Evolutionary journey of the retroviral restriction gene Fv1

George R. Young*, Melvyn W. Yap*, Johan R. Michaux†, Scott J. Steppan‡, and Jonathan P. Stoye*e,†

*Retrovirus-Host Interactions Laboratory, The Francis Crick Institute, London NW1 1AT, United Kingdom; †Laboratoire de Génétique de la Conservation, Université de Liège, 4000 Liège, Belgium; ‡UMR Animal, Santé, Territoires, Risques et Ecosystèmes (ASTRE), Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), Campus International de Baillarguet, Université de Montpellier, 34398 Montpellier, France; ‡Department of Biological Science, Florida State University, Tallahassee, FL 32304; and ‡Department of Medicine, Imperial College London, London SW7 2AZ, United Kingdom

Edited by Stephen P. Goff, Columbia University Medical Center, New York, NY, and approved August 17, 2018 (received for review May 18, 2018)

Both exogenous and endogenous retroviruses have long been studied in mice, and some of the earliest mouse studies focused on the heritability of genetic factors influencing permissivity and resistance to infection. The prototypic retroviral restriction factor, Fv1, is now understood to exhibit a degree of control across multiple retroviral genera and is highly diverse within Mus. To better understand the age and evolutionary history of Fv1, a comprehensive survey of the Murideoa was conducted, allowing the progenitor integration to be dated to ~45 million years. Intact coding potential is visible beyond Mus, and sequence analysis reveals strong signatures of positive selection also within field mice, Apodemus. Fv1’s survival for such a period implies a recurring and shifting retroviral burden imparting the necessary selective pressures—an influence likely also common to analogous factors. Regions of Fv1 adapt cooperatively, highlighting its preference for repeated structures and suggesting that this functionally constrained aspect of the retroviral capsid lattice presents a common target in the evolution of intrinsic immunity.

restoration factor | evolution | host-virus interactions | retrovirus

While a variety of viruses occasionally integrate as endogenous viral elements (1), the absolute requirement for an integrated proviral stage is the defining feature of retroviral replication. When infection occurs within a germ cell, endogenous retroviruses (ERVs) may be inherited in a Mendelian manner and hence, form a partial “fossil record” of historic viral burdens. Although originally unappreciated, retroviruses, as filterable, transmissible pathogens, have been studied since the late 1800s. The earliest breeding of inbred animals both facilitated and was necessitated by the study of ERVs and exogenous retroviruses as the agents of “heritable cancer” (2). Research developing these themes in mice led to the description of Friend virus susceptibility 1 (Fv1), a dominant locus conferring protection from otherwise lethal challenges with murine leukemia virus (MLV) (3, 4). Within common laboratory lines, two alleles can be observed, Fv1A and Fv1B, that were identified in BALB/c and NIH-Swiss mice, respectively. Each allele confers resistance to virus of the opposing N and B tropism and may be additively combined (5, 6).

The molecular cloning of Fv1 revealed its derivation from a retroviral gag gene (7, 8). While many examples of such co-options for host defense have been reported, these are most frequently products of env operating through receptor blockade (9), Fv1’s presumably more unique mode of restriction, indirectly determined to be through capsid (CA) binding (10), has remained elusive. Similarly, while its domain organization has been characterized, the protein has not proven amenable to crystallization, and all studies to date have had a necessarily genetic basis. Nevertheless, recent work has expanded the scope of restriction beyond the gammaretroviruses to lenti- and spumaviruses (11).

Based on instances of absence within certain Mus species and on its absence in Rattus, previous estimates have placed integration of Fv1’s progenitor virus at 4–7 Mya (12, 13). Despite this apparently recent ancestry, the pol gene of the progenitor virus is lacking, and neither LTR has been discerned (8). Searches for intact representatives of the progenitor revealed no closely related ERVs, and Fv1 shares only 45% amino acid identity with its nearest neighbor in the mouse genome, MuERV-L (ERV with a leucine tRNA primer binding site) (8). This paradox may result from incomplete representation of exogenous viruses among those endogenized and fixed but, equally, may suggest a longer and more complex evolutionary history. Indeed, Southern blotting revealed hybridizing digestion fragments within the genus Mastomys (12), although this was never further studied.

Here, we have sought to more accurately determine the origin of Fv1 and to use a phylogenetic approach to inform on the historical selection pressures that have shaped its restriction specificities and preserved the gene through evolutionary time. Results

Resolving the History of the Fv1 Locus. Within Mus, Fv1 (GRCm38 Chr4:147,868,979–147,870,358) is located in an ~5-kb region between Migration and invasion inhibitory protein (Mip) and Mitofusin 2 (Mfn2) (Fig. 1). The shared direction and relative separation of this pair are common among assembled genomes from humans and mice through to chickens (diverging ~310 Mya),

Significance

We have charted the evolution of the capsid-binding retroviral restriction factor Fv1 through murid evolution, extending its age to ~45 million years. Functionality can be found outside of the genus Mus, and shared signatures of positive selection are visible across species. Modeling suggests that maintenance for these extended periods can only be parsimoniously explained by repeated selection events—waves of retroviral infection throughout murid evolution. Our results complement and extend findings with TRIM5α and suggest that conserved features of retroviral capsid lattice assemblies may be common targets in convergent evolution of intrinsic defenses to retroviral infection. Functional constraints on capsid structure may prevent effective escape of host factors and result in cyclical coevolution, which is visible in the evolution of Fv1.

Author contributions: G.R.Y., J.R.M., S.J.S., and J.P.S. designed research; G.R.Y. and M.W.Y. performed research; J.R.M. and S.J.S. contributed new reagents/analytic tools; G.R.Y. and M.W.Y. analyzed data; and G.R.Y. and J.P.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Published under the PNAS license.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. MH001948.69 and MH727610.4).

To whom correspondence should be addressed. Email: jonathan.stoye@crick.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1808516115/-/DCSupplemental.
and they thus present a useful framework within which the presence or absence of Fv1 can be established. To initially investigate the presence of Fv1 immediately beyond the genus Mus, we analyzed three sequences from the genus Apodemus: Apodemus sylvaticus from an archived genome assembly and Apodemus uralensis and Apodemus semomus from targeted assemblies of the region. Fv1 was present in all instances, with complete ORFs visible in A. sylvaticus and A. uralensis. Both ORFs exhibited activity directed against MLVs when assayed for restriction capacity (Table 1). The divergence of the Apodemini and Murini tribes predates that of the Murini and Praomyini (14), represented by Murs species for which genome sequences have been published to date. Encouraged, we thus sought to extend this analysis to reexamine the point of insertion of Fv1’s progenitor virus. Representative assemblies from rodent genera for which genome sequences have been published to date (SI Appendix, Table S1) were compiled and searched for Fv1. Fv1 was further noted in gerbils (Meriones and Psammomys) within the Muridae, in hamsters (Cricetulus, Mesocricetus, and Phodopus) within the Cricetidae, and in the blind mole rat, Spalax, within the Spalacidae. No sequences identified in this screen contained intact ORFs. Twenty-two assemblies contained single contigs bridging Miip and Mfn2 and were used to build an alignment of the region (species for which both genes were not assembled together were excluded, as unassembled regions would otherwise be indistinguishable from genuinely absent sequence). Comparisons revealed a high degree of variability due largely to the activity of transposable elements (TEs) (Fig. 1). This variability, combined with the multiple points of erroneous homology presented by TEs, posed a significant challenge and necessitated the use of a repeat-aware alignment program, FSA (15), which can be used in conjunction with RepeatMasker annotations. The region further displayed a propensity for large deletions; one, spanning Fv1, was visible in Rattus norvegicus and explained the absence of hybridization signals within samples from this genus (7, 12). Similar deletions were also visible within other species (Fig. 1).

Thus, although an ORF was absent in many instances, the progenitor integration could be identified throughout the Muridae, Cricetidae, and Spalacidae. Among currently available assemblies, Fv1 was absent in the Dipodidae (Jaculus jaculus) and in all more distantly diverged groups (Fig. 1). Subsequent deletions between Miip and Mfn2 may have occurred since their speciation from the last common ancestor, however, and indeed, otherwise conserved regions of these genes are absent within both J. jaculus and Dipodomys ordii (Fig. 1), highlighting this possibility. Accordingly, these data suggest a minimum insertion time of ~45 Mya within the common ancestor of the Muroidea (14), which might be further extended to ~50 Mya if insertion occurred more basally within the Myodonta.

The Hunt for a Candidate Progenitor. While Fv1’s derivation from an ERV-L is clear, the precise nature of its progenitor remains obscure. Interestingly, our screening revealed regions of an
A consensus sequence is basal to the (orange), and the Eumuroidea (green). The regions of the ERV-L = (purple), − (13), but selective pressures LTR (Fig. 1). While only this spider derives. We thus conducted a more are present within the species sampled. S. galili 3.14 (Fig. 2). Notably, the = (Fig. 2). Revealing Signatures of Positive Selection. per ∼ and represented a useful comparator for per ∼ and with the above . Overall, comparatively few genera retained with the above Fv1 retention by modeling the genes of Fv1 clustered with the Fv1 Fv1 has been described only in the [17] and rat [per Apodemus and Fv1 ∼ at 45 My. site genus is site that surrounded per Fv1 has been described only in the literature (13). Indeed, duplication of multiple copies of Fv1 is estimated to have diversified 8.3–10.1 Mya (14, 18) (SI Appendix, Table S2). Positive selection has previously been noted in comparisons of the Fv1 genes of Mus (13), but selective pressures operating over the newly determined timescales will necessarily result in differences visible not only between species but also, between more distantly diverged groups. The Apodemus genus is similarly sized to Mus and represented a useful comparator for determining and comparing signatures of positive selection. We thus sequenced the Fv1 genes for 15 species of Apodemus, obtaining complete ORFs for 11 (SI Appendix, Table S2).

Nucleotide sequence identity was lower within Apodemus (median of 94.2%) than within Mus (96.8%), suggesting a greater sequence diversity. Signatures of positive selection were 

The 45 My of Fv1. Overall, comparatively few genera retained Fv1 ORFs, and mutational inactivation or deletion was common. This apparent propensity has previously been noted within Mus, where both gene loss and disruption have occurred (13). Despite this, the observation of any intact coding potentials over such extended periods can likely only be parsimoniously explained by their un-interrupted existence. We thus sought to assess the requirement for and frequency of selection events on Fv1 retention by modeling the likelihood of ORF loss under neutral pressure. Such periods (realized through alterations in the viral burden) may gradually occur as a result of geographic movement, habitat change, or changes in association with other species or more rapidly through the involvement of other cellular factors or through receptor escape (16). The probability of survival of a monoexonic 1.380-bp ORF (for Fv1 b) can be modeled with an exponential decay function. Using empirical estimates of background substitution rates for the mouse [μsub = 5.94e−9 per site per year, μindel = 2.88e−10 per site per year (17)] and rat [μsub = 6.31e−9 per site per year, μindel = 3.14e−10 per site per year (17)] as representative examples, half-life values of 0.576 and 0.538, respectively, were calculated (Fig. 3). As these were generic values, we also explicitly simulated the mutation of Fv1 b with the above μ, yielding half-life values of 0.856 and 0.788, respectively (Fig. 3). Using even these higher values, a mean lifetime of 1.14–1.23 My was calculated, and probability of ORF retention fell below 0.05 within 3.5–3.7 My, reaching 1.5e−10 at 45 My.

The maintenance of an Fv1 ORF in any lineage over such extended time periods thus necessitates either continuous or intermittent selection reoccurring at a frequency not regularly exceeding ∼1.2 My.

The Fv1s of Apodemus Reveal Signatures of Positive Selection. To determine the range of genera retaining coding potential, we conducted a larger survey of the Murinae, with a specific focus on the African murines, a frequently recovered evolutionary grouping containing both the Murini and the Apodemyni, which is estimated to have diversified 8.3–10.1 Mya (14, 18) (SI Appendix, Table S2). Positive selection has previously been noted in comparisons of the Fv1 genes of Mus (13), but selective pressures operating over the newly determined timescales will necessarily result in differences visible not only between species but also, between more distantly diverged groups. The Apodemus genus is similarly sized to Mus and represented a useful comparator for determining and comparing signatures of positive selection. We thus sequenced the Fv1 genes for 15 species of Apodemus, obtaining complete ORFs for 11 (SI Appendix, Table S2).

Nucleotide sequence identity was lower within Apodemus (median of 94.2%) than within Mus (96.8%), suggesting a greater sequence diversity. Signatures of positive selection were
visible in both genera; overall, 14 sites were subject to pervasive positive selection, with another 7 sites under episodic positive selection within a subset of species of either genus (Fig. 4). Pervasive selection analyses with FEL and FUBAR assume that selection pressures for each site are constant throughout a phylogeny, assessing selection across all branches, whereas episodic selection analysis, with MEME, determines if individual sites have been subject to selection within a subset of branches. Likely due to the increased statistical power afforded by the larger number of sequences, an increased number of sites displayed positive selection within this analysis in comparison with previous assessments (13).

Supporting the observation of increased variability within *Apodemus*, five of seven instances of episodic selection were within this genus.

Two known hypervariable areas, V_A (Fv1^b^ residues 248–276) and V_B (344–358) (11), were again prominent and together, included 11 of 21 positively selected residues. A third region, V_C (374–401), while variable within *Mus*, was relatively invariant within *Apodemus* and contained only a single residue under positive selection. Continued support for the annotation of V_A and V_B alone was warranted, therefore, and suggested that these two regions likely form the main contact with CA, with individual downstream residues potentially in structural proximity. To explore this further, we sought to determine a metric for “residue involvement”—the frequency of a particular residue pair changing in combination repeatedly (SI Appendix, Fig. S1A) and hence, to determine potential linkages. This might occur where alteration in the size or charge of a residue necessitates a corresponding supporting alteration of another residue in close structural proximity. An alignment of Fv1 genes and calculated nodal sequences was walked to determine all pairs of changes at each branch or leaf where one or both residues were under positive selection. This revealed linkages both within and between V_A and V_B, with residues frequently found to change in

Fig. 4. Fv1 variability in *Mus* and *Apodemus*. Collapsed representation of the multiple sequence alignment of *Mus* (extending upward) and *Apodemus* (extending downward), with the most frequent residue toward the center. Alignment gaps are shaded gray. Sites under pervasive positive selection are boxed red above and below, and those under episodic positive selection are boxed red on one side only. Sites under pervasive negative selection are boxed blue above and below. For comparison, residues linked to specific restriction activities are shaded red (11). Central coloring represents residue involvement (SI Appendix, Fig. S1B) [above: low (blue) to high (red)] and structural predictions from Jpred4 [below: no prediction/unstructured (black), alpha helix (green), and beta sheet (gold)]. The previously identified variable regions, V_A and V_B (11), are boxed.
combinations (SI Appendix, Fig. S1B). Again, however, this did not support the continued annotation of Vc.

Analyses of likely secondary structures with this larger dataset confirm previous α-helix predictions within the N-terminal region, thought to form a coiled coil required for dimerization, as well as the presence and length of the unstructured linker region (19) (Fig. 4). Overall, 10 of 14 and 6 of 7 sites under pervasive and episodic positive selection, respectively, were within the C-terminal region, which has been shown to confer restriction specificity (19). Only sparse residue involvement was observed within the N-terminal α-helices, highlighting their likely conserved structural function. Indeed, where nucleotide variation was observed within this region, it revealed many instances of negative (purifying) selection of corresponding residues (Fig. 4), and while length differences were observed upstream of the N-terminal α-helices and downstream within the linker, none were within the region predicted to form a coiled coil. Length adjustment within the linker was otherwise observed within eight species, with an additional three-residue difference separating the genera. Length adjustments within this area were also determined within other genera within the Murinae, which extend up to 15 residues in size in Fv1 sequences described separately (20).

Discussion

Comparisons of viruses with differing sensitivities to restriction revealed Fv1 to be a CA-binding factor (10, 21). More recently, ordered assemblies of CA have been shown to direct Fv1 binding (22). Here, a requirement for multimerized CA strikes parallels to the CA-binding factor TRIM5α, also shown to interact only with regular arrays of CA (23–25). Both factors exhibit similar domain organizations—an N-terminal facilitating multimerization and a C-terminal conferring restriction specificity (22, 24)—and hybrid factors have activity in vitro (26). Now approaching 50 y since its discovery, however, the mechanism of Fv1 restriction has not been elucidated beyond the achievement of a block between reverse transcription and nuclear entry (27). In fact, no definitive modes of action have yet been described for any CA-binding restriction factors, but mechanisms to promote degradation by the proteasome or to sequester, stabilize, or destabilize the viral core after entry are most widely suggested. Indeed, CA mutants with increased or decreased lattice stability display somewhat equivalent infectivity impairments (28).

In the absence of mechanistic or accurate structural detail, which would facilitate understanding of specific interactions and allow prediction or design of restriction capacity, great efforts have been made to better understand genetic variation at the Fv1 locus. This has revealed a scope of restriction extending beyond the gammaretroviruses (11), suggesting that a variety of viruses have historically contributed to positive selection of the gene within Mus (13). Here, we have sought to more accurately detail the evolutionary journey of Fv1 and present a number of sequences, many with intact ORFs, for species outside the genus Mus. An ML tree built with all complete gene sequences determined here and to date within the literature (11, 13, 20) confirms the widespread distribution of the Fv1 gene across the Eumuridae (SI Appendix, Fig. S2). Accurate field identification of wild-caught animals is an issue pervasive across this and other published studies, and we note a number of potential inconsistencies between suggested identities and the position of certain sequences within this tree both for sequences described herein and for those from the literature. Regardless, their presentation with caveats can only benefit Fv1 research, and we include them accordingly (SI Appendix, Table S2).

Previously accepted to have integrated ~7 Mya (13), we have now dramatically extended this timeframe to ~45–50 Mya, a conclusion supported by other recent research (20). Assuming Dollo parsimony, where a complex trait arises once but can be lost multiple times, the retention of the Fv1 ORF for such extended periods highlights a role for a continuous or frequently reoccurring selective advantage. In turn, this implies repeated waves of infection by novel viruses, possibly as a result of cross-species transmissions. In the absence of such a pressure, Fv1 would be expected to have a mean lifetime of only ~1.2 My. In support of this, we have found Fv1 to be frequently lost within the Muroidea, and even within Mus, Fv1 is deleted or the ORF lost in several instances (11). This further suggests that the progenitor viral gag itself conferred a selective advantage at the point of integration or that its co-option was both unconvoluted and rapid, especially given that, in the absence of such a selective advantage, intact proviruses are otherwise expected to be deleterious.

A requirement for swift transference or adaptation of restriction potential in the face of successive waves of retroviral infections necessitates significant plasticity in Fv1’s mechanism of CA recognition, complicating attempts to predict target specificity from primary sequence alone or to link restriction specificities to specific variations in the endogenous retroviral complement of host lineages. Nevertheless, linkage analyses suggest a previously unappreciated role for cooperative change across the protein, especially within and between VA and VB, inside of which positive selection is also largely confined. Together, these data point toward restriction determination by a face composed largely of these two regions, and indeed, many conversions between restriction capacities require only single-residue substitutions within these areas (11). Repeated change at certain sites may derive from functional constraints on the restriction factor but also from multifactorial constraints on the viral target, leading to the resampling of specific residues and cyclical host–virus coevolution (29). Together, this likely suggests that shared properties of CA lattice structures, rather than the specific properties of any individual CA monomer, are the primary means of recognition.

Fv1α, an ~52-kDa monomer, has previously been shown to be divided into two ~20-kDa fragments separated by a protease-sensitive linker (19). The N-terminal domain has extensive α-helix predictions; in support of a conserved role in coiled coil formation, N-terminal deletions extending past residue 32 are not tolerated, and at least the first 158 residues are required to confer function in hybrid factors (26, 30). Accordingly, we now reveal frequent negative (purifying) selection within this region. In contrast, extensive deletions can be made within the linker, suggesting that its length is of greater importance than its sequence (30). Correspondingly, we now present evidence of length variations within this region, perhaps suggesting a means of fine-tuning C-terminal domain positioning according to differences in the size or curvature of presented CA lattice structures. This again parallels TRIM5α, where regulated positioning of the SPRY domain is central to attainment of an avid interaction with CA (31, 32). Positive selection of TRIM5α has previously been shown to have occurred over at least ~30 My (33, 34) and to have been shaped by lentiviruses over the last 11–16 My (35). Here, we now show that extended coevolution of host factors and their viral partners is also common to murids and undoubtedly necessitates a shifting and reoccurring burden to maintain selection over such timescales. Such fluid interactions likely also present in the TRIM5 gene of the cotton top tamarin, Saginus oedipus, which has also been shown to restrict multiple genera of retroviruses (36–38). CA-binding restriction factors have evolved multiple times in mammals (39), and it might be hypothesized that analogous factors may be common to any host adaptation to retroviral infection or indeed, to any pathogen presenting regularized structures on infection of a cell. Indeed, it is possible that such historic interactions have also shaped both Fv1 and Trim5α. Drawing a parallel to the recognition of LPS from diverse Gram-negative bacteria (40), CA-binding retroviral restriction factors may represent another class of pattern recognition receptors.

Materials and Methods

Modeling ORF Half-Life. In standard genetic code, 22 of 549 possible alterations of the 61 amino acid-encoding codons produce stop codons, assuming an equal rate of mutation across the alphabet: $\varnothing \in \{A, C, G, T\}$. Base
transitions is, however, twice as frequent as transversion among de novo single-nucleotide variants (41), and only 4 of 183 base transitions produce stop codons. Thus, the probability, $P_{\text{stop}}$, of mutation resulting in a stop codon acquisition can be somewhat more accurately represented (42):

$$P_{\text{stop}} \sim \frac{(2 \times 4) + 18}{(2 \times 183) + 366} = \frac{26}{732} = 0.0355.$$

Allowing three substitutions per codon, incorporating the background per-base substitution rates, $r_{\text{base}}$, and the length of the ORF, $t$, allows derivation of the decay constant $\lambda_{\text{sub}}$. Similarly, the decay constant $\lambda_{\text{indel}}$ is derived from $P_{\text{indel}}$, and the probability, $P_{\text{indel}}$, of an indel not being a multiple of the codon length [empirically estimated at 0.83 (43)]:

$$\lambda_{\text{sub}} = \mu_{\text{sub}} \times 3 \times P_{\text{stop}} \times t$$

$$\lambda_{\text{indel}} = \mu_{\text{indel}} \times P_{\text{shift}} \times t.$$

The half-life, $t_{1/2}$, and mean lifetime, $\tau$, are thus calculated with standard formulas:

$$t_{1/2} = \ln(2) / \lambda_{\text{sub}} + \lambda_{\text{indel}} = \tau = 1 / \lambda_{\text{sub}} + \lambda_{\text{indel}}.$$

**Mutation Simulation.** Two Python programs, mutant (commit c9ae773) and orf, scanner (commit 7d7b014), were written to simulate mutational processes and to assess their impact on ORF length. ORF retention rates for 1,000 replicates were used to fit a standard model for exponential decay within $R$ and to derive mean lifetime, $\tau$, and half-life, $t_{1/2}$.

**Software.** The programs developed for this study are available under permissive license at [https://github.com/A-N-Other/pedestal](https://github.com/A-N-Other/pedestal). We encourage code reuse and comment.

**Other Methods.** Full materials and methods are included in SI Appendix.

**ACKNOWLEDGMENTS.** We acknowledge the museums that provided samples for this study—without these excellent collections, this work would not have been possible. We thank Milos Macholan (Czech Academy of Sciences), Ondrej Mikula (Czech Academy of Sciences), Francois Catzeflis (University of Montpellier), Jean-Pierre Quere (French National Institute for Agricultural Research), and Serge Morand (CIRAD/University of Liege) for samples. We also thank Aris Katzourakis (Oxford University) for his thoughtful comments on the manuscript. G.R.Y., M.W.V., and J.P.S. were supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001162), the UK Medical Research Council (FC001162), and the Wellcome Trust (FC001162). S.J.S. was supported by the US National Science Foundation (DEB-0841447).