functional Analysis of Novel Date Palm Menace: Dubas Bug (*Ommatissus lybicus*)

Rajesh Chikatla, Mercedes Trevino, Angela Young
Dr. Abdul Latif Khan, PhD

UNIVERSITY of HOUSTON
Department of Engineering Technology
College of Technology

## Abstract

The Dubas bug (*Ommatissus lybicus)* is a sap feeding insect found in the Middle East and North African regions. They are most known for sucking the sap from date palms which are culturally and economically important to countries in these regions and secreting honeydew that promotes growth of black sooty mold. The Dubas bug infests date palms by laying their eggs into holes of the tissue in date palm fronds which causes chlorosis. The high infestation levels have led to destruction of palm plantations and attempts for extermination or reduction of the Dubas bug population. Despite these countries' best efforts, their methods have caused negative environmental impacts to non-target species and human health. We have been given the genomic sequence of the Dubas bug.

We used different bioinformatic tools to assemble and annotate this sequence for key genes that may play a role in insecticide resistance. This could also lead to tailoring of more suitable insecticides against the Dubas bug. By using command line bioinformatic tools such as FastQC and Trimmomatic PE, we were able to check the quality of our sequences and trim them accordingly. We were then able to create a hybrid assembly using SPAdes. To check the quality of our assembly we used BUSCO and to predict genes we used Augustus. For further understanding of the species, KEGG pathways and other functional analysis was done.

## Background

*Ommatissus lybicus*, commonly known as the Dubas bug, is a sap feeding insect that affects date palms in the Middle East and North African regions. These insects live in the mountain wadi biomes and near fresh water; they avoid extreme temperatures and direct sunlight. They are mostly known for sucking sap from date palms and secreting honeydew which promotes the growth of black sooty mold [1]. Numerous countries have tried to eliminate these Dubas bug infestations with insecticides but have been unsuccessful due to their negative environmental impacts to non-target species and human health. The ability to characterize the Dubas bug genome through bioinformatic tools allows for more accurate insights into the molecular mechanisms underlying the biological processes of *O. lybicus*. Annotated genomic sequences can help uncover essential genetic elements that have been key in *O. lybicus* resistance to current insecticides [2]. With the development of second-generation sequencing technologies like Illumina, the quality of whole genome sequencing has greatly evolved. However, these short Illumina reads on their own still lead to poor *de novo* assembly. Hence the development of long-sequencing reads, Oxford Nanopore, which helps with analyzing highly repetitive elements. By combining both short and long reads to conduct a hybrid assembly, we were able to resolve any gaps in the sequence.

**Figure 1. *Ommatissus lybicus*, commonly known as the Dubas bug.**

The taxonomy of *O. lybicus* showed that it is part of the Tropiduchidae family of plant hoppers which consists of 39 descendants. The *Ommatissus* genus consists of two other species: *O. lofouensis* and *O. binotatus*. Unfortunately, there were no genes or genome assemblies available for comparison in either the family or genus of *O. lybicus*. For this reason, any reference sequences we used were from the order Hemiptera, or more broadly the Arthropoda phylum. Through genome annotation, we are identifying the location of genes and coding regions to determine their roles.

| Entrez records | |
|---|---|
| Database name | Direct links |
| Nucleotide | 125 |
| Protein | 110 |
| Popset | 6 |
| PubMed Central | 11 |
| Identical Protein Groups | 42 |
| Taxonomy | 1 |

**Figure 2. NCBI Entrez records for *O. lybicus.***

Using Augustus as a gene prediction tool allowed us do an *Ab-initio* and homology-based gene prediction. By using the predicted genes from Augustus to do a KEGG annotation, we were able to identify what genes from the Dubas bug were associated with respective biosynthetic pathways provided by the reconstructive maps of KEGG mapping. The analysis of the assembled and annotated genome will help develop a better understanding of the insect which can be largely helpful in developing more tailored methods of controlling and managing *O. lybicus* infestations.

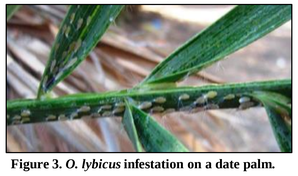

**Figure 3. *O. lybicus* infestation on a date palm.**

**Core Objectives:**

- Research and understand the impact Dubas Bug infections have on date palms
- Hybrid *de novo* genome assembly using bioinformatic tools
- Understand how to verify quality of a genomic assembly
- Genomic annotation using an *ab-initio* approach
- Functional analysis of select proteins and obtaining a better understanding of significant biological pathways

## Methodology

For the assembly workflow, we created both an Illumina-only assembly as well as a hybrid assembly. These assemblies were then compared to a previously prepared sample hybrid assembly. Raw Illumina reads from the Dubas Bug were submitted to FastQC, a quality assessment program, and quality histograms showed that trimming of the reads could be performed for a better assembly. Reads were then submitted to Trimmomatic, a pair-aware trimming program, and ran through FastQC again for comparison with the untrimmed reads. After the reads were determined to be of satisfactory quality, genome assembly could begin. After consideration between various programs, we used the SPAdes genome assembly pipeline for our assembly, which completed successfully. The resulting scaffolds file created by SPAdes was then checked for quality control and completeness through QUAST and Busco analysis, respectively. We then built the hybrid assembly using both the Illumina reads and Nanopore reads.

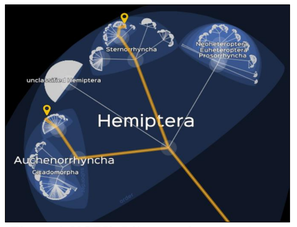


**Figure 4. NCBI's Lifemap picturing the Hemiptera order that *O. lybicus and A. pisum* are both in.**

The same strategy used for Illumina assembly was used for hybrid assembly, albeit with the inclusion of Nanopore reads in SPAdes. An assembly was successfully built, along with a scaffolds file that was also checked for quality control much like the Illumina assembly. Following assembly, we used Augustus to predict genes in the *O. lybicus* sequence. *A. pisum*, a species related on the order level (Hemiptera) was used as the reference genome. After successfully predicting genes with Augustus, we chose any genes that were given a score of 1 to BLAST. Our determination for genes to be annotated with KEGG was they needed to have an alignment score between 50 and 80. The KEGG annotations showed what genes were associated with biosynthetic pathways and provided reconstructive maps of the respective pathways.

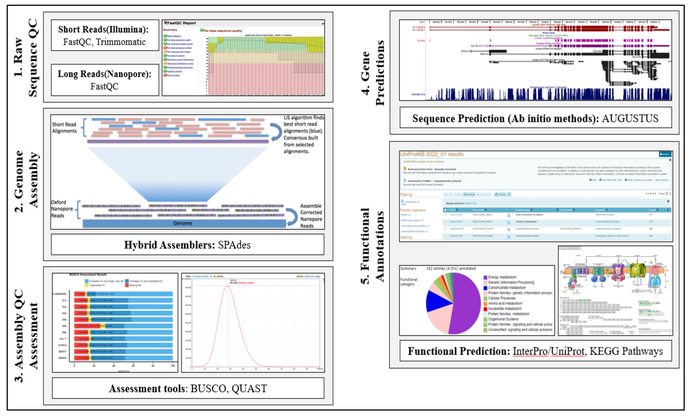


**Figure 5. The methodology used to conduct *O. lybicus* genome assembly and annotation.**

## Results

After trimming the sequences using Trimmomatic PE and checking the quality with FastQC, the following graphs (figures 6a, 6b, 7a, 7b) were given. Quality of the forward trimmed sequence was good, but the reverse trimmed read had a few reads with a quality score lower than 28 (below the green region in figures 6a, 6b, 7a, and 7b). It was not trimmed further as it already dropped 1 GB of the entire sequence.

**Figure 6a. Before trimming forward read of Illumina sequence.**

**Figure 6b. After trimming forward read of Illumina sequence.**

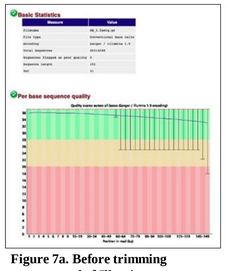

**Figure 7a. Before trimming reverse read of Illumina sequence.**

**Figure 7b. After trimming reverse read of Illumina sequence.**

QUAST analysis on the Illumina assembly showed the assembly was satisfactory, with an N50 value of 5,182, and no mismatches. Analysis on the hybrid assembly (figure 8a), however, had a significant number of mismatches, with nearly 30,000 N's or about 215 N's per 100kbp. Additionally, BUSCO analysis on both assemblies showed about 30% completeness for the Illumina assembly, while the hybrid assembly had an even lower completeness percentage at only 1%. For these reasons, we conducted annotation using the previously assembled an alternative hybrid assembly.

QUAST analysis of the hybrid assembly (figure 8b) showed 19,814 contigs with an N50 value of 15,125 and no mismatches. Annotations were successfully completed with Augustus. According to KEGG pathways analysis, there were 69 matches for metabolic pathways and 23 matches for biosynthesis of secondary metabolites. Described are biological pathways with strong contig associations and suspected relevance to the mechanism in which *O. lybicus* infects date palms.

| Statistics without reference | scaffolds |
|---|---|
| # contigs | 3256 |
| # contigs (>= 0 bp) | 28 752 |
| # contigs (>= 1000 bp) | 2735 |
| # contigs (>= 5000 bp) | 1021 |
| # contigs (>= 10000 bp) | 321 |
| # contigs (>= 25000 bp) | 16 |
| # contigs (>= 50000 bp) | 2 |
| Largest contig | 112 795 |
| Total length | 14 221 540 |
| Total length (>= 0 bp) | 17 531 148 |
| Total length (>= 1000 bp) | 13 828 314 |
| Total length (>= 5000 bp) | 9 751 420 |
| Total length (>= 10000 bp) | 4 708 548 |
| Total length (>= 25000 bp) | 615 281 |
| Total length (>= 50000 bp) | 170 877 |
| N50 | 7500 |
| N90 | 1862 |
| L50 | 599 |
| L90 | 1991 |
| GC (%) | 32.13 |
| **Mismatches** | |
| # N's per 100 kbp | 215.74 |
| # N's | 30 681 |

**Figure 8a. Quast analysis of hybrid assembly**

| Statistics without reference | finalhybridNPI |
|---|---|
| # contigs | 19 814 |
| # contigs (>= 0 bp) | 45 354 |
| # contigs (>= 1000 bp) | 9529 |
| # contigs (>= 5000 bp) | 4766 |
| # contigs (>= 10000 bp) | 2760 |
| # contigs (>= 25000 bp) | 671 |
| # contigs (>= 50000 bp) | 96 |
| Largest contig | 130 647 |
| Total length | 90 535 597 |
| Total length (>= 0 bp) | 98 618 027 |
| Total length (>= 1000 bp) | 83 664 718 |
| Total length (>= 5000 bp) | 72 619 771 |
| Total length (>= 10000 bp) | 58 331 851 |
| Total length (>= 25000 bp) | 25 542 372 |
| Total length (>= 50000 bp) | 6 495 931 |
| N50 | 15 125 |
| N90 | 1503 |
| L50 | 1694 |
| L90 | 7710 |
| GC (%) | 33.4 |
| **Mismatches** | |
| # N's per 100 kbp | 0 |
| # N's | 0 |

**Figure 8b: Quast analysis of hybrid assembly**

Several contigs corresponded to the Hypoxia-inducible factor 1 (HIF-1) pathway indicating that it is highly regulated. Under normal conditions, HIF-1 alpha undergoes hydroxylation at specific prolyl residues leading to immediate ubiquitination followed by proteasomal degradation of the subunit. However, under hypoxia, the alpha subunit becomes stable and interacts with coactivators like p300/CBP to modulate transcriptional activity regulate a collection of hypoxia-inducible genes. The target genes of HIF-1 encode proteins that increase O2 delivery and mediate adaptive responses to O2 deprivation, or lacking nitric oxide, or various growth factors [5]. PDH is pyruvate dehydrogenase E1 component alpha subunit, HK is a hexokinase, GAPDH is glyceraldehyde 3-phosphate dehydrogenase (phosphorylating), and ALDOA is fructose-bisphosphate aldolase, class I. Under conditions in which oxygen consumption is limited, these components of the sequence aid in promoting anaerobic metabolism. Due to the hot, dry nature of the Middle east, it is an adaptation that *O. lybicus* has developed to combat it. Elevated temperatures often induce oxidative stress and antioxidant response in insects like *O. lybicus* [6].

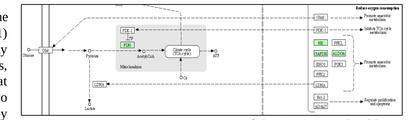

**Figure 9. Reconstructive KEGG map of the Hypoxia-inducible factor 1 (HIF-1) pathway of *O. lybicus***

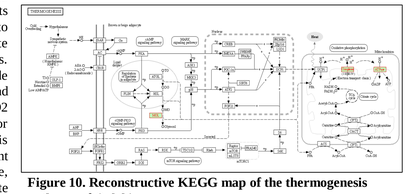

**Figure 10. Reconstructive KEGG map of the thermogenesis pathway of *O. lybicus***

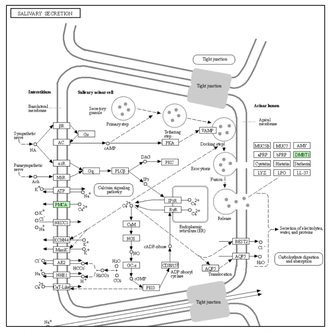


**Figure 11. Reconstructive KEGG map of the salivary secretion pathway of *O. lybicus*.**

Conversely the opposite is also true – *O. lybicus* requires thermogenesis to maintain species population. It has been shown that DBs typically hatch in two batches: once around late April and another time around late September. The most ideal temperature for these biological activities is around 27.5°C [7].

The two matches for salivary secretion obtained from our KEGG pathway results were DMBT1 (deleted in malignant brain tumors 1 protein) and ATP2B (P-type Ca2+ transporter type 2B). Salivary secretion occurs in response to neurotransmitters stimulated from autonomic nerves. Modifications to the salivary secretion pathway may help combat how the pest sucks the sap from date palms.

While feeding on date palms, the insect excretes honeydew that contains sugars and other waste contributing to the development of pathogenic infections [4]. Further studies of *O. lybicus* waste is necessary to pinpoint what is specifically responsible for the decay, but the genes associated to the proteins mentioned serve as a foundational understanding for which proteins are likely relevant and are most prolifically regulated in *O. lybicus.*

## Conclusion

We were successful in assembling an Illumina and hybrid (Illumina + Nanopore) genome sequence. However, due to time constraints and computer storage errors, we were unable to obtain a high-quality hybrid assembly. An alternative hybrid assembly was used for annotation and functional analysis. By utilizing KEGG pathways, we were able to identify key genes in the *O. lybicus* that play a role in metabolism and pathways associated with producing waste known to be detrimental to date palm photosynthesis. By further studying these genes and associated pathways, we can better understand how this organism works metabolically and how it secretes honeydew promoting the growth of a black sooty mold.

In the future, this annotated genome can be used as a resource to isolate key functional components of the *O. lybicus* genome. Targeting the insect's immune response, metabolic pathways, or salivary secretion may be suitable methods to combat infestations and reinvigorate the date palm industry of impacted Middle East countries.

## References

[1] El-Juhany, L. I. (2010). Degradation of Date Palm Trees and Date Production in Arab Countries: Causes and Potential Rehabilitation. *Australian Journal of Basic and Applied Sciences*, *4*, 3998–4010.

[2] Blumberg, D. (2008). Review: Date palm arthropod pests and their management in Israel. *Phytoparasitica*, *36*(5), 411–448. https://doi.org/10.1007/BF03020290

[3] Chen, Z., Erickson, D. L., & Meng, J. (2020). Benchmarking Hybrid Assembly approaches for genomic analyses of bacterial pathogens using Illumina and Oxford Nanopore sequencing. *BMC Genomics*, *21*(1). https://doi.org/10.1186/s12864-020-07041-8

[4] Khan, A. L., Asaf, S., Khan, A., Khan, A., Imran, M., Al-Harrasi, A., … Al-Rawahi, A. (2020). Transcriptomic analysis of Dubas bug (Ommatissus lybicus Bergevin) infestation to Date Palm. *Scientific Reports*, *10*(1), 11505. https://doi.org/10.1038/s41598-020-67438-z

[5] Salceda, S., & Caro, J. (1997). Hypoxia-inducible factor 1alpha (HIF-1alpha) protein is rapidly degraded by the ubiquitin proteasome system under normoxic conditions. Its stabilization by hypoxia depends on redox-induced changes. The Journal of Biological Chemistry, 272(36), 22642–22647. https://doi.org/10.1074/jbc.272.36.22642

[6] Ju, R.-T., Wei, H.-P., Wang, F., Zhou, X.-H., & Li, B. (2014). Anaerobic respiration and antioxidant responses of Corythucha ciliata (Say) adults to heat-induced oxidative stress under laboratory and field conditions. Cell Stress & Chaperones, 19(2), 255–262. https://doi.org/10.1007/s12192-013-0451-x

[7] Al-Kindi, K. M., Kwan, P., Andrew, N., & Welch, M. (2017). Impact of environmental variables on Dubas bug infestation rate: A case study from the Sultanate of Oman. *PLOS ONE*, *12*(5), e0178109. https://doi.org/10.1371/journal.pone.0178109