Diverse Collections

**INTRODUCTION**

**RESULTS**

**Galbut virus clades A and B circulates within the same populations**

To understand the overall diversity of galbut virus infection in *D. melanogaster* we collected flies from eight locations in Larimer County, CO and one location each in Maine, Pennsylvania and Ohio (**Fig. 1A**). Collections of flies in a Larimer County were done using buckets baited with banana and yeast. Flies collected from Maine, Pennsylvania and Ohio were carried out by volunteers using a previously designed 3D printed fly trap [Food Paper]. Volunteers were instructed to collect from these traps for as long as possible before sending the flies to our lab. During the collection period ~1000 flies were collected across all locations with most of the flies collected from Colorado.

RNA was extracted from individual flies, and each was screened for galbut virus via RT-qPCR. Two sets of primers were used to differentiate between galbut virus clade A and clade B (**Fig. 1B & 1C**). Ribosomal protein L32 was used to normalize galbut virus levels. Galbut virus clade A was the most abundant in all locations sampled. However, clade B was still present in every population tested. In all locations except those with minimal sample sizes (Pennsylvania & Ohio) there were low and high infected flies, low positives were indicated by high RT-qPCR Ct value and high positives had low Ct values (**Fig. 1B & 1C**). This genotype has been documented in wild caught populations maintained in our lab, but the effect of this difference biologically is still unknown [Tillie paper?]

Three locations were sampled over two or more months allowing us to albeit, limitedly look at galbut virus infection over time. Overall, galbut virus clade A was more abundant than clade B. However, the samples from CVID showed a shift in which clade was more abundant over time (**Fig. 1D**). In August, clade B was most prevalent with all, but two samples identified as clade B. In September, the prevalence of A increased to roughly the same as clade B [STATS]. By October, the ratios of clade A to clade B had increased to clade A with only one sample identified as clade B. This pattern was not seen in the other two locations where clade A was identified at much higher rates than clade B at all time points (**Fig. 1D**).

A screenshot of a screen shot of a map

Description automatically generated

**Figure 1: Galbut virus clades A and B cocirculate among the same populations**. (A) Locations in the United States that wild *Drosophila* species were collected in 2023. Galbut virus positive screening of flies collected in Larimer County, CO (B) or in other regions of the United States (C). (D) Galbut virus infection over time in three locations of Larimer County, CO. RT-qPCR data is normalized to levels of RpL32 mRNA. Galbut virus genotype is denoted by shape and color.

**Galbut virus genotype A is associated with variable infectivity levels and chaq virus coinfection**

Of the ~1,000 samples collected, 155 were selected for downstream analysis via total RNA metagenomic sequencing. Samples were selected based on: 1) genotype- as many B samples as possible to increase representation of genotype B sequences; 2) samples that had high infection levels based on low ct values; and 3) samples that had low infection levels determined by high ct values. Thirty-two of the sequenced samples did not contain enough galbut virus reads to generate any complete or near complete sequences. However, of the 123 remaining samples, 374 complete or very near complete sequences were identified. Of the samples that yielded no galbut virus sequences, six were from Ohio and sequenced because they were not *D. melanogaster* but could have been infected by a virus similar to galbut virus, 22 were low positives, three were galbut virus genotype B and one was genotype A which was positive by our only genotype A detecting primers but not by the genotype A and B detecting primers.

We first wanted to compare the sequencing results with the RT-qPCR data. Samples identified as genotype A were represented the most across all sample locations. In fact, Pennsylvania was represented solely by clade A (**Fig. 2**). We also saw that most of the RT-qPCR low positive samples were also associated with genotype A, but that genotype A was represented evenly by low and high positive samples (**Fig. 2**). We were also interested in the association of chaq virus to galbut virus genotype A. In all but one case, the sequencing data revealed that chaq virus was associated with genotype A (**Fig. 2**).



**Figure 2: High infectivity and presence of Chaq virus are associated with galbut virus clade A infection.** Normalized RT-qPCR data is shown for samples sequenced via total RNA metagenomics. Color denotes presence of Chaq-virus and shape denotes whether the sample was a low positive sample.

We next wanted to look at the distribution of galbut virus and chaq virus segments identified in the sequencing data. We found that 17 samples had evidence of galbut virus or chaq virus coinfection. Fifteen of the samples with a coinfection only had evidence of multiples of one segment suggesting that coinfection can occur in a segment independent manner. However, the two samples that had complete coinfections by all three segments were the only two samples with multiple RNA1 segments. Multiple RNA3 segments was the most common coinfection, followed by RNA2 and chaq virus at the same frequency (**Fig 3**). It is important to note that samples missing a complete sequence for any of the three galbut virus segments does not necessarily mean that there was no evidence of that segment in the data, it just means that the sequence for that missing segment was too partial to submit to GenBank. For a heatmap of all identified sequences see **Supplemental Figure 1**. The most common missing complete sequence was RNA1 with six samples missing a complete sequence.

A screenshot of a graph

Description automatically generated

**Figure 3: Coinfection of galbut virus does not require all three segments.** The number of near complete or complete sequences submitted to GenBank identified for each segment is shown. Included samples had at least one galbut virus segment or Chaq virus sequence submitted to GenBank. Samples are grouped by location (Colorado (A), Maine (B), Ohio (C), Pennsylvania (D) and unknown (E)).

**Supplemental Figure 1: Coinfection of galbut virus does not require all three segments** **extended data.** Data as in Figure 3 but with data from all sequenced samples included. This figure accounts for partial galbut virus or Chaq virus sequences. Samples are grouped by location (Colorado (A), Maine (B), Ohio (C), Pennsylvania (D) and unknown (E)).

**Materials and Methods**

*Sample Collection:* Flies were collected over the summer of 2023 from Maine, Pennsylvania, Ohio, Colorado and an unknown location. Roughly 1000 flies were collected. Most of the flies were collected in one of eight locations in the area in and around Fort Collins, CO. Flies collected in Colorado were trapped in buckets baited with banana and yeast with organdy separating the flies from the decaying fruit. Flies captured in Maine, Pennsylvania and Ohio were trapped using a 3D printed fly trap [Food Paper]. The flies of unknown origin were isolated from peaches bought at a grocery store in Fort Collins, CO, although the peaches were from California, we were unable to definitively determine where they came from.

*RNA Extraction:* RNA was extracted from individual flies using a magnetic bead-based RNA/DNA capture method modified for use on the KingFisher (ThermoFisher Scientific). Briefly, flies were placed in an assay block [part] with a small BB [part] and 100µL lysis buffer (components). Flies were homogenized at 30 Hz for three minutes in a TissueLyzer [part]. The homogenate was moved to a new deep well plate and combined with a mastermix that included: 90µL homemade magnetic beads (components), 10µL lysis binding enhancer (components) and 60µL 100% isopropanol. In separate 96 well standard plates (part), 150µL of wash buffer one (components), wash buffer two (components) and 50µL of sterile water was added. The extraction was completed using a custom protocol for the KingFisher. Eluted nucleic acids were then stored at -80°C until needed.

*Galbut virus positivity:* To determine which samples were positive for galbut virus, cDNA was generated using 5.5µL RNA, 10µM random hexamer primers (IDT), 10µM dNTPs and water to 13µL and incubated at 65°C for 5 minutes. Next, 4µL 5X FS buffer, 1µL 0.1M DTT and 1µL reverse transcriptase was added and incubated at 50°C for 60 minutes followed by a heat inactivation step at 80°C for 10 minutes after which each sample was diluted in 90µL nuclease-free water. qPCR was conducted using NEB Luna Universal qPCR Mastermix following manufacturer recommendations and cycling conditions (NEB, M3003). Of the 1000 flies captured, 952 were positive for RpL32, suggesting a successful extraction and 641 were positive for galbut virus. Each sample was screened for galbut virus using two primer sets [Table X].

|  |  |  |  |
| --- | --- | --- | --- |
| Primer pair name (#s in lab collection) | Forward | Reverse | Reference |
| Galbut virus 1 (#1600/1601) | CCGTGAAGCAAGGAATCAAT | TGCCGATTTTCTGCTCTTTT | Cross 2020 |
| Galbut virus 2  (#2165/2170) | GTTCTTATCCGCATCGCCTA | AATTCACTACTTCCTTGACACCTC |  |
| RpL32- long (1012/1013) | TGCTAAGCTGTCGCACAAATGG | TGCGCTTGTTCGATCCGTAAC | Cao 2016 |

*NGS Library Preparation:* Of the 641 galbut virus positive samples, 155 were selected for NGS total RNA sequencing. Samples were selected based on their location (novel location), the degree of positivity (a range of high to low positive) and proximity to the positive control melting temperature (very similar to nearly a degree away). Sample concentration was obtained using the Qubit high-sensitivity RNA reagent (ThermoFisher, Q32852) and DNA was removed via a DNAse treatment [reagent]. Prior to library preparation, samples were normalized to 5 ng/µL except where sample concentration was too low. In these cases, either a different sample was selected or the maximum volume of input RNA was used. Due to limited availability of barcodes in our lab, the samples were split into two groups. Samples were prepared for sequencing using the Kapa Biosystems Kapa RNA Hyper Prep kit following manufacturer recommendations (Roche, 08098093702). Eight cycles were used for library amplification. HeLa cell RNA was used as a positive control and water was used as a negative control for both library preparations and sequencing. Sample pooling was determined using High Sensitivity DNA Qubit reagents and final library quality control was done using an Agilent D1000 HS Tapestation and KapaQuant reagents following manufacturers recommendations (Roche, 07960140001). Both pools were sent to Azenta/Genewiz where they were sequenced on a NovaSeq with 2X150 paired end sequencing.

**Data Analysis**

*Identification of virus sequences:* We used our lab’s previously described metagenomic classification pipeline (<https://github.com/stenglein-lab/taxonomy_pipeline>) to identify and validate virus sequences in the NGS datasets. In brief, adapters and low-quality reads were trimmed using cutadapt 3.5. Fastqc was used to assess post-collapsed read quality 0.11.9. Host reads were removed using bowtie2 2.4.5 and the remaining reads were assembled using spades 3.15.4. Virus sequences were identified using BLASTn 2.12.0 and a BLASTx search using diamond 2.0.14 was used to identify novel sequences based off of protein sequences. Draft sequences were validated by remapping of trimmed reads using bwa mem aligner version 0.7.17. Final sequences were submitted to the NCBI nucleotide database and raw NGS data to the NCBI sequence read archive repository.

*Molecular species identification:* To confirm the species identification of the samples, we used our labs species identification pipeline (<https://github.com/stenglein-lab/species_id>). In brief, this pipeline takes preprocessed reads and maps them to a given set of sequences. For our purposes we used a set of 529 cytochrome c oxidase sequences from Drosophilidae. This pipeline identified one *Drosophila affinis*, one *Drosophila simulans* and 147 *D. melanogaster*. Six samples remained undetermined.

**DICUSSION**

Which supports previous data that clade A is overall more abundant than clade B [OC paper].

Since RNA1 encodes the RNA-dependent RNA polymerase this could be a reflection of the decreased number of transcripts needed during viral replication.