Role of Host-Driven Mutagenesis in Determining Genome Evolution of Sigma Virus (DMelSV; Rhabdoviridae) in *Drosophila melanogaster*

Helen Piontkivska[1,*], Luis F. Matos[2,3], Sinu Paul[1,4], Brian Scharfenberg[1,5], William G. Farmerie[6], Michael M. Miyamoto[7], and Marta L. Wayne[7,8]

[1]Department of Biological Sciences and School of Biomedical Sciences, Kent State University, Kent, OH

[2]Department of Entomology & Nematology, University of Florida, Gainesville, FL

[3]Department of Biology, Eastern Washington University, Cheney, WA

[4]Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, La Jolla, CA

[5]Ohio University Heritage College of Osteopathic Medicine, Athens, OH

[6]Interdisciplinary Center for Biotechnology Research University of Florida, Gainesville, FL

[7]Department of Biology, University of Florida, Gainesville, FL

[8]Emerging Pathogens Institute University of Florida, Gainesville, FL

*Corresponding author: E-mail: opiontki@kent.edu.

Accepted: August 29, 2016

Data deposition: This project has been deposited at the NCBI Bioproject under the accession PRJNA326864.

Abstract

Sigma virus (DMelSV) is ubiquitous in natural populations of *Drosophila melanogaster*. Host-mediated, selective RNA editing of adenosines to inosines (ADAR) may contribute to control of viral infection by preventing transcripts from being transported into the cytoplasm or being translated accurately; or by increasing the viral genomic mutation rate. Previous PCR-based studies showed that ADAR mutations occur in DMelSV at low frequency. Here we use SOLiD™ deep sequencing of flies from a single host population from Athens, GA, USA to comprehensively evaluate patterns of sequence variation in DMelSV with respect to ADAR. GA dinucleotides, which are weak targets of ADAR, are strongly overrepresented in the positive strand of the virus, consistent with selection to generate ADAR resistance on this complement of the transient, double-stranded RNA intermediate in replication and transcription. Potential ADAR sites in a worldwide sample of viruses are more likely to be "resistant" if the sites do not vary among samples. Either variable sites are less constrained and hence are subject to weaker selection than conserved sites, or the variation is driven by ADAR. We also find evidence of mutations segregating within hosts, hereafter referred to as hypervariable sites. Some of these sites were variable only in one or two flies (i.e., rare); others were shared by four or even all five of the flies (i.e., common). Rare and common hypervariable sites were indistinguishable with respect to susceptibility to ADAR; however, polymorphism in rare sites were more likely to be consistent with the action of ADAR than in common ones, again suggesting that ADAR is deleterious to the virus. Thus, in DMelSV, host mutagenesis is constraining viral evolution both within and between hosts.

Key words: ADAR, immunity, mutation, quasispecies, phylogeny, RNA virus.

Introduction

The sigma virus of *Drosophila melanogaster* (DMelSV) is a biparentally transmitted parasite which exerts a fitness cost on its hosts in terms of development time and fecundity (Fleuriet 1996; Yampolsky et al. 1999; Wayne et al. 2011; Brusini et al. 2013). DMelSV is a member of the *Mononegavirales*, the virus order that includes the causative agents of rabies, viral hemorrhagic septicemia, infectious hematopoietic necrosis and other economically important diseases of humans and livestock (Hogenhout et al. 2003). A typical rhabdovirus, DMelSV is encoded by a negative single-stranded RNA genome. Its genome is relatively small at ~12 kb and encodes six protein-

coding genes, namely, 3′-N-P-X-M-G-L-5′, which correspond to five structural proteins, namely, the nucleoprotein (N), the polymerase-associated protein (P), the matrix protein (M), the glycoprotein (G), and the polymerase (L), respectively (Teninges et al. 1993; Longdon et al. 2012). The function of the X gene (also referred to as PP3 (Walker et al. 2011)) remains unclear, although it may play a role in innate antiviral responses (Tsai et al. 2008; Longdon et al. 2010). One of the most interesting features of DMelSV is its strictly vertical, biparental mode of transmission (Brun and Plus 1980; Longdon and Jiggins 2010). RNA viruses are notorious for their high mutation rates, although negative strand RNA viruses may have lower mutation rates than positive strand viruses (Domingo and Holland 1997; Drake and Holland 1999; Duffy et al. 2008). The available estimates of the rate of evolution for DMelSV range from $\sim 3.95 \times 10^{-5}$ (Wilfert and Jiggins 2014) to $\sim 4.6 \times 10^{-5}$ substitutions/site/year (Carpenter et al. 2007), which are lower than the estimated average for RNA viruses, with most viruses having rates within an order of magnitude of $\sim 10^{-3}$ substitutions/site/year (Jenkins et al. 2002; Holmes 2003, 2009). Vertically transmitted viruses are also thought to have relatively low substitution rates (Roossinck 2010). Similar to other negative-sense RNA viruses (Chare et al. 2003), no evidence of recombination in DMelSV has been found (Carpenter et al. 2007; Longdon et al. 2012).

In a previous study of genetic variation in DMelSV, a subset of the variants was attributed to host-encoded adenosine deaminases acting on RNA (ADAR) (Carpenter et al. 2009). ADARs are enzymes that target double-stranded RNA (dsRNA), and are a key component of the "editome" (Chen et al. 2014). Although the editome is thought to have some functional importance in terms of metazoan gene regulation, for example (Liu et al. 2014), other analyses highlight that much (though not all) RNA editing is deleterious (Xu and Zhang 2014, 2015). These analyses beg the question of why RNA editing evolved. The presence of dsRNA is often a sign of viral infection, and so some have proposed that ADARs may have evolved as an antiviral response (Keegan et al. 2001).

Changes caused by ADAR can be identified with fair confidence because of their specificity: ADAR changes adenosine (A) to inosine (I). This base change then leads to I pairing with a cytosine (C) instead of a thymidine (T) (Keegan et al. 2001), thereby effectively causing a transition mutation from A to G. Interestingly, the likelihood that a given A will be deaminated is conditioned on its 5′ neighbor. Deamination is more likely to happen when the 5′ neighbor is either an A, U, or C than if it is a G (Lehmann and Bass 2000; Mueller et al. 2006). Furthermore, changes caused by ADAR are introduced into dsRNA in a spatially clustered fashion (Carpenter et al. 2009). Thus, a clumping of A-to-G transitions is another signal of ADAR editing.

In the *Drosophila* genome there is a single ADAR gene, *dADAR* (Palladino et al. 2000). Previously it had been shown that ADAR activity can lead to hypermutation in some rhabdoviruses, for example in DMelSV (Carpenter et al. 2009), which in turn can contribute to a decline in pathogen virulence (Meyers et al. 2003). Reduced virulence could be attributed to ADAR-induced mutations in viral transcripts per se, possibly leading to inactive gene products (e.g., by changing the physico-chemical properties of the encoded amino acids). This in turn can lower the cost of infection because the mutated virus has less of an impact on the host. Alternatively, decreased virulence could be the result of reduced viral load, because the disrupted transcripts may include components necessary for viral proliferation such as the viral polymerase, or by increased mutation in new viral genomes. Disruption of transcripts, as in the first explanation, can be seen as a form of a tolerance strategy within a broader array of host antiviral defenses, where tolerance is defined as differential host fitness for the same viral load. In contrast, reducing cost of infection by slowing down viral proliferation and hence reducing viral load, as in the second explanation, is a form of a resistance to infection (e.g., Medzhitov et al. 2012). At present, it is unclear to what extent these two complementary processes are operating. Finally, because ADAR can also attack the dsRNA replication intermediates of DMelSV, viral genomic mutation may further diminish the cost of infection of the host either by reducing functional gene product or by reducing titer, further muddling the tolerance/resistance dichotomy.

Here we use SOLiD™ deep sequencing from multiple individuals within a single, wild population of the host *D. melanogaster* to comprehensively evaluate the patterns of sigma genome sequence variation within a single population and within hosts, and to determine how much, if any, of the observed variation can be attributable to ADAR. We found that ADAR-driven changes contribute to the observed DMelSV variation both within and between hosts. We also found that sites that are variable among hosts are more likely to be hypervariable in only one or at most two flies, and that such variants are consistent with ADAR activity.

Material and Methods

Fly Collection

Flies were collected from a single location in Athens, GA in August 2007 using banana baits. Flies were allowed to oviposit on prepared *Drosophila* food for 24 h, after which DMelSV infection was determined by exposing the flies to CO_2 (flies infected with this virus become paralyzed upon exposure; reviewed in Brun and Plus 1980). The offspring of each infected fly were propagated as independent isofemale lines, and were kept under standard reading conditions (24°C and 16:8 light: dark).

RNA Extraction

RNA was extracted from each of the five infected females using TRIzol (http://www.thermofisher.com/us/en/home.html; last

accessed August 29, 2016) according to the standard manufacturer's protocol. Purified RNA was quantified using a NanoDrop and for each fly, 500 ng of RNA were used per reaction.

Reverse Transcription and PCR

About 6,370 consecutive nucleotides of the viral genome (partial N, G, M, X, P, and partial L genes) were amplified from each fly in four overlapping pieces 1–2 kb in length using the Superscript III one-step RT-PCR system with Platinum® Taq High Fidelity and following the manufacturer's standard protocol (http://www.thermofisher.com/us/en/home.html; last accessed August 29, 2016).

Libraries were made from each sample, following the standard manufacturer's protocol for SOLiD™ sequencing. Each sample was bar-coded, pooled, and run on a single region of a SOLiDTM 5500×l (Applied Biosystems, Foster City, CA, USA) plate, generating 48 bp reads.

Sequence Assembly

A total of five sigma virus samples (i.e., from five different flies captured from the field in Athens, GA, USA) were sequenced. There was an additional sample that was a technical replicate of one of the flies; however, the results obtained between the two replicates were essentially the same (see fig. 1). Thus, for the rest of the paper, we focus on the five unique biological samples. Genome assembly was performed using CLC Genomics Workbench (https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/; last accessed August 29, 2016), aligning to the published partial genomic sequence of sigma virus NCF strain (isolated in NC, USA; GenBank accession HQ655102), as this sequence was geographically closest to our population. Of NCF's 5,358 nucleotides, a total of 5,258 nucleotide positions corresponding to protein-coding regions were mapped. Unresolved P and M fragments of HQ655102 were filled by mapping reads to the corresponding sections of another strain, 234HRC (GenBank accession X91062, collected in France), using 50 nucleotides on either side of the gap, to ensure the best matches. SOLiD ™ sequencing has a relative low error rate of less than 0.01–0.06% (Glenn 2011), which is relevant given our average sequencing coverage of about 58,000+ reads per site (see supplementary table S1, Supplementary Material online). To minimize the error rate, we used stringent cutoffs for trimming about 39% of the reads prior to mapping (i.e., removing low quality sequences, reads with more than 2 ambiguous nucleotides or those shorter than 46 nucleotides long). We then excluded sites with less than 100 mapped reads and sites corresponding to intergenic noncoding regions. Overall, a total of 5,313 sites contained 100 or more mapped reads in at least one of the strains, and the vast majority of sites produced high-quality mapped reads in all samples.

Because of the context dependency of ADAR mutations, sites whose 5′ neighbors could not be resolved were excluded from considerations, so that for every nucleotide position, the unambiguous designation of ADAR-strong or ADAR-weak site could be made. We also excluded 15 codons that differed between the oldest available sigma strain, X91062 (11 of these were annotated as 3′ UTR of its G gene) and all the other strains. Thus, the analysis was conducted on a set of 5,166 nucleotide positions (total of 1,722 codons) that were shared between a world-wide collection of existing DMelSV sequences and across our sample of wild flies from the Athens GA population (of these, there were 377 codons from N gene, 306 from P, 294 from X, 217 from M, and 528 codons from G gene, respectively). A flow chart illustrating the curating process is provided as supplementary figure S1, Supplementary Material online.

ADAR Designation in the dsRNA Context

ADAR sites were designated for each consensus A nucleotide in the BC7 sample, taking into account 5′ adjacent nucleotides. The BC7 sample was selected because this sample nested within the North American clade and had the largest average number of reads per site. However, the site designations were similar with other strains, with few differences among individual strains. A-harboring sites on the positive sense strand were classified as "strong" or "weak" ADAR sites if they had either A, C, U or G 5′ neighbors, respectively (Lehmann and Bass 2000; Mueller et al. 2006). Because ADAR can act on either strand of the dsRNA target, we also considered U-harboring residues on the coding strand as potential ADAR sites (i.e., they are A nucleotides on the complementary strand), and thus classified them according to their respective complementary 5′ neighbor as well. Thus, coding strand Us were classified as strong ADAR sites if they had a coding strand 3′ A, U or G neighbor, and weak ADAR sites if they had a 3′ C neighbor. There was a total of 2,817 ADAR sites (1,085 and 992 strong and 413 and 327 weak A- and U-sites, respectively). The remaining 2,349 sites harbored either C or G nucleotides. The strong and weak sites, considering both As and Us, were distributed approximately equally across the genes, with the highest proportion of weak sites (29.4%) in the P gene and the lowest (24.1%) in the X gene (see supplementary table S2, Supplementary Material online). We consider the weak ADAR sites as being effectively "resistant" to ADAR-driven modifications, while strong ADAR sites can be thought of as "susceptible". In other words, the susceptible sites may be potentially deleterious, due to their higher mutability, and thus may be eliminated by selection.

Variability Among Viruses

To examine among virus/population sequence variability patterns, we collected 114 DMelSV nucleotide sequences from GenBank. The respective protein coding sequences of five

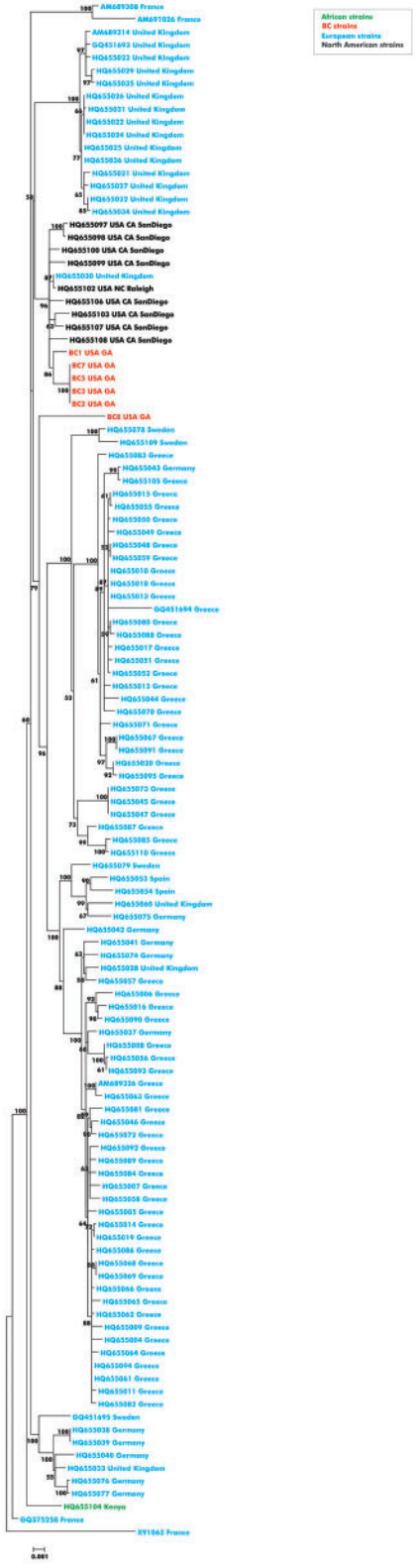

Fig. 1.—Maximum likelihood phylogenetic tree based on the General Time Reversible model, taking into account the gamma distribution and the proportion of invariable sites (GTR + G+I). Numbers on the branches represent the bootstrap support (out of 100 bootstrap replications). Only the bootstrap values above 50% are shown. Each sequence is identified by its GenBank accession number and color-coded according to the place of origin. BC strains are shown in red, while European, African, and North American strains are shown in blue, green, and black, respectively.

genes, N through G, together with five consensus BC sequences, were aligned as per respective amino acid alignment using ClustalW as implemented in MEGA6 (Tamura et al. 2013). We reconstructed a maximum likelihood-based (ML) tree, using the program PhyML 3.0 (Guindon et al. 2009, 2010) (http://atgc.lirmm.fr/phyml/; last accessed August 29, 2016), based on the General Time Reversible model with gamma correction (approximated by four categories), and accounting for presence of invariable sites (GTR + G+I). The reliability of the tree topology was evaluated using 100 bootstrap replications (Felsenstein 1985) (fig. 1, only values above 50% bootstrap support are shown above branches). We have also constructed a phylogeny using the minimum evolution (ME) method in MEGA6 (Tamura et al. 2013). Distances were computed using the maximum composite likelihood method (Tamura et al. 2004). To evaluate the reliability of internal branches, 1,000 bootstrap replications were used. The trees were rooted using the oldest available sigma sequence from France (X91062); however, the exact same topology is obtained when midpoint rooting is used. Because both the ML and ME topologies were essentially the same, only the ML tree is presented here (fig. 1).

We also classified whether or not any individual coding positions were polymorphic in the multiple sequence alignment of all strains. A site was classified as variable if one or more of the viruses surveyed had a different nucleotide at that position than the majority consensus nucleotide.

Of the 595 variable sites, 125 were polymorphic for A-to-G transitions. The degree of spatial clustering among these 125 potential ADAR-edited sites was evaluated on a gene-by-gene basis with the permutation test of Carpenter et al. (2009). For each gene, the mean distance of every A-to-G transition to another (in number of bases) was determined from their absolute differences in alignment position. The alignment positions for these A-to-G transitions were then randomly permutated 1,000 times followed by the calculation of the mean distances for the 1,000 permutations as before. The average distances for the 1,000 permutations were thereafter summarized as the null distribution for the observed mean of the gene.

Resistant and Susceptible ADAR Dinucleotide Frequencies

To test whether or not resistant ADAR dinucleotides are represented according to a random expectation in the genomic sequences, we first used Wordcount of EMBOSS v6.3.1 (Rice et al. 2000) to estimate the dinucleotide frequencies for the two representative sequences, GQ456194 and X91062.

These two sequences were selected for these tests, because they are the two most different genomes in our study according to their pairwise proportional distance ($p = 0.0250$), thus, they can be expected to reflect the breadth of sequence diversity within our sample. In these tests, we focused on the dinucleotide frequencies for GA, which is less susceptible to ADAR editing than AA, CA, or UA dinucleotides. In turn, we also evaluated the dinucleotide frequencies for UC, because this dinucleotide of the positive strand is the complement of GA. By also focusing on UC, we were able to test for an enrichment of both GA and UC dinucleotides in the stems of the folded viral RNA as well as for an overrepresentation of GA in the negative strand. To assess the significance of the observed dinucleotide frequencies, we wrote a custom C ++ program to simulate 1,000 random sequences for each representative genome. These random sequences were simulated according to the corresponding observed lengths and codon-specific base frequencies for the first, second, and third positions of GQ456194 and X91062. The dinucleotide frequencies for the random sequences were then estimated as before and these estimates were summarized as the underlying null distributions for the observed dinucleotide frequencies of the two selected genomes.

Results

This study is focused on coding sites only, as coding sites constitute the vast majority of the viral genome (e.g., 96.3% of the reference genome of AP30, Genbank accession AM689309). Moreover, functional inference based on codon position is straightforward relative to inferences on noncoding sites, particularly in viruses. Average coverage (± standard error of the mean) ranged from 12,438 (±199) reads per site in the BC1 sample to 97,815 (± 1,593) reads per site in BC8, with the smallest nonzero number of reads per site at 188 reads. Median coverage ranged from 5,412 to 51,648 reads per site. Overall, coverage of the coding sequences was extensive, with the overall average of over 58,000 reads per site across all samples (58,360 ±490 reads per site across all samples (supplementary table S1, Supplementary Material online)). Notably, the vast majority of reads had the consensus nucleotide, with only a minor variant fraction. This pattern is consistent with the low substitution rate previously reported for DMelSV (Carpenter et al. 2007; Wilfert and Jiggins 2014). Notably, nonconsensus variants were not uniformly distributed across sites, with certain sites being consistently variable among BC samples, indicating that a common force (or forces) may be driving this pattern.

The DMelSV Genome is Enriched for Resistant ADAR Dinucleotides

The vulnerability of adenosine residues to host-driven ADAR mutation of dsRNA depends on their upstream neighbors: A residues preceded by G residues are the most resistant ("weak" ADAR sites), while A preceded by A, U, or C residues are much more susceptible to editing by ADAR ("strong" ADAR sites; Lehmann and Bass 2000). Accordingly, we hypothesized that GA dinucleotides would be overrepresented in the positive strand RNA (relative to their random sequences). This can be illustrated with the following hypothetical example. In a putative genome that has equal frequencies of nucleotides, one can expect the dinucleotides AA, CA, GA, and UA to occur in equal numbers. Under the assumption of a mutational process equivalent to extreme ADAR editing (i.e., 100% edits of "strong" sites), dinucleotides AA, CA, and UA will all be mutated to AG, CG, and UG, respectively, while GA dinucleotides are not changed. In this case, GA overrepresentation could act as a mechanism to lessen the burden of a potential amino acid substitution (due to ADAR editing) because the "weak" sites will be edited less frequently than the "strong" sites.

However, given that ADAR acts on dsRNA, we also expected that the complement of GA (i.e., UC) would also be enriched in the positive strand. Overrepresentation of both GA and UC dinucleotides would be evidence of the maintenance of GA and UC pairing in folded RNA and/or for the roles of the negative strand in the viral life cycle (Carpenter et al. 2009).

Our dinucleotide tests with X91062 and GQ451694 support our prediction of an overrepresentation of GA dinucleotides on the positive strand (figs. 2 and 3). Specifically, the observed frequencies of GA for X91062 and GQ451694 are 0.0785 and 0.0799, respectively. Except for one random sequence of GQ451694, the observed GA frequencies for the two representative genomes are greater than those for all 1,000 of their simulated sequences. Thus, at $P \leq 0.002$, we find a strong enrichment of resistant GA dinucleotides on the positive strand of both genomes.

Conversely, the observed UC frequencies for X91062 and GQ451694 are 0.0623 and 0.0633, which are exceeded by 27 and 17 of their 1,000 random sequences, respectively (figs. 2 and 3). Thus, we find that the support for UC enrichment on the positive strand is borderline nonsignificant and significant but weak for X91062 and GQ451694 ($P = 0.054$ and 0.034), respectively. Correspondingly, at best, the positive strand is only weakly enriched for UC dinucleotides. Still, regardless of its significance, of greater importance is that the dinucleotide frequency of UC on the positive strand is less than that of GA (χ^2 tests of goodness-of-fit for equal UC and GA counts, df = 1, $P \leq 0.002$ for both X91062 and GQ451694). Collectively, these results support a strong GA enrichment on the positive strand, which is unmatched by the complementary UC dinucleotides.

ADAR Shapes DMelSV Variation Between Hosts

We constructed a multiple sequence alignment of the partial coding sequences including 114 sequences from the GenBank, then used our sequence alignment to identify sites harboring genetic variation. Consistent with prior observations of low substitution rate (Carpenter et al. 2007; Wilfert

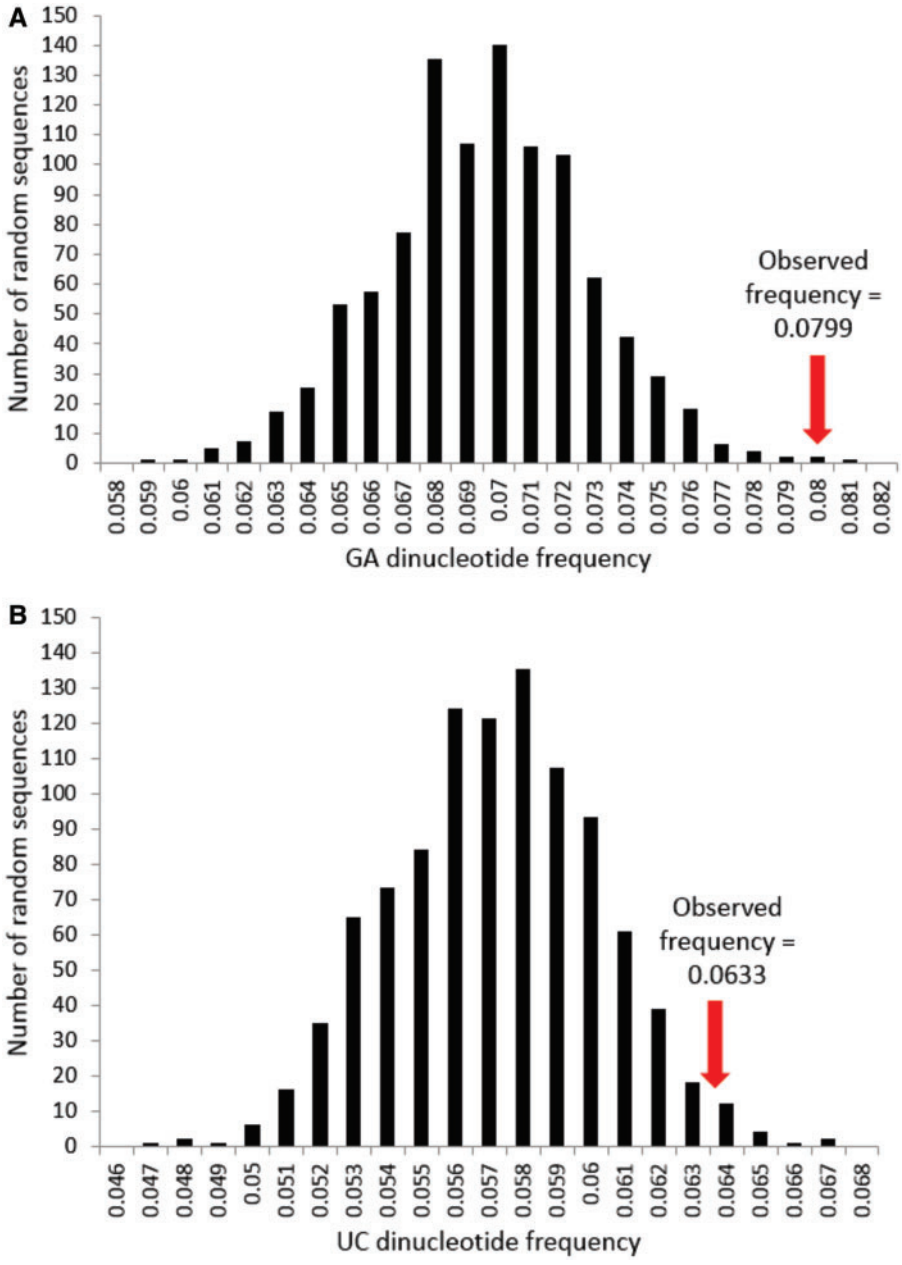

Fig. 2.—Observed GA and UC frequencies for GQ451694 as plotted against the corresponding null distributions for its 1,000 random sequences.

and Jiggins 2014), most sites appear to be highly conserved among all sequences, with only 595 out of 5,166 sites having a polymorphism in at least one out of 114 GenBank genomic sequences. We hereafter refer to the 595 sites with at least one variant as variable and the remaining 4,571 sites as conserved (see multiple sequence alignment supplementary fig. S2, Supplementary Material online). The 595 variable sites include 101, 102, and 392 first, second, and third codon positions, respectively. Thus, third codon positions are ~3.8 times more variable than first and second sites, because of their reduced functional constraints due to the redundancy of the genetic code (Li 1997). Furthermore, of the 53 sites that were polymorphic in BC strains, 20 sites were shared with the 595 sites set. Notably, polymorphic sites were distributed with about equal frequencies among all genes as well as strains (in other words, ~1% of polymorphic sites among BC strains is similar to ~1% of sites that were polymorphic among nine US strains, with comparable gene distributions among genes [Kruskal–Wallis test, $P = 0.347$]).

Next, we asked whether conserved residues differed from variable residues with respect to susceptibility to ADAR activity. Using the consensus sequence of our sample BC7 as a

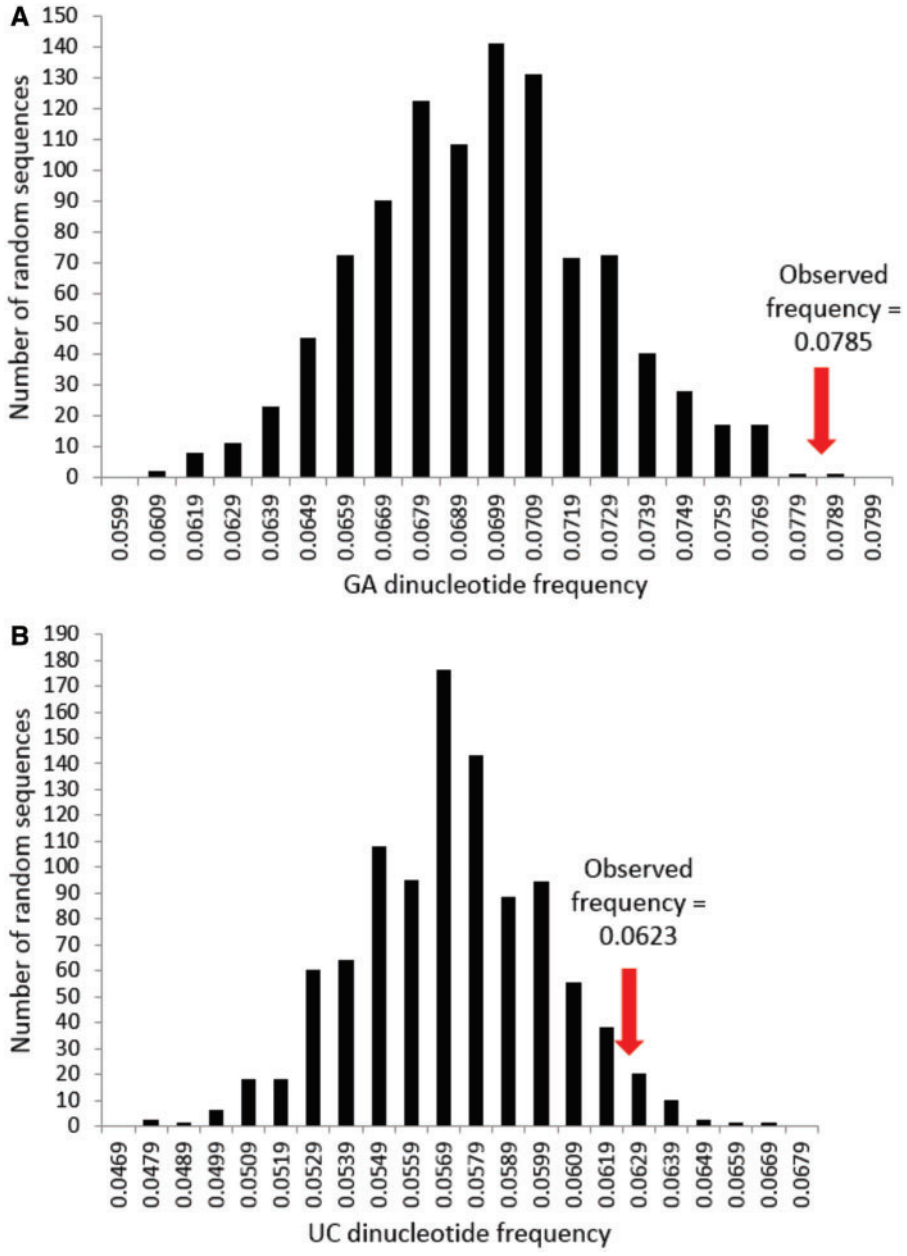

Fig. 3.—Observed GA and UC frequencies for X91062 as plotted against the corresponding null distributions for its 1,000 random sequences.

reference, we identified the 5′ neighbors of all A nucleotides and the 3′ neighbors of all U nucleotides in our alignment (sites whose 5′ or 3′ neighbors could not be identified were excluded), for a total of 331 variable and 2,470 conserved sites that are vulnerable to ADAR (i.e., A and U bases). A 2 × 2 contingency test, contrasting resistant and susceptible ADAR sites versus conserved and variable site designations (table 1), rejects the null hypothesis that both categories of ADAR sites are equally likely to be variable or conserved (Fisher's exact test; $P = 0.0034$). Resistant ADAR sites are overrepresented in the conserved sites (27.1% vs. 19.6% in variable sites). Importantly, this finding holds true even after the 14% greater frequency of non-GA dinucleotides (i.e., AA, CA, and UA) at

Table 1

Variable Sites Across Flies Are More Likely to be Susceptible to ADAR Than Conserved Sites. ADAR Designations Are Per BC7 Consensus Nucleotides. Fisher's Exact Test, One-Tailed $P = 0.0034$

	Conserved	Variable
Susceptible ADAR	1801	266
Resistant ADAR	669	65

the second compared with the first/third codon positions is accounted for by reducing proportionally the number of variable sites for the susceptible ADAR positions (Fisher exact test, $P = 0.045$; see supplementary table S3, Supplementary Material online for details). As this correction specifically deducts from the number of variable susceptible sites, it provides for a conservative test, while accounting for the different base frequencies and the greater tendency for change at third codon positions.

We consider variable sites with A-to-G transitions to be potential ADAR-edited sites. The A-to-G transitions of genes G, M, and N are not significantly clustered according to their one-tailed permutation tests ($P > 0.420$ in every case). Conversely, these changes are significantly clustered for the adjacent genes P and X ($P = 0.022$ and 0.044, respectively). Thus, as reported by Carpenter et al. (2009), we find that the potential ADAR-edited sites of gene X are spatially clumped. In turn, we now provide evidence of such clustering in gene P as well.

We constructed a phylogenetic tree (see fig. 1; details of tree construction are presented in the Materials and Methods). Consensus sequences from four of the five flies from Georgia (BC1, BC3, BC5 and technical replicates BC2 and BC7 which are from a single fly) formed a single clade within North America, with 86% bootstrap support, consistent with the previous identification of geographic structure for viral variation (Carpenter et al. 2007; Wilfert and Jiggins 2014). However, the consensus sequence from the fifth fly (BC8) is on its own branch, outside North America altogether (79% bootstrap support). Our set of five DMelSV genomes contains 53 polymorphic sites, 20 of which had previously been observed to be variable, and 33 of which had not.

To obviate long branch attraction (Felsenstein 1978), the tree we present excludes two of the most extreme cases, one from Florida (HQ655101) and one from Kenya (HQ655096) (Wilfert and Jiggins 2014). We find that BC8 does not group with either of these other known exceptional sequences (data not shown). The significance of such long branches to DMelSV evolution remains unclear.

Within-Fly Variation

The details of DMelSV replication within an individual fly, as well as the number of virions transmitted between generations, remain unclear. We hoped to glean additional information about the evolutionary dynamics of within-host replication by examining the pattern of nucleotide variation within individual flies. Unfortunately, reads from the SOLiD™ platform are relatively short (<50 bases at the time of this experiment), making it impossible to assemble viral haplotypes. Nevertheless, it is possible to draw some conclusions by examining site-by-site variation.

Most importantly, the vast majority of reads within the majority of sites resolved to a single consensus nucleotide, with only a minor fraction of reads harboring a variant nucleotide. The median number of non-consensus nucleotides per site ranged from 5 to 43 between samples (in turn, corresponding to a range of ~0.0711 to 0.0795% of all reads at a given site), indicating that only a tiny fraction of overall reads harbored sequence variants. To address the possibility that some of the observed variants are actually sequencing artifacts, we used the upper-estimate of the SOLID™ error rate of 0.06% (Glenn 2011) to estimate the expected number of erroneous reads at each site. Then we examined the magnitude of differences between the observed and expected number of variant reads (supplementary fig. S3, Supplementary Material online). The results showed that for 65–70% of the sites, the observed number of variants exceeded the expected number of erroneous reads across all samples. This pattern was generally consistent across samples, with 75.5% of sites with the excess of variants being shared across two or more samples (results not shown). Overall, this indicates that while some of the variant reads may be attributed to sequencing errors, the majority of sites harbor minor variants in excess of what can be expected even under the upper limit of the error rate.

Upon closer examination of variants across sites, we noticed that the frequency of the second most abundant allele (i.e., the most common nonconsensus read) varied extensively across our sample, with a maximum of 20.9%; the next most abundant was 16.1% (note, the median across sites and samples was 0.0751%). We thus examined the sites with the top 1% second allele frequency, defined as the sites with the highest fraction of the second allele reads (i.e., 52 sites per sample given 5,166 sites), and noticed that these sites are often shared among flies, such that there are 120 sites rather than 260 as would have been the case if each fly had a unique set of such sites. Hereafter we refer to this set of 120 top 1% sites as the hypervariable sites. Not only do multiple flies share the same site per se, but they also often have the same major and minor alleles at a given site. 12.5% (15) of the hypervariable sites are shared by all five samples; eight additional sites are shared by four out of the five samples. The overall distribution of hypervariable sites was approximately the same between genes, approximately 2–3% in each gene.

To test whether or not the observed pattern of sharing viral hypervariable sites among hosts was due to chance, we randomized the 52 top 1% site assignments within each sample, creating 1,000 pseudosamples, each 5,166 residues long. The randomization results revealed that at most three samples should share a single top 1% site designation, for ~0.02% of sites. In other words, one site out of 5,166 is expected to be shared among three samples by chance alone, and about 2% of sites (i.e., 103) are expected to share the top 1% designation among any two samples. Zero sites are expected to be shared by four or five samples, in contrast to the 23 sites observed in our dataset.

Resistant ADAR sites were less common in hypervariable sites relative to susceptible ADAR sites, significantly less so

Table 2
Resistant ADAR Sites Are Less Common in Hypervariable Sites Than Susceptible ADAR Sites. ADAR Designations Are Per BC7 Consensus Nucleotides. One-Tailed Fisher's Exact Test, $P = 0.0285$

	Hypervariable (top 1%)	Remaining Sites (99%)
Susceptible ADAR	31	2036
Resistant ADAR	4	730

Table 3
Changes That Were Consistent With ADAR Were Significantly More Common at Sites Shared by Fewer Flies Than at Sites Shared by More Flies, Regardless of How Ambivalent Sites Were Considered (See Text for Details)

	ADAR Consistent	ADAR Inconsistent	Ambivalent
Shared by four or five flies	7	6	1
Shared by at most two flies	30	3	19

than the remainder of the genome (11.4% vs. 26.4% at the remaining sites; table 2). There was no difference in the relative abundance of resistant ADAR sites between hypervariable sites shared by four or five flies (2 resistant, 12 susceptible), versus sites shared by one or two flies (11 resistant, 41 susceptible; $P = 0.57$). Such an abundance of strong sites might suggest that hypervariable sites are the result of or at least dominated by ADAR activity. Accordingly, we evaluated whether or not the state of the second allele was consistent with the activity of ADAR (i.e., if the majority allele were A or U, the second allele would be G or C, respectively). This exercise was complicated in that the first or second allele was not always the same between flies. If in at least one of the five flies the change from first to second allele was consistent with ADAR activity, we classified the site as "ambivalent" with respect to ADAR. Changes that were consistent with ADAR were significantly more common at sites shared by fewer flies than at sites shared by more flies (see table 3; $P = 0.0043$ omitting ambivalent sites; $P = 0.0003$ if ambivalent sites are lumped with ADAR-consistent changes; $P = 0.0008$ if ambivalent sites are considered as a separate category). Thus, while ADAR activity may well be driving less common within-fly variants, the widely shared variants cannot be explained by ADAR activity. Moreover, hypervariable sites that were shared by four or five flies (i.e., sites that are consistently variable within the fly) were less likely to be variable among the 114 GenBank strains (i.e., variable in the multiple sequence alignment) than those shared by only one or two flies (shared by four or five: four variable, 19 conserved; unique or shared by two: 34 variable, 46 conserved; $P = 0.028$).

Discussion

A number of different mutational and selective forces drive the molecular evolution of RNA and other biological sequences (Li 1997). Our study now provides new RNA sequence evidence for ADAR editing as being among the multiple mutation processes that underlies the molecular evolution of DMelSV. In particular, we find that 1) resistant ADAR sites are more highly conserved than are susceptible, non-GA positions (table 1 and supplementary table S3, Supplementary Material online); 2) that the positive strand is strongly enriched for weak GA dinucleotides (figs. 2 and 3); and 3) that A-to-G transitions are spatially clustered on the positive strand for gene X, consistent with prior findings (Carpenter et al. 2009), and our results also identified A-to-G spatial clustering on the positive strand for gene P.

Carpenter et al. (2009) suggested that the double-stranded targets upon which ADAR acts could either be the duplex stems created by the secondary structures of single-stranded RNA or the transient double-stranded intermediates in replication and transcription. Our findings of a strong GA overrepresentation, which is unmatched by a similar level of UC enrichment (figs. 2 and 3), is suggestive of ADAR acting to a greater extent on the positive strand. Specifically, enrichment of both GA and UC would be expected if stems were the target of ADAR, because of the functional constraint to maintain the internal base pairing of this secondary structure. In turn, enrichment of both GA and UC would also be expected if both the positive and negative strands of the replication/transcription intermediates were equally operated on by ADAR. Our finding of only a strong GA enrichment may be due to ADAR operating preferentially on the positive antigenome (i.e., ADAR editing may be biased against the negative genome). Indeed, the ADAR mutations identified by Carpenter et al. (2009) were all on the positive strand. However, a mechanism that would bias ADAR activity to the positive strand is not currently known.

Alternatively, we hypothesize that the strong GA enrichment may be because of greater functional constraint and therefore greater selection on either the positive antigenomes, the mRNA transcripts, or both (fig. 4). Once a rhabdovirus successfully infects a cell (fig. 4), the protein-coated negative-sense genome enters and primary transcription begins immediately, using an RNA-dependent RNA polymerase carried by the virion (Lyles et al. 2013). Since the genome is negative sense, transcripts are positive, and may be translated by host machinery directly from the infecting genome. Following sufficient translation of the N protein, nascent, positive RNA can be encapsidated by N proteins. Encapsidated positive RNA is the signal for the switch from transcription of individual mRNAs, to creating full-length positive sense antigenomes to serve as templates for the negative genomes. Though some of these new negative genomes give rise to new virions, many are instead used as templates for a burst of secondary transcription, which produces far more mRNA than primary transcription. Thus, the initial positive sense antigenomes are the core of the rhabdovirus life cycle, and their fidelity is critical to viral fitness. It is this central importance that

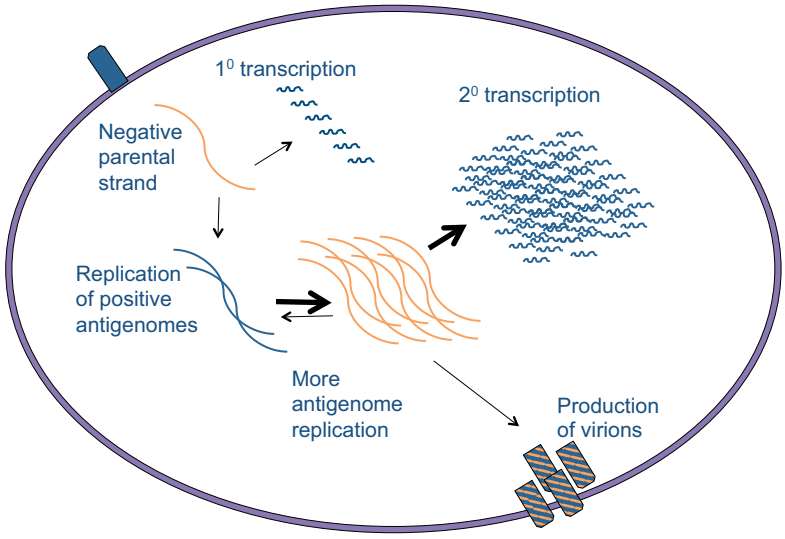

Fig. 4.—Schematic depiction of a DMelSV lifecycle. Larger arrows indicate higher activity or amounts, for example more negative, genomic copies are made from the positive antigenomes and there is little reverse synthesis of positive antigenomes from these new genomes; similarly, more of the new negative genomes are used for production of transcripts than virion production. Note the central role played by the initial full-length positive antigenomes. Any editing of these antigenomes would affect not only the genomes packaged in progeny virions, but all secondary transcription as well.

we hypothesize leads to greater functional constraint, and thereby, selection for weak GA dinucleotides on the positive (but not negative) strand (figs. 2 and 3). This increased selection may also explain why ADAR edits are more evident on the positive strand (Carpenter et al. 2009); that is, as fewer mutations are tolerated, rare ADAR changes become more obvious on the antigenome.

Selection for ADAR resistance provides a straightforward explanation for the enrichment of GA and (possibly) UC dinucleotides. However, the genome is also enriched for CA and UG dinucleotides, which are not thought to be as resistant to ADAR editing as GA or UC. Of course, there are forces other than ADAR likely at play in shaping genome evolution, although their relative contributions to the mutation process remain to be determined. Indeed, these forces may work independently of, in concert with, or in opposition to ADAR. UA and CG dinucleotides, for example, are underrepresented across most taxa (Karlin and Burge 1995). One compelling explanation for underrepresentation of UA and CG dinucleotides across the tree of life, including in DMelSV, is mutational bias, specifically due to transitions, which are the most common mutations (Li 1997). Individual transition mutations will convert CG to UG or CA (transitions of first and second bases, respectively), or UA to CA and UG (transitions of first and second bases, respectively). CA and UG are overrepresented in DMelSV (and other genomes), despite their susceptibility to ADAR attack (see supplementary fig. S4, Supplementary Material online). This overrepresentation might be balancing the underrepresentation of the unmutated, "parental" CG and UA dinucleotides.

In contrast to the spatial clustering of A-to-G transitions in the adjacent genes P and X, no such clumping is found in G, M, and N. To explain this gene-to-gene variation, one obvious possibility is that ADAR is operating to a lesser extent (or even not at all) on the latter three genes. However, given that molecular evolution is mediated by many different factors, which may act simultaneously and/or in opposite directions (Li 1997), we hypothesize instead that other mutational processes are introducing A-to-G transitions in genes G, M, and N in a manner that is more scattered than those caused by ADAR. The scattering of these other transitions would obscure the signal of spatially clustered ADAR edits, thereby reducing the power of the permutation tests.

While there is evidence of selection to resist ADAR changes, we also wondered whether or not ADAR editing is a major source of among-host variation. We found that A residues in conserved sites had an overabundance of the resistant ADAR context (i.e., GA dinucleotides) compared with A residues in variable sites; while variable sites have a smaller fraction of susceptible ADAR sites compared with conserved sites (see table 1 and supplementary table S3, Supplementary Material online). One possible explanation for this observation is that conserved sites are under greater evolutionary constraint than variable sites, and thus have experienced stronger selection. However, it is also possible that the variable sites are variable *because* they are strong ADAR sites, rather than because they are subject to lower selective constraint. Since ADAR results in a predictable substitution of As with Gs (or Us with Cs for targets on the complementary strand), we are able to identify residues with alternate alleles that are likely to be the result of

ADAR mutation. Of the variable sites, variants at 315 out of 335 (~94%) are consistent with ADAR activity, that is have the expected A to G (or U to C) transition as the second variant at a site, while the remaining 20 are not (~6%). It is worth pointing out that multiple mutations per site may have occurred. From these data alone, we cannot distinguish between the hypotheses of lack of constraint and ADAR-driven variability, but certainly these data are consistent with pervasive ADAR activity.

Some insight with respect to evolutionarily unconstrained sites versus hypermutation due to ADAR may be gained from the hypervariable sites and the pattern of sharing of these sites among more or fewer flies. Hypervariable sites which are present in only one or two flies had major and minor alleles that were significantly more likely to be consistent with ADAR activity than sites present in multiple flies. In other words, hypervariable sites that occur with low frequency (i.e., present only within a single or at most two sample(s)) appeared to represent deleterious variation, similar to the expectation that low-frequency segregating sites represent weakly deleterious mutations (e.g., Fay et al. 2001). The rare sites were also more likely to be variable among the 114 GenBank sequences, than the more commonly shared sites. This further supports the interpretation that mildly deleterious mutations are more likely to persist if they occur in less constrained sites than highly conserved ones.

What then is the explanation for the existence of hypervariable sites? Are they merely an experimental artifact, perhaps due to enzyme error? Genomic segments from each fly were amplified and reverse transcribed independently. It seems unlikely that a PCR or RT error should occur at the same base (and result in the same change) five times independently. Furthermore, our randomization test tells us that these shared sites are not the result of chance. One possibility is that these commonly shared sites are the signature of multiple DMelSV infections within a single fly. Multiple infections have been observed previously for DMelSV (Brun 1963; Seecof 1966). Perhaps the most intriguing link between their data and ours is that they observed that the relative proportions of the viruses (determined by plaque size) were heritable across generations. While we cannot precisely determine proportions from counts of reads, the identity of major and minor alleles should reflect relative abundance of virus to some extent. That not only the sites themselves but the specific alleles and their classification as major and minor are identical across multiple flies from the wild north Georgia population is strongly reminiscent of multiple infections seen in the laboratory. However, the short reads of SOLiD™ technology make it impossible to determine whether or not the sites are part of a single haplotype, and thus are likely to be the result of coinfection. Further studies with long read technology will be necessary to draw robust conclusions.

We observe evidence of ADAR at every level of scrutiny, from among flies worldwide, to within a population, to within flies. Our results suggest that ADAR editing is deleterious to DMelSV because: 1) susceptible ADAR sites are underrepresented in the viral genome, 2) ADAR mutations tend to occur at sites that are variable across multiple sequences, and 3) the same within-fly ADAR mutations are less likely to be shared among individual flies. ADAR activity may either have arisen or be conserved due to antiviral function, despite its overall largely nonadaptive nature as was recently shown in human and mouse genomes (Xu and Zhang 2014, 2015). It remains to be seen whether or not ADAR is active against other RNA viruses in *Drosophila melanogaster*, and whether or not the activity is deleterious enough to result in a selective response on the part of the viral genome.

Supplementary Material

Supplementary tables S1–S4 and figures S1–S4 are available at *Genome Biology* and *Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the National Institutes of Health (grant numbers GM083192 (M.W.) and GM86782-01A1 (H.P.)). L.F.M, W.G.F, and M.L.W were also supported by the UF Emerging Pathogens Institute/Interdisciplinary Center for Biotechnology Research Innovative Projects Initiative.

Literature Cited

Brun G. 1963. Édude d'une association du virus et son hôte la Drosophilé: l'état stabilisé. Thèse Paris XI, Orsay, 1963.

Brun G, Plus N. 1980 The viruses of *Drosophila*. In: Ashburner M, Wright T. R. F, editors. The genetics and biology of Drosophila. London: Academic Press. p. 625–702.

Brusini J, et al. 2013. Virulence evolution in a host-parasite system in the absence of viral evolution. Evol Ecol Res. 15:883–901.

Carpenter JA, Keegan LP, Wilfert L, O'Connell MA, Jiggins FM. 2009. Evidence for ADAR-induced hypermutation of the Drosophila sigma virus (Rhabdoviridae). BMC Genet. 10:75.

Carpenter JA, Obbard DJ, Maside X, Jiggins FM. 2007. The recent spread of a vertically transmitted virus through populations of *Drosophila melanogaster*. Mol Ecol. 16:3947–3954. doi: 10.1111/j.1365-294X.2007.03460.x

Chare ER, Gould EA, Holmes EC. 2003. Phylogenetic analysis reveals a low rate of homologus recombination in negative-sense RNA viruses. J Gen Virol. 84:2691–2703.

Chen J-Y, et al. 2014. RNA editome in rhesus macaque shaped by purifying selection. Plos Genet. 10(4):e1004274. doi: 10.1371/journal.pgen.1004274.

Domingo E, Holland JJ. 1997. RNA virus mutations and fitness for survival. Annu Rev Microbiol. 51:151–178. doi: 10.1146/annurev.micro.51.1.151

Drake JW, Holland JJ. 1999. Mutation rates among RNA viruses. Proc Natl Acad Sci U S A. 96:13910–13913. doi: 10.1073/pnas.96.24.13910.

Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet. 9:267–276.

Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. Genetics 158(3):1227–1234.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool. 27:401–410.
Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791. doi: 10.2307/2408678
Fleuriet A. 1996. Polymorphism of the *Drosophila melanogaster*—sigma virus system. J Evol Biol. 9:471–484.
Glenn TC. 2011. Field guide to next generation DNA sequencers. Mol Ecol Res. 11:759–769.
Guindon S, Delsuc F, Dufayard JF, Gascuel O. 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol Biol. 537: 113–137.
Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–321. doi: 10.1093/sysbio/syq010 syq010 [pii]
Hogenhout SA, Redinbaugh MG, Ammar ED. 2003. Plant and animal rhabdovirus host range: a bug's view. Trends Microbiol. 11:264–271.
Holmes EC. 2003. Molecular clocks and the puzzle of RNA virus origins. J Virol 77:3893–3897. doi: 10.1128/jvi.77.7.3893-3897.2003
Holmes EC. 2009. The evolution and emergence of RNA viruses. New York (NY): Oxford University Press.
Jenkins GM, Rambaut A, Pybus OG, Holmes EC. 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J Mol Evol. 54:156–165. doi: 10.1007/s00239-001-0064-3
Karlin S, Burge C. 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 11:283–290. doi: S0168952500890769 [pii]
Keegan LP, Gallo A, O'Connell MA. 2001. The many roles of an RNA editor. Nat Rev Genet. 2:869–878. doi: 10.1038/35098584
Lehmann KA, Bass BL. 2000. Double-stranded RNA adenosine deaminases ADAR1 and ADAR2 have overlapping specificities. Biochemistry 39:12875–12884.
Li W-H. 1997. Molecular evolution. Sunderland, MA: Sinauer Associates, Inc.
Liu H, et al. 2014. Functional Impact of RNA editing and ADARs on regulation of gene expression: perspectives from deep sequencing studies. Cell Biosci 4:44:doi: 10.1186/2045-3701-4-
Longdon B, Jiggins FM. 2010. Quick guide. Paternally transmitted parasites. Curr Biol. 20:R695–R696.
Longdon B, Obbard DJ, Jiggins FM. 2010. Sigma viruses from three species of Drosophila form a major new clade in the rhabdovirus phylogeny. Proc R Soc Lond B: Biol Sci. 270:35–44.
Longdon B, Wilfert L, Jiggins FM. 2012. The sigma viruses of *Drosophila*. In: Dietzgen RG, Kuzmin IV, editors. Rhabdoviruses: molecular taxonomy, evolution, genomics, ecology, host-vector interactions, Cytopathology and Control. Norfolk, UK: Caister Academic Press. p. 117–132.
Lyles DS, Kuzmin IV, Rupprecht CE. 2013. Rhabdoviridae. In: DM Knipe, PM Howley, editors. Fields virology, 6th ed. Philadelphia, PA: Lippincott Williams & Wilkins. p. 885–922.
Medzhitov R, Schneider DS, Soares MP. 2012. Disease tolerance as a defense strategy. Science 335(6071):936–941.
Meyers LA, Levin BR, Richardson AR, Stojiljkovic I. 2003. Epidemiology, hypermutation, within-host evolution and the virulence of *Neisseria meningitidis*. Proc R Soc B Biol Sci. 270:1667–1677. doi: 10.1098/rspb.2003.2416
Mueller S, Papamichail D, Coleman JR, Skiena S, Wimmer E. 2006. Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. J Virol. 80:9687–9696. doi: 10.1128/jvi.00738-06
Palladino MJ, Keegan LP, O'Connell MA, Reenan RA. 2000. dADAR, a *Drosophila* double-stranded RNA-specific adenosine deaminase, is highly developmentally regulated and is itself a target for RNA editing. RNA 6:1004–1018. doi: 10.1017/s1355838200000248
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16:276–277. doi: 10.1016/s0168-9525(00)02024-2
Roossinck MJ. 2010. Lifestyles of plant viruses. Philos Trans R Soc B Biol Sci. 365:1899–1905. doi: 10.1098/rstb.2010.0057
Seecof RL. 1966. Sigma virus content and hereditary transmission in *Drosophila melanogaster*. Virology 29:1–7.
Tamura K, Nei M, Kumar S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci U S A. 101:11030–11035.
Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol. 30:2725–2729. doi: 10.1093/molbev/mst197 mst197 [pii]
Teninges D, Bras F, Dezélée S. 1993. Genome organization of the sigma rhabdovirus: six genes and a gene overlap. Virology 193(2):1018–1023.
Tsai CW, McGraw EA, Ammar ED, Dietzgen RG, Hogenhout SA. 2008. *Drosophila melanogaster* mounts a unique immune response to the Rhabdovirus sigma virus. Appl Environ Microbiol. 74(10):3251–3256.
Walker PJ, Dietzgen RG, Joubert DA, Blasdell KR. 2011. Rhabdovirus accessory genes. Virus Res. 162(1-2):110–125.
Wayne ML, et al. 2011. The prevalence and persistence of sigma virus, a biparentally transmitted parasite of *Drosophila melanogaster*. Evol Ecol Res. 13:323–345.
Wilfert L, Jiggins FM. 2014. Flies on the move: an inherited virus mirrors *Drosophila melanogaster's* elusive ecology and demography. Mol Ecol. 23:2093–2104. doi: 10.1111/mec.12709
Xu GX, Zhang JZ. 2014. Human coding RNA editing is generally nonadaptive. Proc Natl Acad Sci U S A. 111:3769–3774. doi: 10.1073/pnas.1321745111
Xu GX, Zhang JZ. 2015. In search of beneficial coding RNA editing. Mol Biol Evol. 32:536–541. doi: 10.1093/molbev/msu314
Yampolsky LY, Webb CT, Shabalina SA, Kondrashov AS. 1999. Rapid accumulation of a vertically transmitted parasite triggered by relaxation of natural selection among hosts. Evol Ecol Res. 1:581–589.

Associate editor: Dennis Lavrov