The Effect of Species Representation on the Detection of Positive Selection in Primate Gene Data Sets

Ross M. McBee,1 Shea A. Rozmiarek,2 Nicholas R. Meyerson,1 Paul A. Rowley,1 and Sara L. Sawyer*,1
1Department of Molecular Biosciences, University of Texas at Austin, Austin, TX
2Liberty Hill High School, Liberty Hill, TX
*Corresponding author: E-mail: saras@mail.utexas.edu.
Associate editor: Willie Swanson

Abstract

Over evolutionary time, both host- and virus-encoded genes have been continually selected to modify their interactions with one another. This has resulted in the rapid evolution of the specific codons that govern the physical interactions between host and virus proteins. Virologists have discovered that these evolutionary signatures, acquired in nature, can provide a shortcut in the functional dissection of host–virus interactions in the laboratory. However, the use of evolution studies in this way is complicated by the fact that many nonhuman primate species are endangered, and biomaterials are often difficult to acquire. Here, we assess how the species representation in primate gene data sets affects the detection of positive natural selection. Our results demonstrate how targeted primate sequencing projects could greatly enhance research in immunology, virology, and beyond.

Key words: positive selection, simian primates, host–virus arms race, host–virus coevolution, HIV.

The evolution of human genes can be studied in multiple ways (Vitti et al. 2013). One approach uses comparisons of gene orthologs from humans and nonhuman primates to analyze selection over long evolutionary timescales. The main metric used is dN/dS, which summarizes the rates of nonsynonymous (dN) and synonymous (dS) DNA substitutions in gene sequence (Hurst 2002; Nielsen 2005; Kelley and Swanson 2008). Codons where dN/dS > 1 have experienced multiple rounds of natural selection in favor of nonsynonymous mutations. Recurrent positive selection can be driven by many phenomena, with several well-known examples being sperm–egg interactions, mate choice, and host–environment interactions. In addition, viruses and their hosts coevolve over long periods of time and, as a result, the dN/dS > 1 signature can often be detected in gene regions corresponding to physical interaction interfaces between host and virus proteins. The identification of codons with dN/dS > 1 in primate genes has recently become important in guiding genetic studies in the HIV field, having been particularly powerful in dissecting the physical interactions of several human immunity proteins with HIV (Sawyer et al. 2005; Gupta et al. 2009; McNatt et al. 2009; Lim et al. 2010, 2012; Duggal et al. 2011; Lagouette et al. 2012; Compton and Emerman 2013). The identification of this evolutionary signature has been recognized as a shortcut in the laborious functional dissection of host–virus interactions, and is now being used to characterize the interplay between human proteins and other types of viruses as well (Kerns et al. 2008; Elde et al. 2009; Demogines, Farzan, et al. 2012; Demogines, Truong, et al. 2012; Kaelber et al. 2012; Mitchell et al. 2012; Patel et al. 2012; Demogines et al. 2013).

The main limitation to using this analysis is the acquisition of appropriate primate sequence data sets. There are currently nine available simian primate genome projects available through the UCSC genome browser (http://www.genome.ucsc.edu/, last accessed August, 2014), but the acquisition of additional primate sequences is complicated by the fact that many nonhuman primate species are endangered, and even the purchase of immortalized cell lines derived from these species can require a federally issued permit. Because the analysis of dN/dS has become so useful in guiding the genetic study of human genes that interact with viruses, we wished to investigate how many primate sequences are required to reliably detect positive selection.

To test the sensitivity of positive selection analyses when using different primate data sets, we reanalyzed data sets that our group has previously generated for 11 different genes (XLF, XRCC4, MAP4, NBS1, CtIP, WNK1, POLL, NUP153, RANBP2, IBTK, and NUP98/96; supplementary table S1, Supplementary Material online; Demogines et al. 2010; Meyerson et al. 2014). These 11 data sets each consist of 20 orthologous sequences from a matched set of primate species (fig. 1A). In the studies where these data sets were generated, genes were chosen for sequencing based on higher-than-normal levels of protein divergence, or because they are known to encode proteins important for viral lifecycles. Therefore, this is not a random set of genes, but rather a set that is enriched for genes that might be experiencing positive selection.

We generated ten pruned trees representing subsets of these 20 species (fig. 1A). The first four trees were made to reflect the history of primate genome sequencing projects. For instance, the first tree that we analyzed was a three-species tree of human, chimpanzee, and rhesus macaque, representing the first three primate genomes sequenced (Mikkelsen et al. 2005; Rhesus Macaque Genome
Sequencing and Analysis Consortium et al. 2007). The four-species tree also included marmoset, the fourth primate to have its genome sequenced. We then added in the fifth sequenced primate species, Sumatran orangutan (Locke et al. 2011), creating a five-species tree, and then white-cheeked gibbon and gorilla (Scally et al. 2012), creating a seven-species tree. We specifically chose this strategy in order to evaluate the power of early positive selection studies that only had access to a limited number of primate genomes (e.g., Clark et al. 2003; Nielsen et al. 2005; Rhesus Macaque Genome Sequencing and Analysis Consortium et al. 2007; Ortiz et al. 2009; Locke et al. 2011). Beyond this, we made, 10-, 12-, 14-, 15-, 16-, and 18- species subtrees (fig. 1A). The species included in these trees were chosen so that “tree length” scales approximately linearly with the number of species included (fig. 1B). Tree length is the sum of the branch lengths along the tree or, in other words, the average number of nucleotide substitutions per site in an alignment. For any given tree, we find higher tree length in some of our data sets than others (fig. 1B). For instance, the DNA repair gene XLF has the highest level of sequence divergence, and the nuclear pore gene NUP98/96 has the lowest level of divergence. This set of trees was then used to test the effects of species representation on the detection of positive selection.

Because the PAML software suite is now commonly used in virology research, we focused on the detection of positive selection using codon models implemented in PAML’s codeml program (Yang 1997, 2007). These models test for codons that have experienced positive selection for nonsynonymous changes pervasively over the tree being analyzed. They assume that the $dN/dS$ value is constant through time, so an estimate at a site is obtained by essentially averaging the signal over all branches. Each alignment was fit to the codon models M7 (null model, codon values of $dN/dS$ fit to a beta distribution bounded between $0 < dN/dS < 1$) and M8 (positive selection model, similar to M7 but with one extra site class assigned at $dN/dS > 1$). We made comparison to the M7 null model, although M8a is another commonly used null model (Swanson et al. 2003). A likelihood ratio test was then used to determine whether the null model (M7) could be rejected in favor of the model of positive selection (M8). We performed likelihood ratio tests between M7 and M8 for all 11 gene data sets, using each of the 11 possible trees. These genes fell into three classes. First, six genes converged on significant rejection of the null model ($P < 0.05$) as more species were added (fig. 2A). One of these, XRCC4, reached significance after only four species, and stayed significant as more species were added. On the other extreme, MAP4 did not reach significance until the 20 species data set. We conclude that more species allow an increased possibility of detecting positive selection, although some genes are more sensitive than others. A second set of two genes (WNK1 and XLF) did not reach the $P < 0.05$ significance threshold using any of the trees tested (fig. 2B). Further, for these genes there is no clear trend toward significance, suggesting that the null hypothesis would never be rejected even if more sequences were added.

Finally, for a third set of three genes, the null hypothesis is rejected (or very nearly so in the case of NBS1) with smaller data sets, but then the likelihood ratio test loses significance as more species are added (fig. 2C). These genes might be experiencing episodic positive selection only along specific lineages, a signal that PAML might miss (supplementary table S2, Supplementary Material online). We next analyzed the full data sets for each of these three genes with MEME (Murrell et al. 2012). MEME utilizes a model that allows for episodic selection acting on codon sites. Indeed, MEME produced a pattern of increased (or constant) power to detect positive selection in these three genes as more data were added (fig. 2D and E). Sophisticated tests like MEME should
be included in future studies in order to also catch codons experiencing lineage- and clade-specific positive selection.

In the PAML analyses shown in figure 2A–C, there is stochastic noise in the patterns observed. Based on this, we wished to test how likely it would be to get a false signature of positive selection. We next examined more closely the effects of primate species choice on the evolution of one gene from each of the three classes of evolution that were described (RANBP2, WNK1, and IBTK). We again examined likelihood ratio test results for 3-, 6-, 9-, 12-, 14-, 16-, and 18-species trees, but this time we randomly chose species from our 20-species collection to create ten trees of each of these sizes (see Materials and Methods). Those trees and the corresponding sequences were then fit to M7 and M8, and the P-value of the likelihood ratio test is shown (fig. 2A–E). These graphs each represent data for a single gene from panels A to C. In this case, ten alternate species sets were chosen for each number of species shown on the x axis. Each of these trees was used to evaluate M7 and M8, and the P-value is calculated for each model comparison.

If the null model (M7) is rejected in favor of the model of positive selection (M8), the individual codons assigned to the dN/dS > 1 site class can be used to guide genetic studies (Sawyer et al. 2005; Gupta et al. 2009; McNatt et al. 2009; Lim et al. 2010, 2012; Duggal et al. 2011; Schaller et al. 2011; Laguette et al. 2012). The logic is that nonsynonymous mutations in these codons impact function, otherwise selection would not be acting on these sites. For the six genes that pass the likelihood ratio test (fig. 2F) we examined how increasingly rich data sets affect the fraction of codons assigned to the dN/dS > 1 class in M8. In general, this value stabilizes by the time we have included these 20 simian primate species.
The proper assignment of each codon to this class can also be evaluated using a posterior probability (Yang et al. 2005). For instance, if a codon has a posterior probability of $P = 0.95$, there is a 95% chance that this codon is correctly assigned to the dN/dS > 1 class. We then looked at the fraction of sites in each data set assigned to the dN/dS > 1 class with a posterior probability $P \geq 0.95$ (fig. 3B). These values also stabilize by the time we have included these 20 simian primate species. The specific codons identified are illustrated for two of these genes, RANBP2 (fig. 3C) and Polλ (fig. 3D). Interestingly, it appears that many sites identified with as few as 4–5 sequences, even when the likelihood ratio test has low statistical support, are often increasingly supported as more sequences are added. This surprising finding suggests that it may be worth functionally testing codons identified even if only a few sequences are available for analysis, and even before the likelihood ratio test has reached the rigorous $P < 0.05$ value. We summarize this data for all genes in supplementary table S3, Supplementary Material online, and note that there does not appear to be a correlation between the number of sites under positive selection in a gene and the number of species required to reach significance. However, we do note a general trend that the number of selected sites grows as the number of sequences increases, suggesting that more sequences are always better for studies with the goal of identifying sites to be tested in functional assays.

In conclusion, we find that positive selection can be adequately characterized in primate genes with a 20-species data set composed of 8 hominoids, 8 Old World monkeys, and 4 New World monkeys (fig. 1A). Tree length, a measure of overall divergence in an alignment, can also be used as a guide, and our 20-species data sets have tree lengths between approximately 0.3 and 0.65 (fig. 1B). This study should serve both evolutionary biologists and virologists who are interested in the molecular evolution of genes during the course of simian primate speciation. It also underscores the need for more sequenced primates genomes, which would alleviate the burden on individual researchers to obtain these precious primate biomaterials.

**Materials and Methods**

**Primate Sequence Data Sets**

All sequence alignments were previously generated by our lab (Demogines et al. 2010; Meyerson et al. 2014). A description of the genes analyzed is given in supplementary table S1, Supplementary Material online. The primate species analyzed are listed in supplementary table S4, Supplementary Material online.

**PAML Analysis**

Each gene alignment was fit to two codon models, M7 and M8, as implemented in PAML (Yang 1997). A likelihood ratio...
test was then performed, using 2 degrees of freedom, to assess whether M8 (permitting some codons to evolve under positive selection) gives a significantly better fit to the data than M7 (positive selection not allowed). The Bayes Empirical Bayes approach was then used to calculate posterior probability that each codon is properly assigned to the dN/dS > 1 site class (Yang et al. 2005). This entire protocol was performed for each of the 11 data sets, for each of the subtrees analyzed.

MEME Analysis
MEME (Murrell et al. 2012) was used at Datamonkey.org (Delport et al. 2010). A user tree was supplied and the automatic substitution model selection tool was employed. This was performed for each of the three data sets, for each of the subtrees analyzed.

Generation of Random Trees
The analysis shown in panels F–H of figure 2 required the generation of random trees. Ten random trees of each of the seven tree sizes (“number of species” on x axis) were generated for a total of 70 unique trees for each gene. To generate the random trees, each species in each of three primate clades (color coded in fig. 1A) was assigned an integer and random.org was used to generate ten unique sets of nonrepeating integers. For the first four of these sets of trees, 1, 2, 3, and then 4 species were chosen at random from each of the three primate clades resulting in trees containing a total of 3, 6, 9, and 12 species. However, because sequences were available from only four New World monkeys, tree sets past the 12 species size only introduced two new species at each step, one from each of the two other clades.

Supplementary Material
Supplementary tables S1–S4 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments
The authors thank Maryska Kaczmarek, Susan Rozmariak, and Alex Stabell and for critical reading of the manuscript. This work was supported by a grant from the National Institutes of Health (R01-GM-093086). N.R.M. is supported by a National Science Foundation Graduate Research Fellowship. S.L.S. is a Burroughs Wellcome Fund Investigator in the Pathogenesis of Infectious Disease.

References


