Flies on the move: an inherited virus mirrors *Drosophila melanogaster*’s elusive ecology and demography

Lena Wilfert¹

Francis M. Jiggins²

¹) Centre for Ecology and Conservation, University of Exeter, Penryn Campus, Penryn, TR10 9FE, United Kingdom

²) Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, United Kingdom

keywords: migration, overwintering, biogeography, co-evolution

Corresponding Author: Lena Wilfert, Tel. +44 (0) 1326370723, Fax ++44 (0) 1326253638: lena.wilfert@ex.ac.uk

Running title: Flies on the move
Vertically transmitted parasites rely on their host’s reproduction for their transmission, leading to the evolutionary histories of both parties being intimately entwined. Parasites can thus serve as a population genetic magnifying glass for their host’s demographic history. Here, we study the fruitfly *Drosophila melanogaster*’s vertically transmitted sigma virus DMelSV. The virus has a high mutation rate and low effective population size, allowing us to reconstruct at a fine scale how the combined forces of the movement of flies and selection on the virus have shaped its migration patterns. We found that the virus is likely to have spread to Europe from Africa, mirroring the colonization route of *Drosophila*. The North American DMelSV population appears to be the result of a recent single immigration from Europe, invading together with its host in the late 19th century. Across Europe, DMelSV migration rates are low and populations are highly genetically structured, likely reflecting limited fly movement. Despite being intolerant of extreme cold, viral diversity suggests that fly populations can persist in harsh continental climates and that recolonisation from the warmer south plays a minor role. In conclusion, studying DMelSV can provide insights into the poorly understood ecology of *D. melanogaster*, one of the best-studied organisms in biology.
Introduction

The genetic structure of parasite populations provides a record of their past transmission routes. This is shaped by the migration and contact patterns of their hosts, together with selection on the parasite genome and other epidemiological processes. Therefore, with care, parasites that are closely associated with their host can be used as proxies for understanding the phylogeographic or demographic history of their hosts (Wirth et al. 2005; Nieberding & Olivieri 2007). For example, Biek et al. (2006) used feline immune deficiency virus as a marker to infer the recent demographic history of its cougar host. Human history too has been uncovered by parasites – for example, the divergence time of head and body lice has shown that the use of clothing evolved only in the last 100,000 years (72’000 ± 42’000 years, (Kittler et al. 2003) whereas the bacterium Helicobacter pylori reveals details of recent human migrations, such as during the slave trade or the settling of Polynesia (Falush et al. 2003).

Key population genetic parameters shaping the genetic structure of parasite populations can differ dramatically from the host, meaning that the spatial and temporal scale at which demographic processes can be detected in the host and parasite populations can be very different (Nieberding & Olivieri 2007). For example, RNA virus genomes evolve faster than their hosts, allowing the reconstruction of very recent events during their evolution (Nieberding & Olivieri 2007). Some parasites also have smaller effective population sizes ($N_e$), which has the effect of increasing genetic structure within populations (Nieberding & Olivieri 2007). If parasites are to be used to elucidate their host’s evolutionary history, they are most useful when they have transmission routes that link them closely to the movement or reproduction of their hosts, such as vertical or sexual transmission (Nieberding & Olivieri 2007).
The sigma virus DMelSV is a naturally occurring and globally distributed parasite that is specific to *Drosophila melanogaster* (Carpenter *et al.* 2007). This virus is exclusively vertically transmitted to the offspring by both the mother and the father, so its demographic history is intimately linked with – and limited by – that of its host. As a vertically transmitted parasite where there is little genetic variation within hosts and multiple infections are rare (F. Jiggins unpublished data; (Carpenter *et al.* 2007), its effective population size can be estimated in terms of numbers of infected hosts, in a similar way to mitochondrial genes. At its simplest the virus can be treated as a haploid host gene, except that its $N_e$ is likely to be lower because only a minority of flies are infected (prevalences of 0-15% (Carpenter *et al.* 2007) and 7.2 -29.7 % (Fleuriet 1976) have been reported from natural populations).

Furthermore, as a typical RNA virus, DMelSV’s mutation rate is several orders of magnitude higher than that of *Drosophila* (Carpenter *et al.* 2007; Keightley *et al.* 2009). As is common for negative-sense RNA viruses, there is no evidence for genetic recombination between sigma viruses in nature (Carpenter *et al.* 2007), so phylogenetic approaches can be used to investigate its evolution. However, as a pathogen it can be subject to strong selection and has variable prevalence, which will decouple patterns of genetic variation in the host and virus (Fleuriet & Sperlich 1992; Fleuriet & Periquet 1993; Carpenter *et al.* 2007; Wilfert & Jiggins 2013).

Few organisms have been as well studied as the fruitfly *Drosophila melanogaster*. From the earliest days of the field of genetics, *D. melanogaster* has allowed researchers to gain insights into fields ranging from development to aging, physiology to behaviour, and immunology to sexual selection to name but a few. However, to this day, we know surprisingly little about this species’ ecology and natural history in the wild (Keller 2007). Today, *D. melanogaster* is a human commensal with a cosmopolitan distribution. Based on genetic variation at all levels
of the genome, it is clear that its origins lie in sub-Saharan Africa (reviewed in David and Capy (1988) and Stephan & Li (2007)). It has been suggested that its dependency on humans dates back as far as 18,000 years (Lachaise & Silvain 2004), predating its move into Europe sometime after the last glaciation which ended about 16,000 years ago. It then spread across Eurasia, but was only able to reach the Americas, Australia and remote islands with the help of man over the last few centuries, with North America in particular being colonised only in the late 19th century by cold-adapted flies from Europe (David & Capy 1988; Keller 2007; Stephan & Li 2007).

The contemporary migration patterns of *D. melanogaster* are much less understood. The constant global movement of fruitflies with the food trade and the rapid spread of P-elements in the middle of the last century suggest constant gene flow (David & Capy 1988), yet there also is ample evidence of stable latitudinal and longitudinal genetic variation (Rako et al. 2009; Fabian et al. 2012). Appreciable levels of migration are known to occur between *D. melanogaster* populations. While direct observations of fruitflies suggest that movement may be limited to just a few hundreds of meters (reviewed by Dobzhansky 1973), Coyne & Milstead (1987) demonstrated that released alleles of *D. melanogaster* spread ~10 km in 3 months through a continuously habitable and pre-colonised area, potentially by a combination of active dispersal and passive wind transport. High rates of dispersal also occur in the closely related and ecologically similar species *D. simulans*. Here, the vertically transmitted bacterial symbiont *Wolbachia* spread at a rate of more than 100 km per year across California, which allowed Turelli & Hoffmann (1991) to estimate that migration rates in nature are far greater than typically observed in behavioural experiments. One particular aspect of migration that is not fully resolved is the extent to which populations in temperate and cold regions are resident, overwintering in local refugia such as human households, or whether these
populations are partly reseeded annually by immigrants from warmer regions (Keller 2007), generating biased migration from south to north in the Northern hemisphere.

Here, we study the population genetics of DMelSV and demonstrate a link between the demographic histories of this obligate parasite and its host. On the one hand, this close association may be shaped by coevolution between host and parasite. At the same time, the rapidly evolving virus can give insights into its host’s ecology that are hard to discern by studying the fly’s genetics.

Materials and Methods

Virus isolates

As DMelSV is vertically transmitted, it can be isolated by collecting infected flies and establishing infected fly lines in the laboratory. After allowing flies isolated from the field to reproduce in the lab, we diagnosed DMelSV infections through exposure to CO$_2$, a treatment from which uninfected flies fully recover but which leaves infected flies paralysed (Brun & Plus 1980; Wilfert & Jiggins 2010). Details of virus isolate collections can be found in table A1 (supplementary material). Wild isolates of DMelSV were collected from 15 populations of D. melanogaster across Europe (Germany – Southwest and Southeast, Greece – Athens, Spain – Galicia, Sweden – Uppsala, UK – Cambridgeshire, Derby, Essex and Kent) the United States (USA – California, Florida, Georgia and North Carolina) and Africa (Ghana – Accra and Kenya – Thika (n = 2)) between 2005 and 2010 (table A1) (Carpenter et al. 2007; Wilfert & Jiggins 2010), including 8 samples for which we retrieved sequences from
Genbank (3 samples from Athens, Greece: AM689323, AM689330, AM689331; 4 from Essex, UK: AM689310, AM689314, AM689317, AM689318; and 1 from Georgia, USA: AM689320). The populations in Essex, UK and Athens, Greece, were repeatedly sampled across these years. We also included Hap23, a well-studied lab isolate of DMelSV. Infected fly lines collected up to the end of 2007 were maintained in the laboratory for 0.5 to 2.5 years in isofemale lines (with the exception of the 8 lines for which we retrieved sequences from Genbank, see above). During this time, we substituted the 2nd and 3rd wild-type chromosomes with the isogenic line $w^{1118}$ (described by Parks et al. (2004)) using the double balancer line $SM2/Pm; TM3/Sb; spa^{Pol}$ (Wilfert & Jiggins 2010) to allow their use in other experiments. We also obtained eight samples from California (collected by Darren Obbard in 2008), two samples from Kenya (collected by John Pool in 2009), one sample from Ghana (collected by Claire Webster in 2010) and 56 samples from Cambridgeshire and Essex, UK (collected by Heather Cagney in 2009), which were maintained in the lab for 1 to 3 months. These samples were sequenced from the wild-type host.

Sequencing viral isolates

For each virus-infected fly line, we extracted RNA from 10 flies using 300 µl Trizol (Invitrogen, Carlsbad, Ca.) following the manufacturer’s instruction. The RNA was resuspended in 50 µl RNA storage solution (Ambion, Austin, Tx.) and reverse transcribed into cDNA using random hexamers and M-MLV reverse transcriptase (Invitrogen, Carlsbad, Ca.). The viral isolates were sequenced using direct sequencing of PCR products using Sanger sequencing as described by Carpenter et al. (2007) (see Table A2 for primer sequences). The sequences were assembled into contigs and aligned using Sequencher 4.2 (Gene Codes...
In total, we sequenced 5,732 bp of 111 viral isolates spanning the \(n, p, x, m\) and \(g\) genes in the ‘3 prime half of the DMelSV genome and found no evidence for heterozygotes in this data. All sequences have been submitted to Genbank under the numbers GQ451693-GQ451695 and HQ655003-HQ655110. Additionally, we sequenced 1015 bp of the \(p\)- and \(g\)-genes of an additional 57 isolates from the UK and Ghana (559 bp and 456 bp of each gene respectively; Genbank accession numbers JN167249-JN167305).

Estimating the viral mutation rate

To refine the estimated mutation rate reported in Carpenter et al. (2007), we directly measured the mutation rate of DMelSV by re-sequencing virus isolates maintained in live fly stocks. Virus isolates were initially collected and sequenced by Carpenter et al. (2007) in 2005. The viral isolates were then maintained in live fly stocks and transferred to new vials every 3 – 4 weeks for 2.4 years (November 2005 – April 2008), at which point a new sample of viral RNA was isolated for sequencing. We compared a total of 26,857 bp over 15 lines (Supplementary material, table A3), covering partial sequences of the \(p\)- and \(g\)-genes for 13 lines (1208 – 1227 bp) and the 3’ half of the genome (5626 bp and 5729 bp respectively) for two isolates, through an automated search for mismatches in pair-wise alignments of the corresponding sequence pairs. This resulted in a total of 6 substitutions. Additionally, we used the data reported in Carpenter et al. (2007): 3 mutations in 5744 bp for line A3 (excluding 2 deletions in non-coding regions) and none in 5144 bp for line A3-E55 (excluding a 600 bp region of the genome that was modified by ADAR-induced hypermutation (Carpenter et al. 2009)) over a period of 10 – 20 years. The substitution rate was calculated by a maximum likelihood approach, assuming a Poisson process. The confidence intervals are based on a 2-
unit reduction in the log-likelihood, taking account of the uncertainty in the divergence time
of lines A3 and A3-E55 by basing the lower and upper value of the confidence intervals on
the values for the longer and shorter divergence times respectively. As there was no
significant difference between mutations at the 1st and 2nd codon position as opposed to the 3rd
position (Fisher’s Exact test, \( p = 0.453 \)), we interpret this rate as the mutation rate, rather than
the evolutionary rate, assuming that divergence times and selection strength for these lab-
maintained lines may not have been sufficient to purge slightly deleterious mutations (Rocha
et al. 2006). We have implemented this by using the mutation rate as a prior for the 3rd-codon
substitution rate for phylogenetic analyses.

Phylodynamic reconstruction

All phylogenetic analyses were carried out on two data sets: 1) ‘short’: a dataset consisting of
1015 bp of the partial p- and g-genes from 176 viral isolates, which includes the 8 previously
isolated sequences retrieved from Genbank (see under ‘viral isolates’) (58 segregating sites at
codon positions 1&2, 120 at codon position 3); 2) ‘long’: a subset of 111 viral isolates for
which we obtained the 5’ half of the DMelSV genome, covering 5735 bp of the n-, p-, x-, m-
and g-genes (332 segregating sites at codon position 1&2, 424 at codon position 3). Within
the long dataset, the sequences of all isolates were sampled within a 6-month period. In the
short dataset, the earliest samples in Genbank were sampled in November 2005, whereas the
last sample from Ghana was added in February 2010. We have not used tip date information
for phylodynamic reconstruction as tip age provides no information in this data set (an
expected 0.2 mutations between the two sequences with the largest time difference).
We used the coalescent sampling algorithm BEAST v.1.7.5 (Drummond et al. 2012) to reconstruct the phylogeny and demographic history of DMelSV. We ran 3 types of analysis a) a basic analysis of the short data set to determine the time to most recent ancestor (tmrca) for all DMelSV isolates – this analysis revealed that there are a ‘common’ and a ‘rare’ clade of DMelSV (see results) b) discrete phylogeographic analyses within the long data set of the ‘common’ clade to address population structure and migration (Lemey et al. 2009) and c) models with exponential growth rate in the ‘common’ clade both within and across populations using the long data set.

All models use the SRD06 model of sequence evolution (Shapiro et al. 2006) for consistence, as the more complex General Time Reversible model was not supported in analyses with smaller data subsets. We partitioned sites into two categories according to position (1&2 and 3) and separately estimated substitution and rate heterogeneity. Except for the models exploring exponential growth, the phylogenetic tree was reconstructed using a Jeffreys prior distribution of node heights and tree topology under a model of constant population size (Drummond et al. 2002), as this was preferred over a GMRF skyride model (Minin et al. 2008) based on a Monte Carlo Markov Chain equivalent of Akaike’s information criterion (AICM) on the marginal likelihoods as implemented in Tracer v.1.6 (Baele et al. 2012) (http://tree.bio.ed.ac.uk/software/tracer/; difference in AICM 411.93).

To test whether molecular clock rates are uniform across the tree, we first ran models with a relaxed uncorrelated lognormal molecular clock (Drummond et al. 2006). Based on the shape of the marginal posterior distribution of the clock parameter, a strict molecular clock could not be excluded. In accordance with these tests, the models assume a strict molecular clock for sequence evolution. We used the experimentally determined substitution rate as a fully
informative prior for the third codon position, accounting for uncertainty in this parameter by approximating with a lognormal distribution (mean in real space = $3.95 \times 10^{-5}$, log(standard deviation) = 0.5 substitutions/year/site). For all other priors (except location rate prior for phylogeographic models, see below), we used the default priors implemented in BEAUTI 1.7.5. We ran models with 2 runs each of 10 million MCMC generations, sampling the chain every 1000 generations with a burn-in of 1 million generations, to obtain effective sample sizes >200 for all parameters, except for the model exploring exponential growth within the ‘common’ clade, which we ran 3x as long to obtain effective sample sizes > 200. We combined the two runs using Tracer v.1.5 and examined them for convergence. We used TreeAnnotator v. 1.7.5 to produce Maximum Clade Credibility (MCC) trees.

To infer migration rates and geographic origins of branches, we fitted a discrete phylogeographic model with an asymmetric substitution model as described in Lemey et al. (2009) using Ferreira & Suchard (2008) conditional reference prior for the change in location rate. For this analysis, we used the ‘long’ dataset for the ‘common’ clade to maximize information content. We pooled samples per continent as well as into 6 geographic regions (Germany, Greece, Spain, Sweden, UK and US) and removed the lab-derived strain Hap23 (310 segregating sites at codon position 1&2, 364 at codon position 3). Well-supported rates were identified using a Bayes factors analysis based on the Bayesian stochastic search variable selection (Lemey et al. 2009), using a ln Bayes Factor of 3 as a cut-off. To investigate demographic history, we tested whether models with exponential growth rate were supported. As population structure may violate the assumptions of demographic reconstruction and may introduce biases (Ho & Shapiro 2011; Heller et al. 2013), we also performed these analyses in population subsets.
We calculated indicators of population structure and differentiation using DnaSP v5 (Librado & Rozas 2009): $\pi$, the mean pair-wise difference between sequences; $K_{ST}$, a measure of population differentiation based on the proportion of between-population nucleotide differences (Hudson et al. 1992); and the nearest-neighbor statistic $S_{NN}$, which calculates the proportion at which genetic nearest neighbors are found within the same population (Hudson 2000).

Results

The DMelSV phylogeny

The phylogenetic tree of 176 DMelSV isolates isolated from geographically diverse D. melanogaster populations reveals two clades, which we refer to as the ‘common’ and the ‘rare’ clades (Figure 1). The large majority of isolates cluster together in the common clade (fig. 1), and the ‘rare’ clade represents only 1.7% of the DMelSV isolates (AP30 NC_013135, Eir3 JN167273 and Ghana JN167305). During our collections in Cambridgeshire and Essex in the South of England during 2009, we tested 1,610 wild-caught D. melanogaster individuals for DMelSV through the CO$_2$-assay (Brun & Plus 1980; Wilfert & Jiggins 2010) and subsequent sequencing, and found that the prevalence for the ‘common’ clade was 4.9% while that of the ‘rare’ clade was 0.06%, with only one isolate found in the entire sample.

To allow us to date events on the DMelSV tree, we first estimated the rate of sequence evolution of the virus. To do this, we counted the number of sequence changes that had
occurred over several years in viruses that we maintained in Drosophila lines (see methods for details). We estimated that there are $3.95 \times 10^{-5}$ (C.I. $1.5 \times 10^{-5} – 9.5 \times 10^{-5}$) substitutions per site per year. This supports and greatly increases the precision of a previous estimate using some of the same data by Carpenter et al. (2007) of $4.6 \times 10^{-5}$ substitutions per site per year (C.I. $1.8 \times 10^{-4} – 9.5 \times 10^{-6}$), and is similar to the rates found in other rhabdoviruses (Furio et al. 2005; Sanjuan et al. 2010). We found no evidence of heterogeneity in the evolutionary rates across branches on the phylogenetic tree (see methods for details), which together with Carpenter et al.’s analyses (2007) suggests that mutation rates estimated in the lab can be extended to field isolates of DMelSV.

Based on an analysis of sequences of all 176 isolates, the most recent common ancestor of all the DMelSV isolates dates to ~3900 years ago (‘short alignment’; posterior mean: 3873 years, 95% HPD: 778 – 8198 years). The most recent common ancestor of the ‘rare’ clade dates back ~3300 years (posterior mean: 3262 years, 95% HPD 577 – 6940), while the most recent common ancestor of the ‘common’ clade occurred only ~1100 years ago (posterior mean: 1079 years, 95% HPD 177 – 2330), with all the isolates except one in this clade sharing an ancestor 517 years ago (95% HPD 98 – 1114 years). In interpreting these results, it should be noted that using contemporary evolutionary rates to infer the divergence of viral clades in the distant past may lead to underestimates of the true age of these events (Sharp & Simmonds 2011).

Structure and migration between continents

There is a striking difference in the distribution across our phylogeny of isolates from Africa and the other continents (Figure 1). We only have three isolates from Africa, but even this
sample size shows that diversity in Africa is greater than elsewhere. The three African sequences are each found on distant lineages, with two of them having no close relatives among the isolates from other continents. This is reflected in a higher genetic diversity of the African sequences compared to the pooled sample of non-African sequences (short alignment: \( \pi \text{Africa} = 0.06, \pi \text{non-Africa} = 0.01 \)).

The European and North American isolates are also genetically distinct. These isolates are mostly clustered together in the ‘common’ clade, with single sequences from each continent in the ‘rare’ clade (Figure 1). Within the common clade there is substantial genetic structure between Europe and the US, with sequences that are genetic nearest neighbours overwhelmingly being within the same continent (Hudson’s nearest neighbour statistic, short alignment: \( S_{nn} = 0.95, p < 0.001 \)) (Hudson 2000). Genetic diversity is higher in Europe than North America (long alignment, common clade: \( \pi_{\text{Europe}} = 0.008, \pi_{\text{US}} = 0.003 \)). This does not appear to be accounted for by the larger number of populations sampled in Europe, as the genetic diversity within the US, comprising samples from San Diego (California), Raleigh (North Carolina) and, for the short dataset, Athens (Georgia), is less than that of any of the European populations (Table 1). All ‘common’ samples from the US are found on a single branch of the phylogenetic tree, forming a monophyletic clade with one sample from the East of England (figs. 1 & 3). These samples span a geographic distance of 2000 – 2200 miles, whereas the maximum distance between any of the European populations is 1700 miles (Greece and Spain).

To test whether these patterns of migration between continents are supported statistically, we analysed migration between Kenya, Europe and the US using the asymmetric phylogeographic model in BEAST v. 1.7.5 (Drummond et al. 2012). Using only sequences in
the common clade, we found evidence for migration between Europe and Kenya (Bayes Factors: $BF_{\text{Europe->Kenya}} = 31.4$), and Europe and the US ($BF_{\text{Europe->US}} = 31.4$, $BF_{\text{US->Europe}} = 1003.6$). In the common clade, the most recent ancestor of European samples is 385 years (95\% HPD 92 – 786 years), 114 years for the US population (95\% HPD 26 -239 years) and 1082 years for the two samples from Kenya (95\% HPD 253 -2219 years).

Structure and migration between populations

There is also genetic structure between populations on the same continent (Figures 1 and 2). To investigate migration at this finer scale, we analysed isolates in the common clade from Germany, Greece, Spain, Sweden, UK and the US (we excluded Africa due to low sampling). These populations are strongly genetically structured as indicated by a high value of Hudson’s nearest neighbour statistic ($S_{nn} = 0.835$, $p < 0.001$). Some populations, such as the UK and US, are largely monophyletic, whilst others, such as Germany, are more scattered across the tree. This is reflected in our estimates of genetic diversity, which are higher in continental European populations than in the UK and US (Table 1).

We used a Bayesian phylogeographic approach to reconstruct routes of migration between populations. As shown in Fig. 2, we found strong support for migration from Greece to Germany, with the German population serving as a hub with migration supported to all other populations, including the US. It should be noted that immigration from other unsampled populations around the world could inflate diversity in Germany and generate this pattern, so this conclusion should be treated with caution. Additionally, we found support for migration from the US to the UK, due to a single recent back-migration event (Fig. 1 and 3). In accordance with these discrete migration routes, we find little evidence of isolation by
distance: geographic distance explains only 8% of the pairwise genetic distance between samples (Mantel-Test, 10000 replicates, \( p < 0.001 \)).

Effective population size

Sequence variation can also allow us to estimate the effective population size of the virus (\( N_e \)) and reconstruct how it has changed through time. Using the alignment of all 176 sequences, we estimate that the global effective population size of DMelSV is \( 1.7 \times 10^4 \) (95% HPD 0.4 \( x \) \( 10^4 \) – 3.7 \( x \) \( 10^4 \)). Population structure means that this estimate depends on our choice of sampling scheme, so we also estimated \( N_e \) in the monophyletic branches in the largely closed UK and US populations (UK: \( N_e = 783 \), 95% HPD 97 – 1825; US: \( N_e = 3023 \), 95% HPD 455 – 7310). For these estimates, we assume a constant \( N_e \) and 10 generations per year as for the host (e.g. Richardson et al. 2012), as this virus is exclusively vertically transmitted and shows no quasi-species formation (pers. comm. M.L. Wayne).

To investigate how \( N_e \) may have changed through time, we attempted to minimise the confounding effects of population structure (Ho & Shapiro 2011; Heller et al. 2013) by analysing each population separately as well as all the non-African isolates combined. First, we compared a model of exponential growth to a model of constant population size within the common clade. While exponential growth is supported overall (exponential growth rate 0.024, 95% HPD 0.005 – 0.05, doubling time = 29.4 years), at a population level, this pattern is only supported in the Greek and US populations and the rate of growth is very uncertain (common clade only; Greece: growth rate = 0.018, 95% HPD 0.003 – 0.040, doubling time = 37.7 years; US: growth rate = 0.17, 95% HPD 0.023 – 0.404, doubling time = 4.1 years).
Population structure tends to give a false signature of a declining $N_e$ (Heller et al. 2013), so this is unlikely to confound this analysis.

Local and temporal patterns

We found evidence for changes in the genetic composition of the viral population in Greece through time. The Greek isolates are found on two branches of the phylogenetic tree (Fig. 1 & 3). As shown in Fig. 1, the isolates from 2005 ($n = 2$) and 2006 ($n = 9$) are exclusively found on one branch, with the 2007 isolates distributed over both branches ($n = 24$ and $n = 31$ respectively; Fisher’s exact test comparing prevalence in 2005, 2006 and 2007: $p = 0.002$). The change in genotype frequencies is reflected in significant genetic structure between the samples from the different years ($K_{S_{short}} = 0.13173, p < 0.001$).

Populations in the East of England (Cambridgeshire and Essex) were also sampled repeatedly, but here there were no changes through time. Flies were collected from Essex in 2005, 2007 and 2009 ($n = 10$, $n = 10$ and $n = 5$, respectively), and in 2009 from Cambridgeshire ($n = 50$, ~45km from Essex). Here all the years were represented on the same major branch and there was little change through time ($K_{S_{short}} = 0.04667, p < 0.05$). There only limited population structure within the Cambridgeshire sample, comprising of samples from Cambridge town and the surrounding villages Girton, Fulborn, St. Ives and Waterbeach, with only a minority of genetic nearest neighbours coming from the same population ($S_{nn} = 0.36$, $p < 0.05$).

Discussion
DMelSV is a vertically transmitted pathogen that can only be transmitted by the eggs or sperm of *D. melanogaster*. Despite being passed between generations along with the host genome, critical population genetic parameters affecting patterns of genetic variation within the viral populations are very different from its host. We estimate that the effective population size of DMelSV is roughly 10,000, which is about two orders of magnitude smaller than *D. melanogaster* (Andolfatto & Przeworski 2000). This low $N_e$ is expected given that the virus only infects a few percent of flies, so all else being equal its effective population size will be proportionately less than that of fly genes (Carpenter et al. 2007). The viral mutation rate is much greater — we see about $3 \times 10^{-6}$ mutations per site per fly generation in viruses maintained in lab fly stocks (this may be an underestimate of the mutation rate if some deleterious mutations are purged by selection). In contrast, the mutation rate of *D. melanogaster* is three orders of magnitude lower ($3.5 \times 10^{-9}$ per site per generation for nuclear genes (Keightley et al. 2009)). Together these factors mean that patterns of population structure and migration can be seen at a temporal resolution in the DMelSV population that are difficult to detect when studying *D. melanogaster* directly without taking recourse to whole-genome population sequencing (Campo et al. 2013).

As expected given its low effective population size, we see far greater population structure in DMelSV than one would expect to find in the *Drosophila* genome. This variation can therefore potentially be used to provide high resolution insights into the past demography and migration of the flies themselves. This could be valuable, as despite being one of biology’s best-studied organisms, many aspects of *D. melanogaster*’s field ecology remain obscure.

However, this must also be done with caution, as natural selection within the viral population, fluctuations in prevalence and the spread of the virus though uninfected populations may have overwitten the signature of host demography. In particular, there is evidence DMelSV
populations have recently evolved to overcome host resistance due to a polymorphism in the gene \textit{ref(2)P} (Fleuriet & Sperlich 1992; Fleuriet & Periquet 1993; Wayne \textit{et al.} 1996; Bangham \textit{et al.} 2007; Wilfert & Jiggins 2013). Some of the patterns that we see may therefore reflect the spread of these virulent viral genotypes through populations.

At a continental scale, the evolutionary history of the virus mirrors that of the host, with patterns of genetic variation suggesting that the virus originated in Africa, migrated to Europe, and then on to the US. Does this mean that we are seeing the traces of the out of Africa migration of flies in the viral population? At first sight our time-calibrated phylogeny would suggest not, as while the most recent common ancestor of all our isolates dates back ~3900 years, 98\% of the isolates in Europe shared a common ancestor just ~520 years ago. In contrast, even the most recent estimates of the genetic bottleneck that flies experienced during the initial migration of \textit{D. melanogaster} out of Africa date back several thousand years, and other estimates are substantially older (Lachaise \textit{et al.} 1988; Ometto \textit{et al.} 2005; Thornton & Andolfatto 2006). However, it is thought to be common for long-term rates of sequence evolution in viruses to be much slower than are observed over short time periods, and this may frequently result in massive underestimates of the age of distant nodes on viral phylogenies (Sharp & Simmonds 2011). With this bias in mind, along with the substantial errors attached to all the estimates, we would tentatively suggest that the most parsimonious explanation of our data is that we are seeing the traces of the global fly migration (although selection in Europe may also be important; see below). In North America, we estimate that the most recent common ancestor of the ‘common’ viral clade dates back ~114 years to the turn of the last century. This suggests that DMelSV arrived in the United States at the beginning of the invasion of this continent by \textit{D. melanogaster} in the late 19\textsuperscript{th} century (Keller
A recent invasion is also supported by a local pattern of population growth and the lack of genetic structure across the US (albeit with limited sampling).

Within Europe, DMelSV populations are highly genetically structured, with most sequences in some populations forming almost monophyletic clades. This would suggest that there are low levels of migration across Europe, despite there being limited genetic structure at this geographical scale in the Drosophila genome, which has low but significant $F_{ST}$ values across European populations for D. melanogaster (Caracristi & Schlotterer 2003; Verspoor & Haddrill 2011). There is, however, still some ongoing migration, with recent immigrants clearly visible on the phylogeny, and this allows us to reconstruct the movement of the flies carrying DMelSV between populations. Across the phylogeny, in a given year a DMelSV lineage has approximately a 1 in 700 chance of switching to a different population ($\mu=1.39 \times 10^{-3}$ per year; 95% HPD $2.47 \times 10^{-4} - 3.12 \times 10^{-3}$; see Lemey et al. (2009) for details). The fruit trade may mediate this migration. For example, drosophilids are frequently found in shipments of imported fruit in the UK (Reid & Malumphy 2009), a country that in 2011 imported 33% of its fresh fruit from the EU and 55% from outside the EU (DEFRA 2012), including close neighbors of the southern US (Costa Rica: 8%, Colombia: 7%, Dominican Republic: 6%, Ecuador: 4%) (DEFRA 2012).

Overwintering and the persistence of populations in cold climates is a particular aspect on which studying DMelSV population genetics can begin to shed some light. D. melanogaster is not tolerant of extreme cold (Izquierdo 1991) and can only overwinter as adults, with females entering reproductive diapause (Saunders et al. 1989). Bouletreau-Merle et al. (2003) have shown that in the Rhone Valley in Southern France, D. melanogaster is able to overwinter in refugia inside human dwellings, being tolerant to underfeeding at cool
temperatures, thus forming a permanent population in this area. Its sister species *D. simulans*, which does not use such shelters, however is re-invading this area from southern refugia after cold winters (Bouletreau-Merle *et al.* 2003). *D. simulans* does not occur in cold climates north of the Alps in central and northern Europe (Babaaissa *et al.* 1988). The existence of stable latitudinal clines in allele frequencies in *D. melanogaster* indicates that there is some genetic continuity between generations in populations in cold climates. However, the relative importance of recolonisation from the warmer south versus local persistence in these populations is unknown. If a significant proportion of the population is recolonized from the south every generation, then migration should predominantly occur from the south to north.

Our analysis of DMelSV migration patterns suggests that populations persist locally in cold winters and recolonisation from southern Europe is not important. The isolates from the two populations with the harshest winters, Sweden and Germany, have the most phylogenetically diverse DMelSV strains in Europe (Figure 3). Our analysis of migration patterns suggests that central German populations, which experience extreme cold in the winter, show no evidence of immigration from the south. Therefore, rather than being recolonized every generation, this population appears to have persisted over the long term and has probably exported migrants elsewhere. This is despite evidence that DMelSV infection may reduce overwintering success in *D. melanogaster* (Fleuriet 1981b), which further increases the chances of recolonisation from the south. While our data indicates that northern populations are not recolonized from the Iberian or Balkan peninsulas every spring, an alternative explanation of our data is that northern European populations could be recolonised from un-sampled populations in Africa, Asia or South-America through the global fruit-trade, or from Italy. While sampling of these populations would be required to formally reject this hypothesis, it seems improbable to us
that immigration from these areas would be greater than from nearby populations within Europe.

Within the South-East of England we carried out a detailed large-scale study of the local population structure of DMelSV. This revealed very little population structure and variation over years, and in particular showed no evidence of migration other than the long-term seeding from Northern America and central Europe. This suggests that the populations in the South-East of England are long-term resident populations. In Athens, Greece on the other hand, we found a shift in genetic diversity between 2006 and 2007: the 11 samples from 2005 and 2006 were found on only one major branch of the phylogeny, whereas 31 of the 55 samples from 2007 fall onto another major branch. Assuming that this pattern is not a quirk of our limited sampling in the earlier years, this change could be due to seasonal fluctuations, migration from other populations, perhaps reflecting local population structure, or selection favoring viruses from the new clade in 2007. Such rapid changes in viral types are possible in this system (Wilfert & Jiggins 2013).

In this host-parasite system, natural selection due to coevolution with the host may be an important factor shaping patterns of genetic variation (Wilfert & Jiggins 2013). DMelSV infected flies suffer a reduction in egg viability (Fleuriet 1981a) and an increase in larval development time (Seecof 1964). Additionally, infected flies may have reduced survival under stressful conditions such as limited resources (Yampolsky et al. 1999) and overwintering (Fleuriet 1981b). Several polymorphic genes in D. melanogaster populations convey resistance against DMelSV (Gay 1978; Bangham et al. 2008; Magwire et al. 2011; Magwire et al. 2012). As mentioned above, an allele of the ref(2)P gene that can block maternal transmission of DMelSV, has undergone a partial selective sweep in D.
melanogaster (Wayne et al. 1996; Bangham et al. 2007). However, only certain avirulent DMelSV genotypes are affected by ref(2)P, and in Languedoc, France (Fleuriet & Periquet 1993) and Tubingen, Germany (Fleuriet & Sperlich 1992) (also sampled in this study), the avirulent virus type was mostly replaced by virulent viruses, which are largely unaffected by ref(2)P, in the 1980s and early 1990s. We have previously characterized the effect of the resistant allele of ref(2)P on five of the isolates in our phylogeny (Carpenter et al. 2007) (marked on Figure 1). One isolate in the ‘rare’ clade was sensitive to the ref(2)P resistance allele, while four isolates scattered across the common clade were all virulent and are not affected (Carpenter et al. 2007). As it is not known whether DMelSV evolved to evade the effects of ref(2)P once or many times, it is not clear exactly how selection on this trait may have affected the patterns we see. However, it is possible that the common clade is largely virulent, and the migration patterns may be partly driven by selection on this trait.

In conclusion, we show that the vertically transmitted D. melanogaster pathogen DMelSV has been associated with its host for several thousand years. It has a much smaller effective population size and higher mutation rate than its host, resulting in high levels of genetic structure and a detailed record of recent migration events. The virus has spread around the world following the same migration route is its host, and there is a low rate of ongoing migration within and between continents. The degree of genetic structure and direction of migration suggest that D. melanogaster can persist in temperate and cold climates.
We would like to thank Florian Bayer, Mitchel Chewi and Heather Cagney for assistance in the lab. Trevor Bedford, Sam Lycett and John Welch provided invaluable advice on statistics and phylogenetic reconstruction. We are very grateful to the people who helped collecting viral isolates in the field: Natasa Fytrou & family, Jennifer Carpenter, Sandra South, Rebecca Schulte, Florian Bayer & family, Severin Weybora (Bayrische Landesjagdschule), Ben Longdon, Claire Webster, Heather Cagney, Darren Obbard and John Pool. This work was financially supported by a Leverhulme Trust grant and a Royal Society University Research Fellowship to F.J.
References


Fleuriet A (1981a) Comparison of various physiological traits in flies (*Drosophila melanogaster*) of wild origin, infected or uninfected by the hereditary rhabdovirus sigma. *Archives of Virology, 69*, 261-272.


Data Accessibility

DNA sequences with all accompanying information: Genbank accession numbers GQ451693-GQ451695, HQ655003-HQ655110 and JN167249-JN167305. Alignments for the 'short' and 'long' datasets, the BEAST xml files for fig. 1, 2&3 as well as the tree files for fig.1 and 3 are archived on dryad (doi:10.5061/dryad.7n696).

Author contributions

LW collected samples, performed analyses and wrote the paper. FJ collected samples and wrote the paper.
Figure Legends

Figure 1: Maximum clade credibility (MCC) tree of 176 DMelSV isolates reconstructed from the sequence of 1015bp of the p- and g-genes (the ‘short’ dataset). The scale shows estimated divergence dates based on laboratory estimates of the rate of viral evolution. The coloured bar indicates the population of origin of the branches. The grey bars indicate the years in which samples from Greece were collected and arrows indicate 5 isolates of DMelSV that are virulent (black arrow), i.e. can overcome the host’s ref(2)P host resistance mutation, or avirulent (red arrow) (Carpenter et al. 2012). Posterior support >0.6 is indicated for nodes of at least 3rd order.

Figure 2: Migration patterns within the ‘common’ clade in Europe and North America. The weight of the line indicates the Bayes Factor support (from thin to thick arrows: BF = 3 – 10, 10 – 30, >30)

Figure 3. MCC phylogeny of the ‘common’ DMelSV populations in Europe and North America, showing the reconstructed ancestral population states. The tree was reconstructed from 5732bp of sequence from a subset of the isolates shown in Figure 1. The scale shows estimated divergence dates based on laboratory estimates of the rate of viral evolution. The
branches are colored according to reconstructions of the population in which the lineages were found. Posterior support >0.6 is indicated for nodes of at least $3^{rd}$ order.

Tables

Table 1: Genetic diversity in DMelSV populations. Estimates are based on either 1015bp of sequence from 176 isolates ($\pi_{\text{short}}$), or 5732bp of sequence from a subset of 111 isolates ($\pi_{\text{long}}$).

<table>
<thead>
<tr>
<th></th>
<th>$\pi_{\text{long}}$</th>
<th>$\pi_{\text{short}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germany</td>
<td>0.00767</td>
<td>0.00789</td>
</tr>
<tr>
<td>Greece</td>
<td>0.00623</td>
<td>0.00625</td>
</tr>
<tr>
<td>Spain</td>
<td>0.0031</td>
<td>0.00804</td>
</tr>
<tr>
<td>Sweden</td>
<td>0.00927</td>
<td>0.00888</td>
</tr>
<tr>
<td>UK</td>
<td>0.00463</td>
<td>0.00261</td>
</tr>
<tr>
<td>US</td>
<td>0.00266</td>
<td>0.00197</td>
</tr>
<tr>
<td>Total/Europe</td>
<td>0.00822</td>
<td>0.00795</td>
</tr>
</tbody>
</table>