

Determining the Null Model for Detecting Adaptive Convergence from Genomic Data: A Case Study using Echolocating Mammals

Gregg W.C. Thomas¹ and Matthew W. Hahn^{*1,2}

¹School of Informatics and Computing, Indiana University, Bloomington

²Department of Biology, Indiana University, Bloomington

*Corresponding author: E-mail: mwh@indiana.edu.

Associate editor: David Irwin

Abstract

Convergent evolution occurs when the same trait arises independently in multiple lineages. In most cases of phenotypic convergence such transitions are adaptive, so finding the underlying molecular causes of convergence can provide insight into the process of adaptation. Convergent evolution at the genomic level also lends itself to study by comparative methods, although molecular convergence can also occur by chance, adding noise to this process. Parker et al. studied convergence across the genomes of several mammals, including echolocating bats and dolphins (Parker J, Tsagkogeorga G, Cotton JA, Liu Y, Provero P, Stupka E, Rossiter SJ. 2013. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* 502:228–231). On the basis of a null distribution of site-specific likelihood support (SSLS) generated using simulated topologies, they concluded that there was evidence for genome-wide adaptive convergence between echolocating taxa. Here, we demonstrate that methods based on SSLS do not adequately measure convergence, and reiterate the use of an empirical null model that directly compares convergent substitutions between all pairs of species. We find that when the proper comparisons are made there is no surprising excess of convergence between echolocating mammals, even in sensory genes.

Key words: convergence, echolocation, adaptation, parallel evolution.

Convergent evolution provides a unique opportunity to study adaptation, but this phenomenon is not yet well understood at the molecular level (Stern 2013). Finding the molecular causes of convergent phenotypes offers the ability both to study the basis for adaptive evolution and to link phenotype and genotype, making it an exciting crucible for evolutionary biology. Convergence, if widespread, could also pose huge problems for the study of molecular evolution because most phylogenetic methods do not currently accommodate high levels of convergent evolution. This means that studies of species trees (e.g., Castoe et al. 2009), protein evolution (e.g., Williams et al. 2006), positive selection (e.g., Zhang et al. 1997), and many other areas could be adversely affected by the presence of widespread convergent evolution.

Recently, Parker et al. (2013) examined substitutions in 2,326 genes among 21 mammals, reporting “extensive convergent changes” between taxa with independent origins of echolocation. As a test for convergence, Parker et al. (2013) calculated the difference in site-specific likelihood support (Δ SSLS) between the well-supported species phylogeny relating the mammal species and a hypothetical phylogeny in which the echolocating species (four bats and one dolphin) form a monophyletic group. Evidence for adaptive convergence was reported to come from a comparison of these results against a null distribution of the same measure

obtained from simulations. This approach is problematic for a number of reasons, the most important of which is that Δ SSLS is not actually a test for convergence. Convergent topologies are one symptom of convergent molecular evolution, but there are many factors that can generate differences in site-specific likelihood scores across a tree. On the basis of their methods, these authors claimed to find genome-wide signals of adaptive convergence between dolphins and bats, including in sensory genes that may be important for echolocation. If this conclusion is true, it has wide reaching implications for molecular evolution.

In addition to not directly testing for convergence, the analysis of Parker et al. (2013) used simulations to determine whether the observed levels of “convergence” are statistically significant. However, it is known that simulations do not accurately account for observed levels of background convergence (Lartillot et al. 2007; Castoe et al. 2009). A better null model to detect adaptive convergence arises naturally from the data based on the correlation between the number of convergent and divergent amino acid substitutions between all pairs of species, as it accounts for nonadaptive convergence and divergence (Castoe et al. 2009). This method also has the advantage of utilizing information from each pair of species in the full phylogeny, and provides a firm statistical basis to determine whether an excess of convergence is present.

Here, we re-examine patterns of genome-wide convergence among echolocating mammals using an approach that directly compares levels of convergence between all pairs of species in order to generate a null expectation. We find that signals of convergence between echolocating mammals do not exceed expectations, and that the observed convergence in sensory genes is not surprising in a larger genomic context. We also report several surprising details from the original analysis of Parker et al. (2013).

Results and Discussion

Assessing Levels of Overall Convergence

To assess the utility of the approach of Parker et al. (2013) in determining whether there is an excess level of convergence among echolocating species, we counted the number of convergent substitutions and divergent substitutions (those occurring at the same site but resulting in a different amino acid) in 6,400 orthologous genes among all pairs of species in a phylogeny of nine mammals (fig. 1a). These species include an echolocating microbat (*Myotis lucifugus*), a nonecholocating megabat (*Pteropus vampyrus*), and a bottlenose dolphin (*Tursiops truncatus*). Both our analysis and that of Parker et al. are limited to amino acid substitutions, and do not examine silent substitutions. Our analysis contains almost three times as many genes as Parker et al. (2013), providing additional power to detect convergence and suggesting that it should also be more informative about any particular class of convergent genes.

As expected based on the results in Castoe et al. (2009), there is a strong correlation between convergent and divergent substitutions (fig. 1b). The relative amount of

convergence between the two echolocating mammals (labeled in fig. 1b) does not exceed that found between any random pair of mammals. As an independent approach using empirical comparisons between species as a null model, the same distribution of Δ SLS values reported between echolocating mammals was found in a comparison between echolocating bats and the nonecholocating cow (Zou and Zhang, 2015). These results strongly suggest that there is no exceptional genomic signature indicative of adaptive convergence between echolocating species, or at least that there is just as much adaptive convergence between any two species of mammals regardless of the presence of any obvious convergent phenotypes.

In addition, we find that genes with convergent substitutions often overlap between pairwise comparisons, possibly because they are rapidly evolving. This overlap is important to consider when attributing genes with convergence to a specific trait because we would not expect genes that are responsible for echolocation to also be convergent between nonecholocating species. Parker et al. (2013) list 117 genes as convergent between echolocating bats and dolphins. However, they failed to perform similar comparisons among nonecholocating species, or between echolocating species and equally distant nonecholocating species. We examined convergent substitutions in our data set and found 1,372 genes with convergent changes between microbat and dolphin and 1,951 genes between microbat and the nonecholocating cow. We chose the microbat–cow comparison because cow is sister to dolphin in the tree used by Parker et al., allowing us to use two species pairs of equivalent time of divergence (fig. 1a; this comparison is also labeled in fig. 1b). Of the genes with convergence, 738 are overlapping between

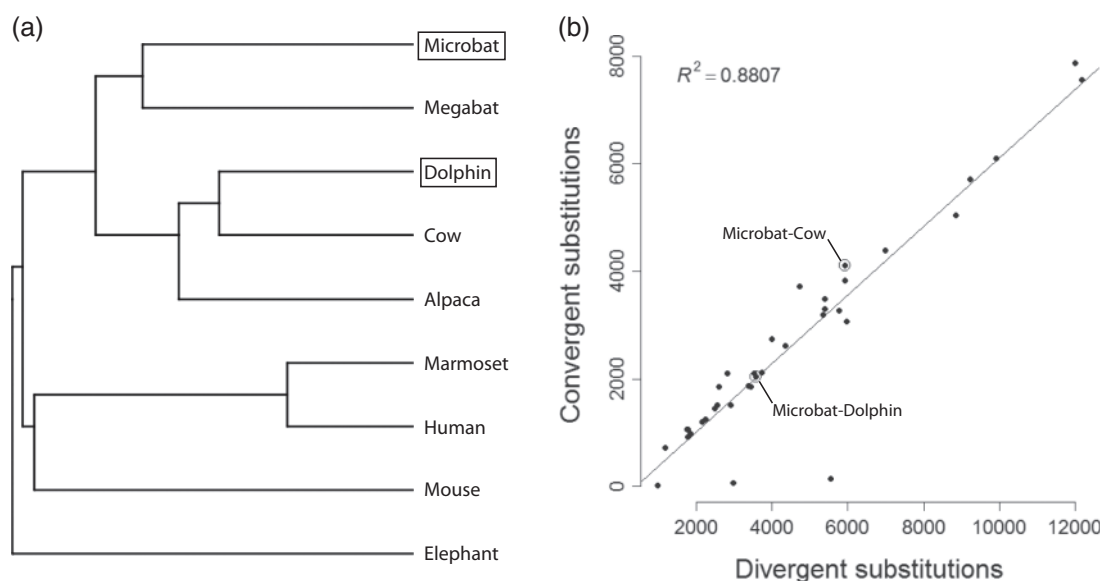


FIG. 1. Convergence between echolocating species does not exceed expectations. (a) The phylogenetic relationships among the mammal species studied here. Echolocating species are shown in boxes. (b) The number of convergent and divergent substitutions along external branches of the tree between all pairs of nine mammal species. The correlation ($R^2 = 0.88$) between these types of substitutions provides an appropriate expectation for convergence (Castoe et al. 2009). The circled dots represent comparisons of interest between microbat–dolphin and microbat–cow. The three dots at the bottom of the panel represent sister-taxon comparisons, for which it is nearly impossible to infer convergent changes (such comparisons were excluded from Castoe et al. 2009).

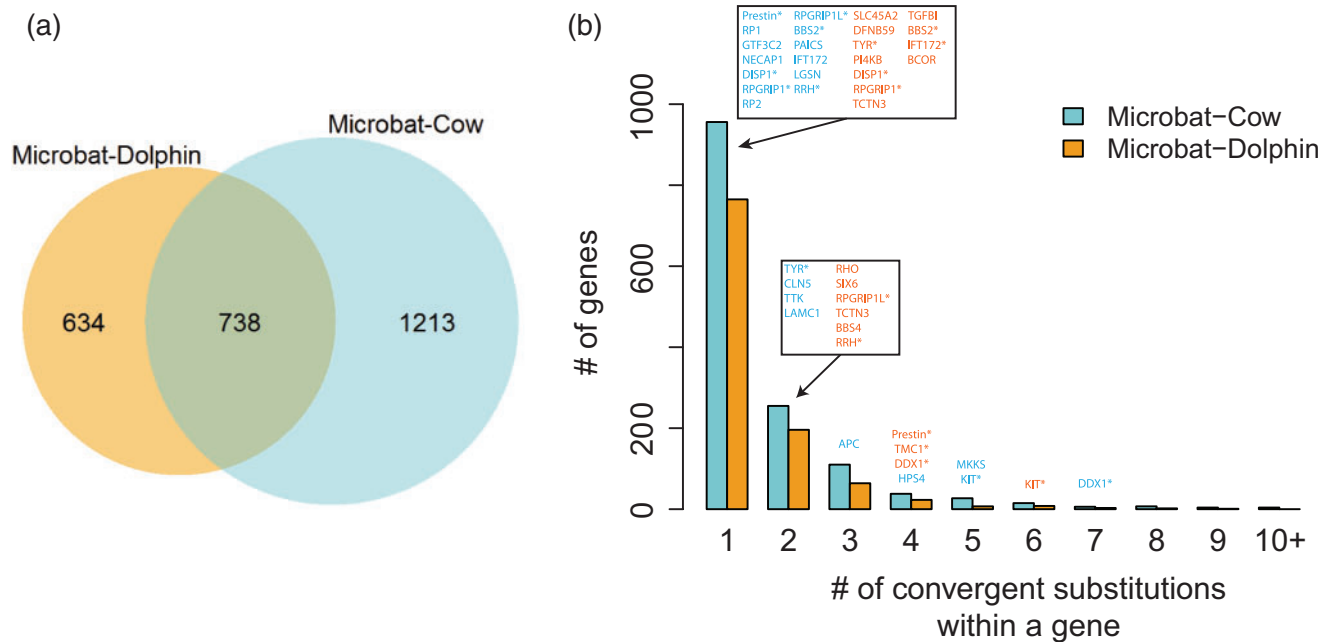


Fig. 2. Convergent substitutions between both microbat–dolphin and microbat–cow. (a) Of the 1,372 genes found with convergent changes between microbat and dolphin, 738 also had convergent changes between microbat and the nonecholocating cow. (b) The number of genes with multiple convergent substitutions in comparisons between microbat–dolphin and microbat–cow. Sensory genes with convergent substitutions are assigned to their corresponding bin indicating the number of convergent substitutions they contain in either comparison. Genes with stars next to their name have convergent substitutions in both comparisons. There are 22 sensory genes in each comparison, indicating no surprising convergence between echolocators.

the two comparisons, which is over half of the genes with convergent changes between echolocators (fig. 2a). We find only 14 genes that are uniquely convergent between the echolocating species when considering all other pairwise comparisons, and none of these are annotated with a function in any sensory system. We also find signatures of positive selection overlapping with those of convergence, but again this is not limited to the echolocating lineages. Of the 1,372 genes we identify with convergent substitutions between microbat and dolphin, 91 (6.63%) show signs of positive selection on both echolocating branches. When considering the 1,951 convergent genes between microbat and the nonecholocating cow, we find 157 (8.05%) that also show significant signatures of positive selection. This indicates that echolocating lineages do not show a statistical excess of convergent sites that have evolved under positive selection.

It could be argued that even within the genes that overlap among comparisons, signals of convergence may be stronger in one species pair than another. For instance, adaptive convergence in a gene may be associated with multiple convergent substitutions. To assess whether the strength of convergence is stronger in echolocating species, we searched for genes with multiple convergent substitutions. We find no excess of genes with multiple convergent substitutions between microbat and dolphin when compared with microbat and cow (fig. 2b), despite our scan having picked up hundreds of genes with such substitutions. If anything, the signal of convergence between microbat and cow appears to be stronger than that between microbat and dolphin (fig. 2b).

In the course of our search for genes with convergent substitutions, we found that only 19 of the 117 loci identified by Parker et al. as convergent using Δ SSLs actually had convergent substitutions in our data set. As mentioned earlier, one problem with the use of Δ SSLs is that it is an indirect measure of convergence. Although convergent substitutions will cause SSLs values to be greater for convergent topologies, they are not the only cause of changes in SSLs. For instance, the site-specific likelihood could be higher for the convergent topology because of divergent substitutions (sometimes called “parallel” substitutions) in the focal species, or because diffuse nonconvergent substitutions occurring anywhere on the tree differentially support the two topologies. Therefore, to further scrutinize this observation, we performed ancestral state reconstruction on the original alignments from Parker et al. (2013). For these 117 genes, we found 11 genes with convergent substitutions between dolphin and the echolocating bats in the suborder Yangochiroptera, and 2 genes with convergent substitutions between dolphin and the echolocating bats in the suborder Yinpterochiroptera. No genes contain convergent substitutions in all echolocating bat lineages and dolphin. Without the presence of convergent substitutions, there is clearly no evidence for convergent evolution, regardless of the Δ SSLs values calculated from these alignments.

Equally surprisingly, despite the claim in the main text of Parker et al. (2013) that the observed values of Δ SSLs and consequent signals of convergence “were not due to neutral processes,” we could find no statistical support for this statement. Close inspection of the supplementary materials

provided by the authors revealed no clear comparison between the observed values of Δ SSLs and the simulated values, and no test for an excess of “convergent topologies” (or any similar test) is reported. The only statistical statement about convergence presented in Parker et al. involves the number and identity of sensory genes, which we address next.

Assessing Levels of Convergence in Particular Gene Classes

The conclusions of Parker et al. (2013) rely heavily on the statistically significant level of “convergence” found among sensory genes. These authors classified 98 loci as sensory genes using Gene Ontology (GO) terms, and also included in their analysis 7 genes previously implicated to play a role in echolocation. Of these 105 genes, 11 fell in the top 5% of genes ranked by mean Δ SSLs uniting echolocating bats and dolphins. For the sound and vision categories, they report the significance of these genes being observed in the tail as $P = 0.041$ and $P = 0.07$, respectively, after carrying out permutations for each class of gene. The authors concluded that the results were indicative of adaptive convergence in these classes of genes. However, they did not do a similar analysis on a closely related pair of nonecholocating species.

We sought to re-evaluate the signatures of convergence reported for sensory genes. We first checked for GO-term enrichment among the 1,372 genes with convergent changes between microbat and dolphin specifically to see if any significant terms were related to hearing, deafness, vision, or blindness. After correcting for multiple tests, we found no GO terms enriched in genes that contain convergent substitutions between microbat and dolphin, including all sensory terms (the lowest nominal P value was 0.016). Next, we looked specifically at the 105 genes classified by Parker et al. as having a role in sensory perception, all of which were included in our full set of 6,400 genes. We found 22 sensory genes to contain convergent substitutions between microbat and dolphin, but we also found an equal number of sensory genes with convergent substitutions between microbat and cow (fig. 2b). These results imply that any signals of convergence at the molecular level are not necessarily driving convergent echolocation.

Conclusions

Recent studies of individual genes have revealed that molecular convergence is a more common and widespread phenomenon than previously thought (Christin et al. 2010; Stern 2013), including among echolocating mammals (Li et al. 2010; Liu et al. 2010; Davies et al. 2012; Shen et al. 2012). However, whether such signals can be found in genomic data remain an open question: It may either be that we lack statistical power to find convergence in genome-wide data (but see Hiller et al. 2012) or that phenotypic convergence is not accompanied by a large number of genes with molecular convergence. Our findings have highlighted the problems inherent in using methods that do not directly test for convergence in determining whether there has been a significant excess of convergence. With the appropriate use of an empirical null model

that examines convergent and divergent substitutions (Castoe et al. 2009), there is no evidence from a genome-wide comparison of echolocating species for exceptional signals of convergent adaptation. Further evidence for molecular convergence will need to come from careful examination of individual molecular changes (e.g., Liu et al. 2014), and the use of comparative methods that use an appropriate null model.

Materials and Methods

We obtained nucleotide sequences for the genes of the following 10 mammals from Ensembl v75: *Monodelphis domestica* (outgroup), *Loxodonta africana*, *Mus musculus*, *Homo sapiens*, *Callithrix jacchus*, *Vicugna pacos*, *Bos taurus*, *T. truncatus*, *P. vampyrus*, and *M. lucifugus*. A total of 6,400 one-to-one orthologs were extracted and aligned with MUSCLE (Edgar 2004) and ambiguous and gap regions removed. Gene trees were inferred using RAxML (Stamatakis 2006), and the average consensus method (Lapointe and Cucumel 1997) implemented in SDM (Crisuolo et al. 2006) was used to generate a species tree. This species tree was used for downstream analyses. The program r8s (Sanderson 2003) was used to generate the ultrametric tree shown in fig. 1a.

We performed ancestral sequence reconstruction using codeml in PAML v4.7 (Yang 2007) and counted the number of divergent and convergent substitutions. We define a convergent substitution as a change from the ancestral state to the same amino acid along two lineages. Following Castoe et al. (2009), changes along both branches at the same site that lead to different amino acids are classified as divergent. Testing for GO-term enrichment was done using Fisher's exact test with a Dunn–Sidak correction for multiple testing.

The alignments for the 117 genes classified as convergent by Parker et al. were provided by the authors. These were masked with Gblocks (Castresana 2000) (as per the settings outlined in the original paper) and stop codons were removed. We then performed ancestral reconstruction using the phylogenetic tree given in Parker et al. (2013), counting convergent substitutions along the branch leading to dolphin and either the branch leading to echolocating bats within the suborder Yinpterochiroptera (*Rhinolophus ferrumequinum* and *Megaderma lyra*) or the branch leading to the echolocating bats in the suborder Yangochiroptera (*M. lucifugus* and *Pteronotus parnellii*).

Positive selection was tested using the branch-site test in PAML, which compares the likelihood of a model in which prespecified (foreground) branches of a phylogeny are constrained from evolving with positive selection with respect to the rest of the tree (background branches) to a model that does not constrain these branches. If the model without constraint is significantly more likely than the one with, then we can conclude that these branches have experienced positive selection. We labeled microbat, dolphin, and cow as the foreground branches in three separate runs of PAML. A critical χ^2 value of 5.41 was used as the cutoff for the likelihood ratio test at a significance level of 0.01 (Yang and dos Reis 2011). Genes were considered as having adaptive convergence if they had a

convergent substitution and were found to be significant for the branch-site test in both species.

Data Availability

The data from our paper is now available through Dryad at: <http://datadryad.org/resource/doi:10.5061/dryad.16qc5>.

Acknowledgments

We thank J. Zhang for comments and J. Cotton et al. for constructive feedback. We also thank the Associate Editor and three anonymous reviewers for suggestions that helped to improve the manuscript. This work was supported by the Indiana University Genetics, Molecular and Cellular Sciences Training Grant from the National Institutes of Health (T32-GM007757) and National Science Foundation grant DBI-0845494 to M.W.H.

References

- Castoe TA, de Koning APJ, Kim H-M, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinsons CL, Pollock DD. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. *Proc Natl Acad Sci U S A*. 106:8986–8991.
- Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 17: 540–552.
- Christin PA, Weinreich DM, Besnard G. 2010. Causes and evolutionary significance of genetic convergence. *Trends Genet*. 26:400–405.
- Crisuolo A, Berry V, Douzery EJP, Gascuel O. 2006. SDM: a fast distance-based approach for (super)tree building in phylogenomics. *Syst Biol*. 55:740–755.
- Davies KTJ, Cotton JA, Kirwan JD, Teeling EC, Rossiter SJ. 2012. Parallel signatures of sequence evolution among hearing genes in echolocating mammals: an emerging model of genetic convergence. *Heredity* 108:480–489.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 32:1792–1797.
- Hiller M, Schaar BT, Indjelan VB, Kingsley DM, Hagey LR, Bejerano G. 2012. A “forward genomics” approach links genotype to phenotype using independent phenotypic losses among related species. *Cell Rep*. 2:817–823.
- Lapointe F-J, Cucumel G. 1997. The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. *Syst Biol*. 46:306–312.
- Lartillot N, Brinkmann H, Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol*. 7:S4.
- Li Y, Liu Z, Shi P, Zhang J. 2010. The hearing gene *Prestin* unites echolocating bats and whales. *Curr Biol*. 20:R55–R56.
- Liu Y, Cotton JA, Shen B, Han X, Rossiter SJ, Zhang S. 2010. Convergent sequence evolution between echolocating bats and dolphins. *Curr Biol*. 20:R53–R54.
- Liu Z, Qi F-Y, Zhou X, Ren H-Q, Shi P. 2014. Parallel sites implicate functional convergence of the hearing gene *prestin* among echolocating mammals. *Mol Biol Evol*. 31:2415–2424.
- Parker J, Tsagkogeorga G, Cotton JA, Liu Y, Provero P, Stupka E, Rossiter SJ. 2013. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* 502:228–231.
- Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19:301–302.
- Shen YY, Lu L, Li GS, Murphy RW, Zhang YP. 2012. Parallel evolution of auditory genes for echolocation in bats and toothed whales. *PLoS Genet*. 8:e1002788.
- Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
- Stern DL. 2013. The genetic causes of convergent evolution. *Nat Rev Genet*. 14:751–764.
- Williams PD, Pollock DD, Blackburne BP, Goldstein RA. 2006. Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comp Biol*. 2:e69.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24:1586–1591.
- Yang Z, dos Reis M. 2011. Statistical properties of the branch-site test of positive selection. *Mol Biol Evol*. 28:1217–1228.
- Zhang J, Kumar S, Nei M. 1997. Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. *Mol Biol Evol*. 14: 1335–1338.
- Zou Z, Zhang J. 2015. No genome-wide protein sequence convergence for echolocation. *Mol Biol Evol*. 32(5):1237–1241.