The Evolution of Seminal Ribonuclease: Pseudogene Reactivation or Multiple Gene Inactivation Events?

Slim O. Sassi, Edward L. Braun,† and Steven A. Benner**

*Foundation for Applied Molecular Evolution, Gainesville, Florida; and †Department of Zoology, University of Florida, Gainesville

Two approaches, one novel, are applied to analyze the divergent evolution of ruminant seminal ribonucleases (RNases), paralogs of the well-known pancreatic RNases of mammals. Here, the goal was to identify periods of divergence of seminal RNase under functional constraints, periods of divergence as a pseudogene, and periods of divergence driven by positive selection pressures. The classical approach involves the analysis of nonsynonymous to synonymous replacements ratios (ω) for the branches of the seminal RNase evolutionary tree. The novel approach coupled these analyses with the mapping of substitutions on the folded structure of the protein. These analyses suggest that seminal RNase diverged during much of its history after divergence from pancreatic RNase as a functioning protein, followed by homoplastic inactivations to create pseudogenes in multiple ruminant lineages. Further, they are consistent with adaptive evolution only in the most recent episode leading to the gene in modern oxen. These conclusions contrast sharply with the view, cited widely in the literature, that seminal RNase decayed after its formation by gene duplication into an inactive pseudogene, whose lesions were repaired in a reactivation event. Further, the 2 approaches, ω estimation and mapping of replacements on the protein structure, were compared by examining their utility for establishing the functional status of the seminal RNase genes in 2 deer species. Hog and roe deer share common lesions, which strongly suggests that the gene was inactive in their last common ancestor. In this specific example, the crystallographic approach made the correct implication more strongly than the ω approach. Studies of this type should contribute to an integrated framework of tools to assign functional and nonfunctional episodes to recently created gene duplicates and to understand more broadly how gene duplication leads to the emergence of proteins with novel functions.

Introduction

According to a standard model (Ohno 1970), the origin of proteins with novel functions begins with a duplication of a gene for a protein performing an ancestral function. Under this model, one duplicate continues to perform the ancestral function, relaxing selective constraints on the second paralog. This allows the second paralog to "explore" regions of "sequence space" without being constrained by natural selection (Zhang 2003; Hurles 2004). Kimura and Ota (1974) proposed that this model represents one of the fundamental "principles governing molecular evolution."

The origin of novel functions is, in some cases, undoubtedly more complicated than suggested by the standard model. For example, gene duplication may be preceded by a period of "gene sharing," where the unduplicated ancestor of the paralogs performs both an ancestral role and a novel role, the latter having arisen whereas the ancestral gene was subject to selective constraints (Hughes 1994). This gene-sharing model postulates that the 2 functions are then partitioned among the paralogs after duplication. Other novel functions may arise from "overprinting," a term used to describe the expression of open reading frames that had not previously encoded a protein (Ohno 1984; Keese and Gibbs 1992; Braun et al. 2000). In fact, examples of exons that resulted from overprinting appear to be more numerous in genes with alternative splicing than exons that arose by duplication (Kondrashov and Koonin 2003). Additional variant models include partial duplications, with or without the formation of a chimera with another gene (Katju and Lynch 2006), and the adaptive change model (Yang and Bielawski 2000; Liberles and Wayne 2002; Bielawski and Yang 2003; Zhang 2003), which invokes positive Darwinian selection on one paralog after duplication. Lastly, the pseudogene reactivation model deserves attention (Balakirev and Ayala 2003); we will explore this model in detail using the example of seminal RNase.

In the standard model, the majority of duplication events are believed to end with the irreversible inactivation of one duplicate (Walsh 1995; Lynch et al. 2001). This belief is consistent with the notion that genes evolving free of constraint have a higher probability of acquiring a mutation that renders the encoded protein pathologically defective (e.g., a nonsense mutation or frameshift) than acquiring a change that results in a novel function. An empirical approach to estimating this probability requires that we examine duplication events in natural history. In the postgenomic age, this has become easier to do, pace the fact that information preserved in the modern genomes is adequate to infer the functional status of duplicates for only recent duplication events. Even so, an integrated framework of tools is needed to help us decide, for reconstructed historical events, whether an ancestral paralogous protein was functional or not.

The tempo of sequence change immediately following duplication has frequently been proposed as a metric to make this decision. Genes that are free of constraint diverge at the rapid rate characteristic of neutral drift and are expected to have a normalized ratio of nonsynonymous to synonymous mutations ($d_N/d_S = \omega$) of unity. Thus, a rapid rate of nonsynonymous sequence divergence ($\omega = 1$) in a duplicate is taken to indicate the absence of constraints (although Balakirev and Ayala (2003) emphasize that assuming $\omega = 1$ is not necessarily warranted for all pseudogenes).

Unfortunately, a pathway giving new function via an episode of positive Darwinian selection will also be characterized by rapid sequence change and a high value of ω that may not significantly differ from unity (although ω values significantly greater than unity are generally accepted as evidence for positive adaptation). Thus, the observation of

Key words: seminal ribonuclease, pseudogene, ruminant, gene duplication, novel function.

E-mail: ssassi@ffame.org.

Mol. Biol. Evol. 24(4):1012–1024. 2007
doi:10.1093/molbev/msm020
Advance Access publication January 30, 2007

rapid sequence evolution in one of the duplicates may also be consistent with this "adaptive model" for the origin of novel functions (Zhang et al. 1998). This leads to the unfortunate possibility that a high ω that does not significantly differ from unity could imply 2 very different conclusions, neutral evolution after the loss of purifying constraint or positive Darwinian selection. Despite this issue, a number of productive efforts have used an elevated ω as a criterion to search for proteins subject to positive Darwinian selection both on a large scale (Endo et al. 1996; Roth et al. 2005) and with specific proteins (Zhang et al. 2002).

In principle, one might distinguish between the adaptive and standard models by determining the period over which rapid sequence evolution took place. If the number of amino acid replacements necessary to shift a protein from one function to another is small (Perutz 1983; Asenjo et al. 1994; Newcomb et al. 1997; Zhang et al. 2002), one can imagine a short period of drift into a fruitful area of sequence space ($\omega \approx 1$), an episode of rapid adaptation ($\omega > 1$), followed by the return to evolution under new functional constraints ($\omega < 1$). Long periods of drift followed by the acquisition of a new function are not expected because genes diverging without constraint for long periods of time are expected to become irretrievably damaged (perhaps with a half life of 5 Myr) (Marshall et al. 1994; Lynch and Conery 2000).

Conversion to a pseudogene is not necessarily synonymous with "gene death." In some cases, pseudogenes can play a regulatory role (Balakirev and Ayala 2003) and are also known to contribute to specific functions, such as generating antibody diversity (Ota and Nei 1995). In other cases, partial reactivation of a pseudogene by exon shuffling produces new functions (Doxiadis et al. 2006). Lastly, pseudogenes could act as donors in interlocus gene conversion that could result in a large number of simultaneous changes to a functional gene (which may or may not be advantageous). An even more extreme example for "life" after conversion to a pseudogene, however, is provided by pseudogene reactivation, which might provide another model for the origin of novel functions. In the pseudogene reactivation model, unconstrained exploration of sequence space continues after mutations render the gene unable to encode a functional protein. Then, lesions incompatible with expression of the pseudogene are repaired. In principle, pseudogene reactivation might allow a gene on one adaptive "peak" to shift to another, even when a required intermediate is toxic. The potential contributions of the reactivation model, whether it is partial or complete reactivation, to the origin of novel functions has led Balakirev and Ayala (2003) to relabel pseudogenes as *poto*genes, for potential genes (using nomenclature Brosius and Gould [1992] originally suggested).

Pseudogene reactivation makes available a longer time to search sequence space, but is believed to generate proteins with new functions only infrequently. The repair of lesions might include the reinsertion of a deleted segment, the removal (in frame) of an inserted segment, or other events that are likely to be improbable. Partial gene conversion with a functional gene as a donor might improve the probability of pseudogene reactivation, but enough of the pseudogene sequence must be preserved to maintain the benefits of expanding the sequence space explored after duplication.

Bovine seminal RNase has been proposed to be an example of a protein encoded by a gene that arose from a reactivated pseudogene (Trabesinger-Ruef et al. 1996). Seminal RNase diverged from the pancreatic RNase family approximately 40 MYA. In the modern ox, seminal RNase is expressed in seminal plasma at a high level (ca. 2% of the soluble protein). The primary function of the RNase in seminal plasma is unclear, but it displays immunosuppressive and other cell-based activities that are highly distinct from the closely related pancreatic ribonucleases (RNases) (Vescia et al. 1980; Laccetti et al. 1992; Kim et al. 1995; Soucek et al. 1996; Sinatra et al. 2000; Lee and Raines 2005).

Orthologs of the gene encoding–bovine seminal RNase in closely related ruminants (e.g., deer, kudu, okapi, and giraffe) have lesions (deletions, insertions, and changes in key residues) expected to be incompatible with production of an active protein (Trabesinger-Ruef et al. 1996; Breukelman et al. 1998; Kleineidam et al. 1999) (fig. 2). Therefore, any role that seminal RNase might play in oxen is not played in other modern ruminants. Further, as seminal RNase pseudogenes are present in multiple lineages branching from the lineage leading to oxen, the gene encoding–seminal RNase was either inactivated multiple times in lineages leading to other modern ruminants (the "multiple inactivation" narrative; see fig. 1*B*) or inactivated only once and was reactivated very recently in an immediate ancestor of oxen (the "pseudogene reactivation" narrative; see fig. 1*A*).

Although the distribution of functional genes and pseudogenes requires invoking 1 of these 2 narratives if the phylogeny shown is correct, a different phylogenetic tree would remove the need for either of these narratives. The topology shown is congruent with other estimates of ruminant phylogeny (Mahon 2004; Hernandez Fernandez and Vrba 2005), but it does remain possible that the seminal RNase phylogeny differs from the ruminant species tree due to incomplete lineage sorting.

Although 3 different narratives can explain the phylogenetic distribution of seminal RNase pseudogenes, initial studies concluded that the pseudogene reactivation narrative was plausible (Trabesinger-Ruef et al. 1996). Consequently, seminal RNase has been viewed as an important example of pseudogene reactivation producing novel function (Harrison and Gerstein 2002; Zhang 2003; Harrison et al. 2005; Katju and Lynch 2006). Here, we examine the value of classical tools, including the estimation of ω by maximum likelihood (ML) methods, to discuss the alternative narratives outlined above. We then introduce additional tools that utilize the 3-dimensional folded structure of the protein as a way to distinguish between these narratives. We found that the evidence strongly supports the multiple inactivation narrative and conclude that bovine seminal RNase should no longer be viewed as an unambiguous example of pseudogene reactivation.

Materials and Methods

Alignment and Phylogenetic Methods

Sequences were obtained from GenBank or sequenced from tissue derived from the Center for Reproduction of

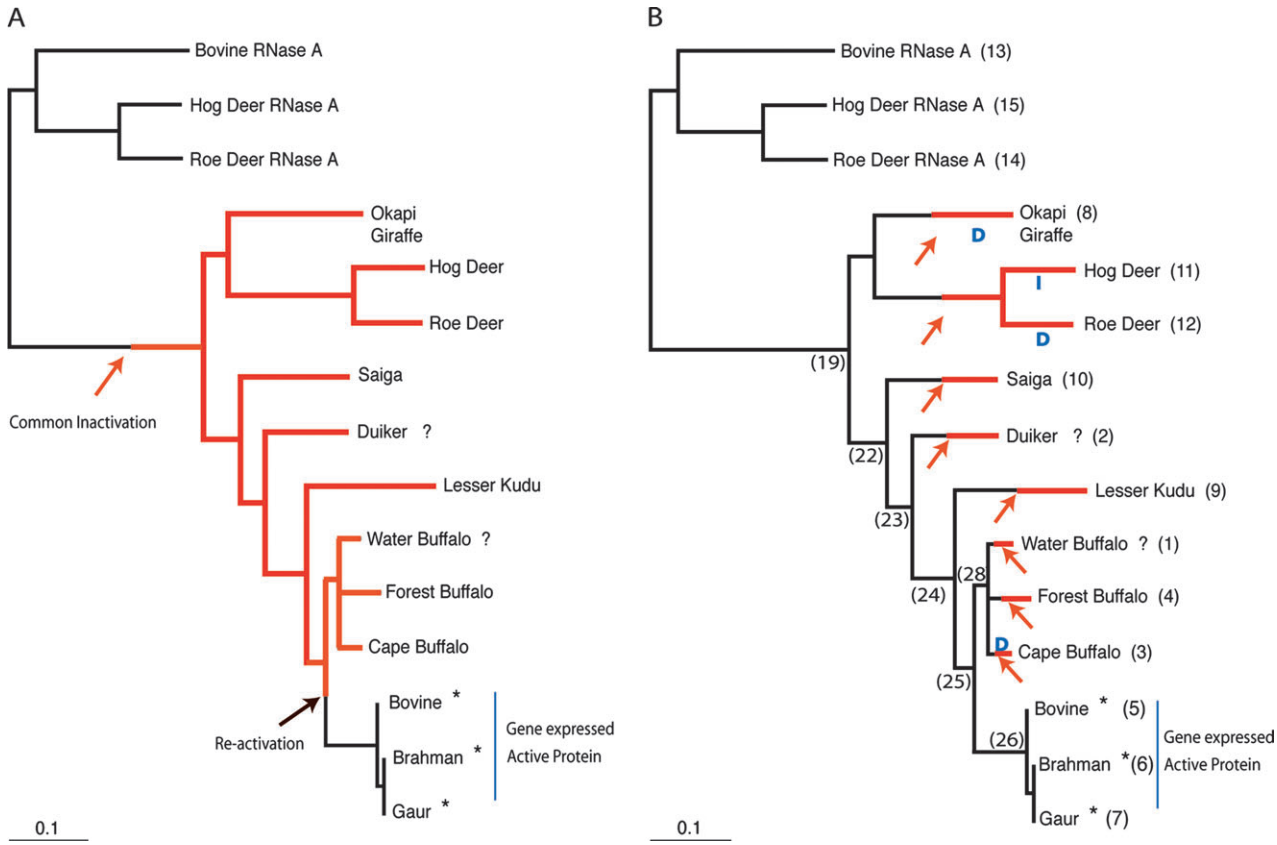

FIG. 1.—Trees showing 2 alternative narratives for the history of ruminant seminal RNase. Episodes during which each narrative postulates that seminal RNase diverged free of functional constraints are indicated by red lines. Periods when each narrative postulates that seminal RNase was under functional constraint are indicated by black lines. Red arrows indicate events where the coding region is proposed by the narrative to have suffered a lesion giving a pseudogene. The black arrow indicates the proposed pseudogene reactivation. * Represents an active and expressed protein in an extant species; ? Indicates a suspected pseudogene (no lesion is present, but examination of available tissues has not indicated the presence of the protein). The appearance of deletions (D) and insertion (I) are shown on the corresponding branches.

Endangered Species (Trabesinger-Ruef et al. 1996). The alignment was produced with ClustalX (Thompson et al. 1994). The DNA sequence alignment was guided by the protein sequence alignment. To capture a comprehensive representation of the phylogeny and its corresponding ambiguity, nucleotide evolution models were selected using Modeltest (Posada and Crandall 1998) using both the likelihood ratio test and Akaike information criterion (AIC). The trees were then calculated using the selected model in PAUP* (Swofford 2001) under the ML optimality criterion. These models were also applied in a Bayesian Markov chain Monte Carlo (MCMC) framework to reconstruct the phylogeny using MrBayes (Ronquist and Huelsenbeck 2003).

Ancestral Reconstruction

The PAML program package was used to reconstruct the ancestral sequences for the seminal RNase genes following an empirical Bayes method (Yang et al. 1995). Three different evolutionary model frameworks were implemented in the reconstruction, a codon model using 2 different procedures to estimate the codon frequencies and an amino acid model. The first codon model (known as 1×4), estimates the frequencies of different codons frequencies by examining the average nucleotide frequencies in the input sequence data as a whole. The second method, 3×4, estimates the frequencies of different codons by examining separately the nucleotide frequencies in the first, second, and third positions in the input data; the frequency of a specific codon is the product of 3 estimated nucleotide frequencies. The third model (the amino acid model) uses an empirical rate matrix (Jones et al. 1992). In addition to using different models to infer ancestral sequences, different tree topologies were considered to reflect uncertainties in the underlying topology. Although the trees shown in figure 1 reflect the estimate based upon Bayesian MCMC analysis using nucleotide data, we also used trees estimated by other methods (ML analysis of nucleotide data).

Structural Mapping and Solvent Accessibility

The solvent accessibility of all residues of seminal RNase was determined using the definition of secondary structure of proteins (DSSP) program (Kabsch and Sander 1983) applied to crystallographic structures of the RNase monomer (pdb:1N3Z [Sica et al. 2003]) and dimer (pdb:1BSR [Capasso et al. 1983], pdb:1R5C [Merlino et al. 2004], and pdb:1R3M [Berisio et al. 2003]). Residues with 10% or greater solvent accessibility were considered

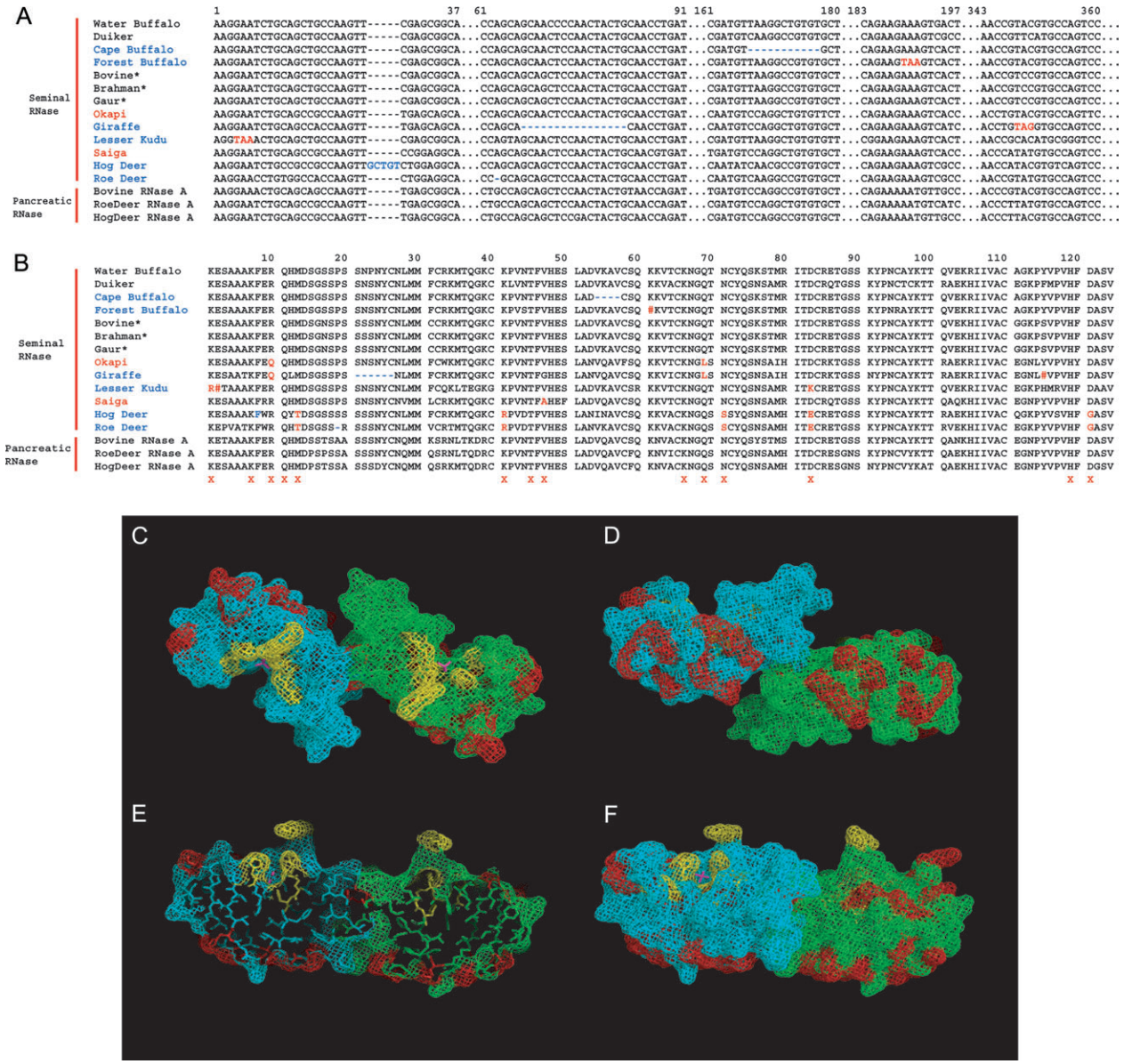

FIG. 2.—Multiple sequence alignment and 3-dimensional crystal (pdb:1R3M [Berisio et al. 2003]) for seminal RNase in its dimeric form. (*A*) DNA sequence alignment. Taxa in blue have an insertion (in blue), deletion (in blue), or a premature stop (in red) relative to the active bovine seminal RNase gene. The indels create frame shifts in the affected genes. Taxa in red have amino acid replacements at the active site likely to render the protein unable to act as a catalyst for the hydrolysis of RNA. (*B*) Protein sequence alignment using the same colors as the DNA sequence alignment. # Represents premature stops. (*C–F*) Different projections of the crystal structure. Blue and green distinguish the subunits of the seminal RNase dimer. Red indicates sites that have amino acids in the last common ancestor of all seminal RNases different from the active bovine seminal RNase. Yellow indicates active site residues. (*C*) Active site is toward the viewer. (*D*) Active site is away from the viewer. (*E*) Cross section showing that none of the sites suffering replacement are inside the folded core of the protein, other than a single site (32) at the dimer interface. (*F*) Active site at the top of the image.

solvent exposed. The solvent accessibility of the ancestrally replaced amino acids was also determined in the same way. The statistical significance of the observed surface distribution of the ancestrally replaced amino acids was determined using a χ^2 test.

Calculations of $\omega = d_N/d_S$

Codeml in the PAML program package was used to calculate all ω (d_N/d_S) values (Yang 1997) under the ML codon model. When different values for ω were calculated for different branches or different branch groupings, the branch model as implemented in PAML was used (Yang 1998; Yang and Nielsen 1998).

Simulations

Simulated data sets under different branch models (Yang 1998; Yang and Nielsen 1998) were produced using the evolver NS branch sites version of Evolver in the PAML (version 3.15) package (Yang 1997). At least 1,000 data sets were simulated for each set of conditions. The seminal

RNase gene and species tree was used in the simulations. Branch lengths, κ (transitions–transversions ratio), and codon frequencies values from the codeml ML analysis of the seminal RNase data set were used in the simulations. Different ω values were also applied to generate the simulated data sets and were varied depending on the simulation conditions as discussed in the Results and Discussion section. The simulated data sets were then analyzed using codeml in the PAML program package to estimate ω and other parameters. The programs Excel and Prism were used to examine the distributions of ω and conduct the statistical analyses.

Incomplete Lineage Sorting

The species/gene tree was compared with a tree that would follow an incomplete lineage-sorting narrative (see Supplementary Material online). The Shimodaira–Hasegawa test as implemented in PAUP* (Swofford 2001) was used to evaluate the topologies.

Gene Conversion

Four different trees representing 4 possible narratives of gene conversion were compared (Supplementary Material online) with each other and to the species/gene tree in a parsimony framework as implemented in PAUP*. Tree lengths for each nonconstant character that vary among the 5 trees were compared and characters that supported one of the trees associated with possible gene conversion events were examined.

Results and Discussion

Seminal RNase Phylogeny is Congruent with Ruminant Phylogeny

The estimate of the RNase gene tree obtained by a Bayesian MCMC analysis here (fig. 1) includes a seminal RNase clade and a pancreatic RNase clade as expected. The seminal RNase clade has a topology that is almost completely congruent with the likely ruminant species tree (Mahon 2004; Hernandez Fernandez and Vrba 2005). This suggests that the history of the seminal RNase gene matches the evolutionary history of ruminants inferred using multiple lines of evidence (morphology as well as mitochondrial and nuclear sequence data) with at most modest topological differences that can be explained by population genetic processes (Pamilo and Nei 1988; Maddison 1997) along with uncertainty in the gene tree and/or species tree.

Despite the general congruence, we wanted to rigorously test the possibility that incomplete lineage sorting might be able to explain the distribution of pseudogenes and functional genes in the extant ruminant species. This narrative postulates that the ancestrally inactivated allele of seminal RNase did not become fixed in the population. Instead, the nonfunctional pseudogene allele was maintained along with a functional allele. In this narrative, the distribution of pseudogenes and functional genes, as observed in the gene tree, is explained by recent losses of polymorphism fixing either the pseudogene or the functional gene in specific lineages. Although instances of deep coalescence that cause modest differences between the gene tree and the species tree are possible, this narrative would require the maintenance of an ancestral polymorphism for an unusually long period of time (through multiple speciation events). The gene tree topology consistent with the deep incomplete lineage-sorting narrative can be excluded (P value = 0.03) using the Shimodaira–Hasegawa topology test (Shimodaira and Hasegawa 1999). On these grounds, we excluded this possibility of deep incomplete lineage sorting and focused on the 2 remaining narratives: pseudogene reactivation and multiple recent inactivations.

Seminal RNase Lesions Support Independent Gene Inactivation Events

The fact that the distribution of functional seminal RNase genes is explained most parsimoniously by the pseudogene reactivation narrative (assuming equal costs for conversion between pseudogenes and functional genes) when combined with the absence of evidence for function of this gene outside of the bovine lineage (suggesting that the seminal RNase gene had no function for ~35 Myr) the best corroborated hypothesis is pseudogene reactivation (Trabesinger-Ruef et al. 1996).

However, the independent inactivation narrative is more consistent with the sequences of the seminal RNase pseudogenes because the specific inactivating lesions differ in each of the ruminant lineages. It is more parsimonious to conclude that none of the lesions in various ruminants were present in the internal nodes of the seminal RNase tree (table 1). These data do not exclude the pseudogene reactivation model because it remains possible that an initially inactivating lesion was lost or occurred outside of the sequenced regions (e.g., the promoter or untranslated regulatory regions). In such a historical narrative, the mutations with the potential to inactivate the gene do not represent events that initially inactivated seminal RNase; instead, they simply reflect the spectrum of mutations expected for pseudogenes after expression was lost. In this version of the narrative, reactivation of an unexpressed seminal RNase gene occurred just before the ox and buffalo diverged through mutation in a regulatory region (fig. 1*A*; also see Trabesinger-Ruef et al. 1996), which was possible because the coding region avoided a lesion (by chance) in the time since divergence, despite being a pseudogene. In fact, as emphasized previously, the pseudogene reactivation narrative is the most parsimonious narrative if one considers only the functional or pseudogene status of the seminal RNase genes (fig. 1*A*).

Nonsynonymous Evolutionary Rates Varied during Seminal RNase History

Estimates of ω were initially obtained for each branch of the tree using a ML method, using a parameter-rich model that allowed each of the 28 branches to be associated with an independent ω value. Many of the estimated ω values were much lower than unity, suggesting that seminal RNase has been subject to purifying selection during most episodes represented by branches in the seminal RNase tree. Some ω values were extremely high ($\omega \gg 1$), however, reflecting either positive Darwinian selection or short branch lengths having few synonymous substitutions

Table 1
Expression of Seminal RNase in the Studied Species and the Lesion Status of the Corresponding Seminal RNase Gene

	Seminal RNase–Coding Sequence Lesions			
Species[a]	Insertions/Deletions	Active Site Replacements	Premature Stop	Protein Expression
Okapi		2 (R10Q and Q69L)		No
Giraffe	16 Nucleotide deletion	3 (R10Q, Q69L, and V47G)	Codon 115	No
Hog deer	5 Nucleotide insertion	5 (M13T, K41R, N71S, D83E, and D121G)		No
Roe deer	1 Nucleotide deletion	5 (M13T, K41R, N71S, D83E, and D121G)		No
Saiga		1 (V47A)		No
Duiker				No
Lesser kudu		2 (K1R and D83K)	Codon 2	No
Water buffalo				No
Cape buffalo	11 Nucleotide deletion			No
Forest buffalo			Codon 62	No
Bovine				Yes
Brahman				Yes
gaur				Yes

[a] Species: Okapi (*Okapia johnstoni*), Giraffe (*Giraffa camelopardalis*), Roe deer (*Capreolus capreolus*), Hog deer (*Axis porcinus*), Saiga (*Saiga tatarica*), Yellow-backed Duiker (*Cephalophus sylvicultor*), Lesser Kudu (*Tragelaphus imberbis*), Water Buffalo (*Bubalis bubalis*), Cape Buffalo (*Syncerus caffer caffer*), Forest Buffalo (*Syncerus caffer nanus*), Bovine (*Bos taurus*), Gaur (*Bos gaurus*), and Brahman (*Bos indicus*) is a breed of Zebu.

(a "division by zero" problem; see Supplementary Material online).

Accordingly, adjacent branches in the tree were grouped to generate a set of ω estimates that were fewer in number than the number of branches in the tree. This grouping decreased the number of free parameters, increased the number of sites useful to estimate individual ω, and consequently decreased the variance of the ω estimates. In the first clustering, all branches with low ω (<1) from the initial analysis (which estimated a separate ω for each branch) were collected into a single group assumed to be described by a single ratio. The branches with ω higher than unity were allowed to have individual ω values unless the branches were adjacent, in which case the adjacent branches were constrained to have a single ω parameter. This resulted in 4 groups of branches, 1 containing the majority of branches and having $\omega < 1$ (the "background ω value") and 3 groups that are candidates for $\omega \geq 1$ (the "high ω" groups). The high ω groups were combinatorially merged into the group with background (low) ω value, ultimately generating a set of 7 models (1 with 3 high ω groups, 3 with 2 high ω groups, and 3 with 1 high ω). This process was designed to cover all possible combinations for calculating ω values from the most complex (each branch with a distinct ω value), to intermediate models (e.g., 3 different high ω groups and the background ω), to the simplest model with more than one ω value (2 group models with 1 high ω and the background ω). These models were also compared with an even simpler model that assumes a single ω value for the entire tree.

These models were then evaluated using the AIC (Burnham et al. 2002; Posada and Buckley 2004), which provides an estimate of the Kullback–Leibler distance between an approximating model under consideration and the unknown "true" model. The AIC provides a way to assess whether the fit of models (based upon likelihood scores) sufficiently improves when parameters are added to justify the increased model complexity.

The best model proved to have 2 ω values, a high ω value ($\omega \approx 1.6$) for 2 recent branches leading to the bovine lineage and a lower one ($\omega \approx 0.3$) for the remainder of the tree (Supplementary Material online). All other groupings increased the complexity of the model and its fit to the data, but that increase was not sufficient to compensate for the cost of having an increased number of parameters. This suggested that the seminal RNase gene was subject to purifying selection during most of its evolutionary history, with the exception of a brief and recent period of positive selection in the immediate ancestry of oxen. Given that $\omega \approx 0.3$ throughout the majority of the tree, these results are more consistent with the multiple gene inactivation narrative than other narratives.

Estimates of ω Suggest That Seminal RNase Genes Were Subject to Selective Constraint

The pseudogene reactivation model predicts that estimates of ω for most branches within the seminal RNase tree will be near unity because the sequences would have undergone neutral evolution after selective constraints were lost. In contrast, independent gene inactivation predicts that the evolution after gene inactivation will be free of constraint ($\omega \approx 1$ for external branches leading to the modern pseudogenes in various ruminants), whereas internal branches will show evidence for selective constraint ($\omega < 1$).

If we assume the independent gene inactivation narrative, the data are inadequate to say where along the external branches the events that created the pseudogenes occurred. Thus, the functional constraints along those branches are unknown. If the event occurred early, then the sequence underwent neutral drift during most of the time represented by that branch, and the branch should have $\omega \approx 1$. If the event occurred late, then the sequence was diverging under functional constraint along most of the branch in question, and ω is expected to be less than unity.

The model with separate ω values for each branch (considered to be overparameterized by the AIC) is consistent with the independent inactivation model as well. In this model, the estimates of ω for the majority of internal branches were substantially less than unity. Likewise, the

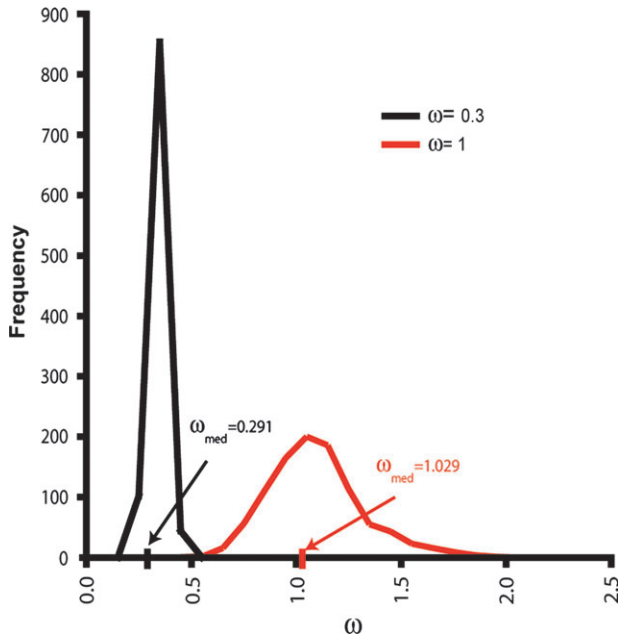

FIG. 3.—Results of simulations assuming constrained and neutral evolution. Black curve shows the distribution of ω estimates if the true ω value for the internal branches is 0.3. Red curve shows the distribution ω estimates if the true ω value for internal branches is unity.

estimates of ω for terminal branches leading to pseudogenes (hog deer and roe deer, okapi, saiga, kudu, Cape buffalo, and forest buffalo) and suspected pseudogenes (duiker and water buffalo) were less than unity.

We then asked how robust these inferences were. To examine rigorously the ability of ML estimates of ω to detect neutral evolution over the tree as a whole, we simulated seminal RNase sequence evolution assuming either constraint ($\omega = 0.3$) or neutral evolution ($\omega = 1$). Although the ω estimates obtained in analyses using the simulated data showed substantial variance (fig. 3), probably reflecting the small number of sites available to calculate ω given the short length of the seminal RNase sequences, the distributions did not overlap. This provides strong support for the contention that the seminal RNase sequences were subject to constraint over much of their history, an observation consistent with the independent inactivation narrative. It is important to note the ω variance increased in the neutral evolution simulations (fig. 3).

Information from RNase Structure Supports Purifying Selection

We then considered a different approach to addressing these questions, one that did not solely rely on linear sequence data. If a protein is divergently evolving without functional constraint, the sites holding amino acid replacements should be randomly distributed with respect to its surface, interior, active site, and other functionally relevant features of its folded structure. In contrast, if selective pressures constrain amino acid replacement, then the distribution of amino acid replacements in the 3-dimensional structure should not be random.

As the structure of seminal RNase is known, this approach could be directly applied to the distribution of amino acid replacements that accumulated along the internal branches in the seminal RNase tree. The sequences of the ancestral proteins were inferred for all ancestral nodes in the RNase tree (Supplementary Material online) using the empirical Bayes method (Yang et al. 1995) implemented in the program PAML (Yang 1997). Ambiguity was incorporated into the reconstruction by examining multiple evolutionary models (2 codon models and an amino acid model) and by varying the tree topology (using other plausible topologies). All of the amino acids that changed along the internal branches were then mapped on the dimeric bovine seminal RNase structure and visualized in 3 dimensions (fig. 2).

This procedure revealed a distribution of ancestral amino acid replacements that was apparently nonrandom. By eye, amino acid replacements appeared to occur almost entirely on the surface of the biomolecule; sites within the folded core of the protein were largely free of replacements. Further, the RNA-binding site and -active site were unchanged (fig. 2). This pattern of replacements indicated that purifying selection did not permit the accumulation of amino acid replacements that would be most likely disrupt the folding (i.e., those in the core of the protein, recognizing that replacements in the core need not disrupt folding, especially if they are associated with compensatory changes) or enzymatic activity of the protein. In contrast, the replacements that have been permitted appear to be those that are least likely to have an impact on folding or activity (i.e., surface residues, although some surface residues may be exceptions to this general rule). Indeed, one surface residue that changed (cysteine-31) is important for seminal RNase quaternary structure (Mazzarella et al. 1993). However, the change at this site is unlikely to indicate the loss of function because it reflects a change to a cysteine residue at this position (table 2). The inference that the ancestral proteins were active was confirmed in a separate study by "resurrecting" those proteins (Sassi S, unpublished data). These same observations regarding the distribution of amino acid changes were true, leading us to the same conclusions, regardless of whether crystal structures based upon the monomer or the dimer were used (pdb:1N3Z [Sica et al. 2003], pdb:1BSR [Capasso et al. 1983], pdb:1R5C [Merlino et al. 2004], and pdb:1R3M [Berisio et al. 2003]).

To place these observations in a quantitative framework, the solvent accessibility of individual amino acid side chains was estimated from the seminal RNase structures using DSSP (Kabsch and Sander 1983). The residues were placed into 2 bins depending upon their solvent accessibility. Residues with solvent accessibility $<10\%$ were considered to be in the core. These core residues represented 36.7% of the 124 amino acids in seminal RNase; the remaining 63.3% of the residues were considered to be surface residues because they are solvent accessible sites. Slightly more than 94% of the sites that accumulated amino acid replacements were solvent accessible, which is significantly more than expected if the residues that accumulated replacements were randomly distributed ($\chi^2 = 7.5130$; degree of freedom (df) = 1; $P = 0.008$).

This structural mapping analysis supports the conclusion that seminal RNase was subject to purifying selection because amino acid replacements appear to have followed

Table 2
Summary of Amino Acid Replacements along Internal Branches of the Seminal RNase Phylogeny

	Branches							
Amino Acids Positions	19..20	20..21	19..22	22..23	23..24	24..25	25..26	25..28
9		E, G → W		G → E				
17							S → N	
19		P → S						
22		N → S					N → S	
31							F → C	
53	D → N		N → D					
55		Q → K		Q → K				
64		T → A						
70	T → S		S → T					
76						N → K		
78						A → T		
80	R → H		H → R					
101	Q → R							
102					A → V			
105					H → R		R → H	H → R
111						E → A	E → G	E → A
113		N → K		N → K				
115							Y → S	

NOTE.—Branches are identified by the 2 delimiting node numbers as in figure 1 (e.g., 19..20).

the expected pattern for an active enzyme. Therefore, these results support the model with multiple gene inactivation events that took place in the recent evolution of the clade, as represented by the external branches of the tree. This independent line of reasoning is entirely consistent with that obtained from ML estimates of ω.

Use of ω and Crystallographic Analysis to Understand Deer Seminal RNases

The lineage leading to the seminal RNase pseudogenes in the deer presents a unique opportunity for further understanding of both the ω estimation and crystallographic tools to determine whether functional constraints were present or absent during an episode of evolutionary history. The seminal RNases pseudogenes from the 2 cervid taxa (hog and roe deer) share common lesions that almost certainly make both unable to encode an active protein; 5 amino acid replacements in the active site (fig. 2). This suggests that those lesions were present in their last common ancestor and that the ancestral protein was also inactive. These 2 deer species are estimated to have diverged ~19 MYA (Hernandez Fernandez and Vrba 2005), a significant period of time for a lineage to have been free of functional constraints.

Indeed, the crystallographic analysis developed here supports that conclusion. Of the 13 amino acid replacements estimated for these branches within the cervid evolutionary history, 5 are buried and 8 are not, nearly exactly the same ratio of buried and surface residues found in the protein as a whole (38.5% buried and 61.5% exposed; $\chi^2 = 0.017$; df = 1; $P = 0.896$). The estimate of ω offers this inference less persuasively, with ω = 0.4 and 0.7 estimated initially for the hog and roe deer branches, respectively. A pairwise ω estimate using a standard d_N and d_S estimator (Nei and Gojobori 1986) for the hog and roe deer alone is 0.73 because d_N is 0.1025 and d_S is 0.140. The best model based upon the AIC did not include a separate ω estimate for the cervid lineage, however, and a model with a separate ω for the deer did not suggest unity as a value (ω = 0.5 for the deer branches and ω = 0.3 for the remaining internal branches). Standard theory would not interpret these values as evidence for an absence of functional constraints.

If the last common ancestor of hog and roe deer indeed contained a seminal RNase pseudogene, then the gene inactivation must have predated the last common ancestor of the deer included in this study. To see how the timing of gene inactivation influenced estimates of ω, seminal RNase evolution was simulated using the gene tree topology and the ML empirical parameters obtained from the RNase sequence alignment (the values estimated by PAML for branch length, transition/transversions ratio (κ), ω = 0.3 for internal branches, and the codon frequencies). The terminal branches leading to the hog deer and roe deer were assumed to have ω = 1, as was a portion of the internal branch leading to their common ancestor. Because the timing of gene inactivation along that internal branch is not constrained by the data, simulations placed the loss of constraint at various points along the internal branch, with a corresponding increase of ω from 0.3 to unity. All simulated data sets were analyzed using codeml from the PAML package (Yang 1997), one ω value calculated for the 3 branches of the deer clade (fig. 4).

As expected, the median ω estimate increased as the position of the transition along the internal branch (C) was made more ancient (fig. 4). Strikingly, the distributions for ω estimates progressively widened as the length of the neutral evolution period was increased; this would be expected to negatively affect the confidence interval. The distribution is widest when ω = 1 is assumed for the full length of all 3 branches, the internal branch (C) and the external branches leading to the hog and roe deer. The distribution is much narrower when all the 3 branches are modeled as having evolved under purifying selection (ω = 0.3).

Comparing the median values of ω from the simulations with the empirical estimate for the deer clade (ω = 0.5)

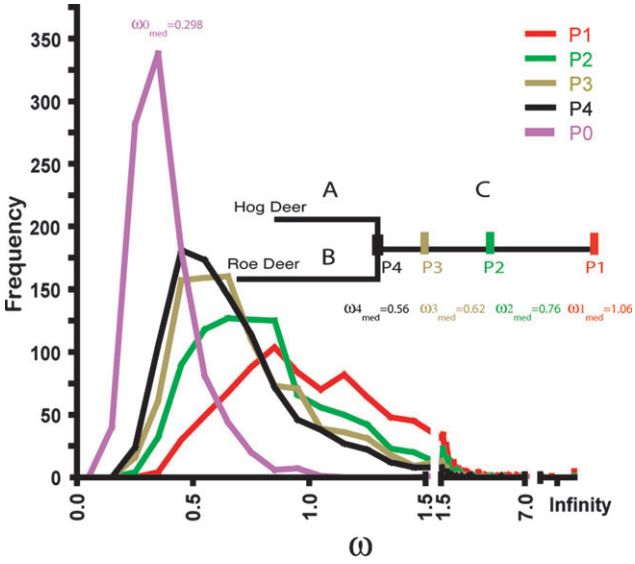

FIG. 4.—Subtree including seminal RNase sequences from hog deer and roe deer and the point where these sequences diverge from the okapi–giraffe clade (point P1). Distributions show the frequency of ω estimates depending upon the point on branch C where the true ω value changes from 0.3 to unity. One thousand data sets have been simulated under each of the shown conditions. The P0 curve shows a distribution of ω values for simulations assuming the true ω value is 0.3 for all branches. Median ω values for P0, P1, P2, P3, and P4 were 0.298, 1.06, 0.76, 0.62, and 0.56, respectively.

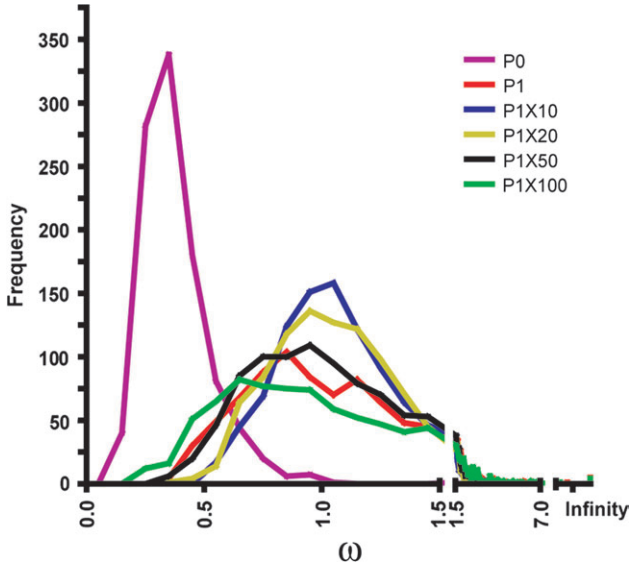

FIG. 5.—Distributions of ω estimates depending upon the length of the internal branch (C) of the deer clade. Branch length of internal branch C has been increased by factors of 10, 20, 50, and 100 when simulating the evolution of seminal RNase. A thousand data sets have been simulated under each of the shown conditions. P0, as before, shows a distribution of ω values for simulations assuming ω to be 0.3 for all branches.

indicated that the empirical value is closest to the simulations that assumed gene inactivation just before the speciation of hog deer and roe deer ($\omega_{med} = 0.56$ for P4; see fig. 4). However, the empirically measured value is within the confidence interval of all simulations, indicating that we cannot constrain well the timing of the inactivation using this approach. In fact, the 5 active site replacements shared by both deer sequences (fig. 2) can be viewed as stronger evidence that the gene inactivation took place long before the hog and roe deer diverged than the estimates of ω.

We then asked whether the small number of substitutions in the seminal RNase lineage was responsible for the inaccurate estimates of ω. To this end, we asked if increasing the expected number of mutations in the deer lineage (by lengthening the branches to mimic a more ancient divergence or a faster rate of evolution than estimated from the data) that occur after loss of purifying selection would alter the width of the ω distribution. To do this, we repeated the simulations but increased the branch lengths for the deer clade by factors of 10, 20, 50, and 100. The ω distribution narrows when the branch is 10-fold longer, but starts to widen again with further increases in the branch length (fig. 5). The distribution is substantially worse when it is 100 times the true lengths, suggesting that such a large number of mutations have resulted in saturation. It is important to note that confidence in ω estimates improves with time up to a factor of 20 compared with the empirically measured branch lengths. Thus, the use of ML estimates of ω alone is unlikely to provide convincing support for a hypothesis of gene inactivation unless the inactivation is more ancient than the ~19 Myr divergence of the deer examined here. Obviously, increased taxon sampling would increase the power of this approach, but there are likely to be fundamental limits like the length of the sequence.

Gene Conversion Did Not Have a Major Impact on Seminal RNase Evolution

Interlocus gene conversion is the most plausible mechanism for pseudogene reactivation when multiple lesions are present. Presumably, a functional paralog remaining in the genome would be the source of the information needed for reactivation (in this case, both pancreatic RNase A and brain RNase are available). Only a single narrative involving a gene conversion in the coding region has the potential to repair a defective seminal RNase pseudogene that also can be reconciled with the RNase gene tree. This gene conversion narrative would involve shifting the occurrence of one or more of the probable inactivating mutations in the lesser kudu to a time prior to the divergence of the kudu and bovine lineage, followed by repair due to gene conversion in the bovine lineage. Clearly, this narrative is less parsimonious than a narrative that postulates that the inactivating mutations in the kudu occurred after the kudu diverged from the bovine lineage. However, evidence for gene conversion involving the functional paralogs, either with the potential to repair a mutation or with no obvious functional consequence, would have general implications for the plausibility of pseudogene reactivation as a mechanism for the origin of genes with novel functions.

The spectrum of gene conversion tract lengths in humans is highly variable and has a relatively short mean tract length, with estimates of ~50 bp (Bosch et al. 2004; Jeffreys and May 2004). Although there is likely to be both among-species and among-locus variation in the rate of gene conversion and the mean tract length, the human data

are likely to provide information about the plausible tract length spectrum. Although the relatively high degree of divergence between seminal RNase and the other RNase genes may have an impact on this tract length spectrum, we felt that the available data on gene conversion tracts indicated that we should expect short tracts where the functional bovine seminal RNase would appear more similar to a functional paralog (RNase A or brain RNase genes) than to the seminal RNase pseudogenes.

To identify sites of this type, we took advantage of the fact that all sites are independent in a maximum parsimony analysis (each site can support trees that differ from the optimal tree). Briefly, the parsimony tree lengths for all characters were measured on the probable "true" tree (fig. 1) and alternative trees that place the bovine lineage within the paralogs ("gene conversion tree"; see Supplementary Material online). Five sites have a shorter parsimony tree length supporting one of the gene conversion trees over the probable true tree. This provided a limited set of candidate sites for gene conversion. The candidate sites, however, were spread throughout the sequence and clearly do not represent a single gene conversion tract (see Supplementary Material online). Further, they are distinct from the position of the kudu lesions, and in fact, the sites are in positions distinct from all of the observed lesions in the cited species.

It is also important to note that the candidate sites are not active site amino acids. Because the candidate sites can also be explained by homoplastic single-base substitutions, it seems reasonable to conclude that interparalog gene conversion has had little or no impact on the evolution of bovine seminal RNase.

Conclusion

This paper introduces a new approach based on crystallographic data for assessing whether purifying selection acted on ancestral genes. Applied to the seminal RNase gene family, the approach suggests inferences consistent with those made by classical analyses based upon estimates of ω. Both inferences are contrary to the inference, frequently mentioned in the literature, that bovine seminal RNase arose in modern ox through the reactivation of an ancestral pseudogene. Instead, it appears that seminal RNase evolved under selective constraints through much of its history, independently suffering lesions that rendered it a pseudogene in the external leaves of multiple ruminants, and underwent an episode of rapid sequence evolution, presumably adaptive, in the lineage connecting modern oxen with their immediate ancestor.

One virtue of combining these 2 types of analysis is that they draw on quite different data sets. Further, they speak to the status of a gene as a potential pseudogene in the event that a chance lesion, like a frameshift or removal of a residue known to be essential for catalysis, does not. Common to both is the need to infer ancestral sequences and the changes along individual branches from extant sequences. Such inferences include uncertainties that are well understood in the community, as well as those that are less well understood (Zhang and Nei 1997; Williams et al. 2006). The crystallographic analysis relies, however, on independent concepts of the physicochemical properties necessary for protein folding and function. Models of amino acid replacement (Jones et al. 1992; Koshi and Goldstein 1998; Whelan and Goldman 2001) also incorporate physicochemical information, but only in an implicit manner reflecting empirical information. Whereas efforts to incorporate 3-dimensional structural information in evolutionary models have been productive for some time (Benner 1989; Thorne et al. 1996; Goldman et al. 1998; Pollock et al. 1999; Fornasari et al. 2002; Robinson et al. 2003; Rodrigue et al. 2005; Yu and Thorne 2006), progress in both sequence genomics and structural genomics is required for these to be universally applicable.

Obviously, sufficient time and consequently a sufficient number of evolutionary events (e.g., nucleotide substitutions and/or amino acid replaces) are required to draw inferences for any segment in a tree. For example, the 19 Myr (for a divergence in time of 38 Myr) separating hog and roe deer was sufficient to be associated with 13 inferred amino acid replacements. This is both sufficient to allow the crystallographic approach to identify the pseudogene status of the hog deer and roe deer seminal RNase genes and is consistent with the number of replacements expected during drift in a typical mammalian lineage. This time period was not, however, sufficiently long to allow estimates of ω to persuasively allow inference of the same status. Indeed, our data suggest that sequence length and taxon sample may limit these inferences as well, as these create large variances for ω.

These examples, combined with the high variance in the estimate of ω found by our simulations, illustrate the need for multiple approaches to complement classical approaches based on the estimation of ω when evaluating the possibilities of constrained evolution, neutral evolution, and adaptive evolution in ancestral lineages. Such approaches can even include experiments, through the resurrection of ancestral proteins within a lineage for study in the laboratory, a field now coming to be called "paleogenetics" (Jermann et al. 1995; Chandrasekharan et al. 1996; Messier and Stewart 1997; Golding and Dean 1998; Zhang et al. 2000, 2002; Thornton 2004; Thomson et al. 2005; Benner et al. 2006; Sassi et al. in press). Once resurrected, ancestral proteins can be studied in the laboratory, allowing the evaluation of their biochemical properties via functional assays. These results may have clear implications for whether changes were neutral or adaptive. Paleogenetic studies become especially important when the confidence in ω values is low.

The combination of analyses focused on the estimation of ω, analyses based upon protein structure and simulations can be interpreted as strong support for a multiple inactivation model for seminal RNase. A corollary of this conclusion is the inference that the ancestral seminal RNase genes had a function and that the function involved the specification of a protein. Although there are examples of pseudogenes that have a regulatory role (Hirotsune et al. 2003), it is difficult to construct a model in which a regulatory pseudogene (i.e., one acting at the RNA level) would show the patterns of conservation evident in seminal RNase (specifically constraints on nonsynonymous sites, especially those that alter residues buried in the protein structure). This finding is suggestive of an environmental

change that rendered seminal RNase either irrelevant or disadvantageous in a variety of ruminants, with the exception of the oxen where the gene shifted to novel function. Further understanding of the ancestral function for seminal RNase in ruminants may require the combination of biochemical experiments from reconstructed ancestors with any information about the ancestral patterns of gene expression that can be obtained. Regardless of the basis for the multiple losses, it is clear that seminal RNase should no longer be viewed as an example of pseudogene inactivation.

Supplementary Material

Supplementary materials are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are indebted to the National Aeronautics and Space Administration Exobiology program for support of this research. We are also indebted to Ross P. Davis for invaluable computer support.

Literature Cited

Asenjo AB, Rim J, Oprian DD. 1994. Molecular determinants of human red/green color discrimination. Neuron. 12:1131–1138.

Balakirev ES, Ayala FJ. 2003. Pseudogenes: are they "junk" or functional DNA? Annu Rev Genet. 37:123–151.

Benner SA. 1989. Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. Adv Enzyme Regul. 28:219–236.

Benner SA, Sassi SO, Gaucher EA. 2006. Molecular paleoscience: systems biology from the past. Adv Enzymol Relat Areas Mol Biol. 75:1–132.

Berisio R, Sica F, De Lorenzo C, Di Fiore A, Piccoli R, Zagari A, Mazzarella L. 2003. Crystal structure of the dimeric unswapped form of bovine seminal ribonuclease. FEBS Lett. 554: 105–110.

Bielawski JP, Yang Z. 2003. Maximum likelihood methods for detecting adaptive evolution after gene duplication. J Struct Funct Genomics. 3:201–212.

Bosch E, Hurles ME, Navarro A, Jobling MA. 2004. Dynamics of a human interparalog gene conversion hotspot. Genome Res. 14:835–844.

Braun EL, Halpern AL, Nelson MA, Natvig DO. 2000. Large-scale comparison of fungal sequence information: mechanisms of innovation in Neurospora crassa and gene loss in Saccharomyces cerevisiae. Genome Res. 10:416–430.

Breukelman HJ, van der Munnik N, Kleineidam RG, Furia A, Beintema JJ. 1998. Secretory ribonuclease genes and pseudogenes in true ruminants. Gene. 212:259–268.

Brosius J, Gould SJ. 1992. On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". Proc Natl Acad Sci USA. 89:10706–10710.

Burnham KP, Anderson DR, Burnham KP. 2002. Model selection and multimodel inference: a practical information-theoretic approach. New York: Springer.

Capasso S, Giordano F, Mattia CA, Mazzarella L, Zagari A. 1983. Refinement of the structure of bovine seminal ribonuclease. Biopolymers. 22:327–332.

Chandrasekharan UM, Sanker S, Glynias MJ, Karnik SS, Husain A. 1996. Angiotensin II-forming activity in a reconstructed ancestral chymase. Science. 271:502–505.

Doxiadis GG, van der Wiel MK, Brok HP, de Groot NG, Otting N, 't Hart BA, van Rood JJ, Bontrop RE. 2006. Reactivation by exon shuffling of a conserved HLA-DR3-like pseudogene segment in a New World primate species. Proc Natl Acad Sci USA. 103:5864–5868.

Endo T, Ikeo K, Gojobori T. 1996. Large-scale search for genes on which positive selection may operate. Mol Biol Evol. 13:685–690.

Fornasari MS, Parisi G, Echave J. 2002. Site-specific amino acid replacement matrices from structurally constrained protein evolution simulations. Mol Biol Evol. 19:352–356.

Golding GB, Dean AM. 1998. The structural basis of molecular adaptation. Mol Biol Evol. 15:355–369.

Goldman N, Thorne JL, Jones DT. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics. 149:445–458.

Harrison PM, Gerstein M. 2002. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. J Mol Biol. 318:1155–1174.

Harrison PM, Zheng D, Zhang Z, Carriero N, Gerstein M. 2005. Transcribed processed pseudogenes in the human genome: an intermediate form of expressed retrosequence lacking protein-coding ability. Nucleic Acids Res. 33:2374–2383.

Hernandez Fernandez M, Vrba ES. 2005. A complete estimate of the phylogenetic relationships in Ruminantia: a dated species-level supertree of the extant ruminants. Biol Rev Camb Philos Soc. 80:269–302.

Hirotsune S, Yoshida N, Chen A, Garrett L, Sugiyama F, Takahashi S, Yagami K, Wynshaw-Boris A, Yoshiki A. 2003. An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. Nature. 423:91–96.

Hughes AL. 1994. The evolution of functionally novel proteins after gene duplication. Proc Biol Sci. 256:119–124.

Hurles M. 2004. Gene duplication: the genomic trade in spare parts. PLoS Biol. 2:E206.

Jeffreys AJ, May CA. 2004. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. Nat Genet. 36:151–156.

Jermann TM, Opitz JG, Stackhouse J, Benner SA. 1995. Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. Nature. 374:57–59.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275–282.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 22:2577–2637.

Katju V, Lynch M. 2006. On the formation of novel genes by duplication in the Caenorhabditis elegans genome. Mol Biol Evol. 23:1056–1067.

Keese PK, Gibbs A. 1992. Origins of genes: "big bang" or continuous creation? Proc Natl Acad Sci USA. 89:9489–9493.

Kim JS, Soucek J, Matousek J, Raines RT. 1995. Catalytic activity of bovine seminal ribonuclease is essential for its immunosuppressive and other biological activities. Biochem J. 308(Pt 2): 547–550.

Kimura M, Ota T. 1974. On some principles governing molecular evolution. Proc Natl Acad Sci USA. 71:2848–2852.

Kleineidam RG, Jekel PA, Beintema JJ, Situmorang P. 1999. Seminal-type ribonuclease genes in ruminants, sequence conservation without protein expression? Gene. 231:147–153.

Kondrashov FA, Koonin EV. 2003. Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. Trends Genet. 19:115–119.

Koshi JM, Goldstein RA. 1998. Models of natural mutations including site heterogeneity. Proteins. 32:289–295.

Laccetti P, Portella G, Mastronicola MR, Russo A, Piccoli R, D'Alessio G, Vecchio G. 1992. In vivo and in vitro growth-inhibitory effect of bovine seminal ribonuclease on a system of rat thyroid epithelial transformed cells and tumors. Cancer Res. 52:4582–4586.

Lee JE, Raines RT. 2005. Cytotoxicity of bovine seminal ribonuclease: monomer versus dimer. Biochemistry. 44:15760–15767.

Liberles DA, Wayne ML. 2002. Tracking adaptive evolutionary events in genomic sequences. Genome Biol. 3: reviews1018.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science. 290:1151–1155.

Lynch M, O'Hely M, Walsh B, Force A. 2001. The probability of preservation of a newly arisen gene duplicate. Genetics. 159: 1789–1804.

Maddison WP. 1997. Gene trees in species trees. Syst Biol. 46: 523–536.

Mahon AS. 2004. A molecular supertree of the Artiodactyla. In: Bininda-Emonds ORP, editor. Phylogenetic supertrees: combining information to reveal the tree of life. Boston: Kluwer Academic Publishers.

Marshall CR, Raff EC, Raff RA. 1994. Dollo's law and the death and resurrection of genes. Proc Natl Acad Sci USA. 91:12283–12287.

Mazzarella L, Capasso S, Demasi D, Di Lorenzo G, Mattia CA, Zagari A. 1993. Bovine seminal ribonuclease: structure at 1.9 Å resolution. Acta Crystallogr Sect D. 49:389–402.

Merlino A, Vitagliano L, Sica F, Zagari A, Mazzarella L. 2004. Population shift vs induced fit: the case of bovine seminal ribonuclease swapping dimer. Biopolymers. 73:689–695.

Messier W, Stewart CB. 1997. Episodic adaptive evolution of primate lysozymes. Nature. 385:151–154.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.

Newcomb RD, Campbell PM, Ollis DL, Cheah E, Russell RJ, Oakeshott JG. 1997. A single amino acid substitution converts a carboxylesterase to an organophosphorus hydrolase and confers insecticide resistance on a blowfly. Proc Natl Acad Sci USA. 94:7464–7468.

Ohno S. 1970. Evolution by gene duplication. Berlin (Germany): Springer-Verlag.

Ohno S. 1984. Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. Proc Natl Acad Sci USA. 81:2421–2425.

Ota T, Nei M. 1995. Evolution of immunoglobulin VH pseudogenes in chickens. Mol Biol Evol. 12:94–102.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. Mol Biol Evol. 5:568–583.

Perutz MF. 1983. Species adaptation in a protein molecule. Mol Biol Evol. 1:1–28.

Pollock DD, Taylor WR, Goldman N. 1999. Coevolving protein residues: maximum likelihood identification and relationship to structure. J Mol Biol. 287:187–198.

Posada D, Buckley TR. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst Biol. 53:793–808.

Posada D, Crandall KA. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics. 14:817–818.

Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL. 2003. Protein evolution with dependence among codons due to tertiary structure. Mol Biol Evol. 20:1692–1704.

Rodrigue N, Lartillot N, Bryant D, Philippe H. 2005. Site interdependence attributed to tertiary structure in amino acid sequence evolution. Gene. 347:207–217.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.

Roth C, Betts MJ, Steffansson P, Saelensminde G, Liberles DA. 2005. The Adaptive Evolution Database (TAED): a phylogeny based tool for comparative genomics. Nucleic Acids Res. 33:D495–D497.

Shimodaira H, Hasegawa M. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 16:1114–1116.

Sica F, Di Fiore A, Zagari A, Mazzarella L. 2003. The unswapped chain of bovine seminal ribonuclease: crystal structure of the free and liganded monomeric derivative. Proteins. 52:263–271.

Sinatra F, Callari D, Viola M, Longombardo MT, Patania M, Litrico L, Emmanuele G, Lanteri E, D'Alessandro N, Travali S. 2000. Bovine seminal RNase induces apoptosis in normal proliferating lymphocytes. Int J Clin Lab Res. 30: 191–196.

Soucek J, Marinov I, Benes J, Hilgert I, Matousek J, Raines RT. 1996. Immunosuppressive activity of bovine seminal ribonuclease and its mode of action. Immunobiology. 195: 271–285.

Swofford DL. 2001. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.

Thomson JM, Gaucher EA, Burgan MF, De Kee DW, Li T, Aris JP, Benner SA. 2005. Resurrecting ancestral alcohol dehydrogenases from yeast. Nat Genet. 37:630–635.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Thorne JL, Goldman N, Jones DT. 1996. Combining protein evolution and secondary structure. Mol Biol Evol. 13:666–673.

Thornton JW. 2004. Resurrecting ancient genes: experimental analysis of extinct molecules. Nat Rev Genet. 5:366–375.

Trabesinger-Ruef N, Jermann T, Zankel T, Durrant B, Frank G, Benner SA. 1996. Pseudogenes in ribonuclease evolution: a source of new biomacromolecular function? FEBS Lett. 382:319–322.

Vescia S, Tramontano D, Augusti-Tocco G, D'Alessio G. 1980. In vitro studies on selective inhibition of tumor cell growth by seminal ribonuclease. Cancer Res. 40:3740–3744.

Walsh JB. 1995. How often do duplicated genes evolve new functions? Genetics. 139:421–428.

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18:691–699.

Williams PD, Pollock DD, Blackburne BP, Goldstein RA. 2006. Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comput Biol. 2:e69.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13:555–556.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 15:568–573.

Yang ZH, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15:496–503.

Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics. 141:1641–1650.

Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol. 46:409–418.

Yu J, Thorne JL. 2006. Testing for spatial clustering of amino acid replacements within protein tertiary structure. J Mol Evol. 62:682–692.

Zhang JZ. 2003. Evolution by gene duplication: an update. Trends Ecol Evol. 18:292–298.

Zhang J, Dyer KD, Rosenberg HF. 2000. Evolution of the rodent eosinophil-associated RNase gene family by rapid gene sorting and positive selection. Proc Natl Acad Sci USA. 97:4701–4706.

Zhang J, Nei M. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J Mol Evol. 44(Suppl 1):S139–S146.

Zhang J, Rosenberg HF, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci USA. 95:3708–3713.

Zhang J, Zhang YP, Rosenberg HF. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat Genet. 30:411–415.

Michele Vendruscolo, Associate Editor

Accepted January 26, 2007