

Evolutionary Genomics of Host Adaptation in Vesicular Stomatitis Virus

Susanna K. Remold,* Andrew Rambaut,† and Paul E. Turner‡

*Department of Biology, University of Louisville; †Department of Zoology, University of Oxford, Oxford, UK; and ‡Department of Ecology and Evolutionary Biology, Yale University

Populations experiencing similar selection pressures can sometimes diverge in the genetic architectures underlying evolved complex traits. We used RNA virus populations of large size and high mutation rate to study the impact of historical environment on genome evolution, thus increasing our ability to detect repeatable patterns in the evolution of genetic architecture. Experimental vesicular stomatitis virus populations were evolved on HeLa cells, on MDCK cells, or on alternating hosts. Turner and Elena (2000. Cost of host radiation in an RNA virus. *Genetics*. 156:1465–1470.) previously showed that virus populations evolved in single-host environments achieved high fitness on their selected hosts but failed to increase in fitness relative to their ancestor on the unselected host and that alternating-host-evolved populations had high fitness on both hosts. Here we determined the complete consensus sequence for each evolved population after 95 generations to gauge whether the parallel phenotypic changes were associated with parallel genomic changes. We also analyzed the patterns of allele substitutions to discern whether differences in fitness across hosts arose through true pleiotropy or the presence of not only a mutation that is beneficial in both hosts but also 1 or more mutations at other loci that are costly in the unselected environment (mutation accumulation [MA]). We found that ecological history may influence to what extent pleiotropy and MA contribute to fitness asymmetries across environments. We discuss the degree to which current genetic architecture is expected to constrain future evolution of complex traits, such as host use by RNA viruses.

Introduction

What shapes the genetic architecture of a complex trait? Random events in evolution can cause populations experiencing the same selection to diverge quickly in a phenotypic trait's genetic architecture (Remold and Lenski 2004). Nevertheless, genetic architecture sometimes evolves in repeatable, predictable patterns. To better understand the evolution of complex phenotypes, in this study we have investigated how the evolution of genetic architecture mediates the relationship between current and future niches.

The genetic architecture of a phenotypic trait consists of the number of loci responsible for the trait, their genomic positions, their direct effects on the trait, and their contributions to genetic variance in the trait. Genetic architecture also includes gene interactions: epistasis among the loci, the dependence of these gene interactions on the environment, and the pleiotropic effects of the genes on other traits (Zeng et al. 1999). Organisms can achieve the same phenotype through quite different underlying genetic architectures, and in nature some of the diversity in genetic architecture is correlated with environmental variability (Lynch and Walsh 1998; Zeng 2005). Differences in genetic architecture, especially dissimilar contributions of individual loci to genetic variance and to interactions, can cause populations with identical current trait means to differ in the alternative phenotypes easily achieved through mutation. As a result, when the trait in question determines niche, populations currently inhabiting the same niche may differ substantially in their potential for niche expansion.

Biological systems involving high mutation rates and large population sizes are most likely to show repeatable patterns of evolution of genetic architecture. In such populations, evolution of complex traits is neither limited by access to mutations nor strongly affected by genetic drift.

Key words: host adaptation, genetic architecture, epistasis, VSV, antagonistic pleiotropy, mutation accumulation.

E-mail: susanna.remold@louisville.edu

Mol. Biol. Evol. 25(6):1138–1147. 2008

doi:10.1093/molbev/msn059

Advance Access publication March 18, 2008

We investigated the influence of historical environment on the evolution of genetic architecture in vesicular stomatitis virus (VSV) populations evolved in replicated tissue culture regimes differing in host availability (Turner and Elena 2000). VSV is a single-stranded negative-sense RNA virus, with a high mutation rate— 10^{-4} mutations per genome replication (Drake and Holland 1999). Experimental evolution occurred through serial transfer (Turner and Elena 2000), where populations expanded from 200 founding virions to approximately 5×10^9 viral progeny, resulting in an effective population size of approximately 2,500 viruses (equal to the harmonic mean) (Hartl and Clark 1989), which is sufficiently large to minimize the effects of genetic drift.

Parallel Host-Use Evolution in RNA Virus Populations

The 12 study populations differed in their historical exposure to hosts and had been divided equally across 3 environments containing tissue culture monolayers of 2 types: human cervical epithelial cells (HeLa) and canine kidney epithelial cells (MDCK). Four populations were evolved solely on HeLa, 4 on MDCK, and 4 on alternating passages on the 2 hosts (Turner and Elena 2000). After 100 generations (25 passages) of virus evolution, fitness relative to the common ancestor when grown on HeLa and on MDCK was measured. Sequence data obtained in this study revealed that some of these fitness values are probably inaccurate (see Materials and Methods, below). However, this discussion and the interpretations of the sequence data presented here are based on overall patterns of increase or decrease in fitness that are likely robust to this inaccuracy. In particular, Turner and Elena (2000) found that the 2 sets of populations propagated on a single host increased in fitness relative to the ancestor on the selected host but showed no such increase in fitness on the unselected host (fig. 1). In contrast, the populations evolved on alternating hosts experienced increased fitness in both habitats (fig. 1). Here, we compare the population consensus RNA sequences to examine whether this broad pattern of parallel phenotypic changes is due to parallel genomic changes.

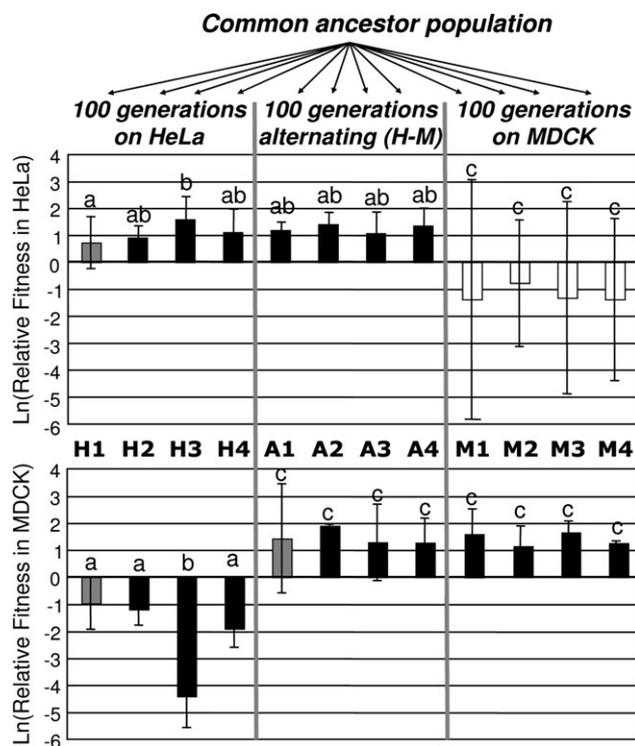


FIG. 1.—Mean and 95% confidence intervals of natural log of estimated fitness of HeLa-evolved, alternating-host-evolved, and MDCK-evolved populations on HeLa and MDCK, redrawn from Turner and Elena (2000). Bars sharing letters above do not differ significantly. Black bars indicate significant difference from the ancestor at $P < 0.05$ by 1-tailed t -test, gray bar indicates $0.05 < P < 0.1$, and white bar indicates $P > 0.1$. Design of evolution is shown at top.

Alternative Genetic Architectures of Phenotypic Patterns

We also compare observed patterns of allele substitutions in the evolved consensus populations with the patterns expected if fitness asymmetry across hosts is caused by true genetic pleiotropy (called antagonistic pleiotropy, AP, especially when the outcome is phenotypic tradeoff) versus the fixation of a mutation with a cost in an unselected environment (sometimes called mutation accumulation, MA). AP occurs when an allele becomes fixed or very common due to its selective advantage to one trait (e.g., fitness in one host) but has a deleterious pleiotropic effect on another trait (e.g., fitness in the other host). More generally, such pleiotropy results in asymmetric effects of a mutation across traits. In contrast, under MA phenotypic asymmetry can occur without any underlying genetic pleiotropy. Under MA, the mutation causing the beneficial effect in the selected host causes a similar benefit in the unselected host. However, a mutation at a different locus that is costly in the unselected host becomes fixed during the same evolutionary period through genetic drift or hitchhiking. As a result, the evolved population shows more phenotypic asymmetry than is actually caused by the beneficial mutation itself. These 2 causes of fitness asymmetry across hosts differ in their consequences for the evolutionary flexibility of the population in the face of new environmental challenges.

Materials and Methods

VSV Populations

We used 12 of the 16 populations whose experimental evolution was described by Turner and Elena (2000). Briefly, all populations were founded using a single common ancestral VSV population, MARM U. Each population was passed independently (every 2 days) in 1 of 3 host-use regimes: transfer on HeLa cells only, transfer on MDCK cells only, and alternating transfers between these 2 cell types. The final host was HeLa for the sequenced populations that were selected on alternating hosts. For these evolutions, American Type Culture Collection was the source for HeLa cells (CCL-2), baby hamster kidney (BHK) cells were from the laboratory of E. Domingo (University of Madrid), and MDCK cells were obtained from the European Collection of Cell Cultures (85011435). These 12 populations (4 replicates \times 3 treatments) were sequenced at 24 passages (roughly 95 generations). Two populations that were founded with the ancestor and that underwent a single passage on BHK (the typical laboratory host) were used to determine the ancestor sequence because the true ancestor was lost in a laboratory flood. Because a sequence comparison between these 2 BHK populations yielded no conflicts, their consensus sequence (based on 4- to 10-fold coverage) was inferred to be the ancestral sequence.

We found that the ancestor was monomorphic at all sites except for one: repeated population level sequences indicated a polymorphism at base 3853 (C and A). This locus in the glycoprotein gene (G) has been demonstrated to be responsible for resistance to the monoclonal antibody I1 (I1-mAb) in MARM U (derived from MARM C) (Holland et al. 1991; Novella and Ebendick-Corpus 2004). Other alleles also confer resistance to I1-mAb (Holland et al. 1991; Novella and Ebendick-Corpus 2004). Deviation from 100% resistance in the ancestor populations will affect the fitness estimates of evolved populations on a given host by a constant. Sensitive genotypes in the evolved populations will lower estimates of population fitness, potentially introducing errors in comparisons among populations.

We used 2 techniques to estimate the percent of each population that is I1-mAb sensitive: plaque assay with and without I1-mAb, the same technique used in the fitness assays, and immunofluorescent assay (IFA), in which all plaques are viewed by fluorescent microscopy, allowing direct estimation of the resistant and sensitive fractions of the population. With one exception (M4, 29 plaques), in all populations at least 64 plaques were screened for sensitivity to I1-mAb. The median number of plaques screened by IFA per population was 148. All the single-host-evolved populations and population A4 are entirely I1-mAb resistant by both methods (mixed linear models $P < 0.05$ for all population-method combinations). However, populations A1–A3 are mostly I1-mAb susceptible (A1 and A2 less than 0.25% resistant; A3 3.5% resistant as estimated by IFA). We conclude that the universally high fitness estimates obtained for the alternating-host-evolved populations reflect only the resistant portion of the populations and are probably lower than the true values. However, the error introduced by the marker polymorphism does not affect the qualitative differences between the 3 sets of populations.

The 2 populations used as proxies for the ancestral population showed no significant difference in pfu detected when plaqued with and without I1-mAb, but IFA of approximately 150 plaques from each indicates that the ancestral population was only approximately 92% resistant. The fact that population sequences repeatedly detected both alleles at the ancestral marker locus shows that a minority allele can be clearly detected with our sequencing technique at a frequency as low as 0.08.

RNA Isolation, Reverse Transcriptase–Polymerase Chain Reaction, and Sequencing of Genomes

Genomic RNA was isolated from all 14 populations (12 evolved and 2 ancestral) using QIAamp MinElute virus spin kit (Qiagen, Hilden, Germany; using either protease or proteinase-K to free RNA from the capsid, and cDNA was generated by reverse transcription with Superscript II (Invitrogen, Carlsbad, CA; using random hexamer primers (1–3 RNA preparations per population). The majority of the genome was amplified via polymerase chain reaction (PCR) using 7 primer pairs, generating overlapping PCR fragments ranging between 2 and 3 kb in length. These were sequenced (dye terminator sequencing, external facilities) to generate 2- to 3-fold coverage of the genomes, such that the replicate sequences were obtained from at least 2 independent PCRs. The trailer region (at the 3' end of the positive sense strand) was sequenced using 5' random amplification of cDNA ends (RACE) kit from Invitrogen, and the leader region was sequenced by the addition of a poly(A) tail to the 3' end of the genomic RNA using poly(A) polymerase from USB Corporation (USB, Cleveland, OH), followed by 3' RACE (Invitrogen).

Because the study's main hypothesis (regarding the relationship between genome sequence and population level fitness in host environments) does not involve within-population estimates of variability, we opted to directly sequence PCR amplified virus population cDNAs rather than to sequence multiple isolated clones from the population. Our sequencing approach detects the dominant (consensus or average) nucleotide at each base position. Reported changes relative to the ancestor must therefore be considered a "minimum set," as new alleles at low frequency will not be detected by this method. The independent estimation of the frequency of the wild-type allele at the marker locus in the 2 populations standing in for the ancestor allows calibration of the sensitivity of our approach for the detection of minority alleles. Whereas both peaks at that locus are clearly visible in 9 out of 10 of our direct population sequencing runs, to detect an allele present in 8% in the population would require sequencing 13 clones on average.

Genome Assembly and Alignment

All sequences were reviewed by hand, and the genomes were assembled using Sequencher (Gene Codes, Ann Arbor, MI). Where multiple, otherwise clean sequencing reads yielded strong secondary peaks at the same locus, the population was concluded to be polymorphic at that locus. With one exception, these polymorphisms involved a new allele present in addition to the ancestral allele.

The exception is base 3853 in population M1, in which the ancestral polymorphism is still present. Consensus sequences from the 12 evolved populations were aligned with the ancestral sequence using Sequencher and ClustalX for Mac OS X. Because the frequency of polymorphism cannot be determined from population level sequences and because other polymorphisms for which the minority genotype did not reach the level of detection by population-reverse transcriptase–polymerase chain reaction undoubtedly exist, for the purposes of this study the data were then reduced to "new allele detected" (and the identity of the allele) versus "only ancestral allele detected."

Statistical Analyses

We used mixed linear models with the competition host (current environment), the evolutionary host regime (ecological history), and population as fixed effects, and experimental replicate as a random effect, to model variability in relative fitness in the evolved populations as measured by Turner and Elena (2000). Within this model, we conducted all pairwise comparisons of evolved populations using Tukey's adjustment for multiple comparisons. Chi square tests and Fisher's exact tests were also used to test for deviations from random in the distributions of mutations. One sample *t*-tests were used to identify fitnesses significantly different from the ancestor.

Finally, to investigate the probability that mutations would occur repeatedly at the same site by chance, we created a simple Monte Carlo simulation. This program models the 12 populations of viruses, repeatedly applying mutations at the observed average per-site rate, generating the distribution of co-occurring mutations at particular sites within or between groups of populations. This program is available from the authors on request.

Results

Evolution of Fitness in 2 Hosts

Using a mixed linear model (SAS Institute Inc. 1999), we characterized how the differing evolutionary histories influenced phenotypic changes within a treatment. Linear contrasts within the mixed model and corrected for multiple comparisons identified 4 host-use phenotypes among the 12 populations (fig. 1). Populations H1, H2, and H4, constituting the first phenotypic group, do not differ significantly from one another in their fitness on either host. H3 has significantly lower fitness on MDCK than any other HeLa-evolved population and significantly higher fitness on HeLa than H1 (but not H2 and H4) (fig. 1). Nevertheless, given the qualitative phenotypic similarity between H3 and the other HeLa-evolved populations, we will hereafter consider H1–H4 to be a single biologically relevant group. There were no significant differences among either A1 through A4 or M1 through M4 (fig. 1). These 3 groups are the same basic groupings identified by Turner and Elena (2000). Within these groups, one sample *t*-tests detected significantly or marginally significantly increased fitness relative to the ancestor of alternating-host-evolved populations on both hosts and of the single-host-evolved populations on their selected host (fig. 1). Tests of HeLa-evolved populations on MDCK revealed significantly or marginally significantly decreased

fitness relative to the ancestor (fig. 1). Although as a group, MDCK-evolved populations show significantly lower fitness than the ancestor on HeLa, at the level of the population this change is not statistically significant due to large variances in the fitness estimates (fig. 1).

Shared New Alleles Are Likely Adaptive

We identified 3–9 new alleles per evolved population consensus sequence, for a total of 67 new alleles occurring at 34 loci in the 12 evolved populations (fig. 2). In all, 52 of these 67 observed mutations (78%) were not unique, and 19 of the 34 polymorphic loci (55%) were affected in multiple independent populations, resulting in a total of 20 non-unique new alleles. We used a Monte Carlo approach to compare these patterns with a null model in which mutations arise and become dominant with equal probability across the ~11-kb genome at a rate based on the total number of observed mutations (5.23×10^{-4} per site per population). The probability of any 2 populations with the same selective environment sharing a mutation at 1 or more sites is 0.02 and of any 3 populations is <0.0001 (only 3 out of 100,000 simulations produced a mutation shared in 3 populations). The probability of 2 of the selective environments having 2 or more populations with co-occurring mutations is therefore much less than 0.0001 and, indeed, never occurred in our simulations. Thus, even a single pair of shared substitutions found among populations of shared selection history indicates an adaptive mutation. Both shared silent and replacement substitutions occurred and are probably adaptive, and there is only a marginally significant excess of replacement substitutions relative to silent substitutions among shared mutations when compared with unique mutations (Fisher's exact test, $P = 0.07$).

Evolution of Monoclonal Antibody Resistance Loci

In the population sequences of 5 out of 12 populations (all 4 HeLa-evolved populations and M4, an MDCK-evolved population), the C (antibody resistant) allele is the only detectable allele at base 3853, the location of the only detected polymorphism in the ancestor (fig. 2). In addition, 2 more populations (A4 and M3) appear fixed for $G_{3846} \rightarrow A$ (fig. 2, table 1), another allele conferring resistance to I1-mAb (Holland et al. 1991). In populations A2 and A3, we detected both alleles at 3846, and M1 has the ancestral polymorphism at 3853 and both alleles at 3846 (data not shown). Interestingly, although we did not detect any known antibody resistance markers in them, the replacement G protein $Leu_{260} \rightarrow Pro$ due to $T_{3856} \rightarrow C$, which is adjacent to the locus of the dominant resistance marker in MARM U is polymorphic in A1 and is the only detectable allele at this locus in M2. A1 is partially and M2 entirely I1-mAb resistant, so this change at 3856 likely confers resistance to I1-mAb. Because all 3 evolutionary environments were antibody free, these changes must be due to selection on some other aspect of G-protein function.

Distribution of Substitutions among Ecological Histories

Populations evolved under the 3 host-use histories did not differ significantly in the number of new alleles present

(1-factor analysis of variance, $P = 0.27$). However, HeLa-evolved populations had proportionally more shared relative to unique mutations (Fisher's exact test, $P = 0.04$) and marginally significantly more replacement relative to silent mutations (Fisher's exact test, $P = 0.07$) than populations evolved in the other 2 histories.

Mutations that occurred in multiple populations fall into 4 groups with respect to their distributions across ecological histories (table 1). Six new alleles ($U_{2401} \rightarrow C$, $A_{2431} \rightarrow G$, $G_{2686} \rightarrow A$, $A_{4167} \rightarrow G$, $C_{9597} \rightarrow U$, and $G_{10048} \rightarrow U$) were detected in multiple HeLa-evolved populations but in none of the other treatments. All but the last of these are replacement substitutions. Three mutations ($C_{696} \rightarrow U$, $G_{1521} \rightarrow A$, and $U_{2937} \rightarrow G$) were found only in the populations evolved on alternating hosts; the first 2 of which are silent substitutions. Given the low probability of multiple shared mutations, this host-history-specific distribution strongly suggests that these mutations are beneficial in the particular historical niches in which they arise, even though 3 of the 9 are silent.

Ten new alleles ($U_{1927} \rightarrow G$, $C_{2151} \rightarrow U$, $A_{2413} \rightarrow G$, $A_{2431} \rightarrow C$, $A_{2736} \rightarrow G$, $C_{3153} \rightarrow A$, $A_{3838} \rightarrow U$, $G_{3846} \rightarrow A$, $U_{3856} \rightarrow C$, and $C_{4981} \rightarrow U$; all but the last replacement substitutions) were detected in 1 or more populations from both the MDCK and alternating-host histories. (Note that the new allele at base pair 2431 in these host histories is different from the one occurring in multiple HeLa-evolved populations [table 1, fig. 2].) Finally, a single new allele, $C_{4180} \rightarrow A$, was detected in both a HeLa and an alternating-host-evolved population. The absence of any shared alleles between HeLa- and MDCK-evolved populations suggests that the majority of mutations shared between alternating-host-evolved populations and single-host-evolved populations are host-use adaptations.

Distribution of Substitutions among Genes

The VSV genome contains 5 genes: the nucleoprotein (N) which encapsidates viral genomic RNA; the phosphoprotein (P); the matrix protein (M), with both structural and nonstructural functions; the glycoprotein (G) which protrudes through the envelope of mature virions; and the large polymerase protein (L) which functions in conjunction with the P protein. Table 1 shows the distribution of shared alleles across genes according to the evolutionary history in which they arose.

New alleles in the evolved populations were distributed disproportionately among these 5 genes, occurring on average in each population every 394 bases in the M gene and every 969 and 869 bases in the G and P genes. Per nucleotide, new alleles were substantially more rare in the N and L genes (per population, one every 5072 and 5842 bases, respectively) (Fisher's exact test, $P < 0.00001$). The ratio of silent to replacement substitutions also differed (Fisher's exact test, $P < 0.00001$), with replacements predominating in G and M and silent substitutions predominating in L (table 1). Thus, genes tend to have either fewer unique and silent mutations or more shared replacement mutations.

Whereas alternating-host-evolved and MDCK-evolved populations showed new alleles in all 5 genes,

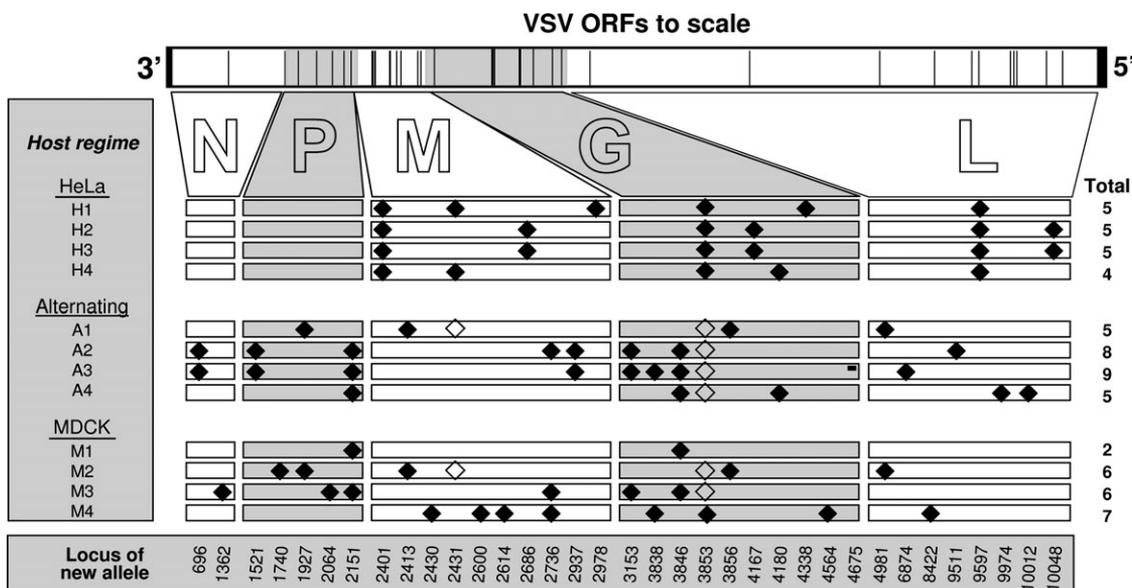


FIG. 2.—Locations of new alleles relative to the ancestral sequence in the 12 evolved populations. Ticks on bar across top indicate physical locations of mutations in the genome. Viral nucleoprotein (N), phosphoprotein (P), matrix protein (M), glycoprotein (G), and large polymerase protein (L) genes are indicated on both scaled and unscaled bars. Within the 12 lower bars, each diamond represents a new nucleotide and the dash represents a single base deletion. Genome location is shown at bottom. Populations sharing symbols in the same locus share the new allele. Filled and open diamonds at the same locus indicate different base substitutions. Two non-new alleles are shown: filled and open diamonds at base 3853 are not new alleles but represent the apparent fixation of 1 of the 2 alleles from the ancestral polymorphism.

HeLa-evolved populations had none in either the N or P genes (fig. 2). Moreover, within a given gene mutations were not equally common across histories (Fisher’s exact test, $P < 0.00001$). In particular, shared mutations in G were more common in the alternating- and MDCK-evolved populations than in the HeLa-evolved populations.

Sets of Co-Occurring Mutations Suggest Epistatic Interactions

Fifteen out of 20 (75%) of the shared new alleles fall into 1 of 5 sets of co-occurring mutations (fig. 3). Sets A and B arose only in HeLa-evolved populations, set C evolved only in alternating-host-evolved populations, and sets

Table 1
Distribution of New Alleles across Genes by Category of Host-regime Experienced

Group	Gene									
	N		P		M		G		L	
	Nucleotide	Amino Acid	Nucleotide	Amino Acid	Nucleotide	Amino Acid	Nucleotide	Amino Acid	Nucleotide	Amino Acid
Shared, HeLa only					U ₂₄₀₁ → C A ₂₄₃₁ → G G ₂₆₈₆ → A 1 ^{NC}	M ₅₁ → T Y ₆₁ → C G ₁₄₆ → E	A ₄₁₆₇ → G	S ₃₆₄ → G	C ₉₅₉₇ → U G ₁₀₀₄₈ → U	P ₁₆₂₂ → L L ₁₇₇₂
Singletons, HeLa only										
Shared MDCK only										
Singletons, MDCK only	1 ^{NC}		2 ^{S,S}		3 ^S					1 ^S
Shared, alternating only	C ₆₉₆ → U	A ₂₁₁	G ₁₅₂₁ → A	E ₄₂	U ₂₉₃₇ → G	* ₂₃₀ → G				
Singletons, alternating only										
HeLa and MDCK										
Shared, HeLa, and alternating							C ₄₁₈₀ → A	T ₃₆₈ → K		
Shared, alternating, and MDCK			U ₁₉₂₇ → G C ₂₁₅₁ → U	Y ₁₇₈ → D Y ₂₅₂ → I	A ₂₄₁₃ → G A ₂₄₃₁ → C A ₂₇₃₆ → G	D ₅₅ → R Y ₆₁ → S N ₁₆₃ → D	C ₃₁₅₃ → A A ₃₈₃₈ → U G ₃₈₄₆ → A U ₃₈₅₆ → C	Q ₂₆ → K E ₂₅₄ → V D ₂₅₇ → N L ₂₆₀ → P	C ₄₉₈₁ → U	A ₈₃

NOTE.—Within each host-regime category and for mutations shared by multiple populations, the nucleotide and amino acid locations and substitutions are shown. Singletons are shown as counts, with each synonymous mutations among a group of singletons indicated with “S”. Mutations in noncoding sequence but within a gene’s mRNA are designated “NC”. No mutations occurred outside of the mRNA regions.

Set	Locus	ORF	HeLa				Alternating				MDCK				Prob.
			H1	H2	H3	H4	A1	A2	A3	A4	M1	M2	M3	M4	
A	G ₂₆₈₆ →A	M		•	•									0.03	
	A ₄₁₆₇ →G	G		•	•										
	G ₁₀₀₄₈ →U	L		•	•										
B	U ₂₄₀₁ →C	M	•	•	•	•								1.00	
	C ₃₅₉₇ →U	L	•	•	•	•									
C	C ₆₉₆ →U	N					•	•						0.03 or 0.001	
	G ₁₅₂₁ →A	P					•	•							
	U ₂₉₃₇ →G	M					•	•							
D	C ₂₁₅₁ →U	P					•	•	•	•	•	•		0.02	
	G ₃₈₄₆ →A	G					•	•	•	•	•	•			
E	U ₁₉₂₇ →G	P					•				•			<0.0001	
	A ₂₄₁₃ →G	M					•				•				
	A ₂₄₃₁ →C	M					•				•				
	U ₃₈₅₆ →C	G					•				•				
	C ₄₉₈₁ →U	L					•				•				

FIG. 3.—New alleles arranged by co-occurring sets in host-evolved populations. Sets are defined such that if one of the constituent new alleles is present in a population, then all other members are also present. The full group of populations sharing the selection history in which the allele set is observed is shown in gray. At right is the probability of detecting the degree of co-occurrence observed if each substitution in a set is laid down randomly among the set's full group of populations. Two probabilities are shown for set C as it cannot be determined whether the correct full group of populations for this set is alternating-host-evolved populations only (4, dark and medium gray bars) or includes all populations exposed to MDCK (all 8 gray bars, bottom).

D and E arose in populations that evolved on alternating hosts and on MDCK alone. Some sets (A and B, C and D) co-occurred within populations. Notably, all sets include mutations in multiple genes, and 2 sets (D and E, both shared by alternating-host-evolved and MDCK-evolved populations) include either confirmed (G₃₈₄₆ → A) or hypothesized (T₃₈₅₆ → C) I1-mAb resistance alleles. One MDCK-only population (M4) did not evolve any of these 5 sets of mutations, although it contains 2 nonunique new alleles (A₂₇₃₆ → G shared with A2 and M3 and A₃₈₃₈ → U shared with A3, fig. 2).

These sets were defined such that if one member occurs in a given population, all other members are also detected. The probability of each set occurring as often as observed if the new alleles interacted only additively can conservatively be calculated as follows. Assuming that each mutation in a set has a probability of 1.0 of appearing in the number of populations in which it was actually observed, we allow the mutation to appear in any of the populations sharing the ecological history of the populations in which the mutations were actually observed. For set A, this means examining the ways in which 3 mutations could each occur in 2 out of the 4 HeLa-evolved populations. We can then calculate the probability that under the possible outcomes so identified we would see only complete sets (fig. 3). These calculations suggest that sets A, C, D, and E are each better explained by epistatic interactions than by additive interactions among loci, though sequential Bonferroni correction indicates that only for set E can we reject additivity. Power to detect epistasis by this approach depends on both the size of the set (more mutations make a set's appearance under additivity less likely) and on the selective advantage associated with the set being sufficiently weak that it is not detected in all populations in which it could have provided a benefit. For this reason, we cannot distinguish epistasis

from additivity in set B as both mutations occurred in all populations sharing the HeLa-only history. As a whole, the 5 sets of co-occurring mutations indicate that epistasis is playing an important role in determining the distribution patterns of these mutations (Fisher's combined probability = 0.019, using higher probability value for set C).

Discussion

The distribution patterns of the genetic changes in these 3 sets of populations elucidate the genetic basis of host adaptation and of the genetic basis of the poor ability of populations selected on a single cell type to use the unselected host. They also shed light on the degree to which populations with the same current niche breadths differ in their ability to expand their host ranges and the genetic basis of these differences. Through these insights into the genetic architecture of host use, the sequences presented here also show that even in organisms like RNA viruses with high mutation rates and large populations, chance can play a significant role in shaping evolutionary trajectories.

Low Fitness on MDCK of HeLa-Evolved Populations

There is only a single allele substitution shared between a HeLa-evolved population and an alternating-host-evolved population (fig. 1 and table 1), indicating that on the whole the individual loci contributing to increased fitness on HeLa depend on the history (HeLa only vs. alternating). The similar phenotypes of high fitness on HeLa cells achieved by these 2 types of populations must have different underlying genetic architectures. To the extent that pleiotropic effects of adaptation to HeLa negatively

influence HeLa-evolved populations' ability to use other hosts, the genetic architecture of their host adaptations in this experiment will constrain them should their ecological circumstances change.

In this overall comparison of the HeLa- and alternating-adapted populations, the large number of accumulated differences makes it difficult to pinpoint a putative allele carrying a cost in MDCK. However, given that the 6 new nonunique alleles in the HeLa-evolved populations appeared in half to all the populations with that history, they would probably have been detected in at least 1 of the alternating-host-evolved populations if they carried no cost in the MDCK environment. Instead, 1 or more of these new alleles probably contribute to increased fitness on HeLa but are deleterious to growth on MDCK (figs. 2 and 3). This is consistent with pleiotropy as the mechanism responsible for the low fitness of HeLa-evolved populations on MDCK cells (fig. 1).

Although the HeLa-evolved populations may have new alleles that contribute to a fitness trade-off, the absence of trade-off in the alternating-host-evolved populations (fig. 1, Turner and Elena 2000) indicates that this outcome was not the only genetic solution to the challenge of improving fitness on HeLa. AP arising in simple, single environment conditions (e.g., Weaver et al. 1999; Hanley et al. 2003; Duffy et al. 2006) may represent the most easily accessible solution to the challenge posed by that environment but not necessarily the only solution. The significantly higher number of unique mutations in the alternating-host-evolved populations relative to the HeLa-evolved populations is consistent with this hypothesis.

Although as suggested by this experiment, exposure to a variable environment might protect some populations from the evolution of genetic trade-offs, clearly such selection cannot always avert trade-offs (e.g., Crill et al. 2000). Interestingly, in this latter case in which alternation between hosts resulted in switching between alleles at 2–3 bases in the capsid gene, the length of time between host switches was quite long (11 days in chemostats in which 100 complete population turnovers occur per day). Presumably, switches at a clearly evolutionary timescale as in the Crill et al. (2000) study should favor such a trade-off, whereas switches occurring at an ecological timescale should select for phenotypic plasticity and the avoidance of trade-off. It would be interesting to determine where the boundary between these 2 outcomes lies and the degree to which it differs among organisms with different mutation rates and population sizes.

Low Fitness on HeLa among MDCK-evolved Populations

In contrast to the HeLa-evolved and alternating-host-evolved populations, the MDCK-evolved and alternating-host-evolved populations share many more mutations (fig. 3). Population M4 shares 2 mutations with other populations but has evolved a unique genetic solution to the challenge of growth on MDCK (fig. 2). Interestingly, M4's phenotype does not reflect this difference from the other MDCK-evolved populations; the MDCK-evolved populations do not differ significantly from one another

in fitness in either host (fig. 1). Unfortunately, as a result of M4's unique genome, we can say very little about the genetic architecture of its increased fitness in MDCK or its low fitness in HeLa.

In contrast, because M1 and M3 share mutation set D with A4, comparing the small group of loci that separate these populations may illuminate which changes are responsible for their differing fitness in HeLa. In M3, we detected differences from the ancestor only at loci in set D, indicating that either individually or epistatically these mutations likely confer most of M3's increased fitness in MDCK as well as its unchanged fitness in HeLa relative to the ancestor.

MA rather than pleiotropy best explains the low fitness suffered by M2 when grown on HeLa. The only detected genetic difference between M2 and A1 is a silent substitution in a serine codon in M2, C₁₇₄₀ → U. The high frequency of this allele in population M2, perhaps through genetic hitchhiking during the selection for the new set E alleles, may have caused low fitness exhibited by M2 (but not by A1) on HeLa.

In general, if there are genetic solutions to the challenge of growth on MDCK without correlated costs on HeLa, we would expect to find some mutations that are fixed in MDCK-evolved populations also occurring in the alternating-host-evolved populations. In fact, all 9 nonunique alleles that occur in any MDCK-evolved population also occur in 1 or more alternating-host-evolved populations, suggesting that these putative host-adaptation alleles do not confer low fitness on HeLa (fig. 2). Alternative interpretations do exist; for example, some singletons observed in the MDCK populations may be beneficial in MDCK and confer a cost in HeLa, and some nonunique mutations may have costs in HeLa that are compensated in the alternating-host-evolved populations. However, finding many instances of shared new alleles between MDCK-evolved and alternating-host-evolved populations is consistent with a MA contributing to the low fitness of the MDCK-evolved populations when grown in HeLa.

High Fitness on Both Hosts of Alternating-Host-Evolved Populations

Population A4 contains all 3 mutations found in M1 (set D mutations, fig. 3) and does not differ significantly from M1 in fitness on MDCK. In addition, A4 carries 3 mutations not found in M1 or in any other MDCK-evolved population. Therefore, 1 or more of these 3 mutations probably makes A4's fitness on HeLa higher than M1's. One possible candidate is C₄₁₈₀ → A, the only new allele occurring in both a HeLa-evolved and an alternating-host-evolved population. A4's increased fitness on HeLa therefore appears to be genetically independent of its increased fitness on MDCK. In contrast, the new alleles in set E (shared by M2 and A1) may confer higher fitness on both hosts through pleiotropic effects on the 2 host-use traits. Set E alleles are the only detected changes from the ancestral sequence in A1, further supporting the hypothesis that this set probably causes A1's high fitness in both hosts.

Genome-Wide Epistasis and the Evolution of Complex Traits

Strikingly, rather than being distributed randomly among populations in this study, many parallel mutations fall within 5 sets of co-occurring mutations among the evolved populations. These sets probably arose in steps. First, a mutation with a large benefit became common in some populations, in parallel. This could occur because RNA virus mutation rates are high enough that multiple populations may find the same beneficial mutation. If this mutation changed the epistatic landscape, then the second step could have occurred: 1 or more mutations that were beneficial in the context of the new epistatic landscape arose and came to dominate in parallel but only in those populations in which the first mutation occurred. Alternatively, because the common ancestor of these populations is known to be polymorphic at one locus, it is theoretically possible that a “hidden ancestor” containing a set of substitutions (A–E, table 2) was among the founding genotypes of some populations. If this were the case, fixation of that set could occur even if the loci in question interacted additively.

Regardless of whether the sets of new alleles arose in the ancestor or in an evolving population, under additivity, a polymorphic population in which the full set was still rare would contain many more genotypes carrying an incomplete set than genotypes carrying the full set because mutants at multiple loci most commonly arise by the sequential addition of mutations to a lineage. Thus, regardless of when the last mutation in the set occurred (before or after the founding of the experimental populations), multiple observations of populations with full sets under an additive model should only have occurred if all substitutions in the set were beneficial. However, if all substitutions in the set were beneficial, generation of the sets via sequential additions of mutations should result in the occurrence of some populations with partial sets for those like A, C, D, and E for which some populations do not have the full set. Because the latter was not observed, a functional interaction (epistasis) among the substitutions is more likely than are additive effects.

Notably, all the sets of co-occurring mutations involved multiple genes. Cuevas et al. (2002), Novella et al. (2004), and Novella and Ebendick-Corpus (2004) also found such multigene mutation sets in evolved VSV genomes, indicating that at least under experimental evolution conditions adaptation may commonly involve epistatic interactions affecting multiple gene products. Such differences in multigene sets across evolutionary conditions suggest testable hypotheses regarding variation in protein–protein interactions within different host cells and could contribute to our understanding of the cell biology of viral infections.

Epistasis Involving Matrix Protein Amino Acid 51

One allele implicated in epistatic interactions in our data deserves particular mention. Group A (fig. 3) includes $U_{2401} \rightarrow C$, which causes the amino acid change M protein $Met_{51} \rightarrow Thr$. Other mutations at this locus have been implicated in a number of host interactions, including the timing of host cell apoptosis (Kopecky et al. 2001; Kopecky

and Lyles 2003), host cell production of interferon (Kopecky et al. 2001; Ahmed et al. 2003), and the relative quantity of 3 M gene translation products in infected cells, one of which begins at amino acid 51 (Jayakar and Whitt 2002). In fact, mutant VSV with altered M protein at amino acid 51 has been proposed for use in treatment of disseminated metastases because of its ability to selectively infect and kill interferon-resistant cancer cells (Lichty et al. 2004; Stojdl et al. 2004; Lun et al. 2006).

Unlike the genetically manipulated M protein mutants used in the studies cited above, populations H1–H4 evolved improved efficiency on an interferon-resistant cell line (HeLa) in response to natural selection. In all, 4 of these populations, $C_{9597} \rightarrow U$, which causes L protein $Pro_{1622} \rightarrow Leu$, accompany the M protein $Met_{51} \rightarrow Thr$ mutation. Either on its own or through an interaction with the mutation at M protein amino acid 51, this change in the major component of the RNA polymerase is beneficial in HeLa but only when there is no exposure to MDCK. This study indicates a need to further explore the interaction of well-described loci of interest with loci at which experimentally evolved populations exhibit co-occurring changes. In light of the proposed therapeutic application of M protein mutants, the interaction between L1622 and M protein amino acid 51 is a candidate for such a study.

RNA Viruses and the Genetic Architecture of Host Adaptation

The lack of symmetry in the genetic architecture of host use in these 3 sets of populations differs strikingly from the symmetry of their fitnesses (fig. 1, Turner and Elena 2000). As a result, although the genomes of the HeLa-evolved populations can be immediately paired with their ecological history from even a partial set of genetic changes relative to the ancestor, the alternating-host-evolved and MDCK-evolved populations cannot be easily distinguished based on their genotypes though their fitnesses on HeLa differ markedly. The much stronger signature of pleiotropy in the adaptive changes seen among HeLa-evolved populations relative to MDCK-evolved populations suggests that although these populations evolved under similarly simple host-regimes they may now differ in their abilities to expand their host range.

Further support for a difference in host range expansion abilities between these 2 types of populations comes from the much higher variance in the fitness estimates of MDCK-evolved populations relative to the other populations (Turner and Elena 2000, fig. 1). Increased “evolutionary error” rather than experimental error may have caused this greater variance for fitness: These populations may be only a single mutational step away from achieving high fitness on MDCK. During fitness assays on HeLa, a sample subpopulation containing the needed mutation would experience expansion of that genotype in the population causing a high fitness estimate, whereas samples without such a mutation would yield low estimates. The ecological pattern of host availability clearly affected the likelihood of a population’s adaptive alleles having pleiotropic effects. However, this likelihood was poorly predicted by the level

of ecological complexity of host availability (single vs. temporally fluctuating).

The large number of silent substitutions that appear to be adaptive indicates that an additional kind of pleiotropy played an important role in the genetic architecture of host adaptation. Nonneutral silent substitutions in coding regions indicate pleiotropy, in that the base in question is both coding for an amino acid and serving some other function (structural, regulatory, or other). Such pleiotropy between coding and noncoding functions, like all pleiotropies, can constrain the evolution of genetic architecture. In this study, silent substitutions were quite common in the evolved populations (table 1); these data are similar to the results of Novella et al. (2004) and Cuevas et al. (2002). In both of those studies parallel replacement substitutions outnumber parallel silent substitutions, but in both cases many of the silent mutations appear to be adaptive.

The other striking result of this study is the large number of parallel changes among the populations. This pattern is the norm in other studies involving RNA viruses (Cuevas et al. 2002; Novella and Ebdick-Corpus 2004; Novella et al. 2004; Greene et al. 2005; Wichman et al. 2005) and is most likely driven by RNA virus' high mutation rates and large population sizes. Nevertheless, we also detected 3 signatures of stochasticity in the evolution of host use in this experiment.

First, we observed that 2–9 mutations were detected in each population, suggesting that an adaptive sweep occurred about every 2.6–8 passages. Consistent with the results of Miralles et al. (1999), this result, in conjunction with the high mutation rate in VSV, suggests that clonal interference was important in our study. Clonal interference is the elimination of beneficial alleles during selective sweeps. Where clonal interference occurs, it enhances the influence of stochastic factors such as the order and timing of mutations on evolutionary trajectories.

Second, detection of M4's unique genetic architecture and of different mutation sets in populations sharing the same selection history indicates that even in an organism with a mutation rate as high as VSV's the evolutionary process is sufficiently mutation limited for stochastic differences to cause populations to diverge. Such stochasticity in evolutionary paths is a hallmark of the influence of epistasis on the evolution of genetic architecture of complex traits. Stochasticity resulting in unique adaptive solutions like that achieved by M4 in its evolution of higher fitness on MDCK may be more common in organisms with large genome sizes and low mutation rates, where the likelihood of parallel fixation of the first mutation in adaptive walks from the same ancestor is low (Remold and Lenski 2004).

Third, the observed sets of co-occurring mutations suggest that epistatic interactions played a prominent role in the genetic architecture of host adaptation. Like clonal interference, epistasis causes stochastic differences in the order or timing of mutations to have cascading effects in the evolutionary trajectories of independent populations. For example, the high frequency of an allele from set D in the MDCK-evolved populations may have excluded any possible benefit of 1 or more alleles from set E to the same population, via a negative epistatic interaction. Similarly, the substitution $A_{2431} \rightarrow G$ in the HeLa-evolved

populations H1 and H4 may have excluded the mutations in set A via negative epistasis. In fact, mutation $G_{2686} \rightarrow A$ causes an amino acid substitution that lies adjacent to the amino acid affected by $A_{2431} \rightarrow G$ (Whitt M, personal communication); they may be alternative but mutually exclusive mutations achieving similar phenotypic change in the protein. These indicators of stochasticity, along with the signs of pleiotropy described above, reveal details of the evolutionary trajectories traversed by the viral populations. In turn, this information illuminates how prior evolution can constrain which phenotypes are evolutionarily accessible to populations.

Acknowledgments

We thank S. Elena for providing strains, I. Novella for primers and much advice on VSV biology, culturing, and sequencing, and D. Dykhuizen, our laboratory members and 2 anonymous reviewers for valuable comments on the manuscript. K. Golemboski and T. Hundley helped develop the IFA and T. Hundley quantified sensitivities to I1-mAb. This work was supported by Yale Donnelley and Ana Fuller Foundation fellowships to S.K.R. and by a Yale Center for Genomics and Proteomics Pilot Grant to P.E.T. and a U.S. National Science Foundation grant (DEB-0452163) to P.E.T and S.K.R.

Literature Cited

- Ahmed M, McKenzie MO, Puckett S, Hojnacki M, Poliquin L, Lyles DS. 2003. Ability of the matrix protein of vesicular stomatitis virus to suppress beta interferon gene expression is genetically correlated with the inhibition of host RNA and protein synthesis. *J Virol.* 77:4646–4657.
- Crill WD, Wichman HA, Bull JJ. 2000. Evolutionary reversals during adaptation to alternating hosts. *Genetics.* 154: 27–37.
- Cuevas JM, Elena SF, Moya A. 2002. Molecular basis of adaptive convergence in experimental populations of RNA viruses. *Genetics.* 162:533–542.
- Drake JW, Holland JJ. 1999. Mutation rates among RNA viruses. *Proc Natl Acad Sci USA.* 96:13910–13913.
- Duffy S, Turner PE, Burch CL. 2006. Pleiotropic costs of niche expansion in the RNA bacteriophage Phi6. *Genetics.* 172:751–757.
- Greene IP, Wang E, Deardorff ER, Milleron R, Domingo E, Weaver SC. 2005. Effect of alternating passage on adaptation of sindbis virus to vertebrate and invertebrate cells. *J Virol.* 79:14253–14260.
- Hanley KA, Manlucu LR, Gilmore LE, Blaney JE, Hanson CT, Murphy BR, Whitehead SS. 2003. A trade-off in replication in mosquito versus mammalian systems conferred by a point mutation in the NS4B protein of dengue virus type 4. *Virology.* 312:222–232.
- Hartl DL, Clark AG. 1989. Principles of population genetics. Sunderland (MA): Sinauer Associates, Inc.
- Holland JJ, de la Torre JC, Clarke DK, Duarte E. 1991. Quantitation of relative fitness and great adaptability of clonal populations of RNA viruses. *J Virol.* 65:2960–2967.
- Jayakar HR, Whitt MA. 2002. Identification of two additional translation products from the matrix (M) gene that contribute to vesicular stomatitis virus cytopathology. *J Virol.* 76: 8011–8018.

- Kopecky SA, Lyles DS. 2003. Contrasting effects of matrix protein on apoptosis in HeLa and BHK cells infected with vesicular stomatitis virus are due to inhibition of host gene expression. *J Virol.* 77:4658–4669.
- Kopecky SA, Willingham MC, Lyles DS. 2001. Matrix protein and another viral component contribute to induction of apoptosis in cells infected with vesicular stomatitis virus. *J Virol.* 75:12169–12181.
- Lichty BD, Power AT, Stojdl DF, Bell JC. 2004. Vesicular stomatitis virus: re-inventing the bullet. *Trends Mol Med.* 10: 210–216.
- Lun XQ, Senger DL, Alain T, et al. 2006. Effects of intravenously administered recombinant vesicular stomatitis virus (VSV^{ΔM51}) on multifocal and invasive gliomas. *J Natl Cancer Inst.* 98:1546–1557 (13 co-authors).
- Lynch M, Walsh B. 1998. *Genetics and analysis of quantitative traits.* Sunderland (MA): Sinauer Associates.
- Miralles R, Gerrish PJ, Moya A, Elena SF. 1999. Clonal interference and the evolution of RNA viruses. *Science.* 285:1745–1747.
- Novella IS, Ebendick-Corpus BE. 2004. Molecular basis of fitness loss and fitness recovery in vesicular stomatitis virus. *J Mol Biol.* 342:1423–1430.
- Novella IS, Zarate S, Metzgar D, Ebendick-Corpus BE. 2004. Positive selection of synonymous mutations in vesicular stomatitis virus. *J Mol Biol.* 342:1415–1421.
- Remold SK, Lenski RE. 2004. Pervasive joint influence of epistasis and plasticity on mutational effects in *Escherichia coli*. *Nat Genet.* 36:423–426.
- SAS Institute Inc. 1999. SAS OnlineDoc®. Version 8. Cary (NC): SAS Institute, Inc.
- Stojdl DF, Lichty BD, tenOever BR, et al. (15 co-authors). 2004. VSV strains with defects in their ability to shutdown innate immunity are potent systemic anti-cancer agents. *Cancer Cell.* 4:263–275.
- Turner PE, Elena SF. 2000. Cost of host radiation in an RNA virus. *Genetics.* 156:1465–1470.
- Weaver SC, Brault AC, Kang W, Holland JJ. 1999. Genetic and fitness changes accompanying adaptation of an arbovirus to vertebrate and invertebrate cells. *J Virol.* 73:4316–4326.
- Wichman H, Millstein J, Bull JJ. 2005. Adaptive molecular evolution for 13,000 phage generations: a possible arms race. *Genetics.* 170:19–31.
- Zeng Z-B. 2005. QTL mapping and the genetic basis of adaptation: recent developments. *Genetica.* 123:25–37.
- Zeng Z-B, Kao C-H, Basten CJ. 1999. Estimating the genetic architecture of quantitative traits. *Genet Res.* 74:279–289.

Edward Holmes, Associate Editor

Accepted March 4, 2008