The discovery, distribution and diversity of DNA viruses associated with *Drosophila melanogaster* in Europe

Authors:

Megan A. Wallace [1,2]	megan.wallace@ed.ac.uk	0000-0001-5367-420X
Kelsey A. Coffman [3]	kcoffman@uga.edu	0000-0002-7609-6286
Clément Gilbert [1,4]	clement.gilbert@egce.cnrs-gif.fr	0000-0002-2131-7467
Sanjana Ravindran [2]	sanjana.ravindran@ed.ac.uk	0000-0003-0996-0262
Gregory F. Albery [5]	gfalbery@gmail.com	0000-0001-6260-2662
Jessica Abbott [1,6]	jessica.abbott@biol.lu.se	0000-0002-8743-2089
Eliza Argyridou [1,7]	argyridou@bio.lmu.de	0000-0002-6890-4642
Paola Bellosta [1,8,9]	paola.bellosta@unitn.it	0000-0003-1913-5661
Andrea J. Betancourt [1,10]	A.Betancourt@liverpool.ac.uk	0000-0001-9351-1413
Hervé Colinet [1,11]	herve.colinet@univ-rennes1.fr	0000-0002-8806-3107
Katarina Eric [1,12]	katarina.eric@ibiss.bg.ac.rs	0000-0002-3456-2576
Amanda Glaser-Schmitt [1,7]	glaser@bio.lmu.de	0000-0002-1322-1000
Sonja Grath [1,7]	grath@bio.lmu.de	0000-0003-3621-736X
Mihailo Jelic [1,13]	mihailoj@bio.bg.ac.rs	0000-0002-1637-0933
Maaria Kankare [1,14]	maaria.kankare@jyu.fi	0000-0003-1541-9050
Iryna Kozeretska [1,15]	iryna.kozeretska@gmail.com	0000-0002-6485-1408
Volker Loeschcke [1,16]	volker@bios.au.dk	0000-0003-1450-0754
Catherine Montchamp-Moreau [1,4]	Catherine.Montchamp@egce.cnrs-gif.fr	0000-0002-5044-9709
Lino Ometto [1,17]	lino.ometto@unipv.it	0000-0002-2679-625X
Banu Sebnem Onder [1,18]	bdalgic@hacettepe.edu.tr	0000-0002-3003-248X
Dorcas J. Orengo [1,19,20]	dorcasorengo@ub.edu	0000-0001-7911-3224
John Parsch [1,7]	parsch@bio.lmu.de	0000-0001-9068-5549
Marta Pascual [1,19]	martapascual@ub.edu	0000-0002-6189-0612
Aleksandra Patenkovic [1,12]	aleksandra@ibiss.bg.ac.rs	0000-0001-5763-6294
Eva Puerma [1,19,20]	e.puerma@ub.edu	0000-0001-7261-187X
Michael G. Ritchie [1,21]	mgr@st-andrews.ac.uk	0000-0001-7913-8675
Omar Rota-Stabelli [1,22,23]	omar.rota@fmach.it	0000-0002-0030-7788
Mads Fristrup Schou [1,6,24]	mads.schou@biol.lu.se	0000-0001-5521-5269
Svitlana V. Serga [1,15,25]	svitlana.serga@gmail.com	0000-0003-1875-3185
Marina Stamenkovic-Radak [1,13]	marina@bio.bg.ac.rs	0000-0002-6937-7282
Marija Tanaskovic [1,12]	marija.tanaskovic@ibiss.bg.ac.rs	0000-0003-1440-2257
Marija Savic Veselinovic [1,13]	marijas@bio.bg.ac.rs	0000-0001-8461-4373
Jorge Vieira [1,26]	jbvieira@ibmc.up.pt	0000-0001-7032-5220
Cristina P. Vieira [1,26]	cgvieira@ibmc.up.pt	0000-0002-7139-2107
Martin Kapun [1,27]	martin.kapun@uzh.ch	0000-0002-3810-0504
Thomas Flatt [1,28]	thomas.flatt@unifr.ch	0000-0002-5990-1503
Josefa González [1,29]	josefa.gonzalez@csic.es	0000-0001-9824-027X
Fabian Staubach [1,30]	fabian.staubach@biologie.uni-freiburg.de	0000-0002-8097-2349
Darren J. Obbard [1,2,*]	darren.obbard@ed.ac.uk	0000-0001-5392-8142

*Author for correspondence

Key Words: DNA virus, Endogenous viral element, Drosophila, Nudivirus, Galbut virus, Filamentous virus, Adintovirus, Densovirus, Bidnavirus

Author affiliations:

[1] The European Drosophila Population Genomics Consortium (DrosEU)
[2] Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, Charlotte Auerbach Road, Edinburgh, EH9 3FL, UK
[3] Department of Entomology, University of Georgia, Athens, Georgia, USA
[4] Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, 91198 Gif-sur-Yvette, France.
[5] Department of Biology, Georgetown University, Washington DC, USA
[6] Department of Biology, Section for Evolutionary Ecology, Lund University, Sölvegatan 37, 223 62 Lund, Sweden
[7] Division of Evolutionary Biology, Faculty of Biology, Ludwig-Maximilians-Universität München, Planegg, Germany
[8] Dept of Cellular, Computational and Integrative Biology - CIBIO University of Trento, Via Sommarive 9, 38123 Trento, Italy
[9] Dept of Medicine&Endocrinology, NYU Langone Medical Center, 550 First Ave, 10016 NY USA
[10] Institute of Integrative Biology, University of Liverpool, Liverpool, L69 7ZB, UK
[11] UMR CNRS 6553 ECOBIO, Université de Rennes1, France
[12] University of Belgrade, Institute for Biological Research "Sinisa Stankovic", National Institute of Republic of Serbia, Bulevar despota Stefana 14211060 Belgrade, Serbia
[13] University of Belgrade – Faculty of Biology, Studentski trg 16, Belgrade, Serbia
[14] Department of Biological and Environmental Science, University of Jyväskylä, Finland
[15] State Institution National Antarctic Scientific Center, Ministry of Education and Science of Ukraine, 16 Shevchenko Ave., 01601, Kyiv, Ukraine
[16] Dept. of Biology, Genetics, Ecology and Evolution, Aarhus University, Ny Munkegade 116,
[17] Department of Biology and Biotechnology, University of Pavia, 27100 Pavia, Italy
[18] Department of Biology, Faculty of Science, Hacettepe University, Ankara, Turkey
[19] Departament de Genètica, Microbiologia i Estadística, Universitat de Barcelona, Barcelona, Spain.
[20] Institut de Recerca de la Biodiversitat (IRBio), Universitat de Barcelona, Barcelona, Spain.
[21] Centre for Biological Diversity, St Andrews University, St Andrews, Scotland HY15 4SS UK
[22] Research and Innovation Center, Fondazione E. Mach, 38010, San Michele all'Adige (TN), Italy
[23] Centre Agriculture Food Environment, University of Trento, 38010, San Michele all'Adige (TN), Italy
[24] Department of Bioscience, Aarhus University, Aarhus, Denmark
[25] Taras Shevchenko National University of Kyiv, 01601, 64 Volodymyrska str, Kyiv, Ukraine
[26] Instituto de Biologia Molecular e Celular (IBMC), University of Porto, Porto, Portugal
[27] Department of Evolutionary Biology and Environmental Studies, University of Zürich, Switzerland; Division of Cell & Developmental Biology, Medical University of Vienna, Austria
[28] Department of Biology, University of Fribourg, CH-1700 Fribourg, Switzerland
[29] Institute of Evolutionary Biology (CSIC-UPF), Barcelona, Spain.
[30] Department of Evolution and Ecology, University of Freiburg, 79104 Freiburg, Germany
DK-8000 Aarhus C, Denmark

Abstract

Drosophila melanogaster is an important model for antiviral immunity in arthropods, but very few DNA viruses have been described in association with the Drosophilidae. This has limited the opportunity to use natural host-pathogen combinations in experimental studies, and may have biased our understanding of the *Drosophila* virome. Here we describe fourteen DNA viruses detectable by metagenomic analysis of 6.5 thousand pool-sequenced *Drosophila*, sampled from 47 European locations between 2014 and 2016. These include three new Nudiviruses, a new and divergent Entomopox virus, a virus related to *Leptopilina boulardi* filamentous virus, and a virus related to *Musca domestica* salivary gland hypertrophy virus. We also find an endogenous genomic copy of Galbut virus, an RNA Partitivirus, segregating at very low frequency. Remarkably, we show that Vesanto virus, a small DNA virus previously described as a Bidnavirus, may be composed of up to 12 segments and represents a new lineage of segmented DNA viruses. Only two of the DNA viruses, Kallithea virus (Nudiviridae) and Vesanto virus (Bidna-virus like) are common, being found in 2% or more of wild flies. The other viruses are rare, with many likely to be represented by a single infected fly in the collection. We find that virus prevalence in Europe reflects that seen in publicly-available datasets, with Kallithea virus and Vesanto virus being commonly detectable in data from wild-caught flies and large population cages, and the others being rare or absent. These analyses suggest that DNA viruses are generally rarer than RNA viruses in *D. melanogaster*, and may be less likely to persist in laboratory cultures. Our findings go some way to redress the earlier bias toward RNA virus studies in *Drosophila*, and lay the foundation needed to harness the power of *Drosophila* as a model system for the study of DNA viruses.

Introduction

Drosophila melanogaster is one of our foremost models for antiviral immunity in arthropods (Huszart and Imler 2008, Mussabekova *et al.* 2017) and more than 100 *Drosophila*-associated viruses have been reported, including at least 30 confirmed to infect *D. melanogaster* (Brun and Plus 1980, Wu *et al.* 2010, Longdon *et al.* 2015, Webster *et al.* 2015, Webster *et al.* 2016, Medd *et al.* 2018). These include RNA viruses with positive sense single-stranded genomes (+ssRNA), such as *Drosophila C virus*, negative sense genomes (-ssRNA), such as *Drosophila melanogaster sigmavirus*, and double-stranded genomes (dsRNA), such as Galbut virus. Many of these RNA viruses are common in laboratory fly cultures and the wild (Webster *et al.* 2015). For example Galbut virus, a segmented and vertically transmitted *Partitivirus*, is carried by more than 50% of wild-caught adult *D. melanogaster* (Webster *et al.* 2015, Cross *et al.* 2020). Overall, more than 20% of wild-caught flies carry multiple RNA viruses, and about one third of laboratory fly lines and almost all *Drosophila* cell cultures are infected by at least one RNA virus (Plus 1978, Brun and Plus 1980, Webster *et al.* 2015, Shi, White, *et al.* 2018). However, in contrast to this wealth of RNA viruses, until relatively recently, DNA viruses of *Drosophila* were entirely unknown (Brun and Plus 1980, Huszart and Imler 2008).

The first described DNA virus of a drosophilid was published in 2011, after discovery by metagenomic sequencing of wild-caught *Drosophila innubila* (Unckless 2011). This virus is a member the Nudiviridae, a lineage of large (120-180Kbp) dsDNA viruses historically best known as pathogens of Lepidoptera and Coleoptera (Harrison *et al.* 2020), but with genomic 'fossil' evidence of a very broad host range (Cheng *et al.* 2020). *Drosophila innubila* Nudivirus infects several *Drosophila* species in North America, with a prevalence of up to 40% in *D. innubila*, where it substantially reduces fecundity (Unckless 2011). The first reported DNA virus of *D. melanogaster* was a closely-related Nudivirus published by Webster et al. (2015), and named 'Kallithea virus' for a collection location. This virus was also initially detected by metagenomic sequencing, but PCR

surveys indicate that it is common in wild *D. melanogaster* and *D. simulans* populations (globally 5% and 0.5% respectively; Webster *et al.* 2015). Kallithea virus has been isolated for experimental study, and reduces male longevity and female fecundity (Palmer *et al.* 2018). Consistent with its presumed niche as a natural pathogen of *Drosophila*, Kallithea virus encodes a suppressor of *D. melanogaster* NF-kappa B immune signalling (Palmer *et al.* 2019). Prior to the work described here, the only other reported natural DNA virus infection of a drosophilid was the discovery (again through metagenomic sequencing) of a small number of RNA reads from *Invertebrate iridescent virus 31* (IIV31; *Armadillidium vulgare* iridescent virus) in *D. immigrans* and *D. obscura* (Webster *et al.* 2016). This virus is known as a generalist pathogen of terrestrial isopods (Piegu *et al.* 2014), but its presence as RNA (indicative of expression) in these *Drosophila* species suggests that it may have a broader host range.

The apparent dearth of specialist DNA viruses infecting *Drosophilidae* is notable (Brun and Plus 1980, Huszart and Imler 2008), perhaps because DNA viruses have historically dominated studies of insects such as Lepidoptera (Cory and Myers 2003), and because DNA viruses are well known from other Diptera, including the Hytrosaviruses of *Musca* and *Glossina* (Kariithi *et al.* 2017), Densoviruses of mosquitoes (Carlson *et al.* 2006), and Entomopox viruses from midges and mosquitoes (Lawrence 2011). The lack of native DNA viruses for *D. melanogaster* has practical implications for research, as the majority of experiments have had to utilise non-native host-parasite combinations (Bronkhorst *et al.* 2014, West and Silverman 2018, but see Palmer *et al.* 2019). Nevertheless, it remains an open question as to whether the *D. melanogaster* virome is really depauperate in DNA viruses.

As part of a large population-genomics study using pool-sequencing of wild *D. melanogaster*, we recently reported the genomes of four new DNA viruses associated with European *Drosophila* samples collected in 2014 (the DrosEU consortium; Kapun *et al.* 2020). These comprised a second *melanogaster*-associated Nudivirus ('Esparto virus'), two Densoviruses ('Viltain virus' and 'Linvill road virus'), and two segments of a putative Bidnavirus ('Vesanto virus'). Here we expand our sampling to encompass 167 short-read pool-sequenced samples from a total of 6668 flies, collected seasonally over three years from 47 different locations across Europe. We combine these data with a small amount of long-read sequencing to complete the genome of a novel and highly divergent Entomopox virus. We also identify a further three *Drosophila*-associated Nudiviruses (two complete genomes, and fragments of a third), fragments of a novel Hytrosa virus most closely related to *Musca domestica* salivary gland hypertrophy virus, fragments of a Filamentous virus distantly related to *Leptopilina boulardi* filamentous virus, and three polinton-like sequences related to 'Adintoviruses'. Our improved assemblies and sampling show that Vesanto virus may be composed of up to 12 segments, and appears to be a representative of a new distinct lineage of multi-segmented ssDNA viruses related to the Bidnaviridae. We use our data to quantify the geographic and temporal distribution of these viruses, and to summarise patterns of genetic diversity for those with highest prevalence. We find that two viruses (Kallithea virus and Vesanto virus) are common in European *D. melanogaster*, but that the majority of DNA viruses appear very rare—most probably appearing once in our sampling.

Methods

Sample collection and sequencing

A total of 6668 adult male *Drosophila* were collected across Europe by members of the DrosEU consortium between 19th June 2014 and 22nd November 2016, using yeast-baited fruit. There were a total of 47 different collection sites spread from Recarei in Portugal (8.4° West) to Alexandrov in Russia (38.7° East), and from Nicosia in Cyprus (36.1° North) to Vesanto in Finland (62.6° North). The majority of sites were represented by more than one collection, with many sites appearing in all three years, and several being represented by two collections per year (early and late in the *Drosophila* breeding season for that location). After morphological examination to infer species identity, a minimum

of 33 and maximum of 40 male flies (mean 39.8) were combined from each site and preserved in ethanol at -20°C or -80°C for pooled DNA sequencing. Male flies were chosen because, within Europe, male *D. melanogaster* should be morphologically unambiguous. Nevertheless, subsequent analyses identified the occasional presence of the sibling species *D. simulans*, and two collections were contaminated with the distant relatives *D. phalerata* and *D. testacea* (below). Full collection details are provided in Supplementary File S1, and the detailed collection protocol is provided as supporting material in Kapun *et al* (2020).

To extract DNA, ethanol-stored flies were rehydrated in water and transferred to 1.5 ml well plates for homogenisation using a bead beater (Qiagen Tissue Lyzer II). Protein was digested using Proteinase K, and RNA depleted using RNAse A. The DNA was precipitated using phenol-chloroform-isoamyl alcohol and washed before being air dried and re-suspended in TE. For further details, see the supporting material in Kapun *et al* (2020). DNA was sequenced in three blocks (2014, most of 2015, remainder of 2015 and 2016) by commercial providers using 151nt paired end Illumina reads. Block 1 libraries were prepared using NEBNext Ultra DNA Lib Prep-24 and NEBNext Multiplex Oligos, and sequenced on the Illumina NextSeq 500 platform by the Genomics Core Facility of the University Pompeu Fabra (UPF; Barcelona, Spain). Block II and III libraries were prepared using the NEBNext Ultra II kit and sequenced on the HiSeq X platform by NGX bio (San Francisco, USA). All raw Illumina read data are publicly available under SRA project accession PRJNA388788.

To improve virus genomes, and following an initial exploration of the Illumina data, we pooled the remaining DNA from four of the collections (samples UA_Yal_14_16, ES_Gim_15_30, UA_Ode_16_47 and UA_Kan_16_57) for long-read sequencing using the Oxford Nanopore Technology 'Minion' platform. After concentrating the sample using a SpeedVac (ThermoFisher), we prepared a single library using the Rapid Sequencing Kit (SQK-RAD004) and sequenced it on an R9.4.1 flow cell, subsequently calling bases with Guppy version 3.1.5 (https://community.nanoporetech.com).

Read mapping and identification of contaminating taxa

We trimmed Illumina sequence reads using Trim Galore version 0.4.3 (Krueger 2015) and Cutadapt version 1.14 (Martin 2011), and mapped trimmed reads as read pairs to reference sequences using Bowtie 2 version 2.3.4 or version 2.4.1 (Langmead and Salzberg 2012), recording only the best mapping position. To remove *Drosophila* reads, and to quantify potentially contaminating taxa such as *Wolbachia* and other bacteria, fungi, and trypanosomatids, we mapped each dataset against a combined '*Drosophila* microbiome' reference. This reference comprised the genomes of *D. melanogaster* (Chang and Larracuente 2019), *D. simulans* (Nouhaud 2018), three *Drosophila*-associated *Wolbachia* genomes, 69 other bacteria commonly reported to associate with *Drosophila* (including multiple *Acetobacter*, *Gluconobacter*, *Lactobacillus*, *Pantoea*, *Providencia*, *Pseudomonas* and *Serratia* genomes), and 16 microbial eukaryotic genomes (including two *Drosophila*-associated trypanosomatids, a microsporidian, the entomopathogenic fungi *Metarhizium anisopliae*, *Beauveria bassiana* and *Entomophthora muscae*, and several yeasts associated with rotting fruit). A full list of the genomes included is provided in Supplementary File S2. To provide approximate quantification we used raw mapped read counts, normalised by target length and fly read counts where appropriate.

During manual examination of *de novo* assemblies (below) we identified a number of short contigs from other taxa, including additional species of *Drosophila*, *Drosophila* commensals such as mites and nematodes, and potential sequencing contaminants such as humans and model organisms. To quantify this potential contamination, we re-mapped all trimmed read pairs to a reference panel of short diagnostic sequences. This panel comprised a region of *Cytochrome Oxidase I* (COI) from 20 species of *Drosophila* (European *Drosophila* morphologically similar to *D. melanogaster*, and *Drosophila* species identified in *de novo* assemblies), 667 species of nematode (including lineages most likely to be associated with *Drosophila*, and a contig identified by *de novo*

assembly), 106 parasitic wasps (including many lineages commonly associated with *Drosophila*), two species of mite (identified in *de novo* assemblies), complete mitochondrial genomes from six model vertebrates, and complete plastid genomes from eight crop species. Because cross-mapping between *D. melanogaster* and *D. simulans* is possible at many loci, we also included a highly divergent but low-diversity 2.3 kbp region of the single-copy nuclear gene *Argonaute-2* to estimate levels of *D. simulans* contamination. Where reads indicated the presence of other *Drosophila* species, this was further confirmed by additional mapping to *Adh*, *Amyrel*, *Gpdh* and *6-PGD*. A full list of the reference sequences included is provided in Supplementary File S2.

Virus genome assembly and annotation

To identify samples containing potentially novel viruses, we retained read pairs that were not concordantly mapped to the combined '*Drosophila* microbiome' reference (above) and used these for *de novo* assembly using SPAdes version 3.14.0 (Nurk *et al.* 2013), after *in silico* normalisation of read depth to a target coverage of 200 and a minimum coverage of 3 using bbnorm (https://sourceforge.net/projects/bbmap/). We performed normalisation and assembly separately for each of the 167 samples. We then used the resulting scaffolds to search a database formed by combining the NCBI 'refseq protein' database with the viruses from NCBI 'nr' database. The search was performed using Diamond blastx (version 0.9.31; Buchfink *et al.* 2014) with an e-value threshold of $1x10^{-30}$, permitting frameshifts, and retaining hits within 5% of the top hit.

The resulting hits were examined to exclude all phage, retroelements, giant viruses (i.e., Mimiviruses and relatives), and likely contaminants such as perfect matches to well-characterised plant, human, pet, and vertebrate livestock viruses (e.g. Ebola virus, Hepatitis B virus, Bovine viral diarrhoea virus, Murine leukemia virus). We also excluded virus fragments that co-occurred across samples with species other than *Drosophila*, such as mites and fungi, as likely to be viruses of those taxa. Our remaining candidate virus list included known and potentially novel DNA viruses, and one previously reported *Drosophila* RNA virus. For each of these viruses we selected at least one representative population sample, based on high coverage, for targeted genome re-assembly.

For targeted re-assembly of each virus we re-mapped all non-normalised reads to the putative virus scaffolds from the first assembly and retained all read pairs for which at least one partner had mapped. Using these virus-enriched read sets we then performed a second *de novo* SPAdes assembly for each target sample, but to aid scaffolding and repeat resolution we additionally included the long reads (Antipov *et al.* 2015) that had been generated separately from UA_Yal_14_16, ES_Gim_15_30, UA_Ode_16_47 and UA_Kan_16_57. We examined the resulting assembly graphs using Bandage version 0.8.1 (Wick *et al.* 2015) and based on inspection of coverage and homology with related viruses we manually resolved short repeat regions, bubbles associated with polymorphism, and long terminal repeat regions. For viruses represented only by a few low-coverage fragments, we concentrated assembly and manual curation on genes and gene fragments that would be informative for phylogenetic analysis.

For Vesanto virus, a Bidna-like virus with two previously-reported segments (Kapun *et al.* 2020), our preliminary manual examination of the assembly graph identified a potential third segment. We therefore took two approaches to explore the possibility that this virus is composed of more than two segments. First, to identify completely new segments, we mapped reads from samples with or without segments S01 and S02 to all high-coverage scaffolds from one sample that contained those segments. This allowed us to identify possible further segments based on their pattern of co-occurrence across samples (e.g. Batson *et al.* 2020, Obbard *et al.* 2020). Second, to identify substantially divergent (but homologous) alternative segments we used a blastp similarity search using predicted Vesanto virus proteins and predicted proteins from *de novo* scaffolds. Again, we examined targeted assembly graphs using Bandage (Wick

et al. 2015), and resolved inverted terminal repeats and apparent mis-assemblies manually.

To annotate viral genomes with putative coding DNA sequences we identified all open reading frames of 150 codons or more that started with ATG, and translated these to provide putative protein sequences. We retained those with significant similarity to known proteins from other viruses, along with those that did not overlap longer open reading frames.

Presence of DNA viruses in publicly available Drosophila datasets

To detect DNA viruses present in publicly available *Drosophila* datasets, we chose 28 SRA 'projects' and mapped these to the virus genomes using Bowtie 2 (Langmead and Salzberg 2012). Among these were several projects associated with the *Drosophila melanogaster* Genome Nexus (Lack *et al.* 2015, Lange *et al.* 2016, Sprengelmeyer *et al.* 2019), the Drosophila Real-Time Evolution Consortium (Dros-RTEC; Machado *et al.* 2019), pooled GWAS studies (e.g. Endler *et al.* 2018), evolve-and-resequence studies (Jalvingh *et al.* 2014, Schou *et al.* 2017, Kelly and Hughes 2019), studies of local adaptation (e.g. Campo *et al.* 2013, Kang *et al.* 2019), and introgression (Kao *et al.* 2015). In total this represented 3003 sequencing 'run' datasets. For each run, we mapped up to 10 million reads to the *Drosophila* DNA viruses identified above (forward reads only for paired-end datasets), and recorded the best-mapping location for each read. Short reads and low complexity regions allow some cross-mapping among the larger viruses, and between viruses and the fly genome. We therefore chose an arbitrary detection threshold of 250 mapped reads to define the presence of each of the larger viruses (expected genome size >100 kbp) and a threshold of 25 reads for the smaller viruses (genome size <100 kbp). Consequently, our estimates may be conservative tests of virus presence, and the true prevalence may be slightly higher. For three of the viruses we additionally selected a subset of the public datasets for *de novo* assembly, using the same assembly approach as outlined for DrosEU data above.

Phylogenetic inference

To infer the phylogenetic relationships among DNA viruses of *Drosophila* and representative viruses of other species, we selected a small number of highly conserved virus protein-coding loci that have previously been used for phylogenetic inference. For Densoviruses we used the viral replication initiator protein, NS1 (Pénzes *et al.* 2020), for Adintoviruses and Bidna-like viruses we used DNA Polymerase B (Krupovic and Koonin 2014, Starrett *et al.* 2020), for Poxviruses we used rap-94, and the large subunits of Poly-A polymerase and the mRNA capping enzyme (Thézé *et al.* 2013), and for Nudiviruses, Filamentous viruses and Hytrosa viruses we used P74, Pif-1, Pif-2, Pif-3, Pif-5 (ODV-e56) and the DNA polymerase B (e.g., Kawato *et al.* 2019). In each case we used a blastp search to identify a representative set of similar proteins in the NCBI 'nr' database, and among proteins translated from publically available transcriptome shotgun assemblies deposited inGenBank. For the Nudiviruses, Filamentous viruses and Hytrosa viruses we combined these with proteins collated by Kawato *et al* (2019). We aligned protein sequences for each locus using t-coffee mode 'accurate', which combines structural and profile information from related sequences (Notredame *et al.* 2000), and manually 'trimmed' poorly aligned regions from each end of each alignment. We did not filter the remaining alignment positions for coverage or alignment 'quality', as this tends to bias toward the guide tree and to give false confidence (Tan *et al.* 2015). We then inferred trees from concatenated loci (where multiple loci were available) using IQtree2 with default parameters (Minh *et al.* 2020), including automatic model selection and 1000 ultrafast bootstraps.

Age of an endogenous viral element

To infer the age of an endogenous copy (EVE) of Galbut virus, we used a strict-clock Bayesian phylogenetic analysis of virus sequences, as implemented in BEAST 1.10.2 (Suchard *et al.* 2018). To make this inference our assumption is that any evolution of the EVE after insertion is negligible relative to RNA virus evolutionary rates. We assembled complete 1.6 kb segment sequences from publicly-available RNA

sequencing datasets (Lin *et al.* 2016, Garlapow *et al.* 2017, Yablonovitch *et al.* 2017, Bost *et al.* 2018, Shi, White*, et al.* 2018, Everett *et al.* 2020), and filtered these to retain unique sequences and exclude possible recombinants identified with GARD (Kosakovsky Pond *et al.* 2006) and RDP5 (Martin *et al.* 2015). The few recombinants were all found in multiply-infected pools, suggesting they may have been chimeric assemblies. For sequences from Shi et al. (2018) we constrained tip dates according to the extraction date, and for other studies we constrained tip dates to the three-year interval prior to project registration. We aligned these sequences with the EVE sequence, and during phylogenetic analysis we constrained most recent date for the EVE to be its extraction date, but left the earliest date effectively unconstrained. Because the range of virus tip dates covered less than 10 years we imposed time information through a strongly informative log-normal prior on the strict clock rate, chosen to reflect the spread of credible evolutionary rates for RNA viruses (e.g., Peck and Lauring 2018). Specifically, we applied a data-scale mean evolutionary rate of $4x10^{-4}$ events/site/year with standard deviation $2.5x10^{-4}$, placing 95% of the prior density between $1x10^{-3}$ and $1.3x10^{-4}$. As our sampling strategy was incompatible with either a coalescent or birth-death tree process, we used a Bayesian Skyline coalescent model to allow flexibility in the coalescence rate, and thereby minimise the impact of the tree prior on the date (although alternative models gave qualitatively similar outcomes). We used the SDR06 substitution model (Shapiro *et al.* 2006) and otherwise default priors, running the MCMC for 100 million steps and retaining every 10 thousandth state. The effective sample size was greater than 1400 for every parameter. BEAST input xml will be provided via Figshare.

Virus quantification, and the geographic and temporal distribution of viruses

To quantify the (relative) amount of each virus in each pooled sample, we mapped read pairs that had not been mapped concordantly to the *Drosophila* microbiome reference (above) to the virus genomes. This approach means that low complexity reads map initially to the fly and microbiota, and are thus less likely to be counted or mismapped among viruses. This slightly reduces the detection sensitivity (and counts) but also increases the specificity. We mapped using Bowtie 2 (Langmead and Salzberg 2012), recording the best mapping location, and using either read count (per million reads) divided by target length (per kilobase) to quantify the viruses, or this value normalised by the equivalent number for *Drosophila* (combined *D. melanogaster* and *D. simulans* reads) to provide an estimate of virus genomes per fly genome in each pool. To quantify Vesanto virus genomes we excluded terminal inverted repeats from the reference, as these may be prone to cross-mapping among segments.

To provide a simple estimate of prevalence we assumed that pools represented independent samples from a uniform global population, and assumed that a pool of *n* flies constituted *n* Bernoulli trials in which the presence of virus reads indicated at least one infected fly. Based on this model, we inferred a maximum-likelihood estimate of global prevalence for each virus, with 2 log-likelihood intervals (e.g., Speybroeck *et al.* 2012). Because some cross-mapping between viruses is possible, and because barcode switching can cause reads to be mis-assigned among pools, we chose to use a virus detection threshold of 1% of the fly genome copy number to define ‘presence’. This threshold was chosen on the basis that male flies artificially infected with Kallithea virus have a relative viral genome copy number of >120 three days post infection (Palmer *et al.* 2018), or around 3% of genome copy number for a single infected fly in a pool of 40. Thus, although our approach may underestimate virus prevalence if titre is low, it provides some robustness to barcode switching while also giving reasonable power to detect a single infected fly.

In reality, pools are not independent of each other in time or space or other potential predictors of viral infection. Therefore, for the three most prevalent viruses (Kallithea virus, Linvill Road virus, and Viltain virus) we analysed predictors of the presence and absence of each viruse within population pools using a binomial generalised linear mixed model approach. We fitted linear mixed models in a spatial framework using R-INLA (Blangiardo *et al.* 2013), taking a Deviance

Information Criterion (DIC) of 2 or larger as support for a spatial or spatiotemporal component in the model. In addition to any spatial random effects, we included one other random-effect and four fixed-effect predictors. The fixed effects were: the level of *D. simulans* contamination (measured as the percentage *D. simulans* Ago2 reads); the amount of *Wolbachia* (measured as reads mapping to *Wolbachia* as relative to the number mapped to fly genomes); the sampling season (early or late); and the year (categorical 2014, 2015, 2016). We included sampling location as a random effect, to account for any additional non-independence between collections made at the same sites or by the same collector. The inclusion of a spatially distributed random effect was supported for Kallithea and Linvill Road viruses, but this did not vary significantly with year. Map figures were plotted and model outputs summarised with the R package ggregplot (https://github.com/gfalbery/ggregplot), and all code to perform these analyses will be provided via Figshare.

Virus genetic diversity

Reads that had initially been mapped to Kallithea virus, Linvill Road virus and Vesanto virus (above) were remapped to reference virus genomes using BWA MEM with local alignment (Li 2013). For the segmented Vesanto virus, we included multiple divergent haplotypes in the reference but excluded terminal inverted repeats, as reads derived from these regions will not map uniquely. After identifying the most common haplotype for each Vesanto virus segment in each of the samples, we remapped reads to a single reference haplotype per sample. For all viruses, we then excluded secondary alignments, alignments with a Phred-scaled mapping quality (MAPQ) <30, and optical and PCR duplicates using picard v.2.22.8 'MarkDuplicates' (http://broadinstitute.github.io/picard/). Finally, we excluded samples that had a read-depth of less than 25 across 95% of the mapped genome.

In addition to calculating per-sample diversity, to calculate total population genetic diversity we created single global pool representative of diversity across the whole population by merging sample bam files for each virus or segment haplotype. To reduce computational demands, each was down-sampled to an even coverage across the genome (no greater read depth at a site than the original median) and no sample contributed more than 500-fold coverage. To produce the final bam files for analyses, bam files for the global pool and each of the population pools were re-aligned around indels using GATK v3.8 (Van der Auwera *et al.* 2013). We created mPileup files using SAMtools (Li *et al.* 2009) to summarise each of these datasets using (minimum base quality = 40 and minimum MAPQ = 30), down-sampling population samples to a maximum read depth of 500. We masked regions surrounding indels using 'popoolation' (Kofler, Orozco-terWengel*, et al.* 2011), and generated allelic counts for variant positions in each using 'popoolation2' (Kofler, Pandey*, et al.* 2011), limiting our search to single nucleotide polymorphisms (SNPs) with a minor allele frequency of at least 1%.

To calculate average pairwise nucleotide diversity at synonymous (π_S) and non-synonymous (π_A) sites we identified synonymous and non-synonymous SNPs using popoolation (Kofler, Orozco-terWengel*, et al.* 2011), excluding SNPs with a minor allele frequency of less than 1%. In general, estimates of genetic diversity from pooled samples, such as those made by popoolation and population2 attempt to account for variation caused by finite sample sizes of individuals each contributing to the pool of nucleic acid. However, such approaches cannot be applied to viruses from pooled samples, as it is not possible to infer the number of infected flies in the pool or even to equate an infected fly with an individual (flies may be multiply infected). For this reason, we calculated π_A and π_S based on raw allele counts derived from read frequencies (code will be made available via Figshare). We did this separately for each gene in the merged global pool, and also for the whole genome in each infected population pool.

Structural variation and indels in Kallithea virus

Large DNA viruses such as Kallithea virus can harbour transposable element (TE) insertions and structural rearrangements (Loiseau *et al.* 2020), and often contain abundant short repeat-

length variation (Zhao *et al.* 2012). To identify large-scale rearrangements, we identified all read pairs for which at least one read mapped to Kallithea virus, and used SPAdes (Bankevich *et al.* 2012) to perform *de novo* assemblies separately for each dataset using both *in silico* normalised and un-normalised reads. We then selected those scaffolds approaching the expected length of the genome (>151 Kbp), and examined the assembly graphs manually using bandage (Wick *et al.* 2015), retaining those in which a single circular scaffold could be seen, with a preference for un-normalised datasets. These were then linearised starting at the DNA Polymerase B coding sequence, and aligned using muscle (Edgar 2004). This approach will miss structural variants at low frequency within each population, but could identify any major rearrangements that are fixed differently across populations.

To detect polymorphic transposable element insertions that were absent from the reference genome, we identified 16 population samples that had more than 300-fold read coverage of Kallithea virus and extracted all reads that mapped to the virus. We aligned these to 135 *D. melanogaster* TEs curated in the November 2016 version of Repbase (Bao *et al.* 2015) using blastn (-task megablast). All reads for which one portion aligned to the virus (Genome reference KX130344.1) and another portion aligned to a *D. melanogaster* TE were identified as chimeric using the R script provided by (Peccoud *et al.* 2018), and those for which the read-pair spanned TE ends were considered evidence of a TE insertion.

Finally, to catalogue short indel polymorphisms in coding and intergenic regions, we used popoolation2 (Kofler, Pandey*, et al.* 2011) to identify the genomic positions (relative to the reference genome) in each of the infected samples for which a gap is supported by at least 5 reads. We used a chi-square test for independence to test if there was an association between the coding status of a position and the probability that an indel was supported at that position in at least one population sample.

Results and Discussion

Host species composition

A total of 8.4 billion read pairs remained after trimming, with between 27.3 and 78 million pairs per population sample. On average, 93% of reads (range 70 - 98%) could be mapped to *Drosophila* or likely components of the *Drosophila* microbial community: *Wolbachia* made up an average of 0.5% of mapped non-fly reads (range 0.0 - 2.9%), other mapped bacterial reads combined were 0.6% (0.0 - 3.2%), and microbial eukaryotes were 0.3% (0.0 - 3.7%). The eukaryotic microbiota included the fungal pathogen *Entomophthora muscae* (Elya *et al.* 2018), with reads present in 42 of 167 samples (up to 1.38 reads per kilobase per million reads, RPKM), a novel trypanosomatid distantly related to *Herpetomonas muscarum* (Sloan *et al.* 2019) with reads present in 80 samples (up to 0.87 RPKM), and the microsporidian *Tubulinosema ratisbonensis* (Niehus *et al.* 2012), which we detected in one sample (0.54 RPKM). We excluded two virus-like DNA Polymerase B fragments from the analyses below because they consistently co-occurred with a fungus very closely related to *Candida* (*Clavispora*) *lusitaniae* (correlation coefficient on >0.94, $p<10^{-10}$; Supplementary File S4). For a detailed assessment of the microbial community in the 2014 collections, see Kapun et al (2020) and Wang et al (2020). Raw and normalised read counts are presented in Supplementary File S3, and raw data are available under project accession PRJNA388788.

The remaining 2% to 30% of reads could include metazoan species associated with *Drosophila*, such as nematodes, mites, or parasitoid wasps. By mapping all reads to small reference panel of *Cytochrome Oxidase I* (COI) sequences (Supplementary File S2), we identified 13 samples with small read numbers mapping to potentially parasitic nematodes, including an unidentified species of *Steinernema*, two samples with reads mapping to *Heterorhabditis bacteriophora* and three with reads mapping to *Heterorhabditis marelatus*. *De novo* assembly also identified an 8.4 kbp nematode scaffold with 85% nucleotide identity to the mitochondrion of *Panagrellus redivivus,* a free-living rhabditid

associated with decomposing plant material. Reads from this nematode were detectable in 73 of the 167 samples, rarely at a high level (up to 0.8 RPKM). Only one sample contained reads that mapped to mite COI, sample UK_Dai_16_23, which mapped at high levels (5.8 and 2.2 RPKM) to two unidentified species of Parasitidae (Mesostigmata, Acari). We excluded two Cyclovirus-like fragments from the analyses below because they occurred only in the sampled contaminated with the two mites, suggesting that they may be associated with the mites or integrated into their genomes (Supplementary File S3, Supplementary File S4).

To detect the presence of drosophilid hosts other than *D. melanogaster*, we mapped all reads to a curated panel of short diagnostic sequences from COI and Argonaute-2, the latter chosen for its ability to reliably distinguish between the close relatives *D. melanogaster* and *D. simulans*. As expected from the samples collected in 2014 (Kapun *et al.* 2020), 30 of the 167 samples contained *D. simulans* at a threshold of >1% of Ago2 reads. Mapping to COI sequences from different species, we identified only three further *Drosophila* species present in any sample at a high level. These included two small yellowish European species; *D. testacea,* which accounted for 2.4% of COI in UA_Cho_15_26 (263 reads), and *D. phalerata*, which accounted for 12.2% of COI in AT_Mau_15_50 (566 reads). Both were confirmed by additional mapping to *Adh*, *Amyrel*, *Gpdh* and *6-PGD* (Supplementary File S3), and their mitochondrial genomes were recovered as 9 kbp and 16 kbp *de novo* scaffolds, respectively. More surprisingly, some of the collections made in 2015 contained reads derived from *D. serrata*, a well-studied model related to *D. melanogaster* and endemic to tropical Australia (Reddiex *et al.* 2018). Samples TR_Yes_15_7 and FR_Got_15_48 had particularly high levels of *D. serrata* COI, with 94% (23,911 reads) and 7% (839 reads) of COI respectively, but reads were also detectable in another 6 pools. The presence of *D. serrata* sequences was confirmed by mapping to *Adh*, *Amyrel*, *Gpdh* and *6-PGD* (Supplementary File S3). However, examination of splice junctions showed that *D. serrata* reads derived from cDNA rather than genomic DNA, and must therefore result from cross-contamination during sequencing, or from barcode switching. Below, we note where conclusions may be affected by the presence of species of other than *D. melanogaster*.

Finally, among *de novo* assembled contigs, we also found evidence for several crop-plant chloroplasts and vertebrate mitochondria that are likely to represent sequencing or barcode-switching contaminants. The amounts were generally very low (median 0.01 RPKM), but a few samples stood out as containing potentially high levels of these contaminants. Most notably sample TR_Yes_15_7, in which only 76% of reads mapped to fly or expected microbiota, had 8.1 RPKM of human mtDNA, 5.1 RPKM of *Cucumis melo* cpDNA, and 3.5 RPKM of *Oryza sativa* cpDNA. We do not believe this contamination has any impact on our findings.

Previously-reported DNA virus genomes

Six different DNA viruses were previously detected among DrosEU samples from 2014 and reported by Kapun *et al*. (2020). These included one known virus (Kallithea virus; Webster *et al.* 2015) and five new viruses, of which four were assembled by Kapun *et al.* (2020). Kallithea virus (Nudiviridae) is a relatively common virus of *D. melanogaster* (Webster *et al.* 2015) that has a circular dsDNA genome of *ca.* 153 kbp encoding approximately 95 proteins (Figure 1), and is closely related to *Drosophila innubila* Nudivirus (Figure 2A). Esparto virus is a second *D. melanogaster* Nudivirus that was present at levels too low to permit assembly by Kapun *et al.* (2020) from the 2014 data, but was instead assembled in that paper from a *D. melanogaster* sample collected in Esparto, California USA (SRA dataset SRR3939042; Machado *et al.* 2019). It has a circular dsDNA genome of *ca.* 183 kbp that encodes approximately 90 proteins, and it is closely related to *Drosophila innubila* Nudivirus and Kallithea virus (Figure 1; Figure 2A). Viltain virus and Linvill Road virus are both small denso-like viruses (Parvoviridae), with ssDNA genomes of approximately 5 kb. Viltain virus is most closely related to *Culex pipiens ambidensovirus* (Jousset *et al.* 2000), and the genome appears to encode at least four proteins—two in each orientation (Figure 1; Figure 2B). As expected, the ends of the genome

are formed of short inverted terminal repeats (Figure 1). Linvill Road virus is most closely related to the unclassified *Haemotobia irritans* densovirus (Ribeiro *et al.* 2019) and appears to encode at least three proteins, all in the same orientation (Figure 1; Figure 2B). As with Esparto virus, Kapun *et al*. (2020) were unable to assemble the Linvill Road virus genome from the DrosEU 2014 data and instead based their assembly on a collection of *D. simulans* from Linvilla, Pennsylvania USA (SRR2396966; Machado *et al.* 2019). Here we identified a DrosEU 2016 collection (ES_Ben_16_32; Benalua, Spain) with sufficiently high titre to permit an improved genome assembly (submitted to Genbank under accession MT490308). This is 99% identical to the previous Linvill Road virus assembly, but by examination of the assembly graph we were able to complete more of the inverted terminal repeats and extend the genome length to 5.4 kb (Figure 1). Table 1 provides a summary of all DNA viruses detectable in DrosEU data.

Vesanto virus is a multi-segmented Bidna-like virus

Kapun *et al*. (2020) also reported two segments of a putative ssDNA Bidnavirus, named Vesanto virus for its collection site in 2014 (submitted to Genbank in 2016 as KX648533 and KX648534). This was presumed to be a complete genome based on homology with *Bombyx mori* bidensovirus (Li *et al.* 2019). Here we have been able to utilise expanded sampling and a small number of long-read sequences to extend these segments and to identify multiple co-occurring segments.

While examining an assembly graph of sample UA_Kan_16_57, we noted a third scaffold with a similarly high coverage (>300-fold) and structure (4.8 kb in length with inverted terminal repeats). This sequence also appeared to encode a protein with distant homology to *Bidnavirus* DNA polymerase B, and we reasoned that it might represent an additional virus. We therefore mapped reads from datasets that had high coverage of segments S01 and S02 to all scaffolds from the *de novo* build of UA_Kan_16_57, with the objective of finding any additional segments based on their co-occurrence across datasets (e.g. as done by Batson *et al.* 2020, Obbard *et al.* 2020). This identified several possible segments, all between 3.3 and 5.8 kbp in length and possessing inverted terminal repeats. We then used their translated open reading frames to search all of our *de novo* builds, and in this way identified a total of 12 distinct segments that show structural similarity and a strong pattern of co-occurrence (Figure 1 and Figure 3; Supplementary File S5). To capture the diversity present among these putative viruses, we made targeted *de novo* builds of three datasets, incorporating both Illumina reads and Oxford nanopore reads (Table 1). We have submitted these sequences to Genbank as MT496850-MT496878, and additional sequences are provided in Supplementary File S6. Because such an assembly is potentially problematic due to the inverted terminal repeats and pools infected with multiple viruses, we also sought to support these structures by identifying individual corroborating Nanopore reads of 2 kbp or more. The challenge of assembling such data means that the inverted terminal repeats should be treated with caution, but it is nevertheless striking that many of these putative segments show sequence similarity in their terminal inverted repeats, as commonly seen for segmented viruses.

Although we identified 12 distinct segments with strongly correlated presence/absence, not all segments were detectable in all affected samples (Figure 3A). Only segment S05, which encodes a putative glycoprotein and a putative nuclease domain protein, was always detectable (in 91 of the 167 samples; Supplementary File S5). Several segments were very commonly detectable, such as S03 (protein with homology to DNA PolB) and S10 (protein with domain of unknown function DUF3472 and a putative glycoprotein) in around 70 samples, and segments S01, S02, S04, S06 and S08 in around 55 samples. Others were extremely rare, such as S12 (a putative NACHT domain protein with homology to S09), which was only seen in five samples. We considered three possible explanations for this pattern.

Our first hypothesis was that Vesanto virus has 12 segments, but that variable copy number among the segments causes some to

occasionally drop below the detection threshold. In support of this, all segments are indeed detectable in the sample with the highest Vesanto virus read numbers (FR_Got_15_49), ranging from 7-fold higher than the fly genome for S07 to 137-fold higher for S05. In addition, 'universal' segment S05 is not only the most widely-detected segment across samples, but also has the highest average read depth within samples. However, despite 1.6 million Vesanto virus reads in the second highest copy number sample (RU_Val_16_20; 125-fold more copies of S6 than of *Drosophila*), no reads at all mapped to S12, which is strongly consistent with the absence of S12 from this sample. Our second hypothesis was that some segments may be 'optional' or satellite segments, or may represent homologous segments that comprise a re-assorting community (as in influenza). This is consistent with the apparent homology between some segments. For example, S01, S03, and S11 all encode DNA Polymerase B-homologs, and S06, S07 and S10 all encode DUF3472 proteins. It is also consistent with the universal presence of S6, which appears to lack homologs. However, two of the DNA PolB homologs are highly divergent (Figure 2C) to the extent it is hard to be confident of polymerase function, and we could not detect compelling negative correlations between homologous segments that might have suggested that they substitute for each other in different populations (Figure 3B). Our third hypothesis was that 'Vesanto virus' in fact represents multiple independent viruses (or phage), and that the superficially clear pattern of co-occurrence is driven by high (hypothetical) prevalence of this virus community in an occasional member of the *Drosophila* microbiota, such as a fungus or trypanosomatid. However, we were unable to detect any correlation with the mapped microbiota reads, and high levels of Vesanto virus are seen in samples with few un-attributable reads. For example, PO_Brz_15_12 has 11-fold more copies of S6 than of the fly, but less than 2% of reads derive from an unknown source (Supplementary File S3).

The complete genome of a new divergent Entomopox virus

Kapun *et al*. (2020) also reported the presence of a pox-like virus in DrosEU data from 2014, but were unable to assemble the genome. By incorporating a small number of long sequencing reads, and using targeted reassembly combined with manual examination of the assembly graph, we were able to assemble this genome from dataset UA_Yal_14_16 (SRR5647764) into a single contig of 219.9 kb. As expected for pox-like viruses, the genome appears to be linear with long inverted terminal repeats of 8.4 kb, and outside of the inverted terminal repeats sequencing coverage was 15.7-fold (Figure 1). We suggest the provisional name 'Yalta virus', reflecting the collection location (Yalta, Ukraine), and we have submitted the sequence to Genbank under accession number MT364305.

Within the Yalta virus genome we identified a total of 177 predicted proteins, including 46 of the 49 core poxvirus genes, and missing only the E6R virion protein, the D4R uracil-DNA glycosylase, and the 35 kDa RNA polymerase subunit A29L (Upton *et al.* 2003). Interestingly, the genome has a higher GC content than the previously published Entomopox viruses, which as a group consistently display the lowest GC content (< 21%) of the Poxvirus family (Perera *et al.* 2010, Thézé *et al.* 2013). Consistent with this, our phylogenetic analysis of three concatenated protein sequences suggests that the virus is distantly related, falling only slightly closer to Entomopox viruses than other pox viruses (Figure 2D). Given that all pox-like viruses infect metazoa, and that no animal species other than *D. melanogaster* appeared to be present in the sample, we believe *D. melanogaster* is likely to be the host.

Two new complete Nudivirus genomes, and evidence for a third

In addition to Kallithea virus and Esparto virus, our expanded analysis identified three novel Nudiviruses that were absent from data collected in 2014. We were able to assemble two of these into complete circular genomes of 112.3 kb (27-fold coverage) and 154.5 kb (41-fold coverage), respectively, based on datasets from Tomelloso, Spain (ES_Tom_15_28; SRR8439136) and Mauternbach, Austria (AT_Mau_15_50; SRR8439127). We suggest the provisional names 'Tomelloso virus' and 'Mauternbach virus', reflecting the collection locations, and we

have submitted the sequences to Genbank under accession numbers KY457233 and MG969167. We predict Tomelloso virus to encode 133 proteins (Figure 1), and phylogenetic analysis suggests that it is more closely related to a beetle virus (Oryctes rhinocerous Nudivirus, Figure 2A; Etebari *et al.* 2020) than to the other Nudiviruses described from *Drosophila*. Mauternbach virus is predicted to encode 95 proteins (Figure 1), and is very closely related to *Drosophila innubila* Nudivirus (Figure 2A; Unckless 2011, Hill and Unckless 2018). However, synonymous divergence (K_S) between these two viruses is approximately 0.7, i.e. nearly six-fold more than that between *D. melanogaster* and *D. simulans*, supporting their consideration as distinct 'species'. The third novel Nudivirus was present at a very low level in a sample from Kaniv, Ukraine (UA_Kan_16_57, SRR8494448), and only small fragments of the virus could be assembled for phylogenetic analysis (Genbank accession MT496841-MT496846). This showed that the fragmentary nudivirus from Kaniv is approximately equally divergent from *D. innubila* Nudivirus and Mauternbach virus (Figure 2A).

The collections from Tomelloso and Kaniv did not contain reads mapping to *Drosophila* species other than *D. melanogaster*, or to nematode worms or mites. Moreover, we identified Tomelloso virus in a number of experimental laboratory datasets from *D. melanogaster* (see below; Riddiford *et al.* 2020), and these lacked a substantial microbiome. Together these observations strongly support *D. melanogaster* as a host for these viruses. In contrast, COI reads suggest that the sample from Mauternbach may have contained with one *Drosophila phalerata* individual (2.4% of diagnostic nuclear reads; Supplementary file S3), and we could not detect Mauternbach virus in any of the public datasets we examined (below), making it uncertain whether *D. melanogaster* or *D. phalerata* was the host.

Evidence for a new filamentous virus and a new Hytrosa virus

Our search also identified fragments of two further large dsDNA viruses from lineages that have not previously been reported to naturally infect Drosophilidae. First, in sample UA_Ode_16_47 (SRR8494427) from Odesa, Ukraine, we identified around 16.6 kb of a novel virus related to the salivary gland hypertrophy viruses of *Musca domestica* and *Glossina palpides* (Figure 2A; Prompiboon *et al.* 2010, Kariithi *et al.* 2013). Our assembled fragments comprised 18 short contigs of only 1 to 3-fold coverage (submitted under accessions MT469997-MT470014). As the *Glossina* and *Musca* viruses have circular dsDNA genomes of 124.3 kbp and 190.2 kbp respectively, we believe that we have likely sequenced 5-15% of the genome. Because this population sample contains a small number of reads from *D. simulans* and an unknown nematode worm related to *Panagrellus redivivus*, and because we were unable to detect this virus in public datasets from *D. melanogaster* (below), the true host remains uncertain. However, given that the closest relatives all infect Diptera, it seems likely that either *D. melanogaster* or *D. simulans* is the host.

Second, in sample ES_Gim_15_30 (SRR8439138) from Gimenells, Spain, we identified around 86.5 kb of a novel virus distantly related to the filamentous virus of *Leptopilina boulardi*, a parasitoid wasp that commonly attacks *Drosophila* (Figure 2; Lepetit *et al.* 2016). The assembled fragments comprised 9 scaffolds of 5.9-16.9 kbp in length and 3 to 10-fold coverage, and are predicted to encode 69 proteins (scaffolds submitted to Genbank under accessions MT496832-MT496840). *Leptopilina boulardi* filamentous virus has a circular genome of 111.5 kbp predicted to encode 108 proteins. This suggests that, although fragmentary, our assembly may represent much of the virus. A small number of reads from ES_Gim_15_30 mapped to a relative of nematode *Panagrellus redivivus* and, surprisingly, to the Atlantic salmon (*Salmo salar*), but we consider these unlikely hosts as the level of contamination was very low and other filamentous viruses are known to infect insects. We were unable to detect the novel filamentous virus in any public datasets from *D. melanogaster* (below), and given that *Leptopilina boulardi* filamentous virus infects a parasitoid of *Drosophila*, it is possible that this virus may similarly infect a parasitoid wasp rather than the fly. However, we were unable to detect any reads

mapping to *Leptopilina* or other parasitoids of *Drosophila* in any of our samples, and we therefore think *D. melanogaster* is a good candidate to be a true host.

Near-complete genomes of three Adinto-like viruses

Based on the presence of a capsid protein, it is thought that some polinton-like transposable elements (also known as mavericks) are actually horizontally-transmitted viruses (Yutin *et al.* 2015). Some of these have recently been proposed as the Adintoviridae, a family of dsDNA viruses related to Bidnaviridae and other PolB-encoding DNA viruses (Starrett *et al.* 2020). We identified three putative Adintoviruses in DrosEU data. The first, provisionally named *Drosophila*-associated Adintovirus 1, occurred in sample UA_Cho_15_26 from Kopachi (Chornobyl Exclusion Zone), Ukraine (SRR8439134) and comprised a single contig of 14.5 kb predicted to encode 12 proteins. Among these proteins are not only a DNA Polymerase B and an integrase, but also homologs of the putative capsid, virion-maturation protease, and FtsK proteins of Adintoviruses (Starrett *et al.* 2020), and possibly very distant homologs of Hytrosavirus gene MdSGHV056 and Ichnovirus gene AsIV-cont00038 (Figure 1). The second, provisionally named *Drosophila*-associated Adintovirus 2, is represented by a 13.3 kb contig assembled using AT_Mau_15_50 from Mauternbach, Austria (SRR8439127). It is very closely related to the first Adintovirus, and encodes an almost-identical complement of proteins (Figure 1). In a phylogenetic analysis of DNA PolB sequences, both fall close to sequences annotated as Polintons in other species of *Drosophila* (Figure 2C). However, it is striking that these two datasets are those that are contaminated by *D. testacea* (1.3%, 1 fly) and *D. phalerata* (2.4%, 1 fly), respectively. We therefore think it likely that *Drosophila*-associated Adintovirus 1 and 2 are associated with those two species rather than *D. melanogaster*, and may potentially be integrated into their genomes. These sequences have been submitted to Genbank under accessions MT496847 and MT496848.

In contrast, *Drosophila*-associated Adintovirus 3 was assembled using sample DK_Kar_16_4 from Karensminde, Denmark (SRR8494437), from which other members of the Drosophilidae were absent. It is similarly 13.8 kb long, and our phylogenetic analysis of DNA PolB places it within the published diversity of insect Adintoviruses—although divergent from other Adintoviruses or polintons of *Drosophila* (Figure 2C; Starrett *et al.* 2020). However, this sequence is only predicted to encode 10 proteins and these are generally more divergent, perhaps suggesting that this virus is associated with a completely different host species, such as the nematode related to *Panagrellus redivivus* or a trypanosomatid—although these species were present at very low levels. The sequence has been submitted to Genbank under accession MT496849

Prevalence varies among viruses, and in space and time

Based on a detection threshold of 1% of the *Drosophila* genome copy-number, only five of the viruses (Kallithea virus, Vesanto virus, Linvill Road virus, Viltain virus and Esparto virus) were detectable in multiple population pools, with the other nine viruses each detectable in only a single pool. For viruses in a single pool, a simple maximum-likelihood estimate of prevalence—assuming independence of flies and pools—is 0.015% (with an upper 2-Loglikelihood bound of 0.07%). Among the intermediate-prevalence viruses, Esparto virus and Viltain virus were detected in 5 pools each, corresponding to a prevalence of 0.08% (0.03-0.17%), and Linvill road virus was detected in 21 pools, indicating a prevalence of 0.34% (0.21-0.51%). The two most common viruses were Kallithea virus, which was detected in 93 pools giving a prevalence estimate of 2.1% (1.6-2.5%), and Vesanto virus, which was detected in 114 pools giving a prevalence estimate of 2.9% (2.4-3.5%)

Kallithea virus, Vesanto virus, and Linvill Road virus were sufficiently prevalent to analyse their presence / absence across populations using a Bayesian spatial Generalised Linear Mixed Model. Our analysis identified a spatial component to the distribution of both Kallithea and Linvill Road viruses that did not differ significantly between years, with a higher prevalence of Kallithea virus in southern and

central Europe, and a higher prevalence of Linvill Road virus in Iberia (Figure 5A and B; ΔDIC of -13.6 and -17.2, respectively, explaining 15.5% and 32.8% of the variance). In contrast, Vesanto virus showed no detectable spatial variation in prevalence, but did vary significantly over time, with a significantly lower prevalence in 2014 compared to the other years (2015 and 2016 were larger by 1.27 [0.42,2.16] and 1.43 [0.50,2.14] respectively). The probability of observing a virus did not depend on the sampling season or the level of *Wolbachia* infection. However, the probability of detecting Linvill Road virus was positively correlated with the level of *D. simulans* contamination (95% credible interval the log-odds ratio [2.9,14.6]). This may suggest that some of these reads actually derived from infections of *D. simulans* (in which the virus can have very high prevalence, see data from Signor *et al.* 2017), or that infections in *D. melanogaster* may be associated with spill-over from *D. simulans*. Sampling location did not explain any significant variation in the probability of detecting any virus, suggesting that—beyond broad geographic trends—there is little temporal consistency in virus prevalence at the small scale.

DNA viruses are detectable in publicly available Drosophila datasets

We wished to corroborate our claim that these viruses are associated with *Drosophila* by exploring their prevalence in laboratory populations and publicly available data. We therefore examined the first 10 million reads of 3003 sequencing runs from 28 *D. melanogaster* and *D. simulans* sequencing projects. In general, our survey suggests that studies using isofemale or inbred laboratory lines tend to lack DNA viruses (e.g., Mackay *et al.* 2012, Grenier *et al.* 2015, Lack *et al.* 2015, Gilks *et al.* 2016, Lange *et al.* 2016). In contrast, studies that used wild-caught or F1 flies (e.g., Endler *et al.* 2018, Machado *et al.* 2019) or large population cages (e.g., Schou *et al.* 2017) were more likely to retain DNA viruses (Supplementary File S7).

Based on our detection thresholds, none of the public datasets we examined appeared to contain Mauternbach virus, Yalta virus, *Drosophila*-associated filamentous virus, *Drosophila*-associated hytrosa virus, or the three *Drosophila*-associated adintoviruses (Supplementary File S7). This is consistent with their extreme rarity in our own sampling, and the possibility that Mauternbach virus and the adinto-like viruses may actually infect species other than *D. melanogaster*. Although some reads from Dros-RTEC run SRR3939056 (99 flies from Athens, Georgia; Machado *et al.* 2019) did map to an adintovirus, these reads actually derive from a distinct virus that has only 82% nucleotide identity to *Drosophila*-associated adintovirus-1. Unfortunately, this closely-related adintovirus cannot corroborate the presence of *Drosophila*-associated adintovirus-1 in *D. melanogaster*, as run SRR3939056 is contaminated with *Scaptodrosophila latifasciaeformis*, which could be the host.

One of our rare viruses was present (but rare) in public data: Viltain virus appeared only once in 3003 sequencing datasets, in one of the 63 libraries from Dros-RTEC project PRJNA308584 (Machado *et al.* 2019). Tomelloso virus, which was rare in our data, was more common in public data, appearing in 5 of 28 projects and 23 of 3003 runs. However, this may explained by its presence in multiple runs from each of a small number of experimental studies (e.g., Liu and Secombe 2015, Siudeja *et al.* 2015, Fang *et al.* 2017, Riddiford *et al.* 2020). Our three most common viruses were also the most common DNA viruses in public data. Linvill Road virus appeared in 10 of the 28 projects we examined, including 363 of the 3003 runs. This virus was an exception to the general rule that DNA viruses tend to be absent from inbred or long-term laboratory lines, as it was detectable in 166 of 183 sequencing runs of inbred *D. simulans* (Signor *et al.* 2017). Kallithea virus appeared in four of the 28 projects, including 60 of the runs, and was detectable in wild collections of both *D. melanogaster* and *D. simulans*. Vesanto virus was detectable in eight of the 28 projects, including 208 of the runs, but only in *D. melanogaster* datasets.

The presence of Vesanto virus segments in public data is of particular value because it could help to elucidate patterns of segment co-occurrence. This virus was highly prevalent in a large experimental evolution study using caged

populations of *D. melanogaster* derived from collections in Denmark in 2010 (Schou *et al.* 2017), where segments S01, S02, S04, S05 and S10 were almost always present, S03, S06, S07 and S08 were variable, and S09, S11 and S12 were always absent. However, because these data were derived from RAD sequencing, absences may reflect absence of the restriction sites. Vesanto virus also appeared in Pooled-GWAS datasets (e.g., Endler *et al.* 2018), for which segments S09 and S12 were always absent and segments S03, S10 and S11 were variable (Supplementary File S7), and in several Dros-RTEC datasets (Machado *et al.* 2019) in which only S12 was consistently absent. Unfortunately, it is difficult to test among the competing hypotheses using pooled sequencing of wild-collected flies or large cage cultures. This is because different flies in the pool may be infected with different viruses or with viruses that have a different segment composition, and because a more complex microbiome may be present. However, we were able to find one dataset from an isofemale line, GA10 collected in Athens, Georgia (USA) in 2009, that had been maintained in the laboratory for at least five generations prior to sequencing (Supporting File S6; ERR705977 from Bergman and Haddrill 2015). From this dataset we assembled 8 of the 12 segments, including two segments encoding PolB-like proteins and two encoding the DUF3472 protein. Mapping identified no reads at all from segments S9 or S12. This most strongly supports a single virus with a variable segment composition between infections and/or re-assortment. Moreover, the low species complexity of this laboratory dataset supports *D. melanogaster* as the host, with over 98% of reads mapped, and with *Drosophila*, *Wolbachia* and *Lactobacillus plantarum* the only taxa present in appreciable amounts. Example Vesanto virus sequences from these datasets are provided in Supplementary file S6.

Genetic diversity varies among viruses and populations

We examined genetic variation in three of the most common viruses; Kallithea virus, Linvill Road virus and Vesanto virus. After masking regions containing indels, and using a 1% minor allele frequency (MAF) threshold for inclusion, we identified 923 single nucleotide polymorphisms (SNPs) across the total global Kallithea virus pool, and 15132 distinct SNPs summed across the 44 population samples. Of these SNPs, 13291 were private to a single population, suggesting that the vast majority of Kallithea SNPs are globally and locally rare and limited to one or a few populations. This is consistent with many of the variants being recent and/or deleterious, but could also reflect a large proportion of sequencing errors—despite the analysis requiring a MAF of 1% and high base quality. Synonymous pairwise genetic diversity in the global pool was very low, at $\pi_S = 0.15\%$, with π at intergenic sites being almost identical (0.14%). Diversity did not vary systematically around the virus genome (Supplementary File S9). Consistent with the large number of low-frequency private SNPs, average within population-pool diversity was 10-fold lower still, at $\pi_S = 0.04\%$, corresponding to a very high F_{ST} of 0.71. In general, the level of constraint on virus genes seemed low, with global π_A/π_S 0.39 and local $\pi_A/\pi_S = 0.58$. These patterns of diversity are markedly different to those of the host, in which π_S (fourfold sites) is on the order of 1% with π_A/π_S (zero-fold and four-fold) around 0.2, and differentiation approximately $F_{ST} = 0.03$ (Tristan *et al.* 2019, Kapun *et al.* 2020). Given that large dsDNA virus mutation rates can be 10-100 fold higher than animal mutation rates (Duffy 2018), the overall lower diversity in Kallithea virus is consistent with bottlenecks during infection and the smaller population size that corresponds to a 2.1% prevalence. The very low within-population diversity and high F_{ST} and π_A/π_S may be indicative of local epidemics, or a small number of infected hosts within each pool (expected to be 1.47 infections in an infected pool, assuming independence) with relatively weak constraint. Alternatively, high F_{ST} and π_A/π_S may indicate a high proportion of sequencing errors.

In Vesanto virus we identified 4059 SNPs across all segments and divergent segment haplotypes in the global pool, with 5491 distinct SNPs summed across all infected population samples, of which 4235 were private to a single population. This corresponded to global and local diversity that was around 7-fold higher than Kallithea virus (global $\pi_S = 1.16\%$, local $\pi_S = 0.24\%$), and to

much higher levels of constraint on the protein sequence (π_A/π_S = 0.10), but a similar level of differentiation (F_{ST}=0.76). Although the prevalence of Vesanto virus appears to be slightly higher than Kallithea virus (2.9% vs. 2.1%), much of the difference in diversity is probably attributable to the higher mutation rates of ssDNA viruses (Duffy 2018). The apparent difference in the allele frequency distribution between these two viruses is harder to explain (73% of SNPs detectable at a global MAF of 1%, versus only 6% in Kallithea virus), but could be the result of the very strong constraint on protein coding sequences keeping non-synonymous variants below the 1% MAF threshold even within local populations. It is worth noting that the difference between Vesanto virus and Kallithea virus in π_A/π_S and the frequency of rare alleles argues against their being purely a result of sequencing error in Kallithea virus, as the error rates would be expected to be similar between the two viruses,

In Linvill road virus, which was only present in 13 populations and has the smallest genome, we identified 178 SNPs across the global pool, and 253 distinct SNPs summed across the infected populations, of which 209 were private to a single population. Although this virus appears at least 6-fold less prevalent than Kallithea virus or Vesanto virus, it displayed relatively high levels of genetic diversity both globally and locally (global π_S = 1.45%, local π_S = 0.21%, F_{ST}=0.86), and intermediate levels of constraint (π_A/π_S = 0.20). Given a mutation rate that is likely to be similar to that of Vesanto virus, this is hard to reconcile with a prevalence that is 6-fold lower. However, one likely explanation is that Linvill Road virus is more prevalent in the sister species *D. simulans* (above), and the diversity seen here represents rare spill-over and contamination of some samples with that species.

Structural variation and transposable elements in Kallithea virus

De novo assembly of Kallithea virus from each sample resulted in 52 populations with complete single-scaffold genomes that ranged in length from 151.7 kbp to 155.9 kbp. Alignment showed these population-consensus assemblies to be co-linear with a few short duplications of 10-100 nt, but generally little large-scale duplication or rearrangement. Two regions were an exception to this: that spanning positions 152,180 to 152,263 in the circular reference genome (between putative proteins AQN78547 and AQN78553; genome KX130344.1), and that spanning 67,903 to 68,513 (within putative protein AQN78615). The first region comprised multiple repeats of around 100 nt and assembled with lengths ranging from 0.2 to 3.6 kbp, and the second comprised multiple repeats of around 140 nt and assembled with lengths between 0.5 and 2.4 kbp. Together, these regions explained the majority of the length variation among the Kallithea virus genome assemblies. We also sought to catalogue small-scale indel variation in Kallithea virus by analysing indels within reads. In total, after indel-realignment using GATK, across all 44 infected samples we identified 2289 indel positions in the Kallithea virus genome that were supported by at least 5 reads. However, only 195 of these indels were at high frequency (over 50% of samples). As would be expected, the majority (1774) were found in intergenic regions (Supplementary File 9).

Pooled assemblies can identify structural variants that differ in frequency among populations, but they are unlikely to identify rare variants within populations, such as those caused by TE insertions. TEs are commonly inserted into large DNA viruses, and these viruses have been proposed as a vector for interspecies transmission of TEs (Gilbert *et al.* 2016, Gilbert and Cordaux 2017). In total we identified 5,169 read pairs (across 16 datasets with >300-fold coverage of Kallithea virus) that aligned to both *D. melanogaster* TEs and Kallithea virus. However, the vast majority of these (5,124 out of 5,169) aligned internally to TEs, more than 5 bp away from the start or end position of the TE, which is inconsistent with insertion (Gilbert *et al.* 2016, Loiseau *et al.* 2020). Instead, this pattern suggests PCR-mediated recombination, and assuming that all chimeras we found were artefactual, their proportion among all reads mapping to the Kallithea virus (0.01%) falls in the lower range of that found in other studies (Peccoud *et al.* 2018). We therefore believe there is no evidence supporting *bona fide* transposition of *D. melanogaster* TEs into

genomes of the Kallithea virus in these natural virus isolates. This is in striking contrast to what was found in the AcMNPV nucleopolyhedrovirus (Loiseau *et al.* 2020) and could perhaps reflect the tropism of Kallithea virus (Palmer *et al.* 2018) to tissues that experience low levels of transposition.

A genomic insertion of Galbut virus is segregating in D. melanogaster.

The only RNA virus we identified among the DNA reads from DrosEU collections was Galbut virus, a segmented and bi-parentally vertically-transmitted dsRNA *Partitivirus* that is extremely common in *D. melanogaster* and *D. simulans* (Webster *et al.* 2015, Cross *et al.* 2020). Based on a detection threshold of 0.1% of fly genome copy number, Galbut virus reads were present in 43 out of 167 samples. There are two likely sources of such DNA reads from an RNA virus in *Drosophila*. First, reads might derive from somatic circular DNA copies that are created as a part of the immune response (Mondotte *et al.* 2018, Poirier *et al.* 2018). Second, reads might derive from a germline genomic integration that is segregating in wild populations (i.e., an Endogenous Viral Element, or EVE; Katzourakis and Gifford 2010, Tassetto *et al.* 2019). We sought to distinguish between these possibilities by *de novo* assembly of the Galbut sequences from high copy-number DrosEU samples and public *D. melanogaster* DNA datasets.

We assembled the Galbut virus sequence from the three DrosEU samples in which it occurred at high depth: BY_Bre_15_13 (Brest, Belarus), PO_Gda_16_16 (Gdansk, Poland), and PO_Brz_16_17 (Brzezina, Poland). We were also able to assemble the sequence from four publicly available sequencing runs: three (SRR088715, SRR098913 and SRR1663569) that we believe are derived from global diversity line N14 (Grenier *et al.* 2015) collected in The Netherlands in 2002 (Bochdanovits and de Jong 2003), and SRR5762793, which was collected in Italy in 2011 (Mateo *et al.* 2018). In every case, the assembled sequence was the same 1.68 kb near full-length copy of Galbut virus segment S03, including the whole of the coding sequence for the viral RNA-dependent RNA polymerase. Also in every case this sequence was inserted into the same location (i.e., identical breakpoints), around 400 bp from the 5' end of a 297 Gypsy-like LTR retrotransposon. This strongly suggests that the Galbut sequence represents a unique germline insertion as, even if the insertion site used in the immune response were constant, the inserted virus sequence would be highly variable across Europe over 14 years. The sequence falls among extant Galbut virus sequences (Figure 6B), and is 5% divergent (18.5% synonymous divergence) from the closest one available in public data. The sequences are provided in Supplementary File S10

Interestingly, populations with a substantial number of Galbut reads (a maximum of 13.8% or 11 chromosomes of 80) appeared geographically limited, appearing more commonly in higher latitudes, and with a different spatial distribution in the early and late collecting seasons (ΔDIC = 26.92; Figures 5C, 6A). Given the absence of this sequence from Dros-RTEC (Machado *et al.* 2019), DGRP (Mackay *et al.* 2012) and the other *Drosophila* Genome Nexus datasets (Lack *et al.* 2015, Lange *et al.* 2016), it seems likely that this insertion is of a recent, likely northern or central European, origin. We used a strict-clock phylogenetic analysis of viral sequences to estimate that the insertion occurred within the last 300 years (posterior mean 138 years ago, 95% highest posterior density interval 20-287 years ago; Figure 6B), i.e. after *D. melanogaster* was spreading within Europe. Unfortunately, the insertion site in a high copy-number transposable element means that we were unable to locate it in the genome. This also means that it was not possible to detect whether the insertion falls within a piRNA-generating locus, which is seen for several endogenous viral elements (EVE) in mosquitoes (Palatini *et al.* 2017) and could perhaps provide resistance to the vertically transmitted virus. Surprisingly, DNA reads from Galbut virus were more likely to be detected at sites with a higher percentage of reads mapping to *Wolbachia* (95% credible interval for the effect [0.074,0.41]; ΔDIC = -5.52). Given that no correlation between Galbut virus and *Wolbachia* has been detected in the wild (Webster *et al.* 2015, Shi, White*, et al.* 2018), we think this most likely reflects a chance association between the

geographic origin of the insertion and the spatial distribution of *Wolbachia* loads (Kapun *et al.* 2020).

Conclusions

Although metagenomic studies are routinely used to identify viruses and virus-like sequences (e.g., Shi, Zhang*, et al.* 2018, Zhang *et al.* 2018), simple bulk sequencing can only show the presence of viral sequences: it cannot show that the virus is replicating or transmissible, nor can it unequivocally identify the host (reviewed in Obbard 2018). This behoves metagenomic studies to carefully consider any additional evidence that might add to, or detract from, the claim that an 'associated virus-like sequence' is indeed a virus. A couple of the DNA viruses described here undoubtedly infect *Drosophila*. Kallithea virus has been isolated and studied experimentally (Palmer *et al.* 2018), and Tomelloso virus is detectable in some long-term laboratory cultures (e.g. Liu and Secombe 2015, Siudeja *et al.* 2015, Fang *et al.* 2017, Riddiford *et al.* 2020). Others, such as Viltain virus, Linvill Road virus, and Vesanto virus, are present at such high copy numbers, and sometimes in laboratory cultures, that a host other than *Drosophila* seems very unlikely. Some, appearing at reasonable copy number but in a single sample, could be infections of contaminating *Drosophila* species (Mauternbach virus, the adinto-like viruses), or spill-over from infections of parasitoid wasps (Yalta virus, the filamentous virus). A few, having appeared at low copy number in a single sample, could be contaminants—although we excluded virus-like sequences that appeared strongly associated with contaminating taxa (Supplementary File S4).

These caveats aside, along with the Nudivirus of *Drosophila innubila* (Unckless 2011) and Invertebrate iridescent virus 31 in *D. obscura* and *D. immigrans* (Webster *et al.* 2016), our study brings the total number of published DNA viruses associated with *Drosophila* to sixteen. Although a small sample, these viruses start to hint at some interesting natural history. First, it is striking that more than a third of the reported DNA viruses are Nudiviruses (six of the 16 published, plus a seventh from *Phortica variegata*; Figure 2). This suggests that the Nudiviridae are common pathogens of *Drosophila*, and may indicate long-term host lineage fidelity with short-term switching among species. Such switching is consistent with the lack of congruence between host and virus phylogenies, and the fact that both *D. innubila* Nudivirus and Kallithea virus infect multiple *Drosophila* species (Figure 2). Second, the majority of DNA viruses seem to be rare. Seven of the 12 viruses confidently ascribable to *D. melanogaster* or *D. simulans* were detected in just one of the 167 population samples, and likely only one of 6668 flies, consistent with a European prevalence less than 0.07%. Only Vesanto virus and Kallithea virus seem common, being detected in more than half of populations and having estimated prevalences of 2.9% and 2.1%, respectively. It is unclear why DNA viruses should have such a low prevalence, on average, as compared to RNA viruses (Webster *et al.* 2015). One possibility is that DNA virus infections are less likely to be persistent than RNA virus infections, or that they reduce fly fitness to the extent that infected flies are less likely to be sampled, reducing the chances of detection. This may also explain why they rarely persist through multiple generations in laboratory fly lines. Alternatively, it may be that the rare viruses represent dead-end spill-over from other taxa that can only be seen here because of the large sample size. Third, although some viruses showed broad geographic patterns in prevalence, a lack of repeatability associated with sampling location and the very high F_{ST} values hint that transient local epidemics may be the norm, with viruses frequently appearing, and then disappearing, from local fly populations.

Finally, it also appears that *Drosophila* do indeed harbour fewer DNA viruses than RNA viruses, supporting an observation that was made before any had been described (Brun and Plus 1980, Huszart and Imler 2008). This cannot be an artefact of reduced sampling effort, as almost all *Drosophila*-associated viruses have been reported from undirected metagenomic studies, and metagenomic studies of RNA are as capable of detecting expression from DNA viruses as they are of detecting RNA viruses (e.g., Webster *et al.* 2015). Instead, it suggests that the imbalance

must reflect some aspect of host or virus biology. For example, it may be a consequence of differences in prevalence: if RNA viruses have higher prevalence in general, or specifically in those adult flies attracted to baits, and/or RNA viruses persist more easily in fly or cell cultures, then this may explain their more frequent detection. Taken together, our analyses of the distribution and diversity of DNA viruses associated with *Drosophila melanogaster* at the pan-European scale provide an essential ecological and evolutionary context for future studies of host-virus interaction in *Drosophila*.

Acknowledgements

We thank all of the members of the DrosEU and Dros-RTEC communities for their ongoing engagement in this collaborative European project. We are especially grateful to the teachers Antonio J. Buendía, Ma Josefa Gómez, Ma Luisa Espinosa and the students of the IES Eladio Cabañero (Tomelloso, Spain), and to the teachers David González, Silvana Castillo and the students of the IES José de Mora (Baza, Spain), who contributed to fly collections in 2016 as part of the “Melanogaster Catch the Fly” citizen science project. We thank Alex Twyford for providing computing time, and Lewis Stevens and Andrew Rambaut for their help with MinION sequencing.

Funding

Megan Wallace was supported by the UK Natural Environmental Research Council through the E3 doctoral training programme (NE/L002558/1), and Sanjana Ravindran was supported by Wellcome Trust PhD programme (108905/Z/15/Z).

Andrea Betancourt received funding from BBSRC grant BB/P00685X/1

Thomas Flatt received funding from Swiss National Science Foundation grants 31003A-182262, PP00P3_165836, and PP00P3_133641/1.

Clément Gilbert received funding from Agence Nationale de la Recherche (grant ANR-15-CE32-0011-01)

Josefa González received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (H2020-ERC-2014-CoG-647900) and from the Fundación Española para la Ciencia y la Tecnologia-Ministerio de Economía y Competitividad (FCT-15-10187).

Sonja Grath received funding from Deutsche Forschungsgemeinschaft grant GR 4495/2

Maaria Kankare received funding from Academy of Finland projects 268214 and 322980.

Martin Kapun received funding from Austrian Science Fund (FWF) grant P32275.

Volker Loeschcke received funding from Danish Research council for natural Sciences (FNU) grant nr 4002-00113B

Banu Sebnem Onder received funding from the Scientific and Technological Research Council of Turkey (TUBITAK) (Grant No. 214Z238)

John Parsch received funding from Deutsche Forschungsgemeinschaft grant PA 903/8

Fabian Staubach received funding from Deutsche Forschungsgemeinschaft grant STA1154/4-1; Projektnummer 408908608

The DrosEU consortium has been funded by a Special Topics Network (STN) grant by the European Society of Evolutionary Biology (ESEB).

Tables

Table 1: DNA Viruses of *Drosophila* present in the DrosEU dataset

Virus name	Relationship	Genome status	Assembly (bp)	First published	Genbank accessions	DrosEU code	Genome collection location	Source data
Kallithea virus	Nudiviruses	Complete	152,388	Webster et al. (2015)	KX130344	UA_Kha_14_46	Kharkiv, Ukraine	SRR5647730
Esparto virus	Nudiviruses	Complete	183,261	Kapun et al. (2020)	KY608910	(Dros-RTEC)	Esparto, California	SRR3939042
Viltain virus	Densoviruses	Complete	5,025	Kapun et al. (2020)	KX648535	FR_Vil_14_07	Viltain, France	SRR5647729
Linvill Road virus	Densoviruses	Complete	5,360	Kapun et al. (2020) Expanded here	MT364305	ES_Ben_16_32	Benalua, Spain	SRR8494475
Vesanto virus	Bidnaviruses	Up to 12 Segments	Up to 52,097	Kapun et al. (2020) Expanded here	MT496850-MT496878	UA_Kan_16_57 UA_Dro_16_56 FR_Got_15_48	Kaniv, Ukraine Drogobych, Ukraine Gotheron, France	SRR8494448 SRR8494441 SRR8439123
Tomelloso virus	Nudiviruses	Complete	112,307	This Paper	KY457233	ES_Tom_15_28	Tomelloso, Spain	SRR8439136
Mauternbach virus	Nudiviruses	Complete	154,465	This Paper	MG969167	AT_Mau_15_50	Mauternbach, Austria	SRR8439127
Yalta virus	Entomopox viruses	Complete	219,929	This Paper	MT364305	UA_Yal_14_16	Yalta, Ukraine	SRR5647764
Drosophila-associated Nudivirus (Kaniv)	Nudiviruses	Fragmentary	4,503	This Paper	MT496841-MT496846	UA_Kan_16_57	Kaniv, Ukraine	SRR8494448
Drosophila-associated Filamentous virus	Filamentous viruses	Fragmentary	86,478	This Paper	MT496832-MT496840	ES_Gim_15_30	Gimenells, Spain	SRR8439138
Drosophila-associated Hytrosavirus	Hytrosaviruses	Fragmentary	16,606	This Paper	MT469997-MT470014	UA_Ode_16_47	Odesa, Ukraine	SRR8494427
Drosophila-associated Adintovirus 1	Adinto-like viruses	Complete	14,567	This Paper	MT496847	UA_Cho_15_26	Kopachi (Chornobyl Exclusion Zone), Ukraine	SRR8439134
Drosophila-associated Adintovirus 2	Adinto-like viruses	Near-Complete	13,277	This Paper	MT496848	AT_Mau_15_50	Mauternbach, Austria	SRR8439127
Drosophila-associated Adintovirus 3	Adinto-like viruses	Near-Complete	13,883	This Paper	MT496849	DK_Kar_16_4	Karensminde, Denmark	SRR8494437

Figures and Legends

Figure 1: Genome structures and read depth. The plots show annotated coding DNA sequences (CDS, red and blue arrows), and terminal Inverted repeat (yellow boxes) for each of the near-complete virus genomes discussed. The read depth (pale blue) is plotted above the genome on a log scale for the population with the highest coverage in the DrosEU dataset. The five largest viruses (top) are plotted according to the 20 kbp scale bar, and the other viruses (bottom) are plotted according to the 2 kbp scale bar. The Nudiviruses are circular, and have been arbitrarily linearized for plotting. Esparto virus was completed using public dataset (SRR3939042). Note that Vesanto virus segments S07 and S11 were absent from the illustrated sample (lower right).

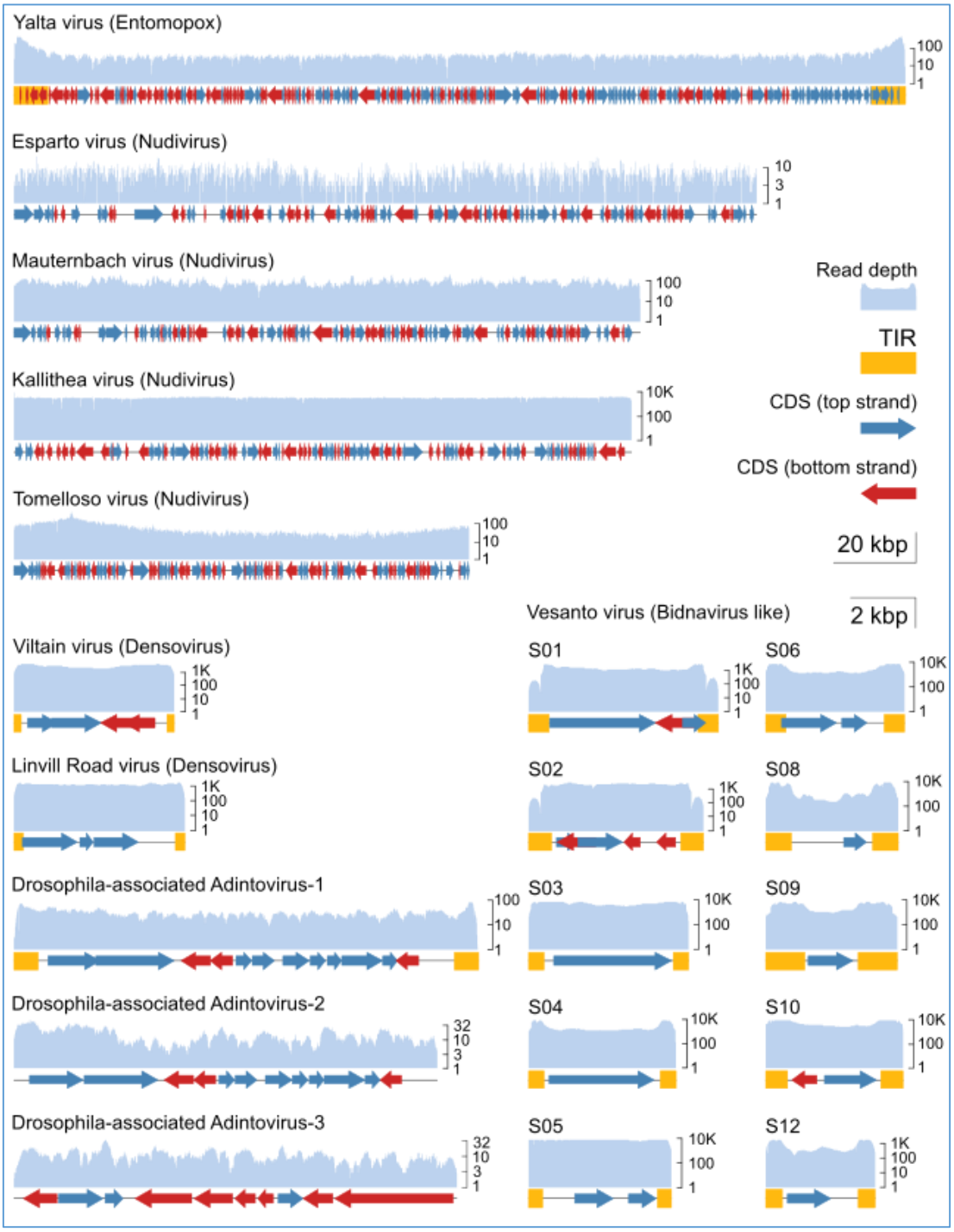
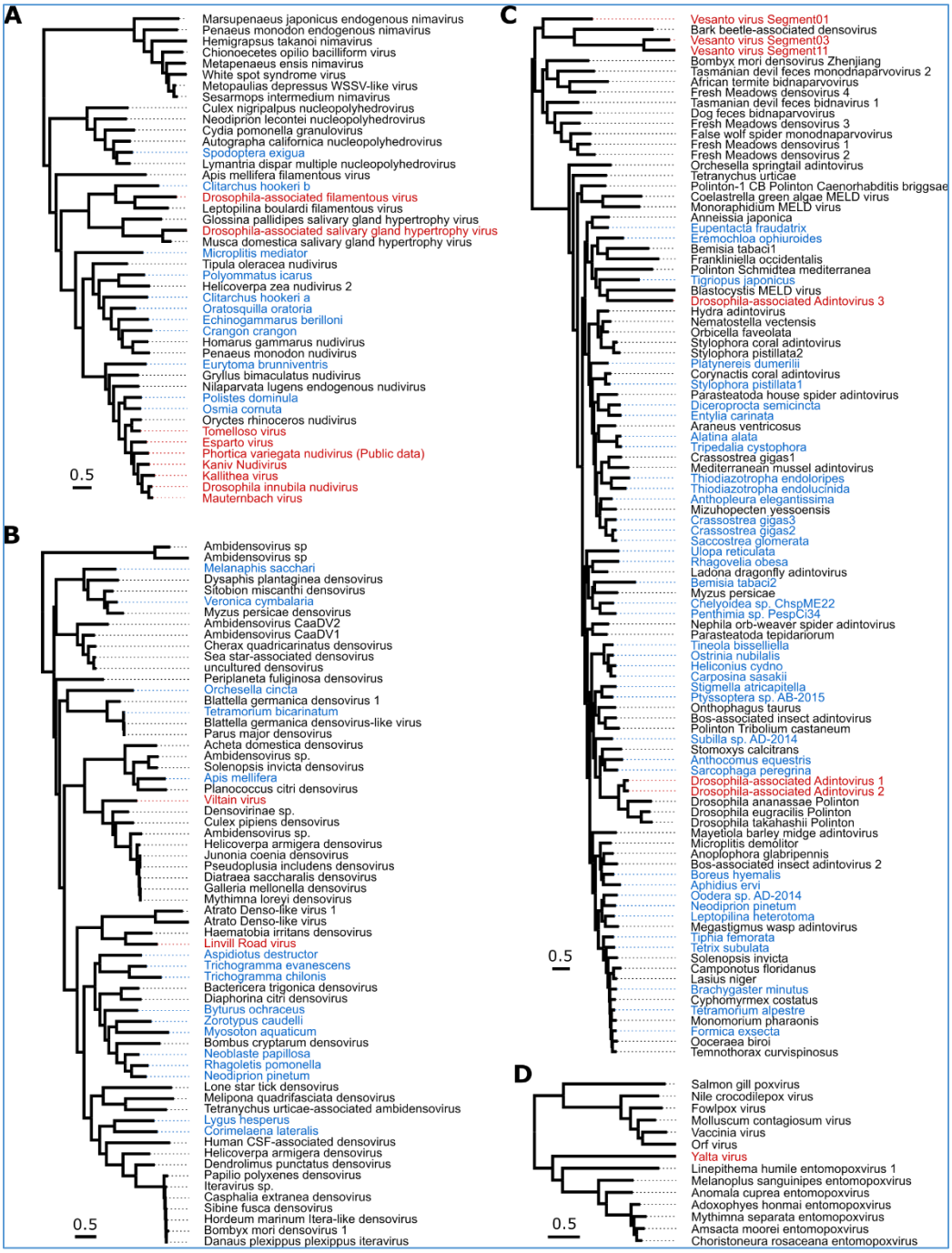


Figure 2: Phylogenetic relationships. (A) Nudiviruses, Hytrosaviruses, Filamentous viruses, Nucleopolyhedrosis viruses and Nimaviruses, inferred from six concatenated protein coding genes. Note that these lineages are extremely divergent, and the alignment is not reliable at deeper levels of divergence. (B) Densoviruses, inferred from NS1. (C) Bidnaviruses (sometimes labelled 'Densovirus') and Adintoviruses (including representative Polintons), inferred from DNA Polymerase B. (D) Pox and Entomopox viruses, inferred from three concatenated protein coding genes. All phylogenies were inferred from protein sequences by maximum likelihood, and scale bars represent 0.5 amino-acid substitutions per site. In each case, trees are mid-point rooted, viruses reported from *Drosophila* are shown in red, and sequences identified from virus transcripts in publicly-available transcriptome assemblies are shown in blue, labelled by host species. The Nudivirus from *Phortica variegata* was derived from PRJNA196337 (Vicoso and Bachtrog 2013). Alignments and tree files with bootstrap support will be made available via Figshare.



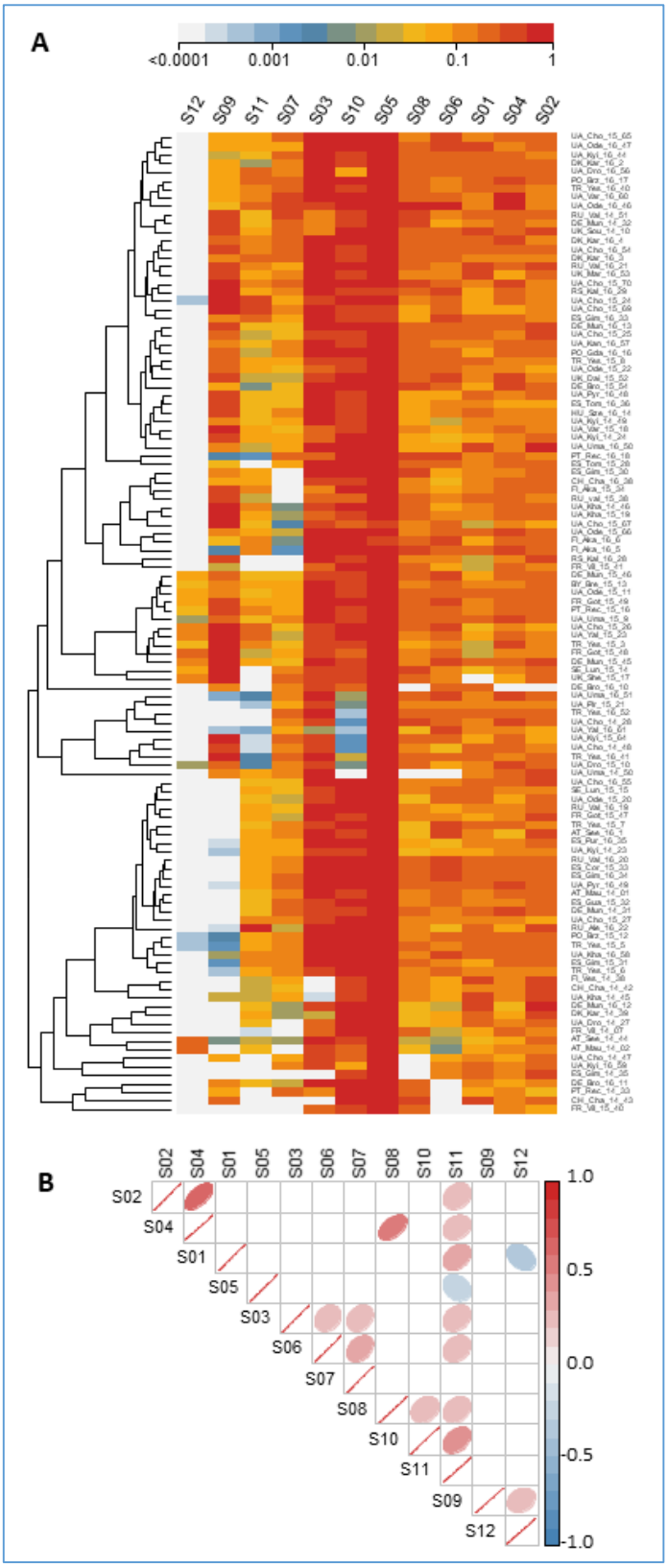


Figure 3: Vesanto virus segment copy-number. (A) Heatmap showing the relative number of sequencing reads from each of the 12 Vesanto virus segments (columns), for each of the population samples (rows). Populations are included if at least one segment appeared at 1% of the fly genome copy-number. Rows and columns have been ordered by similarity (dendrogram) to identify structure within the data. Colours show copy-number relative to the highest-copy segment, on a log scale. (B) Correlations in copy-number among the segments, with 'significant' correlations ($p<0.05$, no corrections) shown with coloured ellipses, according to the direction (red positive, blue negative) and strength of correlation. The absence of strong negative correlations between segments encoding homologous proteins (e.g. S01,S03,S11, which all encode genes with homology to DNA Polymerase B) may indicate that these segments do not substitute for each other.

Figure 4: Geographic distribution of DNA virus reads in European *D. melanogaster*. Maps show the spatial distribution of virus read copy-number (relative to fly genomes) on a non-linear colour scale. Data are shown for the five viruses that were detected more than once (rows), separated by year and whether flies were collected relatively 'early' or 'late' in the season (columns).

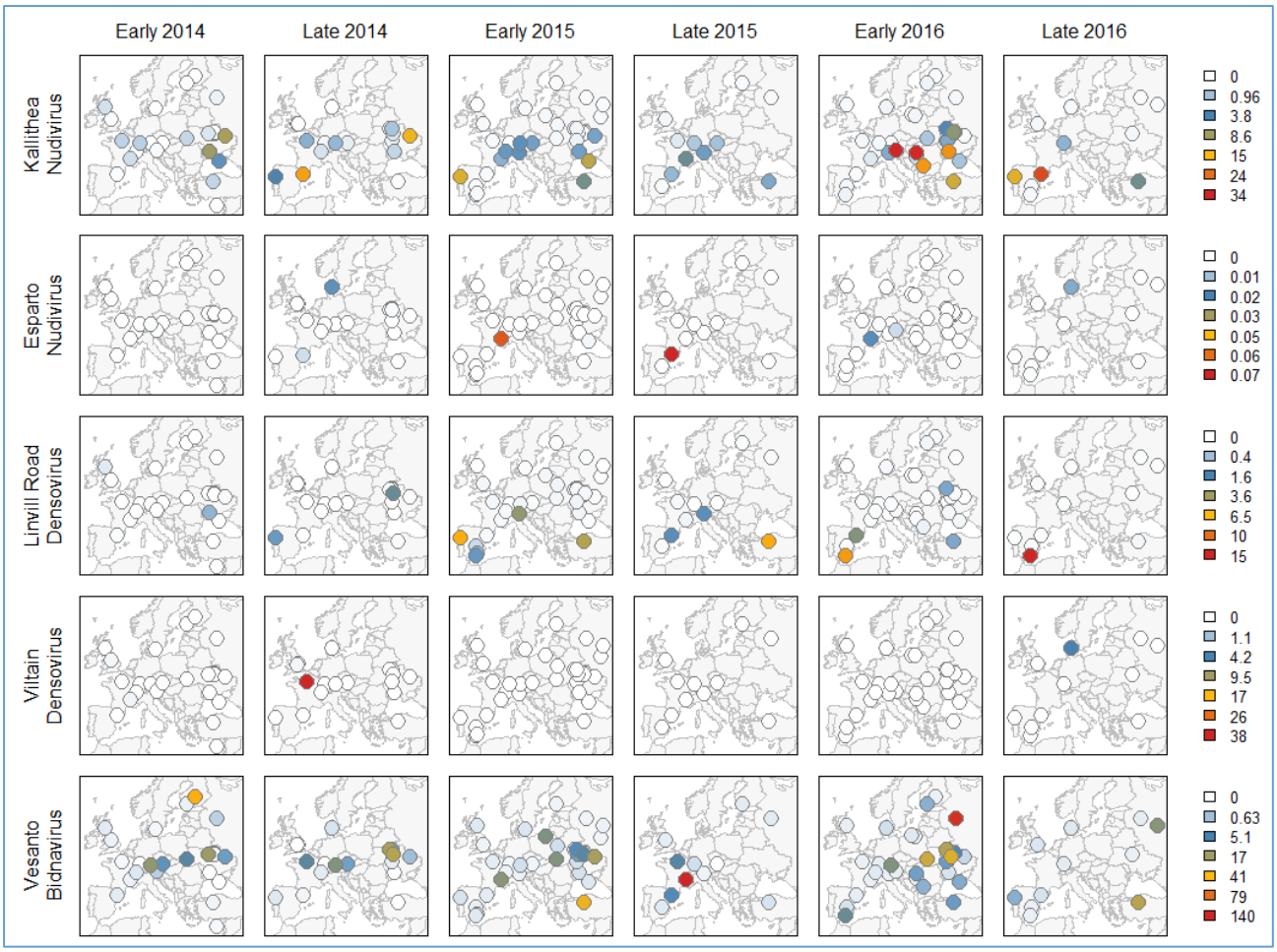


Figure 5: **Geographic variation in estimated prevalence:** Kallithea virus (A), Linvill Road virus (B), and the Galbut EVE (C and D). Sampling sites are marked as white dots, and the colour gradient illustrates predictions from the INLA model, but with scale transformed to the predicted individual-level prevalence (%), assuming independence among individuals and population samples of size 40. Only Kallithea virus, Linvill road virus, and the Galbut EVE displayed a significant spatial component, and only the EVE differed between seasons.

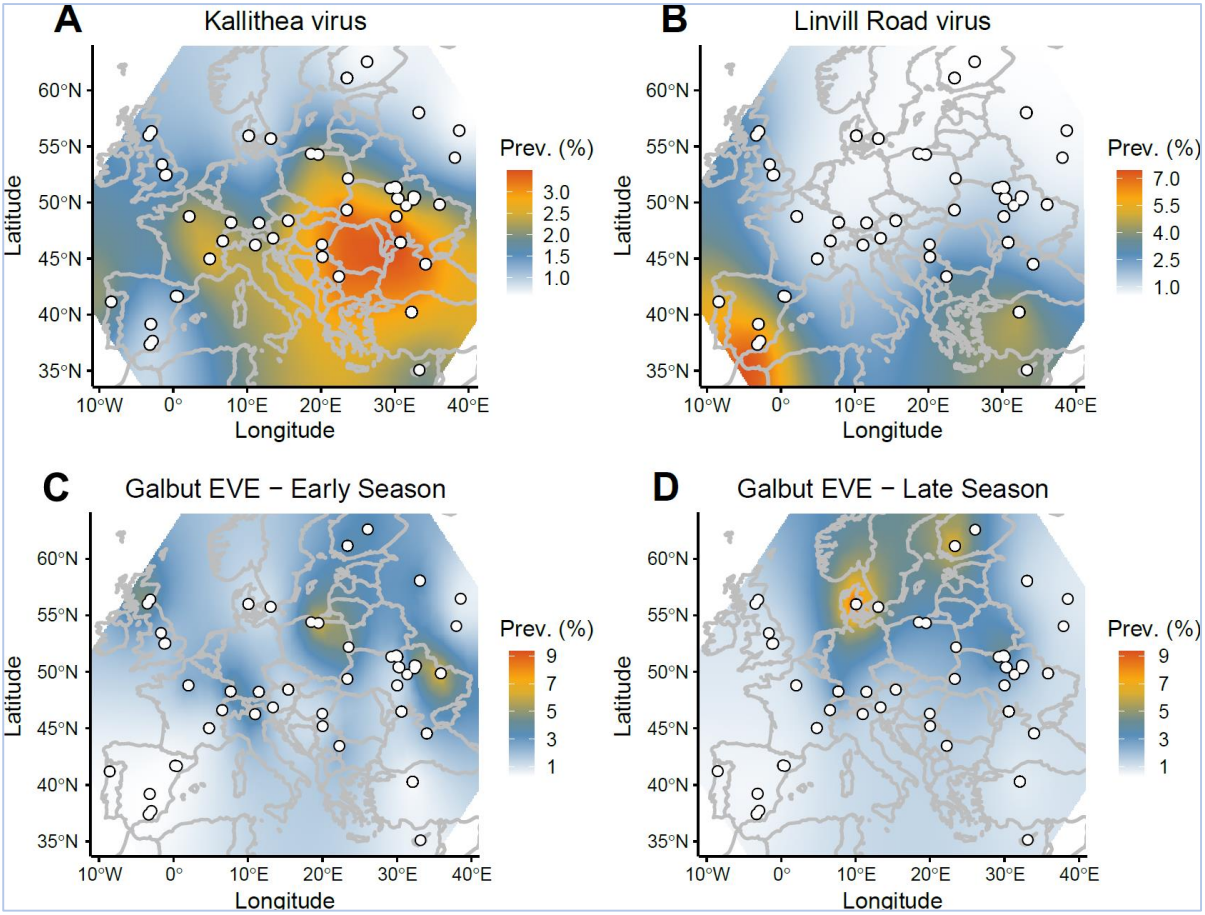
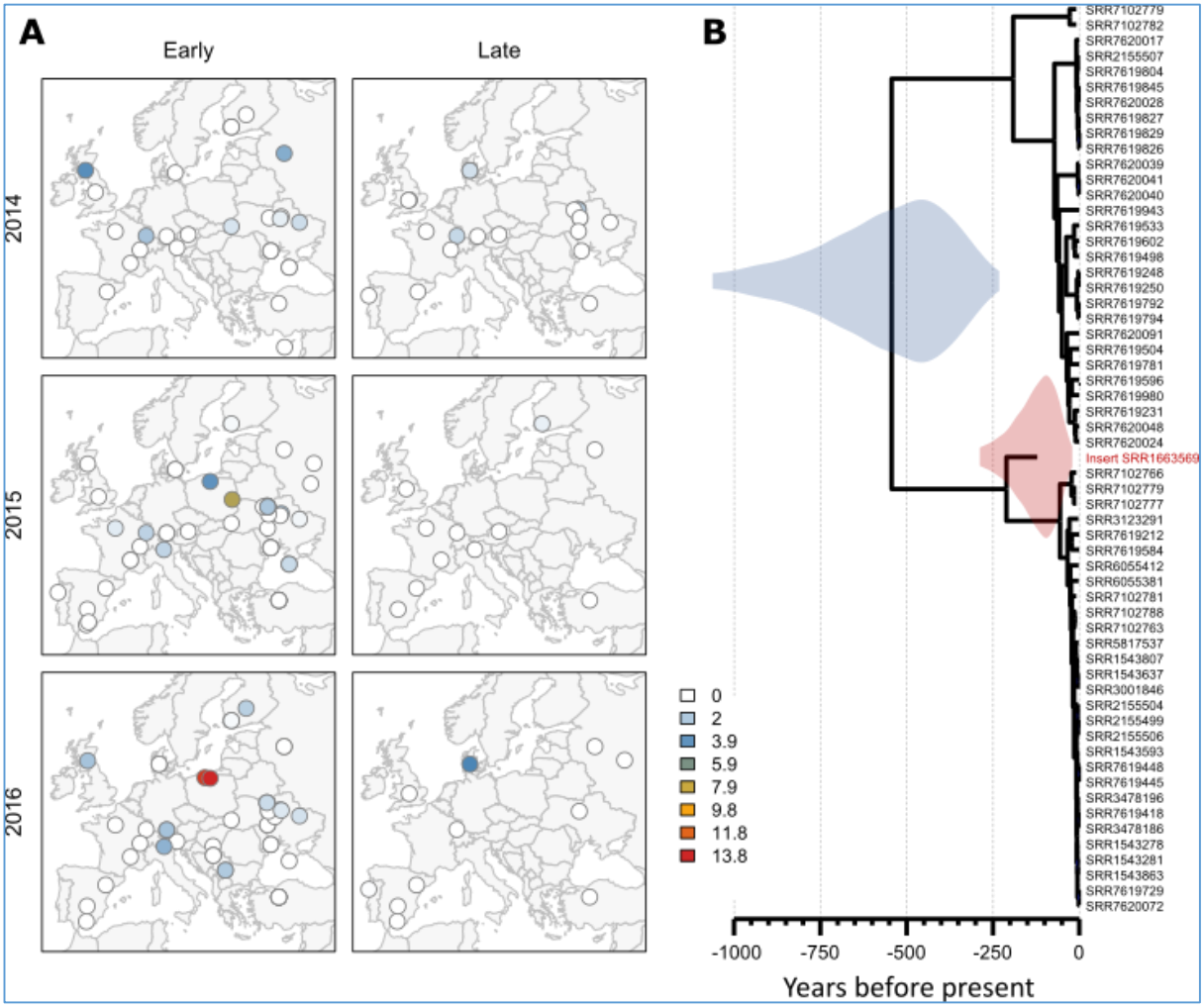


Figure 6: *Drosophila melanogaster* harbours an endogenous genomic copy of Galbut virus. (A) Maps show the spatial distribution of the DNA reads from the Galbut EVE, as a percentage of fly genomes (maximum 13.8%) on colour scale. Rows show years of sampling, and columns show 'early' or 'late' samples in each year (B) The relationship between the Galbut EVE and Galbut virus sequences detectable in public datasets, illustrated by a Bayesian maximum clade-credibility tree inferred under a strict clock, with median-scaled node dates. The 95% highest posterior density for the root date of extant Galbut viruses is shown in blue (230-1060 years before present), and the 95% highest posterior density for the inferred date of insertion, is shown in red (20-290 years before present)



Supplementary Material

Supplementary File S1: Excel spreadsheet detailing collection dates and locations

Supplementary File S2: Text document listing the microorganisms including in the 'Drosophila microbiome' mapping reference, and the mitochondrial and plastid sequences included in the species diagnostic mapping reference.

Supplementary File S3: Excel spreadsheet detailing the mapped read numbers from the DrosEU data. Sheet A gives raw mapped read counts, Sheet B gives counts normalised to read length in reads per kilobase per million reads (RPKM), Sheet C gives raw counts of reads mapping to additional Species-diagnostic loci.

Supplementary File S4: DNA Fasta file of virus fragments thought to be associated with contaminating taxa

Supplementary File S5: Excel spreadsheet detailing the presence and read counts of DNA viruses in DrosEU datasets. Sheet A gives counts normalised to the fly to give virus copy number in genomes fly genome and estimated prevalence at three different detection thresholds, Sheet B provides metadata used for the statistical analysis.

Supplementary File S6: DNA fasta file of assembled Vesanto virus segments, including divergent segments and segments assembled from public datasets.

Supplementary File S7: Excel spreadsheet detailing the presence and read counts of DNA viruses in 28 publicly available *Drosophila* sequencing projects. Sheet 1 summarises the public datasets included, Sheet 2 gives raw mapped read counts

Supplementary File S8: Excel spreadsheet detailing mean and total π_A, π_S and π_A/π_S for each gene (sheet A) and the number of synonymous and non-synonymous SNPs in the genome of Kallithea virus, Linvill Road virus and Vesanto virus (sheet B).

Supplementary File S9: Figure showing A) variation in nucleotide diversity across non-coding and synonymous sites in the Kallithea virus genome, plotted as a sliding window with two window sizes, and B) the percentage of Kallithea virus infected samples that showed evidence of an indel. Intergenic regions of the genome are coloured in grey. A chi-square test for independence found a strong positive association between intergenic regions and InDels (X-squared = 3236, df = 1, p-value < $2.2e^{-16}$).

Supplementary File S10: DNA fasta file of exemplar Galbut virus sequences aligned with the EVE and Gypsy-like LTR retroelement 297.

Bibliography

Antipov, D., A. Korobeynikov, J. S. McLean and P. A. Pevzner (2015). *hybridSPAdes: an algorithm for hybrid assembly of short and long reads*. Bioinformatics **32**(7): 1009-1015.

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham and A. D. Prjibelski (2012). *SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing*. Journal of computational biology **19**(5): 455-477.

Bao, W., K. K. Kojima and O. Kohany (2015). *Repbase Update, a database of repetitive elements in eukaryotic genomes*. Mobile DNA **6**(1): 11.

Batson, J., G. Dudas, E. Haas-Stapleton, A. L. Kistler, L. M. Li, P. Logan, K. Ratnasiri and H. Retallack (2020). *Single mosquito metatranscriptomics recovers mosquito species, blood meal sources, and microbial cargo, including viral dark matter*. bioRxiv: 2020.2002.2010.942854.

Bergman, C. and P. Haddrill (2015). *Strain-specific and pooled genome sequences for populations of Drosophila melanogaster from three continents. [version 1; peer review: 3 approved]*. F1000Research **4**(31).

Blangiardo, M., M. Cameletti, G. Baio and H. Rue (2013). *Spatial and spatio-temporal models with R-INLA*. Spatial and Spatio-temporal Epidemiology **7**: 39-55.

Bochdanovits, Z. and G. de Jong (2003). *Temperature dependent larval resource allocation shaping adult body size in Drosophila melanogaster*. Journal of Evolutionary Biology **16**(6): 1159-1167.

Bost, A., S. Franzenburg, K. L. Adair, V. G. Martinson, G. Loeb and A. E. Douglas (2018). *How gut transcriptional function of Drosophila melanogaster varies with the presence and composition of the gut microbiota*. Mol Ecol **27**(8): 1848-1859.

Bronkhorst, A. W., K. W. R. van Cleef, H. Venselaar and R. P. van Rij (2014). *A dsRNA-binding protein of a complex invertebrate DNA virus suppresses the Drosophila RNAi response*. Nucleic Acids Research **42**(19): 12237-12248.

Brun, G. and N. Plus (1980). *The viruses of Drosophila*. The genetics and biology of *Drosophila*. M. Ashburner and T. R. F. Wright. New York, Academic Press. **2:** 625-702.

Buchfink, B., C. Xie and D. H. Huson (2014). *Fast and sensitive protein alignment using DIAMOND*. Nature Methods **12**: 59.

Campo, D., K. Lehmann, C. Fjeldsted, T. Souaiaia, J. Kao and S. V. Nuzhdin (2013). *Whole-genome sequencing of two North American Drosophila melanogaster populations reveals genetic differentiation and positive selection*. Mol Ecol **22**(20): 5084-5097.

Carlson, J., E. Suchman and L. Buchatsky (2006). *Densoviruses for control and genetic manipulation of mosquitoes*. Insect Viruses: Biotechnological Applications. B. C. Bonning. **68:** 361-+.

Chang, C.-H. and A. M. Larracuente (2019). *Heterochromatin-Enriched Assemblies Reveal the Sequence and Organization of the Drosophila melanogaster Y Chromosome*. Genetics **211**(1): 333-348.

Cheng, R.-L., X.-F. Li and C.-X. Zhang (2020). *Nudivirus Remnants in the Genomes of Arthropods*. Genome Biology and Evolution **12**(5): 578-588.

Cory, J. S. and J. H. Myers (2003). *The ecology and evolution of insect baculoviruses*. Annual Review of Ecology Evolution and Systematics **34**: 239-272.

Cross, S. T., B. L. Maertens, T. J. Dunham, C. P. Rodgers, A. L. Brehm, M. R. Miller, A. M. Williams, B. D. Foy and M. D. Stenglein (2020). *Partitiviruses Infecting Drosophila melanogaster and Aedes aegypti Exhibit Efficient Biparental Vertical Transmission*. Journal of Virology **94**(20): e01070-01020.

Duffy, S. (2018). *Why are RNA virus mutation rates so damn high?* PLOS Biology **16**(8): e3000003.

Edgar, R. C. (2004). *MUSCLE: multiple sequence alignment with high accuracy and high throughput*. Nucleic Acids Research **32**(5): 1792-1797.

Elya, C., T. C. Lok, Q. E. Spencer, H. McCausland, C. C. Martinez and M. Eisen (2018). *Robust manipulation of the behavior of Drosophila melanogaster by a fungal pathogen in the laboratory*. eLife **7**: e34414.

Endler, L., J.-M. Gibert, V. Nolte and C. Schlötterer (2018). *Pleiotropic effects of regulatory variation in tan result in correlation of two pigmentation traits in Drosophila melanogaster*. Molecular Ecology **27**(16): 3207-3218.

Etebari, K., I. Filipović, G. Rašić, G. J. Devine, H. Tsatsia and M. J. Furlong (2020). *Complete genome sequence of Oryctes rhinoceros nudivirus isolated from the coconut rhinoceros beetle in Solomon Islands*. Virus Research **278**: 197864.

Everett, L. J., W. Huang, S. Zhou, M. A. Carbone, R. F. Lyman, G. H. Arya, M. S. Geisz, J. Ma, F. Morgante, G. St Armour, L. Turlapati, R. R. H. Anholt and T. F. C. Mackay (2020). *Gene expression networks in the Drosophila Genetic Reference Panel*. Genome Res **30**(3): 485-496.

Fang, N. L., J. Suresh, K. Chaithanya, X. Linfan, T. Z. Tan, J. Gruber and N. S. Tolwinski (2017). *TGH: Trans-Generational Hormesis and the Inheritance of Aging Resistance*. bioRxiv: 127951.

Garlapow, M. E., L. J. Everett, S. Zhou, A. W. Gearhart, K. A. Fay, W. Huang, T. V. Morozova, G. H. Arya, L. Turlapati, G. St Armour, Y. N. Hussain, S. E. McAdams, S. Fochler and T. F. Mackay (2017). *Genetic and Genomic Response to Selection for Food Consumption in Drosophila melanogaster*. Behav Genet **47**(2): 227-243.

Gilbert, C. and R. Cordaux (2017). *Viruses as vectors of horizontal transfer of genetic material in eukaryotes*. Current opinion in virology **25**: 16-22.

Gilbert, C., J. Peccoud, A. Chateigner, B. Moumen, R. Cordaux and E. A. Herniou (2016). *Continuous Influx of Genetic Material from Host to Virus Populations*. PLOS Genetics **12**(2): e1005838.

Gilks, W., T. Pennell, I. Flis, M. Webster and E. Morrow (2016). *Whole genome resequencing of a laboratory-adapted Drosophila melanogaster population sample [version 1; peer review: 2 approved]*. F1000Research **5**(2644).

Grenier, J. K., J. R. Arguello, M. C. Moreira, S. Gottipati, J. Mohammed, S. R. Hackett, R. Boughton, A. J. Greenberg and A. G. Clark (2015). *Global diversity lines - a five-continent reference panel of sequenced Drosophila melanogaster strains*. G3 (Bethesda) **5**(4): 593-603.

Harrison, R. L., E. A. Herniou, A. Bezier, J. A. Jehle, J. P. Burand, D. A. Theilmanns, P. J. Krell, M. M. van Ere, M. Nakai, S. G. Siddell, A. J. Davison, E. J. Lefkowitz, S. Sabanadzovic, P. Simmonds, D. B. Smith, R. J. Orton, B. Harrach and I. R. Consortium (2020). *ICTV Virus Taxonomy Profile: Nudiviridae*. Journal of General Virology **101**(1): 3-4.

Hill, T. and R. L. Unckless (2018). *The dynamic evolution of Drosophila innubila Nudivirus*. Infection Genetics and Evolution **57**: 151-157.

Huszart, T. and J. L. Imler (2008). *Drosophila Viruses and the Study of Antiviral Host-Defense*. Advances in Virus Research, Vol 72. K. Maramorosch, A. Shatkin and F. Murphy. **72:** 227-265.

Jalvingh, K. M., P. L. Chang, S. V. Nuzhdin and B. Wertheim (2014). *Genomic changes under rapid evolution: selection for parasitoid resistance*. Proceedings of the Royal Society B: Biological Sciences **281**(1779): 20132303.

Jousset, F.-X., E. Baquerizo and M. Bergoin (2000). *A new densovirus isolated from the mosquito Culex pipiens (Diptera: Culicidae)*. Virus Research **67**(1): 11-16.

Kang, L., E. Rashkovetsky, K. Michalak, H. R. Garner, J. E. Mahaney, B. A. Rzigalinski, A. Korol, E. Nevo and P. Michalak (2019). *Genomic divergence and adaptive convergence in Drosophila simulans from Evolution Canyon, Israel*. Proc Natl Acad Sci U S A **116**(24): 11839-11844.

Kao, J. Y., A. Zubair, M. P. Salomon, S. V. Nuzhdin and D. Campo (2015). *Population genomic analysis uncovers African and European admixture in Drosophila melanogaster populations from the south-eastern United States and Caribbean Islands*. Molecular Ecology **24**(7): 1499-1509.

Kapun, M., M. G. Barrón, F. Staubach, D. J. Obbard, R. A. W. Wiberg, J. Vieira, C. Goubert, O. Rota-Stabelli, M. Kankare, M. Bogaerts-Márquez, A. Haudry, L. Waidele, I. Kozeretska, E. G. Pasyukova, V. Loeschcke, M. Pascual, C. P. Vieira, S. Serga, C. Montchamp-Moreau, J. Abbott, P. Gibert, D. Porcelli, N. Posnien, A. Sánchez-Gracia, S. Grath, É. Sucena, A. O. Bergland, M. P. G. Guerreiro, B. S. Onder, E. Argyridou, L. Guio, M. F. Schou, B. Deplancke, C. Vieira, M. G. Ritchie, B. J. Zwaan, E. Tauber, D. J. Orengo, E. Puerma, M. Aguadé, P. S.

Schmidt, J. Parsch, A. J. Betancourt, T. Flatt and J. González (2020). *Genomic analysis of European Drosophila melanogaster populations reveals longitudinal structure, continent-wide selection, and previously unknown DNA viruses*. Molecular Biology and Evolution.

Kariithi, H. M., I. K. Meki, D. G. Boucias and A. M. M. Abd-Alla (2017). *Hytrosaviruses: current status and perspective*. Current Opinion in Insect Science **22**: 71-78.

Kariithi, H. M., M. M. van Oers, J. M. Vlak, M. J. B. Vreysen, A. G. Parker and A. M. M. Abd-Alla (2013). *Virology, Epidemiology and Pathology of Glossina Hytrosavirus, and Its Control Prospects in Laboratory Colonies of the Tsetse Fly, Glossina pallidipes (Diptera; Glossinidae)*. Insects **4**(3): 287-319.

Katzourakis, A. and R. J. Gifford (2010). *Endogenous Viral Elements in Animal Genomes*. PLOS Genetics **6**(11): e1001191.

Kawato, S., A. Shitara, Y. Wang, R. Nozaki, H. Kondo and I. Hirono (2019). *Crustacean Genome Exploration Reveals the Evolutionary Origin of White Spot Syndrome Virus*. J Virol **93**(3).

Kelly, J. K. and K. A. Hughes (2019). *Pervasive Linked Selection and Intermediate-Frequency Alleles Are Implicated in an Evolve-and-Resequencing Experiment of Drosophila simulans*. Genetics **211**(3): 943-961.

Kofler, R., P. Orozco-terWengel, N. De Maio, R. V. Pandey, V. Nolte, A. Futschik, C. Kosiol and C. Schlötterer (2011). *PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals*. PLOS ONE **6**(1): e15925.

Kofler, R., R. V. Pandey and C. Schlötterer (2011). *PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq)*. Bioinformatics **27**(24): 3435-3436.

Kosakovsky Pond, S. L., D. Posada, M. B. Gravenor, C. H. Woelk and S. D. W. Frost (2006). *GARD: a genetic algorithm for recombination detection*. Bioinformatics **22**(24): 3096-3098.

Krueger, F. (2015). *Trim galore*. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files **516**: 517.

Krupovic, M. and E. V. Koonin (2014). *Evolution of eukaryotic single-stranded DNA viruses of the Bidnaviridae family from genes of four other groups of widely different viruses*. Sci Rep **4**: 5347.

Lack, J. B., C. M. Cardeno, M. W. Crepeau, W. Taylor, R. B. Corbett-Detig, K. A. Stevens, C. H. Langley and J. E. Pool (2015). *The Drosophila genome nexus: a population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population*. Genetics **199**(4): 1229-1241.

Lange, J. D., J. B. Lack, J. E. Pool, A. D. Tang and R. B. Corbett-Detig (2016). *A Thousand Fly Genomes: An Expanded Drosophila Genome Nexus*. Molecular Biology and Evolution **33**(12): 3308-3313.

Langmead, B. and S. L. Salzberg (2012). *Fast gapped-read alignment with Bowtie 2*. Nature methods **9**(4): 357.

Lawrence, P. O. (2011). *Gammaentomopoxvirus*. The Springer Index of Viruses. C. Tidona and G. Darai. New York, NY, Springer New York**:** 1533-1539.

Lepetit, D., B. Gillet, S. Hughes, K. Kraaijeveld and J. Varaldi (2016). *Genome Sequencing of the Behavior Manipulating Virus LbFV Reveals a Possible New Virus Family*. Genome Biol Evol **8**(12): 3718-3739.

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. arXiv **1303.3997**.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and S. Genome Project Data Processing (2009). *The Sequence Alignment/Map format and SAMtools*. Bioinformatics **25**(16): 2078-2079.

Li, R., P. Chang, P. Lü, Z. Hu, K. Chen, Q. Yao and Q. Yu (2019). *Characterization of the RNA Transcription Profile of Bombyx mori Bidensovirus*. Viruses **11**(4): 325.

Lin, Y., K. Golovnina, Z. X. Chen, H. N. Lee, Y. L. Negron, H. Sultana, B. Oliver and S. T. Harbison (2016). *Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster*. BMC Genomics **17**: 28.

Liu, X. and J. Secombe (2015). *The Histone Demethylase KDM5 Activates Gene Expression by Recognizing Chromatin Context through Its PHD Reader Motif*. Cell reports **13**(10): 2219-2231.

Loiseau, V., E. A. Herniou, Y. Moreau, N. Lévêque, C. Meignin, L. Daeffler, B. Federici, R. Cordaux and C. Gilbert (2020). *Wide spectrum and high frequency of genomic structural variation, including transposable elements, in large double-stranded DNA viruses*. Virus Evolution **6**(1).

Longdon, B., G. G. R. Murray, W. J. Palmer, J. P. Day, D. J. Parker, J. J. Welch, D. J. Obbard and F. M. Jiggins (2015). *The evolution, diversity, and host associations of rhabdoviruses*. Virus Evolution **1**(1): 12.

Machado, H. E., A. O. Bergland, R. Taylor, S. Tilk, E. Behrman, K. Dyer, D. K. Fabian, T. Flatt, J. González, T. L. Karasov, I. Kozeretska, B. P. Lazzaro, T. J. Merritt, J. E. Pool, K. O'Brien, S. Rajpurohit, P. R. Roy, S. W. Schaeffer, S. Serga, P. Schmidt and D. A. Petrov (2019). *Broad geographic sampling reveals predictable, pervasive, and strong seasonal adaptation in Drosophila*. bioRxiv: 337543.

Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, D. Zhu, S. Casillas, Y. Han, M. M. Magwire, J. M. Cridland, M. F. Richardson, R. R. H. Anholt, M. Barrón, C. Bess, K. P. Blankenburg, M. A. Carbone, D. Castellano, L. Chaboub, L. Duncan, Z. Harris, M. Javaid, J. C. Jayaseelan, S. N. Jhangiani, K. W. Jordan, F. Lara, F. Lawrence, S. L. Lee, P. Librado, R. S. Linheiro, R. F. Lyman, A. J. Mackey, M. Munidasa, D. M. Muzny, L. Nazareth, I. Newsham, L. Perales, L.-L. Pu, C. Qu, M. Ràmia, J. G. Reid, S. M. Rollmann, J. Rozas, N. Saada, L. Turlapati, K. C. Worley, Y.-Q. Wu, A. Yamamoto, Y. Zhu, C. M. Bergman, K. R. Thornton, D. Mittelman and R. A. Gibbs (2012). *The Drosophila melanogaster Genetic Reference Panel*. Nature **482**(7384): 173-178.

Martin, D. P., B. Murrell, M. Golden, A. Khoosal and B. Muhire (2015). *RDP4: Detection and analysis of recombination patterns in virus genomes*. Virus Evolution **1**(1).

Martin, M. (2011). *Cutadapt removes adapter sequences from high-throughput sequencing reads*. EMBnet. journal **17**(1): 10-12.

Mateo, L., G. E. Rech and J. González (2018). *Genome-wide patterns of local adaptation in Western European Drosophila melanogaster natural populations*. Scientific Reports **8**(1): 16143.

Medd, N. C., S. Fellous, F. M. Waldron, A. Xuéreb, M. Nakai, J. V. Cross and D. J. Obbard (2018). *The virome of Drosophila suzukii, an invasive pest of soft fruit*. Virus Evolution **4**(1): vey009-vey009.

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler and R. Lanfear (2020). *IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era*. Molecular Biology and Evolution **37**(5): 1530-1534.

Mondotte, J. A., V. Gausson, L. Frangeul, H. Blanc, L. Lambrechts and M. C. Saleh (2018). *Immune priming and clearance of orally acquired RNA viruses in Drosophila*. Nat Microbiol **3**(12): 1394-1403.

Mussabekova, A., L. Daeffler and J. L. Imler (2017). *Innate and intrinsic antiviral immunity in Drosophila*. Cellular and Molecular Life Sciences **74**(11): 2039-2054.

Niehus, S., P. Giammarinaro, S. Liégeois, J. Quintin and D. Ferrandon (2012). *Fly culture collapse disorder: detection, prophylaxis and eradication of the microsporidian parasite Tubulinosema ratisbonensis infecting Drosophila melanogaster*. Fly **6**(3): 193-204.

Notredame, C., D. G. Higgins and J. Heringa (2000). *T-coffee: a novel method for fast and accurate multiple sequence alignment11Edited by J. Thornton*. Journal of Molecular Biology **302**(1): 205-217.

Nouhaud, P. (2018). *Long-read based assembly and annotation of a Drosophila simulans genome*. bioRxiv: 425710.

Nurk, S., A. Bankevich, D. Antipov, A. Gurevich, A. Korobeynikov, A. Lapidus, A. Prjibelsky, A. Pyshkin, A. Sirotkin, Y. Sirotkin, R. Stepanauskas, J. McLean, R. Lasken, S. R. Clingenpeel, T. Woyke, G. Tesler, M. A. Alekseyev and P. A. Pevzner (2013). *Assembling Genomes and Mini-metagenomes from Highly Chimeric Reads*, Berlin, Heidelberg, Springer Berlin Heidelberg.

Obbard, D. J. (2018). *Expansion of the Metazoan Virosphere: Progress, Pitfalls, and Prospects.* . Current Opinion in Virology **31**: 17-23.

Obbard, D. J., M. Shi, K. E. Roberts, B. Longdon and A. B. Dennis (2020). *A new lineage of segmented RNA viruses infecting animals*. Virus Evolution **6**(1).

Palatini, U., P. Miesen, R. Carballar-Lejarazu, L. Ometto, E. Rizzo, Z. Tu, R. P. Rij and M. Bonizzoni (2017). *Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors Aedes aegypti and Aedes albopictus*. BMC genomics **18**(1): 512.

Palmer, W. H., J. Joosten, G. J. Overheul, P. W. Jansen, M. Vermeulen, D. J. Obbard and R. P. Van Rij (2019). *Induction and Suppression of NF-kappa B Signalling by a DNA Virus of Drosophila*. Journal of Virology **93**(3).

Palmer, W. H., N. C. Medd, P. M. Beard and D. J. Obbard (2018). *Isolation of a natural DNA virus of Drosophila melanogaster, and characterisation of host resistance and immune responses*. PLOS Pathogens **14**(6): e1007050.

Peccoud, J., S. Lequime, I. Moltini-Conclois, I. Giraud, L. Lambrechts and C. Gilbert (2018). *A Survey of Virus Recombination Uncovers Canonical Features of Artificial Chimeras Generated During Deep Sequencing Library Preparation*. G3 (Bethesda, Md.) **8**(4): 1129-1138.

Peck, K. M. and A. S. Lauring (2018). *Complexities of Viral Mutation Rates*. Journal of Virology **92**(14): e01031-01017.

Pénzes, J. J., M. Söderlund-Venermo, M. Canuti, A. M. Eis-Hübinger, J. Hughes, S. F. Cotmore and B. Harrach (2020). *Reorganizing the family Parvoviridae: a revised taxonomy independent of the canonical approach based on host association*. Archives of Virology.

Perera, S., Z. Li, L. Pavlik and B. Arif (2010). *Entomopoxviruses*. Insect Virology. S. Asgari and K. N. Johnson. Norfolk, UK, Caister Academic Press**:** 83–102.

Piegu, B., S. Guizard, Y. P. Tan, C. Cruaud, S. Asgari, D. K. Bideshi, B. A. Federici and Y. Bigot (2014). *Genome sequence of a crustacean iridovirus, IIV31, isolated from the pill bug, Armadillidium vulgare*. Journal of General Virology **95**: 1585-1590.

Plus, N. (1978). *Endonenous viruses of Drosophila melanogaster cell lines - their frequency, identification, and origin*. In Vitro-Journal of the Tissue Culture Association **14**(12): 1015-1021.

Poirier, E. Z., B. Goic, L. Tomé-Poderti, L. Frangeul, J. Boussier, V. Gausson, H. Blanc, T. Vallet, H. Loyd, L. I. Levi, S. Lanciano, C. Baron, S. H. Merkling, L. Lambrechts, M. Mirouze, S. Carpenter, M. Vignuzzi and M. C. Saleh (2018). *Dicer-2-Dependent Generation of Viral DNA from Defective Genomes of RNA Viruses Modulates Antiviral Immunity in Insects*. Cell Host Microbe **23**(3): 353-365.e358.

Prompiboon, P., V.-U. Lietze, J. S. S. Denton, C. J. Geden, T. Steenberg and D. G. Boucias (2010). *Musca domestica salivary gland hypertrophy virus, a globally distributed insect virus that infects and sterilizes female houseflies*. Applied and environmental microbiology **76**(4): 994-998.

Reddiex, A. J., S. L. Allen and S. F. Chenoweth (2018). *A Genomic Reference Panel for Drosophila serrata*. G3: Genes|Genomes|Genetics **8**(4): 1335-1346.

Ribeiro, J. M., H. J. Debat, M. Boiani, X. Ures, S. Rocha and M. Breijo (2019). *An insight into the sialome, mialome and virome of the horn fly, Haematobia irritans*. BMC Genomics **20**(1): 616.

Riddiford, N., K. Siudeja, M. van den Beek, B. Boumard and A. J. Bardin (2020). *Evolution and genomic signatures of spontaneous somatic mutation in Drosophila intestinal stem cells*. bioRxiv: 2020.2007.2020.188979.

Schou, M. F., V. Loeschcke, J. Bechsgaard, C. Schlötterer and T. N. Kristensen (2017). *Unexpected high genetic diversity in small populations suggests maintenance by associative overdominance*. Molecular Ecology **26**(23): 6510-6523.

Shapiro, B., A. Rambaut and A. J. Drummond (2006). *Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences*. Mol Biol Evol **23**(1): 7-9.

Shi, M., V. L. White, T. Schlub, J.-S. Eden, A. A. Hoffmann and E. C. Holmes (2018). *No detectable effect of Wolbachia wMel on the prevalence and abundance of the RNA virome of Drosophila melanogaster*. Proceedings of the Royal Society B-Biological Sciences **In Press**.

Shi, M., Y. Z. Zhang and E. C. Holmes (2018). *Meta-transcriptomics and the evolutionary biology of RNA viruses*. Virus Research **243**: 83-90.

Signor, S. A., F. N. New and S. Nuzhdin (2017). *A Large Panel of Drosophila simulans Reveals an Abundance of Common Variants*. Genome Biology and Evolution **10**(1): 189-206.

Siudeja, K., S. Nassari, L. Gervais, P. Skorski, S. Lameiras, D. Stolfa, M. Zande, V. Bernard, T. R. Frio and A. J. Bardin (2015). *Frequent Somatic Mutation in Adult Intestinal Stem Cells Drives Neoplasia and Genetic Mosaicism during Aging*. Cell Stem Cell **17**(6): 663-674.

Sloan, M. A., K. Brooks, T. D. Otto, M. J. Sanders, J. A. Cotton and P. Ligoxygakis (2019). *Transcriptional and genomic parallels between the monoxenous parasite Herpetomonas muscarum and Leishmania*. PLOS Genetics **15**(11): e1008452.

Speybroeck, N., C. J. Williams, K. B. Lafia, B. Devleesschauwer and D. Berkvens (2012). *Estimating the prevalence of infections in vector populations using pools of samples*. Med Vet Entomol **26**(4): 361-371.

Sprengelmeyer, Q. D., S. Mansourian, J. D. Lange, D. R. Matute, B. S. Cooper, E. V. Jirle, M. C. Stensmyr and J. E. Pool (2019). *Recurrent Collection of Drosophila melanogaster from Wild African Environments and Genomic Insights into Species History*. Molecular Biology and Evolution **37**(3): 627-638.

Starrett, G. J., M. J. Tisza, N. L. Welch, A. K. Belford, A. Peretti, D. V. Pastrana and C. B. Buck (2020). *Adintoviruses: A Proposed Animal-Tropic Family of Midsize Eukaryotic Linear dsDNA (MELD) Viruses*. Virus Evolution.

Suchard, M. A., P. Lemey, G. Baele, D. L. Ayres, A. J. Drummond and A. Rambaut (2018). *Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10*. Virus Evolution **4**(1).

Tan, G., M. Muffato, C. Ledergerber, J. Herrero, N. Goldman, M. Gil and C. Dessimoz (2015). *Current Methods for Automated Filtering of Multiple Sequence Alignments Frequently Worsen Single-Gene Phylogenetic Inference*. Systematic Biology **64**(5): 778-791.

Tassetto, M., M. Kunitomi, Z. J. Whitfield, P. T. Dolan, I. Sánchez-Vargas, M. Garcia-Knight, I. Ribiero, T. Chen, K. E. Olson and R. Andino (2019). *Control of RNA viruses in mosquito cells through the acquisition of vDNA and endogenous viral elements*. eLife **8**: e41244.

Thézé, J., J. Takatsuka, Z. Li, J. Gallais, D. Doucet, B. Arif, M. Nakai and E. A. Herniou (2013). *New insights into the evolution of Entomopoxvirinae from the complete genome sequences of four entomopoxviruses infecting Adoxophyes honmai, Choristoneura biennis, Choristoneura rosaceana, and Mythimna separata*. Journal of virology **87**(14): 7992-8003.

Tristan, B., P. Aurelie and H. Annabelle (2019). *Evidence for purifying selection on conserved noncoding elements in the genome of Drosophila melanogaster*. bioRxiv: 623744.

Unckless, R. L. (2011). *A DNA Virus of Drosophila*. Plos One **6**(10).

Upton, C., S. Slack, A. L. Hunter, A. Ehlers and R. L. Roper (2003). *Poxvirus Orthologous Clusters: toward Defining the Minimum Essential Poxvirus Genome*. Journal of Virology **77**(13): 7590-7600.

Van der Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel and M. A. DePristo (2013). *From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline*. Current Protocols in Bioinformatics **43**(1): 11.10.11-11.10.33.

Vicoso, B. and D. Bachtrog (2013). *Reversal of an ancient sex chromosome to an autosome in Drosophila*. Nature **499**(7458): 332-335.

Wang, Y., M. Kapun, L. Waidele, S. Kuenzel, A. O. Bergland and F. Staubach (2020). *Common structuring principles of the Drosophila melanogaster microbiome on a continental scale and between host and substrate*. Environmental Microbiology Reports **12**(2): 220-228.

Webster, C. L., B. Longdon, S. H. Lewis and D. J. Obbard (2016). *Twenty-Five New Viruses Associated with the Drosophilidae (Diptera)*. Evolutionary Bioinformatics **12**(12): 13-25.

Webster, C. L., F. M. Waldron, S. Robertson, D. Crowson, G. Ferrari, J. F. Quintana, J. M. Brouqui, E. H. Bayne, B. Longdon, A. H. Buck, B. P. Lazzaro, J. Akorli, P. R. Haddrill and D. J. Obbard (2015). *The Discovery, Distribution, and Evolution of Viruses Associated with Drosophila melanogaster*. Plos Biology **13**(7): 33.

West, C. and N. Silverman (2018). *p38b and JAK-STAT signaling protect against Invertebrate iridescent virus 6 infection in Drosophila*. Plos Pathogens **14**(5).

Wick, R. R., M. B. Schultz, J. Zobel and K. E. Holt (2015). *Bandage: interactive visualization of de novo genome assemblies*. Bioinformatics **31**(20): 3350-3352.

Wu, Q., Y. Luo, R. Lu, N. Lau, E. C. Lai, W.-X. Li and S.-W. Ding (2010). *Virus discovery by deep sequencing and assembly of virus-derived small silencing RNAs*. Proceedings of the National Academy of Sciences **107**(4): 1606.

Yablonovitch, A. L., J. Fu, K. Li, S. Mahato, L. Kang, E. Rashkovetsky, A. B. Korol, H. Tang, P. Michalak, A. C. Zelhof, E. Nevo and J. B. Li (2017). *Regulation of gene expression and RNA editing in Drosophila adapting to divergent microclimates*. Nat Commun **8**(1): 1570.

Yutin, N., S. Shevchenko, V. Kapitonov, M. Krupovic and E. V. Koonin (2015). *A novel group of diverse Polinton-like viruses discovered by metagenome analysis*. BMC Biol **13**: 95.

Zhang, Y. Z., M. Shi and E. C. Holmes (2018). *Using Metagenomics to Characterize an Expanding Virosphere*. Cell **172**(6): 1168-1172.

Zhao, X., Y. Tian, R. Yang, H. Feng, Q. Ouyang, Y. Tian, Z. Tan, M. Li, Y. Niu, J. Jiang, G. Shen and R. Yu (2012). *Coevolution between simple sequence repeats (SSRs) and virus genome size*. BMC Genomics **13**(1): 435.